Bayesian analysis of COVID-19 time series data

Gergely Gálfi

6th May, 2020

Abstract

This report is an attempt to determine the unknown parameters of the COVID-19 spread dynamics in 18 different regions of six countries. With the help of Bayesian analysis we investigated how well a SIR model could be fitted on the currently available COVID-19 time series. Additionally to the standard SIR parameters (R_0, γ), we considered the size of the infected population also as unknown as there aren't available any generally homogeneous and reliable data on it. Based on the data of deceased people we've found that even with fairly loose prior distributions a SIR model could be well fitted. The regions/states where this simple model was well fitting were in countries without active control of the virus spread (regions mostly in England, Italy, Spain and US). We observed that in these regions R_0 is ranging between 1.41 and 12.7 at 95% of confidence and has an average of 6.51, definitely larger value than the currently accepted one. We also determined that for these regions the mortality rate was very consistently around 0.1%. The most important outcome of this analysis is that if our prior assumptions are true then within most of the regions only a very small proportion of COVID-19 cases (around 0.3-1%) were discovered and they practically achieved herd immunity as of the beginning of May, 2020.

1 Introduction

In many analytical tasks on COVID-19 data, one of the most desired input parameter is the size of the infected population. Despite its importance the publicly available data on the confirmed cases could be considered as unreliable due to the following reasons:

- There is no way of overall testing of the whole population and the selection methods are changing not only from one region to another but over time as well.
- Certainly the selection for testing is mostly based on (sometimes ad-hoc) diagnostic and spread-prevention considerations which introduces a non-negligible but practically unknown bias.
- There is no unison definition of a COVID-19 case.
- The reporting date of a case could depend on a lot of factors which do not reflect the progress of the disease itself rather due some fluctuating administrative time lags. That could lead to erroneous/useless time series data even if the "grand total" is more or less correct.

To obtain this very important but still unknown parameter we employed Bayesian analysis on the time series data of the people deceased due to COVID-19. The previously depicted problems related to the confirmed cases are also plaguing the data of the deceased ones albeit rather less extent. When we selected the countries for analysis we considered the availability of regional

data, the reliability of data (we haven't taken into account countries where large scale data forgery could be suspected) and having large number of confirmed cases. Beside these we focused our attention to countries which aren't doing active control of the virus spread which makes a homogeneous dynamic assumption more plausible. With these requirements we used the data of five countries: England, France, Germany, Italy, Spain and US. Although Germany has taken very thorough and well-thought measures against the virus, we included it in our analysis to see how our method is working (or not working) on it. In each country we selected the top three regions regarding the number of confirmed cases. Ultimately we have analysed these 18 regions:

Country	Region	Confirmed cases	Deceased	Population
ENG	London	25 357	5 282	8 908 081
ENG	Midlands	21 443	4 325	10 704 906
ENG	North West	21 000	3 289	7 292 093
ESP	Community of Madrid	63 416	8 466	6 663 394
ESP	Catalonia	50 924	5 345	7 675 217
ESP	Castile and León	17 520	1 847	2 399 548
FRA	Île-de-France	7 660	6 347	12 174 880
FRA	Grand Est	3 395	3 037	5 549 586
FRA	Hauts-de-France	3 011	1 480	6 003 815
GER	Bayern	43 371	2 001	13 076 721
GER	Nordrhein-Westfalen	33 977	1 358	17 932 651
GER	Baden-Württemberg	32 576	1 481	11 069 533
ITA	Lombardia	79 369	14 611	9 756 932
ITA	Piemonte	27 939	3 247	4 356 397
ITA	Emilia-Romagna	26 379	3 737	4 459 477
US	New York	380 897	25 124	19 453 561
US	New Jersey	130 593	8 244	8 882 190
US	Massachusetts	70 271	4 212	6 892 503

During the analysis we used the following assumptions:

