PCA and PMF based methodology for air pollution sources identification and apportionment

Marie Chavent, Hervé Guegan, Vanessa Kuentz, Brigitte Patouille, Jérôme Saracco

PCA and PMF based methodology for air pollution sources identification and apportionment*

Marie Chavent[1,2†], Hervé Guégan[3], Vanessa Kuentz[1,2], Brigitte Patouille[1], Jérôme Saracco[1,2,4]

[1] Université de Bordeaux, IMB, CNRS, UMR 5251, France
[2] INRIA Bordeaux Sud-Ouest, CQFD team, France
[3] ARCANE-CENBG, Gradignan, France
[4] Université Montesquieu - Bordeaux IV, GREThA, CNRS, UMR 5113, France

Abstract

Air pollution is a wide concern for human health and requires the development of air quality control strategies. In order to achieve this goal pollution sources have to be accurately identified and quantified. The case study presented in this paper is part of a scientific project initiated by the French Ministry of Ecology and Sustainable Development. For the following study measurements of chemical composition data for particles have been conducted on a french urban site. The first step of the study consists in the identification of the sources profiles which is achieved through Principal Component Analysis completed by a rotation technique. Then the apportionment of the sources is evaluated with a receptor modeling using Positive Matrix Factorization as estimation method. Finally the joint use of these two statistical methods enables to characterize and apportion five different sources of fine particulate emission.

Keywords: Pollution sources, Principal Component Analysis, Receptor modeling, Positive Matrix Factorization.

*Preprint of *Environmetrics*, Vol. 20, pp. 928-942.
†Correspondence to: M. Chavent, IMB, E-mail: chavent@math.u-bordeaux1.fr

1 Introduction

Particulate pollution, also known as particulate matter or PM, comes from various sources such as factory and utility smokestacks, vehicle exhaust, wood burning, mining, construction activity or agriculture. This air pollution is a complex mixture of extremely small particles and liquid droplets suspended in the air we breathe. High concentrations of particles have been found to present a serious danger to human health. Particles of special concern to the protection of lung health are those known as fine particles (PM2.5), less than 2.5 microns in diameter. Development of PM2.5 control strategies is then a wide preoccupation of environmental protection agencies. Since strategies to improve ambient air quality involve the reduction of emissions from primary sources, it is important to be able to identify and apportion the contributions of these sources.

Receptor modeling, using measurements of chemical composition data for particles on a sample site, is often a reliable way to provide information regarding source characteristics [1]. Some multivariate receptor models are based on the analysis of the correlations between measured concentrations of chemical species, assuming that highly correlated compounds come from the same source. One commonly used multivariate receptor model is Principal Component Analysis (PCA) [3], successfully applied to identify sources in several studies. However PCA is not a convenient tool for quantifying sources contributions. Therefore specific methods such as Positive Matrix Factorization (PMF) [5], have been specifically developed in order to address this problem.

The case study presented in this paper is a statistical part of the scientific

program PRIMEQUAL[1], initiated by the MEDD[2] and the ADEME[3], about atmospheric pollution and its impact. We propose and apply a methodology for determining particulate emission sources and their concentrations at the urban site of Anglet located in the south west of France. The following three step process has been implemented:

1. PM2.5 were collected with sequential fine particle samplers on the receptor site and the chemical composition of each sampler was measured with PIXE (Particle Induced X-ray Emission) method. After several pre-treatments a data matrix of chemical compounds concentrations in each sampler was selected.

2. PCA was applied to this data matrix and the standardized principal components were rotated, in order to identify possible sources.

3. PMF was applied to the same data matrix and the results were normalized in order to find components with physical interpretations (concentration of each source in each particle sampler).

Steps 2 and 3 are independent but results of step 2 will be used to validate results of step 3.

2 Data

The air pollution receptor modeling (n, p) data matrix consists of the measurements of p chemical species in n particle samplers. In this application, $n = 61$ samplers of PM2.5 were collected with sequential fine particle samplers by AIRAQ[4] in the french urban site of Anglet, every twelve hours,

[1]Projet de Recherche Interorganisme pour une MEilleure QUalité de l'Air à l'échelle Locale

[2]French ministry of Ecology and Sustainable Development

[3]French Environment and Energy Management Agency

[4]Réseau de surveillance de la qualité de l'air en Aquitaine

in December 2005. There are two samples per 24 hours: one for the day (7AM:7PM) and one for the night (7AM:7PM). The mass and volume, represented by the concentration C in ng/m^3, of each particle sampler were measured with the PIXE method by ARCANE-CENBG[5], as well as the concentrations of $p = 15$ chemical elements (Al, Si, P, S, Cl, K, Ca, Ti, Mn, Fe, Ni, Cu, Zn, Br, Pb). Table 1 gives a subset of the data in their initial form.

