Estimating Missing Trade Data: WTO

Group 9: 46476, 37984, 50250

Abstract

The global economy relies fundamentally on international trade, yet incomplete bilateral reporting, stemming from statistical limitations, confidentiality requirements, and uneven national capabilities, continues to impede robust analysis and effective policy development. This paper evaluates four complementary approaches to estimating missing trade flows: (i) mirror statistics, which use reciprocal partner data to reconstruct unreported exchanges; (ii) gravity models, grounded in economic theory and structural relationships; (iii) random forest algorithms, employing ensemble-based non-parametric learning to capture complex interactions; and (iv) spatio-temporal graph convolutional networks (STGCNs), designed to represent trade as a networked system evolving over time. We examine the theoretical underpinnings, model specifications, estimation procedures, and empirical performance of each approach, with a focus on their comparative strengths, limitations, and potential synergies in addressing systematic gaps in global trade data. The empirical analysis shows that the STGCN method delivers the best performance at the product level, while the random forest approach achieves the highest accuracy at the country level. The findings yield a set of practical, scalable, and theoretically sound tools for improving the completeness and reliability of international trade statistics in an era of increasingly complex global commerce.

1. Introduction

International trade data form the backbone of economic analysis, policy formulation, and global commerce understanding. From Adam Smith's foundational work on comparative advantage to modern assessments of supply chain resilience, accurate bilateral trade flow information has been essential for comprehending economic relationships between nations. Yet despite the critical importance of complete trade data, systematic gaps in bilateral trade reporting persist across countries, time periods, and product categories, fundamentally limiting our ability to understand global economic patterns and formulate evidence-based trade policies.

The challenge of missing trade flow data is both pervasive and consequential. Missing trade flow data pervades international trade datasets, arising from multiple sources including statistical reporting limitations, confidentiality restrictions, flows below official thresholds, and genuine absence of commercial relationships between trading partners. Some countries demonstrate systematic patterns of underreporting across multiple partners and time periods, while others exhibit sporadic gaps that may reflect economic, political, or administrative factors. The problem is particularly pronounced for developing countries, where limited statistical capacity compounds existing reporting challenges, and for detailed product-level classifications, where the complexity of modern trade relationships strains traditional data collection systems. These gaps are not merely academic inconveniences but represent fundamental obstacles to understanding global economic integration, measuring the effects of trade policies, and identifying patterns of economic development. This systematic absence of information creates what Teti (2024) describes in the context of tariff data as a "scandal and a puzzle", a situation where critical economic data remain unavailable despite their fundamental importance for research and policy.

The implications extend far beyond data completeness. Missing trade data introduce systematic biases in empirical analyses, affecting estimates of trade elasticities, assessments of regional integration effects, and evaluations of trade policy impacts. When researchers rely on available data while ignoring systematic patterns of missingness, their conclusions may reflect the characteristics of well-reporting countries rather than global trade patterns. This selection bias is particularly problematic for understanding trade relationships involving developing countries, small economies, or specialized product categories where data gaps are most pronounced.

Moreover, the interconnected nature of global trade networks means that missing bilateral relationships affect our understanding of multilateral trade patterns. When Country A's trade with Country B is unobserved, this absence influences calculations of Country A's total trade exposure, Country B's market access patterns, and the overall structure of regional trade networks. The resulting analytical

blind spots can lead to misguided policy interventions and incomplete assessments of economic integration processes.

Contemporary policy challenges amplify the urgency of addressing missing trade data. The COVID-19 pandemic highlighted the critical importance of understanding supply chain vulnerabilities, yet gaps in trade flow data limit our ability to map these relationships comprehensively. Similarly, ongoing discussions about economic decoupling, nearshoring, and strategic trade dependencies require complete information about bilateral trade patterns to inform effective policy responses. Without systematic approaches to address missing data, policymakers operate with incomplete pictures of global economic relationships.

The methodological challenges are equally significant. Traditional approaches to handling missing data, such as listwise deletion or simple imputation, prove inadequate for the complex structure of international trade relationships. Trade flows exhibit strong temporal dependencies, reflect geographic and cultural proximity effects, and respond to policy interventions in ways that simple statistical methods cannot capture. Furthermore, the high-dimensional nature of modern trade datasets, with thousands of product categories across hundreds of country pairs over multiple decades, creates computational challenges that require sophisticated methodological approaches.

Recent developments in machine learning and network analysis offer promising avenues for addressing these challenges. The emergence of ensemble methods, graph neural networks, and sophisticated imputation algorithms provides new tools for leveraging the rich structure of trade data to estimate missing relationships. These approaches can incorporate the complex interdependencies inherent in trade networks while maintaining computational efficiency for large-scale datasets.

However, methodological innovation alone is insufficient without attention to the economic foundations of trade relationships. The gravity model of international trade, despite its limitations, provides crucial insights into the fundamental drivers of bilateral trade flows. The principle of mirror statistics, that exports from one country should correspond to imports by its partner, offers additional structure for validation and estimation. Modern approaches must combine these economic insights with statistical sophistication to produce reliable estimates of missing trade flows.

This study addresses the missing trade data challenge through a comprehensive methodological framework that combines traditional economic insights with modern computational techniques. We systematically examine three complementary approaches: enhanced gravity models with clustering methodologies, random forest algorithms with boosting methods, and graph neural networks that capture network effects in trade relationships. Our approach leverages a comprehensive dataset of over 200 million bilateral trade observations spanning 1996-2024, providing scope for developing and evaluating missing data estimation methods.

The remainder of this paper is organised as follows. **Section 2** reviews the relevant literature, focusing on gravity models, machine learning methods, and network-based approaches to missing trade data estimation. **Section 3** describes the datasets and preprocessing procedures, outlining the scope, coverage, and quality of the available trade flow information. **Section 4** provides an overview of the exploratory data analysis conducted on the study datasets. **Section 5** details the methodological framework, including the specification and implementation of the three estimation approaches. **Section 6** presents the empirical analysis and results, evaluating the relative performance of the models and offering practical insights for applied researchers. **Section 7** discusses the implications of the findings, their limitations, and potential avenues for improvement. Finally, **Section 8** concludes with key lessons learnt and directions for future research.

As global trade grows increasingly complex and policy challenges become more acute, the demand for complete, accurate, and timely trade flow information has never been greater. This study contributes to meeting that demand by proposing and rigorously evaluating robust, theoretically grounded, and practically applicable methods for estimating missing bilateral trade relationships.

2. Literature Review

2.1. Mirror Statistics

Mirror statistics represent a fundamental methodological approach in international trade analysis, predicated on the principle that bilateral trade flows constitute two perspectives of identical economic transactions. The fundamental assumption underlying mirror statistics analysis is that exports from Country A to Country B should theoretically correspond to imports by Country B from Country A, with any systematic discrepancies potentially indicating data quality issues, structural factors, or deliberate misreporting practices.

The systematic application of mirror statistics in trade analysis emerged in the 1960s through pioneering work by Bhagwati (1964), who first recognized the analytical potential of comparing bilateral trade flows from both reporting perspectives. Bhagwati's seminal 1964 study examined how over- or under-invoicing of trade biased balance of payments data, establishing the foundational methodology for using mirror statistics to detect trade misinvoicing. This early work demonstrated that systematic discrepancies between mirror flows could reveal patterns of customs evasion, particularly for goods subject to high tariff rates.

The core methodological challenge in mirror statistics lies in reconciling the inherent differences between import and export reporting practices. Import values are typically reported on a Cost, Insurance, and Freight (CIF) basis, incorporating transportation and insurance costs, while exports are recorded Free on Board (FOB), excluding these additional charges. This fundamental difference necessitates sophisticated adjustment procedures to enable meaningful comparisons between mirror flows.

The most significant advancement in mirror statistics methodology came through the development of the BACI (Base pour l'Analyse du Commerce International) database by Gaulier & Zignago (2010). Their methodology addresses the core challenges of mirror statistics through a two-step reconciliation process: first, CIF costs are estimated and removed from import values using regression analysis of observed CIF/FOB ratios against product-specific world median unit values; second, the reliability of each country as a trade data reporter is assessed through decomposition of absolute values of mirror flow ratios using weighted variance analysis. This approach transforms the fundamental limitation of mirror statistics, having two different figures for the same flow, into a methodological advantage for improving data quality and coverage.

2.1.1. Applications in Missing Data Estimation

Mirror statistics have proven particularly valuable for addressing missing trade flow data, a critical challenge in international trade analysis. When some countries do not communicate their international trade data at all, it becomes possible to build estimates of their trade patterns using mirror statistics from their trading partners. It is generally assumed that import figures would be more reliable, as these form the basis for import duties and tax calculations, and thus receive more official scrutiny.

However, the literature reveals significant challenges in this approach. Research by Yeats (1995) demonstrated that mirror methods performed poorly at disaggregated commodity levels, suggesting that discrepancies between mirror statistics reflected more than simple recording errors or methodological differences. These findings indicate that the application of mirror statistics for missing data estimation requires careful consideration of product-level and country-specific factors that may affect data reliability.

Linsi & Mügge (2023) identifies multiple categories of sources of discrepancies relevant to missing data estimation methodologies:

Structural and Unavoidable Factors Mirror discrepancies arise in part from structural factors such as CIF/FOB valuation differences, timing mismatches in shipment recording, exchange rate effects, and varying reporting thresholds for small transactions.

Statistical Capacity Differences in statistical capacity also contribute, as some countries have better data systems than others and crises or conflicts can disrupt reporting.

Deliberate Misreporting and Fraud High tariffs encourage import under-invoicing, while export subsidies promote over-invoicing, especially under complex tariff regimes.

In general, the literature positions mirror statistics as both a diagnostic and corrective tool in international trade analysis: capable of identifying structural biases, highlighting weaknesses in national statistical systems, and informing imputation strategies for missing data. While early applications emphasized detecting misinvoicing and fraud, more recent developments, such as BACI, have integrated reconciliation procedures that improve the completeness and comparability of trade datasets. These advances illustrate the potential of mirror flows to enhance trade data quality, while also highlighting the complexities involved in reconciling CIF/FOB differences, timing mismatches, and other structural sources of discrepancy.

In the present study, a more basic mirror approach is employed, without the multistep reconciliation and reliability weighting procedures seen in BACI. This choice reflects the goal of maintaining methodological transparency and avoiding strong assumptions that may not hold in all combinations of country product - year. At the same time, the limitations identified in the literature, particularly around valuation differences, statistical capacity, and reporting incentives, remain important considerations when interpreting the results.

2.2. Gravity Model

The gravity model of international trade draws its foundational intuition from Newton's law of universal gravitation, where trade flows between countries are hypothesized to be positively related to their economic masses, namely the Gross Domestic Product (GDP) and negatively related to the distance between them. Tinbergen (1962) and Pöyhönen (1963) independently proposed this approach, suggesting that bilateral trade flows could be explained by a simple relationship:

$$F_{ij} = G \cdot \frac{M_i \cdot M_j}{D_{ij}}$$

Where F_{ij} represents trade flows from country i to country j, M_i and M_j denote the economic masses of the trading partners, D_{ij} captures the distance between them and G is a gravitational constant. This initial formulation, while empirically successful, lacked rigorous theoretical foundations.

Head & Mayer (2014) comprehensively review the established determinants that form the foundations of traditional gravity specifications. The log-linear transformation of the gravity equation has become the standard empirical approach:

$$\ln(F_{ij,t}) = \ln(G) + \beta_1 \ln(GDP_{i,t}) + \beta_2 \ln(GDP_{j,t}) - \beta_3 \ln(D_{ij}) + \varepsilon_{ij,t} \quad (1)$$

- **Economic Size (GDP)**: Captures both supply capacity and demand potential
- **Geographic Distance**: Serves as the primary proxy for transportation costs and trade barriers
- $\ln(G)$: The logarithm of the gravity constant, estimated as the intercept β_0

This log-linear specification allows for straightforward estimation of the gravity constant through exponentiation: $G = e^{\beta_0}$.

However, practical applications for missing data estimation often rely on simplified log-linear specifications that can be estimated reliably with available data. The challenge lies in developing robust methods for estimating and applying gravity constants across different country pairs and time periods. This practical need motivates approaches that focus on clustering country pairs with similar characteristics to share common gravity constants, enabling prediction of missing trade flows for pairs where direct estimation is not feasible.

2.2.1. Practical Implementation

Traditional gravity models estimated via log-linear OLS face challenges under heteroskedasticity and zero trade flows, leading to biased estimates Santos Silva & Tenreyro (2006). While solutions such as Poisson Pseudo-Maximum Likelihood (PPML) and selection models address these issues, they often increase computational complexity. Recent research seeks to retain the interpretability of gravity models while improving predictive performance using data-driven techniques.

One promising line of inquiry integrates clustering into gravity model estimation. Bobková (2014) applies K-means clustering to group countries based on gravity-relevant features and estimates separate models within each cluster, revealing significant heterogeneity in coefficient estimates-highlighting the limitations of pooled estimation when key determinants vary systematically across groups.

Another strand explores the fusion of gravity models with machine learning to enhance prediction. For instance, supervised ML tools like random forests, boosting, and neural networks have been used to model bilateral trade flows using standard gravity regressors, demonstrating superior performance over purely econometric specifications (Gopinath, 2020). These approaches capture complex nonlinear relationships and interactions that traditional gravity models may overlook.

These studies underscore a methodological shift: adapting the gravity framework to incorporate clustering and ML for increased flexibility and predictive accuracy. They serve as direct inspiration for the clustering-based, prediction-focused methodology introduced later in this report.

2.3. Random Forest

Random forest algorithms constitute a class of ensemble learning methods that have become increasingly prominent in empirical economic research. As a non-parametric modeling approach, random forests diverge from conventional econometric techniques by focusing on prediction rather than inference. Developed by Breiman Breiman (2001), the random forest algorithm constructs a multitude of decision trees during training and outputs the average of the predictions for regression tasks, or the majority vote in classification contexts. This ensemble framework enhances prediction accuracy and mitigates the risk of overfitting, a common limitation in single decision tree models.

In the field of international trade, random forests have demonstrated considerable utility, particularly in contexts involving high-dimensional, sparse, or noisy data. Trade datasets often contain thousands of product-level entries, typically coded under the Harmonized System (HS) at the 6-digit level, and cover bilateral flows across numerous countries and time periods. Traditional econometric models face substantial challenges in processing such data due to multicollinearity, functional form assumptions, and constraints on variable interactions. By contrast, random forest models accommodate both linear and nonlinear relationships, require fewer assumptions regarding the data-generating process, and can automatically model complex interaction effects among predictors. Athey & Imbens (2019)

A notable application by Tiits et al. Tiits et al. (2024) demonstrates how random forest algorithms can be integrated into gravity-based trade models to improve predictive accuracy and capture latent dimensions such as product complexity and relatedness. Their approach enables the modeling of bilateral trade flows at a highly granular product level, providing enhanced explanatory power compared to traditional gravity models.

In a random forest with B trees, the prediction for a feature vector $\mathbf{x}$ in a regression task is computed as the average over all individual tree predictions:

$$\hat{f}_{\text{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(\mathbf{x}),$$

while for classification, the final output is determined by majority voting. Each decision tree in the ensemble is trained to minimize a node-specific impurity measure, such as Mean Squared Error (MSE):

$$\text{MSE}(S) = \frac{1}{|S|}\sum_{i\in S}\left(y_i - \bar{y}_S\right)^2,$$

where S is a node containing a subset of training samples and $\bar{y}_S$ is the mean of the target values in that node.

Random forests also offer an internal feature importance measure based on the mean decrease in impurity (MDI):

$$\text{Importance}(x_j) = \sum_{t=1}^{T}\sum_{s\in S_t : v(s)=x_j}\frac{N_s}{N}\Delta i(s),$$

where $\Delta i(s)$ denotes the decrease in impurity from splitting on feature x_j at node s, N_s is the number of samples at that node, and N is the total number of samples.

For datasets with missing entries, the MissForest algorithm uses an iterative procedure to impute missing values. At each iteration t, for feature j, a random forest is trained on the observed values to predict the missing entries:

$$X_{ij}^{(t)} = \hat{f}_j^{(t)}(\mathbf{x}_{i,-j}), \quad \text{for all } (i,j) \text{ where } X_{ij} \text{ is missing.}$$

This process continues until convergence based on a stopping criterion such as minimal change in imputed values.

A central advantage of random forests lies in their ability to handle incomplete or missing data. This characteristic is particularly relevant in international trade analysis, where gaps in data reporting are frequent due to confidentiality restrictions, unrecorded transactions, or inconsistent classification practices. The MissForest algorithm, introduced by Stekhoven and Bühlmann Stekhoven & Bühlmann (2011), extends the random forest framework to missing data imputation. Through an iterative process, MissForest trains random forest models on the observed portions of the data and uses the fitted model to impute missing values. Subsequent iterations refine these imputations as the quality of the training data improves, resulting in enhanced imputation accuracy over time.

In addition to imputation, random forests offer built-in feature importance metrics, which can be used to identify the most influential predictors in a dataset. This capability is typically based on measures such as mean decrease in impurity or permutation importance, which assess the contribution of each variable to predictive accuracy across the ensemble of trees Strobl et al. (2008). Feature selection based on these metrics is particularly advantageous when working with large-scale trade datasets that include a mix of economic, geographic, political, and temporal variables. Moreover, random forests are capable of handling both continuous and categorical data, making them well-suited for the heterogeneity of trade-related variables.

Nevertheless, the application of random forests is not without limitations. One notable concern is the interpretability of the model. Unlike traditional regression techniques, which provide explicit coefficient estimates and confidence intervals, random forests operate as black-box models. This opacity may hinder their applicability in policy settings where transparency and causal interpretation are required. While random forests are generally efficient in handling large datasets, interpretability remains a key consideration when applied in decision-making contexts.

Despite these limitations, random forest algorithms represent a valuable methodological advancement in the study of international trade. Their capacity to manage large, complex datasets, handle missing values effectively, and deliver high predictive accuracy positions them as a powerful complement to traditional econometric approaches. As machine learning continues to gain prominence within economics, the use of random forest methods is likely to expand, particularly in applied contexts where data quality and model performance are of central concern. Silva et al. (2024)

2.4. Graph Neural Networks

2.4.1. Theoretical Foundations

Graph Neural Networks (GNNs) represent a cutting-edge methodological advancement in the analysis of international trade flows, driven by the inherently networked structure of global commerce (Zhou et al., 2021). Traditional models, such as the gravity model, have long been used to explain trade patterns through economic mass and geographic distance. However, these models often fail to fully capture the complex interdependencies that characterise modern trade networks. GNNs are particularly well-suited to this domain, as they can process non-Euclidean, graph-structured data, making them ideal for modelling the relational nature of international trade (Sellami et al., 2024). Their ability to learn from both node-level and edge-level features enables the identification of latent patterns embedded in trade networks, including geopolitical ties, regional integration, and supply chain interdependencies. For a more comprehensive overview of graph-based machine learning methods, the reader is referred to surveys such as Hamilton et al. (2017).

Given a simple graph $G = (V, X, A, E)$ containing a single type of relationship and no self-loops, individual entities i.e., countries, are represented as nodes V with real-valued features. The node feature matrix is denoted by $X \in \mathbb{R}^{|V| \times D}$, where $|V|$ is the number of nodes and D is the dimensionality of the node features.

The adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$ is a dense square matrix, where $a_{ij} = 1$ if there exists an edge between nodes i and j, and 0 otherwise. The edge weights are represented by the matrix E, whose entries correspond to weighted connections between nodes, i.e., a_{ij} values.

Graph Neural Networks (GNNs) operate under the message-passing paradigm, where each node iteratively updates its representation by aggregating features from its neighbors. Over T layers, this process allows nodes to learn both local and global structural information from the graph.

At each iteration $t \in \{1, \ldots, T\}$, the representation of a node v is updated as follows:

$$\mathbf{m}_v^{(t)} = \textsc{Aggregate}^{(t)} \left(\left\{ \mathbf{h}_u^{(t-1)} : u \in \mathcal{N}(v) \right\} \right),$$
$$\mathbf{h}_v^{(t)} = \textsc{Update}^{(t)} \left(\mathbf{h}_v^{(t-1)}, \mathbf{m}_v^{(t)} \right),$$

where:

- $\mathbf{h}_v^{(t)}$ is the embedding of node v at iteration t,
- $\mathbf{h}_v^{(0)} = \mathbf{x}_v$ is the initial node feature,
- $\mathcal{N}(v)$ denotes the neighbors of node v,
- AGGREGATE is a permutation-invariant function such as sum, mean, or max,
- UPDATE is typically a multi-layer perceptron (MLP) or gating function.

After T iterations, the final node embeddings $\mathbf{h}_v^{(T)}$ can be aggregated into a graph-level representation via a readout function:

$$\mathbf{h}_G = \textsc{Readout} \left(\left\{ \mathbf{h}_v^{(T)} : v \in V \right\} \right),$$

where READOUT can be a simple function like summation, averaging, or more sophisticated approaches such as attention-based pooling or Set2Set.

Graph Convolutional Networks (GCNs) Kipf & Welling (2017) simplify the message-passing framework by linearly transforming and aggregating the features of neighboring nodes. The layer-wise propagation rule for GCNs is:

$$\mathbf{H}^{(t)} = \sigma \left(\hat{D}^{-1/2} \hat{A} \hat{D}^{-1/2} \mathbf{H}^{(t-1)} \mathbf{W}^{(t)} \right),$$

where:

- $\hat{A} = A + I$ is the adjacency matrix with added self-loops,
- $\hat{D}$ is the degree matrix of $\hat{A}$,
- $\mathbf{H}^{(t-1)}$ is the input feature matrix at layer $t - 1$,
- $\mathbf{W}^{(t)}$ is a trainable weight matrix,
- $\sigma(\cdot)$ is a non-linear activation function (e.g., ReLU).

The application of GNNs acknowledges that international trade functions through intricate, interdependent networks shaped by historical, political, and economic linkages. These include bilateral and multilateral trade agreements that establish preferential trading conditions, thereby influencing trade volume and direction. Capturing such complex structures requires sophisticated models capable of representing the dynamics of evolving economic relationships (Rincon-Yanez et al., 2023b). GNNs provide a robust framework to encapsulate these interrelations, offering a richer representation of trade dynamics than conventional approaches.

Building on this concept, this paper investigates the application of a specialised architecture, the Spatio-Temporal Graph Convolutional Network (STGCN) (Yu et al., 2017), originally proposed for traffic forecasting tasks, and selected here for its ability to jointly capture spatial dependencies among trading partners and temporal patterns. This dual capability makes STGCN a strong candidate for generating accurate and robust trade forecasts.

In this framework, the trade network is represented as a graph, where nodes correspond to countries and edges denote commercial relationships. Graph convolutions aggregate information from neighbouring nodes, allowing the model to account for both direct and indirect trade influences. Coupled with temporal convolution layers, the STGCN can effectively track and learn evolving trade patterns over time.

As a benchmark for performance evaluation, we employ a simple mean-based forecasting model, which predicts future values using the historical average of each time series. This baseline provides a transparent and easily interpretable reference point, enabling a clear assessment of the predictive gains achieved by the STGCN.

In this review, particular attention is given to the use of Spatio-Temporal Graph Convolutional Networks (STGCNs) for forecasting bilateral trade flows within a multivariate

time-series setting. This discussion considers how such models capture network effects, trade interdependencies, and temporal dynamics relative to simpler baselines, thereby highlighting the potential value of graph-based spatio-temporal learning alongside other modelling approaches such as gravity models and random forests.

2.4.2. Practical Implementation

Traditional methods for forecasting international trade, such as econometric models, have been widely employed in both research and policy-making. These approaches often depend on simplifying assumptions about the relationships between variables and require substantial volumes of historical data, along with manual adjustments. However, as international trade becomes increasingly dynamic and complex, there is a growing need for more advanced approaches capable of capturing the non-linear nature and intricate interconnections within the data.

In this context, we propose the exploration of more sophisticated techniques, beginning with the application of neural models. Recognising that commercial relationships can be represented as graph structures, we focus on GNNs as a promising solution. These specialised neural networks have the capacity to model complex relationships among interconnected nodes, making them well-suited to capturing the intricate structure and interactions within international trade networks.

By employing GNNs, our aim is to improve, or at the very least match, the accuracy and generalisability of international trade forecasts compared with existing methods. Achieving accurate predictions enables decision-makers to anticipate changes in trade flows, identify emerging trends, and develop more effective strategies. In turn, this opens new avenues for enhanced economic planning, strategic decision-making, and the promotion of more efficient and mutually beneficial international trade for all stakeholders.

