Introduction to Count Time Series

Alejandra Cabaña

Universitat Autònoma de Barcelona

Mini-course ECODEP, February 8 and 15, 2021

Overview

Under-reported data

Under-reporting in data refers to some issue, incident, phenomenon which is responsible for reporting less than the actual level of count data.

The problem of under-reporting is very common in many contexts such as epidemiological, biomedical and social research among others. As for public health, it is well known that some diseases have been traditionally under-reported.

Due to this phenomenon

- inferences might be highly biased,
- assumptions of classical models might be invalidated.

There can be several sources of under-reporting: accuracy of public health registers, social issues, economical interests, ...

CoVID-19

The number of infections by SARS-CoV2 cases, and sometimes even deaths seem to be under-reported worldwide.

Why?

- asymptomatic individuals, or people who experience a mild form of the desease,
- lack of tests to carry out large-scale screenings
- data management, testing policies
- $\vdots$

Under-reported data issue

The New York Times

The Coronavirus Outbreak › LIVE Latest Updates Maps and Cases Reopenings and Closings Governors vs. The Virus

Actual Coronavirus Infections Vastly Undercounted, C.D.C. Data Shows

The prevalence of infections is more than 10 times higher than the counted number of cases in six regions of the United States.

Europe

May 9th 2020 edition ›

Uncounted, unseen

Many covid deaths in care homes are unrecorded

Governments are waking up to a hidden calamity

PARIS

Editor's note: The Economist *is making some of its most important coverage of the covid-19 pandemic freely available to readers of* The Economist Today, *our daily newsletter. To receive it, register here. For our coronavirus tracker and more coverage, see our hub*

- We have devised a very simple methodology for modelling such phenomenon, that has worked well in the past.

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. Copyright © 2016 John Wiley & Sons, Ltd.

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

- It works even for non-stationary series with patterns of trend and/or seasonality by specifying appropriate link functions in the parameters.

- Covariates can also be included.

INAR-hidden Markov chain model

Consider a hidden process X_n with Po-INAR(1) structure:

$$X_n = \alpha \circ X_{n-1} + W_n(\lambda),$$

where $0 < \alpha < 1$ is a fixed parameter, $W_n \sim \text{Poisson}(\lambda)$, i.i.d., independent of X_n and $\circ$ is the binomial thinning operator:

$$\alpha \circ X_{n-1} = \sum_{i=1}^{X_{n-1}} Z_i$$

where Z_i are i.i.d Bernoulli(α).
The INAR(1) process is a homogeneous Markov chain with transition probabilities

$$\mathbf{P}(X_n = i | X_{n-1} = j) = \sum_{k=0}^{i \wedge j} \binom{j}{k} \alpha^k (1-\alpha)^{j-k} \mathbf{P}(W_n = i - k)$$

A simple under-reporting scheme

The under-reported phenomenon is modelled by assuming that the observed counts are

$$Y_n = \begin{cases} X_n & \text{with probability } 1-\omega \\ q \circ X_n & \text{with probability } \omega, \end{cases}$$

where ω and q represent the frequency and intensity of the under-reporting process, respectively.
That is, we observe

$$Y_n = (1-\mathbf{1}_n)X_n + \mathbf{1}_n \sum_{j=1}^{X_n} \xi_j \qquad \mathbf{1}_n \sim \text{Bern}(\omega), \quad \xi_j \sim \text{Bern}(q)$$

Properties of the model

- The stationary distribution of an INAR(1) process X_n with Poisson(λ) innovations is Poisson with mean and variance

$$\mu_X = \sigma_X^2 = \frac{\lambda}{1-\alpha}$$

- Its auto-covariance and auto-correlation functions are

$$\gamma_X(k) = \alpha^{|k|}\lambda \qquad\qquad \rho_X(k) = \alpha^{|k|}$$

- $\mathbf{E}Y_n = \mu_Y = \mu_X(1-\omega(1-q))$.
- The auto-covariance function of the observed process Y_n is

$$\gamma_Y(k) = (1-\omega(1-q))^2\alpha^{|k|}\mu_X$$

Hence, the auto-correlation function of Y_n is a multiple of $\rho_X(k)$:

$$\rho_Y(k) = \frac{(1-\alpha)(1-\omega(1-q))^2}{(1-\alpha)(1-\omega(1-q)) + \lambda(\omega(1-\omega)(1-q)^2)}\alpha^{|k|} = c(\alpha,\lambda,\omega,q)\alpha^{|k|}.$$

Parameter estimation. A naïve method

The marginal probability distribution of Y_n is a **mixture of two Poisson distributions**

$$Y_n \sim \begin{cases} \text{Poisson}\left(\frac{\lambda}{1-\alpha}\right) & \text{with probability } 1-\omega, \\ \text{Poisson}\left(\frac{q\lambda}{1-\alpha}\right) & \text{with probability } \omega. \end{cases}$$

When $q = 0$ the distribution of the observed process $\{Y_n\}$ is a zero-inflated Poisson distribution.

From the mixture we can derive initial estimations for α, λ, ω and q:

- Fit a mixture of two Poisson (i.e. `mixtools` R package) and get $\widehat{\omega}$, $\widehat{\frac{\lambda}{1-\alpha}}$, $\widehat{\frac{q\lambda}{1-\alpha}}$. Then obtaining $\widehat{q}$ is immediate.
- $\widehat{\alpha}$ can be obtained either by using $\widehat{\rho_1}$ and fitting the model

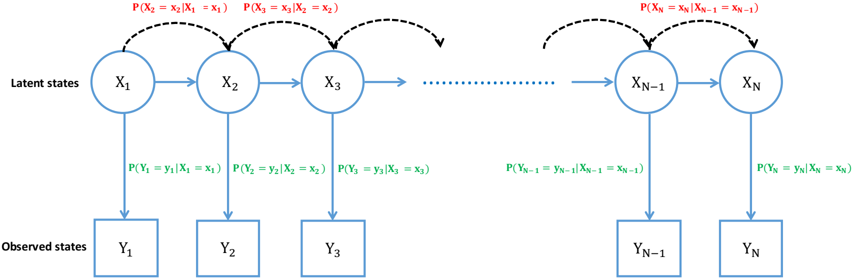
$$\widehat{\rho_1} = c(\alpha, \lambda, \omega, q)\ \alpha \qquad \text{or simply} \qquad \widehat{\alpha} = \frac{\widehat{\rho_2}}{\widehat{\rho_1}}$$

- Finally, estimate λ from $\widehat{\frac{q\lambda}{1-\alpha}}$, $\widehat{q}$ and $\widehat{\alpha}$.

Parameter estimation based on ML

Transition probabilities:

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}) = e^{-\lambda} \sum_{j=0}^{x_n \wedge x_{n-1}} \binom{x_{n-1}}{j} \alpha^j (1-\alpha)^{x_{n-1}-j} \frac{\lambda^{x_n - j}}{(x_n - j)!}$$



Emission probabilities:

$$P(Y_i = j \mid X_i = k) = \begin{cases} 0 & if \quad k < j \\ (1-\omega) + \omega q^k & if \quad k = j \\ \omega \binom{k}{j} q^j (1-q)^{k-j} & if \quad k > j, \end{cases}$$

Parameter estimation based on ML (ii)

The likelihood function of Y is directly intractable,

$$P(Y) = \sum_X P(X, Y) = \sum_x P(Y|X = x)P(X = x)$$

It can be computed recursively as a sum with the **forward algorithm**[1]. Starting with

$$\alpha_1(X_1) = P(X_1)P(Y_1|X_1)$$

and defining the **forward probabilities**

$$\alpha_k(X_k) = P(Y_k|X_k) \sum_{X_{k-1}} P(X_k|X_{k-1})\alpha_{k-1}(X_{k-1}),$$

Hence,

$$P(Y) = \sum_n \alpha_n(X_n).$$

---

[1] T.C. Lystig, J.P. Hughes (2002), *Exact computation of the observed information matrix for hidden Markov models*, Jr of Comp.and Graph. Stat.

