Association between Covid-19 Cases and Inpatient Hospital Bed Occupancy

Emma Hughson

301356242

August 10, 2020

1. Motivation and Background:

1.1 Motivation

The coronavirus pandemic, also known as Covid-19, of 2020 has swept the world by storm. Countries' economies are falling apart, and thousands of people are losing their lives because of this disease. However, most countries' terror has become more manageable. While in comparison the United States has seen many pitfalls in attempts to control the virus. Currently, some of the most populous states, such as Florida, are quickly relaxing measures to return to normal life at the expense of an increase in Covid-19 cases [1]. As shown in figure 1, the choropleth heat map illustrates the amount of covid-19 cases in each state as of July. Places like Florida and Texas have seen high amount of cases when compared to states like Washington. In addition to the number of cases, figure 2, demonstrates the amount of inpatient hospital beds being occupied by covid-19 patients. The number of inpatient beds being occupied are high in states like Florida and Texas while low in Washington. This provides insight into a possible relationship between hospital beds and confirmed cases. As cases rise do beds become more occupied with covid-19 patients?

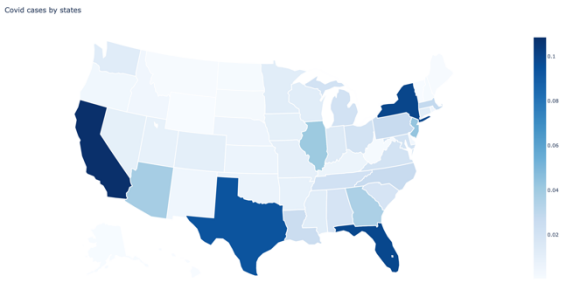

Figure 1. Heat map of confirmed Covid-19 across United States.

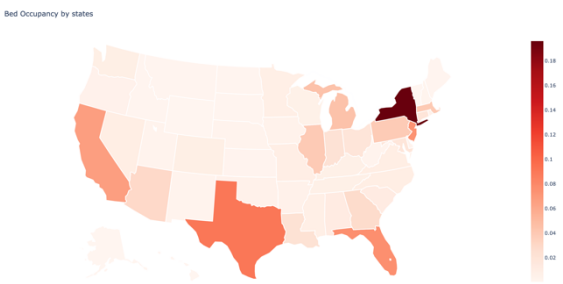

Figure 2. Heat map of bed occupancy of Covid-19 patients across United States.

Due to the novelty of such a virus, many of the worlds scientists are curious as to what Covid-19's effects are in terms of long- and short- effects. In addition, hospital bed capacity has also been a topic of concern for scientists and leaders. Hospital bed capacity is largely important because it can impact the rate of success of saving lives from the virus. If beds run out, then people have nowhere to seek treatment for their symptoms [2]. Knowing the current and future availability of beds is important because states should prepare in the event the virus gets worse in the Fall. Therefore, leaders of the United States (i.e., governors and the president) and scientists, such as Data Scientists, would be highly interested in this project. Possible use cases are leaders using the current project to assess the current state the United States is in, whether confirmed Covid-19 cases can predict hospital bed capacity in case data for hospital bed capacity is not available, and allowing other scientists to use the results of this project to further support their own results.

1.2 Related Work

As the virus is relatively new, very few peer reviewed research has been provided regarding hospital beds and Covid-19. The Center for Disease Control (CDC) does have a dashboard available that currently has information regarding hospital beds in the United States up to July 14, 2020 [3]. The CDC's dashboard provides statistics about inpatient beds occupied by all patients, inpatient beds occupied by covid-19 patients, as well as statistics on ICU beds.

According to the dashboard, as of July 13th, 64,496 patients with covid-19 are occupying inpatient beds. That is approximately 8% of the total beds across the United States.

