UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
August 2016, Vol. 9, No. 2, pp. 255–277

The role of principal component analysis in the evaluation of air quality monitoring networks [1]

El papel del análisis por componentes principales en la evaluación de redes de control de la calidad del aire

Josué M. Polanco Martínez[a]
josue.polanco@bc3research.org

Abstract

One of the most used statistical techniques in environmental sciences is the Principal Component Analysis (PCA). This technique consists in a linear decomposition of a set of correlated variables in terms of orthogonal basis function, so that they reduce the number of variables and remove the correlation between them. The PCA is widely used in the study of environmental phenomena, from the analysis of meteorological fields to the evaluation of air quality monitoring networks (AQMN). Today it's easy to find information about this method in english, but not in Spanish. For these reasons, we are highly motivated to contribute with this paper, which contains the state of the art, to evaluate AQMN by means of PCA. Additionally, some examples (simulated and real-world data) are presented to exemplify the use of this technique.

***Keywords*:** principal component analysis, air quality monitoring networks, redundant sensor detection.

Resumen

Una de las técnicas estadísticas de más amplio uso en estudios ambientales es el análisis por componentes principales (ACP). Esta técnica consiste en la descomposición lineal de un conjunto de variables correlacionadas en términos de funciones de base ortogonal, de tal modo que reducen el número de variables y eliminan la correlación entre ellas. El ACP es utilizado en una amplia gama de aplicaciones en el estudio de fenómenos ambientales, desde el análisis de campos meteorológicos,

[1] Polanco, J. M. The role of principal component analysis in the evaluation of air quality monitoring networks. *Comunicaciones en Estadística*, **9**(2), 271-294.

[a] Basque Centre for Climate Change, Bilbao, España & EPHE, PSL Research University, Laboratoire Paléoclimatologie et Paléoenvironnements Marine, UMR CNRS 5805 EPOC (Environnements et Paléoenvironnements Océaniques et Continentaux), Université de Bordeaux, Pessac, France.

hasta más recientemente (año 2006) en la evaluación de redes de control y vigilancia de la calidad del aire (AQMN). Hoy por hoy, es posible encontrar cierta cantidad de publicaciones en inglés sobre este último tipo de aplicaciones, pero hay una carencia de información en español respecto a su uso en la evaluación de AQMN. Por otro lado, debido a la importancia en muchas ciudades para hacer una adecuada evaluación y control de los contaminantes emitidos al aire, es de crucial importancia contar con un método estadístico de uso práctico para tal fin. Por estas razones, se presenta de manera concisa toda la información pertinente para evaluar AQMN mediante el ACP, así como algunos ejemplos con datos simulados y reales.

***Palabras clave*:** análisis por componentes principales, redes de control de la calidad del aire, detección sensores redundantes.

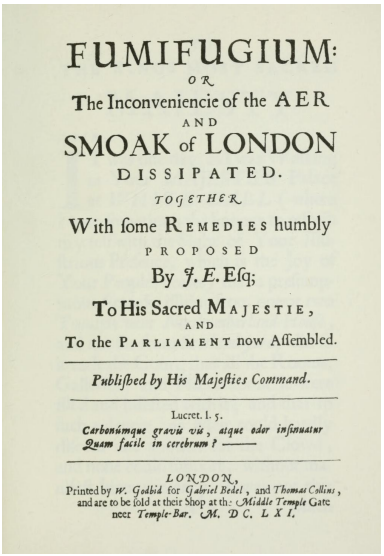
1 Introduction

The term atmospheric contamination refers to the presence of substances or energy forms that imply risk, damage or serious annoyances for humans and material goods (Aránguez et al. 1999). It's important to take into account that atmospheric contamination of natural origin has always existed due to biological, geological, chemical and physical processes that generate particles or pollutant gases, as volcanic eruptions, forest fires, sand storms, biological fermentations, etc. The discovering of the fire by the man originates atmospheric anthropogenic pollution. This type of contamination has acquired importance since the industrial revolution and because of the massive use of fossil fuels as energy sources (Aránguez et al. 1999, Wark & Warmer 1994).

The field of study of atmospheric pollution is very wide, it includes from the studies of greenhouse gases and its relationship with the climatic system, the destruction of the ozone layer due to the chlorofluorocarbons, the impact of accidental releases of chemical, biological or radionuclide contaminants in the atmosphere, to studies about the quality of the air (Seinfeld 1978, Sportisse 2010). However, in this paper we are interested in the concerning to the air quality, and in a particular way, in the statistical existing methods to evaluate in an objective way the air quality of a city.

One of the first descriptive studies related to the air quality dates back to mid of the seventeenth century, the *Fumifugium*, published by Johan Evelyn in 1648 (figure 1). The *Fumifugium* is about the impact of the use of charcoal as a fuel in the London environment and some measures to fight this type of contamination. Later, in 1692, Robert Boyle performed pioneering studies about the atmospheric chemical composition. With the advent of the industrial revolution, the number of studies related to air pollution was increasing. In this period the works of Robert A. Smith in the second half of the nineteenth century on acid rain stand out and the fact that he organized a monitoring network of air pollutants, considered the forerunner of today's Air Quality Monitoring Networks (AQMN) (Seinfeld 1978,

Sportisse 2010).



FUMIFUGIUM:
OR
The Inconveniencie of the AER
AND
SMOAK of LONDON
DISSIPATED.
TOGETHER
With ſome REMEDIES humbly
PROPOSED
By J. E. Eſq;
To His Sacred MAJESTIE,
AND
To the PARLIAMENT now Aſſembled.

Publiſhed by His Majeſties Command.

Lucret. l. 5.
Carbonúmque gravis vis, atque odor inſinuatur
Quam facile in cerebrum? ——

LONDON,
Printed by *W. Godbid* for *Gabriel Bedel*, and *Thomas Collins*, and are to be ſold at their Shop at the *Middle Temple* Gate neer *Temple-Bar*. *M. DC. LXI.*

Figure 1: *Cover of the* Fumifugium *or the inconvenience of the air and the smog disseminated in London and some remedies by J. Evelyn (1648). Source: available in http://www.archive.org/details/fumifugium00eveluoft*

With the massive increase of road traffic and the fast industrial and demographic growth in different places around the world between early and mid-twentieth century, anthropogenic pollutant emissions to the atmosphere soared significantly and with it its consequences (Martínez-Ataz & de Mera-Morales 2004, Seinfeld 1978). The impact on human health must be considered in a special way, as in the case of fatal incidents caused by acute episodes of air pollution, such as the great fog of London in 1952 (Seinfeld 1978, Sportisse 2010).

During the decade of the seventies and eighties the problems of atmospheric pollution were caused mainly by emissions of SO_2 and by particles suspended in the air, emitted mainly by industrial sources and, to a lesser extent, by urban cores (Martínez-Ataz & de Mera-Morales 2004). These problems of pollution of anthropogenic origin called the attention of both the scientific community and society in general, leading to regulation by environmental policies (Henry 1997, Martínez-Ataz & de Mera-Morales 2004).