Reconstructing the hidden chain X_n

In order to reconstruct the hidden series X_n, the **Viterbi algorithm**[2] is used.
The idea is to provide the latent chain $X_1^* = x_1^*, \ldots, X_N^* = x_N^*$ that maximises the likelihood of the latent process given the observed series, assuming all the parameters are known.

Let $P(X_{1:n}|Y_{1:n})$ be the likelihood function of the model, then

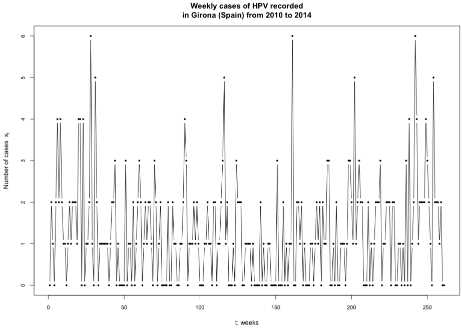
$$P(X_{1:n}|Y_{1:n}) = \frac{P(X_{1:n}, Y_{1:n})}{P(Y_{1:n})}$$

Since $P(Y_{1:n})$ does not depend on X_n, it is enough to maximise the probability $P(X_{1:n}, Y_{1:n})$.
The hidden series is reconstructed as:

$$X^* = \arg\max_X P(X_{1:n}, Y_{1:n}).$$

[2] Viterbi, A.J. (1967), *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Transactions on Information Theory **13**

The HPV data



- Human papillomavirus infection can be severely under-reported since most of the sexually active people carry it without any symptoms.
- The series can be considered stationary.
- Series ranges from 0 to 6 weekly cases, with a mean of 1.27 and a median of 1 case per week. The variance is 1.60. The dispersion index is 1.26 which is statistically different of 1 (p-value=0.0018): **overdispersed series**.
- Recall that a Poisson mix is always overdispersed.

Results of the naïve method

We have applied the naïve approach in order to:

- evaluate whether the stationary series is actually under-reported.
  - Fitting the model $\rho_Y(k) = c(\alpha, \lambda, \omega, q)\alpha^{|k|}$ applying logarithms in order to know whether $log(c(\alpha, \lambda, \omega, q))$ is significant (p-value=0.0002).
  - Modeling Y_n as a unique Poisson distribution or a mixture of two Poisson distributions (the mixture presents a smaller AIC).

- to get initial estimates of the parameters:

$$\widehat{\frac{\lambda}{1-\alpha}} = 2.612, \ \widehat{\frac{q\lambda}{1-\alpha}} = 0.967 \text{ and } \widehat{\omega} = 0.805,$$

and $\widehat{q} = 0.371$, $\widehat{\alpha} = 0.342$ and $\widehat{\lambda} = 1.72$.

ML estimation

The estimated INAR(1) model for the hidden chain X_n is

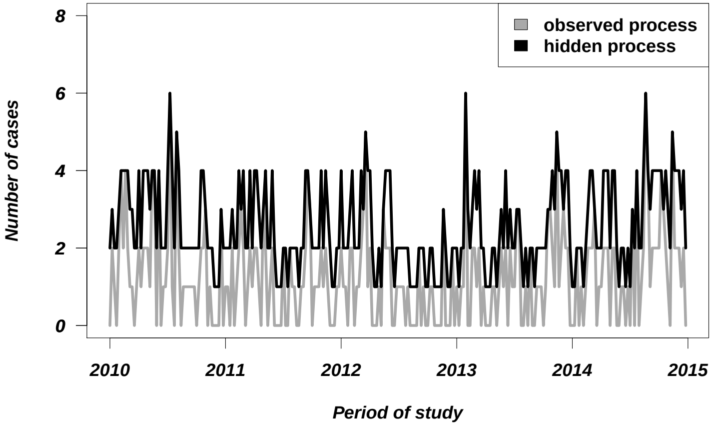
$$X_n = 0.517 \circ X_{n-1} + W_n \qquad W_n \sim \text{Poiss}\,(1.623)\,,$$

with the following structure for the observed chain Y_n,

$$Y_n = \begin{cases} X_n & : \text{with probability } 0.078 \\ 0.327 \circ X_n & : \text{with probability } 0.922. \end{cases}$$

Parameter	ML estimate	s.e.
$\hat{\alpha}$	0.517	0.227
$\hat{\lambda}$	1.623	0.616
$\hat{\omega}$	0.922	0.073
$\hat{q}$	0.327	0.085

Reconstruction of the underlying series



Checking the model

Model validation can be done by using the normal pseudo-residuals. In the discrete case, the normal pseudo-residuals segment $[z_n^-, z_n^+]$ are required,

$$z_n^- = \Phi^{-1}\left(P(Y_n < y_n \mid (Y_1, Y_2, \cdots, Y_{n-1}, Y_{n+1}, \cdots, Y_T))\right) = \Phi^{-1}(u_n^-)$$
$$z_n^+ = \Phi^{-1}\left(P(Y_n \leq y_n \mid (Y_1, Y_2, \cdots, Y_{n-1}, Y_{n+1}, \cdots, Y_T))\right) = \Phi^{-1}(u_n^+)$$

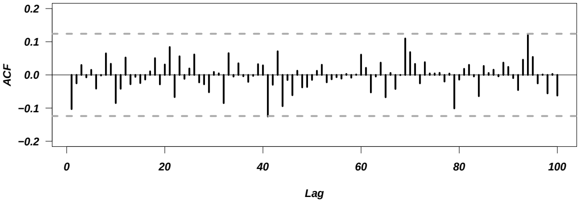
In order to use them in a plot, normally the mid-pseudo residuals

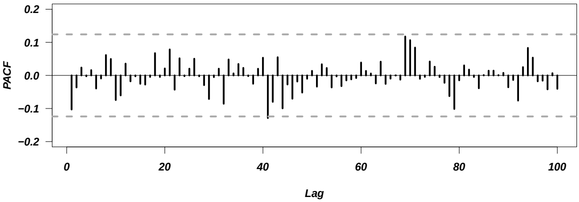
$$z_t^m = \phi^{-1}\left(\frac{u_t^- + u_t^+}{2}\right)$$

are considered.

If the model is adequate, the mid-pseudo-residuals should behave as white noise.

Model validation





Interpretation

- Both the frequency (ω close to 1) and intensity $q \in (0.159, 0.493)$ of the process are statistically significant indicating that the series is actually under-reported, as suspected.
- Although the observed average is 1.27, the hidden series has an estimated average of 3.36 cases per week.
- This means that public health policy makers only recorded 38% of the total HPV cases in the province of Girona during the time of the study.

