Research Article

Statistics in Medicine

Received 17 February 2016, Accepted 9 June 2016 Published online 11 July 2016 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

1. Introduction

The interest in the analysis of count time series has been growing in the past years, and many models have been considered in the literature [1] and applied to several fields like medicine [2] or wage loss claims [3]. The reason for this increasing popularity is, essentially, the limited performance of the classical continuous time series analysis approach when dealing with discrete valued time series. Several challenges appeared with the introduction of discrete time series analysis techniques (unobserved heterogeneity, periodicity, selective or under-reportation, …). Many efforts have been devoted to the introduction of seasonality in these models [4, 5] and coping with unobserved heterogeneity [6], but the problem of under-reporting or under-recording is still in a quite early stage of study. This phenomenon is very common in several contexts such as epidemiological and biomedical research, social research ... Under-reported data might lead to potentially biased inference, as it may invalidate the assumptions of standard models. For instance, [7] explores a Markov chain Monte Carlo-based methodology to study worker absenteeism where two sources of under-reporting are detected: an insufficient surveillance mechanism (if the data are provided by the employer), and a lack of memory if the time series is reconstructed retrospectively by the worker. Also, in public health context, it is well known that some diseases related to occupational or food exposures have been traditionally under-reported [8–10]. Of course in that case, there might be several sources of under-reporting, including accuracy of public health registries, political or economical interests, among others.

We shall be concerned with permanent under-reporting, assuming that there are more cases in the population than the ones known of so far. However, in some applications, under-reporting occurs by a delay in reporting, but as time passes, the counts became more and more complete, as studied in [11].

In the context of biomedical research, the under-reporting of time series data has been studied for instance for norovirus illness [12]. The authors used a negative binomial model without taking into

[a] *Departament de Matemàtiques, Universitat Autònoma de Barcelona, Bellaterra, Spain*

[b] *Unit of Infections and Cancer (UNIC), Cancer Epidemiology Research Program (CERP), Catalan Institute of Oncology (ICO)-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain*

[c] *Grups de Recerca d'Àfrica i Amèrica Llatines (GRAAL), Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra, Spain*

[*] *Correspondence to: Amanda Fernández-Fontelo, Department de Matemàtiques, Universitat Autonoma de Barcelona, Bellaterra, Spain.*

[†] *E-mail: amanda@mat.uab.cat*

account the correlation structure of the time series, perhaps a reasonable assumption for their data, although they claimed that 'the independence assumption of the residuals was slightly violated'.

The model that will be introduced in Section 2 considers two discrete time series: the observed series of counts Y_n, which may be potentially under-reported, and the underlying series X_n (unobserved), having an INAR(1) structure:

$$X_n = \alpha \circ X_{n-1} + W_n, \tag{1}$$

where α is a fixed parameter, $0 < \alpha < 1$ and W_n is assumed to be Poisson(λ) distributed. In addition, X_{n-1} and W_n are assumed to be independent at any time n. The $\circ$ operator in expression (1), called *binomial thinning* or *binomial subsampling*, is defined as follows:

$$\alpha \circ X_{n-1} = \sum_{i=1}^{X_{n-1}} Z_i, \tag{2}$$

where Z_i are independent and identically distributed Bernoulli random variables with probability of success equal to α. The parameter α can be interpreted as the proportion of those events happening at time $n-1$, which are also observed at time n. On the other hand, W_n can be understood as the *innovations* or new events produced at time n.

The integer-autoregressive (INAR) models were first introduced in [13], being a successful count time series model used in practice. There is a wide literature on INAR models [1–6], particularly the reviews [14] and [15] and the references therein.

The INAR(1) process is a homogeneous Markov chain with transition probabilities

$$P(X_n = i | X_{n-1} = j) = \sum_{m=0}^{\min\{i,j\}} \binom{j}{m} \alpha^m (1-\alpha)^{j-m} \cdot P(W_n = i - m), \tag{3}$$

which has a unique stationary solution under mild conditions on the innovations [16]. For Poisson innovations, the stationary marginal distribution is also Poisson distributed with population mean and variance [13],

$$\mu_X = \mathbf{E}(X_n) = \frac{\lambda}{1-\alpha} = \mathbf{Var}(X_n) = \sigma_X^2, \tag{4}$$

where λ is the mean and also the variance of the Poisson innovations. Here, the subscript X indicates a random variable following the corresponding marginal distribution.

Every stationary INAR(1) process is positively correlated because it has positive auto-covariance function given by $\gamma_X(k) = \mathbf{Cov}(X_n, X_{n+k}) = \alpha^{|k|}\sigma_X^2$, and hence, the auto-correlation function (ACF) is also positive

$$\rho_X(k) = \frac{\gamma(k)}{\gamma(0)} = \alpha^{|k|}. \tag{5}$$

The proposed model is described in detail in Section 2. In Section 3, different methods for estimating parameters are presented. In Section 4, we introduce the mid-pseudo-residuals as a method for model validation, and the reconstruction of the hidden process is dealt with the Viterbi algorithm. Some examples of traditionally under-reported data are discussed in Section 5. A simple example regarding the weekly number of cases of human papillomavirus (HPV) in Girona (northeastern Spain) from 2010 to 2014 is considered[‡]. The second example is based on the annual number of deaths attributable to a rare aggressive tumor (pleural and peritoneal mesotheliomas) in Great Britain from 1968 to 2013[§]. Mesothelioma is known to be closely related to occupational and environmental exposure to asbestos [17–19]. The third example is focused on modeling the annual number of botulism cases in Canada from 1970 to 2013 [¶]. In the last two examples, the population growth in the corresponding period of study is taken into account and used as a covariate. Finally, some conclusions and possible future lines of research are discussed in Section 6.

[‡] *Data collected from the open database of Generalitat de Catalunya,* http://www.ics.gencat.cat/sisap/diagnosticat/principal?patologia=Papil%B7loma™lang=ct.

[§] *Data collected from the Office for National Statistics (ONS),* http://www.ons.gov.uk/ons/index.html.

[¶] *Data collected from the Public Health Agency of Canada,* http://dsol-smed.phac-aspc.gc.ca/dsol-smed/ndis/charts.php?c=pl.

2. Model specification

Consider a hidden INAR(1) process X_n satisfying (1), and let Y_n be the observed counts. The way we allow Y_n to be under-reported is by defining the observed counts as

$$Y_n = \begin{cases} X_n & : \text{with probability } 1-\omega \\ q\circ X_n & : \text{with probability } \omega \end{cases} \tag{6}$$

Obviously, the definition (6) means that the observed Y_n coincides with the underlying series X_n (and therefore the observed count at time n is not under-reported) with probability $1-\omega$. Otherwise, Y_n is a *binomial thinning* of the hidden process X_n, representing an under-reported phenomenon coming from an underlying INAR(1) process. The parameter ω in (6) can be interpreted as the proportion of times that X_n is not completely observed (under-reported) while q quantifies the intensity of the under-reportation. The closer to zero is q, the more intense is the problem of under-reporting in the data. Note that when $q=1$ then $Y_n = X_n$ and parameter ω remains superfluous. On the other hand, when $\omega = 0$ again, $Y_n = X_n$, and q is superfluous.

For each n, we observe X_n with probability $1-\omega$, and a $q-$thinning of X_n with probability ω, independently of the past $\{X_j : j \leqslant n\}$. That is, we observe

$$Y_n = (1-\mathbf{1}_n)X_n + \mathbf{1}_n \sum_{j=1}^{X_n} \xi_j \qquad \text{where} \quad \mathbf{1}_n \sim \text{Bern}(\omega) \quad \text{and} \quad \xi_j \sim \text{Bern}(q). \tag{7}$$

Because of independence, the mean of Y_n is

$$\mathbf{E}\left(Y_n\right) = (1-\omega)\,\mathbf{E}\left(X_n\right) + \omega q\ \mathbf{E}\left(X_n\right) = \mu_X\,(1-\omega(1-q)) = \beta\mu_X \tag{8}$$

where $\beta = 1-\omega(1-q)$.

