Non-negative matrix factorization for the analysis of particle number concentrations: Characterization of the temporal variability of sources in indoor workplace

Rachid Ouaret, Anda Ionescu, Olivier Ramalho

Non-negative Matrix Factorization for the analysis of particle number concentrations: characterization of the temporal variability of sources in indoor workplace.

R. Ouaret[a,*], A. Ionescu[b], O. Ramalho[c]

[a] *Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, 4 Allée Emile Monso, 31432 Toulouse, Cedex 4, France*

[b] *University Paris-Est, Center for Study and Research on Thermics, Environment and Systems (CERTES, EA3481), University Paris-Est, 61, avenue du Général de Gaulle, Créteil Cedex, 94010, France*

[c] *French Indoor Air Quality Observatory, Scientific and Technical Center for Building (CSTB), University Paris-Est, 84 Avenue Jean Jaurès, Champs-sur-Marne, 77420, France*

Abstract

The temporal variability of indoor Particle Number (PN) concentrations, their determinants and their relative contributions in an occupied workspace were investigated. The presented study is based on the receptor modeling approach, focusing on Non-negative Matrix Factorization (NMF) to provide new insights on the source time variability. Continuous size distribution from 0.3 µm to 20 µm were collected with a short time step sampling (1 min) over six months in 2015. The measurements were made inside and outside an open-plan office occupied by 6-8 persons. NMF distinguished five major patterns obtained from PN concentrations time series. The apportionment results were expressed as source diurnal profiles and strengths by relating the obtained source contributions to the source information provided by the office occupancy and natural ventilation (the effect of opening windows). Factor 2 contributes to 75% of the total contributions for finer size fraction (< 0.5 µm). Combining NMF results with indoor occupancy and windows states successfully demarcated the main sources of fluctuation. The diurnal profiles of the third factor (F3) and $PN_{0.9-1.8}$ concentrations time series are very similar ($r = 0.95$). The diurnal variation of factor 1 is very similar to that observed for CO_2 variations and $PN_{6.25-12.5}$ time series. Coarse particles (> 17.5 µm) are associated with the $4th$ factor. The latter does not contribute to any of the other particle ranges. The NMF factors interpretation was supported by correlation analysis and statistical tests, as well as by temporal variation comparison.

Keywords: Non-negative Matrix Factorization (NMF), Temporal source apportionment, Particle Number (PN) concentration, Open-plan Office.

1. Introduction

Nowadays, it is becoming increasingly evident that indoor environments play a critical role to understand and assess total human exposure to air pollution. Indoor air quality became a matter of particular interest for the following three main reasons: (*i*) people spend about 85% of their time indoors, (*ii*) indoor pollutant concentrations can be significantly higher from those outdoors

*Corresponding author
Email address: rachid.ouaret@toulouse-inp.fr (R. Ouaret)

Preprint submitted to Elsevier

and (*iii*) many potentially health-hazardous pollutants are emitted from indoor sources [40]. The assessment of the indoor source variability is now necessary for the design and implementation of effective control strategies. Airborne Particulate Matter (PM) is one of the major types of contaminants in indoor air due to their ubiquitous occurrence and toxicity [42, 38]. Particulate matter is a mixture of many different chemical species and the assessment of human exposure to PM requires some knowledge about the sources, and the determination of the time variability of the PM time series.

PM represent a complex mixture of organic and inorganic species which vary in size, composition, and origin. Regarding particle size, PM is usually classified according their aerodynamic diameter such as coarse (2.5 µm-10 µm), fine (<2.5 µm), and ultra-fine particles (< 0.1 µm) [12]. Since different particles originate from different sources and their concentration in a given place is influenced by different factors, scientific knowledge about the occurrence, strength, and temporal variability of the sources is required to reduce the indoor exposure impact. Several studies focused the extent to which human exposure to outdoor PM occurs indoors [64, 41]. From outdoors to indoors, PM enters by infiltrating through cracks and gaps in the building envelope, via natural or mechanical ventilation [11, 26]. Consequently, the temporal patterns of indoor PM concentrations are resulting from both indoor and outdoor sources variability which constantly varies over time. In addition, the PM temporal patterns are influenced by many factors: *(i)* penetration and ventilation efficiency, *(ii)* indoor occupancy and occupants' activity, and *(iii)* building's volume and interaction with surfaces.

Important indoor sources of PM include, but not limited to, household chores (combustion, candles, and cooking, ...) [1, 26] and different human activities [18]. In indoor offices, particles also can be generated from some equipment, such as copier and printers machines, computers, and other electronic devices [39, 9, 56]. Furthermore, many studies claim that particle re-suspension from walking is an important indoor source of PM [53, 57]. The source emissions and sinks are generally variable and depend on many environmental factors, such as temperature, humidity, and air exchange rates. Particle deposition and resuspension within the microenvironments are mechanisms that can extremely change the time variability of the indoor PM concentration. Overall, as detailed in [40], three main factors determine indoor pollutant variations in occupied spaces: (*i*) properties of pollutants, (*ii*) occupants' behavior and (*iii*) building characteristics. In the case of the testing chambers or experimental automated houses, these parameters are usually well defined [68, 61].

To the best of our knowledge there are few studies in the literature focusing on the impact of the occupancy and windows state in offices for a long period with a short time step.

The originality of the study presented in this paper is due to the following features: it is based on a long-term monitoring campaign in real conditions; parameters are monitored with a short time step; it contributes to indoor source identification taking into account the occupancy and the window opening state.

The objective of this work is to get a better understanding of the temporal variability of PM's and their determinants, *i.e.* the influencing factors, as well as their relative contribution to PM indoor concentrations. The measurement campaign was carried out in an occupied open-plan office.

By retracing pollutants to their origins, emission sources variability can be characterized. The source identification and the assessment of their relative contribution to total indoor exposures can provide valuable information for reducing their emission with the aim to protect human health.

The practice of deriving information about pollution sources and the amount they emit is called source apportionment [5].

In principle, there are two basic approaches used in the environmental field to perform a source apportionment analysis: source-oriented models (dispersion model) and receptor-oriented models. Dispersion models require the knowledge of emission rates and dispersion factors together with local topography and meteorology for the estimation of source impacts [16, 22]. It is a direct modeling (from sources to receptors). By contrast, receptor modeling, which is an inverse one (from receptor to sources), is based on the mass conservation and it can be used to identify and apportion sources of contaminants in the air [23]. The main idea in the latter is to solve a mass balance equation using multivariate factor analysis. Thus, the receptor models are used to estimate the contribution of different sources to ambient PM concentrations based on PM measurements and subsequent chemical analysis [49]. Exogenous variables can be used for interpretation.

Indoor environments are modeled sometimes by means of experimental chambers, where climatic parameters are controlled. In this case, physical models can be developed easier than for the real environments. Indeed, in the case of the real buildings the development of such models is more complicated for the following reasons: difficulty in measuring source emissions [58]; unknown or changing airflow patterns [33, 10]; difficulty to model room configuration and lack of information about occupancy and its impact on the indoor environment.

The advantage of receptor models in occupied spaces is that this type of models doesn't require all the information mentioned previously, as input. They are based on the decomposition of measured concentration signals without predetermining the pollutant transformation, transportation, and sink processes [63].

In recent years, there have been tremendous advances in the development of source apportionment techniques based on statistical analysis, particularly the use of Factor Analysis (FA) for temporally varying constituent concentrations [5, 22, 23, 67]. All these techniques are based on the factorization of the initial matrix (database of concentrations in our case) in two matrices: matrix of source profiles and matrix of source contributions, under the constraint of the RMSE minimization and by imposing different constraints, such as: decorrelation (Principal Component Analysis), statistical independence (Independent Component Analysis), non-negativity (Non-negative Matrix Factorization and Positive Matrix Factorization).

Positive Matrix Factorization (PMF), developed by Paatero et al. 1994 [47] is widely used for source apportionment in environmental applications such as atmospheric pollution. This method imposes positivity constraints in the factor computational process by calculating dominant positive factors based on measurements. The factorization is achieved without detailed prior knowledge of source profiles or chemical fingerprints. In contrast to many other linear factorizations such as (PCA) and Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF) [31, 32], as well as PMF, makes positive latent structure explicit.

The Principal Component Analysis (PCA) method is one of the most used tools in factorial methods. In indoor environment, we can find hybridization with other linear methods (regressions, ...) to quantify the factors [3].

In the same vein, we have used NMF and simulated several factorizations, which are obtained by a group of initialization-computation algorithms. For this reason, the different extensions on NMF comprising the PMF method which is considered as a special case [59].

The NMF factors were analyzed and connected with their influencing parameters, such as opening windows and occupancy rate of the open-plan office. In other words, we identify patterns that shared by several sources. To date, few research studies have been conducted to identify temporal patterns of different sources. Most of them focus on the chemical fingerprint. The contribution of this study lies in the characterization of the time variability of the most frequent source-patterns in an occupied open office. The interpretation of NMF factors and time variability patterns were supported by correlation analysis as well as by temporal variation visualization.

2. Materials and Methods

2.1. Materials

2.1.1. PN concentration monitoring

Particle Number concentrations (PN) were acquired continuously using an optical particle counter, a GRIMM 1.108 Dust Monitor (Grimm Technologies, Inc., Douglasville, GA, USA). This monitor counts the number of particles with a diameter within the range of 0.3 to 20 µm, and more in detail in 15 different size channels limited by: 0.3, 0.4, 0.5, 0.65, 0.8, 1.0, 1.6, 2.0, 3.0, 4.0, 5.0, 7.5, 10, 15 and 20 µm diameters (Figure 1, sensor's location). The flow rate of the instrument is 1.2 L/min^{-1}. It can count particles up to 2000 particles.cm^{-3} without coincidence effects with a sensitivity of 0.001 particles.cm^{3} and a reproducibility of 2%. The optical counter uses two laser powers to perform its measurements. Between 0.3 and 2 µm, the high laser power is used. Between 2 and 20 µm, the lowest laser power is used. The measurement at 2 µm is carried out twice at high and low laser power, the final measurement is the average of the two values obtained. During the calibration of each device, the maximum error tolerated is 10% between 0.3 and 2 µm and 20% between 2 and 20 µm.

