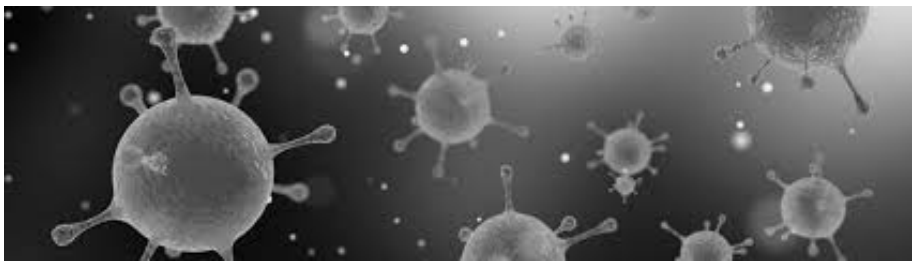

Project Report

Authors

Daniel Juhász Vigild - s161749

Lau Johansson - s164512

Jonas Søbro Christophersen - s153232

Anders Nikolai Fure Nielsen - s192288

May 30, 2020

1 Project scope

We structure the project around the following two purposes:

- Create a PGM that can predict daily new deaths related to COVID-19 for multiple countries.
- Evaluate and compare some of the theories regarding what impacts the variance in morbidity of COVID-19 between countries/regions.

1.1 Methodology

To model daily new deaths related to COVID-19, we implement a Normal Linear Dynamic System. We use information from different data sources to train our models. These can be generally put into three categories: 1) data directly registering COVID-19 deaths and confirmed cases, 2) health and demographic data from OECD, 3) data measuring social distancing from the Google Mobility Reports. Our analysis is structured around the following steps:

- Create a simple Normal LDS model.
- Add complexity in the emission probabilities.
- Add external covariates from OECD data.
- Add external covariates from the Google Mobility Report and confirmed cases of COVID-19.

The steps shown above was done in hope of achieving acceptable prediction accuracy. By incorporating external covariates from OECD regarding health and demographics we wish to evaluate and compare some of the theories regarding characteristics of people and their environments that can impact the morbidity of the virus. These are that: 1) morbidity is higher for people with preexisting conditions relating to heart, lungs and diabetes, 2) morbidity is higher in older populations, 3) areas with high levels of air pollution are more profoundly affected, 4) the amount of exposure to the virus has a significant impact on morbidity. The Google Mobility Data allows us to evaluate the effects of social distancing.

1.2 Countries and variables

We analyse COVID-19 in the following countries: Denmark, Sweden, France, Spain and Italy. We use the following variables:

Variables	Comments
COVID-19 data	
New daily deaths	
Daily new confirmed	
Days since first death	
OECD data	
Acute myocardial infarction	Deaths related to disease pr. 100.000 inhabitant
Asthma	−||−
Chronic obstructive Pulmonary diseases	−||−
Diabetes mellitus	−||−
Diseases of the respiratory system	−||−
Influenza	−||−
Ischaemic heart diseases	−||−
Pneumonia	−||−
Population share 65-80 years	
Population share above 80 years	
Air pollution	
Rooms per person	
Number of people per household	
Single person households	
Google Data	Daily percent change according to baseline from January-February
Retail and recreation	regarding where people are physically present across categories
Grocery and pharmacy	
Parks	
Transit stations	
Workplaces	
Residential	

Table 1: Variables in dataset.

2 Results

The MSE for the five countries and four models are seen in the figure below.

