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

by
Mike Weltevrede (ANR 756479)

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:
August 18, 2020

Abstract

TODO

Acknowledgements

TODO

Contents

1 Introduction

2 Problem description

In this section, we elaborate on the problem at hand, namely the epidemiological spread of SARS-CoV-2 and the disease it causes. By themselves, viruses were responsible for more deaths than all armed conflicts combined in the twentieth century (Adda, 2016). 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 apart from Antarctica 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. On August 2, 2020, 17.8 million people were reported to have been infected with COVID-19, leading to 675 thousand consequent deaths. Only 12 sovereign member states of the United Nations reported no infections. For two of these countries, namely North Korea and Turkmenistan, these reports are suspected to be false.

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 reported 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). Due to the extreme nature of the pandemic in Italy and the availability of enough data, this thesis chooses to focus on Italy. Specifically, we focus on modelling on the level of regions rather than on a nation-wide approach.

We are basing our models on specifications as used by Adda (2016). In the paper, Adda (2016) investigates the spread of several viral diseases in the past, namely for influenza, gastroenteritis, and chickenpox. The key additions made are, firstly, that a spatial spillover effect is considered and, secondly, that some sort of weighting on the parameters is allowed on the basis of region specific variables. With this motivation, Adda (2016) defines three models comprising of a model ignoring interaction between regions, a model taking interaction between regions into account, and a model that expands on the latter by introducing the weights. In addition, we add an intermediate model that ignores interregional dependence but where we do include weights. These models have

not previously been applied to SARS-CoV-2 and can possibly show interesting insights compared to other models.

3 Methodology

In this section, we explain the methodology applied in this thesis. We discuss our models and the thought process behind them. In Section 3.1, we describe the most commonly used model in epidemiology: the SIR model. In Section 3.2, we present a model ignoring effects across regions and for which the transmission rate parameter is determined by the previous infectives. Subsequently, in Section 3.3, we extend this model by allowing the transmission rate parameter to be weighted by several other factors. After this, Section 3.4 presents a model that takes effects across regions into account for which the transmission rate parameter is determined by the previous infectives. Having discussed the models by Adda (2016), we develop our own discrete SIR model in Section 3.5, which is estimated by panel data methods and Bayesian estimation. In Section 3.6, we consider how to do model selection for these four models to determine the best set of regressors to use. Lastly, Section 3.7 describes how undocumented infectives are modelled.

TODO: Section on Discrete SIR

3.1 SIR model

Adda (2016) starts from the Standard Inflammatory Response (SIR) model, the most commonly used model in epidemiology (Anderson & May, 1992; Kermack & McKendrick, 1927). We will follow the notation by Keeling and Rohani (2011). That is, 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. Keeling and Rohani (2011) furthermore define X to be the number of susceptible individuals, Y to be the number of infectives, and Z to be the number of recovered individuals, so that $S = X/N$, $I = Y/N$, and $R = Z/N$, where N is the total population size. As such, at any point in time, we have that

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

$$X, Y, Z \in [0, N] \text{ and } X + Y + Z = N.$$

The SIR model is postulated in continuous time, i.e. the equations in (3.1), (3.5), and (3.3) 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 fraction of infectives) to which a flow is added or subtracted.

$$\frac{dS}{dt} = -\beta SI + wR \tag{3.1}$$
$$\frac{dI}{dt} = \beta SI - \gamma I \tag{3.2}$$
$$\frac{dR}{dt} = \beta I - wR \tag{3.3}$$

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 βSI in equations (3.1) and (3.5), 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 γI in equations (3.5) and (3.3).

Finally, we assume that there is a constant rate of change at which immune individuals lose their immunity. This is denoted by the term wR in equations (3.1) and (3.3). For instance, Adda (2016) mentions that w is set to 0 for chickenpox as individuals acquire a lifetime immunity while w 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). This can be challenged because it is currently still unknown whether immunity is always achieved, especially among those who have had only light to medium symptoms. However, it is estimated that COVID-19 antibodies will remain in a patient's system for two to three years, based on what is known about other coronaviruses but it is too early to know for certain (Leung, 2020). Nonetheless, no definitive results have been shown. For simplicity's sake, we assume that lifelong immunity is achieved, or at least long enough to last through the time-scope of our analysis: we set $w = 0$.

One of the main measures resulting from the SIR model is the estimation of the basic reproduction number $R_0 \coloneqq \beta/\gamma$. An epidemic is said to develop if $R_0 > 1$. This is clear because this is the case when $\beta > \gamma$, i.e. the spread of the virus exceeds the recovery rate: individuals become infected more quickly than they recover. 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.

3.2 Within-Region Spread Model

Recall that the SIR model is postulated in continuous time. Adda (2016) discretizes the SIR model. First of all, we will discuss how the discretization appears to be carried out, due to a lack of explanation by Adda (2016). Recall from (3.5) that $\frac{dI}{dt} = \beta SI - \gamma I$. As such, the discretized version (for a region r) is:

$$I_{r,t} - I_{r,t-1} = \beta S_{r,t-1} I_{r,t-1} - \gamma I_{r,t-1} \tag{3.4}$$

There are a few things to note. Firstly, if we want to estimate this equation's parameters, an error occurs. This is added to the model by an error term denoted by $\eta_{r,t}$:

$$I_{r,t} - I_{r,t-1} = \beta S_{r,t-1} I_{r,t-1} - \gamma I_{r,t-1} + \eta_{r,t} \tag{3.5}$$

Secondly, individuals that get infected do not immediately infect others because there is a so-called latent period, being the period between an infection and the moment that the infective is infectious. For COVID-19, the latent period is estimated to be approximately 2 days shorter than the incubation period (He et al., 2020). The incubation period is 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. 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), or 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. Because the latent period is estimated to be shorter than the incubation period, 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 discuss how we model pre-symptomatic individuals in Section 3.7. Adda (2016) models this transmission lag by making the lag on the right hand side of (3.5) dependent on the incubation period. This is denoted by the parameter τ:

$$I_{r,t} - I_{r,t-1} = \beta S_{r,t-\tau} I_{r,t-\tau} - \gamma I_{r,t-\tau} + \eta_{r,t} \tag{3.6}$$

For instance, Adda (2016) chooses τ equal to one week for acute diarrhea and flu-like illnesses as these have an incubation period of less than a week. Due to the results from Lauer et al. (2020), Q. Li et al. (2020), and Linton et al. (2020), indicating an incubation period of roughly five days, and the result from He et al. (2020) that the latent period is roughly two days shorter than the incubation period, we choose $\tau = 3$.

Thirdly, Adda (2016) adds regressors to the model as control variables, such as the region fixed effects, week effects and year effects in levels. Note that regressors can be added to the model to capture possible effects that would otherwise be included in the error, confounding the estimation of the transmission parameter β. Adda (2016) denotes this matrix of regressors by X, not to be confused with the notation by Keeling and Rohani (2011) for the number of infectives. For this reason, we instead denote the matrix of regressors as used by Adda (2016) by M. This leads to the following formulation:

$$I_{r,t} - I_{r,t-1} = \beta S_{r,t-\tau} I_{r,t-\tau} - \gamma I_{r,t-\tau} + \delta M_{r,t} + \eta_{r,t} \tag{3.7}$$

For our application, the data does not span multiple years. As such, we do not have year effects. Moreover, given that year effects are not available, week effects would capture a time trend. However, we do add a weekend effect. The main reason behind this is that we expect that less people may be detected on the weekend due to possibly some general practitioner practices or testing locations being closed on the weekend, meaning that people who are not willing or able to travel far will not get tested. More information and reasoning is provided in Section 4.3.

Fourthly, there are two other key differences in the model by Adda (2016) that are not properly explained in the paper. First of all, Adda (2016) does not include the term $\gamma I_{r,t-\tau}$ in the model. Presumably, this is because Adda (2016) considers the number of new cases instead of the total number of infectives and, therefore, the number of recovered individuals do not impact that value. Second of all, Adda (2016) replaces the proportion of infectives $I_{r,t-\tau}$ by the number of new cases $Y_{r,t-\tau} - Y_{r,t-\tau-1}$ (where we follow the notation from Keeling and Rohani (2011)). Similarly, Adda (2016) puts the dependent variable to be the number of new cases instead of the number of infectives divided by the population (the incidence rate). In the paper, it is not clearly explained why this is

a correct step to take. To redefine the left-hand side of (3.7), we need to multiply $I_{r,t}$ with $N_{r,t}$ and $I_{r,t-1}$ with $N_{r,t-1}$. On the right-hand side, we need to multiply $I_{r,t-\tau}$ with $N_{r,t-\tau}$ as well as subtracting $\beta S_{r,t-\tau} Y_{r,t-\tau-1}$ to obtain the number of new cases. Mathematically, these two operations are not equivalent and it is unclear why this is a viable operation. Nonetheless, the error that may occur from these mathematical operations is then represented in the error term $\eta_{r,t}$. In this section, we will continue with this model. Similarly, we will follow the model definitions by Adda (2016) in Sections 3.3 and 3.4. In Section , we define our own discretized SIR model.

TODO: Section

Defining $\Delta Y_t \coloneqq Y_t - Y_{t-1}$ and following the notation by Keeling and Rohani (2011), the within-region model as defined by Adda (2016), ignoring effects across regions, is given by:

$$\Delta Y_{r,t} = \beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \delta M_{r,t} + \eta_{r,t} \tag{3.8}$$

The model is estimated by ordinary least squares (OLS). The moment condition that needs to be satisfied due to the strict exogeneity assumption is

$$E\left[\eta_{r,t} \left(\beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \delta M_{r,t}\right)\right] = 0.$$

A general assumption that is made, is that the idiosyncratic error $\eta_{r,t}$ is uncorrelated with the regressors in the matrix $M_{r,t}$. That is, we assume that $E[\eta_{r,t} \mid M_{r,t}] = 0$. Now note that we need to only consider the relation between $\eta_{r,t}$ and $\Delta Y_{r,t-\tau} S_{r,t-\tau}$. The reason why we assume that $E[\eta_{r,t} \mid \Delta Y_{r,t-\tau} S_{r,t-\tau}] = 0$ is that, for a large enough lag τ, the error is not correlated with past data at that lag. That is, the people that are classified as infectives at time $t - \tau$ do not have an effect on the error that we make when considering the infectives at time t under a correct model specification. As such, we assume that the moment condition holds.

3.3 Weighted Within-Region Spread Model

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 β 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 (3.8). The resulting model is a completely new addition. Moreover, note that Adda (2016) defines a model, which is called Full Econometric Model, that takes effects across regions into account as well as allowing for the transmission rate parameter to be weighted by other factors. Due to a lack in available data, this model is not estimated.

Let the tensor W contain K region-specific variables that may influence the transmission rate β. As such, we now allow for β_{within} to differ for these K variables. Unfortunately, we do not have access to a lot of data at the required granular level. For instance, from Eurostat, 2020b, several statistics on the aforementioned characteristics are available, but these are only available on an annual level. Therefore, the selection of variables is limited. We use data from the Google Mobility Report (Google LLC, 2020), which uses anonymous location services to aggregate users' data to obtain a percentage change in the median amount of people that, for instance, visit grocery stores, parks, or use public transport. In Section 4, we elaborate on how these variables included in W are specifically defined. We define M and η in the same way as in Section 3.2. Taking this into account, the weighted within-region model is defined as:

$$\Delta Y_{r,t} = \Delta Y_{r,t-\tau} S_{r,t-\tau} \sum_{k=1}^{K} \beta_{within}^{k} W_{r,t-\tau}^{k} + \delta M_{r,t} + \eta_{r,t} \tag{3.9}$$

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

$$E\left[\eta_{r,t}\left(\Delta Y_{r,t-\tau} S_{r,t-\tau} \sum_{k=1}^{K} \beta_{within}^{k} W_{r,t-\tau}^{k} + \delta M_{r,t}\right)\right] = 0.$$

The same reasoning as in Section 3.2 applies with regards to the assumption that $E[\eta_{r,t} \mid Y_{r,t}] = 0$. On the assumption that $E[\eta_{r,t} \mid \Delta Y_{r,t-\tau} S_{r,t-\tau} \sum_{k=1}^{K} \beta_{within}^{k} W_{r,t-\tau}^{k}] = 0$, we should realize that the only difference with the reasoning in Section 3.2 is the addition of the weighting matrix W. Since these are external factors, such as socioeconomic variables, we believe that these will be uncorrelated with the error. Combining this with the arguments from Section 3.2, we assume that the moment condition holds.

3.4 Within and Between-Region Spread Model

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. The following model is defined:

$$\Delta Y_{r,t} = \beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \beta_{between} S_{r,t-\tau} \sum_{c \in R \setminus r} \Delta Y_{c,t-\tau} + \delta M_{r,t} + \eta_{r,t} \tag{3.10}$$

TODO: Why may this specification not be good

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

$$E\left[\eta_{r,t}\left(\beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \beta_{between} S_{r,t-\tau} \sum_{c \in R \setminus r} \Delta Y_{c,t-\tau} + \delta M_{r,t}\right)\right] = 0.$$

In the same way as in Section 3.2, we can assume that $E[\eta_{r,t} \mid M_{r,t}] = 0$ and $E[\eta_{r,t} \mid \Delta Y_{r,t-\tau} S_{r,t-\tau}] = 0$. Following the same reasoning as before, we assume that the number of infectives who come into contact with susceptibles in other regions at a certain time is not correlated with the error if the lag is large enough. As such, we assume that the moment condition holds.

In (3.10), the transmission parameter β is now allowed to be different within and between regions. Adda (2016) estimates (3.10) 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. For this reason, we only consider OLS for this model.

3.5 Discrete SIR Model

As explained in Section 3.2, there are several steps made by Adda (2016) that are not explained clearly in the paper and that do not seem to mathematically grounded. As such, we provide our own derivation for a discrete SIR model. In this section, we present the model after which we describe our estimation approaches with panel data and Bayesian methods.

Because we are interested in the number of infectives, we will discretize equation (3.5). Recall that this is given by:

$$\frac{dI}{dt} = \beta SI - \gamma I$$

When taking the possible estimation error $\eta_{r,t}$, whether from discretizing or estimating the parameters, into account, this is equivalent to:

$$I_{r,t} - I_{r,t-1} = \beta S_{r,t-1} I_{r,t-1} - \gamma I_{r,t-1} + \eta_{r,t} \tag{3.11}$$

Let us look at this equation more thoroughly. Denote the total number of infectives by $Y_{r,t}$ and let $N_{r,t}$ denote the total population in region r at time t. In Section 3.1, we explained that the SIR model ignores births and deaths so that the population is constant. Births are ignored entirely and deaths are included in the group R. As such, the total population $N_{r,t}$ is actually assumed to be constant over time. Therefore, we now denote the total population for a region r by N_r. Recall that, therefore, $I_{r,t} \coloneqq Y_{r,t}/N_r$ is then the proportion of the population of region r that is infected at a time t. As such, the left-hand side is the change in the proportion of the population that is infected from one day to the next:

$$I_{r,t} - I_{r,t-1} = \frac{Y_{r,t} - Y_{r,t-1}}{N_r}$$

The right-hand side consists of three terms. Firstly, the term $\beta S_{r,t-1} I_{r,t-1}$ relates to the observation that new infectives are formed due to interaction of infectives with the susceptible population, i.e. the people that move from the group X to the group Y. In Section 3.1, we explained that it is assumed that this rate β is constant over time. However, we concede that the addition of introducing a longer lag that Adda (2016) makes is valid. Indeed, when a susceptible person meets an infective and consequently gets infected, this person is not immediately infectious themselves. For the same reasons as laid out in Section 3.2, we replace the lag in the first term by a longer lag τ, which we set to be equal to the estimated latent period of three days:

$$I_{r,t} - I_{r,t-1} = \beta S_{r,t-\tau} I_{r,t-\tau} - \gamma I_{r,t-1} + \eta_{r,t} \tag{3.12}$$

Secondly, the term $-\gamma I_{r,t-1}$ describes the infectives that recover or die from the disease, i.e. the people that move from group Y to group Z. Lastly, as mentioned in the previous paragraph, $\eta_{r,t}$ denotes the error. As was the case for the other models, we assume the error to be idiosyncratic.

3.5.1 Panel data methods

The discrete SIR model in (3.12) is estimated by panel data methods and Bayesian estimation. We first discuss the panel data methods. Panel data refers to a dataset that contains measurements over time for a group of individuals. In our case, the individuals are represented by the R regions and we have observations for T time periods. The main motivation behind using panel data methods is that they comprise more information, making them more efficient. Moreover, because the same individual is observed multiple times over time, they are able to identify unobserved individual effects (heterogeneity) that are persistent over time, such as an aversion to spending money or another effect that causes an individual that experienced some event in the past to experience that effect in the future again with a higher probability. A regular cross-sectional model is usually described as:

$$y_i = x_i'\theta + \epsilon_i, \quad \forall i = 1, \ldots, R \tag{3.13}$$

where y denotes the dependent variable, x denotes the independent variables, θ is the vector of parameters to be estimated, and ϵ is an idiosyncratic error term. A panel data model is usually described in a very similar way:

$$y_{i,t} = x'_{i,t}\theta + \epsilon_{i,t}, \quad \forall i = 1, \ldots, R;\ t = 1, \ldots, T \tag{3.14}$$

where the error component $\epsilon_{i,t} \coloneqq \alpha_i + u_{i,t}$ is a composite error term, comprising a time-invariant individual effect, denoted by α_i, and an idiosyncratic error term, denoted by $u_{i,t} \sim N(0, \sigma_u^2)$. We assume that the individual effect is not correlated with the regressors:

$$E(\alpha_i \mid x_{i,1}, \ldots, x_{i,T}) = 0 \tag{3.15}$$

For the discrete SIR model in (3.12), note that:

$$y_{i,t} = I_{r,t} - I_{r,t-1}, \quad \theta = \begin{pmatrix} \beta \\ -\gamma \end{pmatrix}, \quad x_{i,t} = \begin{pmatrix} S_{r,t-\tau} I_{r,t-\tau} \\ I_{r,t-1} \end{pmatrix}, \quad u_{i,t} = \eta_{r,t}$$

Looking at this notation, we can see that it is safe to impose the assumption that the individual effect and the independent variables are orthogonal: it seems odd to assume that the individual regional effect, which is time-invariant, is correlated with the time-varying incidence rate or susceptible rate of this one specific disease. In the rest of this section, we will follow the notation as in (3.14).

There are three main panel data models that are usually applied: the pooled OLS (POLS), fixed effects (FE), and random effects (RE) models. The choice between these models depends on the assumptions that are placed on the individual effect α_i. For each of these three methods, we define them and discuss the required exogeneity assumption(s) and possible other interesting features.

Pooled OLS ignores the individual effect, hence treating the data as a cross-section. The exogeneity assumption that needs to be satisfied is a simple exogeneity assumption:

$$E(x_{i,t}\epsilon_{i,t}) = 0, \text{ so } E(x_{i,t}u_{i,t}) = 0 \text{ and } E(x_{i,t}\alpha_i) = 0$$

Because the errors depend on α_i for all time periods, the correlation between two errors $\epsilon_{i,t}$ and $\epsilon_{i,s}$ does not decrease as the time periods grow farther apart. This means that, even when the exogeneity assumption is satisfied, POLS has an efficiency problem. Because we expect that there will be a vast difference between the Italian regions, POLS is likely to not be a good model for this thesis. Nonetheless, we estimate the results and compare them to those of the fixed and random effects models.

The fixed effects model makes no assumptions about the time-invariant individual effect but applies a within-transformation, namely time-demeaning the data, to eliminate

it from the model. The effect is that all time-invariant regressors, such as dummy variables gender or one's education level, are also eliminated from the model. The fixed effects model equation is computed in two steps:

1. Compute the time-means, namely $\bar{y}_i$, $\bar{x}_i$, $\bar{u}_i$, and $\bar{\alpha}_i = \alpha_i$ to form the following equation:
$$\bar{y}_i = \bar{x}_i'\theta + \alpha_i + \bar{u}_i$$

2. Subtract the time-means from (3.14) to obtain the fixed effects model equation:
$$(y_{i,t} - \bar{y}_i) = (x_{i,t} - \bar{x}_i)'\,\theta + (u_{i,t} - \bar{u}_i)\,, \quad \forall i = 1, \ldots, R;\ t = 1, \ldots, T \tag{3.16}$$

The fixed effects model is then to apply OLS to (3.16) to obtain the estimates of θ. Note that (3.11) does not contain time-invariant components. As such, the only eliminated component of the discrete SIR model is the individual effect. The fixed effects model requires the strict exogeneity assumption:

$$E\left(u_{i,t} \mid \alpha_i, x_{i,1}, \ldots, x_{i,T}\right) = 0, \quad \forall t = 1, \ldots, T \tag{3.17}$$

This assumption essentially says that the variables are not allowed to depend on any of the errors, whether in the past, present, or future.

The main idea behind the random effects model is to impose a distribution on the individual effects that can then be included in the error term. The random effects model equation is identical to (3.14), where we assume that $\alpha_i \sim N\left(0, \sigma_\alpha^2\right)$. Because the individual effect is included in the error term, we need to impose assumptions on the entire composite error term. As was the case for the fixed effects model, the random effects model also requires the strict exogeneity assumption. However, because the individual effects are included in the error term, we need strict exogeneity between $\epsilon_{i,t}$ and $x_{i,t}$. This is achieved by the assumption that the individual effect is not correlated with the regressors in (3.15). As such, combining the strict exogeneity assumption and the orthogonality assumption, the aforementioned strict exogeneity assumption between $\epsilon_{i,t}$ and $x_{i,t}$ holds. We will not delve into the details of how the random effects model is specifically estimated; suffice to say that it is estimated by generalized least squares (GLS).

3.5.2 Bayesian estimation methods

In the last few months, a lot of research has been done on the spread of SARS-CoV-2 for various locales. Therefore, we may have some idea of what the values of the parameters are likely to be. Continue here

3.6 Model selection

For model selection, we use the Akaike Information Criterion or AIC (Akaike, 1974). The AIC for a particular model is defined as

$$AIC = -2\log(ML) + 2k, \tag{3.18}$$

where ML denotes the maximum likelihood for the model and k denotes the number of parameters in the model. In contrast, one could also consider the Bayesian Information Criterion or BIC (Schwarz et al., 1978). Schwarz et al. (1978) developed it as an alternative to the Akaike Information Criterion. The BIC is defined as

$$BIC = -2\log(ML) + k\log(n), \tag{3.19}$$

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. However, the AIC is efficient in the sense that it will asymptotically select the model that minimizes 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) shows by simulation that the BIC can fail in finite sample sizes even if the true model is in the candidate set. This is because the BIC has a higher maximum risk, defined as the mean squared error of estimating the true covariance matrix. Because we believe that, indeed, the true model generating the data will quite likely not be included in our candidate set, we use the AIC to perform model selection.

3.7 Modelling undocumented infectives

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 infectives 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 infectives should be taken into account.

In this section, we aim to model the undocumented infectives. Note that, by definition, there is no data on the amount of undocumented infectives 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 infectives. 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). However, we do not take this into account in this thesis besides including the indicator variable for the lockdown, as described in Section 4.3. Future research could be done to include these restrictions more robustly.

At a point in time t, we denote the testing capacity by TC_t. In Section 4.2, we explain how a measure of the testing capacity is obtained. The total number of infected people at time t is denoted by Y_t. This group can be subdivided into the documented infectives DI_t and the undocumented infectives UI_t such that $DI_t + UI_t = Y_t$. Therefore, we can denote the documented and undocumented infectives 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:

$$f_t \coloneqq f(TC_t), \tag{3.20}$$

such that

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

Notice that the undocumented infectives can then be written as $UI_t = \frac{1-f_t}{f_t} DI_t$.

