ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)

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

Juliana B. de Souza,

Federal University of Espírito Santo, Vitória, Brazil

Valdério A. Reisen,

Federal University of Espírito 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, Hungary

Pascal Bondon

Centre National de la Recherche Scientifique and CentraleSupélec, Gif-sur-Yvette, and University of Paris–Saclay, France

and Jane Meri Santos

Federal University of Espírito Santo, Vitória, Brazil

[Received March 2016. Revised July 2017]

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 is multi-dimensional, the majority of these studies consider only a single pollutant as a covariate in the GAM model. This model restriction may be because the pollutant variables do not only have serial dependence but also interdependence between themselves. In an attempt to convey a more realistic model, we propose here the hybrid generalized additive model–principal component analysis–vector auto-regressive (GAM–PCA–VAR) model, which is a combination of PCA and GAMs along with a VAR process. The PCA is used to eliminate the multicollinearity between the pollutants whereas the VAR model is used to handle the serial correlation of the data to produce white noise processes as covariates in the GAM. Some theoretical and simulation results of the methodology proposed are discussed, with special attention to the effect of time correlation of the covariates on the PCA and, consequently, on the estimates of the parameters in the GAM and on the relative risk, which is a commonly used statistical quantity to measure the effect of the covariates, especially the pollutants, on population health. As a main motivation to the methodology, a real data set is analysed with the aim of quantifying the

Address for correspondence: Márton Ispány, Faculty of Informatics, University of Debrecen, Debrecen, Kassai út 26, 4028 Debrecen, Hungary.
E-mail: ispany.marton@inf.unideb.hu

0035–9254/17/67000

association between respiratory disease and air pollution concentrations, especially particulate matter PM_{10}, sulphur dioxide, nitrogen dioxide, carbon monoxide and ozone. The empirical results show that the GAM–PCA–VAR model can remove the auto-correlations from the principal components. In addition, this method produces estimates of the relative risk, for each pollutant, which are not affected by the serial correlation in the data. This, in general, leads to more pronounced values of the estimated risk compared with the standard GAM model, indicating, for this study, an increase of almost 5.4% in the risk of PM_{10}, which is one of the most important pollutants which is usually associated with adverse effects on human health.

Keywords: Generalized additive model; Multicollinearity; Principal component analysis; Relative risk; Serial correlation; Vector auto-regressive model

1. Introduction

The effect of air pollutants on human wellbeing 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 (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) and 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 non-linear association between adverse health effects and covariates such as ambient concentrations of air pollutants and meteorological conditions, mainly because it allows for non-parametric adjustments of non-linear confounding effects of seasonality and trends.

In spite of its widespread use, many researchers 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 and by the presence of auto-correlation in the series under study, among others (see for example, Dionisio *et al.* (2016)). Some works that aim to solve these problems include Dominici *et al.* (2002), who proposed a correction to the degrees of freedom in the smooth component, Dominici *et al.* (2006), Lall *et al.* (2011) and 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) and Ramsey *et al.* (2003), who have proposed some approaches to control the problem of concurvity (the non-linear dependence that can remain among the covariates). Additionally, most of the references in the epidemiological research area related to the study of the association between pollution and adverse health effects usually consider only one pollutant whereas the population under study is exposed to a complex mixture of pollutants; a broad discussion of the effect of correlated measurement errors in time series on the relative risk estimates has recently been 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 bias in regression estimates since the presence of multicollinearity (the linear dependence between the covariates) can inflate 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 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. 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 researchers have explored this relevant research direction. For example, Roberts and Martin (2006) evaluated how the pollutants PM_{10}, ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO) affect health, where the issue of multicollinearity was handled by using PCA. Roberts and Martin (2006) 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 by using PCA and a robust method. The relative risk estimates RR 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 that are considered in this paper are supposed to be stationary (including the covariates). As the principal components (PCs) are linear combinations of the covariates, their properties are linearly transferred to the PCs. 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. Zamprogno (2013) showed that the PCs are auto-correlated if the covariates are also auto-correlated. The PCs contain the time structure of the covariates and must therefore be used judiciously in the regression analysis. To remove the temporal correlation structures of PCA, Zamprogno (2013) suggested filtering the series by using a multivariate auto-regressive moving average model in the pollution variables before performing any statistical analysis using PCA. In the same context, Matteson and Tsay (2011) and Hua and Tsay (2014) applied vector auto-regressive (VAR) models to remove the serial correlation of time series of stock returns before carrying out PCA on the residuals of the VAR model. The use of Box–Jenkins methodology to eliminate the serial correlation in the data was also considered in Campbell (1994) who discussed the relationship between sudden infant death syndrome with environmental temperature by using time regression for counts with Poisson marginal distribution.

In the current study, the multicollinearity issue is solved by using PCA on the pollutants, with the components obtained being used as covariates in the GAM. This procedure is called GAM–PCA. Additionally, the problem that is associated with the presence of auto-correlation in the PCs when applying the GAM is circumvented by using a VAR model on the time series of covariates before obtaining the PCs. 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 with the conventional GAM by means of adequate goodness-of-fit statistics and, also, in terms of the relative risk estimate RR, which is a commonly used tool to measure the effect of the covariates, especially the pollutants, on population health. Some results that are related to the methods proposed and the effect of auto-correlated covariates on the PCA are theoretically and empirically discussed. In addition, the estimate of the relative risk RR is evaluated for each model in a real data problem. The objective of estimating RR is to verify whether there is any change in 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 by using a conventional GAM. A simulation study demonstrates that the intercorrelation and auto-correlation that are found in the explanatory pollutant variables may be responsible for this divergence. This is important evidence that prevents use of the standard GAM, from the epidemiological point of view, since the time series phenomena in the explanatory pollutant variables can produce unrealistic risk impacts on the health of the population under study, i.e. this may indicate a false positive result.

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

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. Methodology: generalized additive models, principal component analysis, vector auto-regressive models and relative risk

In this section, we present the methodology that is employed to relate the covariates to the count time series under study. As there are both linear and non-linear relationships between the explanatory variables and the response, a GAM model is used. The procedures are implemented by using count data with a Poisson distribution, as this is a very useful model in practical situations.

We also present, in detail, the PCA and VAR methodologies, to explain how these procedures are linked to solve problems that can occur with the kind of data that we are working with, which means multicollinearity and serial correlation in the explanatory variables.

2.1. *Generalized additive models*

The 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 is widely used to describe non-linear correlations between the variables of interest; see, for example, Schwartz (2000), Ostro *et al.* (1999) and 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 a conditional Poisson distribution with mean μ_t, then

$$p(y_t; \mu_t|\mathcal{F}_{t-1}) = \frac{\exp(-\mu_t)\mu_t^{y_t}}{y_t!}, \qquad y_t = 0, 1, \ldots.$$

Thus, the conditional log-likelihood function of the mutually conditionally independent random variables $Y_1, \ldots, Y_n$ is given by

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

where the vector $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_q)^{\mathrm{T}}$ depends on the covariates and the parameters of the process $\{Y_t\}$. Let $\mathbf{X}_t = (X_{1t}, \ldots, X_{pt})^{\mathrm{T}}$ be the vector of covariates of dimension p at time t, where 'T' 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 what follows, $X_{1t}, \ldots, X_{qt}$ denote the pollutants, whereas $X_{(q+1)t}, \ldots, X_{pt}$ denote the confounding variables at time t ($q \leqslant p$).

The relationship between Y_t and the vector $\mathbf{X}_t$ of covariates is obtained by setting (see, for example, Kedem and Fokianos (2002))

$$\ln(\mu_t)=\sum_{j=0}^{q}\beta_j X_{jt}+\sum_{j=q+1}^{p} f_j(X_{jt}), \qquad q\leqslant p,$$

where $(\beta_0,\boldsymbol{\beta})$ with $\boldsymbol{\beta}:=(\beta_1,\ldots,\beta_q)^{\mathrm{T}}$ is the vector of the coefficients to be estimated (β_j is the coefficient of the jth covariate), and f_j is a smoothing function of an appropriate function space for the jth confounding variable (e.g. the temperature or the humidity variables). Moreover, β_0 denotes the curve intercept and is associated with $X_{0t}=1$ for all t. For simplicity it is assumed that the pollutant covariates are centred. The aforementioned model is usually referred to as a semiparametric model because it involves parametric and non-parametric functions. The parameters of the parametric functions are generally estimated by using maximum likelihood or quasi-likelihood methods, by optimizing the log-likelihood defined in expression (1), with the asymptotic properties given in Kedem and Fokianos (2002). The non-parametric functions are evaluated by using 'splines', 'LOESS' or moving average functions, among others (see Friedman (1991) and Wahba (2001)).

