Generalized additive model with principal component analysis: An application to time series of respiratory disease and air pollution data

Juliana B. de Souza, Valderio A. Reisen, Glaura C. Franco, Marton Ispany, Pascal Bondon, Jane Meri Santos

Generalized additive model with principal component analysis: An application to time series of respiratory disease and air pollution data

Juliana B. de Souza

Federal University of Esprito Santo, Vitória, Brazil

Valdério A. Reisen

Federal University of Esprito Santo, Vitória, Brazil and CentraleSupélec, Gif sur Yvette, France

Glaura C. Franco

Federal University of Minas Gerais, Belo Horizonte, Brazil

Márton Ispány

University of Debrecen, Debrecen, Hungary

Pascal Bondon

CNRS, CentraleSupélec, University of Paris-Saclay, Gif sur Yvette, France

Jane Meri Santos

Federal University of Esprito Santo, Vitória, Brazil

Summary. Environmental epidemiological studies of the health effects of air pollution frequently utilize the generalized additive model (GAM) as the standard statistical methodology, considering the ambient air pollutants as explanatory covariates. Although exposure to air pollutants are multidimensional, the majority of these studies considers only a single pollutant as a covariate in the GAM model. This model restriction may be due to the fact that the pollutant variables do not only possess serial dependence, but also interdependence amongst themselves. In an attempt to convey a more realistic model, we propose here the hybrid GAM-PCA-VAR model, which is the combination of the principal component analysis (PCA) and GAM along with a vector autoregressive (VAR) process. The PCA is used to eliminate the multicollinearity amongst the pollutants while the VAR model is used to handle the serial correlation of the data in order to produce white noise processes as covariates in the GAM. Some theoretical and simulation results of the proposed methodology are discussed, with special attention to the effect of the time correlation of the covariates on the PCA and, consequently, on the estimates of the parameters in the GAM model and on the relative risk (RR), which is a commonly used statistical quantity to measure the impact of the covariates, especially the pollutants, on the population health. As a main motivation to the proposed methodology, a real data set is analysed with the aim to quantify the association between respiratory disease and air pollution concentrations, especially, PM_{10}, SO_2, NO_2, CO and O_3. The empirical results show that the GAM-PCA-VAR model is able to remove the autocorrelations from the principal components. In addition, this method produces estimates of the RR, for each pollutant, which are not affected by the serial correlation present in the data. This, in general, leads to more pronounced values of the estimated risk compared to the standard GAM model, indicating, for this study, an

increase of almost 5.4% in the risk of PM_{10}, one of the most important pollutants which is usually associated with adverse effect on human health.

Keywords: Generalized additive model; Multicollinearity; Principal component analysis; Relative risk; Serial correlation; Vector autoregressive model

1. Introduction

The impact of air pollutants on human well-being has motivated the study and control of atmospheric pollution, which affects human health even for low levels of air pollutants concentrations within air quality guidelines suggested by the World Health Organization, WHO (2006). Many studies have found significant association between daily pollutant concentration levels and hospital admissions for respiratory and cardiovascular diseases, see Schwartz (2000), Ostro *et al.* (1999), Chen *et al.* (2010), among others. The adverse effects of atmospheric pollutants on human health are a source of concern to environmental and public health regulatory agencies. Population studies and epidemiological research have been used to identify these adverse health effects and to guide the development of practices and legislation to control emissions and air quality.

The generalized additive model (GAM) with a Poisson marginal distribution has been the most widely applied method to measure and quantify the nonlinear association between adverse health effects and covariates such as ambient concentrations of air pollutants and meteorological conditions, mainly because it allows for nonparametric adjustments of nonlinear confounding effects of seasonality and trends.

In spite of its widespread use, many authors claim that care is needed when applying the GAM to time series. The fit can be affected, for instance, by a wrong choice of the number of degrees of freedom in the smooth component, by the presence of the autocorrelation in the series under study, among others, see for example, the recent paper by Dionisio *et al.* (2016). Some works that aim to solve these problems include Dominici *et al.* (2002), who proposed a correction on the degrees of freedom in the smooth component; Dominici *et al.* (2006), Lall *et al.* (2011), Michelozzi *et al.* (2007), who have used lag distributed models to relate the response variable to lagged values of a time-dependent predictor; and Figueiras *et al.* (2005), Ramsey *et al.* (2003), who have proposed some approaches to control the problem of concurvity (the nonlinear dependency that can remain among the covariates). Additionally, most of the papers in the epidemiological research area related to the study of the association between pollution and adverse health effects usually consider only one pollutant, while the population under study is exposed to a complex mixture of pollutants; a broad discussion of the impact of correlated measurement errors in time series on the relative risk estimates is recently given in Dionisio *et al.* (2016). The choice of a simple model may be, in general, due to the fact that the pollutants are linearly time correlated variables, which implies in biased regression estimates since the presence of multicollinearity (the linear dependency among the covariates) can inflates the variance of the estimators. This model restriction may not provide the true picture of the scenario in a real problem. As a result, this incorrect analysis may lead to serious consequences on the health of the population under study such as, for example, a false-positive conclusion of the pollution health risk.

One way to circumvent the problem of multicollinearity is to perform a principal component analysis (PCA) on the pollutants covariance matrix. The PCA is a multivariate statistical technique and it is generally used to reduce the dimensionality of a set of data while preserving, as much as possible, the variability in the covariates, see Johnson and Wichern (2007). Evaluating the adverse health effects of a combination of pollutants may be easier to interpret and more feasible than isolating the effects of a single pollutant. Some authors have explored this relevant research direction. For example, Roberts and Martin (2006) evaluated how the pollutants PM_{10}, O_3, SO_2, NO_2 and CO affect health, where the issue of multicollinearity was handled using the PCA. The authors also developed a PCA supervised method in which the relationship between the covariates (the pollutants) and deleterious health effects are determined before the covariates are inserted into the regression model. Recently, Wang and Pham (2011) studied the combined effects of pollutants on daily mortality using a PCA and a robust method. The RR estimates of the results were more significant when the multivariate PCA technique was used. Nevertheless, application of the PCA technique generally requires the data to be obtained through independent replications. All the time series considered in this paper are supposed to be stationary (including the covariates). As the principal components are linear combinations of the covariates, their properties are linearly transferred to the principal components. Therefore, the use of PCA to perform statistical inferences on time-correlated covariates, such as ambient concentration of atmospheric pollutants, should be further examined.

Zamprogno (2013) has addressed this issue by using theoretical and empirical methods to determine the effect of neglecting the time correlation of the covariates in the PCA technique. The author showed that the principal components are autocorrelated if the covariates are also autocorrelated. The principal components contain the time structure of the covariates and must therefore be used judiciously in the regression analysis. To remove the temporal correlation structures of the PCA, Zamprogno (2013) has suggested filtering the series by using a multivariate ARMA model in the pollution variables before performing any statistical analysis using PCA. In the same context, Matteson and Tsay (2011) and Hu and Tsay (2014) have applied VAR models to remove the serial correlation of time series of stock returns before carrying out the PCA analysis on the residual of the VAR model. The use of Box-Jenkins methodology to eliminate the serial correlation in the data was also considered in Campbell (1994) which discusses the relationship between sudden infant death syndrome with environmental temperature using time regression for count with Poisson marginal distribution.

In the current study, the multicollinearity issue is solved using the PCA on the pollutants, with the obtained components being used as covariates in the GAM. This procedure is called GAM-PCA. Additionally, the problem associated with the presence of autocorrelation in the principal components when applying the GAM is circumvented by using a vector autoregressive (VAR) model to the time series of covariates before obtaining the principal components. This new model is called here GAM-PCA-VAR. These two models are formulated theoretically as probabilistic latent variable models in Section 2. The GAM-PCA and GAM-PCA-VAR models are compared to the conventional GAM by means of adequate goodness-of-fit statistics and, also, in terms of the RR estimate, a commonly used tool to measure the impact of the covariates, especially

the pollutants, on the population health. Some results related to the proposed methods and the effect of autocorrelated covariates on the PCA are theoretically and empirically discussed. In addition, the estimate of the RR is evaluated for each model in a real data problem. The objective of estimating the RR is to verify if there is any change on this statistic due to the characteristics of the covariates under study, such as temporal correlation, among others. As a main result of this paper, we find that the two procedures (GAM-PCA and GAM-PCA-VAR) evidenced larger relative risk estimates than those obtained using the conventional GAM. A simulation study demonstrates that the inter and autocorrelation found in the explanatory pollutant variables may be responsible for this divergence. This is an important evidence that prevents the use of the standard GAM, from the epidemiological point of view, since the time-series phenomena present in the explanatory pollutant variables can produce unrealistic risk impacts on the health of the population under study, that is, this may indicates a false-positive result.

The paper is organized as follows. Section 2 presents the statistical models addressed here, such as GAM, PCA and VAR, in some detail. Section 3 discusses some simulations results and the analysis of a real data set. Section 4 concludes the work.

2. Methodology: GAM, PCA, VAR and Relative Risk

In this section, we present the methodology employed to relate the covariates to the count time series under study. As there are both linear and nonlinear relationship between the explanatory variables and the response, a GAM model is used. The procedures are implemented using count data with Poisson distribution, as this is widely used in practical problems.

We also present, in some detail, the PCA and VAR methodologies, in order to explain how these procedures are linked to solve problems that can occur with data exhibiting both multicollinearity and serial correlation in the explanatory variables.

2.1. Generalized Additive Models

The generalized additive model (GAM), see Hastie and Tibshirani (1990), with a Poisson marginal distribution is typically used to relate a discrete outcome variable with a set of covariates in the epidemiological area, for example, to quantify the association between health problems and air pollution concentrations. The GAM model is widely used to describe non-linear correlations among the variables of interest, see for example, Schwartz (2000), Ostro *et al.* (1999), Chen *et al.* (2010).

Let $\{Y_t\} \equiv \{Y_t\}_{t\in\mathbb{Z}}$ be a count time series, i.e., it is composed of non-negative integer valued random variables. The conditional distribution of Y_t, given the past $\mathcal{F}_{t-1}$ which contains the available information up to time $t-1$, is characterized by the weights $p(y_t \,|\, \mathcal{F}_{t-1}) := P(Y_t = y_t \,|\, \mathcal{F}_{t-1})$ where $y_t \in \{0, 1, \ldots\}$. If Y_t has the conditional Poisson distribution with mean μ_t, then

$$p\left(y_t; \mu_t \,|\, \mathcal{F}_{t-1}\right) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!}, \quad y_t = 0, 1, \ldots$$

Thus the conditional log-likelihood function of the mutually conditionally independent

random variables $Y_1, \ldots, Y_n$ is given by

$$\ell(\boldsymbol{\mu}) := \sum_{t=1}^{n} \ln p\left(Y_t; \mu_t \,|\, \mathcal{F}_{t-1}\right) \propto \sum_{t=1}^{n} (Y_t \ln \mu_t - \mu_t), \tag{1}$$

where the vector $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n)^\top$ depends on the covariates and the parameters of the process $\{Y_t\}$. Let $\boldsymbol{X_t} = (X_{1t}, \ldots, X_{pt})^\top$ be the vector of covariates of dimension p at time t, where $\top$ denotes the transpose, which may include past values of Y_t and other auxiliary variables, such as the pollutants and confounding variables (i.e., trends, seasonality and meteorological variables, among others). In the following, $X_{1t}, \ldots, X_{qt}$ denote the pollutants, while $X_{(q+1)t}, \ldots, X_{pt}$ denote the confounding variables at time t ($q \leq p$).

