Stat Methods Appl (2018) 27:407–435
https://doi.org/10.1007/s10260-017-0413-z

ORIGINAL PAPER

Modelling of low count heavy tailed time series data consisting large number of zeros and ones

Raju Maiti[1] · **Atanu Biswas**[2] · **Bibhas Chakraborty**[1]

Accepted: 19 November 2017 / Published online: 1 December 2017

Abstract In this paper, we construct a new mixture of geometric INAR(1) process for modeling over-dispersed count time series data, in particular data consisting of large number of zeros and ones. For some real data sets, the existing INAR(1) processes do not fit well, e.g., the geometric INAR(1) process overestimates the number of zero observations and underestimates the one observations, whereas Poisson INAR(1) process underestimates the zero observations and overestimates the one observations. Furthermore, for heavy tails, the PINAR(1) process performs poorly in the tail part. The existing zero-inflated Poisson INAR(1) and compound Poisson INAR(1) processes have the same kind of limitations. In order to remove this problem of under-fitting at one point and over-fitting at others points, we add some extra probability at one in the geometric INAR(1) process and build a new mixture of geometric INAR(1) process. Surprisingly, for some real data sets, it removes the problem of under and over-fitting over all the observations up to a significant extent. We then study the stationarity and ergodicity of the proposed process. Different methods of parameter estimation, namely the Yule-Walker and the quasi-maximum likelihood estimation procedures are discussed and illustrated using some simulation experiments. Furthermore, we discuss

✉ Raju Maiti
raju.maiti@duke-nus.edu.sg

Atanu Biswas
atanu@isical.ac.in

Bibhas Chakraborty
bibhas.chakraborty@duke-nus.edu.sg

[1] Centre for Quantitative Medicine, Duke-NUS Medical School, 20 College Road, Singapore 169856, Singapore

[2] Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India

the future prediction along with some different forecasting accuracy measures. Two real data sets are analyzed to illustrate the effective use of the proposed model.

Keywords Geometric INAR (1) · Mixture distribution · Strongly stationary · Coherent forecasting · Zero-inflation · Over-dispersion

1 Introduction

Time series of count data arise in various fields of science, especially in social science, medicine and epidemiology. For example, monthly reported cases of a particular water-borne disease, and monthly cases of kidnapping in a city are some examples of count time series data. If the counts are large, data can be well approximated by some continuous distributions and hence the well-known Box-Jenkins' ARMA model can be used. The main justification behind this approximation is that many common discrete distributions, e.g., binomial, Poisson and negative binomial can be well approximated by normal distribution when the means of these distributions are large. However, in practice, it is often observed that the counts are small. For example, for the monthly cases of poliomyelitis data (see Zeger 1988), almost 80% of the total observations lie between 0 and 2. Therefore, in such scenarios, it is not desirable to approximate the data by some continuous time series models. Furthermore, it is very important to use a model that preserves the count property of the data.

In this regard, the most well-known model is the integer-valued auto-regressive process of first order (or INAR(1)) based on binomial thinning operator introduced by McKenzie (1985). This class of models is constructed based on the *binomial thinning operator* of Steutel and Van Harn (1979) defined as follows. Given a discrete random variable X and a constant α lying between 0 and 1, the binomial thinning operator "$\circ$" is defined as $\alpha \circ X = \sum_{i=1}^{X} B_i$, where B_i's are independent and identically distributed (i.i.d.) Bernoulli(α) random variables. Given the above definition of the thinning operator, McKenzie's class of INAR(1) process has the form $Y_t = \alpha \circ Y_{t-1} + \varepsilon_t$, where $\alpha \in (0, 1)$, and $\{\varepsilon_t\}$ is a sequence of i.i.d. discrete random variables. It is also assumed that ε_t is independent of the past lag values of Y_t, i.e, Y_{t-k} for $k \geq 1$. Given the above class of INAR(1) process, it has been shown that all distributions of Y_t that are discrete self-decomposable (DSD) in the sense of Steutel and Van Harn (1979) are stationary solutions of the above equation. For example, Poisson and geometric distributions are stationary solutions of the above INAR(1) process. However, distributions that are defined on a finite support space, e.g., binomial distribution, are not stationary solutions of the above class.

Based on this idea, Al-Osh and Alzaid (1987) introduced the Poisson INAR(1) or PINAR(1) process which was subsequently studied by Freeland and McCabe (2004, 2005), McCabe and Martin (2005), Silva et al. (2009) and many others. This model is widely used in various scientific disciplines because of its nice closed mathematical form. However, when data in practice are under-dispersed (variance is smaller than mean) or over-dispersed (variance is larger than mean), such class of PINAR(1) models does not fit the data well. In such cases, over-dispersed INAR(1) models like geometric INAR(1) (GINAR(1) in short), negative binomial INAR(1) (NBINAR(1) in short)

process proposed by McKenzie (1986), compound Poisson INAR(1) (CPINAR(1) in short) process proposed by Schweer and Weiß (2014) are very useful. However, this over-dispersed class of INAR(1) models also fails when data contains a large number of zeros. For example, the skin lesions data used by Jazi et al. (2012) contains a large number zeros. In such cases, the PINAR(1), GINAR(1) and other over-dispersed models do not fit the data well. As an alternative, Jazi et al. (2012) proposed a class of zero-inflated Poisson INAR(1) (ZINAR(1) in short) models whose innovation distribution is zero-inflated Poisson. Such models can also be used for zero-deflated data. However, the marginal distribution of their model is very complicated and does not have any closed form expression. Recently, Maiti et al. (2015) proposed an another class of zero-inflated Poisson INAR(1) models based on binomial thinning operator for which the marginal distribution of Y_t is zero-inflated Poisson. Such models perform better than the usual over-dispersed models like GINAR(1) in capturing the large number of zeros.

On the other hand, Latour (1998) extended the binomial thinning operator to generalized thinning operator which is defined as follows. Given a discrete random variable X and a constant α lying between 0 and 1, the generalized thinning operator "$\bullet$" is defined as $\alpha \bullet X = \sum_{i=1}^{X} B_i$, where B_i's are i.i.d. non-negative random variables with mean α. Using this operator, Ristić et al. (2009) proposed a new geometric INAR(1) (or NGINAR(1)) process assuming B_i's follow i.i.d. geometric distribution with mean α and they named it *negative binomial thinning operator*. Several other thinning operators and consequent INAR processes can be found in Weiß (2008) and Scotto et al. (2015). In order to accommodate a large number of zeros, Barreto-Souza (2015) proposed a zero-modified geometric INAR(1) or ZMGINAR(1) process based on the negative binomial thinning operator. They showed that the marginal distribution of such models follows a zero-inflated geometric distribution. In addition, such models can be used for both zero-inflation and zero-deflation. However, when a count time series data contains a large number of zeros along with a large number of ones which often arise in practice [e.g., poliomyelitis data used by Zeger (1988)], both the over-dispersed [e.g., GINAR(1)] and zero-inflated Poisson INAR(1) models fail. In such scenarios, it demands some adjustment of probability mass on zero and one observations.

In order to fill this gap, here we propose a new class of one-modified geometric INAR(1) (or OMGINAR(1)) process, extending the idea articulated in Maiti et al. (2015). The applications studied in this article demand a theoretical study of the newly proposed process. We study some structural properties of the proposed model such as mean, variance, dispersion index, autocorrelation function, marginal and joint distributions. We show that the proposed model is strongly stationary and ergodic. We study the parameter estimation using Yule-Walker (YW) and quasi-maximum likelihood estimation (QMLE) methods. We also provide a mathematical proof of consistency of the YW estimators. Furthermore, the robustness of the proposed model is studied in great details with respect to various forecasting measures of accuracy using simulated data from various INAR(1) models like PINAR(1), GINAR(1) and ZINAR(1). Finally, we illustrate the proposed model using two real data sets, namely the monthly cases of poliomyelitis in the US and monthly cases of assault data reported in Pittsburgh, US.

We present the article as follows. In Sect. 2, we describe the proposed model along with its different distributional properties like marginal distribution and autocorrelation. Joint and conditional distributions along with its h-step ahead forecasting distribution are discussed in Sect. 3. In Sect. 4, we discuss two estimation methods, namely the YW and the QMLE to estimate the model parameters. Simulation experiments are presented in Sect. 5. In Sect. 6, we illustrate the methodology using two real data sets. Sect. 7 concludes with some discussions. All the proofs and derivations are relegated to "Appendix".

