This manuscript (permalink) was automatically generated from junrenwang/foodflowproject@e816eeb on December 6, 2020.

Authors

- **Junren Wang**
  · junrenwang
  Department of Civil and Environmental Engineering, University Illinois at Urbana-Champaign

- **Akshay Pandit**
  · Akshay163
  Department of Civil and Environmental Engineering, University Illinois at Urbana-Champaign

Abstract

United States has been a hub of bilateral trade, both internationally and domestically. In our study, we looked specifically at the food flows, i.e. the bilateral trade of food produce, on a domestic scale among its 50 states. We implemented two prevalent machine learning algorithms (Neural Networks and Random Forest) to predict the food flows. We also compared our results with one of the most sought-after models for bilateral trade, the gravity model to validate the relevance of application of machine learning models in our analysis. We find out that the random forest model is the best model out of all the three selected models with a R^2 score of 0.86. The gravity model performed quite poorly because of its linear nature which emphasizes the pertinence of non-linear models in food flow predictions.

Introduction

The article Machine learning in gravity models: An application to agricultural trade written by Munisamy Gopinath et al. employed supervised and unsupervised machine learning (ML) method to decipher patterns of international agricultural trade. Gravity model is one of the most robust empirical models to illustrate the drivers in international trade: bilateral trade between two countries is proportional to size, mostly measured in GDP and inversely proportional to “distance” between them, which commonly fitted through Poisson Pseudo Maximum Likelihood (PPML) method. Munisamy et al. leveraged the decision trees (LightGBM, XGboost), random forests and extra tree regression supervised ML algorithm to predict the bilateral trade. The data used in this project is international bilateral trade information for seven most traded food commodities from 1962 to 2016. The gain and loss was measured by adjusted R-square. The values of maximum depth of the tree, learning rate, number of leaves and feature fraction were tuned to achieve a better prediction. ML methods show more accurate prediction results than gravity model with the adjusted R-square ranged between 45 and 83%. Specifically, LightGBM had the best performance for sugar; Random Forest provided the best fit for corn; and the extra tree regressor yielded highest R-sqaure for beef and milk powder. The size of the two countries has the largest influence to the trade, the distance follows, which is consistent to the gravity model assumptions. Munisamy et al. also employed multilayer perceptron (MLP), one paradigm of unsupervised ML method for the same datasets. The loss was measured by stochastic gradient descent (SGD) method. Unsupervised ML techniques might be a better method for longer-term trade projections than supervised ML.

Munisamy et al. clearly define the quantitative ML algorithm that we can follow to predict the trade and variables we can also consider for national flows. Comparison between the ML algorithm and PPML method shows that ML is a promising method for the topic we interested in. Also, Munisamy’s work indicates a potential problem we might encounter: the zero values prevalent in the trade data might impair the ML model and deviate it to a wrong direction. We suppose that using the qualititive ML method like discriminant analysis and K-Nearest Neighbors to predict the existance of trade and build the quanlititive ML model on the highly potential existing link might produce a better prediction. Also, because Munisamy work focus on the international trade and our targeted scope is the US, we need to find the alternative predictors for tariff, same language and participated international trade organizations.

Food flows between counties in the United States: Xiaowen Lin et al. developed a novel methodology to estimate food transfers between counties in the United States. They exploited the Freight Analysis Framework (FAF) dataset in their analysis to downscale the commodity transfers from FAF zones scale to county scale. FAF data categorizes food commodities into SCTG groups with two-digit codes and in this study, they have studied focused on SCTG 01-07, which are related to food commodity. To achieve this, Lin et al developed a “Food Flow Model” which is a computational algorithm that integrates