The relation between Y_t and the vector $\boldsymbol{X_t}$ of covariates is obtained by setting, see, for example, Kedem and Fokianos (2002),

$$\ln(\mu_t) = \sum_{j=0}^{q} \beta_j X_{jt} + \sum_{j=q+1}^{p} f_j(X_{jt}) \quad \text{with} \quad q \leq p,$$

where $(\beta_0, \boldsymbol{\beta}^\top)$ with $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_q)^\top$ is the vector of the coefficients to be estimated (β_j is the coefficient of the j-th covariate), and f_j is a smoothing function of an appropriate function space for the j-th confounding variable, for example, the temperature or the humidity variable. Moreover, β_0 denotes the curve intercept and is associated with $X_{0t} = 1$ for all t. For the sake of simplicity it is assumed that the pollutant covariates are centered. The aforementioned model is usually referred to as a semi-parametric model because it involves parametric and non-parametric functions. The parameters of the parametric functions are generally estimated using maximum likelihood or quasi-likelihood methods, by optimizing the log-likelihood defined by equation (1), with the asymptotic properties given in Kedem and Fokianos (2002). The non-parametric functions are evaluated using "splines", "loess" or moving average functions, among others, see Friedman (1991) and Wahba (2001).

The RR is frequently used in epidemiological studies to measure the impact of atmospheric pollutant concentrations on the health of the exposed population. The RR of a pollutant covariate X_j, $j = 1, ..., q$, is defined as the relative change in the expected count of respiratory disease events per ξ unit change in the covariate while keeping the other covariates fixed. More precisely, we have

$$\mathrm{RR}_{X_j}(\xi) := \frac{\mathrm{E}(Y \mid X_j = \xi, X_i = x_i, i \neq j)}{\mathrm{E}(Y \mid X_j = 0, X_i = x_i, i \neq j)},$$

see formula (8) in Baxter *et al.* (1997). For Poisson regression, RR does not depend on the values x_i, $i \neq j$, of the other covariates and is given by

$$\mathrm{RR}_{X_j}(\xi) = \exp\left(\beta_j \xi\right).$$

RR is often called relative rate or rate ratio, see, e.g., page 265 in Dalgaard (2008). Note that for binary outcomes, the RR is defined as the ratio of probabilities that an event will occur following a certain exposure/non-exposure to a risk factor, see Zou (2004).

RR can also be interpreted in this study as the ratio of probabilities that a patient is suffering from respiratory diseases per ξ unit change in a pollutant covariate. The RR and its approximate confidence interval (CI) at an α significance level of a covariate X_j, $j = 1, \ldots, q$, in the GAM model with Poisson marginal distribution are estimated as follows:

$$\widehat{\mathrm{RR}}_{X_j}(\xi) = \exp\left(\hat{\beta}_j \xi\right) \quad \text{and} \quad \mathrm{CI}(\mathrm{RR}_{X_j}(\xi)) = \exp\left(\hat{\beta}_j \xi \mp z_{\alpha/2}\, \mathrm{se}(\hat{\beta}_j)\xi\right),$$

where ξ is the variation in the pollutant concentration (for example, a value of $10\mu g/m^3$ of interquartile variation), $\hat{\beta}_j$ is the estimated coefficient for the pollutant X_j being studied with standard error $\mathrm{se}(\hat{\beta}_j)$, and $z_{\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution. At an α significance level, the hypothesis to be tested is defined as $H_0 : \mathrm{RR}_{X_j} = 1$ against $H_1 : \mathrm{RR}_{X_j} > 1$ where $\mathrm{RR}_{X_j} := \mathrm{RR}_{X_j}(1)$, i.e., the RR of unit change in X_j. The rejection of H_0 statistically implies that the respective pollutant has a significant adverse health effect.

2.2. Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical technique that aims, in general, to reduce the dimensionality of a data matrix space through linear transformations of the original variables. The correlation among the variables implies the occurrence of multicollinearity in the regression models. In this study, the PCA technique is used to circumvent the problem of pollutants that are correlated with each other. In general, the whole variability of a system determined by q variables can only be explained using all the q principal components. However, a large part of this variability can be explained using a lower number r of components ($r \leq q$), see Johnson and Wichern (2007).

Consider the following pairs of eigenvalues/eigenvectors of the covariance matrix $\Sigma_{\boldsymbol{X}}$ of the random vector $\boldsymbol{X} = (X_1, ..., X_q)^\top : (\lambda_1, \boldsymbol{a}_1), (\lambda_2, \boldsymbol{a}_2), \ldots, (\lambda_q, \boldsymbol{a}_q)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$. The i-th principal component of $\Sigma_{\boldsymbol{X}}$ is given as follows:

$$Z_i = \boldsymbol{a}_i^\top \boldsymbol{X} = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{qi}X_q, \tag{2}$$

$i = 1, 2, \ldots, q$, where $a_{ji} = (\boldsymbol{a}_i)_j$, $i, j = 1, 2, \ldots, q$, with the properties

$$\mathrm{Var}(Z_i) = \boldsymbol{a}_i^\top \Sigma_{\boldsymbol{X}} \boldsymbol{a}_i = \lambda_i \qquad \text{and} \qquad \mathrm{Cov}(Z_i, Z_j) = \boldsymbol{a}_i^\top \Sigma_{\boldsymbol{X}} \boldsymbol{a}_j = 0,$$

$i, j = 1, 2, \ldots, q$, $i \neq j$, since the eigenvectors are orthogonal.

For a stationary vector time series $\{\boldsymbol{X}_t\} \equiv \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, $\boldsymbol{X}_t = (X_{1t}, ..., X_{qt})^\top$, with the covariance matrix $\Sigma_{\boldsymbol{X}}$, the principal components (PCs) are defined as $Z_{it} = \boldsymbol{a}_i^\top \boldsymbol{X}_t$, $i = 1, \ldots, q$, and

$$\mathrm{Cov}(Z_{it}, Z_{jt}) = \boldsymbol{a}_i^\top \mathrm{Cov}(\boldsymbol{X}_t, \boldsymbol{X}_t)\boldsymbol{a}_j = \boldsymbol{a}_i^\top \Gamma_{\boldsymbol{X}}(0)\boldsymbol{a}_j = \begin{cases} \lambda_i & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and

$$\mathrm{Cov}(Z_{it}, Z_{j(t+h)}) = \boldsymbol{a}_i^\top \mathrm{Cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t+h})\boldsymbol{a}_j = \boldsymbol{a}_i^\top \Gamma_{\boldsymbol{X}}(h)\boldsymbol{a}_j, \tag{4}$$

where $\Gamma_{\boldsymbol{X}}(h)$ denotes the autocovariance matrix function of $\{\boldsymbol{X}_t\}$ at lag h with $\Gamma_{\boldsymbol{X}}(0) = \Sigma_{\boldsymbol{X}}$. This result is proved in Zamprogno (2013).

Equation (3) shows that at zero lag the PCs are uncorrelated, while equation (4) demonstrates that PCA preserves the autocorrelation structures present in time-correlated covariates. That is, for all $i = 1, \ldots, q$, $Z_i \equiv \{Z_{it}\}_{t\in\mathbb{Z}}$ is a time series and the autocorrelation of Z_i, $\rho_{Z_i}(h) \neq 0$, $h = \pm 1, \ldots$, provided the eigenvector $\boldsymbol{a}_i$ is not in the nullspace of the autocovariance matrices $\Gamma_{\boldsymbol{X}}(h)$, $h \neq 0$, which holds clearly, for example, if these matrices have full rank. In addition, Z_i and Z_j, $j \neq i$, are cross-correlated, that is, $\rho_{Z_i,Z_j}(h) \neq 0$ for all $h = \pm 1, \pm 2 \ldots$

Thus, PCA must be used judiciously in time series regression models. We propose in Section 2.4 an alternative method to eliminate the autocorrelation of the principal components.

2.3. GAM-PCA - Generalized Additive Modelling and Principal Component Analysis

One of the research directions developed in this article is the combined use of the PCA technique and the GAM model, which is denoted here as the GAM-PCA model. This hybrid method was previously considered in Wang and Pham (2011) without taking into account the temporal effect in the model parameter estimates. Note that this model is also referred to as PCA-based GAM, see Zhao *et al.* (2014), where the model is applied to quantify the relationships between fish populations and their environment.

In the GAM-PCA model the covariates $Z_{1t}, \ldots, Z_{qt}$ generated by the PCA are linear combinations of the original variables $X_{1t}, \ldots, X_{qt}$. Mathematically, $Z_{it} = \boldsymbol{a}_i^\top \boldsymbol{X}_t$, similarly to (2), but the PCs are now time dependent for all $i = 1, \ldots, q$. These new covariates are used in the GAM model. Let $r \leq q$ and, considering the first r-th pairs of eigenvalues/eigenvectors of the covariance matrix $\Sigma_{\boldsymbol{X}}$, define the matrices $\Lambda_r := \mathrm{diag}\{\lambda_1, \ldots, \lambda_r\}$ and $A_r := (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$, i.e., the eigenvectors form columns of matrix A_r. One can see that A_r is an orthogonal matrix of dimension $q \times r$, i.e., $A_r^\top A_r = I_r$ where I_r is the identity matrix of dimension r. Moreover, $A_r^\top \Sigma_{\boldsymbol{X}} A_r = \Lambda_r$. Let $\Lambda = \Lambda_q$ and $A = A_q$. Then Λ_r is the top-left block of Λ of size $r \times r$ and A_r consists of the first r columns of A, see, e.g., page 11 in Jolliffe (2002). Note that any linear combination of the first r-th new covariates can be expressed as the linear combination of the original covariates in the following way:

$$\sum_{i=1}^{r} \upsilon_i Z_{it} = \sum_{j=1}^{q} \sum_{i=1}^{r} \upsilon_i a_{ji} X_{jt} = \sum_{j=1}^{q} \beta_j^* X_{jt}, \tag{5}$$

where $\boldsymbol{\upsilon} := (\upsilon_1, \ldots, \upsilon_r)^\top$ and $\boldsymbol{\beta}^* := (\beta_1^*, \ldots, \beta_q^*)^\top$ are vectors of dimensions r and q, respectively, and the relation between vectors $\boldsymbol{\upsilon}$ and $\boldsymbol{\beta}^*$ is given by $\boldsymbol{\beta}^* = A_r \boldsymbol{\upsilon}$, and thus $\boldsymbol{\upsilon} = A_r^\top \boldsymbol{\beta}^*$. (Note that $a_{ji} = (\boldsymbol{a}_i)_j$ where a_{ji} denotes the entry of the matrix A in the j-th row and i-th column and $(\boldsymbol{a}_i)_j$ denotes the j-th coordinate of the i-th eigenvector $\boldsymbol{a}_i$.) That is, in the GAM-PCA model, the new parameter vector $\boldsymbol{\beta}^*$ of the original covariates is in the range of matrix A_r. Then, the link function of the GAM-PCA model

using the first r-th PCs is given by

$$\begin{aligned}\mu_t(\upsilon_0, \boldsymbol{\upsilon}, A) &= \exp\left\{\sum_{i=0}^{r} \upsilon_i Z_{it} + \sum_{j=q+1}^{p} f_j(X_{jt})\right\} \\ &= \exp\left\{\upsilon_0 + \boldsymbol{\upsilon}^\top A_r^\top \boldsymbol{X}_t + \sum_{j=q+1}^{p} f_j(X_{jt})\right\}\end{aligned} \tag{6}$$

with $r \leq q \leq p$, where $\boldsymbol{X}_t := (X_{1t}, \ldots, X_{qt})^\top$ is the vector of covariates, υ_0 corresponds to the curve intercept with $Z_{0t} = 1$ for all t, $\boldsymbol{\upsilon}$ is the vector of coefficients of the first r-th principal components, and f_j's are the smoothing functions for the confounding variables (i.e., the temperature and the humidity in this study). In the definition of the link function, we denote only the parameters of the new PC covariates and the transformation matrix of the PCA.