- Within a selected region and within the time frame of the lockdown the dynamics of the spread could be described with a basic SIR model, i.e. with the following set of differential equations:

$$\dot{S} = -\frac{\beta I S}{N};$$

$$\dot{I} = \frac{\beta I S}{N} - \gamma I;$$

$$\dot{R} = \gamma I.$$

  where S is the stock of susceptible population, I is the stock of infected population and E is the stock of removed (recovered or dead). The two constants β and γ are parameters of the model and N is the size of population for which we use the region population. Instead of β we will use the basic reproduction number $R_0 = \beta/\gamma$. The main reason behind using such a simple model (if not the simplest) was the fact that we are fitting model on a very noisy and contaminated data. By the "less parameter - less trouble" principle we usually get to more stable and reliable models.

- The known (or reported) number of COVID-19 deaths equals to the actual number of deaths. We know that this assumption is not entirely true, for example a few countries -

like France - have not included in their regional reports the COVID-19 death cases in care homes. However we may assume that these errors are not causing larger deviances than a factor of two.

- As we consider all the dynamical variables of the SIR model as unknown, we have to connect them to the dead cases. Therefore we assume that for each region there is a constant mortality rate m which connects the predicted number of the deceased people to the removed stock through $D_i^{pred} = mR_i^{pred}$.

- Due to the constraint $S + I + R = N$, we have two independent dynamical variables. To make the initial values well-posed we have to introduce a new parameter i_0 which is the proportion of infected at the time of the first data point. This makes our model a four-parameter model with parameter set (R_0, γ, m, i_0). We don't assume that any of these parameters have the same value for different regions, however it is convincing if you see similar values for γ and m as they aren't depending on the counter-measures against the virus spread.

- The likelihood function was calculated as a product of independent binomial distributions $B(\delta D_i | n = N, p = \delta D_i^{pred}/N)$ for each observations, where δD_i is the reported, δD_i^{pred} is the predicted number of daily deaths.

- For Bayesian analysis we will use the following priors:

$$R_0 \sim lognormal(median = 3, \sigma = 0.921);$$

$$\gamma \sim lognormal(median = 0.05, \sigma = 1.151);$$

$$m \sim lognormal(median = 0.01, \sigma = 2.303);$$

$$i_0 \sim lognormal(median = 0.01, \sigma = 3.454).$$

2 Data preparation

We tried to collect all the data from their original sources[1][2][3][4][6][7]. The only exception was the US data because in that case each state providing their data individually[5]. Therefore we downloaded an aggregated data maintained by Johns Hopkins University.

Visual exploration of the data uncovered a few problems with it. Should these problems remain unhandled it'd lead to biased and unstable models. For example in case of Lombardy region the time series of daily deaths looks like this:

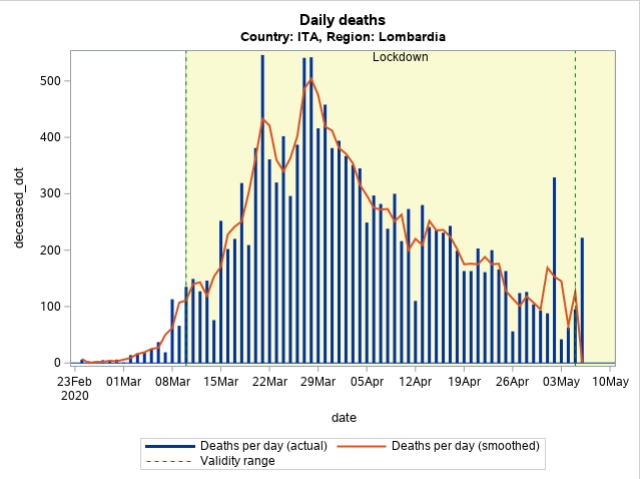


Figure 1: Daily deaths in Lombardy

Obviously the time series has sudden changes, or spikes which cannot be attributed neither to changes of dynamics (it is too fast) nor to statistical fluctuations (too large for a binomial like distribution). We attribute this to some administrative delays or pile-ups of reporting. To get rid of these spikes we smoothed the the cumulative number of deceased people by averaging, i.e.

$$D'_i = (D_{i-1} D_i D_{i+1})^{1/3}$$

We used here geometric mean instead of arithmetic mean because it handles better an exponentially changing time series. Smoothed daily deaths $\delta D'_i$ is calculated from this one as $\delta D'_i = D'_i - D'_{i-1}$. Further on we will use this smoothed time series in our calculations.

