Cumulated burden of Covid-19 in Spain from a Bayesian perspective

David Moriña[1,2*], Amanda Fernández-Fontelo[3], Alejandra Cabaña[2], Argimiro Arratia[4], Gustavo Ávalos[4] and Pedro Puig[2]

[1]Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB), Barcelona, Spain

[2]Centre de Recerca Matemàtica (CRM), Barcelona Graduate School of Mathematics (BGSMath), Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain

[3]Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

[4]Department of Computer Science, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

*Corresponding author: David Moriña

E-mail: dmorina@ub.edu

Facultat d'Economia i Empresa, Universitat de Barcelona, 08034, Barcelona, Spain

Tel. +34 934 021 395436

Abstract

Background: The main goal of this work is to estimate the actual number of cases of Covid-19 in Spain in the period 01-31-2020 / 06-01-2020 by Autonomous Communities. Based on these estimates, this work allows us to accurately re-estimate the lethality of the disease in Spain, taking into account unreported cases. ***Methods***: A hierarchical Bayesian model recently proposed in the literature has been adapted to model the actual number of Covid-19 cases in Spain. ***Results***: The results of this work show that the real load of Covid-19 in Spain in the period considered is well above the data registered by the public health system. Specifically, the model estimates show that, cumulatively until June 1st, 2020, there were 2 425 930 cases of Covid-19 in Spain with characteristics similar to those reported (95% credibility interval: 2 148 261 2 813 864), from which were actually registered only 518 664. ***Conclusions***: Considering the results obtained from the second wave of the Spanish seroprevalence study, which estimates 2 350 324 cases of Covid-19 produced in Spain, in the period of time considered, it can be seen that the estimates provided by the model are quite good. This work clearly shows the key importance of having good quality data to optimize decision-making in the critical context of dealing with a pandemic.

Keywords: Covid-19, Bayesian methods, public health, infections, underreporting

1. Introduction

SARS-CoV-2 belongs to the family of betacoronavirus and has been identified as the cause of Covid-19 disease, which can affect the lower respiratory tract and in some cases progress to pneumonia in humans. In particular, it has been identified as the causative agent of an unprecedented outbreak of pneumonia in Wuhan City, province of Hubei in China starting in December 2019[1] and spreading rapidly all over the world and being declared as a pandemic by World Health Organization (WHO) on 2020 March 11th. Considering that many cases run without developing symptoms beyond those of MERS-CoV, SARS-CoV or pneumonia due to other causes, it is reasonable to assume that the incidence of this disease has been under-registered, especially at the beginning of the outbreak.[2] Similarly, as many other countries' health systems were stressed to the limit of their capacity by the pandemic, it became clear that providing researchers and general public with reliable data was almost impossible. Spain is, to the date, among the most affected European countries in terms of number of registered cases, hospitalizations and deaths and there has been a debate to what extent officially reported data can be trusted.[3] This work aims to estimate the real burden of Covid-19 in Spain by Autonomous Community (CCAA), considering the data officially reported by the Spanish Ministry of Health (which has been reported by each CCAA health department) and to compare these estimates to the results provided by the second wave of the seroprevalence study conducted from May 18th to June 1st.[4] In this study, 63 564 participants were recruited with a participation rate among eligible individuals around 66.5%. Globally, the estimated prevalence of IgG antibodies against SARS-Cov-2 in Spain is around 5.2% (95%CI: 4.6% - 5.4%).

