Predicting local and national COVID-19 outbreaks: the case of Italy

by
Mike Weltevrede (ANR 756479)

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

Tilburg School of Economics and Management
Tilburg University

Supervised by:
Dr. Otilia Boldea

Second reader:
Dr. Pavel Cizek

Date:
September 24, 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, which R. Li et al. (2020) make up a large part of the total number of infectives. 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 regions such as Lombardy. Moreover, we find that the models are able to locally forecast the number of infectives across regions. Lastly, we investigate that the national lockdown in Italy has been effective in reducing the transmission of SARS-CoV-2 over time. Consequently, the approaches in this thesis can also be applied to other nations and for forecasting future epidemic outbreaks.

Acknowledgements

First of all, I would like to say that my heart goes out to all those whose lives or those of loved ones have been impacted by COVID-19. As King Willem-Alexander of the Netherlands beautifully stated on March 20, 2020: *"We deeply sympathise with the relatives of those who have died, and with those who have contracted the virus and are currently at home or in hospital. Our thoughts are with you at this difficult time. [...] 2020 will be a year to remember. Everyone will experience it differently. But I hope and believe that feelings of solidarity and pride will prevail and bring us all closer, as we get through this most difficult of times together."* (Royal House of the Netherlands, 2020)

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, 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 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.

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 shout-out 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 **1**

2 Problem Description **2**

3 Dataset **4**
3.1 Geographical Structure of Italy 5
3.2 Coronavirus Data 6
3.3 Regressors 9

4 SIR Model **10**

5 Modelling Undocumented Infectives **11**

6 Within-Region Spread Model **21**
6.1 Methodology 21
6.2 Results 24

7 Within and Between-Region Spread Model **34**
7.1 Methodology 34
7.2 Results 36

8 Forecasts **40**

9 Conclusion **45**

10 Future Research **46**

Appendices **50**

A Abbreviations **50**

B Tables **51**
B.1 Results for the Within-Region Spread Model 51
B.2 Results for the Within and Between-Region Spread Model 54

C Figures **56**
C.1 Figures for Section 2: Problem Description 56
C.2 Figures for the Within-Region Spread Model 59
C.3 Figures for the Within and Between-Region Spread Model 67
C.4 Figures for Section 8: Forecasts 71

D Discrete SIR Model **83**
D.1 Methodology 83
D.2 Results 84

E Derivations **87**
E.1 Calculation of Population Variables 87
E.2 Functional Forms for Modelling Undocumented Infectives 88
E.2.1 Linear Function 88
E.2.2 General Quadratic Function 88
E.2.3 Special Case Quadratic Formula: Downwards Opening 92
E.2.4 Special Case Quadratic Formula: Upwards Opening 92
E.2.5 Cubic Function 93

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 to two metres distance to one another, causing a massive change in social behaviour.

It is crucial to understand viral diseases inside and out. This 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 decisions. This thesis has the goal to estimate a model to describe and predict local and national outbreaks of SARS-CoV-2. 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 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. 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) found that around 86% of the infectives in China in the early stages of the pandemic went undocumented and that these were also contagious. 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. To that end, we estimate two models as presented by Adda (2016) to take into account these regional effects. 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 on economic activity. 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, Section 5 defines and explores our method for modelling undocumented infectives. Thereafter, in Section 6, we discuss the within-region spread model as presented by Adda (2016). The within and between-region spread model is discussed in Section 7. Next, we provide forecasts in Section 8 to evaluate the predictive power of the models. 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 where it has made its way to nearly every country in the world. At the moment of writing, only twelve sovereign member states of the United Nations reported no infections, of which ten are island countries. The other two countries are North Korea and Turkmenistan. On September 22, 2020, over 31 million people were reported to have been infected with COVID-19. This has led to nearly a million deaths, making it the fourteenth most deadly viral disease to ever have existed (LePan, 2020). The World Health Organization (WHO) declared a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 (WHO, 2020a).After the spread of SARS-CoV-2 only became worse, the WHO declared the virus outbreak to be a pandemic on March 11, 2020 (WHO, 2020b).

Due to the extreme nature of the pandemic and the availability of enough data, this thesis chooses to focus on Italy. Until the end of March, Italy 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 and reported the second highest death-to-cases ratio of 14.45%. 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).

Our contribution is to model the pandemic at the regional level rather than at the national level. We believe that there is likely regional heterogeneity, meaning that there is not one model that can be used to analyze the entire country of Italy. In addition, the regional variation allows for better identification of the average transmission parameters.

Italy is subdivided in several regions. The largest regions are called NUTS 1 regions. These regions contain the so-called NUTS 2 regions, which is the level at which this thesis models. The geographical classification of regions is explained more thoroughly in Section 3.1. To illustrate the regional differences, we look at the number of new cases per capita (also called the incidence rate) for the four NUTS 2 regions that make up the Centro (IT) NUTS 1 region in Figure 2.1. The plots for the other NUTS 1 regions can be found in Appendix C.1.

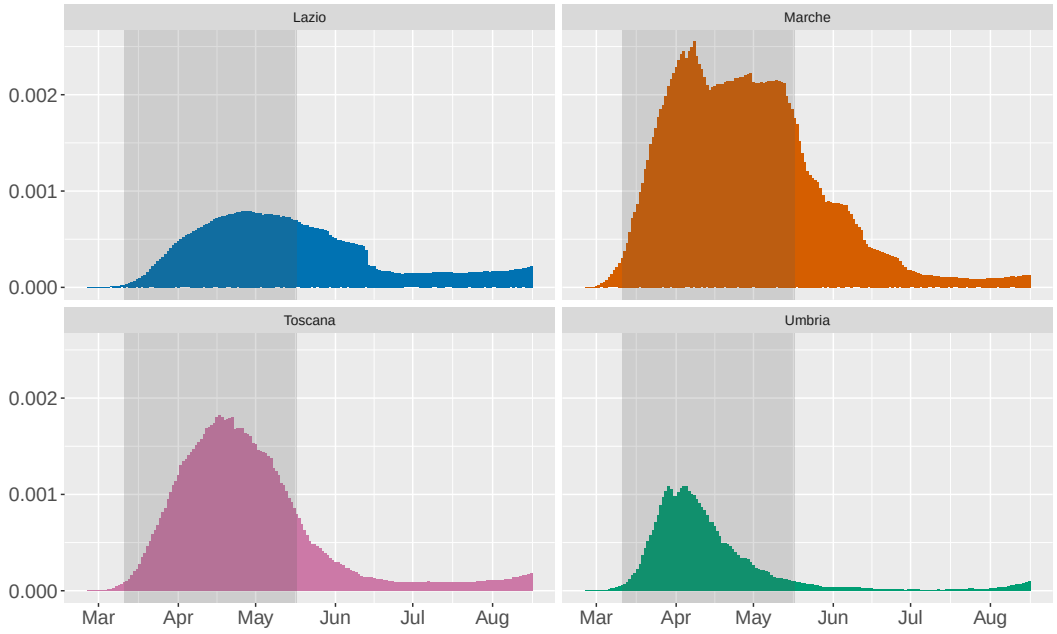


Figure 2.1. Incidence rate per NUTS 2 region for the Centro (IT) NUTS 1 region. The grey area indicates the national lockdown.

In Figure 2.1, we can see that there is a wide difference in the incidence rates between the regions. For the region of Lazio, we see a much lower peak than for the other three regions, especially compared to Marche. Moreover, there is a clearly varying length of the peak. This 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.

When developing models to predict the incidence, one should make sure that past data is selected carefully so that infections from the past do not dominate the latest estimates of the transmission rate parameters. For instance, notice the peaks in the incidence rates around April in Figure 2.1. 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. For this reason, for all models, we use the last 100 days of data, which spans May 9 until August 16. The reason behind choosing 100 days is arbitrary; it is simply a number that retains enough data points while providing variation in the estimates.

In this thesis, we present several models based on specifications presented by Adda (2016). In the paper, Adda (2016) investigates the spread of several viral diseases in the past. The models used are inspired by the Standard Inflammatory Response (SIR) model but deviate from it in the sense that the regressors are constructed with 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, such as the movement between regions and economic indices, are not available. 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. Due to the limited temporal scope of SARS-CoV-2, this information is not available. Consequently, this thesis only discusses the non-weighted models by Adda (2016). To our knowledge, 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 regressors that are included in the models presented 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 sub-regions of the *regioni*. Figure 3.1 presents a map of Italy with the NUTS 2 regions.[1]

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

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

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, 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 policies are often introduced at that level, such as a local lockdown put into place on March 7, 2020 before the national lockdown was instated. Unfortunately, the data outside of the total number of cases was not reported at this granular level. As such, we choose to use the NUTS 2 regions.

For $P = 21$ 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 and the number of hospitalized people who showed symptoms.[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 numbers of positive tests $(totale\_casi_t - totale\_casi_{t-1})$ equal 1, 0, and -1, respectively, while the numbers 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.

[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

There are two variables on the tests executed. The difference between the total amount of tests performed (*tamponi*) and the total amount of people tested *(casi testati)* is that the latter indicates the number of unique persons that were tested because individuals could have been tested more than once. *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*. In addition to the previous remarks, it is important to consider that there is a measurement error in the number of infectives, because there is no possibility that every citizen can be tested for COVID-19. For that reason, the actual number of infectives is higher that the official count as reported in the tables of the Department of Civil Protection. In Section 5, we discuss how the undocumented infectives are modelled.

With respect to the reported death statistics, there is a distinction between Italy and other countries. The Italian numbers include deaths of all patients who were tested positive for COVID-19 before or after their death. Belgian death counts, for instance, also include deaths of people who were suspected of having COVID-19, regardless of whether they were tested (Schultz, 2020). Moreover, Italian data makes no distinction between people who died because of COVID-19 and those who had the disease but 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 deceased patients, 48.6% had three or more comorbidities, 26.6% had two comorbidities, and 23.5% had one comorbidity. 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. 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). In this thesis, we assume that these errors are 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, Italian regions correct mistakes by having the report on a certain day compensate for the errors on the days before. 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 $t-1$, 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 day t and $t+1$. As such, these are left as is. Despite these errors, we assume that the official information is representative of the region for which it has been reported. Other inaccuracies are taken into account by the regression error term in our models.

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 weeks before that date were much lower. The same applies to Sicily, where a negative value of -394 was reported on June 19, 2020. 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 suspect that the Department of Civil Protection imputed the missing values with a value of zero. For instance, on July 5, it was reported that zero tests were executed in the region of Basilicata. On the dates surrounding July 5, however, around 250 tests were executed each day. On July 9, a higher value of 426 was reported. We suspect 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 Regressors

In this section, we describe the regressors that are included in the models by Adda (2016). Both models include regressors that may not directly have an effect on the transmission rate. We cannot include many of these regressors because, as we explained in Section 2, we only have access to relevant data since December 2019. This means that we do not have access to much time-varying information, for instance on seasonality of the virus as well as economic indicators. The only variable that is included is a dummy variable that indicates if the day is on the weekend (Saturday or Sunday). We do not include an intercept because this would be in stark violation of epidemiological models; there is not some nonzero 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 suspect 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 number of individuals who are susceptible to being infected, I denotes the number of individuals who are currently infected, also called infectives, and R denotes the number 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 fraction of susceptible individuals, i to be the fraction of infectives, and r to be the fraction 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 that tell 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. For simplicity's sake, we assume that this is not being influenced sufficiently by public health interventions (Adda, 2016). Finally, the SIR model assumes that there is a constant rate of change ω at which immune individuals lose their immunity. For instance, Adda (2016) sets ω to 0 for chickenpox as individuals acquire a lifetime immunity while ω will be high for gastroenteritis due to almost no immunity emerging. Although it has recently become clear that reinfection with COVID-19 is indeed possible (Bloomberg News, 2020), 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.

$$\frac{dS}{dt} = -\beta SI, \tag{4.1}$$
$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{4.2}$$
$$\frac{dR}{dt} = \gamma I. \tag{4.3}$$

Keeling and Rohani (2011) state that the SIR model can also be described by frequency-dependent transmission instead of density-dependent transmission, meaning that the variables s, i, and r are used instead of S, I, and R. This is given by:

$$\frac{ds}{dt} = -\beta si, \tag{4.4}$$
$$\frac{di}{dt} = \beta si - \gamma i, \tag{4.5}$$
$$\frac{dr}{dt} = \gamma i. \tag{4.6}$$

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 $R_{eff} > 1$ because this implies that $\beta > \gamma$, i.e. the spread of the virus exceeds the recovery rate. 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 reproduction rate for COVID-19 was below 1 in Italy, at between 0.5 and 0.7 (Ministero della Salute, 2020).

5 Modelling Undocumented Infectives

In this section, we discuss our method for modelling undocumented infectives. 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 it is not possible to test the entire population because the testing capacity is simply lacking. Otherwise, then all individuals who were tested positive could be isolated and the spread of the virus would be dampened tremendously. Since this is not possible, there are likely many infectives in society who spread the virus but who are undocumented. In China, at the beginning of the outbreak, 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 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 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. We assume that the number of undocumented individuals monotonically decreases 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 that the number of tests executed (*tamponi*) is used as a measure of the testing capacity. The total number of infected people at time t is denoted by I_t. This group can be subdivided into the documented infectives D_t and the undocumented infectives U_t such that $D_t + U_t = I_t$. Therefore, we can denote the documented and undocumented infectives as proportions of the total number of infected people. This proportion is therefore defined as a function of the testing capacity over time:

$$f_t \coloneqq f(TC_t), \tag{5.1}$$

such that

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

There are some properties and assumptions that equation (5.1) 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

$f(0) = f^{min}$. Denote the total population by N and the total number of removed individuals at time t by R_t (following the terminology and assumptions from the SIR model). The maximum amount of people that can get infected is then equal to $S_t^{max} := N - R_t$. Therefore, at any point in time, it should hold that

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

so f^{min} should be chosen to be larger than $\min_t \left\{ \frac{D_t}{S_t^{max}} \right\}$. 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 that are possibly susceptible, which is not possible.

(A3) If there is enough testing capacity such that the entire susceptible population can be tested, we assume that all infectives will be documented, so that:

$$f(S_t^{max}) = 1.$$

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. For this reason, one could choose to relax the assumption and assume $f(S_t^{max}) = f^{max}$ 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

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

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

- **Linear form**

$$f_t = \frac{1 - f^{min}}{S_t^{max}} TC_t + f^{min}. \tag{5.2}$$

- **Quadratic form**
  We specify three functional forms for a quadratic form. First of all, a general form is given, after which we discuss two special cases.

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

$$f_t = \frac{2 - 4\gamma + 2f^{min}}{\left(S_t^{max}\right)^2} TC_t^2 + \frac{4\gamma - 1 - 3f^{min}}{S_t^{max}} TC_t + f^{min}. \tag{5.3}$$

  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.

  – The first special case is the downwards opening vertex form. We assume that the vertex (also called the extremum) is the point $(S_t^{max}, 1)$, i.e. the parabola is downwards opening. 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:

$$f_t = \frac{f^{min} - 1}{\left(S_t^{max}\right)^2} TC_t^2 - \frac{2(f^{min} - 1)}{S_t^{max}} TC_t + f^{min}. \tag{5.4}$$

  This is equivalent to equation (5.3) 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. We assume that the vertex is the point $(0, f^{min})$, i.e. the parabola is upwards opening. The formula becomes:

$$f_t = \frac{1 - f^{min}}{\left(S_t^{max}\right)^2} TC_t^2 + f^{min}. \tag{5.5}$$

  This is equivalent to equation (5.3) 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 $f\left(\frac{1}{4}S_t^{max}\right) = \gamma_1$ and $f\left(\frac{1}{2}S_t^{max}\right) = \gamma_2$ for some $\gamma_1, \gamma_2 \in (0,1)$ such that $\gamma_1 < \gamma_2$. Then the formula becomes:

$$\begin{aligned} f(TC_t) =& \frac{8 + 64\gamma_1 - 48\gamma_2 - 24f^{min}}{3\left(S_t^{max}\right)^3} TC_t^3 \\ &+ \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14f^{min}}{\left(S_t^{max}\right)^2} TC_t^2 \\ &+ \frac{1 + 32\gamma_1 - 12\gamma_2 - 21f^{min}}{3S_t^{max}} TC_t + f^{min}. \end{aligned} \tag{5.6}$$

  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 $TC_t \in [0, S_t^{max}]$ 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 include regions by considering the regional testing capacity $TC_{p,t}$ and total susceptible population $S_{p,t}^{max} \coloneqq N_p - R_{p,t}$ instead. Then, the function would be dependent on p as well:

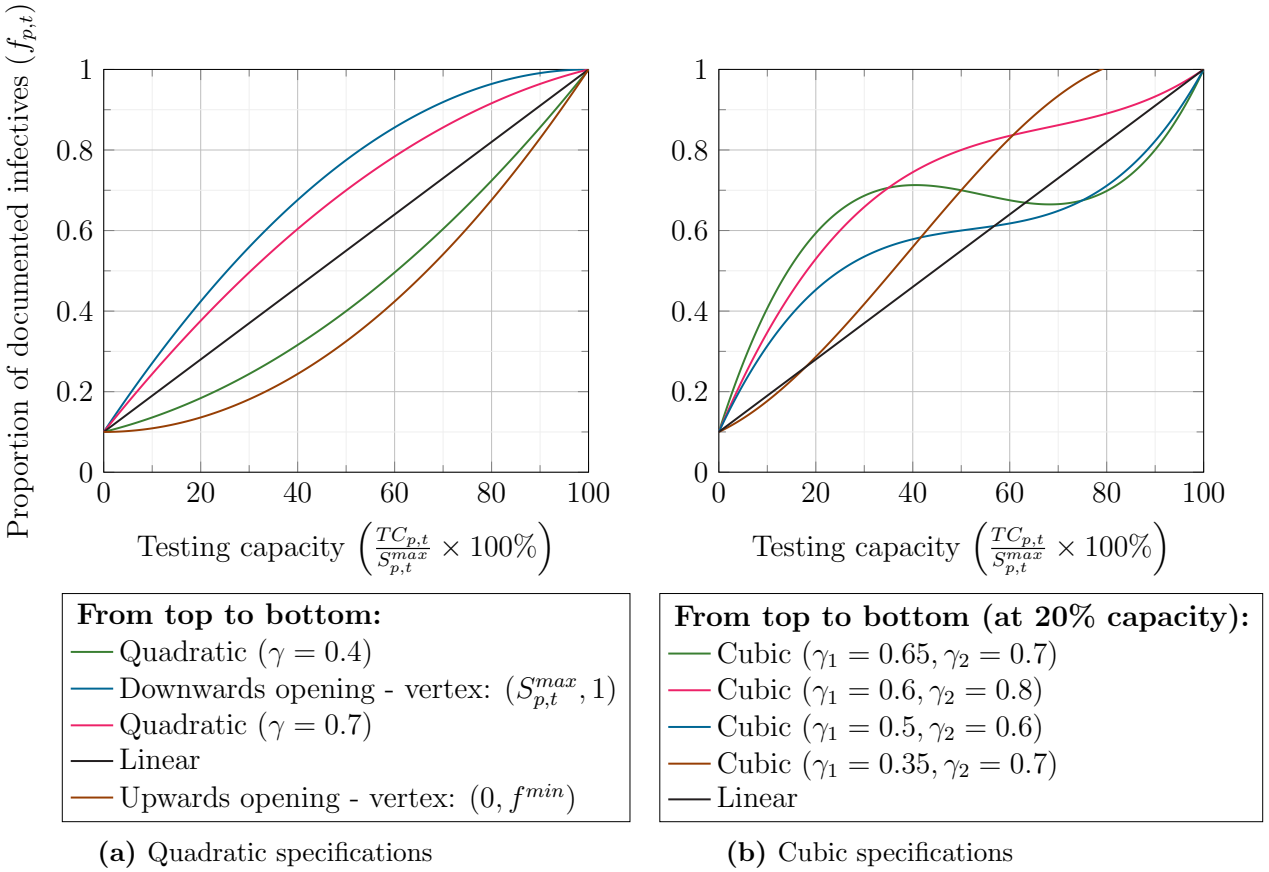
$$f_{p,t} \coloneqq f(TC_{p,t}). \tag{5.7}$$

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 5.1, we specify several functional forms for the specifications as mentioned above. Figure 5.1a shows four different functional forms for the quadratic functional forms while Figure 5.1b shows four different functional forms for the cubic specification.

Not all of the plots in Figure 5.1 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 5.1b 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).



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

Figure 5.1. Functional forms for the proportion of documented infectives ($f^{min} = 0.1$)

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 $R_{eff} > 1$, 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. 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 pool of people who have more severe 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 5.1b. This form also sees a sharp increase at the start of the graph, 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 convince the last proportion of the population to take a test who, at that point, may be infectious. 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. To be general, we choose equation (5.3) 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 $TC_{p,t}/S^{max}_{p,t}$. The results are shown in Figure 5.2.

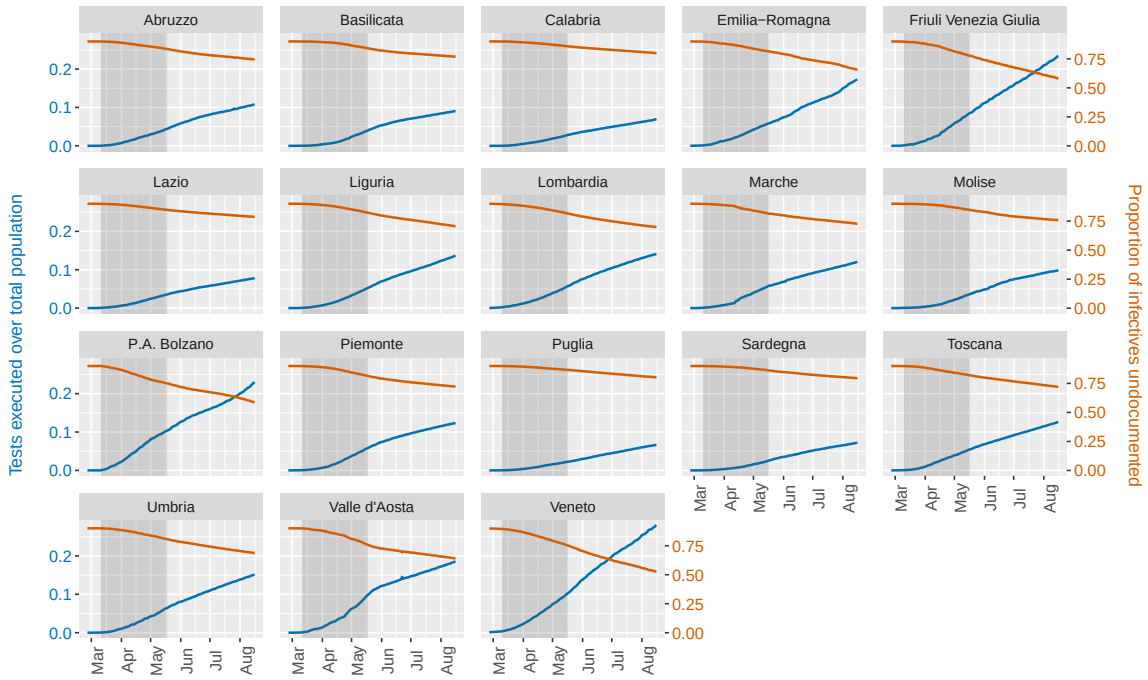
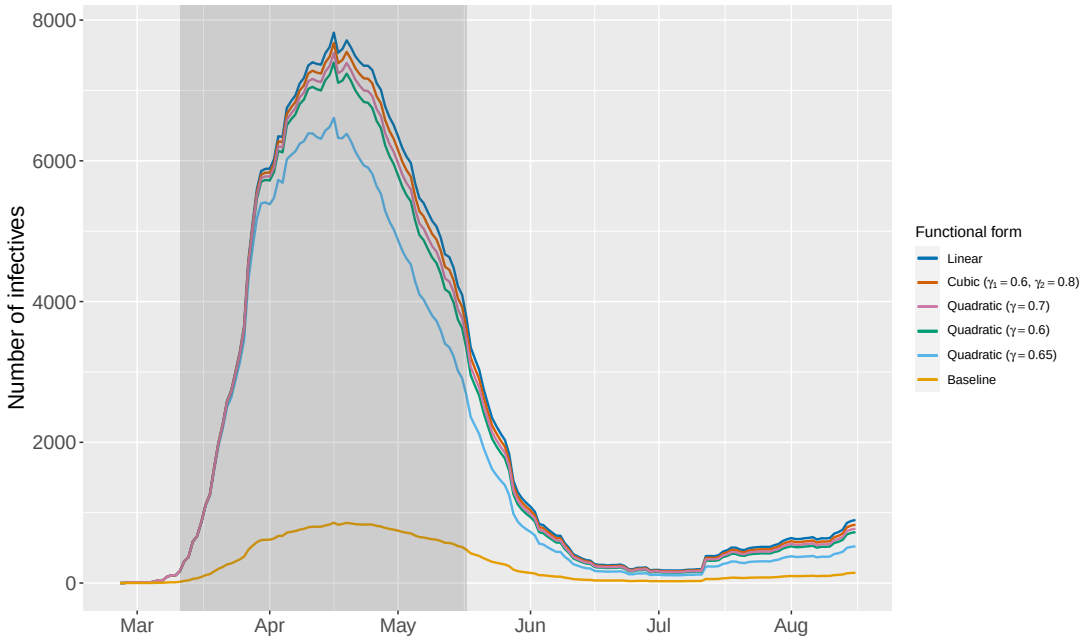


Figure 5.2. Total number of people tested over the total population ($TC_{p,t}/S^{max}_{p,t}$) versus proportion of infectives that are documented $f_{p,t}(\gamma)$. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

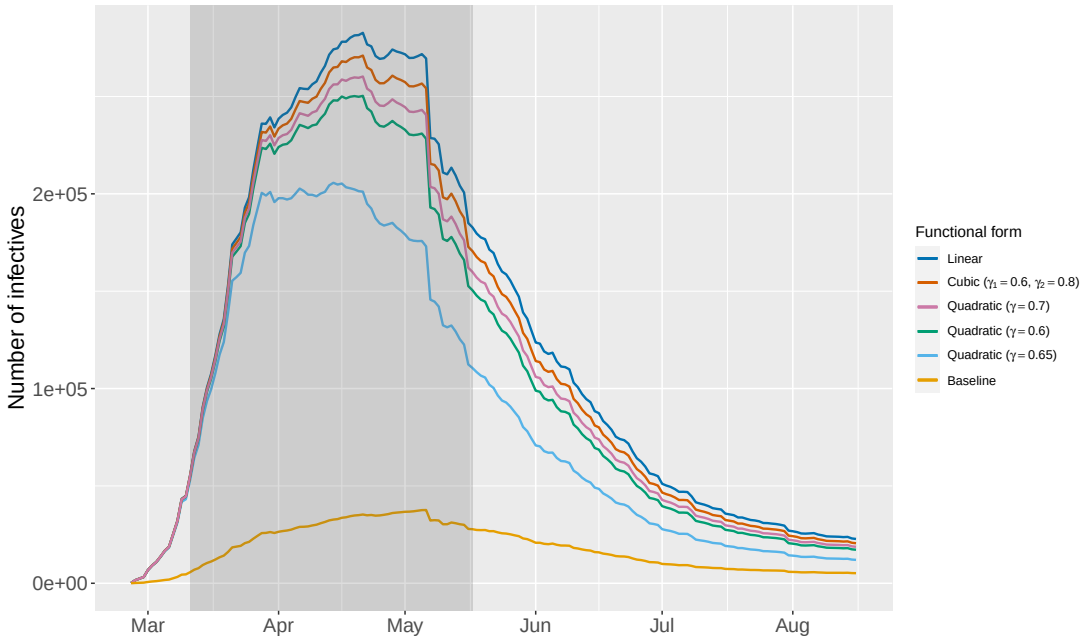
In Figure 5.2, we can see that the profile of the relationship between the two variables is similar over time for different groups of regions; the testing capacity increases over time, while the proportion of infectives that go undocumented decreases. There are also clear differences. Regions such as Friuli Venezia Giulia and Veneto have been testing a higher proportion of their population over time and see a steeper decrease in the proportion of undocumented infectives. 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 over time.

For three regions, namely Calabria, Lombardy, and Veneto, we illustrate the impact of this modelling method by comparing the documented number of infectives (baseline) with several functional forms of $f_{p,t}$ in Figure 5.3. We choose Calabria,

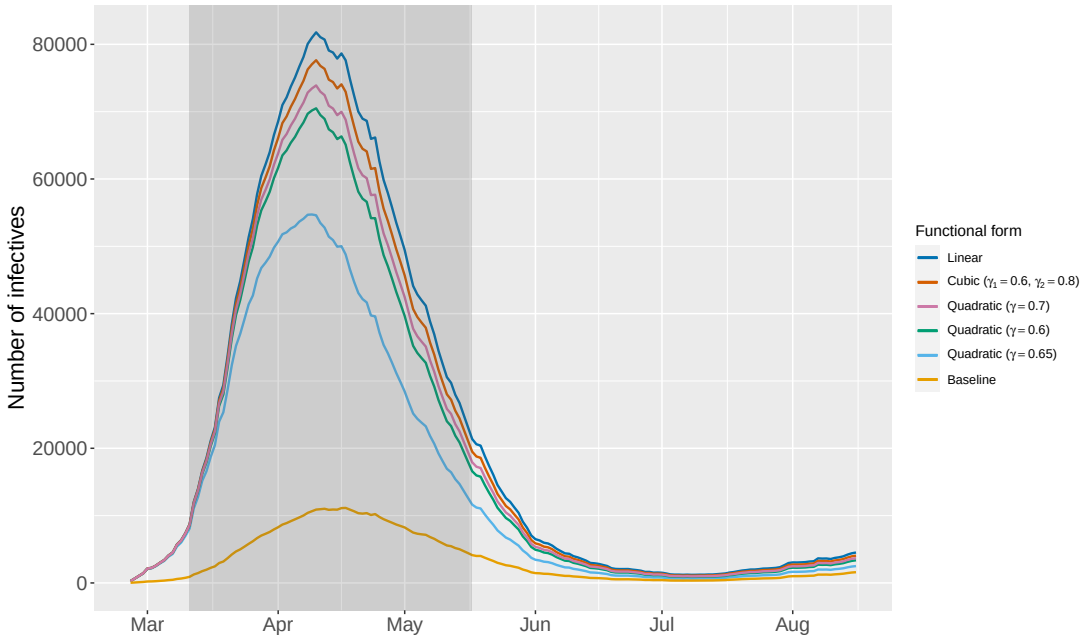
Lombardy, and Veneto because these vary in the proportional amount of tests executed, leading to different profiles in f_t, as can be seen in Figure 5.2.



