arXiv:2104.07575v2 [stat.ME] 17 Aug 2021

Bayesian Synthetic Likelihood Estimation for Underreported Non-Stationary Time Series: Covid-19 Incidence in Spain

David Moriña[1,2], Amanda Fernández-Fontelo[3], Alejandra Cabaña[4], Argimiro Arratia[5], and Pedro Puig[2,4]

[1]Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB), Barcelona, Spain; dmorina@ub.edu
[2]Centre de Recerca Matemàtica (CRM), Cerdanyola del Vallès, Spain
[3]Chair of Statistics, Humboldt-Universität zu Berlin, Berlin, Germany
[4]Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
[5]Department of Computer Science, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Abstract

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, 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. The performance of the proposed methodology is evaluated through a comprehensive simulation study and illustrated by reconstructing the weekly Covid-19 incidence in each Spanish Autonomous Community in 2020, and it reveals that less than 60% of the actual cases in the period 2020/02/19 to 2020/12/15 were registered.

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. There is an enormous global concern around this disease, leading the World Health Organization (WHO) to declare public health emergency [21]. Many methodological efforts have been made to deal with misreported Covid-19 data, following ideas introduced in the literature since the late nineties [6, 4, 20, 1, 24, 10]. These proposals range from the usage of multiplication factors [22] to Markov-based models [5, 14] or spatio-temporal models [23]. Additionally, a new R [19] package able to fitting endemic-epidemic models based on approximative maximum likelihood to underreported count data has been recently published [12]. However, as a large proportion of the cases run asymptomatically [17] 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 [7, 9, 8] although there is a lack of continuous time 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 approach for longitudinal data not accounting for temporal correlations is introduced in [16] and a model capable of dealing with temporal structures using a different approach is presented in [15]. 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 [25] and placed into a Bayesian framework in [18], 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.

2 Methods

Consider an unobservable process X_t following an AutoRegressive-Moving Average ($ARMA(p, r)$) structure, defined by

$$X_t = \phi_0 + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \theta_r \epsilon_{t-r} + \epsilon_t, \quad (1)$$

where ϵ_t is a Gaussian white noise process with $\epsilon_t \sim N(\mu_\epsilon, \sigma^2_\epsilon)$.

In our setting, this process X_t cannot be directly observed, and all we can see is a part of it, expressed as

$$Y_t = \begin{cases} X_t \text{ with probability } 1 - \omega \\ q \cdot X_t \text{ with probability } \omega \end{cases} \quad (2)$$

The interpretation of the parameters in Eq. (2) is straightforward: 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). The parameter 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 in Eq. (1) 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 in [8], the number of affected individuals at time t, $A(t) = I(t) + R(t)$ can be approximated by

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

where $k = \beta - \gamma$ and $M^* = \frac{N(\beta-\gamma)}{\beta-\frac{\gamma}{2}}$, β is the infection rate, γ the recovery rate and N the size of the susceptible population. At any time t the condition $S(t) + I(t) + R(t) = N$ is fulfilled. The expression 3 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 this 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* [3, 2] package for R [19]. 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.

The data and source code underlying this article are available in GitHub, at `https://github.com/dmorinya/BSLCovidSpain`.

3 Results

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.

3.1 Simulation study

A thorough simulation study has been conducted to ensure that the model behaves as expected, including $AR(1)$, $MA(1)$ and $ARMA(1,1)$ structures for the hidden process X_t defined as

$$\begin{aligned} X_t &= \alpha \cdot X_{t-1} + \epsilon_t \text{ (AR(1))} \\ X_t &= \theta \cdot \epsilon_{t-1} + \epsilon_t \text{ (MA(1))} \\ X_t &= \alpha \cdot X_{t-1} + \theta \cdot \epsilon_{t-1} + \epsilon_t \text{ (ARMA(1, 1))} \end{aligned} \tag{4}$$

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

The values for the parameters 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% credibility interval coverage are shown in Table 1. To summarise 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 5 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 expected.

