Exploring surrogate approaches that utilize networks for prediction in epidemic simulation

Student: Lillian Encarnation
Mentors: Dr. Gursharn Kaur, Dr. Sifat Moon

Background

Research Question

CDC surveillance data is often provided at the state or county level, but higher resolution spatial data, such as at a zip code level, can better inform local public health decisions and allocation of resources.

Provided with an epidemic simulation in a contact network, how can we predict infected count trajectories for a zip code given the trajectories for all other zip codes within that county?

- Can we predict $I_{z,t}$ given $\sum I_{z,t}$?
- How can we assess the quality or accuracy of the prediction?

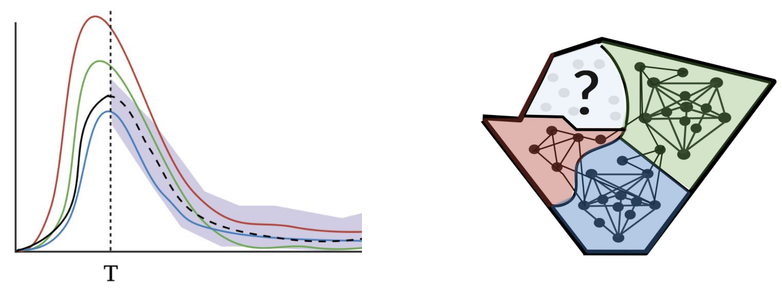


Figure 1. Predictive method visual. Question of whether unknown zip code infected count trajectories can be forecasted with a confidence interval based on county level data.

SIR dynamics on networks

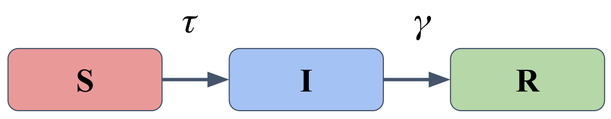


Figure 2. SIR dynamics flow chart. Susceptible (S) individuals become infected (I) at rate τ (transmission rate) and recover (R) at rate γ.

- Agent-based simulation
- Use of EoN model fast SIR simulation in our constructed network with the assumptions of a closed population

Properties of synthetic population in network

- Data is digital twin of a real world network
- There are 19 zip codes, ranging in population from 23235 to 70 nodes

Zip Code	# of Nodes	Mean Degree	Degree IQR
22901	23235	25.48	29.00
22911	15039	24.67	28.00
22903	12676	25.10	28.00
⋮			
22595	928	22.92	26.00
22931	919	24.10	25.00
22904	70	23.61	29.25

Table 1. Zip code attributes in network. Total number of nodes in county is 93569. Total number of zip codes is 19.

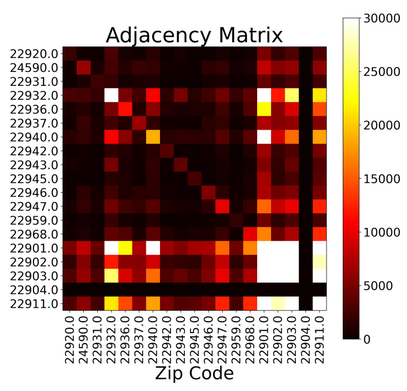


Figure 3. Heatmap of Adjacency Matrix for Synthetic Network. Range is restricted to show detail.

Future Work

- Analyze Var($R_{z,t}$) and MSE(z) as a function of time (t) and τ.
- Investigate if the ratio of **new cases** also stabilizes over time.
- Utilize simulation outputs to explore other forecasting models like **COVID-LSTM** and **MTS-LSTM**.
- Incorporate the distribution of node attributes (age, gender, occupation, etc.) in the prediction model.

Current Work

Analysis of different epicurves

- Simulation is run for three transmission (τ) values (0.06, 0.04, 0.03) with γ=1, with 30 replicates and each replicate has ten random seeds at T=0

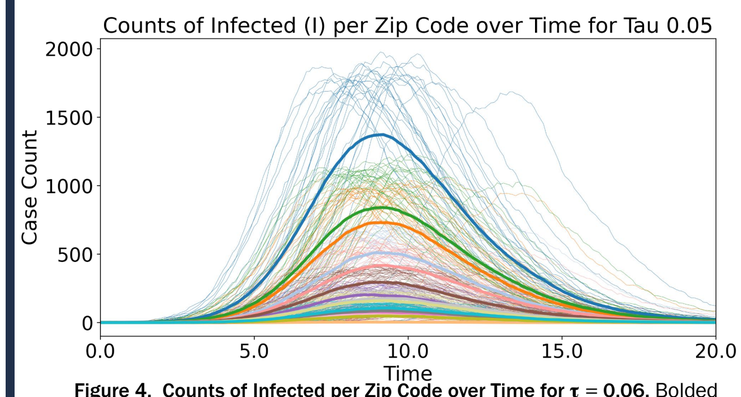


Figure 4. Counts of Infected per Zip Code over Time for τ = 0.06. Bolded lines show mean of replicate infected counts over time and lighter lines of the same color show the 30 replicates for that zip code.

- Observation that infected count curves appear as scaled versions of each other (Figure 4)
- Led to investigation of ratio of infected counts for each zip code to the overall infected counts for all zip codes over time.

