ANA JÚLIA ALVES CÂMARA

MODELO ADITIVO GENERALIZADO PARA DADOS DE CONTAGEM: UMA APLICAÇÃO PARA AVALIAR O IMPACTO DA POLUIÇÃO ATMOSFÉRICA NA SAÚDE

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FEVEREIRO 2019

MODELO ADITIVO GENERALIZADO PARA DADOS DE CONTAGEM: UMA APLICAÇÃO PARA AVALIAR O IMPACTO DA POLUIÇÃO ATMOSFÉRICA NA SAÚDE

ANA JÚLIA ALVES CÂMARA

Orientadora: GLAURA DA CONCEIÇÃO FRANCO – UFMG
Co-Orientador: VALDÉRIO ANSELMO REISEN – UFES

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de MESTRE em ESTATÍSTICA.

FEVEREIRO 2019

Esse trabalho é dedicado a todos que trabalham pela educação no Brasil.

Agradecimentos

Em primeiro lugar, agradeço a Deus por me dar saúde, inteligência e persistência para a realização do presente trabalho e para a minha formação como Mestre em Estatística.

Um terno agradecimento é dedicado à minha família: À minha mãe Edméia que sempre me deu todas as oportunidades e motivação para alcançar meus objetivos e ao meu irmão Igor que há anos tem sido um bom amigo.

Agradeço de forma bastante especial à minha orientadora, Professora Glaura Franco por toda dedicação, disposição, encorajamento e paciência ao longo desses dois anos. Sua orientação foi excepcional. Agradeço também ao Professor Valdério Reisen que com sua co-orientação deu importantes contribuições e motivação para a finalização do trabalho.

Um agradecimento fraterno aos meus colegas de mestrado, que em muitos momentos trouxeram alegria para os meus dias. Agradeço também aos colegas do Stats4Good, que desde o primeiro dia me fizeram sentir bem-vinda e com quem aprendi tanto academicamente, quanto pessoalmente.

Um agradecimento sincero aos meus colegas de laboratório, Victor, Arthur e Uriel que me ensinaram muito de forma altruísta.

Por fim, agradeço à CAPES pelo suporte financeiro desse trabalho.

“In God we trust, all others bring data”

W. Edwards Deming

Resumo

O Modelo Aditivo Generalizado (GAM) tem sido muito utilizado em estudos epidemiológicos, nos quais frequentemente a variável resposta é uma série temporal de números inteiros não negativos. No entanto, o modelo GAM possui a suposição de independência das observações, o que em geral não ocorre em séries temporais. Sendo assim, nessa dissertação propõe-se o Modelo Aditivo Generalizado Autorregressivo Média Móvel (GAM-ARMA), o qual trata-se de uma extensão do GAM com um componente autorregressivo média móvel. O GAM-ARMA é fundamentado no Modelo Linear Generalizado Autorregressivo Média Móvel (GLARMA), com alguns componentes lineares do GLARMA sendo substituídos por *splines* naturais. Neste trabalho são apresentadas as estimações tanto dos elementos paramétricos quanto dos componentes não-paramétricos do modelo. Com o objetivo de avaliar a performance do modelo proposto, foram realizados dois estudos de simulação que mostraram que embora as estimativas apresentem pouco vício e valores baixos de erro quadrático médio, os componentes autorregressivo e média móvel influenciam as estimativas. De forma geral, melhores estimativas foram encontradas quando esses componentes assumiram valores pequenos. A análise de dados reais avaliou o impacto dos poluentes atmosféricos na ocorrência de doença respiratória na região metropolitana de Belo Horizonte, Brasil. O modelo GAM-ARMA apresentou um ajuste melhor que o obtido através do GAM, amplamente utilizado e que não leva em consideração a autocorrelação das observações.

Essa dissertação será apresentada em inglês no formato de artigo.

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

Ana Julia A. Camara,[*,†] Glaura C. Franco,[*,†] and Valderio A. Reisen[*,‡]

†*Federal University of Minas Gerais, Belo Horizonte, Brazil*

‡*Federal University of Espirito Santo, Vitoria, Brazil*

E-mail: anajulia.camara@gmail.com; glaura@est.ufmg.br; valderioanselmoreisen@gmail.com

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 assumes that the observations are independent, which is generally not the case in time series. Therefore, in this paper we propose the Generalized Additive Autoregressive Moving Average (GAM-ARMA) model, which is an extension of GAM with an autoregressive moving average component. The GAM-ARMA is based on the Generalized Linear Autoregressive Moving Average Model (GLARMA), with some linear components of GLARMA being replaced by natural splines. Here we focused on the estimation either for the parametric elements, as well for the non-parametric components. To evaluate the performance of the GAM-ARMA we performed two simulation studies which showed that, although the estimates present small bias and mean squared error, the autoregressive and moving average components influence the estimation. In general, we found better estimates when these components assume small values. In a real data analysis of the effects of air pollution on respiratory disease in

the metropolitan area of Belo Horizonte, Brazil, we observed that the proposed model presented a better fit when compared to the widely used GAM approach, that do not take into account the autocorrelation of the data.

Keywords: Autocorrelation, ARIMA models, Semiparametric Models, Respiratory diseases, Principal component analysis, Relative Risk.

1. Introduction

Many epidemiological studies have been carried out to investigate the impact of atmospheric pollution and meteorological conditions on human health. Pope *et al.* (1995), Dockery and Pope (1996), Villeneuve *et al.* (2003) and other authors indicated a positive association between mortality and particulate matter (PM). Ostro *et al.* (1999), Schwartz (2000) and Chen *et al.* (2010) found a significant association between daily pollutant concentration levels and hospital admission for respiratory and cardiovascular diseases. Besides that, McGeehin and Mirabelli (2001), Ostro *et al.* (2009) and Hertel *et al.* (2009) analyzed temperature effects on mortality in USA and Germany. Studies such as these provide support for health departments in resource allocation and stakeholders in prevention.

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 have, as an important characteristic, the dependency between observations. For modelling this type of data it is necessary to use discrete probability distributions for non-negative integer numbers as Poisson or Negative Binomial distributions.

Nelder and Wedderburn (1972) proposed the Generalized Linear Models (GLM), that are an extension of the normal linear models. The basic idea consists in increasing the possibilities for the distribution of the response variable. In this case, the response variable can assume distributions belonging to the exponential family, e.g. Normal, Poisson, Gamma, Negative Binomial, etc. We can also have more flexibility to the relation between the mean of the dependent variable (μ) and the linear predictor (η), which can assume any monotonous

non-linear function.

Nevertheless, the GLM are not able to capture time dependency structure, which are a characteristic of time series. The Autoregressive Moving Average (ARMA) model, proposed by Box and Jenkins (1976), is one of the most used procedures for modelling time series. However, the ARMA models have the assumption that the series follows the Normal distribution, which is not the case of count data.

New methodologies were then proposed to model time series of counts. Davis *et al.* (2003) proposed the Generalized Linear Autoregressive Moving Average (GLARMA) model, that relates the GLM and ARMA models. This methodology adds an autoregressive moving average structure to the GLM, thus being able to model time series belonging to the exponential family.

Other methods for modelling non-gaussian series also were developed. Benjamin *et al.* (2003) proposed the Generalized Autoregressive Moving Average (GARMA), another extension of GLM models, that presents a different autoregressive structure. Mckenzie(1985) and Al-Osh and Alzaid (1987) introduced the Integer-valued autoregressive (INAR) model. Heinem (2003) proposed the ACP model class (Autoregressive Conditional Poisson) for counting data, able to deal with time dependency and overdispersion. Harvey and Fernandes (1989) used state-space models with conjugate prior distributions, where the counts are modelled as a Poisson distribution whose mean is obtained from a gamma distribution.

However, the relation between the response variable and the covariates can be non-linear, and the previous models were developed over the assumption of linearity. In this case, we have to use some flexible techniques for modelling non-linear time series. Many authors have been using the Generalized Additive Model (GAM) with Poisson marginal distribution in time series to quantify the non-linear association between the effects of air pollution on health and covariates, such as the concentration of air pollutants and meteorological conditions. Proposed by Hastie and Tibishirani (1990), the GAM has been used in time series in many recent studies. Schwartz (2000) analyzed the effect of air pollution in the number of deaths

caused by diseases related to air quality. Aldrin and Haff (2005) employed a GAM in order to model particulate matter concentrations, PM_{10}, $PM_{2.5}$ and $PM_{10} - PM_{2.5}$ (particle sizes between 2.5 and 10 μm), based on meteorological predictors. Belusic *et al.* (2015) used GAM to study the effects of pollutant concentrations for locations across the urban area of Zagreb, Croatia. Despite its widespread use, care is required when the GAM is used in time series. Souza *et al.* (2018) performed a literature review describing the main problems caused by application of the GAM model in data with serial correlation, once GAM assumes that errors are mutually independent. Besides that, many studies in the epidemiologic area have been using the GAM on evaluating the effects associated with a single pollutant on human health, because the air pollutants have a high correlation among them (the paper of Dionisio *et al.* (2016) discuss some challenges of multipollutant exposure).

In this direction, Yang *et al.* (2012) proposed the Generalized Additive model with Autoregressive terms (GAMAR), a methodology for modelling data with time dependence, and following some distributions belonging to the exponential family. The GAMAR model is a GAM with autoregressive structure and with the moving average terms of ARMA methodology omitted. GAMAR is derived from the GARMA model and is a non-parametric model, once all linear terms are replaced by smooth functions.

In the same way of GAMAR, we propose in this article the Generalized Additive Autoregressive Moving Average (GAM-ARMA) model. The GAM-ARMA is based on the GLARMA model, proposed by Davis *et al.* (2003), allowing the fitting of semiparametric models, which includes both linear and nonlinear components in the mean of the process. The gain of GAM-ARMA over GAMAR is the addition of a moving average component, besides the autoregressive term, which makes the model more flexible to capture the autocorrelation structure of the data. Moreover, with GAM-ARMA we have the possibility to adjust semiparametric models, instead of only nonlinear components, such as in the GAMAR approach. The proposed model is presented here focusing on the estimation procedure, either for the parametric counterpart, as well as for the non-parametric components, which are

estimated through some smoothed functions, such as splines. We also derive some properties of the GAM-ARMA model, concerning stationarity conditions.

With the purpose of evaluating the performance of the proposed model for small sample size series, we have performed some simulation studies in order to access the accuracy of parameter estimation. The Monte Carlo studies included models with and without parametric terms in the covariates, from time series generated under the Poisson distribution. We have also analyzed a real time series and compared the fitting with models that do not take into account the autocorrelation of the data. The example includes the fit of GAM-ARMA 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 GLARMA in some detail and the proposed GAM-ARMA model. In Section 3 we show some simulation studies and in Section 4 we present the analysis of a real data set. Finally, section 5 concludes the work.

2. Methodology

In this section we present the GAM-ARMA model, which is based on the work of Davis *et al*. (2003), who has proposed a procedure for modeling data with time dependence and following distributions belonging to the exponential family. For a better understanding of the proposed model, we will first present the GLARMA model and then the GAM-ARMA model.

2.1 Generalized Linear Autoregressive Moving Average Model (GLARMA)

The class of GLARMA models was proposed by Shephard (1995) and generalized by Davis *et al.* (2003). This methodology presents an alternative to model series with the structure of time dependence, being an extension of GLM.

Let y_t be the observations and $\mathcal{F}_{t-1} = (y^{(t-1)}, x^{(t)})$, where $y^{(t-1)}$ is the past of the count

process and $x^{(t)}$ is the past and present of the regressor variables. The conditional distribution of $y_t|\mathcal{F}_{t-1}$ can follow any distribution belonging to the exponential family (EF):

$$y_t|\mathcal{F}_{t-1} \sim EF(\mu_t),$$

where μ_t is the mean of the process.

The linear predictor is given by

$$\eta_t = g(\mu_t) = \mathbf{x}_t^T\boldsymbol{\beta} + Z_t = \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{\infty}\tau_i\varepsilon_{t-i}, \tag{1}$$

where $g(.)$ is a link function, $\mathbf{x}$ is a vector of r explanatory variables, $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)$ is a vector of r parameters and τ_i are the parameters of the error component ε_t. The ε_t term allows the modelling of the autocorrelation structure present in the time series.

The component $Z_t = \sum_{i=1}^{\infty}\tau_i\varepsilon_{t-i}$ can be specified in terms of a finite number of parameters using the methodology of Box and Jenkins (1976),

$$\gamma(B) = \sum_{i=1}^{\infty}\tau_i B^i = \frac{\theta(B)}{\phi(B)} - 1, \tag{2}$$

where B is the backshift operator of the form $B^k(Z_t) = Z_{t-k}$. The autoregressive and moving average components $\phi(B)$ and $\theta(B)$ are polynomials that have their roots outside the unit circle. This model is known as ARMA(p,q). If $q = 0$ we have an AR(p). Thus, based on Equation (2), Z_t can be written as

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

where ε_t is defined as:

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

with $\lambda \in (0, 1]$.

