

IMPROVEMENT OF THE DIMENSION OF AN AIR QUALITY MONITORING NETWORK BY MEANS OF MULTIVARIATE STATISTICAL METHODS

Marta Doval Miñarro¹, Jose A. Egea² y Ginesa Navarro Cobacho¹

¹UNIVERSIDAD POLITÉCNICA DE CARTAGENA. Dpto. de Ingeniería Química e Industrial. Paseo Alfonso XIII, 52 30203 Cartagena (Murcia). Tfno: +34 968 325552. marta.doval@upct.es

²CENTRO DE EDAFOLOGÍA Y BIOLOGÍA APLICADA DEL SEGURA (CEBAS-CSIC). Campus Universitario de Espinardo, 30100, Murcia. Tfno: +34 968 396200. jaegea@cebas.csic.es

<http://dx.doi.org/10.6036/9250>

MEJORA DEL DIMENSIONAMIENTO DE UNA RED DE CALIDAD DEL AIRE MEDIANTE EL EMPLEO DE MÉTODOS ESTADÍSTICOS MULTIVARIANTES

ABSTRACT:

Two multivariate statistical methods (principal component analysis and cluster analysis) have been used in this work to assess potential redundancies in stations and measurements of the air quality monitoring network of the Región de Murcia (Spain). The results show that there are no redundancies, not even in areas with more stations than required by the legislation. This highlights that current criteria for assigning stations to a particular area (based on population numbers) must be complemented with criteria that consider, among others, its industrial or touristic activity. The described methodology is applicable to other existing networks.

Keywords: monitoring networks, redundancies, principal component analysis, cluster analysis

RESUMEN:

En este trabajo se emplean dos métodos estadísticos multivariantes (análisis de componentes principales y análisis cluster) para obtener información sobre posibles redundancias en las estaciones y en las medidas de la calidad del aire de la red de vigilancia ambiental de la Región de Murcia (España). Los resultados muestran que no existen redundancias en la red, incluso en aquellas zonas en las que hay más estaciones que las requeridas por la legislación. Esto pone de manifiesto que los criterios de asignación de estaciones de vigilancia actuales, basados en el número de habitantes de una determinada zona o aglomeración, deben complementarse con criterios de otro tipo que consideren, entre otros, la actividad industrial o turística de la misma. La metodología expuesta es aplicable a otras redes de vigilancia existentes.

Palabras clave: redes de vigilancia atmosférica, redundancias, análisis de componentes principales, análisis cluster

1. INTRODUCTION

Directive 2008/50/EC of 21st May 2008 on ambient air quality and a cleaner air for Europe establishes in its Annex V the minimum number of fixed measurements to monitor and control the air quality in a zone or agglomeration as a function of its population.

Multivariate statistical techniques can be defined as a group of statistical methods that aim at analysing simultaneously data sets where there are multiple variables for each individual or object studied. This is the case of the data collected by air quality monitoring stations: in each station, different pollutants are measured and measurements are obtained every few seconds or minutes. Principal component analysis (PCA) is probably the most used multivariate statistical tool in the field of environmental pollution. Particularly, in air quality, it has been used many times to determine the origin of the pollution [1]–[5] and to optimise the number and location of air monitoring stations [6]–[13]. However, it has been detected that in most cases, it is applied to pollutants individually. As already mentioned, the usual practice is to measure several pollutants in each monitoring station so that pollutants share networks. In this regards, in order to know if a network is oversized, it is important to carry out the statistical analysis with all the pollutants measured at the same time. Only two studies have been found that proceed in this way [14],

[15] and, in none of them, a comparison between the results obtained and the requirement of Directive 2008/50/EC has been carried out.

This work is aimed at showing the usefulness of two multivariate statistical methods (PCA and Cluster analysis (CA)) to detect both redundancies in the stations that take part in an air quality monitoring network and to establish similar behaviour patterns of regulated pollutants. The methodology has been applied to the Air Quality Monitoring Network of the Región de Murcia (AQMNRM), but it is easily applied to any other network. Moreover, the suitability of Directive 2008/50/EC criteria to assign the minimum number of monitoring stations in an agglomeration or zone is also discussed based on the results obtained by PCA.

