Springboard Data Science Capstone Project 2
Final Report

Current and future diversity in global agricultural production and trade networks

Shuangshuang Liu
May, 2020

Contents

1. Introduction

Problem statement: To produce more, or to trade more?
Food security is becoming a pressing issue especially in some developing countries in tropical areas, due to climate change and a fast-growing population that may affect both crop production and demand. Therefore, it is important to understand how food availability may change through domestic production and international trade under future climate scenarios.

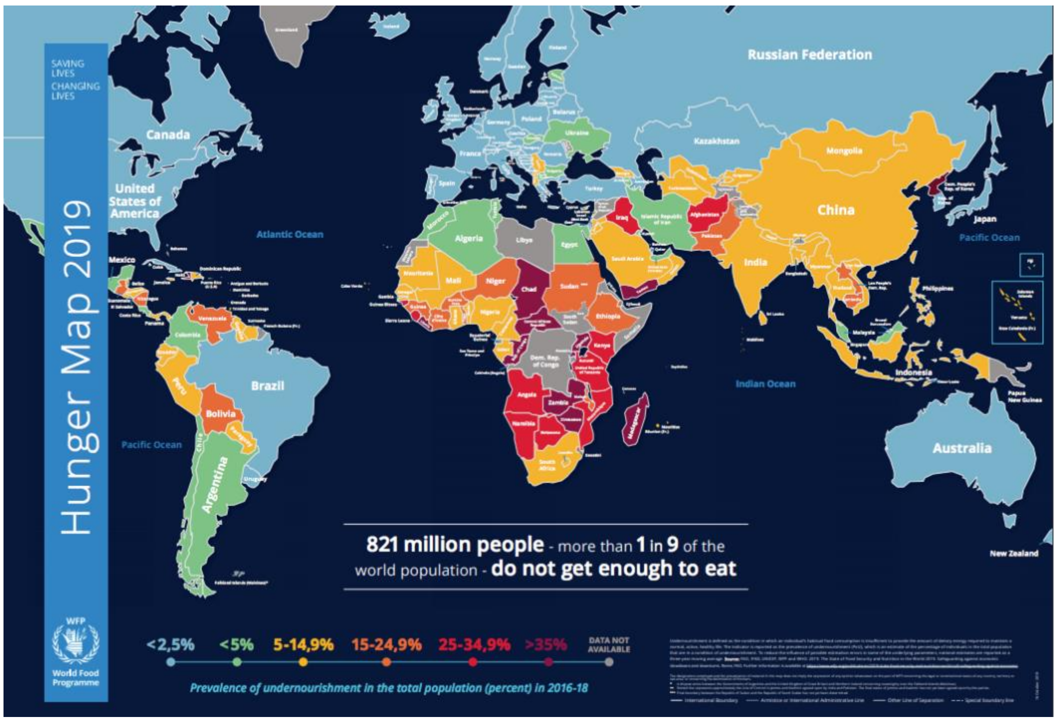

Fig. 1.1 World hunger map (Hunger map 2019. World Food Programme)

Strategies that optimize the balance between local food production and food trade will be essential for resolving food security and promote economic development. Thus an important question for relevant organizations across all levels to consider is: To produce more, or to trade more? This question has been stressed similarly by FAO: "Should the focus of government intervention be primarily on increasing local food production, or should it lean more towards increasing access to food and stimulating rural development in general?" (High-level expert forum, FAO 2009)

Researchers have modeled the projected agricultural production and net trade in 2050 under climate change (see pictures below). However, it is still unclear how production and yield of specific crops (such as maize and rice) as well as their trading network among countries may change. Such information is especially important for susceptible countries (e.g., those in west Africa) to form effective strategies to cope with the challenges and for other countries to offer help such as through trade policies.

