Factorisation matricielle non négative sous contraintes. Application à l'identification de sources industrielles.

Gilles Delmaire[1], Gilles Roussel[1], Dany Hleis[2], Dominique Courcot[2]

[1]Laboratoire d'Informatique Signal Image de la Côte d'Opale,
EA 4491 – Université du Littoral Côte d'Opale
Maison de la Recherche Blaise Pascal
50, rue Ferdinand Buisson, 62228 Calais Cedex, France.
gilles.delmaire@lisic.univ-littoral.fr, gilles.roussel@lisic.univ-littoral.fr

[2]Unité de Chimie Environnementale et Interactions sur le Vivant
EA 4492 – Université du Littoral Côte d'Opale
Maison de la Recherche en Environnement Industriel.
145 Avenue Maurice Schumann. 59140 Dunkerque, France.
dominique.courcot@univ-littoral.fr, dany.hleis@univ-littoral.fr

Résumé— **Notre contribution concerne la factorisation d'une matrice d'observation X en deux matrices inconnues G (matrice de contribution) et F (matrice de profil) permettant d'identifier les différentes sources de pollutions. La recherche de G et F peut s'opérer grâce à des techniques de factorisation (PMF[1] et NMF[2]) qui optimisent la recherche de manière alternée sur G et F. Un travail précédent a permis d'établir dans un contexte de type NMF pondérée les premiers résultats sur la recherche des profils contenus dans la matrice F.**
Cependant, certaines composantes des profils présentent des incohérences. Afin de résoudre ces conflits, il est proposé d'utiliser les techniques d'optimisation sous contraintes afin de laisser libre certains composés chimiques et d'en figer d'autres dont on a la certitude. Le problème est alors équivalent à une famille de sous-problèmes quadratiques sous contraintes dont on peut exprimer la solution. Les contraintes s'expriment dans notre cas comme des contraintes linéaires égalité où une composante est forcée à zéro, ou bien à une valeur constante. Des expressions globales de mise à jours des matrices y sont proposés permettant une rapidité de calcul appréciable.

Ces algorithmes sont utilisés pour une estimation des contributions de sources de particules en suspension dans l'air, alors que de mêmes espèces chimiques se trouvent à la fois dans des émissions naturelles et industrielles. Les résultats permettent de faire disparaître les composantes indésirables de certains profils, mais aussi d'obtenir des informations plus précises sur les profils les plus mal connus.

Mots-clés— **Factorisation matricielle approchée, Diagnostic de sources, Particules en suspension, Qualité de l'air.**

I. Introduction

Les techniques de factorisation matricielle approchée sont d'un grand intérêt dans des domaines aussi variés que la reconnaissance de formes, le traitement d'antennes et aussi le traitement de données environnementales. La NMF, appelée Non negative Matrix Factorization, y est dédiée, elle a plutôt été introduite dans les domaines du traitement de signal et des images. Différentes formes de la NMF ont été développées selon le critère de minimisation choisie. Les versions les plus connues sont entre autres celles de Cichocki et Zdunek [4] et Chi Jen Li [6] basées sur une méthode de gradient projeté. Récemment, des versions pondérées de la NMF ont vu le jour, permettant de tenir compte d'incertitudes spécifiques à chaque mesure. Des incohérences dans les résultats pratiques obtenus nous incitent ici à tenir compte d'informations complémentaires exprimées sous la forme de contraintes.

[1]Postive Matrix Factorization
[2]Non negative Matrix Factorization

II. Une brève revue de méthodes de factorisation approchée.

La factorisation matricielle sous contrainte de positivité s'appuie sur un modèle de produit de matrices dont les éléments sont tous positifs. Elle se situe entre la séparation aveugle de sources d'une part, qui ne suppose aucune information a priori sur les matrices recherchées, et d'autre part les modèles de régression qui supposent connue l'une des matrices du produit. Dans notre cas, la situation sur la connaissance des matrices est intermédiaire dans le sens où certaines matrices peuvent être partiellement connues.

A. Le modèle récepteur

Ce modèle est très générique puisqu'il s'utilise en traitement du signal et de l'image. Dans le domaine de l'environnement, ce modèle de factorisation est plus connu sous le nom de modèle récepteur. Ce modèle (1) fait le lien entre la matrice de données et les différentes sources. Etant donné une matrice de données X de taille n x m, la factorisation procure l'équation approchée suivante :

$$X = GF + E \tag{1}$$

où

- X est la matrice de données de taille n x m, exprimée dans le cas de problèmes environnementaux en ng/m^3. L'élément courant x_{ij} représente la concentration de l'espèce j provenant de l'échantillon i. Une ligne représente donc la composition d'un échantillon complet.

