Forecasting and Modeling of Chicago COVID Data

Kristin Fesmire, Connor Mennenoh, Veera Manohar Reddy Alavalapati

May 3, 2023

MATH 446/546

Introduction/Problem Statement

From the early part of 2020 when the COVID-19 pandemic began, COVID-19 has been a force on Chicago and the world, greatly impacting policy and the basic livelihoods of residents. Up until now, there have been hundreds of thousands of contractions of COVID-19, thousands of hospitalizations of COVID-19, and thousands of deaths as a result of COVID-19 in Chicago. Though this event is an ongoing tragedy, the value of information and data surrounding the pandemic has been recognized since the early days. In order to track the progress of the spread of the virus, news outlets and data inclined individuals have continuously been recording, processing, and analyzing the data that surrounds all of the different aspects of this pandemic. This provides context and information not only for this pandemic as it continues to change and evolve, but also for future pandemics that will inevitably occur.

Given Chicago's COVID-19 data, we intend to produce forecasting models for three main components related to COVID-19: cases, deaths, and hospitalizations. These time series include the number of cases, hospitalization, and deaths in Chicago from March 1st, 2020 to March 29th, 2023. The primary goal will be to produce accurate models for forecasting future values for these series, but also to explore the interactions between these series and how behavior in one corresponds to or precedes behavior in another. We also intend to segment the cases data and create time series models using COVID-19 policy decisions from lawmakers. We hope that this segmentation will give a full understanding of the behavior of COVID-19 throughout time in Chicago.

Prior Work

As COVID-19 has been a substantial influence on the world, many different individuals and groups have spent time studying it through the lens of different approaches and disciplines. In data science, there have been different models utilized in an attempt to provide a model that can predict the future behavior of the infection. In this exploration, many different models have been tried, often in an attempt to better handle confounding factors that can make time series analysis difficult. Some work involved the use of multivariable time series to account for the different types of COVID data. For example, Nick James et al. applied a multivariate model for COVID-19 using cases and deaths as variables. (2022) Other work included work by Tomar and Gupta (2020), where they used a LSTM (Long Short-Term Memory) neural network model to help account for bias in other models. Singh, et al utilized a TBATS (T: Trigonometric seasonality B: Box-Cox transformation A: ARIMA errors T: Trend S: Seasonal components) to better account for seasonality. A common factor among these and other works such as Chyon, et al and Somyanonthanakul, et al is that they experienced either strong initial, or continued success with ARIMA models, even using them as the basis for further refinement. For this reason, along with the ARIMA model being one of the most discussed components of the coursework, the decision was made to model the COVID-19 data using an ARIMA model.

Methodology

To model the three series in the data set, ARIMA models were constructed. ARIMA stands for Auto-Regressive Integrated Moving Average. This model incorporates the

Auto-regressive and Moving Average components of an ARMA model and incorporates the capacity for differencing in order to accommodate non-stationary series. This made the ARIMA model a natural choice to utilize, as it allowed a single common approach to all three series, even when they presented a mixture of stationary and non-stationary data. This model was implemented using the auto.arima() function in R, which allows automatic fitting of an ARIMA model that is the optimal fit for the given data.

We decided to use ARIMA to model the time series for a number of reasons. COVID-19 data has several jumps where there was a large number of deaths, cases, or hospitalizations for a small period of time. With this, the data needs to be smoothed to construct a proper time series model. ARIMA accounts for this through the "moving average" portion of the model. Considering these jumps, the data also needs to be differenced in order to make the data stationary. ARIMA also accounts for this considering that ARIMA is an integrated ARMA model.

Considering the complexity of COVID-19 data, we also produced four segmented ARIMA models for the cases time series according to COVID-19 policy dates:

1. Start of data to Issuance of Stay at Home Order: 3/1/2020-3/26/2020
2. Stay at Home order (Ending with Phase IV: Gradually Resume): 3/26/2020 - 8/20/2021
3. Mask Mandates: 8/20/2021 - 2/28/2022
4. Vaccine Proof Requirement: 12/21/2021 - 2/28/2022

Data Preprocessing and Initial Analysis

The data, shown in Figure 1, was initially processed by removing rows with missing values and missing dates. The dataset was then sorted by time in order to properly visualize the dataset as a time series. In order to better understand the time series data, the data was initially analyzed by visualizing the time series, visualizing the ACF functions, and visualizing the PACF functions. These gave basic information about the time series including potential trends and seasonality that the data may have provided. The initial dataset gave extensive information regarding age, race, ethnicity, and gender for each case, death, and hospitalization. Though this information may produce interesting results, it was removed from the initial dataset for simplicity purposes, as we are only modeling the time series for cases, deaths, and hospitalizations.