There were other regions which were plagued by more serious issues with data. For example, in case of New York state after 23rd Apr the cumulative number of deaths began to decline (which is obviously a marker for erroneous data):

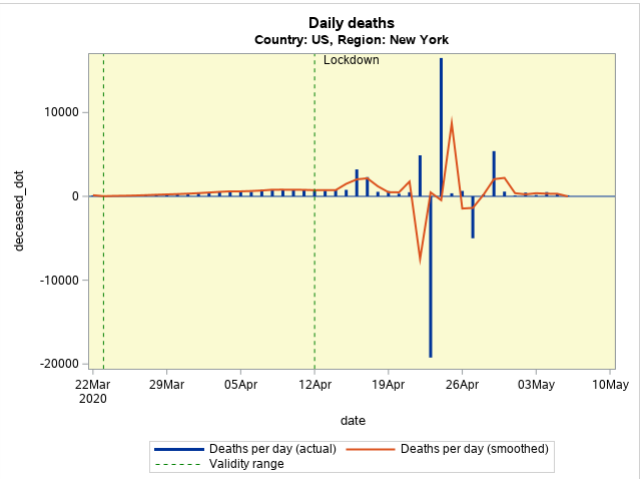


Figure 2: Daily deaths in New York State

In this case there was no other option than to shrink the relevant time interval till it haven't contained this erratic observations.

3 Bayesian analysis

To calculate posterior probabilities we employed Markov Chain Monte Carlo method. Technically we used the **proc mcmc** procedure of SAS. The code, the data and the results are available in dedicated github project[8], please refer to this for the exact parametrization of MCMC.

3.1 Goodness of fit of the SIR model

To quantify the goodness of fit in a meaningful way we defined the median relative deviation as the median of the residual-prediction ratio,

$$\underset{i=day_1 \ldots day_N}{median} \left\{ \frac{\left| \delta D_i' - \delta D_i^{pred} \right|}{\delta D_i^{pred}} \right\},$$

so in layman's terms it is an "average" of model error percentage. We used median instead of mean because the latter is more sensitive on outliers (and our data is widely polluted with them). The posterior analysis of the median relative deviation leads to the following regional statistics:

Country	Region	Mean	StdDev	HPDLower	HPDUpper	Selected
ENG	Midlands	4.05%	0.86%	2.73%	5.91%	Yes
US	New York	4.45%	0.62%	3.25%	5.68%	Yes
ENG	London	4.90%	0.92%	3.27%	6.75%	Yes
ENG	North West	5.55%	0.77%	4.02%	6.97%	Yes
ESP	Community of Madrid	6.26%	0.59%	5.12%	7.55%	Yes
ESP	Castile and León	6.75%	1.05%	4.84%	8.84%	Yes
ITA	Emilia-Romagna	7.90%	0.81%	6.44%	9.56%	Yes
FRA	Grand Est	9.01%	1.02%	7.03%	10.99%	Yes
ITA	Lombardia	11.30%	0.41%	10.50%	12.11%	Yes
ITA	Piemonte	11.69%	1.55%	8.78%	14.64%	Yes
FRA	Île-de-France	12.13%	1.20%	9.94%	14.42%	Yes
ESP	Catalonia	15.92%	0.97%	14.30%	17.89%	No
US	Massachusetts	15.99%	0.94%	14.17%	17.83%	No
FRA	Hauts-de-France	16.44%	1.26%	14.60%	19.12%	No
GER	Baden-Württemberg	17.64%	1.59%	14.52%	20.74%	No
GER	Nordrhein-Westfalen	18.11%	1.52%	15.36%	21.33%	No
US	New Jersey	18.84%	1.05%	16.75%	20.93%	No
GER	Bayern	19.60%	1.09%	17.39%	21.81%	No