– G est la matrice de contributions de taille n x p, avec n le nombre d'échantillons et p le nombre de sources. Dans le cas de l'application aux données environnementales, l'élément g_{ik} est mesuré en $\mu g/m^3$ et représente la contribution massique de la source k à l'échantillon i.
– F est la matrice des profils de taille p x m, avec p nombre de sources et m le nombre d'espèces étudiées. L'élément courant f_{kj} est un rapport de masse, équivalent à un pourcentage de l'espèce j par rapport à la masse totale de la source k exprimé en ng/μg.
– E est la matrice d'erreurs de taille n x m, exprimée dans la même unité que la matrice de données X. e_{ij} représente l'erreur de reconstitution issue de la factorisation de l'espèce j et de l'échantillon i.

Généralement, le nombre de sources p est choisi très inférieur au nombre d'échantillons n et aussi au nombre d'espèces considérées m. Le choix a priori de la valeur appropriée de p dépend essentiellement de la nature des données étudiées [3] et aussi de la connaissance experte sur le nombre de sources en lien avec le site étudié. En l'absence d'information, l'examen du minimum du critère en fonction du nombre de sources permet d'effectuer un choix judicieux.

B. La factorisation en matrices non négatives

La factorisation en matrices non négatives connut ses premiers succès avec les contributions de Lee and Seung [5]. De la même façon que précédemment, on s'intéresse à des factorisations approchées de $X = G\,F$ sous les contraintes de positivité minimisant le critère suivant :

$$\begin{aligned} \{G,F\} &= \arg\min_{G,F} \sum_{i=1}^{n}\sum_{j=1}^{m} \left(x_{ij} - (GF)_{ij}\right)^2 \\ &= \arg\min_{G,F} \|X - GF\|_F^2 \\ \text{sous la contrainte :} \quad & g_{ik} \geq 0 \qquad f_{kj} \geq 0 \end{aligned} \tag{2}$$

où l'indice F désigne la norme matricielle de Frobenius. D'autres fonctions de coût sont aussi populaires en particulier celles utilisant la divergence de Kullback mais nous nous limiterons à cette dernière (2). Différentes implémentations itératives de ce critère sont possibles :

– les plus connues sont les méthodes multiplicatives de mise à jour. Lee et al [5] mentionnent que la mise à jour suivante constitue un bon compromis entre rapidité et facilité d'implémentation. La mise à jour de F (resp. G) s'opère avec un coefficient de mise à jour multiplicatif qui dépend des données X et de la matrice G (resp. F) :

$$F \leftarrow F \circ \frac{(G^T X)}{(G^T G F)} \qquad G \leftarrow G \circ \frac{(X F^T)}{(G F F^T)} \tag{3}$$

où $X \circ Y$ et $\frac{X}{Y}$ désignent respectivement le produit et la division élément par élément.
Avec cette mise à jour, la norme de Frobenius est non croissante, au pire elle devient invariante si l'on a atteint un point stationnaire. Un point stationnaire cependant n'est pas une garantie d'un minimum global du critère.
– Une variante est la méthode des moindres carrés alternés (ALS : Alternative Least Squares). Il s'agit de trouver G (resp. F) en fixant F (resp. G). La formulation mathématique du problème à l'itération $r+1$ s'exprime en fonction des résultats à l'itération r :

$$\begin{aligned} F^{r+1} &= \arg\min_{F\geq 0} \|X - G^r F\|_F^2 \\ G^{r+1} &= \arg\min_{G\geq 0} \left\|X - G F^{r+1}\right\|_F^2 \end{aligned} \tag{4}$$

La résolution de chaque phase de (4) peut être vue comme la résolution d'une collection de problèmes de moindres carrés non négatifs. La recherche de F revient à :

$$\text{j-ème colonne de} \quad F^{r+1} = f_j = \arg\min_{F\geq 0} \|x - G^r f_j\|_F^2 \tag{5}$$

La résolution de ce problème est en général plus gourmande en temps de calcul que la résolution par mise à jour multiplicative. Par ailleurs, la question posée est de savoir s'il existe ou non une séquence convergente de matrices $\{G,F\}$. En optimisation, cette condition est vérifiée si le domaine est borné or il ne l'est pas ici puisque seule la borne minimale 0 est connue. Il apparaît donc important d'utiliser des algorithmes prenant en compte une borne maximale.
– Chi Jen Lin [6] propose de développer des techniques basées sur les gradients projetés pour l'optimisation sous contraintes bornées. La mise à jour itérative de la solution est réalisée selon une approche de type gradient projeté sur l'espace possible :

$$F^{r+1} \leftarrow P\left[F^r - \alpha^r \quad \nabla_{F^r}\left(\|X - G^r F\|_F^2\right)\right] \tag{6}$$

où la projection P permet de ramener le candidat à la frontière du domaine si celui-ci en sort :

$$P(x_i) = \begin{cases} x_i & \text{si } m_i \leq x_i \leq M_i \\ M_i & \text{si } x_i \geq M_i \\ m_i & \text{si } x_i \leq m_i \end{cases}$$

Beaucoup de travaux ont porté sur la recheche du pas α^r qui satisfasse une décroissance suffisamment rapide à chaque étape. Cette étape est aussi celle qui est la plus coûteuse en temps de calcul.

