

Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health

Ana Julia A. Camara, Glaura Franco, Valderio Reisen, Pascal Bondon

▶ **To cite this version:**

Ana Julia A. Camara, Glaura Franco, Valderio Reisen, Pascal Bondon. Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health. Pesquisa Operacional, 2021, 41, pp.e241120. 10.1590/0101-7438.2021.041.00241120 . hal-03437160

HAL Id: hal-03437160

https://centralesupelec.hal.science/hal-03437160

Submitted on 19 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pesquisa Operacional (2021) 41: e241120 p.1-15
doi: 10.1590/0101-7438.2021.041.00241120
© 2021 Brazilian Operations Research Society
Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope
ARTICLES

GENERALIZED ADDITIVE MODEL FOR COUNT TIME SERIES: AN APPLICATION TO QUANTIFY THE IMPACT OF AIR POLLUTANTS ON HUMAN HEALTH

Ana Julia A. Camara[1*], Glaura C. Franco[2], Valderio A. Reisen[3] and Pascal Bondon[4]

Received July 18, 2020 / Accepted May 1, 2021

ABSTRACT. The generalized additive model (GAM) has been used in many epidemiological studies where frequently the response variable is a nonnegative integer-valued time series. However, GAM assume that the observations are independent, which is generally not the case in time series. In this paper, an autoregressive moving average (ARMA) component is incorporated to the GAM. The resulting GAM-ARMA model is based on the generalized linear autoregressive moving average (GLARMA) model where some linear components are replaced by natural splines. Numerical simulations are presented and show that the ARMA component influences the estimation. In a real data analysis of the effects of air pollution on respiratory disease in the metropolitan area of Belo Horizonte, Brazil, it is shown that the proposed model presents a better fit when compared to the classical GAM approach, that does not take into account the autocorrelation of the data.

Keywords: GAM, ARMA model, semiparametric model, Poisson-valued time series.

1 INTRODUCTION

Epidemiological data are frequently treated as time series of counts because they record the relative frequency of certain events that occur in successive time intervals and the observations are correlated.

*Corresponding author
[1]Universidade Federal de Minas Gerais (UFMG), Departamento de Estatística – ICEx, Av. Antonio Carlos, 6627, Pampulha, 31270-901, Belo Horizonte, MG, Brazil – E-mail: anajulia.camara@gmail.com – http://orcid.org/0000-0002-9382-6842
[2]Universidade Federal de Minas Gerais (UFMG), Departamento de Estatística – ICEx, Av. Antonio Carlos, 6627, Pampulha, 31270-901, Belo Horizonte, MG, Brazil – E-mail: glaura@est.ufmg.br – http://orcid.org/0000-0002-7994-8448
[3]Universidade Federal do Espírito Santo (UFES), DEST-CCE-UFES, Av. Fernando Ferrari 514, 29075-910, Vitória, ES, Brazil – E-mail: valderioanselmoreisen@gmail.com – http://orcid.org/0000-0002-8313-7648
[4]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, CNRS, UMR 8506, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France – E-mail: pascal.bondon@l2s.centralesupelec.fr – http://orcid.org/0000-0002-5158-7337

Many epidemiological studies have been carried out to investigate the impact of ambient air pollution concentrations and meteorological conditions on human health. Kelsall *et al.* (1997), Ostro *et al.* (1999), Goldberg *et al.* (2003) and other authors found significant association between daily pollutant concentration levels and mortality. Alonso *et al.* (2010) studied the impact of atmosphere pressure, air humidity, and temperature on the number of hospitalizations. Besides that, Roberts (2004), Stafoggia *et al.* (2008) and other authors found the evidence of interactive effects between temperature and air pollution (e.g., particulate matter and ozone) on mortality and adverse health outcomes. Such studies are an alert about the importance of controlling and reducing air pollutant emissions, and provide support for health departments in resource allocation.

Nevertheless, most of these studies try to model the relation between the occurrence of a disease and the air pollutants using procedures that are not able to capture the dependence inherent to the observations, such as the generalized linear model (GLM) (Nelder and Wederburn, 1972) and GAM (Hastie and Tibishirani, 1990). New methodologies were then proposed to model time series of counts. Shephard (1995) introduced the GLARMA model, then generalized by Davis *et al*. (2003). This methodology adds an ARMA structure to the GLM and is able to model time series belonging to the exponential family. In the same vain, Benjamin *et al*. (2003) proposed the generalized ARMA model. Mckenzie (1985) and Al-Osh and Alzaid (1987) introduced the integer-valued autoregressive model. Heinen (2003) proposed the autoregressive conditional Poisson model for counting data with time dependency and over-dispersion. Gamerman *et al.* (2013) proposed a family of non-gaussian state space models that allows the marginal likelihood to be calculated in an exact way.

The above models assume that the relation between the response variable and the covariates is linear. The GAM offers more flexibility and has been used by many authors to solve real problems in the environmental context, see e.g. Schwartz (2000), Aldrin and Haff (2005), and Belusic *et al.* (2015). Despite its widespread use, care is required when GAM is used in time series due to the serial correlation present in the data. Very few works are concerned with this issue, in particular Yang *et al*. (2012) who proposed GAM with autoregressive terms. Souza *et al*. (2018) have also proposed a hybrid model, including GAM, principal component analysis, and vector autoregression to address the multicolinearity problems that can occur when including several air pollutants in the analysis.

