Monitoring COVID-19 Infections through Wastewater Surveillance

CMPSC 190DD/197A Project, Fall 2020

Motivation & Background

Surveilling SARS-CoV-2 in wastewater appears promising within the public health arsenal for discovering community COVID-19. Monitoring aggregate community wastewater to quantify the circulating virus can motivate investigating upstream sub-sewersheds, intensifying individual human population testing, invoking shelter-in-place orders, and preparing for anticipated hospitalization surges. Globally, studies show compelling correlations between higher wastewater SARS-CoV-2 RNA copy numbers and COVID-19-positive individuals.

It has been observed that *socioeconomic vulnerability* affects the susceptibility to COVID-19 exposure [1, 2]. For example, economically disadvantaged people may be sharing living spaces with many others or people who work jobs (typically low-paying) that require in person interactions are more likely to be exposed to the virus. *Comorbid* (pre-existing) conditions also influence the incidence and severity of COVID-19 disease. Many such diseases or behaviors could be evidenced by chemical wastewater surveillance to differentiate disease susceptibility of individuals within common airsheds. Metabolized and unmetabolized therapeutics, along with chemical markers of comorbid disease and substance use, are excreted into wastewater in addition to SARS-CoV-2. Understanding patterns of comorbidity chemical markers that co-occur with SARS-CoV-2 could indicate in sewersheds where, on more rapidly actionable bases than socioeconomic or demographic factors alone, COVID-19 interventions should be prioritized. This project focuses on this idea, called *Wastewater-based Epidemiology* (WBE). See Figure 1 for a causality graph that explains the assumed relationships between social vulnerability, comoborbidity, disease incidence and viral load.

A number of models have been proposed for modeling the dynamics of COVID-19. The group of "mechanistic" models considers the flow patterns across different compartments to describe the dynamics of the disease e.g., the SIR (S: susceptible, I: Infected, R: Recovered) and the SEIR (adding an additional E: Exposed state) models. Among models in this group are the Imperial College London model and its refinements that stratify the population. The other broad class of models is data-driven. The IHME model, a widely-used model in this group, models the per-capita death rates of a geographical region and utilizes parameterized time series matching with time-series of recovered regions to predict the future dynamics.

In this project you will be modeling the dynamics of COVID-19 using a modified SIR model. Motivated by the previous discussion on the effects of socioeconomic vulnerability and comorbidity, the model divides the population into different compartments with varying levels of social vulnerability and comorbidity. It then constructs the governing equations based on this segmentation. There are multiple parts of the project with each related to a different aspect or task about the model.

