



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

by
Mike Weltevrede (ANR 756479)

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

Tilburg School of Economics and Management
Tilburg University

Supervised by:
dr. Otilia Boldea

Second reader:
dr. George Knox

Date:
September 15, 2020

Abstract

TODO

Acknowledgements

TODO

Contents

1	Introduction	1
2	Problem description	2
3	Dataset	6
3.1	Geographical structure of Italy	6
3.2	Coronavirus data	6
3.3	Independent variables	10
4	SIR model	12
5	Model selection	14
6	Modelling undocumented infectives	15
7	Within-Region Spread Model	23
7.1	Methodology	23
7.2	Results	27
8	Within and Between-Region Spread Model	35
8.1	Methodology	35
8.2	Results	38
9	Conclusion	45
10	Future research	46
	Appendices	51
A	Abbreviations	51
B	Tables	52
B.1	Results for Section 7: Within-Region Spread Model	52
B.2	Results for Section 8: Within and Between-Region Spread Model . . .	54
C	Figures	57
C.1	Figures for Section 2: Problem description	57
C.2	Figures for Section 3: Dataset	61
C.3	Figures for Section 7: Within-Region Spread Model	64
C.4	Figures for Section 8: Within and Between-Region Spread Model . . .	73

D	Derivations	80
D.1	Calculation of population variables	80
D.2	Functional forms for modelling undocumented infectives	80
D.2.1	Linear function	81
D.2.2	General quadratic function	81
D.2.3	Special case quadratic formula: downwards opening	84
D.2.4	Special case quadratic formula: upwards opening	85
D.2.5	Cubic function	86

1 Introduction

Since the beginning of 2020, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has plagued the world. Starting from Wuhan, China, it has made its way to every single continent apart from Antarctica and (nearly) every country in the world. Only 12 sovereign member states of the United Nations reported no infections, of which 10 are island countries. The other two countries are North Korea and Turkmenistan but it is suspected that there are actually cases but that these are not reported for both North Korea (Business Insider, 2020; Nebehay, 2020) and Turkmenistan (Human Rights Watch, 2020; Mackinnon, 2020). In response to SARS-CoV-2, governments have been implementing far-reaching measures to try and contain the virus, such as shutting down schools and restaurants, but also by locking down the entire country.

On August 19, 2020, over 21 million people were reported to have been infected with the infectious respiratory disease caused by SARS-CoV-2, called coronavirus disease 2019 (commonly abbreviated to COVID-19), leading to 771 thousand consequent deaths, while almost 14 million people have already recovered from COVID-19. Therefore, almost 36 million people worldwide have been infected by the virus. On top of this, due to the inherent problem of a limited testing capacity, there are most likely many infections that went and still are going undocumented, meaning that the scope of the problem is much larger than the numbers just mentioned. The World Health Organization (WHO) declared a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 (WHO, 2020a), defined as “*an extraordinary event which is determined to constitute a public health risk to other States through the international spread of disease and to potentially require a coordinated international response*” (WHO, 2019). After the spread of SARS-CoV-2 only became worse, the WHO declared the virus outbreak to be a pandemic on 11 March 2020 (WHO, 2020b), where a pandemic is defined as “*an epidemic occurring on worldwide or over a very wide area, crossing international boundaries, and usually affecting a large number of people*” (Porta, 2014).

2 Problem description

In this section, we elaborate on the problem at hand, namely the epidemiological spread of SARS-CoV-2 and the disease it causes, specifically for the country of Italy. We focus on the additions of this thesis to the existing literature.

Italy has been one of the most intensely struck countries by COVID-19. Until the end of March, it had the highest number of confirmed cases per 100,000 inhabitants. It was subsequently taken over by Spain. Italy remained the second most struck country until May 1, when the United States took over. On July 3, 2020, it had the ninth highest absolute number of confirmed cases, after the United States, Brazil, Russia, India, Peru, Chile, the United Kingdom, and Spain. Italy reported the second highest global death-to-cases ratio of 14.45% (34,818 deaths to 240,961 cases), only after the United Kingdom, which reports a death-to-cases ratio of 15.50% (43,995 deaths to 283,757 cases). The third highest death-to-cases ratio of 12.24% (29,189 deaths to 238,511 cases) was reported by Mexico. The sudden onset of the spread of SARS-CoV-2 put immense pressure on the Italian hospitals, especially in the northern regions such as Lombardy. This forced patients with coronavirus-caused pneumonia to be sent home as well as literal collapses of overworked healthcare workers (Horowitz, 2020).

Due to the extreme nature of the pandemic in Italy and the availability of enough data, this thesis chooses to focus on Italy. Our contribution is to model the pandemic at the regional level rather than at the national level. Even though the regions within Italy are likely more similar to one another than, for instance, to the provinces in the Netherlands, there are large regional differences in Italy as well. This implies that there is not one model that can be used to analyze the entire country of Italy because this would ignore the regional heterogeneous effects. Besides doing so to uncover heterogeneity across regions, the regional variation allows for better identification of the average transmission parameters. To illustrate these regional differences, we will present several figures. Firstly, consider Figure 2.1, which shows the incidence rates categorized by the overarching region that the regions are a part of (called NUTS 1 regions). This regional structure is explained more thoroughly in Section 3.1.

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

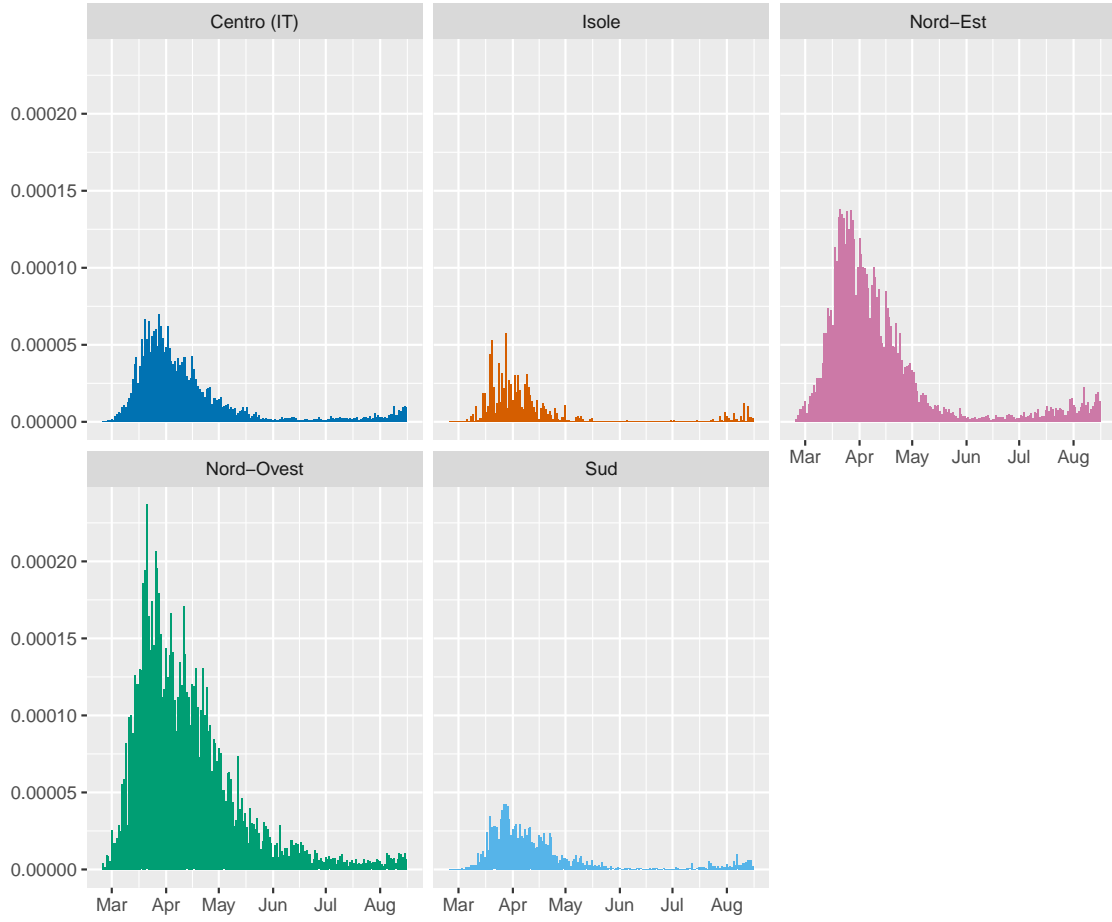


Figure 2.1. Incidence rate per NUTS 1 region

Consider Figure 2.2, where we zoom in on the Centro (IT) NUTS 1 region by looking at the four regions that make it up. These lower level regions are called the NUTS 2 regions. The plots for the other NUTS 1 regions can be found in Appendix C.1. Figure 2.2 shows us that, even among the regions in the NUTS 1 region, there is a vast difference. For the region of Lazio, we see a much lower peak than for the other three regions, especially compared to Marche. Moreover, the varying length of the peak is more visible in Figure 2.2. For instance, compare the regions of Marche and Umbria.

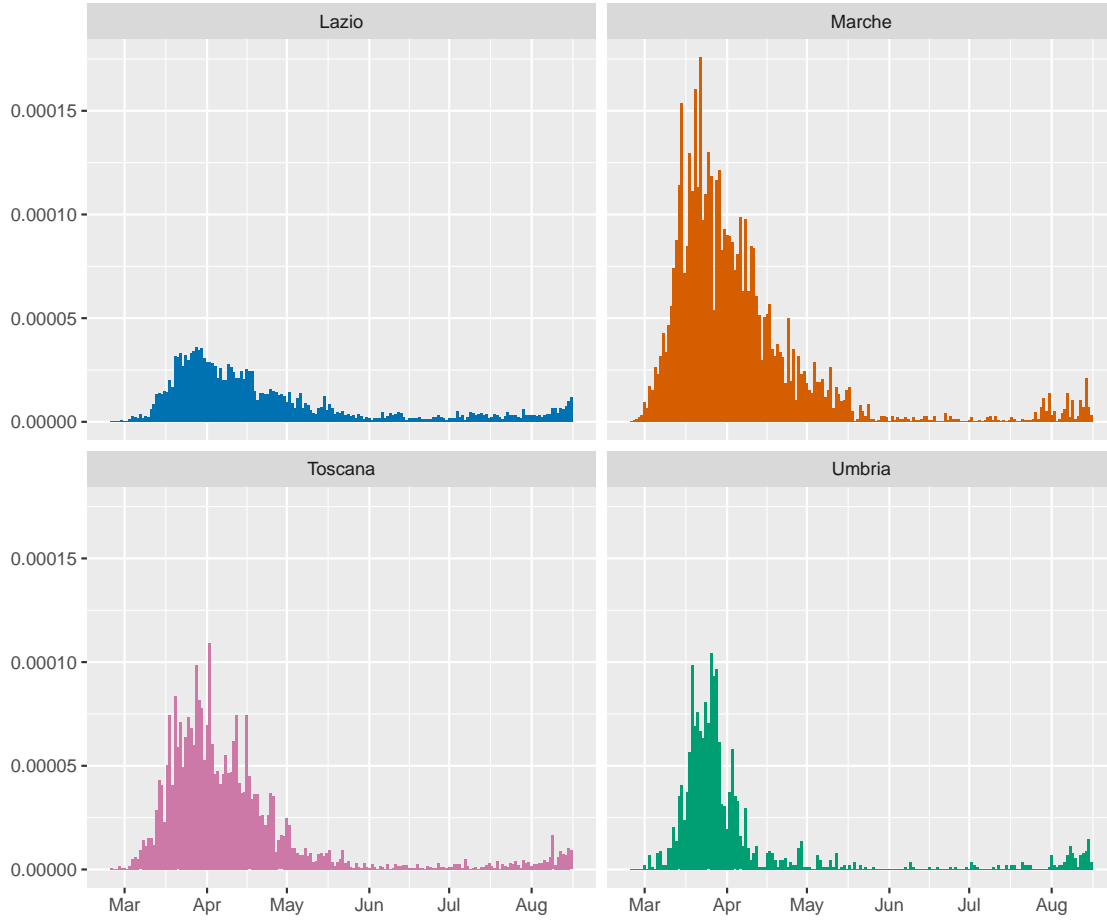


Figure 2.2. Incidence rate per region for the *Centro (IT)* (Centre) NUTS 1 region

We can also investigate the regional heterogeneity by looking at the mean incidence rate over the regions. We consider the mean incidence rate because this allows us to compare regions, even when these have a higher population county. Consider Figure 2.3, which shows the mean incidence rate for the NUTS 2 regions, divided up into their respective NUTS 1 regions.

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

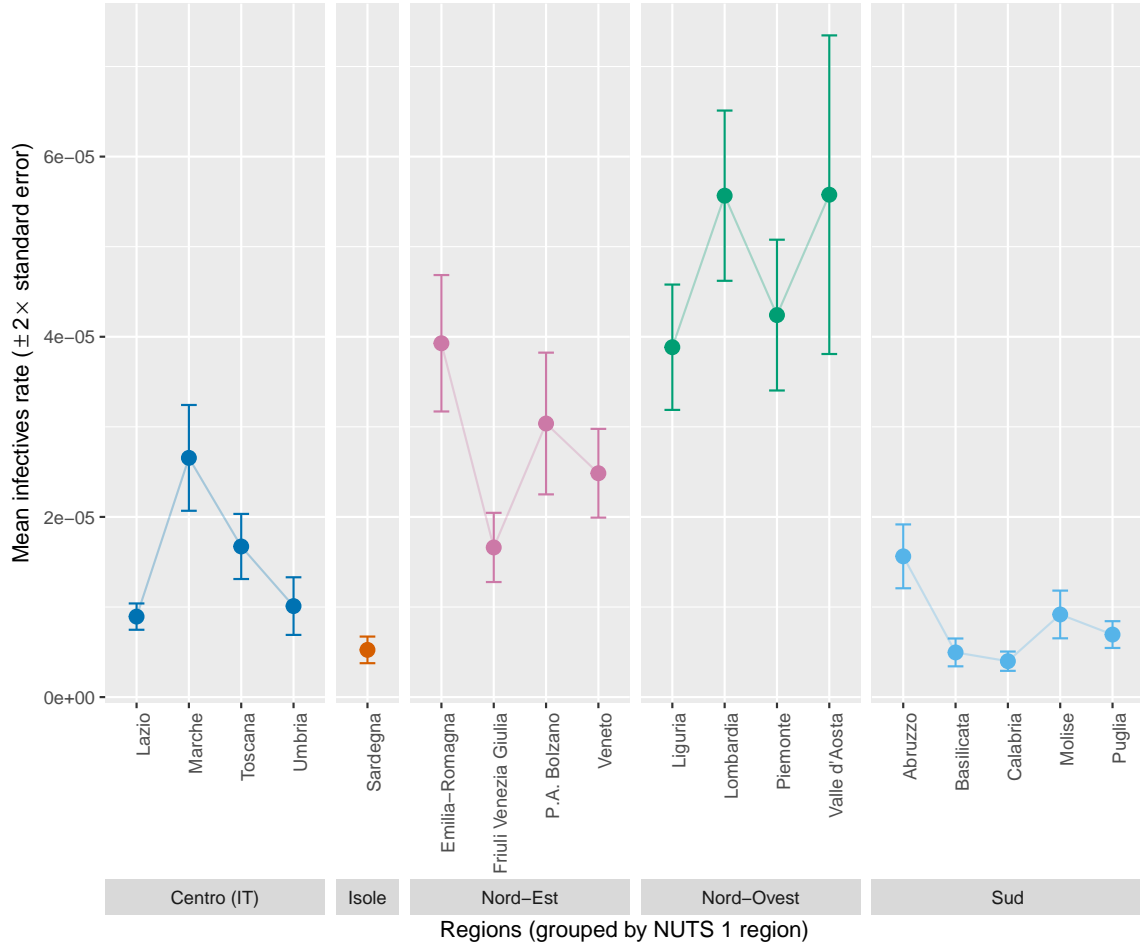


Figure 2.3. Heterogeneity across Italian regions

This shows us that the heterogeneity between regions can likely not be ignored.

In this thesis, we present several models. We are basing the models on specifications as used by Adda (2016). In the paper, Adda (2016) investigates the spread of several viral diseases in the past, namely for influenza, gastroenteritis, and chickenpox. The models used to research this spread are inspired by the Standard Inflammatory Response (SIR) model but deviate from the SIR model in the sense that the right-hand side variables include the number of new cases rather than the absolute number of cases. The key additions made by Adda (2016) are, firstly, that a spatial spillover effect is considered and, secondly, that some sort of weighting on the parameters is allowed on the basis of region specific variables. With this motivation, Adda (2016) defines three models comprising of a model ignoring interaction between regions, a model taking interaction between regions into account, and a model that expands on the latter by introducing the weights. Unfortunately, good weighting variables regard-

ing SARS-CoV-2 are not available due to the temporal limitations of the data. Adda (2016) looks at viruses that have been appearing in society for several years and can, therefore, use weekly information and relevant instruments to quantify the infection rates, such as economic indices. Given that SARS-CoV-2 has only been appearing for around half a year, this information is not available. If this information would be available, one could also estimate a fourth model that Adda (2016) does not discuss, namely an intermediate model that ignores interregional dependence but where weights are included. However, this thesis only discusses the non-weighted models from Adda (2016). These models have not previously been applied to SARS-CoV-2 and can possibly show interesting insights compared to other models.

3 Dataset

In this section, we outline the structure of the data that is used and how it was retrieved. Firstly, we discuss the structure of Italian regions in Section 3.1. Subsequently, we look at the data on COVID-19 such as the incidence rate, reported deaths, and number of recoveries in Section 3.2. Here, we also discuss how possible errors and missing values in the data are handled. Lastly, Section 3.3 discusses the variables that are included in the models by Adda (2016).

3.1 Geographical structure of Italy

In this section, we discuss the structure of Italian regions. Before that, we should understand 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 (EU) and the United Kingdom (Eurostat, 2020a). Italy consists of 21 so-called *regioni* (regions), comparable to Dutch provinces. These constitute the second-level NUTS regions (also called NUTS 2 regions). 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*. Italy’s first-level NUTS regions are defined as groups of regions, namely *Nord-Ovest* (North-West), *Nord-Est* (North-East), *Centro* (Center), *Sud* (South), and *Isole* (Islands). The third-level NUTS regions are 107 provinces, which are administrative subregions of the *regioni*.

3.2 Coronavirus data

In this section, we discuss the data on COVID-19 and how we handled the data processing. The *Presidenza del Consiglio dei Ministri - Dipartimento della Protezione Civile* (Presidency of the Council of Ministers - Department of Civil Protection), hereafter referred to as the Department of Civil Protection, has posted daily reports containing tables with a detailed numerical overview of new cases, active intensive

care (IC) patients, tests executed, and more (Rosini, 2020). This data is divided up between the NUTS 2 regions. Ideally, we would want to have coronavirus data on the NUTS 3 regions since many policies are introduced at that level, such as a lockdown put into place on March 7, 2020 until the strict national lockdown was instated. 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 (*terapia_intensiva*) and the number of hospitalized people who showed symptoms (*ricoverati_con_sintomi*).¹

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

There are two variables on the tests executed. The semantic difference between the total amount of tests performed (*tamponi*) and the total amount of people tested (*casi_testati*) is that the latter indicates the number of unique persons that were tested because individuals could have been tested more than once. Notice that *tamponi* is a good indication of the testing capacity as the number of tests that Italy is able to

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

execute. Henceforth, when the term *testing capacity* is used, this refers to *tamponi*, unless indicated otherwise. In addition to the previous remarks, it is important to consider that there is a measurement error in the number of infectives, as is the case in any other country. This is because there is no possibility that every citizen can be tested for COVID-19. For that reason, the actual number of infectives is higher than the official count as reported in the tables of the Department of Civil Protection.

With respect to the reported death statistics, there is a distinction between Italy and some other European countries. Namely, the Italian numbers include deaths of all patients who were tested positive for COVID-19 before or after their death, regardless of whether they died inside or outside the hospital, assuming that these deaths were reported. In contrast, other countries may only count deaths in hospitals. Belgian death counts, for instance, also include deaths of people who were suspected of having COVID-19, regardless of whether they were tested or not (Schultz, 2020).

Moreover, Italian data makes no distinction between people who died because of COVID-19 and those who simply had the disease but who died from other causes (also referred to as comorbidities). Actually, only 1.2% of the Italian patients who were reported to have died because of COVID-19 until March 19, 2020 did not have a pre-existing condition (European Centre for Disease Prevention and Control, 2020). Of the patients that died and did have at least one comorbidity, 48.6% had three or more comorbidities, 26.6% had two comorbidities, and 23.5% had one comorbidity. European Centre for Disease Prevention and Control (2020) also reports that 73.8% of the deceased patients had hypertension, 33.9% diabetes, 30.1% ischaemic heart disease, 22.0% atrial fibrillation, and 19.5% had a cancer diagnosed in the last five years. As such, it may be the case that a patient died from, for instance, hypertension but because they were infected by SARS-CoV-2 their death was classified as a COVID-19 death instead. Some other countries, such as Germany, do make a distinction between these two groups (Caccia, 2020). In the UK, there is a radical difference between the total number of deaths until June 28 with a positive test result (43,575 deaths), the total number of deaths until June 19 where COVID-19 is mentioned on the death certificate (53,858 deaths), and the total number of deaths until June 19 over and above the usual number at that time of the year (65,132 deaths) (BBC News, 2020). This shows that the UK reports deaths due to COVID-19 on the death certificates even for people who were not tested positive. Moreover, there are many excess deaths over the usual number that may or may not be due to COVID-19 that are now not counted in the official reports. In this thesis, we assume that this error is negligible and that this differing method of counting deaths and cases only applies on a national level and not among a country's regions.

Sometimes, regions correct mistakes by having the numbers in the report on the next day compensate for the error on the day before. In the official publications that

we use, data that was wrongly published on a day $t - 1$ is corrected by subtracting the error from or adding the error to the cases from day t . As such, if the error is larger than the number of new cases, the reported amount of new cases is negative. It happened twenty-two times that the number of confirmed cases was reported to be negative (for eleven different regions). The number of deaths was reported to be negative eight times (for six different regions) and the number of recovered patients was reported with a negative value 62 times (for fourteen different regions). We correct this by subtracting the error from the day before and set the previously negative number to zero. In the case that the error on day t is larger than the number on $t - 1$, for instance if a value of -10 is reported on day t while the value for day $t - 1$ is less than ten, we propagate the error to multiple lags until this issue no longer occurs. An example for the region of Basilicata is given in Table 3.1.

Table 3.1. Example of the propagation of negative values for the region of Basilicata

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

For days where the correction did not cause the number to become negative, we have no way of detecting that a correction took place and we cannot reasonably assume how this number should be split up among day t and $t + 1$. As such, these are left as is. Despite these errors, we assume that the official information is accurate and representative of the region for which it has been reported.

A highly negative value of -229 was reported for the region of Campania on June 12, 2020, whereas the number of new cases in the week before that date only ranges from zero to five. The same applies to Sicily, where a negative value of -394 was reported on June 19, 2020. There, the number of new cases in the week before that date only ranges from zero to two. We assume that this corrects for all errors in the past, not just those close to June 12 and 19. Propagating this error backwards as described before would lead to zero new cases per day for Campania from May 13 until June 12 (31 days) and for Sicily from April 28 until June 19 (53 days). Since we have no reason to know how this error is distributed, we remove the regions of Campania and Sicily from our dataset. Another solution could be to distribute the error according to the daily number of cases relative to the total amount of cases until June 12 for Campania or June 19 for Sicily.

An extreme outlier in the positive direction can be found on June 24 for the region

of Trentino. A value of 387 new infectives was reported even though in the four weeks before, the maximum amount of new infectives was seven. Notably, this value is the highest of all reported values for Trentino, with the second highest value being 172 on March 15 and the third highest value being 154 on April 11. For the same reason as mentioned for the high negative values for Campania and Sicily, we remove the region of Trentino from our dataset. Again, another solution would be to distribute this number across the days prior.

Regarding missing values, there are none. We expect that this is because the Department of Civil Protection imputed the missing values with a value of zero. For instance, on July 5, it was reported that zero tests were executed in the region of Basilicata. On the dates surrounding July 5, however, around 250 tests were executed each day. On July 9, a higher value of 426 was reported. We expect that this is to correct for the reported value of zero of July 5. We could, for instance, distribute the 426 among July 5 and 9. However, in this thesis, we do not deal with these outliers and leave them as is. The reason for this is twofold. Firstly, we do not know if it is actually true that a zero is being used as a filler for a missing value. It may be the case that a value of zero was actually reported. That relates to the second reason, namely that unexpectedly low values unequal to zero are also reported (such as a value of three tests being executed on July 19 for Basilicata among a usual value of around 300). As such, we cannot reasonably assume that these zeros (and which ones) pertain to missing values.

3.3 Independent variables

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

The reason behind including the weekend dummy variable is that we expect that less people may be detected on the weekend due to some general practitioner practices or testing locations being closed on the weekend, meaning that people who are not willing or able to travel far will not get tested. These people will then get tested

during the week, meaning that we expect that the number of infectives during week-ends 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.

To investigate this conjecture, we plot the mean incidence rate per NUTS 2 region per day of the week in Figure 3.1. For brevity's sake, we display the plot for the Centro (IT) NUTS 1 region. Plots for the other NUTS 1 regions are included in Appendix C.2, which all show a similar result.