Date	C	Al	Si	...	K	Ca	...	Br	Pb
23-11-05 day	7300	92	75	...	163	35	...	7	10
23-11-05 night	9600	135	90	...	211	23	...	7	77
24-11-05 day	11000	175	137	...	241	69	...	8	19
24-11-05 night	5300	36	31	...	94	44	...	9	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24-12-05 day	21000	< 2	< 1	...	266	< 1	...	7	18
24-12-05 night	18100	18	< 1	...	307	< 1	...	7	19
25-12-05 day	23300	37	22	...	311	12	...	7	14
25-12-05 night	36100	< 2	< 1	...	277	< 1	...	10	19

Table 1: Subset of the original data table

First Ni and Ti elements which were frequently present in concentrations below the detection limits (BDL) were excluded and only 13 elements were selected. Then the few BDL data remaining in this selected data set were replaced by values corresponding to one-half of the appropriate analytical detection limit. Al, Si, S and Fe elements were respectively replaced by the compounds Al_2O_3, $Si0_2$, $S0_4$, Fe_20_3. Then the remaining concentration, called C_{org}, which was not measured with the previous compounds, was calculated for each particle sampler:

$$C_{org} = C - (Al_20_3 + Si0_2 + P + S0_4 + Fe_20_3 + Cl + K + Ca + Mn + Cu + Zn + Br + Pb).$$

Finally the (n, p) concentration matrix $\mathbf{X} = (x_{ij})$ used in the receptor model

[5]Atelier Régional de Caractérisation par Analyse Nucléaire Elémentaire - Centre d'Etudes Nucléaires de Bordeaux Gradignan

has $n = 61$ rows and $p = 14$ columns (Al_20_3, $Si0_2$, P, $S0_4$, Cl, K, Ca, Mn, Fe_20_3, Cu, Zn, Br, Pb, C_{org}). The coefficient x_{ij} is the concentration of the jth chemical compound in the ith sampler. One can observe that C_{org} represents the largest concentration in the particle samplers and then the largest part (almost all) of PM2.5. The discovery of its origin is a key point in the results.

3 Sources identification

In order to identify the sources of fine particulate emission we applied PCA to the concentration matrix $\mathbf{X}$ and completed it by an orthogonal rotation of the standardized principal components. Then we have associated groups of correlated chemical compounds to air pollution sources.

First we will give a short theoretical reminder of Factor Analysis with PCA estimation method. Then we will interpret the corresponding results on the air pollution data.

3.1 Factor Analysis with PCA estimation method

Notations. We consider a (n, p) numerical data matrix $\mathbf{X}$ where n objects are described on $p < n$ variables $x_1, \ldots, x_p$. We will note $\mathbf{x}_j$ a column of $\mathbf{X}$. Let $\tilde{\mathbf{X}} = (\tilde{x}_{ij})_{n,p}$ be the standardized data matrix with $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ where $\bar{x}_j$ and s_j are respectively the empirical mean and the standard deviation of x_j.

Let $\mathbf{R} = \tilde{\mathbf{X}}'\mathbf{M}\tilde{\mathbf{X}}$ be the empirical correlation matrix of $x_1, \ldots, x_p$, where $\mathbf{M} = \frac{1}{m}\mathbf{I}_n$ with $m = n$ or $n - 1$ depending on the choice of the denominator of s_j. The correlation matrix can also be written $\mathbf{R} = \mathbf{Z}'\mathbf{Z}$ with $\mathbf{Z} = \mathbf{M}^{1/2}\tilde{\mathbf{X}}$.

Let us denote by $r \leq p$ the rank of $\mathbf{Z}$ and consider the Singular Value

Decomposition (SVD) of $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' \tag{1}$$

where:

- Λ is the (r, r) diagonal matrix of the r nonnull eigenvalues λ_k, $k = 1, ..., r$, of the matrix $\mathbf{Z}'\mathbf{Z}$ (or $\mathbf{Z}\mathbf{Z}'$), ordered from largest to smallest;
- $\mathbf{U}$ is the (n, r) orthonormal matrix of the r eigenvectors $\mathbf{u}_k$, $k = 1, ..., r$ of $\mathbf{Z}\mathbf{Z}'$ associated with the first r eigenvalues;
- $\mathbf{V}$ is the (p, r) orthonormal matrix of the r eigenvectors $\mathbf{v}_k$, $k = 1, ..., r$ of $\mathbf{Z}'\mathbf{Z} = \mathbf{R}$ associated with the first r eigenvalues.

From the SVD of $\mathbf{Z}$ we deduce the following decomposition of $\tilde{\mathbf{X}}$:

$$\tilde{\mathbf{X}} = \mathbf{M}^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'. \tag{2}$$

Factor Analysis model. The basic idea underlying Factor Analysis (using correlation matrix) is that the p observed standardized variables $\tilde{x}_1, ..., \tilde{x}_p$ can be expressed, to the exception of an error term, as linear functions of $q < p$ unobserved variables or common factors $f_1, ..., f_q$. The observed standardized matrix $\tilde{\mathbf{X}}$ being given, factor analysis model can be expressed in its simplified form as:

$$\tilde{\mathbf{X}} = \mathbf{F}\mathbf{A}' + \mathbf{E}, \tag{3}$$

where $\mathbf{F}$ is the (n, q) matrix of unobserved values of the factors and $\mathbf{A}$ is the (p, q) matrix of unknown loadings providing the information relating the factors f_k to the original variables $x_1, \ldots, x_p$. The (n, p) matrix $\mathbf{E}$ is the rest of the approximation of $\tilde{\mathbf{X}}$ with $\hat{\tilde{\mathbf{X}}} = \mathbf{F}\mathbf{A}'$.

Several approaches were developed to estimate the model (principal factor, maximum likelihood ...) but PCA is often used in practice.

