Case Study IV

James Vasquez, Javier Saldana, Sara Zaheri

Table of Contents

Table of Figures

1. Introduction

Influenza is a common and serious disease causing almost 30'000 deaths a year in the US. It affects all the age groups, however, children with %25 are more exposed to this virus. This number could increase to %40 during epidemics. Children are also the main reason for spreading the virus to household members. Each year, almost 4 million people suffer from severe illness caused by the flu.

Gradual antigen could bring about the emergence of new types of influenza virus. As a result, this pandemic never ceases, and people would be exposed to it every year. On the other hand, predicting influenza could help governments inform people to get vaccinated before the outbreak. According to aforementioned facts regarding Influenza, further research is required.

There have been several models to forecast flu epidemics. One of the most helpful and strong models is ARIMA which both predicts and estimates the number of flu positive cases. In this study, we will develop an ARIMA model for predicting the number of people who become sick by influenza in Texas.

Autoregressive Integrated Moving Average Models (ARIMA) are well-suited to time series data for both understanding and forecasting. This model enables us to predict a time series using the past values. We specify the model by a triple (p,d,q). Each variable shows the number of terms in the prediction equation. In particular, "p" is the number of autoregressive terms, "d" is the number of nonseasonal differences required for stationary, and "q" is the number of lagged forecast errors. In case of having seasonal patterns, we add an extra term to describe the seasonality. This model is called Seasonal Autoregressive Integrated Moving Average (SARIMA).

2. Methods

2.1. Data Source

We obtained weekly influenza positive cases from FluView Interactive[1]. This application has gathered the weekly flu data worldwide. Our team restricts the scope of study in the US, specifically the cases who live in Texas. The data is collected from late 2010 to 2020 containing 504 weeks. Our model intends to predict the total number of patients in a week.

As flu has a seasonal trend, the number of patients differs from season to season. The number of positive cases per week varies from 8496 to 58329. The average number of patients is 31043 having the large standard deviation 9051 due to being seasonal. There were some missing parts in

[1] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

the dataset which were in the age categories. Since we only need the total number of patients, we will not consider this missing data.

2.2. Exploratory Data Analysis (EDA)

As part of a time series analysis not only do we need to identify if we have missing values which is explained in section 2.1 but we also need to ensure that there are no missing dates. To determine that all dates are included in the data set we find the range of the series and calculate the number of days in the series. The weeks calculated should be equivalent to the original data set.

- Calculated Weeks: 504
- Exact Weeks: 504
  Please see python code for details

To double check this result the team re-sampled the data by seven days. When looking for 'isnull' values it should not have any values. This second check of missing dates determines that the data set is complete.

Plotting the series data allows the team to make several observations[2].

- Plotting by week
  - We cannot determine any trend, but we can see that there are peaks and valleys
  - Though it appears there is a pattern it is not definite
- Plotting by month
  - Showing total infected by month
  - The team can say that during December, January, and February there is an uptick in cases
  - During the summer months May thru August the rate is decreased
- Plotting by year
  - Total infected by year
  - No clear observation can be made from this plot

