Conception d'un réseau de capteurs pour l'estimation des sources de pollution chronique par modélisation inverse

Network design for chronic pollution source monitoring estimated by inverse modeling

Anis KHLAIFI*, Anda IONESCU*, Yves CANDAU*

Résumé

L'objectif de ce papier est l'estimation des émissions des sources chroniques et connues par modélisation inverse. Le cas d'étude correspond à une zone industrielle située en banlieue parisienne, où il y a trois sources émettrices de quantités importantes de SO_2. Premièrement, nous avons testé dans quelle mesure on pouvait résoudre ce problème en utilisant les capteurs du réseau de surveillance de la qualité de l'air existant dans la région (trois stations). Des mesures de paramètres météorologiques sont également disponibles dans le voisinage des sources. Le problème inverse a été abordé par minimisation de l'écart entre un modèle direct (le modèle gaussien de dispersion de Pasquill) et les mesures, en utilisant plusieurs techniques d'optimisation (algorithmes génétiques, Gauss-Newton...). Bien que dans certains cas de figure (récepteurs sous le vent des sources), l'ordre de grandeur des émissions soit bien restitué, parfois avec une erreur inférieure à 10 % pour au moins deux sources, d'autres configurations (direction du vent-positions des capteurs) ont conduit à un problème mal posé, qui ne permet pas d'estimer les émissions des sources, ces derniers cas correspondant à des situations météorologiques fréquentes. Cette constatation a mis en évidence la nécessité de concevoir un réseau de capteurs, optimisé pour estimer les sources pour les situations atmosphériques les plus fréquentes. La première étape de la conception du réseau a été la création des mesures semi-synthétiques aux capteurs fictifs répartis uniformément et en grand nombre sur toute la région autour des sources, à partir des données météorologiques réelles et des caractéristiques des sources (plage de variation des débits d'émission), en utilisant le modèle de Pasquill. Ensuite, on a procédé à la sélection des capteurs les plus « prometteurs » au sens de la précision d'estimation des émissions des sources, la méthode d'inversion étant celle testée dans le cas du réseau existant. La redondance a été réduite en utilisant des méthodes statistiques comme la classification ascendante hiérarchique ou l'analyse en composantes principales à noyaux. Au final, avec une dizaine de capteurs on arrive à obtenir des estimations parfaites des émissions des sources dans environ 70 % des cas étudiés.

Mots clés

Modélisation inverse. Estimation des sources. Modèle gaussien de dispersion. Algorithmes génétiques. Conception d'un réseau de capteurs. Classification. Analyse en Composantes Principales à Noyaux

Abstract

This paper focuses on the estimation of the emissions from known sources by inverse modeling. The study case corresponds to an industrial zone in Paris outskirts, where there are three important sources of SO_2. First, we tested to what extent it was possible to solve this problem by using the 3 existing sensors of the existing air quality monitoring network in the area. Meteorological parameters in the sources neighbourhood are available from measurements. The inverse problem has been solved by coupling a direct diffusion model (Pasquill's Gaussian model) and the minimization of an error criterion, by several techniques (genetic algorithms, Gauss-Newton...). Good results are obtained when the monitoring stations are downwind from the sources, and in these cases, the order of magnitude of emissions is retrieved, sometimes with less than 10% error for at least two sources. Some configurations however lead to an ill-posed problem, where it is not possible to retrieve emissions and these cases correspond to frequent meteorological conditions. The latter situations reveal the need to conceive a specific network of sensors, taking into account the source locations and the most frequent weather patterns. The first stage of this network design consisted in simulating semi-synthetic measurements at virtual sensors, uniformly distributed and in large number over the region around the sources, using Pasquill's model applied to actual meteorological data and source emissions (generated from source characteristics). Then, a selection of the best sensors has been performed, according to their potential to retrieve the source emissions. The redundancy of the virtual sensors has been reduced using statistical methods such as hierarchical classification or kernel principal component analysis. By using about ten sensors, perfect estimations of source emissions are obtained in about 70% of the studied cases.

Keywords

Inverse modeling. Source identification. Gaussian model of dispersion. Genetic algorithms. Network design. Classification. Kernel Principal Component Analysis.

* *Centre d'études et de recherche en thermique, environnement et systèmes – Université Paris Est Créteil – 61, avenue du Général de Gaulle – 94010 Créteil Cedex – E-mail : khlaifi.anis@gmail.com, ionescu@u-pec.fr, candau@u-pec.fr*

1. Introduction

L'identification des sources de pollution et de leurs contributions à partir des mesures dans leur environnement peut être traitée par deux approches, adaptées à deux problématiques différentes.

La première problématique concerne l'identification des sources en aveugle, par leurs profils (ou « signatures »). Il s'agit de sources complexes, dont le profil d'émission est inconnu et comprend plusieurs espèces. La résolution du problème de mélange à partir de la composition chimique des particules échantillonnées passe habituellement par des méthodes statistiques d'analyse factorielle [1, 2]. Les mesures utilisées proviennent le plus souvent des expériences spécialement conçues.

Pour la deuxième problématique, la différenciation entre les sources ne se fait plus par leur profil, car il s'agit d'une seule espèce chimique, et l'objectif est d'estimer l'intensité d'émission des sources. La littérature présente aussi bien des cas de sources chroniques, de localisation connue [3], voire des cadastres d'émission, que des cas de sources accidentelles et non localisées [4]. Souvent, sont utilisées les méthodes de rétro-trajectoires [5-7], mais on rencontre également l'utilisation conjointe des modèles directs de dispersion des polluants et des observations. Cette utilisation conjointe modèle + mesure peut se faire par des techniques plus élaborées, comme celles relevant de l'assimilation des données [8, 9], ou par la résolution d'une minimisation modèle-mesure par une technique d'optimisation plus ou moins sophistiquée [3]. Les modèles de transport utilisés peuvent aller de modèles simples (Pasquill [10, 11]) à des modèles plus complexes (reposant sur la résolution des équations de Navier-Stokes), un module traitant la chimie étant parfois intégré. Les mesures de polluants utilisées dans cette approche proviennent assez souvent des réseaux de surveillance de la qualité de l'air [4, 8].

