ESTIMATED COVID-19 BURDEN IN SPAIN: ARCH UNDER-REPORTED NON-STATIONARY TIME SERIES

David Moriña, Amanda Fernández-Fontelo, Alejandra Cabaña, Argimiro Arratia and Pere Puig

Universitat de Barcelona
Facultat d'Economia i Empresa

Fundación MAPFRE

21 July 2022
International Workshop on Statistical Modelling 2022

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
- As a large proportion of the cases run asymptomatically 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

Previously proposed models (count data)

Independent under-reporting states

Research Article

Statistics in Medicine

Received 17 February 2016, Accepted 9 June 2016 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7026

Under-reported data analysis with INAR-hidden Markov chains

Amanda Fernández-Fontelo,[a*†] Alejandra Cabaña,[a] Pedro Puig[a] and David Moriña[b,c]

In this work, we deal with correlated under-reported data through INAR(1)-hidden Markov chain models. These models are very flexible and can be identified through its autocorrelation function, which has a very simple form. A naïve method of parameter estimation is proposed, jointly with the maximum likelihood method based on a revised version of the forward algorithm. The most-probable unobserved time series is reconstructed by means of the Viterbi algorithm. Several examples of application in the field of public health are discussed illustrating the utility of the models.

Keywords: discrete time series; emission probabilities; integer-autoregressive models; thinning operator; under-recorded data

Previously proposed models (count data)

Serially dependent under-reporting states

RESEARCH ARTICLE

WILEY Statistics in Medicine

Untangling serially dependent underreported count data for gender-based violence

Amanda Fernández-Fontelo[1,2] | Alejandra Cabaña[2] | Harry Joe[3] | Pedro Puig[1,4] | David Moriña[4]

[1]School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

[2]Departament de Matemàtiques, Universitat Autònoma de Barcelona, Barcelona, Spain

[3]Department of Statistics, University of British Columbia, Vancouver, Canada

[4]Barcelona Graduate School of Mathematics, Departament de Matemàtiques, Universitat Autònoma de Barcelona, Bellaterra, Spain

Correspondence
Amanda Fernández-Fontelo, School of Business and Economics, Humboldt-Universität zu Berlin, 10178 Berlin, Germany; or Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain.
Email: fernanda@hu-berlin.de

Underreporting in gender-based violence data is a worldwide problem leading to the underestimation of the magnitude of this social and public health concern. This problem deteriorates the data quality, providing poor and biased results that lead society to misunderstand the actual scope of this domestic violence issue. The present work proposes time series models for underreported counts based on a latent integer autoregressive of order 1 time series with Poisson distributed innovations and a latent underreporting binary state, that is, a first-order Markov chain. Relevant theoretical properties of the models are derived, and the moment-based and maximum-based methods are presented for parameter estimation. The new time series models are applied to the quarterly complaints of domestic violence against women recorded in some judicial districts of Galicia (Spain) between 2007 and 2017. The models allow quantifying the degree of underreporting. A comprehensive discussion is presented, studying how the frequency and intensity of underreporting in this public health concern are related to some interesting socioeconomic and health indicators of the provinces of Galicia (Spain).

Previously proposed models (count data)

Non-stationary processes

PLOS ONE

RESEARCH ARTICLE

Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case

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

1 Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany, **2** Departament de Matemàtiques, Barcelona Graduate School of Mathematics (BGSMath), Universitat Autònoma de Barcelona, Barcelona, Spain, **3** Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona, Barcelona, Spain, **4** Centre de Recerca Matemàtica (CRM), Barcelona, Spain, **5** Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, Spain

* fernanda@hu-berlin.de

Abstract

The present paper introduces a new model used to study and analyse the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) epidemic-reported-data from Spain. This is a Hidden Markov Model whose hidden layer is a regeneration process with Poisson immigration, Po-INAR(1), together with a mechanism that allows the estimation of the under-reporting in non-stationary count time series. A novelty of the model is that the expectation of the unobserved process's innovations is a time-dependent function defined in such a way that information about the spread of an epidemic, as modelled through a Susceptible-Infectious-Removed dynamical system, is incorporated into the model. In addition, the parameter controlling the intensity of the under-reporting is also made to vary with time to adjust to possible seasonality or trend in the data. Maximum likelihood methods are used to estimate the parameters of the model.

