**International Trade Nowcasting Challenge Approach Description**

**Methodology *(Replicability; Scalability; Interpretability)***

Please provide a detailed description of the approach used to calculate the point estimates of the International Trade of Goods statistics. The description should contain (1) the data processing steps, (2) the methods and models used, (3) references to the scientific papers/sources that present the methods and models used, and (4) the time it took to process the data set and classify the job advertisements.

Bear in mind that the workflow will be also evaluated based on the criteria for the Reusability and Innovativity Awards.

*This section will be evaluated for:*

*(1) the Replicability criterion: likeliness that the described approach can successfully reproduce the solution submitted by the team for the Accuracy award*

*(2) the Scalability criterion: amount of modification required for the approach to apply to similar datasets on a potentially larger scale*

*(3) the Interpretability criterion: the extent to which a human could understand and articulate the relationship between the approach’s predictors and its outcome; how well the logical reasoning behind the model which is making the prediction is developed (whether it is mathematically and/or technically sound*

|  |
| --- |
| **A complete code file is also attached „code.R“ as well as a file for exploration, please find a description of the approach below:**  **1. Data Processing Steps**  **Data Acquisition** The code begins by loading Eurostat data through the eurostat::get\_eurostat function, targeting trade statistics as listed in the code file. This data acquisition process includes specifying regions and caching data locally for efficiency.  **Data Cleaning and Transformation** After loading, the clean\_data function is applied to preprocess the dataset. This function:   * Selects relevant columns (geo, time, values). * Performs standardization by calculating z-scores using a helper function scale2, which centers and scales the data by subtracting the mean and dividing by the standard deviation. * Applies a logarithmic transformation to the values for modeling purposes. * Adds columns representing the mean and standard deviation for each country, capturing the variation within each time series.   This multi-step transformation makes the data suitable for time series modeling, improving model stability and interpretability by reducing skew and standardizing value ranges.  **2. Methods and Models Used**  **Modeling Approach** The workflow employs an **ARIMA (Auto-Regressive Integrated Moving Average) model**, specifically chosen for its efficacy in forecasting time series data, as it accounts for both trend and seasonality. Using forecast::auto.arima, this model is trained on different data subsets to capture both short- and long-term trends.  **Subsetting for Cross-Validation** The process\_subset function runs models on subsets of the time series to assess performance over different time horizons. This approach uses **time series cross-validation** (tsCV) to calculate error metrics such as RMSE (Root Mean Square Error) over 2, 6, and 12-month horizons as well as for the model itself. This robust cross-validation procedure ensures that the model performs consistently and avoids overfitting, enhancing replicability.  **Point Estimate Calculation** Three primary point estimates are calculated for each country, grouped by export values:   1. **Entry1**: Best model based on minimum RMSE over a 2-month forecast horizon, mean of predictions from models built with actual and logarithmic data. 2. **Entry2**: Best model based on minimum RMSE over a 12-month forecast horizon, mean of predictions from models built with actual and logarithmic data. 3. **Entry3**: Combined best model based on the minimum values of overall RMSE and various forecast horizons (2, 6, 12 months), mean of predictions from models built with actual and logarithmic data.   Note: I did not use the scaled series because it appeared that they lead to the same results and models as the raw values during the testing phase. Logarithmic transformation, on the other side, gave quite different models and estimation windows. |
| **3. Scientific References**  The modeling and cross-validation approach align with standard practices in time series forecasting and statistical learning. Key references include:   * **Hyndman, R. J., & Athanasopoulos, G.** (2018). *Forecasting: Principles and Practice*. This textbook provides foundational approaches to time series forecasting, including ARIMA and cross-validation techniques. * **Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M.** (2015). *Time Series Analysis: Forecasting and Control*. The ARIMA model, foundational to this code, is thoroughly discussed in this book.   **4. Processing Time**  The code is designed for efficiency:   * **Data Loading**: Approximately a few minutes, depending on Eurostat server response time and the number of countries specified. * **Model Training and Evaluation**: With parallel processing enabled via mclapply and multi-core capabilities detected automatically, the model estimation typically completes in under an hour per data set and task. Processing times may vary based on the subset sizes and computational resources.   **Evaluation Criteria**  **Replicability** The approach is highly replicable, as it leverages Eurostat’s public data and widely used R libraries (e.g., eurostat, forecast). The workflow is modular, with each step’s input and output explicitly defined, enabling other researchers or practitioners to reproduce the results. The inclusion of cross-validation further strengthens replicability by providing objective model validation.  **Scalability** The workflow is designed to scale to larger datasets by:   * Allowing dynamic detection of system cores, optimizing processing via parallel computation. * Supporting modifications to the filtering criteria in get\_eurostat to target broader or more granular datasets. For instance, adjusting the frequency of analysis (e.g., weekly instead of monthly) or expanding the set of EU countries could be done with minimal code changes.   **Interpretability** The ARIMA model selection based on RMSE is interpretable, as it reflects both time series theory and real-world performance. The logarithmic transformation and standardization of data values enhance interpretability by bringing the predictors to a comparable scale, facilitating understanding of the model’s responses. |

