Estimación del infra-reporte en datos de recuentos. Aplicación a la CoVID-19.

Alejandra Cabaña

Universitat Autònoma de Barcelona

based on joint works with
Argimiro Arratia, Amanda Fernández-Fontelo,
David Moriña, Pere Puig

Under-reported data

Under-reporting in data refers to some issue, incident, phenomenon which is responsible to report 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.

Due to this phenomenon

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

Regarding public health, it is well known that some diseases have been traditionally under-reported.

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

Under-reported data issue

Work-related skin diseases in Norway may be underreported: data from 2000 to 2013

Jose H. Alfonso[1], Eva K. Løvseth[2], Yogindra Samant[3] and Jan-Ø. Holm[4]

[1]Department of Occupational Medicine and Epidemiology, National Institute of Occupational Health (STAMI), N-0033 Oslo, Norway, [2]Department of National Work Environment Surveillance, National Institute of Occupational Health, N-0033 Oslo, Norway, [3]Norwegian Labour Inspectorate, 7468, Trondheim, Norway, and [4]Department of Dermatology, Oslo University Hospital and Institute of Clinical Medicine, University of Oslo, 03180, Oslo, Norway

How Much Work-Related Injury and Illness is Missed By the Current National Surveillance System?

Kenneth D. Rosenman
Alice Kalush
Mary Jo Reilly
Joseph C. Gardiner
Mathew Reeves
Zhewui Luo

Reporting of Foodborne Illness by U.S. Consumers and Healthcare Professionals

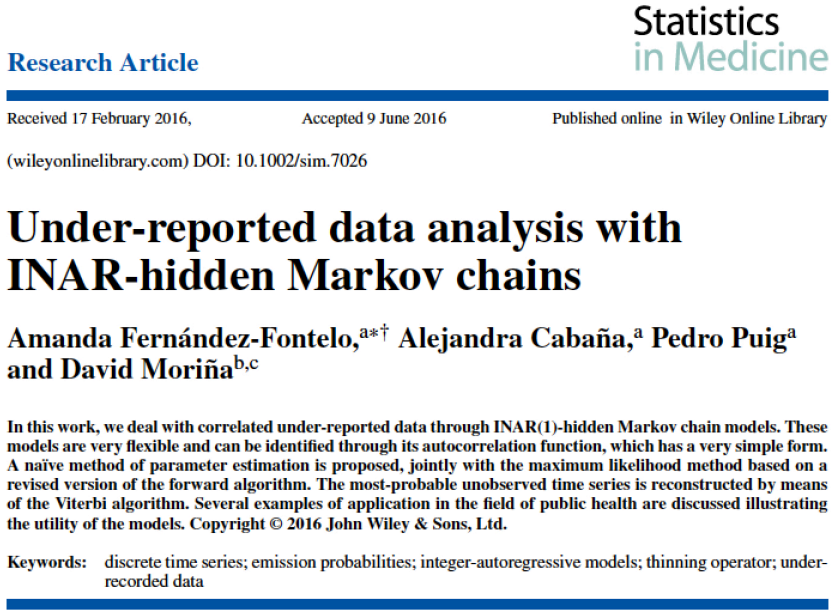
Susan Arendt, Lakshman Rajagopal, [...], and Steven Mandernach

Markov Chain Monte Carlo Analysis of Underreported Count Data With an Application to Worker Absenteeism

RAINER WINKELMANN[1]

Department of Economics, University of Canterbury, Christchurch, New Zealand

- 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

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

- 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 based on ML

The likelihood function of Y is directly intractable,

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

The **forward algorithm**[1] used in the context of HMC is a suitable option.

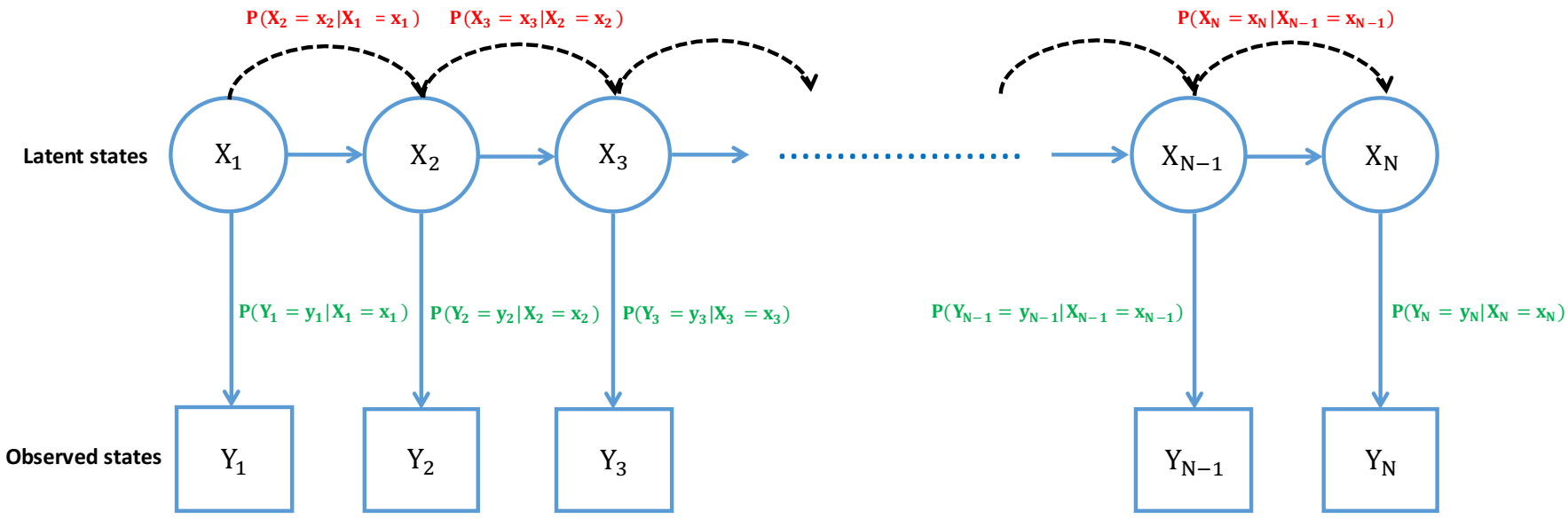
Consider the **forward probabilities**

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

with $\alpha_1(X_1) = P(X_1)P(Y_1|X_1)$. Then, the likelihood function is

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

[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.