Another study by Murray [4], that was published by medRxiv, used forecasting techniques to look at various topics of concern related to Covid-19. One of the concerns they addressed was hospital beds and the forecast of hospital beds into the next 4 months. However, the study was done back in March of 2020 and therefore, can be seen as out of date since the information revolving around the virus is constantly changing. The study reported that hospital bed occupancy will reach a peak around April and May of 2020. They predicted that the capacity of hospital beds will become scarce and hospital systems will reach its maximum capacity making Covid-19 cases hard to manage.

1.3 Problem Statement

The first question being addressed is what current state the United States is in given the amount of confirmed Covid-19 patients. Secondly, given the current state of covid-19 cases, what is the available inpatient hospital bed occupancy and what will the availability look like in the future. Furthermore, can the hospital bed capacity be predicted by confirmed cases. In other words, can the combination of hospital bed data and confirmed case data in the United States provide important insights, or predictions, into what the future looks like.

2. Datasets

I have gathered two datasets. One of the datasets is from the New York Time's Covid-19 GitHub (called "us_states.csv") dataset which addresses confirmed Covid-19 cases in the United States [5]. The other dataset (called "covid19-NatEst.csv) is from the Center for Disease Control and Prevention (CDC) about hospital bed occupancy [3]. The two datasets are related to each other because they both provide information that could potentially impact one another. For instance, as cases increase the amount of hospital beds occupied by Covid-19 patients might increase, as shown in figure 1 and 2. As well, increases in confirmed cases should also go up as bed occupancy also increase as more confirmed patients are diagnosed. Null results were removed from the datasets. The state of District of Columbia was removed from the CDC dataset. Both datasets date attributes were converted to date time values. The data was then grouped by date so that each date could have a sum of the total number of cases and bed occupancy for a given date. The data was then sorted by date to get the rows into chronological order. Certain columns were also renamed from the CDC dataset. Both datasets will be merged on their state and date attributes. Merging will decrease the amount of data available from the New York Times dataset.

2.1 New York Times Dataset

The New York Times is the cumulative number of confirmed Covid-19 cases each day for each state in the United States [5]. It is updated every day and provides up to date information from January 21st to July 25th, 2020. Historical data was used as it contained a complete count for a given day, while live count may have partial counts. Data is available on a county level, state level and country level. The state level data was used because it drilled down on the statewide comparison.

The data has 5 columns: date, state, fips, cases, and deaths. For the current analysis, date, state, and cases are used. The date attribute is the date the data was collected, state attribute was which

state the data was collected for, and the cases attribute is the number of cumulative cases that day for a given state. The data is collected by journalists who extract information from news conferences, data releases, and public officials. The creator(s) of the dataset also noted that government official constantly change case numbers and provide inconsistent information [5]. Confirmed patients are counted by where they are seeking treatment and are counted only if they have been confirmed a laboratory test and is reported by some level of government agency.

2.2 CDC Dataset

The CDC dataset can be found on the CDC's website under current hospital capacity estimates [3]. This page is dedicated to Covid-19 related data and focuses on the current inpatient and ICU bed occupancy estimates. The data is submitted by hospitals in the United States to the NHSN COVID-19 Module and is weighted and imputed to account for non-responses and missing data. Collection occurred between April 4th to July 7th, 2020. The data comes with the estimated number of beds being occupied by Covid-19 patients in both inpatient and ICU areas of hospitals, confidence intervals, and the amount of inpatient and ICU beds occupied by all patients. The project will focus on only inpatient beds not ICU beds. The attributes used for the analysis are 'Day for which estimate is made' (renamed date), 'Number of patients in an inpatient care location who have suspected or confirmed COVID-19, percent estimate (percent of inpatient beds)' as well as, 'Number of patients in an inpatient care location who have suspected or confirmed COVID-19, estimate' (renamed bed amount). The current data for hospital beds is only reported up to July 14th, 2020 and has not been updated since.

3. Methodology

The task at hand is to use various forecasting tools, along with regression models to analysis the association between hospital beds and confirmed Covid-19 cases. I plan to use two datasets one contains hospital bed data and the other contain confirmed case data and combine them. I will first assess what each data sets forecast looks like by using ARIMA. Then I will use regression models to assess the relationship between Covid-19 cases and Bed Occupancy. Finally, I will use a VAR model to forecast Bed Occupancy using Covid-19 data.