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 (%)**
$AR(1)$	$\hat{\phi_0}$	-0.983	5.189	70.10%
	$\hat{\alpha}$	0.043	0.814	92.46%
	$\hat{\omega}$	-0.003	0.111	94.10%
	$\hat{q}$	-0.001	0.014	89.03%
	$\hat{m}$	0.001	0.190	75.17%
	$\hat{\beta}$	0.007	0.192	74.49%
	$\hat{\sigma_\epsilon}$	-1.689	4.718	81.07%
$MA(1)$	$\hat{\phi_0}$	-1.241	5.171	68.31%
	$\hat{\theta}$	0.051	0.818	90.40%
	$\hat{\omega}$	-0.005	0.108	95.06%
	$\hat{q}$	-0.001	0.014	87.24%
	$\hat{m}$	-0.002	0.187	76.95%
	$\hat{\beta}$	0.004	0.190	80.38%
	$\hat{\sigma_\epsilon}$	-1.619	4.679	83.95%
$ARMA(1,1)$	$\hat{\phi_0}$	-1.834	5.107	61.01%
	$\hat{\alpha}$	0.062	0.799	89.39%
	$\hat{\theta}$	0.011	0.873	96.86%
	$\hat{\omega}$	-0.001	0.014	88.32%
	$\hat{q}$	-0.005	0.109	94.97%
	$\hat{m}$	0.002	0.184	78.49%
	$\hat{\beta}$	0.004	0.183	78.01%
	$\hat{\sigma_\epsilon}$	-1.828	4.631	74.74%

3.2 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) [21], 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 (2020/02/19-2020/12/15) excluding the two autonomous cities Ceuta and Melilla, with very low incidences during all considered time period. 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.

