Predicting The Incidence Rate And Case Fatality Rate Of COVID-19

by
Mike Weltevrede (SNR 1257560)

A thesis submitted in partial fulfillment of the requirements for the degree of Master in Econometrics and Mathematical Economics.

Tilburg School of Economics and Management
Tilburg University

Supervised by:
dr. Otilia Boldea

Second reader:
dr. George Knox

Date:
June 24, 2020

Abstract

TODO

Acknowledgements

TODO

Contents

1 Introduction **1**

2 Problem description **2**

3 Methodology **4**
3.1 Modelling undocumented infections 4
3.2 Model 1: Within-Region Spread 7
3.3 Model 2: Weighted Within-Region Spread 8
3.4 Model 3: Within and Between-Region Spread 8
3.5 Model 4: Full Model 9

4 Dataset **10**
4.1 Coronavirus data 10
4.2 Independent variables 12

5 Results **14**
5.1 Model 1: Within-Region Spread 14

6 Conclusion **19**

7 Future research **20**

Appendices **23**

A Tables **23**

1 Introduction

2 Problem description

In this section, we will elaborate on the methods that we apply in order to model the epidemiological spread of SARS-CoV-2 and the disease it causes: COVID-19. We are basing our model on specifications as used by Adda (2016).

Q: Is it common practice to later refer simply to Adda's paper as "Adda" or should we always put "Adda (2016)"?

In the paper, Adda (2016) investigates the spread of several viral diseases in the past, namely influenza, gastroenteritis, and chickenpox. The key additions made are, firstly, that a spatial spillover effect is considered and, secondly, that we allow for some sort of weighting on the parameters on the basis of region specific variables. Adda (2016) starts from the Standard Inflammatory Response (SIR) model, the most commonly used model in epidemiology (Anderson & May, 1992; Kermack & McKendrick, 1927).

The SIR model splits the total population into three groups. S denotes the fraction of individuals who are susceptible to being infected, I denotes the fraction of individuals who are currently infected, also called infectives, and R denotes the fraction of individuals who have been removed from the model, be that because they successfully recovered from the disease or that they have deceased. tadda2016economic defines R to be the group of individuals who have recovered but who are still immune, i.e. the deceased people are not included in R.

Q: All other sources except for Adda look at the group R as the removed, i.e. people who overcame the disease but also deaths. It is not clear how Adda deals with deaths, although I suspect he subtracts them from the susceptible population and adjusts the total population accordingly. I think it is likely not an issue since the number of deaths is negligible compared to the size of the large population. (Also see the next TODO note)

As such, at any point in time, we have that

$$S, I, R \in [0, 1] \text{ and } S + I + R = 1.$$

The SIR model is postulated in continuous time, i.e. the equations in (2.1), (2.2), and (2.3) depict the change in the variables S, I, and R, respectively, for one time period ahead. This type of model is also called a stock-and-flow model because there is a certain stock (for instance the number of infectives) to which a flow is added or subtracted.

$$\frac{dS}{dt} = -\alpha SI + \lambda R \tag{2.1}$$

$$\frac{dI}{dt} = \alpha SI - \beta I \tag{2.2}$$

$$\frac{dR}{dt} = \beta I - \lambda R \tag{2.3}$$

It is important to grasp the main assumptions of the SIR model, which also tell us how these equations are constructed. The first assumption that is made, is that the population is constant, meaning that births and deaths are ignored.

TODO: Should see if this matters. I currently do not take them into account, apart from the deaths due to COVID

Next, note that the spread of the virus is determined by the interaction between the infectives and the susceptible population. The second assumption that is made under the SIR model in this light is that there is a constant rate of change in infectives that is proportional to this interaction between the infectives and the susceptible population. This is represented by the term αSI in equations (2.1) and (2.2), which is also called the transmission term (Keeling & Rohani, 2011). The third assumption that

the SIR model makes is that there is a constant rate of change at which infectives recover or decease. This relates to the term βI in equations (2.2) and (2.3).

TODO: Look into this, if we do use the definition that people who die are included. Reason: the fuller hospitals are, the more people will likely decease.

Finally, we assume that there is a constant rate of change at which immune individuals lose their immunity. This is denoted by the term λR in equations (2.1) and (2.3). For instance, Adda (2016) mentions that λ is set to 0 for chickenpox as individuals acquire a lifetime immunity while λ will be high for gastroenteritis due to almost no immunity emerging. In the case of COVID-19, some studies show that it is likely that individuals who recovered from COVID-19 may be immune to reinfection, at least temporarily (Kirkcaldy et al., 2020). Nonetheless, no definitive results have been shown.

TODO: Explain more later on immunity since this is currently still researched.

One of the main measures resulting from the SIR model is the estimation of the basic reproduction number $R_0 \coloneqq \alpha/\beta$. An epidemic is said to develop if $R_0 > 1$.

TODO: Explain how $R_0 > 1$ is determined (why not 2, for instance).

This measure is widely used to indicate that an ongoing epidemic is dying out if R_0 drops below 1. For instance, the Italian health ministry has posted an article on May 9, 2020 stating that the R_0 reproduction rate for COVID-19 was below 1 in Italy, at between 0.5 and 0.7 (Ministero della Salute, 2020a), showing that this measure is also used communicated to citizens as a way of informing them whether the pandemic is tending to end.

3 Methodology

In this section, we will explain the methodology applied in this thesis. We will discuss our models and the thought process behind them. Before that, it is important to understand the concept of an incubation period. This is defined as the period between an infection and the moment that the infected individual starts showing symptoms, at which point the infective is said to be symptomatic. Note that this is not the same as the period between an infection and the moment that the infective is infectious, which is called the latent period. For COVID-19, the latent period is estimated to be approximately 2 days shorter than the incubation period (He et al., 2020). That is, there are infectives who are able to infect others before showing symptoms. We call these people pre-symptomatic, which is distinctive from asymptomatic people in the sense that asymptomatic people do not develop symptoms and pre-symptomatic people will develop symptoms but they develop a higher viral load just before said symptoms are apparent.