The period of investigation covers 6 months with a measurement interval of 1 minute. The indoor PN concentrations were monitored in an open-plan office occupied by 6-8 persons. Outdoor PN concentrations were collected with the same type of instrument,which housed in a dedicated enclosure outdoors on the roof of the building. Particle number concentrations are recorded every minute in a memory card with an autonomy of more than 45 days and the data is recovered (approximately) every 15 days.

The monitoring period runs from January, 1^{st}, 2015, to June, 30^{th}, 2015 representing 4 344 hours. In our experiment, the database called later raw data matrix $\boldsymbol{X}$, consists of the number of particles for the 15 fractions (size ranges) (15 variables #.cm^{-3}) and 260580 rows (minutes). In the case of this study, there are 485 missing values scattered (randomly). So, about 0.0018% of missing values have been detected. The missing data are imputed using interpolation of the available data or by simple calculation of the medians if there are less than 5 consecutive missing values. For particle measurements, we rarely find more than 30 consecutive minutes of missing values.

2.1.2. Climatic data, occupancy and window opening state

Meteorological, windows opening states and occupancy parameters were collected simultaneously in the same open-plan office.

The presence of occupants is estimated using motion detectors, while the state (opened/closed) of the windows and doors are recorded by contactors. The measuring device has three motion detectors, two contactors on two doors, and five contactors on windows facing exterior environment.

Motion detectors are infrared passive sensors able to detect motion in a half-spherical field (cf. section 2.1.3). The door and windows contactors are soft-leaf switches that record a binary signal at each state change: *opened* → *closed* or *closed* → *opened*. We retrieve the amount of movement which tells us about occupancy, but we do not measure the number of occupants per minute. More specifically, when no movement of occupants is detected, no data is transmitted to the *CSTBox*

(Telemonitoring Box manufactured by CSTB). As soon as motion (of occupants) is detected, the information quantity is recorded during 10 seconds and a frame containing an array of 10 samples is sent to the *CSTBox*. From the recorded raw data of the motion quantity variations, data is transformed and re-coded to binary time series with 1-minute time step.

It should be noted that motion detection greatly underestimates the actual occupancy of the office space. The situation where the occupant is static is considered to be a vacancy state. This would reflect a high vacancy rate over the measurement period ($\approx$ 94%).

2.1.3. The open-plan environment

The concentration of indoor particles is highly variable and indoor-specific. The measurements were performed in a building (Scientific and Technical Center for Building (CSTB), cf. Figure A.1 in supplementary material) located in a suburban area, at 30 km East of Paris. The measurement campaign was conducted in an open-plan office space with a total area of $132\,\mathrm{m}^2$ and a volume of $364\,\mathrm{m}^3$. Figure 1 shows the plan of the open space located at the 2^{nd} floor. A virtual tour of the open space office is given in the supplementary material of this article.

Figure 1: Plan of the office space (132 m^2, 364 m^3). The configuration of the tables varies with the number of occupants. An example of motion detection is materialized by two pink half spheres.

This office has a permanent mechanical exhaust ventilation. A single flow ventilation system provides a constant air extraction rate of 228 $m^3.h^{-1}$) (measured in 2014 at $\pm$ 6%). The six extract units are located by a black cross on the Figure 1. On the opposite side, 10 air inlets are attached to the joinery of the 5 sliding windows. as described in [43]. In addition, natural ventilation is possible due to the windows. Consequently, the effective ventilation rate greatly depends on the opening state of the windows. The actions made by the occupants on the windows (opening / closing) depend on the indoor comfort as well as on the individual sensitivity of the occupants [62, 21]. These actions on windows together with the other environmental factors provide different time-scale variations that are difficult to model by pure physical laws.

The indoor materials were carpets on the floor, painted walls, and ceiling tiles. The furniture comprises typical L-shaped desks melamine-faced particleboard and aluminium closets. A laser multi-function copier was in use in the office plan. No specific major sources of particles, such as combustion, were identified in the office.

2.2. Methods

2.2.1. Related work

The number of existing receptor models applied to the environmental field is relatively large and it includes methods such as Principal Component Analysis [60, 20], Positive Matrix Factorization (PMF) [48, 47], Independent Component Analysis (ICA) [4] and recently Non-negative Matrix Factorization (NMF) in outdoor [17, 59, 27, 28, 37] and indoor [55, 46, 44, 45] environment.

Compared to the PCA and ICA, the PMF and NMF methods have the advantage of more realistic non-negative constraint on factor profiles and contributions. Following this criterion, this study used non-negativity for indoor source apportionment and particularly employed NMF that offers a wide range of algorithms and extensions compared to PMF.

NMF and its generalizations have been used for different purposes such as dimensionality reduction, feature extraction, clustering, blind source separation (BSS), and classification. In this paper, NMF was used only for BSS purposes.

In the environmental field, NMF is a new method of a wide range of receptor modeling. It is used to analyze the series of chemical concentration measurements and to find underlying explanatory sources [59, 28]. Some research papers proposed an extension of the standard NMF form to incorporate some physical proprieties. Limem et al. (2013) proposed an informed NMF with a specific parametrization which involves constraints about some known components of the factorization [35]. Plouvin et al. (2014) extended the latter work by adding some information provided by a physical dispersion model [50]. Limem et al. (2014) introduced the use of basic equality constraints and have derived theoretical expressions of constrained Weighted NMF (WNMF) to characterize industrial source apportionment of PM_{10} [34]. To take into consideration both constraints simultaneously, in the previous works [35, 34] a new parametrization was proposed by incorporating a new unconstrained matrix [13]. The update rules in [13] are based on the framework of the Split

Gradient Method of Lantéri *et al.* [29]. To identify distinctive PM_{10} patterns across Europe on the airborne data obtained from 1097 monitoring stations for 3 years, Žibert *et.al* [70] used NMF and the autocorrelation function (ACF) to enhance the interpretation of factors features. In the indoor air quality (IAQ) field, the positive (non-negative) matrix factorization is not such popular as for outdoor air quality. We find only one study assessing the use of NMF in IAQ, which is conducted by Rösch et al. in 2014 [55]. Also in the aircraft cabins context, PMF coupled with information related to VOC sources have been applied to identify the major VOC sources [65]. The study pointed out the importance of service and humans as major source (29%) of the total VOC emissions. Overall, there is not sufficient research to provide meaningful information on indoor source time variability. As the IAQ continues to face the realities of climate change [40], our understanding of temporal patterns of exposure is crucial and this research gap requires further investigation. Therefore, this study aims to address the elements of the research gap in indoor source identification. This paper is an extended version of the works presented in [44], [46] as well as in [45].

2.2.2. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a multivariate data analysis technique which is aimed to estimate physically meaningful latent components from non-negative data. Mathematically, the factorization is carried out by a linear superposition of non-negative basis components and non-negative weights.

NMF was initially introduced by Paatero's works [48, 47], which refer to the problem as Positive Matrix Factorization (PMF) that corresponds to the mass-balance model. The NMF has gained popularity by the works of Lee and Seung who presented multiplicative update algorithms for computing the NMF to optimize a cost function based on either a Euclidean distance measure or a generalized Kullback–Leibler divergence [30, 31, 32]. A multitude of NMF variants and generalizations is summarised in Cichocki's concise lecture note [15].

In this paper, a matrix is denoted with an uppercase bold letter, *e.g.*, $\boldsymbol{X}$, its elements with the corresponding lowercase letter, e.g., x_{it}, and a column vector in lowercase boldface, e.g., $\boldsymbol{x}_i$.

Given an input non-negative raw data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}_+^{I\times T}$ and a positive integer reduced rank J, $(J \leq \min(I,T))$, the non-negative matrix factorization problem consists in finding two non-negative matrices $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J] \in \mathbb{R}_+^{I\times J}$ and an encoding matrix $\boldsymbol{H} = h_{jt} \in \mathbb{R}_+^{J\times T}$ that approximate $\boldsymbol{X}$, i.e.

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} = \boldsymbol{x}_t \approx \sum_{j=1}^{J} \boldsymbol{w}_j h_{jt}. \tag{1}$$

Depending on the application field, these factors $\boldsymbol{W}$ and $\boldsymbol{H}$ are interpreted differently. In environmental source apportionment, $\boldsymbol{H}$ plays the role of mixing matrix (weights) which represents the source contributions, while $\boldsymbol{W}$ expresses temporal factor profiles (time series, see Appendix C). Thus, if a weight h_{jt} in a column of $\boldsymbol{H}$ is high, then the corresponding basis vector $\boldsymbol{w}_j$ is very important in approximating $\boldsymbol{x}_t$. Geometrically the basis vectors generate a simplicial cone and the columns of the matrix $\boldsymbol{W}$ are basis vectors spanning a subspace in $J \leq I$. Once estimated, the $\boldsymbol{H}$ and $\boldsymbol{W}$ matrices will be presented in sections 3.2.1 and 3.2.2, respectively. Indeed, the $\boldsymbol{H}$ matrix will allow us to analyze the contribution of the factors and the $\boldsymbol{W}$ matrix to visualize the diurnal profiles of the factors (time series).

