International Journal of Geosciences, 2015, 6, 1154-1165
Published Online October 2015 in SciRes. http://www.scirp.org/journal/ijg
http://dx.doi.org/10.4236/ijg.2015.610090

Scientific Research Publishing

Correspondence Analysis on a Space-Time Data Set for Multiple Environmental Variables

Palma Monica

Universitá del Salento, Lecce, Italy
Email: monica.palma@unisalento.it

Received 3 August 2015; accepted 26 October 2015; published 29 October 2015

Abstract

Applications of the multivariate technique called correspondence analysis for environmental studies are relatively new and are limited to spatial multivariate data set. In this paper, a procedure of applying correspondence analysis to a large space-time data set for multiple environmental variables is shown. In particular, nitrogen dioxide and carbon monoxide hourly concentrations measured during January 1999 at several monitored stations in a district of Northern Italy are analyzed. The procedure consists in transforming the continuous variables into categorical ones by the means of appropriate indicator variables, generating special contingency tables and applying correspondence analysis. The use of this classical multivariate technique allows the identification of important relationships among pollution levels and monitoring stations and/or relationships among pollution levels and observation times.

Keywords

Space-Time Data, Indicator Transform, Correspondence Analysis

1. Introduction

Usually environmental monitoring networks collect a huge amount of data such as pollutant concentrations, atmospheric variates, weather conditions, and so on, which are of particular interest for public policies oriented to environmental and human health protection.

Such data sets may have the following features:

- they are multivariate, as several variables are simultaneously measured;
- they present a spatio-temporal structure, since the measurements are taken in several point of the study area

and for a certain period of time.

Classical multivariate techniques represent useful tools for analyzing multiple va-riables. Their main goal is to obtain a summary description of the data: Principal Component Analysis (PCA) finds a smaller number of variates representing all those collected, without loss of essential information; Correspondence Analysis (CA) studies the association between two or more categorical variables by representing the categories of the variables as points in a low-dimensional space; Canonical Correlation Analysis (CCA) describes the relationships between two groups of several variables. Classical multivariate techniques can be also applied to space-time data sets in order to summarize the spatial and temporal profiles which characterize the information, finding relationships among the data. In De Iaco *et al.* [1], the use of PCA allowed summarizing a very large data set of space-time observations for three contaminants. The authors identified a single measure of total air pollution which synthesized the original data without loss of information. Moreover, lately a space-time data set for air pollution and atmospheric variables has been analyzed through CCA in De Iaco [2]. The author emphasized the features of that multiva-riate technique which allowed describing very important relationships between three contaminants (nitrix oxide, nitrogen dioxide and ozone) and atmospheric indicators (humidity, temperature and wind speed).

Hence, when multiple variables are measured at several locations of the area under study and for a period of time, in other words, when a space-time multivariate data set is available, and the aim is studying the simultaneous behaviour of the va-riables in order to understand the relationships among the space-time observations, a multivariate technique is the most useful tool. CA is one of the multivariate techniques with a wide range of applications in several fields such as social and political sciences, marketing research, economy, ecology and biology. This technique is usually applied as an exploratory method, with the aim to describe the structure of the data under study with minimal constraints on the form of the same structure [3].

In this paper, it will be shown that even CA can be applied to a space-time multivariate data set, finding very important results which other techniques may not highlight. In particular, in this paper CA will be applied to an air pollution data set involving two contaminants measured at monitoring stations in northern Italy during January 1999. The analysis will identify relationships in space among pollution levels and monitoring stations and relationships in time among pollution levels and observation times.

After a presentation of CA (Section 2) and a review of its theory (Section 2.1), the description of computational aspects follows (Section 2.2). Then, the data set (Section 3) and the most important results from the applied CA and their interpretation are given (Section 4).

2. Correspondence Analysis

CA is an algebraic technique analogous to PCA, but, while PCA is used for tables of continuous measurements, CA is more appropriate for categorical variates. Hence, CA is suitable for analyzing qualitative information represented by a contingency table. Lebart *et al.* [4] suggest that CA is useful for the analysis of large data matrices, particularly when there is little auxiliary information concerning the data. The original development of the method was driven by the need to analyze occurrence frequencies in a contingency table [5]. This technique can be viewed as finding the best simultaneous representation of two data sets that comprise the rows and columns of a data matrix with non-negative entries [4].

For a long time CA has been applied by European statistical community for psycho-metric and economic studies. This technique has been very popular in France, mainly owing to the efforts of Jean-Paul Benzécri [5]; it came to occupy such a strong position in the analysis methodology that it almost became synonymous with data analysis. In the 80's this technique started to be used in English-speaking countries since some books and papers presented the method in relatively simple form [4] [6] and now several American statistical package, such as SAS [7] and SPSS [8], include procedures to perform correspondence analysis. In the geostatistical context, applications of CA are relatively new. Avila *et al.* in [9] and [10] analyzed a data set consisting of the concentrations of chemical elements measured in a lake; Dutot *et al.* [11] applied this method to an aerosol collected in a simple atmospheric environment; Jiménez-Espinosa *et al.* [12] used CA on 602 soil samples taken in a region of NW Spain to identify geochemical patterns and anomalies.