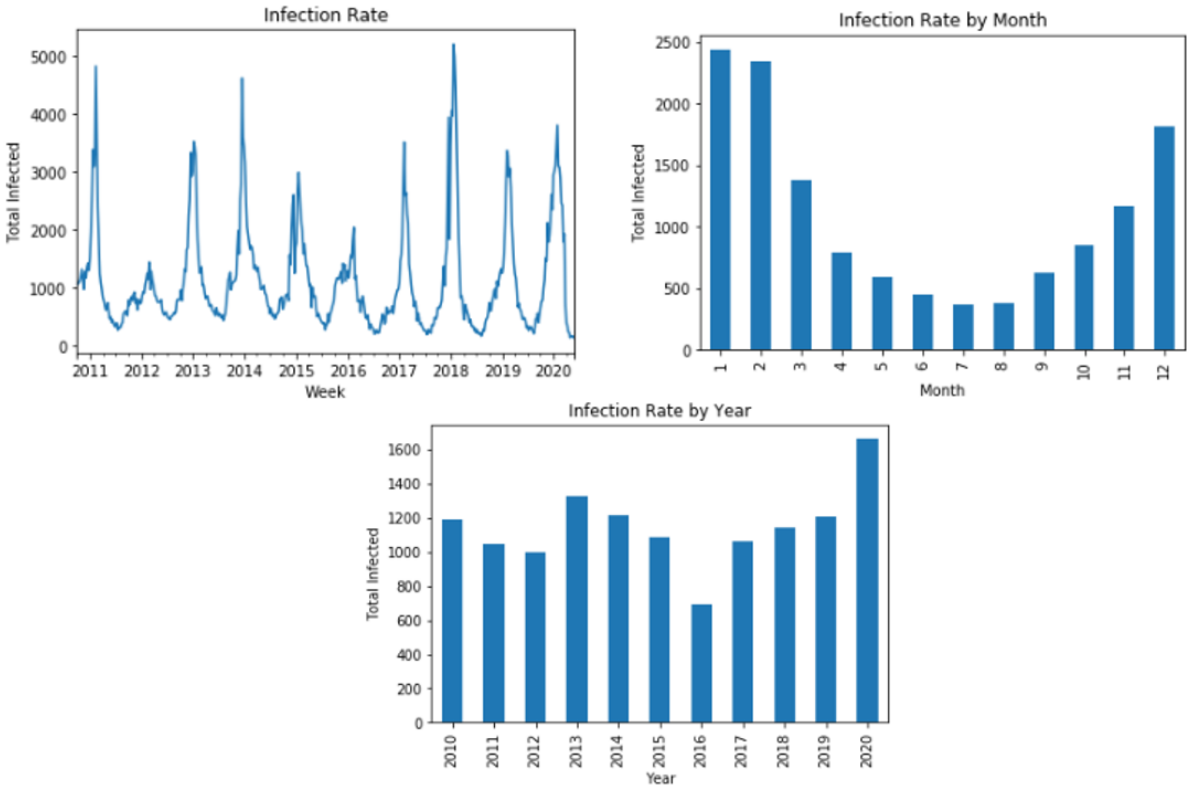

Figure 1. Infection Rate: Plotting data by mid-week and total infected. Infection Rate by Month: Total infected by month. Infection Rate by Year: Total all infected for that year.

Adding a decompositional analysis allows the team to show trends and patterns.

- Observed
  - Data plotted as is
- Trend
  - Between 2012 to mid-2016 there is a decrease of infected rates
  - After mid 2016 there is a general increase infected rate
- Seasonal
  - Clearly see an uptick and valley during the year
  - There is a seasonal pattern
- Residuals
  - The difference between the observed and predicted

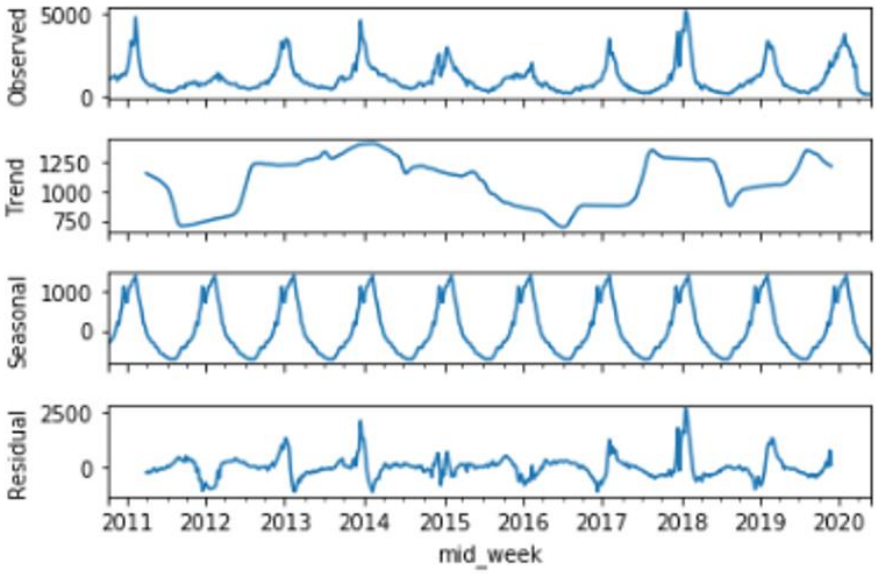

Figure 2. Seasonal Decomposition Plot.