Des réseaux de surveillance ont été conçus au niveau d'un continent, d'un pays ou d'une région. Dans toutes les agglomérations urbaines françaises de plus de 250 000 habitants, la Loi sur l'air de 1996 impose l'implantation d'un réseau de surveillance de la qualité de l'air. La conception du réseau prend en compte une multitude de critères, comme le mode d'occupation des sols, en plus d'une modélisation déterministe donnant des informations sur la variabilité spatiale du phénomène. Cependant, la conception de ces réseaux n'est pas orientée vers la surveillance des sources chroniques.

Le but du travail présenté dans ce papier est d'estimer les émissions des sources chroniques et connues, sources situées dans une zone industrielle, à partir des mesures de pollution effectuées dans leur environnement. Dans un premier temps, nous avons étudié dans quelle mesure on pouvait quantifier les émissions de ces sources à partir des stations de mesure du réseau de surveillance de la qualité de l'air existantes dans la région, en utilisant également les paramètres météorologiques de la région.

La méthodologie a été basée sur l'inversion d'un modèle gaussien de diffusion par linéarisation, optimisation locale ou optimisation combinatoire-algorithmes génétiques, pour un cas d'étude réel : une zone industrielle francilienne. Les concentrations horaires de polluants, mesurées par trois capteurs du réseau de surveillance de la qualité de l'air pendant une journée ont permis dans certains cas une estimation assez précise des trois principales sources existantes, tandis que dans d'autres cas, cette inversion a échoué. L'analyse des résultats, principalement en rapport avec le vent dominant, nous a permis de conclure sur les situations conduisant à un problème mal posé, et sur l'insuffisance du réseau de qualité de l'air existant pour permettre une surveillance des trois sources. Pour répondre à cette question, un réseau doit être conçu spécifiquement avec cet objectif fixé, et une méthodologie est proposée dans ce papier.

2. Description du site et des données disponibles

Le site d'étude est une zone industrielle de la grande banlieue francilienne, d'environ 130 km², située à 50 km au nord-ouest de Paris. Les principales sources de pollution soufrée de la région (99 % des émissions déclarées dans le cadre du principe pollueur-payeur) sont : une centrale EDF, à Porcheville (28 624 tonnes SO_2/an), une usine Renault, à Flins (1 440,7 tonnes SO_2/an) et une chaufferie, à Somec (878,1 tonnes SO_2/an), localisées le long de la vallée de la Seine (Figure 1).

Notre étude concerne une seule journée. Les débits horaires d'émissions sont disponibles *via* l'étude menée par Avila Galarza [12], qui les a estimés à partir des paramètres process (combustion), de même que les températures et les vitesses des gaz à la sortie de la cheminée.

Figure 1.
Site d'étude : points de mesure (■) et sources de pollution (●) (à partir d'une carte IGN, BD ORTHO).
Study zone map: source locations are indicated using the (●) symbol and monitoring stations, by (■) (based on a map from the National Geographic Institute (IGN, BD ORTHO)).

Nous disposons également d'une base de données météorologiques tri-horaires enregistrées par la station de Météo-France située à Trappes, pendant deux ans : vitesse et direction du vent, température ambiante, pression atmosphérique et nébulosité. Malheureusement, la station se trouve un peu loin du site d'étude, à 28 km Sud-Est de la principale source, qui est la centrale EDF de Porcheville.

Des mesures de concentrations horaires de SO_2 sont relevées en continu par le réseau de surveillance de la qualité de l'air de la région d'Ile-de-France, AIRPARIF, sur les sites de Bonnières-sur-Seine, Les Mureaux, Mantes-la-Jolie et Limay. Pendant la journée d'étude, le capteur de Bonnières-sur-Seine n'a jamais été sous le vent d'aucune des trois sources, donc il a été ignoré dans cette étude.

3. Formulation du problème inverse d'estimation des sources

L'objectif est d'identifier les paramètres d'émission pour les trois principales sources de SO_2, à partir des mesures de concentrations de SO_2 effectuées par des capteurs situés à proximité et des paramètres météorologiques. On suppose que les sources sont chroniques, de localisation connue et, de plus, qu'on a une idée de l'ordre de grandeur de l'émission de chaque source (par exemple à travers les émissions annuelles déclarées dans le cadre du principe « pollueur-payeur »).

3.1 Formulation générale

L'identification des paramètres d'émission des principales sources du site d'étude peut se présenter sous la forme d'un problème d'optimisation : on cherche

$$\min_X \| C_{mes,\,i} - C_{th,\,i} \| \text{ pour } i = 1 \ldots N_{capteurs},$$

où $C_{mes,\,i}$ représente la concentration de SO_2 mesurée par le capteur i (µg.m^{-3}), alors que $C_{th,\,i}$ est la concentration théorique de SO_2 calculée au capteur i pour le cas où l'émission est représentée par les valeurs de X.

La solution du problème, X, est composée d'un triplet des paramètres d'émission pour chacune des trois sources : $X = \big((Q_1, T_1, V_1,), (Q_2, T_2, V_2), (Q_3, T_3, V_3)\big)$, où Q représente le débit d'émission (g.s^{-1}), T, la température (K), et V la vitesse d'émission (m.s^{-1}).

Pour obtenir la concentration théorique $C_{th,\,i}$ au niveau du capteur i, on cumule les concentrations $C^j_{th,\,i}$ des trois sources j au niveau du même capteur :

$$C_{th,\,i} = \sum_{j=1}^{Nsources\,=\,3} C^j_{th,\,i}\,(X).$$

La concentration $C^j_{th,\,i}$ provenant d'une source j mesurée par le capteur i peut être calculée à partir du modèle Gaussien [10, 11, 13] par la formule :

$$C^j_{th,\,i}\,(X) = \frac{Q_j}{2\pi\sigma_y\sigma_z U} \times \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right] \text{x}$$

$$\left\{\exp\left[-\frac{1}{2}\left(\frac{z - H(X)}{\sigma_z}\right)^2\right] + \exp\left[-\frac{1}{2}\left(\frac{z + H(X)}{\sigma_z}\right)^2\right]\right\},$$

avec : U vitesse du vent (m.s^{-1}), y distance sous le vent (m), z distance verticale au-dessus du sol (m) ; σ_y et σ_z (m) expriment la diffusion turbulente suivant les directions y et z, alors que H, la hauteur effective d'émission (m), inclut d'une part h_s, la hauteur physique de la cheminée (m) et d'autre part Δh, la surélévation du panache (m) : $H = h_s + \Delta h$.

