Passamani Giuliana

A three-way synthetic air quality index capable of assessing health risks

Passamani Giuliana[1]; Masotti Paola[2]

[1] Department of Economics and Management, University of Trento, Italy.
e-mail address: giuliana.passamani@unitn.it
[2] Department of Economics and Management, University of Trento, Italy.
e-mail address: paola.masotti@unitn.it

Abstract:
Pollution data are information typically stored in a three-way array whose dimensions refer to units, variables and times: units are the monitoring sites, variables are the different pollutants and times are the days (or even hours) on which average pollutant concentrations are measured. Reducing the space and pollutant variability of observed data to a single daily indicator, associated with health risk evaluation, is of primary importance for taking decisions in order to protect people from possible health effects that pollution exerts. In the present paper, an air pollution index is developed and implemented for a given geographical area. Unlike other indexes already suggested in the literature, the one we propose takes into account the simultaneous presence of several pollutants in the atmosphere and their possible combined effects on human health. It also takes into account the space dimension by calculating a weighted average of the measured standardised concentrations over multiple monitoring sites characterized by different pollution conditions. The methodological approach we adopt relies on three-way principal component analysis, a multidimensional multivariate technique applicable symmetrically or asymmetrically, in order to allow an easier interpretation of the data structure. When we apply the technique on our pollution data, we obtain a spatially averaged air quality index combining the fair additive effects of different pollutants: such index results to be a reliable one, avoiding the problems of ambiguity and eclipsicity that usual air quality indexes suffer. The estimated combined effects of pollutants makes use of a matrix of weights based on the normalized space component matrix estimated using a T2 model, after reducing the site dimension. For assessing health risk, the values for the index, estimated using the suggested procedure, are to be classified with respect to the appropriate threshold values defining the health categories.

Keywords:
Air quality index; three-way PCA; T2-Tucker model; pollution health risk

1. Introduction:

Air quality indexes are important synthetic measures aiming to assess the effects of air quality on human health. They are based on data collected at monitoring sites, where several pollutants are observed and measured as hourly or daily concentrations using some averaging technique. Pollutant concentrations are then to be converted into a single numerical index describing the level of pollution and the associated air quality. EPA, EEA and other agencies as well as specialized literature, have suggested different procedures for conversion. Therefore, the main question arising is the following: how the different pollutants concentrations can be converted, at best, in a combined index? In the present work, we suggest an analytical procedure based, first, on data reduction and, second, on aggregating functions. We use this procedure for extracting information from a few major pollutants, monitored over time at multiple sites. The air quality index we propose takes into account the simultaneous effects on human health of the presence of several pollutants in the atmosphere and allows evaluation of air quality.

2. Methodology:

In a given geographical area, air pollution data are concentration values observed at monitoring sites where several pollutants are measured. Though recorded using the same measurement unit, the average daily concentrations represent values varying within ranges that are different for each pollutant. For this reason, they must be standardised before being analysed. The standardised daily values for each pollutant, at the different monitoring sites, are then collected into a three-way data array $\mathbf{X}$, of dimension (T x I x K): its generic element is x_{tik}, where t denotes the time, i the site and k the pollutant. It is an array of values subscripted by three indices, one for each of the A, B and C modes.

In order to reduce the site and pollutant dimensions - that is the B and C modes of the data array - into a smaller number of components, the specialized literature proposes quite a few statistical procedures. Given the asymmetrical reduction objective of our approach, between the two most popular approaches, the Tucker and the Parafac/Candecomp models, we prefer to focus the attention on the Tucker (T2) model (Tucker, 1966) and to use the notation developed by Kiers (2000) and adopted by Giordani et al. (2014).

Considering the $\mathbf{X}$ array as a collection of T matrices of order (I x K), we can define a new matrix $\mathbf{X}_A$ of order (T x IK), and write the T2 model, without loss of generality with respect to the T3 model, as follows:

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C}' \otimes \mathbf{B}') + \mathbf{E}_A \tag{1}$$

where $\mathbf{A}$ is a (T x T) matrix that, differently from the T3 model, will not be reduced in the T2 model. $\mathbf{G}_A$ is the matricized core array, a (T x PQ) matrix whose column elements are the interactions between the reduced components of the P sites and of the Q pollutants. $\mathbf{C}$ is a (K x Q) matrix and $\mathbf{B}$ is a (I x P) matrix, which are the component matrices for the B and C modes, and $\mathbf{E}_A$ is a (T x IK) matrix of errors.