2. TOOLS Y METHODS

2.1. STUDY AREA

The Región de Murcia is divided into 6 homogeneous zones for air quality management purposes (Figure 1), regarding their geographical characteristics, the environmental and human activities taking place and their main type of pollution [16]. The zones *Norte*, *Centro* and *Litoral – Mar Menor* have a great ecological value with an important farming, livestock and medium-size industrial activity. *Murcia Ciudad* area is typically urban, whereas *Valle de Escombreras* is an industrial area. *Cartagena* is an area both urban and industrial. Currently, there are 8 fixed monitoring stations and 2 mobile stations distributed throughout the 6 zones. All the zones have one fixed monitoring station except for *Valle de Escombreras* and *Murcia Ciudad* that have two of them. In this study, only the fixed monitoring stations will be considered.

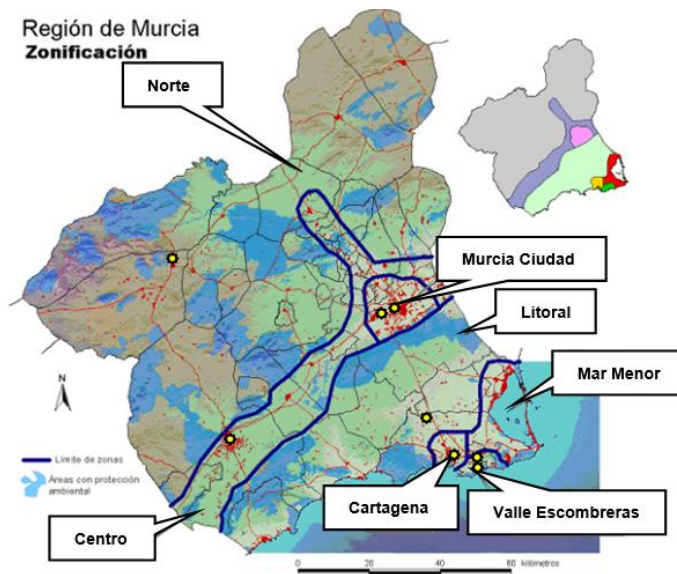



Fig. 1: Zones and agglomerations of the AQMNRM [16]. Yellow dots represent the fixed air monitoring stations. Norte area (Caravaca), Centro area (Lorca), Murcia Ciudad area (San Basilio and Alcantarilla), Litoral-Mar Menor area (Aljorra), Cartagena area (Mompeán) and Valle de Escombreras area (Escombreras and Alumbres).

	IMPROVEMENT OF THE DIMENSION OF AN AIR QUALITY MONITORING NETWORK BY MEANS OF MULTIVARIATE STATISTICAL METHODS	ENVIRONMENTAL TECHNOLOGY AND ENGINEERING
COLLABORATION	Marta Doval Miñarro, Jose A. Egea, Ginesa Navarro Cobacho	Air pollution control

2.2. DATA ANALYSIS AND STATISTICS

2.2.1. Data collection

Data used in this study are publicly available on the website of air quality of *Comunidad Autónoma de la Región de Murcia* (<https://sinclair.carm.es/calidadaire/>).

2.2.2. R programming language

All the calculations performed in this work have been done using the R software [17].

2.2.3. Principal component analysis (PCA)

The main idea of PCA is to reduce the dimensionality of a data set which contains a number of interrelated variables while maximizing the variability of the principal components (PC's). These PC's are not correlated and some of them retain most of the variability present in the original variables [18]. PCA aims to represent the original data in a new orthogonal space of functions that represent the principal modes of variability of the system. The modes of variation are the empirical orthogonal functions (EFOs) and they are related to the eigenvectors of the correlation or covariance matrix of the original data. The projections of the original data on these functions are called principal components, whose main characteristic is the lack of correlation among them [19], [20].

There are different criteria to select the optimal number of principal components. In this work the Kaiser and the percentage of explained variability (which should be higher than 0,8 or 0,9) criteria are used.

2.2.4. Cluster Analysis (CA)

Cluster analysis (CA) aims at grouping elements in homogeneous groups based on a similarity measure. CA methods can be divided into hierarchical and non-hierarchical methods. Among the hierarchical ones, the most used is the Ward method, which is the one used in this study. This popular method starts with an only member in a group and continues clustering in groups of two elements in each step until only one group is formed after $n-1$ steps. The criterion to choose the pair of clustered groups in each step is as follows: among all the possible combinations to cluster groups in pairs, the pair chosen is the one minimizing the squared sum of the distances between the elements and their respective groups. The results of a CA are usually shown through a tree plot or a dendrogram.