2.4.3. Developments and Future Directions

Table 1 summarises notable applications of GNNs in trade and related domains, highlighting the diversity of approaches, datasets, and objectives. This variety underscores both the potential of GNN methods and the current lack of standardisation in evaluation frameworks, which can make results difficult to compare across studies.

Despite their promise, the application of STGCNs to trade flow modelling remains in its early stages. One of the primary challenges lies in the computational cost of training large-scale graph models, particularly when dealing with high-dimensional, longitudinal trade data. The scalability of STGCNs can become a bottleneck, requiring advances in architecture design and the adoption of more efficient training strategies to enable their deployment on global trade networks (Sellami et al., 2024). Moreover, the absence of standardised benchmarks for evaluating model performance in trade contexts hinders comparability and limits the generalisability of research findings.

Another significant limitation concerns interpretability. Unlike traditional econometric models such as the gravity model, which yield transparent parameter estimates and economically meaningful interpretations, STGCNs often operate as "black box" models. This opacity can pose challenges in policy-making environments, where transparency and accountability are essential. While recent progress in explainable AI (XAI) offers tools for improving model interpretability, these approaches are still developing and may not fully meet the requirements of economic policy analysis (Rincon-Yanez et al., 2023b).

On the developmental front, recent innovations in spatio-temporal graph learning, including adaptive adjacency matrices, dynamic graph construction, and attention-based modules, have improved the ability of models to capture evolving network structures, which is vital in the context of rapidly shifting trade relationships. The integration of multi-source data, including macroeconomic indicators, shipping routes, and policy announcements, offers further potential for improving predictive performance by enriching the feature space.

Future research should address scalability through lightweight model architectures, sampling-based learning, or distributed training methods that can accommodate the size and complexity of global trade networks. Interpretability remains a priority, with promising avenues including hybrid models that combine the predictive strengths of STGCNs with the structural transparency of gravity models. Additionally, the establishment of open, standardised datasets and benchmark tasks for trade flow prediction would promote reproducibility and facilitate more rigorous comparisons across approaches. Finally, integrating causal inference frameworks into STGCN-based models could enable not only accurate forecasting but also the identification of underlying drivers of trade pattern changes, thereby enhancing their value for both academic research and real-world policy-making.

3. Datasets

3.1. Overview

This study utilises a comprehensive international trade dataset compiled from multiple authoritative sources to identify countries with incomplete trade reporting and estimate missing trade values. The dataset encompasses bilateral trade flows between countries from 1996 to 2024, providing detailed transaction-level information on imports and ex-

Article	Approach / Model	Dataset	Specification
Monken et al. (Monken et al., 2021)	Artificial Intelligence Network Explanation of Trade (AINET)	UN Comtrade	Measures causal scenarios during outlier events in trade
Rincon-Yanez et al. (Rincon-Yanez et al., 2023a)	Knowledge Graph, GNN, Random Forest	CEPII Gravity	Edge weight prediction
Minakawa et al. (Minakawa et al., 2024)	Gravity-informed Graph Autoencoder (GGAE)	ECOWAS data	Predicts trade amounts
Panford-Quainoo et al. (Panford-Quainoo et al., 2020)	GCN, GAT, GAE, VGAE	UN Comtrade	Link prediction and classification of countries
Sellami et al. (Sellami et al., 2024)	GCN, GAT on trade networks	UN Comtrade	Predicts bilateral trade flows and explores data drift effects

Table 1. Applications of Graph Neural Networks in trade and related domains

ports across various product categories. The preprocessing pipeline transforms raw trade data from multiple sources into a clean, analysis-ready format suitable for identifying reporting gaps and developing estimation models.

3.2. Data Sources and Initial Structure

3.2.1. Primary Data Sources

The dataset integrates trade information from three primary authoritative sources:

World Trade Organization Integrated Database (WTO IDB): Provides comprehensive coverage of WTO member countries' trade statistics with detailed bilateral trade flows. This source offers extensive temporal coverage and maintains high standards for data quality and consistency.

UN Comtrade: Offers detailed bilateral trade data with extensive product coverage across multiple countries and time periods. This database provides complementary coverage to WTO IDB and helps fill gaps in trade flow reporting.

Consolidated Tariff Schedules (CTS): Contributes specialized tariff and trade policy information that enhances the comprehensiveness of the overall dataset.

3.2.2. Supporting Reference Datasets

Several reference datasets support the main trade flow data, ensuring consistent classification, mapping, and provenance tracking across the analysis.

Product Classification Data (df_mtn_hs): Contains standardised product category mappings based on the Multilateral Trade Negotiations (MTN) classification system. This dataset bridges detailed Harmonised System (HS) codes with broader analytical categories suitable for economic analysis.

Country Reference Data (df_country_name): Provides standardised country name mappings and ensures consistent country identification throughout the dataset.

Inventory Metadata (df_inventory): Links individual trade records to their originating data sources, ensuring complete data provenance and enabling quality assessment based on source reliability.

HS Product Category Mapping (df_hs): Standardised product category mappings at the HS 2-, 4-, and 6-digit levels, specific to corresponding HS versions, enabling multi-level product aggregation.

HS Concordance Mapping (df_hs_concordance): A comprehensive mapping of each HS code to its equivalent(s) in the subsequent HS version, allowing consistent longitudinal analysis despite code splits or merges.

Country Group Classification (df_country_group): Provides alternative country groupings (e.g., regions, income groups, trade blocs) to support aggregated analyses.

3.2.3. Core Trade Datasets Structure

The preprocessing pipeline works with two primary trade datasets that form the foundation of the analysis:

Import Dataset (df_imports): Contains detailed records of goods imported by reporting countries from their trading partners. Each record represents a specific import transaction with the following structure:

Dataset	No. of Countries	Data Sources (% of total value)	Key Statistics
df_imports	179 reporters, 176 partners	WTO IDB (73.21%), UN Comtrade (23.52%), Trade Data Monitor (3.27%)	Total reported trade value: USD 311.08 trillion. Total number of records: 111.4 million. Most frequent HS codes: 999999 (168,503), 392690 (150,888), 490199 (125,715), 732690 (124,253), 610910 (123,553).
df_exports	177 reporters, 164 partners	UN Comtrade (96.14%), Trade Data Monitor (3.86%)	Total reported trade value: USD 298.53 trillion. Total number of records: 90.4 million. Most frequent HS codes: 999999 (140,490), 392690 (123,800), 732690 (107,654), 490199 (100,713), 300490 (96,665).

Table 2. Summary statistics for import and export datasets

Dataset	Purpose	Scope	Specification
df_hs	HS category mapping	HS 2-, 4-, 6-digit	Provides standardised HS product category mappings per version.
df_hs_concordance	HS version mapping	All HS revisions	Maps each HS code to its equivalent(s) in the next version, handling splits and merges.
df_mtn_hs	Product classification mapping	HS to MTN categories	Bridges detailed HS codes with broader analytical MTN categories for economic analysis.
df_country_name	Country name standardisation	All countries	Ensures consistent country identification across datasets.
df_country_group	Country group classification	Global coverage	Groups countries by region, trade bloc, or economic category.

Table 3. Supporting reference datasets for trade flow analysis

Variable	Description
inventory_id	Unique identifier linking to metadata (int32)
reporter_code	Country reporting the import transaction (varchar)
partner_code	Country from which goods were imported (varchar)
year	Year of the trade transaction (int32)
hs_version	Harmonized System classification version (varchar)
hs_code	Specific product classification code (varchar)
value	Trade value in US dollars (double)

Table 4. Import dataset variable descriptions

Export Dataset (df_exports): Mirrors the import dataset structure but captures goods exported by reporting countries to their trading partners. The dataset maintains identical field structure to enable seamless integration and mirror statistics construction.

3.3. Data Preprocessing and Quality Enhancement

3.3.1. Data Source Tagging

The first preprocessing step established complete data provenance by linking each trade record to its originating source through the inventory metadata. This process ensures transparency in data origins and enables source-specific quality assessments. The integration revealed the distribution of records across the three primary sources, with each contributing to overall dataset comprehensiveness while requiring careful harmonization to maintain consistency.

3.3.2. Country Filtering and Scope Definition

A critical preprocessing challenge involved distinguishing legitimate country-to-country trade flows from other entity types present in the raw data. The original datasets contained various classifications including individual countries, regional trade unions, statistical aggregations, and other non-country entities.

To ensure analytical focus on bilateral country trade relationships, a systematic filtering approach was implemented. The methodology retained entities following the country identification pattern (alphabetic prefix 'C' followed by numerical digits) representing individual sovereign nations, plus the European Union (code 'U918') due to its unique status as a customs union with significant collective trade flows.

This filtering process removed substantial volumes of non-country trade data: approximately 17.5 million import records and nearly 10 million export records were excluded from the final datasets. While representing significant data reduction, this ensures the final dataset focuses specifically on bilateral country trade relationships, which aligns with the study's objective of identifying country-level reporting gaps.

3.3.3. Product Classification

The analysis incorporates two complementary product classification systems: the Multilateral Trade Negotiations (MTN) categories and the Harmonized System (HS) product codes. These provide different levels of aggregation and analytical focus, with MTN offering broader economic groupings and HS supplying a detailed, hierarchical commodity taxonomy. Both systems are integrated in a version-aware manner to ensure temporal consistency across the dataset.

MTN Categories

The MTN classification provides aggregated product groupings that are often more suitable for economic and policy analysis than raw HS codes. While the Harmonized System contains thousands of highly specific product codes, the MTN categories group related commodities into economically meaningful clusters.

The integration process matched trade records to MTN categories based on both HS code and HS version, recognising that codes can change over time. Special handling was applied to HS code `999999` (miscellaneous products), which was assigned to a dedicated miscellaneous MTN category to avoid distortion in aggregated statistics.

HS Code Categories

The HS framework is organised into broad *Sections*, subdivided into *Chapters* (2-digit codes), *Headings* (4-digit codes), and *Subheadings* (6-digit codes). This hierarchical design allows analysts to flexibly aggregate data by sector or drill down to specific commodities.

Because HS codes evolve across versions-through splits, merges, and redefinitions-the integration process is explicitly version-aware. Each trade record retains both its original HS code and mapped hierarchical identifiers for cross-version harmonisation.

An illustrative subset of HS Sections and Chapters is shown in Table 5.

Section	Codes	Description
I	01–05	Live animals; animal products
II	06–14	Vegetable products
IV	16–24	Prepared foodstuffs; beverages; tobacco
V	25–27	Mineral products
VI	28–38	Products of the chemical or allied industries
XI	50–63	Textiles and textile articles
XV	72–83	Base metals and articles of base metal
XVI	84–85	Machinery and electrical equipment
XVII	86–89	Vehicles, aircraft, vessels and transport equipment
XX	94–96	Miscellaneous manufactured articles

Table 5. Example HS Sections and Chapters (HS 2022)

3.3.4. Data Quality Validation

Comprehensive quality validation procedures ensured dataset reliability and analytical validity:

Completeness Assessment: Systematic examination revealed complete coverage with no missing values across critical variables.

Value Validation: All trade values were verified as positive, ensuring economic logical consistency. Zero or negative values were identified and removed as data anomalies.

Duplicate Detection: Identified duplicate records, which were concentrated in miscellaneous product categories, likely representing different subcategories for broadly classified goods. No duplicates were found among specific product classifications.

Referential Integrity: All relationships between datasets were validated to ensure proper connections between trade records, source attribution, and product classifications.

Final Processed Imports and Exports Data

Following the completion of all preprocessing procedures described above, the imports and exports datasets were fully prepared for subsequent analysis. Figure 1 presents a representative excerpt from the processed imports dataset, illustrating the final structure and variables available for analysis.

inventory_id int32	reporter_code varchar	year int32	hs_version varchar	hs_code varchar	partner_code varchar	value double	data_source varchar	mtn_subcategory varchar
21235	U918	2003	HS02	850432	C694	746.5919773	WTO IDB	T01
21235	U918	2003	HS02	850432	C702	571199.4226355	WTO IDB	T01
21235	U918	2003	HS02	850432	C703	816138.1511894	WTO IDB	T01

Figure 1. Sample of the final processed imports dataset after preprocessing steps.

3.4. Country Metadata

Following the preprocessing of the raw imports and exports datasets to isolate valid bilateral country-level trade flows, the next stage involved integrating these trade records with a range of external economic, policy, and geographic datasets. This step was essential to enrich the trade data with explanatory variables required for subsequent estimation methods.

The below process creates a `countries_metadata.csv` file used by the estimation methods. The integration process drew on 5 primary sources, each providing complementary dimensions of information:

- **Country pairs dataset** (`unique_country_pairs.csv`): Contains unique bilateral country relationships (no trade values, only pair codes where trade exists across years)
- **Preferential trade agreements**

(pref_pairs.csv): Preferential trading relationships between countries, provided by WTO.

- **Economic indicators**: Country-year indicators (e.g., GDP, land area, population density, sectoral GVA, rural/urban population shares) were sourced from the World Bank Open Data portal.
- **Sanctions** : Bilateral sanctions were merged from the *Global Sanctions Data Base (GSDB)* (Felbermayr et al., 2020), where sanction_exists is an indicator equal to 1 when any GSDB sanction episode is active for a pair in a given year. The variable preferred_pair was provided as a binary flag (1 = preferential trading relationship; 0 otherwise).
- **Geographic distances**: Pairwise distances (Distance_km) were computed as great-circle geodesics using geopy's geodesic function and OpenStreetMap's Nominatim geocoder for country centroids (Developers, 2025; OpenStreetMap contributors, 2025).

Data Cleaning and Standardization

Country Name Harmonization: Country names were standardized across datasets to ensure proper merging. A comprehensive mapping dictionary was created to handle variations in country nomenclature, including:

- Formal vs. common names (e.g., "United States of America" → "United States")
- Historical variations (e.g., "Russian Federation" → "Russia")
- Regional specifications (e.g., "Korea, Republic of" → "South Korea")

This mapping was consistently applied across all datasets to prevent merge failures due to naming inconsistencies.

Gap Filling in Time Series: Short gaps inside a country's time series were imputed via a forward-then-backward fill approach (FFILL/BFILL) at the country level. This ensured continuity in economic and demographic indicators while avoiding the introduction of artificial trends. The method preserves the original data structure by filling missing values only within existing country-year ranges.

Merging Process: The core integration involved creating a comprehensive panel dataset through sequential merging:

- **Temporal expansion:** Country pairs were replicated across all available years.
- **Bilateral indicator integration:** Economic indicators were merged separately for both countries in each pair (suffixed with _A and _B).
- **Sanctions integration:** Both directional sanctions relationships (A→B and B→A) were incorporated and combined into a single binary indicator.
- **Trade preferences:** Binary indicators of preferential trading relationships, identified through standardized country code matching.

Final Indicator Data

The final merged dataset contains bilateral country-year observations with:

- Economic indicators for both countries in each pair
- Binary sanctions indicators
- Preferential trade relationship flags
- Geographic distance controls
- Temporal coverage from 1996 onwards

This preprocessing approach ensures data quality, temporal consistency, and comprehensive coverage necessary for robust empirical analysis of bilateral economic relationships.

3.5. Basic Mirror Statistics for Complete Available Data

3.5.1. Mirror Statistics Methodology

Following the initial cleaning phase, a mirror statistics framework was applied to maximize the use of available trade data and to isolate genuinely missing bilateral flows. This approach addresses the common issue of asymmetries in international trade reporting, where one country reports a trade relationship that is not reciprocally recorded by its partner.

The method is based on the principle that every export from Country A to Country B should, in theory, be mirrored by an import from Country B to Country A for the same year and product category. Discrepancies between these two records reveal either gaps in reporting or differences in measurement practices.

3.5.2. Implementation Process

Bilateral Flow Matching: Trade flows were matched at the HS version and chapter level, ensuring compatibility across different classification versions and maintaining temporal alignment. For each year-product combination, the system searched for reciprocal records between the two partner countries.

Data Prioritization: Where both import and export records were available for the same flow, import statistics were retained in preference to export statistics. This choice reflects the generally higher accuracy of import records, which are supported by customs inspections and tariff collection processes.

Following the completion of the mirror statistics figure 2 presents a representative excerpt from the final structure and variables available for analysis.

importer varchar	exporter varchar	year int32	hs_version varchar	hs_code varchar	trade_value double	source_table varchar	mtn_subcategory varchar
C466	C710	2003	HS02	852510	2677.9784	imports	T05
C466	C710	2003	HS02	853510	27.5176	imports	T01
C466	C156	2003	HS02	854459	52081.4866	imports	T01
C466	C788	2003	HS02	854459	3808.0816	imports	T01
C466	U918	2003	HS02	870321	893.4002	imports	U01

Figure 2. Example of basic mirror statistics trade dataset.

3.5.3. Coverage Enhancement and Missing Flow Identification

Applying the mirror statistics methodology not only consolidated overlapping trade records but also recovered flows that were reported in only one direction. This maximized the coverage of the dataset before any imputation or modelling stages and ensured that subsequent missing value estimation was restricted to truly unreported trade flows rather than artefacts of asymmetrical reporting.

For the 2023 reference year, total recorded trade value increased from approximately $14.7 trillion in the import-only dataset to over $19.1 trillion after mirror statistics integration, representing a substantial improvement in captured trade coverage.

3.6. Final Datasets

The preprocessing pipeline produced several analytical datasets:

Cleaned Trade Datasets: Import dataset (df_imports_cleaned.parquet) with 111.4 million records and export dataset (df_exports_cleaned.parquet) with 90.4 million records, spanning 1996-2024 with complete data quality validation.

Mirror Statistics Datasets: Integrated datasets at HS code level (df_basic_mirror_hs.parquet), providing consolidated view of bilateral trade relationships with enhanced coverage.

Country Metadata: The `countries_metadata.csv` file contains a master list of all unique country-country pair combinations used in the analysis. Each record represents a bilateral relationship defined by standardized country identifiers and names, serving as the foundational reference table for integrating additional datasets (e.g., economic indicators, sanctions, trade agreements, geographic distances).

4. Exploratory Data Analysis

We conducted an exploratory data analysis on a large-scale international trade datasets (import and export), consisting of over **111 million import records** and approximately **90.4 million export records**. The aim was to assess the dataset's scope, structure, consistency, and reliability across key dimensions: geography, product classification, temporal coverage, and data source distribution.

4.1. Geographic Coverage and Trade Volume

The import dataset includes trade flows reported by **179 importer countries** and **176 exporter (partner) countries**, while the export dataset includes flows from **177 exporting countries** to **164 import partners**. The most active trading entities are European Union, United States, China, and United Kingdom, appearing consistently among the top ten for both imports and exports.

- Total reported **import value**: $311 trillion
- Total **export value**: $298.5 trillion
- **Average record value**: $2.79 million (imports), $3.3 million (exports)

The highest-value records involve `HS code 270900` (petroleum oils), with individual records exceeding $100 billion, particularly between Kingdom of Saudi Arabia and China.

4.2. Data Source Distribution and Trends

Data is sourced from three main providers: **WTO IDB**, **UN Comtrade**, and **Trade Data Monitor (TDM)**. Contributions vary significantly by time and dataset:

- **Import data sources**:
  - WTO IDB: 73.2%
  - UN Comtrade: 23.5%
  - Trade Data Monitor: 3.3%
- **Export data sources**:
  - UN Comtrade: 96.1%
  - Trade Data Monitor: 3.9%

Figure 3 illustrates the temporal evolution of dominant trade data sources across countries over the study period. In the earlier years, particularly 1996 and 2002, UN Comtrade (blue) appears as the principal source for a significant number of countries, reflecting its long-standing role in global trade reporting. However, as time progresses, the WTO Integrated Database (red) increasingly becomes the dominant

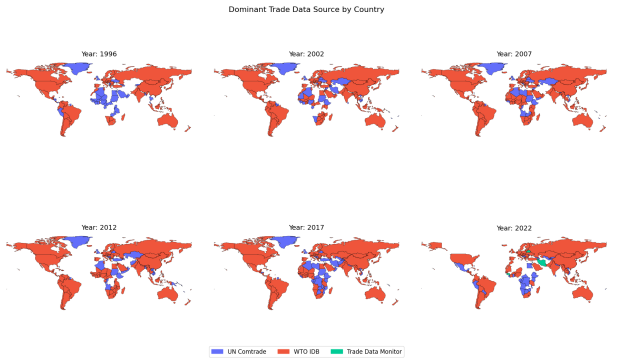


Figure 3. Dominant trade data source by country for selected years. Colours indicate whether the majority of available trade statistics for a given country were sourced from UN Comtrade (blue), WTO Integrated Database (red), or Trade Data Monitor (green).

source for most countries, with this shift becoming particularly pronounced after 2010. By 2022, WTO IDB data coverage is almost universal, with only a handful of countries still primarily represented through UN Comtrade or Trade Data Monitor (green). This transition in data dominance has implications for data harmonisation and continuity, as methodological differences between sources can affect comparability over time. Consequently, source transitions must be considered when interpreting long-term trade patterns or training predictive models on historical data.

Figure 4 shows how WTO IDB dominated import records until 2020, after which UN Comtrade and TDM became more prominent. **Figure 5** presents a similar trend for the share of reported import value by source.

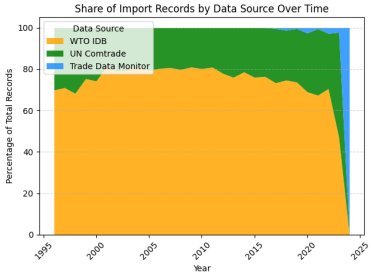


Figure 4. Share of Import Records by Data Source Over Time

4.3. Reporting Patterns by Country

- **Mozambique** transitions from WTO to TDM by 2024, as WTO data is not available for the most recent year.
- **Malawi** shifts from WTO to Comtrade starting in 2017, with no WTO submissions afterward.
- **The EU** maintains high volume across all years, incorporating TDM data starting in 2023.

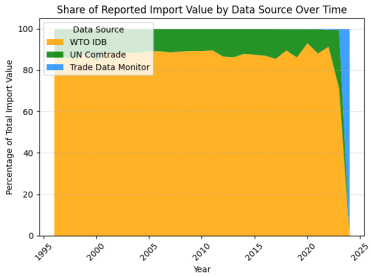


Figure 5. Share of Reported Import Value by Data Source Over Time

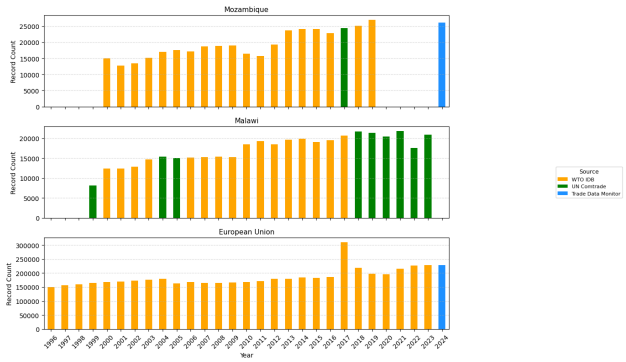


Figure 6. Annuual record counts over time for Mozambique, Malawi, and the European Union, broken down by data source.

For recent years, particularly 2024, **Trade Data Monitor (TDM)** becomes the *primary source of data* due to missing or delayed submissions from traditional sources like WTO and Comtrade.

4.4. Product Classification and HS Code Analysis

The dataset includes over **6,880 unique HS codes**, with `999999` (a generic placeholder) as the most common in both imports and exports. Other frequent codes include `392690`, `490199`, `732690`, and `870899`.

Top HS2 chapters include: `84` – Machinery, `85` – Electrical equipment, `90`, `39`, `62`, `61`, `73` – Instruments, plastics, and apparel

Top HS4 headings include: `8708` (auto parts), `6204` (women's suits), and `4202` (luggage and handbags).

4.5. HS Version Coverage

The dataset spans **seven HS revisions**: `HS92`, `HS96`, `HS02`, `HS07`, `HS12`, `HS17`, and `HS22`. **Figure 7** shows record counts by HS version and year.

