SMU

Data Science Quantifying the World DS7333.4043

Lance Dacy
Reannan McDaniel
Shawn Jung
Jonathan Tan

Unit 8 Case Study

June 30, 2020

Table of Contents

Introduction

Every year, influenza viruses pose remarkable impacts on socio-economics, such as costs of medical care, loss of productivity, and deaths. Since economic considerations are essential for influenza control, decision makers often need to examine the following questions for health interventions:

- How much will an influenza season cost a states/provinces or country?
- Which states/provinces or counties are susceptible high costs?
- Where to distribute vaccines to achieve the maximum returns?

Influenza, is an infectious disease caused by an influenza virus. Three of the four types of influenza viruses affect humans: Type A, Type B, and Type C. Type D has not been known to infect humans, but is believed to have the potential to do so. Usually, the virus is spread through the air from coughs or sneezes.

Influenza spreads around the world in yearly outbreaks, resulting in about three to five million cases of severe illness. In the northern and southern parts of the world, outbreaks occur mainly in the winter, while around the equator, outbreaks may occur at any time of the year.

Larger outbreaks known as pandemics are less frequent. In the 20th century, three influenza pandemics occurred: Spanish influenza in 1918 (17–100 million deaths), Asian influenza in 1957 (two million deaths), and Hong Kong influenza in 1968 (one million deaths). The World Health Organization (WHO) declared an outbreak of a new type of influenza A/H1N1 to be a pandemic in June 2009.

To date, only a small number of studies have estimated the economic impacts of influenza. The Office of Technology Assessment reported that the influenza accounts for $1 ~ 3 billion per year in medical costs. Meltzer, et al. argued that the annual economic burden of pandemic influenza could range from $71.3 ~ 166.5 billion. The latest estimation by Molinari et al. indicated that the short-term costs and long-term burden of seasonal influenza can be amounted to $26.8 ~ $87.1 billion a year. These studies have established systematic methods to analyze influenza economics and offered valuable guidance for interventions.

While most are interested in the medical and socio-economy costs of the flu, the Dacy Pharmacy Supply Chain is taking a different approach. Given that a new distribution center will be coming on-line in India, the organization needs predictable data to determine the the "just in time, just enough" level of electrolyte drinks that should be available to consumers. Given the spike of use in the drink from the US, the assumption is that all countries experience the same. Shipping liquids to India is an expensive proposition. The special instance of customs declaration and weight restrictions require the supply chain to be optimal to save inventory amassing in an expensive warehouse.

Method

The Dacy Pharmacy Supply Chain is looking to correlate the import of electrolyte liquids to the India Distribution Center based on the cases of Flu that aggravate just about any country that is populated with humans. If analysis of this data allows them to predict the number of Flu cases year over year in the future; supply operations can be coordinated to coincide with the cases; optimizing the JITJE (just in time, just enough) model for export/import expenses.

The method chosen for this level of prediction is an autoregressive integrated moving average (ARIMA) model. This model is used to generalize the autoregressive moving average (ARMA). Both of these models are fitted to the Flu time series data (described in the EDA section) to predict future points in the series (forecasting).

ARIMA excels at helping organizations predict:

- a pattern of growth/decline in the data is accounted for (auto-regressive)
- the rate of change of the growth/decline in the data is accounted for (integrated)
- noise between consecutive time points is accounted for (moving average)

