TREND ANALYSIS OF THE U.S. IMPORTS FOR CONSUMPTION OF STEEL PRODUCTS

Xiaoxi Liu, Xiaotong Ding, Liya Li, Chengchen Luo
STA141B Final Project
Department of Statistics
University of California, Davis
1 Shields Ave, Davis, CA 95616
{xixliu, xtding, lyali, dorluo}@ucdavis.edu

1.Introduction

The United States is the world's largest steel importer. U.S. steelmakers don't produce enough steel to meet domestic demand. Imports fill more than a fifth of the nation's steel supply. In 2018, president Trump's administration's new tariffs on steel have made steel more expensive in the U.S. than almost anywhere in the world. If we know the import value and quantities of the past years, we can see whether the new tariff policy exposes huge influence on the steel markets.

In this project, our task is to explore trend of the steel imports from year 2013 to 2018 with respect to import steel's price and quantities and make prediction on the unit price (in thousand dollars/Metric Tons) for future months. (i.e. make projection on our data until September, October, November and December 2018). Our project consists of 5 parts: *Data exploration, Hypothesis tests, Classification, Model selection and Prediction*. This report is used to demonstrate our project.

2. Dataset Analysis

2.1 Data Source

We use the data on *U.S. Census Bureau*, a principal agency of the U.S. Federal Statistical System, responsible for producing data.

We use year-to-date and month-to-date datasets from *year 2013 to 2018 on*

steel imports.

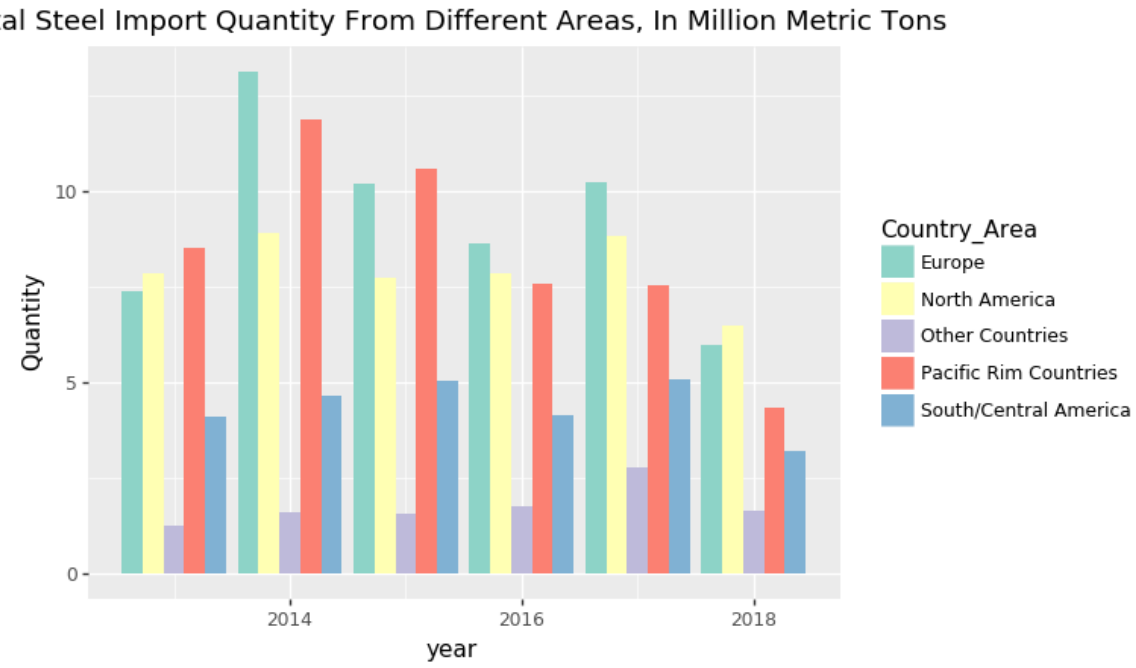
We save the data from the website into *two csv files, p.csv and c.csv*, p.csv is the steel import quantity and value with respect to different countries and c.csv is the steel import is the steel import with respect to product type.

Data Munging: See code in "FinalCode.ipynb".

2.2 Exploratory Analysis and Visualization

In this part visualization was used to help us understanding data and exploring possible hidden factors to analysis from our steel import datasets.