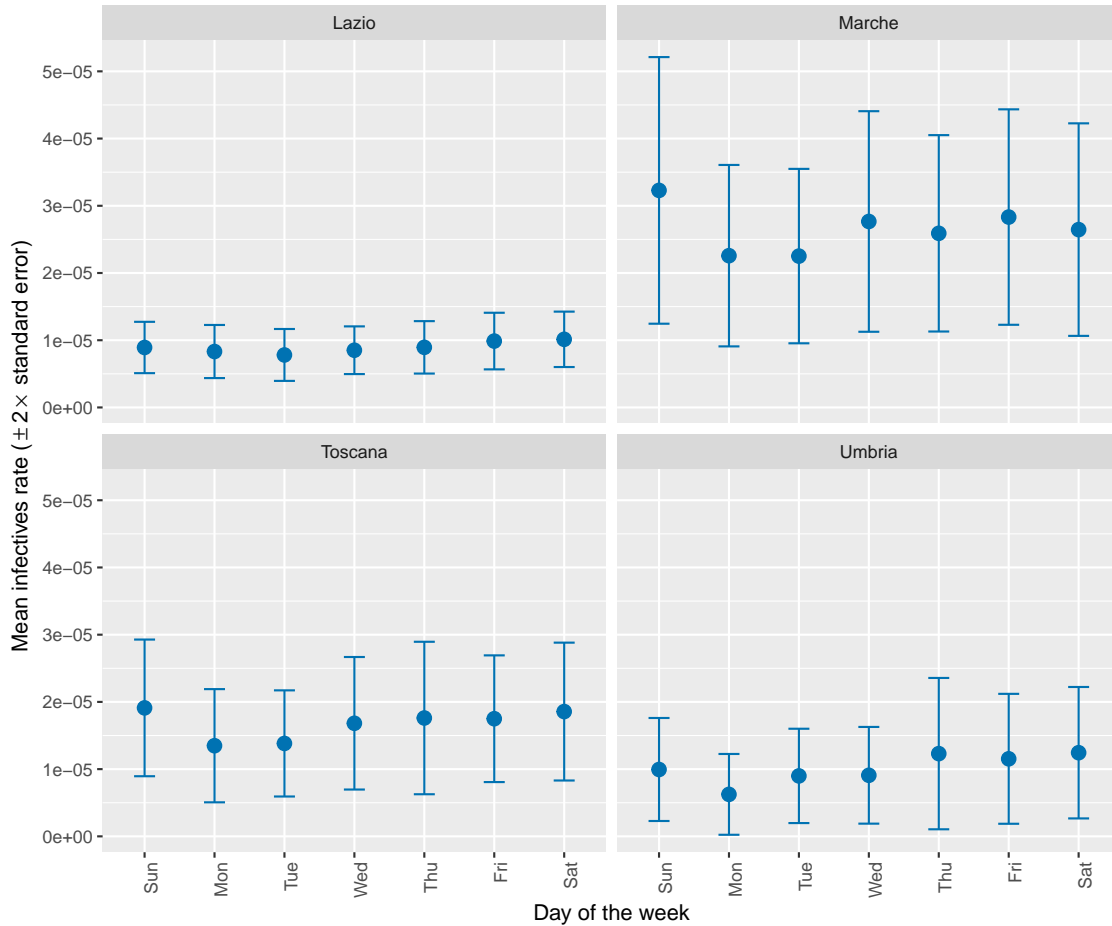


Figure 3.1. Incidence rate per NUTS 2 region per day of the week for the *Centro (IT)* (Centre) NUTS 1 region

Figure 3.1 shows us that, although the mean incidence rate for Sunday is usually

the highest, there is no statistical evidence to infer that there is a higher incidence rate on certain days. This can be concluded because the given error bars of two times the standard deviation overlap. As such, our intuition is likely not as strong on this subject. Nonetheless, we include the variable in the model because an inspection in isolation is not necessarily an indication that the variable is not predictive for the intended dependent variable.

4 SIR model

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

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

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

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

Finally, we assume that there is a constant rate of change ω at which immune individuals lose their immunity. For instance, Adda (2016) mentions that ω is set to 0 for chickenpox as individuals acquire a lifetime immunity while ω will be high for gastroenteritis due to almost no immunity emerging. In the case of COVID-19, some studies show that it is likely that individuals who recovered from COVID-19 may be immune to reinfection, at least temporarily (Kirkcaldy et al., 2020). This can be challenged because it is currently still unknown whether immunity is always

achieved, especially among those who have had only light to medium symptoms. Leung (2020) estimates that COVID-19 antibodies will remain in a patient's system for two to three years, based on what is known about other coronaviruses, indicating a possibility of temporary immunity. However, it has recently become clear that reinfection is indeed possible. Two Chinese persons have been tested positive in August after having recovered from COVID-19 a few months prior (Bloomberg News, 2020). Future research could be done to incorporate this new information. In this thesis, for simplicity's sake, we assume that lifelong immunity is achieved, or at least long enough to last through the temporal scope of our analysis: we set $\omega = 0$.

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

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

$$\frac{dI}{dt} = \beta SI - \gamma I, \quad (4.2)$$

$$\frac{dR}{dt} = \gamma I - \omega R. \quad (4.3)$$

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

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

$$\frac{di}{dt} = \beta si - \gamma i, \quad (4.5)$$

$$\frac{dr}{dt} = \gamma i - \omega r. \quad (4.6)$$

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

rate, for COVID-19 was below 1 in Italy, at between 0.5 and 0.7 (Ministero della Salute, 2020).

5 Model selection

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

$$AIC = -2\log(ML) + 2k, \quad (5.1)$$

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

$$BIC = -2\log(ML) + k\log(n), \quad (5.2)$$

where n denotes the sample size. Both the AIC and BIC are used as the minimizer in the model selection; the model that is picked by the model selection procedure is the one with the lowest AIC or BIC.

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

included in our candidate set, we use the AIC to perform model selection rather than the BIC.

6 Modelling undocumented infectives

In this section, we discuss how our method for modelling undocumented infectives is constructed. We talk about the assumptions, the formulation, and the empirical impact that this has on the total number of infectives. A common concern with the spread of viruses, especially one spreading as rapidly as SARS-CoV-2, is that there is no possibility to test the entire population for the disease because the testing capacity is simply not there. If this were possible, then all individuals who were tested positive could be isolated and the spread of the virus would be dampened tremendously. However, since this is not possible, there are likely many infectives in society who spread the virus but who are undocumented. In China, around 86% of the infectives went undocumented. These undocumented infectives were estimated to also be contagious, at a level of around 55% of the contagiousness of documented infectives (R. Li et al., 2020).

R. Li et al. (2020) carried out their research during the period from January 10 until January 23, 2020, meaning that there was a lack of major restrictions such as travel bans. The same conditions do not apply to Italy during our research period, as it was under a strict national lockdown. This lockdown was imposed on March 10, 2020. The restrictions were relaxed around May 18, when businesses were allowed to reopen and citizens were allowed free movement within the region they live in, although they were still barred from travelling to other regions unless they had an essential motive (Severgnini, 2020). R. Li et al. (2020) make the important note that their results are indeed highly dependent on the specific situation in the country of interest, for instance due to differences in testing, case definition, and reporting. Nonetheless, this research shows that undocumented infectives should be taken into account. Consequently, this thesis aims to model the undocumented infectives. However, we do not account for the lockdown and similar limiting restrictions in our model. Future research could be done to include these restrictions more robustly.

Note that, by definition, there is no data on the number of undocumented infectives because, otherwise, these cases would indeed be documented. As such, some assumptions need to be made since we cannot apply *supervised learning* methods (being models where there is data on a dependent variable to predict) to determine the number of undocumented infectives. Mainly, we assume that the number of undocumented individuals is decreasing as the testing capacity increases. Similarly, the number of documented individuals increases in the testing capacity. The logic behind this is clear: as more people are tested, more infectives move from being undocumented to being documented.

At a point in time t , we denote the testing capacity by TC_t . In Section 3.2, we explained how a measure of the testing capacity is obtained. The total number of infected people at time t is denoted by s_t . This group can be subdivided into the documented infectives DI_t and the undocumented infectives UI_t such that $DI_t + UI_t = s_t$. Therefore, we can denote the documented and undocumented infectives as proportions of the total number of infected people, at any point in time. As mentioned earlier in this section, this proportion may change over time as the testing capacity increases. This proportion is therefore defined as a function of the testing capacity over time:

$$f_t := f(TC_t), \quad (6.1)$$

such that

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

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

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

- (A1) Since f_t is a proportion, we need to have that $f_t \in [0, 1]$.
- (A2) If no one is tested, we assume that there is a certain minimum proportion of infectives who are documented, denoted by $f^{min} \in [0, 1]$. That means that $f(0) = f^{min}$. It should be noted that, at any point in time, it should hold that

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

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

- (A3) Denote the total population at time t as N_t . Then, if there is enough testing capacity such that the entire population can be tested, we assume that all

infectives will be documented, so that:

$$f(N_t) = 1.$$

This also assumes that the tests that are executed are perfect at determining whether someone actually is infected. However, it is common knowledge that such tests have a certain rate of false positives and negatives. In the case of COVID-19 specifically, positive screening tests are not followed-up (as is usually common practice to confirm a diagnosis) because of scarcity in testing resources and/or prioritization of allocating tests to the sickest patients (Frasier, 2020). Moreover, BMJ (2020) reports that serological tests for COVID-19 carry with them risks of bias and heterogeneity in their accuracy. Therefore, they state that these serological tests should only be used cautiously for clinical decision making and epidemiological surveillance. For this reason, one could choose to relax the assumption and assume $f(N_t) = f^{max}$ for some $f^{max} \in [0, 1]$ set to be a more reasonably perceived value.

- (A4) As mentioned earlier in this section, f_t needs to be monotonically increasing in TC_t , i.e. the proportion of infectives that is documented is increasing in the testing capacity. Mathematically, this means that

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

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

- **Linear form**

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

- **Quadratic form**

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

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

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

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

- We assume that the vertex (i.e. the extremum) is the point $(N_t, 1)$, i.e. the parabola is downwards opening. Recall that any quadratic function can be rewritten to the so-called vertex form $f(x) = a(x-h)^2 + k$, where the vertex of the function is (h, k) . Choosing this special case means that there will be no unknown parameters needed to define the function because we know the location of the vertex and a known point $(0, f^{min})$ on the parabola. We can then derive that the formula becomes:

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

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

- For the same reason as for the previous specification, we assume that the vertex is the point $(0, f^{min})$, i.e. the parabola is upwards opening. We can then derive that the formula becomes:

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

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

• Cubic form

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

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

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

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

$$f_{p,t} := f(TC_{p,t}). \quad (6.7)$$

such that

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

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

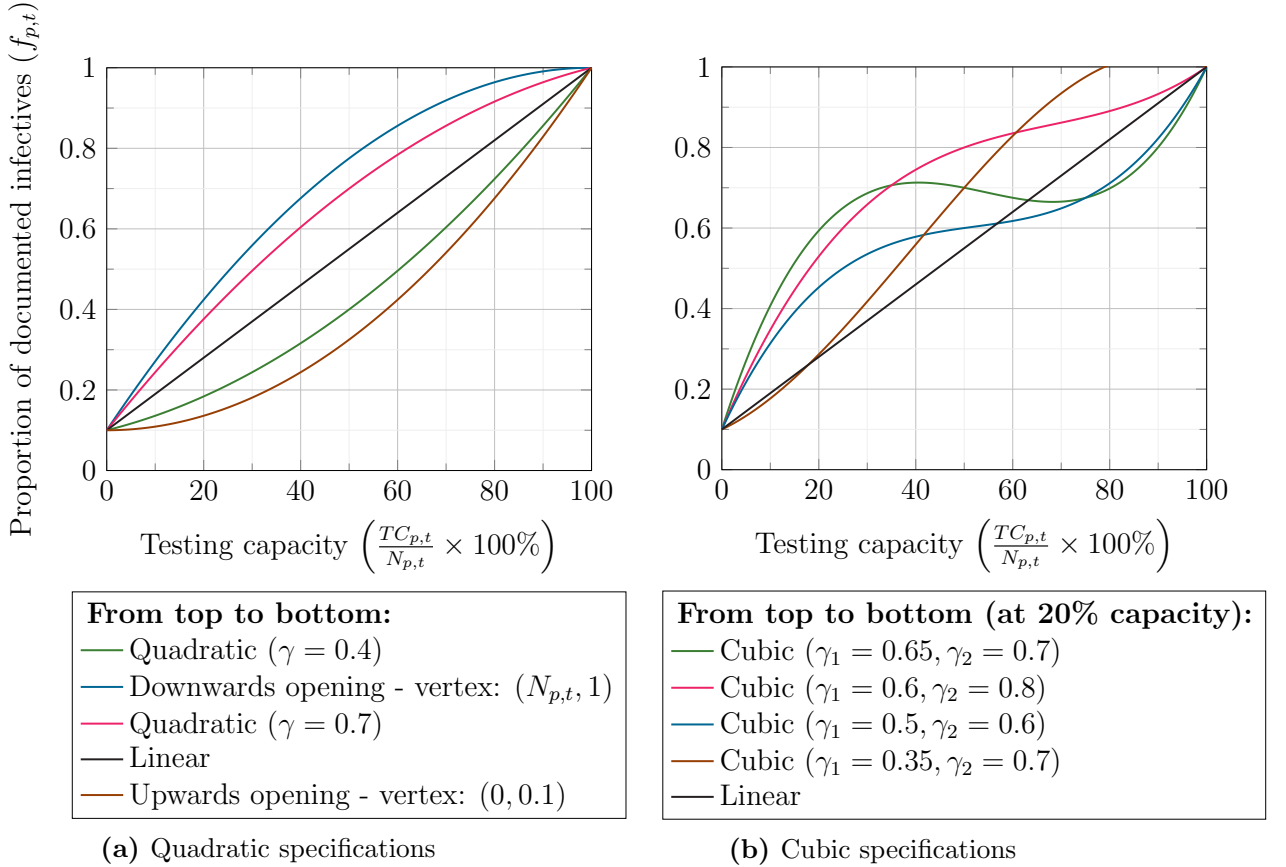


Figure 6.1. Functional forms for the proportion of documented infectives ($f^{min} = 0.1$)

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

if we would use a cubic form, the values of γ_1 and γ_2 should first be tested by means of a plot, for instance.

Next, we argue which of these forms is most appropriate. As mentioned at the beginning of this section, we cannot estimate which form would fit the data best because there is, by definition, no data on the undocumented infectives. As such, we argue which functional form to use by theoretical logic rather than a mathematical approach. Before that, notice that the shape of the functional form may differ depending on the basic reproduction number R_{eff} , as defined in Section 2. R_{eff} estimates how many people an infective will on average infect. If $R_{eff} > 1$, a person is estimated to infect more than one person and an epidemic is expected to develop. In this case, we expect that an increased testing capacity will have a larger immediate effect. We assume that a person who has been tested positive adheres to the common guidelines that they should self-quarantine. Consequently, this infective does not infect other people who would otherwise become undocumented infectives. For the remainder of this argument, we assume that $R_{eff} > 1$. The reason for this is that there are a variety of methods to estimate R_{eff} and that we cannot reasonably make our own model easily dependent on these varying results. Future research could be conducted regarding a two-step approach.

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

We first argue why a downwards opening quadratic function fits the requirements well. Note that when a large proportion of the population has been tested, the pool of untested people, who are potentially infectious, is smaller. The probability that they, in isolation of other effects, are infected is lower. The argument for this is as in the previous paragraph: assuming that the people close to them who were tested positive (be that family, acquaintances, or those that they would perhaps run into at the supermarket) do indeed self-isolate, they would not have been able to be in contact with them and they have a lower chance to be infected. When a small number of people is tested and suddenly the testing capacity is increased, a larger pool of people who had symptoms and could previously not be tested, now have access to a test. The people who are now most likely to get tested positive have strong symptoms. As they are now tested positive, we assume they self-quarantine and cannot infect other people. Therefore, the functional form that fits this argument best is a downwards opening quadratic function.

One could also consider the cubic representation with $\gamma_1 = 0.6$ and $\gamma_2 = 0.8$, or

some similar parameter values, as in Figure 6.1b. There, we see similar behaviour at the start of the graph where there is a sharp increase, after which it levels out. The difference is found when there is the testing capacity to test the last proportion of the population, leading to a sudden sharp increase in the proportion of documented infectives. An argument in favour of this specification is that it may be difficult to track down and convince the last proportion of the population to take a test who, at that point, may be infectious. For instance, these may simply be people who refuse to take such a test, whether those reasons are grounded or not. There may also be people who underestimate their symptoms or their importance and who, even though they are encouraged to get tested, believe that they do not need to be. For instance, they may feel that others need to get the test more. If these infective people do not get tested, the proportion of documented infectives may level out more quickly.

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.

Lastly, the question is what to choose for the parameter γ . Recall that (6.3) and (6.4) are equivalent when $\gamma = \frac{3}{4} + \frac{1}{4}f^{min}$, meaning that (6.4) is the most extreme case possible and that the slope cannot be constructed to be more steep. To be general, we choose (6.3) to be our functional form with an unknown parameter γ , denoted by $f_{p,t}(\gamma)$. The models can then use a specific value for γ or perhaps can choose an approach that incorporates a possibility to leave γ as an unknown parameter, possibly to be estimated.

Now that we have chosen our functional form, we are interested in investigating the relationship between $TC_{p,t}$ and $f_{p,t}(\gamma)$ over time for all regions and to compare these. Because the population size differs over the regions, this is likely to impact the absolute number of tests executed. As such, instead of comparing $f_{p,t}(\gamma)$ to $TC_{p,t}$, we compare it to $TC_{p,t}/N_{p,t}$. The results are shown in Figure 6.2.

In Figure 6.2, we can see that the pattern of the relationship between the two variables is similar over time for different groups of regions. Importantly, we draw attention to the fact that the proportion of infectives that go undocumented decreases over time. This is logical because the testing capacity increases over time and we have assumed a monotonic relationship.

For illustration purposes, we now give an example of the impact of this modelling method. We present the number of infectives at three time points for three regions in Table 6.1. In this Table, i_t is computed as DI_t/f_t .

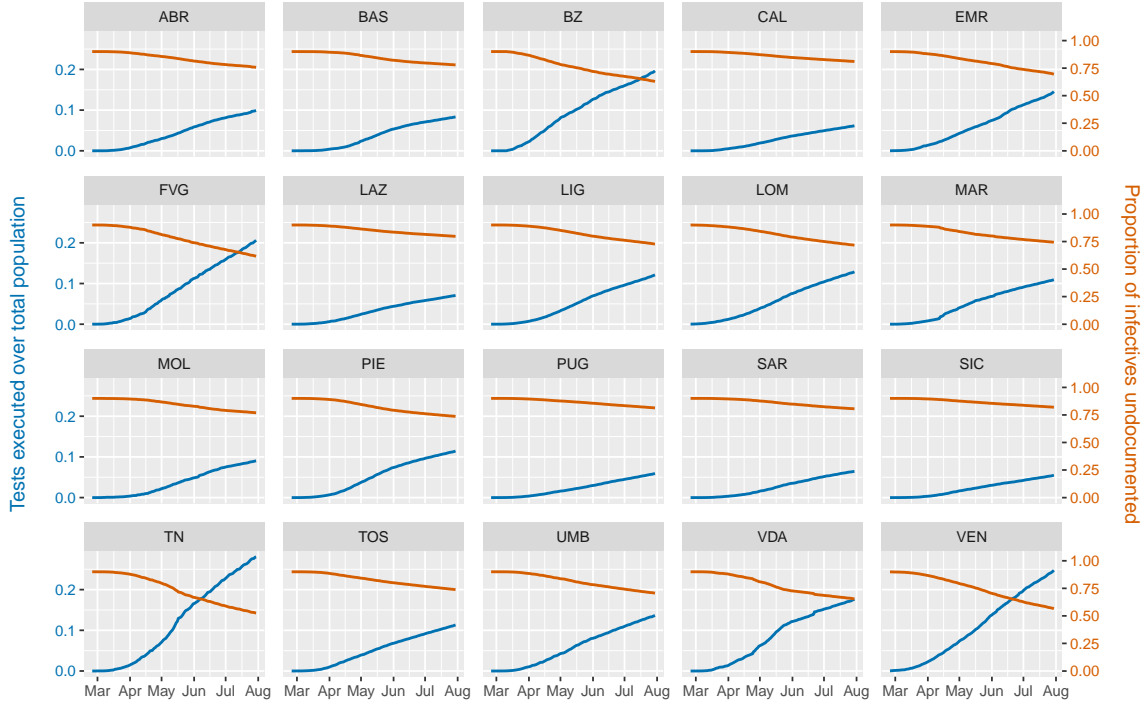


Figure 6.2. Total number of people tested over the total population ($TC_{p,t}/N_{p,t}$) versus proportion of infectives that are documented $f_{p,t}(\gamma)$

Table 6.1. Impact of modelling undocumented infectives over time. The quadratic specification with $\gamma = 0.7$ is used.

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

Table 6.1 shows us that the impact of the proportion of documented infectives f_t differs over the regions. We chose Calabria, Lombardy, and Veneto because these regions vary in the proportional amount of tests executed, leading to different profiles in f_t . We can see this profile in Figure 6.2 as well. When the amount of tests executed grows less steeply, as is the case in Calabria, the number of undocumented infectives in society grows stronger. On the other hand, for a region that invests heavily in testing, such as Veneto, the undocumented infectives are less pronounced. For example, consider the changes in Calabria and Veneto from June 1 to August 1. For Calabria, the growth in the documented infectives accounted for only 17.87% of the total growth in infectives. In contrast, in Veneto the growth in the documented

infectives accounted for 40.46% of the total growth. Of course, Lombardy finds itself in the middle, where documented infectives make up 23.87% of the total growth. Hence, our method correctly incorporates the intuition that a higher testing capacity leads to more infectives being documented.

7 Within-Region Spread Model

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

7.1 Methodology

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

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

There are a few things to notice. Firstly, if we want to estimate (7.1)'s parameters, an error occurs. This is added to the model through an error term denoted by $\eta_{p,t}$:

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

Secondly, individuals that get infected do not immediately infect others because there is a so-called latent period, which is the period between an infection and the moment that the infective is infectious. For COVID-19, the latent period is estimated to be approximately 2 days shorter than the incubation period (He et al., 2020). The incubation period is the period between an infection and the moment that the infected individual starts showing symptoms, at which point the infective is said to be symptomatic. The incubation period for COVID-19 is estimated to be above 2 and below 11.5 (Lauer et al., 2020), 12.5 (Q. Li et al., 2020), or 14 days (Linton et al., 2020). This is a large range, but this is not rare. For instance, the incubation period for chicken pox is estimated to be between 9 and 21 days (Papadopoulos, 2018). While the maximum incubation period is not agreed upon by Lauer et al. (2020) and

Q. Li et al. (2020), their results on the median are similar. Lauer et al. (2020) report a median incubation period of 5.1 days (95% CI: 4.5 to 5.8 days), while Q. Li et al. (2020) report a median incubation period of 5.2 days (95% CI: 4.1 to 7.0 days). For comparison, Linton et al. (2020) give the result of a mean incubation period of 5.0 days (95% CI: 4.2 to 6.0 days) when excluding Wuhan residents and 5.6 days (95% CI: 5.0 to 6.3 days) when including Wuhan residents.

Because the latent period is estimated to be shorter than the incubation period, there are infectives who are able to infect others before showing symptoms. We call these people pre-symptomatic, which is distinctive from asymptomatic people in the sense that asymptomatic people do not develop symptoms and pre-symptomatic people will develop symptoms but they develop a higher viral load just before said symptoms are apparent. On June 9, 2020, the World Health Organization said that pre-symptomatic people may actually be able to infect others (Sutherland & Gretler, 2020). This may be an issue when considering policies such as self-isolation when one is sick, because an infective may have already spread the virus before feeling sick. Sutherland and Gretler (2020) moreover reiterate the WHO’s statement that studies have been done that show that asymptomatic people can spread the virus but that more research needs to be done to show how many of these infectious asymptomatic people exist. We discuss how we model pre-symptomatic individuals in the form of undocumented infectives in Section 6.

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

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

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 14 days, we choose $\tau = 14$.

Adda (2016) adds regressors to the model as control variables, such as the region fixed effects, week effects and year effects in levels. Regressors can be added to the model to capture possible effects that would otherwise be included in the error, confounding the estimation of the transmission parameter β . Adda (2016) denotes this tensor of regressors by X . This leads to the following formulation:

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

For our application, the data does not span multiple years. As such, we do not have year effects. Also note that week effects would capture a time trend because we do not have the same week number multiple times for the same region. Therefore,

we will not include a week effect. We do add a weekend effect. More information and reasoning is provided in Section 3.3.

There are two other key differences in the model specification by Adda (2016) compared to (7.4) that are not clear from the paper. First of all, Adda (2016) replaces the incidence rate $I_{p,t-\tau}$, also called the incidence rate, by the number of new cases $i_{p,t-\tau} - i_{p,t-\tau-1}$. Similarly, Adda (2016) puts the dependent variable to be the number of new cases instead of the incidence rate. Second of all, Adda (2016) does not include the term $\gamma I_{p,t-\tau}$ in the model. Presumably, this is because Adda (2016) considers the number of new cases instead of the total number of infectives and, therefore, the number of recovered individuals do not impact that value. In this section and Section 8.1, we will continue with this model to see how the models perform in the case of COVID-19. Even though the mathematical derivation of the model does not explicitly derive from the SIR model, it may still provide correct estimates although there is no econometric evidence for this so far.

Now we present the final model, ignoring effects across regions, by Adda (2016). Defining $\Delta i_{p,t} := i_{p,t} - i_{p,t-1}$ and following the notation by Keeling and Rohani (2011), the within-region model is given by:

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

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

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

A general assumption that is made, is that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. Therefore, we assume that $E[\eta_{p,t} | X_{p,t}] = 0$. The reason why we assume that $E[\eta_{p,t} | S_{p,t-\tau} \Delta i_{p,t-\tau}] = 0$ is that, for a large enough lag τ , the error is not correlated with past data at that lag. By that, we mean that the people that are classified as infectives at time $t - \tau$ do not have an effect on the error that we make when considering the infectives at time t under a correct model specification. Because we chose τ to exceed the maximum estimated incubation period, we assume that this holds. Therefore, we assume that the moment condition holds.

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

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

We can rewrite (7.6) as follows

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

The moment conditions that then need to hold are:

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

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

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

Just like earlier in this section, we make the assumption that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. Therefore, we assume that $E[\eta_{p,t} | X_{p,t}] = 0$. The reason why we assume that $E \left[\eta_{p,t} \left| S_{p,t-\tau} \left(\frac{DI_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{DI_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right) \right] = 0$ is for the same reason as earlier in this section, namely that for a large enough lag τ , the error is not correlated with past data at that lag. 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.

The issue may come when we consider the condition $E \left[\eta_{p,t} \left| \frac{DI_{p,t-1}}{f_{p,t-1}(\gamma)} \right] = 0$. Again, this is not due to the scaling function $f_{p,t-1}(\gamma)$ but when considering the term $DI_{p,t-1}$. However, recall that the error $\eta_{p,t}$ is made when predicting the amount of infectives at time t . In Section 7.1, we have explained that there is a latent period during which an infected person is not able to infect others yet. As such, when considering only a one-period difference, there should be no correlation between $\eta_{p,t}$ and $DI_{p,t-1}$. Consequently, the scaling of the infectives by using our functional form, has no additional impact on the moment conditions. Therefore, we assume that the moment conditions hold, even when modelling undocumented infectives according to the method explained in this section.

