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:
July 10, 2020

Abstract

TODO

Acknowledgements

TODO

Contents

1 Introduction

Since the beginning of 2020, the novel coronavirus SARS-CoV-2 (causing the viral disease COVID-19) has plagued the world. Starting from Wuhan, China, it has made its way to every single continent and (nearly) every country in the world. In response to SARS-CoV-2, governments have been implementing far-reaching measures to try and contain the virus, such as shutting down schools and restaurants, but also by locking down the entire country. By themselves, viruses were responsible for more deaths than all armed conflicts combined in the 20th century (Adda, 2016).

Italy has been one of the most intensely struck countries by COVID-19. Until the end of March, it had the highest number of confirmed cases per 100,000 inhabitants. It was subsequently taken over by Spain. Italy remained the second most struck country until May 1, when the United States took over. On July 3, 2020, it had the ninth highest absolute number of confirmed cases, after the United States, Brazil, Russia, India, Peru, Chile, the United Kingdom, and Spain. Despite dropping in this ranking, Italy reports the second highest global death-to-cases ratio of 14.45% (34,818 deaths to 240,961 cases), only after the United Kingdom, which reports a death-to-cases ratio of 15.50% (43,995 deaths to 283,757 cases). The third highest death-to-cases ratio of 12.24% (29,189 deaths to 238,511 cases) was reported by Mexico. 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).

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). 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. Adda (2016) 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.

TODO: 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.

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. 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.

One of the main measures resulting from the SIR model is the estimation of the basic reproduction number $R_0 := \alpha/\beta$. An epidemic is said to develop if $R_0 > 1$. 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, 2020), showing that this measure is also used communicated to citizens as a way of informing them whether the pandemic is tending to end.

TODO: Explain more later on immunity?

TODO: Explain how $R_0 > 1$ is determined.

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.

Update text accordingly given that I moved the sections around.

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 3.6.

3.1 Model 1: Within-Region Spread

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

$$I_{r,t} = \alpha_{within} I_{r,t-\tau} S_{r,t-\tau} + X_{r,t}\delta + \eta_{r,t} \tag{3.1}$$

TODO: Explain discretization and why this is fine.

Crucially, there is a difference in definition compared to the continuous-time SIR model, namely that I denotes the number of new cases rather than the fraction of the population that is infected. On the other hand, S still denotes the fraction of the population that is susceptible. The matrix X includes weekend and week of the year dummy variables. Lastly, we include an idiosyncratic error term η. The model is estimated by ordinary least squares (OLS). 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) report a median incubation period of 5.1 days (95% CI: 4.5 to 5.8 days), while Q. Li et al. (2020) report 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 $\tau = 5$.

The moment condition that needs to be satisfied due to the strict exogeneity assumption is

$$E\left[\eta_{r,t}\left(\alpha_{within} I_{r,t-\tau} S_{r,t-\tau} + X_{r,t}\delta\right)\right] = 0.$$

TODO: Add reasoning

3.2 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 (3.1). After defining the between-regions model in section 3.3, we will apply the same methodology to obtain the full weighted model in section 3.4.

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 (3.1). Taking this into account, the weighted within-region model is defined as:

$$I_{r,t} = I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha^k_{within} W^k_{r,t-lag} + X_{r,t}\delta + \eta_{r,t} \tag{3.2}$$

TODO: Possibly update later

The moment condition that needs to be satisfied due to the strict exogeneity assumption is

$$E\left[\eta_{r,t}\left(I_{r,t-lag}S_{r,t-lag}\sum_{k=1}^{K}\alpha_{within}^{k}W_{r,t-lag}^{k}+X_{r,t}\delta\right)\right]=0.$$

TODO: Add reasoning

3.3 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 modelled 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:

$$I_{r,t}=\alpha_{within}I_{r,t-lag}S_{r,t-lag}+\alpha_{between}S_{r,t-lag}\sum_{c\in R\setminus r}I_{c,t-lag}+X_{r,t}\delta+\eta_{r,t} \tag{3.3}$$

The moment condition that needs to be satisfied due to the strict exogeneity assumption is

$$E\left[\eta_{r,t}\left(\alpha_{within}I_{r,t-lag}S_{r,t-lag}+\alpha_{between}S_{r,t-lag}\sum_{c\in R\setminus r}I_{c,t-lag}+X_{r,t}\delta\right)\right]=0.$$

TODO: Add reasoning

In (3.3), the transmission parameter α is now allowed to be different within and between regions. Adda (2016) estimates (3.3) 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.

3.4 Model 4: Full Model

We now incorporate the between-region effects as well as the weighting of the transmission parameter. In addition to (3.2), 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.

$$\begin{aligned} I_{r,t} = I_{r,t-lag} S_{r,t-lag} &\sum_{k=1}^{K} \alpha_{within}^{k} W_{r,t-lag}^{k} \\ &+ S_{r,t-lag} \sum_{c \in R \setminus r} I_{c,t-lag} \sum_{k=1}^{\tilde{K}} \alpha_{between}^{k} \widetilde{W}_{r,c,t-lag}^{k} \\ &+ X_{r,t}\delta + \eta_{r,t} \end{aligned} \tag{3.4}$$

The moment condition that needs to be satisfied due to the strict exogeneity assumption is

$$E\left[\eta_{r,t}\left(I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha_{within}^{k} W_{r,t-lag}^{k} + S_{r,t-lag} \sum_{c \in R \setminus r} I_{c,t-lag} \sum_{k=1}^{\tilde{K}} \alpha_{between}^{k} \widetilde{W}_{r,c,t-lag}^{k} + X_{r,t}\delta\right)\right] = 0.$$

TODO: Add reasoning

3.5 Model selection

For model selection, we will use the Bayesian Information Criterion or BIC (Schwarz et al., 1978). Schwarz et al. (1978) developed it as an alternative to the Akaike Information

Criterion or AIC (Akaike, 1974). The AIC for a particular model is defined as

$$AIC = -2\log(ML) + 2k, \tag{3.5}$$

where ML denotes the maximum likelihood for the model and k denotes the number of parameters in the model. In contrast, the BIC is defined as

$$BIC = -2\log(ML) + k\log(n), \tag{3.6}$$

where n denotes the sample size. Both the AIC and BIC are used as the minimizer in the model selection. That is, the model that is picked by the model selection procedure is the one with the lowest AIC or BIC. When choosing between the two methods, one should realize that they have different properties, particularly related to consistency. The AIC tends to select a larger model than the BIC. Moreover, if the true model is included in the set of candidate models, and under some additional assumptions, the BIC will select the true model with probability one as n goes to infinity, whereas the AIC is not consistent. On the other hand, if the true model is not in the set of candidate models, clearly no method can possibly select the true model since it is not considered. However, the AIC is efficient in the sense that it will asymptotically select the model that minimises the mean prediction error while the BIC is not efficient (Vrieze, 2012). Proponents of using the AIC over the BIC argue that this shows that the AIC is to be preferred because it is virtually impossible for the true model to be constructed because "all models are wrong" (Box, 1976). That does not mean that reality cannot be modelled; some models can be useful despite not being perfectly true. Burnham and Anderson (2002) state that *"A model is a simplification or approximation of reality and hence will not reflect all of reality. [...] While a model can never be "truth," a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless".* Lastly, Vrieze (2012) also shows by simulation that, even if the true model is in the candidate set, the BIC can fail in finite sample sizes because it has a higher maximum risk, being defined as the mean squared error of estimating the true covariance matrix.

These papers make me believe that AIC is to be preferred as I do believe that we are not considering the true model.

3.6 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 infectives 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 regions 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 denote the testing capacity by 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 $DI_t + UI_t = I_t$. Therefore, 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. This proportion is therefore defined as a function of the testing capacity over time:

$$f_t \coloneqq f(TC_t), \tag{3.7}$$

such that

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

There are some properties that (3.7) should satisfy and some assumptions that we make. These are as follows:

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 there are no documented infections. That is,

$$f(0) = 0.$$

   Of course, individuals could be documented as being infected when they show sufficient symptoms without being tested. However, we assume that this is not the case and that only people who tested positive are documented.

   Possibly research if this is true in Italy.

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

$$f(N_t) = 1.$$

   This also 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 usually common practice to confirm a diagnosis) because of scarcity in testing resources and/or prioritization of allocating tests to the sickest patients (Frasier, 2020).

   Research the accuracy of coronatests; these are said to be quite accurate. Even if this is not the case, talk more about FP/FN and their impact.

   Q: I could not find much information that these tests are highly accurate or any sort of indication on how accurate they are, except for https://www.bmj.com/content/370/bmj.m2516. You mentioned that you found some materials on this. Could you see if you can find this?

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. Mathematically, this means that

$$f'(N_t) \geq 0.$$

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 N_t. Then, the function would be dependent on r as well:

$$f_{r,t} \coloneqq f(TC_{r,t}). \tag{3.8}$$

such that

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

We test several functional forms of the function f_t. Derivations are given in appendix C.

NB: There is a different proposal in Appendix C.2.

- **Linear form**

$$f_t = \frac{1}{N_t} TC_t. \tag{3.9}$$

- **Quadratic form**
  We specify three functional forms for a quadratic form. First of all, a general form. After this, we discuss two special cases.

  – For the general quadratic form, we assume without loss of generality that $f(\frac{1}{2}N_t) = \gamma$ for some $\gamma \in \left[\frac{1}{4}, \frac{3}{4}\right]$. Then the formula becomes:

$$f_t = \frac{2 - 4\gamma}{N_t^2} TC_t^2 + \frac{4\gamma - 1}{N_t} TC_t. \tag{3.10}$$

  If $\gamma \in \left[\frac{1}{4}, \frac{1}{2}\right)$, the function is upwards opening. If $\gamma \in \left(\frac{1}{2}, \frac{3}{4}\right]$, the function is downwards opening. If $\gamma = \frac{1}{2}$, then the formula simplifies to the linear specification. In appendix C.1.2, we explain why γ cannot be below $\frac{1}{4}$ or above $\frac{3}{4}$.