3.1 Forecasting Models

Forecasting is a data mining method which uses time series data to assess trends and growth rate to determine future outcomes. There are several forecasting models available, but two are used for this data analysis, ARIMA and VAR. Forecasting models primarily rely on either past data of the attribute being predicted or uses a combination of past data along with another predictor variable. The goal is to predict the target variable with the best predictor variables available to make the model more reliable [6]. More specifically, using forecast models to predict bed occupancy, confirmed Covid-19 cases, and predicting bed occupancy based on confirmed Covid-19 cases, as well as lagged values. In addition, the correct forecasting model is necessary to get more reliable results. As Covid-19 is relatively new, models that rely on seasonal trends should be avoided as it is too soon to understand the seasonal trend of the virus. For both models, mean absolute percentage error (MAPE) scores will be used to evaluate the forecasting models to calculate accuracy [7]. Root Mean Squared Error (RMSE) will also be used to assess root average difference between predicted values and actual values [14].

Initially, ARIMA, which stands for Auto Regressive Integrated Moving Average, was conducted on both datasets separately. ARIMA is used on univariate data and uses the available time series' past values to forecast future values. Other than the time series being univariate, the data should be non-seasonal. ARIMA is calculated using three values: the first value, p, is the Auto Regression; the second value, q, is the Moving Average; finally, the third value, d, is the total amount of differencing which makes the data stationary. The term stationary is used to describe the dataset when the predictors are independent of one another [7]. The Auto Regression is a linear regression model which uses lags for prediction. Lags are the state of a data points at different times throughout the dataset. The moving average also depends on the current and past values of the data points and is calculated by using the linear combination of error terms during times in the past [8]. ARIMA was used because it gives insight into the attributes ability to predict itself. Primarily using it to see if past values of bed occupancy can be used to predict future bed occupancy. Also using it see if past values of confirmed Covid-19 cases can predict future confirmed Covid-19 cases.

In addition to ARIMA, VAR was used to combine the datasets to forecast bed occupancy levels. VAR, which stands for Vector Autoregression, is used when two or more datasets that involve time series influence on another. Similar to ARIMA, VAR uses past values of the datasets(s) to predict the future values of the dataset(s) using autoregression. VAR not only uses a combination of lag values from the target variable but also from the attributes used to predict the variable [9]. As already mentioned, it is expected that bed occupancy of Covid-19 patients should be affected by the number of confirmed cases. While the confirmed cases should also go up as bed occupancy also increase as more confirmed patients are diagnosed. The VAR model uses the past values of both the Covid-19 confirmed cases data and the CDC Covid-19 bed occupancy data to predict the number of Covid-19 bed occupancies at a given time, while also predicting confirmed cases at a given time.

3.2 Regression Models

In addition to well established forecasting models, simple Regression models were also used to assess the prediction capabilities of the two datasets. Regression is a data mining technique generally used to establish the relationship between a target variable and predictor variables [10]. In this case, the relationship between bed occupancy and confirmed Covid-19 cases. Moreover, regression models can be used for time series data and identifying casual relationships. One of the regression methods that will be used is a Linear Regression model, which will be used to estimate how much bed occupancy changes as confirmed Covid-19 cases change using a linear relationship. In addition to Covid-19 cases, the linear regression model will also use previous values of bed amount, which are called lags, to assist in prediction.

Using a linear regression model, the best fit line can be used to determine causal trends and can be mapped onto the data provided by the two datasets. The best fit line is produced by using the predictors (i.e., confirmed cases and lagged data) as a linear combination to predict bed occupancy levels. The coefficients of the line are determined by learning which combination of coefficients work best at predicting the best fit line of the given data points for bed occupancy [11]. The other regression model is Random Forest Regressor which uses an ensemble of Decision Trees. Each tree gets a random sample of the training datasets and predicts a value at the leaf of the tree. The trees will vote on which value is the best, and this value will be used to

predict the bed occupancy [12]. r^2 score will be used to evaluate the performance of the two regression models, as well as mean absolute error. r^2 score assess the goodness of fit of the model and provides evidence as to how much the predictor attributes can explain the target variable. [13] While mean absolute error measures the average absolute difference between prediction values and true values [14].

