C. R. Geoscience 334 (2002) 957–965

S1631-0713(02)01836-9/FLA

Géosciences de surface / Surface Geosciences

Contribution des réseaux de neurones artificiels (RNA) à la caractérisation des pollutions de sol. Exemples des pollutions en hydrocarbures aromatiques polycycliques (HAP)

Adrian Dan*, Jasha Oosterbaan, Philippe Jamet

Centre d'informatique géologique, École nationale supérieure des mines de Paris, 35, rue Saint-Honoré, 77305 Fontainebleau cedex, France

Reçu le 25 août 2002 ; accepté le 3 septembre 2002

Présenté par Ghislain de Marsily

Abstract – Artificial Neural Networks (ANNs) characterisation of soil pollution: the Polycyclic Aromatic Hydrocarbons (PAHs) case study. We develop the ANNs (Artificial Neural Networks) method to explore contaminant concentration profiles observed in soils of polluted sites. ANNs are particularly efficient in simultaneous analysis of numerous parameters and in identification of complex relations involving field data. Applying the ANN models on a PAH (Polycyclic Aromatic Hydrocarbon) database, we extracted the most characteristic components of known contaminations and applied it to identify the source type of similar polluted sites. The performed tests prove the generalisation capability of the selected ANN model. ***To cite this article: A. Dan et al., C. R. Geoscience 334 (2002) 957–965.***

Artificial Neural Networks (ANN) / Polycyclic Aromatic Hydrocarbons (PAH) / soil contamination / coking plant / manufactured gas plant

Résumé – L'analyse exploratoire de données environnementales utilise fréquemment des outils statistiques. Cet article présente l'applicabilité des réseaux de neurones artificiels (RNA) à l'étude des pollutions des sols par les hydrocarbures aromatiques polycycliques (HAP). Les RNA mettent en évidence des relations entre la distribution en polluants observée dans des prélèvements de sol et le type d'activité industrielle génératrice. Ces modèles facilitent, par exemple, l'identification des substances chimiques caractéristiques des pollutions observées sur les sites d'anciennes cokeries et usines à gaz. ***Pour citer cet article : A. Dan et al., C. R. Geoscience 334 (2002) 957–965.***

réseaux de neurones artificiels (RNA) / hydrocarbures aromatiques polycycliques (HAP) / pollution de sol / cokerie / usines à gaz

Abridged version

1. Introduction

Identifying the contaminant source characteristics is fundamental to understanding pollution features. Here, we extract workable relations between the source type and the observed contaminant concentration profiles from audit studies concerning a large database of sites contaminated by PAHs (Polycyclic Aromatic Hydrocarbons).

We develop the ANN (Artificial Neural Networks) method to explore contaminant concentration profiles (i.e., the PAHs content of a soil sample) observed at polluted sites. We make use of the hypothesis that a field data rep-

* Correspondance et tirés à part.
Adresses e-mail : adrian@cig.ensmp.fr (A. Dan), jasha@cig.ensmp.fr (J. Oosterbaan), jamet@isige.ensmp.fr (P. Jamet).

resent the scalar image of a complex reality integrating information about (I) source history, (II) physical and chemical properties of pollutants, (III) transfer processes in the geological environment, and (IV) external environmental conditions.

All these factors, concealed in isolated data, may become accessible in a statistical form by assembling them in a database containing a sufficient number of representative data.

In this sense, we exploit records from many sites involving the same pollutants in variable environmental conditions.

ANN are black-box type mathematical models, formed by a set of computing elements (each of which has a transfer function) arranged in a specific structure (like the one presented in Fig. 1), where the network parameters are represented by the values (weights) associated with the computing element connections. The expression of the ANN output is shown by Eq. (2). Here, we use ANN for a discrimination problem. The objective is to associate input values (i.e., concentrations of a pollutant family) to one among many possible output categories (i.e., the industrial activities susceptible of producing the observed contamination).

Typically, these output categories are coded with values between 0 and 1. For a two-category case, the network output values may be interpreted as the probability of an input dataset to belong to one of the output categories.

The structure and number of the intermediate (hidden) layers and the transfer functions are selected with respect to the problem complexity. We use one hidden layer of two, four and respectively ten computing elements with sigmoid transfer functions – i.e., logistic and hyperbolic tangent; Eq. (1). In all cases, the output element has a logistic transfer function (Fig. 2). For the ANN learning phase, we use the back-propagation with momentum algorithm.

The database is divided into three parts: a training and a test sets of 108 examples each (54 examples for each output category example, i.e., concentrations profiles observed on coking and manufactured gas plant sites), used for the network parameters determination, and an additional verification set, containing 695 performance evaluation examples, discarded in the training phase.

The input data (x_i, $i = 1$ to 10) are unit-normalised logarithm transforms of the PAH concentration measured in a soil sample (cf. Section 3.2).

Table 1 lists the selected ANN models arranged by test performance criterion. We observe that the ANN with ten and four hidden computing elements produce similar performances in processing the test set examples (i.e., around 80% of the examples correctly discriminated). Consequently, we select the four hidden neurons ANN as the resulting compromise in model performance–complexity trade-off.