PCA. In PCA, when $q = r$, equation (2) is written:

$$\tilde{\mathbf{X}} = \mathbf{\Psi}\mathbf{V}' \tag{4}$$

where $\mathbf{\Psi} = \mathbf{M}^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}$ is the principal component scores matrix. The columns of $\mathbf{\Psi}$ are the r principal components $\boldsymbol{\psi}_k = (m\lambda_k)^{1/2}\mathbf{u}_k$, $k = 1, ..., r$. Since $\mathbf{U}$ and $\mathbf{V}$ are orthonormal we have $\boldsymbol{\psi}_k = \tilde{\mathbf{X}}\mathbf{v}_k$ and $\text{Var}(\boldsymbol{\psi}_k) = \lambda_k$.

Estimation of the factor model using PCA. When $q = r$ equation (2) is written:

$$\tilde{\mathbf{X}} = \mathbf{F}\mathbf{A}' \tag{5}$$

where $\mathbf{F} = \mathbf{M}^{-1/2}\mathbf{U}$ is the factor scores matrix and $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{1/2}$ is the loadings matrix. The columns $\mathbf{f}_k = m^{1/2}\mathbf{u}_k$ of the matrix $\mathbf{F}$ are realizations of the r factors f_k, $k = 1, ..., r$. The coefficient a_{jk} of the matrix $\mathbf{A}$ is equal to the empirical correlation between $\mathbf{x}_j$ and $\mathbf{f}_k$. Since $\mathbf{U}$ and $\mathbf{V}$ are orthonormal we have $\mathbf{f}_k = \lambda_k^{-1/2}\tilde{\mathbf{X}}\mathbf{v}_k = \lambda_k^{-1/2}\boldsymbol{\psi}_k$ for $k = 1, ..., r$ and $\text{Var}(\mathbf{f}_k) = 1$. Then $\mathbf{f}_k$ is also the standardized principal component $\boldsymbol{\psi}_k$.

When the user only retains the first $q < r$ eigenvalues of $\mathbf{\Lambda}$ the corresponding approximation of $\tilde{\mathbf{X}}$ in (3) is then:

$$\hat{\tilde{\mathbf{X}}}_q = \mathbf{F}_q\mathbf{A}_q'$$

where $\mathbf{F}_q$ and $\mathbf{A}_q$ are the matrices $\mathbf{F}$ and $\mathbf{A}$ reduced to their first q columns. $\mathbf{F}_q$ is then the matrix of the first q standardized principal components.

Rotation of the standardized principal components. Let $\mathbf{T}$ be an orthogonal transformation matrix corresponding to an orthogonal rotation of the q axes in a p-dimensional space: $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_q$.

The orthogonal rotation is applied to the standardized principal components:

$$\hat{\breve{\mathbf{X}}}_q = \mathbf{F}_q\mathbf{T}(\mathbf{A}_q\mathbf{T})'.$$

The q rotated standardized principal components $\breve{\mathbf{f}}_k^q$ are the q columns of the matrix $\breve{\mathbf{F}}_q = \mathbf{F}_q\mathbf{T}$. They have the property of being mutually orthogonal and of variance equal to 1.

In order to be able to interpret the $\breve{\mathbf{f}}_k^q$'s (also called rotated factors) let us remark that the coefficients $\breve{a}_{jk}^q$ of the matrix $\breve{\mathbf{A}}_q = \mathbf{A}_q\mathbf{T}$ are equal to the empirical correlations between the rotated factors $\breve{\mathbf{f}}_k^q$ and $\mathbf{x}_j$.

From a practical point of view the orthogonal transformation matrix $\mathbf{T}$ is defined in order to construct a matrix $\breve{\mathbf{A}}_q$ such that each variable x_j is clearly correlated to one of the rotated factor $\breve{\mathbf{f}}_{k^*}^q$ (that is $\breve{a}_{jk^*}^q$ close to 1) and not to the other rotated factors (that is $\breve{a}_{jk}^q$ close to 0 for $k \neq k^*$). The most popular rotation technique is varimax which seeks rotated loadings maximizing the variance of the squared loadings in each column of $\breve{\mathbf{A}}_q$.

3.2 Results

We applied the FACTOR procedure of SAS to the data matrix $\mathbf{X}$ introduced in section 2. The following options were used: METHOD=PRIN, ROTATE=VARIMAX and NFACTORS=5. The number $q = 5$ of factors was chosen both because it allows to explain 90,93% of the total variance and because decompositions in a larger number of factors did not give satisfactory interpretations. Table 2 gives the matrix $\breve{\mathbf{A}}_5$ of the loadings after rotation.

This matrix can be used to associate, when possible, sources with the rotated factors. Indeed we observe for each factor the strongly correlated coumpounds. For instance Zn and Pb are strongly correlated to $\breve{\mathbf{f}}_3^5$. Be-