The relative risk RR is frequently used in epidemiological studies to measure the effect of atmospheric pollutant concentrations on the health of the exposed population. RR for a pollutant covariate X_j, $j=1,\ldots,q$, 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, see formula (8) in Baxter *et al.* (1997):

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

For Poisson regression RR does not depend on the values x_i, $i\neq j$, of the other covariates and it can be expressed as

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

RR is often called the relative rate or rate ratio; see, for example, page 265 in Dalgaard (2008). Note that for binary outcomes RR is defined as the ratio of probabilities that an event will occur following a certain exposure or 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. RR and its approximate confidence interval at an α level of significance of a covariate X_j, $j=1,\ldots,q$, in the GAM with Poisson marginal distribution are estimated as follows:

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

where ξ is the variation in the pollutant concentration (e.g. a value of $10\,\mu\mathrm{g\ m}^{-3}$ 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 level of significance, 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. RR for a unit change in X_j. The rejection of H_0 statistically implies that the pollutant has a significant adverse health effect.

2.2. 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 between 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 by using all the q PCs. However, a large part of this variability can be explained by using a lower number r of components ($r \leqslant q$); see Johnson and Wichern (2007).

Consider the following pairs of eigenvalues and eigenvectors of the covariance matrix $\Sigma_{\mathbf{X}}$ of the random vector $\mathbf{X}=(X_1,\ldots,X_q)^{\mathrm{T}}:(\lambda_1,\mathbf{a}_1),(\lambda_2,\mathbf{a}_2),\ldots,(\lambda_q,\mathbf{a}_q)$, where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_q$. The ith PC of $\Sigma_{\mathbf{X}}$ is

$$Z_i=\mathbf{a}_i^{\mathrm{T}}\mathbf{X}=a_{1i}X_1+a_{2i}X_2+\ldots+a_{qi}X_q, \tag{2}$$

$i=1,2,\ldots,q$, with the properties

$$\begin{aligned}\operatorname{var}(Z_i)&=\mathbf{a}_i^{\mathrm{T}}\Sigma_{\mathbf{X}}\mathbf{a}_i=\lambda_i,\\ \operatorname{cov}(Z_i,Z_j)&=\mathbf{a}_i^{\mathrm{T}}\Sigma_{\mathbf{X}}\mathbf{a}_j=0,\end{aligned}$$

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

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

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

and

$$\operatorname{cov}(Z_{it},Z_{j(t+h)})=\mathbf{a}_i^{\mathrm{T}}\operatorname{cov}(\mathbf{X}_t,\mathbf{X}_{t+h})\mathbf{a}_j=\mathbf{a}_i^{\mathrm{T}}\Gamma_{\mathbf{X}}(h)\mathbf{a}_j, \tag{4}$$

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

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

2.3. Generalized additive modelling and principal component analysis

One of the research directions that is developed in this paper is the combined use of the PCA technique and a GAM, which is denoted here as GAM–PCA. 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 a 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 $Z_{1t},\ldots,Z_{qt}$ that are generated by the PCA are linear combinations of the original variables $X_{1t},\ldots,X_{qt}$. Mathematically, $Z_{it}=\mathbf{a}_i^{\mathrm{T}}\mathbf{X}_t$, similarly to equation (2), but the PCs are now time dependent for all $i=1,\ldots,q$. These new covariates are used in the GAM. Let $r\leqslant q$ and, considering the first r pairs of eigenvalues and eigenvectors of the covariance matrix $\Sigma_{\mathbf{X}}$, define the matrices $\Lambda_r:=\operatorname{diag}(\lambda_1,\ldots,\lambda_r)$ and $A_r:=(\mathbf{a}_1,\ldots,\mathbf{a}_r)$, i.e. the eigenvectors form columns of matrix A_r. We can see that A_r is an orthogonal matrix of dimension $q \times r$, i.e. $A_r^{\mathrm{T}} A_r = I_r$ where I_r is an identity matrix of dimension r. Moreover, $A_r^{\mathrm{T}} \Sigma_{\mathbf{X}} A_r = \Lambda_r$. Let $\Lambda = \Lambda_q$ and $A = A_q$. Then Λ_r is the top left-hand block of Λ of size $r \times r$ and A_r consists of the first r columns of A; see, for example, page 11 in Jolliffe (2002). Any linear combination of the first r new covariates can be expressed as a 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 $\boldsymbol{\upsilon} := (\upsilon_1, \ldots, \upsilon_r)^{\mathrm{T}}$ and $\boldsymbol{\beta}^* := (\beta_1^*, \ldots, \beta_q^*)^{\mathrm{T}}$ are vectors of dimensions r and q respectively, and the relationship between vectors $\boldsymbol{\upsilon}$ and $\boldsymbol{\beta}^*$ is given by $\boldsymbol{\beta}^* = A_r \boldsymbol{\upsilon}$ and thus $\boldsymbol{\upsilon} = A_r^{\mathrm{T}} \boldsymbol{\beta}^*$, i.e., 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 PCs is given as

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

with $r \leqslant q \leqslant p$, where $\mathbf{X}_t := (X_{1t}, \ldots, X_{qt})^{\mathrm{T}}$ 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 PCs and f_js 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

$$\begin{gathered} Y_t | \mathcal{F}_{t-1} \sim \mathrm{Po}(\mu_t), \\ \mathbf{X}_t = A \mathbf{Z}_t \end{gathered}$$

with link function (6), where $\mathrm{Po}(\cdot)$ denotes the Poisson distribution, the latent variables $\{\mathbf{Z}_t\}$ 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 $\mathbf{Z}_t = A^{\mathrm{T}} \mathbf{X}_t$ 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, a GAM is fitted by using these new variables. If $\{\mathbf{Z}_t\}$ and thus $\{\mathbf{X}_t\}$ are Gaussian processes then the joint distribution of $(Y_t, \mathbf{X}_t)$ can be expressed as a product of a Poisson and a Gaussian distribution. Thus, given a sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, the log-likelihood, up to a constant, is derived as a hybrid sum of a Poisson and a Gaussian log-likelihood:

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

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 quite complicated the maximization with respect to its parameters is more complex; hence a two-stage method is proposed. Firstly, the parameter matrices A and Λ are estimated by applying the PCA to the estimated covariance matrix $\hat{\Sigma}_{\mathbf{X}}$. Secondly, the parameters υ_0 and $\boldsymbol{\upsilon}$ are estimated by fitting the GAM with link function (6) using the first r PCs. Note that this procedure works without assuming any distribution assumption for the covariates. In the case of Gaussian covariates the maximization of the Gaussian part of the log-likelihood (7) is equivalent to the application of PCA to these covariates. In what follows, the assumption of a normal distribution for covariates is used only in computing the standard information criteria for model selection. The approach that was discussed above is similar to PC regression (see, for example, chapter 8 in Jolliffe (2002)), and it can be considered as a two-stage regression method, which is a procedure that is 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

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

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

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

where $\hat{v}_i$ is the estimated coefficient of the ith PC in equation (6) and $\hat{\mathbf{a}}_i, i=1,\ldots,r$, are the first r estimated eigenvectors. Equation (9) can be easily derived by using equation (5). Since the PCs are uncorrelated the standard error of $\hat{\beta}^*_j$ can be estimated by

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

2.4. Generalized additive modelling, principal component analysis and vector auto-regressive modelling

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

Let now $\{\mathbf{X}_t\}$, $\mathbf{X}_t=(X_{1t}, X_{2t}\ldots, X_{qt})^{\mathrm{T}}$, be a VARMA$(p^*, q^*)$ process defined as the solution to the following system (see Hamilton (1994)):

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

where B is the delay operator, γ is a q-dimensional vector and the innovation process $\{\varepsilon_t\}$ is q-dimensional white noise with $E(\varepsilon_t)=0$ and $\mathrm{var}(\varepsilon_t)=\Sigma_\varepsilon$, where Σ_ε is a $q\times q$ variance matrix. The operators $\Phi(B)=I_q-\Sigma_{i=1}^{p*}\Phi_i B^i$ and $\Theta(B)=I_q+\Sigma_{i=1}^{q*}\Theta_i B^i$ are polynomial matrices of orders p^* and q^* respectively, and the Φ_is and Θ_is are matrices of constants with dimension $q\times q$. If $\det\{\Phi(z)\}\neq 0$ for all complex z such that $|z|\leqslant 1$ then the VARMA model (10) has exactly one stationary solution; see theorem 11.3.1 in Brockwell and Davis (1991). Seasonal VARMA models are built by using the same structure as in equation (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 $\gamma=0$. Therefore, model (10) simplifies to