The GAM-PCA model can be considered as a probabilistic latent variable model defined by

$$Y_t \,|\, \mathcal{F}_{t-1} \sim \text{Poi}(\mu_t) \quad \text{and} \quad \boldsymbol{X}_t = A\boldsymbol{Z}_t$$

with link function (6), where $\text{Poi}(\cdot)$ denotes the Poisson distribution, the latent variables $\{\boldsymbol{Z}_t\}$ form a vector white noise process of dimension q with diagonal variance matrix Λ, see Definition 11.1.2 in Brockwell and Davis (1991), and A is an orthogonal matrix of dimension $q \times q$. The quadruple $(\upsilon_0, \boldsymbol{\upsilon}, A, \Lambda)$ forms the parameters of the GAM-PCA model to be estimated. Clearly, the latent variables can be expressed as $\boldsymbol{Z}_t = A^\top \boldsymbol{X}_t$ for all t. Hence, GAM-PCA can also be interpreted as a two-stage model where, in the first stage, new variables (PCs) are derived by the PCA using the original covariates, and, in the second stage, GAM is fitted by using the first r-th new variables. If $\{\boldsymbol{X}_t\}$ is a Gaussian process, then the joint distribution of $(Y_t, \boldsymbol{X}_t)$ can be expressed as a product of a Poisson and a Gaussian distribution. Thus, given a sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, the log-likelihood, up to a constant, is derived as a hybrid sum of a Poisson and a Gaussian log-likelihood:

$$\ell(\upsilon_0, \boldsymbol{\upsilon}, A, \Lambda) \propto \sum_{t=1}^{n} (Y_t \ln \mu_t - \mu_t) - \frac{1}{2}\sum_{t=1}^{n} (A^\top \boldsymbol{X}_t)^\top \Lambda^{-1} (A^\top \boldsymbol{X}_t) - \frac{n}{2} \ln \det \Lambda, \tag{7}$$

where μ_t depends on the parameters by the link function (6). The parameters of the GAM-PCA model can be estimated, for example, by the maximum likelihood method. Since the log-likelihood (7) is rather complicated, the maximization with respect to its parameters is more complex, and a two-stage method is proposed. Firstly, the parameter matrices A and Λ are estimated by applying the PCA for the estimated covariance matrix $\widehat{\Sigma}_{\boldsymbol{X}}$. Secondly, the parameters υ_0 and $\boldsymbol{\upsilon}$ are estimated by fitting the GAM model with link function (6) using the first r-th PCs. Note that this procedure works without assuming any distribution assumption for the covariates. In case of Gaussian covariates the maximization of the Gaussian part of the log-likelihood (7) is equivalent to the application of PCA for these covariates. In the sequel, the assumption of normal distribution for covariates is used only in computing the standard information criteria

for model selection. The approach discussed above is similar to the principal component regression, see, for example, Chapter 8 in Jolliffe (2002), and it can be considered as a two-stage regression method, which is a procedure well-known in the econometric area, see Amemiya (1985).

In this context, the estimate of RR per ξ unit change in the pollutant concentration for the original covariate X_j, $j = 1, \ldots, q$, is given as follows:

$$\widehat{\mathrm{RR}}^{*}_{X_j}(\xi) = \exp\left(\hat{\beta}^{*}_{j}\xi\right), \tag{8}$$

where ξ is, for example, the interquartile variation. The term $\hat{\beta}^{*}_{j}$ is given by

$$\hat{\beta}^{*}_{j} := \sum_{i=1}^{r} \hat{a}_{ji}\hat{\upsilon}_i, \quad j = 1, \ldots, q, \tag{9}$$

where $\hat{\upsilon}_i$ is the estimated coefficient of the i-th PC in (6) and $\hat{a}_{ji} = (\hat{\boldsymbol{a}}_i)_j$ where $\hat{a}_{ji}$ is the entry of the matrix $\hat{A}$ in the j-th row and i-th column and $\hat{\boldsymbol{a}}_i$, $i = 1, \ldots, r$, are the first r-th estimated eigenvectors. Equation (9) can be easily derived using equation (5). Since the PCs are uncorrelated the standard error of $\hat{\beta}^{*}_{j}$ can be derived as

$$\mathrm{se}^2(\hat{\beta}^{*}_{j}) = \sum_{i=1}^{r} \hat{a}^2_{ji}\,\mathrm{se}^2(\hat{\upsilon}_i).$$

2.4. GAM-PCA-VAR - GAM-PCA and Vector Autoregressive Modelling

As previously discussed, the use of PCA for time series produces autocorrelations and cross-correlations among the principal components. In this paper, we suggest a procedure to eliminate the autocorrelations and cross-correlations of these components by applying a vector autoregressive moving average (VARMA) filter to the original data to obtain a white noise process, see, also, Greenaway-McGrevy *et al.* (2012). The proposed model, called here GAM-PCA-VAR, aims to eliminate the temporal correlation in order to obtain estimates of the regression parameters, and consequently RR estimates, which are free from the serial correlation present in the covariates that could lead to spurious analysis in real applications.

Let now $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{qt})^{\top}$, be a VARMA$(p^*, q^*)$ process defined as the solution to the following system, see Hamilton (1994):

$$\Phi(B)(\boldsymbol{X}_t - \boldsymbol{\gamma}) = \Theta(B)\boldsymbol{\varepsilon}_t, \tag{10}$$

where B is the delay operator, $\boldsymbol{\gamma}$ is a q dimensional vector and the innovation process $\{\boldsymbol{\varepsilon}_t\}$ is a q dimensional white noise with $\mathrm{E}(\boldsymbol{\varepsilon}_t) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}_t) = \Sigma_{\boldsymbol{\varepsilon}}$, where $\Sigma_{\boldsymbol{\varepsilon}}$ is a $q \times q$ variance matrix. The operators $\Phi(B) = I_q - \sum_{i=1}^{p^*} \Phi_i B^i$ and $\Theta(B) = I_q + \sum_{i=1}^{q^*} \Theta_i B^i$ are polynomial matrices of orders p^* and q^*, respectively, and Φ_i's and Θ_i's are matrices of constants with dimension $q \times q$. If $\det \Phi(z) \neq 0$ for all complex z such that $|z| \leq 1$ then the VARMA model (10) has exactly one stationary causal solution, see Theorem 11.3.1 in Brockwell and Davis (1991). Seasonal VARMA models are built using the same structure as in (10), but with the lag time being a multiple of the seasonal period.

The VAR(1) model is a particular case of the VARMA(p^*, q^*) model with $p^* = 1$ and $q^* = 0$. Without loss of generality, it is here assumed that $\boldsymbol{\gamma} = 0$. Therefore, the model in (10) can be written in the form

$$\boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t. \tag{11}$$

A VAR(1) process has unique stationary causal solution provided all the eigenvalues of $\boldsymbol{\Phi}$ are less than 1 in absolute value. In this case, the unique solution of the VAR(1) model can be expressed as the almost surely convergent infinite series $\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{\Phi}^j \boldsymbol{\varepsilon}_{t-j}$, see Example 11.3.1 in Brockwell and Davis (1991). The autocovariance matrix function of $\{\boldsymbol{X}_t\}$ is given by $\Gamma_{\boldsymbol{X}}(h) = \sum_{j=0}^{\infty} \boldsymbol{\Phi}^{j+h} \Sigma_{\boldsymbol{\varepsilon}} (\boldsymbol{\Phi}^\top)^j$, $h = 0, \pm 1, \ldots$. The identification and estimation procedures for model (10) are given in Hamilton (1994) and Brockwell and Davis (1991). The seasonal VAR(1) model with period s, usually denoted by SVAR$_s$(1), is an extension of Model (11) with a seasonal matrix autoregressive coefficient at lag s. This seasonal matrix has to satisfy similar stationary condition to the one of the VAR(1) model, see, for example, Brockwell and Davis (1991). In the following, the model proposed here, which combines PCA, VAR and GAM procedures, is discussed.

The GAM-PCA-VAR model is a combination of the VAR(1) model given in (11), where $\boldsymbol{X}_t$ represents the pollution variables at time t in the context of this paper, and GAM-PCA model by using the white noise error process of (11) as covariates. Mathematically, let $Z_{1t}, \ldots, Z_{qt}$ at time t be given by

$$Z_{it} = \boldsymbol{a}_i^\top \boldsymbol{\varepsilon}_t = \boldsymbol{a}_i^\top (\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1}), \qquad i = 1, \ldots, q, \tag{12}$$

where $(\lambda_i, \boldsymbol{a}_i)$, $i = 1, \ldots, q$, denote the eigenvalues/eigenvectors of the variance matrix $\Sigma_{\boldsymbol{\varepsilon}}$ of the white noise innovation in (11), and, therefore, the PCs vector $\boldsymbol{Z}_t$ has now uncorrelated components $Z_i \equiv \{Z_{it}\}$, $i = 1, \ldots, q$, and these components are white noise processes with variances λ_i, $i = 1, \ldots, q$, respectively. The impact of the VAR(1) filter in the GAM-PCA-VAR model is to eliminate the serial correlation present in the original pollutant covariates. Large positive values in a coordinate of the innovation $\boldsymbol{\varepsilon}_t$ indicate locally high environmental influence according to this pollutant at time t. On the contrary, large negative values indicate negligible influence. The link function of the GAM-PCA-VAR model using the first r-th PCs is defined by

$$\begin{aligned} \mu_t(\upsilon_0, \boldsymbol{\upsilon}, A, \boldsymbol{\Phi}) &= \exp\left\{ \sum_{i=0}^{r} \upsilon_i Z_{it} + \sum_{j=q+1}^{p} f_j(X_{jt}) \right\} \\ &= \exp\left\{ \upsilon_0 + \boldsymbol{\upsilon}^\top A_r^\top \boldsymbol{X}_t - \boldsymbol{\upsilon}^\top A_r^\top \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \sum_{j=q+1}^{p} f_j(X_{jt}) \right\}, \end{aligned} \tag{13}$$

which clearly shows that, in contrast to GAM-PCA, Y_t depends on both $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-1}$ demonstrating the presence of serial dependence in the GAM-PCA-VAR model.