As for the variance of Y_n, observe that

$$Y_n^2 = \left(1-\mathbf{1}_n^2\right)X_n^2 + \mathbf{1}_n^2\left(q\circ X_n\right)^2 = \left(1-\mathbf{1}_n\right)X_n^2 + \mathbf{1}_n\left(q\circ X_n\right)^2$$

so that

$$\mathbf{E}\left(Y_n^2\right) = (1-\omega)\,\mathbf{E}\left(X_n^2\right) + \omega\mathbf{E}\left(q\circ X_n\right)^2 = (1-\omega)\,\mathbf{E}\left(X_n^2\right) + \omega\left(q^2\mathbf{E}\left(X_n^2\right) + q(1-q)\mathbf{E}\left(X_n\right)\right)$$

and hence

$$\begin{aligned}\mathbf{Var}\left(Y_n\right) &= \mathbf{E}\left(Y_n^2\right) - \left(\mathbf{E}\left(Y_n\right)\right)^2 = \left(1-\omega\left(1-q^2\right)\right)\left(\sigma_X^2+\mu_X^2\right) + \omega q(1-q)\mu_X - (1-\omega(1-q))^2\,\mu_X^2 \\ &= \mu_X^2\left(\omega\,(1-\omega)\,(1-q)^2\right) + \sigma_X^2\left(1-\omega(1-q^2)\right) + \mu_X\omega q(1-q)\end{aligned}$$

In the Poisson innovations case, with $\mu_X = \sigma_X^2 = \lambda/(1-\alpha)$,

$$\mathbf{Var}(Y_n) = \frac{\lambda^2}{(1-\alpha)^2}\omega(1-\omega)(1-q)^2 + \frac{\lambda}{(1-\alpha)}\,(1-\omega(1-q))$$

The following proposition shows that the auto-correlations (ACF) at lag k of the observed chain are a multiple of $\alpha^{|k|}$.

Proposition 1

The ACVF of the observed chain Y_n is given by

$$\rho_Y(k) = \frac{\gamma_Y(k)}{\gamma_Y(0)} = \frac{(1-\alpha)\,(1-\omega(1-q))^2}{(1-\alpha)\,(1-\omega(1-q)) + \lambda\left(\omega(1-\omega)(1-q)^2\right)}\alpha^{|k|} = c\,(\alpha,\lambda,\omega,q)\,\alpha^{|k|}. \tag{9}$$

The proof is a plain computation of covariances, and although not rather involved, it is deferred to Appendix A.

Note that the structure of the resulting ACF is very similar to that of the hidden process X_n but damped by $c(\alpha, \lambda, \omega, q)$ that is a constant with respect to the lag k. In fact, $\log(\rho_Y(k))$ is a linear function in k, while $\log(\rho_X(k))$ is also a linear function in k but without intercept. However, a remarkable difference is that the partial auto-correlation function (PACF) of Y_n is not equal to zero for lags $k > 1$.

3. Parameter estimation

3.1. A method based on the marginal distribution

According to (7), the observed process can be written as $Y_n = (1 - \mathbf{1}_n)X_n + \mathbf{1}_n q \circ X_n$. Like in general all the hidden Markov chain processes, this is not a Markovian process. The marginal distribution of X_n is Poisson$\left(\frac{\lambda}{1-\alpha}\right)$, and the q-thinning of a Poisson distribution is also Poisson distributed. Therefore, the marginal distribution of Y_n is a mixture of two Poisson distributions with parameters $\frac{\lambda}{1-\alpha}$ and $\frac{q\lambda}{1-\alpha}$, with probabilities $1 - \omega$ and ω. For $q = 0$, the maximum intensity of the under-reporting mechanism, the marginal distribution would be a zero-inflated Poisson. INAR(1) processes with zero-inflated Poisson innovations are introduced in [20] in order to analyze count time series with many zeros. As opposed to Y_n for $q = 0$, these models do not have zero-inflated Poisson marginals. Note that when $q = 0$, the observed process can be represented as $Y_n = Z_n X_n$, where Z_n are independent Bernoulli random variables with probability of success equal to $1 - \omega$, also independent of X_n. In [21], a similar model is introduced, where the hidden process X_n is not an INAR(1), but a different Poisson-based Markovian process called PPAR(1). Like in general all the hidden Markov chain processes, Y_n is not a Markovian process even in the simpler case of $q = 0$. Accordingly to the marginal distribution, we can obtain an estimation of $\frac{\lambda}{1-\alpha}$, $\frac{q\lambda}{1-\alpha}$ and ω by fitting the whole observed series using a mixture of two Poisson distributions. To do this, we have used in all examples the R package *mixtools* [22] based on the EM-algorithm. Then, from here, we obtain the estimates $\hat{\theta} = \hat{\lambda}/(1 - \hat{\alpha})$, $\hat{q}$ and $\hat{\omega}$.

To estimate α, and then also calculate an estimate of the other parameter of interest λ, several simple approaches can be considered:

(1) Direct calculations allow us to compute the constant term in (9) $\hat{c} = c(\hat{\alpha}, \hat{\lambda}, \hat{\omega}, \hat{q})$ from the estimates $\hat{\theta}$, $\hat{q}$, and $\hat{\omega}$. Then, our first estimator of α is $\hat{\alpha}_1 = \hat{\rho}_Y(1)/\hat{c}$, where $\hat{\rho}_Y(k)$ indicates the empirical autocorrelation coefficient of order k. This approach, as many moment-based estimators, can produce values out of the domain $[0, 1]$.

(2) Expression (9) also leads directly to another simple estimator using the two first autocorrelation coefficients: $\hat{\alpha}_2 = \hat{\rho}_Y(2)/\hat{\rho}_Y(1)$. It is specially useful when $\hat{\alpha}_1$ is out of the parameter domain.

(3) Note that expression (9) can be written as

$$\log(\rho_Y(k)) = \log(c(\alpha, \lambda, \omega, q)) + \mathrm{k}\log(\alpha). \tag{10}$$

Therefore, $\log(\alpha)$ can be estimated as the slope of the least squares regression line, fitting $\log(\hat{\rho}_Y(k))$ against k for the first few lags. This estimator is denoted as $\hat{\alpha}_3$.

All these estimators based on the marginal distribution are useful for exploratory analysis where fast methods are required. Moreover, they may also be used as initial values in the algorithms for numerical computation of the maximum likelihood estimators that will be presented in the next section. Parametric bootstrap could be used to provide standard errors of the estimators and to provide a set of initial values for the algorithms.

3.2. Maximum likelihood estimation

The parameters of interest can be estimated by maximum likelihood using hidden Markov chains methodology. If the process of the observed series is denoted as $Y = Y_{1:n} = (Y_1, Y_2, Y_3, \dots, Y_n)$ and the hidden INAR(1) process as $X = X_{1:n} = (X_1, X_2, X_3, \dots, X_n)$, the likelihood function can be written as

$$P(Y) = P(Y_1, Y_2, \dots, Y_n) = \sum_X P(X, Y) = \sum_x P(Y \mid X = x)P(X = x). \tag{11}$$

That is, the second term is the sum for each possible hidden chains X of the joint probability of Y and X. The explicit computation of this probability is directly intractable, as it happens in general for hidden Markov chains models. Hence, it is necessary to use an indirect method to compute the likelihood function (11). Our choice is a modification of the well-known forward algorithm [23].