Table 2: Correlations between the chemical compounds and the rotated factors

	$\breve{\mathbf{f}}_1^5$	$\breve{\mathbf{f}}_2^5$	$\breve{\mathbf{f}}_3^5$	$\breve{\mathbf{f}}_4^5$	$\breve{\mathbf{f}}_5^5$
Al_2O_3	**0.981**	0.087	-0.042	0.070	-0.038
SiO_2	**0.979**	0.012	-0.055	0.104	-0.074
P	**0.972**	0.090	-0.017	0.071	-0.092
SO_4	-0.028	0.765	0.247	0.180	-0.345
Cl	-0.153	-0.274	-0.136	-0.181	**0.879**
K	0.597	0.716	0.111	0.233	0.031
Ca	0.608	0.091	-0.113	0.560	0.272
Mn	-0.279	0.119	0.604	0.582	-0.238
Fe_2O_3	0.198	0.282	0.289	0.848	-0.112
Cu	0.213	0.359	0.161	0.816	-0.149
Zn	-0.029	0.053	**0.977**	0.129	-0.044
Br	0.490	0.615	0.097	0.281	0.392
Pb	0.004	0.163	**0.969**	0.126	-0.054
C_{org}	-0.018	0.893	0.021	0.222	-0.160

Table 3: Factor-source associations

Factor 1	Soil dust
Factor 2	Combustion
Factor 3	Industry
Factor 4	Vehicle
Factor 5	Sea

cause Zn and Pb are known to have industrial origin this rotated factor is associated to the industrial pollution source. The same way, the element Cl is strongly correlated to $\breve{\mathbf{f}}_5^5$ which is then associated with sea salt pollution. Possible associations between the five rotated factors and five pollution sources are given in Table 3.

In order to confirm these associations we have confronted the rotated factors $\breve{\mathbf{f}}_k^5$ with external parameters such as meteorological data (temperatures and wind directions) and the periodicity night/day of the sampling. The coefficient $\breve{f}_{ik}^5$ represents a "relative" contribution of the source k to the particle sampler i. Fig. 1(a) gives for instance the evolution of the relative contribution of the "vehicle" source associated with $\breve{\mathbf{f}}_4^5$. The night

samplers have been distinguished from the day ones, enabling to notice that the contribution of this source is stronger during the day than during the night. It then confirms that this source corresponds to vehicle pollution. The same way, Fig. 1(b) gives the evolution of the relative contribution of the "combustion" source associated with $\breve{\mathbf{f}}_2^5$. We can notice an increase in the contribution of this source in the middle of the sampling period, which corresponds to a decrease in the temperature measured on the sampling site, see Fig. 1(c). This confirms that this source corresponds to combustion and heatings pollution.

The identification of the sources using PCA is only the first step of a more complex process which consists in quantifying the sources. Although it is essential to identify the sources, the true challenge is to define, in percentage of total fine dust mass, the quantity of each of these sources.

4 Sources apportionment

In order to apportion the sources of fine particulate emission we have applied PMF to the concentration matrix $\mathbf{X}$ and then normalized the results to find components with physical interpretation.

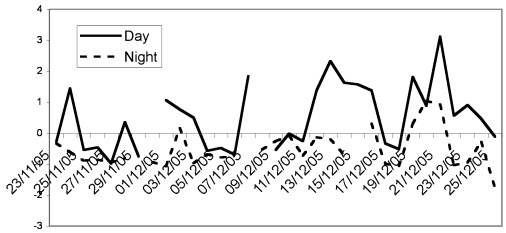
4.1 Receptor modeling with PMF estimation method

The basic problem is to estimate, from the data matrix $\mathbf{X}$, the number q of sources, their compositions and their contributions. To address this problem we consider the mass balance equation:

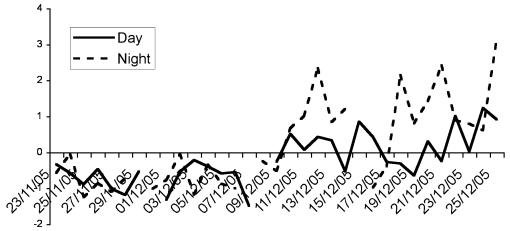
$$x_{ij} = \sum_{k=1}^{q} g_{ik} b_{jk} \tag{6}$$

where

(a) Evolution of the Factor 4 associated to cars pollution



(b) Evolution of the Factor 2 associated to heatings pollution



(c) Evolution of temperatures

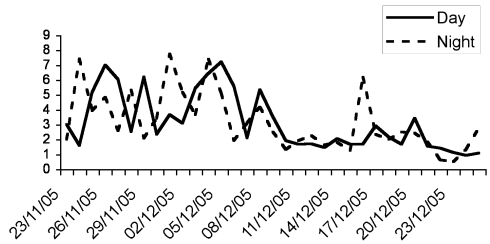


Figure 1: Evolution of factors 2 (a) and factor 4 (b) and evolution of the temperatures (c)

- g_{ik} is the concentration in particles from source k in the particle sampler i;

- b_{jk} is the mass fraction (percentage) of species j in source k.

In the receptor modeling vocabulary the b_{jk}'s are the sources compositions (or sources profiles) and the g_{ik}'s are the sources contributions. The product $g_{ik}b_{jk}$ is then the approximation of the concentration in the sampler i in particles from the jth species coming from the source k. Let m_{ijk} be the mass in the sampler i of species j from source k, and let m_{ik} be the mass in the sampler i from source k. Then $b_{jk} = \frac{m_{ijk}}{m_{ik}}$ is the percentage of species j emitted by source k when sampler i was collected. Since the mass fraction b_{jk} is independent from i the sources profiles are assumed to be constant during the sampling period.