In terms of the single standardised observation x_{tik}, model (1) can be written as:

$$x_{tik} = \sum_{p=1}^{P} \sum_{q=1}^{Q} b_{ip} c_{kq} g_{tpq} + e_{tik} \tag{2}$$

where b_{ip} and c_{kq} are the elements of the component matrices and represent, respectively, the loading of the i-th site on the p-th reduced site component and the loading of the k-th pollutant on the q-th reduced pollutant component.

The advantage of analysing $\mathbf{X}$ using a T2 model, instead of a standard Principal Component Analysis (PCA) fixed-effects factor analysis model, is that T2 takes into account the possible two-way interactions among the data, while information coming from PCA is incomplete.

The estimated parameter matrices of the T2 model (1) are obtained by minimising the sum of the squared residuals, using an Alternating Least Squares (ALS) algorithm implemented in the package ThreeWay available with the R software (Giordani et al., 2014, p.3).

The choice of the number of reduced components in either mode should be based on the maximum variability explained, while keeping their number as low as possible.

The idea we have in mind is to get a final single time series of values measuring the regional level of pollution resulting from the joint combined effects of the pollutants. In order to end up with a single time series, if the application of the technique suggests values larger than one for the number P and Q of reduced components, we have to aggregate the components. Therefore, the procedure will consist of two stages: in the first one, we reduce the B and C modes of matrix $\mathbf{X}$ and, in the second one, we aggregate the resulting reduced columns of the core matrix. The reduction must take into account the characteristics of the monitoring sites and of the pollutants. For the monitoring sites, the differences among them refer to

traffic conditions and to density of population, while, for the pollutants, the differences are more contrasting, so that it is reasonable that more components are needed for explaining their observed variability.
When we empirically analyse space dimension, it emerges a comparable homogeneity of the sites and, therefore, it is reasonable to assume that we can reduce the *I* different monitoring sites to just one component, that is *P* = 1. For aggregating the different sites, we follow an approach connected with PCA. We know that the objective of PCA is to find unit length linear combinations of the variables showing the greatest variances, so that the eigenvectors resulting from a PCA eigen decomposition of a covariance matrix of observed variables, are returned in orthonormal form, that is uncorrelated and normalized. Therefore, our procedure suggests the application of a normalization technique which takes the following form:

$$\hat{x}_{tk} = \sum_{i=1}^{I} (\hat{b}_i)^2 x_{tik} \tag{3}$$

where $\sum(\hat{b}_i)^2 = 1$. This aggregation formula has the advantage of transforming in normalized weights the loadings associated to the different sites.
Instead, for aggregating the different pollutants over the *Q* reduced dimension, we use a formula suggested by Swamee and Tyagi (1999):

$$I_\rho(t) = \left(\sum_{q=1}^{Q} \left(\sum_{k=1}^{K} \hat{c}_{kq} \hat{x}_{tk} \right)^\rho \right)^{\frac{1}{\rho}} \tag{4}$$

where ρ is a positive real number. The aggregation (4) takes into consideration the fair additive effects of combining pollutants and is free from eclipsicity, i.e. false security, and is defined so that ambiguity, i.e. unnecessary alarm, is minimized. Moreover, the aggregation (4) focuses on the pollutants having high pollution levels. Also Ruggieri and Plaia (2011) and Plaia et al. (2013) have applied the same formula for measuring air quality.
Though formula (4) gives a daily measure of the overall pollution index, in order to use it for assessing air quality, we need to define appropriate threshold values for health categories.

3. Result:

Due to the availability of data, we must restrict our empirical analysis to just three pollutants: particulate matter less than 10 micrometers, PM_{10}, nitrogen dioxide, NO_2, and ozone, O_3. According to European Citeair index directives, these represent the mandatory pollutants for calculating any background pollution index. With respect to them, we have daily data covering the period 2014-2015, collected at several monitoring sites: the resulting three-way data array contains concentrations data recorded using the same measurement unit, i.e. in terms of micrograms per cubic meter (µg/m3). However, the average daily concentration values differ both in terms of range and of seasonal patterns. To transform air pollutant concentrations into comparable indexes in the range [0, 100], we use an algorithm involving piecewise linear functions, as in Murena (2004):

$$x_{tik} = \frac{x_H - x_L}{BP_{Hk} - BP_{Lk}} (y_{tik} - BP_{Lk}) + x_L \tag{5}$$

where y_{tik} is the daily concentration of pollutant *k* at site *i* on day *t*, BP_{Hk} (BP_{Lk}) is the breakpoint ≥ (≤) than y_{tik} and x_H (x_L) is the x value corresponding to BP_{Hk} (BP_{Lk}). The resulting x_{tik} value will vary within the same range for any pollutant and will be characterized by the following upper threshold values: 25 for "good air quality"; 50 for "low pollution"; 70 for "moderate pollution"; 85 for "unhealthy for sensitive groups"; 100 for "unhealthy".

