Moriña *et al.*
BMC Medical Research Methodology (2023) 23:75
https://doi.org/10.1186/s12874-023-01894-9

BMC Medical Research Methodology

RESEARCH

Open Access

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] and Pedro Puig[2,4]

Abstract

Background The problem of dealing with misreported data is very common in a wide range of contexts for different reasons. 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 high proportion of asymptomatic cases. In this work, a flexible framework is proposed, with the objective of quantifying the severity of misreporting in a time series and reconstructing the most likely evolution of the process.

Methods The performance of Bayesian Synthetic Likelihood to estimate the parameters of a model based on AutoRegressive Conditional Heteroskedastic time series capable of dealing with misreported information and to reconstruct the most likely evolution of the phenomenon is assessed through a comprehensive simulation study and illustrated by reconstructing the weekly Covid-19 incidence in each Spanish Autonomous Community.

Results Only around 51% of the Covid-19 cases in the period 2020/02/23–2022/02/27 were reported in Spain, showing relevant differences in the severity of underreporting across the regions.

Conclusions The proposed methodology provides public health decision-makers with a valuable tool in order to improve the assessment of a disease evolution under different scenarios.

Keywords Continuous time series, Mixture distributions, Under-reported data, ARCH models, Infectious diseases, Covid-19, Bayesian synthetic likelihood

Background

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. There is an enormous global concern around this disease, leading the World Health Organization (WHO) to declare public health emergency [1]. Many methodological efforts have been made to deal with misreported Covid-19 data, following ideas introduced in the literature since the late nineties [2–7]. These proposals range from the usage of multiplication factors [8] to Markov-based models [9, 10] or spatio-temporal models [11]. Additionally, a new R [12] package able to fitting endemic-epidemic models based on approximative maximum likelihood to underreported count data has been recently published [13]. However, as a large proportion of the cases run asymptomatically [14] and mild symptoms could have been easily confused with those of similar diseases at the beginning of the pandemic, it is reasonable to expect that Covid-19 incidence has been notably underreported. Very recently several approaches based on discrete time series have been proposed [15–17] although there is a lack of continuous time series models capable of dealing with misreporting, a characteristic of the Covid-19

*Correspondence:
David Moriña
dmorina@ub.edu
[1] Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB), Barcelona, Spain
[2] Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
[3] Department of Computer Science, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
[4] Centre de Recerca Matemàtica (CRM), Barcelona, Spain

data and typically present in infectious diseases modeling. In this sense, a new approach for longitudinal data not accounting for temporal correlations is introduced in [18] and a model capable of dealing with temporal structures using a different approach is presented in [19]. 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 [20] and placed into a Bayesian framework in [21], 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). The method takes a vector summary statistic informative about the parameters and assumes it is multivariate normal, estimating the unknown mean and covariance matrix by simulation to obtain an approximate likelihood function of the multivariate normal.

Methods

Auto Regressive 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

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

where $Z_t^2 = \alpha_0 + \alpha_1 \cdot Z_{t-1}^2 + \epsilon_t$, being $\epsilon_t \sim N\left(\mu_\epsilon(t), \sigma_\epsilon^2\right)$. 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

$$Y_t = \begin{cases} X_t \text{ with probability } 1-\omega \\ q \cdot X_t \text{ with probability } \omega, \end{cases} \tag{2}$$

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 $t \in R$ 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 in [17], the number of affected individuals at time t, $A(t) = I(t) + R(t)$ can be approximated by

$$A(t) = \frac{M^*(\beta_0, \beta_1, \beta_2, t) A_0 e^{kt}}{M^*(\beta_0, \beta_1, \beta_2, t) + A_0\left(e^{kt} - 1\right)}, \tag{3}$$

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 (3) allow us to incorporate the behavior 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 model 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* [22, 23] package for R [12]. 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 be uniform on the corresponding feasible region of the parameter space and zero elsewhere.

Results

This section presents the results of the analyses using the proposed methodology over a real data set and they are compared to the most common alternatives. The performance of the method is also studied by means of a comprehensive simulation study, with and without covariates.

The performance and an application of the proposed methodology are studied through a comprehensive simulation study and a real dataset on Covid-19 incidence in Spain on this Section.