One of the most frequent suppositions in relation to the time series is stationarity, that

is, it develops randomly in time around a constant mean, reflecting some form of stable equilibrium (Moretin and Toloi, 1985). Davis *et al.* (2003) proved some properties of the GLARMA model, such as stationarity and ergodicity. They presented only the case of a Poisson distribution and estimated the parameters through the maximum likelihood approach. The authors concluded that the stationarity of the process η_t depends on parameter λ. For $\lambda = 0.5$ there exists a stationary distribution for η_t even if μ_t is not strictly stationarity.

2.2 Generalized Additive Autoregressive Moving Average Model (GAM-ARMA)

In this section, we extend the results of Davis *et al*. (2003) in order to allow the modelling not only of linear terms but also the possibility to include covariates with non-linear correlation with the response variable.

Thus, instead of GLM, we will use the GAM methodology, proposed by Hastie and Tibishirani (1990), combined with the ARMA model, proposed by Box and Jenkins (1976). This model will be denominated GAM-ARMA and the advantage of this proposed methodology is the possibility to adjust semiparametric and non-parametric models to the data, capturing either linear and non-linear relationships, and obtaining better estimates.

As in the GLARMA model, the observations, y_t given the past information, possess any distribution belonging to the exponential family,

$$y_t|\mathcal{F}_{t-1} \sim EF(\mu_t),$$

where μ_t is the mean of the process.

Following the idea of GLARMA models, the predictor in a GAM-ARMA(p,q) model is written with the addition of an autoregressive moving average structure of order p and q to the general form of a GAM predictor. Considering semiparametric form, the predictor η_t combines k variables $(x_1, ..., x_k)$ related linearly with the response variable and r variables $(w_1, ..., w_r)$ related with the response variable through some smooth function, in the following

way:

$$\eta_t = \beta_0 + \beta_1 x_{t,1} + ... + \beta_k x_{t,k} + s_1(w_{t,1}) + ... + s_r(w_{t,r}) + \sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}, \tag{5}$$

where s is a smooth function.

The structure of component $\sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}$ is defined similarly to the GLARMA model in (3) and (4).

There are several approaches in the literature to estimate the functions s. Recent studies have been using reduced rank approaches due to the low computational cost and facilities to obtain good estimators of the smoothing terms. In his book, Wood (2006) presents a review about the estimation of these terms through the GAM methodology proposed by Hastie and Tibishirani (1990), and some approaches as Thin Plate Regression Splines (Wood, 2003), B-splines or Basis Splines (De Boor, 1978; Dierckx, 1993), among others.

In this work we use B-spline curves 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 $B_{j,d}$ depends on the knots $t_j, ..., t_{j+1+d}$, where d is the order of polynomial piece. If the knot vector is given by $\mathbf{t} = (t_j)_{j=1}^{m+d+1}$ for some positive integer number m, we can form m B-splines $\{B_{j,d}\}_{j=1}^{m}$ of degree d associated with this knot vector. A linear combination of B-splines is called *spline* function and is given by

$$s = \sum_{j=1}^{m} \alpha_j B_{j,d}, \tag{6}$$

where $\alpha = (\alpha_j)_{j=1}^{m}$ are m real numbers called the B-spline coefficients or control points of s. For more properties, see De Boor (1978).

The most common spline is a *cubic spline* because it is of low degree, fairly smooth assuming continuity restrictions up to the second derivative, and yet has the power to incorporate several different trends in data simply by increasing the number of knots. A function s is called cubic spline on $[a, b]$, if s is a cubic polynomial s_i in each interval $[w_t, w_{t+1}]$. We use in this paper the *natural cubic splines*, which is a restriction of cubic splines. In this

case, the polynomials before the first knot and after the last knot are modeled through linear functions, which means the second derivative at the two end points are zero. In terms of basis splines, it means that B-splines are constructed from cubic polynomials ($d = 3$) with restrictions before and after of the knots at the ends.

General accounts about splines can be found in the books by Hastie *et al.* (2008), and Ahlberg *et al.* (1967). Wegman and Wright (1983) wrote an important paper about splines in Statistics.

Choosing the optimal number and positions of knots in splines approach, through B-splines, is a complex task. Eilers and Marx (1996) do a review about the main challenges involving the choose of knots. Too many knots lead to the overfitting problem, while few knots lead to the underfitting of data. Friedman and Silverman (1989), and Kooperberg and Stone (1991,1992) proposed automatic methods to optimizing the number and positions of knots, which in general is a difficult numerical problem. Some authors as O'Sullivan (1986, 1988) proposed some penalty to prevent the overfitting, and Eilers and Marx (1986) proposed a simplify and generalized penalty based on O'Sullivan's work called P-spline. In general, the quality of the approach depends more on the number of knots, than their positions. Popularly, a good method is the use of tertiles, quartiles, and percentiles. Harrell (2004) recommends that the number of knots is decided based on the sample size available. For a sample size less than 100, three or four knots usually generate good fitting and a balanced model in relation to flexibility and loss of accuracy. For big samples, five knots is a good starting point. We can use the Akaike's information criterion (AIC), see Akaike (1973), to chose the number of knots, which is defined as:

$$AIC = -2ln(L) + 2p, \tag{7}$$

where p is the number of parameters in the model, and L is the likelihood function of the model.

So the parameter vector of GAM-ARMA is defined as $\delta = (\alpha^T, \beta^T, \tau^T)^T$, with $\tau =$

$(\phi^T, \theta^T)^T$, and the parameter estimation is realized jointly by maximizing the likelihood function by numerical methods.

Let $f(y_t|\mathcal{F}_{t-1})$ be the conditional density of Y_t given $\mathcal{F}_{t-1}$, where f is any distribution in the exponential family. Thus the log-likelihood function can be written as

$$l(\delta, y_t) = \sum_{t=1}^{n} log f(y_t|\mathcal{F}_{t-1}).$$

To facilitate the understanding, the dependence of ε_t and δ are suppressed. The maximization of the log-likelihood is performed by Newton's method, and the initialization is performed with zero initial values for all parameters, including the autoregressive and moving average terms. Most of the times, the convergence occurs approximately within 10 iterations.

For inference in the GAM-ARMA model, the central limit theorem holds, so the asymptotic distribution of the maximum likelihood estimators is

$$\widehat{\delta} \approx N(\delta, \widehat{\Omega}), \tag{8}$$

where the approximate covariance matrix of the estimators is:

$$\widehat{\Omega} = -(\frac{\partial^2 L(\widehat{\delta})}{\partial\delta\partial\delta^T})^{-1}. \tag{9}$$

As the GAM-ARMA is based on the GLARMA model, more details about stationarity, properties and inference can be found in Davis *et al.* (2003).

For illustration, we will use the case where the count time series follows the Poisson distribution. The GAM-ARMA Poisson is defined as

$$y_t|\mathcal{F}_{t-1} \sim Poisson(\mu_t).$$

Omitting terms which do not involve parameters, the log-likelihood for the Poisson distribution is

$$l(\delta, y) = \sum_{t=1}^{n} (y_t\eta_t - e^{\eta_t}), \tag{10}$$

where

$$\eta_t = ln(\mu_t) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{r} s_j(w_{t,j}) + Z_t, \tag{11}$$

with Z_t defined in (3) and

$$\varepsilon_t = \frac{y_t - \mu_t}{\mu_t^{\lambda}} = (y_t - e^{\eta_t})e^{-\lambda\eta_t}. \tag{12}$$

If $\varepsilon_v = 0$ and $Y_v = 0$ for $v \leq 0$, then ε_t have mean

$$E(\varepsilon_t|\mathcal{F}_{t-1}^{\varepsilon}) = 0, \tag{13}$$

variance given by

$$\begin{aligned} Var(\varepsilon_t) = E(\varepsilon_t^2) = E[E(\varepsilon_t^2|\mu_t)] = \\ E[\mu_t^{-2\lambda}E[((y_t - \mu_t)^2|\mu_t]] = E(\mu_t^{-2\lambda}\mu_t) = E(\mu^{1-2\lambda}), t \geq 1 \end{aligned} \tag{14}$$

and covariance

$$Cov(\varepsilon_t, \varepsilon_v) = 0, t \neq v. \tag{15}$$

From the properties above, the mean and variance for η_t in the Poisson GAM-ARMA are

$$E(\eta_t) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{r} s_j(w_{t,j}), \tag{16}$$

and

$$Var(\eta_t) = Var[\beta_0 + \sum_{j=1}^{k} \beta_j x_{t,j} + \sum_{j=1}^{r} s_j(w_{t,j})] + Var[\sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}] = \sum_{i=1}^{\infty} \tau_i^2 E(\mu_t^{1-2\lambda}), \tag{17}$$

and for $v = t + h, h > 0$,

$$Cov(\eta_t, \eta_{t+h}) = \sum_{i=1}^{\infty} \tau_i \tau_{i+h} E(\mu_{t-i}^{1-2\lambda}). \tag{18}$$

If $\lambda = 0.5$, we have $Var(\varepsilon_t) = 1$ and the covariances are not dependent of time t. So we have a stationary distribution for η_t, even if μ_t is not strictly stationary.

In relation to the analysis of the goodness-of-fit of the GAM-ARMA, this can be performed through the Akaike Information Criterion (AIC), defined in (7), and by the Bayesian Information Criterion (BIC), which is most used in time series models and is defined as

$$BIC = -2ln(L) + kln(n), \tag{19}$$

where L is the likelihood function of the model, k is the number of parameters to be estimated and n is the number of observations or the sample size.

3. Simulation study

In order to evaluate the performance of the proposed model with respect to the parameters estimation, some simulation studies were developed. As the Poisson distribution is the most used in practical examples, the simulation studies are considering just this distribution. The first simulation study considers the Poisson GAM-ARMA model with only non-parametric components in the covariates. In the second study, we evaluate semiparametric models, with parametric and non-parametric terms in the covariates, but only with autoregressive terms. The R codes for simulations are included in the Appendix.

3.1 Non-parametric Models

For this simulation, we generated 1000 samples of size 100. The model is given by:

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim Poisson(\mu_t) \\ \eta_t = ln(\mu_t) &= \beta_0 + ns(w_t, 5) + \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i}, \\ \text{where, } ns(w_t, 5) &= \sum_{i=1}^{5} \alpha_i B_i(w_t), \end{aligned} \tag{20}$$

with ns being a natural cubic spline, and w_t is the series of minimum temperature in Vitoria, Brazil, between April 10, 2005 and July 19, 2005. The data set is available on the IEMA website: `https://iema.es.gov.br/qualidadedoar/dadosdemonitoramento/automatica`.

Model 1: Autoregressive Case

In this case, the $Z_t = \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i}$ term is an AR(1), where:

$$Z_t = \phi[Z_{t-1} + (Y_{t-1} - e^{\eta_{t-1}})e^{-\lambda\eta_{t-1}}]. \tag{21}$$

The values of ϕ were fixed at $0.1, 0.4$, and 0.6. The parameter λ assumed the value 0.5, based on the conclusions about stationarity in Section 2.2 and due to distinct values implied in different estimates. The terms B_i, $i = 1, ..., 5$ composed the B-spline basis for the natural cubic spline. The parameters used in this simulation were

$$\beta_0 = 0.8,\ \alpha_1 = 0.1,\ \alpha_2 = -0.2,\ \alpha_3 = 0.5,\ \alpha_4 = -1.0,\ \alpha_5 = 0.8$$

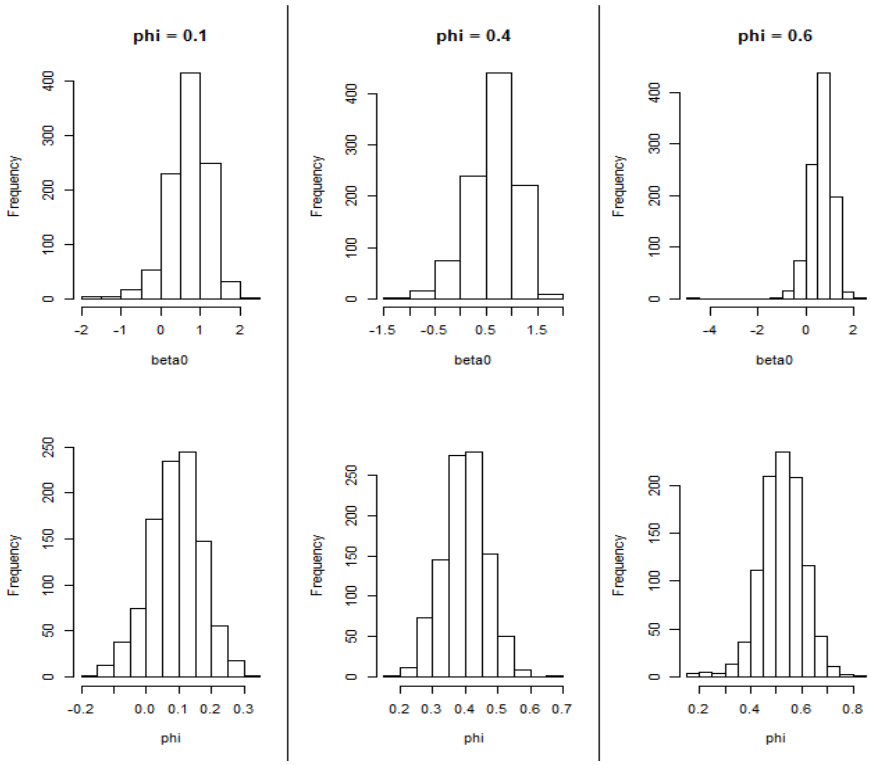
Table 1 shows the mean of parameter estimates in 1000 replicates of Model 1. The values in parenthesis are the Mean Squared Error.

