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 thousands of 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.

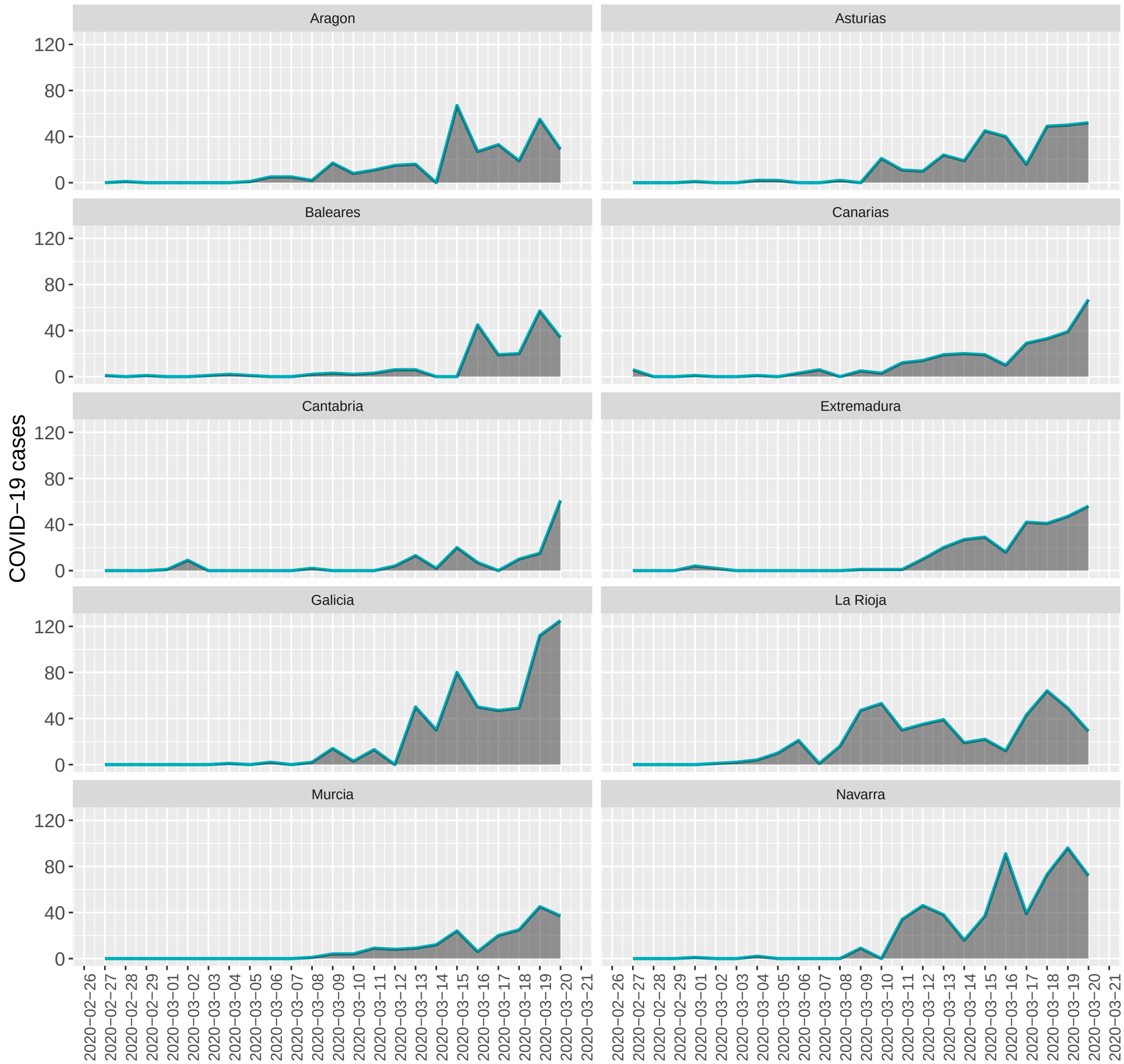

Figure 1 (a): Daily COVID−19 cases from 27−02−2020 to 20−03−2020

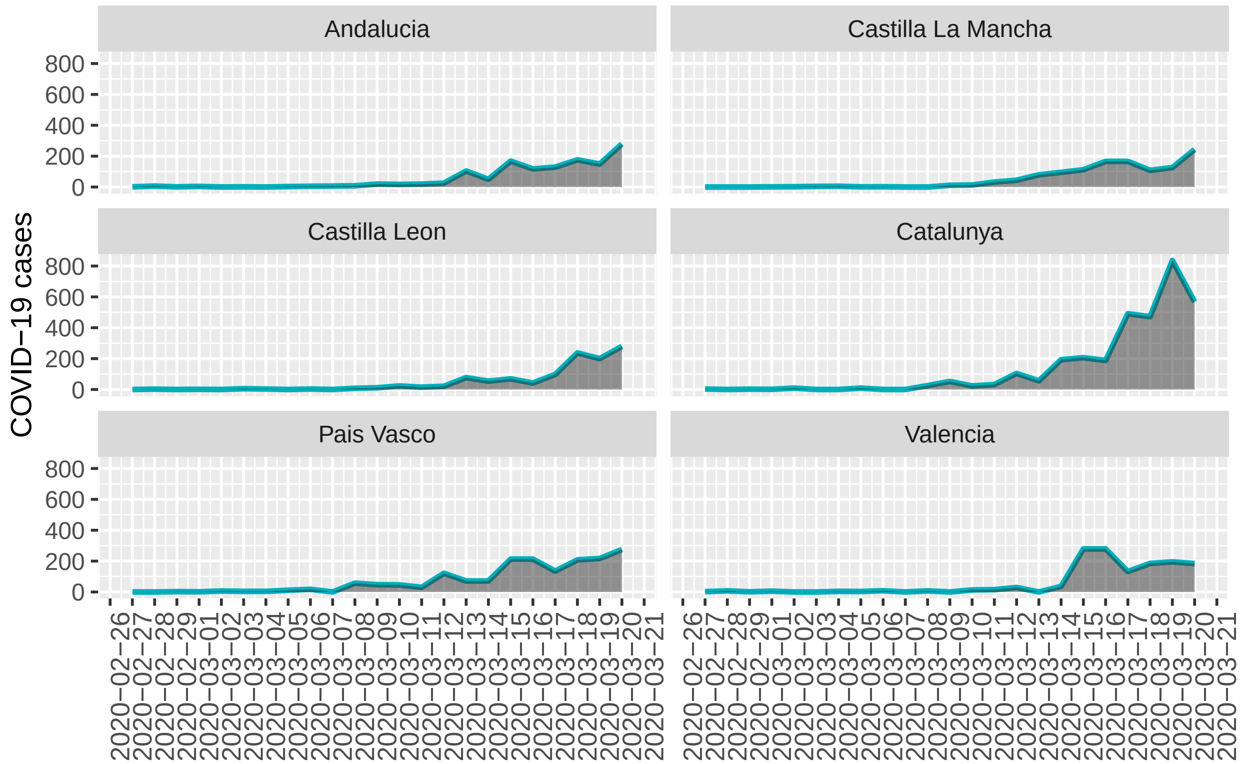

Figure 1 (b): Daily COVID−19 cases from 27−02−2020 to 20−03−2020

	minimum	mean	median	maximum	standard deviation	dispersion index
Andalucia	0.00	55.96	17.00	279.00	77.66	107.79
Aragon	0.00	13.52	5.00	67.00	18.27	24.67
Asturias	0.00	14.96	2.00	52.00	19.01	24.17
Baleares	0.00	8.83	2.00	57.00	15.82	28.35
Canarias	0.00	12.48	6.00	67.00	16.57	21.99
Cantabria	0.00	6.26	0.00	61.00	13.26	28.09
Castilla La Mancha	0.00	52.61	11.00	243.00	70.44	94.31
Castilla Leon	0.00	49.87	11.00	279.00	80.63	130.36
Catalunya	0.00	142.17	25.00	836.00	229.08	369.11
Extremadura	0.00	12.91	1.00	56.00	18.22	25.71
Galicia	0.00	25.13	2.00	125.00	37.50	55.96
La Rioja	0.00	21.61	19.00	64.00	20.14	18.76
Murcia	0.00	8.87	4.00	45.00	12.84	18.58
Navarra	0.00	24.09	2.00	96.00	32.06	42.67
Pais Vasco	0.00	76.04	46.00	275.00	89.35	104.98
Valencia	0.00	60.17	7.00	279.00	95.03	150.08

