Wastewater Predicting COVID-19 Positives Cases in Ottawa

Yuxiao Chen
School of Computer Science
Carleton University
Ottawa Ontario, Canada

Kyle Knobloch
School of Computer Science
Carleton University
Ottawa Ontario, Canada

This project aims to apply wastewater data to predict the positive COVID tests in Ottawa. The methodologies include a linear model and random forest regressor model to analyze and predict COVID-19 cases. Being able to predict COVID-19 positives cases based on the levels of COVID-19 RNA in wastewater testing will give the public health unit time to ramp up testing and ready hospital beds. This information will also allow for a more accurate depiction of COVID-19 case numbers when combined with traditional testing.

***Index Terms*—Artificial Intelligence, COVID-19**

I. Introduction

Many cities in Ontario have started using wastewater as a form of testing for positive COVID-19 cases. This includes Ottawa, Toronto, and Waterloo among others. This project is focused on Ottawa wastewater testing as an indication of the number of positive COVID-19 cases. The hope is to be able to get a better picture of the COVID-19 positives cases in Ottawa as testing rules and availability change throughout the pandemic. Wastewater is collected and tested for specific viral RNA strands and results are posted the following day. COVID-19 case spikes place health workers in a tough position to cater to the increasing needs of the community. This problem leaves little room for preparation and therefore causes a slower response and resource scarcity.

The project aims to focus on this issue by analyzing and predicting future COVID-19 positive cases based on wastewater levels, giving the community and public health unit more time to prepare to be able to better handle the COVID-19 situation in Ottawa. Many challenges did arise in the face of this problem. Wastewater gets collected five days a week, giving two days without data. The objective, however, is to use artificial intelligence models to be able to predict future data sets. The aim is to use the wastewater data results from the Ottawa region and Ottawa public health unit's data set of positive COVID-19 cases to implement the analysis.

II. Approach

A. Data Cleaning and Splitting

The wastewater data and the COVID-19 positive cases data needed to be downloaded into the Python script and cleaned up into usable arrays. Then, the two data needed to be combined on a common index, in this case the date the data was published was the obvious choice. To achieve this, a unique loop was created that would be able to loop over both the wastewater data as well as the COVID-19 positive case data and combine them based on the date. This resulted in a table that has the date, the combined normalized means of the wastewater testing and the total positives COVID-19 cases in Ottawa. With this data, it was into randomized test and training test sets. For both models, the date column was not needed so that column was dropped or only used as an index for the data.

B. Linear Regression Model & Gradient Descent

The first methodology utilized Linear Regression and Gradient Descent. The goal is to build a linear model which can fit COVID-19 tests with the wastewater data as well as attempt to predict the COVID-19 case count based on the wastewater results. The predictive ability of the model is based on the gradient descent model. PyTorch was imported into the environment to help run the model efficiently and calculate the loss.

Before stepping into the model establishment, the wastewater level and COVID-19 tests were graphed to investigate the data trend as seen in Figure 1. Next, the initial linear model was designed and implemented by randomly generating a weight and a bias. The bias was initially set to 0. However, the actual results are far from the predicted results as seen in Figure 2. Therefore, the loss needed to be calculated. With PyTorch, the average of the squares of the difference between the actual data and the predicted results was produced. Next, the derivative of the loss was calculated. Since the number of wastewater levels is very different from that of COVID-19 tests, the learning rate was set to 100. With the new learning rate, the weight and bias values are updated. After the first update, the effectiveness is graphed in Figure 3. The conclusion that was drawn is that there is a need for more updates to get the predictions closer to reality. Hence, the number of iterations was increased to 10000. In each iteration, the gradient of weight and bias was set to 0 to avoid cumulation. At last, as the loss gets smaller, the linear model was tested with the split test data.

Linear regression is relatively fast to build, and it works well with a small volume of data and simple relationships. Additionally, the model is also easy to understand, and the results can be very interpretable, which is beneficial to decision

No funding was received for this project, however it will be used as a project grade for COMP 3106 A - W22 at Carleton University, Ottawa Ontario, Canada.

analysis. However, despite the mentioned advantages, linear regression also has its limitations. For example, it is difficult to model nonlinear data. And when dealing with very complex data, linear regression could not express very well.

