Estimated Covid-19 burden in Spain: ARCH underreported non-stationary time series

David Moriña[1], Amanda Fernández-Fontelo[2], Alejandra Cabaña[2], Argimiro Arratia[3], Pedro Puig[2]

[1] Universitat de Barcelona, Barcelona, Spain
[2] Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain
[3] Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail for correspondence: dmorina@ub.edu

Abstract: The problem of dealing with misreported data is very common in a wide range of contexts. The current situation caused by the Covid-19 worldwide pandemic is a clear example, where the data provided by official sources were not always reliable due to data collection issues and to the large proportion of asymptomatic cases. In this work, we explore the performance of Bayesian Synthetic Likelihood to estimate the parameters of a model capable of dealing with misreported information and to reconstruct the most likely evolution of the phenomenon.

Keywords: under-reported data; ARCH models; infectious diseases; Covid-19; Bayesian synthetic likelihood.

1 Introduction

The Covid-19 pandemic that is hitting the world since late 2019 has made evident that having quality data is essential in the decision making chain, especially in epidemiology but also in many other fields. Many methodological efforts have been made to deal with misreported Covid-19 data, following ideas introduced in the literature since the late nineties. As a large proportion of the cases run asymptomatically (Oran and Topol (2020)) and mild symptoms could have been easily confused with those of similar diseases at the beginning of the pandemic, its reasonable to expect that Covid-19 incidence has been notably underreported. Very recently several approaches based on discrete time series have been proposed (see Fernández-Fontelo et al. (2020)) although there is a lack of continuous time

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

series models capable of dealing with misreporting, a characteristic of the Covid-19 data and typically present in infectious diseases modeling. In this sense, a new model capable of dealing with temporal structures using a different approach is presented by Moriña et al. (2020). A typical limitation of these kinds of models is the computational effort needed in order to properly estimate the parameters. Synthetic likelihood is a recent and very powerful alternative for parameter estimation in a simulation based schema when the likelihood is intractable and, conversely, the generation of new observations given the values of the parameters is feasible. The method was introduced in Wood (2010) and placed into a Bayesian framework in Price et al. (2018), showing that it could be scaled to high dimensional problems and can be adapted in an easier way than other alternatives like approximate Bayesian computation (ABC).

2 Methods

AutoRegressive Conditional Heteroskedasticity (ARCH) models are a well-known approach to fitting time series data where the variance error is believed to be serially correlated. Consider an unobservable process X_t following an AutoRegressive ($AR(1)$) model with ARCH(1) errors structure, defined by

$$X_t = \phi_0 + \phi_1 \cdot X_{t-1} + Z_t,$$

where $Z_t^2 = \alpha_0 + \alpha_1 \cdot Z_{t-1}^2 + \epsilon_t$, being $\epsilon_t \sim N(\mu_\epsilon(t), \sigma^2)$. The process X_t represents the actual Covid-19 incidence. In our setting, this process X_t cannot be directly observed, and all we can see is a part of it, expressed as

$$Y_t = \begin{cases} X_t \text{ with probability } 1-\omega \\ q \cdot X_t \text{ with probability } \omega, \end{cases} \tag{1}$$

where q is the overall intensity of misreporting (if $0 < q < 1$ the observed process Y_t would be underreported while if $q > 1$ the observed process Y_t would be overreported) and ω can be interpreted as the overall frequency of misreporting (proportion of misreported observations). To model consistently the spread of the disease, the expectation of the innovations ϵ_t is linked to a simplified version of the well-known compartimental Susceptible-Infected-Recovered (SIR) model. At any time $t \in \mathbb{R}$ there are three kinds of individuals: Healthy individuals susceptible to be infected ($S(t)$), infected individuals who are transmitting the disease at a certain speed ($I(t)$) and individuals who have suffered the disease, recovered and cannot be infected again ($R(t)$). As shown by Fernández-Fontelo et al. (2020), the number of affected individuals at time t, $A(t) = I(t) + R(t)$ can be approximated by

$$A(t) = \frac{M^*(\beta_0, \beta_1, \beta_2, t) A_0 e^{kt}}{M^*(\beta_0, \beta_1, \beta_2, t) + A_0(e^{kt} - 1)}, \tag{2}$$

where $M^*(\beta_0, \beta_1, \beta_2, t) = \beta_0 + \beta_1 \cdot C_1(t) + \beta_2 \cdot C_2(t)$, being $C_1(t)$ and $C_2(t)$ dummy variables indicating if time t corresponds to a period where a mandatory confinment was implemented by the government and if the number of people with at least one dose of a Covid-19 vaccine in Spain was over 50% respectively. At any time t the condition $S(t) + I(t) + R(t) = N$ is fulfilled. The expression (2) allow us to incorporate the behaviour of the epidemics in a realistic way, defining $\mu_\epsilon(t) = A(t) - A(t-1)$, the new affected cases produced at time t.

