Annali del Dipartimento di Metodi e Modelli
per l'Economia il Territorio e la Finanza
1/2020, pp. 83-94, DOI: 10.13133/2611-6634/512

E-ISSN 2611-6634
ISSN 2385-0825

SAPIENZA
UNIVERSITÀ EDITRICE

Giuliana Passamani*, Paola Masotti*

A SYNTHETIC AIR POLLUTION INDEX USING A T2 APPROACH

Abstract. It is generally acknowledged that exceedance of the safe thresholds of air pollutants significantly affect human health. Hence, we can understand the importance of having a synthetic measure for overall air pollution, in a certain geographical area, computable using an index: to the values taken on by this index are then associated different health risks. In order to compute a reliable and comparable air quality index, in this paper we suggest an analytical procedure based on data reduction functions for summarizing three-way arrays, here consisting of the observations on a few major pollutants monitored over time at multiple sites. These functions reduce the dimensions of the array by applying a joint principal component analysis. Symmetrical three-way PCA has been successfully applied in environmental experimental studies, where the three dimensions of the data array are typically of similar length. Instead, a long time dimension usually characterizes environmental pollution data, where an asymmetrical two-way PCA is preferable: the aim of this paper is to indicate an approach that, by reducing two of the three dimensions, leads to the computation of an informative air quality index. Plaia et al. (2013) have applied an analogous approach, but reducing just one dimension of the array.

Keywords: three-way analysis, principal component analysis, T2-Tucker model, air quality index.

1. Introduction

The specialized literature has devoted a considerable effort towards measuring air pollution: various overall air quality indices (AQIs) have been proposed with the aim of combining daily observations on a variety of pollutants, at multiple monitoring sites, in such a way to give rise to a simple and informative indicator properly describing air quality in a particular area.

Some studies have focused their aim in developing indices based on the maximum operator as an aggregation function, either of sub-indices defined as equivalent measures of the observed air pollutants, or of sub-indices based on order statistics, as percentiles and maxima. The first is the case of the Pollution Standard Index (PSI) introduced by Ott and Hunt (1976), while the second is the case of indices obtained by means of hierarchical aggregation processes based on the median and the maximum, as in Bruno and Cocchi (2002).

Some other studies have proposed aggregate indices. Plaia et al. (2013) suggest an index based, first, on a spatial reduction and then, on a pollutant synthesis. Swamee and Tyagi (1999) point to the calculation for each pollutant of a sub-index expressed as a power function of its concentrations, and to the aggregation on an ordinal scale of the calculated pollutant sub-indices, thus determining a uniform index measuring the severity of overall pollution. According to Swamee and Tyagi, a synthetic air quality index should avoid the typical problems of ambiguity and eclipsicity: ambiguity is described as a characteristic of linear and root sum square aggregation forms and can raise unnecessary alarm, eclipsicity is a characteristic of weighted root mean square aggregation and can give a false sense of security. Swamee and Tyagi also highlight that the maximum operator aggregation function does not take into account changes in the remaining pollutants.

In the present study we focus on the three dimensional nature of the observed pollution data: they generally consist of the observations on a few main pollutants, at some monitoring sites, on different occasions, where the occasions are highly frequent, given that each site monitors air quality continuously

* Department of Economics and Management, University of Trento, Italy

by specialized devices. We suggest an aggregated air quality index based on a three-way principal component analysis (PCA).

Three-way PCA has already been successfully applied in environmental studies analysing hydro-chemical data, in an attempt to assess the quality of waters through the clear identification of factors characterising each of the modes designed for data collection: usually, chemical-physical characteristics together with spatial and temporal variability. As, for example, in Barbieri et al. (2002) for freshwaters or in Giussani et al. (2008) for lake waters, the application of three-way PCA leads to the extraction of relevant information hidden in the data. This type of extracted information is very useful to environmental agencies in developing strategies to manage water resources.

We will use the same approach for the analysis of our ambient dataset, with the aim of assessing the quality of air. However, as explained further on, the same approach will be adapted in order to enhance also the dynamic characteristics of our data. In other words, the aim of our work is to obtain a measure for daily air pollution which can be used not only for evaluating health risks associated to the level of pollution on a particular day, but, also, for monitoring its evolution over time. Therefore, we are not reducing time dimension, but the other two dimensions, the monitoring sites and the pollutants, taking into account possible interactions. The advantage is that we obtain a realistic unambiguous and non-ecliptic synthetic measure representing most of the variability observed in the data.

The structure of the paper is as follows. In Section 2, the Tucker decomposition models (Tucker, 1966) for summarizing three-way arrays in terms of their components are briefly presented: first, the symmetrical Tucker3 (T3) model, where all three dimensions are reduced, then the asymmetrical Tucker2 (T2) (Tucker1 (T1)) model, where just two (one), dimensions are reduced. The T2 model is the one chosen for analysing our dataset. In the same section, we also describe a function for aggregating the detected principal components when their number is greater than one. In Section 3, the main results of the analysis on our pollution data are described and the advantages of the suggested T2-Tucker procedure are outlined and then compared with the results of a conventional PCA. In Section 4 conclusions and suggestions for further research are drawn.

2. Methodological background

The available chemical, spatial and temporal information on pollution can be stored in a three-way array of data of order $(I \times K \times T)$, which is seen as a box with the vertical axis corresponding to the I units, the monitoring sites, the horizontal axis corresponding to the K variables, the pollutants, and the depth axis corresponding to the T occasions, the time frequencies.