C. Random Forest Regressor Model

After analyzing the data with the linear regression models, there was a need to get a better understanding of the data, so the decision was made to also use the SKLearn library Random Forest Regressor (RFR) model on the data. The RFR model was trained on the first 350 data points which are roughly 2 years of data. The RFR Then the test data was made up of the 50 most recent data available, about the most recent 2-3 months. This resulted in some strong results that showed that the current testing is severely lacking compared to the expected results based on the wastewater testing.

RFR is a machine learning technique that uses trees to try and either classify or predict the targets. In these cases, the model is trying to predict the value based on the inputs. This is done by using a decision tree, in ML a node is a decision that attempts to get closer to the desired result. As such, this model needs the training to be able to become more accurate with its decisions. This model generates hundreds of these trees, picks the one that best fits all the data sets and uses it for the prediction phase where the test data is used.

III. Results & Discussion

The wastewater testing data was provided by Big Life Labs and is available on GitHub. The Ottawa Public Health unit provided the data for COVID-19 testing results. The wastewater data needed to simplify the data as the wastewater data included many data points that are useful but were not needed in the research. The two wastewater columns that were focused on were “covN1_nPMMoV_meanNr” and ”covN2_nPMMoV_meanNr”. These are the two major variants of RNA that needed to be tested for, then they are normalized on the Pepper mild mottle virus levels that stay extremely stable in Ottawa’s wastewater. For the subset that’s being used in this research, the two means were combined by adding them together. For the COVID-19 positive tests in Ottawa, the data was split into two groups: the general population and the cases in long-term care homes. The data consisted of how many tests were performed in a day and a positive percentage. The data needed to multiply the total number of tests by the positive percentage to get the number of positive cases. The covid positive numbers were also combined to get a general picture of COVID-19 cases in Ottawa. This resulted in two wastewater sets, one with a combined wastewater normalized mean and one that had the two means in separate columns.

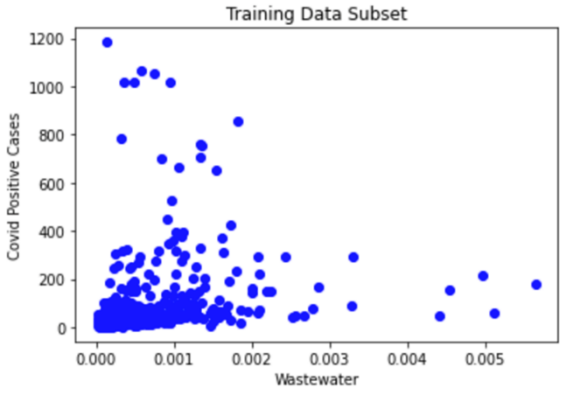

Fig. 1. Training Data Subset of Covid Positive Cases and Wastewater.

In the training dataset graph in Figure 1, a large part of the data cluster on the left side of the graph, which is the side with a lower wastewater level, while the 0 to 200 positive covid tests were the most intensive and restrictions were at their highest. As testing access improved and restrictions were decreased, there were higher positive case counts which correlated with higher wastewater testing mean values. In the subsequent model establishment and testing, it will be important to understand the initial data.

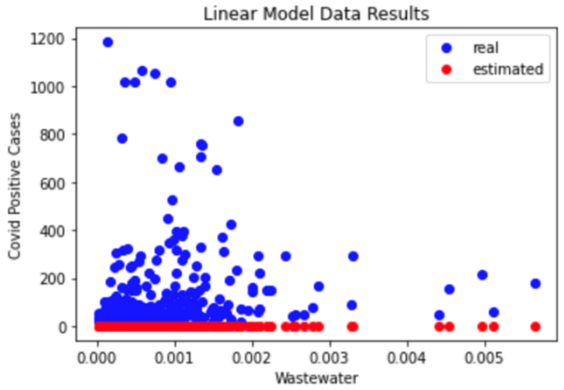

Fig. 2. Linear Model Data Results.