There are some properties that (3.20) should satisfy and some assumptions that we make. These are as follows:

(A1) Since f_t is a proportion, we need to have $f_t \in [0, 1]$.

(A2) If no one is tested, we assume that there are a certain minimum amount of documented infectives, denoted by $f^{min} \in [0, 1]$. That is,

$$f(0) = f^{min}.$$

Note that at any point in time, it should hold that

$$\begin{aligned} & DI_t + UI_t < N_t \\ \iff & DI_t + \frac{1 - f_t}{f_t} DI_t < N_t \\ \iff & \frac{1}{f_t} DI_t < N_t \\ \iff & f_t > \frac{DI_t}{N_t}, \end{aligned}$$

so f^{min} should be chosen to be larger than $\frac{DI_t}{N_t}$. The fact that this is true should be clear. If f_t would be lower than the fraction of the population that is documented to be infective, then the total number of infectives in a population would exceed the total number of people living in that population, which is not possible.

(A3) 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 infectives 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). Moreover, BMJ (2020) reports that serological tests for COVID-19 carry with them risks of bias and

heterogeneity in their accuracy. Therefore, they state that these serological tests should only be used cautiously for clinical decision making and epidemiological surveillance. For this reason, one could choose to relax the assumption and assume $f(N_t) = f^{max}$ for some $f^{max} \in [0, 1]$ set to be a more reasonably perceived value.

(A4) As mentioned earlier in this section, 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

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

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

- **Linear form**

$$f_t = \frac{1 - f^{min}}{N_t} TC_t + f^{min}. \tag{3.21}$$

- **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 $f\left(\frac{1}{2}N_t\right) = \gamma$ for some $\gamma \in \left[\frac{1}{4} + \frac{3}{4}f^{min}, \frac{3}{4} + \frac{1}{4}f^{min}\right]$. Then the formula becomes:

$$f_t = \frac{2 - 4\gamma + 2f^{min}}{N_t^2} TC_t^2 + \frac{4\gamma - 1 - 3f^{min}}{N_t} TC_t + f^{min}. \tag{3.22}$$

  If $\gamma \in \left[\frac{1}{4} + \frac{3}{4}f^{min}, \frac{1}{2} + \frac{1}{2}f^{min}\right)$, the function is upwards opening. If $\gamma \in \left(\frac{1}{2} + \frac{1}{2}f^{min}, \frac{3}{4} + \frac{1}{4}f^{min}\right]$, the function is downwards opening. If $\gamma = \frac{1}{2} + \frac{1}{2}f^{min}$, then the formula simplifies to the linear specification. In appendix D.2.2, we explain why γ cannot be below $\frac{1}{4} + \frac{3}{4}f^{min}$ or above $\frac{3}{4} + \frac{1}{4}f^{min}$.

  – 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). Choosing this special case means that there will be no unknown parameters needed to define the function because we know the location of the vertex and a known point $(0, f^{min})$ on the parabola. We can then derive that the formula becomes:

$$f_t = \frac{f^{min} - 1}{N_t^2} TC_t^2 - \frac{2(f^{min} - 1)}{N_t} TC_t + f^{min}. \tag{3.23}$$

  Note that this is equivalent to (3.22) for $\gamma = \frac{3}{4} + \frac{1}{4}f^{min}$. 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 $(0, f^{min})$, i.e. the parabola is upwards opening. We can then derive that the formula becomes:

$$f_t = \frac{1 - f^{min}}{N_t^2} TC_t^2 + f^{min}. \tag{3.24}$$

Note that this is equivalent to (3.22) for $\gamma = \frac{1}{4} + \frac{3}{4} f^{min}$. 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:

$$\begin{aligned} f(TC_t) = &\frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3N_t^3} TC_t^3 + \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14f^{min}}{N_t^2} TC_t^2 \\ &+ \frac{1 + 32\gamma_1 - 12\gamma_2 - 21f^{min}}{3N_t} TC_t + f^{min}, \end{aligned} \tag{3.25}$$

No bounds on γ_1 and γ_2 have been set. Particularly, there are combinations of γ_1 and γ_2 for which the codomain of f_t on $TC_t \in [0, N_t]$ may not be the interval $[0, 1]$, violating assumption (A1), and for which the function is not monotonically increasing, violating assumption (A4). One could derive explicit conditions on possible combinations for γ_1 and γ_2 such that this is not the case but this is not done in this thesis.

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

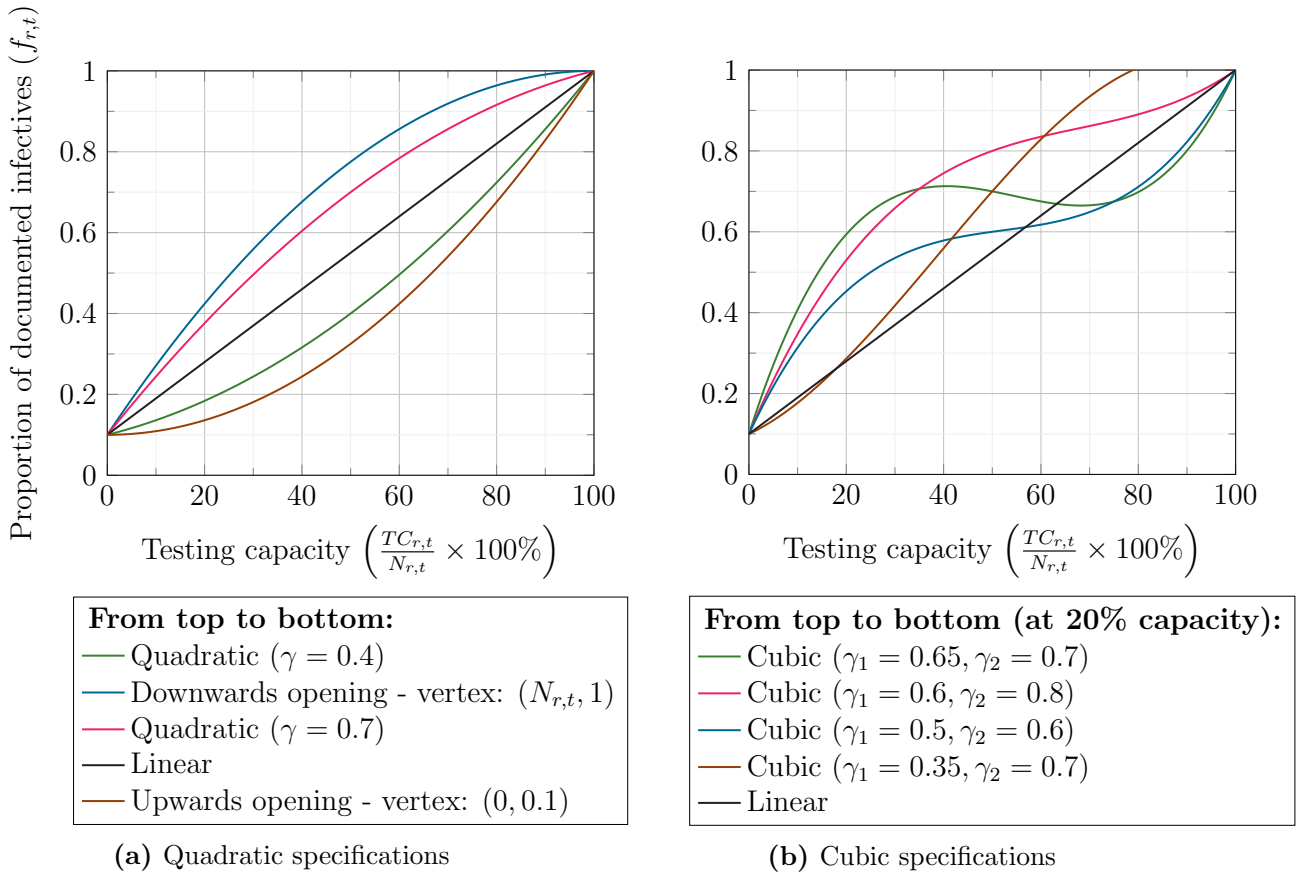
$$f_{r,t} \coloneqq f(TC_{r,t}). \tag{3.26}$$

such that

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

In Figure 3.1, we specify several functional forms for the specifications as mentioned above. Figure 3.1a shows four different functional forms for the quadratic functional forms while 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. Moreover, recall that



(a) Quadratic specifications

(b) Cubic specifications

Figure 3.1. Functional forms for the proportion of documented infectives ($f^{min} = 0.1$)

there are combinations of γ_1 and γ_2 for the cubic representation for which assumptions (A1) and (A4) are violated. Figure 3.1b shows that $\gamma_1 = 0.35$ and $\gamma_2 = 0.7$ cause the function to exceed the maximum value allowed for $f_{r,t}$ of 1, despite decreasing so that $f(N_{r,t}) = 1$. A combination of $\gamma_1 = 0.65$ and $\gamma_2 = 0.7$ creates a non-monotonic functional form. 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 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 infectives. As such, we 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 infectives. 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 (3.22) and (3.23) are equivalent when $\gamma = \frac{3}{4} + \frac{1}{4}f^{min}$, meaning that (3.23) is the most extreme case possible and that the slope cannot be constructed to be more steep. To be general, we choose (3.22) 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 are interested in investigating the relationship between $TC_{r,t}$ and $f_{r,t}(\gamma)$ over time for all regions and to compare these. Because the population size differs over the regions, this is likely to impact the absolute number of tests executed. As such, instead of comparing $f_{r,t}(\gamma)$ to $TC_{r,t}$, we compare it to $TC_{r,t}/N_{r,t}$. The results are shown in Figure 3.2.

In Figure 3.2, we can see that the pattern of the relationship between the two variables is similar over time for different groups of regions. Importantly, we note that the proportion of infectives that go undocumented decreases over time. This is logical because the testing capacity increases over time and we have assumed a monotonic relationship.

For illustration purposes, we now give an example of the impact of this modelling method. We present the number of infectives at three time points for three regions in Table 3.1.

Table 3.1. Impact of modelling undocumented infectives over time. The quadratic specification with $\gamma = 0.7$ is used.

	Calabria			Lombardy			Veneto		
	DI_t	f_t	I_t	DI_t	f_t	I_t	DI_t	f_t	I_t
April 1	669	10.8%	6,448	44,601	11.8%	409,003	9,592	13.4%	82,106
June 1	1,158	15.4%	10,670	88,846	21.0%	717,289	19,121	29.5%	139,610
August 1	1,269	19.2%	11,291	96,102	28.5%	747,691	20,133	44.2%	142,111

Table 3.1 shows us that the impact of the proportion of documented infectives f_t differs over the regions. We chose Calabria, Lombardy, and Veneto because these regions vary in the proportional amount of tests executed, leading to different profiles in

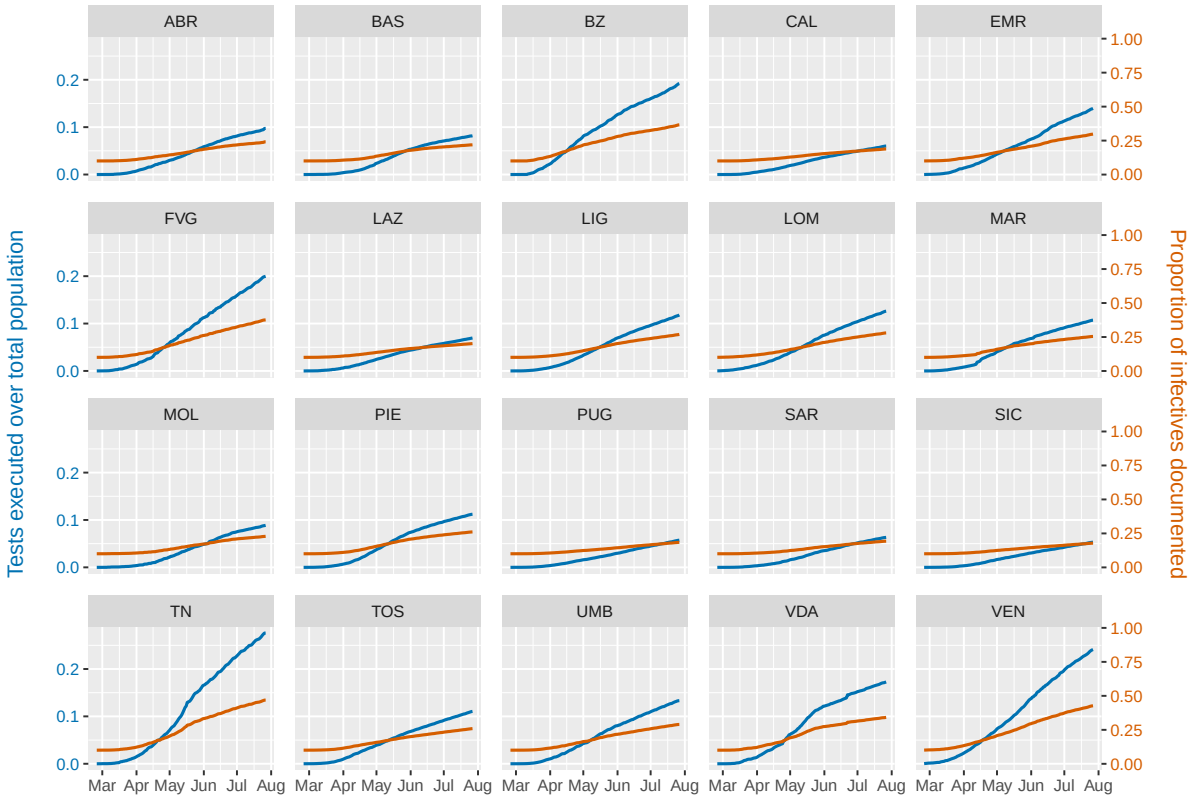


Figure 3.2. Total number of people tested over the total population ($TC_{r,t}/N_{r,t}$) versus proportion of infectives that are documented $f_{r,t}(\gamma)$

f_t. We can see this profile in Figure 3.2 as well. When the amount of tests executed grows less steeply, as is the case in Calabria, the number of undocumented infectives in society grows stronger. On the other hand, for a region that invests heavily in testing, such as Veneto, the undocumented infectives are less pronounced. For example, consider the changes in Calabria and Veneto from June 1 to August 1. For Calabria, the growth in the documented infectives accounted for only 17.87% of the total growth in infectives. In contrast, in Veneto the growth in the documented infectives accounted for 40.46% of the total growth. Of course, Lombardy finds itself in the middle, where documented infectives make up 23.87% of the total growth. Hence, our method correctly incorporates the intuition that a higher testing capacity leads to more infectives being documented.

Using the specification of undocumented infectives, we can now adapt the models (3.8)-(3.10) to include these undocumented infectives. Let us take the within-region spread model (3.8) as an example. Recall that this model was given as

Discrete SIR model

$$\Delta Y_{r,t} = \beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \delta M_{r,t} + \eta_{r,t}.$$

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

$$\frac{DI_{r,t}}{f_{r,t}(\gamma)} = \beta_{within} \frac{DI_{r,t-\tau}}{f_{r,t-\tau}(\gamma)} S_{r,t-\tau} + \delta M_{r,t} + \eta_{r,t}. \tag{3.27}$$

We can rewrite (3.27) as follows

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

The moment conditions that need to hold are:

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

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 (3.8):

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

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 (3.9) and (3.10) so that their moment conditions also do not depend on $f_{r,t}(\gamma)$.

4 Dataset

In this section, we outline the structure of the data that is used and how it was retrieved. Firstly, we discuss the structure of Italian regions in Section 4.1. Subsequently, we look at the data on COVID-19 such as the incidence rate in Section 4.2. Here, we also discuss how we tackled possibly errors in the data, as well as missing values. Lastly, Section 4.3 discusses the variables that are included in the weighted model in Sections 3.3.

4.1 Geographical structure of Italy

The NUTS classification (Nomenclature of territorial units for statistics, from the French *Nomenclature des Unités Territoriales Statistiques*) 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* (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*, comparable to *Het Gooi*, *Twente*, or the *Achterhoek* in the Netherlands.

4.2 Coronavirus data

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 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 (Rosini, 2020). This data is divided up between the NUTS 2 regions. 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 outside of the total number of cases was not reported at this granular level. As such, we choose to use the NUTS 2 regions.

For $R = 21$ Italian regions, we retrieved the data on the coronavirus from February 25, 2020, until August 16, 2020, leading to a total amount of time observations of $T = 174$ and a total amount of observations of $N \times T = 3,654$. The statistics that are of interest to us are:

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

- Total amount of deaths (*deceduti*);
- 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*).[1] 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 or computational error. We take the first difference of the total amount of positive cases to define the number of confirmed cases. Secondly, the semantic 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 because 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). Actually, only 1.2% of the deceased patients in Italy

[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

until March 19, 2020 had a pre-existing condition (European Centre for Disease Prevention and Control, 2020). Of the patients that died and did have at least one comorbidity, 48.6% had three or more comorbidities, 26.6% had two comorbidities, and 23.5% had one comorbidity. European Centre for Disease Prevention and Control (2020) also reports that 73.8% of the deceased patients had hypertension, 33.9% diabetes, 30.1% ischaemic heart disease, 22.0% atrial fibrillation, and 19.5% had a cancer diagnosed in the last five years. As such, it may be likely the case that a patient died from, for instance, hypertension but because they were infected by SARS-CoV-2 their death was classified as a COVID-19 death instead. 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. In this thesis, we assume that this error is negligible.

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 or, otherwise, the error will be assumed to be consistently applied to the data received from that region. 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 twenty-two times that the number of confirmed cases was reported to be negative (for 11 different regions). The number of deaths was reported to be negative eight times (for 6 different regions) and the number of recovered patients was reported with a negative value sixty-two times (for 14 different regions). 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. An example for the region of Basilicata is given in Table 4.1.

Table 4.1. Example of the propagation of negative values for the region of Basilicata

Date	Original values	Step 1	Step 2	Step 3	Final step
May 3	6	6	6	6	2
May 4	0	0	0	0	0
May 5	10	10	10	-4	0
May 6	3	3	-14	0	0
May 7	-16	-17	0	0	0
May 8	-1	0	0	0	0

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 should note that a highly negative value of -229 was 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. The same applies to Sicily, where a negative value of -394 was reported on June 19, 2020. There, the number of new cases in the week before that date only ranges from 0 to 2. We assume that this corrects for all errors in the past, not just those close to June 12 and 19. Propagating this error backwards as described before would lead to zero new cases per day for Campania from May 13 until June 12 (31 days) and for Sicily from April 28 until June 19 (53 days). Since we have no reason to know how this error is distributed, we remove the regions of Campania and Sicily 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 for Campania or June 19 for Sicily.

Regarding missing values, there are none. It is to be expected that the Department of Civil Protection imputed the missing values with a value of zero. For instance, on July 5, it was reported that zero tests were executed in the region of Basilicata. On the dates around this, around 200 tests were executed. On July 9, a higher value of 426 was reported. We expect that this is to correct for the reported value of zero of July 5. We could, for instance, distribute the 426 among July 5 and 9. However, in this thesis, because we do not know for sure if this is indeed correct and other low values are also reported (such as a value of three tests being executed on July 19 for Basilicata), we do not deal with these outliers and leave them as is.

4.3 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 as described in Section 4.1, 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 keep 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.

We distinguish three sets of regressors, as mentioned in Section 3. Firstly, we have a set of control variables included in the tensor $M_{r,t}$ which are not assumed to have a (large) effect on the transmission parameter β. 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; if still applicable

Table 4.2. Specification of regressors

Matrix	Variable	Year	Description
$M_{r,t}$	weekend	n/a	Binary indicator denoting if the day is on the weekend (Saturday or Sunday)
$W_{r,t}$	touristArrivals	2018	Number of tourist arrivals
	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
	deathRatePneumonia	2016	Number of deaths from pneumonia per 100,000 inhabitants
	availableBeds	2018	Number of hospital beds
	populationDensity	2018	Amount of people per square kilometer
$\widetilde{W}_{c,r,t}$			

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.

Figure 4.1 shows us that the largest correlation is 0.64 and occurs between the dis-

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

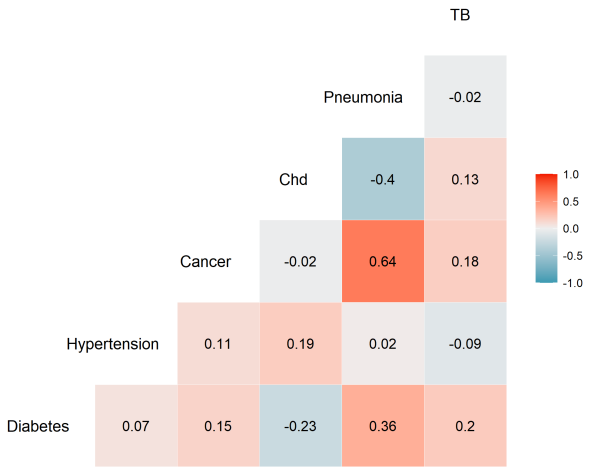


Figure 4.1. Correlation matrix of the discharge rates for various comorbidities of COVID-19

charge rates of pneumonia and cancer. We also see a relatively high correlation of -0.4 between the discharge rates of pneumonia and coronary heart disease. 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. When we speak of statistical significance without specifying a significance level, a level of 0.05 is implied. In Section 5.1, we discuss the results for the within-region spread model. Subsequently, we present the results for the weighted within-region spread model in Section 5.2. After this, Section 5.3 discusses the results for the within and between-region spread model. For the models in Sections 5.1 until 5.3, Adda (2016) explains that the estimated coefficients can be interpreted as the marginal effects of a change in the infection rate on the future infection rate when the entire population is susceptible to the disease. After discussing the models by Adda (2016), we present the results from estimating the discrete SIR model in Section 5.4.

5.1 Within-Region Spread Model

In this section, we present the results for the within-region spread model. Recall that this was given in equation (3.8) as:

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

Firstly, we 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 Akaike Information Criterion (AIC). For both result sets, we present the results from the regular model as well as modelling the undocumented infectives with a quadratic form with $\gamma = 0.7$ and $f^{min} = 0.1$ as in (3.22). We also apply a rolling window of 100 days, meaning that the used data spans May 9 till August 16.

TODO: Write section and refer

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 2, it may achieve good results if regions are sufficiently similar. However, as we have already seen in Section 2, there is indeed a difference among regions and pooling the data is likely not a good solution. The results from estimating (3.8) with OLS for the national data are given in Table 5.1. Model selection with AIC is not applied.

Table 5.1. Estimates from Within-Region Spread Model on a national level. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Regular model				Modelling undocumented infectives			
	Estimate	Std. Error	t value	p value	Estimate	Std. Error	t value	p value
Intercept	110.361	27.954	3.948	0.000***	333.176	111.089	2.999	0.003**

Table 5.1 continues on next page

Table 5.1 continued from previous page

	Regular model				Modelling undocumented infectives			
	Estimate	Std. Error	t-value	p-value	Estimate	Std. Error	t-value	p-value
Weekend	-38.434	32.711	-1.175	0.243	-164.440	147.739	-1.106	0.272
β_{within}	0.691	0.065	10.691	0.000***	0.765	0.049	15.647	0.000***

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

Table 5.1 shows estimates for β_{within} of 0.691 and 0.765 for the models excluding and including undocumented infections, respectively. Both estimated parameters are statistically significant at a 1% significance level, whether undocumented infectives are modelled or not. As mentioned earlier in this section, this model does not take into account effects specific to regions. In Table B.1 in Appendix B, 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 as was explained in Section 3.6. To execute model selection, we use the AIC and we make sure that the term for β_{within} remains in the model. The models also retain an intercept, by definition. As such, model selection is solely performed on whether the weekend dummy should be included. In Table 5.2, we present the results. The results comparing the use of the BIC over the AIC for model selection are presented in Table B.2.

