Analyzing US Food Import Contents

Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586
apd6628@cs.rit.edu

Abstract—Food in today’s world travels a thousand miles before actually reaching its customers. The import and export business has now become a very competitive one. There has been a tremendous growth in the U.S. food imports, as they roughly spent about three billion dollars in the first eight months of 2017. There are many reasons for countries to import food items, the main reasons for this being the citizens’ demands of wanting a variety of food items overseas, globalization and an elevation in technology. The sentiments of Robert Kiyosaki echoes with imports, when the supply exceeds the demand, there is a drop in the commodity’s worth as well as its financial value. Consumers need a variety of food items from overseas and this is known as citizen demands. Due to decreased trade barriers, importing food items is easier. In this project, different analysis are implemented to analyze the US Food import content.

Index Terms—import and export; data-mining; data science; knowledge discovery in database; globalization; ARIMA; machine learning; time series; correlation; inflation; accuracy;

I. INTRODUCTION

The inability of humans to construe and summarize huge amounts of data, and use them for decision-making purposes, has resulted in usage of new visualization tools for analysis of automated and intelligent systems[1]. The controlled nature of knowledge discovery in databases (KDD) allows people to scrutinize data from different perspectives and categorize it accordingly.

The techniques of data science, machine learning, and data mining can be used in many domains. Some key applications where data mining is used extensively are in recommendations systems, prediction of retail sales, category management, and import-export data. The main purpose of the analysis is for accurate prediction of users data. Thus, if the accuracy of the prediction increases, then we can achieve optimum food items, with minimum import cost and thus reduce the problem of over-stocking and under-stocking.

This brings me to the one domain where I am using data mining and its techniques for analysis purpose of US Food import contents. This topic brings up an important question “Why does the US, a country with an abundance of resources and exceptional diverse economy, need to import food from other countries?”[2]. No country is self-sufficient in spite of it extending over multiple climatic zones, and having enormous natural resources and workforce as well as they ability to produce coal, copper, iron and other natural resources. When it comes to food and consumable items the United States has to import it from foreign lands. The import generally consists of around 60% of fresh fruits and 80% of seafood products, thus resulting in an increase in the productivity with

People now can taste food from different parts of the world without actually going there. Despite these positive trends, food importation have more downsides. Importing of food items is harmful to local market and farmer since these local markets are not as competitive in terms of cost as compared to big and famous international brands.

The food quality tends to be compromised during the import process since a lot of preservatives are added in order to keep the food fresh, healthy and long lasting, since it has to go a great distance before it reaches the US. This may not be very healthy for people.

Apart from this importing food is not good for the environment too. While trains, planes and other different modes of transportation are supplying food from one continent to another the carbon-di-oxide and other such gases take a toll on the environment. Another factor which adds to the environment is the packaging of imported food items which is generally more than regular items. Approximately 25% of the people do not recycle items. Therefore, once it is thrown away, it reaches the water bodies thus polluting it[3].

Due to the reasons mentioned above, I am proposing to analyze food products which are imported by the United States from the year 1999-2014 and check if this import cost can be reduced by analyzing all the perspectives. This will in turn reduce the cost which is involved in importing that crop/fruit. Thus, providing an added value to the agricultural business. Doing this could be advantageous since it would increase labor in the US as well as save on billions of dollars which are invested in food import. In section two we see the background of this project and some relevant work done in this field. Followed by section three which highlights the methodology and each step which brought us closer to our results. Section four covers the different results and analysis that has been implemented in the paper. In section five we conclude the paper and talk about our findings and different things learned during the paper. The final section which is section six highlights somethings that can be done in future in order to enhance it and make it even more robust.

II. BACKGROUND

Agriculture has been the oldest form of subsistence and hence import and export play a crucial role in expanding the agricultural productivity. If the export business of a particular country is high it indicates a greater demand for that countries products, thus resulting in an increase in the productivity with
respect to the agricultural products. Import of the agricultural goods also follows the similar outline. Export and import are two sides of the same coin[3].

In spite of agriculture being such an important topic, it has not investigated and analyzed to its fullest potential. There are few papers which have been published in this domain most of which are related to India. Since agriculture is the backbone of that country in particular.