To conclude, one could naively consider constructing a model for the entire nation of Italy. Even though this does not take into account regional differences, as described

in Section 2, it may achieve good results if regions are sufficiently similar. However, as we have already seen in Section 2, there is indeed a difference among regions and pooling the data is likely not a good solution. For good measure, we also estimate a national model. This is done in two ways, the first of which is simply by adding the values of s , i , and r of all regions together to obtain the national numbers, for which the model from (7.5) or (7.6), depending on whether undocumented infectives are modelled, is estimated. The results from this model will be indicated by National (OLS) in the results section. The second method is to apply pooled OLS (POLS), which is a panel data estimation method. This model ignores the individual regional effect, hence treating the data as one large cross-section. This means that the T observations for some region i are actually treated to be cross-sectional observations of T different individuals. These results will be indicated by National (POLS) in the results section.

Q: I also looked into Random Effects but it returns the same results. I Googled and could not find an explanation. I think it may be because there is little regional heterogenous variation, i.e. $\sigma_a^2\alpha = 0$. Is this correct? If so, do you know of a source that I can cite or is this unnecessary and can I just write it down myself?

7.2 Results

In this section, we present the results for the within-region spread model. Firstly, we present the results where the data is pooled to a national level. Subsequently, results are presented for the models per region to which model selection is applied with the Akaike Information Criterion (AIC). For both result sets, we present the results from the regular model as well as modelling the undocumented infectives with a quadratic form with $\gamma = 0.7$ and $f^{min} = 0.1$ as in (6.3).

For all models, we apply a rolling window of 100 days, meaning that the used data spans May 9 until August 16. The reason behind this is that infections from the past should not dominate the latest estimates of the transmission rate parameters. When the (first) wave has passed and the transmission rate is low, including the data from during the peak moment of the pandemic will influence the transmission rate. We will illustrate this at the end of the section.

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

Table 7.1. Estimates from 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	<i>t</i> -value	<i>p</i> -value	Estimate	Std. Error	<i>t</i> -value	<i>p</i> -value
Weekend	92.860	30.241	3.071	0.003***	15.086	2.974	5.074	0.000***
β_{within}	0.493	0.022	21.992	0.000***	0.533	0.019	27.980	0.000***

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

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

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

The test statistic is constructed as follows:

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

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

As mentioned at the end of the previous section, this national model does not take into account effects specific to regions. It is also clear that the same model might not be suitable for all regions; we should apply model selection to the individual models as was explained in Section 5. To execute model selection, we use the AIC and we make sure that the term for β_{within} remains in the model, meaning that model selection is solely performed on whether the weekend dummy should be included. In Table 7.2, we present the results. In Table B.1 in Appendix B, we present the results from running the model on each region separately with the same model specification for each region. The results comparing the use of the BIC over the AIC for model selection are presented in Table B.2, where we see that the BIC selects a smaller model for three regions, being Marche, Apulia, and Veneto, but that the model specification is the same for all of the other regions.

Table 7.2. Estimates from Within-Region Spread Model per region with model selection by AIC. Estimates are given with t -statistics in parentheses. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Region	Regular model		Modelling undocumented infectives	
	β_{within}	Weekend	β_{within}	Weekend
National (OLS)	0.493*** (21.992)	92.860*** (3.071)	0.423*** (27.417)	379.483*** (2.976)
National (POLS)	0.533*** (27.980)	15.086*** (5.074)	0.458*** (22.981)	90.994*** (4.893)
ABR	0.372*** (4.952)	2.751** (2.091)	0.322*** (6.253)	13.454** (2.249)
BAS	0.073 (0.665)		0.077 (0.764)	
BZ	0.269*** (3.468)	2.346*** (3.683)	0.219*** (4.175)	7.117*** (3.961)
CAL	0.242** (2.043)	2.189** (2.614)	0.201* (1.917)	12.897*** (2.828)
EMR	0.358*** (11.673)	22.416*** (4.844)	0.287*** (14.768)	91.467*** (5.266)
FVG	0.342*** (7.559)	2.582*** (3.476)	0.252*** (8.891)	8.578*** (3.508)
LAZ	0.494*** (10.069)	10.172*** (3.441)	0.424*** (11.273)	57.018*** (3.610)
LIG	0.382*** (11.931)	4.179 (1.430)	0.332*** (14.770)	
LOM	0.583*** (18.205)		0.513*** (18.615)	
MAR	0.436*** (9.376)	2.697** (2.144)	0.387*** (11.347)	12.796** (2.286)
MOL	0.140 (1.595)	1.855** (2.456)	0.155* (1.810)	11.448** (2.299)
PIE	0.352*** (23.006)	7.449* (1.896)	0.300*** (25.913)	28.533 (1.492)
PUG	0.376*** (6.165)	3.804** (2.581)	0.335*** (7.302)	23.618*** (2.719)
SAR	0.437*** (4.417)		0.276*** (3.874)	5.301 (1.655)
TOS	0.281*** (6.666)	8.444*** (3.854)	0.237*** (8.744)	35.079*** (3.929)
UMB	0.436*** (3.936)	1.428*** (3.168)	0.291*** (4.134)	5.402*** (3.351)
VDA	0.290*** (4.341)		0.302*** (6.394)	
VEN	0.489*** (6.748)	11.951* (1.920)	0.284*** (7.344)	39.380*** (2.672)

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

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

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

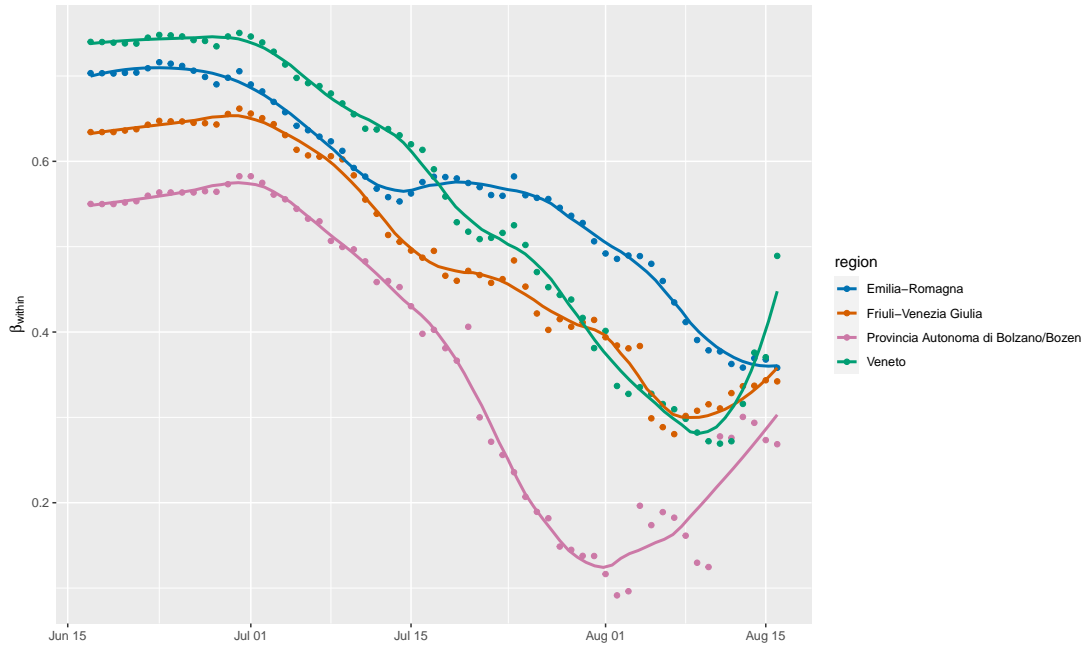
We are again interested in investigating whether the differences in the estimates differ significantly. For this, recall that the standard error of the estimate can be computed by dividing the estimate by the t -statistic. The t -statistic for the national models are equal to 2.572 (OLS) and 2.720 (POLS),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

When we want to interpret the estimates of β_{within} , we should recall that Adda (2016) notices that these can be interpreted as the marginal effects of a change in the

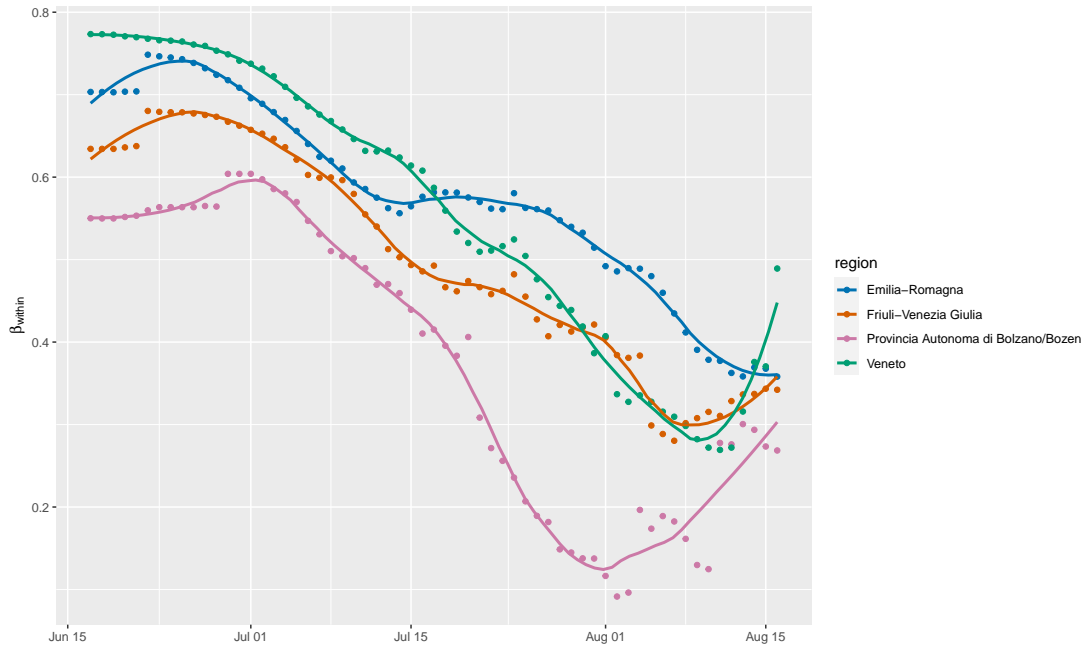
infection rate on the future infection rate when the entire population is susceptible to the disease. However, notice that this is made with the model formulation as explained in Section 7.1, where the removal term is omitted from the model. Because this does not take into account the specific removal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

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

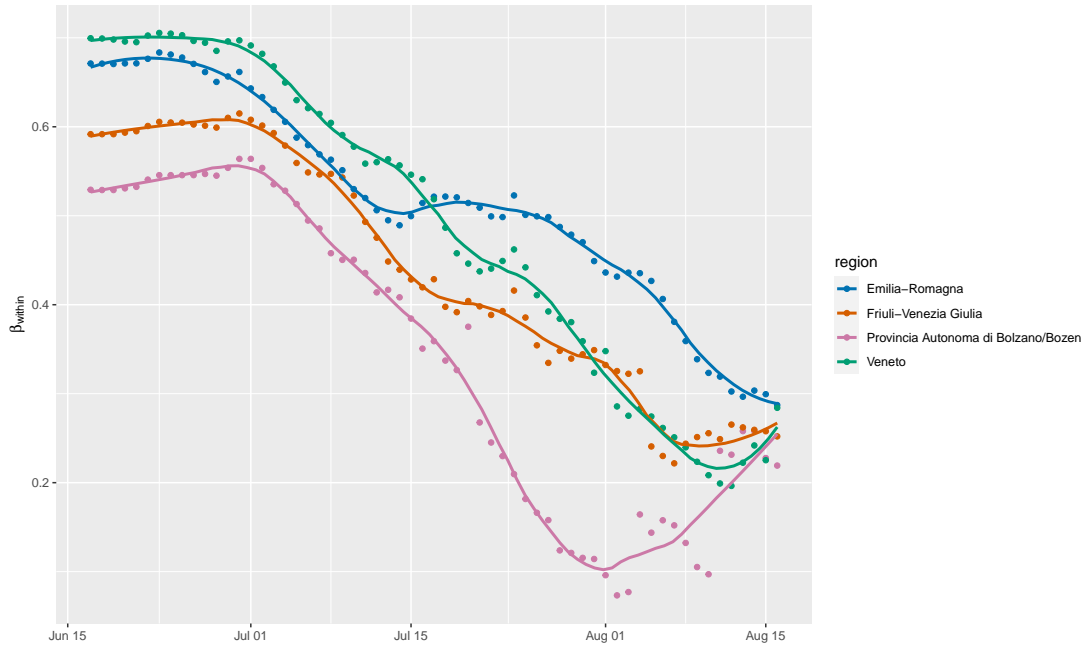
We are also interested in looking at the estimate of β_{within} over time. We expect that it decreases over time, implying that SARS-CoV-2 is transmitted less. In Figure 7.1, we present plots for the regions in the *Nord-Est* (North-East) NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.3, which generally show similar results. Each point in the graphs in Figure 7.1 is the estimate of β_{within} when only the latest 100 data points before that date are used. In addition, a LOESS (locally estimated scatter plot smoothing) curve with span parameter 0.3 is fit to the data points.



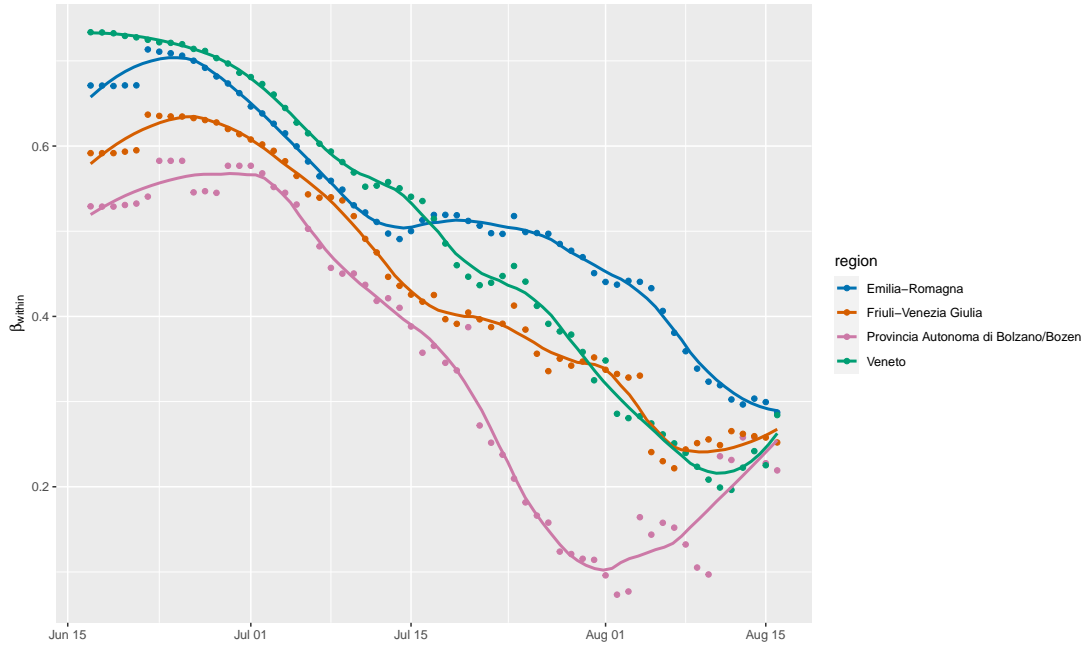
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



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

Figure 7.1. Progression of β_{within} over time for the *Nord-Est* (North-East) NUTS 1 region

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

At the start of this section, we mentioned the application of a rolling window, meaning that the last 100 days of data are used instead of the entire dataset. We noted that this was the case to avoid old data skewing the results. Consider Figure 7.2, where we present the plot of β_{within} over time for the *Nord-Est* NUTS 1 region when a rolling window is not applied. Undocumented infectives are modelled, as before, using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

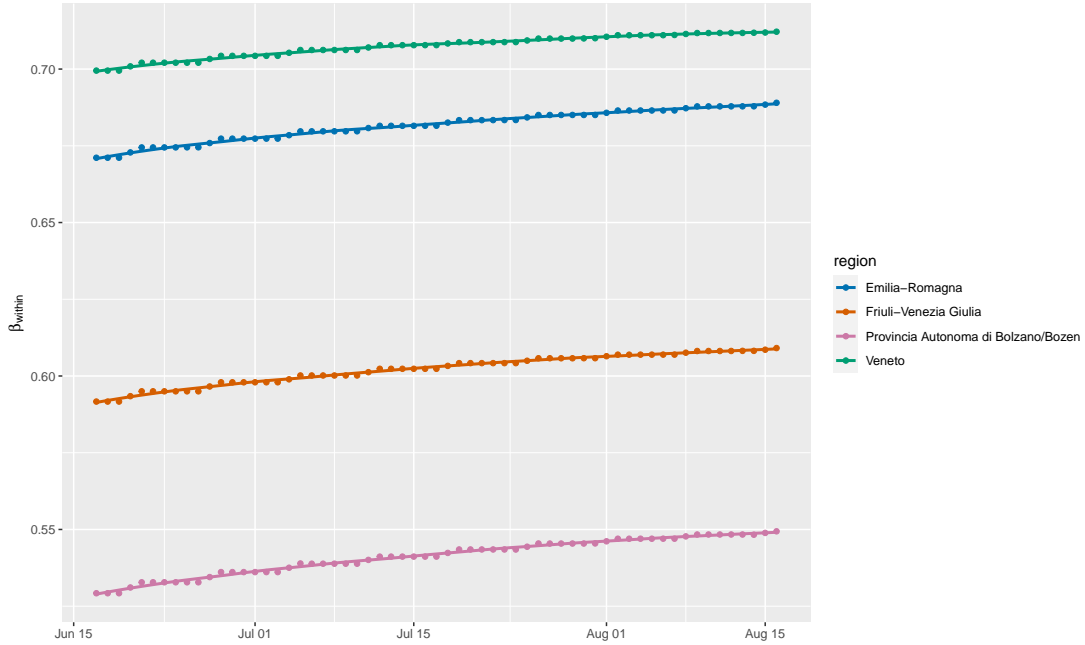


Figure 7.2. Progression of β_{within} over time for the *Nord-Est* (North-East) NUTS 1 region including undocumented infectives without applying a rolling window

Compare Figure 7.2 with Figure 7.1c, which applies the same specifications but without applying the rolling window. It is immediately clear that the variation in β_{within} is minimal and even slightly increasing. This not intuitive for the reason that a decrease in infectives over time would lead to a lower estimated transmission

parameter. As such, it is not logical that the estimate for β_{within} would level out, especially at a level not close to zero, or increase. The reason behind choosing 100 days is arbitrary; it is simply a number that retains enough data points while providing variation in the estimates for β_{within} .

8 Within and Between-Region Spread Model

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

8.1 Methodology

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

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

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

Notice, moreover, that it does not make sense to consider a national model. Because we do not consider countries outside of Italy, the set $R \setminus r$ is empty if we consider r to be the entire country of Italy. This would mean that the national model for the within and between-region spread model is equivalent to the national model for the within-region spread model. As such, in this section, we only consider the model

applied to the regions.

In (8.1), the transmission parameter β is now allowed to be different within and between regions. Adda (2016) estimates (8.1) by OLS and by instrumental variable estimation (IV). Weather episodes, such as the amount of rain and temperature-related instruments, are used as instruments. There is a biological reasoning behind choosing these instruments, for instance that warmer temperatures tend to have a negative effect on the proliferation of some viruses. A social reason is also given, namely that bad weather conditions impact the amount of social interaction between people, meaning that there are less opportunities for viruses to spread. We challenge the choice of these instruments, particularly in the case of SARS-CoV-2. Unfortunately, we do not have sufficient information on the effect of the weather on the virus. SARS-CoV-2 has only been quite apparent since January 2020 and there has not been enough fluctuation over time in temperatures to show a necessary effect that can be disentangled from, for example, policies being effective in driving the virus back. For this reason, we only consider OLS for this model.

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

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

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

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

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

We can rewrite (8.2) as follows

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

The moment conditions that then need to hold are:

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

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

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

Just like before, we make the assumption that the idiosyncratic error $\eta_{p,t}$ is uncorrelated with the regressors in the tensor $X_{p,t}$. Therefore, we assume that $E[\eta_{p,t} | X_{p,t}] = 0$. Now note that there are three additional terms to consider, namely the relation between $\eta_{p,t}$ and $S_{p,t-\tau} \left(\frac{DI_{p,t-\tau}}{f_{p,t-\tau}(\gamma)} - \frac{DI_{p,t-\tau-1}}{f_{p,t-\tau-1}(\gamma)} \right)$, $\beta_{between} S_{p,t-\tau} \sum_{c \in R \setminus r} \left(\frac{DI_{c,t-\tau}}{f_{c,t-\tau}(\gamma)} - \frac{DI_{c,t-\tau-1}}{f_{c,t-\tau-1}(\gamma)} \right)$, and $\frac{DI_{p,t-1}}{f_{p,t-1}(\gamma)}$, the first and last of which have been discussed in Section 8.1. That leaves the middle term, for which the reasoning is identical to the first term: for a large enough lag τ , the error is not correlated with past data at that lag, which is independent of the scaling functions $f_{p,t-\tau}(\gamma)$ and $f_{p,t-\tau-1}(\gamma)$ as these are constructed

without the past infectives in mind. We also say that it does not matter whether we consider infectives within the region or in other regions, as the longer time lag applies in any case. Therefore, we assume that the moment conditions hold.

8.2 Results

In this section, we present the results for the within and between-region spread model. In Table B.3 in Appendix B, we present the results from running the model on each region separately without applying model selection. The results comparing the use of the BIC over the AIC for model selection are presented in Table B.4, where we see that the BIC selects a smaller model for four regions, being Calabria, Liguria, Piedmont, and Umbria, but that the model specification is the same for all of the other regions. In Table 8.1, we present the results where we execute model selection using the AIC and we make sure that the terms for β_{within} and $\beta_{between}$ remain in the model.

Table 8.1. Estimates from Within and Between-Region Spread Model per region with model selection by AIC. Estimates are given with t -statistics in parentheses. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are modelled using the quadratic specification with $\gamma = 0.7$ and $f^{min} = 0.1$.

Region	Regular model			Modelling undocumented infectives		
	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
ABR	-0.011 (-0.079)	$5.871 \times 10^{-3***}$ (3.399)	2.394* (1.909)	-0.047 (-0.495)	$5.943 \times 10^{-3***}$ (4.455)	12.494** (2.279)
BAS	0.040 (0.357)	7.638×10^{-4} (1.286)		0.034 (0.321)	6.299×10^{-3} (1.379)	
BZ	0.246** (2.355)	2.043×10^{-4} (0.317)	2.305*** (3.534)	0.201** (2.559)	1.018×10^{-4} (0.310)	7.030*** (3.848)
CAL	0.186 (1.460)	7.936×10^{-4} (1.171)	1.893** (2.168)	0.128 (1.088)	8.270×10^{-4} (1.341)	11.514** (2.472)
EMR	0.365*** (4.184)	-8.645×10^{-4} (-0.080)	22.410*** (4.817)	0.171*** (2.635)	0.015* (1.873)	90.384*** (5.267)
FVG	0.114* (1.707)	$3.701 \times 10^{-3***}$ (4.346)	1.759** (2.481)	0.034 (0.929)	$3.026 \times 10^{-3***}$ (7.601)	4.330** (2.139)
LAZ	0.342*** (2.671)	7.364×10^{-3} (1.283)	10.573*** (3.568)	0.238** (2.221)	0.010* (1.853)	61.107*** (3.879)
LIG	-0.026 (-0.372)	0.033*** (6.515)		-0.023 (-0.404)	0.029*** (6.613)	
LOM	0.329*** (4.929)	0.226*** (4.257)		0.285*** (4.463)	0.190*** (3.926)	
MAR	0.245** (2.521)	$4.505 \times 10^{-3**}$ (2.229)	2.150* (1.710)	0.181** (2.312)	$4.644 \times 10^{-3***}$ (2.907)	10.483* (1.921)
MOL	-0.085 (-1.168)	$4.063 \times 10^{-3***}$ (8.821)		-0.077 (-1.158)	$4.596 \times 10^{-3***}$ (9.811)	
PIE	0.104***	0.065***		0.091***	0.057***	

Table 8.1 continues on next page

Table 8.1 continued from previous page

Region	Regular model			Modelling undocumented infectives		
	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
PUG	(3.665)	(9.767)		(3.882)	(9.785)	
	0.120	$5.525 \times 10^{-3***}$	3.083**	0.035	$8.078 \times 10^{-3***}$	18.908**
SAR	(1.051)	(2.629)	(2.115)	(0.393)	(3.896)	(2.303)
	0.407***	1.846×10^{-3}		0.252**	1.831×10^{-4}	5.062
TOS	(2.903)	(0.299)		(2.468)	(0.321)	(1.532)
	0.051	0.010***	8.023***	0.030	$9.454 \times 10^{-3***}$	33.901***
UMB	(0.605)	(3.107)	(3.812)	(0.535)	(4.141)	(4.095)
	0.427***	4.689×10^{-5}	1.410***	0.235**	2.449×10^{-4}	4.877***
VDA	(2.981)	(0.102)	(2.904)	(2.467)	(0.876)	(2.832)
	0.023	$1.111 \times 10^{-3***}$		0.055	$7.104 \times 10^{-4***}$	
VEN	(0.259)	(4.193)		(0.775)	(4.374)	
	0.904***	-0.033***	16.184***	0.434***	-7.954×10^{-3}	43.868***
	(6.990)	(-3.771)	(2.721)	(4.370)	(-1.637)	(2.951)

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

We first discuss the statistical evidence for the estimated values of β_{within} and $\beta_{between}$ and how these interact. After this, we will consider the interpretation of the coefficients and the model selection. Before that, there are two things to notice to start with. Firstly, notice that the estimates for $\beta_{between}$ are generally much smaller than the estimates for β_{within} . This is likely the case because Adda (2016) defined the models with the absolute number of new cases instead of a proportion. As such, summing over all regions leads to a large number of total new cases, causing the parameter estimate to be driven down. A notable exception being the region of Lombardy, where the estimate for $\beta_{between}$, namely 0.378, is indeed smaller than the estimate for β_{within} , namely 0.514, but the two estimates are much closer than for other regions. A similar result can be found for the region of Piedmont. For the region of Lombardy, the reason for the estimates being much closer is twofold. Firstly, because it is the largest region in Italy, housing around one-sixth (16.67%) of the total population of Italy, where the second-largest region is Lazio, which houses 9.74% of the total amount of Italians. Secondly, Lombardy was hit the hardest by SARS-CoV-2 of all of the Italian regions and, therefore, the number of infectives there is much higher than in other regions. As such, summing the other regions has a proportionally smaller effect.