The forward probabilities are defined as

$$\begin{aligned} \alpha_k(X_k) &= P(Y_1,\dots,Y_k,X_k) = P(Y_{1:k},X_k) = \sum_{X_{k-1}} P(Y_{1:k},X_k,X_{k-1}) \\ &= \sum_{X_{k-1}} P(Y_k \mid Y_{1:k-1},X_k,X_{k-1})P(Y_{1:k-1},X_k,X_{k-1}) \\ &= \sum_{X_{k-1}} P(Y_k \mid Y_{1:k-1},X_k,X_{k-1})P(X_k \mid X_{k-1},Y_{1:k-1})P(Y_{1:k-1},X_{k-1}) \\ &= \sum_{X_{k-1}} P(Y_k \mid Y_{1:k-1},X_k,X_{k-1})P(X_k \mid X_{k-1},Y_{1:k-1})\alpha_{k-1}(X_{k-1}) \\ &= P(Y_k \mid X_k) \sum_{X_{k-1}} P(X_k \mid X_{k-1})\alpha_{k-1}(X_{k-1}). \end{aligned} \tag{12}$$

Taking into account these probabilities, the expression (11) can be written as

$$P(Y) = P(Y_1,Y_2,\cdots,Y_n) = \sum_{X_n} \alpha_n(X_n), \tag{13}$$

where $\alpha_1(X_1) = P(Y_1,X_1) = P(X_1)P(Y_1 \mid X_1)$. Note that the sum in expression (13) has an infinite number of terms because of the existence of an infinite number of states. In practice, this can be solved by truncation, $P(Y) = \sum_{X_n=0}^{T} \alpha_n(X_n)$, considering a reasonable upper limit T that need to be set for each particular problem. A conservative criterium used and checked in our examples is to take T equal to two times the maximum of the observed series, that is, $T = 2\max\{Y_i\}$.

In our case, using (3), the transition probabilities $P(X_n \mid X_{n-1})$ take the form

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}) = e^{-\lambda} \sum_{j=0}^{min(x_n,x_{n-1})} \binom{x_{n-1}}{j} \alpha^j (1-\alpha)^{x_{n-1}-j} \frac{\lambda^{x_n-j}}{(x_n-j)!}. \tag{14}$$

Moreover, straightforward calculations show that the emission probabilities $P(Y_n \mid X_n)$ are given by

$$P(Y_i = j \mid X_i = k) = \begin{cases} 0 & \mathit{if} \;\; k < j \\ (1-\omega) + \omega q^k & \mathit{if} \;\; k = j \\ \omega\binom{k}{j} q^j (1-q)^{k-j} & \mathit{if} \;\; k \geqslant j, \end{cases} \tag{15}$$

Therefore, we have an efficient method to compute the likelihood function (13) using the recursive relation given in (12). Once we have implemented the likelihood function, it is maximized with respect to the parameters of interest through an iterative approach using the function `nlm` in R. The R script we used for this purpose is available in the Supporting Information.

When q is close to 1 or ω is close to 0, $Y_n \cong X_n$, under-reporting does not represent an issue, and a standard INAR(1) model could be appropriate to describe Y_n. In order to see what happens with the maximization algorithm in these frontier cases, we have simulated series of length 100, with $\omega = 0.01$ or $q = 0.98$, taking the other parameters arbitrary values. In both scenarios, the program was not able to reach the maximum of the likelihood function, failing the convergence of the algorithm.

The model can be extended to time-dependent parameters or to parameters depending on covariates. For instance, the parameters ω and q might be time-dependent if there were a pattern of trend and/or seasonality in the under-reporting phenomenon. In Section 5, we will explore a couple of examples where the parameter λ depends on a covariate. Once the parameters of the hidden process X_n are estimated, it is possible to make forecasting of future values using the methods for INAR models described in the literature, like those proposed in [3] and [4]. These predictions could be very useful from the epidemiological point of view and can be used by public health policy makers.

4. Goodness of fit and reconstruction of the hidden process

From a methodological point of view, model selection can be based on both statistical significance of the parameters and the Akaike information criterion (AIC). The goodness of fit of the selected model can be assessed by using the so-called normal pseudo-residuals, which are a particular case of Cox-Snell residuals [24]. In the discrete case, these residuals are computed from the estimated conditional distribution

$$\hat{P}\left(Y_n = y_n \mid \left(Y_1, Y_2, \dots, Y_{n-1}, Y_{n+1}, \dots, Y_T\right)\right) = \frac{\hat{P}\left(Y_n = y_n, Y_1, Y_2, \dots, Y_{n-1}, Y_{n+1}, \dots, Y_T\right)}{\hat{P}\left(Y_1, Y_2, \dots, Y_{n-1}, Y_{n+1}, \dots, Y_T\right)}, \tag{16}$$

which is a ratio of likelihoods. These likelihoods are computed using the forward algorithm described in (12), evaluating the likelihood function at the maximum likelihood estimators of the parameters. Then, the normal pseudo-residual segment $[z_n^-, z_n^+]$ can be obtained, where

$$z_n^- = \Phi^{-1}\left(\hat{P}(Y_n < y_n \mid (Y_1, Y_2, \dots, Y_{n-1}, Y_{n+1}, \dots, Y_T))\right) = \Phi^{-1}(u_n^-) \tag{17}$$

$$z_n^+ = \Phi^{-1}\left(\hat{P}(Y_n \leqslant y_n \mid (Y_1, Y_2, \dots, Y_{n-1}, Y_{n+1}, \dots, Y_T))\right) = \Phi^{-1}(u_n^+) \tag{18}$$

Note that when $y_n = 0$, the probability u_n^- in (17) is 0, and z_n^- is not defined. A solution, then, is to take a standard normal quantile corresponding to a probability close to 0, for instance, $z_n^- = -4$. If the fitted model is valid, the normal pseudo-residuals z_n^- and z_n^+ are approximately normally distributed. A compromise between both, the mid-pseudo-residuals, which are commonly used in practice [23], are defined as

$$z_n^m = \Phi^{-1}\left(\frac{u_n^- + u_n^+}{2}\right), \tag{19}$$

and they can themselves be used for checking a white noise behavior. The analysis of pseudo-residuals is useful for the assessment of the general fit of a selected model and the detection of outliers. Other approaches can be used for assessing the goodness of fit in this context, as uniform pseudo-residuals, although the described approach was preferred for coherence with classical time series analysis. A comprehensive compilation of these methods can be found in [23].

Once the model has been validated, the reconstruction of the hidden process X_n is very interesting because it shows a picture of the real phenomenon. It also let us identify where the under-reporting effect has been more serious, and, perhaps, it helps for identifying the causes. A simple idea to estimate the hidden process is to find the chain X^* that maximizes the likelihood of X_n given the observed series, assuming that the parameters of the model are known (those estimated by MLE). In other words, $X^* = \arg\max_X \hat{P}(X_{1:n} \mid Y_{1:n})$. Note that

$$\hat{P}(X_{1:n} \mid Y_{1:n}) = \frac{\hat{P}(X_{1:n}, Y_{1:n})}{\hat{P}(Y_{1:n})}. \tag{20}$$

Because the probability $\hat{P}(Y_{1:n})$ does not depend on X_n, it is enough to maximize $\hat{P}(X_{1:n}, Y_{1:n})$ with respect to $X_{1:n}$. The Viterbi algorithm [25, 26] can be used to maximize it, and then to reconstruct the most probable hidden chain X^*.

The Viterbi algorithm has been implemented in several R packages, for instance, in the `HMM` package [27] and in the `HiddenMarkov` package [28]. All these implementations are based on hidden Markov chains having a finite number of states. However, in the present setting, its implementation is a little bit more complex because our hidden chain has infinite states. In this case, we need to specify an upper limit T on the hidden chain in order to compute the transition and emission probabilities. A simple solution is to take $T = 2\max\{Y_i\}$. The upper limit T can also be chosen empirically for each example, starting with an initial value and increasing it by a specified amount on each iteration until there is no change in the output. An R script including the adapted Viterbi algorithm is available in the Supporting Information.