In paper [4] the authors talk about using natural logarithms since it is easy to convert the variables that have a wide range of variance. They also go on to talk about maintaining and plot the data using the time series approach. This is done to check the trend and seasonality in the data.

To achieve a perfect balance between import and export such that there is no over-stocking or under-stocking of food various methods were implemented [5]. The most common being the autoregressive integrated moving average (ARIMA) technique. This algorithm is generally used for Time Series Analysis. As stated by Bandyopadhyay, et al. [6] ARIMA is not capable of handling small variation in the data. If there happens to be a sudden change, then the model would not be able to cope with it thus failing to capture the change. This would lead to the model giving us low accuracy. Another disadvantage of this is the data in the model is considered to be in linear form but in reality, it is not always true. As G. Zhang mentioned in his paper [5] the major setback of using ARIMA is the preconceived notion that the data in the model will be in the linear form. A linear associative structure is assumed between the time series variables; Hence, one could conclude that no nonlinear patterns can be captured by using this model.

III. METHODOLOGY

Fig 1 gives a gist of my model structure, which speaks about the input, algorithm used and its output.

A. Dataset

The data for this project has been obtained from Department of Agriculture which belongs to data.gov [7]. This dataset contains valuable information regarding the import values of different food products entering the U.S ports from different parts of the world. The import products range from crops to oilseeds and even different beverages. The dataset includes 15 years of data from the year 1999-2014. The US food import dataset consists of 18 different food categories along with their subcategories coming from 64 different countries. Therefore, we can run multiple algorithms and find out the hidden patterns in them. A snapshot of the dataset can be seen in Fig 2. Before performing any kind of analysis I wanted to see how correlated all the variables are with each other and their linkages. On using the edgebundle function() in R an image is created with all the its attributes along its perimeter. This gives us an idea regarding the linkage between the attached node. This can be seen in the image below. Fig 3 depicts the attribute bakery is mostly connected to all the attributes in the dataset therefore, showing good correlation. Similarly, the attribute year does not have any link with the rest of the attributes. Thus we can say the year does not have a correlation with other variables. In Fig 4 we see the values as well as the correlation matrix. Due to the difference in correlation values, the size of the circle as well as the intensity varies. Thus, we can say that the correlation coefficients are proportional to the intensity and the size. The closer the coefficient is to positive or negative one, the variables are more closely related. If it is close to zero, means there is no relationship between the variables.

B. Dimensionality Reduction

Due to the presence of so many attributes present in the dataset dimensionality reduction had to be performed. The 18 food groups along with their subgroups were categorized into perishable, semi-perishable and non-perishable food items. Similarly, the 64 different countries were grouped into their respective continents. Thus, the dataset now contains three different food types and the import amount from 1999-
2014 from each continent. The cleaned and reduced data can be seen in Fig 5.

![Fig. 4. Correlation of the variables](image)

![Fig. 5. Snapshot of the cleaned dataset](image)

**C. Algorithm implemented**

Time series is a statistical method that helps detect trends and patterns. We can categorize the data into the following three parts:

- **Time series data:** This data is taken at regular time intervals.
- **Cross-sectional data:** Collection of one or more variables data at the same point in time.
- **Pooled data:** This is a mixture of time series data and cross-sectional data [8].

Some important features of time series are:

1) In order to describe how past events affect the future ones.
2) To forecast succeeding values in the series.

It is believed that if the data points are observed carefully they reveal some internal structure like autocorrelation, trend or seasonal variation that one must pay attention to.

1) **Data Preprocessing:** In order to perform time series method, we only require the total amount and the date attribute. In my dataset, the total amount is the total import amount for a particular food group like perishable, semi-perishable and non-perishable. Similarly, the date attribute is the year. This data was used for predicting the total import amount for that particular food group for three years (i.e. till the year 2017). For the implementation, I have used R.

2) **Time Series object & Plot:** After processing the data according to time series requirements the next step was to store the data in a time series object. To do so, the ts () function in R was used. Using the dygraph, which is an interactive plot for time series data we have created a plot for year-wise import food for each continent as seen in Fig 6. On the similar lines, plots for non-perishable and semi-perishable were also created.