2 The model

Let $\{Y_t\}_{t\in\mathbb{N}}$ be a PINAR(1) process of Al-Osh and Alzaid (1987) based on *binomial thinning operator* and can be written as $Y_t = \alpha \circ Y_{t-1} + \varepsilon_t, \quad t = 0, 1, \ldots$, where Y_t has the Poisson marginal distribution with mean λ. Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $P(X_t = 1) = p = 1 - P(X_t = 0)$. Then the ZIPINAR(1) process of Maiti et al. (2015), $\{Z_t\}_{t\in\mathbb{N}}$, based on the idea of allocating extra weight at 0 in PINAR(1) process, can be written as $Z_t = X_t Y_t$. Using the similar idea, here we propose a new OMGINAR(1) process. We allocate an extra weight at 1 in the GINAR(1) process of McKenzie (1986). Our proposed process can be defined as follows.

Let $\{Y_t\}_{t\in\mathbb{N}}$ be a GINAR(1) process as defined in McKenzie (1986) where Y_t has the geometric marginal distribution in the form $P(Y_t = i) = (1-\theta)\theta^i, \ i = 0, 1, \ldots;$ and let $\{X_t\}$ be a sequence of i.i.d. Bernoulli random variables defined above. Then, the proposed process OMGINAR(1) can be written as follows

$$Z_t = \begin{cases} Y_t & \text{with probability } p \\ 1 & \text{with probability } 1 - p. \end{cases} \tag{1}$$

Assuming that $Y_t^0 = 1$ when $Y_t = 0$, we can write the above process (1) as

$$Z_t = Y_t^{X_t}. \tag{2}$$

Unlike ZMGINAR(1) process of Barreto-Souza (2015) which is based on *negative binomial thinning operator* defined by Ristić et al. (2009), our proposed process is based on *binomial thinning operator* of Steutel and Van Harn (1979). Under the above setup, we can obtain the following result.

Proposition 1 *The marginal distribution of* $\{Z_t\}$ *can be written as*

$$P(Z_t = i) = \begin{cases} 1 - p + p\theta(1-\theta), & \text{for} \quad i = 1 \\ p(1-\theta)\theta^i, & \text{for} \quad i = 0, 2, 3, \ldots. \end{cases} \tag{3}$$

Proof Proof is given in Appendix A. □

Corollary 1 *Using the above result, the marginal mean and marginal variance of* $\{Z_t\}$ *can be obtained as* $E(Z_t) = 1 - p + p\theta^*$ *and* $Var(Z_t) = 1 - p + p\theta^*(1 + 2\theta^*) - (1 - p + p\theta^*)^2$*, respectively. Hence, the dispersion-index (DI) can be computed as*

$$DI = \frac{Var(Z_t)}{E(Z_t)} = p(1 - \theta^*) + \frac{2p\theta^*}{1 - p + p\theta^*}$$

where $\theta^* = \dfrac{\theta}{1 - \theta}$.

Unlike the GINAR(1) process (for which $DI = \dfrac{1}{1 - \theta} > 1$), here DI can take any value between 0 and ∞ depending on the values of θ and p. Therefore, the proposed process can be used for both under- and over-dispersed time series data. However, in this article, we use the process for over-dispersed time series data.

3 Joint and conditional distributional properties

3.1 Auto-correlation structure and weak stationarity

The auto-covariance function (ACVF) of the process can be routinely derived, and is given by

$$\gamma_z(h) = \begin{cases} p^2\theta^*(1 + \theta^*)\alpha^h, & \text{if} \quad h = 1, 2, \ldots \\ 1 - p + p\theta^*(1 + 2\theta^*) - (1 - p + p\theta^*)^2 & \text{if} \quad h = 0, \end{cases} \tag{4}$$

where $\gamma_z(h) = Cov(Z_{t+h}, Z_t)$. This implies that the auto-correlation function of the process decays exponentially to 0 as $h \to \infty$. This phenomena can be used to characterize the process.

From equations (3) and (4), we can see that the marginal mean of Z_t and the auto-covariance function between Z_t and Z_{t+h} do not depend on the time index t. Hence, the proposed process OMGINAR(1) is at least covariance (weakly) stationary.

3.2 Strong stationarity and ergodicity

Under the above setup, it can be shown that the proposed OMGINAR(1) process is strongly stationary. For proof, see Appendix B.

Proposition 2 *The joint distribution of* Z_{t+h} *and* Z_t *for the proposed process can be derived as*

$$P_{Z_{t+h},Z_t}(i,j) = \begin{cases} (1-p)^2 + 2p\bar{p}\theta\bar{\theta} + p^2\theta\bar{\theta}P_{Y_{t+h}|Y_t}(i|j), & \text{if} \quad i=j=1 \\ p\theta^j\bar{\theta}\left\{\bar{p} + pP_{Y_{t+h}|Y_t}(i|j)\right\} & \text{if} \quad i=1, j\neq 1 \\ p\bar{p}\bar{\theta}\theta^i + p^2\theta\bar{\theta}P_{Y_{t+h}|Y_t}(i|j) & \text{if} \quad i\neq 1, j=1 \\ p^2\bar{\theta}\theta^j P_{Y_{t+h}|Y_t}(i|j) & \text{if} \quad i,j\neq 1 \end{cases} \tag{5}$$

where

$$P_{Y_{t+h}|Y_t}(i|j) = \begin{cases} (1-\alpha^h)(1-\theta)\theta^{i-j}\sum_{k=0}^{i}\binom{j}{k}\alpha^{hk}\{(1-\alpha^h)\theta\}^{j-k} & \\ \quad +\binom{j}{i}\alpha^{h(i+1)}(1-\alpha^h)^{j-i}, & i=0,1,\ldots,j \\ (1-\alpha^h)(1-\theta)\theta^{i-j}\{\alpha^h+(1-\alpha^h)\theta\}^j, & i=j+1,j+2,\ldots. \end{cases} \tag{6}$$

Proof Derivation of the above result is given in Appendix C. □

The joint probability generating function (pgf) of Z_{t+1} and Z_t can be derived as

$$\Phi_{Z_{t+1},Z_t}(u,v) = (1-p)^2uv + p(1-p)\left(\frac{1-\theta}{1-\theta u}\right) + p(1-p)\left(\frac{1-\theta}{1-\theta v}\right) + p^2\frac{\lambda(\lambda+\alpha u)}{(\lambda+u)(\lambda+\alpha u+v-\alpha uv)}, \tag{7}$$

which is not symmetric in u and v. Hence the process is not time-reversible. This is also because the hidden process $\{Y_t\}$ is not time-reversible.

Using the joint distribution result of Z_{t+h} and Z_t in (5), we can prove that the proposed OMGINAR(1) process is ergodic. See Appendix D for the proof.

3.3 Conditional distribution

Even though the latent process $\{Y_t\}$ is a Markov Chain of order one, i.e., given $(Y_t,\ldots,Y_1)$, Y_{t+1} depends only on the most present observation Y_t, the observed process $\{Z_t\}$ may not be a Markov chain of order one. In fact, the order of the process $\{Z_t\}$ cannot be assured. For example, suppose $Z_t \neq 1$, then the conditional distribution of Z_{t+1} given $(Z_t, Z_{t-1},\ldots,Z_1)$ is equivalent to the conditional distribution of Z_{t+1} given Z_t. However, if $Z_t = 1$ and $Z_{t-1} \neq 1$, then the conditional distribution of Z_{t+1} given $(Z_t, Z_{t-1},\ldots,Z_1)$ is equal to the conditional distribution of Z_{t+1} given (Z_t, Z_{t-1}). In general, if $Z_t = 1,\ldots,Z_{t-k+1} = 1$ but $Z_{t-k} \neq 1$, then the conditional

distribution of Z_{t+1} given $(Z_t, Z_{t-1}, \ldots)$ is equal to the conditional distribution of Z_{t+1} given $(Z_t, Z_{t-1}, \ldots, Z_{t-k})$.

Again the above result can be generalized to Z_{t+h} from Z_{t+1}, i.e., for any given integer $h \geq 1$, the conditional distribution of Z_{t+h} given $(Z_t = 1, Z_{t-1} = 1, \ldots, Z_{t-k+1} = 1, Z_{t-k} \neq 1, \ldots)$ is equal to the conditional distribution of Z_{t+h} given $(Z_t = 1, Z_{t-1} = 1, \ldots, Z_{t-k+1} = 1, Z_{t-k} \neq 1)$.

Since the process is not a Markov Chain, the conditional distribution of Z_t given past observations does not have any closed form expression. Thus the run distribution of zeros and ones, and expected length of those runs do not have any closed mathematical formula. Therefore, results related to expected length of runs of zeros and ones for the proposed process are difficult to compute.