On June 9, 2020, the World Health Organization (WHO) said that it is unclear whether asymptomatic people can actually spread the virus but that pre-symptomatic people may actually be able to infect others (Sutherland & Gretler, 2020). This may be an issue when considering policies such as self-isolation when one is sick, because an infective may have already spread the virus before feeling sick. Sutherland and Gretler (2020) moreover reiterate the WHO's statement that studies have been done that show that asymptomatic people can spread the virus but that more research needs to be done to show how many of these infectious asymptomatic people exist. We will discuss how we deal with pre-symptomatic individuals in section 3.1. After this, we will discuss the applied models in sections 3.2 until 3.5, where we also discuss the incubation period more.

If discussed earlier, move this piece of text.

3.1 Modelling undocumented infections

A common concern with the spread of viruses, especially one so rapidly spreading as SARS-CoV-2, is that there is no possibility to test the entire population on whether they are infected because the testing capacity is simply not there. If this were possible, then all individuals who were tested to be positive could be isolated and the spread of the virus would be dampened tremendously. However, since this is not possible, there are likely many infectives in society who spread the virus but who are undocumented. In China, around 86% of the infections went undocumented (R. Li et al., 2020). R. Li et al. (2020) also estimate that these were also contagious, with around 55% of the contagiousness of documented infectives. This was investigated during the period from January 10 till January 23, 2020, so considering a lack of major restrictions such as travel bans. R. Li et al. (2020) make the important note that these results are indeed highly dependent on the specific situation in the country of interest, for instance due to differences in testing, case definition, and reporting. Nonetheless, even if these numbers are lower for other cases, such as Italy under lockdown, this research shows that undocumented infections should be taken into account.

In this section, we aim to model the undocumented infections. Note that, by definition, there is no data on the amount of undocumented infections because, otherwise, these cases would indeed be documented. As such, some assumptions need to be made since we cannot apply *supervised learning* methods (being models where there is a data on a dependent variable to predict) to determine the number of undocumented infections. Firstly, we assume that the amount of undocumented individuals is decreasing as the testing capacity increases. Similarly, the amount of documented individuals increases in the testing capacity. The logic behind this is clear: as more people are tested, more infective move from being undocumented to being documented. Secondly, as mentioned, R. Li et al. (2020) consider that there are no major restrictions. As we know, Italy has been under a strict national lockdown. This was imposed on March 10, 2020. The restrictions were relaxed around May 18, when businesses were allowed to reopen and citizens were allowed free movement within the region they live in, although they

were still barred from travelling to other region unless they had an essential motive (Severgnini, 2020). Therefore, the model of undocumented infections should possibly take this into account.

Make sure to either do this or remove it from the text

At a point in time t, we define the testing capacity as TC_t. In section 4.1, we explain how a measure of the testing capacity is obtained. The total number of infected people at time t is denoted by I_t. This group can be subdivided into the documented infections DI_t and the undocumented infections UI_t such that $DI_t + UI_t = I_t$. We can denote the documented and undocumented infections as proportions of the total number of infected people, at any point in time. As mentioned before, this proportion may change over time as the testing capacity increases, among others. As such, we define this proportion as a function of the testing capacity over time:

$$f_t \coloneqq f(TC_t), \tag{3.1}$$

such that

$$\begin{cases} DI_t &= f_t I_t \\ UI_t &= (1 - f_t) I_t. \end{cases} \tag{3.2}$$

Note that we can then rewrite the expression for UI_t as:

$$UI_t = (1 - f_t) I_t = \frac{1 - f_t}{f_t} DI_t.$$

This quantity can be calculated because we have data on the number of documented individuals.

Let us consider some properties that (3.1) should satisfy and some assumptions that we make.

1. Since f_t is a proportion, we need to have $f_t \in [0, 1]$.

2. If no one is tested, we assume that these are not counted as documented infections. That is,

   $$f(0) = 0.$$

   Of course, individuals could be documented as being infected when they show sufficient symptoms. However, we assume that this is not the case.

3. Denote the total population at time t as N_t. Then, if the entire population is tested, we assume that all infections will be documented. That is,

   $$f(N_t) = 1.$$

   This assumes that the tests that are executed are perfect at determining whether someone actually is infected. However, it is common knowledge that such tests have a certain rate of false positives and negatives. In the case of COVID-19 specifically, positive screening tests are not followed-up (as is common practice to confirm a diagnosis) because of scarcity in testing resources and/or prioritization of allocating tests to the sickest patients (Frasier, 2020).

   Q: Should I give a numerical example to show how a high accuracy can still lead to large false positives and negatives or is this considered to be common knowledge?

4. As mentioned before, f_t needs to be monotonically increasing in TC_t. That is, the proportion of infectives that are documented is increasing in the testing capacity.

If $R_0 > 1$, we expect this to have an decreasing growth rate (like the downwards closing parabola), i.e. the more people are tested (assuming a positive test leads to the proper consecutive measures such as self-isolation), the less people are infected by this documented infective so that the rate of increase in the number of undocumented infections also decreases. Therefore, the group of documented infectives slowly dominates the undocumented infectives. However, it is difficult (not possible?) to take R_0 into account in this calculation. Perhaps a two-step approach is possible but likely out of the scope of this thesis.

Clearly, we can generalise this to include regions...

