UPPSALA
UNIVERSITET

PhD project: Towards an e-Science for Environmental Epidemiology

Stefan Engblom
Dept of Information Technology,
Uppsala university
stefane@it.uu.se, +46 18 471 27 54

This project will tackle the considerable challenges of *computational modeling* to *estimate* and *predict* the state of an epidemic. We will do so using data from novel monitoring techniques involving *non-traditional sources of data* in general and sewage water analysis in particular. At present, no data-driven disease-spread model exploiting this data stream is available, despite the fact that it has become a critical information source during the Covid-19 pandemic. The immediate benefits of this approach include low signal latency, adjustable geographical resolution, access to asymptomatic populations, and early detection of upcoming threats, e.g., new variants of already circulating viruses. We will develop nonlinear filters designed specifically with epidemiological models and high noise/low regularity signals in mind. The tools will align with the needs of the multidisciplinary SciLifeLab's Swedish Environmental Epidemiology Center (SEEC), a pandemic preparedness center for research, competence and technology development.

We will develop novel e-Science tools that concretely encompass:

1) **Data-driven computational modeling** underpinning *fusion* of heterogeneous data streams, for situation awareness, and prediction of societal impact of disease outbreaks;
2) **Model-based data analysis** for tracking of disease characteristics and evaluation of the impact of interventions through causal inference methodology.

By *data-driven computational modeling* we mean parametric computational models which are either informed dynamically by a series of measurements or, more broadly, non-parametric models that are inferred from data through machine learning. Topic 1) is based on the observation that *data fusion* is a key enabling technology. The computational modeling envisioned in the project therefore needs to deal with *heterogeneous* data streams, e.g., hospital resource allocation, test positivity, and app-based self-reporting systems. The complementary viewpoint of topic 2) requires the model to be formulated at a level of detail which allows for an interpretation of the variables, e.g., disease fatality or hospitalization rates.

The PhD project described herein will be carried out locally in a cross-disciplinary research environment formed within the CRUSH Covid initiative. The environment comprises a previously recruited PhD student, a Postdoc, and two seniors at the IT Dept, as well as on-going collaborations with SEEC. The long-term goal is to build up an internationally leading competence and experience in methods for epidemic awareness to *inform*, *improve*, and *evaluate* public health interventions.

This PhD position is part of the eSSENCE-SciLifeLab graduate school in data-intensive science. The school addresses the challenge of data-intensive science both from the foundational methodological perspective and from the perspective of data-driven science applications. It is an arena where experts in computational science, data science and data engineering (systems and methodology) work closely together with researchers in (data-driven) sciences, industry and society to accelerate data-intensive scientific discovery.

eSSENCE is a strategic collaborative research programme in e-science between three Swedish universities with a strong tradition of excellent e-science research: Uppsala University, Lund University and Umeå University.

SciLifeLab is a leading institution and national research infrastructure with a mandate to enable cutting-edge life sciences research in Sweden, foster international collaborations, and attract and retain knowledge and talent.

Project

During the Covid-19 pandemic, it was realized that public health decision-making support required an extensive use of multiple epidemiological data sources. The data heterogeneity involves not only varying temporal and geographical locality, but also sampling frequency, measurement noise, as well as ethical aspects and issues around data sensitivity. One of SEECs goals is research in wastewater-based epidemiology and the scope involves not only the quantification of endemic and emerging pathogens, but also other health markers such as drug abuse or the prevalence of antimicrobial resistance. We previously developed the technology for computational Bayesian modeling of realistic epidemiological processes [1] and within CRUSH Covid, this methodology was applied to Covid-19 in Sweden in general, and the Uppsala Region in particular [2]. Our work was communicated on a weekly basis to the Uppsala Region for half a year. Although the achieved *mean precision* of those models was satisfactory, the magnitude of the *uncertainty* is identified as a bottleneck. Wastewater analysis data are notably free from many limitations of the data provided by healthcare such as testing or through symptom quantification apps (e.g., Covid Symptom Study [3,4]). Sewage water analysis can be used locally through monitoring pump stations that serve individual buildings and by sequence analyses, the ratio of selected variants of concerns can also be determined.

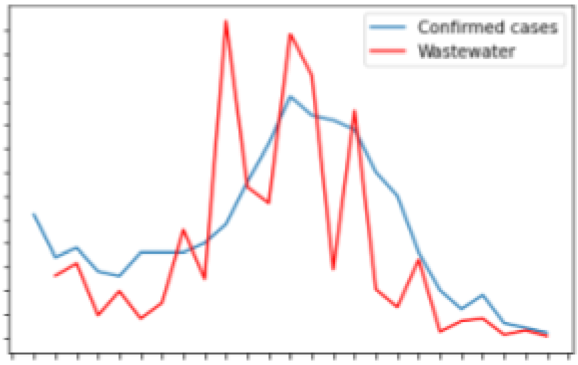

Fig. 1: Weekly confirmed cases of Covid-19 in Uppsala municipality together with data from wastewater measurements.

