A BAYESIAN INAR(1) MODEL WITH ADAPTIVE OVERDISPERSION

HELTON GRAZIADEI, HEDIBERT F. LOPES AND PAULO C. MARQUES F.

Abstract. We propose a first order integer-valued autoregressive (INAR) model in which a mixture of Geometric and Poisson distributions is introduced as a means to learn the appropriate level of overdispersion of the time series as well as capturing inflations of low-counts in the time series. A data-augmentation scheme allows to obtain the approximated posterior distribution of the model parameters. The forecasting performance of the Poisson-INAR(1) and the Adaptive-INAR(1) is compared for a data set of crimes in Pittsburgh.

1. Introduction

Low-count time series arise in a wide range of applications such as epidemiology, econometrics, environmental studies and public policy. The development of such models has attracted significant attention over the past decades, mainly motivated by the seminal integer-valued autoregressive (INAR) model, introduced by [McKenzie, 1985] and [Al-Osh and Alzaid, 1988]. This model has a simple interpretation and can be applied in any time series that have a "birth-and-death" structure. Particularly, it has various advantages over the continuous autoregressive (AR) models, especially in "low" count situations where approximations to Gaussian models may be imprecise.

The INAR(1) structure contains two random components: the number of survivors (or maturations) at the immediate previous time and the current number of immigrations (or innovations). The classical Poisson-INAR(1) model assumes the Binomial thinning operator for the maturations and Poisson distribution for the innovations of the process. However, the latter assumption may not be suitable especially when the series have a large number of zeros or extreme observations, which lead to overdispersion [Mullahy, 1997, Maiti et al., 2015]. In order to overcome the limitations of the Poisson-INAR(1) model, we propose an adaptive model by using a mixture of Poisson and Geometric distributions on the innovations to capture adaptively overdispersion in the series. Also, we introduce a set of latent variables to obtain the posterior distribution of the parameters by means of an efficient data-augmentation scheme. We apply the proposed model in a data set of burglary counts in Pittsburgh from 2007 to 2010.

2. The Adaptive-INAR(1) Model

To accommodate a wider range of innovations distributions, we propose a generalization of the usual Poisson-INAR(1) model [Al-Osh and Alzaid, 1988] and the Geometric-INAR(1) model [Aghababaei Jazi et al., 2012] such that the innovations follow a mixture of Geometric and Poisson distributions. This flexible model allows to learn the appropriate level of overdispersion from the data in a fully Bayesian fashion as well as inflating the probability of extremely small values.

Definition. *For a time series $\{Y_t\}_{t\geq 1}$ of counting data, let the innovations $\{Z_t\}_{t\geq 2}$ be a sequence of independent random variables, which are also independent of a collection $\{B_i(t) : i \geq 0, t \geq 1\}$ of independent Bernoulli(α) random variables. The* Adaptive-INAR(1) *is defined by the functional relation*

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

for $t \geq 2$, in which the binomial thinning operator is defined by $\alpha \circ Y_{t-1} = \sum_{i=0}^{Y_{t-1}} B_i(t)$. Also, Z_t is a mixture of a Geometric and a Poisson distributions such that $p(z_t \mid \theta, \lambda, w) =$ w Geometric$(\theta) + (1 - w)$ Poisson(λ), $w \in [0, 1]$. Notice that if $w = 1$, we have the Geometric-INAR(1) model, whereas $w = 0$ implies the Poisson-INAR(1).

For simplicity, assume that there exists $y_1 \in \mathbb{N}$ such that $P(Y_1 = y_1 \mid \alpha, \theta, \lambda) = 1$. Also, notice that as w becomes large, the innovation is contaminated by the Geometric distribution in the mixture, which increases the variability of the process and the total number of low-counts. As an illustration, Figure 1 shows simulated time series for $w = 0.1$ and $w = 0.9$, while the remaining parameters were fixed as $(\alpha_0, \theta_0, \lambda_0) = (0.10, 0.15, 5.66)$. It is clear that the latter time series is considerably more dispersed than the former. Furthermore, since the process $\{Y_t\}_{t\in\mathbb{N}}$ is Markovian, the joint distribution of $(Y_1, \ldots, Y_T)$, given α and λ, can be decomposed as its conditional distributions, i.e.,