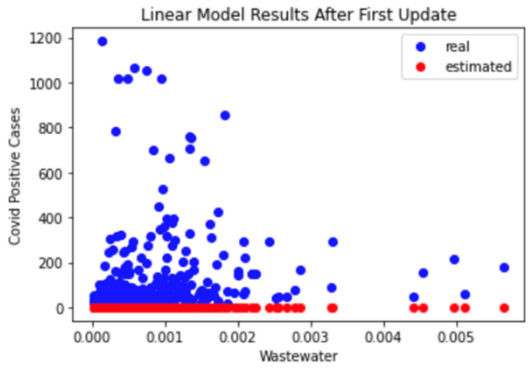

Fig. 3. Linear Model Results After First Update.

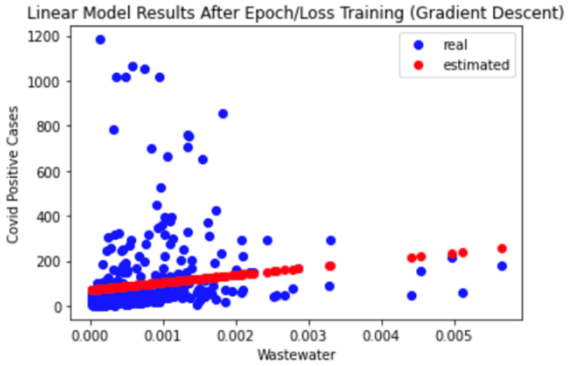

Fig. 4. Linear Model Results After Epoch/Loss Training (Gradient Descent).

In Figure 2, the initial linear model and actual data are displayed. From the graph, the estimated data points are very different from the real ones resulting in a very high error rate. Therefore, the linear model was improved with a gradient descent. Figure 3 shows the real and estimated data after the first update. It is clear to see that the difference between actual and estimated is still high resulting in a poor fit. Finally, in Figure 4, after 10000 iterations of updates, it can be observed that the linear model predicts 0 to 200 COVID tests increasing on the clustering side, which is a reasonable fit.

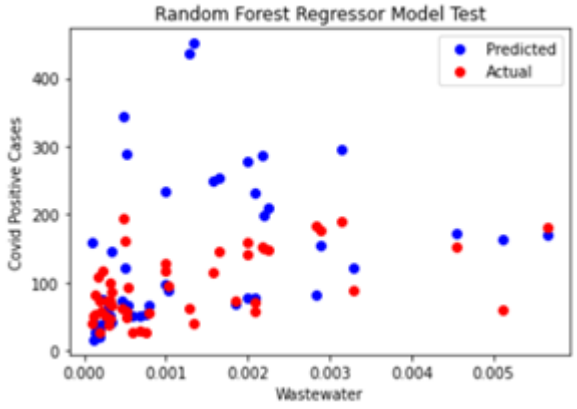

Fig. 5. Random Forest Regressor Model Test.

The Random Forest Regressor model was tested on the last 50 data points of the wastewater collection which is a little over two months in Figure 5. What can be seen is that there is a disconnect between the predicted and actual COVID-19 positive cases. The blue predicted shows that there is a high number of COVID-19 cases when the actual shows there is a small number of positive cases. This indicates that there are likely more covid cases than what is currently being reported and tested for in Ottawa. This result is likely due to the changes in the PCR testing eligibility in Ontario which was dramatically reduced to those who are high risk which includes the elderly, medical workers in high-risk settings and other minority groups.

This data and analysis show that in Ottawa the number of reported cases is likely much lower than the actual number of cases. The other conclusion that could be drawn from this result is that a new wave is about to begin in Ottawa. This can be seen in graphs that overlay the new positive cases counts and the wastewater testing levels. Usually, the wastewater beings to trend upwards before the positive COVID-19 cases do as can be seen in Appendix A, Figure 7.

IV. Conclusion

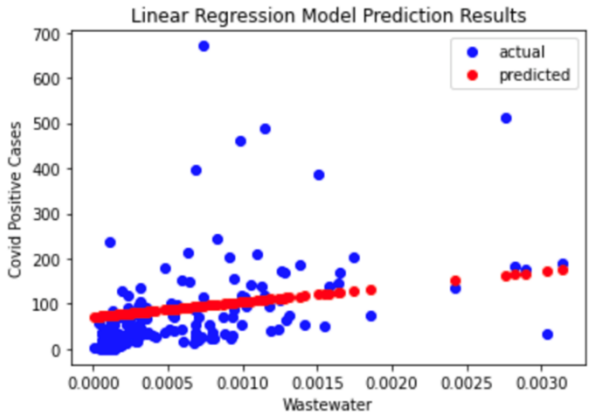

Fig. 6. Linear Regression Model Prediction Results.