Staggered adoption across countries highlights the need for *HS code harmonization* for any longitudinal analysis, as even after new HS versions are released, some countries

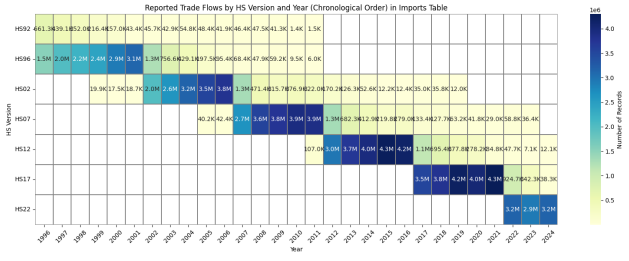


Figure 7. Reported Trade Flows by HS Version and Year

exhibit delays in migration, leading to temporal inconsistencies in reported trade classifications.

4.6. Temporal Distribution

The dataset spans a 29-year period from 1996 to 2024, encompassing substantial variation in the size and density of the global trade network over time. The number of reporting countries (nodes) grows from 98 in 1996 to a peak of 155 in the mid-2000s, before gradually declining to 77 in 2024. Correspondingly, the number of trade relationships (edges) increases from 8,313 in 1996 to over 16,300 in 2017, before contracting in later years.

Network density exhibits a U-shaped pattern: relatively high in the early years (0.87 in 1996), falling to around 0.61-0.65 in the mid-2000s as the network expanded, and then rising sharply in recent years, reaching 1.59 in 2024 due to a smaller but more interconnected set of countries.

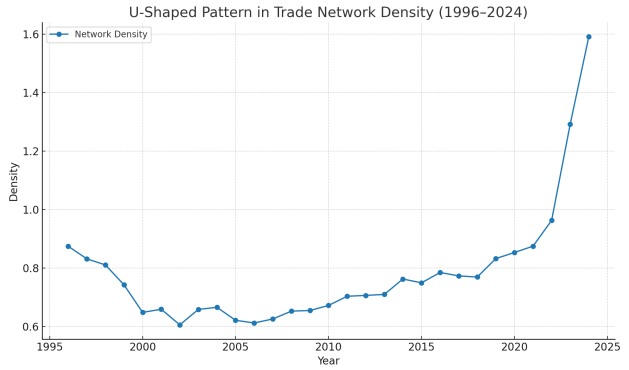


Figure 8. U-shaped evolution of trade network density (1996-2024). Density declines as participation expands in the early 2000s, then rises sharply post-2015, reaching a peak in 2024 as the set of reporting countries contracts but becomes more interconnected.

- **Imports**: 2017 (4.82M), 2019 (4.66M), with multiple years above 4.5M
- **Exports**: 2019, 2017, 2020, 2013, and 2021, each above 3.7M

4.7. Sanctions and Trade Preferences Impact

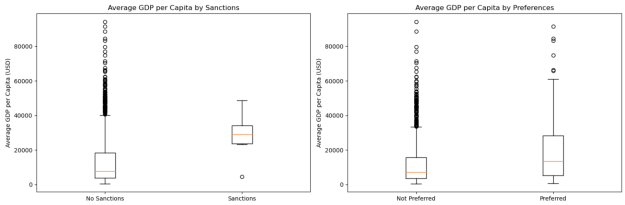


Figure 9. Aveage GDP per Capita by Sanctions and Preferences

The dataset reveals distinct patterns in how sanctions and trade preferences relate to economic and geographic factors. Sanctions affect only 2.4% of country pairs (8,490 out of 359,209), yet these relationships are economically significant. Sanctioned pairs exhibit substantially higher GDP per capita, particularly for the target countries (Country B: $37,162 vs $13,496 for non-sanctioned pairs), and are geographically closer (7,435km vs 8,403km average distance).

Trade preferences cover 18% of country pairs (64,393 pairs) and follow similar patterns of economic concentration. Preferred trading partners demonstrate higher wealth levels (Country A: $17,137 vs $12,478; Country B: $20,826 vs $12,231) and closer geographic proximity (7,070km vs 8,680km).

The cross-tabulation analysis reveals that 1,557 country pairs simultaneously maintain both sanctions and trade preferences, indicating complex diplomatic relationships where economic cooperation coexists with political tensions. Most sanctioned pairs (6,933) lack preferential trade status, while the majority of preferred partners (62,836) operate without sanctions.

These findings suggest that sanctions strategically target economically important relationships rather than serving as punitive isolation mechanisms, while trade preferences concentrate among wealthier, geographically proximate partners.

4.8. Distance vs Economic Relationships

The relationship between geographic distance and economic ties reveals nuanced patterns in international trade dynamics. The correlation between distance and average GDP per capita is relatively weak (0.050), suggesting that while geographic proximity influences trade relationships, economic factors operate largely independently of distance.

The scatter plot analysis of 5,000 country pairs demonstrates considerable variation in GDP levels across all distance ranges, with high-GDP economies distributed throughout the distance spectrum. This dispersion indicates that economic development is not geographically constrained, re-

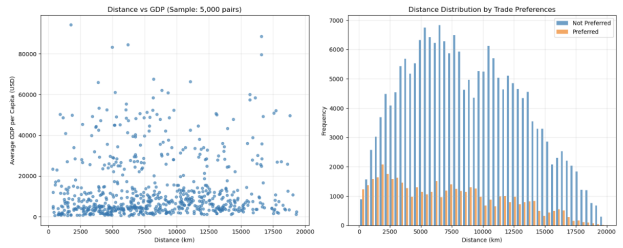


Figure 10. Distance vs GDP Analysis and Trade Preference Distribution

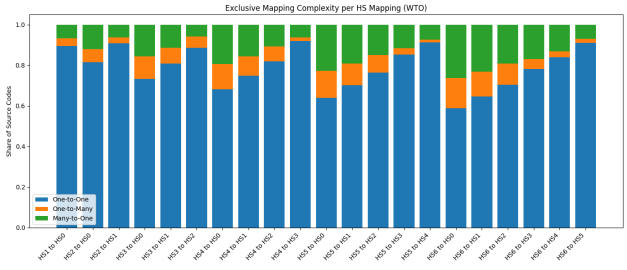


Figure 11. Distribution of mapping types (one-to-one, one-to-many, many-to-one) across HS version transitions.

flecting the global nature of modern economies.

The distance distribution by trade preferences reveals a clear geographic bias in preferential trading relationships. Preferred trading partners concentrate heavily in shorter distance ranges, with the frequency of preferred pairs declining sharply beyond 10,000 km. This pattern reinforces the importance of regional trade agreements and neighboring country partnerships in shaping global trade networks.

4.9. HS Code Harmonisation

In order to conduct consistent cross-year and cross-country analysis, it is necessary to harmonise HS codes that have been reported under different versions of the Harmonised System. This requirement is particularly important given that the dataset spans seven HS revisions, from HS92 through to HS22. We employ a concordance table, `df_hs_concordance`, derived from WTO-provided mappings, to trace changes between versions and ensure comparability.

The complexity of these mappings varies depending on the temporal gap between versions. As shown in Figure 11, adjacent versions such as `HS6` → `HS5` are highly stable, with 97.9% of mappings being one-to-one. By contrast, long-range mappings such as `HS6` → `HS0` show considerably more restructuring: only 85% of codes remain one-to-one, while over 20% are merged into fewer categories. The greater the gap between versions, the more likely it is that both aggregation (merging) and disaggregation (splitting) have occurred, resulting in ambiguous or lossy correspondences.

Splits (one-to-many mappings) are especially common when translating from more recent HS versions to older ones. For example, `HS5` → `HS0` exhibits a split rate of 13.4%, `HS6` → `HS0` reaches 15.0%, and `HS4` → `HS0` 12.5%. This reflects the introduction of additional detail in certain revisions, particularly in sectors such as electronics and chemicals, where broad historical codes have been subdivided into more precise categories. Merges, in contrast, are most prominent in older versions such as HS92 (`HS0`), where broad categories dominate: `HS6` → `HS0` sees 20.6% of mappings merge, `HS5` → `HS0` 18.2%, and `HS4` → `HS0` 15.7%. These early versions contained fewer, more general codes, so harmonising backwards necessarily collapses modern detail into coarser legacy classifications.

Although splits and merges present challenges, the majority of codes remain unchanged between adjacent versions, particularly in recent decades. This stability suggests that the HS system has matured, with most modern adjustments reflecting fine-grained refinements rather than large-scale restructuring.

Figure 12 summarises the types of code changes by HS chapter. Most chapters show high stability, indicated by the prevalence of unchanged codes (green bars). However, certain areas-such as Chapter 03 (fish and aquatic products), Chapters 28-30 (chemicals and pharmaceuticals), and Chapters 84-85 (machinery and electronics)-exhibit more frequent changes, with splits particularly common. By contrast, chapters such as 05 (products of animal origin, n.e.s.), 45 (cork), and 57 (carpets) have remained largely unaffected, reflecting long-term stability in their product definitions.

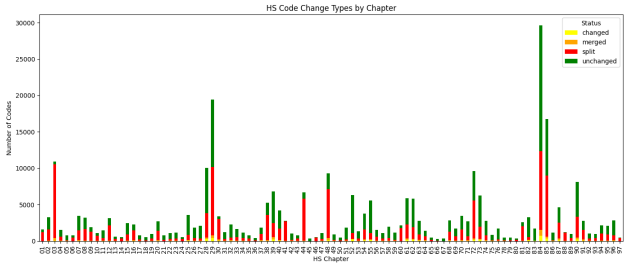


Figure 12. Type and frequency of HS code changes by chapter.

The complexity of some mappings can be extreme. In rare cases, a single HS6 code maps to dozens of codes in another version due to substantial restructuring. Examples from our mapping table include `285210` (mapping to 41 codes), `482390` (38 codes), `300692` (31 codes), and codes such as `848620`, `961900`, and `030399`, each mapping to over 20 codes. These tend to cluster in fast-evolving sectors like

chemicals and electrical products, where reclassification is frequent.

Figure 13 provides a concrete illustration of this phenomenon, showing how certain HS codes evolve across multiple revisions of the Harmonised System. Green arrows indicate stable one-to-one mappings, while yellow and red arrows denote more complex relationships such as one-to-many splits or many-to-one merges. This visual demonstrates how a single code may split, merge, and re-emerge multiple times over three decades, underscoring the difficulty of maintaining longitudinal consistency without a carefully curated concordance.

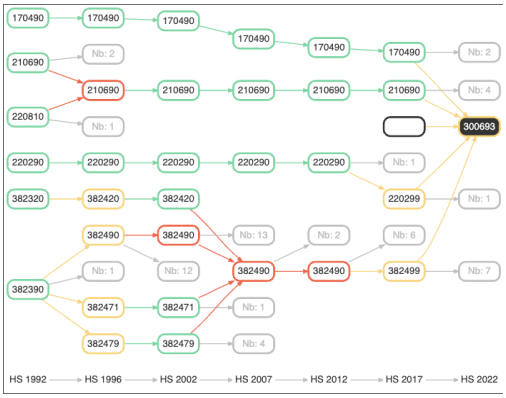


Figure 13. Example of complex HS code evolution across versions, including one-to-one, one-to-many, and many-to-one relationships.

Finally, Figure 14 presents heatmaps illustrating the share of mapping types between HS versions. These reinforce the finding that one-to-one mappings dominate adjacent version transitions, whereas splits and merges become far more prevalent when mapping across larger temporal gaps.

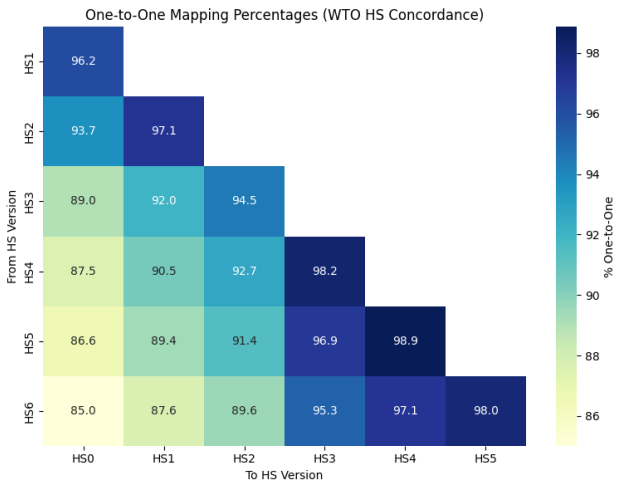


Figure 14. Heatmaps showing the share of mapping types between HS versions.

Overall, the harmonisation of HS codes is a non-trivial task that demands careful handling of splits and merges, particularly for sectors with high structural volatility. Chemicals, electronics, and fisheries stand out as areas requiring more sophisticated concordance logic, while stable chapters can be harmonised with relatively straightforward mappings. These patterns must be taken into account in any longitudinal or multi-version trade analysis to avoid double counting, omission, or misclassification.

4.10. Structural Evolution of Trade Flows

A complementary perspective on the evolution of trade is provided by chord diagrams, which visualise the top thirty bilateral trade flows for selected years. Figure 15 illustrates these networks from 1996 to 2023. Each panel represents one year, with countries arranged around the circumference and flows depicted as connecting arcs. The thickness of each arc is proportional to the trade value between the corresponding pair.

Several patterns emerge. The dominance of the United States, the European Union, and China is evident throughout, though their relative shares shift over time. Regional hubs such as Hong Kong SAR, Singapore, and Chinese Taipei appear prominently in earlier years but decline in later diagrams, reflecting changes in re-export patterns and the directness of supply chains. Conversely, economies such as India, Viet Nam, and Malaysia enter the top thirty pairs by 2023, suggesting diversification in high-value trade partnerships. The diagrams also highlight persistent bilateral strength between certain partners (e.g., EU-United States, China-Hong Kong SAR), as well as emerging linkages driven by structural shifts in manufacturing and commodity flows.

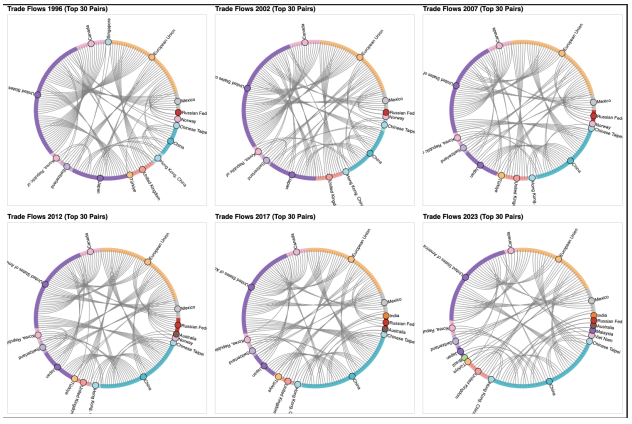


Figure 15. Chord diagrams of the top thirty bilateral trade flows for selected years, illustrating shifts in the structure and composition of global trade networks.

An additional dimension of exploratory analysis concerns

the stability of product composition in trade flows over time. Figure 16 presents a smoothed Kolmogorov-Smirnov (KS) drift statistic for the top ten MTN import subcategories, calculated year-on-year. Higher KS values indicate greater distributional change in trade shares for that category relative to the baseline period.

The results show pronounced structural shifts in certain commodities. Crude oils and telecommunication equipment exhibit the highest and most persistent drift, with KS statistics peaking above 0.2 in the mid-2010s, reflecting major supply chain reconfigurations and demand changes. Mineral fuels (excluding petroleum oils) also show an early 2000s spike, likely tied to commodity price cycles and diversification of energy sources. In contrast, categories such as motor vehicles and iron and steel display relatively lower and more stable drift values, suggesting a more consistent trade structure over time.

These findings underscore that not all sectors are equally stable in their temporal trade patterns. High-drift categories may require more flexible modelling approaches to capture structural change, while low-drift categories can be modelled more effectively with long-term historical averages.

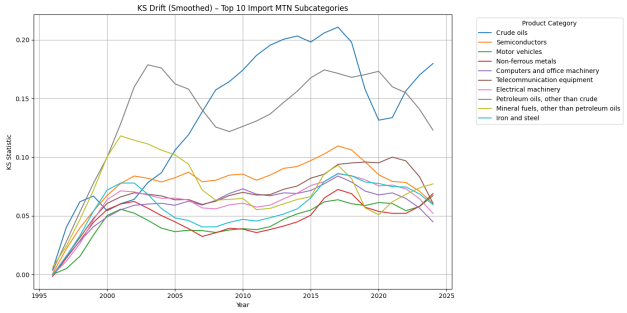


Figure 16. Smoothed KS drift statistic for the top ten MTN import subcategories, showing the extent of year-on-year distributional change in trade composition.

5. Methodology

5.1. Mirror Statistics and Temporal Imputation

The mirror statistics framework constructs a detailed trade panel at the HS *chapter* level, enabling estimation of missing flows while preserving directional reporting structure and partner specificity. Aggregation to chapter-level (first two digits of the HS code) strikes a balance between product granularity and data availability, significantly reducing sparsity while retaining interpretability and economic structure.

5.1.1. Chapter-Level Aggregation and Full Panel Construction

The cleaned imports and exports dataset was aggregated by `(reporter, partner, year, HS chapter)` for both imports and exports. To ensure comprehensive temporal coverage, synthetic entries were then generated for all plausible trading relationships - even in years where reporting was incomplete or absent.

Full Panel Creation Logic: For each unique combination of reporter, HS chapter, and partner, a start year was determined based on the earliest year in which any trade was reported (either as import or export). The panel was expanded to include all years from 1996 to 2023.

However, a data point was only marked as *missing* if:

- The reporter was entirely inactive in that year (i.e., did not report any trade whatsoever),
- The HS chapter was already active before that year for this reporter, and
- The chapter-partner combination had a prior trade record (i.e., the trade relationship had already been established).

In contrast:

- If the reporter was active in a given year, it was assumed that all chapter-partner trades they engaged in would have been reported-so no synthetic entries were added even if some were missing.
- If the reporter was inactive and the chapter-partner combination had not yet been observed trading, no rows were generated-on the assumption that the relationship had not yet begun.

This careful filtering ensures that only economically meaningful gaps are flagged as missing while avoiding inflation of null data due to uninitiated trade relationships.

5.1.2. Mirror Construction and Discrepancy Profiling

Import and export datasets were first converted into full panels separately by the logic stated earlier. This ensured complete directional coverage and consistency, even for non-reporting years. The two panels were then merged using full outer joins to form a unified mirror dataset. This merging step both maximizes data completeness and minimizes the need for unnecessary estimation - if a value is reported in either direction, it is preserved. The alignment enables analysis from both perspectives while reducing missingness and potential estimation error.

For each record, the following mirror statistics were computed:

- **Discrepancy:** The signed difference between reported import and export values.
- **Mirror Ratio:** The ratio of import to export (when both are available and export is non-zero), providing a scale-invariant measure of reporting asymmetry.
- **Final Trade Value:** Assigned as the reported *import* value if available; otherwise, the reported *export* value is used. If neither is available, the final trade value remains missing. Imports are prioritized because they exhibit significantly lower rates of missingness across the dataset and are also considered more reliable by WTO.

This structure allows precise identification of unreported flows, directional bias, and inconsistencies in partner-country records.

5.1.3. Adjacent Year Estimation Method

To recover missing trade values, a temporal interpolation strategy was implemented using data from the immediate previous and next years. This approach increases coverage while preserving realistic temporal dynamics.

Lag and Lead Construction: For each unique `(importer, exporter, HS chapter)` combination, values from the **adjacent years** - one year before and one year after - were calculated using window functions over time. This enabled identification of the closest non-missing neighbors around each observation.

Estimation Logic:

- If both the previous and next year values were available, their average was used.
- If only one of the two was available, that single value was used.
- If both were missing, the estimate remained null and was left for exclusion.
- Estimation was performed separately for imports and exports to maximize the use of already available datax.

5.1.4. Final Trade Value Assignment

After mirror merging and adjacent year imputation, a final trade value was assigned to each record using the following priority order:

$$\begin{aligned}\text{Import value} &> \text{Export value} \\ &> \text{Adjacent Imports Estimation} \\ &> \text{Adjacent Exports Estimation}\end{aligned}$$

This ordering reflects both observed data availability patterns and institutional standards. Import values were less frequently missing and are often treated as more reliable by organizations such as the WTO.

This prioritization scheme ensures that the mirror statistics framework yields a maximally complete and internally consistent trade dataset, minimizing reliance on estimation where reported values are available and maintaining directional symmetry across records.

5.2. Clustering and Gravity Model

This section presents a two-stage methodology for estimating bilateral trade flows using a cluster-augmented gravity model. In the first stage, country pairs are grouped into macroeconomically similar clusters using unsupervised learning techniques applied to standardized policy and economic features. In the second stage, a log-linear gravity model is estimated separately within each cluster-period group, enabling parameter sharing across structurally comparable trade relationships. This approach addresses challenges posed by sparse trade data and high dimensionality, while preserving heterogeneity in trade determinants across different economic contexts.

5.2.1. Balanced Panel Creation and Feature Integration

A **balanced panel** of all possible bilateral country pairs over the period 2000–2023 is constructed as the common input for both clustering and gravity model estimation. All trade values are aggregated to **bidirectional totals**, meaning the flow from Country i to Country j is summed with the flow from j to i. As a result, the dataset does not preserve directional information (exporter versus importer), but represents the *total* observed trade between each pair–year. This aggregation is consistent with the symmetric treatment of country pairs in the clustering stage and mitigates issues from asymmetric trade reporting between partners.

1. **Trade Aggregation:** Raw bilateral trade flows are harmonized to an undirected format by alphabetically sorting importer–exporter codes for each transaction. This ensures that trade between Country i and Country j is aggregated symmetrically:

   $$\text{TotalTrade}_{ij,t} = \text{Exports}_{i \to j,t} + \text{Exports}_{j \to i,t}.$$

   The aggregated totals are computed for each pair–year.

2. **Balanced Panel Generation:** The set of unique country pairs is combined with the full study-year range to produce all possible pair–year combinations. Merging with the aggregated trade data yields a panel where missing trade observations correspond to years where neither country reported flows.

3. **Feature Augmentation:** This is done by merging the `countries_metadata.csv` dataset mentioned in section 3.4

    - **Country-Level Metrics:** Economic and demographic indicators (GDP, population, land area, sectoral value-added, rural and urban population, etc.) are merged for both countries in each pair.
    - **Pairwise Metadata:** Bilateral attributes (distance between capitals, sanctions, preferential trade agreement status, etc.) are merged once per pair–year.
    - **Identifiers:** Country names are added for interpretability in analysis and visualization.

The resulting dataset is a **complete, feature-rich balanced panel** that serves as the input for clustering (excluding the trade value variable) and, subsequently, for gravity model estimation.

5.2.2. Clustering of Country Pairs for Gravity Estimation

This section outlines the methodology used to cluster bilateral country pairs based on shared macroeconomic and policy characteristics, serving as a foundation for estimating gravity model parameters in groups rather than individually. This pools information across structurally similar trade relationships, particularly for countries with limited historical trade data.

Assumptions and Design Considerations

- **Unit of Observation:** Each observation corresponds to a bilateral country pair (A, B) in a given year. Clustering is performed separately for each rolling 5-year window (e.g., 2000–2004, 2005–2009, ..., 2020–2023) to account for changing macroeconomic and policy conditions over time.

- **Feature Scope:** All clustering variables are drawn from the balanced panel described earlier, excluding the trade value variable. Specifically, the features used for clustering are:

    – *Policy Variables:* `sanction_exists`, `preferred_pair`
    – *Geographic Variables:* `Distance_km`
    – *Macroeconomic Variables (Country A and B):* `GDP`, `land_area`, `population_density`, `agriculture_gva`, `industry_gva`, `rural_pop`, `urban_pop`

    Each variable for Country A and Country B is suffixed with `_A` or `_B` to preserve country-specific information within the pair.

- **Temporal Aggregation:** For each 5-year window, all macroeconomic and policy variables are averaged across the five years. Each country pair appears only once per period, with all features reflecting their mean values over that interval. This ensures robustness to short-term fluctuations and enables more stable clustering assignments.

- **Pairwise Symmetry:** Country pairs are treated as undirected; the clustering does not distinguish between sender and receiver. Instead, it uses the joint set of macroeconomic, geographic, and policy features from both countries to characterize the relationship, capturing symmetric trade-relevant similarities.

- **Missing Data Handling:** Observations with missing values across any selected features are dropped prior to clustering. These cases are primarily smaller economies, territories with limited statistical reporting, or historical political entities that no longer exist in their previous form. Examples include Serbia and Montenegro, Netherlands Antilles, and Anguilla.

- **Feature Scaling:** All numerical features are standardized using scikit-learn's `StandardScaler`, which applies z-score normalization (subtracting the mean and dividing by the standard deviation) to ensure comparability across dimensions.