All CA applications for environmental studies are limited to spatial multivariate data sets, where observations for several variables are spatially located [13]. Actually, most, if not all, environmental data are collected in space and time and exhaustive time series are often available for several monitored stations inside the area of interest. One of the major goal for an environmental quality control system is to obtain summary information about pollution conditions [14] [15]. Knowing the area inside the monitored region and/or interval of time

within the observed period which need of closer controls because of frequent exceeding fixed pollution levels, is definitely a very important issue. CA allows achieving this goal simultaneously for several contaminants. Therefore, it is useful to develop a procedure of applying CA to space-time multivariate data sets.

2.1. The Method

The theory of CA is discussed in several books, [4]-[6], so only the main features of the method are reviewed here.

From an initial data matrix $\boldsymbol{X}_{(l\times p)}$ with non-negative entries $x_{hi}, h=1,\cdots,l; i=1,\cdots,p$, CA determines the best simultaneous geometrical representation of rows and columns in a low-dimensional space (usually in a two-dimensional space).

Let $\boldsymbol{F}_{(l\times p)}$ be the relative frequency matrix, whose entries are:

$$f_{hi} = x_{hi} \sum_{h=1}^{l}\sum_{i=1}^{p} x_{hi}. \tag{1}$$

Two different matrices are used to re-scale $\boldsymbol{F}$, these are:

$$\boldsymbol{D}_l = \operatorname{diag}\left[f_{h\cdot} \right] \tag{2}$$

and

$$\boldsymbol{D}_p = \operatorname{diag}\left[f_{\cdot i} \right], \tag{3}$$

where:

$$f_{h\cdot} = \sum_{i=1}^{p} f_{hi}, \qquad \forall\, h=1,\cdots,l \tag{4}$$

and

$$f_{\cdot i} = \sum_{h=1}^{l} f_{hi}, \qquad \forall\, i=1,\cdots,p. \tag{5}$$

CA consists in finding a vector u, in a p-dimensional space, which maximizes

$$\boldsymbol{u}^{\mathrm{T}} \left[\boldsymbol{D}_l^{-1}\boldsymbol{F}\boldsymbol{D}_p^{-1} \right]^{\mathrm{T}} \boldsymbol{D}_l \left[\boldsymbol{D}_l^{-1}\boldsymbol{F}\boldsymbol{D}_p^{-1} \right]\boldsymbol{u}, \tag{6}$$

subject to the constraint:

$$\boldsymbol{u}^{\mathrm{T}} \boldsymbol{D}_p^{-1}\boldsymbol{u} = 1. \tag{7}$$

It is known that this is equivalent to finding the vector $\boldsymbol{v}$, in an l-dimensional space, which maximizes

$$\boldsymbol{v}^{\mathrm{T}} \left[\boldsymbol{D}_p^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}_l^{-1} \right]^{\mathrm{T}} \boldsymbol{D}_p \left[\boldsymbol{D}_p^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}_l^{-1} \right]\boldsymbol{v}, \tag{8}$$

subject to the constraint:

$$\boldsymbol{v}^{\mathrm{T}} \boldsymbol{D}_l^{-1}\boldsymbol{v} = 1. \tag{9}$$

The eigenvectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are related by:

$$\boldsymbol{v} = \frac{1}{\lambda}\boldsymbol{F}\boldsymbol{D}_p^{-1}\boldsymbol{u} \quad \text{and} \quad \boldsymbol{u} = \frac{1}{\lambda}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}_l^{-1}\boldsymbol{v} \tag{10}$$

where λ is the same eigenvalue for either maximization problems in (6) and (8).

This duality formula permits displaying the row and column projections in the same graph (called *biplots*) and this CA feature has been considered as its advantage with respect to others multivariate techniques.

Sequentially, the method searches for new solutions orthogonal to the previous ones; in particular, orthogonality is considered with respect to the inner product defined by the weighting matrices (2) and (3). There will be $\left(\min(l,p)-1\right)$ non-trivial solutions.

The factors

$$\Phi = \boldsymbol{D}_p^{-1}\boldsymbol{u} \tag{11}$$

and

$$\Psi = \boldsymbol{D}_l^{-1}\boldsymbol{v} \tag{12}$$

define the plane where rows and columns of the data matrix are projected.