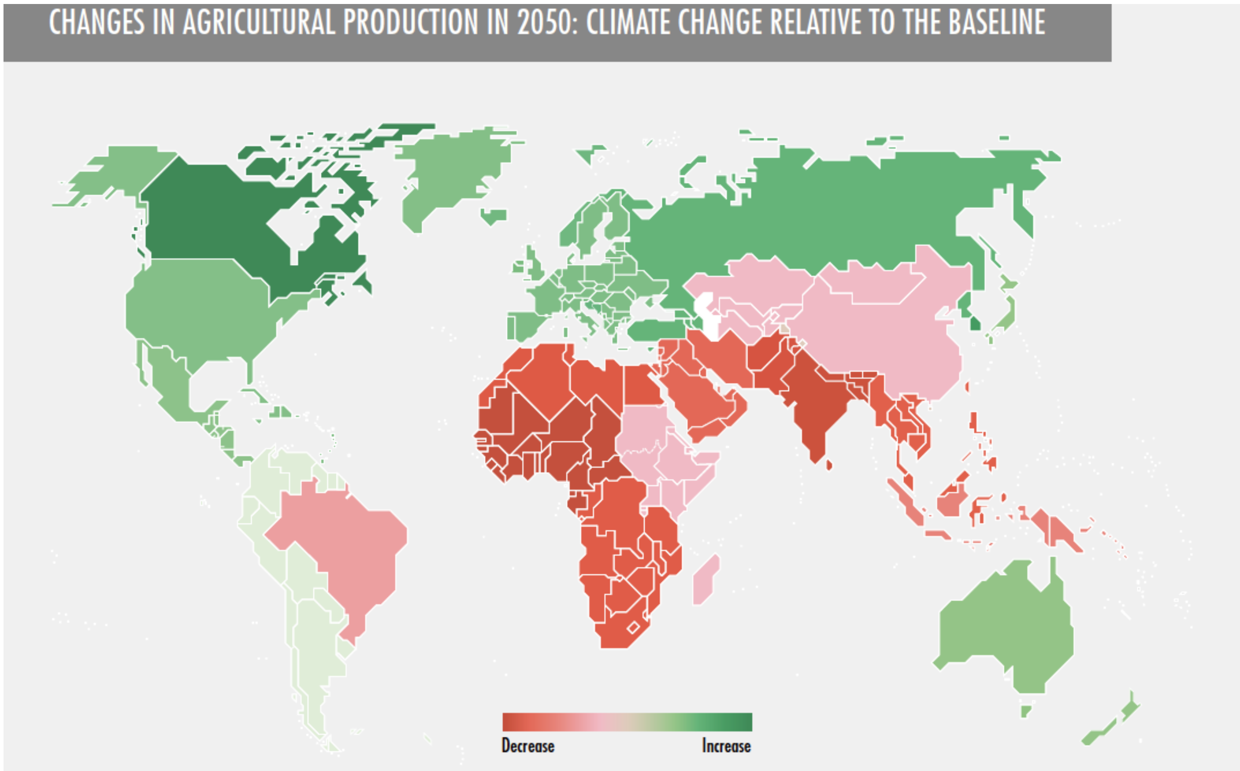

Figure 1.2 Predicted changes in agricultural production in 2050 (http://www.fao.org/3/I9542EN/i9542en.pdf)