  – We assume that the vertex (i.e. the extremum) is the point $(N_t, 1)$, i.e. the parabola is downwards opening. Note that any quadratic function can be rewritten to the so-called vertex form $f(x) = a(x - h)^2 + k$, where the vertex of the function is (h, k). Moreover, notice that we need $h \geq N_t$ to adhere to the monotonicity assumption. Since we know that the point $(N_t, 1)$ is on the parabola, choosing this special case means that there will be no unknown parameters to define the function. Using that $(0, 0)$ is on the parabola, we can then derive that the formula becomes:

$$f_t = -\frac{1}{N_t^2} TC_t^2 + \frac{2}{N_t} TC_t. \tag{3.11}$$

  Note that this is equivalent to (3.10) for $\gamma = \frac{3}{4}$. Therefore, this is a boundary case for a downwards opening quadratic function.

  – For the same reason as for the previous specification, we assume that the vertex is the point $(0, 0)$, i.e. the parabola is upwards opening. Using that $(N_t, 1)$ is on the parabola, we can then derive that the formula becomes:

$$f_t = \frac{1}{N_t^2} TC_t^2. \tag{3.12}$$

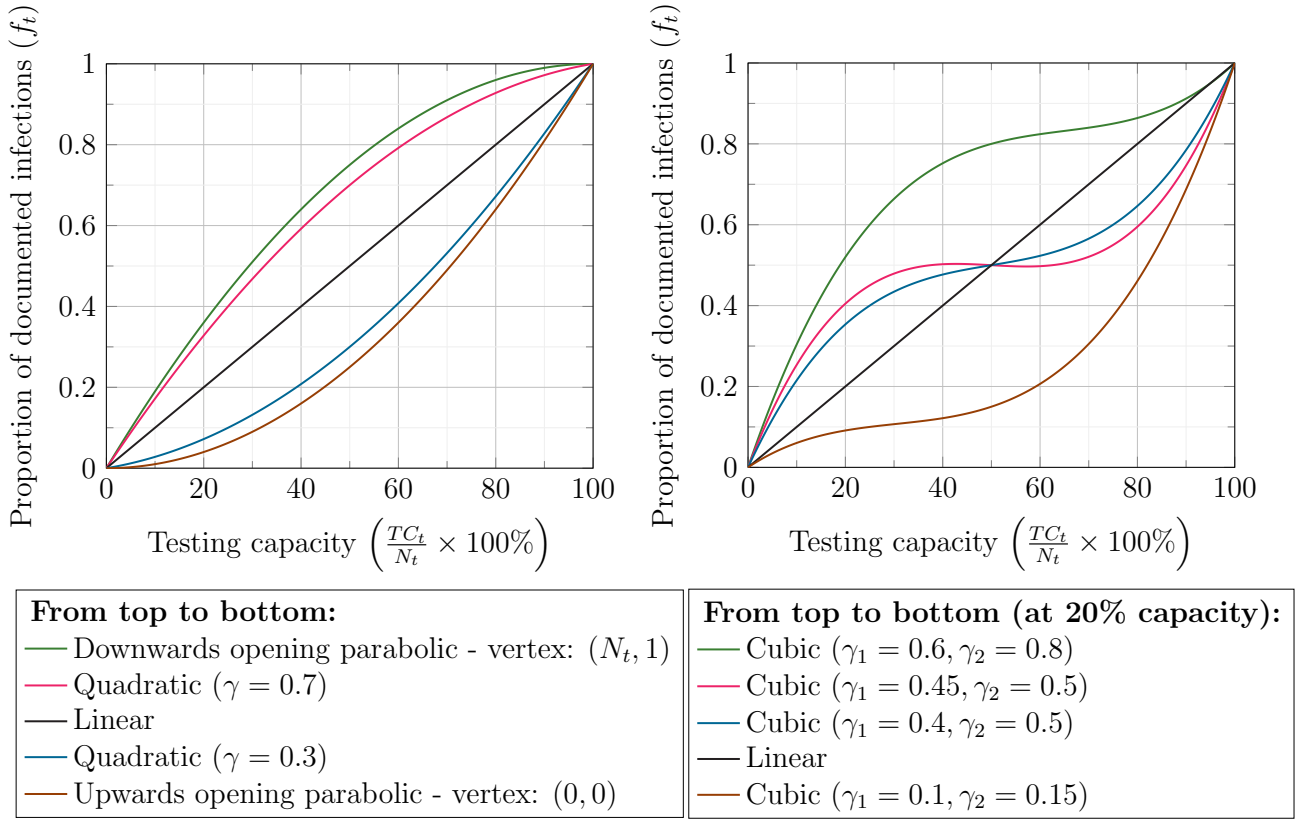
Note that this is equivalent to (3.10) for $\gamma = \frac{1}{4}$. Therefore, this is a boundary case for an upwards opening quadratic function.

- **Cubic form**
  For the cubic form, we assume without loss of generality that $f\left(\frac{1}{4}N_t\right) = \gamma_1$ and $f\left(\frac{1}{2}N_t\right) = \gamma_2$ for some $\gamma_1, \gamma_2 \in (0,1)$ such that $\gamma_1 < \gamma_2$. Then the formula becomes:

$$f_t = \frac{64\gamma_1 - 48\gamma_2 + 8}{3N_t^3}TC_t^3 + \frac{-32\gamma_1 + 20\gamma_2 - 2}{N_t^2}TC_t^2 + \frac{64\gamma_1 - 24\gamma_2 + 2}{6N_t}TC_t. \quad (3.13)$$

Suppose that we take $\gamma_2 = \frac{1}{2}$ as a special case. That is, when half of the population is tested, then half of the infectives are identified. At that point, (3.13) reduces to

$$f_t = \frac{64\gamma_1 - 16}{3N_t^3}TC_t^3 + \frac{8 - 32\gamma_1}{N_t^2}TC_t^2 + \frac{64\gamma_1 - 10}{6N_t}TC_t.$$



From top to bottom:
- Downwards opening parabolic - vertex: $(N_t, 1)$
- Quadratic $(\gamma = 0.7)$
- Linear
- Quadratic $(\gamma = 0.3)$
- Upwards opening parabolic - vertex: $(0, 0)$

From top to bottom (at 20% capacity):
- Cubic $(\gamma_1 = 0.6, \gamma_2 = 0.8)$
- Cubic $(\gamma_1 = 0.45, \gamma_2 = 0.5)$
- Cubic $(\gamma_1 = 0.4, \gamma_2 = 0.5)$
- Linear
- Cubic $(\gamma_1 = 0.1, \gamma_2 = 0.15)$

(a) Quadratic specifications **(b)** Cubic specifications

Figure 3.1. Functional forms for the proportion of documented infectives

In Figure 3.1, we specify several functional forms for the specifications as mentioned above. Figure 3.1a shows functional forms for a downwards opening as well as an upwards opening quadratic version of (3.10). It also shows plots of (3.11) and (3.12). Figure 3.1b shows four different functional forms for the cubic specification. Note that not all of the plots in Figure 3.1 are meant to be realistic portrayals. They simply show how the functions behave as the parameters change. Note that there are some combinations of γ_1 and γ_2 for which the cubic function is non-monotonically increasing. For instance, as can be seen in Figure 3.1b, the combination $\gamma_1 = 0.45$ and $\gamma_2 = 0.5$ creates a functional form for which the proportion of documented infections decreases slightly around the middle. As explained earlier in this section, this is not desirable. Henceforth, if we would use a cubic form, the values of γ_1 and γ_2 should first be tested by means of a plot, for instance.

Next, we will argue which of these forms is most appropriate. As mentioned at the beginning of this section, we cannot estimate which form would fit the data best because there is, by definition, no data on the undocumented infections. As such, we will argue which functional form to use by theoretical logic rather than a mathematical approach. Before that, there are two things to notice. Firstly, note that it is difficult to test 100% of the population without some rigorous metric, such as making it obligatory to get the test. Secondly, the shape of the functional form may differ depending on the basic reproduction number R_0, as defined in section 2. R_0 estimates how many people an infective will on average infect. If $R_0 > 1$, a person is estimated to infect more than one person and an epidemic is expected to develop. In this case, we expect that an increased testing capacity will have a larger immediate effect. We assume that a person who has been tested positive adheres to the common guidelines that they should self-quarantine. Consequently, this infective does not infect other people who would otherwise become undocumented infectives. For the remainder of this argument, we assume that $R_0 > 1$. The reason for this is that there are a variety of methods to estimate R_0 and that we cannot reasonably make our own model easily dependent on these varying results. Future research could be conducted regarding a two-step approach.

The main question that we need to ask ourselves is whether the impact of a change in testing capacity is different relative to the initial testing capacity. That is, if the testing capacity is low and we increase it, does that have a larger effect on the proportion of documented infectives than when testing capacity is high and we increase it by the same amount?

We first argue why a downwards opening quadratic function fits the requirements well. Note that when a large proportion of the population has been tested, the pool of untested people, who are potentially infectious, is smaller. The probability that they, in isolation of other effects, are infected is lower. The argument for this is as in the previous paragraph: assuming that the people close to them who were tested positive (be that

family, acquaintances, or those that they would perhaps run into at the supermarket) do indeed self-isolate, they would not have been able to been in contact with them and they have a lower chance to be infected. When a small number of people is tested and suddenly the testing capacity is increased, a larger pool of people who had symptoms and could previously not be tested, now have access to a test. The people who are now most likely to get tested positive have strong symptoms. As they are now tested positive, we assume they self-quarantine and cannot infect other people. Therefore, the functional form that fits this argument best is a downwards opening quadratic function.

