Quantifying the non-reported new daily cases of COVID-2019 by region in Spain at a real-time

The present outbreak of COVID-19 disease, caused by the SARS-CoV-2 virus, has put the planet in quarantine. On January 30, 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a "public health emergency of international concern", and then a pandemic on March 11.

Spain has become the fifth country worldwide with more infected cases, officially registering over 13 thousand cases in a short time. Although many critical and severe measures have been considered from the authorities to lessen the impact of the outbreak and help flatten the curve, they rely on numbers that could be unreliable and therefore misrepresent the implications of such pandemic.

Counts in Spain due to the protocols used for testing, mainly include individuals with severe symptoms. The authorities have juste announced a new protocol with rapid tests to be implementend in a few days elpais.com.

Given the nature of our data, we can guess that the estimated number of cases that we are finding are in fact potentially severe cases, and presumably the size of the infected population (asymptomatic) is even higher.

Accordingly, the current analysis aims to update the situation concerning COVID-19 daily, and particularly quantify the potential under-reporting in the official registered cases by region in Spain. Results herein can help to have a more realistic picture of the pandemic at a real time as well as to more accurately estimate essential measures such as the basic reproduction number or the fatality rate that are used for practitioners and politicians to make decisions.

The data for the analysis have been extracted from eldiario.es, where official data are gathered.

Notice that this analysis con be easily reproduced for other countries.