OPEN ACCESS

Citation: Fernández-Fontelo A, Moriña D, Cabaña A, Arratia A, Puig P (2020) Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. PLoS ONE 15(12): e0242956. https://doi.org/10.1371/journal.pone.0242956

Editor: Paul K. Newton, University of Southern California, UNITED STATES

Received: July 5, 2020

Accepted: November 12, 2020

Previously proposed models (continuous data)

Non-correlated longitudinal data

Moriña et al. BMC Medical Research Methodology (2021) 21:6
https://doi.org/10.1186/s12874-020-01188-4

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Quantifying the under-reporting of uncorrelated longitudal data: the genital warts example

David Moriña[1,2*], Amanda Fernández-Fontelo[3], Alejandra Cabaña[4], Pedro Puig[4], Laura Monfil[5], Maria Brotons[5] and Mireia Diaz[5]

Abstract

Background: Genital warts are a common and highly contagious sexually transmitted disease. They have a large economic burden and affect several aspects of quality of life. Incidence data underestimate the real occurrence of genital warts because this infection is often under-reported, mostly due to their specific characteristics such as the asymptomatic course.

Methods: Genital warts cases for the analysis were obtained from the Catalan public health system database (SIDIAP) for the period 2009-2016. People under 15 and over 94 years old were excluded from the analysis as the incidence of genital warts in this population is negligible. This work introduces a time series model based on a mixture of two distributions, capable of detecting the presence of under-reporting in the data. In order to identify potential differences in the magnitude of the under-reporting issue depending on sex and age, these covariates were included in the model.

Results: This work shows that only about 80% in average of genital warts incidence in Catalunya in the period 2009-2016 was registered, although the frequency of under-reporting has been decreasing over the study period. It can also be seen that this issue has a deeper impact on women over 30 years old.

Conclusions: Although this study shows that the quality of the registered data has improved over the considered period of time, the Catalan public health system is underestimating genital warts real burden in almost 10,000 cases, around 23% of the registered cases. The total annual cost is underestimated in about 10 million Euros respect the 54 million Euros annually devoted to genital warts in Catalunya, representing 0.4% of the total budget.

Keywords: Genital warts, Estimation, HPV, Under-reporting, Time series

Previously proposed models (continuous data)

Classical time series

scientific reports

Check for updates

OPEN

New statistical model for misreported data with application to current public health challenges

David Moriña[1,2], Amanda Fernández-Fontelo[3], Alejandra Cabaña[4] & Pedro Puig[2,4]

The main goal of this work is to present a new model able to deal with potentially misreported continuous time series. The proposed model is able to handle the autocorrelation structure in continuous time series data, which might be partially or totally underreported or overreported. Its performance is illustrated through a comprehensive simulation study considering several autocorrelation structures and three real data applications on human papillomavirus incidence in Girona (Catalonia, Spain) and Covid-19 incidence in two regions with very different circumstances: the early days of the epidemic in the Chinese region of Heilongjiang and the most current data from Catalonia.

Model

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 $\epsilon_t \sim N(\mu_\epsilon(t), \sigma^2)$.
In our setting, this process X_t cannot be directly observed, and all we can see is a part of it, expressed as

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

Model

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 ($R(t)$). The number of affected individuals at time t, $A(t) = I(t) + R(t)$ can be approximated by

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

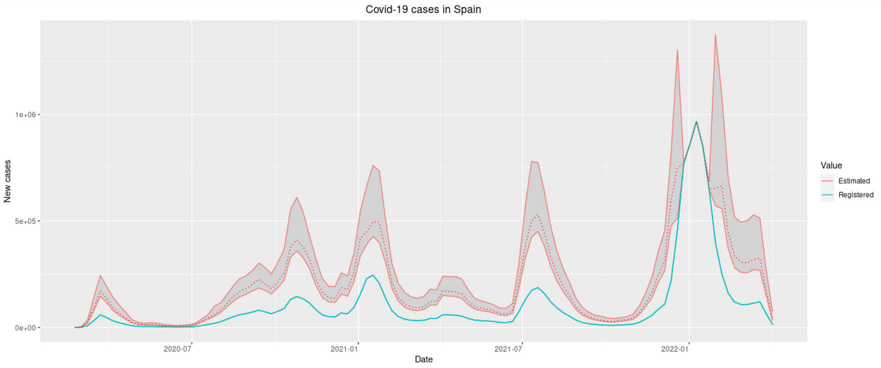
BSL

- Synthetic likelihood is a recent and very powerful alternative for parameter estimation in a simulation based schema when the likelihood is intractable but the generation of new observations given the values of the parameters is feasible
- Introduced by Simon Wood in 2010, Bayesian framework by Leah Price *et al.* in 2018
- 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