4 Parameter estimation

4.1 Yule-Walker estimation

Given a data set $\{Z_1, Z_2, \ldots, Z_n\}$ of size n, we can write the following three moment equations to obtain the YW estimates of α, θ and p:

$$\hat{\mu}'_1 = 1 - p + p\theta^* \tag{8}$$

$$\hat{\mu}'_2 = p\theta^*(1 + 2\theta^*) + (1 - p) \tag{9}$$

$$\hat{\gamma}_z(1) = p^2\theta^*(1 + \theta^*)\alpha \tag{10}$$

where $\hat{\mu}'_1 = \frac{1}{n}\sum_{t=1}^{n} Z_t$, $\hat{\mu}'_2 = \frac{1}{n}\sum_{t=1}^{n} Z_t^2$, and $\hat{\gamma}_z(1) = \frac{1}{n}\sum_{t=2}^{n}(Z_t - \hat{\mu}'_1)(Z_{t-1} - \hat{\mu}'_1)$.

After solving the first two equations, we can obtain the YW estimates of p from the following quadratic equation

$$2p^2 + (5\hat{\mu}'_1 - \hat{\mu}'_2 - 4)p + 2(\hat{\mu}'_1 - 1)^2 = 0. \tag{11}$$

Suppose $\hat{p}_{yw}$ be the YW estimate of p, then from the first equation (8) we can get

$$\hat{\theta}^*_{yw} = \frac{\mu'_1 - 1 + \hat{p}_{yw}}{\hat{p}_{yw}}$$

which implies

$$\hat{\theta}_{yw} = \frac{\mu'_1 - 1 + 2\hat{p}_{yw}}{\mu'_1 - 1 + 2\hat{p}_{yw}}. \tag{12}$$

From the third moment Eq. (10), we get

$$\hat{\alpha}_{yw} = \frac{\hat{\gamma}_z(1)}{\hat{p}^2_{yw}\hat{\theta}^*_{yw}(1 + \hat{\theta}^*_{yw})}. \tag{13}$$

Proposition 3 *Under the above setup, the YW estimators of α, θ and p are consistent, i.e.,*

$$\hat{\alpha}_{yw} \xrightarrow{p} \alpha, \qquad \hat{\theta}_{yw} \xrightarrow{p} \theta, \qquad \hat{p}_{yw} \xrightarrow{p} p,$$

where '$\xrightarrow{p}$' denotes the convergence in probability.

Proof Proof is given in Appendix F. □

4.2 Quasi-maximum likelihood estimation

Suppose $\{Z_1, Z_2, \ldots, Z_n\}$ be set of n observations. In order to obtain the maximum likelihood estimates of OMGINAR(1) process, we have to maximize the log likelihood function

$$\begin{aligned}\ell_n(\alpha, \theta, p) &= \ln p(Z_1, Z_2, \ldots, Z_n) \\ &= \ln \Big(p(Z_1) p(Z_2 \mid Z_1) p(Z_3 \mid Z_2, Z_1) \ldots p(Z_n \mid Z_{n-1}, \ldots, Z_1) \Big)\end{aligned}$$

subject to the constraint $0 < \alpha, \theta, p < 1$, where for the proposed process, the conditional distribution of Z_t given the past observations can be written as

$$p(Z_t|Z_{t-1}, Z_{t-2}, \ldots) = \begin{cases} p(Z_t|Z_{t-1}) & \text{if } Z_{t-1} \neq 1 \\ p(Z_t|Z_{t-1}, Z_{t-2}) & \text{if } Z_{t-1} = 1, Z_{t-2} \neq 1 \\ p(Z_t|Z_{t-1}, Z_{t-2}, Z_{t-3}) & \text{if } Z_{t-1} = 1, Z_{t-2} = 1, Z_{t-3} \neq 1 \\ \vdots \end{cases}$$

In practice, beyond $k = 1$, $p(Z_t|Z_{t-1} = 1, Z_{t-2} = 1, \ldots, Z_{t-k+1} = 1, Z_{t-k} \neq 1)$ has a very cumbersome expression. For example, for $k = 2$, we will have $2^3 = 8$ different cumbersome expressions for the conditional distribution of $p(Z_t|Z_{t-1}, Z_{t-2})$. To avoid that, here we propose to use one-step QMLE where we maximize $\ell^*(\alpha, \theta, p) = \ln p(Z_1) + \sum_{t=2}^{n} \ln p(Z_t|Z_{t-1})$ instead of maximizing the actual likelihood function. In the next section, using some simulated data sets we study the consistency of this one-step QMLE method both with respect to bias and standard error.

5 Simulation study

In this section, we carried out some simulation experiments to compare the proposed model with some other existing INAR(1) models, namely the PINAR(1), GINAR(1), CPINAR(1) with Poisson_2, and ZINAR(1) models. For model validation, we generated samples from the proposed model, and compared the fit of the proposed model to the data with the above five models with respect to AIC and some h-step ahead forecasting accuracy measures, namely predicted root mean squared error or PRMSE(h), predicted

mean absolute error or PMAE(h) and percentage of true prediction PTP(h) which can be obtained using the following formulas:

$$\text{PRMSE}(h) = E\left((Y_{n+h} - \hat{Y}_{n+h})^2 \mid \mathbf{Y}_{n:1}\right) \hat{=} \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - \hat{Y}^{(h)}_{mean,n+i})^2}; \quad h = 1, 2, \ldots,$$

where $\hat{Y}^{(h)}_{mean,n+i} = \widehat{\text{mean}}(Y_{n+i}|Y_{n-h+i})$ be the h-step ahead conditional mean of the fitted process;

$$\text{PMAE}(h) = E\left(\left|Y_{n+h} - \hat{Y}_{n+h}\right| \mid \mathbf{Y}_{n:1}\right) \hat{=} \frac{1}{m}\sum_{i=1}^{m}\left|Y_{n+i} - \hat{Y}^{(h)}_{median,n+i}\right|; \quad h = 1, 2, \ldots,$$

where $\hat{Y}^{(h)}_{median,n+i} = \widehat{\text{median}}(Y_{n+i}|Y_{n-h+i})$ be the h-step ahead conditional median of the fitted process; and

$$\text{PTP}(h) = E\Big(I(Y_{n+h} = \hat{Y}_{n+h}) \mid \mathbf{Y}_{n:1}\Big) \hat{=} \frac{1}{m}\sum_{i=1}^{m} I(Y_{n+i} = \hat{Y}^{(h)}_{mode,n+i}) \times 100; \quad h = 1, 2, \ldots,$$

where $\hat{Y}^{(h)}_{mode,n+i} = \widehat{\text{mode}}(Y_{n+i}|Y_{n-h+i})$ be the h-step ahead conditional mode of the fitted process, and $\mathbf{Y}_{n:1} = (Y_n, Y_{n-1}, \ldots, Y_1)$. To study the robustness of the proposed model, we generated samples from the PINAR(1), GINAR(1), CPINAR(1) with with Poisson$_2$, and ZINAR(1) models.

To begin with, we generated samples from the OMGINAR(1) process. We set the parameter values $\alpha = 0.3, 0.6$, $\theta = 0.6$, and $p = 0.5, 0.7, 0.9$ and sample sizes $n = 100, 500, 1000, 5000$. Note that α, the first order ACF of the hidden process GINAR(1), can take any value between 0 and 1. Therefore, we set $\alpha = 0.3$ for the class of lower ACF values and $\alpha = 0.6$ for the class of higher ACF values. Here $\theta = 0.6$ was chosen based on some real data examples. However, the mixture parameter p that plays the main role in differentiating the proposed model from the GINAR(1) process was varied between 0.5 and 1. Here p close to 1 implies the process almost equals to the GINAR(1) process and close to 0 implies that the resulting process coincides with a degenerate process at 1. So we decided to set p not very close to 0. Besides, samples of size 100 were used to study the small-sample properties, samples of size 5000 were used to get an idea about the large-sample properties, and samples of sizes 500 and 1000 were used for moderate-sample properties. For a fixed sample size and fixed set of parameter values, we generated the samples 500 times and obtained the average estimates of parameters along with their biases and standard errors. The estimated parameters with their biases and standard errors are presented in Table 1. As we can see, there is very little effect on the biases of the QMLE estimates whereas standard errors converge to zero, as the sample size increases.

