UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Agosto 2016, Vol. 9, No. 2, pp. 271–294

El papel del análisis por componentes principales en la evaluación de redes de control de la calidad del aire[1]

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

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

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, hasta más recientemente (año 2006) en la evaluación de redes de control y vigilancia de la calidad del aire (RCVCA). 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 RCVCA. 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 RCVCA 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.

Abstract

[1]Polanco, J. M. El papel del análisis por componentes principales en la evaluación de redes de control de la calidad del aire. *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.

One of the most statistical techniques used in environmental sciences is the Principal Component Analysis (PCA). This technique consist in a linear decomposition of a set of correlated variables into a set of uncorrelated variables named principal components. It is one of the simplest and most robust ways of doing dimensionality reduction. 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). Due to the potential use of this method, more information in Spanish is required. For these reasons, we are highly motivated to contribute with this review 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.

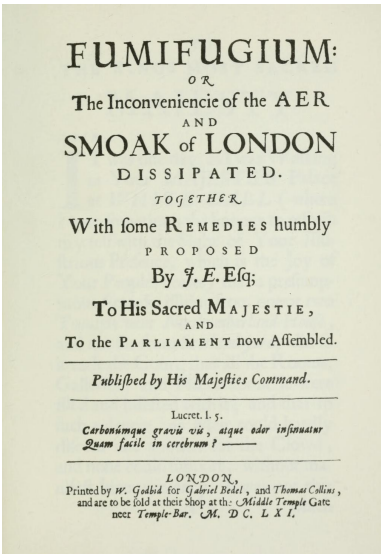
1. Introducción

El término contaminación atmosférica se refiere a la presencia en el aire de materias o formas de energía que impliquen riesgo, daño o molestias graves para los humanos y para los bienes materiales (Aránguez et al. 1999). Es importante tener en cuenta que la contaminación atmosférica de origen natural siempre ha existido debido a procesos biológicos, geológicos, químicos y físicos que generan partículas o gases contaminantes, como las erupciones volcánicas, incendios forestales, tormentas de arena, fermentaciones biológicas, etc. Con el descubrimiento del fuego por el hombre se origina la contaminación atmosférica antropogénica. Este tipo de contaminación ha ido adquiriendo importancia desde la Revolución Industrial y por el uso masivo de combustibles fósiles como fuentes de energía (Aránguez et al. 1999, Wark & Warmer 1994).

El campo de estudio de la contaminación atmosférica es muy amplio, incluye desde los estudios de los gases de efecto invernadero y su relación con el sistema climático, la destrucción de la capa de ozono debido a los clorofluorocarbonos, el impacto de liberaciones accidentales de contaminantes químicos, biológicos o radionucleidos a la atmósfera, hasta estudios de la calidad del aire (Seinfeld 1978, Sportisse 2010). Sin embargo, en este artículo estamos interesados en lo que respecta a la calidad del aire, y de manera particular, en los métodos estadísticos existentes para evaluar de manera objetiva la calidad del aire en una ciudad.

Uno de los primeros estudios descriptivos relacionados con la calidad del aire data de mediados del siglo XVII, el *Fumifugium*, publicado por Johan Evelyn en 1648 (Figura 1). El *Fumifugium* trata sobre el impacto del uso del carbón como un combustible en el medio ambiente de Londres y algunas medidas para combatir este tipo de contaminación. Posteriormente, en 1692, Robert Boyle realizó estudios pioneros sobre la composición química atmosférica. Con la llegada de la revolución industrial, el número de estudios relacionados con la contaminación del aire fue en aumento. En este periodo destacan los trabajos de Robert A. Smith en la

segunda mitad del siglo XIX sobre la lluvia ácida y por qué organizó una red de seguimiento de contaminantes atmosféricos, considerada como la precursora de las actuales redes de control y vigilancia de la calidad del aire (RCVCA)(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 th: *Middle Temple* Gate neer *Temple-Bar*. *M. DC. LXI.*

Figura 1: *Portada del* Fumifugium *o la inconveniencia del aire y el esmog diseminado de Londres y algunos remedios, por J. Evelyn (1648). Fuente: disponible en http://www.archive.org/details/fumifugium00eveluoft*

Con el incremento masivo del tráfico rodado y el rápido crecimiento industrial y poblacional en diferentes lugares alrededor del mundo entre principios y mediados del siglo XX, las emisiones antropogénicas de contaminantes a la atmósfera se dispararon considerablemente y con ello sus consecuencias (Martínez-Ataz & de Mera-Morales 2004, Seinfeld 1978). De modo especial hay que considerar el impacto en la salud humana, como es el caso de los incidentes mortales causados por episodios agudos de contaminación atmosférica, tales como la gran niebla de Londres en 1952 (Seinfeld 1978, Sportisse 2010).

Durante la década de los sesenta y setenta, los problemas de contaminación atmosférica fueron ocasionados principalmente por emisiones de SO_2 y por partículas suspendidas en el aire, emitidas principalmente por fuentes industriales y, en menor medida, por núcleos urbanos (Martínez-Ataz & de Mera-Morales 2004). Estos problemas de contaminación de origen antropogénico llamaron la atención tanto de la comunidad científica como de la sociedad en general, lo cual condujo a la regulación mediante políticas medioambientales (Henry 1997, Martínez-Ataz & de Mera-Morales 2004).

Como consecuencia de la regulación legal para el control de contaminantes at-

mosféricos, los niveles de emisiones de SO_2 y otros contaminantes del aire durante los últimos 25 años han ido disminuyendo en la mayoría de los países occidentales industrializados. En contraste, las emisiones de SO_2 están incrementando en países con economías emergentes (Nunnari et al. 2004). Otra de las consecuencias de las políticas medioambientales es que motivaron el desarrollo de métodos de captación (muestreo) y de análisis para medir tanto las emisiones como las inmisiones (Henry 1997, Martínez-Ataz & de Mera-Morales 2004).

Con lo que respecta a los métodos para captar y analizar la evolución de las inmisiones de diversos tipos de contaminantes y poder llevar a cabo un control y vigilancia de la calidad del aire, hoy en día se utilizan principalmente los analizadores automáticos (Martínez-Ataz & de Mera-Morales 2004, World-Health-Organization 2000). Estos analizadores tienen varias ventajas con respecto a otros métodos de seguimiento de contaminantes. Por ejemplo, tienen una muy alta resolución temporal (se pueden obtener datos cada hora o a menor resolución), pueden ser instalados en cualquier lugar adecuado para ello, etc. Sin embargo, su coste no es bajo, tienden a ser más susceptibles a problemas técnicos cuando no se cuenta con el mantenimiento adecuado y con personal técnico cualificado y requieren de una constante evaluación (Martínez-Ataz & de Mera-Morales 2004).

