Predicting COVID-19 Hospitalizations

IEE 577

Nicholas Seah

Michael Shim

Kesin Dolwani

**Table of Contents**

[Introduction 3](#_Toc70170919)

[Methodology 3](#_Toc70170920)

[Results 6](#_Toc70170921)

[Discussion/Conclusion 8](#_Toc70170922)

[Appendix 9](#_Toc70170923)

# 

# **Introduction**

Over the past year, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or COVID-19 has forced many to pivot their way of living and thinking. As of April 1st, 2021, there have been over 30.7 million documented COVID cases and 554 thousand COVID-related deaths solely in the United States. Due to the considerable impact of the virus, immense amounts of money are flowing towards the funding of research, recovery, and medicine related to COVID-19 to help fight the ongoing pandemic. One of the many challenges for the society amongst this pandemic is that there is not sufficient capacity in hospitals across the country to treat everyone suffering from COVID-19.

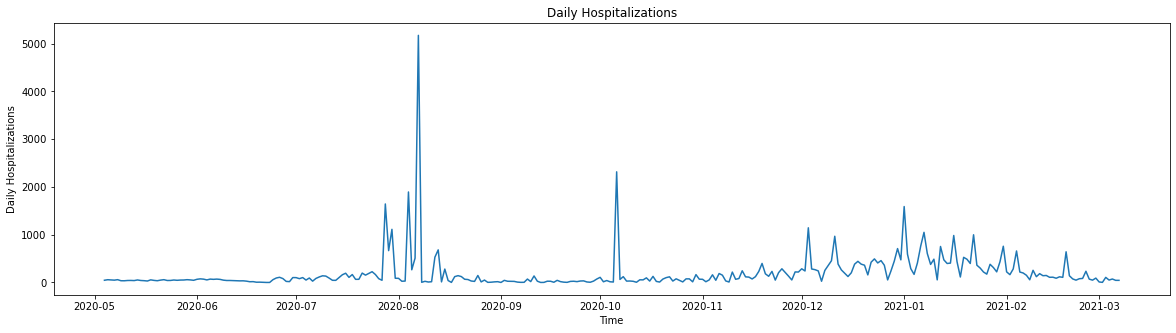
The purpose of this project is to predict the number of COVID-19 Hospitalizations using machine learning algorithms to determine if a data-driven prediction model can be used to predict hospitalizations.

We used public data from The COVID Tracking Project to train our model on daily hospitalizations from May 5th, 2020, to March 7th, 2021. The forecast predicts the daily hospitalizations in Arizona from March 8th, 2021, to March 14th, 2021. We will test our model using the actual daily hospitalizations from the Arizona Department of Health Services.

# **Methodology**

The first step in building our machine learning model was to import and clean the data. We imported the data as a CSV file using the pandas read\_csv function. We obtained the testing data from the Arizona Department of Health Services. To clean the data, we first had to remove the columns that were irrelevant to this project. We removed all columns other than the daily hospitalizations and the dates. The date column had to be converted to the DateTime format from a string format for ease of use.

# 

The next step was to plot daily hospitalizations. We used the matplotlib library for this. The original plot of daily hospitalizations had multiple negative hospitalization values. These seemed like incorrectly entered values, so we took the absolute value instead of removing these rows. Figure 1 shows the daily hospitalization plot. Next, we had to ensure that the time series was stationary. A stationary time series has a constant mean and variance and makes it much easier to predict future values. We decided to explore an ARMA model and ARIMA model after consulting multiple professors at ASU. 

*Figure 1: Daily Hospitalizations Plot*

We then plotted the autocorrelation plot and partial autocorrelation plot. We used these plots to select a range of values for hyperparameter tuning. We used the pandas.plotting library for the autocorrelation plot. We used the statsmodels.graphics.tsaplots library for the partial autocorrelation plot. We used the autocorrelation plot to select the range for the moving average parameter q. We used the partial autocorrelation plot to select the range for the autoregressive parameter p. The autocorrelation and partial autocorrelation plots are in Figures 2 and 3 in the Appendix.

We chose a range of values from 0-10 for both the lag parameter p and moving average parameter q. These were chosen by heuristically analyzing the ACF and PACF plots. Next, we implemented the ARIMA model to produce the 7-day forecast of daily hospitalizations from March 8th - March 14th, 2021. We used the heuristic approach to select initial values of p = 10, d = 1, q = 7 to obtain initial results.