M includes intercept

Table 5.2. Estimates from Within-Region Spread Model per region with model selection by AIC. Estimates are given with t-statistics in parentheses. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Regular model			Modelling undocumented infectives		
Region	β_{within}	Intercept	Weekend	β_{within}	Intercept	Weekend
National	0.691***	110.361***		0.758***	310.400***	
	(10.691)	(3.948)		(15.647)	(2.999)	
ABR	0.256**	3.478***		0.318***	15.030***	
	(2.573)	(4.562)		(3.394)	(4.211)	
BAS	0.037	0.984**		0.038	4.691**	
	(0.356)	(2.212)		(0.374)	(2.282)	
BZ	0.239**	1.525***	1.048*	0.126	5.258***	2.862
	(2.315)	(3.675)	(1.671)	(1.297)	(4.418)	(1.616)
CAL	0.155	1.904***		0.117	10.950***	
	(1.343)	(4.058)		(1.041)	(4.272)	
EMR	0.231***	25.840***	7.747**	0.379***	80.870***	24.090
	(2.701)	(6.917)	(2.138)	(5.693)	(6.158)	(1.618)
FVG	0.485***	2.012***		0.579***	5.163***	
	(5.050)	(4.123)		(7.180)	(3.451)	
LAZ	0.555***	8.063***	4.285*	0.540***	45.450***	21.030

Table 5.2 continues on next page

Table 5.2 continued from previous page

	Regular model			Modelling undocumented infectives		
Region	β_{within}	Intercept	Weekend	β_{within}	Intercept	Weekend
	(5.506)	(3.441)	(1.725)	(6.129)	(3.693)	(1.482)
LIG	0.654***	5.629***		0.737***	18.540**	
	(8.647)	(3.016)		(11.440)	(2.191)	
LOM	0.607***	58.680***		0.647***	230.400***	
	(7.728)	(3.358)		(8.820)	(2.715)	
MAR	0.439***	2.748***		0.477***	10.930***	
	(5.252)	(3.658)		(6.843)	(3.411)	
MOL	0.296***	0.591***		0.312***	2.840**	
	(5.608)	(2.659)		(6.911)	(2.261)	
PIE	0.715***	9.988***	-7.690	0.725***	44.490***	-44.350*
	(12.510)	(2.951)	(-1.506)	(13.980)	(2.650)	(-1.681)
PUG	0.506***	2.595***		0.453***	16.610***	
	(6.332)	(3.513)		(6.348)	(3.926)	
SAR	0.216**	1.215***		0.191*	6.586***	
	(2.153)	(3.492)		(1.958)	(3.700)	
TN	-9.01×10^{-3}	7.344*		-7.83×10^{-3}	19.420*	
	(-0.086)	(1.802)		(-0.075)	(1.865)	
TOS	0.588***	3.444**	4.245**	0.562***	15.770***	17.460**
	(6.659)	(2.513)	(2.231)	(6.908)	(2.635)	(2.070)
UMB	0.489***	0.809***		0.355***	3.491***	
	(4.266)	(3.077)		(3.192)	(3.595)	
VDA	0.064	0.566***		0.163*	1.772***	
	(0.643)	(3.602)		(1.730)	(3.307)	
VEN	0.435***	15.370***		0.421***	39.120***	
	(4.033)	(3.928)		(4.295)	(4.214)	

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

We will discuss the results regarding the estimates of β_{within} first, after which we discuss the model selection that took place. Considering the estimates of β_{within}, we see that these differ vastly over the regions with varying degrees of statistical significance. We first consider the model without modelling undocumented infectives. For four regions, namely Basilicata (BAS), Calabria (CAL), Trentino (TN), and Aosta Valley (VDA), we find no statistical significance at a level of 0.1. All other estimates are significant at a level of at least 0.05. Looking only at the statistically significant estimates, these range from 0.216 for Sardinia (SAR) till 0.715 for Piedmont (PIE). This already shows that a national model is not to be applied to individual regions, despite the statistical significance of the estimate of β_{within} for the national model. The estimated parameters 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 variable $\Delta Y_{r,t}$.

If we model undocumented infectives, we see that the estimates generally differ a bit from the ones discussed previously. Including the national model as well as the statis-

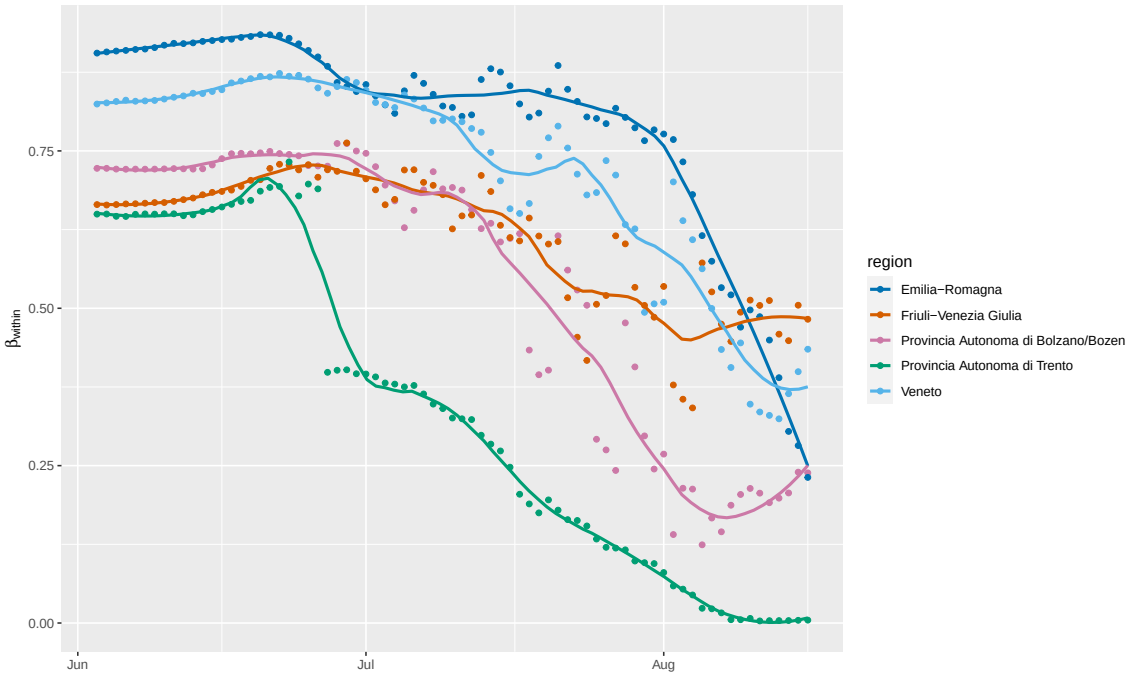
tically insignificant estimates, the estimate is higher in 55% of the cases (eleven out of twenty) and lower in the other 45%. As such, there is no consistent effect of the modelling method used that biases the results in one typical direction. On the subject of statistical significance, there are four regions whose estimate for β_{within} is not statistically significant at a level of 0.1. Three of these are the same as for the previous model: Basilicata, Calabria, and Trentino. In addition, we also find no statistical evidence for the region of Bolzano (BZ). On the other regions, there are two additional regions for which the estimate of β_{within} is only statistically significant at a level of 0.1 and not at a 5% level, namely Sardinia and Aosta Valley. All other estimates are statistically significant at a level of 0.01, ranging from 0.312 for Molise (MOL) till 0.758 for the national model. Excluding the national model, the highest estimate is 0.737 for the region of Liguria (LIG). Again, a national model is not suitable to be applicable to individual regions. Moreover, we once again note that differences in the population count lead to different magnitudes of the other estimated parameters.

Regarding the model selection, we indeed see that the AIC gives a varying model selection per region. However, the set of regressors that is used is the same whether undocumented infectives are modelled are not. We do see that most of the estimates for the weekend dummy's parameter are not statistically significant at a level of 0.05. As mentioned in Section 3.6, this is because the AIC tends to select a larger model. As mentioned, all models retain the intercept and the term $\Delta Y_{t-\tau} S_{t-\tau}$ in the model. In fifteen out of twenty cases, the entire model is selected. In the other five cases, the weekend dummy is excluded.

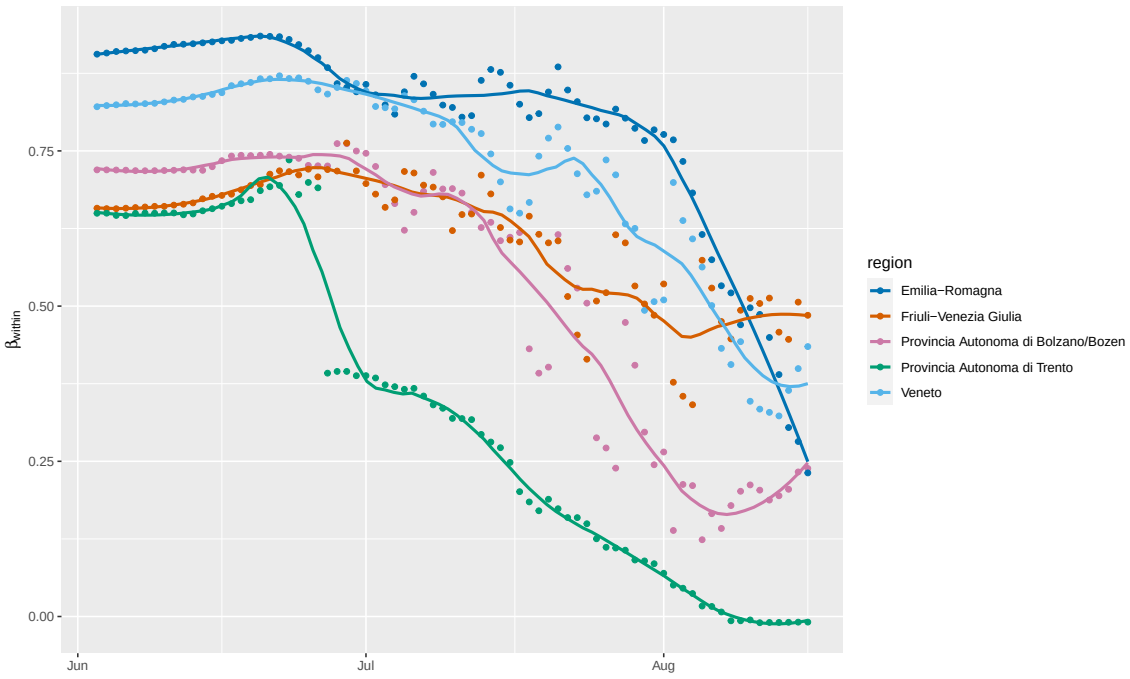
We are also interested in looking at the estimate of β_{within} over time. That is, if we keep adding data, do we see an interesting effect in its progression? Hopefully, it decreases over time, implying that SARS-CoV-2 is transmitted less. This would support the findings in Figure . In Figure 5.1, we present plots for the regions in the *Nord-Est* (North-East) NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.1. Each point in the graphs in Figure 5.1 is the estimate of β_{within} when only the latest 100 data points before that date are used. In addition, a LOESS (locally estimated scatter plot smoothing) curve with span parameter 0.3 is fit to the data points to give a better insight in the progression over time.

Figure: infectives plot

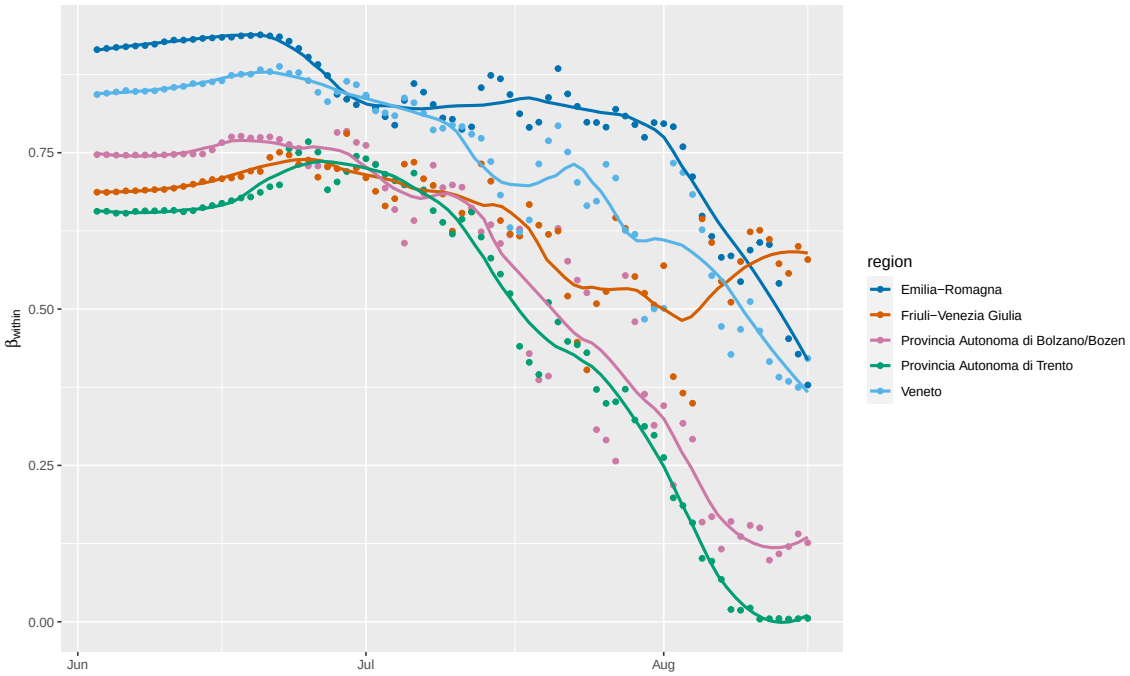
Section



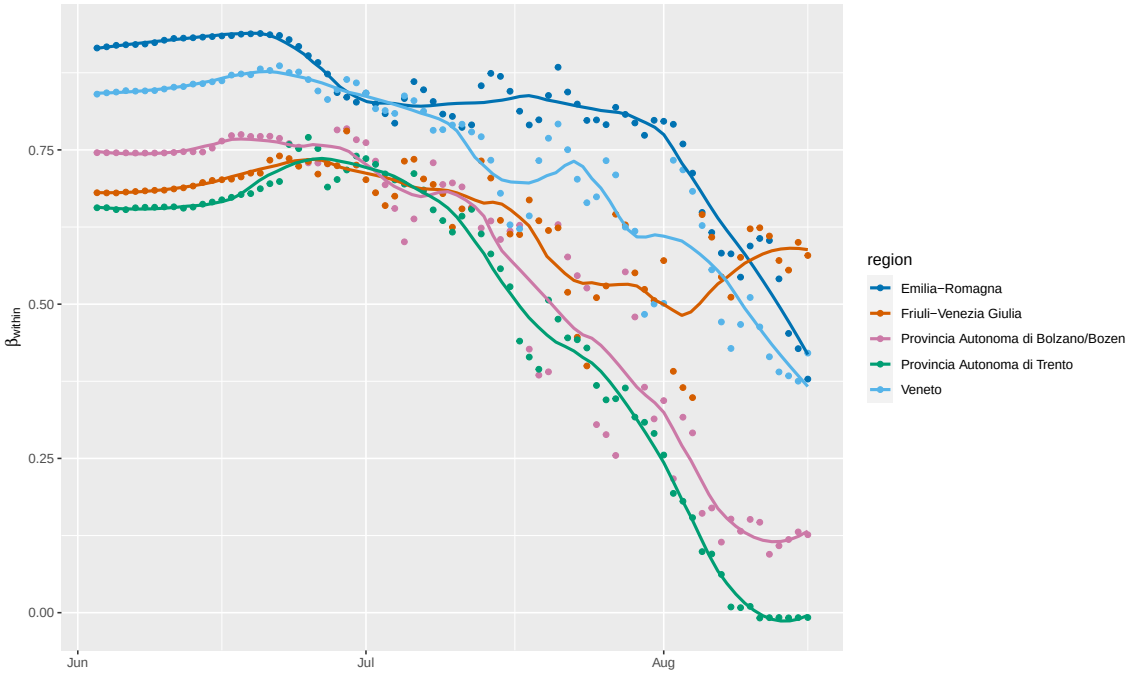
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



(d) With model selection by AIC;
including undocumented infectives

Figure 5.1. Progression of β_{within} over time for the *Nord-Est* (North-East) NUTS 1 region

Considering the progression of β_{within} over time, we indeed see that it decreases over time, as we expected. In addition, at first sight, it may seem like applying model selection does not have an effect on the estimates of β_{within}. However, a slight difference can be seen more clearly when considering the difference between the lines for the regions of Friuli Venezia Giulia and Trentino at the start of the graphs. When comparing the models with their equivalent version when including undocumented infectives, the effect is more pronounced, with higher values for β_{within}.

5.2 Weighted Within-Region Spread Model

In this section, we present the results for the weighted within-region spread model. Recall that this was given in equation (3.9) as:

$$\Delta Y_{r,t} = \Delta Y_{r,t-\tau} S_{r,t-\tau} \sum_{k=1}^{K} \beta_{within}^{k} W_{r,t-\tau}^{k} + \delta M_{r,t} + \eta_{r,t}$$

TODO: Results still need to be inserted.

5.3 Within and Between-Region Spread Model

In this section, we present the results for the within and between-region spread model. Recall that this was given in equation (3.10) as:

$$\Delta Y_{r,t} = \beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \beta_{between} S_{r,t-\tau} \sum_{c \in R \setminus r} \Delta Y_{c,t-\tau} + \delta M_{r,t} + \eta_{r,t}$$

Notice that it does not make sense to consider a national model. Because we do not consider countries outside of Italy, the set $R \setminus r$ is empty if we consider r to be the entire country of Italy. This would mean that the national model for the within and between-region spread model is equivalent to the national model for the within-region spread model. As such, in this section, we only consider the model applied to the regions. Once again, we execute model selection using the AIC and we make sure that the terms for β_{within} and $\beta_{between}$ remain in the model. The models also retain an intercept, by definition. In Table B.3 in Appendix B, we present the results from running the model on each region separately without applying model selection. In Table 5.3, we present the results. The results comparing the use of the BIC over the AIC for model selection are presented in Table B.4.

Table 5.3. Estimates from Within and Between-Region Spread Model per region with model selection by AIC. Estimates are given with t-statistics in parentheses. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Regular model				Modelling undocumented infectives			
Region	β_{within}	$\beta_{between}$	Intercept	Weekend	β_{within}	$\beta_{between}$	Intercept	Weekend
ABR	0.199*	3.763×10^{-3}	2.445**		0.1742*	6.38×10^{-3}***	8.196*	
	(1.833)	(1.297)	(2.220)		(1.668)	(2.753)	(1.927)	
BAS	0.039	-1.286×10^{-3}	1.432*		0.040	-6.479×10^{-4}	5.730*	
	(0.379)	(-0.682)	(1.803)		(0.383)	(-0.448)	(1.846)	
BZ	0.245**	-2.117×10^{-3}*	2.219***	1.136*	0.141	-8.112×10^{-4}	6.422***	2.958*
	(2.396)	(-1.717)	(3.851)	(1.823)	(1.448)	(-1.391)	(4.429)	(1.677)
CAL	0.162	-1.708×10^{-3}	2.486***		0.124	-1.071×10^{-3}	12.590***	
	(1.401)	(-0.954)	(3.228)		(1.093)	(-0.661)	(3.518)	
EMR	0.199**	5.569×10^{-3}	25.370***	7.532**	0.175*	0.022***	81.970***	21.410
	(2.000)	(0.635)	(6.640)	(2.063)	(1.861)	(2.957)	(6.491)	(1.493)
FVG	0.345***	3.882×10^{-3}**	1.187*		0.218**	4.860×10^{-3}***	2.050	
	(3.050)	(2.247)	(1.971)		(2.071)	(4.757)	(1.366)	
LAZ	0.389***	0.017***	5.880**	3.681	0.257**	0.025***	39.65***	18.49
	(3.499)	(3.017)	(2.489)	(1.539)	(2.508)	(4.469)	(3.508)	(1.427)
LIG	0.370***	0.039***	-1.144	-4.453*	0.365***	0.041***	-6.395	-17.86
	(4.450)	(5.708)	(-0.519)	(-1.727)	(4.855)	(6.982)	(-0.738)	(-1.538)
LOM	0.487***	0.330***	29.430	-34.720	0.324***	0.619***	36.300	-167.400
	(5.753)	(3.128)	(1.374)	(-1.411)	(3.732)	(5.503)	(0.396)	(-1.400)
MAR	0.317***	6.051×10^{-3}*	1.318		0.290***	7.460×10^{-3}***	3.872	
	(2.888)	(1.709)	(1.177)		(3.016)	(2.738)	(0.960)	
MOL	0.262***	1.143×10^{-3}	0.2469		0.2637***	1.609×10^{-3}	0.729	
	(4.289)	(1.102)	(0.645)		(4.826)	(1.552)	(0.395)	
PIE	0.417***	0.075***	-2.265	-8.154*	0.301***	0.094***	-13.900	-37.560
	(4.387)	(3.799)	(-0.501)	(-1.707)	(3.004)	(4.775)	(-0.715)	(-1.577)
PUG	0.469***	2.65×10^{-3}	1.890*		0.397***	3.585×10^{-3}	12.980**	
	(5.248)	(0.925)	(1.780)		(4.658)	(1.217)	(2.513)	
SAR	0.224**	-1.188×10^{-3}	1.617***		0.201**	-9.838×10^{-4}	8.086***	
	(2.222)	(-0.872)	(2.797)		(2.041)	(-0.859)	(3.241)	
TN	-0.010	-4.676×10^{-3}	8.936		-7.828×10^{-3}	7.548×10^{-4}	18.230	

Table 5.3 continues on next page

Table 5.3 continued from previous page

	Regular model				Modelling undocumented infectives			
Region	β_{within}	$\beta_{between}$	Intercept	Weekend	β_{within}	$\beta_{between}$	Intercept	Weekend
	(-0.096)	(-0.262)	(1.221)		(-0.075)	(0.100)	(1.153)	
TOS	0.508***	7.268×10^{-3}*	1.887	3.970**	0.382***	0.010***	8.836	16.370**
	(5.103)	(1.677)	(1.147)	(2.098)	(3.842)	(2.944)	(1.421)	(2.016)
UMB	0.498***	-3.378×10^{-4}	0.916**		0.350***	8.692×10^{-5}	3.375**	
	(4.214)	(-0.337)	(2.217)		(2.985)	(0.139)	(2.619)	
VDA	-0.042	2.205×10^{-3}***	-0.124		-0.040	1.762×10^{-3}***	-0.526	
	(-0.425)	(3.547)	(-0.506)		(-0.426)	(5.023)	(-0.795)	
VEN	0.452***	-0.016	20.270***		0.425***	-5.454×10^{-4}	39.700***	
	(4.166)	(-1.204)	(3.594)		(3.997)	(-0.099)	(3.591)	

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

We will discuss the results regarding the estimates of β_{within} and $\beta_{between}$ first, after which we discuss the model selection that took place. There are two things to notice first off the bat. Firstly, notice that the estimates for $\beta_{between}$ are generally much smaller than the estimates for β_{within}, with a notable exception being the region of Lombardy (LOM). This is likely the case because Adda (2016) defined the models with the absolute number of new cases instead of a proportion. As such, summing over all regions leads to a large number of total new cases, causing the parameter estimate to be driven down. The second matter to be noticed is that there are some negative values of the estimated β_{within}, for instance for Trentino (TN) and Aosta Valley (VDA), and $\beta_{between}$, for instance Basilicata (BAS) and Bolzano (BZ). This should not be possible because this means that when infectives meet susceptible people, the incidence rate decreases. Luckily, we see that none of these estimates are statistically significant at a level of 0.05. Noteworthy is the negative estimate for $\beta_{between}$ for Bolzano in the regular model, which is statistically significant at a level of 0.1.