This definition can easily be generalised to be applicable to regions, by considering the total population in that region $N_{r,t}$ instead of the total population. Then, the function would be dependent on r as well:

$$f_{r,t} \coloneqq f(TC_{r,t}). \tag{3.3}$$

such that

$$\begin{cases} DI_{r,t} &= f_{r,t} I_{r,t} \\ UI_{r,t} &= (1 - f_{r,t}) I_{r,t}. \end{cases} \tag{3.4}$$

We test several functional forms of the function f_t:

- Linear form: Since $f(0) = 0$ and $f(N_t) = 1$, the formula becomes:

$$f_t = \frac{1}{N_t} TC_t.$$

- Downwards opening parabola: We assume that the vertex is the point $(N_t, 1)$. Using that $(0, 0)$ is on the parabola, we can then derive that the formula becomes:

$$f_t = -\frac{1}{N_t^2} TC_t^2 + \frac{2}{N_t} TC_t.$$

- Upwards opening parabola: We assume that the vertex is the point $(0, 0)$. Using that $(N_t, 1)$ is on the parabola, we can then derive that the formula becomes:

$$f_t = \frac{1}{N_t} TC_t^2.$$

- Cubic form: We assume that $f(\frac{1}{2}N_t) = \frac{1}{2}$ and that $f(\frac{1}{4}N_t) = L$, where $L \in \left(0, \frac{1}{2}\right)$. Then the formula becomes:

$$f_t = aTC_t^3 + bTC_t^2 + cTC_t + d$$

where

$$\begin{cases} a &= \frac{64L-16}{3N_t^3} \\ b &= \frac{8-32L}{N_t^2} \\ c &= \frac{32L-5}{3N_t} \\ d &= 0. \end{cases}$$

For instance, for $N = 1000$ and $L = \frac{2}{5}$, we have that the formula evaluates to:

$$f_t = 3.2 \times 10^{-9} \times TC_t^3 - 4.8 \times 10^{-6} \times TC_t^2 + 2.6 \times 10^{-3} \times TC_t.$$

TODO: Add a derivation of these functional forms in the appendix.

In Figure 3.1, we have plotted these specifications for a total population at time t of $N_t = 1000$ and a value for $L = \frac{2}{5}$ for the cubic form.

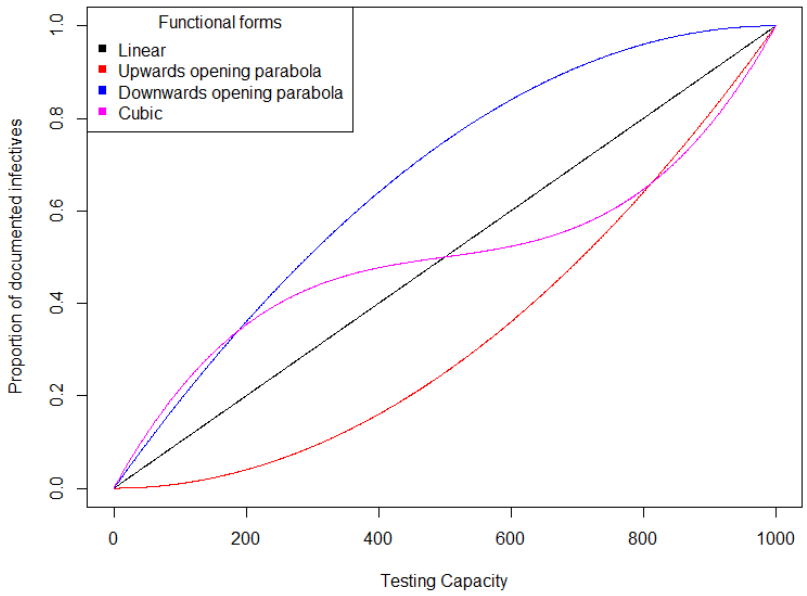

Figure 3.1: Functional forms for the proportion of documented infectives $\left(N_t = 1000, L = \frac{2}{5}\right)$

TODO: Continue here: how is this incorporated in the models and how do we check the specification?

3.2 Model 1: Within-Region Spread

We start with a simple model ignoring effects across regions. The within-region model is henceforth defined as:

$$I_{r,t} = \alpha_{within} I_{r,t-\tau} S_{r,t-\tau} + X_{r,t}\delta + \eta_{r,t} \tag{3.5}$$

where the subscript τ is a lag indicating the length of the incubation period. The incubation period for COVID-19 is estimated to be above 2 and below 11.5 (Lauer et al., 2020), 12.5 (Q. Li et al., 2020), and 14 days (Linton et al., 2020). This is a large range, but this is not rare. For instance, the incubation period for chicken pox is estimated to be between 9 and 21 days (Papadopoulos, 2018). While the maximum incubation period is not agreed upon by Lauer et al. (2020) and Q. Li et al. (2020), their results on the median are similar. Lauer et al. (2020) reports a median incubation period of 5.1 days (95% CI: 4.5 to 5.8 days), while Q. Li et al. (2020) reports a median incubation period of 5.2 days (95% CI: 4.1 to 7.0 days). For comparison, Linton et al. (2020) give the result of a mean incubation period of 5.0 days (95% CI: 4.2 to 6.0 days) when excluding Wuhan residents and 5.6 days (95% CI: 5.0 to 6.3 days) when including Wuhan residents. Due to the results from Lauer et al. (2020), Q. Li et al. (2020), and Linton et al. (2020), we choose $\tau = 5$.

TODO: Possibly test multiple lags again.

The matrix X includes fixed effects for regions, as well as weekend and week of the year dummy variables.

TODO: Elaborate on the definition of X and why we choose these variables, possibly in the Dataset section.

Lastly, we include an idiosyncratic error term η. The model is estimated by ordinary least squares (OLS). Because fixed effects for regions are included in X, note that this means that running OLS is actually a least-squares dummy variables (LSDV) regression. In general, the main issue with LSDV regression is that there needs to be an indicator variable for each observed individual (in our case, these are regions). However, it is feasible to run an LSDV regression since we consider a relatively small number of regions with a large number of time periods. This will be explained more in section 4.

TODO: If we remove the regional dummies, remove this text.

3.3 Model 2: Weighted Within-Region Spread

In the previous model, it has been assumed that the incidence rate within a certain region is only determined by the previous incidence rates plus some other effects. However, the transmission rate α is likely influenced by other factors as well. These may include policies, such as shutting down restaurants or public transport, but also persistent regional characteristics such as metrics on the quality of hospitals or economic development. In this section, we incorporate these factors in the within-region model (3.5). After defining the between-regions model in section 3.4, we will apply the same methodology to obtain the full weighted model in section 3.5.