2. Methods

A new Bayesian hierarchical framework to analyze potentially under-reported count data was recently introduced.[5] It was originally used to estimate unreported cases of tuberculosis in Brasil, but we have adapted it to use it in the context of Covid-19 disease in Spain by CCAA. A limitation of this methodology is that the spatial effect can only be estimated on regions with at least one neighbor, so isolated CCAA cannot be included (Islas Baleares, Canarias, Ceuta and Melilla), although the incidence of the disease in these regions is much smaller than in peninsular ones. All Covid-19 cases reported by the Spanish Ministry of Health through the Instituto de Salud Carlos III by CCAA in the period 01-31-2020 to 06-01-2020 accessed by July 24th, 2020 (the data are being updated retrospectively as new information for some CCAA is available) were included in this work. The model allows for the inclusion of covariates on the true count-generating process and on the underreporting mechanism as well. Average, minimum and maximum temperature per CCAA and day (as reported by the Agencia Estatal de Meteorología[6]) were included to evaluate their potential impact on the number of Covid-19 cases as well as an indicator for the non-pharmaceutical interventions undertaken by the Spanish government (no intervention until March 15th, declaration of the emergency state from March 16th to March 30th and from April 13th to June 1st, mandatory confinement from March 31st to April 12th). Ratio of PCR and antibodies tests per 1 000 habitants were included as covariates that might have an impact on the underreporting mechanism. Technical details are available in Appendix B (Supplementary material).

3. Results

It can be seen that the number of registered cases represent only a small fraction of the actual burden of the disease in all CCAA (Fig. 1). These unreported cases can be interpreted as asymptomatic or with mild symptoms or even cases with similar clinical characteristics than those that were registered, and the causes for un-reporting might be multiple -patients with unusual symptoms could have been misdiagnosed, limit stress of the public health system at some points of time, among others.

By considering these unreported cases as well, it can be seen that the estimates found in this study are very similar to the results provided by the seroprevalence study conducted in Spain[4] and that in most CCAA the projection of seroprevalence study yields 95% confidence intervals with non-empty intersection with 95% credible intervals (CrI) provided by the present study, as can be seen in Table 1.

Having accurate estimates for the actual number of cases is also useful to estimate lethality associated to the disease, as it seems to be overestimated in Spain when using the officially reported cases compared to lethality estimated in other countries. Estimates for each CCAA and Spain are provided in Table 1, and it can be seen that they are much more consistent with those reported in countries with similar characteristics[7], all the cases being around 1% instead of values as high as 6.84% in Castilla La Mancha when using only registered data. The overall estimate for Covid-19 lethality in Spain is 1.10% (95% CrI: 0.95% - 1.25%).

The impact of the considered covariates on the actual Covid-19 incidence is shown in Fig. 2. It can be seen (right bottom) that the incidence rate is increasing until the declaration of the emergency state (1.0) and then decreasing drastically. Regarding the temperature effect, there is no clear pattern. Maximum and minimum temperature seem to have no effect on the Covid-19 incidence, while the decreasing incidence that can be seen when the average temperature increases disappeared in a sensitivity analysis replacing non-pharmacological interventions by a sequential indicator of time as covariate. Therefore, temperature is probably acting just as a confusing factor here.

Additionally, it can be seen (Fig. S1 in the Appendix A, supplementary material) that the probability of reporting a case increases as the number of performed PCR and antibodies tests increases, as could be expected.

3.1. Model checking

The goodness of fit of the proposed model can be checked by obtaining predictions for the registered values and comparing them to the actual registered values. Fig. S2 (Appendix A, supplementary material) shows this comparison for each CCAA, and it can be seen that predicted and actually registered values are very similar (perfect fit would be over the diagonal).

The goodness of fit of the model can also be assessed by checking whether summary statistics of the registered data are fitted properly by the model through replicates. In particular, Fig. S3 (Appendix A, supplementary material) shows how sample mean and variance are captured by the model, comparing the prior (top) and posterior (bottom) predictive distributions of both sample statistics and the mean squared error. It can be seen that the uncertainty in the parameters has been reduced considerably by the data, as the posterior predictive distribution are notably more precise than the corresponding priors, meaning that the model is fitting the data well.