As consequence of the legal regulation for the control of atmospheric pollutants, the levels of emissions of SO_2 and other air pollutants during the last 25 years has been decreasing in the majority of occidental industrialized countries. In contrast, SO_2 emissions are increasing in countries with emerging economies (Nunnari et al. 2004). Other of the consequences of environmental politics is that they motivated the development of catchment methods (sampling) and of analysis to measure

both the emissions and inmissions (Henry 1997, Martínez-Ataz & de Mera-Morales 2004).

With regard to the methods to capture and analyze the evolution of the inmissions of several types of pollutants and to carry out a control and monitoring of air quality automatic analyzers are mainly used today (Martínez-Ataz & de Mera-Morales 2004, World-Health-Organization 2000). These automatic analizers have several advantages regarding other methods for monitoring pollutants. They have for example a very high temporal resolution (Data can be obtained each hour or at a lower resolution), they can be installed in any suitable place, etc. However their cost is not low, they tend to be more susceptible to technical problems when they don't have proper maintenance and technical qualified staff and they require constant evaluation (Martínez-Ataz & de Mera-Morales 2004).

The set of automatic analyzers (nodes) that measures inmissions of pollutants and make a red of sampling is known as a net of control and vigilance of the air quality (AQMN) (Martínez-Ataz & de Mera-Morales 2004). These nets allow making an study and adeaute monitoring of the air quality. However they require a constant evaluation to ascertain and ensure that each one of their nodes provide a suitable characterization of the air quality in the zone where each sensor samples (Pires et al. 2008). Constant evaluations help to determine the right number of nodes of the net so that observation obtained is not redundant, to detect failures in some of the nodes or an inadequate spatial location of the nodes (Lau et al. 2009, Pires et al. 2009).

One of the useful tools to objectively evaluate the AQMN are the statistical multivariate techniques as the principal components analysis (PCA) or mathematical techniques of classification as the analysis of conglomerates. These techniques have been used for the evaluation and proper management of networks of water quality monitoring (Shrestha & Kazama 2007, Singh et al. 2004, Wunderlin et al. 2001). However the combined use of both techniques in the evaluation of a AQMN has been recently made in the year 2006 (Gramsch et al. 2006) to determine the seasonal trend and the spatial distribution of PM_{10} and O_3. Subsequently, the PCA has been used frequently to assess AQMN in different regions of the world (Ibarra-Berastegi et al. 2007, Ibarra-Berastegi et al. 2009, Pires et al. 2009, Pires et al. 2008).

Because of the need to quantitatively assess AQMN by applying multivariate statistical techniques, such as the PCA, it is necessary to have timely information on this subject. With respect to information in English there is a certain amount of publications (see, for example, Gramsch et al. (2006), Ibarra-Berastegi et al. (2007), Ibarra-Berastegi et al. (2009), Lau et al. (2009), Pires et al. (2009) and Pires et al. (2008)).However, that is not the case in Spanish language, with exceptions such as Polanco-Martínez (Polanco-Martínez 2012). The lack of this information in Spanish language has been one of the main motivations for writing this article.

The objective of this paper is to provide a brief revision of the statistical bases of the principal component analysis and its application in the evaluation of networks

of control and vigilance of the air quality. It provides also a case of study with simulated data (synthetic) and another one with real data of a AQMN located in the city of Bilbao for the period 2006-2010. The structure of the paper is the following. In section 2 the mathematical bases of the PCA are introduced. Section 3 provides information to interpret the PCA. Section 4 provides different rules to determine the number of principal components to retain. Section 5 presents the cases of study. Lastly, conclusions are presented in section 6.

2 Principal component analysis

The principal components analysis (PCA) is one of the most popular and ancient multivariate statistical techniques in the data analysis. This technique was developed by Karl Pearson[1] in 1901 (Pearson 1901), but it was not until 1939 when Hotelling made a much more formal presentation and coined the term principal component (PC) (Abdi & Williams 2010, Hotelling 1933). The PCA also receives other names, depending on its application field, v. g., in the theory of stochastic processes it is known as expansion or transformed of Karhunen-Loéve (Monahan et al. 2009), in turbulence as own orthogonal decomposition (Berkooz et al. 1993), in social and economic sciences as principal vectors (Kendall 1980) in atmospheric sciences as empirical orthogonal functions (Von Storch & Zwiers 1999, Wilks 1995).

Before submitting a formal definition and describe its main mathematical characteristics, we can mention that the PCA is a kind of linear transformation applied to a set of multivariate data commonly correlated with each other, to turn them into a smaller number of non correlated and orthogonal[2] variables, it is, to express the contained information in a set of data, with a smaller number of variables (Jolliffe 2002, Wilks 1995). Main objectives of the principal components analysis, according to Abdi & Williams (2010) and Jolliffe (2002), is to extract the most important information of a set of multivariate data, to compress a set of multivariate data keeping only the information considered important (reduce the dimensionality of data), simplify the description of a data set and analyze the structure of observations and variables.

2.1 Mathematical notation and preliminary concepts

Before presenting the methodological part of the PCA it is important to maintain adequate and consistent nomenclature. The methodological presentation is based and is analogous to the one of Abdi & Williams (2010) and Hannachi et al. (2007). To denote matrices, vectors and elements bolded uppercases will be used, bolded lowercases and italics lowercase, respectively. Matrices, vectors and elements of

[1] Although some investigations state that its origins date back to Cauchy (in 1829) and Jordan (in 1874) (Abdi & Williams 2010)

[2] Within the PCA, the concept of orthogonality of time series (the PCs are) correspond to the concept of uncorrelated series.

the same matrix will use the same letter, *v. gr.*, $\mathbf{X}, \mathbf{x}, x$. The transpose of a matrix will be represented with the superscript ^{T}. I will denote the identity matrix with $\mathbf{I}$. The column vector of ones and of length I is given by $\mathbf{1}_{(I\times 1)}$.

The data that are going to be analyzed through PCA contain I observations (samples)[3] and J variables. The number of observations is bigger than the number of variables in the cases of studio (Section 5) in this paper of revision, although it is not necessary. Each observation (sample) is obtained in the times $t_i, i = 1, 2, \cdots, I$ and are represented by the matrix $\mathbf{D}_{(I\times J)}$ where a ij-th element comes given by $d_{i,j}$. Before applying the PCA to the data that come from the AQMN it is required a very simple type of processing focused on the median, this is because by general rule data measuring these networks are in the same scale. Although it is important to verify whether the data analyzed have the same scale and if this were not the case, it is necessary typify (or apply some form of standardization) variables under study. Processing consist in subtract the median of the observations to each variable and work with the "anomalies" $\mathbf{X}_{(I\times J)}$ (Hannachi et al. 2007, Wilks 1995), it is:

$$\mathbf{X}_{(I\times J)} = \mathbf{D}_{(I\times J)} - \mathbf{1}_{(I\times 1)}\overline{\mathbf{D}}_{(1\times J)} \tag{1}$$

t where $\overline{\mathbf{D}}$ (vector of simple means) is given by:

$$\overline{\mathbf{D}}_{(1\times J)} = (\overline{\mathbf{d}}_{\mathbf{1}}, \overline{\mathbf{d}}_{\mathbf{2}}, ..., \overline{\mathbf{d}}_{\mathbf{J}}) = \frac{1}{I}\mathbf{1}^T_{(1\times I)}\mathbf{D}_{(I\times J)} \tag{2}$$

If we replace the third member of the relationship (2) in the relationship (1) and factoring, anomalies can also be represented, according to Hannachi et al. (2007), as follows:

$$\mathbf{X}_{(I\times J)} = \left(\mathbf{I}_{(I\times I)} - \frac{1}{I}\mathbf{1}_{(I\times 1)}\mathbf{1}^T_{(1\times I)}\right)\mathbf{D}_{(I\times J)} = \mathbf{M}_{(I\times I)}\mathbf{D}_{(I\times J)} \tag{3}$$

Where $\mathbf{M}_{(I\times I)}$ is the centering matrix of order I. Hereinafter we will not use (unless necessary) dimensional subscripts I, J to simplify mathematical notation.