La surélévation du panache peut être évaluée par la formule de Holland [13] : $\Delta h = \big(1.5 \cdot V \cdot d_s + 4.10^{-5} C_p \cdot Q \cdot (T - T_a)\big) \cdot U^{-1}$, où d_s est le diamètre de la cheminée (m), C_p la chaleur spécifique du polluant à pression constante (J/g.K) et T_a la température ambiante (K).

3.2 Approches de résolution

Le problème inverse défini dans la section 3.1. est non-linéaire (conséquence de la forme analytique du modèle gaussien) et sous-déterminé (9 paramètres d'émission des sources et 3 capteurs). Il peut être résolu de plusieurs manières.

(i) Une première approche est de considérer la surélévation du panache Δh connue [14, 15], bien qu'elle dépend des paramètres d'émission ; le problème est alors déterminé et linéaire.

(ii) En supposant connues seulement les températures et vitesses d'émission, le problème est déterminé mais non-linéaire, la solution peut être recherchée par une technique d'optimisation locale, au sens des moindres carrés non-linéaires, par exemple la méthode de Gauss-Newton [16, 17].

(iii) La résolution du problème non-linéaire et sous-déterminé (tel que défini dans la section 3.1.) peut se faire par une méthode d'optimisation globale s'appuyant sur les algorithmes génétiques [3, 18].

La résolution de ce problème d'optimisation par des algorithmes génétiques présente quelques avantages, provenant du fait que ces algorithmes :

- utilisent un codage de paramètres et non les paramètres eux-mêmes ; ceci est un critère fondamental de différenciation entre les algorithmes génétiques et ceux d'exploration ;
- travaillent sur une population de points au lieu d'un point unique, ce qui permet une recherche globale et donc plus performante ;
- assurent la convergence vers des solutions optimales (ensembles de solutions) et non pas une solution unique (quasi-solutions du problème d'inversion).

4. Résultats de l'inversion pour l'estimation des sources à partir du réseau existant

Le problème d'identification des paramètres d'émission des trois principales sources du site formulé dans la section 3.1 a été résolu par les trois approches présentées dans la section 3.2.

Sans décrire la méthodologie, on peut faire quelques brefs commentaires avant de présenter les résultats obtenus.

(i) Inversion dans le cas linéaire : une étude de sensibilité avait montré que Δh n'était pas l'un des paramètres les plus influents et que cette hypothèse n'avait pas de conséquences majeures sur les résultats [14].

(ii) La résolution au sens des moindres carrés non-linéaires montre une fluctuation importante par rapport aux débits réels. Les valeurs obtenues sous-estiment les débits réels mesurés.

(iii) Les algorithmes génétiques offrent la possibilité la plus intéressante pour la recherche de l'optimum pour les raisons présentées dans la section 3.2. Leur implémentation (en version mono- et multi-objectif) : codage des paramètres, analyse de configurations et choix d'une solution de la famille des solutions possibles en fonction de leurs caractéristiques, tous ces aspects sont détaillés dans [14] et quelques-uns présentés succinctement dans [3, 18].

Les résultats obtenus à travers ces trois approches sont présentés dans le Tableau 1.

Tableau 1.
Analyse comparative entre les différentes méthodes utilisées pour l'estimation des débits d'émission des trois sources.
Comparative analysis of the various methods used for the estimation of the emission flows of the three sources.

Source	Heure	Débit estimé				Débit réel
		Inversion linéaire (Δh connu)	Optimisation locale (Gauss-Newton)	Algorithmes génétiques mono-objectif	Algorithmes génétiques multi-objectif	
EDF	10	1 165	1 189	1 301	1 377	1 223
	11	11.8	377	423	447	1 219
	12	18	798	777	1 728	1 637
	13	2.3	795	1 375	1 701	2 739
	14	31.7	1 274	1 766	1 576	2 748
	15	5.1	1 027	2 294	2 898	2 743
	16	9.8	768	1 452	1 944	2 744
	17	558	550	1 701	1 779	2 738
	18	0.9	342	398	2 767	2 722
RENAULT	10	255	159	253	279	171
	11	77.6	21.6	18.4	12	171
	12	160	73.4	77	20.3	171
	13	143	131	122	151	171
	14	298	167	127	138	171
	15	154	121	85.8	91.1	171
	16	134.1	79.3	35.3	100	171
	17	35.8	85.1	180	88.3	171
	18	29.9	20.1	32.3	10.4	171
SOMEC	10	144	100	100	97.2	8.9
	11	15.5	16.2	10	61.8	10.6
	12	22.9	15.4	12.5	17.7	10.2
	13	13.6	14.7	13.3	20.1	10.9
	14	12.5	6.8	6.9	9.1	10.9
	15	13.6	12.3	8.8	13	11
	16	14	9.2	7.3	6.4	10.5
	17	14.8	15.8	16	16	26.9
	18	2.6	1	2.5	1.2	30.9

De manière générale, on constate que l'information contenue dans les mesures des différents capteurs du site d'étude était suffisante « en partie » pour restituer les paramètres d'émission des principaux émetteurs de SO_2 du site d'étude.

Dans certains cas, la reconstitution des sources est possible, et ceci, indépendamment de l'importance de leurs rejets. D'autre part, on met aussi en évidence des cas de problème mal posé ; ces cas correspondent aux situations météorologiques dans lesquelles les récepteurs ne sont pas sous le vent des sources, les mesures aux récepteurs sont alors de l'ordre du bruit de fond. Une analyse de ces résultats est détaillée dans [14].

Cette pré-étude permet d'en tirer deux conclusions importantes.

- Premièrement, la méthode d'inversion mise au point par couplage du modèle direct de Pasquill avec une technique d'optimisation a prouvé son efficacité sur des données réelles. En effet, bien que les émissions soient d'ordre de grandeur assez différent pour les trois sources, il existe un ensemble de situations où elles peuvent être reconstituées de manière assez satisfaisante, parfois avec une erreur inférieure à 10 % pour au moins deux sources.
- Deuxièmement, l'analyse des résultats, principalement en rapport avec le vent dominant, nous a permis de conclure sur les situations conduisant à un problème mal posé, et sur l'insuffisance du réseau de qualité de l'air existant pour permettre une surveillance des trois sources. Pour répondre à cette objectif de surveillance des sources, un réseau spécifique doit être conçu.