C. La pondération de la NMF

Récemment, certains auteurs ont proposé une version pondérée aux incertitudes de mesures de la NMF, exprimée soit comme une suite de mise à jour des colonnes de la matrice F [2] ou bien sous forme globale [1]. Le critère à minimiser est une extension du critère de la NMF classique sous forme pondérée.

$$\begin{aligned} \{G,F\} &= \arg\min_{G,F} \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\frac{(x_{ij} - (GF)_{ij})}{\sigma_{ij}}\right)^2 \\ \text{sous les contraintes :} \quad & g_{ik} \geq 0 \qquad f_{kj} \geq 0 \end{aligned} \tag{7}$$

où σ_{ij} désigne l'écart type issu de la mesure de l'échantillon i et de l'espèce j, rassemblée dans la matrice $\Sigma = \{\sigma_{ij}\}$. Une matrice de pondération globale

W peut donc être exprimée : $W = \frac{1_{n\times m}}{\Sigma\circ\Sigma}$.
Blondel et al [1] proposent une expression unifiée d'une solution itérative de ce critère (de type Newton). L'avantage majeur réside dans la mise à jour multiplicative globale de chaque itération :

$$F \leftarrow F \circ \frac{G^T(W\circ X)}{G^T(W\circ(GF))} \qquad G \leftarrow G \circ \frac{(W\circ X)F^T}{(W\circ(GF))F^T} \quad (8)$$

où $X \circ Y$ et $\frac{X}{Y}$ désignent respectivement le produit et la division élément par élément.
Cette méthode est une extension de la méthode multiplicative aux données pondérées. Ceci est particulièrement important en mesures physiques où les données présentent des incertitudes relativement différentes. D'autres critères y sont aussi envisagés, en particulier la minimisation de la divergence de Kullback pondérée. Quelle que soit le critère, la méthode développée dans cette contribution est de type Newton, sa force essentielle réside dans la mise à jour directe du tableau des profils ou des contributions. Le coût de calcul y est donc très restreint.

III. Résolution du problème sous contraintes

Dans la pratique, une recherche de profils n'est jamais totalement aveugle. Il arrive fréquemment que l'on sache que certaines composantes sont nulles. C'est cette connaissance partielle que l'on souhaite intégrer sous forme de contraintes. A ce jour, aucune contribution ne prend en charge à notre connaissance des critères de type pondérés avec contraintes.

A. Présentation des contraintes égalité

Actuellement, les contraintes que nous avons voulu prendre en compte portent sur des connaissances expertes simples. Elles rendent compte simplement de la présence ou de l'absence de certains composés dans une source incluse dans la matrice F. Par contre, nous ne disposons pas de connaissance sur la matrice G. La fomulation globale des contraintes égalité se fait grâce à deux matrices Ω (pxm) et Φ (pxm) :

$$F \circ \Omega - \Phi = 0 \quad (9)$$

Ω est une matrice binaire traduisant la présence ou l'absence de contraintes pour la source et l'espèce concernée :

$$\Omega_{ij} = \begin{cases} 1 & \text{si } F_{ij} \text{ doit être forcée.} \\ 0 & \text{sinon.} \end{cases} \quad (10)$$

Φ est la matrice des valeurs de forçage. Certains profils peuvent être forcés à des valeurs nulles ou à des valeurs non nulles.

B. Expression du problème sous contraintes

La particularité de ce problème est de devoir trouver une suite de matrices qui converge vers la bonne solution. A cela, nous y ajoutons les contraintes égalité définies dans l'équation (9). Nous cherchons donc à obtenir des formulations similaires à (8) qui incluent directement l'expression des contraintes. Nous devons donc pour le résoudre raisonner en découpant la matrice de profils (respectivement de contributions) en colonnes (respectivement en lignes). Nous proposons de formuler le problème pour la matrice de profils sachant que la même technique peut être appliquée à la matrice des contributions (seules les contraintes diffèrent). Considérons f_i^k la ième colonne de la matrice F obtenue à l'itération k. Soit aussi M_i $(l_i \times p)$ la matrice de contraintes issue de la i-ième colonne de la matrice Ω contenant l_i contraintes. Elle vérifie la relation suivante :

$$M_i f_i - \delta_i = 0 \quad (11)$$

où δ_i est l'extraction des valeurs de forçage significatives issues de Φ : $M_i\varphi_i - \delta_i = 0$
et φ_i est la i-ème colonne de la matrice Φ.
Par exemple, un cas à 5 sources et deux contraintes où la deuxième et la quatrième composantes sont forcées à des valeurs :