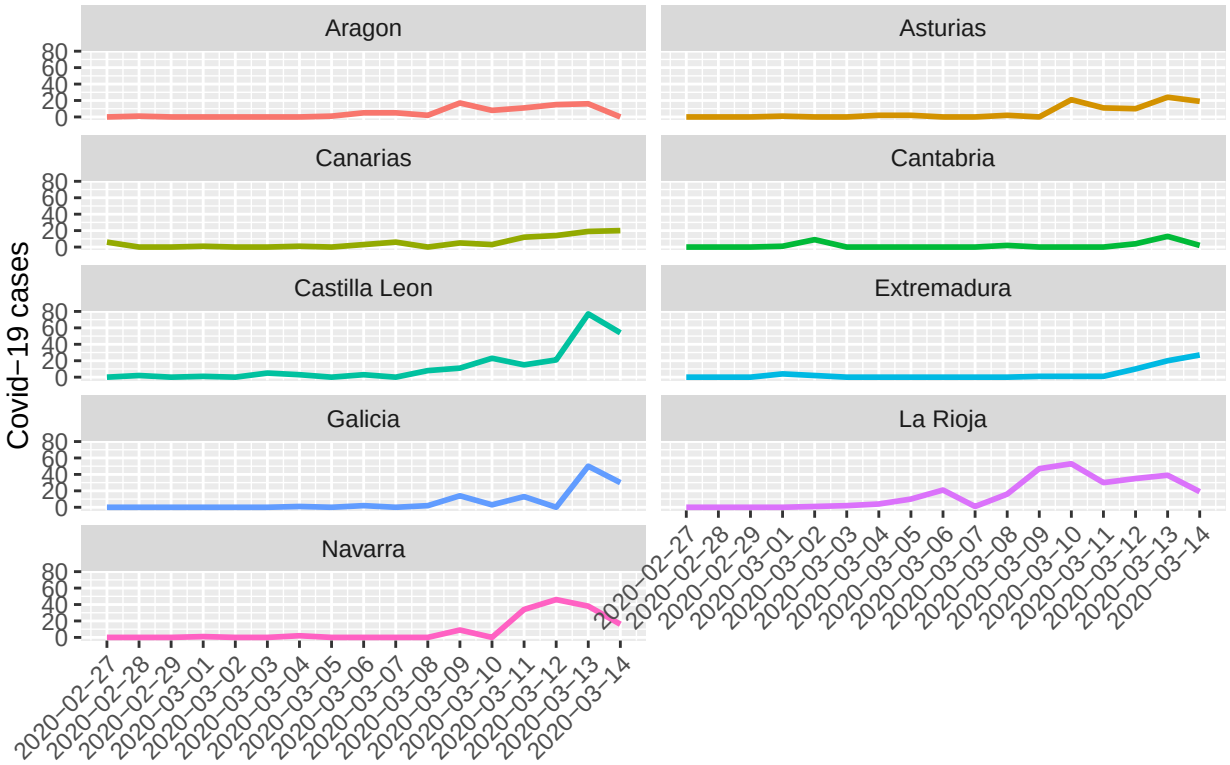

Figure 1 (a): Daily COVID−19 cases from 27−02−2020 to 14−03−2020

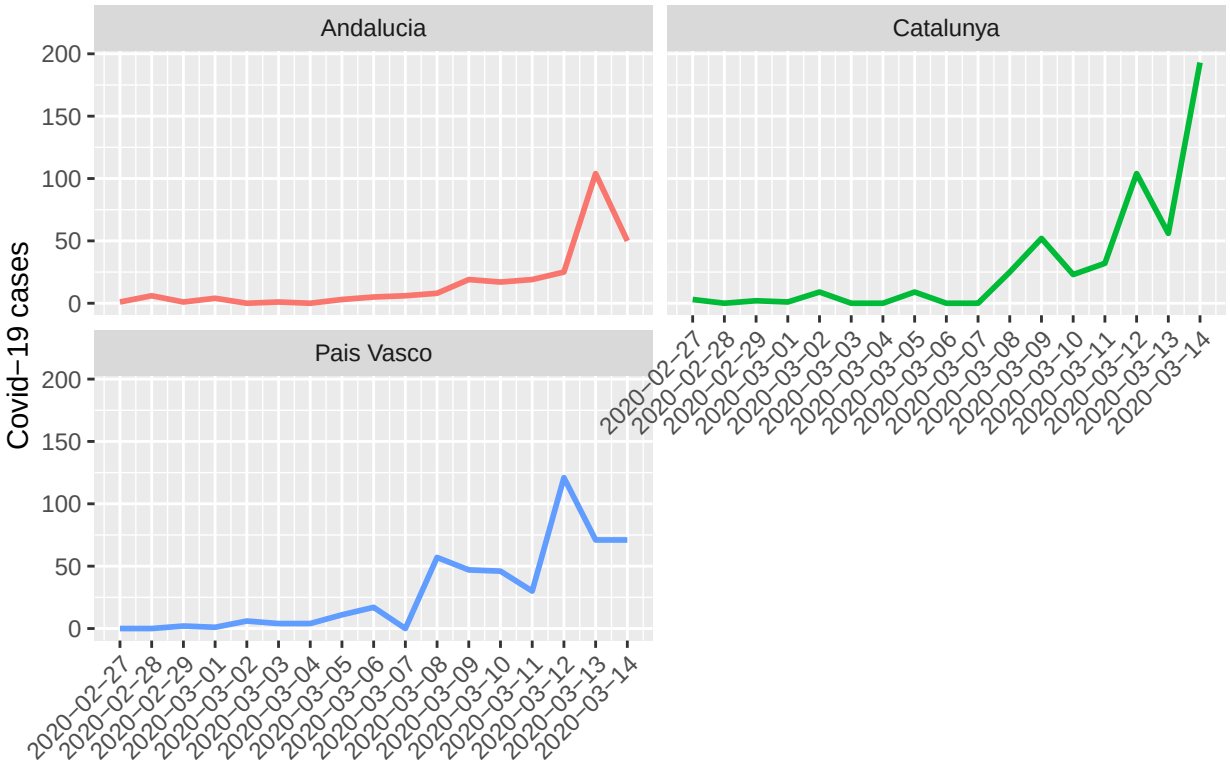

Figure 1 (b): Daily COVID−19 cases from 27−02−2020 to 14−03−2020

% latex table generated in R 3.6.2 by xtable 1.8-4 package % Thu Mar 19 19:15:12 2020

	minimum	mean	median	maximum	standard deviation	dispersion index
Andalucia	0.00	15.82	6.00	104.00	26.08	42.99
Aragon	0.00	4.76	1.00	17.00	6.25	8.20
Asturias	0.00	5.41	1.00	24.00	8.34	12.87
Canarias	0.00	5.29	3.00	20.00	6.81	8.75
Cantabria	0.00	1.82	0.00	13.00	3.70	7.49
Castilla Leon	0.00	13.12	3.00	77.00	21.43	35.01
Catalunya	0.00	29.94	9.00	193.00	50.63	85.62
Extremadura	0.00	3.88	0.00	27.00	7.89	16.03
Galicia	0.00	6.76	0.00	50.00	13.70	27.76
La Rioja	0.00	16.35	10.00	53.00	18.20	20.25
Navarra	0.00	8.59	0.00	46.00	15.42	27.68
Pais Vasco	0.00	28.71	11.00	121.00	35.06	42.83

