Identification of redundant air quality monitoring stations using robust principal component analysis

Higor Cotta, Valdério Reisen, Pascal Bondon, Paulo Prezotti

Environmental Modeling & Assessment manuscript No.
(will be inserted by the editor)

Identification of redundant air quality monitoring stations using robust principal component analysis

Higor Henrique Aranda Cotta · Valdério Anselmo Reisen · Pascal Bondon · Paulo Roberto Prezotti Filho

Received: date / Accepted: date

Abstract Air quality monitoring stations are essentials for monitoring air pollutants and, therefore, are essential to protect the public health and the environment from the adverse effects of air pollution. Two or more stations may monitor the same pollutant behavior. In this scenario, the equipment must be reallocated to provide a better use of public resources and to enlarge the monitored area. The identification of redundant stations can be carried out by the application of principal component analysis (PCA) as a grouping technique. The principal component analysis

Higor Henrique Aranda Cotta · Valdério Anselmo Reisen · Paulo Roberto Prezotti Filho
Graduate program in Environmental Engineering, Departament of Statistics, Federal University of Espírito Santo, Brazil E-mail: higor.cotta@centralesupelec.fr

Higor Henrique Aranda Cotta · Valdério Anselmo Reisen · Pascal Bondon · Paulo Roberto Prezotti Filho
Laboratoire des Signaux et Systèmes (L2S),CNRS-CentraleSupélec-Université Paris-Sud, Université Paris-Saclay, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

Paulo Roberto Prezotti Filho
Federal Institute of Espírito Santo, Brazil

is a set of linear combinations of the original variables constructed to explain the variance-covariance structure of the data. It is well known that outliers affect the covariance structure of the variables. Since the components are computed by using the covariance or the correlation matrix, the outliers also affect the properties of the components. This article proposes a grouping methodology that applies robust PCA to identify air quality monitoring stations that present similar behavior for any pollutant or meteorological measure. To illustrate the usefulness of the proposed methodology, the robust PCA is applied to the management of the automatic air quality monitoring network of the Greater Vitória Region in Brazil consists of 8 stations. It was found that four components could explain 84% of the total variability, and it is possible to create a group composed of at least two stations in each one of the components. Therefore, the redundant stations can be installed in a new site to expand the monitored area.

Keywords Air quality · Monitoring networks · time series · Robust principal component analysis · Outliers

1 Introduction

The concern about air pollution problems has increased considerably in the last 50 years. Especially in developing countries, the air quality has been degraded as a result of industrialization, population growth, high rates of urbanization, and inadequate or nonexistent policies to control air pollution. The problems caused by air pollution produce local, regional, and global impacts. In this context, the particulate matter (PM), especially the PM_{10}, which has an aerodynamic diameter less than 10 μm, is one of the most important pollutants with natural and anthropogenic sources.

Its adverse impacts on humans health may lead to an increment of mortality rates, respiratory and cardiovascular problems for short and longterm exposure at high concentrations (Beelen et al. 2014; Cesaroni et al. 2014; Hoek et al. 2013; Rückerl et al. 2011).

The primary purpose of air quality management is to protect public health and the environment from the adverse effects of air pollution. Adequate control of air quality involves several activities such as risk management, setting standards for emissions and air quality, implementation of control measures, and risk communication (WHO 2005). The monitoring of air quality is essential for any air pollution control policy. The realization of efficient management of air quality is important for identifying and quantifying the pollutants found in a region and their sources. This is accomplished by using stations to monitor different pollutants according to the needs of the regions where the stations are installed.

In Brazil, although the limits for pollutant concentrations are established by the federal legislation CONAMA 003/90 (Conselho Nacional do Meio Ambiente 1990), this decree does not contemplate guidelines on how to construct or how to manage monitoring networks and, thus, entrusting this task to each one of the 27 federative units. In this scenario, an actual overview of Brazil's air quality monitoring networks is given in a recent publication of the Instituto de Energia e Meio Ambiente coauthored by the Brazilian Ministry of the Environment. This publication highlights although essential the air quality monitoring in Brazil is far from being a reality. Due to the dimensions of the country, the non-prioritization of air quality policies and the amount of financial resources destined to the monitoring activities, only 12 out of 27 unity members have an operational air quality monitoring network (IEMA 2014).

The installation and continuous operation of an air quality monitoring station are cost-intensive as it requires finding a suitable place for the installation and personnel for its maintenance. Only one monitoring station should operate in an area characterized by a specific pattern of air pollution. Pires et al. (2008a) indicate the number of stations that constitute a monitoring network must be optimized to reduce costs and expenses. If there are stations with similar patterns of pollution for a specific pollutant, the monitoring equipment could be properly relocated to another area of interest.