For model validation, first we performed a comparison between our proposed model and other competing models mentioned earlier with respect to AIC. We repeated the above simulation experiment and computed the AIC for all the models under comparison. Based on 500 Monte Carlo replications, we reported the percentage of

Table 1 Parameter estimates of the OMGINAR(1) model using quasi maximum-likelihood method along with its bias and standard error

Sample size (n)	α			θ			p		
	$\hat{\alpha}_{qmle}$	bias	se	$\hat{\theta}_{qmle}$	bias	se	$\hat{p}_{qmle}$	bias	se
$\alpha = 0.3, \theta = 0.6, p = 0.5$									
100	0.302	0.0027	0.1562	0.581	0.0185	0.0552	0.485	0.0144	0.0669
500	0.302	0.0021	0.0679	0.585	0.0148	0.0246	0.485	0.0144	0.0296
1000	0.304	0.0044	0.0495	0.586	0.0131	0.0166	0.485	0.0145	0.0206
5000	0.301	0.0019	0.0211	0.587	0.0127	0.0072	0.484	0.0157	0.0091
$\alpha = 0.3, \theta = 0.6, p = 0.7$									
100	0.293	0.0060	0.1102	0.588	0.0116	0.0462	0.684	0.0157	0.0686
500	0.299	0.0006	0.0484	0.590	0.0093	0.0209	0.687	0.0128	0.0296
1000	0.301	0.0017	0.0321	0.591	0.0085	0.0152	0.688	0.0116	0.0210
5000	0.304	0.0043	0.0153	0.591	0.0082	0.0065	0.688	0.0117	0.0097
$\alpha = 0.3, \theta = 0.6, p = 0.9$									
100	0.297	0.0024	0.0847	0.592	0.0078	0.0447	0.895	0.0044	0.0625
500	0.301	0.0010	0.0350	0.596	0.0037	0.0194	0.897	0.0027	0.0276
1000	0.303	0.0031	0.0244	0.597	0.0029	0.0139	0.894	0.0050	0.0199
5000	0.303	0.0035	0.0116	0.597	0.0024	0.0060	0.895	0.0040	0.0089
$\alpha = 0.6, \theta = 0.6, p = 0.5$									
100	0.613	0.0134	0.1559	0.561	0.0385	0.0730	0.401	0.0981	0.0695
500	0.631	0.0312	0.0588	0.569	0.0304	0.0296	0.406	0.0938	0.0338
1000	0.624	0.0245	0.0407	0.569	0.0305	0.0215	0.408	0.0912	0.0229
5000	0.625	0.0259	0.0169	0.571	0.0283	0.0089	0.409	0.0908	0.0107

Table 1 continued

Sample size (n)	α			θ			p		
	$\hat{\alpha}_{qmle}$	bias	se	$\hat{\theta}_{qmle}$	bias	se	$\hat{p}_{qmle}$	bias	se
$\alpha = 0.6, \theta = 0.6, p = 0.7$									
100	0.620	0.0208	0.0947	0.566	0.0337	0.0602	0.635	0.0646	0.0932
500	0.631	0.0311	0.0366	0.577	0.0222	0.0227	0.635	0.0641	0.0413
1000	0.626	0.0269	0.0268	0.576	0.0239	0.0178	0.635	0.0640	0.0304
5000	0.630	0.0306	0.0116	0.577	0.0224	0.0078	0.638	0.0618	0.0135
$\alpha = 0.6, \theta = 0.6, p = 0.9$									
100	0.619	0.0192	0.0604	0.589	0.0108	0.0565	0.886	0.0132	0.0684
500	0.616	0.0166	0.0255	0.585	0.0144	0.0239	0.893	0.0064	0.0265
1000	0.616	0.0164	0.0191	0.589	0.0106	0.0185	0.896	0.0038	0.0210
5000	0.616	0.0163	0.0084	0.589	0.0102	0.0078	0.897	0.0026	0.0099

Table 2 Selection percentages of various models under study through AIC where the data are generated from the OMGINAR(1) process with various sets of parameter values

Sample size (n)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.6, p = 0.5$				
100	0	0	0	0	100.0
500	0	0	0	0	100.0
1000	0	0	0	0	100.0
5000	0	0	0	0	100.0
	$\alpha = 0.3, \theta = 0.6, p = 0.7$				
100	0.2	0	0.2	0	99.6
500	0	0	0	0	100.0
1000	0	0	0	0	100.0
5000	0	0	0	0	100.0
	$\alpha = 0.3, \theta = 0.6, p = 0.9$				
100	0.2	32.7	6.9	1.2	59.0
500	0.0	0.4	0.0	0.0	99.6
1000	0.0	0.0	0.0	0.0	100.0
5000	0.0	0.0	0.0	0.0	100.0
	$\alpha = 0.6, \theta = 0.6, p = 0.5$				
100	0	0	0	0	100.0
500	0	0	0	0	100.0
1000	0	0	0	0	100.0
5000	0	0	0	0	100.0
	$\alpha = 0.6, \theta = 0.6, p = 0.7$				
100	0.5	0.5	0	0	99.0
500	0	0	0	0	100.0
1000	0	0	0	0	100.0
5000	0	0	0	0	100.0
	$\alpha = 0.6, \theta = 0.6, p = 0.9$				
100	0	27.5	7.5	2.5	62.5
500	0	0.0	0.0	0.0	100.0
1000	0	0.0	0.0	0.0	100.0
5000	0	0.0	0.0	0.0	100.0

times AIC selects a particular model among the set of six models in Table 2. It turns out that as the sample size increases, AIC selects the OMGINAR(1) model almost all the time. In other words, the proposed model is consistent with respect to AIC.

We also performed a similar study with respect to point forecasting accuracy measures, namely PRMSE(h), PMAE(h) and PTP(h). In this case, we fixed the sample size $n = 400$ and repeated the above simulation experiment for all the above set of parameter values. To compute the accuracy measures, we divided each sample into two parts. The first part consisting of 300 (training set) observations were used to fit the models

under comparison and the remaining 100 observations (validation set) were used to calculate the above three accuracy measures for $h = 1, 2, 3, 4, 5$. Based on 500 Monte Carlo replications, we computed the average values of these measures and reported them in Tables 3, 4 and 5. As we can see, our proposed model performs better than the competing models considered in this study, as the mixture parameter decreases. In other words, as the mixture parameter decreases, the proposed model deviates from the GINAR(1) model, resulting in higher forecasting accuracy than the GINAR(1) and other four competing models. Furthermore, it is observed that as α increases, the forecasting accuracy also increases across all the models. This is because the mean and variance of the innovation process ε_t of the hidden process GINAR(1) decreases to zero as α increases to one (i.e., the innovation process converges to a degenerate process degenerated at 0). On the other hand, as we can see from Tables 3, 4 and 5, the forecasting accuracy decreases across all the models as h increases; this finding is in conformity with our intuitive expectation that as one goes far ahead from the present, chances of making an accurate forecast will decrease.

To study the robustness of the OMGINAR(1) model, we simulated data from different models, viz, PINAR(1), GINAR(1), CPINAR(1) with Poisson_2, ZINAR(1) and ZMGINAR models, and computed the percentage of times the model is selected by AIC and how it performs with respect to the above forecasting accuracy measures. However, we could not include the ZMGINAR(1) model in the simulation study because for some simulated data sets, the estimated value of α for the ZMGINAR(1) model was out of the parametric space $(\max\{0, \frac{\pi\mu}{1+\pi\mu}\}, \frac{\mu}{1+\mu})$. However, for the Poliomyelitis and assault data analyses in Sect. 6, the estimated parameter α for the ZMGINAR(1) model lies in the above restricted interval. So, we included this important model in the data analysis section but not here. For all the data generating models, we set $\alpha = 0.3, 0.6$ for all the models. Individually, we set $\lambda = 1, 1.5$ for the PINAR(1) and CPINAR(1) models, $\theta = 0.5, 0.6$ for the GINAR(1) model, and $\lambda = 1.5$, $\rho = 0.1, 0.3, 0.5$ for the ZINAR(1) model. For each data generating process (DGP), we repeated the above procedure to compute the h-step ahead forecasting accuracy measures. We only reported the results based on DGP PINAR(1) in Tables 6, 7 and 8; and on DGP GINAR(1) in Tables 9, 10 and 11. For other DGPs, similar kind of observations are observed. So we skipped those results here. As we can see from all those results, our model performs at least as good as (often better than) the GINAR(1) process in terms of the forecasting accuracy measures. Irrespective of whether the data were generated from PINAR(1) or GINAR(1) or CPINAR(1) or ZINAR(1), our proposed model always has lower forecasting errors compared to the GINAR(1) process. This is because the proposed process is a more generalized version of the GINAR(1) process; more specifically, when the mixing parameter p is 1, the proposed process reduces to the GINAR(1) process. While we are considering a more complicated process compared to the GINAR(1) process by introducing an extra parameter p, the added complexity of our proposed process is offset by the improved fitting and forecasting accuracy measures. In addition, note that, here our objective is to improve the fitting of the data, not the inference of the model parameters. In that respect, our approach is quite successful.