In this work a more general model for count data is proposed, which is able to handle both the autocorrelation structure of the time series and the nonlinearity existing in the covariates. This model is composed of a GAM with an ARMA component and is called a GAM-ARMA model. The non-parametric components are estimated through some smoothed functions, such as splines. Numerical simulations are performed to access the accuracy of parameter estimation in small sample size series following a Poisson distribution. Finally, a real-time series is analyzed without taking and taking into account the autocorrelation of the data. The example includes the fit of a GAM-ARMA model to evaluate the impact of air pollutants and meteorological variables

on the number of chronic obstructive pulmonary disease cases in the metropolitan area of Belo Horizonte, Brazil.

The paper is organized as follows. Section 2 presents the GAM-ARMA model, detailing some properties and the inference procedure. Section 3 shows the simulation study. Section 4 presents the analysis of a real series of pulmonary disease counts. Section 5 concludes the work.

2 THE GAM-ARMA MODEL

2.1 Presentation of the model

We combine the GAM with the ARMA model proposed by Box and Jenkins (1976) to model linear and nonlinear relations between the response variable and the covariates, and the time correlation of the response. The advantage of this methodology is the possibility to adjust semiparametric and non-parametric models to the data, capturing either linear and non-linear relationships, and thus obtaining better estimates.

As in the GLARMA model, the conditional distribution of the observation y_t given the past information $\mathscr{F}_{t-1}^{y} = \sigma\{y_s, s \leq t-1\}$ follows a Poisson distribution, i.e.,

$$y_t \mid \mathscr{F}_{t-1}^{y} \sim \text{Poi}(\mu_t), \tag{1}$$

where $\mu_t = \text{E}(y_t \mid \mathscr{F}_{t-1}^{y})$. Here, the predictor $\eta_t = \ln(\mu_t)$ follows the model

$$\eta_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{l} s_j(w_{t,j}) + Z_t, \tag{2}$$

where $(x_{t,1}, \ldots, x_{t,k})$ denotes the covariates related linearly to η_t, $(w_{t,1}, \ldots, w_{t,l})$ denotes the covariates related to η_t via smooth functions $s_1, \ldots, s_l$, and Z_t modelises the time correlation. Following Davis *et al*. (2003),

$$Z_t = \sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}, \tag{3}$$

where, for some $\lambda \in (0,1]$,

$$\varepsilon_t = (y_t - \mu_t)\mu_t^{-\lambda} = (y_t - e^{\eta_t})e^{-\lambda\eta_t}, \tag{4}$$

and the parameters τ_i's are the coefficients in the power series expansion

$$\sum_{i=1}^{\infty} \tau_i z^i = \left(1 - \sum_{i=1}^{p} \phi_i z^i\right)^{-1} \left(1 + \sum_{i=1}^{q} \theta_i z^i\right) - 1, \quad |z| \leq 1, \tag{5}$$

where the polynomials $\phi(z) = 1 - \phi_1 z - \cdots \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots \theta_q z^q$ have no common zeroes and have all their zeros outside the unit circle. It follows from (3) and (5) that Z_t can be calculated recursively with the difference equation

$$Z_t = \phi_1(Z_{t-1} + \varepsilon_{t-1}) + \cdots + \phi_p(Z_{t-p} + \varepsilon_{t-p}) + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}. \tag{6}$$

According to (4), $\mathrm{E}(\varepsilon_t \mid \mathscr{F}_{t-1}^y) = \mu_t^{-\lambda}(\mathrm{E}(y_t \mid \mathscr{F}_{t-1}^y) - \mu_t) = 0$. Now, let $\mathscr{F}_{t-1}^{\varepsilon} = \sigma\{\varepsilon_s, s \leq t-1\}$, (4) implies that $\mathscr{F}_{t-1}^{\varepsilon} \subset \mathscr{F}_{t-1}^y$. Therefore,

$$\mathrm{E}(\varepsilon_t \mid \mathscr{F}_{t-1}^{\varepsilon}) = \mathrm{E}[\mathrm{E}(\varepsilon_t \mid \mathscr{F}_{t-1}^y) \mid \mathscr{F}_{t-1}^{\varepsilon}] = 0,$$

which shows that (ε_t) is a martingale difference sequence. Hence, $\mathrm{cov}(\varepsilon_s, \varepsilon_t) = 0$ for $s \neq t$, and the variance of ε_t is

$$\mathrm{var}(\varepsilon_t) = \mathrm{E}(\varepsilon_t^2) = \mathrm{E}[\mathrm{E}(\varepsilon_t^2 \mid \mathscr{F}_{t-1}^y)] = \mathrm{E}(\mu_t^{-2\lambda}\,\mathrm{E}[(y_t - \mu_t)^2 \mid \mathscr{F}_{t-1}^y]) = \mathrm{E}(\mu_t^{1-2\lambda}). \tag{7}$$

Now, (2), (6) and (7) imply that

$$\mathrm{E}(\eta_t) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{l} s_j(w_{t,j}),$$

$$\mathrm{var}(\eta_t) = \sum_{i=1}^{\infty} \tau_i^2\,\mathrm{E}(\mu_{t-i}^{1-2\lambda}),$$

and

$$\mathrm{cov}(\eta_t, \eta_{t+h}) = \begin{cases} \sum_{i=1}^{\infty} \tau_i \tau_{i+h}\,\mathrm{E}(\mu_{t-i}^{1-2\lambda}), & \text{if } h \geq 0, \\ \sum_{i=1}^{\infty} \tau_i \tau_{i-h}\,\mathrm{E}(\mu_{t+h-i}^{1-2\lambda}), & \text{if } h < 0, \end{cases}$$

When $\lambda = 0.5$, (ε_t) are the Pearson residuals and the covariances of (η_t) do not depend on t, even if (η_t) is not strictly stationary.

