Wavelet-based detection of outliers in Poisson INAR(1) time series*

Isabel Silva and Maria Eduarda Silva

Abstract The presence of outliers or discrepant observations has a negative impact in time series modelling. This paper considers the problem of detecting outliers, additive or innovational, single, multiple or in patches, in count time series modelled by first-order Poisson integer-valued autoregressive, PoINAR(1), models. To address this problem, two wavelet-based approaches that allow the identification of the time points of outlier occurrence are proposed. The effectiveness of the proposed methods is illustrated with synthetic as well as with an observed dataset.

1 Introduction

Time series, as any other data, may contain outliers which are observations that look discordant from most of the observations in the dataset. Neglecting the presence of outliers in a time series hinders statistical inference, leading to model misspecification and biased parameter estimation. Since the seminal work of Fox [7] two major approaches for dealing with outliers in time series, may be distinguished. One approach advocates the use of robust estimators to reduce the effect of the outlying observations. However, this approach often leads to ignoring observations hence eventually masking the presence of important underlying phenomena, precluding risk analysis. Alternatively, several methodologies for detecting and estimating outliers and other intervention effects have been established for ARMA models. The

* Accepted authors manuscript (AAM) published in Recent Studies on Risk Analysis and Statistical Modeling, Contributions to Statistics, 2018. DOI: 10.1007/978-3-319-76605-8_13. Final source webpage: https://doi.org/10.1007/978-3-319-76605-8_13

Isabel Silva
Faculdade de Engenharia, Universidade do Porto and CIDMA, Portugal e-mail: ims@fe.up.pt

Maria Eduarda Silva
Faculdade de Economia, Universidade do Porto and CIDMA, Portugal e-mail: mesilva@fep.up.pt

emphasis has been on iterative procedures and likelihood based statistics, see for instance Chang et al. [5], Chen and Liu [6] and Tsay [17]. Also several tailored procedures have been proposed to some nonlinear time series models. However, the problem of detection and estimation of outliers in time series of counts has received less attention in the literature. Count time series occur in many areas such as telecommunications, actuarial science, epidemiology, hydrology and environmental studies where the detection of outliers may be invaluable in risk assessment.

One of the most popular classes of models for time series of counts is the class of INAR models proposed by Al-Osh and Alzaid [1] and McKenzie [11], extensively studied in the literature and applied to many real-world problems because of its easiness of interpretation. These models are apparently autoregressive models in which the usual multiplication has been replaced by a random operation, called thinning operation (for details see Scotto et al. [13]) and the innovations are discrete-valued random variables. Barczy et al. [2, 3] proposed Conditional Least Squares estimation of the INAR(1) model parameters contaminated with outliers additive and innovational, assuming that the time points of the outliers occurrence are known, but their sizes are unknown. Recently, Silva and Pereira [15] suggested a Bayesian approach in order to detect additive outliers in PoINAR(1) models.

In this work, procedures to identify the times of outlier occurrence in PoINAR(1) time series using wavelets are proposed. Wavelets are basis functions that combine properties such as localization in time and scale, orthonormality, different degrees of smoothness, compact support and fast implementation, for details see Percival and Walden [12]. In particular, Discrete Wavelet Transform (DWT), which is a powerful tool for a time-scale multi-resolution analysis, is applied. DWT can be considered as filters of different cut-off frequencies used to analyse a signal at different scales. In a first approach, similar to that of Grané and Veiga [8], the so called detail coefficients derived from DWT, using the Haar wavelet, are compared with a threshold. In a second approach, the parametric resampling method of Tsay [18] is used in order to obtain the empirical distribution of these detail coefficients.

The remainder of this work is organized as follows. Section 2 presents the first-order Poisson Integer-valued AutoRegressive model contaminated with additive and innovational outliers. A brief description of wavelets and DWT is given in Section 3. The proposed wavelet-based procedures to detect time of outlier occurrence are explained in Section 4. The proposed procedures are illustrated and compared with synthetic data in Section 5. Furthermore, the methods are also applied on an observed dataset. Finally, Section 6 concludes the paper.

2 Poisson INAR(1) model contaminated with outliers

Motivated by the need of modelling correlated series of counts, several models for integer-valued time series were proposed in the literature. One of them is the INteger AutoRegressive model proposed by Al-Osh and Alzaid [1] and McKenzie [11]. This model is based on the binomial thinning operation, proposed by Steutel and Van

Harn [16], which is defined on a non negative integer-valued random variable X by $\alpha \circ X = \sum_{k=1}^{X} Y_k$, where $\alpha \in [0,1]$ and $\{Y_k\},\ k = 1,\ldots,X$, is a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables, independent of X, with $P(Y_k = 1) = 1 - P(Y_k = 0) = \alpha$. This sequence is called the counting series of $\alpha \circ X$. Note that, $\alpha \circ X | X \sim \text{Bi}(X, \alpha)$. For an account of the properties of the thinning operation see Silva and Oliveira [14].