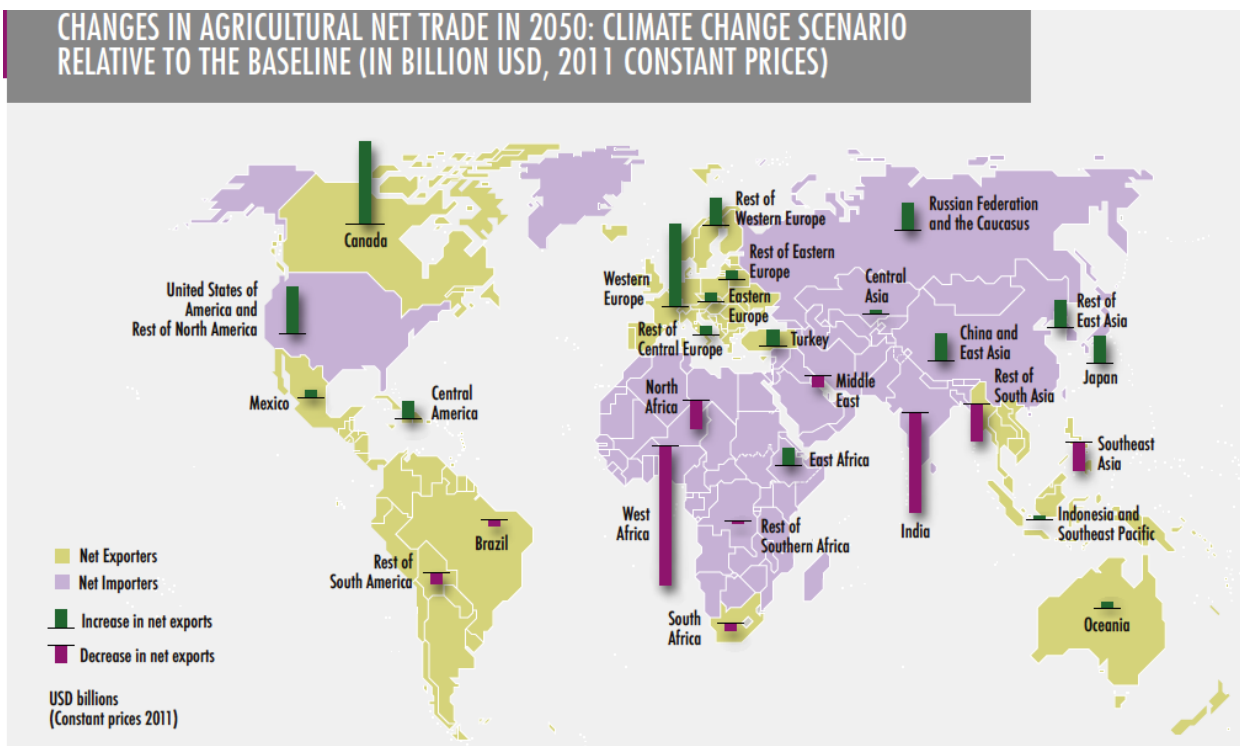

Figure 1.3 Predicted changes in agricultural net trade in 2050 with climate change scenario (http://www.fao.org/3/I9542EN/i9542en.pdf)

Meanwhile, currently there are limited options for organizations with less resources to obtain relevant data and intelligence to aid their decision-making. For example, agricultural production and trade datasets from public databases (such as those maintained by FAO) are served in raw data form and require data expertise to wrangle, visualize and model; insights for specific commodities or countries may be available through academic research papers or books, which would usually require paid access and can be difficult to grasp for non-technical users; agricultural business platforms, though offering valuable data intelligence, may incur considerable costs.

Therefore, a publicly available dashboard/website which can help users conduct data wrangling, visualization and forecasting, would be valuable for those with less technical expertise and resources to quickly access data-driven insights.

In this project, firstly we will examine the time series data of agricultural trade and production and predict future trends. Current and future trade network among countries will also be analyzed and modelled. Lastly, climate change scenarios will be incorporated to improve accuracy of the forecasts. The models will be deployed to an interactive dashboard by which users can extract intelligence with a few clicks.

Potential Clients

Government officials and policy makers, NGOs and research institutions interested in strategies to improve the economic outlook of countries whose agricultural production and food security may be threatened by future climate conditions.

Main Questions

(1) What is the diversity in agricultural production and trade networks among countries?
(2) How will such diversity change over time, as the globe gets warmer and crowdier?
(3) What can countries do to cope with future challenges (recommendations on improving production and trading profiles)?

2. Data wrangling and exploratory analysis

2.1 The Dataset

The food and agricultural trade dataset (http://www.fao.org/faostat/en/#data/TM) is used in the first part of analyses in this project. This data set is collected, processed and disseminated by FAO according to the standard International Merchandise Trade Statistics Methodology. A few steps have been taken by FAO to clean the source data in the normalized data file, including: outliers were checked; missing data were imputed (marked via the 'Flag' column) with trade partner data; data on food aid were added to take into account of total cross-border trade flows.