Further comments

- We have extended the method to some non-stationary series: series with patterns of trend and/or seasonality by specifying appropriate link functions in the parameters λ, ω and q, for instance,

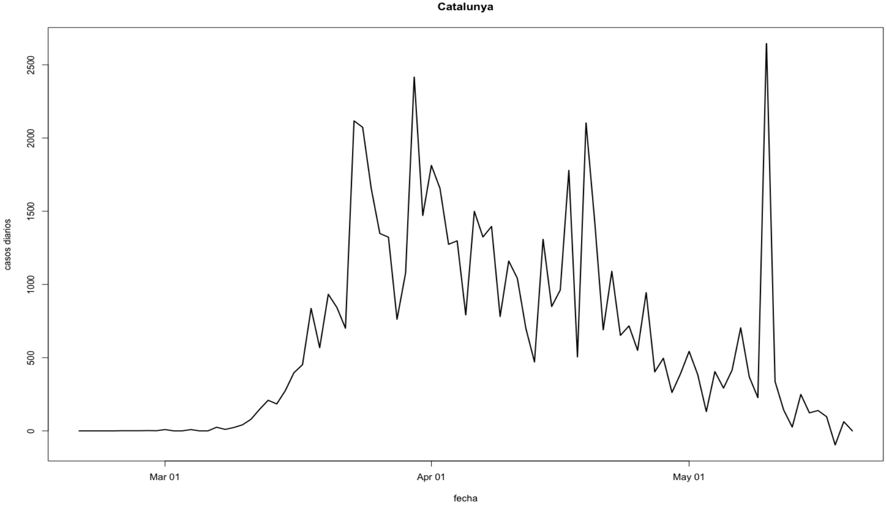
$$\lambda_t = e^{\beta_0 + \beta_1 t}$$

$$\omega_t = logit(\gamma_0 + \gamma_1 t)$$

- Covariates can also be included.

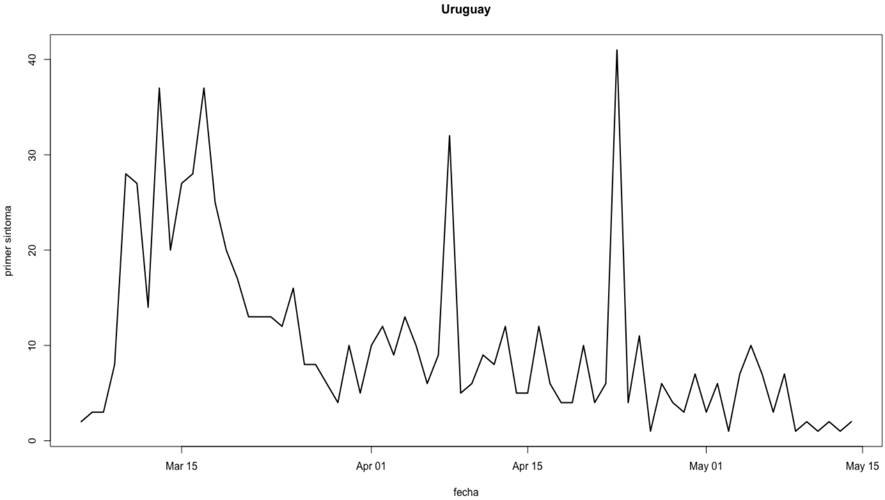
CoVID-19 data

As you already know, data on the number of "cases" of CoVID-19 is far from being stationary.



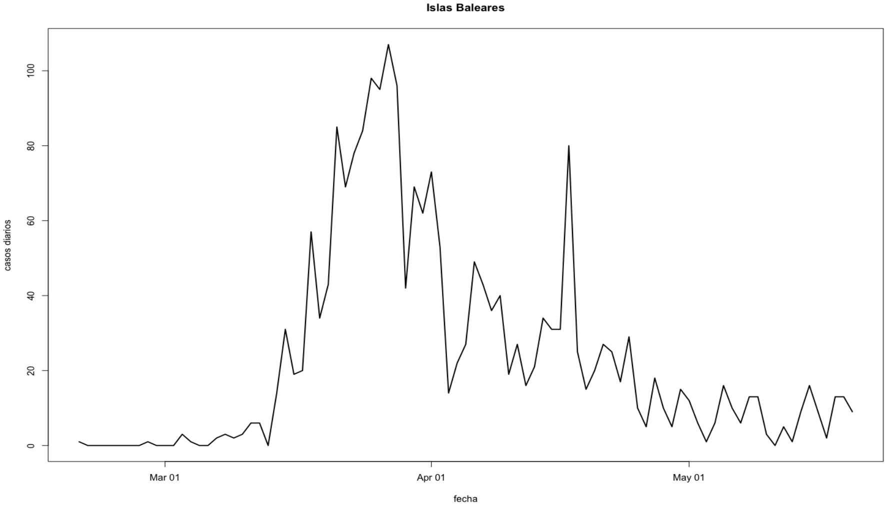
Data form `https://cnecovid.isciii.es/covid19/`.

CoVID-19 in Uruguay



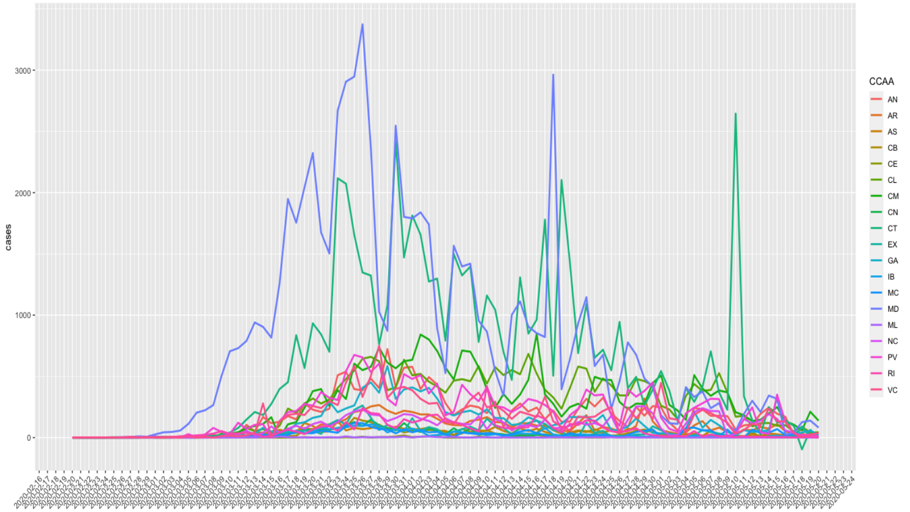
Data obtained from área de Epidemiología del MSP, corresponding to the date of 1st symptom.

CoVID-19 in Balearic Islands



Data form `https://cnecovid.isciii.es/covid19/`.

CoVID-19 in Spain



Data form `https://cnecovid.isciii.es/covid19/`.

Modelling non-stationary under-reported time series including information of the spread of the disease

As a first approach to find a more realistic model for the data we shall consider that

- the mean of the latent process X_n varies in time:

$$X_n = \alpha \circ X_{n-1} + W_n(\lambda_n)$$

and /or

- the intensity of the under-reporting varies in time[3]:

$$Y_n = (1 - \mathbf{1}_n) X_n + \mathbf{1}_n \sum_{j=1}^{X_n} \xi_j \qquad \mathbf{1}_n \sim \text{Bern}(\omega), \quad \xi_j \sim \text{Bern}(q_n)$$

[3] periodic effects?

A model for the mean of the latent process

Consider the simplest SIR model,

$$\begin{aligned}\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\beta\frac{I(t)S(t)}{N}\\ \frac{\mathrm{d}I(s)}{\mathrm{d}t} &= \beta\frac{I(t)S(t)}{N} - \gamma I(t)\\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t)\end{aligned}$$

S healthy but susceptible to get the disease , $I(n)$ infected and thus transmitters of the disease and $R(n)$ the removeded individuals who will not get infected again.
The parameters of interest are the infection rate β, the removal rate γ, and the susceptible population N.