The Bayesian Synthetic Likelihood (BSL) simulations are based on the described and the chosen summary statistics are the mean, standard deviation and the three first coefficients of autocorrelation of the observed process. Parameter estimation was carried out by means of the *BSL* (An et al. (2019)) package for R. Taking into account the posterior distribution of the estimated parameters, the most likely unobserved process is reconstructed, resulting in a probability distribution at each time point. The prior of each parameter is set to a uniform on the corresponding feasible region of the parameter space and zero elsewhere.

3 Results

This work focuses on the weekly Covid-19 incidence registered in Spain in the period (2020/02/23-2022/02/27). It can be seen in Figure 1 that the registered data (turquoise) reflect only a fraction of the actual incidence (red). The grey area corresponds to 95% probability of the posterior distribution of the weekly number of new cases (the lower and upper limits of this area represent the percentile 2.5% and 97.5% respectively), and the dotted red line corresponds to its median.

In the considered period, the official sources reported 11,056,797 Covid-19 cases in Spain, while the model estimates a total of 25,283,406 cases (only 43.73% of actual cases were reported). This work also revealed that while the frequency of underreporting is extremely high for all regions (values of $\hat{\omega}$ over 0.90 in all cases), the intensity of this underreporting is not uniform across the considered regions: Aragón is the CCAA with highest underreporting intensity ($\hat{q} = 0.05$) while Extremadura is the region where the estimated values are closest to the number of reported cases ($\hat{q} = 0.50$). Detailed underreported parameter estimates for each region can be found in Table 1. Although the main impact of the vaccination programmes can be seen in mortality data, the results of this work also showed a significant decrease in the weekly number of cases as well in all CCAA except Aragón. Figure 2 represents the estimated and registered processes globally for Spain.

Acknowledgments: Investigation funded by Fundación MAPFRE.

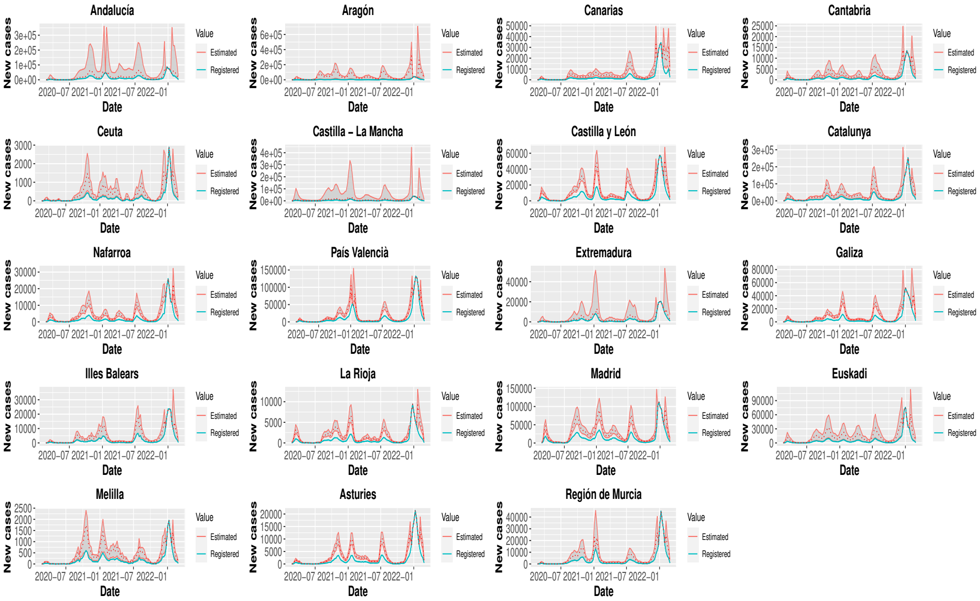

FIGURE 1. Registered and estimated weekly new Covid-19 cases in each Spanish region.