(a) Calabria



(b) Lombardy



(c) Veneto

Figure 5.3. Comparison of the number of active infectives with several functional forms of $f_{p,t}$. The grey area indicates the national lockdown.

We can see that all functional forms have a huge impact on the total number of infectives compared to the baseline case, although they are all relatively close to one another. As such, it seems like the choice of one or the other form is unlikely to impact the estimates much. In the results sections, we will compare the effect on the estimated transmission rate using different functional forms.

We now give a concrete example of the impact in Table 5.1; we show the number of active documented infectives $D_{p,t}$, the proportion of infectives that are documented $f_{p,t}$, and the resulting total number of infectives $\tilde{I}_{p,t} := D_{p,t}/f_{p,t}$.

Table 5.1. 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		
	$D_{p,t}$	$f_{p,t}$	$\tilde{I}_{p,t}$	$D_{p,t}$	$f_{p,t}$	$\tilde{I}_{p,t}$	$D_{p,t}$	$f_{p,t}$	$\tilde{I}_{p,t}$
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 5.1 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 more strongly. 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.

Finally, note that modelling undocumented infectives also has an impact on how many people recover and die, thereby impacting the susceptible population. In this thesis, we do not take this effect into account. Future research can be done to incorporate this modelling method into the calculation of the number of recoveries and deaths.

6 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 6.1 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 Section 6.2, 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.

6.1 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 equation (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:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-1} I_{p,t-1} - \gamma I_{p,t-1}. \tag{6.1}$$

Because a model is never fully able to represent reality, we need to account for statistical errors when estimating the parameters in equation (6.1). This is incorporated in the model through an error term, denoted by $\eta_{p,t}$:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-1} I_{p,t-1} - \gamma I_{p,t-1} + \eta_{p,t}. \tag{6.2}$$

Next, 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. 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).

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. We discussed how we model pre-symptomatic individuals in the form of undocumented infectives in Section 5.

Adda (2016) models the transmission lag by making the lag on the right hand side of equation (6.2) dependent on the incubation period. This is denoted by the parameter τ:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-\tau} I_{p,t-\tau} - \gamma I_{p,t-1} + \eta_{p,t}. \tag{6.3}$$

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 likely independent of the incubation period. For instance, Adda (2016) chooses τ equal to one week for acute diarrhea and flu-like illnesses as these have an incubation period of less than a week. Due to the results from Lauer et al. (2020), Q. Li et al. (2020), and Linton et al. (2020), indicating an incubation period for COVID-19 of at most fourteen 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 these regressors by X. This leads to the following formulation:

$$I_{p,t} - I_{p,t-1} = \beta S_{p,t-\tau} I_{p,t-\tau} - \gamma I_{p,t-1} + \delta X_{p,t} + \eta_{p,t}. \tag{6.4}$$

For our application, the data does not span multiple years. Therefore, we do not add year and week effects as they are trends that do not match the SIR model description over a shorter period of time. We also do not include region fixed effects because these act as a region-specific intercept, which was explained in Section 3.3 to be in stark violation of epidemiological models.

There are three other key differences in the model specification by Adda (2016) compared to equation (6.4). First of all, Adda (2016) uses the number of new cases $\Delta I_{p,t-\tau} \coloneqq I_{p,t-\tau} - I_{p,t-\tau-1}$ instead of the total amount of cases $I_{p,t-\tau}$. This is done on both sides of the equation. Second of all, the susceptible rate $s_{p,t}$ is used instead of the total susceptible population $S_{p,t}$. Lastly, Adda (2016) does not include the term $\gamma I_{p,t-1}$ in the model. There are likely two reasons for this. Firstly, the total number of cases are increasing exponentially and are much larger than the number of new cases, by a factor of 1,000 throughout parts of the dataset. Therefore, classical regression methods would fail to identify meaningful recovery parameters. Secondly, the term $\beta s_{p,t-\tau} I_{p,t-\tau-1}$, arising from first differencing $I_{p,t-\tau}$ on the right-hand side, seems to serve as a proxy for the removed cases, which explains the omission of the removed individuals from the right-hand side of the equation. In this section and Section 7.1, we continue with this model to see how it performs in the case of COVID-19.

Now we present the within-region spread model, ignoring effects across regions, as presented by Adda (2016):

$$\Delta I_{p,t} = \beta_{within} s_{p,t-\tau} \Delta I_{p,t-\tau} + \delta X_{p,t} + \eta_{p,t}. \tag{6.5}$$

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 equation (6.5) becomes:

$$\frac{D_{p,t}}{f_{p,t}(\gamma)} - \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)} = \beta_{within} s_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) + \delta X_{p,t} + \eta_{p,t}. \tag{6.6}$$

The models in equations (6.5) and (6.6) are estimated by ordinary least squares (OLS). For the version including undocumented infections, the moment conditions that need to be satisfied due to the strict exogeneity assumption are:

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

We assume that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. The reason why we assume that $E\left[\eta_{p,t} \,\middle|\, s_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \right] = 0$ 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. 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. Because we chose τ to exceed the maximum estimated incubation period, we assume that this holds.

To conclude this section, one could naively consider constructing a model for the entire nation of Italy. Even though this does not take into account regional heterogeneity, as described in Section 2, it may achieve good results if regions are sufficiently similar. This is done in two ways, the first of which is by adding the values of S, I, and R of all regions together to obtain the national numbers, for which the model from equation (6.5) or (6.6), depending on whether undocumented infectives are modelled, is estimated. The results from this model will be labelled with National (OLS). The second method is to apply pooled OLS (POLS). This is a panel data estimation method that ignores the regional heterogeneity, hence treating the data as one large cross-section. These results will be labelled with National (POLS).

6.2 Results

In this section, we present the results for the within-region spread model. When no statistical significance level is mentioned, we take a significance level of 0.05. Firstly, we present the results where the data is pooled to a national level in Table 6.1. Subsequently, results are presented for the models per region. Lastly, we investigate the progression of the transmission rate over time.

Table 6.1. 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	1031.775	763.668	1.351	0.180	536.730	145.941	3.678	0.000***
β_{within}	0.695	8.377×10^{-3}	82.928	0.000***	0.745	0.015	49.497	0.000***

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

Table 6.1 shows estimates for β_{within}, denoted by $\widehat{\beta}_{within}$, of 0.695 and 0.745 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:

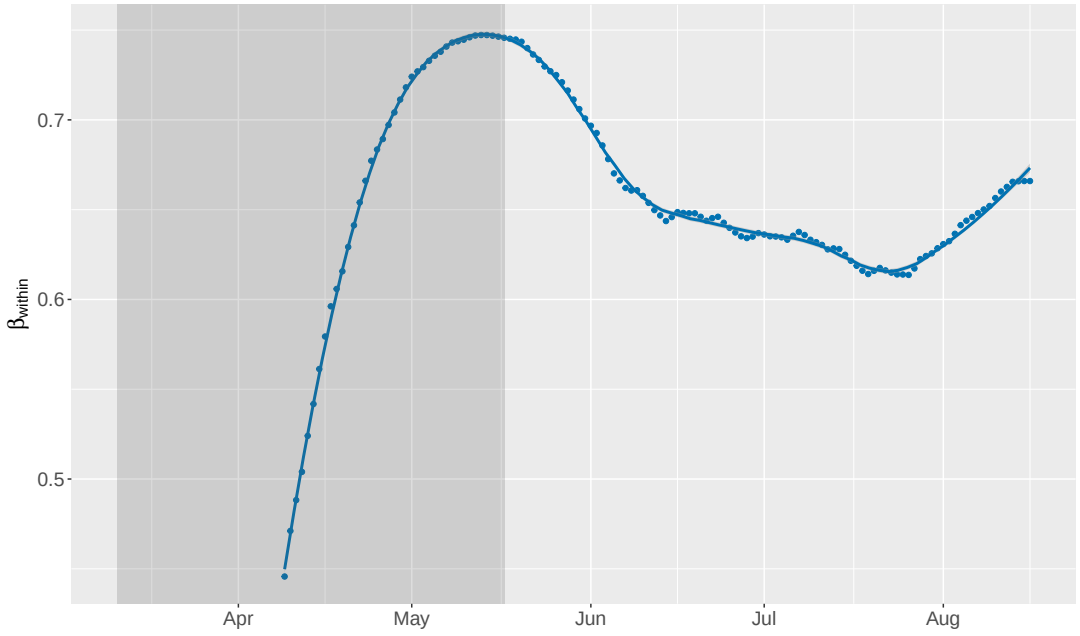
$$H_0 : \ \beta_{within,OLS} - \beta_{within,POLS} = 0.$$

The test statistic is constructed as follows:

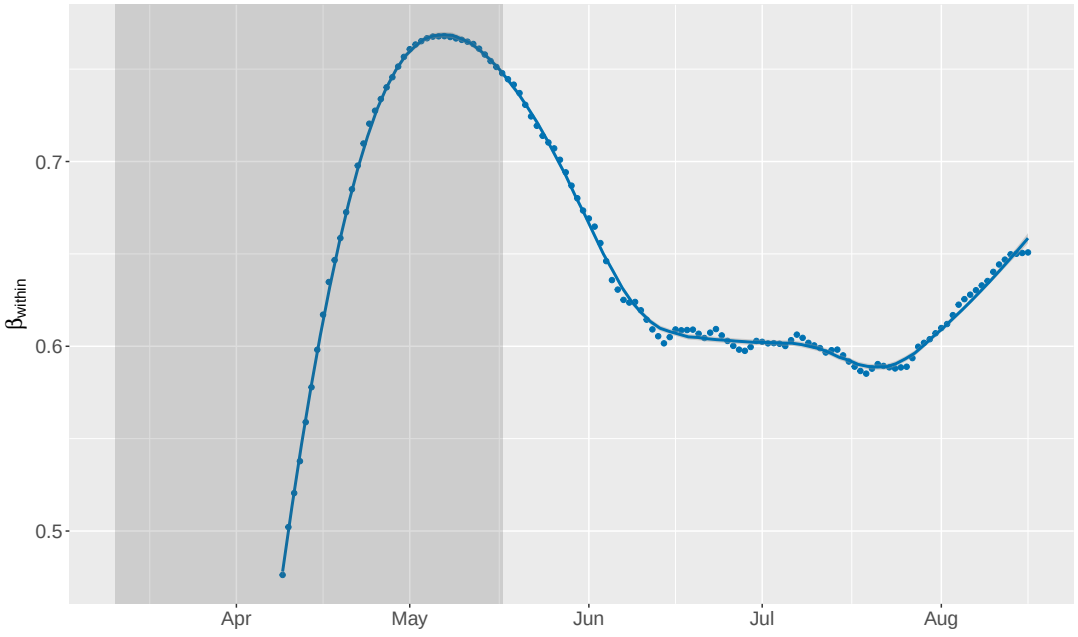
$$t = \left| \frac{\widehat{\beta}_{within,OLS} - \widehat{\beta}_{within,POLS}}{\sqrt{s.e._{within,OLS}^{2} - s.e._{within,POLS}^{2}}} \right|,$$

where $s.e.$ represents the standard error. If we fill out the needed values, we find that $t = 2.910$. Since this exceeds the critical value $t_{T;0.95} = 1.96$, we find statistical evidence in favour of the null hypothesis, at a significance level of 0.05. Therefore, we can conclude that the methods lead to different results.

We are interested in looking at the national transmission rate over time. Because we are able to take into account all regions when using POLS, we need fewer time periods to estimate a model consistently. Therefore, the data used to plot the transmission rate over time in Figure 6.1, uses the thirty latest time observations for each region. In addition, a LOESS (locally estimated scatter plot smoothing) curve with span parameter 0.3 is fit to the data points. We also add a dark grey ribbon around the LOESS curve, representing the 95% confidence interval; the fact that this area is very narrow and almost not visible indicates a high precision.



(a) Infectives exclude undocumented cases



(b) Infectives include undocumented cases

Figure 6.1. Progression of β_{within} over time for the national model (POLS). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

Figure 6.1 shows us that the national transmission rate has been increasing steeply but that it started to slow down during the national lockdown, leading to a decrease from around the end of the lockdown until a sudden rise around the end of July. This sudden increase is likely because the amount of infectives increased a bit over time again for multiple regions, indicating the start of a second wave. Figure 2.1 and the figures in Appendix C.1 indeed illustrate this increase.

As mentioned at the end of the previous section, this national model does not take into account effects specific to regions. In Table 6.2, we present the results for the regional applications. In Appendix B.1 we present Table B.1, including model selection on the weekend dummy with the Akaike Information Criterion (AIC), and Table B.2, comparing the AIC to the Bayesian Information Criterion (BIC) for model selection.

Table 6.2. 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)	0.695***	1031.770	0.626***	3994.670
	(82.928)	(1.351)	(95.942)	(1.146)
National (POLS)	0.745***	536.730***	0.674***	3687.670***
	(49.497)	(3.678)	(39.568)	(4.011)
Abruzzo	0.689***	7.783	0.635***	4.141
	(54.420)	(0.333)	(61.082)	(0.034)
Basilicata	0.526***	10.583**	0.467***	47.536*
	(16.317)	(2.015)	(20.915)	(1.836)
P.A. Bolzano	0.470***	33.685***	0.411***	101.481***
	(34.914)	(3.918)	(47.979)	(4.115)
Calabria	0.576***	16.395	0.543***	71.329
	(27.768)	(1.256)	(31.289)	(0.866)
Emilia-Romagna	0.591***	319.867***	0.527***	1150.100***
	(53.855)	(3.239)	(78.918)	(3.286)
Friuli Venezia Giulia	0.580***	26.799*	0.510***	63.662
	(37.623)	(1.784)	(50.373)	(1.266)
Lazio	0.780***	64.527	0.738***	200.778
	(37.824)	(0.618)	(44.268)	(0.355)
Liguria	0.605***	27.118	0.561***	6.003
	(26.467)	(0.374)	(30.750)	(0.017)
Lombardy	0.785***	18.758	0.712***	460.873
	(115.830)	(0.072)	(132.924)	(0.407)
Marche	0.690***	−6.903	0.646***	−93.780
	(32.563)	(−0.102)	(36.300)	(−0.286)
Molise	0.802***	−0.289	0.766***	−19.838
	(26.102)	(−0.039)	(30.023)	(−0.488)
Piedmont	0.657***	−10.100	0.602***	−219.058
	(44.196)	(−0.048)	(48.251)	(−0.208)
Apulia	0.668***	−6.930	0.642***	−138.834
	(37.115)	(−0.140)	(39.512)	(−0.400)
Sardinia	0.555***	4.890	0.518***	7.216
	(36.721)	(0.535)	(43.088)	(0.126)
Tuscany	0.538***	104.123	0.495***	381.733
	(36.718)	(1.540)	(40.311)	(1.106)
Umbria	0.393***	13.286***	0.337***	49.742***
	(20.123)	(3.786)	(31.442)	(4.247)
Aosta Valley	0.417***	2.316	0.347***	6.776
	(26.810)	(1.029)	(30.647)	(0.779)
Veneto	0.538***	211.985**	0.470***	429.101
	(38.385)	(2.282)	(54.690)	(1.632)

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

For both models, we find statistically significant results for $\widehat{\beta}_{within}$ for all regions, ranging from 0.393 for Umbria to 0.802 for Molise. If we model undocumented infectives, the estimates range from 0.337 for Umbria to 0.766 for Molise. This variation already shows that a national model should not be applied to individual regions, despite the statistical significance of the estimates for the national models. Notice that modelling undocumented infectives seems to impact the order of magnitude between the regions. For instance, where Lombardy has the second-highest estimate for the regular model, it loses that position to Lazio when undocumented infectives are included. Moreover, the estimates for all regions are lower when modelling undocumented infectives. Therefore, it appears that modelling undocumented infectives tends to decrease the estimates of the within-region transmission rate.

We are interested in investigating whether the differences in the estimates differ significantly when modelling undocumented infectives. For this, recall that the standard error of the estimate can be computed by dividing the estimate by the t-statistic. The test statistics for the national models are equal to 6.496 (OLS) and 3.123 (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. Additionally, the estimates differ significantly for ten out of eighteen regions. It is noteworthy that the region of Apulia reports a t-statistic of 1.944, thereby barely falling below the critical value; the estimates are statistically significant different at a significance level of 5.19%.

When we want to interpret the estimates of β_{within}, we should recall that Adda (2016) states 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, because the removal term is omitted from the model, as explained in Section 6.1, and therefore does not take into account the specific removal rate of COVID-19, we cannot interpret the coefficients in the same way. Nevertheless, we can compare the magnitude of the coefficients with one another. For instance, consider the region of Lombardy and the island of Sardinia. We consider the model where undocumented infectives are modelled, for which we find values for $\widehat{\beta}_{within}$ of 0.712 (Lombardy) and 0.518 (Sardinia). Although this cannot tell us much about the spread within the region explicitly, it shows us that the transmission in Lombardy was more severe than on Sardinia. A similar interpretation can be applied to any comparison of regions and for the model without modelling undocumented infectives.

In Table 6.2, we modelled undocumented infectives using the quadratic specification with $\gamma = 0.7$. Of course, we can also apply other specifications as defined in Section 5. Table 6.3 shows the results of the within-region spread model for four different specifications for the national models and three regions. For all specifications, we set $f^{min} = 0.1$.

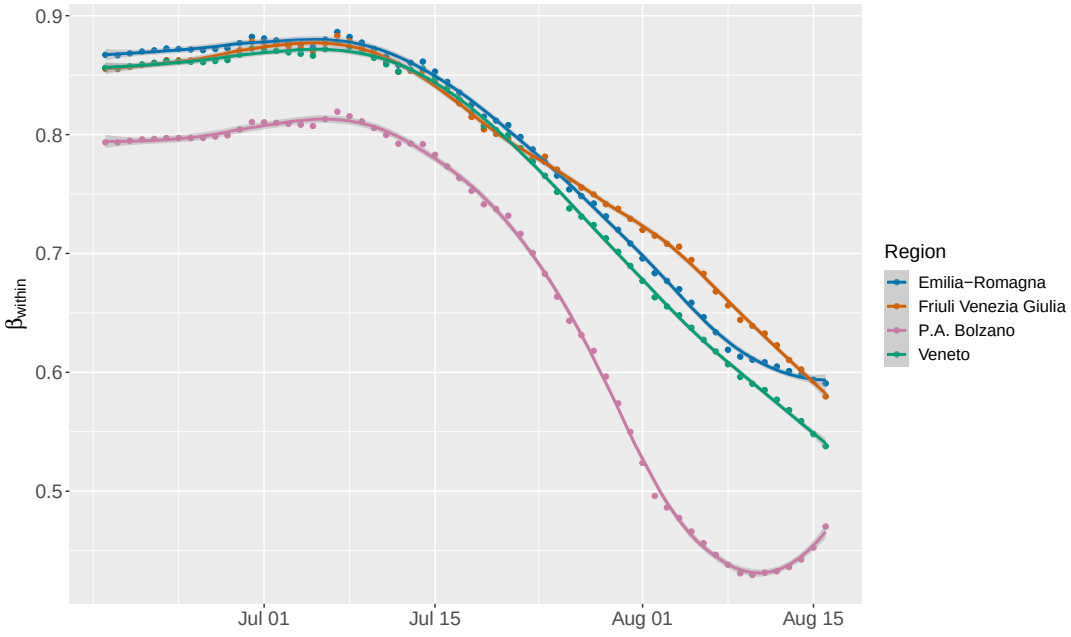
Table 6.3. Estimates for the within-region transmission rate comparing functional forms for modelling undocumented infectives ($f^{min} = 0.1$). Estimates are given with t-statistics in parentheses. Data spans May 9 until August 16, 2020 (100 days).

Region	Linear	Quadratic ($\gamma = 0.6$)	Quadratic ($\gamma = 0.7$)	Cubic ($\gamma_1 = 0.6,\ \gamma_2 = 0.8$)
National (OLS)	0.649***	0.636***	0.626***	0.611***
	(89.995)	(93.264)	(95.942)	(101.179)
National (POLS)	0.695***	0.687***	0.674***	0.650***
	(41.661)	(40.860)	(39.568)	(37.464)
Calabria	0.554***	0.550***	0.543***	0.527***
	(30.196)	(30.584)	(31.289)	(32.987)
Lombardy	0.735***	0.727***	0.687***	0.669***
	(124.856)	(127.818)	(132.924)	(144.560)
Veneto	0.483***	0.478***	0.470***	0.461***
	(52.329)	(53.215)	(54.690)	(55.445)

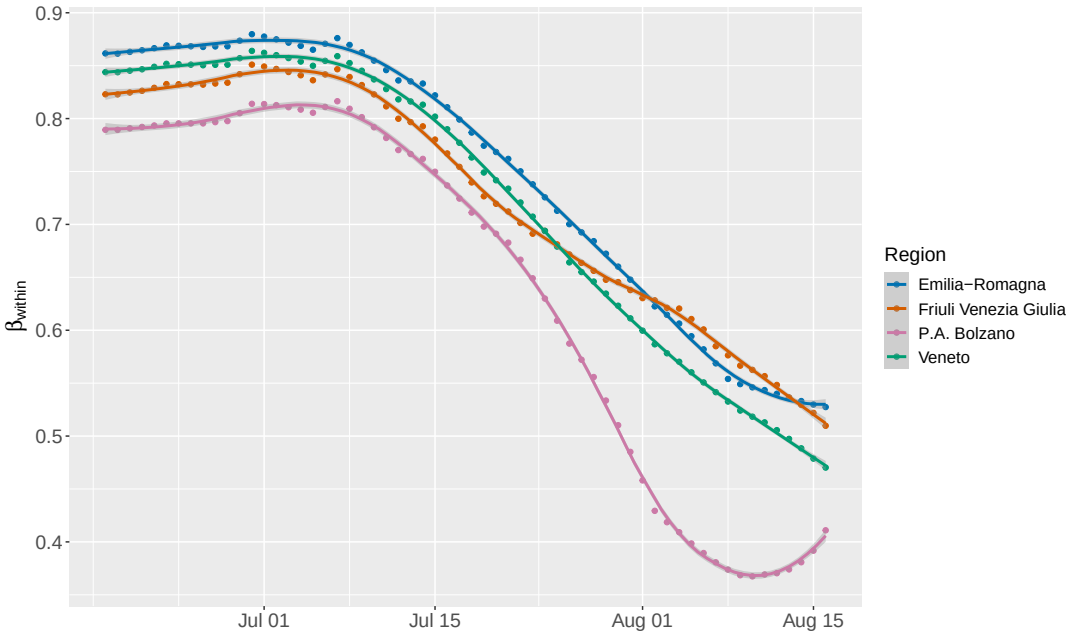
Significance levels: * = 0.1 ** = 0.05, *** = 0.01

Although the estimates in Table 6.3 do not differ much for different specifications, the difference for some specifications may be statistically significant. Consider the linear and quadratic specifications for Lombardy. Comparing the linear and quadratic ($\gamma = 0.6$) specifications, we find a t-statistic of 0.977, resulting in a non-significant difference. The difference between the estimates using the linear and quadratic ($\gamma = 0.7$) specifications, however, does yield a significant difference, with $t = 6.127$. Therefore, there is indeed an impact of the specific functional form and one should choose the one that they deem to be fitting.

We are mainly 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 6.2, 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 6.2 is the estimate of β_{within} 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) Infectives exclude undocumented cases

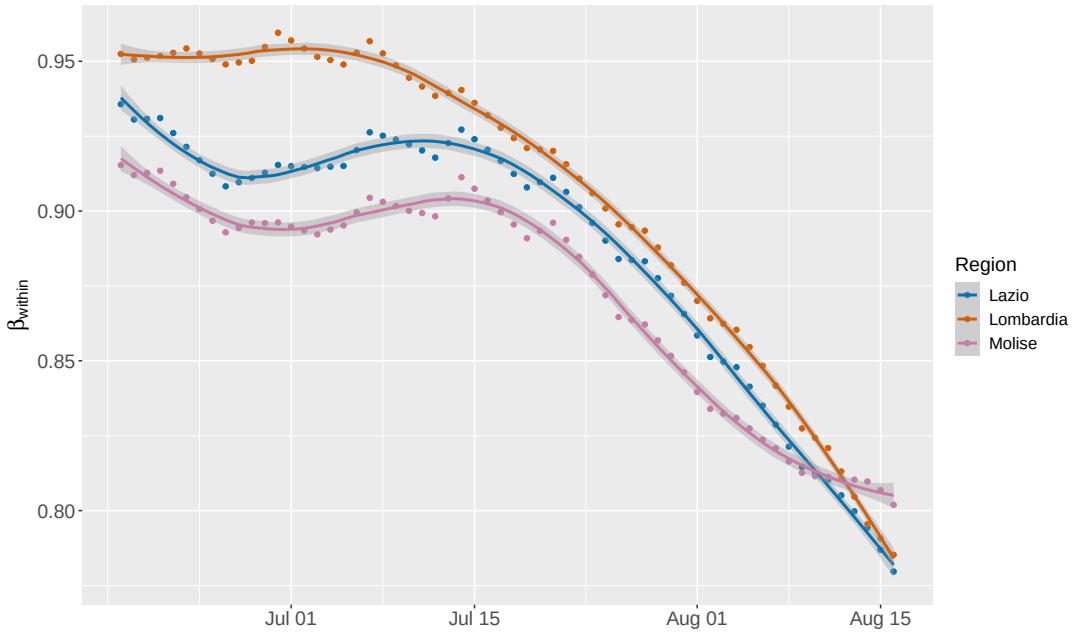


(b) Infectives include undocumented cases

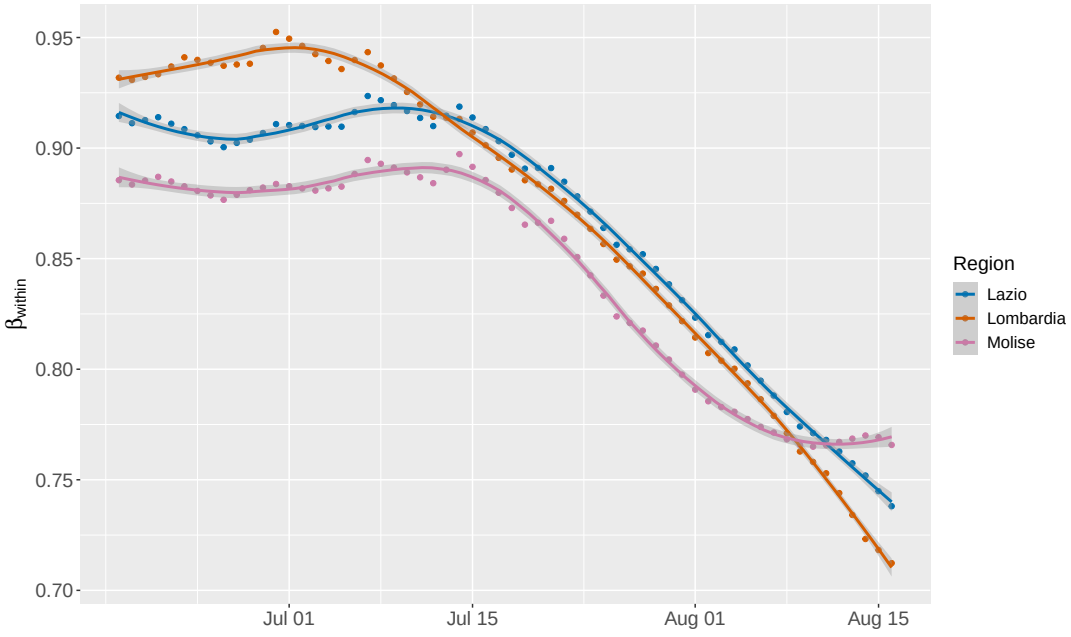
Figure 6.2. Progression of $\widehat{\beta}_{within}$ over time for the Nord-Est NUTS 1 region. The last 100 days are used for estimation. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Considering the progression of $\widehat{\beta}_{within}$ over time, we indeed see that it decreases over time, as we expected. We do see a slight increase or slanting in some estimates of β_{within} towards the end of the timespan. This is likely because of the arising second wave, as discussed earlier in this section. We also notice that the profile of $\widehat{\beta}_{within}$ over time is similar when comparing the models excluding and including undocumented infectives; the difference can be found in the level of $\widehat{\beta}_{within}$ which, as explained earlier in this section, tends to be lower when modelling undocumented infectives.

