Use of multivariate time series techniques to estimate the impact of particulate matter on the perceived annoyance

Milena Machado, Valdério Reisen, Jane Meri Santos, Neyval Costa Reis Júnior, Séverine Frère, Pascal Bondon, Márton Ispány, Higor Cotta

Use of multivariate time series techniques to estimate the impact of particulate matter on the perceived annoyance

Milena Machado[a]*, Valdério Anselmo Reisen[bce], Jane Meri Santos[c], Neyval Costa Reis Junior[c], Severine Frère[d], Pascal Bondon[e], Márton Ispány[f], Higor Henrique Aranda Cotta[be]

[a] Instituto Federal de Ciência e Tecnologia do Espírito Santo, Guarapari -E.S.- Brazil

[b] Department of Statistics, Universidade Federal do Espírito Santo, Vitoria, Brazil

[c] Department of Environmental Engineering, Universidade Federal do Espírito Santo, Vitoria, Brazil

[d] Université du Littoral Côte d'Opale, Maison de la Recherche en Science de l´homme, Dunkerque, France

[e] Laboratoire des Signaux et Systems (L2S), CNRS-CentraleSupélec-Université Paris-Sud, Gif-sur-Yvette, France

[f] University of Debrecen, Debrecen, Hungary

* Tel: +55 (27) *988527717*, e-mail: milenamm@ifes.edu.br

Abstract

As well known, Particulate matter (PM) is an air pollutant that causes damage to the health of humans, other animals, plants, affects the climate and is a potential cause of annoyance through deposition on various surfaces. The perceived annoyance caused by particulate matter is related mainly to the increase of settled dust in urban and residential environments. PM can originate from many sources, i.e., paved and unpaved roads, buildings, agricultural operations and wind erosion represent the largest contributions beyond the relatively minor vehicular and industrial sources emissions. The aim of this paper is to quantify the relationship between perceived annoyance and particulate matter concentration and to estimate the relative risk (RR). The data was collected in the Metropolitan Region of Vitoria (MRV), Brazil. *For this purpose, the variables of interest were modelled using vector time series model (VAR), principal component analysis (PCA), and logistic regression (LOG). The combination of these techniques resulted in a hybrid model denoted as LOG-PCA-VAR which allows to estimate RR by handling multipollutant effects. This study shows that there is a strong association between the*

perceived annoyance and different sizes of PM. The estimates of RR indicate that an increase in air pollutant concentrations significantly contributes in increasing the probability of being annoyed.

Key words: Annoyance, principal component analysis, logistic regression, relative risk.

1- Introduction

Particulate matter, such as dust, dirt, soot, and smoke, are environmental stressors that can cause annoyance, disturbance, stress and impairs well-being (Colls, 2002; Cox, 2000; Dockery and Pope, 1994; Farfel et al., 2005). According to Nordin and Lidén (2006), perceived annoyance can be considered as a community problem even if only a small proportion of the population is annoyed on sparse occasions. The World Health Organization (WHO, 1946) defines health as a state of complete physical, mental and social well-being and not merely the absence of disease.

PM is formed by particles with different composition, form and sizes: ultrafine particles ($PM_{0.1}$) whose effects on human health are still poorly studied, fine particles ($PM_{2.5}$) that are housed in the terminal bronchiole, inhalable particles (PM_{10}) that penetrate the respiratory system, total suspended particles (TSP) which are represented by all particles suspended in the atmosphere (size range from 0.005μm to 100μm), and the sediment particles matter (SPM) that result from the sedimentation or deposition of particles previously suspended in the atmosphere, with different sizes and origin, that accumulate on the surfaces and cause annoyance (Holgate *et al.*, 1999).

The association between air pollutants and perceived annoyance is the subject of interest in several studies. Most of them, have considered regression models to quantify this relationship, for example, in the cases of odours (Blanes-Vidal, 2012), gases (Klaeboe *et al*., 2000; Oglesby *et al.* (2000a), and particles (Klaeboe *et al*., 2003; Rotko *et al*. 2002; Jacquemin *et al*., 2007; Llop *et al*., 2008; Klaeboe, 2008; Amundsen *et al*. 2008; Nikolopoulou *et al*., 2011).

Klaeboe *et al*. (2000) have considered logistic regression to correlate NO_2 concentration and degrees of annoyance due to traffic, and they have found that people are more likely to be annoyed when they are exposed to high air pollution levels. Oglesby *et al.* (2000) have applied a linear regression model to correlate annoyance and concentration levels of NO_2 and PM_{10}, and they have found significant correlations between these variables. Rotko *et al*. (2002) have compared exposures to $PM_{2.5}$ and NO_2 concentrations and perceived annoyance using a linear regression model, and they have observed a high correlation between personal 48h-$PM_{2.5}$ and 48h-NO_2 concentrations exposure and perceived annoyance at home. Jacquemin *et al*. (2007) have applied a linear regression, and they have found a strong positive correlation between the $PM_{2.5}$ concentration and perceived annoyance reported by people. Amundsen *et al*. (2008) have quantified exposure–response relationships between perceived annoyance and PM_{10}, $PM_{2.5}$ and NO_2 concentrations, and they have observed a significant correlation between these variables. Nikolopoulou *et al*. (2011) have used a logistic regression model to correlate air quality perception of pedestrians and $PM_{1\text{-}10}$ concentration measured on sidewalks close to streets, and they have found a positive correlation in this study.

Note that, the above-mentioned studies have applied simple linear regression and logistic regression but have not considered a synergistic effect among pollutants and perceived annoyance. As pointed out by Vanhatalo *et al*. (2016), Souza et al. (2018) among others, this analysis becomes very restrictive and may lead to biased regression estimates because air pollutants covariates are physically and statistically correlated phenomena. In addition, to estimate any multiple regression model without considering the multi-collinearity, the parameter estimates may lead to a spurious model. One way to mitigate the multi-collinearity problem is to apply principal component analysis (PCA). However, as pointed out by Zamprogno *et al*. (2019), to use PCA technique the variables have to be uncorrelated in time.

As well known, the air pollutants concentrations are time series and they can't be assumed to be temporally uncorrelated. Thus, it is necessary to use the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the pollutants to identify the existence of serial correlation, and to apply a Vector Autoregressive Model (VAR) as a filter to mitigate the temporal correlation in the covariates.

In this context, this paper proposes a combination of multivariate statistical techniques to investigate the joint effect of different sizes of particulate matter to the perceived annoyance. Thus, the combination of the statistic tools LOG model, PCA and time series analysis can lead to an estimate of the relative risk of perceived annoyance by handling multipollutant effects. The relative risk is usually the parameter of interest to measure the impact of the covariates, especially the air pollutants on the population health (Zou, 2004). The proposed methodology results in a model called LOG-PCA-VAR. To our knowledge, this is the first work which uses logistic regression with PCA and multivariate time series models to quantify the relationships between particulate matter (PM_{10}, TSP and SPM) and perceived annoyance to estimate the relative risk (RR), which is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group. In the air pollution problems, it is usually to measure the impact of atmospheric pollutants on the health of the exposed population see, for example, (Martin et al., 1987).

2- Material and methods

2.1. Metropolitan Region of Vitoria

The Metropolitan Region of Vitoria (MRV) is located on the east coast of Brazil, in the state of Espirito Santo (Figure 1). MRV is a densely populated region, with 1,500,000 inhabitants and it is a highly industrialized and expanding urban region with various air pollutants emission sources such as steel, pelletizing, mining, cement industries, vehicles, road re-suspension, port and airport operations, and construction (Santos *et al*., 2017).