5. Conception d'un réseau de mesure pour l'estimation des sources

5.1 Création de capteurs virtuels et données semi-synthétiques

Un maillage régulier de 10 × 10 capteurs fictifs a été défini, couvrant le domaine où sont placées les sources, soit une zone de 22 km × 21 km, assurant un bon recouvrement de l'espace d'étude.

On génère aléatoirement des valeurs d'émission (débit, température et vitesse) tri-horaires pour une période de deux ans, en respectant le domaine de variation des émissions de chaque source (déduit à partir des mesures couvrant un mois). Cette période de deux ans a été choisie pour pouvoir utiliser une base de données météorologiques tri-horaires réelles : enregistrées par la station de Météo-France située à Trappes (vitesse et direction du vent, température ambiante, pression atmosphérique et nébulosité).

On crée ensuite des concentrations tri-horaires de SO_2, aux 100 capteurs fictifs, pour les deux ans d'étude. Pour cela, on utilise le modèle gaussien, appliqué aux données météorologiques réellement mesurées à Trappes et aux émissions générées aléatoirement (entre les limites inférieure et supérieure d'émission de chacune des trois sources) ; on peut qualifier ces concentrations de « semi-synthétiques », car elles ont été créées, en partie, en utilisant des données réelles.

Pour chaque capteur fictif, on cherche la valeur maximale enregistrée et on identifie la situation (émissions + météo) correspondante. Les journées où une telle situation s'est produite sont alors gardées pour l'étude. La période de deux ans a ainsi été réduite à 66 jours d'étude, les jours les plus représentatifs, soit 528 scénarios contenant les concentrations aux 100 capteurs fictifs.

5.2 Analyse de la redondance du réseau par un indice de performance

Les 100 capteurs fictifs initialement créés sont uniformément distribués sur le domaine d'étude, et en grand nombre. Dans une première étape, on étudie l'importance du nombre de capteurs pour l'estimation de chaque source, dans le but de restreindre le réseau initial.

L'estimation de chaque source est faite selon la procédure d'inversion présentée dans la section 3, à partir des concentrations semi-synthétiques. Il est évident que la performance de cette inversion dépend des capteurs choisis. Pour cela, on définit un indice de performance I_p propre à chaque source :

$$I_p = \frac{\text{Nombre de cas d'estimation exacte}}{\text{Nombre total de cas étudiés}},$$

indice variable avec le nombre de capteurs.

On a choisi de commencer par un réseau minimal (trois capteurs) et d'augmenter progressivement sa taille. Le choix du capteur ajouté à chaque pas est basé sur un critère que nous avons appelé « fréquence de détection », défini comme nombre de fois où un capteur mesure des concentrations non nulles pendant la période d'étude.

On étudie ensuite la variation de l'indice de performance du réseau en fonction du nombre de capteurs utilisés, pour chaque source (Figure 2). Lorsqu'on rajoute un nouveau capteur, on peut constater :

- une augmentation de l'indice de performance du réseau : sa présence dans le réseau est importante ;
- un palier : le nouveau capteur n'améliore pas les performances obtenues, mettant en évidence une certaine redondance entre les capteurs, dans ce cas, leur nombre pourrait être réduit, un groupe présenté sur un palier remplacé par un seul capteur.

Ces deux situations sont présentes pour les trois sources, mais, il faut noter qu'elles ne concernent pas les mêmes capteurs.

Les configurations des réseaux obtenus en choisissant les dix meilleurs capteurs (du point de vue indice de performance) pour chaque source sont présentés dans la figure 3 : (a) pour la source EdF (« réseau » EdF), (b) pour la source Renault (« réseau Renault ») et (c) pour la source Somec (« réseau Somec »).

