Covid-19 Wastewater Surveillance Prediction
Author: Nantawat Samermit

Executive Summary:

Problem:

The Sars-CoV-19 or 'Covid-19' virus is an ongoing pandemic. Businesses and communities are being affected with each wave and each new variant. Wastewater surveillance is a non-invasive way to measure mRNA concentrations of the virus.

Exploratory Research Question:

Can I use the viral concentration surveilled from wastewater to forecast a trend of Covid-19 infections in a community?

Methods:

1. Using time-series ARIMA model to forecast the concentration of Covid-19 mRNA in surveilled wastewater.

Results:

1. Due to stationarity, various ARIMA models must be implemented case-by-case per county-level dataset
2. Once an optimized ARIMA model is found using supervised machine learning, a forecast of the trend of wastewater samples containing Covid-19 mRNA can be achieved. But it does not have high precision.
3. Instead of focusing on high-precision, advised by the mentor to focus on the slope of that trend line to use as an indicator.

Implications:

I believe that using a forecast of Covid-19 mRNA concentration in wastewater could be used as an indicator to the presence of Covid-19 in that area. However this is with the following assumption:
1. Wastewater will be continuously surveilled
2. Viral shedding persists into wastewater with any future dominant variants

Ideas for Further Research:

- Add exogenous features to time-series such as:
    - Travel log
    - Vaccination rate in the county
- Windowing functions as a means to gradient boost the model

Potential Client Usage:

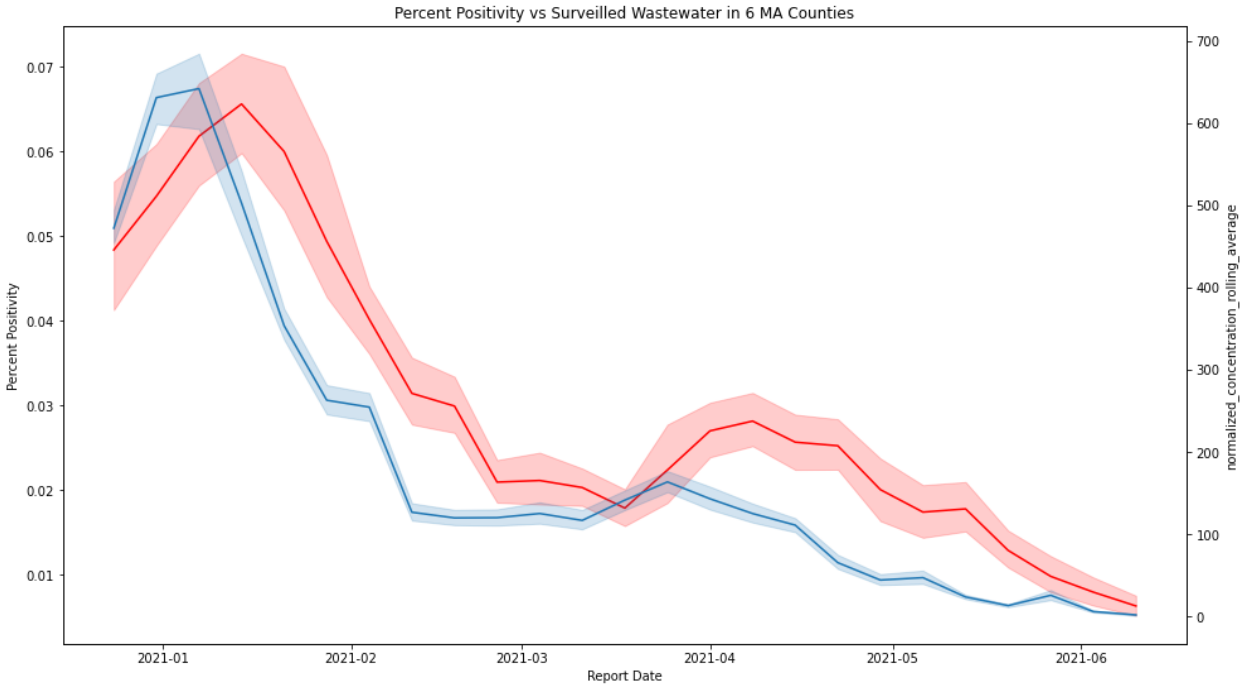
- Advance notice to hospitals to prepare their ICU staff for a surge in the community
- Advance notice to local governments as to where they should concentrate their vaccinations/testing services or mortuaries.
- Advance notices to local businesses of their risk of infection within their locale.

