WS20/21 - Machine Learning: Applications in Economic Research

Predicting Trade Flow

Data Analysis Project

Yangfan Lyu / Andreas Thiele

15.3.2021

Table of Content

1. Research question

Our investigation deals with the question whether machine learning methods can improve out-of-sample prediction of high-dimensional bilateral trade flows. Conventional analysis is usually based on the gravity equation that accounts for economic size of both origin and destination country, geographic distance, price index, and multilateral resistance terms such as tariff, common borders, and variables related to language and political status. In addition, we are also interested in answering the question whether other variables such as greenhouse gas (GHG) emissions and the inflation rate are likewise found to play a part in determining bilateral trade flows. We include GHG because we intuitively expect that emissions are positively related to international trade; that is, countries that trade more intensively with each other, are likewise expected to grow (higher GDP), achieve higher standards of living (GDP/capita), expand to more markets and reach more destinations to expand their market size (Melitz, 2003), and consequently generate more pollution. The importance of including GHG is further reinforced by Cristea et al. (2011), that have pointed out that 33% of measured GHG emissions can be attributed to international trade. Nonetheless, we do not know how important GHG emissions are in trade and to what degree they are contributing to explain trade flows. The inflation rate, which is approximated by the average price index (API) of a particular country, is included because it is considered to potentially alter both the direction and the size of international trade by changing optimality conditions regarding to labor and capital as production inputs (Stockman, 1985).

2. Data used

The primary dataset we have used is TRADHIST from CEPII[1], a bilateral trade historical time series from 1827 to 2014, consisting of five key variables: (1) bilateral nominal trade flows, (2) country level aggregate nominal exports and imports, (3) nominal GDPs, (4) exchange rates, (5) bilateral factors that could favor or hamper trade, such as geographical distance, common borders, colonial and linguistic links, as well as bilateral tariffs and indicator for political and economic union. Subsequently, we extend this dataset to include global greenhouse gas (GHG) emission from Kaggle and an indicator average price index (API) to

[1] http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=32.

represent inflation for a country-pair for given years. After obtaining our final dataset, our study period ranges from 1990 to 2014. Finally, 90% of our data will be used as training set and the remaining 10% for testing purposes.

2.1. GHG Inventory Data

The dataset contains GHG emissions of each country from 1990 to 2014. We merge GHG emissions data with our primary dataset based on the country's iso code and year. As GHG emissions are denominated in total value for each country, we decided to instead use the weighted value of GHG emissions to pin down the emissions induced by bilateral trade. When considering the weights, we need to think about where emissions could come from in this process.

Cristea et al. (2011) have mentioned that emissions are associated with output and transport in every origin-destination-product trade flow. GDP in both origin and destination can therefore be used to indicate output. However, for transport we decided to use bilateral trade flow as a proxy for transportation because increased bilateral trade flow is often connected with more intensive public transportation, which in turn is considered to induce higher GHG emissions. For this reason, we created a weighted value of GHG emissions in each entry using GDP in both origin and destination and bilateral trade flows as the weights to indicate potential GHG emissions induced in bilateral trade. The following equation shows our estimation approach:

$$GHG_{ijt} = \left(\frac{GDP_{it} + GDP_{jt}}{GDP_t} + \frac{bilateral\ trade\ flow_{ijt}}{total\ trade\ flow_{it}} \right) \times total\ GHG_{it},$$

where GHG_{ijt} indicates GHG emission induced by trade flow between origin country i and destination j in time t. GDP_{it}, GDP_{jt}, $bilateral\ trade\ flow_{ijt}$ and $total\ trade\ flow_{it}$ can in turn be obtained from the TRADHIST dataset.; the variable $total\ GHG_{it}$ however is from the GHG inventory dataset.