**Architecture**

Please provide a description of the architecture of your approach. A diagram of the architecture is considered of additional value. Indicate what modifications would be required to apply the approach to similar datasets on a larger scale.

*This section will be evaluated for:*

1. *the Architecture criterion: evaluated based on its modules, their cohesion and their configurability; an architecture which is modular and includes clear connections between modules or components receives a higher score*

|  |
| --- |
| **Architectural Components**   1. **Data Loading and Configuration Module**    * **Files/Functions**: get\_eurostat, write\_rds    * **Description**: This module is responsible for fetching the Eurostat trade data, applying basic filtering parameters, and caching it locally. By setting up a consistent and configurable data pipeline (e.g., specifying regions, trade type, and data transformations), this module ensures that the data aligns with the requirements of the downstream modeling components. 2. **Data Cleaning and Preprocessing Module**    * **Files/Functions**: clean\_data, scale2    * **Description**: In this module, the loaded data undergoes preprocessing, including scaling, transformation, and feature engineering. The clean\_data function prepares the data by scaling values, applying log transformations, and calculating summary statistics, which improves model performance and interpretability. This module outputs a processed data frame suitable for ARIMA modeling. 3. **Modeling Module**    * **Files/Functions**: estimate\_arima\_models, process\_subset, auto.arima, tsCV    * **Description**: This is the core module, where time series modeling is performed using ARIMA models. The process\_subset function iteratively fits models on various data subsets, enabling cross-validation and error calculation. Additionally, parallel processing via mclapply in estimate\_arima\_models optimizes model fitting across multiple cores, reducing processing time. 4. **Model Evaluation and Selection Module**    * **Files/Functions**: entry1, entry2, entry3, mutate, group\_by, summarise    * **Description**: The evaluation module identifies optimal models based on RMSE scores over various forecast horizons (2, 6, and 12 months). Three entry types (Entry1, Entry2, and Entry3) are defined based on the minimum RMSE over these horizons, facilitating model selection for different accuracy requirements. Each entry is output as a CSV file for later aggregation. 5. **Export Module**    * **Files/Functions**: write\_csv, toJSON    * **Description**: This module exports the selected models’ forecasts as both CSV and JSON files. The JSON output is structured hierarchically by country, providing a clean, accessible format for integration with other systems or for reporting purposes. |
| Data Flow and Processing Diagram  Data Loading --> Preprocessing --> Modeling (Parallel) --> Evaluation --> Export |
| **Scalability Modifications**  To adapt this approach for larger or similar datasets, several modifications can enhance scalability and flexibility:   1. **Data Loading**: Extend the filter options to support additional or more granular time series data (e.g., daily rather than monthly) or to fetch multiple trade statistics simultaneously. 2. **Parallel Processing**: The code dynamically detects system cores, but for much larger datasets, it might be beneficial to leverage distributed computing (e.g., using R's parallel package across clusters or integrating with Spark via sparklyr). 3. **Modular Configuration**: The inclusion of a configuration file or environment variables for adjustable parameters (e.g., geographic scope, model types, and ARIMA parameters) would allow users to adapt the workflow without altering the core code. 4. **Dynamic Model Selection**: For even more extensibility, the model selection process could incorporate additional machine learning models (e.g., SARIMA, Prophet, or XGBoost) based on a configurable model pool, allowing users to expand beyond ARIMA.   These adjustments would facilitate the application of this architecture to similar large-scale trade datasets, enabling high reusability and flexibility across domains. |