In this context, the principal component analysis has been successfully used in air pollution for managing a network of monitoring stations in several studies, for instance, Zamprogno (2013) studied PCA with time series models in many different applications related to air pollution data, Zhao et al. (2015) applied PCA to verify redundant air quality monitoring networks in Shanghai (China). Dominick et al. (2012) used PCA and Cluster Analysis (CA) to check the pattern of behavior of the pollutants carbon monoxide (CO), ozone (O_3), particulate matter of diameter $< 10\mu m$ (PM_{10}), sulfur dioxide (SO_2), nitric oxide (NO) and nitrogen dioxide (NO_2) in five different stations in Malaysia. Pires et al. (2009, 2008a,b) applied PCA to identify monitoring sites with similar concentrations of pollutants for PM_{10}, SO_2, CO, NO_2 and O_3 in the metropolitan area of Porto (Portugal). Lu et al. (2011) employed PCA to study the air quality monitoring network of Hong Kong for the pollutants of SO_2, NO_2 and Respirable Suspended Particulate (RSP). The authors found that the monitoring stations located in nearby areas are characterized by the same specific air pollution characteristics and suggested that redundant equipment should be transferred to other monitoring stations allowing for further enlargement

of the monitored area. Other studies include Lau et al. (2009) and Gramsch et al. (2006).

The application of PCA is not exclusive to the management of air quality monitoring networks. Recently, Villas-Boas et al. (2017) used PCA and nonlinear PCA to assess the redundancy of the parameters and monitoring locations of the Piabanha water quality network in Brazil. Phung et al. (2015) applied PCA and other multivariate statistical tools to assess the river surface water quality and also redundant monitoring stations in Can Tho City (Vietnam).

At this point, PCA is one of the main multivariate statistical techniques. The goal of PCA is to explain the covariance structure of the data through auxiliary variables called components. These components are constructed from linear combinations of the original variables and are uncorrelated. Briefly, PCA calculates the eigenvalues and eigenvectors of the covariance or correlation matrix. The main application of PCA is to reduce the dimensionality of a correlated data matrix of n dimension to a m dimension, where $m < n$. The reduction is performed so that the new set of variables captures most of the variability contained in the original data. A review of the fundamentals of PCA using R (R Core Team 2018) can be found in Sergeant et al. (2016).

Besides the use for dimensionality reduction, the PCA technique can be used for the clustering of the variables of a data matrix. Cadima and Jolliffe (1995) discusses the clustering of variables considering the eigenvectors of the PCA. The grouping of variables consists of choosing variables that have similar values for its eigenvectors in absolute value and are highly correlated to the principal component.

In the air pollution context, outliers may arise from different scenarios such as human-made disasters and natural catastrophes, measurement errors due to the fail-

ure of equipment or a sudden change in the atmosphere conditions, human errors, among others. Another critical situation arises when the observed pollutant is under the concentration limits established by legislation standards, but it may be considered as an atypical observation during the statistical analysis.

Furthermore, PCA is sensitive to outliers since the estimation of the mean vector, the covariance matrix, and the correlation matrix are directly influenced by outliers. As a consequence, the estimation of the eigenvalues and eigenvectors of the covariance or correlation matrix will be influenced by outliers present in the data, see, e.g., Filzmoser (1999). It is worthwhile to mention that even a single outlier may affect classical statistics methods. Croux and Haesbroeck (2000) indicate that conclusions obtained from principal component analysis calculated from a dataset with outliers may be misleading.

Under these circumstances, the common choice made by a wide range of scientists and practitioners to mitigate this problem is to delete the observations suspected to be outliers. As pointed out by Maronna et al. (2006, Chapter 1), the removal of an outlier observation may lead to many issues since the deletion is based on a subjective decision. A viable option to attenuate these problems is to use robust statistical methods since these methods still work well even when the presence of outliers is uncertain. Among the methods for robust estimation of the covariance or correlation matrix with time-independent datasets, there is the estimator proposed by Ma and Genton (2001). This estimator uses the so-called $Q_n(.)$ estimator proposed by Rousseeuw and Croux (1993), which is independent of the location parameter of the dataset.

In this paper, the central idea is to robustify the estimation of the covariance matrix before calculating its eigenvalues and eigenvectors in PCA. The methodology

proposed consists of the application of robust principal component analysis and selecting the stations which presented higher correlations to the selected PCs. Then, a decision rule is to be applied to decide to keep the redundant station in the same place or to move it to a new area. This proposed methodology is also adequate when outliers are presented in the dataset. The PM_{10} data of the metropolitan area of the Greater Vitória Region (GVR), Brazil, is analyzed as an illustrative example.

The paper is structured as follows: Section 2 describes the data and the statistical model introducing the proposed estimation method and how to identify monitoring stations that present similar behavior; Section 3 presents the data analysis and its discussion comparing robust PCA to the standard one. Finally, Section 4 presents the closing remarks.

2 Data and methods

2.1 Sampling stations in the Greater Vitoria Region

The Greater Vitória Region is located on the southeast coast of Brazil (latitude 20ž 19S, longitude 40ž20W) with a population of approximately 1.900.000 inhabitants. The climate is tropical humid, with average temperatures ranging from 24žC to 30žC. The region has many ports being an important cargo transport hub in Brazil. Also, there are many industries presented in the region, such as steel plants, iron ore pellet mill, stone quarrying, cement and food industry, and asphalt plant.

The automatic air pollution monitoring network (AAQMN) of GVR is consisted by eight monitoring stations distributed in the cities of this region as follows: two stations in Serra (Laranjeiras and Carapina), three stations in Vitória (Jardim Camburi, Enseada do Suá and Vitória Centro), two stations in Vila Velha (Vila Velha

Centro and Ibes) and one station in Cariacica (at the regional food distribution center, CEASA). The PM_{10}, in $\mu g/m^3$, is monitored in all stations. Figure 1 presents the geographical location of each station. The PM_{10} series corresponds to the daily average (over 24h period) observed at all stations from January 2005 to December 2009.