An important application of this study is the number reduction of PAH used in pollution type classification, made by examining the weights associated with particular constituent (Fig. 3). We noticed that the highest weights are associated with the most discriminating PAH connections (i.e., ANT, FLN, BaP). These observations are validated by a successive presentation of all training set examples, perturbing successively the input values of a specific PAH and examining the influence on the output value.

As a result, we state that, when the PAH concentrations are known, the BgP and IP contents give minor additional discriminating information.

We tested the selected RNA with eight soil samples external to the initial database, four samples from a manufactured gas plant site and four samples from a coking plant site. The results (Fig. 4) show that the RNA model correctly identified the source type for all samples of manufactured gas plant site and that one sample of coking plant site is improperly classified.

In Fig. 5, we illustrate the example profiles used in the training phase. We observed a high similarity between the profiles produced by the same type of source activity – i.e., a ‘square root symbol’ profile for manufactured gas plant sites (Fig. 5a), and a ‘roof shape’ profile for the coking plant sites (Fig. 5b) – with the NAP and ANT contents, showing the most obvious visual difference between the two profile types.

However, only by this simple visual criterion, it is not possible to highlight the other discriminating components indicated by ANN analysis (i.e., FLN and BaP). By this mean, we confirm the advantage of ANN data analysis to determine subtle but nevertheless important discriminating factors.

1. Présentation

La mesure des teneurs en substances polluantes et la localisation des zones polluées constituent la phase initiale, souvent déterminante, de l’audit environnemental d’un site contaminé.

L’exploitation des données de terrain à l’aide de méthodes exploratoires de données permet de mettre en évidence certaines caractéristiques typiques de la pollution en cause. Ces caractéristiques, véritables « signatures » de la pollution, peuvent être utilement exploitées pour des cas de pollution similaire. Les méthodes d’analyse exploratoire sont applicables à la résolution des problèmes caractérisés par une faible connaissance a priori des relations entre les données. Les réseaux de neurones artificiels (RNA) sont des outils de calcul puissants permettant ce type d’analyse de données dans le domaine des sciences de la Terre [3, 5].

Du fait de leur ubiquité, de leur persistance dans l'environnement et de leur toxicité, les pollutions en HAP ont fait l'objet de nombreux travaux scientifiques [1, 6, 8, 12]. Nous présentons ici les résultats principaux de l'investigation d'une base de données issue d'études d'audits réalisées sur plusieurs sites contaminés en HAP [10]. Cet article illustre l'intérêt des RNA dans l'exploitation des informations relatives aux distributions de teneurs en substances polluantes sur des sites pollués dont l'activité industrielle génératrice est connue, afin d'optimiser la caractérisation des contaminations de même type.

2. Problématique générale

Les sites pollués représentent un problème environnemental complexe, tant en termes scientifiques qu'en termes de décision publique. Une caractérisation approfondie et fiable de la pollution est indispensable afin d'estimer les risques associés à ces sites, dans leurs dimensions spatiale et temporelle, puis de prendre des mesures techniques appropriées.

Malheureusement, les données de terrain (sols, eaux) permettant d'asseoir une stratégie efficace sont souvent en nombre trop limité (étant donné le coût des analyses) pour offrir tout le recul nécessaire sur un site donné. C'est la raison pour laquelle il est souhaitable de tirer profit des données d'audits disponibles sur des pollutions similaires.

Comment envisager l'emploi de ces données externes au site étudié ? Nous partons du constat qu'une donnée de terrain (une concentration en polluant observée dans un prélèvement de sol, par exemple) est la traduction scalaire d'une réalité complexe intégrant (*i*) l'historique du terme-source, (*ii*) les propriétés intrinsèques des polluants, (*iii*) la fonction de transfert dans le milieu géologique et (*iv*) les conditions environnementales externes. Ces divers facteurs, masqués dans une donnée isolée, sont susceptibles de se manifester sous une forme statistique accessible, en assemblant dans une base une quantité suffisante de données représentatives de la réalité. Pour ce faire, on utilise des données issues de sites impliquant les mêmes polluants, soumis à l'influence de conditions environnementales variables. Les méthodes d'analyse exploratoire ont donc pour objectif principal « d'épuiser » le contenu informatif de la donnée.

Les réseaux de neurones artificiels (RNA) sont un outil de calcul permettant de développer cette démarche. Ils utilisent, pour l'analyse des informations, un ensemble de fonctions mathématiques, des paramètres organisés et une méthodologie d'apprentissage calquée, selon les concepteurs, sur le fonctionnement du cerveau.

3. Les réseaux de neurones artificiels

3.1. Principe général

Les RNA sont des modèles mathématiques non linéaires, de type « boîte noire », capables de déterminer des relations entre données par la présentation (l'analyse) répétée d'exemples (à savoir des couples constitués par une information d'entrée et une valeur de sortie que l'on voudrait approcher par le modèle). La modélisation à l'aide de RNA (appelée « phase d'apprentissage ») suppose l'adaptation des paramètres du réseau, afin de mettre en évidence les relations qui portent sur les exemples présentés.