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

significant at a level of 0.01. Future research can be done into models that restrict the estimates to be positive.

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

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

Looking only at the statistically significant estimates of β_{within} , these range from 0.104 for Piedmont until 0.904 for Veneto for the regular model and from 0.091 for Piedmont until 0.434 for Veneto when undocumented infectives are included. Excluding Veneto, the highest values for the regular model and the model including undocumented infectives are 0.427 for Umbria and 0.285 for Lombardy, respectively. Therefore, we notice that the estimates for Veneto are outliers and that the values are a bit lower than for the within-region model. When considering the statistically significant estimates of $\beta_{between}$, we see that these range from 1.111×10^{-3} for Aosta Valley until 0.226 for Lombardy for the regular model and from 7.104×10^{-4} for Aosta Valley until 0.190 for Lombardy when undocumented infectives are included. Conclusively, we see that the estimates of $\beta_{between}$ differ much less over the regions than those of β_{within} .

One aspect to pay attention to is regarding the impact of modelling undocumented infectives on the estimates. Notice that modelling undocumented infectives again causes the (significant) estimates of β_{within} to be lower than for the regular model, just like for the within-region spread model. For $\beta_{between}$, however, this is not the case. For four out of ten cases, namely for Abruzzo, Marche, Molise, and Apulia, where both estimates of $\beta_{between}$ are significant, the estimate when modelling undocumented infectives is higher. We believe that this is likely the case because all regions are affected in a similar way by the modelling method of undocumented in-

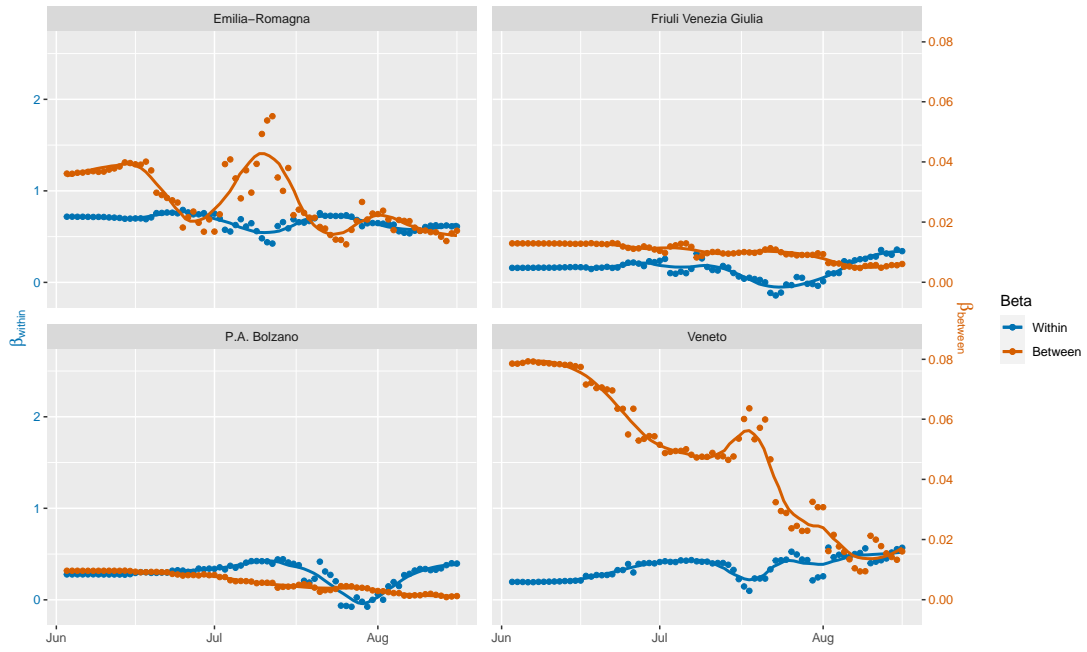
fectives, although the regions that test more rigorously are of course affected less. As such, there is no uniform way in which the estimate is affected. Also consider Figure 6.2, where we plotted the testing capacity versus the proportion of undocumented infectives for our modelling method. We do not see that a certain pattern happens for the four aforementioned regions that does not happen for other regions that find significant estimates of $\beta_{between}$, such as Lombardy, Piedmont, and Tuscany.

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

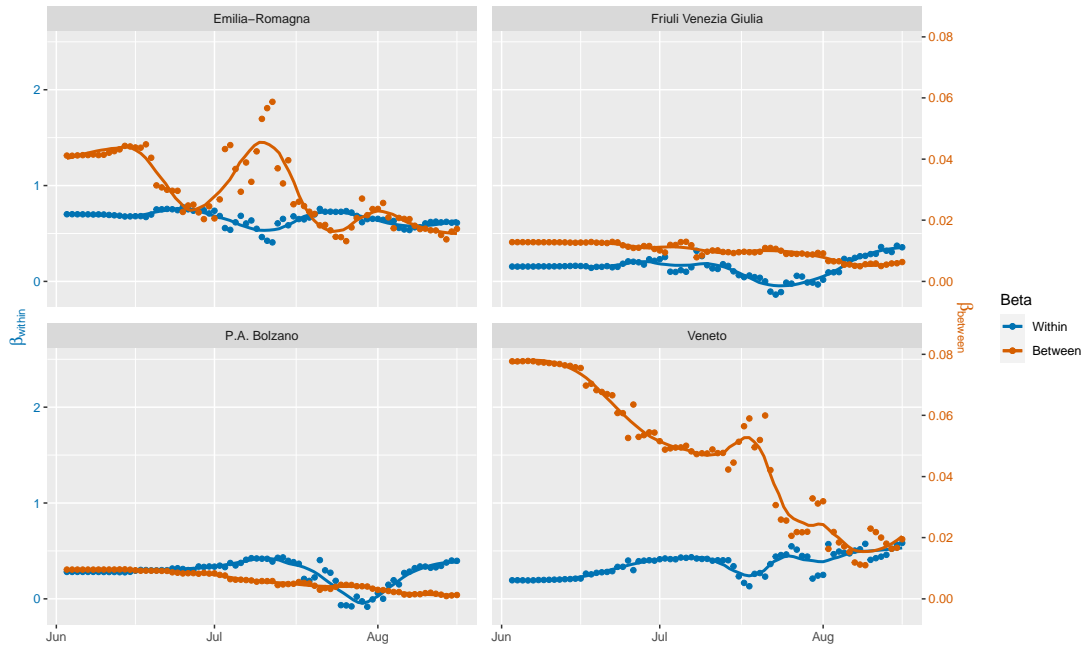
Recall that we cannot interpret the coefficients in the same way as Adda (2016) does but that we can compare the magnitude of the coefficients with one another across regions. For instance, let us compare the regions of Lombardy and Piedmont. For both regions, all estimates are statistically significant.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

Regarding model selection, the set of regressors that is used is the same whether undocumented infectives are modelled or not, except for the island of Sardinia. All models retain the terms related to β_{within} and $\beta_{between}$ in the model. For Sardinia, the weekend dummy is excluded in the regular model but it is included when undocumented infectives are modelled. In eleven out of the remaining seventeen cases, the entire model is selected. In the other six cases, the weekend dummy is excluded.

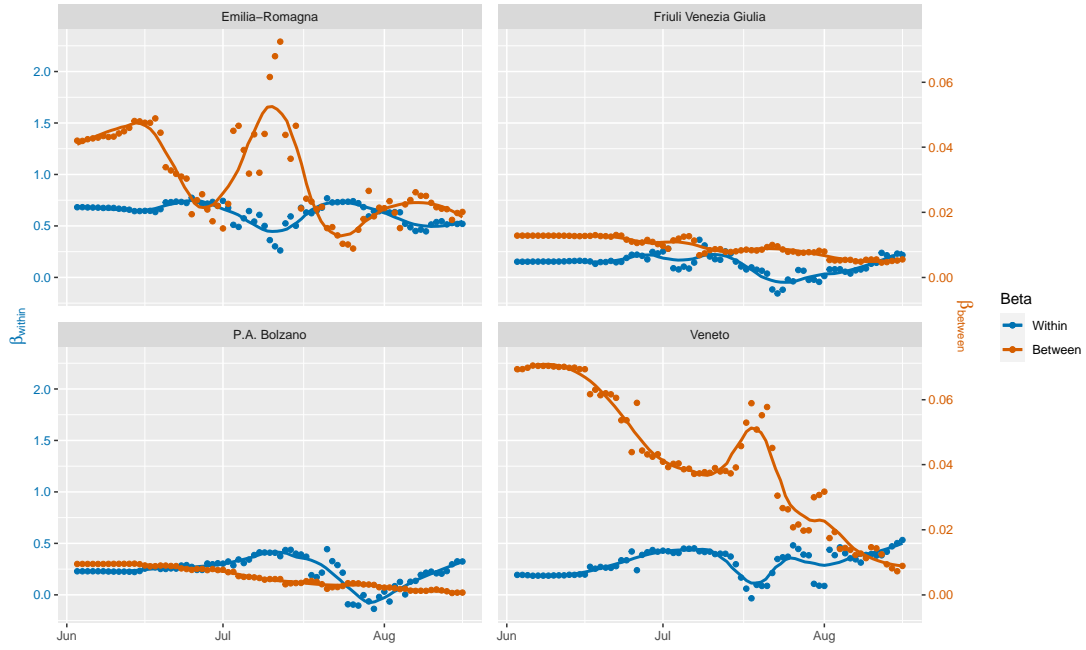
To conclude, we are again interested in looking at the estimates of β_{within} and $\beta_{between}$ over time. In Figure 8.1, we present plots for the regions in the *Nord-Est* (North-East) NUTS 1 region. Plots for the other NUTS 1 regions can be found in Appendix C.4. Each point in the graphs in Figure 8.1 is the estimate of β_{within} or $\beta_{between}$ when only the latest 100 data points before that date are used. In addition, a LOESS curve with span parameter 0.3 is fit to the data points.



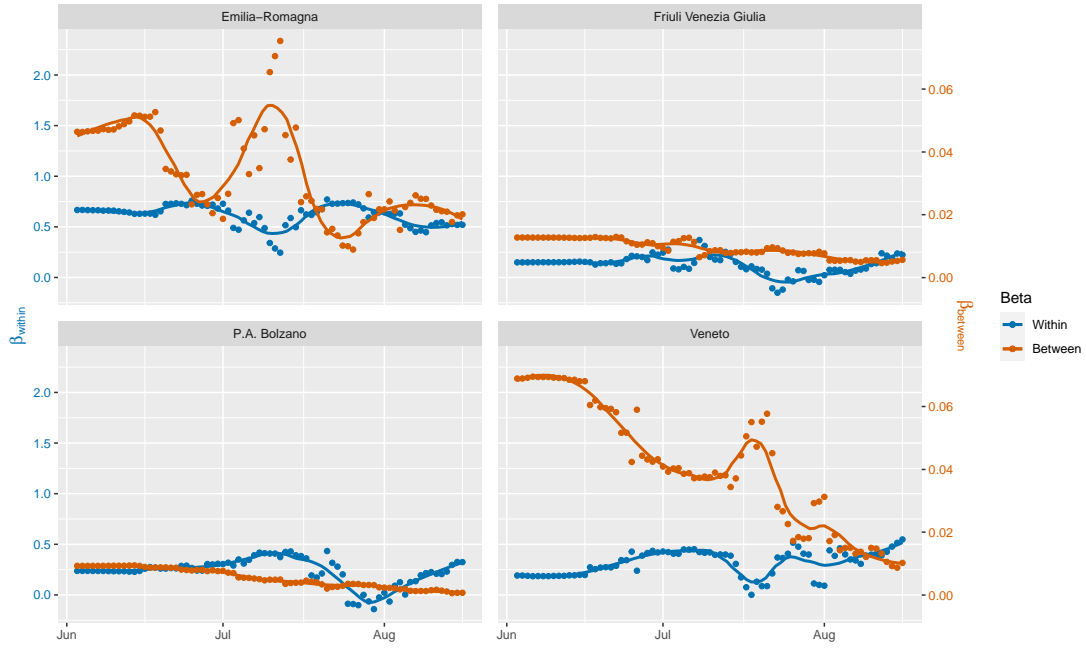
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



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

Figure 8.1. Progression of β_{within} and $\beta_{between}$ over time for the *Nord-Est* (North-East) NUTS 1 region

Figure 8.1 shows very different patterns compared to the within-region spread model, as discussed in the previous section. Firstly, we do not see a decreasing movement for β_{within} . In contrast, the values seem to fluctuate around some value. For $\beta_{between}$, we also see some fluctuation but we do generally see a decreasing movement. For the other NUTS 1 regions, we see some similar patterns although the progression varies a lot over the regions. For instance, the estimate for $\beta_{between}$ for the region of Lazio tends to increase over time, as can be seen in Figure C.14. Model selection by AIC does not seem to impact the estimates of β_{within} and $\beta_{between}$ much.

9 Conclusion

10 Future research

References

- Adda, J. (2016). Economic activity and the spread of viral diseases: Evidence from high frequency data. *The Quarterly Journal of Economics*, 131(2), 891–941.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723.
- Anderson, R. M., & May, R. M. (1992). *Infectious diseases of humans: Dynamics and control*. Oxford University Press.
- BBC News. (2020). *Death rate ‘back to normal’ in UK*. Retrieved July 1, 2020, from <https://www.bbc.com/news/health-53233066/>
- Bloomberg News. (2020). *Two Chinese patients test positive months after virus recovery*. Retrieved September 7, 2020, from <https://www.bloomberg.com/news/articles/2020-08-13/two-chinese-patients-test-positive-months-after-virus-recovery>
- BMJ. (2020). *Diagnostic accuracy of serological tests for COVID-19: Systematic review and meta-analysis*. Retrieved July 13, 2020, from <https://www.bmj.com/content/370/bmj.m2516/>
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.
- Burnham, K. P., & Anderson, D. R. (2002). A practical information-theoretic approach. *Model selection and multimodel inference*, 2nd ed. Springer, New York, 2.
- Business Insider. (2020). *Coronavirus: Nearly 200 North Korea soldiers ‘die from outbreak government refuses to acknowledge’*. Retrieved August 19, 2020, from <https://www.scmp.com/news/asia/east-asia/article/3074377/coronavirus-nearly-200-north-korea-soldiers-die-outbreak>
- Caccia, F. (2020). *Coronavirus, “il conteggio dei morti varia da paese a paese. la Germania esclude chi ha altre patologie”*. Retrieved June 11, 2020, from https://www.corriere.it/cronache/20_marzo_22/coronavirus-il-conteggio-morti-varia-paese-paese-germania-esclude-chi-ha-altre-patologie-6a452e6a-6c19-11ea-8403-94d97cb6fb9f_preview.shtml
- European Centre for Disease Prevention and Control. (2020). *Rapid risk assessment: Coronavirus disease 2019 (COVID-19) pandemic: Increased transmission in the EU/EEA and the UK - seventh update*. Retrieved August 17, 2020, from <https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-coronavirus-disease-2019-covid-19-pandemic>
- Eurostat. (2020a). *Eurostat regional data background*. Retrieved June 11, 2020, from <https://ec.europa.eu/eurostat/web/regions/background>
- Eurostat. (2020b). *Eurostat regional statistics database*. Retrieved June 11, 2020, from <https://ec.europa.eu/eurostat/web/regions/data/database>

- Frasier, S. L. (2020). *Coronavirus antibody tests have a mathematical pitfall*. Retrieved June 19, 2020, from <https://www.scientificamerican.com/article/coronavirus-antibody-tests-have-a-mathematical-pitfall/>
- He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine*, 26(5), 672–675.
- Horowitz, J. (2020). *Italy's health care system groans under coronavirus — a warning to the world*. Retrieved June 11, 2020, from <https://www.nytimes.com/2020/03/12/world/europe/12italy-coronavirus-health-care.html>
- Human Rights Watch. (2020). *Turkmenistan denies apparent covid-19 outbreak*. Retrieved August 19, 2020, from <https://www.hrw.org/news/2020/06/27/turkmenistan-denies-apparent-covid-19-outbreak>
- Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.
- Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 115(772), 700–721.
- Kirkcaldy, R. D., King, B. A., & Brooks, J. T. (2020). COVID-19 and postinfection immunity: Limited evidence, many remaining questions. *JAMA*.
- Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of internal medicine*, 172(9), 577–582.
- Leung, H. (2020). *What we know about coronavirus immunity and reinfection*. Retrieved June 9, 2020, from <https://time.com/5810454/coronavirus-immunity-reinfection/>
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., Et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England Journal of Medicine*.
- Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490), 489–493.
- Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B., Kinoshita, R., & Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of clinical medicine*, 9(2), 538.
- Mackinnon, A. (2020). *Turkmenistan's secretive strongman remains in denial about the pandemic*. Retrieved August 19, 2020, from <https://foreignpolicy.com/2020/04/10/turkmenistan-coronavirus-pandemic-denial-strongman-berdimuhamedov/>

- Ministero della Salute. (2020). *Coronavirus: Contagion rate R_0 below 1. prudence needed in phase two says ISS*. Retrieved June 11, 2020, from http://www.salute.gov.it/portale/news/p3_2_1_1_1.jsp?lingua=italiano&menu=notizie&p=dalministero&id=4717
- Nebehay, S. (2020). *North Korea testing, quarantining for COVID-19, still says no cases: WHO representative*. Retrieved August 19, 2020, from <https://www.reuters.com/article/us-health-coronavirus-northkorea/north-korea-testing-quarantining-for-covid-19-still-says-no-cases-who-representative-idUSKBN21P3C2>
- Papadopoulos, A. (2018). *Chickenpox: Practice essentials, background, pathophysiology*. Retrieved June 22, 2020, from <https://emedicine.medscape.com/article/1131785-overview/>
- Porta, M. (2014). *A dictionary of epidemiology*. Oxford University Press.
- Rosini, U. (2020). *COVID-19*. Retrieved July 4, 2020, from <https://github.com/pcm-dpc/COVID-19/tree/master/legacy/dati-regioni>
- Schultz, T. (2020). *Why Belgium's death rate is so high: It counts lots of suspected COVID-19 cases*. Retrieved September 15, 2020, from <https://www.npr.org/sections/coronavirus-live-updates/2020/04/22/841005901/why-belgiums-death-rate-is-so-high-it-counts-lots-of-suspected-covid-19-cases>
- Schwarz, G. Et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464.
- Severgnini, C. (2020). *Discorso di Conte in conferenza stampa, le riaperture dal 18 maggio: “corriamo un rischio calcolato”*. Retrieved June 18, 2020, from corriere.it/politica/20_maggio_16/discorso-conte-conferenza-stampa-oggi-decreto-18-maggio-1e810142-9785-11ea-ba09-20ae073bed63.shtml
- Sutherland, J., & Gretler, C. (2020). *WHO now says role of silent virus spreaders remains unclear*. Retrieved June 18, 2020, from <https://www.bloomberg.com/news/articles/2020-06-09/who-says-symptomless-spread-is-rare-in-jolt-to-virus-efforts>
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2), 228.
- WHO. (2019). *What are the International Health Regulations and Emergency Committees?* Retrieved August 19, 2020, from <https://www.who.int/news-room/q-a-detail/what-are-the-international-health-regulations-and-emergency-committees>
- WHO. (2020a). *Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)*. Retrieved August 19, 2020, from [https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-\(2005\)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-\(2019-ncov\)](https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov))

- WHO. (2020b). *WHO director-general's opening remarks at the media briefing on COVID-19 - 11 march 2020*. Retrieved August 19, 2020, from <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>
- Worldometer. (2020). *Italy population*. Retrieved August 3, 2020, from <https://www.worldometers.info/world-population/italy-population/>

Appendices

A Abbreviations

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

Table A.1. Abbreviations for the Italian regions.

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

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

Abbreviation	Full name	First 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 2
SIR model	Standard Inflammatory Response model	Section 4
AIC	Akaike Information Criterion	Section 5
BIC	Bayesian Information Criterion	Section 5

Table A.2 continues on next page

Table A.2 continued from previous page

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

B Tables

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

In Section 7.2, we presented the results from the within-region spread model (7.5):

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

This appendix contains additional tables with results for this model. As is the case for Section 7.2, we use the last 100 observations.

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

Region	Regular model		Modelling undocumented infectives	
	β_{within}	Weekend	β_{within}	Weekend
National (OLS)	0.493*** (21.992)	92.860*** (3.071)	0.423*** (27.417)	379.483*** (2.976)
National (POLS)	0.533*** (27.980)	15.086*** (5.074)	0.458*** (22.981)	90.994*** (4.893)
ABR	0.372*** (4.952)	2.751** (2.091)	0.322*** (6.253)	13.454** (2.249)
BAS	0.063 (0.565)	0.612 (0.772)	0.063 (0.615)	3.279 (0.890)
BZ	0.269*** (3.467)	2.346*** (3.683)	0.219*** (4.175)	7.117*** (3.961)
CAL	0.242** (2.043)	2.189** (2.614)	0.201* (1.917)	12.897*** (2.828)
EMR	0.358*** (11.673)	22.416*** (4.844)	0.287*** (14.768)	91.467*** (5.266)
FVG	0.342*** (7.559)	2.582*** (3.476)	0.252*** (8.891)	8.578*** (3.508)
LAZ	0.494*** (10.069)	10.172*** (3.441)	0.424*** (11.273)	57.018*** (3.610)
LIG	0.382*** (11.931)	4.179 (1.430)	0.322*** (13.539)	17.706 (1.242)
LOM	0.566*** (16.318)	26.949 (1.271)	0.502*** (16.989)	114.802 (1.072)

Table B.1 continues on next page

Table B.1 continued from previous page

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

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

Table B.2. Estimates from Within-Region Spread Model per region with model selection by AIC versus BIC. Estimates are given with t -statistics in parentheses. Data spans May 9 until August 16, 2020 (100 days). Undocumented infectives are not modelled.

Region	Model selection with AIC		Model selection with BIC	
	β_{within}	Weekend	β_{within}	Weekend
National (OLS)	0.493*** (21.992)	92.860*** (3.071)	0.493*** (21.992)	92.860*** (3.071)
National (POLS)	0.533*** (27.980)	15.086*** (5.074)	0.533*** (27.980)	15.086*** (5.074)
ABR	0.372*** (4.952)	2.751** (2.091)	0.440*** (6.387)	
BAS	0.073 (0.665)		0.073 (0.665)	
BZ	0.269*** (3.468)	2.346*** (3.683)	0.269*** (3.468)	2.346*** (3.683)
CAL	0.242** (2.043)	2.189** (2.614)	0.242** (2.043)	2.189** (2.614)
EMR	0.358*** (11.673)	22.416*** (4.844)	0.358*** (11.673)	22.416*** (4.844)
FVG	0.342*** (7.559)	2.582*** (3.476)	0.342*** (7.559)	2.582*** (3.476)

Table B.2 continues on next page

Table B.2 continued from previous page

Region	Model selection with AIC		Model selection with BIC	
	β_{within}	Weekend	β_{within}	Weekend
LAZ	0.494*** (10.069)	10.172*** (3.441)	0.494*** (10.069)	10.172*** (3.441)
LIG	0.382*** (11.931)	4.179 (1.430)	0.399*** (13.271)	
LOM	0.583*** (18.205)		0.583*** (18.205)	
MAR	0.436*** (9.376)	2.697** (2.144)	0.473*** (10.775)	
MOL	0.140 (1.595)	1.855** (2.456)	0.140 (1.595)	1.855** (2.456)
PIE	0.352*** (23.006)	7.449* (1.896)	0.363*** (25.048)	
PUG	0.376*** (6.165)	3.804** (2.581)	0.376*** (6.165)	3.804** (2.581)
SAR	0.437*** (4.417)		0.437*** (4.417)	
TOS	0.281*** (6.666)	8.444*** (3.854)	0.281*** (6.666)	8.444*** (3.854)
UMB	0.436*** (3.936)	1.428*** (3.168)	0.436*** (3.936)	1.428*** (3.168)
VDA	0.290*** (4.341)		0.290*** (4.341)	
VEN	0.489*** (6.748)	11.951* (1.920)	0.542*** (7.958)	

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

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

In Section 8.2, we presented the results from the within and between-region spread model (8.1):

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

This appendix contains additional tables with results for this model. As is the case for Section 8.2, we use the last 100 observations.