Table 1: Results of parameter estimates in 1000 replicates of size 100 of Model 1

	ϕ	β_0	α_1	α_2	α_3	α_4	α_5
$\phi = \mathbf{0.1}$	0.0894 (0.0070)	0.7506 (0.2030)	0.1465 (0.2208)	-0.1794 (0.3245)	0.5465 (0.1961)	-0.9709 (1.1107)	0.8078 (0.1506)
$\phi = \mathbf{0.4}$	0.3985 (0.0048)	0.6878 (0.1974)	0.1897 (0.1891)	-0.1002 (0.2739)	0.5461 (0.1591)	-0.8822 (0.8773)	0.8314 (0.1314)
$\phi = \mathbf{0.6}$	0.5161 (0.0149)	0.6706 (0.2599)	0.2579 (0.2774)	0.0007 (0.3628)	0.8259 (0.3544)	-0.6242 (1.1667)	0.7959 (0.1982)

In general, the estimates were close to the real values, but the worst estimates, and consequently the biggest MSE, were obtained when $\phi = 0.6$. So, we had better estimates when ϕ assumed smaller values. This occurred for the estimates of ϕ, as well as to the components of the basis expansion.

Figure 1: Histograms of parameters estimates β_0 and ϕ - Model 1



In Figure 1 we have the histograms of simulations for parameters β_0 and ϕ relative to the Model 1. For parameter β_0, the parameter distribution was approximately asymmetric in all cases and centered around 0.8, the true value of this parameter. Already regarding ϕ, the distribution was approximately symmetric around the true value when ϕ was equal to 0.1 and 0.4. However, to ϕ equal 0.6 the distribution was not centered around the true value of the parameter and was relatively asymmetric.

Model 2: Moving Average Case

In the Model 2, the Z_t term is a MA(1), where

$$Z_t = \theta(Y_{t-1} - e^{\eta_{t-1}})e^{-\lambda\eta_{t-1}}. \tag{22}$$

The parameter θ was fixed at values $0.1, 0.4$, and 0.6, and λ assumed the value 0.5.

The same parameters were used for the basis:

$$\beta_0 = 0.8,\ \alpha_1 = 0.1,\ \alpha_2 = -0.2,\ \alpha_3 = 0.5,\ \alpha_4 = -1.0,\ \alpha_5 = 0.8$$

Table 2: Results of parameter estimates in 1000 replicates of size 100 of Model 2

	θ	β_0	α_1	α_2	α_3	α_4	α_5
$\theta = \mathbf{0.1}$	0.0935 (0.0077)	0.7357 (0.2221)	0.1520 (0.2367)	-0.1656 (0.3437)	0.5639 (0.1927)	-0.9208 (1.2603)	0.8086 (0.1513)
$\theta = \mathbf{0.4}$	0.4281 (0.0052)	0.7123 (0.2805)	0.1675 (0.2689)	-0.1335 (0.3357)	0.5438 (0.1723)	-0.8659 (1.2629)	0.8056 (0.1508)
$\theta = \mathbf{0.6}$	0.6260 (0.0062)	0.7419 (0.1408)	0.1176 (0.1204)	-0.1752 (0.1667)	0.5095 (0.0889)	-0.8776 (0.5815)	0.7536 (0.1042)

From Table 2, we can observe that the estimates are close to the real values, for all values of θ. Differently than observed in the autoregressive case, the estimates do not become worse as θ increases.

Figure 2: Histograms of parameters estimates β_0 and θ - Model 2

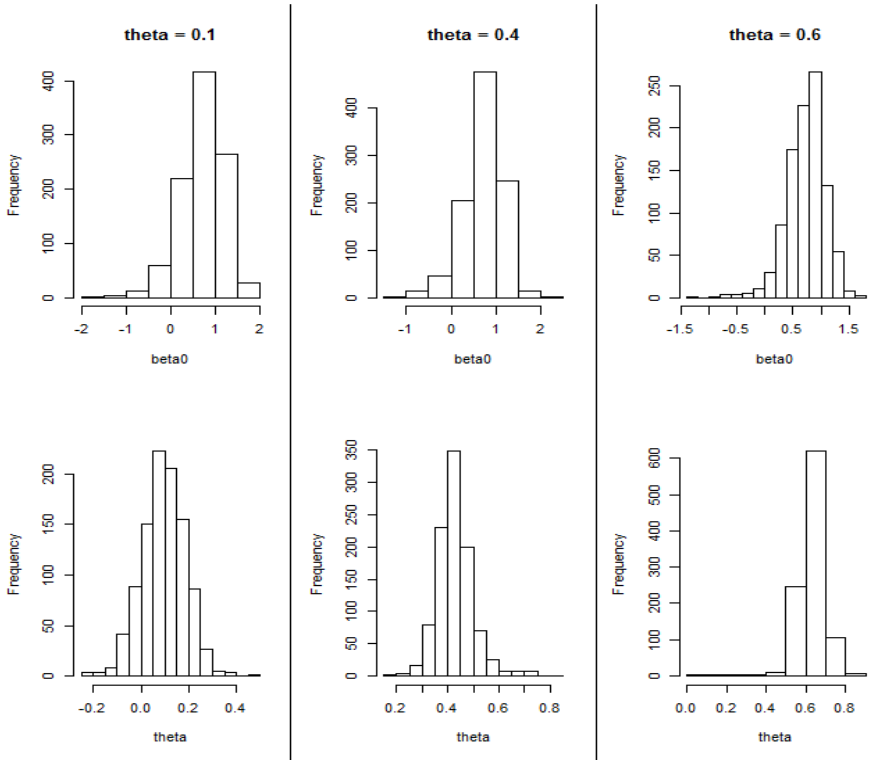


Figure 2 shows the histograms of the simulations for the parameter θ in Model 2. Regarding β_0, the parameter distribution was asymmetric in all cases and centered approximately in 0.8, the true value of this parameter. For parameter θ equal to 0.1 and 0.4 the distribution

was approximately symmetric around the true value of this parameter, while for $\theta = 0.6$ we observed some asymmetry in this parameter distribution.

3.2 Semiparametric Models

In this Monte Carlo study, the model was evaluated with parametric and non-parametric terms in the covariates. Again we simulated 1000 samples of size 100.

Model 3 was generated as:

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim Poisson(\mu_t) \\ ln(\mu_t) = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + ns(w_t, 3) &+ \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i} \\ \text{where}, ns(w_t, 3) &= \sum_{i=1}^{3} \alpha_i B_i(w_t), \end{aligned} \tag{23}$$

In the Model 3, the terms $B_i, i = 1, 2, 3$ composed the B-spline basis for natural cubic spline. The term w_t was the same of the non-parametric models, and the $x_{1,t}$ and $x_{2,t}$ were simulated time series. The series $x_{1,t}$ was generated from an ARMA(1,1) process with $\phi = 0.42$ and $\theta = 0.13$, and $x_{2,t}$ was an ARMA(1,2), with $\phi = 0.30$, $\theta_1 = -0.76$ and $\theta_2 = -0.17$. The autoregressive term was the same as in (21), evaluated when ϕ assumed values equal to 0.1, 0.4 and 0.6. The parameter λ was 0.5, and the parameters α and β were

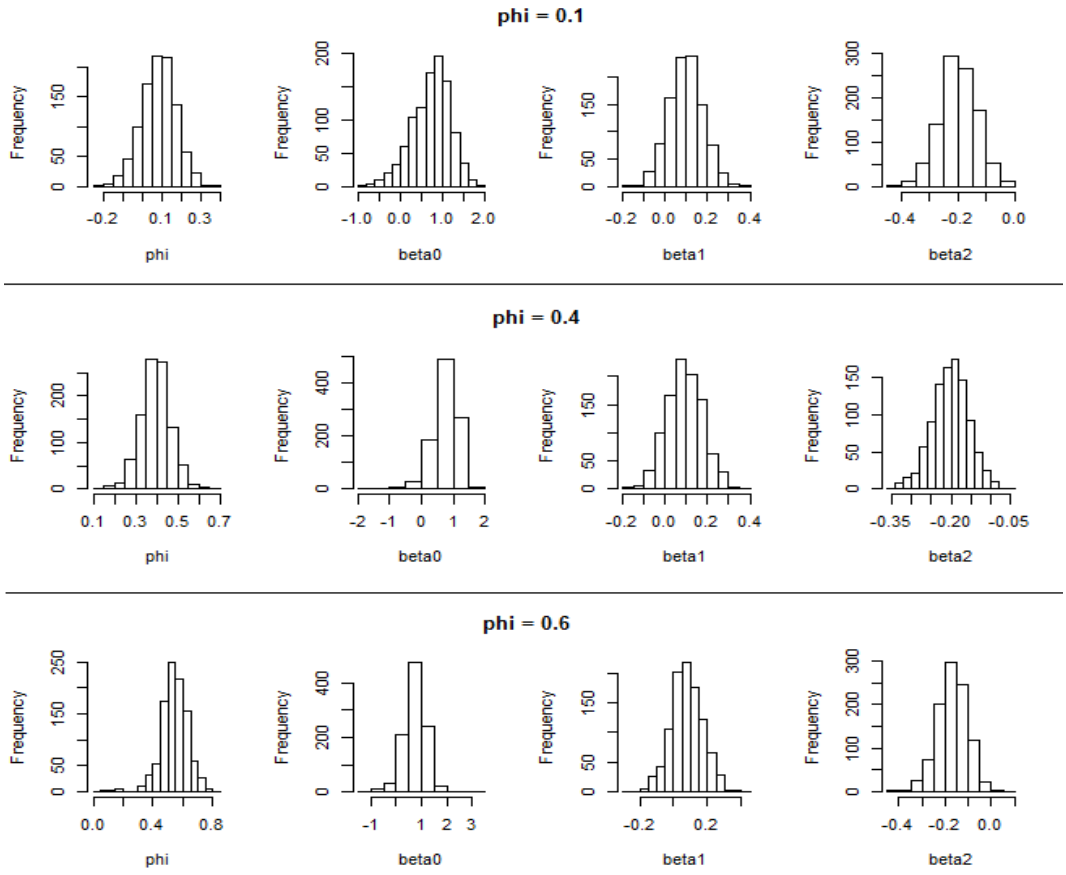
$$\beta_0 = 0.8,\ \beta_1 = 0.1,\ \beta_2 = -0.2,\ \alpha_1 = 0.5,\ \alpha_2 = -1.0,\ \alpha_3 = 0.8.$$

In Table 3, the results of parameter estimation were close to the original values of the parameters. In general, the values of MSE were small, but for smaller values of ϕ we found better estimates, however the estimates are worse (MSE larger) when $\phi = 0.6$.

Table 3: Results of parameter estimates in 1000 replicates of size 100 of Model 3

	ϕ	β_0	β_1	β_2	α_1	α_2	α_3
$\phi = \mathbf{0.1}$	0.0845 (0.0074)	0.7637 (0.1795)	0.0986 (0.0078)	-0.1953 (0.0036)	0.5455 (0.1264)	-0.9812 (0.9071)	0.8134 (0.0917)
$\phi = \mathbf{0.4}$	0.3927 (0.0055)	0.6907 (0.1852)	0.0945 (0.0067)	-0.1956 (0.0028)	0.5332 (0.1236)	-0.8401 (0.7623)	0.9035 (0.0985)
$\phi = \mathbf{0.6}$	0.5311 (0.0128)	0.7078 (0.2084)	0.0362 (0.0145)	-0.2443 (0.0049)	0.2491 (0.2183)	-0.6168 (0.9690)	0.9289 (0.1587)

Figure 3: Histograms of parameters estimates β's and ϕ - Model 3



In Figure 3 we have the histograms of the simulations for parameters ϕ and β's relative to the Model 3. The parameter distribution of ϕ was approximately symmetric around the true value when ϕ was equal 0.1 and 0.4. For ϕ equal to 0.6 the parameter distribution was asymmetric. Regarding β_0, the parameter distribution was asymmetric in all cases and centered around 0.8, the true value of this parameter. Already regarding β_1 and β_2, we

observed distributions approximately symmetric around the true values of the parameters, even for higher values of ϕ.