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \\ f_i \\ \\ \end{bmatrix} = \begin{bmatrix} 80 \\ 30 \end{bmatrix} \quad (12)$$

Le problème se formule donc sous la forme d'une optimisation quadratique sous contraintes :

$$\begin{aligned} J(f_i,\lambda_i) &= (x_i - Gf_i)^T D_{w_i} (x_i - Gf_i) \\ \text{sous les contraintes :} \quad & M_i f_i = \delta_i \quad et \quad f_i \geq 0 \end{aligned} \quad (13)$$

où x_i désigne la ième colonne de la matrice X
et $D_{w_i} = diag(w_i)$ où w_i désigne la i-ème colonne de la matrice $W = \frac{1_{n\times m}}{\Sigma\circ\Sigma}$ et Σ est la matrice des incertitudes en lien avec la matrice de données X.
L'utilisation du lagrangien [10] permet de formuler la solution sous forme d'un critère sans contrainte :

$$J(f_i,\lambda_i) = (x_i - Gf_i)^T D_{w_i} (x_i - Gf_i) + \lambda_i^T (M_i f_i - \delta_i) \quad (14)$$

où λ_i est le vecteur des multiplicateurs de Lagrange de la i-ème colonne de la matrice F.
La dérivation de ce critère quadratique conduit aux relations suivantes :

$$\begin{cases} \frac{\partial J}{\partial f_i} = & -G^T D_{w_i}(x_i - Gf_i) + M_i^T\lambda_i \\ \frac{\partial J}{\partial \lambda_i} = & (M_i f_i - \delta_i) \end{cases} \quad (15)$$

L'annulation du gradient conduit à résoudre les expressions matricielles suivantes :

$$\begin{bmatrix} G^T D_{w_i} G & M_i^T \\ M_i & 0 \end{bmatrix} \begin{bmatrix} f_i \\ \lambda_i \end{bmatrix} = \begin{bmatrix} G^T D_{w_i} x_i \\ \delta_i \end{bmatrix} \quad (16)$$

Cette équation (16) doit être résolue pour toutes les colonnes de la matrice F, soit pour l'indice i variant de 1 à m. Chacune de ces équations nécessite la recherche de $(p + l_i)$ inconnues rassemblées dans autant d'équations.
Soit $Vec(\Gamma_i)$ l'espace supplémentaire aux lignes de M_i tel que $\text{rang}\begin{bmatrix} M_i \\ \Gamma_i^T \end{bmatrix} = p$
Γ_i est de dimensions $(p \times (p - l_i))$, il vérifie la relation d'orthogonalité et de normalisation suivante :

$$\begin{cases} M_i\Gamma_i = & 0_{l_i\times(p-l_i)} \\ \Gamma_i^T\Gamma_i = & I_{(p-l_i)\times(p-l_i)} \end{cases} \quad (17)$$

B.1 Le cas $\delta_i = 0$

Intéressons nous à la solution dans le cas où $\delta_i = 0$. La comparaison de (17) et de la contrainte contenue dans (16) nous indique que f_i est une combinaison linéaire des vecteurs colonnes de Γ_i :

$$f_i = \qquad \Gamma_i \theta_i \tag{18}$$

où θ_i ($(p - l_i) \times 1$) désigne le vecteur de paramètres minimaux à rechercher dans le calcul de f_i. Le problème (14) est donc équivalent à la résolution du problème quadratique sans contrainte :

$$K(\theta_i) = (x_i - (G\Gamma_i)\theta_i)^T D_{w_i} (x_i - (G\Gamma_i)\theta_i) \tag{19}$$

On reconnait dans cette dernière équation la formulation de la NMF pondérée détaillée dans [1] appliquée au vecteur θ_i en remplaçant la matrice G par la matrice $G\Gamma_i$. Blondel & al y expriment le vecteur à l'itération k+1 en fonction du vecteur à l'itération k selon la méthode de Newton. On peut aussi se référer à l'équation (8) pour en déduire l'expression de la solution :

$$\theta_i^{k+1} \leftarrow \theta_i^k \circ \frac{\Gamma_i^T G^T (D_{w_i} x_i)}{\Gamma_i^T G^T (D_{w_i} G \Gamma_i \theta_i^k)} \tag{20}$$