Lastly, we are interested in comparing the transmission rates over time across the three regions with the highest transmission rate in Table 6.2. The reason for this is that we expected the estimated transmission rate for Lombardy to have been the highest but we saw a higher value for Molise. However, this does not mean that Molise has always had the highest estimated transmission rate. Moreover, the number of cases in Molise is actually not that high; on September 24, it only saw a total number of 623 cases while Lazio and Lombardy saw 14,975 and 105,226 total infectives, respectively. In Figure 6.3, we compare Lazio, Lombardy, and Molise.



(a) Infectives exclude undocumented cases

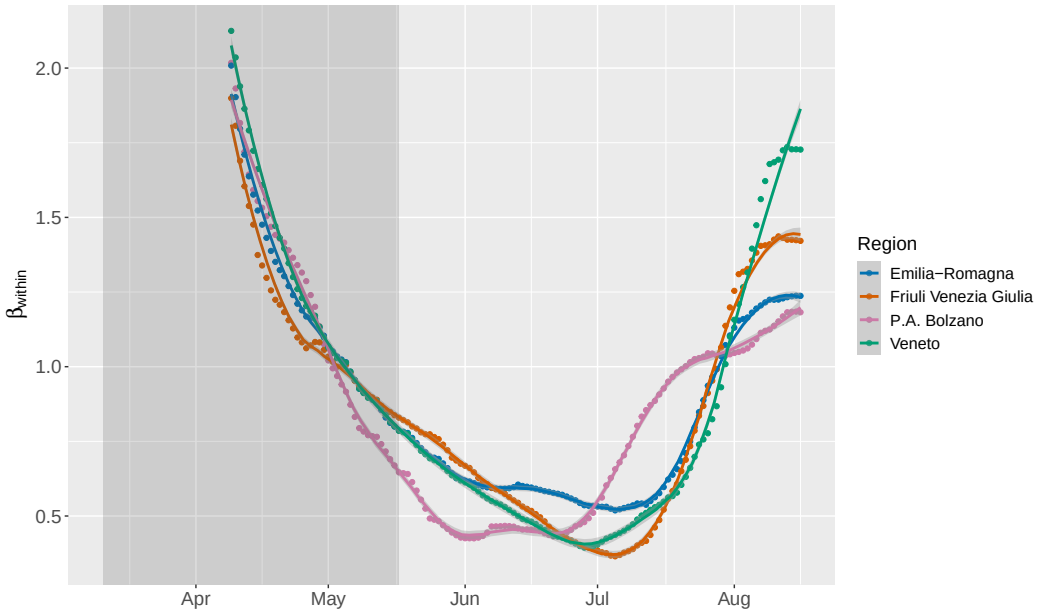


(b) Infectives include undocumented cases

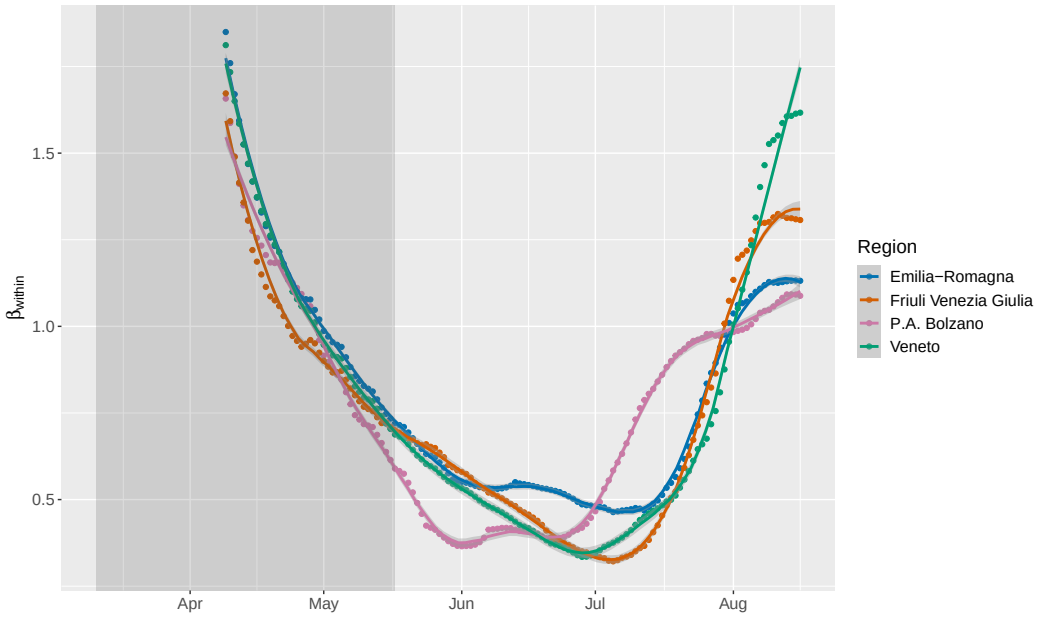
Figure 6.3. Progression of $\widehat{\beta}_{within}$ over time for Lazio, Lombardy, and Molise. The last 100 days are used for estimation. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Firstly, notice that there is a difference in the ranking over time depending on whether undocumented infectives are modelled or not; if they are, then Lombardy is passed by Lazio in the middle of July, although Lazio never exceeds Lombardy for the regular model. Secondly, focusing on the regular model, we indeed see that Lombardy reported the highest transmission rate until a steep decrease was seen and Molise passed it in the middle of August. This is likely because Lombardy did not see an increase indicating a second wave, as can be seen in Figure C.1, while Molise did, as can be seen in Figure C.3.

To conclude the discussion of the within-region spread model, we investigate the progression of the estimated transmission rate over time when using a smaller sample size for our models, so that we can look at the events during the national lockdown. Because we use a smaller sample size, the Type II error probability of our estimation decreases, meaning that we can expect wider confidence intervals. In Figure 6.4, we plot this progression for the Nord-Est NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.2.



(a) Infectives exclude undocumented cases



(b) Infectives include undocumented cases

Figure 6.4. Progression of $\widehat{\beta}_{within}$ over time for the Nord-Est NUTS 1 region. The last 30 days are used for estimation. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

Figure 6.4 shows, as expected, quite different results compared to Figure 6.2. Considering the period during the national lockdown, indicated by the grey area, we see that the estimated transmission rates for all regions has decreased substantially, suggesting that the lockdown has been effective in driving the transmission of the virus back. After the lockdown ended on May 10, the degree of the decline slanted and towards the beginning of July we saw an increase in the transmission rate again.

7 Within and Between-Region Spread Model

In this section, we present the model by Adda (2016) that takes effects across regions into account. Section 7.1 discusses the methodology, including the model formulation, estimation method, moment conditions, and the inclusion of undocumented infectives. 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 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:

$$\Delta I_{p,t} = \beta_{within} s_{p,t-\tau} \Delta I_{p,t-\tau} + \beta_{between} s_{p,t-\tau} \sum_{c \in R \setminus r} \Delta I_{c,t-\tau} + \delta X_{p,t} + \eta_{p,t}. \tag{7.1}$$

It should be noted that the specification in equation (7.1) 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, inhabitants of Lombardy, which lies in north-west Italy, 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 equation (7.1) and explain the other criteria as possible future research in Section 10.

In equation (7.1), the transmission parameter β is now allowed to be different within and between regions. Adda (2016) estimates equation (7.1) 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 adverse weather conditions impact the amount of social interaction between people. It is unlikely that these are valid instruments for 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 December 2019 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.

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 equation (7.1) becomes:

$$\begin{aligned}\frac{D_{p,t}}{f_{p,t}(\gamma)} - \frac{D_{p,t-1}}{f_{p,t-1}(\gamma)} =& \beta_{within} s_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \\ &+ \beta_{between} s_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right) \\ &+ \delta X_{p,t} + \eta_{p,t}.\end{aligned} \tag{7.2}$$

The moment conditions that then need to hold are:

$$\begin{aligned}E\Bigg[\eta_{p,t} \Bigg(& \beta_{within} s_{p,t-\tau} \left(\frac{D_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{D_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \\ & + \beta_{between} s_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{D_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{D_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right) \\ & + \delta X_{p,t} \Bigg) \Bigg] = 0.\end{aligned} \tag{7.3}$$

We assume that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in $X_{p,t}$. In Section 6.1 we explained why we assume that $\eta_{p,t}$ is uncorrelated with the within-region spread term. We assume that the between-region spread term is uncorrelated with $\eta_{p,t}$ for the same reasons: for a large enough lag τ, the error is not correlated with past data at that lag, which is independent of $f_{p,t-\tau}(\gamma)$ and $f_{p,t-\tau-1}(\gamma)$. We also state 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.

7.2 Results

In this section, we present the results for the within and between-region spread model. When no statistical significance level is mentioned, we take a significance level of 0.05.

In Table 7.1, we present the results for the regional applications. In Appendix B.2 we present Table B.3, including model selection on the weekend dummy with the AIC, and Table B.4, comparing the AIC to the BIC for model selection.

Table 7.1. 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	0.123	0.012***	−18.628	−0.130	0.017***	−128.095
	(0.938)	(4.333)	(−0.834)	(−0.944)	(5.545)	(−1.169)
Basilicata	0.736***	-4.135×10^{-4}**	14.312***	0.691***	-5.265×10^{-4}***	70.108***
	(8.458)	(−2.584)	(2.698)	(9.708)	(−3.320)	(2.743)
P.A. Bolzano	0.421***	3.809×10^{-4}	31.133***	0.364***	2.614×10^{-4}*	92.402***
	(10.940)	(1.375)	(3.555)	(14.000)	(1.809)	(3.691)
Calabria	0.753***	-1.313×10^{-3}**	26.157*	0.756***	-1.996×10^{-3}***	132.166
	(10.017)	(−2.450)	(1.960)	(9.088)	(−2.629)	(1.589)
Emilia-Romagna	0.604***	-1.534×10^{-3}	324.983***	0.475***	5.215×10^{-3}	1075.270***
	(11.560)	(−0.264)	(3.214)	(14.276)	(1.404)	(3.002)
Friuli Venezia Giulia	0.763***	-2.095×10^{-3}***	41.340***	0.705***	-1.947×10^{-3}***	139.824***
	(11.896)	(−2.942)	(2.704)	(16.205)	(−4.640)	(2.893)
Lazio	0.106	0.040***	130.051*	0.229***	0.035***	782.017*
	(1.583)	(10.272)	(1.783)	(4.467)	(10.129)	(1.960)
Liguria	0.589***	6.335×10^{-4}	22.016	0.604***	-2.008×10^{-4}	84.998
	(6.381)	(0.184)	(0.282)	(6.684)	(−0.556)	(0.222)
Lombardy	0.642***	0.105***	364.667**	0.591***	0.074***	2067.970**
	(45.526)	(10.773)	(2.027)	(34.369)	(6.588)	(2.190)
Marche	0.052	0.024***	−112.390*	−0.308**	0.035***	−651.703**
	(0.414)	(5.155)	(−1.774)	(−2.578)	(7.987)	(−2.468)
Molise	−0.087*	2.455×10^{-3}***	−5.585*	−0.046	2.506×10^{-3}***	−25.810
	(−1.837)	(19.605)	(−1.671)	(−1.287)	(23.760)	(−1.651)
Piedmont	0.671***	-2.548×10^{-3}	1.848	0.643***	-9.928×10^{-3}	33.469
	(6.923)	(−0.138)	(8.018×10^{-3})	(6.393)	(−0.507)	(0.030)
Apulia	0.599***	2.240×10^{-3}	−21.609	0.185*	0.019***	−596.977*
	(6.024)	(0.704)	(−0.400)	(1.759)	(4.355)	(−1.783)
Sardinia	0.731***	-1.276×10^{-3}***	17.678*	0.757***	-2.273×10^{-3}***	105.529*
	(14.796)	(−3.726)	(1.909)	(15.340)	(−4.996)	(1.930)
Tuscany	0.654***	-6.650×10^{-3}**	159.062**	0.650***	-9.361×10^{-3}***	746.518**
	(12.899)	(−2.386)	(2.274)	(12.999)	(−3.268)	(2.163)
Umbria	0.315***	1.758×10^{-4}*	11.181***	0.275***	1.398×10^{-4}**	40.840***
	(6.746)	(1.835)	(3.062)	(10.138)	(2.425)	(3.399)

Table 7.1 continues on next page

Table 7.1 continued from previous page

	Regular model			Modelling undocumented infectives		
Region	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Aosta Valley	0.356***	1.210×10^{-4}**	0.486	0.281***	1.115×10^{-4}***	−2.003
	(12.707)	(2.580)	(0.211)	(14.380)	(3.755)	(−0.235)
Veneto	0.664***	−0.011***	294.301***	0.604***	-8.899×10^{-3}***	819.111***
	(14.282)	(−2.832)	(3.120)	(21.351)	(−5.019)	(3.331)

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

Firstly, note that the estimates for $\beta_{between}$ are generally much smaller than the estimates for β_{within}. This is likely the case because the models are defined using the absolute number of new cases instead of the new cases per capita. As such, summing over all regions leads to a large number of total new cases, causing the parameter estimates to be driven down. A notable exception being the region of Lombardy, where $\widehat{\beta}_{between} = 0.105$ is much larger than for most other regions. The reason for this may be because Lombardy was arguably hit the hardest by SARS-CoV-2 of all of the Italian regions; 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 $\beta_{between}$. This is not logical because this would imply that the interaction of infectives and susceptible people leads to a reduction in the transmission. Unfortunately, these estimates are not necessarily statistically insignificant. Future research can be done into models that restrict the estimates to be positive. For the within and between-region spread model, we find varying results of statistical significance. Fourteen (which are all positive) and eleven (of which six are negative) out of eighteen regions find a significant value for $\widehat{\beta}_{within}$ and $\widehat{\beta}_{between}$, respectively, when excluding undocumented infectives. For the model including undocumented infectives, fifteen (of which one is negative) and fourteen (of which six are negative) regions find a significant result for $\widehat{\beta}_{within}$ and $\widehat{\beta}_{between}$. It is interesting to note that a negative value of $\widehat{\beta}_{between}$ seems to go hand-in-hand with a relatively high value of $\widehat{\beta}_{within}$.

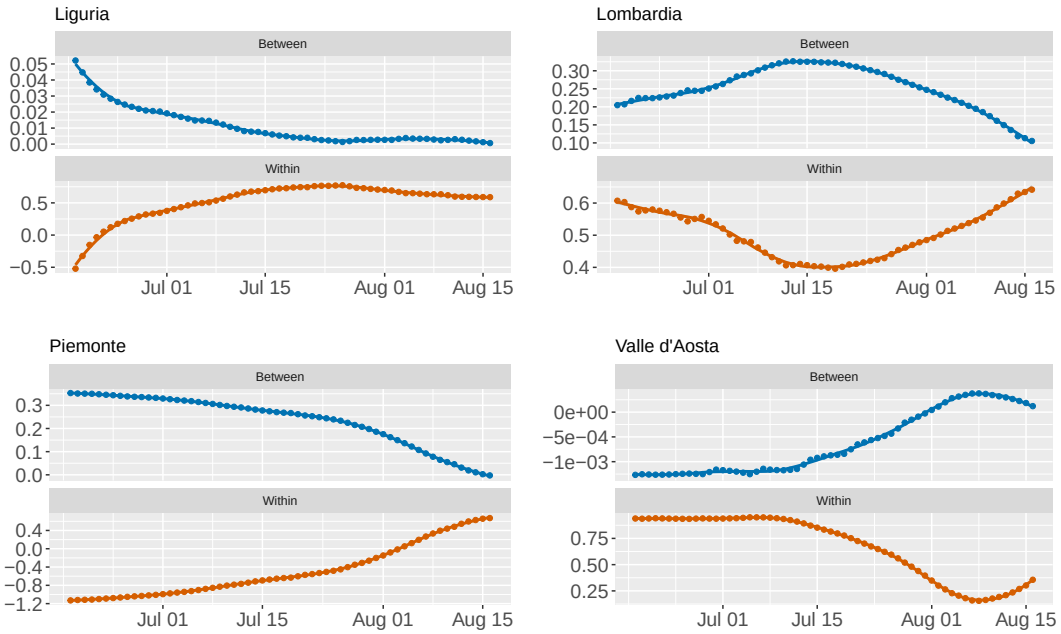
Looking only at the statistically significant and positive estimates of β_{within}, these range from 0.315 for Umbria to 0.763 for Friuli Venezia Giulia for the regular model and from 0.229 for Lazio to 0.757 for Sardinia when undocumented infectives are included. Again, modelling undocumented infectives does not retain the ordering of regions. When considering the statistically significant and positive estimates of $\beta_{between}$, we see that these range from 1.210×10^{-4} for Aosta Valley to 0.105 for Lombardy for the regular model and from 1.115×10^{-4} for Aosta Valley to 0.074 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}.

We are again interested in testing whether the estimates when including undocumented infectives differ significantly from those for the regular model. We only consider the regions for which both models produce a statistically significant estimate. For $\widehat{\beta}_{within}$, a significant difference is only found for Emilia-Romagna ($t = 2.083$), Lombardy ($t = 2.293$), and Aosta Valley ($t = 2.196$). For $\widehat{\beta}_{between}$, only the estimate for Lombardy differs significantly ($t = 2.085$).

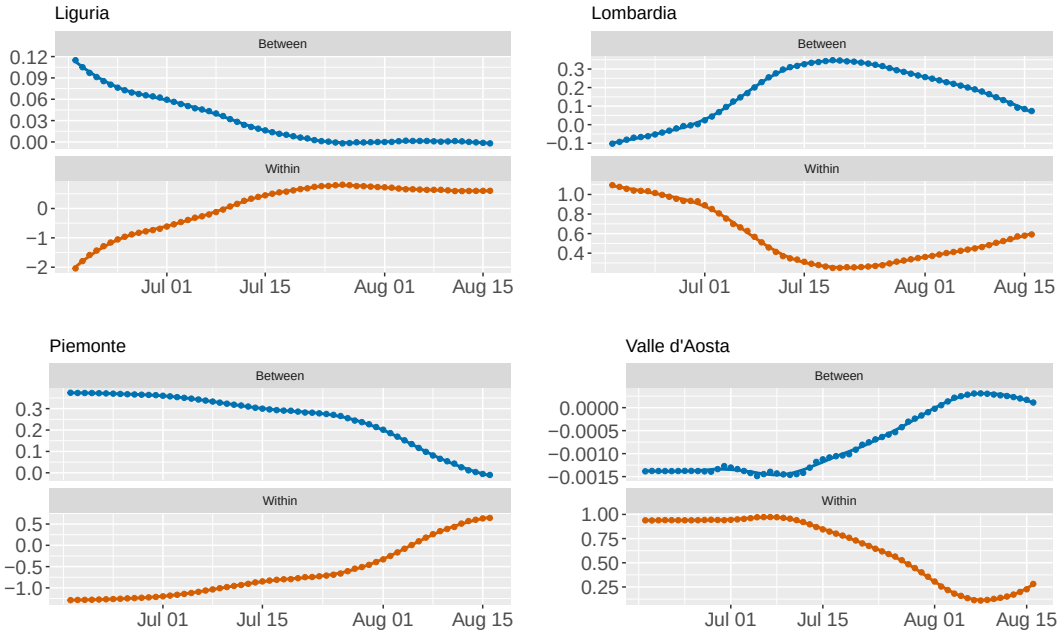
One aspect to pay attention to is regarding the impact of modelling undocumented infectives on the estimates. In the results for the within-region spread model, we saw that the estimated transmission rates were all lower when modelling undocumented infectives. For the within and between-region spread model, this is not the case anymore for Calabria, Liguria, and Sardinia. For the significant estimates of $\beta_{between}$, only three out of six regions find a lower estimate when modelling undocumented infectives.

Recall that we cannot interpret the coefficients in the same way as Adda (2016) does but we can compare the magnitude of the coefficients with one another. For instance, we find significantly differing estimates for β_{within} of 0.229 and 0.591 and for $\beta_{between}$ of 0.035 and 0.074 for Lazio and Lombardy, respectively. We can conclude that the transmission within the region as well as between regions was worse in Lombardy compared to Lazio, although we cannot explicitly interpret the magnitude of that transmission.

To conclude, we are again interested in looking at the estimates of β_{within} and $\beta_{between}$ over time. In Figure 7.1, we present plots for the regions in the Nord-Ovest NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.3, which generally show similar results. Each point in the graphs in Figure 7.1 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) Infectives exclude undocumented cases



(b) Infectives include undocumented cases

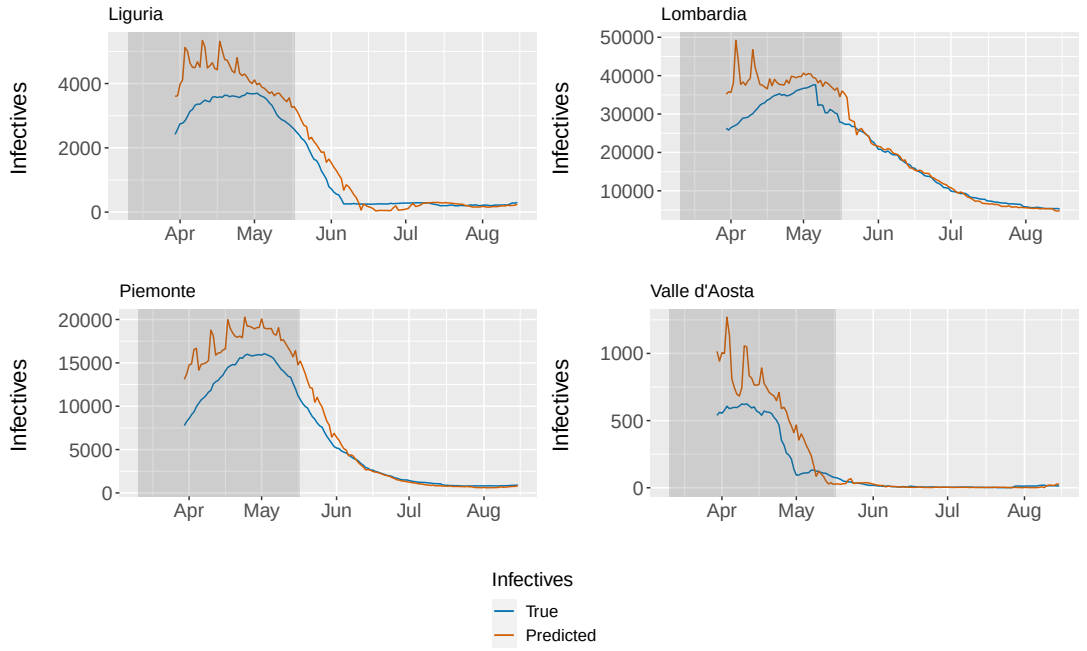
Figure 7.1. Progression of β_{within} and $\beta_{between}$ over time for the Nord-Ovest NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Figure 7.1, most notably, shows profiles for $\widehat{\beta}_{within}$ and $\widehat{\beta}_{between}$ that seem to be the complete opposite of one another. This does not necessarily seem logical: if the transmission between regions increases, it does not logically imply that the transmission within the region decreases. This property is seen across all regions, although the shapes of the profiles differ vastly across regions. We also see many negative estimates. Lastly, we also do not see the same movement for an estimate, e.g. β_{within} decreasing or $\beta_{between}$ increasing over time for all regions consistently. In Figure 7.1, we see, for instance, an increasing profile for β_{within} for Liguria and Piedmont but a mostly decreasing profile for Lombardy and Aosta Valley.

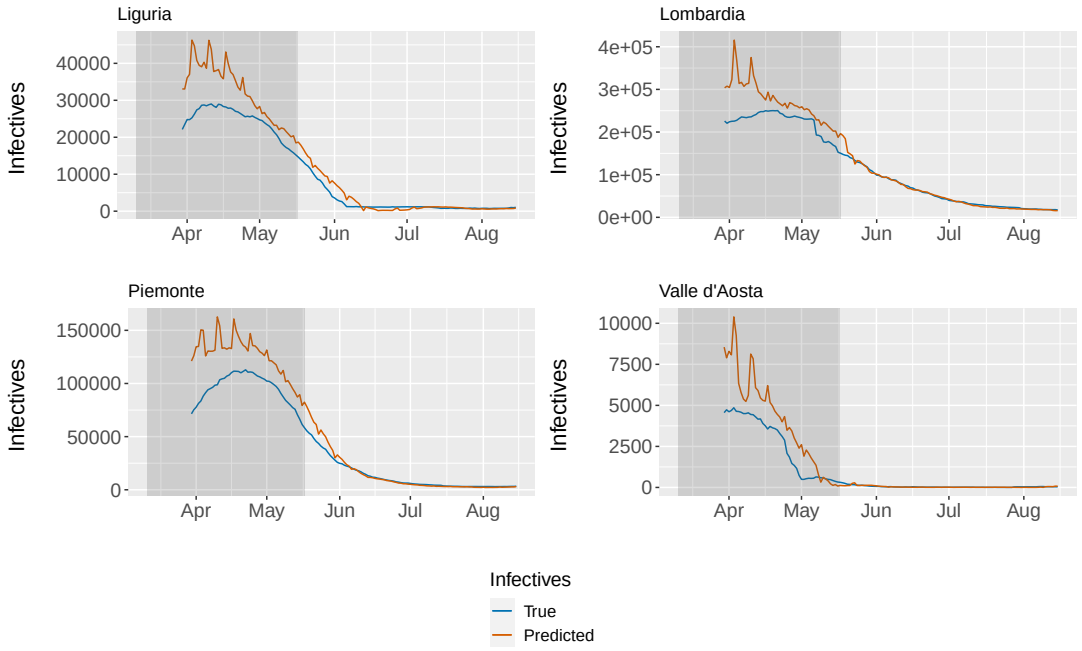
8 Forecasts

In this section, we will provide forecasts for the models to investigate the applicability of our models in addition to the estimation of the transmission rates. Firstly, we will consider one-period ahead forecasts, after which we will consider multi-period ahead forecasts.

Consider Figure 8.1, where we plot one-period ahead forecasts for the Nord-Ovest NUTS 1 region for the within-region spread model. Plots for the other NUTS 1 regions can be found in Appendix C.4.



(a) Infectives and forecasts exclude undocumented cases

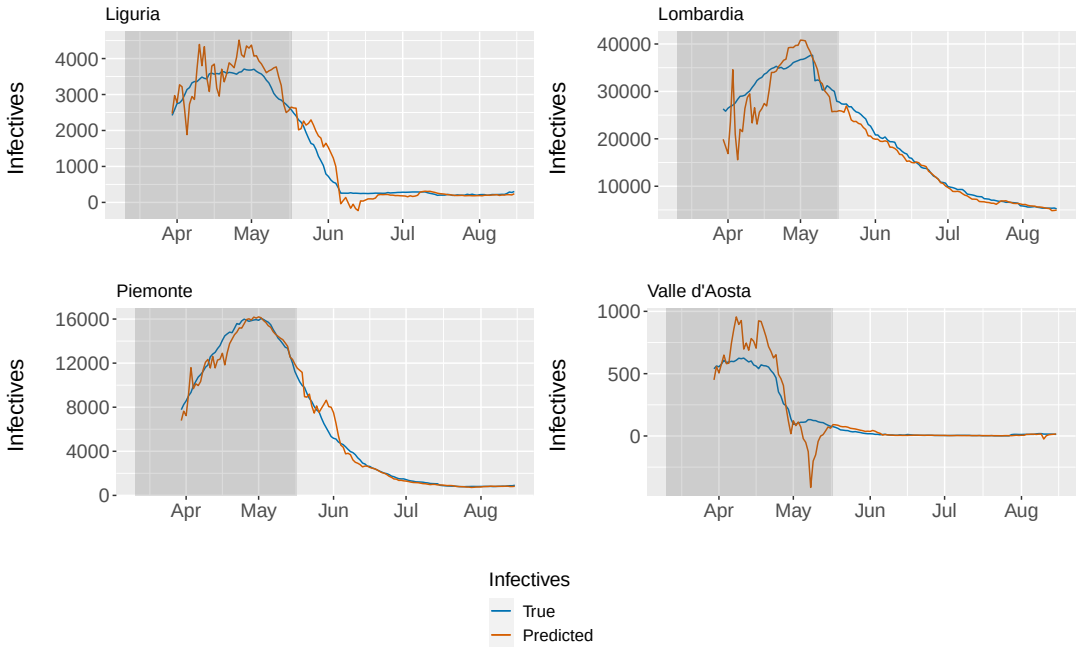


(b) Infectives and forecasts include undocumented cases

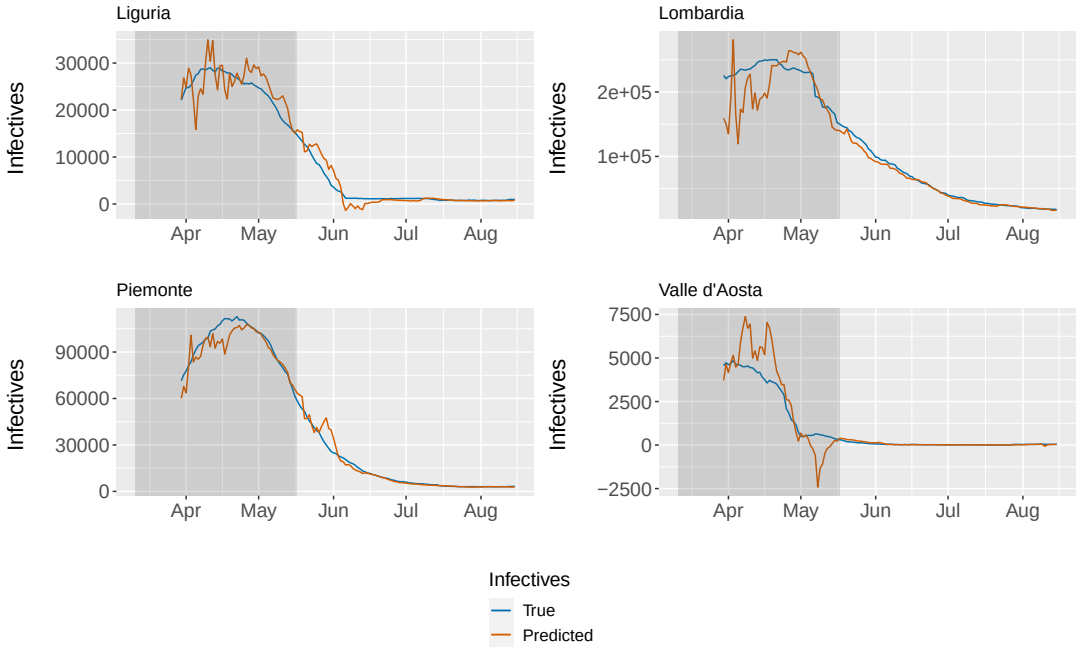
Figure 8.1. One-period ahead forecasts for the within-region spread model for the Nord-Ovest NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Figure 8.1 shows that our one-period ahead forecasts are able to emulate the profile of the infectives but that almost all predictions exceed the true number of infectives. The gap between the true and forecasted values decreases over time, as the number of (documented) infectives decreases.