We present histograms with the empirical distribution of the other estimated parameters of the previous models in the Appendix.

4. Real Data Analysis

In this section, the GAM-ARMA was applied to monthly numbers of Chronic obstructive pulmonary disease (COPD) cases, popularly known as acute bronchitis, in the metropolitan area of Belo Horizonte, Brazil, between the years of 2007 and 2013 ($n = 84$). According to DATASUS, 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 Chronic obstructive pulmonary disease in Belo Horizonte.

Studies concerning air pollution in Belo Horizonte are relatively rare, and even more relating pollutant series with respiratory diseases. Information about the concentration of pollutants in this region is very limited, with all the series presenting a large number of missing observations. Thus, we had to perform some data imputation before proceeding to the analysis. Therefore this work brings not only a theoretical, but also a practical contribution to the study of this important problem that affects the health of inhabitants of the metropolitan area of Belo Horizonte.

Figure 4: Time series of the number of COPD cases and concetrations of air pollutants in the metropolitan area of Belo Horizonte, Brazil

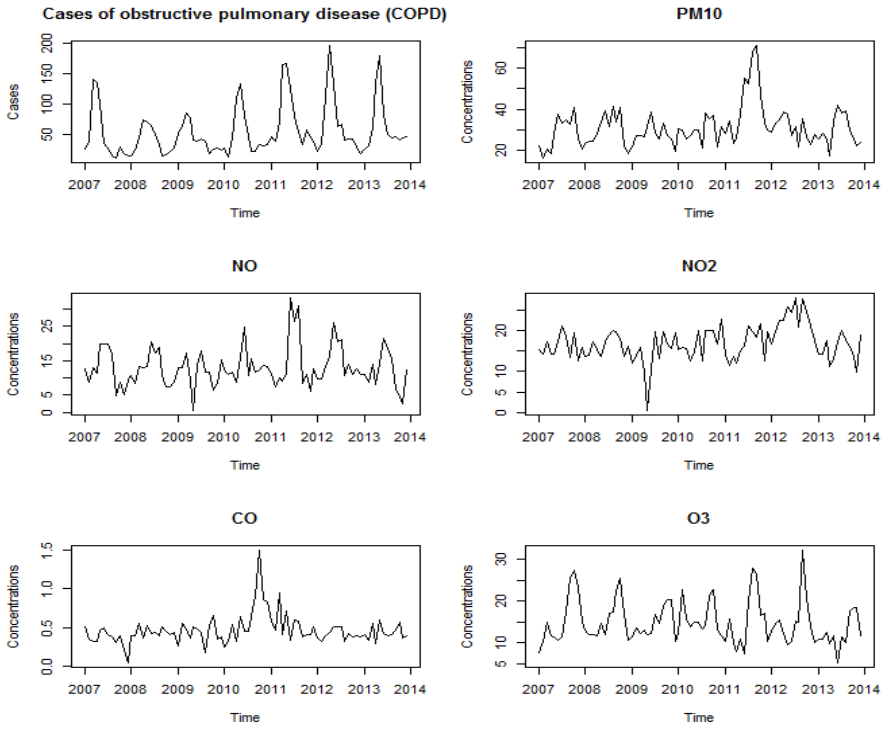


Figure 4 presents the time series of COPD cases between 2007 and 2013 and the time series of the concentrations of the following pollutants in the metropolitan area of Belo Horizonte: Particulate Matter (PM_{10}), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Carbon Monoxide (CO) and Ozone (O_3). Besides that, meteorological information, as temperature and relative humidity of the air, were investigated. The information about the concentration of pollutants and meteorological information was collected from State Environment and Water Resources Institute (IEMA) and the number of cases of COPD were collected from DATASUS (`http://datasus.saude.gov.br/`).

Some descriptive statistics are in Table 4. The average number of COPD cases are 54.93 monthly, with a standard deviation of 42.3. Moreover, the minimum number of monthly cases is 10.00 and the maximum 196.00. The average minimum temperature (T_{min}) is 17.89°C, with a standard deviation of 1.90°C. The average relative humidity (RH) is 61.60%, with a standard deviation of 7.48%.

Table 4: Descriptive statistics for the air pollutants concentrations on Chronic obstructive pulmonary cases

				Percentile			
	Mean	Standard Deviation	Minimum	25	50	75	Maximum
CO	0.44	0.19	0.02	0.35	0.41	0.51	1.50
PM_{10}	30.97	9.51	16.16	25.21	29.88	35.32	70.83
NO	12.97	5.80	0.57	9.05	12.01	15.52	33.11
NO_2	16.59	4.18	0.52	13.74	16.07	19.76	27.90
O_3	15.038	5.33	5.294	11.49	13.77	17.44	32.10
T_{min}	17.89	1.90	13.87	16.43	18.27	19.62	21.15
RH	61.60	7.48	45.83	56.17	61.60	67.55	77.63
Cases	54.93	42.3	10.00	27.00	41.00	66.00	196.00

In Table 5 we present the correlation matrix among air pollutants and the number of Chronic obstructive pulmonary cases. We can observe a significant correlation among some air pollutants, and between air pollutants and meteorological variables. The pollutants O_3 and NO presented the highest correlation with the response variable COPD.

Table 5: Correlation among pollutants and Chronic obstructive pulmonary disease cases

	COPD	CO	PM_{10}	NO	NO_2	O_3
COPD	1.00					
CO	0.03	1.00				
PM_{10}	0.08	0.15	1.00			
NO	0.30*	0.10	0.47	1.00		
NO_2	-0.03	0.15	0.29	0.52	1.00	
O_3	-0.35*	0.07	0.37	-0.19	0.24	1.00

* Correlations with COPD significant at a 5% level

4.1 Adjustment with a single pollutant

As discussed in the introduction, most of the works that evaluates the impact of air pollutants on human health employ the GAM model, but with only a single pollutant, once the air pollutants have time dependence and also posses interdependence among themselves (see

the papers of Pope *et al.* (1995), Schwartz (2000) and Bell *et al.* (2006)). To show the gain of GAM-ARMA over the GAM model in this case, we first adjusted these two procedures considering only the nitrogen monoxide (NO) as a covariate, once this pollutant posses a positive and significant correlation with the response variable (COPD). In the adjustment of all models in this section, the minimum temperature ($Temp$) and the relative humidity (RH) of the air were considered as having a non-linear relation with COPD. We observed an annual and semi-annual seasonality in the response variable, which was incorporated in the model with sine and cosine functions. The trend was also incorporated into the modelling.

The linear preditor of GAM-ARMA Poisson adjusted with a single pollutant was defined as

$$\begin{aligned} \eta_t = \beta_1 * NO_t + \beta_2 * sen12_t + \beta_3 * cos12_t + \beta_4 * sen6_t + \\ \beta_5 * cos6_t + \beta_6 * trend_t + ns(Temp_t, 3) + ns(RH_t, 3) + \sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}, \end{aligned} \tag{24}$$

and the GAM model follow the same structure in (24) without $\sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}$, the autoregressive moving average component. The chose of the optimal number of knots was based on the sample size, for this, as recommended in Section 2.2, we tested adjustments with three and four knots, and comparing the AIC the best model was obtained with just three knots.

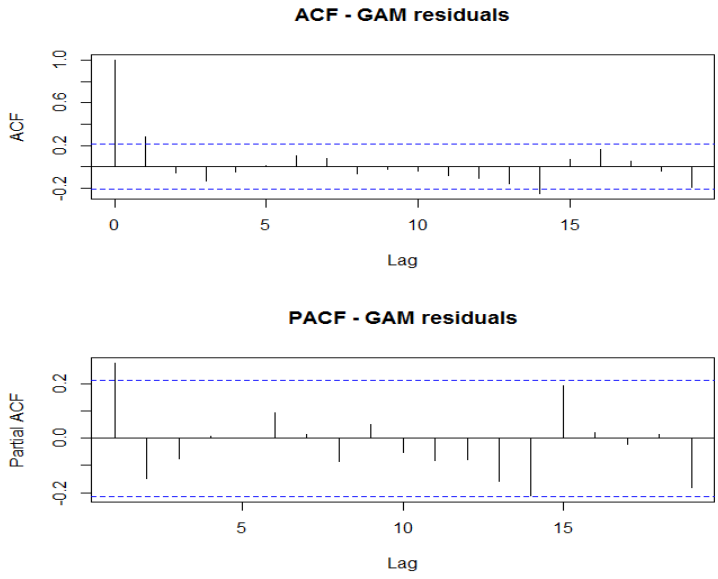
We first present the estimates of the fitted GAM model (Table 6), where all coefficients were significant (p-value < 0.05). The criterion to compare the goodness-of-fit was the BIC, and for this model the value obtained was 1297.514.

Table 6: Results of the GAM model to estimate the effect of NO concentration on the number of Chronic obstructive pulmonary disease cases

Variable	Estimates	Standard Error	p-value
NO	0.0545	0.0032	0.0000
sen12	0.2470	0.0415	0.0000
cos12	-0.5562	0.0611	0.0000
sen6	-0.3137	0.0306	0.0000
cos6	-0.2522	0.0326	0.0000
trend	0.0096	0.0007	0.0000

Figure 5 shows the ACF and PACF graphs for the residuals in the GAM model. The white noise was not obtained, once there are peaks outside the confidence bands in the initial lags. This behavior indicates the need for an adjustment considering the autoregressive structure of the series. For this, we adjusted the proposed model GAM-ARMA.

Figure 5: ACF and PACF of residuals - GAM model



Applying the GAM-ARMA methodology, the best fit was obtained with a GAM-AR(1), which means the coefficient of order 1 in the autoregressive polynomial was significant. Table 7 shows the estimates of the adjusted model, with all coefficients significant (p-value < 0.05).

The BIC obtained was 1155.059, which is a lower value than those obtained with the GAM model with only a single covariate (NO).

The ACF and PACF plots in Figure 6 reveal a good adjustment of GAM-AR(1) model, once we obtained white noise residuals, with a smaller BIC.

Figure 6: ACF and PACF of residuals - GAM-AR(1) model

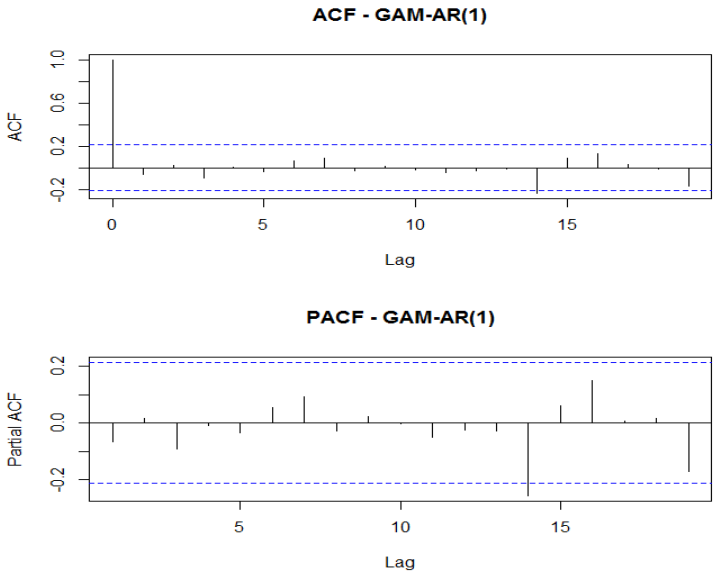
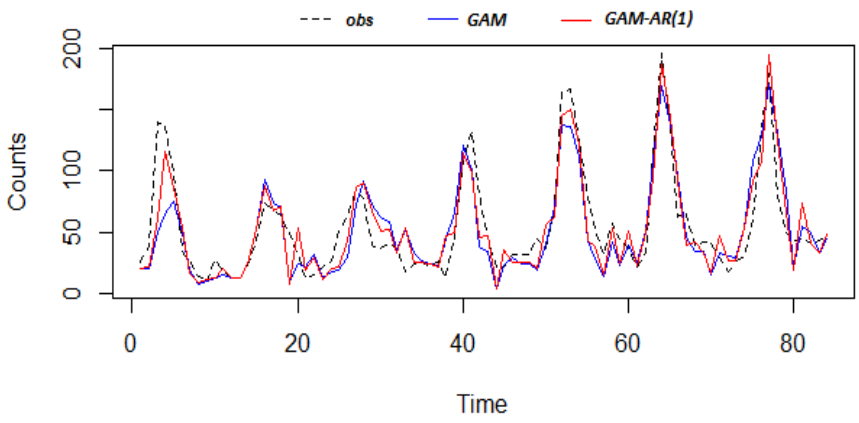


Table 7: Results of GAM-AR(1) model to estimate the effect of NO concentration on the number of Chronic obstructive pulmonary disease cases

Variable	Estimates	Standard Error	p-value
NO	0.0515	0.0029	0.0000
sen12	0.3271	0.0499	0.0000
cos12	-0.5221	0.0615	0.0000
cos6	-0.2324	0.0362	0.0000
trend	0.0127	0.0010	0.0000
ϕ_1	0.0700	0.0053	0.0000

The comparison among the adjustments of GAM and GAM-AR(1) models in Figure 7 revealed that the fitting of GAM-AR(1) described better the number of COPD cases.

