**Daniel Gutierrez**

**University of Padova**

**LM Data Science**

**Thesis Outline**

1. **Research question**

What is the most suitable model for early identification of a pandemic outbreak?

1. **Introduction**

Motivation for the topic: Finding a way to monitor and identify if a virus becomes pandemic to take appropriate action. As part of my work in an insurance carrier, it is important to understand better the speed and severity of a pandemic outbreak, as well as the drivers that allow it to become pandemic. The pandemics take a big toll on lives and resources from countries. If outbreaks are better understood, one can mitigate the next one, or even maybe hedge in some cases the economic losses associated with it.

1. **Related Work**

List some related work and methodologies:

* + Mostly people have focused on improving the SIR model given the last COVID-19 pandemic.
  + But there are some other modelling approaches like Bass Models, Gompertz
  + Network modelling (power law, Barabási, scale free): Consider individual as nodes and links are contagions, fit a model that describes the network growth.
  + Machine learning approaches like SVM, Linear Regression
  + Deep Learning approaches: Deep Neural Networks, Recurrent Neural Networks, Graphical Neural Networks.

1. **History**

Recap of major pandemic and epidemiological events and modelling:

* + Pre Modern:
    - Bernoulli (1760): model for smallpox
    - John snow (1854): spatial epidemiology approach for London cholera outbreak
  + 20th century:
    - SIR model / Kermack and McKendrick (1927): Developed this model for pandemic modelling, used it in past Spanish flu data (1918)
    - Influenza Pandemics and Stochastic Models (1950s – 1970s): In the Asian and Hong Kong Flu, epidemiologist refined stochastic models and the Reed-Frost and chain binomial models became popular.
    - HIV/AIDS and Network Models 1980s: Application of network-based models
  + 21st century:
    - SARS (2002-2004)
      * Agent Based Models (ABMs)
      * SEIR (Susceptible-Exposed-Infectious-Recovered), extension of SIR
    - H1N1 (2009):
      * Bayesian inference methods
      * Phylogenetic models
    - Ebola (2014-2016):
      * Spatial models with mobility data
      * Machine learning introduction
    - COVID-19 (2020)
      * Neural Networks with Bayesian Inference
      * Metapopulation models
      * Deep learning like GNNs

1. **Data**

Given the extensive availability of COVID-19 data, this seems like an attractive choice to experiment with different frameworks. The daily cases by country are found in this Johns Hopkins repository:

* + <https://github.com/CSSEGISandData/COVID-19>

Additional data related to COVID, like government measures to contain spread and mobility are found in the Oxford Government response tracker:

* + <https://github.com/OxCGRT/covid-policy-dataset/tree/main>

External data that can be used as predictors for the speed and severity of contagion on different geographies, like weather, development indicators (access to clean water, housing conditions, etc), population density, and some several indicators which could be tested as covariates in a regression framework sense. The main socioeconomic indicators can be found here:

* + <https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators>

Epidemics data in Colombia for the major outbreaks in the recent history: Dengue, Zika, Chicunguya. The data is collected from the INS and cleaned to just get the weekly new cases. Original data requests can be done in the following link:

* + <https://portalsivigila.ins.gov.co/Paginas/Buscador.aspx>

1. **Methodology**

Models will be fitted to some selected countries, for which evaluation metrics will be considered:

Two modelling perspectives will be done: One on the whole dataset to be able to find the best parameters that describe the pandemic evolution, because we already know the peak.

Another is to do train-test split. The idea is to split the data to the early stages of the pandemic and find the model that best predicts the overall evolution of the pandemic.

For the evaluation of the best models, I will use traditional metrics such as RMSE and R2.

1. **Modelling**

This section will implement different modelling approaches for the COVID-19 data in selected countries:

* Colombia
* Mexico
* Panama
* Dominican Republic
* Chile
* Brazil
* Uruguay

These countries are selected because of the interest of the insurance company to get more knowledge on the evolution of pandemics on the Latin American countries.

For Colombia the Dengue, Zika and Chicunguya will also be modelled.

For a separate regression framework, we can use almost all countries in the world to see which covariates have a stronger effect on the severity of the pandemic. In this way we have a larger sample to do some statistical inference.

**Time Series / Adoption modelling**

**1st part**

This section covers the time series approaches via partial differential equations like Bernoulli, SIR and BASS. The idea is to use GGMs as an alternative to SIR and BASS to account for the dynamic peak size for the case of the BASS, and the possibility to include external information like policy changes, public behavior, and environmental influences.

Comments:

* SIR is not so good because assumes either recover is immune or death / you can contract disease twice: SIS, SIRS, SEIS, SEIR, MSIR are alternatives for immunity.
* Bass assumes only one peak, so either model each “wave” separately or be ok with having bad fit
* Mass vaccination alternative
* The classic epidemic models (as described in the SIR epidemic models module or the SI models) were constructed as sets of ordinary differential equations (EDO), which implies several simplifying hypotheses. Ordinary differential equations with a given set of initial conditions always lead to exactly the same result. These models always consider all individuals to be equal, which implies that two individuals can interact with the same probability. This is hardly ever true in real systems.
* but in practice, each individual has a finite set of contacts to which he can transmit the infection: the set of all these contacts form a network
* Knowledge of the structure of the network allows models to calculate epidemic dynamics at the population level from the behavior of infections.

**2nd part**

In this section will focus on the network modelling approach, like Barabási, Erdős-Rényi, among others.

**3rd part**

Here the focus is on other modelling approaches that are more machine learning intensive: SVMs, Neural Networks (DNNs, RNNs).

**4th part**

Here I will combine some applications from the first part like GGM with external information that can be leveraged, for example spatial information, weather, containment measures to refine the GGM.

**5th part Regression Model: Feature Importance**

This section will focus on understanding which external structural factors have a bigger impact on the severity of the pandemic to different countries. The outcome won’t be the curve but the impact of the pandemic, which could be measured as the toll of deaths (observed now 5 years after) or financial distress.

**6th part Correlation with economic indicators**

Every pandemic is unique, but in this part, we can shortly cover how economic growth and associated indicators were affected, which can help understand in a future pandemic how to mitigate the losses.

1. **Results comparison**

Compare the results among models highlighting accuracy and usability for policy recommendations.

1. **Conclusions and recommendations for further work**