Universitat Autònoma de Barcelona

New models of count data with applications

AMANDA FERNÁNDEZ FONTELO

Doctoral Thesis

New models of count data with applications

AMANDA FERNÁNDEZ FONTELO

Supervised by PERE PUIG CASADO

Departament de Matemàtiques

Doctorat en Matemàtiques

2017-2018

Universitat Autònoma de Barcelona

Caminante, son tus huellas
el camino y nada más;
Caminante, no hay camino,
se hace camino al andar.
Al andar se hace el camino,
y al volver la vista atrás
se ve la senda que nunca
se ha de volver a pisar.
Caminante no hay camino
sino estelas en la mar.

Antonio Machado Ruiz

Acknowledgements

First, I would like to express my sincere gratitude to my supervisor Dr. Pere Puig for his support, patient, motivation, and immense knowledge. Words cannot express how grateful I am to him.

Besides my supervisor, I would like to thank Dr. Harry Joe and Dr. Elizabeth Ainsbury who provided me with the opportunity to join their research groups. Without their support, it would have been impossible to conduct this research.

I would like to express my special appreciation to Dr. Alejandra Cabaña, Dr. David Moriña, Dr. Anna Alba and Dr. Manuel Higueras for encouraging my research and allowing me to grow as a research scientist.

My sincere thanks also go to Paco Carreño, Loli Garcia, Maria José Calejo and Beatriz Díaz. A part of this Ph.D. thesis is yours.

The last but not the least, I would like to thank Oriol, my family and friends for being always by my side.

Amanda Fernández Fontelo
Barcelona, September 2018

Contents

Introduction

Non-negative integer-valued data (or count data) are intrinsically found in the nature of many phenomena, and many models related to essential applications have been presented in the literature in several contexts such as epidemiological or biomedical research, social research, economic research, ecological research, among others. Since the beginning of the last century, some renowned authors have proposed models of count data in many interesting real-world processes. For instance, Tukey (1949) assumed a Poisson distribution for the number of mutants in a bacterial sample, Fisher *et al.* (1922) modelled the number of bacteria in soil samples with a Poisson distribution, Fisher and Mather (1936) used the Binomial distribution in a linkage test with mice data, and Rao *et al.* (1973) considered the Negative Binomial distribution for the number of different compounds identified in water samples.

Since count data are present everywhere, the necessity of high-quality methods and techniques to correctly model and analyse these data is indisputable. In this sense, many comprehensive works can be found in the literature, where both, basic techniques and more extended methods to investigate count data, have fully developed from different perspectives. Since the forties, some of the most influential contributions to the analysis of non-negative integer-valued data can be found in Fisher (1941) who presented a detailed description of the Negative Binomial distribution, Kemp and Kemp (1965) who introduced some relevant properties of the Hermite distribution, the works carried out by

Johnson and Kotz (1969), Johnson and Kotz (1982) and Johnson *et al.* (1997) where an extended review (moments, properties, methods of estimation, applications, among others) of some of the standard and not-so-common discrete-valued data distributions were presented, and Hinde (1982) who proposed an approach to estimate compound Poisson regression models. Moreover, other interesting contributions can be found in D'Agostino and Stephens (1986) who described, among others, the well-known Chi-squared test and some other tests of goodness-of-fit for discrete distributions, Cox and Snell (1989) who set out different approaches to analyse binary data, Karlis and Xekalaki (1999) who considered the problem of mixtures of Poisson distributions, Puig and Valero (2007) who characterized count data distributions involving additivity and binomial subsampling, and the work written by Hilbe (2011) where the negative binomial regression was studied in detail. On the other hand, some works such as Mood (1950) and Cox and Hinkley (1974) provided an introduction to the theory of statistics, including that theory related to count data distributions.

Despite the vast amount of excellent-quality works dealing with the major concerns in non-negative integer-valued data, and the enormous effort of every author who has contributed to the better understanding and modelling of many count data phenomena, some issues related to these data are not entirely solved yet. Among these issues, it is noteworthy the well-documented problem of overdispersion which has been studied by many authors. For instance, Cox (1983) examined the behaviour of maximum likelihood estimators when slightly overdispersion is present in a simple model, Cameron and Trivedi (1986) proposed regression-based tests for overdispersion in the Poisson model, Barron (1992) described some appropriate methods of count data analysis when both, overdispersion and auto-correlation, are present, and finally Ganio and Schefer (1992) presented some diagnostic tools for assessing overdispersion. Other essential works can be found in Del Castillo and Pérez-Casany (1998) who introduced the weighted Poisson distribution for fitting overdispersion in count data, Sellers *et al.* (2010) who reviewed the COM-Poisson distribution, and Weiß and Schweer (2015) who proposed a method for detecting overdispersion in INARCH(1) processes. On the other hand, the less familiar problem of underdispersion has been considered by Del Castillo and Pérez-Casany (1998) who also proposed the weighted Poisson distribution for fitting underdispersion in count data, Faddy and Bosch (2001) who modelled and analysed underdispersed data based on pure birth processes, Sellers *et al.* (2010) who also considered the COM-Poisson distribution when underdispersion occurs, and Weiß (2013) who presented integer-valued autoregressive models for underdispersed counts.

Many authors are also concerned about other issues commonly found in count distributions such as the inflation (or deflation), the truncation and the mixture of distributions, even if data are time-dependent or not. Accordingly, inflated and deflated count distributions have been studied, for example, by Greene (1994) where several modifications of both, the Poisson and Binomial distributions, were presented in order to accommodate zero-inflation, Böhning (1998) where several Zero-inflated Poisson models were reviewed in detail, and Ridout *et al.* (2001) where a score test for evaluating a Zero-inflated Poisson regression model against zero-inflated negative binomial alternatives was proposed. Additionally, David and Johnson (1952), Cohen (1954), Brass (1958), and Grogger and Carson (1991) focused on truncated count distributions. These works studied, among others, the Poisson and Negative Binomial truncated models, and some methods of parameter estimation. Mixtures of distributions of counts have been considered by Blischke (1964, 1965) who described moment-based methods for estimating parameters of a mixture of two Binomial distributions and a mixture of Normal distributions, Barndorff-Nielsen (1965) who studied the identifiability of mixtures of exponential families, and Karlis and Xekalaki (1998, 1999) who introduced improvements of the EM algorithm for finite Poisson mixtures.

A further difficulty appears when dealing with count data correlated over time. The analysis of count time series has been rapidly growing in the past year, and many authors have been contributed to its constant improvement, introducing interesting works. Some examples can be found in McKenzie (1985, 2003) who introduced the INAR(1) model, Alzaid and Al-Osh (1990) and Du and Li (1991) who presented different INAR(p) models, Jung and Tremayne (2006) and Weiß (2008) who proposed comprehensive reviews based on INAR models, Zucchini and MacDonald (2009) who described Hidden Markov models for count time series, Moriña *et al.* (2011) who proposed an INAR(2) model for hospital admissions considering seasonal effects, and Weiß and Puig (2015) who studied the marginal distribution of compound Poisson INAR(1) models.

Currently, researchers and data analysts are paying particular attention to data quality, being one of the more critical issues in data analysis, and hence in count data analysis. According to Oliveira *et al.* (2005), a formal definition of data quality comprises different meanings and interpretations. A complete characterization of data quality may cover concepts related to the accessibility, relevancy, believability, objectivity, and interpretability of data (Oliveira *et al.* 2005). Without forgetting that all these items are equally crucial for seeking to ensure quality data, this Ph.D. thesis is focused on the believability of data, and especially on the phenomenon of under-reporting. Formally, under-reporting refers to some incident which is responsible for reporting less than the actual level of count data, meaning that the believability of data significantly decreases when the under-reporting is present. Some authors have dealt with this phenomenon. For example, Winkelmann (1996) considered a Markov chain Monte Carlo-based methodology to study under-reporting in worker absenteeism data, Rosenman *et al.* (2006) performed a capture-recapture analysis in order to estimate the number of missed cases of work-related injury and illness in Michigan, Park *et al.* (2011) evaluated the global magnitude of reported and under-reported mesothelioma, Gamado *et al.* (2014) studied the effect of under-reporting in epidemics through stochastic epidemic models, and Crowcroft *et al.* (2018) also used capture-recapture methods to estimate the under-reporting of pertussis in Ontario.

Finally, several authors have focused their attention on different topics related to count data: multivariate data (Shanbhag and Rajamannar 1974, Bishop *et al.* 1975, and Johnson *et al.* 1997), spatiotemporal data (Bauer *et al.* 2016, and Neelon *et al.* 2013), over-reported data (Hofman 2013) and missing values (Kaciroti *et al.* 2008, Al-Osh 2009, and Mian and Paul 2016), among others.

When count data analysis is conducted, suitable techniques and methods should be used if some of the above issues (overdispersion and underdispersion, inflation and deflation, truncation, data quality, etc.) might be present in data. Using inappropriate techniques and/or ignoring the presence of some of the above matters, results can likely lead to completely wrong and nonsense conclusions which, in extreme cases, can lead to disproportionate consequences. Even though many authors have contributed to improving methods for count data analysis, further work should be conducted for continually improving the current techniques and providing more appropriate and realistic approaches.

Accordingly, the present thesis is aimed at introducing novel methods and techniques of count data analysis to deal with some of the issues that have been previously described. In this sense, this thesis comprises different papers where novel methods have been presented. In particular, the first and second papers (Fernández-Fontelo *et al.* 2016, and Fernández-Fontelo *et al.* submitted) are focused on the assessment of the under-reporting phenomenon in count time series. Two realistic models are proposed on the basis of the classical count time series models. Real-data applications within different contexts are discussed to show the practicality of these models. The third paper (Fernández-Fontelo *et al.* 2017) is based on a general model of count time series, which considers moderate overdispersion, even if a series is non-stationary. This new model has been applied to the analysis of data of fallen cattle registered at a local scale when series have low counts, many zeros, and slightly overdispersion as part of a project commanded by the Ministry of Agriculture, Food and Environment of Spain. In the fourth paper (Fernández-Fontelo *et al.* 2018), an exact goodness-of-fit test is presented for detecting zero-inflation and zero-deflation in count distributions within the biological dosimetry framework. This test was firstly introduced by Rao and Chakravarti (1956) derived from the problems of occupancy, although we have taken to the biological dosimetry analysis the idea behind the original test. This test is viewed as a complement to the always used u-test when data are not overdispersed or underdispersed, but they may be zero-inflated or zero-deflated.

This thesis is organized as follows. Chapter 1 is an overall presentation of the methodological results of the papers, and also some of their applications to real-world data. A detailed discussion of the importance of these results to improve count data analysis is also included in this chapter. Chapter 2 proposes a definite conclusion based on the achievements of the thesis and some further research. Finally, Chapters 3, 4, 5 and 6 include the manuscripts of the papers mentioned above.

The present thesis was founded by the grants MTM2012-31118, MTM2015-69493-R from the Ministry of Economy and Competitiveness of Spain, MDM-2014-0445 "María de Maeztu" Programme for Units of Excellence in R&D from the Ministry of Economy and Competitiveness of Spain, and the National Sciences and Engineering Research Council from the Government of Canada.

CHAPTER 1

Overall presentation of the results and discussion

This chapter is aimed at presenting and discussing the primary results of the papers previously mentioned. In particular, the chapter consists of the following four sections: (1.1) a section describing the new models of count time series for under-reporting data proposed in Fernández-Fontelo *et al.* (2016) and Fernández-Fontelo *et al.* (submitted); (1.2) a section introducing the HINAR(p) model proposed in Fernández-Fontelo *et al.* (2017); (1.3) a section presenting an exact test for the Poisson distribution focused on the zero-inflation and zero-deflation of data, which was firstly introduced by Rao and Chakravarti (1956) and later on suggested by Fernández-Fontelo *et al.* (2018) in the biological dosimetry context; and finally (1.4) a section discussing the primary results of the thesis, focusing on the influence of these methods in the future count data analysis.

1.1 New models to analyse under-reported count data through INAR(1)-hidden processes

The interest in count time series analysis has been rapidly increasing in the past years as a result of the limited performance of the classical time series analysis when dealing with discrete-valued time series, especially when these series have low counts and many zeros.

These models have become more popular in the literature where many useful models have been applied to different fields such as epidemiology and public health (Allard 1998, Cardinal *et al.* 1999, and Moriña *et al.* 2011), finance (Brännäs and Hellström 2001, Pedeli *et al.* 2011, and Weiß and Kim 2013), environment (Pavlopoulos and Karlis 2008), among others.

Furthermore, many authors have considered trend and/or seasonal components of non-stationary series as well as the problem of heterogeneity (Gourieroux and Jasiak 2004, Monteiro *et al.* 2010, and Moriña *et al.* 2011). Unfortunately, although many series are undoubtedly under-reported, this issue has not been considered yet in count time series analysis. Some of the reasons which explain this under-reporting fact include the accuracy of public health registers, political or economic interests and social issues and stigmas.

Accordingly, in this section, new models of count time series for under-reporting data are briefly introduced. Full details can be found in Chapter 3 (Fernández-Fontelo *et al.* 2016) and Chapter 4 (Fernández-Fontelo *et al.* submitted).

1.1.1

The models

Let X_n be a hidden process following an INAR(1) model which satisfies the following stochastic structure:

$$X_n = \alpha \circ X_{n-1} + W_n, \tag{1.1}$$

where the parameter $0 < \alpha < 1$ is fixed, and the operator $\circ$ is the well-known binomial thinning or binomial subsampling such that $[\alpha \circ X_{n-1} | X_{n-1} = x_{n-1}] =_d \sum_{i=1}^{x_{n-1}} \{\mathrm{B}_i(\alpha)\}$, where $\{\mathrm{B}_i(\alpha)\}$ is a sequence of independent and identically distributed Bernoulli(α) random variables. This binomial thinning operator ensures the integer discreteness of the series. The innovations of the model, that is, W_n, are independent and identically distributed following a Poisson(λ) distribution. The expectation and variance of X_n are $\mu_X = \mathrm{E}(X_n) = \mathrm{Var}(X_n) = \lambda/(1-\alpha) = \sigma_X^2$, respectively. Additionally, if the process X_n has a finite moment of order two, then every stationary INAR(1) model takes the following auto-covariance function $\gamma_X(k) = \mathrm{Cov}(X_n, X_{n+k}) = \alpha^k \sigma_X^2$ and hence, the auto-correlation function of X_n is $\rho_X(k) = \gamma(k)/\gamma(0) = \alpha^k$ (geometrically decreasing at rate α). See, for instance, McKenzie (2003) and Weiß (2008).

Let Y_n be an observed and potentially under-reported process, and I_n be a binary process such that I_n is an indicator of whether the observation Y_n is under-reported or not. The process Y_n satisfies:

$$Y_n = \begin{cases} X_n & \text{with probability } 1-\omega \\ q \circ X_n & \text{with probability } \omega, \end{cases} \tag{1.2}$$

where $0 < \omega < 1$ and $0 < q < 1$. The definition in (1.2) means that, at time n, the observed process Y_n coincides with the hidden process X_n with probability $1-\omega$, and then the process is not under-reported ($I_n = 0$). Otherwise, the observed process Y_n is a binomial thinning of the hidden process X_n with probability ω, and then the process is under-reported ($I_n = 1$). Parameters ω and q can be interpreted, respectively, as the overall frequency and intensity of the under-reporting phenomenon. The closer to one and zero are ω and q, the more frequent and intense is the under-reporting issue in data, respectively.

According to the expression (1.2), we proposed two different models: (a) the "full model" in Fernández-Fontelo *et al.* (submitted) which assumes a dependence structure between the states of under-reporting through a binary discrete-time Markov chain (Zucchini and MacDonald 2009); and (b) the "reduced model" in Fernández-Fontelo *et al.* (2016) which considers independence between the states of under-reporting.

1.1.1.1

Properties and parameter estimation

Under stationarity, the marginal distribution of both models, the "full model" and the "reduced model", is a mixture of two Poisson distributions with parameters $\frac{\lambda}{1-\alpha}$ with probability $1-\omega$, and $\frac{q\lambda}{1-\alpha}$ with probability ω. Accordingly, the expectation and variance of Y_n are $\mathrm{E}(Y_n) = \mu_X\,(1-\omega(1-q))$, and $\mathrm{Var}(Y_n) = \mu_X^2\,(\omega(1-\omega)(1-q)^2) + \sigma_X^2(1-\omega(1-q^2) + \mu_X\omega q(1-q))$, respectively.

The expression of the auto-correlation function (ACF) is slightly different depending on the model. The ACF of the "full model" can be written as follows:

$$\rho_Y(k) = \frac{\alpha^k(1-\omega(1-q))^2 + \lambda_2^k\mu_X\omega(1-\omega)(1-q)^2 + (\alpha\lambda_2)^k\omega(1-\omega)(1-q)^2}{\mu_X\omega(1-\omega)(1-q)^2 + (1-\omega(1-q))}, \tag{1.3}$$

where $\lambda_2 = 1 - p_{01}/\omega$ and p_{01} is the probability of going from the state of not under-reporting to the state of under-reporting. It is important to highlight that when $\omega = p_{01}$ the expression (1.3) results in the structure of the ACF of the "reduced model". Full details related to the computation and interpretation of the ACF of both models can be found in the corresponding papers in Chapters 3 and 4.

In the previously mentioned chapters, the authors introduce two different methods for estimating the parameters of the models: moment-based and likelihood-based methods. The first one is mainly based on the marginal distribution of the process Y_n (a mixture of two Poisson distributions), and the theoretical expression of the ACF (1.3). The second one is based on the likelihood function of the model which is computed using the forward algorithm (see, for example, Zucchini and MacDonald 2009) since the direct computation of the likelihood function is not tractable. The proposed algorithms for estimating parameters by the methods of moments and maximum likelihood are broadly described in Chapters 3 and 4.

1.1.2 Applications to Public Health data

It is well-known that many public health issues are entirely under-reported, and those include diseases related to occupational or food exposures (Rosenman *et al.* 2006, Alfonso *et al.* 2015, and Arendt *et al.* 2013), several sexually transmitted infections (Duron *et al.* 2018), many social phenomenon regarding violence and abuses like gender-based violence (Watts and Zimmerman 2002, Gracia 2004, and Palermo 2014) and alcohol or drug abuses (Holmes *et al.* 2012, and McGregor *et al.* 2003). As a result, their statistics are profoundly mistaken, leading to underestimating their actual magnitudes.