Al conjunto de analizadores automáticos (nodos) que miden inmisiones de contaminantes y que forman una red de muestreo, se les conoce como una red de control y vigilancia de la calidad del aire (RCVCA) (o en Inglés, *Air Quality Monitoring Network – AQMN*) (Martínez-Ataz & de Mera-Morales 2004). Estas redes permiten hacer un estudio y seguimiento adecuado de la calidad del aire. Sin embargo, requieren de una constante evaluación para averiguar y garantizar que cada uno de sus nodos proporcionen una caracterización adecuada de la calidad del aire en la zona donde muestrea cada sensor (Pires et al. 2008). Las constantes evaluaciones ayudan a determinar el número adecuado de nodos de la red de tal modo que no se obtenga información redundante, a detectar fallos en alguno de los nodos o a detectar una inadecuada localización espacial de los nodos (Lau et al. 2009, Pires et al. 2009).

Una de las herramientas útiles para evaluar objetivamente las RCVCA son las técnicas estadísticas multivariantes, como el análisis por componentes principales (ACP) o técnicas matemáticas de clasificación como el análisis de conglomerados. Estas técnicas han sido utilizadas para la evaluación y el manejo adecuado de redes de seguimiento de la calidad del agua (Shrestha & Kazama 2007, Singh et al. 2004, Wunderlin et al. 2001). Sin embargo, el uso combinado de ambas técnicas en la evaluación de una RCVCA ha sido recientemente llevado a cabo en el año 2006 (Gramsch et al. 2006) para determinar la tendencia estacional y la distribución espacial de PM_{10} y O_3. Posteriormente, el ACP ha sido utilizado con mucha frecuencia para evaluar RCVCA en diferentes regiones del planeta (Ibarra-Berastegi et al. 2007, Ibarra-Berastegi et al. 2009, Pires et al. 2009, Pires et al. 2008).

Debido a la necesidad de evaluar cuantitativamente las RCVCA mediante la aplicación de técnicas estadísticas multivariantes, como el ACP, es necesario contar

con información oportuna sobre esta temática. En lo que respecta a información en lengua inglesa exista una cierta cantidad de publicaciones (véase, por ejemplo, Gramsch et al. (2006), Ibarra-Berastegi et al. (2007), Ibarra-Berastegi et al. (2009), Lau et al. (2009), Pires et al. (2009) y Pires et al. (2008)). Sin embargo, no sucede lo mismo en idioma español, con excepciones como la de Polanco-Martínez (Polanco-Martínez 2012). La carencia de esta información en lenguaje español ha sido una de las principales motivaciones para la escritura de este artículo.

El objetivo de este artículo es proporcionar una breve revisión de los fundamentos estadísticos del análisis por componentes principales y su aplicación en la evaluación de redes de control y vigilancia de la calidad del Aire. Se proporciona también un caso de estudio con datos simulados ("sintéticos") y otro con datos reales de una RCVCA localizada en la ciudad de Bilbao para el periodo 2006-2010. La estructura del artículo es la siguiente. En la sección 2 se presentan los fundamentos matemáticos del ACP. La sección 3 proporciona información para interpretar el ACP. En la sección 4 se proporcionan diferentes reglas para determinar el número de componentes principales a retener. La sección 5 presenta los casos de estudio. Por último, en la sección 6 se presentan las conclusiones.

2. Análisis de componentes principales

El análisis de componentes principales (ACP) es una de las técnicas estadísticas multivariantes más populares y antiguas en el análisis de datos. Esta técnica fue desarrollada por Karl Pearson en 1901[1] (Pearson 1901), pero no fue sino hasta 1939 cuando Hotelling hizo una presentación mucho más formal y acuñó el término de *componente principal* (CP) (Abdi & Williams 2010, Hotelling 1933). El ACP también recibe otros nombres, dependiendo de su campo de aplicación, *v. gr.*, en la teoría de procesos estocásticos se conoce como la expansión o transformada de Karhunen-Loéve (Monahan et al. 2009), en turbulencia como descomposición ortogonal propia (Berkooz et al. 1993), en ciencias sociales y económicas como vectores principales (Kendall 1980), en ciencias atmosféricas como funciones empíricas ortogonales (Von Storch & Zwiers 1999, Wilks 1995).

Antes de presentar una definición formal y describir sus principales características matemáticas, podemos mencionar que el ACP es un tipo de transformación lineal aplicada a un conjunto de datos multivariantes habitualmente correlacionados entre sí, para convertirlos en un menor número de variables no correlacionadas y ortogonales[2] entre sí, esto es, expresar la información contenida en un conjunto de datos, con un número menor de variables (Jolliffe 2002, Wilks 1995). Los principales objetivos del análisis de componentes principales, de acuerdo con Abdi & Williams (2010) y Jolliffe (2002), es extraer la información más importante de un conjunto de datos multivariantes, comprimir un conjunto de datos multivariantes

[1] Aunque algunas investigaciones establecen que sus orígenes se remontan hasta Cauchy (en 1829) y Jordan (en 1874) (Abdi & Williams 2010).

[2] En el seno del ACP, el concepto de ortogonalidad de series temporales (las CPs lo son) corresponden al concepto de series incorrelacionadas.

manteniendo solo la información que se considere importante (reducir la dimensionalidad de los datos), simplificar la descripción de un conjunto de datos y analizar la estructura de las observaciones y de las variables.

2.1. Notación matemática y conceptos preliminares

Antes de exponer la parte metodológica del ACP es importante mantener una nomenclatura adecuada y consistente. La presentación metodológica está basada y es análoga a la de Abdi & Williams (2010) y Hannachi et al. (2007). Para denotar matrices, vectores y elementos, se usarán mayúsculas en negritas, minúsculas en negritas y minúsculas en itálicas, respectivamente. Matrices, vectores y elementos de la misma matriz usarán la misma letra, *v. gr.*, $\mathbf{X}, \mathbf{x}, x$. La transpuesta de una matriz se representará con el superíndice ^{T}. La matriz identidad estará denotada por $\mathbf{I}$. El vector columna de unos y de longitud I viene dado por $\mathbf{1}_{(I\times 1)}$.

