Zero-Inflated Count Time Series Regression Models

Mohammed Alqawba, Norou Diawara, and N. Rao Chaganty
Department of Mathematics and Statistics, Old Dominion University,
Norfolk, Virginia, USA

Abstract: Count time series data are frequent in many applied disciplines. In describing them, a specific count may reveal more often than usual. In faming a modeling approach, one must account for the excess count. In this paper, we develop a copula-based time series model for zero-inflated counts with the presence of covariates. Zero-inflated Poisson (ZIP), zero-inflated negative Binomial (ZINB), and zero-inflated Conway-Maxwell-Poisson (ZICMP) distributed marginals will be considered, while the joint distribution is modeled under Gaussian copula with autoregression moving average (ARMA) errors. Likelihood is formulated for inference, under sequential inference method. A simulated study is conducted, and a practical application in environmental setting is described.

Keywords: Conway-Maxwell-Poisson; Count time series; Gaussian copula; Negative binomial; Poisson; Sequential importance sampling; Zero-inflation.

1. INTRODUCTION

Zero-inflated counts time series are found in several fields such as environmental sciences, public health, and economics. For examples, monthly counts of sandstorms in some areas, rare diseases with low infection rates, and crimes such as arson. In these cases, the observed counts may include a considerable frequency of zeros. However, during certain seasons, these counts could take larger values. Additionally, these zero-inflated counts are usually autocorrelated when the data is collected over time. Standard time series models fail to account for such problems. Motivated by these problems, we propose and develop a class of time series models for zero-inflated counts with the presence of covariates using Gaussian copula.

Copulas are multivariate distributions with uniform margins on the unit interval. There are numerous copulas available, and one of the most popular copulas in the literature is the Gaussian copula. The Gaussian copula shares many of the properties of multivariate normal (Gaussian) distribution such as the correlation structure. Therefore, the flexibility to manipulate the association structure by using the Gaussian copula will be taken advantage of. The Gaussian copula function is given by

$$C(u_1, \dots, u_n) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)), \quad \forall u_i \in [0, 1], \tag{1.1}$$

where Φ^{-1} is the inverse CDF of a standard normal and Φ_R is the joint CDF of a standard multivariate normal distribution with covariate matrix equal to the positive definite correlation matrix R.

Address correspondence to N. Diawara, Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23520, USA; E-mail: ndiawara@odu.edu

To accommodate for the zero-inflation in the data, the zero-inflated Poisson (ZIP), zero-inflated negative Binomial (ZINB), and zero-inflated Conway-Maxwell-Poisson (ZICMP) distributions are chosen.

Suppose Y_t denotes a random count at time t with the probability mass function (pmf) and the cumulative distribution function (cdf) are given by f_t and F_t, respectively.

1. ZIP: ω_t zero-inflation parameter, λ_t intensity parameter, and
$$f_t(y_t) = \omega_t I_{\{y_t=0\}} + (1-\omega_t)\frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!},$$
where $I_{\{y_t=0\}}$ is the indicator function, $\omega_t \in [0,1]$, and $\lambda_t > 0$. If $\omega_t \to 0$, the baseline Poisson distribution is obtained.

2. ZINB: ω_t zero-inflation parameter, λ_t intensity parameter, κ_t dispersion parameter, and
$$f_t(y_t) = \omega_t I_{\{y_t=0\}} + (1-\omega_t)\frac{\Gamma(\kappa_t+y_t)}{\Gamma(\kappa_t)y_t!}\Big(\frac{\kappa_t}{\kappa_t+\lambda_t}\Big)^{\kappa_t}\Big(\frac{\lambda_t}{\kappa_t+\lambda_t}\Big)^{y_t},$$
where $I_{\{y_t=0\}}$ is the indicator function, $\omega_t \in [0,1]$, $\lambda_t > 0$, and $\kappa_t \geq 0$. If $\omega_t \to 0$, the baseline NB distribution is obtained.

3. ZICMP: ω_t zero-inflation parameter, λ_t intensity parameter, κ_t dispersion parameter, and
$$f_{Y_t}(y_t) = \omega_t I_{\{y_t=0\}} + (1-\omega_t)\frac{\lambda_t^{y_t}}{(y_t!)^{\kappa_t} Z(\lambda_t,\kappa_t)},$$
where $I_{\{y_t=0\}}$ is the indicator function, $\omega_t \in [0,1]$, $\lambda_t > 0$, and $\kappa_t \geq 0$. If $\omega_t \to 0$, the baseline CMP distribution is obtained, and if $\kappa_t = 1$, the ZIP distribution is obtained.

Covariates can be included through the marginal parameters via generalized linear models (GLM).