To better estimate the scope of these issues and keep the under-reporting phenomenon under control, the models proposed in this thesis have been applied to the following datasets: (1) weekly new diagnosis of human papillomavirus between 2010 and 2014 in Girona (Spain); (2) annual deaths by pleura and peritoneal mesotheliomas between 1968 and 2013 in Great Britain; (3) annual number of botulism cases between 1970 and 2013 in Canada; and (4) quarterly complaints of domestic violence against women recorded in different judicial districts of Galicia (Spain) between 2007 and 2017.

The new diagnoses per week of human papillomavirus in Girona are undoubtedly under-reported since both, the frequency ($\widehat{\omega} = 92.2\%$) and intensity ($\widehat{q} = 32.6\%$) of the under-reporting phenomenon, are statistically significant. Although an average of 1.27 cases per week is observed within the period, a real average of 3.36 cases per week is estimated. This last means that approximately 1 out of 3 (1.27/3.36, 38%) weekly new diagnoses of human papillomavirus has been adequately diagnosed and officially recorded in Girona from 2010 to 2014.

On the other hand, both series, the annual deaths by pleura and peritoneal mesotheliomas in Great Britain and the annual cases of botulism in Canada, are recorded in large periods of time (1968-2013 and 1970-2013, respectively), where population rapidly increased. As a result, positive trends are detected in these series, and then covariates related to the annual population sizes in Great Britain and Canada are included in both series models. In particular, the population sizes (N) are considered in the models as covariates through the following link-function: $\lambda = e^{aN}$, where a is a parameter to be estimated, and λ is the parameter of the Poisson distribution in equation (1.1). The phenomenon of under-reporting is evident in both series which parameters of frequency ($\widehat{\omega} = 93.0\%$ and $\widehat{\omega} = 67.1\%$, respectively) and intensity ($\widehat{q} = 51.7\%$ and $\widehat{q} = 31.7\%$, respectively) are also statistically significant.

Finally, the phenomenon of under-reporting in gender-based violence data is studied in 35 out of 45 judicial districts in Galicia (Spain) where the number of quarterly complaints is officially record. Both, the frequencies and intensities of the under-reporting phenomenon, are statistically significant in each judicial district, demonstrating a severe lack of reporting in the number of complaints of gender-based abuse in Galicia. In particular, frequencies of under-reporting are unfortunately high in the 35 judicial districts of Galicia, while the intensities of this issue are significantly higher in the rural judicial districts compared with the urban judicial districts.

Further details can be found in Chapters 3 and 4 where the results of these examples of application are described in detail as well as some relevant interpretations in the public health context.

1.2

The HINAR(p) model

The autoregressive moving average models (ARMA) are the most commonly used models when dealing with continuous time series, or even discrete time series which can be appropriately approximated by a Normal distribution. Nevertheless, when series have low counts and/or many zeros, these models are entirely inappropriate. In this regard, many authors have proposed several alternatives to these models when series are non-negative integer-valued and cannot be approximated to continuous time series.

The abovementioned INAR(1) model (1.1) was the first one proposed to deal with low-valued count data with many zeros. However, some generalizations of this model have been introduced by the literature in the past years. For instance, a natural extension of the INAR(1) model is the INAR(p) model which was firstly introduced by Alzaid and Al-Osh (1990), and later on, by Du and Li (1991). This INAR(p) process is restricted to model both stationary and equidispersed time series. As a result, several authors have proposed new models which consider overdispersed series (Jazi *et al.* 2012a, Jazi *et al.* 2012b, and Zhu and Joe 2006), or non-stationary series (Moriña *et al.* 2011). However, to the best knowledge of the authors, nobody has proposed a version of the INAR(p) model that would consider overdispersion and non-stationarity at once.

Accordingly, in this section, a more general INAR(p) model is presented. The HINAR(p) model, which was firstly introduced in Fernández-Fontelo *et al.* (2017), considers series with moderate overdispersion and also accommodates trend and seasonal components of non-stationary series through representative covariates introduced in the model.

This study can be found in Chapter 5.

The model

Alzaid and Al-Osh (1990) and Du and Li (1991) introduced the following INAR(p) model which is defined by the equation:

$$X_n = \alpha_1 \circ X_{n-1} + \alpha_2 \circ X_{n-2} + \cdots + \alpha_p \circ X_{n-p} + W_n, \tag{1.4}$$

where $0 < \alpha_1, \alpha_2, \ldots, \alpha_p < 1$ are fixed parameters and W_n is a sequence of independent and identically distributed Poisson(λ) random variables (innovations). Notice that when $p = 1$, this model results in the INAR(1) model introduced in (1.1). As mentioned above, the binomial thinning operator $[\alpha_j \circ X_{n-j} | X_{n-j} = x_{n-j}] =_d \sum_{i=1}^{x_{n-j}} \{\mathrm{B}_i(\alpha_j)\}$, where $\{\mathrm{B}_i(\alpha_j)\}$ are independent and identically distributed random variables with P$(B_j(\alpha_j) = 1) = \alpha_j$, which ensures the discreteness of the series.

While Alzaid and Al-Osh (1990) assumed that the joint distribution of $(\alpha_1 \circ X_{n-1}, \alpha_2 \circ X_{n-2}, \ldots, \alpha_p \circ X_{n-p})$, when $p > 1$, is conditional Multinomial, Du and Li (1991) imposed conditional independence. The second approach is considered in our work because is more tractable and interpretable in practice.

The novel extension of the conventional INAR(p) model proposed by Fernández-Fontelo *et al.* (2017) (Chapter 5) takes the following expression:

$$X_n = \alpha_1(n) \circ X_{n-1} + \alpha_2(n) \circ X_{n-2} + \cdots + \alpha_p(n) \circ X_{n-p} + W_n(a_1(n), a_2(n)), \tag{1.5}$$

where the innovations follow a 2th-Hermite distribution with paramteres $a_1(n)$ and $a_2(n)$, and the parameters of the model (including those of the 2th-Hermite distribution) can be time-dependent. This model is called HINAR(p) where $p \geq 1$.

A 2th-Hermite distribution originates from the following linear combination: $Y = X_1 + 2X_2$, where X_1 and X_2 are two independent Poisson distributions with parameters a_1 and a_2, respectively. The dispersion index of this distribution satisfies $1 \leq \delta \leq 2$, thus the 2th-Hermite distribution is slightly overdispersed compared with the standard Poisson distribution.

Parameters of the model (1.5) can be time-dependent by specifying appropriate link functions with covariates related to trend and seasonal components. Particularly, trend and seasonal covariates are accommodated in the model by a second-order trigonometric polynomial with a linear part. A suitable function is needed to link these covariates to parameters. Accordingly, the chosen link function for parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ takes the following expression:

$$\begin{aligned}\text{logit}(\alpha_j(n)) = \log\left(\frac{\alpha_j(n)}{1-\alpha_j(n)}\right) &= \beta_{j0} + \beta_{j1}n + \beta_{j2}\sin\left(\frac{2\pi n}{T_1}\right) + \beta_{j3}\cos\left(\frac{2\pi n}{T_1}\right) \\ &+ \beta_{j4}\sin\left(\frac{2\pi n}{T_2}\right) + \beta_{j5}\cos\left(\frac{2\pi n}{T_2}\right), \qquad (1.6)\end{aligned}$$

while the chosen link function for parameter a_1 and a_2 (2th-Hermite distribution) would take the following:

$$a_j(n) = \exp\left(\gamma_{j0} + \gamma_{j1}n + \gamma_{j2}\sin\left(\frac{2\pi n}{T_1}\right) + \gamma_{j3}\cos\left(\frac{2\pi n}{T_1}\right) + \gamma_{j4}\sin\left(\frac{2\pi n}{T_2}\right) + \gamma_{j5}\cos\left(\frac{2\pi n}{T_2}\right)\right), \qquad (1.7)$$

In both functions, (1.6) and (1.7), parameters γ_{j1} and β_{j1} capture the possible trend effect, while the other parameters capture the seasonal components of periods T_1 and T_2. In this work, it is considered that $T_1 = 52$ weeks and $T_2 = 26$ weeks due to the nature of data.

Notice that both functions in expressions (1.6) and (1.7) fall in the corresponding domain of the parameters $0 < \alpha_1, \alpha_2, \ldots, \alpha_p < 1$ and $a_1, a_2 \geq 0$, respectively.

Notice also that when parameters are not time-dependent and the parameter $a_2 = 0$, the 2th-Hermite distribution results in the Poisson distribution with parameter a_1. Hence, the model in (1.5) becomes the model in (1.4). In other words, the HINAR(p) model contains the conventional INAR(p) model.

Full details on model specification and properties can be found in Chapter 5.

1.2.1.1 Parameter estimation and forecasting

Parameters of the model $\Theta = (\beta_1, \beta_2, \ldots, \beta_l, \gamma_1, \gamma_2, \ldots, \gamma_m)$ are estimated through the maximum likelihood method.

The authors in Fernández-Fontelo *et al.* (2017) present the following expression of the conditional probability density function (CPDF) of an HINAR(p) model like in (1.5).

Given a sample $\{x_1, x_2, \ldots, x_N\}$, the CPDF of the HINAR(p) model is:

$$P(x_n|x_{n-1}, \cdots, x_{n-p}; \Theta) = e^{-(a_1(n)+a_2(n))} \sum_{i_1=0}^{m_1} \binom{x_{n-1}}{i_1} \alpha_1^{i_1}(n)\,(1-\alpha_1(n))^{x_{n-1}-i_1}$$
$$\cdot \sum_{i_2=0}^{m_2} \binom{x_{n-2}}{i_2} \alpha_2^{i_2}(n)\,(1-\alpha_2(n))^{x_{n-2}-i_2} \cdot \cdots$$
$$\cdot \sum_{i_p=0}^{m_p} \binom{x_{n-p}}{i_p} \alpha_p^{i_p}(n)\,(1-\alpha_p(n))^{x_{n-p}-i_p} \sum_{j=0}^{\left[\frac{M}{2}\right]} \frac{a_1(n)^{M-2j} a_2(n)^j}{(M-2j)!j!}, \tag{1.8}$$

where $M = x_n - i_1 - i_2 - \cdots - i_p$, $m_1 = \min(x_n, x_{n-1})$, $m_2 = \min(x_{n-2}, x_n - i_1), \cdots$, $m_p = \min(x_{n-p}, x_n - (i_1 + \cdots + i_{p-1}))$.

Given the expression (1.8), the likelihood function can be efficiently computed.

All details of the proof can be found in Chapter 5.

Different predictions are considered: (1) the average behaviour of the series; and (2) the k-time-ahead distribution when dealing with an HINAR(1) (case $p = 1$). On the one hand, to forecast the average behaviour of the series, an extension of the methodology described in Moriña *et al.* (2011) is applied. On the other hand, to forecast the k-time-ahead distribution, the following new result is introduced.

Given n observations $x_1, x_2, \ldots, x_n$, to make predictions at k time-ahead based on the HINAR(1) model, the distribution of X_{n+k} is required.

Easily, the distribution of X_{n+2} can be obtained by replacing in the expression $X_{n+2} = \alpha(n+2) \circ X_{n+1} + W(a_1(n+2), a_2(n+2))$ by $X_{n+1} = \alpha(n+1) \circ X_n + W(a_1(n+1), a_2(n+1))$. That is,

$$X_{n+2} = (\alpha(n+1)\alpha(n+2)) \circ X_n + W(\alpha(n+2)(a_1(n+1) + 2(1-\alpha(n+2))a_2(n+1))$$
$$+ a_1(n+2), \alpha^2(n+2)a_2(n+1) + a_2(n+2)), \tag{1.9}$$

because of the property of clousure under convolution of the Hermite distribution. Notice that the distribution of X_{n+2} is determined through the last observed value of the series X_n.

When $k > 2$, the previous recursive procedure can be extended leading to the following result:

$$X_{n+k} = f(n+k) \circ X_n + W\left(c(n+k), d(n+k)\right), \ k = 1, 2, ... \tag{1.10}$$

where

$$f(n+k) = \prod_{i=1}^{k} \alpha(n+i) \tag{1.11}$$

and

$$\begin{aligned} c(n+k) &= \alpha(n+k)(c(n+k-1) + 2d(n+k-1)(1-\alpha(n+k))) \\ &\quad + a_1(n+k) \end{aligned} \tag{1.12}$$

$$d(n+k) = \alpha^2(n+k)d(n+k-1) + a_2(n+k)\,,\; k = 2, 3, ..., \tag{1.13}$$

where $c(n+1) = a_1(n+1)$ and $d(n+1) = a_2(n+1)$. As before, this result can be obtained because of the property of clousure under convolution of the Hermite distribution.

By replacing parameters in equations (1.10), (1.11), (1.12) and (1.13) by their maximum likelihood estimates, regions of prediction of size $1-\alpha$ can be estimated.

See Chapter 5 for a detailed description.

1.2.2 Application to dairy and beef cattle data

The analysis of fallen stock data of cattle has been shown to be a good potential indicator of animal health surveillance (Alba *et al.* 2015). When the baseline patterns of mortality are described and estimated, a suspected outbreak peak can be detected. This unusual event can be strongly related to some animal health issues.

Many authors have studied animal mortality data from large populations using different approaches such as classical time series analysis (Alba *et al.* 2015), or survival analysis (Tapprest *et al.* 2017), among others. However, in many cases, these fallen stock data of cattle should not only be studied at a large scale, but also at a local scale because an outbreak may also occur at a smaller geographical level. Unfortunately, when data are analysed at a large scale, outbreaks at local scales could be completely invisible.

Sometimes, mortality data of cattle collected at a local scale have low values with many zeros. In these situations, for instance, the classical time series analysis is inappropriate, leading to nonsense results.

To accurately describe and estimate the baseline patterns of fallen stock data of cattle at a local scale, the HINAR(p) model is proposed. This model allows accommodating overdispersed series with low counts and many zeros. In particular, two different series are modelled: (1) a series of fallen stock in a small beef cattle population in Spain between 2007 and 2011; and (2) a series of fallen stock in a small dairy cattle population in Spain between 2007 and 2011.

Both series of beef and dairy cattle are slightly overdispersed with empirical dispersion indices of 1.93 and 1.77, respectively. Additionally, series of beef and dairy cattle range from 0 to 9 and 0 to 7, respectively. Hence, the HINAR(p) model seems to be an appropriate novel alternative for modelling these series.

The series of beef cattle is modelled by means of an HINAR(1) model with $\alpha_1(n) = \alpha_1$, and a decreasing trend included in the Hermite parameter a_1 through the link function $a_1(n) = \exp(\gamma_1 n)$. In particular, the maximum likelihood estimates of the parameters are $\widehat{\alpha_1} = 0.104$, $\widehat{\gamma_1} = -0.004$ and $\widehat{a_2} = 0.306$.

On the other hand, the series of dairy cattle is modelled using an HINAR(1) model, where annual seasonality is included in both parameters α_1 and a_2 through the following link functions: logit $(\alpha_1(n)) = \beta_0 + \beta_2 \sin(2\pi n/52)$ and $a_2(n) = \exp(\gamma_0 + \gamma_3 \cos(2\pi n/52))$. The maximum likelihood estimates of these parameters are $\widehat{\beta_0} = -3,894$, $\widehat{\beta_2} = -3.304$, $\widehat{a_1} = 0.388$, $\widehat{\gamma_0} = -1.540$ and $\widehat{\gamma_3} = 0.698$. In both models, the parameters are statistically significant.

Additionally, these models are selected based on several criteria described in Chapter 5.

1.3

The CR-test for the Poisson model

The goal in biological dosimetry is to estimate the dose of radiation that a suspected irradiated individual has received by using chromosome damage in peripheral lymphocytes as biomarkers of exposure (*i.e.* dicentric chromosome aberrations with/without rings). In particular, these doses of radiation are estimated through the response calibration curves which are created by exposure of human blood cells to different and appropriate radiation doses.

A suspected irradiated individual can be exposed to whole or partial body irradiation.

Commonly, in whole body irradiation (WBI) under low-linear energy transfer (LET) radiation exposures, the number of observed dicentrics per cell (chromosome aberrations) is Poisson distributed, whose rate depends on the dose considering a linear-quadratic function with identity link. However, in partial body irradiation (PBI), which occurs when only a fraction of the body is irradiated to an homogeneous dose, the Poisson distribution is completely inappropriate.

When PBI occurs, the number of observed cells is a mixture of a Poisson distribution and structural zeros, that is, a Zero-inflated Poisson distribution. The distribution of the number of chromosome aberrations of the non-irradiated scored cells provides an excess of zeros, comparing with the distribution of those aberrations producing in an homogeneous WBI. As a result, the distribution of the number of chromosome aberrations under PBI exposures is always overdispersed and zero-inflated.

According to the recommendation of the International Atomic Energy Agency (IAEA 2011), the well-known u-test should be used to detect PBI. This test determines whether the ratio of the sample variance to the sample mean (the sample dispersion index) is significantly different from 1 (dispersion index of the Poisson distribution). The u-test only studies the dispersion of the data, indicating PBI exposure when these data are overdispersed (dispersion index > 1). Nevertheless, other features of the data such as the zero-inflation can lead to rejection of the hypothesis of WBI, when the u-test does not reject this hypothesis. In other words, WBI exposures can be rejected due to the overdispersion and/or the zero-inflation of the data.

It is important to remark that there are other possible causes, apart from PBI, producing overdispersion and zero-inflation in data. For instance, this is the case for whole body low-LET-irradiation from different doses (heterogeneous exposures), which can be modelled using a mixed-Poisson distribution. On the other hand, when whole body high-LET-irradiation occurs, data can be overdispersed and/or zero-inflated. As a result, these data can be modelled through a Compound Poisson distribution.

In this thesis, we suggest using an exact zero-inflation goodness-of-fit test for the Poisson distribution in the biological dosimetry context. This test, that the authors term CR-test, was firstly proposed by Rao and Chakravarti (1956) based on the theory of the occupancy problems. Additionally, to demonstrate the usefulness and necessity of this test in biological dosimetry, five examples based on both, *in vitro* and *in vivo* data, are described and discussed.

1.3.1

The test

In Rao and Chakravarti (1956), the authors proposed the CR-test for the Poisson distribution based on the following experiment within the occupancy problems: consider n boxes and S balls. Balls are randomly distributed between boxes with the same probability, that is, $1/n$. The authors were interested in the random variable N_0 which denotes the number of empty boxes. As a result, in the previously mentioned article, an exact expression for computing these probabilities are presented and demonstrated in detail.

In this thesis, the authors translate the idea behind the previous experiment to the biological dosimetry framework. Accordingly, consider that $X = (X_1, X_2, \ldots, X_n)$ is a dosimetry sample where X_j is the number of aberrations (dicentrics with/without rings) found in the cell j for $j = 1, \ldots, n$. Then, $S = \sum_{j=1}^{n} X_j$ is the total number of aberrations found in a sample of cells. Dicentrics are randomly distributed between cells with the same probability, that is, $1/n$. The random variable N_0 denotes the number of cells free of aberrations.