The time series of the sources and their weights are calculated iteratively by minimizing a suitable measure f for the divergence between $\boldsymbol{W}$ and $\boldsymbol{H}$:

$$\underset{\boldsymbol{W},\boldsymbol{H}\geq 0}{\arg\min} f(\boldsymbol{W},\boldsymbol{H}) = \underset{\boldsymbol{W},\boldsymbol{H}\geq 0}{\arg\min} \left[\mathcal{D}(\boldsymbol{X} \parallel \boldsymbol{W}\boldsymbol{H}) + \mathscr{R}(\boldsymbol{W},\boldsymbol{H})\right], \tag{2}$$

where $\mathcal{D} : \mathbb{R}_+^{I\times J} \times \mathbb{R}_+^{J\times T} \to \mathbb{R}_+$ is a loss function and $\mathscr{R}$ is an optional regularization function that enforce desirable properties (e.g. smoothness, sparsity, ...) on matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ [15]. In this study, only $\underset{\boldsymbol{W},\boldsymbol{H}\geq 0}{\arg\min} \left[\mathcal{D}(\boldsymbol{X} \parallel \boldsymbol{W}\boldsymbol{H})\right]$ is optimized. The simplest loss function measure is based on the Frobenius norm:

$$\mathcal{D}_F(\boldsymbol{X} \parallel \boldsymbol{W}\boldsymbol{H}) = \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\right\|_F^2 = \frac{1}{2}\sum_{i}^{I}\sum_{t}^{T}\left(\boldsymbol{X}_{it} - (\boldsymbol{W}\boldsymbol{H})_{it}\right)^2. \tag{3}$$

The Frobenius similarity measure is a special case of the so-called $\beta-$divergence [15]. In this study we will consider only Kullback-Leibler divergence in case of Brunet algorithm [8] as follows:

$$\mathcal{D}_{KL}\left(\boldsymbol{X} \parallel \boldsymbol{W}\boldsymbol{H}\right) = \sum_{i}^{I}\sum_{t}^{T}\boldsymbol{X}_{it}\ln\left(\frac{\boldsymbol{X}_{it}}{(\boldsymbol{W}\boldsymbol{H})_{it}}\right) - \boldsymbol{X}_{it} + (\boldsymbol{W}\boldsymbol{H})_{it}\,. \tag{4}$$

2.2.3. Algorithms for solving the NMF problem

Many numerical algorithms have been developed to solve the NMF problem expressed in the equation (2). They can be divided into three general classes: (i) Alternating Least Squares (ALS) algorithms, (ii) multiplicative update algorithms, and (iii) gradient descent algorithms [6, 15]. The ALS algorithm computes the optimal solution of the unconstrained least squares problem, then it optimizes alternatively over one of the two factors $\boldsymbol{W}$ or $\boldsymbol{H}$ while keeping the other fixed. This subproblem is then reduced to one-factor convex.

By minimizing two criteria the squared Euclidean distance (or equivalently the squared Frobenius norm $\mathcal{D}_F$) and the generalized Kullback-Leibler divergence $\mathcal{D}_{KL}$, Lee and Seung [32] proposed the multiplicative update algorithm to solve the equation (1). Simple multiplicative update formulas based on $\mathcal{D}_F$ are given by

$$w_{ij} \quad \leftarrow \quad w_{ij}\frac{\left[\boldsymbol{X}\boldsymbol{H}^{\top}\right]_{ij}}{\left[\boldsymbol{X}\boldsymbol{H}\boldsymbol{H}^{\top}\right]_{ij}+\varepsilon}, \tag{5}$$

$$h_{jt} \quad \leftarrow \quad h_{jt}\frac{\left[\boldsymbol{W}^{\top}\boldsymbol{X}\right]_{jt}}{\left[\boldsymbol{W}^{\top}\boldsymbol{W}\boldsymbol{H}\right]_{jt}+\varepsilon}. \tag{6}$$

For the implementation purpose, a small positive constant ε is added to the denominator in each update rule to avoid division by zero. Lee and Seung claimed that the above algorithm converges to a local minimum [32], which was later shown to be incorrect (see for example [14, 6, 36]): the above algorithm 5, 6 can only keep the non-increasing property of the objective.

Based on the Kullback-Leibler divergence, Brunet et al. [8] used modified versions of Lee and Seung's (2001) [32] simple multiplicative updates to avoid numerical underflow. At each step, $\boldsymbol{W}$ and $\boldsymbol{H}$ are updated by using the coupled divergence equations:

$$\boldsymbol{W}_{ia} \quad \leftarrow \quad \boldsymbol{W}_{ia} \frac{\sum_{\mu} \left[\frac{\boldsymbol{H}_{a\mu} \boldsymbol{X}_{i\mu}}{(\boldsymbol{W}\boldsymbol{H})_{i\mu}} \right]}{\sum_{\nu} \boldsymbol{H}_{a\nu}}, \tag{7}$$

$$\boldsymbol{H}_{a\mu} \quad \leftarrow \quad \boldsymbol{H}_{a\mu} \frac{\sum_{i} \left[\frac{\boldsymbol{W}_{ia} \boldsymbol{X}_{i\mu}}{(\boldsymbol{W}\boldsymbol{H})_{i\mu}} \right]}{\sum_{j} \boldsymbol{W}_{ja}}. \tag{8}$$

The stopping criterion for Lee and for Burnet optimization is the variance of the connectivity matrix.

2.2.4. Initialization of NMF

Most NMF objectives are not convex and they are sensitive to the initialization of matrices $\boldsymbol{W}$ and $\boldsymbol{H}$. A good initialization can then sidestep some of the convergence problems, especially for a large input dataset. If the starting values for the algorithm are chosen randomly, every run of the NMF algorithm may find a different local minimum of the objective function. Therefore a reasonable initialization of matrix $\boldsymbol{W}$ and $\boldsymbol{H}$ is necessary and helps the physical interpretation of the obtained patterns [59, 50]. Particular emphasis has to be placed on the initialization of NMF because of its local convergence. Several approaches have then been proposed to choose the appropriate NMF initialization (see [2] for a review). Thiem et al. [59] suggested the use of Non-Negative Double Singular Value Decomposition (NNDSVD, developed in [7]) for PM source apportionment and did not recommend random initialization. Meanwhile, Hutchins et al. [24] claim that performing 30-50 runs for random initialization is considered sufficient to get a robust estimate of the factorization.

We tested three different kinds of initialization techniques for the two different NMF algorithms described above. In the first seeding method, a random starting point has been used, where the entries of $\boldsymbol{W}$ and $\boldsymbol{H}$ are drawn from a uniform distribution, within the same range as the $\boldsymbol{X}$ matrix's entries. That is the entries of each factor are drawn from a uniform distribution over $[0, \max\{\boldsymbol{x}\}]$, where $\boldsymbol{x}$ is the column vector of $\boldsymbol{X}$. We used a maximum of 100 runs for each algorithm to achieve stability. The second initialization method that we tested consists in using the results of the Independent Component Analysis (ICA) (FastICA algorithm [25]) where only the positive part is used to initialize the factors. The last tested initialization algorithm is the Non-Negative Double Singular Value Decomposition (NNSVD) [7].

2.2.5. Determining the number of components

One of the critical parameters in NMF is the number of components J to select for the factorization in equation 1. An appropriate decision on the value of J is critical in practice, but it is usually chosen such that $J \ll \min(I, T)$ in which case $\boldsymbol{WH}$ represents a compressed form of the initial matrix $\boldsymbol{X}$. The factorization rank parameter can be estimated by computing the Residual Sum of Squares (RSS) or by the explained variance (EVAR) between a target matrix $\boldsymbol{X}$ and its estimate $\widehat{\boldsymbol{X}}$:

$$\mathrm{RSS} = \sum_i \sum_t \left(\boldsymbol{X}_{it} - \widehat{\boldsymbol{X}}_{it} \right)^2 \tag{9}$$

$$\mathrm{EVAR} = 1 - \frac{\mathrm{RSS}}{\sum_{i,t} \widehat{\boldsymbol{X}}_{it}} \tag{10}$$

The "optimal" rank is chosen using the graph of the EVAR (or RSS); it corresponds to the first point where the graph shows an inflection point, as Hutchins et al. [24] did with the algorithm of Lee et al. [31].

3. Results and discussion

In the following, a preliminary analysis of the original data is conducted (section 3.1). Then, the choice of the number of factors and their profiles are discussed in the section 3.2. Finally, for interpretation purposes, theses factors are analyzed and related to different other parameters such as occupancy and windows state.

In this paper, the terms factor, component and patterns are equivalent. The results and analyses were conducted in R [54] using the NMF R package [19] and figures were produced using the ggplot2 package [66].

Since particles in the fine range (<2.5 µm) dominate the total particle number, the obtained NMF patterns emphasize this range. As the range variation of the number of particles can be very different from one bin to another, we considered very carefully the standardization need of the raw data matrix as input for NMF. This standardization procedure produces negative values that should be avoided due to the non-negativity constraint. For this reason we have made a translation of the standardized values by adding a constant equal to 2 particles.cm^{-3} to each one. This constant shift does not impact the factorization results.

3.1. Data description and preliminary analysis

During the sampling campaign, the indoor temperature values varied from 15°C to 40°C and relative humidity ranged between 18 % and 65%. The highest humidity values were recorded during May and June.

Fifteen size bins were collected, but only five representative fractions are described in this paper: $\mathrm{PN}_{0.35} = \mathrm{PN}_{[0.30-0.40]}$, $\mathrm{PN}_{0.9} = \mathrm{PN}_{[8-10]}$, $\mathrm{PN}_{2.5} = \mathrm{PN}_{[2-3]}$, $\mathrm{PN}_{8.75} = \mathrm{PN}_{[6.25-12.25]}$ and $\mathrm{PN}_{17.5} = \mathrm{PN}_{[15.5-19.5]}$. The indoor $\mathrm{PN}_{0.35}$ varied from 1 to 263 #.cm^{-3} (#.cm^{-3} for number per cm^3) with a median of 16 #.cm^{-3} and 90 % (percentile P_{90}) of the values are less than 18 #.cm^{-3}. There is no significant difference in the monthly levels excepting for March. The range variations for $\mathrm{PN}_{0.9}$, $\mathrm{PN}_{8.75}$ and $\mathrm{PN}_{17.5}$ are 2611, 119 and 26 $\#.L^{-1}$, respectively.

The variability expressed by means of the coefficient of variation ($CV = Standard\,Deviation/Mean$) depends on the size of the particle. Indeed, the CV values are: 113% for $\mathrm{PN}_{0.35}$, 80% for $\mathrm{PN}_{0.9}$, 92% for $\mathrm{PN}_{2.5}$, 250% for $\mathrm{PN}_{8.75}$ and 750% for $\mathrm{PN}_{17.5}$.

