SPANISH JOURNAL OF STATISTICS
Vol. **3** No. 1 2021, Pages 37–44
doi:https://doi.org/10.37830/SJS.2021.1.03

REGULAR ARTICLE

Misreported longitudinal data in epidemiology: Review of mixture-based advances and current challenges

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

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

Received: October 1, 2021. Accepted: December 20, 2021.

Abstract: The problem of dealing with misreported data is very common in a wide range of contexts and for different reasons. This has been and still is an important issue for data analysts and statisticians as not accounting for it could led to biased estimates and conclusions, and in many cases that would have implications in a posterior decision making process, as we all have seen in the current worldwide Covid-19 pandemic. In the last few years, many approaches have been proposed in the literature to accomodate data presenting this issue, especially in the fields of epidemiology and public health but also in other areas as social science. In this work, a comprehensive review of the recently proposed methods based on mixture models for longitudinal data (correlated and uncorrelated) is presented and several examples of application are discussed, including several approaches to the burden of Covid-19 infection cases in Spain and different approaches to deal with underreported registries of human papillomavirus infections and genital warts in Catalunya.

Keywords: epidemiology, modelling, misreported data, longitudinal data, public health

MSC: 37M10, 92D30, 46N30, 62-07

1 Introduction

Dealing with misreported data is a very common problem in many fields. The most usual issue is facing data that is only partially registered or underreported, and this type of misreporting has received some attention in the recent years, with alternative approaches being proposed in epidemiology, social and biomedical research, among many other contexts (Bernard et al., 2014; Arendt et al., 2013; Rosenman et al., 2006; Alfonso et al., 2015; Winkelmann, 1996). The opposite phenomenon, i. e. dealing with overreported data, occurs with less frequency although also arises in some cases (see for instance Mehta (2018)), and has been less studied. The mechanisms leading to misreported data are diverse, and so are the methods proposed in the literature to overcome this issue, ranging from Markov chain Monte-Carlo (Winkelmann, 1996) approaches to time series analysis (Fernández-Fontelo et al., 2016; Fernández-Fontelo et al., 2019) or spatio-temporal modelling (Stoner et al., 2019). From the computational point of view, a new R (R Core Team, 2019) package useful to fit endemic-epidemic models based on approximate maximum likelihood to underreported count data (Bracher, 2019) and the package *MisRepARMA* (Moriña et al., 2021), able to fit misreported time series and to reconstruct the most likely actual series were recently introduced. Additionally a web application allowing its users to fit misreported AutoRegressive Moving Average (ARMA) time series model has been recently published and is accessible at https://dmorina.shinyapps.io/MisRepARMA/.

Misreported data might potentially lead to biased inference, as it may invalidate the assumptions of standard models. For instance, Winkelmann (1996) explores a Markov chain Monte Carlo based methodology to study worker absenteeism where two sources of underreporting 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 underreported (Alfonso et al., 2015; Rosenman et al., 2006; Arendt et al., 2013). Of course in that case there might be several sources of underreporting, including accuracy of public health registries, political or economical interests, among others.

This review is focused on methods recently proposed to deal with permanent underreporting, assuming that there are more cases in the population than the acknowledged ones. In many applications, however, underreporting occurs by a delay in reporting, but as time passes, the counts became more and more complete, as studied in Höhle and an der Heiden (2014). This is also the case of the public health information systems of many countries regarding the Covid-19 pandemic, as they are being updated when the results of the tests are available, but we have seen that unfortunately, waiting until the information is complete is often too late to make the optimal decisions to handle the pandemic in a proper way.

From the methodological point of view, the first attempts to overcome this problem in the context of longitudinal data were based on count time series, with a growing popularity due to the limited performance of the classical continuous time series approach when the variable of interest consists of low counts (Fernández-Fontelo et al., 2016; Fernández-Fontelo et al., 2019). Similar ideas were used to analyze uncorrelated longitudinal data (Moriña et al., 2021). Another additional issue to deal with inspired by the Covid-19 data is non-stationarity. This issue has been treated in Fernández-Fontelo et al. (2020) by means of discrete time series.

2 Proposed models and examples

Some of the most recently proposed methodologies based on mixture models and able to deal with misreported data are described in this Section, together with some examples of application.

2.1 Longitudinal uncorrelated data

In this context, the simplest case is when the longitudinal data we are dealing with present no temporal correlation. This is the case covered in Moriña et al. (2021), where a model is proposed and used to analyze the underreporting issue in genital warts (GW) record in Catalonia (northeastern

Spain). In the particular case of sexually transmitted diseases, it is very important to have reliable information due to their remarkable morbidity, and therefore, the importance of controlling trends over time and priority setting (see McCormack and Koons (2019) for a comprehensive discussion focused on developing countries).

