

Final Project - Tracking COVD-19 from Wastewater

Lalitanjali Bonlidi, Vivek Bhatia, Raghu Soori, and Prakash Bhatt,

Harvard Extension School - 109B - 2021 Spring

Prof Protopapas, Glickman, and Tanner

Contents

[Background: 3](#_Toc70452649)

[Motivation: 3](#_Toc70452650)

[A description of the data: 3](#_Toc70452651)

[Method to explore the data 3](#_Toc70452652)

[What methods have you used to explore the data 4](#_Toc70452653)

[Visualizations and captions that summarize the noteworthy findings of the EDA 4](#_Toc70452654)

[A project question based on the insights you gained through EDA 4](#_Toc70452655)

[A baseline model or clear plan for its implementation 4](#_Toc70452656)

Background: Conventionally COVID-19 is tracked via symptomatically and clinically diagnosed cases. This method has inherent drawback as many of the patients do not report the disease. In the paper titled “SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases” authors gathered data of presence of SARS-CoV-2, the virus responsible for COVID-19, in wastewater. The paper was published in <https://www.medrxiv.org/content/10.1101/2020.06.15.20117747v2> . The authors were successful in uncovering a trend that showed the presence of the virus in wastewater is a leading indicator as in 4-10 days the clinical data follows the wastewater data. The work suggests that longitudinal wastewater analysis can be used to predict COVID-19 cases more accurately than clinical data.

Motivation: As part of the final project for Spring 2021, class of Harvard Extension School for the 109B- Data Science 2, Advance Topic in Data Science, we are attempting to replicate, improve or validate the finding of the author in the original paper.

A description of the data: We are using the data set in a excel file MWRAData20210424-data.xlsx. This data is collected by company called BioBots who is specialized in detecting presences of the virus in the wastewater. The measurement is collected daily, and the data is collected from two of the counties around the Boston. One of them northern and other is southern counties. The data is collected daily and measurement of presence of the data is in copies/ml. For each of these southern and northern counties, the data also provides 7 days average, low confidance interval, and high confidence interval. Dataset also has variance of south/north, but that data is not populated in the provided data set. The example of the data is below.



Picture: data model

Method to explore the data

We used conventional EDA technique to explore the data. Since the data is collected almost daily, we plotted the time series data with 7 days lag in both of the counites.

To be filled by Vivek……

What methods have you used to explore the data

Visualizations and captions that summarize the noteworthy findings of the EDA

A project question based on the insights you gained through EDA

A baseline model or clear plan for its implementation

We propose to predict the future infection date based on the past history. We would like to explore the various model for those prediction

* Regression
* RNN and LSTM
* XGBoost
* ARIMA?
* GAM

Appendix

Meenu’s link

<https://www.kaggle.com/anubhavsanyal/tensorflow-with-lstm-83-accuracy>