If the purpose of data analysis were to produce synthetic and interpretable information of what is stored in the three-way array, we recognise that principal component analysis would be a candidate statistical technique for reduction of data dimensions. In a two dimensional framework, where the data matrix is made up by the observations of a set of variables on a sample of statistical units, conventional PCA is usually applied to reduce the number of variables by estimating a few uncorrelated linear combinations, or components, of the same variables, that contain most of the observed covariance. But, in a three dimensional framework the analysis is performed on a three-way data structure and the reduction problem becomes much more complicated because of the three-way possible interactions among the data, that should be taken into account.

2.1 The models for the principal component analysis in a three-way array

In order to better understanding the complexity of the three-way analysis, we present the Tucker-3 symmetrical PCA model following the approach of Giordani et al. (2014). Denoting with x_{ikt} the observation of the k-th variable on the i-th unit at the t-th occasion, the scalar formulation of the symmetrical three-way PCA model, or T3 model, leading to a limited number of components for all the three dimensions, can be written according to the following equation:

$$x_{ikt} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{kq} c_{tr} g_{pqr} + e_{ikt} \qquad (1)$$

where a_{ip} ,b_{kq} and c_{tr} represent, respectively, the loading of the *i*-th unit on the *p*-th component for the first dimension; the loading of the *k*-th variable on the *q*-th component for the second dimension and the loading of the *t*-th occasion on the *r*-th component for the third dimension. In the same equation g_{pqr} represents the core element and measures the interaction among the principal components of the three modes, and e_{ikt} represents the error term for the model.

As mentioned in the introduction, Barbieri et al. (2002) and Giussani et al. (2008) used three-way PCA formulated in (1) for analysing the chemical characterization of water quality, using monthly observations at sampling sites, over a rather short period. These environmental studies are mainly based on experimental data sampled and collected for the purpose, with the dimensions of the three ways similar in terms of number of observations. In both cited works, the data array is pre-treated, by centring and scaling each observed variable, to remove differences due to the units of measure. The results highlight the main relations between the characteristics of the waters and the sites monitored, as well as the variability over the short period. Conventional PCA would not allow the display of the amount of information provided by multiway PCA for environmental quality monitoring.

However, there are situations like the one we are focusing on in the empirical analysis, in which it could be useful to apply three-way PCA, but reducing only two of the three dimensions. This can be dealt with if we consider that the observed three-way data array **X** can be thought of as a collection of two-way matrices $\mathbf{X}_t$ of order $(I \times K)$, or a collection of matrices $\mathbf{X}_k$ of order $(I \times T)$, or a collection of matrices $\mathbf{X}_i$ of order $(K \times T)$. Thus **X** could be thought as matricized into three super-matrices of orders $(I \times TK)$, $(K \times TI)$, or $(T \times IK)$ respectively, as in Kiers (2000, p. 6).

Typically, air pollution data are recorded at high frequencies by real-time monitoring systems, using advanced electrochemical sensors and particulate matter analysers. As the purpose of analysing air pollution data is to produce daily information on air quality and then to compare its evolution over time, the reduction procedure should be performed on the collection of two-way matrices $\mathbf{X}_t$, by reducing the space dimension across the sites, and by reducing the variable dimension across the pollutants. Following this suggested procedure would then leave unreduced the time dimension.

According to the notation used in (1), the scalar formulation of the asymmetrical two-way PCA model, or T2 model, on a three dimensional dataset, will be as follows:

$$x_{ikt} = \sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip} b_{kq} g_{pqt} + e_{ikt} \qquad (2)$$

where only two of the three dimensions are reduced. This asymmetrical procedure would be the most useful for measuring pollution: in fact, the high frequency of time observations makes the time dimension very large and meaningfully unreducible, unless we monthly aggregate the daily observations or annualize them. Moreover, monitoring pollution over time is probably the main goal of analysing this kind of data, because it allows of properly acting in case of a worsening situation. These are the reasons why, given the just mentioned aim, three-way PCA would preferably be applied according to a T2 model, leading, in any case, to an interesting interpretation of the data structure.

In a T1 model, only one of the three dimensions would be reduced, as follows:

$$x_{ikt} = \sum_{p=1}^{P} a_{ip} g_{pkt} + e_{ikt} \qquad (3)$$

In fact, three-way T1 analysis comes down to a conventional PCA performed on a matricized version of order $(I \times TK)$ of the observed data array, where the dimension I is reduced.

The main advantage of using a T2 model like (2), instead of a T1 model like (3), is that we can still take into consideration the interactions between two of the three dimensions, whereas in a T1 model this is not possible, because only one dimension is analysed. In any case, a preliminary analysis of the data using a conventional PCA on the covariance matrix of each of the two super-matrices of order $(I \times TK)$ and $(K \times TI)$ resulting from matricizing the array $\mathbf{X}$, could be useful for detecting the likely number of components for both dimensions involved in the T2 model. The estimation of the T2 model parameters is then carried out by minimising $\sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{t=1}^{T} e_{ikt}^{2}$, using an Alternating Least Squares (ALS) algorithm implemented in ThreeWay of the R package (Giordani et al., 2014, p.3).

2.2 The aggregation of the components

In this sub-section, the attention is focused on how to derive a synthetic statistic giving a simple interpretation of the information extracted from the data using a T2 model, when the number of components, in at least one direction, is larger than one. In the following discussion, the time dimension is unreduced, while the number of components P, related to the monitoring sites, is assumed to be equal to one, that is P=1. This means that, analysing pollution data, the multiple sites are assumed fairly homogeneous with respect to the major pollutants recorded, and we consider this assumption to be quite reasonable for a limited geographical area.