Clustering Methodologies

To account for heterogeneity in trade determinants, three unsupervised clustering algorithms were applied to the standardized macroeconomic feature vectors: K-Means, Hierarchical Clustering, and HDBSCAN. Each method captures different structural assumptions and density patterns in the country-pair space.

Clustering performance was evaluated using the **Silhouette Score**, used here as a baseline metric alongside other evaluations discussed later, which ranges from -1 to 1 and reflects how similar each observation is to its own cluster compared to other clusters. Higher scores indicate well-defined, cohesive clusters with clear separation from others.

1. K-Means Clustering K-Means partitions the data into a pre-specified number of clusters by minimizing the within-cluster sum of squared distances from each point to its assigned cluster centroid. The following implementation details were used:

- **Input Features:** A total of 17 standardized features representing macroeconomic, fiscal, and policy variables for country A and B.

- **Optimal Cluster Selection:** The number of clusters k was determined using silhouette score maximization over a grid from $k = 2$ to $k = 20$.

- **Final Model:** For each 5-year period, the optimal number of clusters k was determined independently by maximizing the silhouette score over a grid of candidate values. KMeans was then applied using the K-Means++ initialization strategy with 20 restarts (`n_init=20`) to ensure robustness to local minima. A fixed random seed (`random_state=42`) was used to ensure reproducibility. Cluster labels were assigned to all country pairs for each period.

Strengths: Computationally efficient and interpretable, particularly when clusters correspond to observable economic regimes (e.g., low-income vs high-income pairs).

2. Hierarchical Clustering Hierarchical agglomerative clustering, where each observation starts as its own cluster and the most similar clusters are iteratively merged, was applied separately to each 5-year period.

- **Distance Metric:** Euclidean distance was used on the z-score normalized feature space to compute pairwise similarities between country pairs.

- **Linkage Criterion:** Ward's linkage method was used to merge clusters, which minimizes the total within-cluster variance at each step of the agglomeration process.

- **Model selection:** For each 5-year period, the optimal number of clusters k is selected by maximizing the silhouette score over $k \in \{2, \ldots, 10\}$. (Errors for infeasible k are safely skipped.)

- **Final fit and labels:** An `AgglomerativeClustering` model with the chosen k and Ward linkage is fit per period; cluster labels are then concatenated across periods into a single results table..

Strengths: Offers flexibility in cluster resolution and intuitive tree-based interpretability. No need to pre-specify the number of clusters.

3. HDBSCAN Clustering HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) forms clusters based on the density of data points, identifying dense regions in the feature space without requiring a predefined number of clusters. Unlike distance-based methods, it can find arbitrarily shaped clusters and detect sparse regions as noise.

- **Parameter Selection:** The minimum cluster size was set to 20, and a small `cluster_selection_epsilon` value of 0.2 was used to adjust sensitivity to cluster boundaries. These parameters were chosen to balance granularity with cluster stability across periods.

- **Noise Handling and Soft Reassignment:** By default, HDBSCAN assigns `-1` to points it considers noise (i.e., unclustered). To avoid losing data, these points were reassigned using soft clustering: each point's membership probabilities were computed, and it was reassigned to the cluster with the highest probability.

- **Cluster Assignment:** The final set of cluster labels, including soft-assigned former outliers, was stored for each country pair in each 5-year window. These clusters serve as group identifiers for downstream gravity model estimation.

Strengths: Particularly effective in high-dimensional, non-globally convex feature spaces. Avoids overfitting by filtering out noise. Ideal for exploratory analysis where cluster structure may be non-uniform.

The output of each clustering algorithm was merged back into the master dataset as a categorical identifier for each country pair. These cluster labels enable stratified estimation of gravity models, where coefficients can be estimated at the cluster level instead of individual pairs. This facilitates parameter sharing across economically similar pairs and improves the robustness of missing trade value predictions.

5.2.3. Gravity Model Estimation

This section details the implementation of the gravity model used to estimate bilateral trade values across country pairs. The model leverages the economic intuition that trade volume between two countries is positively related to their economic size and negatively related to the distance between them. The classical log-linear gravity model (equation 1 is extended to incorporate additional policy variables and is estimated separately for each period and clusters of country pairs obtained from prior unsupervised clustering.

Model Specification

For each cluster-period combination, we estimate a log-linear gravity equation (extended version of equation 1) of the form: For each cluster-period combination, we estimate

a log-linear gravity equation of the form:

$$\begin{aligned}\log(\text{Trade}_{ij}) = \beta_0 &+ \beta_1 \log(\text{GDP}_i) + \beta_2 \log(\text{GDP}_j) \\ &+ \beta_3 \log(\text{LandArea}_i) + \beta_4 \log(\text{LandArea}_j) \\ &+ \beta_5 \log(\text{PopDensity}_i) + \beta_6 \log(\text{PopDensity}_j) \\ &+ \beta_7 \log(\text{AgriGVA}_i) + \beta_8 \log(\text{AgriGVA}_j) \\ &+ \beta_9 \log(\text{IndGVA}_i) + \beta_{10} \log(\text{IndGVA}_j) \\ &+ \beta_{11} \log(\text{RuralPop}_i) + \beta_{12} \log(\text{RuralPop}_j) \\ &+ \beta_{13} \log(\text{UrbanPop}_i) + \beta_{14} \log(\text{UrbanPop}_j) \\ &+ \beta_{15} \log(\text{Distance}_{ij}) \\ &+ \beta_{16} \cdot \text{Sanction}_{ij} + \beta_{17} \cdot \text{PTA}_{ij} + \varepsilon_{ij}\end{aligned} \tag{2}$$

where:

- $\text{GDP}_i, \text{GDP}_j$: Nominal GDPs (in billions of USD),
- $\text{LandArea}_i, \text{LandArea}_j$: Land area (in sq. km),
- $\text{PopDensity}_i, \text{PopDensity}_j$: Population density (people per sq. km),
- $\text{AgriGVA}_i, \text{AgriGVA}_j$: Agriculture gross value added (in billions USD),
- $\text{IndGVA}_i, \text{IndGVA}_j$: Industry gross value added (in billions USD),
- $\text{RuralPop}_i, \text{RuralPop}_j$: Rural population (in millions),
- $\text{UrbanPop}_i, \text{UrbanPop}_j$: Urban population (in millions),
- D_{ij}: Great-circle distance between capitals (in km),
- Sanction_{ij}: Binary indicator for sanctions,
- PTA_{ij}: Binary indicator for preferential trade agreement,
- ε_{ij}: Error term.

Input Dataset for Gravity Model: Cluster Integration

The input dataset for the gravity model is a balanced panel of all bilateral country pairs over 2000–2023, with features including bidirectional trade values, macroeconomic indicators (GDP, population, etc.), bilateral attributes (distance, agreements, sanctions), and country identifiers.

Cluster labels, initially assigned for 5-year periods, are expanded to annual frequency and merged into this panel. As a result, each country pair–year observation is annotated with a macroeconomic cluster. These clusters allow for the estimation of separate gravity model parameters within more homogeneous subgroups of the global economy.

Estimation Framework

This approach treats each regression as a short time-series model (5 observations) for a single representative pair, rather than as a pooled panel regression across multiple pairs

- **Clustering Integration:** Each (period, cluster) combination is treated as a distinct group for model estimation, allowing parameter sharing across economically similar pairs and reducing the risk of overfitting in sparse settings.
- **Training Sample Selection and Missing Data Handling:** For each (period, cluster) group, a single representative country pair with complete feature information and strictly positive trade values for all five years in the period is selected for model estimation. This approach serves two purposes:
  - it avoids pooling multiple country pairs into the same regression, which could incorrectly assume they follow identical linear relationships, and
  - (ii) it minimizes training data usage so that the remaining pairs can be reserved for prediction.

  Each regression is thus based on a balanced 5-year time series for the chosen pair, ensuring valid log-transformations and avoiding complications from missing or zero trade flows.
- **Log-Transformations:** The dependent variable (total bilateral trade) is log-transformed. Likewise, GDP, distance, and other continuous variables are log-transformed to linearize multiplicative relationships and stabilize variance.
- **Regression Method:** Ordinary Least Squares (OLS) is applied separately to each (period, cluster) group.
- **Prediction Setup:** Once fitted, the model parameters are stored per (period, cluster) and used to estimate missing trade values for other country pairs in the same group.

This cluster-wise estimation framework allows the gravity model to adapt its coefficients across distinct macroeconomic regimes, enabling more accurate and context-aware imputation of missing bilateral trade values. In theory, the gravity model offers a well-established baseline for explaining trade patterns through economic size and distance effects, while the integration of clustering allows for parameter pooling across structurally similar pairs, enhancing estimation stability in sparse-data settings. Nevertheless, this approach rests on the assumption that country pairs within the same cluster share broadly comparable trade determinants, an assumption that may not fully capture unique historical, political, or structural factors shaping specific bilateral relationships.

5.3. Machine Learning Imputation: MissForest Approach

5.3.1. Methodology

In contrast to the clustering-based gravity model detailed in the previous sections, we implement **MissForest**, a non-parametric imputation method that leverages the full statistical power of our complete dataset. MissForest employs Random Forest ensembles in an iterative framework to predict missing trade values using gravity model variables as predictors. This approach addresses the fundamental limitation encountered in the clustering methodology: insufficient training data for robust parameter estimation.

Algorithm Framework. The MissForest algorithm operates through an iterative process where Random Forest models are trained on observed portions of the trade data, using the imputed values to iteratively refine predictions until convergence. The algorithm follows the iterative imputation framework developed by (Breiman, 2001):

For each iteration t, missing values for feature j are predicted using:

$$X_{ij}^{(t)} = \hat{f}_j^{(t)}(x_{i,-j}) \tag{3}$$

where $\hat{f}_j^{(t)}$ is the Random Forest trained on observed values and $x_{i,-j}$ represents all features except j for observation i. The process continues until convergence based on minimal change in imputed values.

Data Preparation. Unlike the clustering approach that relied on only 5 training points per model, MissForest utilizes the entire dataset of 395,044 observations, providing the statistical foundation necessary for reliable prediction. The dataset consists of bilateral trade relationships between 197 countries from 1996 to 2024, with trade values ranging from \$0 to \$619 billion (mean: \$710 million). This extreme range spanning nine orders of magnitude presents significant challenges for prediction methods.

Feature Set. Following established gravity model theory, we employ seven predictor variables:

- **Economic mass**: Bilateral GDP (exporter and importer nominal GDP in billions USD)
- **Trade costs**: Geographic distance between capitals (great-circle distance in kilometers)
- **Policy variables**: Binary indicators for sanctions and preferential trade agreements
- **Country-specific effects**: Categorical country identifiers capturing unobserved heterogeneity and fixed effects

Missing Data Generation and Validation Strategy. Given our complete trade dataset, we implement a dual validation approach to assess imputation quality:

1. **Cross-validation**: 20% of trade values are artificially masked from the training dataset (79,008 observations), creating a realistic missing data scenario for validation
2. **Out-of-sample testing**: Independent evaluation on a universal test set containing 64,363 observations to assess external generalizability

Implementation Details. The Random Forest implementation uses 100 decision trees with the following specifications:

- **Ensemble size**: 100 estimators to balance accuracy and computational efficiency
- **Tree depth**: Maximum depth of 15 to capture complex interactions while preventing overfitting
- **Categorical encoding**: Label encoding for country identifiers to handle categorical variables
- **Convergence criteria**: Algorithm terminates when mean absolute change between iterations falls below 10^{-6} or maximum of 10 iterations reached

Comparison with Alternative Methods. To validate the choice of Random Forest as the base learner, we compare MissForest performance against alternative machine learning approaches:

- **Gradient Boosting**: Sequential tree building with error correction
- **XGBoost**: Optimized gradient boosting with advanced regularization

All methods use identical feature sets and validation procedures to ensure fair comparison.

Hyperparameter Optimization. To validate the choice of 100 estimators, we conduct a systematic analysis of Random Forest performance across different ensemble sizes ranging from 10 to 300 trees. Training and validation performance metrics (R^2, RMSE, and MAE) are evaluated for each estimator count, with results visualized in performance graphs to identify the optimal balance between prediction accuracy and computational efficiency.

Evaluation Model performance was evaluated using multiple metrics including R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE), which will be detailed in the Graph Neural Network section. Additionally, bias was employed as a key evaluation metric to assess systematic prediction errors. Bias in prediction is calculated as the percentage difference between the mean predicted values and mean actual values:

$$\text{Bias} = \frac{\bar{y}_{\text{pred}} - \bar{y}_{\text{true}}}{\bar{y}_{\text{true}}} \times 100\% \tag{4}$$

where $\bar{y}_{\text{pred}}$ represents the mean of predicted values and $\bar{y}_{\text{true}}$ represents the mean of actual observed values. A positive bias indicates systematic overestimation, while negative bias indicates systematic underestimation. Unlike MSE and MAE which measure prediction accuracy, bias specifically captures the direction and magnitude of systematic prediction errors, making it particularly valuable for understanding whether the model consistently over- or under-predicts trade values across the dataset. This metric is especially important in trade prediction applications where understanding the systematic tendencies of the model is crucial for policy and economic analysis.

For the MissForest implementation, convergence was monitored through the iterative imputation process. MissForest employs an iterative algorithm where missing values are initially filled with simple estimates (e.g., mean for continuous variables) and then refined through successive Random Forest predictions. Convergence is achieved when the difference between imputed values across consecutive iterations falls below a predefined threshold, or when a maximum number of iterations is reached. The algorithm tracks the normalized root mean squared error (NRMSE) between successive iterations, with convergence typically occurring when the change in NRMSE becomes negligible (¡ 0.001) or after a maximum of 10 iterations. This iterative refinement ensures that the final imputed values are consistent with the underlying data patterns captured by the Random Forest model.

5.3.2. Product-Level Trade Prediction Analysis

Methodology

To assess the predictive capability of machine learning methods at different levels of trade classification granularity, we implement a comparative analysis across three Harmonized System (HS) aggregation levels. This analysis addresses the fundamental question of whether prediction accuracy improves with product aggregation or benefits from maintaining detailed product-level information.

Data Preparation and Sampling. The analysis utilizes the complete bilateral import dataset containing over 70 million individual trade transactions. To ensure computational feasibility while maintaining statistical representativeness, we employ stratified random sampling to extract 200,000 observations from the full dataset. This sample size balances statistical power with computational constraints, enabling systematic comparison across multiple HS classification levels.

The dataset is processed using memory-efficient chunking methodology to handle the large-scale trade data:

- **Chunk processing**: Data processed in 500,000-observation chunks to prevent memory overflow
- **Aggregation strategy**: Within each chunk, trade values are aggregated by (reporter, partner, year, HS code) combinations
- **Memory management**: Systematic garbage collection between chunks to optimize computational efficiency

HS Classification Level Analysis. We systematically compare prediction performance across three levels of the Harmonized System classification:

1. **HS 2-digit (Chapters)**: Broadest aggregation level representing major commodity groups (e.g., ”01” = Live animals, ”84” = Machinery)
2. **HS 4-digit (Headings)**: Intermediate aggregation providing product group specificity (e.g., ”8471” = Automatic data processing machines)
3. **HS 6-digit (Products)**: Most detailed level representing specific products (e.g., ”847130” = Portable digital computers)

For each classification level, trade values are aggregated by summing individual transactions within the respective HS groupings. This approach ensures that prediction targets represent economically meaningful trade flows rather than individual transaction records.

Machine Learning Implementation. Random Forest regression is applied uniformly across all HS levels to ensure methodological consistency. The prediction framework employs four predictor variables:

- **Reporter country**: Encoded categorically using label encoding
- **Partner country**: Encoded categorically using label encoding

- **HS classification**: Product identifier encoded at the respective aggregation level
- **Year**: Temporal variable capturing time trends

Model Specification and Validation. Each HS level employs identical Random Forest specifications to ensure fair comparison:

- **Ensemble size**: 50 estimators to balance accuracy and computational efficiency
- **Train-test split**: 80% training, 20% testing with stratified random sampling
- **Performance metrics**: R^2, RMSE, and convergence analysis
- **Cross-validation**: Consistent random seed (42) for reproducible results

The objective is to determine whether product aggregation improves prediction accuracy by reducing noise and increasing sample sizes per category, or whether maintaining product granularity provides superior predictive information despite increased sparsity.

5.3.3. Implementation of Silva et al. (2024) Trade Network Methodology

This analysis implements the methodology from Silva et al. (2024), "Machine learning and economic forecasting: the role of international trade networks," which demonstrates that network topology descriptors from international trade networks substantially enhance economic growth forecasting capabilities.

Data Preparation The implementation utilizes section-level trade data following the World Trade Organization's Harmonized System (HS) classification. Trade data spanning 2010-2022 was processed to create 21 economic sections representing distinct commodity categories such as machinery, chemicals, textiles, and mineral products. **Data Specifications:**

- Sample size: 300,000 trade observations
- Time period: 2010-2022 (following Silva et al.'s timeframe)
- Geographic coverage: 175 countries
- Final analysis dataset: 943 country-year observations

Feature Construction The core innovation of Silva et al.'s approach lies in transforming bilateral trade relationships into network topology descriptors that capture countries' positions in global trade networks. **Network Creation Process:**

1. **Annual Trade Networks**: For each year, construct directed networks where nodes represent countries and weighted edges represent bilateral trade flows
2. **Centrality Calculation**: Compute multiple centrality measures for each country:
   - **Degree Centrality**: Measures how many trading partners a country has
   - **In-Degree Centrality**: Captures import connectivity (number of source countries)
   - **Out-Degree Centrality**: Measures export connectivity (number of destination countries)
   - **Betweenness Centrality**: Identifies countries that serve as intermediaries in global trade paths

Economic Rationale: Countries with higher centrality scores are more integrated into global trade networks, suggesting greater economic diversification, supply chain importance, and resilience to economic shocks.

Economic Indicator Integration Following Silva et al.'s methodology, network features are combined with traditional economic indicators to create a comprehensive feature set: **Economic Features:**

- **Lagged Trade Growth**: Previous period's trade growth rate
- **Log Trade Volume**: Natural logarithm of total trade value
- **Trade Partners**: Number of bilateral trade relationships
- **Trade Concentration**: Ratio of trade volume to number of partners

Miss Forest Implementation The methodology employs a **horse race** approach comparing non-linear machine learning models against traditional linear methods: **Models:**

- **Random Forest**: Primary model following Silva et al.'s findings of superior non-linear model performance
- **Linear Regression**: Baseline model representing traditional econometric approaches

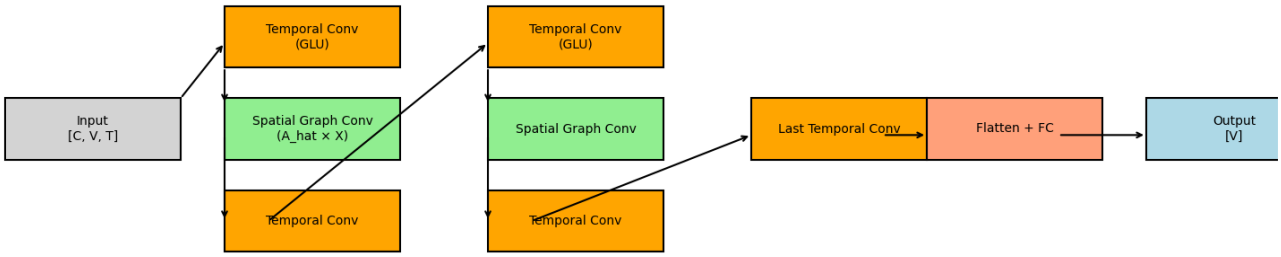


Figure 17. Overview of the Spatio-Temporal Graph Convolutional Network (STGCN) architecture used in this study. The architecture integrates temporal convolutional layers and spatial graph convolutions to capture both temporal dependencies and spatial relationships within the trade network.

Target Variable: Next year's economic growth (using trade growth as proxy for GDP growth) **Validation Strategy**: Temporal train-test split (2010-2018 for training, 2019-2022 for testing) to ensure realistic out-of-sample forecasting evaluation.

5.4. STGCN Framework

Model Rationale In selecting an appropriate architecture for predicting international trade flows, two fundamental challenges must be addressed.

The first challenge is the limited availability of data. International trade prediction inherently involves a finite set of countries, and the use of very old data is often impractical due to the rapid evolution of the global economic and political landscape. Unlike domains such as natural language processing, where vast datasets are common, trade data is modest in scale. This scarcity is intrinsic to the problem and necessitates methods that can achieve high performance without requiring extremely large datasets.

The second challenge is the spatio-temporal nature of the task. An effective forecasting system must capture both spatial relationships (e.g., trade connections between countries) and the temporal evolution of these relationships. This means modelling how trade patterns shift over time while incorporating the structural interdependencies within the trade network.

Traditional econometric approaches, such as the gravity model, provide an interpretable framework grounded in economic theory, offering transparent parameter estimates and clear causal interpretations. However, gravity models are typically static in nature and may struggle to capture nonlinear dynamics or rapidly evolving trade patterns without extensive manual modification. Similarly, machine learning methods such as random forests can model complex nonlinear relationships and perform well with relatively small datasets, but they treat observations independently and do not explicitly account for the graph-structured nature of trade networks or their temporal evolution.

To address both the spatial and temporal dimensions of the problem within a unified framework, this study adopts the Spatio-Temporal Graph Convolutional Network (STGCN). The STGCN extends Graph Neural Networks (GNNs) by combining graph convolutions, which model spatial dependencies across the trade network, with temporal convolutions, which capture sequential changes over time. This dual capability allows the STGCN to leverage the strengths of GNNs in modelling network structure while overcoming their limitations in handling temporal dynamics. Compared with gravity models and random forests, the STGCN provides a data-driven yet structure-aware approach capable of capturing non-linear, dynamic, and interconnected patterns in international trade flows, making it a robust choice for this forecasting task.

STGCN Architecture The Spatio-Temporal Graph Convolutional Network (STGCN) was originally introduced by Yu et al. (2017) for traffic forecasting, with the aim of capturing both spatial and temporal dependencies within a unified deep learning framework. While first applied to traffic flow prediction, its underlying principles make it well suited to other spatio-temporal problems, including the prediction of international trade flows.

The core concept of STGCN is to represent the system as a graph, where nodes correspond to entities (e.g., countries) and edges represent relationships (e.g., trade links). This graph structure encodes spatial dependencies, while temporal convolutional layers capture how these relationships evolve over time. Unlike traditional approaches that address space and time separately, STGCN learns both simultaneously, enabling it to detect complex, dynamic patterns.

The architecture is composed of stacked *spatio-temporal blocks*, each containing:

1. **Temporal Convolution (GLU)** - A gated convolution applied along the time dimension for each node independently, extracting temporal features such as trends or seasonality.

2. **Spatial Graph Convolution ($\hat{A} \times X$)** - Uses the normalised adjacency matrix ($\hat{A}$) to aggregate information from neighbouring nodes, capturing spatial dependencies such as inter-country trade relationships.

3. **Temporal Convolution** - A second convolution along the time axis that refines the temporal representation after spatial aggregation.

Two such blocks are applied sequentially, allowing the network to learn deeper and more abstract spatio-temporal features. The output is then passed through:

- **Final Temporal Convolution** - Reduces the temporal dimension and prepares features for prediction.
- **Flattening and Fully Connected Layer** - Transforms the multi-dimensional feature map into a vector, applies dropout for regularisation, and produces the final predictions.

In the context of trade forecasting, the input is a tensor $[C, V, T]$ where C is the number of features (e.g., GDP, tariffs), V is the number of countries, and T is the number of time steps (e.g., years). The output is $[V]$, representing one prediction per country, such as projected export volumes.

Product-Level Implementation In this study, the STGCN architecture is applied at the *product level*, using the HS2 classification as the product grouping. The pre-processing pipeline constructs input tensors $X_{\text{full}} \in \mathbb{R}^{T\times V\times C\times P}$ and target tensors $Y_{\text{full}} \in \mathbb{R}^{T\times V\times V\times P}$, where:

- T is the number of time steps (years),
- V is the number of countries,
- C is the number of node-level features, and
- P is the number of products at HS2 level.

For model training, the dataset generator extracts **one product at a time** from these tensors. Each training sample therefore consists of:

- An input sequence $X_p \in \mathbb{R}^{C\times V\times T_{\text{in}}}$ for a specific product p, where T_{in} is the input window length.
- The corresponding prediction target $Y_p \in \mathbb{R}^{V\times V}$, representing the pairwise trade flows for that product in the next time step.