Les RNA sont constitués d'un ensemble d'éléments de calcul (neurones artificiels), organisés dans une structure spécifique (par exemple, celle présentée sur la Fig. 1), les paramètres du réseau (les poids) étant représentés par les valeurs associées aux connections de ces éléments de calcul. Un élément de calcul du RNA comporte une ou plusieurs entrées et une sortie. La valeur de sortie est obtenue par l'application d'une relation mathématique (fonction d'activation) sur la somme pondérée d'entrées.

Dans la modélisation à l'aide de réseaux de neurones artificiels, on peut choisir le type de fonctions d'activation, le nombre de neurones et l'arrangement de leurs connexions (à savoir la structure du réseau). Généralement, on utilise des fonctions d'activation de type « sigmoïde » (équation (1)) :

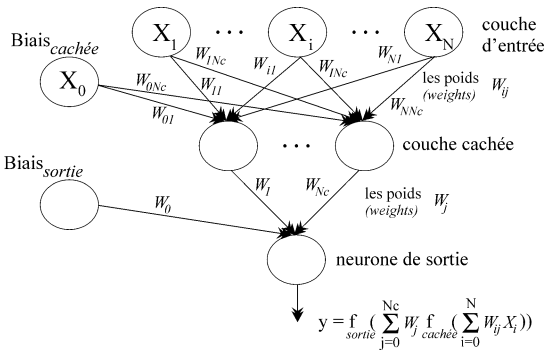


Figure 1. Structure du perceptron multicouche. Celui-ci comporte la couche d'entrée, une ou plusieurs couches intermédiaires (cachées) et la couche de sortie. Chaque couche contient des unités de calcul – neurones – connectés à d'autres neurones par la voie des poids. Les flèches (les connexions des éléments de calcul) indiquent le sens de propagation des données.

Figure 1. Structure of the multilayer perceptron. It consists of an input layer, one or more intermediary (hidden) layers, and one output layer. Each layer assembles processing units – neurones – connected to other neurones. The strength of the connection is measured by a weight. The arrows (connections between neurones) show the direction of information propagation.

logistique (S) :

$$f(x) = 1/\big(1 - \exp(-x)\big)$$

et (1)

tangente hyperbolique (T) :

$$f(x) = \big(\exp(2x) - 1\big)/\big(\exp(2x) + 1\big)$$

qui peuvent adapter le degré de non-linéarité du modèle en fonction de la complexité du problème.

La structure de réseau la plus employée – dite du « perceptron multicouche » (Fig. 1) – consiste en un arrangement en plusieurs niveaux de neurones ou couches, dont l'information se propage unidirectionnellement de la couche d'entrée vers la couche de sortie. Dans ce schéma, à la différence des neurones artificiels de la couche cachée et de la couche de sortie, la fonction d'activation des éléments de la couche d'entrée est la fonction mathématique identité, dont la sortie a la même valeur que l'entrée. Un neurone particulier (nommé « biais ») est connecté à chaque couche de neurones (sauf celle d'entrée). Dans le calcul de la valeur de sortie du RNA (équation (2)), le « biais » joue le même rôle que le « terme libre » dans une équation de régression.

L'équation correspondante pour la sortie du modèle est (équation (2)) :

$$y = f_{\text{sortie}}\left(\sum_{j=0}^{n_{\text{cachée}}} w_j\, f_{\text{cachée}}\left(\sum_{i=0}^{10} w_{ij}\, x_i\right)\right) \qquad (2)$$

où f_{sortie} et $f_{\text{cachée}}$ sont les fonctions d'activation pour la couche de sortie et pour la couche cachée, les facteurs w_j dessinant les poids des connexions entre les neurones de la couche cachée (en nombre de $n_{\text{cachée}} = 2$, 4 ou 10) et la sortie, les w_{ij} dessinant les poids des connexions entre les neurones de la couche d'entrée (indice i) et ceux de la couche cachée (indice j), tandis que les x_i symbolisent les entrées ($x_0 =$ le « biais »).

La propriété principale des RNA est leur capacité de généralisation, c'est-à-dire que ces modèles sont susceptibles de produire une réponse correcte lorsqu'ils sont appliqués sur des exemples différents de ceux utilisés dans la phase d'apprentissage.

Dans cette étude, on utilise des RNA pour un problème de discrimination. L'objectif est d'associer les valeurs d'entrée (concentrations d'une famille de polluants) à une catégorie parmi plusieurs possibles (les activités industrielles susceptibles de causer la pollution). Ces catégories sont souvent représentées par des valeurs de sortie entre 0 et 1. Dans un cas à deux catégories, ces deux valeurs peuvent être interprétées comme une probabilité d'appartenance d'une entrée à une catégorie.

3.2. Prétraitement des données et structuration du réseau

Afin d'être modélisées, les données – ici un ensemble de teneurs (exprimés en $\text{mg}\,\text{kg}^{-1}$) de $n = 10$ HAP observés dans le même échantillon du sol – passent par une phase de prétraitement, destinée à les rapporter à l'intervalle d'utilisation des réseaux de neurones $(-1, 1)$. Pour ce faire, on considère que la distribution des teneurs représente la réalisation d'un vecteur à n composantes. Chaque vecteur d'entrée est transformé en un vecteur à norme unitaire, en appliquant à chaque composant v_i la relation simple :

$$v_i^{\text{norm}} = v_i/\|v\| \quad \text{où } \|v\| = \sqrt{\sum_i v_i^2}, \quad i = 1, \ldots, n.$$

Ce calcul est réalisé séparément en chaque point où on mesure les teneurs en HAP.