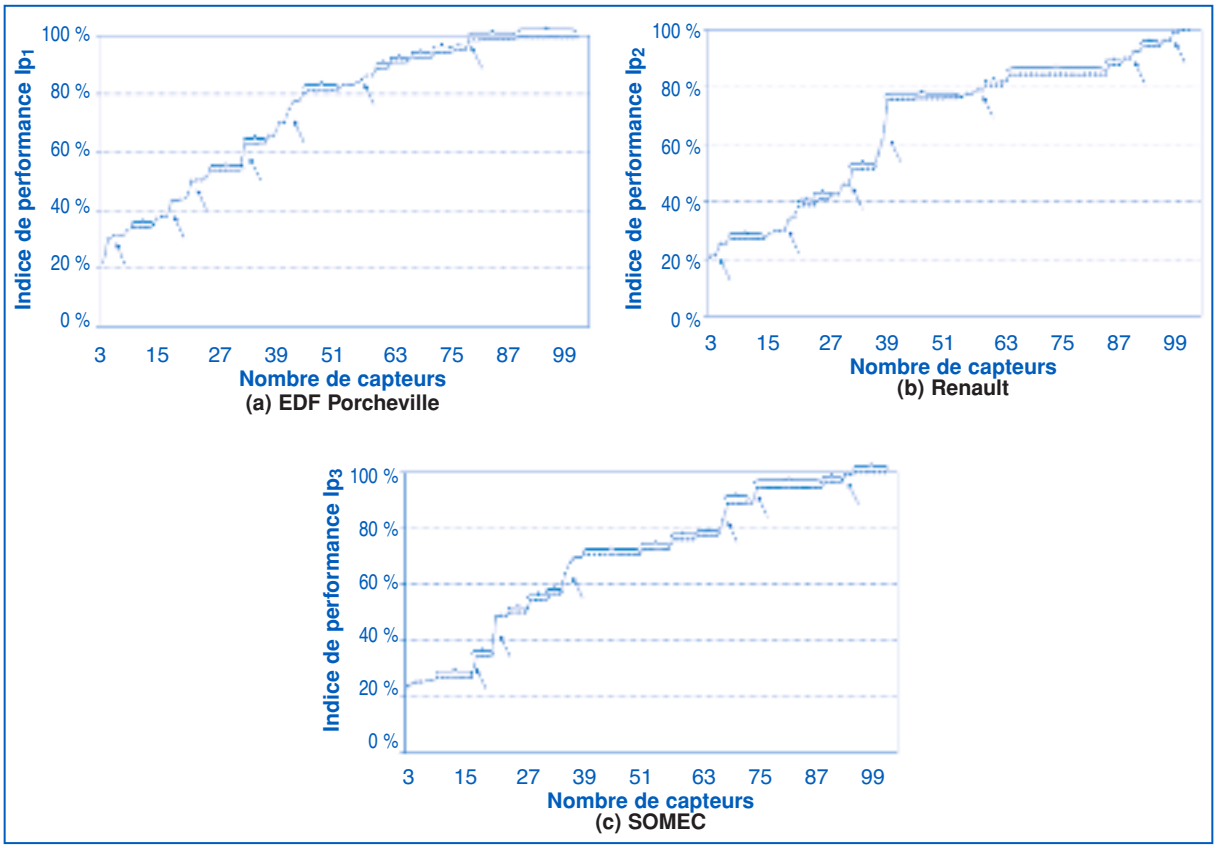

Figure 2.
Variation de l'indice de performance des trois sources sur la période d'étude en fonction du nombre de capteurs utilisés.
({: capteurs redondants, ↖ : capteur apportant une amélioration dans l'inversion) [19].
Performance index variation for the three sources versus the number of sensors used.
({ : redundant sensors, ↖ : sensor improving inversion performance) [19].

5.3 Méthodes de sélection des capteurs

L'analyse présentée dans la section 5.2 a mis en évidence deux catégories de capteurs, dont la présence : *(i)* améliore la performance de l'inversion, *(ii)* est redondante.

C'est en tenant compte de ces deux aspects, que dans la suite, on présente deux méthodes permettant de sélectionner un sous-ensemble optimal à partir du réseau initial.

5.3.1 Amélioration de l'indice de performance

À partir des résultats présentés dans la Figure 2, on peut choisir directement les capteurs apportant une amélioration importante de l'indice de performance de chaque source. Etant donné qu'il ne s'agit pas des mêmes capteurs pour les trois sources, nous avons d'abord choisi les dix premiers capteurs correspondant à chacune des trois sources, à partir de leur indice de performance. Les capteurs communs pour les trois sources sont gardés pour faire partie du réseau (trois capteurs). Parmi les capteurs restants (27), on choisit, dans l'ordre, ceux pour lesquels l'erreur moyenne d'estimation des trois sources est la plus faible (jusqu'à dix capteurs). La configuration du réseau obtenu est présentée dans la figure 3 (d) : « réseau global ».

5.3.2. Diminution de la redondance

La Figure 2 met en évidence des paliers, qui correspondent aux capteurs redondants, donc corrélés. On peut analyser cette redondance directement par des méthodes de classification et reconnaissance de formes.

À chacun des 100 capteurs fictifs, correspond une série temporelle des concentrations semi-synthétiques pour la période d'étude, qui sera une variable dans la méthode statistique multivariée appliquée ensuite : Classification Ascendante Hiérarchique (CAH) ou Analyse en Composantes Principales à Noyaux (ACPN).

La Classification Ascendante Hiérarchique (CAH) [20] permet le regroupement des capteurs en classes construites de manière à maximiser leur cohérence pour une certaine mesure de similarité. Dans ce cas, c'est l'algorithme de Ward, basé sur des distances euclidiennes, qui a été utilisé.

Pour exploiter les structures non-linéaires pouvant exister dans la matrice des mesures de concentrations, la décomposition de cette matrice peut se réaliser par projection des capteurs dans un espace de plus grande dimension (ACPN) [21]. Les composantes obtenues reflètent la structure relationnelle

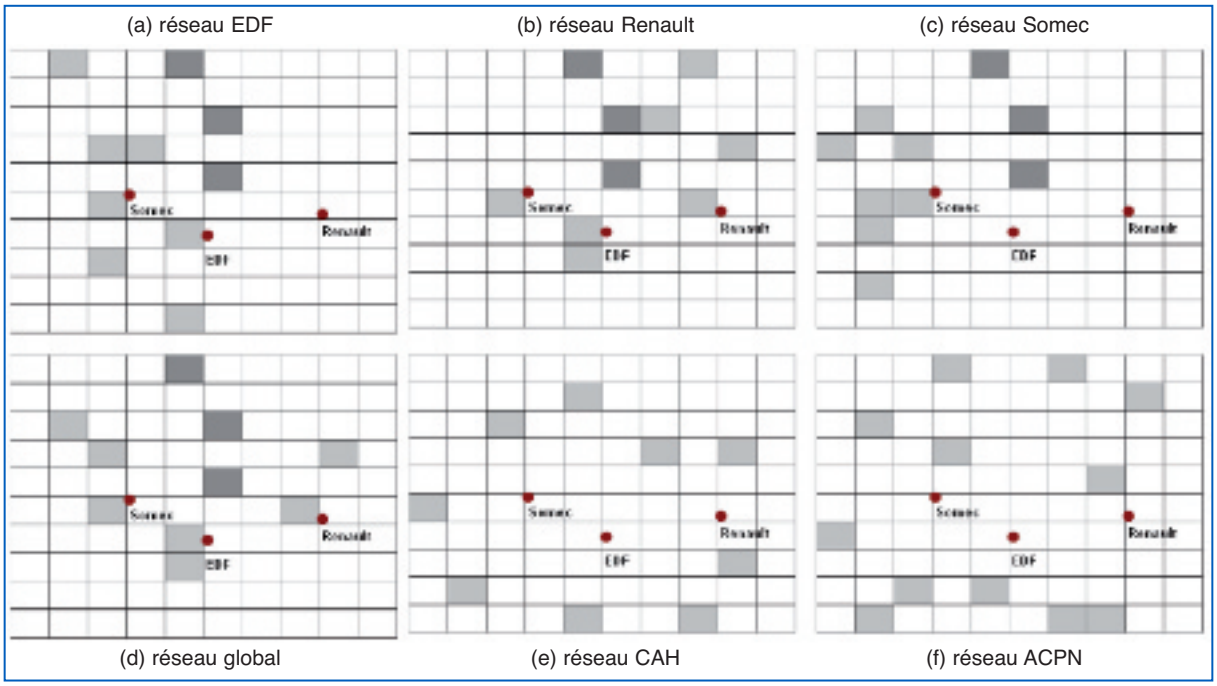

Figure 3.
Configurations de réseaux obtenues par la technique d'amélioration de l'indice de performance de réseau (a), (b), (c), (d) et celle de la diminution de la redondance (e), (f) [19].
Network configurations obtained by improving the network index of performance (a), (b), (c), (d) or by using the redundancy reduction (e), (f) [19].

pouvant exister entre les capteurs par rapport à un critère statistique spécifique (la corrélation non-linéaire).