As commented above, deviations from the Poisson distribution can lead, for instance, to rejection of the hypothesis of WBI in low-LET exposures. One of these departures comes from the excess number of zeros (zero-inflation) which is frequently found in PBI exposures. The CR-test is especially focused on the random variable N_0 (number of cells free of aberrations), taking also into account the number of scored cells n and observed aberrations S. This test contrasts the null hypothesis H_0 : Data are Poisson distributed, against the alternative H_1 : Data are zero-inflated. The exact p-value of this test can be computed by means of the following expression:

$$\begin{aligned} P(N_0 \geq n_0) &= \sum_{i=n_0}^{n} \sum_{j=i}^{n} (-1)^{j-i} \binom{n}{j} \binom{j}{i} \left(1 - \frac{j}{n}\right)^S \\ &= \sum_{i=n_0}^{n} \frac{(-1)^{i-n_0}}{(n_0-1)!(j-n_0)!} \binom{n}{i} \left(1 - \frac{i}{n}\right)^S , \end{aligned} \tag{1.14}$$

where n_0 is the number of observed cells free of aberrations. The CR-test is especially interesting because it allows the researchers to detect the problem of zero-inflation in the data, which cannot be detected through the u-test (just testing overdispersion). Additionally, it is important to highlight that this exact test can be seen as to complement the widely used u-test, being especially useful in scenarios where the overdispersion is not present but the number of zeros is completely anomalous.

Moreover, when dealing with large values of n and S, exact computations based on the CR-test can be tedious since the factorial numbers intensify the calculation of the expression (1.14). As a result, in Chapter 6 an asymptotic approach of the CR-test is proposed. This asymptotic test is essentially the normalised version of the random variable N_0.

Full details can be found in Chapter 6. Also, in this chapter, a version of the CR-test for studying zero-deflation is proposed.

1.3.2

Examples of application

Several examples of application are studied to show the usefulness of the CR-test for detecting PBI exposures.

Two cases of *in vitro* samples exposed to WBI at 10 Gy (Sasaki 2003) and at 0.25 Gy (IAEA 2011) are analysed.

The first example consists of $n = 200$ cells, $S = 705$ aberrations and $n_0 = 3$ cells free of aberrations. As a result, these samples of aberrations seem to be underdispersed (dispersion index of 0.892) and zero-deflated (index $z_i = -0.191 < 0$, introduced by Puig and Valero 2006). However, while the CR-test (p-value $= 0.147$) shows that the sample is not zero-deflated, the u-test (p-value $= 0.044$) shows that the data are underdispersed. This last is because the distribution of the number of aberrations is complex, since the samples were exposed to a high dose, and some mechanism behind those distributions produces underdispersion, but not zero-deflation.

The second example consists of $n = 2008$ cells, $S = 22$ aberrations and $n_0 = 1987$ cells free of aberrations. Although slightly overdispersion is detected (dispersion index of 1.081), the zero-inflation is clearly not significant ($z_i = 0.04 \approx 0$). Finally, while the CR-test leads to non-rejection of the Poisson hypothesis (p-value $= 0.109$), the u-test leads to a significant overdispersion (0.005). This result agrees with the fact that the samples were exposed to WBI at a relatively low dose.

The other examples are based on *in vivo* data collected in different radioactive accidents.

The first one is based on the data from one of the workers involved in the Tokaimura (Japan) criticality accident (Hayata *et al.* 2001). A total of $n = 175$ cells are studied, identifying $S = 537$ aberrations (dicentrics plus rings) and $n_0 = 14$ cells free of aberrations. Both, the sample dispersion index (1.05) and the index $z_i = 0.177 > 0$, show that the sample can be overdispersed and zero-inflated. However, the u-test finally shows that the data are not overdispersed (p-value $= 0.329$), but the CR-test indicates that the data are zero-inflated (p-value $= 0.022$). This last means that the zero-inflation is detected (and not the overdispersion) because the CR-test is more sensitive to the changes in the frequency of zeros of the sample than the u-test, which is not as precise as the CR-test. As a result, the hypothesis of PBI has to be rejected since it can only be accepted when both tests, u and CR, agree. Of course, the WBI hypothesis is also rejected because the data are not Poisson distributed due to the unusual number of zeros in the samples.

The second one is based on the data from one of the workers affected in the radiation accident in Stamboliyski (Bulgaria) (Grégorie *et al.* 2013). After the accident, blood samples were taken and sent to the Institut de Radioprotection et de Sûreté Nucléaire (IRSN, Paris), and to the National Centre of Radiobiology and Radiation Protection (NCRRP, Bulgaria). However, for one of the exposed workers, contradictory results were provided. While the IRSN concluded that the individual was totally irradiated, the NCRRP concluded that that individual was partially irradiated. After the analysis of the data through the CR-test, data from ISRN seem to be useless due to some problems during the experiment (underdispersion is detected). Nevertheless, data from NCRRP is overdispersed, but not zero-inflated. This result leads to rejection of the WBI and PBI exposures, suspecting that the distribution of the data could be a mixture of Poisson distributions since the individual could be exposed to different radiation doses.

The last example is based on a 75-aged man patient exposed to Thorotrast (Sasaki *et al.* 1987). The data consist of $n = 500$ cells, $S = 15$ aberrations, and $n_0 = 486$ cells free of aberrations. These samples seem to be slightly overdispersed (dispersion index of 1.1) and zero-inflated ($z_i = 0.177 > 0$). As a result, the p-value of the CR-test indicates that the data are not zero-inflated (p-value $= 0.191$), but the p-value of the u-test indicates that the data are overdispersed. This last leads to rejection of both, the hypothesis of WBI and PBI. In other words, the overdispersion of these samples can be explained because this patient could be exposed to different doses of radiation, and a mixture of Poisson distributions is more appropriate for these data than the Poisson or Zero-inflated Poisson distributions.

A user-friendly Shiny application based on `R` language has been built to make available the use of the CR-test to all researchers. The app can be found through the link `http://asapps.bcamath.org:5053`.

1.4 Discussion

This thesis is aimed at providing innovative methods to deal with some of the most relevant issues in count data analysis. In particular, our primary objectives consist of: (1) dealing with the problem of under-reporting in count time series (Chapters 3 and 4); (2) introducing a new model of count time series which allows both, time-dependent parameters and moderate overdispersed data (Chapter 5); and (3) suggesting a test of goodness-of-fit which was firstly introduced by Rao and Chakravarti (1956) on the basis of the occupancy problems. This test is presented in the biological dosimetry framework (Chapter 6).

The models introduced in Chapters 3 and 4 allow the accommodation of under-reporting in count time series in very flexible ways. The INAR(1) structure for the underlying process as well as the different structures between the states of under-reporting (independence and dependence through a binary Markov chain) allow a natural interpretation of all the parameters, being general enough to suit many real phenomena.

These models quantify the phenomenon of under-reporting through the parameters of frequency ω and intensity q. Additionally, when the states of under-reporting are correlated, the parameter p_{01} plays a part in this quantification.

The usefulness of these models is demonstrated in the context of public health. Several examples based on real data are studied in detail in Chapters 3 and 4. Although these models were initially designed to deal with stationary series, the results of some of these examples show that they can be extended to other scenarios where trends are relevant. As a result, these trends were considered in the model through appropriate covariates in the innovations of the latent process, that is, the INAR(1) model. These models could also be generalized to describe more complex scenarios where both, covariates of trend and seasonality, are included in the models. Besides, the application to gender-based violence data shows that the phenomenon of under-reporting behaves completely different depending on the area where the data were collected (urban or rural), the average income per month, the poverty risk, among others.

The INAR(1) model was naturally extended to the INAR(p) model by Alzaid and Al-Osh (1990) and Du and Li (1991). During the past years, the INAR(p) model has been studied in detail and applied to different scenarios. The HINAR(p) model, proposed in Chapter 5, is a natural generalization of the simpler INAR(p) with Poisson innovations, allowing both, time-dependent parameters and slightly overdispersed data. Moreover, this new model deals with many count time series with low values and many zeros. These reasons make the HINAR(p) model more versatile than the simpler INAR(p) model.

The HINAR(p) model has been motivated by the study of outbreaks in mortality data collected in beef and dairy cattle farms at small levels. A mortality outbreak can be recorded in a wide geographical area or can be more localised in a small area. However, when such data are only studied at large scales, these outbreaks localised at small areas are overlooked. As a result, the HINAR(p) models are more appropriate than the classical ARMA models when dealing with data at a local scale. This HINAR(p) model allows describing the profiles of the series at local areas.

Completely different models are found depending on whether the mortality data were collected in beef or dairy cattle farms. While the series of beef have been fitted using an HINAR(1) model with a decreasing trend included in the innovations, the series of dairy cattle has been adjusted using an HINAR(1) model with an annual seasonality included in the α-parameter, and also in the innovations. These models have been selected based on different criteria, and their validations are performed through static and dynamic cross-validations based on the crude residuals.

Chapter 6 presents a test for the Poisson distribution which has been built based on the occupancy problems (Rao and Chakravarti 1956). The idea behind this test (CR-test) is translated into the biological dosimetry framework.

When radiation exposure occurs, biological dosimetry practitioners aim to identify the proportion of undamaged lymphocytes coming from the bone marrow and to know what the minimum of these undamaged lymphocytes is necessary to recover from overexposure. This last cannot be directly achieved via goodness-of-fit tests for the Poisson distribution, but reasonably they are relevant to provide a correct diagnosis for those overexposed individuals.

The CR-test is proposed as a complement of the classical u-test to study the possible zero-inflation and zero-deflation in scored sample cells, and thus to identify exposure patterns (*i.e.* whole or partial body irradiation). Additionally, a reasonable approximation of this exact test is presented, especially when dealing with large sample sizes.

Several examples of applications based on both, *in vitro* and *in vivo* data, are thoroughly studied. Discrepancies between results from the CR-test and the u-test are found, showing that the CR-test can correctly work as a complement of the u-test in different data sets.

CHAPTER 2

Conclusions and further research

Despite the large number of novelty methodologies introduced in the past years to cope with count data, many issues in this field remain to be addressed. Necessarily, the researchers should be mainly focused on these issues to develop valuable tools and techniques to deal with them, and finally be able to provide better-quality analysis results.

Accordingly, the primary objective of this Ph.D. thesis consists of introducing innovative methodologies to understand some of the most frequent phenomena in count data. Moreover, this work does not only show the theory behind these methodologies, but also analyses real-world examples demonstrating that these new methods can work better than those previously known.

The methodologies proposed in this thesis can be considered small but relevant signs of progress in count data analysis. They allow studying several issues of count data from entirely different points of view, showing especially good results when dealing with some real-world problems in the public health and biological dosimetry frameworks. However, considerably more efforts have to keep doing to improve the existing techniques and tools continually, and thus being able to understand, shortly, many of the currently under-studied phenomena in count data.

The present research can be extended in different promising ways.

Firstly, new models of count time series to cope with under-reporting data could be studied by considering different latent processes. These new models could combine generalized or random thinning operators, and a wide range of appropriate distributions for the innovations of the process (Zhu and Joe 2006, and Zhu and Joe 2010). Moreover, time-dependent parameters could be included in these models when series are non-stationary. These promising generalizations can provide relevant and interesting works since the under-reporting in public health data is, unfortunately, one of the most frequent causes of poor statistics.

From a public health point of view, almost all the diseases in the world are under-reported for different reasons. This issue under-estimates the magnitude of these diseases in the entire world, motivating the use of inappropriate public health policies.

Then, many of the recent count time series models are not included yet in any common-used software. In this sense, a complete `R` package considering both, some of the simpler models and those proposed in this thesis, could be built using a user-friendly environment. This package would be a useful tool for all the researchers who are not especially familiar with complex programming languages.

Finally, considerably more sophisticated models and methodologies could be introduced by exploring new thinning operators as well as extending those existing. For example, considering new versions of the generalized thining operators introduced by Zhu and Joe (2010), and extending them to INAR(p) models. Additionally, it would be interesting to introduce a count time series model with an AR(1) structure allowing negative correlation. From an epidemiological point of view, this kind of model could provide very interesting results in the area, helping the researchers to understand better the phenomena where the temporal correlation structure is kind of negative, and also accurately estimate it.

CHAPTER 3

Under-reported data analysis with INAR-hidden Markov chains

This chapter corresponds to the contents of: Fernández-Fontelo, A., Cabaña, A., Puig, P. and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*; **35(26)**: 4875-4890. This paper is available through the link:

`https://onlinelibrary.wiley.com/doi/10.1002/sim.7026`

CHAPTER 4

Count time series models with under-reported data for gender-based violence in Galicia (Spain)

This chapter corresponds to the contents of Fernández-Fontelo *et al.* (submitted).

Abstract

Under-reporting in gender-based violence data is a worldwide problem leading to underestimation of the magnitude of this social scourge. This problem can degrade the data quality, providing poor and biased results which lead society to misunderstand the actual scope of this social issue. The present work proposes time series models for under-reported counts based on a latent INAR(1) time series with Poisson distributed innovations and a latent under-reporting binary state that is a first order Markov chain. The most relevant theoretical properties of the models are derived, and the method of moments and maximum likelihood are presented for parameter estimation. The new time series models are applied to quarterly complaints of domestic violence against women recorded in each judicial districts of Galicia (Spain) between 2007 and 2017. The models allow for quantifying of the amount of under-reporting. A comprehensive discussion is presented,

studying how the frequency and intensity of under-reporting relates to socio-economic indicators of the provinces of Galicia.

4.1 Introduction

Gender-based violence is a worldwide problem representing an important violation of women's rights and also constituting a risk factor for developing many physical, physiological and mental health problems. Besides, this is not only a problem of gender inequalities reflected in societies around the world, but also serves for demonstrating the control exercised from man over women (*i.e.*, intimate partners often use domestic violence to demonstrate their superiority in households or relationships) (Watts and Zimmerman 2002). The latter is a sign of serious behavioral problems and/or possible mental disorders of those individuals who want to exert control over women using violence. Accordingly, Gracia (2004) defined domestic violence against women as an important problem of public health, although, unfortunately, most of the corresponding journals and media do not handle it as an actual problem of public health.

The Spanish newspaper *El Pais* published in 2015 that 12.5% of 16-aged Spanish women have been gender-based victims in their lifetimes. 45% of them have gone to the doctor or social services, almost once during the traumatic period, and 29% of them have visited a professional such as a psychologist or psychiatrist.

According to the Article 1 of the Declaration on the Elimination of Violence against Women by the United Nations, violence against women is *"any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women."* (United Nations 1993).

One of the main worldwide issues when dealing with gender-based violence is that the number of official recorded cases seems to be far from the actual number of cases (due to under-reporting), since many victims do not provide information about their condition. This means that, in the entire world, official prevalences of gender-based violence victims do not estimate accurately the actual scope of the problem. This suggests that, as a society, we are not tackling very well with the problem of violence against women (World Health

Organization, 2002). On this matter, Gracia (2004) also introduced the concept of the *"iceberg"* of domestic violence. This is essentially that the official number of cases of violence against women is just the tip of the iceberg, being hidden in the water most of the cases which are not visible yet to the society. Additionally, there are many reasons why cases are not reported such as economical dependency, fear of shame and stigma, cultural beliefs, etc. (Palermo *et al.* 2014).

The same Spanish newspaper *El Pais* also published in 2015 that 67.8% of the gender-based violence victims had never reported their condition to the police. Furthermore, 44.8% of these underlying cases played down the problem believing that domestic violence is normal, 26.2% of them felt worried for possible reprisals, and 21% of them felt embarrassed with their condition. Surprisingly, 24.3% of these victims did not break up the relationship with their batterer.

In order to deal with the problem of under-reporting in gender-based violence, appropriate techniques for quantifying the under-reported cases and estimating the actual number of victims should be introduced and studied. It is important to point out that an accurate information of the number of invisible cases of domestic violence is one of the most important things for estimating the actual magnitude of the problem, and making completely visible the hidden part of the "iceberg". In addition, a good understanding of the problem allows the governments to make more accurate official awareness campaigns trying to avoid the phenomenon of under-reporting, and giving women a voice.

In this sense, some works in the literature deal with the problem of under-reporting in domestic violence data using different approaches. For example, Palermo *et al.* (2014) provide bounds for under-estimation of gender-based violence by analysing demographic and health survey data in 24 countries between 2004 and 2011, Wirtz *et al.* (2013) develop a screening tool to identify female survivors of gender-based violence in a humanitarian setting, or Du *et al.* (1991), who demonstrate a significant change in help-seeking rates of intimate partner violence data in Canada comparing data recorded in 1993 with data recording in 1999, and showing an increasing number of help-seeking rates in this period.

The goal of the present work is to deal with the problem of under-reporting in gender-based violence data in order to estimate the actual number of victims of domestic violence by means of a new model of count time series. This model is an extension of the model in

Fernández-Fontelo *et al.* (2016), consisting of a latent process $\{X_n : n = 1, \dots, N\}$, and an observed process $\{Z_n : n = 1, \dots, N\}$. The assumption of independence between the successive states of under-reporting is relaxed in this new model.

Let X_n be a latent process of actual counts that follow a stationary Integer-Valued AutoRegressive model of order 1 (INAR(1)) with Poisson distributed innovations satisfying the stochastic structure:

$$X_n = \alpha \circ X_{n-1} + W_n, \quad W_n \sim \text{Poisson}(\lambda), \tag{4.1}$$

where the parameter $\alpha \in (0, 1)$ and the operator $\circ$ is the well-known *binomial thinning* or *binomial subsampling*, such that $[\alpha \circ X_{n-1} | X_{n-1} = x_{n-1}] =_d \sum_{i=1}^{x_n-1} B_i(\alpha)$ where $\{B_i(\alpha)\}$ is a sequence of independent and identically distributed Bernoulli(α) random variables. The expectation and variance of the *binomial thinning* operator are $\text{E}\left(\alpha \circ X_{n-1} | X_{n-1} = x_{n-1}\right) = \alpha x_{n-1}$ and $\text{Var}\left(\alpha \circ X_{n-1} | X_{n-1} = x_{n-1}\right) = \alpha(1-\alpha)x_{n-1}$, respectively (Steutel and Van Harn 1979).