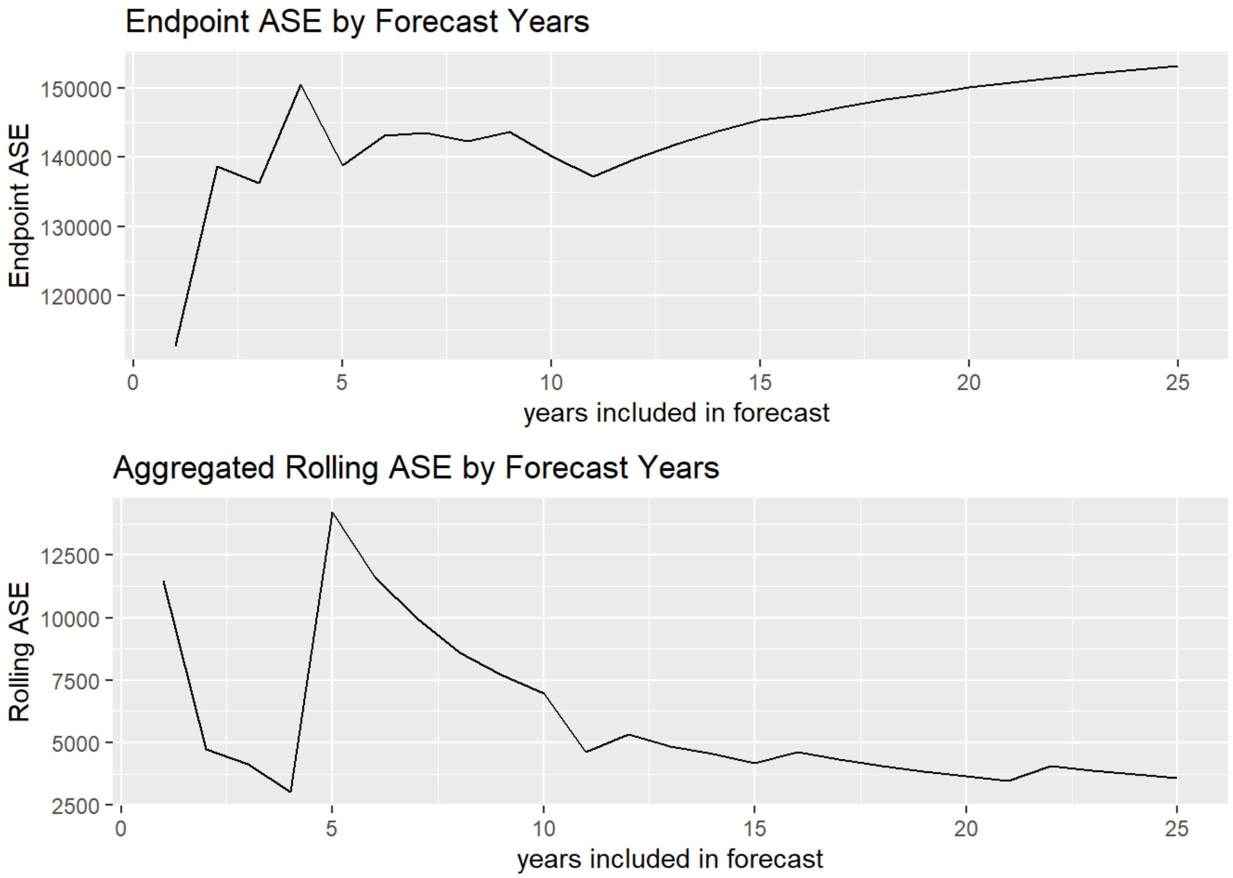
The nature of time series data is that it is made up of a sequence of data points taken at successive, equally spaced points in time. ARIMA models are applied where data show evidence of non-stationarity, an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

ARIMA models are typically expressed like "ARIMA (p,d,q)", with the three terms p, d, and q outlined as follows:

- **p**: the number of preceding ("lagged") Y values that have to be added/subtracted to Y in the model. This is done to make better predictions based on local periods of growth/decline in the data (autoregressive portion)
- **d**: the number of times that the data has to be "differenced" to produce a stationary signal. (integrated portion). If d=0, this means that the data does not tend to go up/down in the long term (i.e., the model is "stationary"). If this is true, then technically the model is just ARMA, not AR-I-MA. If p is 1, then the data is going up/down linearly. If p is 2, the data is going up/down exponentially.
- **q**: the number of preceding/lagged values for the error term that are added/subtracted to Y. This captures the "moving average" part of ARIMA.

This is important to understand given the interpretation of the model and the method presented to The Dacy Pharmacy Supply Chain.