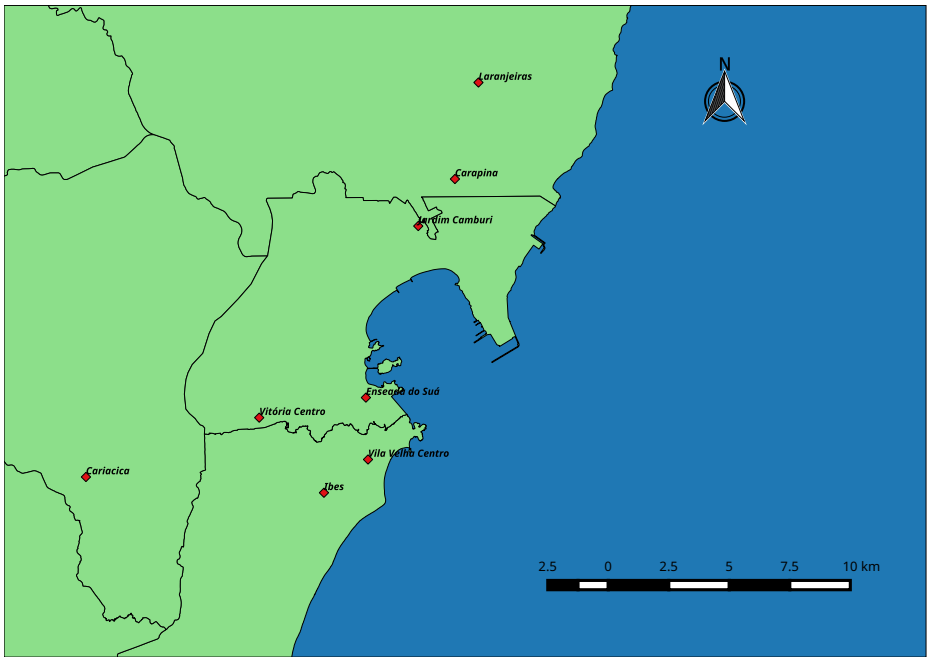


Fig. 1 Geographical location of the stations.

2.2 Principal component analysis

Most of the practitioners employ the standard PCA, which is based on the sample covariance matrix and is summarized in the sequel. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample of size n of an independent and identically distributed multivariate distribution with dimension p, mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. The method of moment

estimator (MME) of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})', \tag{1}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$. As stated by Jollife (2002), the big drawback of PCA tool based on covariance matrices is the sensitivity of the PCs to the units of measurement of the variables. Therefore, if large differences in the variances of variables are found, the variables with large variances will tend to dominate the first PCs. To avoid this problem, the use of PCA based on the correlation matrix is suggested. To this end, the sample correlation matrix $\hat{\boldsymbol{P}}$ can be obtained as $\hat{\boldsymbol{P}} = \hat{\boldsymbol{D}}\hat{\boldsymbol{\Sigma}}_n\hat{\boldsymbol{D}}$, where $\hat{\boldsymbol{D}} = \text{diag}(1/\sqrt{\hat{\boldsymbol{\sigma}}_{11}}, \ldots, 1/\sqrt{\hat{\boldsymbol{\sigma}}_{pp}})$, where $\hat{\sigma}_{ii}$, for $i = 1, \ldots, p$, is the sample covariance. It is straightforward to see that even one outlier will affect the sample mean, and, thus, the whole covariance (or correlation matrix).

Now, consider the random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ with sample covariance matrix $\hat{\boldsymbol{\Sigma}}_n$ and its associated sample eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ with corresponding normed eigenvectors $\hat{\boldsymbol{a}}' = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]$. Let

$$\hat{Y}_i = \hat{\boldsymbol{a}}_i'\boldsymbol{X}. \tag{2}$$

Then, we have

$$\widehat{\text{Var}}(\hat{Y}_i) = \hat{a}_i'\hat{\boldsymbol{\Sigma}}_n\hat{a}_i = \hat{\lambda}_i, \quad i = 1, 2, \ldots, p, \tag{3}$$

$$\widehat{\text{Cov}}(\hat{Y}_i, \hat{Y}_k) = \hat{a}_i'\hat{\boldsymbol{\Sigma}}_n\hat{a}_k = 0, \quad i \neq k, i, k = 1, 2, \ldots, p, . \tag{4}$$

If some $\hat{\lambda}_i$ are equal, the choice of the corresponding eigenvectors $\hat{a}_i$ is not unique.

Associated with (2), it can be shown that

$$\sum_{i=1}^{p}\widehat{\text{Var}}(X_i) = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p = \sum_{i=1}^{p}\widehat{\text{Var}}(\hat{Y}_i). \tag{5}$$

Equation 5 states that the whole variability of $\boldsymbol{X}$ is retained by the principal components $\hat{\boldsymbol{Y}}$. Therefore, if the main goal of the use of PCA is to reduce the number of variables, the scientist may choose to retain only part of the total original variability.