Results from CA consist of graphical representations of the projections of rows and columns of the data matrix onto factorial planes, in order to find and understand underlying relationships [4]. There are also convenient diagnostics that help in the interpretation of the results; in particular:

• the *percentage of explained variation*, which is a measure of fit when a particular factor is retained, so that the *cumulative percentage of explained variation*

$$\frac{\sum_{k=1}^{K}\lambda_k}{\sum_{k=1}^{\min(l,p)-1}\lambda_k}, \tag{13}$$

represents a global measure of fit when K factors, $\left(K < \min(l,p)-1\right)$ are retained, each λ_k giving the contribution of a particular factor. Note that the terminology is similar to that one used in PCA, but in CA the term *variation* does not refer to *variance* in the statistical sense; it is an increasing function of K and it is used to choose the number of factors to be kept;

• the *absolute contributions* of the h-th row $(AC)_h^k$ and the i-th column $(AC)_i^k$ to the k-th factor, $k = 1,\cdots,\left(\min(l,p)-1\right)$, explain the composition of the retained factor. They are respectively:

$$(AC)_h^k = f_{h\cdot}\psi_{hk}^2 \qquad h = 1,\cdots,l, \tag{14}$$

and

$$(AC)_i^k = f_{\cdot i}\phi_{ik}^2 \qquad i = 1,\cdots,p; \tag{15}$$

• the *relative contributions* of a retained factor with the h-th row $(RC)_h^k$ or the i-th column $(RC)_i^k$ provide a measure of the row or column variation explained by the factor. They are respectively:

$$(RC)_h^k = \frac{\lambda_k\psi_{hk}^2}{\sum_{k=1}^{\min(l,p)-1}\lambda_k\psi_{hk}^2} \qquad h = 1,\cdots,l, \tag{16}$$

and

$$(RC)_i^k = \frac{\lambda_k\phi_{ik}^2}{\sum_{k=1}^{\min(l,p)-1}\lambda_k\phi_{ik}^2} \qquad i = 1,\cdots,p. \tag{17}$$

Note that the ACs serve primarily as guides to the interpretation of the dimension defined by the retained factors; whereas the RCs indicate how well a point is described by the retained factors. Usually, a large AC implies a large RC, but not conversely [6].

2.2. Computational Aspects

The application of CA to a space-time data set for multiple environmental variables is based on special contingency matrices generated as follows.

Let $z_r(\boldsymbol{s}_\alpha, t_\omega), r = 1,\cdots,R; \boldsymbol{s}_\alpha = (x,y)_\alpha \in D \subset \Re^2; t_\omega \in T \subset \Re$, be the space-time data for R variables measured at α-th location, $\alpha = 1,\cdots,N_r$ and ω-th observation time, $\omega = 1,\cdots,W$. For semplicity, consider $N_r = N, \forall\, r = 1,\cdots,R,$ although the procedure can be used to analyze variables measured at different sets of spatial locations.

Let $c_j^r, j=1,\cdots,J; r=1,\cdots,R,$ be J non-overlapping classes of values defined for each of the R variables under study.

Through the indicator transform, the belonging of $z_r(s_\alpha, t_\omega)$ to a certain class of values is described:

$$i_r\left(\boldsymbol{s}_\alpha, t_\omega, c_j^r\right)=\begin{cases}1 & \text{if } z_r\left(\boldsymbol{s}_\alpha, t_\omega\right)\in c_j^r,\\ 0 & \text{otherwise.}\end{cases} \tag{18}$$

From the four dimensional matrix (variable, station, time, class of values) obtained after the indicator transformation (18), the following two dimensional matrices are generated.

• Matrix $\boldsymbol{A}_{(R\times N)\times J}=\left[n_{\alpha,j}^r\right]$, where

$$n_{\alpha,j}^r=\sum_{\omega=1}^{W} i_r\left(\boldsymbol{s}_\alpha, t_\omega, c_j^r\right), \tag{19}$$

$$\alpha=1,\cdots,N;\ j=1,\cdots,J;\ r=1,\cdots,R.$$

In $\boldsymbol{A}$, the $(R\times N)$ rows represent all survey stations for each variable and the columns represent the J classes of values, so that the entries (19) indicate the number of times, values belonging to the j-th class, are recorded at the α-th station.

• Matrix $\boldsymbol{B}_{(R\times W)\times J}=\left[m_{\omega,j}^r\right]$, where

$$m_{\omega,j}^r=\sum_{\alpha=1}^{N} i_r\left(\boldsymbol{s}_\alpha, t_\omega, c_j^r\right), \tag{20}$$

$$\omega=1,\cdots,W;\ j=1,\cdots,J;\ r=1,\cdots,R.$$

In $\boldsymbol{B}$, the $(R\times W)$ rows represent the observation times for each variable and the columns represent the J classes of values, so that the entries (20) indicate, for each variable, how many stations in the ω-th observation time have values belonging to the j-th class.

The indicator transform allows the user to categorize continuous variables, synthesizing a large multivariate space-time data set. The above two dimensional matrices relate different classes of values (in the case study pollution levels) to locations (matrix $\boldsymbol{A}$) or to observation times (matrix $\boldsymbol{B}$), jointly for the variables (pollutants) under study. Thus, CA applied to each matrix, $\boldsymbol{A}$ and $\boldsymbol{B}$, will allow describing relationships

• in space, among pollution levels and monitored stations,

• in time, among pollution levels and observation times,

simultaneously for the variables under study.