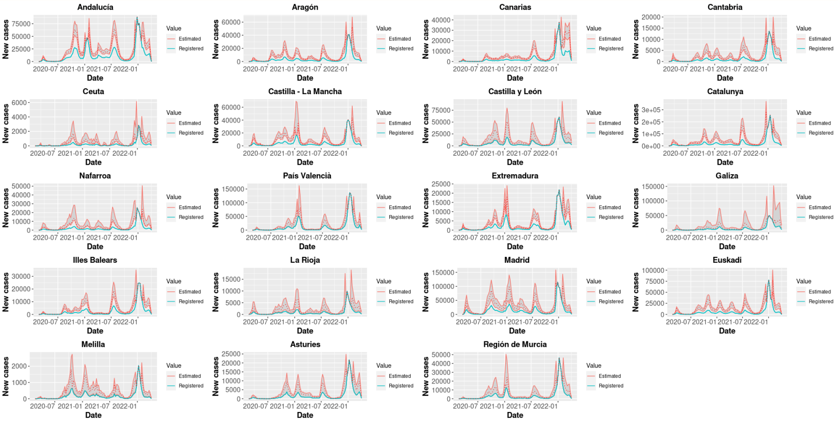
Covid-19 in Spain

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), 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 (2020/02/23-2022/04/03)

12 Covid-19 in Spain



Covid-19 in Spain



Covid-19 in Spain

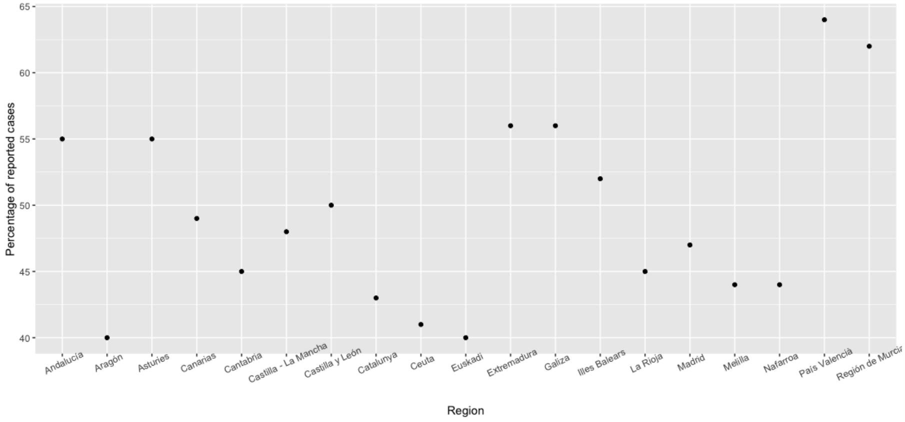
CCAA	Parameter	Estimate (95% CI)
Andalucía	$\hat{\omega}$ $\hat{q}$	0.97 (0.95 - 0.99) 0.44 (0.41 - 0.48)
Aragón	$\hat{\omega}$ $\hat{q}$	0.98 (0.97 - 0.99) 0.28 (0.27 - 0.32)
Asturies	$\hat{\omega}$ $\hat{q}$	0.97 (0.90 - 0.99) 0.40 (0.37 - 0.53)
Cantabria	$\hat{\omega}$ $\hat{q}$	0.97 (0.95 - 0.99) 0.30 (0.28 - 0.35)
Castilla y León	$\hat{\omega}$ $\hat{q}$	0.98 (0.95 - 0.99) 0.36 (0.32 - 0.41)
Castilla - La Mancha	$\hat{\omega}$ $\hat{q}$	0.98 (0.96 - 0.99) 0.33 (0.31 - 0.36)
Canarias	$\hat{\omega}$ $\hat{q}$	0.98 (0.96 - 0.99) 0.35 (0.32 - 0.38)
Catalunya	$\hat{\omega}$ $\hat{q}$	0.98 (0.96 - 0.99) 0.30 (0.27 - 0.34)
Ceuta	$\hat{\omega}$ $\hat{q}$	0.98 (0.95 - 0.99) 0.28 (0.25 - 0.31)

CCAA	Parameter	Estimate (95% CI)
Extremadura	$\hat{\omega}$ $\hat{q}$	0.98 (0.95 - 1.00) 0.40 (0.36 - 0.44)
Galiza	$\hat{\omega}$ $\hat{q}$	0.84 (0.33 - 0.98) 0.41 (0.35 - 0.56)
Illes Balears	$\hat{\omega}$ $\hat{q}$	0.98 (0.96 - 0.99) 0.36 (0.33 - 0.39)
Región de Murcia	$\hat{\omega}$ $\hat{q}$	0.93 (0.45 - 0.98) 0.46 (0.34 - 0.80)
Madrid	$\hat{\omega}$ $\hat{q}$	0.98 (0.96 - 0.99) 0.37 (0.34 - 0.40)
Nafarroa	$\hat{\omega}$ $\hat{q}$	0.99 (0.97 - 1.00) 0.30 (0.26 - 0.32)
Euskadi	$\hat{\omega}$ $\hat{q}$	0.99 (0.97 - 0.99) 0.27 (0.25 - 0.31)
La Rioja	$\hat{\omega}$ $\hat{q}$	0.98 (0.96 - 0.99) 0.31 (0.28 - 0.35)
Melilla	$\hat{\omega}$ $\hat{q}$	0.97 (0.95 - 0.99) 0.34 (0.31 - 0.37)
País Valencià	$\hat{\omega}$ $\hat{q}$	0.95 (0.40 - 0.98) 0.46 (0.40 - 0.67)