2.2 Parameter estimation

There are several approaches in the literature to estimate functions s_j's. Recent studies have used reduced rank approaches due to the low computational cost and facilities to obtain good estimators of the s_j's. Wood (2006) presents a review of methods for choosing the s_j's using the GAM methodology and some approaches as thin plate regression splines (Wood, 2003), B-splines and basis splines (De Boor, 1978; Dierckx, 1993), among others.

In this work, the B-spline curves were used given their simplicity to obtain flexible smoothing. B-splines are constructed from polynomial pieces, joined at control points called knots. By definition, the B-spline $B_{i,d}$ depends on the knots $t_i \leq \cdots \leq t_{i+d+1}$, where d is the order of the polynomial. If the knot vector is $(t_1, t_2, \ldots, t_{m+d+1})$ for some positive integer number m, it is possible to form m B-splines $B_{1,d}, \ldots, B_{m,d}$ of degree d associated with this knot vector. A spline function s_j is a linear combination of B-splines, i.e.,

$$s_j = \sum_{i=1}^{m} \alpha_{i,j} B_{i,d}, \tag{8}$$

where the reals $\alpha_{1,j}, \ldots, \alpha_{m,j}$ are called the B-spline coefficients of s_j. For more properties, see De Boor (1978). Here, we take $d = 3$ and we use natural cubic splines. In this case, the polynomials before the first knot and after the last knot are modeled through linear functions, which

means that the second derivative at the two end points are zero. General accounts about splines can be found in the books by Hastie *et al*. (2008), and Ahlberg *et al*. (1967). The choice of the optimal number of knots is based on the work of Harrell (2004) and depends on the sample size n. Typically, when $n \leq 100$, three or four knots usually generate good fitting and a balanced model in relation to flexibility and loss of accuracy. For large n, five knots is a good starting point. The Akaike's information criterion (AIC) can be used to choose the number of knots, see Akaike (1973).

Combining (2) and (8), and dropping $d = 3$ in the notation, the model of the predictor can be written as

$$\eta_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{l} \sum_{i=1}^{m} \alpha_{i,j} B_i(w_{t,j}) + Z_t, \tag{9}$$

where Z_t is given by (6). Thus, for a fixed integer m and fixed knots $(t_1, t_2, \ldots, t_{m+4})$, the parameter vector of the GAM-ARMA model is defined by

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

According to (1), the conditional log-likelihood function is

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

where $\eta_t(\delta)$ is given by (9) and $Z_t(\delta)$ is obtained by (6). The maximization of $L_n(\delta)$ can be performed by Newton's method initialized with zero values for all parameters. In practice, the convergence occurs approximately within 10 iterations.

Goodness-of-fit measures for the proposed methodology can be calculated with the AIC and the bayesian information criterion (BIC) defined by

$$\text{BIC} = -2\ln(L_n(\widehat{\delta}_n)) + r\ln(n),$$

where $\widehat{\delta}_n$ are the parameter values that maximize $L_n(\delta)$ and r is the number of parameters estimated by the model.

The relative risk (RR) is widely used to measure the impact of air pollution on human health, see Baxter *et al*. (1997). RR for the pollutant covariate $x_j = (x_{t,j})$ in (9) is the relative change in the expected count of respiratory disease event per ξ-unit change in x_j while keeping the other covariates fixed, and is given by

$$\widehat{\text{RR}}_{x_j}(\xi) = \exp(\widehat{\beta}_j \xi).$$

RR and its confidence interval (CI) of level $1 - \alpha$ are estimated as follows,

$$\widehat{\text{RR}}_{x_j}(\xi) = \exp(\widehat{\beta}_j \xi), \tag{10}$$

$$\widehat{\text{CI}}\{\text{RR}_{x_j}(\xi)\} = \exp(\widehat{\beta}_j \xi \pm z_{\alpha/2}\, \text{se}(\widehat{\beta}_i)\xi), \tag{11}$$

where $\widehat{\beta}_j$ is the conditional maximum likelihood estimator $\widehat{\beta}_{j,n}$ of β_j, $\text{se}(\widehat{\beta}_j)$ is the estimated standard deviation (s.d.) of $\widehat{\beta}_j$, and $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the standard normal distribution.

3 SIMULATION STUDY

In our numerical experiment, the sample size is $n = 100$, the number of replications is $N = 1000$, $\lambda = 0.5$ in (4), $(p,q) = (1,0)$ in (6) and $(k,l,m) = (2,1,3)$ in (9). The predictor model is given by

$$\eta_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \alpha_1 B_1(w_t) + \alpha_2 B_2(w_t) + \alpha_3 B_3(w_t) + Z_t, \tag{12}$$

where the B_i's compose the B-spline basis for natural cubic splines and

$$Z_t = \phi[Z_{t-1} + (y_{t-1} - e^{\eta_{t-1}})e^{-\eta_{t-1}/2}]. \tag{13}$$

The covariates $(x_{t,1}, x_{t,2})$ are simulated (one time) with the ARMA models, $x_{t,1} = 0.42x_{t-1,1} + u_t + 0.13u_{t-1}$ and $x_{t,2} = 0.30x_{t-1,2} + v_t - 0.76v_{t-1} - 0.17v_{t-2}$ where (u_t, v_t) is a sequence of independent Gaussian random variables with zero-mean and unit variance. The covariate (w_t) is the real time series of daily minimum temperature in Vitória, Brazil, between April 10, 2005 and July 19, 2005. The parameter values are

$$\beta_0 = 0.8, \quad \beta_1 = 0.1, \quad \beta_2 = -0.2, \quad \alpha_1 = 0.5, \quad \alpha_2 = -1.0, \quad \alpha_3 = 0.8,$$

and three different values of ϕ are considered, $\phi = 0.1, 0.4, 0.6$ corresponding respectively to increasing values of the autocorrelation in the response variable.