The rest of the paper is organized as follows. In Section 2, we describe the Gaussian copula zero-inflated regression models for the zero-inflated count time series. In Section 3, we describe the parameter estimation method applied via sequential importance sampling. Section 4 presents a simulation study and real data examples to illustrate the proposed models. We end the paper with a summary in Section 5.

2. REGRESSION MODELS FOR ZERO-INFLATED COUNT TIME SERIES

Using the ideas in Masarotto and Varin (2012) and Alqawba et al. (2019a), we construct a regression model for zero-inflated time series count data in the presence of covariates. Suppose that the errors ϵ_t for $t = 1, \ldots, n$ follow a stationary $ARMA(p,q)$ process, with Gaussian noise, η_t for $t = 1, \ldots, n$ that are independent and identically distributed normal random variables with variance σ_η^2. Then the error vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_t)'$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $R(\boldsymbol{\rho})$ where $\boldsymbol{\rho} = (\boldsymbol{\varphi}, \boldsymbol{\delta})$ is a function of the $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)'$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)'$, the autoregressive and moving average vector of parameters, respectively. As in Masarotto and Varin (2012), we make the assumption $\sigma_\eta^2 = h(\boldsymbol{\rho})$ so that $R(\boldsymbol{\rho})$ will be a correlation matrix, where $h(\boldsymbol{\rho})$ is a function of the dependence parameters. See page 9, Alqawba (2019) for an implicit form of the function $h(\boldsymbol{\rho})$.

As a special case, consider the process $ARMA(1,0)$ (or $AR(1)$). Then the process ϵ_t is governed by $\epsilon_t = \varphi\epsilon_{t-1} + \eta_t$. With the assumption $\sigma_\eta^2 = 1 - \varphi^2$, the correlation matrix takes the form $R(\boldsymbol{\rho}) = R(\boldsymbol{\varphi}) = \left[\varphi^{|i-j|}\right]$, which is known as autoregressive of order one. Note that the marginally ϵ_t is standard normal and the joint cdf of the vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_t)'$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $R(\boldsymbol{\rho})$. Thus the cdf of $\boldsymbol{\epsilon}$ is $\Phi_{R(\boldsymbol{\rho})}(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ and the induced copula is the Gaussian copula in (1.1), since $u_t = \Phi(\epsilon_t)$ is uniform on $[0,\ 1]$ for $t = 1, 2, \ldots, n$.

Let F_t be one of the cdfs of the ZIP, ZINB or the ZICMP distributions. A general regression model for the zero-inflated count Y_t is

$$Y_t = F_t^{-1}\{\Phi(\epsilon_t)|\mathbf{X}_t; \boldsymbol{\theta}\}, \quad \text{for } t = 1, \ldots, n, \tag{2.1}$$

where

$$F_t^{-1}(u) = \inf\{z \in \mathbb{R} : F_t(z) \geq u\}, \quad u \in (0,1)$$

is the generalized inverse (quantile function) of the cdf F_t. The vector $\mathbf{X}_t = (\mathbf{x}_t, \mathbf{z}_t, \mathbf{w}_t)'$ consists of covariates corresponding to the intensity (mean) parameter λ_t, the zero-inflation parameter ω_t and the dispersion parameter κ_t if needed, respectively. Notice that some of the covariates could be constant across time. The vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})'$ is the unknown regression parameter that needs to be estimated from the data.

We construct the model in (2.1) in such a way that ensures the zero-inflated count Y_t follows the desired distribution $F_t(.)$ by the integral transformation theorem. Such model appears in the literature under different names (see for examples, Masarotto and Varin, 2012, Jia et al., 2018, and Lennon and Yuan, 2019). Generally, the model falls under the class of nonlinear state-space model since the zero-inflated counts, $\{Y_t\}$ are assumed to be generated using a nonlinear function of the latent or state ARMA process, $\{\epsilon_t\}$.

Note that since the counts are zero-inflated, the probability that the count is zero affects the range of ϵ_t such that the range of u_t when $Y_t = 0$ is wider in comparison with $Y_t > 0$. In other words, the zero-inflation parameter ω_t affects the range of u_t when $Y_t = 0$ whereas the intensity parameter λ_t and the dispersion parameter κ_t (if existed) affect the ranges of u_t when $Y_t > 0$.

The joint distribution function of the zero-inflated count time series, Y_t, for $t = 1, \ldots, n$ follows the Gaussian copula given in (1.1), that is,

$$F(y_1, \ldots, y_n) = \Phi_{R(\boldsymbol{\rho})}\Big(\Phi^{-1}(F_1(y_1|\mathbf{X}_1; \boldsymbol{\theta})), \ldots, \Phi^{-1}(F_n(y_n|\mathbf{X}_n; \boldsymbol{\theta}))\Big), \tag{2.2}$$

and it holds only if (2.1) holds.