Descriptive statistics for the five size bins concentrations according to occupancy and windows opening state are shown in Table 1. The mean values show very large variations; mean values are

Table 1: Descriptive statistics of the selected indoor PN size bins (#.cm^{-3} for number per cm^3) according to occupancy (Occup. or Non Occup.) and windows opening state (Open or Close). The percentile P corresponds to percentage of the total values are the same as or below the measurement concentrations: the 25^{th} percentile (P_{50}), the 50^{th} percentile corresponds to the median and the 75th percentile (P_{75}). n represents the number of samples (minutes).

PN	**Group**	n	**Mean(SD)**	**Min**	**Max**	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}
$PN_{0.35}$	Close	230266	28(32.1)	0.68	262.8	5.73	8.76	16.1	34.83	63.6
	Open	29765	20.4(16.4)	3.16	125	7	9.46	14.6	26	41.8
$PN_{0.9}$	Close	230266	0.33(0.27)	0	2.6	0.1	0.15	0.2	0.42	0.64
	Open	29764	0.32(0.19)	0.025	1.4	0.12	0.18	0.2	0.41	0.57
$PN_{2.5}$	Close	230266	0.031(0.02)	0	0.6	0.006	0.013	0.02	0.042	0.06
	Open	29764	0.081(0.05)	0	0.6	0.029	0.04	0.06	0.11	0.15
$PN_{8.75}$	Close	230266	3×10^{-4}(0.001)	0	0.11	0	0	0	0	0.001
	Open	29764	0.001(0.002)	0	0.03	0	0	0	0.002	0.004
$PN_{17.5}$	Close	230266	3×10^{-5}(0.0002)	0	0.02	0	0	0	0	0
	Open	29764	1×10^{-4}(0.0004)	0	0.006	0	0	0	0	0
$PN_{0.35}$	Non Occup	245326	27.25(31)	0.68	262.8	5.9	8.8	16	34	60.8
	Occup	14705	25.54(28.5)	2.74	188.1	6	8.7	14.8	30.15	57.93
$PN_{0.9}$	Non Occup	245326	0.32(0.26)	0	2.6	0.1	0.16	0.26	0.41	0.63
	Occup	14704	0.36(0.29)	0.02	2.4	0.12	0.17	0.28	0.44	0.72
$PN_{2.5}$	Non Occup	245326	0.036(0.03)	0	0.67	0.006	0.01	0.027	0.05	0.073
	Occup	14704	0.061(0.04)	0	0.6	0.02	0.03	0.049	0.08	0.121
$PN_{8.75}$	Non Occup	245326	0.0004(0.001)	0	0.07	0	0	0	0	0.001
	Occup	14704	0.002(0.002)	0	0.11	0	0	0.001	0.002	0.004
$PN_{17.5}$	Non Occup	245326	0.00003(0.0002)	0	0.02	0	0	0	0	0
	Occup	14704	0.0002(0.001)	0	0.02	0	0	0	0	0.001

higher than the median ones. As the data distribution was positively skewed, the median (P_{50}) was preferred for interpretation purposes instead of the arithmetic mean.

For most of the particle sizes, the median values of PN are very similar regardless the windows state, while the 10^{th} and 90^{th} percentile values are different. For instance, 90 % of $PN_{0.35}$ concentrations are less than 64 #.cm^{-3} in the case of closed windows state and 42 #.cm^{-3} when at least one window is opened. These observations are related to the total number of minutes in each opening state. On average, windows remain closed around 85% of the total time throughout the study period (24/7), week-end and holidays included.

Indoor particles can originate from outdoor sources and also from various indoor sources. Thus, it is possible that the levels of indoor particles exceed outdoor ones. Table 2 shows the Indoor/Outdoor (I/O) ratio for the selected five size bins levels in different configurations. Diurnal time variability of the ratio (I/O) is provided in the supplementary material of this paper (cf. Figure B.1).

Median I/O ratios are less than 1 for smaller range ($\leq 2.5\,\mu m$) regardless the windows state or indoor occupancy.

It is worth noting that median I/O ratios of PN during the non-occupancy are very similar to those observed for closed windows. In the absence of known indoor sources, the reported median I/O ratios have been ranged from 0.37 to 0.53 in the case of opening windows.

Table 2: Descriptive statistics of I/O PN ratios for the selected size ranges according to occupancy (Occup. and Non Occup.) and opening windows state (Open and Close). n represents the number of samples (minutes) and P the percentile values.

I/O ratio	**Group**	n	**Mean(sd)**	**min.**	**max.**	P_{25}	P_{50}	P_{75}	P_{90}
$PN_{0.35}$	Close	104393	0.49(0.2)	0.008	4.318	0.363	0.47	0.59	0.71
	Open	496	0.73(0.21)	0.347	1.476	0.601	0.7	0.88	1
$PN_{0.9}$	Close	104393	0.57(0.25)	0.002	7.436	0.425	0.53	0.65	0.81
	Open	496	0.76(0.18)	0.36	1.378	0.643	0.76	0.87	1
$PN_{2.5}$	Close	104393	0.59(4.57)	0.0003	611	0.243	0.37	0.57	0.91
	Open	496	0.62(0.33)	0.1	2.212	0.376	0.57	0.83	1.07
$PN_{8.75}$	Close	104393	2.29(6.22)	0.0004	311	1	1	1	1.91
	Open	496	6(11.33)	0.012	71	0.524	1	3.72	21
$PN_{17.5}$	Close	104393	1.17(1.9)	0.002	201	1	1	1	1
	Open	496	2.43(5.2)	0.032	51	1	1	1	11
$PN_{0.35}$	Non Occup.	99902	0.49(0.2)	0.043	4.318	0.365	0.47	0.56	0.71
	Occup.	4987	0.51(0.26)	0.008	4.143	0.342	0.47	0.63	0.8
$PN_{0.9}$	Non Occup.	99902	0.57(0.25)	0.002	7.436	0.424	0.53	0.65	0.8
	Occup.	4987	0.62(0.27)	0.075	3.849	0.448	0.58	0.73	0.9
$PN_{2.5}$	Non Occup.	99902	0.6(4.6)	0.0003	611	0.239	0.36	0.56	0.8
	Occup.	4987	0.9(4.25)	0.008	291	0.398	0.59	0.2	1.4
$PN_{8.75}$	Non Occup.	99902	2.1(5.67)	0.0004	301	1	1	1	1
	Occup.	4987	6.6(12.7)	0.001	311	0.524	1	11	21
$PN_{17.5}$	Non Occup	99902	1.15(1.7)	0.002	201	1	1	1	1
	Occup	4987	2(3.8)	0.009	71	1	1	1	1

The Figure 2 shows the diurnal variability of the median value of indoor PN concentrations according to occupancy and windows states for the following bins: $PN_{0.35}$, $PN_{0.9}$, $PN_{2.5}$ and $PN_{17.5}$. The diurnal pattern for $PN_{0.35}$ shows a convex shape when all the windows are closed and no occupancy detected. The I/O ratio of the corresponding bin is always less than 0.6 for closed windows and it ranges between 0.6 and 1 in the case of opened windows. The variation profile can be explained by the fact that the outdoor particles penetrate inside by mechanical ventilation and infiltration through the gaps in the building, given that the windows are closed. These fine particles come mostly from outdoor sources, such as traffic which is higher during the rush hours. The diurnal profile of the PN of size 0.9 µm when the windows are opened is characterized by a significant peak between 8:00 and 9:30 a.m. (when usually work starts) and then it decreases until 7:00 p.m. On the other hand, the same variation is observed when a movement of occupants is detected, even when the windows are closed. For $PN_{2.5}$, there is a significant difference between the case when the windows are opened or closed, regardless the occupancy. The values are higher for $PN_{2.5}$ when windows are opened showing the importance of outdoor sources for these size bins. By contrast, for the coarse particles ($PN_{17.5}$), the occupancy variable discriminates the profiles revealing that these particles are generated indoors (occupants' activities, such as walking).

To summarize, opening the windows results in similar trends in PN for both occupation and non-occupation conditions (except for $PN_{17.5}$ due to less contribution of coarse particles in the air). Mostly, PN concentrations with occupancy are higher than PN concentrations without occupancy. However, real contribution of occupation is marked when the windows are closed.

The particulate number concentration decrease in the indoor environment occurs mainly by two mechanisms: ventilation and deposition. In general, ventilation could play a positive role in the loss of particles from indoor air, but sometimes it may cause entering the outdoor pollutants via the supplied air coming indoors.