Simulation study

Although the estimation method is already known and has been tested before, to the best of our knowledge it has never been used in the context of ARCH time series, and therefore a thorough simulation study has been conducted to ensure that the model behaves as expected, including $ARCH(1)$, $AR(1)$, $MA(1)$ and $ARMA(1,1)$ structures for the hidden process X_t defined as

$$\begin{gathered} X_t = \phi_0 + \phi_1 \cdot X_{t-1} + Z_t, Z_t^2 = \alpha_0 + \alpha_1 \cdot Z_{t-1}^2 + \epsilon_t \text{ (ARCH(1))} \\ X_t = \phi_0 + \alpha \cdot X_{t-1} + \epsilon_t \text{ (AR(1))} \\ X_t = \phi_0 + \theta \cdot \epsilon_{t-1} + \epsilon_t \text{ (MA(1))} \\ X_t = \phi_0 + \alpha \cdot X_{t-1} + \theta \cdot \epsilon_{t-1} + \epsilon_t \text{ (ARMA(1, 1))} \end{gathered} \tag{4}$$

where $\epsilon_t \sim N\left(\mu_\epsilon(t), \sigma_\epsilon^2\right)$.

The values for the parameters ϕ_1, α_0, α_1, α, θ, q and ω ranged from 0.1 to 0.9 for each parameter. Average absolute bias, average interval length (AIL) and average 95% credible interval coverage are shown in Table 1. To summarize model robustness, these values are averaged over all combinations of parameters, considering their prior distribution is a Dirac's delta with all probability concentrated in the corresponding parameter value.

For each autocorrelation structure and parameters combination, a random sample of size $n = 1000$ has been generated using the R function *arima.sim*, and the parameters $m = log(M^*)$ and β have been fixed to 0.2 and 0.4 respectively. Several values for these parameters were considered but no substantial differences in the model performance were observed related to the value of these parameters or sample size, besides a poorer coverage for lower sample sizes, as could be expected.

Real incidence of Covid-19 in Spain

The betacoronavirus SARS-CoV-2 has been identified as the causative agent of an unprecedented world-wide outbreak of pneumonia starting in December 2019 in the city of Wuhan (China) [1], named as Covid-19. Considering that many cases run without developing symptoms or just with very mild symptoms, it is reasonable to assume that the incidence of this disease has been underregistered. This work focuses on the weekly Covid-19 incidence registered in Spain in the period (2020/02/23–2022/02/27).

Table 1 Model performance measures (average absolute bias, average interval length and average coverage) summary based on a simulation study

Structure	Parameter	Bias	AIL	Coverage (%)
ARCH(1)	$\widehat{\phi_0}$	-0.377	3.586	68.77%
	$\widehat{\phi_1}$	0.122	0.525	66.08%
	$\widehat{\alpha_0}$	-0.296	1.351	74.72%
	$\widehat{\alpha_1}$	-0.085	0.920	77.34%
	$\widehat{\omega}$	-0.020	0.234	83.71%
	$\widehat{q}$	-0.022	0.167	85.06%
	$\widehat{m}$	-0.226	0.783	79.17%
	$\widehat{\beta}$	-0.734	3.581	76.83%
	$\widehat{\sigma_\epsilon}$	-1.540	3.323	63.65%
AR(1)	$\widehat{\phi_0}$	-0.983	5.189	70.10%
	$\widehat{\alpha}$	0.043	0.814	92.46%
	$\widehat{\omega}$	-0.003	0.111	94.10%
	$\widehat{q}$	-0.001	0.014	89.03%
	$\widehat{m}$	0.001	0.190	75.17%
	$\widehat{\beta}$	0.007	0.192	74.49%
	$\widehat{\sigma_\epsilon}$	-1.689	4.718	81.07%
MA(1)	$\widehat{\phi_0}$	-1.241	5.171	68.31%
	$\widehat{\theta}$	0.051	0.818	90.40%
	$\widehat{\omega}$	-0.005	0.108	95.06%
	$\widehat{q}$	-0.001	0.014	87.24%
	$\widehat{m}$	-0.002	0.187	76.95%
	$\widehat{\beta}$	0.004	0.190	80.38%
	$\widehat{\sigma_\epsilon}$	-1.619	4.679	83.95%
ARMA(1,1)	$\widehat{\phi_0}$	-1.834	5.107	61.01%
	$\widehat{\alpha}$	0.062	0.799	89.39%
	$\widehat{\theta}$	0.011	0.873	96.86%
	$\widehat{\omega}$	-0.001	0.014	88.32%
	$\widehat{q}$	-0.005	0.109	94.97%
	$\widehat{m}$	0.002	0.184	78.49%
	$\widehat{\beta}$	0.004	0.183	78.01%
	$\widehat{\sigma_\epsilon}$	-1.828	4.631	74.74%