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

Region	Regular model			Modelling undocumented infectives		
	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
ABR	-0.011 (-0.079)	$5.871 \times 10^{-3***}$ (3.399)	2.394* (1.909)	-0.047 (-0.495)	$5.943 \times 10^{-3***}$ (4.455)	12.494** (2.279)
BAS	0.039 (0.346)	6.89911×10^{-4} (1.059)	0.2450 (0.283)	0.032 (0.298)	5.547×10^{-4} (1.132)	1.711 (0.435)
BZ	0.246** (2.355)	2.043×10^{-4} (0.317)	2.305*** (3.534)	0.201** (2.559)	1.018×10^{-4} (0.310)	7.030*** (3.848)
CAL	0.186 (1.460)	7.936×10^{-4} (1.171)	1.893** (2.168)	0.128 (1.088)	8.270×10^{-4} (1.341)	11.514** (2.472)
EMR	0.365*** (4.184)	-8.645×10^{-4} (-0.080)	22.410*** (4.817)	0.171*** (2.635)	0.015* (1.873)	90.384*** (5.267)
FVG	0.114* (1.707)	$3.701 \times 10^{-3***}$ (4.346)	1.759** (2.481)	0.034 (0.929)	$3.026 \times 10^{-3***}$ (7.601)	4.330** (2.139)
LAZ	0.342*** (2.671)	7.364×10^{-3} (1.282)	10.573*** (3.568)	0.238** (2.221)	0.010* (1.853)	61.107*** (3.879)
LIG	-0.025 (-0.357)	0.033*** (6.260)	0.345 (0.135)	-0.022 (-0.391)	0.029*** (6.413)	1.408 (0.115)
LOM	0.326*** (4.878)	0.220*** (4.081)	14.034 (0.704)	0.283*** (4.421)	0.186*** (3.799)	65.410 (0.646)
MAR	0.245** (2.521)	$4.505 \times 10^{-3**}$ (2.229)	2.150* (1.710)	0.181** (2.312)	$4.644 \times 10^{-3***}$ (2.907)	10.483* (1.921)
MOL	-0.087 (-1.186)	$4.020 \times 10^{-3***}$ (8.188)	0.163 (0.264)	-0.080 (-1.177)	$4.559 \times 10^{-3***}$ (9.250)	0.981 (0.257)
PIE	0.104*** (3.635)	0.064*** (9.363)	0.260 (0.088)	0.091*** (3.856)	0.056*** (9.514)	0.933 (0.066)
PUG	0.120 (1.051)	$5.525 \times 10^{-3***}$ (2.629)	3.083** (2.115)	0.035 (0.393)	$8.078 \times 10^{-3***}$ (3.896)	18.908** (2.303)
SAR	0.397*** (2.835)	-3.065×10^{-5} (-0.048)	0.803 (1.237)	0.252** (2.468)	1.831×10^{-4} (0.321)	5.062 (1.532)
TOS	0.051 (0.605)	0.010*** (3.107)	8.023*** (3.812)	0.030 (0.535)	$9.454 \times 10^{-3***}$ (4.141)	33.901*** (4.095)
UMB	0.427*** (2.981)	4.689×10^{-5} (0.102)	1.410*** (2.904)	0.235** (2.467)	2.449×10^{-4} (0.876)	4.877*** (2.832)
VDA	0.023 (0.256)	$1.113 \times 10^{-3***}$ (3.937)	-6.610×10^{-3} (-0.024)	0.055 (0.769)	$7.107 \times 10^{-4***}$ (4.169)	-6.958×10^{-3} (-7.896 $\times 10^{-3}$)
VEN	0.904*** (6.990)	-0.033*** (-3.771)	16.184*** (2.721)	0.434*** (4.370)	-7.954×10^{-3} (-1.637)	43.868*** (2.951)

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

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

Region	Model selection with AIC			Model selection with BIC		
	β_{within}	$\beta_{between}$	Weekend	β_{within}	$\beta_{between}$	Weekend
ABR	-0.011 (-0.079)	$5.871 \times 10^{-3***}$ (3.399)	2.394* (1.909)	0.030 (0.226)	$6.147 \times 10^{-3***}$ (3.524)	
BAS	0.040 (0.357)	7.638×10^{-4} (1.286)		0.040 (0.357)	7.638×10^{-4} (1.286)	
BZ	0.246** (2.355)	2.043×10^{-4} (0.317)	2.305*** (3.534)	0.201** (2.559)	1.018×10^{-4} (0.310)	7.030*** (3.848)
CAL	0.186 (1.460)	7.936×10^{-4} (1.171)	1.893** (2.168)	0.249* (1.967)	$1.219 \times 10^{-3*}$ (1.845)	
EMR	0.365*** (4.184)	-8.645×10^{-4} (-0.080)	22.410*** (4.817)	0.365*** (4.184)	-8.645×10^{-4} (-0.080)	22.410*** (4.817)
FVG	0.114* (1.707)	$3.701 \times 10^{-3***}$ (4.346)	1.759** (2.481)	0.114* (1.707)	$3.701 \times 10^{-3***}$ (4.346)	1.759** (2.481)
LAZ	0.342*** (2.671)	7.364×10^{-3} (1.283)	10.573*** (3.568)	0.342*** (2.671)	7.364×10^{-3} (1.283)	10.573*** (3.568)
LIG	-0.026 (-0.372)	0.033*** (6.515)		-0.026 (-0.372)	0.033*** (6.515)	
LOM	0.329*** (4.929)	0.226*** (4.257)		0.329*** (4.929)	0.226*** (4.257)	
MAR	0.245** (2.521)	$4.505 \times 10^{-3**}$ (2.229)	2.150* (1.710)	0.245** (2.496)	$5.179 \times 10^{-3**}$ (2.587)	
MOL	-0.085 (-1.168)	$4.063 \times 10^{-3***}$ (8.821)		-0.085 (-1.168)	$4.063 \times 10^{-3***}$ (8.821)	
PIE	0.104*** (3.665)	0.065*** (9.767)		0.104*** (3.665)	0.065*** (9.767)	
PUG	0.120 (1.051)	$5.525 \times 10^{-3***}$ (2.629)	3.083** (2.115)	0.127 (1.099)	$6.362 \times 10^{-3***}$ (3.029)	
SAR	0.407*** (2.903)	1.846×10^{-3} (0.299)		0.407*** (2.903)	1.846×10^{-3} (0.299)	
TOS	0.051 (0.605)	0.010*** (3.107)	8.023*** (3.812)	0.051 (0.605)	0.010*** (3.107)	8.023*** (3.812)
UMB	0.427*** (2.981)	4.689×10^{-5} (0.102)	1.410*** (2.904)	0.427*** (2.981)	4.689×10^{-5} (0.102)	1.410*** (2.904)
VDA	0.023 (0.259)	$1.111 \times 10^{-3***}$ (4.193)		0.023 (0.259)	$1.111 \times 10^{-3***}$ (4.193)	
VEN	0.904*** (6.990)	-0.033*** (-3.771)	16.184*** (2.721)	0.904*** (6.990)	-0.033*** (-3.771)	16.184*** (2.721)

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

C Figures

C.1 Figures for Section 2: Problem description

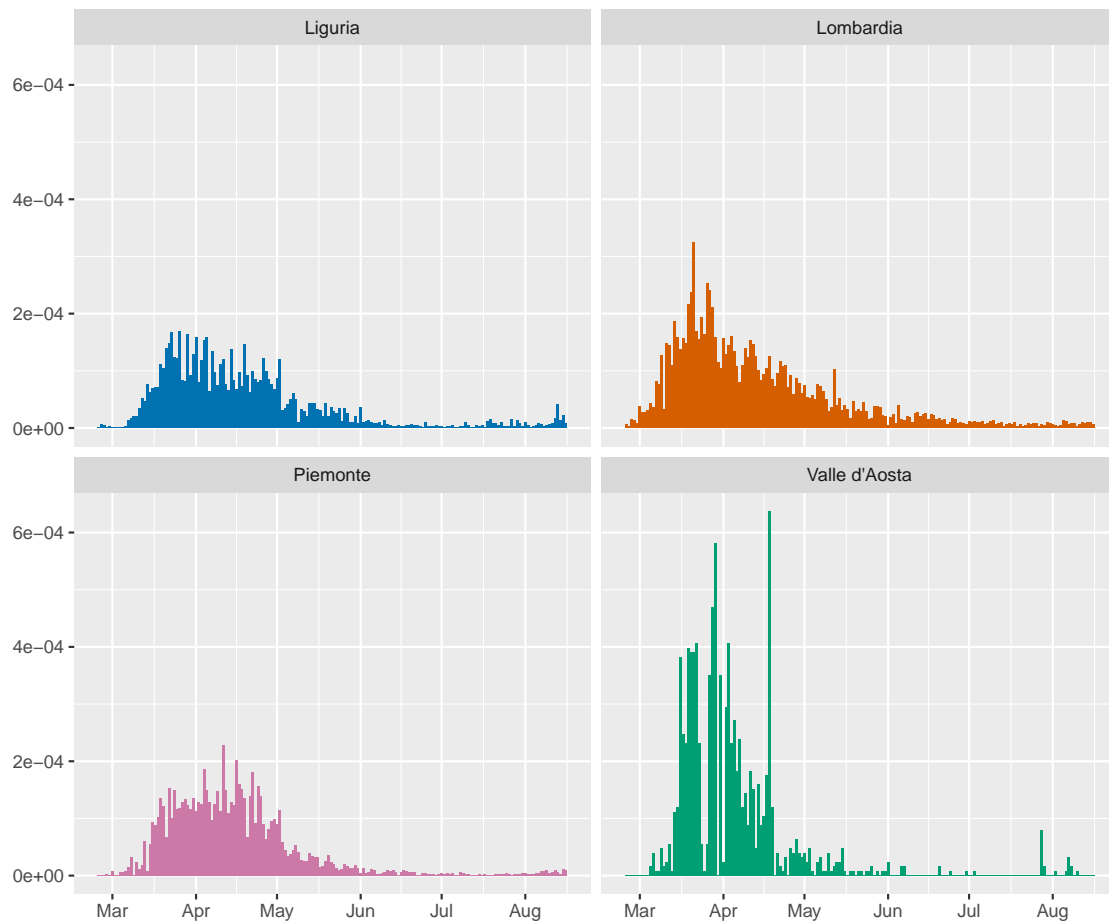


Figure C.1. Incidence rate per region for the *Nord-Ovest* (North-West) NUTS 1 region

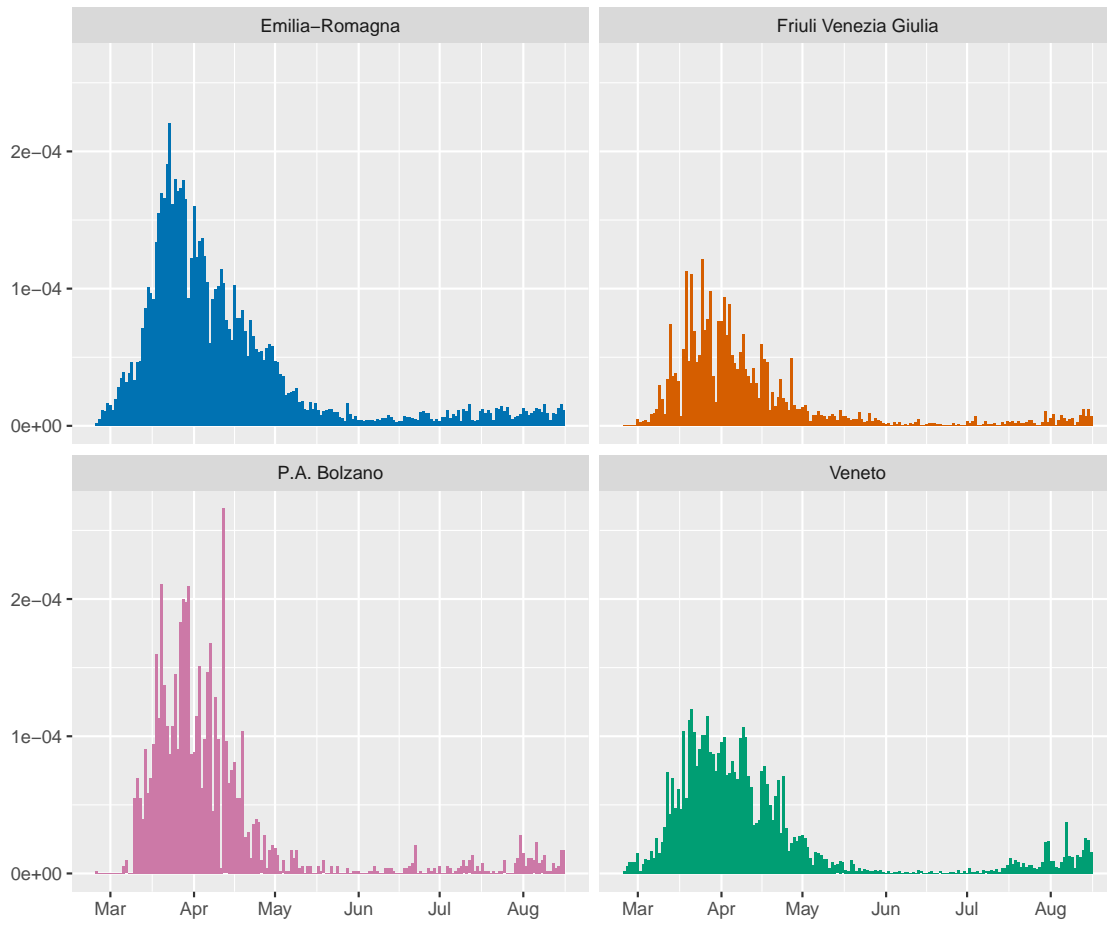


Figure C.2. Incidence rate per region for the *Nord-Est* (North-East) NUTS 1 region

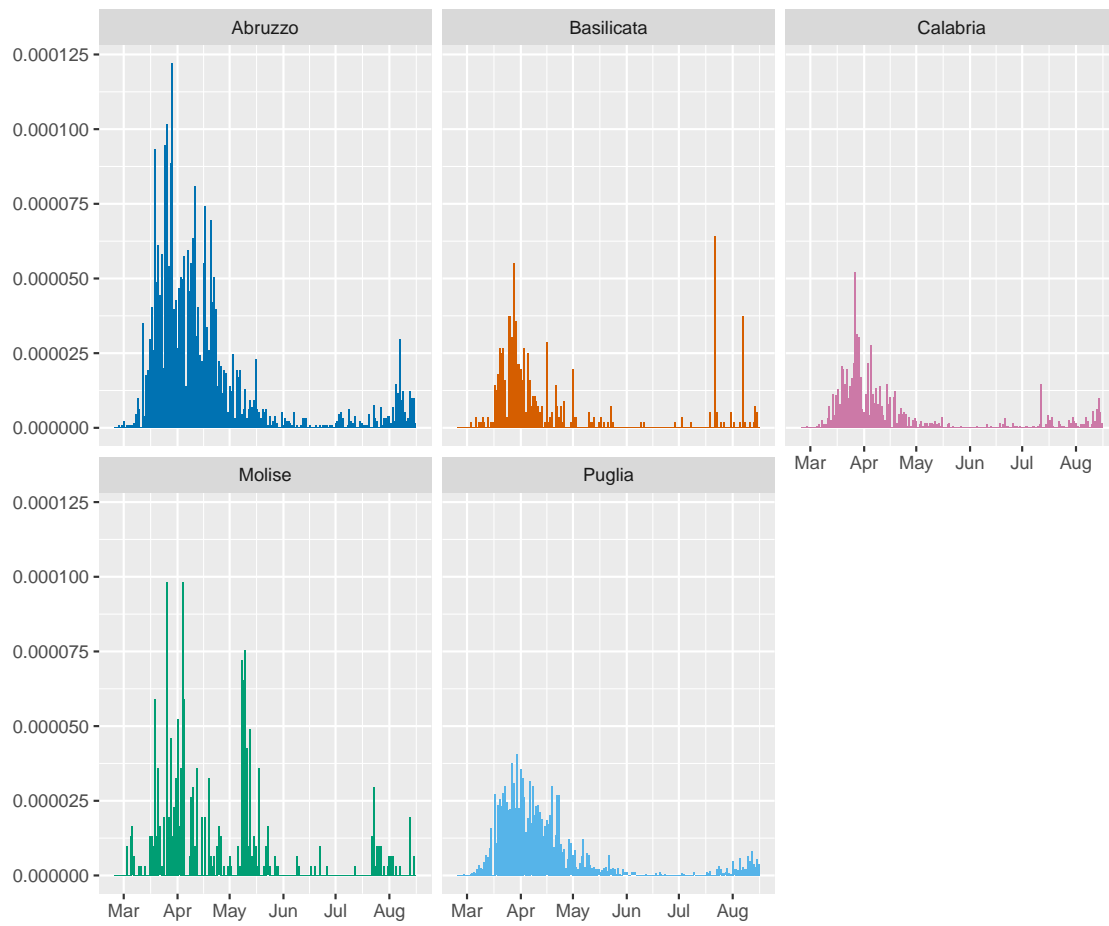


Figure C.3. Incidence rate per region for the *Sud* (South) NUTS 1 region

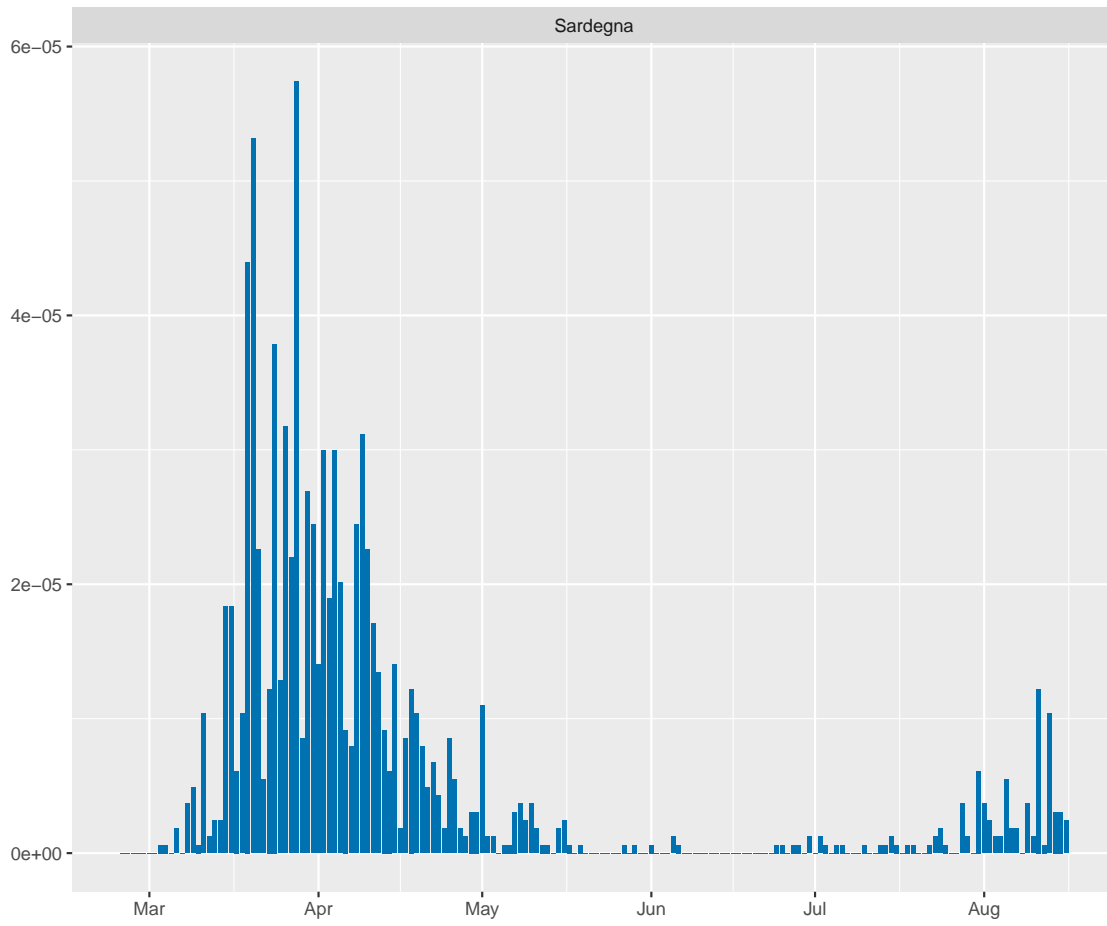


Figure C.4. Incidence rate per region for the *Isole* (Islands) NUTS 1 region

C.2 Figures for Section 3: Dataset

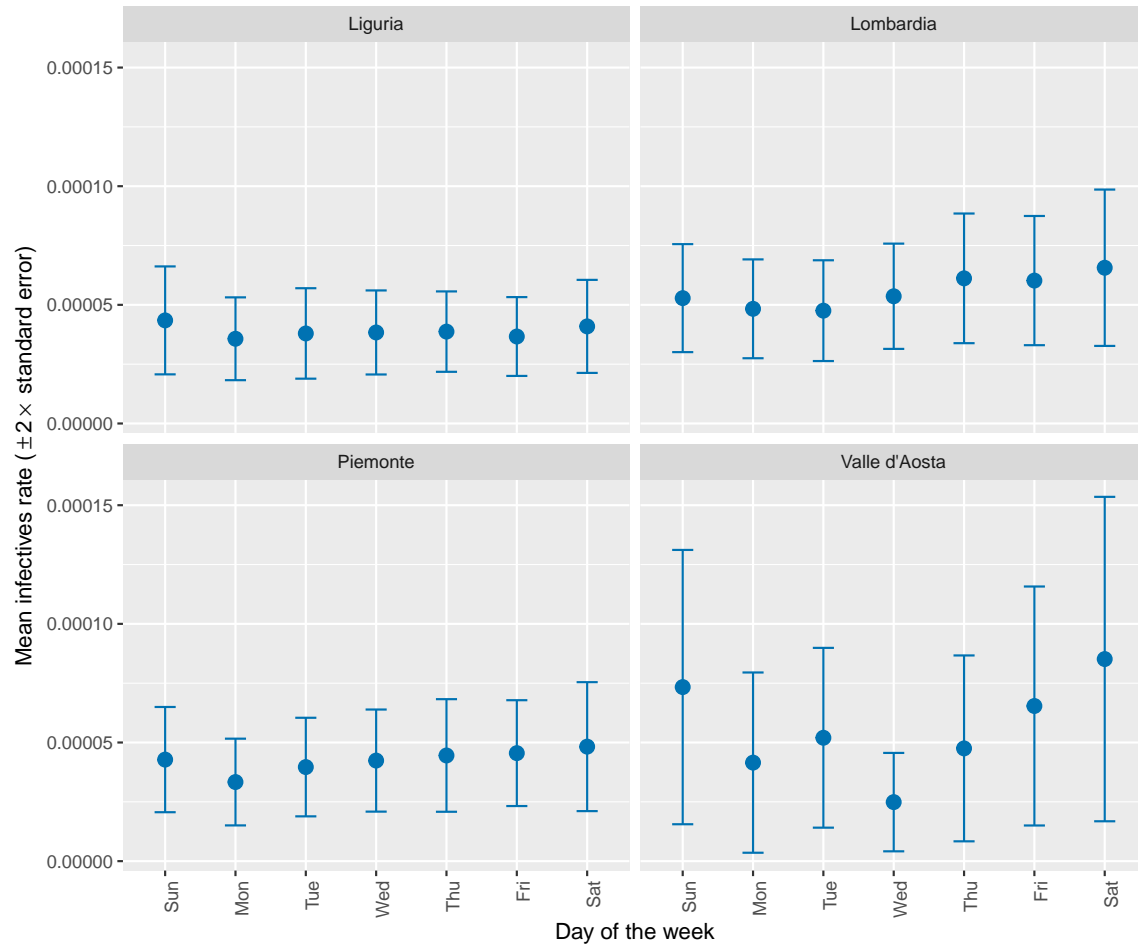


Figure C.5. Incidence rate per NUTS 2 region per day of the week for the *Nord-Ovest* (North-West) NUTS 1 region

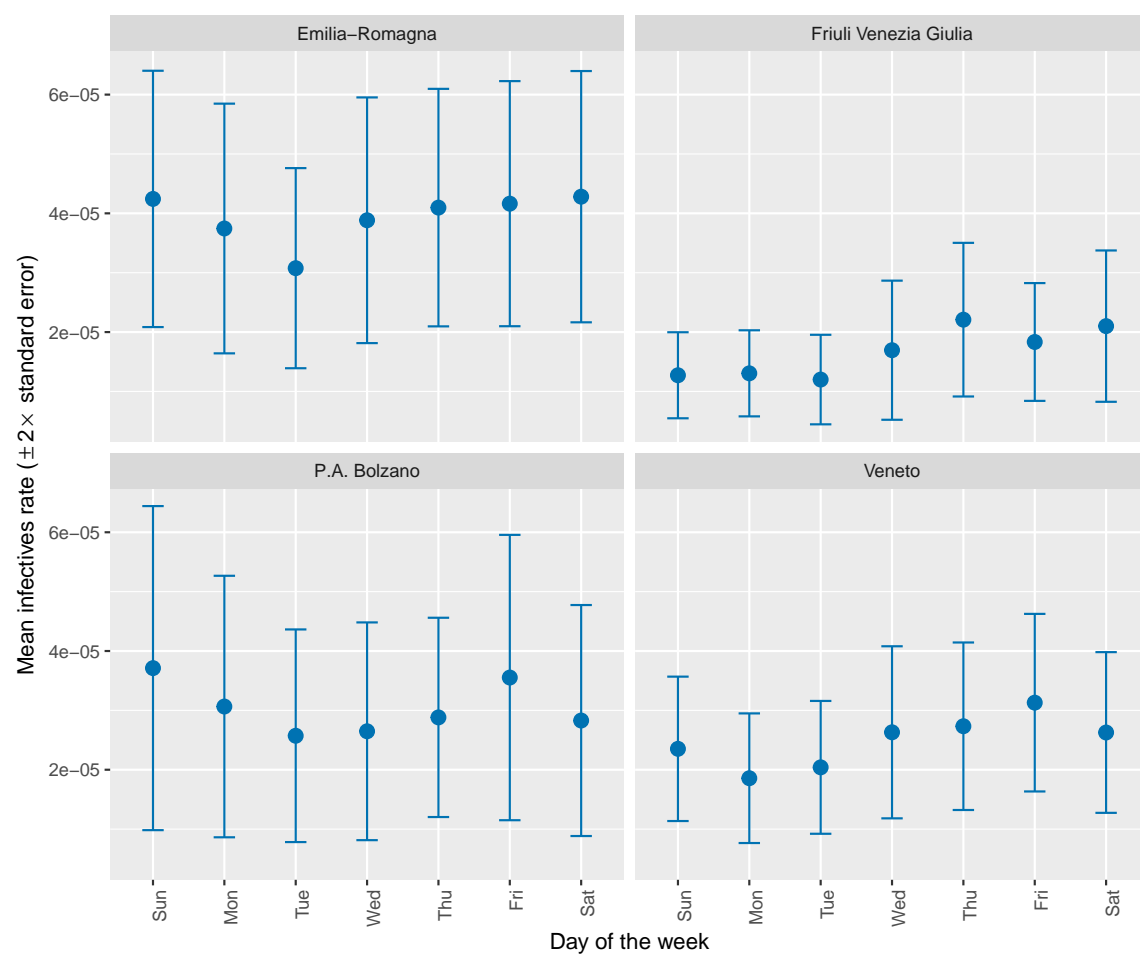


Figure C.6. Incidence rate per NUTS 2 region per day of the week for the *Nord-Est* (North-East) NUTS 1 region