4. Evaluation

Using various forecasting and regression models, the chances of gaining valuable insight into the future of Covid-19 increases. As mentioned previously, since Covid-19 is still novel not many peer reviewed analysis have come out, especially in regard to bed data. One of the previously mentioned studies, was one by Murray [4], published by medRxiv. When this study was released, in March 2020, not a lot of data was available to make valuable predictions. They believed that the peak of the virus would be reached in April, which unfortunately is not the case. Figure 3 shows the Covid-19 cases from January to July. As the figure illustrates, the first peak occurred in April, but the second peak which was much larger than the first occurred in July. As such, given more recent data and using various modelling techniques, the current analysis should be able to capture the more recent trend in data. Using various modeling techniques and more recent data, the current analysis should be able to capture the future of Covid-19, which relatively early studies could not because of the lack of data at the time.

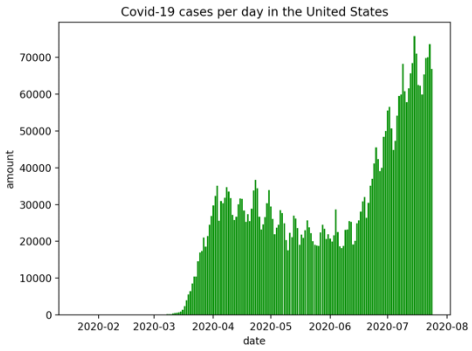

Figure 3. Distribution of confirmed Covid-19 cases a day since January 21st, 2020

5. Results

Initially, ARIMA was conducted on both datasets separately to assess the potential future trend of both confirmed Covid-19 cases and bed occupancy. Taking a closer look at bed occupancy levels (figure 4), the current model is relatively close to the validation data. The mean absolute percentage error (MAPE) score was approximately 0.0488. This gives the current model around 95% accuracy. It also had a RMSE score of 3604.5. This accuracy gives confidence that a future trend of about 50 steps (or 50 days) into the future could be highly insightful. In figure 5, the forecasted results are shown. It shows that over the next 50 days, the bed occupancy level will increase at a slow and steady incline to approximately 100,000 beds into late August.

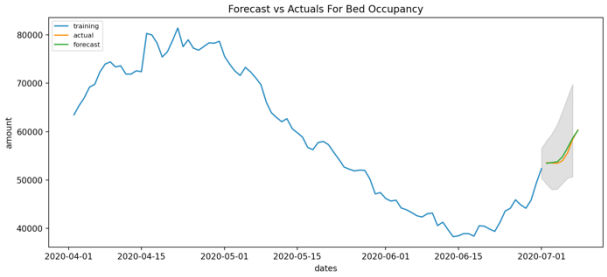

Figure 4. Forecast of bed occupancy vs. Actual bed occupancy.

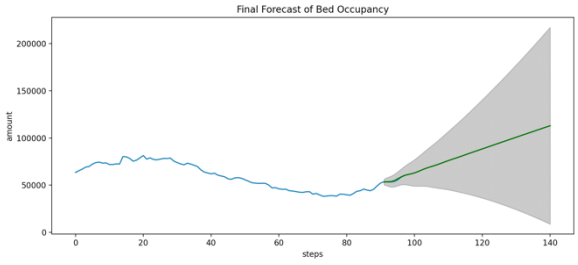

Figure 5. Forecast of bed occupancy data 50 steps into the future using a model with 95% accuracy

The ARIMA results for confirmed Covid-19 cases are shown in Figure 6. The model also performed well at predicting the validation data. The RMSE score was 5809.97 and the highest MAPE score was 0.0032, which is around 99% accuracy. This may, however, be due to overfitting of the data. Altering the ARIMA parameters, the closest second accuracy score was approximately 92% accuracy, with a MAPE score of 0.0748. Figure 7 shows the model with 99% accuracy 50 steps into the future. Relying

on this model, the confirmed cases of Covid-19 are expected to increase to approximately 80,000 confirmed cases a day in the United States. However, using the 92% accuracy, the Covid-19 cases are expected to increase to 100,000 a day.