CA results will also identify clusters of survey stations and intervals of time which need of closer controls when the contaminants frequently exceed fixed thresholds.

3. The Data Set

The data set consists of concentration values of two pollutants over a particular period of time and at stations of the monitoring network in Milan district, Lombardy (this is one of the northern Italy regions which suffers a serious air pollution pro-blem). The air quality monitoring network covers a wide area with about 190 stations where the main atmospheric contaminants, such as sulphur dioxide (SO_2), ozone (O_3), nitric oxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO), and meteorological variates, such as humidity, wind velocity, temperature, solar radiation, are continuously measured.

In the Milan district, air pollution is mainly caused by traffic and industrial activities. Two pollutants, which are primarily generated by the human activities, considered among the most dangerous ones for the atmosphere and human health and have been analyzed in this paper: NO_2 and CO. Nitrogen dioxide is a secondary pollutant generated by the thermic and photochemical reactions among the primary pollutants; it is caused, mainly in winter, by civil and industrial heating systems and by traffic. Therefore its concentration values are very high in urban areas characte-rized by high population density. Carbon monoxide is a primary pollutant caused by the motor vehicles emissions and its values are very high in areas with heavy traffic and poor ventilation. These characteristics are considered to choose the period of the year to be analyzed: January 1999. Indeed, most of the highest values for both pollutants under study were observed during the first month of the year. The box plot of

the hourly averages for each pollutant, measured during January 1999 (**Figure 1**), highlights exceeding the so called *level of attention* for several times during the month.

The national laws, particularly the Premier's Decree of the 12th of November, 1992, according to the European settlements, lay down, for each pollutant, a specific threshold called *level of attention*. When the pollution concentrations exceed this level for a long time and at several monitoring stations, air quality is poor and the situation is considered dangerous for the public health.

The analysis is limited to stations in the Milan district where data for both contaminants are available at all the desidered time points. In **Figure 2**, the 27 selected survey stations are shown. They have been classified, according to the Premier's Decree of the 20th of May, 1991, in two types:

• stations *C*, which are located in areas with heavy traffic and poor ventilation; in these areas the CO plume is more evident;

• stations *B*, which are located in areas with high density population, therefore these areas are subject to both NO_2 and CO pollution.

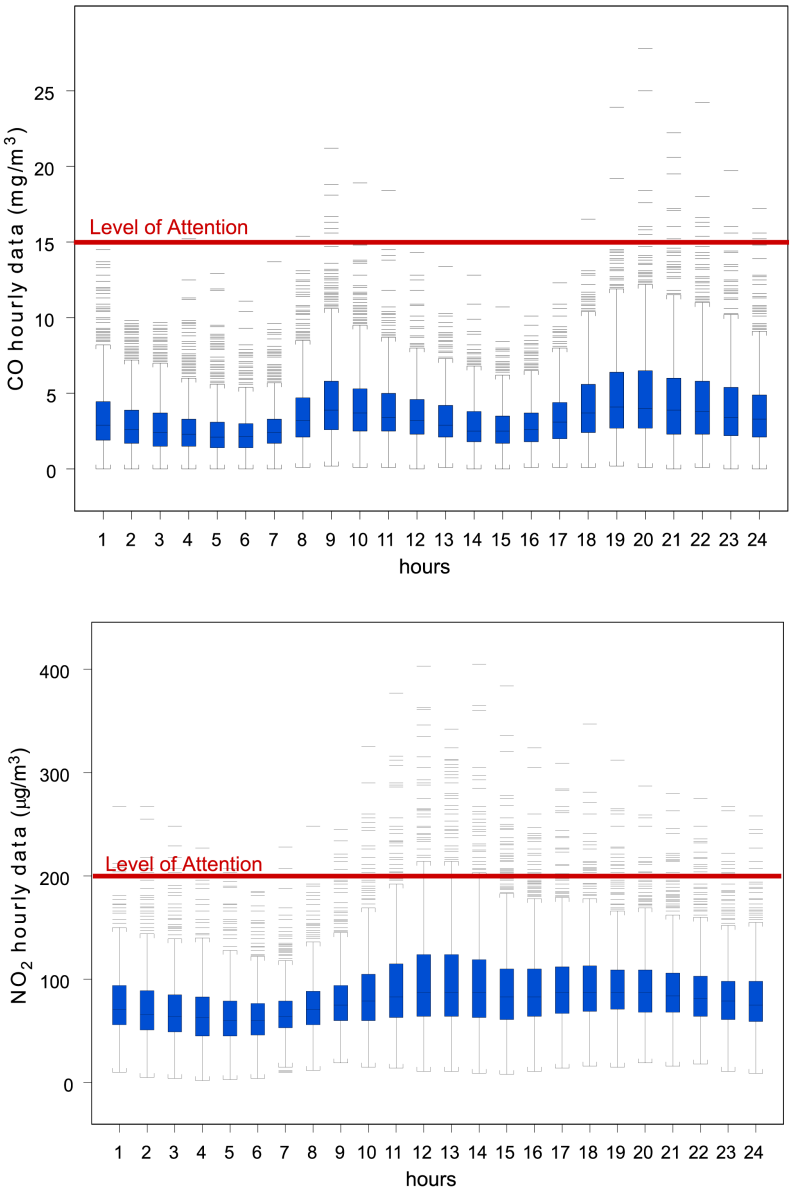

Figure 1. Pollution concentration values for CO and NO_2 during January 1999.