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

A VAR(1) process has a unique stationary solution provided that all the eigenvalues of Φ 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 $\mathbf{X}_t = \Sigma_{j=0}^{\infty} \Phi^j \varepsilon_{t-j}$; see example 11.3.1 in Brockwell and Davis (1991). The autocovariance matrix function of $\{\mathbf{X}_t\}$ is given by $\Gamma_{\mathbf{X}}(h) = \Sigma_{j=0}^{\infty} \Phi^{j+h} \Sigma_{\varepsilon} (\Phi^{\mathrm{T}})^j$, $h = 0, \pm 1, \ldots$. 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, which is usually denoted by $\text{SVAR}_s(1)$, is an extension of model (11) with a seasonal matrix auto-regressive coefficient at lag s. This seasonal matrix must satisfy a similar stationary condition to that of the VAR(1) model; see, for example, Brockwell and Davis (1991). In what follows, 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 (11), where $\mathbf{X}_t$ represents the pollution variables at time t in the context of this paper, and the GAM–PCA model by using the white noise error process (11) as covariates. Mathematically, let $Z_{1t}, \ldots, Z_{qt}$ at time t be given by

$$Z_{it} = \mathbf{a}_i^{\mathrm{T}} \varepsilon_t = \mathbf{a}_i^{\mathrm{T}} (\mathbf{X}_t - \Phi \mathbf{X}_{t-1}), \qquad i = 1, \ldots, q, \tag{12}$$

where $(\lambda_i, \mathbf{a}_i)$, $i = 1, \ldots, q$, denote the first r eigenvalues and eigenvectors of the variance matrix Σ_{ε} of the white noise innovation in equation (11), and, therefore, the PC vector $\mathbf{Z}_t$ 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 effect of the VAR(1) filter in the GAM–PCA–VAR model is to eliminate the serial correlation in the original pollutant covariates. Large positive values in a co-ordinate of the innovation ε_t indicate locally high environmental influence according to this pollutant at time t. In contrast, large negative values indicate negligible influence. The GAM–PCA–VAR model that is based on the first r PCs is defined by

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

which clearly shows that, in contrast with GAM–PCA, Y_t depends on both $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$, 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

$$\begin{gathered} Y_t | \mathcal{F}_{t-1} \sim \text{Po}(\mu_t), \\ \mathbf{X}_t = \Phi \mathbf{X}_{t-1} + A \mathbf{Z}_t \end{gathered}$$

with link function (13), where the latent variables $\{\mathbf{Z}_t\}$ 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 $\mathbf{Z}_t = A^{\mathrm{T}}(\mathbf{X}_t - \Phi \mathbf{X}_{t-1})$ 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 $\varepsilon_t = \mathbf{X}_t - \Phi \mathbf{X}_{t-1}$ at time t; in the second stage new uncorrelated variables (PCs) $\{\mathbf{Z}_t\}$ are derived by using the PCA for the innovation process $\{\varepsilon_t\}$, and in the third stage a GAM is fitted by using the first r PCs as covariates. The order of models in the term GAM–PCA–VAR corresponds to these stages starting with the third and finishing with the first, 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 $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, is derived as

$$l(\upsilon_0, \boldsymbol{\upsilon}, A, \Lambda, \Phi) \propto \sum_{t=2}^{n} \{Y_t \ln(\mu_t) - \mu_t\} - \frac{1}{2} \sum_{t=2}^{n} (\mathbf{X}_t - \Phi\mathbf{X}_{t-1})^{\mathrm{T}} A \Lambda^{-1} A^{\mathrm{T}} (\mathbf{X}_t - \Phi\mathbf{X}_{t-1}) - \frac{n-1}{2} \ln\{\det(\Lambda)\}, \tag{14}$$

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

Remark 1. Another model, called hereafter GAM–VAR–PCA, can be derived by interchanging the order of the VAR filter and PCA. Namely, the multicollinearity between 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 $\mathbf{Z}_t^{(r)} = A_r^{\mathrm{T}}\mathbf{X}_t$ for all t. We fit a VAR(1) model to the r-dimensional process $\{\mathbf{Z}_t^{(r)}\}$, i.e. $\mathbf{Z}_t^{(r)} = \Psi_r \mathbf{Z}_{t-1}^{(r)} + \mathbf{W}_t^{(r)}$, where Ψ_r is a matrix of dimension $r \times r$ and $\{\mathbf{W}_t^{(r)}\}$, $\mathbf{W}_t^{(r)} = (W_{1t}^{(r)}, \ldots, W_{rt}^{(r)})^{\mathrm{T}}$, is an r-dimensional white noise process. The link function of the GAM–VAR–PCA model is

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

which looks like equation (13). Nevertheless, there is an important difference between these two formulations. Whereas, in the GAM–PCA–VAR model, the vector $\mathbf{Z}_t^{(r)} = (Z_{1t}, \ldots, Z_{rt})^{\mathrm{T}}$ in equation (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 $\mathbf{W}_t^{(r)}$ in equation (15) is also a white noise process but its components are not necessarily uncorrelated. Hence, the new covariates of the GAM–VAR–PCA model that are involved in 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 correct alternative.

2.5. Goodness of fit

A comparison of the procedures proposed 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 MSE is defined as

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

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

$$\mathrm{AIC}=-2\hat{l}+2k,$$
$$\mathrm{BIC}=-2\hat{l}+k\ln(n),$$

where $\hat{l}$ is the maximized value of the log-likelihood function defined by expressions (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 models 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 resulting from the proposed two- and three-stage estimation methods for GAM–PCA and GAM–PCA–VAR models respectively.

3. Results

3.1. Simulation study

To evaluate the effect in the parameter estimation, and hence in the RR-estimates, of a GAM in the presence of temporal correlation in both the dependent Y_t and independent $\mathbf{X}_t=(X_{1t},\ldots, X_{qt})^{\mathrm{T}}$ vector, a simple simulation study was conducted. The data were generated under three scenarios: independent data (scenario 1); the dependent variable is a time series and the covariates are independent random vectors in time (scenario 2); both the dependent and independent variables are time series (scenario 3). For the three scenarios, the data were generated from a conditional Poisson model, $Y_t|\mathbf{X}_t\sim\mathrm{Po}(\mu_t)$.

Initially, only one covariate X_1 was considered. In this case, for (scenario 1), the predictor is given by $\log(\mu_t)=\beta_0+\beta_1X_{1t}$ 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 $\log(\mu_t)=\beta_0+\beta_1X_{1t}+\epsilon_t$, where $\{\epsilon_t\}\sim\mathrm{AR}(1)$ with auto-regressive coefficient $\varphi=0.1,0.5,0.9$. The difference between scenarios 2 and 3 is that, for the first, $X_{1t}\sim N(0,1)$ for all t and, for the latter, $\{X_{1t}\}\sim\mathrm{AR}(1)$ 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 is $n=100$ and the number of Monte Carlo simulations was equal to 1000. The empirical values of the mean, bias and MSE are displayed in Table 1. All results were obtained by using R code.

In the case of independent data (scenario 1), 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 auto-correlation structure 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 tends to underestimate β_1 severely. As 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 auto-correlation structure increases. Hence, the correlation structure 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 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 relationship between the variables.

Next, we evaluate the effect in the parameter estimation of a GAM when there are two

Table 1. Simulation results for a single covariate

Scenario	*Parameter*	*Mean*	*Bias*	*MSE*
1, independent	$\beta_0 = 1$	0.9958	−0.0042	0.0049
	$\beta_1 = 1.5$	1.5010	0.0010	0.0026
2, $\varphi = 0.1$	$\beta_0 = 1$	1.4873	0.4873	0.2921
	$\beta_1 = 1.5$	1.4457	−0.0543	0.0671
2, $\varphi = 0.5$	$\beta_0 = 1$	1.6084	0.6084	0.4782
	$\beta_1 = 1.5$	1.4091	−0.0909	0.1116
2, $\varphi = 0.9$	$\beta_0 = 1$	2.7779	1.7779	4.7168
	$\beta_1 = 1.5$	1.3189	−0.1811	0.2544
3, $\varphi = 0.1$	$\beta_0 = 1$	1.4732	0.4732	0.3673
	$\beta_1 = 1.5$	1.3903	−0.1097	0.1180
3, $\varphi = 0.5$	$\beta_0 = 1$	1.6512	0.6512	0.5727
	$\beta_1 = 1.5$	1.3790	−0.1210	0.1528
3, $\varphi = 0.9$	$\beta_0 = 1$	2.8475	1.8475	5.0797
	$\beta_1 = 1.5$	1.2518	−0.2482	0.2918

Table 2. Simulation results for two covariates X_1 and X_2

Scenario	*Parameter*	*Mean*	*Bias*	*MSE*
1, 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
2	$\beta_0 = 1$	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
3, $\phi_{11} = 0.7$, $\phi_{12} = 0$,	$\beta_0 = 1$	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
3, $\phi_{11} = 0.7$, $\phi_{12} = 0.4$,	$\beta_0 = 1$	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

covariates, $\mathbf{X}_t = (X_{1t}, X_{2t})^{\mathrm{T}}$. The set-up is the same as described previously for scenarios 1, 2 and 3, with two covariates instead of a single one. Thus, under scenario 1 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 scenarios 2 and 3, the predictor is given by $\log(\mu_t) = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t$, where $\{\epsilon_t\} \sim \mathrm{AR}(1)$, with $\varphi = 0.5$. Now the difference between scenarios 2 and 3 is that, for the first, $X_{1t}, X_{2t} \sim N(0,1)$, mutually independent, and, for the latter, $(X_{1t}, X_{2t})^{\mathrm{T}}$ forms a VAR(1) process with auto-regressive coefficient matrix Φ of dimension 2×2. The results are displayed in Table 2.