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T \mid \alpha, \lambda) = \prod_{t=2}^{T} P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha, \theta, \lambda, w).$$

By applying the functional relation (1), the law of total probability and the independence of $\{B_i(t)\}_{i\geq 0, t\geq 1}$ and $\{Z_t\}_{t=2}^{T}$, we obtain that

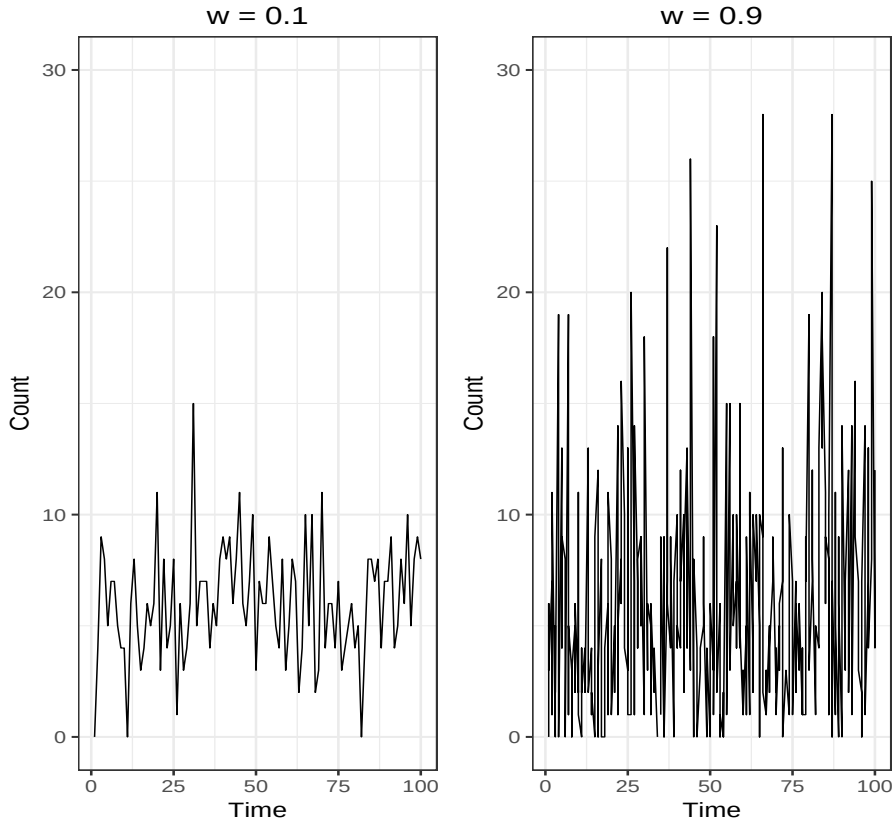

FIGURE 1. Typical simulated series for $w = 0.1$ ($\overline{y} = 6.02$ and $s.d. = 2.57$) and $w = 0.9$ ($\overline{y} = 6.08$ and $s.d. = 5.83$).

$$
\begin{aligned}
P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha,\ \theta,\ \lambda,\ w) &= P(\alpha \circ Y_{t-1} + Z_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha,\ \theta,\ \lambda,\ w) \\
&= P(\sum_{i=1}^{Y_{t-1}} B_i(t) + Z_t = y_t \mid Y_{t-1} = y_{t-1},\ \alpha,\ \theta,\ \lambda,\ w) \\
&= \sum_{m_t=0}^{\min\{y_t, y_{t-1}\}} P(\sum_{i=1}^{y_{t-1}} B_i(t) = m_t, Z_t = y_t - m_t \mid \alpha,\ \theta,\ \lambda,\ w) \\
&= \sum_{m_t=0}^{\min\{y_t, y_{t-1}\}} P(\sum_{i=1}^{y_{t-1}} B_i(t) = m_t \mid \alpha) P(Z_t = y_t - m_t \mid \theta,\ \lambda,\ w).
\end{aligned}
$$

Consequently, the likelihood function of the Adaptive-INAR(1) model is given by

$$
L_y(\alpha, \theta, \lambda, w) = \prod_{t=1}^{T} \sum_{m_t=0}^{\min\{y_{t-1}, y_t\}} \binom{y_{t-1}}{m_t} \alpha^{m_t} (1-\alpha)^{y_{t-1}-m_t} \left[w\ \theta(1-\theta)^{y_t - m_t} + (1-w)\ \frac{e^{-\lambda} \lambda^{y_t - m_t}}{(y_t - m_t)!} \right]. \quad (2)
$$