In a linear regression model with normal errors, the correlation of the responses, say Y_t and Y_s, agrees with the correlation of the corresponding errors, ϵ_t and ϵ_s for $t \neq s$. However, in our model the function, F^{-1}, is nonlinear, hence the correlation of Y_t and Y_s is not necessarily linear function of the correlation of ϵ_t and ϵ_s. Jia et al. (2018) studied the relationship between the autocorrelations of the two processes $\{Y_t\}$ and $\{\epsilon_t\}$ and defined a function that links the autocorrelations of the two processes $\{Y_t\}$ and $\{\epsilon_t\}$ using Hermite expansions.

3. PARAMETER ESTIMATION

We are interested in estimating the parameter vectors $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\rho})'$ using a maximum likelihood estimation (MLE) method. Based on the probability density function define in (2.2), the likelihood function is given by

$$L(\boldsymbol{\vartheta}; \mathbf{y}) \;=\; \Pr(Y_1 = y_1, \ldots, Y_n = y_n)$$

$$= \sum_{j_1=0}^{1} \cdots \sum_{j_n=0}^{1} (-1)^{j_1+\cdots+j_n} F(y_1 - j_1, \ldots, y_n - j_n), \qquad (3.1)$$

where $F(y_1, \ldots, y_n)$ for $j_t = 0, 1$ is given in (2.2), and can be expressed as

$$\Phi_{R(\boldsymbol{\rho})}(\mathcal{D}_1^+, \ldots, \mathcal{D}_n^+) = \int_{-\infty}^{\mathcal{D}_1^+} \cdots \int_{-\infty}^{\mathcal{D}_n^+} \phi_{R(\boldsymbol{\rho})}(\epsilon_1, \ldots, \epsilon_n) d\epsilon_1 \ldots d\epsilon_n, \qquad (3.2)$$

where $\mathcal{D}_t^+ = \Phi^{-1}\{F_t(y_t|\mathbf{X}_t; \boldsymbol{\theta})\}$. Therefore, maximizing (3.1) requires the evaluation of 2^n multivariate distribution functions, and with time series data usually n is quite large so the number of functions will be astronomically large and almost impossible to be optimized. In addition, straightforward optimization methods of the likelihood function are not available yet due to the many-to-one mapping given in (2.1). In addition, calculating the finite difference in (3.1) numerically might result in negative values when the dimension is large (Nikoloulopoulos, 2016).

However, for some cases where the copula functions do not have a closed form, the probability density function can be evaluated by integration over a rectangle (Panagiotelis et al., 2012). In fact, for the Gaussian copula with discrete margins, the likelihood function is given by the following n-dimensional rectangular integral

$$\begin{aligned} L(\boldsymbol{\vartheta}; \mathbf{y}) &= \Pr(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \int\limits_{\mathcal{D}_1(y_1;\boldsymbol{\theta})} \cdots \int\limits_{\mathcal{D}_n(y_n;\boldsymbol{\theta})} \phi_{R(\boldsymbol{\rho})}(\epsilon_1, \ldots, \epsilon_n) d\epsilon_1 \ldots d\epsilon_n, \qquad (3.3) \end{aligned}$$

where

$$\mathcal{D}_t(y_t; \boldsymbol{\theta}) = [\Phi^{-1}\{F_t(y_t^-|\mathbf{X}_t; \boldsymbol{\theta})\}, \Phi^{-1}\{F_t(y_t|\mathbf{X}_t; \boldsymbol{\theta})\}] \qquad (3.4)$$

for $t = 1, \ldots, n$ and $\phi_{R(\boldsymbol{\rho})}(.)$ is the probability density function of an n-dimensional normal distribution with zero mean vector and a variance covariance matrix given by $R(\boldsymbol{\rho})$. For small n, notable works have been done on precisely approximating the normal integral given in (3.3) (see for examples, Joe 1995 and Genz 1992). However, for large n, as of the case for time series data, evaluating the likelihood function using these deterministic approximations is computationally intensive and is inefficient especially when the number of covariates is large. Masarotto and Varin (2012) argued that applying simple Monte Carlo approximations of the likelihood given in (3.3) used in importance sampling (IS) are quite inefficient. However, they suggested sequential importance sampling method inspired by the popular Geweke-Hajivassiliou-Keane (GHK) algorithm (Geweke, 1991; Hajivassiliou et al., 1996; Keane, 1994) which was proven to be quite efficient in approximating the multivariate probability integral given in (3.3). They assumed sampling from the following truncated normal density given by

$$f_t(\epsilon_t|y_t, \epsilon_{t-1}, \ldots, \epsilon_1; \boldsymbol{\rho}), \quad t = 1, \ldots, n \qquad (3.5)$$

as a replacement of the difficult to control, $f_t(\epsilon_t|y_t, y_{t-1}, \ldots, y_1; \boldsymbol{\rho})$ over the interval given in (3.4). In addition, since we assume that the joint distribution of the errors is multivariate normal distribution with variance covariance matrix $R(\boldsymbol{\rho})$, the conditional density $\phi(\epsilon_t|\epsilon_{t-1}, \ldots, \epsilon_1; \boldsymbol{\rho})$ is of univariate normal distribution with mean $m_t = \mathrm{E}(\epsilon_t|\epsilon_{t-1}, \ldots, \epsilon_1)$ and variance $v_t^2 = \mathrm{Var}(\epsilon_t|\epsilon_{t-1}, \ldots, \epsilon_1)$, for $t = 1, \ldots, n$.