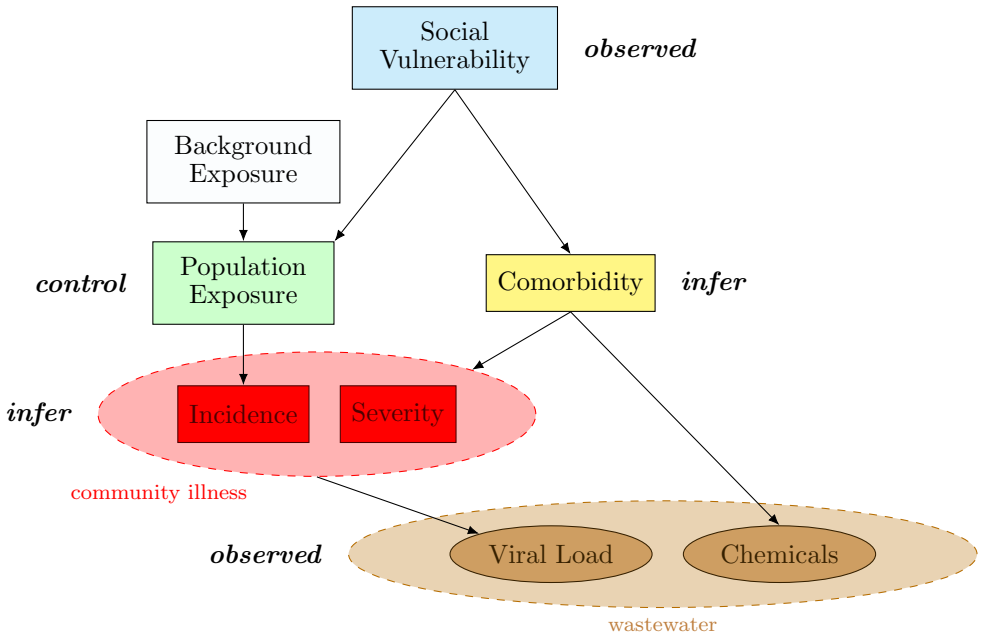

Figure 1: Social vulnerability (observed, from census tract data) affects COVID-19 exposure and comorbidity (observed from wastewater data through chemicals). COVID-19 incidence across locations is risk-based, affected by exposure to the virus and comorbid conditions, a hypothesis aligned with the literature showing that populations with specific pre-existing conditions (e.g. cardiovascular disease, diabetes, asthma, old age) or substance usage (e.g. smoking tobacco or marijuana, or vaping) are more likely to be COVID-19 symptomatic and excrete a higher viral load associated with severe COVID-19. The far right arrow indicates that comorbidity affects the wastewater chemistry (metabolites and pharmaceutical risk factor indicators) assessed by proxy of medication affecting the whole chemical composition. The incidence and severity are expected to be correlated with viral discharge, hence wastewater SARS-CoV-2 concentrations. There are thus two observations likely to be correlated: spatially mapped social vulnerability and wastewater composition.

Logistics

– Please don't hesitate to ask questions if you're having a difficult time and start early.
– Create a public repository on Github for the project files. For each part create a Jupyter notebook file, and your write explanations and code in the notebook then push them to the public repo you created. Please be clear in your explanations and answers. You can use any visualization tool that you want to make your point.
– If you want to collaborate easier with your group mates, you can use Google Colab or JupyterHub.
– Before you do anything, explore the given data files. What data do they contain? What are the shapes and data types of arrays? What do they represent and how do they come together? Make sure you understand before delving into the details.
– At the end of the quarter, you will compile your findings into a presentation and present it to your peers.

Project Details

You are given a dataset containing measurements of the viral load density L (observed from wastewater) from

- A tree representing a city's sewer system:
  - 100 leaf regions
  - Binary tree T
  - The tree structure is used in Parts 3 and 4(optional) only.
- For each leaf node, the population (same across leaf nodes) is divided into segments according to their social vulnerability and comorbidity:
  - measure of social vulnerability is discretized to 4 values.
  - measure of comorbidity is discretized to 4 values.

For a given leaf node, let the entire population be N. The population is divided into S, I and R components following the SIR model. Furthermore, the susceptible population S is divided into compartments $S_{v,c}$ based on social vulnerability and comorbidity (So $\sum_{v,c} S_{v,c} = S$). Similarly, the infected population I is partitioned into separate compartments I_c based on comorbidity (So $\sum_c I_c = I$). We parametrize a transition rate from the susceptible compartments to the infected compartments as follows:

$$\frac{dS_{v,c}}{dt} = -\frac{\beta_{v,c} \cdot S_{v,c} \cdot I}{N}$$

$$\frac{dI_c}{dt} = -\sum_v \frac{dS_{v,c}}{dt} - \gamma \cdot I_c$$

$$\frac{dR}{dt} = \gamma \cdot I$$

Here $\beta_{v,c}$ is the reciprocal of the typical time between contacts and γ is the reciprocal of the typical time until an infected person recovers (or succumbs).

The total viral load L is assumed to be:

$$L = \sum_c I_c \cdot L_c \tag{1}$$

where L_c is the viral load for a person in the comorbidity compartment c.

The known model parameters are the initial susceptible populations in each compartment (i.e. $S_{v,c}^{(0)}$) in each leaf node, the initial infected populations $I_c^{(0)}$ in each leaf node, L_c values and γ. You can assume that initial recovered populations are zero (i.e. $R^{(0)} = 0$) ~~and that initially $S_{v,c}^{(0)} >> I_c^{(0)}$ for all v and c, therefore $N \approx \sum_{v,c} S_{v,c}^{(0)}$~~ This is wrong. Please disregard it and compute $S_{v,c}^{(0)}$ by subtracting the initial infected population $I_c^{(0)}$ from N. What is not known are the model parameters $\beta_{v,c}$. It is important to emphasize that the modeled behavior of the disease is the same across different leaf nodes. In other words, **the model parameters $\beta_{v,c}$, L_c and γ are shared across the leaf nodes**.