Several models proposed in the literature able to deal with misreported longitudinal data are based on more sophisticated versions of the following schema, replacing the usual product by other operators in the discrete case (usually by the binomial thinning operator, see Section 2.2).

Consider X_t the series of real GW incidence, where $t = 1, 2, \ldots$ is the time, following a normal distribution with mean μ and variance σ^2. In this setting, the X_t process cannot be directly observed, and all we can see is a part of it, expressed as

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

The interpretation of the parameters in Eq. (1) is straightforward: q is the overall intensity of misreporting (if $0 < q < 1$ the observed process Y_t would be underreported while if $q > 1$ the observed process Y_t would be overreported). The parameter ω can be interpreted as the overall frequency of misreporting (proportion of misreported observations).

The series Y_t represents the registered values corresponding to GW incidence. According to Eq. (1), the registered observations series Y_t is a mixture of two normally distributed random variables $Y_t = Y_{1t}$ with probability $(1 - \omega_t)$ and $Y_t = Y_{2t}$ with probability ω_t, where Y_{1t} coincides with the unobserved process X_t and Y_{2t} is a normal random variable with mean $q \cdot \mu$ and variance $q^2 \cdot \sigma^2$. The parameter ω_t is modeled as $logit(\omega_t) = \alpha_0 + \alpha_1 \cdot t$ and can be interpreted as the frequency of underreporting at a time t, while q can be interpreted as the intensity of such underreporting, both taking values between 0 and 1. When $q = 0$ the observed incidence is $Y_t = 0$ and when $q = 1$ there is no underreporting. A value of ω_t equal to 0 indicates that the observed value at time t is not underreported, and a value of ω_t equal to 1 means that underreporting is for sure happening. In order to detect potential differences in GW incidence depending on sex (men and women) and age group (16-29 and 30-94), these covariates were included in the model, so the mean of the observed process Y_{1t} was modeled as $\mu_{1,t} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s$ (where a is the age, s is the sex and $a * s$ is the interaction between age and sex). The average of the second component Y_{2t} can be recovered as $\mu_{2,t} = q \cdot (\beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s)$.

In the particular case of these data, after diagnostic checks a seasonal cycle of 3 months was observed. This behavior was included in the model through the following trigonometric funtion
$f(t) = \beta_5 \cdot sin(\frac{2 \cdot \pi \cdot t}{3}) + \beta_6 \cdot cos(\frac{2 \cdot \pi \cdot t}{3})$ on the terms $\mu_{1,t}$ and $\mu_{2,t}$.

The loglikelihood function corresponding to the final model considered in this paper can be written as

$$l(Y, \theta) = \sum_{t=1}^{n} \log \left((1 - \omega_t) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_t - \mu_{1,t})^2}{2\sigma^2}} + \omega_t \frac{1}{\sqrt{2\pi} q \sigma} e^{\frac{(y_t - \mu_{2,t})^2}{2q^2\sigma^2}} \right), \tag{2}$$

where $Y = y_1, \ldots, y_n$ is the observed series, $\theta = (\alpha_0, \alpha_1, \gamma, \beta_0, \ldots, \beta_6, \sigma)$, $\omega_t = \frac{e^{\alpha_0 + \alpha_1 t}}{1 + e^{\alpha_0 + \alpha_1 t}}$, $q = \frac{e^{\gamma}}{1 + e^{\gamma}}$ and $\mu_{1,t}$ and $\mu_{2,t}$ are as defined before.

The estimates and associated standard errors can be obtained maximizing Eq. 2 in the usual way.

The results of this work show that in relative terms, the underreporting issue has a deeper impact on people over 30 years old (where GW incidence is lower), especially among women. Nonetheless, the relative difference between registered and estimated annual averages range between 13.3% and 24.9%. It is also remarkable that the quality of SIDIAP register regarding GW in Catalunya has been significantly improving during the study period, as the frequency of underreported observations has been decreasing over time, from around 95% in 2009 to 21% in 2016.

