Disease Forecasting – wrt Covid-19

by R Greiner … [Overview, Publications] .. Sept 2024

Disease Forecasting is the task of predicting some "population property" of a disease, in a population, some time in the future. For example, predicting the number of Covid-19 cases (or alternatively, #deaths, or #ventilators_needed), in the city of Edmonton (alternatively, in Alberta, or Canada, or …) in 4 weeks (or perhaps 2 weeks, or 2 months). One obvious question is: What information is used for this prediction? One could use just the "current" number of Covid-19 cases at time t_0 (this is the time when we are making the prediction) – call that quantity c_0.[1] Alternatively, we might *also* include the number of Covid-19 cases yesterday ($[c_{-1}, c_0] = c_{-1:0}$), or perhaps also include 2 days ago (call these 3 values: $c_{-2:0}$), or maybe everything for last 1 week, the 7-tuple $c_{-6:0}$. This is a more powerful approach, since it allows us to detect trends in the data, including how quickly things are changing. We can go one step further and include not only the number of cases in the past 7 days, but also the number of deaths ($d_{-6:0}$), weather patterns, policy decisions (eg, whether schools are closed, whether masking is required, etc.), etc. This will give us more information, which we anticipate will allow us to make a more accurate prediction. Note the only requirement is that this information is known time t_0. Of course, we can include the occurrence of future holidays – eg, on t_0 = 11/Dec, we know that t_{14} = 25/Dec is Christmas, which means we can use the fact that "2 weeks from now is a Holiday". (Of course, we do not know whether the policy then will allow large gatherings, etc.)

In general, we can view a "forecasting scenario" as a 4-tuple

[quantity-to-estimate, region, forecast_interval, input]

– so if we used the #cases, #deaths over the past 1 week, to forecast the number of cases, in Edmonton, in 4 weeks:

[#Covid19_cases, Edmonton, 4weeks, $[c_{-6:0}, d_{-6:0}]$] .

One other issue: instead of just using the past 7 days of data, we might instead use all of the data after some fixed start time – say ४ = 1/Apr/2020. If so, write $c_{1/Apr/2020:0}$. Note this means different lengths of input – so the prediction associated with t_0 = 7/Apr, $c_{1/Apr/2020:0}$ would involve 7 values (from 1–7 / Apr), but for t_0 = 30/Apr, $c_{1/Apr/2020:0}$ would involve 30 values (from 1–30 / Apr), etc.

Intuitively, the more information we have, the better. However, it is very hard for humans to process and find patterns when dealing with so many variables. Therefore, we take a machine learning approach to this problem, where we attempt to learn a model for this prediction task. Of course, these techniques can use clever ideas from the (multi-dimensional) signal processing community (and the Natural Language community) as much of the data – like

$c_{-k:0} = \langle c_{0-k}, c_{1-k}, \ldots c_0 \rangle$

– is a time series; and from the general epidemiological models, using partial differential equations on quantities that correspond to the number of individuals within various compartments, etc.

There are some subtleties here, about

---

[1] To motivate this, imagine there are 3,000 new reported cases today. How many cases can we expect tomorrow? A good guess would be a number around 3,000 but we don't know if we should expect more or less cases. Is the number of cases going up? Is it going down? Or is it flat?

- what data will we train on – eg, just the single geographic region (say Edmonton), or perhaps also other regions as well (St Albert, Calgary, … Toronto, …, San Francisco, London?) ?
- how often we train – eg, do we train a model from one set of early data sequences (+labels), then use that model to make many predictions, at various time points,
  vs
  train a model from a set of early data sequences (+labels), use it once for a **single** prediction (so produce one model for 1/Apr, based on data before 1/Apr), then for the next day, learn a new model from a different set of sequence, and use that model once (so produce a new model for 2/Apr), etc.
- do we view the forecast_interval as an input to the learned model (so a single model could predict 7 days, or 14 days or …. by just changing that forecast_interval input), vs learning a model specific to single value of forecast_interval
- do we let the models learn everything from just the data itself, or do we incorporate priors that reflect what we already know about epidemics and their behavior?
- …