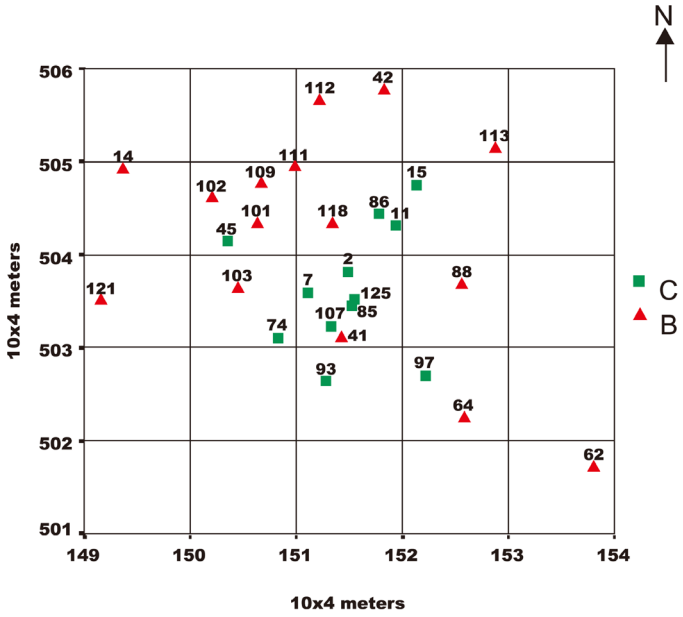

Figure 2. Posting map of the selected survey stations in Milan district.

In order to split each spatial-temporal distribution into non-overlapping classes of values, the following thresholds:

a) 1.6 2.3 3 3.9 5.4 mg/m^3

b) 52 64 75 90 115 μg/m^3

corresponding to the 0.17, 0.33, 0.50, 0.67, 0.83 quantiles of the distributions of CO a) and NO_2 b) hourly averages, are considered. Hence, six classes of CO and NO_2 concentrations are defined as follows:

$$c_1^r = \left[\text{minimum } z_r\left(\boldsymbol{s}_\alpha, t_\omega\right); 0.17 \text{ quantile value}\right]$$

$$c_q^r = \left]\frac{(q-1)}{6} \text{ quantile value}; \frac{q}{6} \text{ quantile value}\right], \quad q = 2,\cdots,5;$$

$$c_6^r = \left]0.83 \text{ quantile value}; \text{ maximum } z_r\left(\boldsymbol{s}_\alpha, t_\omega\right)\right],$$

$$r = 1,2;\ \alpha = 1,\cdots,27;\ \omega = 1,\cdots,744.$$

Then, through the indicator transform, two dimensional matrices are generated as described in (2.2); so that:

• ***A*** is a $(2\times 27)\times 6 = 54\times 6$ matrix, whose entries are:

$$n_{\alpha,j}^r = \sum_{\omega=1}^{744} i_r\left(\boldsymbol{s}_\alpha, t_\omega; c_j^r\right) \tag{21}$$

$$\alpha = 1,\cdots,27;\ j = 1,\cdots,6;\ r = 1,2\ ;$$

• ***B*** is a $(2\times 24)\times 6 = 48\times 6$ matrix, whose entries are $m_{\omega,j}^r$, as defined in (20), cumulated every 24 hours, that is:

$$\hat{m}_{\omega,j}^r = \sum_{\delta=0}^{30}\sum_{\alpha=1}^{27} i_r\left(\boldsymbol{s}_\alpha, t_\omega + 24\delta; c_j^r\right) \tag{22}$$

$$\omega = 1,\cdots,24;\ j = 1,\cdots,6;\ r = 1,2.$$

CA is applied to these matrices.

4. Results

A French package software, *SPAD* [16], is used for the data analysis since it performs most of multivariate

techniques, giving graphical results and diagnostics, in a very simple and fast manner.

Even if it is a commercial software, it is a very powerful software for data mining, indeed it can perform many statistical data analysis, as Factorial Analysis, Classification, Segmentation, as well as Textual analysis. Moreover, SPAD has a good graphical tools and is easy to use (user-friendly) [17].

CA is applied to matrix ***A*** and matrix ***B***, since information from both analysis are useful for the aim of the paper, as it will be shown.

The results from CA are displayed in a series of tables and graphs. In particular, **Table 1** and **Table 2** show the eigenvalues and the percentages of variation explained by the five non-trivial factors from CA applied to the matrix ***A*** and ***B***, respectively.