From Table 2, similar conclusions are drawn to those in the case of a single covariate (Table 1), i.e. 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 with 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 equation

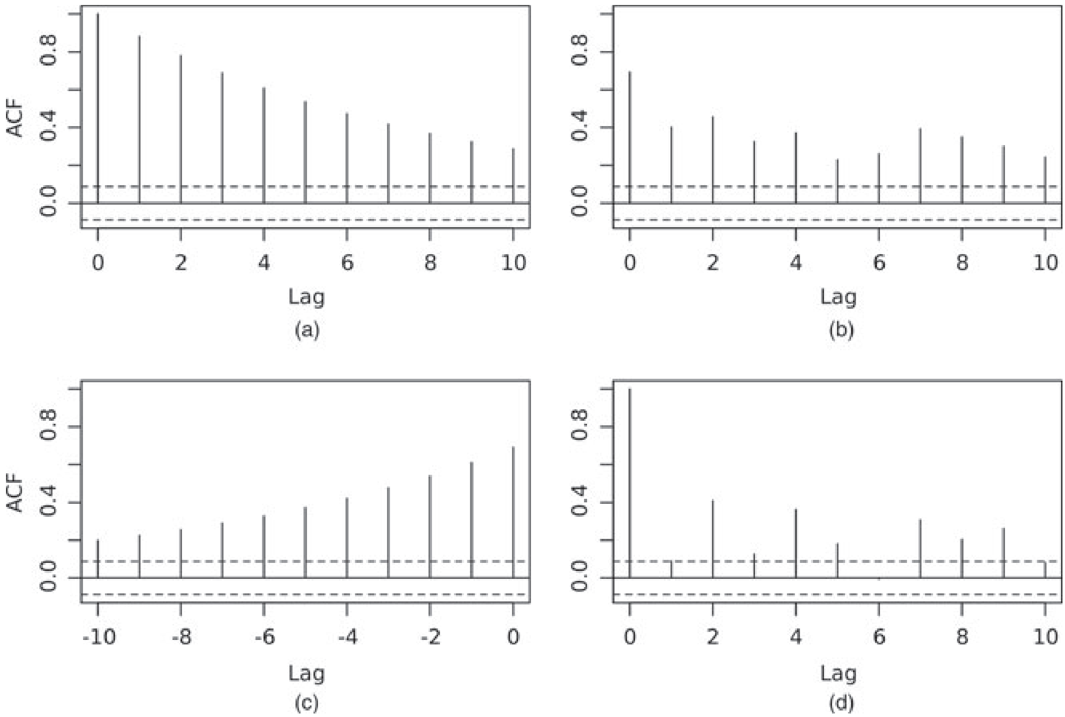

Fig. 1. Sample auto-correlation function and cross-correlation function of the PCs: (a) PC 1; (b) PC 1 × PC 2; (c) PC 2 × PC 1; (d) PC 2

(4). For this one sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_{500}\}$ was generated from the process $\{\mathbf{X}_t\}$ in equation (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, i.e. $\hat{Z}_{1t}, \hat{Z}_{2t}, t = 1, \ldots, 500$, were computed from the 2×2 sample covariance matrix of $\{\mathbf{X}_1, \ldots, \mathbf{X}_{500}\}$. The sample correlation and cross-correlation functions of the PCs are displayed in Fig. 1, in which $\hat{Z}_{1t}$ and $\hat{Z}_{2t}$ correspond to PC 1 and PC 2 respectively. As can be seen, the plots clearly indicate that the correlation structure of the models is transferred to the PCs as shown in equation (4). On the basis of the above empirical evidence as well as on the discussion of the previous sections, it is clear that the temporal correlation cannot 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 Section 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 daily hospital admissions for respiratory diseases was obtained from the main childrens' emergency department in the Vitória metropolitan area (called the Hospital Infantil Nossa Senhora da Gloria). Respiratory diseases are classified according to

Table 3. Descriptive statistics for the variables under study (Vitória metropolitan area, January 2005–December 2010)

Statistic	*Mean*	*Standard deviation*	*Minimum*	*25th percentile*	*50th percentile*	*75th percentile*	*Maximum*
PM_{10} ($\mu g\ m^{-3}$)	33.45	8.83	8.98	27.90	32.75	38.39	86.74
SO_2 ($\mu g\ m^{-3}$)	12.44	3.11	4.89	10.06	12.16	14.57	26.48
O_3 ($\mu g\ m^{-3}$)	31.86	8.36	12.10	25.97	30.73	36.58	72.34
NO_2 ($\mu g\ m^{-3}$)	24.82	6.93	9.03	19.59	24.13	29.37	62.59
CO ($\mu g\ m^{-3}$)	885.79	231.28	295	724.82	866.60	1031.09	2141.50
Minimum temperature (°C)	20.86	2.47	13.10	19.08	21.15	22.80	25.98
Average temperature (°C)	24.43	2.45	17.00	22.62	24.40	26.35	30.80
Maximum temperature (°C)	29.35	3.28	19.40	27.20	29.41	31.60	39.70
Air relative humidity (%)	77.43	6.03	61.60	73.24	77.19	81.14	97.28
Number of treatments for respiratory diseases	27.09	6.15	1.00	13.00	24.00	37.00	121.00

the 10th international classification of diseases, and the group investigated consisted of children under 6 years old. The study was performed between January 1st, 2005, and December 31st, 2010 ($n=2191$).

The following atmospheric pollutants were studied: PM_{10} particulate matter, SO_2, NO_2, O_3 and CO. Information on the daily levels of these pollutants and data for meteorological variables were obtained from the State Environment and Water Resources Institute, where the data were collected at eight monitoring stations.

The data collection for all the pollutants occurred over a 24-h period that began in the first half-hour of the day. The following data were obtained at each station: the 24-h average concentration for PM_{10} and SO_2, 8-h moving average concentrations for CO and O_3, and the 24-h 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 that are 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 that were considered exceeded neither the primary air quality standard recommended by the Brazilian National Council for the Environment, nor the guidelines suggested by the World Health Organization. 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).

The average maximum temperature that was 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%.

Fig. 2 shows that the series of air pollutants concentration and the number of hospital admissions for respiratory diseases have seasonal behaviour, which was to be expected for these phenomena. Another characteristic that was observed in the series was an apparently weak stationarity. This result is confirmed in the graphs of the sample auto-correlation functions shown in Fig. 3.

Table 4 shows the correlations between 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 between 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