Consider Figure 8.2, where we plot the forecasts for the Nord-Ovest NUTS 1 region for the within and between-region spread model instead.



(a) Infectives and forecasts exclude undocumented cases



(b) Infectives and forecasts include undocumented cases

Figure 8.2. One-period ahead forecasts for the within and between-region spread model for the Nord-Ovest NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

In Figure 8.2, we see that the predicted values are generally closer to the true values compared to the within-region spread model. There are two important notes to make. Note that negative estimated values of the number of infectives occur for Liguria and Aosta Valley, among other regions, which is not possible. This is likely the case due to a steeper decrease in the number of infectives $\tau = 14$ days prior.

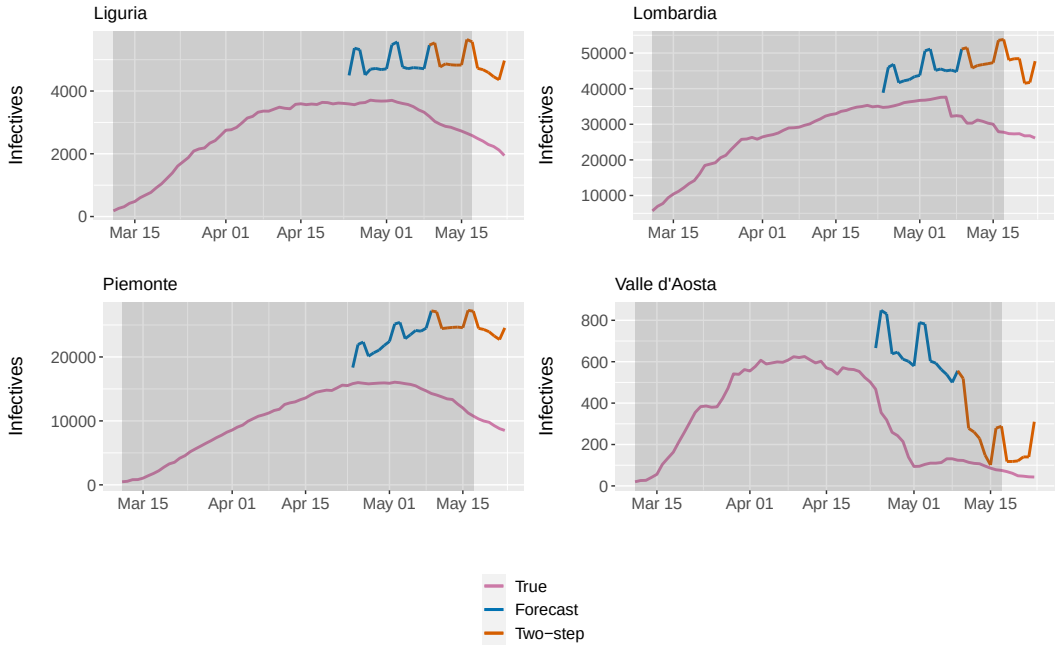
We now consider forecasts for more than one period ahead for the within-region spread model. Firstly, we will discuss a method by which we can forecast up to τ periods ahead, after which we talk about an iterative method to forecast more than τ periods ahead. Because the right-hand side of our model formulation in equation (6.5) depends on data with a time lag of τ and that only a weekend effect is included in the regressors, which is readily available, we can directly forecast τ periods ahead. When we have estimated our model with data up to period t, we obtain estimates $\widehat{\beta}_{within}$ and $\widehat{\delta}$ that can be used to estimate the future number of infectives:

$$\widehat{\Delta I_{p,t}} = \widehat{\beta}_{within} s_{p,t-\tau} \Delta I_{p,t-\tau} + \widehat{\delta} X_{p,t}.$$

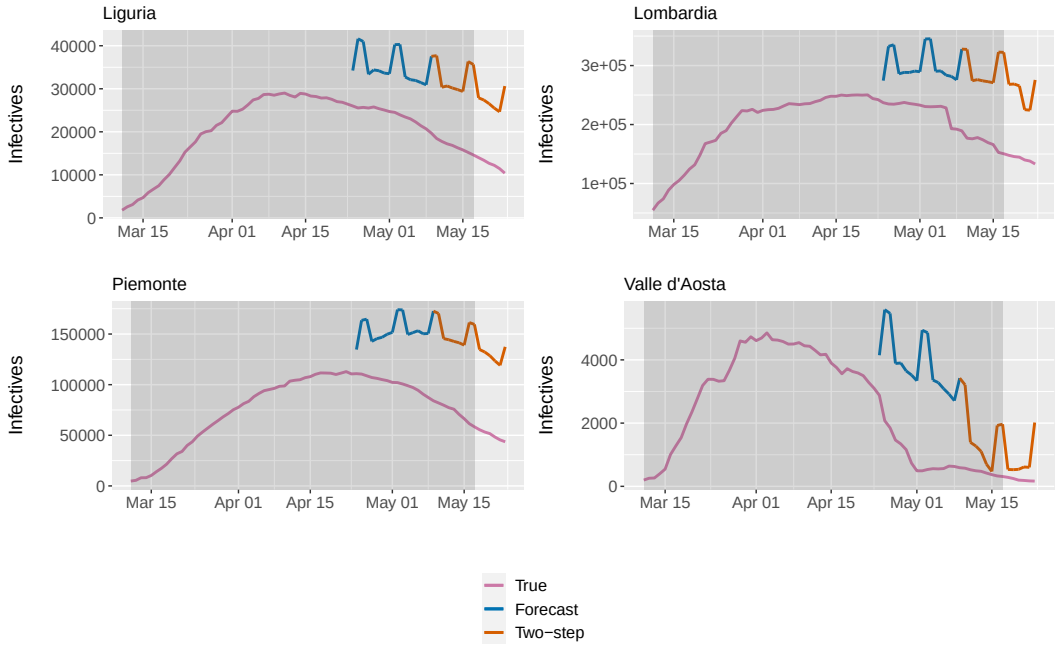
The results from applying this method are labelled with *Forecast* in the plots below. We can also iteratively forecast more than τ periods ahead by using the estimated number of infectives as the true data and assuming that the susceptible population changed only by the number of removed infectives, i.e. deaths and recoveries are ignored:

$$S_{p,t+1} = S_{p,t} - \widehat{\Delta I_{p,t}}.$$

Then, we can again use the estimated values $\widehat{\beta}_{within}$ and $\widehat{\delta}$ to compute the number of infectives. The results from applying this method are labelled with *Two-step* in the plots below. In Figure 8.3, we present the results for the Nord-Ovest NUTS 1 region for the within-region spread model. Plots for the other NUTS 1 regions can be found in Appendix C.4.



(a) Infectives and forecasts exclude undocumented cases



(b) Infectives and forecasts include undocumented cases

Figure 8.3. Iterated forecasts for the within-region spread model for the Nord-Ovest NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

Figure 8.3 shows that the multi-period ahead forecasts are grossly overestimating the number of infectives. This is because the estimates were retreived using data when the number of infectives was increasing steeply, although it has since been decreasing. We conclude that this method of forecasting is not to be preferred.

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 31 million people worldwide and caused almost a million deaths so far. The goal of this thesis was to apply models which recognize that the Italian regions are not homogeneous such that 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 presented by Adda (2016). When not taking spatial spillover into account, it became clear that the regions of Lazio, Lombardy, and Molise sported the highest within-region transmission rate. However, Molise did not see a high number of cases. The results that do consider a possible spatial spillover indicated that the within-region transmission rate for Lazio and Molise is not significant but that both have a significant between-region transmission rate. For Lombardy, a significant within-region and between-region transmission rate was found. For the other regions, the between-region transmission was generally shown to have been less strong. Lastly, the models can be used to forecast the number of infectives. A multi-period ahead forecast extrapolating data from during the national lockdown to forecast infections after the lockdown is not suitable.

Conclusively, this thesis has exposed 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 actions to tackle the pandemic. There are consistent regional effects that models should take into account, The results from those models also differ vastly across regions. Consequently, regional effects should be kept in mind 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 suitable data is available. As highlighted in Section 7, 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 more accurate results.

We also made the assumption that individuals that successfully clear COVID-19 gain immunity for 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 national lockdown. 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, with the intention to also make conclusions about the effective reproduction number R_{eff}. In Appendix D, we present a short explanation on the methodology that was tried and some results. 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, both in the data and due to the reporting process. The underlying data is often nonstationary because of a slow start, after which there is an exponential increase in the number of cases. If one is able to take the nonstationarity into account, this may be a promising field of research.

Lastly, this thesis did not discuss the full weighted model presented by Adda (2016) due to a lack of data. As more 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.

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/

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.

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/

Horowitz, J. (2020). *Italy's health care system groans under coronavirus — a warning to the world*. Retrieved June 11, 2020, from https://www.nytimes.com/2020/03/12/world/europe/12italy-coronavirus-health-care.html

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

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

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/

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

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

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

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

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

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

Royal House of the Netherlands. (2020). *Speech by His Majesty the King in light of the coronavirus*. Retrieved September 22, 2020, from https://www.royalhouse.nl/documents/speeches/2020/03/20/speech-by-his-majesty-the-king-in-light-of-the-coronavirus

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

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

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, 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 or 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 6.2
BIC	Bayesian Information Criterion	Section 6.2

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 6.1
POLS	Pooled Ordinary Least Squares	Section 7.1

B Tables

B.1 Results for the Within-Region Spread Model

In Section 6.2, 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 (6.5) 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 with model selection by AIC. Estimates are given with t-statistics in parentheses. Data spans 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)	0.700***		0.629***	
	(92.704)		(105.586)	
National (POLS)	0.745***	536.730***	0.674***	3687.670***
	(49.497)	(3.678)	(39.568)	(4.011)
Abruzzo	0.691***		0.635***	
	(60.481)		(66.997)	
Basilicata	0.526***	10.583**	0.467***	47.536*
	(16.317)	(2.015)	(20.915)	(1.836)
P.A. Bolzano	0.470***	33.685***	0.411***	101.481***
	(34.914)	(3.918)	(47.979)	(4.115)
Calabria	0.586***		0.549***	
	(30.517)		(34.049)	
Emilia-Romagna	0.591***	319.867***	0.527***	1150.100***
	(53.855)	(3.239)	(78.918)	(3.286)
Friuli Venezia Giulia	0.580***	26.799*	0.514***	
	(37.623)	(1.784)	(54.200)	
Lazio	0.786***		0.741***	
	(43.142)		(49.820)	
Liguria	0.608***		0.561***	
	(28.836)		(33.085)	
Lombardy	0.785***		0.713***	
	(132.494)		(149.266)	

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	0.689***		0.644***	
	(35.744)		(39.446)	
Molise	0.801***		0.761***	
	(29.160)		(32.996)	
Piedmont	0.657***		0.601***	
	(48.479)		(52.338)	
Apulia	0.667***		0.639***	
	(40.380)		(42.668)	
Sardinia	0.558***		0.519***	
	(39.712)		(46.239)	
Tuscany	0.538***	104.123	0.499***	
	(36.718)	(1.540)	(43.521)	
Umbria	0.393***	13.286***	0.337***	49.742***
	(20.123)	(3.786)	(31.442)	(4.247)
Aosta Valley	0.422***		0.349***	
	(28.892)		(32.698)	
Veneto	0.538***	211.985**	0.470***	429.101
	(38.385)	(2.282)	(54.690)	(1.632)

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)	0.700***		0.700***	
	(92.704)		(92.704)	
National (POLS)	0.745***	536.730***	0.745***	536.730***
	(49.497)	(3.678)	(49.497)	(3.678)
Abruzzo	0.691***		0.691***	
	(60.481)		(60.481)	
Basilicata	0.526***	10.583**	0.551***	
	(16.317)	(2.015)	(18.180)	
P.A. Bolzano	0.470***	33.685***	0.470***	33.685***
	(34.914)	(3.918)	(34.914)	(3.918)
Calabria	0.586***		0.586***	
	(30.517)		(30.517)	
Emilia-Romagna	0.591***	319.867***	0.591***	319.867***
	(53.855)	(3.239)	(53.855)	(3.239)
Friuli Venezia Giulia	0.580***	26.799*	0.590***	
	(37.623)	(1.784)	(41.081)	

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	0.786***		0.786***	
	(43.142)		(43.142)	
Liguria	0.608***		0.608***	
	(28.836)		(28.836)	
Lombardy	0.785***		0.785***	
	(132.494)		(132.494)	
Marche	0.689***		0.689***	
	(35.744)		(35.744)	
Molise	0.801***		0.801***	
	(29.160)		(29.160)	
Piedmont	0.657***		0.657***	
	(48.479)		(48.479)	
Apulia	0.667***		0.667***	
	(40.380)		(40.380)	
Sardinia	0.558***		0.558***	
	(39.712)		(39.712)	
Tuscany	0.538***	104.123	0.546***	
	(36.718)	(1.540)	(39.980)	
Umbria	0.393***	13.286***	0.393***	13.286***
	(20.123)	(3.786)	(20.123)	(3.786)
Aosta Valley	0.422***		0.422***	
	(28.892)		(28.892)	
Veneto	0.538***	211.985**	0.538***	211.985**
	(38.385)	(2.282)	(38.385)	(2.282)

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

B.2 Results for the Within and Between-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 (7.1) 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 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	0.149	0.011***		-0.098	0.017***	
	(1.169)	(4.274)		(-0.727)	(5.411)	
Basilicata	0.736***	-4.135×10^{-4}**	14.312***	0.691***	-5.265×10^{-4}***	70.108***
	(8.458)	(−2.584)	(2.698)	(9.708)	(−3.320)	(2.743)
P.A. Bolzano	0.421***	3.809×10^{-4}	31.133***	0.364***	2.614×10^{-4}*	92.402***
	(10.940)	(1.375)	(3.555)	(14.000)	(1.809)	(3.691)
Calabria	0.753***	-1.313×10^{-3}**	26.157*	0.756***	1.996×10^{-3}***	132.166
	(10.017)	(−2.450)	(1.960)	(9.088)	(−2.629)	(1.589)
Emilia-Romagna	0.604***	-1.534×10^{-3}	324.983***	0.475***	5.215×10^{-3}	1075.270***
	(11.560)	(−0.264)	(3.214)	(14.276)	(1.404)	(3.002)
Friuli Venezia Giulia	0.763***	-2.095×10^{-3}***	41.340***	0.705***	-1.947×10^{-3}***	139.824***
	(11.896)	(−2.942)	(2.704)	(16.205)	(−4.640)	(2.893)
Lazio	0.106	0.040***	130.051*	0.229***	0.035***	782.017*
	(1.583)	(10.272)	(1.783)	(4.467)	(10.129)	(1.960)
Liguria	0.582***	9.788×10^{-4}		0.598***	-1.727×10^{-3}	
	(6.558)	(0.306)		(6.930)	(−0.514)	
Lombardy	0.642***	0.105***	364.667**	0.591***	0.074***	2067.970**
	(45.53)	(10.773)	(2.027)	(34.369)	(6.588)	(2.190)
Marche	0.052	0.024***	−112.390*	−0.308**	0.035***	−651.703**
	(0.414)	(5.155)	(−1.774)	(−2.578)	(7.987)	(−2.468)
Molise	−0.087*	2.455×10^{-3}***	−5.585*	−0.046	2.506×10^{-3}***	−25.810
	(−1.837)	(19.605)	(−1.671)	(−1.287)	(23.760)	(−1.651)
Piedmont	0.670***	-2.492×10^{-3}		0.642***	-9.712×10^{-3}	
	(7.326)	(−0.146)		(6.777)	(−0.536)	
Apulia	0.611***	1.748×10^{-3}		0.185*	0.019***	−596.977*
	(6.508)	(0.598)		(1.759)	(4.355)	(−1.783)
Sardinia	0.731***	-1.276×10^{-3}***	17.6778*	0.757***	-2.273×10^{-3}***	105.529*
	(14.796)	(−3.726)	(1.909)	(15.340)	(−4.996)	(1.930)
Tuscany	0.654***	-6.650×10^{-3}**	159.062**	0.650***	-9.361×10^{-3}***	746.518**
	(12.899)	(−2.386)	(2.274)	(12.999)	(−3.268)	(2.163)

Table B.3 continues on next page

Table B.3 continued from previous page

	Regular model			Modelling undocumented infectives		
Region	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Umbria	0.315***	1.758×10^{-4}*	11.181***	0.275***	1.398×10^{-3}**	40.840***
	(6.746)	(1.835)	(3.062)	(10.138)	(2.425)	(3.399)
Aosta Valley	0.355***	1.240×10^{-4}***		0.282***	1.096×10^{-4}***	
	(12.796)	(2.794)		(14.499)	(3.860)	
Veneto	0.664***	−0.011***	294.301***	0.604***	-8.899×10^{-3}***	819.111***
	(14.282)	(−2.832)	(3.120)	(21.351)	(−5.019)	(3.331)

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	0.149	0.011***		0.149	0.011***	
	(1.169)	(4.274)		(1.169)	(4.274)	
Basilicata	0.736***	-4.135×10^{-4}**	14.312***	0.736***	-4.135×10^{-4}**	14.312***
	(8.458)	(−2.584)	(2.698)	(8.458)	(−2.584)	(2.698)
P.A. Bolzano	0.421***	3.809×10^{-4}	31.133***	0.421***	3.809×10^{-4}	31.133***
	(10.940)	(1.375)	(3.555)	(10.940)	(1.375)	(3.555)
Calabria	0.753***	-1.313×10^{-3}**	26.157*	0.726***	-9.995×10^{-4}*	
	(10.017)	(−2.450)	(1.960)	(9.685)	(−1.926)	
Emilia-Romagna	0.604***	-1.534×10^{-3}	324.983***	0.604***	-1.534×10^{-3}	324.983***
	(11.560)	(−0.264)	(3.214)	(11.560)	(−0.264)	(3.214)
Friuli Venezia Giulia	0.763***	-2.095×10^{-3}***	41.340***	0.763***	-2.095×10^{-3}***	41.340***
	(11.896)	(−2.942)	(2.704)	(11.896)	(−2.942)	(2.704)
Lazio	0.106	0.040***	130.051*	0.128*	0.039***	
	(1.583)	(10.272)	(1.783)	(1.924)	(10.043)	
Liguria	0.582***	9.788×10^{-4}		0.582***	9.788×10^{-4}	
	(6.558)	(0.306)		(6.558)	(0.306)	
Lombardy	0.642***	0.105***	364.667**	0.651***	0.101***	
	(45.53)	(10.773)	(2.027)	(48.017)	(10.417)	
Marche	0.052	0.024***	−112.390*	0.110	0.021***	
	(0.414)	(5.155)	(−1.774)	(0.900)	(4.790)	
Molise	−0.087*	2.455×10^{-3}***	−5.585*	−0.091*	2.438×10^{-3}***	
	(−1.837)	(19.605)	(−1.671)	(−1.906)	(19.357)	
Piedmont	0.670***	-2.492×10^{-3}		0.670***	-2.492×10^{-3}	
	(7.326)	(−0.146)		(7.326)	(−0.146)	
Apulia	0.611***	1.748×10^{-3}		0.611***	1.748×10^{-3}	
	(6.508)	(0.598)		(6.508)	(0.598)	
Sardinia	0.731***	-1.276×10^{-3}***	17.678*	0.707***	-1.033×10^{-3}***	
	(14.796)	(−3.726)	(1.909)	(14.611)	(−3.207)	

Table B.4 continues on next page

Table B.4 continued from previous page

	Model selection with AIC			Model selection with BIC		
Region	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
Tuscany	0.654***	-6.650×10^{-3}**	159.062**	0.654***	-6.650×10^{-3}**	159.062**
	(12.899)	(−2.386)	(2.274)	(12.899)	(−2.386)	(2.274)
Umbria	0.315***	1.758×10^{-4}*	11.181***	0.315***	1.758×10^{-4}*	11.181***
	(6.746)	(1.835)	(3.062)	(6.746)	(1.835)	(3.062)
Aosta Valley	0.355***	1.240×10^{-4}***		0.355***	1.240×10^{-4}***	
	(12.796)	(2.794)		(12.796)	(2.794)	
Veneto	0.664***	−0.011***	294.301***	0.664***	−0.011***	294.301***
	(14.282)	(−2.832)	(3.120)	(14.282)	(−2.832)	(3.120)

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

C Figures

C.1 Figures for Section 2: Problem Description

In this appendix, we present additional plots as referenced in Section 2.

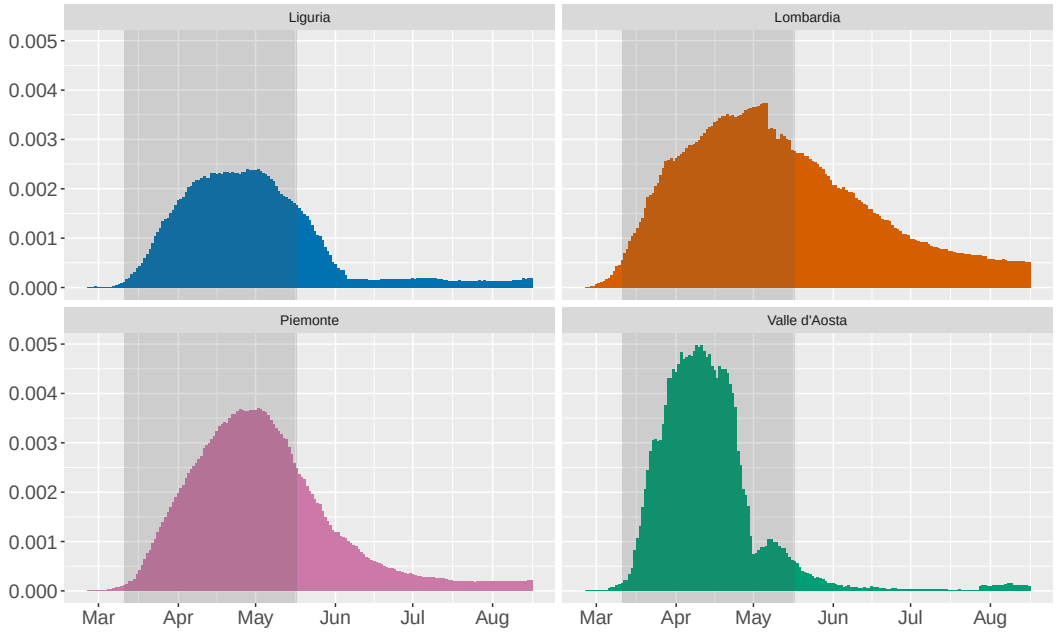


Figure C.1. Incidence rate per region for the Nord-Ovest NUTS 1 region. The grey area indicates the national lockdown.

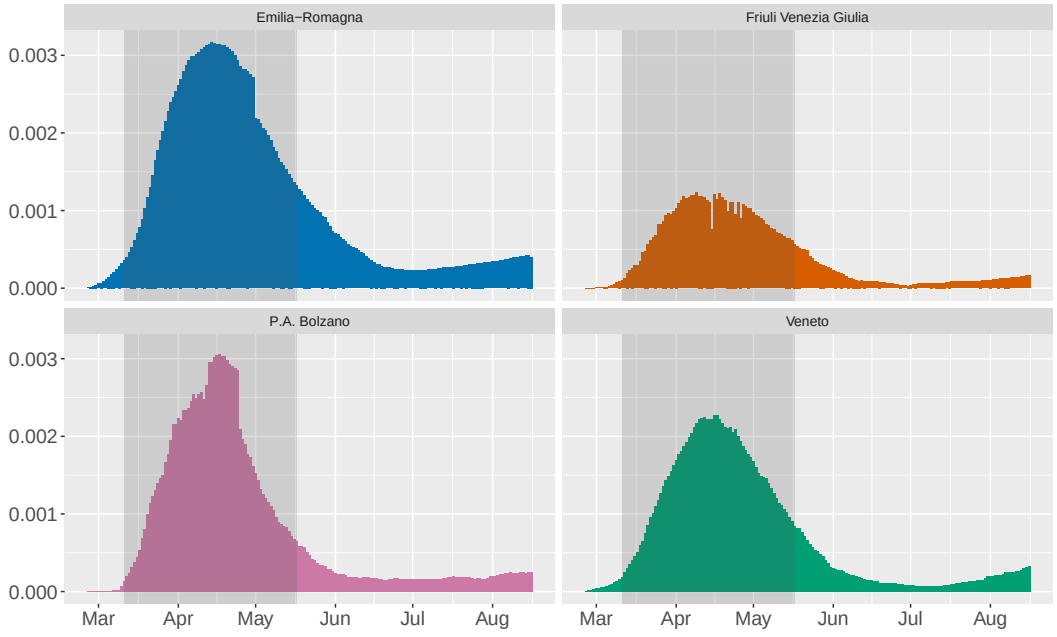


Figure C.2. Incidence rate per region for the Nord-Est NUTS 1 region. The grey area indicates the national lockdown.

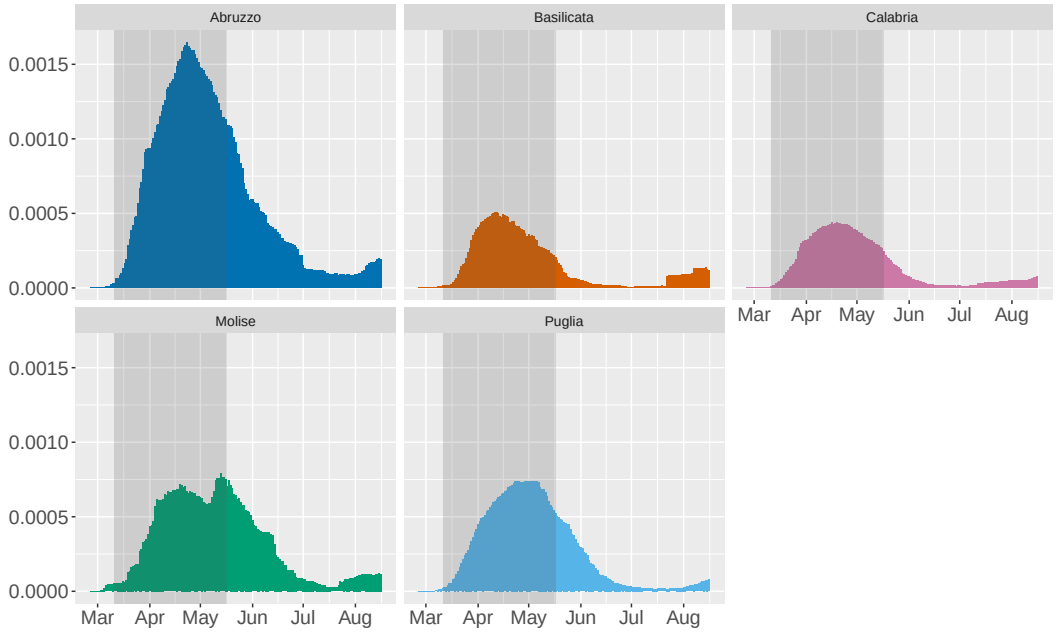


Figure C.3. Incidence rate per region for the Sud NUTS 1 region. The grey area indicates the national lockdown.