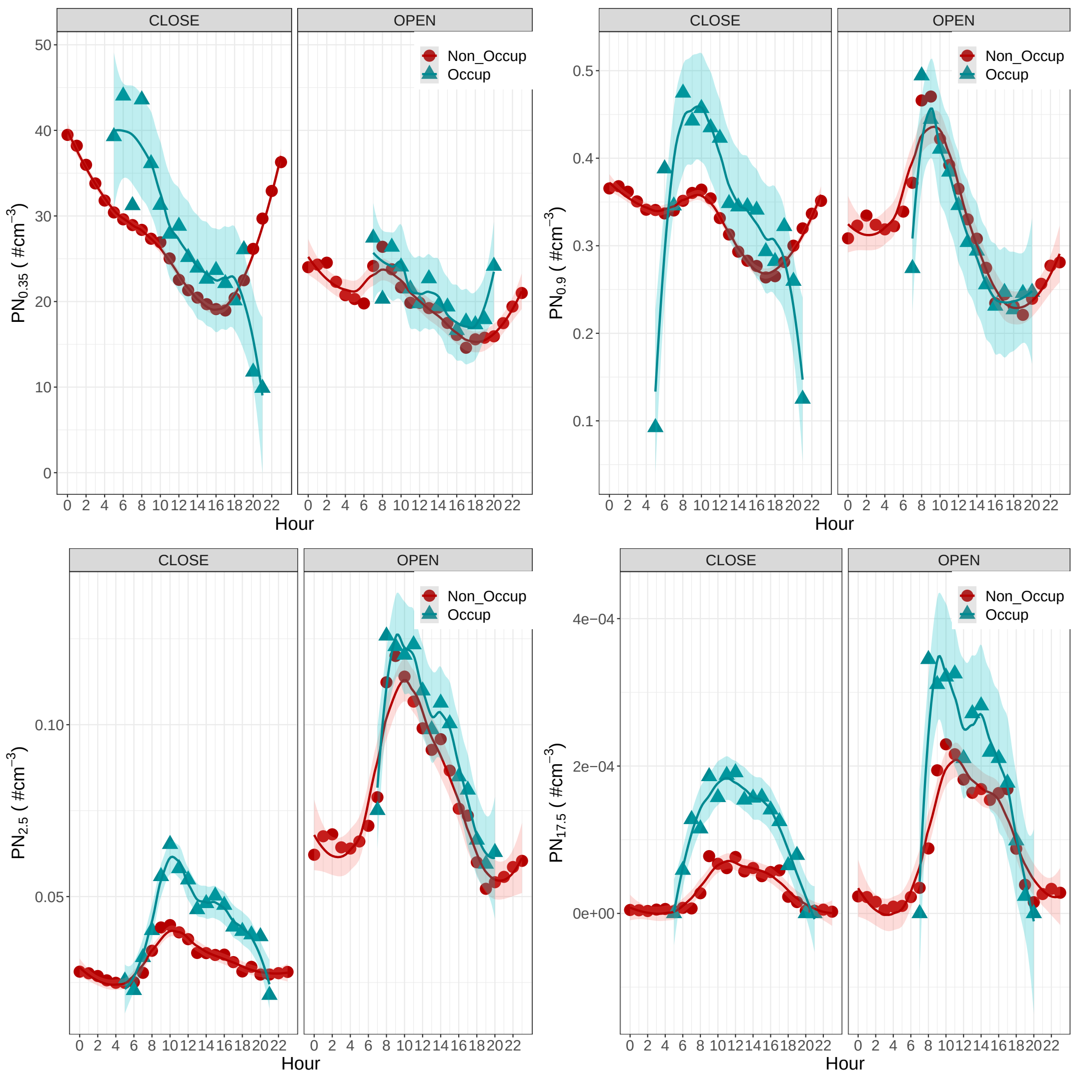

Figure 2: Diurnal variability of the median value of indoor PN concentrations for the following size ranges: 0.35 µm, 0.9 µm 2.5 µm and 17.5 µm. $PN_{0.35} = PN_{[0.30-0.40]}$, $PN_{0.9} = PN_{[8-10]}$, $PN_{2.5} = PN_{[2-3]}$ and $PN_{17.5} = PN_{[15.5-19.5]}$.

3.2. NMF results

As presented in the subsection 2.2.4 and 2.2.5, the method should be initialized and an optimal number of components should be determined.

Values in the range $J = 2, \ldots, 10$ were tested based on 20 000 randomly sampled 1-minute dataset using three different initialization algorithms: ICA, NNSVD and Random for the two optimization methods proposed by Brunet [8] and Lee [31, 32].

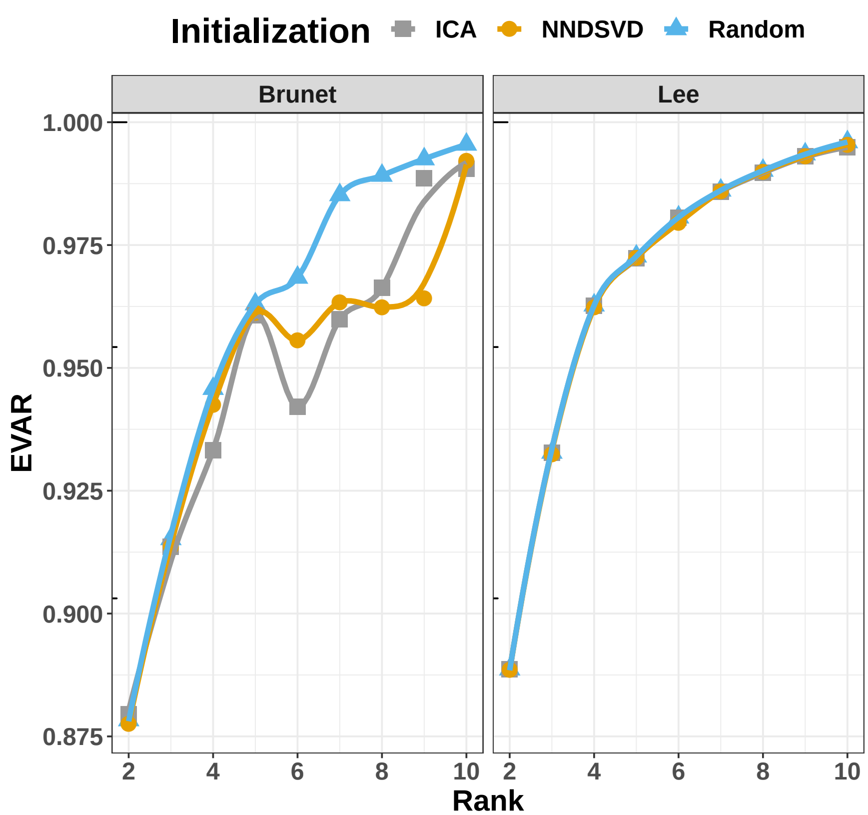

Figure 3: Explained Variance (EVAR) variations of Brunet [8] or Lee [31, 32] algorithms according to different initializations.

Figure 3 shows the evolution of the explained variance according to initialization algorithms (ICA, NNSVD, and Random) according to the rank number. Unexpectedly, for each initialization method, NMF generated extremely resembling results for Lee's algorithm. A similar observation is observed for the first 5 rank factors in Brunet's algorithm. Note that random initialization may make the experiments unrepeatable because of their local minima property. Several studies have revealed that methods with non-random seeding demonstrate their superiority either in the fast convergence in prophase or the structure preservation [69, 7]. This is yet another reason for not choosing random initialization in our experiments.

The interpretability of the factors can be a selection criterion for the rank choice. The use of expert insight in the case of indoor air quality was taken into account. It can be observed that using the interpretability criterion of the value of J by increasing or decreasing by 1, either the new NMF-factor shows a mixture of parts of existing factors or contains completely new environmental patterns. To represent as many different patterns as possible and taking into account the explained variance variation, we found for the considered dataset that J should be about 4-6. To draw comparisons between different results, the best physical interpretation and factorization error at the same time were obtained for $J = 5$. With this rank corresponding thus to 5 factors, the explained variance corresponds to 96% and 97.5% for Brunet's and Lee's algorithms, respectively.

3.2.1. Factors contributions

The time series of the factors obtained by NMF are provided in the supplementary material of this article (cf. Figure C.1). The relative contribution of factors to each particle size is shown in Figure 4. The relative contributions of the five components are distributive *i.e.* all the fractions can be found in at least one factor.

For the fine size fractions ($0.3\,\mu\text{m}, 0.45\,\mu\text{m}, 0.57\,\mu\text{m}$), the contribution of the component 2 obtained by NMF is nearly 75 % of the total contributions. By contrast, this component contributes the least for larger sizes of particles, so it is specific to the fine size.

The fourth component is related to coarse fraction, which represents 80 % of all the other contributions. For fractions between 1.3 µm and 4.5 µm, the components 3 and 5 share over 50 % of all contributions. The first component is involved in about 20% for particles smaller than 2.5 µm and can reach 80% for particles with a size of 12.5 µm.

To facilitate the interpretation of each component, we compute the correlations between the original data (the monitored time series corresponding to each fraction) and each NMF component. The interpretation of the NMF components is based on finding which PN time series are the most strongly correlated with each component. Figure 5 shows the correlation values between the components time series and fraction variables time series. The color variations express the strength and direction of the correlation relationship.

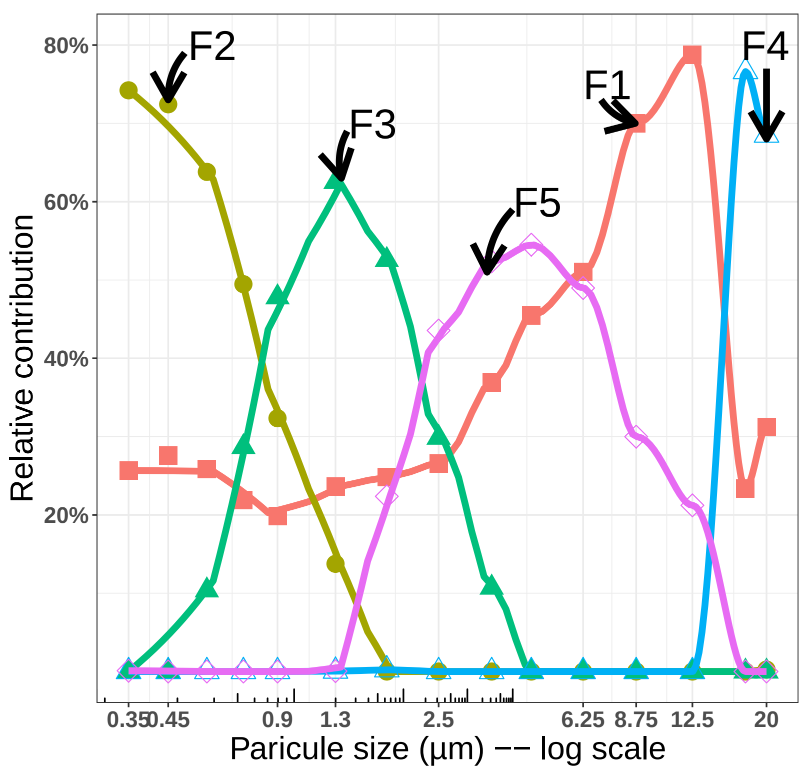

Figure 4: Relative contributions of the factors to temporal variability of the 15 fractions.