Table 3 Values of PRMSE(h) for different models and for varying h where the data were simulated from the OMGINAR(1) process with various sets of parameter values

h-step	PRMSE(h)					PRMSE(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.6, P = 0.5$					$\alpha = 0.3, \theta = 0.6, p = 0.7$				
1	1.318	1.357	1.388	1.319	1.316	1.544	1.547	1.559	1.552	1.542
2	1.329	1.344	1.369	1.329	1.331	1.585	1.586	1.600	1.590	1.586
3	1.329	1.333	1.348	1.329	1.329	1.586	1.587	1.603	1.591	1.587
4	1.329	1.329	1.338	1.329	1.329	1.585	1.585	1.597	1.586	1.585
5	1.329	1.329	1.332	1.329	1.329	1.585	1.585	1.597	1.586	1.585
	$\alpha = 0.3, \theta = 0.6, P = 0.9$					$\alpha = 0.6, \theta = 0.6, p = 0.5$				
1	1.779	1.772	1.773	1.773	1.774	1.122	1.180	1.182	1.186	1.105
2	1.861	1.852	1.859	1.856	1.854	1.166	1.189	1.211	1.201	1.157
3	1.874	1.866	1.875	1.871	1.871	1.163	1.176	1.192	1.187	1.161
4	1.875	1.868	1.879	1.873	1.873	1.160	1.166	1.183	1.175	1.159
5	1.875	1.870	1.878	1.875	1.874	1.158	1.160	1.172	1.166	1.158
	$\alpha = 0.6, \theta = 0.6, P = 0.7$					$\alpha = 0.6, \theta = 0.6, p = 0.9$				
1	1.346	1.372	1.358	1.369	1.342	1.514	1.511	1.510	1.512	1.510
2	1.441	1.448	1.462	1.461	1.445	1.707	1.704	1.705	1.705	1.706
3	1.456	1.460	1.470	1.466	1.464	1.769	1.765	1.767	1.767	1.769
4	1.460	1.463	1.477	1.470	1.469	1.799	1.788	1.792	1.790	1.792
5	1.459	1.458	1.473	1.466	1.466	1.811	1.800	1.809	1.807	1.805

Table 4 Values of PMAE(h) for varying h where the data were simulated from the OMGINAR(1) process with various sets of parameter values

h-step	PMAE(h)					PMAE(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.6, p = 0.5$					$\alpha = 0.3, \theta = 0.6, p = 0.7$				
1	0.621	0.761	0.776	0.622	0.612	0.878	0.929	0.950	0.927	0.868
2	0.612	0.747	0.679	0.613	0.612	0.870	0.942	0.893	0.880	0.870
3	0.612	0.757	0.629	0.612	0.612	0.870	0.874	0.872	0.871	0.870
4	0.612	0.744	0.615	0.612	0.612	0.870	0.871	0.870	0.870	0.870
5	0.612	0.721	0.612	0.612	0.612	0.870	0.870	0.870	0.870	0.870
	$\alpha = 0.3, \theta = 0.6, p = 0.9$					$\alpha = 0.6, \theta = 0.6, p = 0.5$				
1	1.148	1.125	1.140	1.134	1.127	0.545	0.562	0.588	0.586	0.493
2	1.175	1.171	1.178	1.176	1.170	0.523	0.627	0.651	0.646	0.493
3	1.175	1.169	1.170	1.170	1.169	0.502	0.629	0.594	0.624	0.493
4	1.175	1.169	1.169	1.169	1.169	0.495	0.776	0.522	0.535	0.493
5	1.175	1.169	1.169	1.169	1.169	0.493	0.863	0.505	0.500	0.493
	$\alpha = 0.6, \theta = 0.6, p = 0.7$					$\alpha = 0.6, \theta = 0.6, p = 0.9$				
1	0.736	0.722	0.745	0.736	0.727	0.947	0.832	0.843	0.836	0.832
2	0.773	0.847	0.855	0.867	0.766	1.096	1.069	1.088	1.077	1.071
3	0.762	0.871	0.787	0.813	0.760	1.125	1.152	1.127	1.140	1.130
4	0.757	0.874	0.772	0.768	0.757	1.126	1.138	1.128	1.125	1.124
5	0.757	0.855	0.761	0.758	0.757	1.127	1.123	1.125	1.124	1.122

Table 5 Values of PTP(h) for varying h where the data were simulated from the OMGINAR(1) process with various sets of parameter values

h-step	PTP(h)					PTP(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.6, P = 0.5$					$\alpha = 0.3, \theta = 0.6, p = 0.7$				
1	61.07	39.25	49.96	60.99	63.58	46.97	35.54	34.63	36.53	48.22
2	63.58	23.98	25.44	63.39	63.58	48.14	29.40	30.54	32.87	48.14
3	63.58	21.09	22.77	63.37	63.58	48.14	28.42	28.64	30.75	48.14
4	63.58	20.28	20.97	63.37	63.58	48.14	28.41	28.41	30.53	48.14
5	63.58	20.24	20.35	63.37	63.58	48.14	28.41	28.41	30.53	48.14
	$\alpha = 0.3, \theta = 0.6, P = 0.9$					$\alpha = 0.6, \theta = 0.6, p = 0.5$				
1	31.03	41.13	41.09	41.12	41.35	63.94	63.81	62.43	62.55	70.10
2	31.51	36.46	36.61	36.56	36.69	63.10	23.71	41.44	42.42	70.10
3	31.51	36.01	36.05	36.02	35.91	66.39	21.19	22.50	22.86	70.10
4	31.51	36.00	36.00	36.00	35.62	68.47	19.55	21.47	21.63	70.10
5	31.51	36.00	36.00	36.00	35.75	69.13	18.07	20.04	20.17	70.10
	$\alpha = 0.6, \theta = 0.6, P = 0.7$					$\alpha = 0.6, \theta = 0.6, p = 0.9$				
1	55.51	55.65	55.68	55.67	57.78	51.96	52.59	52.53	52.55	52.67
2	51.76	32.39	31.66	31.82	53.43	35.13	40.85	40.59	40.71	42.54
3	54.01	28.87	30.55	30.20	54.29	34.08	37.89	38.06	37.80	38.83
4	54.27	26.74	28.49	28.22	54.32	34.21	35.78	36.91	36.03	37.06
5	54.27	25.94	26.97	26.69	54.32	34.18	35.11	35.82	35.26	36.06

Table 6 Values of PRMSE(h) for varying h where the data were simulated from the PINAR(1) process with various sets of parameter values

h-step	PRMSE(h)					PRMSE(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \lambda = 1$					$\alpha = 0.3, \lambda = 1.5$				
1	0.988	0.990	0.987	0.988	0.991	1.197	1.21	1.197	1.197	1.207
2	1.004	1.004	1.003	1.004	1.004	1.287	1.30	1.285	1.286	1.301
3	1.004	1.004	1.004	1.004	1.004	1.313	1.31	1.312	1.313	1.314
4	1.004	1.004	1.004	1.004	1.004	1.315	1.31	1.313	1.318	1.316
5	1.004	1.004	1.004	1.004	1.004	1.315	1.31	1.317	1.316	1.316
	$\alpha = 0.6, \lambda = 1$					$\alpha = 0.6, \lambda = 1.5$				
1	0.855	0.873	0.856	0.855	0.866	1.019	1.040	1.021	1.020	1.037
2	0.972	0.995	0.975	0.972	0.975	1.172	1.212	1.170	1.171	1.211
3	0.988	0.991	0.991	0.989	0.990	1.241	1.292	1.242	1.240	1.268
4	0.993	0.991	0.993	0.993	0.992	1.282	1.317	1.282	1.282	1.308
5	0.992	0.991	0.992	0.991	0.991	1.301	1.320	1.304	1.302	1.318