2.2 On how to calculate the principal components

The most usual way to present the calculation of the principal components (PC) in the texts about analysis of environmental data (specifically climatic or meteorological data) is through the solution of an eigenvalue problem through the matrix of covariances of anomalies $\mathbf{X}$ of the data $\mathbf{D}$ under study (Hannachi et al. 2007, Von Storch & Zwiers 1999, Wilks 1995). This procedure can be mathematically expressed and according to Hannachi et al. (2007) as follows. The matrix of sample covariances of the matrix of anomalies $\mathbf{X}$ (relation (3)) is defined (Hannachi

[3]The term observation is used only for practical purposes, the data under analysis via PCA are not limited to observational data.

et al. 2007) by the relation:

$$\mathbf{S} = \frac{1}{I}\mathbf{X}^T\mathbf{X} \tag{4}$$

Where each element of $\mathbf{S}$ is formed by the covariances between each pair of variables of $\mathbf{X}$ of dimensions $I \times J$. However, the purpose of principal component analysis is to find a new set of variables (linear combinations) not correlated with each other to explain the maximum variance. This equates to find a unit vector $\mathbf{q} = (q_1, ..., q_J)^T$ such that $\mathbf{Xq}$ has the maximum variability (Hannachi et al. 2007, Von Storch & Zwiers 1999). It is

$$max\{\mathbf{q}^T\mathbf{Sq}\} \tag{5}$$

subject to condition $\mathbf{q}^T\mathbf{q} = 1$.

The eigenvectors or empirical orthogonal functions (EOFs) are obtained according to Hannachi et al. (2007), as a solution to the problem of eigenvalues

$$\mathbf{Sq} = \lambda\mathbf{q} \tag{6}$$

Where eigenvalues $\lambda_l, l = 1, 2, ..., N$ with $N = min(I, J)$, are given by:

$$\lambda_l = \mathbf{q}_l^T\mathbf{Sq}_l = \frac{1}{I}\|\mathbf{Xq}_l\|^2 \tag{7}$$

Eigenvalues λ_l provide a measure of the variance of $\mathbf{X}$ in the direction of $\mathbf{q}_l$. Once the problem of eigenvalues is resolved (relation 6), by general rule, these are arranged in a decreasing way (Hannachi et al. 2007), this is, $\lambda_1 \geq \lambda_2 ... \geq \lambda_N$. A common way to express the percentage of variance corresponding to each eigenvalue it through the relation

$$\frac{100\lambda_l}{\sum_{l=1}^{N}\lambda_l} \tag{8}$$

The l-th principal components are given by the projection of $\mathbf{X}$ on the I-th eigenvector $\mathbf{q}_l = (q_{1l}, q_{2l}, ..., q_{Jl})^T$ (Hannachi et al. 2007) and are expressed through the following relationship:

$$\mathbf{p}_l = \mathbf{Xq}_l \tag{9}$$

Elements $(p_{tl}, t = 1, ..., I)$ of the relation (9) can be expressed as:

$$p_{tl} = \sum_{j=1}^{J} x_{tj}q_{jl} \tag{10}$$

Such that the I-th eigenvalue λ_l represents the variance of the I-th principal component $\mathbf{p}_l = (p_{l1}, p_{l2}, ..., p_{lI})^T$ (Hannachi et al. 2007).

However, to calculate the main components, in practice, not the matrix sample covariance (relation (4)) is calculated nor the eigenvalue problem (relation (7)) is solved, but is calculated by the decomposition of $\mathbf{X}$ for singular values (Singular Value Decomposition - SVD) (Abdi & Williams 2010, Hannachi et al. 2007). The fact of using the SVD is mainly due to issues of computational efficiency. On the other hand, the SVD have the advantage over the eigenvalue problem due to the possibility of operating on non-square matrices, as well as calculate and display sequentially the first l eigenvectors in an orderly manner without having to calculate all the eigenvalues of the matrix of sample covariances $\mathbf{S}$.

This procedure is described below (Abdi & Williams 2010, Hannachi et al. 2007). The matrix $\mathbf{X}_{(I\times J)}$ is decomposed by singular values as follows:

$$\mathbf{X}_{(I\times J)} = \mathbf{P}_{(I\times r)}\mathbf{\Sigma}_{(r\times r)}\mathbf{Q}^T_{(r\times J)} \tag{11}$$

Where matrices $\mathbf{P}$ (known as the matrix of principal components) and $\mathbf{Q}$ (known also a the matrix of loads or of projection) are such that their columns $\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p}_r$ and $\mathbf{q_1}, \mathbf{q_2}, ..., \mathbf{q}_r$, are orthogonal and are called singular vectors left and right, respectively. The rank of $\mathbf{X}$ is r and $\mathbf{\Sigma} = Diag(\sigma_1, \sigma_2, ..., \sigma_r)$ is a diagonal matrix whose elements $\sigma_1 \geq \sigma_2 ... \geq \sigma_r \geq 0$ are the singular values of $\mathbf{X}$ (Abdi & Williams 2010, Hannachi et al. 2007). However, due to the property of orthogonality of the eigenvectors (they form a base), the relation (11) can be expanded (decomposed) as a linear combination (Hannachi et al. 2007, Wilks 1995), that is:

$$\mathbf{X} = \sum_{l=1}^{r} \sigma_l \mathbf{p}_l \mathbf{q}_l^T \tag{12}$$

It is also possible to express the matrix of covariances $\mathbf{S}$ in terms of the SVD (Abdi & Williams 2010, Hannachi et al. 2007). If relation (11) is replaced in (4) we obtain:

$$\mathbf{S} = \frac{1}{I}\mathbf{Q}\mathbf{\Sigma}^2\mathbf{Q}^T \tag{13}$$

where $\mathbf{\Sigma}^2 = Diag(\sigma_1^2, \sigma_2^2, ..., \sigma_r^2)$ and the singular values are arranged according to $\sigma_1^2 \geq \sigma_2^2 ... \geq \sigma_r^2$. Eigenvalues are related to the singular values through $\lambda_l = \frac{\sigma_l^2}{I},\ l = 1, ..., r$.