Figure 1, Basic Visualization of the Dataset:

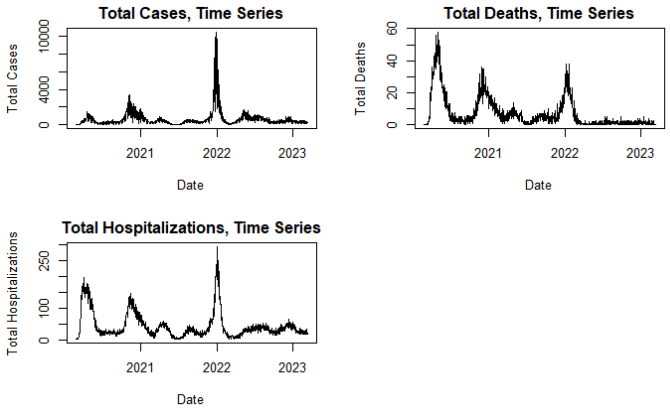
	Date	Cases...Total	Deaths...Total	Hospitalizations...Total
424	2020-06-13	120	18	23
810	2020-06-14	92	12	28
458	2020-06-15	235	20	31
642	2020-06-16	210	11	24
144	2020-06-17	219	21	28

In order to utilize a segmented data strategy to our models, the model was divided by the dates specified in our introduction. The first date range was taken from the first recorded case in Chicago on 03/01/2020 until the beginning of the initial stay at home order issuance on 03/26/2020. This consists of 28 total days. The second date range began on 03/26/2020 and

ended on 08/20/2021. This period of time was defined by the stay at home order that was instituted. It consisted of 514 total days. The third date range began on 08/20/2021 and ended on 02/28/2022. Mask mandates were the defining aspect of this period. This period consisted of 126 total days. The final date range began on 02/28/2022 and has continued to the current point in time. For the purpose of our model, the current point in time would be defined as 03/10/2023, which was the latest date recorded when the dataset was acquired. This period of time consisted of 378 total days.

The data used to create the model was obtained from publicly available COVID-19 datasets for the City of Chicago. Specifically, the data was obtained from "Chicago Data Portal," a website that provides publicly available datasets for the City of Chicago. The data was collected from March 01, 2020 until March 10, 2023.

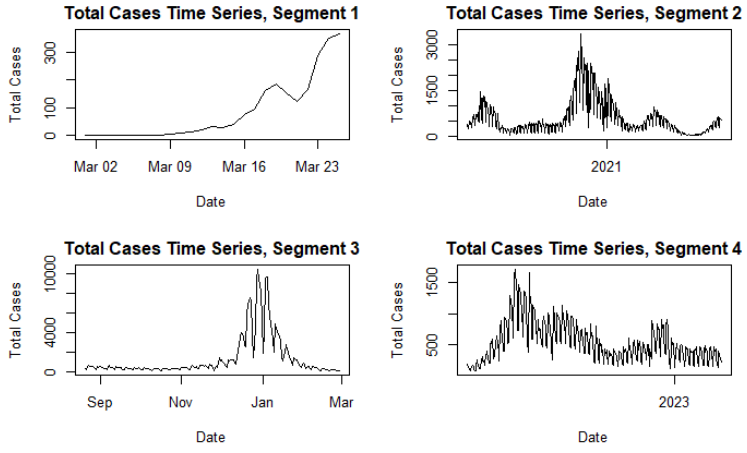
Figure 2, Time Series for Cases, Deaths, and Hospitalizations:



Initial time series analysis involved visualizing the three initial time series and comparing them. One notable difference between the time series is the spike of cases at the beginning of

2022. The other two series also showed spikes at this time period, but the spike of cases was much larger and much more significant than the spike of cases for the other two time series. Another notable insight from the three time series was the spikes in deaths and hospitalizations at the beginning of the pandemic. The cases time series also experienced this spike, but the spike was not as significant as the other two time series. This may be due to low case reporting at the beginning of the pandemic. The three time series all seemed to experience spikes at the beginning of 2021 and the beginning of 2022, but there was no spike at the beginning of 2023. This may have been an indicator of seasonality, if not for the lack of spike at the beginning of 2023. Overall, the three time series tended to flatten after the peak at the beginning of 2022.