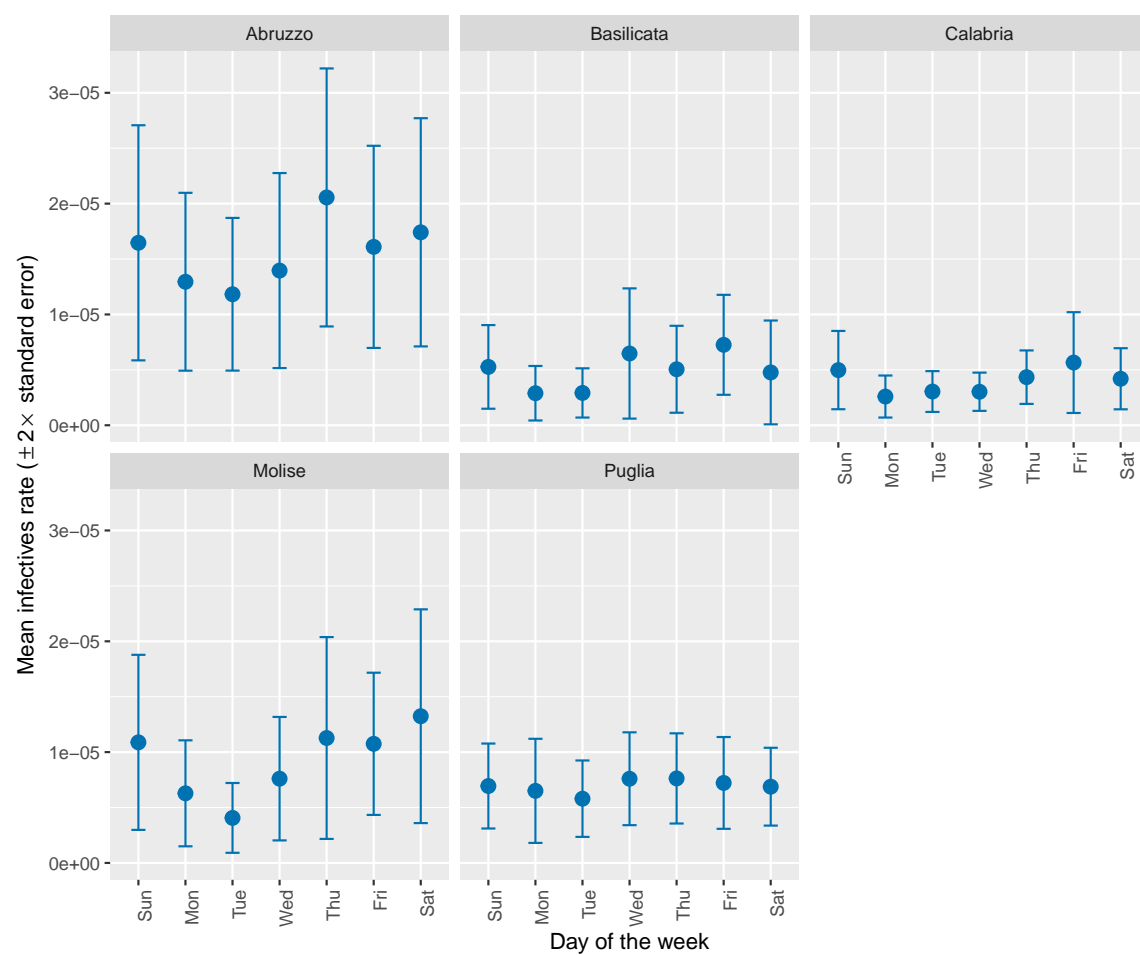


Figure C.7. Incidence rate per NUTS 2 region per day of the week for the *Sud* (South) NUTS 1 region

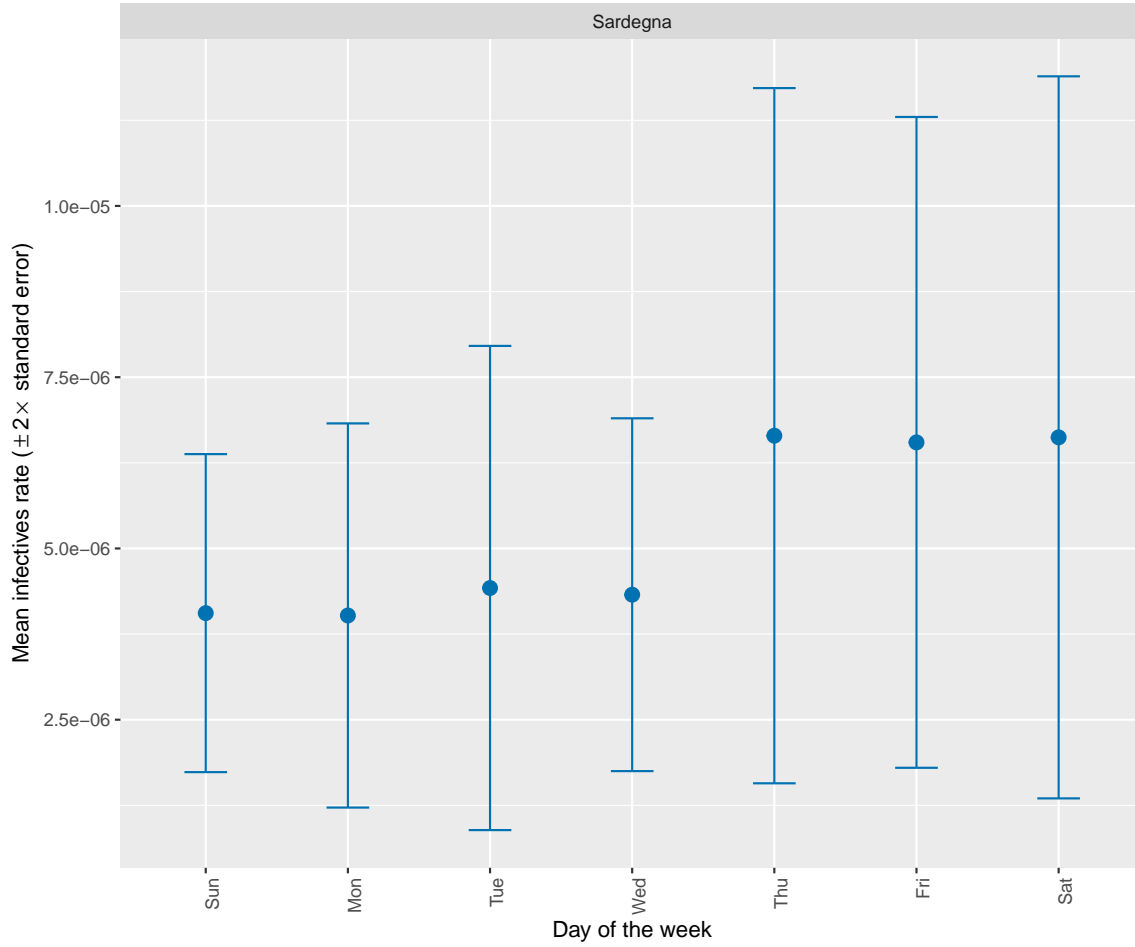
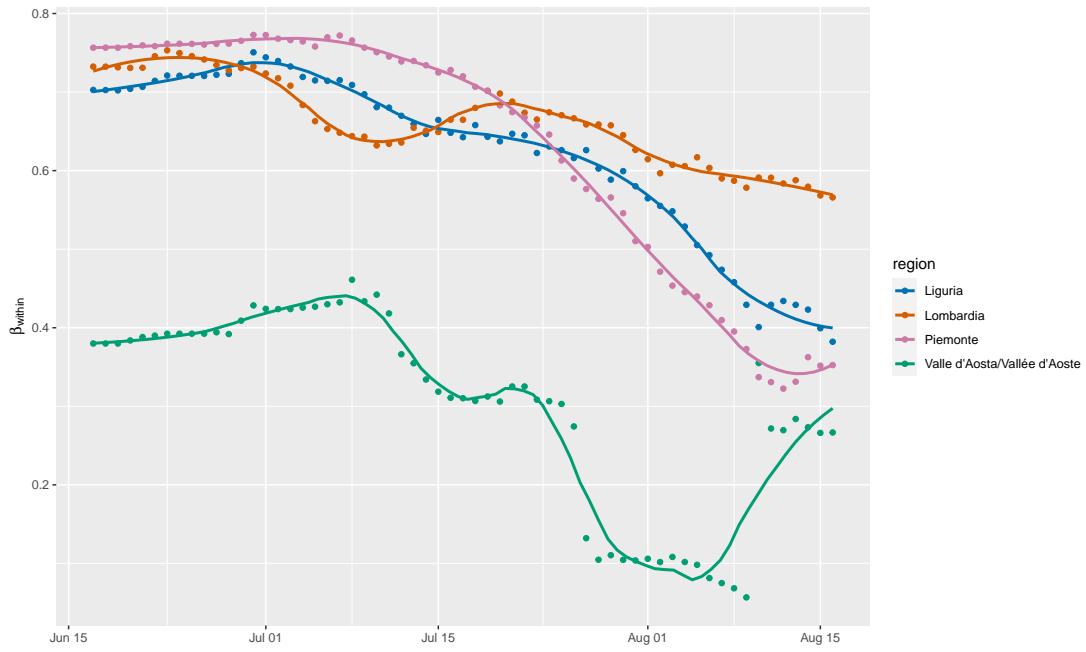


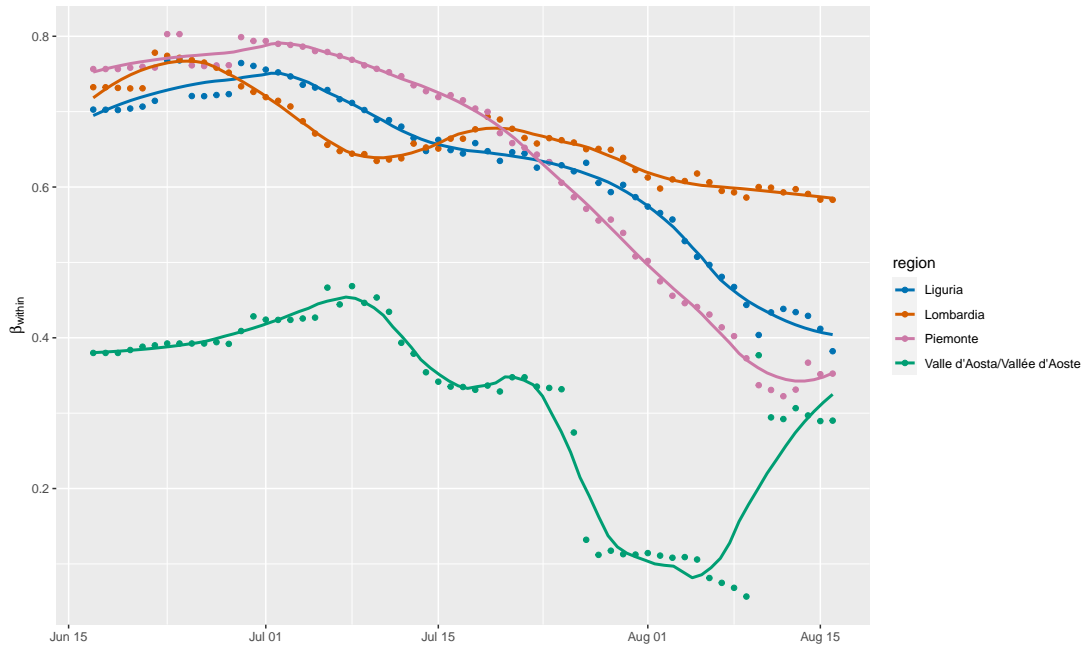
Figure C.8. Incidence rate per NUTS 2 region per day of the week for the *Isole* (Islands) NUTS 1 region

C.3 Figures for Section 7: Within-Region Spread Model

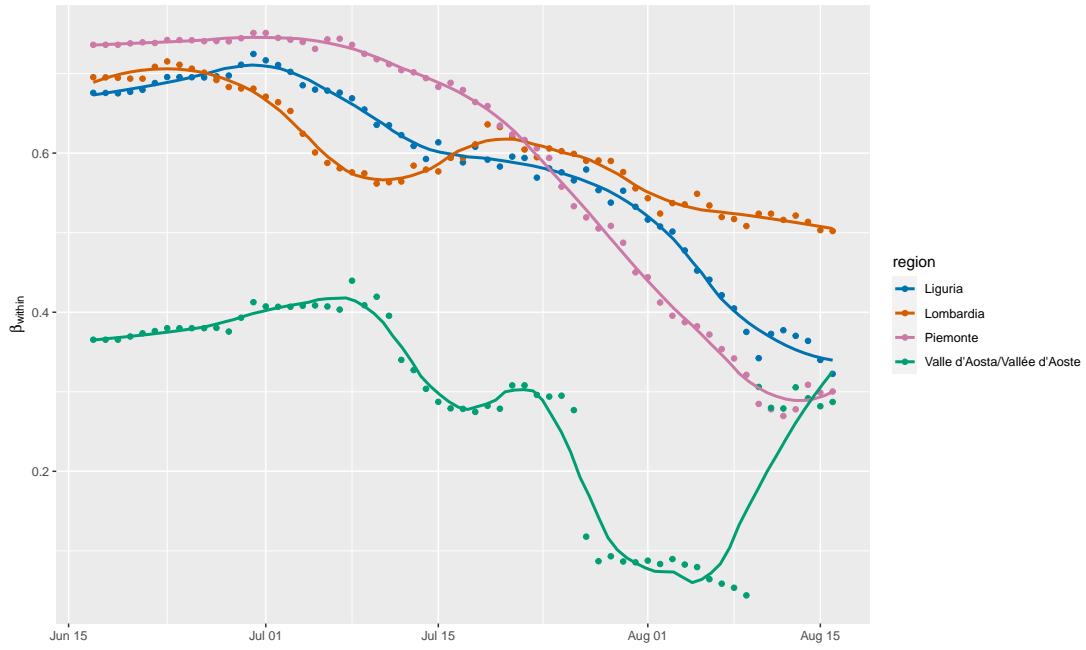
In Section 7.2 we presented the plots of β_{within} over time for the *Nord-Est* (North-East) NUTS 1 region for Within-Region Spread Model. In this section, we present the plots for the other NUTS 1 regions. As is the case for Section 7.2, we use the last 100 observations.



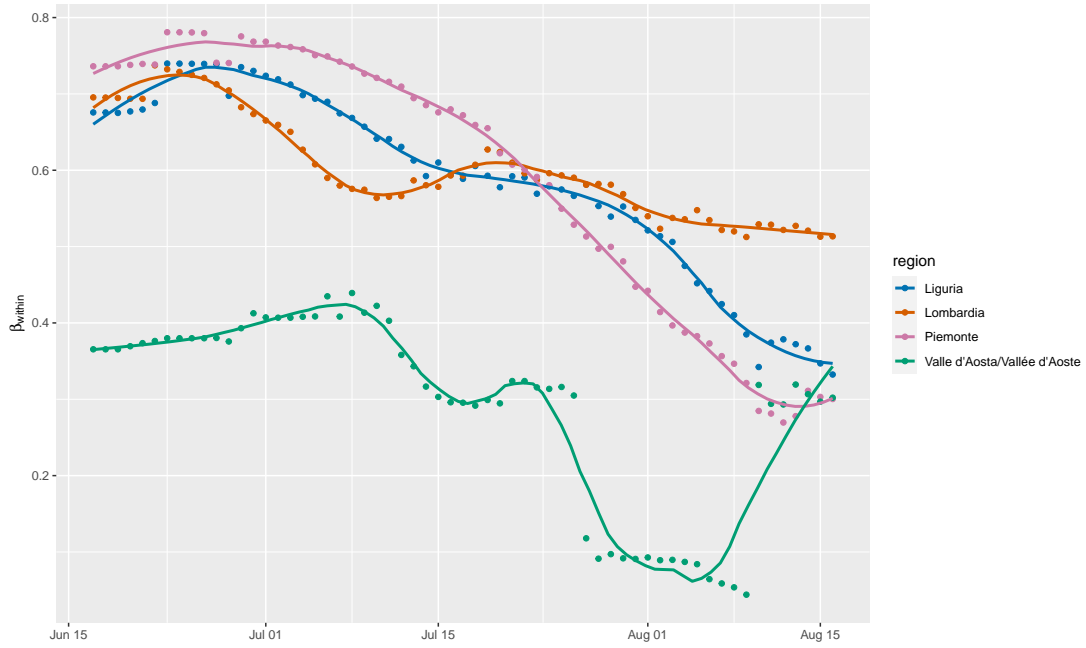
(a) Without model selection



(b) With model selection by AIC

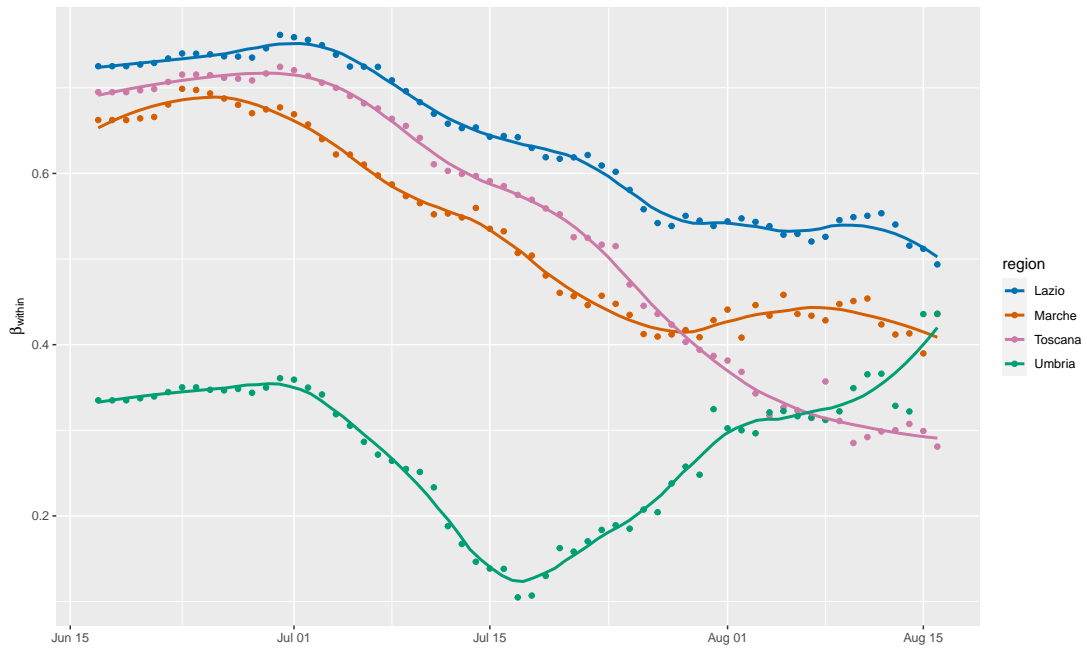


(c) Without model selection;
including undocumented infectives

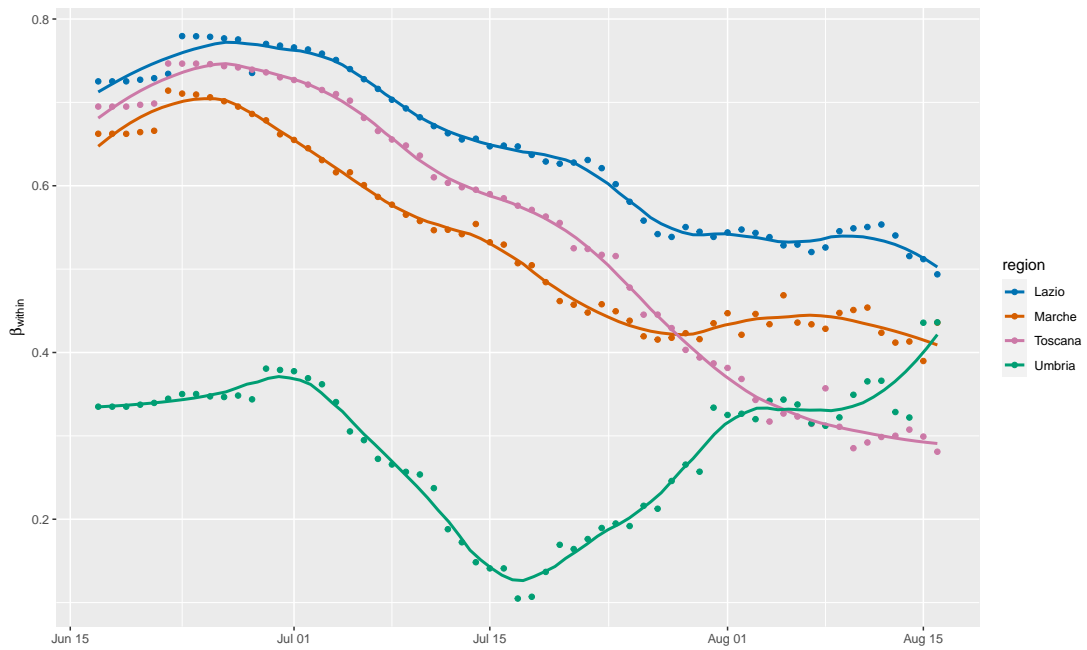


(d) With model selection by AIC;
including undocumented infectives

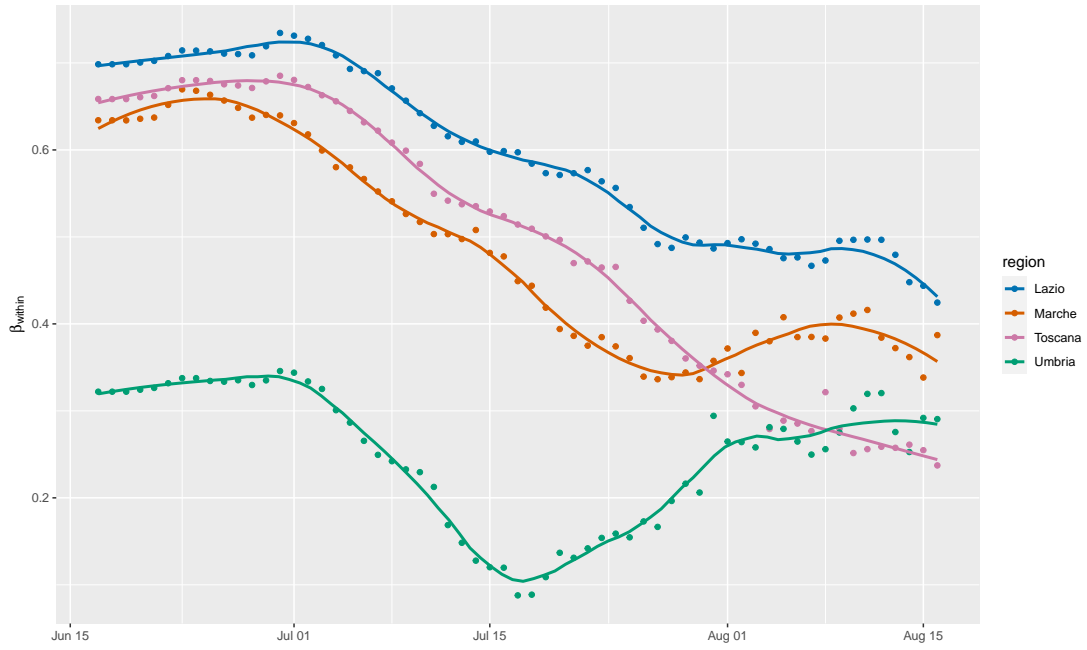
Figure C.9. Progression of β_{within} over time for the *Nord-Ovest* (North-West) NUTS 1 region