Los datos que van a ser analizados mediante ACP contienen I observaciones (muestras)[3] y J variables. El número de observaciones es mayor que el número de variables en los casos de estudio (Sección 5) en este artículo de revisión, aunque esto no es necesario. Cada observación (muestra) es obtenida en los tiempos $t_i, i = 1, 2, \cdots, I$ y están representadas por la matriz $\mathbf{D}_{(I\times J)}$ (filas por columnas), donde un ij-ésimo elemento viene dado por $d_{i,j}$. Antes de aplicar el ACP a los datos que provienen de las RCVCA se requiere de un tipo de procesamiento muy simple centrado en la media, esto es debido a que por regla general los datos que miden estas redes están en la misma escala. Aunque es importante corroborar si los datos analizados tienen la misma escala y, si este no fuera el caso, es necesario tipificar (o aplicar alguna forma de estandarización) las variables en estudio. El procesamiento consiste en sustraer la media de las observaciones a cada variable y trabajar con las "anomalías" $\mathbf{X}_{(I\times J)}$ (Hannachi et al. 2007, Wilks 1995), esto es:

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

t donde $\overline{\mathbf{D}}$ (vector de medias muestrales) viene dado por:

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

Si sustituimos el tercer miembro de la relación 2 en la relación 1 y factorizando, las anomalías también pueden representarse, de acuerdo con Hannachi et al. (2007), del siguiente modo:

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

Donde $\mathbf{M}_{(I\times I)}$ es la matriz de centrado de orden I. De aquí en adelante no utilizaremos (a menos que sea necesario) los subíndices dimensionales I, J para simplificar la notación matemática.

[3]Se utiliza el término observación solo por cuestiones prácticas, los datos bajo análisis vía ACP no se limitan a datos observacionales.

2.2. Sobre cómo calcular las componentes principales

La forma más usual como se presenta el cálculo de las componentes principales (CP) en los textos de análisis de datos ambientales (específicamente datos climáticos o meteorológicos) es mediante la solución de un problema de autovalores a tráves de la matriz de covarianzas de las anomalías $\mathbf{X}$ de los datos $\mathbf{D}$ bajo estudio (Hannachi et al. 2007, Von Storch & Zwiers 1999, Wilks 1995). Este procedimiento puede expresarse matemáticamente y de acuerdo con Hannachi et al. (2007) de la siguiente forma. La matriz de covarianzas muestral de la matriz de anomalías $\mathbf{X}$ (relación 3) está definida (Hannachi et al. 2007) por la relación

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

Donde cada elemento de $\mathbf{S}$ está formado por las covarianzas entre cada par de variables de $\mathbf{X}$ de dimensiones $I \times J$.

Ahora bien, el objetivo del análisis por componentes principales es encontrar un nuevo de conjunto de variables (combinaciones lineales) no correlacionadas entre sí que expliquen la máxima varianza. Esto equivale a encontrar un vector unitario $\mathbf{q} = (q_1, ..., q_J)^T$ tal que $\mathbf{Xq}$ tenga la máxima variabilidad (Hannachi et al. 2007, Von Storch & Zwiers 1999). Esto es

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

sujeta a la condición $\mathbf{q}^T\mathbf{q} = 1$.

Los autovectores o funciones empíricas ortogonales (*Empirical Orthogonal Functions* – EOFs) se obtienen, de acuerdo con Hannachi et al. (2007), como una solución al problema de autovalores

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

donde los autovalores $\lambda_l, l = 1, 2, ..., N$ con $N = min(I, J)$, vienen dados por:

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

Los autovalores λ_l proporcionan una medida de la varianza de $\mathbf{X}$ en la dirección de $\mathbf{q}_l$. Una vez que se resuelve el problema de autovalores (relación 6), por regla general, estos son ordenados de modo decreciente (Hannachi et al. 2007), esto es, $\lambda_1 \geq \lambda_2 ... \geq \lambda_N$. Una forma común de expresar el porcentaje de varianza correspondiente a cada autovalor es por medio de la relación

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

Las l-ésimas componentes principales están dadas por la proyección de $\mathbf{X}$ sobre el l-ésimo autovector $\mathbf{q}_l = (q_{1l}, q_{2l}, ..., q_{Jl})^T$ (Hannachi et al. 2007) y se expresan

mediante la siguiente relación:

$$\mathbf{p}_l = \mathbf{X}\mathbf{q}_l \tag{9}$$

Los elementos $(p_{tl}, t = 1, ..., I)$ de la relación 9 pueden expresarse como:

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

De tal modo que el l-ésimo autovalor λ_l representa la varianza de la l-ésima componente principal $\mathbf{p}_l = (p_{l1}, p_{l2}, ..., p_{lI})^T$ (Hannachi et al. 2007).

Sin embargo, para calcular las componentes principales, en la práctica, no se calcula la matriz de covarianza muestral (relación 4) ni se resuelve el problema de autovalores (relación 7), sino que se calculan por medio de la descomposición de $\mathbf{X}$ por valores singulares (*Singular Value Decomposition* – SVD) (Abdi & Williams 2010, Hannachi et al. 2007). El hecho de utilizar la SVD se debe fundamentalmente a cuestiones de eficiencia computacional. Por otro lado, la SVD tienen la ventaja frente al problema de autovalores debido a la posibilidad de operar sobre matrices no cuadradas, así como calcular y presentar de manera secuencial los l primeros autovectores de manera ordenada sin tener que calcular todos los autovectores de la matriz de covarianzas muestral $\mathbf{S}$.

Este procedimiento se describe a continuación (Abdi & Williams 2010, Hannachi et al. 2007). La matriz $\mathbf{X}_{(I\times J)}$ se descompone por valores singulares de la siguiente manera:

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

donde las matrices $\mathbf{P}$ (conocida como la matriz de las componentes principales) y $\mathbf{Q}$ (conocida también como matriz de cargas o de proyección) son tales que sus columnas $\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p}_r$ y $\mathbf{q_1}, \mathbf{q_2}, ..., \mathbf{q}_r$, son ortogonales y se llaman vectores singulares izquierdo y derecho, respectivamente. El rango de $\mathbf{X}$ es r y $\mathbf{\Sigma} = Diag(\sigma_1, \sigma_2, ..., \sigma_r)$ es una matriz diagonal cuyos elementos $\sigma_1 \geq \sigma_2 ... \geq \sigma_r \geq 0$ son los valores singulares de $\mathbf{X}$ (Abdi & Williams 2010, Hannachi et al. 2007). Ahora bien, debido a la propiedad de ortogonalidad de los autovectores (constituyen una base), la relación 11 se puede expandir (descomponer) como una combinación lineal (Hannachi et al. 2007, Wilks 1995), esto es:

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

También es posible expresar la matriz de covarianzas $\mathbf{S}$ en términos de la SVD (Abdi & Williams 2010, Hannachi et al. 2007). Si se sustituye la relación 11 en 4 se obtiene:

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

t donde $\mathbf{\Sigma}^2 = Diag(\sigma_1^2, \sigma_2^2, ..., \sigma_r^2)$ y los valores singulares están ordenados de acuerdo a $\sigma_1^2 \geq \sigma_2^2 ... \geq \sigma_r^2$. Los autovalores se relacionan con los valores singulares mediante $\lambda_l = \frac{\sigma_l^2}{I},\ l = 1, ..., r$.