The quantities m_t and v_t^2 can be efficiently obtained through the Cholesky decomposition of $R(\boldsymbol{\rho})$. Therefore, the conditional density given in (3.5) is of a truncated normal distribution over the interval given in (3.4), and a random sample can be obtained setting

$$\epsilon_t = \epsilon_t(u_t) = m_t + v_t\Phi^{-1}\{(1-u_t)a_t + u_t b_t\}, \quad t = 1, \dots, n, \tag{3.6}$$

where $u_1, \dots, u_n$ are n i.i.d. uniform random variable on the unit interval $(0,1)$, and

$$a_t = \Phi\left[\frac{\Phi^{-1}\{F_t(y_t^-|\mathbf{X}_t;\boldsymbol{\theta})\} - m_t}{v_t}\right], \qquad b_t = \Phi\left[\frac{\Phi^{-1}\{F_t(y_t|\mathbf{X}_t;\boldsymbol{\theta})\} - m_t}{v_t}\right],$$

for $t = 1, \dots, n$.

The likelihood function is then approximated by the following sequential sampler algorithm.

1. For $k = 1, \dots, K$,
   - (a) generate n independent uniform$(0,1)$ random variables, $u_1^{(k)}, \dots, u_n^{(k)}$;
   - (b) compute the randomized errors $\epsilon_t^{(k)} = \epsilon_t(u_t^{(k)})$ using (3.6);
2. estimate the likelihood by:

$$\widehat{L}(\boldsymbol{\vartheta};\mathbf{y}) = \frac{1}{K}\sum_{k=1}^{K}\left\{\prod_{t=1}^{n}\frac{\phi(\epsilon_t^{(k)}|\epsilon_{t-1}^{(k)}, \dots, \epsilon_1^{(k)};\boldsymbol{\vartheta})}{f_t(\epsilon_t^{(k)}|y_t, \epsilon_{t-1}^{(k)}, \dots, \epsilon_1^{(k)};\boldsymbol{\vartheta})}\right\}, \tag{3.7}$$

where K denotes the number of replication. Börsch-Supan and Hajivassiliou (1993) showed that $\widehat{L}(\boldsymbol{\vartheta};\mathbf{y})$ is an unbiased estimator of $L(\boldsymbol{\vartheta};\mathbf{y})$.

Thus, the maximum likelihood estimate of $\boldsymbol{\vartheta}$ can be obtained by:

$$\widehat{\boldsymbol{\vartheta}} = \underset{\boldsymbol{\vartheta}}{\arg\max}\ \widehat{L}(\boldsymbol{\vartheta};\mathbf{y}). \tag{3.8}$$

This optimization will yield a Hessian matrix that can be inverted to obtain standard errors for the model parameters.

4. DATA ANALYSIS

4.1. Simulation

To evaluate the performance of the proposed method, we performed a comprehensive simulation study in R (R Core Team, 2013). Based on the proposed model given in (2.1), the simulation process is given as follows.
1) simulate $\boldsymbol{\epsilon} \sim \Phi_{R(\boldsymbol{\rho})}(\epsilon_1, \dots, \epsilon_n)$, 2) compute $\mathbf{U} = (\Phi(\epsilon_1), \dots, \Phi(\epsilon_n))$, 3) compute $\mathbf{Y} = (F_1^{-1}\{U_1|\mathbf{X}_1;\boldsymbol{\theta}\}, \dots, F_n^{-1}\{U_n|\mathbf{X}_n;\boldsymbol{\theta}\})$, where $\mathbf{X_t}$ is the set of covariates for $t = 1, \dots, n$. Due to the space limitation and the computational requirements of the estimation algorithm, we summarize the study results by considering three models and each assumes one of the zero-inflated distributions presented in the introduction. We consider the $MA(1)$ dependence structure. No covariates are considered, so the intensity parameter λ, zero-inflated parameter ω, and the dispersion parameter κ (if existed) are constant across time. The dependence parameter of the latent $MA(1)$ process is chosen to be $\delta = 0.5$ across all three marginals. The marginal parameters are then given by ZIP with $\lambda = 4.3$ and $\omega = 0.25$; ZINB with $\lambda = 4.3$, $\omega = 0.25$ and $\kappa = 0.5$; and ZICMP with $\lambda = 3$, $\omega = 0.2$ and $\kappa = 0.25$.

