Novel disease prediction

By Douglas Domingo

The problem

Can you use sewer water to predict an outbreak?

The Solution

Use past COVID-19 spikes with historical waste water data to predict future outbreaks

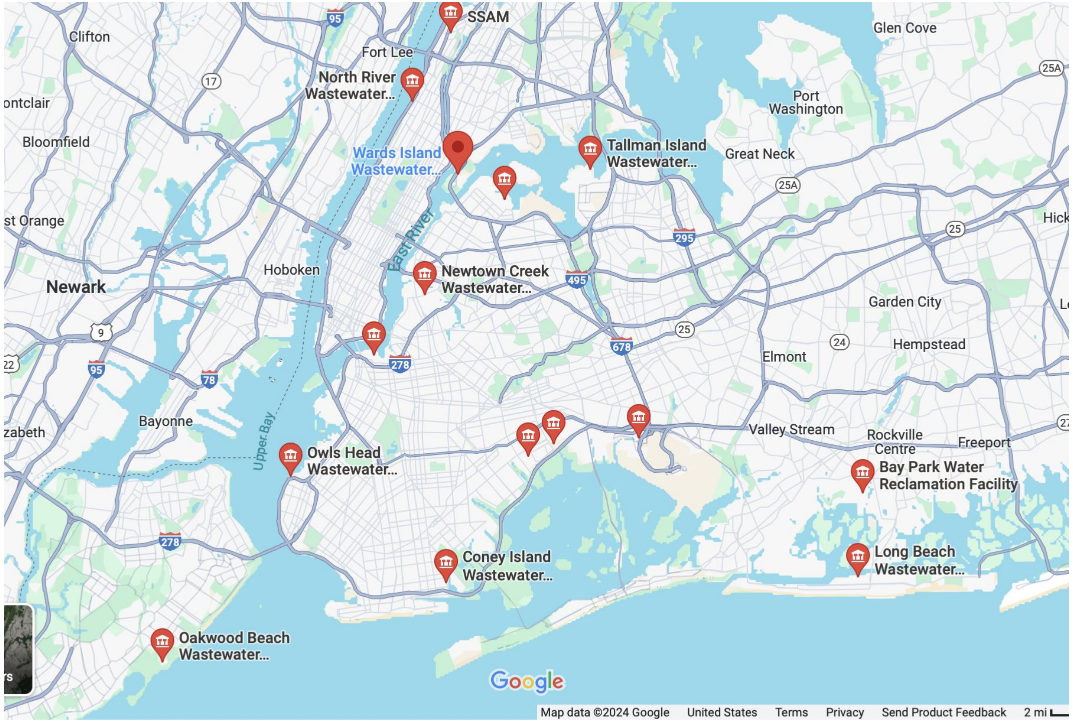
Waste water locations

NYC offers a variety of data it collects open to the public

14 collection sites in NYC covers all neighborhoods in NYC

Samples are taken twice a week

Samples are looking for genetic material not active virus.