The three-way data array $\mathbf{X}$ available for the empirical analysis is of dimensions (730 x 5 x 3). With the aim of showing the kind of data we are working with and their seasonal pattern, in Figure 1 we represent the standardised data for the three pollutants at a single chosen monitoring site, located in the main town of the region. We recall that we have corresponding data for other four sites located in the geographical area of interest.

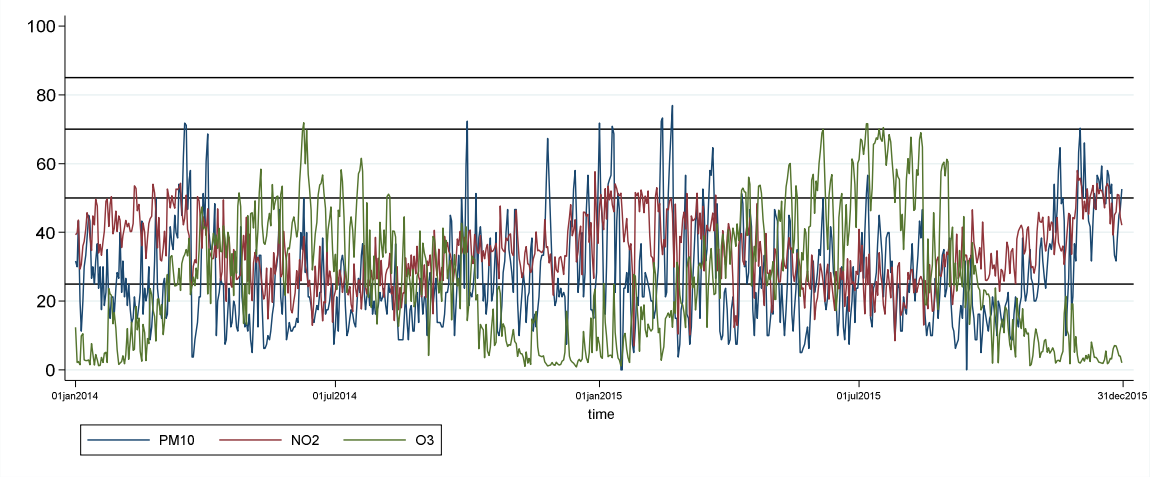

Figure 1: Standardised daily observations, over 2014 and 2015, for the pollutants: PM_{10} (navy), NO_2 (maroon) and O_3 (green), at Trento PSC monitoring site (threshold lines in black).

In the figure, we report the standard threshold reference lines for assessing air quality with respect to each pollutant. As can be noticed, most days can be classified in the categories "low pollution" (threshold value of 50) and "moderate pollution" (threshold value of 70), for any of the three pollutants, but, there are also days classified as "unhealthy for sensitive groups".

Given this datast, we aim to estimate an overall air quality index capable of taking into account the possible combined effects of different pollutants on human health.

If we impose on model (2) the restrictions $P = Q = 1$, the estimation procedure would result in a pollution time series index that does not need further aggregation of pollutant components.

Instead, if we allow the pollutant C mode to be reduced to $Q = 2$ components - which is more realistic given the different characteristics of the pollutants - and then we aggregate the pollutant components using the formula (4) of Swamee and Tyagi, with ρ = 3, the procedure would result in another index measuring overall air pollution condition on that particular day. The two pollution indexes obtained as just described are represented in Figure 2: the first one, IndexT211, in red and the second one, IndexT212i3ms, in blue.

