ISSN: 2317-0840

Modeling count time series: a comparative case study

Gisele O. Maia[†1], Glaura C. Franco[1], Thiago R. Santos[1], Ana Júlia A. Câmara[2].

[1] *Departamento de Estatística, Universidade Federal de Minas Gerais (UFMG).*
[2] *Departamento de Estatística, Universidade Federal do Espírito Santo (UFES).*

Abstract: *This paper presents an application for counting data, where the observation-driven and parameter-driven models are compared. To this purpose, the Generalized Additive Autoregressive Moving Average (GAM-ARMA) and Non-Gaussian State Space with Exact Marginal Likelihood (NGSSEML) models are used. Model parameters are estimated using the maximum likelihood method. The ability of the procedure to model and forecast real data is presented for the number of chronic obstructive disease (COPD) cases.*

Keywords: *Observation-driven model; Parameter-driven model; GAM-ARMA; NGSSEML; count data.*

Modelagem de séries temporais de contagem: um estudo comparativo de caso

Resumo: *Esse artigo apresenta uma aplicação a dados de contagem, onde os modelos observation-driven e parameter-driven são comparados. Com esse propósito, os modelos Generalizado Aditivo Autorregressivo Média Móvel (GAM-ARMA) e Espaço de Estados Não-Gaussiano com Verossimilhança Marginal Exata (NGSSEML) são utilizados. Parâmetros dos modelos são estimados utilizando o método de máxima verossimilhança. A capacidade do procedimento de modelar e prever dados reais é apresentada para o número de casos de doença obstrutiva crônica (COPD).*

Palavras-chave: *Modelo Observation-driven; GAM-ARMA; NGSSEML; Modelo Parameter-driven; Dados de contagem.*

Introduction

Cox (1981) classified two classes of models for Non-Gaussian time-dependent data: observation-driven model and parameter-driven model. The main difference between the two models is the way the dependence structure is incorporated into the model. Let $\{Y_t\}_{t\in\mathbb{N}}$ be a time series and $\mathcal{F}_t \equiv \sigma\{Y_{t-1}, Y_{t-2}, \ldots\}$ the past observation. In the observation-driven model, some conditional distributions for Y_t given $\mathcal{F}_t$ are assumed and current parameters are deterministic functions of lagged dependent variables as well as contemporaneous and lagged exogenous variables.

Pioneering work is due to Zeger e Qaqish (1988), who proposed a model for continuous and counting time series where only the first and second conditional moments are specified. In Benjamin, Rigby e Stasinopoulos (2003), the authors introduced the Generalized Autoregressive Moving Average (GARMA) model, where the conditional distribution belongs to the exponential family. Albarracin, Alencar e Ho (2019) proposed the GAR-M model to reduce the problem of multicollinearity in the GARMA model. Davis e Liu (2012) presented a class of models where the conditional distribution belongs to the uniparametric exponential family. Rocha e Cribari-Neto (2009) proposed the Beta Autoregressive Moving Average (βARMA) model for time series that assume values in the $(0, 1)$ interval. Melo e Alencar (2020) proposed the CMP-ARMA model, that allows the modeling of underdispersed, equidispersed, and overdispersed data.

[†]Autora correspondente: ⟨giseleemaia07@gmail.com⟩.

Another important work in the class of observation-driven models is Davis, Dunsmuir e Streett (2003), where the GLARMA model was introduced to model count time series. The authors assume that the observations, conditioned on the past information, follow a Poisson distribution. Camara et al. (2021) proposed the GAM-ARMA model, which combines the structure of the Generalized Additive Model (GAM) (HASTIE; TIBSHIRANI, 1990) with the GLARMA models of Davis, Dunsmuir e Streett (2003) for count time series, incorporating the possibility of including auxiliary variables that possess a non-linear relationship with the response variable.

In the parameter-driven class, $\{Y_t\}_{t\in\mathbb{N}}$ given a latent process $\{\alpha_t\}_{t\in\mathbb{N}}$ is assumed to be conditionally independent and parameters vary over time as dynamic processes, assuming a particular probability distribution. The first parameter-driven model was proposed by Zeger (1988) for a time series of counts. Davis, Dunsmuir e Wang (2000) also proposed a model for the time of counts, which is driven by a Gaussian latent factor. Davis e Wu (2009) extended the work of Davis, Dunsmuir e Wang (2000) to data with overdispersion, using the Negative Binomial distribution. Maia et al. (2021) proposed the modeling of a wide class of time series, including nonnegative, count, bounded, binary, and real-valued.