**List of Data Sources with Descriptions**

For each country, list the data sources (and their description) that were used to calculate the point estimates for the selected country. Please use the template below to provide the information for each source. If multiple data sources were used, please copy paste the template below and fill it in.

Bear in mind that the data sources will also be evaluated based on its openness, availability, coverage and consistency.

**1. Eurostat time Series**

**Number of data points collected from the data source (for each reference period)**

|  |  |
| --- | --- |
| October 2024 | Eurostat time series: 8624 observations |
| November 2024 |  |
| December 2024 |  |
| January 2025 |  |
| February 2025 |  |
| March 2025 |  |

**Structure of the data used to predict the point estimates**

|  |  |
| --- | --- |
| **Attribute Name** | **Attribute Description** |
| geo | Country code, indicating the reporting country |
| time | Time of the reported trade values, in monthly increments |
| values | Export values, in millions of euros, non-seasonally adjusted |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* **Coverage**: Covers all EU member states for the years relevant to the nowcasting process, with data available up to the most recent month of reporting.
* **Openness and Availability**: The Eurostat trade data is publicly accessible and available for download without restrictions, meeting high standards of openness and accessibility.
* **Consistency**: Eurostat data is consistently updated, quality-controlled, and adheres to EU statistical regulations, ensuring accuracy and reliability across reporting periods and member states.

**Hardware Specifications *(Replicability; Scalability; Interpretability)***

Please describe the hardware specifications of the machines that were used to run the methodology.

*This section will be evaluated for:*

*(1) the Replicability criterion*

*(2) the Scalability criterion*

*(3) the Interpretability criterion*

**Machine 1**

|  |  |
| --- | --- |
| CPUs | Apple Silicon M2 Pro, 10 shared cores, 16 GB Memory |
| GPUs | Apple Silicon M2 Pro, 10 shared cores, 16 GB Memory (not needed for current approach) |
| TPUs | TPU name and capacity |
| Disk space | 8 MB |

**Libraries *(Maintainability)***

Please provide the libraries used for approach, if any, as well as the links to these libraries, if available.

*This section will be evaluated for:*

*(1) the Maintainability and openness criterion: use of libraries which are regularly maintained will yield higher scores. (Examples include pytorch, tensorflow, scikit-learn, pandas, numpy, etc.) The use of libraries which are openly available will yield higher scores.*

|  |
| --- |
| The following R libraries were used in the approach, focusing on data manipulation, statistical modeling, and parallel processing. All libraries are open-source, widely maintained, and available on CRAN, ensuring maintainability and openness.   1. **tidyverse**    * **Description**: A collection of R packages for data science that simplifies data manipulation, visualization, and programming. It includes dplyr, ggplot2, and other packages.    * **Link**: <https://www.tidyverse.org/>    * **Maintainability**: Highly maintained with regular updates, and it is well-documented with extensive community support. 2. **eurostat**    * **Description**: Provides access to Eurostat’s statistical data, allowing users to directly query and download datasets, including the international trade statistics used here.    * **Link**: <https://cran.r-project.org/web/packages/eurostat/index.html>    * **Maintainability**: Actively maintained, with periodic updates to align with Eurostat’s data sources, ensuring reliability in accessing up-to-date data. 3. **forecast**    * **Description**: A robust package for time series forecasting, including ARIMA modeling, automatic model selection (auto.arima), and error evaluation.    * **Link**: <https://cran.r-project.org/web/packages/forecast/index.html>    * **Maintainability**: Regularly updated with active community and developer support, making it a dependable choice for time series analysis. 4. **parallel**    * **Description**: Part of R’s core package suite, parallel enables multi-core and distributed processing, essential for handling large datasets and improving processing efficiency.    * **Link**: <https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf>    * **Maintainability**: Supported as a core R package, parallel is highly stable and maintained alongside R’s primary development. 5. **jsonlite**    * **Description**: Facilitates JSON handling in R, providing tools to read, write, and parse JSON data structures, used here for exporting results.    * **Link**: <https://cran.r-project.org/web/packages/jsonlite/index.html>    * **Maintainability**: Regularly updated with strong community support, ensuring reliable integration of JSON data formats.   These libraries contribute to the approach's maintainability by providing essential, well-supported tools for data handling, statistical modeling, and export functionality. |