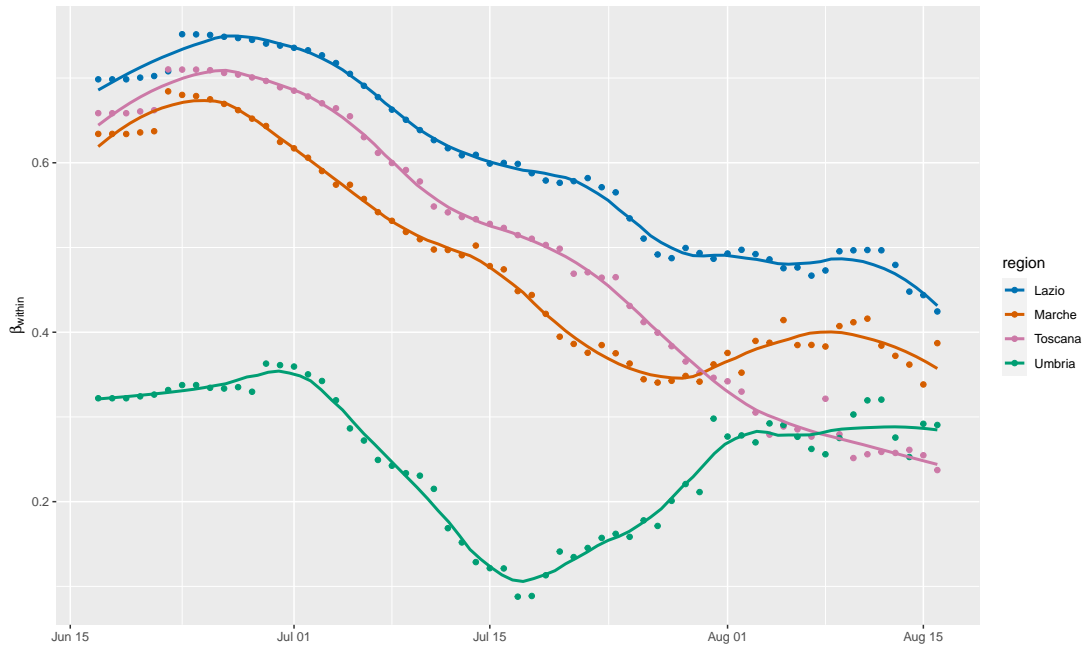
(a) Without model selection



(b) With model selection by AIC

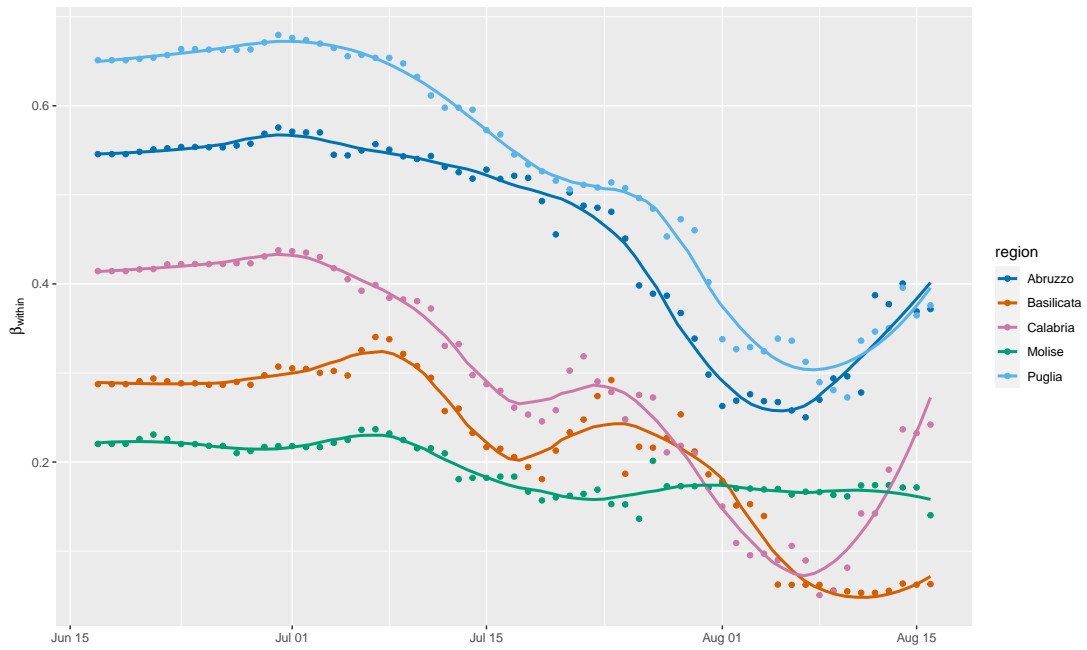


(c) Without model selection;
including undocumented infectives

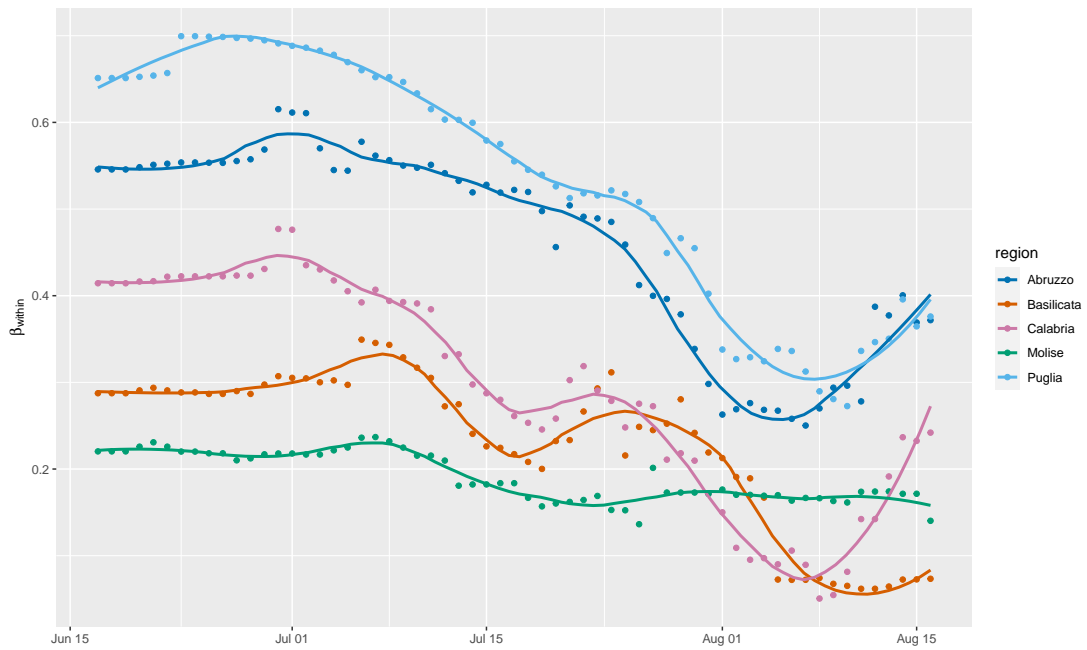


(d) With model selection by AIC;
including undocumented infectives

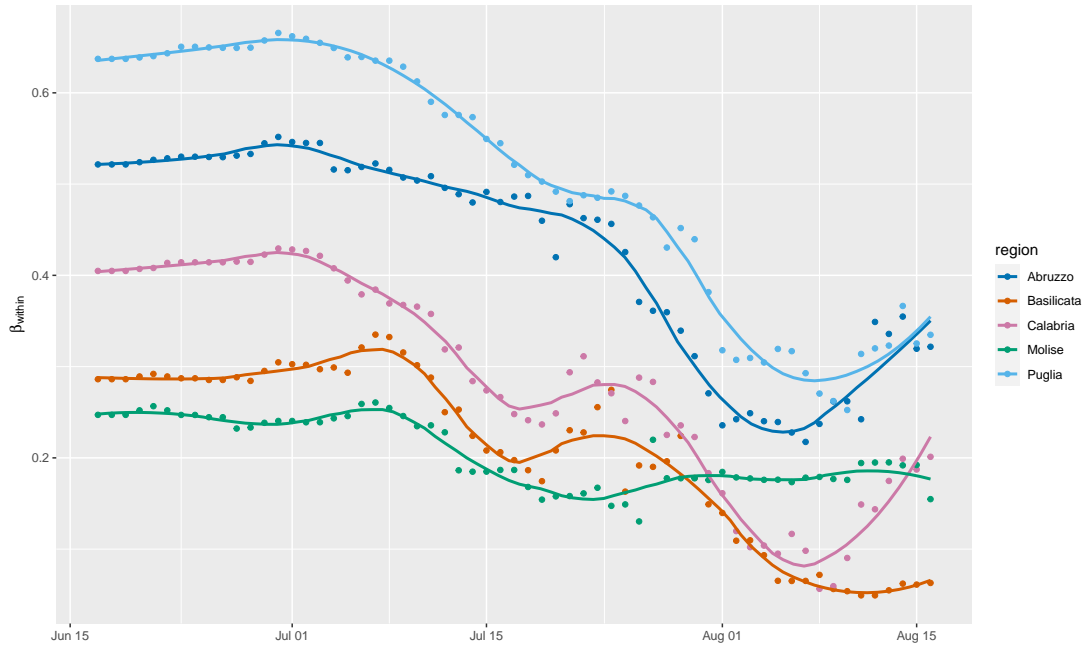
Figure C.10. Progression of β_{within} over time for the *Centro (IT)* (Centre) NUTS 1 region



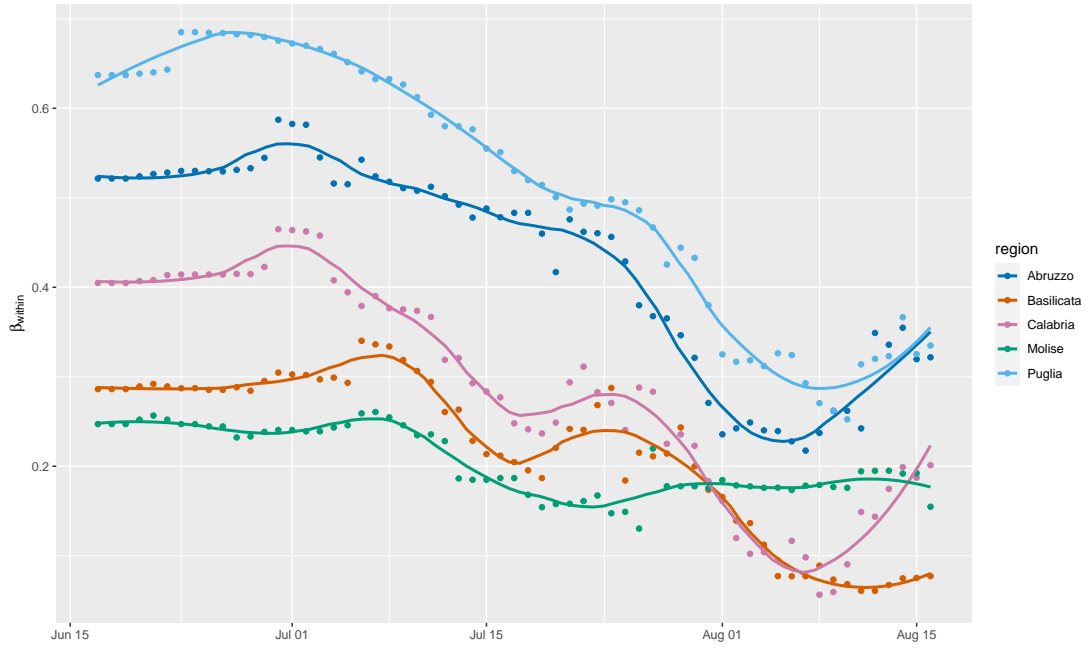
(a) Without model selection



(b) With model selection by AIC

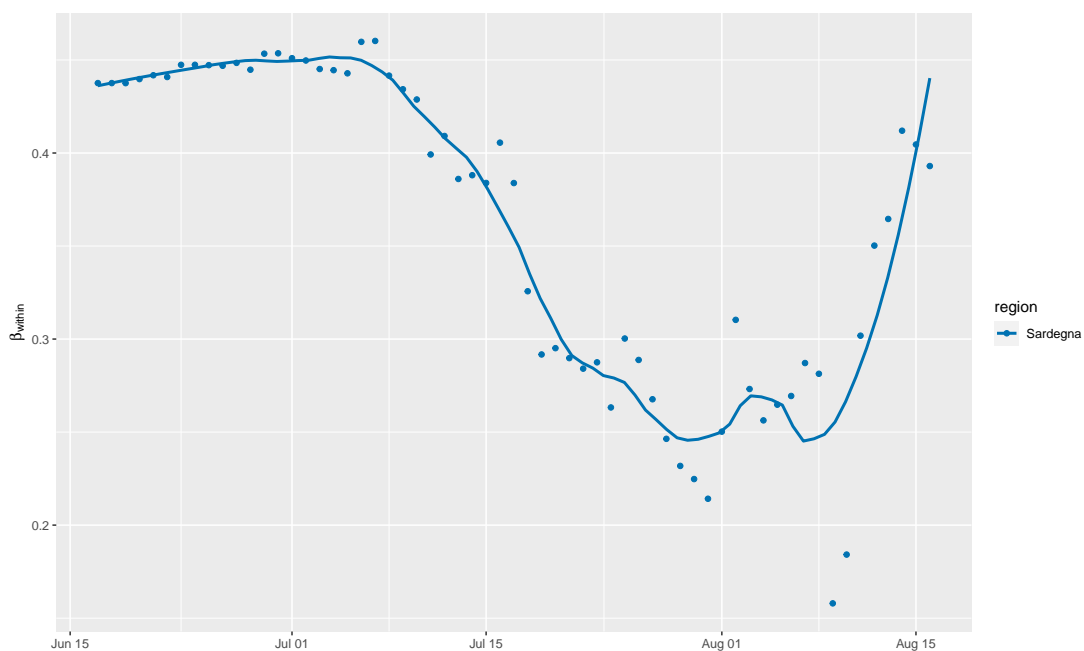


(c) Without model selection;
including undocumented infectives

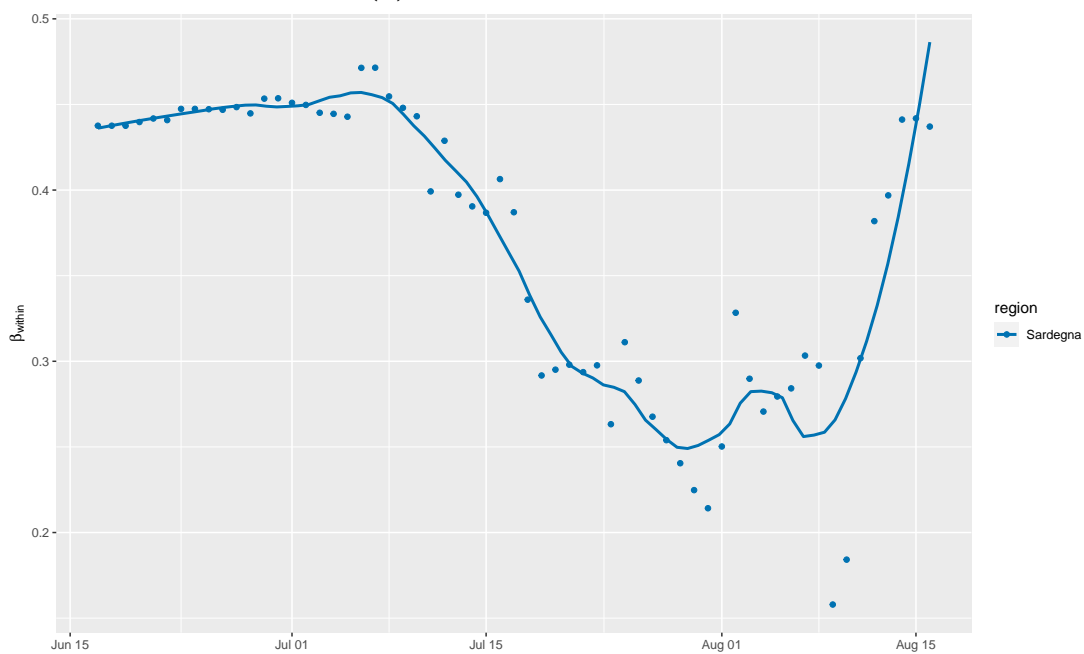


(d) With model selection by AIC;
including undocumented infectives

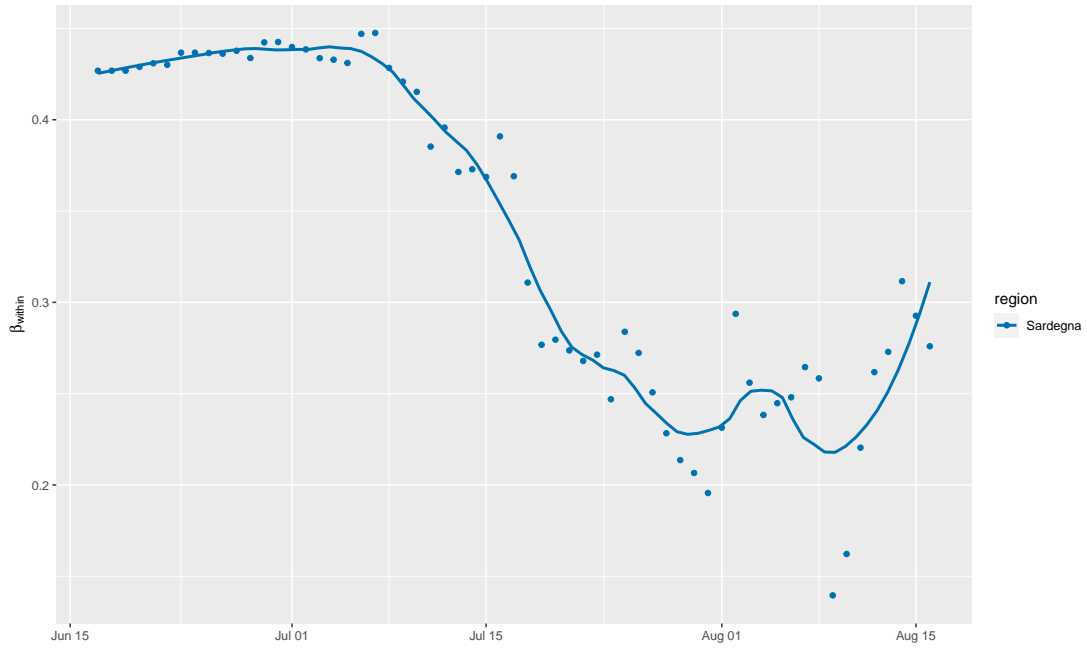
Figure C.11. Progression of β_{within} over time for the *Sud* (South) NUTS 1 region



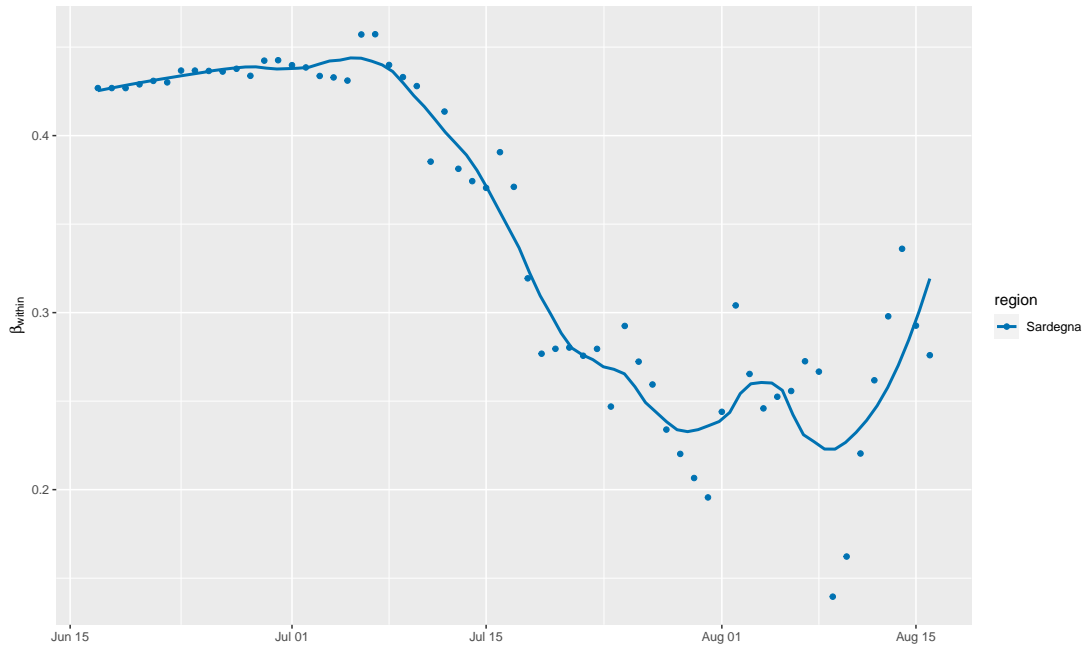
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives

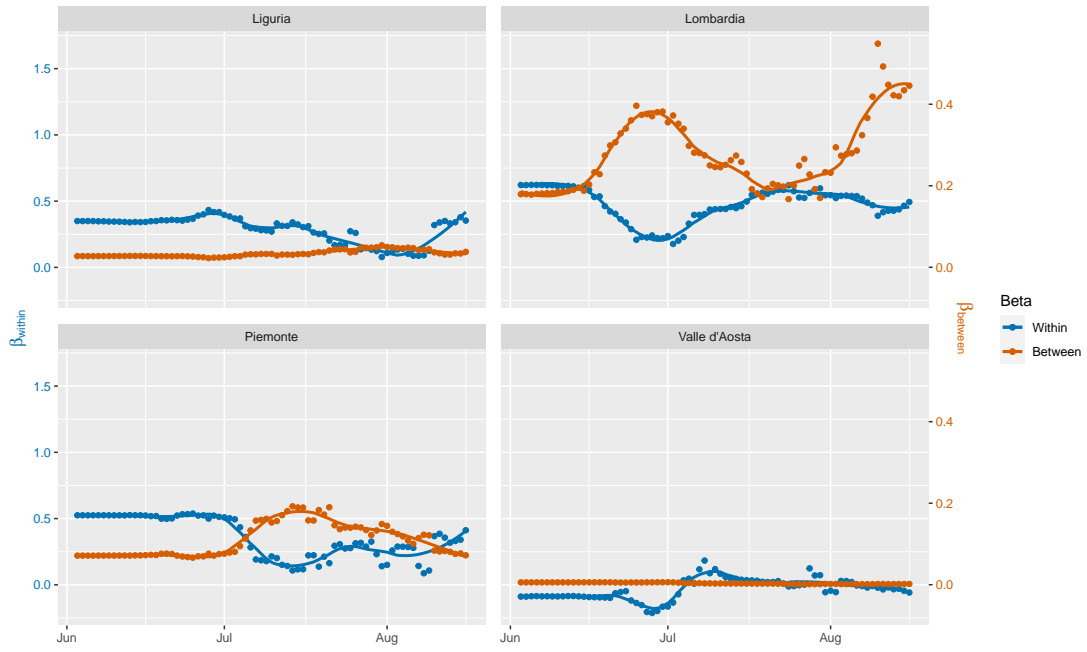


(d) With model selection by AIC;
including undocumented infectives

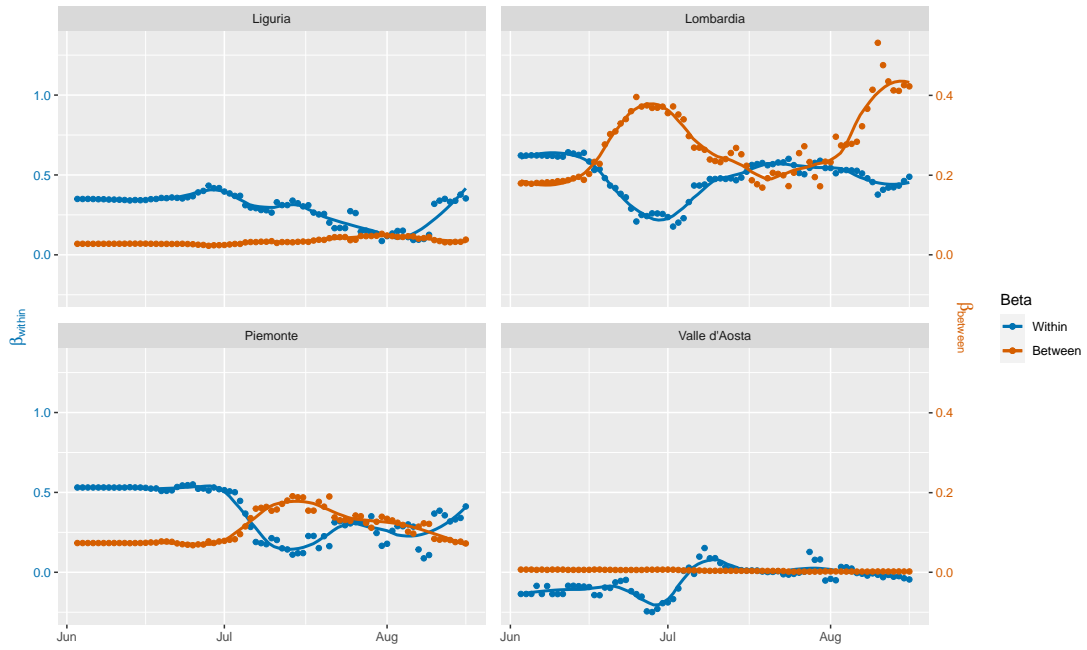
Figure C.12. Progression of β_{within} over time for the *Isole* (Islands) NUTS 1 region

C.4 Figures for Section 8: Within and Between-Region Spread Model

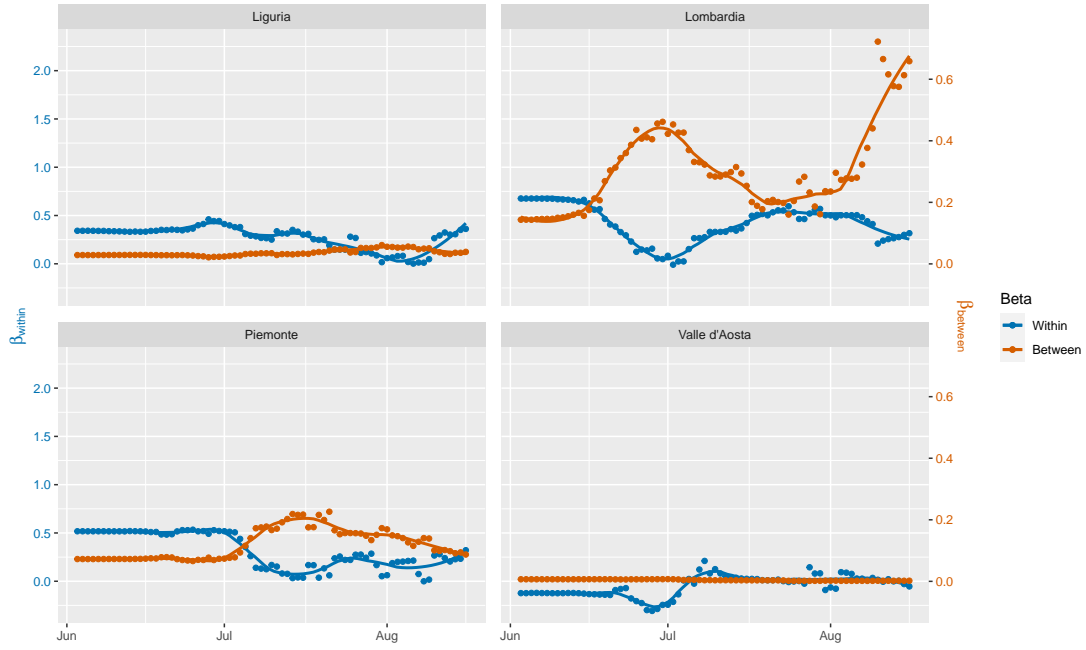
In Section 8.2 we presented the plots of β_{within} and $\beta_{between}$ over time for the *Nord-Est* (North-East) NUTS 1 region for Within and Between-Region Spread Model. In this section, we present the plots for the other NUTS 1 regions. As is the case for Section 8.2, we use the last 100 observations.