machine learning, linear programming, network constraints, and mass balance. This model incorporates a gamma mixture hurdle model that uses supervised learning to develop a functional form of regression models at the FAF zone scale which is then utilized for calculating potential food transfer between counties. A key assumption made here is that network properties remain consistent across scales (Konar et al). The gamma mixture hurdle model is a two-part model where a) Hurdle model uses logistic regression to predict the presence or absence of a link, and b) Gamma mixture model estimates the mass of the estimated link in the previous part following the assumption that food flux distributions follow gamma distribution across scales (Konar et al). Finally, they use linear programming to minimize transportation distance of food flows between counties and maintain mass balance at FAF zone level to solve the flow system at the county scale. Lin et al. used Fourier amplitude sensitivity test (FAST) to determine the most influential variables. They found out that counties in California and the Great Lakes region have the highest outflow of commodities. Also, the network density of the county scale (0.016) is much less than the network density of FAF scale (0.675). To validate their results, they compared their results with a study performed by Smith et al. where they modeled county-scale corn flows. They used R2-squared values and simple matching coefficient (SMC) to compare the results. SMC values turned out to be 1 for potential links, addressing identical estimation of presence or absence of county flow links in both models. R2-squared values are highest for outflows with a value of 0.46. Through the global sensitivity and uncertainty analysis (GSUA), where they used FAST method, they found out the distance is the most influential variable. The paper is an extensively rigorous effort to understand the food supply chain dynamics at a much finer resolution. The lack of existing literature at such finer scale, with the only literature at county-scale for corn flows by Smith et al., was another hurdle which they surpassed using a data-driven framework by bringing supervised machine learning, mass balance, and linear programming together to predict food flows for all food commodities. In their article, they addressed the shortcomings of their models, which we will be addressing in our project. Firstly, the model estimate food flows only for 2012, which means that regression models are specific to each time period and cannot be compared across different periods. Second, the model does not capture the non-linearity between environmental variables and food flow that can be captured using deep learning. Also, they used the great-circle distance between counties, which is also used in the gravity model of international trade, which is a simplification of transport pathways. This can also be addressed by using the roadway network for which we have data from FAF. In our project, we will be focusing on these three key shortcomings to further research in this field.

Method

To estimate the bilateral food flow between state pairs, we employ Random forest and Neural Network model. A gravity model is used as baseline model.

Data sources

We used diverse data sources to obtain data for major years and those years are 1997, 2002, 2007, 2012 and 2017. Those data sources are described briefly below.

Bilateral Trade: We obtain data on agricultural and food commodity transfers between FAF zones in the United States. The Freight Analysis Framework Version 4 (FAF4) database provides empirical agricultural and food commodity transfers between FAF zones for the year 2012□(Oak Ridge National Laboratory 2015). Second, we obtain statistical information on economic production within each US county.

Production: We used United States Department of Agriculture (USDA) data to determine state level production values for 10 selected food and crop categories in accordance with SCTG categorization,

namely corn, dairy, honey, milk, oats, rice, rye, wheat, aquaculture and sorghum. The value is measured in tons.

Distance: To calculate the distance between states, we used the latitude and longitude data and used the haversine formula.

Income and GDP: We obtained economic data, i.e. income and GDP data from Bureau of Economic Analysis portal. For income we used income per capita values. All the values are measured in million dollars.

Data Cleaning

EDA

The EDA is divided into two sections. First is Data Cleaning where we go through various key datasets and tidy them up so that they all can work together. In the second section, we will be looking at their distributions and their possible correlation between each other. To begin with, we load the data for food flow. The data has coded features which has relevant meaning. They are described below. fr: Foreign. Trade across borders. We won't be considering this data for our analysis as we are focussed on food flows among US states.

orig: Origin

dms: Domestic

dest: Destination

st: State

inmode: mode of import of international flows

outmode: mode of export of international flows

sctg2: Food category. This is a standard code used by Bureau of Trade Statistics for classifying food [BTS Guide]

value: value of food in dollars. tons: food value in weight.

After introducing the flow data, we utilized the latitude and longitude of states to calculate the haversine distance between origin and destination of food flows. Next, we introduced the remaining data from their respective files.

Data Preprocessing

In the exploratory data analysis, we manipulated the data to find appropriate features for the model. There are 39 features in total. In this analysis, we removed self-loops in the food flow network, i.e. the origin and destination of the domestic flows being the same. Before moving towards building the neural network, we formatted the data type of different features to meet the model requirements (i.e. categorical or numerical). To obtain categorical features, we create a function that takes the string value of the column and return the whole data-frame with a new one-hot encoded feature for the feature fed to the function. We also carefully removed some null values in many features to not lose any valuable data values. This was done by removing the features with missing percentage greater

than 30%. To spot any collinearity and anomalies that might affect our results, we produced the correlation matrix and histogram of all variables (refer figures). Most of the crop production values are right tailed which means that we need to normalize the data before we feed it to our model. Also, we notice that the “value” feature is highly concentrated at one bin. The standard deviation for the dataset is very high and the quantile values are quite low, which suggests that we need to remove the major outliers from the dataset.

Model

Neural network

In this analysis, we used neural network to predict the food flow between US states. We used two hidden layers with 30 and 12 neurons respectively with ReLu activation function. We used Adam optimizer function with a learning rate of 0.001, “mean squared error” loss function and RMSE as the metric. Batch size of 100 was considered with a validation split of 0.2 and number of epochs equals to 25. The results of the model run are discussed in the results and discussion section.