For the T2 model[1], in the limit case when the number of components of both dimensions can be taken as equal to one, that is P=Q=1, the application of the reduction procedure will result in a core vector $\boldsymbol{g}$ of order $(T \times 1)$, where each element represents an overall spatial and pollutant synthesis on day t. This synthetic daily measure of pollution, for the area considered, is given by a single linear combination, with different weights, of the site and of the pollutant values, weights based on the contribution of each site and of each pollutant to the total variability.

Instead, when the number of components is P=1 for one dimension and $1 \leq Q \leq K$ for the other, the application of the same procedure will result in a score matrix $\boldsymbol{G}$ of order $(T \times Q)$, where each element represents a spatial and pollutant sub-index. This sub-index has still to be aggregated over the dimension of the reduced number Q of pollutants, using a proper aggregation function, in order to have a daily pollution measure, or pollution index.

In Swamee and Tyagi (1999), the suggested formula for aggregating pollutant sub-indices is:

$$I_{\rho}(t) = \left(\sum_{k=1}^{K} (g_k(t))^{\rho} \right)^{\frac{1}{\rho}} \tag{4}$$

where ρ is a parameter in the range $[1, \infty)$. If ρ=1, the resulting index is just the linear combination of the estimated pollutant scores. The authors show that for ρ=2.5 the resulting index does not suffer of eclipsicity and ambiguity is minimum. It should be noted that the formula does give the same weights to pollutants with high levels and with low levels, therefore the index should be carefully evaluated before any conclusion is drawn.

[1] For a T1 model, when P=1, the reduction procedure will result in a $(T \times K)$ matrix $\boldsymbol{G}$, where each element represents a spatial synthesis, but has to be aggregated over all the pollutants, in order to have the pollution index.

3. **The synthetic AQI resulting from the asymmetrical T2 analysis on pollution data**

In order to better understand the implications and the advantages of considering the T2 model instead of a conventional PCA for analysing a three dimensional array of pollution data, we perform the analysis that we describe in the following.

3.1 Pre-processing and standardizing the data array

Before actually carrying out a three-way analysis, it is useful to pre-process the data. Observations on aerosol and gaseous pollutants are usually recorded at hourly frequencies. It follows that, depending on their diurnal characteristics, the first step of any pollution data analysis is to aggregate the hourly observations by calculating daily syntheses of the selected pollutants, using an appropriate function according to the guidelines of the environment agencies. Given that each pollutant is measured on an appropriate and different scale, once the daily syntheses are computed they have to be standardized prior to any joint analysis of their possible relations, in particular if the aim is to derive a combined measure of the possible effects of pollutants on health.

The set of raw data originally available for our analysis consists of hourly observations on several pollutants recorded at seven monitoring sites over a certain time.

In order to have data on the same K pollutants for a certain number of monitoring sites, we have to focus just on three pollutants: particulate matter, PM_{10}, nitrogen dioxide, NO_2 and ground-level ozone, O_3. This is a constrained but reasonable choice[2], since it is well acknowledged by international organizations that exceedances of air pollutants pose serious health risks and PM_{10}, NO_2 and O_3 are generally recognized as the three pollutants that most significantly affect human health.

First, the hourly values obtained from continuous measurements of each pollutant at the various monitoring sites, have been transformed in daily observations by means of 24-hour running means in the case of PM_{10}**,** while, in the case of NO_2 and of O_3, the maximum hourly concentrations within the day have been taken as daily observations. The unit of measure is $\mu g/m^3$ for all three pollutants.

Of the seven monitoring sites for which we have the availability of data, the attention in the analysis is restricted to five of them that we consider as more homogeneous and to well represent the geographical area of the Italian province of Trento[3]. The chosen sites refer to the two main urban areas, Trento and Rovereto, to an area with quite a significant road traffic, Borgo Valsugana, to a touristic area not directly influenced by urban sources, Riva del Garda, and to a suburban area suffering for some road traffic, Piana Rotaliana.

The daily observations cover a period of two years, 2014 and 2015. Therefore, the overall three-way data array is of order $(3 \times 5 \times 730)$.

With the purpose of analysing comparable observations, we have to compute standardized indexed new daily values, PI_{ikt}, calculated according to the segmented linear principle based on the following formula[4]:

$$\mathrm{PI}_{ikt} = \frac{\mathrm{PI_H} - \mathrm{PI_L}}{\mathrm{BP_H} - \mathrm{BP_L}}(x_{ikt} - \mathrm{BP_L}) + \mathrm{PI_L} \tag{5}$$

where x_{ikt} is the daily concentration of pollutant k at site i on day t, $\mathrm{BP_H}$ ($\mathrm{BP_L}$) is the breakpoint $\geq$ ($\leq$) x_{ikt} and $\mathrm{PI_H}$ ($\mathrm{PI_L}$) is the PI value corresponding to $\mathrm{BP_H}$ ($\mathrm{BP_L}$). The resulting PI_{ikt} will be in the range [0, 100] for each pollutant, with the following breakpoint values: 25 for "good air quality"; 50 for "low pollution"; 70 for "moderate pollution"; 85 for "unhealthy for sensitive groups"; 100 for "unhealthy".