2.2.1 Where do steel products come from? - Area of the world and import



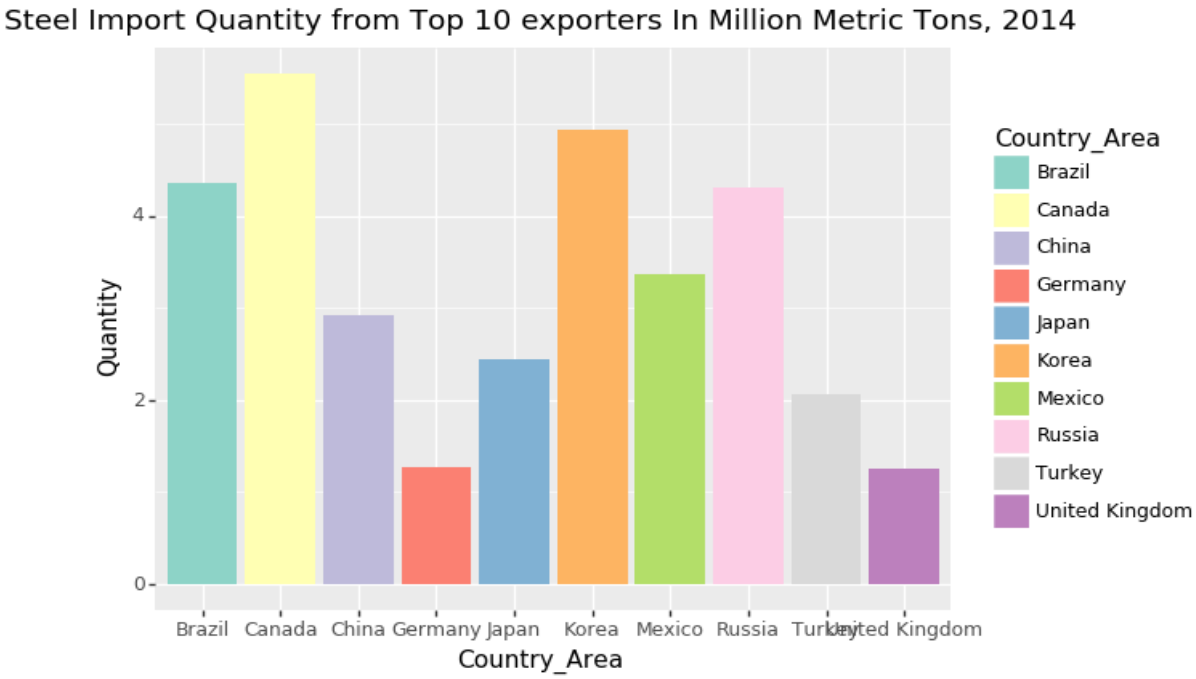
From the plot it can be seen that in terms of large areas, Europe has been one of the largest steel product exporting areas, along with Pacific Rim Countries (Asia).

2.2.2 Which countries are the top 10 exporters?

In terms of countries and territories, the top 10 exporters were visualized with respect to quantity from 2013-2018, respectively. The results are as follows, listed in *descending* order:

Years	Top 10 steel import countries
2013	Canada, Brazil, Korea, Mexico, Japan, China, Russia, Turkey, German, Taiwan
2014	Canada, Korea, Russia, Brazil, Mexico, Japan, China, Turkey, United Kingdom, German
2015	Canada, Brazil, Korea, Turkey，Mexico, Japan, China, Russia, German, Taiwan
2016	Canada, Brazil, Korea, Turkey，Mexico, Japan, Russia, German, Taiwan, Other
2017	Canada, Brazil, Korea, Turkey，Mexico, Japan, China, Russia, German, Taiwan
2018	Canada, Brazil, Mexico, Korea, Turkey, Japan, China, Russia, German, Taiwan