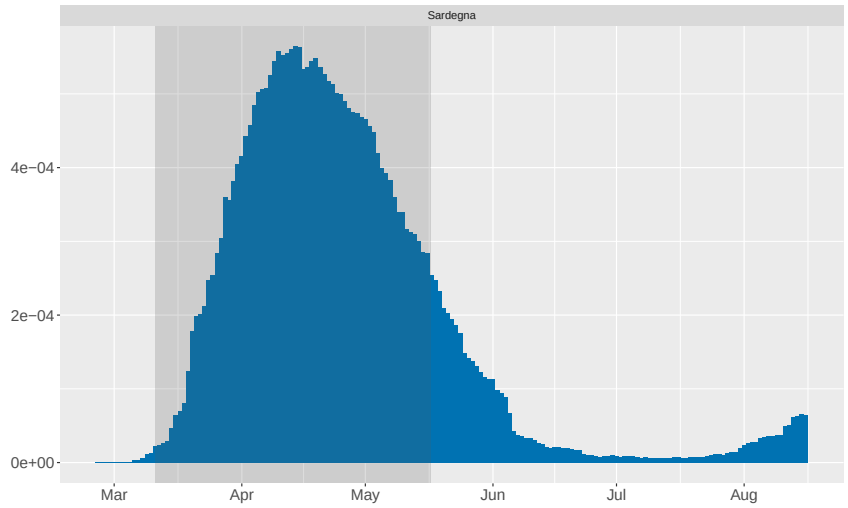
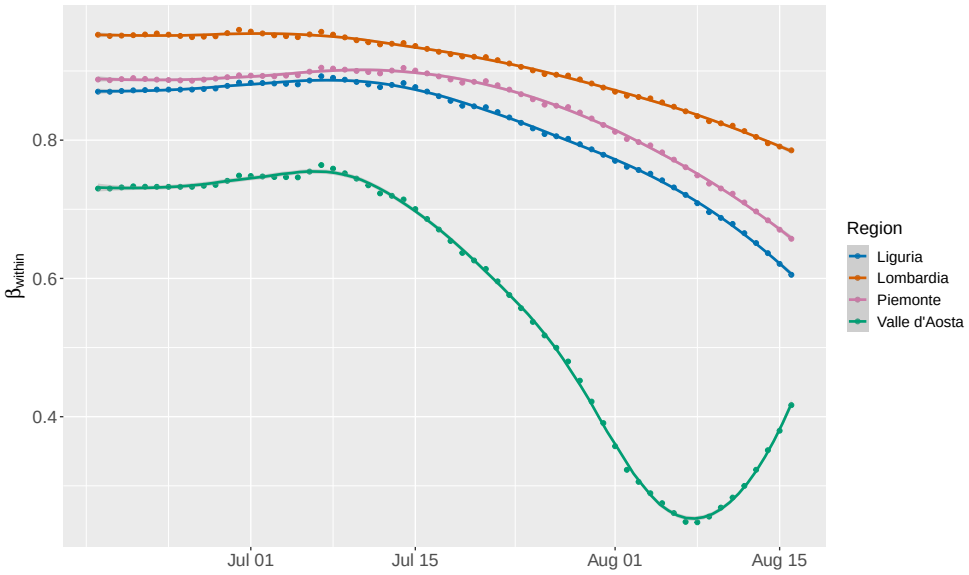


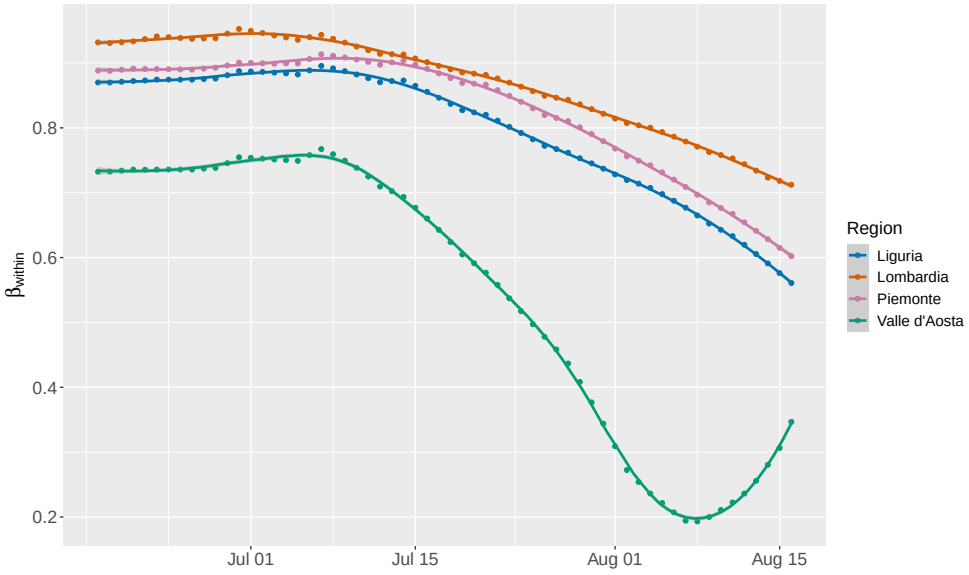
Figure C.4. Incidence rate per region for the Isole NUTS 1 region. The grey area indicates the national lockdown.

C.2 Figures for the Within-Region Spread Model

In this appendix, we present additional plots as referenced in Section 6.2.

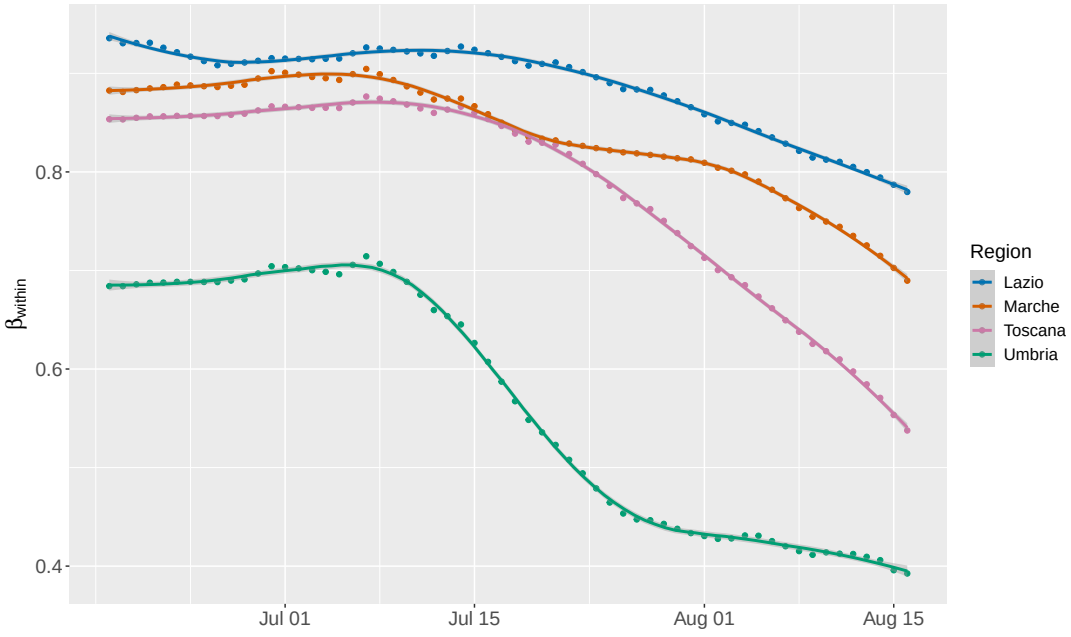


(a) Infectives exclude undocumented cases

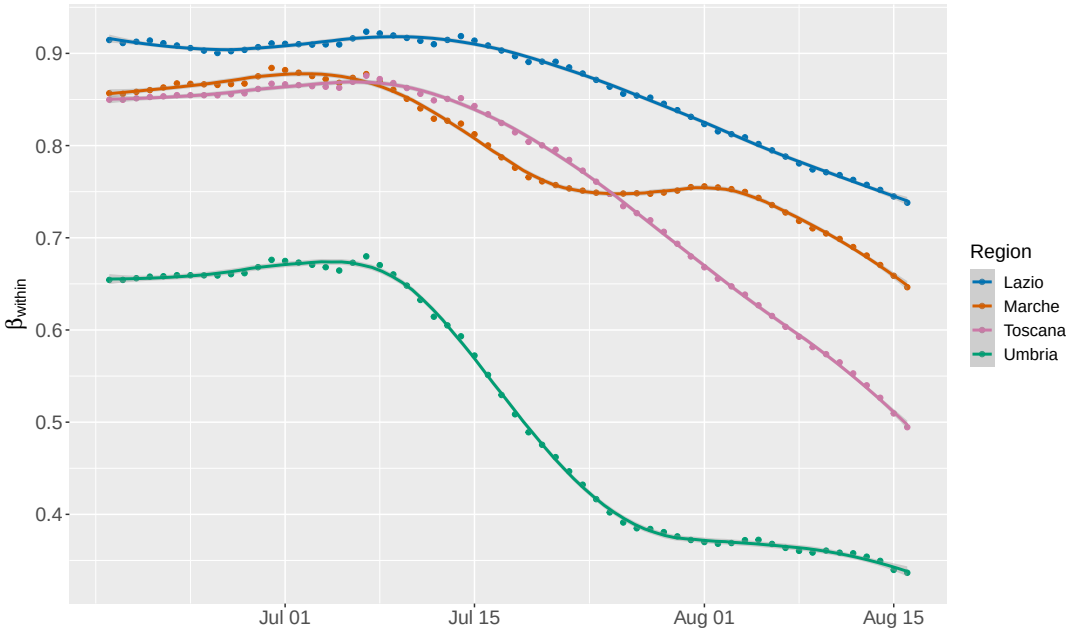


(b) Infectives include undocumented cases

Figure C.5. Progression of β_{within} over time for the Nord-Ovest NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

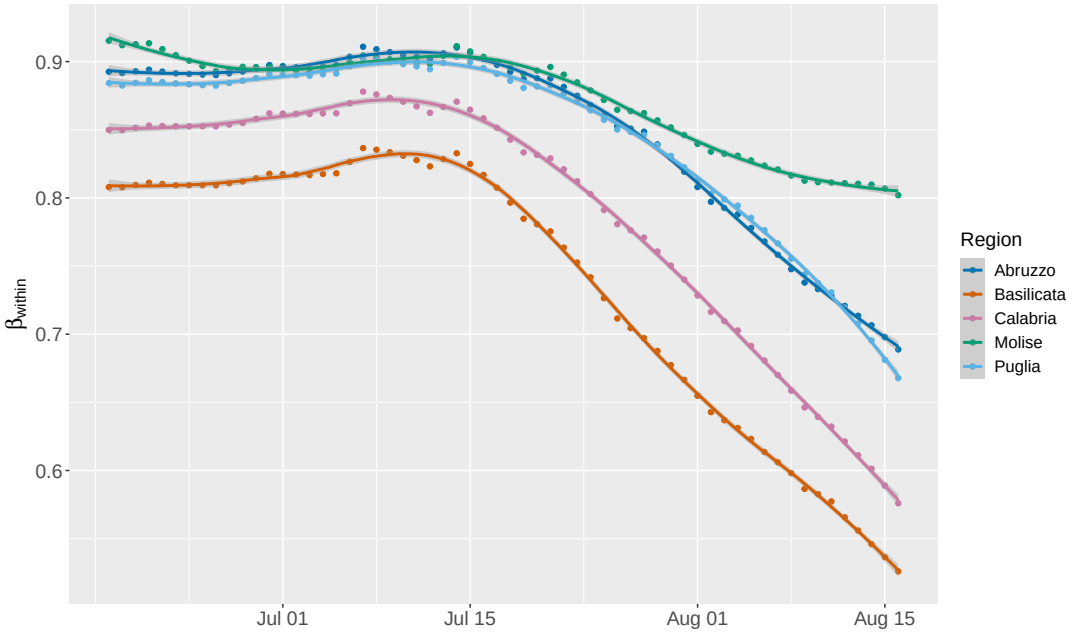


(a) Infectives exclude undocumented cases

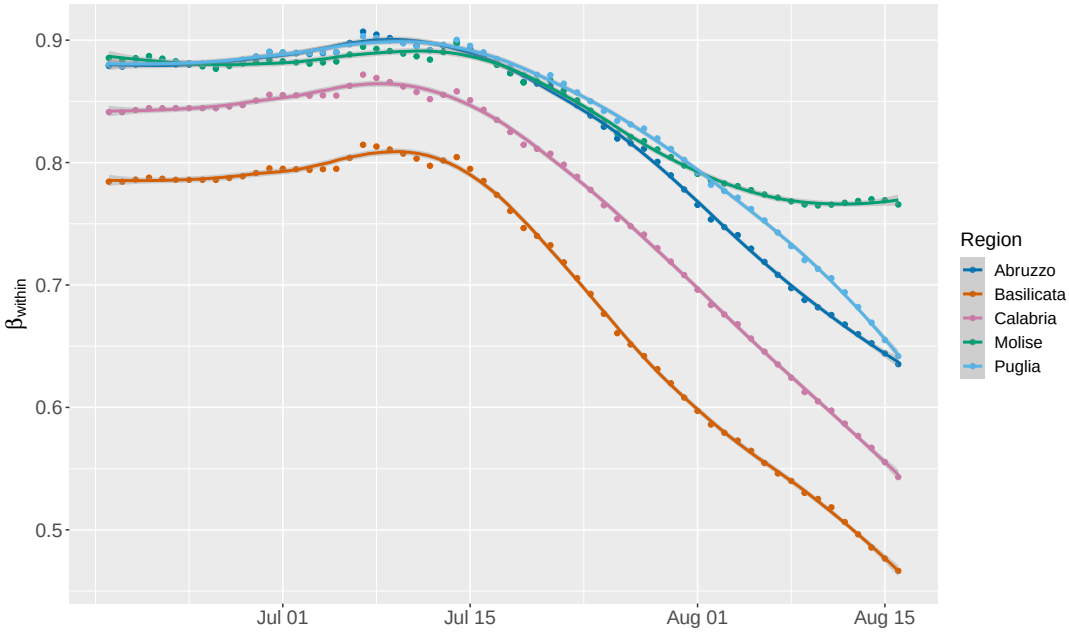


(b) Infectives include undocumented cases

Figure C.6. Progression of β_{within} over time for the Centro (IT) NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

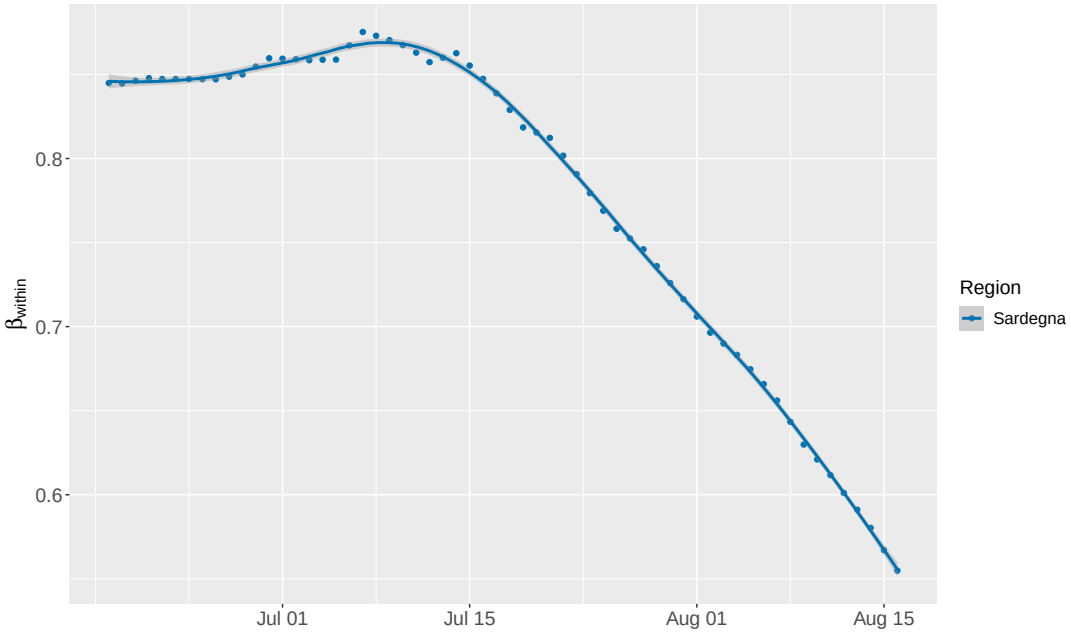


(a) Infectives exclude undocumented cases

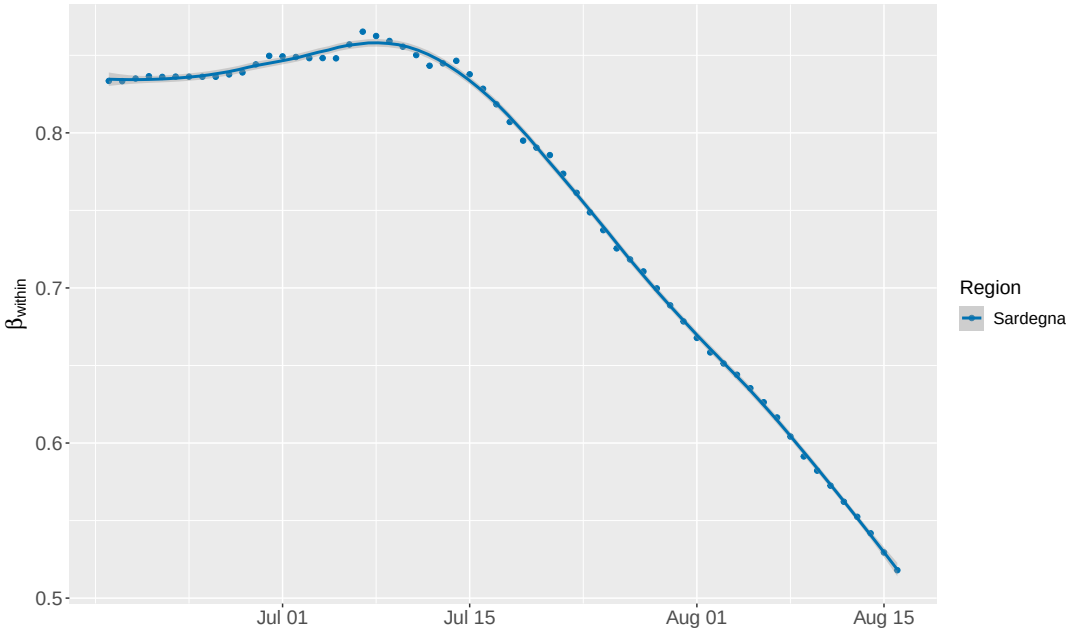


(b) Infectives include undocumented cases

Figure C.7. Progression of β_{within} over time for the Sud NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

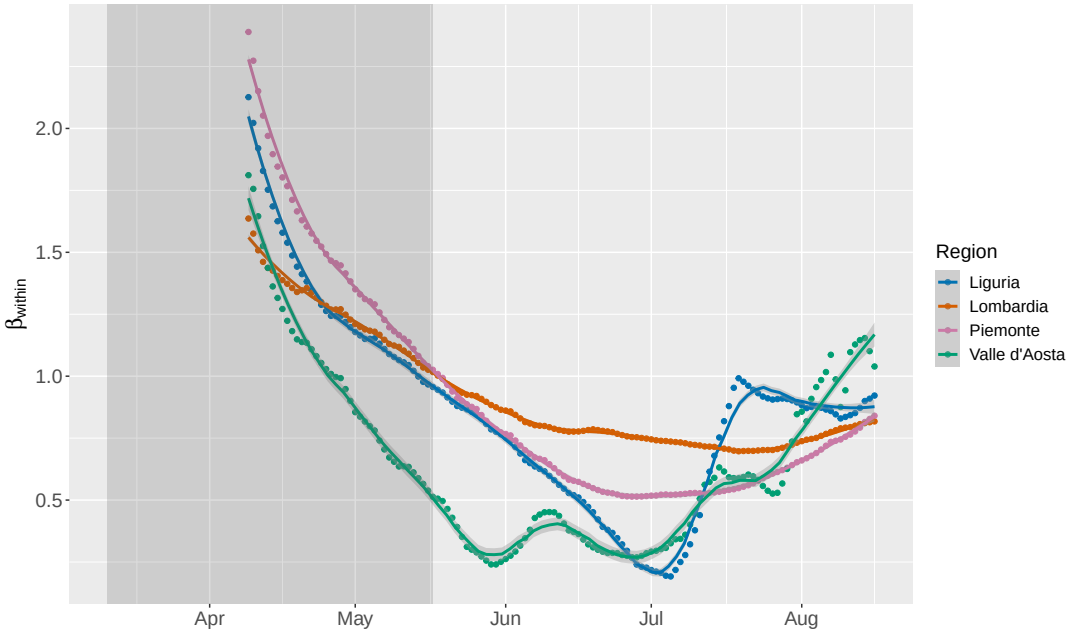


(a) Infectives exclude undocumented cases

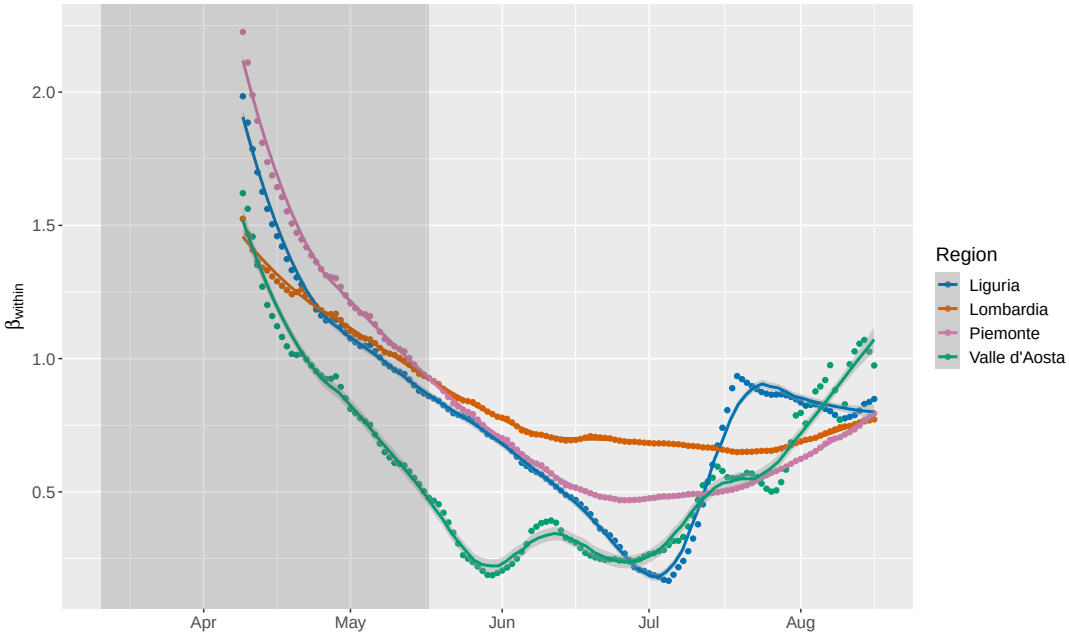


(b) Infectives include undocumented cases

Figure C.8. Progression of β_{within} over time for the Isole NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

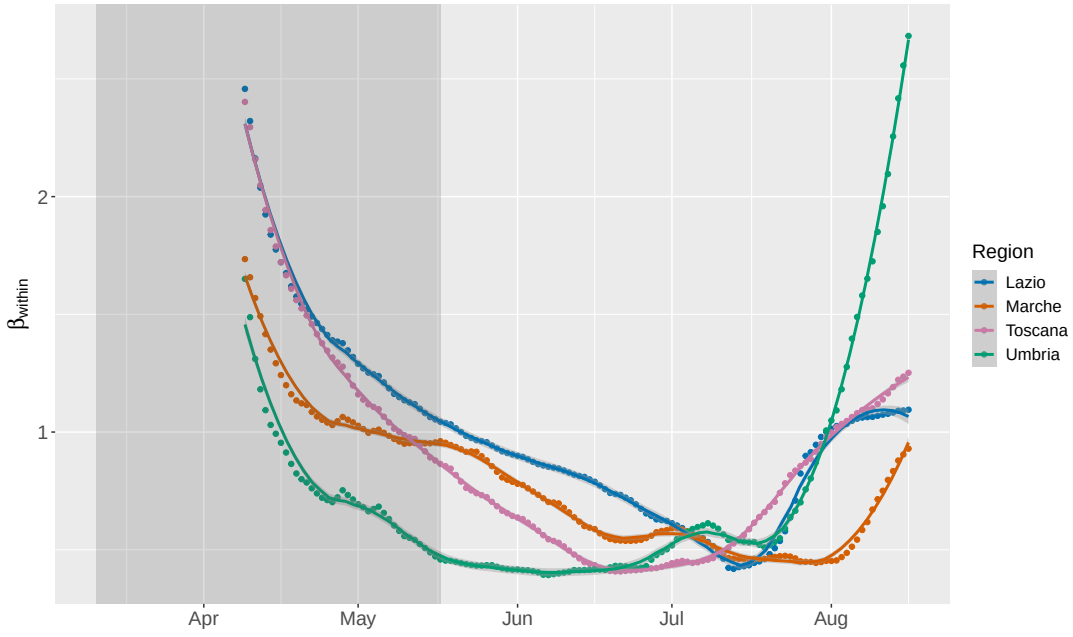


(a) Infectives exclude undocumented cases

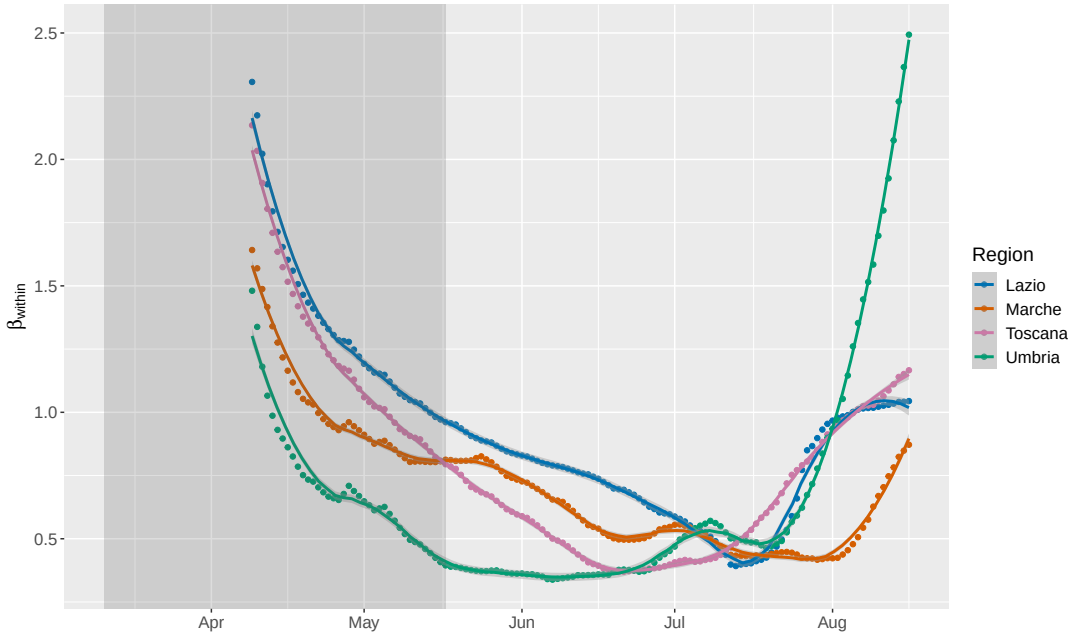


(b) Infectives include undocumented cases

Figure C.9. Progression of β_{within} over time for the Nord-Ovest NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

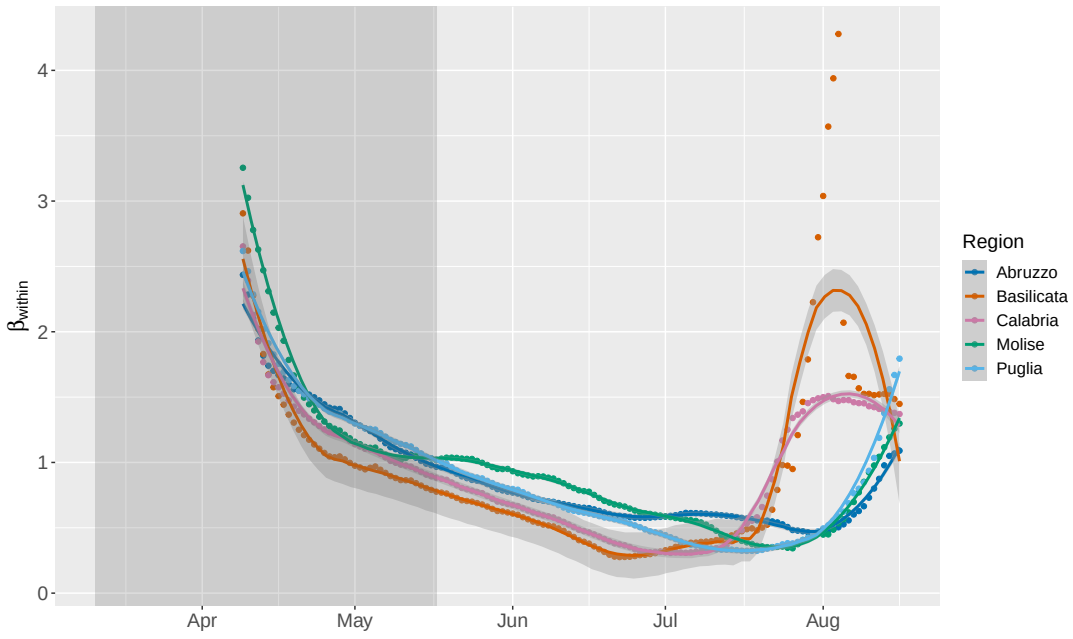


(a) Infectives exclude undocumented cases

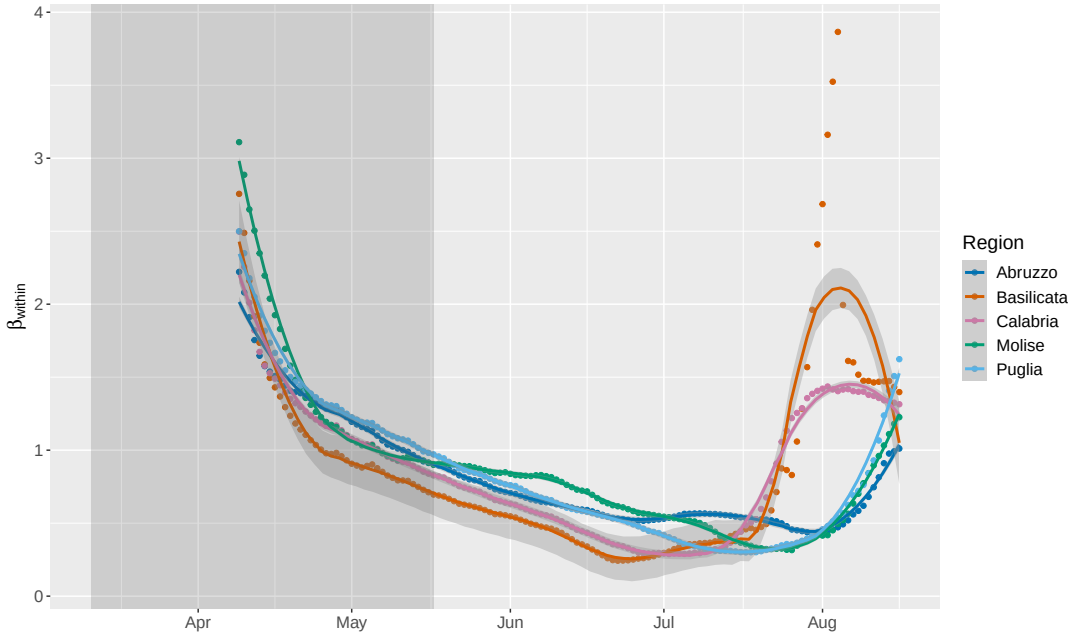


(b) Infectives include undocumented cases

Figure C.10. Progression of β_{within} over time for the Centro (IT) NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