Project Parts

Part 1 (due November 9^{th} 11:59 PM)

(a): Initial $S_{v,c}^{(0)}$ probability mass distribution (PMF) as well as initial infected population sizes $I_c^{(0)}$, and model parameters L_c and γ for a single leaf node are given in `part1a.npz`. Choose arbitrary $\beta_{v,c}$ values between in $[0, 1]$ such that $\beta_{v,c}$ are non decreasing as v or c increases. In other words, $\beta_{v,0} \leq \beta_{v,1} \leq \beta_{v,2} \leq \beta_{v,3}$ for all v and $\beta_{0,c} \leq \beta_{1,c} \leq \beta_{2,c} \leq \beta_{3,c}$ for all c.

1. Using the model dynamics described in the previous section, and all the given and chosen parameters, simulate the behavior of the disease for 120 days for this single leaf node.

2. Plot $S_{v,c}$ and I_c values over time. (You can plot all $S_{v,c}$ on the same plot. Same for I_c.). Also plot the overall S, I, R and L values.

3. Do the shapes of S, I, R look similar to what you expected? Which of these plots is the "curve" people refer to when they say "flatten the curve"?

4. Do you observe that some $S_{v,c}$ compartments converged to zero while others converged to a positive value? Why do you think that is?

5. Print the percentages of population that never got infected for all compartments (i.e. all values of v and c).

6. Multiply all the $\beta_{v,c}$ values by $1/4$. What happened to the S, I, R plots? Did the "curve" flatten compared to the previous case? Print the percentages of population that never got infected with these $\beta_{v,c}$ values.

(b): Now suppose we don't know the model parameters $\beta_{v,c}$ and we are trying to estimate them from observed data—as would happen in the real world. To make the estimation of parameters easier, from all the leaf nodes the ones with **uniform single social vulnerability** are picked. So you are given 5 leaf nodes each for the four different values of social vulnerability in the `part1b.npy` file. The PMF of susceptible population $S_{v,c}^{(0)}$ in each comorbidity compartment areas follows:

5 Leaf nodes with vulnerability = 0.2: PMF of comoborbidity: (0.5, 0.3, 0.1, 0.1) for all nodes.
5 Leaf nodes with vulnerability = 0.4: PMF of comoborbidity: (0.4, 0.3, 0.2, 0.1) for all nodes.
5 Leaf nodes with vulnerability = 0.6: PMF of comoborbidity: (0.3, 0.3, 0.2, 0.2) for all nodes.
5 Leaf nodes with vulnerability = 0.8: PMF of comoborbidity: (0.1, 0.2, 0.3, 0.4) for all nodes.

1. Using these distributions, total population sizes and observations of viral load densities L for 20 consecutive days (day 0 through day 19) for the 20 leaf nodes that are given to you, estimate the 16 parameters $\beta_{v,c}$ using grid search and MMSE on the "validation" data. In other words try to minimize MSE between $L_{\text{validation}}$ and $L_{\text{estimated}}$. Compare the $\beta_{v,c}$ you estimated with the ground truth given in the file. This step is to make sure your code and logic works correctly.

Tips:
– For the grid search, don't search over 16 values of $\beta_{v,c}$. Remember that nodes in different social vulnerability do not share any $\beta_{v,c}$ values. So $\beta_{i,c}$ and $\beta_{j,c}$ are independent for $i \neq j$.
– Divide [0,1] interval to 21 values for each $\beta_{v,c}$. Since the $\beta_{v,c}$ values are non-decreasing as v or c increases, *a lot* of the points on the grid are *invalid*. Use this to your advantage while doing the grid search.
– Don't worry if your MSE is not zero. There is noise in the measurements L.

2. Now use the "test" data (20 L values for 20 days) from the file and estimate the $\beta_{v,c}$. This time you won't have access to the ground truth $\beta_{v,c}$ values. Print the $\beta_{v,c}$ values you estimated.

3. Use the $\beta_{v,c}$ you found to predict the disease behaviour for the future. Solve the equations for 100 days and plot S, I, R, L values for the first node in each social vulnerability case. On L graphs, also plot the corresponding observed values for the first 20 days (with a circle marker).

Part 2 (due November 24^{th} 11:59 PM)

(a): Suppose that upon hearing the news of the epidemic on day 10, people start to take precautions on their own. $\beta_{v,c}$ values were the same for all nodes before day 9, and *immediately* on day 10, each node's $\beta_{v,c}$ values get multiplied by a constant $\alpha_{own} \in [0, 1]$. This α_{own} is different for different nodes but the same for different v and c compartments within the same node. Figure 2 demonstrates the effect of α_{own} on S, I and R plots.