2.2 Discrete stationary time series

As stated before, the underreporting structure defined in Eq. 1 has been adapted to the discrete time series case in Fernández-Fontelo et al. (2016), by means of the *binomial thinning* operator. The proposed model assumes, as before, that the actual values of the series X_t are not fully observed and follow an integer-valued autoregressive model of order 1 (INAR(1)) (Jung and Tremayne, 2006):

$$X_t = \alpha \circ X_{t-1} + W_t(\lambda), \tag{3}$$

where $\alpha \in (0, 1)$ and W_t is assumed to follow a Poisson distribution with a fixed mean λ. In addition, X_{t-1} and W_t are assumed to be independent at any time t. The $\circ$ operator in expression (3), called *binomial thinning*, is defined as follows,

$$\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} Y_i, \tag{4}$$

where Y_i are iid Bernoulli random variables with a probability of success equal to α. Therefore, if $X_{t-1} = x_{t-1}$, then $\alpha \circ x_{t-1}$ is binomially distributed with the number of successes equal to x_{t-1}.

The observed process Y_t is then defined as

$$Y_t = \begin{cases} X_t & : \text{with probability } 1-\omega \\ q \circ X_t & : \text{with probability } \omega \end{cases} \tag{5}$$

The interpretation of the parameters is very similar to that of the previous models (notice that in this case the frequency of underreported observations ω is not time dependent), and two methods for estimation are provided in Fernández-Fontelo et al. (2016). The first method is based on the fact that the marginal distribution of Y_t 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, so an estimation of $\frac{\lambda}{1-\alpha}$, $\frac{q\lambda}{1-\alpha}$ and ω can be found by fitting the whole observed series using a mixture of two Poisson distributions. The parameters α and λ can be estimated using for instance the first autocorrelation coefficients, which are explicitly derived in the paper.

The estimates obtained by using the previously described method could be used as initial values to improve the speed of the maximum likelihood algorithm, based on an adapted version of the forward algorithm (Zucchini and MacDonald, 2009).

One strength of this methodology is that it allows to recover the most likely hidden process X_t using the Viterbi algorithm (Viterbi, 1967; Forney, D. G., 1973), which is implemented in several R packages for situations with a finite number of states, and has to be adapted to cover the present case also.

This model was used in Fernández-Fontelo et al. (2016) to analyze three examples in the field of public health, illustrating the model performance in the case of a stationary series (weekly HPV cases in Girona in the period 2010-2014) but also in the presence of trends (number of annual deaths by mesothelioma in Great Bretain in the period 1968-2013 and number of annual cases of botulism in Canada in the period 1970-2013). According to the reported results, it can be seen that although only an average of 1.27 HPV cases per week were registered, the most likely reconstructed series estimated an average of 3.36 cases per week. Regarding botulism in Canada, an average of 9 cases per year were registered although a proportion of $\hat{\omega} = 0.671$ underreported observations was estimated by the model. Mesothelioma is a disease related to asbestos exposure and is difficult to diagnose so although an average of 14 annual deaths in Great Britain in the period 1968-2013 it is supposed to be highly underreported. It is confirmed by an estimated proportion of $\hat{\omega} = 0.930$ underreported observations.

A more sophisticated underreporting structure was introduced in Fernández-Fontelo et al. (2019). In this case, the hidden process X_t is again assumed to follow an INAR(1) structure but the observed series Y_t is defined by

$$Y_t = \begin{cases} X_t & : \text{If } I_t = 0, \\ q \circ X_t & : \text{If } I_t = 1, \end{cases} \tag{6}$$

where $\circ$ is again the binomial thinning operator and I_t is a binary discrete time Markov chain indicating whether the observation Y_t is underreported or not. The difference with respect to the previous case is that in this case the states of underreporting I_t are serially dependent. Two methods of parameter estimation are provided in the model, one based on moments method and another based on maximum likelihood, and the most likely hidden process can be recovered by a Viterbi-like algorithm.

This model was used to analyze the underreporting issue on the gender-based violence reports in several Galician districts, and it was shown that as expected, it is an important problem related to these kind of data that should be accounted for, but also that the underreporting is severely non-uniformly distributed across the considered districts, as the frequency of underreported observations ranged from $\hat{\omega} = 0.078$ (very few observations are underreported in this area) to $\hat{\omega} = 0.976$ (almost all observations are underreported).

2.3 Discrete non-stationary time series

The outbreak of the Covid-19 pandemic in 2020 has made very clear that models able to handle non-stationary time series with potential underreporting issues were absolutely necessary to help decision-makers. The previously discussed models for discrete stationary time series can be extended in several ways to be able to deal with non-stationary discrete time series. A possible way is explored in Fernández-Fontelo et al. (2020), where the authors consider that the true (unobserved) counts X_t follow again an INAR(1) structure and the observed process Y_t is, as before, defined by

$$Y_t = \begin{cases} X_t & \text{with probability } 1 - \omega \\ q \circ X_t & \text{with probability } \omega. \end{cases} \tag{7}$$