2.3. METHODOLOGY

PCA and CA have been applied to determine common emission patterns of regulated atmospheric pollutants and to study potential information redundancies in the monitoring stations. All the stations of the AQMNRM with data of nitric oxide (NO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃) and particulate matter < 10 μm (PM₁₀) for the years 2010-2015 were considered. This is the time period with data of all pollutants in all stations, with the following exceptions:

1. *Escombreras* station, where O₃ is not measured. To avoid removing this station from the analysis, a second study was performed focused on the area of *Campo de Cartagena*, that comprises the stations of *Aljorra*, *Alumbres*, *Mompeán* and *Valle de Escombreras*.
2. *Caravaca* station (*Norte* area), where SO₂ is not measured. This station has not been considered in any of the two studies carried out.

In Table 1 the two sets of data used are shown.

	STUDY 1: Pollutants: NO, NO ₂ , SO ₂ , O ₃ y PM ₁₀	STUDY 2 – <i>Campo de Cartagena</i> : Pollutants: NO, NO ₂ , SO ₂ , y PM ₁₀
Stations	Aljorra	Aljorra
	Alumbres	Alumbres
	Mompeán	Mompeán
	Alcantarilla	Escombreras
	San Basilio	
	Lorca	

Table 1. Data sets used to evaluate atmospheric pollutant emission patterns and the potential redundancy of monitoring stations.

For each station and pollutant, daily average concentrations in the studied years were downloaded. Subsequently, annual mean concentrations were obtained and, finally, a global mean for the whole period for each pollutant and station was calculated. The global means are considered to be representative of the time period studied due to the low values of standard deviation obtained (shown in the next section).

3. RESULTS

3.1. GLOBAL MEANS OF POLLUTANTS CONCENTRATION

The global means for each pollutant concentration and each station are presented in Table 2

Station	NO	NO ₂	SO ₂	O ₃	PM ₁₀
Aljorra	5,37 (1,04)	13,40 (3,23)	6,52 (0,91)	62,24 (9,66)	28,83 (1,65)
Alumbres	4,52 (1,24)	15,67 (2,78)	12,38 (3,08)	66,95 (4,95)	23,25 (0,94)
Mompeán	10,48 (1,32)	26,18 (3,82)	9,29 (1,87)	54,03 (7,75)	24,86 (1,93)
San Basilio	20,10 (2,82)	39,34 (4,97)	5,97 (1,63)	50,66 (3,98)	31,85 (3,56)
Alcantarilla	11,63 (3,12)	26,19 (3,47)	4,69 (1,68)	59,41 (2,01)	24,24 (1,67)
Lorca	3,66 (1,21)	12,97 (2,24)	7,62 (1,57)	76,02 (9,74)	25,82 (2,67)

(a)

Station	NO	NO ₂	SO ₂	PM ₁₀
Aljorra	5,37 (1,04)	13,40 (3,23)	6,52 (0,91)	28,83 (1,65)
Alumbres	4,52 (1,24)	15,67 (2,78)	12,38 (3,08)	23,25 (0,94)
Mompeán	10,48 (1,32)	26,18 (3,82)	9,29 (1,97)	24,86 (1,93)
Escombreras	8,72 (2,29)	21,32 (4,20)	14,49 (2,10)	24,68 (2,32)

(b)

Table 2. Global means ($\mu\text{g}/\text{m}^3$) of pollutant concentrations of stations in Study 1 (a) and stations in Study 2 (b). Standard deviations in brackets.

3.2. PCA RESULTS

3.2.1. Study 1

The PCA analysis performed with data of Study 1 revealed that two principal components explained at least 87% of the original data variability. Besides, the Kaiser criterion (standard deviation equal or greater than 1) would confirm the selection of two components (Table S1). Figure 2 shows the principal components 1 and 2 in terms of pollutants (Figure 2a) and stations (Figure 2b).

PC1 relates nitrogen oxides (NO and NO₂) and compares them with O₃ and, to a lesser extent, with SO₂ and PM₁₀. Thus, locations on the right side of Figure 2b (e.g., *San Basilio*) show the highest concentrations of NO and NO₂ compared with the rest, whereas locations on the left side of Figure 2b (e.g., *Lorca* and *Alumbres*) show the highest concentrations of O₃.

