ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Indoor air pollutant sources using blind source separation methods

Rachid Ouaret[1], Anda Ionescu[1], Olivier Ramalho[2] and Yves Candau[1]*

1- University Paris-Est, Center for Study and Research on Thermics, Environment and Systems
University Paris-Est Créteil, 61 avenue du Général de Gaulle, F-94010 Créteil Cedex - France
2- University Paris-Est, Scientific and Technical Centre for Building (CSTB)
84 avenue Jean Jaurès, 77420 Champs-sur-Marne - France

Abstract. The objective of this study is to separate different sources of variability of air pollutant concentrations time series of particulate matter (PM) monitored in real indoor environments. Different blind source separation (BSS) methods (ICA, PMF, NMF) were applied in order to identify the PM sources and their contributions. The source profiles were characterized by their autocorrelation functions (ACF) which were compared to the ACFs of other variables. Their interpretation was completed by the analysis of polar plots including exogenous factors. Source contributions were also quantified.

1 Introduction

People spend more than 80% of their time indoors. Meanwhile, the indoor source temporal variability is still very little studied. The information on pollutant emission sources is transmitted through their mixtures when monitoring air pollutant concentrations. Thus, the dataset collected from this mixture represent the integrated of underlying latent factors.

Particulate number concentrations (PM) have been measured with a high time resolution during a long period. The PM source time variability was explored using BSS methods such as Independent Component Analysis (ICA), Positive Matrix Factorization (PMF) and Non-negative Matrix Factorization (NMF). These methods decompose the measured data sets into factor profiles and factor contributions.

2 Data sets

The first monitoring campaign was performed in an individual office, covering a period of 45 days (from March 1st to April 4th, 2011). The indoor air particle number and CO_2 concentrations were sampled every 10 minutes. The environmental measurements and occupational parameters in the second campaign were performed in an open-plan office occupied by 6-8 persons over six months (from January 1st to June 30, 2015). The data include particle number concentrations, the windows opening and the human presence with 1-min time step. The particle number measurements were performed using an optical particle counter (Dust Monitor 1.108, Grimm) for fifteen size bins within a range of 0.3-20 µm.

* This work received financial support from the French research program on air quality (PRIMEQUAL) through the grant n° 12-MRES-PRIMEQUAL-4-CVS-09.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

3 Methodology

3.1 The mixing model

The mixing instantaneous linear models describe can be described as follows:

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{m \times T}$ is the observation matrix, $\mathbf{A} \in \mathbb{R}^{m \times p}$ represents the mixing matrix, $\mathbf{S} \in \mathbb{R}^{p \times T}$ is the source profiles matrix and $\mathbf{E} \in \mathbb{R}^{m \times T}$ corresponds to the additive error-perturbation matrix. Given $\mathbf{X}$, the aim is to estimate the matrices $\mathbf{A}$ and $\mathbf{S}$ by factorization under different constraints by the minimization of different cost functions f (Euclidian or Frobenius distances, Kullbach-Leibler divergence, negentropy…):

$$\min_{c} \left\{ f\left(\mathbf{X}, \mathbf{AS}\right) \right\} \quad (2)$$

3.2 Blind Source Separation (BSS) methods

For source identification and apportionment based on the chemical fingerprint, the most commonly used method in outdoor air pollution has been, until now, the Positive Matrix Factorization (PMF), developed by Paatero [1, 2]. Compared to the Principal Components Analysis (PCA) or Independent Component Analysis (ICA) [3,4], it has the advantage of more realistic non-negative constraints on factor profiles and contributions, and a better scaling of the data by individually assigned uncertainties.

Following the main idea of positivity constraint, the Non-Negative Matrix Factorization (NNMF or NMF) was introduced later by Lee and Seung [5]. Under certain conditions (by introducing weights in the classic NMF), the PMF and NMF methods can be equivalent.

By contrast to the constraint of positivity, the ICA requires statistical independence of the components. The components can be also interpreted as source profiles and the loadings as their contributions, but the problem is that they can be also negative.

Finally, the PCA can be considered also as a method for matrix factorization, requiring non-correlated components. It has the same inconvenient of negativity as the ICA when used for source identification and apportionment.