Furthermore, suppose the following independent prior distributions:

$$w \sim Beta(a_0^{(w)}, b_0^{(w)}),\ \alpha \sim Beta(a_0^{(\alpha)}, b_0^{(\alpha)}),\ \theta \sim Beta(a_0^{(\theta)}, b_0^{(\theta)}),\ \lambda \sim Gamma(a_0^{(\lambda)}, b_0^{(\lambda)}).$$

By using the Bayes' Theorem, we have the joint posterior distribution:

$$\pi(\alpha, \theta, \lambda, w \mid y) \propto L_y(\alpha, \theta, \lambda, w)\ \pi(\alpha)\ \pi(\lambda)\ \pi(\theta)\ \pi(w),$$

which does not seem straightforward to handle computationally. Therefore, we propose data-augmentation scheme [Tanner and Wong, 1987] which allows to obtain samples from the joint posterior in a direct manner.

3. Data-augmentation and full conditionals

Data-augmentation stands for constructing efficient sampling methods by introducing latent variables in the algorithm [Tanner and Wong, 1987, Van Dyk and Meng, 2001]. Our proposed data augmentation scheme treats the vector of maturations $m = (m_1, \ldots, m_T)$ and the indicator variables of the mixture components $u = (u_2, \ldots, u_T)$ as latent variables. Specifically, each u_t is defined as $u_t = 1$, if $z_t \mid \theta \sim Geometric(\theta)$ or $u_t = 0$, if $z_t \mid \lambda \sim Poisson(\lambda)$. Postulate that, for $t = 2, \ldots, T$

$$\begin{aligned} p(y_t \mid m_t, u_t = 1) &= \theta(1-\theta)^{y_t - m_t} \mathbb{I}_{\{m_t, m_{t+1}, \ldots\}}(y_t), \\ p(y_t \mid m_t, u_t = 0) &= \frac{e^{-\lambda} \lambda^{y_t - m_t}}{(y_t - m_t)!} \mathbb{I}_{\{m_t, m_{t+1}, \ldots\}}(y_t), \\ p(m_t \mid y_{t-1}, \alpha) &= \binom{m_t}{y_{t-1}} \alpha^{y_{t-1}} (1-\alpha)^{m_t - y_{t-1}}. \end{aligned} \tag{3}$$

Also, the model structure implies that:

(1) M_t depends on Y_{t-1} only through α,
(2) Given $M_t = m_t$ and $U_t = u_t$, Y_t only depends on the innovation parameter.
(3) Each indicator variable u_t, $t = 2, \ldots, T$, is independent of m_t, y_{t-1} , given w.
(4) Since $M_t \leq Y_{t-1}$ and $M_t \leq Y_t$ we have $M_t \leq \min\{y_t, y_{t-1}\}$.

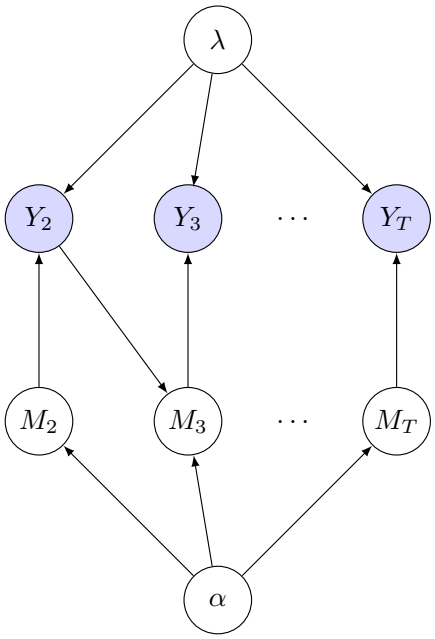

FIGURE 2. Graphical representation of the data-augmented Poisson-INAR(1) Model ($w = 0$).

The graph in Figure 2 illustrates the proposed data-augmentation for the Poisson-INAR(1) model by a plate representation where the shaded variables are observable while open nodes are latent variables and parameters [Jordan et al., 2004].