From this table one can conclude that there are two clusters of regions regarding the median relative deviation, one with values below 15% and the other with larger than 15% and there is a large gap between the two. It is remarkable that of all the German regions are in the group of weaker fit: this could be explained by the active control employed by German authorities which makes the dynamics far more complex to be described by a simple SIR model. However there are other regions falling into this group like Catalonia or New Jersey. For these it seems likely that the quality of the data is the real culprit: in New Jersey the weekly seasonality is so strong that the within week changes are larger than the monthly change, rendering the model fitting unreliable. So we decided to focus our attention on the stronger group and the overall statistics of the parameters are based only on the eleven members of this one:

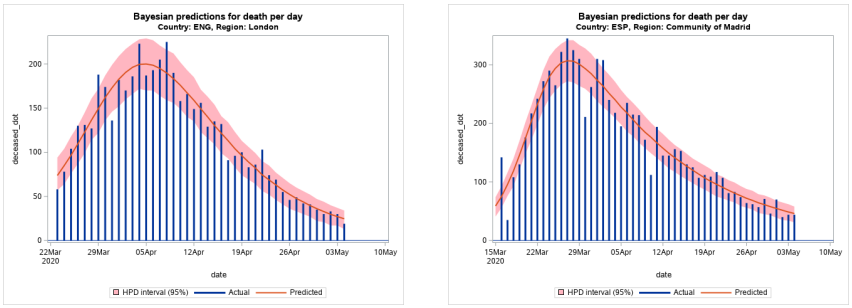
- England: London, Midlands, North West

- France: Île-de-France, Grand Est
- Italy: Lombardia, Piemonte, Emilia-Romagna
- Spain: Community of Madrid, Castile and León
- US: New York

From now on, we will refer to these regions as the *selected group*.

3.2 Posterior predictions

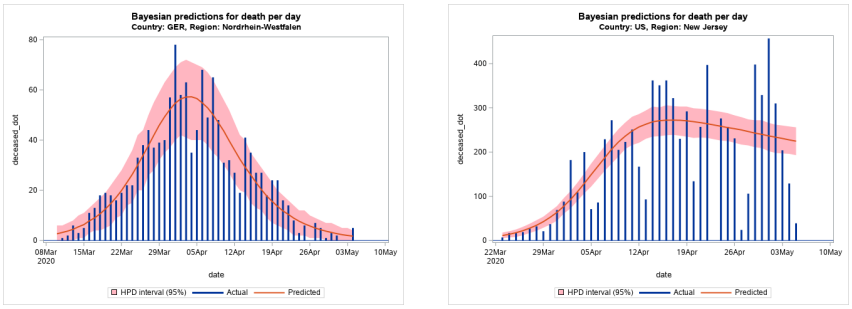
With the help of posterior sample we can produce the prediction estimates. To give some insight on the accuracy, here are two regions where the actual model fits very well:



(a) London (b) Madrid

Figure 3: Daily death counts with good model fit(predicted vs actual)

Apart from a few spikes - which were more or less eliminated by the geometrical mean smoothing - the actual values are very close to the predicted ones and additionally, the uncertainty (HPD) range is fairly narrow. A significantly worse (though not very bad) fit could be seen for the nonselected regions. Two representative examples could be seen below:



(a) Nordrhein-Westfalen (b) New Jersey

Figure 4: Daily death counts with bad model fit(predicted vs actual)

In case of Nordrhein-Westfallen the actual data is not extremely noisy, so one may suspect that here the SIR model itself was too simple to provide a good fit. Contrary, on the time series of New Jersey signs of extreme seasonality and other noise could be seen and that is an important factor behind bad model fitting.