En se référant à l'équation (18), on peut déduire l'expression de la ième colonne de la matrice de profil :

$$f_i^{k+1} \leftarrow f_i^k \circ \frac{\Gamma_i \Gamma_i^T G^T (D_{w_i} x_i)}{\Gamma_i \Gamma_i^T G^T (D_{w_i} G f_i^k)} \tag{21}$$

En remarquant que $\Gamma_i \Gamma_i^T = diag(1_{p\times 1} - \omega_i)$ avec ω_i la ième colonne de Ω, on peut constater qu'on sélectionne les composantes actives des coefficients multiplicatifs de (8). L'expression précédente peut être donc synthétisée au niveau matriciel sous la forme :

$$F \leftarrow F \circ \left((1_{p\times m} - \Omega) \circ \left[\frac{G^T (W \circ X)}{G^T (W \circ (GF))} \right] \right) \tag{22}$$

Dans cette dernière équation, $(1_{p\times m} - \Omega)$ doit être vu comme le masque de contraintes appliqué aux coefficients multiplicatifs de F.

B.2 Le cas δ_i quelconque

Reprenons l'équation (16) que nous allons entreprendre de résoudre dans le cas où δ_i est quelconque. Les équations (11) et (17) montrent qu'une colonne du profil peut s'exprimer :

$$f_i = \varphi_i + \Gamma_i \theta_i \tag{23}$$

Le critère à minimiser devient alors sans contrainte, il est aussi à rapprocher de l'équation (19) :

$$K(\theta_i) = (x_i - G\varphi_i - (G\Gamma_i)\theta_i)^T D_{w_i} (x_i - G\varphi_i - (G\Gamma_i)\theta_i) \tag{24}$$

On remarque l'analogie du critère avec l'équation (19) dans laquelle les données x_i sont remplacées par les données équivalentes $x_i - G\varphi_i$. Il en découle en utilisant cette remarque et en s'inspirant de l'équation (20) la nouvelle expression des paramètres réduits à l'itération $k + 1$:

$$\theta_i^{k+1} \leftarrow \theta_i^k \circ \frac{\Gamma_i^T G^T D_{w_i} (x_i - G\varphi_i)}{\Gamma_i^T G^T (D_{w_i} G \Gamma_i \theta_i^k)} \tag{25}$$

L'expression (23) permet de mettre à jour la ième colonne de la matrice de profil F :

$$f_i^{k+1} - \varphi_i \leftarrow \left(f_i^k - \varphi_i\right) \circ \frac{\Gamma_i \Gamma_i^T G^T D_{w_i} (x_i - G\varphi_i)}{\Gamma_i \Gamma_i^T G^T (D_{w_i} G (f_i^k - \varphi_i))} \tag{26}$$

L'expression précédente peut être unifiée dans une expression globale similaire à l'équation (22) en remarquant que la matrice $\Gamma_i \Gamma_i^T$ agit comme un operateur de sélection de composantes équivalent au masque de contraintes $(1_{p\times m} - \Omega)$:

$$F - \Phi \leftarrow (F - \Phi) \circ \left((1_{p\times m} - \Omega) \circ \left[\frac{G^T (W \circ (X - G\Phi))}{G^T (W \circ (G(F - \Phi)))} \right] \right) \tag{27}$$

Cette équation est l'expression globale unificatrice qui permet de mettre à jour les itérées des matrices de profil. Elle est la généralisation de l'équation (22) et est à comparer à l'expression (8) qui correspond à la version sans contrainte.

C. Résumé de l'algorithme

Au vu des propriétés précédentes, l'algorithme peut être synthétisé sous la forme suivante :

```
Initialiser G et F
normalisation de F et G
Tant que le critère d'arrêt n'est pas atteint    {
    recherche de F à G constant selon (27)
    normalisation de F et G
    recherche de G à F constant selon (8)        }
```

La normalisation de F et G est faite ici à chaque pas d'itération en normalisant chaque profil de source puis en réactualisant les contributions associées dans G. Elle permet juste de conserver une cohérence dans la matrice des profils, en aucun cas, elle ne modifie la valeur du critère à optimiser.