4. Discussion

Dealing with under-reported data is very common in several fields including epidemiology, biomedical and social research among many others. It is known that predictions based on under-reported data might be severely biased if this issue is not taken into account at the modeling stage.[8] This is especially important when dealing with diseases with a huge number of asymptomatic cases, as the Covid-19, where the proportion of infected individuals developing no symptoms can be as high as 40-45%.[9] This concern has received a lot of attention recently in the biomedical and methodological literature, and several proposals have been done in order to model under-reported data, from Markov chain Monte Carlo methods[10] to time series analysis.[8,11]

This work shows that Covid-19 cases in Spain are severely under-reported and that an estimation of the unreported cases consistent with the results provided by the second wave of a nationwide seroprevalence study[4] can be achieved by means of a

hierarchical Bayesian methodology proposed very recently.[5] The results of this study also show that non-pharmaceutical interventions undertaken by the Spanish government and by regional administrations had a significant impact on Covid-19 incidence, as a monotonous decrease in the disease incidence following the implantation of mobility restrictions can clearly be seen. No impact of temperature could have been detected, as the apparent decrease in Covid-19 incidence for higher average temperatures was better explained by the sequential pass of time. The methodology also allows for the inclusion of covariates that might explain the under-reporting mechanism, and so, it can provide public health decision-makers with ways of improving the way data are registered. In this case, it can be seen that the more PCR and antibodies tests are conducted, the more likely is to report a case.

It is important to notice that the considerable differences in coincidence between the estimated number of cases provided by the Bayesian methodology and the seroprevalence study can be partially explained by the differences in how CCAA reported their data to the Spanish Ministry of Health. Some of them reported prevalent cases and some included, at least lately, asymptomatic cases tested positive while others reported only cases that required some kind of medical attention, so asymptomatic and some mild symptoms cases might be missing in the data provided by these regions.

One of the lessons that should certainly be learned from the current Covid-19 pandemic is that it is crucial to provide researchers with reliable data under extremely complex circumstances, in order to be able to assure public health decision makers are provided with the most reliable information at any time. When this is by no ways possible, the issue should be at least taken into account by using a model capable of accommodating under-reported data like the one used in this study.

Key points

- Only around 21% of Covid-19 cases were reported in Spain in the period 01-31-2020 to 06-01-2020.
- Decision making in the context of Public Health should be based on accurate data, whereas this work shows that this was hardly achieved in the Covid-19 pandemic.
- Temperature does not seem to have a relevant impact in Covid-19 incidence rate.
- Non-pharmaceutical interventions like mandatory confinement did effectively reduce Covid-19 incidence in Spain.

Data availability

All data used in this paper are publicly available from the cited sources and from the GitHub repository https://github.com/dmorinya/BayesCovidSpain.

Conflict of interest

None of the authors has any conflict of interest.

Funding

This work was supported by grant COV20/00115 from Instituto de Salud Carlos III. David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and Fundación Santander Universidades. A. Arratia and G. Ávalos acknowledge financial support from the Spanish Ministry of Science and Innovation (contract TIN2017-89244-R) and the Centre of Cooperation for Development (CCD-UPC). This work was partially supported by grant RTI2018-096072-B-I00 from the Spanish Ministry of Science and Innovation.

References

1. Sohrabi C, Alsafi Z, O'Neill N, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int J Surg*. 2020;76:71-76. doi:10.1016/j.ijsu.2020.02.034

2. Zhao, Musa, Lin, et al. Estimating the Unreported Number of Novel Coronavirus (2019-nCoV) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early Outbreak. *J Clin Med*. 2020;9(2):388. doi:10.3390/jcm9020388

3. Hyafil A, Moriña D. Análisis del impacto del confinamiento en el número de reproducción del SARS-Cov-2 en España. *Gac Sanit*. 2020. doi:10.1016/j.gaceta.2020.05.003

4. Ministerio de Ciencia e Innovación, Gobierno de España. Estudio ENE-COVID19: Segunda ronda estudio nacional de sero-epidemiología de la infección por SARS-COV-2 en España. 2020. Available at: https://www.mscbs.gob.es/ciudadanos/ene-covid/docs/ESTUDIO_ENECOVID19_SEGUNDA_RONDA_INFORME_PRELIMINAR.pdf (24 July 2020, date last accessed).