Covid-19 in Spain

- In the considered period, the official sources reported 11,612,568 Covid-19 cases in Spain, while the model estimates a total of 23,674,309 cases (only 51% of actual cases were reported)
- While the frequency of underreporting is extremely high for all regions (minimum value of $\hat{\omega} = 0.84$ in Galiza), the intensity of this underreporting is not uniform across the considered regions. It can be seen that Aragón and Ceuta are the regions with highest underreporting intensity ($\hat{q} = 0.28$) while País Valencià and Región de Murcia are the ones where the estimated values are closest to the number of reported cases ($\hat{q} = 0.46$)

Covid-19 in Spain



Impact of covariates

CCAA	Covariate	Estimate (95% CI)
Andalucía	$\hat{Vacc}$ $\hat{Conf}$	-1.71 (-2.66, -0.68) -1.67 (-2.31, -0.39)
Aragón	$\hat{Vacc}$ $\hat{Conf}$	-1.06 (-1.36, -0.69) 0.76 (0.17, 1.43)
Asturies	$\hat{Vacc}$ $\hat{Conf}$	-0.90 (-1.77, -0.63) 0.44 (0.11, 0.69)
Cantabria	$\hat{Vacc}$ $\hat{Conf}$	-0.53 (-1.29, -0.25) -0.44 (-0.71, 0.002)
Castilla y León	$\hat{Vacc}$ $\hat{Conf}$	-1.22 (-1.88, -0.60) -0.84 (-1.33, -0.23)
Castilla - La Mancha	$\hat{Vacc}$ $\hat{Conf}$	-0.80 (-1.11, -0.40) 0.06 (-0.18, 0.44)
Canarias	$\hat{Vacc}$ $\hat{Conf}$	-1.34 (-1.78, -1.06) -0.92 (-2.06, -0.29)
Catalunya	$\hat{Vacc}$ $\hat{Conf}$	-1.51 (-1.97, -0.94) -0.25 (-0.52, 0.21)
Ceuta	$\hat{Vacc}$ $\hat{Conf}$	-1.38 (-1.93, -0.84) 0.007 (-0.52, 0.34)

CCAA	Parameter	Estimate (95% CI)
Extremadura	$\hat{Vacc}$ $\hat{Conf}$	-0.72 (-1.30, -0.37) 1.45 (1.24, 1.83)
Galiza	$\hat{Vacc}$ $\hat{Conf}$	-2.03 (-3.07, -1.34) -0.20 (-0.53, 0.18)
Illes Balears	$\hat{Vacc}$ $\hat{Conf}$	-0.72 (-1.16, -0.34) 0.74 (0.43, 1.01)
Región de Murcia	$\hat{Vacc}$ $\hat{Conf}$	-1.97 (-3.07, -0.59) 0.62 (-0.02, 1.36)
Madrid	$\hat{Vacc}$ $\hat{Conf}$	-0.35 (-0.77, -0.07) 0.35 (-0.39, 0.59)
Nafarroa	$\hat{Vacc}$ $\hat{Conf}$	-2.05 (-3.20, -1.33) -1.71 (-1.92, -0.53)
Euskadi	$\hat{Vacc}$ $\hat{Conf}$	-0.10 (-0.24, 0.00) -0.42 (-0.69, -0.21)
La Rioja	$\hat{Vacc}$ $\hat{Conf}$	-0.43 (-0.71, -0.22) -0.83 (-1.08, -0.35)
Melilla	$\hat{Vacc}$ $\hat{Conf}$	-1.59 (-2.05, -0.93) -0.48 (-0.82, -0.11)
País Valencià	$\hat{Vacc}$ $\hat{Conf}$	-1.70 (-2.64, -0.52) 1.45 (1.24, 1.83)

Conclusions

- Having accurate data is key in order to provide public health decision-makers with reliable information
- 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
- The analysis of the Spanish Covid-19 data shows that in average only about 51% of the cases in the period 2020/02/23-2022/04/03 were reported

Next steps

- Model diagnostics
- Simulation study
- Underlying process model selection procedure

David Moriña, Amanda Fernández-Fontelo, Alejandra Cabaña, Argimiro Arratia and Pere Puig

Thank you!

dmorina@ub.edu