3.3 Posterior statistics of model parameters

3.3.1 Posteriors for R_0

Regional statistics:

Country	Region	Mean	StdDev	HPDLower	HPDUpper	Selected
ENG	London	2.5809	0.3538	1.8904	3.2661	Yes
ENG	Midlands	4.6868	0.7222	3.2744	6.0624	Yes
ENG	North West	4.4788	0.6322	3.3443	5.8111	Yes
ESP	Castile and León	8.4515	1.0105	6.5409	10.4934	Yes
ESP	Catalonia	15.9582	1.6285	12.8363	19.1596	No
ESP	Community of Madrid	6.8405	0.4470	6.0063	7.7180	Yes
FRA	Grand Est	6.0669	1.0298	4.2226	8.2106	Yes
FRA	Hauts-de-France	5.0371	0.8531	3.4261	6.7114	No
FRA	Île-de-France	6.5734	1.1141	4.4018	8.7178	Yes
GER	Baden-Württemberg	2.4625	0.3228	1.7783	3.0429	No
GER	Bayern	3.2071	0.3653	2.5329	3.9791	No
GER	Nordrhein-Westfalen	1.6916	0.2301	1.2448	2.1053	No
ITA	Emilia-Romagna	7.7059	1.3599	5.3132	10.5080	Yes
ITA	Lombardia	6.7552	0.3832	5.9764	7.4867	Yes
ITA	Piemonte	2.2933	0.5579	1.2858	3.3286	Yes
US	Massachusetts	2.7025	0.7358	1.4578	4.0579	No
US	New Jersey	22.3195	5.5975	13.6574	33.6495	No
US	New York	15.2264	8.1769	4.4892	30.8782	Yes

Overall statistics for the selected group:

Mean	StdDev	HPDLower	HPDUpper
6.5145	4.2203	1.4069	12.7140

It could be seen that for the selected the regions, R_0 is more consistent than for the nonselected ones. This reassuring about the correctness of the SIR model as our prior distribution on R_0 would have allowed far larger excursions. It is also remarkable that the mean of R_0 is 6.51 which is larger than most of the values given in the literature which - as of writing this report - is ranging between 1.4 and 5.7. It also could be seen that the mean values are highly differing between some regions (two extremes are London with 2.6 and New York with 15.2). This is normal as counter-measures and local social habits are differing as well and these factors could have significant impact on R_0. For the nonselected group we can even see more extremal values which supports our previous decision of omitting them from further analysis.

3.3.2 Posteriors for γ

Regional statistics:

Country	Region	Mean	StdDev	HPDLower	HPDUpper	Selected
ENG	London	0.1206	0.0176	0.0886	0.1537	Yes
ENG	Midlands	0.0622	0.00731	0.0493	0.0774	Yes
ENG	North West	0.0703	0.00859	0.0546	0.0882	Yes
ESP	Castile and León	0.0491	0.00368	0.0420	0.0562	Yes
ESP	Catalonia	0.0282	0.00174	0.0249	0.0317	No
ESP	Community of Madrid	0.0554	0.00189	0.0518	0.0590	Yes
FRA	Grand Est	0.0507	0.00388	0.0436	0.0583	Yes
FRA	Hauts-de-France	0.0593	0.00883	0.0432	0.0765	No
FRA	Île-de-France	0.0527	0.00451	0.0448	0.0620	Yes
GER	Baden-Württemberg	0.1668	0.0288	0.1181	0.2244	No
GER	Bayern	0.1271	0.0145	0.1006	0.1549	No
GER	Nordrhein-Westfalen	0.3185	0.1112	0.1628	0.5549	No
ITA	Emilia-Romagna	0.0324	0.00335	0.0263	0.0391	Yes
ITA	Lombardia	0.0417	0.00141	0.0392	0.0447	Yes
ITA	Piemonte	0.1102	0.0498	0.0476	0.2054	Yes
US	Massachusetts	0.1199	0.0535	0.0405	0.2250	No
US	New Jersey	0.0133	0.00275	0.00827	0.0189	No
US	New York	0.0290	0.0137	0.00733	0.0565	Yes

Overall statistics for the selected group:

Mean	StdDev	HPDLower	HPDUpper
0.0613	0.0327	0.0131	0.1304

The value of 0.061 1/day is equivalent to an infectious halving time of 11 days. It is noticeable that the German regions has unrealistically high values for γ, although it doesn't contradict our previous observation that SIR model isn't eligible to describe the German situation. However London had a very good and also Piemonte had a fairly good fit, and they still have about a twice of the γ of other regions. We don't have any explanation for that at the moment. Nevertheless, the nine other selected regions has fairly consistent values.