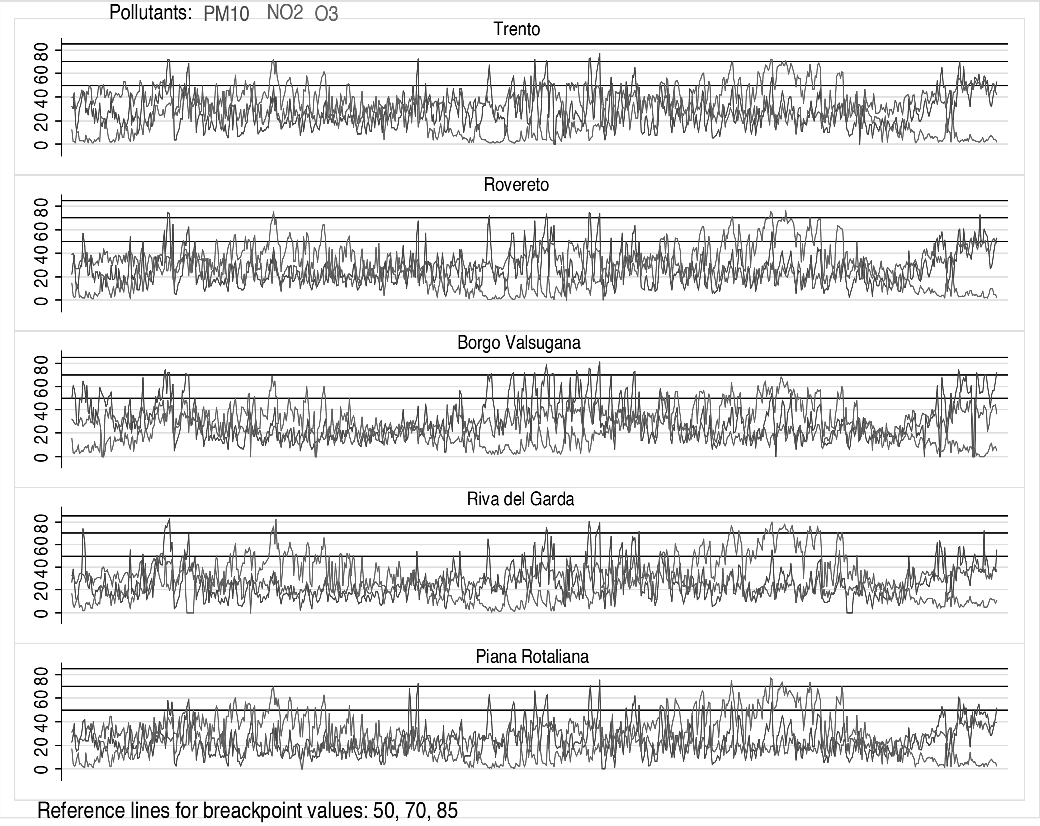
[2] According to European Citeair index directives, PM_{10} and NO_2, are mandatory pollutants for calculating roadside pollution indices and PM_{10}, NO_2 and ozone, O_3, are mandatory pollutants for calculating background pollution indices**.**

[3] The dataset for the empirical analysis has been provided by "Agenzia Provinciale per la Protezione dell'Ambiente (APPA)" of the Province of Trento (Italy).

[4] Data have been pre-processed and standardized as in Murena (2004) after having adjusted the boundaries of the classes according to the most recent European Environment Agency directives.

The behaviour over time of the newly calculated time series is shown in Figure 1, where we represent the PI_{ikt} values for the three pollutants, indicated with different colours, at each of the five monitoring sites. As can be noticed, most days can be classified in the categories "low pollution" (breakpoint value of 50) and "moderate pollution" (breakpoint value of 70), for any of the three pollutants, but, there are also many days "unhealthy for sensitive groups" in the case of PM_{10} - mainly, Borgo Valsugana - and in the case of O_3. Moreover, the three pollutants appear to be characterized by a clear seasonal variation, with PM_{10} and NO_2 positively correlated and both negatively correlated with O_3.

Figure 1. Standardized indexed daily observations over 2014 and 2015, for the three pollutants: PM_{10} *(navy),* NO_2 *(maroon) and* O_3 *(green), at the different monitoring sites (breakpoint lines in black).*



3.2 A comparison between Plaia's AQI and the T2 AQI

In this sub-section we compare the air quality index calculated by following the procedure suggested and used by Plaia et al. (2013)[5], with the one calculated by applying the T2 model approach that we propose in this paper.

[5] Working with daily data collected over the year 2006, for four main pollutants, at nine monitoring sites in the city of Palermo, instead of adopting an asymmetrical model like (2) for reducing the dataset, Playa et al. (2013) perform an asymmetrical PCA analysis reducing just one dimension of their three-way pollution data array.

With the newly calculated array **X** of order $(I \times K \times T)$, made up by the standardized measures PI_{ikt}, the reduction procedure using conventional PCA requires the matricization of the array **X** into the $(I \times TK)$ matrix **Y**, made up by juxtaposing the matrices $\mathbf{X}_t$ next to each other. PCA is then applied on the $(I \times I)$ covariance matrix, with the aim of aggregating monitoring sites. The resulting $(TK \times 1)$ vector containing the scores on the first principal component (PC) is then reordered in a $(T \times K)$ matrix containing a single daily index for each of the recorded pollutant, index representing the entire geographical area.