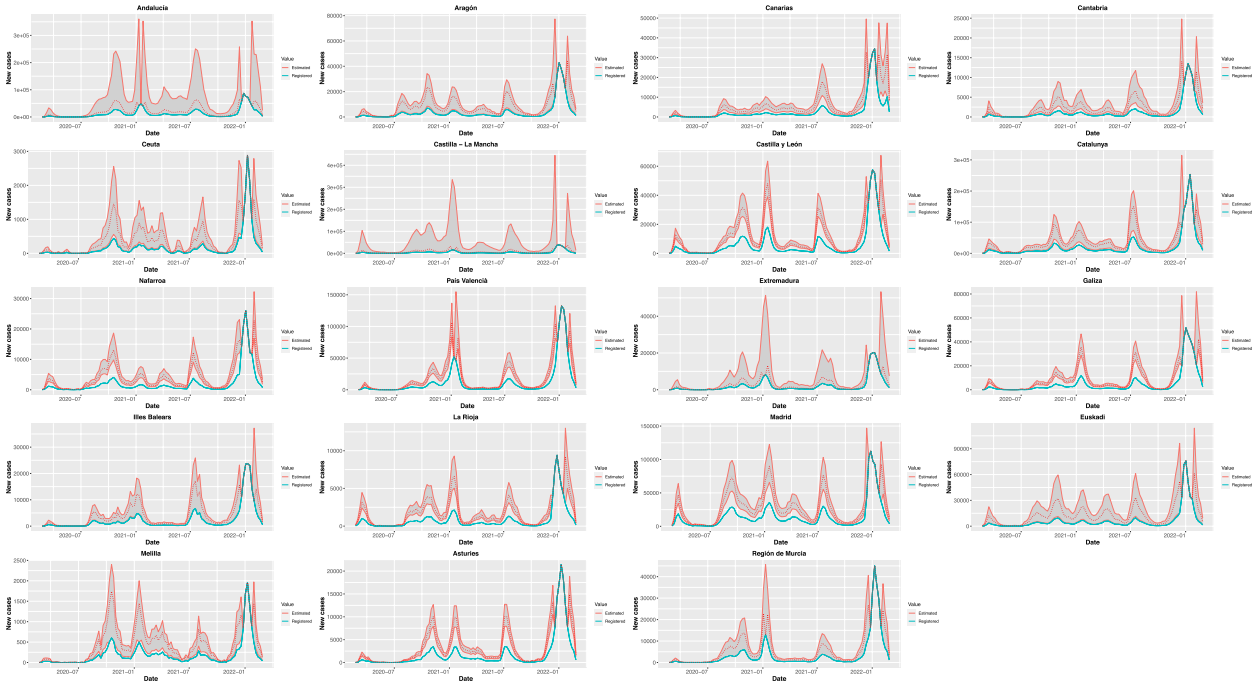

Fig. 1 Registered and predicted weekly new Covid-19 cases in each Spanish region

It can be seen in Fig. 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 predicts a total of 21,639,627 cases (only 51.10% of actual cases were reported). This work also revealed that while the frequency of underreporting is extremely high for all regions (values of $\widehat{\omega}$ over 0.80 in all cases), the intensity of this underreporting is not uniform across the considered regions: Aragón and Ceuta are the CCAAs with highest underreporting intensity ($\widehat{q} = 0.28$) while Región de Murcia and País Valencià are the regions where the predicted values are closest to the number of reported cases ($\widehat{q} = 0.46$). Detailed underreporting parameters estimates for each region can be found in Table 2. 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 Euskadi, as can also be seen in Table 2 through the estimates corresponding to parameter β_2. Figure 2 represents the predicted and registered Covid-19 weekly incidence globally for Spain.