2.2. Average price index

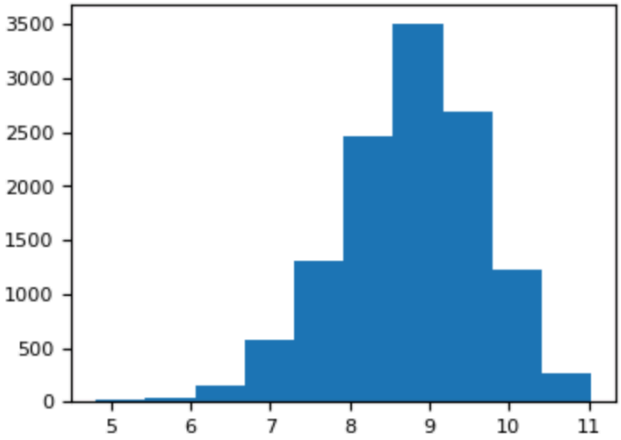
The dataset on average price indices (API) of countries for time period 1960 to 2019 is obtained from the online database[2] of "The World Bank". The underlying reasoning for extending our original dataset lies in the economic consideration that inflation is deemed to impact international trade patterns (Stockman, 1985). After downloading the dataset, the given columns involve "Country Name", "Country Code", "Indicator Name", "Indicator Code" and the remaining columns detailing the time periods. As merging requires applying an inner join with a key that identifies entries in both the original and the new dataset, we eliminate all columns except "Country Code" and the time periods. Finally, the structure of the columns for the time periods (horizontally) does not match the vertical layout present in our original dataset. For this reason, we conduct a transposing measure for all columns except "Country Code" to translate the horizontal axis into a vertical one. After completing this step, we finally apply the inner join and successfully merge our original dataset with our new column "API" detailing the average price index for any given country from time period 1960 to 2019.

2.3. Data cleaning techniques

In order to reserve as much information as possible for further analysis, we have applied several steps to clean our dataset. First, we check the share of NaNs ("not a number" in Python format) in each variable and drop the columns which have a value of above 80%. Furthermore, we use the reduced primary dataset to do a pre-test of random forest (RF) to check the feature importance and eliminate variables which are 1) deemed less significant for prediction and 2) likewise exhibit a high share of NaNs. Afterwards, we remove identified NaNs in the remaining columns of our final dataset. Finally, we take the logarithm of our data to eliminate the long tail present in our distribution (see figure 1). After successfully merging it with the other two datasets, we have 12227 observations left.

[2] https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG

Figure 1: frequency distribution of log (Flow)



3. Methods

We are going to employ linear regression, Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression Tree (CART) and RF in our methodological approach. Linear regression, including country fixed effects and time fixed effects, will constitute our benchmark model. Further, we want to check whether linear shrinkage and nonlinear nonparametric models can improve our prediction of bilateral trade flows.

3.1. LASSO

This linear shrinkage model penalizes the coefficients associated with irrelevant variables to zero. Therefore, the consistency for selecting the most relevant variables each time will be achieved only when irrepresentable conditions are fulfilled (Zhao and Yu 2006, Meinshausen and Yu 2009), meaning that the selected variables should be as less-correlated as possible to the unselected ones, because otherwise our estimators could be biased; this is also referred to as sparsity in the observations.

The penalty is given as follows:

$$\sum_{i=1}^{n} p(\beta_{h,i};\ \lambda, \omega_i) \coloneqq\ \lambda \sum_{i=1}^{n} |\beta_{h,i}|,$$

where $p(\beta_{h,i};\ \lambda, \omega_i)$ is a penalty function, $\beta_{h,i}$ is the coefficient we want to penalize, λ is the penalty parameter and ω_i is the associated weighting factor. In the case of LASSO, the penalty term treats every coefficient equally, using the same penalty parameter.

3.2. CART and RF

The RT splits the observations according to the available covariates in a way that when reaching the terminal nodes (which are also referred to as regions in Hastie et al. (2001)), the outcome variable in each set is as homogenous as possible, meaning that we need to split the observations according to a certain value of covariates k for minimizing the squared error loss. The function below illustrates our point:

$$\min \sum_{k \in left} (y_k - \hat{y}_{left})^2 + \sum_{k \in right} (y_k - \hat{y}_{right})^2,$$

where y_k is the output values in each terminal node from either left branch or right branch, and $\hat{y}_{left}$ and $\hat{y}_{right}$ is the average of the sample y_k. Therefore, RT is a non-parametric model that approximates an unknown nonlinear relation by splitting the covariates continuously until the smallest value in loss function is achieved. The RF model, in turn, was proposed by Breiman (2001) to reduce the variance of RT by averaging the noise out among randomly constructed RTs.