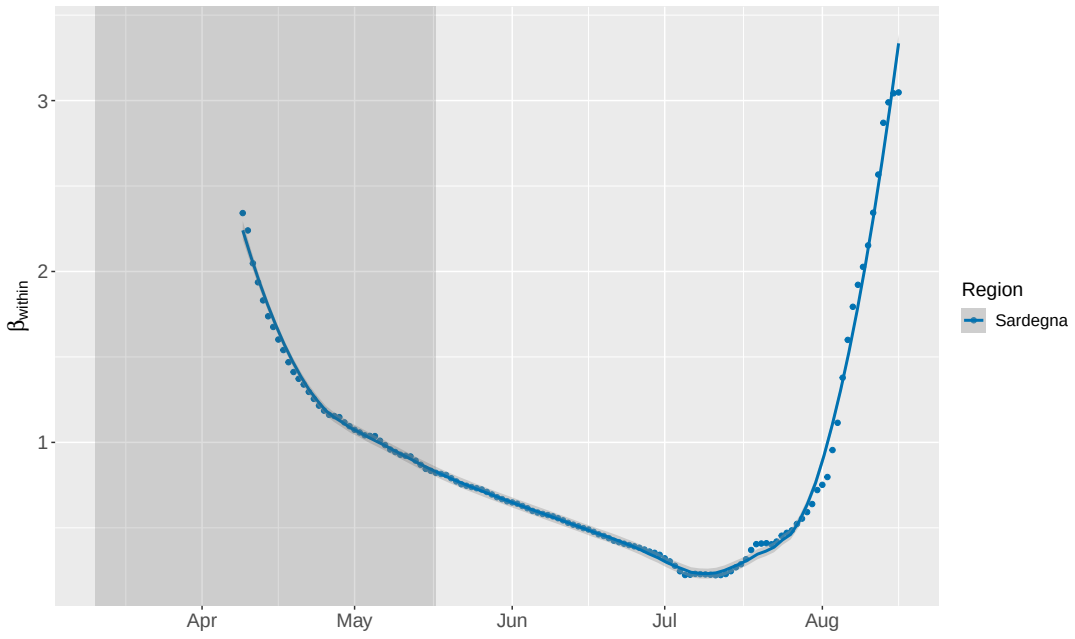


(a) Infectives exclude undocumented cases

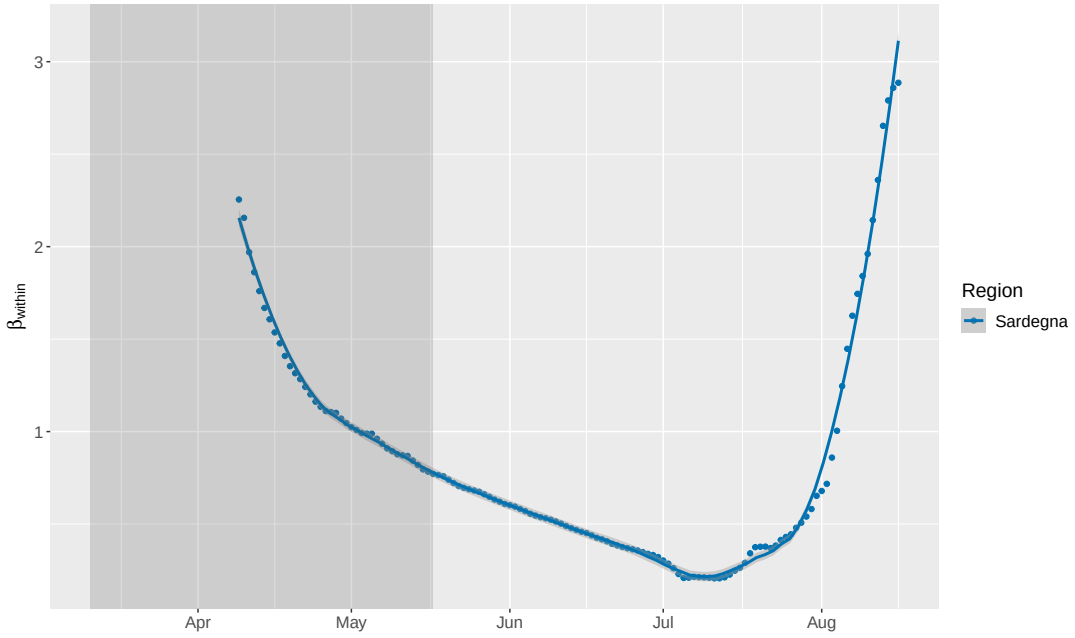


(b) Infectives include undocumented cases

Figure C.11. Progression of β_{within} over time for the Sud NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.



(a) Infectives exclude undocumented cases

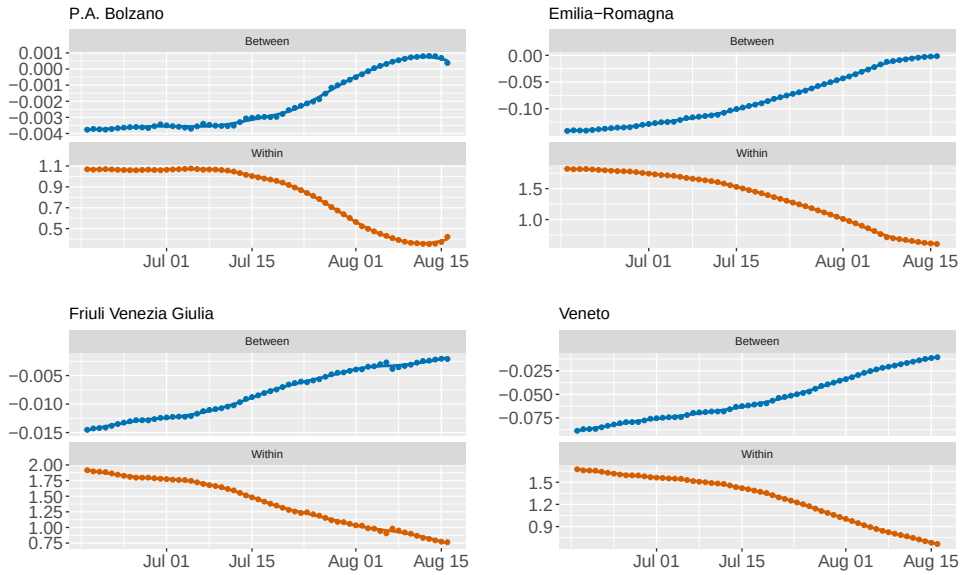


(b) Infectives include undocumented cases

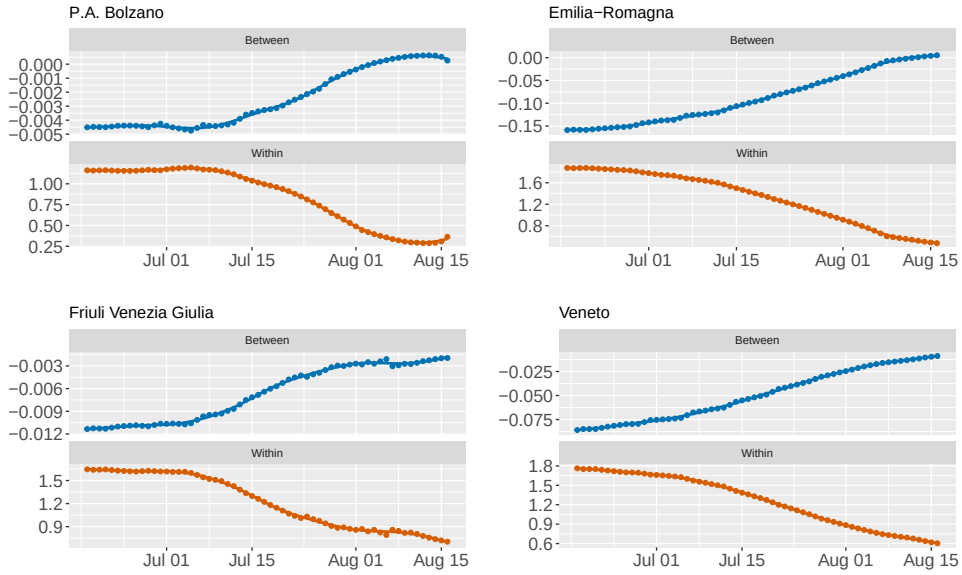
Figure C.12. Progression of β_{within} over time for the Isole NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

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

In this appendix, we present additional plots as referenced in Section 7.2.

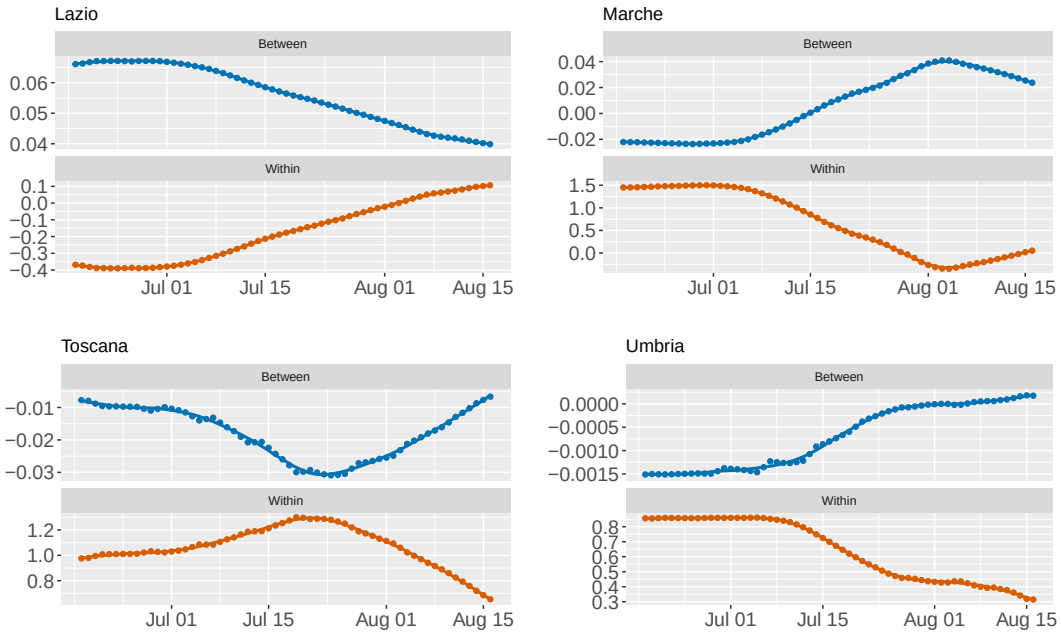


(a) Infectives exclude undocumented cases

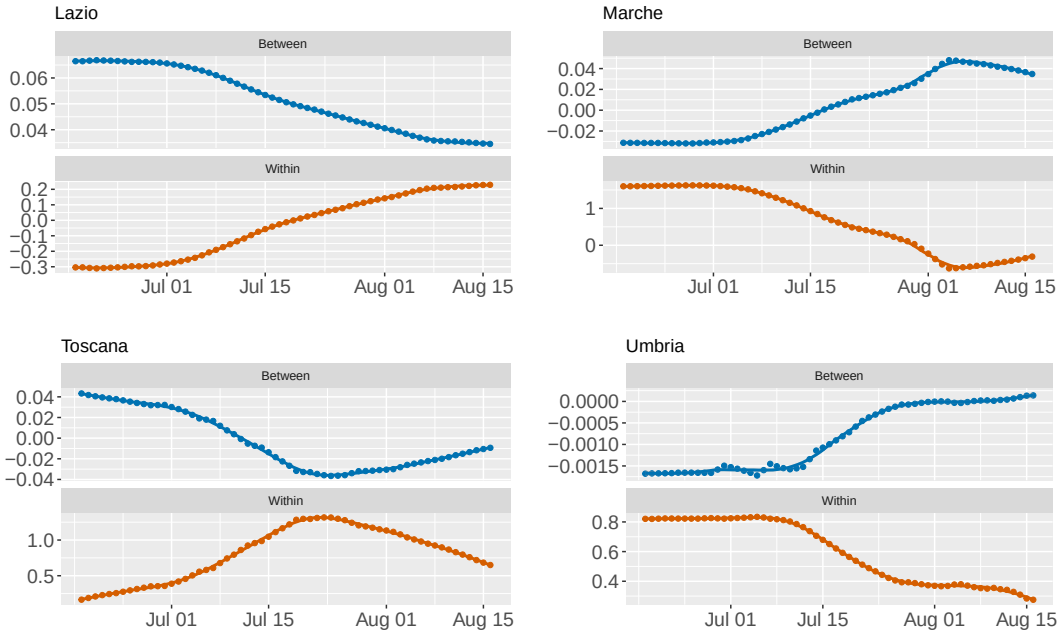


(b) Infectives include undocumented cases

Figure C.13. Progression of β_{within} and $\beta_{between}$ over time for the Nord-Est NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

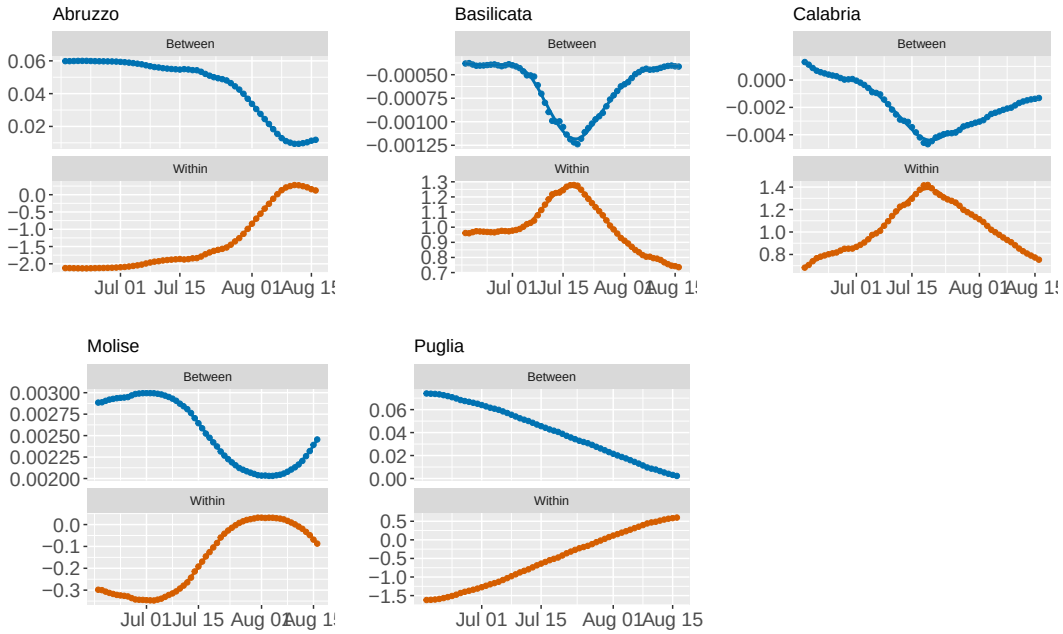


(a) Infectives exclude undocumented cases

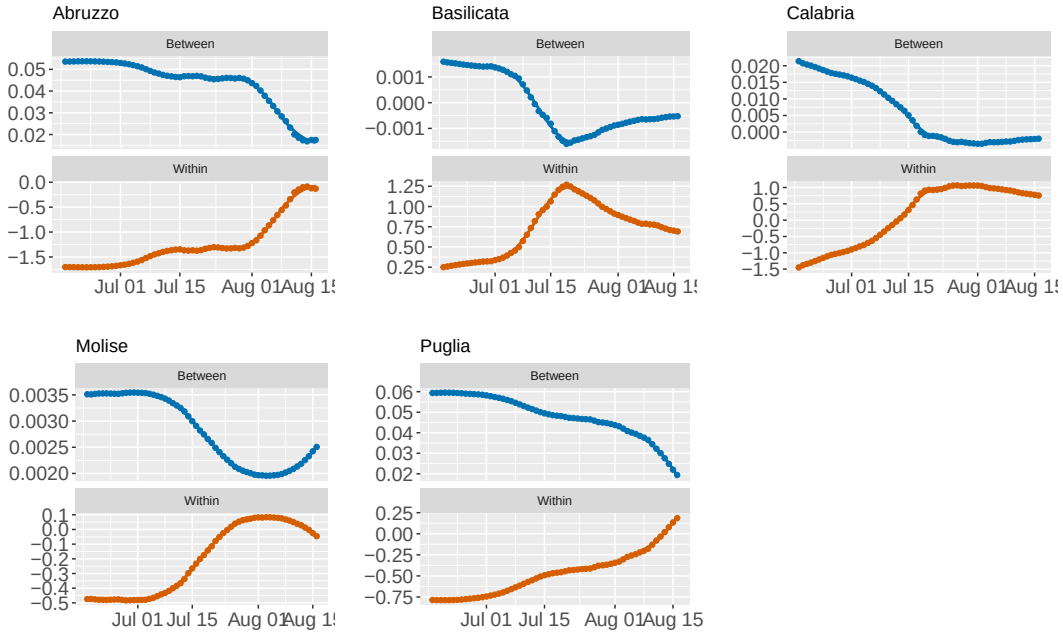


(b) Infectives include undocumented cases

Figure C.14. Progression of β_{within} and $\beta_{between}$ over time for the Centro (IT) NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

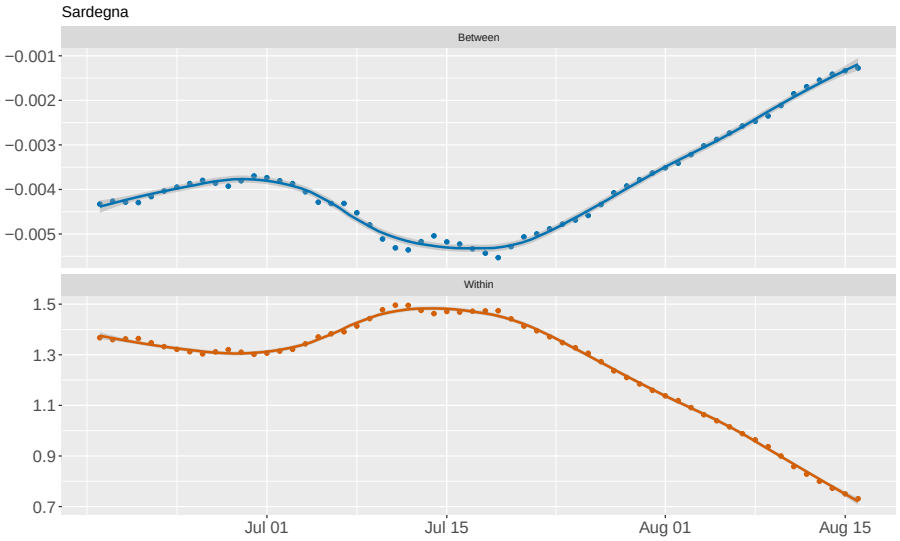


(a) Infectives exclude undocumented cases

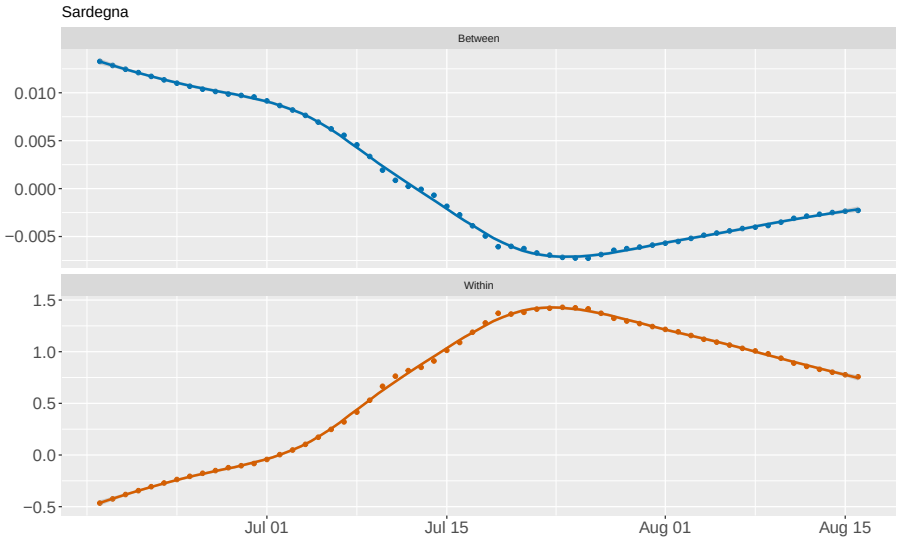


(b) Infectives include undocumented cases

Figure C.15. Progression of β_{within} and $\beta_{between}$ over time for the Sud NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



(a) Infectives exclude undocumented cases

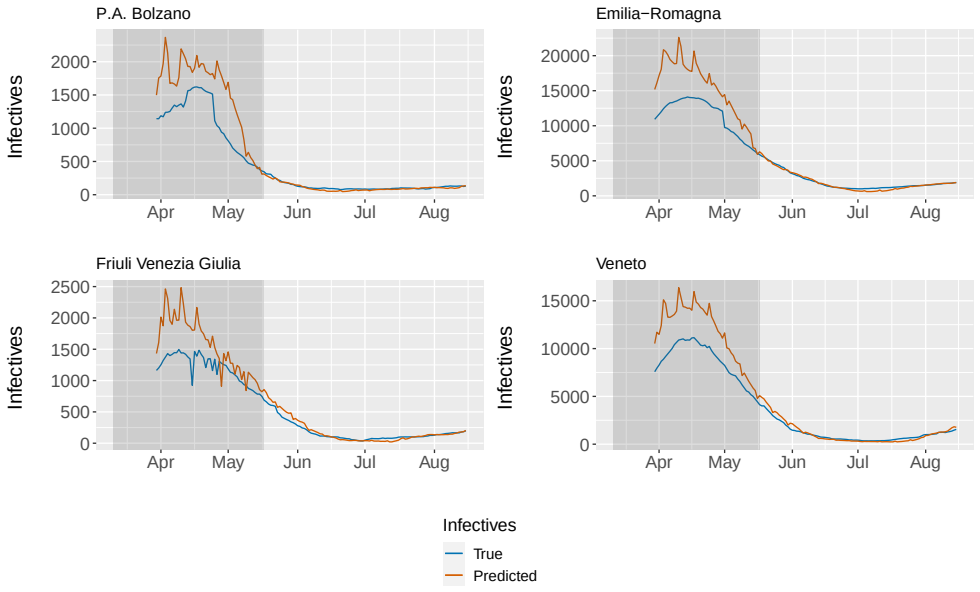


(b) Infectives include undocumented cases

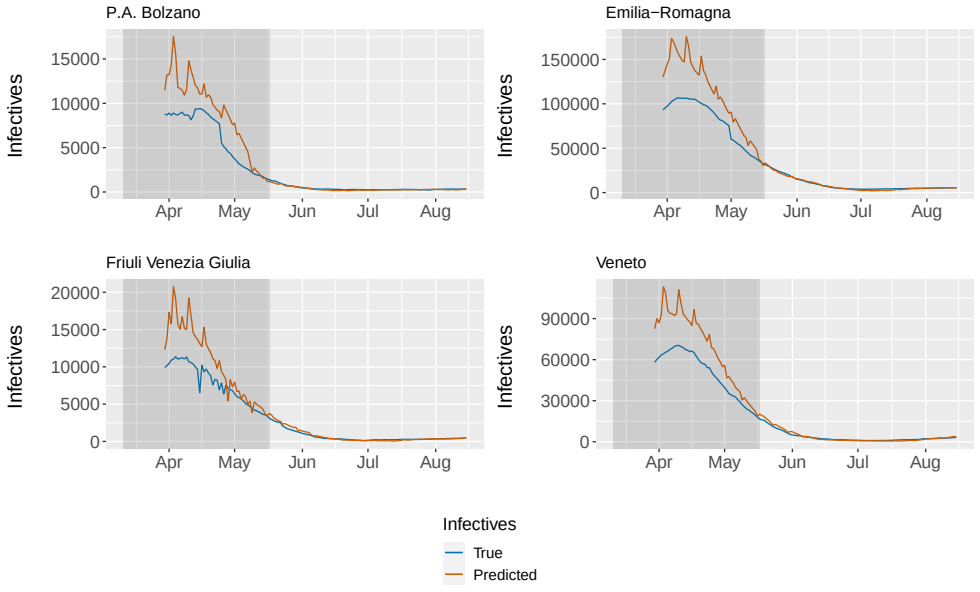
Figure C.16. Progression of β_{within} and $\beta_{between}$ over time for the Isole NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

C.4 Figures for Section 8: Forecasts

In this appendix, we present additional plots as referenced in Section 8.

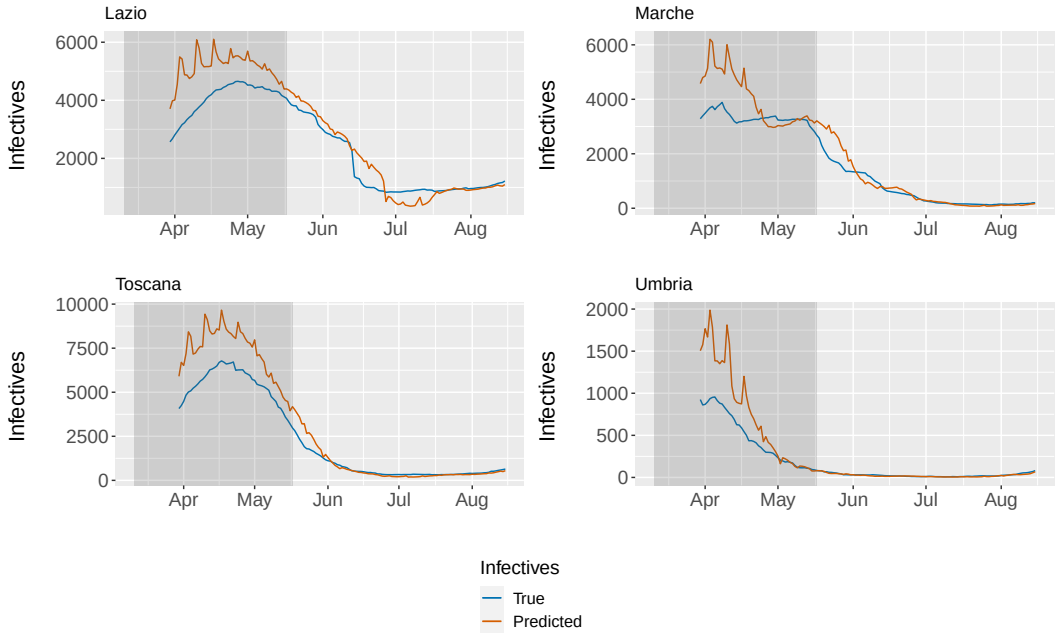


(a) Infectives exclude undocumented cases

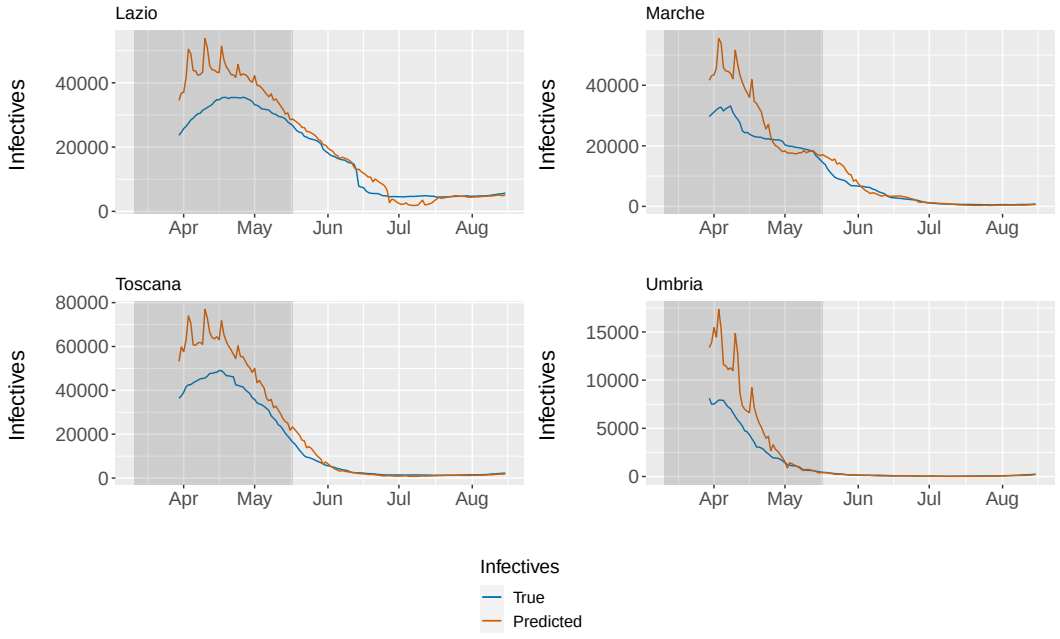


(b) Infectives include undocumented cases

Figure C.17. One-period ahead forecasts for the within-region spread model for the Nord-Est NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the lockdown.