The GAM-PCA-VAR model can also be considered as a probabilistic latent variable model defined by

$$Y_t \,|\, \mathcal{F}_{t-1} \sim \text{Poi}(\mu_t) \quad \text{and} \quad \boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + A\boldsymbol{Z}_t$$

with link function (13), where the latent variables $\{\boldsymbol{Z}_t\}$ form a vector white noise process of dimension q with diagonal variance matrix Λ, A is an orthogonal matrix of dimension $q \times q$, and Φ is a matrix of dimension $q \times q$. The quintuplet $(\upsilon_0, \boldsymbol{\upsilon}, A, \Lambda, \Phi)$ forms the parameters of the GAM-PCA-VAR model to be estimated. Clearly, the latent variable can be expressed as $\boldsymbol{Z}_t = A^\top(\boldsymbol{X}_t - \Phi \boldsymbol{X}_{t-1})$ for all t, see also equation (12). Hence, GAM-PCA-VAR can be interpreted as a three-stage model, where in the first stage the temporal dependence is eliminated by taking the new serially uncorrelated variable $\boldsymbol{\varepsilon}_t = \boldsymbol{X}_t - \Phi \boldsymbol{X}_{t-1}$ at time t; in the second stage new uncorrelated variables (PCs) $\{\boldsymbol{Z}_t\}$ are derived by using the PCA for the innovation process $\{\boldsymbol{\varepsilon}_t\}$; and in the third stage GAM is fitted by using the first r-th PCs as covariates. The order of models in the acronym GAM-PCA-VAR corresponds to these stages starting with the third one and finishing with the first one, which is generally accepted in the time series literature.

Under the assumption that the distribution of the innovation vector is multivariate normal, the conditional log-likelihood of the GAM-PCA-VAR model, given a sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, is derived as:

$$\ell(\upsilon_0, \boldsymbol{\upsilon}, A, \Lambda, \Phi) \propto \sum_{t=2}^{n} (Y_t \ln \mu_t - \mu_t) - \frac{1}{2} \sum_{t=2}^{n} \boldsymbol{\varepsilon}_t^\top A \Lambda^{-1} A^\top \boldsymbol{\varepsilon}_t - \frac{n-1}{2} \ln \det \Lambda, \tag{14}$$

where $\boldsymbol{\varepsilon}_t = \boldsymbol{X}_t - \Phi \boldsymbol{X}_{t-1}$ and μ_t depends on the parameters by the link function (13). Since the maximization of this log-likelihood is also rather computationally expensive, a three-stage estimation method is proposed: firstly, VAR(1) model is fitted to the original covariates by applying standard time series techniques; secondly, using PCA for the residuals defined by $\hat{\varepsilon}_t = \boldsymbol{X}_t - \hat{\Phi} \boldsymbol{X}_{t-1}$, $t = 2, \ldots, n$, where $\hat{\Phi}$ denotes the estimated autoregressive coefficient matrix in the fitted VAR(1) model, the first r-th PCs are computed; thirdly, GAM model is fitted using these PCs by maximizing the Poisson part of the log-likelihood (14). The relative risk of the GAM-PCA-VAR model, computed similarly to (8), is denoted here by $\widehat{\mathrm{RR}}^{**}$.

Remark 1. *Another model to GAM-PCA-VAR, called hereafter GAM-VAR-PCA can be derived by interchanging the order of VAR filter and PCA. Namely, the multicollinearity amongst the original covariates is eliminated by PCA firstly and then the serial dependence is handled by VAR modelling. More precisely, let A_r be defined as in Section 2.3 and $\boldsymbol{Z}_t^{(r)} = A_r^\top \boldsymbol{X}_t$ for all t. We fit a VAR(1) model to the r-dimensional process $\{\boldsymbol{Z}_t^{(r)}\}$, that is $\boldsymbol{Z}_t^{(r)} = \Psi_r \boldsymbol{Z}_{t-1}^{(r)} + \boldsymbol{W}_t^{(r)}$, where Ψ_r is a matrix of dimension $r \times r$ and $\{\boldsymbol{W}_t^{(r)}\}$, $\boldsymbol{W}_t^{(r)} = (W_{1t}^{(r)}, \ldots, W_{rt}^{(r)})^\top$, is an r-dimensional white noise process. The link function of the GAM-VAR-PCA model is*

$$\begin{aligned} \mu_t(\upsilon_0, \boldsymbol{\upsilon}, A_r, \Psi_r) &= \exp\left\{ \sum_{i=0}^{r} \upsilon_i W_{it}^{(r)} + \sum_{j=q+1}^{p} f_j(X_{jt}) \right\} \\ &= \exp\left\{ \upsilon_0 + \boldsymbol{\upsilon}^\top A_r^\top \boldsymbol{X}_t - \boldsymbol{\upsilon}^\top \Psi_r A_r^\top \boldsymbol{X}_{t-1} + \sum_{j=q+1}^{p} f_j(X_{jt}) \right\}, \end{aligned} \tag{15}$$

which looks like (13). *Nevertheless, there is an important difference between these two*

formulations. Whereas in the GAM-PCA-VAR model, the vector $\boldsymbol{Z}_t^{(r)} = (Z_{1t}, \ldots, Z_{rt})^\top$ *in* (13) *is a white noise process with uncorrelated components* $Z_i \equiv \{Z_{it}\}$, $i = 1, \ldots, r$, *in the GAM-VAR-PCA model, the vector* $\boldsymbol{W}_t^{(r)}$ *in* (15) *is also a white noise process but its components are not necessarily uncorrelated. Hence, the new covariates of the GAM-VAR-PCA model involved into the GAM model are no longer uncorrelated and thus, the estimators of its parameters may present bias and high variance. For this reason, the GAM-VAR-PCA model is not a true alternative.*

2.5. Goodness-of-fit

The comparison of the proposed procedures is performed by means of some goodness-of-fit statistics, such as the mean square error (MSE), Akaike information criterion (AIC) and Bayesian information criterion (BIC). The estimated mean square error is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

where $\hat{Y}_i$ is the predicted value of Y_i, the number of hospital treatments. The Akaike information criterion (AIC), see Akaike (1973), and the Bayesian information criterion (BIC), see Schwarz (1978), which are widely applied for model selection, are defined as:

$$\text{AIC} = -2\widehat{\ell} + 2k \quad \text{and} \quad \text{BIC} = -2\widehat{\ell} + k\ln(n),$$

where $\widehat{\ell}$ is the maximized value of the log-likelihood function defined by (1), (7) and (14) for the GAM, GAM-PCA and GAM-PCA-VAR models, respectively, k is the number of free parameters to be estimated and n is the sample size. Note that $k = 1 + r(q+2) - r(r+1)/2$ for GAM-PCA and $k = 1 + r(q+2) + q^2 - r(r+1)/2$ for GAM-PCA-VAR since the degree of freedom in $q \times r$ orthogonal real matrices is $rq - r(r+1)/2$. In this study, the log-likelihood ℓ is evaluated at the parameter values resulted from the proposed two- and three-stage estimation methods for GAM-PCA and GAM-PCA-VAR, respectively.

3. Results

3.1. Simulation study

In order to evaluate the effect in the parameter estimation, and hence in the RR estimates, of a GAM model in the presence of temporal correlation in both, the dependent Y_t and independent $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{qt})^\top$ vector, a simple simulation study was conducted. The data were generated under three scenarios: independent data (S1); the dependent variable is a time series and the covariates are independent random vectors in time (S2); and both the dependent and independent variables are time series (S3). For the 3 scenarios, the data were generated from a conditional Poisson model, $Y_t|\boldsymbol{X}_t \sim \text{Poi}(\mu_t)$.

Initially, only one covariate X_1 was considered. In this case, for (S1), the predictor is given by $\log(\mu_t) = \beta_0 + \beta_1 X_{1t}$ where $X_{1t} \sim N(0,1)$ for all t, which means that neither $\{Y_t\}$ nor $\{X_{1t}\}$ are time series. Under Scenarios 2 and 3, the predictor is given by $\log(\mu_t) = \beta_0 + \beta_1 X_{1t} + \epsilon_t$, where $\{\epsilon_t\} \sim AR(1)$ with autoregressive coefficient $\varphi = 0.1, 0.5$ and 0.9. The difference between Scenarios 2 and 3 is that, for the first, $X_{1t} \sim N(0,1)$

Table 1. Simulation results for a single covariate

Model	Parameter	Mean	Bias	MSE
S1: Independent	$\beta_0 = 1$	0.9958	-0.0042	0.0049
	$\beta_1 = 1.5$	1.5010	0.0010	0.0026
S2: φ=0.1	$\beta_0 = 1$	1.4873	0.4873	0.2921
	$\beta_1 = 1.5$	1.4457	-0.0543	0.0671
S2: φ=0.5	$\beta_0 = 1$	1.6084	0.6084	0.4782
	$\beta_1 = 1.5$	1.4091	-0.0909	0.1116
S2: φ=0.9	$\beta_0 = 1$	2.7779	1.7779	4.7168
	$\beta_1 = 1.5$	1.3189	-0.1811	0.2544
S3: φ=0.1	$\beta_0 = 1$	1.4732	0.4732	0.3673
	$\beta_1 = 1.5$	1.3903	-0.1097	0.1180
S3: φ=0.5	$\beta_0 = 1$	1.6512	0.6512	0.5727
	$\beta_1 = 1.5$	1.3790	-0.1210	0.1528
S3: φ=0.9	$\beta_0 = 1$	2.8475	1.8475	5.0797
	$\beta_1 = 1.5$	1.2518	-0.2482	0.2918

for all t and, for the later, $\{X_{1t}\} \sim AR(1)$ with $\phi = 0.5$. Thus, Scenario 2 represents the case where $\{Y_t\}$ is a time series, but $\{X_{1t}\}$ is not and Scenario 3 represents the case where both $\{Y_t\}$ and $\{X_{1t}\}$ are time series. For these three scenarios, $\beta_0 = 1$, $\beta_1 = 1.5$, the sample size $n = 100$ and the number of Monte Carlo simulations was equal to 1000. The empirical values of mean, bias and mean square error (MSE) are displayed in Table 1. All results were obtained by using R-code.

In the case of independent data (S1), the estimate of β_1 is very close to the true value, as expected. However, the picture changes dramatically especially in Scenario 3. It can be seen that the estimate of β_1 is heavily affected by the autocorrelation structure present in the data, by presenting a negative bias which increases in absolute value as φ increases positively. Hence, the estimated MSE also increases substantially with φ. In particular, for the last scenario when both $\{Y_t\}$ and $\{X_{1t}\}$ are time series, it can be seen that the fitted GAM model tends to severely underestimate β_1. As the RR is a function of β_1, its bias also introduces bias in the RR estimates in the sense that it tends to decrease when the autocorrelation structure increases. Hence, the correlation structure present in the data may attenuate the true RR estimate, which can lead to a false positive conclusion (this empirical evidence was also discussed in Dionisio *et al.* (2016) in a different simulation scenario). Thus, if a GAM model is fitted to time series variables, without mitigating the temporal correlation structure of the covariates as, for example, by removing this from the data, the RR estimate may not correspond to the true relation between the variables.