**Open license *(Maintainability)***

Please provide the open license of the provided code, if any.

*This section will be evaluated for:*

*(1) the Maintainability and openness criterion: whether the approach is open and under an open licens*e

|  |
| --- |
|  |

**Similarities/differences to State-of-the-Art techniques *(Originality)***

Please provide a list of similarities and differences between the used methodology and to the state-of-the-art techniques.

*This section will be evaluated for:*

*(1) the Originality of the approach criterion: compare the approach used to the state-of-the-art; the extent to which the submission represents an improvement over these pre-existing approaches*

|  |
| --- |
| **Similarities to State-of-the-Art Techniques**   1. **ARIMA Modeling for Time Series**    * **Description**: ARIMA (Auto-Regressive Integrated Moving Average) is a classic, well-established model in time series forecasting due to its effectiveness in handling seasonal and trend-based patterns. It remains a frequently used baseline method for economic forecasting.    * **Comparison**: State-of-the-art methods, especially in short-to-mid-term forecasting, often use ARIMA models or seasonal ARIMA (SARIMA) variants due to their transparency and reliable performance in structured time series like trade data. This approach’s use of auto.arima aligns with standard practices. 2. **Cross-Validation for Model Selection**    * **Description**: The approach uses time series cross-validation (tsCV) to evaluate and select the best-fitting model based on RMSE at different horizons (2, 6, and 12 months). This method assesses model robustness across several subsets.    * **Comparison**: Cross-validation is a standard technique for hyperparameter tuning in machine learning models. In time series, cross-validation is especially critical to avoid overfitting, and this project’s focus on multiple horizons aligns with advanced, horizon-specific performance assessment techniques used in state-of-the-art methodologies. 3. **Parallel Processing for Scalability**    * **Description**: The code leverages the parallel package to handle subset processing across multiple cores, speeding up model training and evaluation.    * **Comparison**: Parallelization is common in high-performance computing solutions for time series forecasting, and state-of-the-art methods often use multi-core processing (or distributed computing) to manage large datasets efficiently. This setup aligns well with those standards. |
| **Differences and Innovations**   1. **Hierarchical Cross-Validation for Entry-Based Selection**    * **Innovation**: The approach’s three-entry system (Entry1, Entry2, and Entry3) allows for specific model selection based on minimum RMSE values across different horizons (2, 6, and 12 months) as well as the model’s RMSE. Each entry provides a unique perspective on model accuracy, enhancing the adaptability of forecasts to varying timeframes.    * **Comparison**: Most state-of-the-art models apply a single model across all horizons or optimize strictly for long-term or short-term accuracy. This hierarchical, entry-based approach is more granular, offering flexibility in selecting models based on forecast needs—a key innovation for trade nowcasting. 2. **Simplified Transformation Pipeline with Standardization and Log Scaling**    * **Innovation**: The project applies standardization and log scaling as simple yet effective transformations, rather than using complex feature engineering or advanced deep learning architectures. This maintains interpretability and prepares the data efficiently for ARIMA modeling.    * **Comparison**: Modern state-of-the-art techniques, especially in machine learning, often rely on deep learning models (e.g., LSTMs, Transformers) which can capture complex, non-linear patterns in the data. However, this approach balances simplicity and interpretability with effective transformations that are computationally lighter. This focus on simplicity over deep learning models can be an advantage when interpretability is crucial, as it is in economic forecasting. 3. **ARIMA Model as a Transparent, Interpretable Baseline in Trade Nowcasting**    * **Innovation**: While many current models in time series forecasting have shifted toward machine learning or neural networks, this approach maintains ARIMA’s traditional statistical foundation. This choice provides better interpretability, particularly valuable for policymakers and trade analysts.    * **Comparison**: State-of-the-art time series forecasting has shifted toward non-linear, data-driven models (e.g., Prophet, LSTM). However, this project’s ARIMA-based method remains robust and is complemented by transformations that improve model stability, providing a more transparent alternative to black-box models. This is a notable difference, as it maintains clarity in how forecasts are generated. |
| 1. **Country-Specific and Series-Specific Estimation Strategy**    * **Innovation**: The estimation function estimate\_arima\_models applies models specifically to each country and data series (e.g., original or log-transformed values), enabling country-specific and series-specific model fitting. This tailored modeling approach respects individual country trends and series characteristics.    * **Comparison**: While many modern models are designed to generalize across multiple regions or variables simultaneously, this approach’s segmentation by country and data series aligns more closely with national statistical office methodologies. This level of segmentation is original in that it optimizes forecasts for each country individually, which is often missing in generalizable models.   **Summary of Originality**  The originality in this approach comes from balancing traditional ARIMA-based modeling with innovative, efficient, and interpretable preprocessing and model selection techniques. While many state-of-the-art techniques in time series nowcasting rely on machine learning, this approach favors statistical transparency, country-specific adaptation, and error-based entry selection, providing a reliable, scalable, and interpretable solution for trade statistics nowcasting. |