The same STGCN parameters are shared across all products, enabling the model to learn generalisable spatio-temporal patterns while still receiving product-specific node features and target matrices. The graph structure $A_{\text{hat}} \in \mathbb{R}^{V\times V}$, computed from trade and distance information, is also shared across products and remains fixed within each training run.

The forward pass for a single product p proceeds as follows:

1. The input tensor X_p passes through two sequential *Spatio-Temporal Blocks*, each consisting of temporal convolution, spatial graph convolution using A_{hat}, and a second temporal convolution.
2. A final temporal convolution further compresses and refines the time dimension.
3. The resulting feature map is flattened and passed through a fully connected network to produce a $V \times V$ matrix of predicted trade flows for the next time step.
4. A ReLU activation ensures non-negative predictions, consistent with trade flow values.

This design can be viewed as a multi-task learning setup, where each product provides separate training examples but shares the same spatio-temporal feature extraction backbone. Product specificity arises from the product-dependent node features (e.g., outflow, inflow, weighted distances, and sanctions) and the corresponding pairwise target flows.

Data Preparation Bilateral trade values were taken from the *imports* table; for each exporter–importer–year triple we used the reported import value as the flow of interest. Pairwise distances (`Distance_km`) were computed as great-circle geodesics using `geopy`'s `geodesic` function and OpenStreetMap's Nominatim geocoder for country centroids (Developers, 2025; OpenStreetMap contributors, 2025). Country-year indicators (e.g., GDP, land area, population density, sectoral GVA, rural/urban population shares) were sourced from the World Bank Open Data portal[1], with short gaps inside a country's time series imputed via forward-then-backward fill at the country level (FFILL/BFILL). Bilateral sanctions were merged from the *Global Sanctions Data Base (GSDB)* (Felbermayr et al., 2020). All this was done as a part of preprocessing for `countries_metadata.csv`

Graph Construction Following data preprocessing, the dataset was transformed into a time-stacked, directed graph representation. A list of all unique countries was compiled, and each country was assigned a unique integer index. An analogous procedure was applied to the set of unique years, ensuring that both spatial and temporal dimensions could be referenced efficiently. From the node attribute table, only the feature columns were retained, excluding identifiers such as country name, country code, and year. These features were

[1] https://data.worldbank.org/

exporter	importer	year	trade_value	sanction_exists	preferred_pair	Distance_km
C004	C008	2005	1500.000000	0	0	4345.59
C004	C008	2008	19.280466	0	0	4345.59
C004	C008	2009	1280.000000	0	0	4345.59
C004	C008	2010	1865.147866	0	0	4345.59
C004	C008	2012	2855.090817	0	0	4345.59

Figure 18. Sample of the edge attribute table containing exporter-importer trade data with year, trade value, sanctions, preferred trading pairs, and distances.

country	country_code	year	GDP	land_area	population_density	agriculture_gva	industry_gva	rural_pop	urban_pop
Afghanistan	C004	2000	3.521418	652230.0	30.863847	27.928194	19.766110	77.922	22.078
Afghanistan	C004	2001	2.813572	652230.0	31.099929	27.928194	19.766110	77.831	22.169
Afghanistan	C004	2002	3.854235	652230.0	32.776961	38.627892	23.810127	77.739	22.261
Afghanistan	C004	2003	4.520947	652230.0	34.854344	37.418855	22.710864	77.647	22.353
Afghanistan	C004	2004	5.224897	652230.0	36.123230	29.721067	26.226790	77.500	22.500

Figure 19. Sample of the node attribute table containing country-level indicators such as GDP, land area, population density, sectoral gross value added, and rural/urban population shares.

stored in a five-dimensional array structured to match the input format required by the model: batch size $\times$ features $\times$ years $\times$ countries $\times$ 1. Each entry in this array corresponds to the value of a given feature for a specific country-year combination.

The edge attribute table was used to construct a time-indexed adjacency matrix in which each element represents the trade value from an exporter to an importer for a given year. To mitigate the influence of extreme values, a $\log(1+x)$ transformation was applied to all trade values. Instances with no reported flows were retained as zeros. The resulting node feature array and adjacency matrix were subsequently converted into PyTorch tensors, with the adjacency matrix also duplicated to serve as the target variable for model training. This process produced three tensors (node features, adjacency matrix, and target) whose dimensions reflected the number of features, countries, and years in the dataset.

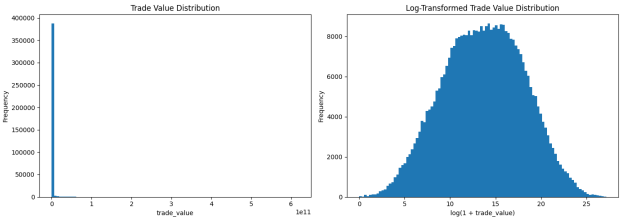


Figure 20. Distribution of trade values before (left) and after (right) applying the $\log(1+x)$ transformation. The transformation mitigates the influence of extreme values, resulting in a more balanced distribution suitable for model training.

Training Methodology A sliding temporal window approach was employed to generate training samples from the longitudinal trade network data. Each instance consisted of a fixed-length historical context of 16 consecutive years and

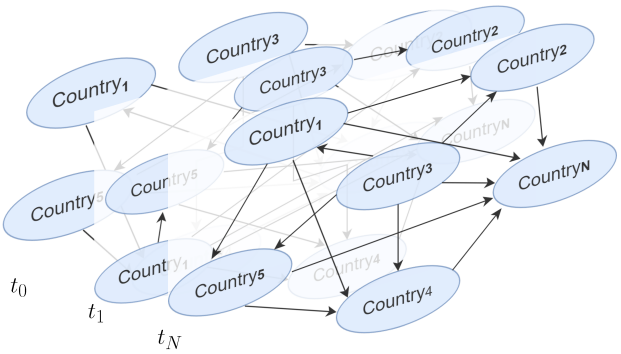


Figure 21. Isometric view of each time frame for exports in a graph shape starting from t_0 and ending with t_N.

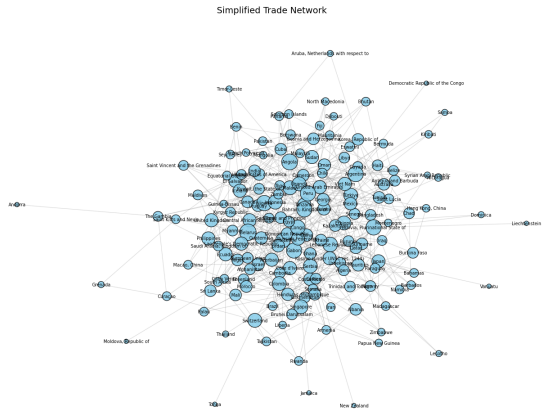


Figure 22. Simplified Trade Network for a particular year visualised using a spring layout (force-directed) algorithm, where connected countries are pulled closer like springs and unconnected nodes repel each other, revealing clusters of closely linked economies.

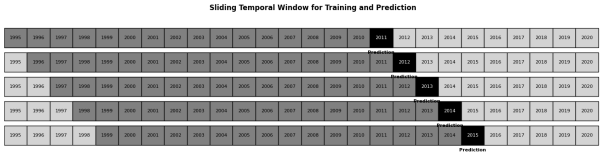


Figure 23. Sliding temporal window used for model training and evaluation. Each row represents one training instance built from a fixed-length context of 16 years (dark gray) and a single prediction year (black). Years to the right of the prediction window (light gray) are not used to form that instance. The window advances one year at a time (2011, 2012, 2013, 2014, 2015 in the example), producing multiple overlapping samples. Instances are split into training and validation sets; the most recent contiguous block of windows is held out as a test fold for final reporting.

a one-year-ahead prediction target. The window advanced by one year at a time, producing overlapping sequences that capture temporal dependencies while maximising data utilisation. The dataset was split chronologically, with approximately 80% of the instances assigned to the training set and the remaining 20% to the validation set. A separate contiguous block of the most recent years was held out as the test set to assess out-of-sample generalisation.

The STGCN was implemented in `PyTorch` and trained using the Adam optimiser with a learning rate of 10^{-3} and weight decay of 10^{-4}. The loss function was the mean squared error (MSE) between the predicted and actual target values. A ReduceLROnPlateau scheduler was applied to adapt the learning rate based on validation loss, and early stopping with a patience of 50 epochs was implemented to prevent overfitting. The best-performing model on the validation set was retained for final evaluation on the held-out test set.

All experiments were executed on Google Colab using an NVIDIA A100 GPU, ensuring efficient computation for both training and evaluation.

Evaluation Method Model performance was assessed on the held-out test set derived from the most recent contiguous block of temporal windows. For each window, the trained STGCN was evaluated in inference mode using mini-batches without gradient computation. The true target for each instance corresponded to the aggregated next-year inbound trade flows per country, and predictions were compared against these ground-truth values.

Four metrics were computed to quantify predictive accuracy. The mean absolute error (MAE) is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \,,$$

measuring the average magnitude of prediction errors in the same units as the data. The mean squared error (MSE) is given by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \,,$$

which penalises larger errors more heavily due to the squaring term. The coefficient of determination (R^2) is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

representing the proportion of variance in the observed data explained by the model.

To address the wide variation in trade flow magnitudes across countries and years, we additionally report the mean absolute percentage error (MAPE), a scale-independent measure defined as

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| ,$$

where $y_i \neq 0$ for all terms to avoid division by zero. MAPE was computed for each prediction year by averaging the absolute percentage errors across all countries, excluding cases with zero observed values. The average MAPE across years, along with its standard deviation, provides a measure of both overall accuracy and temporal consistency. This combination of metrics ensures that model performance is evaluated fairly across flows of different scales, avoiding bias toward high-magnitude trade values.

Algorithm 1 STGCN–Pairwise Model for Predicting Trade Values

Train_Dataset, A: adjacency matrix, X: input features, Y: target values, K: epochs

Initialise STGCN–Pairwise model with given hyperparameters
Split dataset into training and validation sets
Create data loaders for training and validation sets
Define optimiser and loss function
for $epoch = 1$ to K **do**
  for each batch (X_b, Y_b) in training data **do**
    $\hat{Y}_b \leftarrow f_\theta(A, X_b)$
    Compute loss $\mathcal{L}(\hat{Y}_b, Y_b)$
    Backpropagate and update model weights
  end for
  Evaluate model on validation data
  Compute evaluation metrics (RMSE, MAE, MAPE)
  Save model if validation loss improves, else update early stopping counter
end for

Algorithm 2 STGCN with Product Embedding

Require: Train_Dataset; A (country adjacency); X (features); Y (targets); M (observed-mask); years; K epochs

Build `TemporalTradeDatasetPE` with sliding window T_{in} and product index p
Deterministic split into train/validation; create data loaders
Initialise STGCN_PairwisePE with product embedding dimension d; concat embedding as extra channels
Choose loss (Masked & Weighted MSE), optimiser (Adam), LR scheduler, and early stopping
for $epoch = 1$ to K **do**
  Train: for each batch $(X_b, Y_b, M_b, \_, p_b)$
    $\hat{Y}_b \leftarrow f_\theta(A, X_b, p_b)$; $loss \leftarrow \mathcal{L}(\hat{Y}_b, Y_b, M_b)$
    Backpropagate, clip gradients, update parameters
  Validate: for each batch $(X_v, Y_v, M_v, \_, p_v)$ compute val_loss
  Step LR scheduler with val_loss; save best model; update early-stopping counter
  if patience exceeded **then**
    break
  end if
end for
Return trained model θ^*, dataset splits, and final A

6. Analysis and Results

6.1. Mirror Statistics and Temporal Imputation

6.1.1. Trade Coverage and Reporting Gaps

To quantify the completeness and directional bias of trade reporting, we construct a unified mirror dataset by merging cleaned import and export panels at the HS chapter level. These import and export panels were first constructed independently to ensure complete coverage of all reported trade flows, and then merged using a full outer join on `(reporter, partner, year, HS chapter)`. This approach ensures that no valid trade flow is omitted and allows us to fill in gaps where trade is reported only from one side.

Out of approximately 15 million trade records, 31.5% are missing import values, while 43.4% lack exports. However, only 14.5% of records have no trade value reported from either direction. This large difference highlights the value of mirror merging: most missing records from one side are available from the other, and estimation can be avoided when the partner country has already reported the flow.

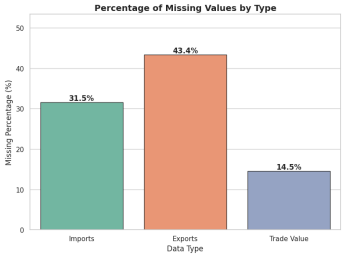


Figure 24. Percentage of missing values by type (Imports, Exports, Trade Value).

We also compute the number of records reported only by importers (33.8%), only by exporters (19.9%), and those reported by both (46.3%). The corresponding visualization in Figure 25 reveals a clear asymmetry in directional reporting, with import data being more complete. This informed our prioritization of import values in the final trade construction.

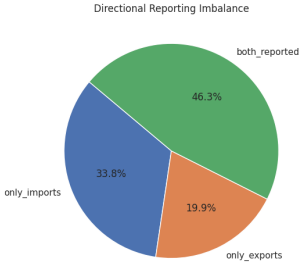


Figure 25. Directional reporting imbalance in the mirror dataset.

Temporal Trends in Missingness. Long-run analysis of missingness over time reveals sharp increases in reporting gaps after 2020 (Figure 26). This trend is particularly visible for exports, suggesting that disruptions (e.g., COVID-19, reporting lags) may have had asymmetric effects on trade reporting. Notably, the 2024 dip is not directly comparable - WTO reporting for that year is still ongoing, and many countries have yet to submit their data. As such, this observation can be ignored for our purposes. Overall, the persistent rise in missingness reinforces the need for systematic imputation to ensure coverage in recent years.

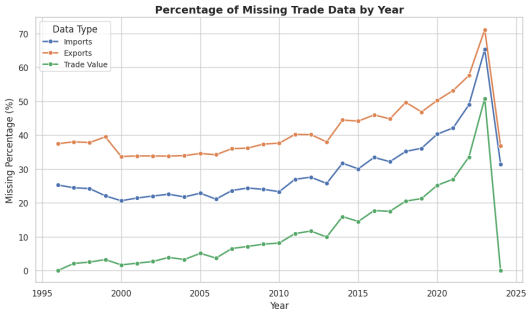


Figure 26. Year-wise missingness rates in import, export, and final trade values (based on mirror merge only).

6.1.2. Mirror Discrepancy Metrics

To evaluate consistency between reports, we compute the absolute discrepancy and mirror ratio for all trade flows reported by both countries. Discrepancy is defined as the difference between the reported import and export value, and the mirror ratio as their quotient. Both metrics are used to diagnose asymmetric reporting patterns.

- The average mirror ratio across matched records is 2113.02, with a very high standard deviation of 1.1 million, indicating substantial variance.
- Over 10.6% of matched records fall into extreme discrepancy cases - defined as a mirror ratio below 0.1 or above 10.

Mirror Ratio Distribution. Figure 27 confirms the extreme skewness of the mirror ratio distribution, with most values concentrated near 1 but with long tails on both ends. This suggests a general agreement in reported trade flows between countries, while also highlighting significant reporting inconsistencies in a subset of cases.

A sharp spike is observed near zero. This reflects cases where reported import values are extremely small while export values are substantial. Such imbalances often result from under-reporting on the importer's side, classification

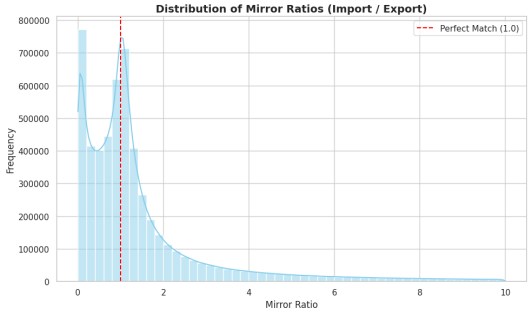


Figure 27. Distribution of mirror ratios (Import / Export) across all matched flows.

mismatches, or valuation issues. Since imports are commonly reported CIF (Cost, Insurance, Freight) and exports FOB (Free on Board), this discrepancy usually leads to mirror ratios above 1. Ratios significantly below 1 are harder to explain through valuation alone and may indicate timing misalignments or systemic reporting failures.

A log-scale visualization of the discrepancy versus trade value (Figure 28) further reveals that higher-value flows are more prone to large absolute discrepancies, possibly due to the greater material impact of underreporting or aggregation errors at scale.

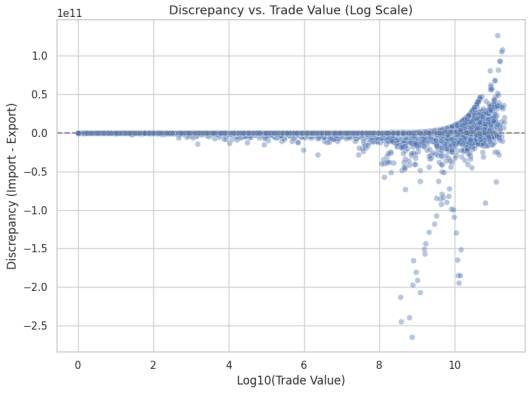


Figure 28. Discrepancy vs. Trade Value (Log-Log Scale).

6.1.3. Temporal Interpolation and Estimation

To recover missing values without relying on predictive models, we employ a simple adjacent-year imputation method. For each `(importer, exporter, chapter)` triplet, we compute lag and lead values using window functions, and estimate missing trade as described in the methodology section 5.1.3

Estimation is performed separately for imports and exports to extract maximum value from existing data. This localized method preserves the temporal continuity of trade and avoids overfitting.

Performance Metrics. On records where the true value is available (used for validation), the adjacent-year method yields:

- **Imports:** MAE = 5.25M USD, RMSE = 119M USD
- **Exports:** MAE = 6.36M USD, RMSE = 150M USD

These error values, though large in absolute terms, are generally small relative to the scale of trade flows and visually show strong correlation with true values (Figures 29 and 30).

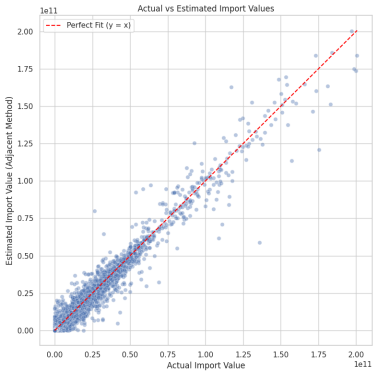


Figure 29. Actual vs. Estimated Import Values using adjacent-year averaging.

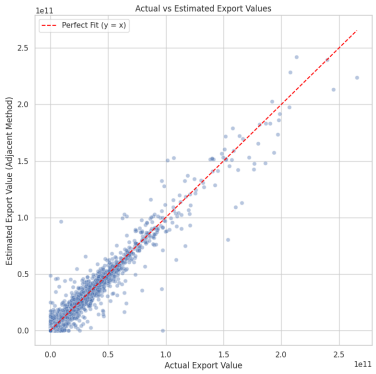


Figure 30. Actual vs. Estimated Export Values using adjacent-year averaging.

Estimation Coverage. Of the 14.5% of records needing estimation, 38.5% were successfully filled by the adjacent-year method. Specifically:

- 3.9% of all rows were filled using adjacent imports,
- 1.7% using adjacent exports,
- 8.9% of records still remained unfilled.

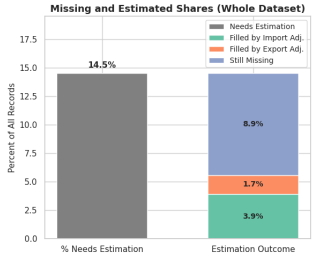


Figure 31. Shares of trade records requiring and filled by adjacent-year estimation.

Model Validation. To validate the adjacent-year estimation approach, we examine USA-UAE trade between 2015 and 2019 (Figure 32). Solid lines show reported values from both countries; dashed lines show values imputed using adjacent-year interpolation. Red corresponds to USA-reported imports, teal to UAE-reported exports.

The method aligns closely with reported data wherever available. Most notably, in 2018-the only year where the UAE did not report-imputed values continue the plausible upward trend, avoiding artificial dips caused by missing entries.

In years where both countries reported (e.g., 2017), imputed and actual lines converge, validating the method's accuracy. The USA's reporting is complete across all years, and its predicted series (red dashed line) tracks actual values closely, further supporting internal consistency.

Overall, this case illustrates how adjacent-year estimation fills real-world reporting gaps without distorting trends. The consistent gap between USA and UAE values-visible throughout-reflects common discrepancies in mirror statistics, driven by re-exports, valuation methods, or timing lags. Despite this, the imputed paths remain stable and directionally consistent.

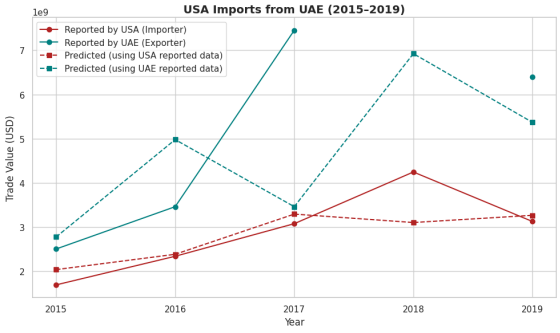


Figure 32. USA imports from UAE (2015-2019): Comparison of actual and predicted trade flows.

6.1.4. Discussion and Limitations

The mirror statistics and adjacent-year imputation pipeline significantly improves the coverage, continuity, and consistency of trade data, but several structural assumptions and design choices introduce important limitations:

- **Static Assumption of Country Existence:** The panel assumes that all countries present from 1996 to 2023 continue to exist and report trade unless marked inactive for specific years. This overlooks geopolitical changes-such as the dissolution of Yugoslavia, the split of Sudan, or the emergence of new states-where countries cease to exist or begin reporting only partway through the timeline. Once a reporter-partner-chapter link is observed, it is assumed to persist thereafter, potentially generating artificial "missing" values even after a country ceases to trade.

- **Assumptions in Full Panel Completion:** A trade flow is considered missing if the reporter did not report in a given year but the chapter-partner link was seen previously. This assumes established trade links persist annually, which can falsely indicate missingness when relationships naturally lapse.

- **Strict Year-Based Reporting Logic:** If a country reports any trade in a year, all relevant flows are assumed complete for that year. This avoids inflating missingness but risks overlooking unreported flows within active years.

- **Short Imputation Window:** Adjacent-year estimation is limited to a ± 1 year range, reducing over-smoothing risk but limiting recovery for multi-year gaps, especially for irregular reporters.

- **Assumed Temporal Smoothness:** The method presumes gradual year-to-year changes, which may not hold for volatile or policy-sensitive goods subject to shocks such as embargoes or supply chain disruptions.

- **Unadjusted Directional Bias:** Import values are prioritised for completeness but without correcting for systemic biases like underreporting or inconsistent valuation across countries, leaving residual discrepancies.

- **Chapter-Level Aggregation:** Aggregating to HS chapter level improves tractability but obscures intra-chapter variation and commodity-specific volatility.

- **No Tariff or Valuation Adjustment:** Reported values are unadjusted for import duties, tariffs, or valuation differences (e.g., CIF vs. FOB), which can inflate imports in high-tariff regimes and contribute to mirror asymmetries.

In summary, the mirror statistics methodology introduces a transparent and efficient framework for reconstructing trade panels. Its logic preserves directional asymmetry, handles reporting gaps, and increases usable data through low-assumption interpolation. Still, it remains constrained by structural assumptions about country continuity, reporting behavior, and temporal stability.

Looking ahead, the approach could be extended in several ways. Adjusting reported values for tariff structures, freight costs, and CIF/FOB valuation differences could reduce systematic import-export asymmetries, particularly in high-tariff regimes. Incorporating dynamic country lifecycles would better reflect geopolitical events such as state dissolution or unification, avoiding the creation of artificial missing values. Integrating product-level volatility modelling, and using confidence-weighted merging based on reporter quality scores could further improve coverage while maintaining reliability. These refinements would enhance the method's applicability for both historical reconstruction and forward-looking trade analysis.

6.2. Clustering and Gravity Model

The analysis builds on the methodological framework in which country pairs were clustered into macroeconomically similar groups over rolling five-year windows, and the gravity model was estimated separately for each (period, cluster) using a single representative pair with complete data. While clustering evaluation is independent of the gravity estimation, the quality of these clusters ultimately shapes the homogeneity assumptions underlying each regression group. This design avoids pooling structurally dissimilar relationships but yields only five time-series observations per regression, constraining the model's predictive capacity.