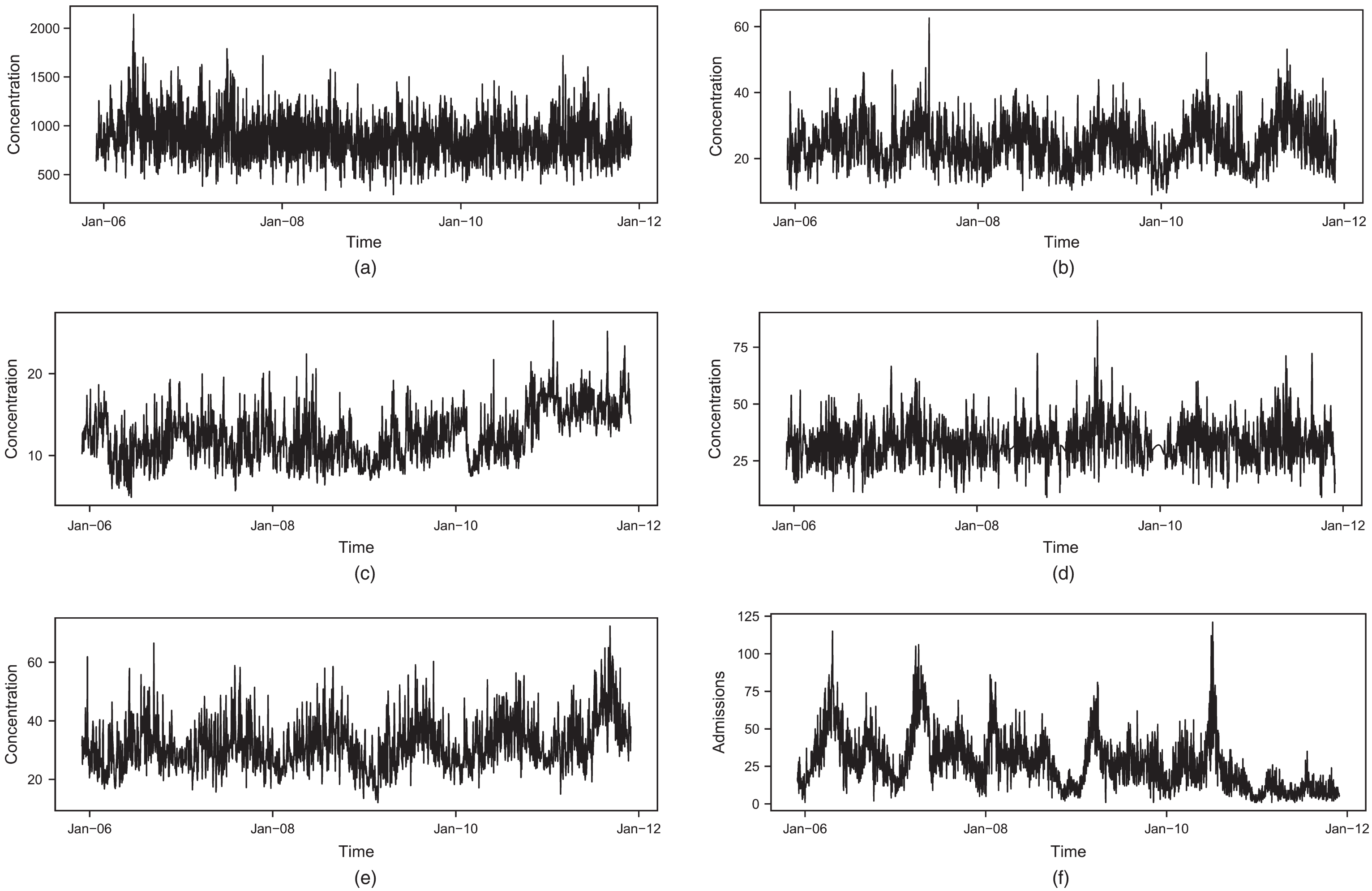

Fig. 2. Concentration of (a) CO, (b) NO_2, (c) SO_2, (d) PM_{10} and (e) O_3, and (f) number of treatments for respiratory diseases

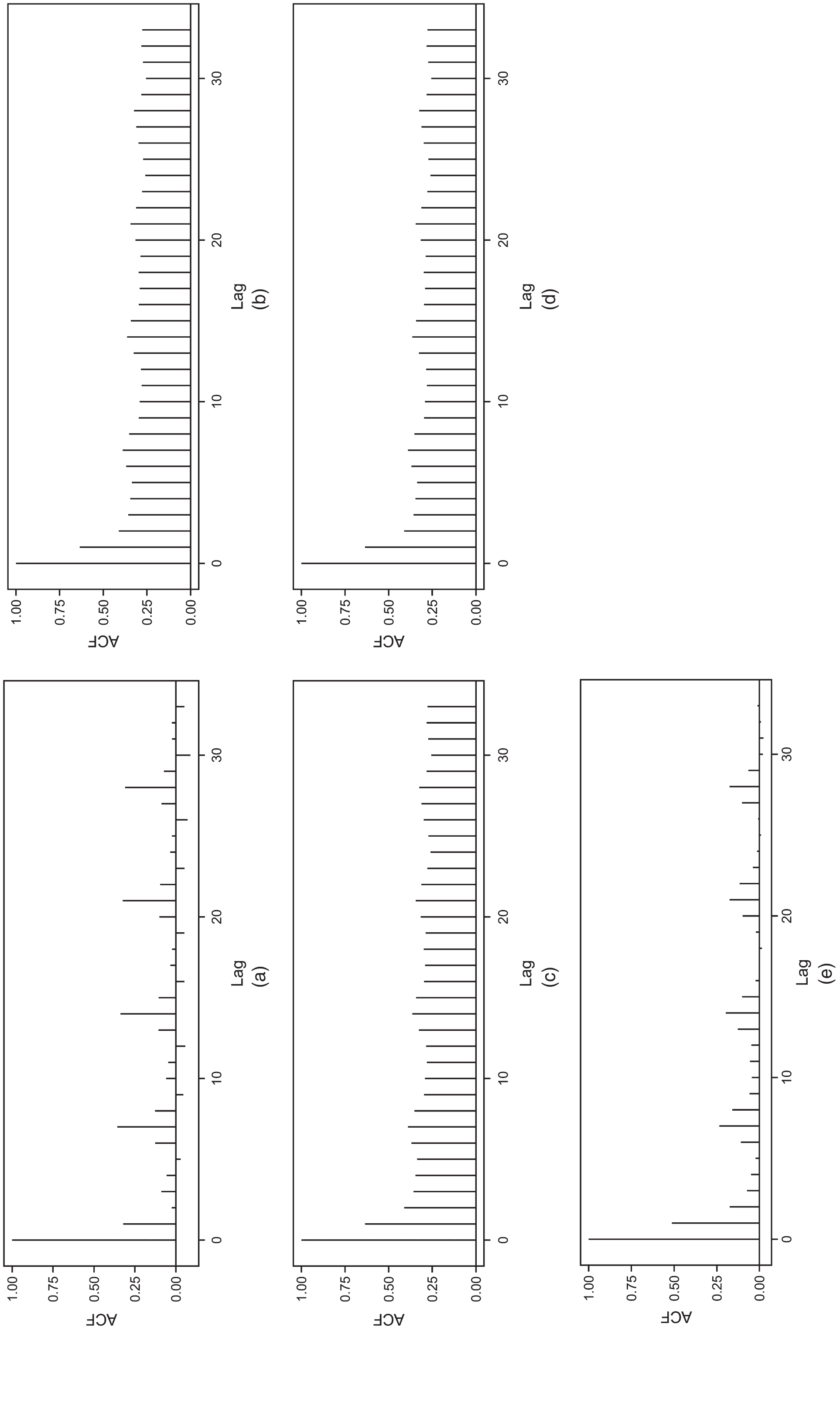

Fig. 3. Sample auto-correlation function of the pollutants: (a) CO; (b) O_3; (c) SO_2; (d) NO_2; (e) PM_{10}

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

	PM_{10}	SO_2	NO_2	CO	O_3	T(max)	T(min)	RH	Number of treatments
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	
Number of treatments	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 (%); all correlations were significant at a 5% level.

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

	PC 1	*PC 2*	*PC 3*	*PC 4*	*PC 5*
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.
‡Possible cluster.

found for the correlation between PM_{10} and 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 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, to keep the notation consistent with the previous sections, PC 1, ..., PC 5 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 PCs 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, a double-dagger symbol indicates the possible clusters for each PC.

Fig. 4 shows the time behaviour of some PCs obtained from the pollutant concentration series, i.e. the original data. Fig. 4 shows that PC 1 is auto-correlated and that the cross-correlations are