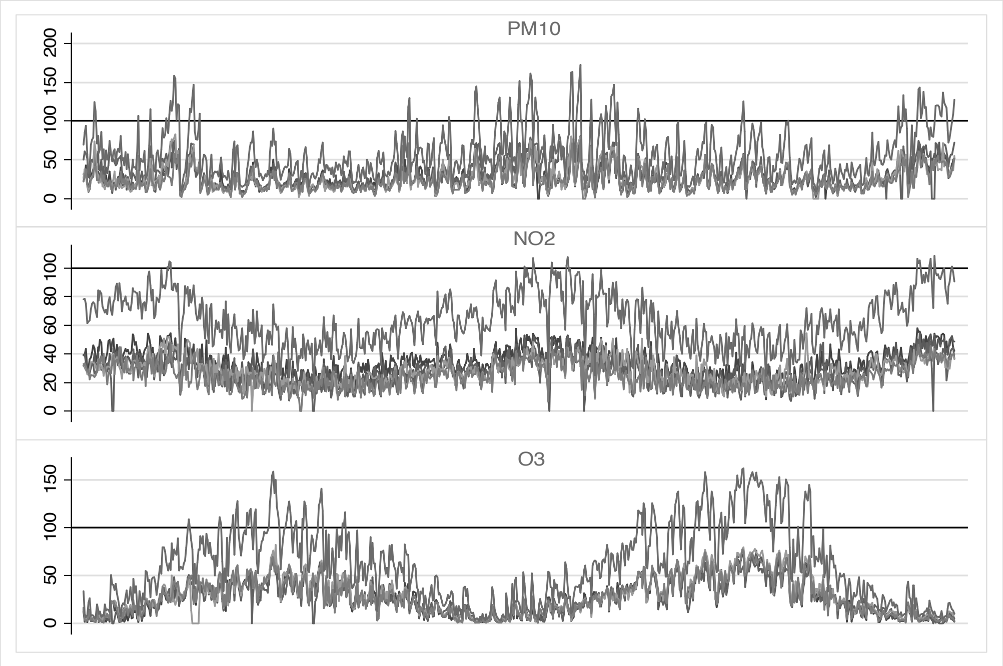
The procedure just outlined has been applied to our $(5 \times 3 \times 730)$ data array **X** of PI_{ikt} values. The main assumption for performing PCA on the covariance matrix of the matricized matrix **Y** is that the monitoring sites can be aggregated with respect to pollution, in other words, that they are quite homogeneous in terms of pollution. This seems to be a reasonable assumption according to the loadings/weights of each site on the first PC, shown in Table 1, which appear to vary within a narrow range, 0.408 and 0.478. It is to be noticed that the first PC results to explain 88.45% of the observed variability among the monitoring sites, which is quite a large value.

Table 1. Spatial aggregation using PCA: loadings on the first PC (explained variance 88.45%).

Sites	Trento	Rovereto	Borgo Valsugana	Riva del Garda	Piana Rotaliana
Loadings	0.464	0.453	0.478	0.408	0.429

The scores on the first PC, for each pollutant, are represented by the time series in red in Figure 2.

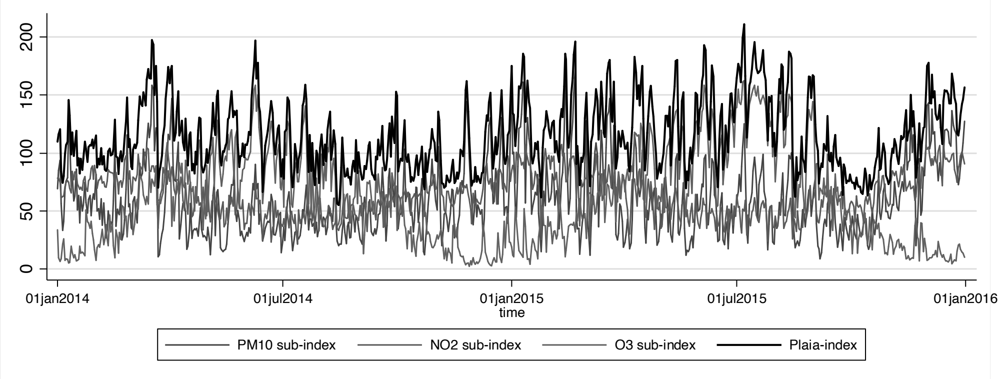
Figure 2. Air quality sub-indices (red) for PM_{10}, NO_2 *and* O_3, *over 2014 and 2015, calculated as scores on the first PC aggregating the monitoring sites. The black reference line highlights the limit value of 100.*



According to Plaia et al. (2013), the time series in red are to be considered as air quality sub-indices for each pollutant, PM_{10}, NO_2 and O_3: any time series value is the spatial synthesis for pollutant *k* at time *t*. In each graph of Figure 2, indicated with different colours, we can observe the standardized values whose weighted linear combination originates the synthetic value. As can be noticed, the synthetic values are much higher than the original standardized values and many of them are even higher than the limit value of 100 of the "unhealthy" category, according to the definition given by Murena (2004). Therefore, the newly calculated air quality sub-indices cannot be evaluated using the known pollution categories, contrary to what Playa et al. (2013, p. 389) state. Moreover, conventional PCA applied on the covariance matrix of the two-way matrix **Y**, does not allow to exploit the correlation among pollutants, but it just enhances the conjoint variability of the monitoring sites.