In the MRV area, there is an interest to investigate the impact caused by PM due to population reports of being constantly annoyed (approximately 25% of the complaints to environmental agency in 2008 are about air pollution), specially by the amount of dust in surfaces (Souza, 2014; Melo *et al*., 2015). Recently, Machado *et al*. (2018) have developed a survey where showed that, in the MRV, more than 90% of the respondents

have complained about perceived annoyance caused by the air pollution and, the most of these complaints were related to the amount of dust in their houses.

2.2. The particulate matter data

In the MRV area the weather conditions and the air quality are monitored via two complementary sets of monitoring network stations: automatic air quality monitoring and the manual SPM monitoring. Figure 1 shows the map of the urbanized area divided by municipality (Cariacica, Serra, Viana, Vila Velha e Vitoria), the main roads, the main industrial sources of PM (point red) and the air quality monitoring stations networks (blue points). They are: (M1) Laranjeiras, (M2) Carapina, (M3) Jardim Camburi, (M4) Enseada, (M5) Vitória, (M6) Vila Velha, (M7) Ibes, (M8) Cariacica. The coverage areas are 1.5 km around of each air quality monitoring station.

The monitoring station networks are managed by the local environmental agency (IEMA) that measure automatically hourly concentrations of different pollutants, specifically the PM_{10} (particulate matter less than 10$\mu g/m^3$) and TSP (total suspend particles). The SPM (sediment particulate matter) are measured monthly only. Therefore, for a coherence analysis, the maximum mean of PM_{10} and TSP concentrations were also monthly computed and used in the regression model.

The datasets used are the flow of monthly average sediment particulate matter (SPM) as well as monthly maximum and average values of particulate matter (PM_{10}) and total suspended particle (TSP) from the eight air quality monitoring stations measured during 3 years (from July 11 to July 2014).

2.3. The Survey

Measurements of PM and perceived annoyance were performed monthly from July 11 to July 2014. Perceived annoyance was collected in two steps: face-to-face interview to the first contact with respondent and monthly telephone updates (panel survey). The face-to-face interviews randomly selected surrounding 1.5 km of each air-quality monitoring station (Figure 1). On the face-to-face interview the respondent confirmed in continuing the interviews in the following months (panel survey) about perceived annoyance (details in Machado, 2018).

The monthly panel survey questionnaire only included two questions were applied to 220 respondents (over 16 years old) from July 11 to July 2014. Telephone questions aimed at monitoring the evolution of perceived annoyance over time-related to PM in the environment.

To quantify the perceived annoyance, categorical and numerical scales were considered and applied according to the context of the question (for example, "*Do you feel annoyed by dust during this last month?*" With the categorical answers option: *not annoyed, slightly annoyed, moderate annoyed, very annoyed, extremely annoyed* and "do *not know*". And a second question with a numerical scale: "*What is the score that represents your perceived annoyance last month? from 1 to 10 points scales, where 1 is not annoyed and 10 is extremely annoyed.*"). These questions were formulated based on the following studies Rotko *et al*. (2002), Klaeboe, (2008) and Amundsen *et al.* (2008).

From these questions, the average levels of perceived annoyance reported by all respondents was calculated. The results were dichotomized to be used as the dependent variable in the logistic regression model discussed in Section 2.4. The cut-off sample score of the perceived annoyance was the median 7, i.e., the scores levels of perceived annoyance attributed high scores (≥7) was codified by 1 while the average levels of annoyed reported low scores (<7) was codified as 0. Similar approach was used by Rotko *et al*. (2002), Egondi *et al.* (2013) and Whittle *et al*. (2014).

2.4. Statistical Techniques

As previously mentioned, the main objective of this paper is to quantify the association between perceived annoyance (response) and pollutants (covariates) variables using data observed in the Metropolitan Region of Vitoria (MRV). The response variable is binary. Therefore, the logistic regression becomes the appropriate regression method to describe the association among variables. However, for this statistic model, some assumptions are required, and, among them, the covariates should be independent from each other and independent of time. And, the air pollutants do not follow these assumptions. From this matter raised one of the main contribution of this papers which is to proposed a hybrid logistic regression model (LOG_VAR_PCA) to quantify the association between the perceived annoyance and pollutant variables using the data set referred in the previous section.

Since the covariates (air pollutants) are time series, the use of time series models can help to understand the dynamic of the data and, additionally, to give a more precise statistical support in quantifying and discussing the association between particulate matter concentrations and perceived effects (Schwartz *et al*., 2000, Gouveia *et al.*, 2004).

Multivariate techniques are also required for the purpose of this paper as justified as follows. To analyse the perceived annoyance caused by particulate matter a joint analysis of sediment particulate matter (SPM), particulate matter (PM_{10}) and total suspended particles (TSP) is required. In this context, an analysis of the multivariate data set will be performed without simply isolating the effects of a single pollutant.

Since the covariates are time series and cross-correlated, the data requires a prior treatment using principal component analysis, see Zamprogno *et al.* (2019), Souza *et al.* (2018) Vanhatalo *et al.* (2016) and reference therein. Although the components obtained from PCA are not correlated, they can also present autocorrelation, which is transferred to the residuals of the fitted model. Thus, in this work, data are filtered through a multivariate time series model (the VAR model see, for example, Wei (2006)) before applying the PCA technique, as suggested by Souza et al. (2018) and Zamprogno *et al*. (2019). The models and techniques are summarized in the next subsections.

2.4.1 The Logistic Regression model

In many practical situations, the response variable in a regression model is categorical, for example, when the variable is binary, indicating the presence or absence of a

characteristic. Therefore, the logistic regression model becomes an important statistical tool to measure and quantify the relationship between perceived annoyance and a set of explanatory variables (particulate matter).

The logistic regression model and its parameter estimates are summarized. For more details see, for example, Abraham and Ledolter (2006).

Let $\mathbf{X} = (X_1, X_2, \dots, X_p)^t$ be a vector containing p explanatory variables. Suppose that the response variable Y is dichotomic (binary), that is, $\mathrm{Y} = 1$ or $\mathrm{Y} = 0$ for the outcome to be success or failure, respectively. Let the probability of Y to have success or failures, with respect to $\mathbf{X}$, be defined as $P(\mathrm{Y} = 1|\boldsymbol{X}) = \pi(\boldsymbol{X})$ and $P(\mathrm{Y} = 0|\boldsymbol{X}) = 1 - \pi(\mathbf{X})$, respectively.

For the explanatory vector $\mathbf{X}$, with the parameter vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^t$, and the response Y, the probability of success is parameterized as

$$P(Y = 1) = \pi(\boldsymbol{X}) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}. \quad (1)$$

Since this probability is a logistic function of the vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^t$, it can be shown that the logit of the multiple logistic regression model is given by

$$\ln\left(\frac{\pi(\boldsymbol{X})}{1 - \pi(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \quad (2)$$

The parameter $\beta_i, i = 0, \dots, p$, are unknown and have to be estimated based on sample data by the iteratively reweighted least squares approach. Let now $\boldsymbol{X_1}, \dots, \mathbf{X_n}$ be a sample of observations of the vector of covariates $\mathbf{X}$ and $Y_1, \dots, Y_n$ are the corresponding response variables. It can be shown that the vector parameter $\boldsymbol{\beta}$ can be estimated by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{P}'\widehat{\boldsymbol{W}}\boldsymbol{P}\right)^{-1}\boldsymbol{P}'\widehat{\boldsymbol{W}}\boldsymbol{Z}, \quad (3)$$

where the matrix $\mathbf{P}$ is the matrix of regressors which has one in the first column for the intercept parameter and $\widehat{\boldsymbol{W}}$ is a diagonal matrix of dimension $n\ x\ n$ with elements given by $\hat{\pi}_i(1 - \hat{\pi}_i), i = 1, \dots, n$, where $\hat{\pi}_i$ have to be estimated using the maximum likelihood method based on sample data, $\boldsymbol{Z}$ is a $n\ x\ 1$ matrix which elements are

$$Z_i = \ln\left\{\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right\} + \left\{\frac{Y_i - \hat{\pi}_i}{\hat{\pi}_i(1 - \hat{\pi}_i)}\right\}. \quad (4)$$

It can be demonstrated that

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left(\boldsymbol{P}'\widehat{\boldsymbol{W}}\boldsymbol{P}\right)^{-1} \quad (5)$$

Regarding to Equations (4) and (5) it is possible to identify a problem that may occur: the multicollinearity. The exact multicollinearity occurs when the matrix of covariates is not a full rank matrix, i.e., when the maximal number of linearly independent columns of P is less than the number of columns. Hence, the determinant of the matrix $\left(\boldsymbol{P}'\widehat{\boldsymbol{W}}\boldsymbol{P}\right)^{-1}$ is 0 and the matrix is not invertible.