There are four main categories: export/import quantity (units: tonnes, heads for live animals) and export/import value (units: $1,000). All food and agricultural products imported/exported annually by all the countries in the world are included. Time coverage is annually from year 1961 to 2013.

2.2 Data wrangling and exploratory analysis

The datafile for all-countries data contains 35,976,124 rows and 13 columns (a sample of 5 rows is shown in Fig. 2.1). Missing values in the 'Value' column were checked and one row which contains missing data was dropped. Outliers and most missing values in this datafile has been filled by imputation with partner country data by the data curator (i.e., FAO).

	Reporter Country Code	Reporter Countries	Partner Country Code	Partner Countries	Item Code	Item	Element Code	Element	Year Code	Year	Unit	Value	Flag
27702338	200	Singapore	237	Viet Nam	220	Chestnut	5910	Export Quantity	2017	2017	tonnes	1.0	Im
5856829	96	China, Hong Kong SAR	216	Thailand	1164	Meat, dried nes	5610	Import Quantity	1993	1993	tonnes	2.0	NaN
24609836	173	Poland	162	Norway	892	Yoghurt, concentrated or not	5910	Export Quantity	2017	2017	tonnes	1.0	NaN
30395509	210	Sweden	229	United Kingdom	892	Yoghurt, concentrated or not	5922	Export Value	1989	1989	1000 US$	59.0	NaN
17445396	110	Japan	151	Netherlands Antilles (former)	828	Cigarettes	5922	Export Value	1995	1995	1000 US$	1047.0	NaN

Figure 2.1 Sample rows from the food and agricultural trade dataset (All country normalized).

Based on this dataset, a total of 424 unique commodities were traded, among 184 reporter countries and 255 partner countries. In order to analyze the diversity (total number of) commodities traded and trading partners per reporter country, the dataset is aggregated by items or trading partners per reporter country, respectively. Afterwards, the resulted data were reformatted from long to wide shape with years values as columns, which are suitable for time series visualization and analyses (Fig. 2.2).

Reporter Countries	Item	Element	Unit	Item Code	Y1986	Y1987	Y1988	Y1989	Y1990	...	Y2009	Y2010	Y2011	Y2012	Y2013	Y2014	Y2015	Y2016	Y2017
Benin	Juice, lemon, single strength	Export Quantity	tonnes	996	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
Sri Lanka	Pineapples	Import Quantity	tonnes	5166	0	0	0	0	0	...	0	9	5	0	1	0	0	3	2
Finland	Milk, whole condensed	Export Quantity	tonnes	11557	0	30	0	0	0	...	0	0	0	0	0	0	0	0	0
Lebanon	Cinnamon (cannella)	Export Quantity	tonnes	40887	0	0	0	0	0	...	4	0	20	5	0	0	5	5	1
Paraguay	Pigs	Import Quantity	tonnes	1034	0	0	0	0	0	...	0	0	0	0	0	88	259	72	177

Figure 2.2 Sample rows from the food and agricultural trade dataset (All country normalized), aggregated by 'Item' and reformatted for time series analysis.

The diversity (i.e., total number) of exported/imported agricultural commodities is heavily skewed to the right, with majority of countries export/import more than 250 items (Fig. 2.3).

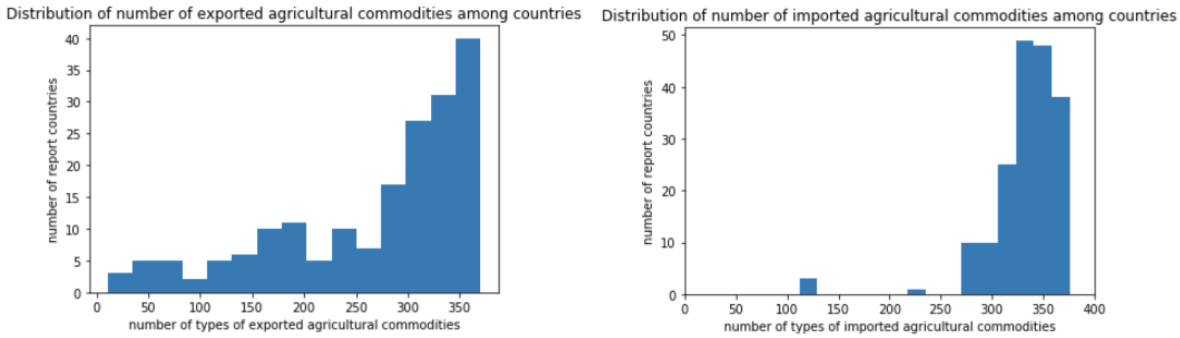

Figure 2.3 Diversity of exported and imported agricultural commodities among countries.