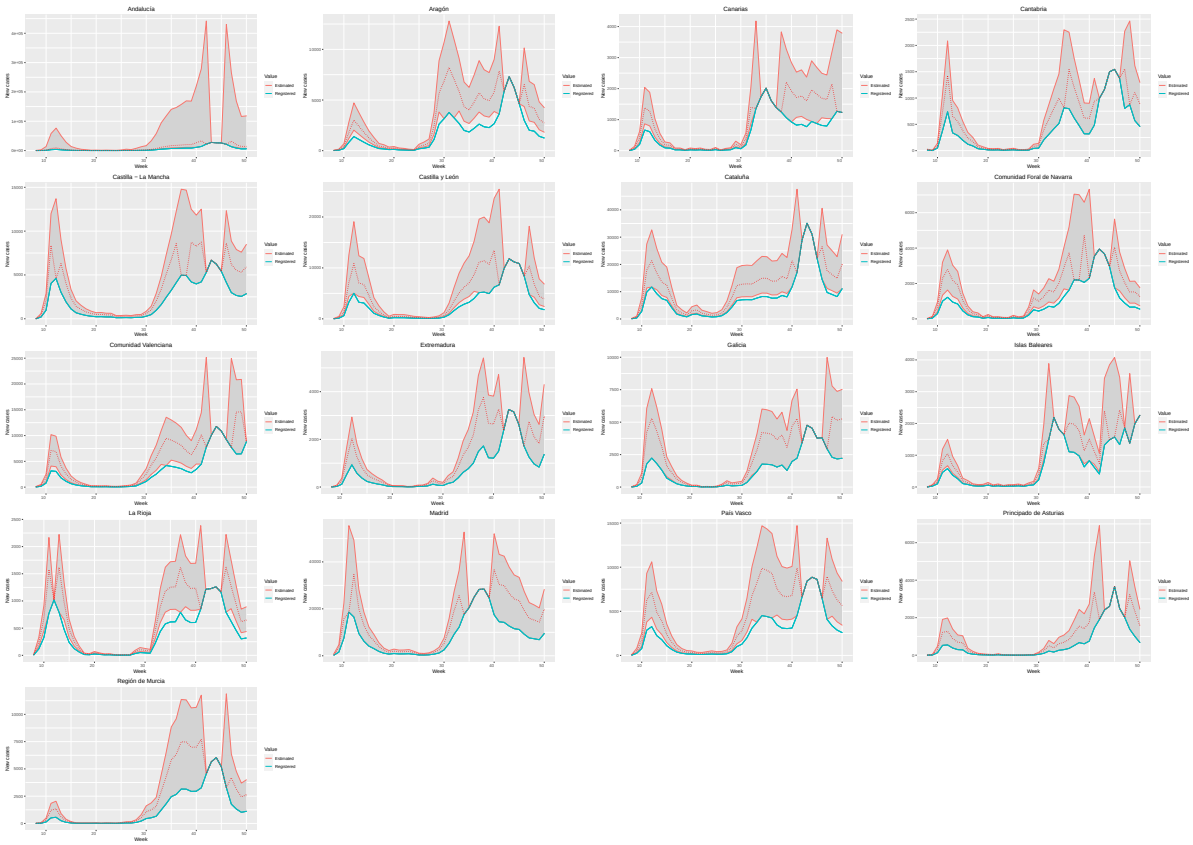

Figure 1: Registered and estimated weekly new Covid-19 cases in each Spanish region.

In the considered period, the official sources reported 1,819,982 Covid-19 cases in Spain (excluding Ceuta and Melilla), while the model estimates a total of 3,078,858 cases (only 59.11% of actual cases were reported). These work also shows that while the frequency of underreporting is ex-

tremely high for all regions (values close to 1) with the exception of Andalucía, the intensity of this underreporting is not uniform across the considered regions, as shown in Table 2. It can be seen that Andalucía, Galicia and Región de Murcia are the CCAA with highest underreporting intensity ($\hat{q} = 0.42$) while Islas Baleares and Catalunya are the regions where the estimated values are closest to the number of reported cases ($\hat{q} = 0.55$).

Figure 2 shows the evolution of the registered (turquoise) and estimated (red) weekly number of Covid-19 cases in Spain in the period 2020/02/19-2020/12/15, excluding the autonomous cities of Ceuta and Melilla.

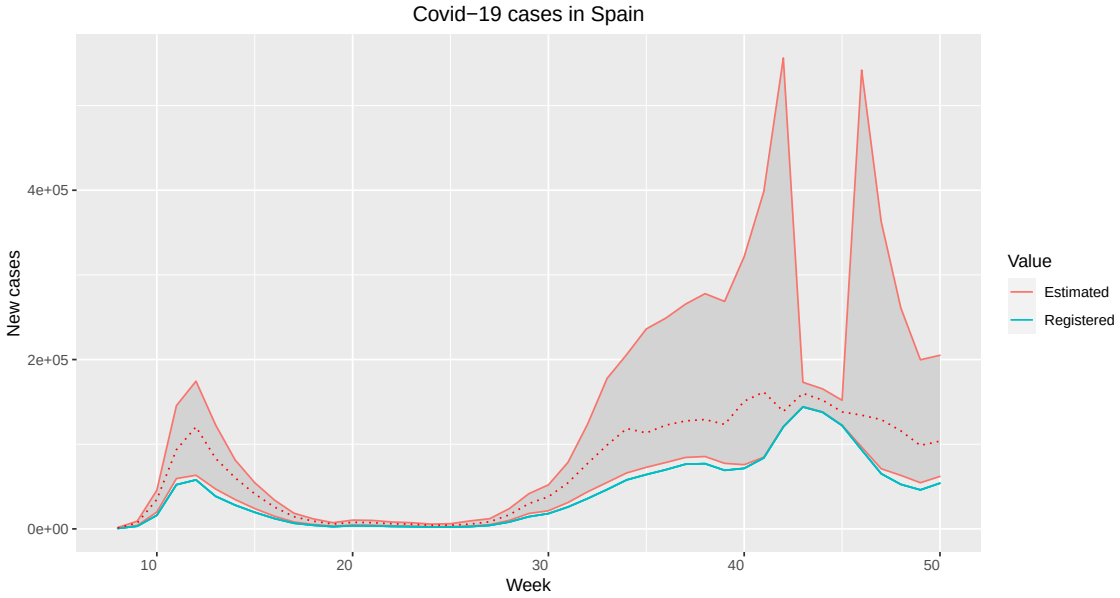

Figure 2: Registered and estimated weekly new Covid-19 cases in Spain.

4 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 [13, 11]). However, as the Covid-19 pandemic situation has made evident, it is not always possible to implement these recommendations in a proper way.

The analysis of the Spanish Covid-19 data shows that in average less than 60% of the cases in the period 2020/02/19-2020/12/15 were reported. Having accurate data is key in order to provide public health decision-

Table 2: Estimated underreported frequency and intensity for each Spanish CCAA. Reported values correspond to the median and percentiles 2.5% and 97.5% of the corresponding posterior distribution.