Ratio curves for simple forecasting

- Ratios of zip code to county infected counts over time were calculated by the following equation, scaling by population:

$$R_{z,t} = \frac{N}{P_z} \times \frac{I_{z,t}}{\sum_z I_{z,t}}$$

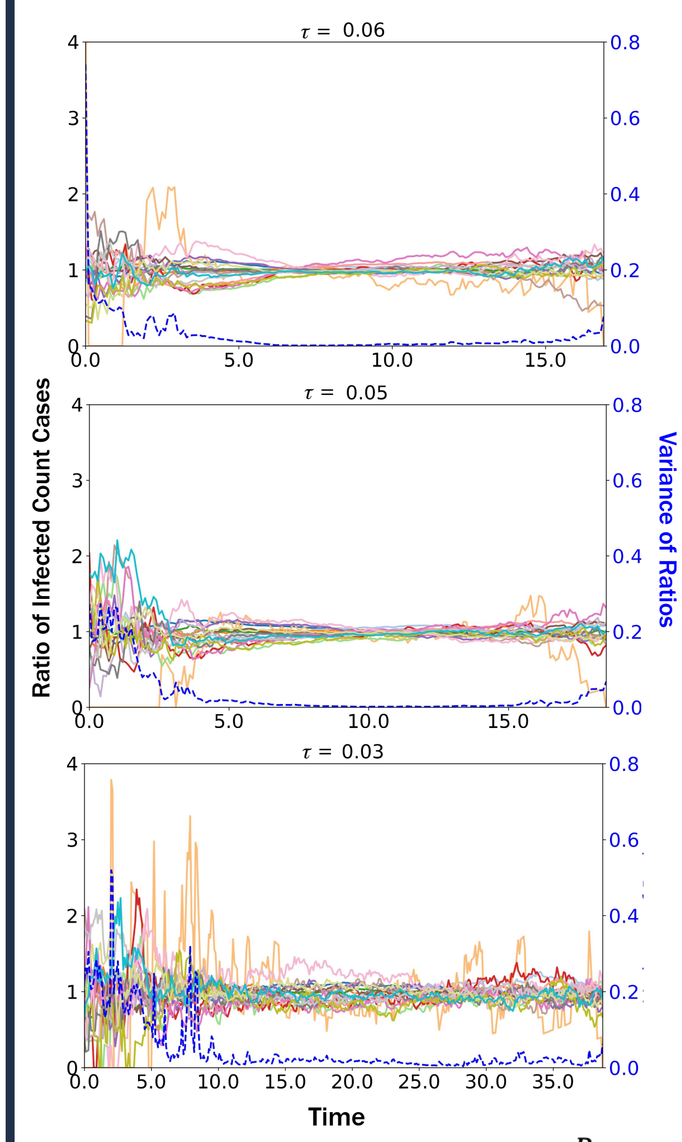


Figure 5. Ratios of Population-Scaled Infected Zip Code Epicurves to Overall County Epicurve for Differing τ Values. Solid colored lines represent differing zipcode ratios. Dashed line shows plotted variance between zip code ratios per time step. The maximum x-axis value corresponds to the first time a zip code reaches zero mean infected cases. Note time axis differences.

- Figure 5 indicates that $R_{z,t} \approx R_z$ for t > T, where T depends on τ. A smaller τ and t corresponds to a higher T and higher Var($R_{z,t}$).
- Suppose the infection counts are known only for some zip codes, and the total infection count is known. We estimate the infection counts for a zip code z as follows:

$$\hat{I}_{z,t} = E[R_{zt}] \times \frac{P_z}{N} \times \sum_z I_{z,t}$$

Here, $E[R_{zt}]$ represents the average ratios with the known zip codes.

Ratio-based prediction assessment

- MSE (mean squared error) appears larger for zip codes with larger populations (r=0.78 for MSE and population size, for τ=0.06)

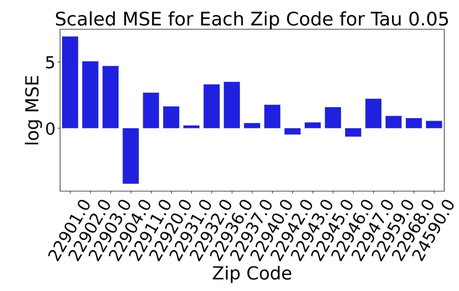
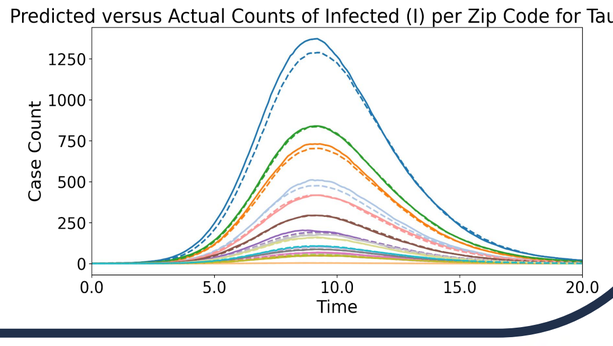


Figure 6. Scaled MSE for Zip Codes.

Figure 7. Predicted vs Actual Counts of Infected per Zip Code for τ = 0.05.

- Method of prediction shows potential merit in accurately predicting zip code infected count trajectories after time T, when τ and county information is known



References

Wang, L., Chen, J., & Marathe, M. (2019). DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9607-9612. https://doi.org/10.1609/aaai.v33i01.33019607

Lucas, B., Vahedi, B. & Karimzadeh, M. A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA. *Int J Data Sci Anal* 15, 247–266 (2023). https://doi.org/10.1007/s41060-021-00295-9

Nikparvar, B., Rahman, M.M., Hatami, F. *et al.* Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network. *Sci Rep* 11, 21715 (2021). https://doi.org/10.1038/s41598-021-01119-3