To conclude, the linear regression model was able to estimate data points close to the actual ones in the training section as shown in Figure 4 as well as in the test dataset Figure 6. Although the predicted results do not fit the actual data completely, the results are still reasonable. Overall, the number of positive tests for COVID-19 is increasing as the wastewater level increases. As for the unpredictably high or low number of positive COVID-19 tests, many other factors may have contributed to this. These include increased or decreased transmissibility of COVID-19 as variants were identified and spread in communities. There were changes in the PCR testing accuracy, availability, reporting and eligibility which could have dramatically reduced the reported number of cases compared to the actual number of cases. It also disproportionately reported those who are high risk which including the elderly, medical workers in high-risk settings and other minority groups as COVID-19 positive and under-reported those who are at low risk of getting COVID-19.

The results from this analysis show that the linear model is an OK fit for the relationship between wastewater and positive COVID-19 tests results. The linear model fit is not perfect but does work well to display the correlation between the two data points. The random forest regression model displayed that we are likely in another peak of this pandemic that is being under-reported by public health unit testing in Ottawa. The RFR model showed that there is a high level of COVID-19 RNA in the wastewater testing which predicts a high level of COVID-19 positive cases. However, as Figure 5 shows, the difference between actual and predicted is very high.

User Manual

The programming environment that was used is Google Collab as it supports many useful AI-related Python libraries.

To view the code, visit the GitHub repository linked below. Inside of the Jupyer Notebook, each code block must be executed in order as the previous block contains imports and variables for the next block. The code is commented on and organized into three sections for readability.

https://github.com/comp3106-w22-covid/Code/blob/main/COMP_3106_Project_Code.ipynb

APPENDIX A

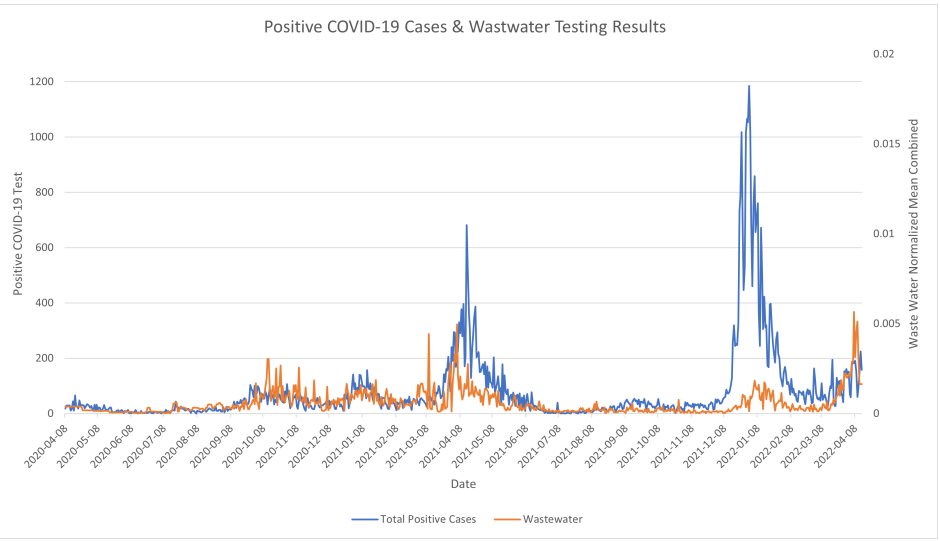

Fig. 7. Linear Regression Model Prediction Results.

REFERENCES

[1] X. Liao, "Learn Deep Learning with PyTorch", 2019, https://github.com/L1aoXingyu/code-of-learn-deep-learning-with-pytorch.

[2] Big Life Lab "The Public Health Environmental Surveillance Database (PHESD)", 2022, https://github.com/Big-Life-Lab/PHESD.

[3] Ottawa Public Health Unit " Daily number of Ottawa residents tested for COVID-19 and the percentage of residents tested with laboratory-confirmed COVID-19", https://www.arcgis.com/home/item.html?id=26c902bf1da44d3d90b099392b544b81.