3.3 Algorithms, initialization and number of factors

In this study, Multilinear Engine (ME) algorithm was used for the PMF application. Based on Kullback-Leibler divergence, simple multiplicative updates algorithm is employed for the standard NMF. FastICA algorithm [4] has been employed in the ICA application.

In general, BSS problems are non-convex and non-linear in the variables **A** and **S**. Therefore a reasonable approach consists in a preprocessing step for matrix initialization to find local solution. A Nonnegative Double Singular Value Decomposition (NNDSVD) [6] was applied for NMF procedure.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

In this study, the retained number of factors corresponds to the inflection point on the RMSE curve between **X** and estimated **AS**; a number of four factors corresponds to the local optimum.

3.4 Source characterization

3.4.1 Source characterization using the autocorrelation function (ACF)

One of the most important criticisms made on matrix factorization methods such as ICA and PCA lays in the lack of physical interpretation of negative factors. However, by using the inherent variability structure as given by the ACF, this inconvenient can be avoided. Thus, the source variability, obtained by a BSS method can be characterized by its ACF. The rate at which ACF decays to zero may be then interpreted as a measure of the memory of the process. In this work, the ACF decay structure of observed time series (PM, CO_2 and climatic parameters) is compared to the ACF structure of the estimated factor profiles. Similar ACF is than interpreted as similar variability, when comparing an estimated source to another factor.

3.4.2 Polar plots with different exogenous parameters

In order to analyze the influence of different parameters, such as wind speed and direction, polar plots of the estimated sources can be drawn. Some other parameters such as window opening or human presence in the office can be considered, too.

4 Results

4.1 Individual office- monitoring campaign 2011

The PCA and ICA were applied to the standardized database composed of 15 fractions of PM. In order to avoid the fact that the resulted components can be also negative, their ACFs were calculated.

Figure 1 shows an example of ACF for the first two principal components, PC1, PC2 (Fig.1c) and of two independent components, IC2, IC4 (Fig.1d), the most representative ones. One can notice the same type of ACF decay for PC1, IC4, CO_2, $PM_{2.5}$ and $PM_{4.5}$. The source revealed by PCA and ICA (PC1 and IC4) generates or simply corresponds to medium-sized particles such as $PM_{2.5}$ and $PM_{4.5}$ (Fig.1b); in addition the same structure characterizes the ACF of CO_2 (Fig.1a), which is an indicator of human presence. Most ACF decay behaves as a mixture of damped exponential-sinusoidal functions, especially for medium-sized particles and for CO_2. This reveals an important aspect of the source variability: the seasonality.

The ACFs for PC2 and IC2 are very similar. They can represent a source of fine particles, because the ACF of $PM_{0.35}$ (Fig.1b) is characterized by the same feature. For finer particle number concentrations, the autocorrelation values are relatively high with a very slow decrease (power low). This indicates the persistence characteristics of fine particle sources.

Figure 2 shows the relative contributions of different factors obtained with three factorization methods: ICA, NMF and PMF. The curves can be interpreted in terms of granulometry of the different sources. The components obtained with ICA (Fig.2a)

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

and NMF (Fig.2b) factorizations contribute in a distributed manner to almost the whole range of particle size, but the PMF method (Fig. 2c) reveals two distinct groups: a cluster for coarse particles, and another one for finer particles.

