Bayesian analysis of COVID-19 time series data

Gergely Gálfi

27th April, 2020

Abstract

This report is an attempt to determine the unknown parameters of the COVID-19 spread dynamics in 15 different regions of five 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 homogenous 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 in Italy, Spain and US). We observed that in these regions R_0 is ranging between 2.98 and 25.28 with 95% of confidence, much larger than the currently accepted values. We also determined that for these regions the mortality rate was very consistently around 0.15%. The most important outcome of this analysis is that if our prior assumptions are true then with the analysed regions of Italy, Spain and the US only a very small proportion of COVID-19 cases (around 0.5-1%) were discovered and they practically achieved herd immunity as of the end of April, 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 unkown 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 aviability of regional data,

the reliability of data (we don'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 homogenous dynamic assumption more plausible. With these requirements we used the data of five countries: 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 regions:

Country	Region	Confirmed cases	Deceased	Population
ESP	Community of Madrid	59 421	7 986	6 663 394
ESP	Catalonia	47 755	4 699	7 675 217
ESP	Castile and León	16 404	1 690	2 399 548
FRA	Île-de-France	7 660	5 690	12 174 880
FRA	Grand Est	3 395	2 779	5 549 586
FRA	Hauts-de-France	3 011	1 315	6 003 815
GER	Bayern	41 070	1 621	13 076 721
GER	Nordrhein-Westfalen	31 879	1 131	17 932 651
GER	Baden-Württemberg	31 043	1 249	11 069 533
ITA	Lombardia	73 479	13 449	9 756 932
ITA	Piemonte	25 098	2 878	4 356 397
ITA	Emilia-Romagna	24 662	3 431	4 459 477
US	New York	380 897	23 998	19 453 561
US	New Jersey	109 038	5 938	8 882 190
US	Massachusetts	54 938	2 899	6 892 503

During the analysis we used the following assumptions:

- Within a selected region and timeframe 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 wich 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 polluted data. Using the "less parameter - less trouble" principle usually leads to more stable and reliable model fitting.

- The known 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 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 number of the deceased to the removed stock through $D = mR$.
- Due to the constraint $S + I + R = N$, we have two independent dynamical variables. To make the inital 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 make our model a four-parameter model with paramter 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.
- 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]. 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 promblems with it. Should these problems remain unhandled it'd lead to biased and unstable models. For example in case of Madrid region the time series of daily deaths looks like this:

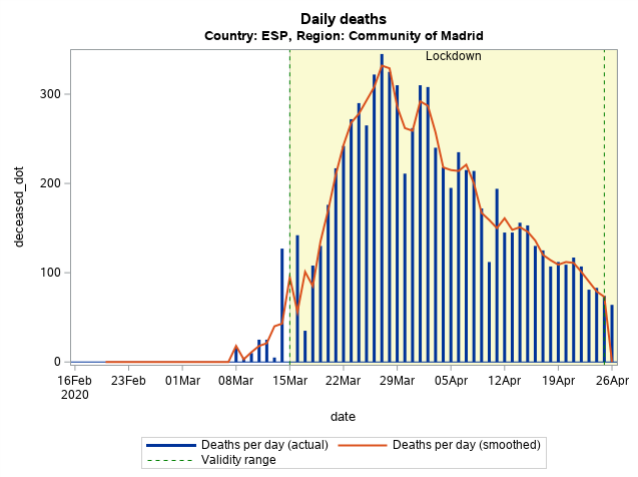


Figure 1: Daily deaths in Madrid

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 with smoothed the curve by averaging, i.e.

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

Instead of arithmetic mean we used here geometric mean because it handles better an exponentially changing time series.