Let $\{X_t\}$ be a discrete time, positive integer-valued stochastic process. It is said to be a PoINAR(1) process if it satisfies the following equation,

$$X_t = \alpha \circ X_{t-1} + e_t, \tag{1}$$

where $e_t \sim \text{Poisson}(\lambda)$, is the so called arrival process, $0 < \alpha < 1$, and for each t, all counting series of $\alpha \circ X_{t-1}$ are mutually independent and independent of $\{e_t\}$. Under these conditions, the process is strictly stationary and $X_t \sim \text{Poisson}(\frac{\lambda}{1-\alpha})$ if $X_0 \sim \text{Poisson}(\frac{\lambda}{1-\alpha})$.

A time series is affected by an additive outlier (AO) if an external error or exogenous change occurs on a certain time point, affecting only this observation and not entering the dynamics of the process. Formally, a contaminated PoINAR(1) with $I \in \mathbb{N}$ additive outliers with magnitude $\omega_i \in \mathbb{N}$ at time points $s_i \in \mathbf{N}$, $i = 1,\ldots,I$ can be defined as follows

$$Y_t = X_t + \sum_{i=1}^{I} \delta_{i,s_i} \omega_i,$$

where X_t is a PoINAR(1) model satisfying (1) and $\delta_{k,m} = 1$, if $k = m$; $\delta_{k,m} = 0$, if $k \neq m$, is an indicator function.

On the other hand, an innovational outlier (IO) can be considered as an internal change or endogenous effect on the noise process, affecting all subsequent observations. Thus, the observed time series $Y_1,\ldots,Y_n$ is a PoINAR(1) process contaminated with $I \in \mathbb{N}$ innovational outliers with size ω_i at time points s_i, $i = 1,\ldots,I$ if it satisfies the following equation

$$Y_t = \alpha \circ Y_{t-1} + \eta_t,$$

with $\eta_t = e_t + \sum_{i=1}^{I} \delta_{i,s_i} \omega_i$, where $e_t \sim \text{Poisson}(\lambda)$ and I, s_i, ω_i and $\delta_{k,m}$ are defined as before.

Note that in both cases, the underlying outlier free process X_t is unobserved.

3 Brief description of discrete wavelet transform

A wavelet is a function that can be considered as a small wave which grows and decays in a limited time period, for details see Percival and Walden [12]. Simi-

larly to Fourier analysis that uses sinusoidal functions to find the frequency components contained in a signal, wavelet analysis uses shifted and scaled versions of a so called wavelet mother to provide the time localization of each spectral component. Formally, a (mother) wavelet is any real-valued function $\psi(\cdot)$ defined on $\mathbb{R}$ satisfying $\int_{-\infty}^{\infty} \psi(u)\, du = 0$, $\int_{-\infty}^{\infty} \psi^2(u)\, du = 1$, and $0 < \int_0^{\infty} \frac{|\mathbf{\Psi}(f)|^2}{f}\, df < \infty$, where $\mathbf{\Psi}(f) = \int_{-\infty}^{\infty} \psi(u)\, \mathrm{e}^{-i2\pi f u}\, du$ is the Fourier transform of $\psi(\cdot)$.

Following Percival and Walden [12], let $\mathbf{X} = \{X_t, t = 0, \ldots, N-1\}$ be a time series (or signal), with $N = 2^J$, $J \in \mathbb{N}$. The DWT coefficients $\mathbf{W} = \{W_n, n = 0, \ldots, N-1\}$ are defined by

$$\mathbf{W} = \mathscr{W}\mathbf{X} \quad \Leftrightarrow \quad [\mathbf{W}_1 \ldots \mathbf{W}_J\, \mathbf{V}_J]^T = [\mathscr{W}_1 \ldots \mathscr{W}_J\, \mathscr{V}_J]^T \mathbf{X},$$

where $\mathscr{W}$ is a $N \times N$ orthonormal matrix of dilations and translations of the mother wavelet $\psi(\cdot)$, defined as $\frac{1}{\sqrt{d}} \psi\left(\frac{u-t}{d}\right)$ with dilation d and translation t parameters taking dyadic values, i.e., $d = 2^j$ and $t = k2^j$, for $j, k \in \mathbb{Z}$. Note that, for $j = 1, \ldots, J$, $\mathbf{W}_j$ is a column vector with $N/2^j$ elements that contains all the DWT coefficients for scale $\tau_j = 2^{j-1}$, $\mathbf{V}_J$ contains the scaling coefficients $\mathbf{W}_{N-1}$, associated with average on scale $d_J = 2^J$, $\mathscr{W}_j$ has dimension $N/2^j \times N$ and $\mathscr{V}_J$ is $1 \times N$.