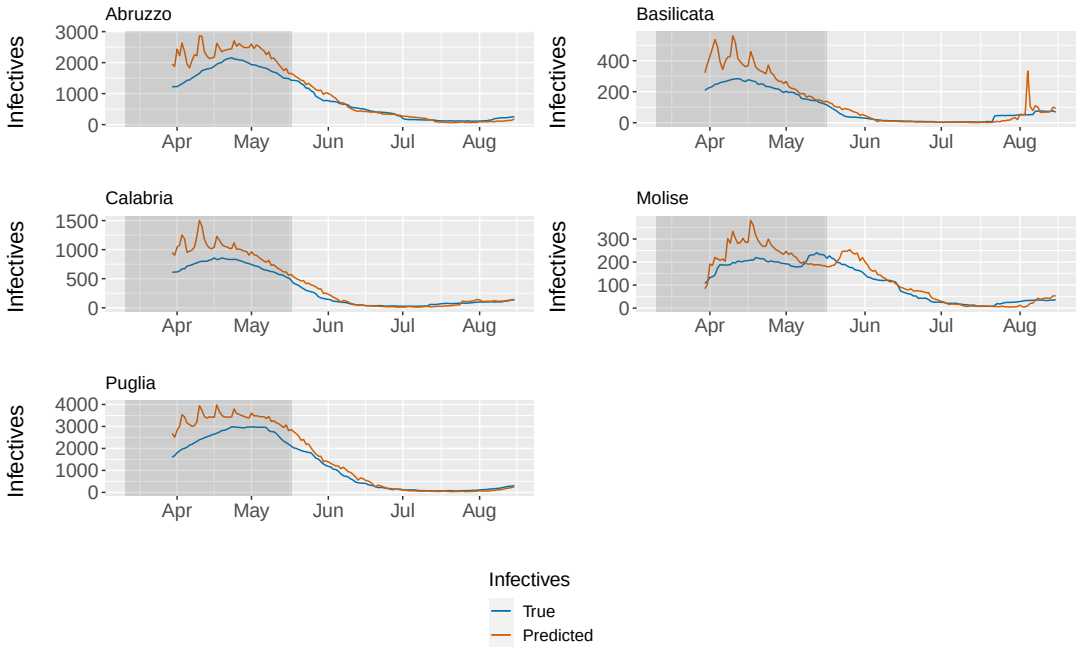


(a) Infectives exclude undocumented cases

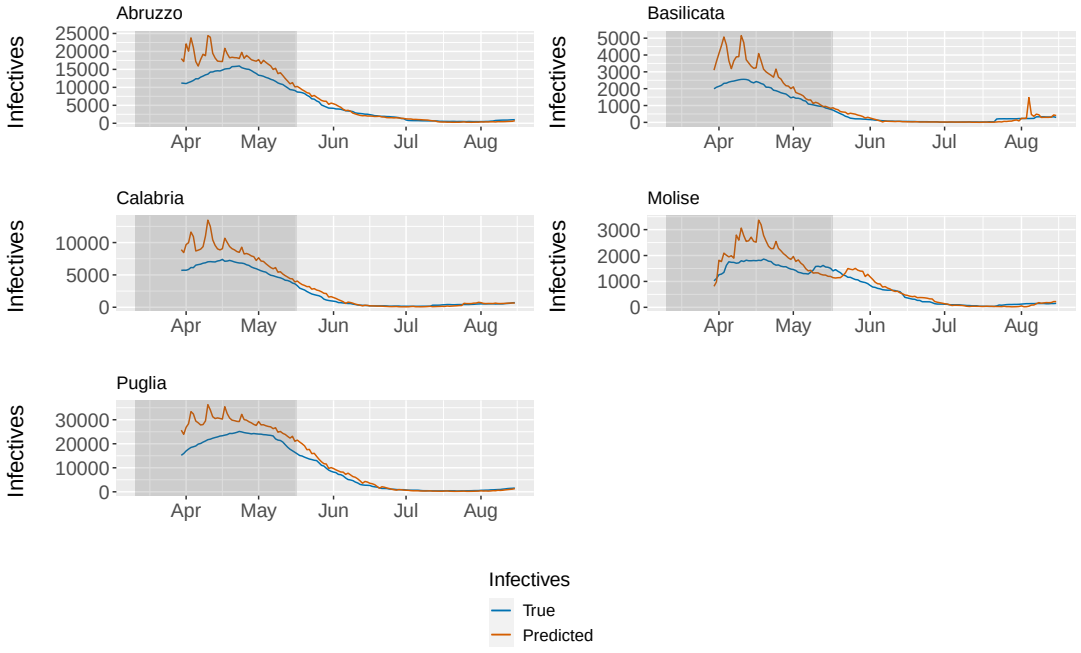


(b) Infectives include undocumented cases

Figure C.18. One-period ahead forecasts for the within-region spread model for the Centro (IT) NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the lockdown.

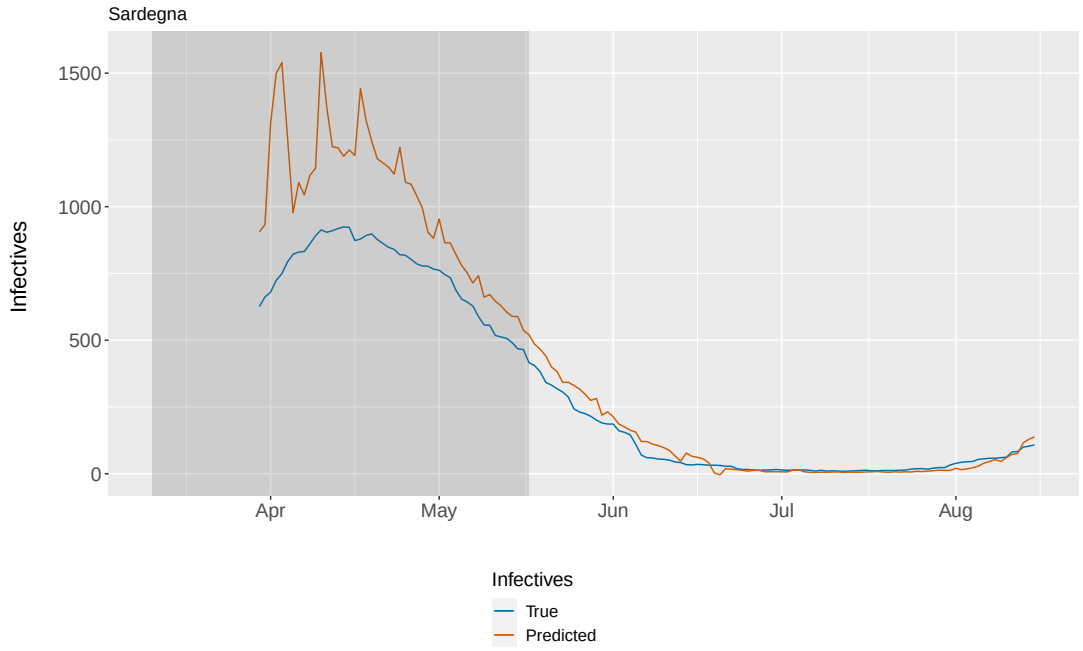


(a) Infectives exclude undocumented cases

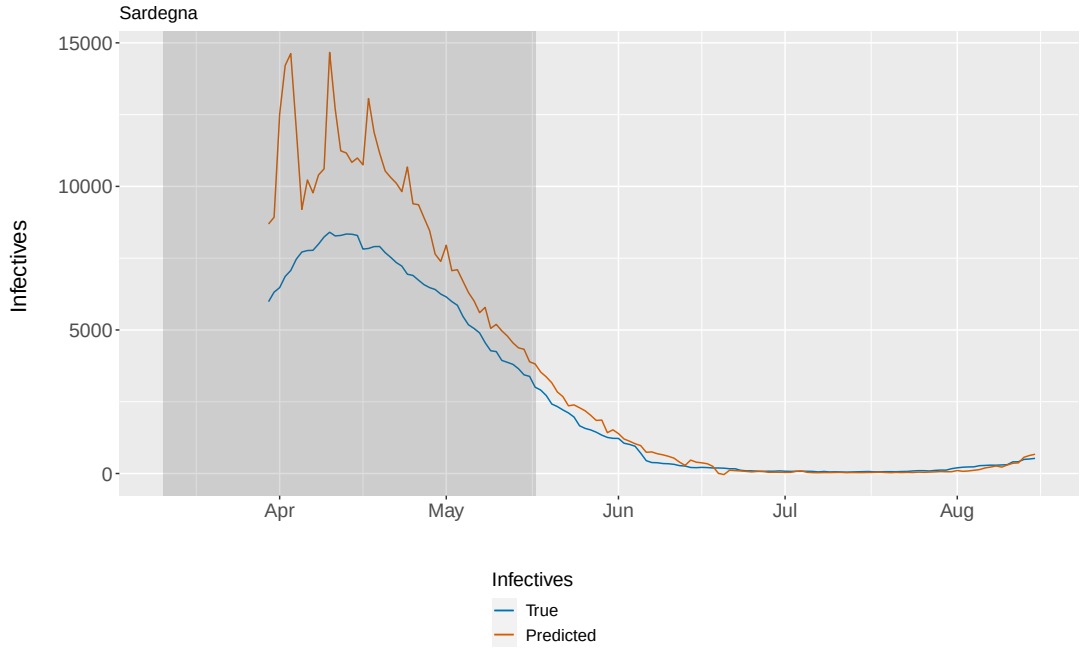


(b) Infectives include undocumented cases

Figure C.19. One-period ahead forecasts for the within-region spread model for the Sud NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the lockdown.

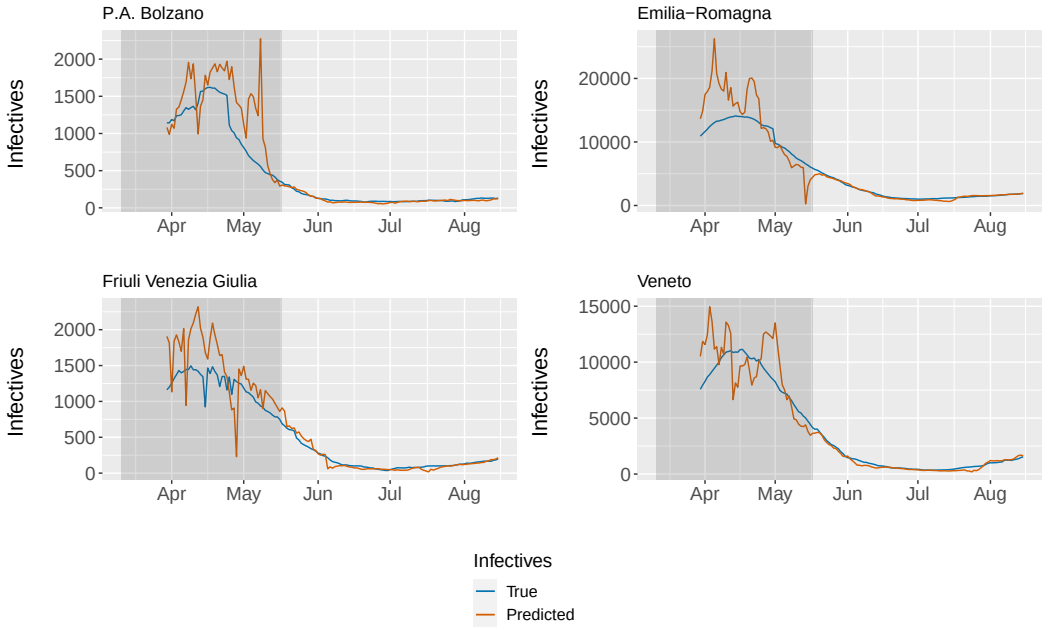


(a) Infectives exclude undocumented cases

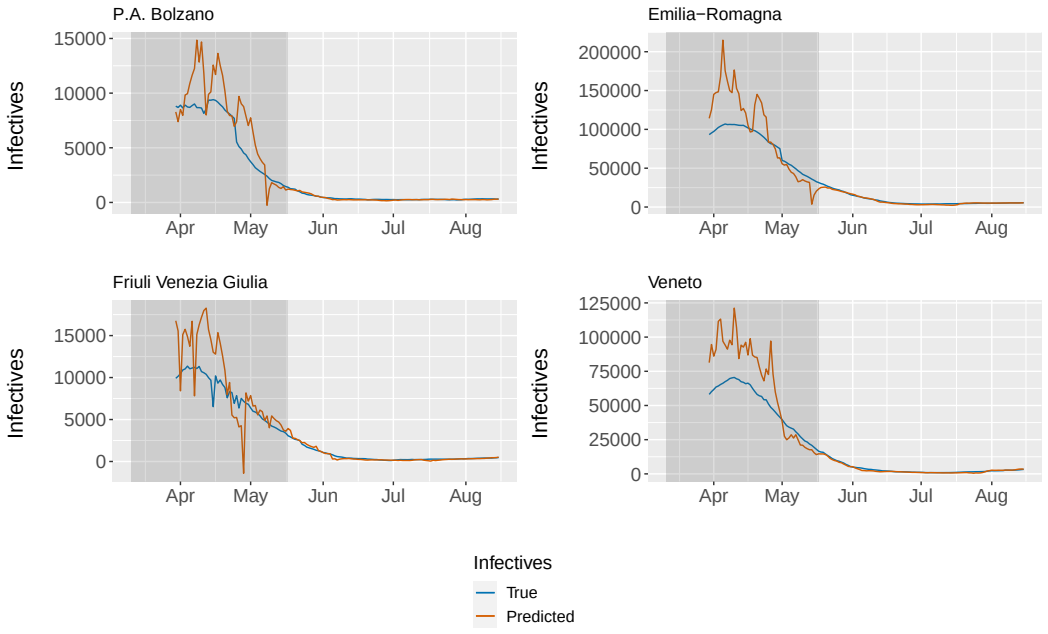


(b) Infectives include undocumented cases

Figure C.20. One-period ahead forecasts for the within-region spread model for the Isole NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the lockdown.

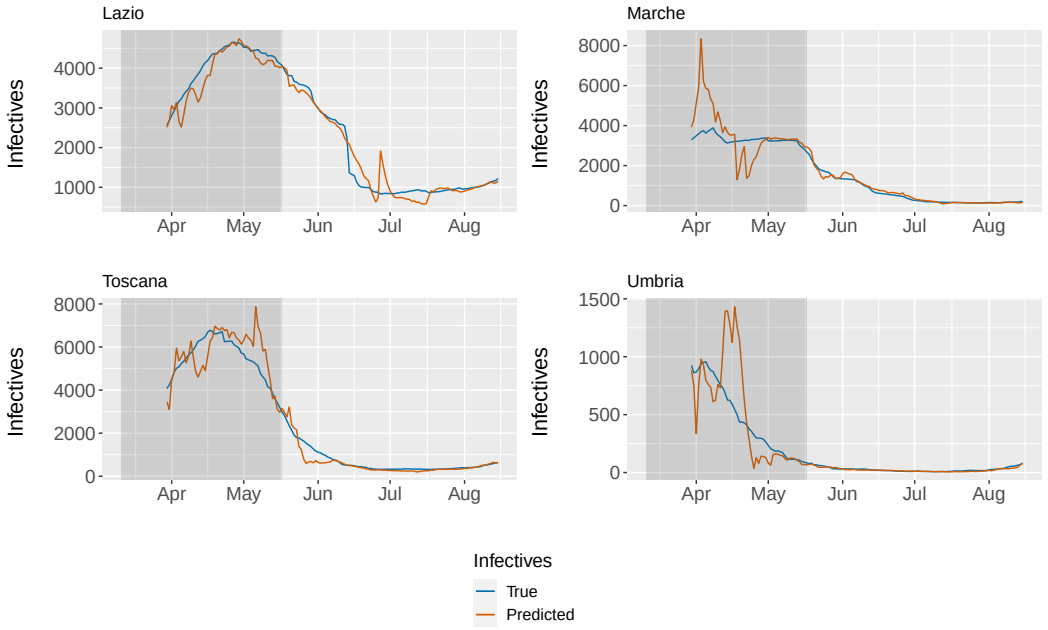


(a) Infectives exclude undocumented cases

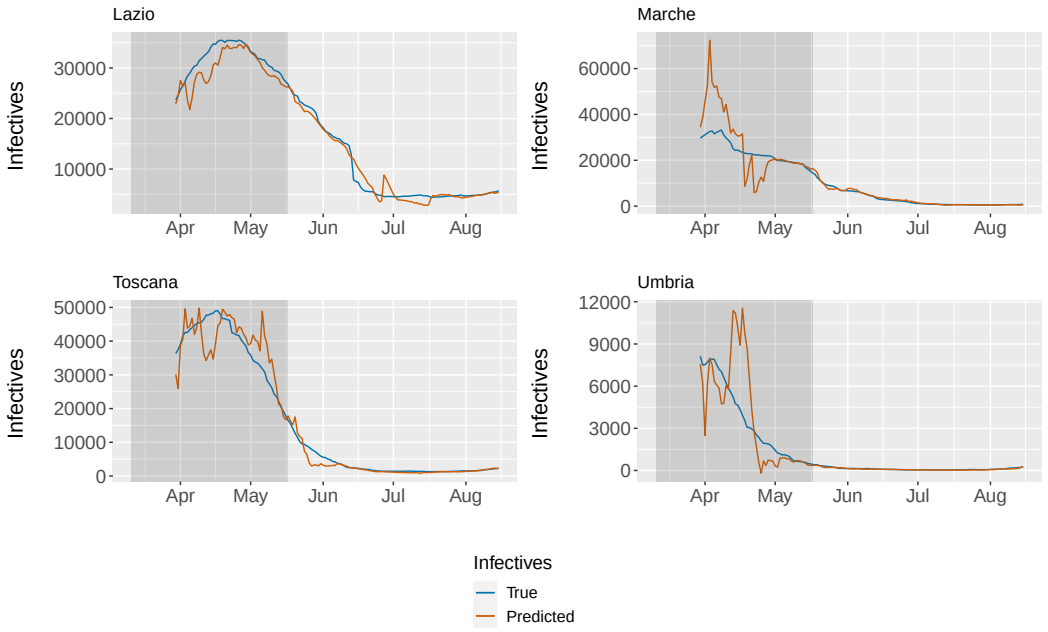


(b) Infectives include undocumented cases

Figure C.21. One-period ahead forecasts for the within and between-region spread model for the Nord-Est NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

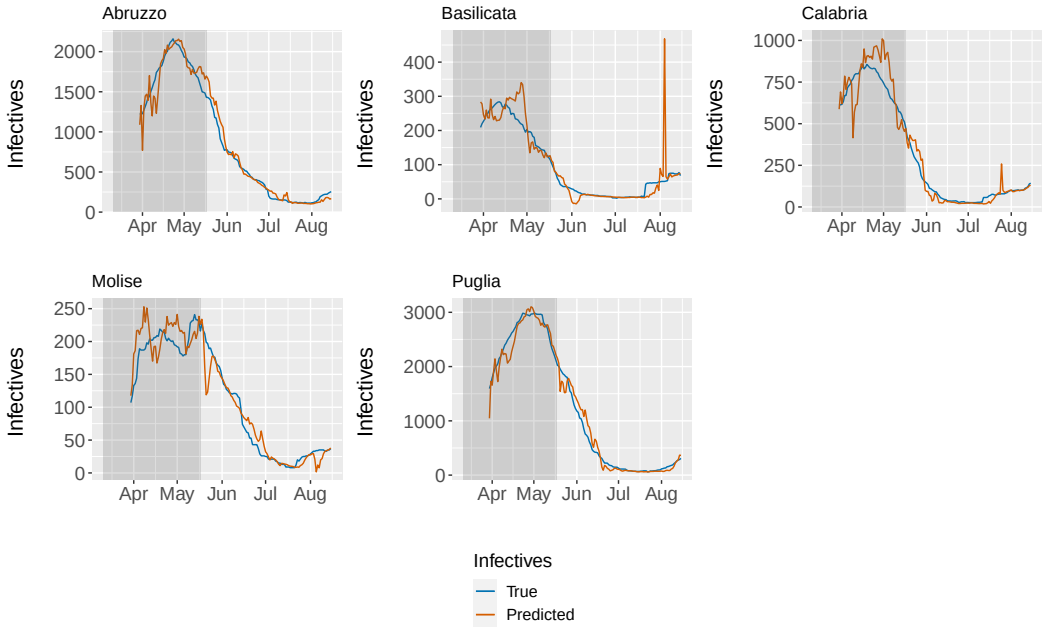


(a) Infectives exclude undocumented cases

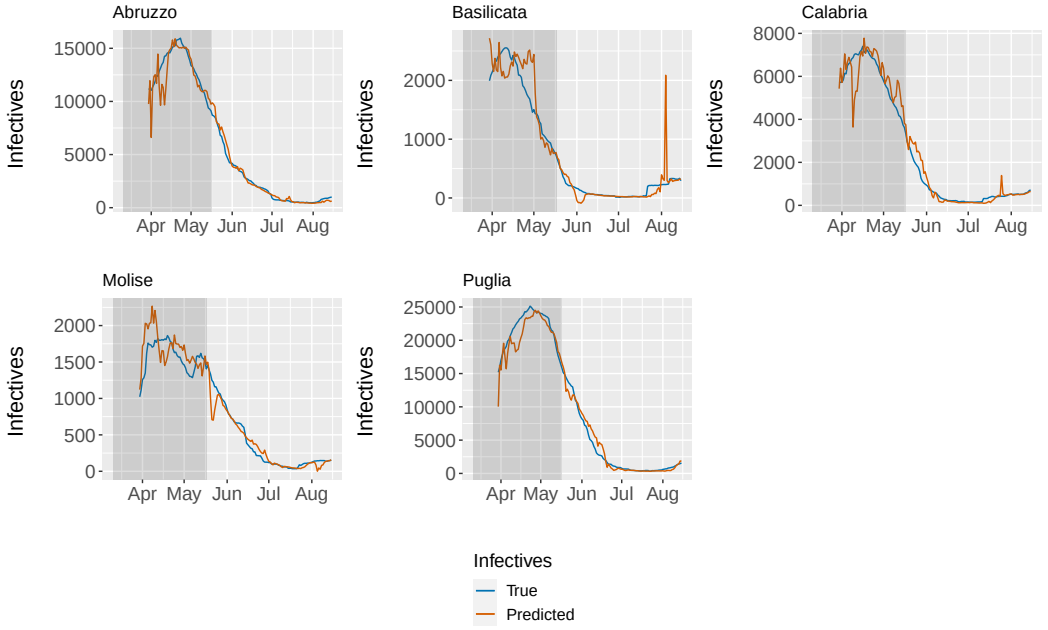


(b) Infectives include undocumented cases

Figure C.22. One-period ahead forecasts for the within and between-region spread model for the Centro (IT) NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

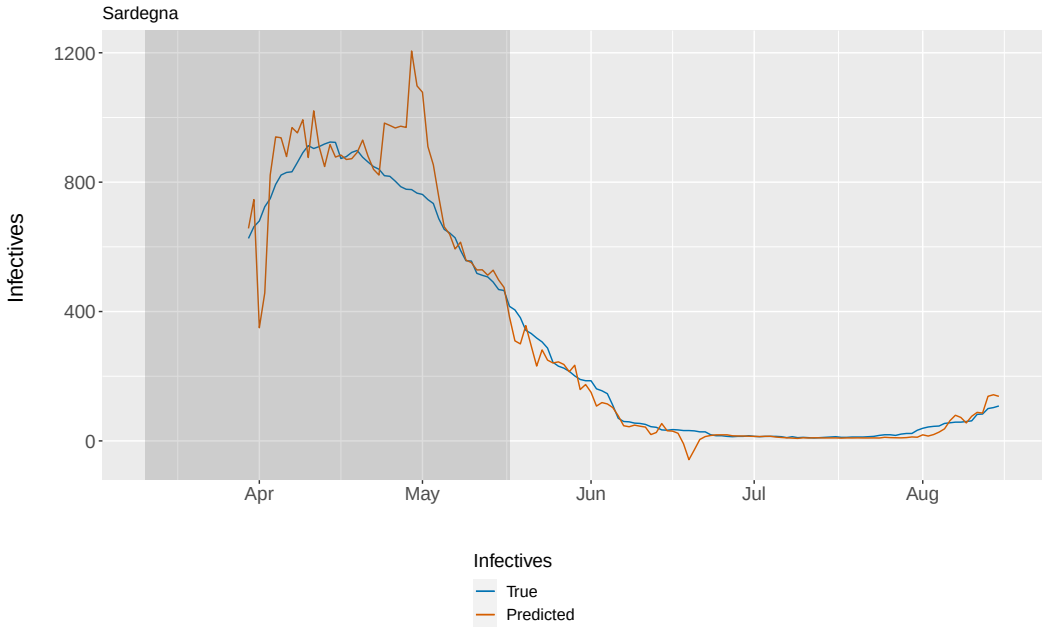


(a) Infectives exclude undocumented cases

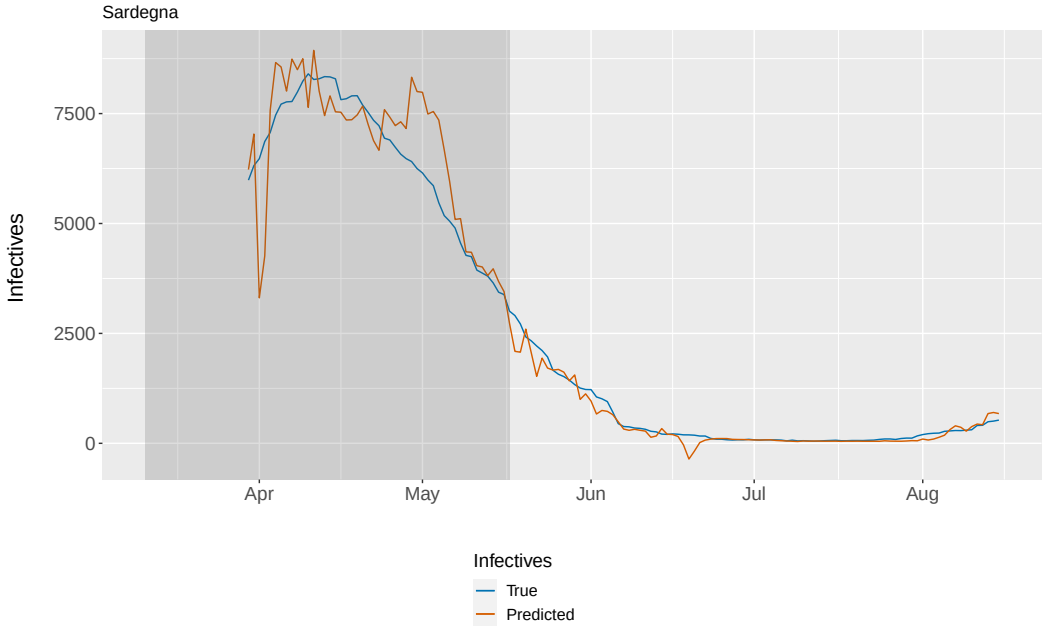


(b) Infectives include undocumented cases

Figure C.23. One-period ahead forecasts for the within and between-region spread model for the Sud NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

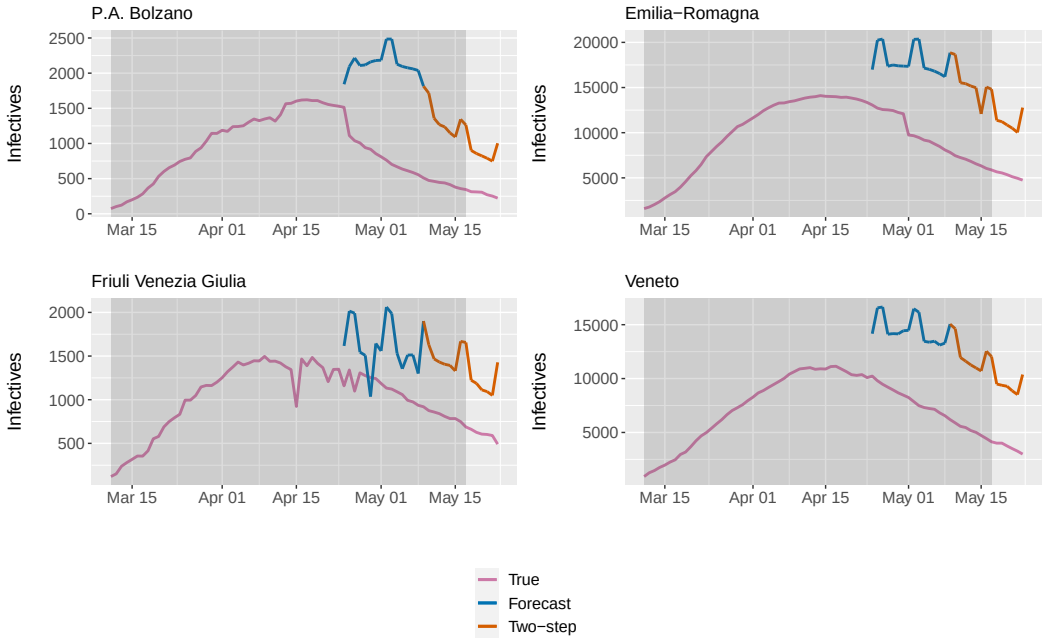


(a) Infectives exclude undocumented cases

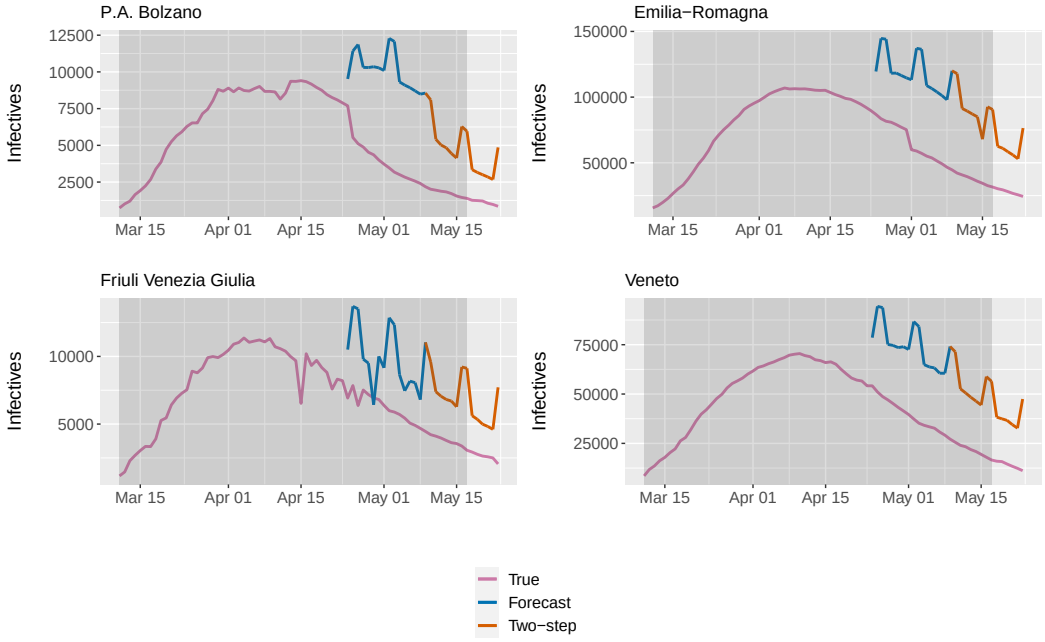


(b) Infectives include undocumented cases

Figure C.24. One-period ahead forecasts for the within and between-region spread model for the Isole NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.