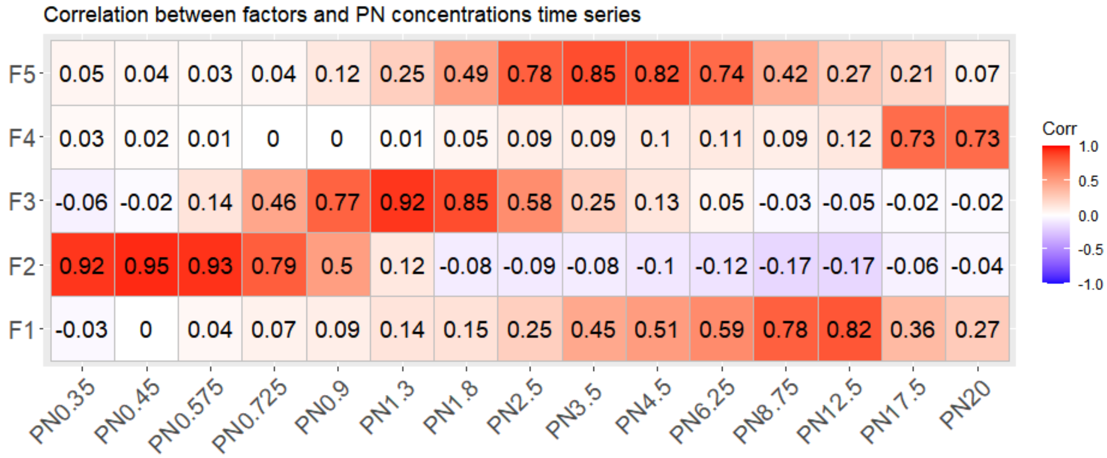

Figure 5: Pearson correlation coefficient between NMF factors ($\boldsymbol{H}$ matrix) and the PN concentrations time series of each fraction.

Note that all the significant correlations, i.e. greater than 0.6, are positive. The first NMF component is strongly correlated with particles ranging from 6.25 µm to 12.5 µm, the correlation coefficients vary form +0.6 and +0.82. The contribution of this component to the variability of the of particles of sizes between 3.5 µm and 4.5 µm is not negligible ($\sim$ 40%) although their correlations are not very strong (+0.45, +0.51).

The 2^{nd} factor is associated with particle size ranges below 0.8 µm with a strong positive correlation coefficient (> 0.8). Its contribution is higher than 63% for particles of sizes 0.575 µm, 0.45 µm and 0.35 µm.

Factor 3 contributes (up to 50%) to sizes between 0.9 µm and 2.5 µm, with a correlation varying between +0.58 and +0.95. We notice that the diurnal profiles of the F3 and $\mathrm{PN}_{0.9-1.8}$ are very similar (cf. Figure 2); both of them are characterized by a significant peak between 8:00 a.m. and 9:30 a.m. followed by a decreases until 7:00 p.m.

Note that for particles of size 2.5 µm, two factors contribute simultaneously to its variation: F3 (correlation = 0.58) and F5 (correlation =0.78). The same observation can be made for the particles of size 6.25 µm: two factors contribute to its variation (F1 and F5). Coarse particles (>17.5 µm) are associated with the 4^{th} factor with a correlation of +0.73. The latter does not contribute to any of the other particle ranges. In Figure 3, there is a peak of contribution of about 80% for particle sizes 17.5 µm and > 20 µm. Particles in the intermediate size range (between 2.5 and 6.25 µm) are associated to F5 with correlations ranging from +0.74 to +0.85. The contribution of F5 to the variability of particles of 3.5 µm size, is close to 50%.

3.2.2. Time variation of the factors

Having examined the overall contributions of NMF components, we now take a closer look to interpretate them. This section sheds light on the nature of the hidden components obtained using NMF. To do so, we combine three sources of information: *(i)* diurnal variation of different time series (NMF factores and CO_2), *(ii)* windows states and *(iii)* indoor occupancy. According to the states of the windows (i.e. open/closed) and to the occupancy (occupied/ unoccupied), we further subset the factor's time series and plotted the diurnal variation in Figure 6. More specifically, two figures are presented for each component corresponding to the window states (closed or open) and in each figure, two curves (red dot and green triangle) are associated with the occupancy status. To facilitate the interpretation, diurnal variations in CO_2 concentrations have been added in the same Figure. The CO_2 variations are used here as a fingerprint of occupant presence in order to allow identification of similar factors with this type of variation.

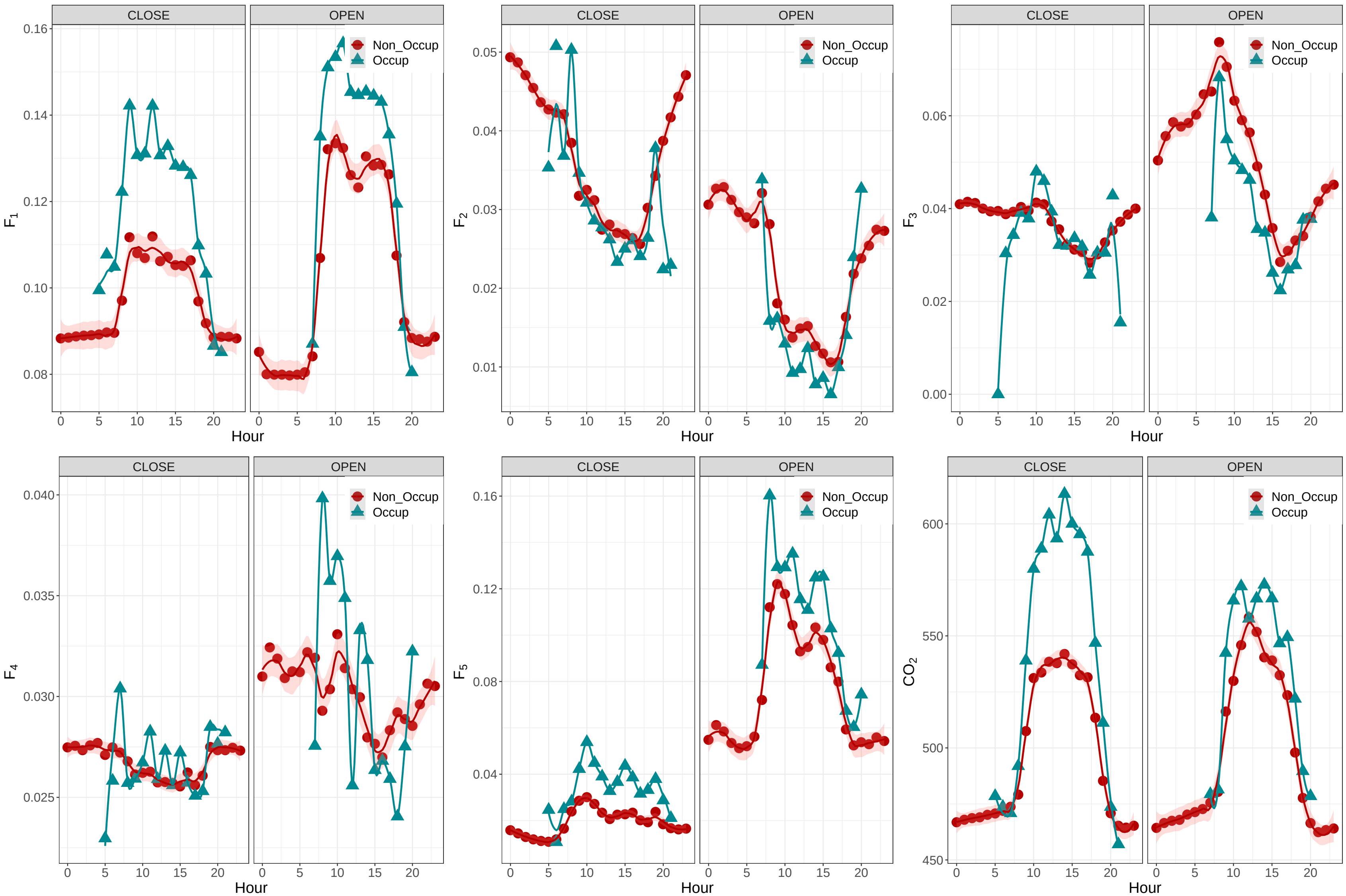

Figure 6: Diurnal variation of the five NMF factors and CO_2 concentrations according to windows state and occupancy.

Table 3 shows statistical tests performed to determine whether the differences between the average values of all the NMF factors are significant in two situations: (i) when all windows are closed *vs.* when at least one window is open, and (ii) in the case of occupancy *vs.* non-occupancy. For a *p-value* <0.05, the null hypothesis "H_0 : *No difference in mean*" is rejected. For example, for the first factor, we can say that there is no significant difference when all the windows are closed or at least one window is open (*p-value* >0.05, gray cells in the table 3). Whereas occupancy is an important parameter for component 1 (F1): the variability of the two profiles (occupancy *vs.* non-occupancy) are different (*p-value* <0.05).

Table 3: Significant differences between diurnal profiles to discriminate the role of windows and occupancy parameters using statistical tests.

$p-value$	Windows		Occupancy	
	$t-test$	Wilcoxon	$t-test$	Wilcoxon
F1	0.36	0.98	4.69E-06	2.54E-06
F2	0.0003301	0.0006599	0.0005568	0.0009668
F3	0.0001015	3.69E-05	0.00245	0.004618
F4	6.77E-08	1.97E-08	0.1147	0.02706242
F5	0E-10	0E-10	0.02706242	0.002432

The diurnal profile of the first factor is very similar to the diurnal variations of indoor CO_2 concentrations (cf. Figure 6), suggesting that this component is related to the presence of the occupants, who are the main source of CO_2. When all the windows are closed, the occupancy curve (red dots) is separated (*p-value* < 0.05) from the non-occupancy one.

When at least one window is open, the values of component 1 are higher. It is as if another source coming from outside is added or the ventilation changes the transport mechanisms of large particles (resuspension). Nevertheless, several studies report that the mechanism of resuspension is especially related to particles of sizes less than 10 µm [52, 51]. As previously mentioned (Figure 5), the peak contribution of the first component is associated with particles of sizes between 8.75and 12.5 µm.

The diurnal profile of the second component illustrates a decreasing effect during the period from 6 a.m. to 7 p.m.. For F2, there is no significant difference for the "occupancy" variable, the two curves overlap. As shown in previous sections, this component is related to fine particles (less than 0.75 µm).