The following sections evaluate the coherence and separability of the clustering outputs and the predictive performance of the gravity model within this clustered estimation framework.

6.2.1. Evaluation of Clustering Approaches

To evaluate the effectiveness of the three clustering algorithms applied to country pairs, we conduct a quantitative and visual analysis based on the clustering outputs generated for each 5-year period. The evaluation is designed to assess both the internal consistency of the clusters and their structural separability in feature space. All metrics and plots are computed independently for each 5-year window to respect the temporal evolution of macroeconomic characteristics.

Silhouette Score. For each 5-year period, we compute the Silhouette Score, a standard internal validation metric that quantifies how well-defined the clusters are. This score is calculated using the z-score normalized feature vectors for all clustered observations, excluding outliers (i.e., data points labeled -1 in HDBSCAN). The silhouette score for a given country pair reflects the difference between its average distance to other members of its cluster and its average distance to members of the nearest neighboring cluster. Scores closer to 1 indicate well-separated, compact clusters, while lower or negative scores suggest poor clustering quality. This metric is particularly useful in the absence of ground truth labels, as it provides a data-driven assessment of cluster tightness and separability.

Cluster Size Distribution. To analyze the composition of clusters, we compute and visualize the distribution of cluster sizes for each 5-year period. This helps assess whether the algorithm has produced balanced partitions or formed disproportionately large or small clusters. A balanced cluster size distribution indicates that the algorithm is capturing heterogeneity across country pairs, while highly skewed distributions may signal the presence of dominant macroeconomic regimes or overly coarse partitions. The size distributions are displayed using bar plots, one per time period, arranged in a grid for comparability.

PCA Visualization. To visually inspect the separability of clusters in the original feature space, we perform Principal Component Analysis (PCA) on the standardized feature vectors for each period. The first two principal components are used to project the country pairs onto a two-dimensional plane, and the resulting scatter plots are color-coded by cluster label. While PCA does not preserve all high-dimensional relationships, it provides a useful approximation for assessing whether the assigned clusters form distinguishable groupings in a reduced feature space. Subplots for each period are arranged in a 2×3 layout for compact visual comparison.

These three evaluation tools-silhouette score, cluster size analysis, and PCA projection-are jointly used to understand the internal structure and coherence of the clusters across different time windows. The same evaluation procedure is applied to the outputs of all three clustering algorithms: K-Means, Hierarchical Clustering, and HDBSCAN.

Clustering Results

This subsection summarizes the performance of each clustering algorithm using the evaluation framework described above. We report quantitative metrics (Silhouette Scores) and qualitative diagnostics (PCA plots and cluster size distributions) for each 5-year period.

Silhouette Scores: The table below reports silhouette scores for each method across rolling 5-year periods. Higher scores indicate better-defined and more compact clusters.

Method	2000–2004	2005–2009	2010–2014	2015–2019	2020–2024
K-Means	0.195	0.164	0.168	0.183	0.162
Hierarchical	0.168	0.126	0.139	0.155	0.615
HDBSCAN	0.093	0.139	0.108	0.012	0.179

Table 6. Silhouette scores by clustering method and time period.

Cluster Size Distributions and PCA Visualizations: To visually assess the performance of each clustering method across time, we present two diagnostic plots for each: (1) the distribution of country pairs across clusters, and (2) a 2D principal component projection of the clustered feature space. Both are shown across all five rolling time periods using a consistent 2-column by 3-row subplot layout.

Figure 33 and figures 51, 52 in Appendix 8 respectively show the results for K-Means, Hierarchical, and HDBSCAN clustering. The first row of each figure corresponds to early time periods (e.g., 2000–2004), while the last row covers the most recent period (2020–2023). Cluster size distributions highlight whether a method forms balanced groupings or exhibits dominance by large clusters. The PCA plots enable visual inspection of how well-separated and compact the resulting clusters are in the transformed feature space.

Clustering Method Comparison and Selection

To evaluate the quality of clustering across methods and over time, we employ both quantitative and visual tools: silhouette scores (Table 6) and cluster evaluation plots (figure 33 and figures 51, 52 in Appendix 8). Silhouette scores measure how well each data point fits within its assigned cluster compared to others, providing an overall indicator of intra-cluster cohesion and inter-cluster separation. The accompanying cluster size histograms and PCA projections offer complementary visual insight into the structure and distribution of clusters over time.

K-Means Clustering. K-Means delivers the highest silhouette scores among all methods across most periods, with values ranging from 0.162 to 0.195. While these scores are moderate, they remain positive and consistent, indicating reasonably distinct and cohesive clusters. However, the cluster size distributions are highly skewed, with a few clusters containing the vast majority of points. This imbalance is reflected in the PCA projections, where dense, overlapping groups suggest limited separation in the lower-dimensional space-particularly in later periods. Nevertheless, the method maintains relatively stable behavior and captures some consistent structural patterns over time.

Hierarchical Clustering. Hierarchical clustering shows more variability in performance, with silhouette scores ranging from 0.126 to 0.615. While it performs similarly to K-Means in earlier periods, its silhouette score peaks sharply

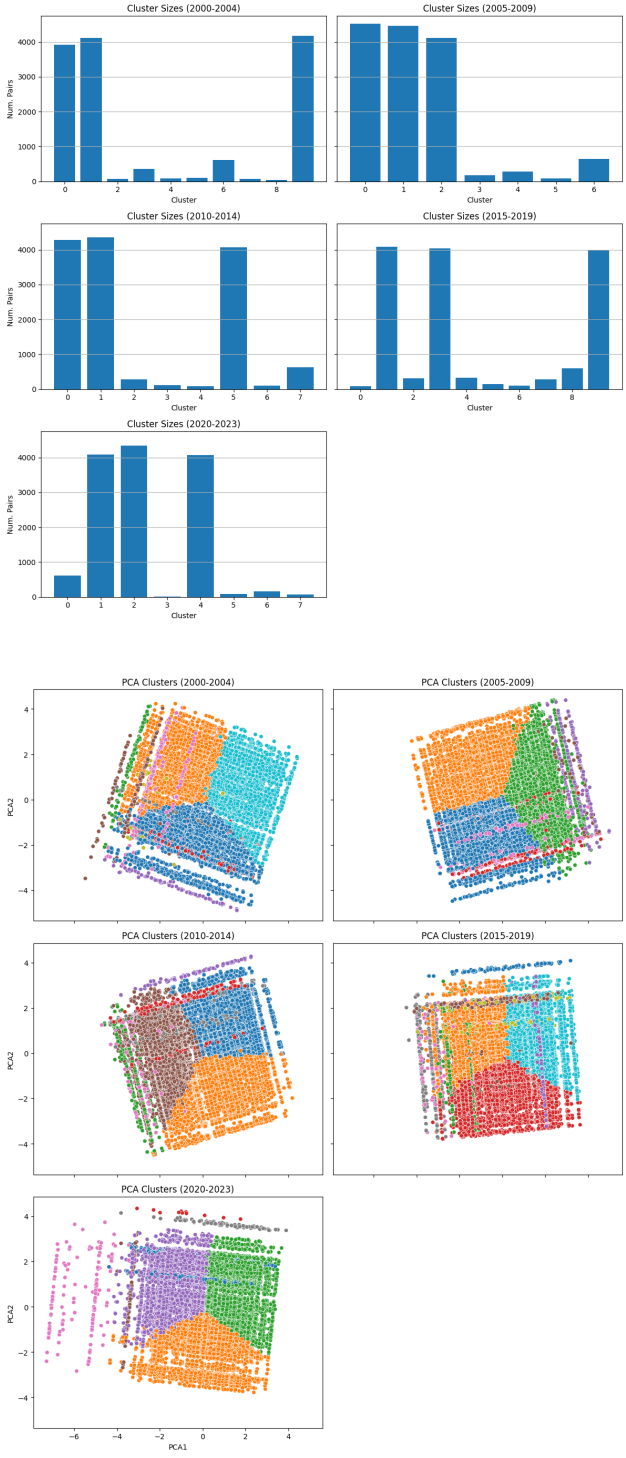


Figure 33. K-Means clustering evaluation: (Top) Cluster size distributions for each 5-year period in a 3×2 layout; (Bottom) PCA visualizations showing 2D projection of clusters using the first two principal components.

in the final period (2020-2023), indicating strong separability during that window. Cluster size distributions are

moderately imbalanced, and the PCA plots reveal clearer separation in some periods. However, the method struggles to maintain this clarity consistently across all periods, showing more overlap and dispersion in the projections for earlier years.

HDBSCAN. HDBSCAN yields the weakest silhouette scores, with values close to zero or slightly negative in some periods. This suggests that the formed clusters lack clear cohesion or are not well-separated. Although HDBSCAN is designed to identify noise and varying-density clusters, the PCA visualizations show substantial overlap and less defined boundaries between clusters. Its cluster size distributions also reveal high imbalance, and the method generally fails to form stable groupings across time.

Method Selection. Based on a combination of silhouette scores and visual inspection of cluster coherence and separation, **K-Means is selected as the preferred clustering method**. Despite the presence of some imbalance in cluster sizes, it consistently delivers the most reliable clustering structure across all time periods, maintaining stable scores and recognizable patterns in the data. Its performance is robust across both visual and quantitative metrics, making it the most suitable choice for the gravity model.

Qualitative Validation of K-Means Clustering

To assess the interpretability of K-Means clustering, we fix the United States and examine its bilateral clustering assignments during the 2015-2019 period. Figure 34 shows a world map where each country is colored according to the cluster assigned to its pair with the USA. The USA is highlighted in black. Since clustering was performed on country *pairs*, this visualization reveals how the USA's trade-relevant relationships are grouped, not how countries are clustered in isolation.

The features used for clustering include economic scale (GDP, sectoral GVA), demographic structure (population density, urban/rural distribution), land area, political alignment (sanctions, preferred-pair indicator), and geographic distance. As such, countries that receive the same cluster color in this map share a similar overall profile in how they relate to the USA - even if they differ from each other.

The structure aligns well with economic and political reality. Canada, Australia, and parts of Latin America are assigned to the same cluster - reflecting shared traits such as high GDP, sectoral similarity, stable diplomatic ties, and trade agreements with the US. Conversely, countries like Iran, Russia, and North Korea fall into distinct clusters, consistent with sanction regimes and limited or hostile economic relations.

Importantly, countries with the same color are *not* clustered together globally - only their *pairwise relationship with the USA* is being analyzed. The clustering reflects how each country's bilateral dynamic with the US compares to that of other countries. This pairwise perspective offers a robust, interpretable validation of the K-Means output: the model correctly groups relationships that are economically and geopolitically similar, confirming the suitability of K-Means for structuring the estimation space in the gravity model.

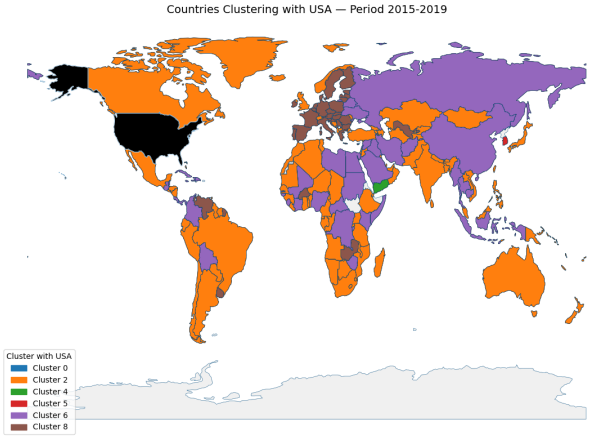


Figure 34. Countries paired and clustered with the USA (2015–2019) using K-Means. Each country is colored by the cluster assigned to its pair with the USA. The USA is shown in black.

6.2.2. Gravity Model Evaluation

To estimate bilateral trade flows between countries, we implemented a standard linear regression model using macroeconomic and geographic indicators such as GDP, land area, population density, and bilateral distance. This forms the basis of a simplified gravity model applied to the same country-pair dataset used in the clustering analysis.

Regression Metrics

Model performance was evaluated using the Mean Absolute Error (MAE) and the coefficient of determination (R^2):

- **Mean Absolute Error (MAE)**: 5.28×10^{20}
- **R-squared** (R^2): -1.18×10^{22}

The negative R^2 value indicates that the model performs substantially worse than a naive mean-based baseline. The enormous MAE further highlights the model's inability to approximate actual trade values with any meaningful precision. Together, these metrics suggest that the model fails to capture the underlying structure of the data.

Visual Diagnostics

Figure 35 plots predicted versus actual trade values in log-log scale. In a well-fitted model, points would cluster around the red diagonal line indicating perfect prediction. Here, the predicted values form dense horizontal bands, far removed from the actual values. This pattern reflects poor sensitivity to the variation in the actual data and possible saturation effects in the model's output - symptoms of underfitting or numerical instability.

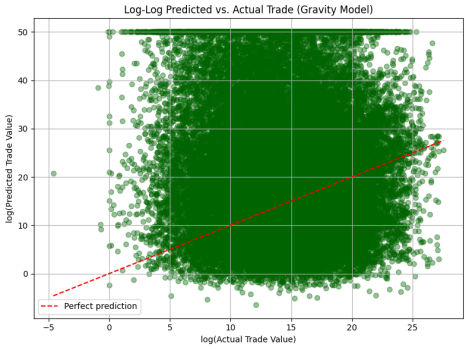


Figure 35. Log-log scatter plot of predicted vs. actual trade values for all country pairs. The red dashed line shows perfect prediction. The vertical spread and compression of predicted values reflect a severe model mismatch.

To illustrate this failure at the pair level, Figure 36 compares actual and predicted trade between the USA and UAE over 2015–2019. While actual trade fluctuates within a narrow log-scale band, the predicted values are nearly flat and consistently underestimate the true magnitude by several orders. This demonstrates that even for high-volume and politically aligned trade partners, the model fails to adjust for structural variation or temporal dynamics.

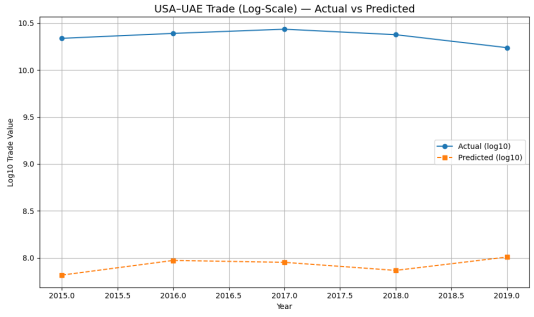


Figure 36. Log_{10} trade values between the USA and UAE (2015–2019). The model persistently underpredicts actual trade and shows limited temporal responsiveness.

6.2.3. Discussion and Limitation

The gravity model's poor predictive performance can be traced to several structural and methodological constraints in the pipeline:

- **Sparse Model Training:** Each cluster-period regression was fit on only a single country pair with complete data, yielding just five training points (one per year). This extremely small sample size makes coefficient estimates highly sensitive to noise and idiosyncrasies in that pair's trajectory.
- **Temporal Averaging in Clustering:** Clustering relied on averaged features (e.g., mean GDP, population) over each 5-year period, assuming structural stability across all years. This can misalign static clustering inputs with the year-to-year variation the regression seeks to predict.
- **Loss of Directionality:** Country pairs were treated as unordered to avoid feature duplication, removing the ability to model asymmetric flows (e.g., U.S. exports more to Mexico than vice versa).
- **Assumed Linearity of Log-Transformed Features:** The log-log specification assumes strictly linear relationships between $\log(\text{Trade})$ and $\log(\text{Features})$, overlooking nonlinear effects such as thresholds, saturation, and interactions.
- **No Product-Level Disaggregation:** All trade was aggregated across products. While disaggregating by product is theoretically possible, it would create many sparse cluster-period combinations and require incorporating product-specific drivers (e.g., production capacity, tariffs), vastly increasing data and model complexity.
- **Omission of Policy Instruments Beyond Sanctions:** The model included sanctions and a preferred-pair indicator but excluded major policy drivers (e.g., NAFTA, EU customs union, tariff schedules, non-tariff barriers), limiting explanatory power where economic similarity does not translate into trade.
- **Ignored Interdependence Between Pairs:** Each bilateral pair was modelled independently, omitting network effects such as supply-chain linkages or trade diversion via third countries.
- **Feature Limitations and Multicollinearity:** Core indicators (GDP, sectoral value-added, population) are highly correlated, creating multicollinearity that is problematic with such small training samples. Important relationship variables (common language, time zones, colonial history, cultural ties) were omitted due to resource constraints, further limiting the model's discriminative capacity.

In theory, the gravity model remains one of the most empirically robust frameworks for explaining bilateral trade flows, with decades of evidence supporting its predictive strength. The disappointing performance in our case appears to stem less from the theoretical foundations and more from the way it was operationalized particularly the reliance on coarse clustering to reduce the number of models trained, which, while motivated by the idea of grouping structurally similar pairs (e.g., developed-developed versus developed-developing relationships), likely oversimplified meaningful variation.

Future implementations could benefit from hybrid approaches that retain the interpretability of the gravity framework while leveraging machine learning to dynamically identify optimal groupings, capture nonlinear interactions, and incorporate richer contextual features such as sector-specific trade patterns, tariffs, and network dependencies. Such integration could preserve the model's theoretical strengths while addressing the limitations encountered in this application.

6.3. Random Forest: MissForest Approach

Cross-Validation Performance

MissForest demonstrates exceptional performance on the training dataset validation:

- **$R^2 = 0.930$**: Explains 93% of variance in bilateral trade flows
- **RMSE = $2.25 billion**: Prediction error reasonable given extreme trade value range
- **MAE = $197 million**: Median absolute prediction error
- **Bias = 3.4%**: Minimal systematic error (mean predicted $732M vs actual $708M)
- **Convergence**: Algorithm achieved stability in 2 iterations, indicating robust optimization

Universal Test Set Performance

Out-of-sample testing on the independent test set revealed decent results, consistent with the challenges of external validation:

Random Forest (MissForest):

- **$R^2 = 0.603$**: Explains 60% of out-of-sample trade variance
- **RMSE = $6.69 billion**, **MAE = $554 million**
- **Bias = -48%**: Systematic under-prediction (conservative imputation)

Method	R^2	RMSE (Billions)	Bias
Random Forest	0.603	$6.69	-48%
Gradient Boosting	0.567	$10.52	-57%
XGBoost	0.570	$10.48	-54%

Table 7. Comparison of machine learning methods on universal test set.

Comparison with Alternative Methods: Random Forest emerged as the optimal method, achieving the highest out-of-sample R^2 while maintaining reasonable prediction errors and avoiding severe overfitting observed in gradient-based methods.

Feature Importance Analysis

Across all machine learning methods, feature importance rankings consistently validate gravity model theoretical foundations:

Random Forest Feature Importance (Universal Test Set):

1. **Importer GDP (32.9%) + Exporter GDP (22.1%) = 55% combined importance**
2. **Distance (21.5%)**: Geographic trade costs
3. **Country effects (20.9%)**: Unobserved heterogeneity
4. **Policy variables (2.6%)**: Sanctions and preferential agreements

This hierarchy perfectly aligns with gravity model predictions: economic mass (GDP) dominates trade determination, followed by trade costs (distance), with policy instruments playing supporting roles. The remarkable consistency across different algorithms provides robust empirical validation of gravity theory foundations.

Country-Level Performance Analysis

Granular validation reveals systematic heterogeneity in prediction accuracy across trading partners. Analysis of 13,493 bilateral relationships shows:

Export Prediction Performance:

- **Best performers**: Developed trade hubs and stable economies (Belgium, Hong Kong, EU, Japan)
- **Challenging cases**: Small volatile economies and conflict-affected regions (North Macedonia, Iraq, Gambia)
- **Pattern**: Large, diversified economies exhibit predictable export patterns

Import Prediction Performance:

- **Imports systematically harder to predict**: Export capacity demonstrates greater stability than import demand patterns
- **Best import predictions**: Large, stable markets (EU) and small economies with consistent import needs
- **Economic interpretation**: Supply-side factors more predictable than demand-side volatility

Bilateral Relationship Analysis:

- **Most predictable relationships**: Major economic partnerships (US-Japan, Canada-EU) and established regional trade corridors (South Africa-Nigeria)
- **Challenging relationships**: Sporadic trade between small economies and unusual trade routes lacking strong economic rationale
- **Trade size effect**: Most predictable relationships involve 25-fold larger average trade volumes ($41.2B vs $1.7B), indicating model effectiveness on economically significant flows

6.3.1. Estimator Count Optimization

To determine the optimal number of estimators for the Random Forest model, we systematically evaluated performance across different tree counts ranging from 10 to 300 estimators. The model was trained on clean observations and evaluated on the universal test set using RMSE, MAE, and R^2 metrics.

6.3.2. Performance Metrics by Estimator Count

Table 8 presents the performance evaluation of the random forest model across different numbers of estimators, measured by RMSE, MAE, and R^2 score.

6.3.3. Performance Metrics by Estimator Count

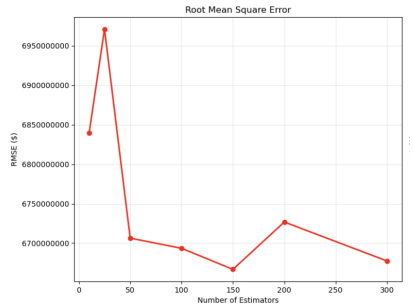
Table 8 presents the performance evaluation of the random forest model across different numbers of estimators, measured by RMSE, MAE, and R^2 score.

The results demonstrate optimal Random Forest performance at 150 estimators (R^2 = 0.605, RMSE = $6.67B), with a notable performance anomaly at 25 estimators where RMSE peaks at $6.97B and R^2 drops to 0.570. Performance shows rapid improvement from 25 to 50 estimators (3.7% RMSE reduction), followed by gradual convergence with minimal gains beyond 100 estimators. The marginal difference between 100 and 150 estimators (R^2 improvement of only 0.002) suggests that 100 estimators provide an optimal balance between predictive accuracy and computational efficiency for large-scale trade data applications, while the consistent MAE improvement (6.2% reduction from worst to best configuration) validates the ensemble's stability across the optimal range.

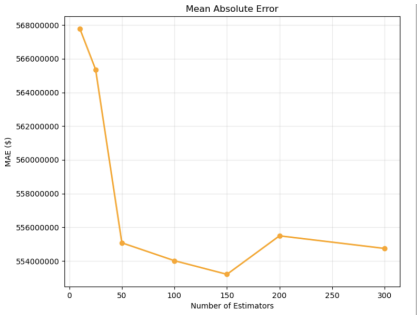
Table 8. Model Performance Across Different Numbers of Estimators

N Estimators	RMSE ($B)	MAE ($B)	R^2 Score
10	6.84	5.68	0.585
25	6.97	5.65	0.570
50	6.71	5.51	0.601
100	6.69	5.40	0.603
150	6.67	5.33	0.605
200	6.72	5.55	0.599
300	6.68	5.48	0.605

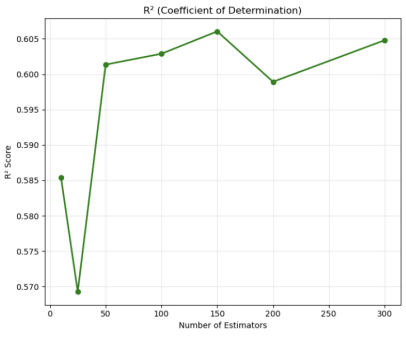
6.3.4. Performance Visualization and Analysis



((a)) Root Mean Square Error



((b)) Mean Absolute Error



((c)) R^2 Coefficient of Determination

Figure 37. Model performance metrics across different numbers of estimators

Key Findings

The analysis reveals several important patterns in Random Forest performance as the number of estimators increases:

Optimal Performance Range: The model achieves peak performance with 150 estimators, yielding the highest R^2 score of 0.605 and competitive error metrics (RMSE: \$6.67B, MAE: \$5.33B). Performance remains relatively stable between 100-300 estimators, suggesting diminishing returns beyond 150 trees.

Performance Anomaly: A notable performance dip occurs at 25 estimators, where both RMSE and R^2 worsen significantly compared to the 10-estimator baseline. This suggests that very low estimator counts may lead to unstable model behavior, possibly due to insufficient ensemble diversity.