2.2.1 Robust PCA

Outliers affect the estimation of the location (mean) and the scale (variance) of random variables. To address this problem, Rousseeuw and Croux (1993) proposed a robust estimator, Q_n, for the dispersion of a dataset. Let $X_1, \ldots, X_n$ be n i.i.d. copies of a random variable X, the estimator Q_n is the k-th order

$$Q_n(x) = d\ \{|X_i - X_j|; i < j\}_{\{k\}}, \tag{6}$$

where $i, j = 1, \ldots, n$, and d is a value for consistency of the estimator. The k-th order statistic is the integer value $k = \lfloor (\binom{n}{2} + 2)/4 \rfloor + 1$.

It is known that for any univariate second-order random variables X and Y it is possible to compute the covariance between them as follows

$$\mathrm{Cov}(X, Y) = \frac{\alpha\beta}{4}\big(\mathrm{Var}(X/\alpha + Y/\beta) - \mathrm{Var}(X/\alpha - Y/\beta)\big), \tag{7}$$

for any $\alpha, \beta \in \mathbb{R}$, see, Huber (2004). In order to robustify (7), Ma and Genton (2001) proposed to use the estimator Q_n instead of the sample variance obtaining

$$\hat{\sigma}_{Q_n}(X, Y) = \frac{\alpha\beta}{4}\left[Q_n^2\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) - Q_n^2\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)\right], \tag{8}$$

where $\alpha = Q_n(X)$ and $\beta = Q_n(Y)$.

The correlation between the univariate second-order random variables X and Y can be estimated by

$$\hat{\rho}_{Q_n}(X, Y) = \frac{Q_n^2\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) - Q_n^2\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)}{Q_n^2\left(\frac{X}{\alpha} + \frac{Y}{\beta}\right) + Q_n^2\left(\frac{X}{\alpha} - \frac{Y}{\beta}\right)}, \tag{9}$$

where X, Y, α and β are defined in (8).

Let $\boldsymbol{X}$ be a random vector of $p \geq 2$ variables. The robust sample covariance and correlation matrix of the random vector $\boldsymbol{X}$, namely, $\hat{\boldsymbol{\Sigma}}_{Q_n}$ and $\hat{\boldsymbol{P}}_{Q_n}$, respectively, are obtained by estimating every covariance or correlation pairs between X_i and X_j, $i, j = 1, \ldots, p$. In this work, the robustified principal component analysis is achieved by replacing the standard covariance (or correlation matrix) with $\hat{\boldsymbol{\Sigma}}_{Q_n}$ and $\hat{\boldsymbol{P}}_{Q_n}$.

It is worthwhile to mention that the robust estimation procedure discussed above will provide similar results to the ones estimated using the standard sample estimator when there are no outliers presented in the dataset. Therefore, its usage is recommended.

2.2.2 PCA clustering and station selection

PCA technique can also be used for clustering of the variables. A method for clustering variables using PCA is discussed in Cadima and Jolliffe (1995). The grouping of variables consists of choosing variables that have similar values for its eigenvectors in module and are highly correlated to the principal component. The correlation between a retained PC group and the related full PC (containing all the variability of $\hat{Y}_i$) is given by

$$\hat{r}_k = \hat{\lambda}_j^{1/2} (\hat{\boldsymbol{a}}_j^{k\prime} \hat{\boldsymbol{\Sigma}}_{n,k}^{-1} \hat{\boldsymbol{a}}_j^k)^{1/2}, \tag{10}$$

where $\hat{\lambda}_j$ is eigenvalue of j-th component, $\hat{\boldsymbol{a}}_j^k$ is the clustered vector of $\hat{\boldsymbol{a}}_j$ containing k variables and $\hat{\boldsymbol{\Sigma}}_{n,k}^{-1}$ is the sub-matrix of $\hat{\boldsymbol{\Sigma}}_n$, which lines and columns correspond to the k grouped variables.

The main idea behind the method is to address monitoring stations which present similar behaviors for the PM_{10} pollutant (the technique is easily expanded to any

other pollutant or meteorological parameter). Thus, a decision rule can be applied to decide to keep the redundant station in the same place or to move it to a new area.

As a possible decision rule, Pires et al. (2009) suggested three criteria: (i) sites should be monitoring the highest possible pollutant concentrations; (ii) the number of pollutants being monitored at each site should be maximized; and (iii) the distribution should maximize distances between locations.

In this context, the following methodology for addressing monitoring stations which present similar behavior for a given pollutant is proposed:

1. Perform a descriptive statistical analysis of the data to verify the occurrence of possible outliers and to check for different scale of the measured variables;
2. Compute the robust PCA using the covariance or the correlation matrix;
3. Select a desirable number of PCs to be retained, e.g., 80% or more of the total variability;
4. Arbitrarily choose a cutoff point for the absolute values of the eigenvectors;
5. Create a group of variables whose coefficient of eigenvectors are equal or greater than the cutoff point in the component;
6. Using (10) compute the correlation between the selected variables in the PC and the full component. If the chosen variables and the component are not correlated, verify the cutoff point and redo the steps 4-6;
7. Apply the decision criteria of Pires et al. (2009) to decide to keep or to move to a new area the monitoring equipment of the pollutant considered in the study.

