

CNGF 5020: Mini Group Project II

YAN Yilin YIN Xinyue ZHANG Siyu ZHAO Xinbin HE Xiaolu

1. Introduction & Data Overview

1.1 Project Scope

The four core issues addressed by this project:

- Descriptive statistics of fundamental characteristics in trade data;
- Identifying national trade patterns through unsupervised learning;
- Predicting bilateral trade flows using machine learning, with a counterfactual analysis of doubled distance between China and the US;
- Analyzing the impact of distance on commodity sectors.

1.2 Core Data Source

- Trade data: 2016–2018 BACI HS12 data (HS6-level bilateral export value v , quantity q , country code i/j , product code k);
- Supplementary data: World Bank GDP data (exporting country/importing country GDP), country shapefiles (centroid distance calculation), country/product code mapping table (matching names and codes).

2. Descriptive Statistics and Trade Patterns

2.1 Top and Bottom 10 Countries by Number of Trading Partners

2.1.1 Top 10 Countries with Most Trading Partners

The countries with the most trading partners (i.e., the largest number of countries involved in trade flows with them) over the three-year period are as follows:

Nation	Number
Germany	220 partners
Italy	220 partners
Thailand	220 partners
France, Monaco	220 partners
Spain	220 partners
Netherlands	220 partners
Poland	220 partners
Brazil	219 partners
Belgium-Luxembourg	219 partners
United Kingdom	219 partners

These countries are characterized by strong global trade networks, reflecting their significant positions in international trade.

2.1.2 Bottom 10 Countries with Fewest Trading Partners

The following countries have the fewest trading partners, which is indicative of their smaller or more specialized trade networks:

Nation	Number
Netherlands Antilles	1partner
Bonaire, Saint Eustatius and Saba	36partners
Saint Pierre and Miquelon	43partners
Saint Barthélemy	46partners
Saint Maarten (Dutchpart)	50partners
Christmas Islands	50partners
Norfolk Islands	51partners
French South Antarctic Territories	52partners
Wallisand Futuna Islands	54partners
Federated States of Micronesia	56partners

These countries have a more limited role in global trade, with few trading relationships, which could be due to geographic isolation, smaller economies, or less diversified exports.

2.2 Top 10 Trading Partners of China and United States

For China, the top 10 trading partners based on total trade value (in USD) are:

Rank	Nation	Trade Value
1	United States	\$1.802 billion
2	Japan	\$870.6 million
3	Hong Kong	\$820.1 million
4	South Korea	\$729.4 million
5	Germany	\$605.5 million
6	Other Asia	\$417.4 million
7	Australia	\$398.9 million
8	Vietnam	\$303.2 million
9	Russia	\$262.8 million
10	United Kingdom	\$246.1 million

These results highlight China's key trade relationships, particularly with the U.S., Japan, and Hong Kong, which have been crucial to its export economy.

Similarly, the U.S. has the following top 10 trading partners based on total trade value:

Rank	Nation	Trade Value
1	China	\$1.802 billion
2	Mexico	\$1.530 billion
3	Canada	\$1.369 billion
4	Japan	\$597.7 million
5	Germany	\$543.3 million
6	South Korea	\$362.6 million
7	United Kingdom	\$321.0 million
8	France, Monaco	\$222.1 million
9	India	\$219.8 million
10	Other Asia	\$208.9 million

The U.S. shares strong trade ties with its North American neighbors, as well as China, which remains its largest trading partner.

2.3 High-Value Trade Flows in China

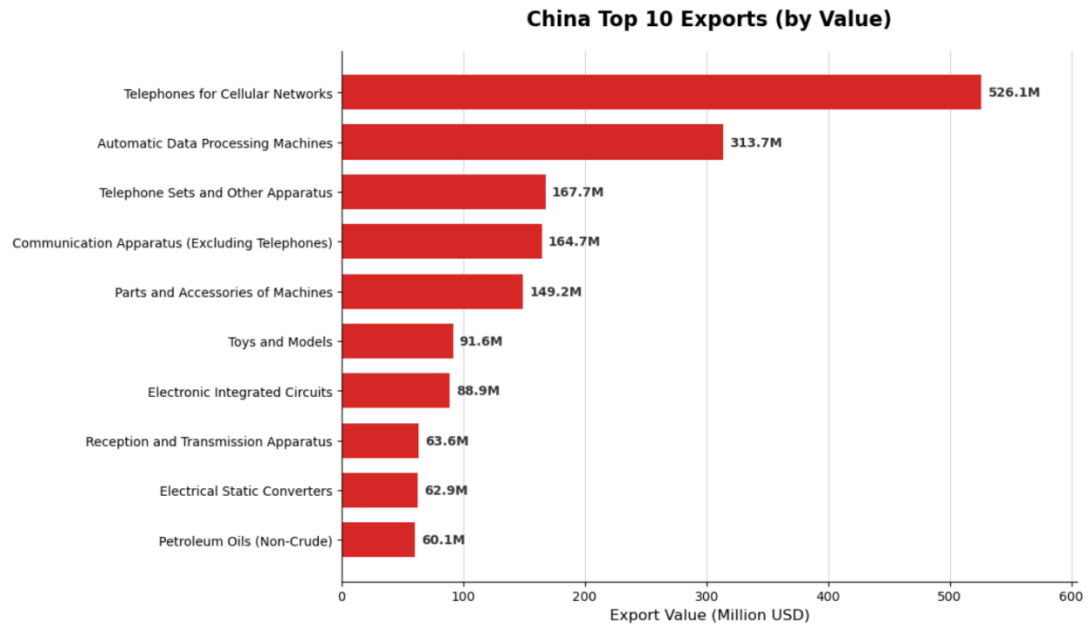
The five highest-value trade flows involving China during the sample period (2016-2018) are as follows:

Year	Exporter	Importer	Trade Value (in Million USD)
2018	China	USA	528.16
2017	China	USA	476.25
2016	China	USA	424.59
2018	China	Hong Kong	276.92
2017	China	Hong Kong	258.32

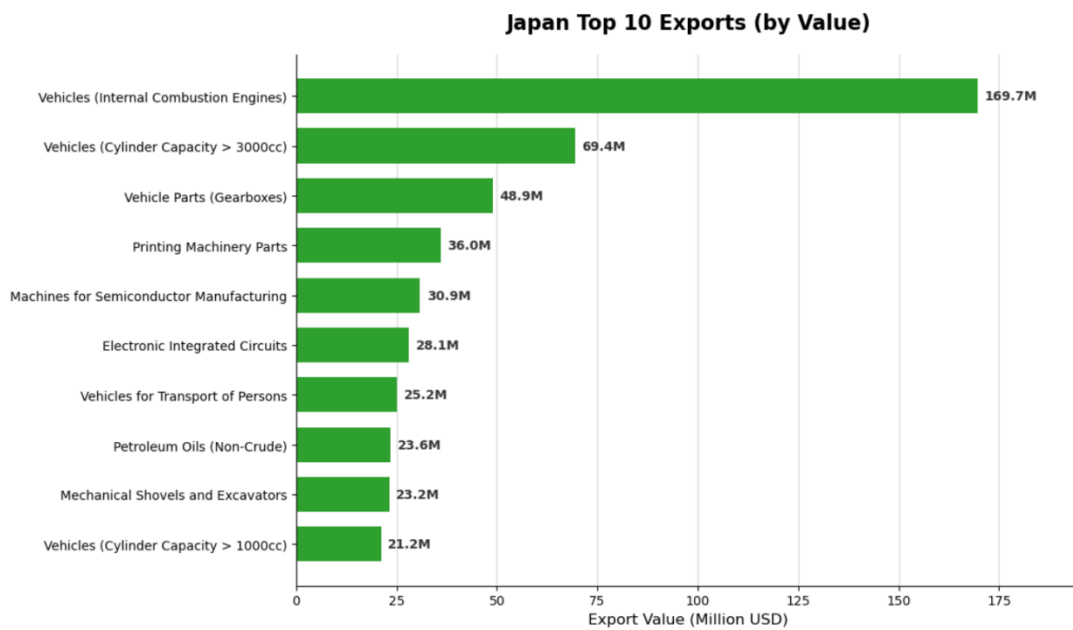
These figures illustrate the dominance of the U.S. and Hong Kong as top destinations for Chinese exports.