Figure 7: Comparing the fit of GAM-AR(1) and GAM models



We also calculate the relative risk (RR) for the NO pollutant, as this is an important information for the regulatory agencies to measure the impact of this component in the population health. Table 8 presents the comparison of the relative risks and 95% intervals for interquartile variation in the NO pollutant among the GAM and GAM-AR(1) models. The estimate of RR per interquartile variation (ξ) in the pollutant concentrations X_j, $j = 1, ..., k$, is

$$\widehat{RR}_{X_j}(\xi) = exp(\widehat{\beta}_j \xi), \tag{25}$$

where $\widehat{\beta}_j$ is the estimated coefficient of the j-th pollutant. The confidence interval (CI) is given by

$$CI(RR_{X_i}(\xi)) = exp(\widehat{\beta}_j \xi \pm z_{\alpha/2} se(\widehat{\beta}_j)\xi), \tag{26}$$

where $z_{\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the standard normal distribution, and $se(\widehat{\beta}_j)$ is the standard error of $\widehat{\beta}_j$.

Table 8: Comparison of the Relative Risk and 95% confidence intervals for an interquartile variation of the NO concentrations among GAM and GAM-AR(1) models

NO	Relative Risk	CI
GAM	1.0627	[1.0553;1.0702]
GAM-AR(1)	1.0591	[1.0524;1.0658]

In Table 8, the RR estimates were significant for the NO pollutant, which means that this pollutant contributed significantly to the increase in the number of Chronic obstructive pulmonary disease cases. Comparing the relative risks (RR) of GAM and GAM-AR(1) we can observe that the RR is slightly smaller for the GAM-AR(1) model. As the adjustment obtained with this model was better, i.e., smaller BIC and with noise residuals, the relative risk of GAM-AR(1) is more reliable.

4.2 Adjustment with all pollutants

The study including only a single pollutant is very restrictive, mainly when there is more information available about other pollutants that can impact the number of respiratory diseases. However, a high number of covariates can lead to identification problems, and the correlation between them may imply multicollinearity. A possible solution to this problem is the Principal Component Analysis (PCA), which is a methodology where the variance and covariance of a random vector are explained through linear combinations of the original variables (Pearson, 1901), and the combinations called Principal Components (PC), are not correlated with each other. Wang and Pham (2011) proposed an hybrid method called GAM-PCA considering the combined use of the PCA technique and the GAM model without considering the temporal effect in the estimation.

The main problem with this approach is that PCA requires independent observations. Zamprogno (2013), showed that the Principal Components are autocorrelated if the covariates are autocorrelated because they preserve the structure of the original variables. To

remove the time correlation structure, Matteson and Tsay (2011), and Hu and Tsay (2014) applied the Vector Autoregression (VAR) to the covariates and realized the PCA in the residuals of the model. Based on this, Souza *et al*. (2018) proposed a model in which the time dependency of data is removed through the VAR methodology, then Principal Components are derivated from the residuals of VAR, and finally, the GAM model is applied to the PCs as covariates. Although this procedure presents a gain over the methodologies employed in the area literature, it still does not include the autocorrelation structure present in the response variable.

To analyze the impact of the air pollutants under study on the occurrence of Chronic obstructive pulmonary disease in Belo Horizonte, we have followed a similar procedure to that proposed by Souza *et al*. (2018), but adding the moving average autoregressive structure described in Section 2.2. We call this model a GAM-PCA-ARMA. For this, we removed the time dependency of the pollutant series using the VAR methodology, then derived the Principal Components (PC) from the residuals of the VAR model. Finally, we applied the proposed methodology to the data set of COPD with PCs, meteorological variables, sine and cosine functions, and trend as covariates.

Table 9 presents the results of applying the PCA to the residuals of the pollutant series obtained from the VAR model. The first three components correspond to 81.16% of the total variability. Following the parsimony criterium, the simplest model, with the first three principal components (PC's) as covariates, can handle with the complex structure of the data. As a complement, a cluster division was performed for each component group. In Table 9, the (*) symbol indicates the possible clusters for each principal component.

Table 9: Results of factor loadings and statistics applying PCA for the pollutants

	PC1	PC2	PC3	PC4	PC5
Standard deviation**	1.4063	1.0979	0.9355	0.8052	0.5416
Proportion of variance	0.3956	0.2411	0.1750	0.1297	0.0587
Cumulative proportion of variance	**0.3956**	**0.6366**	**0.8116**	0.9413	1.0000
CO	-0.3194	0.3321	0.8270*	0.3198	-0.0383
PM_{10}	-0.4823	0.1953	-0.4924*	0.6256	0.3088
NO	-0.5486*	-0.4651	-0.0627	0.0143	-0.6918
NO_2	-0.5751*	-0.1540	0.0923	-0.5886	0.5390
O_3	-0.1834	0.7821*	-0.2473	-0.3994	-0.3661

** Standard deviation is the square root of the eigenvalue

The linear preditor of GAM-ARMA Poisson adjusted with the first three principal components was defined as

$$\eta_t = \beta_1 * PC1_t + \beta_2 * PC2_t + \beta_3 * PC3_t + \beta_4 * sen12_t + \beta_5 * cos12_t + \beta_6 * sen6_t + \beta_7 * cos6_t + \beta_8 * trend_t + ns(Temp_t, 3) + ns(RH_t, 3) + \sum_{i=1}^{\infty} \tau_i \varepsilon_{t-i}, \tag{27}$$

and the GAM model follows the same structure in (27) without the autoregressive moving average component. As in section 4.1 the chose of the optimal number of knots was based on the sample size and the best model was obtained with three knots.

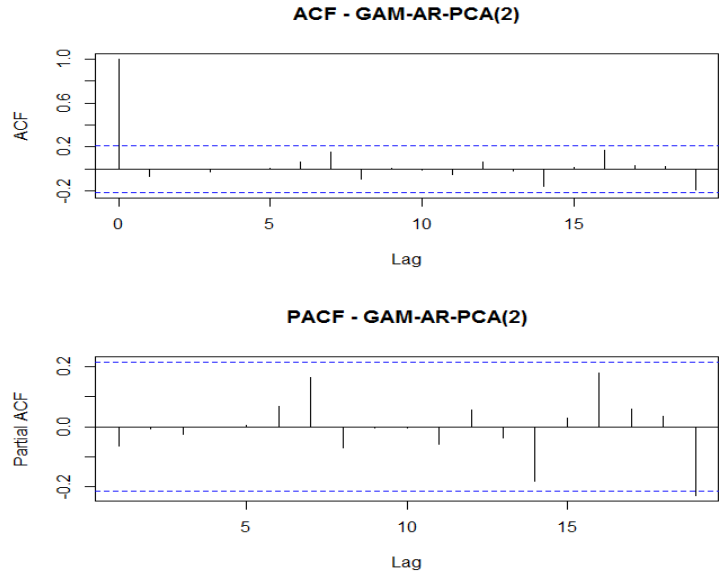
Applying the proposed methodology, the best fit was obtained with a GAM-PCA-AR(2), where the coefficients of order 1 and 2 in the autoregressive polynomial were significant. Table 10 reveals that all coefficients were significant (p-value < 0.05) and the BIC was 1257.422.

Table 10: Results of GAM-AR-PCA(2) model to estimate the effect of pollutants concentrations on the number of Chronic obstructive pulmonary disease cases

Variable	Estimates	Standard Error	p-value
PC1	-0.0758	0.0131	0.0000
PC2	0.0400	0.0183	0.0287
PC3	-0.1699	0.0199	0.0000
sen12	0.1626	0.0544	0.0028
cos12	-0.9668	0.0595	0.0000
sen6	-0.2339	0.0347	0.0000
cos6	-0.1927	0.0398	0.0000
trend	0.0162	0.0020	0.0000
ϕ_1	0.0724	0.0077	0.0000
ϕ_2	0.0351	0.0095	0.0000

Figure 8 presents the ACF and PACF graphs, revealing a good adjustment of the GAM-PCA-AR(2), once we found white noise.

Figure 8: ACF and PACF of residuals - GAM-AR-PCA(2)



To asses the gain of our model over the one proposed by Souza *et al.* (2018), i.e., without the autocorrelation structure, we adjusted their model to the same data set. Table 11

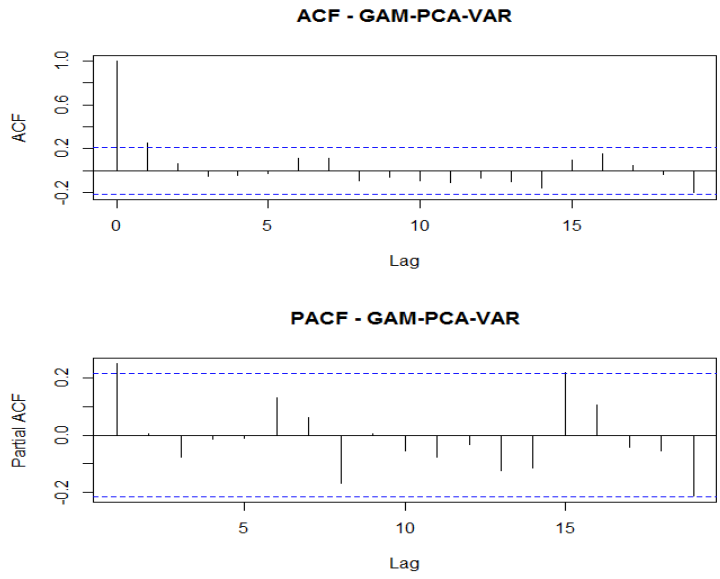
presents the results, where we can see that the coefficients of "PC2" and "sen12" were not significant. The ACF and PACF graphs in Figure 8 reveal that the residuals are not white noise. Added to this, the BIC was equal to 1397.059, larger than the one obtained with the GAM-PCA-AR(2) model. Then, we can conclude that the addition of the autoregressive component leads to a better fit of the data.

Table 11: Results of GAM-PCA-VAR model to estimate the effect of pollutants concentrations on the number of Chronic obstructive pulmonary disease cases

Variable	Estimates	Standard Error	p-value
PC1	-0.1217	0.0125	0.0000
PC2	0.0234	0.0209	0.263
PC3	-0.2222	0.0225	0.0000
sen12	0.0582	0.0396	0.141
cos12	-0.9291	0.0564	0.0000
sen6	-0.2213	0.0302	0.0000
cos6	-0.2894	0.0348	0.0000
trend	0.0093	0.0007	0.0000

The ACF and PACF graphs in Figure 9 reveal that the residuals were not white noise. Added to this, the BIC of GAM-PCA-VAR was larger than obtained with the GAM-AR-PCA(2). Then, we can conclude that the addition of the autoregressive component leads to the best adjustment of the data.

Figure 9: ACF and PACF of residuals - GAM-PCA-VAR



In Figure 10 we have the comparison among the adjustments of GAM-PCA-VAR and GAM-AR-PCA(2) models, revealing that the fitting of GAM-AR-PCA(2) described better the number of COPD cases.

