Forecasting COVID-19 Hospitalizations with Wastewater

Ryan Zhang advised by Dr. Will Townes

Abstract

In this project, we aimed to determine whether or not wastewater could help us forecast COVID hospitalizations. We used an ARIMA(2,1,0) and a quantile regression model with the same orders. We used these as baseline models and compared the resulting predictions when we added lag 2 and lag 3 wastewater covariates. We found that root mean squared error appeared to slightly decrease when we included wastewater covariates, but an F-test on the ARIMA model with wastewater covariates and the corresponding ARIMA model without wastewater covariates failed to reject the null (p=0.8105). We also noticed that our model tended to perform the worst when the response time series variable steeply increased, which is also when hospitals are most at risk.

1 Problem

Occasionally hospitals get overwhelmed by surges in patient volume. As a result, the quality of care provided by these hospitals may be diminished, or they may need to turn patients away. A prominent example in recent years was the COVID-19 pandemic. In order to ameliorate this problem, it is important to predict when these surges will occur so that hospitals can prepare accordingly.

Our main interest in this project is to determine whether or not wastewater data may be beneficial as a covariate to improve patient volume forecasts.

2 Data

For this project, we will only use data concerning Pennsylvania. Our wastewater data comes from Centers for Disease Control and Prevention [2]. Our response variable is "pcr_conc_lin", which is the normalized SARS-CoV-2 virus concentration detected in the wastewater. There were two normalization methods: flow-population and microbial. We decided to focus only on data that was normalized using flow-population, since there were rather few data points using microbial normalization and the units wouldn't be comparable between normalization types. We also had 34 unique wastewater plants identified by their "key_plot_id". We could either include each wastewater plant as a separate covariate, or aggregate these wastewater plants into a single summarized covariate. To decide which method to go forward with, we decided to plot the data availability of each wastewater plant over time. This can be seen in figure 1, where a black dot or line means we have data for that day or time span for the given plant on the y-axis. As we can see, a lot of plants didn't start sampling until halfway through our entire date range. As such, if we were to include each wastewater plant as a separate covariate, we would need a way to handle the large amount of missing data. Consequently, we decided to aggregate all 34 plants into a single covariate.

Another issue we found was that the wastewater data was sampled in an alternating schedule, with samples being taken at intervals of 2 days and then 5 days, resulting in an alternation of 2-day and 5-day gaps between each sample. On the other hand, we had hospitalization data for every day. In order to avoid excessive imputation on the daily scale, we decided to aggregate our data to the

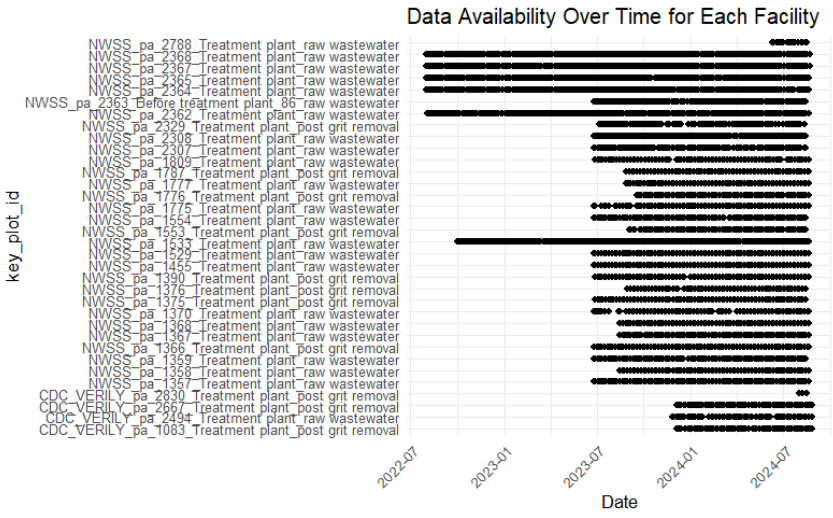
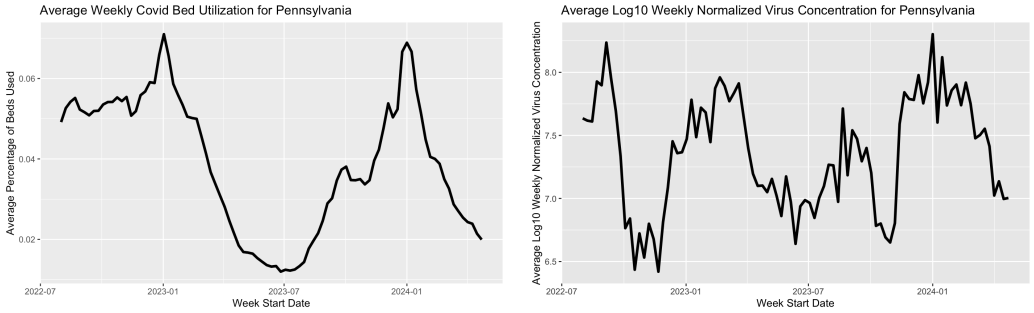


Figure 1: Data availability for each wastewater plant.

weekly scale. We would do this by cutting the data from Monday to Sunday, and the value for that week was the arithmetic mean over all the data present within that week.