5. Stoner O, Economou T, Drummond Marques da Silva G. A Hierarchical Framework for Correcting Under-Reporting in Count Data. *J Am Stat Assoc*. 2019:1-17. doi:10.1080/01621459.2019.1573732

6. Agencia Estatal de Meteorología. Available at: http://www.aemet.es/ca/portada (24 July 2020, date last accessed).

7. Català M, Pino D, Marchena M, et al. Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers. *medRxiv* 2020; Preprint. doi:10.1101/2020.05.01.20087023

8. Fernández-Fontelo A, Cabaña A, Puig P, Moriña D. Under-reported data analysis with INAR-hidden Markov chains. *Stat Med*. 2016;35(26):4875-4890. doi:10.1002/sim.7026

9. Oran DP, Topol EJ. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Ann Intern Med*. 2020. doi:10.7326/m20-3012

10. Winkelmann R. Markov Chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empir Econ*. 1996;21(4):575-587. doi:10.1007/BF01180702

11. Moriña D, Fernández-Fontelo A, Cabaña A, Puig P. New statistical model for misreported data with application to current public health challenges. *arXiv* 2020; Preprint.

Table 1. Registered, estimated and projected from the ENE-COVID19 study cumulated COVID-19 cases by CCAA in the period 01-31-2020/06-01-2020. CrI stands for credible interval and CI stands for confidence interval. Spain excluding Islas Baleares, Canarias, Ceuta and Melilla.*

CCAA	Registered	Estimated (95% CrI)	Projection from ENE-COVID19 Study (95%CI)
Andalucía	32 878	142 294 (126 387 – 164 403)	244 013 (210 356 – 286 084)
Aragón	12 616	43 877 (37 333 – 52 029)	64 645 (51 452 – 83 115)
Cantabria	4 620	34 282 (27 959 – 41 659)	18 594 (12 203 – 27 311)
Castilla - La Mancha	43 080	211 286 (185 594 – 244 486)	209 385 (176 859 – 248 009)
Castilla y León	52 316	180 408 (161 644 – 204 663)	179 966 (155 971 – 206 361)
Cataluña	108 358	347 729 (316 690 – 395 782)	468 188 (399 111 – 552 616)
Comunidad Foral de Navarra	15 326	62 639 (53 621 – 73 665)	41 870 (32 056 – 54 300)
Comunidad Valenciana	29 302	132 087 (115 561 – 155 241)	135 102 (110 083 – 170 128)
Extremadura	11 150	37 611 (31 953 – 44 561)	35 234 (25 625 – 46 979)
Galicia	21 334	80 723 (70 711 – 93 246)	59 389 (45 891 – 75 586)
La Rioja	7 956	34 389 (28 780 – 41 473)	12 355 (8 870 – 16 790)
Madrid	141 312	938 391 (841 130 – 1 086 685)	759 627 (666 339 – 866 241)
País Vasco	28 704	129 780 (109 621 – 154 185)	81 688 (61 818 – 108 181)
Principado de Asturias	4 854	20 913 (17 238 – 25 289)	16 365 (11 251 – 23 524)
Región de Murcia	4 858	29 522 (24 041 – 36 495)	23 902 (16 433 – 37 347)
Spain*	518 664	2 425 930 (2 148 261 – 2 813 864)	2 350 324 (1 984 319 – 2 802 574)

Table 2. Observed and estimated lethality by CCAA and globally. CrI stands for credible interval. Spain excluding Islas Baleares, Canarias, Ceuta and Melilla.*