One could also consider the cubic representation with $\gamma_1 = 0.6$ and $\gamma_2 = 0.8$, or some similar parameter values, as in Figure 3.1b. There, we see similar behaviour at the start of the graph where there is a sharp increase, after which it levels out. The difference is found when the last proportion of the population is tested, leading to a sudden sharp increase in the proportion of documented infections. An argument in favour of this specification is that it may be difficult to track down and convince the last proportion of the population to take a test who, at that point, may be infectious. For instance, these may simply be people who refuse to take such a test, whether those reasons are grounded or not. There may also be people who underestimate their symptoms or their importance and who, even though they are encouraged to get tested, believe that they do not need to be. For instance, they may feel that others need to get the test more. If these people are to be reached, a more rigorous program is needed and this may cause the sharp rise as a high testing capacity is reached.

Weighing these two specifications off, we believe that the former argument is more general and stable, where the second argument is quite specific and whose validity may differ across countries. In general, of all possible fitting solutions, the one with the least number of assumptions needed is to be preferred. Therefore, we opt to use a downwards opening quadratic functional form over a cubic form.

Lastly, the question is what to choose for the parameter γ, if anything. Recall that (3.10) and (3.11) are equivalent when $\gamma = \frac{3}{4}$, meaning that (3.11) is the most extreme case possible and that the slope cannot be constructed to be more steep. To be general, we choose (3.10) to be our functional form with an unknown parameter γ, denoted by $f_{r,t}(\gamma)$.

Now that we have chosen our functional form, we can adapt the models (3.1)-(3.4) to include these undocumented infections. Let us take the within-region spread model (3.1) as an example. Recall that this model was given as

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

Using that $I_{r,t} = \frac{DI_{r,t}}{f_{r,t}}$, this becomes

$$\frac{DI_{r,t}}{f_{r,t}(\gamma)} = \alpha_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} + X_{r,t}\delta + \eta_{r,t}. \tag{3.14}$$

We can rewrite (3.14) as follows

$$\begin{aligned} & DI_{r,t} = f_{r,t}(\gamma) \left(\alpha_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} + X_{r,t}\delta + \eta_{r,t} \right) \\ \iff & DI_{r,t} = \alpha_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} f_{r,t}(\gamma) + X_{r,t}\delta f_{r,t}(\gamma) + \eta_{r,t} f_{r,t}(\gamma). \end{aligned}$$

The moment conditions that need to hold are:

$$\begin{aligned} & E\left[\eta_{r,t} f_{r,t}(\gamma) \left(\alpha_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} f_{r,t}(\gamma) + X_{r,t}\delta f_{r,t}(\gamma) \right) \right] = 0 \\ \iff & E\left[\eta_{r,t} f_{r,t}^2(\gamma) \left(\alpha_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} + X_{r,t}\delta \right) \right] = 0 \\ \iff & f_{r,t}^2(\gamma) E\left[\eta_{r,t} \left(\alpha_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} + X_{r,t}\delta \right) \right] = 0. \end{aligned}$$

Since $f_{r,t}^2(\gamma)$ is simply a scaling function, regardless of the chosen parameter, it has no influence on the dependence between the error and the regressors. As such, it can be taken out of the expectation term. Subsequently, we can divide both sides of the equation by $f_{r,t}^2(\gamma)$ to obtain the original moment condition of (3.1):

$$E\left[\eta_{r,t} \left(\alpha_{within} I_{r,t-\tau} S_{r,t-\tau} + X_{r,t}\delta \right) \right] = 0.$$

Therefore, the scaling of the infectives by using our functional form, has no additional impact on the moment conditions. A similar logic applies to the moment conditions for (3.2)-(3.4) so that their moment conditions also do not depend on $f_{r,t}(\gamma)$.

TODO: Continue here

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 3.2 and 3.4.

The *Presidenza del Consiglio dei Ministri - Dipartimento della Protezione Civile* (Presidency of the Council of Ministers - Department of Civil Protection), hereafter referred to as the Italian Department of Civil Protection, has posted daily reports containing tables with a detailed numerical overview of new cases, active intensive care (IC) patients, tests executed, and more. This data is 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, where the region of *Trentino-Alto Adige* (Trento-South Tyrol) 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 7, 2020 until the strict national lockdown was instated. Unfortunately, the data was not reported at this granular level. As such, we choose to use the NUTS 2 regions.

4.1 Coronavirus data

For $R = 21$ Italian regions, we retrieved the data on the coronavirus from February 24, 2020, until June 29, 2020. Some time gaps occurred in the beginning of the data, leading to a total amount of time observations of $T = 127$. It was retrieved from the Italian Department of Civil Protection, who publish daily reports containing a number of statistics (Rosini, 2020)[1]. The statistics that are of interest to us are:

- New amount of current positive cases (*nuovi_positivi*);
- Total amount of deaths (*deceduti*);

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

[1] Official data descriptions of all variables can be found at https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-covid19-italia.md

- Total amount of recoveries (*dimessi_guariti*);
- Total amount of positive cases (*totale_casi*);
- Total amount of tests performed (*tamponi*);
- Total number of people tested (*casi_testati*).

The report also contains, for instance, the number of active ICU cases (*terapia_intensiva*) and the number of hospitalized people who showed symptoms (*ricoverati_con_sintomi*). There are two notes to make. Firstly, the data source states that the new amount of current positive cases at time t is defined as the first difference of the total amount of positive cases: $(totale\_casi_t - totale\_casi_{t-1})$. However, this is not always the case. To illustrate, we consider the region of Abruzzo on June 16 till June 18. The daily number of positive tests equal 1, 0, and -1, respectively, while the number of new confirmed cases equal 2, 2, and 1, respectively. This is likely a small measurement error. We will take the first difference of the total amount of positive cases to define the number of confirmed cases. Secondly, the difference between the total amount of tests performed (*tamponi*) and the total amount of people 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 Department of Civil Protection. 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 positive for COVID-19 before or after their death, regardless of whether they died inside or outside the hospital, assuming that these deaths were reported. 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). In the UK, there is a radical difference between the total number of deaths until June 28 with a positive test result (43,575 deaths), the total

number of deaths until June 19 where COVID-19 is mentioned on the death certificate (53,858 deaths), and the total number of deaths until June 19 over and above the usual number at that time of the year (65,132 deaths) (BBC News, 2020). This shows that the UK reports deaths due to COVID-19 on the death certificates even for people who were not tested positive. Moreover, there are many excess deaths over the usual number that may or may not be due to COVID-19 that are now not counted in the official reports.

We also make the note that it is unclear how the Department of Civil Protection 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 or do so inaccurately. For instance, different regions may adhere to different principles when deciding whether a death is classified as being due to COVID-19. 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 the number of confirmed cases or the number of deaths was reported to be negative. 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 close to June 12. Propagating this error backwards as described before would lead to zero new cases per day for Campania from May 13, 2020 until June 12. Since we have no reason to know how this error is distributed, we remove the region of Campania from our dataset. Another solution could be to distribute the error according to the daily number of cases relative to the total amount of cases until June 12.

TODO: Update accordingly if this changes

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. Because we have no way of knowing how these values are distributed over the two days, missing data is simply imputed with a value of 0. One could, on the other hand,

TODO: Update accordingly if this changes

assume a certain spread, such as fifty-fifty.

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 only available on an annual basis and is often not up-to-date. That is, sometimes data is available only up to 2016. For each variable, we kept the most recent data and assume 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
$W_{r,t}$	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

Table 4.2 continues on next page

Table 4.2 continued from previous page

Matrix	Variable	Year	Description
	deathRateChd	2016	Number of deaths from coronary heart disease per 100,000 inhabitants
	deathRateCancer	2016	Number of deaths from cancer per 100,000 inhabitants
	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

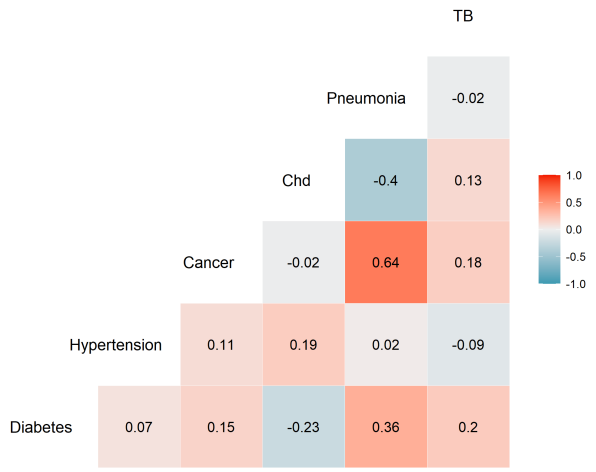


Figure 4.1. Correlation matrix of the discharge rates for various comorbidities of COVID-19

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 this section, we present the results for the within-region spread model. Recall that this was given in equation (3.1) as:

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

Firstly, we will present the results where the data is pooled to a national level. Subsequently, results are presented for the models per region to which model selection is applied with the Bayesian Information Criterion (BIC). For both result sets, we present the results from the regular model as well as modelling the undocumented infections with a quadratic form with $\gamma = 0.7$ as in (3.10).

Naively, one could consider constructing a model for the entire nation of Italy. Even though this does not take into account regional differences, as described in section X, it may be sufficient enough if regions are sufficiently similar. The results from estimating (3.1) with OLS for the national data are given in Table 5.1.

