Predicting The Transmission 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:
September 17, 2020

Abstract

This thesis makes two main contributions to the existing literature on the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Firstly, a method for modelling undocumented infectives is developed. Secondly, the models by Adda (2016) are applied to estimate the transmission rate of SARS-CoV-2 across Italian regions. We find that the models correctly indicate that the virus has been transmitted much more severely within and between northern regions, such as Lombardy, than in southern regions, such as Calabria. Consequently, the approaches in this thesis can also be applied to other regions and when another pandemic occurs in the future.

Acknowledgements

I would like to start by thanking my supervisor: dr. Otilia Boldea. After I made the choice to terminate my thesis at the National Library after one month, she approached me with the proposal to write my thesis on the very interesting and topical subject of COVID-19. Throughout the process, despite the inconvenience that we could not meet in person due to the pandemic, she offered expert advice and thorough answers to my questions.

On a personal note, I would like to thank my loving partner Fenna for supporting me throughout. She has celebrated positive times together with me, comforted me when times were a bit more dreary, and provided critical feedback on this thesis. A quick thank-you also goes out to my parents, Edwin and Monique, and my sister Lieke. Even though you often joke that you do not understand much about what I study, you have supported me nonetheless. Also in my search for a first job you have stood by me and helped me to land a beautiful position.

Lastly, I want to express my gratitude to the institution of Tilburg University and everyone that has made my time there a blast. I can genuinely say that I enjoyed my time as a student at Tilburg University. The first time that I set foot on campus, I instantly felt at home and knew that this was the right location to pursue my studies. The great people that I got to meet within the econometrics program and at the Tilburg Debating Society Cicero (a special shoutout goes out to Jos, Lisa, Lotte, Isis, and Roel) have made this phase of my life one to never forget and to always look back on with joy. Now, it is time to close this chapter of my life and to start a brand new one.

Contents

1 Introduction

Since December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has plagued the world, having infected almost 30 million people. The infectious respiratory disease caused by SARS-CoV-2, called coronavirus disease 2019 (commonly abbreviated to COVID-19) has consequently been responsible for nearly a million deaths. The virus is primarily spread through close human contact and respiratory droplets generated by breathing, sneezing, coughing, etcetera (European Centre for Disease Prevention and Control, 2020b). This has spurred governments worldwide to try and contain the virus by implementing far-reaching measures, such as shutting down schools and restaurants or even by locking down the entire country. Moreover, countries have encouraged or even enforced social distancing, where individuals should keep at least one and a half metres distance to one another. Social distancing has been employed in most public places, such as supermarkets, department stores, and schools, causing a massive change in social behaviour.

It is crucial to understand viral diseases inside and out; it allows policy makers to decide which policies to implement and to look back at which policies were effective in driving the viruses back. To this extent, exact models should be built that can represent the situation at hand and accurately inform those who need the information with the insights needed to make proper and impactful decisions. Adda (2016) develops econometric models to analyze the incidence of influenza, chickenpox, and gastroenteritis using high-frequency data from France. These models aim to incorporate spatial spillover effects between regions and the impacts of these viral diseases on economic activity. Moreover, R. Li et al. (2020) found that around 86% of the infectives in China went undocumented and that these were also contagious. This thesis has two goals, namely to apply the models developed by Adda (2016) to SARS-CoV-2 and to develop a method to model undocumented infectives. We focus on the country of Italy, which has been one of the most severely affected countries in the world. For 21 Italian regions, we gather data on coronavirus-related variables from the Italian Department of Civil Protection (Rosini, 2020). This data is publicly available and spans the time from February 26, 2020 onward. We use data until August 16.

We make two major contributions in this thesis. First of all, we recognize that a large portion of the people that are infectious are not tested and, hence, go undocumented. A major issue when considering epidemics and pandemics is the inherent problem of a limited testing capacity. An exponential increase in the number of cases, especially for a viral disease that is novel and about which very little is known, means that more and more people should get tested but that it is difficult to meet the demand for tests. The impact of this is that there are many infections that went and still are going undocumented, meaning that the scope of the problem is much larger than the numbers reported (R. Li et al., 2020). In effect, an increasing number of

people becomes infected and is able to infect others, meaning that the virus becomes more difficult to control. In this thesis, we develop a method that estimates the number of undocumented infective by using the testing capacity as a measure.

The second contribution that this thesis makes is that we acknowledge that there are likely structural regional differences that make it difficult to estimate a general national model that is applicable to all regions within that country. To that end, we estimate two models as presented by Adda (2016) to take into account these regional effects. The first model, called the within-region spread model, ignores interaction between regions. The second model is the so-called within and between-region spread model, which takes into account the infectives in other regions as well.

In Section 2, we describe the history of the COVID-19 pandemic, the magnitude of the situation in Italy, and discuss the incidence across the Italian regions. Section 3 talks about the dataset that we use and how it was processed for analysis. Subsequently, Section 4 introduces one of the most commonly used models in epidemiology, namely the Standard Inflammatory Response (SIR) model. After this, we discuss the method of model selection that is applied in this thesis in Section 5. Section 6 defines and explores our method for modelling undocumented infectives. Thereafter, in Section 7, we discuss the within-region spread model as presented by Adda (2016). The within and between-region spread model is discussed in Section 8. Finally, a conclusion is given in Section 9 and proposals for future research are presented in Section 10.

2 Problem Description

The spread of SARS-CoV-2 started in Wuhan, China, from which it has made its way to nearly every country in the world. On September 16, 2020, almost 30 million people were reported to have been infected with COVID-19. This has lead to nearly a million consequent deaths, making it the fourteenth most deadly viral disease to ever have existed (LePan, 2020). At the moment of writing, only 12 sovereign member states of the United Nations reported no infections, of which 10 are island countries. The other two countries are North Korea and Turkmenistan but it is suspected that there are actually cases but that these are not reported for both North Korea (Nebehay, 2020) and Turkmenistan (Human Rights Watch, 2020). It is even rumored that North Korean authorities have issued shoot-to-kill orders to prevent the coronavirus entering the country from China (Agence France-Presse, 2020). In Turkmenistan, despite no official cases having been registered, the hospitals are overflowing with patients showing symptoms that align with those of COVID-19 (RFE/RL, 2020).

The World Health Organization (WHO) declared a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 (WHO, 2020a), defined as *"an*

extraordinary event which is determined to constitute a public health risk to other States through the international spread of disease and to potentially require a co-ordinated international response" (WHO, 2019). After the spread of SARS-CoV-2 only became worse, the WHO declared the virus outbreak to be a pandemic on 11 March 2020 (WHO, 2020b), where a pandemic is defined as *"an epidemic occurring on worldwide or over a very wide area, crossing international boundaries, and usually affecting a large number of people"* (Porta, 2014).

Due to the extreme nature of the pandemic and the availability of enough data, this thesis chooses to focus on Italy. Italy has been one of the most intensely struck countries by SARS-CoV-2. Until the end of March, it had the highest number of confirmed cases per 100,000 inhabitants in the world, before being surpassed by Spain. 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. Italy reported the second highest death-to-cases ratio of 14.45% (34,818 deaths to 240,961 cases), only after the United Kingdom, which reported a death-to-cases ratio of 15.50% (43,995 deaths to 283,757 cases). 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 instead of being admitted to the hospital, as well as literal collapses of overworked healthcare workers (Horowitz, 2020). In the city of Bergamo, Lombardy's fourth largest city, the virus spread like wildfire; the army even had to be sent in to handle the rapid increase in the number of deaths (Scarr & Sharma, 2020).

Our contribution is to model the pandemic at the regional level rather than at the national level. Even though the regions within Italy are likely more similar to one another than, for instance, to the provinces in the Netherlands, there are large regional differences in Italy as well. This implies that there is not one model that can be used to analyze the entire country of Italy because this would ignore the regional heterogenous effects. Besides doing so to uncover heterogeneity across regions, the regional variation allows for better identification of the average transmission parameters.

To illustrate the regional differences, we present several figures. Figure 2.1 shows the incidence rates over time categorized by the five overarching regions (called NUTS 1 regions) that the regions are a part of. The geographical classification of regions is explained more thoroughly in Section 3.1. The incidence rate is defined as the total number of people currently infected divided by the total population.

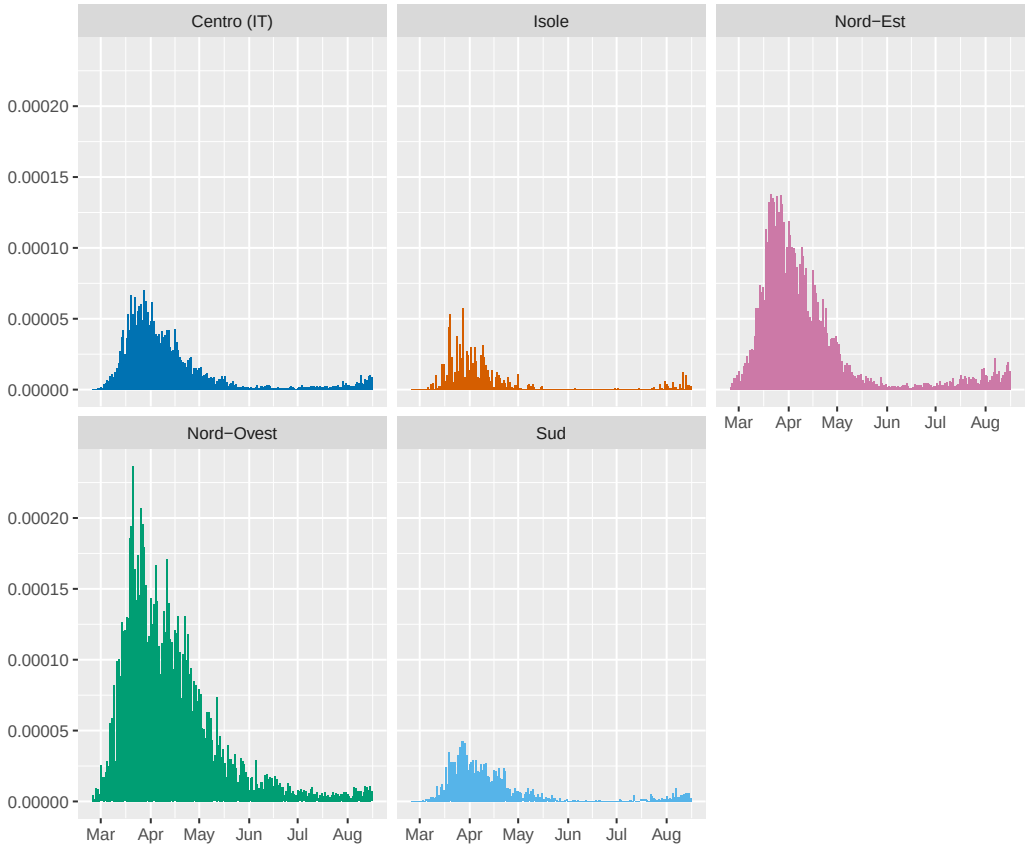


Figure 2.1. Incidence rate per NUTS 1 region.

In Figure 2.1, we can see that there is a wide difference in the incidence rates between these larger regions. Not only do we see that the heights of the peaks differ, we also notice that the length of the peaks differ slightly over the regions. This already shows that models that pool these regions together, to form Italy as a whole, are likely less suitable than models that take these differences into account.

Consider Figure 2.2, where we zoom in on the Centro (IT) NUTS 1 region by looking at the four regions that make it up. These lower level regions are called the NUTS 2 regions. The plots for the other NUTS 1 regions can be found in Appendix C.1.

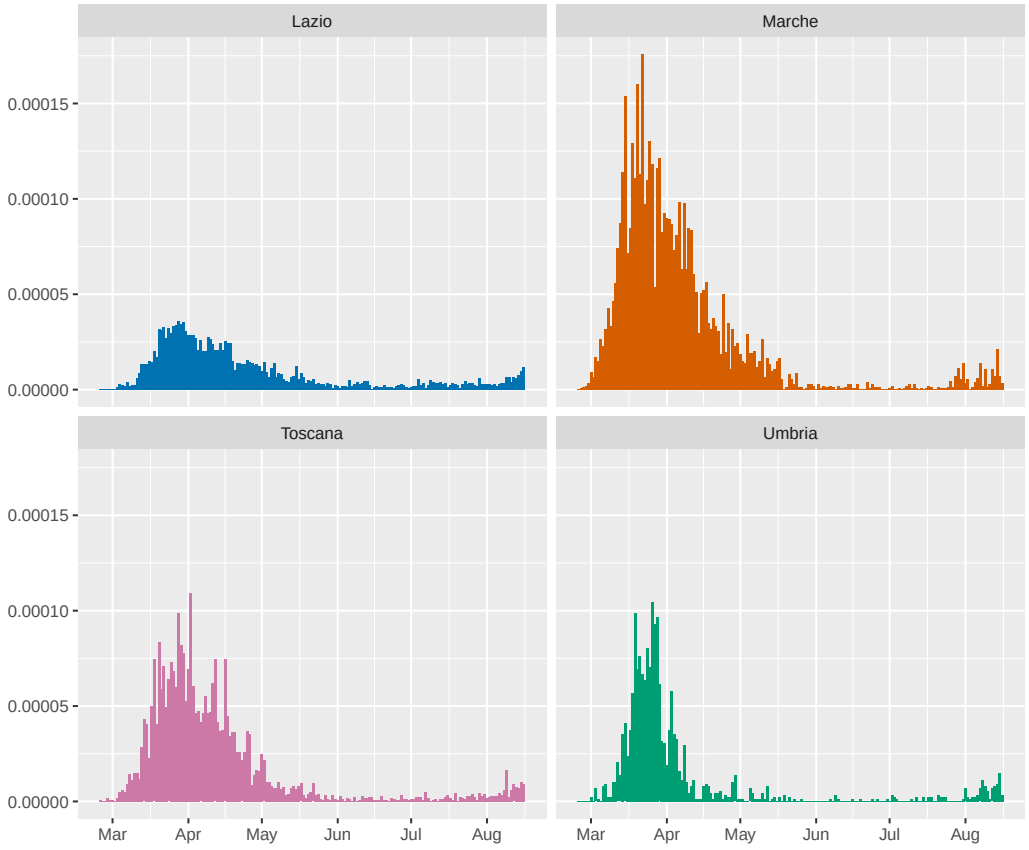


Figure 2.2. Incidence rate per NUTS 2 region for the Centro (IT) NUTS 1 region.

Figure 2.2 shows that, even among the regions in the NUTS 1 region, there is a vast difference. For the region of Lazio, we see a much lower peak than for the other three regions, especially compared to Marche. Moreover, the varying length of the peak is more visible in Figure 2.2, for example when comparing the regions of Marche and Umbria.

We also investigate the regional heterogeneity by looking at the mean incidence rate over time across the regions. We consider the mean incidence rate because this allows us to compare regions, even when these have a higher population county. A caveat should be made that there may be heterogeneity in the transmission that may be independent of the mean amount of cases. However, investigating the mean incidence rate may give us some sort of idea of how the pandemic has impacted the regions, at least on average. Consider Figure 2.3, which shows the mean incidence rate for the NUTS 2 regions, divided up into their respective NUTS 1 regions. Along with the mean incidence rate, we present an error bar the size of two times the standard error.

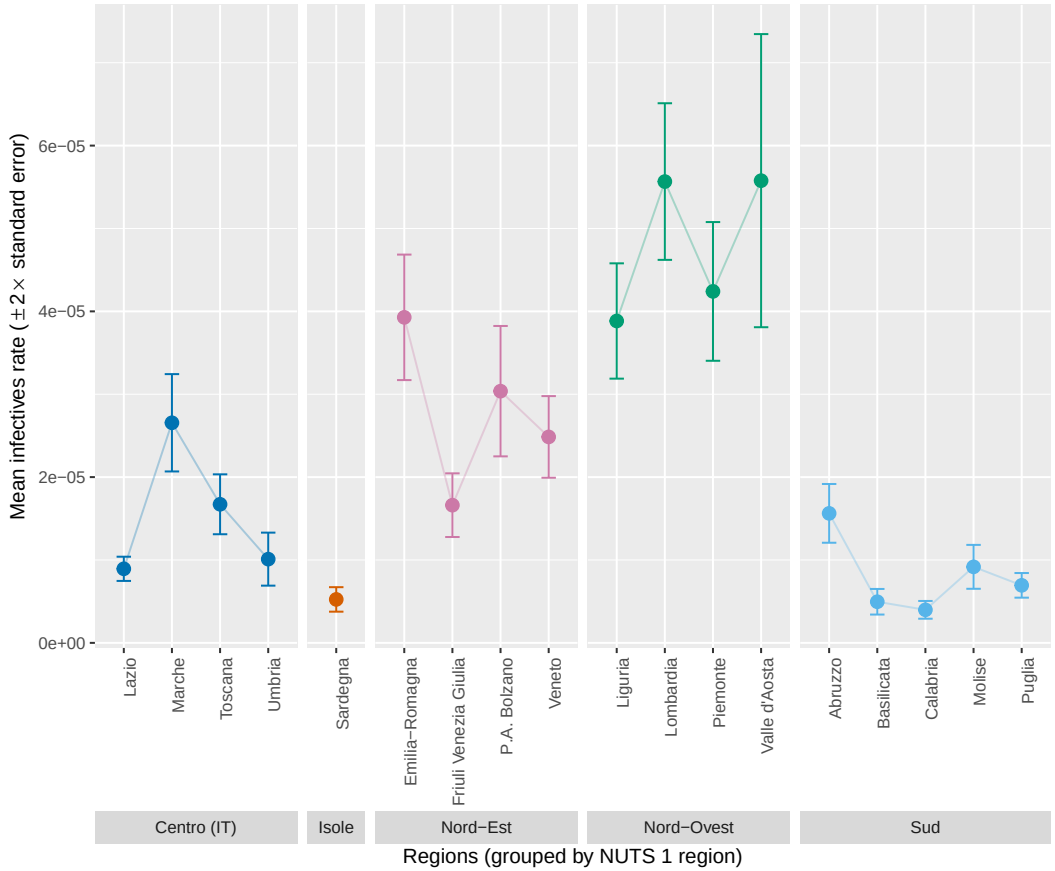


Figure 2.3. Heterogeneity across Italian regions.

From Figure 2.3, we can see that there is a large difference in the mean incidence rate over the different regions. Although we see that the mean incidence rates within the same NUTS 1 region are more similar than outside these regions, there is still a proper difference among regions, even in the same NUTS 1 region. To illustrate, consider the Nord-Est NUTS 1 region. The mean of P.A. Bolzano (Autonomous Province of Bolzano) cannot be statistically distinguished with that of Emilia-Romagna and Veneto, which can be concluded because the given error bars overlap. However, it can be distinguished from the mean incidence rate of Friuli Venezia Giulia, because these error bars do not overlap. This shows us that the heterogeneity between regions may be an issue.

In this thesis, we present several models. We are basing the models on specifications presented 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 models used to research this spread are inspired by the Standard Inflammatory Response (SIR) model but deviate from the SIR model in the sense that the variables include the number of new cases rather than the absolute number

of cases. The key additions made by Adda (2016) 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. Unfortunately, good weighting variables regarding SARS-CoV-2 are not available due to the temporal limitations of the data. Adda (2016) looks at viruses that have been appearing in society for several years and can, therefore, use weekly information and relevant instruments to quantify the infection rates, such as economic indices. Given that SARS-CoV-2 has only been appearing for around half a year, this information is not available. Consequently, this thesis only discusses the non-weighted models by Adda (2016). These models have not previously been applied to SARS-CoV-2 and can possibly show interesting insights compared to other models.

3 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 3.1. Subsequently, we look at the data on COVID-19 such as the incidence rate, reported deaths, and number of recoveries in Section 3.2. Here, we also discuss how possible errors and missing values in the data are handled. Lastly, Section 3.3 discusses the independent variables that are included in the models by Adda (2016).

3.1 Geographical Structure of Italy

In this section, we discuss the structure of Italian regions according to the NUTS classification (Nomenclature of Territorial Units for Statistics, from the French *Nomenclature des Unités Territoriales Statistiques*). This is a hierarchical system for dividing up the economic territory of the European Union and the United Kingdom (Eurostat, 2020a). Italy consists of five first-level NUTS regions (also called NUTS 1 regions), namely *Nord-Ovest* (North-West), *Nord-Est* (North-East), *Centro (IT)* (Center), *Sud* (South), and *Isole* (Islands). These larger regions are subdivided into 21 second-level NUTS regions (also called NUTS 2 regions), known as *regioni*. These *regioni* are comparable to Dutch provinces. The *regioni* of *Trentino-Alto Adige* (Trento-South Tyrol) is split into two NUTS 2 regions: *Provincia Autonoma di Bolzano/Bozen* and *Provincia Autonoma di Trento*. The third-level NUTS regions (also called NUTS 3 regions) are 107 administrative subregions of the *regioni*. Figure 3.1 presents a map of Italy with the NUTS 2 regions.[1]

[1]Source: https://www.geocurrents.info/cartography/customizable-base-maps-of-italy

Figure 3.1. Map of Italy and the NUTS 2 regions that make it up.

3.2 Coronavirus Data

In this section, we discuss the data on COVID-19 and how we handled the data processing. 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. Un-

fortunately, 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 $P = 21$ Italian regions, we retrieved the data on COVID-19 from February 25, 2020, until August 16, 2020, leading to observations for $T = 174$ days and a total number of $P \times T = 3,654$ observations. The statistics that are of interest to us are:

- New number of current positive cases (*nuovi_positivi*);
- Total number of deaths (*deceduti*);
- Total number of recoveries (*dimessi_guariti*);
- Total number of positive cases (*totale_casi*);
- Total number of tests performed (*tamponi*);
- Total number of people tested (*casi_testati*).

In addition to these variables, 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*).[2]

The data source states that the new number of current positive cases at time t, namely *nuovi_positivi*, is calculated as the first difference of the total number of positive cases: $(totale\_casi_t - totale\_casi_{t-1})$. However, in the data these two are not always equal. To illustrate, we consider the region of Abruzzo on June 16 until June 18. The daily number of positive tests $(totale\_casi_t - totale\_casi_{t-1})$ equal 1, 0, and -1, respectively, while the number of new confirmed cases (*nuovi_positivi*) equal 2, 2, and 1, respectively. This is likely a measurement or computational error. We take the first difference of the total number of positive cases to define the number of confirmed cases rather than looking at the new number of positive cases reported.

There are two variables on the tests executed. 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. Notice 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. In addition to the previous remarks, it is important to consider 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

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

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. In Section 6, we discuss how the undocumented infectives are modelled.

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. On the other hand, Belgian death counts, for instance, also include deaths of people who were suspected of having COVID-19, regardless of whether they were tested or not (Schultz, 2020). Moreover, Italian data makes no distinction between people who died because of COVID-19 and those who simply had the disease but who died from other causes (also referred to as comorbidities). Actually, only 1.2% of the Italian patients who were reported to have died because of COVID-19 until March 19, 2020 did not have a pre-existing condition (European Centre for Disease Prevention and Control, 2020a). 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 (2020a) 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 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. Some other countries, such as Germany, do make a distinction between these two groups (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 and that this differing method of counting deaths and cases only applies on a national level and not among a country's regions.

Sometimes, regions correct mistakes by having the numbers in the report on a day compensate for the errors on the days before. In the official publications that we use, data that was wrongly published on a day $t-1$ is corrected by subtracting the error from or adding the error to the cases from day t. As such, if the error is larger than the number of new cases, the reported amount of new cases is negative. It happened twenty-two times that the number of confirmed cases was reported to be negative (for eleven different regions). The number of deaths was reported to be negative eight times (for six different regions) and the number of recovered patients was reported

with a negative value 62 times (for fourteen different regions). We correct this by subtracting the error from the day before and set the previously negative number to zero. 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 ten, we propagate the error to multiple lags until this issue no longer occurs. An example for the region of Basilicata is given in Table 3.1.

Table 3.1. 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 days where the correction did not cause the number to become negative, we have no way of detecting that a correction took place and we cannot reasonably assume how this number should be split up among the days concerned. As such, these are left as is. Despite these errors, we assume that the official information is accurate and representative of the region for which it has been reported.

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 zero to five. 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 zero to two. 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.

An extreme outlier in the positive direction can be found on June 24 for the region of Trentino. A value of 387 new infectives was reported even though in the four weeks before, the maximum amount of new infectives was seven. Notably, this value is the highest of all reported values for Trentino, with the second highest value only being 172 on March 15. For the same reason as mentioned for the high negative values for Campania and Sicily, we remove the region of Trentino from our dataset. Again, another solution would be to distribute this number across the days prior.

Regarding missing values, there are none. We expect that this is because 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 surrounding July 5, however, around 250 tests were executed each day. 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, we do not deal with these outliers and leave them as is. The reason for this is twofold. Firstly, we do not know if it is actually true that a zero is being used as a filler for a missing value. It may be the case that a value of zero was actually reported. That relates to the second reason, namely that unexpectedly low values unequal to zero are also reported (such as a value of three tests being executed on July 19 for Basilicata among a usual value of around 300). As such, we cannot reasonably assume that these zeros (and which ones) pertain to missing values.

3.3 Independent Variables

In this section, we describe the independent variables, or regressors, that are included in the models by Adda (2016). Both models include a tensor X with variables that may not directly have an effect on the transmission rate. We noticed in Section 2 that there is a difference between the case of SARS-CoV-2 and the viruses investigated by Adda (2016), namely due to the time dimension. SARS-CoV-2 has only been appearing in society since December 2019 and, hence, we do not have much time-varying information, for instance on seasonality of the virus as well as economic indicators for the Italian regions. Therefore, we cannot include many variables in our tensor X. The only variable that is included is a dummy variable that denotes if the day t is on the weekend (Saturday or Sunday). We do not include an intercept because there is not some (non-zero) mean number of new cases that is persistent throughout time for a certain region.

The reason behind including the weekend dummy variable is that we expect that less people may be detected on the weekend due to 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. These people will then get tested during the week, meaning that we expect that the number of infectives during weekends will be lower. On the other hand, it is unknown whether the reported number of positive tests on a certain day is the amount of people that got tested on that day or the amount of tests that were processed on that day that turned out to be positive. The difference is that there is a time lag between people being tested and the results of that test being processed and announced. Therefore, there could be a delay of one or multiple days.

4 SIR Model

In this section, we explain the most commonly used model in epidemiology, namely the Standard Inflammatory Response (SIR) model (Anderson & May, 1992; Kermack & McKendrick, 1927). The SIR model splits the total population into three groups. S denotes the fraction of individuals who are susceptible to being infected, I denotes the fraction of individuals who are currently infected, also called infectives, and R denotes the fraction of individuals who have been removed from the model, be that because they successfully recovered from the disease or because they have deceased. We furthermore define s to be the number of susceptible individuals, i to be the number of infectives, and r to be the number of recovered individuals, so that $S = s/N$, $I = i/N$, and $R = r/N$, where N is the total population size. As such, at any point in time, we have that

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

$$s, i, r \in [0, N] \text{ and } s + i + r = N.$$

The SIR model makes four main assumptions. Understanding these assumptions also tells us how the model is constructed. The first assumption is that the population is constant, meaning that births and deaths are ignored. There exist other models in epidemiology that take both of these into account but these are not considered in this thesis due to a lack of data. The second assumption that is made under the SIR model is that there is a time-constant rate of change in infectives, proportional to the interaction between the infectives and the susceptible population. This is represented by the parameter β, also called the *transmission rate* or the *force of infection* (Keeling & Rohani, 2011). The third assumption that the SIR model makes is that there is a constant rate of change γ at which infectives recover or decease. This is a biological parameter that depends on the type of the virus and the strain, thereby not being influenced much by public health interventions (Adda, 2016).

Finally, we assume that there is a constant rate of change ω at which immune individuals lose their immunity. For instance, Adda (2016) mentions that ω is set to 0 for chickenpox as individuals acquire a lifetime immunity while ω 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. Leung (2020) estimates that COVID-19 antibodies will remain in a patient's system for two to three years, based on what is known about other coronaviruses, indicating a possibility of temporary immunity. However, it has recently become clear that reinfection is indeed possible. Two Chinese persons have been tested positive in August after having recovered from COVID-19 a few months prior (Bloomberg News, 2020).

Future research could be done to incorporate this new information. In this thesis, for simplicity's sake, we assume that lifelong immunity is achieved, or at least long enough to last through the temporal scope of our analysis: we set $\omega = 0$.

Now that the assumptions are clear, we present the definition of the SIR model. The SIR model is postulated in continuous time, i.e. the equations in (4.1), (4.2), and (4.3) depict the change in the variables S, I, and R, respectively, for one time period ahead.

$$\frac{dS}{dt} = -\beta SI + \omega R, \tag{4.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{4.2}$$

$$\frac{dR}{dt} = \gamma I - \omega R. \tag{4.3}$$

This type of model is also called a stock-and-flow model because there is a certain stock at some point in time (for instance the number of infectives) to which a flow is added and/or subtracted. Keeling and Rohani (2011) state that the SIR model can also be described by density-dependent transmission instead of frequency-dependent transmission. By that, they mean that the variables s, i, and r are used instead of S, I, and R. This is given by:

$$\frac{ds}{dt} = -\beta si + \omega r, \tag{4.4}$$

$$\frac{di}{dt} = \beta si - \gamma i, \tag{4.5}$$

$$\frac{dr}{dt} = \gamma i - \omega r. \tag{4.6}$$

One of the main measures resulting from the SIR model is the estimation of the effective reproduction number $R_{eff} := \beta/\gamma$. An epidemic is said to develop if $R_{eff} > 1$. This is clear because $R_{eff} > 1$ implies that $\beta > \gamma$, i.e. the spread of the virus exceeds the recovery rate: individuals in a society become infected more quickly than they recover. The reproduction number R_{eff} is widely used to indicate whether an ongoing epidemic is dying out. For instance, the Italian health ministry has posted an article on May 9, 2020 to communicate that the R_0, being the basic reproduction rate, for COVID-19 was below 1 in Italy, at between 0.5 and 0.7 (Ministero della Salute, 2020).

5 Model Selection

In this section, we explain how model selection is carried out. One can imagine that the same model specification does not apply to all Italian regions. For this reason, we apply model selection. We use the Akaike Information Criterion or AIC (Akaike, 1974). The AIC for a particular model is defined as

$$AIC = -2\log(ML) + 2k, \tag{5.1}$$

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

$$BIC = -2\log(ML) + k\log(n), \tag{5.2}$$

where n denotes the sample size. Both the AIC and BIC are used as the minimizer in the model selection; 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 rather than the BIC.