In matrix form equation (6) can be written:

$$\mathbf{X} = \mathbf{G}\mathbf{B}' \tag{7}$$

where $\mathbf{G}$ is a (n, q) matrix of sources contributions and $\mathbf{B}$ is a (p, q) matrix of sources compositions. Approximations of $\mathbf{G}$ and $\mathbf{B}$ are obtained from the data matrix $\mathbf{X}$ and a previously selected number q of sources by applying the two following steps:

- *PMF step.* The matrix $\mathbf{X}$ is factorized in a product $\mathbf{H}\mathbf{C}'$ of rank q under constraints of positivity of the coefficients. This condition is required by physical reality of non negativity of sources compositions and contributions: $g_{ik} \geq 0$ and $b_{jk} \geq 0$.

- *Scaling step.* The columns of the approximations $\hat{\mathbf{H}}$ and $\hat{\mathbf{C}}$ obtained in the previous step are scaled in order to get the approximations $\hat{\mathbf{G}}$

and $\hat{\mathbf{B}}$. The scaling coefficients are defined to fulfill other physical constraints of the sources compositions and contributions.

Let us assume now that there are at least as many species as sources.

PMF step. Given a matrix $\mathbf{X}$ and a previously selected rank $q \leq p$ the aim of PMF (or Non Negative Matrix Factorization) is to approximate $\mathbf{X}$ by a product of two matrices $\mathbf{HC}'$ (with $\mathbf{H}$ of dimension (n, q) and $\mathbf{C}$ of dimension (p, q)) subject to $h_{ik} \geq 0$ and $c_{jk} \geq 0$. Matrices $\mathbf{H}$ and $\mathbf{C}$ are obtained by minimization of a least squares function $Q(\mathbf{H}, \mathbf{C})$ under constraints of positivity.

When the constraints of positivity are ignored the ordinary SVD of $\mathbf{X}$, that is $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'$ with $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$, provides a sequence of approximations $\mathbf{HC}'$ of rank $q = 1, \ldots, r$ which minimizes the square of the euclidean norm of the residual matrix $\mathbf{L} = \mathbf{X} - \mathbf{HC}'$:

$$Q_1(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n}\sum_{j=1}^{p} l_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk})^2. \tag{8}$$

In (8) the rows and the columns of $\mathbf{X}$ have the same weight. Let us denote now ω_i the weight of the ith row and ϕ_j the weight of the jth column of $\mathbf{X}$. Let Ω and Φ be two diagonal matrices respectively with elements ω_i, $i = 1, \ldots, n$ and ϕ_j, $j = 1, \ldots, p$. The generalized SVD of $\mathbf{X}$, that is $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'$ with $\mathbf{U}'\Omega\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}'\Phi\mathbf{V} = \mathbf{I}_r$, provides a sequence of approximations $\mathbf{HC}'$ which minimizes:

$$Q_2(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n}\sum_{j=1}^{p} \omega_i\phi_j(x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk})^2. \tag{9}$$

Note that this generalized SVD of $\mathbf{X}$ is obtained by finding the ordinary SVD of $\Omega^{1/2}\mathbf{X}\Phi^{1/2}$.

A third type of approximation of $\mathbf{X}$ is defined by minimizing:

$$Q_3(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij}(x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk})^2 \tag{10}$$

but this approximation can not be obtained by SVD unless the w_{ij}'s can be written as products $w_{ij} = \omega_i \phi_j$. Gabriel and Zamir (1979) suggest a number of ways in which special cases of this weighted least squares analysis may be used.

The PMF algorithm developed by Paatero and Tapper (1994) in the context of receptor modeling minimizes (10) with $w_{ij} = 1/\sigma_{ij}^2$. The coefficient σ_{ij} is a measure of uncertainty of the observation x_{ij}. Given the σ_{ij}'s this method searches $\mathbf{H}$ and $\mathbf{C}$ minimizing:

$$Q_4(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk}}{\sigma_{ij}} \right)^2 \tag{11}$$

subject to $h_{ik} \geq 0$ and $c_{jk} \geq 0$.

Polissar et al. (1998) propose several definitions for calculating the σ_{ij}'s from a matrix $\mathbf{X}$ of chemical species concentrations. The one used in the PMF program of the US Environment Protection Agency[6] is the following:

$$\sigma_{ij} = \begin{cases} 2LD & \text{if} \quad x_{ij} \leq LD, \\ \sqrt{(\theta_j x_{ij})^2 + LD^2} & \text{if} \quad x_{ij} > LD, \end{cases} \tag{12}$$

where LD is the limit of detection for the jth species and θ_j is a percentage of uncertainty associated with the jth species. One can note the subjectivity of this definition which changes from an article to another using this PMF method for sources apportionment.

In this case study we have made a different choice for the σ_{ij}'s. Indeed, dealing with variables measured on very different scales is a problem when approximating $\mathbf{X}$ globally on all the variables. Minimizing the unweighed

[6]E.P.A. PMF 1.1 Users's guide, http://www.epa.gov/heasd/products/pmf/pmf.htm

quadratic error Q_1 in (8) gives better approximations for the columns of $\mathbf{X}$ corresponding to variables with large dispersion. Hence we have chosen to use Q_4 with $\sigma_{ij} = s_j$, the empirical standard deviation of the jth variable.

