Modelling under-reported data through INAR-hidden Markov chains

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

amanda@mat.uab.cat, acabana@mat.uab.cat, ppuig@mat.uab.cat, david.morina@uab.cat

[1]Departament de matemàtiques, Universitat Autònoma de Barcelona, 08193, Bellaterra, Spain.
[2]Unit of Infections and Cancer (UNIC), Cancer Epidemiology Research Program (CERP), Catalan Institute of Oncology (ICO)-IDIBELL, LHospitalet de Llobregat, Barcelona, Spain
[3]Grups de Recerca dfrica i Amrica Llatines (GRAAL), Unitat de Bioestadstica, Facultat de Medicina, Universitat Autnoma de Barcelona, Bellaterra, Spain

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 (Al-Osh and Alzaid, 1987). The main reason for this increasing popularity is the limited performance of the classical time series analysis approach when dealing with discrete valued time series. With the introduction of discrete time series analysis techniques, several challenges appeared such as unobserved heterogeneity, periodicity, under-reporting, among others. Many efforts have been devoted in order to introduce seasonality in these models (Moriña *et al.*, 2011) and also coping with unobserved heterogeneity. However, the problem of under-reported data is still in a quite early stage of study in many different fields. This phenomenon is very common in many contexts such as epidemiological and biomedical research. It might lead to potentially biased inferences and may also invalidate the main assumptions of the classical models. Especially, in public health context it is well known that several diseases have been traditionally under-reported (occupational realted diseases, food exposures diseases, ...). The model we will present in this work considers two discrete time series: the observed series of counts Y_t which may be under-reported, and the underlying series X_t with an INAR(1) structure $X_n = \alpha \circ X_{n-1} + W_n$, where $0 < \alpha < 1$ is a fixed parameter and W_n are the innovations which are Poisson(λ) distributed. The binomial thinning operator (or binomial subsampling) is defined as $\alpha \circ X_{n-1} = \sum_{i=1}^{X_{n-1}} Z_i(\alpha)$; where Z_i are i.i.d Bernoulli random variables with probability of success equal to α. The way we allow the observed process Y_n to be under-reported is by defining that Y_n is X_n with probability $1-\omega$ or is $q\circ$ with probability ω. Obviously, this definition means that the observed Y_n coincides with the underlying series X_n, and therefore the count at time n is not under-reported with probability $1-\omega$. Several applications in the field of public health will be discussed, using real data regarding incidence and mortality attributable to diseases related to occupational and environmental exposures and known toxics and traditionally under-reported. Full details of the work can be found in Fernández-Fontelo *et al.*, (2016).

[1] Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive INAR(1) process. *Journal of Time Series Analysis*, **8**, 261–275.
[2] Moriña, D., Puig, P., Ríos, J., Vilella, A. and Trilla, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine*, **30**, 3125–3136.
[3] Fernández-Fontelo, A., Cabaña, A., Puig, P. and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, **35**, 4875–4890.