6 Modelling Undocumented Infectives

In this section, we discuss how our method for modelling undocumented infectives is constructed. We talk about the assumptions, the formulation, and the empirical impact that this has on the total number of infectives. A common concern with the spread of viruses, especially one spreading as rapidly as SARS-CoV-2, is that there is no possibility to test the entire population for the disease because the testing capacity is simply not there. If this were possible, then all individuals who were tested 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. These undocumented infectives were estimated to also be contagious, at a level of around 55% of the contagiousness of documented infectives (R. Li et al., 2020).

R. Li et al. (2020) carried out their research during the period from January 10 until January 23, 2020, meaning that there was a lack of major restrictions such as travel bans. The same conditions do not apply to Italy during our research period, as it was under a strict national lockdown. This lockdown 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). R. Li et al. (2020) make the important note that their 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, this research shows that undocumented infectives should be taken into account. Consequently, this thesis aims to model the undocumented infectives. However, we do not account for the lockdown and similar limiting restrictions in our model. Future research could be done to include these restrictions more robustly.

Note that, by definition, there is no data on the number 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 data on a dependent variable to predict) to determine the number of undocumented infectives. Mainly, we assume that the number of undocumented individuals is decreasing as the testing capacity increases. Similarly, the number 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.

At a point in time t, we denote the testing capacity by TC_t. In Section 3.2, we explained how a measure of the testing capacity is obtained. The total number of

infected people at time t is denoted by s_t. This group can be subdivided into the documented infectives D_t and the undocumented infectives U_t such that $D_t + U_t = s_t$. 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 earlier in this section, this proportion may change over time as the testing capacity increases. This proportion is therefore defined as a function of the testing capacity over time:

$$f_t \coloneqq f(TC_t), \tag{6.1}$$

such that

$$\begin{cases} D_t &= f_t i_t, \\ U_t &= (1 - f_t) i_t. \end{cases}$$

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

There are some properties and assumptions that (6.1) should satisfy. These are as follows:

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

(A2) If no one is tested, we assume that there is a certain minimum proportion of infectives who are documented, denoted by $f^{min} \in [0, 1]$. That means that $f(0) = f^{min}$. It should be noted that, at any point in time, it should hold that

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

so f^{min} should be chosen to be larger than $\min_t \left\{ \frac{D_t}{N_t} \right\}$. 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, so that:

$$f(N_t) = 1.$$

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, i.e. the proportion of infectives that is documented is increasing in the testing capacity. Mathematically, this means that

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

We test several forms of the function f_t. Derivations are given in Appendix E.2.

- **Linear form**

$$f_t = \frac{1 - f^{min}}{N_t} TC_t + f^{min}. \tag{6.2}$$

- **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:

$$f_t = \frac{2 - 4\gamma + 2f^{min}}{N_t^2} TC_t^2 + \frac{4\gamma - 1 - 3f^{min}}{N_t} TC_t + f^{min}. \tag{6.3}$$

  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 E.2.2, we explain why γ cannot be below $\frac{1}{4} + \frac{3}{4}f^{min}$ or above $\frac{3}{4} + \frac{1}{4}f^{min}$.

  – The first special case is the downwards opening vertex form. We assume that the vertex (the extremum) is the point $(N_t, 1)$, i.e. the parabola is downwards opening. Recall 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:

$$f_t = \frac{f^{min} - 1}{N_t^2} TC_t^2 - \frac{2(f^{min} - 1)}{N_t} TC_t + f^{min}. \tag{6.4}$$

This is equivalent to (6.3) for $\gamma = \frac{3}{4} + \frac{1}{4} f^{min}$. Therefore, this is a boundary case for a downwards opening quadratic function.

– The second special case is the upwards opening vertex form. 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:

$$f_t = \frac{1 - f^{min}}{N_t^2} TC_t^2 + f^{min}. \tag{6.5}$$

This is equivalent to (6.3) 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:

$$\begin{aligned} f(TC_t) = &\frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3N_t^3} TC_t^3 \\ &+ \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14f^{min}}{N_t^2} TC_t^2 \\ &+ \frac{1 + 32\gamma_1 - 12\gamma_2 - 21f^{min}}{3N_t} TC_t + f^{min}. \end{aligned} \tag{6.6}$$

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.

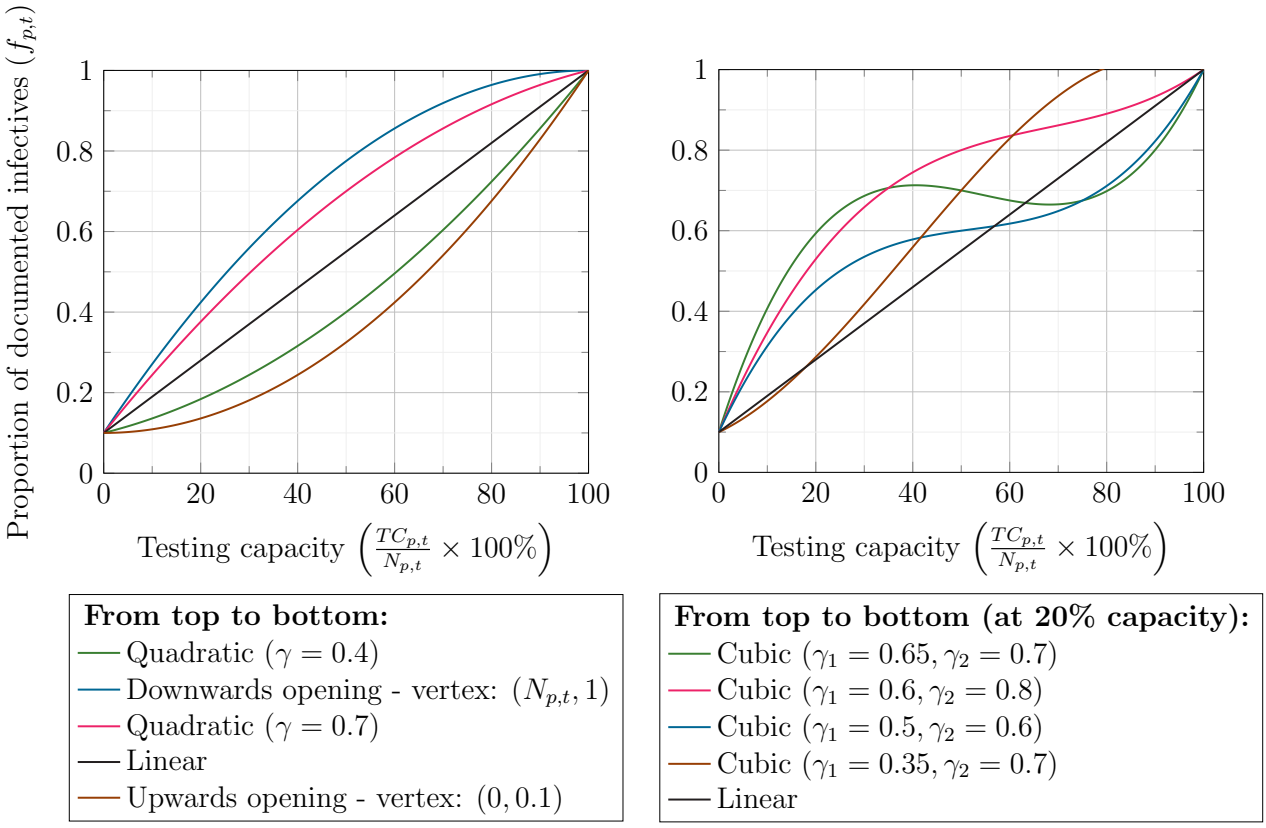
These definitions can easily be generalized to be applicable to regions by considering the total population in a region $N_{p,t}$ instead of the total population N_t. Then, the function would be dependent on p as well:

$$f_{p,t} \coloneqq f(TC_{p,t}). \tag{6.7}$$

such that

$$\begin{cases} D_{p,t} &= f_{p,t} i_{p,t}, \\ U_{p,t} &= (1 - f_{p,t}) i_{p,t}. \end{cases}$$

In Figure 6.1, we specify several functional forms for the specifications as mentioned above. Figure 6.1a shows four different functional forms for the quadratic functional forms while Figure 6.1b shows four different functional forms for the cubic specification.



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

Figure 6.1. Functional forms for the proportion of documented infectives ($f^{min} = 0.1$)

Not all of the plots in Figure 6.1 are meant to be realistic portrayals. They simply show how the functions behave as the parameters change. Moreover, recall that there are combinations of γ_1 and γ_2 for the cubic representation for which assumptions (A1) and (A4) are violated. Figure 6.1b shows that $\gamma_1 = 0.35$ and $\gamma_2 = 0.7$ cause the function to exceed the maximum value allowed for $f_{p,t}$ of 1, violating (A1). A combination of $\gamma_1 = 0.65$ and $\gamma_2 = 0.7$ creates a non-monotonic functional form, which violates (A4).

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 a theoretical rather than an empirical approach. Before that, note that the shape of the functional form may differ depending on the effective reproduction number R_{eff}, as defined in Section 2. R_{eff} estimates how many people an infective will on average infect. If $R_{eff} > 1$, 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 $R_{eff} > 1$. The reason for this is that the results from this thesis will be most important during an epidemic. Future research could be conducted into a two-step approach, where R_{eff} is estimated first so that the method of modelling undocumented infectives can be adapted accordingly.

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 follows: 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 6.1b. 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 there is the testing capacity to test the last proportion of the population, 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 do not believe that they should get tested, whether their reasons are grounded or not. Perhaps these people underestimate their symptoms or their importance. They may,

even though they are encouraged to get tested, believe that they do not need to be. For instance, these people may feel that others need to get the test more. If these infective people do not get tested, the proportion of documented infectives may level out more quickly. Moreover, these people will only turn up to the testing location if they are convinced that the testing capacity is high enough, leading to a final increase as the capacity approaches 100%.

Weighing these two specifications off, we believe that the argument in favour of a quadratic form is more general and stable, whereas the argument in favour of a cubic form is more specific. In general, of all possible fitting solutions, the one with the least number of assumptions needed is often to be preferred. Therefore, we opt to use a downwards opening quadratic functional form over a cubic form. Now that we have chosen our functional form, the question is what to choose for the parameter γ. Recall that (6.3) and (6.4) are equivalent when $\gamma = \frac{3}{4} + \frac{1}{4}f^{min}$, meaning that (6.4) is the most extreme case possible and that the slope cannot be constructed to be more steep. To be general, we choose (6.3) to be our functional form with an unknown parameter γ, denoted by $f_{p,t}(\gamma)$. A specific value for γ can then be chosen or an approach that incorporates a possibility to leave γ as an unknown parameter, such as nonlinear least squares (NLS), can be applied.

We investigate the relationship between $TC_{p,t}$ and $f_{p,t}(\gamma)$ over time and compare these across regions. Because the population size differs over the regions, this is likely to impact the absolute number of tests executed. As such, instead of comparing $f_{p,t}(\gamma)$ to $TC_{p,t}$, we compare it to $TC_{p,t}/N_{p,t}$. The results are shown in Figure 6.2. In Figure 6.2, we can see that the pattern of the relationship between the two variables is similar over time for different groups of regions; the testing capacity increases over time, while the proportion of infectives that go undocumented decreases. However, there are also clear differences across regions. Regions such as Friuli Venezia Giulia and Veneto have been testing a higher proportion of their population over time. In effect, they see a steeper decrease in the proportion of infectives that go undocumented. On the other hand, regions that do not test a large proportion of the population, such as Calabria and Apulia, see a less strong decrease in the proportion of undocumented infectives over time.

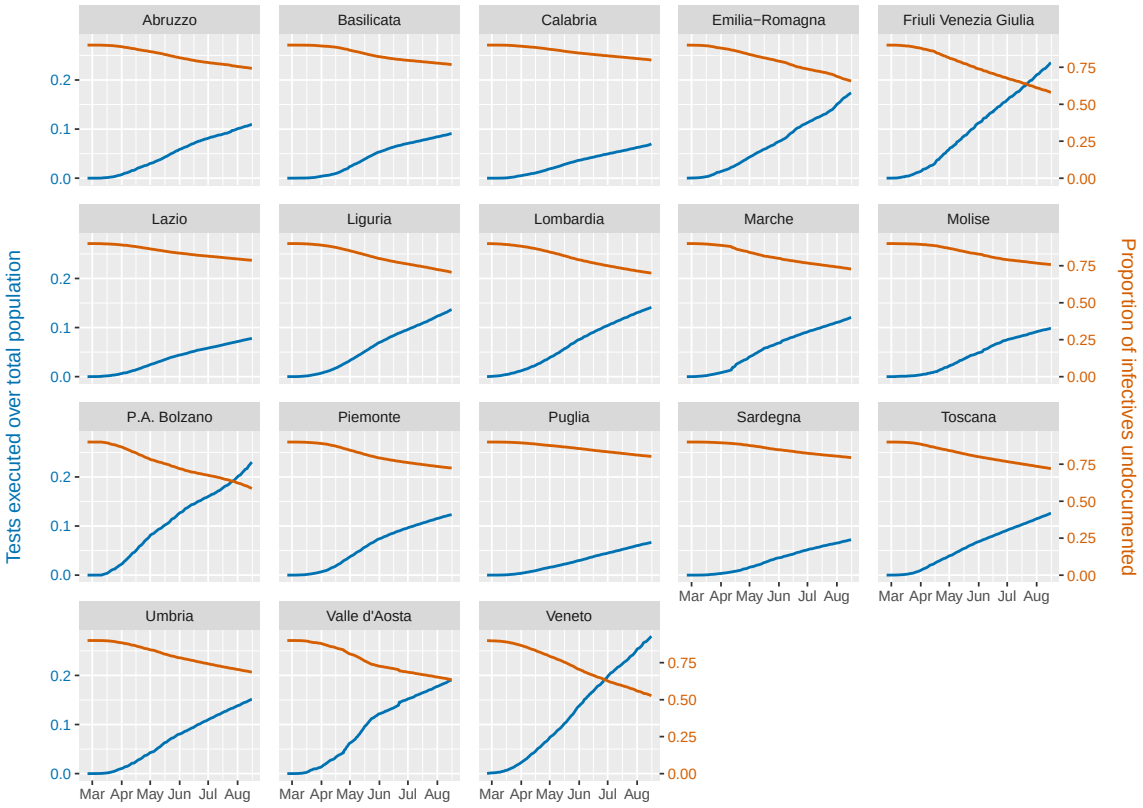


Figure 6.2. Total number of people tested over the total population ($TC_{p,t}/N_{p,t}$) versus proportion of infectives that are documented $f_{p,t}(\gamma)$. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

To illustrate the impact of this modelling method, we give an example for three regions in Table 6.1. We show the number of documented infectives $D_{r,t}$, the proportion of infectives that are documented $f_{r,t}$, and the resulting total number of infectives $i_{r,t}$, which is computed as $D_{r,t}/f_{r,t}$. For the three regions, we choose Calabria, Lombardy, and Veneto because these vary in the proportional amount of tests executed, leading to different profiles in f_t, as can be seen in Figure 6.2.

Table 6.1. Impact of modelling undocumented infectives over time. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

	Calabria			Lombardy			Veneto		
	$D_{r,t}$	$f_{r,t}$	$i_{r,t}$	$D_{r,t}$	$f_{r,t}$	$i_{r,t}$	$D_{r,t}$	$f_{r,t}$	$i_{r,t}$
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 6.1 shows us that the impact of the proportion of documented infectives f_t differs over the regions. 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. 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.

7 Within-Region Spread Model

In this section, we present the within-region spread model as presented by Adda (2016), which ignores effects across regions. Section 7.1 discusses the methodology, including the derivation of the model from the SIR model, moment conditions, and the inclusion of undocumented infectives. Subsequently, the results are presented in 7.2, where we discuss the statistical evidence, the magnitude of the estimates, and how they compare across the regions. Moreover, we investigate the progression of the estimates over time.

7.1 Methodology

In this section, we present the methodology of the within-region spread model. Recall that the SIR model is postulated in continuous time. Adda (2016) provides a discrete-time model that is based on the SIR model. Adda (2016) does not discuss how the discretization is carried out. Therefore, we discuss how the discretization appears to be carried out. Recall from (4.2) that $\frac{dI}{dt} = \beta SI - \gamma I$. As such, the discretized version (for a region p) for a single time period, without numerical integration, is:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-1} I_{p,t-1} - \gamma I_{p,t-1}. \tag{7.1}$$

There are a few things to notice. Firstly, because a model is never fully able to represent reality, we need to account for statistical errors when estimating the parameters in (7.1). This is incorporated in the model through an error term, denoted by $\eta_{p,t}$.

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-1} I_{p,t-1} - \gamma I_{p,t-1} + \eta_{p,t}. \tag{7.2}$$

Secondly, individuals that get infected do not immediately infect others because there is a so-called latent period, which is 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 chickenpox 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). Linton et al. (2020) do not report the median incubation period but instead 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 while pre-symptomatic people will develop symptoms. A key characteristic of pre-symptomatic people is that they develop a higher viral load just before said symptoms become apparent. On June 9, 2020, the World Health Organization said 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 discussed how we model pre-symptomatic individuals in the form of undocumented infectives in Section 6.

Adda (2016) models the transmission lag by making the lag on the right hand side of (7.2) dependent on the incubation period. This is denoted by the parameter τ:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-\tau} I_{p,t-\tau} - \gamma I_{p,t-1} + \eta_{p,t}. \tag{7.3}$$

Note that we have not included the lag τ in the term $\gamma I_{p,t-1}$, which represents the recovery rate. This is because when a person recovers from the disease, they immediately move to that group and this is independent of the incubation period. 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. A value of τ equal to 3 weeks is chosen for chickenpox. Due to the results from Lauer et al. (2020), Q. Li et al. (2020), and Linton et al. (2020), indicating an incubation period for COVID-19 of at most 14 days, we choose $\tau = 14$.

Adda (2016) adds regressors to the model as control variables, such as the region fixed effects, week effects and year effects in levels. 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 tensor of regressors by X. This leads to the following formulation:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-\tau} I_{p,t-\tau} - \gamma I_{p,t-1} + \delta X_{p,t} + \eta_{p,t}. \tag{7.4}$$

For our application, the data does not span multiple years. As such, we do not have year effects. Also note that week effects would capture a time trend because we do not have the same week number multiple times for the same region. Therefore, we do not include a week effect. We do add a weekend effect. More information and reasoning is provided in Section 3.3.

There are two other key differences in the model specification by Adda (2016) compared to (7.4) that are not clearly motivated in their paper. First of all, Adda (2016) replaces the incidence rate $I_{p,t-\tau}$ by the number of new cases $i_{p,t-\tau} - i_{p,t-\tau-1}$ and uses the number of new cases $i_{p,t} - i_{p,t-1}$ as the dependent variable instead of $I_{p,t} - I_{p,t-1}$. Second of all, Adda (2016) does not include the term $\gamma I_{p,t-1}$ 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. In this section and Section 8.1, we continue with this model to see how the models perform in the case of COVID-19. Even though the mathematical derivation of the model does not explicitly derive from the SIR model, it may still provide correct estimates although there is no econometric evidence for this so far.

Now we present the final model, ignoring effects across regions, by Adda (2016). Defining $\Delta i_{p,t} \coloneqq i_{p,t} - i_{p,t-1}$, the within-region spread model is given by:

$$\Delta i_{p,t} = \beta_{within} S_{p,t-\tau} \Delta i_{p,t-\tau} + \delta X_{p,t} + \eta_{p,t}. \tag{7.5}$$

One could naively 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. For good measure, we also estimate a national model. This is done in two ways, the first of which is simply by adding the values of s, i, and r of all regions together to obtain the national numbers, for which the model from (7.5) or (7.6), depending on whether undocumented infectives are modelled, is estimated. The results from this model will be labelled with National (OLS) in Section 7.2. The second method is to apply pooled OLS (POLS), which is a panel data estimation method. This model ignores the individual regional effect, hence treating the data as one large cross-section. This means that the T observations for some region p are actually treated to be cross-sectional

observations of T different individuals. These results will be labelled with National (POLS) in Section 7.2.

The model in 7.5 is estimated by ordinary least squares (OLS). The moment condition that needs to be satisfied due to the strict exogeneity assumption is:

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

A general assumption that is made, is that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. Therefore, we assume that $E\left[\eta_{p,t}\mid X_{p,t}\right]=0$. The reason why we assume that $E\left[\eta_{p,t}\mid S_{p,t-\tau}\Delta i_{p,t-\tau}\right]=0$ is that, for a large enough lag τ, the error is not correlated with past data at that lag. By that, we mean that 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. Because we chose τ to exceed the maximum estimated incubation period, we assume that this holds. Therefore, we assume that the moment condition holds.

Using the specification of undocumented infectives, we can now adapt the model to include these undocumented infectives. Using that $\Delta i_{p,t}=\frac{D_{p,t}}{f_{p,t}(\gamma)}-\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}$, the model in (7.5) becomes:

$$\frac{D_{p,t}}{f_{p,t}(\gamma)}-\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}=\beta_{within}S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)}-\frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)+\delta X_{p,t}+\eta_{p,t}. \quad (7.6)$$

We can rewrite (7.6) as follows

$$D_{p,t}=f_{p,t}(\gamma)\left(\beta_{within}S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)}-\frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)+\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}+\delta X_{p,t}+\eta_{p,t}\right).$$

The moment conditions that then need to hold are:

$$E\left[\eta_{p,t}f_{p,t}(\gamma)\left(\beta_{within}S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)}-\frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)+\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}+\delta X_{p,t}\right)\right]=0$$
$$\iff f_{p,t}(\gamma)E\left[\eta_{p,t}\left(\beta_{within}S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)}-\frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)+\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}+\delta X_{p,t}\right)\right]=0.$$

Since $f_{p,t}(\gamma)$ 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 $f_{p,t}(\gamma)$ to obtain the following moment condition:

$$E\left[\eta_{p,t}\left(\beta_{within}S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)}-\frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)+\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}+\delta X_{p,t}\right)\right]=0. \quad (7.7)$$

Just like earlier in this section, we make the assumption that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. Therefore, we assume that $E\left[\eta_{p,t} \,|\, X_{p,t}\right] = 0$. We assume that $E\left[\eta_{p,t} \,\middle|\, S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)\right] = 0$ for the same reason as earlier in this section, namely that for a large enough lag τ, the error is not correlated with past data at that lag. This is independent of the scaling functions $f_{p,t-\tau}(\gamma)$ and $f_{p,t-\tau-1}(\gamma)$ as these are constructed without the past infectives in mind.

Recall that the error $\eta_{p,t}$ is made when predicting the amount of infectives at time t. In Section 7.1, we have explained that there is a latent period during which an infected person is not able to infect others yet. As such, when considering only a one-period difference, there should be no correlation between $\eta_{p,t}$ and $D_{p,t-1}$. Moreover, as was explained in the previous paragraph, $\eta_{p,t}$ is independent of the scaling function $f_{p,t-1}(\gamma)$. Therefore, we can assume that $E\left[\eta_{p,t} \,\middle|\, \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}\right] = 0$. Consequently, the scaling of the infectives by using our functional form, has no additional impact on the moment conditions. Therefore, we assume that the moment conditions in (7.7) hold, even when modelling undocumented infectives.

To conclude this section, for all models, we use the last 100 days of data, which spans May 9 until August 16. The reason behind this is that infections from the past should not dominate the latest estimates of the transmission rate parameters. When the (first) wave has passed and the transmission rate is low, including the data from during the peak moment of the pandemic will influence the transmission rate. Consider Figure 7.1, where we present the plot of the estimated values of β_{within} over time for the Nord-Est NUTS 1 region when all T data points are used.

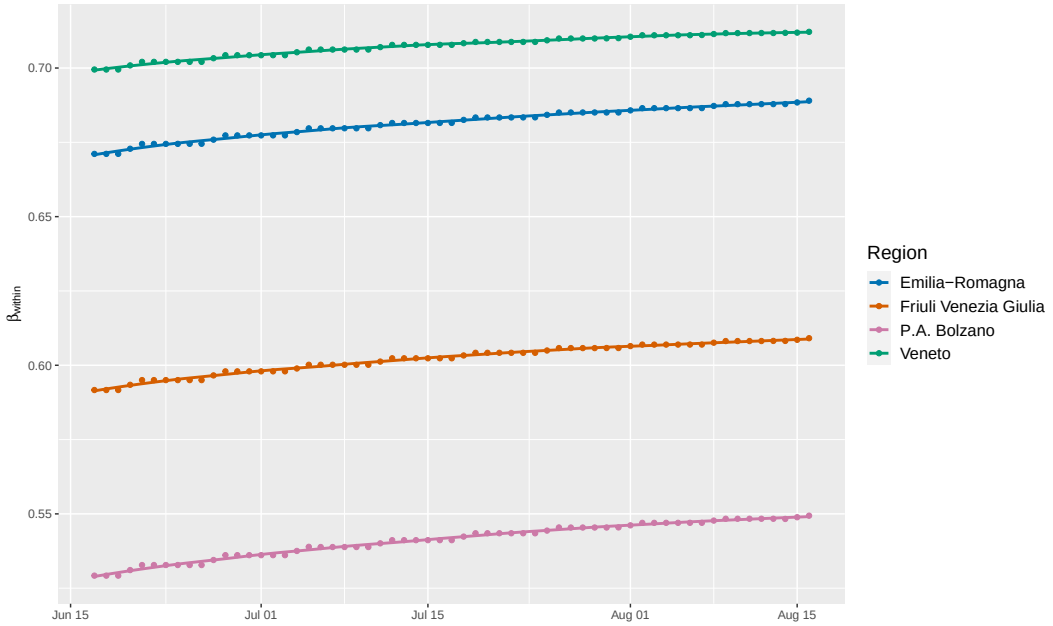


Figure 7.1. Progression of β_{within} over time for the Nord-Est NUTS 1 region using all T data points. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

It is immediately clear that the variation in β_{within} is minimal and even slightly increasing. This not intuitive for the reason that a decrease in infectives over time would lead to a lower estimated transmission parameter. As such, it is not logical that the estimate for β_{within} would level out, especially at a level not close to zero, or increase. The reason behind choosing 100 days is arbitrary, however; it is simply a number that retains enough data points while providing variation in the estimates for β_{within}.

7.2 Results

In this section, we present the results for the within-region spread model. Firstly, we present the results where the data is pooled to a national level. Subsequently, results are presented for the models per region after applying model selection 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 (6.3). If no statistical significance level is mentioned, we take a significance level of 0.05.

The results from estimating a national-level model, as mentioned at the end of the previous section, are given in Table 7.1. Model selection with AIC is not applied and undocumented infectives are not included.

Table 7.1. Estimates from the within-region spread model on a national level. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are not modelled.

	OLS				Pooled OLS			
	Estimate	Std. Error	t-value	p-value	Estimate	Std. Error	t-value	p-value
Weekend	92.860	30.241	3.071	0.003***	15.086	2.974	5.074	0.000***
β_{within}	0.493	0.022	21.992	0.000***	0.533	0.019	27.980	0.000***

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

Table 7.1 shows estimates for β_{within}, denoted by $\hat{\beta}_{within}$, of 0.493 and 0.533 for the national model estimated by OLS and POLS, respectively. Both estimated parameters are statistically significant at a 1% significance level. This seems to imply that the POLS model estimates that the transmission is a bit worse than the OLS model estimates. We apply a t-test to test whether these estimates are statistically different from one another. The null hypothesis is given by:

$$H_0 : \ \beta_{within,OLS} - \beta_{within,POLS} = 0.$$

The test statistic is constructed as follows:

$$t = \frac{\hat{\beta}_{within,OLS} - \hat{\beta}_{within,POLS}}{\sqrt{s.e._{within,OLS}^2 - s.e._{within,POLS}^2}},$$

where $s.e.$ represents the standard error. If we fill out the needed values, we find that $t = -1.376$. Since the absolute value of t does not exceed the critical value $t_{T;0.95} = 1.96$, we do not find statistical evidence in favour of the null hypothesis, at a significance level of 0.05, and we cannot say that the methods lead to different results.

As mentioned at the end of the previous section, this national model does not take into account effects specific to regions. It is also clear that the same model might not be suitable for all regions; we should apply model selection to the individual models as was explained in Section 5. To execute model selection, we use the AIC and we make sure that the term for β_{within} remains in the model, meaning that model selection is solely performed on whether the weekend dummy should be included. In Table 7.2, we present the results. In Table B.1 in Appendix B, we present the results without applying model selection. The results comparing the use of the BIC over the AIC for model selection are presented in Table B.2.

Table 7.2. Estimates from the within-region spread model per region with model selection by AIC. Estimates are given with t-statistics in parentheses. Data spans May 9 until 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}	Weekend	β_{within}	Weekend
National (OLS)	0.493***	92.860***	0.423***	379.483***
	(21.992)	(3.071)	(27.417)	(2.976)
National (POLS)	0.533***	15.086***	0.458***	90.994***
	(27.980)	(5.074)	(22.981)	(4.893)
Abruzzo	0.372***	2.751**	0.322***	13.454**
	(4.952)	(2.091)	(6.253)	(2.249)
Basilicata	0.073		0.077	
	(0.665)		(0.764)	
P.A. Bolzano	0.269***	2.346***	0.219***	7.117***
	(3.468)	(3.683)	(4.175)	(3.961)
Calabria	0.242**	2.189**	0.201*	12.897***
	(2.043)	(2.614)	(1.917)	(2.828)
Emilia-Romagna	0.358***	22.416***	0.287***	91.467***
	(11.673)	(4.844)	(14.768)	(5.266)
Friuli Venezia Giulia	0.342***	2.582***	0.252***	8.578***
	(7.559)	(3.476)	(8.891)	(3.508)
Lazio	0.494***	10.172***	0.424***	57.018***
	(10.069)	(3.441)	(11.273)	(3.610)
Liguria	0.382***	4.179	0.332***	
	(11.931)	(1.430)	(14.770)	
Lombardy	0.583***		0.513***	
	(18.205)		(18.615)	
Marche	0.436***	2.697**	0.387***	12.796**
	(9.376)	(2.144)	(11.347)	(2.286)
Molise	0.140	1.855**	0.155*	11.448**
	(1.595)	(2.456)	(1.810)	(2.299)
Piedmont	0.352***	7.449*	0.300***	28.533
	(23.006)	(1.896)	(25.913)	(1.492)
Apulia	0.376***	3.804**	0.335***	23.618***
	(6.165)	(2.581)	(7.302)	(2.719)
Sardinia	0.437***		0.276***	5.301
	(4.417)		(3.874)	(1.655)
Tuscany	0.281***	8.444***	0.237***	35.079***
	(6.666)	(3.854)	(8.744)	(3.929)
Umbria	0.436***	1.428***	0.291***	5.402***
	(3.936)	(3.168)	(4.134)	(3.351)
Aosta Valley	0.290***		0.302***	
	(4.341)		(6.394)	
Veneto	0.489***	11.951*	0.284***	39.380***
	(6.748)	(1.920)	(7.344)	(2.672)

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

We first discuss the statistical evidence for the estimated values of β_{within}, where we start by considering the model excluding undocumented infectives. After this, we consider the interpretation of the coefficients and the model selection. Considering the estimates of β_{within}, we see that these differ vastly over the regions with varying degrees of statistical significance. For only two regions, we find no statistically significant result, namely Basilicata and Molise. This may be due to not enough regional variation but the results may still be significant for policy purposes even if they are not statistically significant. For the region of Calabria, the estimate of β_{within} is statistically significant at a significance level of 0.05. For all other regions, the estimates are statistically significant at a significance level of 0.01. Looking only at the statistically significant estimates, these range from 0.242 for Calabria until 0.583 for Lombardy. This already shows that a national model should not be applied to individual regions, despite the statistical significance of the estimates of β_{within} for the national models.