Finally, Gamerman, Santos e Franco (2013) introduced a non-Gaussian State Space model for non-Gaussian time series and reliability analysis, with exact marginal likelihood, named NGSSEML. The proposed approach has the characteristic of being simple and flexible, and results are obtained using the exact marginal likelihood, which is not achieved in other non-Gaussian state space models. In addition, several distributions belong to the class of NGSSEML models, such as Poisson, Gamma, Weibull, Laplace, and Normal, among others.

Comparative studies between the observation-driven and parameter-driven classes can be found in the works of Davis, Dunsmuir e Wang (1999), where review of existing models for the two classes is performed; Jung e Tremayne (2011) and Jung, Kukuk e Liesenfeld (2006) compare the two classes based on real data applications and Franco, Migon e Prates (2015) perform a study comparing the GLARMA and state space models for count data, using Bayesian and frequentest approaches. In this direction, the purpose of the paper is to give some contribution in the comparison between the two classes, focusing on time series of count. In the observation-driven class, we adopt the GAM-ARMA model, while the NGSSMEL model is employed in the parameter-driven class.

In this work, the comparison between the two classes is based on real data application. In this application, we adjust the two models to the data and calculate measures to evaluate the performance of each model. The same assessment is made for predicting future observations, considering the approach of each model.

The paper is organized as follows. In Section Methodology, we present the two models that will be compared. In section Parameter Estimation, likelihood inference is developed to obtain parameter estimates. In Section Forecast, we discuss forecasting procedures based on the two models. In Section Real Data Application, we analyze the data set, namely, the monthly number of chronic obstructive pulmonary (COPD) cases. Concluding remarks are addressed in Section Conclusion.

Methodology

In this section, we present the basic properties of the GAM-ARMA and NGSSEML models that will be essential for characterizing these models. First, we present the GAM-ARMA, a generalized additive model for count time series proposed by Camara et al. (2021). Let $\{Y_t\}_{t\in\mathbb{N}}$ be a count time series, and considering the past observations $\mathcal{F}_{t-1} = \sigma(Y_s, s \leq t-1)$, define

$$Y_t|\mathcal{F}_{t-1} \sim Poisson(\mu_t), \tag{1}$$

where μ_t is the mean of the conditional distribution. The linear predictor, according to the definition of the GAM-ARMA model in Camara et al. (2021), is given by

$$W_t = \log(\mu_t) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{r} s_j(w_{t,j}) + Z_t, \tag{2}$$

where $(x_1, \ldots, x_k)$ represent the k explanatory variables linearly related to the response variable, $(\beta_0, \ldots, \beta_k)$ the vector of coefficients of the linear component, $(w_1, \ldots, w_r)$, represent the r explanatory variables related in a non-linear way with the response variable through the smooth curves $(s_1, \ldots, s_r)$ and Z_t has the same structure defined by the GLARMA model (DAVIS; DUNSMUIR; STREETT, 2003), that is,

$$Z_t = \phi_1(Z_{t-1} + e_{t-1}) + \cdots + \phi_p(Z_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}, \tag{3}$$

where $(\phi_1, \ldots, \phi_p)$ and $(\theta_1, \ldots, \theta_q)$ are vector parameters and $e_t = \frac{Y_t - \mu_t}{\mu_t^{\lambda}}$, where $\lambda \in (0, 1]$. The parameter λ is treated as a constant instead of estimating it.

The curves are estimated using the methodology proposed by (BOOR, 1978), where B-spline curves are constructed from polynomial parts joined to certain point values (knots). Consider m (positive integer) B-splines $\{B_{j,d}\}_{j=1}^{m}$ depending on knots $t_j, \ldots, t_{j+1+d}$, where d is the order of the polynomial pieces. The linear combination of B-splines is denoted as spline function and is given by

$$s = \sum_{j=1}^{m} \alpha_j B_j(w_t), \tag{4}$$

where $\{\alpha_j\}_{j=1}^{m}$ are coefficients of the B-splines. The B-splines will be a combination of third-degree polynomials, thus $d = 3$. The choice of the number and position of the knots is a subject of debate in several works. In this work, knots are chosen according to a pre-established interval and are equidistant. However, other selection methods can be found in Eilers e Marx (1996), Green e Silverman (1994), Friedman e Silverman (1989), Kooperberg e Stone (1991), and Harrel (2004). The Akaike's information criterion (AIC) can also be used in choosing the number of knots, as performed in Camara et al. (2021).

Stationarity results are difficult to obtain in observation-driven models. Davis, Dunsmuir e Streett (2003) show that stationarity results can be obtained for W_t, when Z_t is a moving average process, λ parameter assumes values within the interval $[1/2, 1]$ and model without covariates. For the GAM-ARMA model, these specific cases remain valid, but we can show that e_t is stationary and W_t is homogeneous non-stationary. For more details, see Camara et al. (2021) and Davis, Dunsmuir e Streett (2003).