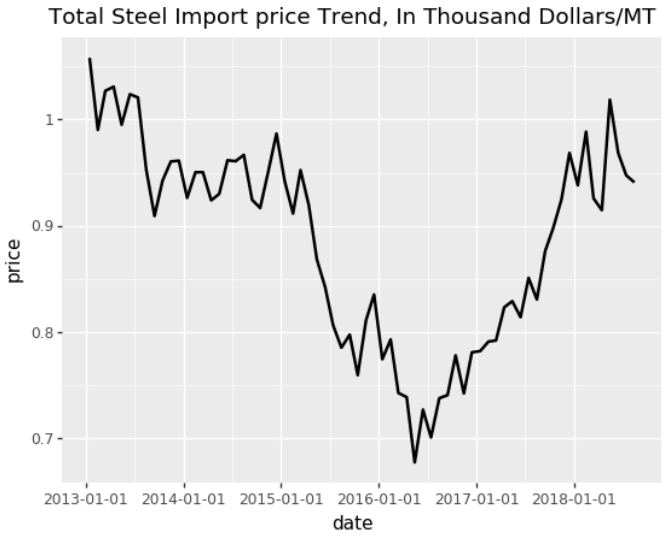
From the chart it can be seen that ***Canada*** has been the largest exporter for the last 5 years, followed by Brazil, South Korea, Mexico, etc. United Kingdom was replaced by Taiwan in 2015, and China was replaced by Other countries in 2016.

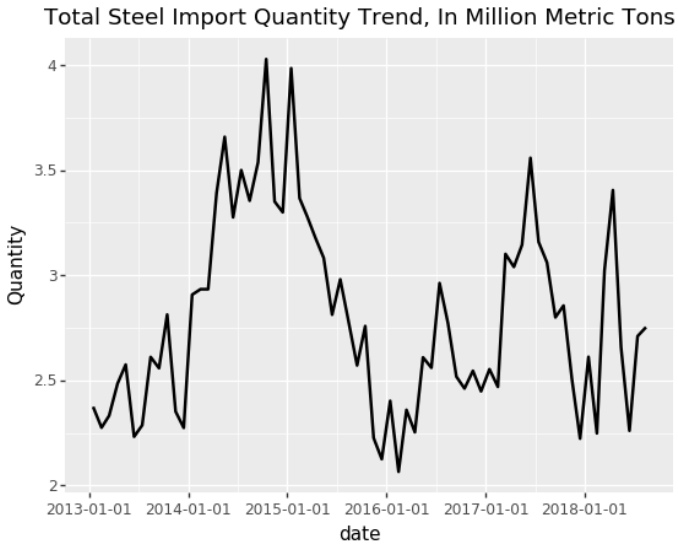


Example image: top 10 exporters, 2014. See full set of images in "FinalCode.ipynb".

2.2.3 Quantity trend and price trend

We draw two time series plots to get a big picture of the import quantity and price changes over year 2013 to 2018. The trends are as follows:





From plots above, both plots have a striking drop in year of 2015 and reached it lowest point in the beginning of year 2016; the import quantity is very high in the year of 2014; however, the price stays relatively stable.

Possible explanation: Since United States is the world largest steel importer, the high demand of steel and stable price in 2014 triggered the chain reaction which led to the sharp drop of import quantity and price in 2015. The significant decrease of the top ten import sources of steel of United States attributed to the decline of the global steel price.

After some research on Google we conjecture this situation happened because of the long going problem of excesses steel produced around the world, which is also the reason why the price was declining before year 2015. After this “winter” season of the entire steel industry, price slowly went up in the next few years.

Also, from the trend plot of steel price and quantity, we observe that when price increases the import quantity increases also, so we guess that there may exists a rough positive between the import price and import quantity. Then, we start investigating this possible relationship as below.

2.2.4 Relationship between Steel import Quantity and Value

Firstly, we draw a scatter plot to get a general idea of the relationship between quantity and value.

From the plot it can be seen that the relationship between quantity and value is almost linear, and it seems that the year may affect this relationship, which indicates that '*year*' and '*quantity*' may be factors to consider when we predict the price. Thus, hypothesis tests would be conducted in next section.