Let the tensor W contain K region-specific variables that may influence the transmission rate α. As such, we now allow for α_{within} to differ for these K variables. In section 4, we elaborate on how these variables included in W are specifically defined and selected. For instance, we include the number of rail travellers, which changes over time, but also a measure of the development of health care through the number of available hospital beds, which does not change over time. We define X and η in the same way as for (3.5). Taking this into account, the weighted within-region model is defined as:

TODO: Possibly update later

$$I_{r,t} = I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha_{within}^{k} W_{r,t-lag}^{k} + X_{r,t}\delta + \eta_{r,t} \tag{3.6}$$

3.4 Model 3: Within and Between-Region Spread

A key addition made by Adda (2016) is recognizing that there is spatial spillover between regions. That is, there may be infectives in one region that travel to another region and then infect individuals there. As such, the number of new cases would be modeled as $\alpha_{within} SI + \alpha_{between} S\tilde{I}$ where $\tilde{I}$ is the fraction of infectives from outside the region of interest who meet susceptible people from within the region. Clearly, this is an important addition to the model and we acknowledge and incorporate this in this thesis.

TODO: Consider the difference in definition in I between the SIR model and Adda. Possibly use the notation from Keeling and Rohani (but this includes X)

The following model is defined:

$$\begin{aligned} I_{r,t} = {} & \alpha_{within} I_{r,t-lag} S_{r,t-lag} \\ & + \alpha_{between} S_{r,t-lag} \sum_{c \in R \setminus r} I_{c,t-lag} \\ & + X_{r,t}\delta + \eta_{r,t} \end{aligned} \tag{3.7}$$

In (3.7), the transmission parameter α is now allowed to be different within and between regions. Adda (2016) estimates (3.7) by OLS and by instrumental variable estimation (IV). Weather episodes, such as the amount of rain and temperature-related instruments, are used as instruments. There is a biological reasoning behind choosing these instruments, for instance that warmer temperatures tend to

have a negative effect on the proliferation of some viruses. A social reason is also given, namely that bad weather conditions impact the amount of social interaction between people, meaning that there are less opportunities for viruses to spread. We challenge the choice of these instruments, particularly in the case of SARS-CoV-2. Firstly, we do not have sufficient information on the effect of the weather on the virus. That is, SARS-CoV-2 has only been quite apparent since January 2020 and there has not been enough fluctuation over time in temperatures to show a necessary effect that can be disentangled from, for example, policies being effective in driving the virus back. Secondly, we challenge the social reasons entirely, although not quantitatively. In our view, bad weather conditions in themselves are not likely to be strong enough instruments for the viral spread. That is, even if they are indeed exogenous with respect to the error term and that they are correlated with the viral spread, we expect this to not be quite strong.

TODO: This is currently a claim and I have not looked at Adda's quantitative tests for these instruments.

For this reason, we only consider OLS for this model.

3.5 Model 4: Full Model

We now incorporate the between-region effects as well as the weighting of the transmission parameter. In addition to (3.6), we now also put weights on the between-region transmission parameter by some possibly influential variables. Let the tensor $\widetilde{W}$ contain $\tilde{K}$ variables that now can influence the transmission rate α_{within} between two regions r and c.

TODO: Possibly consider not following Adda's notation with the tildes and use something like V and L instead of $\widetilde{W}$ and $\tilde{K}$, respectively.

$$\begin{aligned} I_{r,t} = I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha_{within}^{k} W_{r,t-lag}^{k} \\ + S_{r,t-lag} \sum_{c \in R \setminus r} I_{c,t-lag} \sum_{k=1}^{\tilde{K}} \alpha_{between}^{k} \widetilde{W}_{r,c,t-lag}^{k} \\ + X_{r,t}\delta + \eta_{r,t} \end{aligned} \tag{3.8}$$

4 Dataset

In this section, we will outline the structure of the data that is used and how it was retrieved. Firstly, we discuss the structure of Italian regions and the reasons why we chose to use Italy as our region of interest. Subsequently, we will look at the data on COVID-19, such as the incidence rate. Here, we also discuss how we tackled possibly errors in the data, as well as missing values. Lastly, we consider the variables that are included in the weighted models in sections 3.3 and 3.5.

Italy has been one of the most intensely struck countries in the world. On June 6th, 2020, it had the seventh highest absolute number of cases, after the United States, Brazil, Russia, the United Kingdom, Spain, and India. Despite dropping in this positioning, Italy reports the highest death-to-cases ratio of 14.47%, followed closely by the United Kingdom, which reports a death-to-cases ratio of 14.18%. The sudden onset of the spread of SARS-CoV-2 put immense pressure on the Italian hospitals, especially in the northern regions such as Lombardy. This forced patients with coronavirus-caused pneumonia to be sent home as well as literal collapses of overworked healthcare workers (Horowitz, 2020).

TODO: Add table

The Italian ministry of Health Services (Ministero della Salute) has posted daily reports containing tables with a detailed numerical overview of new cases, active intensive care (IC) patients, and tests executed, all divided up between the second-level NUTS regions (also called NUTS 2 regions). The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the European Union (EU) and the United Kingdom (Eurostat, 2020a) as used by Eurostat, the statistical office of the EU. Italy consists of 21 so-called *regioni* (regions), comparable to Dutch provinces. These constitute the second-level NUTS regions (also called NUTS 2 regions), where the region of *Trentino-Alto Adige* is split into two regions: *Provincia Autonoma di Bolzano/Bozen* and *Provincia Autonoma di Trento*. Italy's first-level NUTS regions are defined as groups of regions, namely *Nord-Ovest* (North West), *Nord-Est* (North East), *Centro* (Center), *Sud* (South), and *Isole* (Islands). The third-level NUTS regions are 107 provinces, which are subregions of the *regioni*. Ideally, we would want to have coronavirus data on the NUTS 3 regions since many policies are introduced at that level, such as a lockdown put into place on March 7th, 2020 until the strict national lockdown was instated. Unfortunately, the data was not reported at this granular level. As such, we chose to use the NUTS 2 regions.