It turns out that the parameter "opening windows" perfectly discriminates the component 5 ($p-value$ <0.05), but the importance of occupancy is not conclusive. When at least one window is open, the component 5 levels are high. The diurnal profiles of the two components 1 and 5 have comparable patterns, for the case of open windows. As a matter of fact, the relative contribution of F1 and F5 are important when the open-plan is occupied (7 a.m.-7 p.m.).

For component 3, a peak at 9 a.m. is observed in the case of open windows and the diurnal profile gives a sinusoidal appearance. On the other hand, NO_x emissions vary according to the time of day: they are very high in the morning and late afternoon (the rush hours). A hypothesis can be put forward suggesting that the peak of component 3 is of external origin, related to some sources such as road traffic.

Components F3 and F5 could be attributed to the influence of outdoor sources on the indoor environment. The seasonal appearance (of component 3) is mainly due to diurnal variations of outdoor sources. This finding provides additional insight into the sources of particles ranging in size from 1.8 µm to 6.25 µm. Thus, by combining them with the information provided in Table 2, we notice the important influence of outdoor conditions on indoor concentration levels. At this stage, one of the major points of interest has been the outdoor source identification i.e. when outdoor concentrations and characteristics are the main contributing factors. Overall, outdoor sources have been mainly associated with fine particles in accumulation mode (0.1–1 µm), probably because these particles can persist in the air since they are too small for inertial deposition and too large for diffusion removal processes. These particles are capable of entering in the buildings and remaining airborne for longer periods.

The forth component (F4) is characterized mainly by the coarse particles ($>$ 17.5 µm). Figure 6 shows a strong variability in the diurnal profile during the occupation period. The profile is very random because it is mainly associated with the activities of the occupants. These activities, and in particular walking, are responsible of resuspension of the coarse particles. We remind that the office is equipped with a carpet covering the entire floor. The statistical tests cannot be used because the probability density is bimodal for this profile. We notice that when at least one window is open, the profile is further modified. Statistical tests confirm this observation (*p-value* $<5\%$).

3.3. Limitations and future directions

Firstly, a generic NMF, which is a method to solve a linear system, was employed and the underlying model was used as a set of linear Chemical Mass Balance (CMB) to estimate individual source components from their mixtures. However, the indoor mixing phenomena are better represented by non-linear relationships, such as infiltration, sink, inertial deposition, and diffusive removal processes. A research question that remains is how these latter processes can be considered in a more general model. In the same vein, future work would include the hybridization of physical models with factorial source separation methods.

Secondly, although NMF has a realistic non-negative constraint on factor profiles and contributions, it still does not cover all the indoor environment characteristics. Thus, many parameters as occupant behavior and windows opening, which have *"random impacts"* over the primary determinants of the decay rate, have to be incorporated in the NMF optimization problems. Similarly, the including time-activity patterns could improve the interpretability of the NMF results.

One critical question that remains to be answered about the NMF method for indoor air quality is: how to integrate inherent indoor specificity and constraints in the NMF formulation?

Recently the NMF field expanded to multidimensional data arrays, called Non-negative Tensor Factorization (NTF) [15], which could offer valuable new insights on IAQ modeling issues. Future work also involves validating the performance of the NMF model using data over a much longer period and above all including more information about the different activities such as walking, cleaning, printing, ...

From the instrumentation point of view, it is clear that the motion detection has been largely underestimated by the measurement method. It might be more appropriate to introduce the "*number of occupants*" as a parameter in the data processing as well as in the modeling step.

3.4. Conclusions

Evidence continues to mount that indoor particles concentrations are one of the major determinants for individual exposure. That is why it is necessary to characterize the source time variability and to assesses the different source impacts on the total personal exposure. Attempts to estimate the source of indoor particles concentrations are complicated because indoor air quality is both building-specific and occupant-specific. Many researchers are using experimental chambers; however quantifying continuous ventilation rates and penetration factors for an occupied building can be tedious, time-consuming and expensive. While most receptor-oriented methods use the pollutant compositions (chemical fingerprint) to apportion the contribution of individual sources, our work focuses on the time variability source characterization, namely *"temporal fingerprint"*. This study outlines the basic temporal characterization and source apportionment using Non-negative Matrix Factorization (NMF).

This study has shown that continuous measurements of indoor particle number concentrations with additional information of occupancy and windows opening are useful for the source apportionment models. The factorization has been successfully employed to study particle number time series fluctuations in an occupied open-plan office. Five distribution profiles were resolved using different initialization and optimization algorithms. Thanks to NMF and correlation analysis, the impact of occupancy and windows states associated with outdoor traffic were identified. The outdoor sources are captured by two components (very likely the traffic impact) with major number modes at 1.3 µm. Besides, a potential association with primary outdoor pollutants (NO_x) has been captured and a diurnal pattern similar to traffic can be associated. A common pattern between F1 and $PN_{0.9-1.8}$ has been identified, such as diurnal profile which is characterized by a significant peak between 8:00 a.m. and 9:30 a.m, decreasing until 7:00 p.m.

Taking into consideration the results presented above, throughout the six-month measurements of indoor and outdoor PN concentrations, it can be concluded the followings:

- This study demonstrates the importance of recording real-time concentrations over a longer duration (i.e., several months with a short time step). The exploitation of such data has the potential to extract and capture the different patterns of temporal variability and their major determinants.
- These results contribute significantly to the very small data set available in the literature on

the time variability source characterization.

- The NMF technique -and its variants- shows considerable promise for further application to the indoor environment and the possibility to identify other sources and their contributions.
- Continuous monitoring of climatic parameters as well as active instrumentation of the building (windows, occupancy) are necessary for the evaluation of the total exposure to indoor pollutants.

To sum up, the results of this study entail that the degree of human's exposure to different sources varies with many parameters (building and occupants specific). These sources could be captured using NMF and its variants.

Acknowledgments

This work received financial support from the French research program on air quality (PRIMEQUAL) through the Grant No 12-MRES-PRIMEQUAL-4-CVS-09.

References

[1] Eileen Abt, Helen H Suh, Paul Catalano, and Petros Koutrakis. Relative contribution of outdoor and indoor particle sources to indoor concentrations. *Environmental science & technology*, 34(17):3579–3587, 2000. 1

[2] Russell Albright, James Cox, David Duling, Amy N Langville, and C Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, Tech. rep. 919. NCSU Technical Report Math 81706. http://meyer. math. ncsu. edu/Meyer/Abstracts/Publications. html, 2006. url: http://citeseerx. ist. psu. edu/viewdoc/download, 2006. 2.2.4

[3] Mohd Yasreen Ali, Marlia M Hanafiah, Md Firoz Khan, and Mohd Talib Latif. Quantitative source apportionment and human toxicity of indoor trace metals at university buildings. 121:238–246. 1

[4] JP Barnard and C Aldrich. Modelling of air pollution in an environmental system by use of non-linear independent component analysis. *Computer Aided Chemical Engineering*, 9:81–86, 2001. 2.2.1

[5] Claudio A Belis, Bo R Larsen, Fulvio Amato, Imad El Haddad, Olivier Favez, Roy M Harrison, Philip K Hopke, Silvia Nava, Pentti Paatero, and André Prevot. European guide on air pollution source apportionment with receptor models, 2014. 1

[6] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007. 2.2.3, 2.2.3

[7] Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008. 2.2.4, 3.2

[8] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004. 2.2.2, 2.2.3, 3.2, 3

[9] Carmen Cacho, G Ventura Silva, Anabela O Martins, Eduardo O Fernandes, Dikaia E Saraga, Chrysanthi Dimitroulopoulou, JB Bartzis, Diana Rembges, Josefa Barrero-Moreno, and Dimitrios Kotzias. Air pollutants in office environments and emissions from electronic equipment: A review. *Fresen. Environ. Bull*, 22(9), 2013. 1

[10] Christopher Y Chao and Eddie C Cheng. Source apportionment of indoor pm2. 5 and pm10 in homes. *Indoor and Built Environment*, 11(1):27–37, 2002. 1

[11] Chun Chen and Bin Zhao. Review of relationship between indoor and outdoor particles: I/o ratio, infiltration factor and penetration factor. *Atmospheric Environment*, 45(2):275–288, 2011. 1

[12] Judith C Chow, John G Watson, et al. Guideline on speciated particulate monitoring. Technical report, US Environmental Protection Agency, 1998. 1

[13] Robert Chreiky, Gilles Delmaire, Matthieu Puigt, Gilles Roussel, Dominique Courcot, and Antoine Abche. Split gradient method for informed non-negative matrix factorization. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 376–383. Springer, 2015. 2.2.1

[14] Moody Chu, Fasma Diele, Robert Plemmons, and Stefania Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. In *SIAM Journal on Matrix Analysis*. Citeseer, 2004. 2.2.3

[15] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009. 2.2.2, 2.2.2, 2.2.2, 2.2.3, 3.3

[16] John A Cooper and John G Watson Jr. Receptor oriented methods of air particulate source apportionment. *Journal of the Air Pollution Control Association*, 30(10):1116–1125, 1980. 1

[17] Gilles Delmaire, Gilles Roussel, Dany Hleis, and Frédéric Ledoux. Une version pondérée de la factorisation matricielle non negative pour l'identification de sources de particules atmospheriques. application au littoral de la mer du nord. *Apii Jesa Journal Europeen Des Systemes Automatises*, 44(4):547, 2010. 2.2.1

[18] Andrea R Ferro, Royal J Kopperud, and Lynn M Hildemann. Source strengths for indoor human activities that resuspend particulate matter. *Environmental science & technology*, 38(6):1759–1764, 2004. 1

[19] Renaud Gaujoux and Cathal Seoighe. A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):367, 2010. 3

[20] Hai Guo. Source apportionment of volatile organic compounds in hong kong homes. 46(11):2280–2286. 2.2.1

[21] Per Heiselberg, Erik Bjørn, and Peter V Nielsen. Impact of open windows on room air flow and thermal comfort. 1(2):91–100. 2.1.3

[22] Ronald C Henry, Charles W Lewis, Philip K Hopke, and Hugh J Williamson. Review of receptor model fundamentals. *Atmospheric Environment (1967)*, 18(8):1507–1515, 1984. 1