The wavelet coefficients of white noise or Gaussian data are themselves white noise or Gaussian random variables, respectively, see Percival and Walden [12]. Furthermore, as referred by Bilen and Huzurbazar [4] and Percival and Walden [12], wavelet coefficients in $\mathbf{W}_j$ are approximately uncorrelated even when the data is highly correlated and they allow the reconstruction of the time series. The synthesis of $\mathbf{X}$ (inverse DWT) is given by $\mathbf{X} = \mathscr{W}^T\mathbf{W} = \sum_{j=1}^{J} \mathscr{W}_j^T \mathbf{W}_j + \mathscr{V}_J^T \mathbf{V}_J = \sum_{j=1}^{J} \mathscr{D}_j + \mathscr{A}_J$, where $\mathscr{D}_j$ is called the jth *level wavelet detail* and $\mathscr{A}_J$ has all its elements equal to the sample mean of the time series. For $1 \le j \le J-1$, the jth *level wavelet smooth* is $\mathscr{A}_j = \sum_{k=j+1}^{J} \mathscr{D}_k + \mathscr{A}_J$, and can be considered as an approximation (smoother version) of $\mathbf{X}$.

In practice, the discrete wavelet transform (DWT) matrix $\mathbf{W}$ is computed through a so called pyramid algorithm introduced by Mallat [9] that uses linear filtering and downsampling operations. More specifically, for a even width L, consider a wavelet filter $\{h_l : l = 0, \ldots, L-1\}$, which is a high-pass filter, and a scaling filter $g_l = (-1)^{l+1} h_{L-1-l}$, that is a low-pass filter. In the first step of the pyramidal algorithm, two sets of coefficients are produced by the convolution of $\mathbf{X}$ with the low-pass filter $\{g_l\}$ (producing the first level approximation coefficients $c\mathbf{A}_1$) and with the high-pass filter $\{h_l\}$ (deriving the first level detail coefficients $c\mathbf{D}_1$), and then a downsample is performed (retain every other filtered value). The next step divides the first level approximation coefficients in two sequences using the same procedure, replacing $\mathbf{X}$ by $c\mathbf{A}_1$ and computing $c\mathbf{A}_2$ and $c\mathbf{D}_2$. Therefore, at level j, the decomposition of $\mathbf{X}$ has the following structure $[c\mathbf{A}_j, c\mathbf{D}_j, c\mathbf{D}_{j-1}, \ldots, c\mathbf{D}_1]$.

The detail coefficients capture certain features of the time series, such as sudden changes, peaks, or spikes, presenting large values in the presence of these singularities, and therefore they can be used to detect outliers. In general, the first level

of decomposition is enough to analyse time series contaminated with outliers Bilen and Huzurbazar [4] and Grané and Veiga [8].

There are many mother wavelets. In this work, the Haar wavelet (among the many mother wavelets) is used. Since it can be considered as a square wave defined by

$$\psi(t) = \begin{cases} -1/\sqrt{2}, & -1 \leq t \leq 0 \\ 1/\sqrt{2}, & 0 < t \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

the Haar wavelet is more suitable for count data. In this case, low-pass filters correspond to moving averages of the observations and high-pass filters correspond to moving differences of the observations.

4 Procedures to detect the time of outliers occurrence

In this section, two wavelet-based methods for detecting the time of occurrence of outliers in PoINAR(1) processes are described. The procedures can be summarised in the following steps:

Step 1 Given an observed time series of counts, $\mathbf{Y} = \{Y_t, t = 0, \ldots, N\}$, fit a PoINAR(1) model and compute the resulting Pearson residuals[3] $\mathbf{Z} = \{\hat{z}_t, t = 1, \ldots, N-1\}$, given by $\hat{z}_t = \dfrac{Y_t - (\hat{\alpha} Y_{t-1} + \hat{\lambda})}{\sqrt{\hat{\alpha}(1-\hat{\alpha}) Y_{t-1} + \hat{\lambda}}}$.

Step 2 The DWT is applied to the Pearson residuals to obtain the first level detail coefficients, $c\mathbf{D}_1 = (d_1, d_2, \ldots, d_{N/2})$.

Step 3a **Threshold approach**:

- **(i)** Set the threshold k_1^a (discussed in Subsection 4.1).
- **(ii)** The set of (ordered) indices, $\mathbf{S} = \{s_1, \ldots, s_I\}$, containing the positions of the detail coefficients which are above the threshold k_1^a is obtained. As in Grané and Veiga [8], the problem of masking[4] is avoided by searching the outliers recursively. This means that for each outlier detected, $\mathbf{Z}$ is reconstructed applying the inverse discrete wavelet transform (IDWT) to modified detail coefficients where the largest (in absolute value) detail coefficient above the threshold is set to zero. The procedure ends when no more outliers are detected.