4.1 Coronavirus data

For $R = 21$ Italian regions, we retrieved the data on the coronavirus from January 31st, 2020, until June 13, 2020. Some time gaps occurred in the beginning of the data, leading to a total amount of time observations of $T = 116$. It was retrieved from the Ministero della Salute, who publish daily reports under a title similar to *Covid-19, i casi in Italia: 10 giugno ore 18*, where *10 giugno* (June 10th) would be updated to the relevant date (Ministero della Salute, 2020b). These daily reports were posted with tables containing the following information per region:

TODO: Update accordingly if this changes, also T=x

- Hospitalized with symptoms *(Ricoverati con sintomi)*
- Active intensive care patients *(Terapia intensiva)*
- Home isolated active cases *(Isolamento domiciliare)*
- Total number of active cases *(Totale attualmente positivi)*
- Dismissed/recovered *(Dimessi/guariti)*
- Deceased *(Deceduti)*
- Total confirmed cases *(Casi totali)*

- Increase in total confirmed cases - compared to the previous day *(Incremento casi totali - rispetto al giorno precedente)*
- Total amount of tests executed *(Tamponi)*
- Total amount of persons tested *(Casi testati)*
- Increase in total amount of tests executed *(Incremento tamponi)*

The difference between the total amount of tests executed *(tamponi)* and the total amount of persons tested *(casi testati)* is that the latter indicates the number of unique persons that were tested. That is, individuals could have been tested more than once. Do note that *tamponi* is a good indication of the *testing capacity* as the number of tests that Italy is able to execute. Henceforth, when the term *testing capacity* is used, this refers to *tamponi*, unless indicated otherwise.

It should be noted that there is a measurement error in the number of infectives, as is the case in any other country. This is because there is no possibility that every citizen can be tested for COVID-19. For that reason, the actual number of infectives is higher that the official count as reported in the tables of the Ministero della Salute. With respect to the reported death statistics, there is a distinction between Italy and some other European countries. Namely, the Italian numbers include deaths of all patients who were tested for COVID-19 before or after their death, regardless of whether they died inside or outside the hospital. In contrast, other countries may only count deaths in hospitals. French death counts, for instance, only have included deaths at hospitals and clinics caring for patients, excluding people who die at home or in care homes, although the French president Emmanuel Macron did announce that these centers would be tracked from the first week of April onward (Sevillano, 2020). Moreover, Italian data makes no distinction between people who died because of COVID-19 or simply had the disease but who died from other causes (also referred to as comorbidities). Patients who had pre-existing conditions actually make up around 96% of the total death count in Italy (Istituto Superiore di Sanità, 2020). In some other countries, such as Germany, a distinction between these two groups is actually made (Caccia, 2020).

We also make the note that it is unclear how the Ministero della Salute collects its information. If regions or provinces submit this information to the government each day, there may be areas that fail to submit their data for a certain day. Despite this, we assume that this official information is accurate and representative of the region for which it has been reported. If this is not the case, the numbers in the report on the next day will compensate for the error on the day before. In the official publications that we use, data that was wrongly published on a day $t-1$ is corrected by subtracting the error from or adding the error to the cases from day t. As such, if the error is larger than the number of new cases, the reported amount of new cases is negative. It happened fifteen times that a number was reported to be negative. Since negative numbers are not logical in the context of confirmed cases or deaths, we correct this by subtracting the error from the day before and set the previously negative number to 0. In the case that the error on day t is larger than the number on $t-1$, for instance if a value of -10 is reported on day t while the value for day $t-1$ is less than 10, we propagate the error to multiple lags until this issue no longer occurs. For non-negative corrected numbers, we do not have a way to detect which these are and we cannot reasonably assume how this number should be split up among day t and $t+1$. As such, these are left as is. One note that should be made is a highly negative value of -229 reported for the region of Campania on June 12, 2020, whereas the number of new cases in the week before that date only ranges from 0 to 5. We assume that this corrects for all errors in the past, not just those near to June 12. Propagating this error backwards as described before would lead to zero new cases per day for Campania from May 13, 2020 onward.

TODO: Update accordingly if this changes

TODO: Because the report above is highly unrealistic and we have no way of knowing how this correction is distributed across the days, should we delete the region of Campania entirely? We can also distribute it to some measure. See notes of 10th thesis meeting.

Regarding missing data, there are only three cases, namely for Abruzzo on March 10, Puglia on March 16, and Campania on March 18. Given that faulty data is also corrected as described before, we assume that the cases missing on day t are added to those of day $t+1$. This is confirmed by higher values compared to the expected trend, as seen in Table 4.1. As such, missing data is simply imputed with a value of 0.

TODO: Update accordingly if this changes

Table 4.1: Number of confirmed cases around a day t with missing data

	Abruzzo	Puglia	Campania
Day $t-1$	8	64	60
Day $t+1$	46	110	192
Day $t+2$	5	43	97

4.2 Independent variables

Independent variables, or regressors, were obtained from Eurostat, which is the statistical office of the European Union (Eurostat, 2020b). Statistical data, broken down to the three NUTS levels, are published on their website. The data can be freely filtered according to year, geolocation (being the NUTS regions), and other aspects depending on the data, such as sex, age, or the unit of measure. Unfortunately, this data is not available daily and is often not up-to-date. That is, sometimes data is available up to 2016. For each variable, we kept the most recent data and assumed that this would be representative for the present. In Table 4.2 we mention per variable in what year the most recent observations were.