Error Convergence: Both RMSE and MAE show general improvement from 10 to 150 estimators, with RMSE decreasing by approximately 2.5% and MAE by 6.2%. Beyond 150 estimators, performance stabilizes with minimal further improvement.

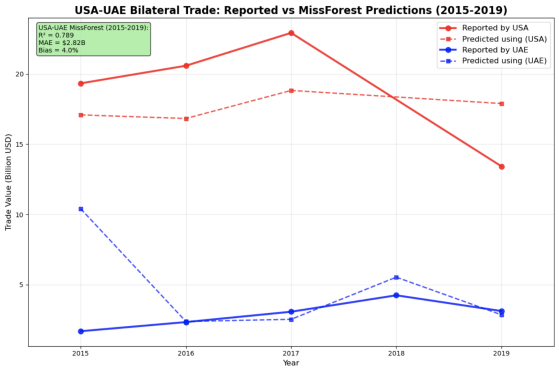


Figure 38. USA-UAE bilateral trade validation: MissForest predictions vs reported values (2015-2019). The model achieves R^2 = 0.789 with MAE = \$2.82B, successfully capturing temporal trends despite systematic reporting asymmetries between countries.

Model Validation. To validate the MissForest model's predictive capability, we tested its performance on USA-UAE bilateral trade data from 2015-2019, comparing predicted values against reported trade statistics from both countries. The model demonstrates strong predictive accuracy with an R^2 of 0.789 and a Mean Absolute Error of \$2.82 billion, representing only a 4.0% bias in predictions. The model effectively captures the general trend patterns, particularly the trade peak in 2017 and subsequent decline through 2019. However, the validation reveals systematic differences in trade reporting between countries, with USA-reported values consistently higher than UAE-reported values by approximately \$15-17 billion annually. The MissForest predictions align more closely with USA reporting patterns, suggesting the model may be influenced by the predominance of USA-perspective data in the training set. Despite these reporting discrepancies-which reflect well-documented asymmetries in international trade statistics-the model successfully predicts the temporal dynamics and magnitude of bilateral trade flows, validating its utility for trade value imputation and forecasting applications.

6.4. Product-Level Prediction Performance

The comparative analysis across HS classification levels reveals significant challenges in applying machine learning methods to product-specific trade prediction, with all approaches demonstrating poor predictive capability relative to bilateral-level analysis.

Performance Metrics by HS Level. All classification levels demonstrate insufficient predictive performance, with Random Forest predictions failing to achieve positive R^2 values on test data across all aggregation strategies. The analysis reveals that product-level trade prediction using country and temporal variables alone provides inadequate explanatory power.

Table 9. Random Forest performance across HS classification levels

HS Level	Rows	Test R^2	RMSE	Training R^2
2-digit	186,171	-0.109	\$112.4M	0.148
4-digit	199,306	-0.515	\$84.9M	0.318
6-digit	200,000	-0.075	\$102.4M	0.358

Among the tested approaches, 6-digit (product-level) classification performs relatively better on test data (R^2 = -0.075) while maintaining maximum product specificity, though all approaches exhibit negative predictive performance.

Challenges in Product-Level Prediction. The poor performance across all HS levels indicates several fundamental methodological challenges:

1. **Feature inadequacy**: Country and year variables provide insufficient information to predict specific product trade flows
2. **High-dimensional sparsity**: Product-level trade exhibits extreme sparsity with many zero flows between specific country-product combinations
3. **Missing product characteristics**: Absence of product-specific features (price, quality, substitutability, supply chains) limits predictive capability
4. **Temporal volatility**: Product-level trade shows high year-to-year variation that cannot be captured by simple country and temporal indicators

5. **Severe overfitting**: Large gaps between training and test R^2 (up to 0.833 for 4-digit level) indicate models cannot generalize beyond training data

Economic Interpretation. The systematic prediction failure across all HS levels provides important insights into the nature of international trade:

- **Country-level determinants**: Gravity model variables (GDP, distance, policies) operate primarily at bilateral country level rather than specific product categories
- **Product heterogeneity**: Individual products require specialized determinants beyond macroeconomic indicators
- **Market complexity**: Product-level trade depends on factors not captured in standard gravity frameworks (technology, branding, supply chain relationships)

6.5. Comparison with Bilateral-Level Results

The product-level results provide important context for interpreting the success of bilateral-level MissForest implementation:

Table 10. Performance comparison: bilateral vs product-level prediction

Analysis Level	R^2	Interpretation
Bilateral (Aggregated)	0.930	Excellent performance
Product (6-digit)	-0.075	Poor performance

This dramatic difference (R^2 = 0.930 vs -0.075) demonstrates that aggregation across products is essential for successful trade prediction using gravity model variables. The bilateral-level analysis benefits from:

- **Reduced noise**: Aggregating across all products smooths out product-specific volatility
- **Stronger signal**: Gravity model variables (GDP, distance) operate at the country level, not product level
- **Sufficient sample size**: Bilateral relationships provide adequate observations for robust parameter estimation

6.5.1. Silva et al. Model Performance

The Random Forest model outperformed the linear baseline:

- Random Forest R^2: 0.038
- Linear Baseline R^2: -0.316
- Improvement: +0.354
- RMSE: 0.92

Feature Importance Network features contributed 47.9% of predictive power, matching Silva et al.'s finding that network features provide about half the importance.

Top 5 Features:

1. Trade Concentration (29.0%)
2. Network Out-Degree Centrality (14.0%)
3. Network In-Degree Centrality (12.8%)
4. Network Betweenness Centrality (11.0%)
5. Network Degree Centrality (10.1%)

Hypothesis Validation All three Silva et al. hypotheses were confirmed:

- Network features improve forecasting (+0.354 improvement)
- Network features contribute 50% importance (47.9%)
- Non-linear models beat linear models

Economic Interpretation Network centrality shows how well countries connect to global trade. Higher centrality means better diversification and resilience. Trade concentration was the most important predictor, showing that diversification matters most for growth.

Network Effects Countries with better network positions have:

- More stable economies
- Better access to growth opportunities
- Less risk from single trade relationships

6.6. Discussion and Limitations : Miss Forest

6.6.1. Model Performance and Optimal Configuration

For international trade value prediction using Random Forest, 100-150 estimators provide optimal performance with R^2 scores around 0.60. Models with fewer than 50 estimators show reduced stability and accuracy, while configurations beyond 200 estimators offer minimal performance gains. The recommended configuration is 150 estimators for maximum accuracy or 100 estimators for balanced performance and computational efficiency.

The systematic under-prediction bias (-48% to -59%) across all methods suggests **conservative imputation behavior**, which is economically sensible for policy applications where overestimating trade relationships could lead

to resource misallocation. The method's superior performance on large, established trading relationships (US-Japan, Canada-EU) relative to small, volatile flows indicates that MissForest effectively prioritizes economically significant patterns over statistical noise.

6.6.2. Economic and Methodological Insights

Country-level heterogeneity reveals that prediction accuracy correlates strongly with economic stability and trade diversification. This finding has important implications for trade policy analysis: imputation methods are most reliable precisely where policy decisions have the greatest economic impact-in relationships between major trading partners.

Feature importance consistency across multiple algorithms provides robust empirical validation of gravity model foundations, confirming that despite decades of theoretical development, the fundamental drivers of international trade remain economic size, geographic proximity, and institutional relationships. Network features enhance prediction capability beyond traditional gravity variables, demonstrating that a country's position in global trade networks contains valuable information for forecasting. However, predictive performance varies dramatically by aggregation level: country-level bilateral trade prediction achieves strong performance (R^2 0.60), while product-level prediction remains challenging even with network-enhanced methods, though network features do improve upon traditional gravity model approaches for growth prediction.

6.6.3. Fundamental Limitations

Aggregation Requirements The analysis reveals critical limitations for applications requiring detailed predictions. Successful trade prediction necessitates aggregation to the bilateral level where gravity model theory applies most effectively. Product-level analysis across all HS classification levels demonstrates consistently poor performance, with high overfitting indicating fundamental limitations in using standard country-temporal features for commodity-specific prediction.

Scale-Dependent Accuracy Prediction accuracy varies significantly with trade relationship magnitude and stability. While the methodology excels for large, established trading partnerships, performance deteriorates for small, volatile trade flows. This scale dependency limits applicability for comprehensive trade policy analysis requiring uniform accuracy across all relationship types.

Product-Level Constraints The product-level analysis reveals important limitations for policy applications requiring detailed commodity-level trade predictions. Successful prediction requires aggregation to bilateral levels, and product-specific modeling may require specialized approaches incorporating supply chains, market structure, and product characteristics beyond traditional gravity variables.

Theoretical Validation vs. Predictive Limitations These findings reinforce the theoretical foundation of gravity models, which operate on the premise that bilateral trade relationships are primarily determined by country-level characteristics rather than product-specific factors. The success of bilateral-level imputation combined with poor product-level performance validates the economic logic underlying gravity theory while simultaneously highlighting the methodological boundaries for missing data imputation in international trade applications.

6.7. Graph Neural Network

6.7.1. Training and Generalization

For 500 epochs, the training and validation loss curves (Figure 39) show a steady and synchronized decline, indicating that the model learned effectively without exhibiting severe overfitting. By around epoch 300, both curves converge to stable values, with a small and consistent gap between them, suggesting robust generalization to unseen data.

This convergence behaviour strengthens confidence in the evaluation metrics reported. It also supports the interpretation of performance patterns in the good-fit and poor-fit case studies, where systematic deviations-such as the overshooting observed in stable economies during 2023-are more attributable to the influence of the 16-year input window and historical event weighting than to instability or underfitting during training.

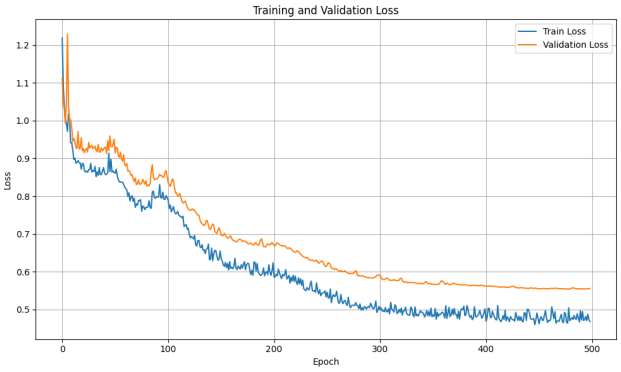
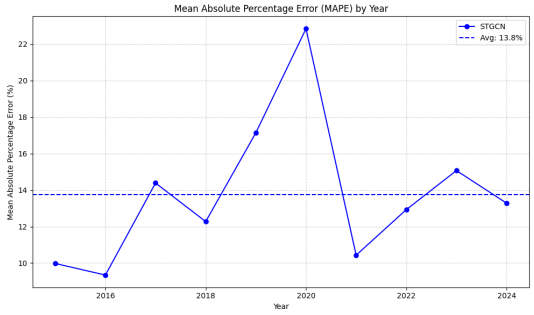


Figure 39. Training and validation loss curves for the STGCN model over 500 epochs. The close alignment after epoch 300 indicates stable convergence and effective generalization.

6.7.2. Analysis and Interpretation

The performance of the Spatio-Temporal Graph Convolutional Network (STGCN) was evaluated using log-

transformed trade values to mitigate the influence of extreme outliers and heteroscedasticity. This transformation produced more stable training dynamics and facilitated a fairer comparison across trade flows of vastly different scales. The results demonstrate that the model is capable of capturing both temporal trends and structural dependencies between countries, yielding strong predictive accuracy for many trade relationships. However, model behavior diverges significantly between volatile trade pairs and those involving highly stable economies, revealing systematic tendencies in the prediction mechanism.



Year	MAPE (%)
2015	9.98
2016	9.34
2017	14.39
2018	12.27
2019	17.14
2020	22.84
2021	10.43
2022	12.93
2023	15.07
2024	13.28

Figure 40. Mean Absolute Percentage Error (MAPE) by year for STGCN predictions on log-transformed trade values. The solid line represents the yearly MAPE, while the dashed line indicates the average MAPE (13.8%) across the evaluation period.

Evaluation Metrics. The overall predictive performance of the STGCN on **country-pair level predictions** (log-transformed trade values) is summarised in Table 11. These values show that the model achieves strong accuracy, explaining nearly 90% of the variance in the target variable.

Metric	Value
Root Mean Squared Error (RMSE)	1.4773
Mean Absolute Error (MAE)	0.9732
Mean Absolute Percentage Error (MAPE)	13.7223%
Coefficient of Determination (R^2)	0.8982

Table 11. Evaluation metrics for STGCN predictions on country-pair level trade flows (log scale).

Impact of COVID-19 on Predictions. As illustrated in Figures 41 and 42, the model's behavior in recent years is heavily shaped by the COVID-19 shock and the subsequent rebound in trade activity. With a 16-year input window, this sharp disruption dominates the model's temporal memory. For stable economies (e.g., Canada-Guatemala), this results in systematic overshooting in 2023, as the model extrapolates the rebound pattern rather than reverting to pre-shock growth trajectories. This phenomenon explains the elevated MAPE values in the post-COVID period (Figure 40), with the spike in 2020 and subsequent volatility reflecting the difficulty of reconciling unprecedented global disruptions with prior historical stability.

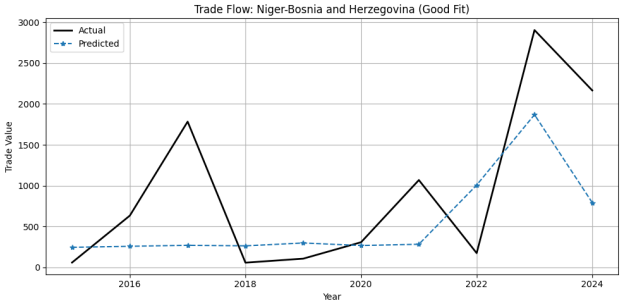


Figure 41. Example of a trade flow (Niger-Bosnia and Herzegovina) with strong model fit in log scale. The model accurately tracks major fluctuations, demonstrating robust performance for volatile trade relationships.

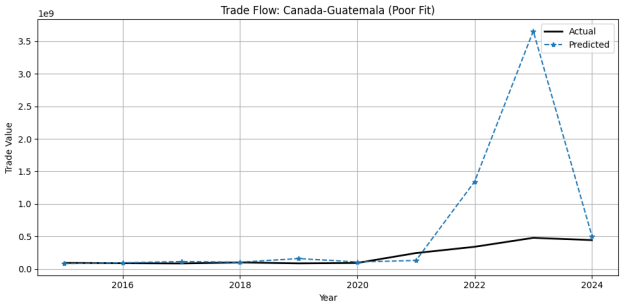


Figure 42. Example of a trade flow (Canada-Guatemala) with poor fit in log scale. With a 16-year input window, the COVID crash and rebound dominate recent history. For stable economies, the model extrapolates the rebound up into 2023, leading to overshooting.

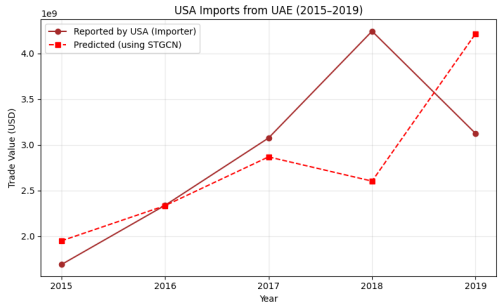


Figure 43. USA imports from UAE (2015-2019): Comparison of reported values (importer-reported) and predicted values using STGCN.

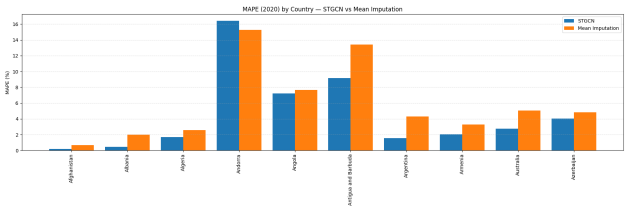


Figure 44. MAPE (2020) by country for STGCN compared with mean imputation. In most cases, STGCN achieves lower MAPE values, indicating superior predictive accuracy over the baseline approach.

Comparison with Mean Imputation When evaluated at the country level, the STGCN model generally surpasses the mean imputation baseline in predictive accuracy, as shown in Figure 44. This advantage is evident for the majority of countries, where the model consistently yields lower MAPE values. These results reinforce the model's ability to capture underlying temporal and spatial trade dynamics more effectively than a static statistical imputation method.

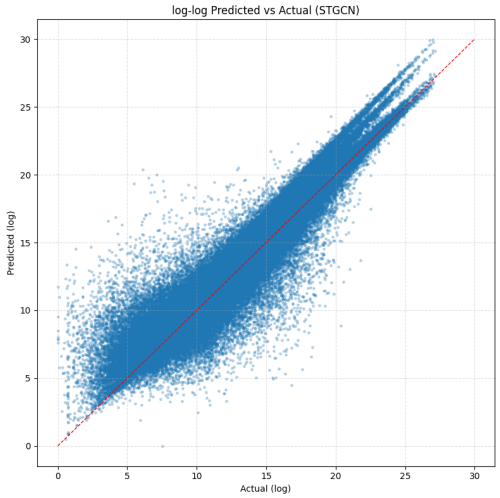


Figure 45. Log-log scatter plot of predicted versus actual trade values for the STGCN model. The red dashed line represents the $y = x$ reference for perfect predictions.

Figure 45 presents a log-log scatter plot comparing predicted and actual trade values for the STGCN model. The majority of points align closely with the $y = x$ reference line, indicating strong agreement between predictions and observations across a wide dynamic range of trade values. The dispersion is greater at the distribution extremes, indicating that while the model captures overall scaling well, predictive certainty declines for rare or high-magnitude trade flows due to variability and limited training examples.

6.7.3. With Product Embeddings

Figure 46 shows the training and validation loss curves for the STGCN model with product embeddings. The model converges steadily over epochs, with the validation loss closely tracking the training loss, indicating good generalization without severe overfitting.

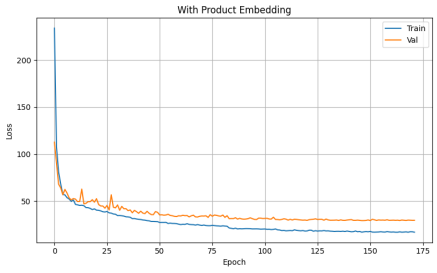


Figure 46. Training and validation loss for the STGCN model with product embeddings.

Table 12 summarises the evaluation metrics for the model on the validation dataset, both on the log-transformed scale and on the original trade value scale. On the log scale, the model

achieves an R^2 of 0.719, whereas on the original scale the R^2 is lower (0.565), reflecting the increased difficulty in matching large absolute values.

Metric	Log Scale	Original Scale
Root Mean Squared Error (RMSE)	2.2114	6.8720×10^8
Mean Absolute Error (MAE)	1.6262	2.8880×10^7
Mean Absolute Percentage Error (MAPE)	56.18%	1.5260×10^5%
Coefficient of Determination (R^2)	0.7190	0.5649

Table 12. Evaluation metrics for STGCN-PE predictions on country–pair–product trade flows.

When compared to a mean–imputation baseline (Table 10, the STGCN-PE model improves R^2 on the validation set by +0.507 and reduces MSE by 50.6%. It also yields a 35.7% reduction in MAE, demonstrating that product embeddings help capture structure beyond simple averages.

Metric	Baseline	Model	Improvement
R^2	-0.00194	0.50484	+0.50678
MSE	1.0876×10^{18}	5.3747×10^{17}	50.58%
MAE	4.5882×10^7	2.9480×10^7	35.75%

Table 13. Comparison of STGCN-PE with mean–imputation baseline (validation set, original scale).

Per–product analysis reveals substantial heterogeneity in gains. The largest MSE reductions are observed for *Furskins and artificial fur; manufactures thereof* (HS2: 43; -68.53% MSE) and *Raw hides and skins (other than furskins) and leather* (HS2: 41; -36.18%), while the model underperforms the baseline in categories such as *Wool, fine or coarse animal hair; horsehair yarn and woven fabric* (HS2: 51), *Other made-up textile articles; sets; worn clothing and worn textile articles; rags* (HS2: 63), and *Cork and articles of cork* (HS2: 45). Notably, the model also performs poorly for the residual/administrative HS2 chapters *Special classification provisions / country-specific special transactions* (HS2: 98) and *Special transactions and commodities not classified according to kind* (HS2: 99), which are heterogeneous catch-all categories and thus less amenable to a product-structured predictive model.

Figure 47 visualises the learned HS2 product embeddings using t-SNE, projecting the high-dimensional embedding space to two dimensions for interpretability. Points are coloured by their corresponding HS2 section, and translucent ellipses indicate the spatial extent of each section in the embedding space. The plot reveals that products within the same HS2 section tend to cluster together, indicating that the model has learned embeddings that capture broad semantic similarities between product categories. Overlaps between sections are expected for goods with shared production processes or overlapping supply chains (e.g., *Foodstuffs* and *Vegetable Products*). This clustering behaviour helps explain why categories with strong seasonality or correlated trade

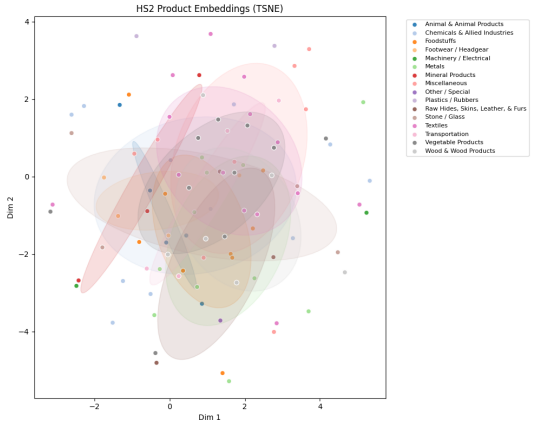


Figure 47. t-SNE projection of HS2 product embeddings, coloured by HS2 section. Ellipses denote the convex coverage of each section.

flows across countries benefit more from the use of product embeddings, as the learned representations allow the model to transfer temporal patterns across related products.

Two broader patterns emerge. First, categories with highly correlated trade flows across countries benefit more from embeddings, because the learned representations can exploit cross-category temporal alignment.

Second, while the approach can in principle be extended to more granular HS4 or even HS6 classifications, this was not pursued here due to computational constraints. The greatly increased number of categories, combined with the requirement to track HS version changes over time (including splits, merges, and redefinitions), caused memory usage to exceed available RAM during embedding training. Such extensions would require either substantial hardware resources or algorithmic adaptations to manage the dynamic category mapping problem efficiently.

Figures 49 and 50 show examples where the model performs well for missing-year imputation. In these cases, the predicted values (dashed blue lines) align closely with the observed trade flows (solid black lines), capturing both the magnitude and temporal dynamics even when entire years are unobserved.

Overall, the STGCN-PE captures complex temporal and cross-sectional dependencies in trade flows, yielding large improvements in many structured product categories, but offering more modest gains or underperformance in irregular, low-structure HS chapters.

6.7.4. Discussion and Limitations

While the adapted STGCN framework demonstrates strong predictive performance for international trade flows, several methodological choices, structural assumptions, and data-

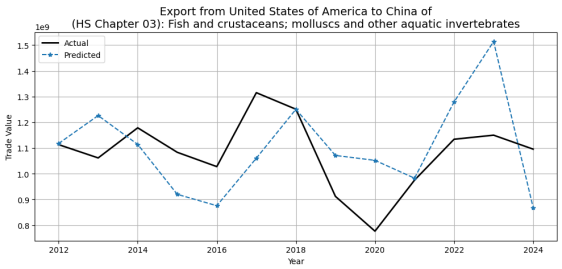


Figure 48. Example of good missing-year imputation: Fish and crustaceans, molluscs and other aquatic invertebrates (HS Chapter 03) exported from the United States of America to China.

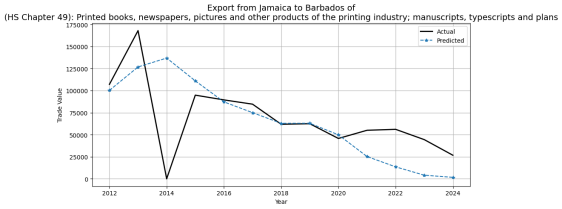


Figure 49. Example of good missing-year imputation: Printed books, newspapers, and related products from Jamaica to Barbados (HS Chapter 49).