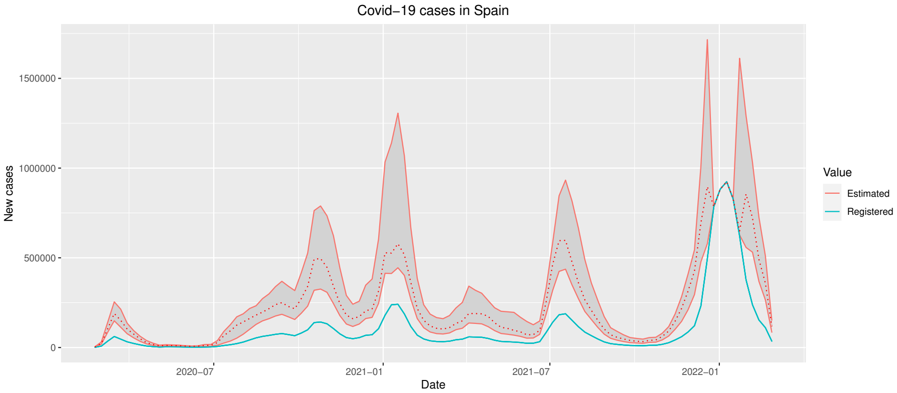

FIGURE 2. Registered and estimated weekly new Covid-19 cases globally in Spain.

References

An, Z., South, L.F. and Drovandi, C. (2019). BSL: An R Package for Efficient Parameter Estimation for Simulation-Based Models via Bayesian Synthetic Likelihood. *arXiv preprint*.

Fernández-Fontelo, A., Moriña, D., Cabaña, A., Arratia, A. and Puig P. (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE*, **15**, e0242956.

Moriña, D., Fernández-Fontelo, A., Cabaña, A. and Puig P. (2021) New statistical model for misreported data with application to current public health challenges. *Scientific Reports*, **11**, 23321.

Oran, D.P. and Topol, E.J. (2020). Prevalence of asymptomatic SARS-CoV-2 infection. *Annals of Internal Medicine*, **173(5)**, 362 – 367.

Price, L.F., Drovandi, C.C., Lee, A. and Nott D.J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, **27(1)**, 1 – 11.

Wood S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466(7310)**, 1102 – 1104.

TABLE 1. Estimated underreported frequency and intensity for each Spanish region. CI stands for Credible Interval.

Region	Parameter	Estimate (95% CI)
Andalucía	$\hat{\omega}$	0.96 (0.89 - 0.98)
	$\hat{q}$	0.45 (0.41 - 0.51)
Aragón	$\hat{\omega}$	0.97 (0.97 - 0.98)
	$\hat{q}$	0.05 (0.05 - 0.29)
Principado de Asturias	$\hat{\omega}$	0.98 (0.97 - 0.99)
	$\hat{q}$	0.35 (0.33 - 0.37)
Cantabria	$\hat{\omega}$	0.97 (0.95 - 0.99)
	$\hat{q}$	0.31 (0.28 - 0.35)
Castilla y León	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.38 (0.34 - 0.40)
Castilla - La Mancha	$\hat{\omega}$	0.96 (0.93 - 0.99)
	$\hat{q}$	0.36 (0.30 - 0.39)
Canarias	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.32 (0.29 - 0.36)
Catalunya	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.35 (0.33 - 0.39)
Ceuta	$\hat{\omega}$	0.97 (0.94 - 0.99)
	$\hat{q}$	0.30 (0.27 - 0.35)
Extremadura	$\hat{\omega}$	0.90 (0.49 - 0.97)
	$\hat{q}$	0.50 (0.39 - 0.70)
Galiza	$\hat{\omega}$	0.98 (0.97 - 0.99)
	$\hat{q}$	0.33 (0.30 - 0.35)
Illes Balears	$\hat{\omega}$	0.96 (0.88 - 0.98)
	$\hat{q}$	0.39 (0.35 - 0.61)
Región de Murcia	$\hat{\omega}$	0.97 (0.92 - 0.99)
	$\hat{q}$	0.43 (0.38 - 0.48)
Madrid	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.40 (0.36 - 0.42)
Melilla	$\hat{\omega}$	0.97 (0.95 - 0.99)
	$\hat{q}$	0.35 (0.32 - 0.38)
Comunidad Foral de Navarra	$\hat{\omega}$	0.98 (0.97 - 0.99)
	$\hat{q}$	0.31 (0.29 - 0.34)
Euskadi	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.30 (0.28 - 0.35)
La Rioja	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.32 (0.29 - 0.35)
País Valencià	$\hat{\omega}$	0.99 (0.97 - 0.99)
	$\hat{q}$	0.38 (0.37 - 0.41)