TODO: Fix this and look up the actual maximum year per variable

We distinguish three sets of regressors, as mentioned in section 3. Firstly, we have a set of control variables included in the tensor $X_{r,t}$ which are not assumed to have a (large) effect on the transmission parameter α. Secondly, the tensor $W_{r,t}$ consists of variables that are assumed to affect the transmission within regions. Lastly, the matrix $\widetilde{W}_{c,r,t}$ contains variables that are assumed to affect the transmission between regions. The specification of these regressors can be found in Table 4.2.

TODO: Insert $\widetilde{W}$ variables and possibly move around variables to X

Table 4.2: Specification of regressors

Matrix	Variable	Year	Description
$X_{r,t}$	weekend	n/a	Binary indicator denoting if the day is on the weekend (Saturday or Sunday)
	weekNumber	n/a	The calendar week number
$W_{r,t}$	airPassengersArrived	2018	Number of air passengers arrived
	touristArrivals	2018	Number of tourist arrivals
	broadbandAccess	2019	Percentage of population that has access to broadband internet
	deathRateDiabetes	2016	Number of deaths from diabetes per 100,000 inhabitants
	deathRateInfluenza	2016	Number of deaths from influenza per 100,000 inhabitants
	deathRateChd	2016	Number of deaths from coronary heart disease per 100,000 inhabitants
	deathRateCancer	2016	Number of deaths from cancer per 100,000 inhabitants

Table 4.2 continues on next page

Table 4.2 continued from previous page

Matrix	Variable	Year	Description
	deathRatePneumonia	2016	Number of deaths from pneumonia per 100,000 inhabitants
	availableBeds	2018	Number of hospital beds
	riskOfPovertyOrSocialExclusion	2018	Percentage of population at risk of poverty or social exclusion
$\widetilde{W}_{c,r,t}$			

One of the most important aspects in interpreting the results of a regression analysis is that interpretations are made under the *ceteris paribus* assumption. That is, we look at the effect of a change in one variable while holding all other variables constant. Because of this, there should be no large correlation between our independent variables. If there would be a large correlation between some regressors, then it is not possible to consider a change in one variable without causing a change in some other variable(s). Specifically for our case, we concur that there are people who often have multiple diseases at the same time and that there is likely a large correlation between the various death rates. To investigate this, we consider the correlation matrix in Figure 4.1. As described before, these variables are unfortunately not varying over time but they do vary over the regions. Because we are using the region-wise correlation, do note that a small sample size of $R = 21$ is used. Therefore, the numbers should be taken with a grain of salt.

TODO: Cite a source on low sample size w.r.t. correlations

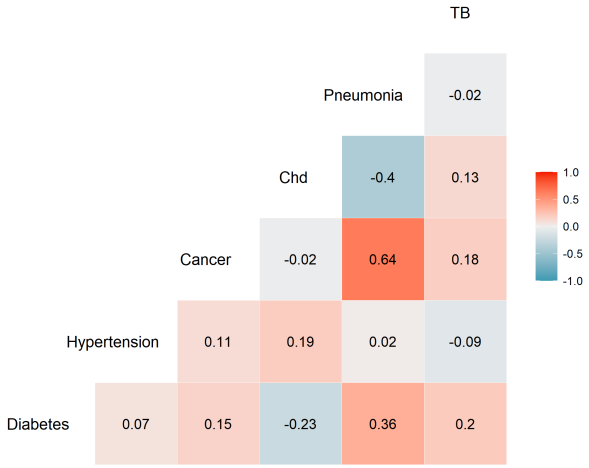

Figure 4.1: Correlation matrix of the discharge rates for various comorbidities of COVID-19

Figure 4.1 shows us that the largest correlation is 0.64 and occurs between the discharge rates of pneumonia and cancer. We also see a relatively high correlation of -0.4 between the discharge rates of pneumonia and CHD. For this reason, we remove the discharge rate of pneumonia from the model.

5 Results

In this section, we present the results from the models as presented in section 3.

5.1 Model 1: Within-Region Spread

In Table 5.1, we present the results from a nationwide model.

Table 5.1: Estimates from Model 1: Within-region spread on a national level

	Estimate	Std. Error	t value	p value	
Intercept	577.6258	196.0201	2.947	0.003812	**
Weekend	516.5431	138.8212	3.721	0.000295	***
Week number	-37.3267	11.4075	-3.272	0.001370	**
α_{within}	0.9150	0.0342	26.757	$< 2 \times 10^{-16}$	***

Significance levels: * = 0.05, ** = 0.01, *** = 0.001

The estimate for α_{within} is 0.9150 and is statistically highly significant. Note that it seems to be quite small compared to the other estimates. This is because this represents the estimated effect of only a unit change in $I_{t-\tau}S_{t-\tau}$. Because Italy has many inhabitants, and $S_{t-\tau}$ is generally quite close to 1, this means that a unit change is relatively small.

We have now seen the results for the national model. Of course, this does not take into account effects specific to regions. In Table 5.2, we present the results from running the model on each region separately with the same model specification for each region. After that, we present the results from applying model selection with the Bayesian Information Criterion (BIC) in Table 5.3.

TODO: Replace p-values with standard errors and add stars

Table 5.2: Estimates from Model 1: Within-region spread per region without model selection. Estimates are given with p-values in parentheses.