In Table 1, $\widehat{\mu}_{\widehat{\delta}_j}$ represents the average of the N estimates of the parameter δ_j and the corresponding mean squared errors (MSE) in parenthesis for $\phi = 0.1, 0.4, 0.6$. We see that the estimates are close to the true values of the parameters. In general, the values of MSE are small, but increase as ϕ increases.

Figure 1 presents the histograms of the N estimates of ϕ and the β_j's for $\phi = 0.1, 0.4, 0.6$. While the empirical distribution of the estimates of ϕ is approximately symmetric about the true value when $\phi = 0.1, 0.4$, this distribution is asymmetric when $\phi = 0.6$. The empirical distribution of the estimates of β_0 is asymmetric about the true value for all values of ϕ. Concerning β_1 and β_2, the distributions are approximately symmetric about their true values, even when $\phi = 0.6$.

Table 1 – Parameter estimates in Model (12)–(13) with MSE in parenthesis.

	$\widehat{\mu}_{\widehat{\phi}}$	$\widehat{\mu}_{\widehat{\beta}_0}$	$\widehat{\mu}_{\widehat{\beta}_1}$	$\widehat{\mu}_{\widehat{\beta}_2}$	$\widehat{\mu}_{\widehat{\alpha}_1}$	$\widehat{\mu}_{\widehat{\alpha}_2}$	$\widehat{\mu}_{\widehat{\alpha}_3}$
$\phi = 0.1$	0.0845	0.7637	0.0986	-0.1953	0.5455	-0.9812	0.8134
	(0.0074)	(0.1795)	(0.0078)	(0.0036)	(0.1264)	(0.9071)	(0.0917)
$\phi = 0.4$	0.3927	0.6907	0.0945	-0.1956	0.5332	-0.8401	0.9035
	(0.0055)	(0.1852)	(0.0067)	(0.0028)	(0.1236)	(0.7623)	(0.0985)
$\phi = 0.6$	0.5311	0.7078	0.0362	-0.2443	0.2491	-0.6168	0.9289
	(0.0128)	(0.2084)	(0.0145)	(0.0049)	(0.2183)	(0.9690)	(0.1587)

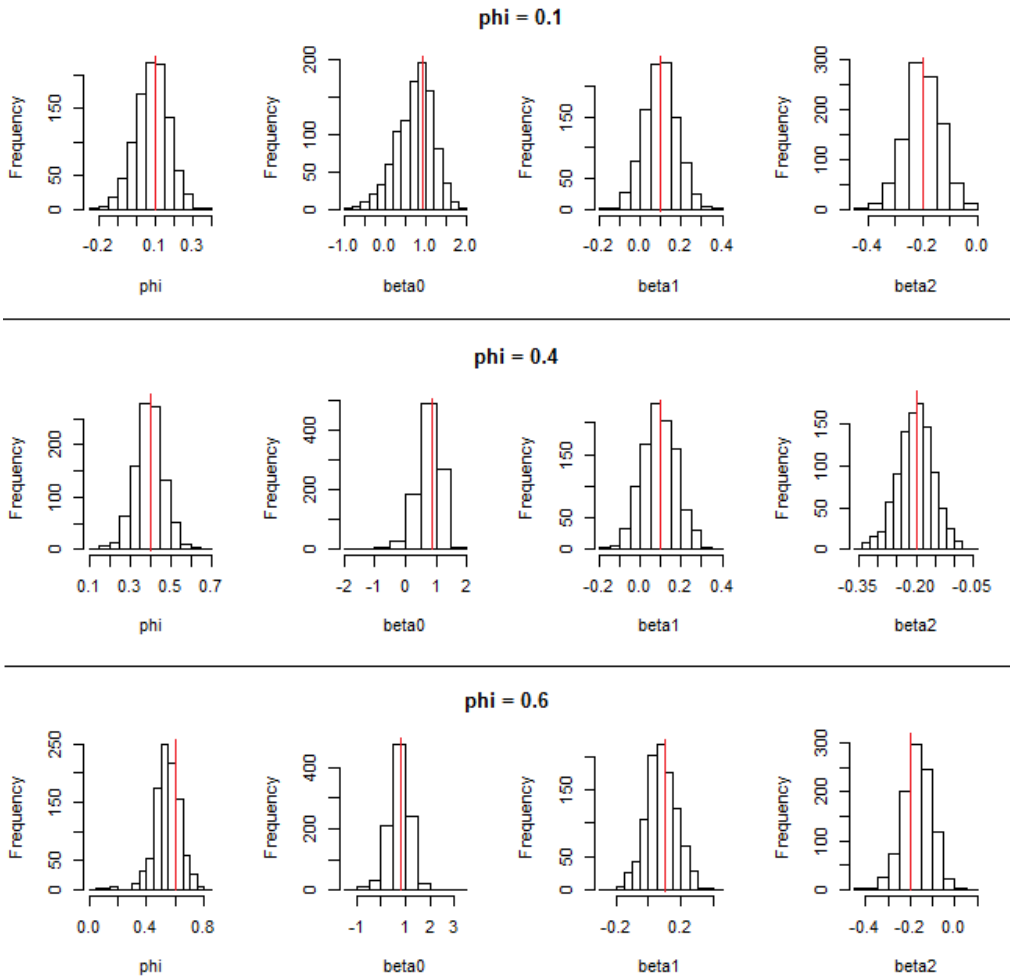


Figure 1 – Histograms of parameter estimates of ϕ and the β_j's in Model (12)–(13).