Hence, the likelihood function of the Adaptive-INAR(1) model can also be written as:

$$\begin{aligned} L_y(\alpha, \theta, \lambda, w) &= \prod_{t=2}^{T} p(y_t \mid y_{t-1}, \alpha, \theta, \lambda, w) \\ &= \prod_{t=2}^{T} \sum_{m_t=0}^{\min\{y_t, y_{t-1}\}} \sum_{u_t=0}^{1} p(y_t, m_t, u_t \mid y_{t-1}, \alpha, \theta, \lambda, w) \\ &= \prod_{t=2}^{T} \sum_{m_t=0}^{\min\{y_t, y_{t-1}\}} \sum_{u_t=0}^{1} p(y_t \mid m_t, u_t, \theta_{u_t})\ p(m_t \mid y_{t-1}, \alpha)\ p(u_t \mid w). \end{aligned}$$

Notice that in the case the conditional distributions $p(y_t \mid m_t, u_t, \theta)$, and $p(m_t \mid y_{t-1}, \alpha)$ are the postulated probability functions given in 3, we have the same likelihood function from 2. In addition, from the prior independence assumption, it follows that

$$p(y, m, u, \alpha, \theta, \lambda, w) = p(y, m, u \mid \alpha, \theta, \lambda, w)\ \pi(\alpha)\ \pi(\theta)\ \pi(\lambda)\ \pi(w),$$

where $y = (y_2, \ldots, y_T)$, $m = (m_2, \ldots, m_T)$ and $u = (u_2, \ldots, u_T)$. It is easy to see that

$$p(y, m, u \mid \alpha, \theta, \lambda, w) = \prod_{t=2}^{T} p(y_t \mid m_t, u_t, \theta_{u_t})\ p(m_t \mid y_{t-1}, \alpha)\ p(u_t \mid w).$$

Hence,

$$p(y, m, u, \alpha, \theta, \lambda, w) = \Big[\prod_{t=2}^{T} p(y_t \mid m_t, u_t, \theta_{u_t})\ p(m_t \mid y_{t-1}, \alpha)\ p(u_t \mid w)\Big]\pi(\alpha)\ p(\theta)\ p(\lambda)\ p(w).$$

Therefore, the full conditional distributions of α, θ, λ, w and u are given by:

$$\begin{aligned} p(\alpha \mid \ldots) &\propto p(y, m, u, \alpha, \theta, \lambda, w) \propto \Big[\prod_{t=2}^{T} p(m_t \mid y_{t-1}, \alpha)\Big]\ \pi(\alpha) \\ &\propto \alpha^{a_0^{(\alpha)} + \sum_{t=2}^{T} m_t - 1}(1-\alpha)^{b_0^{(\alpha)} + \sum_{t=2}^{T}(y_{t-1} - m_t)} \\ \alpha \mid \ldots &\sim Beta\left(a_0^{(\alpha)} + \sum_{t=2}^{T} m_t,\ b_0^{(\alpha)} + \sum_{t=2}^{T}(y_{t-1} - m_t)\right), \end{aligned} \tag{4}$$

$$\begin{aligned} p(\theta \mid \ldots) &\propto p(y, m, u, \alpha, \theta, \lambda, w) \propto \Big[\prod_{\{t:u_t=1\}} p(y_t \mid m_t, u_t, \theta)\Big]\ \pi(\theta) \\ &\propto \theta^{a_0^{(\theta)} + \sum_{t=2}^{T} u_t - 1}(1-\theta)^{b_0^{(\theta)} + \sum_{t=2}^{T}(y_t - m_t)\mathbb{I}_{\{1\}}(u_t)}. \\ \theta \mid \ldots &\sim Beta\left(a_0^{(\theta)} + \sum_{t=2}^{T} u_t,\ b_0^{(\theta)} + \sum_{t=2}^{T}(y_t - m_t)\mathbb{I}_{\{1\}}(u_t)\right), \end{aligned} \tag{5}$$

$$\begin{aligned} p(\lambda \mid \ldots) &\propto p(y, m, u, \alpha, \theta, \lambda, w) \propto \Big[\prod_{\{t:u_t=0\}}^{T} p(y_t \mid m_t, u_t, \lambda)\Big]\ \pi(\lambda) \\ &\propto \lambda^{a_0^{(\lambda)} + \sum_{t=2}^{T}(y_t - m_t)\mathbb{I}_{\{0\}}(u_t)} e^{b_0^{(\lambda)} + T - \sum_{t=2}^{T} u_t - 1} \\ \lambda \mid \ldots &\sim Gamma\left(a_0^{(\lambda)} + \sum_{t=2}^{T}(y_t - m_t)\mathbb{I}_{\{0\}}(u_t),\ b_0^{(\lambda)} + T - \sum_{t=2}^{T} u_t - 1\right), \end{aligned} \tag{6}$$