In contrast to diversity in commodities, the diversity in export/import trading partners is more evenly distributed, especially for export partners. There are 110 and 106 countries which have fewer than 180 export or import partners, respectively (Fig. 2.4).

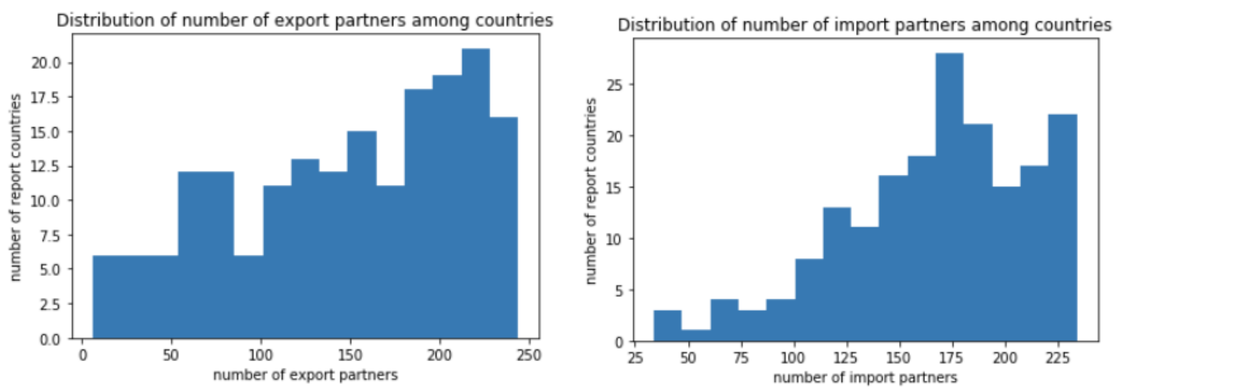

Figure 2.4 Diversity of export and import partners of agricultural commodities among countries.

Top-racked countries in terms of exporting or importing quantities and values were explored, as well as top-traded commodities in terms of quantities or values. Time series for trading of specific items were visualized to find patterns of interests. In addition, different features with rolling windows, such as 3-year rolling means, minimums, maximums and standard deviations were compared. The 3-year rolling mean was chosen to smooth the data for modeling, as it best captures the data profile while reduces variation (Fig 2.5).