4 RESULTS

Here, we fit a GAM-ARMA model to the monthly number of chronic obstructive pulmonary disease (COPD) cases, popularly known as acute bronchitis, in the metropolitan area of Belo Horizonte, Brazil, between January 2007 and December 2013 ($n = 84$). According to the department of information technology of the Brazilian public health system, each hour three Brazilian citizens die as a result of this disease. The objective of this analysis is to evaluate the association among the concentration of atmospheric pollutants and meteorological conditions with the occurrence of COPD in Belo Horizonte.

Studies concerning air pollution in Belo Horizonte are relatively rare, even rarer regarding the relation between pollutant series and respiratory diseases. Information about the concentration

of pollutants in this region is very limited, with all the series presenting missing observations. Among the pollutants measured at the state environment and water resources institute, we select the nitrogen monoxide (NO) as the explanatory variable in this study since it presents the largest significative correlation coefficient $\rho = 0.3$ related to COPD. Some data imputations are performed before fitting the model, in order to handle the missing observations. We use a robust procedure for imputation in time series using Kalman smoothing and state space model (Harvey, 1989) and the package "imputeTS" from software R (Moritz, S., Package "imputeTS" - Time series missing value imputation).

Figure 2 presents the time series of COPD cases, NO concentration, minimum temperature (T_{min}) and relative humidity (RH) of the air. A positive trend can be detected in the number of COPD cases and NO concentration. Furthermore, all time series present a seasonal behaviour. Table 2 contains some descriptive statistics of the data, where Q1 and Q3 denote the first and third quartile, respectively.

Table 2 – Descriptive statistics of the data.

	Min	Max	Q1	Q3	Mean	Median	s.d.
Cases	10	196	27	66	54.93	41	42.3
NO ($\mu g/m^3$)	0.57	33.11	9.05	15.52	12.97	12.01	5.80
T_{min}(°C)	13.87	21.15	16.43	19.62	17.89	18.27	1.90
RH (%)	45.83	77.63	56.17	67.55	61.60	61.60	7.48

In our model, NO concentration is related linearly to η_t, while T_{min} and RH have a non-linear relation with η_t. Besides these explanatory variables, a trend component and sine and cosine functions are also incorporated in the model. The trend is included to modelise the slight positive trend in the cases of COPD. The sine and cosine functions are necessary to handle the annual and semi-annual seasonality in the response variable. Therefore, the model writes

$$\begin{aligned}\eta_t = \beta_1 x_{t,1} + \beta_2 \sin(2\pi t/12) + \beta_3 \cos(2\pi t/12) + \beta_4 \sin(2\pi t/6) + \beta_5 \cos(2\pi t/6) + \beta_6 t + \\ + \alpha_{1,1} B_1(w_{t,1}) + \alpha_{2,1} B_2(w_{t,1}) + \alpha_{3,1} B_3(w_{t,1}) + \\ + \alpha_{1,2} B_1(w_{t,2}) + \alpha_{2,2} B_2(w_{t,2}) + \alpha_{3,2} B_3(w_{t,2}) + Z_t,\end{aligned} \quad (14)$$

where t is the month number, $x_{t,1}$ is the NO concentration, $(w_{t,1})$ is T_{min} and $(w_{t,2})$ is RH. A simple GAM model where Z_t is removed in (14) is also adjusted, to show the benefict of modeling the data autocorrelation through Z_t in the GAM-ARMA model. The choice of the optimal number of knots is based on the sample size. Thus, as recommended in Section 2, three and four knots are tested, and comparing the AIC, the best model is obtained with three knots.

Table 3 presents the estimates $\widehat{\beta}_i$'s of the parameters β_i's in the fitted GAM model with the corresponding standard errors given by the software R. All estimates are significant at 5% level of significiance. On the other hand, the value of BIC is 1297.514 and the in-sample MSE between the fitted values and the observed values of COPD cases (see figure 4) is 531.642.