Regarding principal component 2, its analysis is not so straightforward because it combines variables PM₁₀, SO₂ and O₃. Locations at the top of Figure 2b (e.g., *Aljorra* and *Lorca*) show the following combination: high values of PM₁₀, low values of SO₂ and high values of O₃, whereas locations at the bottom of Figure 2b (e.g., *Alumbres*, *Mompeán*) show high levels of SO₂, low levels of PM₁₀ and relatively high levels of O₃. Observing Figure 2 it is inferred that *Alumbres* and *Aljorra* are very opposed stations. The same applies to *Lorca* and *San Basilio*.

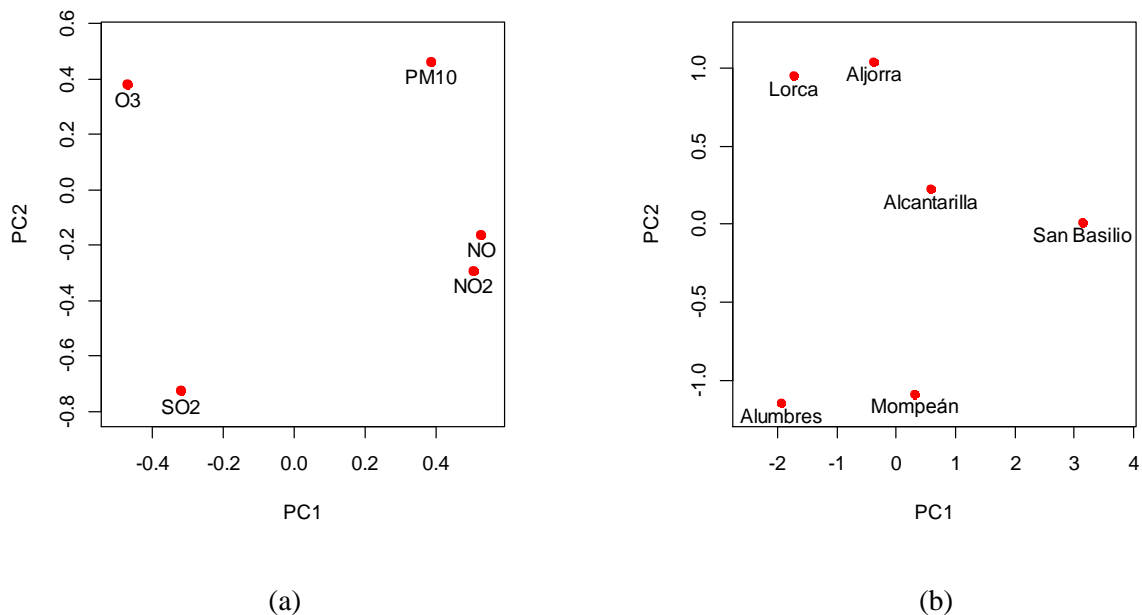


Fig. 2: PCA plots for the pollutants (a) and stations (b), for Study 1.

3.2.2. Study 2

Table S2 shows that the accumulated variance is higher than 0,9 for two components. The Kaiser criterion confirms that two components are enough to explain most of the information contained in the original data. Figure 3 shows the principal components 1 and 2 in terms of pollutants (Figure 3a) and stations (Figure 3b).

PC1 compares NO₂ with PM₁₀, thus, locations on the right side of Figure 3b (e.g., *Mompeán*) present high values of NO₂ and low values of PM₁₀, whereas locations on the left side of Figure 3b (e.g., *Aljorra*) present high values of PM₁₀ and low values of NO₂.

PC2 compares NO with SO₂, thus, locations at the top of Figure 3b (e.g., *Mompeán*, *Aljorra*) show high values of NO and low values of SO₂, whereas locations at the bottom of Figure 3b (e.g., *Alumbres*) show low values of NO and high values of SO₂.

Analysing Figures 3a and 3b it can be inferred that *Mompeán* and *Aljorra* are very opposed locations. *Alumbres* and *Valle de Escombreras* present higher concentrations of SO₂ while *Aljorra* presents very high values of PM₁₀. Further, *Mompeán* presents high values of NO and NO₂.