The difference with respect to the previous models for discrete time series is that the mean of the innovations of the hidden process X_t and the underreporting parameter q are modelled as functions of time and thus allowing for non-time-homogeneous processes, and that only maximum likelihood estimation is suitable in the case of non-stationary time series. Both underreporting-related parameters ω and q could be considered time-dependent, but in this work, to avoid potential convergence issues, only time-dependent q is considered. In particular, as a weekly seasonal behavior was observed in the reported data and to ensure that the estimates are within the interval $(0, 1)$, it was adjusted by means of the logistic function

$$q_t = \frac{\exp\left(\gamma_0 + \gamma_1 t + \gamma_2 \sin\left(\frac{2\pi t}{7}\right) + \gamma_3 \cos\left(\frac{2\pi t}{7}\right)\right)}{1 + \exp\left(\gamma_0 + \gamma_1 t + \gamma_2 \sin\left(\frac{2\pi t}{7}\right) + \gamma_3 \cos\left(\frac{2\pi t}{7}\right)\right)}, \tag{8}$$

where γ_1 indicates whether q increases or decreases over time and γ_2 and γ_3 indicate whether the series has a seasonal pattern with period $p = 7$ (weekly). Notice that if $\gamma_1 = \gamma_2 = \gamma_3 = 0$, then the previous logistic function becomes constant and thus $q_t = q$, resulting in the model (9). Hence, considering this function for the intensity of the underreporting, the underreporting process in this model was defined by:

$$Y_t = \begin{cases} X_t & \text{with probability } 1-\omega \\ q_t \circ X_t & \text{with probability } \omega. \end{cases} \tag{9}$$

At this point, however, the model is not able to handle non-stationarity yet. In order to be able to fit non-stationary discrete time series and answer the main question of researchers facing an epidemic, the innovations of the hidden INAR(1) process X_t are linked to a compartimental epidemiological model. In this case, a Susceptible-Infected-Recovered (SIR) model was used with that purpose. This model considers three compartments were individuals are included in at each time t: Those who are healthy but susceptible to get the disease ($S(t)$), those who are infected and thus transmitters of the disease ($I(t)$), and those individuals who have been removed from the system and will not get infected again ($R(t)$) (Anderson and May, 1992; Vynnycky and White, 2010). The parameters of interest are the infection rate β, the removal rate γ, and the total susceptible population N.

It can be seen that the number of individuals affected by the disease at time t can be represented by:

$$A(t) = \frac{M^* A(0) \mathrm{e}^{kt}}{M^* + A(0)(\mathrm{e}^{kt} - 1)}, \tag{10}$$

where $k = \beta - \gamma$ and $M^* = \frac{N(\beta-\gamma)}{\beta-\gamma/2}$.

The information on the spread of the disease can be included in the hidden process X_t by considering that the expectation of the innovations is not constant as in the previous models but time varying as $\lambda_t = \text{new}(t) = A(t) - A(t-1)$, where $\text{new}(t)$ are the new affected cases at time t.

Therefore, the unobserved process X_t becomes:

$$X_t = \alpha \circ X_{t-1} + W_t(\lambda_t), \tag{11}$$

As in the previous models described, an appropriate version of the Viterbi algorithm could be used to reconstruct the most likely hidden process, which is specially interesting in this case as it provides a more realistic picture of the pandemic situation in a given moment of time.

Additionally, two forecasting methods are provided; one based on average point predictions given the sample observations and another based on the conditional distribution of Y_{t+k} given the last value of the latent process X_t. Standard errors for the first forecasting method could be obtained using a numerical mechanism as the Delta method. For the second method forecasts, prediction regions of size $1-\alpha$ can be found through the values l (lower limit) and u (upper limit) satisfying $\sum_{j=1}^{l} \mathrm{P}(Y_{t+k} = j | X_t = x_t) \approx \frac{\alpha}{2}$ and $\sum_{j=1}^{u} \mathrm{P}(Y_{t+k} = j | X_t = x_t) \approx 1 - \frac{\alpha}{2}$.

This model was used to analyze the Covid-19 data registered in several regions in Spain, and the results reported in the paper confirm that the underreporting issue is indeed present in Covid-19 data from various regions conditioned to different management, policies, and climate conditions, and that it also varies across geographic areas, with registered coverages ranging from 33.7 % in Ourense (Galicia) to 71.8 % in Málaga (Andalucía) of the estimated cases.

3 Discussion

Facing misreported information from public health registers is very common in many situations, for instance data regarding potentially asymptomatic diseases like HPV infection or Covid-19, or difficult to diagnose as mesothelioma.