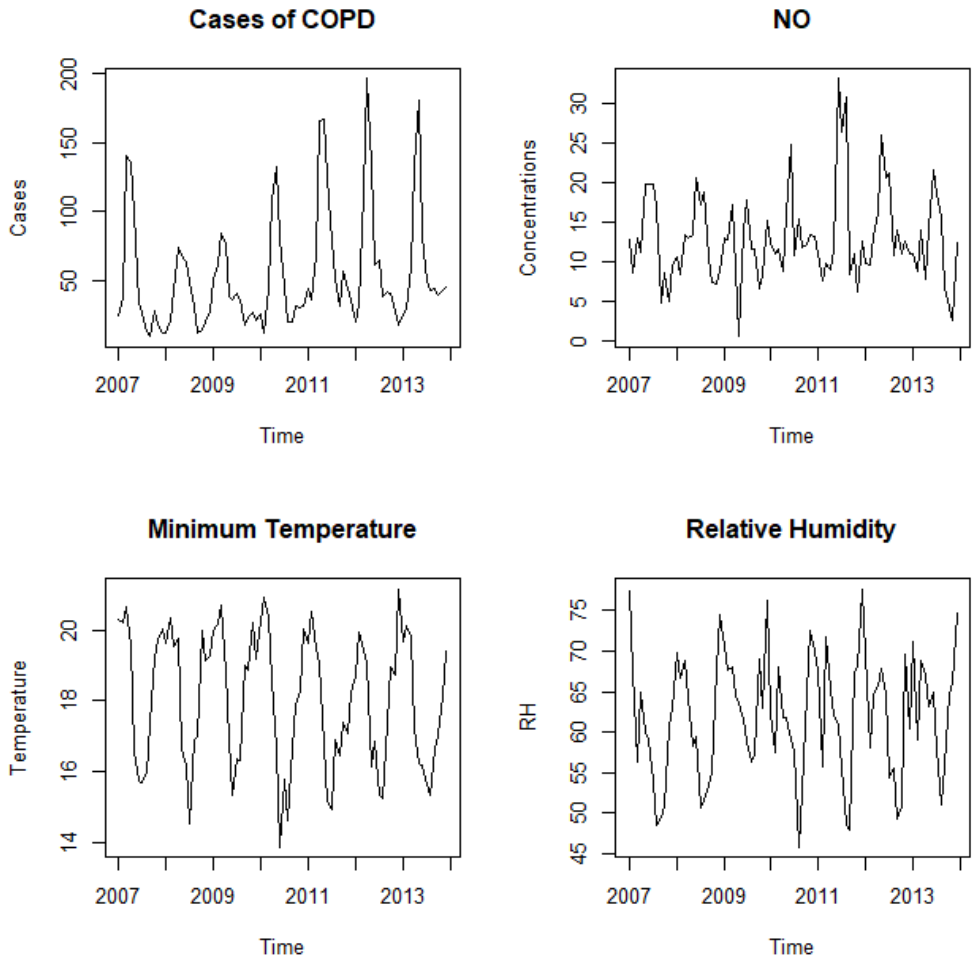
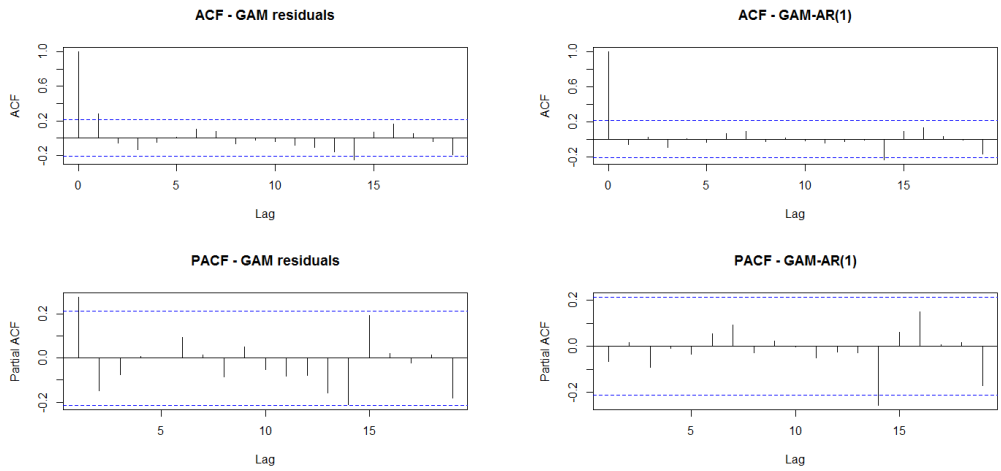


Figure 2 – Number of COPD cases, concentration of NO, minimum temperature and relative humidity of the air in the metropolitan area of Belo Horizonte, Brazil, between January 2007 and December 2013.

Table 3 – Parameter estimates of a GAM model (14) ($Z_t = 0$) fitted to the COPD cases.

Parameter	β_1	β_2	β_3	β_4	β_5	β_6
Estimate	0.0545	0.2470	-0.5562	-0.3137	-0.2522	0.0096
Standard error	0.0032	0.0415	0.0611	0.0306	0.0326	0.0007

Figure 3(a) plots the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) of the residuals in the GAM model. Some correlation is still present in these residuals, indicating the need for a more elaborated model.



a) GAM model. b) GAM-AR(1) model.

Figure 3 – Sample ACF and PACF of the residuals in the GAM and GAM-AR(1) models.

Applying the GAM-ARMA methodology, the best fit is obtained with a GAM-AR(1) model. Table 4 shows the estimates $\widehat{\beta}_i$'s and $\widehat{\phi}$ of the parameters β_i's and ϕ in the fitted GAM-AR(1) model with the corresponding standard errors given by the software R. Again, all estimates are significant at 5% level of significiance. The value of BIC is 1155.059 and the in-sample MSE between the fitted values and the observed values of COPD cases (see figure 4) is 356.169. Both values are smaller than the corresponding values obtained with the GAM model. Furthermore, the sample ACF and PACF plots in figure 3(b) show no difference with a white noise which reveals a good adjustment of the GAM-AR(1) model.

Table 4 – Parameter estimates of a GAM-AR(1) model (14) fitted to the COPD cases.