Now call $A(t)$ the population affected by the desease:

$$A(t) = I(t) + R(t) \Rightarrow S(t) = N - A(t)$$

Adding the last two equations, and replacing $I(t) = A(t) - R(t)$ in the last one we get

$$\begin{aligned}
\frac{\mathrm{d}A(t)}{\mathrm{d}t} &= \frac{\beta}{N}\left(N - A(t)\right)\left(A(t) - R(t)\right) \\
\frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma(A(t) - R(t))
\end{aligned}$$

so that

$$\frac{\mathrm{d}R}{\mathrm{d}A} = \frac{\mathrm{d}R}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}A} = \frac{\gamma}{\beta}\frac{N}{N - A(t)} \Rightarrow R(t) = \frac{N\gamma}{\beta}\log\left(\frac{N - A_0}{N - A(t)}\right) + R_0$$

and replacing $R(t)$ in the first equation to get

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = \frac{\beta}{N}\left(A(t) - \frac{N\gamma}{\beta}\log\left(\frac{N - A_0}{N - A(t)}\right) - R_0\right)(N - A(t)) \approx$$

$$(\beta - \gamma)A(t) - \frac{(\beta - \gamma/2)}{N}A^2(t)$$

considering that $R_0 = 0$ and $A_0 \approx 0$.

The previous equation has the form

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = kA(t)\left(1 - \frac{A}{M^*}\right)$$

where $k = \beta - \gamma$ and $M^* = \dfrac{N(\beta - \gamma)}{\beta - \gamma/2}$ and its solution is the logistic function

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

Close to the origin, $A(t) \approx A_0 \mathrm{e}^{kt}$, and A_∞ can be found solving

$$\frac{\beta}{N}\left(A_\infty - \frac{N\gamma}{\beta}\log\left(\frac{N - A_0}{N - A_\infty}\right) - R_0\right)(N - A_\infty) = 0$$

Finally, the expectation of the innovations

We will consider that the expectation of the hidden process X_n at time n is the number of new infected individuals

$$\lambda_n = A(n) - A(n-1)$$

hence λ_n grows exponentially at the begining, and after reaching its maximum A_∞ decreases exponentially.

The parameters of the model are then

- α, $m = \log(M^*)$, k for the latent process
- ω and q (or q_n) for the under-reporting

Parameters are estimated by Maximum Likelihood, and the likelihood is computed with the Forward Algorithm again.

Forecasting Expected values

In order to predict the values of $Y_{n+1}, Y_{n+2}, \ldots$ given $Y_1, Y_2, \cdots, Y_n$, recall that

$$\mathbf{E}(Y_{n+k}) = \mathbf{E}\left(X_{n+k}\right)(1-\omega(1-q_n))$$

where $\mathbf{E}(X_{n+k}) = \frac{\lambda_{n+k}}{1-\alpha}$ and $\lambda_{n+k} = A(n+k) - A(n+k-1)$.
Since

$$\mathbf{E}(X_{n+k}) = \alpha^k \mathbf{E}(X_n) + \sum_{i=1}^{k} \alpha^{k-i} \lambda_{n+i}$$

if we have estimates for the corresponding parameters at a given time $n+k$, the average point prediction of Y_{n+k} is

$$\mathbf{E}(Y_{n+k}) = \varphi_{n+k} \alpha^k Y_n + \frac{1-\omega(1-q_{n+k})}{1-\omega(1-q_n)} \left(1-\omega(1-q_n)\right) \sum_{i=1}^{k} \alpha^{k-i} \lambda_{n+i}$$

and the standard errors are computed via the *Delta method*.

Forecasting with the Conditional distribution

The conditional distribution of Y_{n+k} given X_n is a mixture of two components, corresponding to the situations in which the latent process is fully observed or under-reported respectively, and each of them is a sum of a Binomial rv and a Poisson rv.

$$Y_{n+k}|X_n = x_n \sim \begin{cases} \text{Bin}\left(\alpha^k, x_n\right) + \text{Poiss}\left(\sum_{i=1}^k \alpha^{k-i}\lambda_{n+i}\right) & 1-\omega \\ \text{Bin}\left(q_{n+k}\alpha^k, x_n\right) + \text{Poiss}\left(q_{n+k}\sum_{i=1}^k \alpha^{k-i}\lambda_{n+i}\right) & \omega \end{cases}$$

Replacing the parameters by their maximum likelihood estimates, we can also estimate regions of prediction of size $1-\alpha^*$ finding the lower and upper limits r_1 and r_2 that satisfy:

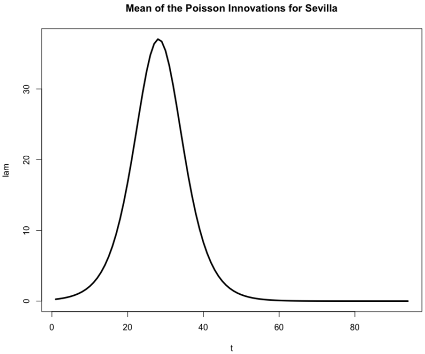
$$\sum_{j=1}^{r_1} \mathrm{P}(Y_{n+k} = j|X_n = x_n) \approx \alpha^*/2$$

$$\sum_{j=1}^{r_2} \mathrm{P}(Y_{n+k} = j|X_n = x_n) \approx 1-\alpha^*/2$$

Sevilla data

Using data from Portal de la Junta de Andalucía[4], from February 26th 2020 to May 29th 2020 for the number of daily counts of PCR+ tests, we get

$$X_n = 0.82 \circ X_{n-1} + W_n(\lambda_n) \qquad \text{where } \lambda_n \text{ is ,}$$

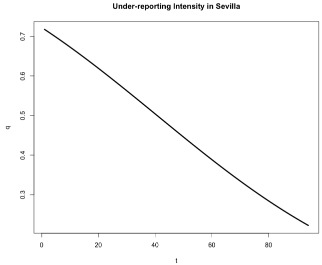


[4]https://www.juntadeandalucia.es/institutodeestadisticaycartografia/badea/informe/anual?CodOper=b3_2314&idNode=42348

And the model for the observations

$$Y_n = \begin{cases} X_n & \text{with probability } 0.29 \\ q_n \circ X_n & \text{with probability } 0.71, \end{cases}$$