Ultimately, we decided to first aggregate each wastewater plant to the weekly level, and then for each week we took an arithmetic mean of all the wastewater plants with data for that week. We ended up with a single wastewater value for every week. Finally, the data appeared to be skewed towards larger numbers, so we decided to apply a log-10 transformation at the very end.

Our hospitalization data was provided by the COVID-19 Hospitalization by State API from Carnegie Mellon University Delphi Group [1]. In particular, we are interested in the variable: ”inpatient_bed_covid_utilization”. This is the percentage of total inpatient beds currently utilized by patients who have suspected or confirmed COVID-19 in Pennsylvania. However, this number only accounts for hospitals that report both ”inpatient_beds_used_covid” and ”inpatient_beds”, which are used to get our utilization percentage. This data was once again aggregated to the weekly level from the daily level using the same methodology seen with our wastewater data. The dates in which we had both wastewater data and hospitalization data were from August 1st, 2022 to April 22nd, 2024. The final time series can be seen in figure 2.



(a) Weekly Bed Utilization Time Series.

(b) Transformed Wastewater Time Series.

Figure 2: Data after transformations

3 Determining Model Orders

While not necessary since our project was more so to compare equivalent models with and without wastewater covariates, we decided to perform some exploratory data analysis to decide our time series model orders such as the autoregressive order, the moving average order, and the differencing.

Note that this analysis was done on the daily bed utilization data as opposed to the weekly, since it was performed at the onset of the project. Future work could involve using the weekly data to perform this same analysis.

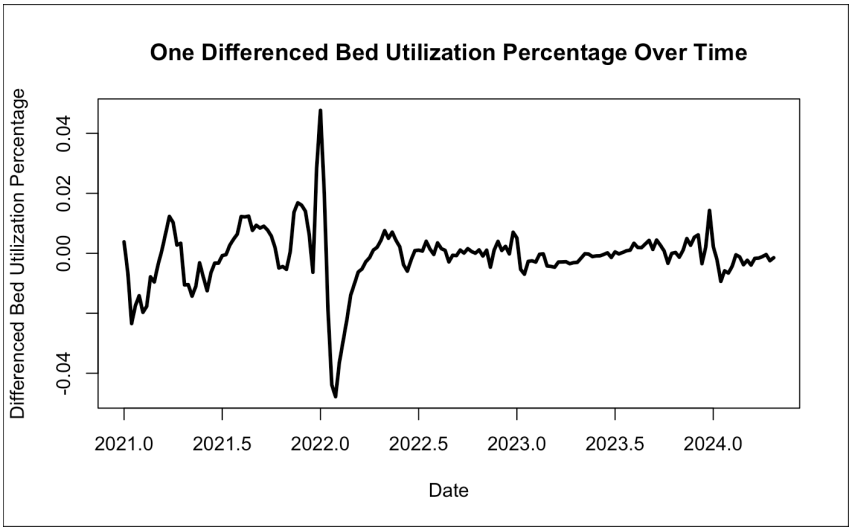
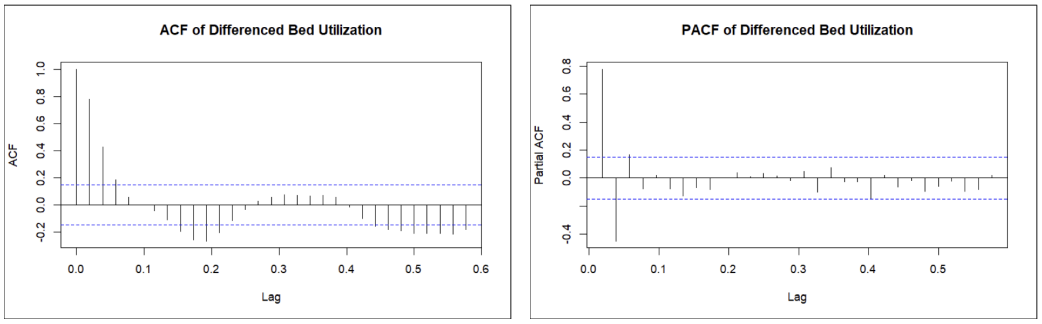


Figure 3: One Differenced Weekly Bed Utilization.