5. Examples

The proposed methodology is applied to three examples in the field of public health. The first example is based on the number of cases per week of HPV in Girona (a province of Catalonia, Spain) from 2010 to 2014. This example has been chosen because it is a stationary series, so it is a good way to illustrate the methodology in the simplest situation. The second and the third examples are based on the number of annual deaths by cancer of pleura and peritoneum (mesotheliomas) in Great Britain between 1968 and 2013 and the number of annual cases of botulism in Canada from 1970 to 2013, respectively. These other two examples have also been chosen because they are more complex than the previous, in the sense that they allow us to illustrate how the method can be used when the series presents a trend explainable by using a covariate.

5.1. *Weekly number of cases of human papillomavirus recorded in Girona*

Human papillomavirus is one of the most prevalent sexually transmitted infections. It is so common that nearly all sexually active men and women have it at some point in their lives, according to the information provided by the Centre of Control of Diseases in [29]. Generally, the infection disappears on its own without inducing any health problem, but in some cases, it can be related to several cancers (cervical, vulva, vagina, penis, anus, …), which usually are diagnosed many years after the infection. Accordingly, it seems reasonable to consider that the HPV could be severely under-reported because most of the sexual active people get it without symptoms and health problems.

Because of the stationarity of the series, the naïve approach shown in Section 3.1 may be useful to determine whether the series is under-reported or not by means of the estimation of the ACF.

The data were obtained from the open database of the *Generalitat de Catalunya* and represent the weekly number of cases of HPV recorded in Girona (Spain) from 2010 to 2014. The series consists in 260 observations, ranging from 0 to 6 cases, with a mean of 1.27 cases per week. The variance of the series is 1.60, slightly greater than the mean, giving a dispersion index of 1.26. Taking into account that $\hat{\rho}_Y(1) = 0.133$, the test described in [16] rejects an equidispersed Poisson INAR(1) process in favor of an overdispersed marginal distribution ($p = 0.0018$). This result is consistent because a mixture of two Poisson distributions is always overdispersed. Figure 1 shows how the series behaves during the period of study. At a first glance, the series shows no pattern of trend and/or seasonality; that is, the series seems to be stationary.

Figure 2 shows that Y_n could be compatible with an INAR(1) process. However, this simple graphical description is not enough, and Y_n has to be studied in more detail.

Therefore, the naïve approach described in Section 3.1 can be used to assess whether a stationary series is actually under-reported or not, at least in two different ways. One approach is to fit the observed series Y_n using a Poisson distribution and also a mixture of two Poisson distributions, and then comparing both models by means of their AIC values. A better performance of the Poisson mixture model should be interpreted as a sign of under-reporting. An alternative approach may be to fit a simple linear model like in (10), considering $\log(\hat{\rho}_Y(k))$ against k. If the regression line has a statistically significant intercept, the

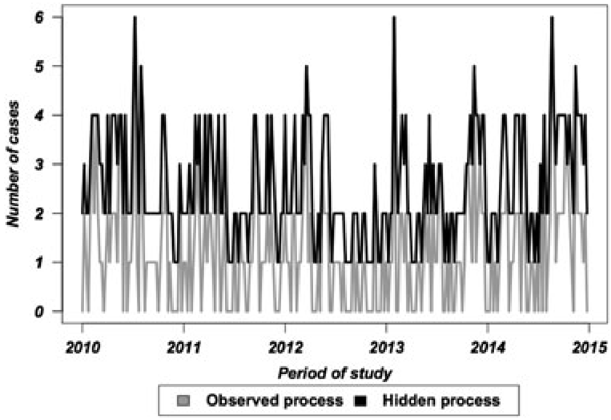

Figure 1. Observed process Y_n and estimated hidden process X_n of the number of weekly cases of HPV recorded in Girona (Spain) from 2010 to 2014.

series could be under-reported. For our data, fitting a mixture of two Poisson distributions, we obtained the following estimates:

$$\widehat{\frac{\lambda}{1-\alpha}} = 2.765, \ \widehat{\frac{q\lambda}{1-\alpha}} = 1.042 \text{ and } \hat{\omega} = 0.868, \tag{21}$$

and using these three values, we also obtained $\hat{q} = 0.377$. Taking into account the first empirical auto-correlation coefficients, according to Section 3.1, we estimated α and λ obtaining the following values: $\hat{\alpha}_1 = 0.370$, $\hat{\lambda}_1 = 1.744$, $\hat{\alpha}_2 = 0.504$, and $\hat{\lambda}_2 = 1.372$. We have also modelled Y_n by means of a unique Poisson distribution, obtaining that the AIC of the mixture of two Poisson distributions is smaller (783.843) than the AIC of a unique Poisson (787.075), so the mixture seems to fit the phenomenon better. We have also fitted (10) by ordinary least squares in order to analyze whether the intercept of the regression is statistically significant (p-value= 0.002). The p-values of the parameters of the regression line are reliable because the asymptotic variance of $\log(\hat{\rho}_Y(k))$ does not depend on the temporal lag k. In this sense, Figure 3 shows the predicted regression line being clear that it does not pass through the origin. Thus, both naïve methods suggest that the observed series might come from and under-reported INAR(1) model, or, at least, its behavior is compatible with this assumption. Table I shows the maximum likelihood estimators and their standard errors.

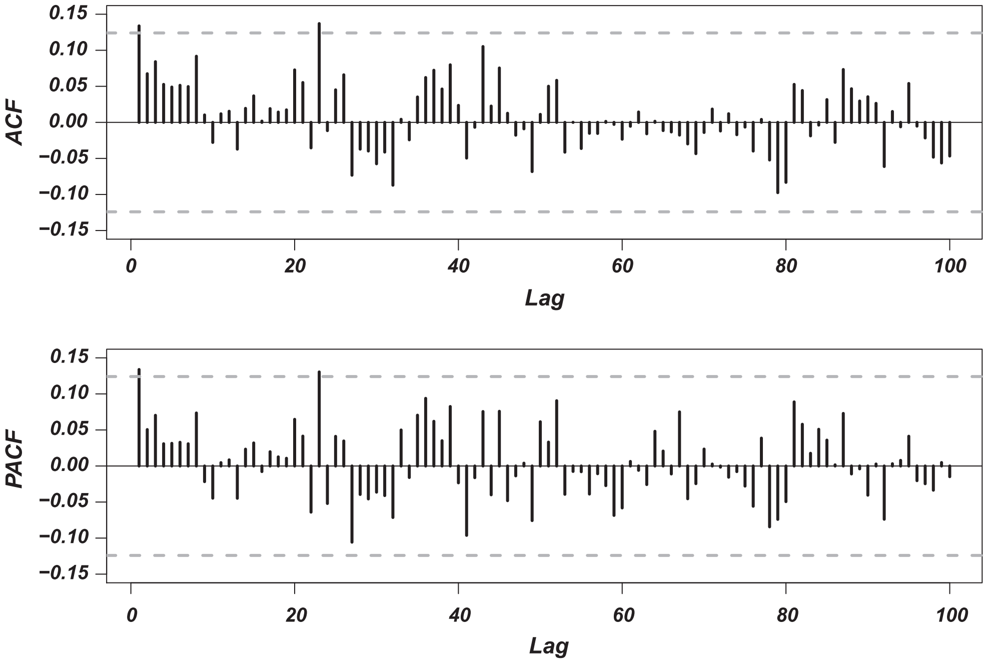

Figure 2. Auto-correlation function (ACF) and partial auto-correlation function (PACF) of the observed process.

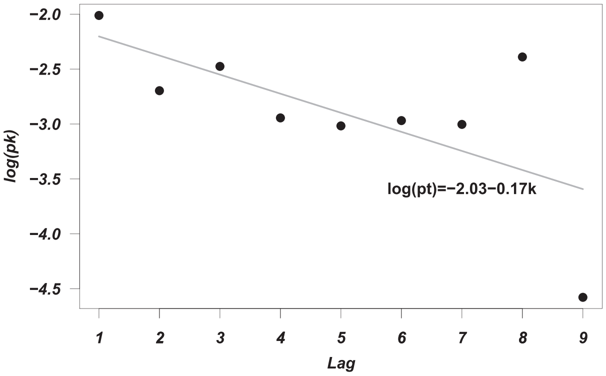

Figure 3. Observed values and predicted regression line of $\log(\rho_k) = \log(c(\alpha, \lambda, \omega, q)) + \text{k} \log(\alpha)$.