Scaling step. Let $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{C}}'$ be the product calculated by PMF. Since $\hat{x}_{ij} = \sum_{k=1}^{q} \hat{h}_{ik}\hat{c}_{jk} = \sum_{k=1}^{q} \hat{h}_{ik}\frac{\beta_k}{\beta_k}\hat{c}_{jk}$ the matrix $\hat{\mathbf{X}}$ can be written:

$$\hat{\mathbf{X}} = \breve{\mathbf{H}}\breve{\mathbf{C}}' \tag{13}$$

with $\breve{h}_{ik} = \hat{h}_{ik}\beta_k$ and $\breve{c}_{jk} = \frac{\hat{c}_{jk}}{\beta_k}$.

The aim of scaling is then to define the scaling constants β_k, $k = 1, \ldots, q$ such that $\breve{\mathbf{H}}$ and $\breve{\mathbf{C}}$ verify the physical conditions of the matrices $\mathbf{G}$ and $\mathbf{B}$ of the mass balance equation (6). We are going to use the two following conditions.

- Let γ_i be the concentration in the ith sampler:

$$\gamma_i = \sum_{k=1}^{q} g_{ik}. \tag{14}$$

  In other words the sum of the concentrations of the sources adds up to the total concentration of the samplers.

- If the sum of the concentrations of the observed species adds up to (resp. is lower than) the total concentration of the samplers, then the sum of all species in each source profile is equal to (resp. lower than) unity:

$$\begin{cases} \sum_{j=1}^{p} b_{jk} = 1 & \text{if } \sum_{j=1}^{p} x_{ij} = \gamma_i, \\ \sum_{j=1}^{p} b_{jk} < 1 & \text{otherwise.} \end{cases} \tag{15}$$

First we consider the case where $\forall i, \sum_{j=1}^{p} x_{ij} = \gamma_i$. From the physical constraints (15) the scaling coefficients β_k can be calculated in two different ways.

- Directly from $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{C}}'$ we get

$$\gamma_i = \sum_{j=1}^{p} x_{ij} = \sum_{j=1}^{p} \left(\sum_{k=1}^{q} \hat{h}_{ik}\hat{c}_{jk} + \hat{l}_{ij} \right) = \sum_{k=1}^{q} \hat{h}_{ik} \left(\sum_{j=1}^{p} \hat{c}_{jk} \right) + \sum_{j=1}^{p} \hat{l}_{ij},$$

we can set

$$\hat{\beta}_k = \sum_{j=1}^{p} \hat{c}_{jk}. \tag{16}$$

Then we have the following approximations $\hat{\mathbf{G}}$ and $\hat{\mathbf{B}}$ of $\mathbf{G}$ and $\mathbf{B}$: $\hat{b}_{jk} = \frac{\hat{c}_{jk}}{\sum_{j=1}^{p} \hat{c}_{jk}}$ which satisfies constraint (15), and $\hat{g}_{ik} = \hat{h}_{ik}(\sum_{j=1}^{p} \hat{c}_{jk})$ which satisfies constraint (14) with an error sum of squares equal to $\sum_{i=1}^{n}(\sum_{j=1}^{p} \hat{l}_{ij})^2$.

- Considering the linear approximation of γ_i

$$\gamma_i = \sum_{k=1}^{q} \beta_k \hat{h}_{ik} + e_i \tag{17}$$

we search $\beta = (\beta_1, ..., \beta_q)'$ minimizing the error sum of squares:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\gamma_i - \sum_{k=1}^{q} \beta_k \hat{h}_{ik})^2.$$

A wellknown solution to this minimization problem is:

$$\hat{\hat{\beta}} = (\hat{\mathbf{H}}'\hat{\mathbf{H}})^{-\mathbf{1}}\hat{\mathbf{H}}\gamma \tag{18}$$

with $\gamma = (\gamma_1, \ldots, \gamma_n)'$. The corresponding approximations $\hat{\hat{\mathbf{G}}}$ and $\hat{\hat{\mathbf{B}}}$ are such that $\hat{\hat{b}}_{jk} = \frac{\hat{c}_{jk}}{\hat{\hat{\beta}}_k}$ does not satisfy constraint (15), and $\hat{\hat{g}}_{ik} = \hat{h}_{ik}\hat{\hat{\beta}}_k$ satisfies constraint (14) with an error sum of squares equal to $\sum_{i=1}^{n}(\hat{e}_{ij})^2$ with $\hat{e}_{ij} = \gamma_i - \sum_{k=1}^{q} \hat{\hat{\beta}}_k \hat{h}_{ik}$.

Since $\sum_{i=1}^{n}(\hat{e}_{ij})^2$ is the minimum error sum of squares we have:

$$\sum_{i=1}^{n} (\hat{e}_{ij})^2 \leq \sum_{i=1}^{n} (\sum_{j=1}^{p} \hat{l}_{ij})^2. \tag{19}$$

Obviously, in case of equality in (19), we get $\hat{\beta} = \hat{\hat{\beta}}$ which means that we have simultaneously the sum of all species in each source profile which is unity and the sum of the concentrations of the sources best fitting (for the least sum of squares error) the total concentration of the samples.

Comparing $\sum_{i=1}^{n}(\hat{e}_{ij})^2$ with $\sum_{i=1}^{n}(\sum_{j=1}^{p}\hat{l}_{ij})^2$ or equivalently comparing $\hat{\beta}$ with $\hat{\hat{\beta}}$ provides a confirmation that the information given by the columns of $\hat{\mathbf{H}}$ are coherent with the physical model we try to approximate. It is hence a first good way to validate the results.