On the topic of statistical significance, recall that there were only four regions that did not have a statistically significant estimate for β_{within} for the within-region spread model at a level of 0.05, namely Basilicata, Calabria (CAL), Trentino, and Aosta Valley. In both models in Table 5.3, we see that these regional models once again do not find statistically significant estimates for β_{within}. In addition, there are more regions that no longer have a statistically significant estimate for β_{within}. For the regular model, this is only the case for the region of Abruzzo (ABR), of which the estimate is only significant at a level of 0.1. When modelling undocumented infections, the four regions that did not find a significant estimate of β_{within} at a level of 0.05 for the within-region spread model were Basilicata, Calabria, Trentino, and Bolzano. In addition to these regions, the models now also do not find statistically significant estimates for three other regions: Abruzzo, Emilia-Romagna (EMR), and Aosta Valley. For the regional models that do find a statistically significant estimate for β_{within} in both models, which are twelve regions, we find a higher estimate when modelling undocumented infectives for only one region, namely the region of Molise (MOL). For all other regions, the estimate is lower. Looking only at the statistically significant estimates of β_{within}, these range from 0.199 for Emilia-Romagna (EMR) till 0.508 for Tuscany (TOS) for the regular model and from 0.201 for Sardinia (SAR) till 0.425 for Veneto (VEN) when undocumented infectives are included. As such, the bounds are much tighter than for the within-region model.

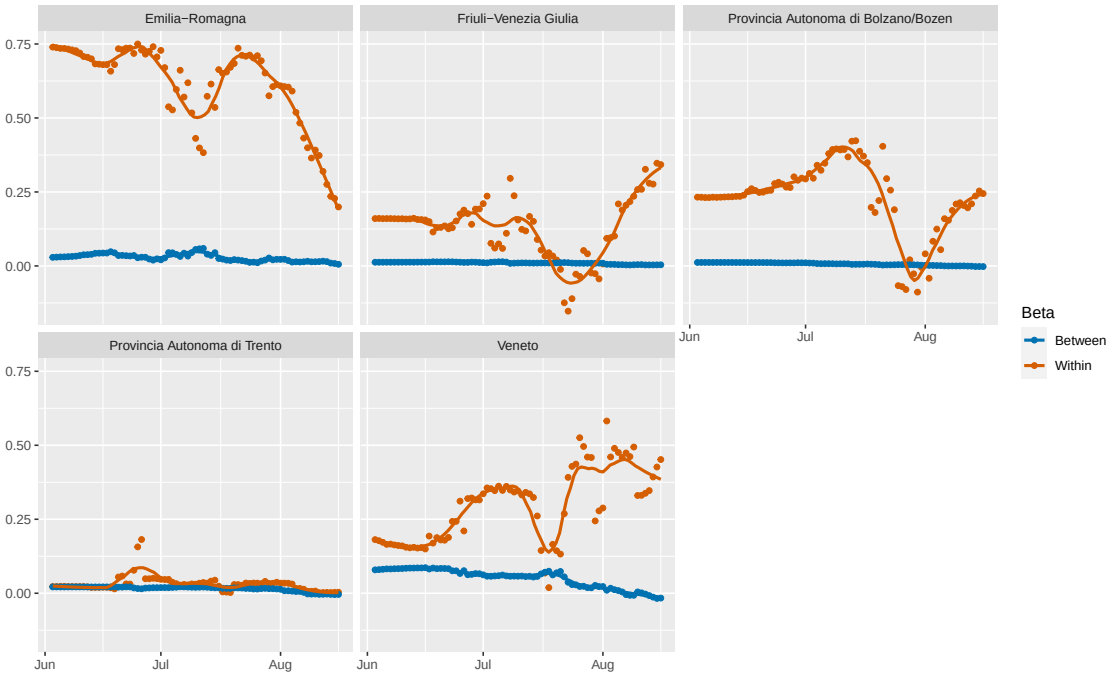
We now consider the estimate of $\beta_{between}$, possibly in combination with conclusions on the estimate of β_{within}. Firstly, it is interesting to note that the estimate for $\beta_{between}$ for the regular model is statistically significant at a level of 0.01 for five regions, at a level of 0.05 only for the region of Friuli Venezia Giulia, and at a level of 0.1 for three regions, namely Bolzano, Marche (MAR), and Tuscany. For the other ten regions, we do not find statistical evidence for the regular model that a between-region effect should be taken

into account. We notice that, for the six regions with a significant estimate of $\beta_{between}$, only one of these regions does not find a significant estimate for β_{within}, namely Aosta Valley. As such, it seems that a significant between-region effect goes hand-in-hand with a significant within-region effect but not necessarily vice versa.

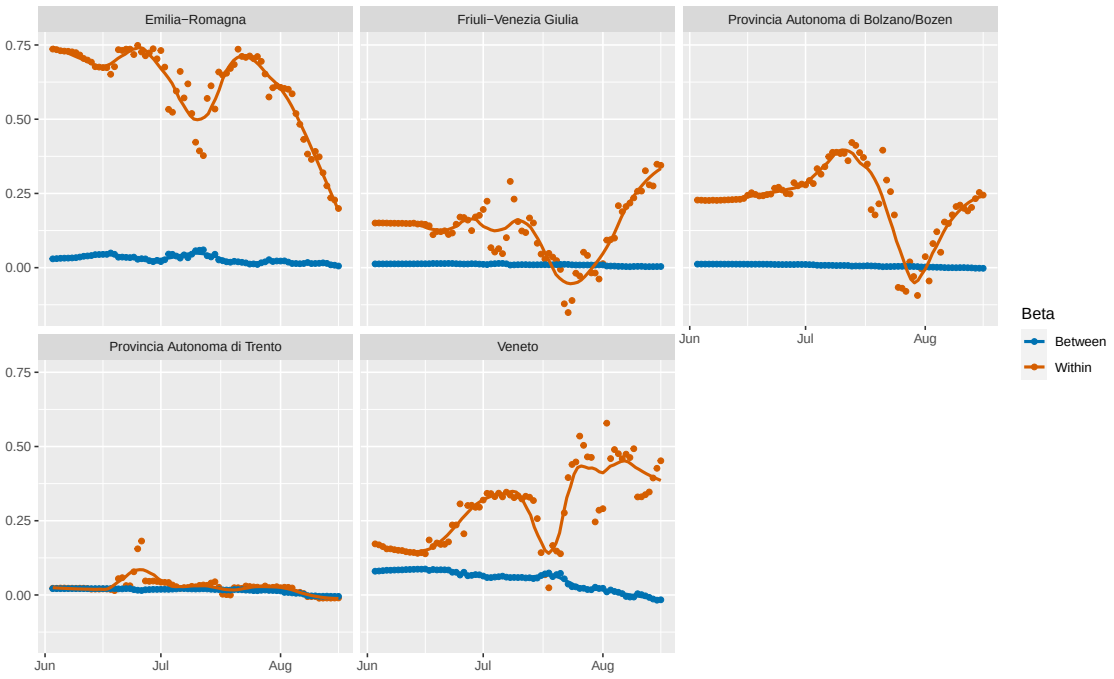
Considering the model where we include undocumented infectives, the estimate for $\beta_{between}$ is statistically significant at a level of 0.01 for nine regions. For the other ten regions, the estimate is not significant. Of these nine regions, three regional models did not find a significant within-region effect, namely for Abruzzo, Emilia-Romagna, and Aosta Valley. Although there are proportionally more regions that sport both a significant within-region and between-region effect, we again see a tendency that a statistically significant between-region effect appears together with a significant between-region effect. Looking only at the statistically significant estimates of $\beta_{between}$, these range from 2.205×10^{-3} for Aosta Valley till 0.330 for Lombardy for the regular model and from 1.762×10^{-3} for Aosta Valley till 0.619 for Lombardy when undocumented infectives are included. Excluding Lombardy, the highest estimates of $\beta_{between}$ are 0.075 and 0.094 for the regular model and the model including undocumented infectives, respectively, both for the region of Piedmont (PIE). These estimates differ much less over the regions.

Regarding the model selection, we again see that the AIC gives a varying model selection per region. Once again, the set of regressors that is used is the same whether undocumented infectives are modelled are not. Most of the estimates for the weekend dummy's parameter are once again not statistically significant at a level of 0.05. When modelling undocumented infectives, none of the estimated weekend dummy parameters are statistically significant. As mentioned, all models retain the intercept and the terms related to β_{within} and $\beta_{between}$ in the model. In seven out of nineteen cases, the entire model is selected. In the other twelve cases, the weekend dummy is excluded.

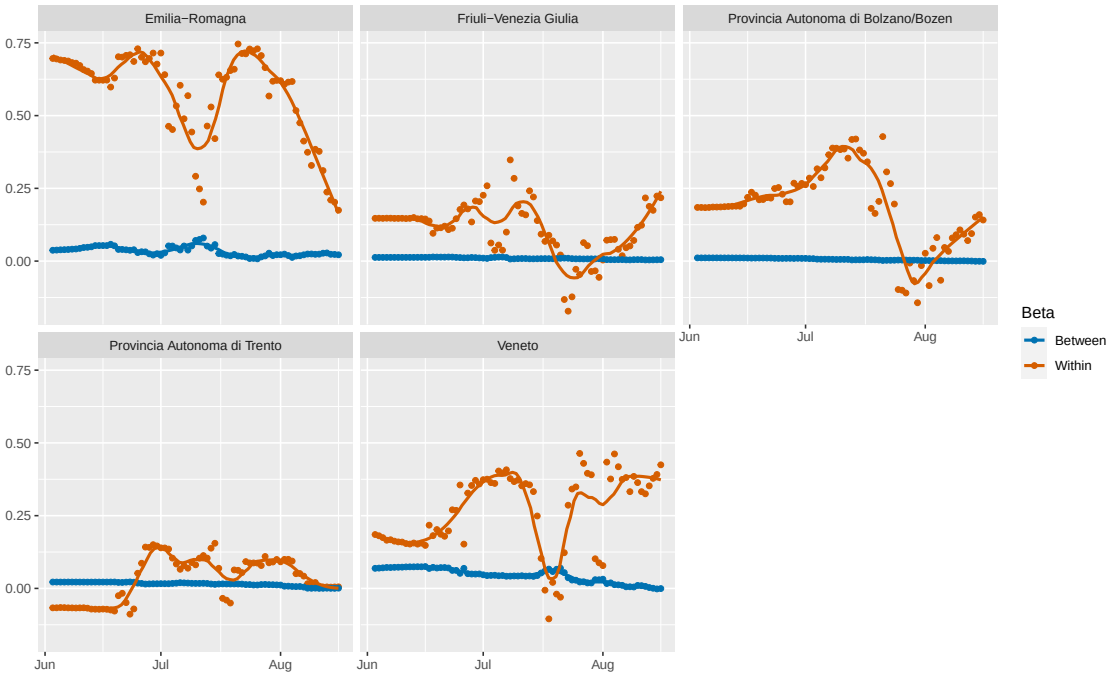
To conclude, we are interested in looking at the estimates of β_{within} and $\beta_{between}$ over time. In Figure 5.2, we present plots for the regions in the *Nord-Est* (North-East) NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.1. Each point in the graphs in Figure 5.2 is the estimate of β_{within} or $\beta_{between}$ when only the latest 100 data points before that date are used. In addition, a LOESS curve with span parameter 0.3 is fit to the data points to give a better insight in the progression over time.



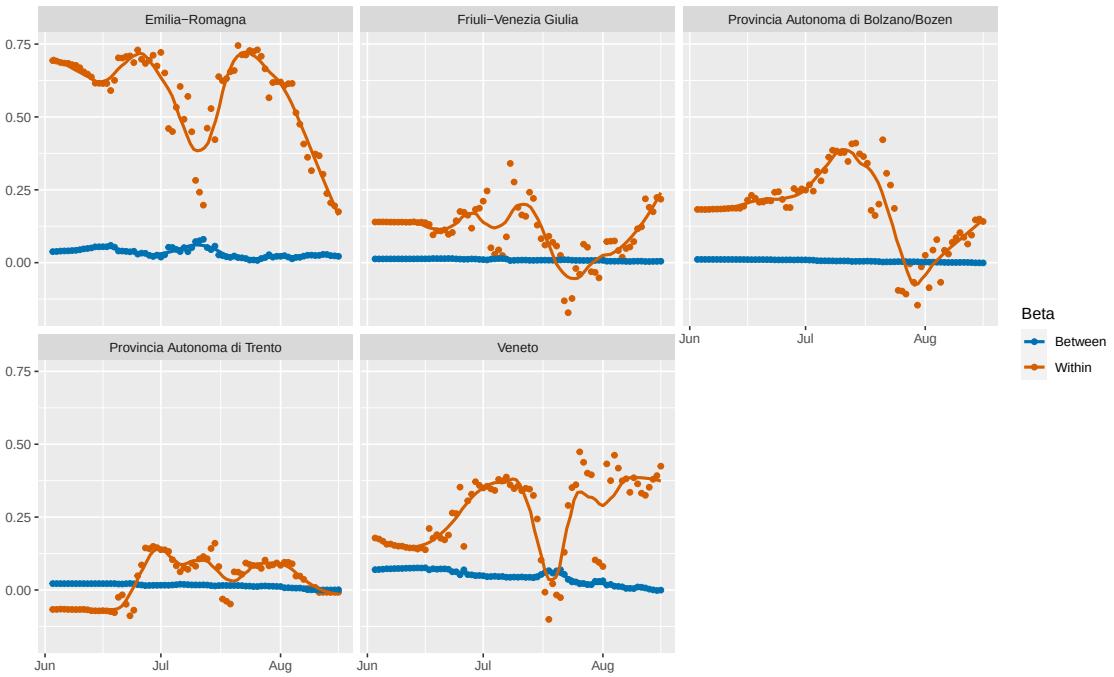
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



(d) With model selection by AIC;
including undocumented infectives

Figure 5.2. Progression of β_{within} and $\beta_{between}$ over time for the *Nord-Est* (North-East) NUTS 1 region

First of all, notice that negative values are found at some point for all regions except for Emilia-Romagna. This also seems to take place around the same time, namely the end of July. Considering the estimated values for $\beta_{between}$, we cannot really distinguish a large difference over time, except for the regions of Emilia-Romagna and Veneto, where we see a clear decrease over time. This is due to the small scale as described earlier in this section. Regarding the estimates of β_{within}, a much more erratic pattern over time is discovered where we do not see a clear decrease over time. As described in Section 5.1, we expect that these values decrease over time to concur with a decreasing amount of infectives over time.

Why?

Plot BZ infectives over time; second wave?

Once again, model selection by AIC does not seem to impact the estimates of β_{within} and $\beta_{between}$ much. However, a distinct pattern is discovered when undocumented infectives are included. This is most pronounced for the regions of Trentino and Veneto. For both regions, the amplitude of the distribution seems to be increased. However, the progression of β_{within} and $\beta_{between}$ seem to follow a similar pattern over time if we disregard the change in magnitude.

5.4 Model 4: Discrete SIR model

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/

BMJ. (2020). *Diagnostic accuracy of serological tests for COVID-19: Systematic review and meta-analysis*. Retrieved July 13, 2020, from https://www.bmj.com/content/370/bmj.m2516/

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

European Centre for Disease Prevention and Control. (2020). *Rapid risk assessment: Coronavirus disease 2019 (COVID-19) pandemic: Increased transmission in the EU/EEA and the UK - seventh update*. Retrieved August 17, 2020, from https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-coronavirus-disease-2019-covid-19-pandemic

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/

Google LLC. (2020). *Google COVID-19 community mobility reports*. https://www.google.com/covid19/mobility/

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

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.

Leung, H. (2020). *What we know about coronavirus immunity and reinfection*. Retrieved June 9, 2020, from https://time.com/5810454/coronavirus-immunity-reinfection/

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.

Worldometer. (2020). *Italy population*. Retrieved August 3, 2020, from https://www.worldometers.info/world-population/italy-population/

Appendices

A Abbreviations

The tables in this appendix present commonly used abbreviations in this thesis, including the regional abbreviations.

Table A.1. Abbreviations for the Italian regions.

Abbreviation	Italian name	English name
ABR	Abruzzo	Abruzzo
BAS	Basilicata	Basilicata
BZ	Alto Adige or Provincia Autonoma di Bolzano/Bozen	South Tyrol or Province of Bolzano
CAL	Calabria	Calabria
CAM	Campania	Campania
EMR	Emilia-Romagna	Emilia-Romagna
FVG	Friuli Venezia Giulia	Friuli Venezia Giulia
LAZ	Lazio	Lazio
LIG	Liguria	Liguria
LOM	Lombardia	Lombardy
MAR	Marche	Marche
MOL	Molise	Molise
PIE	Piemonte	Piedmont
PUG	Puglia	Apulia
SAR	Sardegna	Sardinia
SIC	Sicilia	Sicily
TN	Trentino or Provincia Autonoma di Trento	Trentino or Province of Trento
TOS	Toscana	Tuscany
UMB	Umbria	Umbria
VDA	Valle d'Aosta/Vallée d'Aoste	Aosta Valley
VEN	Veneto	Veneto

Table A.2. Commonly used abbreviations in this thesis.

Abbreviation	Full name	Defined in...
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2	Section 1
COVID-19	Coronavirus Disease 2019	Section 1
SIR model	Standard Inflammatory Response model	Section 3.1
OLS	Ordinary Least Squares	Section 3.2
AIC	Akaike Information Criterion	Section 3.6
BIC	Bayesian Information Criterion	Section 3.6

Table A.2 continues on next page

Table A.2 continued from previous page

Abbreviation	Full name	Defined in...
NUTS	Nomenclature des Unités Territoriales Statistiques	Section 4.1

B Tables

B.1 Results from Within-Region Spread Model

In Section 5.1, we presented the results from the within-region spread model (3.8):

$$\Delta Y_{r,t} = \beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \delta M_{r,t} + \eta_{r,t}.$$

This appendix contains additional tables with results for this model. As is the case for Section 5.1, we use the last 100 observations.

Table B.1. Estimates from Within-Region Spread Model per region without model selection. Estimates are given with t-statistics in parentheses. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Regular model			Modelling undocumented infectives		
Region	β_{within}	Intercept	Weekend	β_{within}	Intercept	Weekend
National	0.691***	110.400***	-38.430	0.765***	333.200***	-164.400
	(10.690)	(3.948)	(-1.175)	(15.650)	(2.999)	(-1.106)
ABR	0.254**	3.366***	0.412	0.317***	14.170***	3.067
	(2.545)	(3.961)	(0.307)	(3.367)	(3.549)	(0.480)
BAS	0.043	1.157**	-0.621	0.044	5.404**	-2.569
	(0.417)	(2.222)	(-0.644)	(0.427)	(2.249)	(-0.579)
BZ	0.239**	1.525***	1.048*	0.126	5.258***	2.862
	(2.315)	(3.675)	(1.671)	(1.297)	(4.418)	(1.616)
CAL	0.151	1.675***	0.818	0.114	9.585***	4.851
	(1.305)	(3.139)	(0.900)	(1.009)	(3.293)	(0.985)
EMR	0.231***	25.840***	7.747**	0.379***	80.870***	24.090
	(2.701)	(6.917)	(2.138)	(5.693)	(6.158)	(1.618)
FVG	0.483***	1.974***	0.167	0.579***	5.189***	-0.098
	(4.958)	(3.799)	(0.223)	(7.134)	(3.178)	(-0.041)
LAZ	0.555***	8.063***	4.285*	0.540***	45.450***	21.030
	(5.506)	(3.441)	(1.725)	(6.129)	(3.693)	(1.482)
LIG	0.664***	6.655***	-4.158	0.743***	23.800**	-19.840
	(8.784)	(3.332)	(-1.395)	(11.560)	(2.579)	(-1.391)
LOM	0.609***	65.710***	-25.620	0.647***	265.600***	-123.400
	(7.753)	(3.490)	(-1.003)	(8.819)	(2.842)	(-0.903)
MAR	0.439***	2.637***	0.385	0.478***	9.938***	3.369

Table B.1 continues on next page

Table B.1 continued from previous page

	Regular model			Modelling undocumented infectives		
Region	β_{within}	Intercept	Weekend	β_{within}	Intercept	Weekend
	(5.226)	(3.108)	(0.286)	(6.829)	(2.715)	(0.569)
MOL	0.296***	0.686***	-0.328	0.311***	3.341**	-1.706
	(5.587)	(2.652)	(-0.725)	(6.866)	(2.267)	(-0.655)
PIE	0.715***	9.988***	-7.690	0.725***	44.490***	-44.350*
	(12.510)	(2.951)	(-1.506)	(13.980)	(2.650)	(-1.681)
PUG	0.506***	2.515***	0.278	0.454***	16.050***	1.886
	(6.299)	(3.033)	(0.215)	(6.321)	(3.353)	(0.253)
SAR	0.217**	1.309***	-0.327	0.191*	6.996***	-1.404
	(2.148)	(3.258)	(-0.473)	(1.943)	(3.400)	(-0.403)
TN	4.549×10^{-3}	9.117*	-6.484	5.350×10^{-3}	23.880*	-16.340
	(0.043)	(1.911)	(-0.720)	(0.050)	(1.961)	(-0.712)
TOS	0.588***	3.444**	4.245**	0.562***	15.770***	17.460**
	(6.659)	(2.513)	(2.231)	(6.908)	(2.635)	(2.070)
UMB	0.492***	0.635**	0.590	0.360***	2.979***	1.689
	(4.303)	(2.117)	(1.200)	(3.233)	(2.668)	(0.932)
VDA	0.058	0.622***	-0.182	0.158	1.947***	-0.558
	(0.573)	(3.346)	(-0.576)	(1.659)	(3.070)	(-0.519)
VEN	0.435***	15.420***	-0.212	0.421***	39.660***	-1.910
	(4.010)	(3.556)	(-0.032)	(4.274)	(3.860)	(-0.126)

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

Table B.2. Estimates from Within-Region Spread Model per region with model selection by AIC versus BIC. Estimates are given with t-statistics in parentheses. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Model selection with AIC			Model selection with BIC		
Region	β_{within}	Intercept	Weekend	β_{within}	Intercept	Weekend
National	0.763***	288.400***		0.763***	288.400***	
	(15.600)	(2.785)		(15.600)	(2.785)	
ABR	0.318***	15.030***		0.318***	15.030***	
	(3.394)	(4.211)		(3.394)	(4.211)	
BAS	0.038	4.691**		0.038	4.691**	
	(0.374)	(2.282)		(0.374)	(2.282)	
BZ	0.126	5.258***	2.862	0.116	6.159***	
	(1.297)	(4.418)	(1.616)	(1.179)	(5.811)	
CAL	0.117	10.950***		0.117	10.950***	
	(1.041)	(4.272)		(1.041)	(4.272)	
EMR	0.379***	80.870***	24.090	0.377***	88.030***	
	(5.693)	(6.158)	(1.618)	(5.626)	(7.060)	
FVG	0.579***	5.163***		0.579***	5.163***	
	(7.180)	(3.451)		(7.180)	(3.451)	

Table B.2 continues on next page

Table B.2 continued from previous page

	Model selection with AIC			Model selection with BIC		
Region	β_{within}	Intercept	Weekend	β_{within}	Intercept	Weekend
LAZ	0.540***	45.450***	21.030	0.540***	51.570***	
	(6.129)	(3.693)	(1.482)	(6.086)	(4.420)	
LIG	0.737***	18.540**		0.737***	18.540**	
	(11.440)	(2.191)		(11.440)	(2.191)	
LOM	0.647***	230.400***		0.647***	230.400***	
	(8.820)	(2.715)		(8.820)	(2.715)	
MAR	0.477***	10.930***		0.477***	10.930***	
	(6.843)	(3.411)		(6.843)	(3.411)	
MOL	0.312***	2.840**		0.312***	2.840**	
	(6.911)	(2.261)		(6.911)	(2.261)	
PIE	0.725***	44.490***	-44.350*	0.719***	32.630**	
	(13.980)	(2.650)	(-1.681)	(13.770)	(2.122)	
PUG	0.453***	16.610***		0.453***	16.610***	
	(6.348)	(3.926)		(6.348)	(3.926)	
SAR	0.191*	6.586***		0.191*	6.586***	
	(1.958)	(3.700)		(1.958)	(3.700)	
TN	-7.826×10^{-3}	19.420*		-7.826×10^{-3}	19.420*	
	(-0.075)	(1.865)		(-0.075)	(1.865)	
TOS	0.562***	15.770***	17.460**	0.561***	20.860***	
	(6.908)	(2.635)	(2.070)	(6.779)	(3.759)	
UMB	0.355***	3.491***		0.355***	3.491***	
	(3.192)	(3.595)		(3.192)	(3.595)	
VDA	0.163*	1.772***		0.163*	1.772***	
	(1.730)	(3.307)		(1.730)	(3.307)	
VEN	0.421***	39.120***		0.421***	39.120***	
	(4.295)	(4.214)		(4.295)	(4.214)	

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

B.2 Results from Weighted Within-Region Spread Model

In Section 5.2, we presented the results from the within and between-region spread model (3.9):

$$\Delta Y_{r,t} = \Delta Y_{r,t-\tau} S_{r,t-\tau} \sum_{k=1}^{K} \beta_{within}^{k} W_{r,t-\tau}^{k} + \delta M_{r,t} + \eta_{r,t}$$

This appendix contains additional tables with results for this model. As is the case for Section 5.2, we use the last 100 observations.