CCAA	Parameter	Estimate (95% CI)
Andalucía	$\hat{\omega}$	0.62 (0.36, 0.97)
	$\hat{q}$	0.42 (0.33, 0.50)
Aragón	$\hat{\omega}$	0.97 (0.92, 0.99)
	$\hat{q}$	0.46 (0.40, 0.52)
Principado de Asturias	$\hat{\omega}$	0.98 (0.95, 0.99)
	$\hat{q}$	0.43 (0.38, 0.49)
Cantabria	$\hat{\omega}$	0.96 (0.92, 0.99)
	$\hat{q}$	0.52 (0.46, 0.62)
Castilla y León	$\hat{\omega}$	0.98 (0.95, 0.99)
	$\hat{q}$	0.46 (0.40, 0.51)
Castilla - La Mancha	$\hat{\omega}$	0.97 (0.93, 0.99)
	$\hat{q}$	0.48 (0.42, 0.59)
Canarias	$\hat{\omega}$	0.97 (0.92, 0.99)
	$\hat{q}$	0.48 (0.43, 0.55)
Cataluña	$\hat{\omega}$	0.97 (0.92, 0.99)
	$\hat{q}$	0.55 (0.49, 0.62)
Extremadura	$\hat{\omega}$	0.97 (0.93, 0.99)
	$\hat{q}$	0.46 (0.41, 0.53)
Galicia	$\hat{\omega}$	0.98 (0.95, 0.99)
	$\hat{q}$	0.42 (0.37, 0.48)
Islas Baleares	$\hat{\omega}$	0.96 (0.91, 0.99)
	$\hat{q}$	0.55 (0.49, 0.63)
Región de Murcia	$\hat{\omega}$	0.98 (0.94, 0.99)
	$\hat{q}$	0.42 (0.37, 0.48)
Madrid	$\hat{\omega}$	0.98 (0.94, 0.99)
	$\hat{q}$	0.48 (0.41, 0.55)
Comunidad Foral de Navarra	$\hat{\omega}$	0.98 (0.95, 0.99)
	$\hat{q}$	0.44 (0.40, 0.49)
Pais Vasco	$\hat{\omega}$	0.97 (0.90, 0.99)
	$\hat{q}$	0.46 (0.40, 0.55)
La Rioja	$\hat{\omega}$	0.96 (0.92, 0.99)
	$\hat{q}$	0.49 (0.44, 0.55)
Comunidad Valenciana	$\hat{\omega}$	0.97 (0.94, 0.99)
	$\hat{q}$	0.44 (0.39, 0.50)

makers with reliable information, which can also be used to improve the accuracy of dynamic models aimed to estimate the spread of the disease [26] and to predict its behavior. 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.

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.

Acknowledgements

This work was supported by grant COV20/00115 from Instituto de Salud Carlos III (Spanish Ministry of Health). This work was partially supported by grant RTI2018-096072-B-I00 from the Spanish Ministry of Science and Innovation.

References

[1] Jose H. Alfonso, Eva K. Løvseth, Yogindra Samant, and Jan-Ø. Holm. Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis*, 72(6):409–412, jun 2015.

[2] Ziwen An, Leah F South, and Christopher Drovandi. BSL: An R Package for Efficient Parameter Estimation for Simulation-Based Models via Bayesian Synthetic Likelihood. *arXiv*, 2019.

[3] Ziwen An, Leah F. South, and Christopher C. Drovandi. *BSL: Bayesian Synthetic Likelihood*, 2019. R package version 3.0.0.

[4] Susan Arendt, Lakshman Rajagopal, Catherine Strohbehn, Nathan Stokes, Janell Meyer, and Steven Mandernach. Reporting of foodborne illness by U.S. consumers and healthcare professionals. *International journal of environmental research and public health*, 10(8):3684–714, aug 2013.

[5] Amin Azmon, Christel Faes, and Niel Hens. On the estimation of the reproduction number based on misreported epidemic data. *Statistics in medicine*, 33(7):1176–92, mar 2014.

[6] Helen Bernard, Dirk Werber, and Michael Höhle. 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 Infectious Diseases*, 14(1), mar 2014.