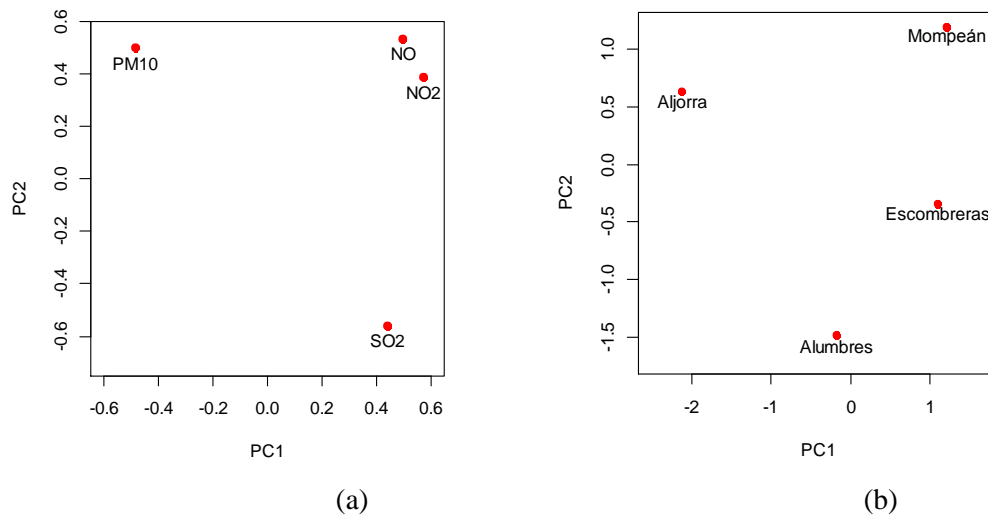


Fig. 3: PCA plots for the pollutants (a) and stations (b), for Study 2.

3.3. CA RESULTS

3.3.1. Study 1

Dendrogram of Figure 4a shows two or three differentiated groups, depending on the rescaled distance considered.

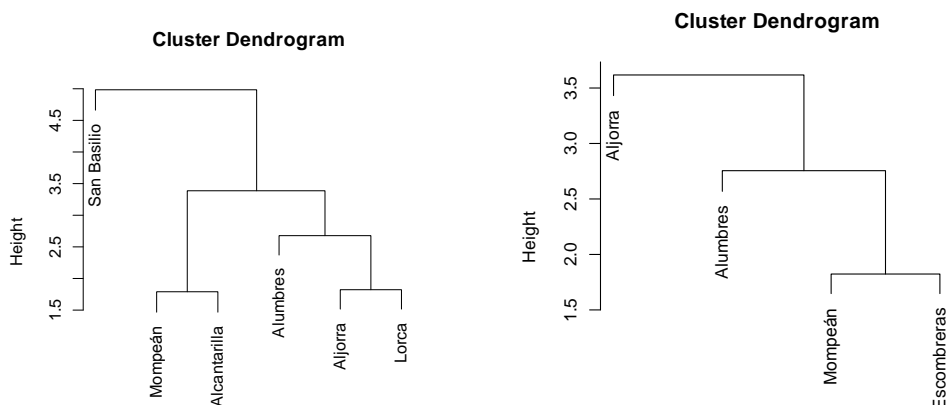



Fig. 4: Dendrogram for study 1 (a) and study 2 (b).

The dendrogram corresponds to PC1 of the PCA analysis in previous sections, where *San Basilio* forms an only group, specially due to its high ratio NO_x/O₃ compared with the rest of locations. As we move towards the right side of the dendrogram this ratio decreases.

	IMPROVEMENT OF THE DIMENSION OF AN AIR QUALITY MONITORING NETWORK BY MEANS OF MULTIVARIATE STATISTICAL METHODS	ENVIRONMENTAL TECHNOLOGY AND ENGINEERING
COLLABORATION	Marta Doval Miñarro, Jose A. Egea, Ginesa Navarro Cobacho	Air pollution control

3.3.2. Study 2

Dendrogram of Figure 4b represents the data for study 2. The number of selected clusters in this case is again between two and three. As it can be observed, *Escombreras* and *Mompeán* are in one side of the dendrogram while *Aljorra* is in the opposite. *Alumbres* would stay in the middle of both groups although in the same branch as *Mompeán* and *Escombreras*.