4. Results

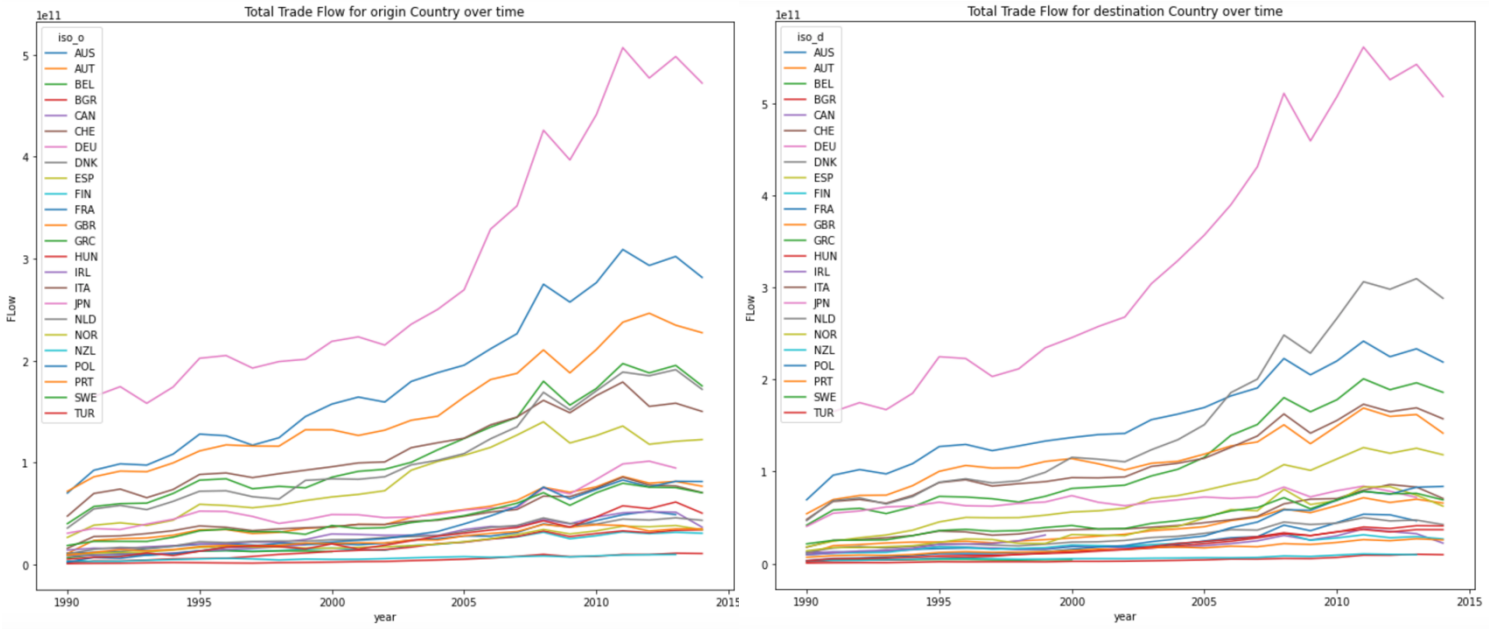
Before looking at the out-of-sample prediction from each model, we will shortly delve into the observed trade pattern for each country: Figure 2 shows the changes in trade volume for both origin and destination countries from 1990 to 2014. While Japan has the highest trade flow both as origin and destination, France has a larger trade flow as an origin, but the Netherlands started to have higher trade volume since 2006 as a destination. Overall, trade flows exhibit a persistent increasing trend and is found to be approximately linear over time. For this reason, we expect that linear regression is likely to achieve a good result in predicting bilateral trade.

The linear regression has achieved a R-squared of 0.9763, which is better than 0.9385 in the LASSO case. One possible reason for this could be that LASSO only selects 9 out of 32 features, which potentially reduces the model's predictive power by eliminating less relevant variables when the trend in outcome is easier to track. CART and RF achieve even better out-of-sample prediction results, 0.993 and 0.999 respectively, if we use the optimal tree depth (10 in this case) and the optimal number of trees grown in a forest (15

in this case). What is more, in RF we have a mean squared error of 0.002, which is smaller than 0.015 in the CART case. This could be related to the fact that the RF algorithm is a bagging of CART trees. By averaging the prediction results over multiple trees, we preserve the stable relationship among bootstrap samples and average out the noise.