Table 1: Summary of the daily COVID-19 cases from 27-02-20 to 20-03-2020 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.

	α	β_0	β_1	ω	q	AIC
Andalucia		0.3279	0.2479	0.8572	0.6122	212.7
s.e. (Andalucia)		0.1728	0.0087	0.0883	0.0211	
Aragon		-0.3013	0.2046	0.3709	0.2282	166.5
s.e. (Aragon)		0.33	0.0175	0.1347	0.0526	
Asturias		-0.3282	0.2051	0.4396	0.2662	134.1
s.e. (Asturias)		0.3922	0.0202	0.1492	0.0684	
Baleares		-0.7201	0.2237	0.8452	0.2877	120
s.e. (Baleares)		0.6296	0.0294	0.1078	0.0559	
Canarias		0.0592	0.1659	0.3799	0.284	121.6
s.e. (Canarias)		0.4031	0.021	0.1475	0.0848	
Cantabria		1.8292	0.0344	0.6844	0.0411	101.5
s.e. (Cantabria)		0.4494	0.0244	0.1019	0.022	
Castilla La Mancha		-0.1119	0.2631	0.5525	0.478	179.6
s.e. (Castilla La Mancha)		0.1969	0.0102	0.1294	0.0261	
Castilla Leon		-0.6833	0.2945	0.7692	0.5552	164.8
s.e. (Castilla Leon)		0.2506	0.0121	0.1235	0.0286	
Catalunya		-0.3254	0.31972	0.502	0.552	274.2
s.e. (Catalunya)		0.1155	0.0057	0.1148	0.0187	
Extremadura		0.0904	0.173	0.518	0.0471	108.7
s.e. (Extremadura)		0.4445	0.0229	0.114	0.0292	
Galicia		-0.8934	0.2656	0.5915	0.4145	171.8
s.e. (Galicia)		0.3198	0.0162	0.1364	0.0478	
La Rioja		2.1865	0.0921	0.5259	0.2519	208.8
s.e. (La Rioja)		0.2126	0.0118	0.1116	0.0357	
Murcia		-2.2174	0.2724	0.1708	0.2922	87.9
s.e. (Murcia)		0.4695	0.024	0.1459	0.1265	
Navarra		-0.7687	0.2661	0.7357	0.5638	186.3
s.e. (Navarra)		0.262	0.0133	0.1173	0.0353	
Pais Vasco		1.2253	0.2143	0.7241	0.6253	260.2
s.e. (Pais Vasco)		0.141	0.0072	0.1089	0.0196	
Valencia		4.4696	0.0447	0.7727	0.0572	377.4
s.e. (Valencia)		0.2823	0.014	0.0893	0.0096	

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.