Step 3b **Parametric resampling approach**:

- **(i)** Compute the acceptance envelope (discussed in Subsection 4.2).

[3] $Z_t = \frac{Y_t - \mathrm{E}[Y_t|Y_{t-1}]}{\sqrt{\mathrm{Var}(Y_t|Y_{t-1})}}$

[4] Masking occurs when one outlier prevents others from being detected.

(ii) The set of (ordered) indices, $\mathbf{S} = \{s_1, \ldots, s_I\}$, containing the positions of the detail coefficients which are outside of the acceptance envelope is calculated.

Step 4 The exact position of the outlier in the residual series is obtained as in Grané and Veiga [8]: let s be a generic element of $\mathbf{S}$, compute the sample mean of $\mathbf{Z}$ without the observations $2s$ and $2s-1$, given by $\bar{z}_{N-2} = \dfrac{1}{N-2} \sum_{i \neq 2s, 2s-1} \hat{z}_i$; the time of the outlier occurrence in the residual series is $2s$ if $|\hat{z}_{2s} - \bar{z}_{N-2}| > |\hat{z}_{2s-1} - \bar{z}_{N-2}|$, or equal to $2s-1$ otherwise.

As noted by Bilen and Huzurbazar [4] and Grané and Veiga [8], the first level coefficients detect only the beginning of an outliers patch and therefore, when searching for patches of outliers it is necessary to use the second level detail coefficients, $c\mathbf{D}_2$. Thus, in **Step 3a** there are two thresholds $k_1^{a_1}$ and $k_2^{a_2}$, corresponding to the first and second levels of detail coefficients, respectively. Similarly, there are two acceptance envelopes, one for $c\mathbf{D}_1$ and one for $c\mathbf{D}_2$, in **Step 3b**.

4.1 Setting the threshold

In the non-Gaussian context of this work, there are no results available for the distribution of the detail coefficients. Thus Monte Carlo simulations are used to obtain the empirical distribution of the maximum of the detail coefficients (in absolute value) for the Pearson residuals of PoINAR(1) models. Then a threshold is computed as follows. For each (α, λ) in the set $\{(\alpha, \lambda) : \alpha = (2k+1) \times 10^{-1}, k = 0, \ldots, 4; \lambda = 2k+1, k = 0, \ldots, 14\}$, 20000 replications of the corresponding PoINAR(1) process are generated for each sample size $N = 2^J + 1$, for $J = 7, \ldots, 10$. The model is fitted, the Pearson residuals, $\hat{z}_i$, for $i = 1, \ldots, N-1$, are computed and the maximum of the first and second level detail coefficients are obtained. The thresholds $k_1^{\mathrm{a}_1}$ and $k_2^{\mathrm{a}_2}$ are set as the $100(1-\mathrm{a})$th percentiles of the corresponding empirical distributions, for $\mathrm{a} = \mathrm{a}_1$ or $\mathrm{a} = \mathrm{a}_2$. The results[5] indicate that the thresholds vary not only with the sample size N but also with the specific combination of the parameters α and λ. Therefore, adopting a conservative strategy, for each sample size N the thresholds are set to the minimum obtained for all the combinations of parameters in each level of decomposition. The obtained thresholds are shown in Table 1.

4.2 Computing the acceptance envelope

Tsay [18] proposed to obtain the empirical distribution of a chosen functional using bootstrap samples generated from a fitted model, and then compare the observed

[5] Available from the authors.

Table 1 Empirical threshold values corresponding to 90th and 95th percentiles of the maximum of the detail coefficients (first and second level), in absolute value, for PoINAR(1) Pearson residuals.

N	128	256	512	1024
$k_1^{0.05}$	3.469	3.694	3.886	4.118
$k_1^{0.1}$	3.182	3.450	3.657	3.840
$k_2^{0.05}$	3.157	3.347	3.518	3.691
$k_2^{0.1}$	2.936	3.138	3.320	3.504

value for the series with this distribution. For this purpose, an acceptance envelope is obtained from the $100(1-\mathsf{a}/2)$th and $100\mathsf{a}/2$th percentiles of this empirical distribution. If the fitted model is adequate, the functional of interest of the original data should be within the envelope. In this work, the functionals of interest are the first and second level detail coefficients of the Pearson residuals of PoINAR(1) model. Thus, for several sample sizes $N = 2^J + 1, J = 7,8,9$, and parameter values $\{(\alpha,\lambda) : \alpha \in \{0.1,0.5,0.9\}; \lambda \in \{1,5,9,13\}\}$, 20000 realizations of PoINAR(1) process are generated and the corresponding Pearson residuals are estimated. For each series of Pearson residuals, the DWT is applied to obtain the first and second level detail coefficients, $c\mathbf{D}_1$ and $c\mathbf{D}_2$, and the acceptance envelopes are constructed from the 0.01th and 99.99th percentiles[6] of the empirical distribution of $c\mathbf{D}_1$ and $c\mathbf{D}_2$, respectively. Once again, the results[7] show that the acceptance envelopes vary not only with the sample size N but also with the combination of the parameter values (α,λ). Therefore, assuming a conservative strategy, for each sample size, an acceptance envelope with the minimum amplitude is chosen. The acceptance envelopes are available from the authors upon request.