After a brief presentation of the GAM-ARMA model, we begin the presentation of the Non-Gaussian State-Space with Exact Marginal Likelihood model (NGSSEML) proposed by Gamerman, Santos e Franco (2013). Let $\{y_t\}_{t \in \mathbb{N}}$ denote a time series. If the following assumptions are satisfied, then the time series belongs to the NGSSEML class of models:

Assumption 1 *The probability function can be written in the following form:*

$$p(y_t|\mu_t, Y_{t-1}, \psi) = a(y_t, \psi)\mu_t^{b(y_t,\psi)} \exp(-\mu_t c(y_t, \psi)). \tag{5}$$

Several distributions belong to the NGSSEML class as Poisson, Gamma, Normal, Weibull, and Pareto, among many others. However, in this work, we focus on the Poisson distribution. Then, $y_t \in \{0, 1, \ldots\}$, $Y_t = \{Y_0, y_1, \ldots, y_t\}$, for $t = 1, 2, \ldots,$, where Y_0 represents previously available information, $a(y_t, \psi) = (y_t!)^{-1}$, $b(y_t, \psi) = y_t$, $c(y_t, \psi) = 1$, $\mu_t > 0$, for all t, and $\psi = (w, \beta)^\top$.

Assumption 2 *If x_t is a covariate vector, then we have the relation $\mu_t = \lambda_t g(x_t, \beta)$, where g is the link function, β are the regression coefficients and λ_t is the latent state variable to the description of the dynamic level.*

Assumption 3 *The following distribution is defined:*

$$w\frac{\lambda_{t+1}}{\lambda_t}|\lambda_t, Y_t, \psi \sim Beta(wa_t, (1-w)a_t),$$

where $0 < w \leq 1$ and a_t will be specified later.

Assumption 4 *The prior distribution $\lambda_0|Y_0 \sim Gamma(a_0, b_0)$ is used for the dynamic level λ_t.*

If the Equation 5 is satisfied, the following results are obtained:

- The prior distribution $\lambda_t|Y_{t-1}$ follows a $Gamma(a_{t|t-1}, b_{t|t-1})$, where $a_{t|t-1} = wa_{t-1}$ and $b_{t|t-1} = wb_{t-1}$.
- $\mu_t = \lambda_t g(x_t, \beta)|Y_{t-1} \sim Gamma(a^*_{t|t-1}, ab^*_{t|t-1})$, where $a^*_{t|t-1} = wa_{t-1}$ and $b^*_{t|t-1} = wb_{t-1}[g(x_t^\top \beta)]^{-1}$.
- The posterior distribution of $\mu_t|\mathbf{Y}_t$ is $Gamma(a^*_t, b^*_t)$, where $a^*_t = a^*_{t|t-1} + b(y_t, \psi)$ and $b^*_t = b^*_{t|t-1} + c(y_t, \psi)$.
- $\lambda = \mu_t \left[g(x_t^\top \beta)\right]^{-1} |\mathbf{Y}_t \sim Gamma(a_t, b_t)$, where $a_t = a_{t|t-1} + b(y_t, \psi)$ and $b_t = b_{t|t-1} + c(y_t, \psi)g(x_t, \beta)$.
- The one-step-ahead predictive density function is given by

  $$p(y_t|Y_{t-1}, \psi) = \frac{\Gamma(b(y_t, \psi) + a_{t|t-1})a(y_t, \psi)(b_{t|t-1})^{a_{t|t-1}}}{\Gamma(a_{t|t-1})\left[c(y_t, \psi) + b_{t|t-1}\right]^{b(y_t,\psi)+a_{t|t-1}}},$$

  where $\Gamma(\cdot)$ is a gamma function and $t \leq n$, where n is the size of the time series.

Further details on results are provided in Gamerman, Santos e Franco (2013).

Parameter Estimation

In this section, we briefly discuss the approaches used by GAM-ARMA and NGSSEML models to obtain parameter estimates. Combining (2) and (4), we can write W_t as follows

$$W_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{i=1}^{r}\sum_{j=1}^{m} \alpha_{i,j} B_j(w_{t,i}) + Z_t. \tag{6}$$

Thus, the parameter vector of the GAM-ARMA model is defined by

$$\delta = (\beta_0, \ldots, \beta_k, \alpha_{1,1}, \ldots, \alpha_{r,m}, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q).$$

Define $l_n = \log f(Y_t|\mathcal{F}_{t-1})$, where the conditional density f of Y_t given $\mathcal{F}_{t-1}$ is Poisson. Therefore, we can write the following log-likelihood function

$$L_n(\delta) = \sum_{t=1}^{n}(y_t W_t(\delta) - e^{W_t(\delta)}), \tag{7}$$

where W_t is given by (6). All parameters will be estimated together through the likelihood function. The Newton-Raphson optimization algorithm is used to obtain the numerical solution initialized with zero values for the parameters of term Z_t, and for the regression coefficients, we used the estimates obtained in the GLM adjustment without the term Z_t.