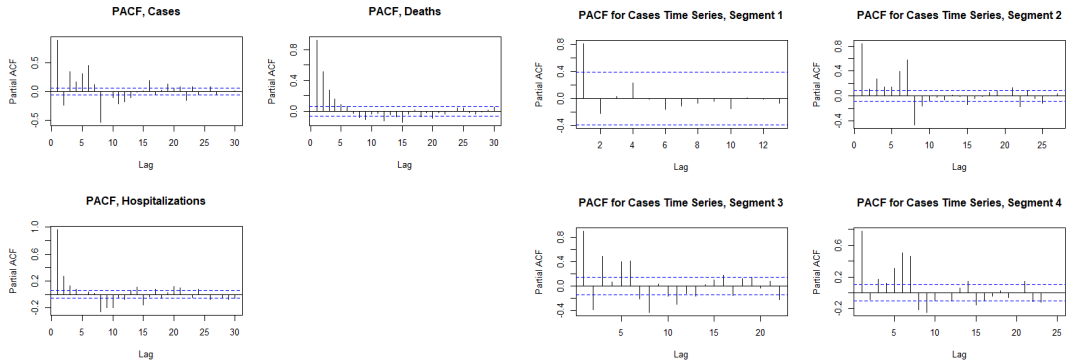
Figure 3, Segmented Cases Time Series:



The four segmented time series also provide interesting information about the general behavior of COVID-19 cases in Chicago. The first segment gives a general idea of the behavior of the spread of COVID-19 when the virus started to spread in Chicago. March 2020 shows an upward trend throughout the entire month. The second segment consists of the first main spike of

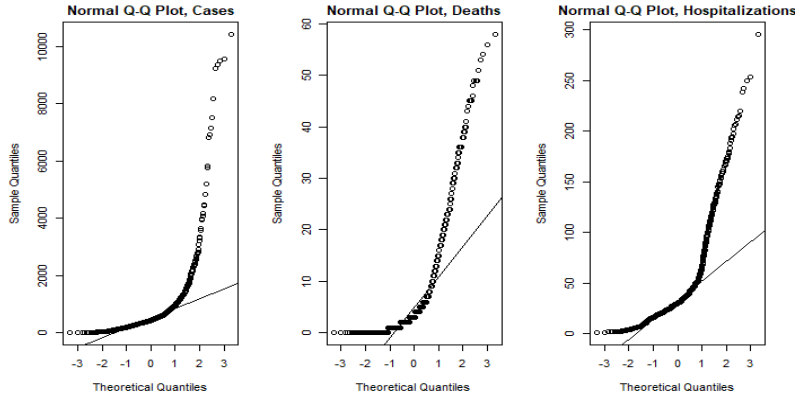
COVID-19 in early 2021 and the tail end of when cases started to be reported. Generally, it can be seen that there is no specific trend in the data, just spikes. The third segment involves the large spike of case reports that occurred in early 2021. Other than that spike, there is no general trend in the data. The fourth segment includes the wide adoption of at-home COVID-19 tests and the end of mask mandates. These factors may give an understanding of why the time series is flattened compared to the two previous segmentations. Overall, the four segmentations allow for better and more in depth understanding and analysis of the behavior of COVID-19 case data.

Figure 4, PACF Plots of the Time Series:



The PACF plots shown above give credence to the validity of the approach through the ARIMA model. Each of the PACF plots shows diminishing over time, and most do so quickly after only a handful of lags. Furthermore, in the case of the segmented plots the behavior becomes more favorable, with segment one especially falling within the boundary after only one lag. Taken as a whole the plots of the PACF support the use of ARIMA, and imply that the models produced will be relatively manageable in terms of the number of coefficients they will require.