2.3. Analysis

To perform the analysis, the model was split into 3 data sets. The training data set contains total infections from January 1, 2011 through December 31, 2018. This training set was used to model the time series and then tested on the testing data set, which contains total infections from January 1, 2019 through December 31, 2019. Once the model is trained and tested, it is validated by forecasting the total infections from January 1, 2020 through May 26, 2020.

The autocorrelations and partial autocorrelation of the data echo the decomposition plot in Figure 2. In the autocorrelations, there appears to be evidence of slowly dampening and oscillating behavior. The oscillating behavior is evidence of a seasonality. In this instance, we can expect $s = 52$ since the data was resampled with a weekly frequency. The partial autocorrelation shows a high correlation at lag 1. To address the correlation, the data is differenced by 1, which means that $d = 1$ in our model. Once the data is differenced by the first order and seasonally differenced by 52, we can see the seasonality and correlation has been removed (see Fig. 4).

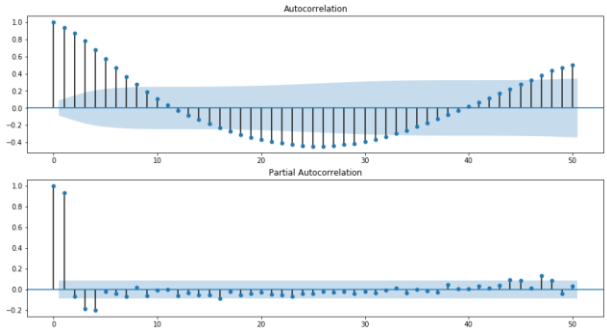

Figure 3. ACF and PACF of raw data

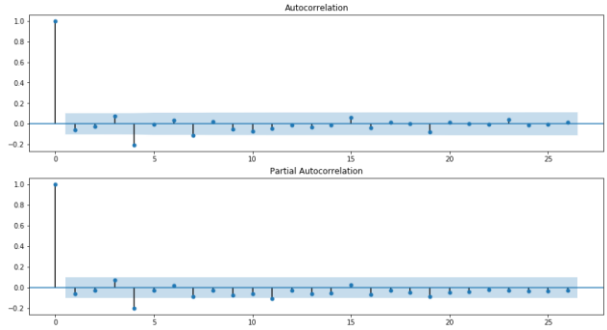

Figure 4. ACF and PACF of differenced and seasonally differenced data

By taking the differenced and seasonal order values, a stepwise search then can provide the order values with the lowest AIC. A stepwise parameter search was performed for the ARIMA model and the SARIMA model. The parameter search for the ARIMA model recommended an ARIMA (0,1,0), which is known as a random walk model. The parameter search for the SARIMA model recommended a SARIMA $(0,1,0)\ (3,1,1)_{52}$ model.

3. Results

With the recommended order values, we fit the models into the training data set. Once fit, we predict the total infections in 2019 and compare the predictions from the ARIMA and SARIMA models to the actual values. Based on the fit, the ARIMA model generated a 917 root mean squared error and the SARIMA model generated a 900 root mean squared value. Yet, the ARIMA model produced a -1% r^2 while the SARIMA model explains 2% of the variance. Comparing the ARIMA and SARIMA we can see the lack of seasonality impact the performance of the ARIMA model.

When the ARIMA model is differenced, it is modeling white noise which is why it optimizes itself by simply modeling a straight line. In doing so, it plots the mean and just remains constant as it strives to minimize the RMSE. The SARIMA model performs much better by matching the seasonality of the data. However, the uptick in total infections during 2019 was not matched by the model. As a result, we anticipate this issue to persist moving forward if the number of total infections doesn't return to its normal level. Considering the SARIMA model performed better than the ARIMA model, we moved forward with the SARIMA model and used it to predict the 2020 total infections.