This problem can be seen by writing $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{W}}^{\frac{1}{2}}\widehat{\boldsymbol{W}}^{\frac{1}{2}}$ and $\boldsymbol{L} = \widehat{\boldsymbol{W}}^{\frac{1}{2}}\boldsymbol{P}$ then

$$\boldsymbol{Var}\left(\widehat{\boldsymbol{\beta}}\right) \quad = (\boldsymbol{L}'\boldsymbol{L})^{-1}. \tag{6}$$

It can be shown that $rank(L) = rank(\boldsymbol{P})$, where *rank* (.) denotes the operator which counts the quantity of linear independent lines. Therefore, if $\boldsymbol{P}$ has not full rank or its columns are very close to being linearly dependent (highly correlated), this will have an effect on $(\boldsymbol{L}'\boldsymbol{L})^{-1}$ matrix, thus, affecting the estimated parameters (Lutkepohl, 1991).

2.4.2 Principal Component Analysis

As well known, 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.

In this study, the PCA technique is used to circumvent the problem of pollutants that are correlated with each other, i.e., the multicollinearity phenomenon. In general, the whole variability of a system determined by *p* variables can only be explained using all the *p* principal components. However, a large part of this variability can be explained using a lower number *r* of components (r <p) see for example, Johnson and Wichern (2007).

As mentioned before, the use of PCA requires attention regarding the covariates that are correlated in time (serial correlation) as it is the case of air pollutants. The time correlation of the vector **X** will lead to PCs auto-correlated and cross-correlated in time. As pointed by Souza *et al.* (2018) and Zamprogno *et al.* (2019), the effect of time correlation in atmospheric pollutants strongly influences the estimates of the principal components, increasing the total variability of the data and increasing the retained variability of the first component. This can be mitigate using a multivariate time series to filter the data, as suggested in Souza *et al.* (2018) and Zamprogno *et al.* (2019).

In Equation (2), the vector X will be the PCA variables generated from the sample covariance matrix of the filtered pollutants using a multivariate autoregressive time series model of order 1 (VAR(1)) (see, for example, Wei (2006)).

This is addressed in the Result and discussion Section. More details of the use of PCA in regression models can be recently found in Souza *et al.* (2018) Zamprogno *et al.* (2019), Hu and Tsay (2014) and Roberts and Martin (2006).

2.4.3 Relative Risk

The relative risk (RR) is frequently used in epidemiological studies to measure the impact of atmospheric pollutant concentrations on the health of the exposed population. The RR can be defined as the association that an effect (annoyance) can occur following a certain exposure to a risk factor, which corresponds to the exposure to particulate matter concentration levels in this study. The relative risk is used in data analysis with binary outcomes (0 or 1) as in the case of annoyance. According to Bishop (2007) the relative

290 risk is the result of dividing the probability of the event (being annoyed when exposed –
291 A|B) by the probability of the event (being annoyed when not exposed – A|B^C), i.e.:

$$RR(A,B) = \frac{P(A|B)}{P(A|B^c)} \tag{7}$$

292

293 According to Baxter (1997), by analogy, the relative risk function at level x of the desired
294 pollutant, denoted RR (x), is defined as:

$$RR(x) = \frac{E(Y|X = x)}{E(Y|X = 0)} \tag{8}$$

295

296 It is the ratio of the expected value of the response variable at level x of the independent
297 variable to the expected of the response if the independent variable was 0.

298 In this context, for the logistic regression, it can be shown that the RR can be estimated
299 by

$$\widehat{RR}(x_i) \approx e^{x_i \hat{\beta}_i} \tag{9}$$

300

301 where x_i is the interquartile variation (3st quantile - 1st quantile from Table 1) in the ith
302 pollutant concentration and $\hat{\beta}_i$ is represented by:

$$\hat{\beta}_i = \sum_{j=1}^{r} \hat{\alpha}_{ij} \hat{\gamma}_j \quad i = 1,2,\dots,p, \tag{10}$$

303 where $\hat{\alpha}_j = (\hat{\alpha}_{ji})$ is the *j*-th estimated eigenvector of the covariates matrix (from Table
304 3); $\hat{\gamma}_j$ is the estimated coefficient of the *j*-th PC calculated in the logistic regression (from
305 Table 4). Through the coefficient $\hat{\beta}_i$ it is computed the individual contribution of each
306 pollutant to the perceived annoyance see, for example, Souza *et al.* (2018).

307

308 **1- Results and discussion**

309 Table 1 presents the descriptive statistics (minimum, maximum, average and standard
310 deviation) of the pollutants monthly measured in the Vitoria region from 2011 to 2014.
311 Note that, the maximum particulate matter concentrations observed for PM_{10} and TSP
312 pollutants can be very dangerous for the health system since its values are above the limits
313 set by the World Health Organization (WHO, 2006). The maximum value for SPM is also
314 higher than the annoyance standard values considered in many countries see, for example,
315 (Vallack and Shilitto, 1998; Melo *et al*., 2018).

316

317 In the standard regression model, the basic assumption is that the covariates are not
318 correlated and not time-dependent. However, in the case studied here, the predicable
319 variables do not satisfy these properties, since the pollutant variables are serially and time

dependent. As shown in Table 2, the pollutants are contemporaneously correlated, for example, the sample correlation between SPM x PM_{10} is $\hat{\rho}_{SPM,PM_{10}} = 0.424$. The pollutants are time series, and their behaviours over time are displayed in Figures 2 to 6. These figures show the monthly data time series of each air pollutant (particles deposition rate, monthly averages of PM_{10} and TSP, monthly maximum averages of PM_{10} and TSP) from July 2011 to October 2014. These also display the sample autocorrelation (ACF) and partial autocorrelation (PACF) functions which clearly show that the pollutants are time-dependent. In the ACF and Partial ACF plots (Figures 2-11), the vertical axis measures the strength of the correlation and the horizontal axis is the time lag at which the correlation was calculated. The dashed lines represent the 95% confidence intervals for uncorrelated data.

The sample ACF measures the dependence between the observations of the same time series at different delays, usually detonated as lags in time series methods. Figures 7 to 11 show that the VAR (1) removed the time correlations. From these, it appears that the series have a very weak yearly seasonality. However, it should be noted that the seasonal yearly effect (if any) may be reduced by the smoothing of the monthly mean average of the pollutants PM_{10} and TSP.

Since the covariates do not meet the regression basic assumption, one way to mitigate the problem is to remove the time correlation (serial-correlation) of the series. In this context, it is suggested here to use a linear time series filter as a procedure to transform the data into a "white noise" process. This problem and how to mitigate it are well-addressed in the recent publications Souza *et al.* (2018), Vahatalo and Kulahci (2016), and Zamprogno *et al*. (2019).