Table 7 Values of PMAE(h) for varying h where the data were simulated from the PINAR(1) process with various sets of parameter values

h-step	PMAE(h)					PMAE(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \lambda = 1$					$\alpha = 0.3, \lambda = 1.5$				
1	0.735	0.735	0.736	0.736	0.736	0.895	0.974	0.962	0.896	0.985
2	0.742	0.909	0.790	0.742	0.760	0.950	1.046	0.946	0.950	0.954
3	0.742	0.980	0.743	0.742	0.743	0.950	0.955	0.952	0.949	0.950
4	0.742	0.998	0.742	0.742	0.743	0.949	0.950	0.950	0.950	0.950
5	0.742	1.000	0.742	0.742	0.743	0.950	0.950	0.950	0.950	0.950
	$\alpha = 0.6, \lambda = 1$					$\alpha = 0.6, \lambda = 1.5$				
1	0.554	0.562	0.556	0.555	0.562	0.726	0.778	0.725	0.726	0.777
2	0.710	0.740	0.711	0.714	0.713	0.880	0.975	0.901	0.880	0.973
3	0.738	0.875	0.761	0.738	0.774	0.926	1.041	0.924	0.925	1.038
4	0.737	0.932	0.743	0.737	0.803	0.949	1.091	0.947	0.949	0.990
5	0.736	0.978	0.736	0.736	0.803	0.960	1.121	0.958	0.960	0.960

Table 8 Values of PTP(h) for varying h where the data were simulated from the PINAR(1) process with various sets of parameter values

h-step	PTP(h)					PTP(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \lambda = 1$					$\alpha = 0.3, \lambda = 1.5$				
1	42.70	40.67	41.12	42.48	41.60	33.48	29.13	30.85	33.46	31.87
2	37.82	37.06	37.21	37.90	37.69	33.10	23.08	25.65	33.10	26.91
3	36.97	37.02	37.03	36.80	37.03	33.11	22.25	22.87	33.11	23.86
4	36.85	37.02	37.02	36.30	37.06	33.11	22.25	22.30	33.11	22.67
5	36.77	37.02	37.02	36.50	37.04	33.11	22.25	22.25	33.11	22.36
	$\alpha = 0.6, \lambda = 1$					$\alpha = 0.6, \lambda = 1.5$				
1	52.78	52.91	52.92	52.78	52.61	42.95	41.41	42.53	42.95	41.44
2	43.50	41.39	42.13	43.49	43.02	34.36	29.78	33.44	34.33	32.12
3	40.17	37.40	38.92	40.24	39.58	33.13	24.52	28.84	32.99	28.88
4	38.76	37.04	37.40	38.62	38.16	32.85	22.63	27.03	32.87	26.54
5	37.81	37.03	37.06	37.86	37.63	32.95	22.45	25.68	32.95	24.63

Table 9 Values of PRMSE(h) for varying h where the data were simulated from the GINAR(1) process with various sets of parameter values

h-step	PRMSE(h)					PRMSE(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.5$					$\alpha = 0.3, \theta = 0.6$				
1	1.347	1.344	1.344	1.343	1.344	1.838	1.832	1.833	1.832	1.832
2	1.379	1.379	1.379	1.379	1.379	1.947	1.941	1.944	1.938	1.940
3	1.378	1.378	1.378	1.378	1.378	1.965	1.964	1.962	1.962	1.965
4	1.378	1.378	1.378	1.378	1.378	1.964	1.966	1.965	1.965	1.967
5	1.378	1.378	1.378	1.378	1.378	1.964	1.966	1.963	1.965	1.967
	$\alpha = 0.6, \theta = 0.5$					$\alpha = 0.6, \theta = 0.6$				
1	1.175	1.162	1.164	1.163	1.162	1.535	1.530	1.531	1.531	1.530
2	1.338	1.335	1.336	1.335	1.335	1.767	1.757	1.759	1.758	1.757
3	1.374	1.373	1.373	1.374	1.373	1.869	1.855	1.854	1.857	1.856
4	1.385	1.386	1.385	1.386	1.385	1.910	1.897	1.897	1.898	1.897
5	1.385	1.386	1.386	1.386	1.386	1.934	1.922	1.919	1.920	1.924

Table 10 Values of PMAE(h) for varying h where the data were simulated from the GINAR(1) process with various sets of parameter values

h-step	PMAE(h)					PMAE(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.5$					$\alpha = 0.3, \theta = 0.6$				
1	0.970	0.828	0.828	0.829	0.828	1.248	1.167	1.223	1.170	1.168
2	0.999	0.952	0.967	0.951	0.950	1.297	1.287	1.289	1.288	1.287
3	0.999	0.981	0.988	0.981	0.980	1.301	1.288	1.288	1.288	1.288
4	0.999	0.988	0.995	0.987	0.991	1.301	1.288	1.288	1.288	1.288
5	0.999	0.992	0.994	0.988	0.992	1.302	1.288	1.288	1.288	1.288
	$\alpha = 0.6, \theta = 0.5$					$\alpha = 0.6, \theta = 0.6$				
1	0.593	0.590	0.592	0.591	0.591	0.979	0.817	0.818	0.817	0.817
2	0.881	0.792	0.791	0.792	0.792	1.218	1.087	1.168	1.087	1.087
3	0.992	0.883	0.890	0.883	0.883	1.282	1.224	1.269	1.224	1.224
4	1.004	0.927	0.959	0.927	0.926	1.311	1.286	1.290	1.290	1.286
5	1.003	0.950	0.983	0.953	0.949	1.324	1.288	1.297	1.291	1.288

Table 11 Values of PTP(h) for varying h where the data were simulated from the GINAR(1) process with various sets of parameter values

h-step	PTP(h)					PTP(h)				
	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)	PINAR(1)	GINAR(1)	CPINAR(1)	ZINAR(1)	OMGINAR(1)
	$\alpha = 0.3, \theta = 0.5$					$\alpha = 0.3, \theta = 0.6$				
1	49.76	52.17	52.22	52.22	52.21	23.84	44.27	44.35	44.41	44.24
2	41.66	50.31	50.33	50.32	50.21	23.54	41.45	41.35	41.42	41.50
3	38.93	50.30	50.30	50.30	50.31	23.63	41.38	41.39	41.39	41.41
4	38.88	50.30	50.30	50.30	50.30	23.63	41.39	41.39	41.39	41.39
5	38.99	50.30	50.30	50.30	50.30	23.63	41.39	41.39	41.39	41.39
	$\alpha = 0.6, \theta = 0.5$					$\alpha = 0.6, \theta = 0.6$				
1	62.18	62.33	62.34	62.35	62.34	54.84	55.50	55.46	55.47	55.45
2	51.76	54.47	54.47	54.45	54.41	26.84	46.64	46.40	46.64	46.57
3	47.48	51.85	51.86	51.76	51.88	23.08	43.34	42.68	43.38	43.31
4	42.75	51.20	51.17	51.09	51.15	23.17	42.15	40.98	42.17	42.07
5	39.38	50.88	50.85	50.85	50.87	23.32	42.05	41.05	41.99	42.04

Table 12 Estimated parameters, AIC, χ^2-goodness of fit, and different h-step ahead forecasting accuracy measures for the monthly US poliomyelitis data set where $h = 1$

Model	Estimated values	AIC	χ^2-goodness of fit	PRMSE	PMAE	PTP
PINAR(1)	$\hat{\alpha}_{mle} = 0.184$	584.81	12.37	1.431	0.950	30.00
	$\hat{\lambda}_{mle} = 1.337$					
GINAR(1)	$\hat{\alpha}_{mle} = 0.041$	538.67	0.19	1.431	0.950	45.00
	$\hat{\theta}_{mle} = 0.569$					
CPINAR(1)	$\hat{\alpha}_{mle} = 0.141$	563.36	0.76	1.431	0.950	45.00
	$\hat{\lambda}_{mle} = 0.815$					
ZINAR(1)	$\hat{\alpha}_{mle} = 0.177$	569.01	1.30	1.431	0.950	45.00
	$\hat{\lambda}_{mle} = 1.593$					
	$\hat{\rho}_{mle} = 0.308$					
ZMGINAR(1)	$\hat{\alpha}_{cls} = 0.306$	582.598	0.542	1.746	0.950	45.00
	$\hat{\mu}_{cls} = 1.473$					
	$\hat{\pi}_{cls} = 0.113$					
OMGINAR(1)	$\hat{\alpha}_{qmle} = 0.068$	534.01	0.019	1.431	0.950	45.00
	$\hat{\theta}_{qmle} = 0.577$					
	$\hat{p}_{qmle} = 0.882$					

6 Data analysis

6.1 Poliomyelitis data

We consider the monthly cases of poliomyelitis data in the US for a period of 14 years from 1970 to 1983. This data set was first analyzed by Zeger (1988). In particular, it has 168 observations; out of which 64 (38%) observations are zero, 55 (32%) observations are one, and remaining 49 (30%) observations have monthly cases more than one. The marginal mean and marginal variance are computed as 1.33 and 3.50, and hence the dispersion index which is defined as the ratio of variance and mean is given as 2.63. It indicates that the data is over-dispersed. The raw data along with its ACF and PACF are plotted in Fig. 1 to see the characteristic of the data.