Next, we evaluate the effect in the parameter estimation of a GAM model when there are two covariates, $\boldsymbol{X}_t = (X_{1t}, X_{2t})^\top$. The setup is the same one as described previously for scenarios S1, S2 and S3, with two covariates instead of a single one. Thus, under S1 the predictor is given by $\log(\mu_t) = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t}$, where $X_{1t}, X_{2t} \sim N(0, 1)$ are independent for all t. Under S2 and S3, the predictor is given by $\log(\mu_t) = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t$, where $\{\epsilon_t\} \sim AR(1)$, with $\varphi = 0.5$. Now the difference between S2 and S3 is that, for the first one, $X_{1t}, X_{2t} \sim N(0, 1)$, mutually independent, and, for the later, $(X_{1t}, X_{2t})^\top$ forms a VAR(1) process with autoregressive coefficient matrix Φ of

Table 2. Simulation results for two covariates, X_1 and X_2

Model	Parameter	Mean	Bias	MSE
S1: Independent	$\beta_0 = 1$	0.9964	-0.0036	0.0048
	$\beta_1 = 1.5$	1.5015	0.0015	0.0026
	$\beta_2 = 0.5$	0.4999	-0.0001	0.0020
S2	$\beta_0 = 1$	1.5955	0.5955	0.5180
	$\beta_1 = 1.5$	1.4799	-0.0201	0.0701
	$\beta_2 = 0.5$	0.4719	-0.0281	0.0621
S3: $\phi_{11} = 0.7$, $\phi_{12} = 0$	$\beta_0 = 1$	1.6254	0.6254	0.7941
$\phi_{21} = 0$, $\phi_{22} = 0.5$	$\beta_1 = 1.5$	1.3708	-1.1292	0.1208
	$\beta_2 = 0.5$	0.4596	-0.0404	0.0701
S3: $\phi_{11} = 0.7$, $\phi_{12} = 0.4$	$\beta_0 = 1$	1.6654	0.6654	1.3042
$\phi_{21} = 0$, $\phi_{22} = 0.5$	$\beta_1 = 1.5$	1.3559	-0.1441	0.1299
	$\beta_2 = 0.5$	0.4487	-0.0513	0.0933

dimension 2×2. The results are displayed in Table 2.

From Table 2, similar conclusions are drawn as in the case of a single covariate (Table 1), that is, the coefficients of X_1 and X_2 are always underestimated when the process is generated by time series, either in the response or in the covariate vector. Nevertheless, the bias in the estimates is much larger in a more complex model structure compared to the case of a single covariate.

The next empirical study has the aim to illustrate, with a simple simulated model, the time-correlation effect in the PCA as discussed in Section 2.2, more specifically, the result of (4). For this purpose one sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{500}\}$ was generated from the process $\{\boldsymbol{X}_t\}$ in (11) that follows a two-dimensional VAR(1) model with $\phi_{11} = \phi_{22} = 0.5$, $\phi_{12} = 0.1$ and $\phi_{21} = 0.8$ and Gaussian white noise vector with

$$\Sigma_\epsilon = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}.$$

The estimated PCs, that is, $\hat{Z}_{1t}, \hat{Z}_{2t}$, $t = 1, ..., 500$, were computed from the 2×2 sample covariance matrix of $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{500}\}$. The sample correlation and cross-correlation functions of the PCs are displayed in Figure 1, in which $\hat{Z}_{1t}$ and $\hat{Z}_{2t}$ correspond to PC1 and PC2, respectively. As can be seen, the plots clearly indicate that the correlation structure of the models is transferred to the principal components as shown in equation (4). Based on the above empirical evidences, as well as on the discussion of the previous sections, it is clear that the temporal correlation can not be neglected when using PCA in regression models with covariates being time series data, otherwise the conclusions can be totally erroneous and lead to severe consequences. Therefore, the use of the proposed methodology discussed in Subsection 2.4 can be an alternative approach to mitigate this problem. These issues are also discussed in the next section, but with a real data set.

3.2. Data Analysis

In this study, the number of hospital admissions for respiratory diseases (RD) was obtained from the main childrens emergency department in the Vitória Metropolitan Area (called Hospital Infantil Nossa Senhora da Gloria). Respiratory diseases are classified

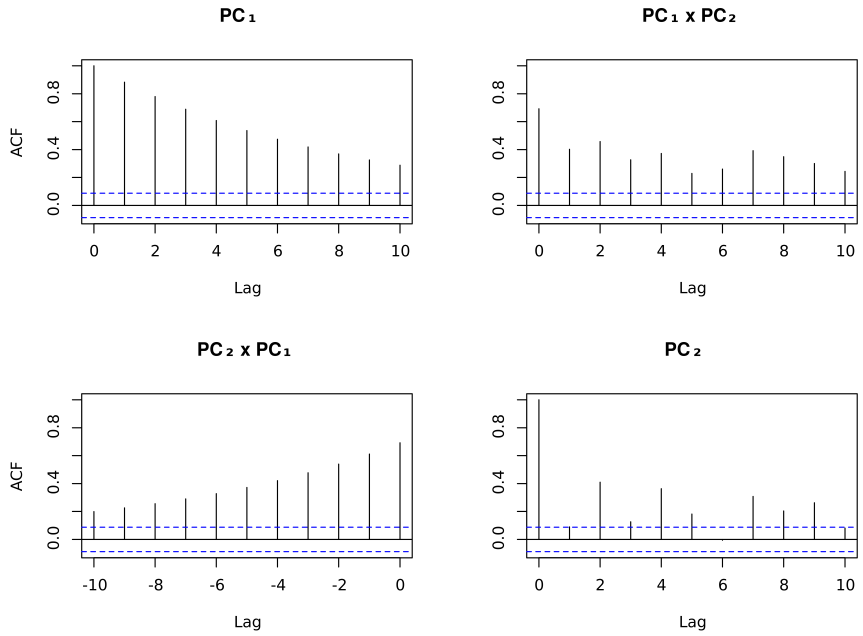

Fig. 1. Sample autocorrelation function (ACF) and cross-correlation function (CCF) of the PCs.

according to the International Classification of Diseases (ICD-10), and the investigated group consisted of children under 6 years old. The study was performed between January 1, 2005 and December 31, 2010 (n=2191).

The following atmospheric pollutants were studied: particulate material (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). Information on the daily levels of the aforementioned pollutants and data for meteorological variables were obtained from the State Environment and Water Resources Institute (IEMA), where the data were collected at 8 monitoring stations (RAMQAr).

The data collection for all the pollutants occurred over a 24-hour period that began in the first half-hour of the day. The following data were obtained at each station: the 24-hour average concentration for PM_{10} and SO_2, 8-hour moving average concentrations for CO and O_3, and the 24-hour maximum concentration for NO_2. The daily averages among the stations at which these variables were recorded were used as the covariates in the regression approaches suggested here.

Table 3 shows the descriptive statistics (i.e., the averages, standard deviations, and quantile values, among others) of the variables considered. The average number of daily treatments was 27.1 with a standard deviation of 6.15. The concentrations of the pollutants considered exceeded neither the primary air quality standard recommended by the Brazilian National Council for the Environment (CONAMA), nor the guidelines suggested by the World Health Organization (WHO). However, other studies have shown that human exposure to air pollutants levels below the acceptable standards can also cause deleterious human health effects, see Bakonyi *et al.* (2004).

Table 3. Descriptive statistics for the variables under study (Vitória Metropolitan Area, Jan 2005 to Dec 2010) †

				Percentile			
	Mean	Standard deviation	Minimum	25	50	75	Maximum
PM_{10}	33.45	8.83	8.98	27.90	32.75	38.39	86.74
SO_2	12.44	3.11	4.89	10.06	12.16	14.57	26.48
O_3	31.86	8.36	12.10	25.97	30.73	36.58	72.34
NO_2	24.82	6.93	9.03	19.59	24.13	29.37	62.59
CO	885.79	231.28	295	724.82	866.60	1031.09	2141.50
T_{min}	20.86	2.47	13.10	19.08	21.15	22.80	25.98
T_{ave}	24.43	2.45	17.00	22.62	24.40	26.35	30.80
T_{max}	29.35	3.28	19.40	27.20	29.41	31.60	39.70
RH	77.43	6.03	61.60	73.24	77.19	81.14	97.28
NT	27.09	6.15	1.00	13.00	24.00	37.00	121.00

† T= Temperature ($°C$); RH= Air relative humidity (%); NT= Number of treatments for respiratory diseases. The measure of concentration of pollutants is $\mu g/m^3$.

Table 4. Correlation among pollutants, meteorological variables and number of treatments †

	PM_{10}	SO_2	NO_2	CO	O_3	T_{max}	T_{min}	RH	NT
PM_{10}	1.00								
SO_2	0.31	1.00							
NO_2	0.34	0.04	1.00						
CO	0.35	0.22	0.61	1.00					
O_3	-0.04	-0.08	0.04	-0.40	1.00				
T_{max}	0.20	0.44	-0.43	-0.06	-0.23	1.00			
T_{min}	-0.10	0.16	-0.48	-0.10	-0.16	0.62	1.00		
RH	-0.28	-0.29	0.23	0.26	-0.22	-0.44	-0.03	1.00	
NT	0.05	-0.33	0.09	0.09	-0.08	-0.15	-0.19	0.14	1.00

† T= Temperature ($°C$); RH= Air relative humidity (%); NT= Number of treatments for respiratory diseases.
All correlations were significant at a 5% level.

The average maximum temperature used in the model was $29.35°C$ with a standard deviation of $3.28°C$, and the average relative humidity of the air was 77.43% with a standard deviation of 6.03%.

The graphs in Figure 2 show that the series of air pollutants concentration and the number of hospital admissions for RD possess seasonal behaviour, which was to be expected for these phenomena. Another characteristic observed in the series was an apparently weak stationarity. This result is confirmed in the graphs of the sampling functions of the autocorrelations shown in Figure 3.

Table 4 shows the correlations among the atmospheric pollutants, the meteorological variables and the treatments. Although some sample correlations appear not to be numerically significant, the non-parametric Pearson correlation test indicated that correlation among the atmospheric pollutants is significant for all pairs of variables at level 5% and for most pairs at level 0.1%. For example, the test displayed 0.0476 as the maximum empirical level, which was found for the correlation between PM_{10} and

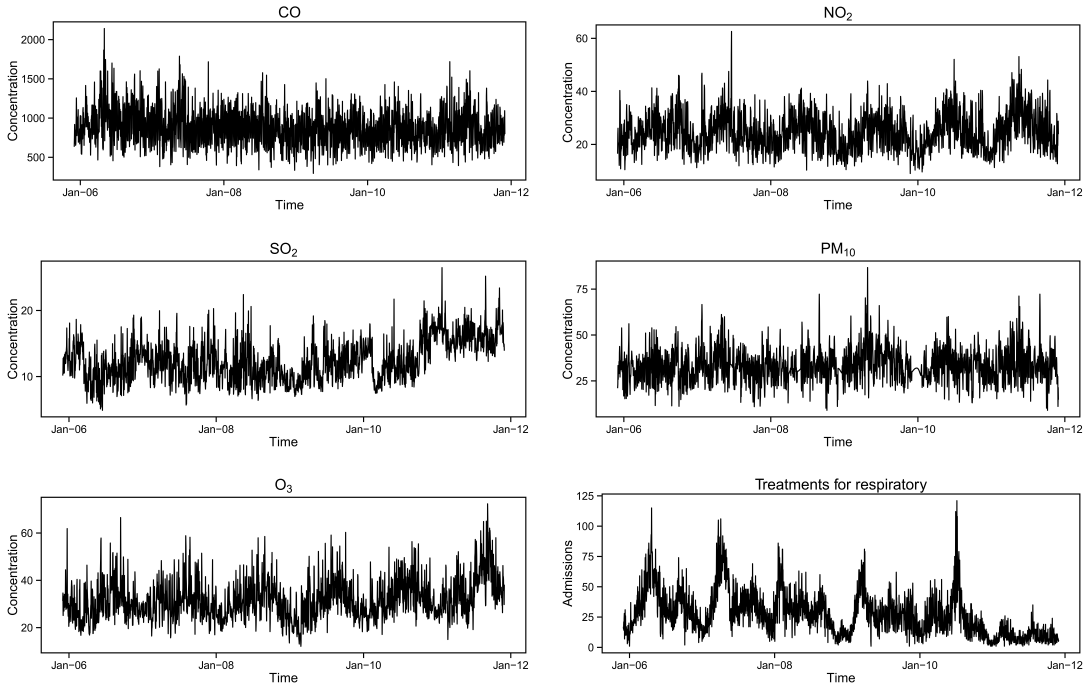

Fig. 2. Concentration of CO; NO_2; SO_2; PM_{10}; O_3 and the Number of treatments for RD.