Random Forest

First of all, the FAF data needed to be expanded, because it does not include the case of no transaction, we will add the case of zero. To meet Kaggle’s memory limitations, whether the products are imported, exported or produced domestically is ignored in RF model. We group the bilateral flow data by origination state, destination state, commodity type, transportation mode and year. As shown in Figure 1, the dotted-line is the no transaction situation, the flow from S2 to S1 in simplified figure is the sum of the flow export from foreign countries (transshipment via S2) and domestic flow produced in S2.

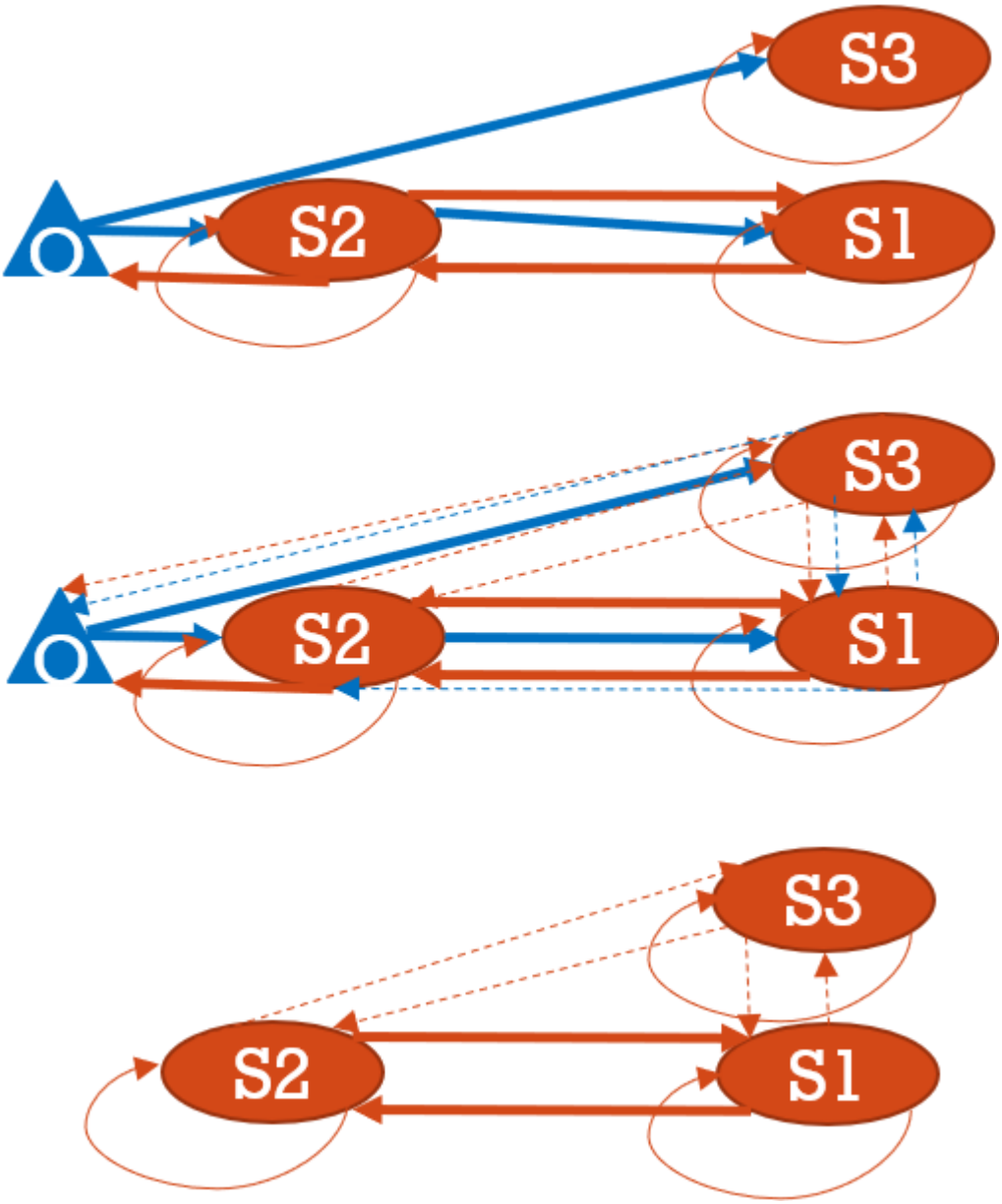

Figure 1: Problem definition: (up) FAF data, (middle) real situation, (down), simplified for random forest.