Parameter	β_1	β_2	β_3	β_4	β_5	β_6	ϕ
Estimate	0.0515	0.3271	-0.5221	-0.3068	-0.2324	0.0127	0.0700
Standard error	0.0029	0.0499	0.0615	0.0349	0.0362	0.0010	0.0053

Figure 4 shows that the GAM-AR(1) model fits better the observed number of COPD cases than the GAM model.

The RR for the NO is an important information for the regulatory agencies to quantify the impact of this pollutant on the population health. Table 5 presents the estimated RR and CI for the NO, $\widehat{\text{RR}}$ and $\widehat{\text{CI}}$ given by (10) and (11) where $\alpha = 5\%$, respectively, obtained with the GAM and GAM-AR(1) models. In both cases, $\widehat{\text{RR}}$ is significant which means that NO contributes significantly to the increase in the number of COPD cases; $\widehat{\text{RR}}$ is slightly smaller for the GAM-AR(1) model. Although $\widehat{\text{RR}}$ are comparable in the two models, the adjustment with the GAM-

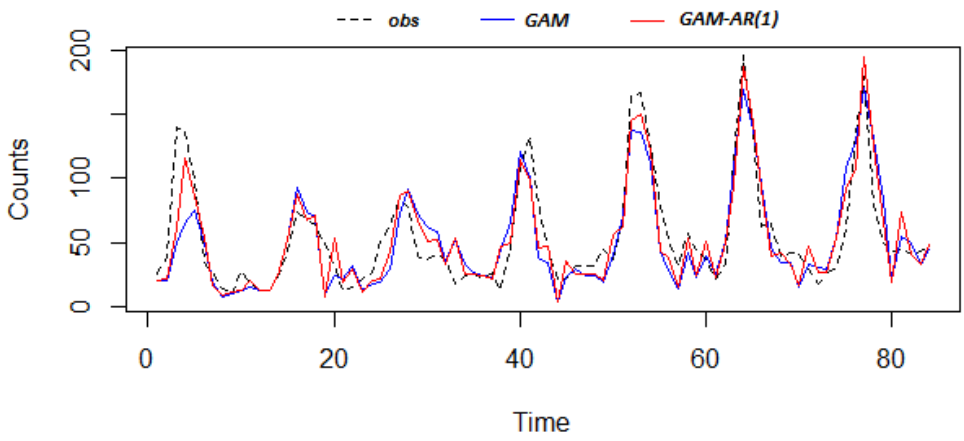


Figure 4 – Fits of GAM and GAM-AR(1) models to the number of COPD cases.

AR(1) model is the best in view of the measures of BIC and MSE, and the correlation of the residuals.

Table 5 – Estimated RR and 95% CI for the NO in the GAM and GAM-AR(1) models.

NO	GAM	GAM-AR(1)
$\widehat{\text{RR}}$	1.0627	1.0591
$\widehat{\text{CI}}$	[1.0553;1.0702]	[1.0524;1.0658]

5 CONCLUSIONS

In this work, a new methodology called GAM-ARMA was proposed, based on the GLARMA model introduced by Davis *et al.* (2003). The GAM-ARMA model allows the fitting of semiparametric models, accommodating covariates with linear and non-linear relation with the response variable in count data with time correlation.

A numerical simulation study showed that the estimates of the parameters are close to the true values for a moderate sample size of $n = 100$, and that the preciseness of the estimation degrades as the correlation in the data increases.

The model was applied to the monthly number of COPD cases in Belo Horizonte, Brazil, to quantify the impact of NO concentrations and meteorological variables on the occurrence of this disease. The best fit was obtained with a GAM-AR(1) model. This model presented white noise residuals and smaller measures of BIC and MSE compared to the GAM. The RR analysis revealed that NO contributed significantly to the increase of COPD cases.

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), Minas Gerais State Research Foundation (Fundação de Amparo à Pesquisa do estado de Minas Gerais — FAPEMIG), and Espírito Santo State Research Foundation (Fundação de Amparo à Pesquisa do Espírito Santo — FAPES). This research was partially supported by CentraleSupélec and by the iCODE Institute, research project of the IDEX Paris-Saclay, and by the Hadamard Mathematics LabEx (LMH) through the grant number ANR-11-LABX-0056-LMH in the Programme des Investissements d'Avenir. The authors are very grateful to the anonymous referee and the editor for their comments, which improved this paper.

References

[1] AHLBERG J, NILSON E & WALSH J. 1967. *The Theory of Splines and Their Application*. New York: Academic Press Inc.

[2] AKAIKE H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281.

[3] AL-OSH M & ALZAID A. 1987. First order integer valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, pp. 261–275.

[4] ALDRIN M & HOBÆK HAFF I. 2005. Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmospheric Environment*, **39**: 2145–2155.

[5] ALONSO J, ACHCAR J & HOTTA L. 2010. Climate changes and their effects in the public health: use of Poisson regression models. *Pesquisa Operacional*, **30**: 427–442.

[6] BAXTER L, FINCH S, LIPFERT F & YU Q. 1997. Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. *Risk Anal.*, **17**(13): 273–278.

[7] BELUSIC A, HERCEG-BULIC I & KLAIC Z. 2015. Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. *Geofizika*, **32**: 47–77.

[8] BENJAMIN M, RIGBY R & STASINOPOULOS D. 2003. Generalized autoregressive moving average models. *Journal of the American Statistical association*, pp. 214–223.

[9] BOX G & JENKINS G. 1976. *Time series analysis*. San Francisco: Holden-Day.