(a) ACF. (b) PACF.

Figure 4: PACF and ACF of the one differenced bed utilization time series.

Figure 3 shows the weekly bed utilization time series after being differenced once. We can see that the mean appears to be reasonably centered around 0. However, there does appear to be heteroscedasticity, suggesting a GARCH model. However, to keep the models simple, we decided to go forward with an ARIMA(P,1,Q) model. To determine P and Q, we plotted the ACF and PACF of the differenced time series in figure 4. Here we see that the PACF seems to cut off after lag 2. It does appear that the third lag is just barely significant, but we opted to choose the simpler, lower order model. It also appears that the ACF is somewhat geometrically decaying. This is evidence for a second order auto-regressive model. Thus, we will be using an ARIMA(2,1,0) model in this project. We also wanted to test our data with other models, settling on an auto-regressive quantile regression model (tau=0.5). We also arbitrarily decided to use lag 2 and lag 3 wastewater covariates.

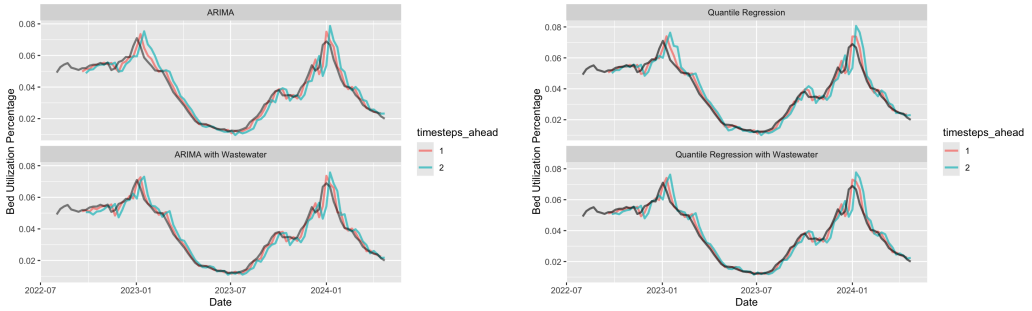
Ultimately, we end up with 4 models: ARIMA(2,1,0), ARIMA(2,1,0) with lag 2 and lag 3 wastewater covariates, Quantile(2,1,0), and Quantile(2,1,0) with lag 2 and lag 3 wastewater covariates. Note that in order to implement an auto-regressive quantile regression model, we simply lagged the response time series variable and added them as covariates in a tabular dataset.

4 Prediction Methodology

We first created a list of model fits, each trained on an increasing subset of data. We decided to have our first model train on the first 7 weeks. Every model after that was trained on the first n weeks, where we incremented n by one from 7 until the end of the dataset. We essentially tried to replicate how the model may be trained and used in practice, where we get to retrain the model every week on the previous week's data.

After obtaining this list of model fits, we obtained 1-step-ahead and 2-step-ahead forecasts for each model. These forecasts would then be compared to the actual value for that day, giving us the root mean squared error.

5 Results



(a) ARIMA forecasts vs. original values.

(b) Quantile forecasts vs. original values.

Figure 5: Forecasts vs. Original Values

In figure 5, we plotted the forecasted values alongside the original values, where the original values are plotted in black. We can see that the 1-step-ahead and 2-step-ahead forecasts lag slightly behind the original line, especially at the peaks. Whereas the original data reverses direction, the models continues to predict an increase, and doesn't adjust until after observing more data. We also wanted to compare the RMSE of the models without wastewater and the models with wastewater covariates. In addition, we want to see how the model performs during stable patient volume, decreasing patient volume, and steeply increasing patient volume.

If we look at figure 6, there are two important things to notice. Firstly, the models with wastewater are the green and purple bars. For the most part, the bars are slightly lower than their non-wastewater counterparts. Secondly, we can see that for 1-step-ahead, the model appears to perform the worst from July 2023 to January 2024. This is not ideal since this date range corresponds to the steep incline in bed utilization percentage, meaning our model performs the worst when it matters most.

Figure 7 shows us how the absolute error changes over time, and we can see that our model has the highest absolute error around both peaks, also suggesting that our model performs worst when the hospitals are at high occupancy.

Model	Multiple R-squared
ARIMA	0.2579
ARIMA w wastewater	0.2607

Table 1: Multiple R-squared for ARIMA models

Finally, since our ARIMA model was implemented as a linear model, we can extract the multiple R-squared value to see how much variance is explained by our model. Table 1 shows that we are explaining around one-fourth of the total variance in the data. Finally, to see whether or not our additional wastewater covariates are significant, we decided to perform a partial F-test on the ARIMA model without wastewater and the ARIMA model with wastewater. The results can be seen in figure 8, where we find a p-value of 08105, meaning we fail to reject the null that the wastewater covariate coefficients are nonzero.