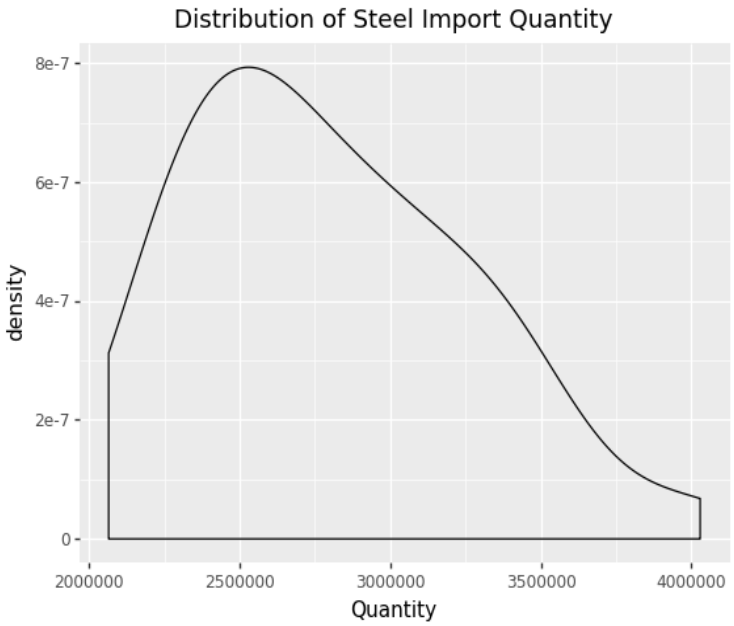
3. Statistical Hypothesis Testing

In March 2018, Trump-administration imposed tariffs on steel (25%) from most countries. Since we try to predict the unit price after this policy was enacted, hypothesis testing is needed here to test whether this policy affects the import steel products' quantity and price or not. We perform permutation test and Kruskal-Wallis test in this section.

3.1 Does import quantity decrease since the tariff has been imposed?

In this part, we try to perform appropriate hypothesis test on price and quantity from the c.csv dataset to explore whether the import quantity decreases due to the increasing tariff.

First, we visualize the distribution of the import quantity and price to choose the appropriate hypothesis testing methods we can use.



From the plot, we see that the quantity data points are not normally distributed**.** For here, we have a situation in which none of the standard mathematical statistical approximations apply. We have computed a summary statistic, such as the difference in mean, but do not have a useful approximation, such as that provided by the CLT. Thus, we choose to use permutation test here.

Permutation Test:

H0: $\mu 1 = \mu 2$ $\mu 1$: mean calculated from quantity before 3/2018

Ha: $\mu 1 > \mu 2$ $\mu 2$: mean calculated from quantity after 3/2018

Test procedure: We reject the null if p-value is less than $\alpha = 0.05$.

At $\alpha = 0.05$**,** we perform the permutation test and repeat the test procedure for 100000 times.

We then get the following results. (For test functions: See permutation test function and simulation in “FinalCode.ipynb”, part 2.1.)

Test statistic	51372

P-value	0.51372

Test Result: Since the p-value is greater than significance level, we cannot reject H0 at α = 0.05. Which indicates that there is no significant evidence to claim that the tariff in March 2018 made the quantity decrease.

3.2 Does mean price increase since the 2018, march tariff has been imposed?

Same reason as above, we choose to use permutation test here.

Permutation Test

H0: μ1= μ2 μ1: mean calculated from price before 3/2018

Ha: μ1 < μ2 μ2: mean calculated from price after 3/2018

At α = 0.05**,** We get the following results.

Test statistic	3046
P-value	0.03046

Test Result: Since the p-value is smaller than significance level, we reject H0 at α = 0.05. Therefore, we claim that the tariff in March 2018 made the steel price increase.

Conclusion: From the 2 permutation tests we can see that the there is no significant evidence to claim that new tariffs affect the total import quantity, but unit price did increase. This result indicates that, under new tariffs policy, it's unclear for us to see how much imports will fall. Therefore, manufacturers might

have to keep importing these materials and pay the tariffs. The fact increases their costs, making their products less profitable, or if they raise their prices, less competitive. Experts say: “it's not clear that the existing American industry can supply all the steel and aluminum now imported from abroad, since restarting closed factories will be difficult, expensive and could take years in some cases.” Thus, we need to consider year (2018) when we try to predict the future price.

3.3 Do exporting areas affect price?

For this part, we choose to use *Kruskal-Wallis test*, since we need to compare five independent areas of same sample sizes here.