$$p(w \mid \ldots) \propto p(y, m, u, \alpha, \theta, \lambda, w) \propto \Big[\prod_{t=2}^{T} p(u_t \mid w)\Big]\ \pi(w)$$
$$\propto w^{a_0^{(w)}+\sum_{t=2}^{T} u_t-1}(1-w)^{b_0^{(w)}+T-\sum_{t=2}^{T} u_t-1}$$
$$w \mid \ \ldots \ \sim \ Beta\left(a_0^{(w)}+\sum_{t=2}^{T} u_t,\ b_0^{(w)}+T-\sum_{t=2}^{T} u_t-1\right), \tag{7}$$

$$p^* = p(u_t = 1 \mid \ldots) \propto p(u_t = 1 \mid w)\ p(y_t \mid u_t = 1, m_t, \theta) \propto w\ \theta(1-\theta)^{y_t-m_t}$$
$$p(u_t = 0 \mid \ldots) \propto p(u_t = 0 \mid w)\ p(y_t \mid u_t = 0, m_t, \lambda) \propto (1-w)\ \frac{e^{-\lambda}\lambda^{y_t-m_t}}{(y_t-m_t)!}.$$
$$u_t \sim \text{Bernoulli}(p^*). \tag{8}$$

$$p(m_t|\ldots) \propto p(m_t \mid y_t, \alpha) \times p(y_t \mid m_t, \theta)$$
$$\propto \frac{1}{(y_{t-1}-m_t)!\ m_t!}\left(\frac{\alpha}{(1-\theta)(1-\alpha)}\right)^{m_t} \mathbb{I}_{\{0,1,\ldots,min(y_t,y_{t-1})\}}(m_t),\ \text{if } u_t = 1. \tag{9}$$
$$p(m_t \mid \ \ldots) \propto p(m_t \mid y_t, \alpha) \times p(y_t \mid m_t, \lambda)$$
$$\propto \frac{1}{(y_t-m_t)!\ (y_{t-1}-m_t)!\ m_t!}\left(\frac{\alpha}{\lambda\ (1-\alpha)}\right)^{m_t} \mathbb{I}_{\{0,1,\ldots,min(y_t,y_{t-1})\}}(m_t),\ \text{if } u_t = 0. \tag{10}$$

The proposed Gibbs sampler cycles by sampling from the conditionals given from (4) to (10) until it reaches convergence.

4. Forecasting

Forecasting is often the main reason for applying time series models. Under the Bayesian approach, a pointwise forecast is made through a statistical functional of the (posterior) predictive distribution, whose calculation is described in this section. Let $(\alpha^{(s)}, \theta^{(s)}, \lambda^{(s)}, w^{(s)})$, $s = 1, \ldots, M$, be the posterior parameter samples from the MCMC output obtained by the Gibbs sampler algorithm proposed in Section 3. To obtain the k-step ahead predictive distribution of the Adaptive-INAR(1) Model, we generate $Y_{t+k}^{(s)}$ recursively applying the posterior samples on the functional relation (1), that is,

$$Y_{T+k}^{(s)} = \alpha^{(s)} \circ Y_{T+k-1}^{(s)} + Z_{T+k}^{(s)},\ s = 1, \ldots, M,$$

where $Z_{T+k}^{(s)} \sim Poisson(\lambda^{(s)})$. Then, we approximate the predictive probabilities $p(i \mid y_1, \ldots, y_T)$ by empirical averages, i.e.,

$$\hat{p}(i \mid y_1, \ldots, y_T) = \frac{1}{M} \sum_{s=1}^{M} \mathbb{I}(y_{T+k}^{(s)} = i),\ i = 0, 1, \ldots.$$