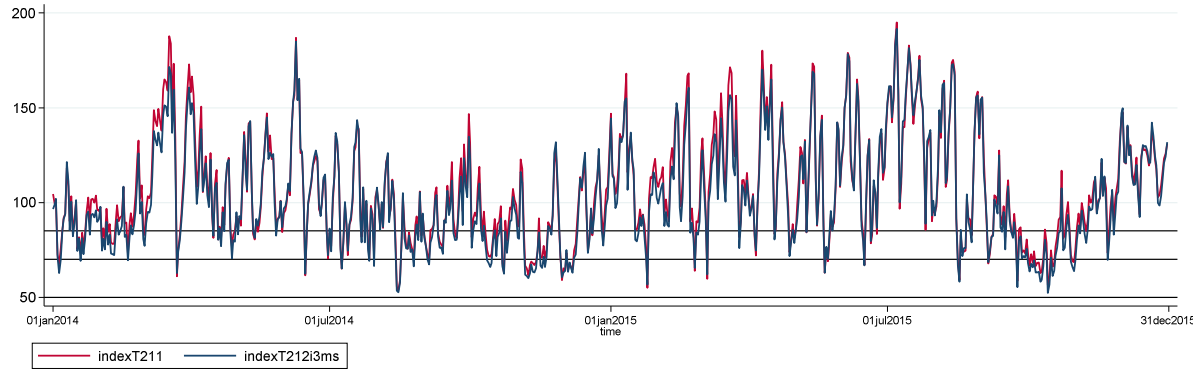

Figure 2: Estimated overall pollution indexes, without normalizing the site loadings (threshold lines in black).

What we can notice is that both pollution indexes, which are very similar, take on values that are well beyond the threshold values defined for any single standardised pollutant. In other words, if we use these overall indexes measuring the combined effects of pollutants on health, we need new threshold values for delimiting health risk categories.

Instead, if we apply the following suggested procedure, we end up with very interesting estimated values for the overall pollution index. The procedure consists of these steps:

– estimate model (1) choosing the smallest combination of P and Q which gives the highest fit;
– normalize the site loadings by calculating their squares and then apply formula (3) for space aggregation and formula (4) for pollutant aggregation: the result is a single time series index measuring the combined pollution levels;
– for assessing health risks, compare the values of the resulting index with the threshold values used for standardising observed pollutant concentrations.

It is interesting to consider more closely the air quality index resulting from the application of the just suggested estimation procedure. In Figure 3 we represent the standardised observations on the pollutants at Trento PSC monitoring site, as in Figure 1, together with the estimated new overall pollution index, indexT2122i3ms.

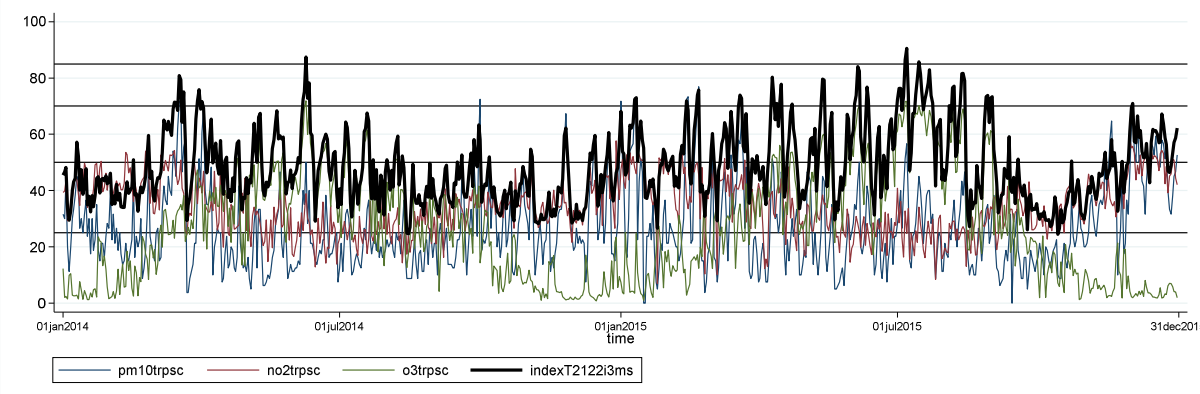

Figure 3: The overall estimated pollution index (black) and the standardised daily observations for PM_{10} (navy), NO_2 (maroon) and O_3 (green), at Trento PSC monitoring site (threshold lines in black).

As can be easily detected, the air quality index appears to move along the maximum values taken on by the pollutants and shows that overall pollution levels are mostly classified as “low pollution” and “moderate pollution” with many values classified as “unhealthy for sensitive groups” and few even “unhealthy” in summer 2015. Other studies have developed 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) and Ott(1978), 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).
What we can conclude is that the proposed procedure, which uses, for the loadings, a normalization technique based on the same constraint applied when finding the principal components in PCA, gives a very reliable air quality index comparable with the ones suggested in the literature. Moreover, the threshold reference values for health categories are just the standard ones.

4. Discussion and Conclusion:

Highlighting the importance of having a trustworthy combined air pollution index measuring the actual air quality in a certain geographical area on which are located multiple monitoring

sites, we, nevertheless, must pay particular attention to the definition of consistent threshold values for the associated health categories. This is the focus of the present work in which we suggest a procedure for estimating an aggregate reliable air quality index whose values can be easily compared with the associated appropriate health categories. Some interesting indications have emerged, leading the path to further research.

References:

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.

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., Tyagi, 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.