Table I. Maximum likelihood estimates (with their standard errors) of the proposed model for the number of weekly cases by HPV in Girona (Spain) from 2010 to 2014.

Parameter	ML estimate	SE
$\hat{\alpha}$	0.517	0.227
$\hat{\lambda}$	1.623	0.616
$\hat{\omega}$	0.922	0.073
$\hat{q}$	0.326	0.085

The parameters of the model have been also estimated by means of the maximum likelihood method (MLE) (Section 3.2). It could be computationally intensive depending on the length of the series and the magnitude of the observed values, but in this case, our program is quite fast. Also, the initial values for the program that maximizes the likelihood are important for its good convergence. Using the three naïve methods proposed in Section 3.1, this program converges to the same values of the parameters α, λ, ω, and q. Of course, the closer the initial values are to the MLE, the faster will be the convergence of the algorithm. In this example, the methods in which are computed $\hat{\alpha}_i$ and $\hat{\lambda}_i$ are equivalent because the estimates are quite similar. However, this conclusion may vary depending on the example we deal with. The estimated INAR(1) model for the hidden chain X_n was

$$X_n = 0.517 \circ X_{n-1} + W_n(1.623), \tag{22}$$

with the following specification for the observed chain Y_n,

$$Y_n = \begin{cases} X_n & : \text{ with probability } 0.078 \\ 0.326 \circ X_n & : \text{ with probability } 0.922. \end{cases} \tag{23}$$

The interpretation of this model is straightforward. Although the observed average is 1.27, the real (hidden) phenomenon has an estimated average of $\hat{\mu}_X = \hat{\lambda}/(1-\hat{\alpha}) = 3.36$ cases per week. Consequently, public health policy makers could learn that the registry of HPV cases might be improved because there might be only recording 38% (1.27/3.36) of the total HPV cases in Girona province. Both the frequency and intensity of the under-reporting process are statistically significant. Particularly, the estimated frequency of under-reporting ω is very high (close to 1), and a 95% confidence interval of the intensity q is $(0.159, 0.493)$. This high value for ω is reasonable given the nature of the data. Note that the value of the standard error of $\hat{\omega}$ would indicate that this is not significantly different from 1. Therefore, a simpler model setting $\omega = 1$, having the form $Y_n = q \circ X_n$, could be considered, assuming that all the observations are under-reported. However, when $\omega = 1$ the marginal distribution is Poisson distributed (not a mixture of Poissons), contradicting the detected overdispersion.

Once the model is fitted, the most likely sequence has been reconstructed using the Viterbi algorithm described in Section 4. The reconstructed chain of the weekly number of cases of HPV in Girona from 2010 to 2014 is shown in Figure 1 together with the observed process. Note that a first conclusion is that all the recorded zeros are in fact under-reported cases.

Finally, the goodness-of-fit of the model detailed in (22) and (23) is checked by means of the mid-pseudo-residuals (Section 4). Figure 4 shows the ACF and PACF of the mid-pseudo-residuals of the model showing a result compatible with a white noise as expected.

5.2. *Modeling with covariates*

5.2.1. Annual deaths by cancer of pleura and peritoneum (mesotheliomas). Asbestos is an environmental carcinogen, and asbestos-related diseases represent a global-scale environmental issue [30]. A particularly concerning asbestos-related disease is malignant mesothelioma, a rare and highly aggressive tumor for which the only known cause is the exposure to asbestos fibers [31]. This tumor can be placed in the mesothelium of several organs like the peritoneum, the pericardium or the tunica vaginalis, although the

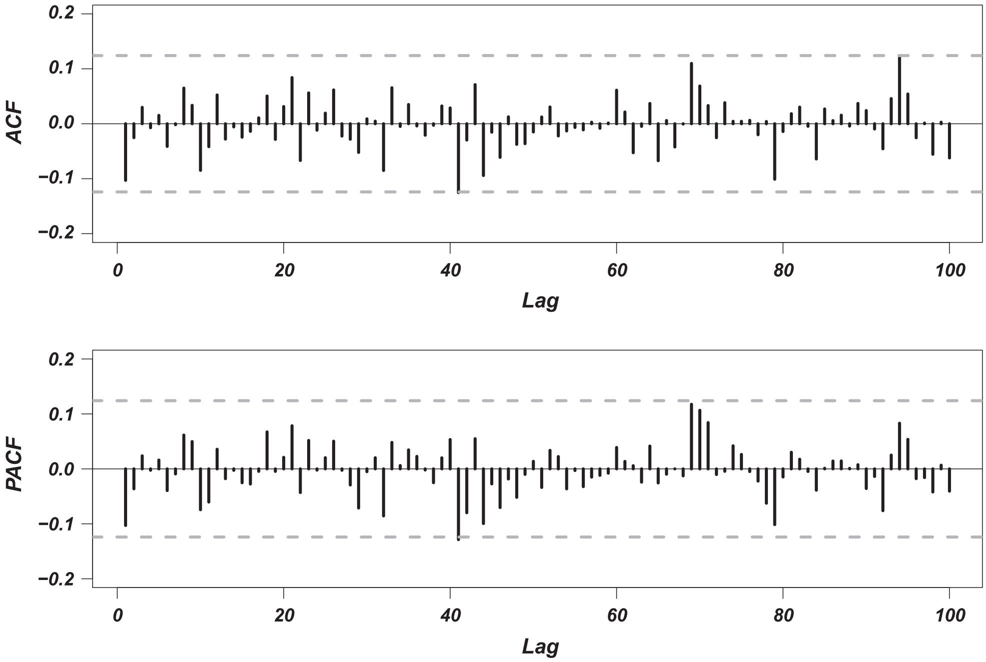

Figure 4. Auto-correlation function (ACF) and partial auto-correlation function (PACF) of the mid-pseudo-residuals of the selected model for the number of weekly cases of human papillomavirus in Girona 2010–2014.

most common is the pleural mesothelioma. It can also be spread to different organs (as in our example). This rare form of cancer is difficult to diagnose, and its prevalence is currently increasing. For instance, in [32], the authors predict that numbers of deaths among men due to meshotelioma in Spain will increase until 2016 because of the cohort effect. From an historical point of view, there are two important dates on the asbestos issue in the United Kingdom. First, in 1985, the United Kingdom banned the import and use of the most lethal forms of asbestos (this rule was replaced in 1992 with a law that also banned some uses of other forms of asbestos, considered less lethal). After that, in 1999, the import and use of all forms of asbestos were finally banned in the country. In the United Kingdom, all type of asbestos-related diseases are considered to be under-reported, specially several years ago, as the use of asbestos was permitted long after other countries banned the mineral's use. Nevertheless, several recent works have remarked that these diseases are still being under-reported in many countries [33]. The data used in this example were extracted from the website of the Office for National Statistics. It corresponds to the number of annual deaths by cancer of pleura and peritoneum (mesotheliomas) in Great Britain between 1968 and 2013. The range of this series is from 3 to 30 (Figure 5, with a mean of 14, a median of 15, and a variance of 34, so a moderate overdispersion is detected. In this regard, Figure 5 shows the evolution of the series, suggesting an increasing trend probably related to the growth of the population during the period of study.