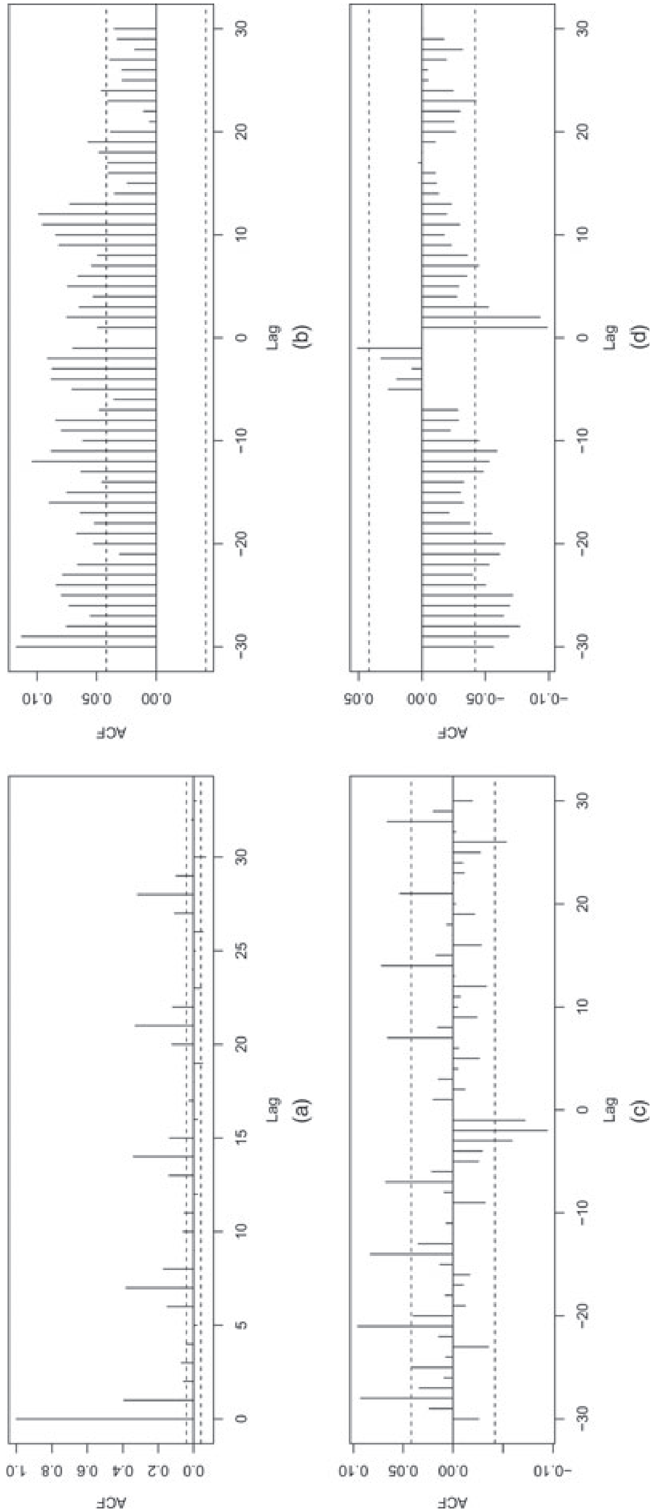

Fig. 4. Cross-correlation function of the main components of the pollutants studied: (a) PC 1; (b) PC 1 × PC 2; (c) PC 1 × PC 3; (d) PC 2 × PC 3

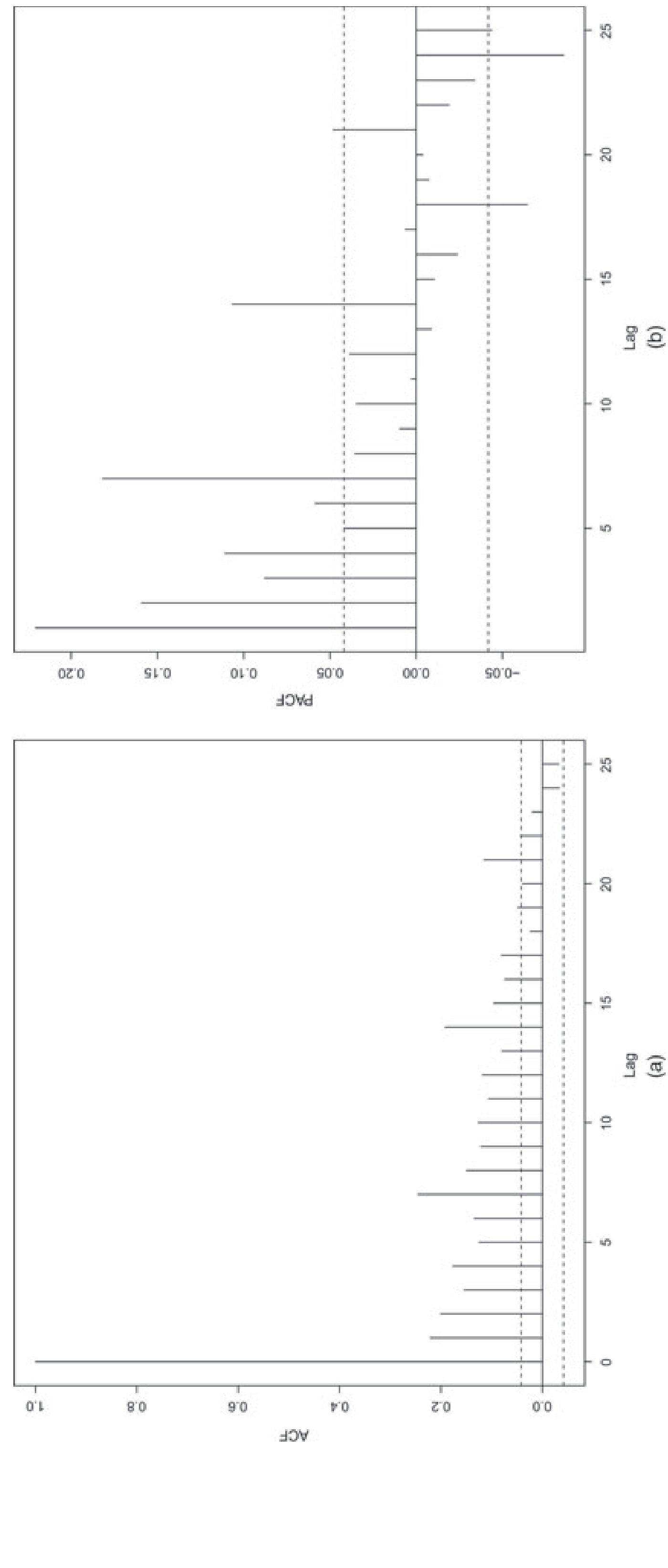

Fig. 5. (a) Sample auto-correlation function and (b) sample partial auto-correlation function of the residuals of the GAM–PCA model

Table 6. Results of the final GAM–PCA model to estimate the effects of pollutants concentrations on hospital admissions in the Vitória metropolitan area

Variable†	*Estimate*	*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‡
Holiday 2	0.1886	0.0440	4.29	0.0000‡
Holiday 3	0.3189	0.0384	8.30	0.0000‡
Air relative humidity	−0.0061	0.0009	−6.83	0.0000‡
PC 1	−0.0244	0.0040	−6.16	0.0000‡
PC 2	0.0163	0.0055	2.99	0.0028§
PC 3	−0.0157	0.0056	−2.79	0.0052‡

†Holiday 2, Corpus Christ plus Our Lady of Penha; holiday 3, carnival plus holiday (Tiradentes day) plus Brazil's Independence day.
‡Significant at the 0.001 level.
§Significant at the 0.01 level.

non-null, corroborating the results that were discussed in Sections 2.2 and 3.1. The components also clearly exhibited the seasonal behaviour of the pollution variables, as expected, i.e. the graphs show that the auto-correlation structure of the pollutants persists in the components. Therefore, the PCA technique should be applied carefully even for processes with an apparently weak auto-correlation 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 criterion for the components to be included in the GAM. Thus, following the parsimony criterion, the first three components were chosen as covariates (highlighted in italics 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 on the basis of the assumption that the count of respiratory disease events (i.e. hospital admissions) followed a conditional Poisson distribution.

The analysis involved several procedures implemented in stages. Initially, the short-term seasonality was treated by using indicator variables for week days 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 by using spline smoothing curves (see Friedman (1991) and Wahba (2001)). The best GAM–PCA fit was obtained on the basis of a residual analysis and the Akaike information criterion AIC (Akaike, 1973).

As previously mentioned, the unfiltered PCs are auto-correlated. Consequently, this property was transferred to the residuals of the GAM–PCA model (Fig. 5). Therefore, as a post-processing