Table 1: Summary of the daily COVID-19 cases from 27-02-20 to14-03-2020 by region in Spain

If the under-reporting is ignored, the daily counts can be appropriately modeled following: $exp(\alpha_0 + \alpha_1 t)$, since the number of daily COVID-19 cases overtime properly growths exponentially according to Figure 1.

However, if we consider that the official number of daily cases does not reflect the total number of cases (e.g., a proportion of the cases is not observed, and thus the data are misreported), the model above does not make any sense, and therefore a more appropriate alternative should be considered.

We shall base all the subsequent analysis in a model introduced by Fernández-Fontelo et al. (2016).

In that model, two different processes are considered: X_n which is the true process but unobserved (latent),

and Y_n which is observed and potentially under-reported. In this application, the latent process is assumed to be Poisson distributed with time-dependent rate, $\lambda_t = exp(\beta_0 + \beta_1 t)$. The observed process will always be lower or equal than the latent process (due to the under-reporting) in such a way that Y_n will be equal than X_n (non under-reporting) with probability $1 - \omega$; or Y_n is $q \circ X_n$ with probability ω. Parameters ω and q quantify the overall frequency and intensity of the phenomenon, which roughly speaking describe respectively the number of times the observed counts are not equal to the real ones, and the distance between the real and observed processes.

Table 2 shows the estimates of the models parameters by region. They can be interpreted as follows quickly. For instance, for Catalunya, the overall frequency and intensity of under-reporting are roughly 0.55 and 0.37, respectively. This means that 55% of the counts throughout the period are not entirely reported and that averagely the proximity between the real and observed processes is 0.37 (being 1 when two processes are identical).

% latex table generated in R 3.6.2 by xtable 1.8-4 package % Thu Mar 19 19:15:12 2020

	α	β_0	β_1	ω	q	AIC
Andalucia		0.361	0.2684	0.8442	0.3371	100.2
s.e. (Andalucia)		0.3528	0.0218	0.1014	0.0395	
Aragon		-1.4641	0.2973	0.7565	0.7077	70.6
s.e. (Aragon)		0.5553	0.0384	0.2829	0.112	
Asturias		-2.2792	0.3476	0.8187	0.6445	81.6
s.e. (Asturias)		0.6227	0.0398	0.1783	0.0913	
Canarias		0.2255	0.1601	0.4628	0.1397	87
s.e. (Canarias)		0.6873	0.0453	0.1806	0.0864	
Cantabria		1.5157	0.0662	0.8773	0.0693	61.9
s.e. (Cantabria)		0.5268	0.0819	0.0281	0.0374	
Castilla Leon		-1.503	0.3614	0.6468	0.4852	94.6
s.e. (Castilla Leon)		0.5358	0.0336	0.1999	0.0532	
Catalunya		0.0078	0.3032	0.548	0.3705	141.1
s.e. (Catalunya)		0.2634	0.0172	0.1376	0.0411	
Extremadura		-0.1971	0.0487	0.1969	0.6811	62.7
s.e. (Extremadura)		0.7041	0.0446	0.1229	0.0309	
Galicia		-3.665	0.465	0.573	0.3378	83.6
s.e. (Galicia)		0.8122	0.0508	0.2087	0.0563	
La Rioja		0.8927	0.2165	0.7927	0.4244	137.6
s.e. (La Rioja)		0.279	0.0186	0.1096	0.0366	
Navarra		-2.0804	0.3724	0.5651	0.1729	83.7
s.e. (Navarra)		0.6226	0.04	0.1553	0.0428	
Pais Vasco		-0.1955	0.3327	0.3879	0.3776	127.2
s.e. (Pais Vasco)		0.2757	0.0198	0.1633	0.036	

Table 2: Estimates of under-reporting parameters by region in Spain

Using the Viterbi algorithm, the model also enables reconstructing the most likely sequence of real COVID-19 cases throughout the study. This allows us to have an estimated time series of truly daily cases and evaluate the impact of under-reporting over measures such as the basic reproduction number. Figure 2 shows the observed and reconstructed series over time by region.