Next, we used Grid Search to optimize these parameters. We had to start by creating a function to evaluate our models. We used the Root Mean Error Statistic (RMSE) as our error statistic. This function outputs the RMSE value for a given set of parameters p, d, and q. We then validated that the evaluation function worked as intended. We then created a second function to implement Grid Search to optimize the parameters p, q, and d. We were able to implement a function to produce optimal parameter values of p = 3, q = 9, and d = 2 for the ARIMA model and p = 3, d = 0, and q = 5 for the ARMA model.

After we obtained the optimal parameter values, we used our forecast to create diagnostic plots. We used the statsmodels.tsa plot diagnostics function to plot the standardized residuals over time, a histogram plus the estimated density of standardized residuals, a Q-Q plot, and a correlogram. The ARIMA and ARMA plot diagnostics are in Figures 4 and 5 in the Appendix.

From these plot diagnostics, we can see that the ARIMA model performed better than the ARMA model. The standardized residual plots are nearly identical and indicate that we may have some outliers in our dataset. These are likely from holidays or large gatherings that promoted the rapid spread of COVID-19. The histogram plot tells you if the data is normally distributed. The ARIMA model is almost normally distributed, however, the ARMA model is not. The QQ plot also tells us if the data is normally distributed. From the plots, we can see that the data is close to normally distributed but not perfect. The correlogram shows the serial correlation in data that changes over time. As expected, the lag at 0 is equal to 1, while the lag from 1-10 is approximately 0.

# **Results**

We compared the predicted hospitalizations to the actual hospitalizations in both the ARIMA and ARMA models to compute the RMSE. The predicted and actual number of daily hospitalizations from March 8th to March 14th with the ARMA and the ARIMA models are in Figure 6. As shown by the comparison of the RMSE values, the performance of the ARIMA model was four times as better than the ARMA model, thus indicating that the model to be used for predicting hospitalizations is the ARIMA model.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** |  | **Daily Hospitalizations** | | | | | | | **Total** | **RMSE** |
|  | **3/8/21** | **3/9/21** | **3/10/21** | **3/11/21** | **3/12/21** | **3/13/21** | **3/14/21** |
| **ARMA** | Prediction | 84 | 80 | 87 | 99 | 102 | 105 | 112 | 669.9 | 47.8 |
| Actual | 68 | 74 | 66 | 36 | 57 | 45 | 37 | 383 |
| **ARIMA** | Prediction | 45 | 64 | 54 | 49 | 56 | 46 | 44 | 358 | 11.8 |
| Actual | 68 | 74 | 66 | 36 | 57 | 45 | 37 | 383 |

*Figure 6: ARIMA and ARMA Results*

However, the RMSE of the ARIMA model is a non-zero number: 11.8. This indicates that on average, there is an error rate of nearly 12 newly hospitalized patients per day. Between the number of totals predicted and actual hospitalized numbers using the ARIMA model, there is a percent error of nearly 6.5%. The current ARIMA model is by no means perfect. It can be further improved to reduce the percent error by performing more data cleaning and data transformation. However, the current percent error of nearly 6.5% is an error rate that hospitals in Arizona can prepare ahead, given the predicted hospitalizations of the next day. Though it is not perfect and can be further improved, the currently implemented ARIMA model is of the acceptable range for predicting the hospitalizations due to COVID-19 in Arizona.

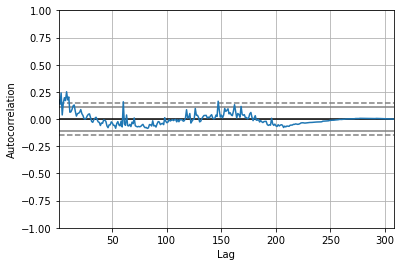
# **Discussion/Conclusion**

The ARIMA model can be applied to each state and be broken down into individual hospitals, making it very useful in real-world decision making. The model is relatively accurate to about a week out state-wide. Using population density around each area, the state data can be broken down further to be accurate for each hospital. Based on the results, we recommend that a 20% safety factor be added to the predicted new COVID-19 patients to be fully confident in staffing numbers. This safety factor could prove costly for hospitals. We could improve the existing ARIMA model by adding different factors like vaccination numbers and if non-essential businesses are open if hospitals want to decrease this safety factor. These inputs likely heavily affect hospitalization numbers in reality and can be implemented into a future ARIMA model. Another way to improve the safety factor is by exploring other types of models. A model implementing Kalman filters could be a more accurate predictor in the future. Kalman filters are good at looking at time series data with a lot of noise and predicting future results in the actual data.