$P(Y_k|X_k)$ and $P(X_k|X_{k-1})$ are the so-called emission and transition probabilities.

Parameter estimation based on ML (II)

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:

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

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** 260–269

An application

Modelling the number of weekly cases by human papillomavirus in Girona from 2010 to 2014:

- The human papillomavirus could 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.

ML estimation

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

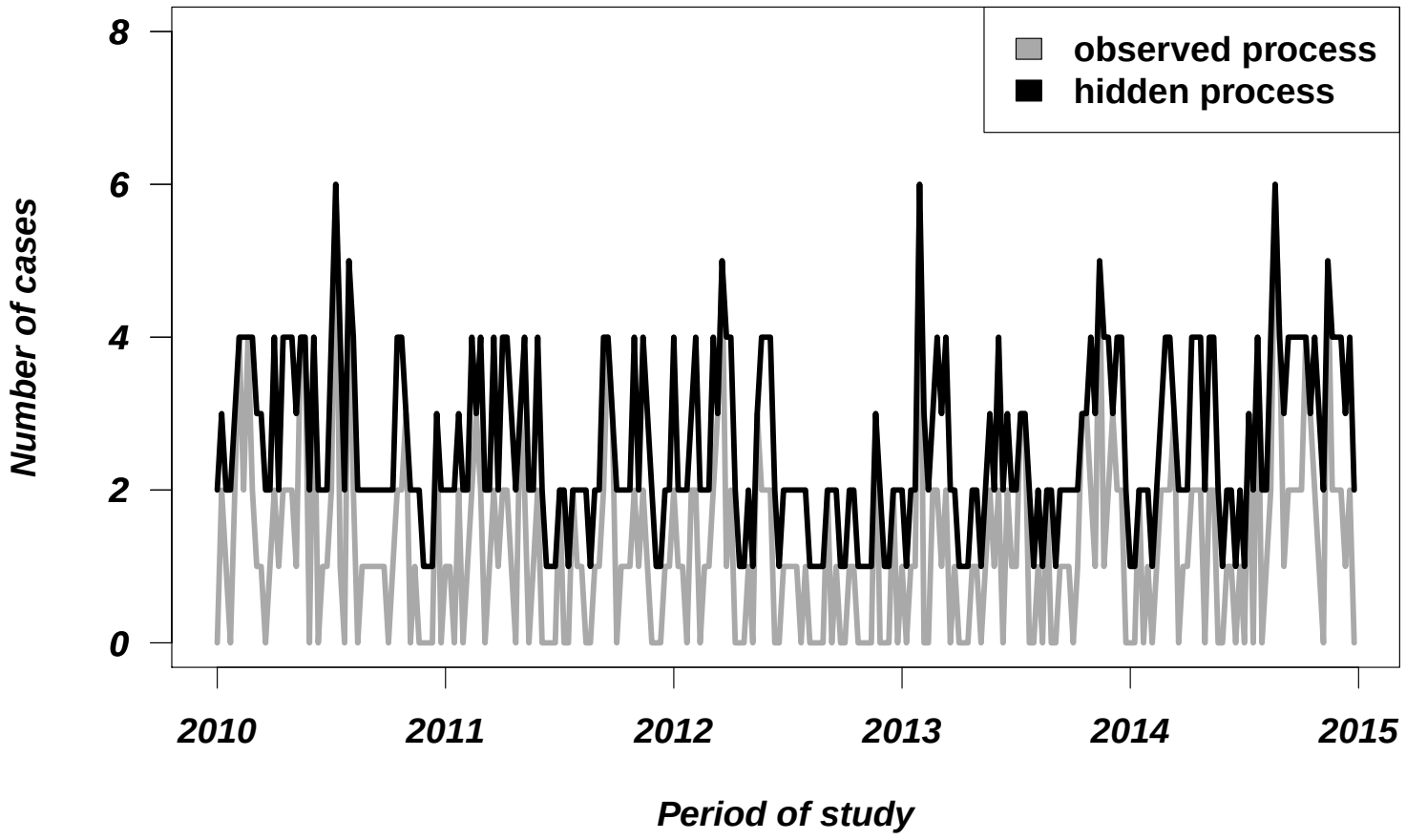
$$X_n = 0.517 \circ X_{n-1} + W_n \qquad W_n \sim \text{Poiss}\,(1.623)\,,$$

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,

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

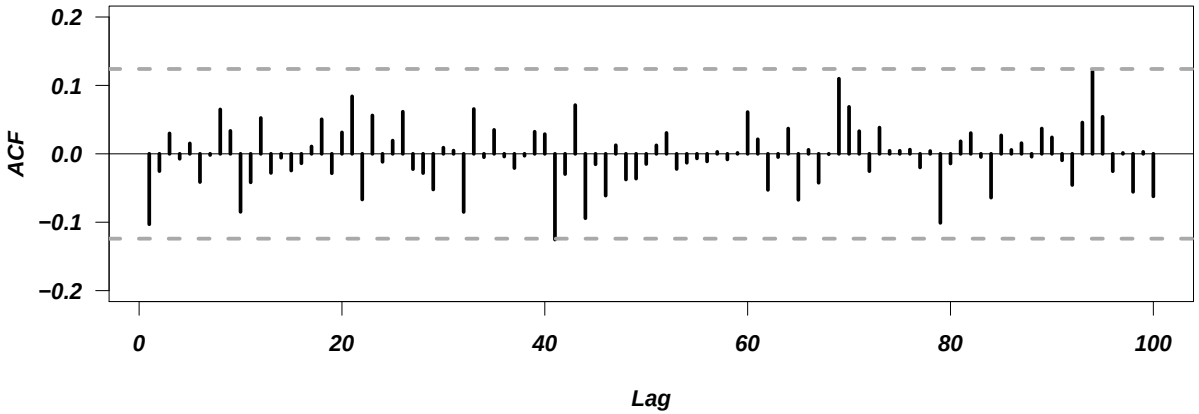
In order to use them in a plot, normally the mid-pseudo residuals