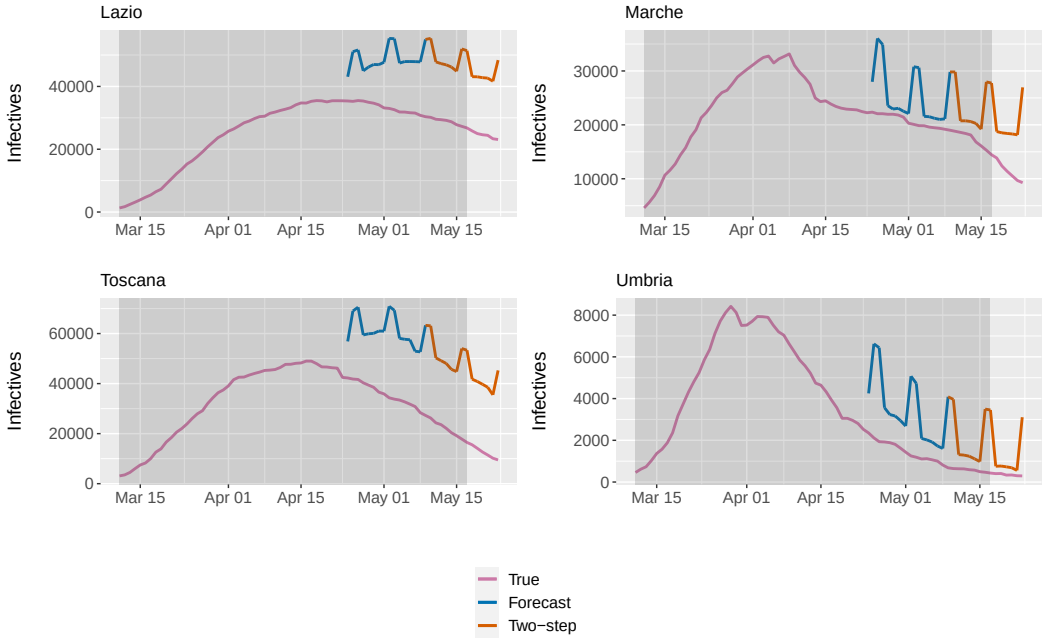
(a) Infectives exclude undocumented cases



(b) Infectives include undocumented cases

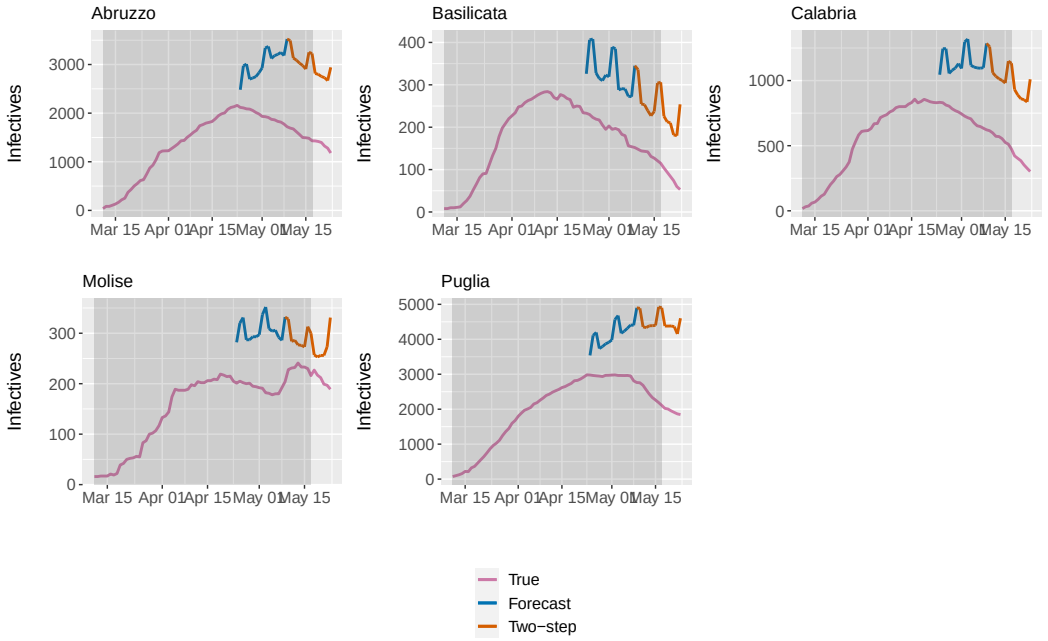
Figure C.25. Multi-period ahead forecasts for the within-region spread model for the Nord-Est NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

(a) Infectives exclude undocumented cases

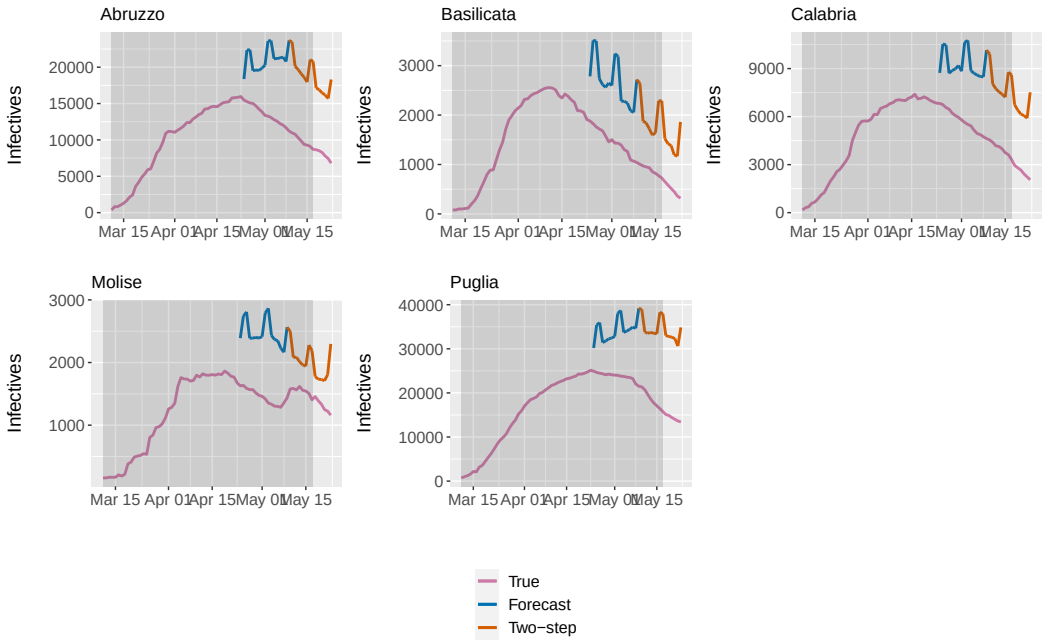


(b) Infectives include undocumented cases

Figure C.26. Multi-period ahead forecasts for the within-region spread model for the Centro (IT) NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

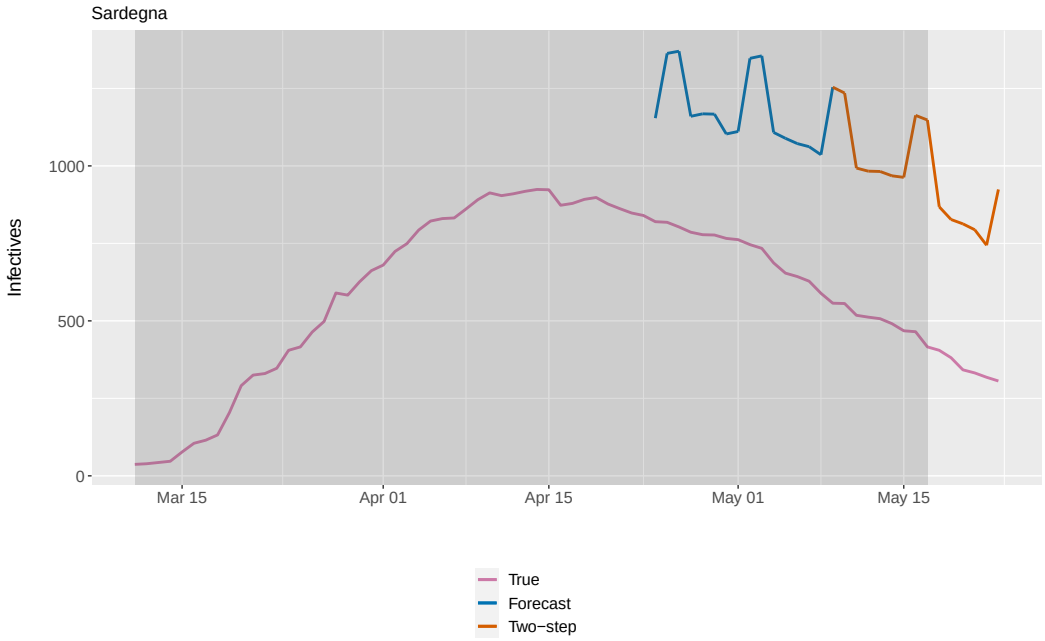


(a) Infectives exclude undocumented cases

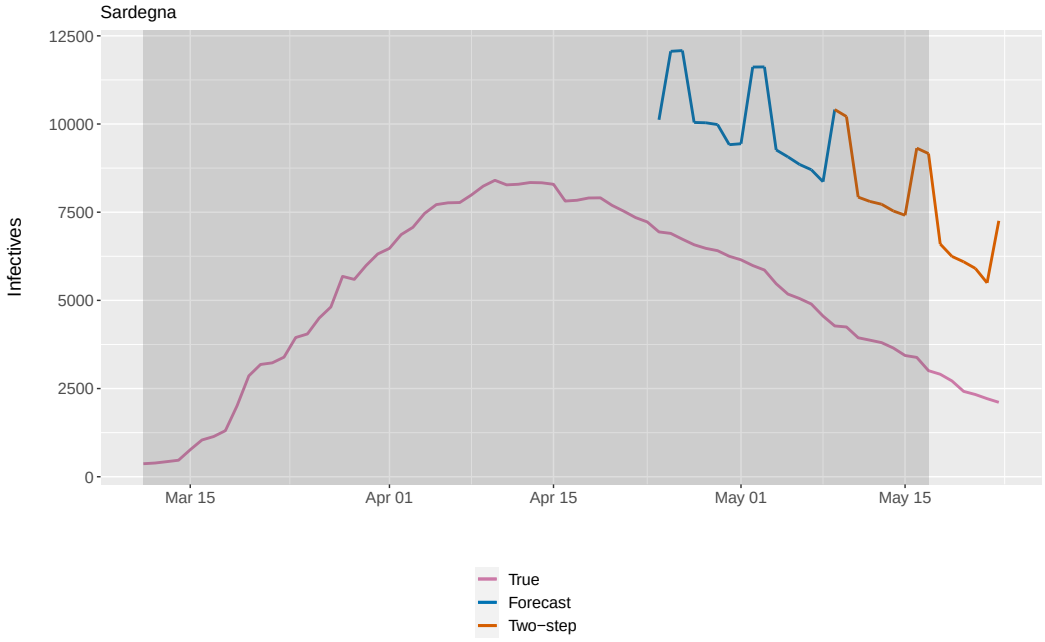


(b) Infectives include undocumented cases

Figure C.27. Multi-period ahead forecasts for the within-region spread model for the Sud NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.



(a) Infectives exclude undocumented cases



(b) Infectives include undocumented cases

Figure C.28. Multi-period ahead forecasts for the within-region spread model for the Isole NUTS 1 region. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$. The grey area indicates the national lockdown.

D Discrete SIR Model

In Section 10, we referenced an approach that aims 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. In particular, the approach aims 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 include the addition by Adda (2016) of a longer lag to take the incubation period into account. This leads to the following discretized equations, using frequency-dependent transmission:

$$s_{p,t} - s_{p,t-1} = -\beta s_{p,t-\tau} i_{p,t-\tau} + \eta_{p,t}, \tag{D.1}$$
$$i_{p,t} - i_{p,t-1} = \beta s_{p,t-\tau} i_{p,t-\tau} - \gamma i_{p,t-1} + \eta_{p,t}, \tag{D.2}$$
$$r_{p,t} - r_{p,t-1} = \gamma i_{p,t-1} + \eta_{p,t}. \tag{D.3}$$

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

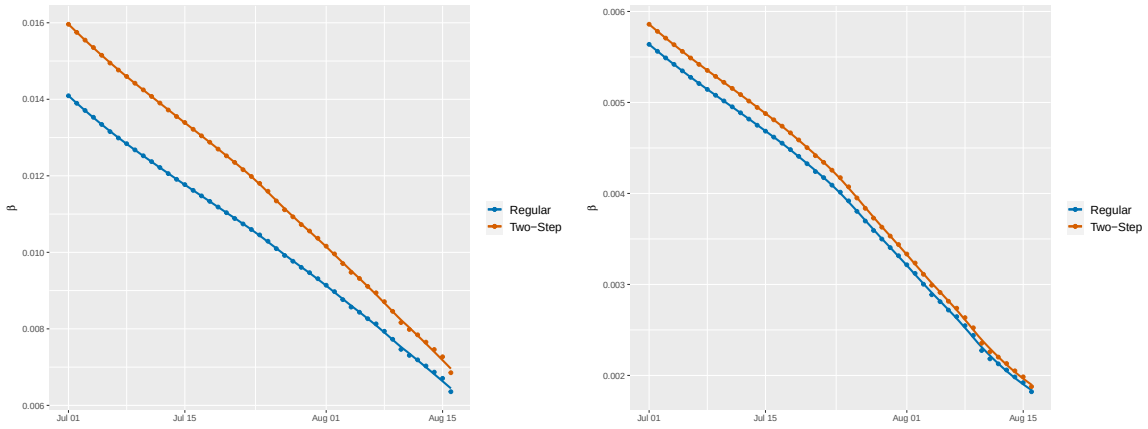
Significance levels: * = 0.1 ** = 0.05, *** = 0.01

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 is not possible. As such, the random effects model is not applicable to this situation. This also becomes clear

when considering the results on R_{eff} for the random effects model, as they are too low to be realistic. For the pooled OLS model, it is interesting to compare the values of R_{eff} when modelling undocumented infectives. As expected, the values of R_{eff} are larger when undocumented infectives are included.

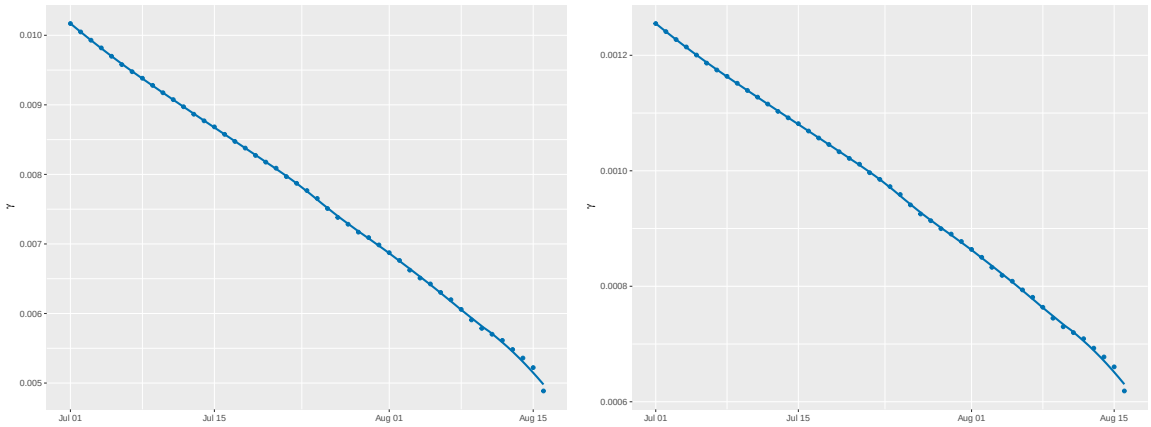
Unfortunately, all estimates are quite low. Consider the estimate of $\gamma = 4.886 \times 10^{-3}$. This implies that the average infectious period is $\gamma^{-1} = 204.67$ days, which is not credible. This means that the estimates are not individually interpretable. However, it may be possible that the resulting value of R_{eff} is still 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.



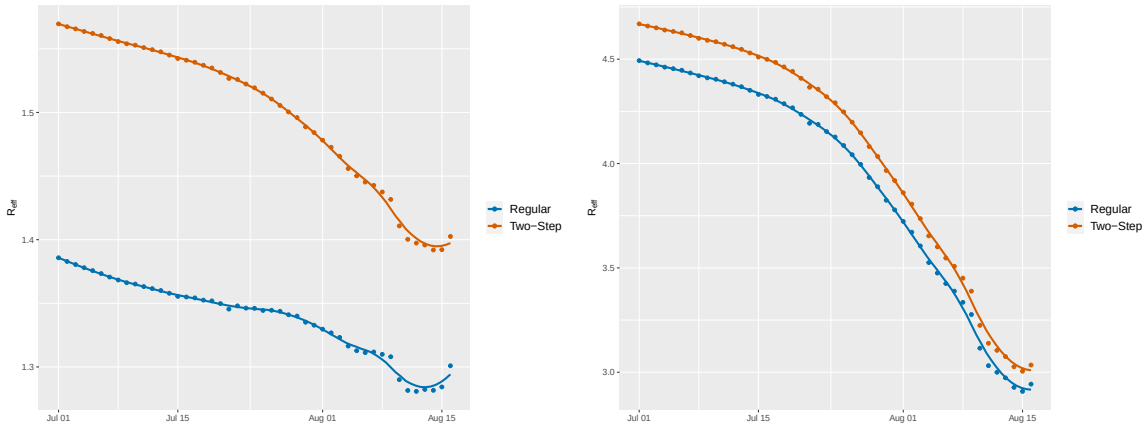
(a) Infectives exclude undocumented cases **(b)** Infectives include undocumented cases

Figure D.1. Progression of the transmission rate β over time. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



(a) Infectives exclude undocumented cases **(b)** Infectives include undocumented cases

Figure D.2. Progression of the recovery rate γ over time. Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.



(a) Infectives exclude undocumented cases **(b)** Infectives include undocumented cases

Figure D.3. Progression of the effective reproduction number R_{eff} over time. 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 to 4.5 (for the regular β). On the other hand, if undocumented infectives are not modelled, it only ranges from 1.3 to 1.4. The values for β and R_{eff} decrease over time, indicating that the pandemic became less severe. 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 to 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 5.

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 $R_{p,t}$ 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 5.

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) $f(S_t^{max}) = 1$

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:

$$aS_t^{max} + f^{min} = 1.$$

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

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

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) $f(S_t^{max}) = 1$,

(IV) $f(\beta S_t^{max}) = \gamma$ for $\beta, \gamma \in (0, 1)$,

(V) The vertex of the parabola should be to the right of S_t^{max} 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 S_t^{max}. The set of equations that we need to solve are:

$$\begin{cases} a\left(S_t^{max}\right)^2 + bS_t^{max} + f^{min} & = 1 \text{ (from assumption (III))} \\ a\beta^2\left(S_t^{max}\right)^2 + b\beta S_t^{max} + f^{min} & = \gamma \text{ (from assumption (IV))} \end{cases} \tag{E.3}$$

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

$$\begin{pmatrix} \left(S_t^{max}\right)^2 & S_t^{max} & | & 1 - f^{min} \\ \beta^2\left(S_t^{max}\right)^2 & \beta S_t^{max} & | & \gamma - f^{min} \end{pmatrix}$$

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

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

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

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

As such, we have derived that:

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

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:

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

where we use that $\beta(1-\beta)\left(S_t^{max}\right)^2 > 0$. 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:

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

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

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

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

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

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

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

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

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

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, condition (E.6a) tells us

that both the numerator and denominator of the fraction are negative. Therefore, to satisfy condition (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:

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

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

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

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 condition (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 $a = \frac{2 - 4\gamma + 2f^{min}}{(S_t^{max})^2} = 0$.

Conclusively, we have derived that:

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

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) $f(S_t^{max}) = 1$,

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

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), $h = S_t^{max}$ and $k = 1$. Therefore:

$$f(TC_t) = a(TC_t - S_t^{max})^2 + 1.$$

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

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

Therefore, the formula becomes:

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

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) $f(S_t^{max}) = 1$,

(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:

$$\begin{aligned} a\left(S_t^{max}\right)^2 + f^{min} &= 1 \\ \iff a &= \frac{1 - f^{min}}{\left(S_t^{max}\right)^2} \end{aligned}$$

Therefore, the formula becomes:

$$f(TC_t) = \frac{1 - f^{min}}{\left(S_t^{max}\right)^2} TC_t^2 + f^{min},$$

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) $f(S_t^{max}) = 1$,

(IV) $f(\beta_1 S_t^{max}) = \gamma_1$ and $f(\beta_2 S_t^{max}) = \gamma_2$ 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 S_t^{max}. The set of equations that we need to solve are:

$$\begin{cases} a\left(S_t^{max}\right)^3 + b\left(S_t^{max}\right)^2 + cS_t^{max} + f^{min} & = 1 \text{ (from assumption (III))} \\ a\beta_1^3\left(S_t^{max}\right)^3 + b\beta_1^2\left(S_t^{max}\right)^2 + c\beta_1 S_t^{max} + f^{min} & = \gamma_1 \text{ (from assumption (IV))} \\ a\beta_2^3\left(S_t^{max}\right)^3 + b\beta_2^2\left(S_t^{max}\right)^2 + c\beta_2 S_t^{max} + f^{min} & = \gamma_2 \text{ (from assumption (IV))} \end{cases} \tag{E.8}$$

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

$$\left(\begin{array}{ccc|c} (S_t^{max})^3 & (S_t^{max})^2 & S_t^{max} & 1 - f^{min} \\ \beta_1^3 (S_t^{max})^3 & \beta_1^2 (S_t^{max})^2 & \beta_1 S_t^{max} & \gamma_1 - f^{min} \\ \beta_2^3 (S_t^{max})^3 & \beta_2^2 (S_t^{max})^2 & \beta_2 S_t^{max} & \gamma_2 - f^{min} \end{array}\right)$$

$$= \left(\begin{array}{ccc|c} (S_t^{max})^3 & (S_t^{max})^2 & S_t^{max} & 1 - f^{min} \\ \frac{1}{64} (S_t^{max})^3 & \frac{1}{16} (S_t^{max})^2 & \frac{1}{4} S_t^{max} & \gamma_1 - f^{min} \\ \frac{1}{8} (S_t^{max})^3 & \frac{1}{4} (S_t^{max})^2 & \frac{1}{2} S_t^{max} & \gamma_2 - f^{min} \end{array}\right)$$

$$\xrightarrow{r_2 \times 64} \xrightarrow{r_3 \times 8} \left(\begin{array}{ccc|c} (S_t^{max})^3 & (S_t^{max})^2 & S_t^{max} & 1 - f^{min} \\ (S_t^{max})^3 & 4 (S_t^{max})^2 & 16 S_t^{max} & 64\gamma_1 - 64 f^{min} \\ (S_t^{max})^3 & 2 (S_t^{max})^2 & 4 S_t^{max} & 16\gamma_2 - 64 f^{min} \end{array}\right)$$

$$\xrightarrow{r_2 - r_1} \xrightarrow{r_3 - r_1} \left(\begin{array}{ccc|c} (S_t^{max})^3 & (S_t^{max})^2 & S_t^{max} & 1 - f^{min} \\ 0 & 3 (S_t^{max})^2 & 15 S_t^{max} & -1 + 64\gamma_1 - 63 f^{min} \\ 0 & (S_t^{max})^2 & 3 S_t^{max} & -1 + 8\gamma_2 - 7 f^{min} \end{array}\right)$$

$$\xrightarrow{r_2 \leftrightarrow r_3} \left(\begin{array}{ccc|c} (S_t^{max})^3 & (S_t^{max})^2 & S_t^{max} & 1 - f^{min} \\ 0 & (S_t^{max})^2 & 3 S_t^{max} & -1 + 8\gamma_2 - 7 f^{min} \\ 0 & 3 (S_t^{max})^2 & 15 S_t^{max} & -1 + 64\gamma_1 - 63 f^{min} \end{array}\right)$$

$$\xrightarrow{r_1 - r_2} \xrightarrow{r_3 - 3r_2} \left(\begin{array}{ccc|c} (S_t^{max})^3 & 0 & -2 S_t^{max} & 2 - 8\gamma_2 + 6 f^{min} \\ 0 & (S_t^{max})^2 & 3 S_t^{max} & -1 + 8\gamma_2 \\ 0 & 0 & 6 S_t^{max} & 2 + 64\gamma_1 - 24\gamma_2 - 42 f^{min} \end{array}\right)$$

$$\xrightarrow{r_1 + \frac{1}{3} r_3} \xrightarrow{r_2 - \frac{1}{2} r_3} \left(\begin{array}{ccc|c} (S_t^{max})^3 & 0 & 0 & \frac{8 + 64\gamma_1 - 48\gamma_2 - 24 f^{min}}{3} \\ 0 & (S_t^{max})^2 & 0 & -2 - 32\gamma_1 + 20\gamma_2 + 14 f^{min} \\ 0 & 0 & 6 S_t^{max} & 2 + 64\gamma_1 - 24\gamma_2 - 42 f^{min} \end{array}\right)$$

$$\xrightarrow{r_1 \div (S_t^{max})^3} \xrightarrow{r_2 \div (S_t^{max})^2} \xrightarrow{r_3 \div 6 S_t^{max}} \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{8 + 64\gamma_1 - 48\gamma_2 - 24 f^{min}}{3 (S_t^{max})^3} \\ 0 & 1 & 0 & \frac{-2 - 32\gamma_1 + 20\gamma_2 + 14 f^{min}}{(S_t^{max})^2} \\ 0 & 0 & 1 & \frac{2 + 64\gamma_1 - 24\gamma_2 - 42 f^{min}}{6 S_t^{max}} \end{array}\right)$$

Conclusively, we have derived that:

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

so that:

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

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