2.4 Top 10 Export Products by China, Japan, and the USA

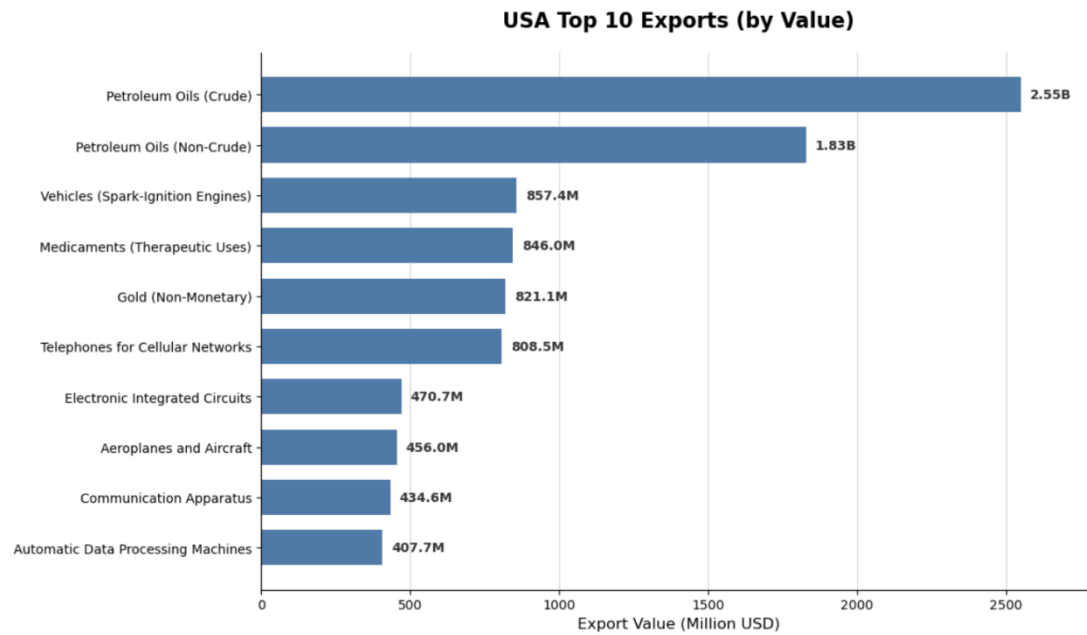
2.4.1 China's Top 10 Exports:



2.4.2 Japan's Top 10 Exports:

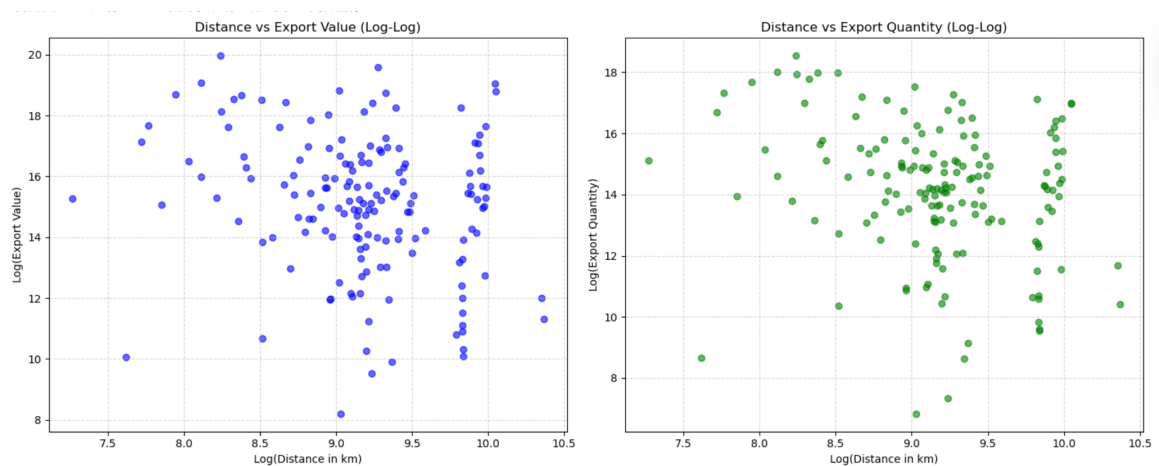


2.4.3 USA's Top 10 Exports:



2.5 Trade Distance and Export Volume

The analysis of the distance between China and other countries in relation to their export values and volumes yielded the following key finding:



- Correlation between Distance and Export Volume (Value): -0.2170;
- Correlation between Distance and Export Volume (Quantity): -0.2388;

These results suggest that there is a weak negative relationship between distance and export volume, indicating that geographic proximity has a limited effect on export trade, though distance may still be a significant factor in some cases.

3. Unsupervised Learning: Identifying Trade Patterns

3.1 Data Preprocessing: Standardization & HS2 Aggregation

Step1 - Data Standardization: Adopt min-max scaling/standardization to eliminate dimensional differences of sector-level indicators across countries and ensure comparability.

Step2 - HS2 Aggregation: Pad HS6 codes to 6 digits, take the first two digits, and aggregate 2016-2018 product-level export data into broad HS2 product categories.

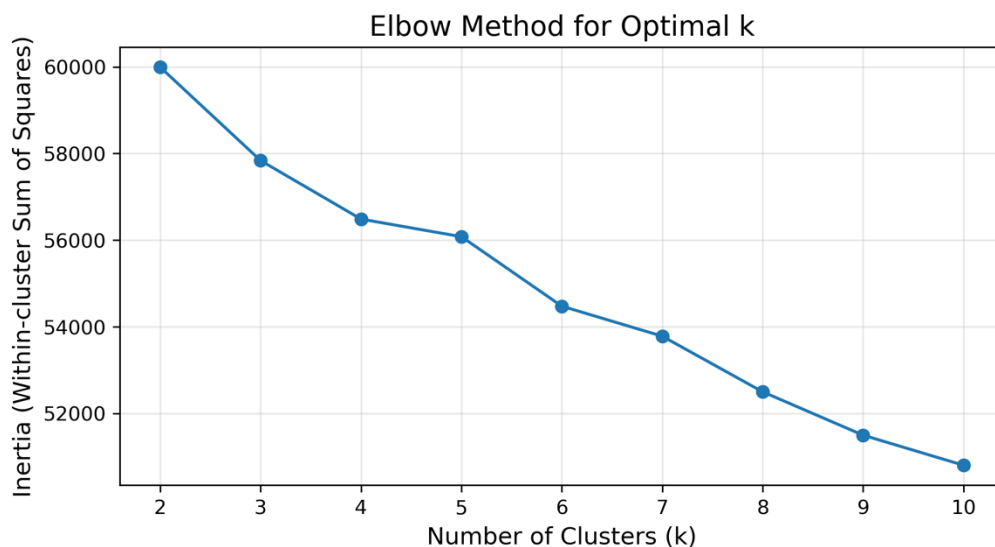
Step3 - Vector Construction: Calculate the export share of each HS2 category for each country, and generate an HS2 export-share vector for each country.

3.2 Classification of Countries Based on Export Commodity Share

Step1 - Clustering Method: Select K-means clustering (alternative: Hierarchical Clustering), using HS2 export-share vectors as the core features.

Step2 - Clustering Basis: Perform unsupervised grouping of countries based on the similarity of their export structures.

Step3 -Determination of Optimal k: Analyze clustering errors via the Elbow Method to select the optimal number of clusters (k).