Table 1 shows a summary of the simulation results for the ZIP, ZINB, and ZICMP models with with stationary $MA(1)$ errors. The summary shows that the proposed estimation method performs well with the latent process $\{\epsilon_t\}$ following MA(1) process.

Table 1. Mean of estimates, MADEs (within parentheses) for zero-inflated models with MA(1) dependence structure.

Model	n	λ	ω	κ	δ
ZIP	100	4.3223(0.1903)	0.2533(0.0347)		0.5167(0.0869)
	200	4.3205(0.1290)	0.2521(0.0272)		0.5038(0.0578)
	500	4.3088(0.0933)	0.2514(0.0176)		0.4961(0.0368)
ZINB	100	4.4940(0.8777)	0.2462(0.1281)	0.6227(0.2387)	0.5289(0.1102)
	200	4.3317(0.7072)	0.2293(0.1099)	0.5556(0.1712)	0.4958(0.0661)
	500	4.2904(0.4367)	0.2413(0.0721)	0.5305(0.1041)	0.4955(0.0456)
ZICMP	100	3.2587(0.5566)	0.3421(0.1421)	0.2585(0.0491)	0.5119(0.0842)
	200	3.1520(0.3189)	0.3400(0.1400)	0.2544(0.0375)	0.4992(0.0554)
	500	3.1229(0.2183)	0.3397(0.1397)	0.2549(0.0241)	0.4978(0.0366)

4.2. Real Data Example

The data set used in this example consists of the monthly count of strong sandstorms recorded by the AQI airport station in Eastern Province, Saudi Arabia, which was originally studied in Alqawba et al. (2019b). The station happens to be is located in one of the major dust producing regions in the world (Idso, 1976). Sandstorm is a weather event that results from strong wind releasing dust from the ground and transfers it long distances (Goudie and Middleton, 2006). Sandstorms can cause many environmental and human-related hazards. For examples, sandstorms impact the air quality, disturb daily activities, and transportations. Hence, studying and accurately analyzing the behavior of these phenomena is important to successfully forecast such events.

The monthly counts studied here are characterized as strong sandstorms by the AQI airport station. Tao et al. (2002) stated that a strong sandstorm reduces the level of visibility to less than 500 meters and with average wind speed of 17.2 to 24.4 meters/seconds. The counts of these events contain zero inflation. Several works have been applied on handling rare events such as strong sandstorms (see for examples Tan et al., 2014 and Ho and Bhaduri, 2015). Here, we apply the proposed zero-inflated count time series regression models using Gaussian copula.

The data set consists of 348 monthly counts of strong sandstorms, starting from January 1978 to December 2013. The main objective was to apply the proposed models and investigate if there were any significant seasonal and trend components. Additionally, we investigated if there were any other predictors that affected the frequency of sandstorms such as the monthly counts of dust haze events, maximum wind speed, temperature, and relative humidity.

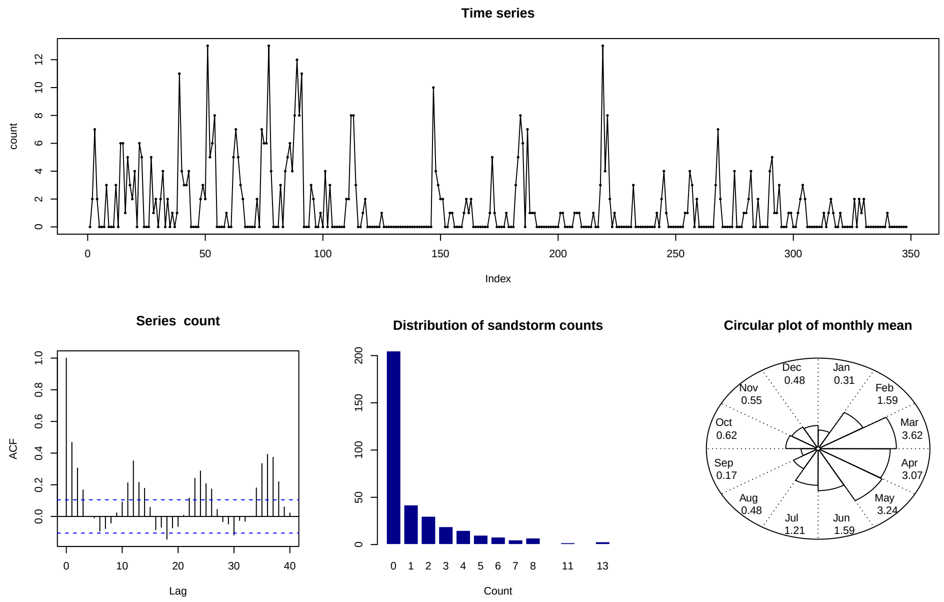

Figure 1. Time series plot of monthly count of sandstorms, the autocorrelation function, bar-plot of distribution of sandstorm counts, and circular plot of the monthly mean count of sandstorms.