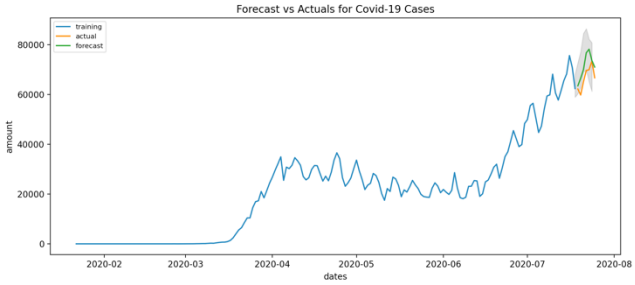

Figure 6. Forecast of confirmed Covid-19 cases vs. Actual confirmed Covid-19 cases

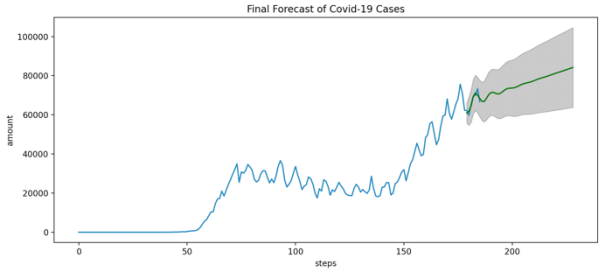

Figure 7. Forecast of confirmed Covid-19 cases 50 steps into the future using a model with 99% accuracy

After conducting ARIMA forecasting, regression models using Linear Regression and Random Forest Regressor were used to observe the relationship between Covid-19 cases and Bed Occupancy. Out of the two regression models, Linear Regression performed best, receiving a r^2 score of 0.753 and a mean absolute error of 1721. Figure 8 shows the test data is relatively close to the predicted data, when compared side by side. While, Random Forest Regressor received a r^2 score of -1.75 and a mean absolute error of 3761. The Linear Regression demonstrates that bed occupancy can potentially be predicted using confirmed Covid-19 cases along with past values of bed occupancy, up to 7 days in the past. However, using Covid-19 cases alone produced worse results for both models with a r^2 score of -153 for Linear Regression and a r^2 score of -12. Therefore, Covid-19 cases alone cannot be used to predict bed occupancy. Figure 9 shows the actual values and the predicted values of the validation dataset. The predictions values almost line up but there is still a lot of error. As such, extending the amount of data available to train the model prediction gets substantially better, with an r^2 score of 0.93 and MAE of 513 (figure 10).

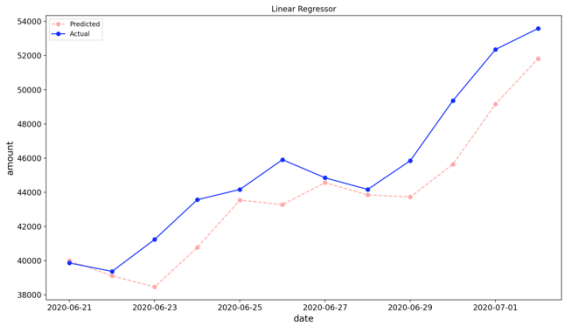

Figure 8. Linear regression of predicted values vs. actual values of bed occupancy on test dataset

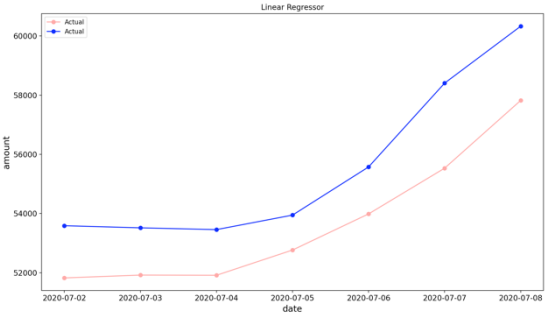

Figure 9. Linear regression of predicted values vs. actual values of bed occupancy on validation dataset