Moving, then, from the air quality sub-indices to the calculation of a synthetic daily index, if we follow the procedure suggested by Plaia et al. (2013), the overall time series of daily air quality indices is obtained by means of aggregating the pollutant sub-indices using the formula suggested in Ruggieri and Plaia (2011). Such aggregation by pollutants applied to our dataset, has given the time series represented in black in Figure 3. As can be observed, the overall daily Plaia-index takes on values well above the limits, even that for the "unhealthy" pollution category, and this occurs over the entire period of observation. In other words, if we want to evaluate overall pollution and we use the breakpoints indicated by the EC directives and categorized according to Murena (2004), the pollution daily index would suffer of the drawback of ambiguity, a problem of overestimation where the aggregate index is too high and crosses a limit level. This would imply unnecessary alarm, according to the definition given by Ott (1978).

Figure 3. Representation, in black, of the overall AQI time series calculated according to Plaia et al. (2013), together with the three PM_{10}, NO_2 *and* O_3 *air quality sub-indices used in the calculation.*



Now we apply the procedure proposed by us in this paper, to the same dataset.

First, we choose the options P=1 and Q=1 for the two reduced dimensions, which means just one principal component for both modes, the sites and the pollutants. The loadings a_{i1} of formula (2) take on values very similar to the ones in Table 1, while the loadings b_{k1} take on the values reported in Table 2, which shows similar weights for all three pollutants. The resulting overall explained variance is 82.50%.

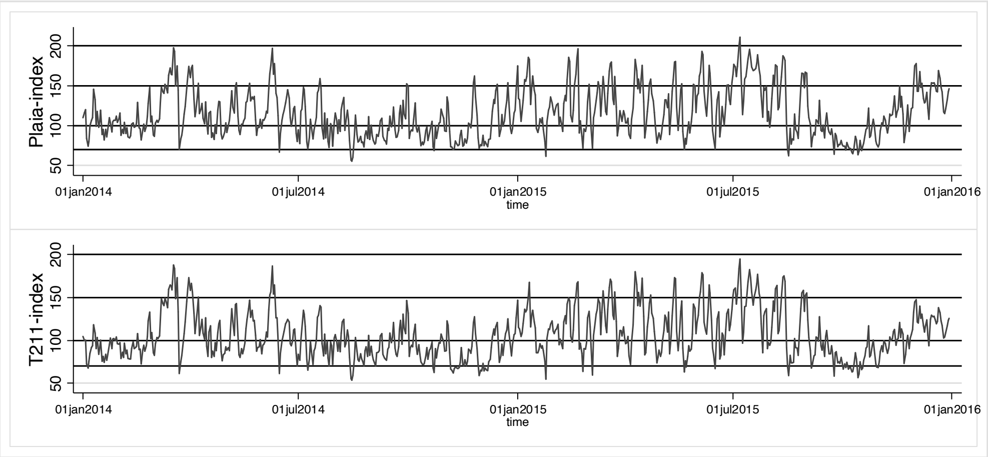
Table 2. Pollution aggregation using T2: loadings on the first pollutant PC.

Pollutants	PM_{10}	NO_2	O_3
Loadings	0.57	0.57	0.58

The results of the application of the T2 model approach[6] discussed in Section 2, consist in the overall air pollution index, corresponding to the core vector ***g***, represented in Figure 4, where it is shown together with the Plaia-index already discussed. It is to be noticed that the differences between the two indices are quite small: the behaviour of both is very similar, with the T2 index taking on values just a little bit smaller than those of the Plaia-index. This can be recognised if we look at the reference black lines corresponding to the values 70, 100, 150 and 200. In this particular case, the advantage of the T2 procedure, which aggregates over space and pollutants, is that the index is obtained within a single run of the software package, and taking into account, at the same time, the interactions between sites and pollutants.

The graphs in Figure 4 show that the calculated T2-index, as well as the Plaia-index, takes into account some additive effects among the pollutants, but the resulting values are such that we need to simulate new breakpoint values if we want to be able to evaluate the newly calculated index in terms of health risk categories. In fact, as can be observed, most values exceed the extreme breakpoint of 100 and therefore we need a new categorization system for the air quality index calculated according to either procedure.

*Figure 4. Representation of the Plaia's AQI time series and of the overall T2 AQI for P=*1 *and Q=*1.



Anyway, looking carefully at Figure 4, it is interesting to note that year 2015 is worse than 2014 in terms of pollution, with a clear deteriorating condition in summer 2015, when the combination of the effects of all three pollutants makes a really bad air.

It's also to be underlined that if we would have applied twice a conventional PCA retaining just one principal component, first on matrix $\mathbf{Y}$ and then on the $(T \times K)$ matrix of the scores, we would have obtained an index dominated mainly by O_3 and its seasonal behaviour[7], therefore an unreliable index.

3.3 Aggregation by pollutants in the T2 AQI

When using the T2 model approach in the case *P*=1 for one mode, and $1 \leq Q \leq K$ for the other mode, we need to aggregate the *Q* components in order to obtain the overall AQI. The case *P*=1 and *Q*=2 is quite interesting for our dataset, because it allows the presence of two principal components for the pollutant space, which is reasonable given that two pollutants, PM_{10} and NO_2 are positively correlated but, at the same time negatively correlated with the other pollutant O_3. In fact, allowing *P*=1 and Q=2 components in