The INAR(1) process is an homogeneous Markov chain with the following conditional probability density function:

$$\text{P}(X_n = x_n | X_{n-1} = x_{n-1}) = \sum_{j=0}^{\min(x_n, x_{n-1})} \binom{x_{n-1}}{j} \alpha^j (1-\alpha)^{x_{n-1}-j} \text{P}(W_n = x_n - j). \tag{4.2}$$

The expression (4.1) has a unique stationary solution depending on the distribution of the innovations. When W_n are independent Poisson(λ), the marginal distribution of X_n is also Poisson with $\mu_X = \text{E}(X_n) = \lambda/(1-\alpha) = \text{Var}(X_n) = \sigma_X^2$. Furthermore, if X_n have finite second order moments, the auto-covariance function of every stationary INAR(1) model is $\gamma_X(k) = \text{Cov}(X_n, X_{n+k}) = \alpha^k \sigma_X^2$ and hence, the auto-correlation function (ACF) of X_n is $\rho_X(k) = \gamma(k)/\gamma(0) = \alpha^k$, which is geometrically decreasing at rate α.

The INAR(1) model was firstly introduced by McKenzie (1985) and later on by Al-Osh and Alzaid (1987). Several higher-order versions of this model, such as INAR(p) with Poisson or other distributed innovations, were introduced by Al-Osh and Alzaid (1990) and Du and Li (1991). Jazi *et al.* (2012a), Zhu and Joe (2006) and Zhu and Joe (2010) proposed extended versions of the INAR(1) model assuming different discrete probability laws in the innovations, and/or introducing new thinning operators. Additionally, Moriña *et al.* (2011) and Zhu and Joe (2006) proposed extensions for these model when series are non-stationary by introducing seasonal covariates and trends in the models. Other authors such as Freeland and McCabe (2004) proposed forecasting methods based on integer-valued time series processes. Several interesting reviews of these models can be found in Weiß (2008), Fokianos (2011) and Scotto *et al.* (2015).

The remainder of this paper is organized as follows. The quarterly count data for the number of complaints of domestic violence in regions of Galicia are described in Section 4.2. The models for under-reported counts and their probabilistic properties are presented in Section 4.3. Parameter estimation via the moment-based method and the maximum likelihood method are introduced in Section 4.4. The results of the models fitted to the Galicia data are presented in Section 4.5; the models are validated and latent processes are reconstructed using well-known techniques of integer-valued time series models. Section 4.6 is devoted to draw some general conclusions.

4.2 Data description

The data we use to illustrate the usage of the proposed models are based on the quarterly number of complaints of domestic violence against women between 2007 and 2017 from the 45 judicial districts which constitute the 4 provinces (14 from *A Coruña*, 9 from *Lugo*, 9 from *Ourense* and 13 from *Pontevedra*) of the autonomous community of Galicia [1] in Spain. The judicial districts in Spain are essentially territorial units devoted for administering justice. At least, they have to be formed of one municipality within the same province of the autonomous community.

Different ways for reporting the gender-based maltreatment are available. This abuse can be directly reported by the victim or friend/family member, by the police when an official report is presented by the victim (*i.e.*, a report describing physical lesions), or, also, by the social services when they have relevant information related to this maltreatment. All these ways for reporting maltreatment against women allow the victims to officially

[1] In Galicia, the official languages are Galician and Castilian Spanish

record the abuse in form of complaints. The data on gender-based violence complaints in Galicia are completely open and available to all citizens from *La Delegación del Gobierno de España para la Violencia de Género* and *El Consejo General del Poder Judicial de España* [2].

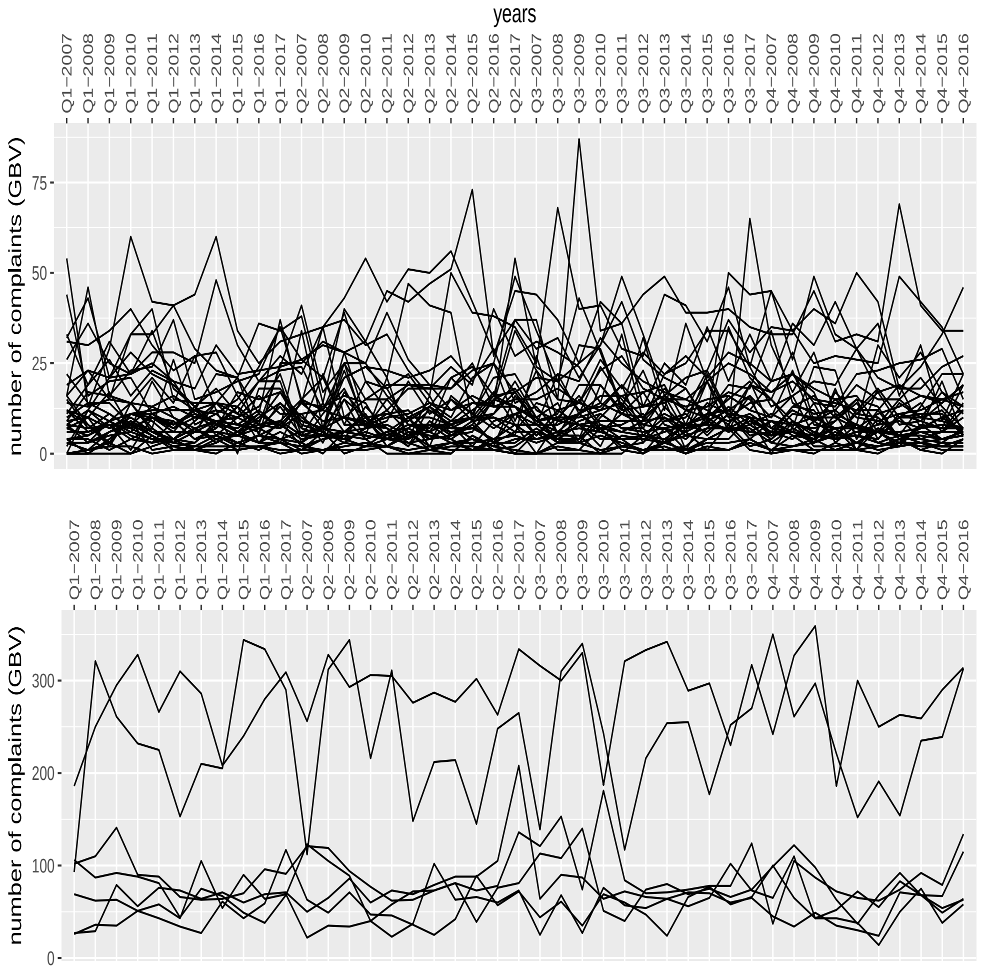


Figure 4.1: Time series plots for the number of quarterly complaints of gender-based violence between 2007 and 2017 in the 45 judicial districts of Galicia (Spain).

Figure 4.1 shows the quarterly series of complaints of domestic violence against women recorded between 2007 and 2017 in the 45 judicial districts of Galicia ($n = 43$ quarters). The series with larger counts belong to the main cities of Galicia, where pop-

[2] `http://estadisticasviolenciagenero.msssi.gob.es` and `http://www.poderjudicial.es/cgpj/es/Temas/Violencia-domestica-y-de-genero/Actividad-del-Observatorio/Datos-estadisticos`

ulation is much larger than in the other areas (*Vigo, A Coruña, Santiago de Compostela, Lugo, Pontevedra, Ferrol* and *Ourense*). Excluding these series, in general, the number of complaints quarterly recorded in Galicia presents relatively low counts over the period of study (between 0 and 87 per quarter).

Figure 4.2 shows the distributions of the averages (left) and dispersion indeces (right) of the registered complaints in the judicial districts of Galicia.

One of the main goals of the present work consists on studying and quantifying the under-reporting in series of complaints of gender-based violence. Count data are generally over-dispersed relative to Poisson, that is, the dispersion index, defined as the variance to mean ratio, exceeds 1. Over-dispersion can occur because of under-reporting if the counts from the under-reporting states and non-under-reporting states are Poisson with different means, or because observed data are a mixture of Poisson distributions based on a latent variable that causes heterogeneity.

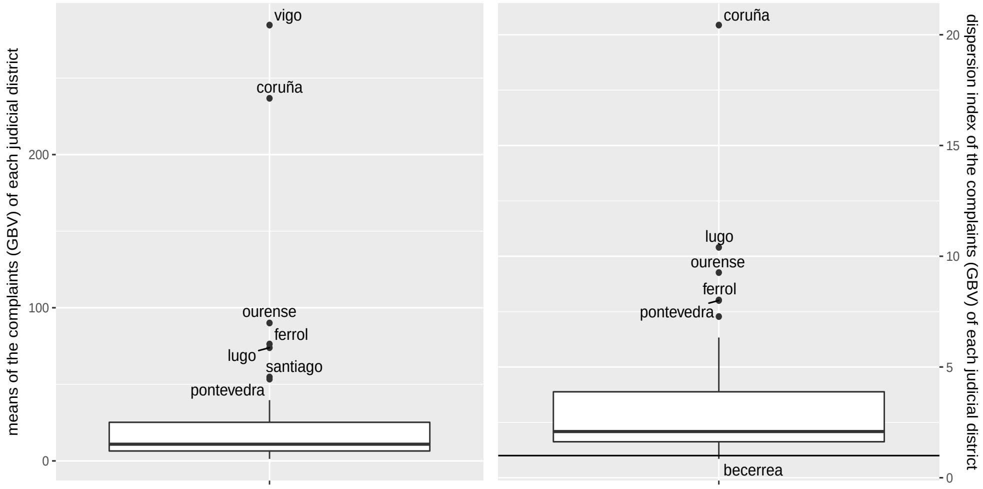


Figure 4.2: The boxplot on the left shows the average number of quarterly complaints over judicial districts of Galicia. The boxplot on the right shows the dispersion indices of number of quarterly complaints over judicial districts of Galicia.

Figure 4.2 (right) shows that nearly all dispersion indices of the series corresponding to the 45 judicial districts of Galicia are greater than 1, that is, most of the series are over-dispersed. In order to be able to identify the under-reporting issue, 10 judicial districts with dispersion indices near 1 or with low counts were removed.

A total of 35 out of 45 judicial districts (77.8%) are used for the analysis of under-reporting of complaints of gender-based violence (71.4% from *A Coruña*, 60.0% from *Lugo*, 60.0% from *Ourense* and 100% from *Pontevedra*). In other words, a total of 58510 complaints of domestic violence against women were officially registered between 2007 and 2017 among the 45 judicial districts of Galicia. After removing 10 out of 45 (22.2%) judicial districts, a total of 56173 complaints remain.

Full and detailed results are presented and discussed in Section 4.5.

4.3

Model specification

In this section, theoretical details and properties are given for a new time series model for under-reported counts. The model assumes a latent process of actual counts that follow an INAR(1) (temporal dependence), and an under-reporting binary state process. The states of under-reporting are assumed to follow a first order Markov chain.

4.3.1

The model and its nested models

Consider a latent INAR(1) process $\{X_n : n = 1, \ldots, N\}$ satisfying (4.1). Let $\{Z_n : n = 1, \ldots, N\}$ be an observed and potentially under-reported process, and $\{I_n : n = 1, \ldots, N\}$ be a binary process such that I_n is an indicator of whether the observation Z_n is under-reported or not. The process Z_n satisfies:

$$Z_n = \begin{cases} X_n & \text{if} \quad I_n = 0, \\ q \circ X_n & \text{if} \quad I_n = 1, \end{cases} \tag{4.3}$$

where $q \in (0, 1)$. The model in the expression (4.3), named under-reported count process model (UCPM), means that if the observed Z_n is equal to X_n, then there is no under-reporting at time n ($I_n = 0$). On the other hand, if Z_n is a binomial thinning of X_n (i.e., $q \circ X_n$), then the process is under-reported at time n ($I_n = 1$). The overall frequency of being under-reported over time is $\text{P}(I_n = 1) = \omega$, while the parameter q represents the overall intensity of the under-reporting.

Consider the situation that the states of under-reporting I_n are serially dependent. As a simple plausible model, we assume they follow a binary discrete-time Markov chain (Zucchini and MacDonald 2009), in which every state results in a success (under-reporting) or failure (no under-reporting). A binary discrete-time Markov chain depends on the transition probabilities, that is, the probabilities of going from one state (under-reporting or no under-reporting) at time $n-1$ to other state (under-reporting or no under-reporting) at time n. In this case, the transition probabilities $\mathrm{P}(I_n = j|I_{n-1} = i) = p_{ij}$ for $i, j = 0, 1$, lead to the transition probability matrix: $\boldsymbol{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$. For instance, the parameter p_{01} of the transition probability matrix is the probability of going from the state of no under-reporting at time $n-1$ to the state of under-reporting at time n. Under stationarity, we have $\Omega\boldsymbol{P} = \Omega$ where $\Omega = (1-\omega, \omega)$. Therefore, $\boldsymbol{P}$ can be simplified in terms of ω and one other parameter, for example p_{01}; this leads to the following re-parametrized matrix:

$$\boldsymbol{P} = \begin{bmatrix} \mathrm{P}(I_n = 0|I_{n-1} = 0) & \mathrm{P}(I_n = 1|I_{n-1} = 0) \\ \mathrm{P}(I_n = 0|I_{n-1} = 1) & \mathrm{P}(I_n = 1|I_{n-1} = 1) \end{bmatrix} = \begin{bmatrix} 1 - p_{01} & p_{01} \\ p_{01}\frac{1-\omega}{\omega} & 1 - p_{01}\frac{1-\omega}{\omega} \end{bmatrix}. \tag{4.4}$$

Together, equations (4.1)–(4.4) specify a times series model of under-reported counts based on a latent INAR(1) times series X_n with Poisson distributed innovations and a latent under-reporting process I_n that is a binary first order Markov chain. We refer to this as the full model with five parameters $\alpha, \lambda, q, \omega, p_{01}$. We also consider three nested models with fewer parameters.

(M4a) If I_n is an independent sequence with $\mathrm{P}(I_n = 1) = \omega$ for all n, then $\omega = p_{01} = p_{11}$ and the simpler model introduced in Fernández-Fontelo *et a.* (2016) is obtained. This nested model has four parameters $\alpha, \lambda, q, \omega$.

(M4b) If X_n in (4.1) is an independent sequence with $\alpha = 0$, then the serial dependence in the observed counts Z_n comes only from the Markov dependence in I_n. This nested model has four parameters $\lambda, q, \omega, p_{01}$.

(M3) If both I_n and X_n are independent sequences so that there is no serial dependence in the observed counts Z_n, then Z_n is a mixture of two Poisson distributions. The resulting nested model has three parameters λ, q, ω,

Model properties

The expectation and variance of the process Z_n are $\mathrm{E}(Z_n) = \mu_X(1-\omega(1-q))$ and $\mathrm{Var}(Z_n) = \mu_X^2\omega(1-\omega)(1-q)^2 + \mu_X(1-\omega(1-q))$, respectively, coinciding with those of the model of Fernández-Fontelo *et al.* (2016). The expression of the ACF is presented in the following proposition, whose proof is given in the Appendix B.

Proposition 1 *The ACF of the observed and under-reported process Z_n is for positive integers k:*

$$\rho_Z(k) = \frac{\alpha^k(1-\omega(1-q))^2 + \lambda_2^k\mu_X\omega(1-\omega)(1-q)^2 + (\alpha\lambda_2)^k\omega(1-\omega)(1-q)^2}{\mu_X\omega(1-\omega)(1-q)^2 + (1-\omega(1-q))}, \tag{4.5}$$

where $\lambda_2 = 1 - p_{01}/\omega$ is the second eigenvalue of $\boldsymbol{P}$ given in (4.4).

From the equation (4.5), when $\alpha > |\lambda_2|$, the ACF decreases geometrically at rate α. However, when $\alpha < |\lambda_2|$ and $\lambda_2 > 0$ (higher chance of remaining in the same state of under-reporting), the ACF decreases geometrically at rate λ_2. Finally, when $\alpha < |\lambda_2|$ and $\lambda_2 < 0$ (higher chance to change the state of under-reporting), the absolute value of the ACF decreases geometrically at rate $|\lambda_2|$. Note that when $p_{01} = \omega$, the expression (4.5) results in the ACF of nested model M4a, as given in Fernández-Fontelo *et al.* (2016).

Under stationarity, the marginal distribution of Z_n is a mixture of two Poisson distributions such that:

$$Z_n \sim \begin{cases} \text{Poisson}\left(\frac{\lambda}{1-\alpha}\right), & \text{probability } 1-\omega, \\ \text{Poisson}\left(\frac{q\lambda}{1-\alpha}\right), & \text{probability } \omega. \end{cases} \tag{4.6}$$

Notice that when $q = 0$, that is, the overall intensity of the under-reporting is maximum, the distribution of the observed process Z_n would be a Zero-inflated Poisson distribution. Several works in the literature deal with INAR models with innovations following a Zero-inflated Poisson (*i.e.* Jazi *et al.* 2012a).

4.4 Model estimation

In this section, two different methods for estimating the parameters of the full model and its nested models are presented: a simpler one based on the method of moments, and another based on the log-likelihood function. The method of moments is computed using moments related to the marginal distribution of the observed process Z_n (mixture of two Poisson distributions), and also the expression of the auto-correlation function (4.5) of Z_n. The likelihood function of Z_n is not directly tractable, and the forward algorithm for hidden Markov chains (HMC) is used for its computation.

4.4.1 Moments-based estimation

Based on the marginal distribution of Z_n in expression (4.6), a moment-based method for computing point estimates of parameters can be used. Additionally, parametric bootstrap is proposed in order to compute the corresponding 90% confidence limits of parameters. The bias-corrected and accelerated bootstrap method (BCa) for computing bootstrap confidence intervals is considered (Efron and Tibshirani 1986, and DiCiccio and Efron 1996).

Point estimates based on the method of moments are obtained as follows:

1. Based on an appropriate method like the Expectation-Maximisation (EM-algorithm) (Zucchini and MacDonald 2009), the marginal distribution of Z_n can be fitted, obtaining the estimates $\widehat{\omega}$, $\widehat{\theta_1} = \widehat{\lambda}/(1-\widehat{\alpha})$ and $\widehat{\theta_2} = \widehat{q\lambda}/(1-\widehat{\alpha})$. It is then straightforward to estimate the overall intensity of the under-reporting as $\widehat{q} = \widehat{\theta_2}/\widehat{\theta_1}$.