Random forest (RF) is a tree-based algorithm. The RF algorithm generates several independent trees through bootstrapping. Each tree randomly select predictor variables. The final output is determined by the average results of all trees. There are 637245 examples in the datasets 7(categories)x51(importers)x51(exporters)x5(years)x7(transportation modes) and 5 categorical variables ('origst' , 'destst', 'mode', 'sctg2', 'year') and 27 numeric variables. We remove the variables with missing rate larger than 25% like rice production, rye production and sorghum production. Because we want to predict the trade "value", we need to find which variables are relevant to the trade. Meanwhile, for those variables have high correlation, we need to carefully consider it to avoid the multimillionaires problems. Here are some findings for correlations. The trade weight is inversely proportional to the distance between the two states and their total GDP, which conforms to the gravity model of transactions. Here are some observations from the histgram: firstly all the production values follows a edge peak distribution. We might need to normalize it when we build the model. The value is very similar for different transcations. There are few extreme large numbers. After deleting the variables with high missing ratio, we still have several nans for production variables in the datasets. There are two different possible explanations (1) This kind of product is not grown or produced there (2)Information is not disclosed for some reason. We cannot figure out the specific reason. So we assume that all the nans are due to reason(1) then we set all nans as zeros. Also in previous section, we have already dropped some columns with large missing data which decrease the bias introduced by set all nans as zeros. The nominal categorical variables in this Food flow datasat

are transformed to a format that works better with the regression algorithms using one hot encoding method. Random Forests are good at dealing with outliers and different scale features but we also need to fit a linear model as baseline. In order to prevent close to 0 weights in base model due to different scales of features, we standardize the features to center the feature columns at mean 0 with standard deviation 1. After preparing the data for Random Forest model, we decide the hyperparameters for the RF model by applying the brute force grid search.

Baseline Model

To validate our model, we also try the gravity model. The linear model like Ridge model is not suitable for the bilateral food flow estimation. Gravity model is one of the most robust empirical models to illustrate the drivers in international trade: bilateral trade between two countries is proportional to size, mostly measured in GDP and inversely proportional to "distance" between them, which commonly fitted through Poisson Pseudo Maximum Likelihood (PPML) method to deal with the zeros. We use the python GLE library to do Poisson Psedo Maximum Likelihood regression. The gravity model calculate the $origst \times year$ and $destst \times year$ fixed effect and the impact of variables used in RF model.

Result

Neural Network

As we saw in the last section, the selected value for the number of layers and neurons is 2, and 30 and 12 respectively. These specific values were selected by trying out multiple different layers and neurons combination to find the most efficient structure. The learning rate was kept at 0.001 with a batch size of 100. We find out that the R^2 score for the model came out to be 0.43. The figure below shows the scatter plot of the real test values vs the prediction.

Figure 2: Neuron Network Result.

Random Forest

Through GridSearching, the hyperparameters we selected was 20 for max depth of tree and 50 for number of independent tree. Figure 3 shows the estimation results. The validation r^2 is 0.86. RF model shows the better performance for the large flow.