Define the following:

Area	Europe	Asia	South/Central America	North America	Other countries
Mean Import Price	μ1	μ2	μ3	μ4	μ5

H0: μ1 = μ2= μ3 =μ4 = μ5
Ha: otherwise

At α = 0.05 we get the results as follows:

Statistics	167.0228055450827
P-Value	4.553799555189723e-35

Test Result: Since p-value is extremely small (<<0.05), we reject H0, which indicates that different exporting areas have different prices, and we need to consider that when we predict the price.

4. Classification

4.1 Why choose to use classification?

From previous sections we assume that different exporting area is a factor in predicting the future price. Plot of quantity and price from different areas is as

follows:

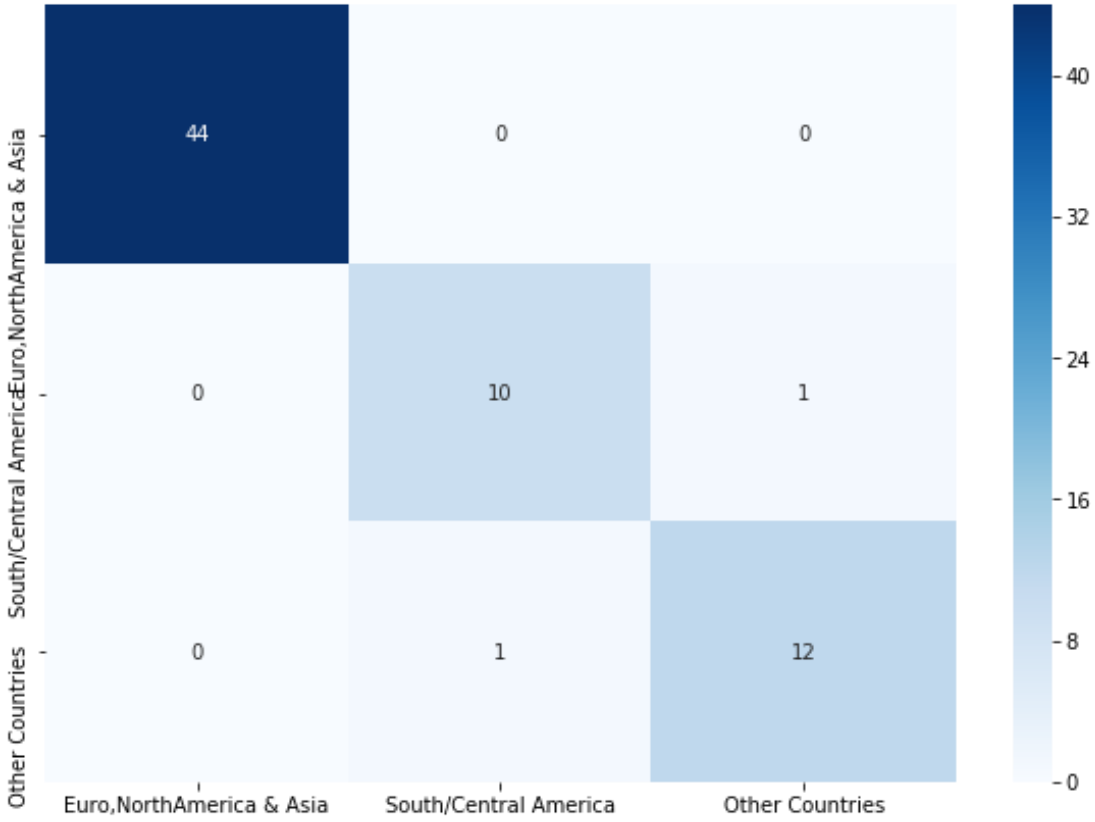
It is clear that the quantity and price distribution is separated by 3 larger area: *Europe-North America-Asia (ENP), South/Central America (SCA), and Other Countries (OC)*. Knowing this, we want to classify area given a quantity, and we may be able to determine the price from here.

4.2 Classifier selection and confusion matrix

Our goal is to use features *('Quantity', 'Value', 'year', 'month')* to classify the area. We prepared 10 classification methods as candidates and used *10-fold cross validation* to evaluate the models. The result is as follows: (See detailed functions and packages in "FinalCode.ipynb", part 3.2.)