Data:

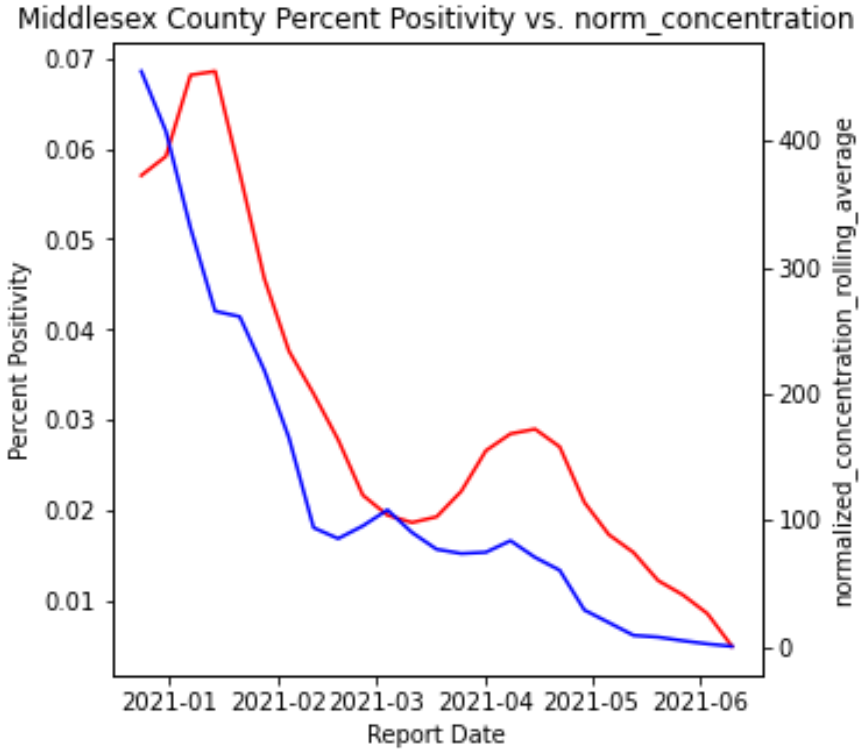
1. Cambridge MA wastewater data archive
    a. 'https://data.cambridgema.gov/resource/ayt4-g2ye.json'
2. Massachusetts Department of Health
    a. 'https://www.mass.gov/doc/covid-19-raw-data-june-15-2021/download'
3. Biobot's Github
    a. 'https://github.com/biobotanalytics/covid19-wastewater-data/blob/master/wastewater_by_county.csv'

Findings:

- The trend of the Covid-19 positivity percentage seems to follow after the surveilled wastewater mRNA measurements. (Positivity Percentage in Red).