Figure 5, QQ-Plots of Cases, Deaths, and Hospitalizations:



Generally, to get a basic understanding of the three initial time series, the data was plotted using qqnorm() in R, as shown in Figure 5. This was done to obtain an understanding of the distributions of the time series. Since the three series do not follow the normality line that was plotted with the quantiles, it can be seen that the three series do not follow a normal distribution. The series that follows closest to a normal distribution, though, may be hospitalizations. This was considering that the quantiles of hospitalizations follow the linear line for theoretical quantiles -2 to 1.

To determine the stationarity of the three time series, the Augmented Dickey-Fuller Test was utilized. This test was used through the adf.test() function from the tseries package. The Augmented Dickey-Fuller Test is a statistical test that is used to determine the stationarity of a time series. This test has a null hypothesis of H_0: p < alpha and an alternative hypothesis of H_a: p > alpha. If the null hypothesis is rejected, there is enough information to conclude that the time series is stationary. The Augmented Dickey-Fuller Test was computed for each of the three models, as shown in Figure 6.

Figure 6, Augmented Dickey-Fuller Test:

Data	P-value	Conclusion
Cases	0.01	Null hypothesis can be rejected
Deaths	0.07824	Null hypothesis cannot be rejected
Hospitalizations	0.01	Null hypothesis can be rejected

Given that the Augmented Dickey Fuller Test concluded that non-stationarity could be deduced for the deaths dataset, the deaths data was differenced to account for this. This was done using the diff() function in R. From this, the deaths data was differenced by one. Though this is true, from Figure 2, it can be seen that there are clear jumps in the data. We will keep this in our considerations when modeling the data.

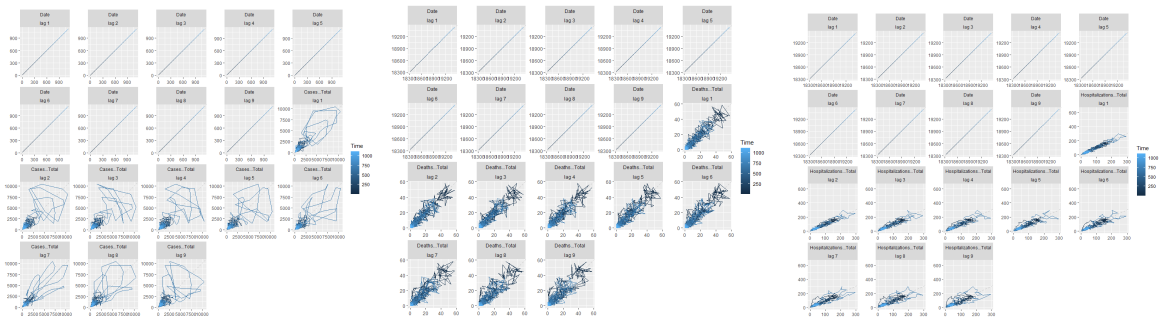
Figure 7, Augmented Dickey-Fuller Test, After Differencing:

Data	P-value	Conclusion
Cases	0.01	Null hypothesis can be rejected
Deaths	0.01	Null hypothesis can be rejected
Hospitalizations	0.01	Null hypothesis can be rejected

From Figure 7, differencing the deaths time series by one and recomputing the Augmented Dickey-Fuller Test concluded stationarity for the dataset. This allowed us to model the deaths time series using an ARIMA model with an understanding that the data needed to be differenced.

The time series function, ts(), in R was used to obtain a time series object of the data set. This object was utilized to create additional visualizations of the time series. One of these visualizations is shown in Figure 8. This visualization produced plots that plotted the time series against their lagged time series. Along with the PACF and ACF functions, this helps to give an understanding of how the time series should be modeled.

Figure 8, Lagged Time Series for Cases, Deaths, and Hospitalizations:



Modeling

Models for the time series were produced in two specific ways. First, the time series were modeled using ARIMA as is. Then, models were formulated by segmenting the cases time series by COVID policy.

After the initial exploration of the data and determining the stationarity and other key behaviors of each series, all three full time series were modeled using the auto.arima() function in R. This resulted in three models corresponding to the total cases, deaths, and hospitalizations respectively.