Based on the sample ACF plots, the residual analysis and the Akaike information criterion (AIC), which is an estimator of the relative quality of statistical models for a given set of data, a Vector Autoregressive Model of order 1, denoted by VAR (1), was chosen to model the vector of all pollutants time series (particles deposition rate, monthly averages of PM_{10} and TSP, monthly maximum averages of PM_{10} and TSP). The sample ACF plots of the filtered data are displayed in Figures 7 to 11. From these plots, it can be seem that the time-correlation of the series was removed, and the filtered data displays a similar behaviour of a white noise process, that is, the correlations of the residuals are nulls. In addition, the residuals do not show any anomaly (results are available upon request). Therefore, this indicates that the VAR (1) model well-fitted the data. For a more details of multivariate linear time series models see, for example, Wei (2006).

Table 3 displays the results of the PCA technique applied to the filtered series. The total cumulative variance was used as a criterion for choosing the number of components resulted by the PCA. Thus, the first three components were chosen, which explain 86% of the total variability. In the PC1, the higher contributions come from TSP, PM_{10} TSP. In the case of PC2, SP gives most of the variability and, for the PC3, PM_{10} gives the highest contribution. The pollutants indicated by (*) are the ones that give more contributions to the variability of the PC. For more details on PCA and its application see, for example, Cadima and Jolliffe, (1995).

In the multiple logistic regression model, the response variable (perceived annoyance) was associated with the covariates PC1, PC2 and PC3 resulting in the hybrid LOG-PCA-VAR fitted model and its parameter estimates are in Table 4.

The relative risk (RR) of annoyance results were expressed by the interquartile variation range. The RR analysis was performed for different levels of pollutants concentrations to test the null hypotheses H_0: RR =1 against H_1: RR > 1, using significance level of 5%. Tor each pollutant, Table 5 displays the results of the estimates of RR and the respectively confidence interval (CI), for the standard and the proposed methodology, that is, $\widehat{RR*}$ refers to the estimated RR using the standard logistic regression, and $\widehat{RR}$ corresponds to RR estimate based on the LOG-PCA-VAR model. Note that, the $\widehat{RR*}$ was considered in the study for comparison purpose, that is, to quantify (if any) the impact on the RR when the multivariate time series properties (multicollinearity and time and cross-correlation structures) of the covariates are ignored.

According to Table 5, the estimate of the RR for SPM increases approximately by a factor of 1.5 considering the interquartile variation equal to 2g/m² 30 days whereas, for PM_{10} (monthly mean), $\widehat{RR}$ increases by a factor of 1.6 considering the interquartile variation equal to 5µg/m³. In the case of TSP (monthly mean), $\widehat{RR}$ can be interpreted as a factor that increases 2.2 when exposed to the interquartile variation equal to 13 µg/m³. For PM_{10} (monthly maximum) variable, $\widehat{RR}$ growths by a factor of 2.4 considering the interquartile variation equal to 8 µg/m³ whereas, for the variable TSP (monthly maximum), $\widehat{RR}$ is equal to 1.8 considering the interquartile variation equal to 20 µg/m³. The estimated confidence intervals were calculated based on the central limit theorem as showed by Souza *et al.* (2018). The $\widehat{RR}$ values indicate that, all pollutants contributes significantly for the increase of the probability of being annoyed with 95% of confidence. It is interesting to note that the values of $\widehat{RR*}$ was not significant in any case. This is not a surprising result since the temporal correlation in data was not considered in the regression model which lead to underestimating the regression parameter and inflating the intercept. Consequently, this gives a spurious result in the sense that the pollutants don't make any impact on the perceived annoyance.

The proposed hybrid LOG-PCA-VAR model, in addition to the estimation of the impact of particulate matter on the perceived annoyance, which indicated significantly contribution of the pollutant to this response variable, it contributed to show the spurious result when the temporal correlation structure in the data is not considered to obtain the estimates of a logistic regression model. This corroborates the use of the proposed methodology when dealing with regression models in which the covariates are multivariate time series and all results are in accordance with Souza *et al*. (2018).

2- Conclusion

This study proposes the application of multivariate statistical techniques (time series models, principal component analysis and logistic regression) to estimate the effect between exposure to particulate matter concentrations (SPM, PM_{10} and TSP) and response of the population measured by the perceived annoyance levels.

The descriptive and graphical analysis motivated the use of the PCA technique for the air pollutant data by the initial indication of cross-correlation between the covariates (pollutants). The VAR(1) model was used to transform the original time series of air pollutants, resulting in time uncorrelated data (white noise) before applying the PCA technique. Based on these modelling steps, the PCA variables becomes uncorrelated and not cross-correlated.

The logistic regression model was applied with the level of annoyance as the dependent variable and the air pollutants as covariates. Moreover, by the new methodology developed in this study (*LOG-PCA-VAR*), the combined effect of particulate matter was analysed and the relative risk of annoyance for each original air pollutants was calculated. The estimates of relative risk, i.e, $\widehat{RR}$, showed that, in general, an increase in air pollutant concentrations (i.e., the particulate matter metrics examined here: TSP, PM_{10} and SPM) significantly contributes in increasing the probability of being annoyed.

In summary, the results obtained in this study provide evidence of a significant correlation between particulate matter and perceived annoyance levels, also indicating that, at least for particulate matter, perceived annoyance is not only related to one pollutant but to a group of pollutant. In future work, this methodology should be used to analysis with other pollutants. Other methodologies, such as bootstrap techniques, could also be used to estimate the confidence intervals more precisely, and GLARMA modelling could be used to solve the data autocorrelation problem.

3- Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the manuscript. The results in this paper were part of the PhD thesis of the first author under supervision of Valderio A. Reisen and Jane M. Santos, at PPGEA-UFES, Brazil, 2015 (Machado 2015). The authors would like to thank CNPq, CAPES and FAPES for their financial support. Part of this paper was revised when Valdério Reisen, Márton Ispány and Milena Machado were visiting CentraleSupélec in July 2018, January and July 2019. These authors are indebted to CentraleSupélec and Université Paris-Sud for their financial supports. This research was also partially supported by the iCODE Institute, research project of the IDEX Paris-Saclay, and by the Hadamard Mathematics LabEx (LMH) through the grant number ANR-11-LABX-0056-LMH in the Programme des Investissements d'Avenir. The work of Márton Ispány is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-financed by the European Union and the European Social Fund.

6- References

1) Abraham B. & Ledolter J. Introduction to Regression Modeling. Thomson Brooks/Cole, 2006.
2) Amundsen A.H., Klaeboe R. & Fyhri A. (2008). Annoyance from vehicular air pollution: Exposure–response relationships for Norway. Atmospheric Environment, 42, 679-688.