[10] CHOCK D & CHEN C. 2000. A study of the association between daily mortality and ambient air polluatant concentrations in Pittsburg, Pennsylvania. *J, Air Waste Manage, Assoc,*, **50**: 1481–1500.

[11] CIFUENTES L, KOPFER K & LAVE L. 2000. Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago, Chile. *J, Air Waste Manage, Assoc,*, **50**: 1287–1298.

[12] DAVIS R, DUNSMUIR W & STREETT S. 2003. Observation driven models for Poisson counts. *Biometrika*, **90**.

[13] DE BOOR C. 1978. *A practical guide to Splines*. Berlin: Springer.

[14] DIERCKX P. 1993. *Curve and surface fitting with Splines*. Berlin: Springer.

[15] GAMERMAN D, SANTOS T & FRANCO G. 2013. A non-Gaussian family of state-space models with exact marginal likelihood. *Journal of Time Series Analysis*, **34**: 625–645.

[16] GOLDBERG M, BURNETT R, VALOIS M, FLEGEL K, BAILAR J & BROOKS J. 2003. Associations between ambient air pollution and daily mortality among persons with congestive heart failure. *Environ. Res.*, **91**: 8–20.

[17] GREENAWAY-MCGREVY R & SUL D. 2012. Estimating the number of common factors in serially dependent approximate factor models. *Econ. Lett.*, **116**: 531–534.

[18] HAMILTON J. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

[19] HARREL F. 2004. *Bioestatistical Modeling*. Nashvile, TN.

[20] HARVEY A. 1989. *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

[21] HARVEY A. 1993. *Time Series Models*. 2nd ed.. Cambridge, MA: MIT Press.

[22] HARVEY A & FERNANDES C. 1989. Time series models for count or qualitative observations. *Journal of Business Economic Statistics*, **7**: 407–417.

[23] HASTIE T & TIBSHIRANI R. 1990. *Generalized additive models*. London: Chapman and Hall.

[24] HASTIE T, TIBSHIRANI R & FRIEDMAN J. 2008. *The elements of statistical learning*. California: Springer.

[25] HEINEN A. 2003. Modelling Time Series Count Data: Autoregressive Conditional Poisson Model observations. *Munich Personal RePEc Archive*, .

[26] HU Y & TSAY R. 2014. Principal volatility component analysis. *J. Bus. Econ. Stat.*, **32**: 153–164.

[27] KELSALL J, SAMET J, ZEGER S & XU J. 1997. Air pollution and mortality in Philadelphia. *Am. J. Epidemiol.*, **146**: 750–762.

[28] LI G, SUN J, JAYASINGHE R & PAN X. 2012. Temperature modifies the effects of particulate matter on non-accidental mortality: A comparative study of Beijing, China and Brisbane, Australia. *Public health Research*, **2**: 21–27.

[29] MATTESON D & TSAY R. 2011. Dynamic orthogonal components for multivariate time series. *J. Am. Stat. Assoc.*, **106**: 1450–1463.

[30] MCKENZIE E. 1985. Some simple models for discrete variate time series. *Water Resources Bulletin*, **21**: 645–650.

[31] NELDER J & WEDDERBURN R. 1972. Generalized linear models. *J. Roy.Statist. Soc. Ser. A*, **135**: 370–384.

[32] OSTRO B, ESKELAND G, SANCHEZ J & FEYZIOGLU T. 1999. Air pollution and health efects: A study of medical visits among children in Santiago, Chile. *Environ. Health Persp.*, **107**: 69–73.

[33] PEARSON K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**: 559–572.

[34] REN C. 2007. *Evaluation of interactive effects between temperature and air pollution on health outcomes*. Ph.D. thesis. School of Public Health, Queensland University of Technology.

[35] ROBERTS S. 2004. Interactions between particulate air pollution and temperature in air pollution mortality time series study. *Environmental Research*, **96**: 328–337.

[36] SCHWARTZ J. 2000. Harvesting and long term exposure effects in the relationship between air pollution and mortality. *Am. J. Epidemiol.*, pp. 440–448.

[37] SHEPHARD N. 1995. *Generalized Linear Autoregressions*. Technical report, Nuffield College.

[38] SOUZA J, REISEN V, FRANCO G, ISPANY M, BONDON P & SANTOS J. 2018. Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, **67**: 453–480.

[39] STAFOGGIA M, SCHWARTZ J, FORASTIERE F, PERUCCI C & GROUP S. 2008. Does temperature modify the association between air pollution and mortality? A multicity case-crossover analysis in Italy. *Am. Journal of Epidemiology*, **167**: 1476–1485.

[40] WANG Y & PHAM H. 2011. Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. *Int. J. Syst. Assur. Eng. Manag.*, **2**: 253–259.

[41] WOOD S. 2003. Thin plate regression splines. *Royal Statistical Society*, **65**: 95–114.

[42] WOOD S. 2006. *Generalized Additive Models: An Introduction With R*. Boca Raton, FL: Chapman and Hall/CRC Press.

[43] YANG L, QIN G, ZHAO N, WANG C & SONG G. 2012. Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *Medical Research Methodology*, **12**.

[44] ZAMPROGNO B. 2013. *PCA in time series with short and long-memory time series*. Ph.D. thesis. Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória, Brazil.

How to cite

CAMARA AJA, FRANCO GC, REISEN VA & BONDON P. 2021. Generalized additive model for count time series: an application to quantify the impact of air pollutants on human health. *Pesquisa Operacional*, **41**: e241120. doi: 10.1590/0101-7438.2021.041.00241120.