5 Simulation study and illustration

This section presents the results of a simulation study designed to evaluate and compare the performance of the procedures described above (implemented in Matlab [10]). For these purposes, the percentage of correct detections and the average number of false outliers detected in 1000 repetitions are computed. In each repetition, a realization of a PoINAR(1) process with parameters in the set $\{(\alpha,\lambda) : \alpha \in \{0.1,0.5,0.8\}; \lambda \in \{1,3,5\}\}$ is contaminated with single (1) or multiple (3) outliers either additive or innovational, randomly placed, with integer-valued magnitude $\omega = \lceil 5\sigma_X \rceil, \lceil 10\sigma_X \rceil$, where $\lceil x \rceil$ is the smallest integer greater than or equal to x. The Pearson residual series are obtained and the procedures described in Section 4 are applied. Several sample sizes are considered, $N = 128,256,512$. Some

[6] In the performed simulation study, the detail coefficients present a large variability. Therefore, as a compromise between correct and false detection of outliers, it is found that a reasonable acceptance envelope is constructed from the 0.01th and 99.99th extreme percentiles.

[7] Available from the authors.

of the results are shown in Tables 2 and 3 for the threshold $k_1^{0.05}$ and the acceptance envelope constructed from the 0.01th and 99.99th percentiles of the empirical distribution of $c\mathbf{D}_1$.

Table 2 Percentage of correct detections and average number of false outliers detected, in 1000 repetitions of PoINAR(1) models with sample sizes $N+1$ for some parameter values, contaminated with 1 additive outlier or 1 innovational outlier, with magnitude $\lceil 5\sigma_X \rceil$ and $\lceil 10\sigma_X \rceil$.

			1 Additive Oulier				1 Innovational Oulier			
			% Correct		Average False		% Correct		Average False	
(α,λ)	N	ω	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.
(0.1, 1)	128	$\lceil 5\sigma_X \rceil = 6$	81.8	72.5	0.088	0.05	69.9	63.4	0.092	0.069
		$\lceil 10\sigma_X \rceil = 11$	98.2	97.8	0.07	0.05	99.7	98.8	0.094	0.061
	256	$\lceil 5\sigma_X \rceil = 6$	64	81.8	0.114	0.128	67.4	63.6	0.168	0.147
		$\lceil 10\sigma_X \rceil = 11$	98.7	99.1	0.102	0.105	99.9	99	0.122	0.144
	512	$\lceil 5\sigma_X \rceil = 6$	78.1	91.8	0.185	0.268	60.3	66.7	0.163	0.293
		$\lceil 10\sigma_X \rceil = 11$	100	100	0.166	0.239	100	100	0.18	0.284
(0.5, 3)	128	$\lceil 5\sigma_X \rceil = 13$	73	99	0.047	0.03	73.6	63.4	0.096	0.046
		$\lceil 10\sigma_X \rceil = 25$	100	99.9	0.002	0.013	99.9	100	0.077	0.049
	256	$\lceil 5\sigma_X \rceil = 13$	64.7	99.6	0.064	0.059	67.4	84.2	0.098	0.086
		$\lceil 10\sigma_X \rceil = 25$	99.8	99.9	0.085	0.143	100	100	0.132	0.103
	512	$\lceil 5\sigma_X \rceil = 13$	98.5	99.3	0.095	0.152	64.9	86.1	0.123	0.26
		$\lceil 10\sigma_X \rceil = 25$	99.7	100	0.158	0.087	100	100	0.113	0.225
(0.8, 5)	128	$\lceil 5\sigma_X \rceil = 25$	97.9	97.7	0.023	0.026	98.1	95.7	0.049	0.04
		$\lceil 10\sigma_X \rceil = 50$	100	100	0.51	0	100	100	0.053	0.028
	256	$\lceil 5\sigma_X \rceil = 25$	91.5	94.4	0.391	0.404	97.2	96.1	0.071	0.064
		$\lceil 10\sigma_X \rceil = 50$	100	100	0	0	100	100	0.059	0.067
	512	$\lceil 5\sigma_X \rceil = 25$	92.5	96.5	0.524	0.087	98.7	98.9	0.068	0.156
		$\lceil 10\sigma_X \rceil = 50$	100	100	0.001	0.004	100	100	0.077	0.12