2. The parameter α is estimated using the theoretical expression of the auto-correlation function (ACF) of Z_n in (4.5) as follows. The expression of the ACF can be written as:

$$\rho_Z(k) = C_1\alpha^k + C_2\lambda_2^k\left(\mu_X + \alpha^k\right) \tag{4.7}$$

where

$$C_1 = \frac{(1-\omega(1-q))^2}{\mu_X\omega(1-\omega)(1-q)^2 + (1-\omega(1-q))}, \quad C_2 = \frac{\omega(1-\omega)(1-q)^2}{\mu_X\omega(1-\omega)(1-q)^2 + (1-\omega(1-q))}$$

are completely determined by replacing parameters by their moment-based estimates obtained in the step 1. Notice that $\widehat{\mu}_X = \widehat{\sigma}^2_X = \widehat{\theta}_1$. Replacing in equation (4.7) by the two first coefficients of the empirical ACF, that is, $\rho_Z(1)$ and $\rho_Z(2)$, a system of two equations, which depends on α and λ_2, has to be solved. This system leads to the following fourth-degree equation: $A\alpha^4 + B\alpha^3 + D\alpha^2 + E\alpha + F = 0$, where $A = C_1(C_1 + C_2)$, $B = 2C_1(C_2\mu_X - \rho_Z(1))$, $D = C_1\mu_X\left(C_2\mu_X + C_1\right) - C_2\rho_Z(2) + \rho^2_Z(1)$, $E = -2\mu_X\left(C_1\rho_Z(1) + C_2\rho_Z(2)\right)$ and $F = \mu_X\left(\rho^2_Z(1) - C_2\mu_X\rho_Z(2)\right)$. The coefficients A and E are always positive and negative respectively, while the coefficients B, D and F can be positive or negative. From a general point of view, this equation can have a minimum of 0 and a maximum of 4 real positive roots, depending on the signs of the coefficients B, D and F. Empirically, focusing on the gender-based violence data, this equation has always two positive real roots within $(0, 1)$. This means that there are two latent dependent processes that contribute to the ACF of the observed process Z_n. Hence, the parameter α might not be completely identifiable based on $\rho_Z(1)$ and $\rho_Z(2)$, being needed higher serial correlations such as $\rho_Z(3)$ and $\rho_Z(4)$. That is, the better estimate of α corresponds to the positive real root of the equation (4.7) which provides the theoretical coefficients $\rho(3)$ and $\rho(4)$ closer to the empirical ones.

3. Finally, the parameter λ can be directly obtained through $\widehat{\lambda} = \widehat{\theta}_1(1 - \widehat{\alpha})$.

If $\hat{\alpha}$ becomes 0, then nested model M4b is obtained. For nested model M4a, slight changes in the previous method of moments provide the moment-based estimates. Accordingly, when independence is considered between the states of under-reporting, parameters $(\alpha, \lambda, \omega, q)$ can be estimated modifying step 2. in the following way:

2′. The parameter α can be estimated in different ways. The first one is based on the theoretical expression of the ACF which is that one in the expression (4.5) when $\lambda_2 = 0$. Taking the first empirical coefficient of the ACF, and replacing the parameters ω and q by their moment-based estimates obtained in the previous step 1, α is directly estimated. The second way for estimating the parameter α consists of using the two first empirical coefficients of the ACF such that $\widehat{\alpha} = \widehat{\rho(2)}/\widehat{\rho(1)}$. This model implies that the ACF is positive and decreasing in the first few lags, and hence $\widehat{\alpha}$ is in (0,1).

For nested model M3, the parameteres (λ, ω, q) are directly obtained by fitting a mixture of two Poisson distributions by means, for instance, of the EM-algorithm (similarly as in the step 1).

In order to provide estimates of the standard errors of the moment-based estimates of the model parameters, 90% confidence limits based on the bias-corrected and accelerated bootstrap method (BCa) are computed (Efron and Tibshirani 1986, and DiCiccio and Efron 1996).

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

4.4.2

Maximum likelihood method

The parameters of the model can be estimated by maximum likelihood using the methodology for hidden Markov chains. Consider a realization $(Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N)$ or $\boldsymbol{Z}_{1:N} = \boldsymbol{z}_{1:N}$, a latent process $(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$ or $\boldsymbol{X}_{1:N} = \boldsymbol{x}_{1:N}$, and a correlation structure between the states of under-reporting $(I_1 = i_1, I_2 = i_2, \ldots, I_N = i_N)$ or $\boldsymbol{I}_{1:N} = \boldsymbol{i}_{1:N}$ based on a binary discrete-time Markov chain. The likelihood function takes the expression:

$$\begin{aligned}\mathrm{P}(\boldsymbol{Z}_{1:N} = \boldsymbol{z}_{1:N}) &= \sum_{\boldsymbol{x}_{1:N}} \mathrm{P}\Big[\{X_1 = x_1, \ldots, X_N = x_N\}, \{Z_1 = z_1, \ldots, Z_N = z_N\}\Big] \\ &= \sum_{\boldsymbol{x}_{1:N}, \boldsymbol{i}_{1:N}} \mathrm{P}\Big[\{Z_1 = z_1, \ldots, Z_N = z_N\}\Big|\{X_1 = x_1, \ldots, X_N = x_N\}, \{I_1 = i_1, \ldots, I_N = i_N\}\Big] \\ &\quad \mathrm{P}\Big[\{X_1 = x_1, \ldots, X_N = x_N\}, \{I_1 = i_1, \ldots, I_N = i_N\}\Big]. \end{aligned} \tag{4.8}$$

The direct computation of this function (4.8) is untractable. Some indirect method should be used for computing (4.8). Different recursive methods can be used for this purpose (Zucchini and MacDonald 2009) and we introduce the notation $\boldsymbol{z}_{1:j}, \boldsymbol{x}_{1:j}, \boldsymbol{i}_{1:j}$ for the first j elements. The choice in this work is the well-known forward algorithm which is based on the forward probabilities:

$$
\begin{aligned}
&\gamma_n\left(\boldsymbol{z}_{1:n}, x_n, i_n\right) = \mathrm{P}\left(\boldsymbol{Z}_{1:n} = \boldsymbol{z}_{1:n}, X_n = x_n, I_n = i_n\right) \qquad (4.9)\\
&= \sum_{x_{n-1}, i_{n-1}} \mathrm{P}(X_n = x_n, I_n = i_n, X_{n-1} = x_{n-1}, I_{n-1} = i_{n-1}, \boldsymbol{Z}_{1:n} = \boldsymbol{z}_{1:n})\\
&= \sum_{x_{n-1}, i_{n-1}} \mathrm{P}\left(Z_n = z_n | \boldsymbol{Z}_{1:n-1} = \boldsymbol{z}_{1:n-1}, X_n = x_n, X_{n-1} = x_{n-1}, I_n = i_n, I_{n-1} = i_{n-1}\right)\\
&\quad \times \mathrm{P}\left(\boldsymbol{Z}_{1:n-1} = \boldsymbol{z}_{1:n-1}, X_n = x_n, X_{n-1} = x_{n-1}, I_n = i_n, I_{n-1} = i_{n-1}\right)\\
&= \sum_{x_{n-1}, i_{n-1}} \mathrm{P}\left(Z_n = z_n | X_n = x_n, I_n = i_n\right)\\
&\quad \times \mathrm{P}\left(X_n = x_n, I_n = i_n | \boldsymbol{Z}_{1:n-1} = \boldsymbol{z}_{1:n-1}, X_{n-1} = x_{n-1}, I_{n-1} = i_{n-1}\right)\\
&\quad \times \mathrm{P}\left(\boldsymbol{Z}_{1:n-1} = \boldsymbol{z}_{1:n-1}, X_{n-1} = x_{n-1}, I_{n-1} = i_{n-1}\right)\\
&= \sum_{x_{n-1}, i_{n-1}} \mathrm{P}\left(Z_n = z_n | X_n = x_n, I_n = i_n\right)\\
&\quad \times \mathrm{P}\left(X_n = x_n, I_n = i_n | \boldsymbol{Z}_{1:n-1} = \boldsymbol{z}_{1:n-1}, X_{n-1} = x_{n-1}, I_{n-1} = i_{n-1}\right)\\
&\quad \times \gamma_{n-1}\left(\boldsymbol{z}_{1:n-1}, x_{n-1}, i_{n-1}\right).
\end{aligned}
$$

Since the processes X_n and I_n are mutually independent, then:

$$
\begin{aligned}
\gamma_n\left(z_{1:n}, x_n, i_n\right) = &\sum_{x_{n-1}, i_{n-1}} \mathrm{P}\left(Z_n = z_n | X_n = x_n, I_n = i_n\right) \mathrm{P}\left(X_n = x_n | X_{n-1} = x_{n-1}\right)\\
&\times \mathrm{P}\left(I_n = i_n | I_{n-1} = i_{n-1}\right) \gamma_{n-1}\left(\boldsymbol{z}_{1:n-1}, x_{n-1}, i_{n-1}\right), \qquad (4.10)
\end{aligned}
$$

where $\mathrm{P}(X_n = x_n | X_{n-1} = x_{n-1})$, which is the conditional probability density function of an INAR(1) model with Poisson(λ) innovations, are the transition probabilities of the model following expression (4.2). On the other hand, $\mathrm{P}(I_n = i_n | I_{n-1} = i_{n-1})$ comes from the transition probability matrix $\boldsymbol{P}$ in (4.4), and $\mathrm{P}(Z_n = z_n | X_n = x_n, I_n = i_n)$, which are the emission probabilities, take the following expression:

$$
\mathrm{P}(Z_n = z_n | X_n = x_n, I_n = i_n) = \begin{cases} 0 & \text{if} \quad x_n < z_n \\ 0 & \text{if} \quad i_n = 0, x_n > z_n \\ 1 & \text{if} \quad i_n = 0, x_n = z_n \\ \binom{x_n}{z_n} q^{z_n} (1-q)^{x_n - z_n} & \text{if} \quad i_n = 1, x_n \geq z_n \end{cases} \qquad (4.11)
$$

According to equation (4.10), the likelihood function of the process Z_n can be computed recursively through:

$$\mathrm{P}(\boldsymbol{Z}_{1:N} = \boldsymbol{z}_{1:N}) = \mathrm{P}\left(Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N\right) = \sum_{x_N = z_N}^{\infty} \gamma_N\left(\boldsymbol{z}_{1:N}, x_N, i_N\right), \tag{4.12}$$

starting from $\gamma_1(z_1, x_1, i_1) = \mathrm{P}\left(X_1 = x_1\right)\mathrm{P}\left(Z_1 = z_1 | X_1 = x_1\right)\mathrm{P}\left(I_1 = i_1\right)$, and $\mathrm{P}\left(X_1 = x_1\right) = e^{-\nu}\nu^{x_1}/x_1!$ with $\nu = \lambda/(1-\alpha)$. On the other hand, it is assumed that $\mathrm{P}(I_1 = 0) = 1$ and $\mathrm{P}(I_1 = 1) = 0$, that is, it is assumed that at time $n = 1$, the process is not under-reported.

Slight modifications in the expression of the forward probabilities (4.10) are needed for the nested models. For instance, for nested model M4a with parameter vector $(\alpha, \lambda, \omega, q)$, the forward probabilities take the expression:

$$\gamma_n\left(\boldsymbol{z}_{1:n}, x_n\right) = \sum_{x_{n-1}} \mathrm{P}\left(Z_n = z_n | X_n = x_n\right)\mathrm{P}\left(X_n = x_n | X_{n-1} = x_{n-1}\right)\gamma_{n-1}\left(\boldsymbol{z}_{1:n-1}, x_{n-1}\right). \tag{4.13}$$

The transition probabilities remain invariant taking the expression (4.2), but the emission probabilities are simplified as follows:

$$\mathrm{P}(Z_n = z_n | X_n = x_n) = \begin{cases} 0 & \text{if} \quad x_n < z_n \\ (1-\omega) + \omega q^{x_n} & \text{if} \quad x_n = z_n \\ \omega\binom{x_n}{z_n} q^{z_n}(1-q)^{x_n - z_n} & \text{if} \quad x_n > z_n \end{cases} \tag{4.14}$$

Full details can be found in Fernández-Fontelo *et al.* (2016).

For nested model M3, the forward probabilities are computed according to (4.13), with the same emission probabilities in expression (4.14), but now the transition probabilities are computed using a Poisson distribution with parameter λ.

For nested model M4b with parameter vector $(\lambda, \omega, q, p_{01})$, the forward probabilities can be computed according to (4.10) with the emission probabilities in (4.11), but the transition probabilities are computed using a Poisson distribution with parameter λ.

Notice also that the full model is a Hidden Markov chain with an infinite number of states, since the actual counts are always considered equal or greater than those observed over that period, that is, $x_n \geq z_n$ for all n. Empirically, expression (4.12) can be solved by specifying and upper threshold, that is, $\mathrm{P}(\boldsymbol{Z}_{1:N} = \boldsymbol{z}_{1:N}) = \sum_{x_n=z_n}^{T} \gamma_n \left(\boldsymbol{z}_{1:n}, x_n, i_n\right)$, where T can be three times the maximum value of the observed series, that is, $T = 3\max\left(Z_n\right)$.

4.4.3

Goodness-of-fit and reconstruction of the latent process

In this section, the mid-pseudo residuals and the Viterbi algorithm are described, for the sake of completeness. The mid-pseudo residuals are widely used for validating Hidden Markov Chains (HMC) models, while the Viterbi algorithm, which is also commonly used in HMC contextes, is used for reconstructing the most probable sequence of states (actual number of complaints) (Zucchini and MacDonald 2009).

The normal pseudo-residuals (also called quantile residuals by Dunn and Smyth 1996) are used for assessing the general fit of a model, and also for detecting outliers. They are a case of the well-known Cox-Snell residuals (Cox and Snell 1968). In continuous cases, they are computed based on the following probability: $v_n = \Phi^{-1}(\mathrm{P}(Z_n \leq z_n | Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}, Z_{n+1} = z_{n+1}, \ldots, Z_N = z_N)) = \Phi^{-1}\left(u_n\right)$. If the model is valid, these residuals are realizations of a standard normal distribution. However, in discrete cases, the so-called pseudo-residual segments $[v_n^-, v_n^+]$ are needed, which are built as follows:

$$\begin{aligned} v_n^- &= \Phi^{-1}\left(\mathrm{P}\left(Z_n < z_n | Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}, Z_{n+1} = z_{n+1}, \ldots, Z_N = z_N\right)\right) \\ &= \Phi^{-1}\left(u_n^-\right), \qquad (4.15) \\ v_n^+ &= \Phi^{-1}\left(\mathrm{P}\left(Z_n \leq Z_n | Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}, Z_{n+1} = z_{n+1}, \ldots, Z_N = z_N\right)\right) \\ &= \Phi^{-1}\left(u_n^+\right), \qquad (4.16) \end{aligned}$$

where

$$\begin{aligned} &\mathrm{P}\left(Z_n = z_n | Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}, Z_{n+1} = z_{n+1}, \ldots, Z_N = z_N\right) = \\ &= \frac{\mathrm{P}\left(Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}, Z_n = z_n, Z_{n+1} = z_{n+1}, \ldots, Z_N = z_N\right)}{\mathrm{P}\left(Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}, Z_{n+1} = z_{n+1}, \ldots, Z_N = z_N\right)}. \qquad (4.17) \end{aligned}$$

These probabilities are computed using the forward algorithm described in equation (4.10), evaluating the likelihood function at the maximum likelihood estimators of the parameters.

In discrete cases, it is complicated to check the normality of the normal-pseudo residuals segments. Accordingly, the mid-pseudo residuals can be calculated from the normal-pseudo residuals segments in the following way:

$$v_n^m = \Phi^{-1}\left(\frac{u_n^- + u_n^+}{2}\right). \tag{4.18}$$

As in the continuous time series case, if a model is valid, the mid-pseudo residuals are realizations of the standard normal distribution. Notice that when $z_n = 0$ the lower pseudo-residual segment is not defined. In this cases, when computing this segment, the lower limit can be replaced by a standard normal quantile of a probability very close to 0. For instance, $v_n^- = 6.3613$ $(p = 10^{-10})$. In addition, this is not relevant when computing the mid-pseudo residuals since $u_n^- = 0$, and then $v_n^m = \Phi^{-1}\left(u_n^+/2\right)$.

Other methods can be used for checking the assessment of these model based on HMC. A comprehensive review can be found in Zucchini and MacDonald (2009).

The Viterbi algorithm (Viterbi 1967, and Forney 1973) is a common method used in HMC contexts to determine the most likely sequence of latent states. After models are validated, this method allows us to reconstruct the most likely sequence of the latent states.

The idea behind the Viterbi algorithm is to provide the latent chain $X^* = (X_1^* = x_1^*, X_2^* = x_2^*, \ldots, X_N^* = x_N^*)$ that maximises the likelihood function of the latent process $(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$ given the observed series $(Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N)$, assuming also that the parameters of the model are completely known (those computed by the maximisation of the likelihood function). That is:

$$\begin{aligned} X^* &= (X_1^* = x_1^*, X_2^* = x_2^*, \ldots, X_N^* = x_N^*) \\ &= \text{argmax}_{z_{1:N}} \text{P}(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N | Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N) \\ &= \text{argmax}_{z_{1:N}} \text{P}(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N, Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N), \end{aligned} \tag{4.19}$$

because the probabilities $\text{P}(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N | Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N)$ and $\text{P}(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N, Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N)$ are proportional since the probability $\text{P}(Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N)$ does not depend on the latent process $(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$.

4.5 Results for Galicia data

4.5.1 Under-reporting analysis in gender-based violence

The autonomous community of Galicia (Spain) is made up of 45 judicial districts within the 4 provinces (14 from *A Coruña*, 9 from *Lugo*, 9 from *Ourense* and 13 from *Pontevedra*). Each judicial district provides an official series of the number of quarterly complaints of gender violence against women from 2007 to 2017. Before analysing the complaints data using the models in Sections 4.3 and 4.4, series coming from these judicial districts are conveniently described and explored by using some descriptive statistics such as the dispersion index (ratio between the sample variance and sample mean) (Figure 4.2).

Accordingly, 8 out of 45 series (18%) are removed from the further analysis because they show dispersion indices very close to 1 (*Noia* (*A Coruña*), *Muros* (*A Coruña*), *Becerreá* (*Lugo*), *Sarria* (*Lugo*), *Ortigueira* (*A Coruña*), *Bande* (*Ourense*), *Xinzo de Limia* (*Ourense*) and *Ribadavia* (*Ourense*)). Additionally, the series of the judicial districts of *Villalba* (*Lugo*), *Viveiro* (*Lugo*), *Carballiño* (*Ourense*) and *Arzúa* (*A Coruña*). Also the series of *Villalba* (*Lugo*) and *Arzúa* (*A Coruña*) are removed because the low counts mean that the model and its nested versions are not well estimated.

For the remaining 35 series included in the analysis of under-reporting, two different methods introduced in Section 4.4 are considered for estimating parameters of the full model and its nested models. In these 35 regions, there are 56173 complaints out of 58510 (96.01%) recorded between 2007 and 2017.