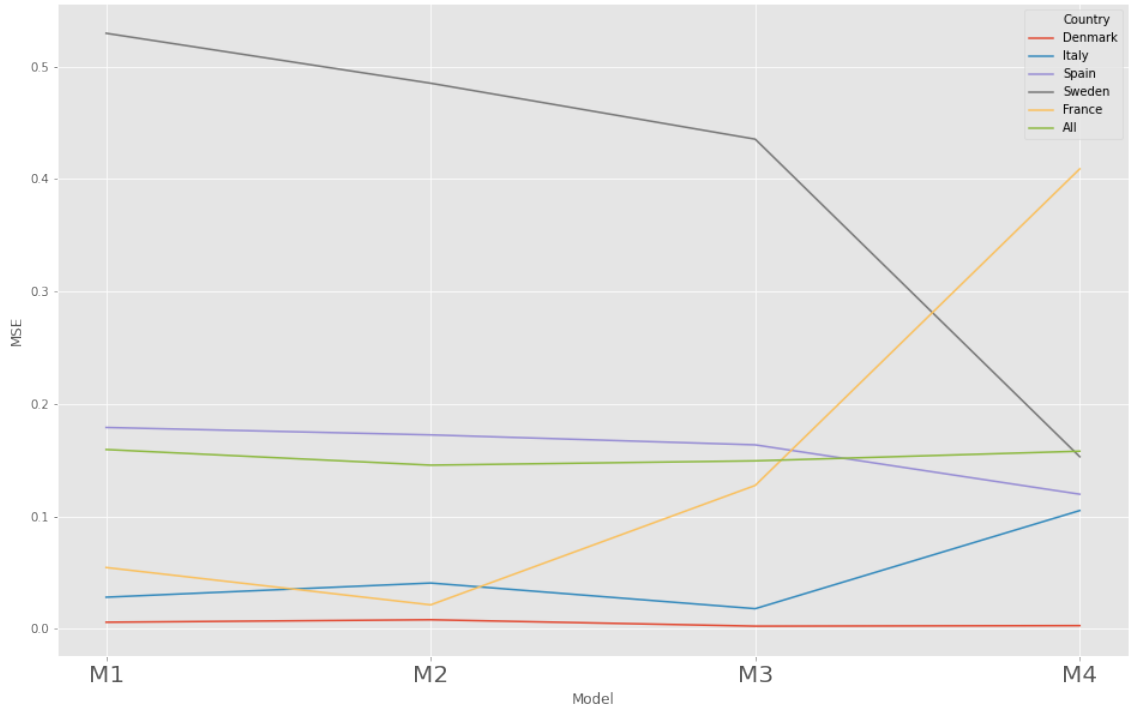

Figure 1: MSE for the five countries and four models. From simpler to more complex (more data sources) from the left.

As the figure shows adding complexity to the model did not improve performance in prediction of the validation data. As seen the prediction-performance was actually lowered when utilizing the Google-data, which might be a sign of overfitting to the training data, or that there weren't a sufficient amount of data available to thoroughly test hypothesis regarding social distancing (Google mobility data) or country sociodemographics (OECD data).

The results from Figure 1 shows that the mean MSE across the countries for each model extension is decreased by changing the emission probability. When including OECD and google data, the MSE is increased. Results are ambiguous when looking at country specific MSE's as it is seen that for Sweden the MSE performance improves whereas the MSE for France deteriorates with addition of external covariates.

2.1 Insights and conclusions

Since we did not observe any noticeable increase in performance when adding external OECD-variables to the model, one is compelled to conclude that these variables are not useful in determining the mortality of COVID-19. Since other research projects have shown the opposite, it might also be due to some of the modelling choices we made. Rather, we conclude that it is not possible - based on our models - to confidently compare and evaluate the theories on what variables impact COVID-19 morbidity.

The model we proposed is (for the sake of simplicity, and knowing we did not have a lot of data at hand) relying on simplistic linear relations, and we did not investigate whether there might exist non-linear relations which would improve model predictions. Additionally it might be the case that a sudden addition of many external covariates caused the model to overfit as the degrees of freedom for the model decreased drastically. The modelling approach might have been better when utilizing more data (whether it being longer time-series, or addition of countries), but unfortunately neither of these two were possible.

3 Links

Hyperlink to explanatory notebook (notebook viewer)
Hyperlink to explanatory notebook (Google Colab gist)
Hyperlink to Appendix Notebook (notebook viewer)
Hyperlink to Appendix notebook (Google Colab gist)