To end this subsection it is important to consider a pair of questions. From a computational point of view, it must be taken into account that as $\mathbf{q}_j$ (j-th column of $\mathbf{Q}$) as -$\mathbf{q}_j$ are right solutions of the SVD of matrix $\mathbf{X}$. If the computational package provides as solution $\mathbf{q}_j$ or -$\mathbf{q}_j$, the sign of the j-th column of $\mathbf{P}$ will appear changed. Sometimes a subsequent step is performed, it consists in the rotation of the matrix of projection $\mathbf{Q}$ through a matrix of rotation $\mathbf{R}$, it is, $\mathbf{Q}' = \mathbf{R}^T\mathbf{Q}$ (Abdi & Williams 2010, Hannachi et al. 2007). The matrix of rotation $\mathbf{R}$ can be orthogonal (orthogonal rotation) or not (oblique rotation). One of the main objectives of rotation of principal components is to ease up the orthogonality

condition, to have more localized structures in the space of easier intepretation (Hannachi et al. 2007, Jolliffe 2002). One of the most used methods to make the rotation is the Varimax, which was developed by Kaiser in 1958, although there are another methods of rotation, like the quartimax or promaz (Jolliffe 2002).

In the cases of study of this article, rotation has not helped in the interpretation of results (section 5) that's why it is not applied.

3 Interpretation of principal components

This section presents in a concise way three subsections with information (eg, terminology) for a proper interpretation of the results obtained when applying the PCA.

3.1 Contribution of an observation to a principal component

The importance of an observation for a principal component can be obtained by the ratio of the square of the principal component corresponding to this observation between the eigenvalue associated with this component. This ratio is known as the contribution of the i-th observation to the l-th component (Abdi & Williams 2010), is denoted by $ctr_{i,l}$, and it is expressed as:

$$ctr_{i,l} = \frac{p_{i,l}^2}{\sum_i p_{i,l}^2} = \frac{p_{i,l}^2}{\lambda_l} \tag{14}$$

where $p_{i,l}$ is the l-th principal component and λ_l its eigenvalue associated. $ctr_{i,l}$ can take values between 0 and 1 and for a certain l-th component the sum of the contributions of all observations is equal to 1. An useful suggestion is to base the interpretation of a component in the observations whose contribution is much bigger tan the average of the contribution, it is, observations whose contribution is bigger than $1/I$ (where I is the number of observations).

3.2 Contribution of a principal component to an observation

The importance of a component for a given observation may be estimated through the square cosine and indicates the contribution of a component to the distance squared of the observation to the origin. This corresponds to the square of cosine of the angle of the right triangle formed with the origin, observation and its projection on the principal component where sum of the squared cosines is equal to 1, and is thus calculated:

$$\gamma_{i,l}^2 = \frac{p_{i,l}^2}{\sum_l p_{i,l}^2} = \frac{p_{i,l}^2}{d_{i,\mathbf{g}}^2} \tag{15}$$

where $d^2_{i,\mathbf{g}}$ is the square of the distance of an observation given to the origin. In other words $d^2_{i,\mathbf{g}}$ is calculated as the sum of the squares of all the principal components for this observation. Components with big values of $\gamma^2_{i,l}$ contribute a to a good portion of the total distance, therefore, these components have importance for his observation (Abdi & Williams 2010, Jolliffe 2002).

3.3 Correlation of a component and a variable; loading factors.

The correlation between a principal component and a variable is known in the PCA jargon as factor loadings or spatial coefficients. Note that the sum of the squares of coefficients of correlation between a variable and all the components is equal to 1 (Abdi & Williams 2010). Although the use of loading factors seems to simplify the interpretation, the fact of being squared involves loss of information. It is, often a component appears to separate two or more groups, whose elements tend to have high scores of opposite sign, being its contribution far greater than the rest of observations. Sometimes components can be very useful elements to detect atypical multivariate specially in cases where the "weird" is not the individual values of variables but its joint appearance. It is therefore advisable to analyze both cases, it is, the loading factors and these squared.

Special care must be taken when using the term loading because it has several interpretations and can be a potential source of confusion. For this reason it is important to corroborate the meaning of loading either reading texts on PCA or in the outputs of the computer program used to calculate PCA (Abdi & Williams 2010, Wilks 1995). In this sense is quite useful the table 9.3 of Wilks (1995), where a variety of terminology and its synonyms for the PCA is provided.

4 On the number of principal components

One of the main applications of principal component analysis is to reduce the dimensionality of the variables in a data set; therefore, a problematic point is to determine how many principal components must be retained (Note that the maximum number of principal components that can be retained is equal to the number of variables) (Dray 2008, Wilks 1995). The decision of the number of components may lead to a loss of information (underestimation) or introduce random noise (overfitting) (Dray 2008). Although the PCA technique has more than a century of existence, this issue remains open until today. However, there are some rules known as stopping that can be helpful to determine the number of principal components to retain (Jolliffe 2002, Wilks 1995).

Recently Peres-Neto et al. (2005) made a compilation of 20 objective rules that divide into two large families. The rules based on intervals of confidence (v. Gr., Parallel analysis, methods of bootstrap based on eigenvalues or Bartlett's test)

and the rules based on average values of statistical tests (v. Gr., the rule Kaiser-Guttman, the broken-stick model or of the minimum average partial correlation). At the same time Peres-Neto et al. (2005) make a comparative analysis between rules. They found when using simulated data that, more than the stopping rule used or the number of elements and the grade of gaussianity of analyzed data, these rules are much more dependent on the existing correlations between the observations or between the variables (Peres-Neto et al. 2005). These results (dependence on the correlation) are similar to the obtained in previous studies but with real data (climatic data) (Preisendorfer 1988, Wilks 1995).

With regard to the use of stopping rules in similar studies to the case of study presented in this paper (Section 5), the two frequently used criteria are the criteria of Kaiser (Guttman) and the criteria of the percentage of accumulated variance (PVA) (Lau et al. 2009, Pires et al. 2008). The criterion of Kaiser considers only to retain the PCs with eigenvalues bigger than one (Jolliffe 2002, Wilks 1995). The problem with this criteria is that it can be very restrictive (even taking into account the suggestion of Jolliffe (2002), it is, to retain the PCs whose eigenvalues are equal or greater than 0,7) (Lau et al. 2009, Pires et al. 2008). The criterion PVA considers retaining the PCs whose accumulated percentages of variance exceed a certain value. For example, some authors (Pires et al. 2009, Pires et al. 2008) recommend taking into account only the PCs that exceeds at least the 90% of accumulated variance (PVA_{90}). This criteria (PVA_{90}) was used in the cases of study of this paper (Section 5). However it must be taken into account that these rules provide a good indication to know the number of PC to retain and, each case of study should be analyzed with the whole information (for example, the loading factors) provided by the PCA.

5 Cases of study

In this section two cases of study are introduced to simplify the use of the PCA in the evaluation of AQMN. first, a case of study with synthetic data is presented; it is, created data whose statistical characteristics we know a priori. In second place comes an example with real data of values of SO_2 emissions of a AQMN located in the metropolitan area of Bilbao, Spain. For the principal components analysis the computational package `FactoMineR` (Lê et al. 2008) was used in R (R Development Core Team 2009), whereas to represent graphically in 3D the space of the principal components the `R scatterplot3d` package was used (Ligges & Mächler 2003).