Section

TODO: Update dates in table caption

Table 5.1. Estimates from Model 1: Within-region spread on a national level. Data spans February 12 till July 4, 2020.

	Regular model				Modelling undocumented infections (Q)			
	Estimate	Std. Error	t value	p value	Estimate	Std. Error	t value	p value
Intercept	577.63	196.02	2.95	0.0038**	1.91×10^4	3.01×10^3	6.33	0.0000***
Weekend	516.54	138.82	3.72	0.0003***	4.77×10^3	1.28×10^3	3.73	0.0003***
Week number	-37.33	11.41	-3.27	0.0014**	-884.88	133.37	-6.63	0.0000***
α_{within}	0.92	0.03	26.76	0.0000***	0.77	0.04	19.38	0.0000***

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

Table 5.1 shows an estimate for α_{within} of 0.9150 that is statistically highly significant at a 1% significance level. Note that it is 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, this means that a unit change is relatively small. A similar effect will be observed for the regional models.

Of course, this model does not take into account effects specific to regions. In Table A.1 in Appendix A, we present the results from running the model on each region separately with the same model specification for each region. It is clear that the same model

might not be suitable for all regions. That is, we should apply model selection to the individual models .To execute model selection, we use the BIC and we make sure that the term for α_{within} remains in the model. The models also retain an intercept. As such, model selection is performed on the weekend dummy and the week number variable. In Table 5.2, we present the results. To compare the results from using AIC over BIC, a table to compare the two is presented in Table A.2

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

TODO: Update dates in table caption

These results were found with $f^{min} = 0.1, \beta = 0.5, \gamma = 0.7$; see section C.2.

Table 5.2. Estimates from Model 1: Within-region spread per region with model selection by stepwise BIC. Estimates are given with p-values in parentheses. Data spans February 12 till July 4, 2020. We set $f^{min} = 0.1$ (see section C.2).

	Regular model				Modelling undocumented infections (Q)			
Region	α_{within}	Intercept	Weekend	Week number	α_{within}	Intercept	Weekend	Week number
National	0.77	1,935.54	481.56	-89.84	0.77	19,065.81	4,774.00	-883.88
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABR	0.46	51.85		-2.09	0.46	507.69		-20.30
	0.00	0.00		0.00	0.00	0.00		0.00
BAS	0.50	6.65		-0.28	0.50	65.78		-2.74
	0.00	0.00		0.00	0.00	0.00		0.00
BZ	0.54	42.78		-1.83	0.54	410.43		-17.42
	0.00	0.00		0.00	0.00	0.00		0.00
CAL	0.43	22.04		-0.93	0.43	216.57		-9.03
	0.00	0.00		0.00	0.00	0.00		0.00
CAM	0.67	54.54		-2.34	0.67	541.07		-23.15
	0.00	0.00		0.00	0.00	0.00		0.00
EMR	0.74	260.82	52.14	-12.00	0.74	2,542.02	529.20	-116.74
	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
FVG	0.48	55.57		-2.31	0.48	540.88		-22.34
	0.00	0.00		0.00	0.00	0.00		0.00
LAZ	0.76	68.12		-2.89	0.76	658.13		-27.41
	0.00	0.00		0.00	0.00	0.00		0.00
LIG	0.66	101.03		-4.10	0.66	992.95		-40.06
	0.00	0.00		0.00	0.00	0.00		0.00
LOM	0.61	1,053.59	162.07	-44.60	0.61	10,384.96	1,605.69	-439.53
	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00
MAR	0.65	86.43	15.46	-3.98	0.65	850.75	157.49	-39.18
	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
MOL	0.25	7.37		-0.26	0.25	73.43		-2.62
	0.00	0.00		0.01	0.00	0.00		0.01
PIE	0.77	234.09	63.42	-10.70	0.77	2,319.76	621.18	-105.98
	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00

Table 5.2 continues on next page

Table 5.2 continued from previous page

	Regular model				Modelling undocumented infections (Q)			
Region	α_{within}	Intercept	Weekend	Week number	α_{within}	Intercept	Weekend	Week number
PUG	0.67	50.46		-2.14	0.67	495.67		-20.89
	0.00	0.00		0.00	0.00	0.00		0.00
SAR	0.49	23.92		-1.01	0.49	235.12		-9.91
	0.00	0.00		0.00	0.00	0.00		0.00
SIC	0.66	39.51		-1.72	0.66	386.97		-16.77
	0.00	0.00		0.00	0.00	0.00		0.00
TN	0.30	80.22		-2.94	0.31	785.49		-28.91
	0.00	0.00		0.00	0.00	0.00		0.00
TOS	0.73	96.84	27.93	-4.55	0.73	955.73	278.80	-44.92
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UMB	0.64	21.41		-0.95	0.64	207.67		-9.19
	0.00	0.00		0.00	0.00	0.00		0.00
VDA	0.32	24.43	6.66	-1.10	0.32	242.13	64.99	-10.91
	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00
VEN	0.67	227.05		-9.81	0.67	2,196.05		-94.25
	0.00	0.00		0.00	0.00	0.00		0.00

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

We see that the BIC gives a varying model selection per region. As mentioned, all models retain the intercept and the term $I_{t-\tau}S_{t-\tau}$ in the model. Interestingly, we also see that the week number dummy is retained in all models. In 7 out of 22 cases, the entire model is selected (including the national model).

Table 5.2 shows that the estimate for α_{within} varies vastly over the regions, from 0.25 for Molise till 0.77 for the national model and Piemonte. The estimates for the other variables vary a bit more, but this is likely due to the differences in the population count, thereby affecting the magnitude of the dependent variable $I_{r,t}$.

Update accordingly

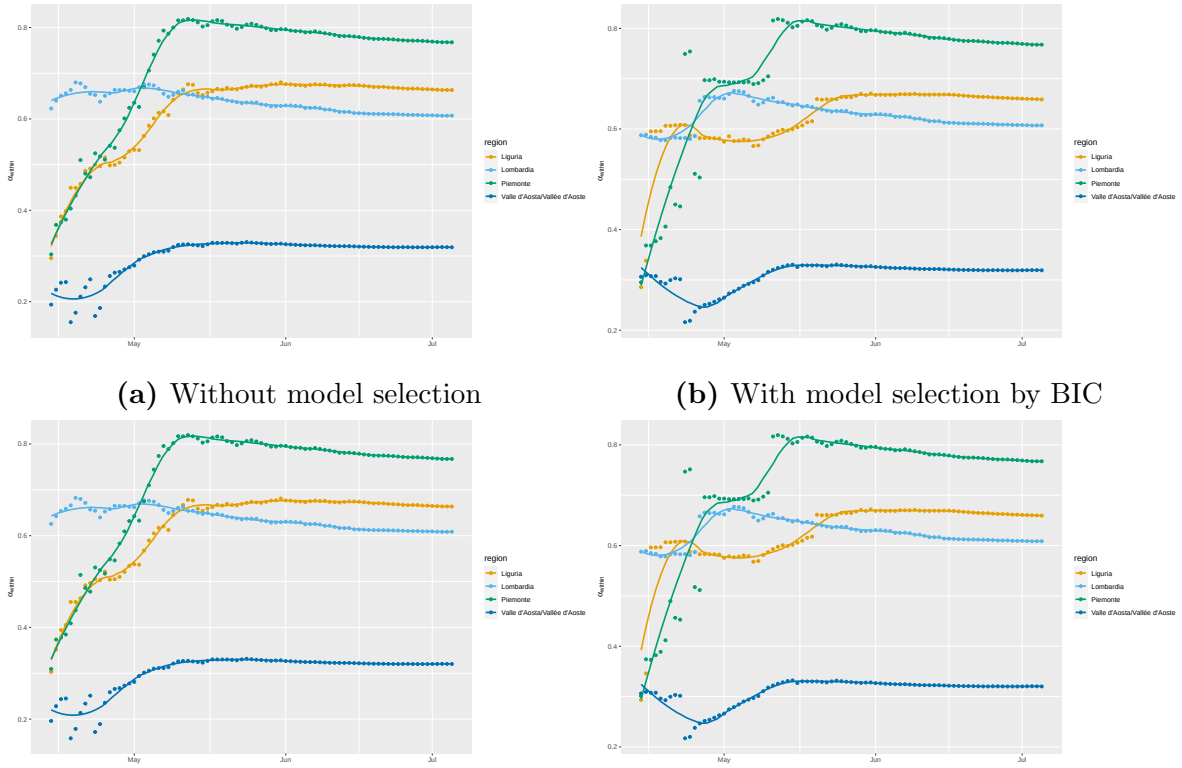
We are also interested in looking at the estimate of α_{within} over time.

TODO: Why?

That is, if we keep adding data, do we see an interesting effect in its progression? We use at least 50 data points. In Figure 5.1 we present plots for the regions in the *Nord-Ovest* (North-West) NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix B.1. Each point in the graphs in Figure 5.1 is the estimate of α_{within} when only data before that date was used.

TODO: Add national plot and text

TODO: Explain better



(a) Without model selection

(b) With model selection by BIC

(c) Without model selection; with modelling undocumented infections

(d) With model selection by BIC; with modelling undocumented infections

Figure 5.1. Progression of α_{within} over time for the *Nord-Ovest* (North-West) NUTS 1 region

In these figures, we see that the value for α_{within} levels out as more data is added. Unfortunately, this is not a positive progression.

TODO: Explain more and possibly consider using a rolling window.

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

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

BBC News. (2020). *Death rate 'back to normal' in UK*. Retrieved July 1, 2020, from https://www.bbc.com/news/health-53233066/

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Burnham, K. P., & Anderson, D. R. (2002). A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed. Springer, New York*, *2*.