4. DISCUSSION

As it can be seen in Figures 2b and 3b, all the stations of the AQMNRM and the ones in *Campo de Cartagena*, according to the first two principal components, are relatively spread. There are no groups of stations too close in any case. This shows that, despite being some stations geographically close, these do not detect the same concentrations of pollutants, so it can be concluded that there are no redundancies in the network or an excessive number of stations in certain areas. The number of stations considered in the Study 1 (7) coincides with the minimum number of stations according to Directive 2008/50/EC as a function of the number of inhabitants of each area. However, it is worth remembering that *Caravaca* has not been taken into account in the analysis. Thus, the AQMNRM has one more station than the required by law. A more complete analysis would require data of SO₂ in *Caravaca* to check if this station is redundant with other stations in the network. In any case, the population criterion to determine the minimum number of stations in an area should be complemented with others that consider, among others, the existence of an important industrial network or touristic activity. In this regard, in the area of *Valle de Escombreras*, according to the legislation only one monitoring station is needed; however, the PCA has revealed that the measurements in *Alumbres* and *Escombreras* are not redundant.

For both studies (1 and 2), it is shown that the pollutants measured have different sources, with the exception of NO and NO₂, that appear relatively close in the PCA plots. These two species come primarily from road traffic and, although NO is emitted in higher proportion than NO₂, the former is rapidly oxidized to NO₂ in contact with ambient air. Regarding SO₂, although part of it comes from traffic, it can also be emitted in industrial activities involving combustion of fossil fuels. This is the case of *Valle de Escombreras* area, where apart from many chemical industries, there is one of the biggest petrol refineries of the country, and it has the highest levels of SO₂ in our Region. PM₁₀ is partially emitted by road traffic but it has also a natural origin, as the typical intrusions of Saharan dust that often affect the Region. Lastly, O₃ is a secondary pollutant with formation dynamics different from the rest of pollutants measured. As it was seen in the Study 1, the PC1 shows that O₃ concentrations are opposed to the nitrogen oxides ones, which is reasonable as the former generates from the latter.

The CA has revealed that the stations of Study 1 can be classified into three big groups; one of them with the only station of *San Basilio*; another one with *Mompeán* and *Alcantarilla*; and a third one with *Alumbres*, *Aljorra* and *Lorca*, according to the ratio NO₂ / O₃. Regarding the Study 2, among 2 and 3 groups are observed, in this case according to the ratio NO₂ / PM₁₀. On the one hand, there is *Aljorra* station; on the other hand, *Mompeán* and *Escombreras*; and, in between them, *Alumbres*. Although the stations that form each group have similar characteristics, the PCA has revealed that in no case there are redundancies.

4. CONCLUSIONS

It has been shown after applying PCA and CA that there are no redundancies in the information provided by the studied stations, so that all of them are necessary to completely characterize the air quality of the *Región de Murcia*. This fact highlights that it is important to consider other criteria such as an important industrial presence to assign the number of fixed stations in a zone, as it is the case of *Valle de Escombreras* area where, according to its population, only one station is required by law. Thus, it can be concluded that, in occasions, the number of monitoring stations specified in the legislation can be insufficient to completely characterize the air quality of an area.

Moreover, the PCA has revealed that, with the exception of NO and NO₂, that share the same emission sources, the origin of the rest of pollutants measured is diverse, so there are no groups of pollutants in this regard. Therefore, it can be concluded that, apart from being compulsory, it is necessary to measure all of the regulated pollutants in the *Región de Murcia*.

It is worth mentioning that it would be very interesting to carry out further studies where potential redundancies were investigated as a function of the season of the year, so that meteorological variables, that may influence the results, were taken into account.