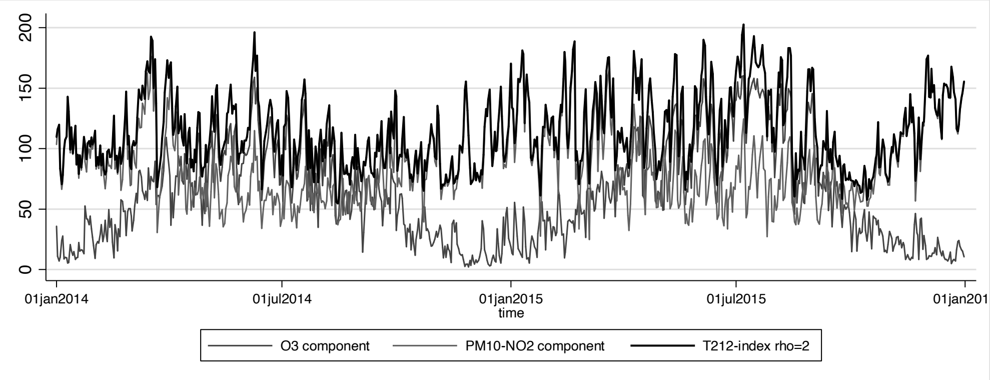
[6] Empirical analyses are performed using the package ThreeWay of R software.

[7] For reasons of space, we avoid the graph with the index time series, which is available upon request.

the T2 model, they will explain 95.06 % of the total variability. Moreover, of the two *Q* components one is associated with PM_{10} and NO_2, while the other to O_3, as can be seen after rotation of the pollutant space.

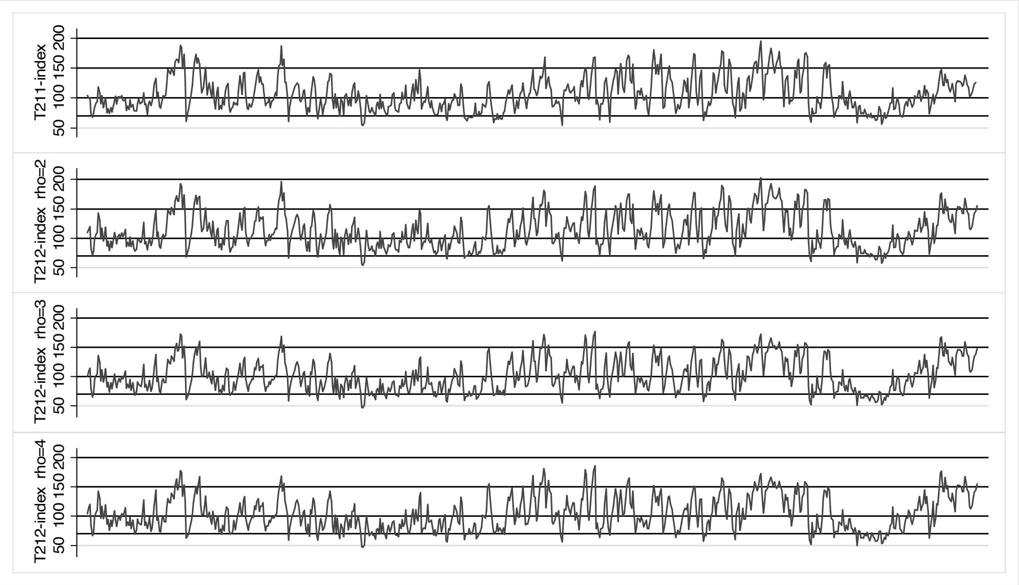
Following Swamee and Tyagi (1999) we aggregate the *Q* components using formula (4) for different values of ρ. In Figure 5 we graph the two pollutant components together with the T2-aggregate AQI calculated for ρ=2. As can be observed, the values taken on by the index represent quite closely the two components, resembling the maximum of either one, with the alternating of the seasons.

Figure 5. The two pollutant components and the T2-aggregate AQI for ρ=2



It is interesting to observe that the T2-aggregate AQI takes on values which are very close to the ones of the T2 index calculated for *P*=1 and *Q*=1, which is shown in the first graph in Figure 6. In the same figure, we also compare the T2-aggregate AQIs for different values of ρ.

Figure 6. Comparison among AQI indices: the T2-index and the T2-aggregated indices for ρ=2, 3, 4



What we can realize moving from the second to the fourth graph is that increasing ρ, the index values decrease and become smooth very slowly, showing a tendency to converge as ρ increases.

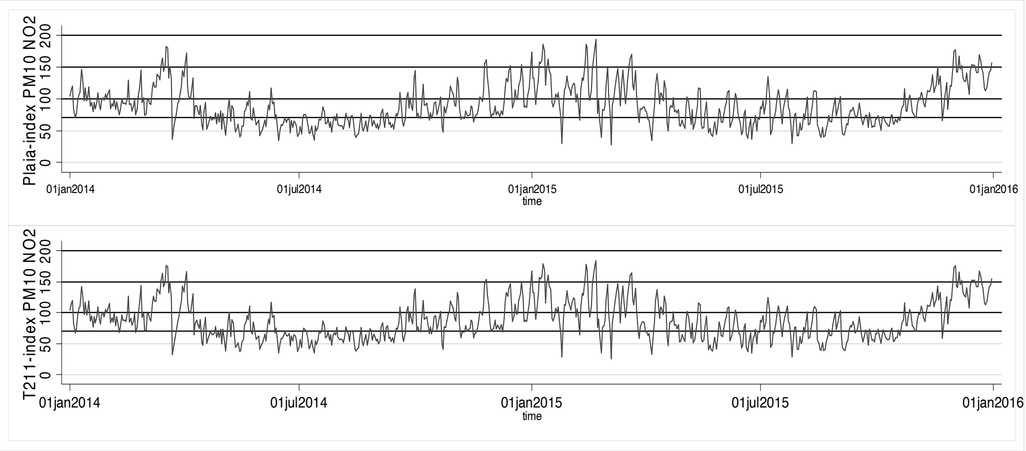
3.4 The breakpoints question for the aggregate AQI

As already stressed, one major question arising from the calculated air quality indices is how to evaluate and categorize their daily values in order to have an information interpretable in terms of health risks. We have seen that the range of their calculated values is wider than the range used for the standardized single pollutant, for which we have precise breakpoints and health risk categories (Murena, 2004). In other words, we need to define new breakpoints, since the original ones are conceived for one pollutant at a time.