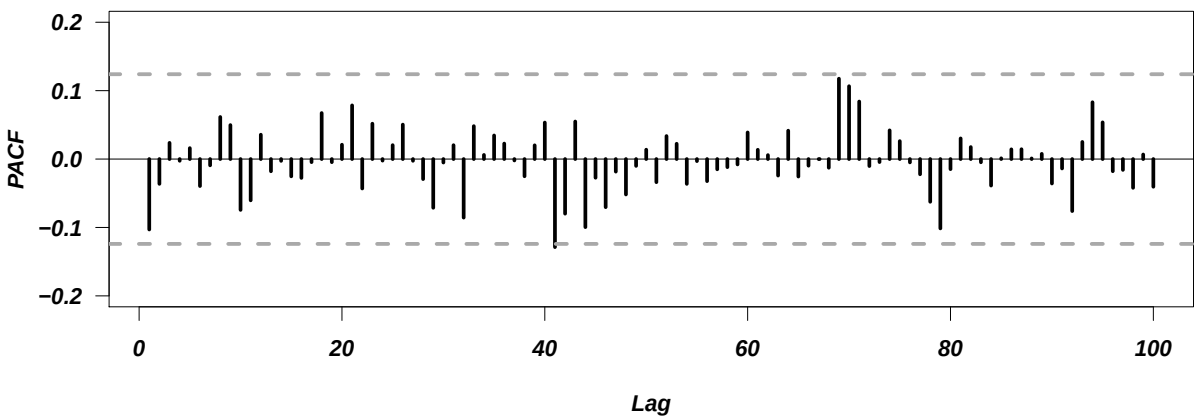
$$z_t^m = \phi^{-1}\left(\frac{u_t^- + u_t^+}{2}\right)$$

are considered.