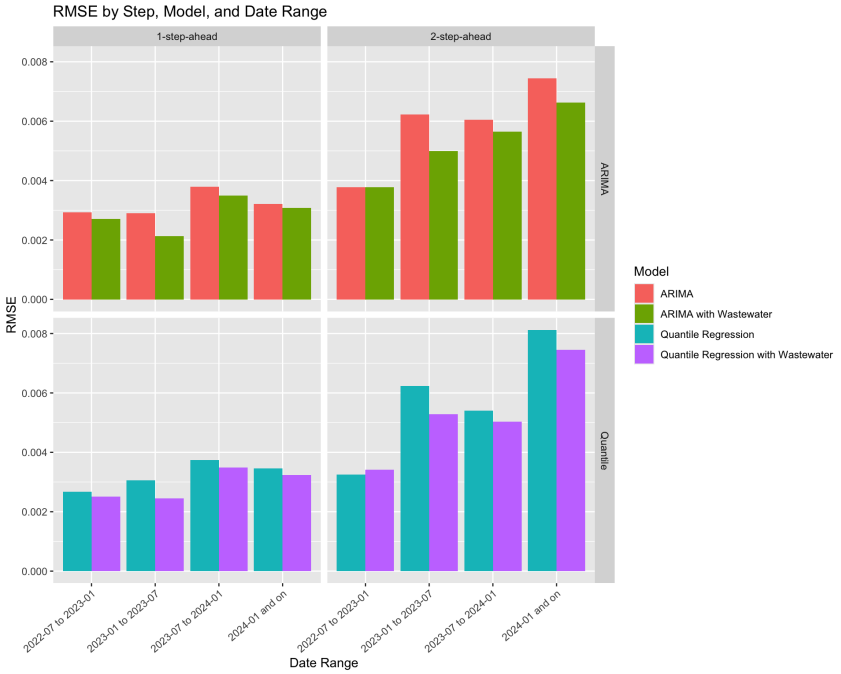
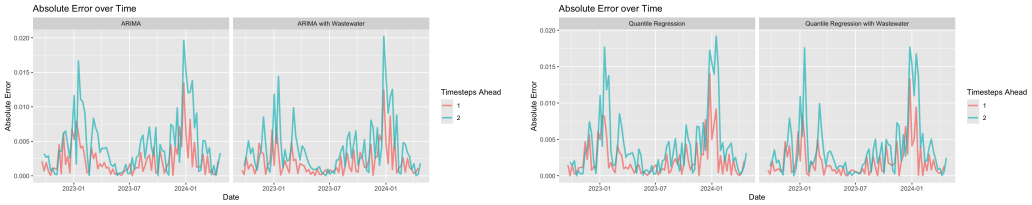


Figure 6: RMSE by step, model, and date range.



(a) ARIMA.

(b) Quantile.

Figure 7: Absolute error over time.

```
Analysis of Variance Table

Model 1: week_avg_cov_util_lag0 ~ week_avg_cov_util_lag1 + week_avg_cov_util_lag2
Model 2: week_avg_cov_util_lag0 ~ week_avg_cov_util_lag1 + week_avg_cov_util_lag2 +
    week_avg_pcr_conc_lin_lag2 + week_avg_pcr_conc_lin_lag3
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1     84 0.00074371
2     82 0.00073991  2 3.8013e-06 0.2106 0.8105
```

Figure 8: Partial F-test to determine significance of wastewater covariates in an ARIMA model.

6 Further Work

As previously mentioned, one may want to redo our analysis for determining our model orders on the transformed weekly data as opposed to the original daily data. We are also curious as to how our results may change if we weight the value of each wastewater plant by its population served, where wastewater plants serving larger populations may be more representative of the state. We may also want to explore methods for including each of the 34 wastewater plants as separate covariates instead of summarizing them into a single covariate. Finally, we would like to see if our results are similar in other states.

References

[1] Carnegie Mellon University Delphi Group. *COVID-19 Hospitalization Data API*. `https://cmu-delphi.github.io/delphi-epidata/api/covid_hosp.html`. Accessed: 2024-12-05. n.d.

[2] Centers for Disease Control and Prevention. *NWSS Public SARS-CoV-2 Concentration in Wastewater*. `https://data.cdc.gov/Public-Health-Surveillance/NWSS-Public-SARS-CoV-2-Concentration-in-Wastewater/g653-rqe2/about_data`. Accessed: 2024-12-05. n.d.