In the NGSSEML class, the parameters are divided into two types: latent states parameters λ_t and fixed parameters ψ. For the latent state parameters, on-line and smoothed inferences can be performed. In the smoothed inference, the estimation of level component is based on all available information Y_n and smoothing techniques should be used, while in the on-line inference, the main interest is the on-line distribution $\lambda_t|Y_t$, for all t. For more details on inferences see Sections 2 and 3 of Gamerman, Santos e Franco (2013). For the fixed parameters, the inference can be performed using the marginal log-likelihood function. Therefore, consider the marginal log-likelihood function

$$\begin{aligned}\ln L(\psi;\mathbf{Y}_n) &= \ln \prod_{t=\tau+1}^{n} p(y_t|\mathbf{Y}_{t-1},\psi)\\ &= \sum_{t=\tau+1}^{n} \ln\Gamma(a^*_{t|t-1}+b(y_t,\psi)) - \ln\Gamma(a^*_{t|t-1}) + a^*_{t|t-1}\ln b^*_{t|t-1} + \ln(a(y_t,\psi))\\ &\quad - (b(y_t,\psi)+a^*_{t|t-1})\ln(c(y_t,\psi)+b^*_{t|t-1}),\end{aligned}$$

where τ is the instant of the first non-zero observation, $\mathbf{Y}_n = (y_1,\ldots,y_n)^\top$, and parameters of the specific model can compose ψ, in addition to w,β. The maximization of the marginal log-likelihood can be performed using a numerical solution.

Forecast

Following the approach discussed in (DUNSMUIR; SCOTT, 2015), the one-step-ahead forecast Y_{n+1}, given observations up to time n and covariates x_{n+1} at time $n+1$, is obtained as follows, for GAM-ARMA model. First, we calculate

$$\hat{W}_{n+1} = x_{n+1}^\top\hat{\boldsymbol{\beta}} + \sum_{i=1}^{r}\sum_{j=1}^{m}\hat{\alpha}_{i,j}B_j(w_{t,i}) + \hat{Z}_{n+1},$$

where

$$\hat{Z}_{n+1} = \sum_{j=1}^{p}\hat{\phi}_j(\hat{Z}_{n+1-j}-\hat{e}_{n+1-j}) + \sum_{j=1}^{q}\hat{\theta}_j\hat{e}_{n+1-j}. \tag{8}$$

The term Z_{n+1} is determined using values of Z_t and e_t for $t \leq n$. The predictive distribution of Y_{n+1} is estimated to be $f(y|\hat{W}_{n+1})$, where the density f is given by the conditional distribution used (in this case Poisson). The estimated density provides a complete description of the forecast, so we can obtain prediction intervals as well as a single-point forecast (mean, mode, or median).

For the NGSSEML model, the following predictive distributions of future observations are defined:

- Consider the following evolution equation:

  $$w\frac{\lambda_{t+h+1}}{\lambda_{t+h}}|\lambda_{t+h},\mathbf{Y}_t \sim Beta(w^{h+1}a_t,(1-w)w^h a_t),$$

  for $h = 0,1,2,\ldots.$ $\mathbf{Y}_t = \{Y_0,y_1,\ldots,y_t\}$, for $t = 1,2,\ldots,$ and Y_0 represents the past history.

- If $h > 0$, the h-step-ahead evolution can be approximated by

  $$\lambda_{t+h}|\mathbf{Y}_t,\psi \approx Gamma(a_{t+h|t},b_{t+h|t}),$$

  where $a_{t+h|t} = w^h a_t$ and $b_{t+h|t} = w^h b_t$.

- Finally, the predictive density function of the observation h steps ahead is given by

$$p(y_{t+h}|\mathbf{Y}_t,\psi) \approx \frac{\Gamma(b(y_{t+h},\psi)+a_{t+h|t})a(y_{t+h},\psi)(b_{t+h|t})^{a_{t+h|t}}}{\Gamma(a_{t+h|t})\left[c(y_{t+h},\psi)+b_{t+h|t}\right]^{b(y_{t+h},\psi)+a_{t+h|t}}},$$

  where $y_{t+h} \in \{0,1,2,\dots\}$.

Analytic expressions for these predictive distributions are not available, so all results presented for forecast in the NGSSEML model are approximate.