For the case of contamination with 1 outlier (Table 2), the complete set of results shows that the procedures are sensitive to the increasing of the magnitude of the outlier (AO or IO) but none of the approaches presents better performance than the other. The percentage of correct detection is similar for both types of outliers. When the outlier magnitude is equal to $\lceil 10\sigma_X \rceil$, for the threshold approach the minimum percentage of correct detections is 98.2 % and 99.1 % for AO and IO cases, respectively; while for the parametric resampling approach, the minimum values are 97.8 % for the AO case and 98.8 % for the IO case. The average number of false outlier detection is slightly bigger for the AO cases, where the maximum average number of false outliers detected is 0.794 for the threshold approach and 0.985 for the parametric resampling approach. In the IO cases, the values are 0.184 and 0.379 for the first and second approaches, respectively.

In the case of contamination with 3 outliers, the results presented in Table 3 show that the percentage of correct detections decreases marginally with respect to Table 2. The analysis of the complete set of results for multiple outliers shows that in general for IO case the threshold approach seems preferable since it leads to a higher

Table 3 Percentage of correct detections and average number of false outliers detected, in 1000 repetitions of PoINAR(1) models with sample sizes $N+1$ for some parameter values, contaminated with 3 additive outlier or 3 innovational outlier, with magnitude $\lceil 5\sigma_X \rceil$ and $\lceil 10\sigma_X \rceil$.

			3 Additive Ouliers				3 Innovational Ouliers			
			% Correct		Average False		% Correct		Average False	
(α,λ)	N	ω	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.
(0.5, 1)	128	$\lceil 5\sigma_X \rceil = 8$	80.9	30.3	0.021	0.041	83.1	73.7	0.082	0.05
		$\lceil 10\sigma \rceil = 15$	100	99.9	0.009	0.002	99.8	99.9	0.039	0.02
	256	$\lceil 5\sigma_X \rceil = 8$	79.9	77.4	0.032	0.052	77.8	78.6	0.124	0.146
		$\lceil 10\sigma_X \rceil = 15$	66.1	100.0	0.078	0.011	99.9	99.9	0.09	0.096
	512	$\lceil 5\sigma_X \rceil = 8$	88.2	79.5	0.085	0.158	69.9	66.1	0.198	0.383
		$\lceil 10\sigma_X \rceil = 15$	99.9	99.9	0.031	0.055	99.6	100.0	0.171	0.282
(0.8, 3)	128	$\lceil 5\sigma_X \rceil = 20$	99.6	89.6	0.011	0.026	97.2	97.3	0.021	0.01
		$\lceil 10\sigma_X \rceil = 39$	100	100	0.011	0.006	100	100	0.011	0.003
	256	$\lceil 5\sigma_X \rceil = 20$	90.0	90.5	0.165	0.199	99.3	98.6	0.05	0.056
		$\lceil 10\sigma_X \rceil = 39$	100	100	0.087	0.077	100	100	0.028	0.022
	512	$\lceil 5\sigma_X \rceil = 20$	91.4	94.3	0.384	0.576	97.7	97.0	0.062	0.128
		$\lceil 10\sigma_X \rceil = 39$	100	100	0.833	0	100	100	0.063	0.083
(0.1, 5)	128	$\lceil 5\sigma_X \rceil = 12$	57.5	44.0	0.026	0.013	54.7	47.7	0.042	0.026
		$\lceil 10\sigma_X \rceil = 24$	99.6	99.5	0.034	0.021	99.8	99.8	0.027	0.012
	256	$\lceil 5\sigma_X \rceil = 12$	54.2	50.7	0.039	0.043	51.2	51.9	0.057	0.054
		$\lceil 10\sigma_X \rceil = 24$	99.9	99.9	0.027	0.025	99.9	99.8	0.058	0.062
	512	$\lceil 5\sigma_X \rceil = 12$	39.9	57.3	0.07	0.114	43.6	29.5	0.062	0.129
		$\lceil 10\sigma_X \rceil = 24$	99.8	99.9	0.028	0.098	99.9	99.8	0.057	0.112

percentage of correct detections while the mean number of false detections is comparable to the parametric approach. On the other hand, for AO case the parametric approach leads to a higher percentage of correct detections but also to an increase of 70% in the mean number of false detections.