IV. Application à un cas réel

A. Introduction

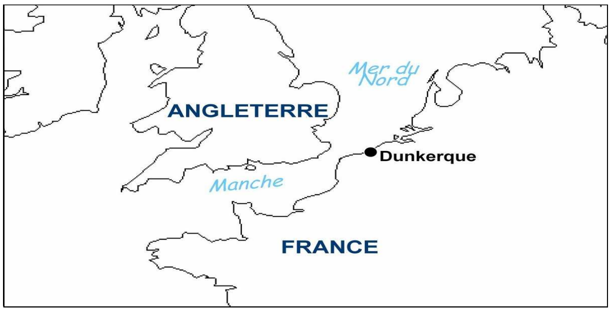
Nous avons travaillé sur des données recueillies à Dunkerque, un centre urbain situé sur le littoral de la Mer du Nord (Figure 1) où siège une importante activité industrielle, et notamment des installations sidérurgiques. Ce travail constitue la deuxième étape d'interprétation des origines des particules en suspension dans l'air, alors que la première fut consacrée à des influences essentiellement naturelles. Cette fois, nous nous sommes focalisés sur des échantillons collectés en un même site, mais lorsque les directions de vent couvraient à la fois le secteur marin et la zone industrielle de Dunkerque. 89 échantillons ont ainsi été extraits d'un ensemble d'échantillons collectés au cours de 2 campagnes de mesure. Les teneurs en 19 éléments et ions (donnés dans la table I) ont été quantifiées dans chacun des échantillons.

Dans [2], nous avons mis en évidence la présence de 4 sources naturelles que nous avons appelées sources « Mer », « Mer Agée », « Continent » et « Crustale ». Leurs profils ont été validés dans [2] et dans la littérature spécialisée [7] [8] [9]. Nous avons donc considéré que ces sources émettaient de manière permanente avec un profil relativement stable que nous avons figé dans nos recherches.

La zone industrielle de Dunkerque possède un site

sidérurgique important, qui rassemble les installations qui assurent i) la fabrication du coke ii) l'agglomération des minerais de fer, iii) l'obtention de la fonte dans des hauts-fourneaux, iv) l'élaboration de l'aciérie et v) le traitement des co-produits (laitiers). Ces installations émettent des particules de taille inférieure à 10 μm. Néanmoins, ces particules présentent des différences de composition chimique d'une source à l'autre [11] et ne sont pas émises, ni simultanément, ni de façon constante en terme de flux. Pour ces raisons, il n'est pas approprié de retenir un profil unique pour ce type de site industriel. Il est alors plus judicieux de chercher à isoler les différentes sources qui ont un impact significatif sur la qualité de l'air.

Pour y parvenir, nous avons adopté une démarche progressive, en considérant 2 sources industrielles puis 3 mais nous reportons ici uniquement le cas le plus significatif c'est à dire 3 sources industrielles.



les secteurs industriels

Fig. 1. Localisation de la zone d'étude

B. *Identification des principales sources*

Nous proposons de présenter le cas considérant 3 sources industrielles en plus des 4 sources naturelles, soit 7 sources au total. Les matrices relatives aux contraintes à savoir Ω et Φ sont données respectivement en table I et II. Elles s'expriment en colonnes en fonction des espèces et en lignes en fonction des sources : les quatre premières sont les sources naturelles, le trois autres représentent les sources industrielles.

TABLE I

Matrice de contraintes Ω

Al	Ca	Cr	Cu	Fe	K	Mg	Mn	Na	Ni	Pb	Sn	Ti	V	Zn	Cl^-	NO_3^-	SO_4^{2-}	NH_4^+
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	1	1	0	1
0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	1	1	1	1
0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0

TABLE II

Matrice de Forçage Φ

Al	Ca	Cr	Cu	Fe	K	Mg	Mn	Na	Ni	Pb	Sn	Ti	V	Zn	Cl^-	NO_3^-	SO_4^{2-}	NH_4^+
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	421	316	263
367	140	.2	.1	147	136	64	2.5	122	.1	.1	0	15	.3	.3	3	2.4	.5	.5
0	10	0	0	0	9	30	0	253	0	0	0	0	0	0	200	280	217	0
0	12	0	0	0	12	37	0	312	0	0	0	0	0	0	550	0	78	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Au total, 36 paramètres sont libres dans la matrice de profil F répartis dans les trois dernières lignes qui correspondent aux trois sources industrielles. Concernant la matrice de contribution, aucune contrainte n'est envisagée. Elle contient donc 89×7 paramètres libres.

L'estimation de ces quantités est rassemblée par source, à savoir sur une figure, on donne en premier lieu une colonne de G représentant la contribution en masse de chaque source puis la même quantité instantanée exprimée en pourcentage de la masse totale instantanée. Enfin, la dernière courbe représente une ligne de F composant le profil chimique de la source. La cinquième source (appelée Hauts fourneaux) est représentée sur la figure 2 tandis que la source 6 (appelée Laitiers d'aciérie) est représentée sur la figure 3. Enfin la source 7 appelée Unité d'agglomération est donnée figure 4.