REFERENCES

- [1] Gratani L, Crescente MF, Varone L. "Long-term monitoring of metal pollution by urban trees". *Atmos. Environ.* 2008. Vol.42-35 p.8273–8277. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2008.07.032>
- [2] Juda-Rezler K., Reizer M., Oudinet JP. "Determination and analysis of PM10 source apportionment during episodes of air pollution in Central Eastern European urban areas: The case of wintertime 2006". *Atmos. Environ.* 2011. Vol.45-36 p.6557–6566. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2011.08.020>
- [3] Slezakova K., Pires JCM, Castro D. *et al.* "PAH air pollution at a Portuguese urban area: Carcinogenic risks and sources identification". *Environ. Sci. Pollut. Res.* 2013. Vol.20-6 p.3932–3945. DOI: <http://dx.doi.org/10.1007/s11356-012-1300-7>
- [4] Vellingiri K., Kim KH, Jeon JY, *et al.* "Changes in NO_x and O₃ concentrations over a decade at a central urban area of Seoul, Korea". *Atmos. Environ.* 2015. Vol.112 p.116–125. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2015.04.032>
- [5] Fang GC, Chiang HC, Hsu CY, *et al.* "Characteristics of Concentrations and Metal Compositions for PM2.5 and PM2.5–10 in Yunlin County, Taiwan during Air Quality Deterioration". *Aerosol Air Qual. Res.* 2015. Vol.15-7 p.2571–2583. DOI: 10.4209/aaqr.2015.04.0261
- [6] Pires JCM, Sousa SIV, Pereira MC, *et al.* "Management of air quality monitoring using principal component and cluster analysis-Part I: SO₂ and PM₁₀". *Atmos. Environ.* 2008. Vol.42-5 p.1249–1260. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2007.10.044>
- [7] Pires JCM, Sousa SIV, Pereira MC, *et al.* "Management of air quality monitoring using principal component and cluster analysis-Part II: CO, NO₂ and O₃". *Atmos. Environ.* 2008. Vol.42-5 p.1261–1274. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2007.10.041>
- [8] Pires JCM, Pereira MC, Alvim-Ferraz MCM, *et al.* "Identification of redundant air quality measurements through the use of principal component analysis". *Atmos. Environ.* 2009. Vol.43-25 p.3837–3842. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2009.05.013>
- [9] Lu WZ, Di He H, Yun Dong L. "Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis". *Build. Environ.* 2011. Vol.46-3 p.577–583. DOI: <http://dx.doi.org/10.1016/j.buildenv.2010.09.004>
- [10] Zhao L, Xie Y, Wang J, *et al.* "A performance assessment and adjustment program for air quality monitoring networks in Shanghai". *Atmos. Environ.* 2015. Vol.122 p.382–392. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2015.09.069>
- [11] Wang C, Zhao L, Sun W, *et al.* "Identifying redundant monitoring stations in an air quality monitoring network". *Atmos. Environ.* 2018. Vol.190 p.256–268. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2018.07.040>
- [12] Caggiano R, Di Leo S, D'Emilio M, *et al.* "Statistical tools for data optimization in air quality monitoring networks". *Fresenius Environ. Bull.* 2007. Vol.16 p.364–371.
- [13] Ibarra-Berastegi G, Sáenz J, Ezcurra A, *et al.* "Assessing spatial variability of SO₂ field as detected by an air quality network using Self-Organizing Maps, cluster, and Principal Component Analysis". *Atmos. Environ.* 2009. Vol.43 p.3829–3836. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2009.05.010>
- [14] Dominick D, Juahir H, Latif MT, *et al.* "Spatial assessment of air quality patterns in Malaysia using multivariate analysis". *Atmos. Environ.* 2012. Vol.60 p.172–181. DOI: <http://dx.doi.org/10.1016/j.atmosenv.2012.06.021>
- [15] Gómez-Losada A, Lozano-García A, Pino-Mejías R, *et al.* "Finite mixture models to characterize and refine air quality monitoring networks". *Sci. Total Environ.* 2014. Vol.485–486 p.292–299. DOI: 10.1016/j.scitotenv.2014.03.091
- [16] CARM. *La calidad del aire en la Comunidad Autónoma de la Región de Murcia. Informe anual 2014*. Disponible en Web: <https://sinqlair.carm.es/calidadaire/documentos/documentacion/Informeanual_2014.pdf>
- [17] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2014. Disponible en Web: <<http://www.R-project.org/>>
- [18] I. T. Jolliffe, *Principal component analysis*. 2ª edición. Springer Series in Statistics. Nueva York: Springer, 1986. 271p. ISBN: 978-1-4757-1906-2.
- [19] R. B. Cattell. "The Scree Test For The Number Of Factors". *Multivariate Behav. Res.* 1966. Vol.1-2 p.245–276. DOI: http://dx.doi.org/10.1207/s15327906mbr0102_10
- [20] S. D. Wilks. *Statistical Methods In The Atmospheric Sciences*. 3ª edición. International Geophysics Series. Vol.59. Burlington: Elsevier Academic Press, 2006. 704p. ISBN: 978-0-1238-5022-5