5.1 Synthetic data

The construction of the synthetic series for this case of study is made as follows. Three couples of bivariated autoregressive processes of order 1 AR1) are constructed, such that the members of each couple (X, Y), have a certain correlation

– $\text{CORR}[X(t), Y(t)] = \rho_{XY}$ – between them. An AR1 bivariated process is defined (Mudelsee 2014) as follows:

$$\begin{aligned} X(1) &= \mu^X_{N(0,1)}(1), \\ Y(1) &= \mu^Y_{N(0,1)}(1), \\ X(t) &= \rho_X X(t-1) + \mu^X_{N(0,1-\rho_X^2)}(t), \quad t = 2, ..., T, \\ Y(t) &= \rho_Y Y(t-1) + \mu^Y_{N(0,1-\rho_Y^2)}(t), \quad t = 2, ..., T, \end{aligned} \tag{16}$$

$$\begin{aligned} CORR[\mu^X_{N(0,1)}(1), \mu^Y_{N(0,1)}(1)] &= \rho_\mu \\ CORR[\mu^X_{N(0,1)}(t), \mu^Y_{N(0,1)}(t)] &= \frac{1-\rho_X\rho_Y}{[(1-\rho_X^2)(1-\rho_Y^2)]^{1/2}}\rho_\mu, \\ & \qquad t = 2, ..., T \end{aligned}$$

$$CORR[\mu^X_{N(0,1)}(t), \mu^Y_{N(0,1)}(u)] = 0, \quad t, u = 1, ..., T,\ t \neq u \tag{17}$$

Where μ_X and μ_Y are two white noise Gaussian processes, ρ_X and ρ_Y are the parameters of the autoregressive processes X and Y, respectively. The bi-variated model AR1 (equation 16) have to be strictly stationary, therefore it has to meet (Mudelsee 2014): $E[X(t)] = E[Y(t)] = 0$, $VAR[X(t)] = VAR[Y(t)] = 1$ and $CORR[X(t), Y(t)] = \rho_{XY} = \rho_\mu$.

In the first example, the correlations of the couples of the bivariate processes AR1 have the values (arbitrary, although correlations with high values, both positive and negative and medium values have been chosen) of 0.95, 0.50 and -0.75, whereas the parameters of the AR1 have the values of (0.92, 0.93), (0.45, 0.47) and (-0.72, -0.73) for each couple of bivariate processes AR1, respectively. The length of the series T is equal to 2191 elements, and it is the same for the 6 synthetic series. 2191 elements were used because it is the number of days in an interval between the 1-1-2005 and 31-12-2010 ("arbitrary dates"). On the other hand monthly averages (figure 2) of the 3 couples of bivariate processes AR1 generated through the relations 16 and 17 have been calculated. This is an habitual praxis in the pre-process of data measured in the AQMN to avoid possible effects of masking due to the possible existence of high frequency cycles (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012). All this is made in this case of study with the purpose of constructing the simulated data (synthetic) the closest as possible to real data. For the same reason and to maintain the terminology used in the analysis of AQMN, we will use the Word "sensors" to refer the synthetic series.

The result of applying the PCA to simulated data (figure 2) is presented in the table 1 and in the figure 3. As you can see in the table 1 a) the first 3 principal components explain almost the 90% of accumulated variance (in fact they explain a 86,35%), whereby the principal components from 4 to 6 explain only a small percentage of variance (in a particular way the PC-6 which explains only a 0,95% of variance). This information provides an idea grosso modo that only 3 principal components- and thus 3 "sensors"- are necessary to explain the biggest part of the variance of analyzed data. However the key information to determine the optimal number of "sensors" is provided by the loading factors (table 1b). As you can

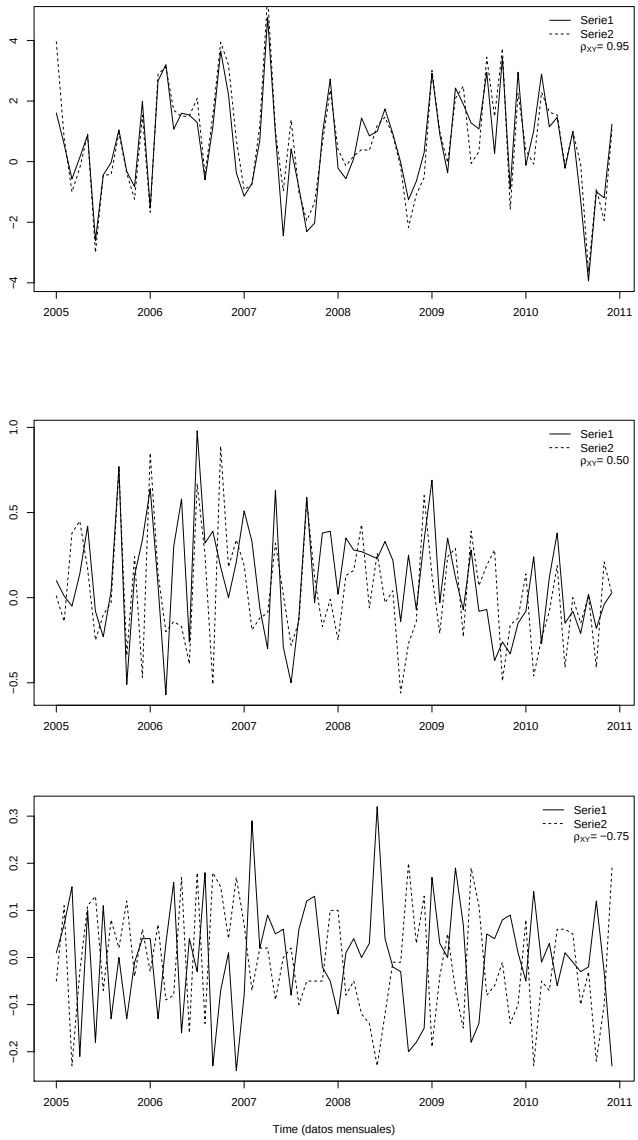


Figure 2: *Synthetic temporal series ("sensors") (72 monthly values). The correlations for each couple of bivariate processes AR1 are 0,95 (above), 0,50 (middle) and -0,75 (below). Parameters of the couples of processes AR1 are 0,92 (continuous line) and 0,93 (discontinuous line) (above), 0,45 (continuous line) and 0,47 (discontinuous line) (middle), -0,72 (continuous line) and -0,73 (discontinuous line) (below). Source: own elaboration.*

see, in the first two couples (1-2 and 3-4) of "sensors", the loading factors have very close values, of this can be interpreted that both the "sensors" 1 and 2, as the 3 and 4 measure the same information, therefore they are redundant. On the other hand the loading factors of the "sensors" 5 and 6 are numerically similar, but of opposite signs, which could indicate that they don't measure redundant information. However, they do it, just that one measures it directly and the other conversely. This information can be visually checked in figure 3, in which you can see that the "sensors" in the place of the principal components tend to make groups. On one hand, the sensors 1 and 2, on the other hand the sensors 3 and 4 and sensors 5 and 6, but reversely spaced. For all this, you can confirm the hypothesis that only 3 "sensors" measure not redundant information.