Une fois les classes de capteurs obtenues, en partant d'un réseau minimal de trois capteurs, on rajoute, pour chaque classe un seul capteur (choisi aléatoirement), les autres capteurs de la même classe étant alors éliminés.

Les résultats obtenus en utilisant la méthode de classification sont présentés dans la Figure 3 (e) : « réseau CAH » et ceux, basés sur l'ACPN, dans la Figure 3 (f), « réseau ACPN ».

5.4 Résultats

5.4.1 Performances du réseau optimisé

Les réseaux obtenus en utilisant la technique ascendante basée sur l'amélioration de l'indice de performance (section 5.3.1) ou celle de la diminution de la redondance (section 5.3.2) sont rassemblés dans la Figure 3.

Les différentes configurations obtenues sont assez similaires, avec les capteurs placés à l'extérieur et autour des sources. Néanmoins, on note que les réseaux obtenus par les techniques de diminution de la redondance par CAH et ACPN donnent des configurations plus dispersées des capteurs.

On remarque également que la position des capteurs choisis en fonction de l'indice de performance propre à chaque source (Figures 3a, 3b, 3c) reflète la rose des vents sur la période d'étude choisie (Figure 4b), sur cette période le vent souffle majoritairement de SSE.

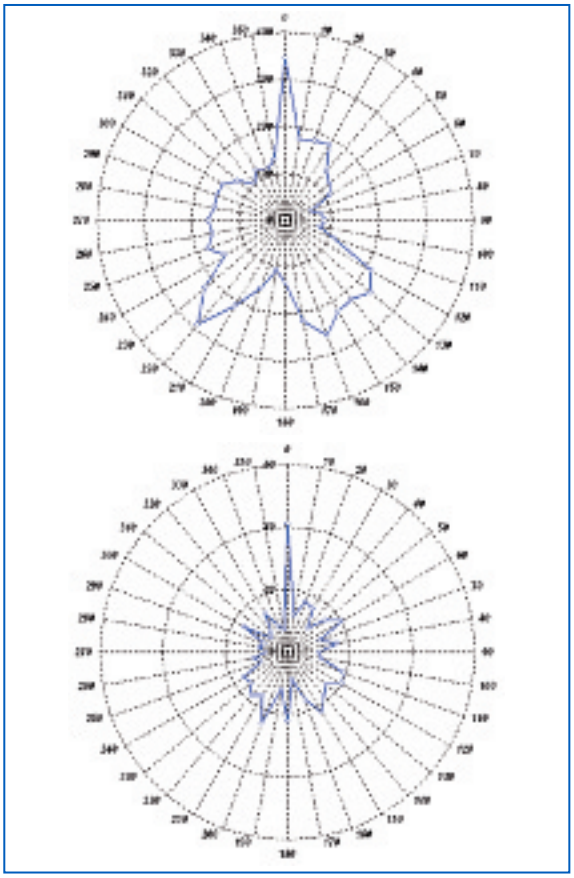

Figure 4.
Rose de vents : (a) sur toute la période (2 ans) ; (b) sur la période de détection maximale des capteurs (66 jours) [19].
Wind rose: (a) during all the period (2 years); corresponding to the period of maximum detection of the sensors [19].

Tableau 2.
Indice de performance de l'inversion pour les différentes sources pour différents réseaux de mesure [19].
Performance index for the source inversion using different networks [19].

Configuration	Méhode	Nombre de de capteurs	Indice de performance pour la source		
			EDF	Renault	Somec
Réseau EDF (Fig. 3a)	Amélioration I_p pour EDF	10	62 %	39 %	59 %
Réseau Renault (Fig. 3b)	Amélioration I_p pour Renault	10	58 %	72 %	50 %
Réseau Somec (Fig. 3c)	Amélioration I_p pour Somec	10	56 %	38 %	79 %
Réseau global (Fig. 3d)	Amélioration I_p pour toutes les sources	10	66 %	66 %	68 %
Réseau CAH (Fig. 3e)	Diminution redondance par CAH	9	72 %	73 %	62 %
Réseau ACPN (Fig. 3f)	Diminution redondance par ACPN	12	62 %	72 %	69 %