Figure 10: Comparing the fit of GAM-AR-PCA(2) and GAM-PCA-VAR models

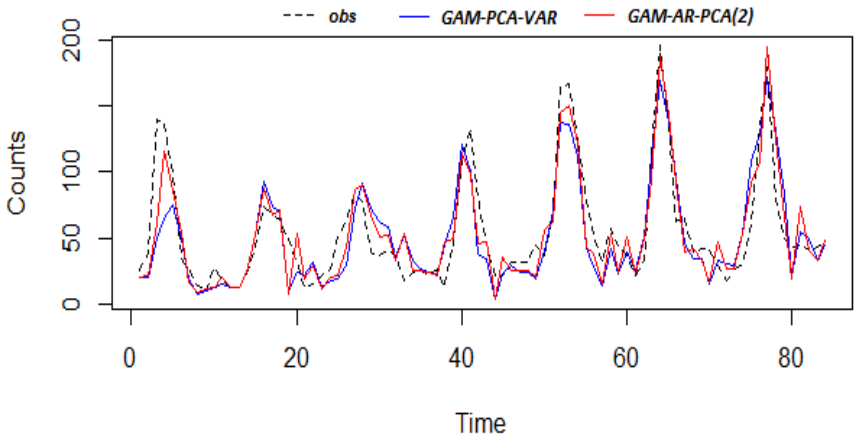


Table 12 presents the comparison among the relative risks (RR) of the GAM-AR-PCA(2) and GAM-PCA-VAR models. In this case, the estimate of RR per interquartile variation (ξ) in the pollutant concentrations X_j, $j = 1, ..., k$, is

$$\widehat{RR}^{*}_{X_j}(\xi) = exp(\widehat{\beta}^{*}_{j}\xi). \tag{28}$$

The estimated coefficient of the j-th pollutant, $\widehat{\beta}^{*}_{j}$, is given by

$$\widehat{\beta}^{*}_{j} = \sum_{i=1}^{l} \widehat{a}_{ji}\widehat{\upsilon}_{i}, \tag{29}$$

where l is the number of Principal Components (PC), $\widehat{\upsilon}_i$, $i = 1, ..., l$ is the estimated coefficient of the i-th PC, and $\widehat{a}_i$ are the first l-th estimated eigenvectors.

The confidence interval (CI) is given by

$$CI(RR^{*}_{X_i}(\xi)) = exp(\widehat{\beta}^{*}_{j}\xi \pm z_{\alpha/2}se(\widehat{\beta}^{*}_{j})\xi), \tag{30}$$

where $z_{\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution, and the standard error of $\widehat{\beta}^{*}_{j}$ is

$$se^2(\widehat{\beta}^{*}_{j}) = \sum_{i=1}^{r} \widehat{a}^{2}_{ji}se^2(\widehat{\upsilon}_i). \tag{31}$$

The RR estimates were significant for the PM_{10}, NO, NO_2 and O_3, which means that these pollutants contributed significantly to the increase in the number of Chronic obstructive pulmonary disease cases. Moreover, the risk relative estimates decreases with the addition of the autoregressive structure.

Table 12: Comparison of the Relative Risk and 95% confidence intervals for an interquartile variation of the pollutant concentrations

	$\widehat{RR}$	$\widehat{RR}^*$
CO	0.9235 (0.8986;0.9492)	0.8995 (0.8724;0.9275)
PM_{10}	1.1386 (1.1113;1.1667)	1.1914 (1.1604;1.2231)
NO	1.0383 (1.0131;1.0640)	1.0811 (1.0532;1.1097)
NO_2	1.0302 (1.0076;1.0533)	1.0647 (1.0419;1.0881)
O_3	1.1007 (1.0649;1.1376)	1.1108 (1.0699;1.1532)

$\widehat{RR}$: GAM-AR-PCA(2) and $\widehat{RR}^*$: GAM-PCA-VAR

5. Conclusion

In this work, we proposed a new class of models, called GAM-ARMA, which is an extension of the GAM based on GLARMA methodology. The structure of GAM-ARMA allows the fitting of non-parametric and semiparametric models, accommodating covariates with non-linear and linear relation with the response variable in count data with time dependence. The time structure is modelled by an autoregressive and moving average component. We presented the estimation procedure both the parametric and non-parametric components.

To evaluate the accuracy of parameter estimation of the proposed model, two simulation studies were performed. First, we analyzed models with just non-linear covariates from time series following the Poisson distribution. In the second study, we evaluated a model with linear and non-linear covariates, also generated under the Poisson distribution. In both cases, we observed estimates close of the real values of the parameters. However, the simulation studies revealed that the values of autoregressive and moving average components, ϕ and θ respectively, can influence the accuracy of estimates (especially the parameter ϕ). In general, as the value of these components increases, we have the worst estimates.

The proposed model was also applied in an epidemiological dataset. In this real data analysis, we studied the impact of air pollutants and meteorological factors in the monthly

numbers of Chronic obstructive pulmonary disease (COPD) cases, in Belo Horizonte, Brazil. For this, we proposed two approaches: first, following the previous studies we adjusted a model with a single air pollutant as covariate, once the pollutants presented correlation among them, and this may imply identification problems and multicollinearity. In the second approach, we adjusted a model with more than one pollutant, based on the methodology proposed by Souza *et al.* (2018).

In the approach with a single pollutant (NO) as covariate, we adjusted a GAM and GAM-ARMA model to compare the effects of the autocorrelation structure in the modelling. The results revealed that the GAM-AR(1), with the coefficient of order 1 in the autoregressive polynomial being significant, presented the best fit. This model presented smaller BIC and white noise residuals, which was not observed in the GAM model. Besides that, we calculate relative risks (RR) and 95% confidence intervals for an interquartile variation of the pollutants concentrations. The RR analysis revealed that the NO contributed significantly to the increase of COPD cases in Belo Horizonte. Moreover, the RR value decreased when the autoregressive term was added.

The second approach considered the following pollutants: PM_{10}, NO, NO_2, CO and O_3. The autocorrelation and multicolinearity present in the pollutants were modelled using PCA and VAR models. The best fit was obtained with the GAM-PCA-AR(2) with the coefficients of order 1 and 2 in the autoregressive polynomials significant. This model presented white noise residuals and smaller BIC in comparison with the model that did not include the autocorrelation structure of the response variable. The RR analysis revealed that the PM_{10}, NO, NO_2 and O_3 contributed significantly to the increase of COPD cases, and once again the addition of the autoregressive component contributed to decreasing the relative risks. This suggests that without this component, the RR can be overestimated. The results revealed that the proposed model provides better adjustments than the conventional models applied in count time series.

References

1. Ahlberg, J.H., Nilson, E.N. and Walsh, J.L. (1967) *The Theory of Splines and Their Application*. Academic Press Inc., New York.

2. Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

3. Al-Osh, M. A., Alzaid, A. A. (1987). First order integer valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, 8(3), 261-275.

4. Aldrin, M. and Hobk Haff, I. (2005). Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmospheric Environment*, Vol 39, pp. 2145-2155.

5. Belusic, A., Herceg-Bulic, I. and Klaic, Z. (2015). Using a Generalized additive model to quantify the influence of local meteorology on air quality in Zagreb. *Geofizika*, Vol 32, pp. 47-77.

6. Benjamin, M. A., Rigby, R. A., Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical association*, 98(461), 214-223.

7. Bell M., Kim J., Dominici F. (2006). Potential confounding of particulate matter on the short-term association between ozone and mortality in multisite time-series studies. *Environ Health Perspect*, 1591-1595.

8. Box, G. E., Jenkins, G. M. (1976). *Time series analysis*, Revised Edition. San Francisco.

9. Chen, R. J., Chu C., Tan, J., Cao, J., Song, W., Xu, X., Jiang, C., Ma W., Yang, C., Chen, B., Gui, Y. and Kan, H. (2010) Ambient air pollution and hospital admission in Shanghai, China. *J. Hazard. Mater.*, 181(1-3), 234-240.

10. Davis, R. A., Dunsmuir, W. T., Streett, S. B. (2003). Observation driven models for Poisson counts. *Biometrika*, 90(4), 777-790.

11. De Boor, C. (1978). *A practical guide to Splines*. Springer, Berlin.

12. Dierckx, P. (1993). *Curve and surface fitting with Splines*. Springer, Berlin.

13. Dockery, D. and Pope, C. (1996). Epidemiology of Acute Health Effects: Summary of time series study. In Richard Wilson and John Spengler, eds., *Particles in our air*. Cambridge, MA: Harvard University Press.

14. Eilers, P. and Marx, B. (1996). Flexible Smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89-121.

15. Friedman, J. and Silverman, B. (1989). Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, 31(1), 3-21.

16. Harrel, F. E. (2004). *Bioestatistical Modeling*. Nashvile TN USA.

17. Harvey, A.C., Fernandes, C. (1989). Time series models for count or qualitative observations. *Journal of Business Economic Statistics*, 7(4), 407-417.

18. Hastie, T.J.; Tibshirani, R.J. (1990) *Generalized additive models*. London, Chapman Hall, 335.

19. Hastie, T.; Tibshirani, R. and Friedman, J. (2008) *The elements of statistical learning*. Standford, California. Springer, Second edition.

20. Heinen, A. (2003). Modelling Time Series Count Data: Na Autoregressive Conditional Poisson Model. *Munich Personal RePEc Archive*.

21. Hertel, T., Lee, H., Rose, S. and Sohngen, B. (2009). Modelling land use related greenhouse gas sources and sinks and their mitigation potential. *Economic Analysis of land use in global climate change policy*. Abingdon: Routledge, chapter 6.

22. Hu, Y. P. and Tsay, R. S. (2014) Principal volatility component analysis. *J. Bus. Econ. Stat.*, 32(2), 153-164.

23. Kooperberg, C. and Stone, C. (1991) A Study of Logspline Density Estimation. *Computational Statistics and Data Analysis*, 12, 327-347.

24. Kooperberg, C. and Stone, C. (1992) Logspline Density Estimation for Censored Dat. *Journal of Computational and Graphical Statistics*, 1, 301-328.

25. Matteson, D. S. and Tsay, R. S. (2011) Dynamic orthogonal components for multivariate time series. *J. Am. Stat. Assoc.*, 106(496), 1450-1463.

26. McGeehin, M. and Mirabelli, M. (2001) The potential impacts of climate variability and change on temperature-related morbidity and mortality in the United States. *Environ Health Perspect*, 109(2), 185-189.

27. McKenzie, E. (1985) Some simple models for discrete variate time series, *Water Resources Bulletin* 21, 645-650.

28. Nelder, J.A.; Wedderburn, R.W.M. (1972) Generalized linear models. *J. Roy.Statist. Soc.* Ser. A,135, 370-384.

29. Ostro, B. D., Eskeland, G. S., Sanchez, J. M. and Feyzioglu, T. (1999) Air pollution and health efects: A study of medical visits among children in Santiago, Chile. *Environ. Health Persp.*, 107(1), 69-73.

30. Ostro, B., Roth, L., Malig, B. and Marty, M. (2009) The Effects of Fine Particle Components on Respiratory Hospital Admissions in Children. *Environ. Health Persp.*, 117(3), 475-480.

31. O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1, 505-527.

32. O'Sullivan, F. (1988) Fast computation of fully automated logdensity and log-hazard estimators. *SIAM J. Sci. Statist. Comput.*, 9, 363-379.

33. Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559-572.

34. Pope, C., Thun, M., Namboodiri, M., Dockery, D., Evans, J., Speizer, F. and Heath, C. (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal Respiratory Critical Care Medicine*, Vol 151, 669-674.

35. Schwartz, J. (2000) Harvesting and long term exposure efects in the relationship between air pollution and mortality. *Am. J. Epidemiol.*, 151, 440-448.

36. Shephard, N. (1995) *Generalized Linear Autoregressions*. Technical report, Nuffield College, Oxford University.

37. 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.

38. Villeneuve, P., Burnett, G., Aronov, D., Shi, Y., Krewski, D., Goldberg, M., Hertzman, C. and Brook, J. (2003). A time-series study of air pollution, socioeconomic status, and mortality in Vancouver, Canada. *J Expo Anal Environ. Epidemiol.*, 13(6), 427-435.

39. Wang, Y. and 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.

40. Wegman, E. and Wright, I. (1983) Splines in Statistics. *Journal of the American Statistical Association*, Vol 78, 351-365.

41. Wood, S. (2003) *Generalized Additive Models: An Introduction With R*. Boca Raton, Florida. Chapman and Hall/CRC Press.

42. Wood, S. (2006) Thin plate regression splines. *Royal Statistical Society*. Vol 65, 95-114.

43. Yang, L., Qin, G., Zhao, N., Wang, C. and Song, G. (2012) Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *Medical Research Methodology*. Vol 12:165.

44. Zamprogno, B. (2013) PCA in time series with short and long-memory time series. *PhD Thesis at the Programa de Ps-Graduao em Engenharia Ambiental do Centro Tecnolgico*, UFES, Vitria, Brazil.