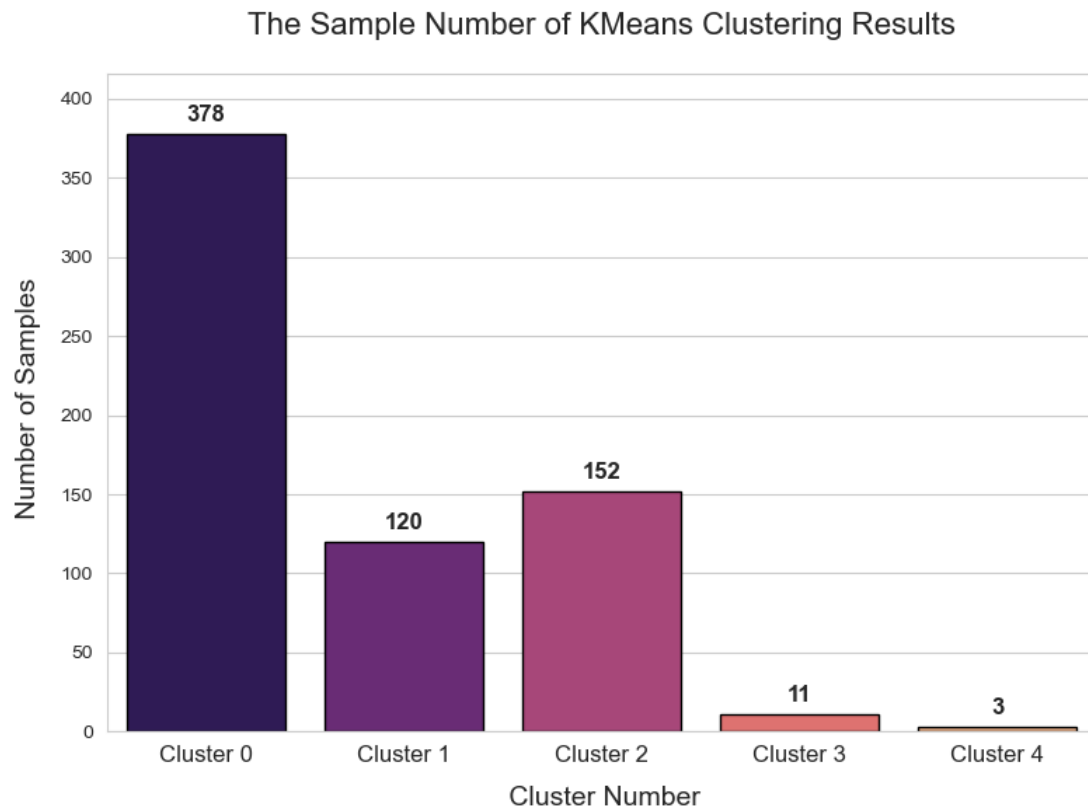


3.3 Key Observations Per Cluster and Visualization

3.3.1 Key Observations Per Cluster

- **Cluster 0:** The largest cluster (378 samples) specializes in HS2 07 (edible vegetables), 05 (other animal products), and 53 (other plant fibers). It has the shortest export distance from China and the lowest HHI, indicating diversified exports.
- **Cluster 1:** Medium-sized (152 samples) with core products in HS2 14 (vegetable plaiting materials), 65 (hats), and 46 (basketwork). Moderate distance and HHI reflect a balance between specialization and diversification.
- **Cluster 2:** Medium-sized (120 samples) focusing on HS2 27 (mineral fuels), 71 (jewelry), and 26 (ores). Similar distance to Cluster 1 but slightly higher HHI, suggesting mild specialization in resource-based products.
- **Cluster 3:** Small cluster (11 samples) specializing in HS2 61 (knitwear), 84 (nuclear reactors/machinery), and 59 (impregnated fabrics). Long export distance and high HHI indicate concentrated exports of manufactured goods.
- **Cluster 4:** Smallest cluster (3 samples) with core products in HS2 39 (plastics), 42 (articles of leather), and 96 (miscellaneous manufactured articles). The longest distance and highest HHI reflect highly specialized exports to distant markets.

3.3.2 K-Means Cluster Sample Distribution

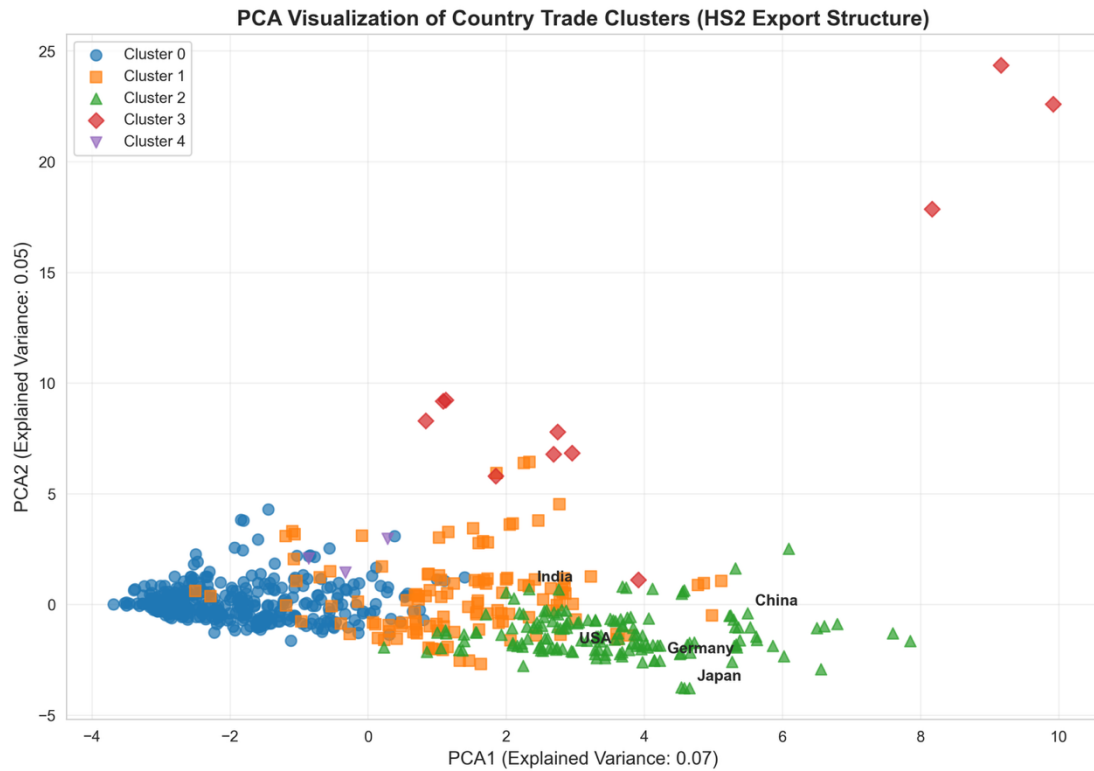


The bar chart of K-Means clustering results reveals an uneven sample distribution:

- Cluster 0 dominates with 378 samples (73% of total), indicating a prevalent trade pattern centered on diversified, short-distance agricultural and plant-based exports.
- Clusters 3 and 4 have only 11 and 3 samples, respectively, representing niche trade patterns with specialized, long-distance exports.

Implication: Most countries share similar export structures (diversified, low-distance), while a small subset adopts specialized strategies for distant markets.

3.3.3 PCA Visualization of Clusters



The PCA plot (explained variance: PCA1=7%, cumulative ~12%—derived from code) shows partial separation of clusters:

- The spatial distribution characteristics of each cluster:
- **Cluster 0:** Concentrated in the left area of the graph (PCA1 \approx -4 to 0), it is the most densely populated cluster (corresponding to the largest cluster analyzed previously, with 378 samples). This cluster represents "diversified agricultural / primary product exports" countries, with a high degree of spatial concentration, indicating that the export structure of these countries is highly similar.
- **Cluster 1:** Distributed to the right of Cluster 0 (PCA1 \approx 0 to 2), partially overlapping with Cluster 0, with an average sample size. Corresponding to "moderately diversified, light industrial product exports" countries, the export structure has some similarity to Cluster 0, but has already shown differentiation.
- **Cluster 2:** Concentrated in the middle-right area of the graph (PCA1 \approx 2 to 6), including the core economies marked (USA, Germany, Japan), and China is also close to this area. Corresponding to "resource / industrial product exports" countries, the export structure is mainly composed of medium-high value-added

products, and is a gathering area of major global trading countries.

