PhD project: Computational Epidemics

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

The recent Covid-19 pandemic saw a boom of novel attempts for monitoring and predicting epidemics. These new developments have been both (i) on the data collection side with new data becoming available from, e.g., wastewater analysis, self-reporting symptom apps, and mobility data from cell phones, and (ii) on the modeling side where models of varying complexity have been proposed to make the most of the available data. The challenge of choosing a suitably complex model given the available data is treated in *system identification* and *machine learning*, while computational modeling itself is of course studied in *computational science.* This project sits at the intersection of these fields. Within real-life applications such as epidemiology, it is crucial to choose a sufficiently complex model that can make the most of the information in the available data, while avoiding overly complex models that cannot be trusted in decision-making throughout an epidemic. Another key aim is the ability to estimate and communicate modeling uncertainties, as this is vital to consider in crisis management.

The project will explore ways of navigating this fresh and exciting landscape in research concerning data-driven computational models for epidemics. The specific emphasis of the project will be flexible to the candidate's background and interests.

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 in a team comprising a previously recruited PhD student, a Postdoc, and two seniors at the IT Dept, as well as on-going collaborations with SEEC. The student will belong to either the division of Systems and Control or the division of Scientific Computing. 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

The following describes in more detail some of the current challenges and possible avenues for future research. We previously developed the technology for computational Bayesian modeling of realistic epidemiological processes [1] and this methodology was applied to Covid-19 in Sweden on the county level [2]. Although the achieved *mean precision* of those models was satisfactory, the magnitude of the *uncertainty* was identified as a bottleneck. Going forward, the hope is that non-traditional data streams along with new modeling approaches will improve lead to a shrinking of these uncertainties.

Challenge A: Data fusion in mechanistic models

The standard approach in epidemiological modeling is to consider data as outputs of a dynamical system driven by a stochastic process. One advantage of such models is that they offer a description of the underlying process and that their parameters are interpretable and meaningful in the real world. Recent developments have seen Kalman filters used as a natural way to fuse data from multiple data sources. This is an efficient way of making the most of all available data, which is important in epidemiology where the amount of data is typically scarce. A key aspect of Kalman filters is the *measurement model*, which reflects how we interpret the measured data in terms of the states of the underlying state space model. It is therefore a pressing matter for research to formulate suitable measurement models for each new data stream.

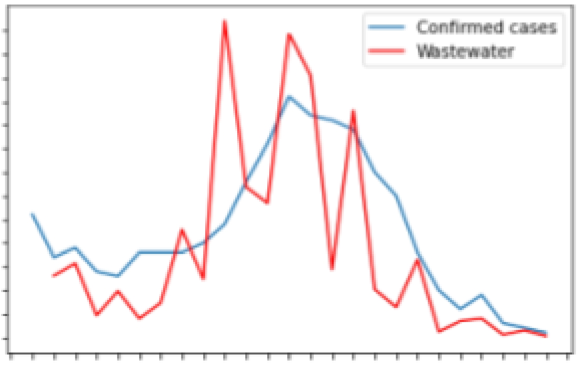

Fig. 1: Weekly confirmed cases of Covid-19 in Uppsala municipality together with data from wastewater measurements.

Challenge B: Development of input-output model

While the mechanistic models described above have many advantages, 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. This means they can often be out-performed by very simple naive models when it comes to real-time predictions. The goal here is to design a parsimonious model with a minimal number of model parameters and variables. For example, 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. This work could combine methods from both traditional system identification and modern machine learning approaches.

Challenge 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.

Finding suitable algorithms for parameter inference is a key area of research in computational epidemiology. In practice, the computational efficiency of the inference algorithm is a key factor. Methods combining aspects of Markov Chain Monte Carlo (MCMC), Approximate Bayesian Inference (ABC), particle- and Kalman filters have been proposed previously, and many other possibilities exist. The goal here is to formulate a coherent setup that is both computationally feasible and allows for a satisfactory time resolution of the time-varying parameters of interest.

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.