3.3.3 Posteriors for mortality rate

Regional statistics:

Country	Region	Mean	StdDev	HPDLower	HPDUpper	Selected
ENG	London	0.06%	0.003%	0.06%	0.07%	Yes
ENG	Midlands	0.05%	0.001%	0.04%	0.05%	Yes
ENG	North West	0.05%	0.001%	0.05%	0.05%	Yes
ESP	Castile and León	0.09%	0.003%	0.08%	0.09%	Yes
ESP	Catalonia	0.10%	0.003%	0.09%	0.11%	No
ESP	Community of Madrid	0.14%	0.002%	0.13%	0.14%	Yes
FRA	Grand Est	0.06%	0.001%	0.05%	0.06%	Yes
FRA	Hauts-de-France	0.03%	0.001%	0.03%	0.03%	No
FRA	Île-de-France	0.06%	0.001%	0.05%	0.06%	Yes
GER	Baden-Württemberg	0.01%	0.001%	0.01%	0.02%	No
GER	Bayern	0.02%	0.000%	0.02%	0.02%	No
GER	Nordrhein-Westfalen	0.01%	0.002%	0.01%	0.02%	No
ITA	Emilia-Romagna	0.10%	0.004%	0.10%	0.11%	Yes
ITA	Lombardia	0.17%	0.002%	0.16%	0.17%	Yes
ITA	Piemonte	0.10%	0.016%	0.09%	0.13%	Yes
US	Massachusetts	0.10%	0.008%	0.09%	0.11%	No
US	New Jersey	0.29%	0.050%	0.21%	0.39%	No
US	New York	0.22%	0.096%	0.10%	0.40%	Yes

Overall statistics for the selected group:

Mean	StdDev	HPDLower	HPDUpper
0.10%	0.06%	0.04%	0.19%

That is very convincing about the righteousness of our assumptions: the mortality rates for the selected regions are very close to each other despite we've given a very loose prior for the mortality rate. The extremal values are belonging to regions excluded from the analysis.

3.3.4 Posteriors for i_0

Regional statistics:

Country	Region	Mean	StdDev	HPDLower	HPDUpper	Selected
ENG	London	0.1094	0.0155	0.0794	0.1400	Yes
ENG	Midlands	0.1260	0.0110	0.1039	0.1470	Yes
ENG	North West	0.0592	0.00585	0.0479	0.0704	Yes
ESP	Castile and León	0.0339	0.00573	0.0230	0.0452	Yes
ESP	Catalonia	0.0286	0.00381	0.0217	0.0365	No
ESP	Community of Madrid	0.1183	0.00680	0.1050	0.1309	Yes
FRA	Grand Est	0.3722	0.0252	0.3234	0.4214	Yes
FRA	Hauts-de-France	0.0387	0.00579	0.0271	0.0495	No
FRA	Île-de-France	0.2403	0.0154	0.2104	0.2697	Yes
GER	Baden-Württemberg	0.0101	0.00150	0.00697	0.0129	No
GER	Bayern	0.00733	0.000900	0.00558	0.00903	No
GER	Nordrhein-Westfalen	0.00475	0.00162	0.00158	0.00775	No
ITA	Emilia-Romagna	0.1603	0.0150	0.1330	0.1918	Yes
ITA	Lombardia	0.1281	0.00556	0.1167	0.1382	Yes
ITA	Piemonte	0.0254	0.00923	0.00795	0.0428	Yes
US	Massachusetts	0.00881	0.00297	0.00249	0.0142	No
US	New Jersey	0.0340	0.00289	0.0282	0.0393	No
US	New York	0.0442	0.00481	0.0343	0.0530	Yes

Overall statistics for the selected group:

Mean	StdDev	HPDLower	HPDUpper
0.1288	0.0984	0.0113	0.3722

As the i_0 parameter is an initial value of the SIR model and highly depends on the more or less arbitrarily selected starting date, it doesn't have too much descriptive value. However we present it for the sake of completeness.