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. (2020). *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

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

Rosini, U. (2020). *COVID-19*. Retrieved July 4, 2020, from https://github.com/pcm-dpc/COVID-19/tree/master/legacy/dati-regioni

Schwarz, G. Et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

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

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, *17*(2), 228.

Appendices

A Tables

A.1 Results from Model 1: Within-Region Spread

In section 5.1, we presented the results from the within-region spread model (3.1):

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

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

TODO: Update dates in table caption

TODO: Add estimates from undocumented infections

Table A.1. Estimates from Model 1: Within-region spread per region without model selection. Estimates are given with p-values in parentheses. Data spans February 12 till July 4, 2020.

Region	α_{within}	Intercept	Weekend	Week number
National	0.77	19065.81	4774.00	-883.88
	0.00	0.00	0.00	0.00
ABR	0.47	480.56	69.44	-20.08
	0.00	0.00	0.16	0.00
BAS	0.51	60.55	13.51	-2.70
	0.00	0.00	0.11	0.00
BZ	0.54	410.43		-17.42
	0.00	0.00		0.00
CAL	0.44	203.89	33.17	-8.93
	0.00	0.00	0.16	0.00
CAM	0.67	541.07		-23.15
	0.00	0.00		0.00
EMR	0.74	2542.02	529.20	-116.74
	0.00	0.00	0.01	0.00
FVG	0.48	540.88		-22.34
	0.00	0.00		0.00
LAZ	0.76	620.25	104.00	-27.20
	0.00	0.00	0.06	0.00
LIG	0.66	947.02	139.01	-39.95
	0.00	0.00	0.09	0.00
LOM	0.61	10384.96	1605.69	-439.53
	0.00	0.00	0.01	0.00
MAR	0.65	850.75	157.49	-39.18
	0.00	0.00	0.01	0.00
MOL	0.28	65.24	22.00	-2.57

Table A.1 continues on next page

Table A.1 continued from previous page

Region	α_{within}	Intercept	Weekend	Week number
	0.00	0.00	0.05	0.01
PIE	0.77	2319.76	621.18	-105.98
	0.00	0.00	0.01	0.00
PUG	0.67	495.67		-20.89
	0.00	0.00		0.00
SAR	0.49	235.12		-9.91
	0.00	0.00		0.00
SIC	0.66	386.97		-16.77
	0.00	0.00		0.00
TN	0.33	729.82	136.87	-28.46
	0.00	0.00	0.15	0.00
TOS	0.73	955.73	278.80	-44.92
	0.00	0.00	0.00	0.00
UMB	0.64	207.67		-9.19
	0.00	0.00		0.00
VDA	0.32	242.13	64.99	-10.91
	0.00	0.00	0.01	0.00
VEN	0.68	2060.40	281.16	-92.28
	0.00	0.00	0.11	0.00

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

Table A.2. Estimates from Model 1: Within-region spread per region with model selection by stepwise AIC versus BIC. Estimates are given with p-values in parentheses. Data spans February 12 till July 4, 2020. We set $f^{min} = 0.1$ (see section C.2).

	Model selection with AIC				Model selection with BIC			
Region	α_{within}	Intercept	Weekend	Week number	α_{within}	Intercept	Weekend	Week number
National	0.77	19065.81	4774.00	-883.88	0.77	19065.81	4774.00	-883.88
National_pvals	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABR	0.47	480.56	69.44	-20.08	0.46	507.69		-20.30
ABR_pvals	0.00	0.00	0.16	0.00	0.00	0.00		0.00
BAS	0.51	60.55	13.51	-2.70	0.50	65.78		-2.74
BAS_pvals	0.00	0.00	0.11	0.00	0.00	0.00		0.00
BZ	0.54	410.43		-17.42	0.54	410.43		-17.42
BZ_pvals	0.00	0.00		0.00	0.00	0.00		0.00
CAL	0.44	203.89	33.17	-8.93	0.43	216.57		-9.03
CAL_pvals	0.00	0.00	0.16	0.00	0.00	0.00		0.00
CAM	0.67	541.07		-23.15	0.67	541.07		-23.15
CAM_pvals	0.00	0.00		0.00	0.00	0.00		0.00
EMR	0.74	2542.02	529.20	-116.74	0.74	2542.02	529.20	-116.74
EMR_pvals	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
FVG	0.48	540.88		-22.34	0.48	540.88		-22.34
FVG_pvals	0.00	0.00		0.00	0.00	0.00		0.00
LAZ	0.76	620.25	104.00	-27.20	0.76	658.13		-27.41

Table A.2 continues on next page

Table A.2 continued from previous page

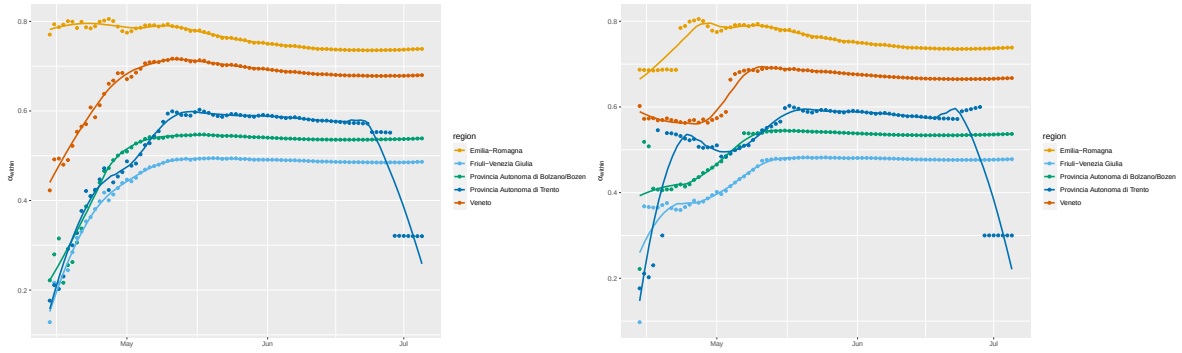
	Model selection with AIC				Model selection with BIC			
Region	α_{within}	Intercept	Weekend	Week number	α_{within}	Intercept	Weekend	Week number
LAZ_pvals	0.00	0.00	0.06	0.00	0.00	0.00		0.00
LIG	0.66	947.02	139.01	-39.95	0.66	992.95		-40.06
LIG_pvals	0.00	0.00	0.09	0.00	0.00	0.00		0.00
LOM	0.61	10384.96	1605.69	-439.53	0.61	10384.96	1605.69	-439.53
LOM_pvals	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
MAR	0.65	850.75	157.49	-39.18	0.65	850.75	157.49	-39.18
MAR_pvals	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
MOL	0.28	65.24	22.00	-2.57	0.25	73.43		-2.62
MOL_pvals	0.00	0.00	0.05	0.01	0.00	0.00		0.01
PIE	0.77	2319.76	621.18	-105.98	0.77	2319.76	621.18	-105.98
PIE_pvals	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
PUG	0.67	495.67		-20.89	0.67	495.67		-20.89
PUG_pvals	0.00	0.00		0.00	0.00	0.00		0.00
SAR	0.49	235.12		-9.91	0.49	235.12		-9.91
SAR_pvals	0.00	0.00		0.00	0.00	0.00		0.00
SIC	0.66	386.97		-16.77	0.66	386.97		-16.77
SIC_pvals	0.00	0.00		0.00	0.00	0.00		0.00 TN
0.33	729.82	136.87	-28.46	0.31	785.49		-28.91	
TN_pvals	0.00	0.00	0.15	0.00	0.00	0.00		0.00
TOS	0.73	955.73	278.80	-44.92	0.73	955.73	278.80	-44.92
TOS_pvals	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UMB	0.64	207.67		-9.19	0.64	207.67		-9.19
UMB_pvals	0.00	0.00		0.00	0.00	0.00		0.00
VDA	0.32	242.13	64.99	-10.91	0.32	242.13	64.99	-10.91
VDA_pvals	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
VEN	0.68	2060.40	281.16	-92.28	0.67	2196.05		-94.25
VEN_pvals	0.00	0.00	0.11	0.00	0.00	0.00		0.00

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

B Figures

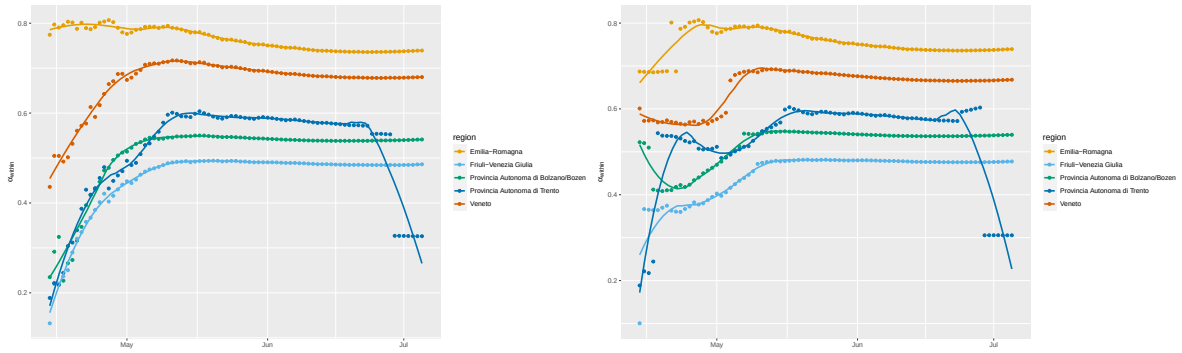
B.1 Plots of α_{within} over time

In section 5.1 we presented the plots of α_{within} over time for the *Nord-Ovest* (North-West) NUTS 1 region. In this section, we will present the plots for the other four NUTS 1 regions.