![Fig. 6. Year-wise perishable food import for each continent.](image)

![Fig. 7. Decomposition of perishable food data](image)

3) **Decomposition of data:** Decomposition of time series data is a mathematical procedure which converts the single time series into multiple time series. This contains the seasonality which checks if the pattern is recurring or not. It also tells us about the trend of the data which indicates whether the pattern is increasing or decreasing over a particular time span. The last graph which can be seen in decomposition is random noise and if present would not be categorized as seasonality or trend.

There are two ways of performing decomposition:

1) Additive: We use the additive method when the seasonal component has a particular pattern and is constant in spite of an increase in time series. It is calculated using the formula given below

\[
\text{Time series} = \text{seasonal} + \text{trend} + \text{random}
\]

2) Multiplicative: We use the multiplicative method when the seasonal component is directly proportional to the increase in time series. It is calculated using the formula given below

\[
\text{Time series} = \text{seasonal} \times \text{trend} \times \text{random}
\]

On decomposing the data the trend and seasonality were analyzed as seen in Fig 7. This was done by using the decompose() function in R.

4) **Applying model and forecast:** Exponential forecasting is a famous forecasting method. Weighted averages of previous results are predicted using exponential smoothing methods and as the results get older the value of the weights decrease exponentially[9]. Basically, it means that if the observation is newer its associated weight is larger. This method is advisable.
for a longer period and produces an accurate forecast. There are multiple types of exponential smoothing. The most popular one being the Holt-Winters method.

The Holt-Winters method is advisable for seasonal time series prediction. There are multiple techniques to implement this method. The different approaches provide different forecast results. In R, the HoltWinters() function shows the best-fitted model for the data. This method contains a forecast equation with three smoothing parameters: alpha which is for the level, beta for trend, and gamma for the seasonal component.

### Where:

\( \alpha \): The alpha component specifies how to smooth the level component. If the value approaches one then higher is its associated weight on the latest value. Similarly, if the values approach zero its associated weight on older values is heavy.

\( \beta \): The beta component specifies how to smooth the trend component. If the value approaches one then higher is its associated weight on the latest value. Similarly, if the values approach zero its associated weight on older values is heavy.

\( \gamma \): The gamma component specifies how to smooth the seasonality component. If the value approaches one then higher is its associated weight on the latest value. Similarly, if the values approach zero its associated weight on older values is heavy.

In the above case, the level value is very high (0.84), which gives an estimate of the local mean, or level of the data-generating process (DGP). The seasonality index is one, which indicates that there is seasonality present in the data.

From the obtained model, the next three years data was forecasted in a quarterly manner (i.e., data till 2017) using the predict() function in R as seen in Fig 8.

### D. Results

In Fig 9, we plot the results which the model predicted from the year 2015 to 2017 along with the original values (i.e., from 1999 to 2014). Therefore, Fig 10 is the import value for each of the continents. For validation purposes, I checked for the root mean square error (RMSE) and the mean absolute percentage error (MAPE). In order to do so, the accuracy function was used on the predicted data against the testing data.

The RMSE value is 725.468 while the MAPE value is so low, which indicates that this model is the best fit for this data.

#### IV. RESULTS & DISCUSSIONS

##### A. Inflation

There are two kinds of economy situations one which is open economy situation while the other is a closed economy. Inflation is a cause of excess demand when dealing with a closed economy situation while in an open economy case inflation is influenced by overseas trading [10]. According to Dexter et al [11], native inflation is directly affected due to the availability and cost of the imports, and is indirectly affected due to the competition with domestic items.

In Fig 12 we see the graph of inflation which has been calculated for each food item every year. On closely observing it we see that Live Farm Animal Import has the highest inflation value. This can be supported with the values seen in Fig 13.