Figure 1: End Point and Rolling Forecasted Data



Endpoint ASE by Forecast Years

Aggregated Rolling ASE by Forecast Years

Manipulating the data for the model, graphs become increasingly important to help visualize and interpret the actions taken. In the graphs above, "Endpoint ASE" performs the best at 1, 3, and 5 years of the data used. It actually shows degradation of performance the more data that is included. This is likely due to the model's inability to adequately predict the large outbreaks approximately every 5 years. "Rolling ASE" performs the best at 4 years, while adding more years of data is likely due to a large number of NA's or sparse data in the collection.

The data science team will iterate through multiple instances of the model to determine the best p, d, q qualifiers for the prediction. This analysis can be seen in more detail within the appendix.

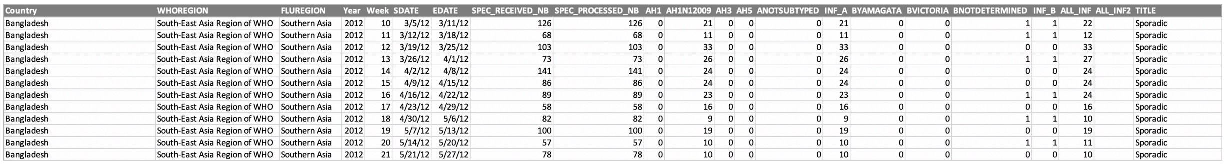
Data

Data Source

There are myriad datasources available with flu data and they can all be fairly specific. Given the nature of the flu strains and the seasonal aspect of the data, it is important to select a consistent and proven source. For the question of interest, the data selected will be provided by the World Health Organization (WHO). The data is categorized by region. The entire Southeast Asia region will be chosen given the expansion plans and distribution center reach. The time frame selected initially will be for 1995-2020.

The dataset can be organized by various categories such as Country (each country within the WHORegion (Southeast Asia)) and FluRegion which is sub-regions from WHHORegion. There are 11 countries within 3 sub-regions in the dataset.

Figure 2: Raw Data from the WHO



Country	WHOREGION	FLUREGION	Year	Week	SDATE	EDATE	SPEC_RECEIVED_NB	SPEC_PROCESSED_NB	AH1	AH1N12009	AH3	AH5	ANOTSUBTYPED	INF_A	BYAMAGATA	BVICTORIA	BNOTDETERMINED	INF_B	ALL_INF	ALL_INF2	TITLE
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	10	3/5/12	3/11/12	126	126	0	21	0	0	0	21	0	0	1	1	22		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	11	3/12/12	3/18/12	68	68	0	11	0	0	0	11	0	0	1	1	12		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	12	3/19/12	3/25/12	103	103	0	33	0	0	0	33	0	0	0	0	33		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	13	3/26/12	4/1/12	73	73	0	26	0	0	0	26	0	0	1	1	27		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	14	4/2/12	4/8/12	141	141	0	24	0	0	0	24	0	0	0	0	24		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	15	4/9/12	4/15/12	86	86	0	24	0	0	0	24	0	0	0	0	24		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	16	4/16/12	4/22/12	89	89	0	23	0	0	0	23	0	0	1	1	24		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	17	4/23/12	4/29/12	58	58	0	16	0	0	0	16	0	0	0	0	16		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	18	4/30/12	5/6/12	82	82	0	9	0	0	0	9	0	0	1	1	10		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	19	5/7/12	5/13/12	100	100	0	19	0	0	0	19	0	0	0	0	19		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	20	5/14/12	5/20/12	57	57	0	10	0	0	0	10	0	0	1	1	11		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	21	5/21/12	5/27/12	78	78	0	10	0	0	0	10	0	0	0	0	10		Sporadic

The World Health Organization (WHO) collects this data in weekly intervals over the course of 24.5 years. The information contains number of virus strands and individual cases. Personal demographics are left out for various reasons, but not necessarily required for the purposes of predicting supply needs.