where $q_n = \text{logit}(0.95 - 0.02n)$

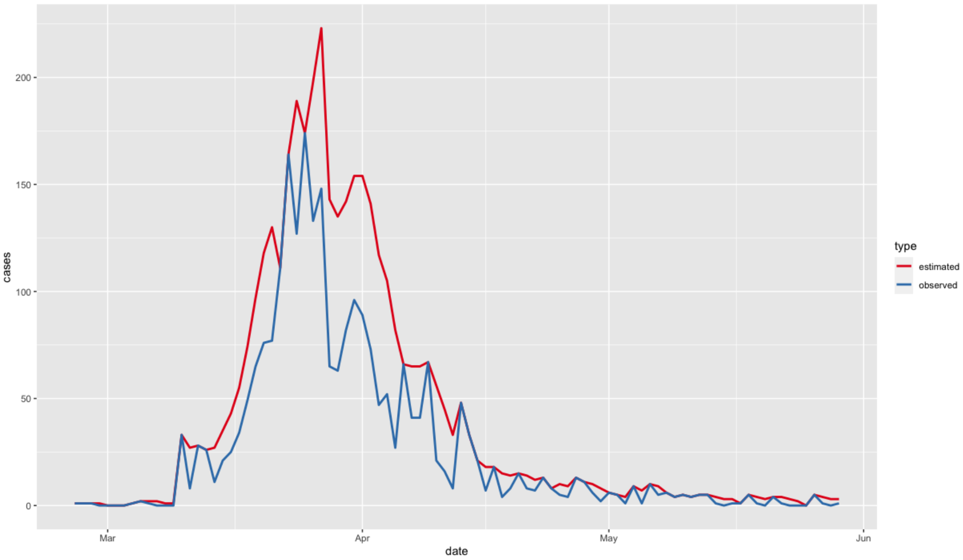


2489 observed cases

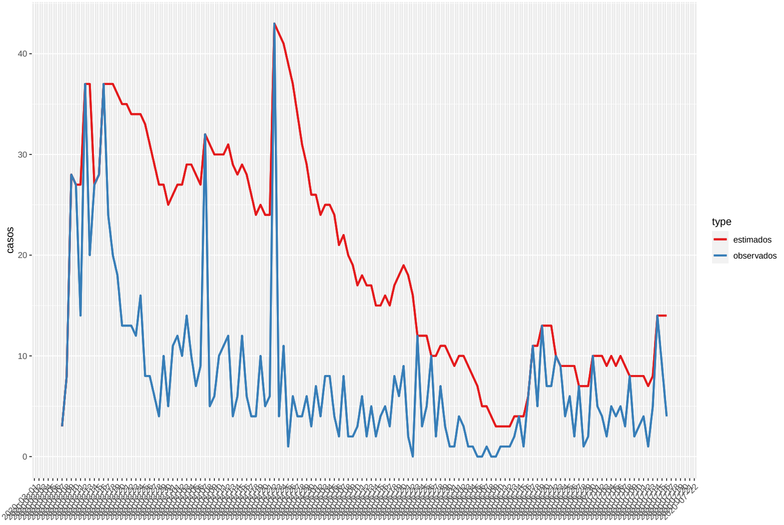
3755 estimated cases

that is, the system is overlooking 1266 individuals (51%), or in other words, assuming that the Viterbi estimation is the true number of infected individuals, only **66%** are reported.

Sevilla daily number of PCR+ cases

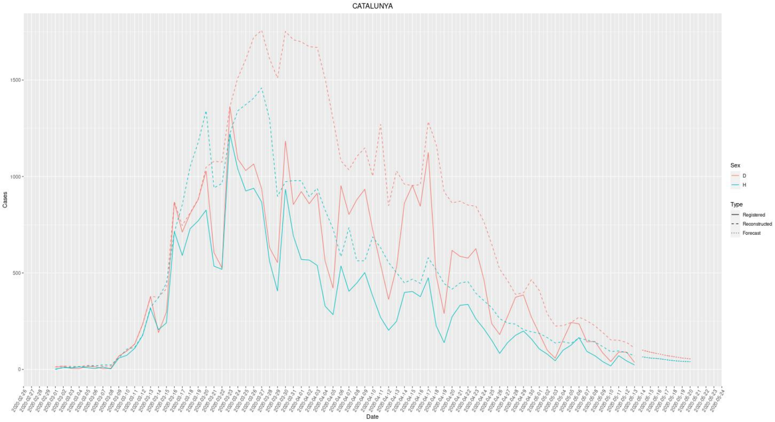


Uruguay data



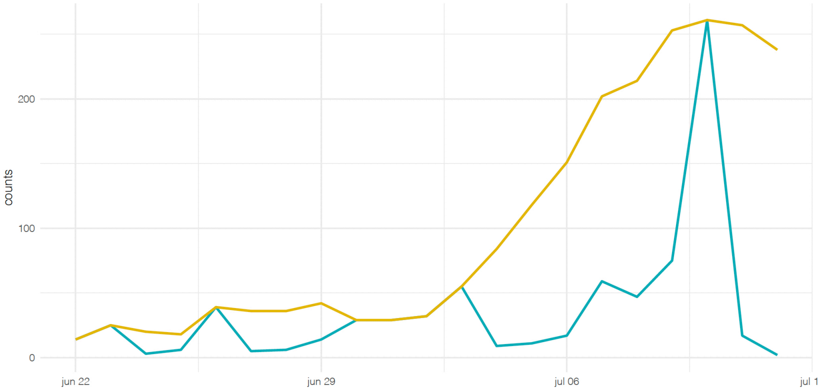
Casos segun fecha del primer sintoma

Cases in Catalunya by gender



The model estimates that in the period 01/03/2020 to 13/05/2020, 92469 cases of COVID-19 have occurred in Catalunya, of which 59887 were registered . That is, 54.41% of cases (32582) would not have been registered in the system.

Barcelona immediately after lockdown

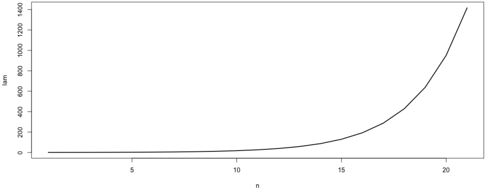


755 observed cases

2153 estimated cases, from June 22 till July 13.

The model for Barcelona after lockdown

$$X_n = 0.8240 \circ X_n + W(\lambda_n) \qquad \text{with } \lambda_n$$



And the model for the observations is

$$Y_n = \begin{cases} X_n & \text{with probability } 0.3810 \\ q_n \circ X_n & \text{with probability } 0.6190, \end{cases}$$

where

$$q_n = \operatorname{logit}\left(-1.0169 - 0.0427n + 0.1245 \sin\left(\frac{2\pi n}{7}\right) + 0.0947 \cos\left(\frac{2\pi n}{7}\right)\right)$$

Predictions

Red lines indicate 95% CI; blue: mean; yellow: median

Serially dependent under-reporting scheme

Recall we have a latent series

$$X_n = \alpha \circ X_{n-1} + W_n(\lambda)$$

and observe

$$Z_n = (1 - \mathbf{1}_n)X_n + \mathbf{1}_n \sum_{j=1}^{X_n} \xi_j \qquad \mathbf{1}_n \sim \text{Bern}(\omega), \quad \xi_j \sim \text{Bern}(q)$$

Imagine that the under-reporting indicators $\{\mathbf{1}_n, n \geq 1\}$ are serially dependent.

Take the simplest scheme: a binary discrete time Markov Chain[5].

[5] Zucchini & McDonald (2009)

Consider a binary discrete-time Markov chain with transition probabilities

$$\mathrm{P}(\mathbf{1}_n = j|\mathbf{1}_{n-1} = i) = p_{ij} \quad , i, j = 0, 1, \text{ with matrix:}$$

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

Using the stationary distribution $\Omega P = \Omega$ where $\Omega = (1 - \omega, \omega)$, the transition matrix P can be written in terms of only one parameter, for example p_{01}:

$$P = \begin{bmatrix} 1 - p_{01} & p_{01} \\ p_{01}\frac{1-\omega}{\omega} & 1 - p_{01}\frac{1-\omega}{\omega} \end{bmatrix}, \tag{1}$$

which also depends on ω, the overall frequency of under-reporting.

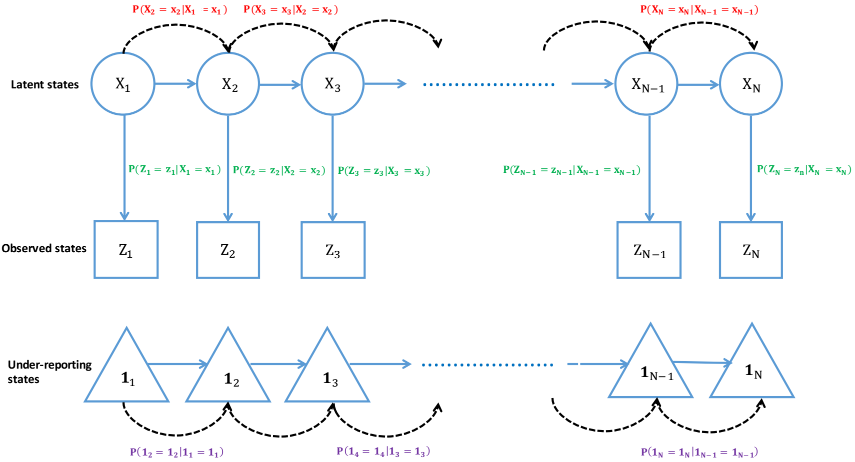
This new model is general in the sense that it considers both temporal dependence ($\alpha \neq 0$), and a dependence structure between the states of under-reporting.