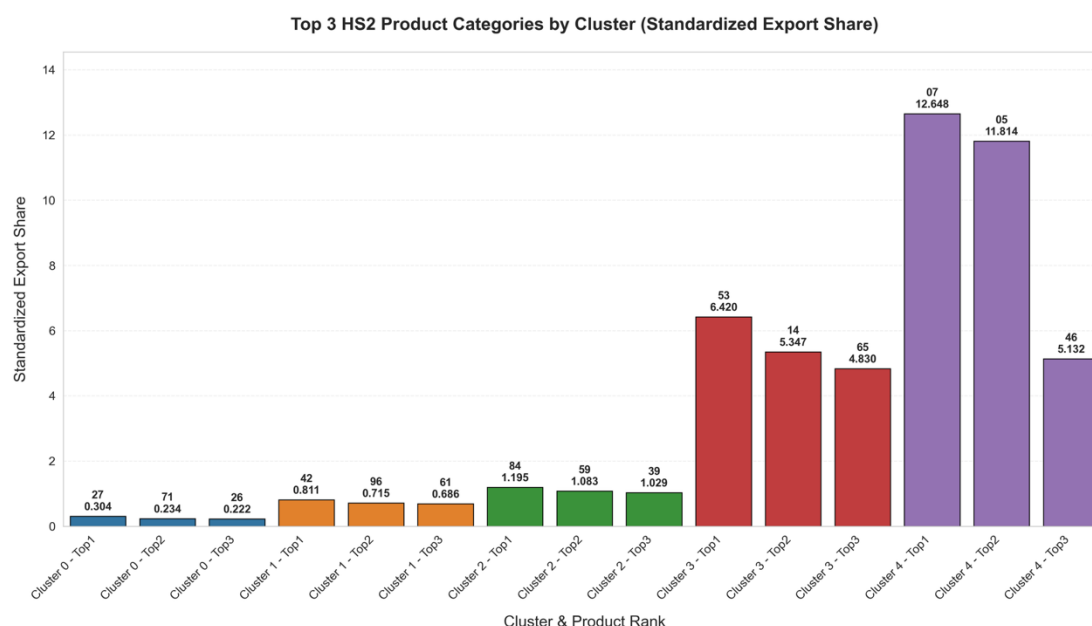
- **Cluster 3:** Significantly isolated from other clusters ($PCA1 \approx 2$ to 10 , $PCA2 \approx 5$ to 25), without overlap with other clusters. Corresponding to "specialized manufacturing exports" countries, the export structure is highly unique, and is very different from most countries (with a small sample size, only 11 samples).
- **Cluster 4:** Occurs only in small quantities in the graph (corresponding to 3 samples analyzed previously), belonging to a very niche "highly specialized export" cluster, with relatively isolated spatial distribution.

3.3.4 The meaning of the positions of core countries

- **China:** Located at the junction of Cluster 0 and Cluster 2 — Consistent with China's trade characteristics: It is both an agricultural / primary product exporter and an industrial / high-value-added product exporter, with an export structure that combines "diversification" and "industrialization".
- **USA/Germany/Japan:** Located in the core area of Cluster 2 — Representing the export structure of these countries as mainly industrial / high-value-added products, being a typical "industrialized export model" in global trade.
- **India:** Located between Cluster 1 and Cluster 2 — Reflecting that India's export structure is in the transitional stage from "light industry" to "industrialization".

3.4 Cluster Characteristics Analysis

3.4.1 Top 3 HS2 Products per Cluster

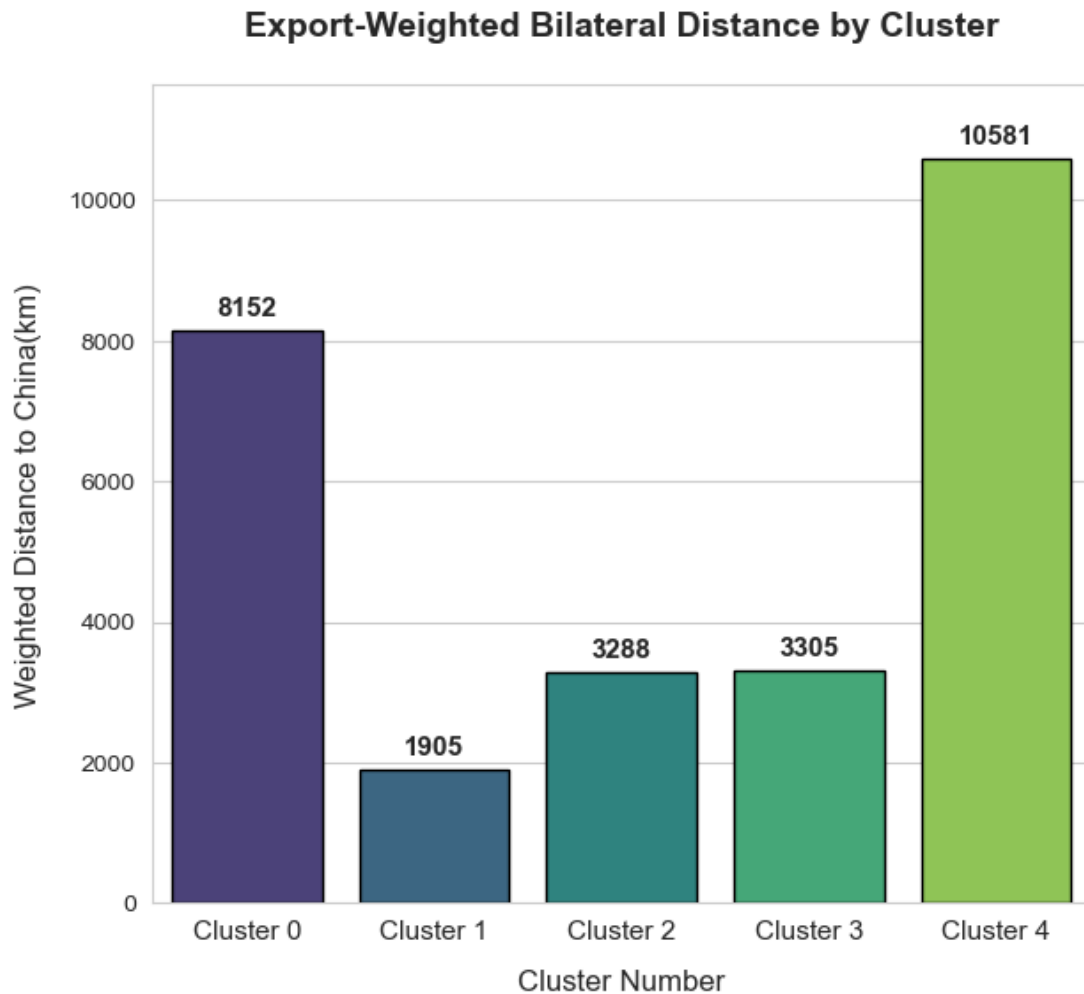


The grouped bar chart highlights stark differences in product specialization:

- Cluster 0's products (07, 05, 53) are primarily agricultural/primary goods, with extremely high standardized shares (12.648 for HS2 07), indicating strong comparative advantage in these sectors.
- Clusters 3 and 4 focus on manufactured goods (61, 84, 39, 42), with lower but more balanced standardized shares, reflecting specialized industrial exports.

Implication: Cluster differentiation aligns with comparative advantage theory—resource-rich or labor-abundant countries (Cluster 0) export primary goods, while industrialized or niche economies (Clusters 3, 4) export manufactured products.