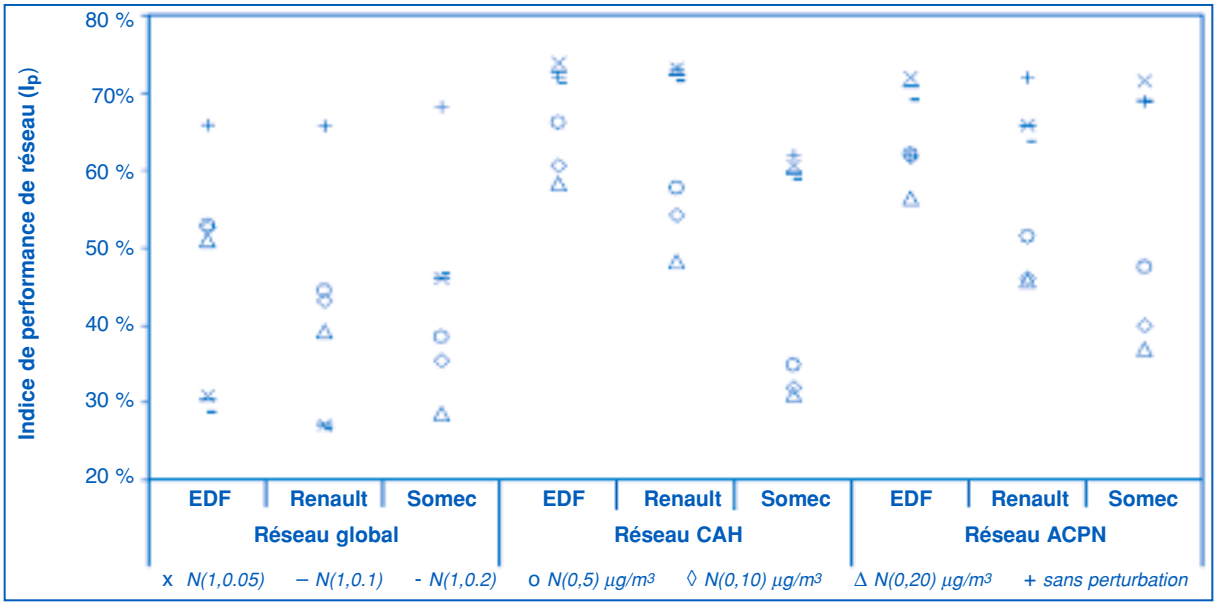

Figure 5.
Indice de performance des réseaux : Global, CAH et ACPN avec des données bruitées et parfaites [19].
Performance index for different networks using noisy and perfect data [19].

Pour chacune des sources prises indépendamment des autres, les capteurs qui ont un impact majeur sur l'estimation des débits sont ceux situés dans ces directions. Néanmoins, par comparaison avec la rose de vents de toute la période (deux ans) on note qu'il y a un vent fréquent de SSW. Un choix plus approprié de la période d'étude devrait tenir compte, en plus du critère de détection maximale, de la fréquence des vents dominants.

Les résultats obtenus par les différentes approches sont présentés dans le Tableau 2. Les différentes méthodes permettent de réduire le réseau initial de 100 capteurs à un réseau d'au maximum 12 capteurs, permettant l'identification des sources dans 70 % des cas environ (entre 40 % et 80 %).

Le Tableau 2 montre que, parmi les trois sources, c'est toujours la source EDF qui est la mieux reconstituée, pour les six configurations de réseaux, ceci peut être expliqué par la position « centrale » de la source, favorable à toutes les situations météorologiques, mais aussi par le fait qu'elle représente la source la plus importante (un facteur de 12 par rapport à l'émission moyenne de la source Renault et un facteur de 30 par rapport à l'émission moyenne de la source Somec).

Si on compare les performances globales de tous les réseaux, en regardant l'erreur d'estimation pour toutes les trois sources, on constate que ce sont les techniques de diminution de redondance par CAH et et ACPN qui conduisent aux réseaux les plus performants.

La conception et les tests effectués sur les réseaux précédents ont été réalisés en utilisant des données synthétiques « parfaites », on rappelle que les concentrations aux capteurs virtuels ont été simulées en utilisant le modèle gaussien. Dans la suite, on se propose de bruiter ces données et de réévaluer les performances des réseaux dans ces nouvelles conditions, plus proches des mesures réelles.

5.4.2 Résultats avec données bruitées

Les données synthétiques obtenues par le modèle gaussien ont été bruitées avec un bruit gaussien additif $C_i^{\text{bruité}} = C_i + \varepsilon_i$ ou multiplicatif : $C_i^{\text{bruité}} = C_i \times \zeta_i$, avec C_i : concentration prédite par le modèle gaussien sur le capteur i, ζ_i : bruit gaussien de moyenne unité et d'écart type σ = 0.05, 0.1, 0.2 et ε_i : bruit gaussien centré d'écart type σ = 5, 10, 20 $\mu g \cdot m^{-3}$.

Pour la comparaison des réseaux, on a défini un indice de performance (%) représentant le nombre de cas où l'estimation du débit réel des trois sources étudiées se fait avec une erreur inférieure à 20 %, calcul concernant toujours la période d'étude de 66 jours. Sur la figure 5, on a représenté cet indice de performance pour chacune des trois sources dans les cas correspondant aux réseaux : « global » (figure 3d), « CAH » (Figure 3e) et « ACPN » (Figure 3f); sur la même verticale, on retrouve pour une source, l'indice de performance correspondant aux données parfaites et aux données bruitées.

On constate que le réseau global est très sensible au bruit introduit, qui entraîne parfois une très forte détérioration des résultats. Par contre, les réseaux CAH et ACPN sont beaucoup moins sensibles au bruit, le réseau le moins sensible étant celui obtenu par la classification ascendante hiérarchique (CAH).

On note que la méthode CAH donne les meilleures estimations pour les deux sources EDF et Renault, alors que la méthode ACPN donne la meilleure estimation de la source la plus faible en émission, Somec.

Cette analyse nous pousse donc vers la recherche d'un indice global d'évaluation de réseaux de mesures (par exemple une moyenne pondérée en fonction des distances sources-capteurs, des erreurs moyennes entre débit réel et débit estimé, mais aussi vers la recherche de combinaison de différentes méthodes (ACPN + CAH par exemple) pour la sélection des capteurs les plus « prometteurs ».

6. Conclusions

Cette étude a démontré la possibilité d'estimer, à partir des mesures aux récepteurs, les émissions des sources chroniques et connues, en couplant un modèle Gaussien de dispersion avec une méthode de minimisation modèle-mesure (Gauss-Newton, algorithmes génétiques), sur un cas réel : une zone industrielle située en région francilienne, où sont implantées trois sources importantes de SO_2.

Dans un premier temps, le problème a été abordé en utilisant les mesures fournies par le réseau de surveillance de la qualité de l'air dans la région (trois capteurs), ainsi que les paramètres météorologiques mesurés à proximité. Dans certains cas de figure (récepteurs sous le vent des sources), l'ordre de grandeur des émissions a bien été restitué, parfois avec une erreur inférieure à 10 % pour au moins deux sources. Néanmoins, d'autres configurations (direction du vent-positions des capteurs) ont conduit à un problème mal posé, qui ne permet pas d'estimer les émissions des sources, ces derniers cas correspondant à des situations météorologiques fréquentes. Cette constatation a mis en évidence la nécessité de concevoir un réseau de capteurs, optimisé pour estimer les sources pour les situations atmosphériques les plus fréquentes.