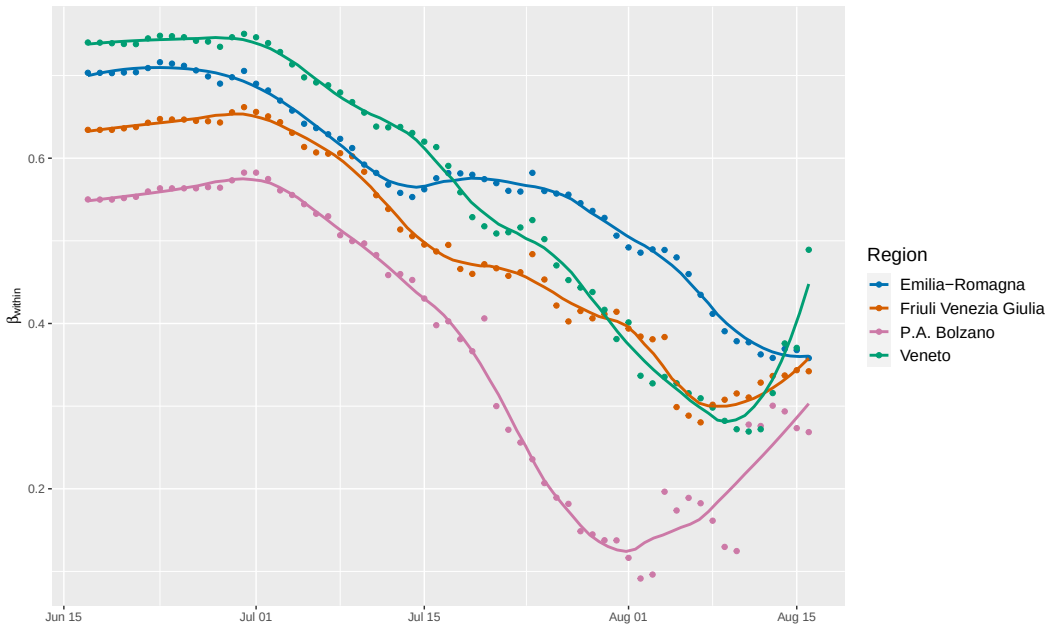
If we model undocumented infectives, we again see that no statistical evidence is found for a nonzero transmission rate for Basilicata and Molise, although the latter is now significant at a significance level of 0.1. Moreover, the estimate for Calabria is not significant anymore at a significance level of 0.05. For all other regions, the estimate is statistically significant at a significance level of 0.01, ranging from 0.219 for Bolzano until 0.513 for Lombardy. Notice that modelling undocumented infectives does not seem to impact the order of magnitude between the regions. For instance, where Lazio has the second-highest estimate for the case when undocumented infectives are not modelled, excluding the national models, it retains that position when undocumented infectives are included. Moreover, notice that the estimates for all regions, excluding the statistically insignificant ones, are all lower. Therefore, it appears that modelling undocumented infectives tends to decrease the estimates of the within-region transmission rate β_{within}.

We are again interested in investigating whether the differences in the estimates differ significantly. For this, recall that the standard error of the estimate can be computed by dividing the estimate by the t-statistic. The t-statistic for the national models are equal to 2.572 (OLS) and 2.720 (POLS), thereby exceeding the critical value $t_{T;0.95} = 1.96$ and indicating that modelling undocumented infectives yields statistically different results at a significance level of 0.05. Besides the national models, there is only one other region for which the estimates differ significantly, namely Veneto, which has a t-statistic of 2.496. It is noteworthy to mention that the region of Emilia-Romagna has a t-statistic of 1.955, thereby barely not exceeding the critical value. With this t-statistic, the estimates are statistically significant different at a significance level of 5.06%. This can be computed by using the density function of the Student's t-distribution.

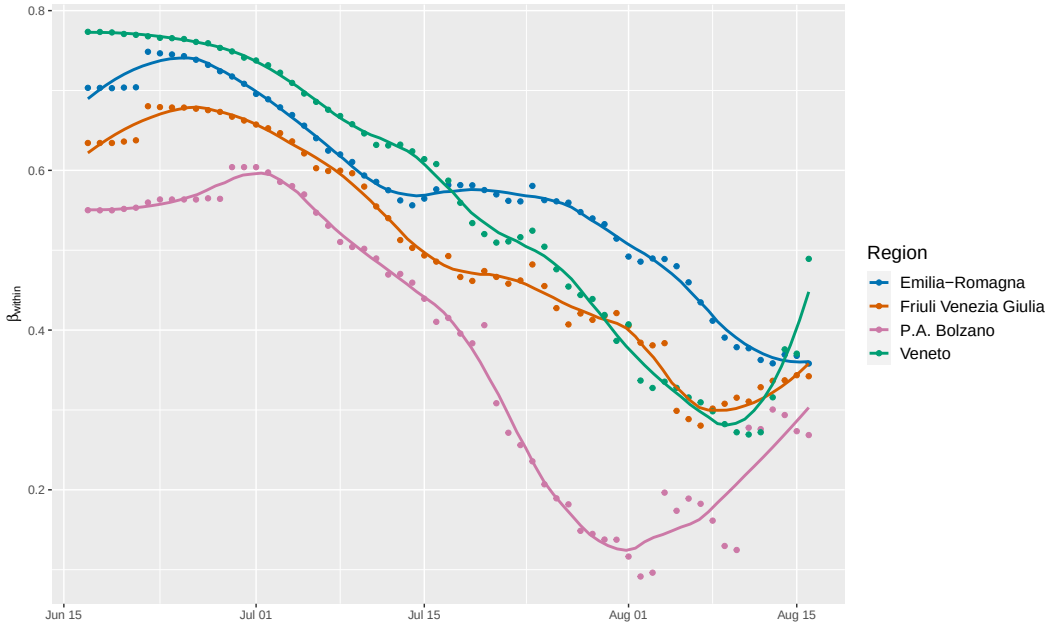
When we want to interpret the estimates of β_{within}, we should recall that Adda (2016) notices that these can be interpreted as the marginal effects of a change in the infection rate on the future infection rate when the entire population is susceptible to the disease. However, notice that this is made with the model formulation as explained in Section 7.1, where the removal term is omitted from the model. Because this does not take into account the specific removal rate of COVID-19, we cannot interpret the coefficients in the same way. However, we can compare the magnitude of the coefficients with one another. For instance, let us compare the regions of Lombardy and the island of Sardinia. We consider the model where undocumented infectives are modelled. For this model, we find estimates of β_{within} of 0.513 and 0.276, respectively. Although this cannot tell us much about the spread within the region as explicitly, this does show us that the transmission in Lombardy was much more severe than on Sardinia. A similar interpretation can be applied to any comparison of regions and for the model without modelling undocumented infectives.

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 generally the same whether undocumented infectives are modelled are not. The only two exceptions that we see are for the region of Liguria and the island of Sardinia. When excluding undocumented infectives, the weekend dummy is included in the model for Liguria, whereas it is excluded when undocumented infectives are modelled. For Sardinia, it is the other way around. We do see that some of the estimates for the weekend dummy's parameter are not statistically significant at a significance level of 0.05. As mentioned in Section 5, this is because the AIC tends to select a larger model. Regardless of whether undocumented infectives are included or not, the entire model is selected for sixteen out of twenty models, and, in the other four cases, the weekend dummy is excluded.

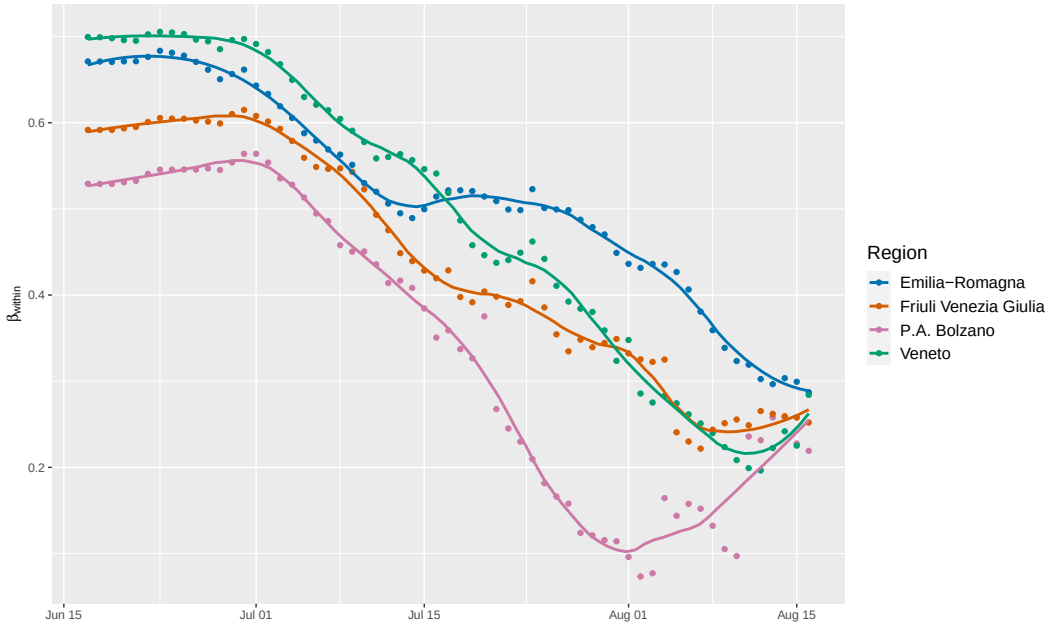
We are also interested in looking at the estimate of β_{within} over time. We expect that it decreases over time, implying that SARS-CoV-2 is transmitted less. In Figure 7.2, we present plots for the regions in the Nord-Est NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.2, which generally show similar results. Each point in the graphs in Figure 7.2 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.



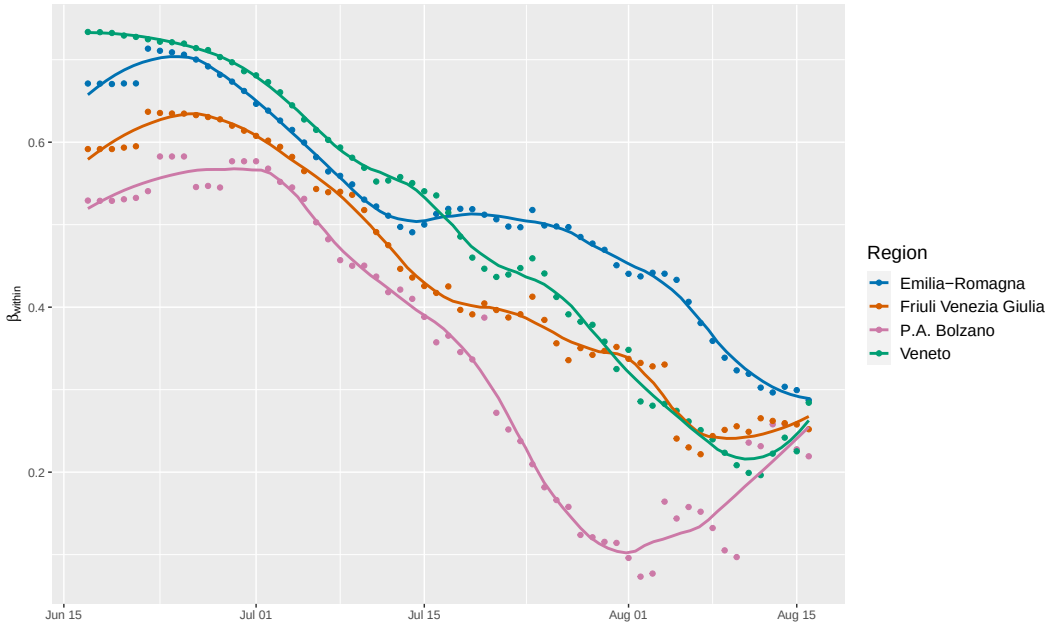
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



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

Figure 7.2. Progression of β_{within} over time for the Nord-Est NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Considering the progression of β_{within} over time, we indeed see that it decreases over time, as we expected. We do see a slight increase in the estimates of β_{within} towards the end of the timespan. This is likely because the amount of infectives increased a bit over time again for multiple regions from mid July onward. Please consider the Figures in Appendix C.1, which indeed illustrate this increase. We also notice that the pattern of β_{within} over time is similar when comparing the models excluding and including undocumented infectives; the difference can be found in the value of β_{within} which, as explained earlier in this section, tends to be lower when modelling undocumented infectives.

8 Within and Between-Region Spread Model

In this section, we present the model by Adda (2016) that takes effects across regions into account. Section 8.1 discusses the methodology, including the model formulation, estimation method, moment conditions, and the inclusion of undocumented infectives. Subsequently, the results are presented in 8.2, where we discuss the statistical evidence, the magnitude of the estimates, and how they compare across the regions. Moreover, we investigate the progression of the estimates over time.

8.1 Methodology

In this section, we present the methodology of the within and between-region spread model. A key addition made by Adda (2016) is recognizing that there is spatial spillover between regions; there may be infectives in one region that travel to another region and then infect individuals there. The following model is defined:

$$\Delta i_{p,t} = \beta_{within} S_{p,t-\tau} \Delta i_{p,t-\tau} + \beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \Delta i_{c,t-\tau} + \delta X_{p,t} + \eta_{p,t}. \tag{8.1}$$

It should be noted that the specification in (8.1) assumes that individuals from all regions are able to meet one another at the same rate. Of course, this assumption is likely not satisfied. Consider, for example, the region of Lombardy, which lies in north-west Italy. Inhabitants of Lombardy are much more likely to travel to bordering regions, such as Piedmont or Veneto, than to regions in the far south, such as Campania or Apulia, or to the islands. As such, it would be better to consider introducing a method by which we only take a certain number of regions that are the closest to another region into account. Another criterion could be to look at economic ties, since SARS-CoV-2 can not only be transmitted by regular civilians meeting each other but also by the exchange of goods, for example. Spatiotemporal models exist that could be applied when a suitable matrix of weighting measures is available. Nonetheless, in this section, we follow the specification that Adda (2016) provides as in (8.1) and explain the other criteria as possible future research in Section 10.

Notice, moreover, that it does not make sense to consider a national model. Because we do not consider countries outside of Italy, the set $R \backslash r$ 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.

In (8.1), the transmission parameter β is now allowed to be different within and between regions. Adda (2016) estimates (8.1) 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. Unfortunately, we do not have sufficient information on the effect of the weather on the virus. 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. Using a spatiotemporal analysis, Briz-Redón and Serrano-Aroca (2020) even show that no evidence of a relationship between COVID-19 cases and temperature was found, although these results should be interpreted carefully due to data uncertainty and confounders. For these reasons, we only consider OLS for this model.

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

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

In the same way as in Section 7.1, we can assume that $E\left[\eta_{p,t} \mid X_{p,t}\right] = 0$ and $E\left[\eta_{p,t} \mid S_{p,t-\tau}\Delta i_{p,t-\tau}\right] = 0$. Following the same reasoning as before, we assume that the number of infectives who come into contact with susceptible people in other regions at a certain time is not correlated with the error if the lag is large enough, therefore assuming that $E\left[\eta_{p,t} \,\middle|\, \beta_{between} S_{p,t-\tau}\sum_{c\in R\backslash r}\Delta i_{c,t-\tau}\right] = 0$. As such, we assume that the moment condition holds.

Using the specification of undocumented infectives, we can adapt the within and between-region spread model to include these undocumented infectives as well. Using that $\Delta i_{p,t} = \frac{D_{p,t}}{f_{p,t}(\gamma)} - \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}$, the model in (8.1) becomes:

$$\begin{aligned}\frac{D_{p,t}}{f_{p,t}(\gamma)} - \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)} =& \beta_{within} S_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \\ &+ \beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right) \\ &+ \delta X_{p,t} + \eta_{p,t}.\end{aligned} \tag{8.2}$$

We can rewrite (8.2) as follows

$$\begin{aligned}D_{p,t} = f_{p,t}(\gamma) \Bigg(& \beta_{within} S_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \\ &+ \beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right) \\ &+ \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)} + \delta X_{p,t} + \eta_{p,t} \Bigg).\end{aligned}$$

The moment conditions that then need to hold are:

$$\begin{aligned}E \Bigg[f_{p,t}(\gamma) \eta_{p,t} \Bigg(& \beta_{within} S_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \\ &+ \beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right) \\ &+ \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)} + \delta X_{p,t} \Bigg) \Bigg] = 0.\end{aligned} \tag{8.3}$$

Since $f_{p,t}(\gamma)$ 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 $f_{p,t}(\gamma)$ to obtain the following moment condition:

$$\begin{aligned}E \Bigg[\eta_{p,t} \Bigg(& \beta_{within} S_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \\ &+ \beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right) \\ &+ \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)} + \delta X_{p,t} \Bigg) \Bigg] = 0.\end{aligned}$$

Just like before, we make the assumption that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. Therefore, we assume that $E\left[\eta_{p,t} \mid X_{p,t}\right] = 0$. Now note that there are three additional terms to consider, namely the relation between $\eta_{p,t}$ and $S_{p,t-\tau}\left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)}\right)$, $S_{p,t-\tau}\sum_{c\in R\setminus r}\left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)}\right)$, and $\frac{D_{p,t-1}}{f_{p,t-1}(\gamma)}$, the first and last of which have been discussed in Section 8.1. That leaves the middle term, for which the reasoning is identical to the first term: for a large enough lag τ, the error is not correlated with past data at that lag, which is independent of the scaling functions $f_{p,t-\tau}(\gamma)$ and $f_{p,t-\tau-1}(\gamma)$ as these are constructed without the past infectives in mind. We also say that it does not matter whether we consider infectives within the region or in other regions, as the longer time lag applies in any case. Therefore, we assume that the moment conditions in (8.3) hold, even when modelling undocumented infectives.

8.2 Results

In this section, we present the results for the within and between-region spread model. If no statistical significance level is mentioned, we take a significance level of 0.05. Once again, we use the last 100 days of data, which spans May 9 until August 16.

In Table 8.1, we present the results where we execute model selection using the AIC and we make sure that the terms for β_{within} and $\beta_{between}$ remain in the model, meaning that model selection is solely performed on whether the weekend dummy should be included. In Table B.3 in Appendix B, we present the results without applying model selection. The results comparing the use of the BIC over the AIC for model selection are presented in Table B.4.

Table 8.1. Estimates from the 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 until 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}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Abruzzo	−0.011	5.871×10^{-3}***	2.394*	−0.047	5.943×10^{-3}***	12.494**
	(−0.079)	(3.399)	(1.909)	(−0.495)	(4.455)	(2.279)
Basilicata	0.040	7.638×10^{-4}		0.034	6.299×10^{-3}	
	(0.357)	(1.286)		(0.321)	(1.379)	
P.A. Bolzano	0.246**	2.043×10^{-4}	2.305***	0.201**	1.018×10^{-4}	7.030***
	(2.355)	(0.317)	(3.534)	(2.559)	(0.310)	(3.848)
Calabria	0.186	7.936×10^{-4}	1.893**	0.128	8.270×10^{-4}	11.514**
	(1.460)	(1.171)	(2.168)	(1.088)	(1.341)	(2.472)
Emilia-Romagna	0.365***	-8.645×10^{-4}	22.410***	0.171***	0.015*	90.384***
	(4.184)	(−0.080)	(4.817)	(2.635)	(1.873)	(5.267)

Table 8.1 continues on next page

Table 8.1 continued from previous page

	Regular model			Modelling undocumented infectives		
Region	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Friuli Venezia Giulia	0.114*	3.701×10^{-3}***	1.759**	0.034	3.026×10^{-3}***	4.330**
	(1.707)	(4.346)	(2.481)	(0.929)	(7.601)	(2.139)
Lazio	0.342***	7.364×10^{-3}	10.573***	0.238**	0.010*	61.107***
	(2.671)	(1.283)	(3.568)	(2.221)	(1.853)	(3.879)
Liguria	−0.026	0.033***		−0.023	0.029***	
	(−0.372)	(6.515)		(−0.404)	(6.613)	
Lombardy	0.329***	0.226***		0.285***	0.190***	
	(4.929)	(4.257)		(4.463)	(3.926)	
Marche	0.245**	4.505×10^{-3}**	2.150*	0.181**	4.644×10^{-3}***	10.483*
	(2.521)	(2.229)	(1.710)	(2.312)	(2.907)	(1.921)
Molise	−0.085	4.063×10^{-3}***		−0.077	4.596×10^{-3}***	
	(−1.168)	(8.821)		(−1.158)	(9.811)	
Piedmont	0.104***	0.065***		0.091***	0.057***	
	(3.665)	(9.767)		(3.882)	(9.785)	
Apulia	0.120	5.525×10^{-3}***	3.083**	0.035	8.078×10^{-3}***	18.908**
	(1.051)	(2.629)	(2.115)	(0.393)	(3.896)	(2.303)
Sardinia	0.407***	1.846×10^{-3}		0.252**	1.831×10^{-4}	5.062
	(2.903)	(0.299)		(2.468)	(0.321)	(1.532)
Tuscany	0.051	0.010***	8.023***	0.030	9.454×10^{-3}***	33.901***
	(0.605)	(3.107)	(3.812)	(0.535)	(4.141)	(4.095)
Umbria	0.427***	4.689×10^{-5}	1.410***	0.235**	2.449×10^{-4}	4.877***
	(2.981)	(0.102)	(2.904)	(2.467)	(0.876)	(2.832)
Aosta Valley	0.023	1.111×10^{-3}***		0.055	7.104×10^{-4}***	
	(0.259)	(4.193)		(0.775)	(4.374)	
Veneto	0.904***	−0.033***	16.184***	0.434***	-7.954×10^{-3}	43.868***
	(6.990)	(−3.771)	(2.721)	(4.370)	(−1.637)	(2.951)

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

We first discuss the statistical evidence for the estimated values of β_{within} and $\beta_{between}$ and how these interact. After this, we consider the interpretation of the coefficients and the model selection. Before that, there are two things to notice to start with. Firstly, notice that the estimates for $\beta_{between}$ are generally much smaller than the estimates for β_{within}. 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. A notable exception being the region of Lombardy, where the estimate for $\beta_{between}$, namely 0.378, is indeed smaller than the estimate for β_{within}, namely 0.514, but the two estimates are much closer than for other regions. A similar result can be found for the region of Piedmont. For the region of Lombardy, the reason for the estimates being much closer is twofold. Firstly, because it is the largest region in Italy, housing around one-sixth (16.67%) of the total population of Italy, where the second-largest region is Lazio, which houses 9.74% of the total amount of Italians. Secondly, Lombardy was hit the hardest by SARS-CoV-2 of all of the

Italian regions and, therefore, the number of infectives there is much higher than in other regions. As such, summing the other regions has a proportionally smaller effect.

The second matter to be noticed is that there are negative values of the estimates of β_{within} and $\beta_{between}$, which happens three and two out of eighteen times, respectively, for the regular model. For the model including undocumented infectives, this happens three times (for the same three regions) and once (the estimate for Emilia-Romagna is now positive), respectively. This should not be possible because this means that when infectives meet susceptible people, the incidence rate decreases. Luckily, these estimates are generally not statistically significant, with an exception for the estimate of $\beta_{between}$ for Veneto in the regular model, for which the estimate is significant at a significance level of 0.01. Future research can be done into models that restrict the estimates to be positive.

On the topic of statistical significance, recall that there was only one region that did not have a statistically significant estimate for β_{within} for the within-region spread model at a significance level of 0.05, namely Basilicata. For the within and between-region spread model, we also find no statistically significant results for Basilicata, neither for β_{within} and $\beta_{between}$. This also happens for the region of Calabria. In addition, nine and eleven out of eighteen regions find a significant estimate for β_{within} and $\beta_{between}$, respectively, when excluding undocumented infectives. For the model including undocumented infectives, ten regions find a significant result for $\beta_{between}$, the difference being that the (negative) result for Veneto is no longer significant. The same nine regions as for the regular model once again find a significant result for β_{within}.

However, if a significant result for $\beta_{between}$ is found, this does not necessarily go hand-in-hand with a significant estimate for β_{within} or vice versa. For the regular model, it only happens four times that the estimates are jointly significant, namely for Lombardy, Marche, Piedmont, and Veneto. For the model including undocumented infectives, this happens only three times, namely for the same regions excluding Veneto.

Looking only at the statistically significant estimates of β_{within}, these range from 0.104 for Piedmont until 0.904 for Veneto for the regular model and from 0.091 for Piedmont until 0.434 for Veneto when undocumented infectives are included. Excluding Veneto, the highest values for the regular model and the model including undocumented infectives are 0.427 for Umbria and 0.285 for Lombardy, respectively. Therefore, we notice that the estimates for Veneto are outliers and that the values are a bit lower than for the within-region model. When considering the statistically significant estimates of $\beta_{between}$, we see that these range from 1.111×10^{-3} for Aosta Valley until 0.226 for Lombardy for the regular model and from 7.104×10^{-4} for

Aosta Valley until 0.190 for Lombardy when undocumented infectives are included. Conclusively, we see that the estimates of $\beta_{between}$ differ much less over the regions than those of β_{within}.

One aspect to pay attention to is regarding the impact of modelling undocumented infectives on the estimates. Notice that modelling undocumented infectives again causes the (significant) estimates of β_{within} to be lower than for the regular model, just like for the within-region spread model. For $\beta_{between}$, however, this is not the case. For four out of ten cases, namely for Abruzzo, Marche, Molise, and Apulia, where both estimates of $\beta_{between}$ are significant, the estimate when modelling undocumented infectives is higher. We believe that this is likely the case because all regions are affected in a similar way by the modelling method of undocumented infectives, although the regions that test more rigorously are of course affected less. As such, there is no uniform way in which the estimate is affected. Also consider Figure 6.2, where we plotted the testing capacity versus the proportion of undocumented infectives for our modelling method. We do not see that a certain pattern happens for the four aforementioned regions that does not happen for other regions that find significant estimates of $\beta_{between}$, such as Lombardy, Piedmont, and Tuscany.

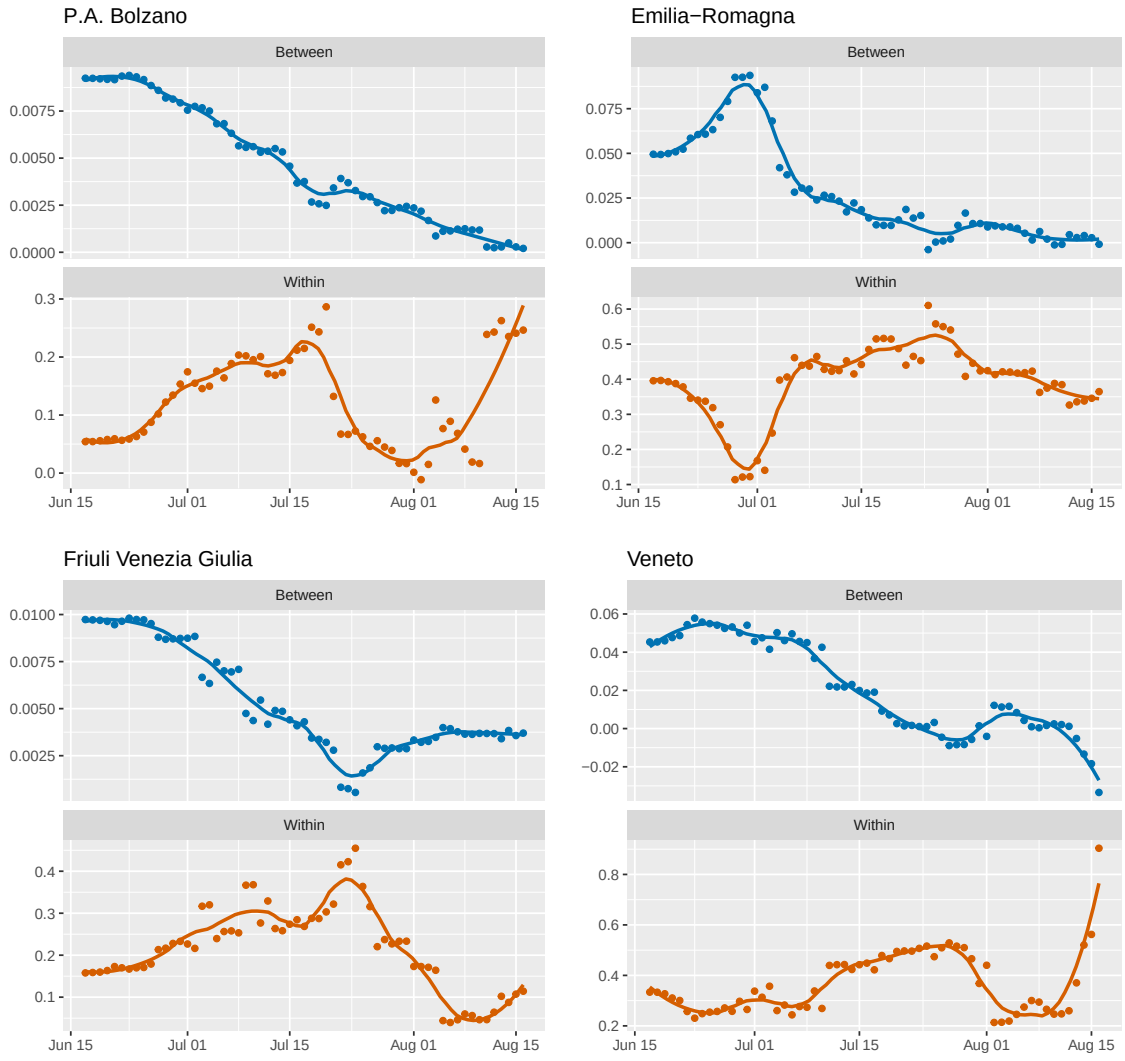
To investigate whether the differences in the estimates differ significantly, we consider the regions of Lombardy, Marche, and Piedmont since these are the only regions that find statistically significant results for all four estimates. For β_{within}, we find t-statistics of 0.476, 0.513, and 0.353, respectively. None of these exceed the critical value of $t_{T;0.95} = 1.96$, meaning that modelling undocumented infectives does not cause a significant difference in these estimates. There is one exception. For Veneto, the t-statistic of 2.882 does exceed the critical value. For $\beta_{between}$, we find t-statistics of 0.501, -0.054, and 0.905, respectively. Once again, no statistical difference can be found. One can check that none of the estimates for $\beta_{between}$, when both are significant, differ significantly.