M4a Assuming independence between the states of under-reporting it reduces to the previous model (Y_n) depending on $(\alpha, \lambda, \omega, q)$

M4b If there is no temporal dependence in the latent process ($\alpha = 0$), the parameters are $(\lambda, \omega, q, p_{01})$. The dependence in $\{Y_n\}$ is due to $\{\mathbf{1}_n\}$.

M3 If there is neither temporal dependence (α=0) nor dependence between the states of under-reporting, the model becomes a mixture of two Poisson distributions which depends on (λ, ω, q), where λ is the expected number of counts, and ω and q are the overall frequency and intensity of the under-reporting, respectively.

The hiden Markov Model



Model properties

The expectation and variance of the observed process are

$$\mathbf{E}(Z_n) = \mu_X(1-\omega(1-q)) \quad \mathbf{Var}(Z_n) = \mu_X^2\omega(1-\omega)(1-q)^2 + \mu_X(1-\omega(1-q))$$

and coincide with those of the previous model (**M4a**). The auto-correlation function of $\{Z_n\}$ is slightly more complicated than that of **M4a**: it depends on the smallest eigenvalue (λ_2) of the transition matrix P:

- When $p_{01} = \omega$, it coincides with the ACF of **M4a**.
- When $\alpha > |\lambda_2|$, the ACF decreases geometrically at rate α.
- When $\alpha < |\lambda_2|$ and $\lambda_2 > 0$ (higher chance of remaining in the same state of under-reporting), the ACF decreases geometrically at rate λ_2.
- When $\alpha < |\lambda_2|$ and $\lambda_2 < 0$ (higher chance to change the state of under-reporting), the absolute value of the ACF decreases geometrically at rate $|\lambda_2|$.

Parameter estimation

Two different methods for estimating the parameters of the full model and its nested models:

- The method of moments, using moments related to the marginal distribution of the observed process Z_n (mixture of two Poisson distributions), and the expression of the auto-correlation function of Z_n.
- Confidence intervals are obtained with bootstrap.
- MLE: since the likelihood function of Z_n is not directly tractable, the forward algorithm for hidden Markov chains (HMC) is used for its computation.

Moment estimators

The marginal distribution of $\{Z_n\}$ can be fitted using EM, obtaining the estimates $\widehat{\omega}$, $\quad \widehat{\theta}_1 = \dfrac{\widehat{\lambda}}{(1-\widehat{\alpha})} \quad \widehat{\theta}_2 = \dfrac{\widehat{q}\widehat{\lambda}}{(1-\widehat{\alpha})}$ and the overall intensity of the under-reporting $\widehat{q} = \dfrac{\widehat{\theta}_2}{\widehat{\theta}_1}$.

The parameter α is estimated using the ACF of $\{Z_n\}$:

$$\rho_Z(k) = C_1\alpha^k + C_2\lambda_2^k\left(\mu_X + \alpha^k\right)$$

where

$$C_1 = \frac{(1-\omega(1-q))^2}{\mu_X\omega(1-\omega)(1-q)^2 + (1-\omega(1-q))}$$

$$C_2 = \frac{\omega(1-\omega)(1-q)^2}{\mu_X\omega(1-\omega)(1-q)^2 + (1-\omega(1-q))}$$

are completely determined by replacing parameters by their moment-based estimates obtained before. Notice that $\widehat{\mu}_X = \widehat{\sigma}_X^2 = \widehat{\theta}_1$.

With the two first coefficients of the empirical ACF, that is, $\rho_Z(1)$ and $\rho_Z(2)$, a system of two equations, which depend on α and λ_2, has to be solved.
This system can be re-written as:

$$A\alpha^4 + B\alpha^3 + D\alpha^2 + E\alpha + F = 0$$

and all the coefficients depend on quantities known or already estimated:

$$C_1 \quad C_2 \quad \mu_X \quad \rho_Z(1) \quad \rho_Y(2)$$

Hence, α might not be completely identifiable based on $\rho_Z(1)$ and $\rho_Z(2)$, and higher order serial correlations $\rho_Z(3)$ and $\rho_Z(4)$ might be needed. The better estimate of α corresponds to the positive real root of the equation which provides the theoretical coefficients $\rho_Z(3)$ and $\rho_Z(4)$ closer to the empirical ones.
This means that there might be up to four latent dependent processes that contribute to the ACF of the observed process $\{Z_n\}$.
Finally, the parameter λ can be directly obtained through

$$\widehat{\lambda} = \widehat{\theta}_1(1 - \widehat{\alpha})$$

Method of Moments on the nested models

When independence is considered between the states of under-reporting (**M4a**), parameters $(\alpha, \lambda, \omega, q)$ can be estimated modifying the previous set-up.

If both $\{X_n\}$ and $\{\mathbf{1}_n\}$ are independent (**M3**), the parameteres (λ, ω, q) are directly obtained by fitting a mixture of two Poisson distributions by means, for instance, of the EM-algorithm.

In order to provide estimates of the standard errors of the moment-based estimates of the model parameters, 90% confidence limits based on bootstrap samples are computed.

The moment-based estimates are useful as initial values for the algorithm that numerically maximises the log-likelihood function of the model.

Maximum Likelyhood Estimation

The likelihood function of the process $\{Z_n\}$ can be computed recursively through:

$$\mathrm{P}(\boldsymbol{Z}_{1:N} = \boldsymbol{z}_{1:N}) = \sum_{x_N = z_N}^{\infty} \alpha_N \left(\boldsymbol{z}_{1:N}, x_N, \mathbf{1}_N\right)$$

starting from

$$\alpha_1(z_1, x_1, \mathbf{1}_1) = \mathrm{P}\left(X_1 = x_1\right) \mathrm{P}\left(Z_1 = z_1 | X_1 = x_1\right) \mathrm{P}\left(\mathbf{1}_1 = i_1\right)$$

and

$$\mathrm{P}\left(X_1 = x_1\right) = \frac{e^{-\nu} \nu^{x_1}}{x_1!} \quad \text{with} \quad \nu = \frac{\lambda}{(1-\alpha)}$$

On the other hand, it is assumed that $\mathrm{P}(\mathbf{1}_1 = 0) = 1$ and $\mathrm{P}(\mathbf{1}_1 = 1) = 0$, that is, it is assumed that at time $n = 1$, the process is not under-reported.

Since the sequences $\{X_n\}$ and $\{\mathbf{1}_n\}$ are mutually independent, then

$$\alpha_n\left(z_{1:n}, x_n, i_n\right) = \sum_{x_{n-1}, i_{n-1}} \mathrm{P}\left(Z_n = z_n | X_n = x_n, \mathbf{1}_n = i_n\right) \mathrm{P}\left(X_n = x_n | X_{n-1} = x_{n-1}\right) \\ \times \mathrm{P}\left(\mathbf{1}_n = i_n | \mathbf{1}_{n-1} = i_{n-1}\right) \alpha_{n-1}\left(\boldsymbol{z}_{1:n-1}, x_{n-1}, i_{n-1}\right)$$

where $\mathrm{P}(X_n = x_n | X_{n-1} = x_{n-1})$, which is the conditional probability density function of an INAR(1) model with Poisson(λ) innovations, are the **transition probabilities** of the model.
On the other hand, $\mathrm{P}(\mathbf{1}_n = i_n | \mathbf{1}_{n-1} = i_{n-1})$ comes from the transition probability matrix $\boldsymbol{P}$, and the **emission probabilities** are

$$\mathrm{P}(Z_n = z_n | X_n = x_n, \mathbf{1}_n = i_n) = \begin{cases} 0 & \text{if} \quad x_n < z_n \\ 0 & \text{if} \quad i_n = 0, x_n > z_n \\ 1 & \text{if} \quad i_n = 0, x_n = z_n \\ \binom{x_n}{z_n} q^{z_n} (1-q)^{x_n - z_n} & \text{if} \quad i_n = 1, x_n \geq z_n \end{cases}$$