La première étape de la conception du réseau a été la création des mesures semi-synthétiques aux capteurs fictifs répartis uniformément et en grand nombre sur toute la région autour des sources, à partir des données météorologiques réelles et des caractéristiques des sources, en utilisant le modèle gaussien de Pasquill. La méthodologie développée pour la conception du réseau a consisté dans la sélection des capteurs les plus « prometteurs » au sens de la précision d'estimation des émissions des sources (la méthode d'inversion étant celle testée lors du réseau de surveillance de la qualité de l'air).

Trois critères ont été considérés pour le choix des capteurs : (i) la fréquence de détection du capteur ; (ii) la performance du capteur vis-à-vis de l'estimation des sources et (iii) la redondance des capteurs. Le premier critère est assuré par le choix du capteur présentant la fréquence de détection des valeurs au-dessus du bruit de mesure la plus élevée durant la période d'étude. Le critère de performance du capteur concerne les capteurs dont l'ajout au réseau engendre une amélioration importante dans l'estimation des débits d'émissions de l'une des trois sources. Cette amélioration dépend en premier lieu de la source (position géographique et importance d'émission). Pour le critère de redondance des capteurs, le choix s'est effectué à base de groupement de capteurs en utilisant des méthodes statistiques de classification et reconnaissance de formes : classification ascendante hiérarchique (CAH) et analyse en composantes principales à noyaux (ACPN).

L'analyse de la qualité des réseaux choisis s'est basée en premier lieu sur leur capacité à estimer les sources. Cette analyse nous a permis de conclure qu'on pouvait obtenir des résultats satisfaisants pour plusieurs configurations testées. En effet, avec une dizaine de capteurs, on arrive à obtenir des estimations exactes dans environ 70 % des cas étudiés.

Les configurations les plus prometteuses ont été testées ensuite avec des données bruitées, en introduisant du bruit gaussien multiplicatif ou additif sur les mesures initiales de concentration issues du modèle gaussien. Les résultats se sont avérés assez sensibles au bruit introduit, les réseaux les plus robustes étant ceux obtenus en utilisant l'ACPN, mais surtout la CAH.

Ces premiers résultats obtenus pour la conception d'un réseau optimisé pour la surveillance des sources chroniques et connues sont plutôt encouragents et la méthodologie de travail développée peut être utilisée comme une bonne base de départ pour des études futures.

References

[1] Hopke P. The application of receptor modeling to air quality data. *Pollution Atmosphérique* 2010.

[2] Paatero P. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems* 1997 ; 37 : 23-35.

[3] Khlaifi A, Ionescu A, Candau Y. Pollution source identification using a coupled diffusion model with a genetic algorithm. *Mathematics and Computers in Simulation* 2009 ; 79 (12) : 3500-10.

[4] Bocquet M. Modélisation inverse des sources de pollution atmosphérique accidentelle : progrès récents. *Pollution Atmosphérique* 2010.

[5] Sauvage S, Plaisance H, Locoge N, Wroblewski A, Coddeville P, Galloo JC. Long term measurement and source apportionment of non-methane hydrocarbons in three French rural areas. *Atmospheric Environment* 2009 ; 43 : 2430-41.

[6] Veron A, Chruch T, Patterson C, Erel Y, Merril J. Continental origin and industrial sources of trace metals in the northwest Atlantic troposphere. *J. Atmos. Chem.* 2000 ; 14 : 339-51.

[7] Paris JD, Stohl A, Ciais P, Ramonet M, Nédélec P. Relations source-récepteur transcontinentales identifiées avec un modèle Lagrangien de dispersion et une analyse en clusters. *Pollution Atmosphérique* 2010.

[8] Pison I, Menut L, Blond N. Inverse modeling of emissions for local photo-oxidant pollution: Testing a new methodology with kriging constraints. *Annales Geophysics* 2006 ; 24 : 1523-35.

[9] Bousquet P. Transport atmosphérique et inversion des sources et puits de gaz à effet de serre. Habilitation à diriger des recherches. Université de Versailles Saint-Quentin-en-Yvelines 2006.

[10] Pasquill F. The estimation of the dispersion of windborne material. *The Meteorological Magazine Society* 1961 ; 70 (1) : 33-49.

[11] Pasquill F, Smith F. Study of the dispersion of windborne material from industrial and other sources. Ed. Ellis horwood, Chichester, England, 1983.

[12] Avila Galarza A. Diffusion des polluants atmosphériques dans une zone à topographie complexe. Validation d'un modèle à l'aide des mesures d'AIRPARIF. Thèse de doctorat, Université Paris XII - Val-de-Marne, 1996 : 172 p.

[13] Hanna S. Chang J. Boundary-layer parameterizations for applied dispersion modelling over urban areas, *Bound.-Layer Meteorol.* 1992 ; 58 : 229-59.

[14] Khlaifi A. Estimation des sources de pollution par modélisation inverse. Thèse de Doctorat, Université Paris XII-Val-de-Marne, 2007 : 353 p.

[15] Khlaifi A, Ionescu A. Impact de la vitesse et de la direction du vent sur l'estimation des principaux émetteurs de la vallée de la Seine en France, XXe colloque de l'Association Internationale de Climatologie. *Climat, Tourisme et Environnement* 2007 : 342-7.

[16] Coleman T, Li Y. On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds. *Mathematical Programming* 1994 ; 67 (2) : 189-224.

[17] Coleman T, Li Y. An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM Journal on Optimization* 1996 ; 6 : 418-45.

[18] Khlaifi A, Ionescu A, Candau Y. Source Identification Based on Coupled Gaussian Model with a Multi-Objective Genetic Algorithm, The Third International Conference on Environmental Science and Technology (IC EST 2007) 2007 Houston, Texas, the United States of America, Book of Abstracts p. 31.

[19] Khlaifi A, Ionescu A, Candau Y. Conception optimale d'un réseau de mesure permettant de caractériser les principaux émetteurs – application dans une zone industrielle. *Revue électronique Sciences et Technologies de l'Automatique* 2008 ; 2 : 7-12.

[20] Saporta G. Théories et méthodes de la statistique, Technip, Paris, 1995.

[21] Boser B, Guyon I, Vapnik V. A Training Algorithm for Optimal Margin Classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory ; 1995(5) : 144-52.