Figure 1 shows the sandstorms series plot, the autocorrelation function, bar-plot of the distribution of sandstorm counts, and circular plot of the monthly mean count of sandstorms. From the time series plot and the bar-plot, we could see that the distribution of the sandstorm counts had more zeros relative to a Poisson distribution with the same empirical mean. These zeros represented about 59% of the sample. Decreasing trend could also be observed from the time series plot. Additionally, seasonality was also captured from the autocorrelation function and circular plot. In fact, from the circular plot, we concluded that most sandstorms occurred during spring time, i.e. March, April, and May months. Thus, trend and seasonal covariates were added to the models.

Hence, we fit several models to investigate the trend and seasonality effects along with the other covariates mentioned above. After performing model selection based on AIC, we ended up with the following models taking the form of (2.1), with the log-linear function of the intensity parameter given by

$$\log(\lambda_t) = \beta_0 + \beta_1 \left(t \times 10^{-3}\right) + \beta_2 x_{1t} + \beta_3 x_{2t} + \beta_4 x_{3t},$$

and the logit function for the zero-inflation parameter given by

$$\text{logit}(\omega_t) = \gamma_0 + \gamma_1 z_{1t} + \gamma_2 z_{2t} + \gamma_3 z_{3t},$$

for $t = 1, \ldots, n$, where $x_{1t} = z_{1t} = \cos\left(\frac{2\pi t}{12}\right)$, $x_{2t} = z_{2t} = \sin\left(\frac{2\pi t}{12}\right)$, and $x_{3t} = z_{3t}$ is the monthly count of dust haze events. The log-function of the dispersion parameter (if existed) is given by $\log(\kappa) = \alpha$, i.e. it was chosen to be constant across time. Thus, the main model was given by

$$Y_t = F_t^{-1}\{\Phi(\epsilon_t)|\mathbf{X}_t; \boldsymbol{\theta}\}, \quad \text{for } t = 1, \ldots, 348,$$

where $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_4, \gamma_0, \ldots, \gamma_3, \alpha)'$ and $\mathbf{X}_t = (\mathbf{x}_t', \mathbf{z}_t', \mathrm{w}_t)'$, in which the intensity covariates were $\mathbf{x}_t = (t \times 10^{-3}, x_{1t}, x_{2t}, x_{3t})'$, the zero-inflation covariates were $\mathbf{z}_t = (z_{1t}, z_{2t}, z_{3t})'$, and no covariates with the dispersion effect, i.e. $\mathrm{w}_t = 1$ for $t = 1, \ldots, n$. The latent random process, the errors, were generally given by the

$ARMA(p, q)$ process. However, after fitting multiple models, we considered the dependence structure that followed $AR(1)$ autocorrelation.

Table 2 shows the three copula-based zero-inflated models we proposed in this paper along with the copula-based Poisson and NB models introduced in Masarotto and Varin (2012), all with the AR(1) correlation structure. The results of all models are comparable. However, the Poisson and NB model seem to perform moderately less than the other models because they fail to account for the overdispersion in the counts caused by the zero inflation and the zero inflation itself. On the other hand, adding more probability to the event zero improves the performance of the fitted model because it addresses the problem of zero inflation and over dispersion. This is why the ZIP, ZINB, and ZICMP models are better fit than the ordinary Poisson and NB distributions in this application.

Table 2. Parameter estimates (standard errors) for the copula-based models fit to the sandstorms count series.

Parameter	ZIP	ZINB	ZICMP	Poisson	NB
β_0	0.9977(0.1175)	0.9709(0.1570)	0.7978(0.1888)	0.2003(0.1147)	0.2996(0.1965)
β_1	-4.1493(0.6065)	-4.7477(0.7976)	-2.4523(0.5772)	-5.1453(0.5517)	-5.8397(0.9643)
β_2	-0.2004(0.0885)	-0.1813(0.1243)	-0.1089(0.0723)	-0.4634(0.0814)	-0.4385(0.1391)
β_3	0.3461(0.0938)	0.4231(0.1239)	0.2093(0.0786)	0.7879(0.0888)	0.7751(0.1352)
β_4	0.0627(0.0088)	0.0645(0.0123)	0.0435(0.0094)	0.0974(0.0085)	0.0950(0.0163)
γ_0	0.7647(0.2622)	0.5656(0.3047)	0.6629(0.2119)		
γ_1	0.6163(0.2460)	0.6648(0.2925)	-1.0047(0.2132)		
γ_2	-0.8931(0.2401)	-0.8363(0.2736)	-0.1496(0.0344)		
γ_3	-0.1489(0.0424)	-0.1659(0.0524)	-0.2466(0.1613)		
α		0.6400(0.2437)	1.1733(0.2230)		0.9195(0.2009)
φ	0.2580(0.0623)	0.2503(0.0724)	0.2870(0.078)	0.1539(0.0419)	0.2488(0.0740)
AIC	910.9	**895.62**	905.6	1017.3	923.06