3) Bishop, Y., Fienberg, S., Holland, P. (2007). Discrete Multivariate Analysis: Theory and Practice. Cambrige: MIT, 575 p.
4) Baxter L., Finch S., Lipfert F., and Yu Q. (1997). Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. Risk Analysis, 17, 273–278.
5) Blanes-Vidal, V., Suh H., Nadimi E. S., Løfstrøm P., Ellermann T., Andersen H. V., Schwartz J., (2011) Residential exposure to outdoor air pollution from livestock operations and perceived annoyance among citizens. Environment International, 40, 44-50.
6) Cadima J., Joliffe I.T. (1995). Loadings and correlations in the interpretation of principal components. Journal of Applied Statistics, 22(2), 203-214.
7) Colls, J. Air Pollution. 2ed. USA: SPON Press Taylor & Francis Group, 2002.
8) Cox, L. (2000). Statistical issues in the study of air pollution involving airborne particulate matter. Environmetrics 11, 611-626.
9) Dockery, D.W. and Pope, C.A. (1994). Acute respiratory effects of particulate air pollution. Annual Review of Public Health, 15, 107–132.
10) Egondi, T., Kyobutungi, C., Ng, N., Muindi, K., Oti, S., van de Vijver, S., Ettarh, R., Rocklöv, J., 2013. Community perceptions of air pollution and related health risks in Nairobi slums. Int. J. Environ. Res. Publ. Health 10, 4851–4868.
11) Farfel, M.R., Orlova, A.O., Lees, P.S.J., Rohde, C., Ashley, P.J., Julian Chisolm, J., 2005. A study of urban housing demolition as a source of lead in ambient dust on sidewalks, streets, and alleys. Environ. Res. 99, 204–213. doi:10.1016/j.envres.2004.10.005
12) Gouveia, N., Bremner, S.A., Novaes, H.M. (2004). Association between ambient air pollution and birth weight in São Paulo, Brazil. Journal of Epidemiology and Community Health, 58, 11-17.
13) Holgate, S.T., Samet, J.M., Koren, H.S., Maynard, R.L., 1999. Air pollution and health. Academic Press.
14) Jacquemin B., Sunyer J., Forsberg B., Gotschi T., Oglesby L., Ackermann-Liebrich U., De Marco R., Heinrich J., Jarvis D., Toren K., Kunzli N. 2007. Annoyance due to air pollution in Europe. International Journal of Epidemiology, 36, 809–820.
15) Johnson, R.A., Wichern, D.W. (2007). Applied Multivariate Statistical Analysis. 6th edition. Prentice Hall, New Jersey, 800 p.
16) Klaeboe, R., Kolbenstvedt, M., Clench-Aas, J., Bartonova, A. (2000). Oslo trafic study - part 1: an integrated approach to assess the combined effects of noise and air pollution on annoyance. Atmospheric Environment, 34, 4727-4736.
17) Klæboe R., Öhrström E., Turunen-Rise I., Bendsten H., Nykänen H. (2003). Vibration in dwellings from road and rail traffic – Part III: towards a common methodology for socio-vibrational surveys. Applied Acoustics, 64, 111–120.
18) Klaeboe, R., Amundsen A.H., Fyhri A. (2008). Annoyance from vehicular air pollution: A comparison of European exposure–response relationships. Atmospheric Environment, 42, 7689-7694.
19) Hu Y-P., Tsay R.S. (2014) Principal volatility component analysis. Journal of Business and Economic Statistics, 32(2), 153-164.
20) Llop S., Ballester F., Estarlich M., Esplugues A., Fernández-Patier R., Ramón R., Marco A., Aguirre A., Sunyer J., Iñiguez C., on behalf of INMA-Valencia cohort

(2008). Ambient air pollution and annoyance responses from pregnant women. Atmospheric Environment, 42, 2982-2992.
21) Lutkepohl, H. (1991). Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin.
22) Machado M., Santos J.M., Reisen V. A., Reis N.C., Mavroidis I. Lima A. T. A new methodology to derive settleable particulate matter guidelines to assist policy-makers on reducing public nuisance. Atmospheric Environment 182 (2018) 242–251.
23) Martin, S.W., Meek, A.H., Willeberg, P., 1987. Veterinary epidemiology. Principles and methods. Iowa StateUniversity Press, Ames, IA, p. 343.
24) Melo, M.M., Santos, J.M., Frere, S., Reisen, V.A., Jr., N.C.R., Leite, M.F.S. de F.S., 2015. Annoyance Caused by Air Pollution: A Comparative Study of Two Industrialized Regions. World Acad. Sci. Eng. Technol. Int. J. Environ. Ecol. Eng. 2, 182–187.
25) Nikolopoulou M., Kleissl J., Linden P.F., Lykoudis S. (2011). Pedestrians' perception of environmental stimuli through field surveys: Focus on particulate pollution. Science of the Total Environment, 409(13), 2493-202.
26) Nordin S., Lidén E. (2006). Environmental odor annoyance from air pollution from steel industry and bio-fuel processing. Journal of Environmental Psychology, 26, 141–145.
27) Oglesby, L., Kunzli, N., Monn, C., Schindler, C., Ackermann-Liebrich, U., Leuenberger, P. (2000). Validity of annoyance scores for estimation of long term air pollution exposure in epidemiologic studies: The Swiss study on air pollution and lung diseases in adults (SAPALDIA). American Journal of Epidemiology, 152, 75–83.
28) Roberts S, Martin M. Using supervised principal components analysis to assess multiple pollutant effects. Environmental Health Perspectives, Vol. 116, No. 12. 2006.
29) Rotko T., Oglesby L., Kunzli N., Carrer P., Nieuwenhuijsen M.J., Jantunen M. (2002). Determinants of perceived air pollution annoyance and association between annoyance scores and air pollution (PM2.5, NO2) concentrations in the European EXPOLIS study. Atmospheric Environment, 36, 4593–4602.
30) Santos, J.M., Reis Jr, N.C., Galvão, E.S., Silveira, A., Goulart, E.V., Lima, A.T., 2017. Source apportionment of settleable particles in a mining-impacted urban and industrialized region in Brazil. Environ. Sci. Pollut. Res. doi:10.1007/s11356-017-9677-y.
31) Schwartz, J. (2000). Harvesting and long-term exposure effects in the relationship between air pollution and mortality. American Journal of Epidemiology, 151(5), 440- 448.
32) Stenlund, T., Lidén, E., Anderson, K., Garvill, J., Nordin, S. (2009). Annoyance and health symptoms and their influencing factors: A population-based air pollution intervention study. Public Health. Vol. 123, p. 339-345.
33) Souza, J. B., Reisen, V. A., Santos, J.M., Franco, G. C. (2014). Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. Rev Saúde Pública 48(3):451-458.

34) Souza, J. B., Reisen, V. A., Franco, G. C., Ispány, M., Bondon, P., Santos, J. M. (2018). Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. Journal of the Royal Statistical Society: Series C (Applied Statistics) (67), 453-480, 2018.
35) Vallack, H.., Shillito, D., 1998. Suggested guidelines for deposited ambient dust. Atmos. Environ. 32, 2737–2744. doi:10.1016/S1352-2310(98)00037-5
36) Vanhatalo E., Kulahci M., Impact of autocorrelation on principal components and their use in statistical process control, Quality and Reliability Engineering International 32 (2016) 1483–1500.
37) Wei, W.W.S. (2006). Time Series Analysis: Univariate and Multivariate Methods. Pearson Addison Wesley.
38) Whittle, N., Peris, E., Condie, J., Woodcock, J., Brown, P., Moorhouse, A.T., Waddington, D.C., Steele, A., (2015). Development of a social survey for the study of vibration annoyance in residential environments: good practice guidance. Appl. Acoust. 87, 83–93.
39) WHO, 1946. Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.
40) WHO, 2005. WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulphur Dioxide. Summary of Risk Assessment. Geneva, 2006.
41) Zamprogno, B., Reisen, V. A., Reis Junior, Neyval Costa, C. H. H. A., and Bondon,P. (2019). Principal component analysis with autocorrelated data (pre-print available with the authors).
42) ZOU, G. A modified Poisson regression approach to prospective studies with binary data. American Journal of Epidemiology 159(7), 702–706. 2004.
43)