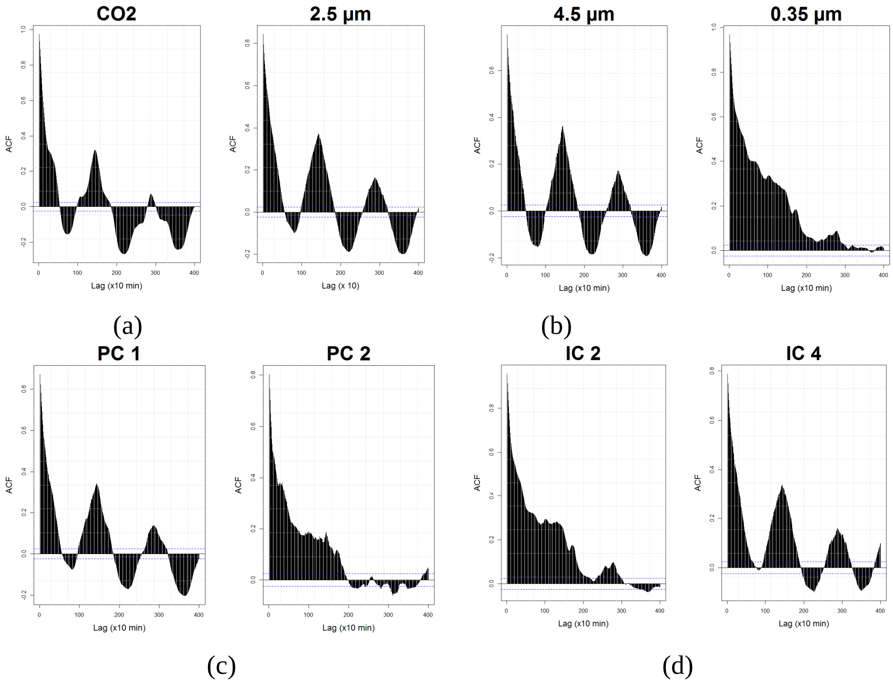
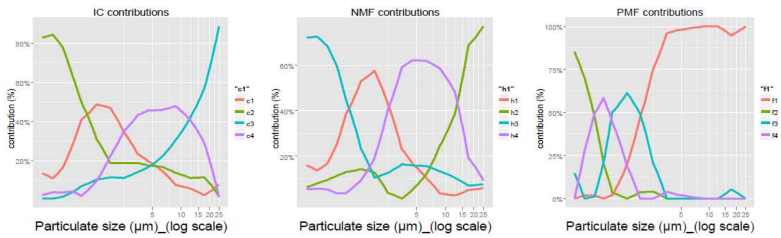


Fig 1. ACF of (a) indoor CO_2; (b) PM fractions; (c) 1^{st} and 2^{nd} principal components; (d) 2^{nd} and 4^{th} independent components.



(a) (b) (c)

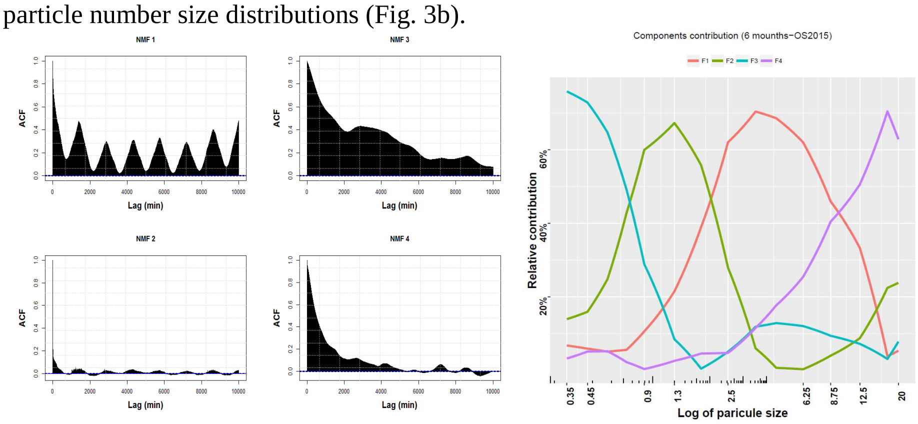
Fig. 2: Relative source contributions in the individual office obtained using different BSS methods: (a) ICA; (b) NMF; (c) PMF.

Due to the similarity of the results between the different BSS methods, only the NMF results are presented for the second campaign.

4.2 Open-plan office- monitoring campaign 2015

The NMF method is applied to the 15 ranges of indoor particle number concentrations measured in the open-plan office. The autocorrelation function of each estimated factor, as well as their relative contributions to the fifteen size ranges are presented in Figure 3.

The ACF of the first factor (NMF1) decreases as sinusoidal but it is still positive for the first 167 hours lag (~7days), while the ACF of the second factor (NMF2) decreases faster (Fig.3a). This indicates that the first one captures the phenomena with a very marked diurnal profile, depending probably on climatic parameters. The second one could be associated to stochastic phenomena. The ACF of the third and the forth factors are decreasing as a hyperbolic function. These components capture the deterministic trends of sources and can be interpreted as permanent (persistent) factors. The relative contributions obtained with NMF are distributive according to particle number size distributions (Fig. 3b).