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

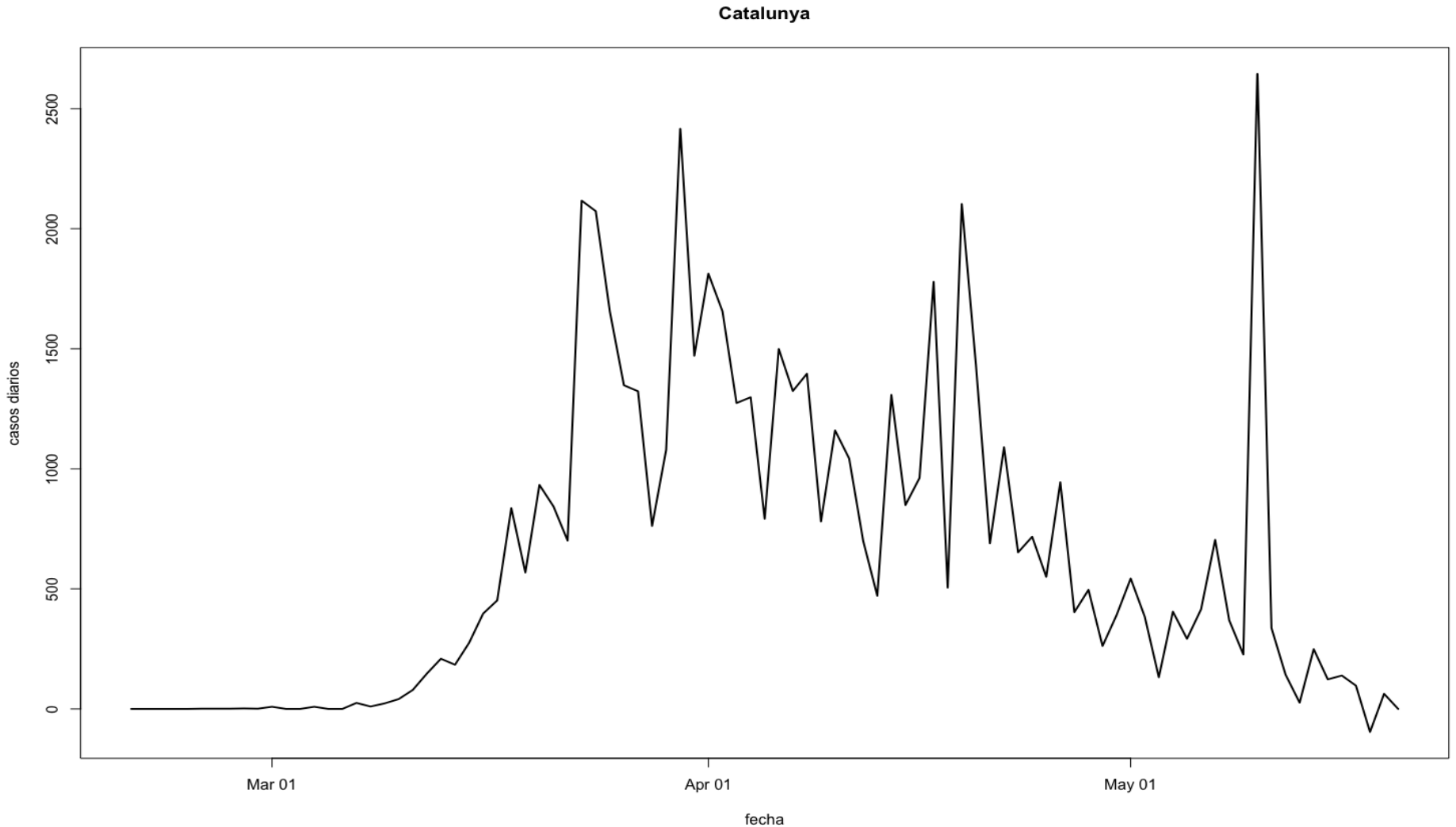
Model validation





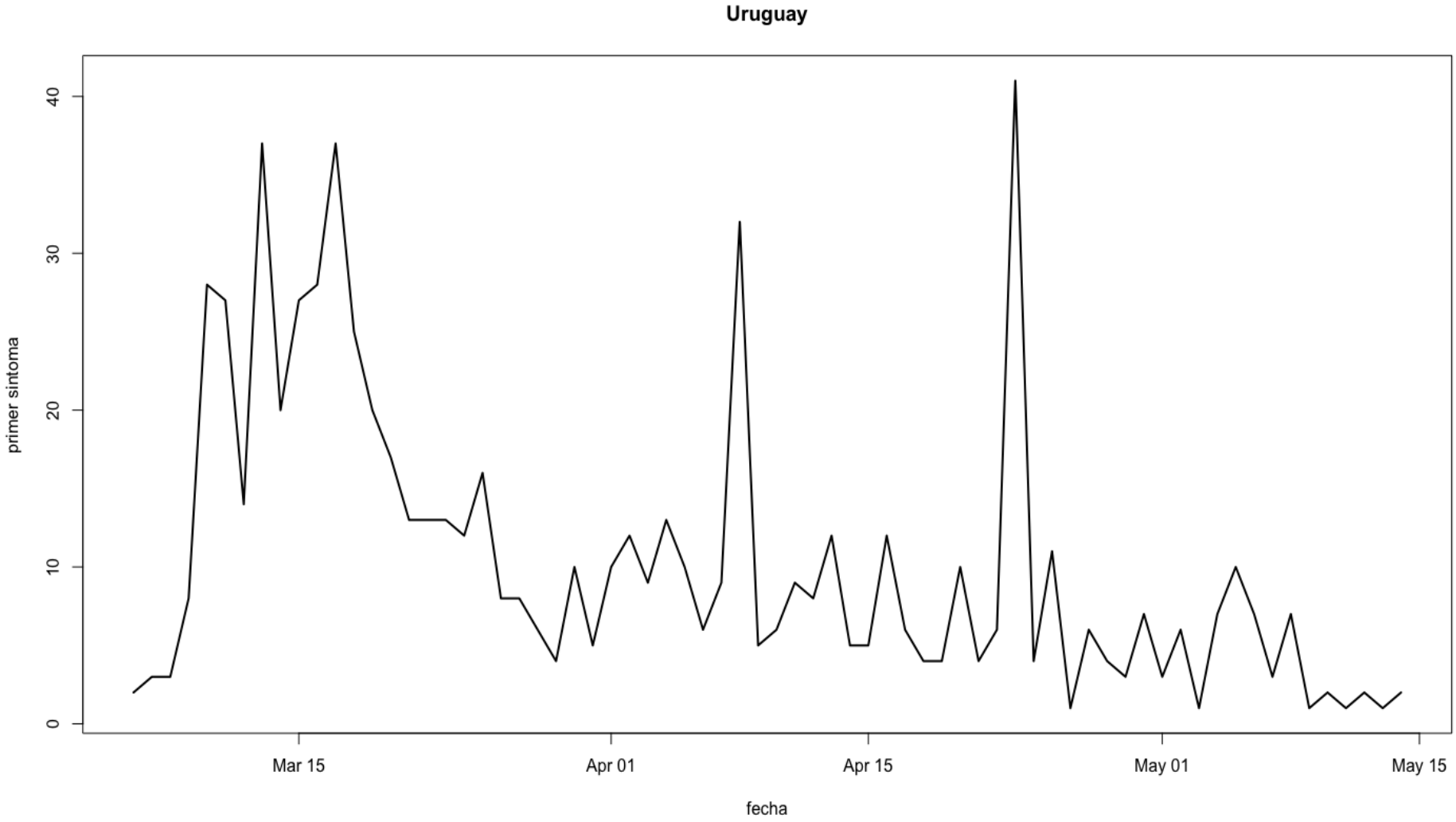
CoVID-19 data

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



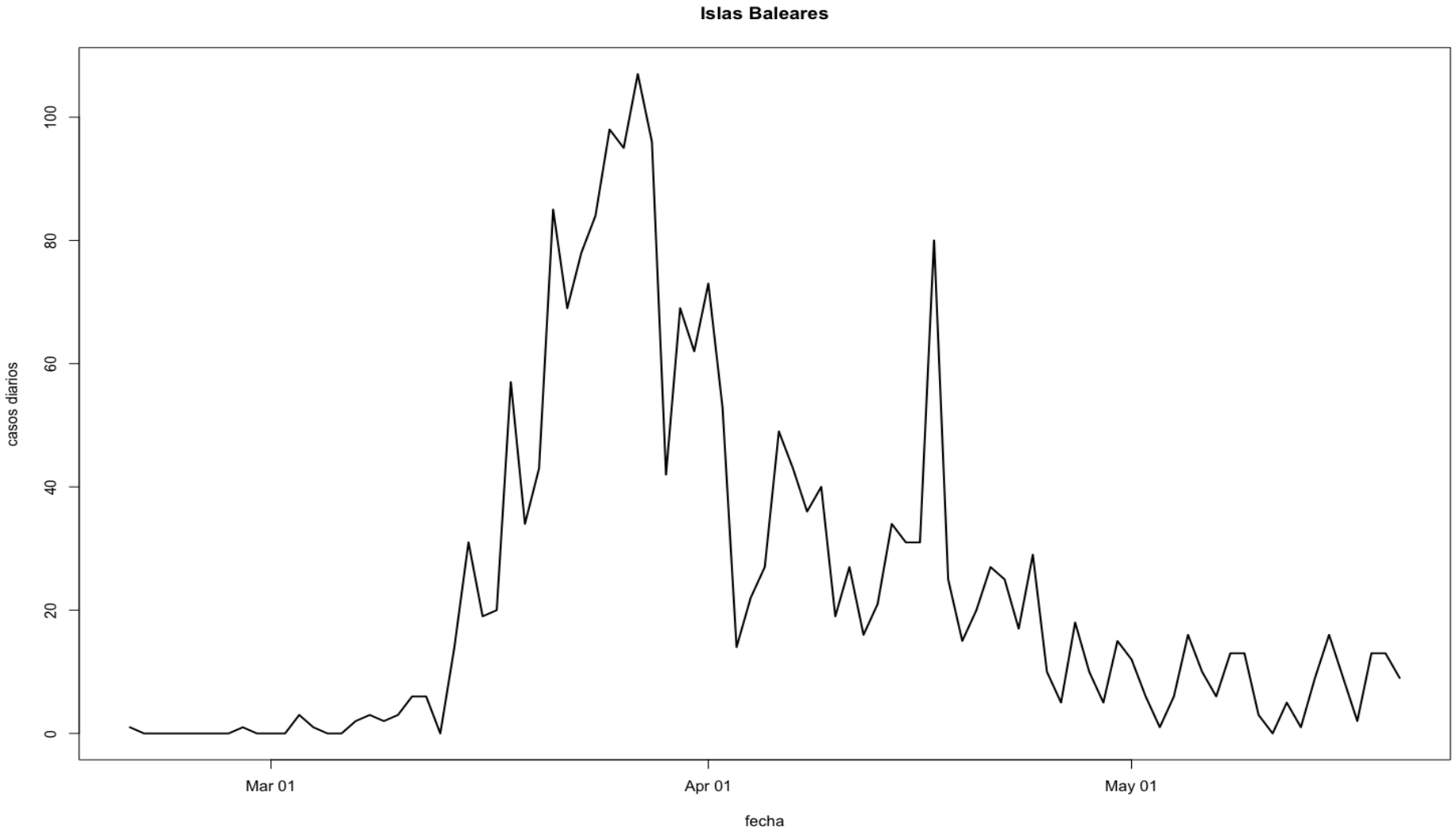
Data form `https://cnecovid.isciii.es/covid19/`.