Region	α_{within}	Intercept	Weekend	Week number
National	0.92	577.63	516.54	-37.33
	(0.00)	(0.00)	(0.00)	(0.00)
ABR	0.63	15.10	7.36	-0.54
	(0.00)	(0.03)	(0.14)	(0.16)
BAS	0.61	2.08	1.42	-0.09
	(0.00)	(0.06)	(0.08)	(0.16)
BZ	0.68	14.02	2.94	-0.56
	(0.00)	(0.01)	(0.47)	(0.08)
CAL	0.57	6.81	3.42	-0.27
	(0.00)	(0.03)	(0.14)	(0.15)
CAM	0.79	20.33	1.81	-0.89
	(0.00)	(0.01)	(0.75)	(0.05)
EMR	0.88	81.40	53.69	-4.60
	(0.00)	(0.01)	(0.01)	(0.00)
FVG	0.67	16.35	5.12	-0.64
	(0.00)	(0.01)	(0.27)	(0.08)
LAZ	0.89	21.22	9.33	-1.10

Table 5.2 continues on next page

Table 5.2 continued from previous page

Region	α_{within}	Intercept	Weekend	Week number
	(0.00)	(0.01)	(0.09)	(0.01)
LIG	0.81	29.77	13.71	-1.29
	(0.00)	(0.01)	(0.10)	(0.05)
LOM	0.83	269.44	179.70	-13.02
	(0.00)	(0.01)	(0.01)	(0.01)
MAR	0.84	23.86	16.54	-1.33
	(0.00)	(0.01)	(0.01)	(0.01)
MOL	0.35	1.64	2.22	-0.01
	(0.00)	(0.24)	(0.03)	(0.88)
PIE	0.87	75.51	62.28	-4.14
	(0.00)	(0.02)	(0.01)	(0.03)
PUG	0.80	16.33	3.88	-0.70
	(0.00)	(0.02)	(0.42)	(0.07)
SAR	0.61	7.96	2.44	-0.31
	(0.00)	(0.03)	(0.34)	(0.13)
SIC	0.78	13.56	3.72	-0.59
	(0.00)	(0.02)	(0.35)	(0.06)
TN	0.71	15.24	20.75	-0.76
	(0.00)	(0.06)	(0.00)	(0.10)
TOS	0.85	29.77	27.75	-1.72
	(0.00)	(0.02)	(0.00)	(0.02)
UMB	0.74	7.19	3.71	-0.36
	(0.00)	(0.04)	(0.14)	(0.08)
VDA	0.46	7.96	6.40	-0.33
	(0.00)	(0.03)	(0.02)	(0.11)
VEN	0.83	66.63	32.94	-3.42
	(0.00)	(0.01)	(0.06)	(0.01)

Table 5.2 shows that the estimate for α_{within} varies vastly over the regions, from to 0.32 for Molise till 0.91 for the national model. Not counting the national model, the highest estimate for α_{within} is attained for the regions of Emilia-Romagna and Lazio, namely 0.88.

In Table 5.3, we see that the BIC gives a varying model selection per region. All models retain the intercept and the term $I_{t-\tau}S_{t-\tau}$ in the model. In 6 out of 22 cases, the entire model is selected (including the national model). In 12 other cases, only the intercept and the term $I_{t-\tau}S_{t-\tau}$ are kept in the model. The other 4 models select either Weekend of Week number on top of this but not both.

Table 5.3: Estimates from Model 1: Within-region spread per region with model selection by stepwise BIC. Estimates are given with p-values in parentheses.

Region	α_{within}	Intercept	Weekend	Week number
National	0.91	565.05	494.80	-35.32
	(0.00)	(0.00	(0.00)	(0.00)
ABR	0.62	9.03		
	(0.00)	(0.00)		
BAS	0.61	1.13		
	(0.00)	(0.01)		

Table 5.3 continues on next page

Table 5.3 continued from previous page

Region	α_{within}	Intercept	Weekend	Week number
BZ	0.69	5.96		
	(0.00)	(0.01)		
CAL	0.57	3.69		
	(0.00)	(0.00)		
CAM	0.80	6.85		
	(0.00)	(0.03)		
EMR	0.88	81.40	53.69	-4.60
	(0.00)	(0.01)	(0.01)	(0.00)
FVG	0.67	7.87		
	(0.00)	(0.00)		
LAZ	0.88	24.00		-1.10
	(0.00)	(0.00)		(0.01)
LIG	0.80	14.61		
	(0.00)	(0.01)		
LOM	0.83	269.44	179.70	-13.02
	(0.00)	(0.01)	(0.01)	(0.01)
MAR	0.84	23.86	16.54	-1.33
	(0.00)	(0.01)	(0.01)	(0.01)
MOL	0.32	2.18		
	(0.00)	(0.00)		
PIE	0.87	75.51	62.28	-4.14
	(0.00)	(0.02)	(0.01)	(0.03)
PUG	0.80	6.60		
	(0.00)	(0.02)		
SAR	0.62	3.77		
	(0.00)	(0.01)		
SIC	0.79	5.37		
	(0.00)	(0.02)		
TN	0.72	3.33	20.74	
	(0.00)	(0.38)	(0.00)	
TOS	0.85	29.77	27.75	-1.72
	(0.00)	(0.02)	(0.00)	(0.02)
UMB	0.75	2.59		
	(0.00)	(0.05)		
VDA	0.47	2.72	6.39	
	(0.00)	(0.08)	(0.02)	
VEN	0.82	77.51		-3.43
	(0.00)	(0.00)		(0.02)

TODO: Put the other plots of alpha over time in. Also make the plots larger (put them next to each other?).

We are interested in looking at the estimate of α_{within} over time. That is, if we keep adding data, do we see an interesting effect in its progression? We use at least 50 data points. Each point in the graphs in Figures 5.1 and 5.2 is the estimate of α_{within} when only data before that date was used.