(a) Without model selection

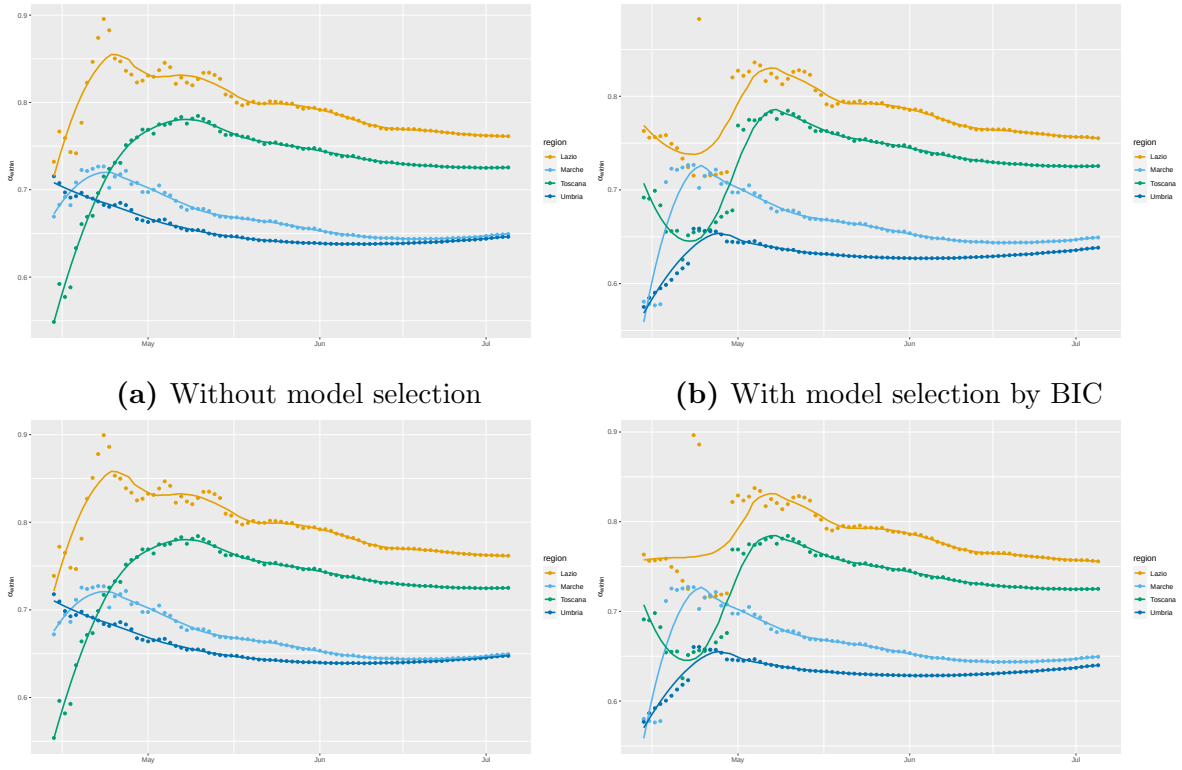
(b) With model selection by BIC



(c) Without model selection; with modelling undocumented infections

(d) With model selection by BIC; with modelling undocumented infections

Figure B.1. Progression of α_{within} over time for the *Nord-Est* (North-East) NUTS 1 region



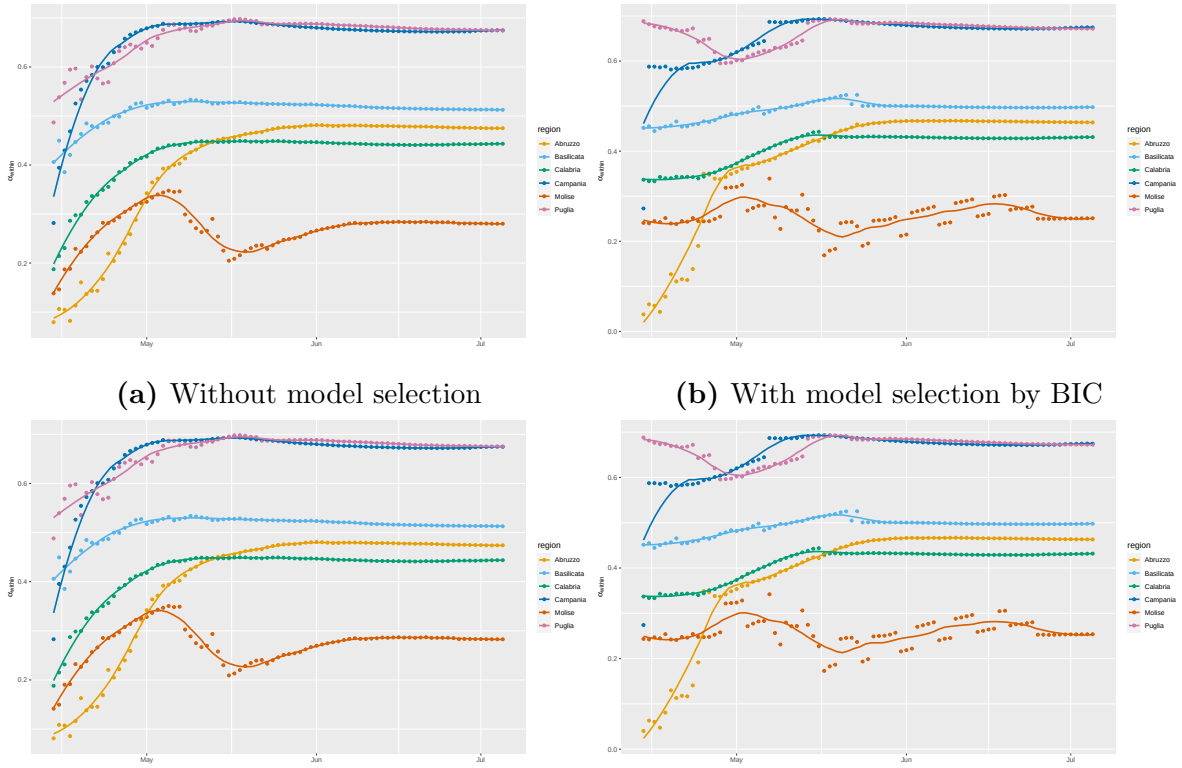
(a) Without model selection

(b) With model selection by BIC

(c) Without model selection; with modelling undocumented infections

(d) With model selection by BIC; with modelling undocumented infections

Figure B.2. Progression of α_{within} over time for the *Centro (IT)* (Centre) NUTS 1 region



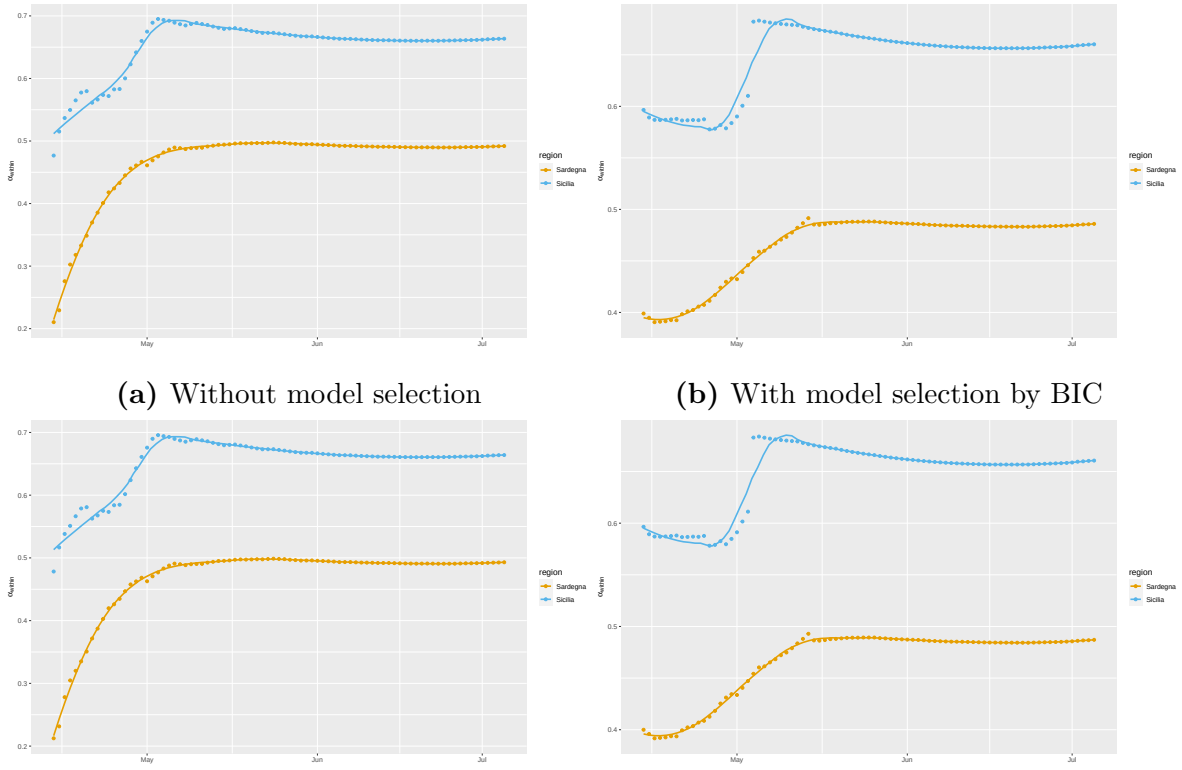
(a) Without model selection

(b) With model selection by BIC

(c) Without model selection; with modelling undocumented infections

(d) With model selection by BIC; with modelling undocumented infections

Figure B.3. Progression of α_{within} over time for the *Sud* (South) NUTS 1 region