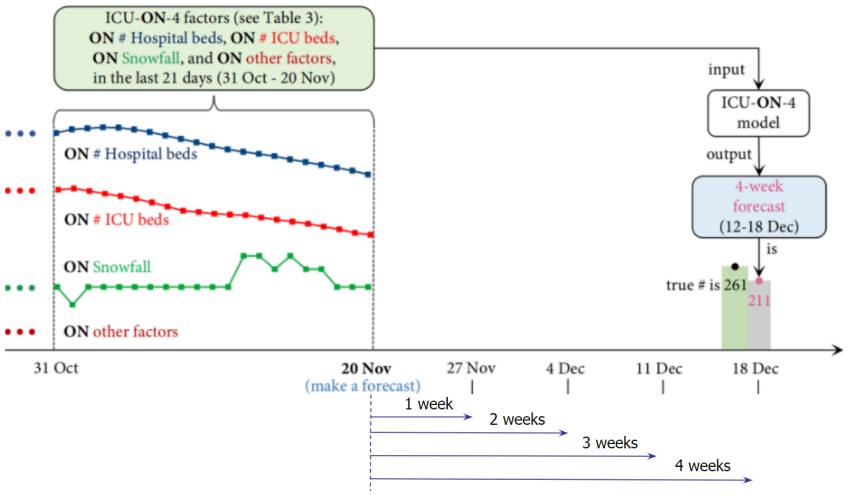
(Minor: each time could be a single day, or perhaps a week – ie, the sum of the events over those 7 days, etc.)

Note that the task of forecasting always involves predicting a quantity (or set of quantities) in the future (such as #case, on 29/Apr), using information available in the present (on or before 1/Apr). This learning approach is related to (although slightly different from) the task of identifying the features that contribute to accurate forecasts (eg, finding that "required masking" is relevant for a 2 week forecast of #Covid19_cases, in Alberta; or that "the weather for the past week" is NOT relevant). There is also work that attempts to predict the time until a disease outbreak, based on predicting when $R_0(t)$ will first exceed 1, etc. Our analysis differs in several ways – eg, we can measure quantities like #deaths, but R_0 is only estimated, etc. (Defining that start of an outbreak as "the first time t* when $R_0(t\ ^*) > 1$" is appropriate if we also know that $R_0(t + \kappa) > 1$ for all $\kappa > 0$, but if this is not true, it is not clear whether this time $t\ ^*$ is really the start of a true outbreak.)

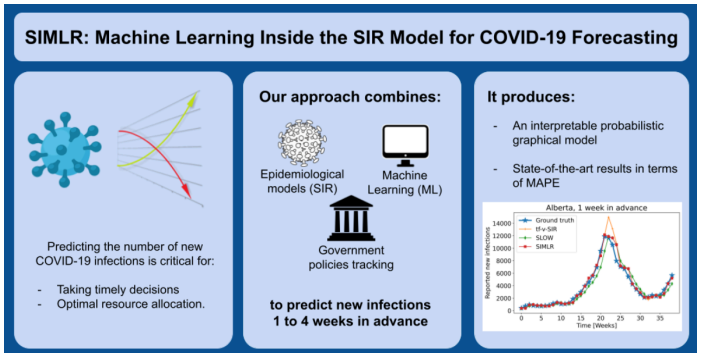
Recent results related to Covid-19 forecasting

- **Learning Models for Forecasting Hospital Resource Utilization for COVID-19 Patients in Canada**
  - (npj Scientific Report, 2022.11)
  - Characteristics .. every week from Oct 2020 to July 2021:
    - Predicting: #hospital_beds, #ICU_beds, #ventilators. #COVID-19_cases, #COVID-19_deaths
      - each WEEK (not daily amount)
    - Location: {Canada, Alberta, British Columbia, Manitoba, Ontario, Québec, Saskatchewan}
    - Forecast time: {1,2,3,4} weeks
    - Input: resource utilization, pandemic progress, population mobility, weather condition, public policy
      - Various different horizons .. typically 2 or 3 weeks

- Considered various sequence models, many from NLP… we found that TCN was the most accurate (MAPE)



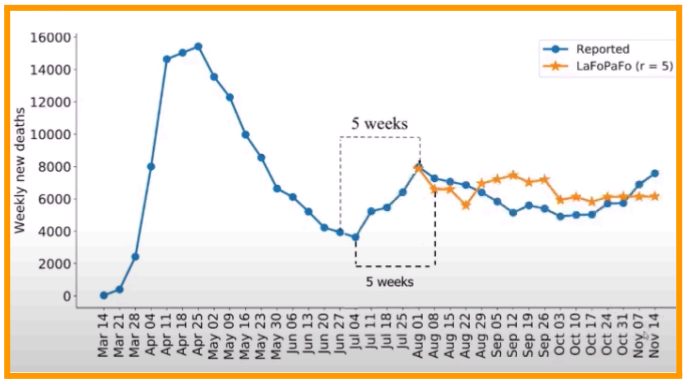
- [siMLr: Machine Learning inside the SIR model for COVID-19 forecasting]
  - (Forecasting, 2022.01)
  - Characteristics: every week from Oct 2020 to July 2021:
    - Predicting: #COVID-19_cases, #COVID-19_deaths
      - each week
    - Location: {US, Canada, Alberta, British Columbia, Manitoba, Ontario, Québec, Saskatchewan}
    - Forecast time: {1,2,3,4} weeks
    - Input: #COVID-19_cases, #COVID-19_deaths, public policy
      - Previous 3 weeks
  - Use ideas from the SIR model [Wiki]... but noticed this only worked when the policy was constant
  - So also considered a simple way to estimate when policies will change, and how (more severe, or less)
  - Won the **Forecasting 2022 Best Paper Award**. (link)