There were other regions which were plagued by more serious issues with data. For example, in case New York state, the cumulative number of deaths began to incline:

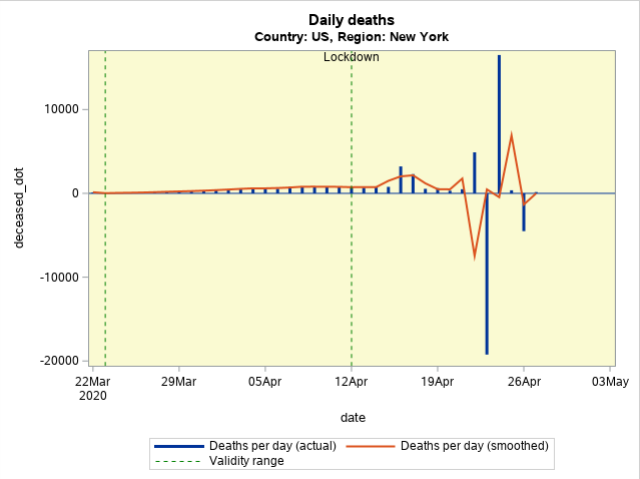


Figure 2: Daily deaths in New York State

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. Techically we used the **proc mcmc** procedure of SAS. The code, the data and the results are available in dedicated github project[6], please refer to this for the exact parametrizaton of MCMC.

From the MCMC algorithm we can obtain the marginal posterior distribution for the model parameters.

3.1 Posteriors for R_0

Country	Region	Mean	StdDev	HPD Lower	HPD Upper
ESP	Castile and León	8.3345	1.2154	6.0865	10.6466
ESP	Catalonia	18.3602	2.5142	13.5290	23.1283
ESP	Community of Madrid	7.3452	0.5739	6.2119	8.4491
FRA	Île-de-France	8.4080	2.1777	4.5706	12.6850
FRA	Hauts-de-France	4.3596	1.0810	2.3325	6.5053
FRA	Grand Est	4.3947	1.0868	2.4007	6.5689
GER	Nordrhein-Westfalen	1.6882	0.2705	1.1924	2.2252
GER	Baden-Württemberg	1.9108	0.2842	1.3447	2.4435
GER	Bayern	2.8746	0.3702	2.1480	3.5804
ITA	Piemonte	10.8869	5.5342	2.9129	22.1919
ITA	Lombardia	6.0635	0.4257	5.2155	6.8629
ITA	Emilia-Romagna	11.0336	2.3228	7.0679	15.8116
US	Massachusetts	5.6716	3.7441	1.5304	13.0989
US	New Jersey	15.1361	6.4361	5.9875	28.6320
US	New York	15.6660	9.0917	4.7117	31.7366

It could be seen that for most of the regions, R_0 is much larger than the values in the literature which (currently) ranging between 1.4 and 5.7. It also could be seen that the mean values are highly differing between regions.This is normal as counter-measures and local social habits are differing as well and these factors could have significant impact on R_0. It is also remarkable that all the German and one French regions have definitely smaller value than the others. That's one reason why we may assume that while the Italian, Spanish and US regions are similar to each other but differing from the France and Germany. So from now on, we will refer to the regions of Italy,Spain and US as the *selected group*. and Calculating the overall statistics for R_0 in gives the following result:

Mean	StdDev	HPD Lower	HPD Upper
10.9441738	6.2238214	2.97937	25.2863

3.2 Posteriors for γ

Country	Region	Mean	StdDev	HPD Lower	HPD Upper
ESP	Castile and León	0.0498	0.00537	0.0391	0.0599
ESP	Catalonia	0.0254	0.00249	0.0203	0.0300
ESP	Community of Madrid	0.0526	0.00242	0.0481	0.0573
FRA	Île-de-France	0.0463	0.00718	0.0335	0.0607
FRA	Hauts-de-France	0.0719	0.0199	0.0384	0.1080
FRA	Grand Est	0.0644	0.0112	0.0468	0.0859
GER	Nordrhein-Westfalen	0.3534	0.1572	0.1593	0.6312
GER	Baden-Württemberg	0.2812	0.0906	0.1610	0.4458
GER	Bayern	0.1591	0.0237	0.1174	0.2086
ITA	Piemonte	0.0208	0.00954	0.00629	0.0401
ITA	Lombardia	0.0456	0.00225	0.0414	0.0500
ITA	Emilia-Romagna	0.0250	0.00338	0.0187	0.0316
US	Massachusetts	0.0619	0.0501	0.00601	0.1636
US	New Jersey	0.0213	0.00765	0.00826	0.0367
US	New York	0.0289	0.0144	0.00596	0.0562

For the selected group, the overall summary is:

Mean	StdDev	HPD Lower	HPD Upper
0.0368016	0.0232719	0.010883	0.084218

The value of 0.03680 1/day is equivalent to an infectious half-life of 19 days. It is also noticable that Germany has unrealistically high values, and France has slightly high values. That also makes us to separate the regions of this two countries from the selected group.