##### B. Country Rankings for US Import

Before performing a ranking of all the countries who the US imports food from, I decided to first see the total import amount that the United States of America spends per year. This can be seen in Fig 14. As we can see a continuous growth in the graph apart from the year 2009 which on further analysis I realized was the time when the recession hit the US. As quoted in the Economic Policy Institute where the United States Census Bureau reported that "U.S. goods and services trade fell from an overall $695.9 billion in 2008 to $380.7 billion in 2009, a decline of $315.3 billion (45.3%)". The reason for this major fall in trade indicates reflects the consequences of recession and financial crisis on the demand.
Since my data was grouped into continents during the data cleaning phase, I wanted to see which is the continent that the United States imports maximum food products from. In Fig 15 we can see that United States import maximum of their food from North America followed by Asia. To further drill down as to which countries within the respective continents do the United States invest their money in, I performed the aggregate() function in R to sum the total import amount of each country. Once this was done the list was sorted in descending order using the order function(). The top 10 countries were chosen for convenience purposes. These are the countries which the United States invests a lot of their money in.

C. Beverage Analysis

Beverages in this dataset were a combination of alcoholic drinks such as malt beer, wine and different liquors and liqueurs as well as non-alcoholic drinks which included water and other liquids which can be imbibed liquids, unsweetened and sweetened drinks, such as soft drinks. As per the analysis, wine is imported maximum from France followed by Italy. France bags the first place with a total amount of $18419.234 followed by Italy which is not far away $18162.373. To confirm the accuracy of the analysis I did read up the top wine producing countries and the result from wine institute in the year 2014 can be seen in Fig 17. As we can see in Fig 17 France and Italy are the top two countries that have an almost equal amount. It is said that the top three wine producing countries, are responsible for manufacturing around half of the wine which is made in the world. With the amount of wine, they make one can fill around 5,127 Olympic sized swimming pools. For the value and quality purposes, it is a good practice to be informed regarding the most popular types of wines in the top producing countries[13].

Mexico has absorbed its skills from German immigrants in the 1800s. 16.5% of the world’s beer is produced by Mexico alone, which is more than any other nation. FEMSA and Grupo Modelo are the two main beer companies [14]. The second on my list is the United Kingdom but that ranks eight as per 2016 statistics. Due to soft drink enthusiasts as well as the competition in the market, there are a countless number of flavors which are quenching the thirst of these enthusiasts around the globe. There are a range of flavors ranging from bubble gum to teriyaki carbonated drinks. Having said this Switzerland tops the market when it comes to the production of soft drinks followed by Austria. Switzerland in 2016 exported the US $1.9 billion value worth of soft drinks(18% of the total exported soft drinks)[15]. The last category mentioned in the beverages was liquors and liqueurs. It was not a surprise to see European countries top this list. It was France which made it to the first position.
followed by United Kingdoms. Thus we can conclude that it is the European continent that dominates when it comes to the production of beverages and US invests maximum of their dollars in importing the beverages from the European continent.

In Fig 18 all the top countries who import their beverages to the United States have been highlighted on the map. To do this I used the googleVis library in R. Within this library the gvisGeoChart() function helped plot the countries on the map. As we can see it is mostly the European countries that the drinks come in from.

V. Conclusion

Time series analysis is popularly used, yet historical data does not always give the true picture of the current scenario. There are a lot of factors that affect the import, like cost, quality, and recession. The 2009 recession period did hit the trade in America greatly. On doing the inflation analysis the output/results should be used constructively so that they can be used for the preparation of anti inflationary policies while taking in considering the import and export decisions. By doing so one would definitely build and stabilize the countries economic growth.

VI. Future Work

This is an interesting domain to explore and since it hasn’t been dug into too much a lot of work can be done. Future scope of this project includes getting the soil and weather data of the places where crops are imported from can be helpful since we can check for that perfect combination of soil and weather in the United States and thus grow it locally rather than importing it. Also, we can use heat maps and perform image convolutions to get more detail regarding the weather soil and growth of a crop.

Acknowledgment

This project would have not been a success without the timely help and feedback from my capstone advisor Dr. Carol Romanowski and instructor Dr. Joe Giegel. I am also grateful to my friends and family who kept me motivated time and again. And last but definitely not the least a big thank you to the RIT CS Department, because of which, I will be receiving my Masters degree.

References