Mini Group Project II

Free International Trade

CNGF 5020

Due date: Dec 3rd (Presentation) & Dec 10th (Other Materials)

For the CNGF 5020 Environmental Modeling and Big Data Analytics, this mini-project is designed to bridge the gap between theoretical knowledge and practical application in the context of empirical social science research.

This project will challenge you to examine international trade by analyzing a comprehensive dataset of bilateral trade flows. The primary goal is to apply data science techniques, including machine learning and econometric analysis, to preprocess trade data, uncover patterns, and investigate the effects of distance and economic indicators on trade outcomes. Students will explore the gravity equation, estimate trade flow models, and apply supervised and unsupervised learning methods to predict export values and identify trade patterns across countries. Through this project, students will gain a deeper understanding of the interaction between geography, economics, and trade while practicing data manipulation, visualization, and statistical analysis.

Requirement:

- Group project: work in groups of 3–5 students.
- Submission consists of two components:
  - A 10-minute in-class presentation on **Dec 3rd**, demonstrating your methodology and key findings.
  - A project repository uploaded to Canvas, including all provided data, well-documented Python scripts/notebooks, and a comprehensive PDF file that serves as your final report. This repository is due on **Dec 10th**.

Data

The folder `trade_data.zip` contains annual bilateral trade records at the product level. Each file `BACI_HS12_YXXXX` reports all export flows from one country to another in a given year at the HS6 product level. The key variables in each file are:

- **t**: the year of the trade flow (e.g., 2017).
- **i**: exporter country code (ISO numeric code).
- **j**: importer country code (ISO numeric code).
- **k**: HS6 product code, a six-digit Harmonized System classification of goods.
- **v**: trade value, measured in thousands of current US dollars.
- **q**: trade quantity, typically measured in metric tons.

The folder also includes two reference files: `country_codes_V202001`, which links ISO numeric codes to country names, and `product_codes_HS12_V202001`, which provides descriptions for each HS6 product. HS6 codes can also be grouped into broader HS2 categories by padding each code to six digits and taking the first two digits.

In addition, the project directory includes a World Bank GDP dataset, which reports annual GDP (current US dollars) for all countries. Students may use this GDP information as a country-level feature in the machine learning section by merging it with the bilateral panel dataset using country codes and year.

Questions are as follows:

1 Descriptive (20)

- Read in data and generate the following descriptive statistics: who were the top 10 countries with the most trading partners, what about the bottom 10 (for total three years)? (5 points).
- Match up the trade data with the product and country codes; describe the trade volume of the whole dataset in terms of value. For China and the United States, identify the top 10 partner countries in terms of total trade value and report the corresponding trade values. In addition, list the five highest-value China–partner trade flows over the sample period. (5 points)
- Calculate the top 10 export products (in terms of value) of China, Japan, and the United States over the sample period, and calculate the top 10 goods with the highest trade volume for both exports and imports in terms of value and quantity (5 points).
- Using the country shapefile, calculate the distance in kilometers between the centroid of China and all the other centroids. You may use either haversine distance with the geographic CRS or use a projected CRS and use the `distance` function of `geopandas`.Please also create a scatterplot of distance and export volume (quantity and value) in logarithm form (5 points).

2 Unsupervised Learning: Identifying Trade Patterns (40)

1. Normalize the trade data (e.g., using min–max scaling or standardization) to make sector-level measures comparable across countries . Aggregate export values into broad two-digit product categories (the first two digits of the HS code) for 2016–2018 and convert them into export-share vectors for each country. (10 points)
2. Apply K-means or hierarchical clustering to group countries based on these HS2 export-share vectors (10 points).
3. Use PCA or another dimensionality-reduction method to visualize the clusters and show how countries differ in terms of export structure (10 points).
4. Analyze the characteristics of each cluster (10 points), including: (a) main product categories in each group; (b) export-weighted bilateral distance; (c) export concentration based on the Herfindahl–Hirschman Index (HHI).

Hints:

- HS6 product codes may appear as 5-digit numbers when opened in Excel. If so, pad a leading zero to restore the 6-digit code before extracting HS2 (e.g., 10121 → 010121).
- Each country is represented by an HS2 export-share vector; clustering is performed in this vector space.
- PCA can be applied to the HS2 share matrix to visualize how countries separate across principal components.
- HHI is computed as the sum of squared export shares across HS2 categories.

3 Machine Learning: Predicting Trade Flows (40)

1. Construct a panel dataset at the exporter–importer–year level by aggregating HS6 product flows into total bilateral export values for each pair of countries in each year . Sort the panel by year and split it along the time dimension (for example, use 2016–2017 as the training set and 2018 as the testing set). Use the bilateral export value **v** as the prediction target. (10 points)
2. Build a machine learning model (e.g., Random Forest, Gradient Boosting, or a simple Neural Network) to predict bilateral export values using features such as geographic distance, GDP, and product or country characteristics (10 points). Apply reasonable feature engineering and hyperparameter tuning.
3. Evaluate the model on the testing set using MAE, MSE, and R-squared (10 points). Discuss how well the model fits the data and comment on which features appear to be most important for explaining bilateral trade flows.
4. Conduct a counterfactual exercise (10 points). Double the bilateral distance between China and the United States while keeping all other features constant. Use the trained model to predict the new export value and compare it with the baseline prediction. Comment on the sensitivity of predicted trade flows to changes in distance.

Hints:

- Start from the BACI HS12 files and first aggregate export values over HS6 products so that each observation represents a country pair in a given year (t, i, j) with a total export value v_{tij}^{total}. Each year can be viewed as a 221×221 matrix of bilateral exports; reshape this matrix into a long-form panel for modelling.
- Perform the time-based train–test split at the panel level, not at the individual HS6-product level, to avoid mixing information across years.

4 Bonus questions

An analyst asks you "for which commodity sectors is distance most important in terms of exporting?" Describe what your economic intuition would say about the answer, and then also map out an analysis strategy for answering their question. As usual, describe:

- What data you would need.
- How you would manipulate the data.
- What analysis you would run.