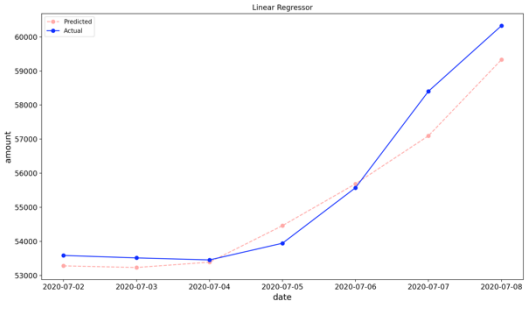

Figure 10. Linear regression of predicted values vs. actual values of bed occupancy on validation dataset combined with test dataset

Finally, VAR was conducted on the two datasets. Since Linear Regression was fairly successful at predicting bed occupancy with lag data, VAR can be used to further assess this relationship between Covid-19 cases and bed occupancy. Initially a Granger Causality test was used to assess the relationship between the two attributes. The results of the Granger Causality test produced a $p < 0.05$ between Covid-19 cases and Bed Occupancy. However, $p > 0.05$ when Bed Occupancy and Covid-19 cases were compared. After conducting Granger Causality, Ad Fuller test was used to assess the stationary state of the data.

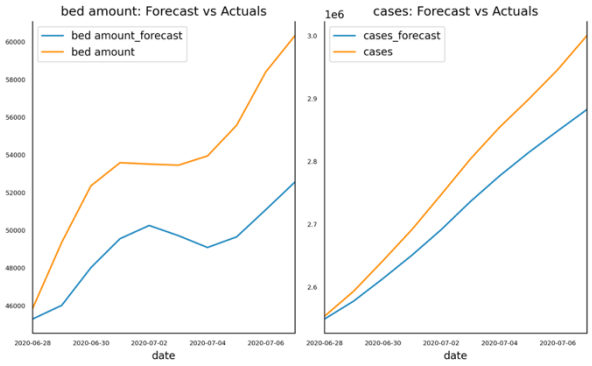

Figure 11. Forecast of bed occupancy data and confirmed Covid-19 cases vs. actual bed occupancy and confirmed Covid-19 cases

The data was initially not stationary and was, therefore, differenced twice to make it stationary. Once the data was stationary, the predictions for both Covid-19 cases and Bed Occupancy were displayed (figure 11). Figure 11 shows that the prediction for Covid-19 cases was relatively close to the validation dataset. However, the bed occupancy was not as close to the validation dataset. Furthermore, the MAPE value for Covid-19 forecasting was 0.0207 and the MAPE value for Bed Occupancy forecasting was 0.0819. As such, predicting Covid-19 cases from Bed Occupancy resulted in a 98% accuracy, which is higher than 92% accuracy produced by predicting Bed Occupancy from Covid-19 cases. However, the RMSE score for bed occupancy was 4935, while the RMSE score was 68548 for Covid-19 forecasting. Which means that the error for Covid-19 forecasting was higher than bed occupancy.

In conclusion, the amount of confirmed Covid-19 cases may not help predict Bed Occupancy. However, using ARIMA, the best Bed Occupancy results were produced showing that using past values of Bed Occupancy should be used to predict future Bed Occupancy. While, Covid-19 cases may be somewhat related to Bed Occupancy, as demonstrated by the Linear Regression, it should be used with hesitancy as there seems not to be a strong causal relationship as originally thought. It is important to note that for the states with the highest confirmed case load, their hospital bed occupancy seems to be going down as their number of cases continue to increase (figure 12, 13). This might explain the lack of predictive power confirmed Covid-19 cases have on predicting bed occupancy. One thing that can be taken away from using the VAR model is that forecasting Covid-19 cases can be helped by using Bed Occupancy. Figure 14 shows the predicted bed occupancy and errors from VAR, as well as predicted Covid-19 cases and errors. It seems that Covid-19 is going to continue a similar trend into the future,