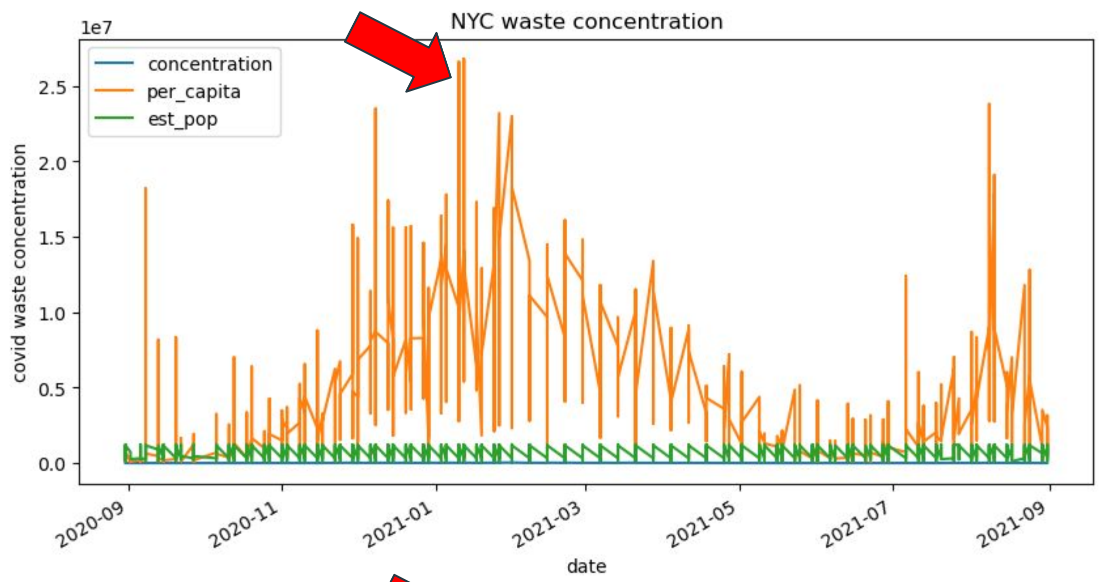


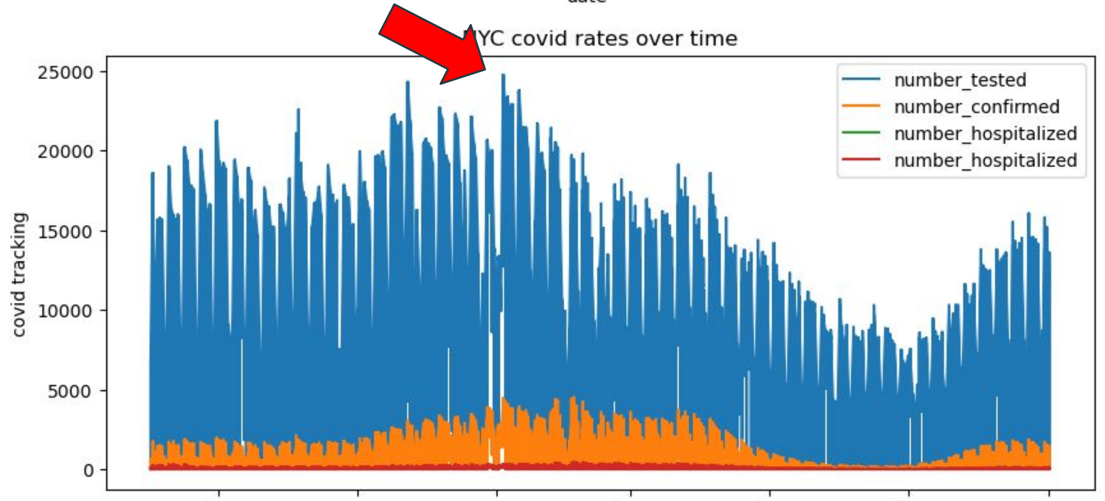
Data

NYC offers data on concentration of COVID-19 genetic material in waste water.

NYC also offers an estimation of how many people each plant serves

We will use a second source of NYC data on COVID-19 cases as our measure of if there is an outbreak

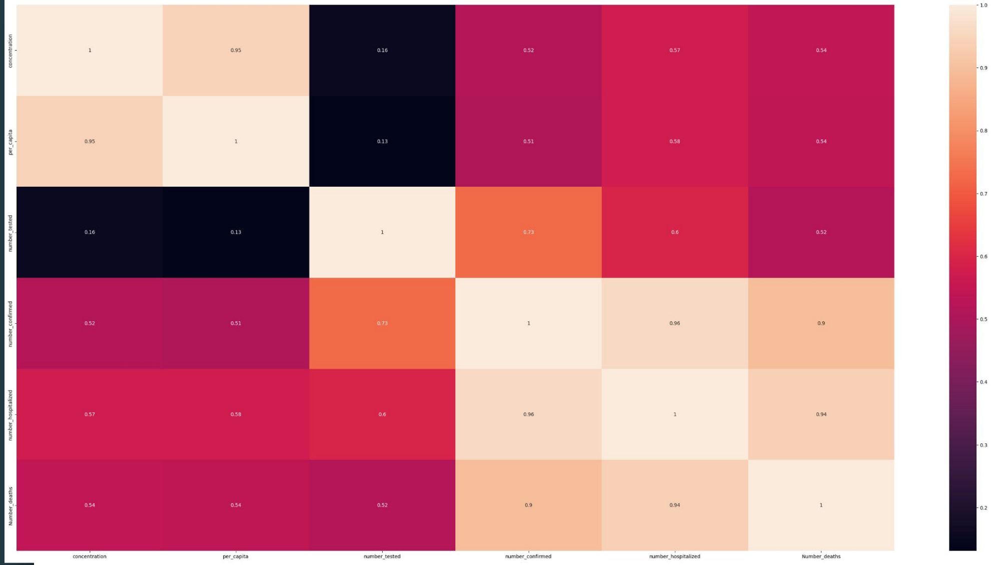




After visualizing the data

It looks like there is some lag time between the spikes in the COVID rates and the waste concentration

There may be a relationship between these features.



	concentration	per_capita	number_tested
number_confirmed	0.515754	0.508627	0.730787
number_hospitalized	0.573271	0.575328	0.598586
Number_deaths	0.543993	0.540407	0.516646

After cleaning and wrangling both data sets the data was combined into one dataframe

A correlation heat map was performed

There did seem to have a correlation between the concentration of COVID waste water and the number of covid cases

Modeling Issues

As the data was explored, no matter what model was used poor accuracy scores are returned

I tried doing a time series split of the data to further investigate the relationship between the features that had the best correlation.

I tried decision tree classifier, random forest classifier, AdaBoost classifier, KNeighbors Classifier

I tried to use gridsearchcv to figure out the best parameters for KNN or randomforest

In conclusion we can’t use waste water to predict outbreaks of COVID-19

Future research

We could try to find more complimentary data.

The CDC has a similar waste water tracking across the entire US.

With more data it may be possible to make a prediction of COVID outbreaks

Considerations

A blind spot in this data is that some rural communities may not have public sewer systems and those areas may not be properly represented.

Testing of sewer water may by politically problematic.

The public or voters may find monitoring wastewater for COVID may be the start of a trend to monitor for other kinds of illnesses or possible drug use and may cause resentment in government overreach.

Thank you
NYC open data
My mentor Rahul Sagrolikar

Questions?

sources

https://www.nyc.gov/site/dep/whats-new/covid-19-wastewater-testing.page

https://data.cityofnewyork.us/Health/SARS-CoV-2-concentrations-measured-in-NYC-Wastewat/f7dc-2q9f/about_data

https://data.cityofnewyork.us/Health/COVID-19-Outcomes-by-Testing-Cohorts-Cases-Hospita/cwmx-mvra/about_data

https://freakonomics.com/podcast/water-water-everywhere-but-you-have-to-stop-and-think/