Real Data Application

In order to compare the ability of the GAM-ARMA and NGSSEML model, we apply both models in the analysis of count time series, more specifically, the monthly number of chronic obstructive pulmonary disease (COPD) cases in the metropolitan area of Belo Horizonte, Brazil. The period analyzed is from January 2007 to December 2013, however, the last 6 observations are excluded from the fit with the purpose of comparing the forecasts, and thus $n = 78$. This dataset has been analyzed by Camara et al. (2021) under the GAM-ARMA model. Furthermore, works such as Davis, Dunsmuir e Streett (2003) and Maia et al. (2021) present applications where they analyze the association between the concentration of atmospheric pollutants and the occurrence of respiratory diseases. Figure 1 presents the time series of COPD cases and the respective autocorrelation (ACF) and partial autocorrelation (PACF) functions.

From Figure 1, we observe that the time series presents a seasonal behavior and a positive trend; thus, the following covariates are considered here: concentration of nitrogen monoxide (NO), terms of annual and semi-annual seasonality, trend, minimum temperature ($Temp$) and relative humidity (RH) of the air. Therefore, we define the predictor as follows for the GAM-ARMA model

$$\begin{aligned} W_t = {} & \beta_1 NO_t + \beta_2 \sin(2\pi t/12) + \beta_3 \cos(2\pi t/12) + \beta_4 \sin(2\pi t/6) + \beta_5 \cos(2\pi t/6) + \\ & \beta_6 t + \alpha_{1,1} B_1(Temp_t) + \alpha_{2,1} B_2(Temp_t) + \alpha_{3,1} B_3(Temp_t) + \alpha_{1,2} B_1(RH_t) + \\ & \alpha_{2,2} B_2(RH_t) + \alpha_{3,2} B_3(RH_t) + Z_t, \end{aligned} \tag{9}$$

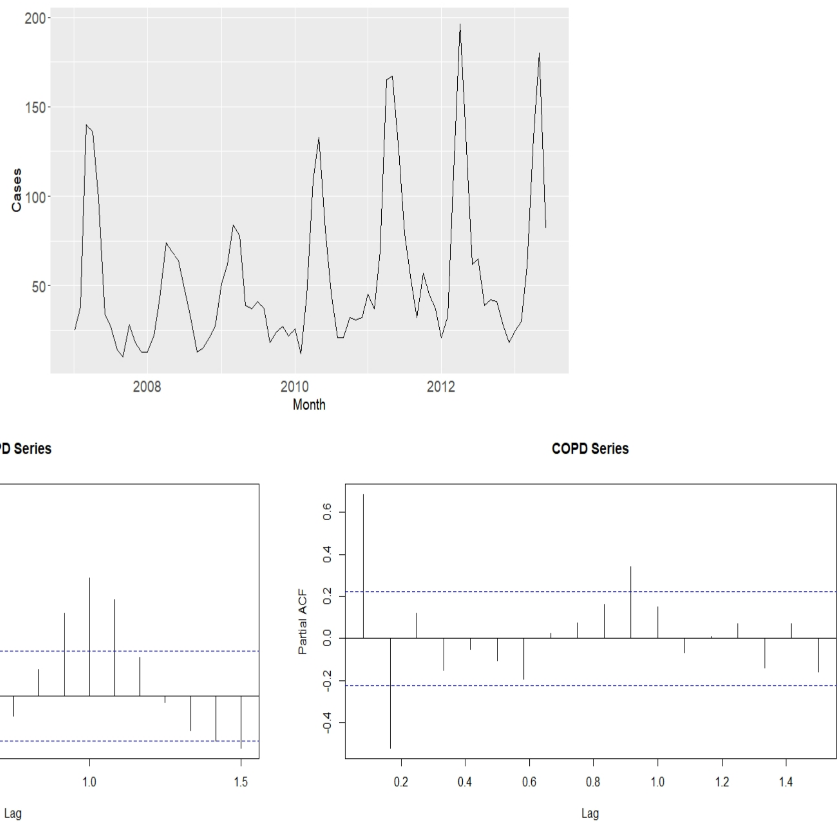
where NO_t and t (trend) show a linear relation, while $Temp$ and RH show a non-linear relation with COPD. The term Z_t is defined as an autoregressive process of order 1. The same covariates are used in the NGSSEML model besides the definition of the prior distribution $\lambda_0|Y_0 \sim Gamma(0.2, 0.1)$.

Table 1: Parameter estimates and standard errors of the GAM-AR(1) for the COPD time series.

Covariate/par.	Estimate	Standard Error
NO	0.051	0.003
$\sin(2\pi t/12)$	0.329	0.051
$\cos(2\pi t/12)$	-0.527	0.064
$\sin(2\pi t/6)$	-0.304	0.036
$\cos(2\pi t/6)$	-0.260	0.040
trend	0.012	0.001
ϕ_1	0.073	0.006

Source: Authors.