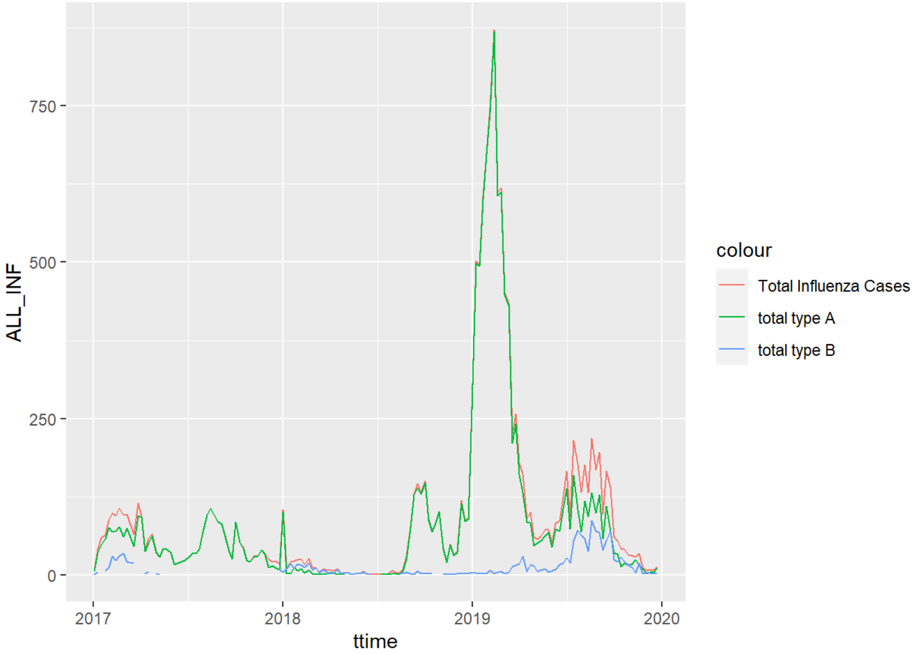
As the process matures, it might be valuable to foster a dataset that includes ages for special inventory for children and senior citizens. For unknown reasons, the years 1995-2010 are absent of the data required to predict the rate of flu in the future, therefore, the dataset was narrowed down to analyze the years 2017-2019 for the most common flu strands, A and B.

The data selected is not only important to predict cases by using the fields describing if someone had flu or not, but also the date/time it took place. Time series is a series of data points indexed (or listed or graphed) in time order and can be accounted for seasonality of the data.

Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

All of these are important pieces of data separately, but most important is the time in which the measurement occurred since that can relatively affect the prediction.