CoVID-19 in Uruguay



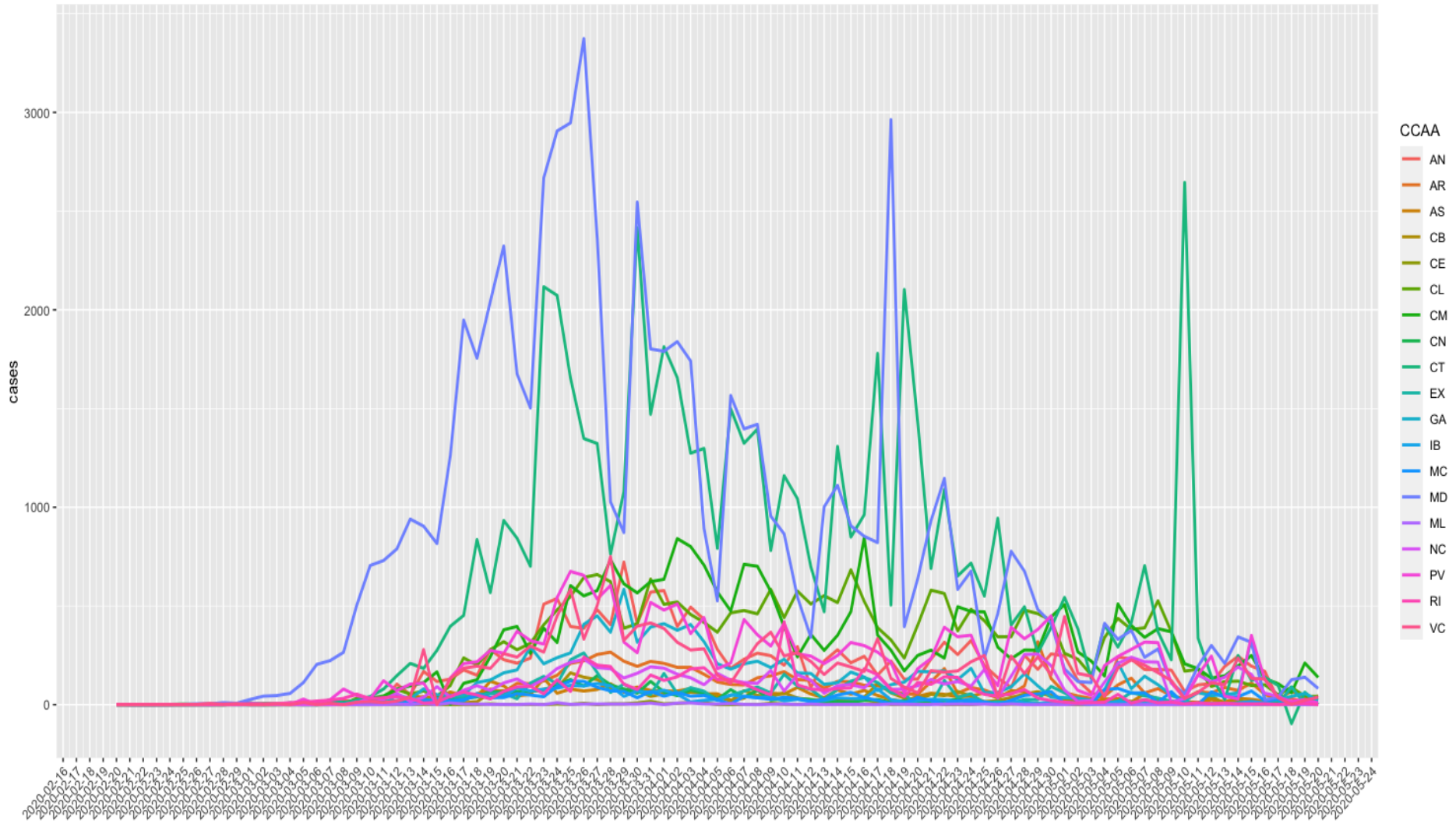
Data obtained from área de Epidemiología del MSP, corresponding to the date of 1st symptom.

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 removed 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

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

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

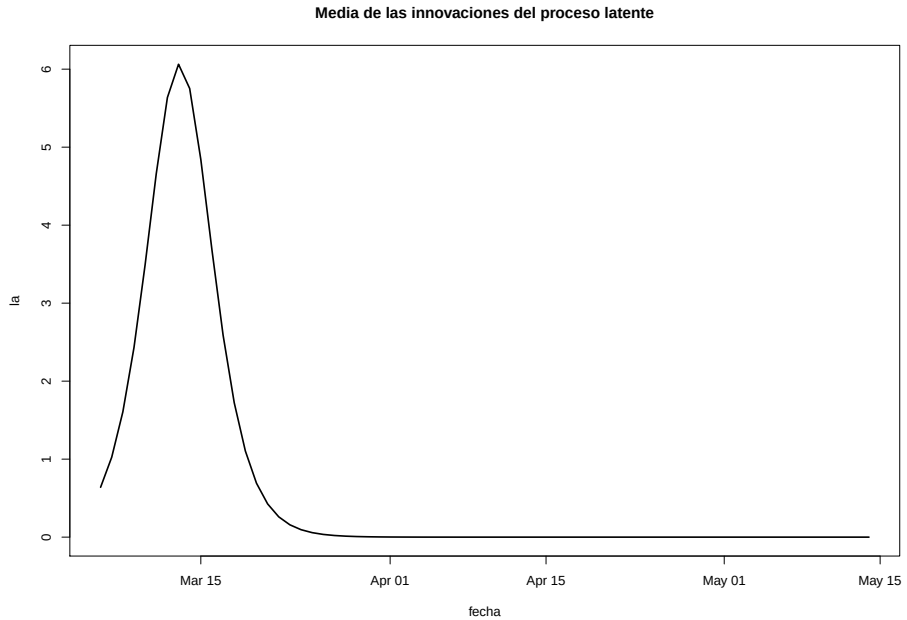
- α, $m = \log(M^*)$, k for the latent process
- ω and q (or q_n) for the under-reporting