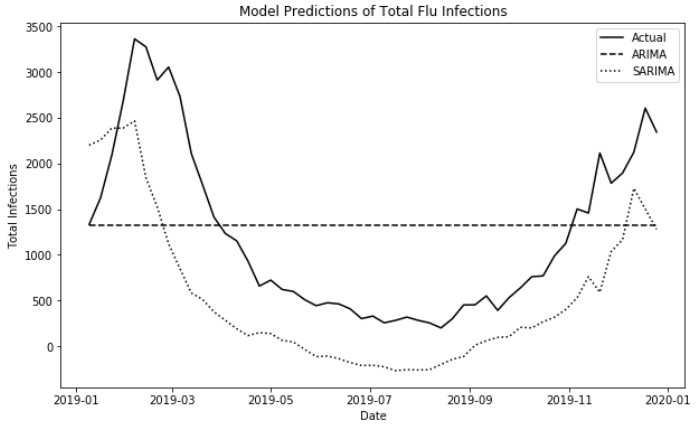

Figure 5. ARIMA and SARIMA 2019 total infection predictions

As evident in the model comparison predictions from 2019, the output for 2020 also proved to be problematic. The RMSE was 1198 which is higher than it was for the entire year of 2019. Yet, the model explains 22% of the variance of total infections which is considerably higher than the 2% in the test. The model captures the seasonal peak and valley, but it fails to match the raw count of total infections. The main reason for this lies primarily in the SARIMA orders (p,d,q). Like the ARIMA, the SARIMA orders modeled a random walk model with seasonality. This means that the model was able to fit the seasonality but believes the variance within that seasonality is white noise.

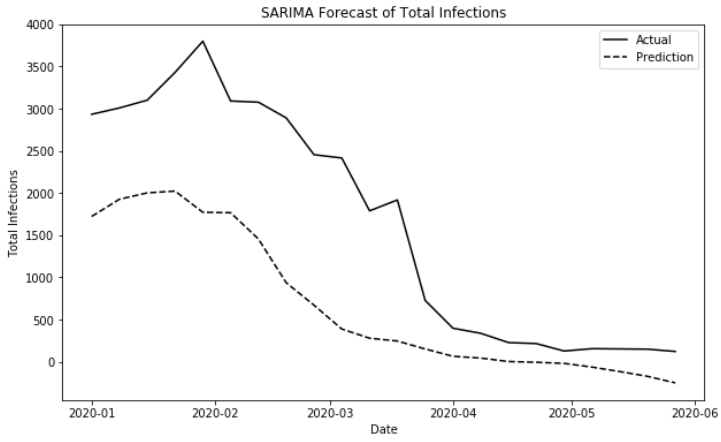

Figure 6. SARIMA total infection predictions of 2020

4. Conclusion

With the SARIMA model outperformed the ARIMA model with a RMSE of 900 compared to the RMSE of 917, the selected model for interpretation is the SARIMA. Though issues with seasonality may occur, it is the team's belief that this issue is not a trigger to use the ARIMA model. As with any data set, as additional data becomes available being able to compare the models and adjusting as needed should be applied when re-evaluating interpretations.

The data set is concentrated to only Texas, additional southern states could potentially be used to help improve the model. The ability to predict the infection rate allows for a high-level overview, but by adding ages to the patients would give a more granular insight to the population that could be at higher risk. This possibly could enable medical personnel and social services to direct resources to the age group at higher risk to themselves.

5. References

1. Business, F. (n.d.). Introduction to ARIMA: Nonseasonal models. Retrieved June 25, 2020, from https://people.duke.edu/~rnau/411arim.htm

2. Brownlee, J. (2019, August 28). How to Decompose Time Series Data into Trend and Seasonality. Retrieved June 26, 2020, from https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

6. Appendix (Code)

The code is implemented in Jupyter notebook. We leave this part blank and submit the. ipython file together with the project."