Figure 2 shows the evolution of the registered (turquoise) and predicted (red) weekly number of Covid-19 cases in Spain in the period 2020/02/23–2022/02/27.

As can be seen in Figs. 1 and 2, there are 4 weeks (2021–12-26, 2022–01-02, 2022–01-09 and 2022–01-16) for which the predicted values coincide with those registered in all simulations, so no underreporting is detected these weeks. This behavior might be due to the breakout of a new variant with different characteristics (for instance producing more symptomatic cases and therefore reducing the underreporting) around these dates.

The registered values predicted by the model can also be obtained as $\widehat{Y_t} = \left(1 - \widehat{\omega} + \widehat{\omega} \cdot \widehat{q}\right) \cdot \widehat{X_t}$, and compared to the actual registered values Y_t. That allows computing standard forecasting error measures as Root Mean Squared Error (RMSE) or Mean Absolute Percentage Error (MAPE). Globally, the RMSE was 113,145.4 and MAPE was around 8%, ranging between 4 to 13% across regions. The specific RMSE and MAPE for each region are described in Table S1 in the Supplementary Material.

The global differences in underreporting magnitudes across regions can be represented by the percentage of reported cases in each CCAA (compared to model estimates), as shown in Fig. 3.

Table 2 Estimated underreporting parameters and impact of the considered covariates (β_1 and β_2 are the coefficients for confinement and vaccination respectively) for each Spanish CCAA. Reported values correspond to the median and percentiles 25% and 75% of the corresponding posterior distribution

CCAA	Parameter	Estimate (P25-P75)
Andalucía	ω	0.97 (0.95—0.99)
	q	0.44 (0.41—0.48)
	β_1	-1.67 (-2.31, -0.39)
	β_2	-1.71 (-2.66, -0.68)
Aragón	ω	0.98 (0.97—0.99)
	q	0.28 (0.27—0.32)
	β_1	0.76 (0.17, 1.43)
	β_2	-1.06 (-1.36, -0.69)
Asturies	ω	0.97 (0.90—0.99)
	q	0.40 (0.37—0.53)
	β_1	0.44 (0.11, 0.69)
	β_2	-0.90 (-1.77, -0.63)
Cantabria	ω	0.97 (0.95—0.99)
	q	0.30 (0.28—0.35)
	β_1	-0.44 (-0.71, 0.002)
	β_2	-0.53 (-1.29, -0.25)
Castilla y León	ω	0.98 (0.95—0.99)
	q	0.36 (0.32—0.41)
	β_1	-0.84 (-1.33, -0.23)
	β_2	-1.22 (-1.88, -0.60)
Castilla – La Mancha	ω	0.98 (0.96—0.99)
	q	0.33 (0.31—0.36)
	β_1	0.06 (-0.18, 0.44)
	β_2	-0.80 (-1.11, -0.40)
Canarias	ω	0.98 (0.96—0.99)
	q	0.35 (0.32—0.38)
	β_1	-0.92 (-2.06, -0.29)
	β_2	-1.34 (-1.78, -1.06)
Catalunya	ω	0.98 (0.96—0.99)
	q	0.30 (0.27—0.34)
	β_1	-0.25 (-0.52, 0.21)
	β_2	-1.51 (-1.97, -0.94)
Ceuta	ω	0.98 (0.95—0.99)
	q	0.28 (0.25—0.31)
	β_1	0.007 (-0.52, 0.34)
	β_2	-1.38 (-1.93, -0.84)
Extremadura	ω	0.98 (0.95—1.00)
	q	0.40 (0.36—0.44)
	β_1	1.45 (1.24, 1.83)
	β_2	-0.72 (-1.30, -0.37)
Galiza	ω	0.84 (0.33—0.98)
	q	0.41 (0.35—0.56)
	β_1	-0.20 (-0.53, 0.18)
	β_2	-2.03 (-3.07, -1.34)
Illes Balears	ω	0.98 (0.96—0.99)
	q	0.36 (0.33—0.39)
	β_1	0.74 (0.43, 1.01)
	β_2	-0.72 (-1.16, -0.34)
Región de Murcia	ω	0.93 (0.45—0.98)
	q	0.46 (0.34—0.80)
	β_1	0.62 (-0.02, 1.36)
	β_2	-1.97 (-3.07, -0.59)
Madrid	ω	0.98 (0.96—0.99)
	q	0.37 (0.34—0.40)
	β_1	0.35 (-0.39, 0.59)
	β_2	-0.35 (-0.77, -0.07)
Nafarroa	ω	0.99 (0.97—1.00)
	q	0.30 (0.26—0.32)
	β_1	-1.71 (-1.92, -0.53)
	β_2	-2.05 (-3.20, -1.33)
Euskadi	ω	0.99 (0.97—0.99)
	q	0.27 (0.25—0.31)
	β_1	-0.42 (-0.69, -0.21)
	β_2	-0.10 (-0.24, 0.00)
La rioja	ω	0.98 (0.96—0.99)
	q	0.31 (0.28—0.35)
	β_1	-0.83 (-1.08, -0.35)
	β_2	-0.43 (-0.71, -0.22)
Melilla	ω	0.97 (0.95—0.99)
	q	0.34 (0.31—0.37)
	β_1	-0.48 (-0.82, -0.11)
	β_2	-1.59 (-2.05, -0.93)
País Valencià	ω	0.95 (0.40—0.98)
	q	0.46 (0.40—0.67)
	β_1	1.45 (1.24, 1.83)
	β_2	-1.70 (-2.64, -0.52)