Appendix

Histograms of simulation studies

Model 1: Autoregressive Case

Figure 11: Histogram of Model 1 ($\phi = 0.1$)

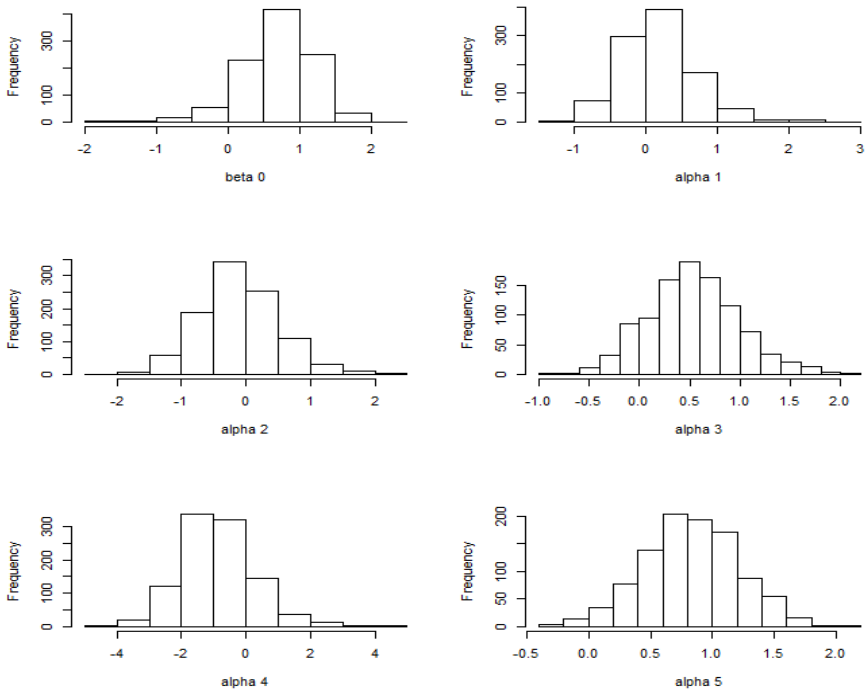


Figure 12: Histogram of Model 1 ($\phi = 0.4$)

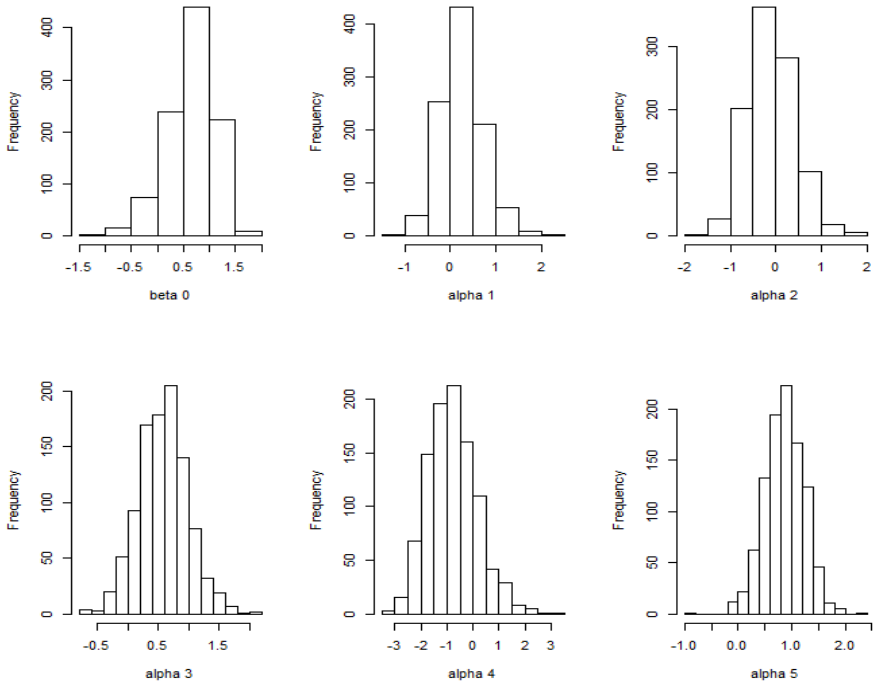
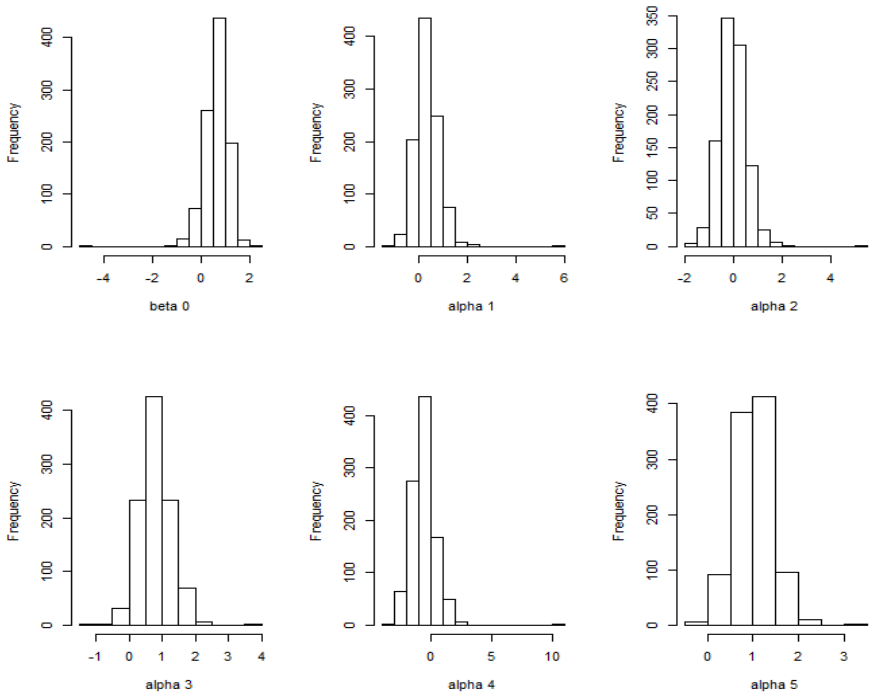


Figure 13: Histogram of Model 1 ($\phi = 0.6$)



Model 2: Moving Average Case

Figure 14: Histogram of Model 2 ($\theta = 0.1$)

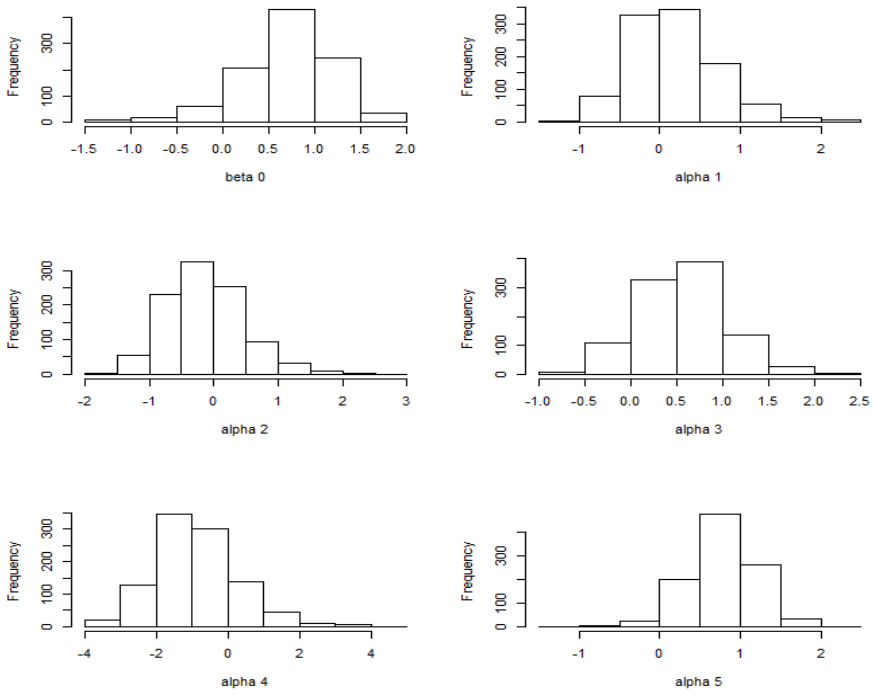


Figure 15: Histogram of Model 2 ($\theta = 0.4$)

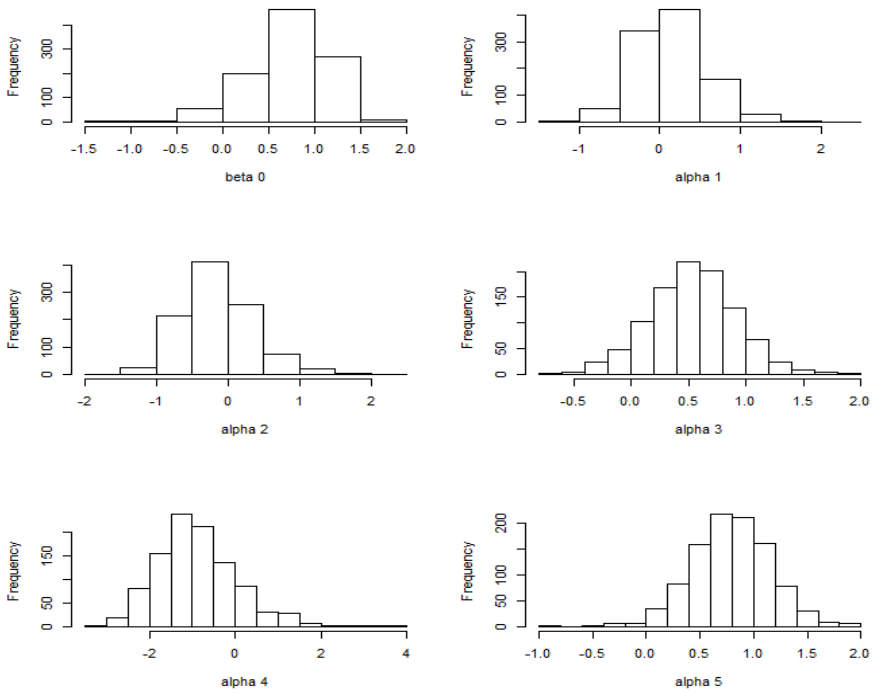
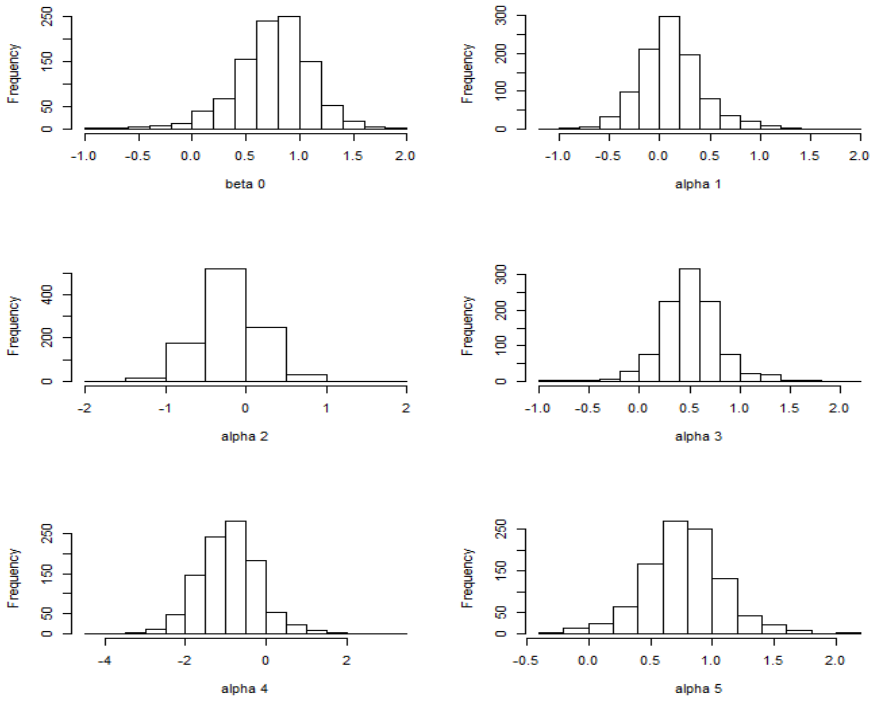


Figure 16: Histogram of Model 2 ($\theta = 0.6$)



Semiparametric Model - Model 3

Figure 17: Histogram of Model 3 ($\phi = 0.1$)

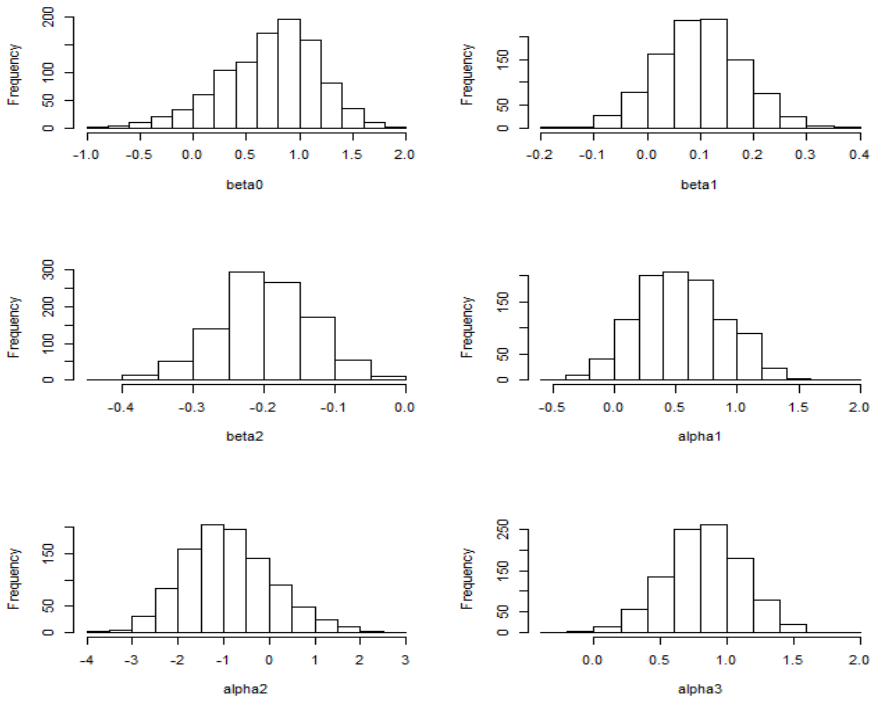


Figure 18: Histogram of Model 3 ($\phi = 0.4$)

