Master Thesis Corona - Notes

Mike Weltevrede

March 2020

1 Model overview

Define:

- $I(t)$: fraction of infected individuals at time t,
- $R(t)$: fraction of recovered individuals at time t,
- $S(t) = 1 - I(t) - R(t)$: fraction of the population that is susceptible at time t,
  - *This assumes that recovery implies immunity*
- a : rate at which new cases develop,
- b : rate of recovery.

Then $R_0 \coloneqq \frac{a}{b}$ is the effective reproduction rate of the virus. If $R_0 < 1$, the virus can be said to be subsiding. Policies are made to halt a.

Standard Inflammatory Response model (SIR)

$$\begin{cases} \frac{dI(t)}{dt} &= aS(t)I(t) - bI(t) \rightarrow \text{daily increase in the fraction of } \textbf{infected} \text{ individuals} \\ \frac{dR(t)}{dt} &= bI(t) \rightarrow \text{daily increase in the fraction of } \textbf{recovered} \text{ individuals} \end{cases}$$

Incidence rate at time t
I.e. the number of new infections in a population P.

$$\begin{aligned} Inc(t) &= \left(\frac{dI(t)}{dt} - \frac{dR(t)}{dt}\right) P \\ &= aS(t)I(t)P \end{aligned}$$

Incidence rate in a region r at time t (Adda, page 922)

$$
\begin{aligned}
Inc(r,t) &= Inc(r,t-lag)\cdot S(r,t-lag)\cdot \sum_k a(k,within)W(k,r,t-lag)\\
&+ \sum_{k,c\neq r} a(k,between)\widetilde{W}(k,c\neq r,t-lag)\cdot Inc(k,c\neq r,t-lag)\cdot S(k,c\neq r,t-lag)\\
&+ X(r,t)\cdot d + e(r,t) \qquad (1)
\end{aligned}
$$

where

- $a(k, within)W(k, r, t-lag)$: k within-region spatial weights,
- $a(k, between)\widetilde{W}(k, c \neq r, t-lag)$: k across-region spatial weights for each other region c,
- lag is defined by Adda as the incubation period,

and

- $W(k, t)$: known spatially heterogenous weights for which we will gather data. For connections within Europe, free data is available from EUROSTAT and for the world from the World Bank.
  - EUROSTAT: this data only contains up to 2017 or 2018.
  - EUROSTAT and World Bank: this data only contains per country aggregated data, not a spatial matrix from country to country.
  - WTTC. This has PDFs with data (so not nicely importable) per country with the top 5 inbound and outbound travel in 2019. They say "Note: Data are average shares over the 2015-2017 period. Source: Oxford Economics, national sources and UNWTO"
  - UNWTO, which provides data free of charge to students and academic researchers. So, unfortunately, not open source.
  - NS, partially open source but not in a nice CSV format. Also, this is only available for 2018.
- $a(k, \cdot)$: unknown coefficients to be estimated from the data,
- $X(r, t)$: includes political regimes, development index and population density, but also includes region-time dummies to capture the effects of potentially unobserved characteristics, such as cultural norms and news, medical capacity shortages etc.

2 Our contributions

1. **New data:**

- **COVID-19 data**: Daily data on new cases, recoveries and deaths from COVID-19 for all countries and provinces within many countries is available at https://github.com/CSSEGISandData/COVID-19 free of charge. However, airline transportation data is only partially available.
- **Transportation data**: We will construct the spatial weights separately for different transportation means: airline, railway transport and road transport. For connections within Europe, free data is available from EUROSTAT, and for the world from the World Bank. We will carefully assess all freely available data, and while there are private data sources which would lead to more accurate modelling, we refrain from using these, as we want to produce a model that can be used in real-time, free of charge by all researchers and policy-makers in the case of COVID-19 but also in case we experience a new viral disease outbreak in the future.
- **Other characteristics subsumed in** $X(r, t)$: These are freely available from the World Bank. We will address the issue of whether these characteristics should be included as additional covariates or as spatial weights.

2. **Real-time prediction of the infection rates:**
   The model in equation (1) allows prediction of infection rates many days ahead with and without policy measures.

   - **No endogeneity concerns**: Unlike Adda (2016), who was interested in the effect of a transportation strike or school closure on the coefficients $a(k, within)$ and $a(k, between)$, we are interested in forecasting the infection rate, and therefore endogeneity (contemporaneous changes in policy measures and shocks in the new infection rates, such as availability of test kits) is not a problem in our analysis. We can also allow for time changes in the coefficients $a(k, within)$ and $a(k, between)$ as the virus spreads, as long as enough time-series observations are available for a particular country.
   - **Heterogeneity in spatial transmission**: we can allow for the coefficients $a(k, within)$ and $a(k, between)$ to depend on the region, as long as there are enough time-series observations in that region.
   - The data is counts of new infections with many zeros (or missing data) for many regions early on, therefore offering the possibility to model this via a **count maximum likelihood model** for spatial data with truncated observations.

3. **Predicting case fatality rates heterogeneously across countries:**
   The case fatality rate can be estimated with higher accuracy as the outcomes of the patients can be predicted based on jointly modelling infection and recovery data.

3 Findings

3.1 Which country do we analyse?

- https://github.com/CSSEGISandData/COVID-19 → Only has region-based information for China and the USA (with information split by overseas territory for countries like mainland France and the Netherlands).
- World Bank → For the Netherlands, railway passenger data is confidential.
- https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Italy#Statistics → has detailed data on Italy by region, sex/age, and date (per region).
- https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_the_Netherlands#Statistics → has detailed data on the Netherlands per date (per province).
- Dr. Boldea asked the RIVM for more detailed data.

We have decided to **use data on Italy** (first).