Le nombre de neurones de la couche d'entrée et de la couche de sortie étant imposé (respectivement par le nombre de teneurs et le nombre de catégories), il reste à choisir le nombre de neurones de la (ou des) couche(s) cachée(s). Ce choix résultera d'un compromis entre la performance discriminatoire du réseau (croissante avec le nombre de neurones de cette couche intermédiaire) et le souci de limiter le caractère artificiel du réseau. On préférera à ce titre un schéma basé sur un nombre de neurones intermédiaires significativement inférieur au nombre de neurones d'entrée.

3.3. Phases d'apprentissage et de test du réseau

Le calage des paramètres du modèle (essentiellement le poids des liaisons entre les différents neurones) est réalisé d'après un algorithme de calcul qui utilise la présentation répétée d'un ensemble de plusieurs couples entrée–sortie connus (exemples qui constituent l'ensemble d'apprentissage). L'objectif de ce calcul et la minimisation d'une fonction d'erreur entre la réponse désirée et la réponse obtenue à la sortie du modèle. Par exemple, l'algorithme de « rétropropagation » estime le gradient de la fonction d'erreur par rapport aux poids du modèle et réalise l'adaptation de ces paramètres successivement de la couche de sortie vers la couche d'entrée.

La validation du modèle se réalisera ensuite sur des exemples (l'ensemble de test) non utilisés dans le calcul des poids. La performance du réseau est déterminée en fonction du nombre de succès et d'échecs dans la discrimination. Les paramètres d'ajustement du réseau sont le nombre de neurones cachés et les fonctions d'activation ; ce travail d'apprentissage et de test est donc opéré sur un nombre important de configu-

rations possibles, lesquelles sont classées en fonction de leur performance.

4. Démonstration sur le cas de pollutions du sol par des HAP

La démarche d'investigation des données par les réseaux de neurones artificiels sera illustrée sur le cas de pollutions industrielles résultant de deux activités anciennes ayant mis en œuvre des procédés de pyrolyse de la houille, à savoir les cokeries et les usines à gaz. Le choix de ces polluants a été motivé par la relative facilité à assembler la base de données de travail.

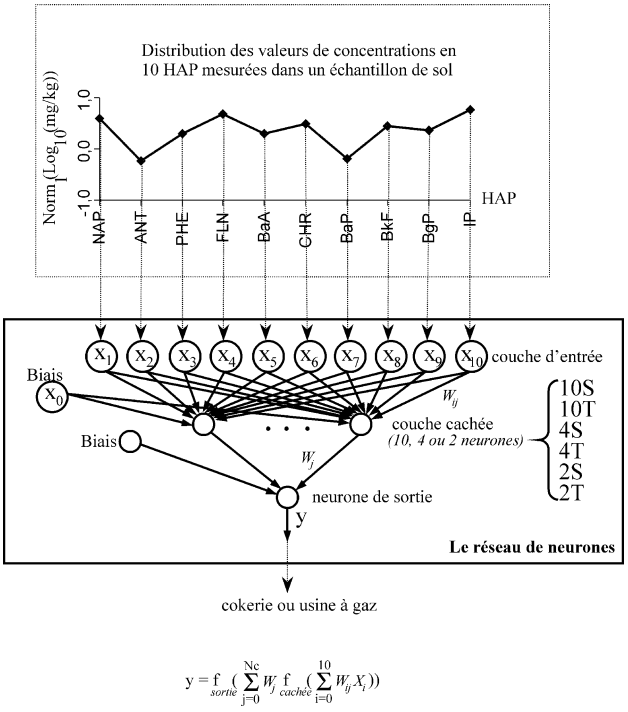
Dans ce cas précis, l'objectif de discrimination entre ces deux activités n'est pas un enjeu majeur, car bien souvent l'activité réalisée sur le site est connue à l'avance. Nous la présentons néanmoins pour deux raisons. Elle permet, en premier lieu, d'illustrer la performance de discrimination par les RNA sur la situation défavorable où les pollutions induites par les deux activités (cokéfaction et gazéification) sont très voisines. En second lieu, un sous-produit de l'analyse par réseaux de neurones, issu de l'examen des poids respectifs affectés aux teneurs d'entrée, est la réduction du nombre de variables caractéristiques de la pollution, c'est-à-dire du nombre de polluants représentatifs qu'il est nécessaire de doser pour obtenir, d'une façon optimale, l'information souhaitée (le potentiel de discrimination, par exemple) sur la pollution en cause.

4.1. La base de données

L'étude repose sur les données de 37 dossiers d'audit relatifs à des sites pollués, à savoir 32 usines à gaz et cinq cokeries. Ces dossiers ont été récoltés en France, en Allemagne et aux Pays-Bas. La base de données ainsi établie contient des données de teneurs en dix HAP dans 911 échantillons de sol (803 pour les usines à gaz et 108 pour les cokeries). Les teneurs, exprimées en $\mathrm{mg\,kg^{-1}}$ de sol, sont obtenues d'après un protocole de dosage appliqué à l'analyse de HAP, à savoir l'extraction Soxhlet, suivie de chromatographie GC ou HPLC [4].