Figure 1: Plots of the monthly number of COPD cases in the metropolitan area of Belo Horizonte from January 2007 to December 2013 (top) and its associated ACF (bottom to the left) and PACF (bottom to the right).



Source: Authors.

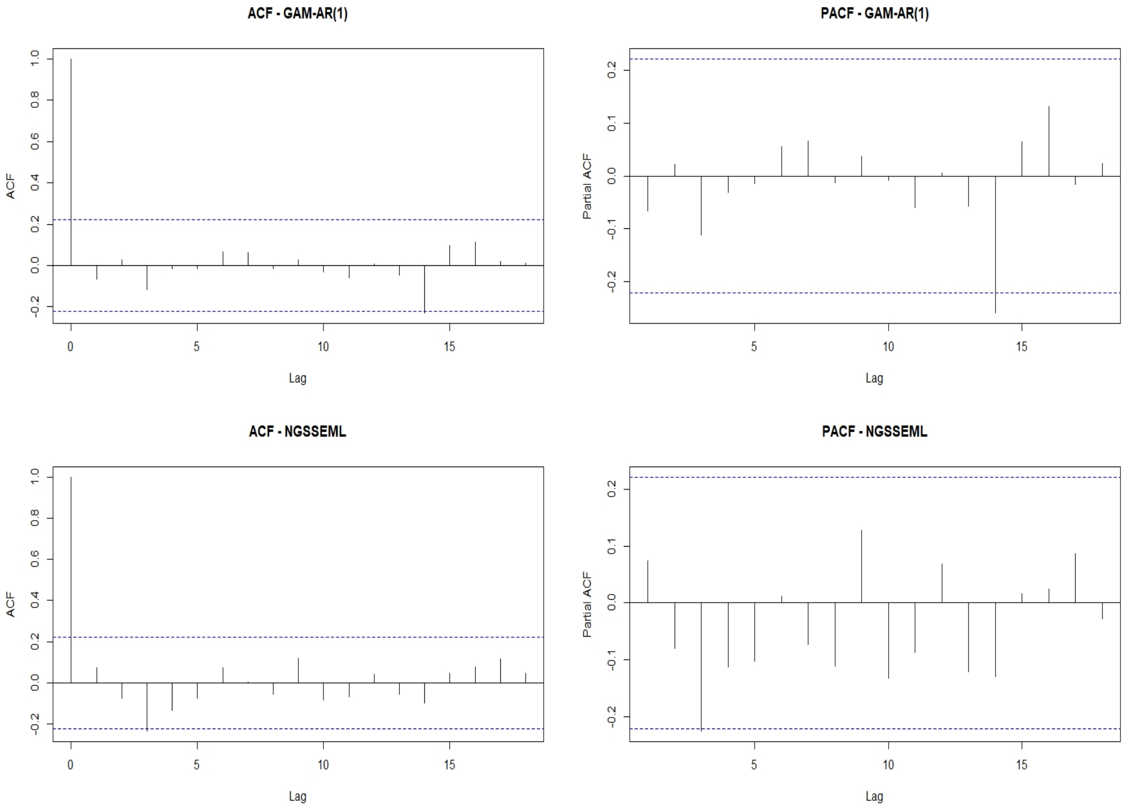
In Tables 1 and 2, we present the maximum likelihood estimates of the parameters and the respective standard errors for the GAM-AR(1) and NGSSEML models, respectively. Figure 2 presents the plots of the residuals for the two adjustments.

Table 2: Parameter estimates and respective standard errors of the NGSSEML for the COPD time series.

Covariate/par.	Estimate	Standard Error
ω	0.171	0.027
NO	0.018	0.007
$\sin(2\pi t/12)$	0.513	0.127
$\cos(2\pi t/12)$	-0.413	0.156
$\sin(2\pi t/6)$	-0.345	0.064
$\cos(2\pi t/6)$	-0.108	0.076
trend	0.008	0.037

Source: Authors.

Figure 2: Plots of the ACF and PACF of the residuals in the GAM-AR(1) (top) and NGSSEML models (bottom).



Source: Authors.

We observe, from Figure 2, that the residuals are a white noise process, which means that they are not correlated, revealing a good adjustment. For comparison purposes, we calculated the following statistics to evaluate the performance of the two fitted models: square root of the mean square error, SMSE $= \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$, the mean absolute error, MAE $= \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$ and the mean absolute percentage error, MAPE $= \frac{1}{n}\sum_{t=1}^{n}|(y_t - \hat{y}_t)/y_t|$, where $\hat{y}_t, t = 1, \ldots, n$, are the fitted values. In Table 3, we present the statistics for each model.