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

Uruguay data

Using data from Área de Epidemiología del MSP, from March 6 2020 till May 14 2020 for the number of daily cases according to the date of apearence of the 1st symptoms,

$$X_n = 0.98 \circ X_{n-1} + W_n(\lambda_n) \qquad \text{where} \lambda_n \text{is,}$$



And the model for the observations

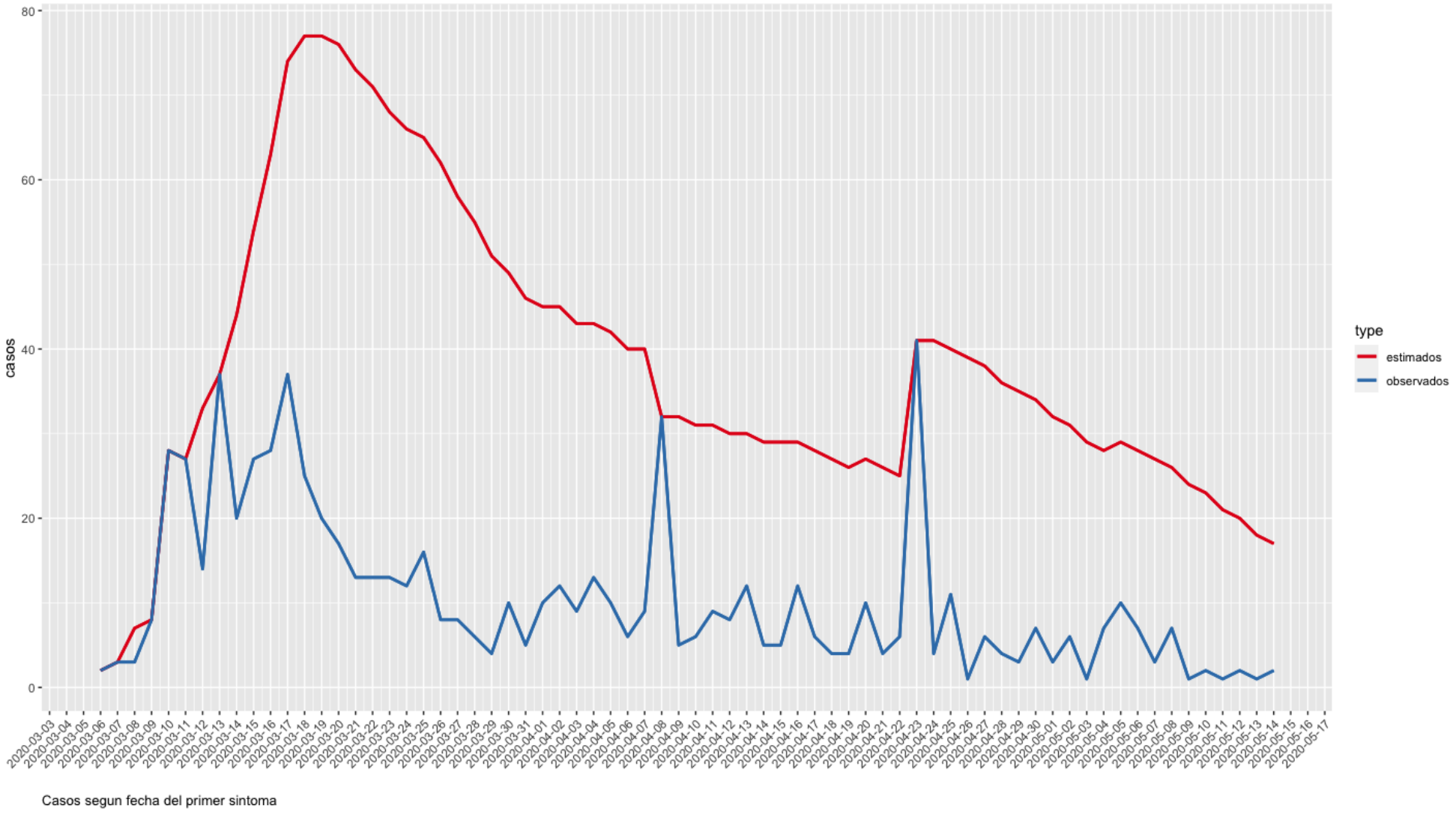
$$Y_n = \begin{cases} X_n & \text{with probability } 0.1 \\ 0.24 \circ X_n & \text{with probability } 0.9, \end{cases}$$