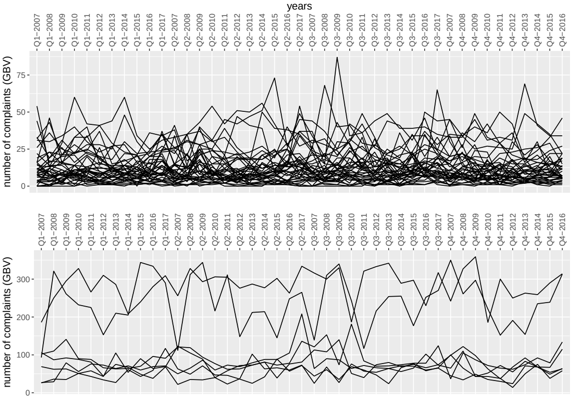
Application: Gender based violence complaints in Galicia, Spain

We analyse official data[6] form the number of quarterly complaints on gender violence against women from 2007 to 2017 in 35 judicial districts spread in the 4 provinces of the Autonomous Community of Galicia. Some districts have been removed from the study.

- Full model and nested models were fitted by the method of moments and by MLE (when counts were not too large).
- When estimating by MM, the criteria for selecting the best model is via ACF and empirical standard errors obtained through bootstrap.
- For the MLE the model is chosen using AIC and BIC criteria.
- Models are validated with pseudo-residuals.
- The latent series is reconstructed using the Viterbi Algorithm.

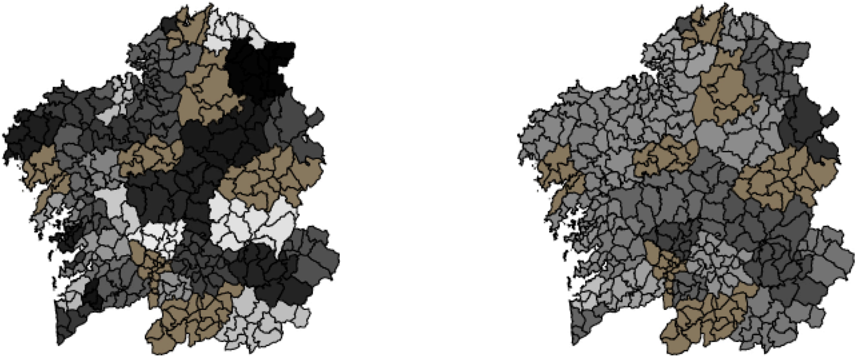
[6] Instituto Galego de Estatística

The data



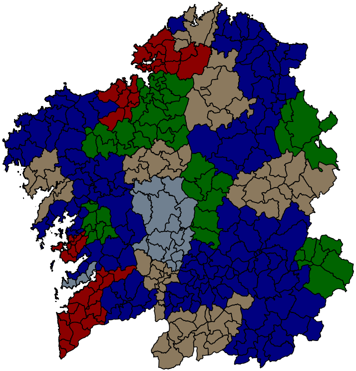
The series with larger counts belong to the main cities of Galicia, where population is much larger than in the other areas (*Vigo, A Coruña, Santiago de Compostela, Lugo, Pontevedra, Ferrol* and *Ourense*).

Overall frequency (ω) and intensity (q) of under-reporting

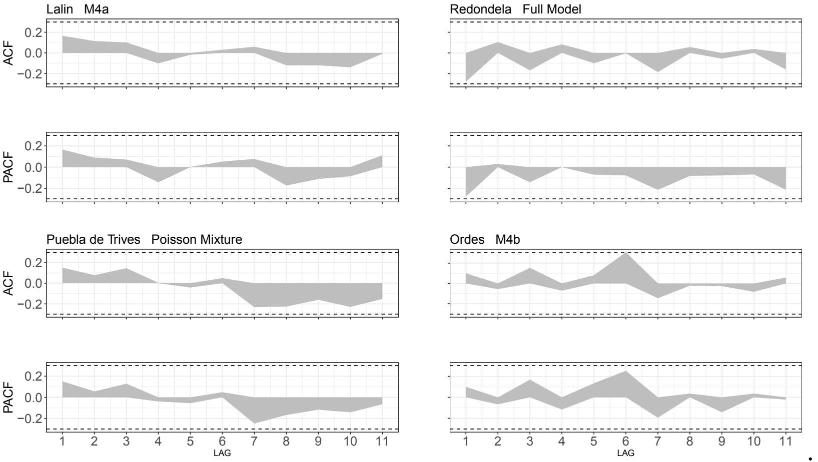


Left: estimated overall under-reporting **frequency**. The darker the area, the more frequent is the under-reporting (ω closer to 1). **Right:** estimated overall under-reporting **intensity**. The darker the area, the more intense, q closer to 0. Light brown corresponds to the removed areas.

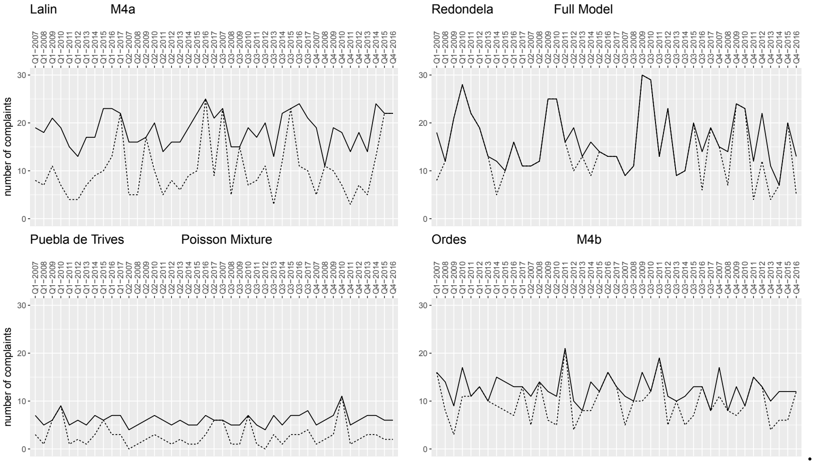
Geographical distribution of the different models



Blue: **M3** with mixtures of Poisson distributions. **Gray**: **M4a** (independence of under-reporting states). **Red**: full model. **Green**: nested model **M4b** (serially uncorrelated latent process). Light brown corresponds to regions that have been removed from the analysis.



ACF and PACF of the mid-pseudo residuals for validating models of *Lalin* (**M4a**), *Redondela* (**Full model**), *Puebla de Trives* (**Poisson Mixture**) and *Ordes* (**M4b).**



Most likely sequences of states (number of quarterly complaints) (dark line) in the observed series (dotted line) in *Lalin* (**M4a**), *Redondela* (**Full model**), *Puebla de Trives* (**Poisson Mixture**) and *Ordes* (**M4b)**.

Moment-based estimates for series with large counts (over 100) and 90% BCa CIs.

Table 1. Moment-based estimates for series with large counts and 90% BCa confidence intervals.