If we consider now the case where $\sum_{j=1}^{p} x_{ij} < \gamma_i$ the scaling coefficients can not be directly calculated from $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{C}}'$ since $\sum_{j=1}^{p} b_{jk} < 1$. They are then evaluated with (18).

A second way to validate the results is based on the regression of γ_i either on $\hat{\hat{\gamma}}_i = \sum_{k=1}^{q}\hat{\hat{\beta}}_k\hat{h}_{ik}$ or $\hat{\gamma}_i = \sum_{k=1}^{q}\hat{\beta}_k\hat{h}_{ik}$, depending on the choice of the scaling coefficients.

4.2 The results

We have applied the PMF algorithm to the concentration matrix $\mathbf{X}$ with $q = 5$ sources. The choice of the number of sources rises from the PCA results. The introduction of C_{org} yields $\sum_{j=1}^{p} x_{ij} = \gamma_i$, then the scaling coefficients $\hat{\beta}_k$ have been calculated from (16).

We thus have the following numerical results:

- the $(61, 5)$ matrix $\hat{\mathbf{G}}$ of the approximated concentrations of the 5 sources in the 61 samples,

- the $(14, 5)$ matrix $\hat{\mathbf{B}}$ of the approximated compositions (profiles) of the 5 sources on the 14 compounds.

Quality of the model approximation. Since we are in the case where $\forall i, \sum_{j=1}^{p} x_{ij} = \gamma_i$ we can evaluate the quality of the approximation of $\mathbf{X}$ by $\hat{\mathbf{G}}\hat{\mathbf{B}}'$ using the two methods mentioned above. We can compare the scaling coefficients $\hat{\beta}_k$ and $\hat{\hat{\beta}}_k$. Table 4 clearly shows that the $\hat{\beta}_k$'s are close to the $\hat{\hat{\beta}}_k$'s. Moreover figure 2 also shows a good fitting of the γ_i's by the $\hat{\gamma}_i$'s.

Table 4: The scaling coefficients

	$\hat{\beta}_k$	$\hat{\hat{\beta}}_k$
$k=1$	147.1	158.6
$k=2$	91.5	89.6
$k=3$	73.5	76.8
$k=4$	251.9	251.9
$k=5$	51.1	73.2

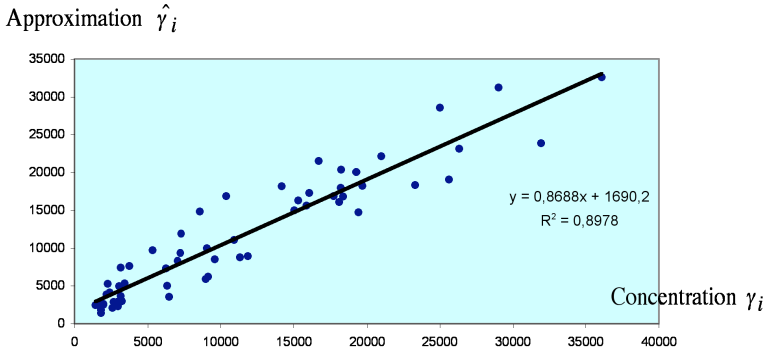


Figure 2: Adjustment of γ by $\hat{\gamma}$.

Sources identification. In practice the knowledge of $\hat{\mathbf{G}}$ and $\hat{\mathbf{B}}$ does not give direct indications on the nature of the sources. To try to discover the nature of the five sources we want to calculate their relative contribution to each of the 14 chemical compounds. In order to do that we need to work with the masses instead of the concentrations. Then we calculate, from $\hat{\mathbf{G}}$,

the approximation of the total mass of particulate emitted from source k in the 61 samplers. This mass is multiplied by $\hat{b}_{jk}$ hence resulting in the percentages reported in Table 5.

Table 5: Relative contributions of the sources to the chemical compounds

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Al_2O_3	100.0	0.0	0.0	0.0	0.0
$Si0_2$	100.0	0.0	0.0	0.0	0.0
P	81.5	0.5	3.9	8.2	6.0
SO_4	4.5	9.5	10.7	67.9	7.5
Cl	0.0	0.0	0.0	0.0	100.0
K	38.8	0.0	4.4	56.7	0.2
Ca	42.0	39.6	0.0	0.0	18.4
Mn	0.0	54.9	33.1	8.5	3.5
Fe_2O_3	19.0	59.2	14.4	7.4	0.0
Cu	18.5	56.8	9.1	15.6	0.0
Zn	9.0	0.5	87.5	0.0	3.1
Br	19.4	12.1	5.7	33.4	29.4
Pb	10.7	0.0	81.4	7.9	0.0
C_{org}	0.0	8.0	0.0	92.0	0.0

Table 5 is used to identify the nature of the sources. For instance Al_20_3 and SiO_2 are emitted exclusively by source 1. Because Al_20_3 and SiO_2 are known to have natural origin this source is associated to the soil dust pollution source. We proceed the same way for the other sources. We deduce possible identifications of the five pollution sources, see Table 6.

Table 6: Receptor model sources identification

$k=1$	Soil dust
$k=2$	Vehicles
$k=3$	Industry
$k=4$	Combustion
$k=5$	Sea

One can notice that the sources identified in Table 6 are the same than those found with PCA in Table 3. To verify the coherence of these sources identifications we have calculated, in Table 7, the correlations between the

factors (the columns of $\breve{\mathbf{F}}_5$) and the sources obtained by receptor modeling (the columns of $\hat{\mathbf{G}}$). We observe that the factors match well with the receptor model sources.