**Contribution to scientific field *(Future orientation)***

Please describe how your submission contributed to the scientific field, what impact it could have and what could potentially be future work to improve the solution.

*This section will be evaluated for:*

*(1) the Future orientation and impact criterion: the potential effect of the approach used will be evaluated; this includes the scale of impact it has on the problem of nowcasting; the impact will be evaluated based on potential efficiency improvements and cost reductions.*

|  |
| --- |
| **Impact on the Field of Nowcasting**   1. **Enhanced Forecasting Efficiency and Practical Application**    * **Description**: The approach focuses on trade nowcasting for EU countries, leveraging accessible data from Eurostat and straightforward statistical models to produce accurate, scalable forecasts. By using ARIMA models and modular, efficient code, this method is computationally lighter and cost-effective compared to many machine learning-based approaches. This makes it an ideal solution for institutions and researchers seeking timely, interpretable trade forecasts with limited computational resources.    * **Impact**: This methodology can be readily applied by government agencies, economists, and trade analysts in need of high-frequency, interpretable forecasts, providing them with a cost-efficient alternative that is easy to replicate and understand. Its design promotes public sector adoption, particularly in statistical offices and organizations that rely on monthly or quarterly trade data to make policy decisions. 2. **Improved Scalability and Customization for Country-Level Analysis**    * **Description**: By structuring the analysis to accommodate each country’s specific economic patterns, this approach acknowledges the heterogeneity in EU member economies, delivering more accurate, customized forecasts. This segmentation promotes scalability, as the modular design can be adapted to new countries or additional trade indicators with minimal changes.    * **Impact**: In economic analysis, especially when countries’ trade conditions vary significantly, tailored forecasting models such as these improve data reliability and relevance. This method’s scalability allows it to be adapted beyond the EU context, potentially benefiting other trade blocs or organizations that require individualized forecasts. 3. **Contribution to Transparent and Interpretable Economic Forecasting**    * **Description**: ARIMA, while traditional, is known for its transparency and interpretability, qualities that are increasingly valuable as machine learning models become more complex. This method balances robust statistical rigor with clarity, enhancing the interpretability of forecasts for non-specialist stakeholders.    * **Impact**: Transparent methodologies like ARIMA-based nowcasting enhance trust in economic forecasting by providing users with a clear understanding of model outputs. For researchers, policymakers, and data scientists, this approach can serve as a foundational model that can be easily benchmarked and further improved upon. |
| **Future Work and Potential Enhancements**   1. **Incorporating Machine Learning for Hybrid Approaches**    * **Potential Enhancement**: Future work could integrate machine learning models, such as LSTM (Long Short-Term Memory) networks or Prophet, in a hybrid approach with ARIMA. These models capture non-linear patterns and seasonal variations, potentially improving accuracy for highly volatile or irregular time series data.    * **Expected Impact**: A hybrid model would retain ARIMA’s interpretability while incorporating the flexibility of machine learning, potentially enhancing both forecast accuracy and responsiveness to sudden economic shifts. 