Improving Wastewater Surveillance: Using Computer Modeling to Improve COVID Transmission Tracking

(Technical Topic)

Investigating the Socio-Technical in Model Creation

(STS Topic)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Colin Crowe

May 1, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kathryn A. Neeley, Department of Engineering and Society

Nathan Brunelle, Department of Computer Science

1. Introduction

Following its initial global outbreak in 2020, COVID-19 has required us to create new, effective methods to track disease spread. One such method that emerged during the pandemic relies on detecting traces of COVID-19 in wastewater, in which samples are collected from sewer lines and we try to deduce which houses are infected based on what traces are found. This method has proven useful for managing COVID-19, though it is not without its fair share of uncertainty. In addition to unavoidable issues, such as sample transportation time, storage, sewage system leaks, and so forth, there are also several logistical issues unique to wastewater monitoring, such as sampling location, frequency of sample collection, and homogenization of the sewage flow all having an impact on our ability to "reverse-engineer" where each sample came from (Wade, 2022, p. 3). For my technical topic, I will investigate ways we can improve our accuracy when doing wastewater management by constructing a computer simulation that will model wastewater transmission and determine how, when, and where samples should be collected from sewage lines to get the most accurate information. Creating such a model will inevitably require making assumptions and simplifications, however, and so for my STS topic I will explore the social and epistemological concerns that go into creating such a model. I will investigate this using a relational view of data, in which I will assume that the data I use to inform my model as well as the data I collect from it have no inherit meaning and can instead only inform through comparisons to other data sets. I will use this framework to investigate how other models navigate the assumptions and simplification that must be made, and how this investigation both informs how I create my own model and the field of computer simulation as a whole.

2. Technical Topic

As discussed above, my technical report will focus on answering the question of how to best conduct wastewater surveillance by using a computer simulation to identify locations and times that conducting a sample would capture the most useful data. To do this, a two-layer simulation will be constructed. On the "top" layer, an agent-based model tracking the spread of COVID through a neighborhood or section of a city will be constructed, with enough granularity to detect individual infections. On the "bottom" layer, a corresponding model of the sewer system will be constructed, using fluid dynamics and the infected/susceptible status of individuals on the top layer to inform how wastewater containing traces of COVID moves throughout the system. This allows us to "cheat", using the top layer to check how accurate the data collected by samples in the second layer really is, and identifying the locations/times that yield the highest accuracy.

The top layer will use an agent-based susceptible-exposed-infected-recovered model to simulate the spread of disease, meaning that the model will simulate individuals as "agents" who can be in one of three states. Individuals who are "infected" or "exposed" have the ability to transmit the disease to individuals who are "susceptible", changing their status in the process. Individuals who are "exposed" do not know they have the disease, and so will act similarly to individuals who are "susceptible", while individuals who are "infected" know they have the disease and will quarantine themselves until they are better. Individuals who are "recovered" are not currently sick and cannot receive the disease. However, if enough time passes, "recovered" individuals can return to the "susceptible" state and be infected again. Constructing such a model requires extensive research and overcoming technical challenges, most notably in compiling data from different sources to base the simulation on and in simplifying the situation enough to allow a computer to give an output in a reasonable timeframe (Bissett, 2016, p. 629). This aspect of model creation will be discussed in-depth during the STS part of the report.

The bottom layer will use a directed flow graph to simulate the sewage network that corresponds to the top layer, with nodes to represent manholes and edges to represent the sewers that connect them. The main challenge here will be in determining the physical properties of COVID-19 and wastewater – how long traces of COVID-19 last, how diluted do they become as sewage lines connect up, and so forth. Consulting how data scientists solved other problems can be useful here. One study that is eerily similar to this discussed how unlawful discharge of harmful chemicals can be traced through sewage lines, in which the authors discuss how they overcame the issues of chemicals getting diluted the longer they stay in the sewage lines and the difficulty in tracing back the detection of chemicals to the unlawful discharge (Solano, 2022).

3. STS Topic

Following the initial outbreak of COVID-19, predictive models gained importance as a means of predicting the impact of the disease and planning accordingly. Indeed, such models were deployed by the Indian government, but they had a fatal flaw. As a team at the University of Michigan points out, this flaw was not a result of any technical problem with the models, but a social one – underreporting of cases. They state that "[insufficient data] limits modelers' ability to predict the course of the pandemic, gauge its impact, and estimate health care resource needs—including oxygen supplies and hospital beds" (Zimmermann, 2021, p. 560). Creating a successful model that creates accurate and useful data does not depend solely on proper application of mathematical formulas, but also on properly understanding the societal context in which the model exists.

For my model to avoid these pitfalls, I have two questions I want to answer: what datasets will I use to construct my model, and how can I be confident that my model gives accurate results? To answer these questions, I will adopt a relational view of data. This framework posits that data, by itself, carries no significant meaning. Instead, data "…consist of a specific way of expressing and presenting information, which is produced and/or incorporated in research practices […] and whose scientific significance depends on the situation in which it is used." (Leonelli, 2015, p. 811). The relational view of data shines the spotlight on the factors surrounding data over the data itself – how it was produced, who produced it, and what methods they used to get it. It is my hope that investigating these two questions using this framework will lead to a more accurate and useful model.