Recall that we cannot interpret the coefficients in the same way as Adda (2016) does but that we can compare the magnitude of the coefficients with one another across regions. For instance, let us compare the regions of Lombardy and Piedmont. For both regions, all estimates are statistically significant. We consider the model including undocumented infectives. We find estimates of β_{within} of 0.285 and 0.091 and estimates of $\beta_{between}$ of 0.190 and 0.057 for Lombardy and Piedmont, respectively. We can conclude that the transmission within the region as well as between regions was worse in Lombardy compared to Piedmont, although we cannot explicitly interpret the magnitude of that transmission.

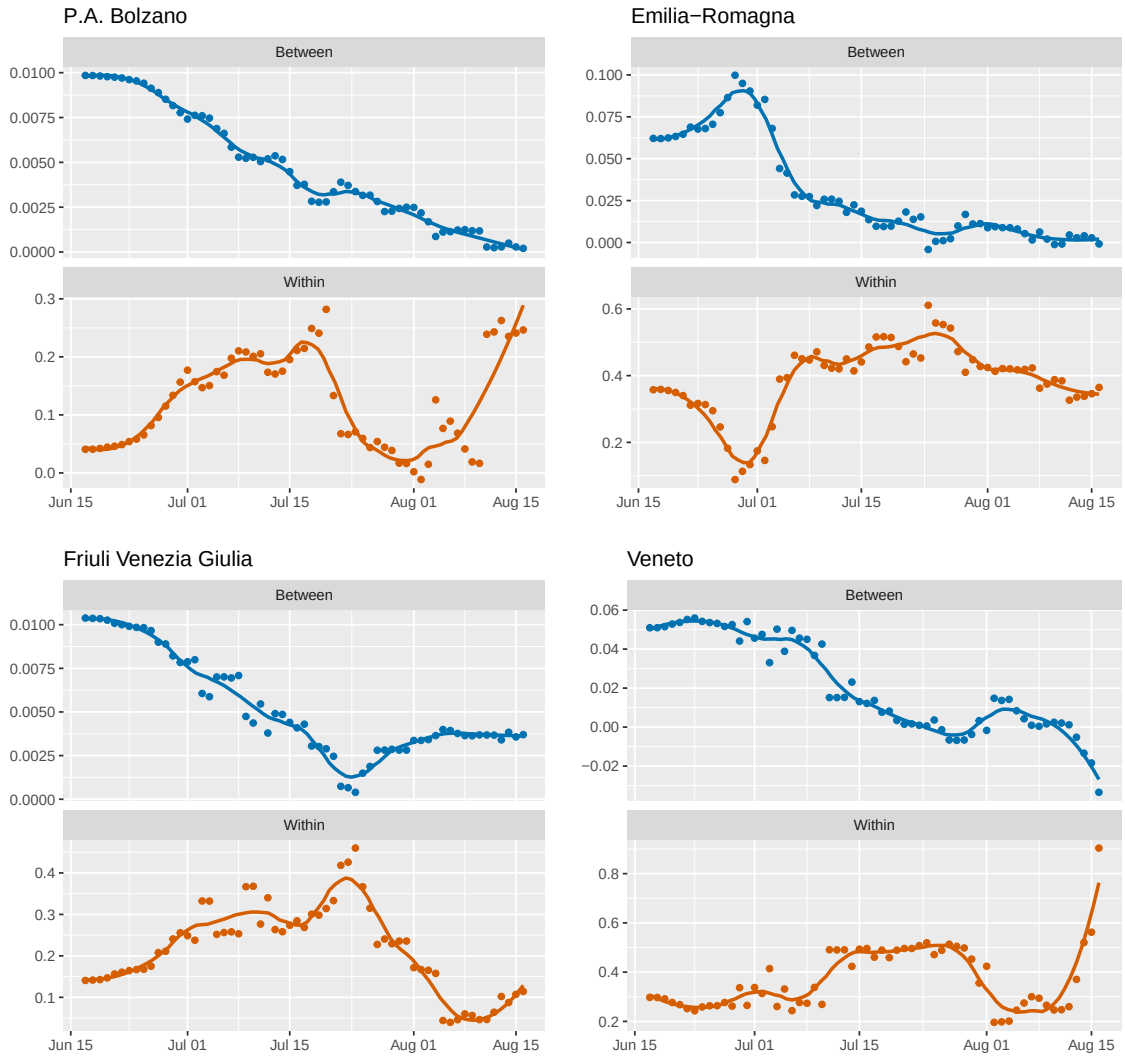
Regarding model selection, the set of regressors that is used is the same whether undocumented infectives are modelled are not, except for the island of Sardinia. All

models retain the terms related to β_{within} and $\beta_{between}$ in the model. For Sardinia, the weekend dummy is excluded in the regular model but it is included when undocumented infectives are modelled. In eleven out of the remaining seventeen cases, the entire model is selected. In the other six cases, the weekend dummy is excluded.

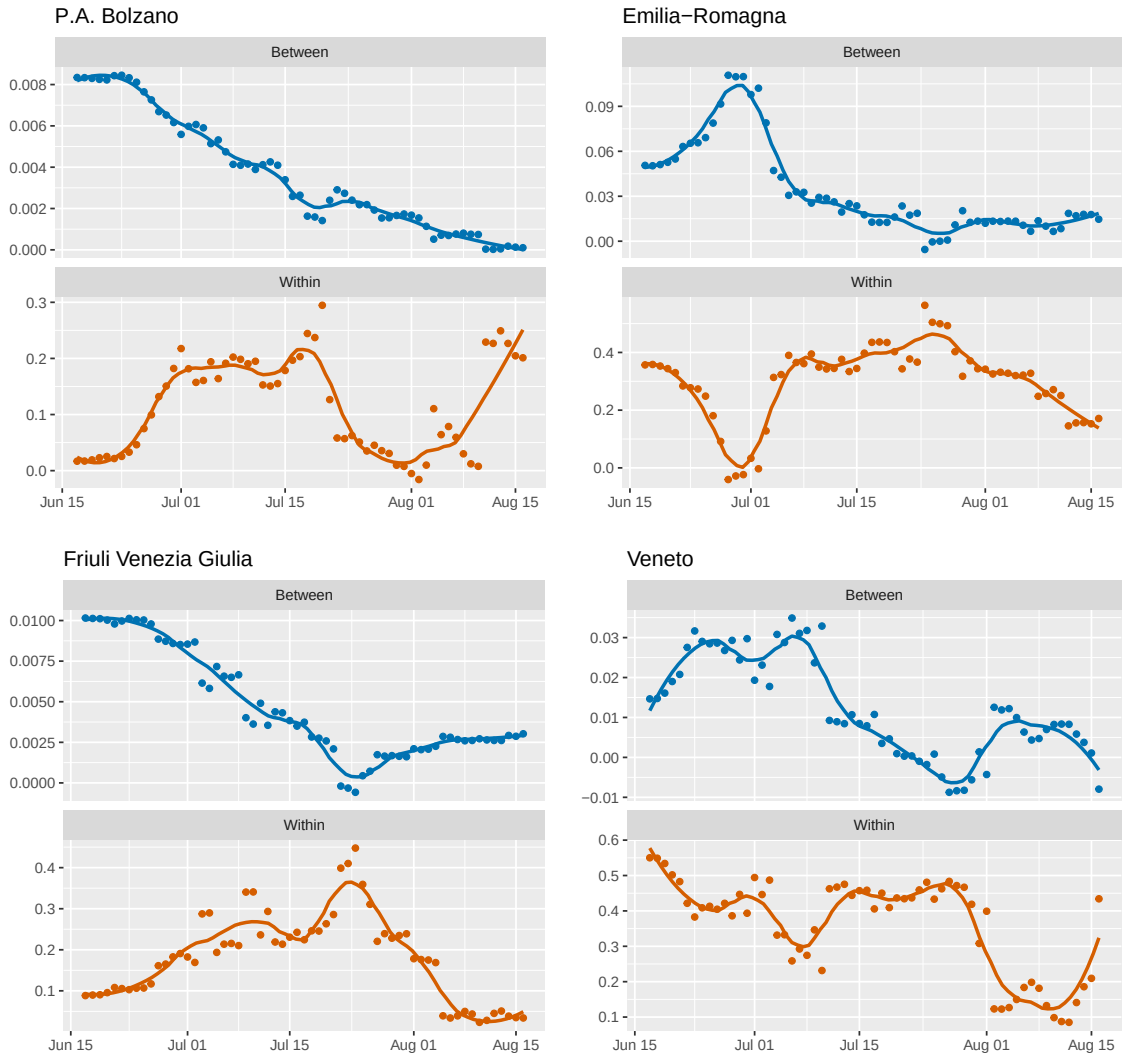
To conclude, we are again interested in looking at the estimates of β_{within} and $\beta_{between}$ over time. In Figure 8.1, we present plots for the regions in the Nord-Est NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.3. Each point in the graphs in Figure 8.1 is the 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.



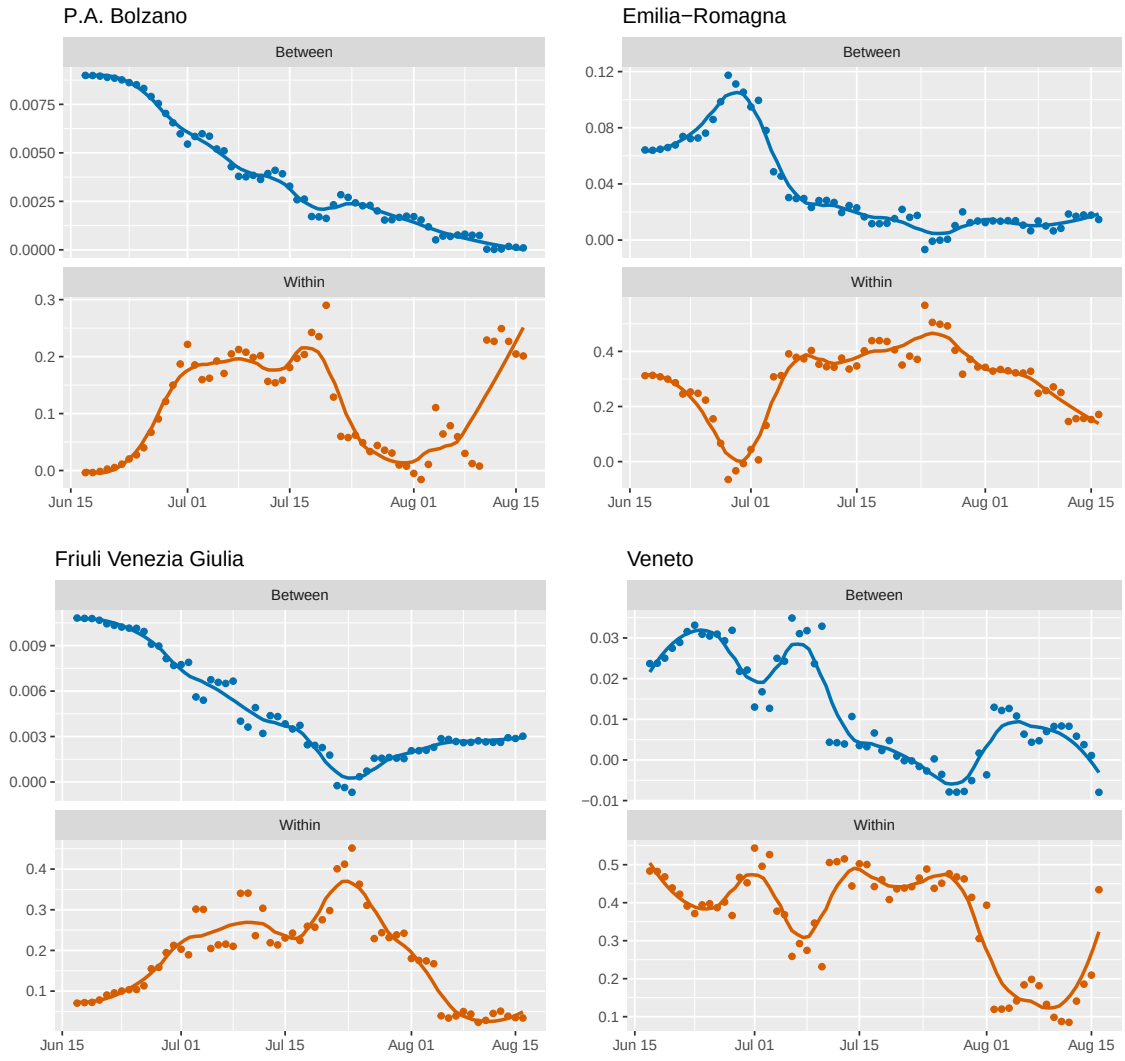
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



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

Figure 8.1. Progression of β_{within} and $\beta_{between}$ over time for the Nord-Est NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Figure 8.1 shows very different patterns compared to the within-region spread model, as discussed in the previous section. Generally, we do not see a decreasing movement for β_{within}. In contrast, the values seem to fluctuate. For $\beta_{between}$, we also see some fluctuation but we do see a decreasing movement. For some regions, we even see that β_{within} and $\beta_{between}$ seem to move in opposite directions. For example, consider Figure 8.1d where model selection is applied and undocumented infectives are included. For Emilia-Romagna and Veneto, this observation is immediately visible.

For Friuli Venezia Giulia, it also seems to hold, although it is less pronounced. The transmission rates for Bolzano seem to not adhere to this observation. For the other NUTS 1 regions, we see some similar patterns although the progression varies a lot over the regions. For instance, the estimate for $\beta_{between}$ for the region of Lazio tends to increase over time, as can be seen in Figure C.10.

Conclusively, it seems like neither model selection by AIC and modelling undocumented infectives have a large impact on the profiles of the estimates of β_{within} and $\beta_{between}$ much, besides the impact on the magnitude of the values, as explained in the discussion of Table 8.1.

9 Conclusion

In this thesis, we have explored methods to model the transmission rate of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), in Italy. This deadly viral disease has had a hold on the entire globe since December 2019. In its raze, it has infected around 30 million people worldwide and caused almost a million deaths so far. The goal of this thesis was apply models which recognize that the Italian regions are not homogeneous such that the regional variation and spatial spillover effects need to be taken into account. This was done through the models presented by Adda (2016). We also aimed to come up with a modelling method to include undocumented infectives. This was done by constructing a functional form for the proportion of total infectives that are documented.

This thesis has shown that there is a vast difference between the transmission rates across regions in both models by Adda (2016). When not taking spatial spillover into account, it became clear that the region of Lombardy sported the highest within-region transmission rate. The results that do consider a possible spatial spillover indicated that the within-region transmission rate within Lombardy was slightly lower but that there was a high transmission between Lombardy and the other regions. For the other regions, the between-region transmission was generally shown to have been less strong.

All things considered, this thesis has been able to expose differences in the transmission rates across the Italian regions by applying models by Adda (2016) and by modelling undocumented infectives. Those involved with processing the results from models for the COVID-19 pandemic, such as policy makers, should not only trust models that focus on a nationwide level when thinking about what to do to tackle the pandemic. There are consistent regional effects that do not only require the modelling methods to differ across regions but the results from those models also differ vastly. Consequently, regional effects should be taken into account when interpreting and acting upon model results.

10 Future Research

In this thesis, we applied our models to the country of Italy and the 21 *regioni* that make it up. Of course, these methods were not specifically tailored to Italy and could easily be applied to other countries, if data is available that is granular enough. As highlighted in Section 8, this thesis does not take into account the specific manners in which regions interact with one another. For instance, the virus may have spread more quickly between regions that have closer economic ties or that are closer geographically. Some sort of spatiotemporal analysis to take these factors into account may be desirable. Giuliani et al. (2020) aim to model the spatiotemporal dimension for the early spread across Italian NUTS 3 regions. Combining their research with the approaches in this thesis could reach results that are more accurate in modelling and predicting the spread of viral diseases.

We also made the assumption that individuals gain immunity when they successfully cleared COVID-19 or at least long enough to last throughout our analysis. Models could be developed that do not assume immunity, given the recent news that reinfection within several months is indeed possible (Bloomberg News, 2020). Finally, this thesis did not take into account the strict lockdown that was instated on March 10. and that lasted for over two months, until around May 18. Future research could be conducted into methods that account for these strict movement limiting regulations.

One major limitation of the models by Adda (2016) is that they do not allow for the estimation of the recovery rate in addition to the transmission rate. Therefore, we tried to develop our own model as well, derived by discretizing the SIR model. A two-step approach was developed to estimate both the transmission rate β and the recovery rate γ, hopefully with the conclusion that we could also make conclusions about the effective reproduction number R_{eff}. Unfortunately, the resulting estimates are too low to be interpreted. This problem is likely the result of nonstationarity in the data. Indeed, Castle et al. (2020) state that nonstationarity is often a problem when modelling pandemics. The underlying data is often nonstationary because of a slow start, after which there is an exponential increase in the number of cases. The problem is especially pernicious due to the stacking of the nonstationarity of the data with the nonstationarity of the reporting process. In Appendix D, we present a short explanation on the methodology that was tried and some results. If one is able to take the nonstationarity into account, this may be a promising field of research.

In this thesis, we also did not discuss the full weighted model presented by Adda (2016) due to a lack of data. As time progresses and more qualitative and suitable data becomes available, it may be worth estimating this model. One data source that could be useful are the Community Mobility Reports by Google, which use

anonymized data of users to investigated changes in movements in the population compared to a median baseline (Google LLC, 2020). For instance, the reports include data on the change in the number of people that have visited public transport hubs, parks, and supermarkets. If this source can be included, the results can show the effects that policy has had on the spread of the virus.

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.

Agence France-Presse. (2020). *North Korea issues shoot-to-kill orders to prevent virus: US*. Retrieved September 17, 2020, from https://www.bangkokpost.com/world/1983779/north-korea-issues-shoot-to-kill-orders-to-prevent-virus-us

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/

Bloomberg News. (2020). *Two Chinese patients test positive months after virus recovery*. Retrieved September 7, 2020, from https://www.bloomberg.com/news/articles/2020-08-13/two-chinese-patients-test-positive-months-after-virus-recovery

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.

Briz-Redón, Á., & Serrano-Aroca, Á. (2020). A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Science of the Total Environment*, 138811.

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

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2020). *Short-term forecasting of the coronavirus pandemic*. Retrieved September 16, 2020, from https://forecasters.org/blog/2020/04/30/short-term-forecasting-of-the-coronavirus-pandemic/

European Centre for Disease Prevention and Control. (2020a). *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

European Centre for Disease Prevention and Control. (2020b). *Transmission of COVID-19*. Retrieved September 17, 2020, from https://www.ecdc.europa.eu/en/covid-19/latest-evidence/transmission

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/

Giuliani, D., Dickson, M. M., Espa, G., & Santi, F. (2020). Modelling and predicting the spread of coronavirus (covid-19) infection in nuts-3 italian regions. *arXiv preprint arXiv:2003.06664*.

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

Human Rights Watch. (2020). *Turkmenistan denies apparent covid-19 outbreak*. Retrieved August 19, 2020, from https://www.hrw.org/news/2020/06/27/turkmenistan-denies-apparent-covid-19-outbreak

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.

LePan, N. (2020). *Visualizing the history of pandemics*. Retrieved September 17, 2020, from https://www.visualcapitalist.com/history-of-pandemics-deadliest/

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

Nebehay, S. (2020). *North Korea testing, quarantining for COVID-19, still says no cases: WHO representative*. Retrieved August 19, 2020, from https://www.reuters.com/article/us-health-coronavirus-northkorea/north-korea-testing-quarantining-for-covid-19-still-says-no-cases-who-representative-idUSKBN21P3C2

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

Porta, M. (2014). *A dictionary of epidemiology*. Oxford University Press.

RFE/RL. (2020). *COVID-19: Hospitals overwhelmed in 'coronavirus-free' Turkmenistan*. Retrieved September 17, 2020, from https://www.rferl.org/a/hospitals-overwhelmed-in-coronavirus-free-turkmenistan/30820433.html

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

Scarr, S., & Sharma, M. (2020). *A deluge of death in northern Italy*. Retrieved September 17, 2020, from https://graphics.reuters.com/HEALTH-CORONAVIRUS-LOMBARDY/0100B5LT46P/index.html

Schultz, T. (2020). *Why Belgium's death rate is so high: It counts lots of suspected COVID-19 cases*. Retrieved September 15, 2020, from https://www.npr.org/sections/coronavirus-live-updates/2020/04/22/841005901/why-belgiums-death-rate-is-so-high-it-counts-lots-of-suspected-covid-19-cases

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

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.

WHO. (2019). *What are the International Health Regulations and Emergency Committees?* Retrieved August 19, 2020, from https://www.who.int/news-room/q-a-detail/what-are-the-international-health-regulations-and-emergency-committees

WHO. (2020a). *Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)*. Retrieved August 19, 2020, from https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)

WHO. (2020b). *WHO director-general's opening remarks at the media briefing on COVID-19 - 11 march 2020*. Retrieved August 19, 2020, from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

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 or P.A. Bolzano	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, Provincia Autonoma di Trento or P.A. 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
NUTS	Nomenclature des Unités Territoriales Statistiques	Section 3.1
SIR model	Standard Inflammatory Response model	Section 4
AIC	Akaike Information Criterion	Section 5
BIC	Bayesian Information Criterion	Section 5

Table A.2 continues on next page

Table A.2 continued from previous page

Abbreviation	Full name	First mentioned in...
OLS	Ordinary Least Squares	Section 7.1
POLS	Pooled Ordinary Least Squares	Section 8.1

B Tables

B.1 Results for Section 7: Within-Region Spread Model

In Section 7.2, we presented the results from the within and between-region spread model. This appendix contains additional tables with results for this model. Recall that the within and between-region spread model was defined in equation (7.5) as:

$$\Delta i_{p,t} = \beta_{within} S_{p,t-\tau} \Delta i_{p,t-\tau} + \delta X_{p,t} + \eta_{p,t}.$$

Table B.1. Estimates from the within-region spread model per region without model selection. Estimates are given with t-statistics in parentheses. Data spans May 9 until 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}	Weekend	β_{within}	Weekend
National (OLS)	0.493***	92.860***	0.423***	379.483***
	(21.992)	(3.071)	(27.417)	(2.976)
National (POLS)	0.533***	15.086***	0.458***	90.994***
	(27.980)	(5.074)	(22.981)	(4.893)
Abruzzo	0.372***	2.751**	0.322***	13.454**
	(4.952)	(2.091)	(6.253)	(2.249)
Basilicata	0.063	0.612	0.063	3.279
	(0.565)	(0.772)	(0.615)	(0.890)
P.A. Bolzano	0.269***	2.346***	0.219***	7.117***
	(3.467)	(3.683)	(4.175)	(3.961)
Calabria	0.242**	2.189**	0.201*	12.897***
	(2.043)	(2.614)	(1.917)	(2.828)
Emilia-Romagna	0.358***	22.416***	0.287***	91.467***
	(11.673)	(4.844)	(14.768)	(5.266)
Friuli Venezia Giulia	0.342***	2.582***	0.252***	8.578***
	(7.559)	(3.476)	(8.891)	(3.508)
Lazio	0.494***	10.172***	0.424***	57.018***
	(10.069)	(3.441)	(11.273)	(3.610)
Liguria	0.382***	4.179	0.322***	17.706
	(11.931)	(1.430)	(13.539)	(1.242)
Lombardy	0.566***	26.949	0.502***	114.802
	(16.318)	(1.271)	(16.989)	(1.072)

Table B.1 continues on next page

Table B.1 continued from previous page

	Regular model		Modelling undocumented infectives	
Region	β_{within}	Weekend	β_{within}	Weekend
Marche	0.436***	2.697**	0.387***	12.796**
	(9.376)	(2.144)	(11.347)	(2.286)
Molise	0.140	1.855**	0.155*	11.448**
	(1.595)	(2.456)	(1.810)	(2.299)
Piedmont	0.352***	7.449*	0.300***	28.533
	(23.006)	(1.896)	(25.913)	(1.492)
Apulia	0.376***	3.804**	0.335***	23.618***
	(6.165)	(2.581)	(7.302)	(2.719)
Sardinia	0.393***	0.795	0.276***	5.301
	(3.761)	(1.278)	(3.874)	(1.655)
Tuscany	0.281***	8.444***	0.237***	35.079***
	(6.666)	(3.854)	(8.744)	(3.929)
Umbria	0.436***	1.428***	0.291***	5.402***
	(3.936)	(3.168)	(4.134)	(3.351)
Aosta Valley	0.267***	0.354	0.287***	1.054
	(3.862)	(1.287)	(5.873)	(1.157)
Veneto	0.489***	11.951*	0.284***	39.380***
	(6.748)	(1.920)	(7.344)	(2.672)

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

Table B.2. Estimates from the 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 until August 16, 2020 (100 days). Undocumented infectives are not modelled.

	Model selection with AIC		Model selection with BIC	
Region	β_{within}	Weekend	β_{within}	Weekend
National (OLS)	0.493***	92.860***	0.493***	92.860***
	(21.992)	(3.071)	(21.992)	(3.071)
National (POLS)	0.533***	15.086***	0.533***	15.086***
	(27.980)	(5.074)	(27.980)	(5.074)
Abruzzo	0.372***	2.751**	0.440***	
	(4.952)	(2.091)	(6.387)	
Basilicata	0.073		0.073	
	(0.665)		(0.665)	
P.A. Bolzano	0.269***	2.346***	0.269***	2.346***
	(3.468)	(3.683)	(3.468)	(3.683)
Calabria	0.242**	2.189**	0.242**	2.189**
	(2.043)	(2.614)	(2.043)	(2.614)
Emilia-Romagna	0.358***	22.416***	0.358***	22.416***
	(11.673)	(4.844)	(11.673)	(4.844)
Friuli Venezia Giulia	0.342***	2.582***	0.342***	2.582***
	(7.559)	(3.476)	(7.559)	(3.476)

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}	Weekend	β_{within}	Weekend
Lazio	0.494***	10.172***	0.494***	10.172***
	(10.069)	(3.441)	(10.069)	(3.441)
Liguria	0.382***	4.179	0.399***	
	(11.931)	(1.430)	(13.271)	
Lombardy	0.583***		0.583***	
	(18.205)		(18.205)	
Marche	0.436***	2.697**	0.473***	
	(9.376)	(2.144)	(10.775)	
Molise	0.140	1.855**	0.140	1.855**
	(1.595)	(2.456)	(1.595)	(2.456)
Piedmont	0.352***	7.449*	0.363***	
	(23.006)	(1.896)	(25.048)	
Apulia	0.376***	3.804**	0.376***	3.804**
	(6.165)	(2.581)	(6.165)	(2.581)
Sardinia	0.437***		0.437***	
	(4.417)		(4.417)	
Tuscany	0.281***	8.444***	0.281***	8.444***
	(6.666)	(3.854)	(6.666)	(3.854)
Umbria	0.436***	1.428***	0.436***	1.428***
	(3.936)	(3.168)	(3.936)	(3.168)
Aosta Valley	0.290***		0.290***	
	(4.341)		(4.341)	
Veneto	0.489***	11.951*	0.542***	
	(6.748)	(1.920)	(7.958)	

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

B.2 Results for Section 8: Within and Between-Region Spread Model

In Section 8.2, we presented the results from the within and between-region spread model. This appendix contains additional tables with results for this model. Recall that the within and between-region spread model was defined in equation (8.1) as:

$$\Delta i_{p,t} = \beta_{within} S_{p,t-\tau} \Delta i_{p,t-\tau} + \beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \Delta i_{c,t-\tau} + \delta X_{p,t} + \eta_{p,t}.$$

Table B.3. Estimates from the within and between-region spread model per region without model selection. Estimates are given with t-statistics in parentheses. Data spans May 9 until 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}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Abruzzo	-0.011	5.871×10^{-3}***	2.394*	-0.047	5.943×10^{-3}***	12.494**
	(-0.079)	(3.399)	(1.909)	(-0.495)	(4.455)	(2.279)
Basilicata	0.039	6.89911×10^{-4}	0.2450	0.032	5.547×10^{-4}	1.711
	(0.346)	(1.059)	(0.283)	(0.298)	(1.132)	(0.435)
P.A. Bolzano	0.246**	2.043×10^{-4}	2.305***	0.201**	1.018×10^{-4}	7.030***
	(2.355)	(0.317)	(3.534)	(2.559)	(0.310)	(3.848)
Calabria	0.186	7.936×10^{-4}	1.893**	0.128	8.270×10^{-4}	11.514**
	(1.460)	(1.171)	(2.168)	(1.088)	(1.341)	(2.472)
Emilia-Romagna	0.365***	-8.645×10^{-4}	22.410***	0.171***	0.015*	90.384***
	(4.184)	(-0.080)	(4.817)	(2.635)	(1.873)	(5.267)
Friuli Venezia Giulia	0.114*	3.701×10^{-3}***	1.759**	0.034	3.026×10^{-3}***	4.330**
	(1.707)	(4.346)	(2.481)	(0.929)	(7.601)	(2.139)
Lazio	0.342***	7.364×10^{-3}	10.573***	0.238**	0.010*	61.107***
	(2.671)	(1.282)	(3.568)	(2.221)	(1.853)	(3.879)
Liguria	-0.025	0.033***	0.345	-0.022	0.029***	1.408
	(-0.357)	(6.260)	(0.135)	(-0.391)	(6.413)	(0.115)
Lombardy	0.326***	0.220***	14.034	0.283***	0.186***	65.410
	(4.878)	(4.081)	(0.704)	(4.421)	(3.799)	(0.646)
Marche	0.245**	4.505×10^{-3}**	2.150*	0.181**	4.644×10^{-3}***	10.483*
	(2.521)	(2.229)	(1.710)	(2.312)	(2.907)	(1.921)
Molise	-0.087	4.020×10^{-3}***	0.163	-0.080	4.559×10^{-3}***	0.981
	(-1.186)	(8.188)	(0.264)	(-1.177)	(9.250)	(0.257)
Piedmont	0.104***	0.064***	0.260	0.091***	0.056***	0.933
	(3.635)	(9.363)	(0.088)	(3.856)	(9.514)	(0.066)
Apulia	0.120	5.525×10^{-3}***	3.083**	0.035	8.078×10^{-3}***	18.908**
	(1.051)	(2.629)	(2.115)	(0.393)	(3.896)	(2.303)
Sardinia	0.397***	-3.065×10^{-5}	0.803	0.252**	1.831×10^{-4}	5.062
	(2.835)	(-0.048)	(1.237)	(2.468)	(0.321)	(1.532)
Tuscany	0.051	0.010***	8.023***	0.030	9.454×10^{-3}***	33.901***
	(0.605)	(3.107)	(3.812)	(0.535)	(4.141)	(4.095)
Umbria	0.427***	4.689×10^{-5}	1.410***	0.235**	2.449×10^{-4}	4.877***
	(2.981)	(0.102)	(2.904)	(2.467)	(0.876)	(2.832)
Aosta Valley	0.023	1.113×10^{-3}***	-6.610×10^{-3}	0.055	7.107×10^{-4}***	-6.958×10^{-3}
	(0.256)	(3.937)	(-0.024)	(0.769)	(4.169)	(-7.896×10^{-3})
Veneto	0.904***	-0.033***	16.184***	0.434***	-7.954×10^{-3}	43.868***
	(6.990)	(-3.771)	(2.721)	(4.370)	(-1.637)	(2.951)

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

Table B.4. Estimates from the 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 until August 16, 2020 (100 days). Undocumented infectives are not modelled.