Les dix HAP dosés sont les suivants : naphtalène (NAP), anthracène (ANT), phénanthrène (PHE), fluoranthène (FLN), benz[a]anthracène (BaA), chrysène (CHR), benzo[a]pyrène (BaP), benzo[k]fluoranthène (BkF), benzo[ghi]pérylène (BgP), indeno[1,2,3-cd]pyrène (IP).

Notre objectif est de parvenir à une relation entre la distribution de teneurs des dix HAP mesurées en un point et l'activité génératrice de la pollution, caractérisée par la valeur 0 pour les cokeries et 1 pour les usines à gaz.



$$y = f_{sortie}\left(\sum_{j=0}^{Nc} W_j\, f_{cachée}\left(\sum_{i=0}^{10} W_{ij} X_i\right)\right)$$

Figure 2. La représentation du cas étudié. Le modèle RNA cherche une relation entre la distribution en HAP observée dans un prélèvement de sol et l'activité industrielle génératrice.

Figure 2. The case study representation. The ANN model searches for a relation between the PAH distribution observed in a soil sample and the industrial activity that has generated it.

4.2. Les réseaux utilisés : structure, apprentissage et test

Pour la modélisation, nous avons utilisé des réseaux de neurones ayant dix neurones (le nombre de HAP considérés) dans la couche d'entrée et un neurone dans la couche de sortie, dont la valeur désigne la classe d'appartenance : cokerie ou usines à gaz (Fig. 2).

La structure et le nombre de couches cachées (intermédiaires) et leurs fonctions d'activation sont choisies en fonction de la complexité du problème à résoudre. Nous utiliserons une seule couche cachée, contenant deux, quatre ou dix neurones, avec des fonctions d'activation sigmoïde, soit logistique (S), soit tangente hyperbolique (T). Dans tous les cas, le neurone de sortie a une fonction d'activation logistique.

Pour l'apprentissage des ces réseaux, nous avons utilisé l'algorithme de rétropropagation. Dans cette application de RNA, les ensembles d'apprentissage et de test possèdent chacun 108 exemples (54 pour les cokeries et 54 pour les usines à gaz). Pour une vérification ultérieure, on constitue l'ensemble de vérification, contenant 695 exemples non utilisés pour l'apprentissage et pour le test. Les données d'entrée x_i sont les logarithmes des teneurs en chaque HAP, normalisées suivant la procédure exposée au Section 3.2.

Tableau 1. Réseaux les plus performants à dix, quatre et deux neurones dans la couche cachée.

Table 1. The best neural networks with ten, four or two neurones in the hidden layer.

	L'ensemble				
RNA	App. coke	App. u. gaz	Test coke	Test u. gaz	Vérif.
10S	100%	100%	87%	74%	79%
4T	98%	98%	83%	74%	81%
2S	91%	83%	89%	57%	83%

App. : Ensemble d'apprentissage. **Test** : Ensemble de test. **Vérif.** : Ensemble de vérification. **u. gaz** : Profils observés sur sites d'usines à gaz. **Coke** : profils observés sur sites de cokeries.

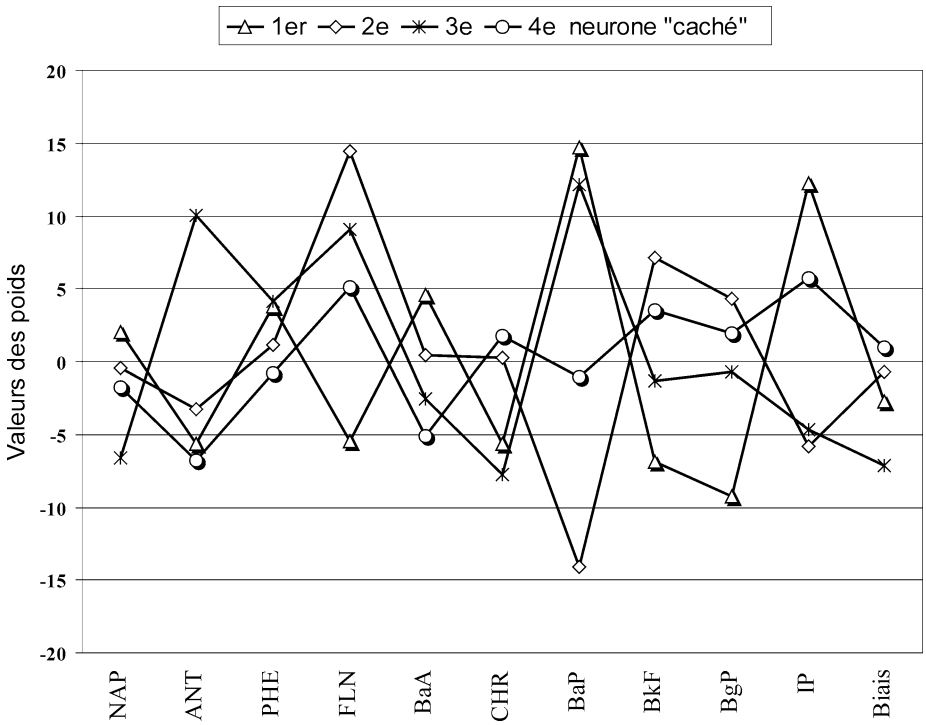


Figure 3. Le réseau 4T : poids des liaisons.