For the method of moments, the full model, nested model M4a and nested model M3 are fitted. For those series for which the under-reporting is quantified using the method of moments, the criteria for selecting the best model are: (a) the empirical ACF of the series and, (b) the empirical standard errors of the parameters computed using parametric bootstrap. When the series has insignificant serial correlation of small lags, the mixture of two Poisson distributions is directly selected. Otherwise, the choice between the full model and nested model M4a depends on which leads to smaller empirical standard errors.

For computing these empirical standard errors, 99 repetitions of 99 parametric bootstrap samples are generated, and the averages of the empirical bootstrap standard errors are provided for each parameter (italic results in Table 4.1).The BCa confidence limits are constructed based on 999 bootstrap samples.

The second method for quantifying the under-reporting in the complaints data is the method of maximum likelihood, introduced in Section (4.4.2). Generally, both methods can be used for studying the under-reporting in the series included in this analysis. However, in some cases, the maximum likelihood method is computationally intensive and slow (especially when counts are very large) and parameter estimates based on the method of moments are reported. When both methods provide sensible results, parameter estimates based on the method of maximum likelihood are reported. The best model among the four under consideration is selected by means of the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC).

Tables 4.1, 4.2 and 4.3 show the moment-based and maximum likelihood estimates of the 35 series considered in the analysis. A total of 19 out of 35 series (54.3%) are modelled based on nested model M3 or the mixture of two Poisson distributions, 3 out of 35 (8.6%) are based on the nested model, 7 out of 35 (20.0%) are based on the full model, and 6 out of 35 (17.1%) are based on the nested model M4b.

From a different point of view, Table 4.4 shows the proportions of each model in every province of Galicia (*A Coruña*, *Lugo*, *Ourense* and *Pontevedra*).

Figure 4.3 show the distribution of the estimated overall frequencies $\widehat{\omega}$ (top) and the estimated intensities $\widehat{q}$ (bottom) of the under-reporting in quarterly complaint of domestic violence against women in the 45 judicial districts of Galicia. The areas in green are those not included in the analysis.

Figure 4.4 shows the distribution of the models in Tables 4.1, 4.2 and 4.3 in each of the 45 judicial districts in Galicia.

Table 4.1: Moment-based estimates for series with large counts and 90% BCa confidence intervals.

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

Table 4.2: Maximum likelihood estimates for series without large counts, and 90% confidence intervals (part I).

	α	λ	ω	q	p_{01}	AIC/BIC
Ordes	-	14.366	0.774	0.562	0.305	241.183
(A Coruña)	-	(11.716, 17.016)	(0.674, 0.874)	(0.457, 0.667)	(0.020, 0.590)	248.228
Padrón	-	13.830	0.573	0.531	-	255.993
(A Coruña)	-	(10.733, 16.926)	(0.240, 0.906)	(0.434, 0.629)	-	261.276
Negreira	-	9.707	0.637	0.505	-	228.144
(A Coruña)	-	(6.849, 12.565)	(0.272, 1.000)	(0.386, 0.624)	-	233.428
Carcubión	-	16.054	0.875	0.546	-	238.872
(A Coruña)	-	(10.593, 21.515)	(0.679, 1.000)	(0.385, 0.706)	-	244.156
Carballo	-	29.916	0.758	0.529	-	286.954
(A Coruña)	-	(25.830, 34.002)	(0.613, 0.903)	(0.460, 0.598)	-	292.238
Betanzos	-	35.957	0.689	0.566	0.356	322.630
(A Coruña)	-	(6.391, 65.523)	(0.454, 0.924)	(0.503, 0.628)	(0.074, 0.644)	329.675
Fonsagrada	-	2.835	0.705	0.208	0.276	136.386
(Lugo)	-	(1.776, 3.894)	(0.416, 0.994)	(0.095, 0.321)	(0.041, 0.511)	143.431
Chantada	-	13.406	0.859	0.394	0.478	228.231
(Lugo)	-	(9.876, 16.936)	(0.690, 1.000)	(0.300, 0.493)	(0.082, 0.874)	235.276
Mondoñedo	-	27.805	0.976	0.360	-	238.253
(Lugo)	-	(18.714, 36.896)	(0.938, 1.000)	(0.240, 0.482)	-	243.536
Monfor. L.	-	13.798	0.105	0.324	-	260.468
(Lugo)	-	(12.716, 14.880)	(0.005, 0.205)	(0.138, 0.509)	-	265.752
Viveiro	-	20.775	0.121	0.545	-	275.563
(Lugo)	-	(18.986, 22.563)	(-0.084, 0.327)	(0.282, 0.809)	-	280.847
Barco V.	-	16.917	0.665	0.466	0.232	249.175
(Ourense)	-	(14.715, 19.119)	(0.394, 0.936)	(0.392, 0.540)	(0.033, 0.431)	256.220
Verin	-	17.170	0.243	0.475	-	281.182
(Ourense)	-	(15.504, 18.837)	(0.058, 0.429)	(0.340, 0.610)	-	286.090
Puebla T.	-	6.953	0.852	0.298	-	177.652
(Ourense)	-	(4.462, 9.444)	(0.717, 0.988)	(0.190, 0.405)	-	182.935
Carballiño	0.396	5.974	0.078	0.263	-	242.095
(Ourense)	(0.228, 0.564)	(4.295, 7.653)	(-0.003, 0.159)	(0.079, 0.447)	-	249.140
Celanova	-	4.465	0.427	0.390	-	190.579
(Ourense)	-	(2.982, 5.948)	(-0.041, 0.895)	(0.154, 0.625)	-	195.862
Caldas R.	-	23.200	0.681	0.524	0.311	280.961
(Pontevedra)	-	(20.173, 26.226)	(0.433, 0.929)	(0.453, 0.595)	(0.038, 0.584)	288.006
Porriño	0.372	21.456	0.917	0.634	0.427	284.719
(Pontevedra)	(0.063, 0.681)	(10.363, 32.549)	(0.769, 1.000)	(0.521, 0.747)	(0.024, 0.830)	293.525

Table 4.3: Maximum likelihood estimates for series without large counts, and 90% confidence intervals (part II).

	α	λ	ω	q	p_{01}	AIC/BIC
Redondela	0.611	7.749	0.559	0.497	0.271	273.958
(Pontevedra)	(0.419, 0.803)	(3.701, 11.800)	(0.309, 0.809)	(0.415, 0.579)	(0.093, 0.449)	282.764
Lalin	0.724	5.347	0.823	0.424	-	253.235
(Pontevedra)	(0.494, 0.954)	(0.625, 10.069)	(0.716, 0.930)	(0.355, 0.493)	-	260.280
Cangas M.	0.502	16.217	0.488	0.527	-	348.936
(Pontevedra)	(0.333, 0.671)	(10.625, 21.810)	(0.346, 0.616)	(0.478, 0.576)	-	355.981
Tui	0.487	11.960	0.766	0.405	0.346	284.252
(Pontevedra)	(0.214, 0.760)	(5.014, 18.906)	(0.555, 0.977)	(0.333, 0.477)	(0.106, 0.586)	293.059
A Estrada	-	11.085	0.213	0.402	-	244.782
(Pontevedra)	-	(9.851, 12.319)	(0.037, 0.388)	(0.232, 0.572)	-	250.065
Marin	-	23.893	0.783	0.594	-	292.212
(Pontevedra)	-	(14.940, 32.847)	(0.389, 1.000)	(0.452, 0.737)	-	297.495

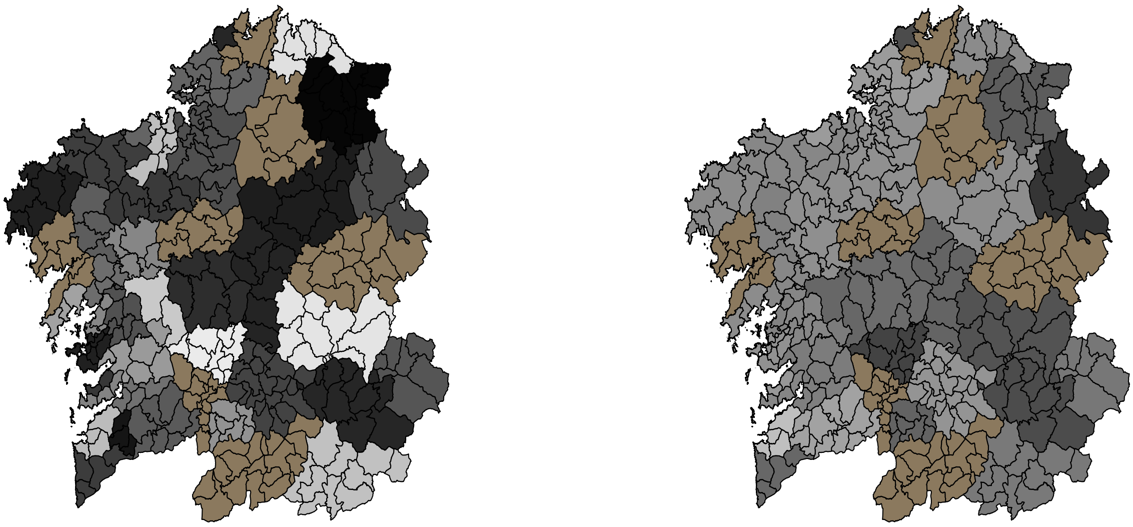
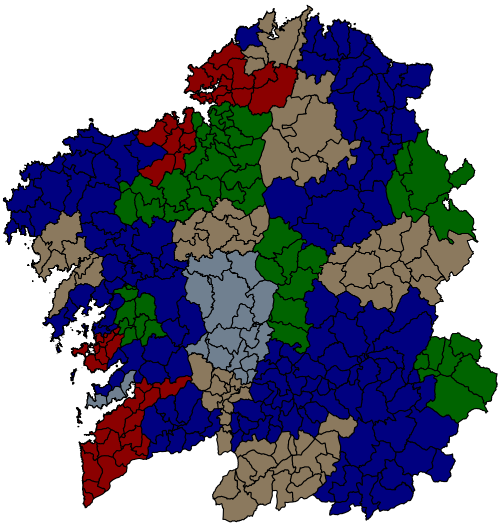


Figure 4.3: The map (left) shows the estimated overall under-reporting frequency in each judicial district of Galicia. The map (right) shows the estimated overall under-reporting intensity in each judicial district of Galicia. In both figures, the darker is the area, the more frequent (ω closer to 1) and intense (q closer to 0) is the phenomenon of under-reporting. The light brown corresponds to the removed areas (10 series).

Table 4.4: Distribution of the models in Tables 4.1, 4.2 and 4.3 among the provinces of Galicia.

	A Coruña	*Lugo*	*Ourense*	*Pontevedra*	(Subtotal)
M3: Poisson Mixture	6 (31.5%) (60.0%)	4 (21.1%) (66.7%)	4 (21.1%) (66.7%)	5 (26.3%) (38.5%)	19 (54.3%)
M4a	0 (0.0%) (0.0%)	0 (0.0%) (0.0%)	1 (33.3%) (16.7%)	2 (67.7%) (15.4%)	3 (8.6%)
full model	2 (28.6%) (20.0%)	0 (0.0%) (0.0%)	0 (0.0%) (0.0%)	5 (71.4%) (38.5%)	7 (20.0%)
M4b	2 (33.3%) (20.0%)	2 (33.3%) (33.3%)	1 (16.7%) (16.7%)	1 (16.7%) (7.7%)	6 (17.2%)
(Subtotal)	10 (28.6%)	6 (17.1%)	6 (17.1%)	13 (37.2%)	35 (100%)



.

Figure 4.4: The map shows the distribution of the models of Tables 4.1, 4.2 and 4.3. Blue corresponds to nested model M3 with mixtures of Poisson distributions. Gray corresponds to nested model M4a. Red corresponds to the full model. Green corresponds to nested model M4b. Light brown corresponds to regions that have been removed from the analysis of under-reporting.

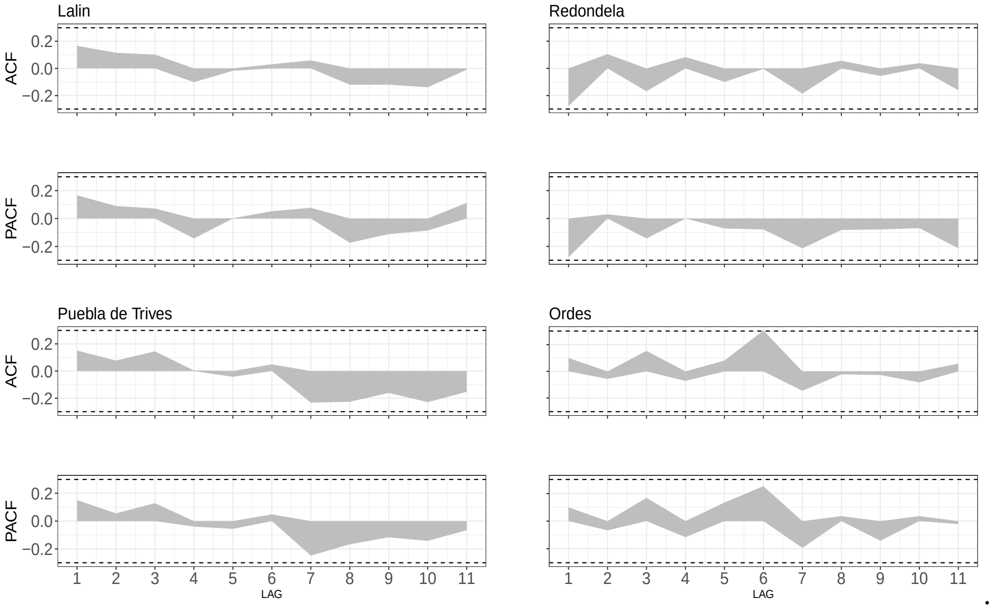


Figure 4.5: ACF and PACF of the mid-pseudo residuals for validating models of *Lalin* (*Pontevedra*), *Redondela* (*Pontevedra*), *Puebla de Trives* (*Ourense*) and *Ordes* (*A Coruña*).

4.5.2 Validation of models and latent process reconstruction

According to Section 4.4.3, the goodness of fit of the models presented in Tables 4.1, 4.2 and 4.3 can be assessed by exploring how its mid-pseudo residuals behave. The model with mid-pseudo residuals like white noise is considered a good-fitting model for the data. In order to explore how these models fit the complaints data, an example in Tables 4.2 and 4.3 (maximum likelihood method) of each of the four models is selected: (a) a mixture of two Poisson distributions through the series of *Puebla de Trives*, (*Ourense*); (b) nested model M4a thorugh the series of *Lalin* (*Pontevedra*); (c) full model through the series of *Redondela* (*Pontevedra*); and (d) nested model M4b through the series of *Ordes* (*A Coruña*). Figure 4.5 shows both the empirical ACF and PACF of the mid-pseudo residuals coming from those models. Residuals of these models behave like white noise, since non-interesting patterns are detected in Figure 4.5.

Latent series from *Puebla de Trives* (*Ourense*), *Lalin* (*Pontevedra*), *Redondela* (*Pontevedra*) and *Ordes* (*A Coruña*) are also reconstructed using the Viterbi (see Section 4.4.3). Results are provided in Figure 4.6.

4.5.3 Interpretations

[1]

In Galicia, the urban areas are essentially located in the provinces of *A Coruña* and *Pontevedra*, where there are the largest cities in the community: *Vigo* (292986 people), *A Coruña* (244099 people), *Santiago de Compostela* (96459 people), *Pontevedra* (82671 people) and *Ferrol* (67569 people). On average, these provinces have younger population with ages of 46.46 and 45 in *A Coruña* and *Pontevedra* respectively, and lower ageing rates. In fact, in *A Coruña* there are 1.5 times more people over 65 than children, while in *Pontevedra* there are 1.25 times more people over 65 than children. Curiously, although *A Coruña* is the province of Galicia with the lowest poverty risk rate (13.39%) and the highest average salary (2095 euros per month), *Pontevedra* is one of those provinces (jointly with *Ourense*) where the poverty risk rate is higher (19.01%), and the average salary is lower (1970 euros per month). However, the phenomenon of under-reporting in *Ourense* seems to be more intense rather than in *Pontevedra*. Additionally, both provinces (*A Coruña* and *Pontevedra*) have the lowest percentages of people who are illiterate or do not finish their primary studies (10.95% *A Coruña* and 9.44% *Pontevedra*), and the highest percentage of people who achieve superior studies (undergraduates, master, doctorate, ...) (15.37% *A Coruña* and 13.17% *Pontevedra*).

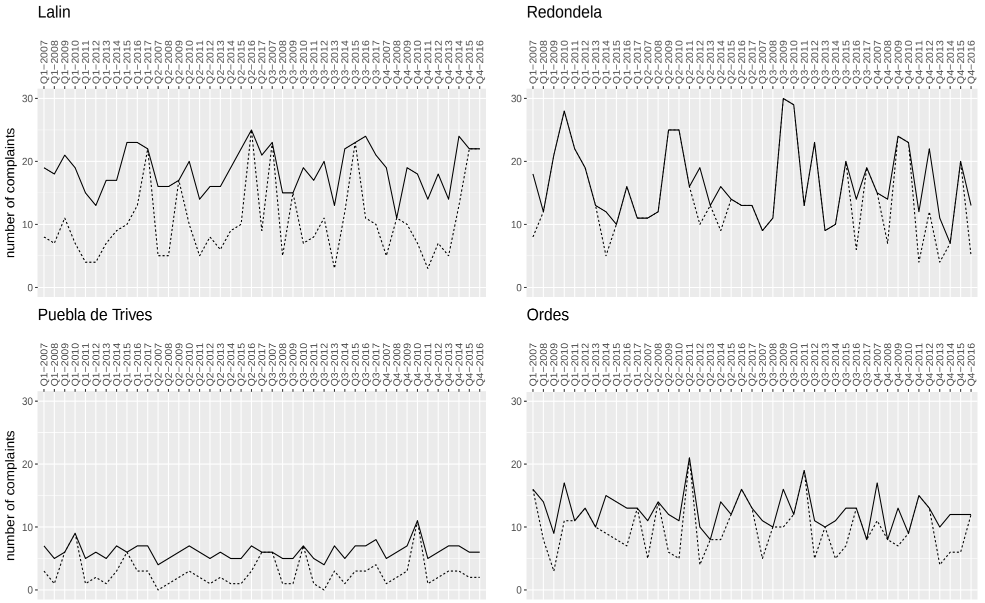
According to the results in Tables 4.1, 4.2, 4.3 and 4.4, models with temporal dependence, that is, with $\alpha \neq 0$ (9 out of 10) and/or a correlation structure between the states of under-reporting (10 out of 13) are essentially selected for modelling the series of *A Coruña* and *Pontevedra*. The distributions of the estimated frequencies ($\widehat{\omega}$) of the judicial districts of both provinces are more homogeneous, showing coefficients of variation of 31.9% (*A Coruña*) and 37.8% (*Pontevedra*). The median estimated frequency of under-reporting is 0.623 in *A Coruña* and 0.648 in *Pontevedra*. The median estimated intensity ($\widehat{q}$) is 0.547 in *A Coruña* and 0.524 in *Pontevedra*. These are those ones closer to one, meaning that the under-reporting in these provinces is less intense than in *Lugo* and *Ourense*.