One of the lessons that should certainly be learned from the current Covid-19 pandemic is that it is crucial to provide researchers with reliable data under extremely complex circumstances, in order to be able to assure public health decision makers are provided with the most reliable information at any time. When this is by no ways possible, the issue should be at least taken into account by using a model capable of accommodating underreported data like the one used in this study.

Although several efforts have been done in this direction in the last years, several challenges still remain. From the methodological point of view, there are a few works dealing with underreported discrete time series, but they are a bit limited for instance in the allowed structure for the hidden process. These models could be extended in different ways, such as considering more complex correlation structures in the underlying process (for instance INAR(p) or INARMA(p, q) structures), or considering more general thinning operators for representing the observed process.

Also from the methodological point of view, the extension of the model introduced in Moriña et al. (2021) to underreported continuous time series is not available yet, nor for stationary or non-stationary processes. Additionally, the considered methods could certainly be useful in the development of new, more general methodologies able to deal with overreporting as well, as this is another issue that, although with minor frequency, appears in the practice of epidemiology and public health.

The availability of packages for commonly used software as R (R Core Team, 2019) would also help to make these methods reach a wider public of potential users, and some efforts in this sense are currently been done (Bracher (2019) for instance).

From the applied point of view, it would be very interesting to use these kind of models to analyze other issues that might be potentially underreported and to analyze more thoroughly the examples used to illustrate the performance of the discussed models. For instance, the differences across geographic areas observed in Fernández-Fontelo et al. (2019) related to underregistered reports of gender-based violence could be better explained if covariates (as rurality index, socioeconomic variables, spatial correlation...) were included in the model.

References

Alfonso, J. H., E. K. Løvseth, Y. Samant, and J. Holm (2015). Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis 72*(6), 409–412.

Anderson, R. M. and R. M. May (1992). *Infectious diseases of humans: dynamics and control*. Oxford University Press.

Arendt, S., L. Rajagopal, C. Strohbehn, N. Stokes, J. Meyer, and S. Mandernach (2013). Reporting of foodborne illness by U.S. consumers and healthcare professionals. *International journal of environmental research and public health 10*(8), 3684–3714.

Bernard, H., D. Werber, and M. Höhle (2014). Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104: H4 in 2011 - a time series analysis. *BMC Infectious Diseases 14*(1).

Bracher, J. (2019). hhh4u: Fit an endemic-epidemic model to underreported data.

Fernández-Fontelo, A., A. Cabaña, H. Joe, P. Puig, and D. Moriña (2019). Untangling serially dependent underreported count data for gender-based violence. *Statistics in Medicine 38*(22), 4404–4422.

Fernández-Fontelo, A., A. Cabaña, P. Puig, and D. Moriña (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine 35*(26), 4875–4890.

Fernández-Fontelo, A., D. Moriña, A. Cabaña, A. Arratia, and P. Puig (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE 15*, e0242956.

Forney, D. G. (1973). The viterbi algorithm. *Proceedings of the IEEE 61*(3), 268–278.

Höhle, M. and M. an der Heiden (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics 70*(4), 993–1002.

Jung, R. C. and A. R. Tremayne (2006). Binomial thinning models for integer time series. *Statistical Modelling 6*(2), 81–96.

McCormack, D. and K. Koons (2019). Sexually Transmitted Infections. *Emergency Medicine Clinics of North America 37*(4), 725–738.

Mehta, R. (2018). Allergy and Asthma: Food Allergies. *FP essentials 472*, 16–19.

Moriña, D., A. Fernández-Fontelo, A. Cabaña, and P. Puig (2021). MisRepARMA: Misreported time series analysis.

Moriña, D., A. Fernández-Fontelo, A. Cabaña, P. Puig, L. Monfil, M. Brotons, and M. Diaz (2021). Quantifying the under-reporting of genital warts cases. *BMC Medical Research Methodology 21*(1), 6.

R Core Team (2019). R: A Language and Environment for Statistical Computing.

Rosenman, K. D., A. Kalush, M. J. Reilly, J. C. Gardiner, M. Reeves, and Z. Luo (2006). How much work-related injury and illness is missed by the current national surveillance system? *Journal of Occupational and Environmental Medicine 48*(4), 357–365.

Stoner, O., T. Economou, and G. Drummond Marques da Silva (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association 114*(528), 1481–1492.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*(2), 260–269.

Vynnycky, E. and R. White (2010). *An introduction to infectious disease modelling*. Oxford University Press.

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

Zucchini, W. and I. L. MacDonald (2009). *Hidden Markov Models for Time Series : An Introduction Using R*. Chapman & Hall/CRC.