Discussion

Although it is very common in biomedical and epidemiological research to get data from disease registries, there is a concern about their reliability, and there have been some recent efforts to standardize the protocols in order to improve the accuracy of health information registries (see for instance [24, 25]). However, as the Covid-19 pandemic situation has made evident, it is not always possible to implement these recommendations in a proper way.

Another work analyzing the cumulated burden of Covid-19 in Spain [26] estimated that only around 21% of the cases were reported in the period 2020/01/01–2020/06/01, but it should be taken into account that it

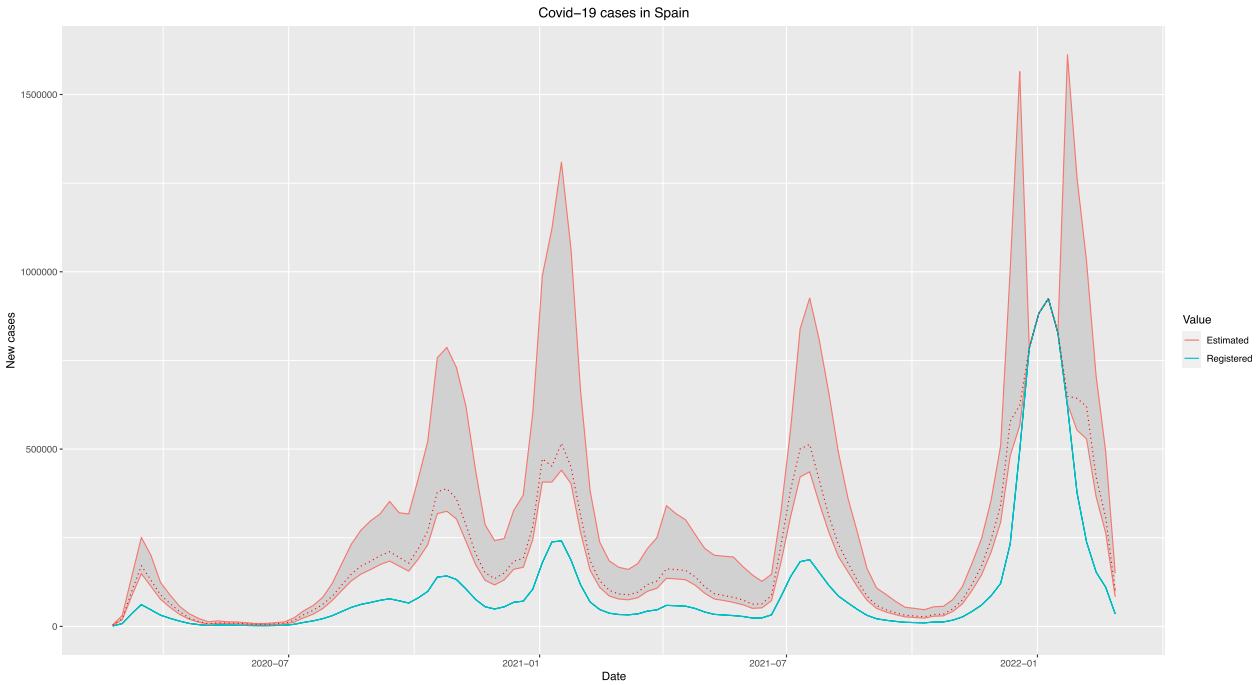

Fig. 2 Registered and predicted weekly new Covid-19 cases globally in Spain

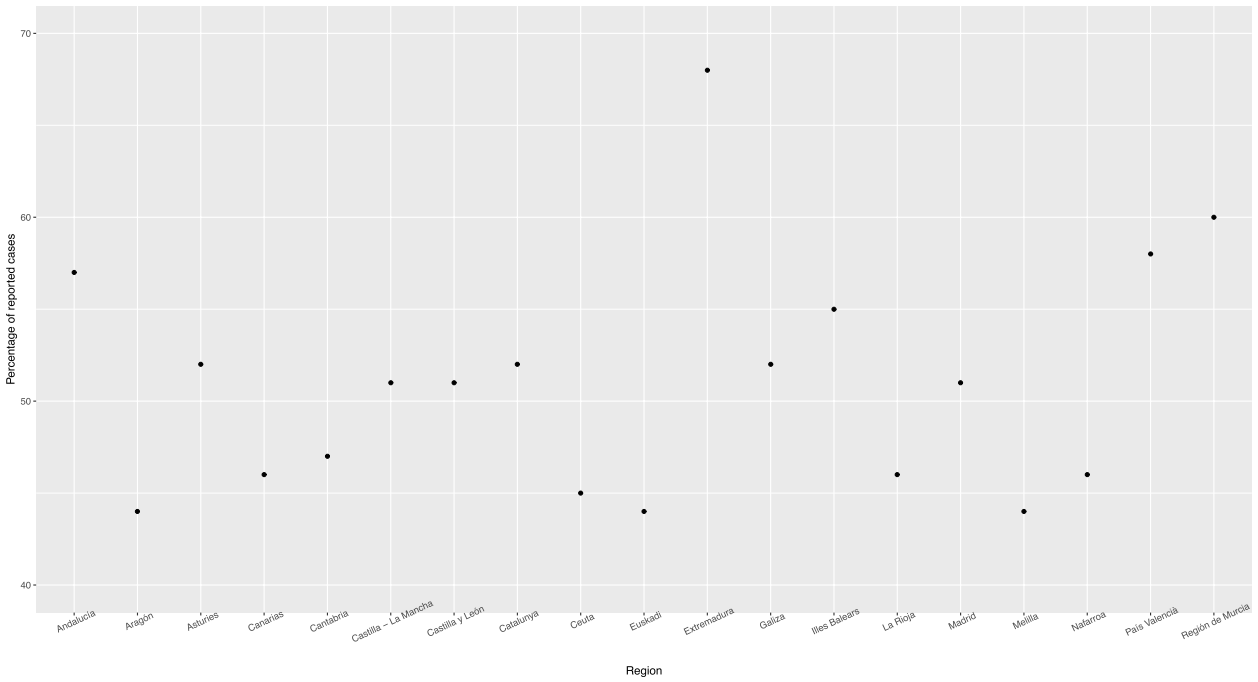

Fig. 3 Percentage of reported cases in each CCAA

seems reasonable to assume that the underreporting intensity was higher at the early stages of the pandemic, and therefore a lower overall underreporting is expected in the longer period considered in this work. Additionally, the presented methodology allows for a real time monitoring and not only cumulated over a time period.

Having accurate data is key in order to provide public health decision-makers with reliable information, which can also be used to improve the accuracy of dynamic models aimed to estimate the spread of the disease [27] and to predict its behavior.

Conclusions

The proposed methodology can deal with misreported (over- or under-reported) data in a very natural and straightforward way and is able to reconstruct the most likely hidden process, providing public health decision-makers with a valuable tool in order to predict the evolution of the disease under different scenarios.

Using a flexible approach for the underlying hidden process, such as ARCH time series, are a natural extension to recent developments (see for instance [19]) proposed for fitting underreported time series but restricted to the case when the underlying process has an ARMA structure and allow us to model phenomena presenting more complex behavior like Covid-19 in the long time period considered in the present work.