[1] Statistics in this subsection were provided by Instituto Galego de Estatística, based on the year 2016.

On the other hand, the rural areas in Galicia are located in the provinces of *Ourense* and *Lugo*, which their main cities are *Ourense* (105636 people) and *Lugo* (97995 people). On average, these provinces have older people (50.20 years in *Ourense*) and (49.52 years in *Lugo*), and higher rates of older people. In fact, in *Ourense* there are 2.34 times more people older than 65 than children, while in *Lugo* there are 2.15 times more people older than 65 than children. The province of *Ourense* is that one where the average income is the lowest in Galicia (1817 euros per month), and the poverty risk rate is the highest in the community (19.35%). Additionally, both provinces are those ones where the percentages of people who are illiterate or do not finish their primary studies are the highest (13.66% in *Ourense* and 14.70% in *Lugo*), but the percentages of people who obtain advanced studies are the lowest (11.55% in *Ourense* and 11.87% in *Lugo*).

Series of these provinces are mainly modelled by a mixture of Poisson distributions (8 out of 12). That is, generally, neither the temporal dependence, nor the dependence structure between the states of under-reporting are significant in the judicial districts within these provinces. The distributions of the estimated frequencies ($\widehat{\omega}$) in these provinces are more heterogeneous, showing coefficients of variation of 60.4% and 64.7% in *Ourense* and *Lugo*, respectively. The median estimated frequency of under-reporting in *Ourense* is 0.546, while in *Lugo* it is 0.782, which is the highest one. The median estimated intensity ($\widehat{q}$) in *Ourense* is 0.428, while in *Lugo* is 0.377. These are those ones closer to zero, meaning that the under-reporting is more intense in these provinces.

To conclude, among the 4 provinces of Galicia there are no differences in the frequency of under-reporting according to whether the area is urban or not, whether there are more older or younger people, or whether the percentage of illiterate people or people who do not finish their primary studies is higher or lower. It seems that the frequency of under-reporting is quite higher in the entire community. In fact, the province whose frequency of under-reporting is the lowest is *Ourense* which is a rural area, with the lowest average income per month, the highest poverty risk ratio and the highest ageing rate. However, there is a remarkable difference between the intensities of under-reporting between those urban and rural provinces. The under-reporting seems to be much more intense in those rural area which the percentages of older people are higher and the percentages of illiterate people or people who do not finish their primary studies are also higher.



.

Figure 4.6: Figure shows the most likely sequences of states (number of quarterly complaints) (dark line) in the observed series (dotted line) in *Lalin* (*Pontevedra*), *Redondela* (*Pontevedra*), *Puebla de Trives* (*Ourense*) and *Ordes* (*A Coruña*).

4.6 Discussion

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

The phenomenon of under-reporting in cases of gender-based violence is a world-wide problem exposed by many authors (Watts and Zimmerman 2002, Gracia 2004, and Palermo *et al.* 2014), and also relevant institutions (United Nations 1993, and WHO 2002). This lack of information provides poor and completely biased statistics, leading to undervalue the magnitude of this social scourge, which sadly affects the entire world population.

The goal of the present work consists of demonstrating how this phenomenon of under-reporting in cases of violence against women affects both rural and urban areas, and also providing to the scientific community with a new model which properly quantifies this phenomenon in both areas. This new model could be a useful tool for easily detecting the under-reporting in official cases of domestic violence against women through the number of official complaints, but also for other type of data related to this social issue.

In the supplemental material is included an `R` code which computes the moment-based estimates for each version of the model (4.3). This code also computes the parametric bootstrap confidence intervals based on the bias-corrected and accelerated method (BCa) of the models parameters (999 bootstrap samples).

Additionally, `R` code files are provided for computing the likelihood functions of the full and nested M4b models. An `R` function calling `C` codes is used, since the computation of their likelihood functions in the recursive forward algorithm involves several loops. However, for nested models M4a and M3, a standalone function in `R` is sufficient for computing the likelihood functions. Numerical optimization of the log-likelihood is carried out with a quasi-Newton method.

Finally, `R` code files are provided for computing the mid-pseudo residuals and the Viterbi algorithm. There are packages in `R` which computes the Viterbi algorithm when dealing with HMC with a finite number of latent states (*i.e.* `HMM` package). However, the model introduced in this work is a HMC with an infinite number of states, and hence specific codes in `R` are provided.

Appendix B: Proof of the Proposition 2

The following is the proof of the Proposition (4.5):

Proof: For $k \neq 0$:

$$\begin{aligned}\mathrm{E}\left(Z_n, Z_{n+k}\right) &= \mathrm{E}\left(X_n(1-I_n)X_{n+k}(1-I_{n+k})\right) + \mathrm{E}\left(X_n(1-I_n)q\circ X_{n+k}I_{n+k}\right)\\ &\quad + \mathrm{E}\left(X_{n+k}(1-I_{n+k})q\circ X_nI_n\right) + \mathrm{E}\left(q\circ X_n q\circ X_{n+k}I_nI_{n+k}\right).\end{aligned}$$

Since processes $\{X_n\}$ and $\{I_n\}$ are mutually independent, but the states of under-reporting $\{I_1, I_2, \ldots, I_k, \ldots\}$ are dependent, each of the previous terms can be computed in the same way than the following one:

$$\mathrm{E}\left(X_n(1-I_n)X_{n+k}(1-I_{n+k})\right) = \mathrm{E}(X_nX_{n+k})\mathrm{P}(I_n=0, I_{n+k}=0),$$

where $\mathrm{P}(I_n = 0, I_{n+k} = 0) = \mathrm{P}(I_{n+k} = 0|I_n = 0)\mathrm{P}(I_n = 0)$ and $\mathrm{P}(I_n = 0) = 1 - \omega$. The probability $\mathrm{P}(I_{n+k} = 0|I_n = 0) = p_{00}^{(k)}$ comes from the k-step transition probability matrix, that is, $\boldsymbol{P}^k$. This is the probability of going from the state of no under-reporting to the state of no under-reporting in k steps. Notice that from the equation (4.4), the probability $p_{00}^{(k)} = (1 - p_{01}^{(k)})$.

On the other hand, the transition probability matrix $\boldsymbol{P}$ can be represented in terms of its eigenvalues (Karlin and Taylor 1981). It is straightforward to see that the first eigenvalue of this matrix is 1, and the second eigenvalue is a real number denoted as $\lambda_2 = 1 - p_{01}/\omega$. In this sense, the transition probability matrix $\boldsymbol{P}$ can be written as $\boldsymbol{P} = \boldsymbol{W}\begin{pmatrix} 1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\boldsymbol{W}^{-1}$, where the matrix $\boldsymbol{W} = \begin{pmatrix} 1 & 1 \\ 1 & -\frac{1-\omega}{\omega} \end{pmatrix}$ has the first and second eigenvectors of the matrix P. This representation allows us to write $\boldsymbol{P}^k$ in terms of $\lambda_2^k = 1 - p_{01}^{(k)}/\omega$. Accordingly:

$$\begin{aligned}
&\mathrm{E}\left(X_n(1-I_n)X_{n+k}(1-I_{n+k})\right) = \mathrm{E}(X_nX_{n+k})(1-\omega)\left(1-\omega(1-\lambda_2^k)\right), \\
&\mathrm{E}\left(X_n(1-I_n)q \circ X_{n+k}I_{n+k})\right) = \mathrm{E}(X_nX_{n+k})\, q\omega(1-\omega)(1-\lambda_2^k), \\
&\mathrm{E}\left(q \circ X_n q \circ X_{n+k}I_nI_{n+k}\right) = \mathrm{E}(X_nX_{n+k})\, q^2\omega\left(1-(1-\omega)(1-\lambda_2^k)\right),
\end{aligned}$$

and making some computations:

$$\mathrm{E}\left(Z_nZ_{n+k}\right) = \mathrm{E}\left(X_nX_{n+k}\right)\left(1-\omega(1-q^2)-\omega(1-\omega)(1-q)^2(1-\lambda_2^k)\right), \quad (4.20)$$

where $\mathrm{E}(X_nX_{n+k}) = \mathrm{Cov}(X_n, X_{n+k}) + \mathrm{E}(X_n)\mathrm{E}(X_{n+k}) = \alpha^k\sigma_X^2 + \mu_X^2$. Hence, the auto-covariance function is:

$$\begin{aligned}
\gamma_Z(k) &= \left(\alpha^k\sigma_X^2+\mu_X^2\right)\left(1-\omega(1-q^2)-\omega(1-\omega)(1-q)^2(1-\lambda_2^k)\right) - \mu_X^2(1-\omega(1-q))^2 \\
&= \left(\alpha^k\sigma_X^2+\mu_X^2\right)(1-\omega(1-q))^2 + \left(\alpha^k\sigma_X^2+\mu_X^2\right)\lambda_2^k\omega(1-\omega)(1-q)^2 - \mu_X^2(1-\omega(1-q))^2 \\
&= \alpha^k\sigma_X^2(1-\omega(1-q))^2 + \lambda_2^k\mu_X^2\omega(1-\omega)(1-q)^2 + (\alpha\lambda_2)^k\sigma_X^2\omega(1-\omega)(1-q)^2. \quad (4.21)
\end{aligned}$$

The conclusion follows because $\gamma_Z(0) = \mathrm{Var}(Z_n)$ and $\rho_Z(k) = \gamma_Z(k)/\gamma_Z(0)$. □

CHAPTER 5

Integer-valued AR processes with Hermite innovations and time-varying parameters: An application to bovine fallen stock surveillance at a local scale

This chapter corresponds to the contents of: Fernández-Fontelo, A., Fontdecaba, S., Alba, A. and Puig, P. (2017). Integer-valued AR processes with Hermite innovations and time-varying parameters: An application to bovine fallen stock surveillance at a local scale. *Statistical Modelling*; **17(3)**: 172-195. This paper is available through the link:

`http://journals.sagepub.com/doi/10.1177/1471082X16683113`

CHAPTER 6

An exact goodness-of-fit test based on the occupancy problems to study zero-inflation and zero-deflation in biological dosimetry data

This chapter corresponds to the contents of: Fernández-Fontelo, A., Puig, P., Ainsbury, E.A. and Higueras, M. (2018). An exact goodness-of-fit test based on the occupancy problems to study zero-inflation and zero-deflation in biological dosimetry data. *Radiation Protection Dosimetry*: 1-10. This paper is available through the link:

https://academic.oup.com/rpd/article-abstract/179/4/317/4802349?redirectedFrom=fulltext

Bibliography

[1] Alba, A., Dórea, F.C., Arinero, L., Sánchez, J., Cordón, R., Puig, P. and Revie, C.W. (2015). Exploring the Surveillance Potential of Mortality Data: Nine Years of Bovine Fallen Stock Data Collected in Catalonia (Spain). *Plos One*; **27**.

[2] Allard, R. (1998). Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization*; **76(4)**: 327-333.

[3] Al-Osh, M. and Alzaid, A.A. (1987). First-Order Integer-Valued Autoregressive (INAR(1)) Process. *Journal of Time Series Analysis*; **8(3)**: 261-275.

[4] Al-Osh, M. (2009). The Impact of Missing Data in a Generalized Integer-Valued Autoregression Model for Count Data. *Journal of Biopharmaceutical Statistics*; **19(6)**: 1039-1054.

[5] Alzaid, A.A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure INAR(p) process. *Journal of Applied Probability*; **27(2)**: 314-324.

[6] Alfonso, J.H., Løvseth, E.K., Samant, Y. and Holm, J.Ø. (2015). Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis*; **72**: 398-421.

[7] Arendt, S., Rajagopal, L., Strohbehn, C., Stokes, N., Meyer, J. and Mandernach, S. (2013). Reporting of foodborne illness by U.S. consumers and healthcare profes-

sionals. *International journal of environmental research and public health*; **10(8)**: 3684-3714.

[8] Barndorff-Nielsen, O. (1965). Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*; **12(1)**: 115-121.

[9] Barron, D. (1992). The Analysis of Count Data: Over-dispersion and Autocorrelation. *Sociological Methodology*; **22**: 179-220.

[10] Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z. and Wang, Y. (2016). Bayesian Penalized Spline Models for the Analysis of Spatio-Temporal Count Data. *Statistics in Medicine*; **35(11)**: 1848-1865.

[11] Benaglia, T., Chauveau, D., Hunter, D.R. and Young, D. (2009). `mixtools`: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*; **32(6)**: 1-29.

[12] Bernard, H., Werber, D. and Höhle, M. (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**: 116.

[13] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA, MIT Press.

[14] Blischke, W.R. (1964). Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*; **59**: 510-528.

[15] Blischke, W.R. (1965). *Mixtures of discrete distributions, Classical and Contagious Discrete Distributions*. Statistical Publishing Society Oxford.

[16] Böhning, D. (1998). Zero-inflated Poisson models and C.A. MAN: a tutorial collection of evidence. *Biometrical J*; **7**: 833-843.

[17] Boulanger, M., Morlais, F., Bouvier, V., Galateau-Salle, F., Guittet, L., Marquignon, M.F., Paris, C., Raffaelli, C., Launoy, G. and Clin, B. (2015). Digestive

cancers and occupational asbestos exposure: incidence study in a cohort of asbestos plant workers. *Occupational and environmental medicine*; **72(11)**: 792-797.

[18] Brännäs, K. and Hellström, J. (2001). Generalized integer-valued autoregression. *Econom. Rev.*; **20(4)**: 425-443.

[19] Brass, W. (1958). Simplified methods of fitting the truncated negative binomial distribution. *Biometrika*; **45**: 59-68.

[20] Cameron, A.C. and Trivedi, P.K. (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics*; **1**: 29-53.

[21] Casella, G., and Berger, R. L. (2002). *Statistical inference*. Belmont, CA: Duxbury.

[22] Cardinal, M., Roy, R. and Lambert, J. (1999). On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine*; **18(15)**: 2025-2039.

[23] Cohen, A.C. (1954). Estimation of the Poisson parameter from truncated samples and from censored samples. *Journal of the American Statistical Association*; **49**: 158-168.

[24] Conti, S., Minelli, G., Ascoli, V., Marinaccio, A., Bonafede, M., Manno, V., Crialesi, R. and Straif, K. (2015). Peritoneal mesothelioma in Italy: Trends and geography of mortality and incidence. *American journal of industrial medicine*; **58(10)**: 1050-1058.

[25] Cox, D.R. and Snell, J.E. (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society. Series B*; **30(2)**: 248-275.

[26] Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. New York. Chapman and Hall/CRC.

[27] Cox, D.R. (1983). Some remarks on overdispersion. *Biometrika*; **70**: 269-274.

[28] Cox, D.R. and Snell, J.E. (1989). *Analysis of Binary Data*. London, Chapman and Hall.

[29] Crowcroft, N.S., Johnson, C., Chen, C., Li, Y., Marchand-Austin, A., Bolotin, S., Schwartz, K., Deeks, S.L., Jamieson, F., Drews, S., Russell, M.L., Svenson, L.W., Simmonds, K., Mahmud, S.M. and Kwon, J.C. (2018). Under- reporting of pertussis in Ontario: A Canadian Immunization Research Network (CIRN) study using capture-recapture. *PLoS ONE*; **13(5)**: e0195984.

[30] D'Agostino, R.B. and Stephens, M.A. (1986). *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

[31] David, F.N. and Johnson, N.L. (1952). The truncated Poisson distribution. *Biometrics*; **8**: 275-285.

[32] Del Castillo, J. and Pérez-Casany, M. (1998). Weighted Poisson Distributions for Overdispersion and Underdispersion Situations. *Annals of the Institute of Statistical Mathematics*; **50**: 567-585.

[33] DiCiccio, T.J. and Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*; **11 (3)**: 189-228.

[34] Dolphin, G.W. (1969). Biological Dosimetry with Particular Reference to Chromosome Aberration Analysis: A Review of Methods. International Atomic Energy Agency (IAEA): IAEA.

[35] Du, J.-G. and Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis*; **12(2)**: 129-142.

[36] Du, M.J., Forte, T., Cohen, M.M., Hyman, I. and Romans, S. (2005). Changing help-seeking rates for intimate partner violence in Canada. *Women Health.*; **41(1)**: 1-19.

[37] Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *J. Comp. Graphical Statist.*; **5**: 236-244.

[38] Dunne, E.F., Markowitz, L.E., Saraiya, M., Stokley, S., Middleman, A., Unger, E.R., Williams, A. and Iskander, J. (2014). CDC grand rounds: Reducing the burden of HPV-associated cancer and disease. *Morbidity and mortality weekly report*; **63(4)**: 69-72.

[39] Duron, S., Panjo, H., Bohet, A., Bigaillon, C., Sicard, S., Bajos, N., Meynard, J-B., Mérens, A. and Moreau, C. (2018). Prevalence and risk factors of sexually transmitted infections among French service members. *PLoS ONE*; **13(4)**: e0195158.

[40] Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*; **1(1)**: 54-75.

[41] Eilers, P.H.C., Marx, B.D. and Durban, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*; **39(2)**: 149-186.

[42] Faddy, M.J. and Bosch, R.J. (2001). Likelihood-Based Modeling and Analysis of Data Underdispersed Relative to the Poisson Distribution. *Biometrics*; **57(2)**: 620-624.

[43] Feller, W. (1957). *An Introduction to Probability Theory and Its Applications* (Vol. 1). Wiley.

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

[45] Fernández-Fontelo, A., Fontdecaba, S., Alba, A. and Puig, P. (2017). Integer-valued AR processes with Hermite innovations and time-varying parameters: An application to bovine fallen stock surveillance at a local scale. *Statistical Modelling*; **17(3)**: 172-195.

[46] Fernández-Fontelo, A., Puig, P., Ainsbury, E.A. and Higueras, M. (2018). An exact goodness-of-fit test based on the occupancy problems to study zero-inflation and zero-deflation in biological dosimetry data. *Radiation Protection Dosimetry*: 1-10.