O_3. The minimum and maximum temperatures were negatively correlated with the pollutants O_3 and NO_2 and positively correlated with the pollutant PM_{10}. The positive correlation between the maximum and minimum temperatures and the pollutant PM_{10} could be explained by the acceleration of the pollutant dispersion during the hotter periods and the accumulation of pollutants in the air at low temperatures, which impeded the dispersion of the particles and kept them at the atmospheric level.

The aforementioned descriptive and graphical analysis motivated the use of the PCA technique in the GAM for the atmospheric pollutant data, even though the pollutants had an apparently weak correlation and self-correlation structure.

Table 5 shows the results of applying the PCA technique to the correlation matrix of the PM_{10}, SO_2, NO_2, O_3 and CO data. Here, in order to keep the notation consistent with the previous sections, PC1,..., PC5 correspond to $\hat{Z}_{1t}, \ldots, \hat{Z}_{5t}$, respectively. The first three components correspond to 83.2% of the total variability. The highest coefficients (in eigenvectors) of principal components 1, 2 and 3 are those of the pollutants CO, O_3 and SO_2, respectively. As a complement to the analysis in Table 5, a cluster division was performed for each component to group, for example, the pollutants with factor loadings higher than 0.45. In Table 5, the (*) symbol indicates the possible clusters for each principal component.

Figure 4 shows the time behaviour of some principal components obtained from the pollutant concentration series, i.e., the original data. The figure shows that PC1 is autocorrelated and that the cross-correlations are non-null, corroborating the results

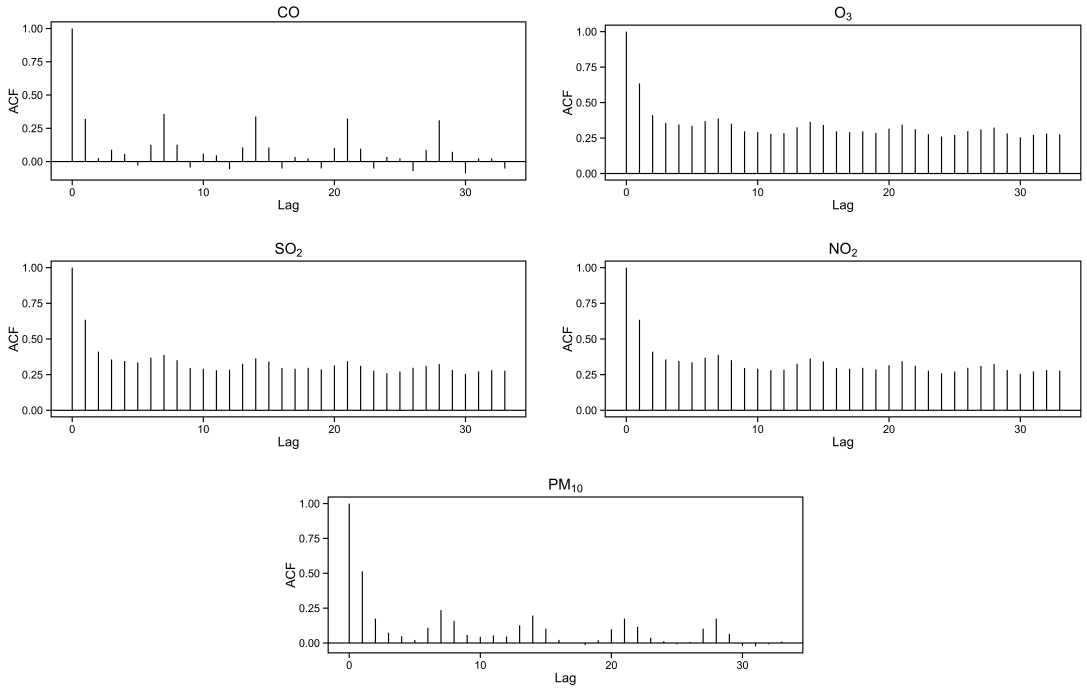

Fig. 3. Sample autocorrelation function (ACF) of the pollutants.

discussed in Subsections 2.2 and 3.1. The components also clearly exhibited the seasonal behaviour of the pollution variables, as expected. That is, the graphs show that the autocorrelation structure of the pollutants persists in the components. Therefore, the PCA technique should be applied carefully even for processes with an apparently weak autocorrelation structure. This is an argument contrary to page 299 in Jolliffe (2002), in which the author argues that *"when the main objective of PCA is only descriptive, complications such as non-independence (temporal) does not seriously affect this objective"* (see, also, Zamprogno (2013) and Vanhatalo and Kulachi (2016)).

The cumulative proportion of the variance was the choice criterium for the components to be included in the GAM. Thus, following the parsimony criterium, the first three components were chosen as covariates (highlighted in bold in Table 5), which corresponds to 83.2% of the total variability and the simplest model as possible to handle the complex correlation structure of the data. The number of daily treatments for respiratory diseases was considered to be the dependent variable, and each outcome was modelled based on the assumption that the count of respiratory disease events (i.e., hospital admissions) followed a Poisson distribution.

The analysis involved several procedures implemented in stages. Initially, the short-term seasonality was treated using indicator variables for weekdays and holidays. A loess smoothing function, see Friedman (1991), was used to model the long-term seasonality to control for the non-linear dependence. The confounding covariates (i.e., the temperature and the relative humidity) were modelled using splines smoothing curves (see Friedman

Table 5. Results of factor loadings and statistics applying PCA for the pollutants

	PC1	PC2	PC3	PC4	PC5
Standard deviation †	1.4315	1.0431	1.0115	0.7741	0.4904
Proportion of variance	0.4098	0.2176	0.2046	0.1198	0.0481
Cumulative proportion of variance	**0.4098**	**0.6274**	**0.8320**	0.9519	1.0000
CO	-0.6074*	-0.1999	-0.2311	-0.2146	-0.7012
NO_2	-0.5058*	0.3316	-0.4786	-0.2599	0.5810
O_3	0.2523	0.8615*	-0.0363	-0.1995	-0.3911
PM_{10}	-0.4680*	0.3213	0.2784	0.7746	-0.0151
SO_2	-0.3041	0.0680	0.7992*	-0.4966	0.1327

† Standard deviation is the square root of the eigenvalue.

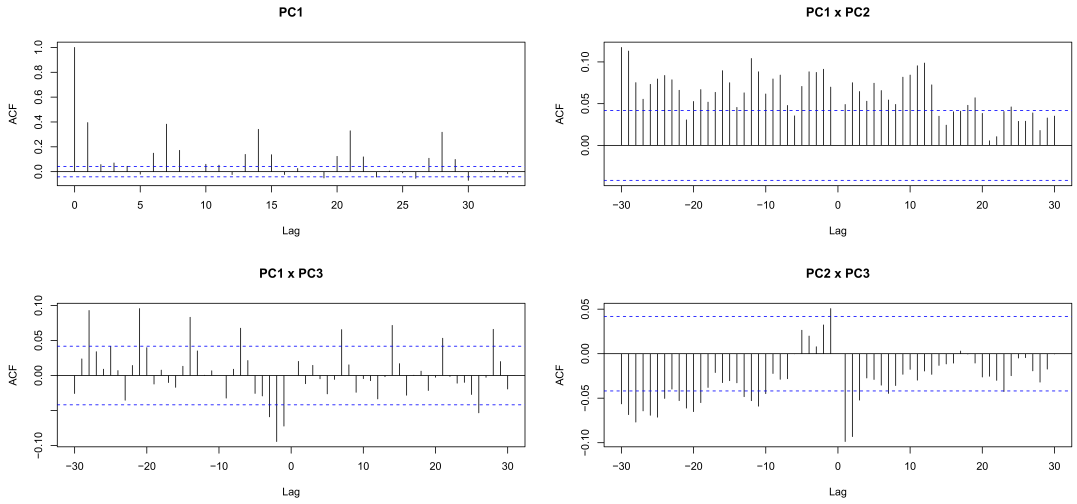

Fig. 4. Cross-correlation function (CCF) of the main components of the pollutants studied.

(1991) and Wahba (2001)). The best GAM-PCA fit was obtained based on a residual analysis and the Akaike information criterion (AIC), Akaike (1973).

As previously mentioned, the unfiltered principal components are autocorrelated. Consequently, this property was transferred to the residuals of the GAM-PCA model (see Figure 5). Therefore, as a post-processing step, a $SAR(1)(1)_7$ model was fitted to the residuals of the GAM-PCA, resulting in the final GAM-PCA model with SAR residuals, shortly the final GAM-PCA model, which has eliminated the autocorrelation in the data. The parameter estimates for this model are given in Table 6. It should be noted that the temperature was no longer significant and thus was dropped from the final model. Figure 6 shows that there is no autocorrelation structure in the residuals of the final GAM-PCA model after SAR filtering the residuals of the GAM-PCA model.

For the GAM-PCA-VAR model proposed in this paper, a seasonal vector autoregressive model with a 7-day period, $SVAR_7(1)$, was used to adjust the pollutant vector. Although the model discussed in the previous section is related to VAR(1), its extension

Table 6. Results of the final GAM-PCA model to estimate the effects of pollutants concentrations on the hospital admission in the Vitória Metropolitan Area †

Variables	Estimates	Standard error	Z value	p-value
(Intercept)	4.4871	0.0901	49.82	0.0000***
Tuesday	-0.1596	0.0152	-10.50	0.0000***
Wednesday	-0.2176	0.0154	-14.14	0.0000***
Thursday	-0.1321	0.0151	-8.76	0.0000***
Friday	-0.1571	0.0154	-10.22	0.0000***
Saturday	-0.1204	0.0150	-8.04	0.0000***
Sunday	-0.0860	0.0154	-5.59	0.0000***
Holiday2‡	0.1886	0.0440	4.29	0.0000***
Holiday3§	0.3189	0.0384	8.30	0.0000***
Air relative humidity	-0.0061	0.0009	-6.83	0.0000***
PC1	-0.0244	0.0040	-6.16	0.0000***
PC2	0.0163	0.0055	2.99	0.0028 **
PC3	-0.0157	0.0056	-2.79	0.0052***

† Significant: '***' 0.001 '**' 0.01

‡ Holiday2 = Corpus Christ + Our Lady of Penha

§ Holiday3 = Carnival + holiday (Tiradentes day) + Brazil's Independence day

using $SVAR_7(1)$ model instead is straightforward obtained. The seasonal VAR(1) model was selected based on the standard fitting tools of multivariate time series model, e.g., the VAR package of R. Table 7 displays the results of applying the PCA technique to the residual matrix of the seasonal VAR(1) model. It shows that the time structure of the pollutants did not alter the cumulative proportion of the variance, that is, the variability in the first three components explain 83% of the variability in the filtered data, which is equivalent to the results in Table 5. This may be explained by the fact that the serial dependence of the pollutants was not strong enough to produce an impact on the PCA, Zamprogno (2013), or because of the effect of the high levels of the pollutant on the estimation of the covariance matrix (see, for example, Reisen *et al.* (2017), Cotta *et al.* (2017) and Zamprogno (2013)).