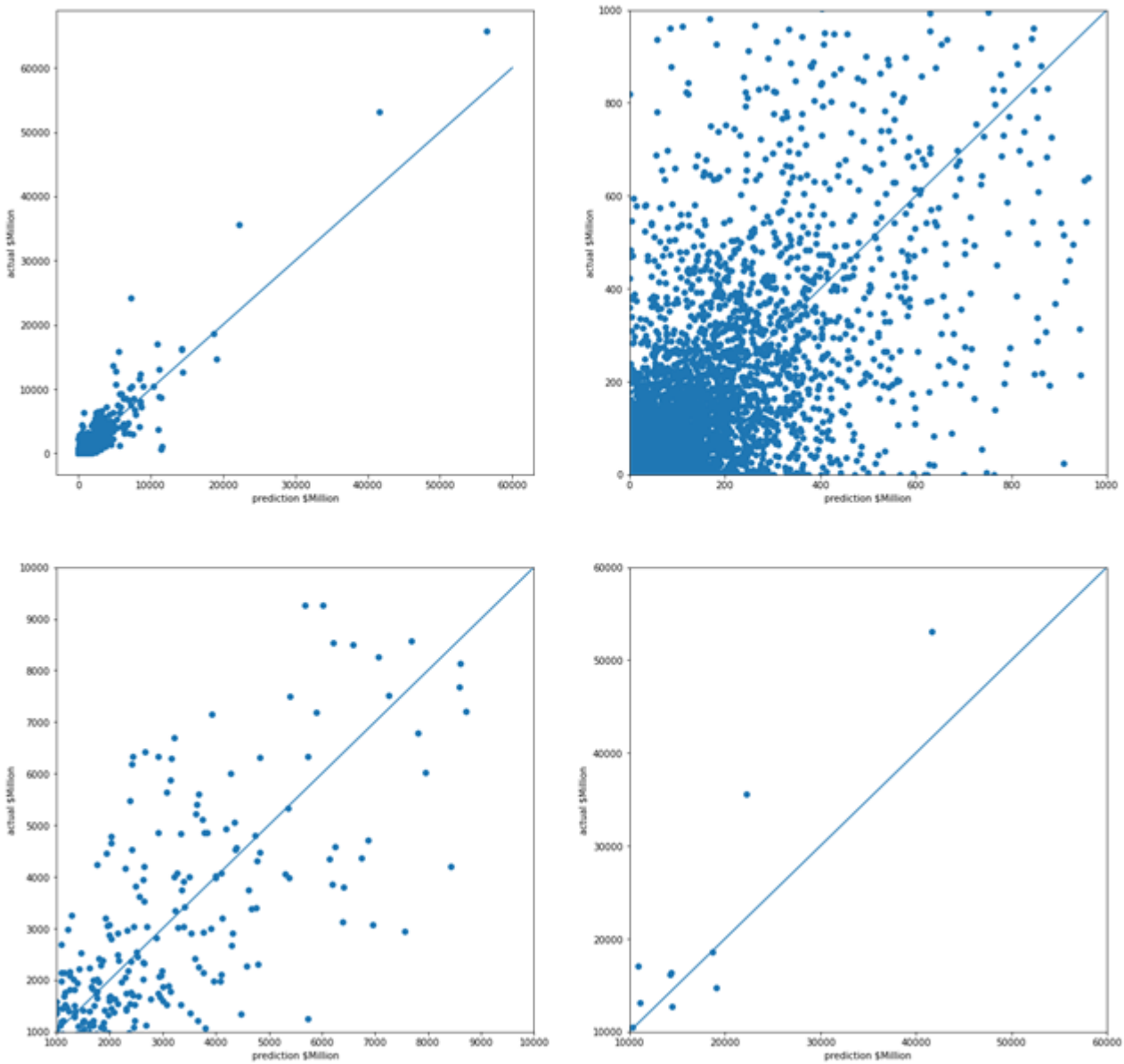

Figure 3: Random Forest Prediction Result.

Figure 4 shows the importance of features on this bilateral food flow regression problem.

The red bars are the importance of the forest with inter-trees variability.
As expected, commodity type, gdp of importors and gdp of exporters are informative. Surprisingly, distance and year shows less impact on the regressiom results. The performance of the distance variable is inconsistent with the law of the gravity model. We consider that this may be becaues we use the log of harversine distance.