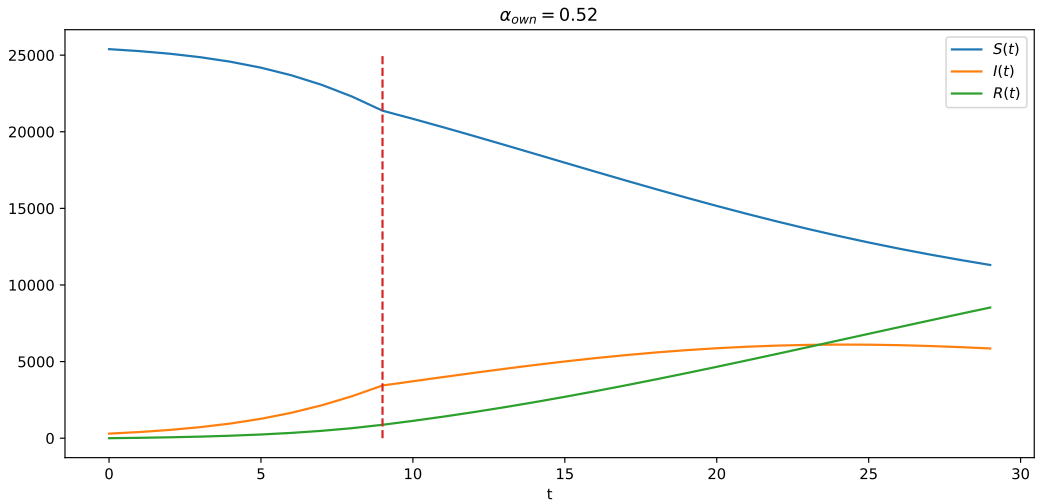

Figure 2: The effect of α_{own} starts on day 10.

Determine the α_{own} values for each node and rank nodes by their α_{own} values. You are going to use a slightly more sophisticated search method which uses an off-the-shelf function to do the mean squared error minimization. You will write a function that takes in a guess of α and returns the mean squared error between $L_{predicted}(t)$ and $L_{observed}(t)$ **between days 10 and 29**. You will find $L_{predicted}(t)$ the same way you did in Part 1 except the **initial conditions will be the $S_{v,c}$, I_c and R values on day 9**. Once this function is written, you can use `scipy.optimize.fmin` to find the α value that gives the minimum mean squared error. You can choose the starting point (second argument to `fmin`) as 0.5.

(b): According to CDC, *non-pharmaceutical interventions* (NPI) are actions, apart from getting vaccinated and taking medicine, that people and communities can take to help slow the spread of

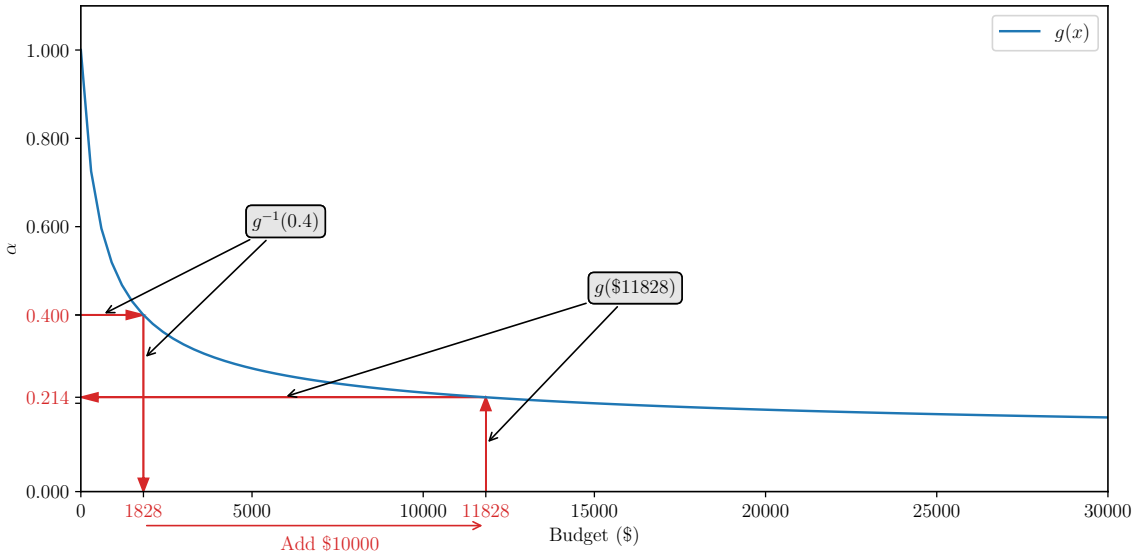

Figure 3: Start from $\alpha_{own} = 0.4$, go to \$1,828 by $g^{-1}(0.4)$, add \$10,000 NPI spending and find new α after NPI by $g(\$11,828) = 0.214$.