Initial modeling of the cases time series produced using auto.arima() gave:

(1) $X_t + (0.7708)X_{t-1} + (0.3100)X_{t-2} + (-0.1172)X_{t-3} = Z_t + (0.2121)Z_{t-1} + (-0.7616)Z_{t-2} + (-0.5886)Z_{t-3} + (0.3306)Z_{t-4} + (0.6299)Z_{t-5}$

Initial modeling of the deaths time series produced using auto.arima() gave:

(2) $X_t + (0.9367)X_{t-1} = Z_t + (-1.7376)Z_{t-1} + (0.8321)Z_{t-2} + (-0.0456)Z_{t-3}$

Initial modeling of the hospitalizations time series produced using auto.arima() gave:

(3) $X_t + (1.3374)X_{t-1} + (-0.4405)X_{t-2} = Z_t + (-1.7675)Z_{t-1} + (0.8791)Z_{t-2}$

After computation of the three initial models were computed, the additional four models for each section were computed.

Modeling of the cases time series, segment 1, produced using auto.arima() gave:

(4) $X_t = Z_t + (0.7141)Z_t$

Modeling of the cases time series, segment 2, produced using auto.arima() gave:

(5) $X_t + (0.2914)X_{t-1} + (-1.0309)X_{t-2} + (0.0693)X_{t-3} + (-0.5674)X_{t-4} + (-0.4602)X_{t-5} = Z_t + (-0.8562)Z_{t-1} + (1.1789)Z_{t-2} + (-0.8041)Z_{t-3} + (0.7077)Z_{t-4} + (-0.2052)Z_{t-5}$

Modeling of the cases time series, segment 3, produced using auto.arima() gave:

(6) $X_t + (0.1724)X_{t-1} + (-0.9297)X_{t-2} + (-0.1426)X_{t-3} + (-0.3614)X_{t-4} + (-0.5066)X_{t-5} = Z_t + (0.1096)Z_{t-1} + (0.4859)Z_{t-2}$

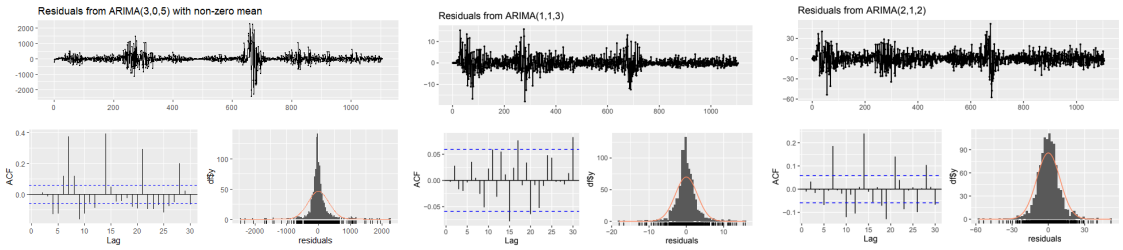
Modeling of the cases time series, segment 4, produced using auto.arima() gave:

(7) $X_t + (0.9045)X_{t-1} + (-0.5329)X_{t-2} + (-0.2466)X_{t-3} = Z_t + (-1.4360)Z_{t-1} + (0.7200)Z_{t-2}$

Results

To understand the performance and quality of the models further, we used the checkresiduals() function in R to investigate the residual values for each model and to aid in determining if the models captured as much information about the series as possible.

Figure 9, Residual Plot for Cases ARIMA Model:



The residual plots for the three original time series all produced distributions that spiked above a normal distribution, giving non ideal residuals. The residuals histogram for hospitalizations, the third plot shown in Figure 9, gives the closest distribution to normal among the three models, indicating better residuals and therefore a better fit of the model. The Ljung-Box test was further computed for all three models to give an understanding of whether or not the produced residuals were white noise. For the cases time series, the test suggested that the residuals were white noise with a p-value that was less than 0.05. For the deaths time series, the test suggested that the residuals were not white noise with a p-value of 0.294, which is greater than the threshold value of 0.05. The last time series, hospitalizations, also suggested that the

residuals were white noise with a p-value that was less than 0.05. The deaths time series, from the plot in the middle in Figure 9, gave the best autocorrelations of the three models, with most autocorrelations within the shown interval.