Table 7: Correlations between the sources of the receptor model and the factors of PCA after rotation

	Source 1	Source 2	Source 3	Source 4	Source 5
Factor 1	**0.98**	-0.18	-0.11	-0.02	-0.18
Factor 2	0.11	0.12	0.06	**0.95**	-0.30
Factor 3	-0.05	-0.09	**0.98**	0.02	-0.15
Factor 4	0.12	**0.96**	0.10	0.11	-0.22
Factor 5	-0.02	-0.13	-0.10	-0.27	**0.88**

Sources descriptions. The matrix $\hat{\mathbf{B}}$ of the sources profiles is reported in Table 8. We notice that, according to these profiles, C_{org}, which represents almost the total concentration in PM2.5, is only emitted by the Vehicle and Combustion sources.

Table 8: The sources profiles

	Soil dust	Vehicles	Industry	Combustion	Sea
Al_2O_3	41.6	0.0	0.0	0.0	0.0
$Si0_2$	18.5	0.0	0.0	0.0	0.0
P	6.2	0.0	0.7	0.0	0.6
SO_4	10.1	15.3	59.6	12.2	22.6
Cl	0.0	0.0	0.0	0.0	74.5
K	12.9	0.0	3.6	1.5	0.1
Ca	2.4	1.6	0.0	0.0	1.4
Mn	0.0	0.2	0.3	0.0	0.0
Fe_2O_3	6.7	15.0	12.7	0.2	0.0
Cu	0.3	0.7	0.4	0.0	0.0
Zn	0.7	0.0	16.3	0.0	0.3
Br	0.2	0.1	0.2	0.0	0.5
Pb	0.3	0.0	6.2	0.0	0.0
C_{org}	0.0	67.1	0.0	85.9	0.0

Sources apportionments. From matrix $\hat{\mathbf{G}}$ of the source contributions we can deduce some interesting comments. First we can focus on the relative contribution of each source in each particle sampler. For instance Figure 3 represents the relative contributions of the Combustion source in the 61 particle samplers. We can notice the increase in the percentage of this source in the second period of sampling, corresponding to a decrease in the temperature (see Fig. 1(c)).

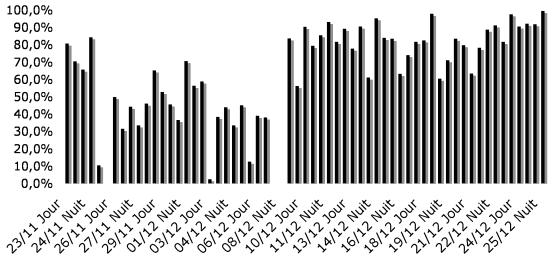


Figure 3: Relative contribution of the source Combustion to the samples.

We can also focus on the contribution of the sources to the PM2.5 dust contamination during the sampling period. Figure 4 shows the predominance of the Combustion source during this winter sampling period.

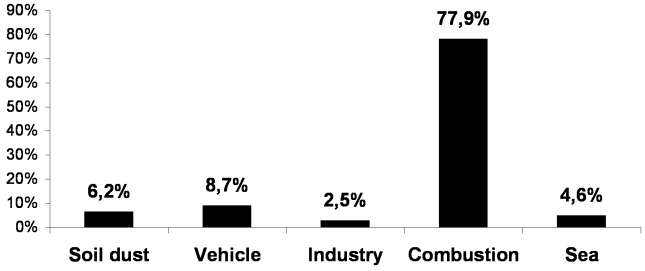


Figure 4: Global sources contributions to the PM2.5 dust contamination.

5 Conclusion

We propose in this case study a methodology for identifying and apportionning air pollution sources in a French urban site. The first step consists in Factor Analysis followed by a rotation technique and enables to identify the profiles of five principal sources: soil dust, vehicles, industry, combustion and sea. Then a receptor modeling approach, based on Positive Matrix Factorization, is used to evaluate their contributions to the fine particles dust contamination. Thus we highlight, during winter, the predominance of combustion source over dust pollution. The interest of the approach lies in the fact that we do not use prior knowledge on the sources (number, nature, profiles), which means that this work can be applied to more complex sampling site. Finally this methodology is not specific to pollution and can be used for other sources detection problems.

References

[1] Hopke, P.K., 1991. Receptor Modeling for Air Quality Management. Elsevier, Amsterdam.

[2] Jianhang, L., Laosheng, W., 2004. Technical details and programming guide for a general two-way positive matrix factorization algorithm. Journal of Chemometrics 18, 519-525.

[3] Jolliffe, I.T., 2002. Principal Component Analysis. Springer Verlag, New York.

[4] Gabriel, K.R., Zamir, S. 1979. Lower Rank Approximation of Matrices by Least Squares with any Choice of Weights, Technometrics 21, 489-498.

[5] Paatero, P., Tapper, U., 1994. Positive Matrix Factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111-126.

[6] Polissar, A.V., Hopke, P.K., Malm, W.C., Sisler, J.F., 1998. Atmospheric Aerosol over Alaska:2. Elemental Composition and Sources. Journal of Geophysical Research 103: 19,045-19,057