Table 3: In sample SMSE, MAE, and MAPE for the GAM-AR(1) and NGSSEML models.

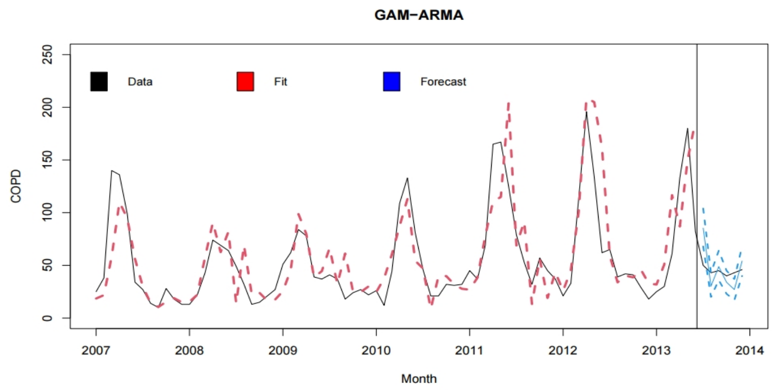
	SMSE	MAE	MAPE
GAM-AR(1)	359.393	13.664	0.395
NGSSEML	524.458	16.540	0.329

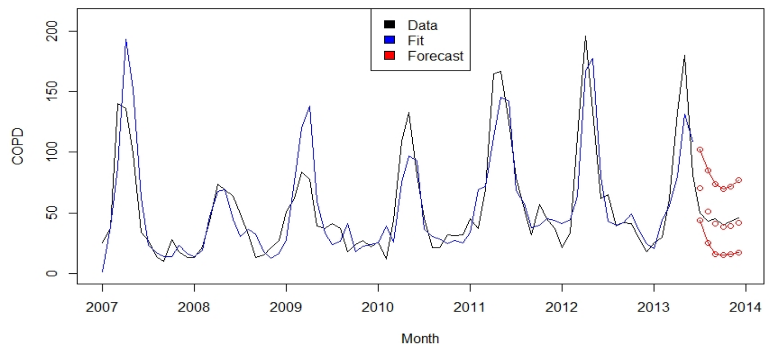
Source: Authors.

From Table 3, the statistics reveal that the adjustment with the GAM-AR(1) model provides better results.

After verifying the adequacy of the models, we are able to make forecasts. We perform the one-step-ahead forecast for the last 6 observations of the COPD time series, using the GAM-ARMA and NGSSEML approaches. Figure 3 presents the one-step-ahead forecast of the COPD time series (at a 5% significance level), based on the two approaches discussed in Section Forecast and Table 4 provides the statistics to assess the quality of each of the forecasting approaches.

Figure 3: One-step-ahead predictions of the COPD time series from $t = 79$ to $t = 84$ under the GAM-ARMA (top) and the NGSSEML (bottom) approach.





Source: Authors.

Table 4: Forecast SMSE, MAE and MAPE for the GAM-AR(1) and NGSSEML models.

	SMSE	MAE	MAPE
GAM-AR(1)	219.543	12.682	0.023
NGSSEML	87.258	6.885	0.148

Source: Authors.

The opposite result is observed for the forecast. From Table 4, we obtain better results for the NGSSEML model.

Conclusion

Based on a real data application, a comparison between the observation-driven and parameter-driven models was presented. For this, the GAM-ARMA models, proposed by Camara et al. (2021), and NGSSEML, proposed by Gamerman, Santos e Franco (2013), were used. The GAM-ARMA and NGSSEML models, were fitted using the GLARMA and NGSSEML packages, respectively. Both models use the maximum likelihood method to estimate parameters. The forecast carried out was in accordance with the approach of each model, using the forecast and NGSSEML packages.

According to the application to the COPD time series, we concluded that, to fit a model for the data, the GAM-ARMA model proved to be more appropriate, i.e. , the fitted model is closer to the real data observations, while in the forecast for the last 6 data observations of COPD time series, we observed an opposite behavior. The NGSSEML model presented better forecasts when compared to the GAM-ARMA model.

We highlight that these results are valid for this real data application and that simulation studies should be conducted with the aim of obtaining more information for comparison between the two models. Furthermore, more real data applications should be evaluated.

Acknowledgements

The authors thank the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-CAPES), National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq) and Minas Gerais State Research Foundation (Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG).

References

ALBARRACIN, O. Y. E.; ALENCAR, A. P.; HO, L. L. Generalized autoregressive and moving average models: multicollinearity, interpretation and a new modified model. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 89, n. 10, p. 1819–1840, 2019.

BENJAMIN, M. A.; RIGBY, R. A.; STASINOPOULOS, D. M. Generalized autoregressive moving average models. *Journal of the American Statistical association*, Taylor & Francis, v. 98, n. 461, p. 214–223, 2003.