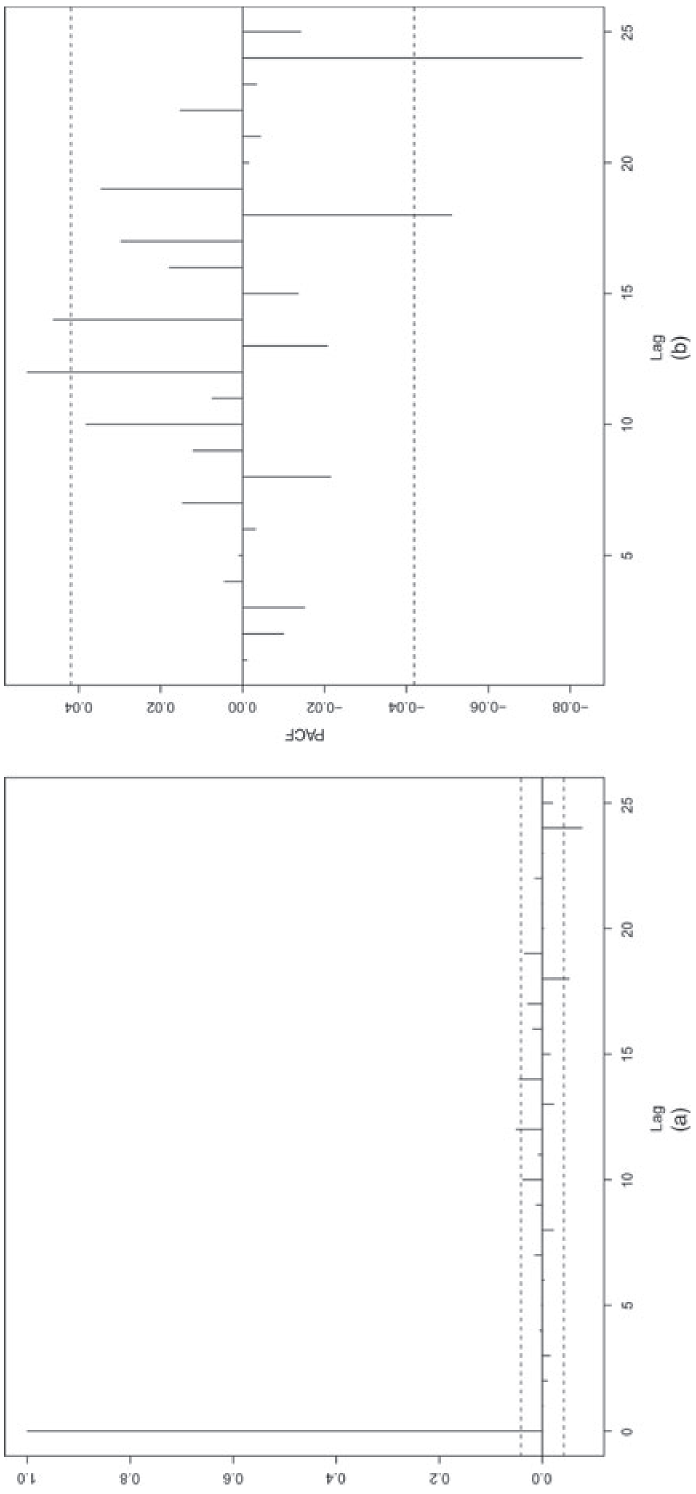

Fig. 6. (a) Sample auto-correlation function and (b) partial auto-correlation function of the residuals of the final GAM–PCA model

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

	PC 1	*PC 2*	*PC 3*	*PC 4*	*PC 5*
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

†The standard deviation is the square root of the eigenvalue.
‡Possible cluster.

step, a seasonal auto-regressive $SAR(1)(1)_7$, model was fitted to the residuals of the GAM–PCA, resulting in the final GAM–PCA model with SAR residuals, briefly the final GAM–PCA model, which has eliminated the auto-correlation 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. Fig. 6 shows that there is no auto-correlation 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 that is proposed in this paper, a seasonal VAR model with a 7-day period, $SVAR_7(1)$, was used to adjust the pollutant vector. Although the model that was discussed in the previous section is related to VAR(1), its extension using an $SVAR_7(1)$ model instead is straightforwardly obtained. The seasonal VAR(1) model was selected on the basis of the standard fitting tools of multivariate time series models, 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, i.e. the variability in the first three components explains 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 sufficiently strong to produce an effect 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 a double-dagger symbol in the analysis. The results showed cross-correlations between the NO_2 and O_3 pollutants that were 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 NO_2.

Fig. 7 shows that the fitting of the seasonal VAR(1) model virtually eliminated the auto-correlation of PC 1 and the cross-correlation, as expected from the aforementioned discussion. The residual plots (auto-correlation function and partial auto-correlation function) of GAM–PCA–VAR displayed similar behaviour to that of the GAM–PCA with SAR residuals (the final GAM–PCA) shown in Fig. 6. These plots are available on request. Additionally, Fig. 8 shows the fit (predicted values) that was obtained by using the GAM–PCA–VAR model. This graph

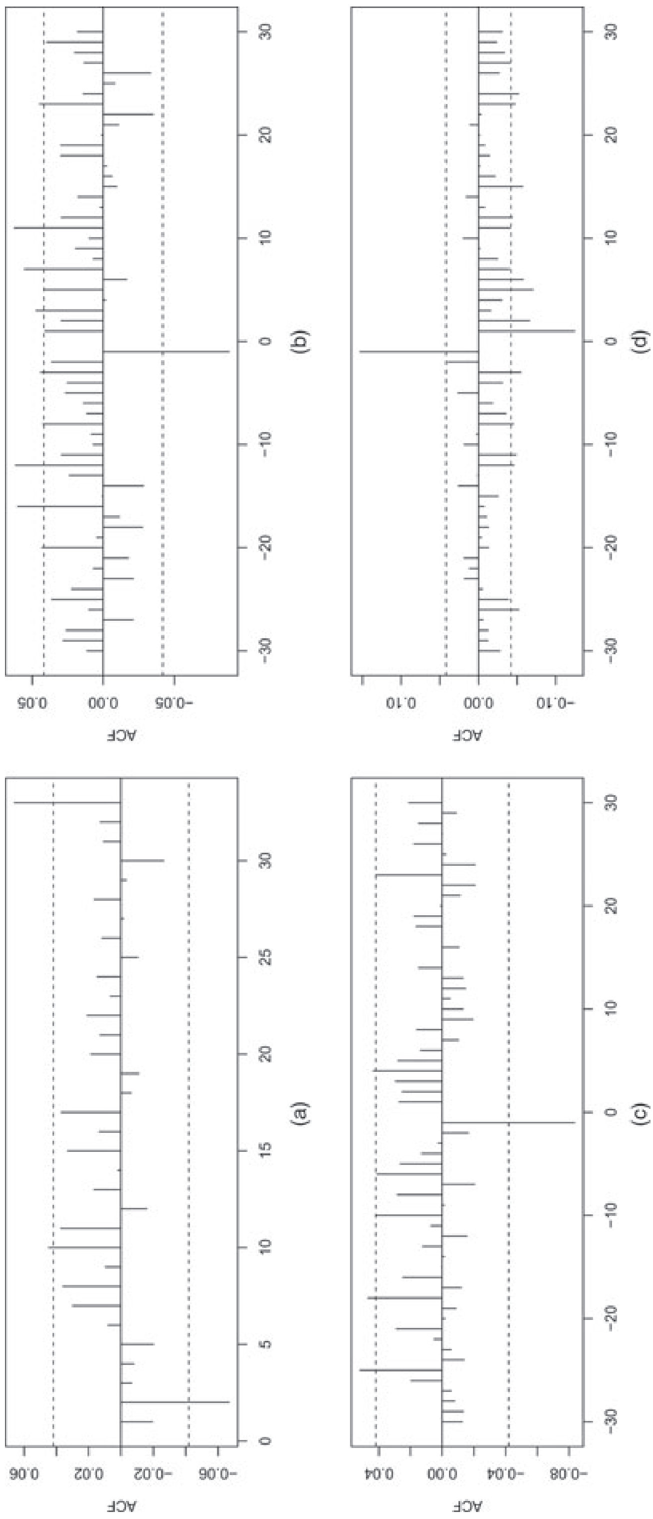

Fig. 7. Cross-correlation function of the main components of the filtered pollutants: (a) PC 1; (b) PC 1 × PC 2; (c) PC 1 × PC 3; (d) PC 2 × PC 3