3 Data analysis and discussion

In this study, the robust PCA was applied as a classification tool to group monitoring sites with redundant measurements of PM_{10} concentrations from January 1st of 2005 to December 31 of 2009 ($n = 1826$). All the plots and analysis were performed using the computing environment R. $\hat{\boldsymbol{\Sigma}}_{Q_n}$ and $\hat{\boldsymbol{P}}_{Q_n}$ are available in the package *tsqn* (Cotta et al. 2017). The dataset and the R codes are available upon request.

Table 1 shows the descriptive statistics (i.e., the averages, standard deviations, and quantile values, among others) of the variables considered. The concentrations of PM_{10} pollutants exceeded hourly and annually, the guidelines suggested by the World Health Organization (WHO 2005). It is observed a high range for all stations.

The boxplot of the data and the series of PM_{10} are shown in Figures 2 and 3, respectively. From the boxplot and the plots of the series, one can observe higher levels of PM_{10} pollutant compared to WHO's guidelines, where the established limit is 50 $\mu g/m^3$ for 24-hour concentrations. Although the high levels of PM_{10} are essential information that should be considered in the context of the air pollution and its impact on human health, these observations can be identified, from a statistical point of view, as being outliers. Therefore, the high levels of PM_{10} presented in the series justify the use and comparison of the robust PCA.

Tables 2 and 3 show the correlations and the robust correlations (as in Section 2.2) between the monitoring stations in the study. From both tables, we observe strong correlations between the variables, e.g., 0.78 for Ibes and Enseada do Suá stations.

The grouping of stations with redundant measurements for the PM_{10} pollutant was carried out following the methodology proposed in Section 2.2.2. That is, stations

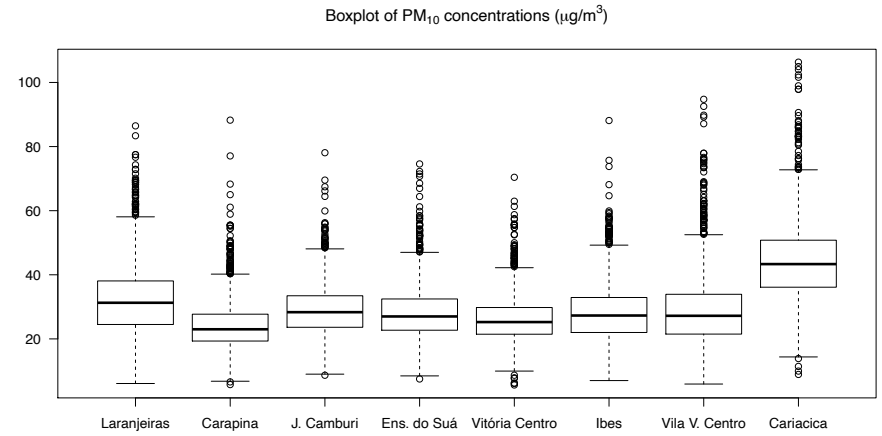


Fig. 2 Boxplot of PM_{10}'s concentrations of the AAQMN of the GVR.

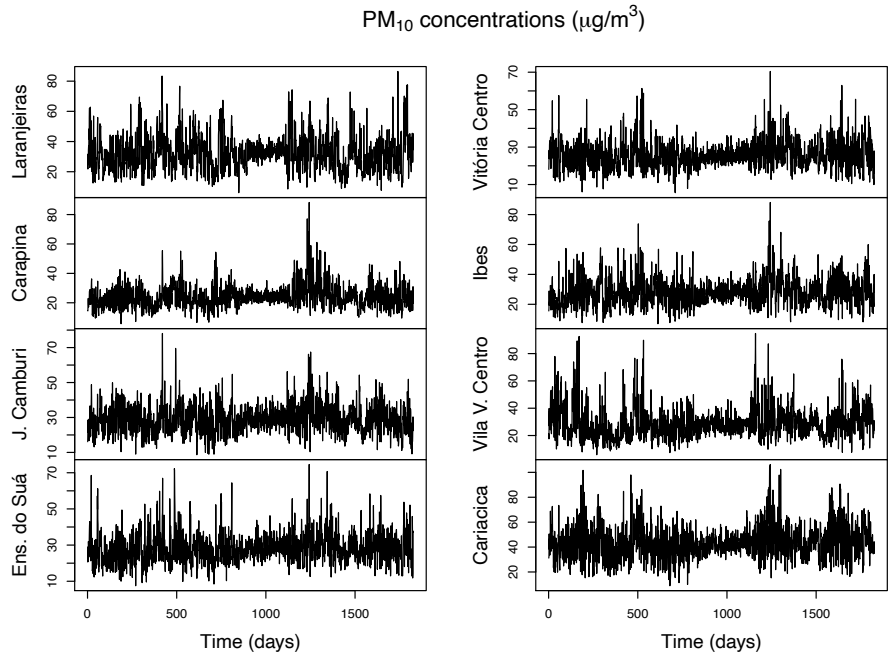


Fig. 3 PM_{10}'s concentrations of the AAQMN of the GVR.