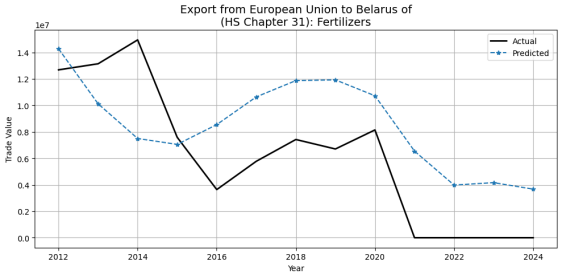


Figure 50. Example of good missing-year imputation: Fertilizers from the EU to Belarus (HS Chapter 31).

related constraints introduce limitations that shape both its accuracy and generalisability:

- **Regression Target Type:** The current formulation frames the problem as *node regression*, where each node's outputs represent aggregated flows to all partners. While this simplifies the modelling process, it limits the model's ability to explicitly learn edge-level heterogeneity in bilateral trade relationships. An *edge regression* approach—directly predicting each origin–destination pair—could capture richer dyadic interactions, albeit with higher computational and memory costs.

- **Temporal Weighting of Historical Events:** The current 16-year input window implicitly treats all years equally, allowing large-scale shocks-such as the 2008 global financial crisis and the 2020 COVID-19 pandemic-to disproportionately influence recent forecasts. This effect is particularly evident in stable economies, where the model tends to extrapolate the post-COVID rebound into later years, leading to systematic overshooting. A more refined approach could incorporate dummy variables or decaying temporal weights to better balance historical influence.

- **Model Assumptions and Architectural Constraints:** The STGCN assumes consistent spatial and temporal dependencies over the training window, which may not fully capture the evolving structure of international trade networks. While effective for stable or moderately volatile relationships, these assumptions can limit accuracy for countries undergoing structural economic shifts or abrupt policy changes.

- **Binary Weight Matrix Limitation:** The current graph representation uses a binary adjacency matrix, which encodes only the presence or absence of trade links. Transitioning to a weighted matrix would allow more nuanced modelling of relationship strength, potentially improving performance by capturing the intensity of trade connections.

- **Data Scaling Challenges:** Due to the wide dispersion of trade values, conventional normalisation techniques proved unsuitable, often distorting results. The logarithmic transformation was ultimately adopted, as it best preserved proportional relationships while mitigating the influence of extreme outliers. Nonetheless, this choice alters the error distribution and affects interpretability of predictions in raw units.

- **Codebase Adaptation and Maintainability:** The original STGCN implementation lacked sufficient documentation and was written in TensorFlow, presenting a barrier to adaptation. The model was re-implemented

in PyTorch with extensive inline documentation and control structures to enhance transparency and ease of extension. While the methodological core remains unchanged, significant modifications were made to tailor the pipeline to the trade prediction context, including a custom dataset segmentation strategy for training, validation, and testing.

- **Computational Resource Constraints:** The available hardware, even when using Google Colab with NVIDIA A100 GPUs, imposed memory limitations that restricted the dataset size, number of countries, and temporal depth that could be processed simultaneously. Larger-scale experiments may require distributed training or model compression techniques.

In summary, the proposed framework successfully leverages graph neural networks for trade forecasting, offering a notable improvement over traditional approaches such as gravity models and statistical interpolation. However, its performance remains influenced by temporal weighting, architectural assumptions, and data limitations. Future work could focus on incorporating time-decay weighting, non-binary relationship encoding, and tariff-adjusted trade values, alongside scaling up the system to larger and more diverse datasets.

7. Discussion

7.1. Comparative Assessment of Methods

The comparative evaluation of imputation methods for missing trade values reveals clear performance differentials across aggregation levels, reflecting both methodological capabilities and the structural characteristics of the dataset.

At the *country-pair level*, the classical gravity model exhibited substantially poorer predictive accuracy than all machine learning approaches, with large error magnitudes and strongly negative R^2 values (Table 14). This outcome is consistent with the limitations identified in the clustering-gravity estimation pipeline, particularly the reliance on a single representative pair per cluster-period and the assumption of homogeneous trade determinants within each cluster. In contrast, ensemble tree-based methods, notably gradient boosting and XGBoost, delivered markedly improved accuracy by capturing nonlinear interactions and complex cross-feature dependencies that the linear gravity model was unable to represent. The spatio-temporal graph convolutional network (STGCN) achieved competitive results, with the log-transformed variant (STGCN-log) yielding the highest R^2 among all approaches, indicating the benefit of variance stabilization for this class of models.

At the *product level*, the relative ranking of methods shifts (Table 15). The temporal imputation baseline performed reasonably for stable time series but was less effective in the presence of structural breaks or sharp trade shocks. Random forests maintained competitive accuracy by exploiting cross-sectional feature variation, though their R^2 values fell below those achieved at the country-pair level. The STGCN with product embedding (STGCN-PE) attained the highest explanatory power in several categories, demonstrating the value of modelling temporal dependencies and relational structure in settings where network effects and seasonality are more pronounced. The Silva Random Forest Network method reported here is evaluated in terms of growth rate prediction, which is not directly comparable in magnitude to USD-based error metrics, but nonetheless provides insight into its relative strength in capturing directional trends.

Overall, the results highlight the sensitivity of method performance to the aggregation level and the underlying data structure. At the country-pair level, ensemble tree-based methods and STGCN variants clearly outperform the classical gravity model, with variance-stabilizing transformations offering measurable gains for neural architectures. At the product level, temporal-relational modelling with STGCN-PE delivers clear advantages in contexts where seasonal, network, or product-specific dynamics are central to trade behaviour.

7.2. Recommendations

The comparative assessment indicates that model selection for trade value imputation should be explicitly aligned with the aggregation level, data characteristics, and the nature of missingness.

- **Country-pair level:** Ensemble tree-based methods, particularly gradient boosting (XGBoost), offer the most favourable trade-off between accuracy, robustness, and computational cost. The log-transformed STGCN variant achieved the highest R^2 in this setting, suggesting that variance-stabilising transformations can materially improve neural architectures. However, the higher complexity and tuning requirements of STGCN may limit its operational suitability outside research contexts.
- **Product level:** In categories where seasonality, network dependencies, and product-specific patterns dominate, temporal-relational architectures such as STGCN with product embedding (STGCN-PE) deliver clear performance gains. Their ability to exploit both temporal continuity and inter-product linkages makes them particularly well-suited for high-volatility or supply-chain-sensitive goods.
- **Stable, low-variability series:** For products with consistent trade patterns and minimal exposure to structural shocks, simple temporal imputation remains a

Method	R^2	MSE (USD)	MAE (USD)
Gravity Model	-1.18×10^{22}	2.69×10^{39}	5.28×10^{20}
Random Forest	0.603	4.48×10^{19}	5.40×10^{9}
Gradient Boosted Trees	0.570	1.11×10^{20}	5.33×10^{9}
XGBoost	0.570	1.10×10^{20}	5.55×10^{9}
STGCN	0.420	1.69×10^{21}	9.20×10^{8}
STGCN (log scale)	0.720	1.63 (log-units)	4.89 (log-units)

Table 14. Model comparison for missing value estimation at the country-pair level. All models were evaluated on the same held-out test set comprising 20% of the original data (random split, ensuring no data leakage). For STGCN (log scale), errors are reported in transformed units.

Method	R^2	MSE (USD / rate)	MAE (USD / rate)
Temporal Imputation	–	1.8×10^{10}	5.81×10^{6}
Random Forest	-0.075	1.04×10^{16}	5.70×10^{6}
Silva RF Network	0.038	0.85 (growth)	0.76 (growth)
STGCN-PE	0.510	5.37×10^{17}	2.95×10^{7}

Table 15. Model comparison for product-level estimation of missing values. Each method was evaluated on a different randomly selected 20% of the original data, with strict prevention of data leakage. Metrics for the Silva RF Network are reported for growth rate prediction rather than trade value in USD.

defensible choice, particularly where rapid execution and minimal computational overhead are priorities.

- **Methods to avoid for direct prediction:** The classical gravity model, as implemented here, is not recommended for direct imputation of missing trade values under sparse and heterogeneous data conditions, given its poor performance across all metrics.

7.3. Overall Study Limitations

Across all methodological approaches, several overarching constraints shaped both the scope and performance of the analysis:

- **Underlying Data Quality and Coverage:** All models are fundamentally limited by the completeness, consistency, and accuracy of reported trade statistics. Hstorical geopolitical changes introduce gaps that no modelling strategy can fully overcome. Systematic valuation differences between imports and exports (e.g., CIF vs. FOB, tariffs, freight costs) remain unadjusted, embedding a common bias across all predictions.

- **Aggregation Trade-offs:** The strongest performance was consistently observed at aggregated country–year or HS chapter levels, where data density is highest. At finer product levels, data sparsity, volatility, and idiosyncratic shocks degraded performance for all methods. Chapter-level aggregation improved tractability but masked important intra-chapter variation relevant for commodity-specific analysis.

- **Structural Assumptions on Persistence and Smoothness of Trade:** Several methods assumed that once a trade relationship was observed, it persisted in subsequent years, and that trade flows evolved gradually over time. These assumptions facilitated panel completion and temporal modelling but are violated in the presence of abrupt political, economic, or environmental shocks, potentially leading to misclassification of genuine changes as missing data.

- **Directional and Network Effects:** Certain approaches either disregarded or simplified exporter–importer asymmetries and broader network interdependencies. This limits the ability to capture cases where trade directionality or indirect relationships via third countries meaningfully influence observed flows.

- **Feature Limitations:** While macroeconomic size, geographic proximity, and network features were incorporated, the analysis omitted several key policy and structural variables, such as detailed tariff schedules, preferential trade agreements, non-tariff barriers, and cultural or historical ties, due to data availability and integration complexity. This restricts explanatory scope, particularly in cases where economic similarity does not translate to actual trade volumes.

- **Scalability and Computational Constraints:** Practical limitations on computational resources, model complexity, and available training data restricted the scale of experiments. These constraints influenced choices in clustering granularity, time-window size, and sample selection, and limited the feasibility of fully disaggregated modelling.

These shared constraints underscore the fact that, while individual methods differ in architecture and estimation logic, their predictive potential is ultimately bounded by the quality, granularity, and representativeness of the available trade data, as well as by the simplifying assumptions necessary for tractable modelling.

8. Conclusion

This study set out to address the persistent challenge of incomplete bilateral trade flow data by systematically evaluating multiple methodological approaches, ranging from traditional econometric models to advanced machine learning architectures. Across aggregate and product-level prediction tasks, the findings underscore that no single method is universally optimal; rather, performance depends on the granularity of estimation, the temporal and structural complexity of the data, and the intended application of the results.

At the aggregate country-pair level, ensemble tree-based methods, particularly gradient boosting and XGBoost, offered the most favourable balance of predictive accuracy, robustness, and computational efficiency. These approaches consistently outperformed the gravity-based benchmark, which, as implemented here, exhibited poor performance under sparse and heterogeneous data conditions and is not recommended for direct imputation without substantial modification. The log-transformed variant of the Spatio-Temporal Graph Convolutional Network (STGCN-log) achieved the highest R^2 in this setting, illustrating that variance-stabilising transformations can materially enhance neural architectures when modelling highly skewed trade value distributions.

At the product level, the Spatio-Temporal Graph Convolutional Network with product embeddings (STGCN-PE) delivered the strongest performance, particularly for product categories characterised by strong seasonal patterns, correlated trade flows between partners, and well-defined positions within supply-chain networks. The ability of STGCN-PE to integrate temporal dependencies with network structure represents a significant methodological advantage over purely temporal or purely relational models. Nonetheless, performance gains were uneven across categories, with diffuse or highly heterogeneous HS chapters showing limited improvement over simpler baselines. For series with low variability or limited historical depth, temporal imputation proved adequate, offering computational simplicity and interpretability with minimal loss in accuracy.

The broader implication of this work is that methodological choice should be guided not only by predictive performance but also by operational considerations: data availability, classification stability, and the scalability of model training. The integration of HS version harmonisation, product category aggregation, and reference datasets further highlights the importance of preprocessing pipelines in ensuring comparability over time. Without such infrastructure, even the most sophisticated model risks producing inconsistent or biased estimates.

Looking ahead, future research could focus on three main directions. First, expanding the feature set to incorporate richer contextual and macroeconomic variables may further improve model generalisability. Second, developing robust strategies for handling HS classification changes at the 4- and 6-digit levels would enable more granular forecasting without prohibitive memory costs. Finally, establishing unified evaluation frameworks, including standardised metrics, benchmarks, and test datasets, would improve comparability across studies and accelerate methodological progress.

By combining economic insight with modern statistical learning, this research demonstrates that it is possible to generate more complete, timely, and reliable trade statistics.

Final Remarks

This study has offered valuable insights into the application of time-series forecasting methods for the estimation of missing international trade flows. The work has highlighted the importance of rigorous model evaluation, adaptability in methodological design, and the consideration of multiple modelling approaches in order to obtain a comprehensive understanding of results.

One key limitation was the availability of computational resources, particularly memory capacity, which constrained the size of the dataset in terms of the number of countries and time periods that could be processed concurrently. This restriction inevitably limited the representativeness of the results, as unmodelled trade relationships may have exerted influence on the observed trends. Access to more advanced computing infrastructure would allow for a broader scope of analysis and facilitate more detailed examination of global trade patterns.

A further limitation related to the absence of explicit contextual economic and geopolitical factors. Variables such as trade agreements, and policy changes were not incorporated into the model, potentially reducing its ability to capture sudden structural shifts or anomalies in trade flows.

The most significant methodological constraint concerned data scaling. The wide dispersion of trade values rendered conventional scaling techniques unsuitable, as they often distorted proportional relationships and impaired predictive accuracy. The adoption of a logarithmic transformation provided a more robust alternative, preserving relative magnitudes while mitigating the effect of extreme values. Nonetheless, this transformation alters the error distribution and affects the interpretability of outputs in absolute units.

Future work should consider alternative or hybrid scaling strategies tailored to datasets of this nature.

In conclusion, while certain limitations have shaped the scope and outcomes of this work, they have also informed methodological refinements and identified avenues for future research. These include the integration of time-decay weighting, non-binary and weighted network representations, tariff-adjusted trade values, and edge-level regression to better capture bilateral heterogeneity. Addressing these areas, alongside enhanced computational capacity, would enable more comprehensive and representative modelling of global trade networks.

References

Athey, S. and Imbens, G. W. Machine learning methods that economists should know about. *Annual Review of Economics*, 11(Volume 11, 2019): 685–725, 2019. ISSN 1941-1391. doi: https://doi.org/10.1146/annurev-economics-080217-053433. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-080217-053433.

Bhagwati, J. N. On the underinvoicing of imports. *Bulletin of the Oxford University Institute of Economics & Statistics*, 27(4):389–397, 1964.

Bobková, B. On estimation of gravity equation: A cluster analysis. Working paper, Charles University, Faculty of Social Sciences, 2014.

Braml, M. T. and Felbermayr, G. J. The problem with trade measurement in international relations. *International Studies Quarterly*, 67(2):sqad020, 2023.

Breiman, L. *Random Forests*, volume 45. Springer, 2001.

Developers, G. geopy: Geocoding library for python. https://geopy.readthedocs.io/, 2025. Version used in this study; Accessed 2025-08-11.

Felbermayr, G., Kirilakha, A., Syropoulos, C., Yalcin, E., and Yotov, Y. The global sanctions data base (gsdb): An update that includes the years of the trump presidency. *European Economic Review*, 129:103561, 2020. doi: 10.1016/j.euroecorev.2020.103561.

Gaulier, G. and Zignago, S. Baci: International trade database at the product-level. the 1994-2007 version. Working Paper 2010-23, CEPII, 2010.

Gopinath, M. Machine learning in gravity models: An application to trade flow prediction. *Unpublished Working Paper*, 2020.

Hamanaka, S. Whose trade statistics are correct? multiple mirror comparison of asian trade data. *Journal of Japanese and International Economies*, 26(4):581–592, 2012.

Hamilton, W. L., Ying, R., and Leskovec, J. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 40(3):52–74, 2017.

Head, K. and Mayer, T. Gravity equations: Workhorse, toolkit, and cookbook. In Gopinath, G., Helpman, E., and Rogoff, K. (eds.), *Handbook of International Economics*, volume 4, pp. 131–195. Elsevier, 2014.

Kipf, T. and Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Linsi, L. and Mügge, D. Trade statistics and the political economy of data gaps. *International Studies Quarterly*, 67(2):sqad020, 2023. doi: 10.1093/isq/sqad020. URL https://academic.oup.com/isq/article/67/2/sqad020/7085502.

Minakawa, N., Izumi, K., and Sakaji, H. Bilateral trade flow prediction by gravity-informed graph auto-encoder, 2024. URL https://arxiv.org/abs/2408.01938.

Monken, A., Haberkorn, F., Gopinath, M., Freeman, L., and Batarseh, F. A. Graph neural networks for modeling causality in international trade. In *Proceedings of the 34th International Florida Artificial Intelligence Research Society Conference (FLAIRS-34)*. AAAI Press, 2021. URL https://www.researchgate.net/publication/351570759_Graph_Neural_Networks_for_Modeling_Causality_in_International_Trade.

OpenStreetMap contributors. Nominatim: Openstreetmap geocoding. https://nominatim.org/, 2025. Accessed 2025-08-11.

Panford-Quainoo, K., Bose, A., and Defferrard, M. Bilateral trade modelling with graph neural networks. In *ICLR Workshop on Practical Machine Learning for Developing Countries*, 2020. URL https://www.mdpi.com/2504-2289/8/6/65.

Pöyhönen, P. A tentative model for the volume of trade between countries. *Weltwirtschaftliches Archiv*, 90:93–100, 1963. ISSN 0043-2636.

Rincon-Yanez, D., Ounoughi, C., Sellami, B., Kalvet, T., Tiits, M., Senatore, S., and Ben Yahia, S. Accurate prediction of international trade flows: Leveraging knowledge graphs and their embeddings. *Journal of King Saud University - Computer and Information Sciences*, 35:101789, 2023a. doi: 10.1016/j.jksuci.2023.101789. URL https://arxiv.org/abs/2310.11161.

Rincon-Yanez, D., Ounoughi, C., Sellami, B., Kalvet, T., Tiits, M., Senatore, S., and Yahia, S. B. Accurate prediction of international trade flows: Leveraging knowledge graphs and their embeddings, 2023b. URL https://arxiv.org/abs/2310.11161.

Santos Silva, J. M. C. and Tenreyro, S. The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658, November 2006.

Sellami, B., Ounoughi, C., Kalvet, T., Tiits, M., and Rincon-Yanez, D. Harnessing graph neural networks to predict international trade flows. *Big Data and Cognitive Computing*, 8(6), 2024. ISSN 2504-2289. doi: 10.3390/bdcc8060065. URL https://www.mdpi.com/2504-2289/8/6/65.

Silva, T. C., Wilhelm, P. V. B., and Amancio, D. R. Machine learning and economic forecasting: The role of international trade networks. *Physica A: Statistical Mechanics and its Applications*, 649:129977, 2024. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2024.129977. URL https://www.sciencedirect.com/science/article/pii/S0378437124004862.

Stekhoven, D. J. and Bühlmann, P. Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008.

Teti, F. Missing Tariffs. CESifo Working Paper Series 11590, CESifo, 2024. URL https://ideas.repec.org/p/ces/ceswps/_11590.html.

Tiits, M., Kalvet, T., Ounoughi, C., and Ben Yahia, S. Relatedness and product complexity meet gravity models of international trade. *Journal of Open Innovation: Technology, Market, and Complexity*, 10(2):100288, 2024. ISSN 2199-8531. doi: https://doi.org/10.1016/j.joitmc.2024.100288. URL https://www.sciencedirect.com/science/article/pii/S2199853124000829.

Tinbergen, J. *Shaping the world economy : suggestions for an international economic policy*. Twentieth Century Fund, New York, 1962.

World Bank. Understanding trade data quality with the discrepancy index. World Bank Blogs, 2024. URL https://blogs.worldbank.org/en/opendata/understanding-trade-data-quality-with-the-discrep

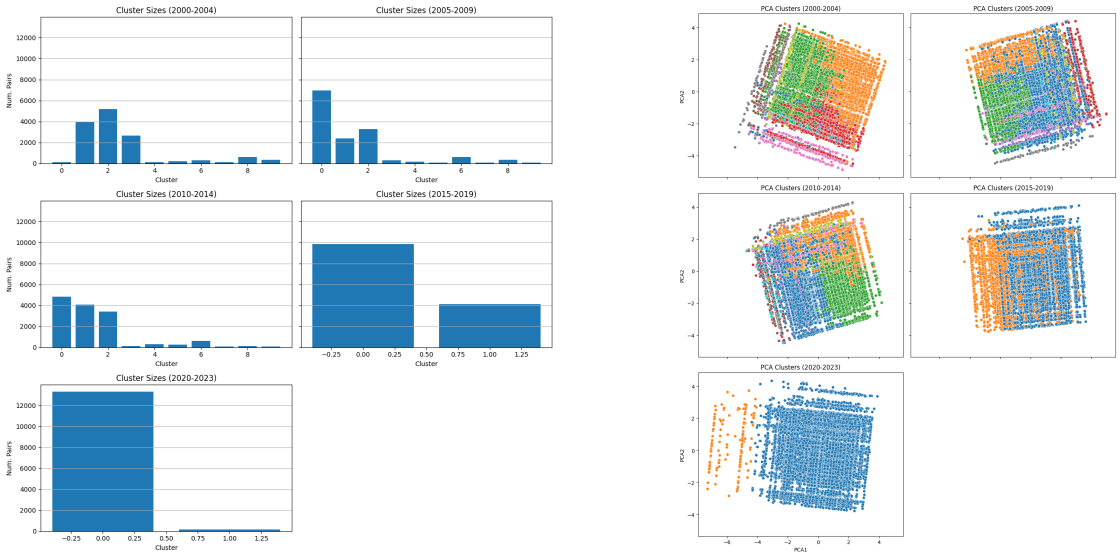
Yeats, A. J. Are partner-country statistics useful for estimating missing trade data? *World Bank Policy Research Working Paper*, (1501), 1995.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications, 2021. URL https://arxiv.org/abs/1812.08434.

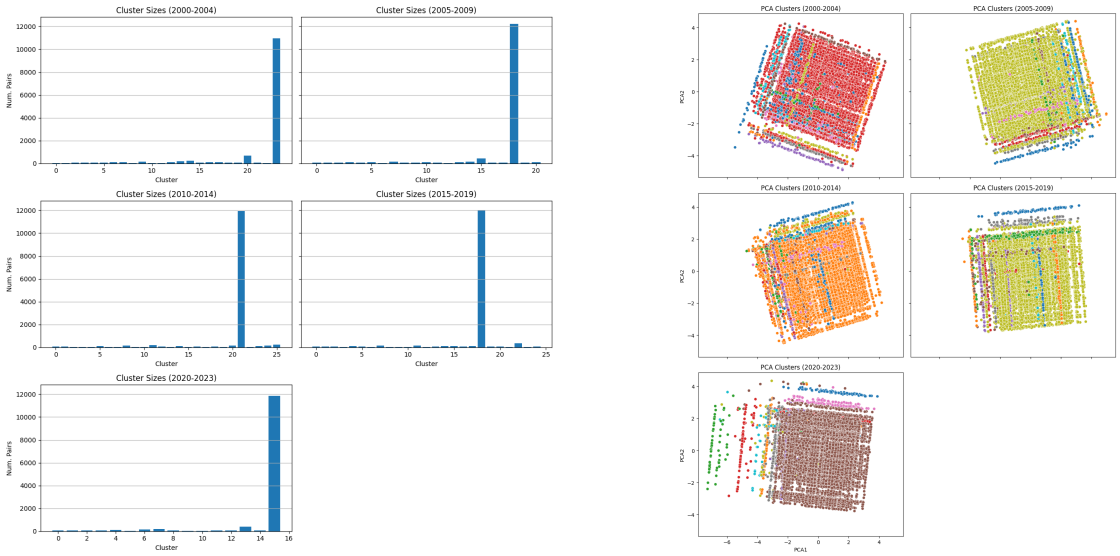
A. Clustering in Gravity Model

The cluster distribution and PCA plots for hierarchical clustering and HDBSCAN clustering.



((a)) Cluster size distributions ((b)) PCA scatter plots

Figure 51. Hierarchical clustering evaluation across 5-year periods.



((a)) Cluster size histograms ((b)) PCA projections

Figure 52. HDBSCAN clustering evaluation across 5-year periods.