We adopt here a generalized median as the k-step ahead pointwise forecast, denoted as $\hat{y}_{T+k}$, which is defined by:

$$\hat{y}_{T+k} = \arg \max_{y_{T+k} \geq 0} |0.5 - \sum_{r=0}^{y_{T+k}} p(r \mid y_1, \ldots, y_T)|,\ k \geq 1. \tag{11}$$

5. Application

We consider time series of weekly counts of burglary in Pittsburgh from 2007 to 2010 (available in `http://forecastingprinciples.com/index.php/crimedata`). The dataset is grouped by clusters of neighborhoods with a total of 34 time series. Our objective lies in comparing the forecasting performance of the Adaptive-INAR(1) model and the Poisson-INAR(1) model. We use a training set, composed by the first $T = 94$ observations, to fit the model and a test set, with the remaining $h = 50$ observations, to evaluate how well the model is to forecast future data. Let $y_{training} = (y_1, \ldots, y_T)$ be the training set and $y_{test} = (y_{T+1}, \ldots, y_{T+h})$ the test set. First we approximate the predictive distribution associated to the Poisson-INAR(1) and Adaptive-INAR(1) models. Recall that the Mean Absolute Error (MAE) for the k-step ahead forecasts is given by:

$$MAE(k) = \sum_{i=T+k}^{T+h} |y_i - \hat{y}_i|,$$

where $\hat{y}_i$ is the generalized median (11) of the predictive distribution at time i, $i = T + 1, \ldots, T+h$. We present in Table 1 the Mean Absolute Error (MAE) for both the Adaptive-INAR(1) and the Poisson-INAR(1) models calculated to each cluster of neighborhood. According to this criterion, the Adaptive-INAR(1) is the favoured model in most of the clusters. As an example, Figure 3 shows the one-step ahead pointwise forecasts for the Poisson-INAR(1) and the Adaptive-INAR(1) in the 54 th district of Pittsburgh. Notice that the Adaptive-INAR(1) outperforms the Poisson-INAR(1) especially in forecasting low-count observations.

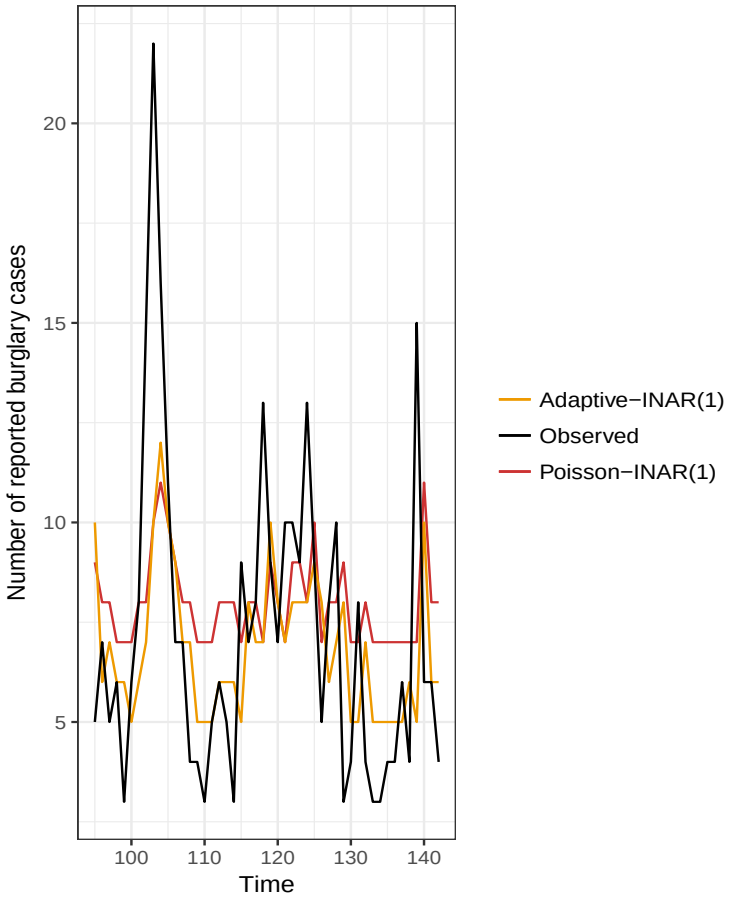

FIGURE 3. One-step ahead forecasts for the burglary time series in the 54th district of Pittsburgh. The black lines represent the observed values in the test set; the yellow line depicts the generalized median of the predictive distribution for the Adaptive-INAR(1) in each time point, and the red line indicates the generalized median according to the Poisson-INAR(1).