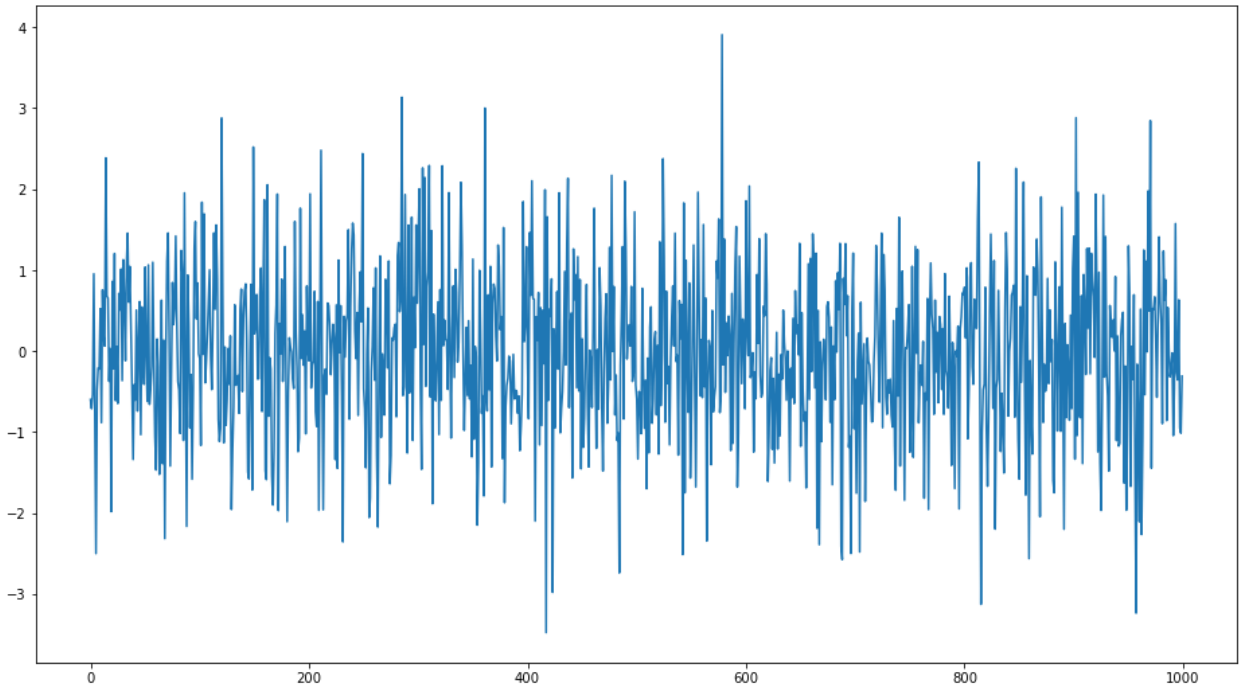
- This can be further broken down to see that the trend does follow at the county-level data as well. This is an example from Middlesex County, MA.



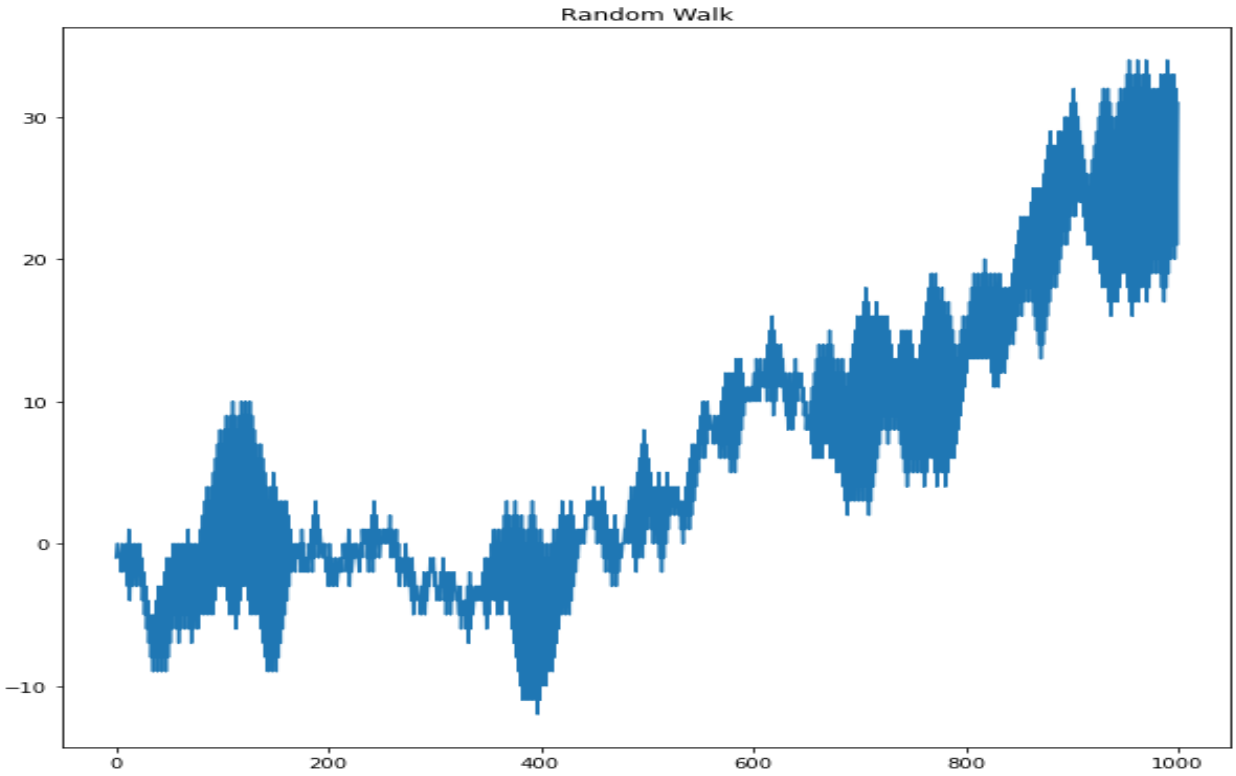
- At the county levels, there are stronger correlations seen between the wastewater samples and the Percent Positivity. Here are there correlation scores:
  - Berkshire: 0.240292
  - Essex: 0.675854
  - Hampshire: 0.392312
  - Middlesex: 0.642016
  - Nantucket: 0.777101
  - Suffolk: 0.779547
- I choose to use a time-series of the wastewater data because:
  - the Percent Positivity follows after the wastewater data
  - It is less invasive to sample wastewater data
  - Wastewater data does not need voluntary participation to sample (everyone poops)
- However, stationarity becomes an issue when wanting to use wastewater data, both at the aggregated state level, and clearly seen at the county level. Here's a comparison of white noise - stationary, random walk - non-stationary, and the county-level wastewater data.