Figure 3. The weights of 4T ANN.

4.3. Résultats et discussion

4.3.1. Sélection du réseau le plus performant

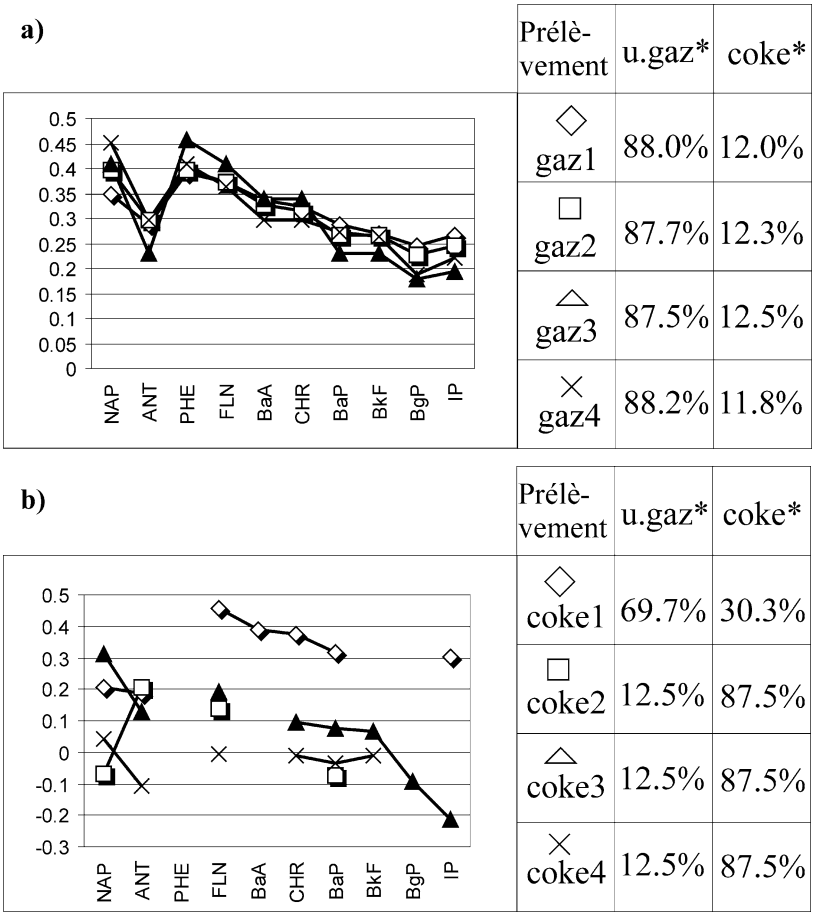
Le classement des réseaux testés est opéré par la propriété de généralisation, mesurée par la performance de la discrimination d'exemples de test. Les meilleurs modèles sélectionnés d'après ce critère sont présentés dans le Tableau 1.

On observe que les réseaux à dix et à quatre neurones dans la couche cachée présentent des performances similaires sur l'ensemble de test (environ 80% d'exemples correctement discriminés). Une performance, à généraliser, légèrement supérieure pour le réseau à quatre neurones est perceptible sur l'ensemble de vérification (81% contre 79% pour le réseau à dix neurones). En revanche, le réseau avec deux neurones dans la couche cachée s'avère un modèle trop simpliste pour pouvoir discriminer avec fiabilité les exemples cokeries dans l'ensemble de test (57%).

4.3.2. Les HAP caractéristiques

Une application importante de cette analyse est la réduction du nombre de HAP utilisés pour la discrimination. Elle s'opère à partir de l'examen des poids respectifs des constituants dans le réseau. La Fig. 3 présente les poids déterminés pour les liaisons entrées–couche cachée et couche cachée–sortie pour le réseau avec quatre neurones dans la couche cachée (4T). Les valeurs absolues maximales des poids de leurs connexions caractérisent les HAP les plus utiles à la discrimination. Il s'agit de ANT, FLN, BaP, BgP et IP.

Le simple critère des poids d'entrée n'est pas suffisant, car chacun des neurones de la couche cachée possède un poids propre sur le neurone de sortie. Une étude de sensibilité est nécessaire pour évaluer la conjugaison des deux niveaux de poids. On réalise, pour ce faire, la présentation successive de tous les exemples de l'ensemble d'apprentissage, en faisant varier à chaque fois les valeurs de teneurs en un des



Prélèvement	u.gaz*	coke*
◇ gaz1	88.0%	12.0%
□ gaz2	87.7%	12.3%
△ gaz3	87.5%	12.5%
× gaz4	88.2%	11.8%

Prélèvement	u.gaz*	coke*
◇ coke1	69.7%	30.3%
□ coke2	12.5%	87.5%
△ coke3	12.5%	87.5%
× coke4	12.5%	87.5%

Figure 4. L'identification de l'activité génératrice d'une pollution en HAP avec le RNA 4T. Prélèvements de test effectués (**a**) sur un site d'usine à gaz [7] et (**b**) sur un site de cokerie ; les valeurs des logarithmes des concentrations ont été normalisées suivant la procédure exposée au Section 2, ∗ : Probabilité, estimée par le RNA 4T, que la source polluante du sol analysé soit du type « usine à gaz » (première colonne) et probabilité qu'elle soit de type « cokerie » (deuxième colonne).