2. **Expanding Cross-Validation and Model Selection Techniques**    * **Potential Enhancement**: This method uses time series cross-validation, but exploring adaptive validation techniques that adjust to real-time trade data fluctuations (e.g., rolling windows or dynamic adjustment) could improve model reliability, especially for monthly or high-frequency data.    * **Expected Impact**: Enhanced cross-validation could allow the model to adapt dynamically to recent data, making it more responsive to current economic trends and less prone to issues with outlier periods, which is valuable for near-term economic forecasting. 3. **Adding External Macroeconomic Indicators as Predictors**    * **Potential Enhancement**: Adding macroeconomic predictors, such as exchange rates, inflation, or global economic indicators, could improve the model’s responsiveness to contextual factors affecting trade. For example, using these predictors in a vector autoregression (VAR) model could yield deeper insights into the interactions between trade and broader economic conditions.    * **Expected Impact**: By incorporating external factors, this approach would become more holistic, capturing the impact of global economic shifts on trade, which is particularly relevant in increasingly interconnected economies. This could lead to more accurate and timely forecasts that respond to macroeconomic events. 4. **Implementing Real-Time Data Collection and Automated Forecasting Pipelines**    * **Potential Enhancement**: Automating data collection, cleaning, and model deployment through an integrated pipeline could create a continuous nowcasting system for trade statistics, capable of providing near real-time updates as new data becomes available.    * **Expected Impact**: This automation would allow for a substantial improvement in efficiency, reducing manual intervention and enhancing timeliness. By continuously updating predictions, it would create a proactive forecasting system, aiding policy and business decisions in near real-time. |
| **Summary of Future Orientation**  By presenting a transparent, efficient, and highly adaptable model for nowcasting EU trade data, this approach provides a foundation for both immediate use and future expansion. Its impact lies in its ability to offer robust, accessible, and cost-effective trade forecasting, with clear potential for further research and enhancements through hybrid modeling, automation, and expanded predictors. These future-oriented improvements position this methodology as a valuable contribution to the fields of economic forecasting and trade analytics, facilitating more responsive and data-driven decision-making. |

**Lessons Learned *(Future orientation)***

Please state any lessons learned during the competition.

*This section will be evaluated for:*

*(1) the Future orientation and impact criterion: what were the lessons learnt during the competition, and what could potentially be future work to improve the solution.*

|  |
| --- |
| * Lessons Leaned   + Importance of Data Quality and Preprocessing   + Model Transparency Matters for End-Users   + Computational Efficiency with Parallel Processing   + Cross-Validation Techniques Are Key to Robust Forecasts * Potential Future Improvements   + Implementing Real-Time Data Pipelines   + Hybrid Models with Machine Learning Components   + Use of Additional Macroeconomic Indicators   + Cloud-Based Processing for Scalability |

**Short description of the Team – area of expertise**

Please provide a description of the team, your area of expertise and contact information.

|  |
| --- |
| Christian Url - Data Scientist, MSc in Statistics, [christian.url@protonmail.com](mailto:christian.url@protonmail.com), Experience in IT Consulting, Prediction of up-times for medical equipment, Fraud detection |