Para terminar esta subsección es importante considerar un par de cuestiones. Desde un punto de vista computacional, hay que tener en cuenta que tanto $\mathbf{q}_j$ (columna j-ésima de $\mathbf{Q}$) como -$\mathbf{q}_j$ son soluciones adecuadas de la SVD de la matriz $\mathbf{X}$. Si el paquete computacional proporciona como solución $\mathbf{q}_j$ o -$\mathbf{q}_j$, el signo de la columna j-ésima de $\mathbf{P}$ aparecerá cambiado. En ocasiones se realiza un paso ulterior, que consiste en la rotación de la matriz de proyección $\mathbf{Q}$ mediante una matriz $\mathbf{R}$ de rotación, esto es, $\mathbf{Q}' = \mathbf{R}^T\mathbf{Q}$ (Abdi & Williams 2010, Hannachi et al. 2007). La matriz de rotación $\mathbf{R}$ puede ser ortogonal (rotación ortogonal) o no (rotación oblicua). Uno de los objetivos principales de la rotación de las componentes principales es suavizar la condición de ortogonalidad, para con ello tener estructuras más localizadas en el espacio de más fácil interpretación (Hannachi et al. 2007, Jolliffe 2002). Uno de los métodos más utilizados para realizar la rotación es el Varimax, el cual fue desarrollado por Kaiser en 1958, aunque existen otros métodos de rotación, como el quartimax o promax (Jolliffe 2002).

En los casos de estudio de este artículo, la rotación no ha ayudado en la interpretación de resultados (Sección 5), por lo que no se aplica.

3. Interpretación de las componentes principales

En esta sección se presenta de manera concisa tres subsecciones con información (*v.gr.*, la terminología) para una adecuada interpretación de los resultados obtenidos al aplicar el ACP.

3.1. Contribución de una observación a una componente principal

La importancia de una observación para una componente principal puede ser obtenida por la razón del cuadrado de la componente principal correspondiente a esta observación entre el autovalor asociado con esta componente. Esta razón es conocida como la *contribución* de la i-ésima observación a la l-ésima componente (Abdi & Williams 2010), es denotada por $ctr_{i,l}$, y se expresa como

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

t donde $p_{i,l}$ es la l-ésima componente principal y λ_l es su autovalor asociado. $ctr_{i,l}$ puede tomar valores entre 0 y 1 y para una determinada l-ésima componente la suma de las contribuciones de todas las observaciones es igual a 1. Una sugerencia útil, es basar la interpretación de una componente en las observaciones cuya con-

tribución es mucho mayor que el promedio de la contribución, *i. e.*, observaciones cuya contribución sea mayor que $1/I$ (donde I es el número de observaciones).

3.2. Contribución de una componente principal a una observación

La importancia de una componente para una observación dada puede ser estimada por medio del *coseno cuadrado*, e indica la contribución de una componente a la distancia al cuadrado de la observación al origen. Esto corresponde al cuadrado del coseno del ángulo del triángulo rectángulo formado con el origen, la observación y su proyección en la componente principal donde suma de los cosenos cuadrados es igual a 1, y es calculado de este modo

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

t donde $d_{i,\mathbf{g}}^2$ es el cuadrado de la distancia de una observación dada al origen. En otras palabras, $d_{i,\mathbf{g}}^2$ es calculado como la suma de los cuadrados de todas las componentes principales para esta observación. Componentes con grandes valores de $\gamma_{i,l}^2$ contribuyen a una buena parte de la distancia total, por tanto, esas componentes tienen importancia para esa observación (Abdi & Williams 2010, Jolliffe 2002).

3.3. Correlación de una componente y una variable; factores de carga.

La correlación entre una componente principal y una variable es conocida en la jerga del ACP como *factores de carga* o *coeficientes espaciales* (*factor loadings*). Nótese que la suma de los cuadrados de los coeficientes de correlación entre una variable y todas las componentes es igual a 1 (Abdi & Williams 2010). Aunque el uso de los factores de carga al cuadrado parecen simplificar la interpretación, el hecho de estar al cuadrado implica perdida de información. Esto es, en muchas ocasiones una componente aparece para separar dos o más grupos, cuyos elementos suelen tener puntuaciones elevadas de signo opuesto, siendo su contribución mucho mayor que el resto de las observaciones. A veces las componentes pueden ser elementos muy útiles para detectar atípicos multivariantes sobre todo en casos en los que lo "raro" no son los valores individuales de las variables sino su aparición conjunta. Por ello, es recomendable analizar ambos casos, esto es, los factores de carga y estos elevados al cuadrado.

Especial cuidado hay que tener cuando se usa el término *loading* porque tiene varias interpretaciones y puede ser una fuente potencial de confusión. Por esta razón es importante corroborar el significado de *loading* ya sea en la lectura de textos sobre la técnica ACP o en las salidas del programa computacional empleado para calcular ACP (Abdi & Williams 2010, Wilks 1995). En este sentido, es bastante útil la

tabla 9.3 del Wilks (1995), donde se proporciona una variedad de terminología y sus sinónimos para el ACP.

4. Sobre el número de componentes principales

Una de las principales aplicaciones del análisis de componentes principales es reducir la dimensionalidad de las variables de un conjunto de datos; por tanto, un punto problemático es determinar cuántas componentes principales han de ser retenidas (téngase en cuenta que el número máximo de componentes principales que pueden ser retenidas es igual al número de variables) (Dray 2008, Wilks 1995). La decisión del número de componentes puede llevar a una pérdida de información (subestimación) o introducir ruido aleatorio (sobreajuste) (Dray 2008). A pesar de que la técnica ACP tiene ya más de un siglo de existencia, este problema permanece abierto hasta hoy en día. Sin embargo, existen algunas reglas conocidas como *stopping rules* que pueden servir de ayuda para determinar el número de componentes principales a retener (Jolliffe 2002, Wilks 1995).