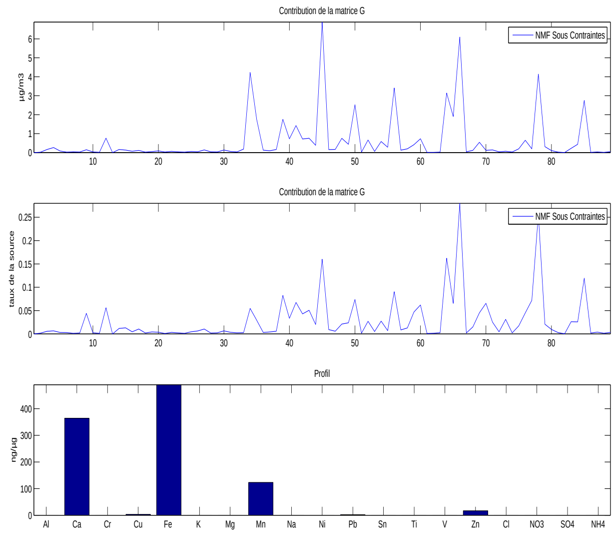


Fig. 2. Profil de la source Hauts Fourneaux

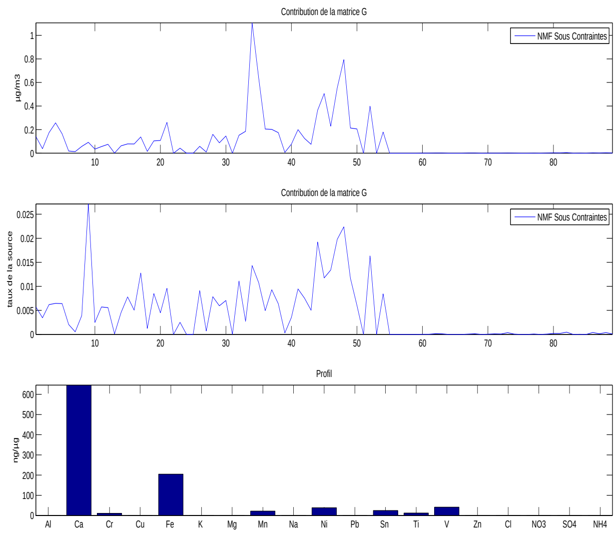


Fig. 3. Profil de la source laitiers d'aciérie

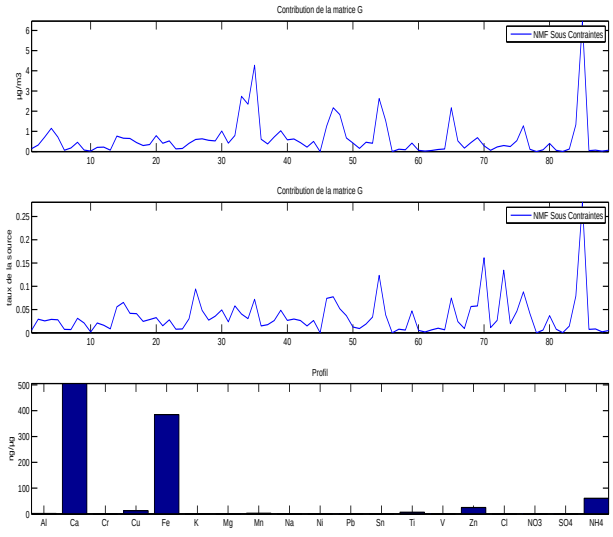


Fig. 4. Profil de la source Unité d'agglomération

Les différents profils obtenus sont caractérisés en particulier par la présence des éléments Fe, Ca, Al, Mg, Mn, Pb, Cu, Zn, Cr, K et NH_4^+ dans des proportions différentes. En confrontant ces profils avec des données expérimentales définies pour chaque source [11], le profil de la figure 4 se réfère aux particules riches en fer émises au niveau des hauts-fourneaux ou de l'aciérie, qui correspondent à deux sources présentant des grandes similitudes de composition chimique. Ces caractéristiques ont également été reportées pour les mêmes installations présentes sur d'autres sites dans le monde [12] [13]. Ces particules riches en fer correspondent à des émissions diffuses qui se produisent lors des phases de coulée et de brassage de la fonte ou de l'acier, alors que ces matières sont dans un état liquide à haute température.

Le deuxième profil (figure 3) est en revanche marqué par une contribution élevée de Ca, puis outre le fer, un ensemble d'impuretés métalliques sont observées. Sans ambiguïté, ce profil peut être attribué aux fines particules de co-produits, qui regroupent l'ensemble des éléments non désirés dans la composition de l'acier et qui sont séparés sous forme de laitiers. Il s'agirait ici de laitiers d'aciérie, qui présentent en général une granulométrie faible.
Enfin, le troisième profil (figure 4) se distingue par la présence d'éléments Ca supérieurs au Fe et l'ion NH_4^+ en quantité non négligeable. Lors de la réalisation des analyses chimiques de différents échantillons de source [11], ces caractéristiques ont effectivement été retrouvées dans l'Unité d'Agglomération des minerais où se produit le chauffage des matières premières (minerais, carbonate de calcium et autres fondants). Cette étape permet d'obtenir une matière agglomérée introduite ensuite dans les hauts-fourneaux. Bien que des éléments comme Al, K, Mg et Pb en faibles teneurs, étaient aussi attendus dans cette source, cette attribution reste la plus vraisemblable.