illnesses. After seeing the rapid spread of the epidemic, the local administration decides to use an NPI (ad campaign) to slow the spread which takes effect *immediately* on day 30. For this intervention they allocate a budget of \$1,000,000 in total, for all 100 nodes. Your job is to allocate the budget across the nodes *fairly* using some criteria you determine.

We assume the relationship between the α value (accounting for both own intervention and NPI) is given by this function:

$$\alpha = g(x) = \frac{1}{log_2(2 \cdot 10^{-4}x + 2)}$$

where x is the budget in dollars. For reference, the inverse of this function is given by:

$$x = g^{-1}(\alpha) = \frac{2^{1/\alpha} - 2}{2 \cdot 10^{-4}}$$

For example, if a node had $\alpha_{own} = 0.4$ and we spend \$10,000 for NPI for this node, it will have an α value of $g(g^{-1}(0.4) + \$10,000) = g(\$1,828 + \$10,000) = 0.214$. Figure 3 illustrates this example.

Using this relationship between money spent on NPI and α, answer these questions for all policies:

i. How do the α values change when you do this? Make a scatter plot of all 100 nodes where x-axis is α_{own}, y-axis is $\alpha_{after\ NPI}$, marker size is node's population size and marker color is determined by average β value of that node. This plot might make it easy to see some patterns.

ii. What is the change in the number of recovered/succumbed people in total after 200 days compared to the case if there was no NPI at all?

Here are policies you should consider:

1. Spend the same amount of money for each node.

2. Spend the budget proportional to each node's population.

3. Spend the budget such that $\alpha_{own} - \alpha_{after\ NPI}$ is the same for all nodes, that is the change in α induced by the NPI is the same for all nodes. To do this, you can again use `scipy.optimize.fmin`. Write a function that takes in a value for change in $\alpha := \Delta\alpha$ and returns the difference between the budget that corresponds to this $\Delta\alpha$ and \$1,000,000. By using `fmin` on this function we find the value of $\Delta\alpha$ that corresponds to the budget of \$1,000,000.

4. Find another way to spend your budget. This is an open ended question so there are no right or wrong answers. You need to decide on a way to measure the effect of your NPI and optimize it while still being fair in terms of a criteria you determine. Some suggested policies might be:

   - Spend the budget based on each node's α_{own} and try to bring all nodes α's to a similar level.
   - Spend the budget proportional to each node's average beta value weighted by the susceptible population size in comorbidity and vulnerability compartments.
   - Hard: Minimize the total recovered population across nodes after 200 days.

(c): For each of the 4 policies above, explain why you think doing this might be fair or unfair. Justify your case and visualize as you see necessary. One way to measure fairness is the individual fairness idea we talked about in class (individuals being the nodes in this case). The question is what is the metric by which you measure two individuals' similarity. You can measure it such that if two nodes' α_{own} values are similar, the budget spent or change in α should be similar. Or you can say the metric of similarity is the population's social vulnerability distribution, therefore nodes with similar SV distribution should have similar levels of NPIs action. Alternatively, you can cluster the nodes based on their SV and/or comorbidity distribution (for example based on the centroid of their PMF–use `scipy.ndimage.center_of_mass`) and see if they receive similar NPI actions.

References

[1] JA Patel, FBH Nielsen, AA Badiani, S Assi, VA Unadkat, B Patel, R Ravindrane, and H Wardle. Poverty, inequality and covid-19: the forgotten vulnerable. Public Health, 183:110, 2020.

[2] Emily E Wiemers, Scott Abrahams, Marwa AlFakhri, V Joseph Hotz, Robert F Schoeni, and Judith A Seltzer. Disparities in vulnerability to severe complications from covid-19 in the united states. Technical report, National Bureau of Economic Research, 2020.