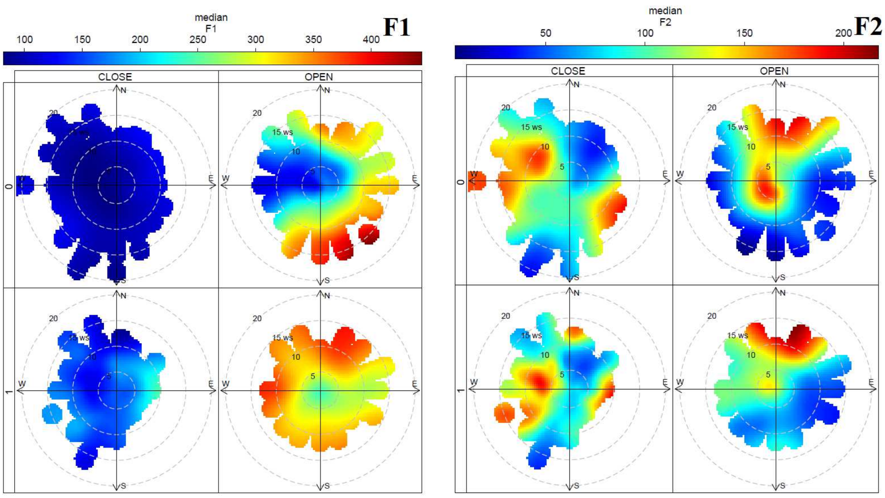
(a) (b)

Fig 3. Autocorrelation function (ACF) of the four NMF components (a) and their relative contribution according to size particles diameter.

In order to examine more in detail the origin of these fluctuations, the factorization results were combined with the human presence and window opening. Thus, Figure 4 shows the polar variability of two NMF factors according to different parameters: the direction and the wind speed, the human presence and the windows opening. The high values for the first factor are observed only when at least one window is opened; this suggests the impact of outdoor sources on the indoor levels.

The ACF of the first factor exhibits the same pattern as the wind speed ACF. This factor can be assigned to outdoor sources, which impact the indoor environment *via* the wind when windows are opened. Furthermore, the wind parameters play a predominant role in case of the absence of occupants. Thus the first factor corresponds to outdoor sources. For the second factor and in the case of opened windows, the highest values correspond to the situations when the wind blows from north-east, where is located an expressway with high traffic. The source of these levels can be thus attributed to the road traffic.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.



(a) (b)

Fig 4. PolarPlot of NMF1 and NMF2 estimated sources variability according to windows opening and human presence.

5 Conclusions

This study showed the potential of matrix factorization methods to separate sources of indoor air pollutant concentrations of particulate matter. Two approaches were employed to give and interpretation to the factors. The ACF factors were compared with the ACF of other parameters (e.g. CO_2-human presence indicator) or to the ACF of the different particle fractions. The ACF shape gives some information about the source type: chronic (diffuse) or intermittent. Factor interpretation (source identification) could be achieved in some cases *via* a multivariate analysis presented as a polar plot according to wind speed and direction, as well as other parameters: windows opening, human presence.

References

[1] P. Pentti and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, Wiley, 1994.

[2] P. Pentti, A weighted non-negative least squares algorithm for three-way ‘PARAFAC’ factor analysis, *Chemometrics and Intelligent Laboratory Systems*, 38(2):223–242, Elsevier, 1997.

[3] P. Comon. Independent component analysis, a new concept?, *Signal processing*, 36(3):287–314, Elsevier, 1994.

[4] A. Hyvärinen, E. Oja, Independent component analysis: Algorithms and applications, *Neural Networks*, 13 (4–5): 411–430, Elsevier, 2000.

[5] D. Daniel. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755): 788–791, 1999.

[6] C. Boutsidis, E. Gallopoulos, SVD-based initialization: A head start on nonnegative matrix factorization, *Pattern Recognition*, 41(4): 1350–1362, Elsevier, 2008.