(a) Without model selection

(b) With model selection by BIC

(c) Without model selection; with modelling undocumented infections

(d) With model selection by BIC; with modelling undocumented infections

Figure B.4. Progression of α_{within} over time for the *Isole* (Islands) NUTS 1 region

C Derivations

C.1 Section 3.6: Modelling undocumented infections

C.1.1 Linear function

For modelling the undocumented infections, we want to construct a formula for a linear function that obeys the following assumptions:

(I) $f(TC_t) = aTC_t + b$ for some $a, b \in \mathbb{R}$,

(II) $f(0) = 0$,

(III) $f(N_t) = 1$

From assumption (II), we obtain that $b = 0$. From assumption (III), we can then derive the value of a. The equation that we need to solve is:

$$aN_t = 1.$$

This is readily solved as $a = \frac{1}{N_t}$. As such, we have derived that

$$f(TC_t) = \frac{1}{N_t} TC_t.$$

C.1.2 General quadratic function

For modelling the undocumented infections, we want to construct a general formula for a quadratic function that obeys the following assumptions:

(I) $f(TC_t) = aTC_t^2 + bTC_t + c$ for some $a, b, c \in \mathbb{R}$,

(II) $f(0) = 0$,

(III) $f(N_t) = 1$,

(IV) $f(\beta N_t) = \gamma$ for some $\beta, \gamma \in (0, 1)$,

(V) The vertex of the parabola should be to the right of N_t in the case of a downwards opening parabola ($\gamma > \frac{1}{2}$) and to the left of the origin in the case of an upwards opening parabola ($\gamma < \frac{1}{2}$). That is:

$$\begin{cases} TC_t|_{f'(TC_t)=0} \geq N_t & \text{if } \gamma > \frac{1}{2} \\ TC_t|_{f'(TC_t)=0} \leq 0 & \text{if } \gamma < \frac{1}{2}. \end{cases}$$

From assumption (II), we obtain that $c = 0$. From assumptions (III) and (IV), we can then derive the values of a and b in terms of β, γ and N_t. The set of equations that we need to solve are:

$$\begin{cases} aN_t^2 + bN_t & = 1 \text{ (from assumption (III))} \\ a\beta^2 N_t^2 + b\beta N_t & = \gamma \text{ (from assumption (IV))} \end{cases} \tag{C.1}$$

To solve (C.1), we can apply row reduction as follows:

$$\begin{aligned} \left(\begin{array}{cc|c} N_t^2 & N_t & 1 \\ \beta^2 N_t^2 & \beta N_t & \gamma \end{array}\right) &\xrightarrow{r_2 - \beta^2 r_1} \left(\begin{array}{cc|c} N_t^2 & N_t & 1 \\ 0 & \beta(1-\beta)N_t & \gamma - \beta^2 \end{array}\right) \\ &\xrightarrow{r_1 - \frac{1}{\beta(1-\beta)} r_2} \left(\begin{array}{cc|c} N_t^2 & 0 & 1 - \frac{\gamma - \beta^2}{\beta(1-\beta)} \\ 0 & \beta(1-\beta)N_t & \gamma - \beta^2 \end{array}\right) \\ &= \left(\begin{array}{cc|c} N_t^2 & 0 & \frac{\beta - \gamma}{\beta(1-\beta)} \\ 0 & \beta(1-\beta)N_t & \gamma - \beta^2 \end{array}\right) \\ &\xrightarrow[r_2 \times \frac{1}{\beta(1-\beta)N_t}]{r_1 \times \frac{1}{N_t^2}} \left(\begin{array}{cc|c} 1 & 0 & \frac{\beta - \gamma}{\beta(1-\beta)N_t^2} \\ 0 & 1 & \frac{\gamma - \beta^2}{\beta(1-\beta)N_t} \end{array}\right) \end{aligned}$$

As such, we have derived that

$$\begin{cases} a &= \frac{\beta-\gamma}{\beta(1-\beta)N_t^2} \\ b &= \frac{\gamma-\beta^2}{\beta(1-\beta)N_t}. \end{cases} \tag{C.2}$$

Now note that our function is continuous. As such, without loss of generality, we choose $\beta = \frac{1}{2}$ and do the following derivations to deduce the values of γ for which assumption (V) holds. That is, we want to find the values of γ for which

$$f'(TC_t) = 0 \iff \begin{cases} TC_t &\geq N_t \text{ for } \gamma > \frac{1}{2} \\ TC_t &\leq 0 \text{ for } \gamma < \frac{1}{2}. \end{cases}$$

Firstly, assuming $\beta = \frac{1}{2}$, the expressions for a and b as in (C.2) reduce to:

$$\begin{cases} a &= \frac{2-4\gamma}{N_t^2} \\ b &= \frac{4\gamma-1}{N_t}. \end{cases} \tag{C.3}$$

We now need to derive the values of γ such that assumption (V) holds. That is:

$$\begin{aligned} & f'(TC_t) = 0 \\ \iff & \frac{\partial aTC_t^2 + bTC_t}{\partial TC_t} = 0 \\ \iff & 2aTC_t + b = 0 \\ \iff & TC_t = -\frac{b}{2a}. \end{aligned}$$

Using (C.3), we can fill out a and b to obtain:

$$TC_t = \frac{1-4\gamma}{4-8\gamma}N_t.$$

Let $\gamma > \frac{1}{2}$. Then:

$$\begin{aligned} & \frac{1-4\gamma}{4-8\gamma}N_t \geq N_t \\ \iff & 1-4\gamma \leq 4-8\gamma \\ \iff & \gamma \leq \frac{3}{4}. \end{aligned}$$

Let $\gamma < \frac{1}{2}$. Then:

$$\begin{aligned} & \frac{1-4\gamma}{4-8\gamma}N_t \leq 0 \\ \iff & 1-4\gamma \leq 0 \\ \iff & \gamma \geq \frac{1}{4}. \end{aligned}$$

As such, we should have that $\gamma \in \left[\frac{1}{4}, \frac{3}{4}\right]$. When $\gamma \in \left[\frac{1}{4}, \frac{1}{2}\right)$, the parabola we receive is upwards opening. On the other hand, when $\gamma \in \left(\frac{1}{2}, \frac{3}{4}\right]$, the parabola we receive is downwards opening. Clearly, when $\gamma = \frac{1}{2}$, the function we receive is linear, since $a = \frac{2-4\gamma}{N_t^2} = 0$.

Conclusively, we have derived that

$$f(TC_t) = \frac{2-4\gamma}{N_t^2} TC_t^2 + \frac{4\gamma - 1}{N_t} TC_t,$$

under the assumption that $\beta = \frac{1}{2}$.

C.1.3 Special case quadratic formula: downwards opening

For modelling the undocumented infections, we want to construct a formula for a downwards opening quadratic function that obeys the following assumptions:

(I) $f(x) = ax^2 + bx + c$ for some $a, b, c \in \mathbb{R}$,

(II) $f(0) = 0$,

(III) $f(N_t) = 1$,

(IV) $f'(N_t) = 0$, i.e. the vertex of the parabola is found at $TC_t = N_t$.

From assumption (II), we obtain that $c = 0$. Now, consider that any quadratic formula can be written as $f(TC_t) = a(TC_t - h)^2 + k$, which is called the vertex form, where the vertex (i.e. the extremum) of the function is (h, k). By assumptions (III) and (IV), $h = N_t$ and $k = 1$. Therefore,

$$f(TC_t) = a(TC_t - N_t)^2 + 1.$$

Using assumption (II), we can solve this equation for a:

$$\begin{aligned} & a(0 - N_t)^2 + 1 = 0 \\ \iff & aN_t^2 = -1 \\ \iff & a = -\frac{1}{N_t^2} \end{aligned}$$

Therefore, the formula becomes:

$$\begin{aligned} f(TC_t) &= -\frac{1}{N_t^2}(TC_t - N_t)^2 + 1 \\ &= -\frac{1}{N_t^2}(TC_t^2 + N_t^2 - 2N_t TC_t) + 1 \\ &= -\frac{1}{N_t^2} TC_t^2 - \frac{N_t^2}{N_t^2} + \frac{2}{N_t} TC_t + 1 \\ &= -\frac{1}{N_t^2} TC_t^2 + \frac{2}{N_t} TC_t. \end{aligned}$$

C.1.4 Special case quadratic formula: upwards opening

For modelling the undocumented infections, we want to construct a formula for an upwards opening quadratic function that obeys the following assumptions:

(I) $f(x) = ax^2 + bx + c$ for some $a, b, c \in \mathbb{R}$,

(II) $f(0) = 0$,

(III) $f(N_t) = 1$,

(IV) $f'(0) = 0$, i.e. the vertex of the parabola is found at $TC_t = 0$.

From assumption (II), we obtain that $c = 0$. Just as in appendix C.1.4, we use the vertex form $f(TC_t) = a(TC_t - h)^2 + k$. By assumptions (III) and (IV), $h = 0$ and $k = 0$. Therefore,

$$f(TC_t) = a(TC_t - 0)^2 + 0 = aTC_t^2.$$

Using assumption (II), we can solve this equation for a:

$$\begin{aligned} aN_t^2 &= 1 \\ \iff a &= \frac{1}{N_t^2} \end{aligned}$$

Therefore, the formula becomes:

$$f(TC_t) = \frac{1}{N_t^2} TC_t^2,$$

which is already in the form as in assumption (I).