Figure 4. Identification of the pollution source type with the 4T ANN. Soil test sample (**a**) of a manufactured gas plant [7] and (**b**) of coking plant sites; the profiles are unit-normalised logarithm transform of the PAH concentration measure of a soil sample (see Section 2).

HAP et en observant l'effet produit sur la valeur de sortie. Par ce moyen, on confirme l'importance pour la discrimination des informations sur les teneurs en ANT, FLN et BaP. On constate que, les teneurs en ces HAP étant connues, la connaissance des teneurs en BgP et IP présente une faible valeur ajoutée.

Le paragraphe 4.4 montrera comment les résultats de la discrimination à l'aide des RNA sont confirmés par les observations qualitatives des distributions de teneurs en HAP associées à un type de source et par les aspects phénoménologiques du problème.

Afin d'estimer la performance du réseau 4T dans l'analyse de données en dehors de la base initiale, nous exploiterons un ensemble de quatre prélèvements de sol effectués sur un site d'usine à gaz et quatre prélèvements de sol effectués sur un site de cokerie. Les analyses des prélèvements réalisés sur le site d'usine à gaz ont révélé la présence des dix HAP utilisés dans notre étude, tandis que, dans les échantillons de sol prélevés sur le site de cokerie, certains HAP ne sont pas détectés.

Les résultats du test (Fig. 4) montrent que le réseau 4T identifie correctement le type de source pour tous les échantillons prélevés sur le site d'usine à gaz et qu'un seul échantillon prélevé sur le site de cokerie est incorrectement classé.

4.3.3. Les distributions caractéristiques

Le prétraitement des données (passage au logarithme, normalisation) permet de comparer visuellement les distributions de teneurs des dix HAP, pour chaque type d'activité industrielle génératrice.

Les distributions utilisées dans l'apprentissage sont présentées sur la Fig. 5, dans laquelle les ordonnées indiquent les logarithmes des concentrations normalisées associées aux HAP, présentés en abscisse selon l'ordre croissant de leur poids moléculaire.

On observe la ressemblance entre les distributions générées par le même type d'activité, à savoir la forme du symbole mathématique « racine carrée » pour les usines à gaz (la valeur associée à l'anthracène (ANT) est inférieure aux autres valeurs de HAP), présenté en Fig. 5a, et une forme en « toit » pour les cokeries (la valeur associée au naphtalène (NAP), inférieure aux autres valeurs de HAP, Fig. 5b). Ainsi, la différence la plus significative entre la forme de distributions de « cokeries » et d'« usines à gaz » est marquée par les teneurs en NAP et ANT, ce qui indique leur importance pour la discrimination. Toutefois, il n'est pas possible de discerner par ce simple critère visuel les autres composants discriminants mis en évidence par l'analyse basée sur les réseaux neuronaux (FLN, BaP). Ceci confirme l'importance de ce type d'ana-

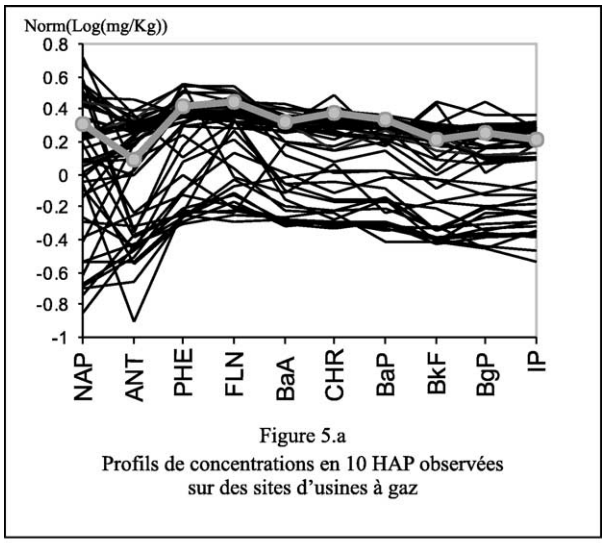


Figure 5.a
Profils de concentrations en 10 HAP observées sur des sites d'usines à gaz

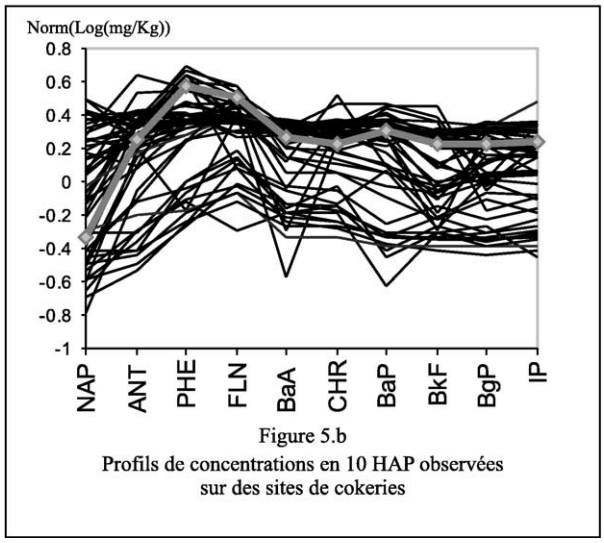


Figure 5.b
Profils de concentrations en 10 HAP observées sur des sites de cokeries

Figure 5. Morphologie des distributions en dix HAP de l'ensemble d'apprentissage « usines à gaz » (**a**) et « cokeries » (**b**).