Table 1 Descriptive statistics of PM_{10} data

	Laranjeiras	Carapina	Jardim Camburi	Enseada do Suá	Vitória Centro	Ibes	Vila Velha Centro	Cariacica
Mean	32.26	24.13	28.97	28.08	26.01	28.13	28.94	44.16
Std. Dev	11.29	7.67	8.01	8.29	7.37	9.20	11.33	13.12
Min.	6.08	5.75	8.67	7.50	5.62	7.00	5.92	8.92
25th perc.	24.50	19.33	23.64	22.71	21.46	22.01	21.51	36.14
50th perc.	31.27	23.00	28.33	27.00	25.25	27.29	27.21	43.33
75th perc.	38.07	27.71	33.46	32.46	29.78	32.91	33.92	50.79
Max.	86.46	88.25	78.08	74.58	70.42	88.12	94.75	106.30

Table 2 Correlation matrix ($\hat{\boldsymbol{P}}$) between the stations

	Laranjeiras	Carapina	Jardim Camburi	Enseada do Suá	Vitória Centro	Ibes	Vila Velha Centro	Cariacica
Laranjeiras	1.00							
Carapina	0.35	1.00						
Jardim Camburi	0.52	0.55	1.00					
Enseada do Suá	0.53	0.54	0.53	1.00				
Vitória Centro	0.45	0.63	0.59	0.67	1.00			
Ibes	0.58	0.61	0.61	0.72	0.64	1.00		
Vila Velha Centro	0.38	0.49	0.44	0.46	0.61	0.46	1.00	
Cariacica	0.42	0.70	0.56	0.54	0.71	0.69	0.46	1.00

Table 3 Robust correlation matrix ($\hat{\boldsymbol{P}}_{Q_n}$) between the stations

	Laranjeiras	Carapina	Jardim Camburi	Enseada do Suá	Vitória Centro	Ibes	Vila Velha Centro	Cariacica
Laranjeiras	1.00							
Carapina	0.40	1.00						
Jardim Camburi	0.59	0.57	1.00					
Enseada do Suá	0.59	0.58	0.59	1.00				
Vitória Centro	0.45	0.65	0.61	0.71	1.00			
Ibes	0.66	0.61	0.62	0.78	0.66	1.00		
Vila Velha Centro	0.44	0.55	0.48	0.54	0.60	0.56	1.00	
Cariacica	0.46	0.70	0.55	0.60	0.73	0.69	0.51	1.00

having the same contribution in a given component will have similar values for their eigenvectors, and they will also be correlated to the component.

In the PCA tool, the estimates of the eigenvalues and their corresponding eigenvectors using $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{P}}_{Q_n}$ are given in Table 4 where, for each component, the grouped stations are highlight by boldface. For both estimators, four components could explain approximately 85 % of the total variability of the dataset, leading to a dimension reduction of the data. It is observed that PCA computed by using $\hat{\boldsymbol{P}}_{Q_n}$ preserved a higher percentage of variability in the components.

For both PCAs, the cutoff point was selected to be 0.37 in absolute value, which led to the highest correlation values. In the standard PCA, this cutoff led to a correlation between the selected PCs groups and the original PCs of 0.96, 0.88, 0.66, and 0.96, for the four PCs, respectively. In the case of robust PCA, correlations of 0.96, 0.89, 0.66, and 0.95 were found. The values are close in both standard and robust PCA.

Thus, for the method of moments estimator for the first component, it is possible to visualize the existence of a group of stations formed by Ibes, Vila Velha Centro, and Cariacica. In the second component, the group is formed by Laranjeiras and Carapina. For the third component, Vila Velha Centro forms a group. Finally, the fourth component is the group formed by Jardim Camburi, and Enseada do Suá.

For the grouping through robust PCA, in the first component, Ibes, Enseada do Suá, and Vitória Centro can be grouped. For the second component, Laranjeiras and Cariacica form a group. In the third component, Vila Velha Centro is the only station in the group. For the fourth component, the group is formed by Enseada do Suá and Jardim Camburi. Therefore, the proposed method allocated groups differently from

Table 4 PCA results for PM_{10} of AAQMN of the GVR

Stations	PCA - $\hat{\boldsymbol{P}}$				PCA - $\hat{\boldsymbol{P}}_{Q_n}$			
	1	2	3	4	1	2	3	4
Laranjeiras	-0.3002	**0.7193**	-0.1756	0.1460	-0.3123	**0.6998**	0.0533	0.0683
Carapina	-0.3554	**-0.4004**	0.2628	0.1750	-0.3488	**-0.4144**	-0.1961	0.2701
Jardim Camburi	-0.3472	0.1700	0.0502	**0.7019**	-0.3446	0.2356	-0.2115	**0.7037**
Enseada do Suá	-0.3632	0.2163	0.0406	**-0.6118**	**-0.3722**	0.1519	-0.0045	**-0.5144**
Vitória Centro	**-0.3864**	-0.2265	-0.1026	-0.1629	**-0.3745**	-0.2867	-0.0211	-0.1276
Ibes	**-0.3869**	0.1787	0.2359	-0.2271	**-0.3863**	0.1902	-0.0881	-0.3395
Vila Velha Centro	-0.3055	-0.2942	**-0.8391**	0.0141	-0.3203	-0.1838	**0.8942**	0.1475
Cariacica	**-0.3721**	-0.2766	0.3542	0.0507	-0.3625	-0.3283	-0.3259	-0.0962
Eigenvalue	4.8971	0.7744	0.6282	0.4973	5.146	0.7568	0.5334	0.4612
Proportion	61.22	9.68	7.85	6.22	64.25	9.46	6.67	5.77
Cumulative	61.22	70.90	78.75	84.97	64.25	73.71	80.38	86.14

$\hat{\boldsymbol{P}}$. However, base on boxplot (Figure 2) and descriptive statistics (Table 1), the grouping based on $\hat{\boldsymbol{P}}_{Q_n}$ is suggested here.