Table 1: *Summary of the principal components analysis. (a) Eigenvalues and percentage of the total variance explained by each PC and (b) loading factors. Source: own elaboration.*

a)

PC	Eigen Val.	% of var.	% of accum. var.
PC-1	2,12	35,39	35,39
PC-2	1,85	30,84	66,23
PC-3	1,21	20,13	86,35
PC-4	0,55	9,25	95,60
PC-5	0,21	3,44	99,05
PC-6	0,06	0,95	100,00

b)

"Sensor"	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6
1	0,58	0,79	0,06	-0,01	0,03	-0,17
2	0,64	0,74	0,09	0,00	-0,01	0,17
3	-0,46	0,23	0,68	0,52	-0,01	0,00
4	-0,50	0,25	0,64	-0,53	0,04	0,00
5	0,66	-0,56	0,37	0,03	0,32	0,00
6	-0,68	0,49	-0,44	0,06	0,31	0,02

5.2 Real data

The case of study with real data belongs to a AQMN located in the metropolitan area of Bilbao (Autonomous Community of the Basque Country, Spain - ACBC or CAPV in Spanish). This geographic zone belongs to one of the main regions with sources of atmospheric pollutants (due mainly to the industry and the road traffic)not only of the CAPV but at a national level (Gangoiti et al. 2002, Ibarra-Berastegi et al. 2007, Polanco-Martínez 2012). Like other industrialized cities of the end of the seventies and early eighties, metropolitan Bilbao showed problems of atmospheric pollution of industrial origin. For this reason a network of control and vigilance of the air quality was installed and one of the first in Europe (remote stations of the network of vigilance of air quality (*Estaciones remotas de la red de vigilancia de la calidad del aire. Departamento de Medio Ambiente y Política Territorial, Gobierno Vasco* 2016, Cambra et al. 2005)). One of the first

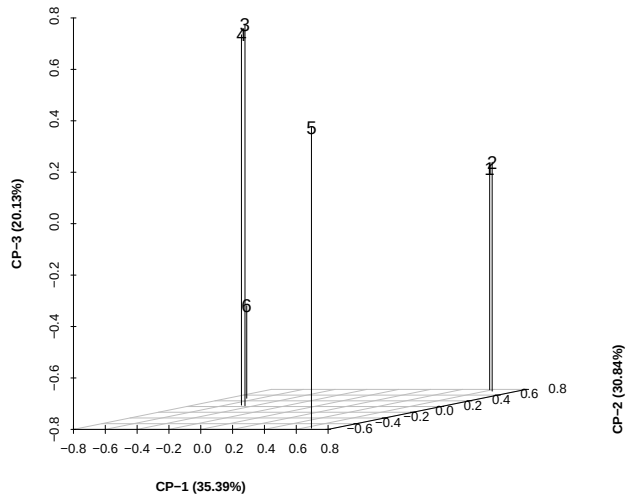


Figure 3: *Graphic representation for the 6 synthetic temporal series ("sensors") in the space of the principal components. Source: own elaboration.*

objectives of this network was to monitor the emissions of industrial origin of SO_2. However, in the last years the SO_2 levels have been declining due to changes in the industrial fabric and to the fuels currently used. However, the current environmental legislation forces to continue with the measurements of SO_2 emissions and other pollutants (Ibarra-Berastegi et al. 2009). Today the stations covering the urban area and the metropolitan zone of Bilbao make part of a bigger network covering the whole ACBC with a total of 51 stations (*Estaciones remotas de la red de vigilancia de la calidad del aire. Departamento de Medio Ambiente y Política Territorial, Gobierno Vasco* 2016).

Since the installation of the AQMN of metropolitan Bilbao till now, the network has suffered important modifications (for example, in the location of the stations or in the increase of these) with the purpose of capturing the representative fields of the emissions of the different pollutants, due to the changes in the location or the increase or decrease of the sources of emissions of SO_2 and other pollutants (Albizuri 2008). This kind of changes is not limited to the AQMN of metropolitan Bilbao, but they are practically present in the majority of the networks of control and vigilance of the air quality of other geographic areas. Because of these reasons, both the process of design and the control and evaluation of a AQMN is a dynamic and interactive process (Ibarra-Berastegi et al. 2009). Therefore it is important

to maintain a constant evaluation of these networks to capture the representative trajectory of the species of pollutants (Lau et al. 2009, Pires et al. 2009, Pires et al. 2008).

Previous studies (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012) proved that the four sensors of the AQMN from the metropolitan area of Bilbao don't measure redundant information of emissions of SO_2 for the period 1996-2001. Therefore all those sensors are necessary to a right evaluation of the emissions of SO_2 for that area of study. In this case of study of this paper of review I present an evaluation of the sensors that measure SO_2 emissions and are located in metropolitan Bilbao but for the period 2006-2010 (figure 4). This with the purpose of corroborate if these four sensors are still necessary for an accurate measurement of the SO_2 emissions.

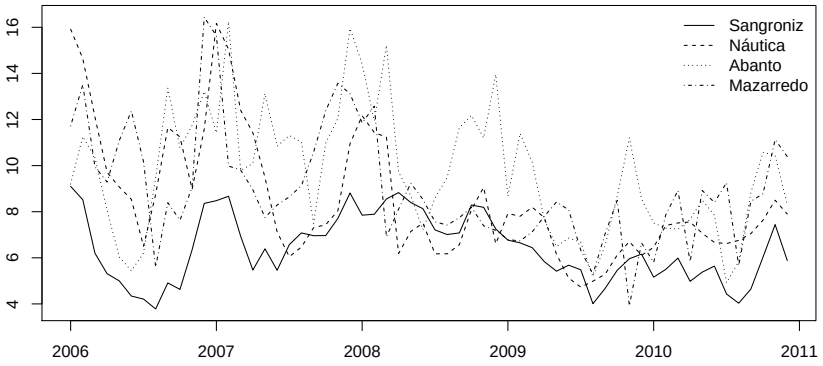


Figure 4: *Temporal series (60 monthly values) of SO_2 immmissions for the four stations. Sangroniz, Náutica, Abanto and Mazarredo. Unities of the $SO2_2$ emissions are in μg^{-m3}. Source: own elaboration.*

The result of the principal component analysis for the monthly series of emissions of SO_2 is presented in table 2. As you can see (table 2 a) the first three principal components explain a high percentage of variance (93.53%), which could indicate that only three sensors are necessary to an accurate characterization of the SO_2 emissions for metropolitan Bilbao. However, according to the loading factors (table 2 b), these have only similar values in the loading factors of the first principal component[4]. But the loading factors for the other components differ, both in sign and in value. This is, the first PC can be interpreted as the result of the mixture of SO_2 emissions that cross metropolitan Bilbao. It is, it represents the contribution of SO_2 emissions of different sources that get mixed and distributed

[4]Which indicates that the first principal component represents a common variability of the SO_2 emissions in metropolitan Bilbao.

uniformly in the area of study, measured by the four sensors installed in the zone. The influence of the first PC to the sensors (2 c) varies relatively little from one to another, with values from 52,52 (Mazarredo) to 67,58 (Náutica). On the other hand, the contribution of each sensor to the first CO (2 c) also varies little, with values oscillating between 21,13 (Mazarredo) and 27,19 (Náutica).