The analysis of the Spanish Covid-19 data shows that in average only around 51% of the cases in the period 2020/02/23–2022/02/27 were reported, and that there are important differences in the severity of underreporting across the Spanish regions. The impact of the vaccination program can also be assessed, achieving a significant decrease in the Covid-19 incidence in almost all regions after 50% of the population had one dose at least (although these results would probably be notably different if including SARS-CoV-2 immunity-escape variants like BA.4 or BA.5, which are currently predominant in many countries), while the impact of the mandatory lockdown could only be detected by the model in 7 out of 19 regions.

The simulation study shows that the proposed methodology behaves as expected and that the parameters used in the simulations, under different autocorrelation structures, can be recovered, even with severely underreported data.

Abbreviations

ABC	Approximate Bayesian computation
AIL	Average interval length
AR	AutoRegressive model
ARCH	AutoRegressive Conditional Heteroskedasticity model
ARMA	AutoRegressive Moving Average model
BSL	Bayesian Synthetic Likelihood
CCAA	Spanish autonomous community
CrI	Credible interval
EM	Expectation–Maximization
MA	Moving average model
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Squared Error
SIR	Susceptible-Infected-Recovered model

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-01894-9.

Additional file 1: Table S1. Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) for the predicted number of cases in each Spanish CCAA.

Acknowledgements

Not applicable.

Authors' contributions

DM, AF-F, AC, AA and PP participated in the development of the statistical model. DM and PP derived the described properties, and DM implemented the model in R software and conducted the analyses. All authors have read and approved the manuscript.

Funding

Research funded by Fundación MAPFRE. This work was partially supported by grant RTI2018-096072-B-I00 from the Spanish Ministry of Science and Innovation and by the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020–001084-M). A.F-F acknowledges Agencia Estatal de Investigación for the financial support IJC2020-045188I/AEI/10.13039/501100011033. AC was partially financed by PID2021-123733NB-I00 (Ministerio de Ciencia e Innovación, Spain). AC and AA were partially supported by Project "EcoDep" CY-AAP2020-0000000013 ("Investissements d'Avenir" ANR-16-IDEX-0008, France).

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the GitHub repository, https://github.com/dmorinya/Mapfre/blob/main/Papers/Paper%202%20(BSL)/BMC%20MRM/GitHub/Data/cases.xls.

Declarations

Ethics approval and consent to participate

Not applicable (no human or animal experimentation and data and source codes fully available).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 3 October 2022 Accepted: 17 March 2023
Published online: 28 March 2023

References

1. Sohrabi Catrin, Alsafi Zaid, O'Neill Niamh, Khan Mehdi, Kerwan Ahmed, Al-Jabir Ahmed, Iosifidis Christos, Agha Riaz. World Health Organization declares global emergency: a review of the 2019 Novel Coronavirus (COVID-19). Int J Surg. 2020. https://doi.org/10.1016/j.ijsu.2020.02.034.
2. Bernard H, Werber D, Höhle M. Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104: H4 in 2011 - a time series analysis. BMC Infect Dis. 2014;14(1):116. https://doi.org/10.1186/1471-2334-14-116.
3. Arendt S, Rajagopal L, Strohbehn C, Stokes N, Meyer J, Mandernach S. Reporting of foodborne illness by U.S. consumers and healthcare professionals. Int J Environ Res Public Health. 2013;10(8):3684–714. https://doi.org/10.3390/ijerph10083684.
4. Rosenman KD, Kalush A, Reilly MJ, Gardiner JC, Reeves M, Luo Z. How much work-related injury and illness is missed by the current national surveillance