[7] Amanda Fernández-Fontelo, Alejandra Cabaña, Pedro Puig, and David Moriña. Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, 35(26):4875–4890, nov 2016.

[8] Amanda Fernández-Fontelo, David Moriña, Alejandra Cabaña, Argimiro Arratia, and Pere Puig. Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE*, 15(12 December):e0242956, dec 2020.

[9] Amanda Fernández-Fontelo, Alejandra Cabaña, Harry Joe, Pedro Puig, and David Moriña. Untangling serially dependent underreported count data for gender-based violence. *Statistics in Medicine*, 38(22):4404–4422, sep 2019.

[10] Cheryl L Gibbons, Marie-Josée J Mangen, Dietrich Plass, Arie H Havelaar, Russell John Brooke, Piotr Kramarz, Karen L Peterson, Anke L Stuurman, Alessandro Cassini, Eric M Fèvre, Mirjam E E Kretzschmar, and Burden of Communicable diseases in Europe (BCoDE) consortium. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC public health*, 14(1):147, feb 2014.

[11] Sonja Harkener, Jürgen Stausberg, Christiane Hagel, and Roman Siddiqui. Towards a Core Set of Indicators for Data Quality of Registries. *Studies in health technology and informatics*, 267:39–45, 2019.

[12] Johannes Bracher. *hhh4underreporting: Fitting endemic-epidemic models to underreported data*, 2021. R package version 0.0.0.0.9002.

[13] Yllka Kodra, Jérôme Weinbach, Manuel Posada-De-La-Paz, Alessio Coi, S Lydie Lemonnier, David van Enckevort, Marco Roos, Annika Jacobsen, Ronald Cornet, S Faisal Ahmed, Virginie Bros-Facer, Veronica Popa, Marieke van Meel, Daniel Renault, Rainald von Gizycki, Michele Santoro, Paul Landais, Paola Torreri, Claudio Carta, Deborah Mascalzoni, Sabina Gainotti, Estrella Lopez, Anna Ambrosini, Heimo Müller, Robert Reis, Fabrizio Bianchi, Yaffa R Rubinstein, Hanns Lochmüller, and Domenica Taruscio. Recommendations for improving the quality of rare disease registries, aug 2018.

[14] Pierre Magal and Glenn Webb. The parameter identification problem for SIR epidemic models: identifying unreported cases. *Journal of Mathematical Biology*, 77(6-7):1629–1648, dec 2018.

[15] David Moriña, Amanda Fernández-Fontelo, Alejandra Cabaña, and Pedro Puig. New statistical model for misreported data with application to current public health challenges. *Scientific Reports (Under review)*, 2021.

[16] David Moriña, Amanda Fernández-Fontelo, Alejandra Cabaña, Pedro Puig, Laura Monfil, Maria Brotons, and Mireia Diaz. Quantifying the under-reporting of uncorrelated longitudal data: the genital warts example. *BMC Medical Research Methodology*, 21(1):6, dec 2021.

[17] Daniel P. Oran and Eric J. Topol. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Annals of Internal Medicine*, jun 2020.

[18] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian Synthetic Likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, jan 2018.

[19] R Core Team. R: A Language and Environment for Statistical Computing, 2019.

[20] Kenneth D Rosenman, Alice Kalush, Mary Jo Reilly, Joseph C Gardiner, Mathew Reeves, and Zhewui Luo. How much work-related injury and illness is missed by the current national surveillance system?

Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine, 48(4):357–65, apr 2006.

[21] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World Health Organization declares Global Emergency: A review of the 2019 Novel Coronavirus (COVID-19). *International journal of surgery (London, England)*, feb 2020.

[22] Theresa Stocks, Tom Britton, and Michael Höhle. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics*, sep 2018.

[23] Oliver Stoner, Theo Economou, and Gabriela Drummond Marques da Silva. A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, pages 1–17, mar 2019.

[24] Rainer Winkelmann. Markov Chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, 21(4):575–587, 1996.

[25] Simon N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, aug 2010.

[26] Zhao, Musa, Lin, Ran, Yang, Wang, Lou, Yang, Gao, He, and Wang. 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. *Journal of Clinical Medicine*, 9(2):388, feb 2020.