C.1.5 Cubic function

For modelling the undocumented infections, we want to construct a general formula for a cubic function that obeys the following assumptions:

(I) $f(x) = ax^3 + bx^2 + cx + d$ for some $a, b, c, d \in \mathbb{R}$,

(II) $f(0) = 0$,

(III) $f(N_t) = 1$,

(IV) $f(\beta_1 N_t) = \gamma_1$ and $f(\beta_2 N_t) = \gamma_2$ for some $\beta_1, \beta_2, \gamma_1, \gamma_2 \in (0, 1)$ and $\beta_1 < \beta_2, \gamma_1 < \gamma_2$.

From assumption (II), we obtain that $d = 0$. From assumptions (III) and (IV), we can then derive the values of a, b, and c in terms of the βs, γs, and N_t. The set of equations that we need to solve are:

$$\begin{cases} aN_t^3 + bN_t^2 + cN_t & = 1 \text{ (from assumption (III))} \\ a\beta_1^3 N_t^3 + b\beta_1^2 N_t^2 + c\beta_1 N_t & = \gamma_1 \text{ (from assumption (IV))} \\ a\beta_2^3 N_t^3 + b\beta_2^2 N_t^2 + c\beta_2 N_t & = \gamma_2 \text{ (from assumption (IV))} \end{cases} \tag{C.4}$$

In appendix C.1.2, we first solved these equations and then assumed a value for β afterwards, without loss of generality. In this case, the equations would become immensely populated if we were to stay general. As such, we now first assume without loss of generality that $\beta_1 = \frac{1}{4}$ and $\beta_2 = \frac{1}{2}$. To solve (C.4), we can then apply row reduction as follows:

$$\left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 \\ \beta_1^3 N_t^3 & \beta_1^2 N_t^2 & \beta_1 N_t & \gamma_1 \\ \beta_2^3 N_t^3 & \beta_2^2 N_t^2 & \beta_2 N_t & \gamma_2 \end{array}\right) = \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 \\ \frac{1}{64} N_t^3 & \frac{1}{16} N_t^2 & \frac{1}{4} N_t & \gamma_1 \\ \frac{1}{8} N_t^3 & \frac{1}{4} N_t^2 & \frac{1}{2} N_t & \gamma_2 \end{array}\right)$$

$$\xrightarrow[r_3 \times 8]{r_2 \times 64} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 \\ N_t^3 & 4N_t^2 & 16N_t & 64\gamma_1 \\ N_t^3 & 2N_t^2 & 4N_t & 16\gamma_2 \end{array}\right)$$

$$\xrightarrow[r_3 - r_1]{r_2 - r_1} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 \\ 0 & 3N_t^2 & 15N_t & 64\gamma_1 - 1 \\ 0 & N_t^2 & 3N_t & 8\gamma_2 - 1 \end{array}\right)$$

$$\xrightarrow{r_2 \leftrightarrow r_3} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 \\ 0 & N_t^2 & 3N_t & 8\gamma_2 - 1 \\ 0 & 3N_t^2 & 15N_t & 64\gamma_1 - 1 \end{array}\right)$$

$$\xrightarrow[r_3 - 3r_2]{r_1 - r_2} \left(\begin{array}{ccc|c} N_t^3 & 0 & -2N_t & 2 - 8\gamma_2 \\ 0 & N_t^2 & 3N_t & 8\gamma_2 - 1 \\ 0 & 0 & 6N_t & 64\gamma_1 - 24\gamma_2 + 2 \end{array}\right)$$

$$\xrightarrow[r_2 - \frac{1}{2} r_3]{r_1 + \frac{1}{3} r_3} \left(\begin{array}{ccc|c} N_t^3 & 0 & 0 & \frac{64\gamma_1 - 48\gamma_2 + 8}{3} \\ 0 & N_t^2 & 0 & -32\gamma_1 + 20\gamma_2 - 2 \\ 0 & 0 & 6N_t & 64\gamma_1 - 24\gamma_2 + 2 \end{array}\right)$$

$$\xrightarrow[r_2 \div N_t^2]{r_1 \div N_t^3} \xrightarrow{r_3 \div 6N_t} \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{64\gamma_1 - 48\gamma_2 + 8}{3N_t^3} \\ 0 & 1 & 0 & \frac{-32\gamma_1 + 20\gamma_2 - 2}{N_t^2} \\ 0 & 0 & 1 & \frac{64\gamma_1 - 24\gamma_2 + 2}{6N_t} \end{array}\right)$$

Conclusively, we have derived that

$$\begin{cases} a &= \frac{64\gamma_1 - 48\gamma_2 + 8}{3N_t^3} \\ b &= \frac{-32\gamma_1 + 20\gamma_2 - 2}{N_t^2} \\ c &= \frac{64\gamma_1 - 24\gamma_2 + 2}{6N_t}. \end{cases} \tag{C.5}$$

so that

$$f(TC_t) = \frac{64\gamma_1 - 48\gamma_2 + 8}{3N_t^3} TC_t^3 + \frac{-32\gamma_1 + 20\gamma_2 - 2}{N_t^2} TC_t^2 + \frac{64\gamma_1 - 24\gamma_2 + 2}{6N_t} TC_t,$$

under the assumption that $\beta_1 = \frac{1}{4}$ and $\beta_2 = \frac{1}{2}$.

C.2 New formulation of undocumented individuals

Q: Could you please read this section and tell me what you think?

Notice that f_t in the original definition starts out extremely low. As such, the number of undocumented infections is estimated to be extremely high and, subsequently, the total number of infections will quickly exceed the total population which is, clearly, impossible. We can execute a few fixes for this, for instance:

1. Change the assumption $f(0) = 0$ to be $f(0) = f_{min}$, for instance $f(0) = 0.1$. Relating this to the next point, this would imply $\delta_t^{max} = \frac{1-f(0)}{f(0)} = \frac{0.9}{0.1} = 9$ (under the monotonicity assumption). Using that $\delta_t < \frac{N_t}{DI_t} - 1$ we need:

$$\begin{aligned} \frac{1}{f_t} - 1 &< \frac{N_t - DI_t}{DI_t} \\ \frac{1}{f_t} &< \frac{N_t}{DI_t} \\ f_t &> \frac{DI_t}{N_t}. \end{aligned}$$

   so that $f_t^* = \max\left\{\frac{DI_t}{N_t}, f_t\right\}$. The issue is that $f_t = \frac{DI_t}{N_t}$ would imply an entirely infected population, which is not realistic. As such, I think setting a lower bound that is quite a bit higher than $\frac{DI_t}{N_t}$ is better.

2. Define the model differently, namely $UI_t = \delta_t DI_t$ (instead of $UI_t = \frac{1-f_t}{f_t} DI_t$) for $\delta_t \in [0, \delta_t^{max}]$ for some $\delta_t^{max} \in \mathbb{R}_+$. That is, the undocumented infections are a certain factor of the documented infections. This may be preferable because it may be more intuitive to set the maximum multiplication factor than the minimum percentage.

   Note that we need

$$\begin{aligned} DI_t + UI_t &< N_t \\ (1+\delta_t) DI_t &< N_t \\ \delta_t &< \frac{N_t}{DI_t} - 1 \end{aligned}$$

   Given a certain functional form for δ, we could say that the final factor δ_t^* is given by:

$$\delta_t^* = \min\{\delta_t, \delta_t^{max}\}.$$

   The issue is that $\delta_t = \frac{N_t}{DI_t} - 1$ would imply an entirely infected population, which is not realistic. As such, I think setting an upper bound that is quite a bit lower than $\frac{N_t}{DI_t} - 1$ is better.

We take the first option as an example. We can derive that a quadratic form is given by:

$$f(TC_t) = \frac{\beta - \gamma + (1-\beta) f^{min}}{\beta(1-\beta)N_t^2} TC_t^2 + \frac{\gamma - \beta^2 - (1-\beta^2) f^{min}}{\beta(1-\beta)N_t} TC_t + f^{min}.$$

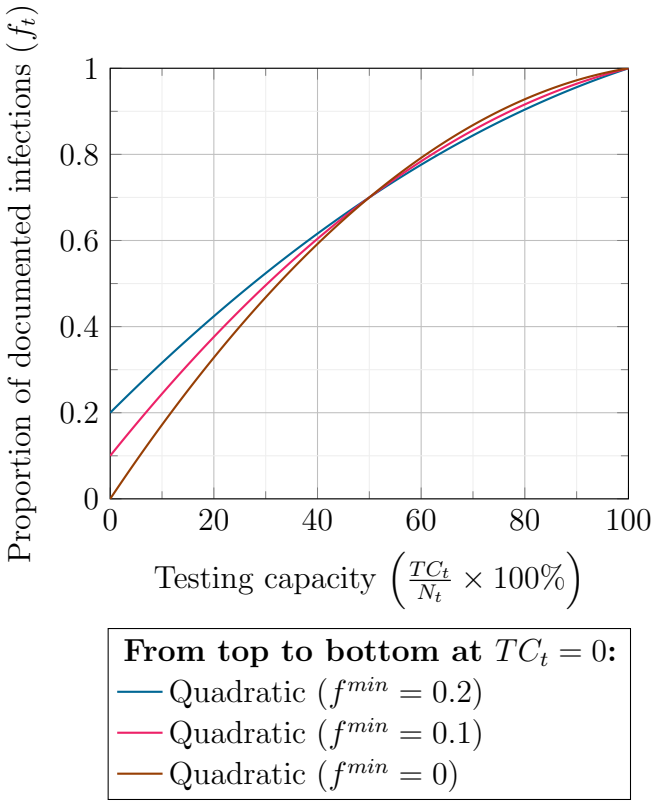


Figure C.1. Functional forms for the proportion of documented infectives ($\beta = 0.5, \gamma = 0.7$)