Figure 3: Initial Visual of the Data



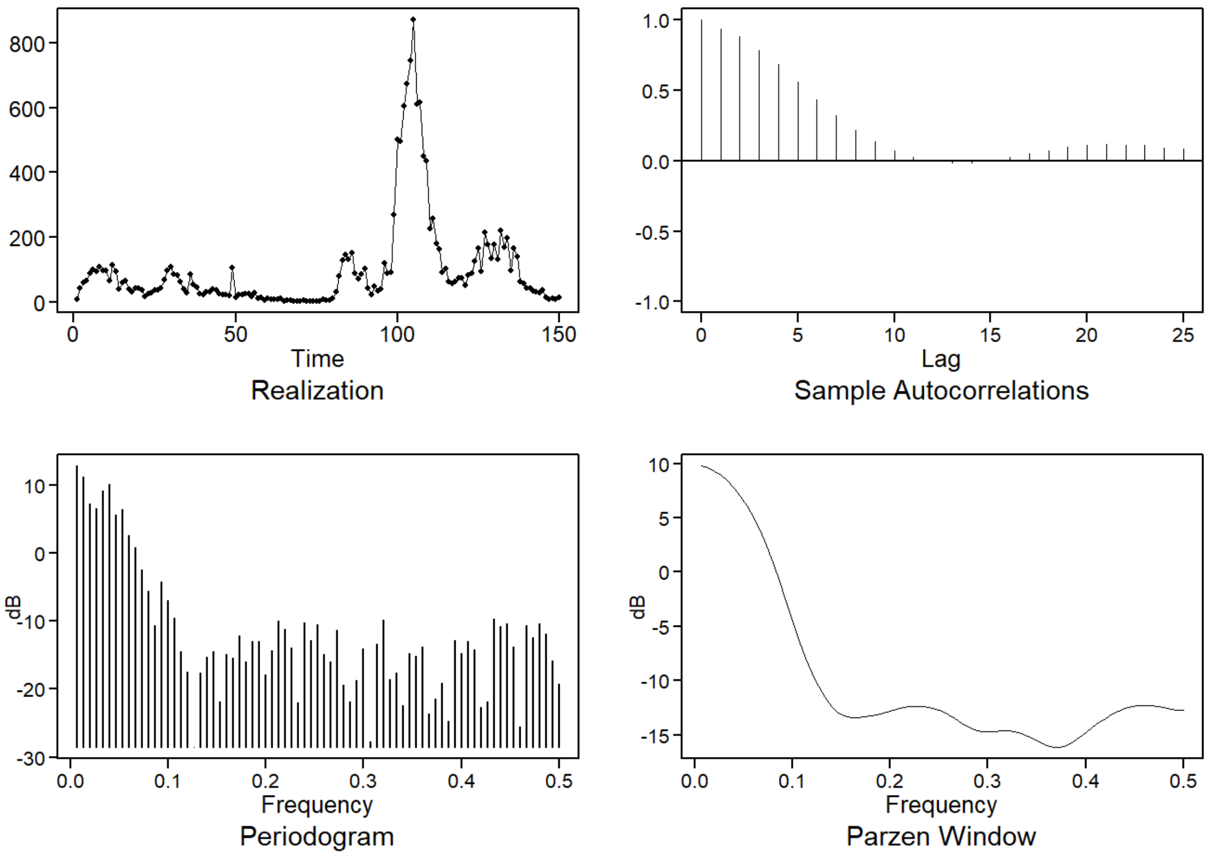
The clearest way to examine a regular time series manually is with a line chart such as the one above. The initial visualization of the data shows that the Type A and B Flu strains are the most common and that there was a huge spike in 2019 most likely attributed to the Swine Flu outbreak. This is why 2017-2019 will be the best data to use for a consistent time-series measurement.

While there is seasonality to pandemics and outbreaks, that is the not the focus of this question of interest. Dacy Pharmacy Supply Chain can react to those in times of need. The business question of interest for this purpose is to allow the supply chain to be prepared for the yearly influx that occurs consistently.

Exploratory Data Analysis

Selecting the appropriate years to forecast is an important step in the data analysis. Various methods will be employed to help determine what the data says about the years. Figure 4 provides some insight into the diagnostics and stationarity of the data.

Figure 4: Time Series Fit Graphs



Stationarity implies that the statistical properties of a process generating a time series do not change over time, not necessarily the data itself, but the properties of the data.

Lag is essentially delay. Just as correlation shows how much two time series are similar, autocorrelation describes how similar the time series is with itself. Consider a discrete sequence of values, for lag 1. Compare the time series with a lagged time series. In other words, shift the time series by 1 before comparing it with itself. Proceed doing this for the entire length of time series by shifting it by 1 every time. An autocorrelation function is then born of the data which basically visualizes how much it correlates with itself.