Furthermore, Table 2 shows that the zero-inflated models are capable of accounting for first order autocorrelations. The autocorrelation coefficients, $\widehat{\varphi}$'s, are similar across models although the zero-inflation models suggest stronger autocorrelation among the observations. For the marginal parameters, $\boldsymbol{\theta}$, the estimates are quite similar between the ZIP and ZINB, and slightly different from the ZICMP. All models suggest significant decreasing trend in the number of strong sandstorms since $\beta_1 < 0$. Seasonality also significant at annual frequencies since $\beta_2, \beta_3, \gamma_1$ and γ_2 are significantly different from zero. Finally, the affect of dust haze is significant since both β_4 and γ_3 are significantly different from zero.

Figure 2 shows the randomized quantile residuals in normal probability and autocorrelation plots of the copula-based ZIP, ZINB and ZICMP models. The normal probability plots suggest the randomized quantile residuals of these three models follow the normal distribution, and the autocorrelation plots indicate the absence of the serial dependence in the residuals. These findings suggest that the proposed models in this paper fit the data adequately. Models with more complicated correlation structures such as $AR(2)$ and $ARMA(1, 1)$ were also considered and fitted to the data with the same covariates. No significant improvements were found and thus we recommend using $AR(1)$. However, dropping the trend and seasonality covariates and running the models with only the dust haze covariate yields significant $AR(2)$ and $ARMA(1, 1)$ dependence structures. Figure 3 shows the predicted values of the sandstorm counts from the three proposed models. The predicted values were calculated using the conditional expectation of Y_t given the past $Y_{t-1} = y_{t-1}, \ldots, Y_1 = y_1$. The plots indicate that our copula-based zero-inflated models adequately predict the injury counts.

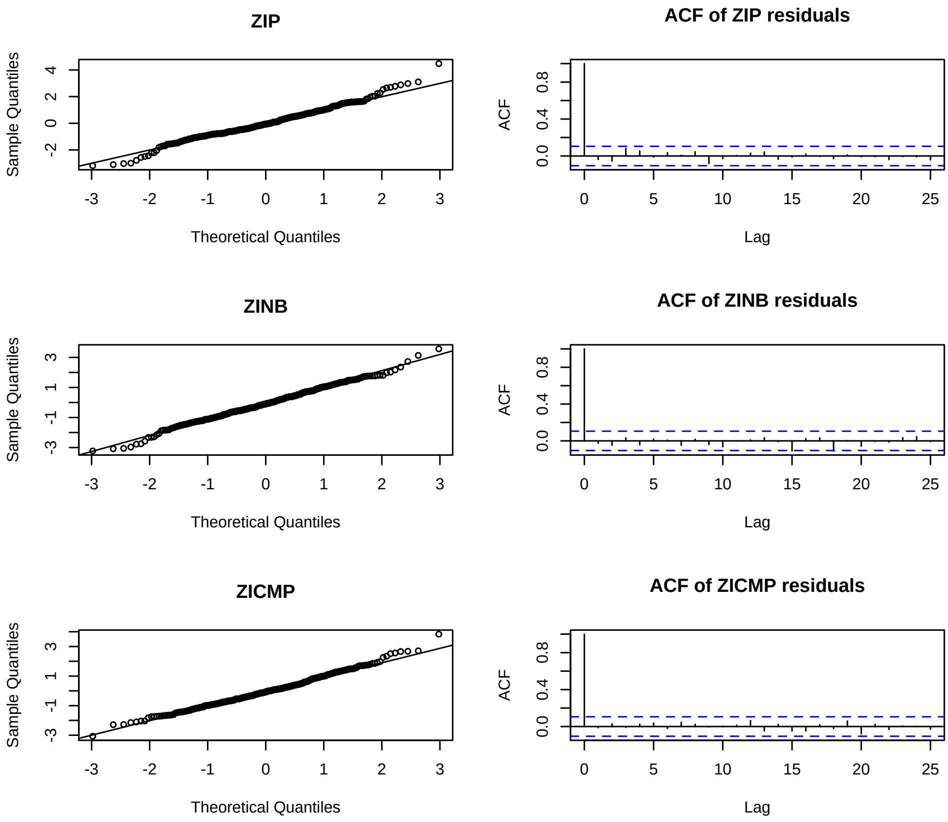

Figure 2. Sandstorm counts series: q-q plots (left) and autocorrelation plots (right) for sets of randomized residuals of the ZIP, ZINB and ZICMP models.