The trend could be implemented in the model by means of a covariate that might explain its behavior (i.e., a simple increasing time-covariate, the total annual population, …). Moreover, we have to decide which parameters could be affected by the covariate and how. Expression (8) shows that in the stationary case the mean of the marginal distribution of the observed process is proportional to λ. Therefore, a natural choice is to implement the trend into parameter λ expressing this as a linear function of the covariate. However, other options would be possible, like to implement the covariate into the under-reporting parameters ω and q. The decision can be made comparing the values of the AIC of several competitor models, choosing the one in which the AIC is the lowest. Another approach would be to model flexible trends using P-splines techniques [34].

In our case, the most suitable model able to describe the time series is that considering $\lambda_n = aN_n$, where N_n indicates the total population at time n. The fitted model is,

$$X_n = 0.873 \circ X_{n-1} + W_n \left(0.053 \frac{N_n}{1000000} \right), \tag{24}$$

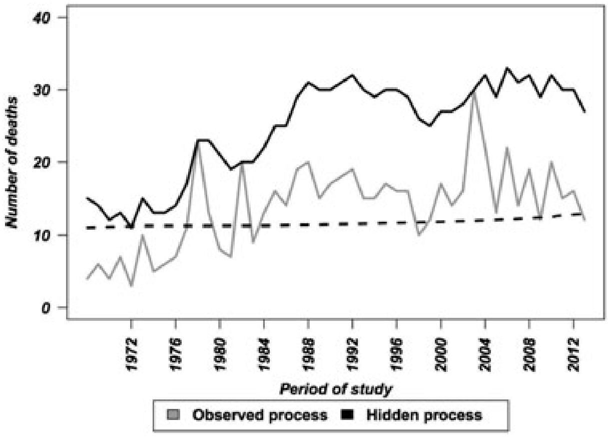

Figure 5. Observed and estimated hidden processes for the number of annual deaths by pleural and peritoneal mesotheliomas in Great Britain from 1968 to 2013. The dotted line represents the yearly total population (in fives of millions of habitants).

Table II. Maximum likelihood estimates (and their standard errors) of the model proposed for the number of annual deaths by pleural and peritoneal mesotheliomas in Great Britain from 1968 to 2013.

Parameter	ML estimate	SE
$\hat{\alpha}$	0.873	0.050
$\hat{a}$	0.053	0.022
$\hat{\omega}$	0.930	0.040
$\hat{q}$	0.517	0.036

with

$$Y_n = \begin{cases} X_n & : \text{with probability } 0.070 \\ 0.517 \circ X_n & : \text{with probability } 0.930. \end{cases} \tag{25}$$

Table II shows the maximum likelihood estimators and their standard errors. Notice that all the parameters of the model are statistically significant, with an AIC statistic equal to 277.18. The interpretation of this model is slightly more complicated than in the previous example. Note that now the average of the innovations of the hidden-INAR(1) process is a linear function of the total population. The estimated frequency of under-reporting is high ($\hat{\omega} = 0.930$), while its intensity q is moderate, providing a 95% confidence level of $(0.446, 0.588)$. The high value of $\hat{\omega}$ and its standard error would indicate that a simpler model of the form $Y_n = q \circ X_n$ could be explored (setting $\omega = 1$). However, to assume that all the observations are under-reported is a hard decision to make on the basis of a short series of only 46 observations.

Lastly, Figure 5 also shows the reconstructed hidden chain of the series and Figure 6 shows the ACF and PACF of the mid-pseudo-residuals of the model in order to evaluate the goodness of fit of the selected model. The chosen model seems to be suitable because their residuals are like white noise.

5.2.2. Annual number of botulism cases. Botulism is a rare but extremely fatal illness induced by a toxin produced by the *clostridium botulinum* bacteria, and it causes progressive flaccid paralysis in humans and animals. Foodborne botulism occurs when the toxin is located in some particular aliments like yogurt, garlic oil, and foil-wrapped baked potatoes. Wound botulism starts when an anaerobic wound is contaminated with the toxin and the organism grows and it is able to produce the toxin. It is typical of inject drug abusers, as it may be caused by contaminated needles or drugs. Infant botulism appears in less than

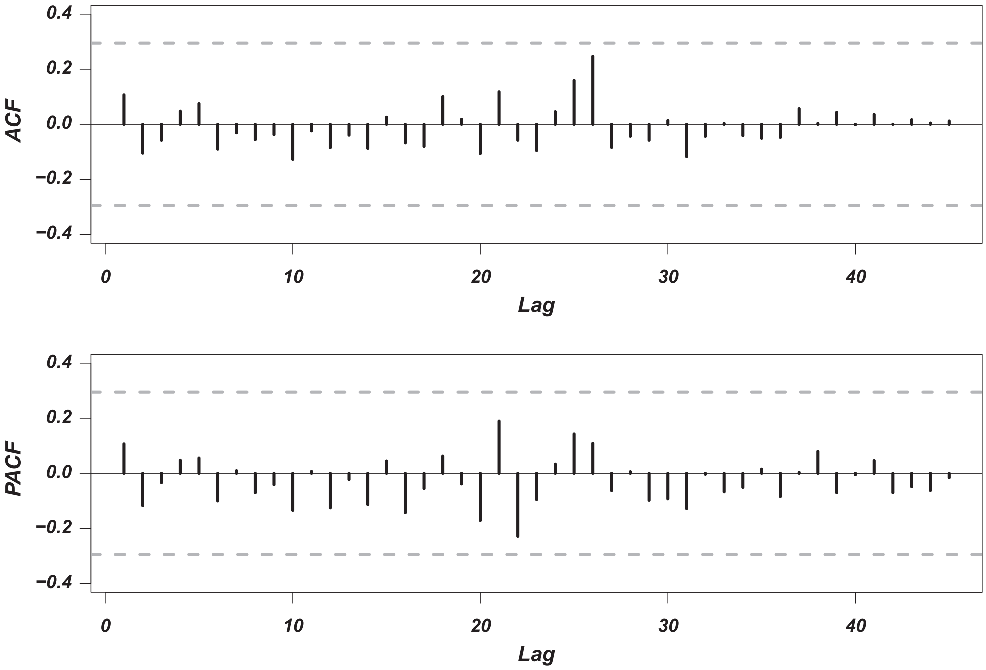

Figure 6. Auto-correlation function (ACF) and partial auto-correlation function (PACF) of the mid-pseudo-residuals of the selected model for the number of annual deaths by pleural and peritoneal mesotheliomas in Great Britain 1968–2013.

a year children because the spores germinate their intestinal tract and produce toxin. The disease is more common in children because of the immaturity of their intestines. Both foodborne and infant botulism are documented as an under-reported form of that disease. In particular, the World Health Organization has been repeatedly remarking that the foodborne illnesses are highly under-reported in several countries [35]. For instance, in [36], the author highlighted that the under-reporting phenomenon in foodborne illnesses is a serious problem in Canada because the patterns of these illnesses are changing and it makes the recognition of the diseases more complicated. In the case of botulism in children, in [37], the authors concluded that several countries have not reported yet their cases of infant botulism making this illness under-reported around the world. This data set was obtained from the Public Health Agency of Canada. It includes the number of botulism cases recorded annually from 1970 to 2013 in Canada. This series ranges from 0 to 36 cases, with a median of 7 cases per year, an average of 9 cases per year and a variance of 44.7 (the series is overdispersed as the two previous examples). At a first glance, Figure 7 does not show any pattern of trend and seasonality. However, during the period of study, the population has been growing significantly. Given the nature of the data, it seems reasonable to think that an increment in the population will produce an increment of the number of botulism cases, so that the yearly population has been included as a covariate.

Several INAR(1) models for the hidden process X_n were analyzed using different functions for the innovation parameter λ and for the under-reporting parameters q and ω. After several explorations, according to the AIC selection criteria, the best model among all was the one in which α, ω, and q were constants and $\lambda_n = aN_n$, with a a constant and N_n the yearly total population (AIC=296.72). The estimated model is,

$$X_n = 0.592 \circ X_{n-1} + W_n\left(2.475\frac{N_n}{10000000}\right), \tag{26}$$

that leads to

$$Y_n = \begin{cases} X_n & : \text{with probability } 0.329 \\ 0.317 \circ X_n & : \text{with probability } 0.671. \end{cases} \tag{27}$$

Table III shows the maximum likelihood estimates of the parameters of the model and their standard errors. Again, all parameters are statistically significant.