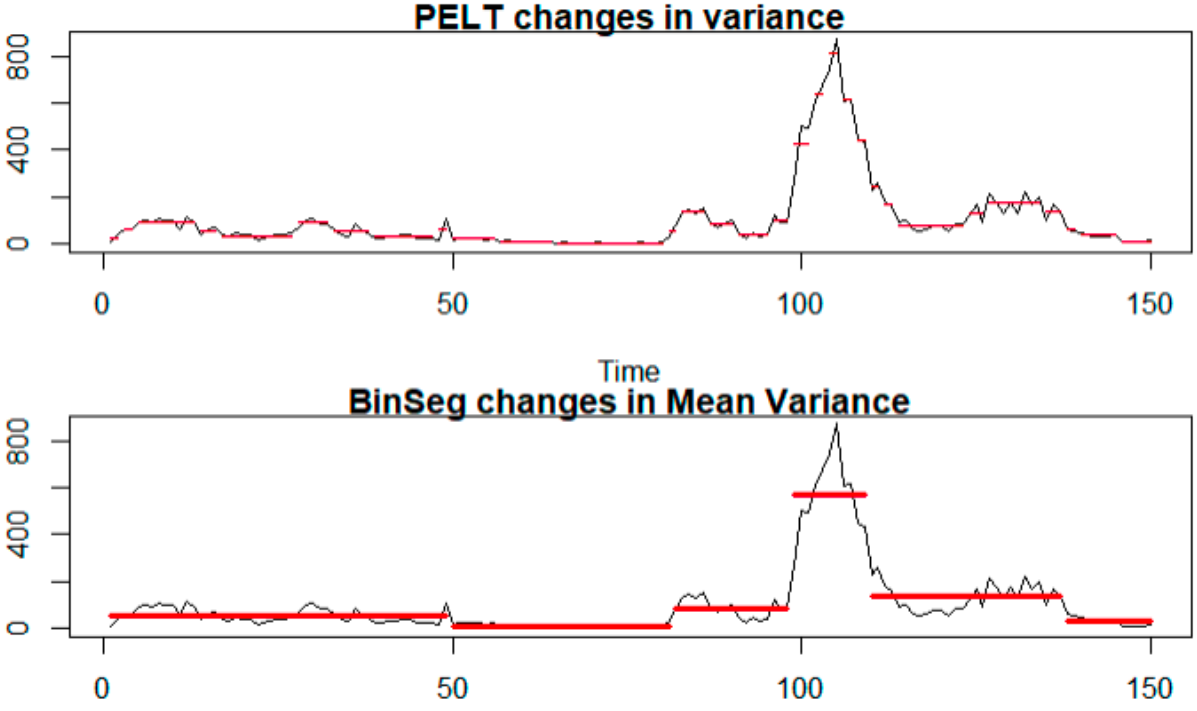
Parzen windows classification is a technique for nonparametric density estimation, which can also be used for classification. The technique approximates a given training set distribution via a linear combination of kernels centered on the observed points.

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.

For any time series, there will be a perfect correlation at lag/delay = 0, given the comparison of the same values with each other. As the time series shifts, the correlation values will begin decreasing. If time series comprises of completely random values, the only correlation will be at lag=0, and no correlation will exist anywhere else. In most of the time series datasets, this is not the case, as values tend to decrease over time, having some correlation at low lag values.

For this study, the autocorrelation can be used to interpret a linear relationship with the statistics of the data. An autocorrelation of +1 represents a perfect positive correlation (an increase seen in one time series leads to a proportionate increase in the other time series). An autocorrelation of negative 1 represents perfect negative correlation (an increase seen in one time series results in a proportionate decrease in the other time series).

Figure 5: Binary Segmentation and Pruned Exact Linear Time Graphs



The BinSeg (Binary Segmentation) / PELT (Pruned Exact Linear Time) changes in variance graphs are a form of change-point analysis that looks at differences in the mean variance over time. The binary segmentation method uses a function to find the the optimal number of change-points for a given variance in the data.

Each uninterrupted red line in the graph represents a time period where the mean value of the data fits the formula's definition of stable. Given the time series nature of this data, an ARIMA (Auto Regressive Integrated Moving Average) model will best allow for the prediction.

They are most easily interpretable on stationary data, meaning analysis should be directed towards these "stable" time periods shown in the graphic. Major spikes in variance identify time periods the model might not predict well for the business purposes of this exercise.

In summary, the mean variance for the spike in Flu cases around the beginning of 2019 (2 years, 100 units) is significantly higher than the mean variance for the rest of the time frame.

The change point analysis above was used to gain a better understanding of the mean variance of the total influenza cases recorded changed over time. Significant increases, like the early 2019 H1N1 (Swine Flu) outbreak can drastically affect conditions of stationarity, and subsequently affect an ARIMA model's ability to accurately predict these spikes.