Recientemente Peres-Neto et al. (2005) hacen una recopilación de 20 reglas objetivas que dividen en dos grandes familias. Las reglas basadas en intervalos de confianza (*v. gr.*, el análisis paralelo, métodos de *bootstrap* basados en autovalores o el test de Bartlett) y las reglas basadas en valores promedios de *tests* estadísticos (*v. gr.*, la regla de Kaiser-Guttman, el modelo *broken-stick* o de la mínima correlación parcial promedio). Al mismo tiempo Peres-Neto et al. (2005) hacen un análisis comparativo entre estas reglas. Ellos encontraron al utilizar datos simulados que, más que la *stopping rule* utilizada o el número de elementos y el grado de Gaussianidad de los datos analizados, estas reglas son mucho más dependientes de las correlaciones existentes entre las observaciones o entre las variables (Peres-Neto et al. 2005). Estos resultados (dependencia de la correlación) son parecidos a los obtenidos en estudios anteriores pero con datos reales (datos climáticos) (Preisendorfer 1988, Wilks 1995).

Con lo que respecta al uso de las *stopping rules* en estudios similares a los casos de estudio presentados en este artículo de revisión (Sección 5), los dos criterios frecuentemente utilizados son el criterio de Kaiser (Guttman) y el criterio del porcentaje de varianza acumulada (PVA) (Lau et al. 2009, Pires et al. 2008). El criterio de Kaiser solo considera retener las CPs con autovalores mayores de uno (Jolliffe 2002, Wilks 1995). El problema con este criterio es que puede ser muy restrictivo (aun teniendo en cuenta la sugerencia de Jolliffe (2002), esto es, retener las CPs cuyos autovalores sean iguales o mayores que 0,7) (Lau et al. 2009, Pires et al. 2008). El criterio PVA considera retener las CPs cuyos porcentajes de varianza acumulados superen un determinado valor. Por ejemplo, algunos autores (Pires et al. 2009, Pires et al. 2008) recomiendan tener en cuenta solo las CPs que superen al menos el 90 % de la varianza acumulada (PVA_{90}). Este criterio (PVA_{90}) se utilizó en los casos de estudio de este artículo (Sección 5). Sin embargo, téngase en cuenta que estas reglas proporcionan una buena indicación para saber el número

de CP a retener y, cada caso de estudio debería analizarse con toda la información (por ejemplo, los factores de carga) proporcionada por el ACP.

5. Casos de estudio

En esta sección se presentan dos casos de estudio para ejemplificar el uso del ACP en la evaluación de RCVCA. En primer lugar se presenta un caso de estudio con datos sintéticos, esto es, datos creados cuyas características estadísticas conocemos *a priori*. En segundo lugar, se presenta un ejemplo con datos reales de valores de inmisiones de SO_2 de una RCVCA ubicada en la zona metropolitana de Bilbao, España. Para el análisis por componentes principales se utilizó el paquete computacional `FactoMineR` (Lê et al. 2008) en R (R Development Core Team 2009), mientras que para representar gráficamente en 3D el espacio de las componentes principales se utilizó el paquete `R scatterplot3d` (Ligges & Mächler 2003).

5.1. Datos sintéticos

La construcción de las series sintéticas para este caso de estudio se realiza de la siguiente manera. Se construyen tres parejas de procesos autorregresivos bivariados de orden 1 (AR1), de tal modo que los miembros de cada pareja (X, Y), tengan una cierta correlación – $\text{CORR}[X(t), Y(t)] = \rho_{XY}$ – entre ellos. Un proceso AR1 bi-variado está definido (Mudelsee 2014) de la siguiente manera:

$$\begin{aligned}
X(1) &= \mu^X_{N(0,1)}(1), \\
Y(1) &= \mu^Y_{N(0,1)}(1), \\
X(t) &= \rho_X X(t-1) + \mu^X_{N(0,1-\rho_X^2)}(t), \quad t = 2, ..., T, \\
Y(t) &= \rho_Y Y(t-1) + \mu^Y_{N(0,1-\rho_Y^2)}(t), \quad t = 2, ..., T, \qquad (16) \\
CORR[\mu^X_{N(0,1)}(1), \mu^Y_{N(0,1)}(1)] &= \rho_\mu \\
CORR[\mu^X_{N(0,1)}(t), \mu^Y_{N(0,1)}(t)] &= \frac{1-\rho_X\rho_Y}{[(1-\rho_X^2)(1-\rho_Y^2)]^{1/2}}\rho_\mu, \\
& \qquad t = 2, ..., T \\
CORR[\mu^X_{N(0,1)}(t), \mu^Y_{N(0,1)}(u)] &= 0, \quad t, u = 1, ..., T,\ t \neq u \qquad (17)
\end{aligned}$$

Donde μ_X y μ_Y son dos procesos de tipo ruido blanco Gaussiano, ρ_X y ρ_Y son los parámetros de los procesos autoregresivos X e Y, respectivamente. El modelo AR1 bi-variado (ecuación 16) requiere se estrictamente estacionario, por lo cual tiene que cumplir (Mudelsee 2014): $E[X(t)] = E[Y(t)] = 0$, $VAR[X(t)] = VAR[Y(t)] = 1$ y $CORR[X(t), Y(t)] = \rho_{XY} = \rho_\mu$.

En el primer ejemplo, las correlaciones de las parejas de los procesos bi-variados AR1 tienen los valores (arbitrarios, aunque se han elegido correlaciones con altos valores, tanto positivos como negativos y valores medios) de 0.95, 0.50 y -0.75,

mientras que los parámetros de los AR1 tienen los valores de (0.92, 0.93), (0.45, 0.47) y (-0.72, -0.73) para cada pareja de procesos bi-variados AR1, respectivamente. La longitud de las series T es igual a 2191 elementos, y es la misma para las 6 series sintéticas. Se ha utilizado 2191 elementos debido a que es el número de días en un intervalo entre el 1-1-2005 y 31-12-2010 ("fechas arbitrarias"). Por otro lado, se han calculado los promedios mensuales (figura 2) de las 3 parejas de procesos bi-variados AR1 generados mediante las relaciones 16 y 17. Esta es una práctica habitual en el pre-proceso de datos medidos en las RCVCA para evitar posibles efectos de enmascaramiento debido a la posible existencia de ciclos de alta frecuencia (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012). Todo esto, en este caso de estudio, se realiza con el objeto de construir los datos simulados (sintéticos) lo más cercanamente posible a datos reales. Por esta misma razón y para mantener la terminología utilizada en el análisis de RCVCA, utilizaremos la palabra "sensores" para referirnos a las series sintéticas.