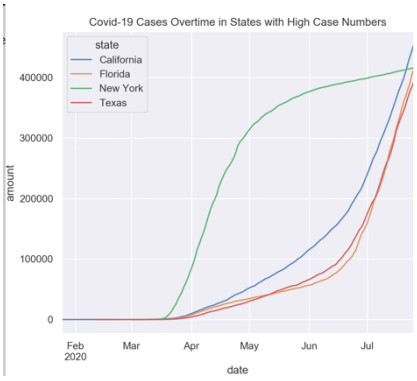

Figure 12. Covid-19 cases over time in the 4 most affected states

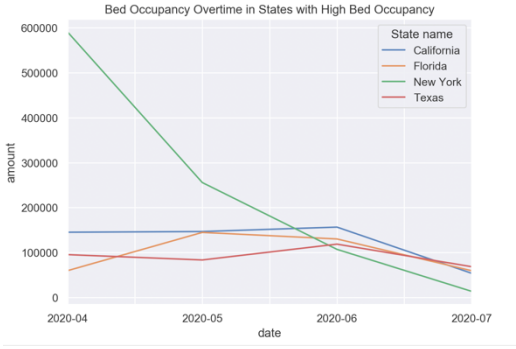

Figure 13. Bed's occupied by Covid-19 cases over time in the 4 most affected states

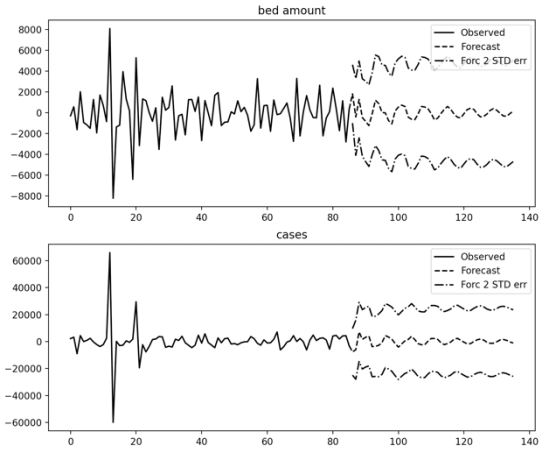

Figure 14. Predicted bed occupancy and Covid-19 cases 50 days into the future using VAR.

while bed occupancy will decrease. However, these forecasting models should be taken with caution as the RMSE scores were high.

In addition to these findings, I learned that forecasting models are very specific to certain types of data. For instance, using a SARIMAX model on this data would not provide good insights as a seasonal trend has not yet been established. As such, using the right forecast model is important in order to produce relatively reliable results. Overall the best model would be the ARIMA on predicting bed occupancy using lagged values instead of using Covid-19 cases. The reason ARIMA is better is because not only did it have lower RMSE scores, it also had high accuracy for predicting bed occupancy. The relationship between bed occupancy and confirmed Covid-19 cases will need to be explored further as there does seem to be somewhat of an inverse relationship and this could provide powerful insight for governors, as well as scientists.

6. Summary

Altogether the results produced by the forecasting and regression models provide insight into what could potentially happen at the beginning of Fall. It is expected, according to the ARIMA model that confirmed Covid-19 cases per day is expected to increase, while bed occupancy is also expected to increase. However, bed occupancy cannot be determined confidently by Covid-19 cases as shown by the VAR model. Although the Linear Regression did show that there might be some relationship, it looks more likely that Bed Occupancy is better predicted using previous Bed Occupancy, while confirmed Covid-19 cases can be potentially predicted by Bed Occupancy. It is important to note however, that the data used to build these models is small. More data is needed to make a more accurate prediction, but the current data provides more insight than the previous amount of data available during the study done by medRxiv in March [4]. Further assessment is required as more data is collected.