	α	λ	ω	q	p_{01}
A Coruña (A Coruña)	0.322 (0.078, 0.602) *0.161*	182.087 (109.690, 247.802) *43.179*	0.256 (0.140, 0.501) *0.103*	0.535 (0.509, 0.565) *0.021*	0.143 (0.052, 0.263) *0.069*
Ferrol (A Coruña)	0.516 (0.067, 0.723) *0.205*	48.536 (27.790, 93.696) *20.382*	0.609 (0.379, 0.949) *0.172*	0.608 (0.567, 0.678) *0.032*	0.178 (0.046, 0.381) *0.101*
Santiago C. (A Coruña)	- - -	69.330 (66.622, 73.068) *1.960*	0.475 (0.339, 0.604) *0.079*	0.560 (0.518, 0.596) *0.025*	- - -
Ribeira (A Coruña)	- - -	42.740 (40.575, 45.317) *1.483*	0.364 (0.234, 0.512) *0.084*	0.548 (0.495, 0.618) *0.036*	- - -
Lugo (Lugo)	- - -	128.155 (120.383, 142.000) *7.324*	0.885 (0.837, 0.977) *0.056*	0.522 (0.465, 0.558) *0.040*	- - -
Ourense (Ourense)	- - -	130.675 (125.268, 137.082) *3.658*	0.739 (0.606, 0.835) *0.067*	0.579 (0.547, 0.614) *0.020*	- - -
Cambados (Pontevedra)	0.355 (0.187, 0.815) *0.159*	40.997 (10.580, 54.021) *10.565*	0.873 (0.730, 0.973) *0.100*	0.570 (0.511, 0.736) *0.067*	0.566 (0.264, 0.856) *0.169*
Pontevedra (Pontevedra)	- - -	67.119 (64.403, 70.000) *1.685*	0.397 (0.291, 0.534) *0.075*	0.485 (0.447, 0.525) *0.024*	- - -
Vigo (Pontevedra)	0.237 (0.090, 0.554) *0.119*	232.486 (135.721, 276.663) *36.372*	0.263 (0.140, 0.428) *0.084*	0.749 (0.720, 0.781) *0.019*	0.201 (0.098, 0.349) *0.073*
Vilagarcia A. (Pontevedra)	- - -	36.718 (34.051, 39.020) *1.540*	0.472 (0.330, 0.602) *0.085*	0.509 (0.456, 0.566) *0.033*	- - -
Ponteareas (Pontevedra)	- - -	43.613 (39.908, 48.609) *2.582*	0.648 (0.478, 0.817) *0.102*	0.624 (0.565, 0.679) *0.037*	- - -

ML estimates for series with less than 100 counts.

Table 2. Maximum likelihood estimates for series without large counts, and 90% confidence intervals.

	α	λ	ω	q	p_{01}	AIC/BIC
Ordes	-	14.366	0.774	0.562	0.305	241.183
(A Coruña)	-	(11.716, 17.016)	(0.674, 0.874)	(0.457, 0.667)	(0.020, 0.590)	248.228
Padrón	-	13.830	0.573	0.531	-	255.993
(A Coruña)	-	(10.733, 16.926)	(0.240, 0.906)	(0.434, 0.629)	-	261.276
Negreira	-	9.707	0.637	0.505	-	228.144
(A Coruña)	-	(6.849, 12.565)	(0.272, 1.000)	(0.386, 0.624)	-	233.428
Carcubión	-	16.054	0.875	0.546	-	238.872
(A Coruña)	-	(10.593, 21.515)	(0.679, 1.000)	(0.385, 0.706)	-	244.156
Carballo	-	29.916	0.758	0.529	-	286.954
(A Coruña)	-	(25.830, 34.002)	(0.613, 0.903)	(0.460, 0.598)	-	292.238
Betanzos	-	35.957	0.689	0.566	0.356	322.630
(A Coruña)	-	(6.391, 65.523)	(0.454, 0.924)	(0.503, 0.628)	(0.074, 0.644)	329.675
Fonsagrada	-	2.835	0.705	0.208	0.276	136.386
(Lugo)	-	(1.776, 3.894)	(0.416, 0.994)	(0.095, 0.321)	(0.041, 0.511)	143.431
Chantada	-	13.406	0.859	0.394	0.478	228.231
(Lugo)	-	(9.876, 16.936)	(0.690, 1.000)	(0.300, 0.493)	(0.082, 0.874)	235.276
Mondoñedo	-	27.805	0.976	0.360	-	238.253
(Lugo)	-	(18.714, 36.896)	(0.938, 1.000)	(0.240, 0.482)	-	243.536
Monfor. L.	-	13.798	0.105	0.324	-	260.468
(Lugo)	-	(12.716, 14.880)	(0.005, 0.205)	(0.138, 0.509)	-	265.752
Viveiro	-	20.775	0.121	0.545	-	275.563
(Lugo)	-	(18.986, 22.563)	(-0.084, 0.327)	(0.282, 0.809)	-	280.847
Barco V.	-	16.917	0.665	0.466	0.232	249.175
(Ourense)	-	(14.715, 19.119)	(0.394, 0.936)	(0.392, 0.540)	(0.033, 0.431)	256.220
Verin	-	17.170	0.243	0.475	-	281.182
(Ourense)	-	(15.504, 18.837)	(0.058, 0.429)	(0.340, 0.610)	-	286.090
Puebla T.	-	6.953	0.852	0.298	-	177.652
(Ourense)	-	(4.462, 9.444)	(0.717, 0.988)	(0.190, 0.405)	-	182.935
Carballiño	0.396	5.974	0.078	0.263	-	242.095
(Ourense)	(0.228, 0.564)	(4.295, 7.653)	(-0.003, 0.159)	(0.079, 0.447)	-	249.140
Celanova	-	4.465	0.427	0.390	-	190.579
(Ourense)	-	(2.982, 5.948)	(-0.041, 0.895)	(0.154, 0.625)	-	195.862
Caldas R.	-	23.200	0.681	0.524	0.311	280.961
(Pontevedra)	-	(20.173, 26.226)	(0.433, 0.929)	(0.453, 0.595)	(0.038, 0.584)	288.006
Porriño	0.372	21.456	0.917	0.634	0.427	284.719
(Pontevedra)	(0.063, 0.681)	(10.363, 32.549)	(0.769, 1.000)	(0.521, 0.747)	(0.024, 0.830)	293.525
Redondela	0.611	7.749	0.559	0.497	0.271	273.958
(Pontevedra)	(0.419, 0.803)	(3.701, 11.800)	(0.309, 0.809)	(0.415, 0.579)	(0.093, 0.449)	282.764
Lalin	0.724	5.347	0.823	0.424	-	253.235
(Pontevedra)	(0.494, 0.954)	(0.625, 10.069)	(0.716, 0.930)	(0.355, 0.493)	-	260.280
Cangas M.	0.502	16.217	0.488	0.527	-	348.936
(Pontevedra)	(0.333, 0.671)	(10.625, 21.810)	(0.346, 0.616)	(0.478, 0.576)	-	355.981
Tui	0.487	11.960	0.766	0.405	0.346	284.252
(Pontevedra)	(0.214, 0.760)	(5.014, 18.906)	(0.555, 0.977)	(0.333, 0.477)	(0.106, 0.586)	293.059
A Estrada	-	11.085	0.213	0.402	-	244.782
(Pontevedra)	-	(9.851, 12.319)	(0.037, 0.388)	(0.232, 0.572)	-	250.065
Marin	-	23.893	0.783	0.594	-	292.212
(Pontevedra)	-	(14.940, 32.847)	(0.389, 1.000)	(0.452, 0.737)	-	297.495

Concluding remarks

This methodology can be used for detecting and quantifying the under-reporting in series of counts based on an INAR(1) model with Poisson distributed innovations and a latent under-reporting binary state that is a first order Markov chain.

We hope to extend it in several directions:

- by using different thinning operators in the latent process,
- by considering innovations with other distributions ,
- by using Conditional GLM models instead of thinning-based models,
- by considering general ARMA models for large counts,
- by modelling the under-reporting indicators, *i.e.* $\mathbf{1}_n$ with a more complicated dependence structure. (I-VAR?)
- $\vdots$