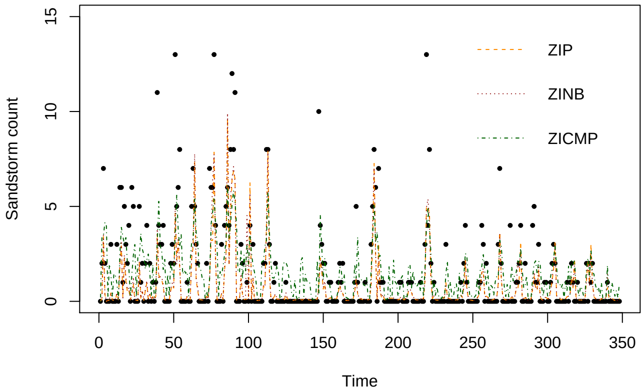

Figure 3. Prediction plot using the conditional expectations of the ZIP, ZINB and ZICMP models. Dots represent the observed sandstorm counts

5. SUMMARY

Zero-inflated count time series data are found in different areas. Applying ordinary Poisson and NB distributions to these time series of counts might not be appropriate due to the frequent occurrence of zeros. In this paper, we have extended the work done by Masarotto and Varin (2012) to include a class of models that accounts for zero inflation. The marginals are assumed to follow one of the ZIP, ZINB, and ZICMP distributions, and the serial dependence was modeled by a Gaussian copula with correlation matrix that of a stationary ARMA process. Likelihood inference was carried out using sequential importance sampling. Simulated studies were conducted to evaluate the parameter estimation procedures. Model assessment to check the goodness of fit for the proposed models was done via residual analysis. The proposed models were applied to the sandstorm data, and according to the residual analysis the models fit the data adequately, but both ZINB and ZICMP seem to have a slight advantage over ZIP distribution. Future direction is to consider different model construction methods from the marginal regression such as Markov models to handle zero-inflated count time series data.

REFERENCES

Alqawba, M. S., 2019. *Copula-Based Zero-Inflated Count Time Series Models*. Doctor of Philosophy (PhD), dissertation, Mathematics and Statistics, Old Dominion University,.

Alqawba, M., N. Diawara, and N. R. Chaganty, 2019a. Zero-inflated count time series models using gaussian copula. *Sequential Analysis* 38 (3):342–357.

Alqawba, M., N. Diawara, and J.-M. Kim, 2019b. Copula directional dependence of discrete time series marginals. *Communications in Statistics - Simulation and Computation* 1–18.

Börsch-Supan, A., and V. A. Hajivassiliou, 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of econometrics* 58 (3):347–368.

Genz, A., 1992. Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics* 1 (2):141–149.

Geweke, J., 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities.

Goudie, A. S., and N. J. Middleton, 2006. *Desert dust in the global system*. Springer Science & Business Media.

Hajivassiliou, V., D. McFadden, and P. Ruud, 1996. Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of econometrics* 72 (1-2):85–134.

Ho, C.-H., and M. Bhaduri, 2015. On a novel approach to forecast sparse rare events: applications to parkfield earthquake prediction. *Natural Hazards* 78 (1):669–679.

Idso, S. B., 1976. Dust storms. *Scientific American* 235 (4):108–115.

Jia, Y., S. Kechagias, J. Livsey, R. Lund, and V. Pipiras, 2018. Latent gaussian count time series modeling. *arXiv preprint arXiv:181100203* .

Joe, H., 1995. Approximations to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association* 90 (431):957–964.

Keane, M. P., 1994. A computationally practical simulation estimator for panel data. *Econometrica: Journal of the Econometric Society* 95–116.

Lennon, H., and J. Yuan, 2019. Estimation of a digitised gaussian arma model by monte carlo expectation maximisation. *Computational Statistics & Data Analysis* 133:277–284.

Masarotto, G., and C. Varin, 2012. Gaussian copula marginal regression. *Electronic Journal of Statistics* 6:1517–1549.

Nikoloulopoulos, A. K., 2016. Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic environmental research and risk assessment* 30 (2):493–505.

Panagiotelis, A., C. Czado, and H. Joe, 2012. Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association* 107 (499):1063–1072.

R Core Team, 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tan, S., M. Bhaduri, and C.-H. Ho, 2014. A statistical model for long-term forecasts of strong sand dust storms. *Journal of Geoscience and Environment Protection* 2 (03):16.

Tao, G., L. Jingtao, Y. Xiao, K. Ling, F. Yida, and H. Yinghua, 2002. Objective pattern discrimination model for dust storm forecasting. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling* 9 (1):55–62.