To visually confirm the grouping results for both estimators, the daily averages of PM_{10} for the groups are shown in Figure 4. It is seen that the grouping using $\hat{\boldsymbol{P}}_{Q_n}$ is superior since, for the first component, the grouped stations have similar concentrations.

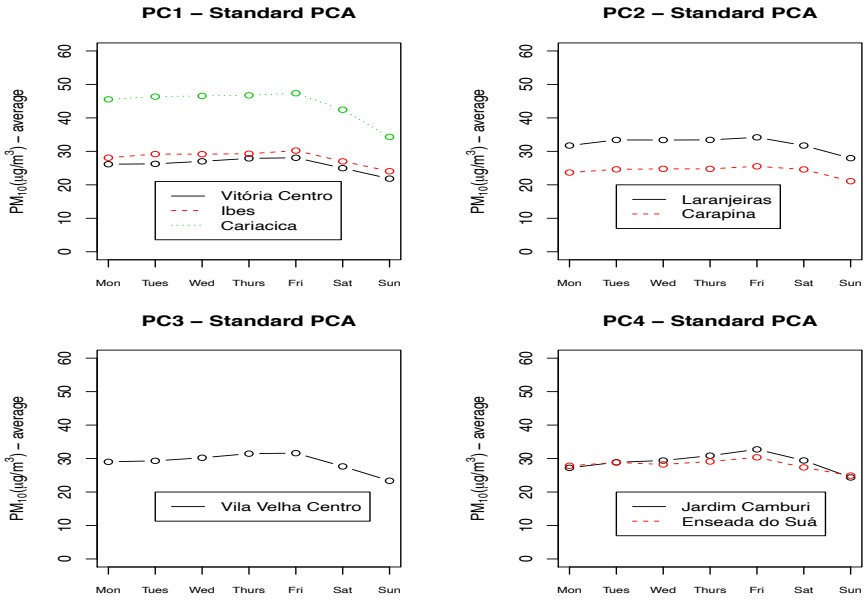
To end this analysis and continuing with the procedure of the methodology discussed in Section 2.2.2, the stations of Vitória Centro and Enseada do Suá may be selected to be moved to a new area to enlarge the total monitored area. It is highlighted that although Cariacica has no important contribution to the robust cluster, it is the only station located in Cariacica municipality and, therefore, must be kept.

4 Conclusions

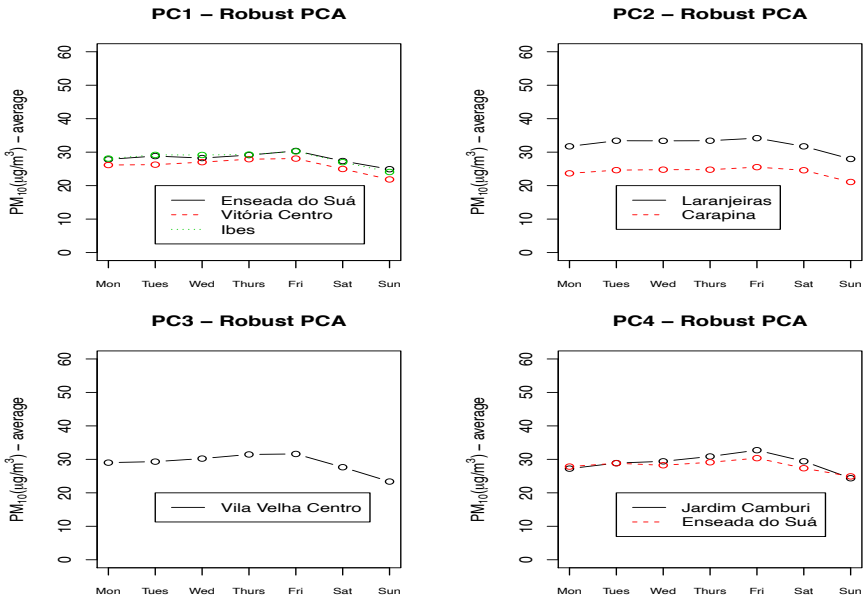
This article proposed and applied a grouping methodology to identify monitoring stations that present similar behavior for a given pollutant. As a case of study, the AAQMN of GVR (Brazil), which monitors the PM_{10} pollutant was considered in order to enable better management of the local monitoring network.

The methodology proposed consists of the application of robust principal component analysis and selecting the stations which presented higher contributions to the chosen PCs. Then, a decision rule is to be applied to decide to keep the redundant station in the same place or to move it to a new area.