However, the clustering of the pollutants by factor loadings resulted in a different interpretation, which is more coherent with the behaviour of the variables considered. The clusters are indicated with (**) in the analysis. The results showed a correlation between the NO_2 and O_3 pollutants that was not observed in the previous case. These two pollutants are physically associated with each other because the formation of O_3 depends on the release of the NO_2 particle.

Figure 7 shows that the fitting of the seasonal VAR(1) model practically eliminated the autocorrelation of PC1 and the cross-correlation, as expected from the aforementioned discussion. The residual plots (ACF and PACF) of GAM-PCA-VAR displayed similar behaviour as the GAM-PCA with SAR residuals (the final GAM-PCA) shown in Figure 6. These plots are available upon request. Additionally, Figure 8 shows the fit (predicted values) obtained using the GAM-PCA-VAR model. This graph shows that the model provided a good fit to the data for the variable of interest, i.e., the number

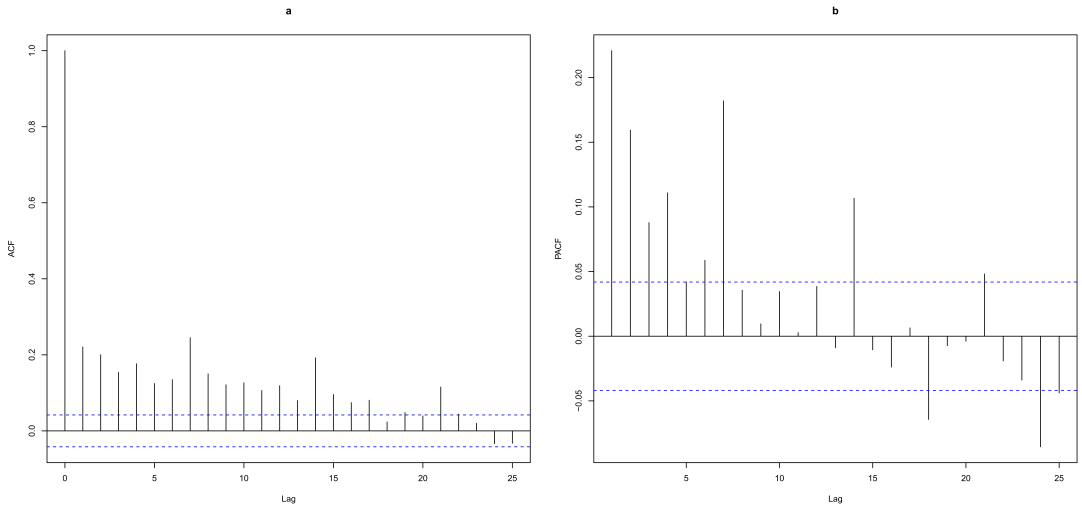

Fig. 5. (a) Sample autocorrelation function and (b) sample partial autocorrelation function of the residuals of the GAM-PCA model.

Table 7. Results of factor loadings and statistics applying PCA for the filtered pollutants

	PC1	PC2	PC3	PC4	PC5
Standard deviation †	1.4774	1.0223	0.9628	0.7228	0.5680
Proportion of variance	0.4366	0.2090	0.1854	0.1045	0.0645
Cumulative proportion of variance	**0.4366**	**0.6456**	**0.8310**	0.9355	1.0000
CO	0.5711**	-0.1431	0.2918	-0.1469	0.7393
NO_2	0.4205	-0.6527**	0.2543	-0.0905	-0.5695
O_3	-0.3693	-0.5801**	-0.4685	-0.4896	0.2606
PM_{10}	0.4012	-0.1409	-0.7040**	0.5663	0.0532
SO_2	0.4468	0.4441	-0.3675	-0.6402	-0.2414

† Standard deviation is the square root of the eigenvalue.

of daily treatments for children under 6 years old in the metropolitan area.

The goodness-of-fit results for the three models (GAM, GAM-PCA and GAM-PCA-VAR) using the MSE, AIC and BIC statistics, are given in Table 8. MSE of the GAM model is approximately 35% higher than the MSE of the other two models, which was an expected results since a more complex model may yield a better residual fit. The AIC and BIC information criteria indicate that the GAM-PCA-VAR model is the best to fit the data. All these empirical analyses, that is, the plots of the ACF and PACF of the residuals which were shown to be uncorrelated, the behaviour of the estimated PCs (Figure 7), the final fit (Figure 8) and the results in Table 8 support the fact that the proposed model GAM-PCA-VAR is suitable, for the purpose of the paper, to model this data. The final performance of this procedure to quantify the association between respiratory disease and pollution is evaluated by means of the estimated RR as follows.

The RR estimates for each pollutant and model were calculated to compare the

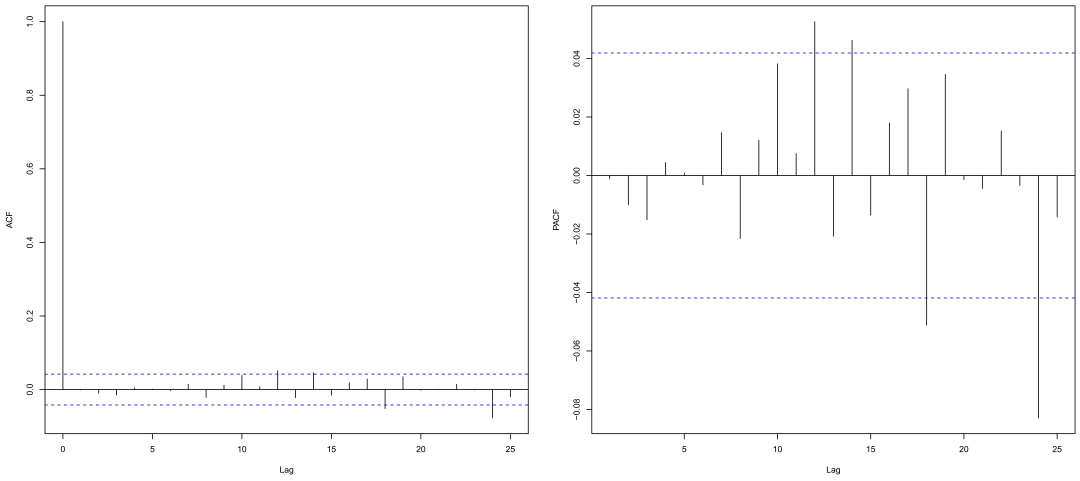

Fig. 6. (a) Sample autocorrelation function and (b) partial autocorrelation function of the residuals of the final GAM-PCA model.

Table 8. Goodness-of-fit statistics for the estimated models

Model	MSE	AIC	BIC
GAM	1.480	24610	24720
GAM-PCA	1.143	24442	24245
GAM-PCA-VAR	1.144	24166	24190

performances of the GAM ($\widehat{\mathrm{RR}}$), GAM-PCA ($\widehat{\mathrm{RR}}^{*}$) and GAM-PCA-VAR ($\widehat{\mathrm{RR}}^{**}$) models for the variables under consideration. The results are displayed in Table 9 in terms of the increase in the interquartile variation, which was based on performing the RR analysis for pollutants at different scales. Most RR estimates were significant for all of the considered models, i.e., in general, the pollutants contributed significantly to the increase in the number of treatments for respiratory diseases. In the majority, the most significant RR estimates were obtained using the developed GAM-PCA-VAR model.

As an example of a specific and comparative analysis of the RR values, the RR estimates for the pollutant PM_{10} increased from approximately 2% ($\widehat{\mathrm{RR}}$) to 3% ($\widehat{\mathrm{RR}}^{*}$) and 7% ($\widehat{\mathrm{RR}}^{**}$). Substantial increases in the RR estimates were also observed for the pollutant CO. In this case, $\widehat{\mathrm{RR}} = 1.020$, $\widehat{\mathrm{RR}}^{*} = 1.048$ and $\widehat{\mathrm{RR}}^{**} = 1.077$.

Therefore, the developed GAM-PCA and GAM-PCA-VAR models generally showed more pronounced results than the conventional GAM for the expected increase in the number of treatments for respiratory diseases, since the procedure allows a set of pollutants to be the explanatory variable.

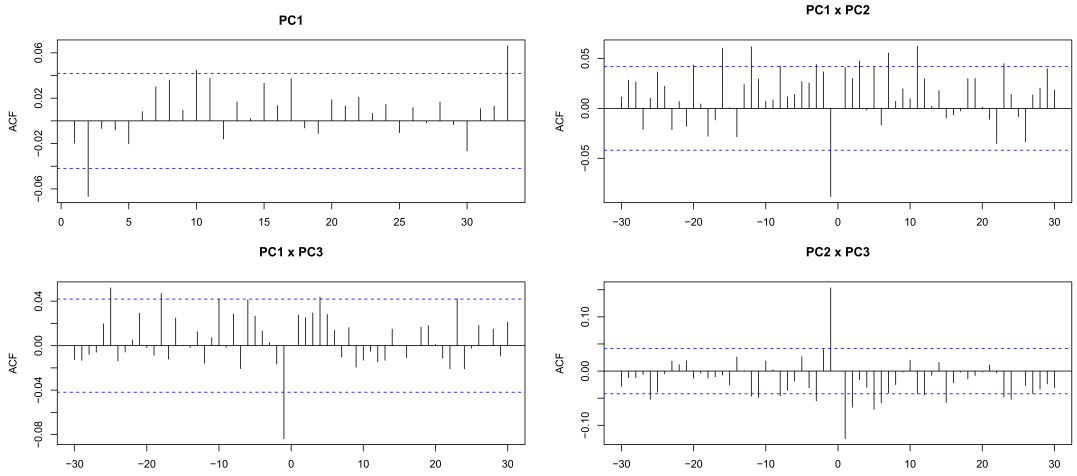

Fig. 7. Cross-correlation function (CCF) of the main components of the filtered pollutants.