6. CONCLUSIONS

The proposed data-augmentation algorithm allows us to develop a simple Gibbs sampler for the Adaptive-INAR(1) model. Also, such model captures inflation of zeros and low-counts in the process which is not the feature of the classical Poisson-INAR(1). Consequently, the forecasting performance of the proposed model may be improved as illustrated in a burglary data set in Pittsburgh.

TABLE 1. Mean Absolute Errors (MAE) of the Poisson-INAR(1) Model and the Adaptive-INAR(1) Model for the one, two and three-step ahead forecasts for the burglary dataset in Pittsburgh (divided by clusters of neighbours)

	Poisson-INAR(1)			Adaptive-INAR(1)		
cluster	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
11	1.21	1.43	1.43	1.23	1.43	1.43
12	3.83	3.87	3.85	3.92	4.40	4.93
13	2.79	3.06	3.35	2.81	2.77	2.89
14	2.25	2.51	2.52	2.23	2.28	2.46
15	2.73	3.11	3.24	3.04	3.21	3.33
16	2.40	2.21	2.22	2.33	2.21	2.22
17	2.33	2.36	2.37	2.31	2.49	2.33
21	1.58	1.60	1.59	1.50	1.23	1.24
22	2.29	2.30	2.33	2.10	2.06	2.11
23	3.21	3.23	3.28	3.15	3.40	3.46
24	2.15	2.32	2.76	1.83	2.11	2.02
25	1.58	1.68	1.63	1.48	1.45	1.41
26	2.75	3.70	3.78	2.46	2.74	3.04
27	1.46	1.62	1.61	1.40	1.40	1.63
28	0.83	0.81	0.80	0.88	0.81	0.80
29	2.56	2.77	2.78	2.48	2.81	2.78
31	3.67	3.43	3.61	3.56	3.53	3.52
32	3.27	3.36	3.37	3.17	3.38	3.41
33	2.19	2.30	2.30	1.96	2.23	2.30
34	3.15	3.34	3.39	3.04	3.38	3.39
35	1.10	1.11	1.13	1.10	1.11	1.13
41	2.33	2.36	2.28	2.33	2.30	2.28
42	3.10	3.11	3.22	3.08	3.26	3.20
43	2.06	2.04	2.09	2.04	2.17	2.09
44	1.67	1.64	1.63	1.63	1.64	1.63
45	2.40	2.43	2.50	2.40	2.47	2.52
46	2.46	2.45	2.46	2.46	2.38	2.48
47	2.23	2.17	2.15	2.23	2.17	2.17
51	2.69	2.53	2.54	2.79	2.57	2.50
52	4.52	4.60	4.74	4.35	4.36	4.57
53	2.77	2.83	2.74	2.90	3.26	3.33
54	2.92	3.38	3.52	2.54	2.89	3.22
55	5.81	5.36	5.37	4.79	4.49	4.76
56	2.56	3.06	3.26	2.15	2.62	2.57
57	1.96	2.04	2.07	2.08	2.06	2.07
58	3.25	3.49	3.74	3.21	3.74	3.85

References

[Aghababaei Jazi et al., 2012] Aghababaei Jazi, M., Jones, G., and Lai, C.-D. (2012). Integer valued ar (1) with geometric innovations. *Journal of the Iranian Statistical Society*, 11(2):173–190.

[Al-Osh and Alzaid, 1988] Al-Osh, M. and Alzaid, A. (1988). First-order integer-valued autoregressive (inar(1)) process: distributional and regression properties. *Statistica Neerlandica*, 42:53–61.

[Jordan et al., 2004] Jordan, M. I. et al. (2004). Graphical models. *Statistical Science*, 19(1):140–155.

[Maiti et al., 2015] Maiti, R., Biswas, A., and Das, S. (2015). Time series of zero-inflated counts and their coherent forecasting. *Journal of Forecasting*, 34(8):694–707.

[McKenzie, 1985] McKenzie, E. (1985). Some simple models for discrete variate time series. *JAWRA Journal of the American Water Resources Association*, 21(4):645–650.

[Mullahy, 1997] Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, pages 337–350.

[Tanner and Wong, 1987] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

[Van Dyk and Meng, 2001] Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.