Since first lag of PACF plot in Fig. 1 is significant, we fitted most of the existing INAR(1) models, namely Poisson INAR(1), over-dispersed models like GINAR(1) and CPINAR(1), zero-inflated models like ZINAR(1) and ZMGINAR(1), and our proposed OMGINAR(1) model to the data to facilitate model comparison. Based on these fitted models, we computed the respective expected frequencies and plotted them in Fig. 2 with the observed frequencies. As we can see, while the PINAR(1) process underestimates the zero cases, the GINAR(1) process overestimates the zero observations and underestimates the one observations. This kind of limitations is seen

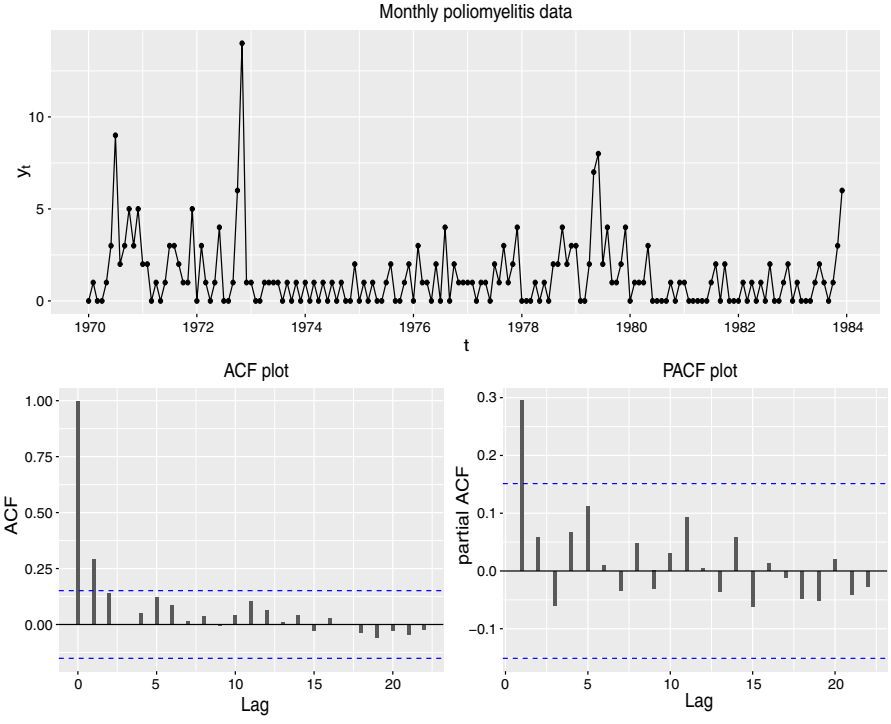

Fig. 1 Monthly cases of poliomyelitis in US during 1970 to 1983, and its ACF and PACF plots

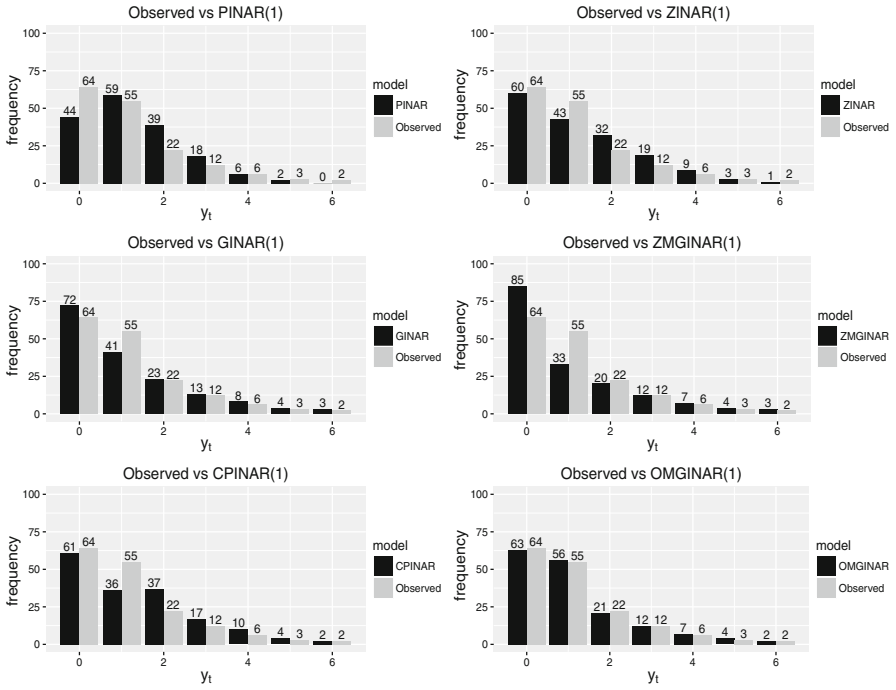

Fig. 2 Observed vs expected frequency distribution for the monthly poliomyelitis data

with the other existing models under consideration. However, as we can see from Fig. 2 and χ^2-goodness of fit statistic from Table 12, our proposed OMGINAR(1) model outperforms the other models with respect to the observed-expected frequency distributions comparison.

Furthermore, we examined the effectiveness of the proposed model over other existing INAR(1) models mentioned above with respect to some forecasting accuracy measures. To compute the forecasting accuracy measures, we divided the data set into two parts, the first part consisting of the first 148 observations (training set) were used to fit the models under comparison and the remaining 20 observations (validation set) were used to compute all the three forecasting accuracy measures. We presented the results based on the forecasting accuracy measures and AIC in Table 12. The results show that except PINAR(1) model, other models have same forecasting accuracy measures. However, our proposed model has the lowest AIC value which indicates that it fits the data best among all the existing models considered in this study.

6.2 Aggravated assault data

In our second application, we analyzed a monthly aggravated assault data set that gives the monthly cases of aggravated assault reported in the 34th police car beat in Pittsburgh, US. The data was first analyzed by Barreto-Souza (2015) using a zero-

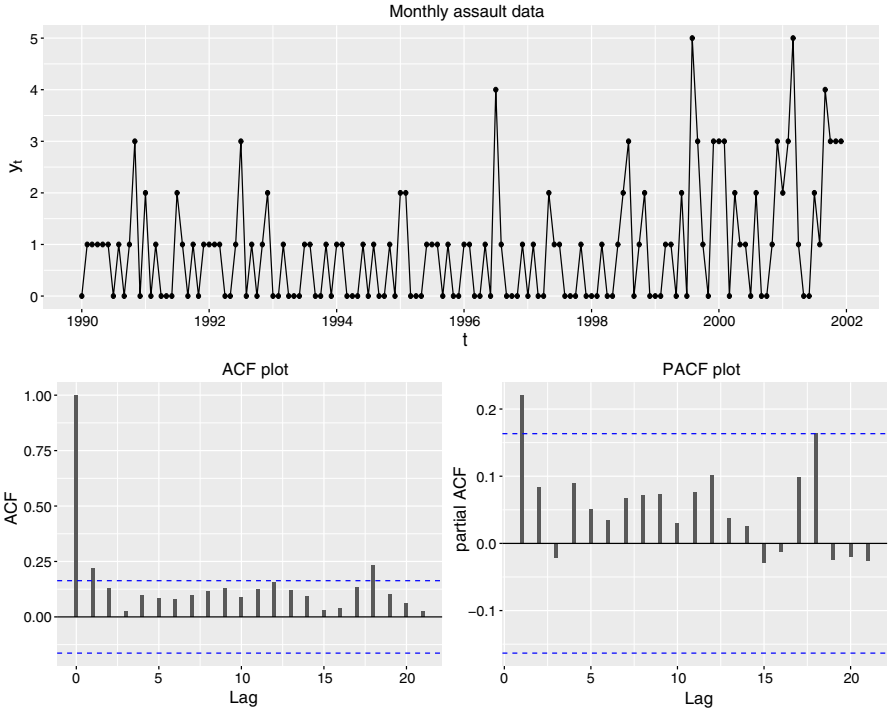

Fig. 3 Monthly cases of aggravated assault data reported in Pittsburgh, and its ACF and PACF plots

mixture of geometric INAR(1) process. The marginal mean and variance for the data set are 0.845 and 0.997, and hence the dispersion index was computed as 1.179. The data set contains 144 observations from January 1990 to December 2000. We presented the data along with its ACF and PACF in Fig. 3.