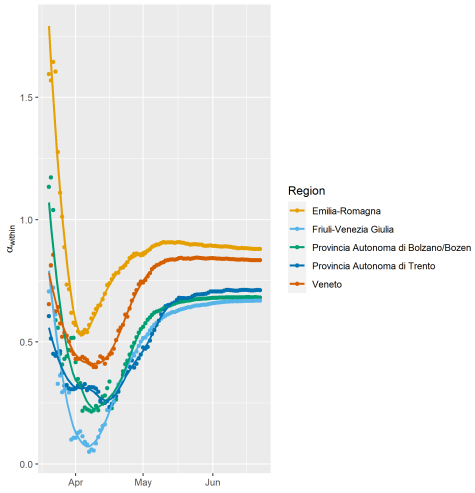

Figure 5.1: Progression of α_{within} over time for the North-East NUTS-1 region (without model selection)

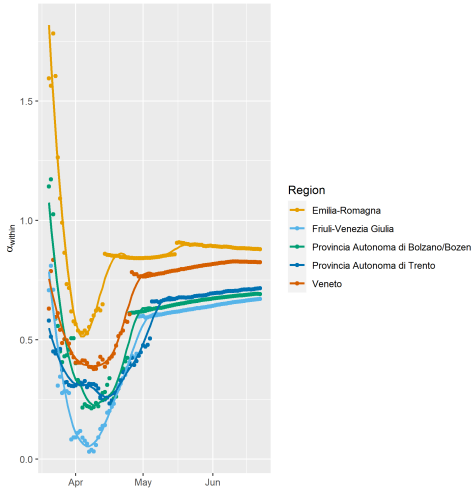

Figure 5.2: Progression of α_{within} over time for the North-East NUTS-1 region (with model selection)

Table 5.4: Estimates from Model 3: Within and Between-Region Spread

	Estimate	Std. Error	t value	p value	
Intercept	1.198×10^{-3}	1.161×10^{-3}	1.032	0.302247	
Weekend	1.282×10^{-5}	1.664×10^{-6}	7.704	1.84×10^{-14}	***
Week Number	-6.939×10^{-7}	1.391×10^{-7}	-4.989	6.47×10^{-7}	***
Median Age	-2.517×10^{-5}	2.455×10^{-5}	-1.025	0.305426	
α_{within}	-0.1104	1.972×10^{-2}	-5.598	2.39×10^{-8}	***
$\alpha_{between}$	0.04845	1.467×10^{-3}	33.036	$< 2 \times 10^{-16}$	***

Significance levels: * = 0.05, ** = 0.01, *** = 0.001

TODO: These results are outdated and should still be updated.

The estimate for α_{within} is -0.1104 and is statistically highly significant. Comparing this to the results from model 1, the sign has flipped. We also notice that the sign has flipped for the region of Trento. The estimate for $\alpha_{between}$ is 0.04845, which is also statistically highly significant.

6 Conclusion

7 Future research

References

Adda, J. (2016). Economic activity and the spread of viral diseases: Evidence from high frequency data. *The Quarterly Journal of Economics*, *131*(2), 891–941.

Anderson, R. M., & May, R. M. (1992). *Infectious diseases of humans: dynamics and control*. Oxford University Press.

Caccia, F. (2020). *Coronavirus, "il conteggio dei morti varia da paese a paese. La Germania esclude chi ha altre patologie"*. Retrieved June 11, 2020, from https://www.corriere.it/cronache/20_marzo_22/coronavirus-il-conteggio-morti-varia-paese-paese-germania-esclude-chi-ha-altre-patologie-6a452e6a-6c19-11ea-8403-94d97cb6fb9f_preview.shtml

Eurostat. (2020a). *Eurostat regional data background*. Retrieved June 11, 2020, from https://ec.europa.eu/eurostat/web/regions/background

Eurostat. (2020b). *Eurostat regional statistics database*. Retrieved June 11, 2020, from https://ec.europa.eu/eurostat/web/regions/data/database

Frasier, S. L. (2020). *Coronavirus antibody tests have a mathematical pitfall*. Retrieved June 19, 2020, from https://www.scientificamerican.com/article/coronavirus-antibody-tests-have-a-mathematical-pitfall/

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine*, *26*(5), 672–675.

Horowitz, J. (2020). *Italy's Health Care System Groans Under Coronavirus — a Warning to the World*. Retrieved June 11, 2020, from https://www.nytimes.com/2020/03/12/world/europe/12italy-coronavirus-health-care.html

Istituto Superiore di Sanità. (2020). *Caratteristiche dei pazienti deceduti positivi all'infezione da SARS-CoV-2 in Italia*. Retrieved June 11, 2020, from https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, *115*(772), 700–721.

Kirkcaldy, R. D., King, B. A., & Brooks, J. T. (2020). COVID-19 and Postinfection Immunity: Limited Evidence, Many Remaining Questions. *JAMA*.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, *172*(9), 577–582.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., Et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, *368*(6490), 489–493.

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B., Kinoshita, R., & Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of clinical medicine*, *9*(2), 538.

Ministero della Salute. (2020a). *Coronavirus: Contagion rate R0 below 1. Prudence needed in phase two says ISS*. Retrieved June 11, 2020, from http://www.salute.gov.it/portale/news/p3_2_1_1_1.jsp?lingua=italiano&menu=notizie&p=dalministero&id=4717

Ministero della Salute. (2020b). *Covid-19, i casi in Italia: 10 giugno ore 18*. Retrieved June 11, 2020, from http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioNotizieNuovoCoronavirus.jsp?lingua=italiano&menu=notizie&p=dalministero&id=4877

Papadopoulos, A. (2018). *Chickenpox: Practice essentials, background, pathophysiology*. Retrieved June 22, 2020, from https://emedicine.medscape.com/article/1131785-overview/

Severgnini, C. (2020). *Discorso di Conte in conferenza stampa, le riaperture dal 18 maggio: "corriamo un rischio calcolato"*. Retrieved June 18, 2020, from corriere.it/politica/20_maggio_16/discorso-conte-conferenza-stampa-oggi-decreto-18-maggio-1e810142-9785-11ea-ba09-20ae073bed63.shtml

Sevillano, E. (2020). *Tracking the coronavirus: why does each country count deaths differently?* Retrieved June 11, 2020, from https://english.elpais.com/society/2020-03-30/tracking-the-coronavirus-why-does-each-country-count-deaths-differently.html

Sutherland, J., & Gretler, C. (2020). *WHO now says role of silent virus spreaders remains unclear*. Retrieved June 18, 2020, from https://www.bloomberg.com/news/articles/2020-06-09/who-says-symptomless-spread-is-rare-in-jolt-to-virus-efforts

Appendices

A Tables