Finally, to examine the performance of the procedures to detect patches of outliers, Table 4 presents the percentages of correct (complete) detections and partial detections and the average number of false patches detected, in 1000 repetitions. As before, in each repetition, the Pearson residuals series are obtained from a realization of a PoINAR(1) model, for several samples sizes and combinations of parameter values. In each realization, a patch with 3 additive outliers, with magnitude equal to $\lceil 10\sigma_X \rceil$, is placed randomly. The threshold approach has been applied with the 90th percentiles of the empirical distribution of the maximum of the absolute value of $c\mathbf{D}_1$ and $c\mathbf{D}_2$, respectively $k_1^{0.1}$ and $k_2^{0.1}$ (see Table 1). For each level of decomposition, in the parametric resampling approach, the acceptance envelopes are constructed from the 0.01th and 99.99th percentiles of the empirical distribution of $c\mathbf{D}_1$ and $c\mathbf{D}_2$, respectively. The results indicates that the threshold approach presents a better performance. However, the percentage of the partial detection obtained in the parametric resampling approach indicates that the results can be improved by tuning the acceptance envelope of the second level of decomposition of DWT.

Table 4 Percentage of correct and partial detections and average number of false outliers detected, in 1000 repetitions of PoINAR(1) models with sample sizes $N+1$ for some parameter values, with a patch of 3 additive outliers, with magnitude $\lceil 10\sigma_X \rceil$.

			% Correct		% Partial		Average False	
(α,λ)	ω	N	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.
(0.8, 1)	$\lceil 10\sigma_X \rceil = 23$	128	100	63.9	0	34.7	0.001	1
		256	100	71.3	0	0	0	0.779
		512	100	99.9	0	0.1	0.001	0.986
(0.1, 3)	$\lceil 10\sigma_X \rceil = 19$	128	69.1	60.8	0.1	38.5	0.077	1
		256	98.8	100	0	0	0.053	0.313
		512	99.5	99.9	0	0.1	0.02	0.616
(0.5, 5)	$\lceil 10\sigma_X \rceil = 32$	128	100.0	99.9	0	0	0.023	0.999
		256	100.0	99.7	0	0	0.011	0.012
		512	99.8	100	0	0	0.01	0.167

Note that, since the outliers (single, multiple or patch) are placed randomly, if they appear in the first observation, both approaches have a poor performance. The same happens when two outliers are placed in subsequent observations, since it can be considered as a patch.

As a final illustration of the described procedures, consider the real dataset with 241[8] observations concerning the number of different IP addresses (in periods of 2 minutes length) at the server of the Department of Statistics of the University of Würzburg on November 29th, 2005, between 10 a.m. and 6 p.m., represented in Figure 1 and studied by Silva and Pereira [15] and Weiß[19]. The values of sample mean ($\bar{x} = 1.32$) and sample variance ($\hat{\sigma}^2 = 1.39$) and the analysis of the sample autocorrelation and partial autocorrelation functions, indicate that a PoINAR(1) model can be fitted to this dataset. By applying both approaches to outlier occurrence time detection to this dataset, an outlier is detected at $t = 224$ (corresponding to $S = \{112\}$). Figure 2 represents the threshold and the acceptance envelope for this illustration. The detection of the outlier at $t = 224$ agrees with the results in Weiß[19] and Silva and Pereira [15]. The former reference indicates as true value $X_{224} = 1$ while in the latter reference the authors use a Bayesian approach that detects an outliers at $t = 224$ with probability 0.99 and estimates $\hat{\alpha} = 0.27$, $\hat{\lambda} = 0.89$ and $\omega = 7$.

6 Final remarks

Parametric wavelet-based methods for the detection of outlier occurrences are described. The procedures use the Haar DWT of the Pearson residuals of the

[8] Since 241 is not a power of two, by default Matlab extends the signal by using symmetric-padding (symmetric boundary value replication).

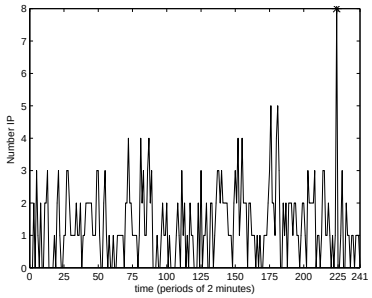

Fig. 1 Cronogram of the IP dataset.

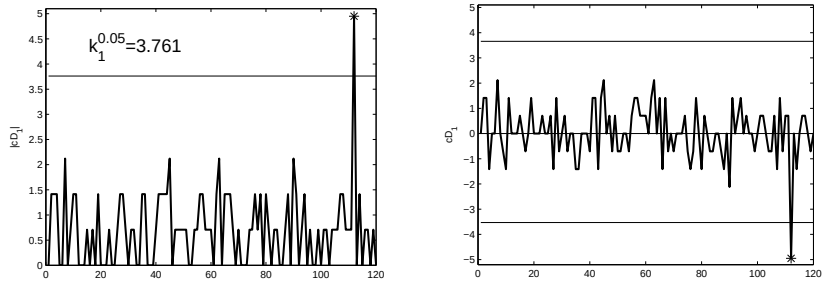

Fig. 2 Results of threshold approach (left panel) and parametric resampling approach (right panel) on the IP dataset.