TODO: Results still need to be inserted.

B.3 Results from Within and Between-Region Spread Model

In Section 5.3, we presented the results from the within and between-region spread model (3.10):

$$\Delta Y_{r,t} = \beta_{within} \Delta Y_{r,t-\tau} S_{r,t-\tau} + \beta_{between} S_{r,t-\tau} \sum_{c \in R \setminus r} \Delta Y_{c,t-\tau} + \delta M_{r,t} + \eta_{r,t}$$

This appendix contains additional tables with results for this model. As is the case for Section 5.3, we use the last 100 observations.

Table B.3. Estimates from Within and Between-Region Spread Model per region without model selection. Estimates are given with t-statistics in parentheses. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Regular model				Modelling undocumented infectives			
Region	β_{within}	$\beta_{between}$	Intercept	Weekend	β_{within}	$\beta_{between}$	Intercept	Weekend
ABR	0.1983*	3.719×10^{-3}	2.375**	0.298	0.174	6.358×10^{-3}***	7.459	2.716
	(1.820)	(1.272)	(2.065)	(0.222)	(1.655)	(2.731)	(1.629)	(0.446)
BAS	0.045	-1.202×10^{-3}	1.563*	-0.5749	0.04512	-6.251×10^{-4}	6.392*	-2.516
	(0.433)	(-0.633)	(1.890)	(-0.593)	(0.434)	(-0.431)	(1.920)	(-0.564)
BZ	0.245**	-2.117×10^{-3}*	2.219***	1.136*	0.141	-8.112×10^{-4}	6.422***	2.958*
	(2.396)	(-1.717)	(3.851)	(1.823)	(1.448)	(-1.391)	(4.429)	(1.677)
CAL	0.158	-1.843×10^{-3}	2.284***	0.891	0.121	-1.117×10^{-3}	11.270***	4.947
	(1.366)	(-1.026)	(2.863)	(0.976)	(1.063)	(-0.689)	(2.957)	(1.001)
EMR	0.199**	5.569×10^{-3}	25.370***	7.532**	0.175*	0.022***	81.970***	21.410
	(2.000)	(0.635)	(6.640)	(2.063)	(1.861)	(2.957)	(6.491)	(1.493)
FVG	0.343***	3.878×10^{-3}**	1.155*	0.147	0.218**	4.860×10^{-3}***	2.053	-0.011
	(3.001)	(2.233)	(1.840)	(0.201)	(2.058)	(4.732)	(1.270)	(-5.322×10^{-3})
LAZ	0.389***	0.017***	5.880**	3.681	0.257**	0.025***	39.650***	18.490
	(3.499)	(3.017)	(2.489)	(1.539)	(2.508)	(4.469)	(3.508)	(1.427)
LIG	0.370***	0.039***	-1.144	-4.453*	0.365***	0.041***	-6.395	-17.860
	(4.450)	(5.708)	(-0.519)	(-1.727)	(4.855)	(6.982)	(-0.738)	(-1.538)
LOM	0.487***	0.330***	29.430	-34.720	0.324***	0.619***	36.300	-167.400
	(5.753)	(3.128)	(1.374)	(-1.411)	(3.732)	(5.503)	(0.396)	(-1.400)
MAR	0.317***	6.010×10^{-3}*	1.285	0.147	0.292***	7.385×10^{-3}***	3.262	2.309
	(2.873)	(1.678)	(1.104)	(0.110)	(3.022)	(2.691)	(0.754)	(0.402)
MOL	0.259***	1.230×10^{-3}	0.332	-0.381	0.261***	1.657×10^{-3}	1.242	-1.957
	(4.226)	(1.178)	(0.837)	(-0.839)	(4.760)	(1.591)	(0.631)	(-0.757)
PIE	0.417***	0.075***	-2.265	-8.154*	0.301***	0.094***	-13.900	-37.560
	(4.387)	(3.799)	(-0.501)	(-1.707)	(3.004)	(4.775)	(-0.715)	(-1.577)
PUG	0.470***	2.616×10^{-3}	1.849*	0.176	0.397***	3.556×10^{-3}	12.590**	1.427
	(5.223)	(0.905)	(1.665)	(0.136)	(4.638)	(1.199)	(2.256)	(0.192)
SAR	0.224**	-1.144×10^{-3}	1.683***	-0.281	0.200**	-9.694×10^{-4}	8.445***	-1.305
	(2.212)	(-0.834)	(2.790)	(-0.405)	(2.024)	(-0.842)	(3.145)	(-0.374)
TN	3.427×10^{-3}	-3.951×10^{-3}	10.430	-6.369	5.389×10^{-3}	8.565×10^{-4}	22.550	-16.390

Table B.3 continues on next page

Table B.3 continued from previous page

	Regular model				Modelling undocumented infectives			
Region	β_{within}	$\beta_{between}$	Intercept	Weekend	β_{within}	$\beta_{between}$	Intercept	Weekend
	(0.032)	(-0.221)	(1.365)	(-0.703)	(0.050)	(0.114)	(1.328)	(-0.710)
TOS	0.508***	7.268×10^{-3}*	1.887	3.970**	0.382***	0.010***	8.836	16.370**
	(5.103)	(1.677)	(1.147)	(2.098)	(3.842)	(2.944)	(1.421)	(2.016)
UMB	0.504***	-4.46×10^{-4}	0.771*	0.610	0.357***	5.832×10^{-5}	2.903**	1.681
	(4.274)	(-0.445)	(1.797)	(1.230)	(3.035)	(0.093)	(2.093)	(0.922)
VDA	-0.057	2.283×10^{-3}***	-0.050	-0.314	-0.053	1.793×10^{-3}***	-0.261	-0.983
	(-0.569)	(3.649)	(-0.198)	(-1.046)	(-0.559)	(5.094)	(-0.367)	(-1.024)
VEN	0.452***	-0.016	20.180***	0.367	0.425***	-5.265×10^{-4}	40.210***	-1.869
	(4.140)	(-1.198)	(3.437)	(0.055)	(3.976)	(-0.095)	(3.389)	(-0.122)

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

Table B.4. Estimates from Within and Between-Region Spread Model per region with model selection by AIC versus BIC. Estimates are given with t-statistics in parentheses. Data spans May 9 till August 16, 2020 (100 days). Undocumented infectives are not modelled.

	Model selection with AIC				Model selection with BIC			
Region	β_{within}	$\beta_{between}$	Intercept	Weekend	β_{within}	$\beta_{between}$	Intercept	Weekend
ABR	0.174*	6.38×10^{-3}***	8.196*		0.174*	6.380×10^{-3}***	8.196*	
	(1.668)	(2.753)	(1.927)		(1.668)	(2.753)	(1.927)	
BAS	0.040	-6.479×10^{-4}	5.730*		0.040	-6.479×10^{-4}	5.730*	
	(0.383)	(-0.448)	(1.846)		(0.383)	(-0.448)	(1.846)	
BZ	0.141	-8.112×10^{-4}	6.422***	2.958*	0.129	-7.727×10^{-4}	7.297***	
	(1.448)	(-1.391)	(4.429)	(1.677)	(1.318)	(-1.314)	(5.343)	
CAL	0.124	-1.071×10^{-3}	12.590***		0.124	-1.071×10^{-3}	12.590***	
	(1.093)	(-0.661)	(3.518)		(1.093)	(-0.661)	(3.518)	
EMR	0.175*	0.022***	81.970***	21.410	0.167*	0.022***	88.350***	
	(1.861)	(2.957)	(6.491)	(1.493)	(1.771)	(3.038)	(7.385)	
FVG	0.218**	4.860×10^{-3}***	2.050		0.218**	4.860×10^{-3}***	2.050	
	(2.071)	(4.757)	(1.366)		(2.071)	(4.757)	(1.366)	
LAZ	0.257**	0.025***	39.650***	18.490	0.252**	0.025***	44.930***	

Table B.4 continues on next page

Table B.4 continued from previous page

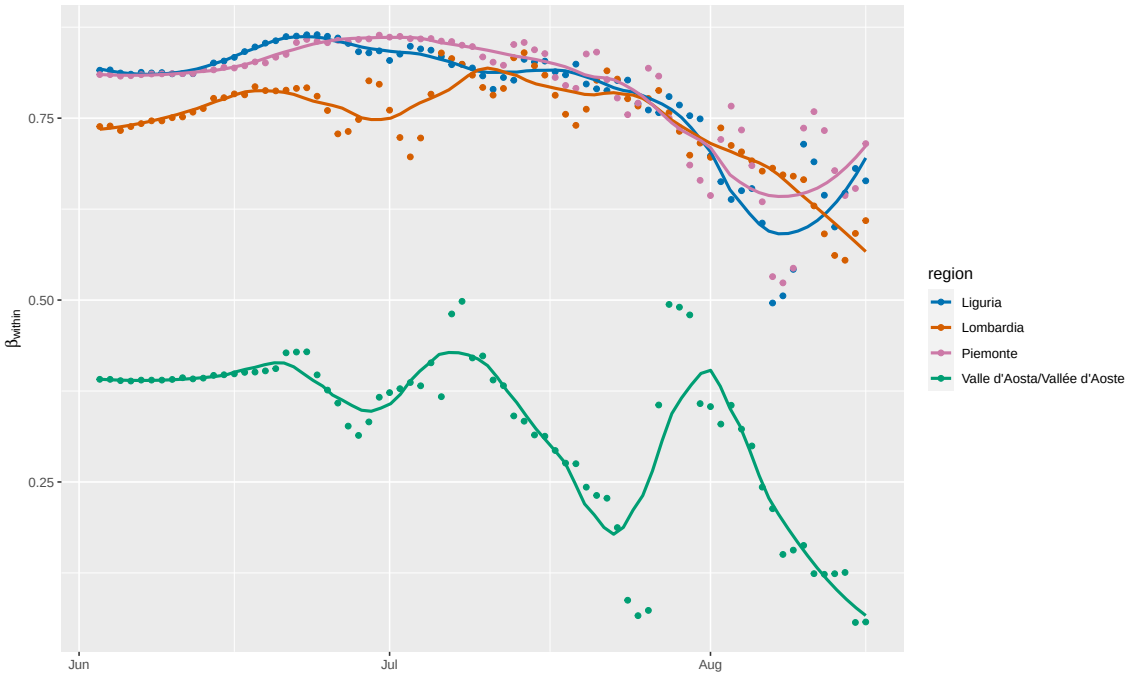
	Model selection with AIC				Model selection with BIC			
Region	β_{within}	$\beta_{between}$	Intercept	Weekend	β_{within}	$\beta_{between}$	Intercept	Weekend
	(2.508)	(4.469)	(3.508)	(1.427)	(2.454)	(4.511)	(4.185)	
LIG	0.365***	0.041***	-6.395	-17.860	0.358***	0.041***	-11.280	
	(4.855)	(6.982)	(-0.738)	(-1.538)	(4.736)	(6.971)	(-1.389)	
LOM	0.324***	0.619***	36.300	-167.400	0.329***	0.608***	-7.339	
	(3.732)	(5.503)	(0.396)	(-1.400)	(3.771)	(5.394)	(-0.085)	
MAR	0.290***	7.460×10^{-3}***	3.872		0.290***	7.46×10^{-3}***	3.872	
	(3.016)	(2.738)	(0.960)		(3.016)	(2.738)	(0.960)	
MOL	0.264***	1.609×10^{-3}	0.729		0.264***	1.609×10^{-3}	0.729	
	(4.826)	(1.552)	(0.395)		(4.826)	(1.552)	(0.395)	
PIE	0.301***	0.094***	-13.900	-37.560	0.288***	0.096***	-25.060	
	(3.004)	(4.775)	(-0.715)	(-1.577)	(2.864)	(4.840)	(-1.372)	
PUG	0.397***	3.585×10^{-3}	12.980**		0.397***	3.585×10^{-3}	12.980**	
	(4.658)	(1.217)	(2.513)		(4.658)	(1.217)	(2.513)	
SAR	0.201**	-9.838×10^{-4}	8.086***		0.201**	-9.838×10^{-4}	8.086***	
	(2.041)	(-0.859)	(3.241)		(2.041)	(-0.859)	(3.241)	
TN	-7.828×10^{-3}	7.548×10^{-4}	18.230		-7.828×10^{-3}	7.548×10^{-4}	18.230	
	(-0.075)	(0.100)	(1.153)		(-0.075)	(0.100)	(1.153)	
TOS	0.382***	0.010***	8.836	16.370**	0.375***	0.010***	13.380**	
	(3.842)	(2.944)	(1.421)	(2.016)	(3.717)	(2.991)	(2.272)	
UMB	0.350***	8.692×10^{-5}	3.375**		0.350***	8.692×10^{-5}	3.375**	
	(2.985)	(0.139)	(2.619)		(2.985)	(0.139)	(2.619)	
VDA	-0.040	1.762×10^{-3}***	-0.526		-0.040	1.762×10^{-3}***	-0.526	
	(-0.426)	(5.023)	(-0.795)		(-0.426)	(5.023)	(-0.795)	
VEN	0.425***	-5.454×10^{-4}	39.700***		0.425***	-5.454×10^{-4}	39.700***	
	(3.997)	(-0.099)	(3.591)		(3.997)	(-0.099)	(3.591)	

Significance levels: * = 0.1 ** = 0.05, *** = 0.01

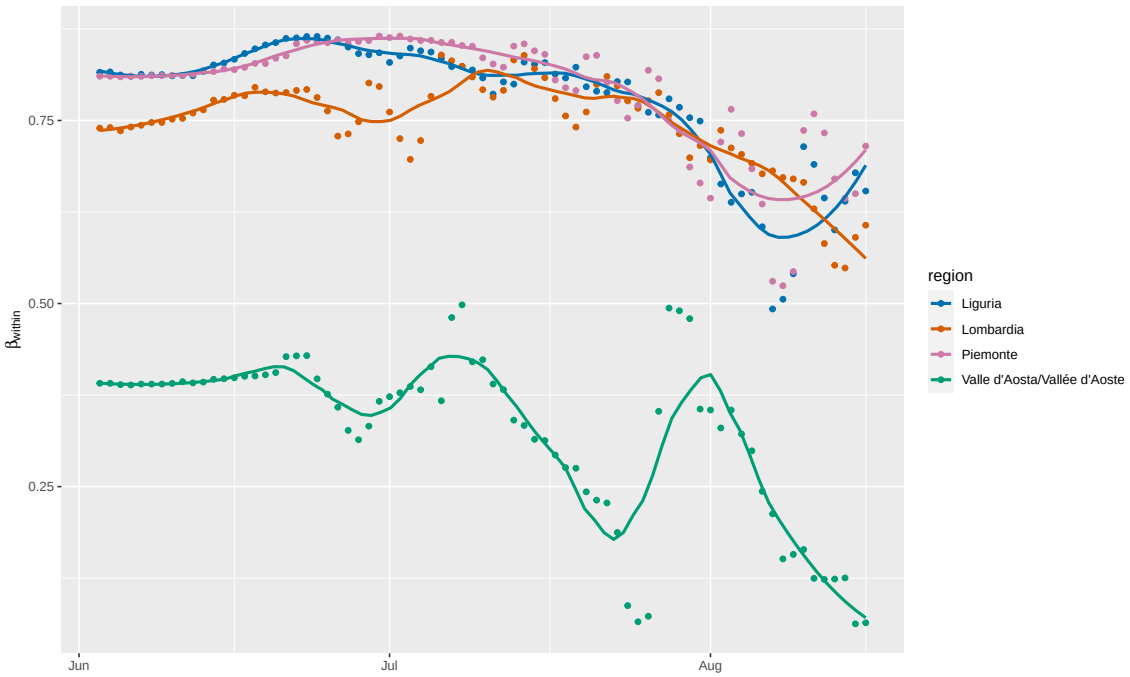
C Figures

C.1 Plots of β_{within} over time

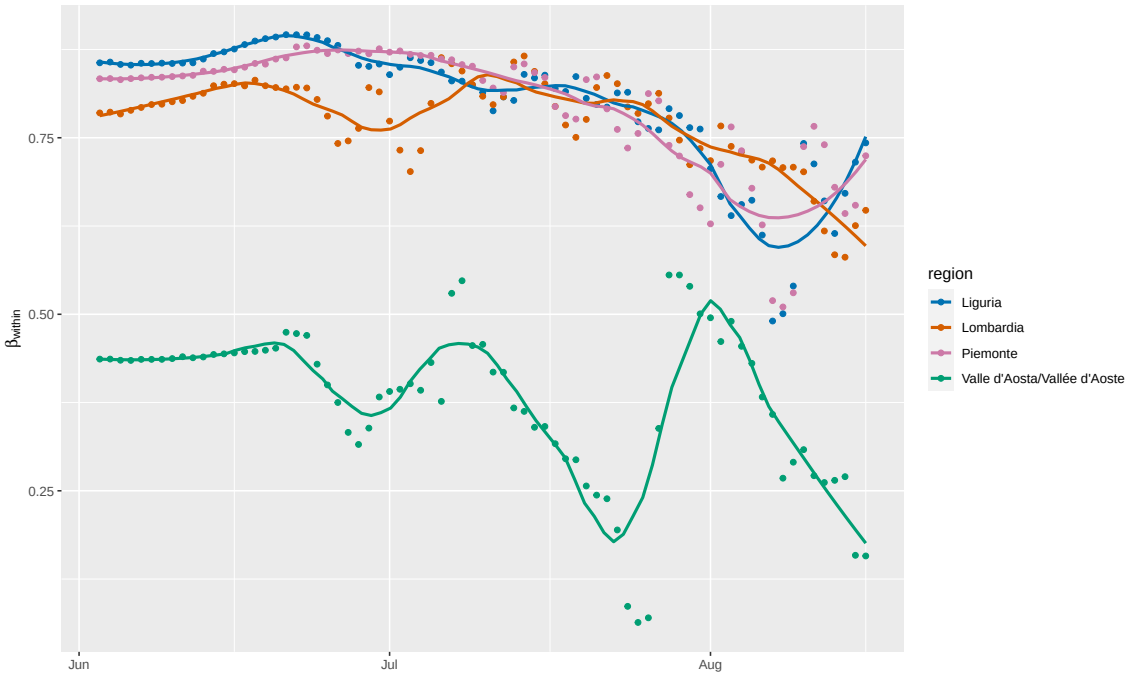
In Section 5.1 we presented the plots of β_{within} over time for the *Nord-Est* (North-East) NUTS 1 region for Within-Region Spread Model. In this section, we present the plots for the other NUTS 1 regions. As is the case for Section 5.1, we use the last 100 observations.