From our three initial COVID-19 models, it became clear that producing a model using the entirety of COVID-19 data, from the beginning of the pandemic to the current point in time, gave unusable results. The models did not capture all of the information of the model, and they gave considerably high autocorrelations between the residual values. Part of the reason for this was hypothesized to be the fact that the data spanned more than three years, over the course of which Chicago had a wide variety of policies that affected both transmission as well as testing. With this, COVID-19 data is generally very complex. As seen in Figure 2, there are many hard-to-predict spikes in the data. This is especially true for the case time series, which produces a very large spike at the beginning of 2022. To combat this, the three main series were partitioned into segments corresponding to the main changes in policy that influenced COVID-19 transmission. Those included the stay at home order, mask mandates, and vaccine mandates. Similar models were then generated for cases for each of these time segments. We then used each segmented time series and checked their performance in comparison to the unsegmented time series. The performance was checked in terms of both accuracy and residual quality.

Figure 10, Model Error Comparison:

Model	Log-Likelihood	BIC	RMSE
Cases	-7965.53	16001.13	326.12
Cases, Segment 1	-111.48	232.49	24.31

Cases, Segment 2	-3426.47	6921.53	199.1143
Cases, Segment 3	-1463.39	2968.75	529.1941
Cases, Segment 4	-2417.87	4871.29	154.63
Deaths	-2764.45	5563.93	2.95679
Hospitalizations	-4071.98	8178.99	9.66

To analyze the computed models, we specifically looked at the log likelihood, the bayesian information criterion, and root mean squared error provided by the ARIMA models. As shown in Figure 10, the initial COVID cases model, (1), gave a relatively large RMSE error of 326.12. The deaths and hospitalizations models, (2) and (3), gave much smaller RMSE values of 2.957 and 9.66. These values were not ideal, but they still indicated much better modeling for the Deaths and Hospitalizations time series. Considering the segmented cases time series, all three gave better modeling of the cases time series than the initial unsegmented cases time series. All of the four segments gave a larger log-likelihood and smaller BIC values. These both indicate better fit for the four models. All of the four segments, except for the third segment, also gave smaller RMSE values for the time series.

After the initial analysis of the time series, the seven series were finally forecasted. This is shown in Figure 11 and Figure 12. These forecasted time series generally predicted flattened values for the future. For the segmented time series, the forecasted values generally seemed to give an accurate representation of the future behavior of the time series. Segment 1's forecasting suggested that the cases would continue to rise in the future, which is what is represented in Segment 2. The other time series suggest flattened future values, which is generally consistent with how the time series behaved.

Figure 11, forecasting for 200 days future of the time series:

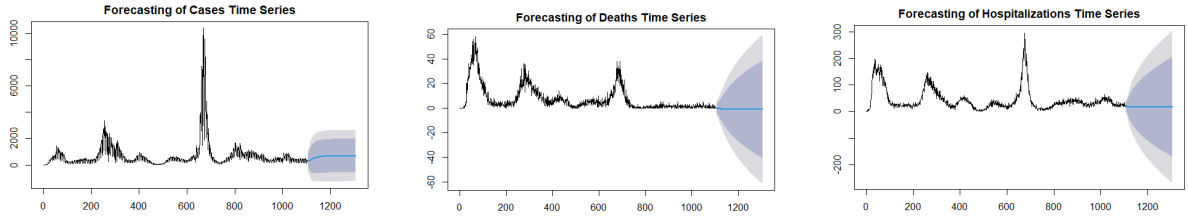
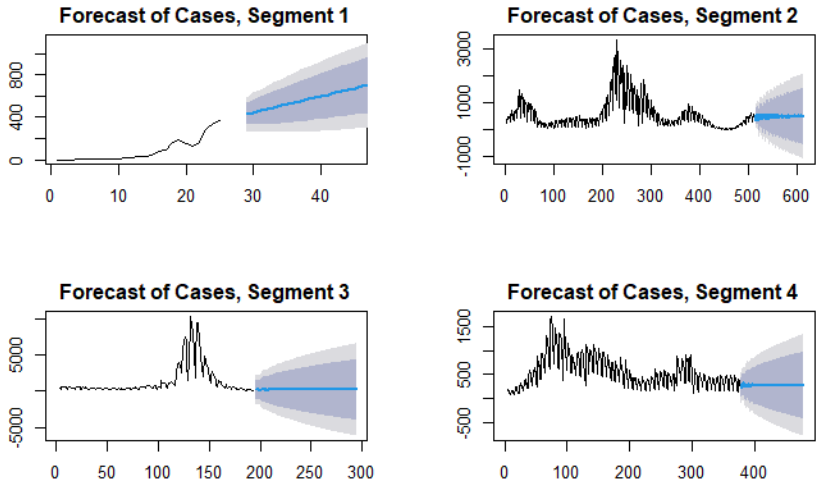