BOOR, C. D. *A practical guide to splines*. [S.l.]: springer-verlag New York, 1978. v. 27.

CAMARA, A. J. A.; FRANCO, G. C.; REISEN, V. A.; BONDON, P. Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health. *Pesquisa Operacional*, SciELO Brasil, v. 41, 2021.

COX, D. R. Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, JSTOR, p. 93–115, 1981.

DAVIS, R. A.; DUNSMUIR, W. T.; STREETT, S. B. Observation-driven models for poisson counts. *Biometrika*, Oxford University Press, v. 90, n. 4, p. 777–790, 2003.

DAVIS, R. A.; DUNSMUIR, W. T.; WANG, Y. Modeling time series of count data. *Statistics textbooks and monographs*, MARCEL DEKKER AG, v. 158, p. 63–114, 1999.

______. On autocorrelation in a poisson regression model. *Biometrika*, Oxford University Press, v. 87, n. 3, p. 491–505, 2000.

DAVIS, R. A.; LIU, H. Theory and inference for a class of observation-driven models with application to time series of counts. *arXiv preprint arXiv:1204.3915*, 2012.

DAVIS, R. A.; WU, R. A negative binomial model for time series of counts. *Biometrika*, Oxford University Press, v. 96, n. 3, p. 735–749, 2009.

DUNSMUIR, W. T.; SCOTT, D. J. The glarma package for observation-driven time series regression of counts. *Journal of Statistical Software*, v. 67, p. 1–36, 2015.

EILERS, P. H. C.; MARX, B. D. Flexible smoothing with B-splines and penalties. *Statistical Science*, Institute of Mathematical Statistics, v. 11, n. 2, p. 89 − 121, 1996. Disponível em: ⟨https://doi.org/10.1214/ss/1038425655⟩.

FRANCO, G. C.; MIGON, H. S.; PRATES, M. O. A monte carlo study and data analysis for time series of counts. *Relatório Técnico Relatório Técnico RTP-01/2015, Belo Horizonte: Departamento de Estatıstica-UFMG*, 2015.

FRIEDMAN, J. H.; SILVERMAN, B. W. Flexible parsimonious smoothing and additive modeling. *Technometrics*, [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], v. 31, n. 1, p. 3–21, 1989. ISSN 00401706. Disponível em: ⟨http://www.jstor.org/stable/1270359⟩.

GAMERMAN, D.; SANTOS, T. R. dos; FRANCO, G. C. A non-gaussian family of state-space models with exact marginal likelihood. *Journal of Time Series Analysis*, Wiley Online Library, v. 34, n. 6, p. 625–645, 2013.

GREEN, P. J.; SILVERMAN, B. W. *Nonparametric regression and generalized linear models: a roughness penalty approach*. [S.l.]: Crc Press, 1994.

HARREL, F. E. Bioestatistical modeling. *Nashvile TN USA*, 2004.

HASTIE, T.; TIBSHIRANI, R. *Generalized Additive Models*. Taylor & Francis, 1990. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9780412343902. Disponível em: ⟨https://books.google.com.br/books?id=qa29r1Ze1coC⟩.

JUNG, R. C.; KUKUK, M.; LIESENFELD, R. Time series of count data: modeling, estimation and diagnostics. *Computational Statistics & Data Analysis*, Elsevier, v. 51, n. 4, p. 2350–2364, 2006.

JUNG, R. C.; TREMAYNE, A. R. Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, Springer, v. 95, p. 59–91, 2011.

KOOPERBERG, C.; STONE, C. J. A study of logspline density estimation. *Computational Statistics & Data Analysis*, Elsevier, v. 12, n. 3, p. 327–347, 1991.

MAIA, G. de O.; BARRETO-SOUZA, W.; BASTOS, F. de S.; OMBAO, H. Semiparametric time series models driven by latent factor. *International Journal of Forecasting*, Elsevier, v. 37, n. 4, p. 1463–1479, 2021.

MELO, M.; ALENCAR, A. Conway–maxwell–poisson autoregressive moving average model for equidispersed, underdispersed, and overdispersed count data. *Journal of Time Series Analysis*, Wiley Online Library, v. 41, n. 6, p. 830–857, 2020.

ROCHA, A. V.; CRIBARI-NETO, F. Beta autoregressive moving average models. *Test*, Springer, v. 18, p. 529–545, 2009.

ZEGER, S. L. A regression model for time series of counts. *Biometrika*, Oxford University Press, v. 75, n. 4, p. 621–629, 1988.

ZEGER, S. L.; QAQISH, B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, JSTOR, p. 1019–1031, 1988.