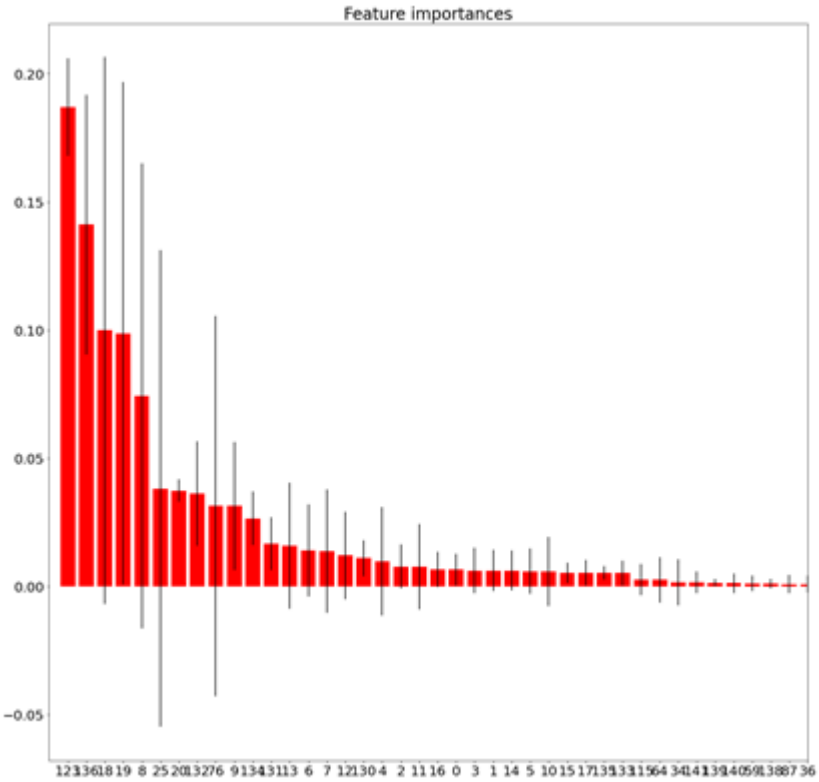

Figure 4: Random Forest Feature Importance.features 0:21('value','income_dms_origst', 'income_dms_deststt', 'animaltotal_dms_origst', 'animaltotal_dms_deststt', 'barley_dms_origst', 'barley_dms_deststt', 'corn_dms_origst', 'corn_dms_deststt', 'crop_total_dms_origst', 'crop_total_dms_deststt', 'honey_dms_origst', 'honey_dms_deststt', 'milk_dms_origst', 'milk_dms_deststt', 'oats_dms_origst', 'oats_dms_deststt', 'wheat_dms_origst', 'wheat_dms_deststt', 'gdp_dms_origst', 'gdp_dms_deststt', 'distance'),22:72 (dms_origst),73:123 (dms_deststt),124:130 (transportation mode), 131:137 (commodity type), 138:142 (year)

Baseline

Using the same data as RF model mentioned above, the validation accuracy r^2 of Ridge model is 0.049 because there are so many zeros in the dataset. We fitted gravity model to bilateral food flow data with PPML estimator, as shown in Figure 5. As we specified the commodity type and transportation mode, there is no group variance. But the average r^2 for different commodity and transportation condition is about 0.25.

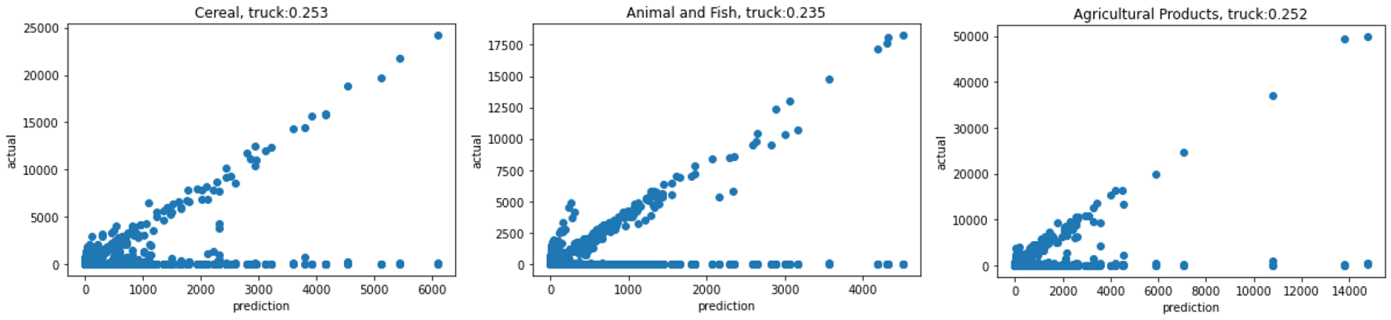

Figure 5: Gravity model for specific commodity and transportation mode.

Discussion

Comparing the models

We compared the different model results. Among three models we employed, Random Forest showed the best prediction performance. We assume that the good performance of RF is due to (1) RF is good

at deal with noise (2) RF can capture the nonlinear relationship, Also, NN and RF models both showed better prediction performance for larger trade flow. We thinks that small trade flow might be impacted on more complex variables, which cannot be predicted by the

Limitations

There are several limitations in this study. Firstly, the distance used in this study is haversine distance. Also we didn't adjust the food flow value using GDP deflator to eliminate the impact of inflation. In other words, we don't consider the transportation network connected each state. We don't consider the commodities trade mode(Domestic, imports or exports) in Random Forest model. Meanwhile, we cannot predict zero value weight flow with Neural Network model.

References

[**???**]doi:10.3386/w27151

[**???**]doi:10.1088/1748-9326/ab29ae

[**???**]http://dx.doi.org/10.17016/IFDP.2020.1296

[**???**]https://arxiv.org/abs/1910.03112

[**???**]https://doi.org/10.1371/journal.pone.0199498

[**???**]https://doi.org/10.1073/pnas.1703793114