Table III. Maximum likelihood estimates (and their standard errors) of the proposed model for the yearly number of cases of botulism in Canada from 1970 to 2013.

Parameter	ML estimate	SE
$\hat{\alpha}$	0.592	0.117
$\hat{a}$	2.475	0.686
$\hat{\omega}$	0.671	0.094
$\hat{q}$	0.317	0.038

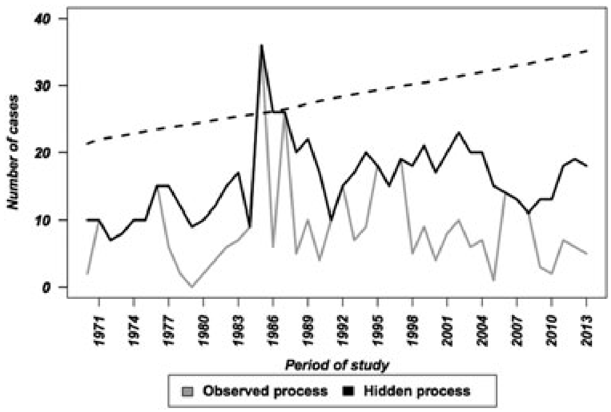

Figure 7. Observed and hidden process of the number of annual cases of botulism in Canada from 1970 to 2013. The dotted line represents the yearly total population (in tens of millions of habitants).

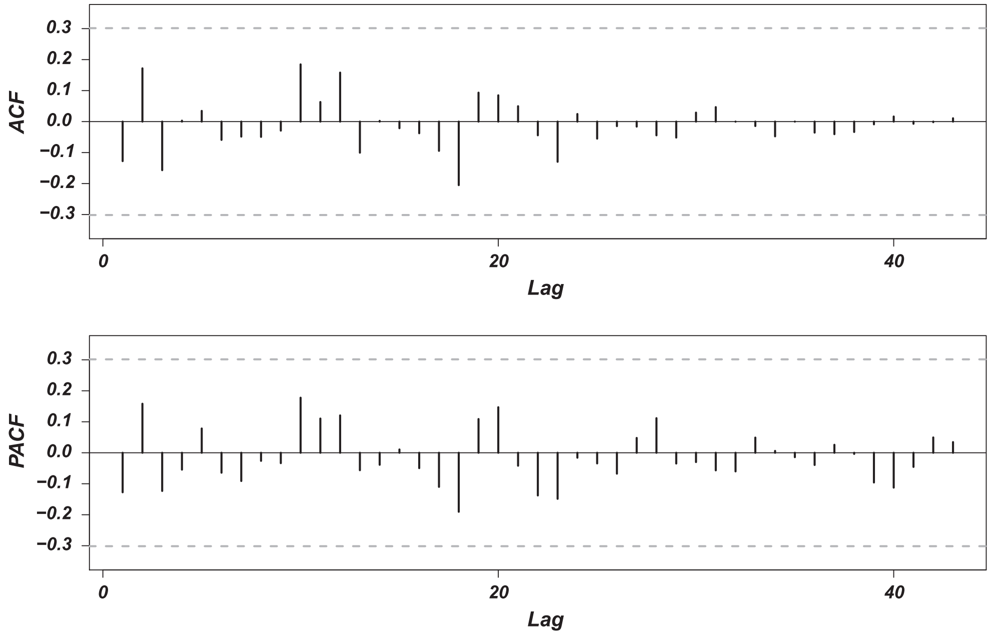

Figure 8. Auto-correlation function (ACF) and partial auto-correlation function (PACF) of the mid-pseudo-residuals of the selected model for the number of annual cases of Botulism in Canada 1970–2013.

Note that in this example the estimated under-reporting frequency is moderate ($\hat{\omega} = 0.671$) with a 95% confidence interval of $(0.487, 0.855)$. Otherwise, the estimated intensity is higher than in the previous example (closer to zero), providing a a 95% confidence interval of $(0.243, 0.391)$.

Figure 7 shows the estimated hidden chain jointly with the observed series. The proportion of times that both series take the same values is approximately $1 - \omega$. The plots of the ACF and PACF of the mid-pseudo-residuals in Figure 8 agree with a white noise behavior showing goodness of fit of the model.

6. Discussion and further work

Dealing with under-reported data is a quite common problem in public health practice. The methodology introduced in this work allows the incorporation of under-reportation in time series data in a very flexible way, following a different approach that introduced in [7] for a socioeconomic Poisson panel data. The INAR structure for the hidden process X_n allows a very natural interpretation of all the parameters [1, 13, 38], and it is general enough to suite a wide range of real phenomena. On the one hand, our methodology allows the detection of the phenomenon of under-reporting on a time series, estimating its frequency and intensity. On the other hand, we are able to reconstruct the most likely non-observed process by means of the Viterbi algorithm. In addition, the examples discussed in Section 5 show that these models, designed to deal with stationary time series, can also be extended to other situations with possible trends, introducing covariates in the innovations of the INAR(1) hidden process. The models could be also extended to describe more complex patterns using time-dependent under-reporting parameters.

The described methodology opens a wide field for future research lines. For instance, other distributions different than Poisson can be used for the innovations of the hidden process, like that introduced in [20]. Moreover, higher-order dependencies in the hidden process could be considered using general INAR(p) ($p > 1$) time series. Starting from the multivariate INAR(1) processes described in [39], under-reported multivariate time series could be also modelled in a similar way than the one presented here.

Appendix A: Proof of Proposition 1

The following is the proof of the Proposition (9):

Proof
Consider $k \neq 0$

$$\begin{aligned}\mathbf{E}\left(Y_nY_{n+k}\right) =& \mathbf{E}\left(X_n\left(1-\mathbf{1}_n\right)X_{n+k}\left(1-\mathbf{1}_{n+k}\right)\right) + \mathbf{E}\left(X_n\left(1-\mathbf{1}_n\right)q\circ X_{n+k}\mathbf{1}_{n+k}\right)\\ &+\mathbf{E}\left(X_{n+k}\left(1-\mathbf{1}_{n+k}\right)q\circ X_n\mathbf{1}_n\right)+\mathbf{E}\left(q\circ X_n\, q\circ X_{n+k}\mathbf{1}_n\mathbf{1}_{n+k}\right).\end{aligned}$$

We can compute each of the terms in the right separately. Because $\mathbf{1}_n$ and $\mathbf{1}_m$ are independent of the chain $\{X_j\}$ and have expectation ω,

$$\mathbf{E}\left(X_n\left(1-\mathbf{1}_n\right)X_{n+k}\left(1-\mathbf{1}_{n+k}\right)\right) = (1-\omega)^2\,\mathbf{E}\left(X_nX_{n+k}\right).$$

Similarly, using the fact that $\mathbf{E}\left(q\circ X_n|X_n\right) = \mathbf{E}\left(\sum_{j=1}^{X_n}\xi_j|X_n\right) = qX_n$,

$$\mathbf{E}\left(X_n\left(1-\mathbf{1}_n\right)q\circ X_{n+k}\left(\mathbf{1}_{n+k}\right)\right) = (1-\omega)\,\omega q\mathbf{E}\left(X_nX_{n+k}\right).$$

Finally,

$$\mathbf{E}\left(q\circ X_n q\circ X_{n+k}\mathbf{1}_n\mathbf{1}_{n+k}\right) = \omega^2q^2\mathbf{E}\left(X_nX_{n+k}\right).$$