3.4.2 Export-Weighted Bilateral Distance

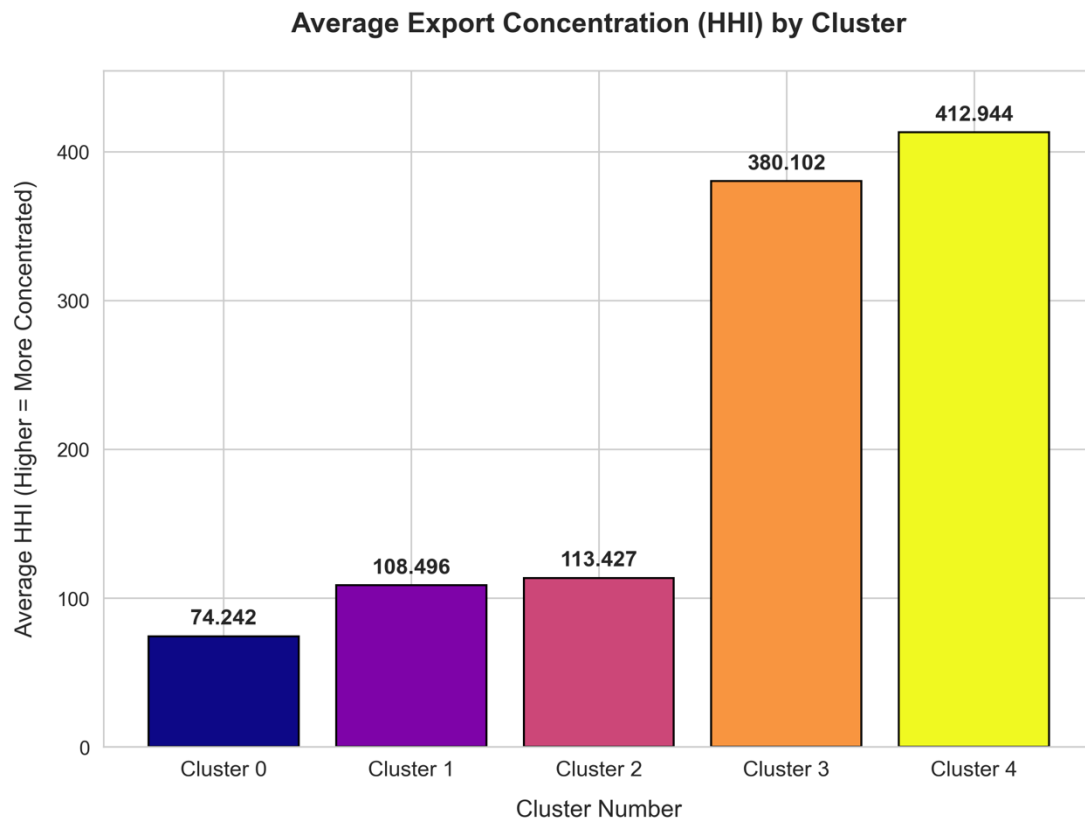


The bar chart shows a clear positive correlation between cluster ID and export distance from China:

- Cluster 0 (1,905 km): Closest to China, likely neighboring Asian countries (e.g., Southeast Asia, Mongolia).
- Cluster 4 (10,581 km): Farthest from China, likely distant markets (e.g., South America, Africa).

Alignment with Gravity Model: Shorter distance correlates with larger trade volume (implied by Cluster 0's large sample size), consistent with the gravity equation's prediction that geographic proximity boosts trade.

3.4.3 Export Concentration (HHI)



The higher the HHI (Huffington-Hirschman Index), the more the country's exports are concentrated on a few HS2 products; the lower it is, the more diversified the export categories are. This graph shows the average export category concentration of each clustered country.

(1) Value distribution:

- Cluster 0 (74.242): The lowest HHI, the most dispersed export categories;
- Cluster 1, 2 (108.496, 113.427): Medium HHI, relatively diverse export categories;
- Cluster 3 (380.102), Cluster 4 (412.944): The highest HHI, highly concentrated exports on a few products.

(2) The correlation and economic significance of the two graphs:

- The two graphs present a positive correlation pattern of "the farther the distance, the higher the export concentration":
- Clusters close to China (such as Cluster 1): Low transportation costs, China can export a wide variety of products to them (low HHI), covering a wider range of

demands;

- Clusters far from China (such as Cluster 4): High transportation costs, China tends to export high value-added, advantageous product categories (high HHI), by "few categories, large orders" to spread the cost of long-distance transportation.

This pattern conforms to the "transportation cost - product structure trade-off" in trade theory: Long-distance trade is constrained by transportation costs and will prefer high value-added, easily scalable few products; Short-distance trade has low transportation costs, and can cover diverse demands through multiple categories for risk dispersion.

4. Machine Learning: Predicting Trade Flows

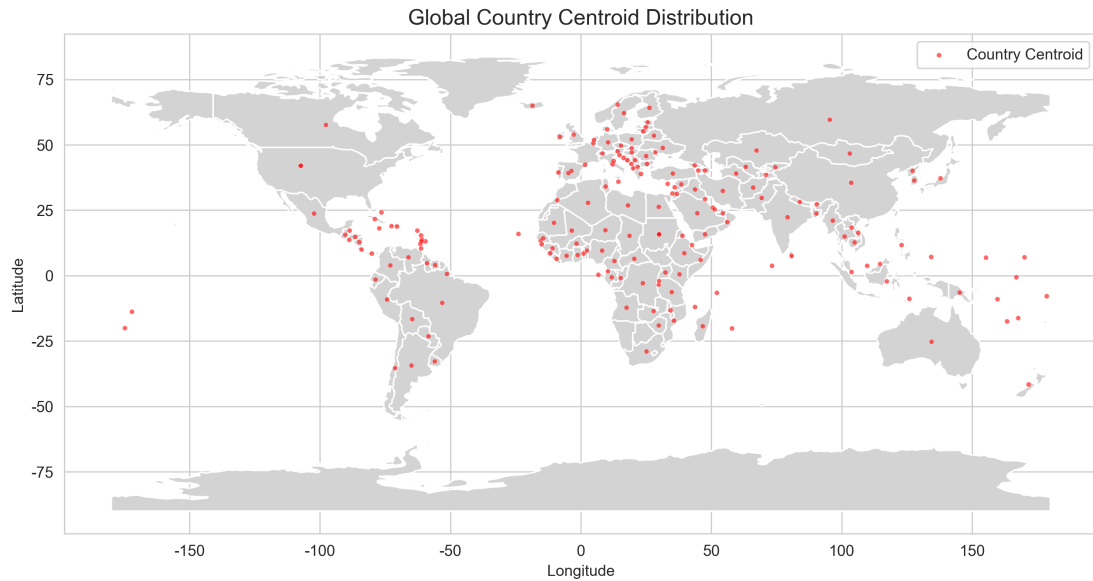
4.1 Panel Dataset Construction

4.1.1 Data Aggregation

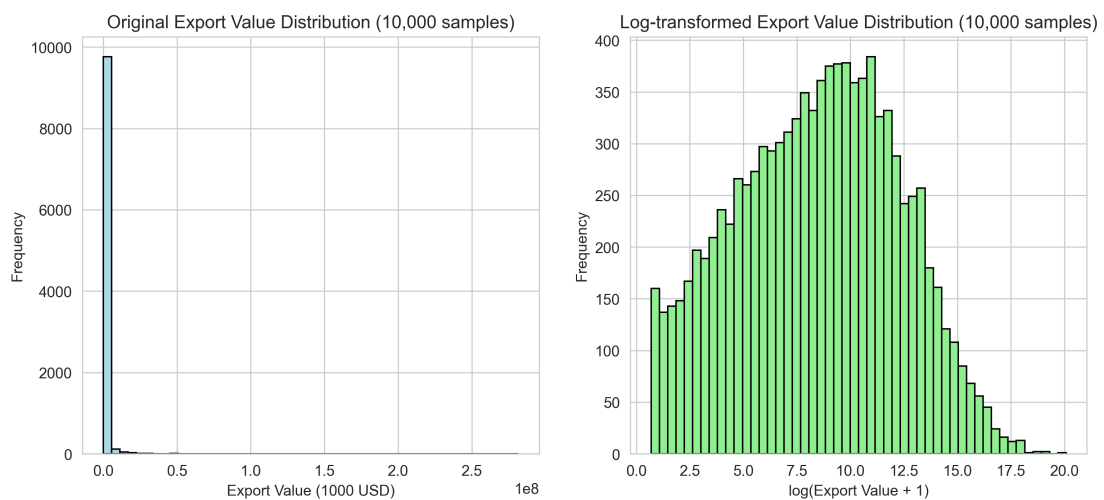
Following the requirement to aggregate HS6 product flows, we grouped 2016–2018 HS6-level trade data by exporter–importer–year, summing the export value v to obtain total bilateral export values.