In addition, the results from this study, although preliminary, can provide support for others' predictions. In other words, if other scientists find that they have similar results, this project can provide more support that their results are reliable. Furthermore, leaders can use the results from this project to assess the current state of Covid-19 in the United States. Forecasting methods provide support that the virus, given the current situation, is only going to get worse into the fall. However, bed occupancy cannot be reliably predicted from confirmed Covid-19 cases. This may be due, as already mentioned, to the possible inverse relationship present between Covid-19 cases and Bed Occupancy.

References

[1] C. Maxouris, "In the US, five states account for more than 40% of the country's nearly 5 million Covid-19 cases," *CNN*, 09-Aug-2020. [Online]. Available: https://edition.cnn.com/2020/08/09/health/us-coronavirus-sunday/index.html. [Accessed: 29-Jul-2020].

[2] M. Sanger-katz, S. Kliff, and A. Parlapiano, "These Places Could Run Out of Hospital Beds as Coronavirus Spreads," *The New York Times*, 17-Mar-2020. [Online]. Available: https://www.nytimes.com/interactive/2020/03/17/upshot/hospital-bed-shortages-coronavirus.html. [Accessed: 01-Aug-2020].

[3] “COVID-19 Data Dashboard - Hospital Capacity Snapshot,” *Centers for Disease Control and Prevention*, 16-Jul-2020. [Online]. Available: https://www.cdc.gov/nhsn/covid19/report-patient-impact.html. [Accessed: 19-Jul-2020].
[4] “Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator days and deaths by US state in the next 4 months,” *Institute for Health Metrics and Evaluation*, 21-Apr-2020. [Online]. Available: http://www.healthdata.org/research-article/forecasting-covid-19-impact-hospital-bed-days-icu-days-ventilator-days-and-deaths. [Accessed: 01-Aug-2020].
[5] Nytimes, “nytimes/covid-19-data,” *GitHub*. [Online]. Available: https://github.com/nytimes/covid-19-data. [Accessed: 13-Jul-2020].
[6] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. Heathmont, Victoria: OTexts, 2018.
[7] S. Prabhakaran, “ARIMA Model - Complete Guide to Time Series Forecasting in Python: ML ,” *[[Machine Learning Plus]]*, 28-Apr-2020. [Online]. Available: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/#:~:text=ARIMA, short for 'Auto Regressive,used to forecast future values. [Accessed: 25-Jul-2020].
[8] “Autoregressive integrated moving average,” *Wikipedia*, 05-Jul-2020. [Online]. Available: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average. [Accessed: 25-Jul-2020
[9] S. Prabhakaran, “Vector Autoregression (VAR) - Comprehensive Guide with Examples in Python,” *[[Machine Learning Plus]]*, 27-Apr-2020. [Online]. Available: https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/. [Accessed: 28-Jul-2020].
[10] Sunil RayI am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years., “Regression Techniques in Machine Learning,” *Analytics Vidhya*, 15-Apr-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/#:~:text=Regression analysis is a form,effect relationship between the variables. [Accessed: 01-Aug-2020].
[11] J. Brownlee, “Linear Regression for Machine Learning,” *Machine Learning Mastery*, 12-Aug-2019. [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/. [Accessed: 05-Aug-2020].
[12] G. Drakos, “Random Forest Regressor explained in depth,” *GDCoder*, 05-Feb-2020. [Online]. Available: https://gdcoder.com/random-forest-regressor-explained-in-depth/. [Accessed: 08-Aug-2020].
[13] J. Frost, “How To Interpret R-squared in Regression Analysis,” *Statistics By Jim*, 16-Jul-2020. [Online]. Available: https://statisticsbyjim.com/regression/interpret-r-squared-regression/. [Accessed: 10-Aug-2020].
[14] Jj, “MAE and RMSE - Which Metric is Better?,” *Medium*, 23-Mar-2016. [Online]. Available: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d. [Accessed: 10-Aug-2020].