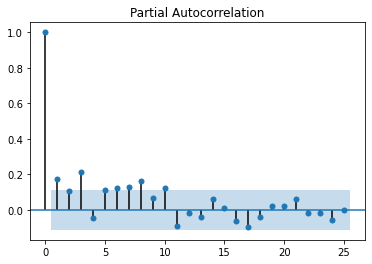
The ARIMA model also contains a lot of good signs about the COVID-19 pandemic moving forward. Both models show that the number of patients being admitted to the hospital is in a decreasing trend. This can contain a lot of information for decision-making with a wide variety of companies who want to know if they can get ready for a post-COVID-19 society. The ARIMA model can alert companies of trends moving into future weeks that could impact business decisions. In conclusion, the model results are meaningful and can influence real-world decision-making.

# **Appendix**

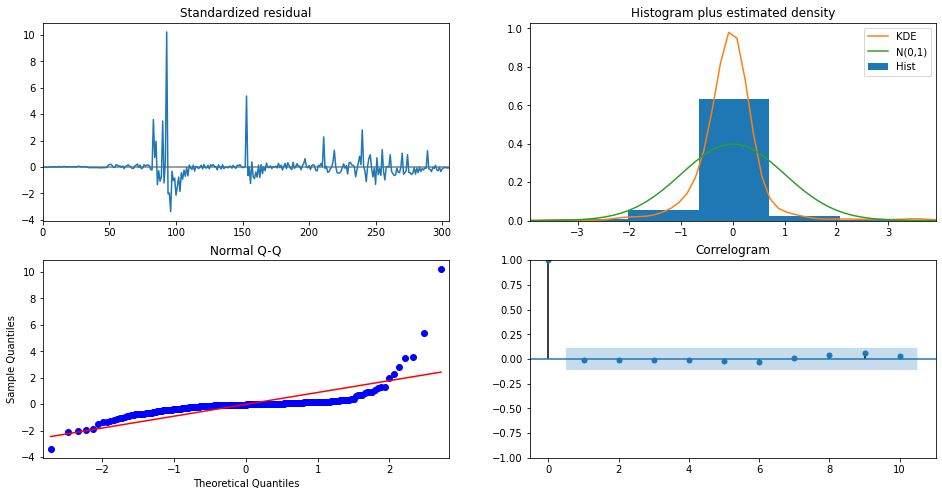
*Figure 1: Daily Hospitalizations Plot*



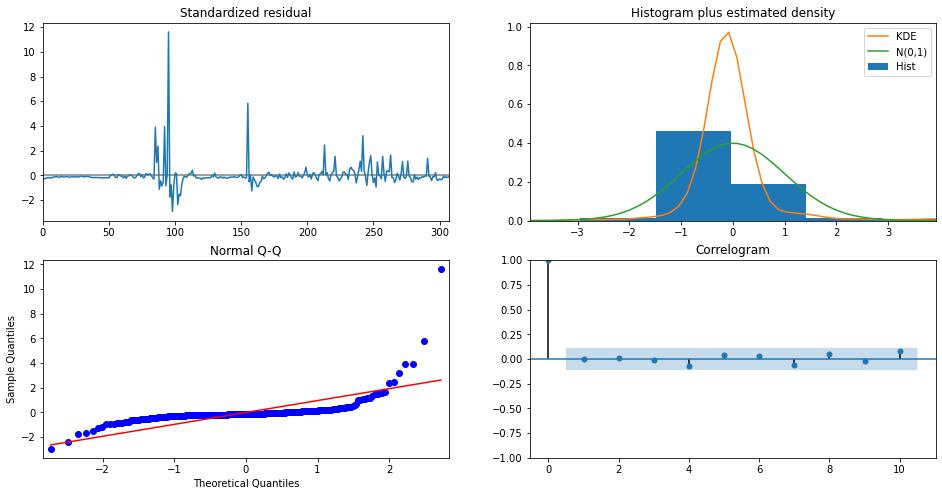
*Figure 2: Autocorrelation Plot*

**

*Figure 3: Partial Autocorrelation Plot*



*Figure 4: ARIMA Plot Diagnostics*



*Figure 5: ARMA Plot Diagnostics*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** |  | **Daily Hospitalizations** | | | | | | | **Total** | **RMSE** |
|  | **3/8/21** | **3/9/21** | **3/10/21** | **3/11/21** | **3/12/21** | **3/13/21** | **3/14/21** |
| **ARMA** | Prediction | 84 | 80 | 87 | 99 | 102 | 105 | 112 | 669.9 | 47.8 |
| Actual | 68 | 74 | 66 | 36 | 57 | 45 | 37 | 383 |
| **ARIMA** | Prediction | 45 | 64 | 54 | 49 | 56 | 46 | 44 | 358 | 11.8 |
| Actual | 68 | 74 | 66 | 36 | 57 | 45 | 37 | 383 |

*Figure 6: ARIMA and ARMA Results*