On the other hand, **Table 3** and **Table 4** list, for the first two factors from CA applied to the matrix ***A***, stations/pollutants with the highest absolute (**Table 3**) and relative (**Table 4**) contributions. Similarly **Table 5** and **Table 6**, which refer to the diagnostics from CA applied to the matrix ***B***.

Figure 3 and **Figure 4** show the projections of rows and columns of each matrix on the respective first factorial plane. Note that in **Figure 3**, which refers to matrix ***A***, columns (points labeled $c1,\cdots,c6$) and rows (points labeled with the station code and a symbol related to the pollutants) are displayed together on the same plane. Similarly in **Figure 4**, where columns (points $c1,\cdots,c6$) and rows (observation hours, $1,\cdots,24$, labeled with different symbols for the pollutants under study) of the matrix ***B*** are projected together on the same plane.

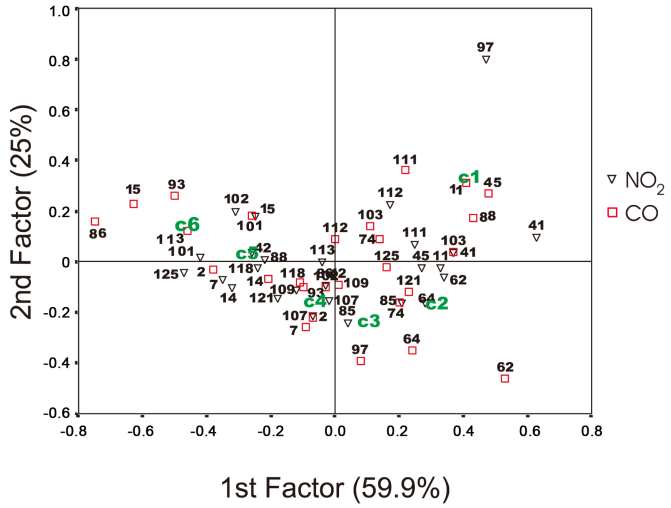

Figure 3. Plot of the first two factors from CA applied to the matrix ***A***.

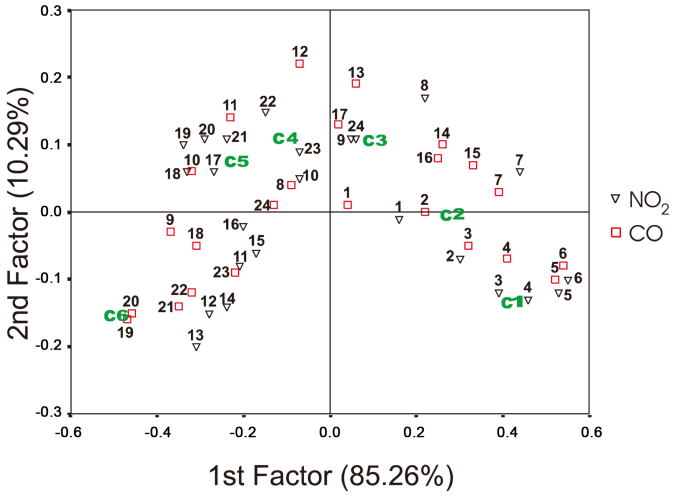

Figure 4. Plot of the first two factors from CA applied to the matrix ***B***.

Table 1. Eigenvalues and percentages of variation explained by the factors from CA applied to the matrix ***A***.

	Eigenvalues	Variation Explained (%)
Factor 1	0.0975	59.90
Factor 2	0.0408	25.05
Factor 3	0.0161	9.89
Factor 4	0.0048	2.94
Factor 5	0.0036	2.22

Table 2. Eigenvalues and percentages of variation explained by the factors from CA applied to the matrix ***B***.

	Eigenvalues	Variation Explained (%)
Factor 1	0.0951	85.26
Factor 2	0.0115	10.29
Factor 3	0.0030	2.69
Factor 4	0.0013	1.21
Factor 5	0.0006	0.56

Table 3. Highest absolute contributions to the first two factors from CA applied to the matrix ***A*** (percentages in parentheses).

Factors	Stations/Pollutants	Classes of Values
Factor 1	86/CO (11) 15/CO (8) 41/NO_2 (8) 93/CO (5) 45/CO (4) 125/NO_2 (4)	c6 (39) c1 (29)
Factor 2	97/NO_2 (27) 62/CO (9) 97/CO (7) 111/CO (6)	c1 (50) c3 (21)

Table 4. Highest relative contributions to the first two factors from CA applied to the matrix ***A*** (percentages in parentheses).

Factors	Stations/Pollutants	Classes of Values
Factor 1	41/NO_2 (95) 101/NO_2 (95) 86/CO (93) 103/NO_2 (91) 125/NO_2 (89) 42/NO_2 (87)	c6 (84) c5 (76)
Factor 2	85/NO_2 (92) 107/CO (89) 97/CO (86)	c3 (65) c1 (41)

The position of the points and the absolute and relative contributions suggest the following comments.