	Model selection with AIC			Model selection with BIC		
Region	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Abruzzo	-0.011	5.871×10^{-3}***	2.394*	0.030	6.147×10^{-3}***	
	(-0.079)	(3.399)	(1.909)	(0.226)	(3.524)	
Basilicata	0.040	7.638×10^{-4}		0.040	7.638×10^{-4}	
	(0.357)	(1.286)		(0.357)	(1.286)	
P.A. Bolzano	0.246**	2.043×10^{-4}	2.305***	0.201**	1.018×10^{-4}	7.030***
	(2.355)	(0.317)	(3.534)	(2.559)	(0.310)	(3.848)
Calabria	0.186	7.936×10^{-4}	1.893**	0.249*	1.219×10^{-3}*	
	(1.460)	(1.171)	(2.168)	(1.967)	(1.845)	
Emilia-Romagna	0.365***	-8.645×10^{-4}	22.410***	0.365***	-8.645×10^{-4}	22.410***
	(4.184)	(-0.080)	(4.817)	(4.184)	(-0.080)	(4.817)
Friuli Venezia Giulia	0.114*	3.701×10^{-3}***	1.759**	0.114*	3.701×10^{-3}***	1.759**
	(1.707)	(4.346)	(2.481)	(1.707)	(4.346)	(2.481)
Lazio	0.342***	7.364×10^{-3}	10.573***	0.342***	7.364×10^{-3}	10.573***
	(2.671)	(1.283)	(3.568)	(2.671)	(1.283)	(3.568)
Liguria	-0.026	0.033***		-0.026	0.033***	
	(-0.372)	(6.515)		(-0.372)	(6.515)	
Lombardy	0.329***	0.226***		0.329***	0.226***	
	(4.929)	(4.257)		(4.929)	(4.257)	
Marche	0.245**	4.505×10^{-3}**	2.150*	0.245**	5.179×10^{-3}**	
	(2.521)	(2.229)	(1.710)	(2.496)	(2.587)	
Molise	-0.085	4.063×10^{-3}***		-0.085	4.063×10^{-3}***	
	(-1.168)	(8.821)		(-1.168)	(8.821)	
Piedmont	0.104***	0.065***		0.104***	0.065***	
	(3.665)	(9.767)		(3.665)	(9.767)	
Apulia	0.120	5.525×10^{-3}***	3.083**	0.127	6.362×10^{-3}***	
	(1.051)	(2.629)	(2.115)	(1.099)	(3.029)	
Sardinia	0.407***	1.846×10^{-3}		0.407***	1.846×10^{-3}	
	(2.903)	(0.299)		(2.903)	(0.299)	
Tuscany	0.051	0.010***	8.023***	0.051	0.010***	8.023***
	(0.605)	(3.107)	(3.812)	(0.605)	(3.107)	(3.812)
Umbria	0.427***	4.689×10^{-5}	1.410***	0.427***	4.689×10^{-5}	1.410***
	(2.981)	(0.102)	(2.904)	(2.981)	(0.102)	(2.904)
Aosta Valley	0.023	1.111×10^{-3}***		0.023	1.111×10^{-3}***	
	(0.259)	(4.193)		(0.259)	(4.193)	
Veneto	0.904***	-0.033***	16.184***	0.904***	-0.033***	16.184***
	(6.990)	(-3.771)	(2.721)	(6.990)	(-3.771)	(2.721)

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

C Figures

C.1 Figures for Section 2: Problem Description

This appendix contains additional plots as referenced in Section 2.

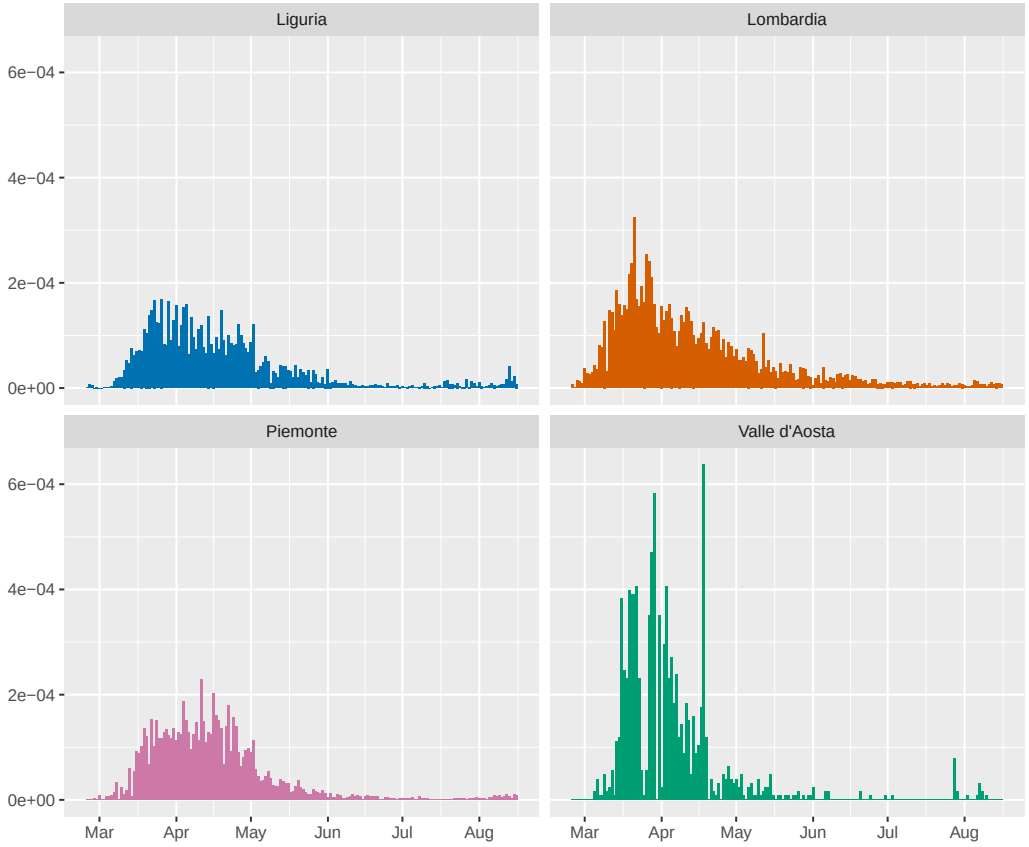


Figure C.1. Incidence rate per region for the Nord-Ovest NUTS 1 region.

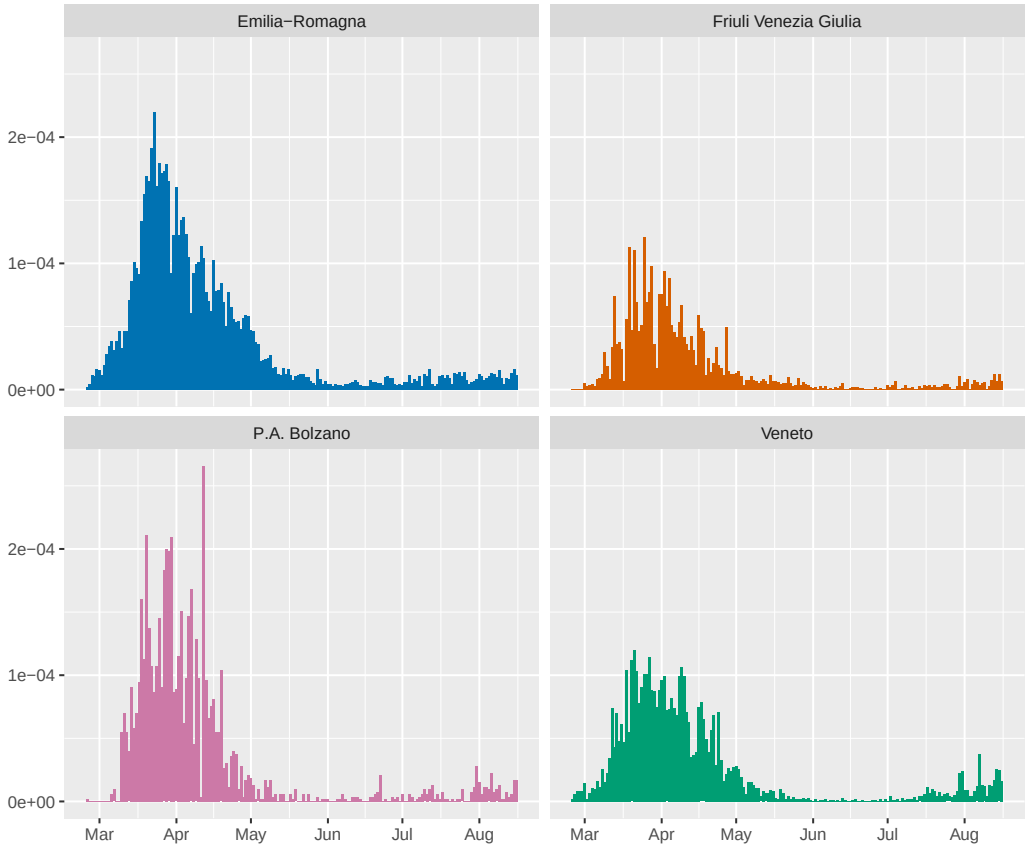


Figure C.2. Incidence rate per region for the Nord-Est NUTS 1 region.

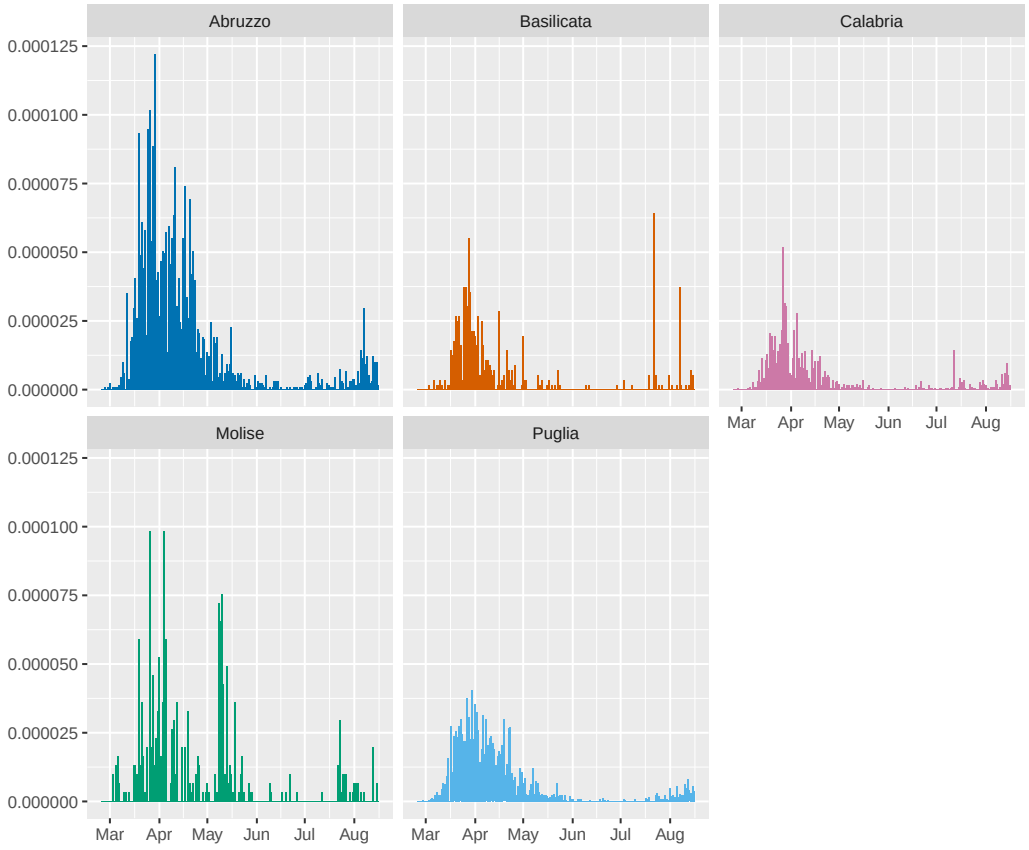


Figure C.3. Incidence rate per region for the Sud NUTS 1 region.

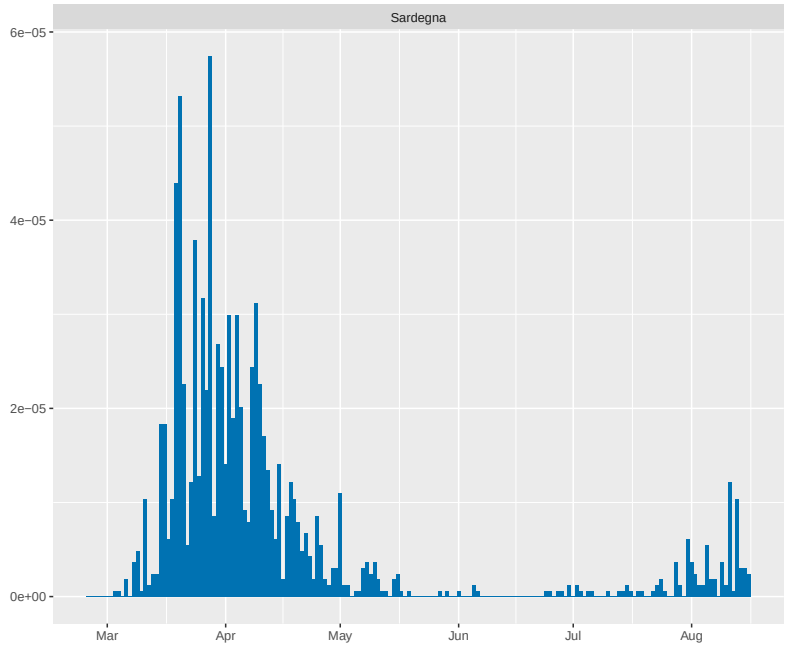
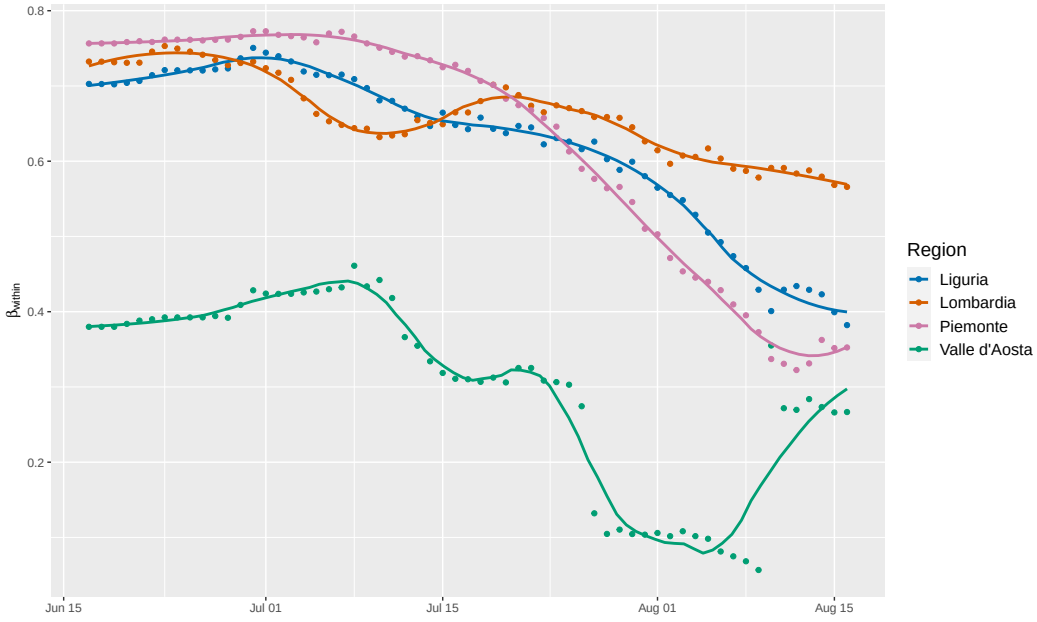


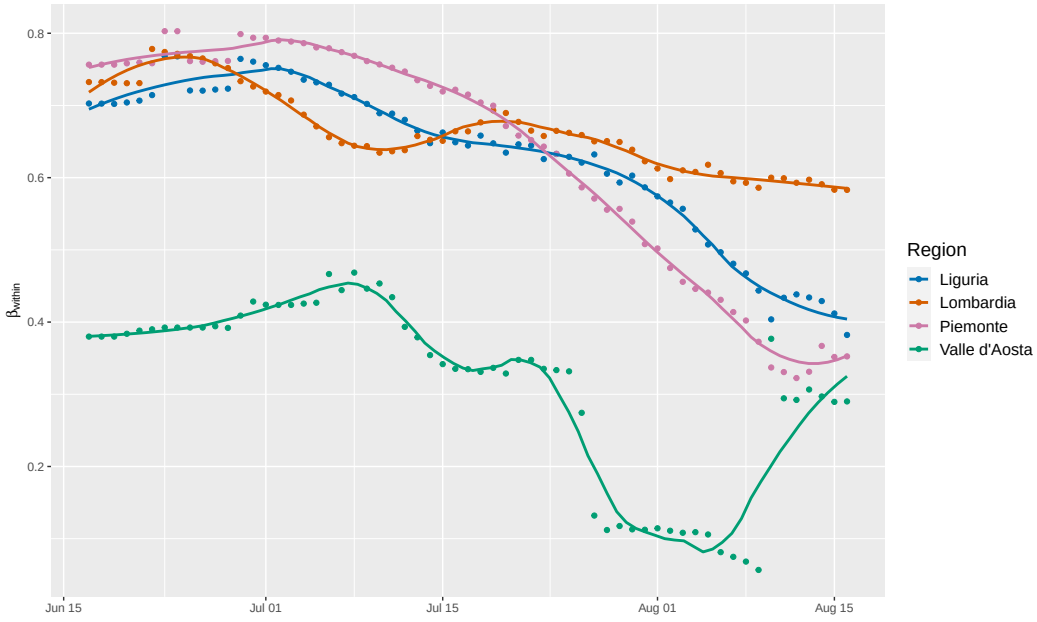
Figure C.4. Incidence rate per region for the Isole NUTS 1 region.

C.2 Figures for the Within-Region Spread Model

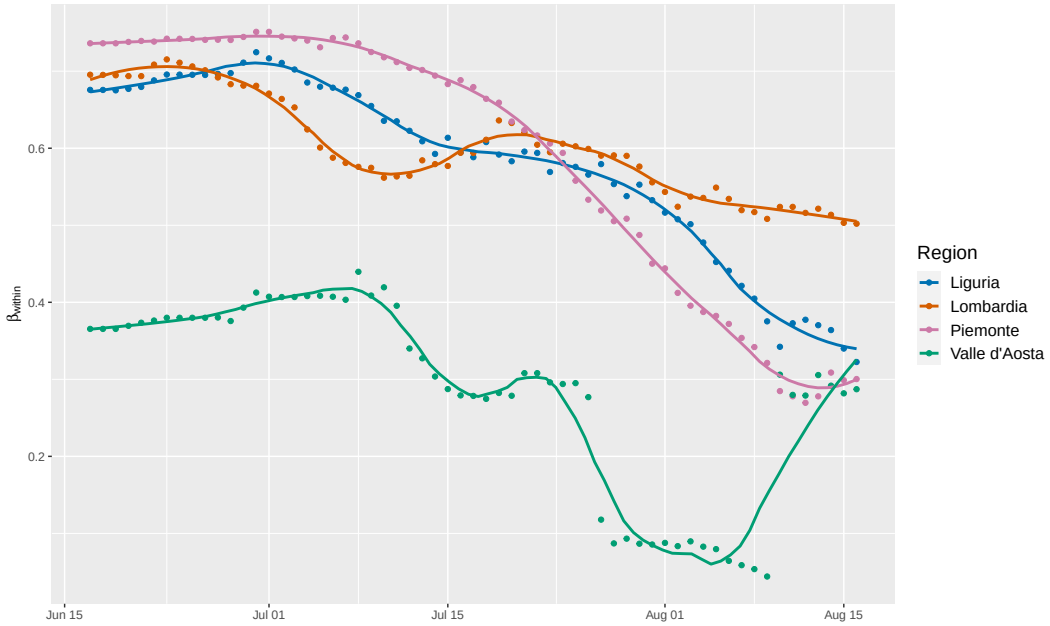
In Section 7.2 we presented the plots of β_{within} over time for the Nord-Est NUTS 1 region for the within-region spread model. In this appendix, we present the plots for the other NUTS 1 regions.



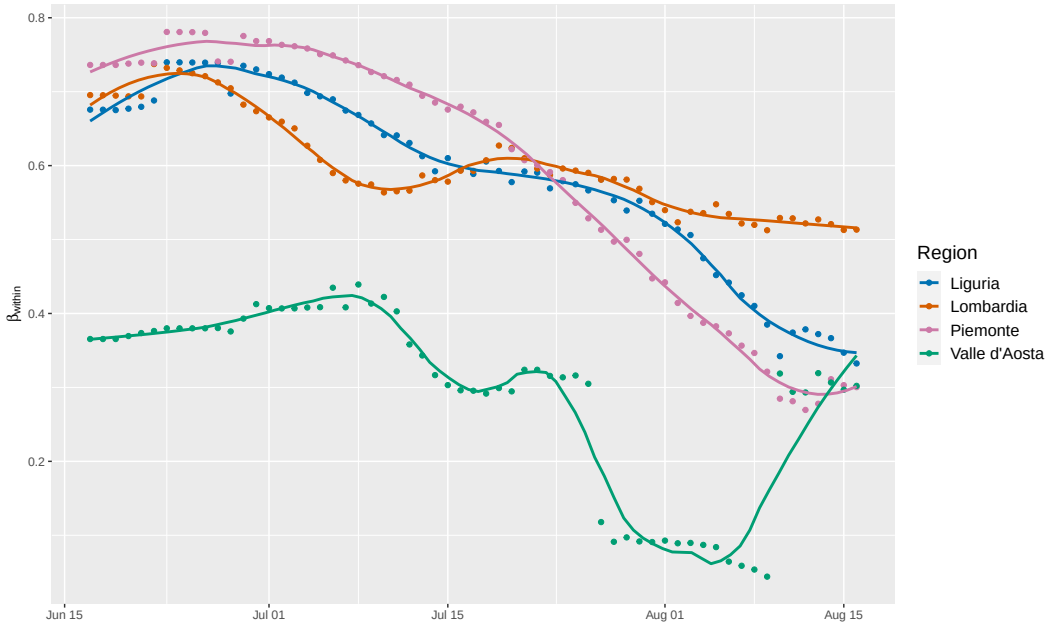
(a) Without model selection



(b) With model selection by AIC

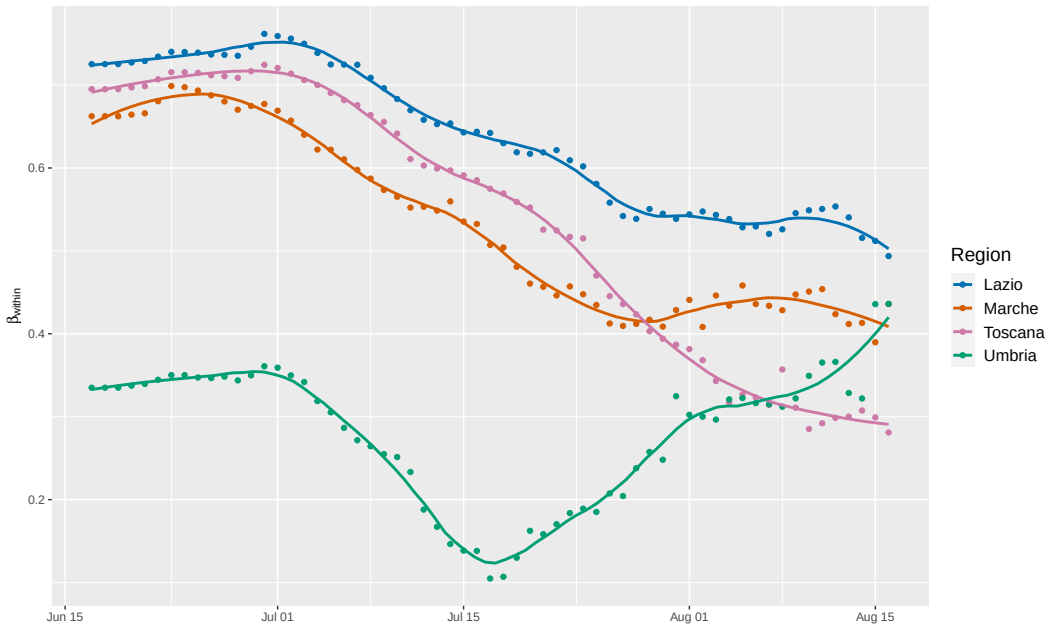


(c) Without model selection;
including undocumented infectives

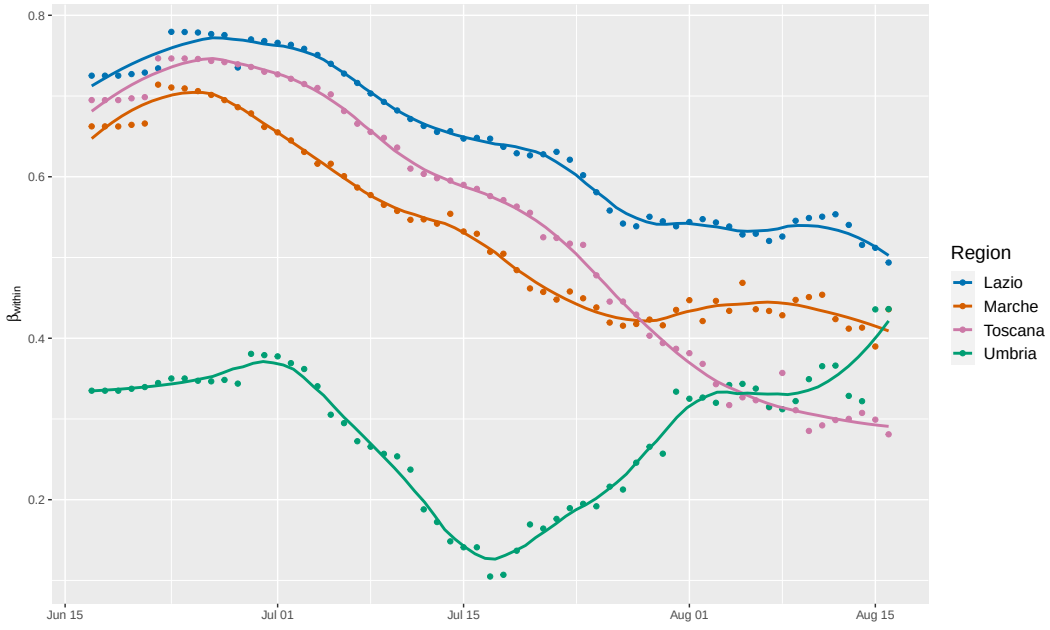


(d) With model selection by AIC;
including undocumented infectives

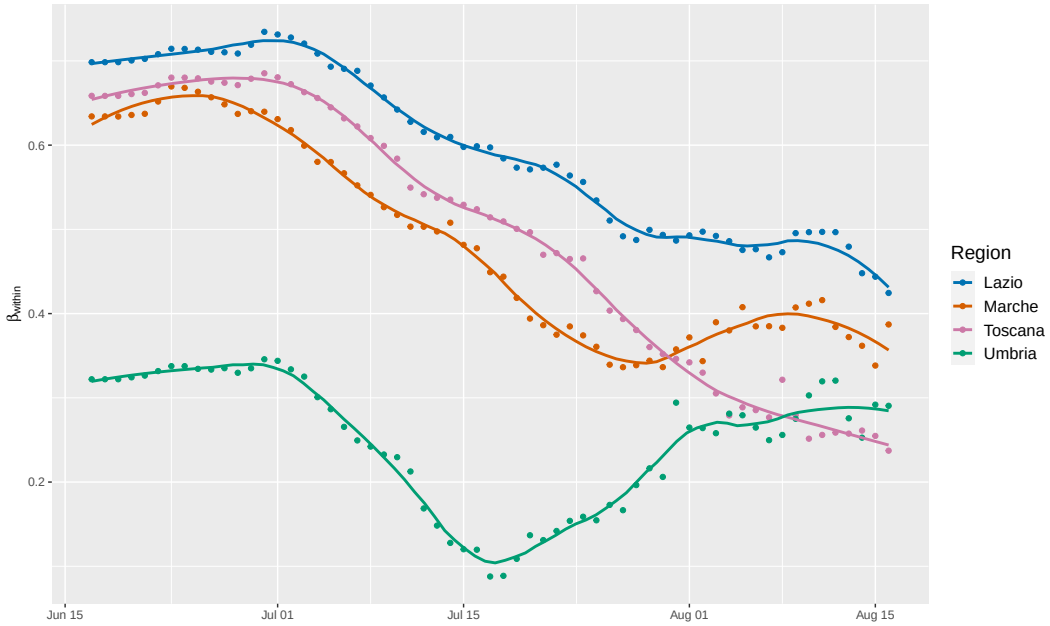
Figure C.5. Progression of β_{within} over time for the Nord-Ovest NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