Digging deeper into how I intend to answer these questions, for the first I will have to decide which aspects of the real world – what *parameters* - are important enough to be included in my model. As Paul Edwards states in his book about climate modeling, "a parameter is kind of a proxy – a stand-in for something that cannot be modeled directly but can still be estimated, or at least guessed" (Edwards*,* 2010, p. 338). This "proxy" quality is the crux of what makes computer modeling so difficult – every parameter requires a tight balance between simplicity and accuracy. Some parameters, such as transmission and mortality rates, are rather obvious inclusions with clear mathematical ramifications. Others may be less obvious. As an example, a model constructed by members of the Information and Cognition Division at Cambridge constructed a pandemic model that accounted for media influence, and describes in detail the assumptions and mathematics that went into incorporating this (Kim, 2019). My model will no doubt have to account for similarly messy factors.

Zooming out towards the second question, I will also have to evaluate how useful the results my model gives can be. After all, my model is only a simplified version of the real thing – how can I be certain that useful conclusions can be drawn from what I create?

To answer this second half of my STS topic, I will once again return to climate models, as climatologists have been struggling to communicate their answers to this question for decades. Once again owing to the level of scrutiny place upon climate models, climatologists have found many metrics to evaluate the validity of their models. Elisabeth Lloyd, in a paper published to the reputable journal *Philosophy of Science* identified four major ones: robustness, or comparing many independently created models to see if they agree; variety of evidence, or displaying accurate behavior with regard to many independent variables; independent support, or how well the model matches with data that wasn't considered when constructing it; and model fit, or how accurately the model predicted real world happenings before they occurred (Lloyd, 2010).

However, there is some disagreement over whether these methods actually work. Also published in *Philosophy of Science*, Wendy Parker assesses the extent to which robustness is a useful indicator of model accuracy, and concludes that "while there are conditions under which robust predictive modeling results have special epistemic significance, scientists are not in a position to argue that those conditions hold in the context of present-day climate modeling" (Parker, 2011, p. 597). Evaluating whether or not my model is successful, it turns out, will also require evaluating my evaluation methods.

In fact, it might be easier to argue for the usefulness of my model rather than its accuracy. It might seem strange to draw conclusions from inaccurate models, but epistemologists seem to believe that this is the best way to treat statistical models. One article published about half a year into the COVID-19 pandemic reflected on the role of predictive models, concluding that “While all of these models are bound to be ‘wrong,’ some will be ‘useful’; and, together, the best of them offer complementary insights into the nature of the disease” (Ellison, 2020, p. 510). Ultimately, there’s a lot of conflicting information out there about how to evaluate computer models, and I intend to untangle this discourse in my STS paper to be able to more effectively evaluate my computer model.

4. Conclusion

For my technical topic, I intend to create a two-layer simulation that will identify methods of wastewater sampling that yield the most accurate data. My STS topic will focus on investigating what data I can use to inform this model and justifying the decisions and assumptions made in forming it. From this investigation I aim to contribute to the need for better contact tracing through improved wastewater surveillance, all the while gaining an improved understanding of the various social and technical concerns that go into creating a scientific model.

References

K. Bissett, J. Cadena, M. Khan, C. J. Kuhlman, B. Lewis and P. A. Telionis (21 April 2016). An integrated agent-based approach for modeling disease spread in large populations to support health informatics*. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).* 629-632

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Ellison, G. T. H. (2020, September 1). COVID-19 and the epistemology of epidemiological models at the dawn of AI. *Annals of Human Biology*, 47(6), 506 - 513.

Kim, L., Fast, S. M., & Markuzon, N. (2019, February 4). Incorporating media data into a model of infectious disease transmission. *PLoS ONE*, 14(2), 1 - 13.

Leonelli, S. (2015, December 1). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810 – 820.

Lloyd, E. (2010). Confirmation and Robustness of Climate Models. *Philosophy of Science*, 77(5), 971-984. doi:10.1086/657427

Parker, W. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science*, 78(4), 579-600. doi:10.1086/661566

Solano, F., Krause, S., & Wollgens, C. (2022, January 1). An Internet-of-Things Enabled Smart System for Wastewater Monitoring. *IEEE Access, Access, IEEE*, 10, 4666 - 4685.

Wade, M. J., Lo Jacomo, A., Armenise, E., Brown, M. R., Bunce, J. T., et al. (2022, February 15). Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the United Kingdom national COVID-19 surveillance programmes*. Journal of Hazardous Materials*, 424(Part B).

Zimmermann, Lauren V. et al. (2021, July 27). Estimating COVID-19– Related Mortality in India: An Epidemiological Challenge With Insufficient Data. *American Journal of Public Health,* 111(S2), S59 – S62.