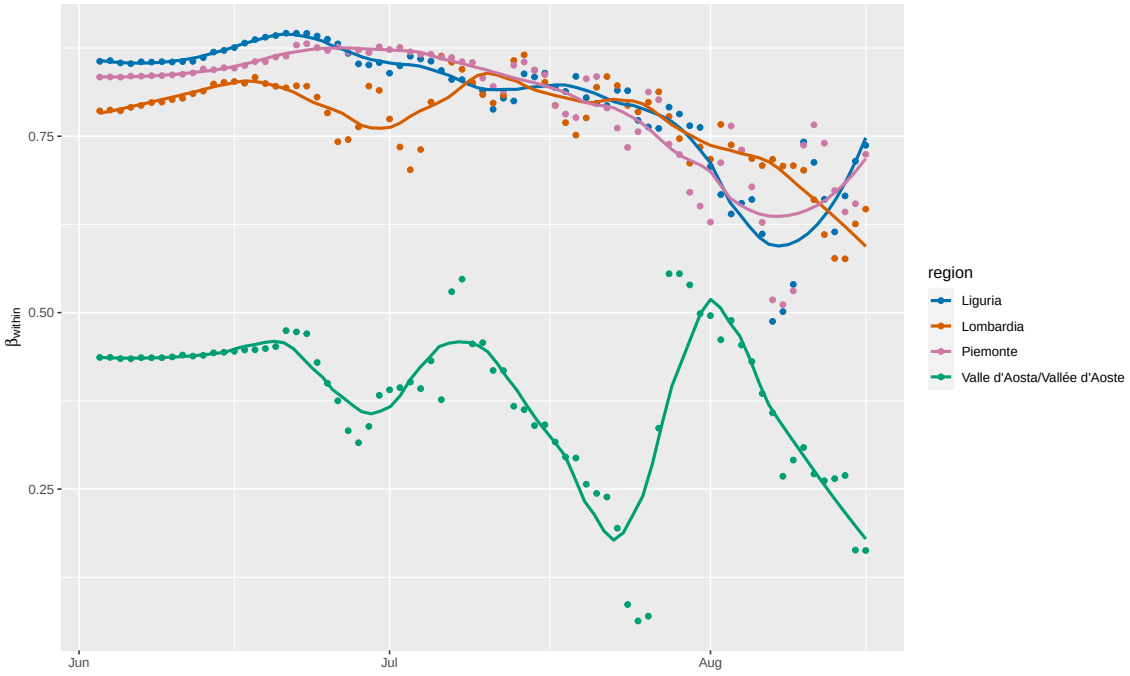
(a) Without model selection



(b) With model selection by AIC

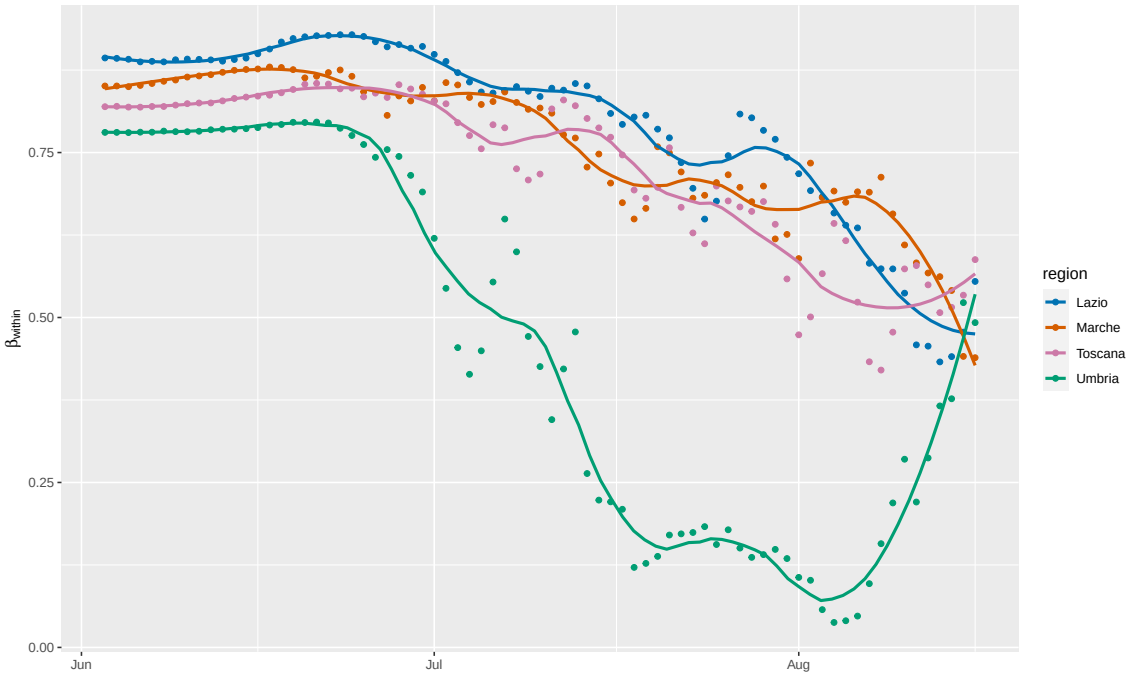


(c) Without model selection;
including undocumented infectives

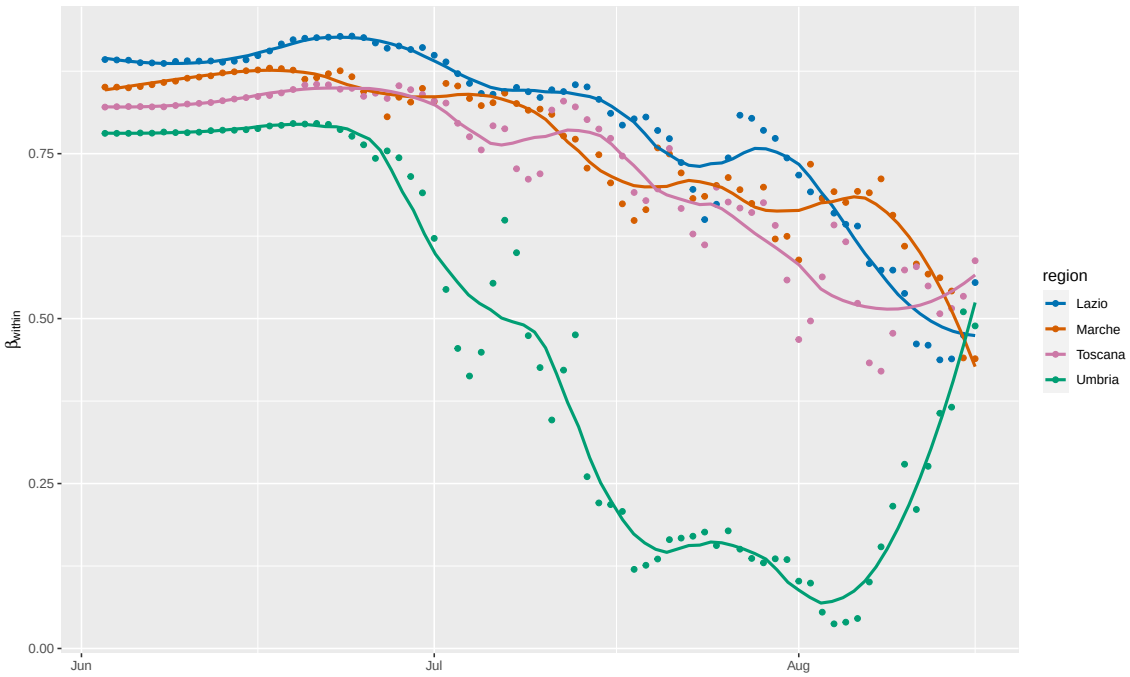


(d) With model selection by AIC;
including undocumented infectives

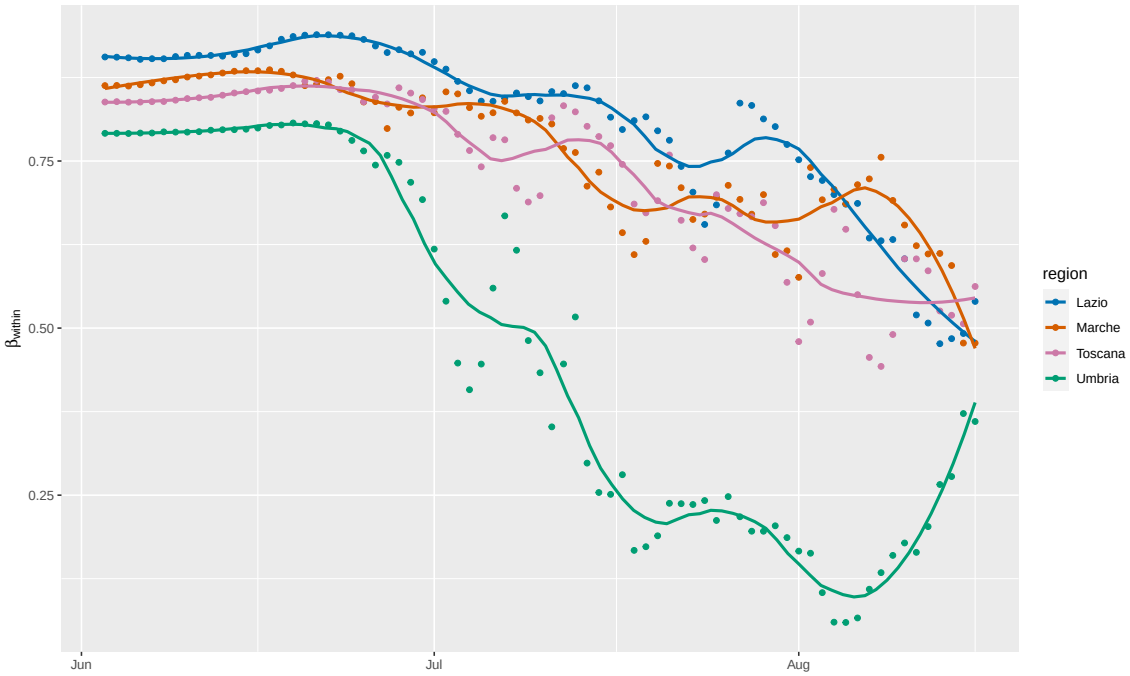
Figure C.1. Progression of β_{within} over time for the *Nord-Ovest* (North-West) NUTS 1 region



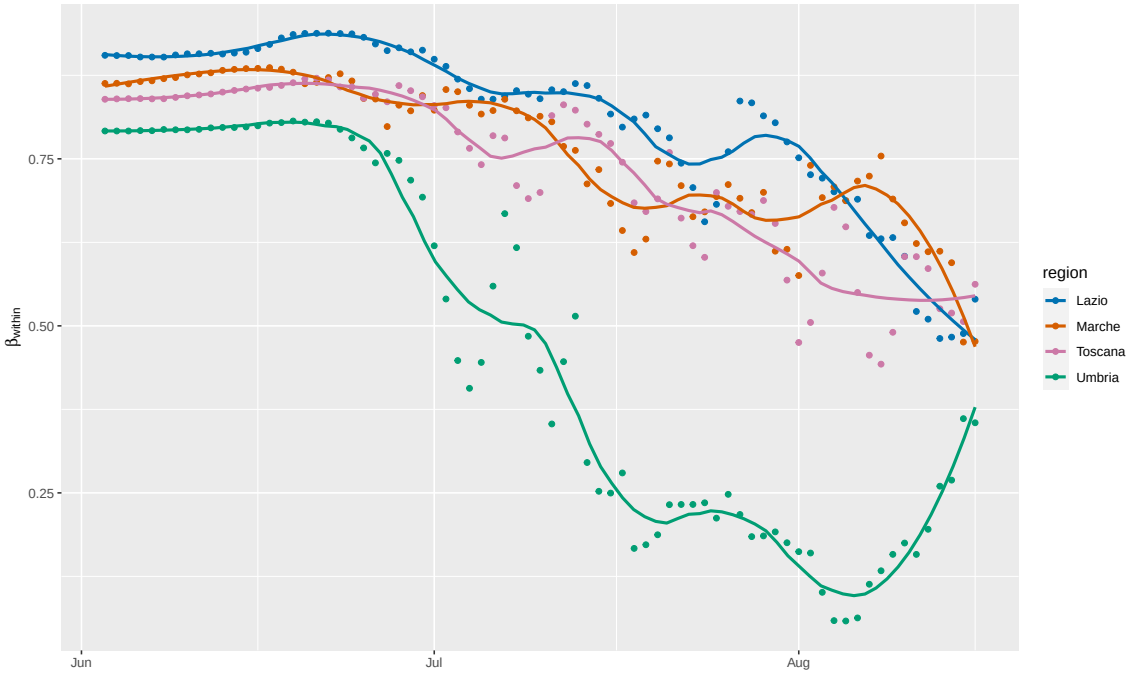
(a) Without model selection



(b) With model selection by AIC

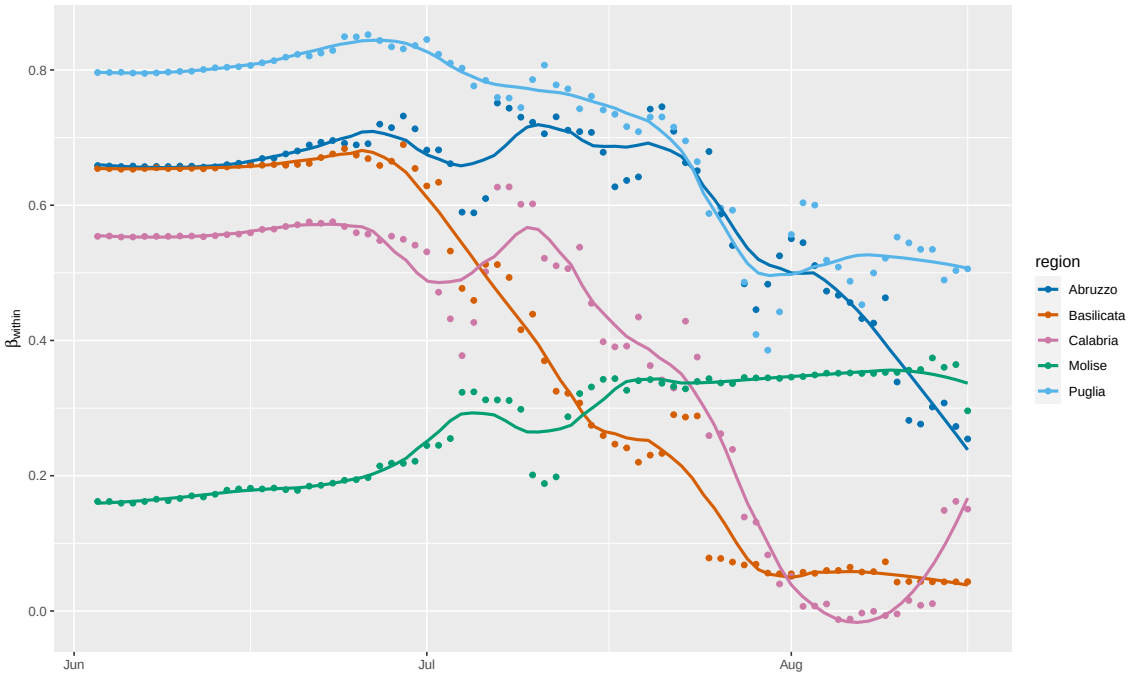


(c) Without model selection;
including undocumented infectives

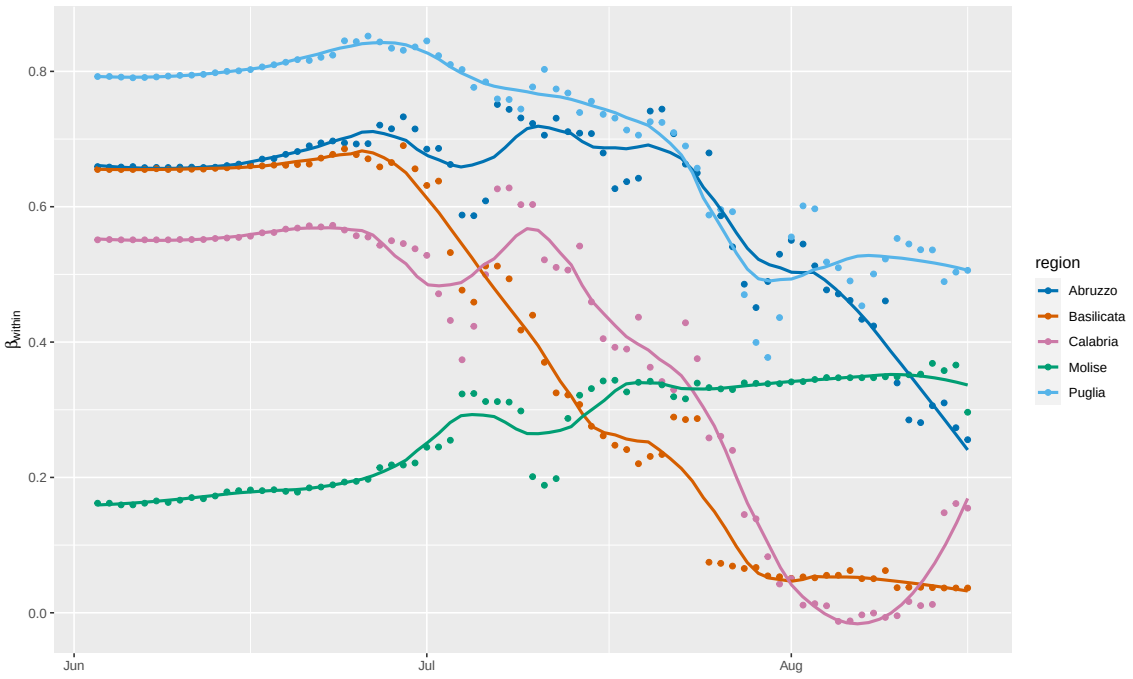


(d) With model selection by AIC;
including undocumented infectives

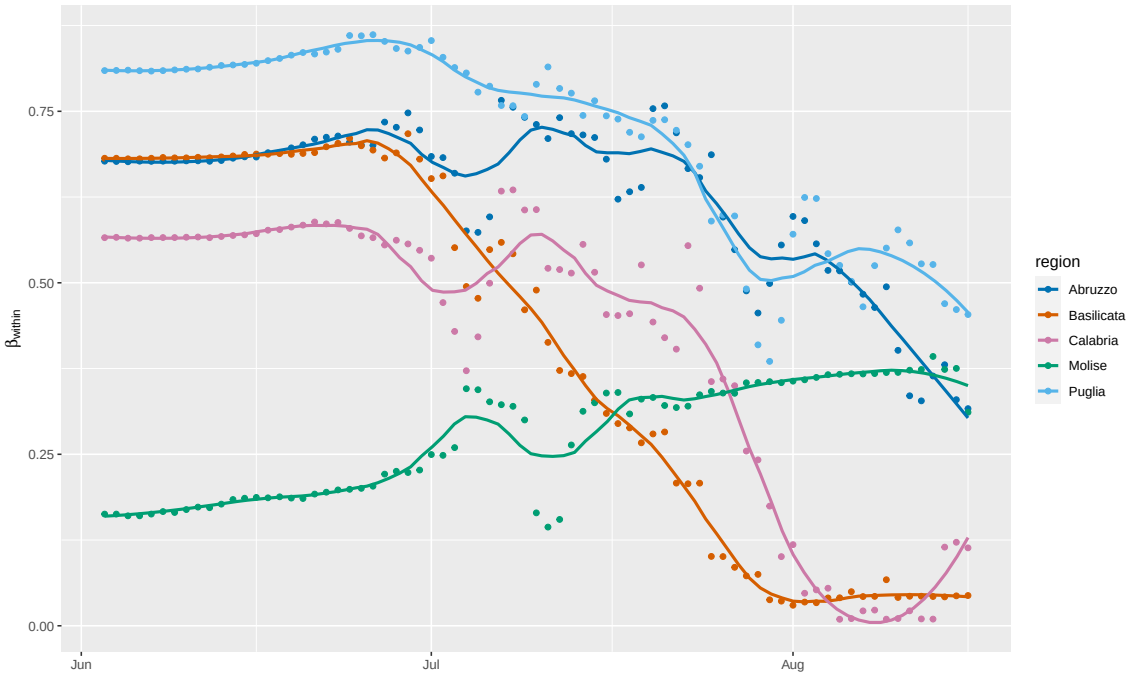
Figure C.2. Progression of β_{within} over time for the *Centro (IT)* (Centre) NUTS 1 region