Figure

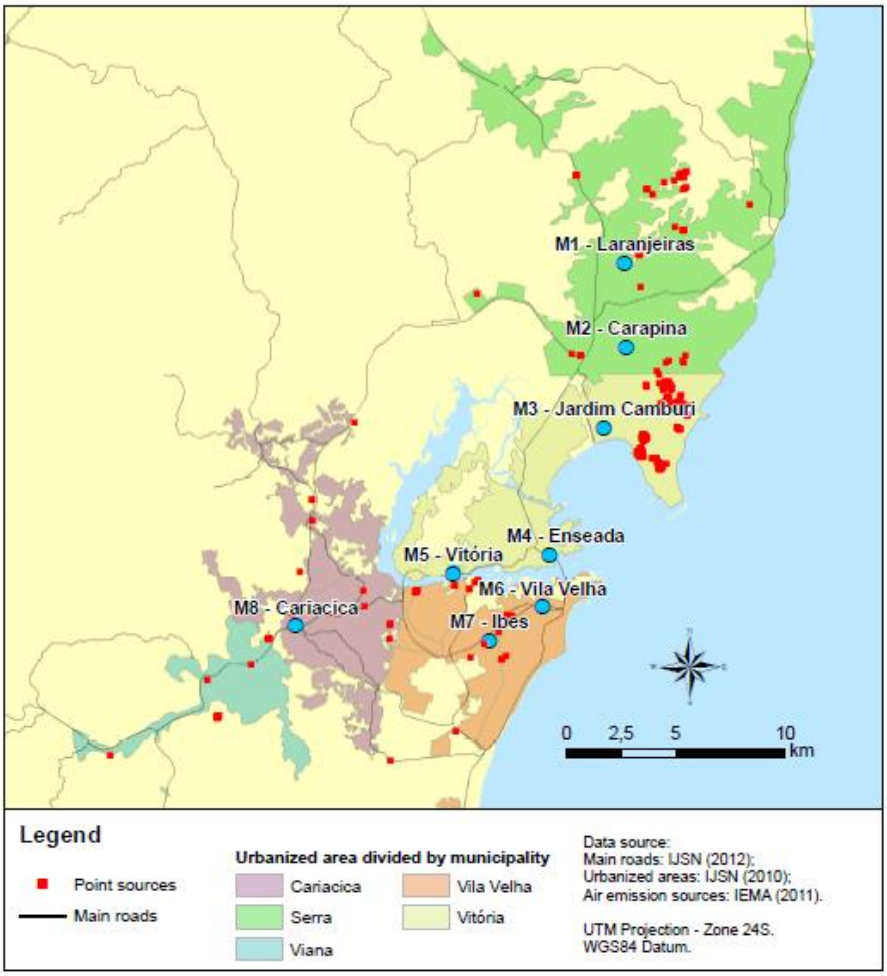
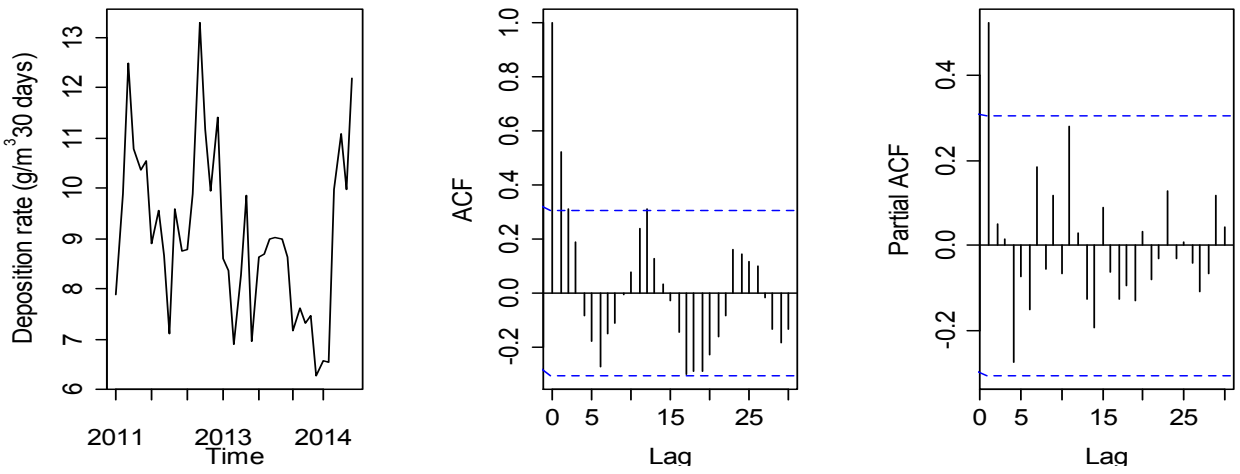
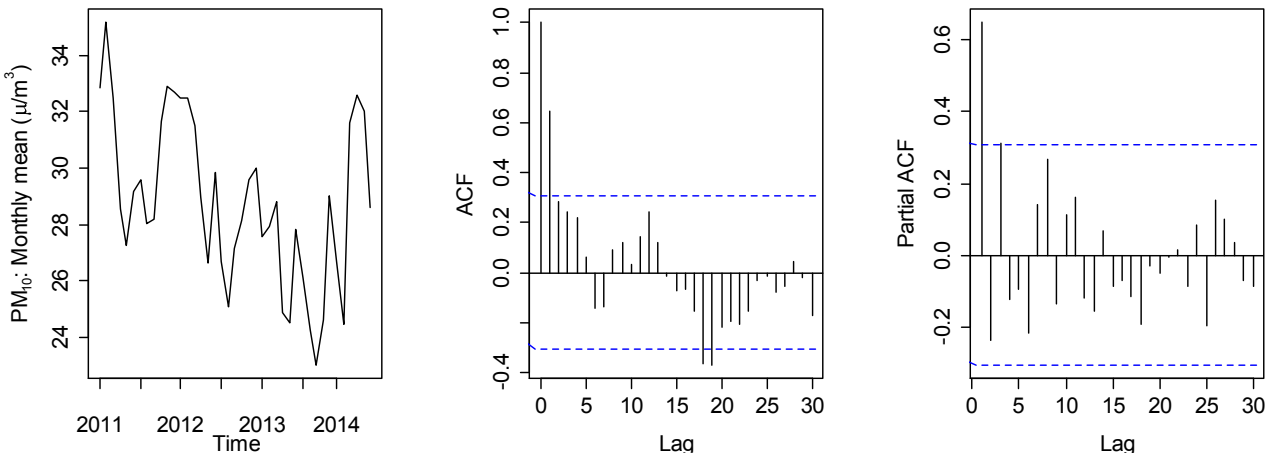


Figure 1- Metropolitan Region of Vitoria, the main sources, the main roads and the air quality monitoring stations network: (M1) Laranjeiras, (M2) Carapina, (M3) Jardim Camburi, (M4) Enseada, (M5) Vitoria, (M6) Vila Velha, (M7) Ibes, (M8) Cariacica.



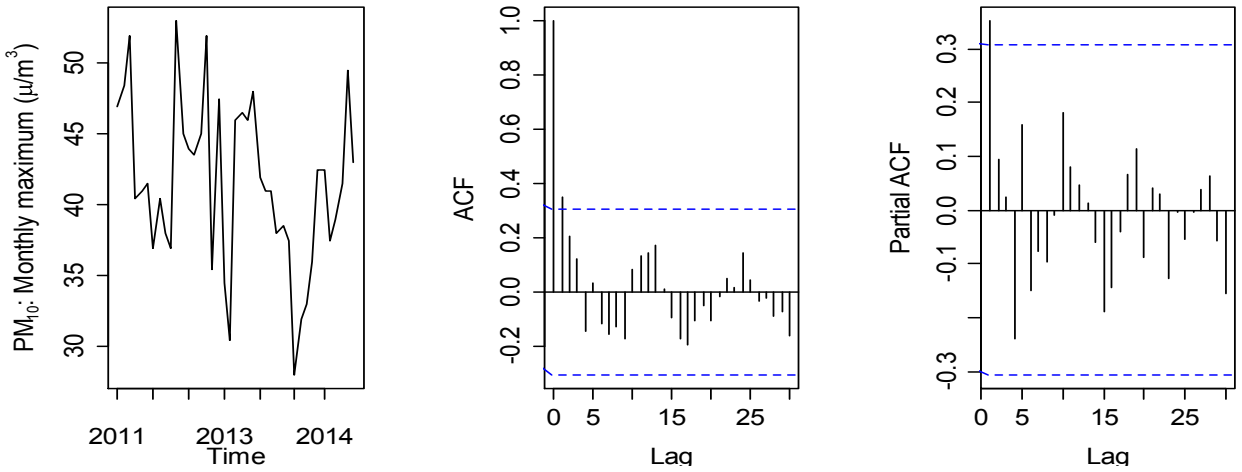
(a) (b) (c)

Figure 2 – Time series (a), autocorrelation function (b) and partial autocorrelation function (c) for SPM from 2011 to 2014.