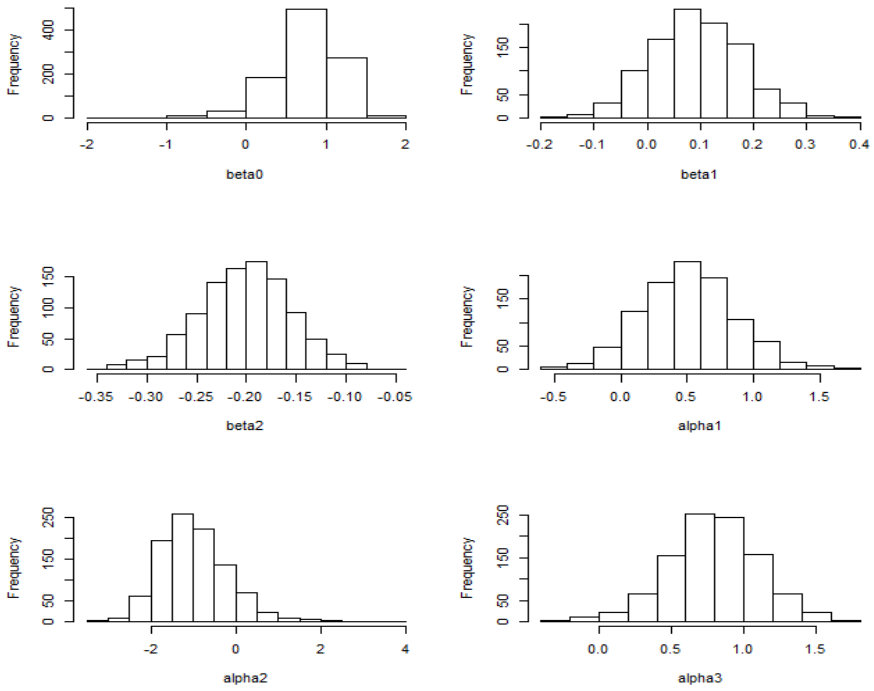
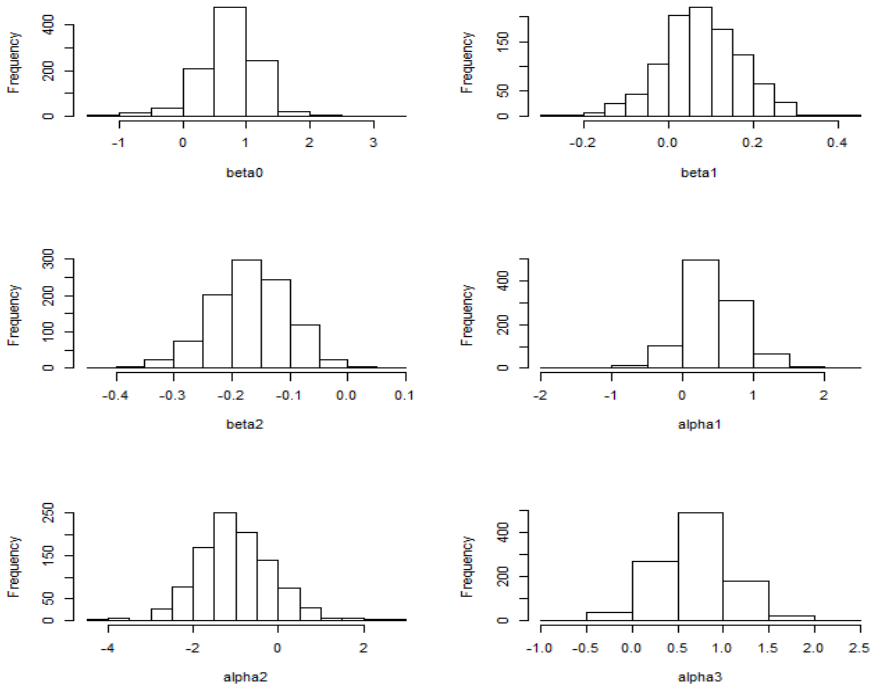


Figure 19: Histogram of Model 3 ($\phi = 0.6$)



Code in R for the simulation studies

Model 1: Autoregressive Case

```
# Run modifying the parameter phi for 0.1, 0.4 and 0.6 n=100
Temp_min <- as.data.frame(ns(DataRD_filter$Tmin,df=5))
colnames(Temp_min) <-
c("Temp_min_1","Temp_min_2","Temp_min_3","Temp_min_4","Temp_min_5")
Intercept <- rep(1,100) X <- as.matrix(Temp_min) XI <-
as.matrix(cbind(Intercept,Temp_min)) XI <- as.data.frame(XI)
beta_0 <- 0.8 alpha_1 <- 0.1 alpha_2 <- -0.2 alpha_3 <- 0.5
alpha_4 <- -1.0 alpha_5 <- 0.8 n <- 100 N <- 1000 W <- matrix(0,
nrow=n, ncol=N) Y <- matrix(0, nrow=n, ncol=N) mu <- matrix(0,
nrow=n, ncol=N) Z <- matrix(0, nrow=n, ncol=N) for (t in 1:N){
  mu[1,t] <- exp(W[1,t])
} beta <- beta_0 + alpha_1*Temp_min$Temp_min_1 +
alpha_2*Temp_min$Temp_min_2 +
  alpha_3*Temp_min$Temp_min_3 + alpha_4*Temp_min$Temp_min_4 + alpha_5*Temp_min$Temp_min_
lambda <- 0.5 phi <- 0.1 betas <- matrix(0, nrow=N, ncol=ncol(XI))
tetas <- matrix(0, nrow=N, ncol=1) phis <-  matrix(0, nrow=N,
ncol=1) for (i in 1:N) {
  for (t in 2:n) {
    W[t,i] <- beta[t] + phi*(Z[t-1,i]+((Y[t-1,i]-exp(W[t-1,i]))/exp(lambda*W[t-1,i])))
    mu[t,i] <- exp(W[t,i])
    Y[t,i] <- rpois(1,mu[t,i])
    Z[t,i] <- phi*(Z[t-1]+((Y[t-1,i]-exp(W[t-1,i]))/exp(lambda*W[t-1,i])))
  }
```

```
} W <- data.frame(W) mu <- data.frame(mu) Y <- data.frame(Y) XI <-
as.matrix(cbind(Intercept,Temp_min)) AR1 <- rep(0,N) for (i in
1:N) {
  AR1<-GLARMA.IC(Y[,i], preditora=XI, Vphi=c(1), Vteta=NULL, lambda=0.5)
  betas[i,] <- AR1$betas
  phis[i,] <- AR1$phi
  print(i)
}
```

Model 2: Moving Average Case

```
# Run modifying the parameter theta to 0.1,0.4 and 0.6 n=100
Temp_min <- as.data.frame(ns(DataRD_filter$Tmin,df=5))
colnames(Temp_min) <-
c("Temp_min_1","Temp_min_2","Temp_min_3","Temp_min_4","Temp_min_5")
cor(Temp_min) X <- as.matrix(Temp_min) Intercept <- rep(1,100) XI
<- as.matrix(cbind(Intercept,Temp_min)) XI <- as.data.frame(XI)
beta_0 <- 0.8 alpha_1 <- 0.1 alpha_2 <- -0.2 alpha_3 <- 0.5
alpha_4 <- -1.0 alpha_5 <- 0.8 n <- 100 N <- 1000 W <- matrix(0,
nrow=n, ncol=N) Y <- matrix(0, nrow=n, ncol=N) mu <- matrix(0,
nrow=n, ncol=N) Z <- matrix(0, nrow=n, ncol=N) for (t in 1:N){
  mu[1,t] <- exp(W[1,t])
} beta <- beta_0 + alpha_1*Temp_min$Temp_min_1 +
alpha_2*Temp_min$Temp_min_2 +
  alpha_3*Temp_min$Temp_min_3 + alpha_4*Temp_min$Temp_min_4 + alpha_5*Temp_min$Temp_min_
lambda <- 0.5 theta <- 0.1 betas <- matrix(0, nrow=N,
ncol=ncol(XI)) tetas <- matrix(0, nrow=N, ncol=1) for (i in 1:N) {
```

```
  for (t in 2:n) {
    W[t,i] <- beta[t] + theta*(Y[t-1,i]-exp(W[t-1,i]))/exp(lambda*W[t-1,i])
    mu[t,i] <- exp(W[t,i])
    Y[t,i] <- rpois(1,mu[t,i])
  }
} W <- data.frame(W) mu <- data.frame(mu) Y <- data.frame(Y) XI <-
as.matrix(cbind(Intercept,Temp_min)) MA1 <- rep(0,N) for (i in
1:N) {
  MA1<-GLARMA.IC(Y[,i], preditora=XI, Vphi=NULL, Vteta=c(1), lambda=0.5)
  betas[i,] <- MA1$betas
  tetas[i,] <- MA1$teta
  print(i)
}
```

Mixed Model - Model 3

```
# Run modifying the parameter phi for 0.1, 0.4 and 0.6 n=100
sim_PM10 <- arima.sim(n = n, list(ar = c(0.42), ma = c(0.13)))
sim_O3 <- arima.sim(n = n, list(ar = c(0.30), ma =
c(-0.76,-0.17))) PM10 <- as.vector(sim_PM10) O3 <-
as.vector(sim_O3) Polut <- cbind(PM10,O3) Polut <-
as.data.frame(Polut) Temp_min <-
as.data.frame(ns(DataRD_filter$Tmin,df=3)) colnames(Temp_min) <-
c("Temp_min_1","Temp_min_2","Temp_min_3") Intercept <- rep(1,100)
X <- as.matrix(Temp_min) XI <-
as.matrix(cbind(Intercept,Polut$PM10,Polut$O3,Temp_min)) XI <-
as.data.frame(XI) beta_0 <- 0.8 beta_1 <- 0.1 beta_2 <- -0.2
```

```
alpha_1 <- 0.5 alpha_2 <- -1.0 alpha_3 <- 0.8 n <- 100 N <- 1000 W
<- matrix(0, nrow=n, ncol=N) Y <- matrix(0, nrow=n, ncol=N) mu <-
matrix(0, nrow=n, ncol=N) Z <- matrix(0, nrow=n, ncol=N) for (t in
1:N){
  mu[1,t] <- exp(W[1,t])
} beta <- beta_0 + beta_1*XI$‘Polut$PM10‘ + beta_2*XI$‘Polut$O3‘ +
alpha_1*XI$Temp_min_1 + alpha_2*XI$Temp_min_2 +
  alpha_3*XI$Temp_min_3
lambda <- 0.5 #phi <- as.matrix(rep(0.1,n)) phi <- 0.1 betas <-
matrix(0, nrow=N, ncol=ncol(XI)) tetas <- matrix(0, nrow=N,
ncol=1) phis <-  matrix(0, nrow=N, ncol=1) for (i in 1:N) {
  for (t in 2:n) {
    W[t,i] <- beta[t] + phi*(Z[t-1,i]+((Y[t-1,i]-exp(W[t-1,i]))/exp(lambda*W[t-1,i])))
    mu[t,i] <- exp(W[t,i])
    Y[t,i] <- rpois(1,mu[t,i])
    Z[t,i] <- phi*(Z[t-1]+((Y[t-1,i]-exp(W[t-1,i]))/exp(lambda*W[t-1,i])))
  }
} W <- data.frame(W) mu <- data.frame(mu) Y <- data.frame(Y) XI <-
as.matrix(cbind(Intercept,Polut$PM10,Polut$O3,Temp_min)) AR1 <-
rep(0,N) for (i in 1:N) {
  AR1<-GLARMA.IC(Y[,i], preditora=XI, Vphi=c(1), Vteta=NULL, lambda=0.5)
  betas[i,] <- AR1$betas
  phis[i,] <- AR1$phi
  print(i)
}
```