El resultado de aplicar el ACP a los datos simulados (figura 2) se presenta en la tabla 1 y en la figura 3. Cómo se puede apreciar en la Tabla 1 a) las primeras 3 componentes principales explican casi el 90 % de la varianza acumulada (de hecho explican un 86,35 %), por lo cual las componentes principales de la 4 a la 6 explican sólo un pequeño porcentaje de varianza (de manera particular la CP-6 que explica tan sólo un 0,95 % de varianza). Esta información proporciona una idea *grosso modo* de que sólo 3 componentes principales –y por ende 3 "sensores"– son necesarias para explicar la mayor parte de la varianza de los datos analizados. Sin embargo, la información clave para determinar el número óptimo de "sensores", la proporcionan los factores de carga (tabla 1 b). Cómo se puede apreciar, en las primeras dos parejas (1-2 y 3-4) de "sensores", los factores de carga tienen valores muy cercanos. Lo cual puede interpretarse que tanto los "sensores" 1 y 2, así cómo el 3 y 4, miden la misma información, por lo cual son redundantes. Por otro lado, los factores de carga de los "sensores" 5 y 6 son parecidos numéricamente, pero de signos opuestos, lo cual podría indicar que no miden información redundante. Sin embargo, si lo hacen, solo que uno lo mide de manera directa y el otro de forma inversa. Esta información puede corroborarse visualmente en la figura 3. En la cuál se puede observar que los "sensores" en el espacio de las componentes principales tienden a formar grupos. Por un lado, los sensores 1 y 2, por otro lado los sensores 3 y 4 y los sensores 5 y 6, aunque distanciados de manera inversa. Por todo esto, se puede confirmar la hipótesis de que sólo 3 "sensores" miden información no redundante.

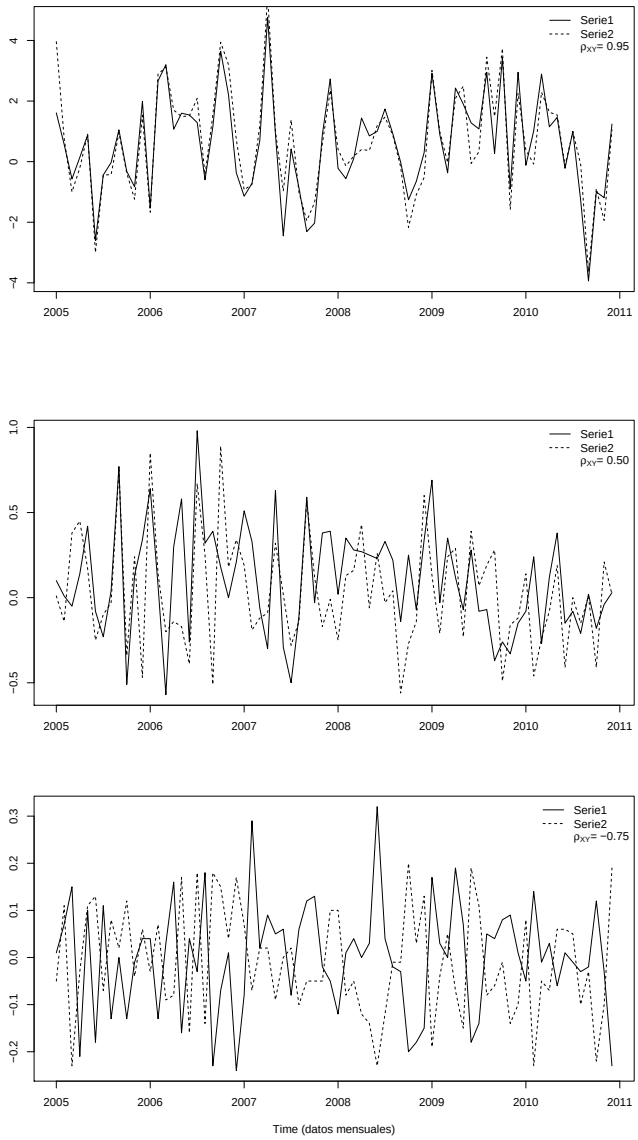


Figura 2: *Series temporales sintéticas ("sensores") (72 valores mensuales). Las correlaciones para cada pareja de procesos bi-variados AR1 son 0,95 (arriba), 0,50 (centro) y -0,75 (abajo). Los parámetros de las parejas de los procesos AR1 son 0,92 (línea continua) y 0,93 (línea discontinua) (arriba), 0,45 (línea continua) y 0,47 (línea discontinua) (centro), -0,72 (línea continua) y -0,73 (línea discontinua) (abajo). Fuente: elaboración propia.*

Tabla 1: *Resumen del análisis por componentes principales. (a) Valores propios y porcentaje de la varianza total explicada por cada CP y (b) factores de carga. Fuente: elaboración propia.*

a)

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

b)

"Sensor"	CP-1	CP-2	CP-3	CP-4	CP-5	CP-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

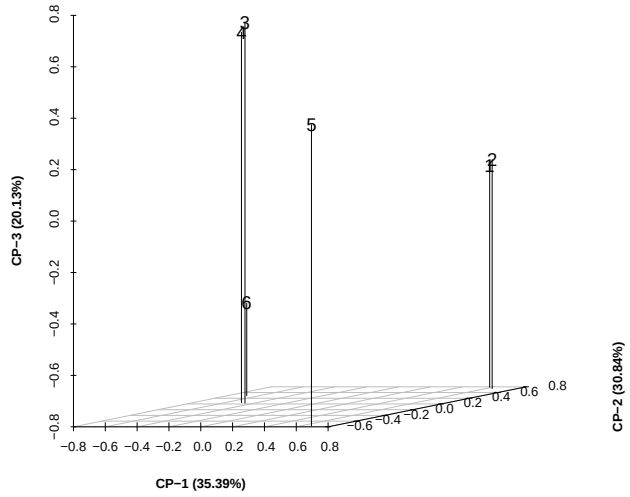


Figura 3: *Representación gráfica para las 6 series temporales sintéticas ("sensores") en el espacio de las componentes principales. Fuente: elaboración propia.*

5.2. Datos reales

El caso de estudio con datos reales pertenece a una RCVCA ubicada en el área metropolitana de Bilbao (Comunidad Autónoma del País Vasco – CAPV, España). Esta zona geográfica pertenece a una de las principales regiones con fuentes de contaminantes atmosféricos (debido principalmente a la industria y al tráfico rodado) no sólo de la CAPV sino a nivel nacional (Gangoiti et al. 2002, Ibarra-Berastegi et al. 2007, Polanco-Martínez 2012). Al igual que otras ciudades industrializadas de finales de la década de los setenta y principios de los ochenta, Bilbao metropolitano presentó problemas de contaminación atmosférica de origen industrial. Por esta razón se instaló una red de control y vigilancia de la calidad del aire, la primera en España y una de las primeras en Europa (*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). Uno de los primeros objetivos de esta red fue dar seguimiento a emisiones de origen industrial de SO_2. No obstante, en los últimos años los niveles de SO_2 han ido disminuyendo debido a los cambios en el tejido industrial y a los combustibles actualmente utilizados. Sin embargo, la actual legislación ambiental obliga a seguir realizando mediciones de inmisiones de SO_2 y otros contaminantes (Ibarra-Berastegi et al. 2009). Hoy en día, las estaciones que cubren el área urbana y la zona metropolitana de Bilbao forman parte de una red mayor que cubre toda la CAPV con un total de 51 estaciones (*Estaciones remotas de la red de vigilancia de la calidad del aire. Departamento de Medio Ambiente y Política Territorial, Gobierno Vasco* 2016).