The second and third PC can be physically interpreted from local patterns of emissions or dispersions of SO_2 (table 2 c and d) but in an opposite way, since its loading factors have opposite signs (couples Sangroniz-Náutica and Abanto and Mazarredo, table 2 b). Take into account that, on the one hand, for the case of the sensors Abanto and Mazarredo, these are located in zones that present different meteorological or orographic conditions. Abanto is located in the periphery of metropolitan Bilbao in a zone with a strong presence of big industrial infrastructures, whereas Mazarredo is in the downtown of Bilbao. On the other hand, Sangroniz is located in the periphery of metropolitan Bilbao where there are not big population's cores or industrial infrastructures, while Náutica is close to the sea.

On the other hand, according to the information provided by the space of the principal components (figure 5), it is possible to appreciate that the sensors are far from each other and are not grouped. However, it is remarkable the way in which the sensors in Sangroniz y Mazarredo show an inverse relationship, which suggests that this couple of sensors could be measuring redundant information, although in an inverse way. This can be explained because Mazarredo is in the downtown of Bilbao, therefore this sensor measures emissions of pollutants directly related to the traffic. While Sangroniz is in the periphery of metropolitan Bilbao in a sparsely populated area and without big industrial infrastructures in its surrounding, therefore the emissions being measured in Sangroniz are not in a direct way, which could be influenced by topographic factors and meteorological conditions.

An important information that must be taken into account is that the installed sensor in Sangroniz is measuring emissions of SO_2 with less intensity than the other 3 sensors. This can be observed in figure 4, which shows that the sensor in Sangroniz is the sensor that measures the minor values of emissions of SO_2. Taking into account these analysis we can establish that it's not relevant to have a sensor in Sangroniz. In fact, currently the Department of Environment and Territorial Politics of the Basque Government does not measure the values of emissions of SO_2 in Sangroniz, although it measures another pollutants species.

Given the above, it can be established that not all four sensors are required for accurate characterization of SO_2 emissions values for metropolitan Bilbao, so the sensor that measures SO_2 emissions in Sangroniz should be removed. These results do not seem to agree with previous studies in the same area and with practically the same sensors (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012). However it must be taken into account that this work is focused on the temporal interval 2006-2011, which is much more recent than the interval of study analyzed by (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012) between 1996-2002. Therefore, it is not expected that the results of both studies are similar and even more because

the areas highly influenced by the anthropogenic component (and metropolitan areas of big cities are influenced) are constantly changing.

Regarding the fourth PC, one can see that the loading factors (Table 2 b) oscillate between relatively low values (correlations) from -0.25 to 0.30, like the contributions of the principal components to each sensor variability (table 2 c), which have values between 5.21 and 8.87.Although the contributions of each sensor to the total variability of each principal component (Table 2 d) present contributions with values (between 20.13 and 34.27) similar to PC1. Therefore, the interpretation of the fourth PC is not obvious.

Table 2: *Summary of the principal component analysis. (a) Eigenvalues and percentage of total variance explained by each PC and (b) loading factors. Source: own elaboration.*

a)

Principal componente	Eigen Val.	% of var.	% of accum. var.
PC-1	2,48	62,15	62,15
PC-2	0,77	19,31	81,46
PC-3	0,48	12,07	93,53
PC-4	0,26	6,47	100,00

b)

Sensor	PC-1	PC-2	PC-3	PC-4
Sangroniz	0,82	-0,28	0,43	-0,25
Náutica	0,82	0,24	-0,46	-0,24
Abanto	0,78	-0,52	-0,18	0,30
Mazarredo	0,72	0,61	0,23	0,23

c)

Sensor	PC-1	PC-2	PC-3	PC-4	$\sum$
Sangroniz	67,18	7,96	18,63	6,23	100
Náutica	67,58	5,57	21,27	5,57	100
Abanto	61,31	26,64	3,18	8,87	100
Mazarredo	52,52	37,09	5,18	5,21	100
$\sum$/100	2,48	0,77	0,48	0,28	

d)

Sensor	PC-1	PC-2	PC-3	PC-4
Sangroniz	27,02	10,30	38,60	24,07
Náutica	27,19	7,21	44,07	21,53
Abanto	24,66	34,48	6,59	34,27
Mazarredo	21,13	48,00	10,74	20,13
$\sum$	100	100	100	100

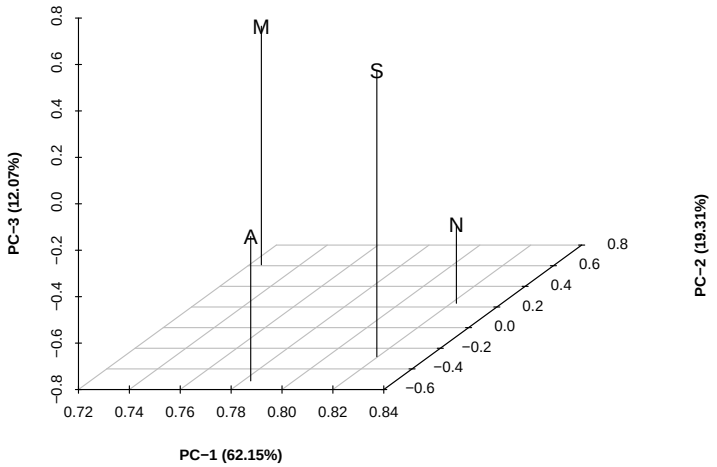


Figure 5: *Graphic representation for the four sensors in the space of the principal components. Abanto (A), Mazarredo (M), Na?utica (N) and Sangroniz (S). Source: own elaboration.*

6 Conclusions

The use of the analysis of principal components for the evaluation of networks of control and vigilance of the air quality is today one of the most used statistical tools. However, the biggest part of the information is available in English. Due to the potential uses, not only academicals, but also in practical cases (v. gr., civil protection, environment secretaries or agencies, etc.) it is necessary to have this information in Spanish, reason why it is presented in this work of review. On the one hand in this work the statistical bases of PCA are presented in such a way that it is accessible to a wide sector of scientific disciplines (with a strong focus on natural sciences), but without loosing mathematical rigor. On the other hand, the key points for an accurate interpretation of the PCA when evaluating AQMN are presented.

This paper presented also two cases of study to exemplify in a practical way the evaluation of AQMN through PCA. In the first example I used simulated

data knowing that they present redundant information, which was verified when applying PCA. While in the second example monthly values of SO_2 emissions coming from a AQMN located in metropolitan Bilbao for the period 2006-2010 were evaluated. Based on the results it is concluded that only three of the four sensors of the AQMN of the metropolitan area of Bilbao are needed. Therefore, it is recommended to remove the sensor that measures SO_2 emissions located in Sangroniz.

Acknowledgement

I am grateful to Prof. Dr. J. Saenz and Dr. G. Ibarra, who initiated me into the world of analysis of networks of control and monitoring of air quality by multivariate statistical techniques and to the reviewer of this work for their thoughtful suggestions to greatly improve this paper. The financing of post-doctoral support of the Basque Government (Ref. No. `POS_2015_1_0006`) is appreciated.