One interesting thing to notice is that none of the three ML models give higher weights to GDP in both origin and destination when weighted value of GHG emissions are included in the models. Particularly in the LASSO case, the coefficients of GDP in a country-pair are penalized to zero, which is counterintuitive in terms of the gravity model. Furthermore, CART and RF consider the interaction term of API between origin and destination to be significant in our models, indicating that instead of the impact of inflation from a single country, country-pair inflation is deemed more important for determining bilateral trade flows because of taking interacting country price levels for explaining trade patterns into account:

Figure 2: Total Trade Flow for both Origin and Destination



5. Discussion and Conclusion

We expect that our analysis would have benefited from including country-pair fixed effects in our benchmark model to better account for geographical differences. However, as computational implementation in Python seemed complex, this step had necessarily to

be omitted. The same would likely also hold true for the Poisson Pseudo Maximum Likelihood (PPML) method, which could contribute to tackling observed g zero trade flows and potential heteroskedasticity present in our error terms. To better reflect the mechanics of the gravity equation, we also pursued to implement the Gravity Modeling Environment (gme) package, however unsuccessfully.

By applying LASSO, its underlying algorithm is connected to two potential problems that could impose caveats to our results. First, the penalty term does not depend on the weight ω_i and treats every feature homogeneously (i.e., with the same degree), which in turn could pose a problem for time series data. Figure 2 further allows us to discern that varying trade patterns are displayed by a large number of countries from 2007 to 2010, which seems to indicate that the financial crisis of 2007-08 influenced countries in the data had a differing impact depending on the country, the most severe ones being Japan, France, Great Britain, and the Netherlands. Therefore, we should expect that the coefficients of our variables that are related to this financial shock should be penalized differently in different countries over time, but LASSO fails to do so.

One possible solution is to use the Adaptive Least Absolute Shrinkage and Selection Operator (adaLASSO) proposed by Zou (2006) to include weighting parameters ω_i in the penalty terms for the purpose of avoiding heteroskedasticity (Medeiros and Mendes 2016). Secondly, as LASSO relies on the irrepresentability condition, which leads to the problem of inconsistent model selection, we can instead perform the Ridge Regression (RR), which shrinks the coefficient of less relevant variables to nearly zero for a given λ, thus relaxing the condition. But in any case, LASSO is considered a good starting point for regularization. Moreover, from our results we can see that linear regression has already achieved a good out-of-sample prediction and ML methods only slightly outperform our linear regression in predicting bilateral trade flows.

Two reasons for explaining this pattern could be considered: Firstly, trade flow is persistent and increases over time, meaning that we can use history data to predict future; furthermore, observations may not be completely independent of each other, which is an important formal requirement of non-parametric methods. If this assumption fails (i.e., in time-series analysis), parametric models are expected to render better predictions. Secondly, the out-performance of CART and RF comes from the nonlinearity which allows interaction among variables (i.e., country experiences that experience higher GDP growth

or have larger GHG emission are also more likely to engage more in international trade). Therefore, ML methods are deemed to perform better than conventional econometric methods when the output variable of interest is displaying a highly stochastic pattern, that is dynamic and unpredictable. Finally, we further expect ML methods, in particular RF, to provide superior predictive capabilities by capturing the interaction among covariates to better fit the nonlinear trend present in the data.

References

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.

Cristea A. D., Hummels D. L., Puzzello L., and Avetisyan M. (2011). Trade and the Greenhouse Gas Emissions from International Freight Transport. NBER Working Paper No. w17117.

Hastie, T., Tibshirami, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction, New York: Springer.

Medeiros, M., and Mendes, E. (2016). $\ell 1$-Regularization of High-Dimensional Time-Series Models with Non-Gaussian and Heteroskedastic Errors. Journal of Econometrics, 191, 255–271.

Meinshausen N., and Yu B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist. 37(1), 246 – 270.

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. Econometrica, 71(6), 1695-1725.

Stockman, A. C. (1985). Effects of Inflation on the Pattern of International Trade. The Canadian Journal of Economics, 18(3), 587-601.

Zhao, P., and Yu B. (2006). On model selection consistency of Lasso. Journal of Machine Learning Research, 7, 2541–2563.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, 101, 1418–1429.