731 observed cases

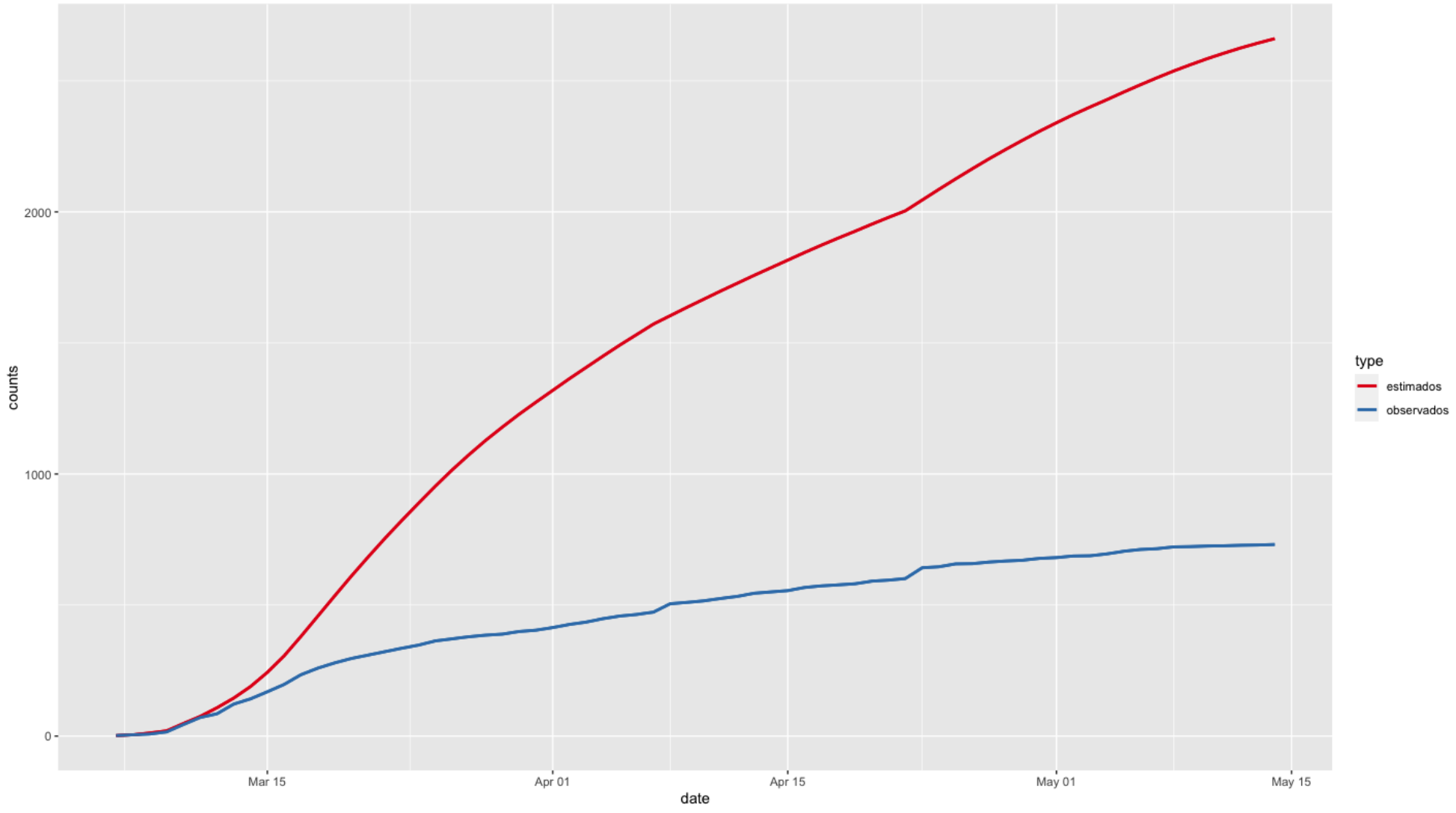
2661 estimated cases

that is, the system is overlooking 1930 individuals (264%), or in other words, assuming that the Viterbi estimation is the true number of infected individuals, **only 27.5% are reported**.

Daily number of cases



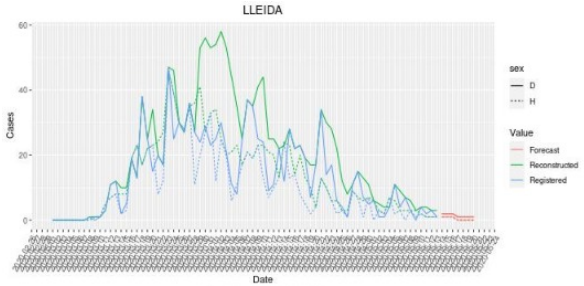
Acumulated number of cases

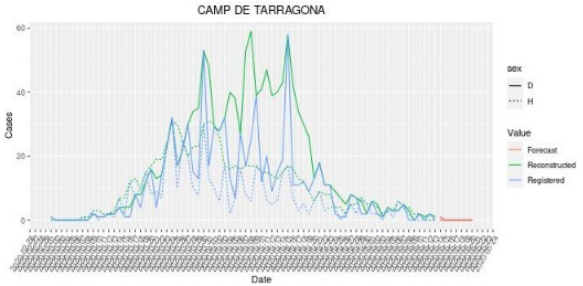


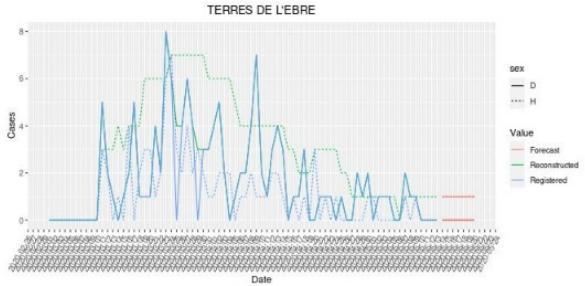
Cases in Catalunya by gender

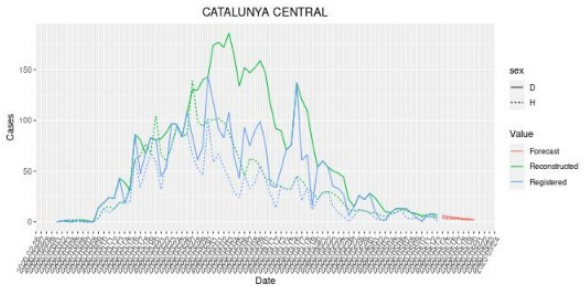
The model estimates that in the period 01/03/2020 to 13/05/2020, 92469 cases of COVID-19 have occurred in Catalunya, of which 59887 were registered . That is, 54.41% of cases (32582) would not have been registered in the system.