Figure 12, Forecasting for 100 days in the Future (20 Days in the Future for Segment 1):



Conclusion/Future Work

Other potential avenues for our work could have included delving into more complex models and algorithms for predicting the time series data for COVID-19. Most of these are outlined in "Prior Work." Generally, though, it may have been useful to try adapting models like LSTM and comparing these to our generated ARIMA models. We could have also tried implementing a multivariate time series. This may have given interesting information considering

that the data provided three main variables: deaths, hospitalizations, and cases. Interesting insights may have been derived from this, as it would be interesting to see how the three variables impact one another.

Along with the segmented cases time series calculation, we could have also segmented the other two time series: deaths and hospitalizations. As segmenting gave more information on the cases time series, segmenting also may have given more information on the deaths and hospitalizations time series. Considering that segmentation was conducted on the cases time series partially due to the unpredictability of that time series, utilizing the segmentation strategy may have given even more accurate information on the other time series since the other time series generally showed better modeling.

Segmentations could also have been further distributed. There are an uncountable number of factors influencing the spread of COVID-19. From this, we could have looked further into policy regarding COVID-19. We could have also split the time series by the wide adoption of things like stay-at-home tests. The widespread use of stay-at-home tests may have accounted for the underreporting of the true number of COVID-19 cases. This may have also been an interesting factor to look at in comparison to the other two main time series, deaths and hospitalizations.

Overall this analysis demonstrated that COVID-19 data is difficult to handle, especially when taken as a whole. The data has been affected by changes in policy, changes in measurement procedures, and even changes in the virus itself. This makes for a situation where applying a single model naively will rarely yield a wealth of useful information, and indeed the only case where that was even close to the case was for Deaths, where the leveling out of the curve in recent data caused the behavior of the model to be much more consistent. Despite the challenges

of applying a model to this data, the results shown above demonstrate the usefulness of ARIMA models to explain many of the key points about the pandemic, and they serve as a solid foundation for other techniques that seek to build more refined approaches to understanding COVID-19.

References

Chyon, F. A., Suman, M. N. H., Fahim, M. R. I., & Ahmmed, M. S. (2022). Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *Journal of virological methods*, *301*, 114433. https://doi.org/10.1016/j.jviromet.2021.114433

Singh, S., Chowdhury, C., Panja, A. K., & Neogy, S. (2021). Time series analysis of COVID-19 data to study the effect of lockdown and unlock in India. *J Inst Eng India Ser B*, *106*(6), 1275–1281. https://doi.org/10.1007/s40031-021-00585-7

Somyanonthanakul, R., Warin, K., Amasiri, W. *et al.* Forecasting COVID-19 cases using time series modeling and association rule mining. *BMC Med Res Methodol* **22**, 281 (2022). https://doi.org/10.1186/s12874-022-01755-x

Tomar, A., & Gupta, N. (2020). Prediction for the spread of covid-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, *728*. https://doi.org/10.1016/j.scitotenv.2020.138762

James, N., & Menzies, M. (2022). Estimating a continuously varying offset between multivariate time series with application to COVID-19 in the United States. *The European Physical Journal Special Topics*, *231*(18-20), 3419–3426. https://doi.org/10.1140/epjs/s11734-022-00430-y

City of Chicago. (2023, April 26). *Covid-19 daily cases, deaths, and hospitalizations: City of chicago: Data Portal*. Chicago Data Portal. Retrieved March 10, 2023, from

https://data.cityofchicago.org/Health-Human-Services/COVID-19-Daily-Cases-Deaths-and-Hospitalizations/naz8-j4nc