Task A: Time-series modeling and data-fusion

Mathematical models describing the spread of an infectious disease are typically designed in terms of *compartments* representing the involved populations, e.g., those individuals that are susceptible to the disease, infected by, or have recovered from it. The transitions from one compartment to another are usually formulated in terms of differential equations. To reflect the inherent randomness of the disease dynamics, the models are driven by stochastic processes.

Current solution As a concrete example, the compartments of our model for the spread of Covid-19 in Sweden are depicted in Fig. 2. The data for the compartments of hospitalized (H), intensive care (W), and deceased (D) are obtained from public healthcare, and estimates for recovered individuals (R) may be found in serological studies. Infected individuals are difficult to estimate due to biases in testing protocols and due to the presence of non-asymptomatic cases. As an alternative, these states can be estimated indirectly via the *infectious pressure* (φ) for which sewage water is a proxy signal. The model depends on parameters that are estimated in a Bayesian sense by combining prior information with that from the data. The forward evolution is defined by driving the dynamics using state-dependent Poissonian processes.

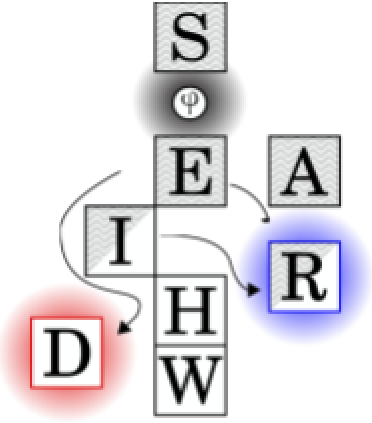

Fig. 2: Compartment model for Covid-19 in Sweden. Susceptible individuals become exposed via the infectious pressure variable.

Planned work By means of sewage water analysis, a signal directly related to the infectious pressure, and hence indirectly to *all* infected categories is obtained. Using only healthcare data and/or symptom apps, the size of the non-symptomatic population cannot be assessed. Therefore, the overall predictive performance of the data-fused model is expected to increase considerably with the incorporation of sewage water data. Non-symptomatic cases are particularly frequent in areas with predominance of young individuals, e.g., students, or when a high proportion of the population is vaccinated.

Task B: Development of input-output model

The model structure described in Task A follows the time series paradigm, where the measured variables are seen as the outputs of a dynamical system driven by a stochastic process. In epidemiology, such models are primarily used for illustrating and simulating infection spread scenarios and investigating the effects of mitigation interventions. Time series models possess high flexibility and can be effectively fitted to epidemiological datasets by manipulating model parameters and disturbance distributions.Yet, they are cumbersome in online use and produce volatile predictions. The main reasons for this are the lack of deterministic (measurable) inputs and the relatively high number of model parameters and variables.

Planned work The goal is to design a parsimonious model with minimal number of model parameters and variables. Using sewage water analysis data as a measured input related to the total number of infected individuals will result in a simpler dynamical model of increased fidelity. The expected performance improvement is thanks to the reduction of uncertainty in the compartments in the 'upstream' direction, and also conditioned by the novel input-output modeling paradigm that is not conventionally employed in epidemiology. Task B builds upon Task A since both require an evaluation of time delays arising in healthcare data with respect to the sewage water sampling.

Task C: Model parameter estimation

All mathematical models of practical use rely on parameters. In epidemiological models, the parameters involve human behavior and are therefore uncertain and changing over time, sometimes drastically so. When a mathematical model is fit to historical data, this does not pose problems since the parameters can be selected on average and yield a satisfactory approximation provided the model structure is flexible enough. The situation changes when the model is used for prediction and the current values of the model parameters have to be estimated from data in a streaming context. Then the model has to be parsimonious and satisfy identifiability criteria.

Current solution Bayesian estimation of the model parameters in the discussed Covid-19 model has been successfully developed by us. The dataset is split into intervals within which the parameters are assumed constant and an estimate is provided for each interval. This procedure results in piecewise-constant parameters with only a rudimentary ability to track the dynamics.

Planned work The goal is to devise tools for online adaptation of the models developed in Task A and Task B to the current epidemiological status. A recursive algorithm updating the parameter estimates as soon as new data have arrived will be developed. It has to satisfy at least two conditions: the estimates have to be sensible from an epidemiological perspective (e.g., positive) and should not diverge when the information content in the data is low (so-called insufficient excitation scenario). Together with a state estimation algorithm implemented on the basis of the models described above, the recursive model parameter estimation algorithm will provide a complete analysis pipeline for monitoring and predicting the epidemiological state.

References

[1] S. Engblom, et al.: Bayesian epidemiological modeling over high-resolution network data. *Epidemics*, 2020.
[2] R. Marin, et al.: Bayesian Monitoring of COVID-19 in Sweden. https://arxiv.org/abs/2205.00859, 2022.
[3] B. Kennedy, et al.: App-based COVID-19 syndromic surveillance and prediction of hospital admissions in COVID Symptom Study Sweden. *Nature Commun.,* 2022
[4] C. Menni, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature*, 2020.