CA applied to the matrix *A*.

As previously described, matrix ***A*** relates six non-overlapping classes of values to CO and NO_2 survey stations, so that, by analyzing this matrix, it is possible to finding underlying relationships in space among

Table 5. Highest absolute contributions to the first two factors from CA applied to the matrix ***B*** (percentages in parentheses).

Factors	Hours/Pollutants	Classes of Values
Factor 1	6/NO_2 (7) 5/NO_2 (6) 6/CO (6) 5/CO (6) 4/NO_2 (5) 19/CO (5) 20/CO (5) 4/CO (4)	c6 (42) c1 (32)
Factor 2	12/CO (9) 13/CO (7) 13/NO_2 (7) 12/NO_2 (4)	c6 (34) c1 (25)

Table 6. Highest relative contributions to the first two factors from CA applied to the matrix ***B*** (percentages in parentheses).

Factors	Hours/Pollutants	Classes of Values
Factor 1	6/NO_2 (97) 6/CO (97) 4/CO (96) 5/CO (95) 5/NO_2 (92) 19/CO (89) 20/CO (88) 4/NO_2 (86)	c2 (92) c6 (91)
Factor 2	13/CO (83) 12/CO (82)	c3 (44) c4 (43)

different pollution levels and monitored locations. The first two factors are retained since they explain together about 85% of the total variation (**Table 3**).

The last class of values (*c*6) and the first one (*c*1) have the highest absolute contributions to the first factor, respectively, 39% and 29% (**Table 2**); whereas the first class (*c*1) and the third one (*c*3) have the highest absolute contribution to the second factor (respectively, 50% and 21%). Hence, the first factor better explains the variation of high pollution levels, *i.e.* levels which are greater than the last quantile values (5.4 mg/m^3 for CO and 115 μg/m^3 for NO_2), while the second factor better explains the variation of low pollution levels, *i.e.* levels which are smaller than 1,6 mg/m^3 for CO and 52 μg/m^3 for NO_2. **Table 2** list also the stations/pollutants which have the highest absolute contributions to the retained factors. The first factor better represents stations located at the central area of the Milan district (86, 125, 41, 93, 15), while the second factor better represents stations located at the peripheral areas (62, 97, 111).

The cumulative relative contributions to the first two factors (**Table 4** shows those stations/pollutants and classes with the highest relative contributions) are always greater than 80%, highlighting the good quality of representation of rows and columns in the space determining by the first two factors.

The projection of the classes to the first factorial plane (**Figure 3**) shows a *horseshoe* effect [6] which corresponds to a non-linear relationships between the two axes, even if they are linearly orthogonal.

In **Figure 3**, classes and stations/pollutants are displayed together so that and it is possible to identify two clusters of stations/pollutants:

1) points 111, 11, 45, 81, referred to CO, and point 41, referred to NO_2, with positive first and second co-ordinate;

2) points 86, 15, 93, 113, 101, referred to CO, and points 102, 15, referred to NO_2, with negative first co-ordinate.

The position of the second cluster on the factorial plane, being the points closer to point *c*6 with respect to the other points, highlights that most of the highest pollutant concentrations was read during January 1999 at those locations.

CA applied to the matrix *B*.

Matrix ***B*** summarizes the spatial aspect for each hour, since in this matrix each entry indicates how many

monitoring stations, at a fixed hour, have recorded pollution levels belonging to a given class of values. Hence, by analyzing this matrix, underling relationships among observation times (hours) and different pollution levels can be identified.

Table 2 shows the eigenvalues and the percentages of variation explained by each of the 5 non-trivial factors. In this case, the first factor explains a greater part of the total variation (85.26%) than in the previous analysis. The greater the percentage of explained variation, the greater the association between rows and columns of the data matrix, then the high percentage of variation explained by this factor is due to a strong association between observation times and classes of pollution levels. Once more, the first two factors are retained since they explain together more than 95% of the total variation.

Figure 4 shows the projections of rows (hours/pollutants) and columns (classes of values) to the first factorial plane. Now, a *horseshoe* effect is evident not only in the projections of the classes, but also in the projections of the hours/pollutants on the factorial plane: this means that distant pairs of hours can be considered as equidistant, while neighbouring hours are progressively dissimilar.

Table 5 and **Table 6** list the classes of values and the hours/pollutants with the highest absolute (**Table 5**) and relative (**Table 6**) contributions. Hours 4, 5, 6 referred to both contaminants have the highest absolute contributions to the first factor and, by looking to the factorial plane (**Figure 4**), the position of these observation times closer to point $c1$ with respect to the other points highlights that most of CO and NO_2 low readings was measured from the 4-th to the 6-th hour. Instead, most of the high pollution concentrations was observed during the evening, particularly during the 19-th to the 22-nd hour, for CO and from the 12-th to 14-th hour, for NO_2.