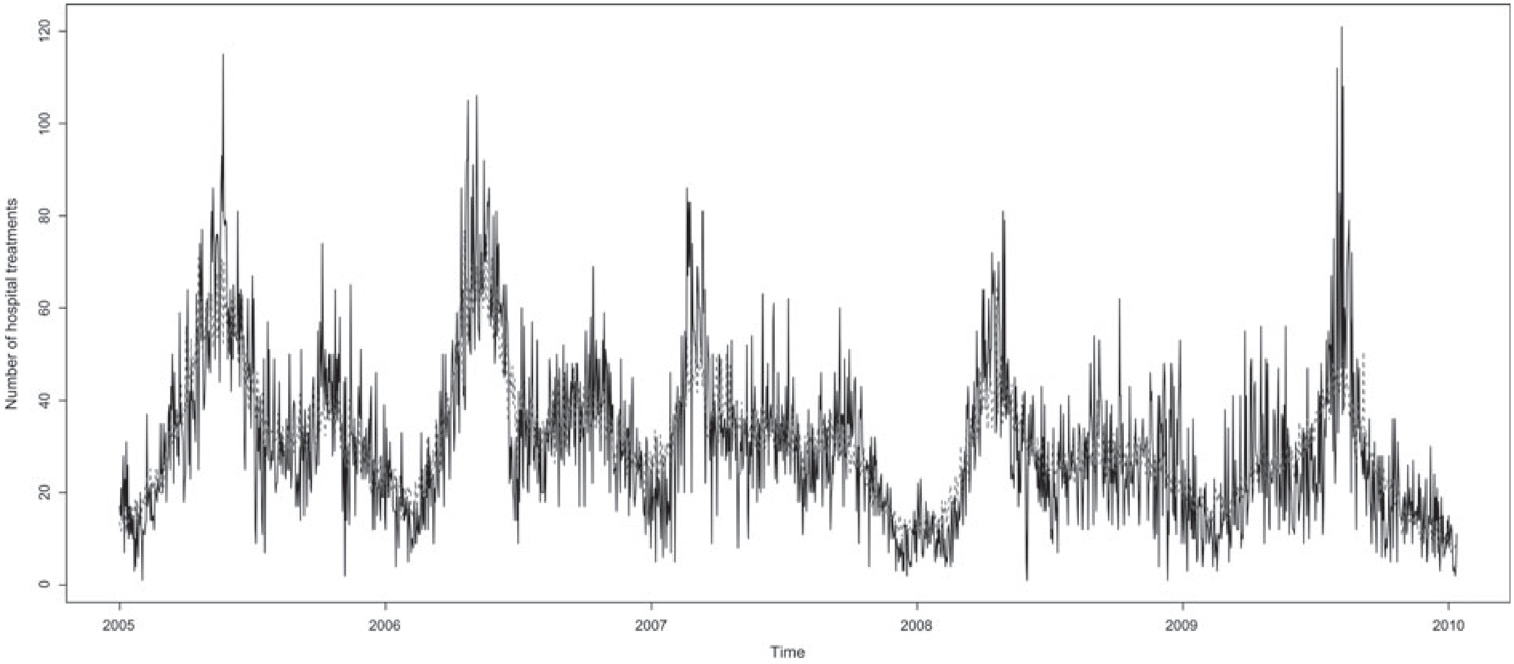

Fig. 8. Fitted GAM–PCA–VAR model to the number of treatments for respiratory disease: ———, orginal series; · · · · · ·, adjusted series

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

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 January 2005 to December 2010†

Pollutant	$\widehat{RR}$	$\widehat{RR}^*$	$\widehat{RR}^{**}$
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)

†$\widehat{RR}$, GAM; $\widehat{RR}$*, GAM–PCA; $\widehat{RR}$**, GAM–PCA–VAR.

shows that the model provided a good fit to the data for the variable of interest, i.e. the number of daily treatments for children under 6 years old in the Vitória metropolitan area.

The goodness-of-fit results for the three models (GAM, GAM–PCA and GAM–PCA–VAR) by using the MSE-, AIC- and BIC-statistics, are given in Table 8. MSE for the GAM model is approximately 35% higher than MSE for the other two models, which was an expected result since a more complex model may yield a better residual fit. AIC and BIC indicate that the GAM–PCA–VAR model is the best to fit the data. All these empirical analyses, i.e. the plots of the auto-correlation function and partial auto-correlation function of the residuals which were shown to be uncorrelated, the behaviour of the estimated PCs (Fig. 7), the final fit (Fig. 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 these data. The final performance of this procedure to quantify the association between respiratory diseases 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 performances of the GAM ($\widehat{RR}$), GAM–PCA ($\widehat{RR}^*$) and GAM–PCA–VAR ($\widehat{RR}^{**}$) 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 the models considered, 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 by 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% ($\widehat{RR}$) to 3% ($\widehat{RR}^*$) and 7% ($\widehat{RR}^{**}$). Substantial increases in the RR-estimates were also observed for the pollutant CO. In this case, $\widehat{RR} = 1.020$, $\widehat{RR}^* = 1.048$ and $\widehat{RR}^{**} = 1.077$.

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.

4. Conclusion

A hybrid of three statistical tools, the VAR model, PCA and the 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. Because of the complexity of the real data, a marginal Poisson assumption would not be the most appropriate choice in this case since the series presented an overdispersion problem, which may come from many features of the data such as changes in the mean and variance, and observations with high levels (which increase substantially the variance) among others. Overdispersion is common in this kind of data, and the negative binomial and the generalized Poisson 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 the negative binomial or generalized Poisson alternatives. However, the Poisson distribution is 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 that were included in the fit. Therefore, the use of a model that also handles overdispersion is an interesting and important issue to be considered in the context of the data analysed here. 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 our future work.

The models developed were denoted here by GAM–PCA and GAM–PCA–VAR. The first model used the PCs of the original pollutants as covariates in the GAM. The residuals of this model were fitted by 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, resulting in the hybrid model defined as GAM–PCA–VAR. In this model, the auto-correlation and cross-correlations of the PCs were removed by the VAR model.

A simulation study was conducted to evaluate the effect in the parameter estimates of GAMs when the explanatory variables have serial correlation. The results showed that, if the autocorrelation 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 population health can be underestimated if the model is not correctly adjusted. This issue was also recently explored in a different scenario by Dionisio *et al.* (2016).

The adequacy of fit of the aforementioned models was compared by means of goodness-of-fit statistics, such as MSE, AIC and BIC. On the basis of 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 relative risk RR of the GAM, GAM–PCA and GAM–PCA–VAR regression models. In general, the RR-estimates were significant for all the models that were considered in the study. It should be stressed here that, in most cases, the estimated RR is larger for GAM–PCA–VAR when compared with the GAM. This can be explained by the results obtained in the simulation study. Thus, the real effect of these pollutants in a number of respiratory diseases can be underestimated if we use the standard GAM under an inappropriate scenario as was the case of the data used here. For example, for the pollutant PM_{10}, the estimated relative risk increased from approximately 2% ($\widehat{\mathrm{RR}}$) to 3% ($\widehat{\mathrm{RR}^*}$) and 7% ($\widehat{\mathrm{RR}^{**}}$). For the GAM–PCA model, an increase of $10.49\,\mu\mathrm{g\,m}^{-3}$ (interquartile range) of the particulate material (PM_{10}) resulted in an $\widehat{\mathrm{RR}}^*$-value of 1.029 with 95% confidence interval (1.001,1.09), whereas for the GAM–PCA–VAR model a higher $\widehat{\mathrm{RR}}^{**}$-value of 1.075 with 95% confidence interval (1.001,1.092). Similar interpretations could be made for the other pollutants and models developed.

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

Acknowledgements

The authors thank the following agencies for their support: the National Council for Scientific and Technological Development (the Conselho Nacional de Desenvolvimento Científico e Tecnológico), the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Espírito Santo State Research Foundation (Fundação de Amparo à Pesquisa do Espírito Santo) and the Minas Gerais State Research Foundation (the Fundação de Amparo à Pesquisa do Estado de Minas Gerais). Pascal Bondon thanks 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 their financial support.

The authors are very grateful to the two referees and the Joint Editor for their suggestions that led to a markedly improved paper.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: 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**, 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**, 273–278.

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

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**, 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. Hazrd. Mater.*, **181**, 234–240.

Cotta, H., Reisen, V., Bondon, P. and Stummer, W. M. (2017) Robust estimation of covariance and correlation functions of a stationary multivariate process. In *Proc. Int. Conf. Time Series*, *Granada*.

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. Hlth*, **15**, article 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. Epidem.*, **156**, 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. Ass.*, **295**, 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. Epidem. Commty Hlth*, **59**, 881–884.

Friedman, J. (1991) Multivariate adaptive regression splines. *Ann. Statist.*, **19**, 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**, 531–534.

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

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

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

Johnson, R. A. and Wichern, D. W. (2007) *Applied Multivariate Statistical Analysis*, 6th edn. Englewood Cliffs: 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. New York: 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. Hlth Perspect.*, **119**, 455–460.

Matteson, D. S. and Tsay, R. S. (2011) Dynamic orthogonal components for multivariate time series. *J. Am. Statist. Ass.*, **106**, 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. Hlth*, **6**, no. 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. Hlth Perspect.*, **107**, 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. and Toledo de Alburquerque, T. (2017) Long-memory model under outliers: an application to air pollution levels. In *Environmental Science and Engineering: Air and Noise Pollution*, vol. 3 (eds B. R. Gurjar, P. Kumar and J. N. Govil), pp. 211–243. New Delhi: Studium.

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

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

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

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

Wahba, G. (2001) Splines in nonparametric regression. In *Encyclopedia of Environmetrics*, vol. 4, 2nd edn (eds A. H. El-Shaarawi and W. W. Piegorsch), pp. 2099–2112. New York: 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. Engng Mangmnt*, **2**, 253–259.

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

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. *Biometr. J.*, **49**, 565–584.

Zamprogno, B. (2013) PCA in time series with short and long-memory time series. *PhD Thesis*. Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, Universidade Federal do Espirito Santo, Vitória.

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. Epidem.*, **159**, 702–706.