Au niveau des contributions de chacune de ces sources, il n'est pas observé de répercussion permanente sur la qualité de l'air ambiant, mais leur impact se traduit plutôt sous la forme de pics. De plus, la contribution de ces différentes sources industrielles n'est pas détectée de façon simultanée, ce qui confirme qu'un tel site doit être considéré comme un site multi-sources.
L'étude du cas deux sources tendrait à montrer que les profils des sources 5 et 7 sont rassemblés en une seule source, hypothèse qui parait beaucoup moins vraisemblable.
En conclusion, l'analyse de la configuration à 3 sources industrielles semble la plus réaliste.

V. Conclusion

Nous avons présenté dans cet article comment l'intégration de contraintes égalité modifiait d'un point de vue théorique les algorithmes de factorisation matricielles non négatives pondérés. Notre originalité est d'abord de proposer des formulations globales d'actualisation des matrices de profils dont le coût calculatoire est très réduit. Notre seconde originalité réside dans l'application elle même qui vise à rechercher les profils d'émissions de polluants naturels et industriels. La connaissance partielle de ces profils permet de centrer la recherche sur les composés des sources les plus mal connus. Les résultats sont tout à fait encourageants et permettent de cerner avec plus de précision les profils industriels.

Références

[1] Blondel V., Ngoc-Diep Ho et Van Dooren P. Algorithms for Weighted Non Negative Matrix Factorization. *Image and Vision Computing*, soumis.

[2] Delmaire G., Roussel G., Hleis D. et Ledoux F. Une version pondérée de la Factorisation Matricielle Non négative pour l'identification de sources de poussières. Application au littoral de la Mer du Nord. Conférence STIC et Environnement 2009, 14-16 Juin 2009, Calais, France.

[3] Guillamet D., Vitria J. et Schiele B.. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, vol. 24, n° 14, pp. 2447-2454, 2003.

[4] Cichocki A. et Zdunek R.. NMFLAB for Signal Processing. Laboratory for Advanced Brain Signal Processing, BSI RIKEN, Saitama, Japan *www Report*, Décember 2006.

[5] Lee Daniel D. et Sebastian Seung H. Learning the parts of objects by non negative matrix factorization.*Nature*. vol. 401, n° 6755, pp. 788-791, 1999.

[6] Chih Jen Lin. Projected Gradients Methods for Non-Negative Matrix Factorization.*Neural Computation*. vol. 19, n° 10, pp. 2756-2779, 2007.

[7] Evans M. C. et S. W. Campbell et V. Bhethanabotla et N. D. Poor. Effect of sea salt calcium carbonate interactions with nitric acid on the direct dry deposition of nitrogen to Tampa Bay, Florida. *Atmospheric Environment*. vol. 38, n° 29, pp 4847-4858, 2004.

[8] Ten-Harkel M. J.. The effects of particle-size distribution and chloride depletion of sea-salt aerosols on estimating atmospheric deposition at a coastal site. *Atmospheric Environment*. vol. 31, n° 3, pp 417-427, 1997.

[9] Zhao Y. et Y. Gao. Acidic species and chloride depletion in coarse aerosol particles in the US east coast. *Science of the total Environment*. vol.407, n° 1, pp 541-547, 2008.

[10] Boyd S. et Vandenberghe L., *Convex optimization*. Cambridge University Press, 2004.

[11] Laversin H. .Traceurs et formes chimiques du fer dans les particules émises dans l'atmosphère depuis un site sidérurgique : Etude spectroscopique et caractérisation de composés de référence et de particules collectées dans l'environnement. *Thèse de l'Université du Littoral Côte d'Opale*, 2006.

[12] Machemer S. D. .Characterization of airborne and bulk particulate from iron and steel manufacturing facilities. *Environ. Sci. Techn.*, vol 38, n°2, pp 381-389, 2004.

[13] K. Oravisjarvi, K.L. Timonen, T. Wiikinkoski, A.R. Ruuskanen, K. Heinanen et J. Ruuskanen. Source contributions to PM2.5 particles in the urban air of a town situated close to a steel works. *Atmospheric Environment*. vol 37, pp 1013-1022, 2003.