Desde la instalación de la RCVCA de Bilbao metropolitano hasta la fecha, la estructura de la red ha sufrido modificaciones importantes (por ejemplo, en la localización de las estaciones o en el aumento de estas) con el objeto de poder captar las campos representativos de las inmisiones de los diversos contaminantes, debido a los cambios de localización o al aumento o disminución de las fuentes de emisiones tanto de SO_2 como de otros contaminantes (Albizuri 2008). Este tipo de cambios no se limitan a la RCVCA de Bilbao metropolitano, sino que prácticamente están presentes en la mayoría de las redes de control y vigilancia de la calidad del aire de otras áreas geográficas. Es por estas razones, tanto el proceso de diseño como el control y la evaluación de una RCVCA es un proceso dinámico e interactivo (Ibarra-Berastegi et al. 2009). Por tanto, es importante mantener una constante evaluación de estas redes para poder captar las trayectorias representativas de las especies de contaminantes(Lau et al. 2009, Pires et al. 2009, Pires et al. 2008).

Estudios previos (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012) demostraron que los cuatros sensores de la RCVCA del área metropolitana de Bilbao no miden información redundante de inmisiones de SO_2 para el periodo 1996-2001. Por tanto, todos esos sensores son necesarios para una correcta evaluación de las inmisiones de SO_2 para esa área de estudio. En este caso de estudio de este artículo de revisión, se presenta una evaluación de los sensores que miden inmisiones de SO_2 y que están localizados en Bilbao metropolitano pero para el periodo 2006-2010 (figura 4). Esto, con el objeto de corroborar si estos cuatro sensores siguen siendo necesarios para una adecuada medición de las inmisiones de SO_2.

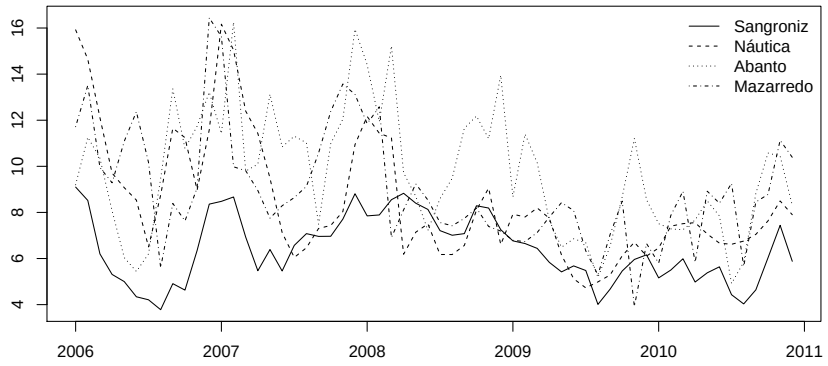


Figura 4: *Series temporales (60 valores mensuales) de inmisiones de SO_2 para las cuatro estaciones. Sangroniz, Náutica, Abanto y Mazarredo. Las unidades de las emisiones de SO_2 están en $\mu g\, m^{-3}$. Fuente: elaboración propia.*

El resultado del análisis por componentes principales para las series mensuales de inmisiones de SO_2 se presenta en la tabla 2. Cómo se puede apreciar (tabla 2 a), las tres primeras componentes principales explican un alto porcentaje de varianza (93.53 %), lo cual podría indicar que solo 3 sensores son necesarios para una adecuada caracterización de las inmisiones de SO_2 para Bilbao metropolitano. Sin embargo, de acuerdo con los factores de carga (tabla 2 b), estos solo tienen valores parecidos en los factores de carga de la primera componente principal[4]. Pero los factores de carga para las otras componentes difieren, tanto en signo como en valor. Esto es, la primera CP puede interpretarse como el resultado de la mezcla de emisiones de SO_2 que atraviesan Bilbao metropolitano. Es decir, representa la contribución de emisiones de SO_2 de diferentes fuentes que se mezclan y se distribuyen uniformemente en el área de estudio, medidas por los cuatro sensores instalados en esta zona. La influencia de la primera CP a los sensores (2 c) varía de uno a otro relativamente poco, con valores que van de 52,52 (Mazarredo) hasta 67,58 (Náutica). Por otro lado, la contribución de cada sensor a la primera CP (2 d) también varia poco, con valores que oscilan entre 21,13 (Mazarredo) y 27,19 (Náutica).

La segunda y tercera CP se puede interpretar físicamente a partir de patrones locales de emisiones o dispersiones de SO_2 (tabla 2 c y d) pero de modo opuesto, pues sus factores de carga tienen signos opuestos (parejas Sangroniz-Náutica y Abanto y Mazarredo, tabla 2 b). Téngase en cuenta que, por un lado, para el caso de los sensores Abanto y Mazarredo, estos se encuentran ubicados en zonas que

[4]Lo cual indica que la primera CP representa una variabilidad común de los valores de inmisiones de SO_2 para Bilbao metropolitano.

presentan diferentes condiciones meteorológicas u orográficas. Abanto se encuentra ubicada en la periferia de Bilbao metropolitano en una zona con fuerte presencia de grandes infraestructuras industriales, mientras que Mazarredo se encuentra en el centro de la ciudad de Bilbao. Por otro lado, Sangroniz se encuentra localizada en las periferias de Bilbao metropolitano donde no hay grandes núcleos de población o infraestructuras industriales, mientras que Náutica se encuentra localiza cerca del mar.

Por otro lado, de acuerdo con la información proporcionada por el espacio de las componentes principales (figura 5), es posible apreciar que los sensores distan mucho entre sí y no se agrupan. Sin embargo, es notable la manera en la que los sensores en Sangroniz y Mazarredo muestran una relación inversa. Lo que sugiere que esta pareja de sensores podrían estar midiendo información redundante, aunque de manera inversa. Esto puede explicarse debido a que Mazarredo se encuentra en el centro de la ciudad de Bilbao, por lo cual este sensor mide inmisiones de contaminantes directamente relacionados con el tráfico rodado. Mientras que Sangroniz se encuentra en la periferia de Bilbao metropolitano en un área poco poblada y sin grandes infraestructuras industriales en los alrededores, por lo cual las inmisiones que se están midiendo en Sangroniz no son de manera directa, las cuales podrían esta influenciadas por factores topográficos y condiciones meteorológicas.