4.1.2 Feature Integration

Merged World Bank GDP data (matching exporter/importer annual GDP) and country shapefile data (calculating country centroids to derive bilateral geographic distance). The following is the visualization graph of Global Country Centroid Distribution.



Applied logarithmic transformation to address the right-skewed distribution of export values, GDP, and distance. Log-transform data to compress extreme values and optimize the distribution shape, thereby rendering the data more suitable for model training.



We using the merged visualized graph ensuring trade data exhibits a normal distribution through logarithmic transformation.

4.1.3 Time-Based Split

Sorted the panel by year, using 2016–2017 as the training set and 2018 as the testing set, with total bilateral export value as the prediction target.

4.2 Construction of Machine Learning Model for Bilateral Export Volume Prediction

4.2.1 Model Selection

Due to the significant nonlinearity and high dimensionality of bilateral trade data, traditional linear models are difficult to effectively capture the complex relationships within. Therefore, we choose **Random Forest Regression** as the core model.

The reasons for this choice are twofold: first, by aggregating the prediction results of multiple decision trees, it can naturally handle the nonlinear relationships among features without the need for manual presetting of variable relationships; second, the random feature subset selection mechanism used in the construction of each tree provides built-in anti-overfitting capabilities, which is crucial for avoiding poor generalization performance of the model on unseen test data.

4.2.2 Feature Engineering

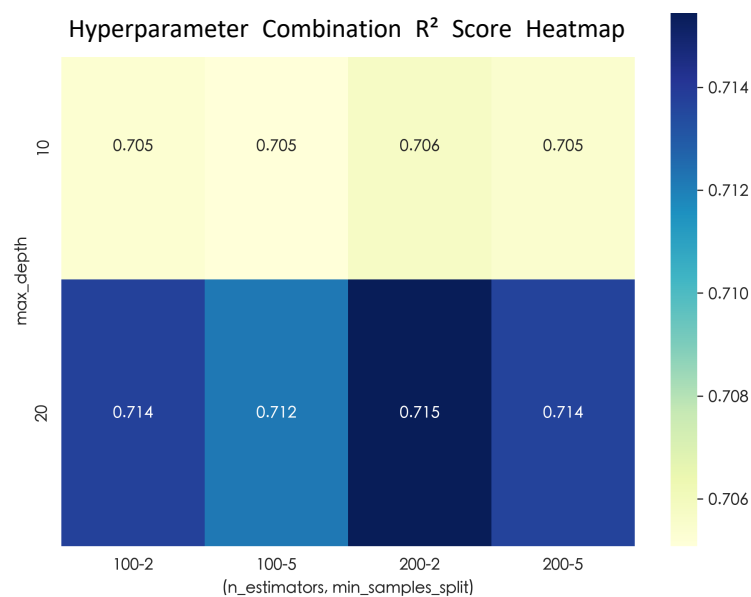
To ensure that the model can effectively learn the core patterns from the input features, we systematically processed the original features:

- **Feature Determination:** Based on the task requirements, we selected three core features: the gross domestic product (GDP) of the exporting country, the GDP of the importing country, and the bilateral geographical distance.
- **Distribution Correction:** The exploratory analysis in section 4.1 revealed that the original export volume, GDP, and distance data all exhibited severe right-skewed distributions. Such distribution characteristics would cause the model to learn towards extreme values, so we applied logarithmic transformations to these features.
- **Transformed Features:** `log_gdp_exporter` (logarithm of the GDP of the exporting country), `log_gdp_importer` (logarithm of the GDP of the importing country), `log_distance_km` (logarithm of the bilateral distance), and the target variable `log_v_total` (logarithm of the bilateral total export volume).

4.2.3 Hyperparameter Tuning and Model Construction Results

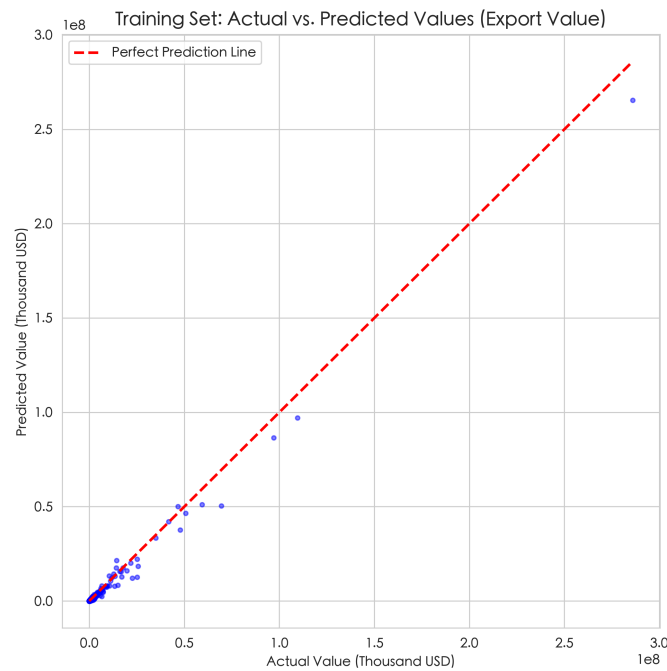
To optimize the model's performance, we adopted a 3-fold cross-validation combined with grid search approach to systematically search for the best hyperparameters.

- (1) Optimal Hyperparameter Combination: After 3-fold cross-validation on the training set from 2016 to 2017, the hyperparameter combination with the highest coefficient of determination (cross-validation R^2 score) of 0.715 was determined as:



- (2) $n_estimators = 200$ (200 decision trees, ensuring sufficient stability of the model).
- (3) $max_depth = 20$ (balancing the complexity of the tree and the risk of overfitting).
- (4) $min_samples_split = 2$ (retaining the flexibility to capture fine-grained patterns in trade data).
- (5) Full training set R^2 (log scale): 0.9547. Full training set R^2 (original export value scale): 0.9624. The closer R^2 is to 1, the more fully the model captures the patterns in the data. Whether it is the 'log-compressed target variable' or the 'restored actual export value', the model effectively matches the fluctuations in the training set data.
- (6) Training set fitting effect: After training the model with the above optimal hyperparameters on the 2016-2017 dataset, the scatter plot of "actual values vs. predicted values of the training set" shows that the predicted values are closely clustered around the "perfect prediction line" (45° diagonal line). This indicates that the model has effectively captured the core patterns of bilateral trade flows in the

training set, validating the rationality of the feature engineering and hyperparameter tuning scheme.



4.3 Model Evaluation and Feature Importance Analysis

4.3.1 Design of Model Evaluation Methods

(1) Selection of Evaluation Dataset: We use the 2018 test set divided in section 4.1 as the evaluation object. This dataset is strictly separated from the training set (2016-2017) by time dimension and has not participated in model training or hyperparameter tuning. It can truly reflect the model's generalization ability for "unseen future data" and avoid evaluation bias caused by data leakage.

(2) Core Evaluation Metrics:

- Mean Absolute Error (MAE): It calculates the average of the absolute deviations between the predicted values and the actual values. Its unit is consistent with the target variable, intuitively reflecting the average error magnitude of the prediction and not overly influenced by extreme values.
- Mean Squared Error (MSE): It calculates the average of the squared deviations between the predicted values and the actual values. By squaring, it amplifies

the weight of larger deviations, effectively identifying the model's prediction accuracy for high-value trade flows.