Received: 29 February 2016
Accepted: 29 April 2016

References

Abdi, H. & Williams, L. J. (2010), 'Principal Component Analysis', *Wiley Interdisc. Rev.: Comp. Stat.* **2**(4), 433–459.

Albizuri, A. (2008), *in* 'Caracterización de patrones meteorológicos a escala regional y local y su relación con los niveles de calidad del aire registrados en la C.A.P.V. Análisis de episodios', Memorias de la 3a. Jornada técnica sobre contaminación atmosférica, Dept. de Medio Ambiente, Planificación Territorial, Agricultura y Pesca, Gobierno Vasco.

Aránguez, E., Ordóñez, J. M., Serrano, J., Aragonés, N., Fernández-Patier, R., Gandarillas, A. & Galán, I. (1999), 'Contaminantes atmosféricos y su vigilancia', *Revista Española de Salud Pública* **73**(2), 123–132.

Berkooz, G., Holmes, P. & Lumley, J. L. (1993), 'The proper orthogonal decomposition in the analysis of turbulent flows', *An. Rev. of Fluid Mech.* **25**(1), 539–575.

Cambra, E., Alonso, E., F., C. & Martínez-Rueda, T. (2005), 'Health impact assessment of air pollution', ENHIS-1 project WP5 health impact assessment, Local City Report Bilbao.

Dray, S. (2008), 'On the number of principal components: A test of dimensionality based on measurements of similarity between matrices', *Comp. Stat. and Data Analysis* **52**(4), 2228–2237.

Estaciones remotas de la red de vigilancia de la calidad del aire. Departamento de Medio Ambiente y Política Territorial, Gobierno Vasco (2016), `http://www.ingurumena.ejgv.euskadi.eus/informacion/la-red-de-control-de-calidad-del-aire/r49-3614/es/`.

Gangoiti, G., Alonso, L., Navazo, M., Albizuri, A., Pérez-Landa, G., Matabuena, M., Valdenebro, V., Maruri, M., Antonio García, J. & Millán, M. M. (2002), 'Regional transport of pollutants over the Bay of Biscay: analysis of an ozone episode under a blocking anticyclone in west-central Europe', *Atm. Env.* **36**(8), 1349–1361.

Gramsch, E., Cereceda-Balic, F., Oyola, P. & Von Baer, D. (2006), 'Examination of pollution trends in Santiago de Chile with cluster analysis of PM10 and Ozone data', *Atm. Env.* **40**(28), 5464–5475.

Hannachi, A., Jolliffe, I. T. & Stephenson, D. B. (2007), 'Empirical orthogonal functions and related techniques in atmospheric science: A review', *Int. J. of Clim.* **27**(9), 1119–1152.

Henry, R. C. (1997), 'History and fundamentals of multivariate air quality receptor models', *Chem. and Intel. Lab. Syst.* **37**(1), 37–42.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *J. of Educational Psychology* **24**(6), 417–441.

Ibarra-Berastegi, G., Elías, A., Barona, A., Sáenz, J., Ezcurra, A. & Díaz de Argandoña, J. (2007), 'From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao', *Env. Mod. and Soft.* **23**(5), 622–637.

Ibarra-Berastegi, G., Sáenz, J., Ezcurra, A., Ganzedo, U., Díaz de Argandoña, J., Errasti, I., Férnandez-Ferrero, A. & **Polanco-Martínez, J.** (2009), 'Assessing spatial variability of SO_2 field as detected by an air quality network using Self-Organizing Maps, cluster, and Principal Component Analysis', *Atm. Env.* **43**(25), 3829–3836.

Jolliffe, I. T. (2002), *Principal component analysis*, Springer-Verlag, New York.

Kendall, S. M. (1980), *Multivariate analysis*, Charles Griffin, London.

Lau, J., Hung, W. T. & Cheung, C. S. (2009), 'Interpretation of air quality in relation to monitoring station's surroundings', *Atm. Env.* **43**(4), 769–777.

Lê, S., Josse, J. & Husson, F. (2008), 'FactoMineR: an R package for multivariate analysis', *J. of Stat. Soft.* **25**(1), 1–18.

Ligges, U. & Mächler, M. (2003), 'Scatterplot3d–an R package for Visualizing Multivariate Data', *J. of Stat. Soft.* **8**(11), 1–20.

Martínez-Ataz, E. M. & de Mera-Morales, Y. D. (2004), *Contaminación atmosférica*, Ed. Universidad de Castilla-La Mancha,.

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. B. & North, G. R. (2009), 'Empirical orthogonal functions: The medium is the message', *J. Clim.* **22**(24), 6501–6514.

Mudelsee, M. (2014), *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*, Springer.

Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R. & Chatterton, T. (2004), 'Modelling SO_2 concentration at a point with statistical approaches', *Env. Mod. and Soft.* **19**(10), 887–905.

Pearson, K. (1901), 'On lines and planes of closest fit to systems of points in space', *Phil. Mag.* **2**(11), 559–572.

Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. (2005), 'How many principal components? Stopping rules for determining the number of non-trivial axes revisited', *Comp. Stat. and Data Analysis* **49**(4), 974–997.

Pires, J. C. M., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2009), 'Identification of redundant air quality measurements through the use of principal component analysis', *Atm. Env.* **43**(25), 3837–3842.

Pires, J. C. M., Sousa, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008), 'Management of air quality monitoring using principal component and cluster analysis Part I: SO_2 and PM10', *Atm. Env.* **42**(6), 1249–1260.

Polanco-Martínez, J. (2012), Aplicación de técnicas estadísticas en el estudio de fenómenos ambientales y ecosistémicos, PhD thesis, University of Basque Country, España.
*https://addi.ehu.es/handle/10810/11295

Preisendorfer, R. W. (1988), *Principal components analysis in Metereology and Oceanography*, Elsevier, Amsterdam.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org

Seinfeld, J. H. (1978), *Contaminación atmosférica. Fundamentos físicos y químicos*, Inst. de Estudios de Adm. Local, Madrid.

Shrestha, S. & Kazama, F. (2007), 'Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan', *Env. Mod. and Soft.* **22**(4), 464–475.

Singh, K. P., Malik, A., Mohan, D. & Sinha, S. (2004), 'Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) a case study', *Water Res.* **38**(18), 3980–3992.

Sportisse, B. (2010), *Fundamentals in air pollution: from processes to modelling*, Springer, Heidelberg.

Von Storch, H. & Zwiers, F. W. (1999), *Statistical analysis in climate research*, Cambridge University Press, Cambridge, U.K.

Wark, K. & Warmer, C. F. (1994), *Contaminación del aire: origen y control*, Limusa, México.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, Academic Press, London.

World-Health-Organization, W. H. O. (2000), *Air Quality Guidelines for Europe*, number 91, WHO Reg. Pub. European series; No. 91.

Wunderlin, D. A., Diaz, M. P., Ame, M. V., Pesce, S. F., Hued, A. C., Bistoni, M. A. et al. (2001), 'Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina)', *Water Res.* **35**(12), 2881–2894.