CCAA	Registered lethality (%)	Estimated lethality (%) (95% CrI)
Andalucía	4.27	0.99 (0.85 – 1.11)
Aragón	6.55	1.88 (1.59 – 2.21)
Cantabria	4.37	0.59 (0.48 – 0.72)
Castilla - La Mancha	6.84	1.39 (1.20 – 1.59)
Castilla y León	3.68	1.07 (0.94 – 1.19)
Cataluña	5.16	1.61 (1.41 – 1.76)
Comunidad Foral de Navarra	3.20	0.78 (0.67 – 0.91)
Comunidad Valenciana	4.55	1.01 (0.86 – 1.15)
Extremadura	4.56	1.35 (1.14 – 1.59)
Galicia	2.85	0.75 (0.65 – 0.86)
La Rioja	4.54	1.05 (0.87 – 1.25)
Madrid	6.15	0.93 (0.80 – 1.03)
País Vasco	4.96	1.10 (0.92 – 1.30)
Principado de Asturias	6.39	1.48 (1.23 – 1.80)
Región de Murcia	3.05	0.50 (0.41 – 0.62)
Spain*	5.16	1.10 (0.95 – 1.25)

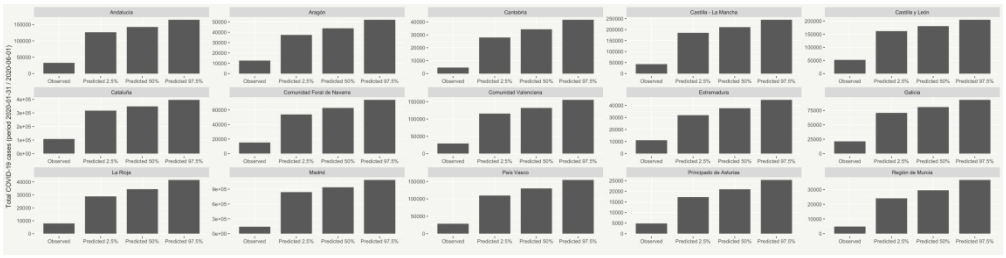

Figure 1: Registered (first bar) and estimated (median and 2.5th and 97.5th percentiles of the posterior distribution) cumulated Covid-19 cases in the period 01-31-2020/06-01-2020 in each CCAA.

299x74mm (300 x 300 DPI)

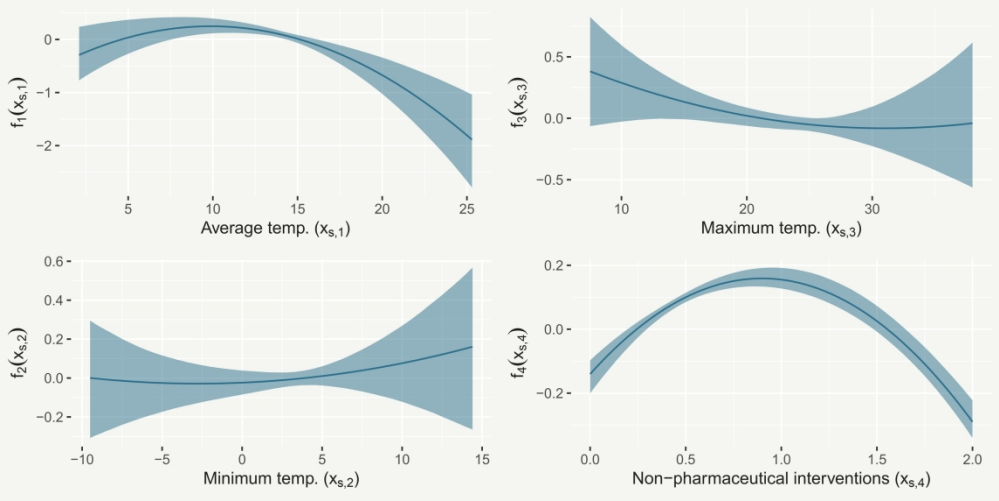

Figure 2: Posterior mean predictions (solid lines) of the effects of average, maximum, minimum air temperature and non-pharmaceutical interventions on the rate of Covid-19 incidence in Spain, with associated 95% CrIs.

228x114mm (300 x 300 DPI)