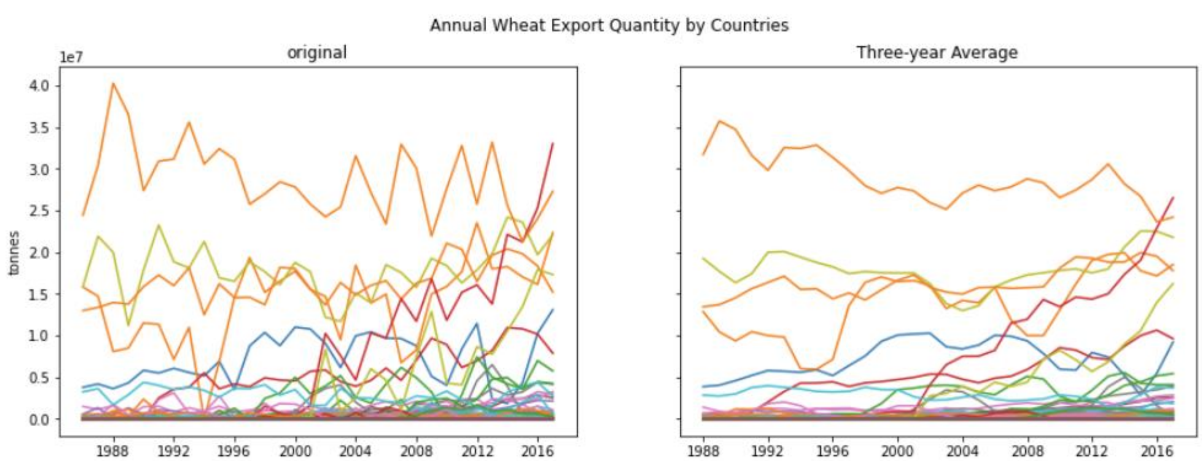

Figure 2.5 Annual wheat export quantity (tonnes) across reporter countries (left: original data; right: three-year rolling means).

3 Statistical Analysis and Modeling

1) ARIMA model

ARIMA (Auto-regressive integrated moving average) is a class of models that 'explains' a given time series based on its own past values. Correlations of 'self' with past values being analyzed include lags and the lagged forecast errors, and the resulted equation can be used to forecast future values. The model can be expressed as:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

Non-seasonal ARIMA models are generally denoted as ARIMA(p,d,q), where parameters p, d, and q are non-negative integers. p is the order (number of time lags) of the autoregressive (AR) model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average (MA) model (Wikipedia). In times series analysis, d is usually optimized to transform non-stationary data to become stationary. Intuitively, stationarity means that the statistical properties of the data-generating process do not change over time. Being (wide-sense) stationary is an assumption for time series analysis. In this project, the Augmented Dickey-Fuller test was applied to evaluate stationarity of data before future analysis.

2) Model development and evaluation

A machine learning model pipeline on time series data is distinct in a few aspects due to the chronological order of data. The order of data points should be maintained when splitting data into train-test sets. In addition, the cross-validation process requires a rolling-one-step approach. The steps in this project can be summarized as bellows:

a). Data is split into training set (70%) and testing set (30%) and the chronological order of data is maintained.

b). Train an ARIMA model with initial parameters (p,d,q) and make a one-step prediction, store the prediction in a separate list.

c). The first data point from testing set is added to training set, repeat step b).

d). After all data points in the testing data have been added into the training data and modeled, compare the list of predicted values with testing data and calculate evaluation metrics. Four evaluation metrics useful for time-series models are calculated, include the mean squared error (MSE), mean absolute percentage error (MAPE), correlation coefficient between predicted values and test data (CORR), and the min-max error (MinMax). In this case MSE is used as the main evaluation metric and the others are serve as references.

e). With grid search, update ARIMA model with the list of parameters, repeat steps b)-d) until the best model which minimizes MSE is identified.

Independent modeling is applied for each item-by-country selection with the above steps.

4. Application development

The modeling process is streamlined into a function for automation and deployed as an app served via Amazon Web Services (AWS) (Fig. 4.1-4.2). The app is built with Streamlit (https://www.streamlit.io/), an open-source app framework using Python.