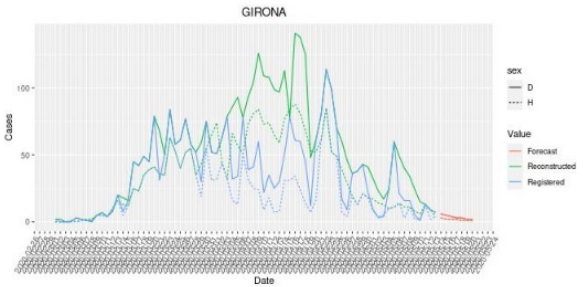
Cases in Catalunya by region

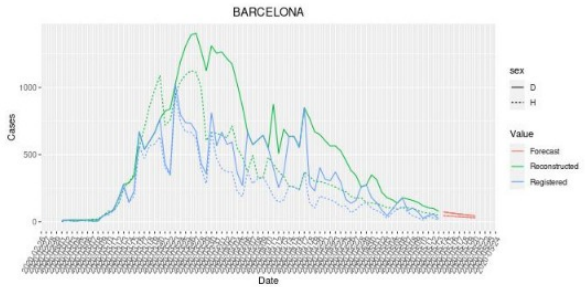


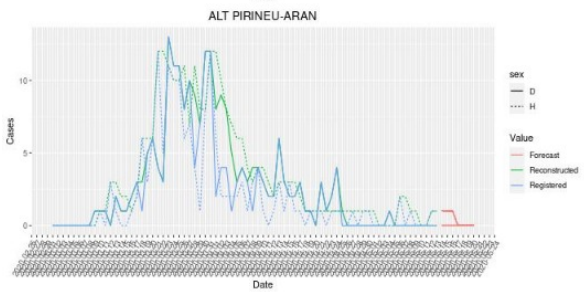




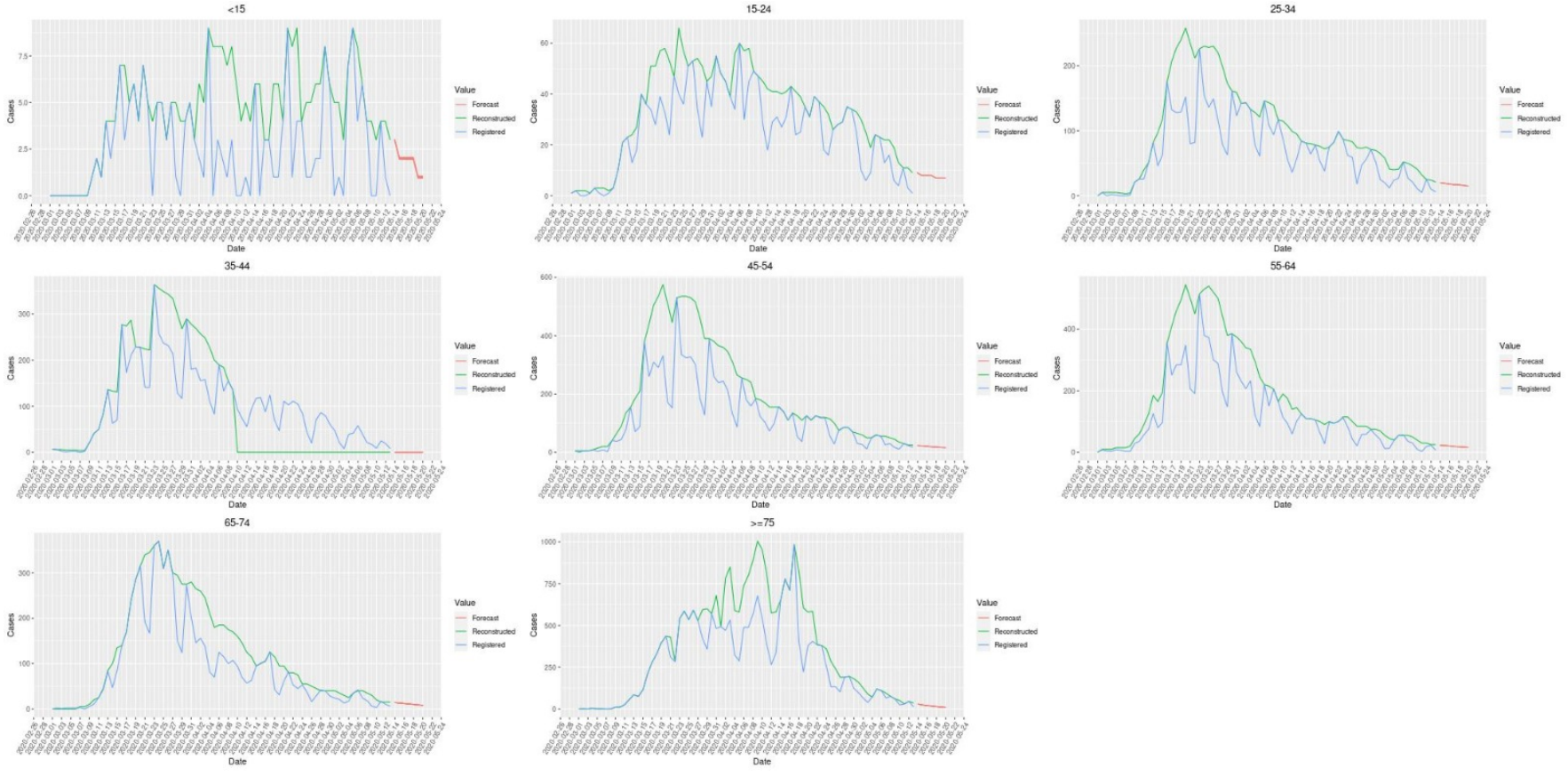








Cases in Catalunya by groups of age



Under-reported ARMA models

Consider now that the model for the unobservable process is an ARMA(p, r)

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_r \varepsilon_{t-r}$$

where ε is a Gaussian White Noise process and the observed process is again

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

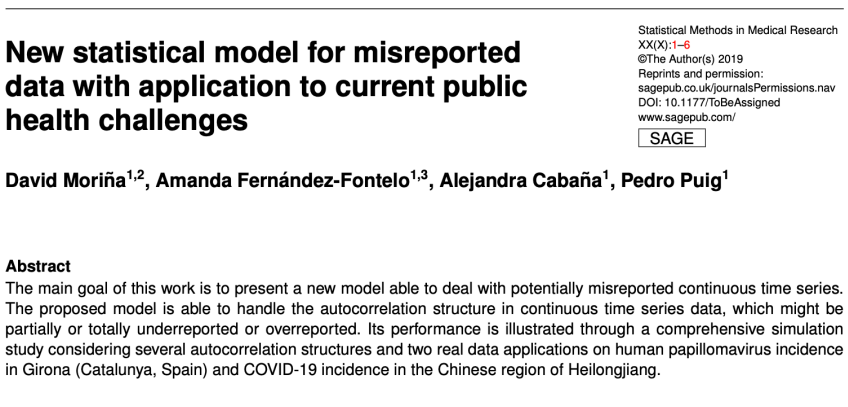
The moments of the observed process can be easily computed and the autocorrelations of the observed process are

$$\rho_Y(k) = c(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_r, \mu_\varepsilon, \sigma_\varepsilon^2, \omega, q)\rho_X(k).$$

- The likelihood of Y is not easily computable, but it can be estimated by means of an iterative algorithm based on its marginal distribution (a mixture of two normals).

- The reconstruction of the unobserved series is a by-product of the estimation procedure. We have used this methodology recently to estimate the incidence of HPV incidence in Girona, and of CoVID-19 in Heilongjiang (China).

And we are working in an extension, modelling the mean of the Gaussian innovations of the ARMA as a GAM.



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

Statistical Methods in Medical Research
XX(X):1–6

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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

Abstract
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 two real data applications on human papillomavirus incidence in Girona (Catalunya, Spain) and COVID-19 incidence in the Chinese region of Heilongjiang.

References

`https://underreported.cs.upc.edu/`

The team, working from confinement in late March 2020

acabana@mat.uab.cat , argimiro@lsi.upc.edu,
fernanda@hu-berlin.de, ppuig@mat.uab.cat,dmorina@mat.uab.cat