Classification Methods	**Training Set Accuracy**	**Testing Set Accuracy**	**Accuracy evaluated by 10-fold cross validation**
Logistic Regression	0.9963235294117647	0.9705882352941176	0.9911764705882353
LDA: Linear discriminant analysis	0.9742647058823529	0.9558823529411765	0.95
SVM: Support vector machine	0.9963235294117647	0.9705882352941176	0.9911764705882353
Decision Tree	1.0	0.9705882352941176	0.9588235294117649
K-nearest Neighbor	0.9375	0.9558823529411765	0.9588235294117646
Random Forest	1.0	0.9852941176470589	0.9705882352941178
Adaboost	0.8786764705882353	0.8676470588235294	0.8235294117647058
Quadratic Discriminative Analysis	0.9889705882352942	0.9852941176470589	0.9617647058823531
Boosting Gradient Descend	0.9926470588235249	0.9852941176470589	0.9647058823529411

From this result, we choose support vector machine (SVM) to be our classifier, and its confusion matrix is plotted as a heatmap:



From this plot we can see that the performance of SVM is satisfying. However, since the dataset is quite small, this level of accuracy is doubtful and worth future exploration.

5. Model Selection

5.1 Outlier Detection and Normalization

Based on previous sections, we decided our explanatory variables are: *'Quantity','year','month','Grouping_code' (transformed from area)*;
Our response variable is unit price. We used *Support vector machine (SVM), Local outlier factor (LOF)* and *Isolation Forest (IF)* to detect outliers and removed outliers that were 0.294% of the data. See details in “FinalCode.ipynb”, part 4.1.

Then *normalization* is applied to the dataset for model selection. See details in “FinalCode.ipynb”, part 4.2.

5.2 Model Selection

Our goal is to predict the price by *'Quantity', 'year', 'month' and 'Grouping_code'.* We chose 2 models as candidates: linear regression model and polynomial regression model.

Using validation size of 0.2, we examined MSE, the average squared difference between the estimated values and what is estimated, of both training set and testing set, and chose the model with smaller MSE. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. The result is as follows:

	Linear Regression	Polynomial Regression
training set MSE	0.05221174070167548	0.047934651472849116
testing set MSE	0.08042093581758672	0.0656832222610577

Test Result: Since the MSE of PR model is approximately 0.065 < 0.08, MSE of LR. The smaller value means error terms are smaller. Thus, we chose Polynomial Regression model as our predicting model.

6. Price Prediction

There are 2 steps in price prediction:

Step 1: Use SVM('Quantity','Value','year','month') to classify Grouping code.

Step 2: Use Polynomial('Quantity','year','month','Grouping_code') to predict the price.

Among these variables, ‘year’ = {2018}, ‘month’ = {9,10,11,12}, ‘Quantity’ = mean quantity of the same month over the past years, ‘Value’ = mean value of the same month over the past years.
After fitting the models (See computation details and code in “FinalCode.ipynb”, part 5), the result is as follows:

Year	Month	Predicted Price (thousand dollars/MT)
2018	9	1.04356478
2018	10	1.050353
2018	11	1.08887015
2018	12	1.11659354

According to this prediction, in the last 4 months of 2018, the unit price of steel import will slightly increase. As we assumed, the newly imposed tariff may be part of the reason, and our prediction explicitly proved the correctness of our assumption. This projection on the price also reflects negative influence on many industries or markets depend on import steel products like building constructions and car production. The increase of the import price will increase their costs, making their products less profitable, or if they raise their prices, less competitive. However, the further results need to be reexamined when data is available.

Conclusion:

In this project we explored trend of the steel imports from year 2013 to 2018 and made prediction on the unit price (in thousand dollars/Metric Tons) for future months.
This report is organized as follows:

- Section 1 is the background introduction.

- Section 2 is about *basic explorations and visualizations* on the dataset we use to do trend analysis.

- Section 3 shows *statistical hypothesis testing*, permutation test and Kruskal-Wallis test, we used to test whether new tariffs on steel has negative or positive influence on the steel import quantity or price.

- Section 4 compares accuracy among 10 different classification methods we use on our data to *find out the best classifier*.

- Section 5 compares the MSE between linear regression model and polynomial regression model to *find the best fitted model*.

- Section 6 uses the best classifier and the model we selected to *make prediction* on unit price (in thousand dollars/Metric Tons) for September, October, November and December 2018.

Data source:
https://www.census.gov/foreign-trade/Press-Release/steel_index.html