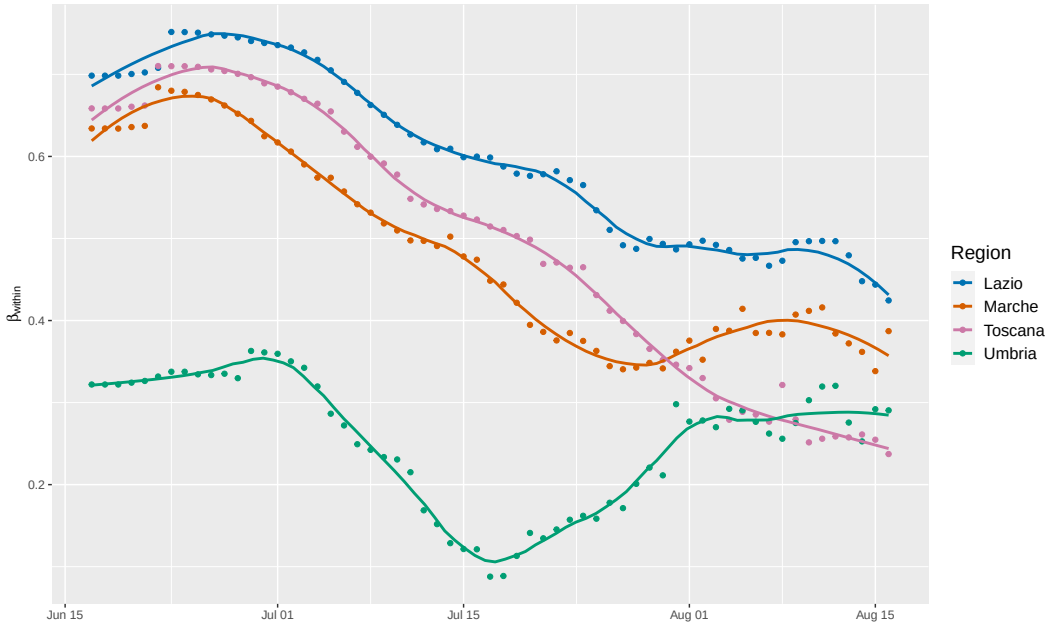
(a) Without model selection



(b) With model selection by AIC

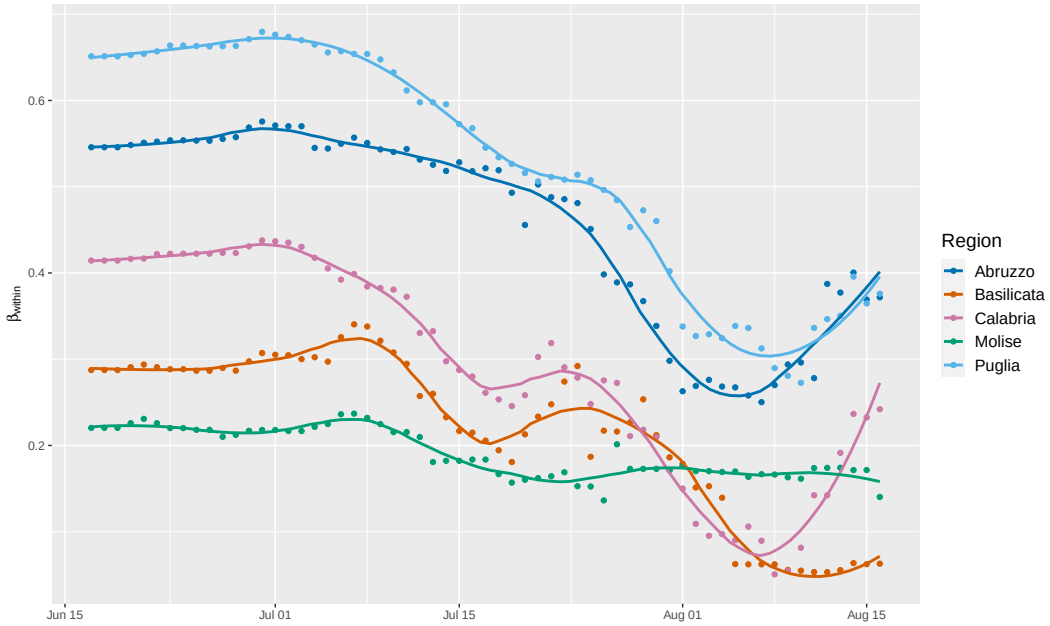


(c) Without model selection;
including undocumented infectives

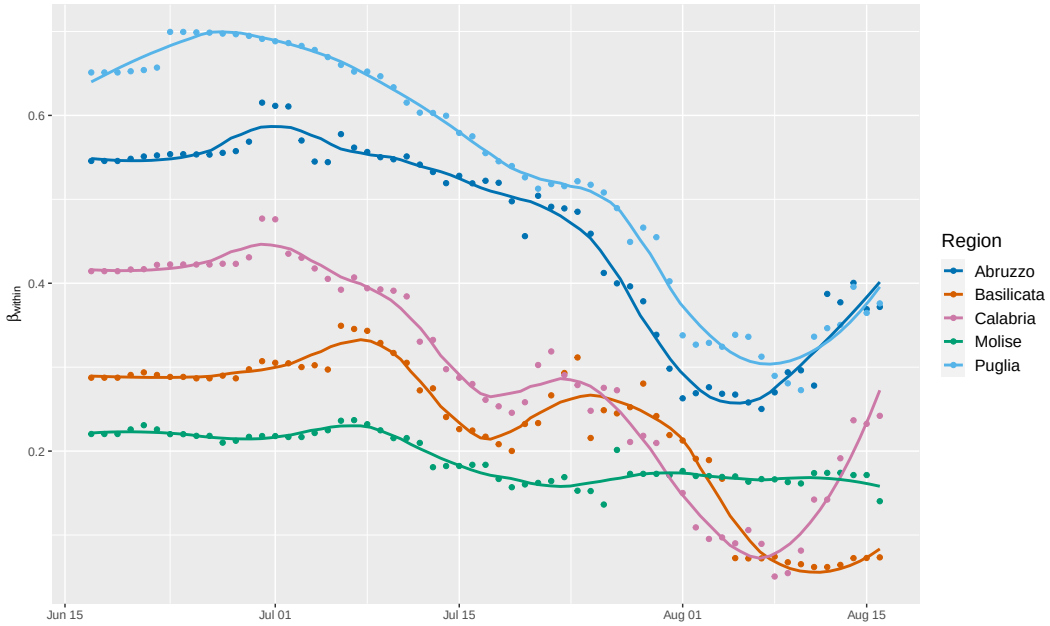


(d) With model selection by AIC;
including undocumented infectives

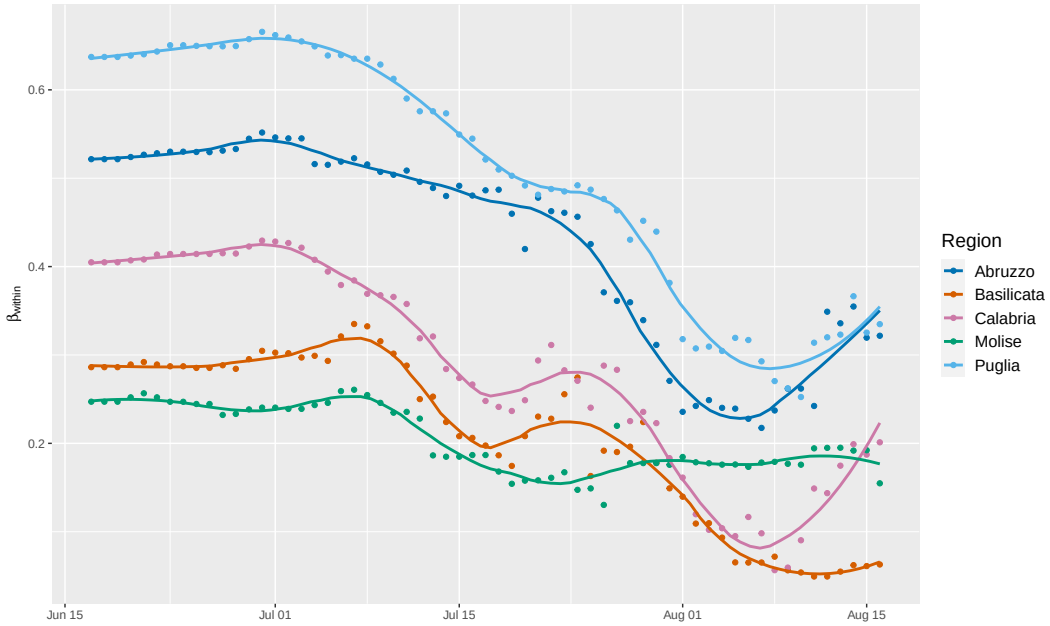
Figure C.6. Progression of β_{within} over time for the Centro (IT) NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



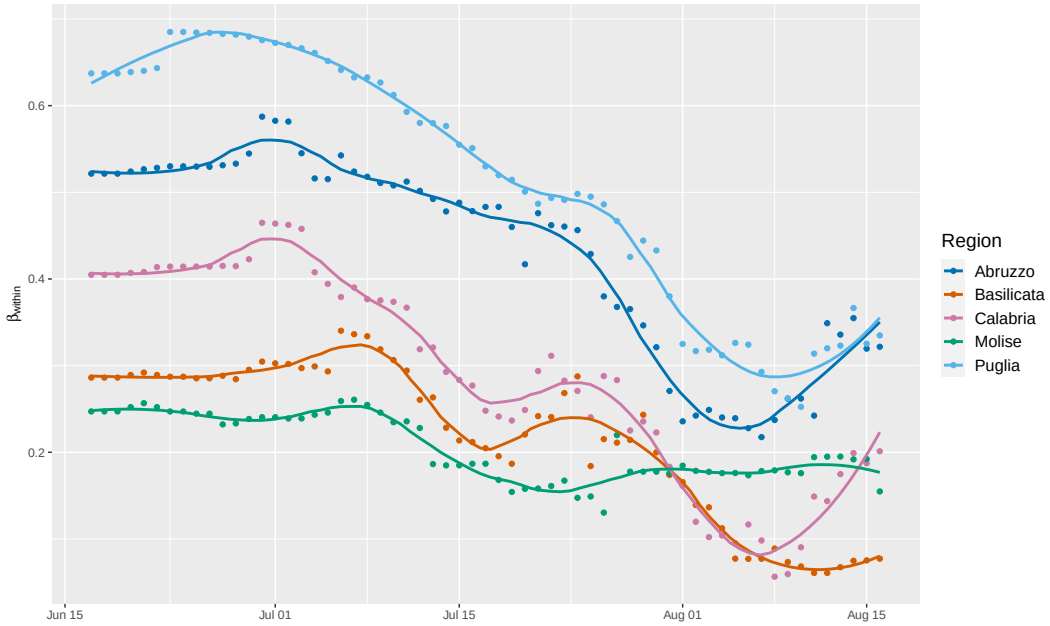
(a) Without model selection



(b) With model selection by AIC

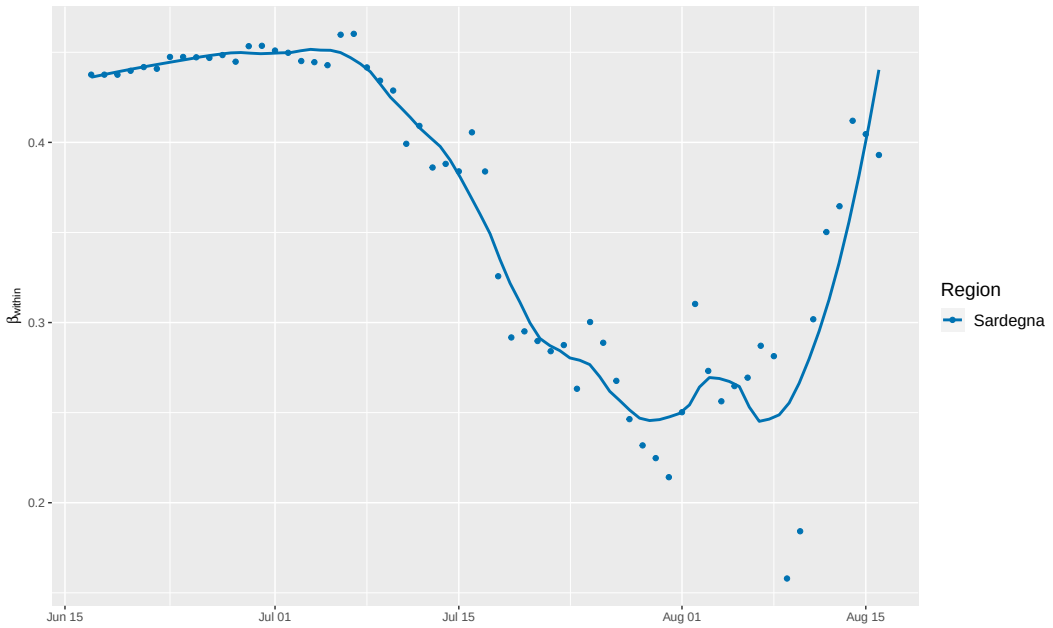


(c) Without model selection;
including undocumented infectives

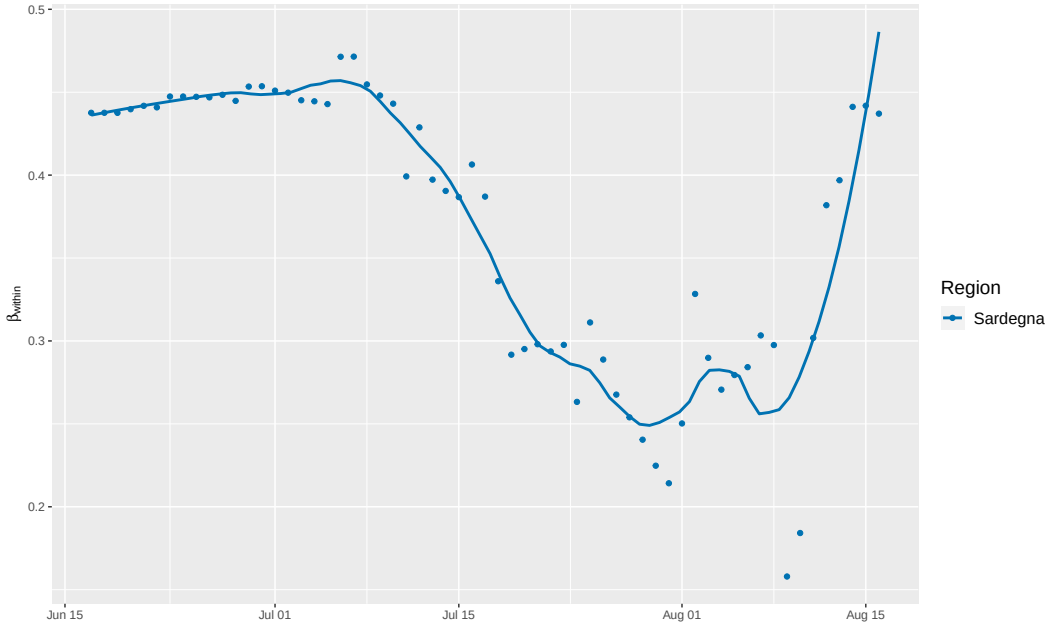


(d) With model selection by AIC;
including undocumented infectives

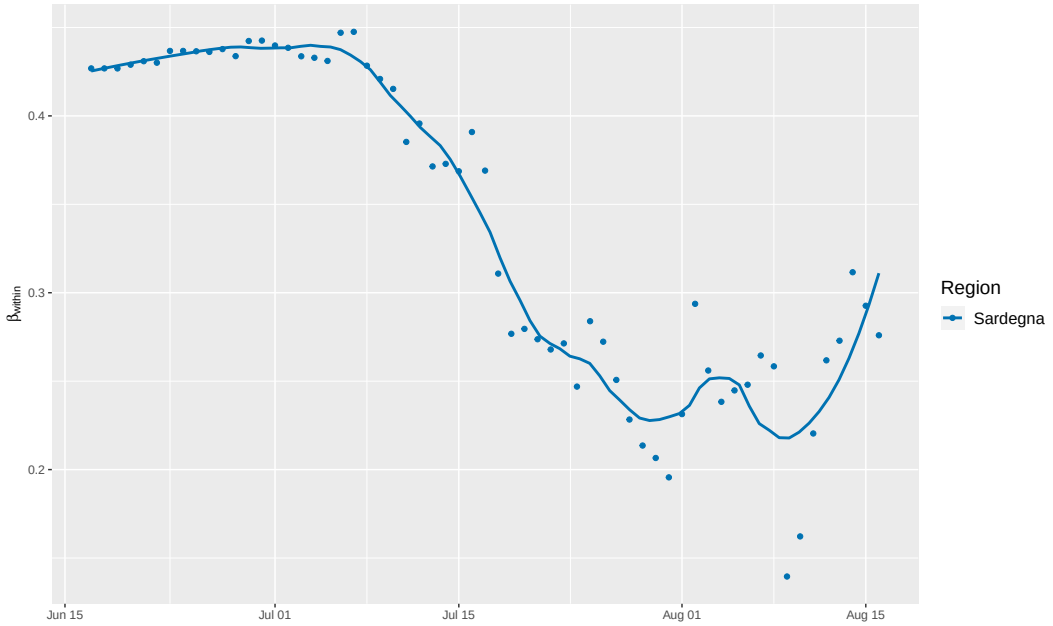
Figure C.7. Progression of β_{within} over time for the Sud NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



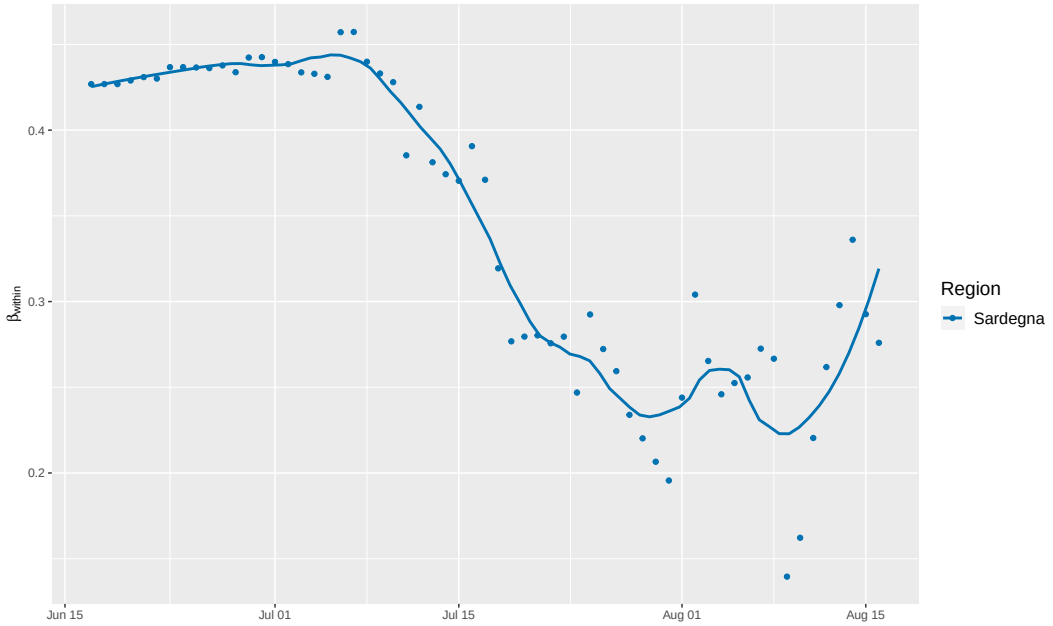
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives

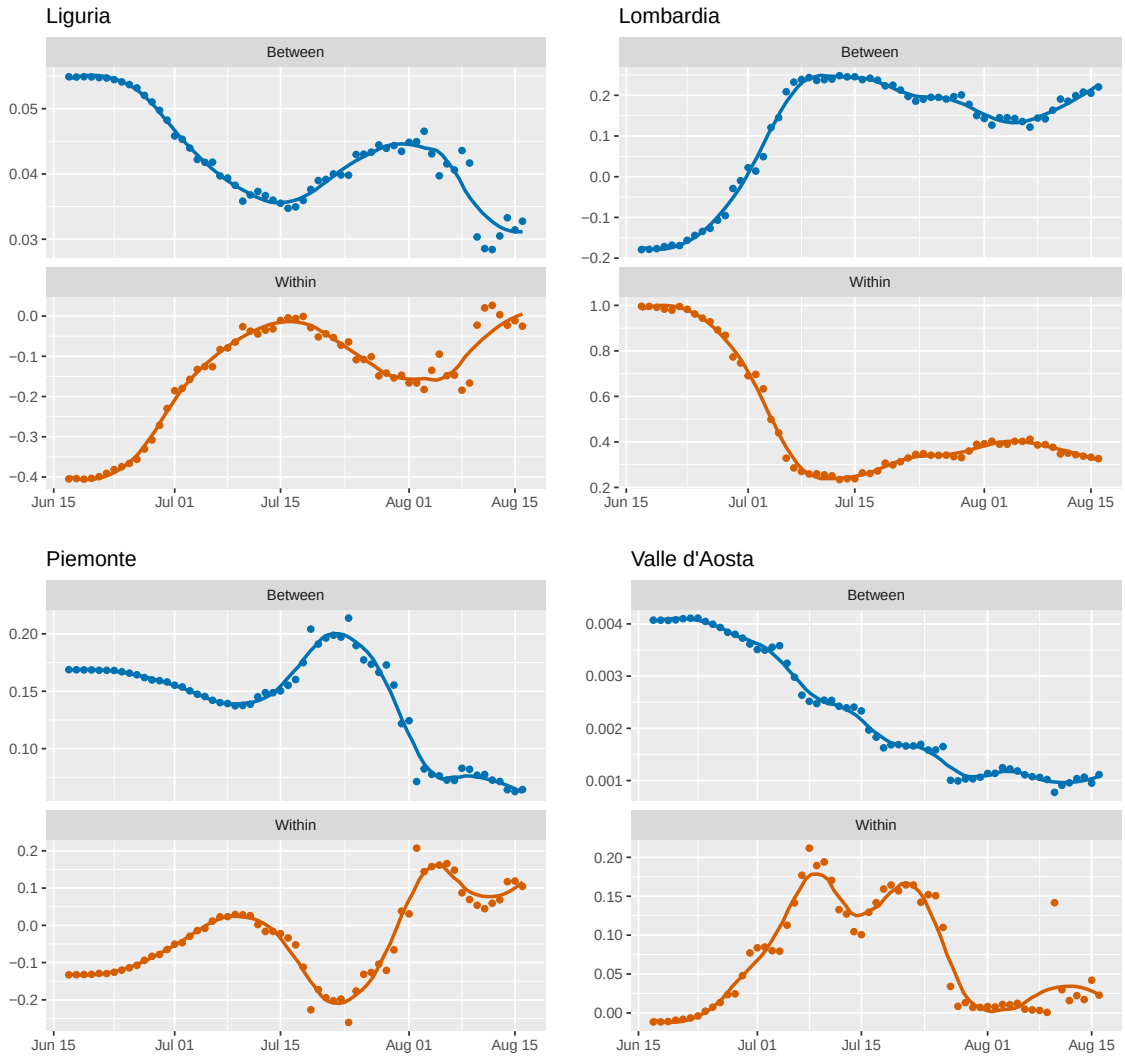


(d) With model selection by AIC;
including undocumented infectives

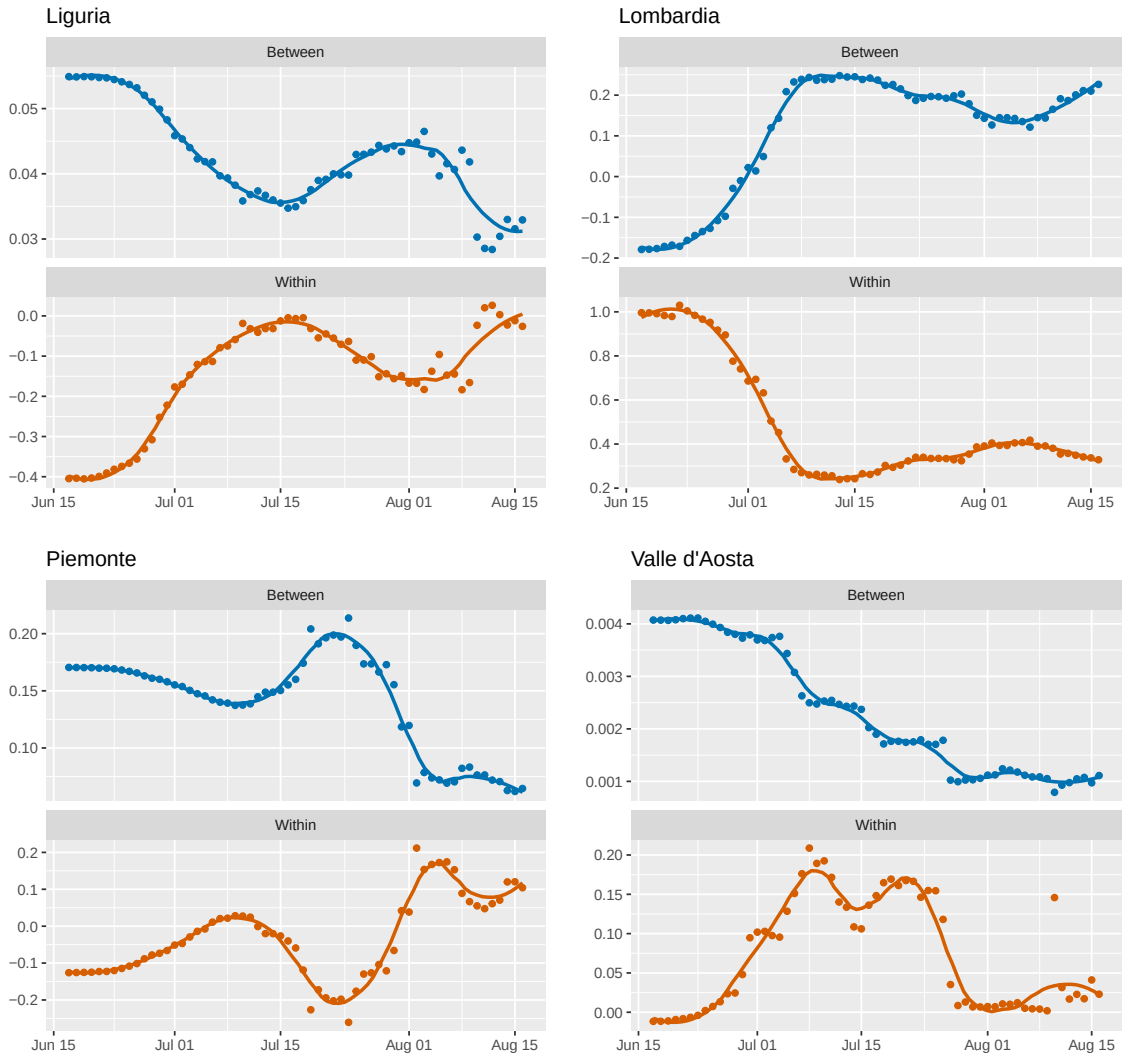
Figure C.8. Progression of β_{within} over time for the Isole NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

C.3 Figures for the Within and Between-Region Spread Model

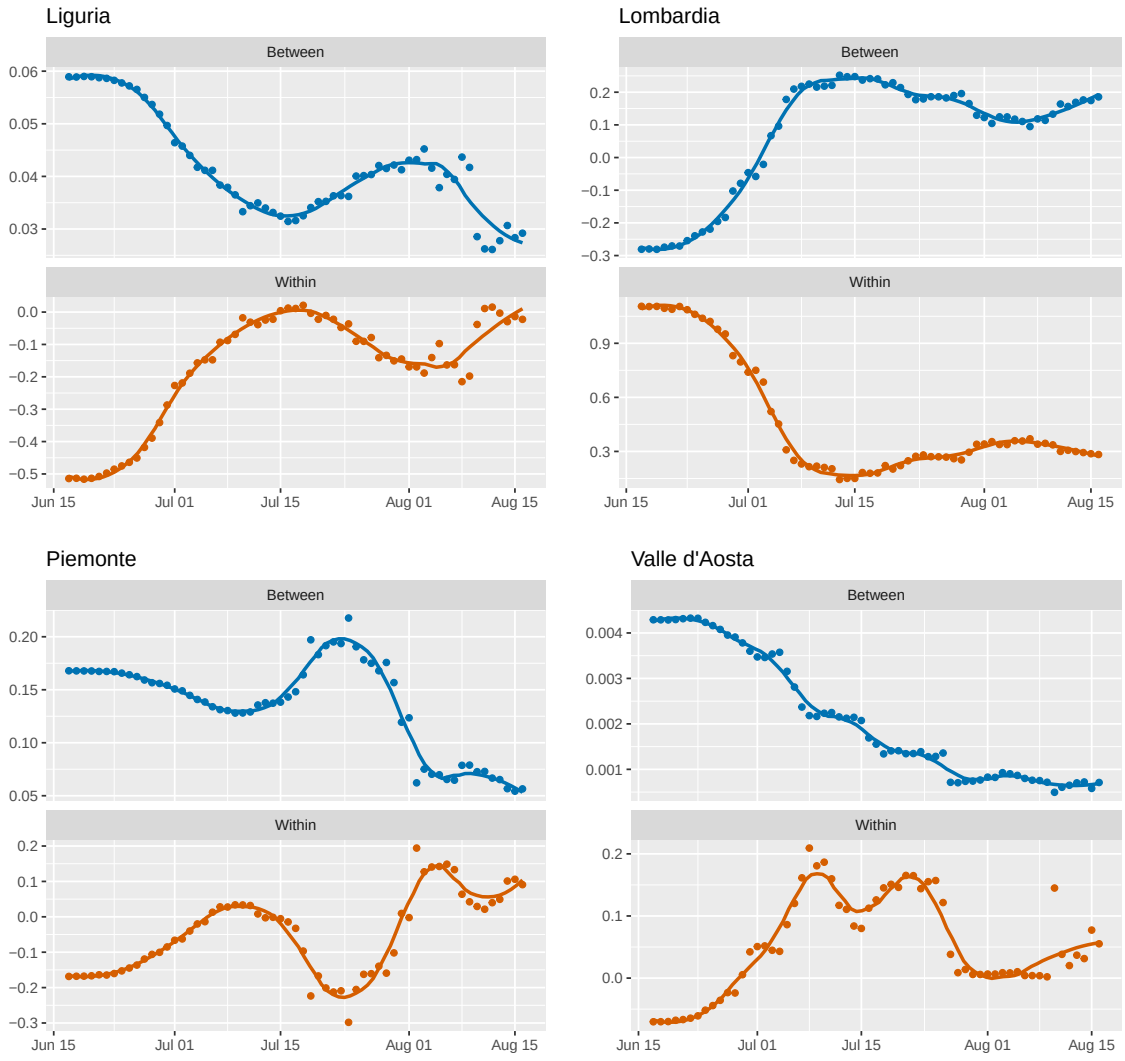
In Section 8.2 we presented the plots of β_{within} and $\beta_{between}$ over time for the Nord-Est NUTS 1 region for the within and between-region spread model. In this appendix, we present the plots for the other NUTS 1 regions.