What we must assume in defining new categories is that an overall AQI should consider the simultaneous presence in the atmosphere of several pollutants affecting human health, with possible additive and correlated effects. Moreover, the higher the number of pollutants analysed together, the larger the breakpoints values of the health risk categories must be, especially when the pollutants are inversely correlated.

As an example, we show in Figure 7 the graphs of the Plaia's AQI and the T2 AQI in the simple case of just two pollutants positively correlated, PM_{10} and NO_2. We can observe that there are many more values under the breakpoint value of 70 than in Figure 6, especially in summer.

Figure 7. Plaia's and T2 AQIs calculated on PM_{10} *and* NO_2



4. Conlusions

The proposed T2 approach, for calculating a reliable daily air pollution index, has been applied to a dataset made up by daily observations on three main pollutants, recorded at five monitoring sites over a period of two years. The procedure is asymmetrical because it reduces two dimensions of the three-way data array. The two dimensions chosen in this paper are the monitoring sites and the pollutants. With the proposed approach, the dimension reduction is performed jointly on both dimensions: in the particular case of P=1 principal component for the sites and Q=1 principal component for the pollutants, the daily AQI time series is obtained with a single run of the implemented procedure. Quite surprisingly, this time series is very close to the two step AQI time series suggested by Plaia et al. (2013), where they first apply a conventional PCA to the matricized data in order to obtain a space synthesis for each pollutant, and then aggregate the pollutant sub-indices. Similar AQI time series are obtained also when P=1 and Q=2, in which case we have to aggregate the pollutant components.

The comparison between the T2 air quality index proposed in this paper, and Plaia's AQI allows us to better understand how the proposed procedure works.

An important question arising when trying to evaluate the resulting AQIs, is the lack, for the calculated values, of breakpoints defining categories in terms of health risks. These categories for the overall AQI should consider the simultaneous presence in the atmosphere of several pollutants affecting human health, with possible additive and correlated effects. In addition, the number of monitoring sites considered in the calculation of the AQI, for a particular area of interest, plays a role. In fact, if the monitoring sites are quite homogeneous, the higher their number makes the weight of each one decrease, thus increasing the single pollutant sub-index resulting from conventional PCA.

Therefore, in order to have an overall air pollution index not giving misleading information, we have to work on the definition of breakpoints increasing their values with the number of pollutants analysed and with the number of monitoring sites covering the geographical area.

Given the main question just outlined, in order to compute an overall daily air quality index with the desired properties of no ambiguity and no eclipsicity, after having worked on the statistical procedure for calculating the index, the research needs to focus further on defining appropriate breakpoint values.

References

BARBIERI P., ADAMI G., PISELLI S., GEMITI F., REISENHOFER E. (2002), A three-way principal factor analysis for assessing the time variability of freshwaters related to a municipal water supply, *Chemometrics and Intelligent Laboratory Systems*, 62(1), 89-100.

BRUNO F., COCCHI D. (2002), A unified strategy for building simple air quality indices, E*nvironmetrics*, 13, 243–261.

GIORDANI P., KIERS H.A.L., DEL FERRARO M.A. (2014), Three-Way Component Analysis Using the R Package ThreeWay, *Journal of Statistical Software*, 57(7), 1-23.

GIUSSANI B., MONTICELLI D., GAMBILLARA R., POZZI A., DOSSI C. (2008), Three-way principal component analysis of chemical data from lake Como, *Microchemical Journal*, 88, 160-166.

GIUSSANI B., RONCORONI S., RECCHIA S., POZZI A. (2016), Bidimensional and Multidimensional Principal Component Analysis in Long Term Atmospheric Monitoring, Atmosphere, 7(155), 1-14.

KIERS H.A.L. (2000), Towards a standardized notation and terminology in multiway analysis, *Journal of Chemometrics*, 14, 105-122.

MURENA F. (2004), Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples, Atmospheric Environment, 38, 6195-6202.

OTT W. R. (1978), *Environmental Indices: Theory and Practice*, Ann Arbor Science Publishers: Ann Arbor.

OTT W.R., HUNT W.F. (1976), A quantitative evaluation of the pollutant standards index, *Journal of the Air Pollution Control Association*, 26(11), 1050-1054.

PLAIA A., DI SALVO F., RUGGIERI M., AGRÓ G. (2013), A Multisite-Multipollutant Air Quality Index, *Atmospheric Environment*, 70, 387-391.

RUGGIERI M., PLAIA A. (2011), An aggregate AQI: Comparing different standardizations and introducing a variability index, *Science of the Total Environment*, 420, 263-272.

SWAMEE P. K., TYAGIi, A. (1999), Formation of an Air Pollution Index, *Journal of the Air & Waste Management Association*, 49, 88-91.

TUCKER L.R. (1966), Some mathematical notes on three-mode factor analysis, *Psychometrika*, 31, 279-311.