3.3 Posteriors for mortality rate

Country	Region	Mean	StdDev	HPD Lower	HPD Upper
ESP	Castile and León	0.000869	0.000039	0.000796	0.000945
ESP	Catalonia	0.00106	0.000065	0.000932	0.00118
ESP	Community of Madrid	0.00138	0.000024	0.00134	0.00143
FRA	Île-de-France	0.000596	0.000034	0.000528	0.000658
FRA	Hauts-de-France	0.000267	0.000015	0.000242	0.000300
FRA	Grand Est	0.000526	0.000014	0.000501	0.000554
GER	Nordrhein-Westfalen	0.000098	0.000026	0.000068	0.000141
GER	Baden-Württemberg	0.000149	0.000024	0.000121	0.000195
GER	Bayern	0.000134	5.215E-6	0.000125	0.000145
ITA	Piemonte	0.00158	0.000455	0.000965	0.00248
ITA	Lombardia	0.00163	0.000025	0.00158	0.00167
ITA	Emilia-Romagna	0.00118	0.000082	0.00104	0.00135
US	Massachusetts	0.00152	0.000642	0.000882	0.00277
US	New Jersey	0.00215	0.000643	0.00126	0.00344
US	New York	0.00225	0.00105	0.000994	0.00407

For the selected group, the overall summary is:

Mean	StdDev	HPD Lower	HPD Upper
0.0015129	0.000654976	0.000839543	0.003240725

That is very convincing about the rightousness of our assumptions: apart from Germany and France, the regions mortality rates are very close to each other however we've given to the mortality rate a very loose prior.

3.4 Posteriors for i_0

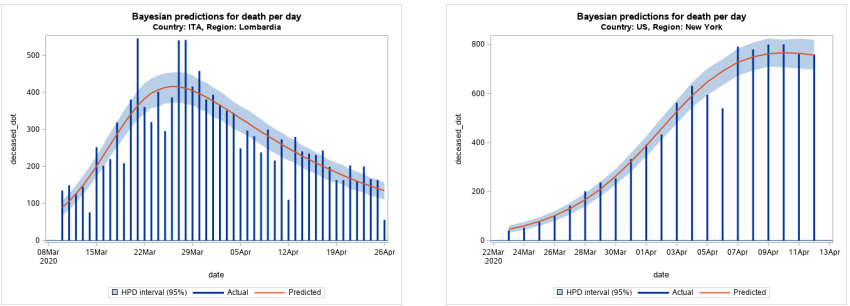
Country	Region	Mean	StdDev	HPD Lower	HPD Upper
ESP	Castile and León	0.0341	0.00564	0.0238	0.0452
ESP	Catalonia	0.0282	0.00388	0.0205	0.0356
ESP	Community of Madrid	0.1201	0.00730	0.1063	0.1342
FRA	Île-de-France	0.2514	0.0195	0.2147	0.2910
FRA	Hauts-de-France	0.0354	0.00676	0.0225	0.0490
FRA	Grand Est	0.3206	0.0397	0.2421	0.3950
GER	Nordrhein-Westfalen	0.00465	0.00191	0.000832	0.00806
GER	Baden-Württemberg	0.00599	0.00167	0.00266	0.00923
GER	Bayern	0.00584	0.000809	0.00440	0.00752
ITA	Piemonte	0.0696	0.0108	0.0502	0.0911
ITA	Lombardia	0.1228	0.00560	0.1120	0.1338
ITA	Emilia-Romagna	0.1717	0.0178	0.1394	0.2083
US	Massachusetts	0.0148	0.00535	0.00334	0.0237
US	New Jersey	0.0319	0.00300	0.0268	0.0384
US	New York	0.0444	0.00478	0.0349	0.0533