5. Conclusion

In this work, an application of CA to an air pollution space-time data set for CO and NO_2 hourly concentrations, recorded at some monitoring stations in Milan district, is given. The transformation of the original continuous variables into new categorical ones has been formally presented in this paper by the means of the indicator approach. By counting the indicator data over both spatial locations and observation times, two contingency matrices are generated. Each of them accounts information of both pollutants examined in this paper. CA is applied to these matrices providing a summary description of spatial and temporal profiles, simultaneously for the contaminants under study. The data analysis allows identifying relationships in space among CO and NO_2 pollution levels and monitored stations and relationships in time among CO and NO_2 pollution levels and observation times. The aim of each air quality control system is to obtain information about the atmospherical conditions and evaluate the opportunity of major restrictions and closer controls. The application of CA carried out in this paper makes it possible, since its graphical results and diagnostics help in identifying stations inside the area under study and intervals of time during the day for which the contaminants of interest need closer controls because of joint exceeding of fixed pollution levels.

Acknowledgements

The author would like to thank Prof. Donato Posa of University of Salento, Apulian region (Italy), whose suggestions have been helpful and improved this paper.

References

[1] De Iaco, S., Myers, D.E. and Posa, D. (2000) Total Air Pollution and Space-Time Modeling. In: Monestiez P., Allard D. and Froidevaux R., Eds., *GeoEnv III, Geostatistics for Environmental Applications*, Kluwer Academic Publishers, Norwell, 45-56.

[2] De Iaco, S. (2011) A New Space-Time Multivariate Approach for Environmental Data Analysis. *Journal of Applied Statistics*, **38**, 2471-2483. http://dx.doi.org/10.1080/02664763.2011.559206

[3] Blasius, J., Greenacre, M., Groenen, P.J.F. and van de Velden, M. (2009) Special Issue on Correspondence Analysis and Related Methods. *Computational Statistics and Data Analysis*, **53**, 3103-3106. http://dx.doi.org/10.1016/j.csda.2008.11.010

[4] Lebart, L., Morineau, A. and Warwick, K.M. (1984) Multivariate Descriptive Statistical Analysis. John Wiley & Sons, New York.

[5] Benzécri, J.P. (1983) Histoire et préhistoire de l'analyse des données. Dunod, Paris.

[6] Greenacre, M.J. (1989) Theory and Applications of Correspondence Analysis. Academic Press, London.

[7] SAS/Stat (1990) SAS Institute Inc., Cary.

[8] (1999) SPSS 8.0, SPSS Inc., Chicago.

[9] Avila, F. and Myers, D.E. (1991) Correspondence Analysis Applied to Environmental Data Sets: A Study of Chautauqua Lake sediments. *Chemometrics and Intelligent Laboratory Systems*, **11**, 229-249. http://dx.doi.org/10.1016/0169-7439(91)85002-7

[10] Avila, F., Myers, D.E. and Palmer, C. (1991) Correspondence Analysis and Adsorbate Selection for Chemical Sensor Arrays. *Journal of Chemometrics*, **5**, 455-465. http://dx.doi.org/10.1002/cem.1180050505

[11] Dutot, A.L., Bergametti, G. and Buat-Menard, P. (1988) Application of Correspondence Analysis to Apportion Sources of Ambient Particles. *Athmospheric Environment*, **22**, 1737-1743. http://dx.doi.org/10.1016/0004-6981(88)90403-9

[12] Jiménez-Espinosa, R., Sousa, A.J. and Chica-Olmo, M. (1992) Application of Correspondence Analysis and Factorial Kriging Analysis: A Case Study on Geochemical Exploration in *Geostatistics*. 2, Soares, A. Ed., Troia, 853-864.

[13] Greenacre, M.J. and Primicerio, R. (2013) Multivariate Analysis for Ecological Data. Fundación BBVA, Bilbao.

[14] De Iaco, S., Palma, M. and Posa, D. (2013) Prediction of Particle Pollution through Spatio-Temporal Multivariate Geostatistical Analysis: Spatial Special Issue. *AStA Advanced Statistical Analysis*, **97**, 133-150. http://dx.doi.org/10.1007/s10182-012-0199-0

[15] De Iaco, S., Maggio, S., Palma, M. and Posa, D. (2012) Advances in Spatio-Temporal Modeling and Prediction for Environmental Risk Assessment. In: Haryanto, B., Ed., *Air Pollution—A Comprehensive Perspective*, InTech, Croazia, 365-390. http://dx.doi.org/10.5772/51227

[16] (1999) SPAD 4.0, Cisia, Montreuil Cedex, France.

[17] Morineau, A. and Lebart, L. (1986) Specific Clustering Algorithms for Large Data Sets and Implementation in SPAD Software, in Classification as a Tool of Research. In: Gaul, W. and Schader, M., Eds., *Classification as a Tool of Research: Proceedings of the 9th Annual Meeting of the Classification Society* (*F.R.G.*), University of Karlsrube, F.R.G., North Holland, 321-329.