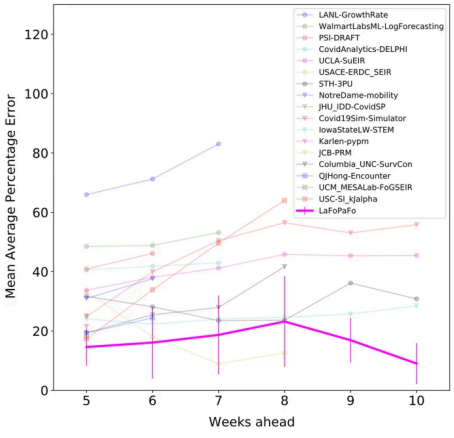


- **A Longitudinal Dataset of Incidence and Intervention Policy Impacts Regarding the COVID-19 Pandemic in Canadian Provinces**
  - (Data In Brief, 2021.10)
  - This did not present a specific forecast-learning model, but instead provides the information needed for forecasting models: here data on
    - socio-demographic, climatic, mobility (via Google cell phone records), daily numbers of new cases and deaths due to COVID-19, as well as COVID-19 intervention policies
    - implemented by the 10 provincial governments in Canada
    - from the beginning of February 2020 to the end of February 2021

- **Dataset of COVID-19 outbreak and potential predictive features in the USA**
  - (Data In Brief, 2021.10)
  - This did not present a specific forecast-learning model, but instead provided the information needed for forecasting models: here data on …
    - the daily number of COVID-19 confirmed cases and deaths,
      + 46 features that may be relevant to the pandemic dynamics: demographic, geographic, climatic, traffic, public-health, social-distancing-policy adherence, and political characteristics
    - from each of 3142 counties in the United States
    - from the beginning of the outbreak (January 2020) until June 2021.

- **Accurate Long-Range Forecasting of COVID-19 Mortality in the USA**
  - (npj Scientific Report, 2021.07)
  - See Researchers aim to forecast COVID-19 cases further into the future | Folio (2021.11)
  - Characteristics .. every week from January 2020 to November 2020
    - Predicting: #COVID-19_cases, #COVID-19_deaths
    - Location: US
    - Forecast time: {5,6,7,8,9,10} weeks
    - Input: #COVID-19_cases, #COVID-19_deaths, #COVID-19 tests, average daily temperature, average daily precipitation,
      + several social distancing related covariates (using Google mobility data) #individuals visiting parks, transit stations, residences, workplaces, grocery stores and pharmacies, retail shops and recreation centers.
      - from Jan 2020 until time t0





  - SotA for long range #cases predictions, 5weeks or more
  - Training: use LaFoPaFo (LAst FOld PArtitioning FOrecaster), not Cross-validation ..

- **Deep Learning for Disease Outbreak Prediction: A parallel LSTM-CNN model**
  - (Journal of Royal Society Interface; 2025.08)

    - Explores ways to identify EWS (Early Warning Signals) for predicting disease outbreak, that are robust to various noise sources

- **Early detection of disease outbreaks and non-outbreaks using incidence data: A framework using feature-based time series classification and machine learning**
    - (PLOS Computational Biology, 2025.02)
    - Describes a general model that can accurately forecast whether a temporal sequence of cases will later lead to an outbreak or not

- **An early warning indicator trained on stochastic disease-spreading models with different noises**
    - (Journal of Royal Society Interface; 2024.08)
    - Addresses the challenge of effectively forecasting an impending transition in a time series of disease outbreaks, even if the values are corrupted by various types of noise

NOTE: This g'doc lists only the results explicitly related to *Forecasting*. Our lab also had several other results related to Covid-19 – see Covid-19 Results – Forecasting and others…)