Hence,

$$\begin{aligned}\mathbf{E}\left(Y_nY_{n+k}\right) &= \mathbf{E}\left(X_nX_{n+k}\left((1-\omega)^2+2\,(1-\omega)\,\omega q+\omega^2q^2\right)\right) = (1-\omega(1-q))^2\,\mathbf{E}\left(X_nX_{n+k}\right)\\ &= (1-\omega(1-q))^2\left(\alpha^{|k|}\sigma_X^2+\mu_X^2\right) = (1-\omega(1-q))^2\left(\alpha^{|k|}\sigma_X^2+\frac{\lambda^2}{(1-\alpha)^2}\right),\end{aligned}$$

so

$$\gamma_Y(k) = \mathbf{Cov}\left(Y_n, Y_{n+k}\right) = (1-\omega(1-q))^2 \left(\alpha^{|k|}\sigma_X^2 + \frac{\lambda^2}{(1-\alpha)^2}\right) - (1-\omega(1-q))^2 \frac{\lambda^2}{(1-\alpha)^2}$$
$$= (1-\omega(1-q))^2 \alpha^{|k|}\sigma_X^2 = (1-\omega(1-q))^2 \alpha^{|k|} \frac{\lambda}{1-\alpha}.$$

The proof follows immediately, because $\gamma_Y(0) = \mathbf{Var}(Y_n)$. □

Acknowledgements

This work was funded by grants MTM2012-31118 and MTM2015-69493-R from the Spanish Ministry of Economy and Competitiveness. The authors thank the two referees for their valuable comments and suggestions.

References

1. McKenzie E. *Stochastic Processes: Modelling and Simulation, Handbook of Statistics*, Vol. 21. Elsevier, 2003. DOI:10.1016/S0169-7161(03)21018-X.
2. Cardinal M, Roy R, Lambert J. On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine* 1999; **18**(15):2025–2039.
3. Freeland R, McCabe B. Forecasting discrete valued low count time series. *International Journal of Forecasting* 2004; **20**(3):427–434.
4. Moriña D, Puig P, Ríos J, Vilella A, Trilla A. A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine* 2011; **30**(26):3125–3136.
5. Monteiro M, Scotto MG., Pereira I. Integer-valued autoregressive processes with periodic structure. *Journal of Statistical Planning and Inference* 2010; **140**(6):1529–1541.
6. Gourieroux C, Jasiak J. Heterogeneous INAR(1) model with application to car insurance. *Insurance: Mathematics and Economics* 2004; **34**(2):177–192.
7. Winkelmann R. Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics* 1996; **21**(4):575–587.
8. Alfonso JH, Løvseth EK, Samant Y, Holm J. Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis* 2015; **72**(6):409–412.
9. Rosenman KD, Kalush A, Reilly MJ, Gardiner JC, Reeves M, Luo Z. How much work-related injury and illness is missed by the current national surveillance system? *Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine* 2006; **48**(4):357–365.
10. Arendt S, Rajagopal L, Strohbehn C, Stokes N, Meyer J, Mandernach S. Reporting of foodborne illness by U.S. consumers and healthcare professionals. *International Journal of Environmental Research and Public Health* 2013; **10**(8):3684–3714.
11. Höhle M, an der Heiden M. Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics* 2014; **70**:993–1002.
12. Bernard H, Werber D, Hoehle M. Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104: H4 in 2011-a time series analysis. *BMC Infectious Diseases* 2014; **14**:116.
13. Al-Osh MA, Alzaid AA. First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis* 1987; **8**(3):261–275.
14. Jung RC, Tremayne A. Binomial thinning models for integer time series. *Statistical Modelling* 2006; **6**(2):81–96.
15. Scotto MG, Weiss CH, Gouveia S. Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling* 2015; **15**(6):590–618.
16. Schweer S, Weiß CH. Compound Poisson INAR (1) processes: stochastic properties and testing for overdispersion. *Computational Statistics & Data Analysis* 2014; **77**:267–284.
17. Boulanger M, Morlais F, Bouvier V, Galateau-Salle F, Guittet L, Marquignon MF, Paris C, Raffaelli C, Launoy G, Clin B. Digestive cancers and occupational asbestos exposure: incidence study in a cohort of asbestos plant workers. *Occupational and Environmental Medicine* 2015; **72**(11):792–797.
18. Frontario SCN, Loveitt A, Goldenberg-Sandau A, Liu J, Roy D, Cohen LW. Primary peritoneal mesothelioma resulting in small bowel obstruction: a case report and review of literature. *The American Journal of Case Reports* 2015; **16**:496–500.
19. Conti S, Minelli G, Ascoli V, Marinaccio A, Bonafede M, Manno V, Crialesi R, Straif K. Peritoneal mesothelioma in Italy: trends and geography of mortality and incidence. *American Journal of Industrial Medicine* 2015; **58**(10):1050–1058.
20. Jazi MA, Jones G, Lai CD. First-order integer valued AR processes with zero inflated Poisson innovations. *Journal of Time Series Analysis* 2012; **33**(6):954–963.
21. Maiti R, Biswas A, Guha A, Ong SH. Modelling and coherent forecasting of zero-inflated count time series. *Statistical Modelling* 2014; **14**(5):375–398.
22. Benaglia T, Chauveau D, Hunter DR, Young D. mixtools: an R package for analyzing finite mixture models. *Journal of Statistical Software* 2009; **32**(6):1–29.
23. Zucchini W, MacDonald IL. *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press: Boca Raton, FL, 2009.

24. Cox DR, Snell JE. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)* 1968; **30**(2):248–275.
25. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 1967; **13**(2):260–269.
26. Forney G. The viterbi algorithm. *Proceedings of the IEEE* 1973; **61**(3):268–278.
27. Himmelmann L. *HMM: HMM – Hidden Markov Models*, 2010. http://cran.r-project.org/package=HMM.
28. Harte D. *HiddenMarkov: Hidden Markov Models*, 2015. http://cran.r-project.org/package=HiddenMarkov.
29. Dunne EF, Markowitz LE, Saraiya M, Stokley S, Middleman A, Unger ER, Williams A, Iskander J. CDC grand rounds: reducing the burden of HPV-associated cancer and disease. *MMWR. Morbidity and mortality weekly report* 2014; **63**(4): 69–72.
30. Imai M, Hino O. Environmental carcinogenesis – 100th anniversary of creating cancer, *Cancer science*. John Wiley & Sons: New York, NY, 2015.
31. Petersen R, Petersen JA, Mikkelsen S. [Non-occupational pleural mesothelioma]. *Ugeskrift for Laeger* 2015; **177**(3):2–3.
32. Pitarque S, Clèries R, Martínez JM, López-Abente G, Kogevinas M, Benavides FG. Mesothelioma mortality in men: trends during 1977–2001 and projections for 2002–2016 in Spain. *Occupational and Environmental Medicine* 2008; **65**(4): 279–282.
33. Park EK, Takahashi K, Hoshuyama T, Cheng TJ, Delgermaa V, Le GV, Sorahan T. Global magnitude of reported and unreported mesothelioma. *Environmental Health Perspectives* 2011; **119**(4):514–518.
34. Eilers MBD PHC, Durban M. Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions* 2015; **39**(2):149–186.
35. Rocourt J, Moy G, Vierk K, Schlundt J. The present state of foodborne disease in OECD countries. *Techical Report*, Food Safety Department – World Health Organization Geneva, Switzerland, 2003.
36. Tamblyn SE. The frustrations of fighting foodborne disease. *CMAJ : Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne* 2000; **162**(10):1429–1430.
37. Koepke R, Sobel J, Arnon SS. Global occurrence of infant botulism, 1976–2006. *Pediatrics* 2008; **122**(1):e73–e82.
38. Weiß CH. Thinning operations for modeling time series of counts-a survey. *Advances in Statistical Analysis* 2008; **92**(3):319–341.
39. Pedeli X, Karlis D. Some properties of multivariate INAR(1) processes. *Computational Statistics & Data Analysis* 2013; **67**:213–225.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.