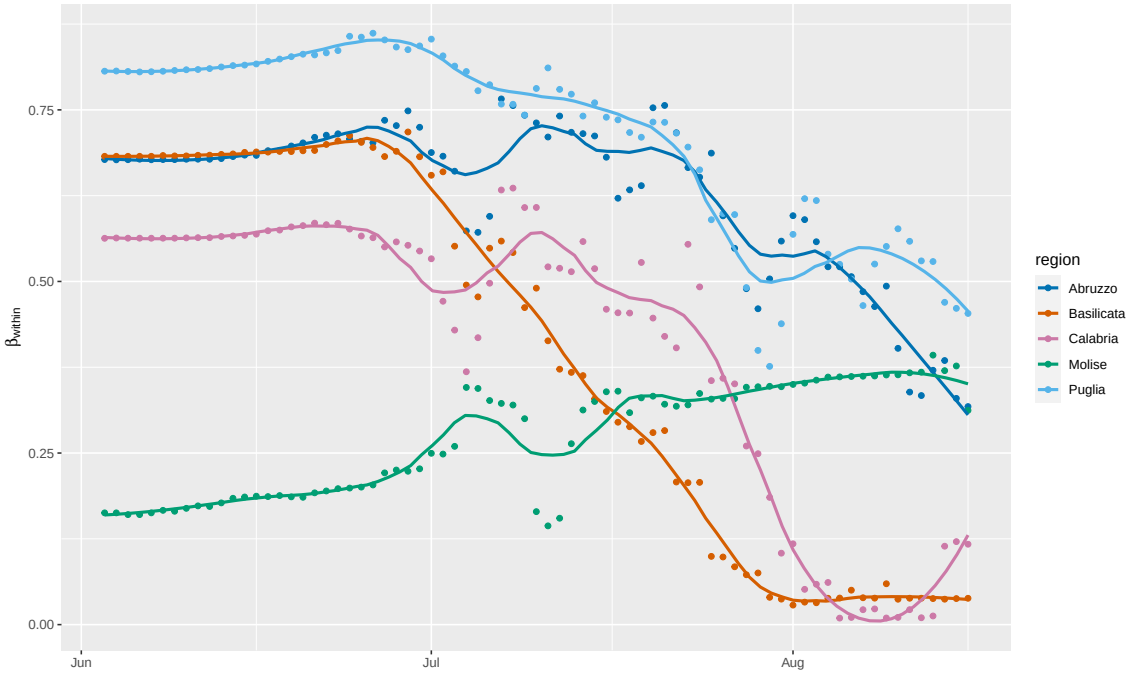
(a) Without model selection



(b) With model selection by AIC

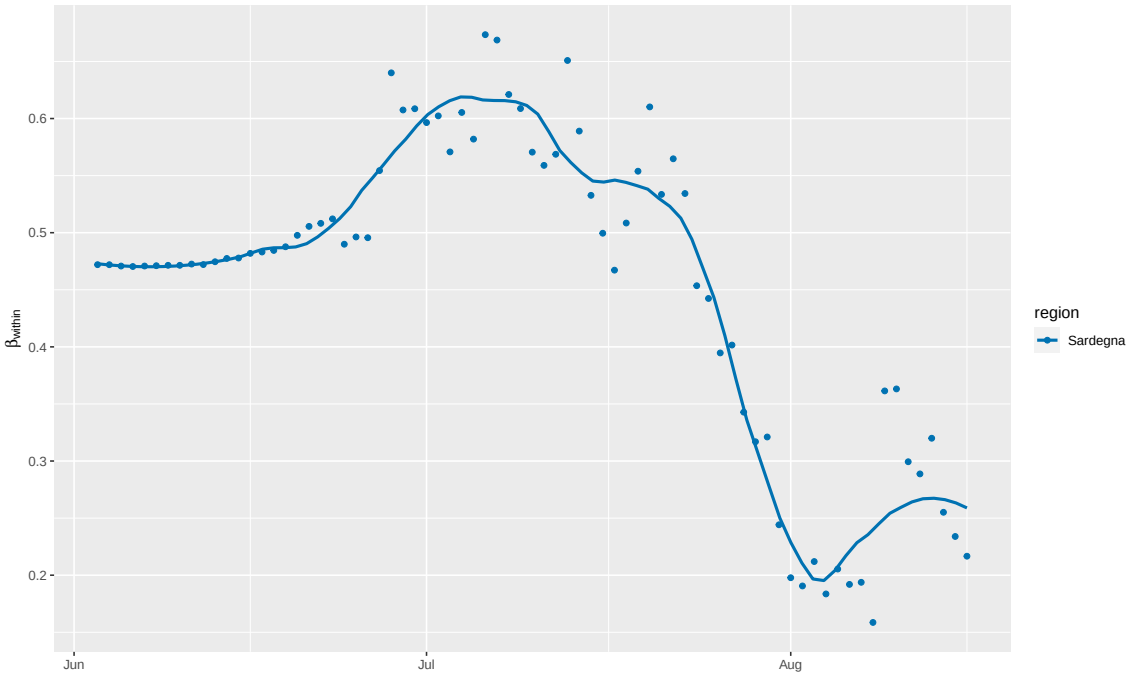


(c) Without model selection;
including undocumented infectives

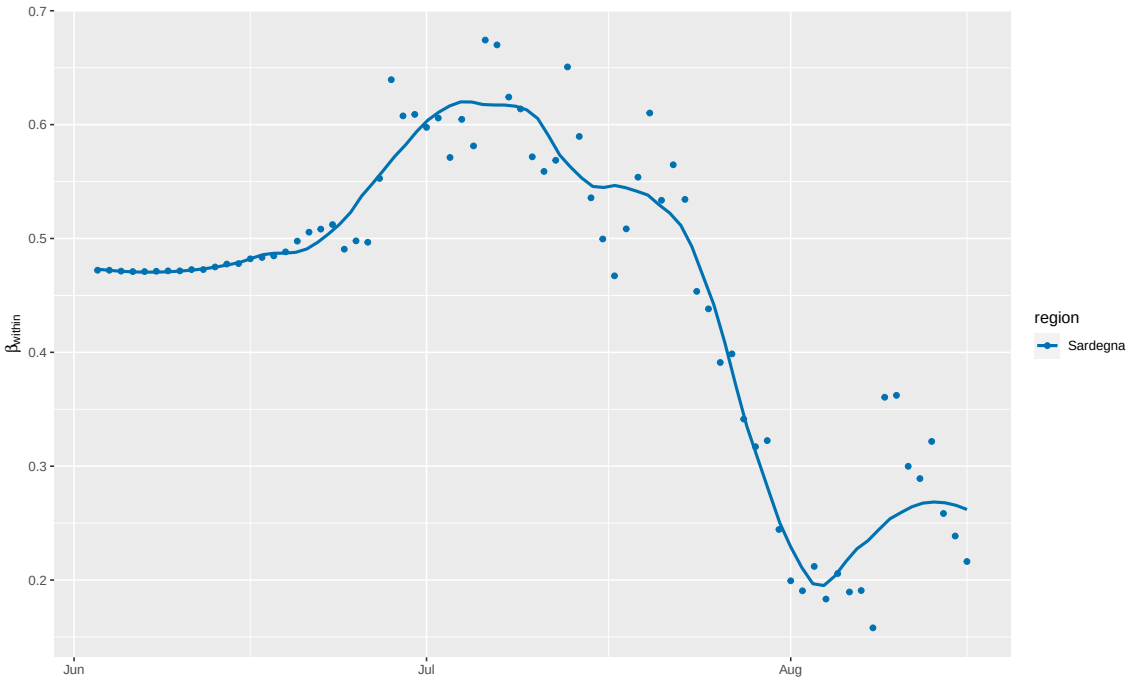


(d) With model selection by AIC;
including undocumented infectives

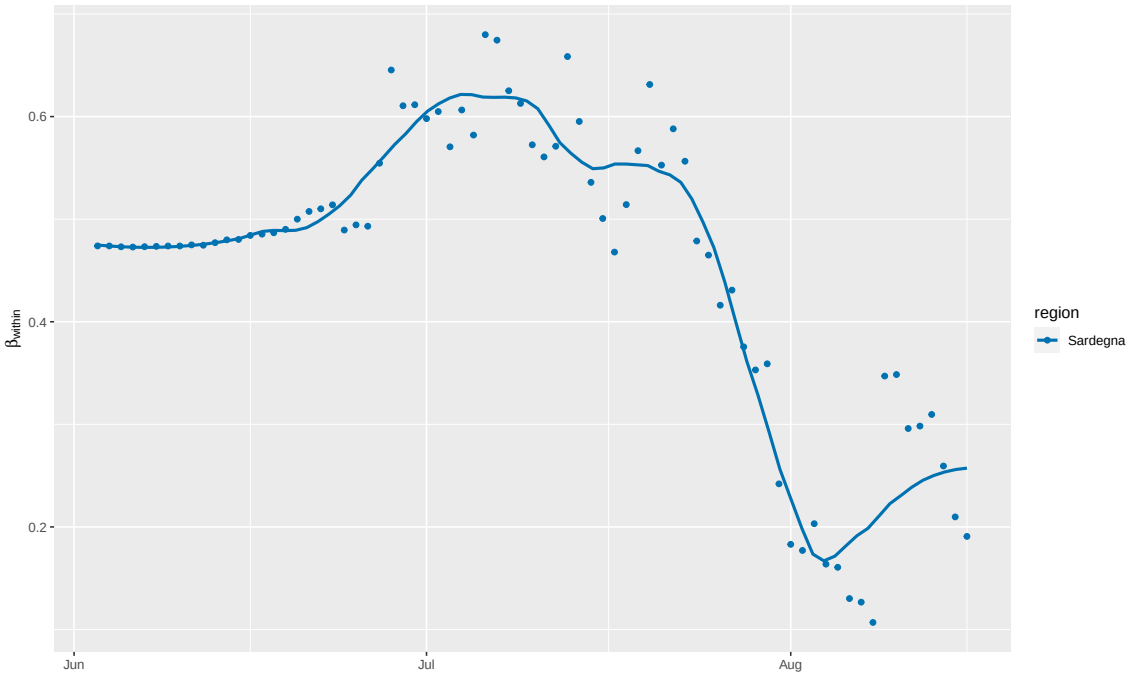
Figure C.3. Progression of β_{within} over time for the *Sud* (South) NUTS 1 region



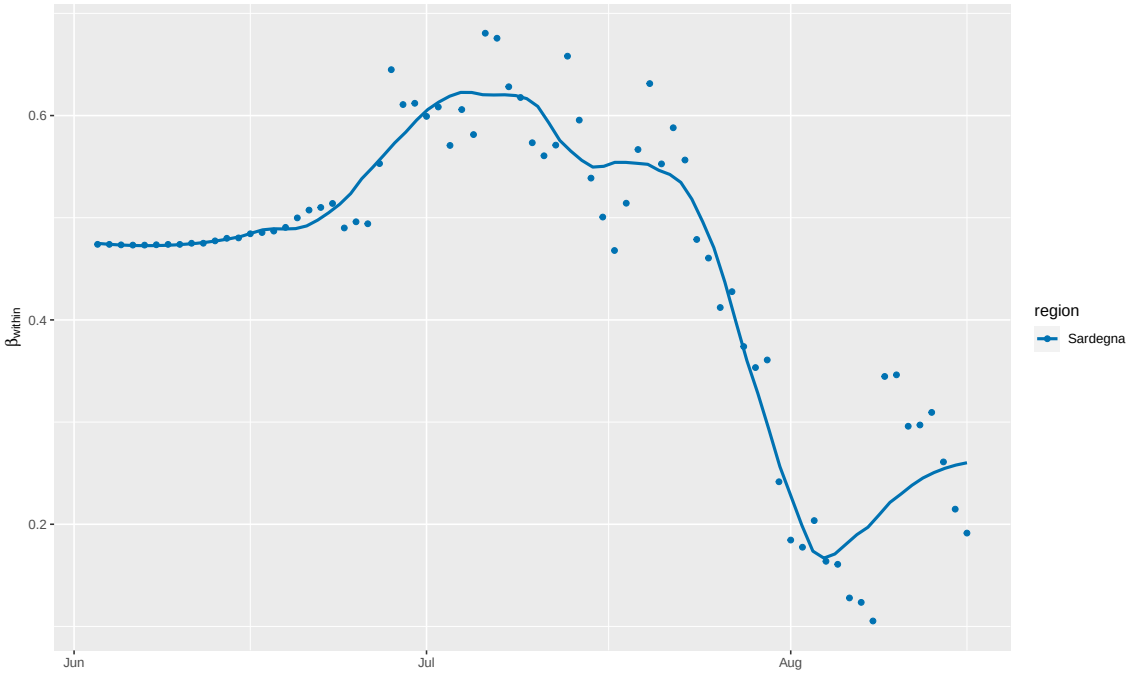
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



(d) With model selection by AIC;
including undocumented infectives

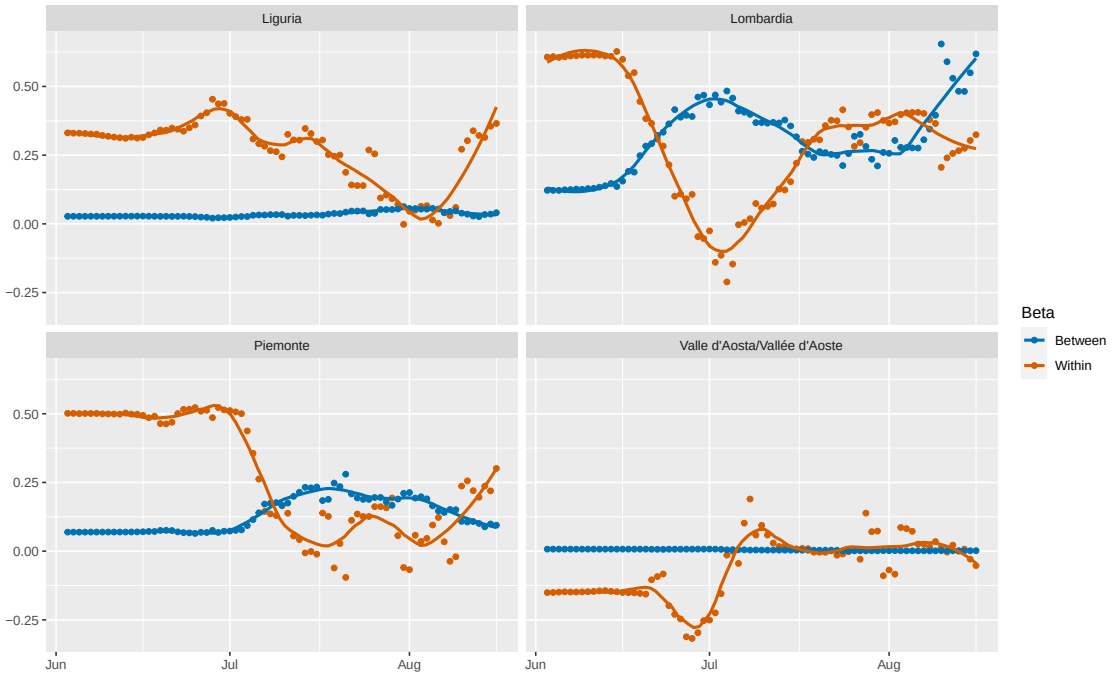
Figure C.4. Progression of β_{within} over time for the *Isole* (Islands) NUTS 1 region

C.2 Plots for Within and Between-Region Spread Model

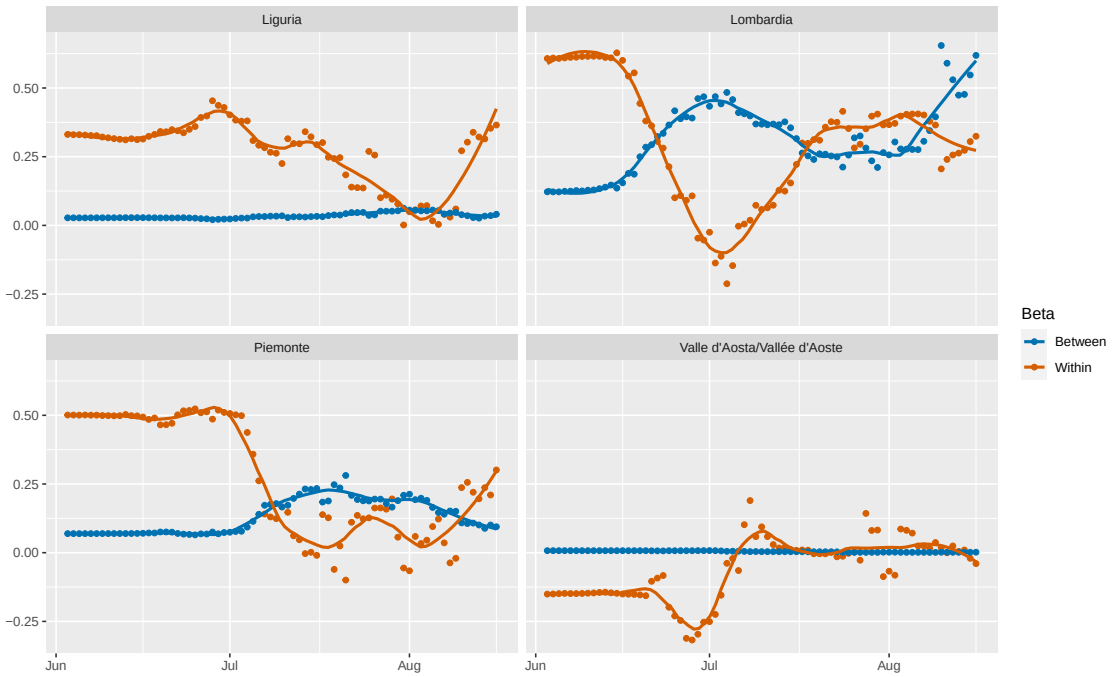
In Section 5.3 we presented the plots of β_{within} and $\beta_{between}$ over time for the *Nord-Est* (North-East) NUTS 1 region for Within and Between-Region Spread Model. In this section, we present the plots for the other NUTS 1 regions. As is the case for Section 5.3, we use the last 100 observations.

(a) Without model selection

(b) With model selection by AIC

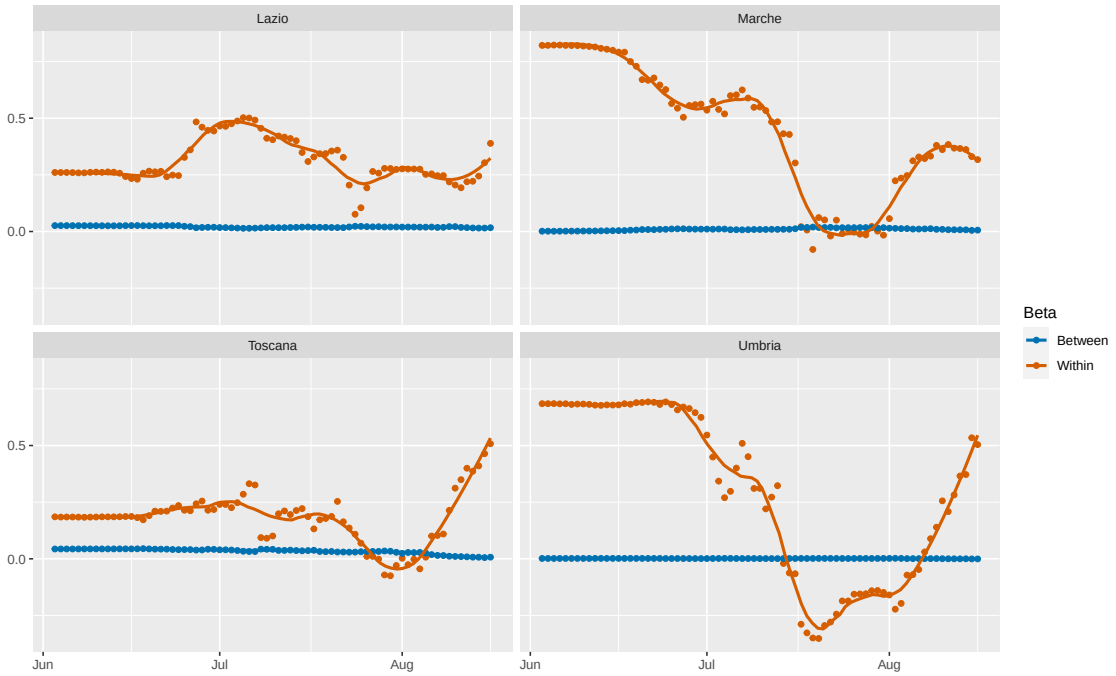


(c) Without model selection;
including undocumented infectives

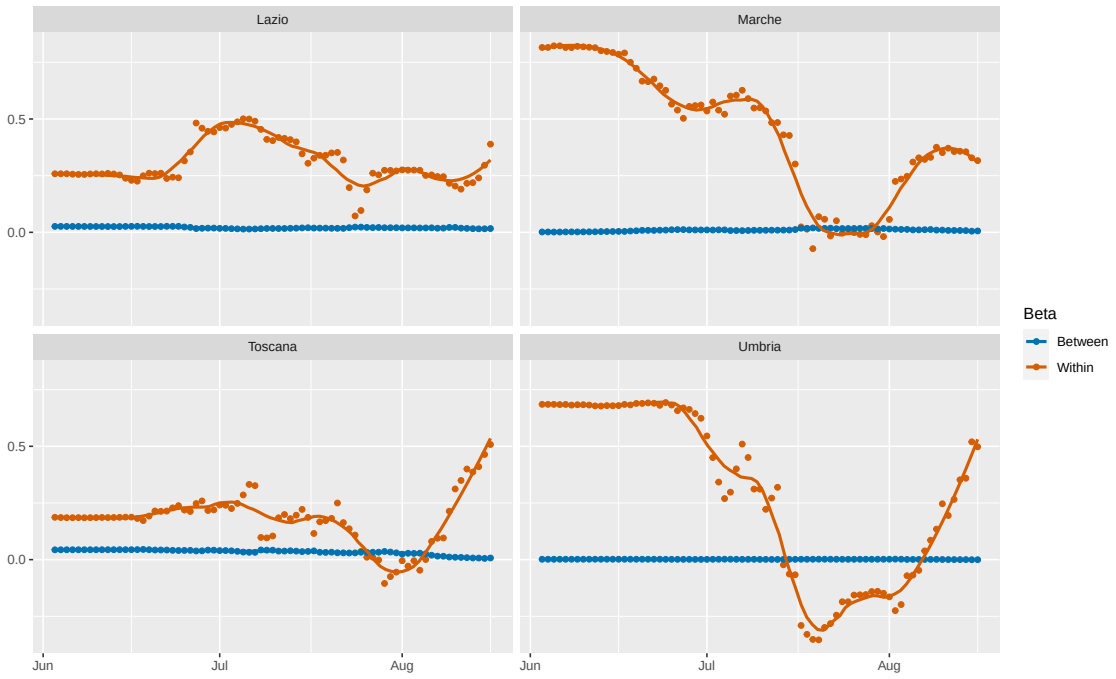


(d) With model selection by AIC;
including undocumented infectives

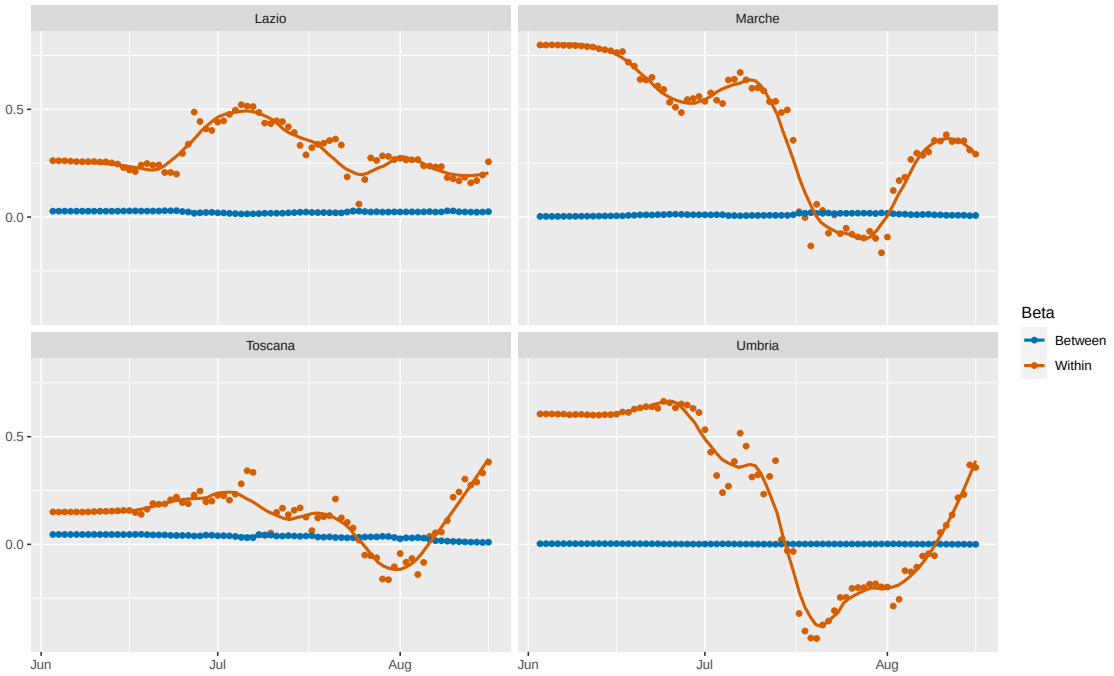
Figure C.5. Progression of β_{within} and $\beta_{between}$ over time for the *Nord-Ovest* (North-West) NUTS 1 region