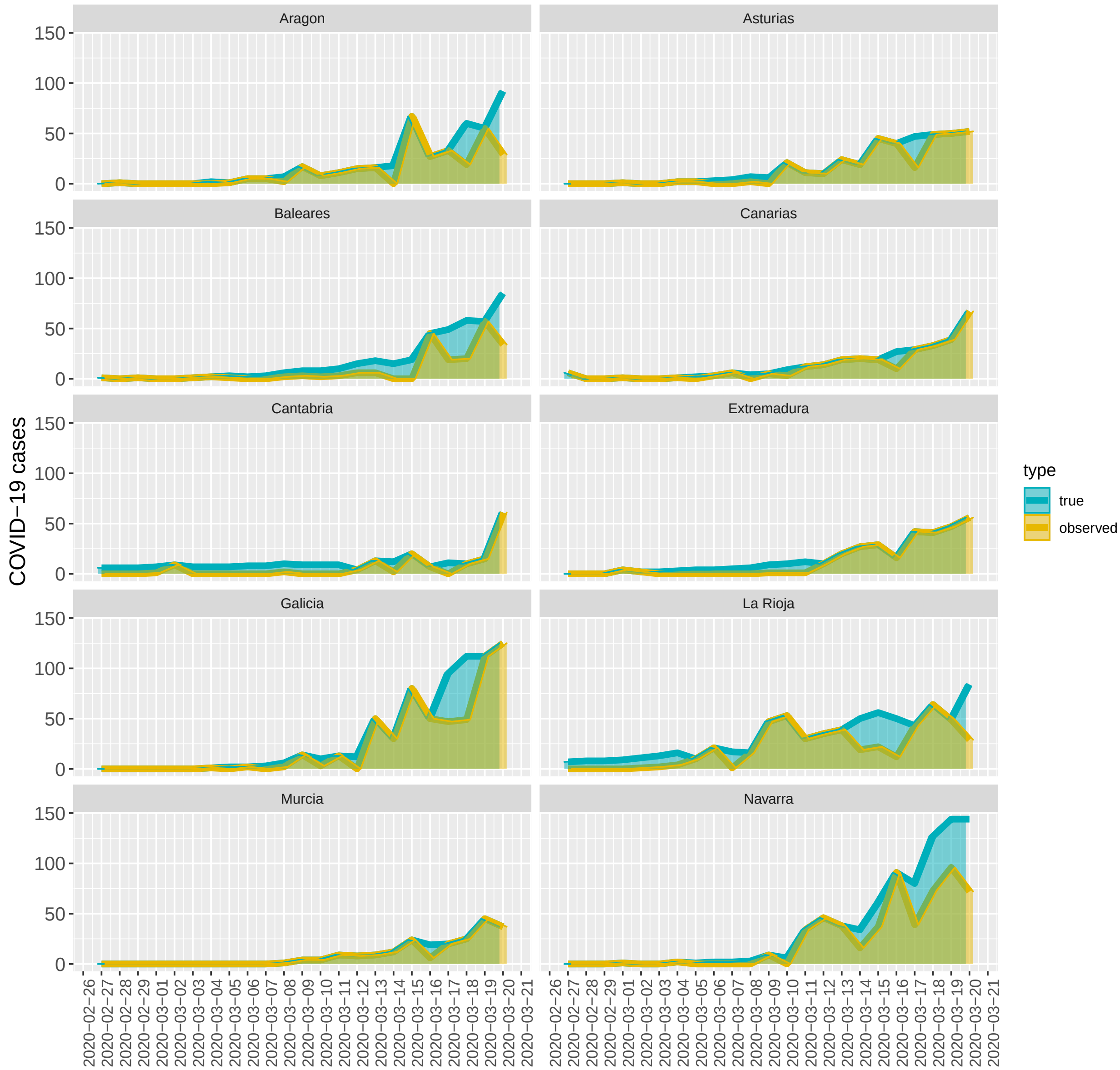

Figure 2 (a): Observed and truly daily cases

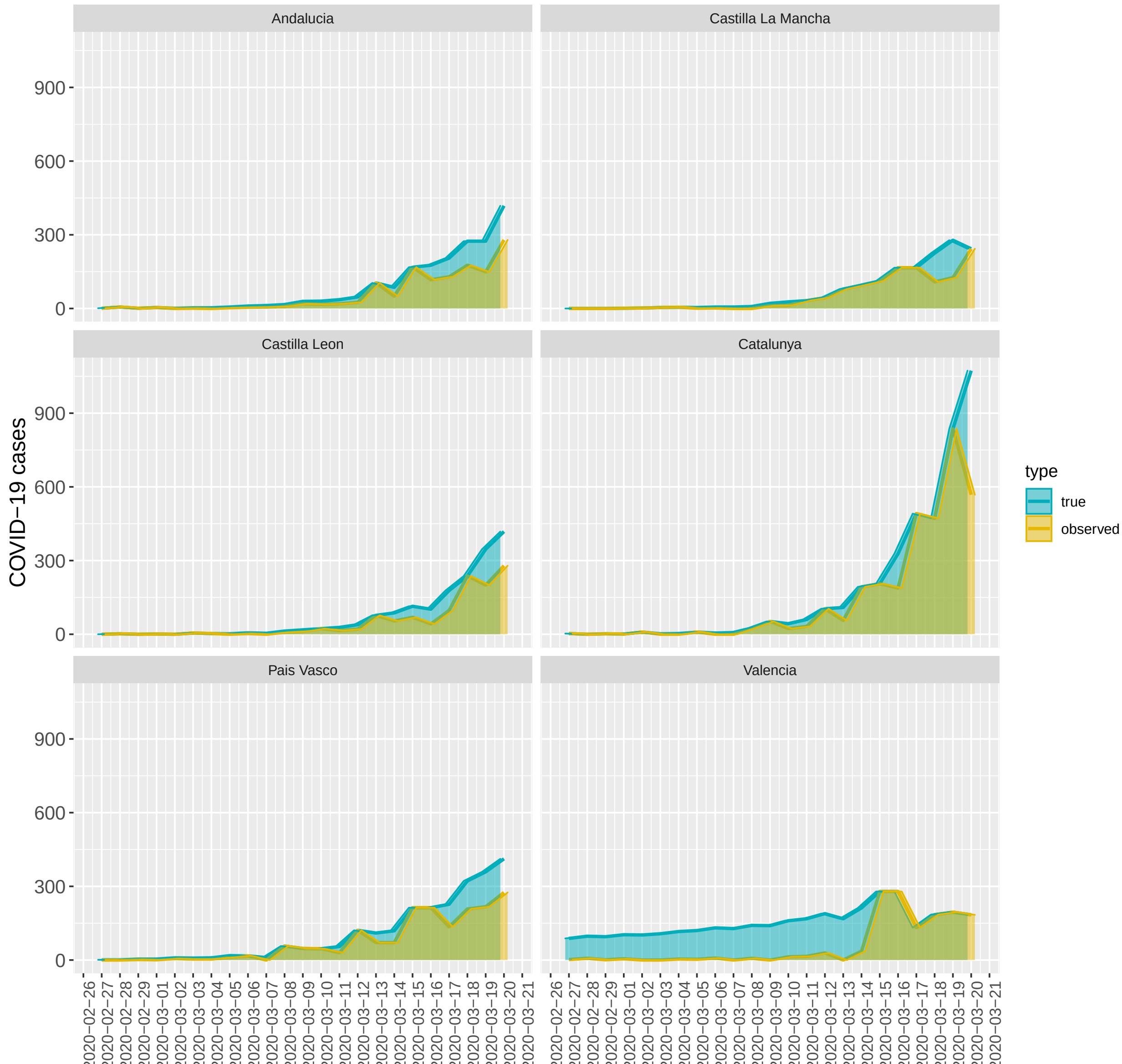

Figure 2 (b): Observed and truly daily cases

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.

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).

	observed mean	true mean	% not covered
Andalucia	55.96	83.04	32.62
Aragon	13.52	19.13	29.32
Asturias	14.96	17.09	12.47
Baleares	8.83	17.65	50.00
Canarias	12.48	13.74	9.18
Cantabria	6.26	11.35	44.83
Castilla La Mancha	52.61	66.26	20.60
Castilla Leon	49.87	74.04	32.65
Catalunya	142.17	175.35	18.92
Extremadura	12.91	15.17	14.90
Galicia	25.13	31.17	19.39
La Rioja	21.61	32.00	32.47
Murcia	8.87	9.43	5.99
Navarra	24.09	35.87	32.85
Pais Vasco	76.04	103.57	26.57
Valencia	60.17	153.00	60.67

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