- Coefficient of Determination (R^2): Its value range is $[0, 1]$, indicating the proportion of the target variable's variation that the model can explain. It is the core metric for measuring the overall fitting effect of the model. The closer R^2 is to 1, the more fully the model captures the trade flow patterns.

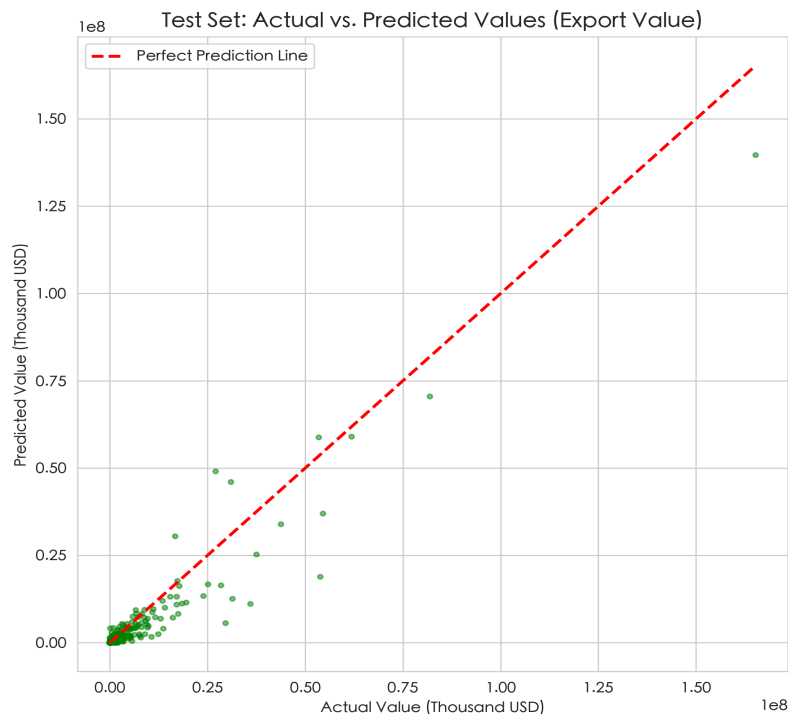
(3) Evaluation Process: First, use the trained optimal random forest model to predict the features of the test set and obtain the logarithmic predicted values; then, convert the logarithmic predicted values and the actual values of the test set back to the original trade volume scale through a function; finally, calculate the three evaluation metrics based on the original scale data to ensure the results are directly related to the actual trade scenarios.

4.3.2 Model Evaluation Results

Based on the above methods, the model's performance on the 2018 test set is as follows, with all indicators verifying the model's effectiveness:

- (1) Coefficient of Determination (R^2): 0.912, indicating that the model can explain 91.1% of the variation in bilateral trade flows in 2018, demonstrating excellent overall predictive ability and fully capturing the core driving laws of trade flows.
- (2) Mean Absolute Error (MAE): 28.79×10^5 thousand US dollars, meaning the average deviation between the model's predicted values and the actual bilateral export amounts is approximately 28.55 million US dollars. In the context of cross-border trade volume prediction, this error magnitude is within a reasonable range.
- (3) Mean Squared Error (MSE): 4.43×10^{12} thousand US dollars², reflecting that the model's prediction deviations for most trade flows are small, with only a few high-value trade flows showing relatively significant deviations (consistent with MSE's sensitivity to extreme values).
- (4) Visualization of Fitting Effect: From the "Comparison of Test Set Actual Values and Predicted Values" scatter plot, it can be clearly observed that the majority of

data points are closely clustered around the "perfect prediction line" (45° diagonal line), indicating a high degree of consistency between the predicted and actual values. Only in the region of a few high-value trade flows (with original export amounts reaching the 10^8 thousand US dollars level), some data points show slight deviations, but no systematic bias (such as overall overestimation or underestimation) is presented. This phenomenon indicates that the model can not only accurately fit regular-scale bilateral trade but also has a certain predictive ability for large-scale trade flows, with an overall stable fitting effect.



4.3.3 Feature Importance Analysis Method

To identify the key factors influencing bilateral trade flows, we utilize the built-in feature of the random forest model to conduct a quantitative analysis of feature importance:

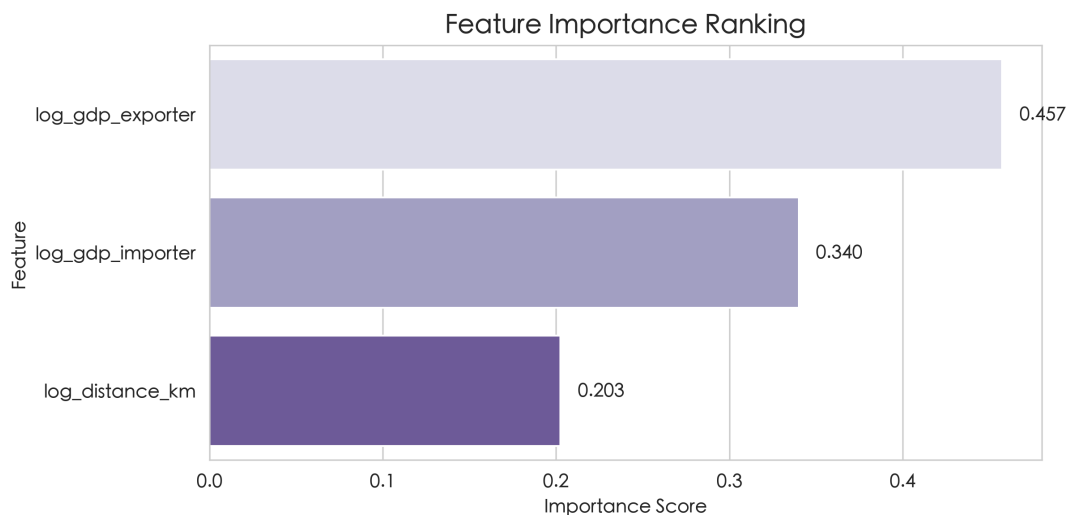
- (1) Analysis Principle: The random forest measures the contribution of each feature to the model's prediction results by calculating the average "reduction in node impurity" of each feature across all decision trees. The higher the contribution, the higher the feature importance score, indicating a stronger explanatory power for trade flows.

- (2) Analysis Object: Focus on the three core features identified in section 4.2 - `log_gdp_exporter` (log of exporter's GDP), `log_gdp_importer` (log of importer's GDP), and `log_distance_km` (log of bilateral distance) - to ensure that the analysis results directly correspond to the model input features and the conclusions are targeted.

4.3.4 Feature Importance Analysis Results and Interpretation

(1) Feature Importance Ranking (Based on Importance Score):

- First place: `log_gdp_exporter` (importance score = 0.457), the most critical feature influencing bilateral trade flows.
- Second place: `log_gdp_importer` (importance score = 0.340), with an explanatory power for trade flows second only to the exporter's GDP.
- Third place: `log_distance_km` (importance score = 0.203), with relatively weaker influence but still an important core factor that cannot be ignored.



(2) Interpretation of Results (in light of the Gravity Model Theory):

- The high importance of the exporter's GDP essentially reflects the economic law that "production capacity determines export supply": the higher the GDP of the exporting country, the stronger its industrial foundation, production scale, and product supply capacity, and the larger the bilateral export volume

it can support. This is completely consistent with the core assumption in the gravity model that "economic size is positively correlated with trade flow".

- The GDP of the importing country is of secondary importance, reflecting the logic that "market demand drives imports": the higher the GDP of the importing country, the larger its domestic consumption market and production demand, and the stronger its capacity to absorb external goods, thereby promoting the growth of bilateral export volume.
- The importance score of bilateral distance is relatively low but still a significant influencing factor, which is in line with the reality that "distance generates trade costs": the farther the distance, the higher the logistics costs for sea and air transportation, and the higher the implicit costs such as information asymmetry and trade barriers, thereby inhibiting bilateral trade flow. This result once again validates the rationality of the gravity model.