Covid-19 Wastewater Surveillance Prediction
Author: Nantawat Samermit

- White Noise:

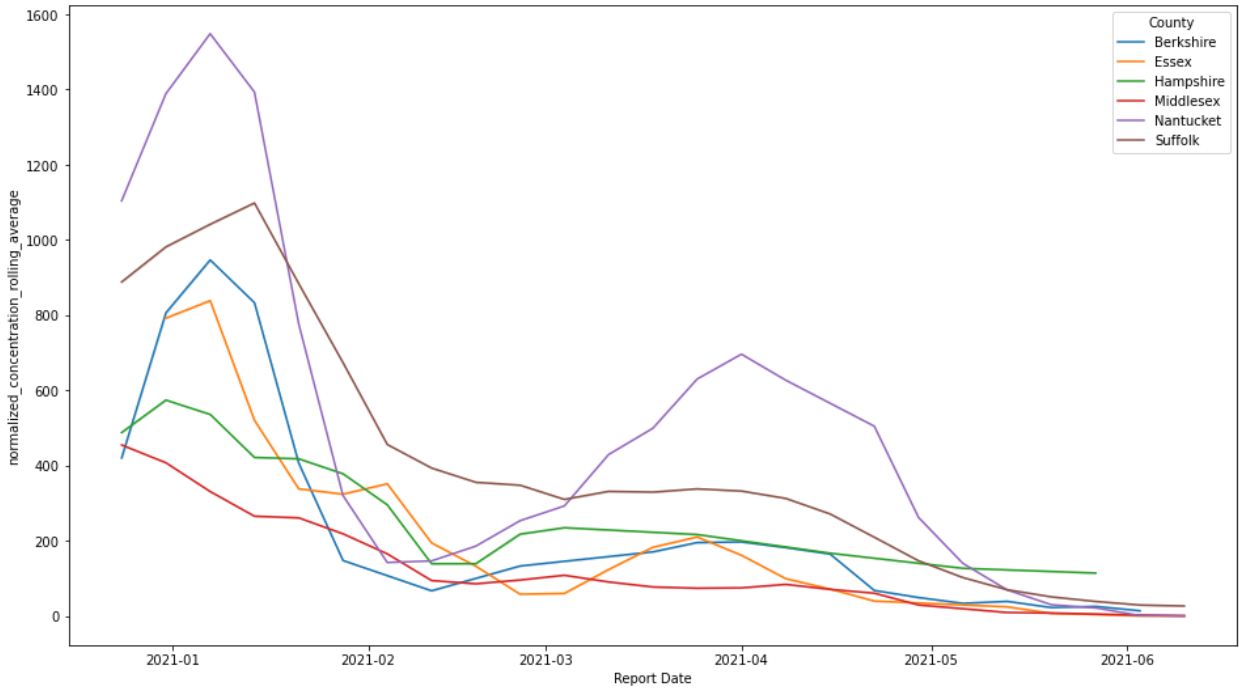


- Random Walk:

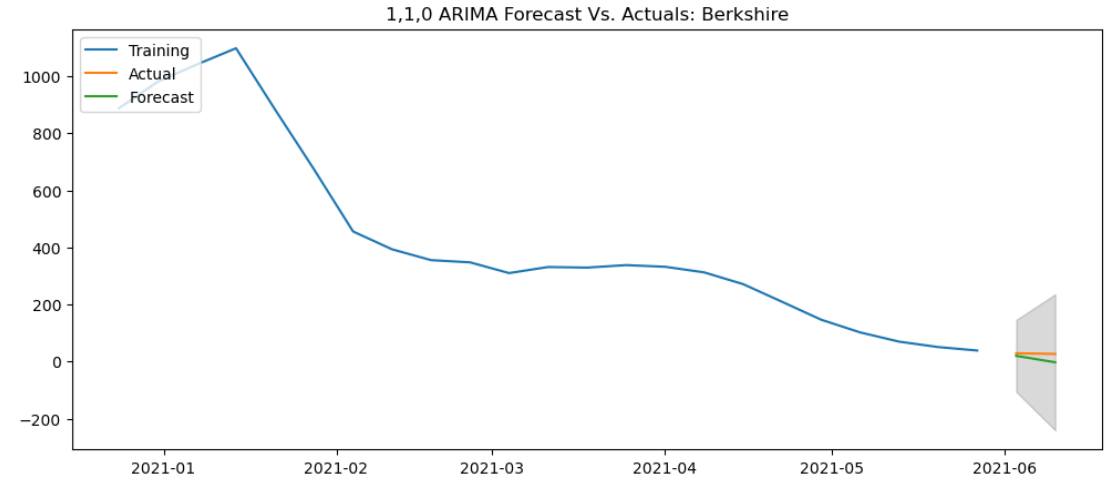


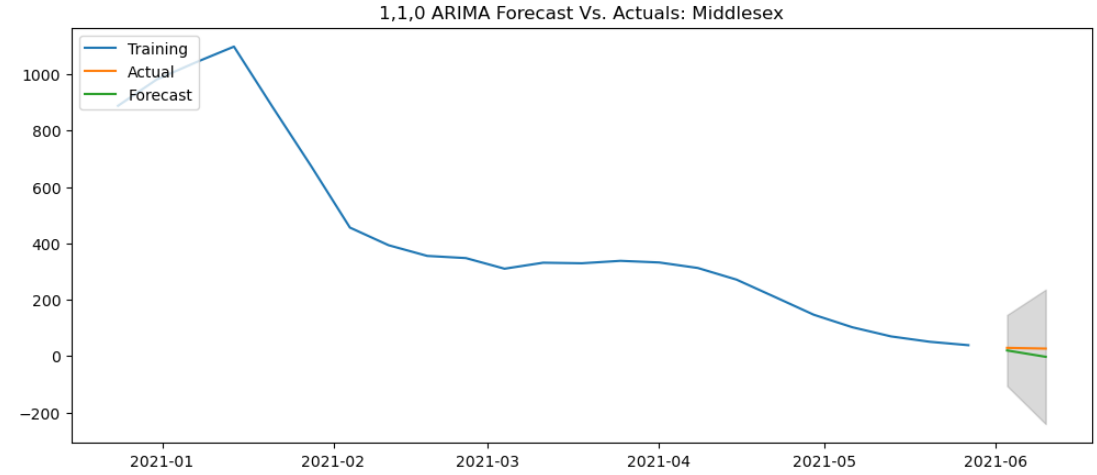
Covid-19 Wastewater Surveillance Prediction
Author: Nantawat Samermit

- County-level wastewater data:



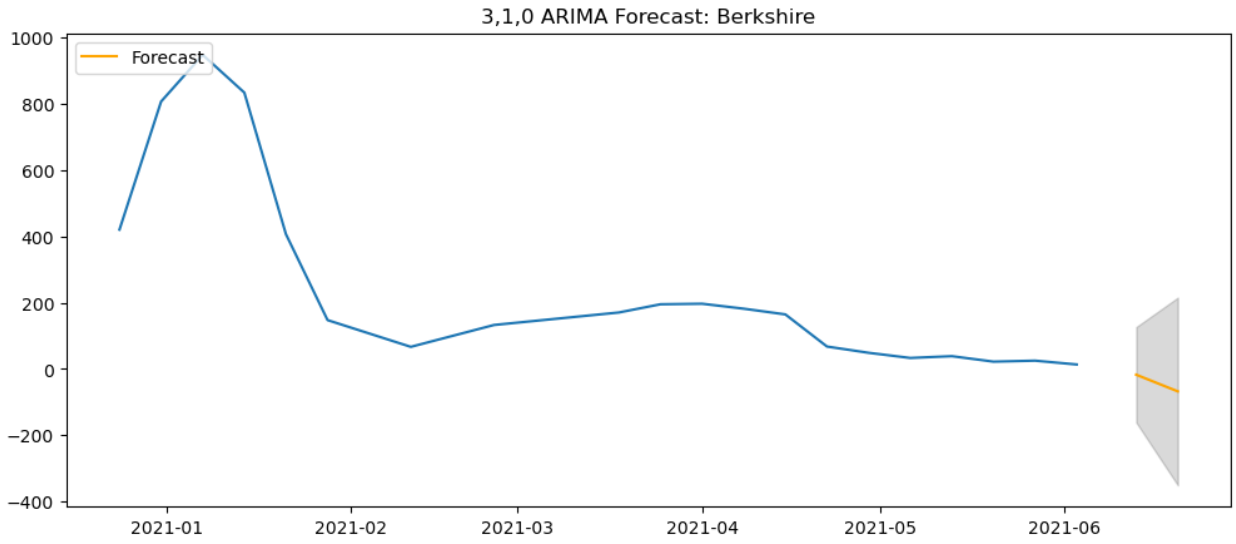
- I used the Augmented Dickey Fuller Test to check for stationarity at the county level. Many county datasets exhibited non-stationarity:
    - Berkshire {'test-statistic': -8.14872810372554, 'p-value': 9.826948680796276e-13}
    - Essex {'test-statistic': -1.5389302303555474, 'p-value': 0.5141728049273325}
    - Hampshire {'test-statistic': -1.1417617418031638, 'p-value': 0.6981338694357283}
    - Middlesex {'test-statistic': -1.7064520585626328, 'p-value': 0.4277912857757429}
    - Nantucket {'test-statistic': 0.005477996686873882, 'p-value': 0.9589827747521857}
    - Suffolk {'test-statistic': -3.030362872805752, 'p-value': 0.032154544360425225}
- For those counties exhibiting non-stationarity, I had to pre-process them using differencing.
- Using a grid-search method for ARIMA models, the supervised machine learning model found optimized parameters based on these specifications:
    - Start AR term: 1
    - End AR term: 3
    - Start MA term: 1
    - End MA term: 3
    - Start Differencing Term: 1
    - Test Parameter: Augmented Dickey Fuller
    - Seasonality: False
- Auto_Arima() grid search yielded these models for these counties:
    - Berkshire: ARIMA(3,1,0)
    - Essex: ARIMA(0,1,0)
    - Hampshire: ARIMA(0,1,1)

    - Middlesex: ARIMA(1,1,0)
    - Nantucket: ARIMA(2,1,0)
    - Suffolk: ARIMA(1,1,0)
- The model metric used was the Mean Average Precision Error: 0.70935
    - This is not an accurate model, but as discussed with my mentor, when forecasting trends sometimes the accuracy isn't the focus but rather the trend.
    - 

    - 

    - These forecast were using in-sample data to determine the accuracy of the model

- Here is an out-of-sample forecast:



Special Thanks:

I'd like to notate thanks specifically to the two who influenced this project the most. First being my mentor, Luka Anicin, who encouraged me to try whatever I could. This was a crazy first project and I couldn't have done it without his guidance and encouragement.

The second being Dr. Patrawat Samermit. She helped me keep it real. Thanks Sis.