We fitted all the six models mentioned above and reported their respective estimated parameter values along with AIC, PRMSE, PMAE and PTP. As we can see, the ZMGINAR(1) model has the lowest AIC value, however our newly proposed model has the second lowest AIC value. Also to see the difference more closely, we used pairwise bar plot for all the models against the observed data; these plots are displayed in Fig. 4. As like the poliomyelitis data, here also PINAR(1) process underestimates zero, and overestimates one, whereas both CPINAR(1) and GINAR(1) processes overestimate zero and underestimate one. However, ZINAR(1) and ZMGINAR(1) processes estimate zeros and ones better than the other existing processes but poorly perform on the other observations. In contrast, our proposed model fits all the observations better than its competitors. This is also clear from the χ^2-goodness of fit statistic given in Table 13.

To compute the forecasting accuracy measures, we divided the data set into two parts, the first part consisting of the first 124 observations (training set) were used to fit the models under comparison and the remaining 20 observations (validation set) were used to compute all the three forecasting accuracy measures. From Table 13, we

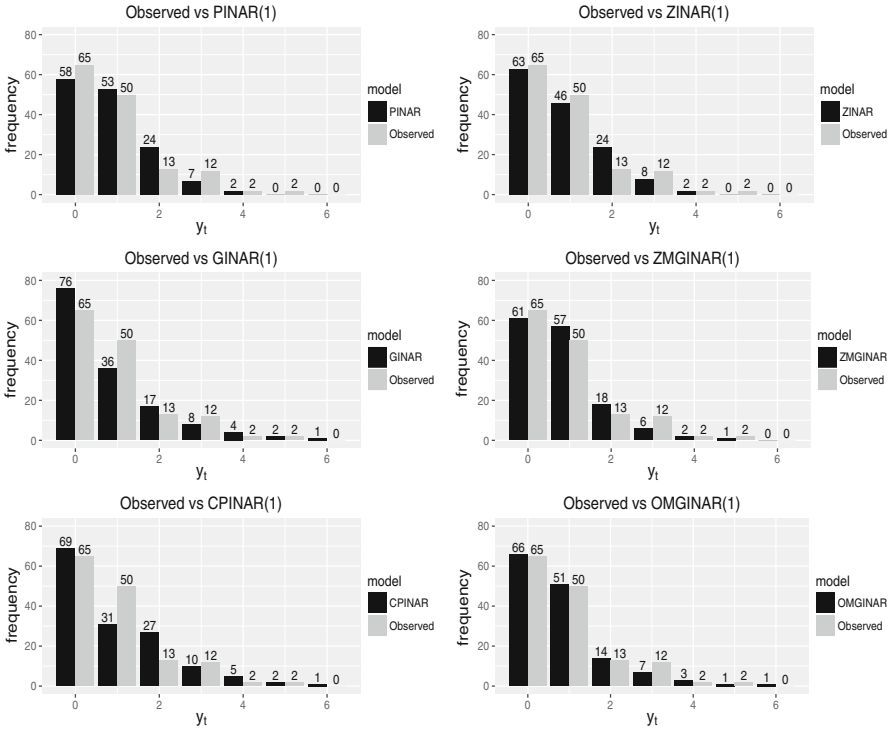

Fig. 4 Observed vs expected frequency distribution for the monthly assault data

see that there is not much difference among these models in terms of the forecasting measures except the ZMGINAR(1) model. The ZMGINAR(1) model has the highest forecasting accuracy in terms of the PTP measure but it has the lowest forecasting accuracy with respect to both PRMSE and PMAE measures. On the other hand, our proposed model along with PINAR(1) and ZINAR(1) models jointly perform better than the others in terms of the PRMSE and PMAE measures. Therefore, the proposed model can be an alternative choice in this case as well.

7 Discussion

In this paper, we proposed a new mixture of geometric INAR(1) process for modeling under and over-dispersed count time series data. We studied the stochastic properties, such as stationarity and ergodicity of the proposed process. We also discussed the h-step ahead coherent forecasting for the proposed model. Some simulated experiments and two real data analyses showed that the proposed model performs at least as good as (and often better than) the GINAR(1) and some other existing over-dispersed and zero-inflated processes.

In particular, we studied two different methods of parameter estimation, namely YW and QMLE for the proposed model. Mathematically we proved the consistency

Table 13 Estimated parameters, AIC, χ^2-goodness of fit, and different h-step ahead forecasting accuracy measures for the monthly aggravated assault data set where $h = 1$

Model	Estimated values	AIC	χ^2-goodness of fit	PRMSE	PMAE	PTP
PINAR(1)	$\hat{\alpha}_{mle} = 0.148$	377.74	17.91	1.378	1.100	25.00
	$\hat{\lambda}_{mle} = 0.904$					
GINAR(1)	$\hat{\alpha}_{mle} = 0.055$	382.59	0.88	1.378	1.600	25.00
	$\hat{\theta}_{mle} = 0.470$					
CPINAR(1)	$\hat{\alpha}_{mle} = 0.197$	393.45	1.09	1.378	1.600	25.00
	$\hat{\lambda}_{mle} = 0.549$					
ZINAR(1)	$\hat{\alpha}_{mle} = 0.145$	377.64	5.94	1.378	1.100	25.00
	$\hat{\lambda}_{mle} = 0.940$					
	$\hat{\rho}_{mle} = 0.175$					
ZMGINAR(1)	$\hat{\alpha}_{cls} = 0.226$	354.27	1.89	1.50	1.700	30.00
	$\hat{\mu}_{cls} = 0.623$					
	$\hat{\pi}_{cls} = -0.425$					
OMGINAR(1)	$\hat{\alpha}_{qmle} = 0.095$	376.90	0.84	1.378	1.100	25.00
	$\hat{\theta}_{qmle} = 0.465$					
	$\hat{p}_{qmle} = 0.861$					

of the YW estimators. While the consistency of the QMLE estimators are not proved theoretically, we empirically illustrated their consistency via extensive simulation experiments.

Although, our study is restricted to the allocating of weight (or probability mass) at one point, the proposed method can easily be extended for more than one points depending on the nature of the data. Here, we made our weight distribution using an i.i.d. Bernoulli process, however a data-driven structure can also be employed by replacing the i.i.d. Bernoulli process with a two-state Markov chain on $\{0, 1\}$ to potentially improve the forecasting performance even better. Since we wanted to keep things relatively simple, we did not pursue this in this current article. However, we recognize this as a promising future research direction.

Acknowledgements The authors would like to thank the reviewer and the associate editor for their careful reading and constructive suggestions which led to this improved version of the paper.

References

Al-Osh M, Alzaid AA (1987) First-order integer-valued autoregressive (INAR(1)) process. J Time Ser Anal 8(3):261–275

Barreto-Souza W (2015) Zero-modified geometric INAR(1) process for modelling count time series with deflation or inflation of zeros. J Time Ser Anal 36:839–852

Freeland RK, McCabe B (2005) Asymptotic properties of CLS estimators in the Poisson AR(1) model. Stat Probab Lett 73(2):147–153

Freeland RK, McCabe BP (2004) Forecasting discrete valued low count time series. Int J Forecast 20(3):427–434

Jazi MA, Jones G, Lai CD (2012) First-order integer valued AR processes with zero inflated Poisson innovations. J Time Ser Anal 33(6):954–963
Latour A (1998) Existence and stochastic structure of a non-negative integer-valued autoregressive process. J Time Ser Anal 19(4):439–455
Maiti R, Biswas A, Das S (2015) Time series of zero-inflated counts and their coherent forecasting. J Forecast 34(8):694–707
McCabe B, Martin GM (2005) Bayesian predictions of low count time series. Int J Forecast 21(2):315–330
McKenzie E (1985) Some simple models for discrete variate time series. JAWRA J Am W Resour Assoc 21(4):645–650
McKenzie E (1986) Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. Adv Appl Probab 18:679–705
Ristić MM, Bakouch HS, Nastić AS (2009) A new geometric first-order integer-valued autoregressive (NGINAR(1)) process. J Stat Plan Inference 139(7):2218–2226
Schweer S, Weiß CH (2014) Compound Poisson INAR(1) processes: stochastic properties and testing for overdispersion. Comput Stat Data Anal 77:267–284
Scotto MG, Weiß CH, Gouveia S (2015) Thinning-based models in the analysis of integer-valued time series: a review. Stat Model 15(6):590–618
Silva N, Pereira I, Silva ME (2009) Forecasting in INAR(1) model. REVSTAT-Stat J 7(1):119–134
Steutel F, Van Harn K (1979) Discrete analogues of self-decomposability and stability. Ann Probab 7:893–899
Weiß CH (2008) Thinning operations for modeling time series of counts-a survey. AStA Adv Stat Anal 92(3):319–341
Zeger SL (1988) A regression model for time series of counts. Biometrika 75(4):621–629