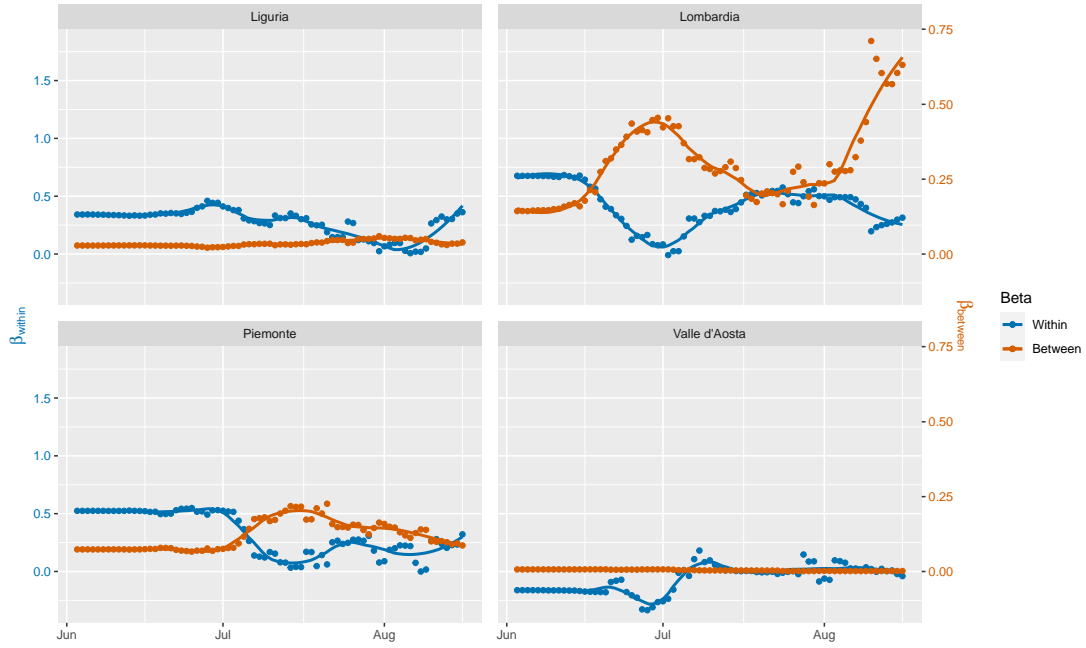
(a) Without model selection



(b) With model selection by AIC

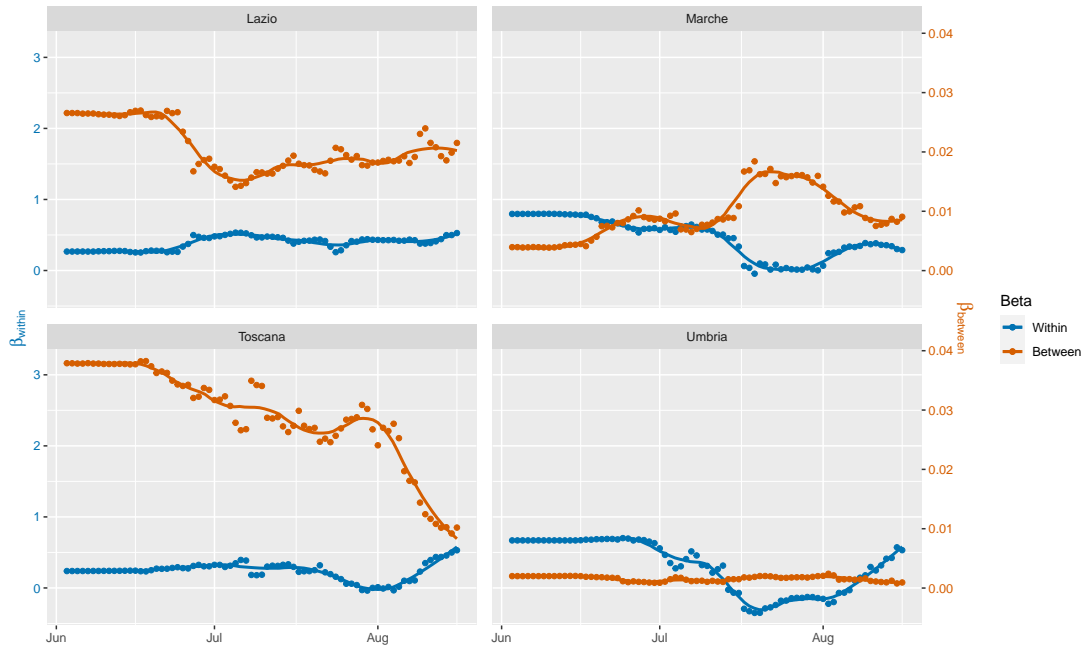


(c) Without model selection;
including undocumented infectives

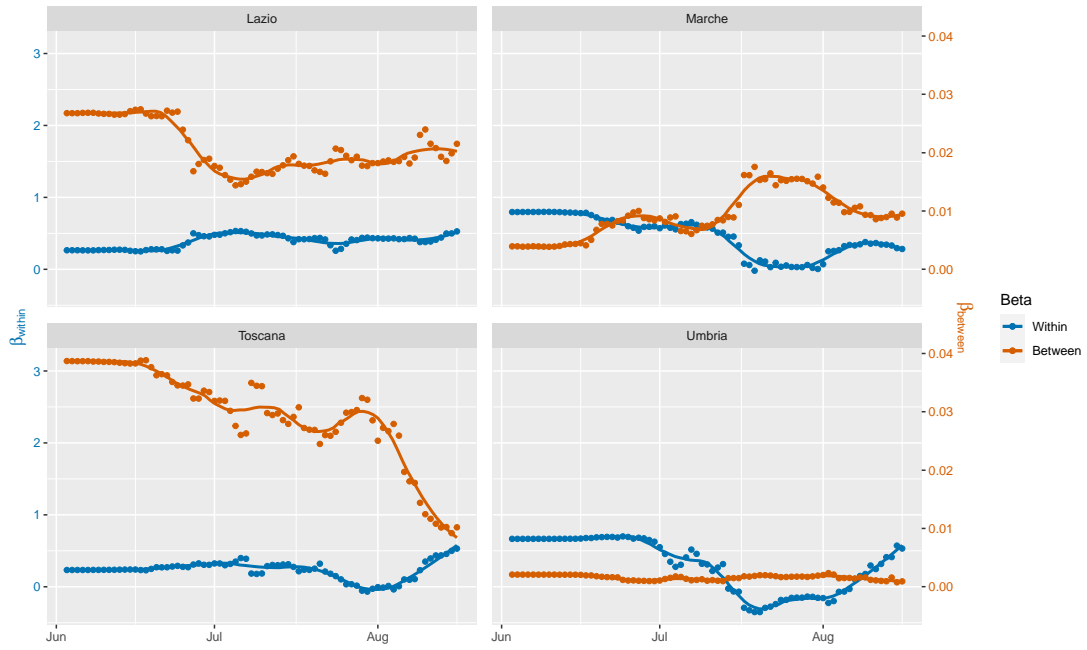


(d) With model selection by AIC;
including undocumented infectives

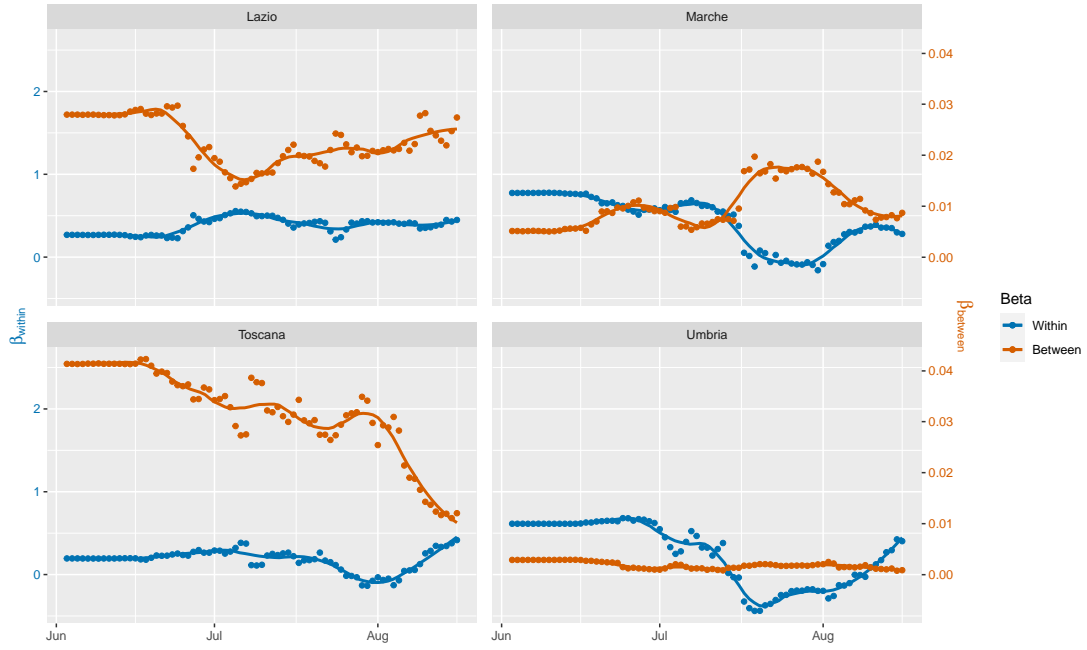
Figure C.13. Progression of β_{within} and $\beta_{between}$ over time for the *Nord-Ovest* (North-West) NUTS 1 region



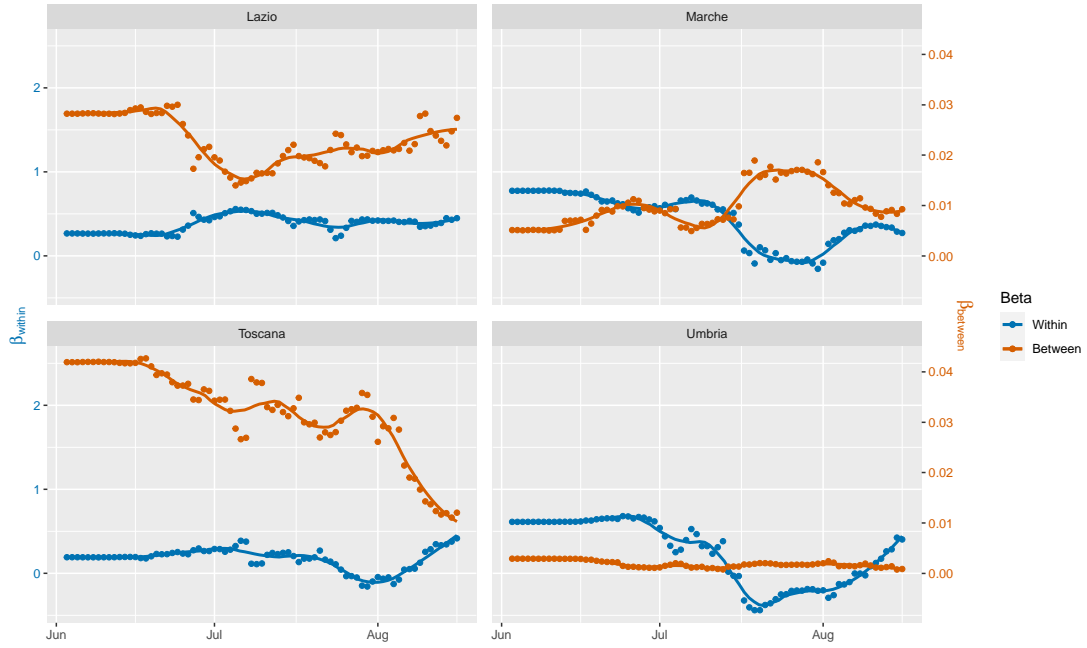
(a) Without model selection



(b) With model selection by AIC

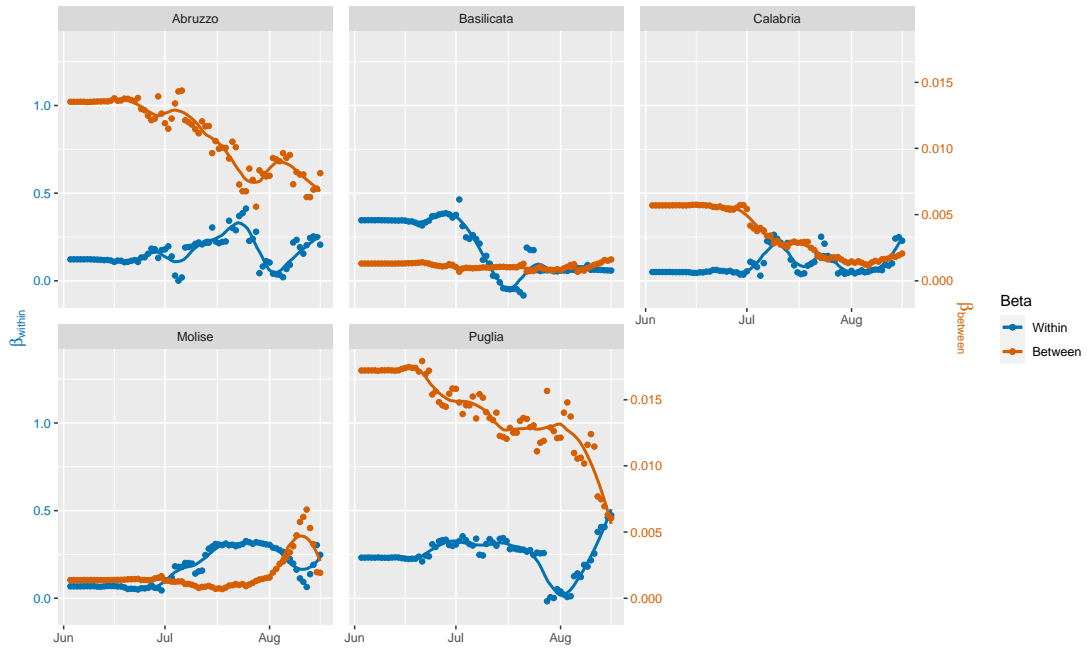


(c) Without model selection;
including undocumented infectives

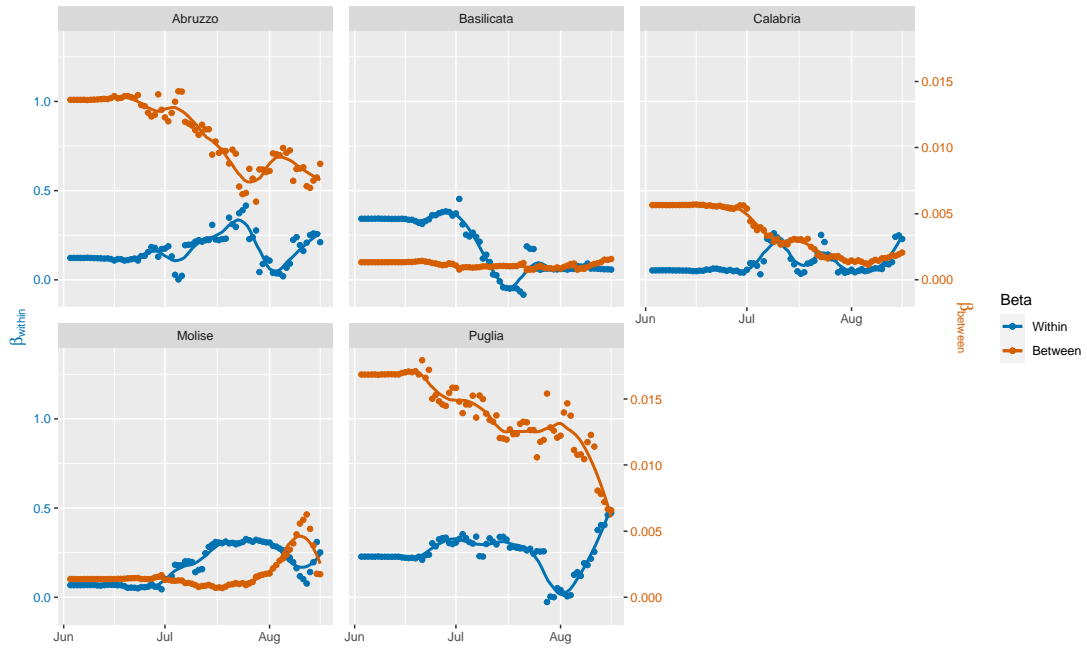


(d) With model selection by AIC;
including undocumented infectives

Figure C.14. Progression of β_{within} and $\beta_{between}$ over time for the *Centro (IT)* (Centre) NUTS 1 region



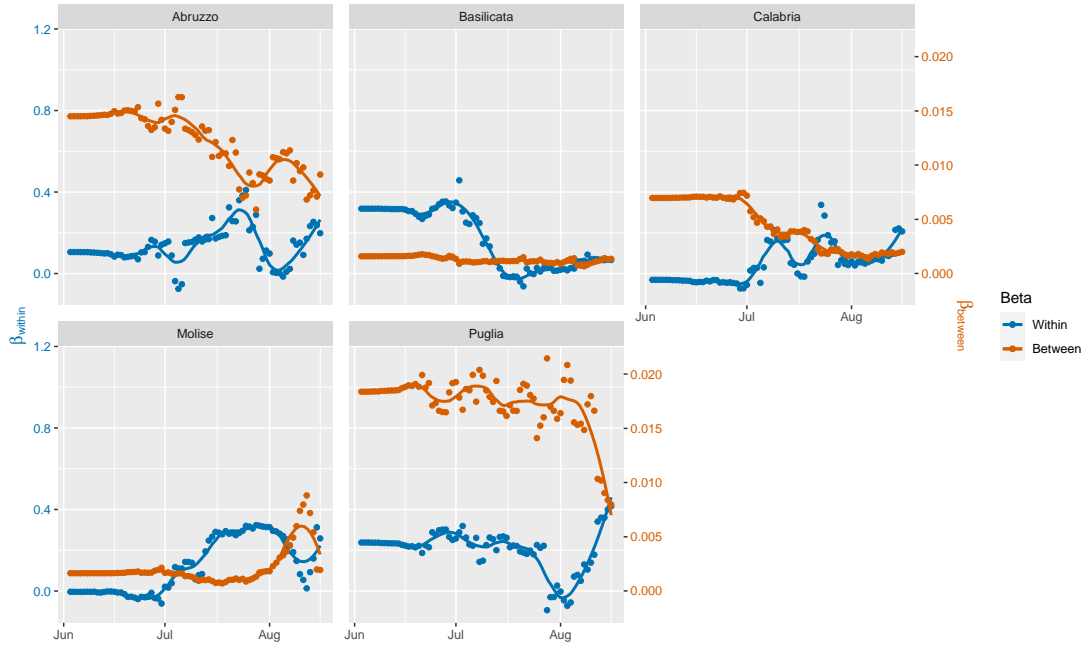
(a) Without model selection



(b) With model selection by AIC



(c) Without model selection;
including undocumented infectives



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

Figure C.15. Progression of β_{within} and $\beta_{between}$ over time for the *Centro (IT)* (Centre) NUTS 1 region

D Derivations

D.1 Calculation of population variables

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

We denote the population of region 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 54th day of 2020, a leap year. As such, the population of region p on February 25, 2020 is calculated as:

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

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:

$$s_{p,t} = N_{p,t} - i_{p,t} - r_{p,t} \quad (\text{D.2})$$

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

D.2 Functional forms for modelling undocumented infectives

In this appendix, we give the derivations for the functional forms for modelling undocumented infectives as discussed in Section 6.

D.2.1 Linear function

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

- (I) $f(TC_t) = aTC_t + b$ for some $a, b \in \mathbb{R}$,
- (II) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,
- (III) $f(N_t) = 1$

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

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

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

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

D.2.2 General quadratic function

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

- (I) $f(TC_t) = aTC_t^2 + bTC_t + c$ for some $a, b, c \in \mathbb{R}$,
- (II) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,
- (III) $f(N_t) = 1$,
- (IV) $f(\beta N_t) = \gamma$ for some $\beta, \gamma \in (0, 1)$,
- (V) The vertex of the parabola should be to the right of N_t in the case of a downwards opening parabola and to the left of the origin in the case of an upwards opening parabola.

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

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

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

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

As such, we have derived that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This is only the case if two conditions are satisfied:

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

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, (D.6a) tells us that both the numerator and denominator of the fraction are negative. Therefore, to satisfy (D.6a), we need that

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

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

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

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

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

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

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

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

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

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

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

Conclusively, we have derived that

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

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

D.2.3 Special case quadratic formula: downwards opening

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

- (I) $f(x) = ax^2 + bx + c$ for some $a, b, c \in \mathbb{R}$,
- (II) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,
- (III) $f(N_t) = 1$,
- (IV) $f'(N_t) = 0$, i.e. the vertex of the parabola is found at $TC_t = N_t$.

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

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

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

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

Therefore, the formula becomes:

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

D.2.4 Special case quadratic formula: upwards opening

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

- (I) $f(x) = ax^2 + bx + c$ for some $a, b, c \in \mathbb{R}$,
- (II) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,
- (III) $f(N_t) = 1$,
- (IV) $f'(0) = 0$, i.e. the vertex of the parabola is found at $TC_t = 0$.

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

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

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

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

Therefore, the formula becomes:

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

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

D.2.5 Cubic function

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

- (I) $f(x) = ax^3 + bx^2 + cx + d$ for some $a, b, c, d \in \mathbb{R}$,
- (II) $f(0) = f^{min}$ for some $f^{min} \in [0, 1]$,
- (III) $f(N_t) = 1$,
- (IV) $f(\beta_1 N_t) = \gamma_1$ and $f(\beta_2 N_t) = \gamma_2$ for some $\beta_1, \beta_2, \gamma_1, \gamma_2 \in [0, 1]$ and $\beta_1 < \beta_2, \gamma_1 < \gamma_2$.

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

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

In appendix D.2.2, we first solved these equations and then assumed a value for β afterwards, without loss of generality. In this case, the equations would become immensely populated if we were to keep the derivation general. As such, we first assume without loss of generality that $\beta_1 = \frac{1}{4}$ and $\beta_2 = \frac{1}{2}$. To solve (D.8), we can then apply row reduction as follows:

$$\begin{aligned}
\left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ \beta_1^3 N_t^3 & \beta_1^2 N_t^2 & \beta_1 N_t & \gamma_1 - f^{min} \\ \beta_2^3 N_t^3 & \beta_2^2 N_t^2 & \beta_2 N_t & \gamma_2 - f^{min} \end{array} \right) &= \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ \frac{1}{64} N_t^3 & \frac{1}{16} N_t^2 & \frac{1}{4} N_t & \gamma_1 - f^{min} \\ \frac{1}{8} N_t^3 & \frac{1}{4} N_t^2 & \frac{1}{2} N_t & \gamma_2 - f^{min} \end{array} \right) \\
&\xrightarrow{r_2 \times 64} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ N_t^3 & 4N_t^2 & 16N_t & 64\gamma_1 - 64f^{min} \\ N_t^3 & 2N_t^2 & 4N_t & 16\gamma_2 - 64f^{min} \end{array} \right) \\
&\xrightarrow{r_3 \times 8} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ 0 & 3N_t^2 & 15N_t & -1 + 64\gamma_1 - 63f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 - 7f^{min} \end{array} \right) \\
&\xrightarrow{r_2 - r_1} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ 0 & 3N_t^2 & 15N_t & -1 + 64\gamma_1 - 63f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 - 7f^{min} \end{array} \right) \\
&\xrightarrow{r_3 - r_1} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 - 7f^{min} \\ 0 & 3N_t^2 & 15N_t & -1 + 64\gamma_1 - 63f^{min} \end{array} \right) \\
&\xrightarrow{r_2 \leftrightarrow r_3} \left(\begin{array}{ccc|c} N_t^3 & N_t^2 & N_t & 1 - f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 - 7f^{min} \\ 0 & 3N_t^2 & 15N_t & -1 + 64\gamma_1 - 63f^{min} \end{array} \right) \\
&\xrightarrow{r_1 - r_2} \left(\begin{array}{ccc|c} N_t^3 & 0 & -2N_t & 2 - 8\gamma_2 + 6f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 \\ 0 & 0 & 6N_t & 2 + 64\gamma_1 - 24\gamma_2 - 42f^{min} \end{array} \right) \\
&\xrightarrow{r_3 - 3r_2} \left(\begin{array}{ccc|c} N_t^3 & 0 & -2N_t & 2 - 8\gamma_2 + 6f^{min} \\ 0 & N_t^2 & 3N_t & -1 + 8\gamma_2 \\ 0 & 0 & 6N_t & 2 + 64\gamma_1 - 24\gamma_2 - 42f^{min} \end{array} \right) \\
&\xrightarrow{r_1 + \frac{1}{3}r_3} \left(\begin{array}{ccc|c} N_t^3 & 0 & 0 & \frac{8+64\gamma_1-48\gamma_2-24f^{min}}{3} \\ 0 & N_t^2 & 0 & -2 - 32\gamma_1 + 20\gamma_2 + 14f^{min} \\ 0 & 0 & 6N_t & 2 + 64\gamma_1 - 24\gamma_2 - 42f^{min} \end{array} \right) \\
&\xrightarrow{r_2 - \frac{1}{2}r_3} \left(\begin{array}{ccc|c} N_t^3 & 0 & 0 & \frac{8+64\gamma_1-48\gamma_2-24f^{min}}{3} \\ 0 & N_t^2 & 0 & -2 - 32\gamma_1 + 20\gamma_2 + 14f^{min} \\ 0 & 0 & 6N_t & 2 + 64\gamma_1 - 24\gamma_2 - 42f^{min} \end{array} \right) \\
&\xrightarrow{r_1 \div N_t^3} \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{8+64\gamma_1-48\gamma_2-24f^{min}}{3N_t^3} \\ 0 & 1 & 0 & \frac{-2-32\gamma_1+20\gamma_2+14f^{min}}{N_t^2} \\ 0 & 0 & 1 & \frac{2+64\gamma_1-24\gamma_2-42f^{min}}{6N_t} \end{array} \right) \\
&\xrightarrow{r_2 \div N_t^2} \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{8+64\gamma_1-48\gamma_2-24f^{min}}{3N_t^3} \\ 0 & 1 & 0 & \frac{-2-32\gamma_1+20\gamma_2+14f^{min}}{N_t^2} \\ 0 & 0 & 1 & \frac{2+64\gamma_1-24\gamma_2-42f^{min}}{6N_t} \end{array} \right) \\
&\xrightarrow{r_3 \div 6N_t} \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{8+64\gamma_1-48\gamma_2-24f^{min}}{3N_t^3} \\ 0 & 1 & 0 & \frac{-2-32\gamma_1+20\gamma_2+14f^{min}}{N_t^2} \\ 0 & 0 & 1 & \frac{2+64\gamma_1-24\gamma_2-42f^{min}}{6N_t} \end{array} \right)
\end{aligned}$$

Conclusively, we have derived that

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

so that

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

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