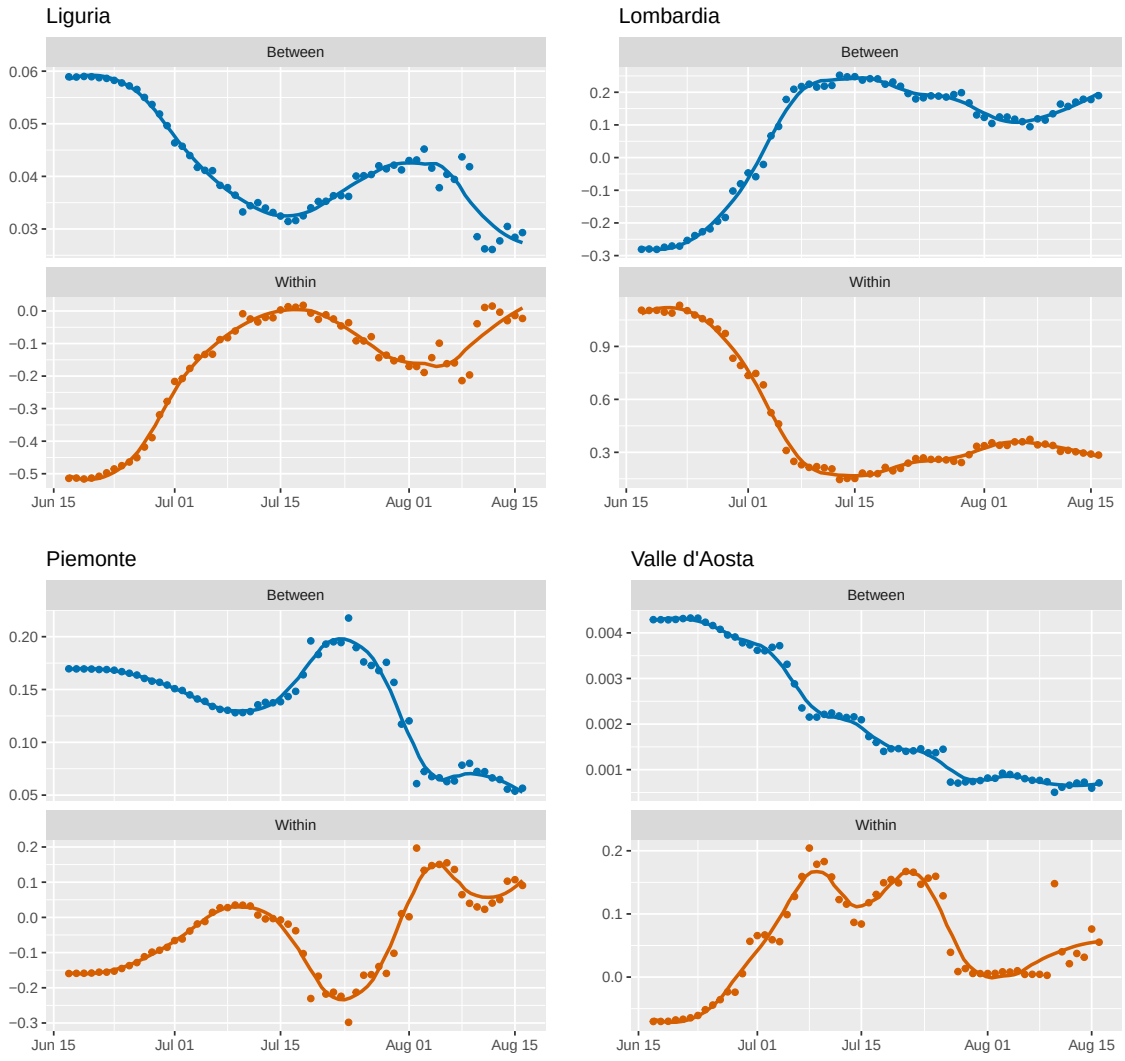
(a) Without model selection



(b) With model selection by AIC

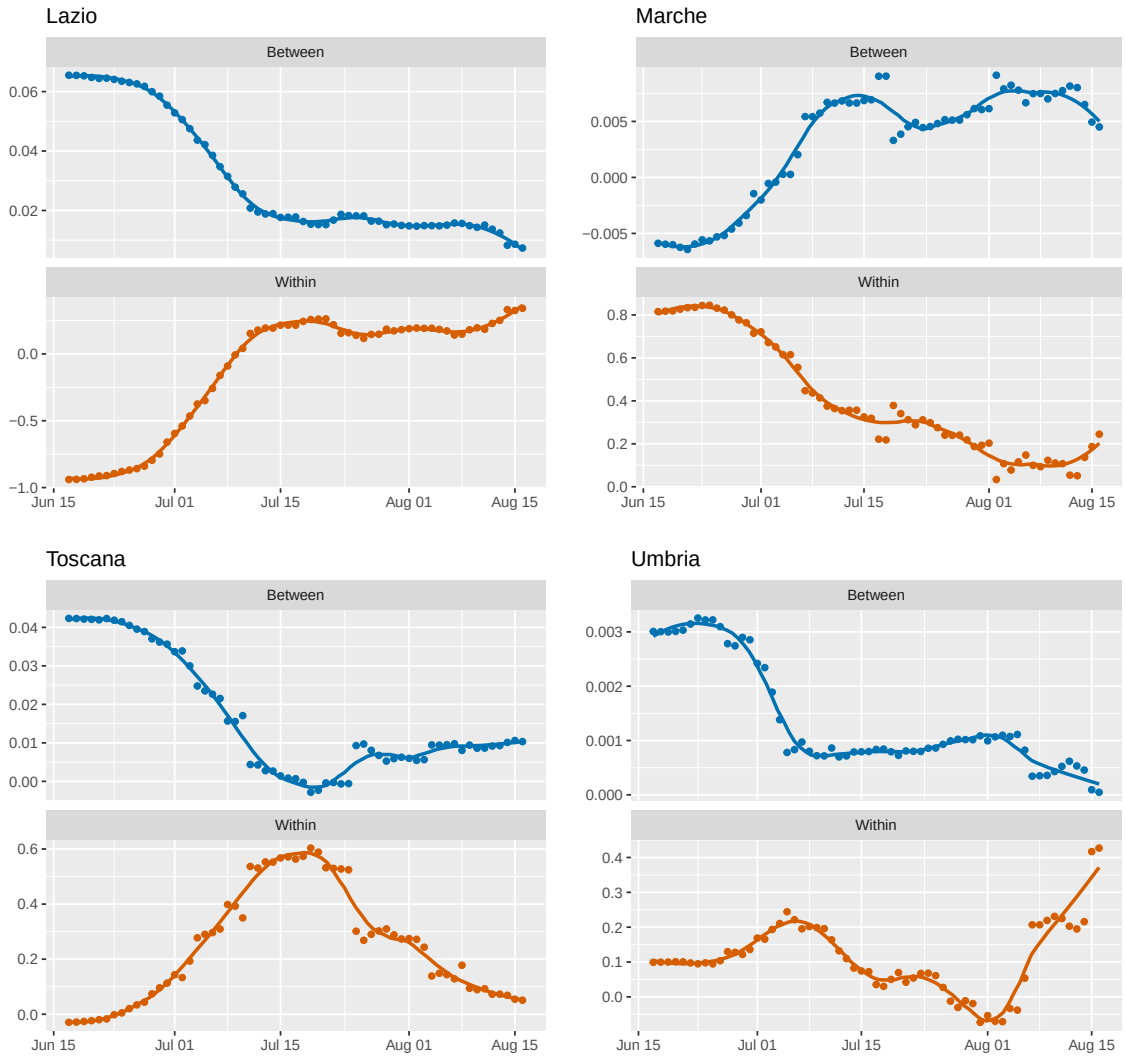


(c) Without model selection;
including undocumented infectives

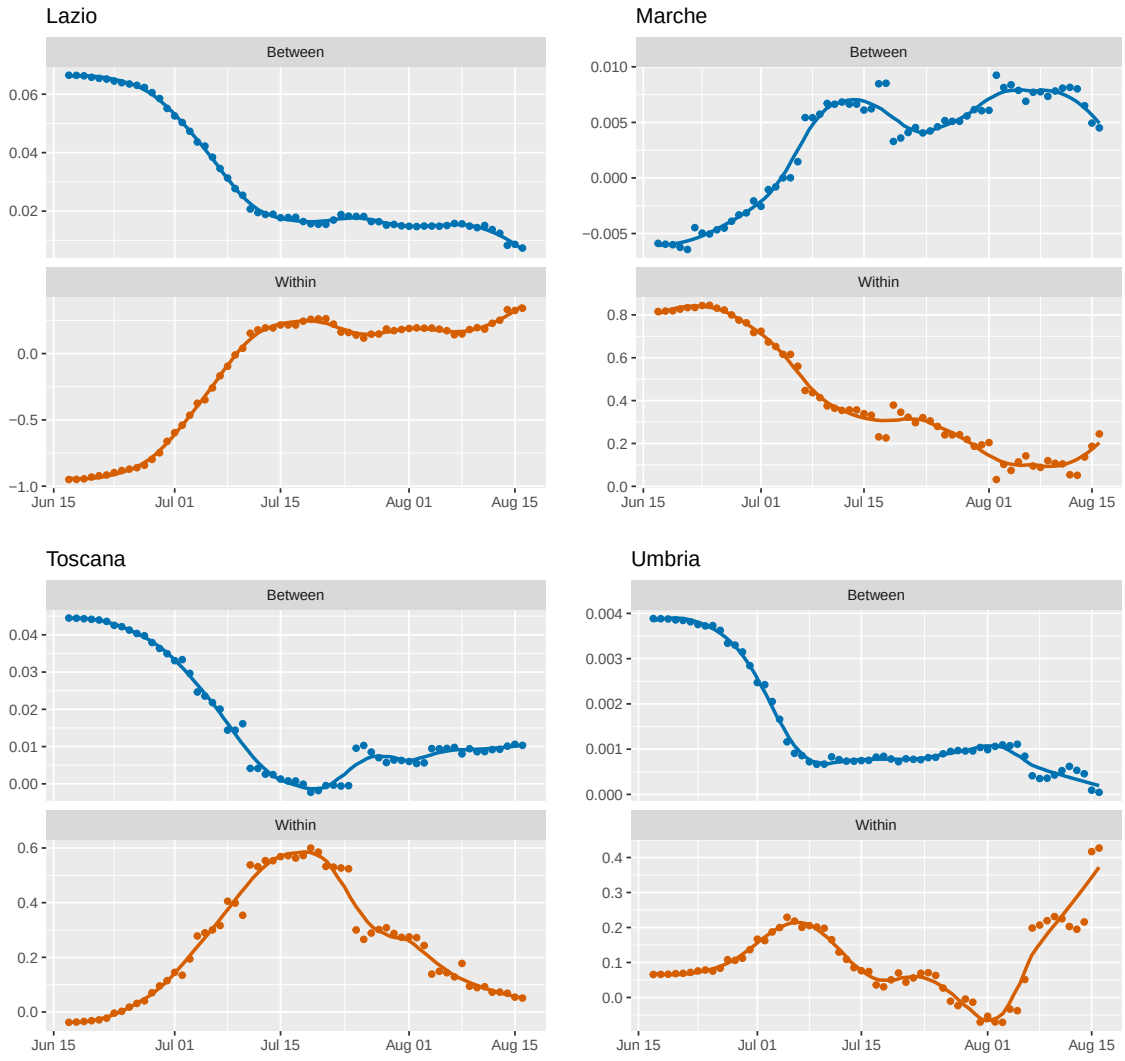


(d) With model selection by AIC;
including undocumented infectives

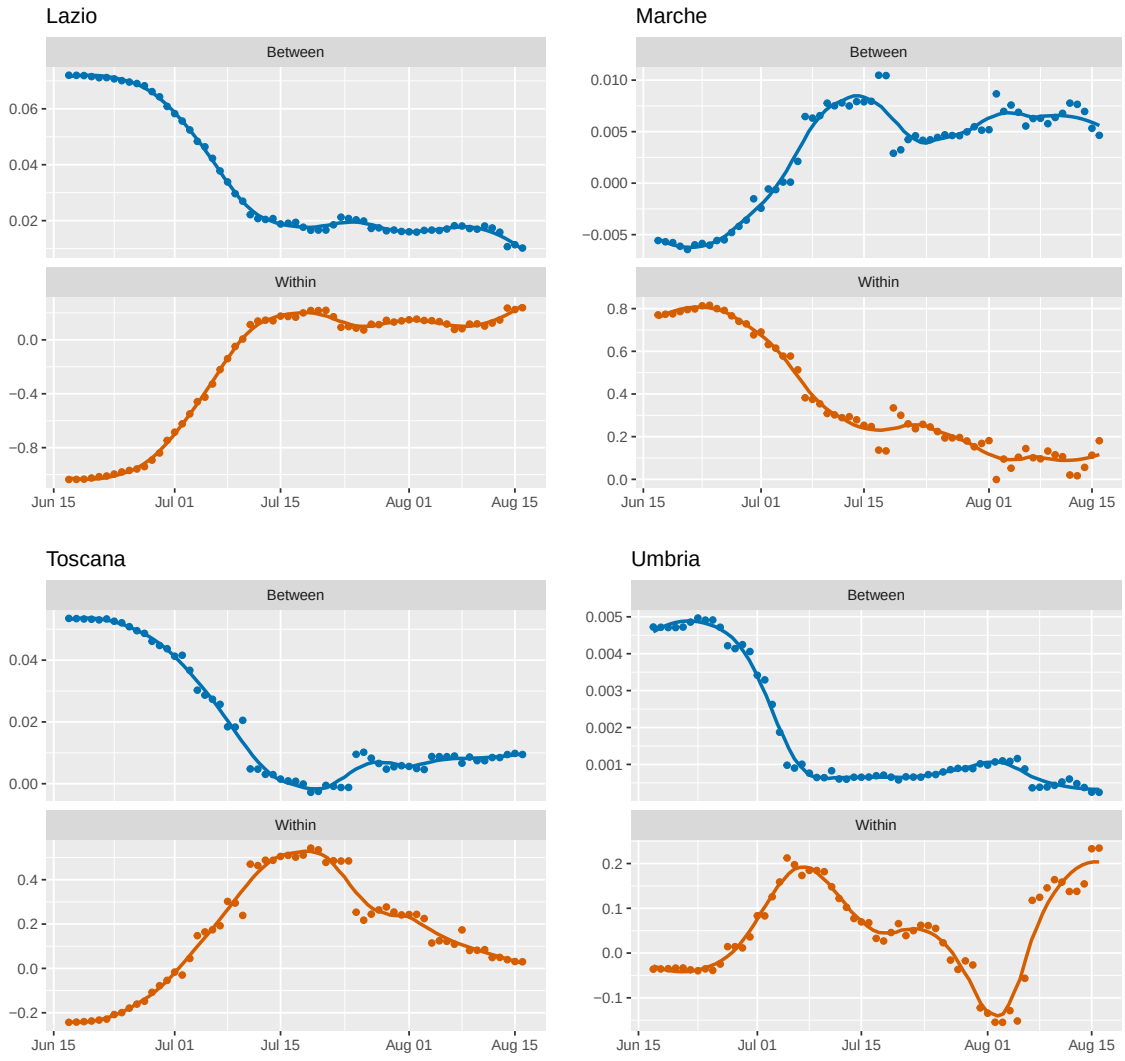
Figure C.9. Progression of β_{within} and $\beta_{between}$ over time for the Nord-Ovest NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



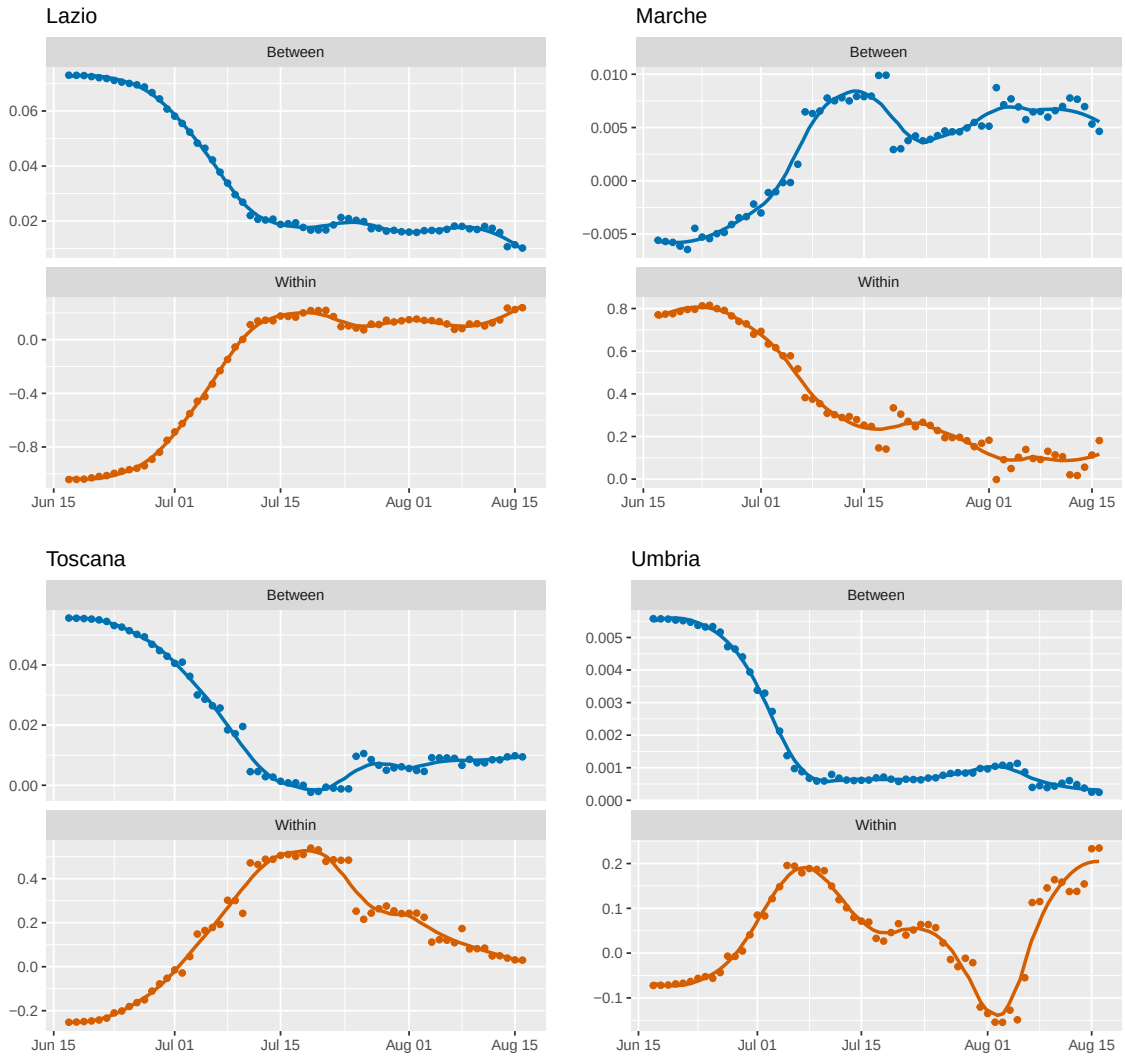
(a) Without model selection



(b) With model selection by AIC

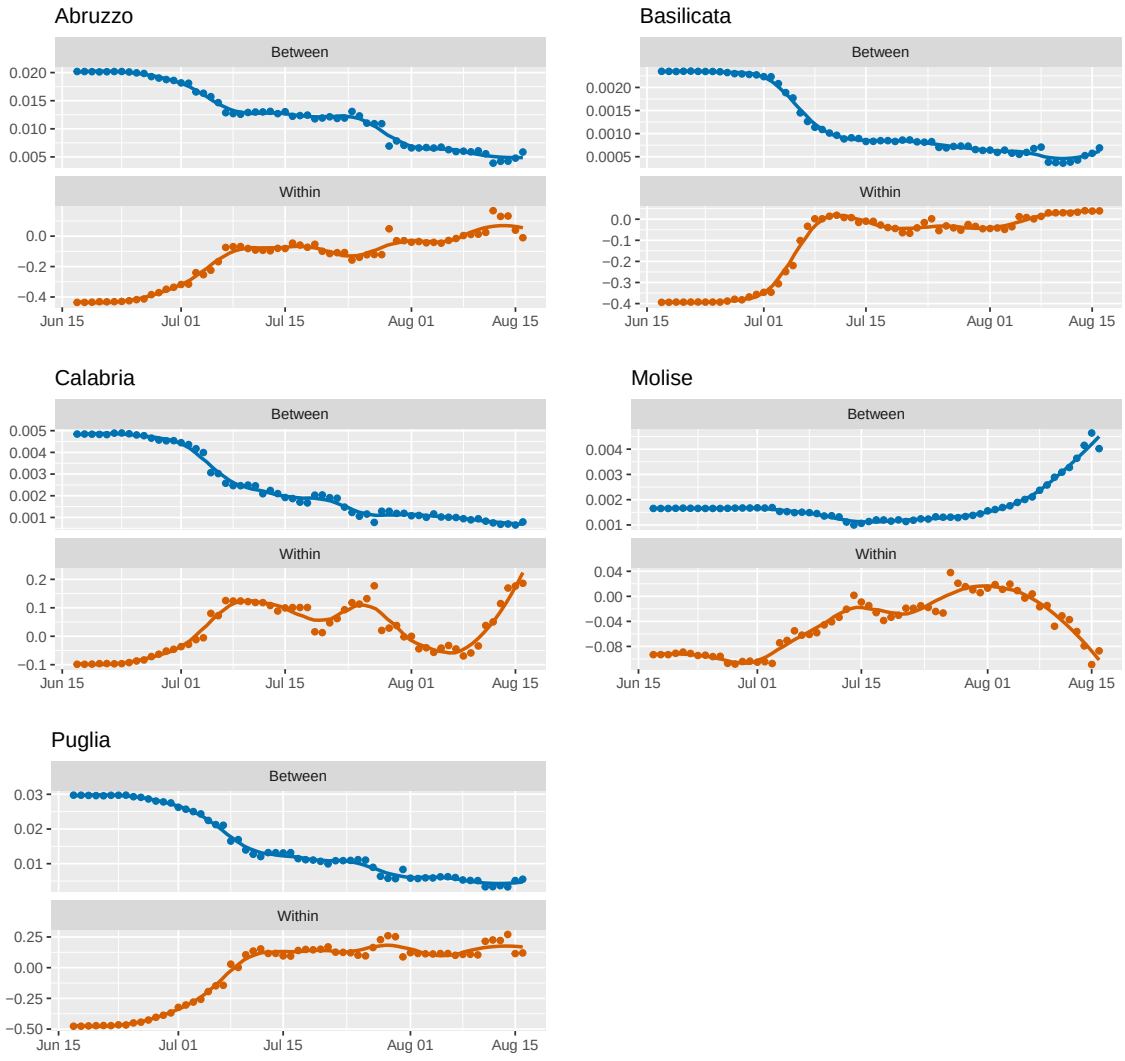


(c) Without model selection;
including undocumented infectives

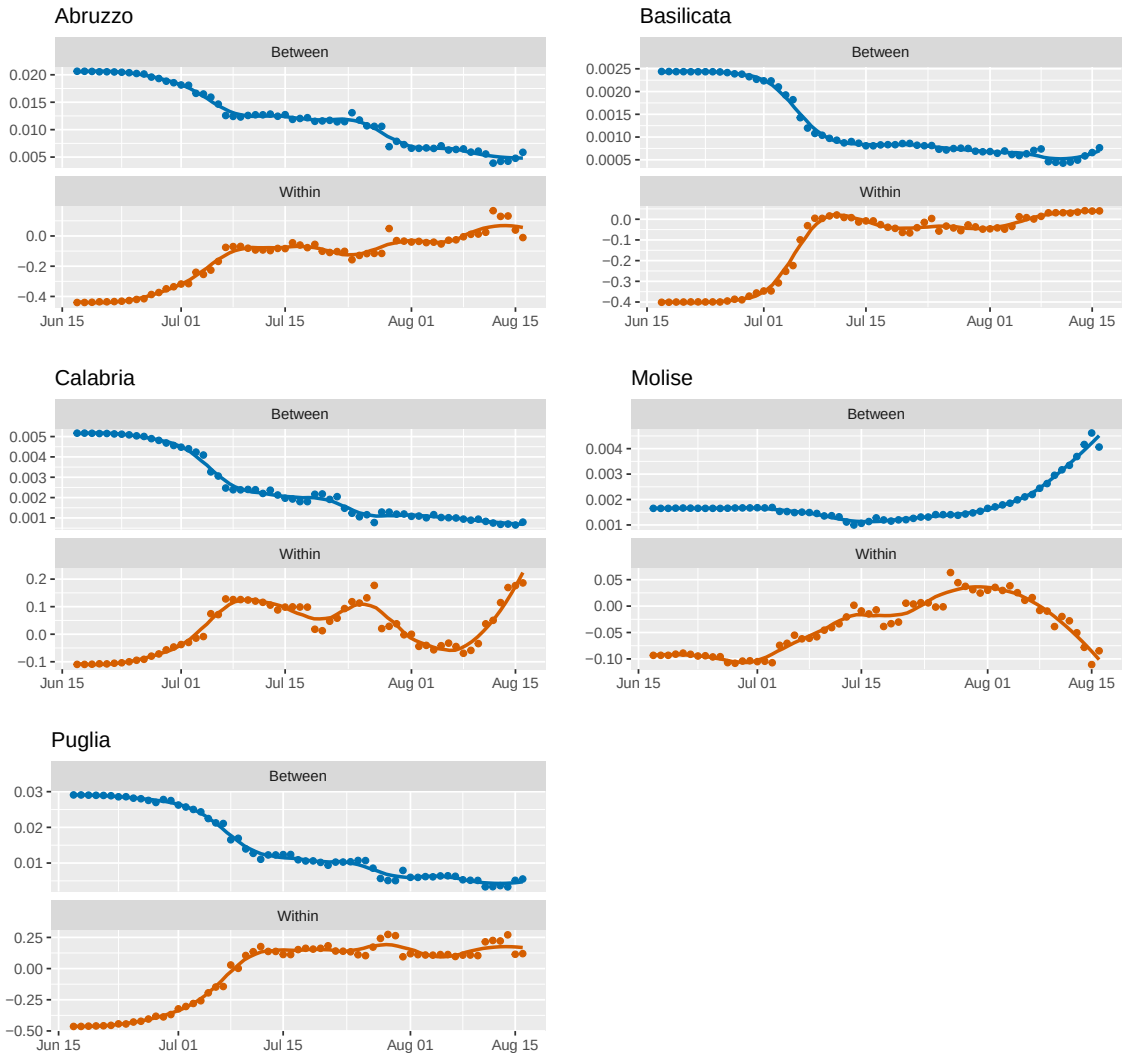


(d) With model selection by AIC;
including undocumented infectives

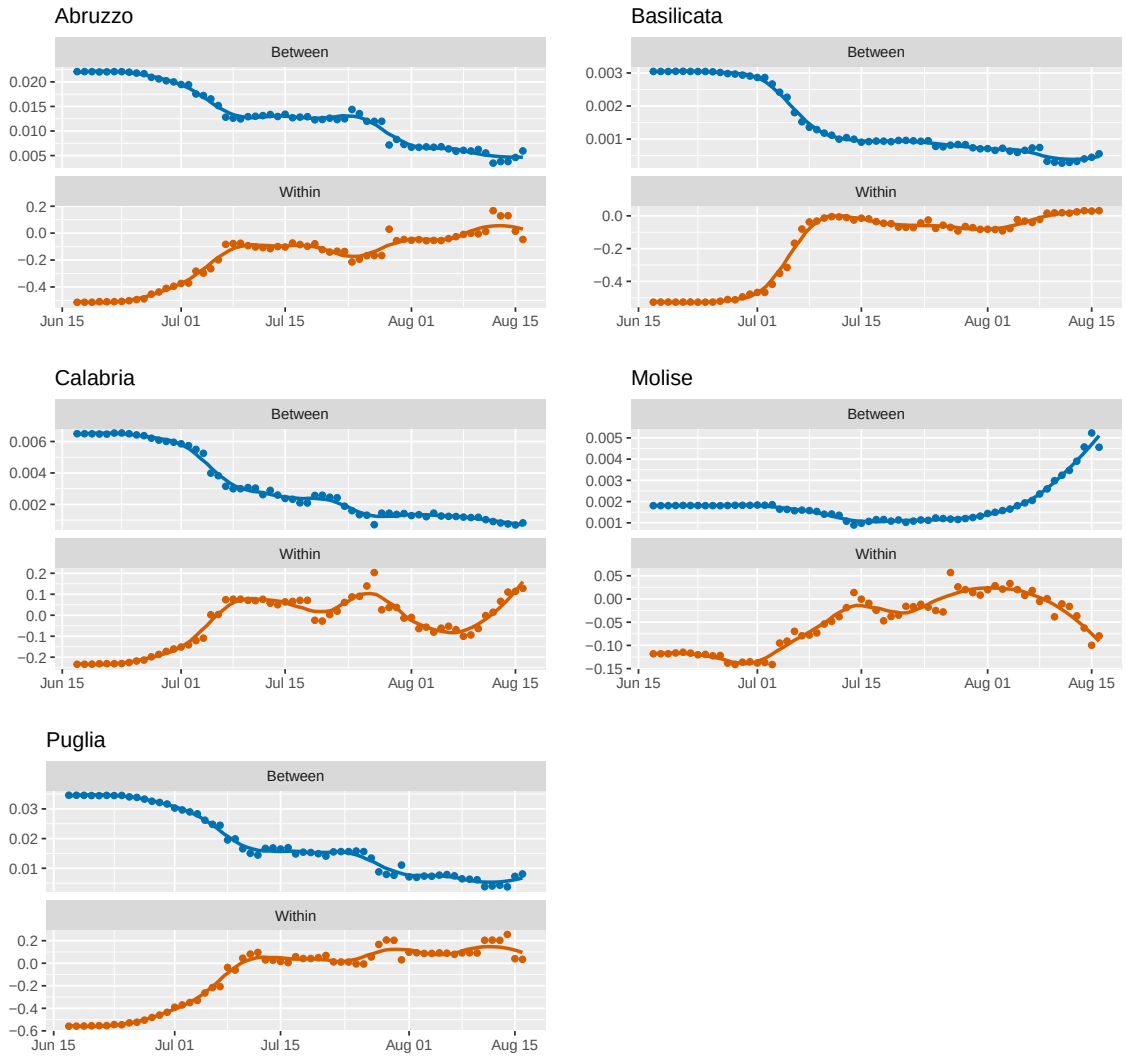
Figure C.10. Progression of β_{within} and $\beta_{between}$ over time for the Centro (IT) NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