In the case study, it was found the occurrence of possible outliers observations during the descriptive analysis of the PM_{10} data, which justified the comparison between the robust and standard PCA. It was found that Ibes, Enseada do Suá, and



(a) Standard PCA



(b) Robust PCA

Fig. 4 The daily average of the PM_{10} data.

Vitória Centro presented similar behavior and thus can be grouped. Also, Jardim Camburi and Enseada do Suá form another group. Therefore, two stations, Ibes and Enseada do Suá, are the candidates to be moved to a new site to enlarge the monitored area.

Acknowledgements The authors would like to thank the support from CNPq, ERASMUS, CAPES and FAPES. Part of this paper was revised when Professor Valdério Reisen was visiting CentraleSupélec in July 2018. This author is indebted to CentraleSupélec for its financial support.

References

Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, Wolf K, Samoli E, Fischer P, Nieuwenhuijsen M, et al. (2014) Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project. The Lancet 383(9919):785–795

Cadima J, Jolliffe IT (1995) Loading and correlations in the interpretation of principle compenents. Journal of Applied Statistics 22(2):203–214

Cesaroni G, Forastiere F, Stafoggia M, Andersen ZJ, Badaloni C, Beelen R, Caracciolo B, de Faire U, Erbel R, Eriksen KT, et al. (2014) Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 european cohorts from the escape project. BMJ 348:f7412

Conselho Nacional do Meio Ambiente (1990) Resolução CONAMA 003/90. Conama Brasília

Cotta HHA, Reisen VA, Bondon P, Lévy-Leduc C (2017) tsqn: Applications of the Qn Estimator to Time Series (Univariate and Multivariate). R Package version 1.0.0

Croux C, Haesbroeck G (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functons and efficiencies. Biometrika 87:603–618

Dominick D, Juahir H, Latif MT, Zain SM, Aris AZ (2012) Spatial assessment of air quality patterns in malaysia using multivariate analysis. Atmospheric Environment 60:172–181

Filzmoser P (1999) Robust principal component and factor analysis in the geostatistical treatment of environmental data. Environmetrics 10:363–375

Gramsch E, Cereceda-Balic F, Oyola P, Von Baer D (2006) Examination of pollution trends in santiago de chile with cluster analysis of PM_{10} and ozone data. Atmospheric environment 40(28):5464–5475

Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman JD (2013) Long-term air pollution exposure and cardio-respiratory mortality: a review. Environmental Health 12(1):1

Huber P (2004) Robust Statistics. Wiley Series in Probability and Statistics–Applied Probability and Statistics Section Series, Wiley

Instituto de Energia e Meio Ambiente (IEMA) (2014) 1ž Diagnóstico da Rede de Monitoramento da Qualidade do Ar no Brasil. Instituto de Energia e Meio Ambiente

Jollife IT (2002) Principal component analysis, 2nd edn. Prentice Hall

Lau J, Hung WT, Cheung CS (2009) Interpretation of air quality in relation to monitoring station's surroundings. Atmospheric Environment 43(4):769–777

Lu WZ, He HD, yun Dong L (2011) Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. Building and Environment 46(3):577–583

Ma Y, Genton MG (2001) Highly robust estimation of dispersion matrices. Journal of Multivariate Analysis 78:11–36

Maronna R, Martin D, Yohai V (2006) Robust statistics. John Wiley & Sons, Chichester

Phung D, Huang C, Rutherford S, Dwirahmadi F, Chu C, Wang X, Nguyen M, Nguyen NH, Do CM, Nguyen TH, et al. (2015) Temporal and spatial assessment of river surface water quality using multivariate statistical techniques: a study in can tho city, a mekong delta area, vietnam. Environmental monitoring and assessment 187(5):229

Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG (2008a) Management of air quality monitoring using principal component and cluster analysis-part i: SO_2 and PM_{10}. Atmospheric Environment 42(6):1249–1260

Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG (2008b) Management of air quality monitoring using principal component and cluster analysis-part ii: CO, NO_2

and O_3. Atmospheric Environment 42(6):1261–1274

Pires JCM, Pereira MC, Alvim-Ferraz MCM, Martins FG (2009) Identification of redundant air quality measurements through the use of principal component analysis. Atmospheric Environment 43(25):3837–3842

R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. Journal of the American Statistical Association 88(424):1273–1283

Rückerl R, Schneider A, Breitner S, Cyrys J, Peters A (2011) Health effects of particulate air pollution: a review of epidemiological evidence. Inhalation toxicology 23(10):555–592

Sergeant CJ, Starkey EN, Bartz KK, Wilson MH, Mueter FJ (2016) A practitioners guide for exploring water quality patterns using principal components analysis and procrustes. Environmental monitoring and assessment 188(4):249

Villas-Boas MD, Olivera F, de Azevedo JPS (2017) Assessment of the water quality monitoring network of the piabanha river experimental watersheds in rio de janeiro, brazil, using autoassociative neural networks. Environmental monitoring and assessment 189(9):439

World Health Organization (WHO) (2005) Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization

Zamprogno B (2013) PCA applied in time series data with applications to air quality data. PhD thesis, PPGEA - Universidade Federal do Espírito Santo, in press

Zhao L, Xie Y, Wang J, Xu X (2015) A performance assessment and adjustment program for air quality monitoring networks in shanghai. Atmospheric Environment 122:382–392