[47] Fernández-Fontelo, A., Cabaña, A., Joe, H., Puig, P. and Moriña, D. Count time series models with under-reported data for gender-based violence in Galicia (Spain). Submitted.

[48] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Phil. Trans. R. Soc. Lond. A*; **222**: 309-368.

[49] Fisher, R.A. and Mather, K. (1936). A linkage test with mice, *Annals of Eugenics*; **7**: 265-280.

[50] Fisher, R.A. (1941). The negative binomial distribution, *Annals of Eugenics*; **11**: 182-187.

[51] Fisher, R.A. (1950). *Methods for Research Workers*. 11th edn. New York: Hafner Publishing Company.

[52] Fokianos, K. (2011). Some recent progress in count time series. *Statistics*; **45(1)**: 49-58.

[53] Fokianos, K., Gombay, E. and Hussein, A. (2014). Retrospective change detection for binary time series models. *Journal of Statistical Planning and Inference*; **145**: 102-112.

[54] Forney, G. (1973). The viterbi algorithm. *Proceedings of the IEEE*; **61(3)**: 268-278.

[55] Freeland, R. and McCabe, B. (2004). Forecasting discrete valued low count time series. *International Journal of Forecasting;* **20(3)**: 427-434.

[56] Frontario, S.C.N., Loveitt, A., Goldenberg-Sandau, A., Liu, J., Roy, D. and Cohen, L.W. (2015). Primary Peritoneal Mesothelioma Resulting in Small Bowel Obstruction: A Case Report and Review of Literature. *The American journal of case reports*; **16**: 496-500.

[57] Gamado, K.M., Streftaris, G. and Zachary S. (2014). Modelling under-reporting in epidemics. *J Math Biol.*; **69(3)**: 737-765.

[58] Ganio, L.M. and Schafer, D.W. (1992). Diagnostics for Overdispersion. *Journal of the American Statistical Association*; **87(419)**: 795-804.

[59] Gilbert, P. and Varadham, R. (2012). `numDeriv`: Accurate Numerical Derivatives (R package). `http://CRAN.R-project.org/package=numDeriv`

[60] Gomes, D. and Canto e Castro, L. (2009). Generalized integer-valued random coefficient for a first order structure autoregressive (RCINAR) process. *Journal of Statistical Planning and Inference*; **139**: 4088-4097.

[61] Gourieroux, C. and Jasiak, J. (2004). Heterogeneous INAR(1) model with application to car insurance. *Insurance: Mathematics and Economics*; **34(2)**: 177-192.

[62] Gracia, E. (2004). Unreported cases of domestic violence against women: towards and epidemiology of social silence, tolerance, and inhibition. *J. Epidemiol. Community Health*; **58**: 536-537.

[63] Greǵoire, E., Hadjidekova, V., Hristova, R., Gruel, F.G., Roch-Lefevre, S., Voisin, P., Staynova, A., Deleva, S., Ainsbury, E.A., Lloyd, D.C. and Barquinero, J.F. (2013). Biological dosimetry assessments of a serious radiation accident in Bulgaria in 2011. *Radiation Protection Dosimetry*; **155(4)**: 418-422.

[64] Greene, W.H. (1994). *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*. NYU Working Paper No. EC-94-10.

[65] Grogger, J.T. and Carson, R.T. (1991). Models for truncated counts. *Journal of applied Econometrics*; **6(3)**: 225-238.

[66] Harte, D. (2015). `HiddenMarkov`: Hidden Markov Models. `https://cran.r-project.org/web/packages/HiddenMarkov/HiddenMarkov.pdf`

[67] Hayata, I., Kanda, R., Minamihisamatsu, M., Furukawa, A. and Sasaki, M.S. (2001). *Proceedings of the International Symposium on the Criticality Accident in Takaimura: Medical aspects of Radiation Emergency*. Tsuji H. and Akushi M., Eds., National Institute of Radiological Science; 82-89.

[68] Higueras, M., Puig, P., Ainsbury, E.A., Vinnikov, V.A. and Rothkamm, K. (2016). A new bayesian model applied to cytogenetic partial body irradiation estimation. *Radiation Protection Dosimetry*; **168(3)**: 330-336.

[69] Hilbe, J.M. (2011). *Negative binomial regression*. New York, Cambridge University Press.

[70] Himmelmann, L. (2010). `HMM`: HMM-Hidden Markov Models. https://CRAN.R-project.org/package=HMM

[71] Hinde, J. (1982). *Compound poisson regression models. In R.* Springer-Verlag, New York.

[72] Hofman, M.S. (2013). Thyroid nodules: time to stop over-reporting normal findings and update consensus guidelines. *BMJ*; **347**: f5742.

[73] Höhle, M. and an der Heiden, M. (2014). Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics*; **70**: 993-1002.

[74] Holmes, J., Meier, P.S. , Booth, A., Guo, Y. and Brennan, A. (2012). The temporal relationship between per capita alcohol consumption and harm: A systematic review of time lag specifications in aggregate time series analyses. *Drug and Alcohol Dependence*; **123**: 7-14.

[75] Hudecová, Š. (2013). Structural changes in autoregressive models for binary time series. *J.Statist.Plann.Inference*; **143**: 1744-1752.

[76] Hudecová, Š., Hušková, M. and Meintanis, S. (2015a). Tests for time series of counts based on the probability generating function. *Statistics*; **49(2)**: 316-337.

[77] Hudecová, Š., Hušková, M. and Meintanis, S. (2015b). Detection of Changes in INAR Models. *Stochastic models, statistics and their applications, pp. 11-18.* Springer Proceedings in Mathematics and Statistics 122.

[78] IAEA (2011). *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies.* International Atomic Energy Agency.

[79] Imai, M. and Hino, O. (2015). Environmental carcinogenesis - 100th anniversary of creating cancer. *Cancer science*; **106(11)**: 1483-1485.

[80] Jazi, M.A., Jones, G. and Lai, C.D. (2012a). First-order integer valued AR processes with zero inflated Poisson innovations. *J. Time Series Anal.*; **33(6)**: 954-963.

[81] Jazi M.A., Jones, G. and Lai, C. (2012b). Integer Valued AR(1) with Geometric Innovations *Journal of The Iranian Statistical Society*; **11**: 173-190.

[82] Johnson, N.L. and Kotz, S. (1969). *Distributions in Statistics—Discrete Distributions*. John Wiley and Sons, New York.

[83] Johnson, N.L. and Kotz, S. (1977). *Urn Models and Their Applications*. John Wiley and Sons, New York.

[84] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York Wiley.

[85] Johnson, N.L. and Kotz, S. (1982). Developments in Discrete Distributions, 1969-1980. *International Statistical Review*; **50**: 70-101.

[86] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics, New Jersey.

[87] Jung, R.C. and Tremayne, A. (2006). Binomial thinning models for integer time series. *Statistical Modelling*; **6(2)**: 81-96.

[88] Kaciroti, N.A., Raghunathan, T.E., Schork, M.A. and Clark, N.M. (2008). A Bayesian model for longitudinal count data with non-ignorable dropout. *Journal of the Royal Statistical Society. Series C, Applied Statistics*; **57(5)**: 521-534.

[89] Karlis, D. and Xekalaki, E. (1998). Minimum Hellinger distance estimation for Poisson mixtures. *Computational Statistics and Data Analysis*; **29**: 81-103.

[90] Karlis, D, and Xekalaki, E. (1999). On testing for the number of components in finite Poisson mixtures. *Annals of the Institute of Statistical Mathematics*; **51**: 149-161.

[91] Karlin, S. and Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, London.

[92] Kemp C.D. and Kemp A.W. (1965). Some Properties of the Hermite Distribution. *Biometrika Trust*; **52**: 381-394.

[93] Kemp, A.W. (1978). On probability generating functions for matching and occupancy distributions. *Zastosowania Matematyki*; **16**: 207-213.

[94] Koepke, R., Sobel, J. and Arnon, S.S. (2008). Global occurrence of infant botulism, 1976-2006. *Pediatrics*; **122(1)**: e73-82.

[95] Maxima, a Computer Algebra System. (2014). Version 5.34.1. `http://maxima.sourceforge.net/`

[96] Maiti, R., Biswas, A., Guha, A. and Ong, S.H. (2014). Modelling and coherent forecasting of zero-inflated count time series. *Statistical Modelling*; **14(5)**: 374-398.

[97] McCabe, B.P.M., Martin, G.M. and Harris, D. (2011). Efficient probabilistic forecasts for counts *Journal of the Royal Statistical Society: Series B*; **73(2)**: 253-272.

[98] McGregor, K., Makkai, T. and Graycar, A. (2003). Self-Reported Drug Use: How Prevalent Is Under-Reporting?. *Trends and Issues in Crime and Criminal Justice*; no. 260.

[99] McKenzie, E. (1985). Some simple-models for discrete variate time-series. *Water Resources Bulletin*; **21(4)**: 645-650.

[100] McKenzie, E. (2003). *Stochastic Processes: Modelling and Simulation, Handbook of Statistics*, vol. 21. Elsevier.

[101] Mian, R. and Paul, S. (2016). Estimation for zero-inflated over-dispersed count data model with missing response. *Statistics in Medicine*; **35(30)**: 5603-5624.

[102] Mood, A.M. (1950). *Introduction to the Theory of Statistics*. New York, McGraw-Hill.

[103] Monteiro, M., Scotto, M.G. and Pereira, I. (2010). Integer-valued autoregressive processes with periodic structure. *Journal of Statistical Planning and Inference*; **140(6)**: 1529-1541.

[104] 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(26)**: 3125-3136.

[105] Moriña, D., Higueras, M., Puig, P. and Oliveira, M. (2015a). `hermite`: Generalized Hermite Distribution (R package), `http://CRAN.R-project.org/package=hermite`

[106] Moriña, D., Higueras, M., Puig, P. and Oliveira, M. (2015b). Generalized hermite distribution modelling with the R package hermite. *The R Journal*; **7**: 263-274.

[107] Neelon B., Gosh P. and Loebs P. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *J. Roy. Stat. Soc. A.*; **176**: 389-413.

[108] Nicholson, W.L. (1961). Occupancy probability distribution critical points. *Biometrika*; **48**: 175-180.

[109] Oliveira, P., Rodrigues, F., Henriques P. and Galhardas H. (2005). A Taxonomy of Data Quality Problems. *In 2nd Int. Workshop on Data and Information Quality (in conjunction with CAiSE 2005)*, Porto, Portugal.

[110] Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothkamm, K. (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*; **58(2)**: 259-279.

[111] Palermo, T., Bleck, J. and Peterman, A. (2014). Tip of the Iceberg: Reporting and Gender-Based Violence in Developing Countries. *American Journal of Epidemiology*; **179(5)**: 602-612.

[112] Parzen, E. (1960). *Modern Probability Theory and Its Applications*. Wiley Series, USA.

[113] Park, E.K., Takahashi, K., Hoshuyama, T., Cheng, T.J., Delgermaa, V., Le, G.V. and Sorahan, T. (2011). Global magnitude of reported and unreported mesothelioma. *Environmental health perspectives*; **119(4)**: 514-518.

[114] Pavlopoulos, H. and Karlis, D. (2008). INAR(1) modeling of overdispersed count series with an environmental application. *Environmetrics*; **19**: 369-393.

[115] Pedeli, X., Hoek, G. and Katsouyanni, K. (2011). Risk assessment of diesel exhaust and lung cancer: combining human and animal studies after adjustment for biases in epidemiological studies. *Environmental Health*; **10**: 30.

[116] Pedeli, X. and Karlis, D. (2013). Some properties of multivariate INAR(1) processes. *Computational Statistics and Data Analysis*; **67**: 213-225.

[117] Perrin, J.B., Ducrot, C., Vinard J.L., Morignat, E., Calavas, D. and Hendrick, P. (2012). Assessment of the utility of routinely collected cattle census and disposal data for syndromic surveillance. *Preventive Veterinary Medicine*; **105**: 244-252.

[118] Petersen, R., Petersen, J.A. and Mikkelsen, S. (2015). Non-occupational pleural mesothelioma. *Ugeskrift for laeger*; **177(3)**: V09140480.

[119] Pitarque, S., Clèries, R., Martínez, J.M., López-Abente, G., Kogevinas, M. and Benavides, F.G. (2008). Mesothelioma mortality in men: trends during 1977-2001 and projections for 2002-2016 in Spain. *Occupational and environmental medicine*; **65(4)**: 279-282.

[120] Puig, P. and Valero, J. (2006). Count data distributions. *J. Am. Stat. Assoc.*; **101**: 332-340.

[121] Puig, P. and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*; **13(2)**: 544-555.

[122] Puig, P. and Barquinero, J.F. (2011). An application of compound Poisson modelling to biological dosimetry. *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences*; **467(2127)**: 897-910.

[123] Pujol M., Barquinero J.F., Puig P., Puig R., Caballín M.R. and Barrios L. (2014). A New Model of Biodosimetry to Integrate Low and High Doses. *PlosOne*; 1-19.

[124] Pujol M., Barrios L., Puig P., Caballín M.R. and Barquinero J.F. (2016). A new model for biological dose-assessment in cases of heterogeneous exposures to ionizing radiation. *Radiation Research*; **185(2)**: 151-162.

[125] Rao, C.R. and Chakravarti, I.M. (1956). Some small sample tests of significance for a Poisson distribution. *Biometrics*; **12**: 264-282.

[126] Rao, B.R., Mazumdar, S., Waller, J.H. and Li, C.C. (1973). Correlation between the numbers of two types of children in a family, *Biometrics*; **29**: 271-279.

[127] Ridout, M., Hinde, J. and Demétrio, C.G. (2001). A score test for testing a Zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*; **57(1)**: 219-223.

[128] Rocourt, J., Moy, G., Vierk, K. and Schlundt, J. (2003). The present state of foodborne disease in OECD countries. *Technical Report, Food Safety Department-World Health Organization, Geneva, Switzerland.*

[129] Rosenman, K.D., Kalush, A., Reilly, M.J., Gardiner, J.C., Reeves, M. and Luo, Z. (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.

[130] Sasaki, M.S. and Miyata, H. (1968). Biological dosimetry in atom bomb survivors. *Nature*; **220**: 1189-1193.

[131] Sasaki, M.S., Takatsuji, T., Ejima, Y., Kodama, S. and Kido, C. (1987). Chromosome aberration frequency and radiation dose to lymphocytes by alpha-particles from internal deposit of Thorotast. *Radiat. Environ. Biophys*; **26**: 227-238.

[132] Sasaki, M.S. (2003). Chromosomal biodosimetry by unfolding a mixed Poisson distribution: a generalized model. *Int. J. Radiat. Biol.*; **79(2)**: 83-97.

[133] Schweer, S. and Weiß, C.H. (2014). Compound Poisson INAR(1) processes: stochastic properties and testing for overdispersion. *Comput. Statist. Data Anal.*; **77**: 267-284.

[134] Scotto, M.G., Weiß, C.H. and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling*: **15(6)**: 590-618.

[135] Sellers, K.F. and Shmueli, G. (2010). A Flexible Regression Model for Count Data. *Annals of Applied Statistics*; **4**: 943-961.

[136] Shanbhag, D.N. and Rajamannar, G. (1974). Some characterizations of the bivariate distribution of independent Poisson variables. *Australian and New Zealand Journal of Statistics*; **16(2)**: 119-125.

[137] Steutel, F.W. and Van Harn, K. (1979). Discrete analogs of self-decomposability and stability. *Annals of Probability*; **7(5)**: 893-899.

[138] Tamblyn, S.E. (2000). The frustrations of fighting foodborne disease. *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne*; **162(10)**: 1429-1430.

[139] Tapprest, J., Morignat, E., Dornier, X., Borey, M., Hendrikx, P., Ferry, B., Calavas, D. and Sala, C. (2017). Fallen stock data: An essential source of information for quantitative knowledge of equine mortality in France. *Equine Veterinary Journal*; **49(5)**: 596-602.

[140] Tukey, J.W. (1949). Moments of random group size distributions, *Annals of Mathematical Statistics*; **20**: 523-539.

[141] United Nations. Declaration on the elimination of violence against women (1993). New York: United Nations General Assembly.

[142] van den Broek, J. (1995). A score test for zero inflation in a poisson distribution. *Biometrics*; **51(2)**: 738-743.

[143] Vinnikov, V.A., Ainsbury, E., Maznyk, N.A. and Rothkamm, K. (2010). Limitations associated with analysis of cytogenetic data for biological dosimetry. *Radiation Research*; **174(4)**: 403-414.

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

[145] Watts, C. and Zimmerman, C. (2002). Violence agains women: global scope and magnitude. *Lancet*; **359**: 1232-1237.

[146] Weiß, C.H. (2008). Thinning operations for modeling time series of counts-a survey. *Advances in Statistical Analysis*; **92(3)**: 319-341.

[147] Weiß, C.H. (2013). Integer-valued Autoregressive Models for Counts Showing Underdispersion. *Journal of Applied Statistics*; **40(9)**: 1931-1948.

[148] Weiß, C.H. and Kim, H.Y. (2013). Parameter estimation for binomial AR(1) models with applications in finance and industry. *Stat. Papers*; **54**: 563-590.

[149] Weiß, C.H. and Puig, P. (2015). The marginal distribution of compound Poisson INAR(1) processes. *Stochastic Models, Statistics and Their Applications*; **122**: 351-359.

[150] Weiß, C.H. and Schweer, S. (2015). Detecting overdispersion in INARCH(1) processes. *Statistica Neerlandica*; **69(3)**: 281-297.

[151] Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*; **21(4)**: 575-587.

[152] Wirtz, A.L., Glass, N., Pham, K., Aberra, A., Rubenstein, L.S., Singh, S. and Vu, A. (2013). Development of a screening tool to identify female survivors of gender-based violence in a humanitarian setting: qualitative evidence from research among refugees in Ethiopia. *Confl Health.*; **7(1)**: 1-13.

[153] World Health Organization. *World report on violence and health*. World Health Organization, Geneva, 2002.

[154] Zheng, H., Basawa, I.V. and Datta, S. (2005). Inference for *p*th-order random coefficient integer-valued autoregressive processes. *Journal of Time Series Analysis*; **27(3)**: 411-440.

[155] Zheng, H., Basawa, I.V. and Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference*; **173**: 212-229.

[156] Zhu, R. and Joe, H. (2006). Modelling count data time series with markov processes based on binomial thinning. *Journal of Time Series Analysis*; **27(5)**: 725-738.

[157] Zhu, R. and Joe, H. (2010). Negative binomial time series models based on expectation thinning operators. *Journal of Statistical Planning and Inference*; **140(7)**: 1874-1888.

[158] Zucchini, W. and MacDonald, I.L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press.