PoINAR(1) model. In a first approach, a threshold based on the empirical distribution of the maximum of the (first and second levels) detail coefficients is used. In a second approach, an acceptance envelope constructed from the empirical distribution of these detail coefficients is obtained through parametric resampling methods. The procedures do not require previous knowledge on the number of outliers and are adequate to detect one or multiple outliers, of different types, additive or innovational and patches of additive outliers. However, an open issue is the discrimination of the two types of outliers.

DWT can only be applied when the sample size of the time series is a power of two. To overcome this limitation, the proposed approaches to outlier detection can use the modified version of DWT, designated by Maximum Overlap DWT (MODWT), introduced by Percival and Walden [12], since MODWT can be applied for a time series of any length.

The performance of the proposed procedures is illustrated with synthetic and real count data. The results show that the methods are efficient and reliable. As far as it is known, this is the first work that treats patches of outliers in the counting time series context. Improvements are still possible by calibrating the percentiles of the empirical distributions used to detect the time of outlier occurrence, either in the threshold approach or in the parametric resampling approach. Different applications may need different significance levels.

The procedures proposed can be applied in other contexts and can also be extended to detect changes in the structure and dynamics of the processes.

Acknowledgements The authors would like to thank the referees for their comments which helped to improve the paper and to Aurea Grané for supplying the programs of the paper [8]. This work is partially supported by Portuguese funds through the CIDMA and the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), within project UID/MAT/04106/2013.

References

1. Al-Osh, M. A., Alzaid, A. A.: First-order integer-valued autoregressive (INAR(1)) process. J. Time Ser. Anal. **8**, 261–275 (1987)
2. Barczy, M., Ispny, M., Pap, G., Scotto, M., Silva, M. E.: Innovational Outliers in INAR(1) Models. Commun. Stat. - Theor. M. **39**, 3343–3362 (2010)
3. Barczy, M., Ispny, M., Pap, G., Scotto, M., Silva, M. E.: Additive outliers in INAR(1) models. Stat. Pap. **53**, 935–949 (2011)
4. Bilen, C., Huzurbazar, S.: Wavelet-Based Detection of Outliers in Time Series. J. Comp. Graph. Stat. **11**, 311–327 (2002)
5. Chang, I., Tiao, G. C.,Chen, C.: Estimation of time series parameters in the presence of outliers. Technometrics **30**, 193–204 (1988)
6. Chen, C., Liu, L. M.: Joint estimation of model parameters and outlier effects in time series. J. Am. Stat. Assoc. **88**, 284–297 (1993)
7. Fox, A. J.: Outliers in Time Series. J. R. Stat. Soc. Ser. B Stat. Methodol. **34**, 350–363 (1972)
8. Grané, A., Veiga, H.: Wavelet-based detection of outliers in financial time series. Comput. Stat. Data An. **54**, 2580–2593 (2010)
9. Mallat, S. G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989)
10. MATLAB Release 2012a. The MathWorks, Inc., Natick, Massachusetts, United States.
11. McKenzie, E.: Some simple models for discrete variate time series. Water Resources Bull. **21**, 645–650 (1985)
12. Percival, D., Walden, A.: *Wavelet methods for time series analysis*, Cambridge Series in Statistical and Probabilistic Mathematics, New York, Cambridge University Press (2006)
13. Scotto, M. G., Weiß, C. H., Gouveia, S.: Thinning-based models in the analysis of integer-valued time series: a review. Stat. Modelling **15**, 590–618 (2015)
14. Silva, M. E., Oliveira, V. L.: Difference equations for the higher-order moments and cumulants of the INAR(1) model. J. Time Ser. Anal. **25**, 317–333 (2004)
15. Silva, M. E., Pereira, I.: Detection of additive outliers in Poisson INAR(1) time series. In Bourguignon, J. P. *et al.* (Eds.) *CIM Series in Mathematical Sciences - Mathematics of Energy and Climate Change*, Springer: 377–388 (2015)
16. Steutel, F. W., Van Harn, K.: Discrete analogues of self-decomposability and stability. Ann. Probab. **7**, 893–899 (1979)
17. Tsay, R. S.: Time series model specification in the presence of outliers. J. Am. Stat. Assoc. **81**, 132–141 (1986)
18. Tsay, R. S.: Model checking via parametric bootstraps in time series analysis. J. R. Stat. Soc. Ser. C Appl. Stat. **41**, 1–15 (1992)
19. Weiß, C. H.: Controlling correlated processes of Poisson counts. Qual. Reliab. Eng. Int. **23**, 741–754 (2007)