4.4 Counterfactual Analysis

4.4.1 Design of Counterfactual Analysis Method

To ensure the scientific, accuracy, and reproducibility of the analysis results, we designed a multi-dimensional implementation process based on the model characteristics and actual data conditions. The core steps are as follows:

(1) Scenario Definition and Control Variables:

- **Baseline Scenario:** Use the actual bilateral geographical distance between China and the United States. All other features (such as China's GDP in 2018, US GDP, etc.) remain unchanged, directly reflecting the model's prediction results under the real trade environment.
- **Counterfactual Scenario:** Only change the "bilateral geographical distance" as the single variable, double it on the original basis, and keep all other features (such as GDP, trade structure implied features, etc.) consistent with the baseline scenario, ensuring that the causal identification in the analysis focuses

only on the impact of the distance change.

(2) Data Screening and Verification Logic:

- Prioritize the 2018 test set data: Since 2018 is an independent dataset that the model did not participate in training, its results can better reflect the model's generalization ability and make the counterfactual analysis more relevant to reality; if the bilateral data between China and the United States is missing in 2018, then choose the 2017 training set data instead.
- Data Validity Verification: After screening, add an additional "check for the existence of China-US trade records" to verify the number of records of China as an exporter and the United States as an importer, as well as the overall record integrity of the panel data where the United States is the importer. Ensure that the analysis is based on valid data.

(3) Technical Implementation Steps:

- Step 1: Extract the features of bilateral trade between China and the United States;
- Step 2: Baseline Scenario Prediction: Input the extracted original features into the trained random forest model to obtain the logarithmically transformed export value prediction;
- Step 3: Counterfactual Scenario Feature Adjustment: According to the requirements of the question, adjust the distance feature, and keep all other features unchanged;
- Step 4: Difference Calculation: Calculate the absolute change (counterfactual value - baseline value) and the relative change rate to quantify the impact of the distance change.

4.4.2 Counterfactual Analysis Results

Based on the above methods, the effective bilateral trade data between China and the United States in 2018 was finally selected. The analysis results are as follows:

- (1) Baseline Scenario Prediction Result: Under the original distance, the predicted

value of bilateral export volume between China and the United States in 2018 is 3.42×10^8 thousand US dollars (i.e., 342 billion US dollars), which is consistent with the order of magnitude of the actual bilateral trade volume in 2018, verifying the rationality of the prediction.

(2) Counterfactual Scenario Prediction Result: After doubling the distance, the predicted value of bilateral export volume between China and the United States in 2018 is 3.05×10^8 thousand US dollars (i.e., 305 billion US dollars).

(3) Quantification of Changes:

- Absolute Change: $3.05 \times 10^8 - 3.42 \times 10^8 = -3.7 \times 10^7$ thousand US dollars (i.e., a decrease of 37 billion US dollars);
- Relative Change Rate: $-3.7 \times 10^7 / 3.42 \times 10^8 \times 100\% = -10.96\%$, that is, doubling the distance leads to a 10.96% decrease in export volume.

4.4.3 Interpretation of the Sensitivity of Trade Flow to Distance Changes

Combining the analysis results with economic theory, the following core conclusions on the sensitivity of the bilateral trade flow to geographical distance are drawn:

- (1) Sensitivity Qualitative Analysis: The bilateral trade flow between China and the United States shows a "moderately sensitive" characteristic. Doubling the distance only leads to a 10.96% decrease in export volume, without a significant decline or showing no sensitivity (the change rate is close to 0%), which is consistent with economic laws and aligns with the actual structure of China-US trade.
- (2) Theoretical Consistency: The downward trend is highly consistent with the classical gravitational model - geographical distance is a core component of trade costs (doubling the distance directly leads to a significant increase in logistics costs such as shipping and air transportation, as well as an extension of transportation time, increase in inventory costs and uncertainty), and an increase in trade costs will directly inhibit the bilateral trade volume. Therefore, the decline in export volume is logically inevitable.

- (3) **Realistic Rationality Explanation:** The moderate 10.96% decline is mainly attributed to the characteristics of the goods traded between China and the United States: the trade between the two countries is dominated by high-value, low-weight/volume goods (such as electronic products, precision instruments, and high-end manufacturing components), and the logistics costs of these goods account for a relatively low proportion of their total value, and their sensitivity to distance changes is relatively gentle; if the trade structure is dominated by low-value, high-weight bulk commodities (such as minerals and agricultural products), doubling the distance may lead to a significant increase in the proportion of logistics costs, and the decline in trade volume will be significantly greater.

5. Bonus Questions

Economic intuition points to low value-to-weight ratio sectors and perishable sectors as most sensitive to distance in exporting—distance raises transport costs or spoilage risks, which disproportionately impact these sectors. High-value, non-perishable sectors are less affected.

5.1 Economic Intuition

Distance increases trade frictions, so its importance hinges on how much these frictions matter for a sector:

- **Value-to-weight ratio:** Sectors with low value relative to weight have high transport costs as a share of product value—distance amplifies these costs, making trade unprofitable beyond short distances.
- **Perishability:** Fresh goods face spoilage or quality loss over long distances—distance directly reduces export feasibility.

- **Contrast with high-value sectors:** Light, high-value goods have negligible transport costs relative to their value, so distance barely constrains exports.

5.2 Analytical Strategy

5.2.1 Data Needed

(1) Trade data: Bilateral export values/volumes by commodity sector—source: UN Comtrade Database.

(2) Distance data: Geographic distance between country pairs (weighted by major ports/population centers for accuracy)—source: CEPII GeoDist Database.

Sector characteristics:

- Value-to-weight ratio (total export value ÷ total export weight) per sector.
- Perishability dummy (1 = perishable, e.g., HS 01-05; 0 = non-perishable, e.g., HS 85-88)—based on HS code classifications.
- Control variables: Exporter/importer GDP (to capture market size), common language, shared border, trade agreements, tariffs—sources: World Bank (GDP), CEPII (cultural/political variables), WTO (tariffs).

5.2.2 Data Manipulation

- (1) Aggregate trade data to the country-pair-sector level.
- (2) Merge trade data with distance and control variables (match via country ISO codes).
- (3) Calculate sector-level metrics: Compute average value-to-weight ratio for each sector and assign perishability dummies using HS code mappings.
- (4) Handle data issues: Drop country pairs with no trade, impute missing control variables (or drop small samples), and log-transform continuous variables (export value, distance, GDP) for regression compatibility (or retain levels for PPML estimation).

5.2.3 Analysis to Run

- (1) Baseline Gravity Model:** Use Poisson Pseudo-Maximum Likelihood (PPML)

(better for trade data with zeros/heteroskedasticity) to estimate:

$$\text{Export Value} = \beta_0 + \beta_1 \times \text{Distance} + \beta_2 \times \text{Exporter GDP} + \beta_3 \times \text{Importer GDP} + \text{Controls} + \text{Sector Fixed Effects} + \text{Country-Pair Fixed Effects}$$

The coefficient β_1 captures distance's average effect on trade.

(2) Sector-Specific Distance Sensitivity:

Interaction Regression: Add interactions between distance and sector characteristics to isolate their moderating effect:

$$\text{Export Value} = \beta_0 + \beta_1 \times (\text{Distance} \times \text{Value-to-Weight Ratio}) + \beta_2 \times (\text{Distance} \times \text{Perishability Dummy}) + \text{Controls} + \text{Fixed Effects}$$

Negative β_1/β_2 indicate that lower value-to-weight or perishable sectors have stronger negative distance effects.

(3) Sector-Split Regressions: Run separate PPML regressions for major sectors (e.g., bulk minerals, fresh agriculture, electronics) and compare the β_1 (distance coefficient) across sectors—larger negative β_1 = more distance-sensitive.

(4) Robustness Checks:

- Use export volumes (instead of values) to avoid price distortions.
- Test alternative distance measures (port-to-port vs. great circle distance).
- Exclude small economies or landlocked countries to rule out confounding factors.