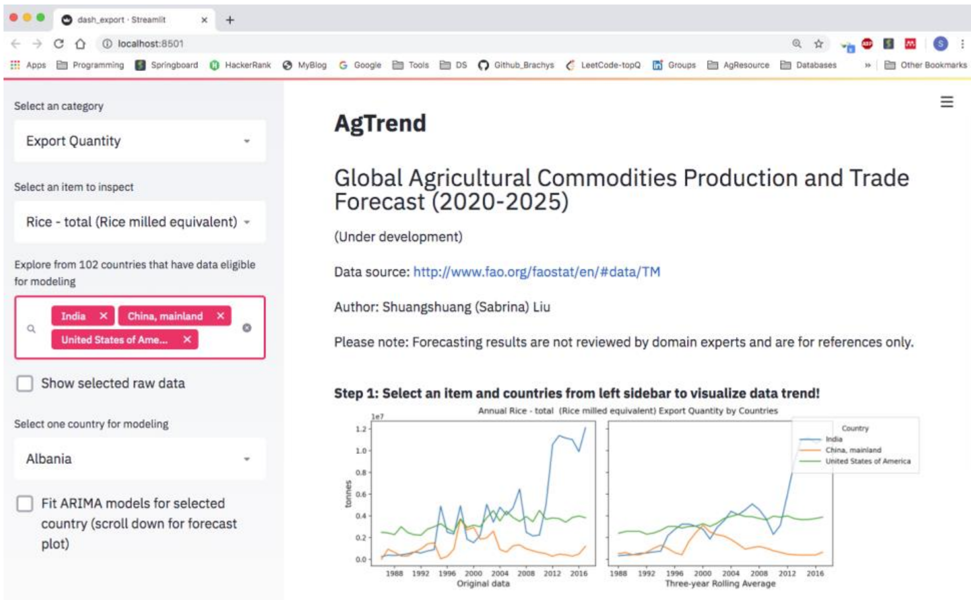

Figure 4.1 Interface of the app "AgTrend", showing data visualization.

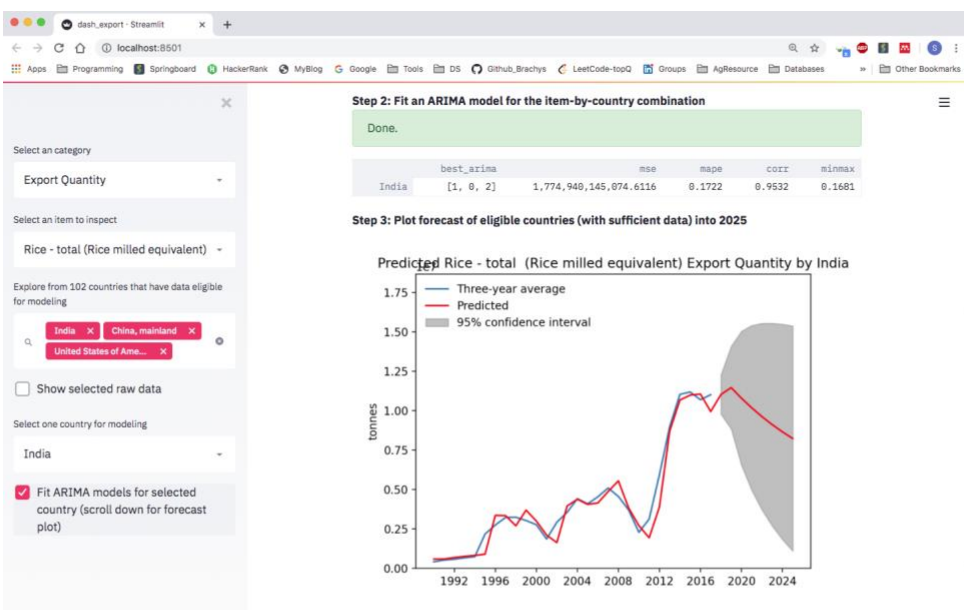

Figure 4.2 Interface of the app "AgTrend", showing example ARIMA modeling results.

5. Summary

In this capstone project, we identified the problem that many organizations especially in developing countries lack access to agricultural intelligence for decision making, such as international trade data and forecasts. We focused on using global agricultural trade data to build forecasting models. Data wrangling and exploratory analyses were performed, and machine learning model pipeline for time series analysis was constructed. Finally, the whole analysis pipeline was streamlined and delivered through an interactive app, AgTrend. The app was served via AWS and can be accessed at: http://agtrend.org:8501/.

AgTrend greatly improved the convenience of visualizing and forecasting global agricultural trade data, and can be updated regularly. Future improvements in the app include dynamic visualization of a country's trade network for selected items and its evolution over years, as well as a recommender system to suggest candidate trading items and/or partners. Lastly, climate change scenarios may be incorporated to improve accuracy of the forecasts.

References

Autoregressive integrated moving average, Wikipedia (https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average)

How to feed the world 2050, High-level expert forum, FAO 2009 (http://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf)

Hunger map, World Food Programme 2019.(https://www.wfp.org/publications/2019-hunger-map)

The state of agricultural commodity markets, FAO 2018. *(http://www.fao.org/3/I9542EN/i9542en.pdf)*