For the selected group, the overall summary is:

Mean	StdDev	HPD Lower	HPD Upper
0.0708498	0.0521161	0.0104	0.18473

3.5 Posterior predictions

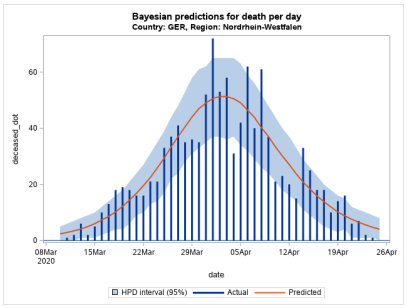
With the help of posterior sample we can produce the prediction estimates as well. To give some insight on the accuary, here are two regions where the actual model fits very well:



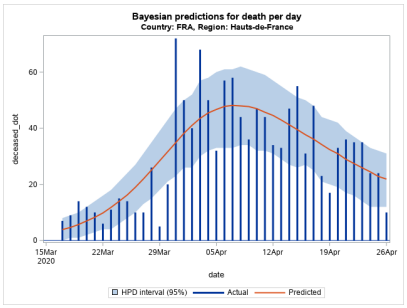
(a) Lombardy (b) New York State

Figure 3: Daily death counts (predicted vs actual)

Certainly in German and French regions the fitting was not so good (though not very bad), which can be seen on this two examples:



(a) Nordrhein-Westfalen

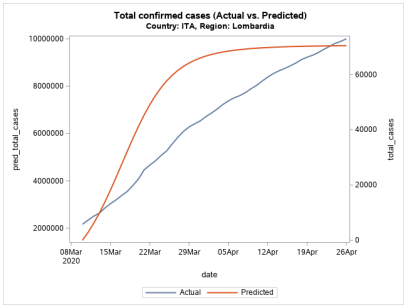


(b) Hauts-de-France

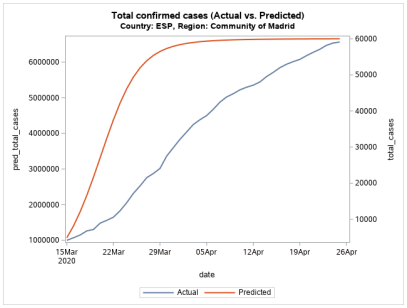
Figure 4: Daily death counts (predicted vs actual)

3.6 Modeling the infectious stock

Puting all these analytics together 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 and we can even compare it with the number of confirmed cases. What we can see on regions which began to strugle with the virus is that they have practically achieved herd immunity, the new confirmed cases and deaths are 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:



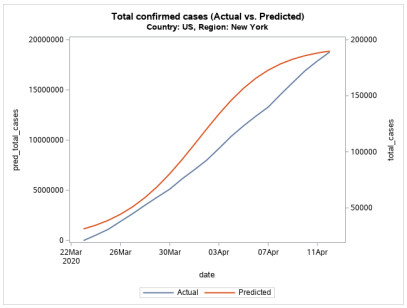
(a) Lombardy



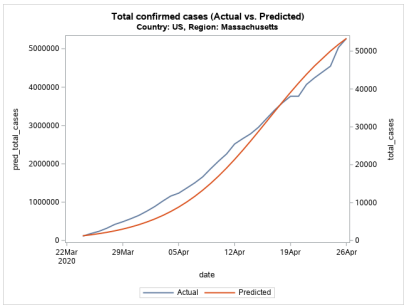
(b) Madrid

Figure 5: Total case counts (predicted vs actual)

On the other hand US has 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) New York State



(b) Massachusetts

Figure 6: 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.

3.7 Conclusions

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

- The R_0 value is much larger than previously thought
- The average length of infectious phase is slightly longer than previously thought (half-life is about 19 days)
- Mortality is about 0.15 %
- 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] Spanish Government, `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] Codes and data used for this report, `https://github.com/ggalfi/COVID19_SAS`