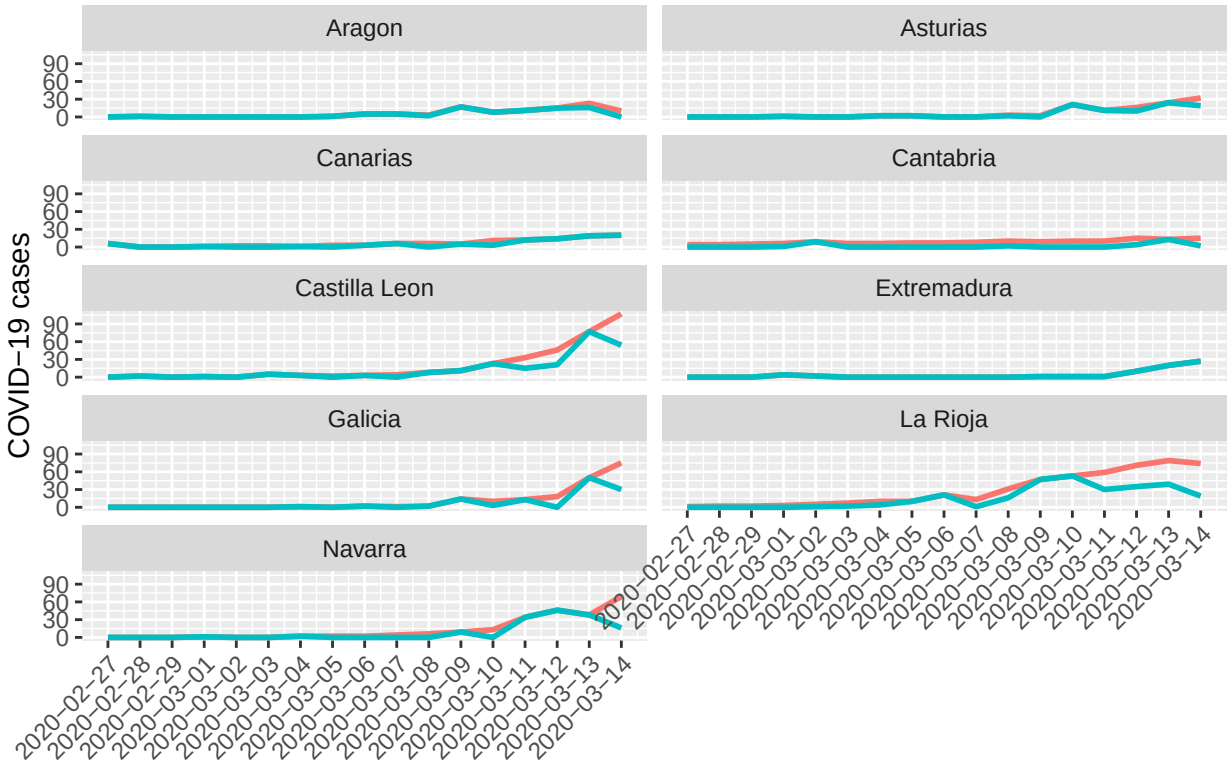

Figure 2 (a): Truly daily cases from 27−02−2020 to 14−03−2020

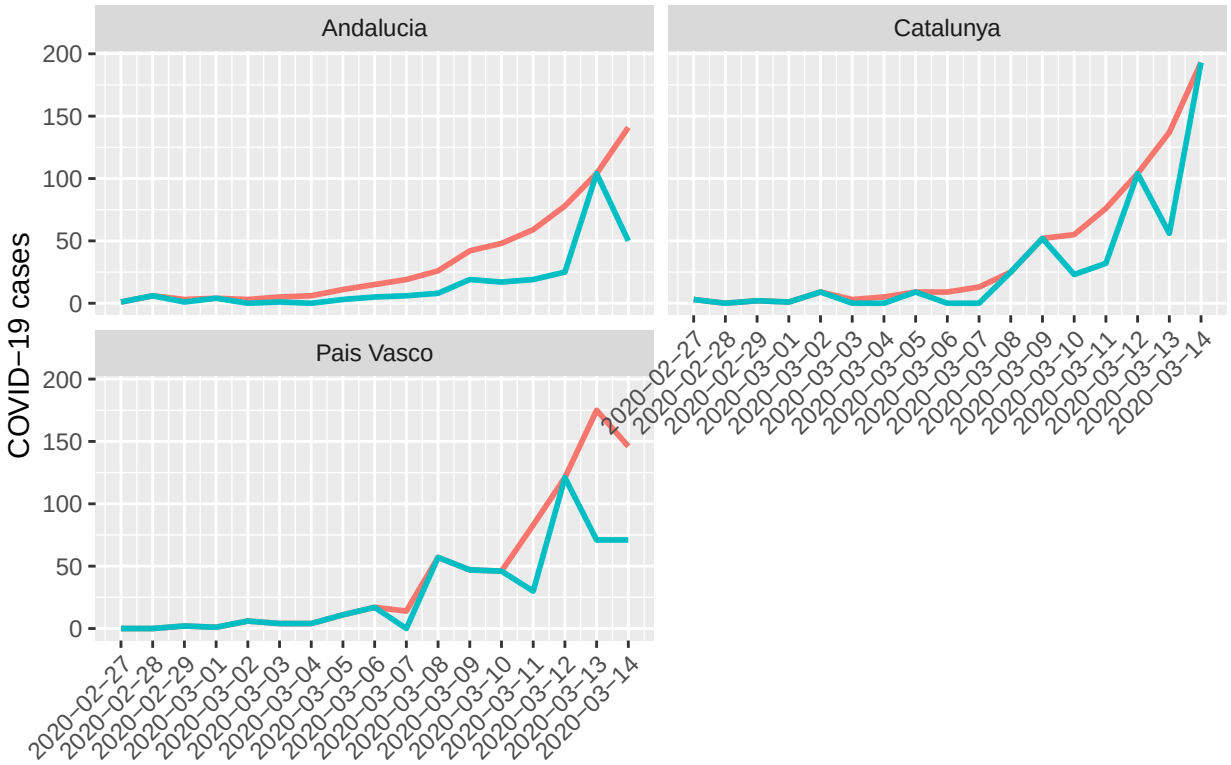

Figure 2 (b): Truly daily cases from 27−02−2020 to 14−03−2020

Using the Viterbi algorithm, the model also enables reconstructing the most likely sequence of real COVID-19 cases throughout the study. This allows us to have an estimated time series of truly daily cases and evaluate the impact of under-reporting over measures such as the basic reproduction number. Figure 2 shows the observed and reconstructed series over time by region.

Table 3 shows the percentages of means counts that are not covered by the official registers. Thus, the highest the rate, the lower is the coverage, and therefore the severe is the impact of the under-reporting.

% latex table generated in R 3.6.2 by xtable 1.8-4 package % Thu Mar 19 19:15:13 2020

	observed mean	true mean	% not covered
Andalucia	15.82	33.59	52.89
Aragon	4.76	5.82	18.18
Asturias	5.41	6.71	19.30
Canarias	5.29	6.53	18.92
Cantabria	1.82	8.47	78.47
Castilla Leon	13.12	19.12	31.38
Catalunya	29.94	40.94	26.87
Extremadura	3.88	3.88	0.00
Galicia	6.76	10.94	38.17
La Rioja	16.35	28.71	43.03
Navarra	8.59	13.29	35.40
Pais Vasco	28.71	43.18	33.51

Table 3: Estimate mean of non-coverage of cases of COVID-19 in Spain

It is instructive to see what the difference would be on epidemic spread by fitting an epidemic model to the reconstructed series of counts and the observed counts recorded by public agencies. We fit the classic

SIR (Susceptible-Infectious-Recovered) model. Table 4 shows the basic reproduction rate by using the reconstructed series (RE) and the observed (RR).