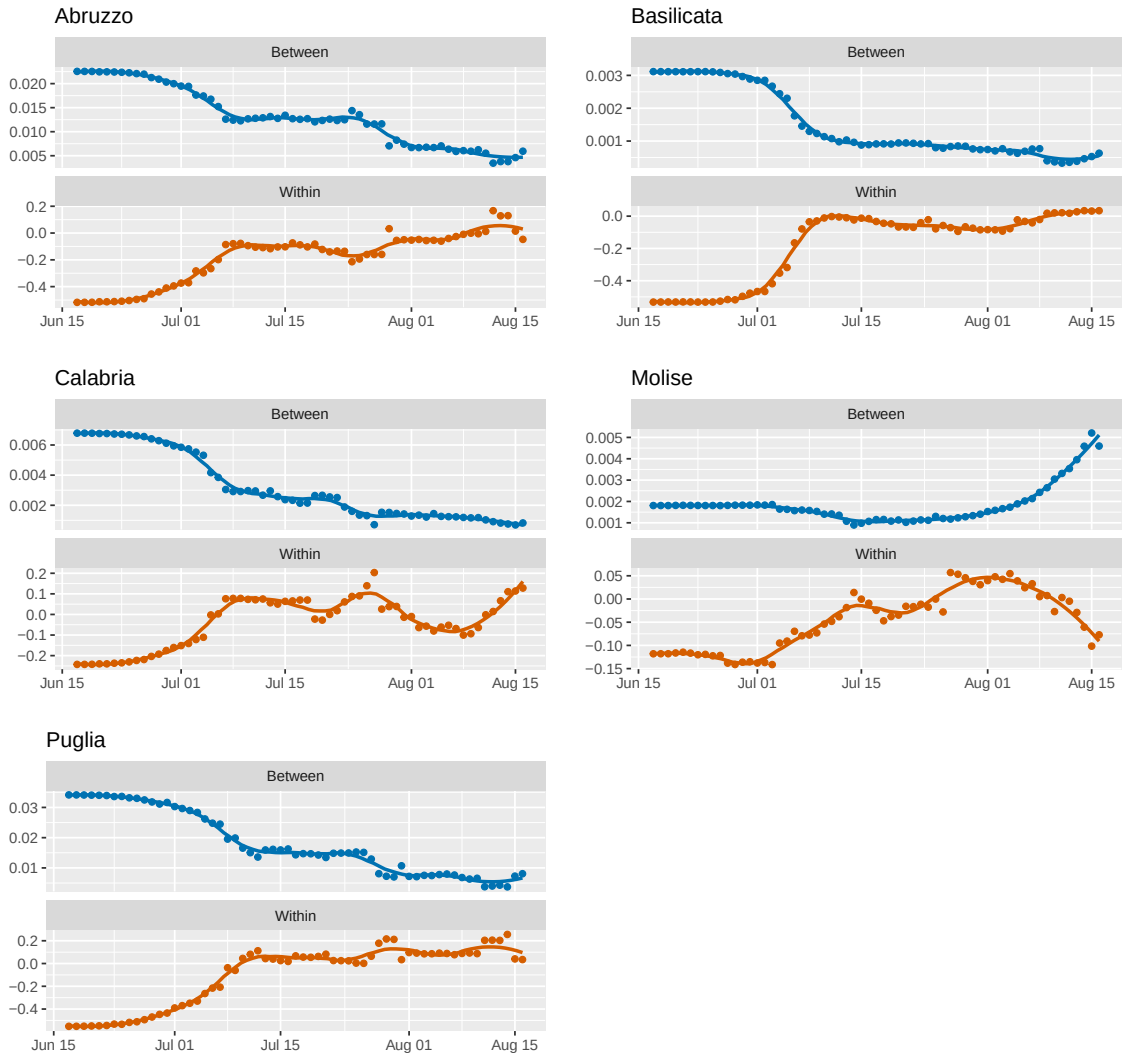
(a) Without model selection



(b) With model selection by AIC

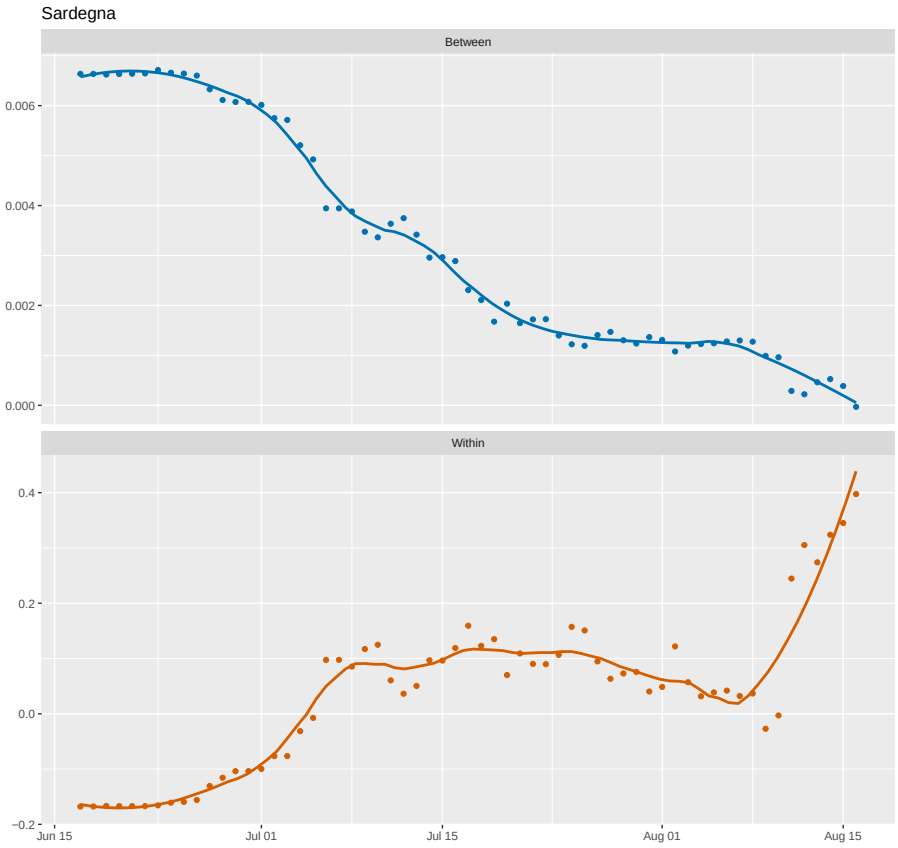


(c) Without model selection;
including undocumented infectives

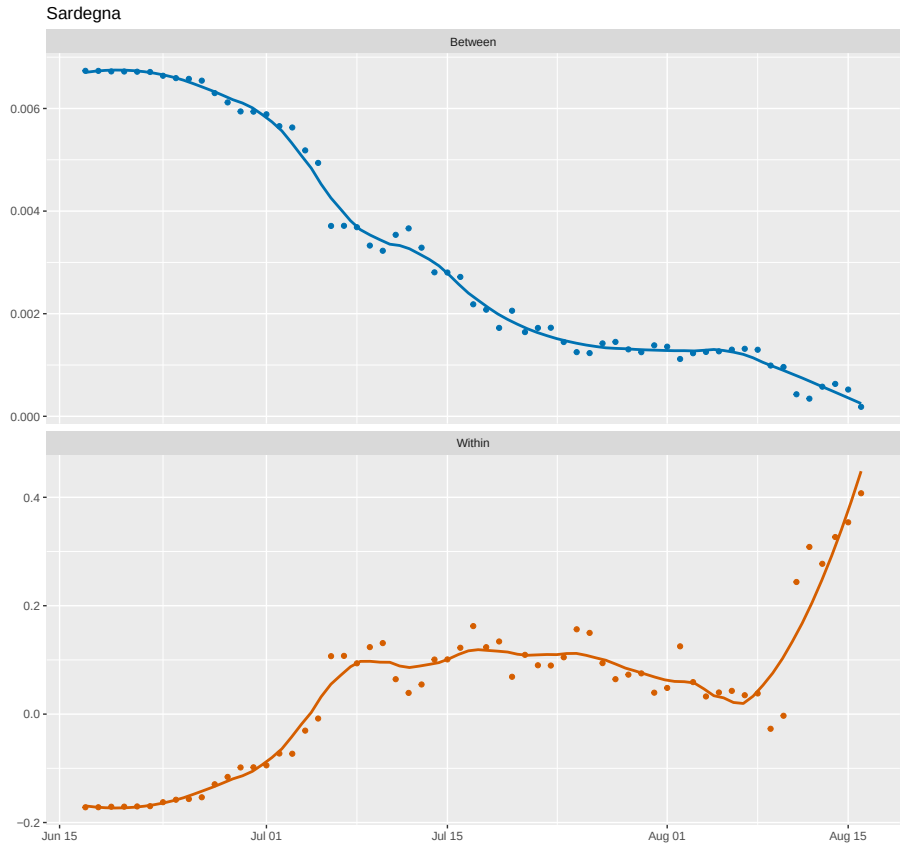


(d) With model selection by AIC; including undocumented infectives

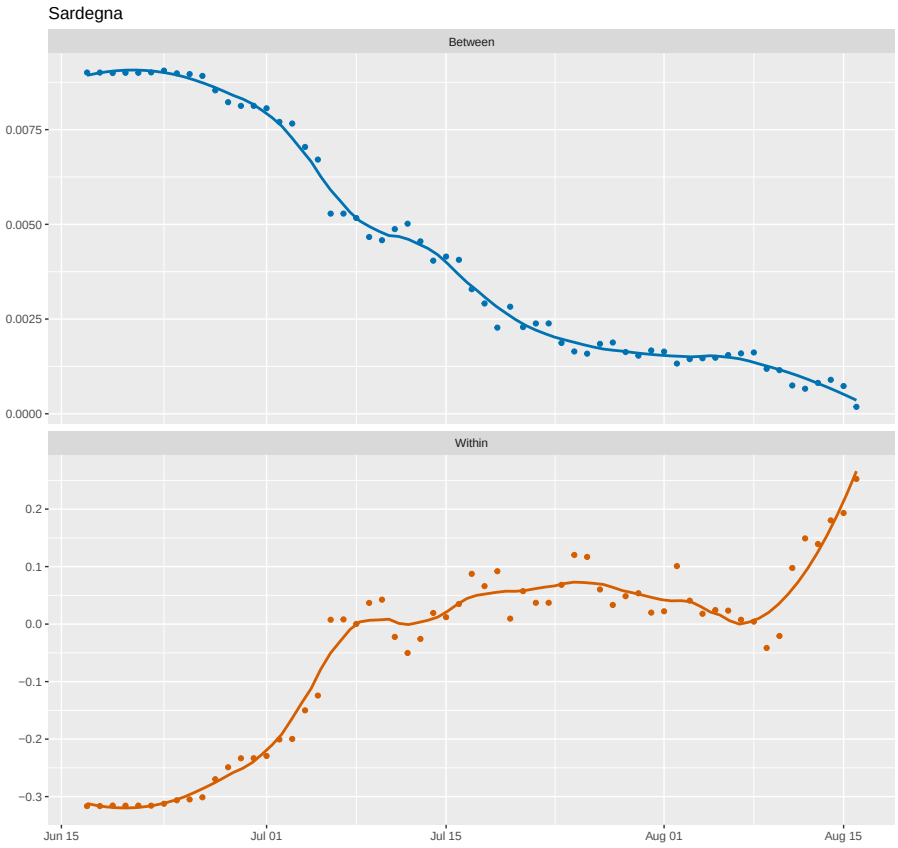
Figure C.11. Progression of β_{within} and $\beta_{between}$ over time for the Sud NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



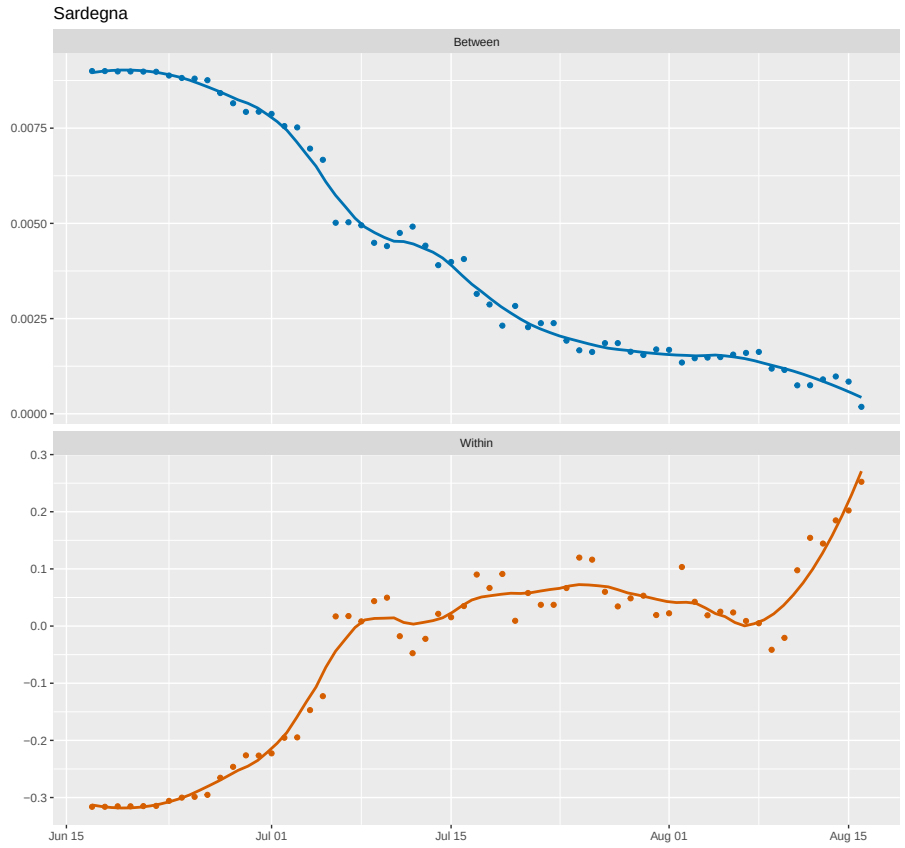
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



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

Figure C.12. Progression of β_{within} and $\beta_{between}$ over time for the Isole NUTS 1 region. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

D Discrete SIR Model

In Section 10, we referenced an approach that aimed to discretize the SIR model to allow for the estimation of both the transmission rate and recovery rate, with the goal of obtaining an estimate of the effective reproduction number. Especially, the approach aimed to take into account the spatiotemporal nature of the data by applying panel data estimation techniques. In this appendix, we highlight the most important parts of this method, focusing on the methodology and presenting some results.

D.1 Methodology

In this appendix, we discuss the methodology behind the discretized SIR model. Starting from the equations in Section 4, we acknowledge the addition by Adda (2016) of a longer lag to take the incubation period into account. This leads to the following discretized equations:

$$S_{p,t} - S_{p,t-1} = -\beta S_{p,t-\tau} I_{p,t-\tau} + \eta_{p,t}, \tag{D.1}$$
$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-\tau} I_{p,t-\tau} - \gamma I_{p,t-1} + \eta_{p,t}, \tag{D.2}$$
$$R_{p,t} - R_{p,t-1} = \gamma I_{p,t-1} + \eta_{p,t}. \tag{D.3}$$

Two methods were explored to estimate the parameters. Firstly, we can apply estimation methods to (D.1) and (D.3) individually to obtain the estimates for β and γ individually. We would call this the regular model. The second method would be to first estimate one of the parameters by means of (D.3) (D.1) and then to we would fill in this estimate in (D.2) to estimate the remaining parameter. If we use equation D.2 to estimate γ and use the resulting estimate $\hat{\gamma}$ in (D.2) to obtain:

$$\begin{aligned} I_{p,t} - I_{p,t-1} &= \beta S_{p,t-\tau} I_{p,t-\tau} - \hat{\gamma} I_{p,t-1} + \eta_{p,t} \\ \iff I_{p,t} - (1-\hat{\gamma})\, I_{p,t-1} &= \beta S_{p,t-\tau} I_{p,t-\tau} + \eta_{p,t}. \end{aligned} \tag{D.4}$$

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. The fixed effects model, in essence, assumes that each individual (region) has a time-constant intercept. The SIR model, on the other hand, does not include an intercept in its formulation. The reason behind this is intuitive: there is not some non-zero mean number of new cases that is persistent throughout time for a certain region. Because of this, the fixed effects model is not suitable for our estimation.

The main idea behind the random effects model is to impose a distribution on the regional effects that can then be included in the error term: $\alpha_i \sim N(0, \sigma_\alpha^2)$. This

assumption may indeed be in line with the SIR model because the mean heterogeneous effect is assumed to be zero. Lastly, the pooled OLS model ignores the regional effect, hence treating the data as one large cross-section. This means that the T observations for some region p are actually treated to be cross-sectional observations of T different individuals. We apply both random effects and pooled OLS and compare the results.

D.2 Results

In this section, we present the results for the discretized SIR model. Table D.1 shows the results of estimating the parameters excluding and including undocumented infectives for two values of the lag τ, namely $\tau = 1$ as in the original SIR model and $\tau = 14$ as in the models by Adda (2016). Note that the estimate of γ is the same regardless of the choice of τ. Recall that this is because γ is estimated from equation (D.3), in which no lag other than 1 is used.

Table D.1. Estimates for the discretized SIR model with panel data methods. Estimates are given with t-statistics (for POLS) or z-statistics (for RE) in parentheses. Data spans May 9 until 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	
τ	Parameter	*POLS*	*RE*	*POLS*	*RE*
	γ	4.886×10^{-3}***	4.425×10^{-3}***	6.187×10^{-4}***	5.888×10^{-4}***
		(26.859)	(13.888)	(27.537)	(15.606)
1	β	6.501×10^{-3}***	4.879×10^{-3}***	1.886×10^{-3}***	9.089×10^{-6}***
		(30.680)	(9.310)	(29.100)	(0.029)
	$\beta_{two\text{-}step}$	6.769×10^{-3}***	3.958×10^{-3}***	1.920×10^{-3}***	-1.198×10^{-4}***
		(107.921)	(14.120)	(36.995)	(−0.420)
	R_{eff}	1.331	1.103	3.048	1.544
	$R_{eff;two\text{-}step}$	1.385	0.894	3.103	−0.203
14	β	6.356×10^{-3}***	2.028×10^{-3}***	1.821×10^{-3}***	-3.641×10^{-4}***
		(28.459)	(3.403)	(26.925)	(−0.525)
	$\beta_{two\text{-}step}$	6.853×10^{-3}***	1.901×10^{-5}***	1.877×10^{-3}***	-3.358×10^{-3}***
		(96.760)	(0.060)	(34.469)	(-11.315)
	R_{eff}	1.301	0.458	2.943	−0.618
	$R_{eff;two\text{-}step}$	1.403	4.296×10^{-3}	3.034	−5.703

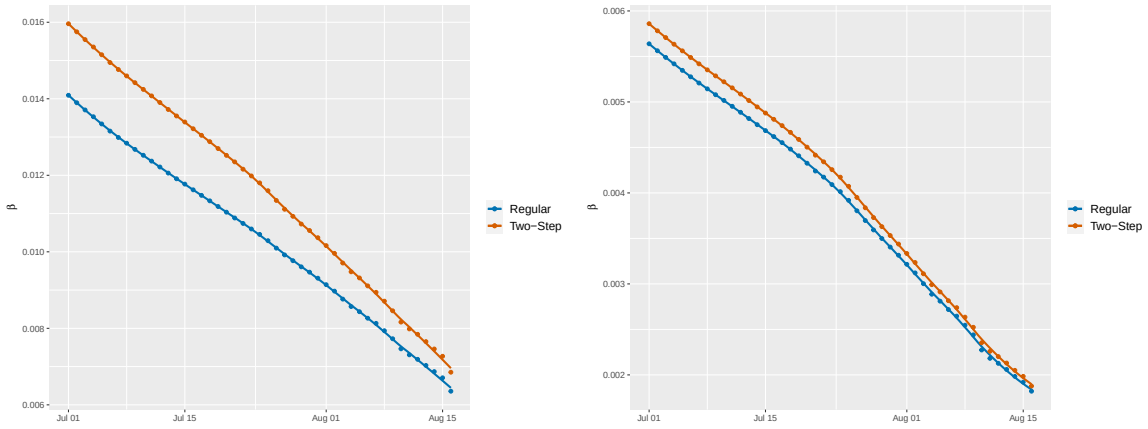
Significance levels: * = 0.1 ** = 0.05, *** = 0.01

Table D.1 shows us that all estimates are statistically significant at a significance level of 0.01. Firstly, notice that the random effects model when including undocumented infectives yields negative estimates, which are not possible. As such, it is not applicable to this situation. This also becomes clear when considering the results on R_{eff}, which are too low for the random effects model. It is also interesting to compare

the values of R_{eff} when modelling undocumented infectives. Indeed, the effect that the values of R_{eff} are larger when undocumented infectives are included is higher, is to be expected.

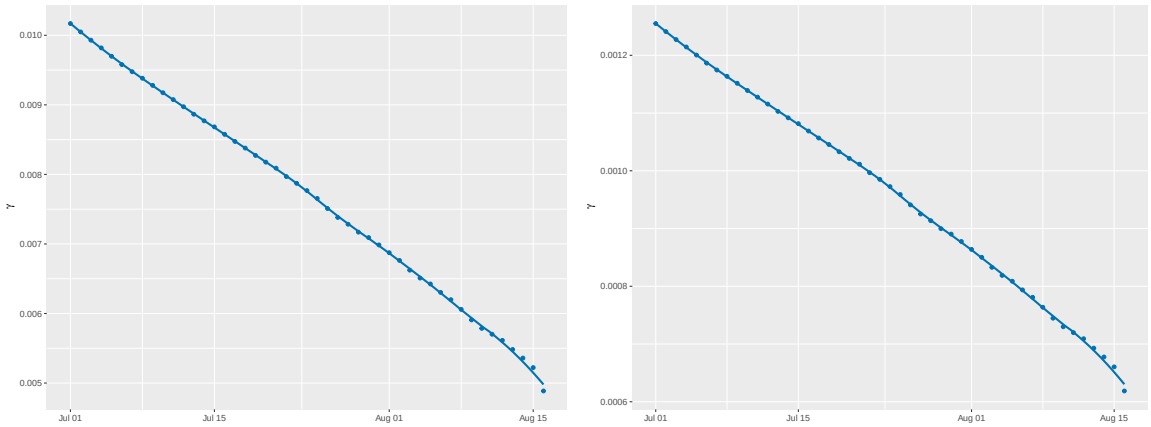
Unfortunately, all estimates are quite low. Consider the estimate of $\gamma = 4.886 \times 10^{-3}$. This implies that the average infectious period is $\gamma^{-1} = 204.67$ days, which is not credible. Unfortunately, this means that the estimates are not individually interpretable. However, it may be possible that the resulting value of R_{eff} is indeed correct if the estimates of β and γ both differ from their true value by the same factor.

Figures D.1, D.2, and D.3 show the progression of the transmission rate β, the recovery rate γ, and the effective reproduction number R_{eff} over time, respectively. Each point in the graphs is the estimate of β, γ, or the resulting value of R_{eff} 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.



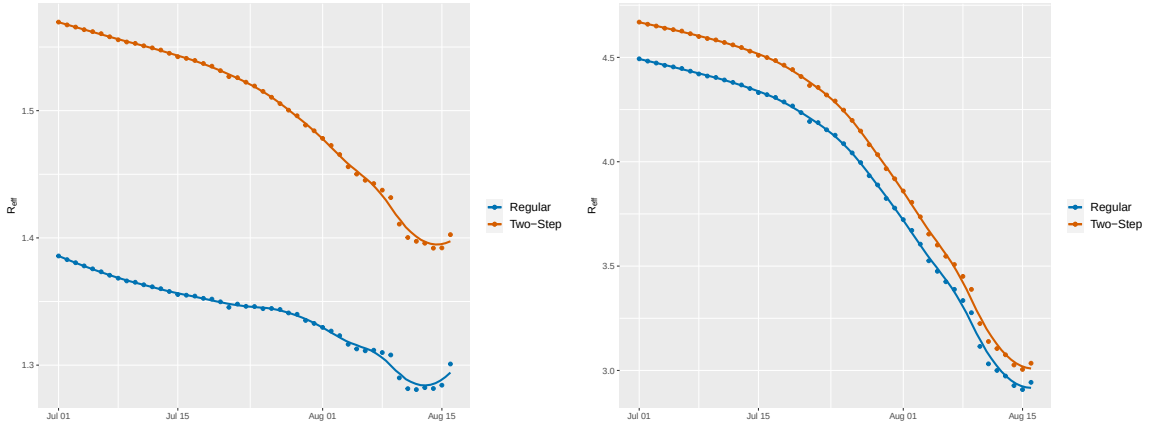
(a) Excluding undocumented infectives **(b)** Including undocumented infectives

Figure D.1. Progression of the transmission rate β over time. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



(a) Excluding undocumented infectives **(b)** Including undocumented infectives

Figure D.2. Progression of the recovery rate γ over time. With a rolling window, data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



(a) Excluding undocumented infectives **(b)** Including undocumented infectives

Figure D.3. Progression of the effective reproduction number R_{eff} over time. With a rolling window, data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Figure D.3 shows that the values for R_{eff} vary more when undocumented infectives are included, namely from around 2.8 until 4.5 (for the regular β). On the other hand, if undocumented infectives are not modelled, it only ranges from 1.3 until 1.4. Luckily, the values for β and R_{eff} decrease over time, indicating a less bad situation of the pandemic. On the other hand, the recovery rate γ also decreases over time, although the range of values that it equals does not change much, namely from around 0.005

until 0.01. We explained in Section 4 that the recovery rate is expected to stay constant over time since it is a biological parameter.

E Derivations

In this appendix, we provide mathematical derivations. Appendix E.1 explains how the total population as well as the susceptible population is computed. In Appendix E.2 we give the derivations for the functional forms for modelling undocumented infectives as discussed in Section 6.

E.1 Calculation of Population Variables

In this appendix, we 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 p at time t by $N_{p,t}$. 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 p on February 25, 2020 is calculated as:

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

where $d_{p,t}$ denotes the number of deaths in region p 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 p at time t, denoted by $s_{p,t}$, is therefore calculated as follows:

$$s_{p,t} = N_{p,t} - i_{p,t} - r_{p,t} \tag{E.2}$$

where $i_{p,t}$ denotes the number of infectives and $r_{p,t}$ denotes the number of removed individuals. Recall that r is made up by adding the recovered individuals and the

deceased individuals. Because we use the calculation of $N_{p,t}$ as in the previous paragraph, the error discussed propagates into the calculation of $i_{p,t}$. However, as before, we assume that this error is negligible.

E.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 6.

E.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:

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

E.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 $\beta, \gamma \in (0, 1)$,

(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:

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

To solve (E.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:

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

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 recall 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:

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

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

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

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

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

Using (E.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:

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

This is only the case if two conditions are satisfied:

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

Now 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, (E.6a) tells us that both the numerator and denominator of the fraction are negative. Therefore, to satisfy (E.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 (E.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:

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

This is only the case if one of the following two conditions is satisfied:

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

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 (E.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 $\gamma \in \left[\frac{1}{4} + \frac{3}{4} f^{min}, \frac{3}{4} + \frac{1}{4} f^{min}\right]$.

E.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:

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

E.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 E.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).

E.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 $\beta_1, \beta_2, \gamma_1, \gamma_2 \in [0, 1]$ 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:

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

In appendix E.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 (E.8), we can then apply row reduction as follows:

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

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

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

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

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

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

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

Conclusively, we have derived that:

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

so that:

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

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