Conclusion

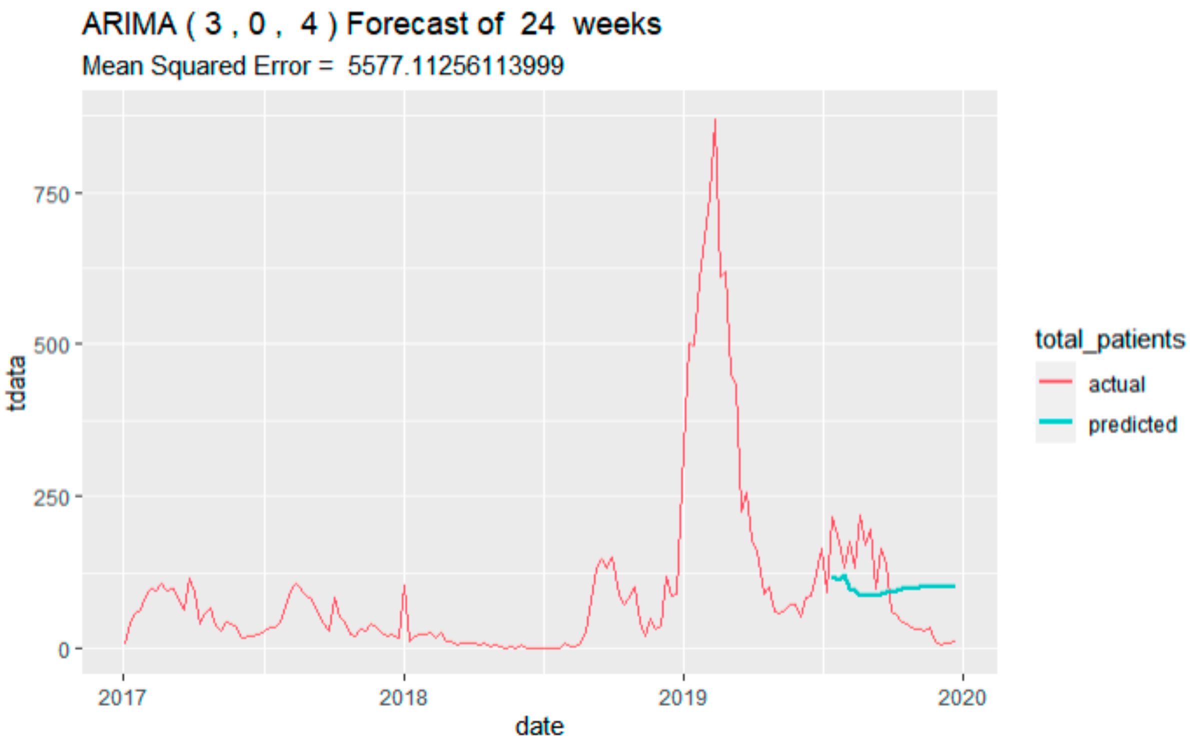
The ARIMA and its time series components definitely complicate matters of prediction. Given the analysis of the data and the various methods used, there are a few answers to the best p, d, and q qualifiers for this business question. To support the Supply Chain specialist's process, one method will be selected.

Given that the Rolling ASE calculated a different p, d, and q combination for each step of the analysis, it is determined to simply use the ARIMA model for a 3 year time frame.

The best model is ARIMA (3,0,4) :

- **p**: 3 auto-regressive components
- **d**: 0 differencing (the long-term data was technically "stationary", even though outliers existed)
- **q**: 4 moving average components

Figure 6: Predicted Cases Using ARIMA



The output coefficients will be used to analyze the supply chain data and "difference" the load factor for India based on the analysis performed. The data science team and the supply chain team will be meeting early next week on next-steps to apply these coefficients to the load factoring model in electrolyte drink distribution.

Appendix

As with any technology; the system required to generate meaningful data, cleaning of the data, as well as the models that were implemented require various software packages. The raw “code” used to explore the methods detailed above can be found in the following libraries:

- GitHub Repository