system? J Occup Environ Med. 2006;48(4):357–65. https://doi.org/10.1097/01.jom.0000205864.81970.63.
5. Alfonso JH, Løvseth EK, Samant Y, Holm JØ. Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. Contact Dermatitis. 2015;72(6):409–12. https://doi.org/10.1111/cod.12355.
6. Winkelmann R. Markov Chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. Empir Econom. 1996;21(4):575–87. https://doi.org/10.1007/BF01180702.
7. Gibbons CL, Mangen M-J, Plass D, Havelaar AH, Brooke RJ, Kramarz P, Peterson KL, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health. 2014;14(1):147. https://doi.org/10.1186/1471-2458-14-147.
8. Stocks T, Britton T, Höhle M. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. Biostatistics. 2018. https://doi.org/10.1093/biostatistics/kxy057.
9. Azmon A, Faes C, Hens N. On the estimation of the reproduction number based on misreported epidemic data. Stat Med. 2014;33(7):1176–92. https://doi.org/10.1002/sim.6015.
10. Magal P, Webb G. The parameter identification problem for SIR epidemic models: identifying unreported cases. J Math Biol. 2018;77(6–7):1629–48. https://doi.org/10.1007/s00285-017-1203-9.
11. Stoner O, Economou T, Drummond Marques da Silva G. A hierarchical framework for correcting under-reporting in count data. J Am Stat Assoc. 2019. https://doi.org/10.1080/01621459.2019.1573732.
12. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019. (https://www.r-project.org/).
13. Bracher J. hhh4underreporting: fitting endemic-epidemic models to underreported data. 2021. https://rdrr.io/github/jbracher/hhh4underreporting/man/hhh4u.html.
14. Oran DP, Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection. Ann Intern Med. 2020. https://doi.org/10.7326/m20-3012.
15. Fernández-Fontelo A, Cabaña A, Puig P, Moriña D. Under-reported data analysis with INAR-hidden Markov chains. Stat Med. 2016;35(26):4875–90. https://doi.org/10.1002/sim.7026.
16. Fernández-Fontelo A, Cabaña A, Joe H, Puig P, Moriña D. Untangling serially dependent underreported count data for gender-based violence. Stat Med. 2019;38(22):4404–22. https://doi.org/10.1002/sim.8306.
17. Fernández-Fontelo Amanda, Moriña David, Cabaña Alejandra, Arratia Argimiro, Puig Pere. Estimating the real burden of disease under a pandemic situation: the SARS-CoV2 case. PLoS One. 2020;15(12 December):e0242956. https://doi.org/10.1371/journal.pone.0242956.
18. Moriña D, Fernández-Fontelo A, Cabaña A, Puig P, Monfil L, Brotons M, Diaz M. Quantifying the under-reporting of uncorrelated longitudal data: the genital warts example. BMC Med Res Methodol. 2021;21(1):6. https://doi.org/10.1186/s12874-020-01188-4.
19. Moriña D, Fernández-Fontelo A, Cabaña A, Puig P. New statistical model for misreported data with application to current public health challenges. Sci Rep. 2021;11(1):23321. https://doi.org/10.1038/s41598-021-02620-5.
20. Wood SN. Statistical inference for noisy nonlinear ecological dynamic systems. Nature. 2010;466(7310):1102–4. https://doi.org/10.1038/nature09319.
21. Price LF, Drovandi CC, Lee A, Nott DJ. Bayesian synthetic likelihood. J Comput Graph Stat. 2018;27(1):1–11. https://doi.org/10.1080/10618600.2017.1302882.
22. An, Ziwen, Leah F. South, and Christopher C. Drovandi. 2019b. BSL: Bayesian synthetic likelihood. https://CRAN.R-project.org/package=BSL.
23. An, Ziwen, Leah F South, and Christopher Drovandi. 2019a. BSL: an r package for efficient parameter estimation for simulation-based models via bayesian synthetic likelihood. arXiv. http://arxiv.org/abs/1907.10940v1.
24. Kodra Y, Weinbach J, Posada-De-La-Paz M, AlessioCoi S, Lemonnier L, van Enckevort D, Roos M, et al. Recommendations for improving the quality of rare disease registries. MDPI AG. 2018. https://doi.org/10.3390/ijerph15081644.
25. Harkener S, Stausberg J, Hagel C, Siddiqui R. Towards a core set of indicators for data quality of registries. Stud Health Technol Inform. 2019;267:39–45. https://doi.org/10.3233/SHTI190803.
26. Moriña D, Fernández-Fontelo A, Cabaña A, Arratia A, Ávalos G, Puig P. Cumulated burden of COVID-19 in Spain from a Bayesian perspective. Eur J Pub Health. 2021;31(4):917–20. https://doi.org/10.1093/eurpub/ckab118.
27. Zhao M, Lin R, Yang W, Lou, 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. https://doi.org/10.3390/jcm9020388.

Publisher's Note