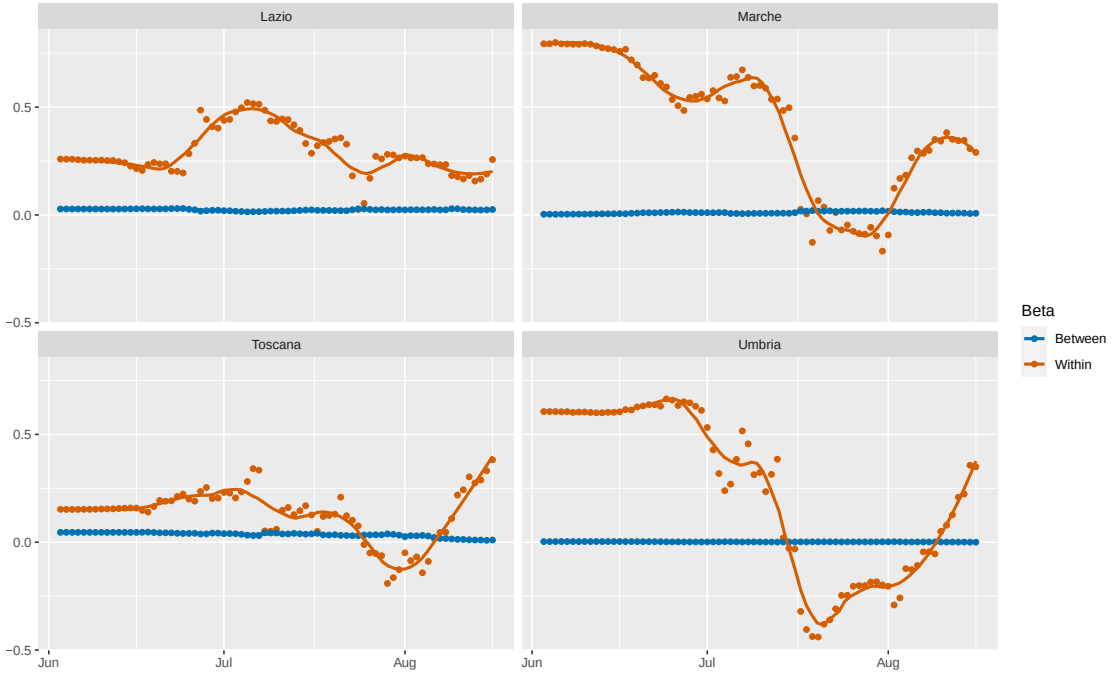
(a) Without model selection



(b) With model selection by AIC

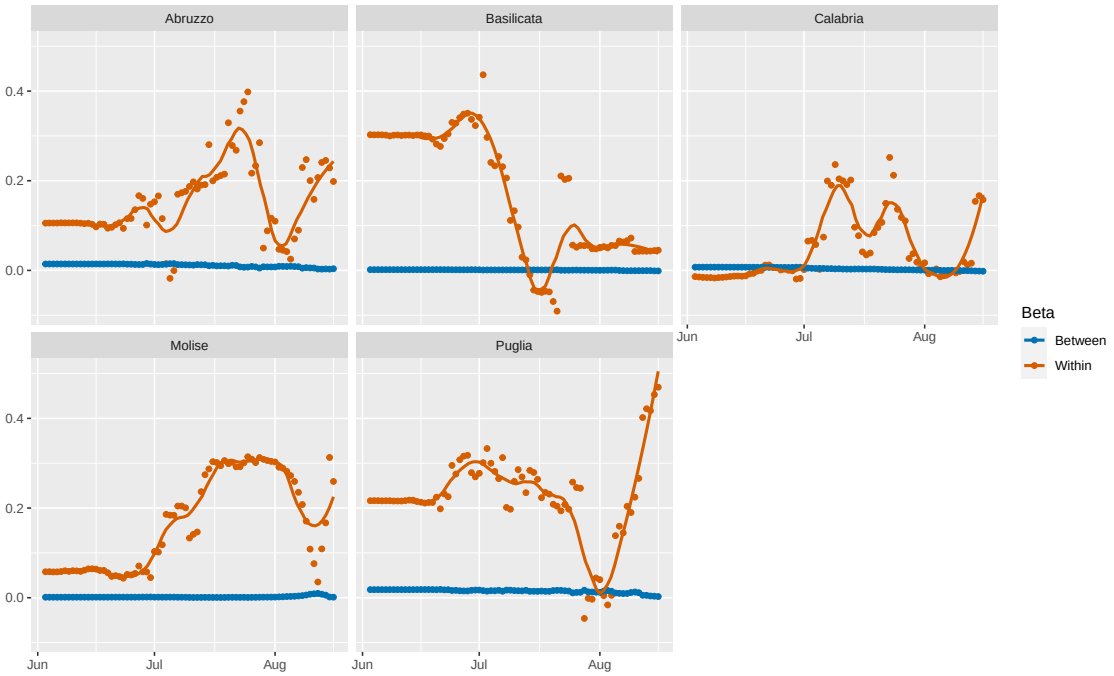


(c) Without model selection;
including undocumented infectives

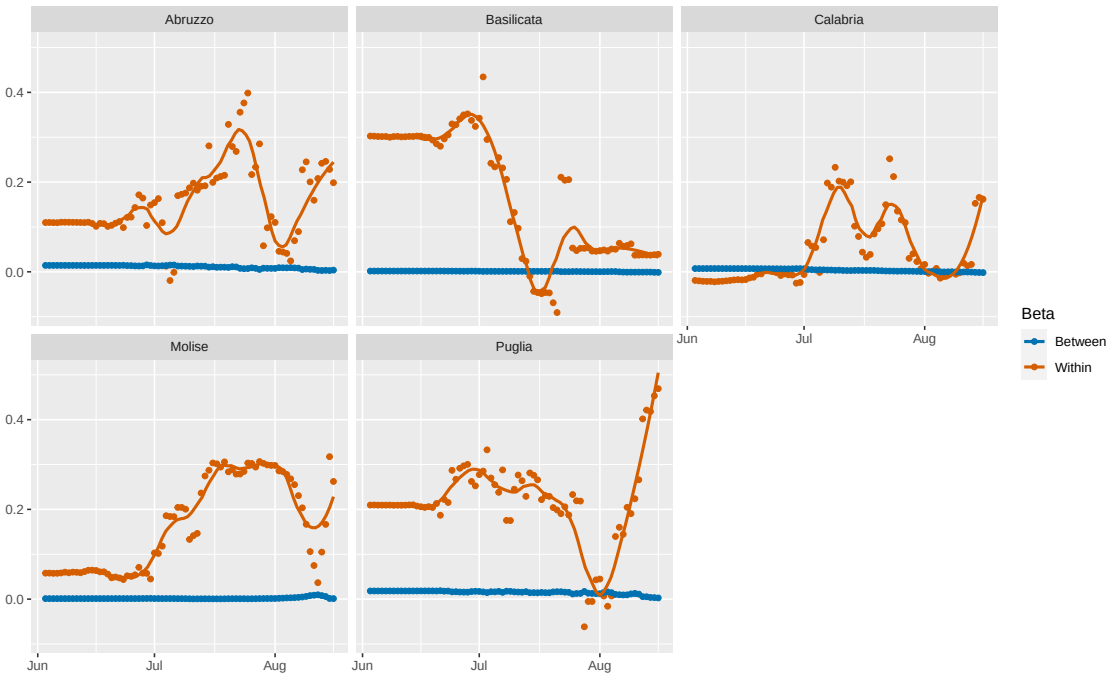


(d) With model selection by AIC;
including undocumented infectives

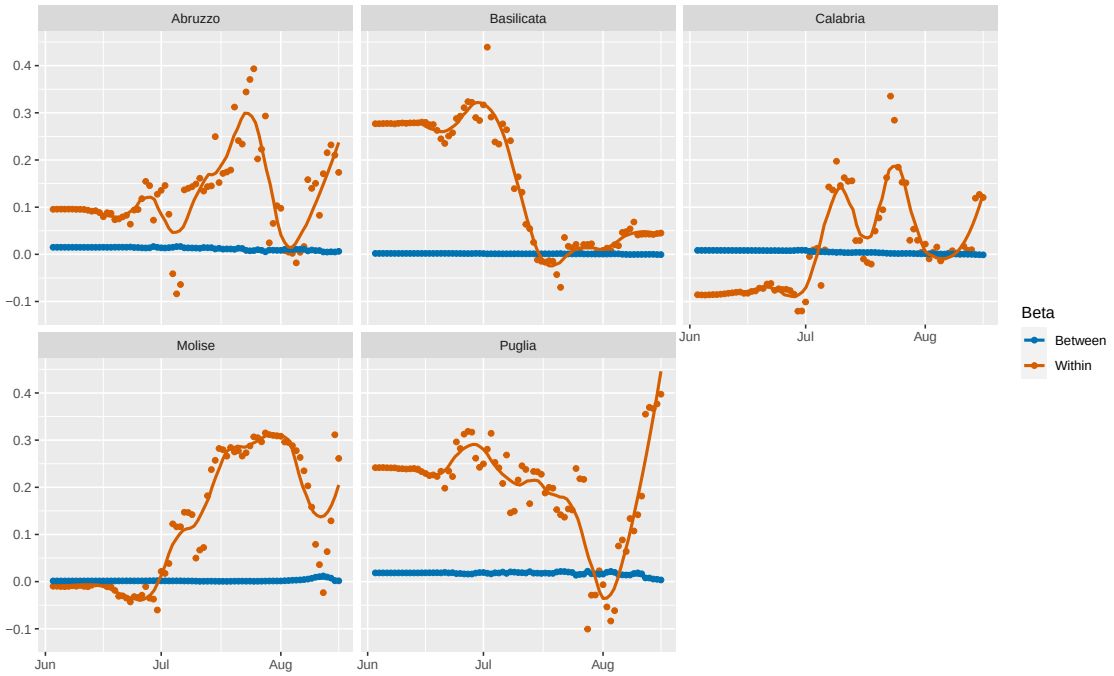
Figure C.6. Progression of β_{within} and $\beta_{between}$ over time for the *Centro (IT)* (Centre) NUTS 1 region



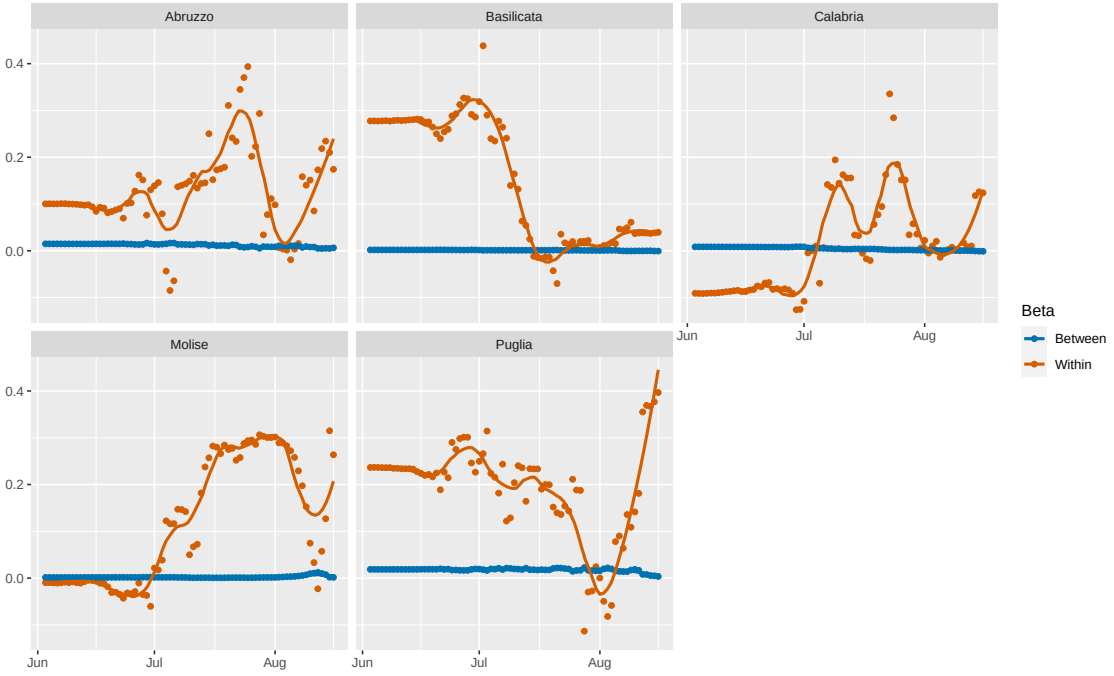
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



(d) With model selection by AIC;
including undocumented infectives

Figure C.7. Progression of β_{within} and $\beta_{between}$ over time for the *Centro (IT)* (Centre) NUTS 1 region

D Derivations

D.1 Calculation of population variables

In this appendix, we will explain how the susceptible population and total population are calculated. Unfortunately, we do not have data on the total population per day. For this reason, we retrieve the latest population numbers per region from Eurostat (2020b), which are from January 1, 2019, and the yearly population growth rates for 2019 and 2020 from Worldometer (2020). For 2019, growth rate was equal to -0.13% and for 2020, excluding the deaths due to the pandemic, it was estimated to be equal to -0.15%. We only have the population growth rates available for the whole of Italy, not per region, unfortunately. As such, we assume that the growth rates are uniformly applicable to all regions. Of course, this is likely to introduce a small error since these growth rates differ over the regions. We assume that this error is negligible.

We denote the population of region r at time t by $N_{r,t}$. We denote the yearly population growth rates for 2019 and 2020 by g_{2019} and g_{2020}, respectively. Lastly, recall that the data for the pandemic starts at February 25, 2020. This is the 54^{th} day of 2020, a leap year. As such, the population of region r on February 25, 2020 is calculated as:

$$N_{r,\text{2020-02-25}} = (1 + g_{2019})(1 + g_{2020})^{\frac{54}{366}} N_{r,\text{2019-01-01}} - D_{r,\text{2020-02-25}} \tag{D.1}$$

where $D_{r,t}$ denotes the number of deaths in region r at time t.

Recall that the data reported at time t is reported with respect to the last 24 hours. As such, the susceptible population at time t can be calculated with the data at that same time. The susceptible population of region r at time t, denoted by $X_{r,t}$, is therefore calculated as follows:

$$X_{r,t} = N_{r,t} - Y_{r,t} - Z_{r,t} \tag{D.2}$$

where $Y_{r,t}$ denotes the number of infectives and $Z_{r,t}$ denotes the number of removed individuals. Recall that Z is made up by adding the recovered individuals $R_{r,t}$ and the deceased individuals $D_{r,t}$. Because we use the calculation of $N_{r,t}$ as in the previous paragraph, the error discussed propagates into the calculation of $X_{r,t}$. However, as before, we assume that this error is negligible.

D.2 Functional forms for modelling undocumented infectives

In this appendix, we give the derivations for the functional forms for modelling undocumented infectives as discussed in Section 3.7.

D.2.1 Linear function

For modelling the undocumented infectives, 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) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,

(III) $f(N_t) = 1$

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

$$aN_t + f^{min} = 1.$$

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

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

D.2.2 General quadratic function

For modelling the undocumented infectives, 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) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,

(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 and to the left of the origin in the case of an upwards opening parabola.

From assumption (II), we obtain that $c = f^{min}$. 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:

$$\begin{cases} aN_t^2 + bN_t + f^{min} & = 1 \text{ (from assumption (III))} \\ a\beta^2 N_t^2 + b\beta N_t + f^{min} & = \gamma \text{ (from assumption (IV))} \end{cases} \tag{D.3}$$

To solve (D.3), we can apply row reduction as follows:

$$
\begin{pmatrix} N_t^2 & N_t \\ \beta^2 N_t^2 & \beta N_t \end{pmatrix} \left| \begin{matrix} 1-f^{min} \\ \gamma - f^{min} \end{matrix} \right. \xrightarrow{r_2 - \beta^2 r_1} \left(\begin{matrix} N_t^2 & N_t \\ 0 & \beta(1-\beta)N_t \end{matrix} \right| \left. \begin{matrix} 1-f^{min} \\ \gamma - f^{min} - \beta^2 + \beta^2 f^{min} \end{matrix} \right)
$$

$$
\xrightarrow{r_2 \div \beta(1-\beta)} \left(\begin{matrix} N_t^2 & N_t \\ 0 & N_t \end{matrix} \right| \left. \begin{matrix} 1-f^{min} \\ \frac{\gamma - f^{min} - \beta^2 + \beta^2 f^{min}}{\beta(1-\beta)} \end{matrix} \right)
$$

$$
\xrightarrow{r_1 - r_2} \left(\begin{matrix} N_t^2 & 0 \\ 0 & N_t \end{matrix} \right| \left. \begin{matrix} \frac{\beta - \gamma + (1-\beta) f^{min}}{\beta(1-\beta)} \\ \frac{\gamma - f^{min} - \beta^2 + \beta^2 f^{min}}{\beta(1-\beta)} \end{matrix} \right)
$$

$$
\begin{matrix} \xrightarrow{r_1 \div N_t^2} \\ \xrightarrow{r_2 \div N_t} \end{matrix} \left(\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right| \left. \begin{matrix} \frac{\beta - \gamma + (1-\beta) f^{min}}{\beta(1-\beta)N_t^2} \\ \frac{\gamma - f^{min} - \beta^2 + \beta^2 f^{min}}{\beta(1-\beta)N_t} \end{matrix} \right)
$$

As such, we have derived that

$$
\begin{cases} a &= \frac{\beta - \gamma + (1-\beta) f^{min}}{\beta(1-\beta)N_t^2} \\ b &= \frac{\gamma - f^{min} - \beta^2 + \beta^2 f^{min}}{\beta(1-\beta)N_t} \\ c &= f^{min}. \end{cases} \tag{D.4}
$$

Firstly, note that this function is an upwards opening parabola if $a > 0$ and a downwards opening parabola if $a < 0$. For instance, we have that:

$$
\begin{aligned}
& a > 0 \\
\iff & \frac{\beta - \gamma + (1-\beta) f^{min}}{\beta(1-\beta)N_t^2} > 0 \\
\iff & \beta - \gamma + (1-\beta) f^{min} > 0 \\
\iff & \gamma < \beta + (1-\beta) f^{min}
\end{aligned}
$$

where we use that $\beta(1-\beta)N_t^2 > 0$. Similarly, we have that $a < 0$ if $\gamma > \beta + (1-\beta) f^{min}$.

Now note that our function is continuous. As such, we assume without loss of generality that $\beta = \frac{1}{2}$ 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

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

Firstly, assuming $\beta = \frac{1}{2}$, the expressions for a and b as in (D.4) reduce to:

$$
\begin{cases} a &= \frac{\frac{1}{2} - \gamma + \frac{1}{2} f^{min}}{\frac{1}{4} N_t^2} \\ &= \frac{2 - 4\gamma + 2 f^{min}}{N_t^2} \\ b &= \frac{\gamma - f^{min} - \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 f}{\frac{1}{4} N_t} \\ &= \frac{4\gamma - 1 - 3 f^{min}}{N_t}. \end{cases} \tag{D.5}
$$

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

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

Using (D.5), we can fill out a and b to obtain:

$$TC_t = \frac{1 - 4\gamma + 3f^{min}}{4 - 8\gamma + 4f^{min}} N_t.$$

Let $\gamma > \frac{1}{2} + \frac{1}{2}f^{min}$. Then, we need to derive γ such that

$$\begin{aligned} & \frac{1 - 4\gamma + 3f^{min}}{4 - 8\gamma + 4f^{min}} N_t \geq N_t \\ \iff & \frac{1 - 4\gamma + 3f^{min}}{4 - 8\gamma + 4f^{min}} \geq 1. \end{aligned}$$

Note that this is only the case if two conditions are satisfied:

$$\begin{cases} sign(1 - 4\gamma + 3f^{min}) & = sign(4 - 8\gamma + 4f^{min}) & \text{(D.6a)} \\ |1 - 4\gamma + 3f^{min}| & \geq |4 - 8\gamma + 4f^{min}| & \text{(D.6b)} \end{cases}$$

Note that our assumption that $\gamma > \frac{1}{2} + \frac{1}{2}f^{min}$ is equivalent to $\gamma > \frac{2+2f^{min}}{4}$ which, in turn, is equivalent to $4 - 8\gamma + 4f^{min} < 0$. As such, (D.6a) tells us that both the numerator and denominator of the fraction are negative. Therefore, to satisfy (D.6a), we need that

$$\begin{aligned} & 1 - 4\gamma + 3f^{min} < 0 \\ \iff & \gamma > \frac{1 + 3f^{min}}{4} \end{aligned}$$

Since we assumed that $\gamma > 2 + 2f^{min}$, this is always satisfied because $f^{min} \in [0, 1]$ so that $1 + 3f^{min} < 2 + 2f^{min} < \gamma$. That brings us to the second condition (D.6b). Because we know that both parts of the fractions are negative, we can now solve for γ as follows:

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

Let $\gamma < \frac{1}{2} + \frac{1}{2}f^{min}$. Then, we need to derive γ such that

$$\begin{aligned} &\frac{1 - 4\gamma + 3f^{min}}{4 - 8\gamma + 4f^{min}} N_t \leq 0 \\ \iff &\frac{1 - 4\gamma + 3f^{min}}{4 - 8\gamma + 4f^{min}} \leq 0. \end{aligned}$$

Note that this is only the case if one of the following two conditions is satisfied:

$$\begin{cases} 1 - 4\gamma + 3f^{min} \leq 0 & \text{and } 4 - 8\gamma + 4f^{min} > 0 \quad \text{(D.7a)} \\ 1 - 4\gamma + 3f^{min} \geq 0 & \text{and } 4 - 8\gamma + 4f^{min} < 0 \quad \text{(D.7b)} \end{cases}$$

As before, note that our assumption that $\gamma > \frac{1}{2} + \frac{1}{2}f^{min}$ is equivalent to $4 - 8\gamma + 4f^{min} > 0$. As such, we know that the only condition that can be satisfied is (D.7a). Therefore, we need that

$$\begin{aligned} &1 - 4\gamma + 3f^{min} \leq 0 \\ &\gamma \geq \frac{1 + 3f^{min}}{4} = \frac{1}{4} + \frac{3}{4}f^{min}. \end{aligned}$$

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

Conclusively, we have derived that

$$f(TC_t) = \frac{2 - 4\gamma + 2f^{min}}{N_t^2} TC_t^2 + \frac{4\gamma - 1 - 3f^{min}}{N_t} TC_t + f^{min},$$

under the assumption that $\beta = \frac{1}{2}$, with $\gamma \in \left[\frac{1}{4} + \frac{3}{4}f^{min}, \frac{3}{4} + \frac{1}{4}f^{min}\right]$.

D.2.3 Special case quadratic formula: downwards opening

For modelling the undocumented infectives, 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) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,

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

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

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:

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

Therefore, the formula becomes:

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

D.2.4 Special case quadratic formula: upwards opening

For modelling the undocumented infectives, 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) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,

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

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

Just as in appendix D.2.4, we use the vertex form $f(TC_t) = a(TC_t - h)^2 + k$. By assumptions (III) and (IV), $h = 0$ and $k = f^{min}$. Therefore,

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

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

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

Therefore, the formula becomes:

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

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

D.2.5 Cubic function

For modelling the undocumented infectives, 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) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,

(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 $d = f^{min}$. 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:

$$\begin{cases} aN_t^3 + bN_t^2 + cN_t + f^{min} & = 1 \text{ (from assumption (III))} \\ a\beta_1^3 N_t^3 + b\beta_1^2 N_t^2 + c\beta_1 N_t + f^{min} & = \gamma_1 \text{ (from assumption (IV))} \\ a\beta_2^3 N_t^3 + b\beta_2^2 N_t^2 + c\beta_2 N_t + f^{min} & = \gamma_2 \text{ (from assumption (IV))} \end{cases} \tag{D.8}$$

In appendix D.2.2, 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 keep the derivation general. As such, we first assume without loss of generality that $\beta_1 = \frac{1}{4}$ and $\beta_2 = \frac{1}{2}$. To solve (D.8), we can then apply row reduction as follows:

$$
\begin{pmatrix} N_t^3 & N_t^2 & N_t & 1-f^{min} \\ \beta_1^3 N_t^3 & \beta_1^2 N_t^2 & \beta_1 N_t & \gamma_1 - f^{min} \\ \beta_2^3 N_t^3 & \beta_2^2 N_t^2 & \beta_2 N_t & \gamma_2 - f^{min} \end{pmatrix} = \begin{pmatrix} N_t^3 & N_t^2 & N_t & 1-f^{min} \\ \frac{1}{64} N_t^3 & \frac{1}{16} N_t^2 & \frac{1}{4} N_t & \gamma_1 - f^{min} \\ \frac{1}{8} N_t^3 & \frac{1}{4} N_t^2 & \frac{1}{2} N_t & \gamma_2 - f^{min} \end{pmatrix}
$$

$$
\xrightarrow[r_3 \times 8]{r_2 \times 64} \begin{pmatrix} N_t^3 & N_t^2 & N_t & 1-f^{min} \\ N_t^3 & 4N_t^2 & 16N_t & 64\gamma_1 - 64f^{min} \\ N_t^3 & 2N_t^2 & 4N_t & 16\gamma_2 - 64f^{min} \end{pmatrix}
$$

$$
\xrightarrow[r_3 - r_1]{r_2 - r_1} \begin{pmatrix} N_t^3 & N_t^2 & N_t & 1-f^{min} \\ 0 & 3N_t^2 & 15N_t & -1 + 64\gamma_1 - 63f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 - 7f^{min} \end{pmatrix}
$$

$$
\xrightarrow{r_2 \leftrightarrow r_3} \begin{pmatrix} N_t^3 & N_t^2 & N_t & 1-f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 - 7f^{min} \\ 0 & 3N_t^2 & 15N_t & -1 + 64\gamma_1 - 63f^{min} \end{pmatrix}
$$

$$
\xrightarrow[r_3 - 3r_2]{r_1 - r_2} \begin{pmatrix} N_t^3 & 0 & -2N_t & 2 - 8\gamma_2 + 6f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 \\ 0 & 0 & 6N_t & 2 + 64\gamma_1 - 24\gamma_2 - 42f^{min} \end{pmatrix}
$$

$$
\xrightarrow[r_2 - \frac{1}{2} r_3]{r_1 + \frac{1}{3} r_3} \begin{pmatrix} N_t^3 & 0 & 0 & \frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3} \\ 0 & N_t^2 & 0 & -2 - 32\gamma_1 + 20\gamma_2 + 14f^{min} \\ 0 & 0 & 6N_t & 2 + 64\gamma_1 - 24\gamma_2 - 42f^{min} \end{pmatrix}
$$

$$
\xrightarrow[r_3 \div 6N_t]{r_2 \div N_t^2} \begin{pmatrix} 1 & 0 & 0 & \frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3N_t^3} \\ 0 & 1 & 0 & \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14f^{min}}{N_t^2} \\ 0 & 0 & 1 & \frac{2 + 64\gamma_1 - 24\gamma_2 - 42f^{min}}{6N_t} \end{pmatrix}
$$

Conclusively, we have derived that

$$
\begin{cases} a &= \frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3N_t^3} \\ b &= \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14f^{min}}{N_t^2} \\ c &= \frac{2 + 64\gamma_1 - 24\gamma_2 - 42f^{min}}{6N_t} = \frac{1 + 32\gamma_1 - 12\gamma_2 - 21f^{min}}{3N_t} \\ d &= f^{min} \end{cases} \tag{D.9}
$$

so that

$$
\begin{aligned} f(TC_t) = &\frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3N_t^3} TC_t^3 + \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14f^{min}}{N_t^2} TC_t^2 \\ &+ \frac{1 + 32\gamma_1 - 12\gamma_2 - 21f^{min}}{3N_t} TC_t + f^{min}, \end{aligned}
$$

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