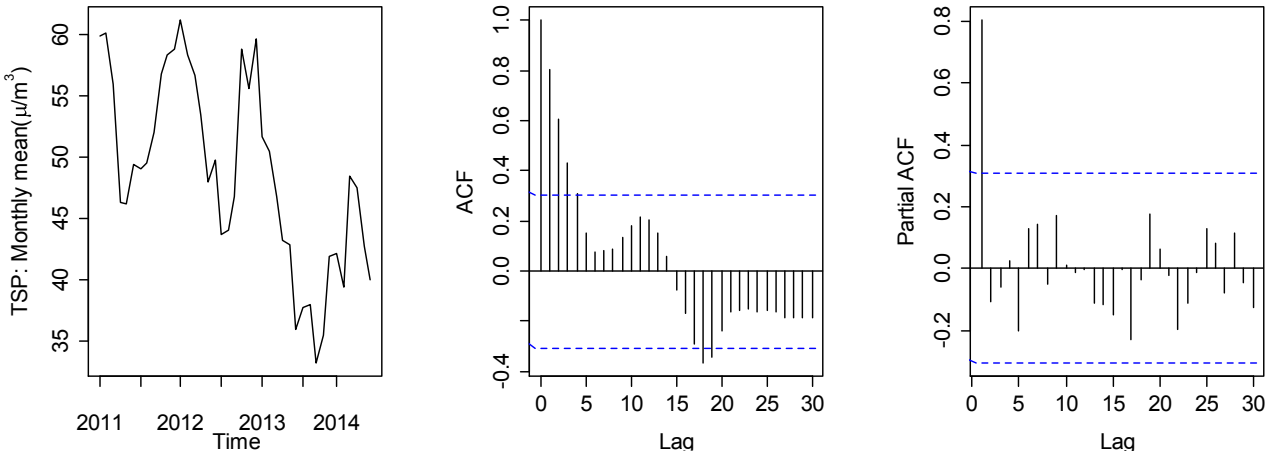
(a) (b) (c)

Figure 3- Time series (a), autocorrelation function (b) and partial autocorrelation function (c) for monthly mean concentration of PM_{10} from 2011 to 2014.



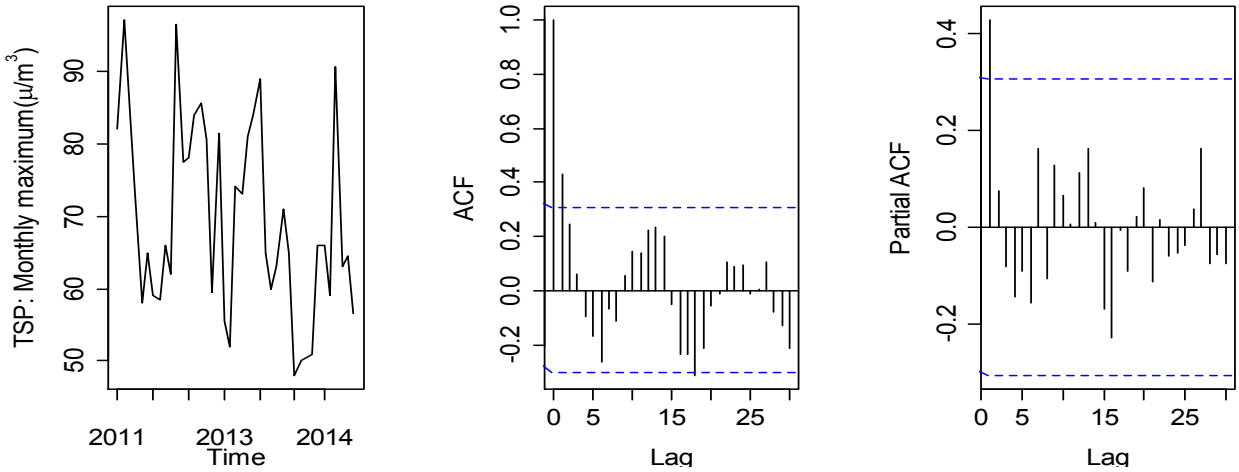
(a) (b) (c)

Figure 4- Time series, autocorrelation function and partial autocorrelation function for monthly maximum PM_{10} concentration from 2011 to 2014.



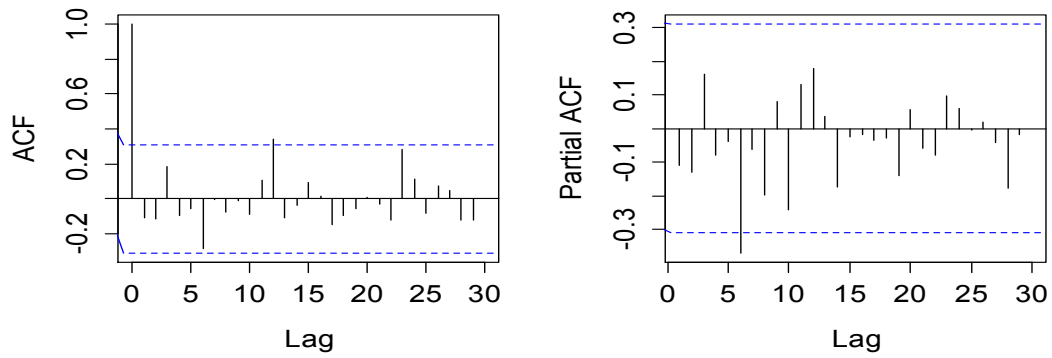
(a) (b) (c)

Figure 5- Time series, autocorrelation function and partial autocorrelation function for monthly mean TSP concentration from 2011 to 2014.



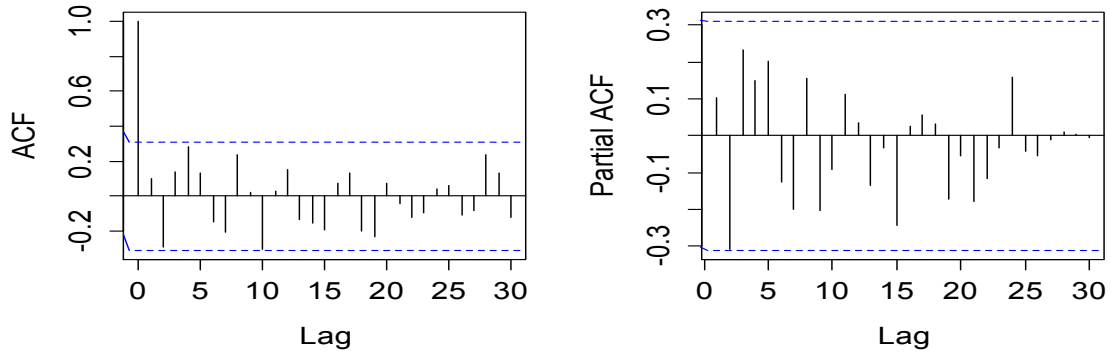
(a) (b) (c)

Figure 6- Time series, autocorrelation function and partial autocorrelation function for monthly maximum TSP concentration from 2011 to 2014.