3.4 Association between the model parameters

We can visualize the connection between the model parameters through some scatter plots:

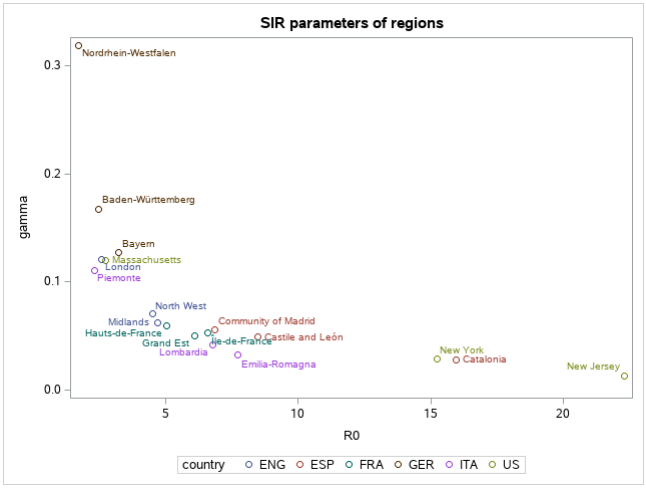


Figure 5: Posterior means of R_0 and γ

This strong inverse connection between R_0 and γ could be explained simply as if a certain region quickly separates the infectious cases (so achieves a higher γ) then it will reduce the reproduction ratio, provided by the β parameter (the daily disease transfer rate) doesn't varying too much. The latter happens to be true, which could be seen on the diagram below, as β is ranging between 0.23 and 0.41. It is also noticeable that there aren't any significant association between γ and β.

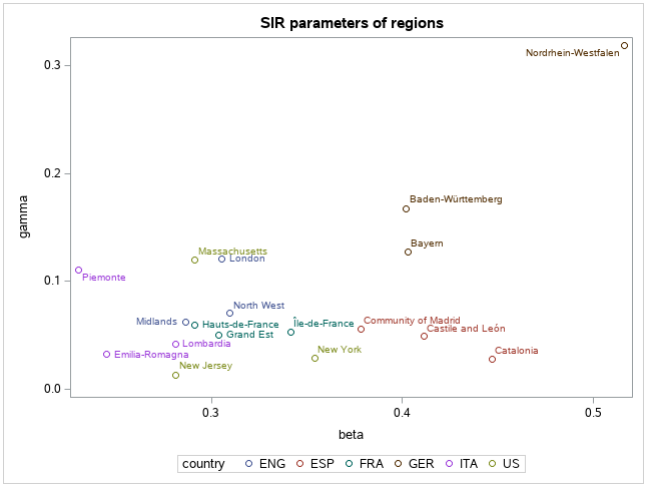


Figure 6: Posterior means of β and γ

The connection between the mortality and γ could be seen on the diagram below:

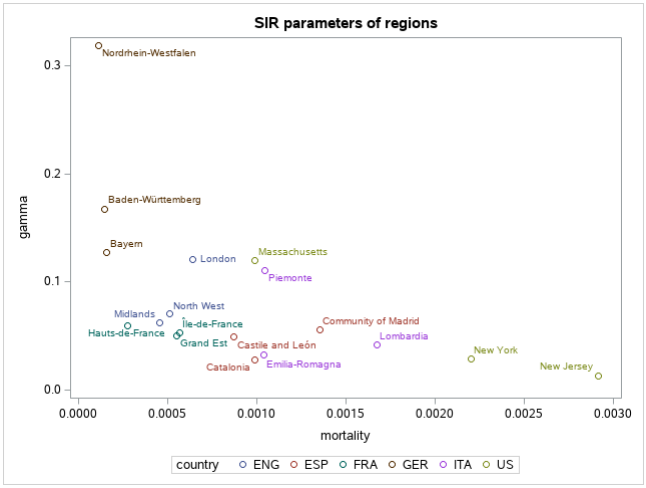
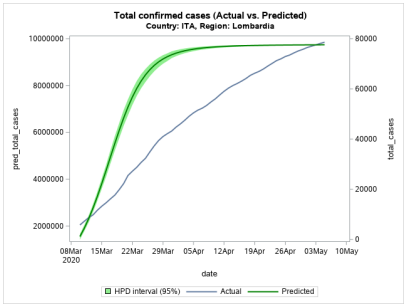