[23] Philip K Hopke. Review of receptor modeling methods for source apportionment. *Journal of the Air & Waste Management Association*, 66(3):237–259, 2016. 1

[24] Lucie N Hutchins, Sean M Murphy, Priyam Singh, and Joel H Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24(23):2684–2690, 2008. 2.2.4, 2.2.5

[25] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000. 2.2.4

[26] Christina Isaxon, Anders Gudmundsson, EZ Nordin, Leif Lönnblad, Andreas Dahl, Gunilla Wieslander, Mats Bohgard, and Aneta Wierzbicka. Contribution of indoor-generated particles to residential exposure. *Atmospheric Environment*, 106:458–466, 2015. 1

[27] Adib Kfoury, Frédéric Ledoux, Abdelhakim Limem, Gilles Delmaire, Gilles Roussel, and Dominique Courcot. The use of a non negative matrix factorization method combined to pm2. 5 chemical data for a source apportionment study in different environments. In *Air Pollution Modeling and its Application XXIII*, pages 79–84. Springer, 2014. 2.2.1

[28] Adib Kfoury, Frédéric Ledoux, Cloé Roche, Gilles Delmaire, Gilles Roussel, and Dominique Courcot. Pm 2.5 source apportionment in a french urban coastal site under steelworks emission influences using constrained non-negative matrix factorization receptor model. *Journal of Environmental Sciences*, 40:114–128, 2016. 2.2.1

[29] Henri Lantéri, Céline Theys, Cédric Richard, and Cédric Févotte. Split gradient method for nonnegative matrix factorization. In *Signal Processing Conference, 2010 18th European*, pages 1199–1203. IEEE, August 23-27, 2010 2010. 2.2.1

[30] Daniel D Lee and H Sebastian Seung. Unsupervised learning by convex and conic coding. *Advances in neural information processing systems*, pages 515–521, 1997. 2.2.2

[31] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 1, 2.2.2, 2.2.5, 3.2, 3

[32] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001. 1, 2.2.2, 2.2.3, 2.2.3, 3.2, 3

[33] Weihui Liang and Xudong Yang. Indoor formaldehyde in real buildings: Emission source identification, overall emission rate estimation, concentration increase and decay patterns. *Building and Environment*, 69:114–120, 2013. 1

[34] A Limem, G Delmaire, Matthieu Puigt, G Roussel, and D Courcot. Non-negative matrix factorization under equality constraints—a study of industrial source identification. *Applied Numerical Mathematics*, 85:1–15, 2014. 2.2.1

[35] Abdelhakim Limem, Gilles Delmaire, Matthieu Puigt, Gilles Roussel, and Dominique Courcot. Non-negative matrix factorization using weighted beta divergence and equality constraints for industrial source apportionment. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6. IEEE, 2013. 2.2.1

[36] Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596, 2007. 2.2.3

[37] Haitao Liu, Chongguo Tian, Zheng Zong, Xiaoping Wang, Jun Li, and Gan Zhang. Development and assessment of a receptor source apportionment model based on four nonnegative matrix factorization algorithms. 197:159–165. 2.2.1

[38] Christopher M Long, Helen H Suh, Lester Kobzik, Paul J Catalano, Yao Yu Ning, and Petros Koutrakis. A pilot investigation of the relative toxicity of indoor and outdoor fine particles: in vitro effects of endotoxin and other particulate properties. *Environmental Health Perspectives*, 109(10):1019, 2001. 1

[39] Marianna Luoma and Stuart A Batterman. Characterization of particulate emissions from occupant activities in offices. *Indoor air*, 11(1):35–48, 2001. 1

[40] William W Nazaroff. Exploring the consequences of climate change for indoor air quality. *Environmental Research Letters*, 8(1):015022, 2013. 1, 2.2.1

[41] Ioannis Nezis, George Biskos, Konstantinos Eleftheriadis, and Olga-Ioanna Kalantzi. Particulate matter and health effects in offices-a review. 156:62–73. 1

[42] S Oeder, S Dietrich, I Weichenmeier, W Schober, G Pusch, RA Jörres, R Schierl, D Nowak, H Fromme, H Behrendt, et al. Toxicity and elemental composition of particulate matter from outdoor and indoor air of elementary schools in munich, germany. *Indoor air*, 22(2):148–158, 2012. 1

[43] Rachid Ouaret, Anda Ionescu, Viorel Petrehus, Yves Candau, and Olivier Ramalho. Spectral band decomposition combined with nonlinear models: application to indoor formaldehyde concentration forecasting. 32(4):985–997. 2.1.3

[44] Rachid Ouaret, Anda Ionescu, Viorel Petrehus, Yves Candau, and Olivier Ramalho. Particulate matter variability sources in an open-plan office: comparison of two monitoring campaigns. In *European Aerosol Conference*, page 1, Tours, France, September 2016. 2.2.1

[45] Rachid Ouaret, Anda Ionescu, Olivier Ramalho, and Yves Candau. Indoor air pollutant sources using blind source separation methods. In *25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.*, Bruges, Belgium, April 2017. 2.2.1

[46] Rachid Ouaret, Anda Ionescu, Olivier Ramalho, Yves Candau, Evelyne Gehin, and Viorel Petrehus. Analysis of the temporal variability of indoor particulate matter concentrations using Blind Source Separation methods: a comparative study. In *International Aerosol Conference*, page 1, Busan, South Korea, August 2014. 2.2.1

[47] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994. 1, 2.2.1, 2.2.2

[48] Pentti Paatero, Unto Tapper, Pasi Aalto, and Markku Kulmala. Matrix factorization methods for analysing diffusion battery data. 22:S273–S276. 2.2.1, 2.2.2

[49] Pallavi Pant. *Receptor modelling studies of airborne particulate matter in the United Kingdom and India.* PhD thesis, University of Birmingham, 2014. 1

[50] Marc Plouvin, Abdelhakim Limem, Matthieu Puigt, Gilles Delmaire, Gilles Roussel, and Dominique Courcot. Enhanced nmf initialization using a physical model for pollution source apportionment. In *22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014)*, pages 261–266, 2014. 2.2.1, 2.2.4

[51] Jing Qian and Andrea R Ferro. Resuspension of dust particles in a chamber and associated environmental factors. 42(7):566–578. 3.2.2

[52] Jing Qian, Andrea R Ferro, and Kathleen R Fowler. Estimating the resuspension rate and residence time of indoor particles. 58(4):502–516. 3.2.2

[53] Jing Qian, Jordan Peccia, and Andrea R Ferro. Walking-induced particle resuspension in indoor environments. *Atmospheric Environment*, 89:464–481, 2014. 1

[54] R Core Team. *R: A language and environment for statistical computing*, 2019. 3

[55] Carolin Rösch, Tibor Kohajda, Stefan Röder, Martin von Bergen, and Uwe Schlink. Relationship between sources and patterns of vocs in indoor air. *Atmospheric Pollution Research*, 5(1):129–137, 2014. 2.2.1

[56] Heidi Salonen, Tunga Salthammer, and Lidia Morawska. Human exposure to ozone in school and office indoor environments. *Environment international*, 119:503–514, 2018. 1

[57] Norbert Serfozo, Sofia Eirini Chatoutsidou, and Mihalis Lazaridis. The effect of particle resuspension during walking activity to pm 10 mass and number concentrations in an indoor microenvironment. *Building and Environment*, 82:180–189, 2014. 1

[58] Ken Sexton and Steven B Hayward. Source apportionment of indoor air pollution. 21(2):407–418. 1

[59] A Thiem, U Schlink, X-C Pan, M Hu, A Peters, A Wiedensohler, S Breitner, J Cyrys, B Wehner, C Rösch, et al. Using non-negative matrix factorization for the identification of daily patterns of particulate air pollution in beijing during 2004–2008. *Atmospheric Chemistry and Physics Discussions*, 12(5):13015–13052, 2012. 1, 2.2.1, 2.2.4

[60] George D Thurston and John D Spengler. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan boston. *Atmospheric Environment (1967)*, 19(1):9–25, 1985. 2.2.1

[61] Bruce A Tichenor, Leslie A Sparks, James B White, and Merrill D Jackson. Evaluating sources of indoor air pollution. 40(4):487–492. 1

[62] Eus JW Van Someren, Kim Dekker, Bart HW Te Lindert, Jeroen S Benjamins, Sarah Moens, Filippo Migliorati, Emmeke Aarts, and Sophie van der Sluis. The experienced temperature sensitivity and regulation survey. 3(1):59–76. 2.1.3

[63] M Viana, TAJ Kuhlbusch, X Querol, A Alastuey, RM Harrison, PK Hopke, W Winiwarter, M Vallius, S Szidat, ASH Prévôt, et al. Source apportionment of particulate matter in europe: a review of methods and results. *Journal of Aerosol Science*, 39(10):827–849, 2008. 1

[64] Lance Wallace and Wayne Ott. Personal exposure to ultrafine particles. *Journal of Exposure Science and Environmental Epidemiology*, 21(1):20–30, 2011. 1

[65] Chao Wang, Xudong Yang, Jun Guan, Zheng Li, and Kai Gao. Source apportionment of volatile organic compounds (vocs) in aircraft cabins. 81:1–6. 2.2.1

[66] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. 3

[67] Elen Yakovleva, Philip K Hopke, and Lance Wallace. Receptor modeling assessment of particle total exposure assessment methodology data. *Environmental Science & Technology*, 33(20):3645–3652, 1999. 1

[68] Xudong Yang. *Study of building material emissions and indoor air quality*. PhD thesis. 1

[69] Zhonglong Zheng, Jie Yang, and Yitan Zhu. Initialization enhancer for non-negative matrix factorization. *Engineering Applications of Artificial Intelligence*, 20(1):101–110, 2007. 3.2

[70] Janez Žibert, Jure Cedilnik, and Jure Pražnikar. Particulate matter (pm10) patterns in europe: An exploratory data analysis using non-negative matrix factorization. *Atmospheric Environment*, 132:217–228, 2016. 2.2.1