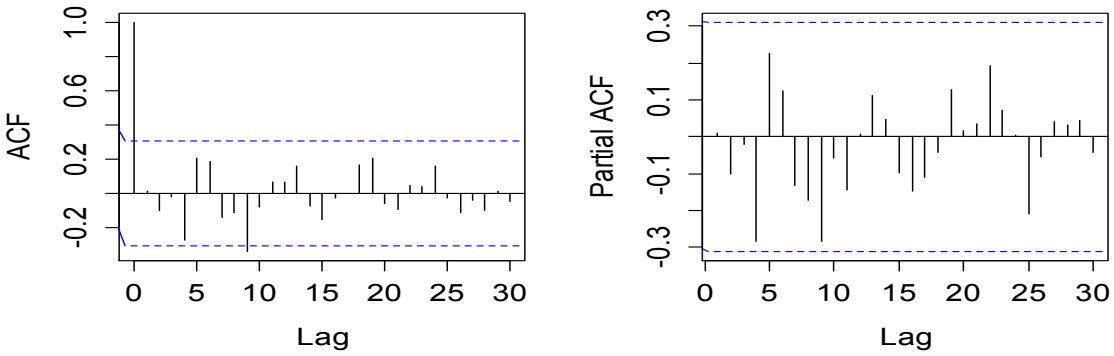
(a) (b)

Figure 7 - Autocorrelation function (a) and partial autocorrelation function (b) for particles deposition rate from 2011 to 2014 after filtering.



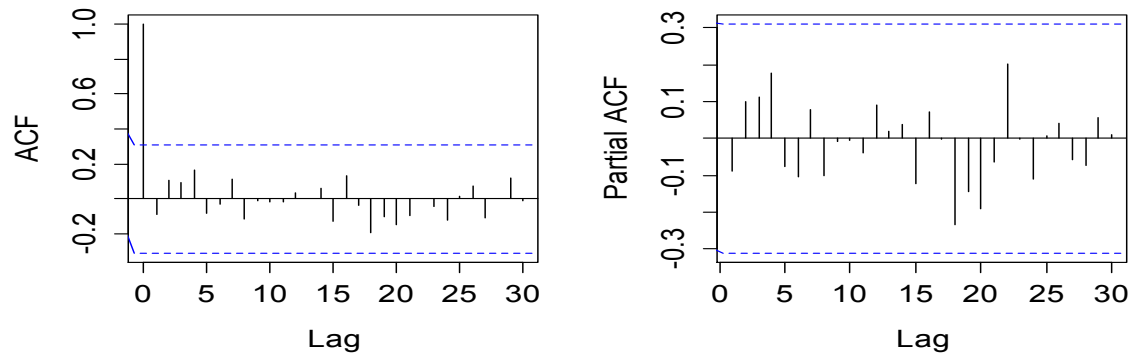
(a) (b)

Figure 8- Autocorrelation function (a) and partial autocorrelation function (b) for monthly mean concentration of PM_{10} from 2011 to 2014 after filtering.



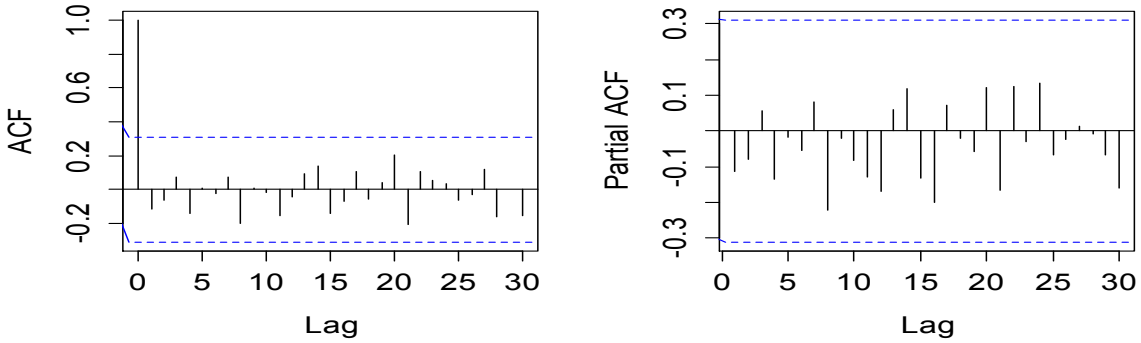
(a) (b)

Figure 9- Autocorrelation function (a) and partial autocorrelation function (b) for monthly maximum PM_{10} concentration from 2011 to 2014 after filtering.



(a) (b)

Figure 10- Autocorrelation function (a) and partial autocorrelation function (b) for monthly mean TSP concentration from 2011 to 2014 after filtering.



(a) (b)

Figure 11- Autocorrelation function (a) and partial autocorrelation function (b) for monthly maximum TSP concentration from 2011 to 2014 after filtering.

Table 1 – Descriptive statistics of air pollutants (from July 2011 to November 2014)

Variable	Minimum	Maximum	Mean	Std. Dev.	1st quantile	3st quantile	90th percentile
SPM (g/m² 30 days)	6.267	13.283	9.097	1.680	7.683	9.969	11.173
PM_{10} (µg/m³)	23.002	35.167	28.818	2.962	26.670	31.575	32.590
TSP (µg/m³)	33.166	61.167	48.665	7.808	42.705	55.899	58.830

Table 2 – Correlation matrix for the original variables (before time series analysis)

Variables	SPM	PM_{10} (mean)	TSP (mean)	PM_{10} (maxim)	TSP (maxim)
SPM	1.				
PM_{10} (mean)	0.424**	1			
TSP (mean)	0.278	0.764**	1		
PM_{10} (maxim)	0.409**	0.681**	0.654**	1	
TSP (maxim)	0.342*	0.701**	0.754**	0.772**	1

**p-value=0,01
*p-value=0,05

Table 3- Results of factor loadings statistics and application of PCA

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Eigenvalue	2.576	1.071	0.681	0.396	0.276
Variability (%)	51.528	21.426	13.622	7.913	5.510
Cumulative %	**51.528**	**72.955**	**86.577**	94.490	100.000
SP (monthly rate)	0.267	0.733*	-0.554	-0.269	-0.112
PM_{10} (monthly mean)	0.495*	-0.257	-0.365	0.674	-0.319
TSP (monthly mean)	0.400*	-0.583	-0.318	-0.607	0.172
PM_{10} (monthly maxim)	0.492*	0.104	0.611*	-0.254	-0.557
TSP (monthly maxim)	0.531*	0.214	0.293	0.200	0.739

*High contributions

Table 4- Parameters estimated by the multiple logistic model estimated for the first three components

	$\hat{\beta}$	Standard error	Exp($\hat{\beta}$)
PC1	0.053	0.202	1.054
PC2	0.058	0.309	1.060
PC3	-0.245	0.390	0.783
Intercept	0.204	0.320	-

Table 5- The estimate RR of annoyance for each pollutant and the respective interval confidence

Pollutants	$\widehat{RR}*$ (standard methodology)	**CI (95%)**	$\widehat{RR}$ (LOG-PCA-VAR)	**CI (95%)**
SPM	0.865	(0.582;1.283)	1.462	(1.070; 1.854)
PM_{10} (monthly mean)	0.819	(0.650; 1.031)	1.649	(1.061; 2.237)
TSP (monthly mean)	0.953	(0.875; 1.037)	2.181	(1.471; 2.891)
PM_{10} (monthly maxim)	0.977	(0.877; 1.088)	2.411	(1.401; 3.421)
TSP (monthly maxim)	0.965	(0.918; 1.014)	1.822	(1.52; 3.052)