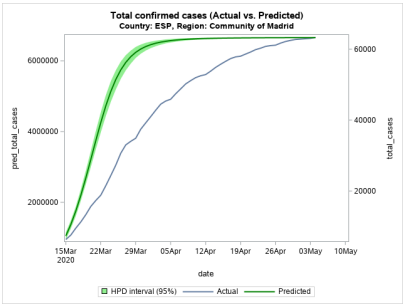
Figure 7: Posterior means of mortality and γ

4 Modeling the infectious stock

By putting together all these analytics we are in the position to answer the original question: how many COVID-19 cases we have? We can calculate the (predicted) number of the infectious cases and we can even compare it with the number of confirmed cases. What we can see on selected regions which began to struggle with the virus at an earlier time they have practically achieved herd immunity and the new confirmed cases and deaths which we can see even now are the remnants from the past three-four weeks when the outbreak was still in it's active phase. These are two examples for that type of region:



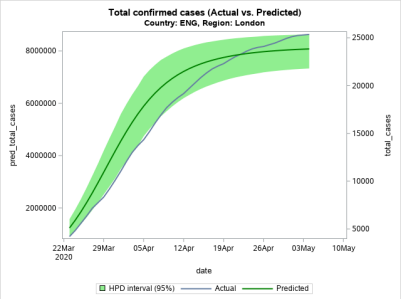
(a) Lombardy



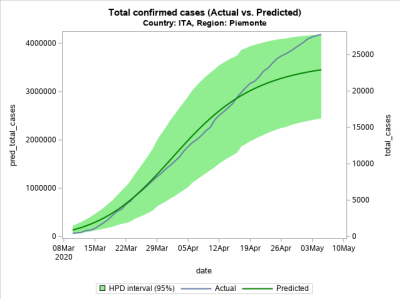
(b) Madrid

Figure 8: Total case counts (predicted vs actual)

On the other hand there are regions which have not yet arrived to herd immunity, although they are very close (within a few weeks) to it. The time series of the total cases looks like this:



(a) London



(b) Piemonte

Figure 9: Total case counts (predicted vs actual)

It is important to note, that the predicted and actual cases are running on totally different scales, practically less than 1% is detected by virus test.

5 Conclusions

Provided by that our assumptions are not totally wrong, the following conclusions could be drawn from this analysis:

- The R_0 value is larger than previously thought;
- The average of infectious halving time is about 11 days;
- Mortality is about 0.1 %;
- A very large proportion of the infected population (>99%) remains unnoticed.

References

[1] Presidenza del Consiglio dei Ministri, Italy, `https://github.com/pcm-dpc/COVID-19`.

[2] Robert Koch Institut, Germany, `https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74`

[3] Spain official COVID-19 data, `https://covid19.isciii.es/resources/serie_historica_acumulados.csv`

[4] OpenCOVID19 France, `https://github.com/opencovid19-fr/data`

[5] Johns Hopkins University, `https://github.com/CSSEGISandData/COVID-19`

[6] UK official COVID-19 data, `https://coronairus.data.gov.uk/downloads/csv/coronavirus-cases_latest.csv`

[7] NHS England data on daily deaths, `https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/`

[8] Codes and data used in this report, `https://github.com/ggalfi/COVID19_SAS`