Table 9. Relative risk (RR) and 95% confidence intervals for treatments for respiratory diseases in children under 6 years old for an interquartile variation in the pollutants PM_{10}, SO_2, NO_2, O_3 and CO in the Vitória Metropolitan Area from Jan 2005 to Dec 2010 †

	$\widehat{\text{RR}}$	$\widehat{\text{RR}}^*$	$\widehat{\text{RR}}^{**}$
PM_{10}	1.020(1.010,1.039)	**1.029(1.001,1.090)**	**1.075(1.001,1.092)**
SO_2	1.040(1.010,1.080)	0.982(0.972,1.001)	1.027(1.010,1.040)
CO	1.020(1.010,1.030)	**1.048(1.002,1.071)**	**1.077(1.020,1.100)**
NO_2	1.000(0.990,1.020)	1.028(1.010,1.040)	1.012(1.010,1.030)
O_3	0.980(0.972,1.001)	1.081(1.003,1.093)	0.992(0.992,1.020)

† $\widehat{\text{RR}}$: GAM, $\widehat{\text{RR}}^*$: GAM-PCA and $\widehat{\text{RR}}^{**}$: GAM-PCA-VAR

4. Conclusion

A hybrid of three statistical tools, the vector autoregressive model (VAR), the principal component analysis (PCA), and the generalized additive model (GAM), with Poisson marginal distribution, was developed in this study to correlate the effect of atmospheric exposure to pollutants PM_{10}, SO_2, NO_2, O_3 and CO with the number of treatments for respiratory diseases in children under 6 years old in the Vitória metropolitan area, Brazil, between 2005 and 2010. It should be noticed that, due to the complexity of the real data, a marginal Poisson assumption would not be the most appropriated choice in this case since the series presented overdispersion problem, which may come from many features of the data such as changes in the mean and variance, observations with high levels (which increases substantially the variance) among others. The overdispersion is a common phenomenon in this kind of data, and the negative binomial (NB) and the generalized Poisson (GP) models are frequently used to account for this problem. For example, several statistics were proposed by Yang *et al.* (2007) in testing for a Poisson regression model against NB or GP alternatives. However, the Poisson distribution is

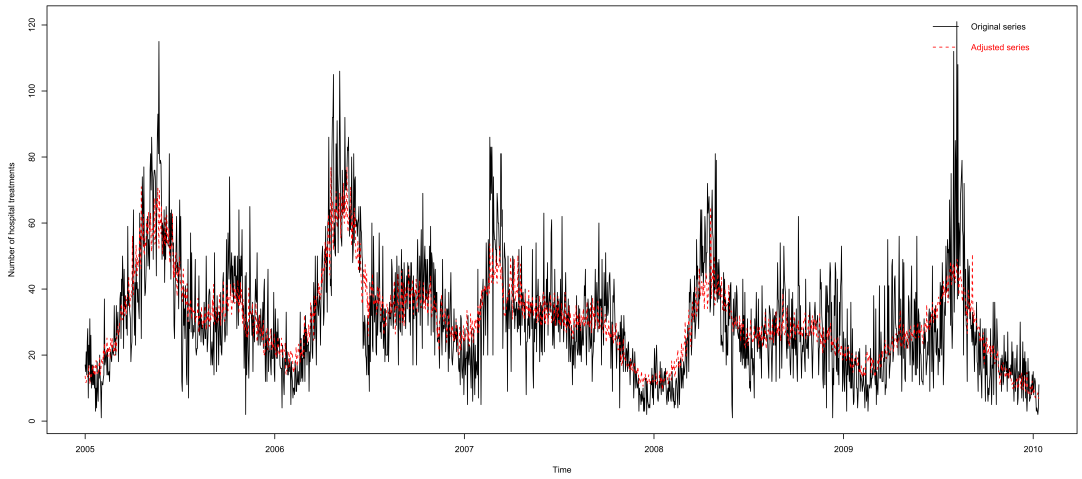

Fig. 8. Fitted GAM-PCA-VAR model to the number of treatments for RD.

the most popular distribution used in real applications when dealing with association between pollution and health adverse problems. Besides, in this work the main objective was to investigate the effect of serial and cross-correlation of the pollutants included in the fit. Therefore, the use of a model that also handles the overdispersion problem is an interesting and important issue to be considered in the context of the data here analysed. Hence this point, and the effect of high concentration levels of the pollutants in the estimate of RR and bootstrap intervals for this quantity will be part of future work of the authors of the paper.

The developed models were denoted here by GAM-PCA and GAM-PCA-VAR. The first model used the principal components (PCs) of the original pollutants as covariates in the GAM model. The residuals of this model were fitted using the SAR(1)$(1)_7$ model, resulting in the final GAM-PCA model. In the second approach, a seasonal VAR(1) model was used to filter the original pollutants, before building the PCs. These modified PCs were then used as covariates in the GAM model, resulting in the hybrid model defined as GAM-PCA-VAR. In this later model, the autocorrelation and cross-correlations of the PCs were removed by the VAR model.

A simulation study was conducted to evaluate the effect in the parameter estimation of GAM models when the explanatory variables possess serial correlation. The results showed that, if the autocorrelation present in the independent variables is not taken into account, the GAM fit tends to underestimate the true value of the coefficients, and consequently, it leads to biased RR estimates. This means that a true effect of a pollutant in the population health can be underestimated if the model is not correctly adjusted. This issue was also recently explored in different scenario by Dionisio *et al.* (2016).

The fitting adequacy of the aforementioned models was compared by means of goodness-of-fit statistics, such as MSE, AIC and BIC. Based on these quantities, in general, the three methods displayed close results, where the standard GAM presented the worst

performance.

The deleterious health effects of the exposure to pollutants for the population of children in the Vitória metropolitan area were obtained by estimating the RR of the GAM, GAM-PCA and GAM-PCA-VAR regression models. In general, the RR estimates were significant for all models considered in the study. It should be stressed here that, in most cases, the estimated RR is larger for GAM-PCA-VAR when compared to the GAM model. This can be explained by the results obtained in the simulation study. Thus, the real effect of these pollutants in the number of respiratory diseases can be underestimated if we use the standard GAM model under an inappropriate scenario as it was the case of the data used here. For example, for the pollutant PM_{10}, the estimated relative risk increased from approximately 2% ($\widehat{\text{RR}}$) to 3% ($\widehat{\text{RR}}^{*}$) and 7% ($\widehat{\text{RR}}^{**}$). For the GAM-PCA model, an increase of 10.49 $\mu g/m^3$ (interquartile range) of the particulate material (PM_{10}) resulted in a $\widehat{\text{RR}}^{*}$ value of 1.029 with 95% CI (1.001,1.09), while for the GAM-PCA-VAR model a higher $\widehat{\text{RR}}^{**}$ value of 1.075 with 95% CI (1.001,1.092). Similar interpretations could be made for the other pollutants and developed models.

In this study, the results obtained using the GAM and GAM-PCA models were coherent with those reported in Wang and Pham (2011), in which the morbidity was correlated with the atmospheric pollutant concentrations using data registered in Korea. Although the serial correlation of the data was ignored by the authors when using PCA, the study also shows that the PCA technique improved the final relative risk estimates.

Acknowledgments

The authors thank the following agencies for their support: the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq), the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES), Espírito Santo State Research Foundation (Fundação de Amparo à Pesquisa do Espírito Santo - FAPES) and Minas Gerais State Research Foundation (Fundação de Amparo à Pesquisa do estado de Minas Gerais - FAPEMIG). Márton Ispány was supported by the EFOP-3.6.1-16-2016-00022 project. Pascal Bondon thanks to the Institute for Control and Decision of the Université Paris-Saclay. Part of this paper was written when Valderio A. Reisen was a visiting professor at CentraleSupélec. Valderio A. Reisen is indebted to CentraleSupélec for the financial support. Some application results presented in this paper were part of the Master Thesis of Juliana B. Souza at PPGEA-UFES under supervision of Valderio A. Reisen and Jane M. Santos.

The authors are very grateful to the two anonymous referees and the editor for their suggestion that led to a markedly improved paper.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.

Bakonyi, S. M. C., Danni-Oliveira, I. M., Martins, L. C. and Braga, A. L. F. (2004) Air pollution and respiratory diseases among children in the city of Curitiba, Brazil. *Rev. Saude Publ.*, **38(5)**, 675–700.

Baxter, L. A., Finch, S. J., Lipfert, F. W. and Yu, Q. (1997) Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. *Risk Anal.*, **17(3)**, 273–278.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*. Springer Series in Statistics. New York: Springer-Verlag.

Campbell, M. J. (1994) Time series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature. *J. R. Statist. Soc. A*, **157(2)**, 191–208.

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.

Cotta, H. H. A., Reisen, V. A., Bondon, P. and Stummer, W. M. (2017) Robust estimation of covariance and correlation functions of a stationary multivariate process. to apper in *25rd European Signal Processing Conference (EUSIPCO)*.

Dalgaard, P. (2008) *Introductory Statistics with R*, 2nd edn. New York: Springer.

Dionisio, K. L., Chang, H. H. and Baxter, L. K. (2016) A simulation study to quantify the impacts of exposure measurement error on air pollution healh risk estimates in copollutant time-series models. *Environ. Health*, 15:114.

Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002) On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidemiol.*, **156(3)**, 193–203.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L. and Samet, J. M. (2006) Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *J. Am. Med. Assoc.*, **295(10)**, 1127–1134.

Figueiras, A., Roca-Pardiñas, J. and Cadarso-Suárez, C. (2005) A bootstrap method to avoid the effect of concurvity in generalized additive models in time series of air pollution. *J. Epidemiol. Commun. H.*, **59**, 881–884.

Friedman, J. (1991) Multivariate adaptive regression splines. *Ann. Stat.*, **19(1)**, 1–67.

Greenaway-McGrevy, R., Han, Ch. and Sul, D. (2012) Estimating the number of common factors in serially dependent approximate factor models. *Econ. Lett.*, **116(3)**, 531–534.

Hamilton, J. D. (1994) *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman & Hall.

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

Johnson, R. A. and Wichern, D. W. (2007) *Applied Multivariate Statistical Analysis*, 6th edn. New Jersey: Prentice Hall.

Jolliffe, I. T. (2002) *Principal Component Analysis*, 2nd edn. New York: Springer.

Kedem, B. and Fokianos, K. (2002) *Regression Models for Time Series Analysis*, 2nd edn. USA: Wiley.

Lall, R., Ito, K. and Thurston, G. D. (2011) Distributed lag analysis of daily hospital admissions and source-apportioned fine particle air pollution. *Environ. Health Persp.*, **119(4)**, 455–460.

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

Michelozzi, P., Kirchmayer, U., Katsouyanni, K., Biggery, A., McGregor, G., Menne, B., Kassomenos, P., Anderson, H. R., Baccini, M., Accetta, G., Analytis, A. and Kosatsky, T. (2007) Assessment and prevention of acute health effects of weather conditions in Europe, the PHEWE project: background, objectives, design. *Environ. Health*, **6(12)**, 1–10.

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

Ramsey, T. O., Burnett, R. T. and Krewski, D. (2003) The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**, 18–23.

Reisen, V. A., Lévy-Leduc, C., Cotta, H. H. A., Alburquerque, T. and Stummer, W. (2017) Long-memory model under outliers: An application to air pollution levels. In *Environmental Science and Engineering: Air and Noise Pollution*, pp. 211–243. USA: Studium Press LLC.

Roberts, S. and Martin, M. (2006) Using supervised principal components analysis to assess multiple pollutant effects. *Environ. Health Persp.*, **114(12)**, 1877–1882.

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

Schwarz, G. E. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6(2)**, 461–464.

Vanhatalo, E. and Kulachi, M. (2016) Impact of autocorrelation on principal components and their use in statistical process control. *Qual. Reliab. Eng. Int.*, **32**, 1483–1500.

Wahba, G. (2001) Splines in nonparametric regression. In *Encyclopedia of Environmetrics* (eds A. H. El-Shaarawi and W. W. Piegorsch), vol. 4, 2nd edn, pp. 2099-2112. Wiley.

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.

WHO (2006) *WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulphur Dioxide. Global update 2005. Summary of Risk Assessment.* Geneva: WHO Press, World Health Organization.

Yang, Z., Hardin, J. W., Addy, C. L. and Vuong, Q. H. (2007) Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. *Biometrical J.*, **49(4)**, 565–584.

Zamprogno, B. (2013) *PCA in time series with short and long-memory time series.* PhD Thesis at the Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória, Brazil.

Zhao, J., Cao, J., Tian, S., Chen, Y., Zhang, Sh., Wang, Zh. and Zhou, X. (2014) A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices. *Aquat. Ecol.*, **48**, 297–312.

Zou, G. (2004) A modified Poisson regression approach to prospective studies with binary data. *Am. J. Epidemiol.*, **159(7)**, 702–706.