Una información importante que hay que tener en cuenta es que el sensor instalado en Sangroniz está midiendo inmisiones de SO_2 con menor intensidad que los otros 3 sensores. Esto puede observarse en la figura 4, la cual muestra que el sensor en Sangroniz es el que mide los menores valores de inmisiones de SO_2. Teniendo en cuenta estos análisis, podemos establecer que no es relevante tener un sensor en Sangroniz. De hecho, actualmente el Departamento de Medio Ambiente y Política Territorial del Gobierno Vasco ya no hace mediciones de valores de inmisiones de SO_2 en Sangroniz, aunque sí de otras especies de contaminantes. Por todo lo anterior, se puede establecer que no todos los cuatro sensores son necesarios para una adecuada caracterización de valores de inmisiones de SO_2 para Bilbao metropolitano, por lo cual se debería remover el sensor que mide inmisiones de SO_2 en Sangroniz. Estos resultados no parecen concordar con estudios previos en la misma zona y con prácticamente los mismo sensores (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012). Sin embargo, hay que tener en cuenta que en este trabajo se enfoca en el intervalo temporal 2006–2011, el cual es mucho más reciente que el intervalo de estudio analizado por (Ibarra-Berastegi et al. 2009, Polanco-Martínez 2012), el cual tuvo lugar entre 1996–2002. Por tanto, no se esperan que los resultados de ambos estudios sean similares y más aun debido a que las zonas altamentemente influenciadas por la componente antropogénica (y las áreas metropolitanas de las grandes ciudades lo están) están en continuo cambio.

Con lo que respecta a la cuarta CP, se puede apreciar que los factores de carga (tabla 2 b) oscilan entre valores (correlaciones) relativamente bajos de -0,25 a 0,30, al igual que las contribuciones de las componentes principales a la variabilidad de cada sensor (tabla 2 c), los cuales presentan valores entre 5,21 y 8,87. A pesar de que las contribuciones de cada sensor a la variabilidad total de cada componente

principal (tabla 2 d) presentan unas contribuciones con valores (entre 20,13 y 34,27) similares a la CP1. Por ello, la interpretación de la cuarta CP no es evidente.

Tabla 2: *Resumen del análisis de componentes principales. (a) Valores propios y porcentaje de la varianza total explicada por cada CP y (b) factores de carga. Fuente: elaboración propia.*

a)

Componente principal	Val. propios	% de var.	% de var. acum
CP-1	2,48	62,15	62,15
CP-2	0,77	19,31	81,46
CP-3	0,48	12,07	93,53
CP-4	0,26	6,47	100,00

b)

Sensor	CP-1	CP-2	CP-3	CP-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	CP-1	CP-2	CP-3	CP-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	CP-1	CP-2	CP-3	CP-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

6. Conclusiones

El uso del análisis por componentes principales para la evaluación de redes de control y vigilancia de la calidad del aire es a día de hoy una de las herramientas estadísticas más utilizadas. Sin embargo, la mayor parte de la información disponible está en inglés. Debido a las potenciales usos, no solo académicos, sino también en casos prácticos (*v. gr.*, protección civil, secretarias o agencias de medio ambiente, etc.) es necesario contar con esta información en español; razón por la cual, se presenta este trabajo de revisión. Por un lado, en este trabajo se presentan

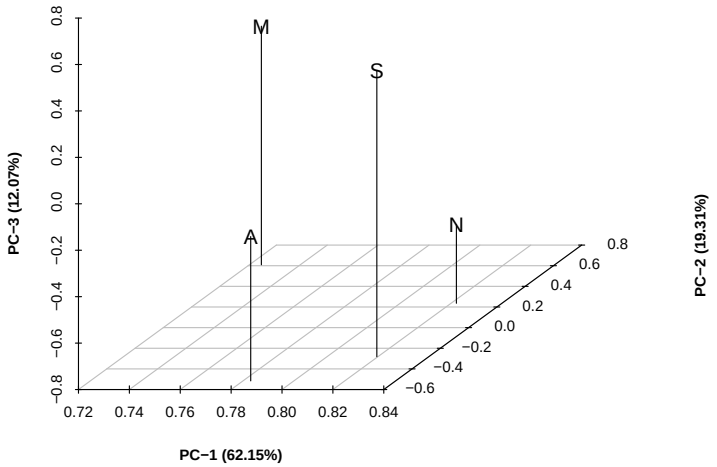


Figura 5: *Representación gráfica para los cuatro sensores el espacio de las componentes principales. Abanto (A), Mazarredo (M), Naútica (N) y Sangroniz (S). Fuente: elaboración propia.*

las bases estadísticas del ACP de tal modo que sea accesible a un amplio sector de disciplinas científicas (aunque con un fuerte enfoque a las ciencias ambientales) pero sin perder rigurosidad matemática. Por otro lado, se presentan los puntos clave para una adecuada interpretación del ACP al evaluar RCVCA.

En este trabajo también se presentaron dos casos de estudio para ejemplificar de forma práctica la evaluación de RCVCA mediante ACP. En el primer ejemplo, se utilizaron datos simulados con conocimientos *a priori* de que presentaban información redundante, lo cual se verificó al aplicar ACP. Mientras que en el segundo ejemplo se analizaron valores mensuales de inmisiones de SO_2 provenientes de una RCVCA ubicada en Bilbao metropolitano para el periodo 2006-2010. Con base en los resultados obtenidos se concluye que solo tres de los cuatros sensores de la RCVCA del área metropolitana de Bilbao son necesarios. Por tanto, se recomienda remover el sensor que mide inmisiones de SO_2 que se encuentra localizado en Sangroniz.

AGRADECIMIENTOS

Se agradece a los Prof. Dr. J. Sáenz y Dr. G. Ibarra, quienes me iniciaron en el mundo del análisis de redes de control y vigilancia de la calidad del aire mediante técnicas estadísticas multivariantes. Al revisor de este trabajo por sus atinadas sugerencias para mejorar enormemente este artículo. Se agradece la financiación de la ayuda post-doctoral del Gobierno Vasco (Ref. No. `POS_2015_1_0006`).

Recibido: 29 de febrero del 2016
Aceptado: 29 de abril del 2016

Referencias

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.