Figure 5. PAH distribution morphologies: 'gas plants' learning set (**a**) and 'coking plants' learning set (**b**).

lyse, qui permet de déterminer des facteurs discriminants subtils, mais néanmoins significatifs.

4.4. Les caractéristiques de la source et les propriétés des polluants

Les facteurs susceptibles de différencier les contaminations engendrées par les deux types d'activité sont principalement la composition du produit polluant initial et l'évolution de la contamination dans le milieu environnant. Les températures et cinétiques de pyrolyse différentes utilisées en cokerie et en usine à gaz [11] aboutissent en effet à la formation de sous-produits de composition chimique variable.

Des procédés très voisins étaient utilisés pour les deux types d'activité discutés ici ; toutefois, leurs finalités industrielles étaient renversées (le gaz est un produit pour les usines à gaz et un sous-produit pour les cokeries, le coke est un produit pour les cokeries et un sous-produit pour les usines à gaz). Ainsi, dans la plupart des usines à gaz, un lavage des gaz à l'anthracène était effectué pour en extraire une partie du naphtalène. Dans les cokeries, le traitement des gaz était moins fréquent [8, 9]. Ces aspects expliquent certaines différences dans les proportions en HAP (notamment en NAP et ANT) observées dans les sous-produits générés par ces activités industrielles.

La structure moléculaire des HAP joue, en outre, un rôle essentiel dans leur stabilité [2]. Un agencement angulaire ou compact des cycles benzéniques (cas du FLN, du BaP, du BaA et du CHR) confère une plus grande rémanence (notamment une moindre mobilité) des composés dans l'environnement. Ainsi, ces produits chimiques sont susceptibles de conserver plus durablement les caractéristiques discriminantes des pollutions étudiées.

5. Conclusion

Nous avons illustré l'applicabilité des réseaux de neurones artificiels à l'étude d'une base de données environnementales pour déterminer les facteurs discriminants des pollutions observées.

Les réseaux de neurones se montrent efficaces dans l'analyse simultanée d'un grand nombre de paramètres. Ils permettent en particulier d'identifier des relations entre données de terrain difficiles à percevoir simplement.

L'étude des modèles obtenus sur une base de données de HAP a mis en évidence les composés les plus caractéristiques : ANT, FLN et BaP. Un réseau à quatre neurones intermédiaires s'avère suffisant pour l'obtention de ce résultat.

En utilisant les valeurs de teneurs en HAP dosés dans un prélèvement de sol d'un site contaminé, les RNA estiment immédiatement le type d'activité polluante le plus probable.

Les performances sur les tests effectués confirment la capacité de généralisation de ces modèles. La recherche se développera sur les pollutions caractérisées par des termes sources plus complexes, impliquant des solvants chlorés industriels, en utilisant la méthodologie testée dans le cas des pollutions en HAP.

Références

[1] B.N. Ames, J. McIann, E. Yamashaki, Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test, Mutat. Res. 31 (1975) 347–364.

[2] M. Blumer, Polycyclic aromatic compounds in nature, Scientific American 234 (1976) 35–46.

[3] C. Calderon-Macias, M.K. Sen, P.L. Stoffa, Artificial neural networks for parameter estimation in geophysics, Geophys. Prospect. 48 (2000) 21–47.

[4] F. Cazier, M.N. Duval, H.C. Dubourguier, E. Degans, Identification of pollutants in soils of former coal industries by GC/MS and HPLC/MS, in : H. Verachtert, W. Verstraete (Eds.), Proc. Int. Symp. Environ. Biotechnol., Antwerpen, 1997, pp. 1895–1902.

[5] F.U. Dowla, L.L. Rogers, Solving problems in environmental engineering and geosciences with neural networks, MIT Press, Cambridge, UK, 1995, 240 p.

[6] K.C. Jones, J.A. Stratford, K.S. Waterhouse, N.B. Vogt, Organic compounds in Welsh soils: Polynuclear Aromatic Hydrocarbons, Environ. Sci. Technol. 5 (1989) 540–550.

[7] N.R. Johnston, R. Sadler, et al., Environmental modification of PAH composition in coal tar containing samples, Chemosphere 27 (1993) 1151–1158.

[8] L.H. Keith, W.A. Teillard, Priority polluants – a perspective view, Environ. Sci. Technol. 13 (1979) 416–423.

[9] J. Oosterbaan, Utilisation des méthodes de l'analyse exploratoire de données environnementales pour l'établissement de descripteurs macroscopiques de la dynamique des termes sources polluants dans la géosphère, thèse, École des mines de Paris, Fontainebleau, 2000, pp. 38–44.

[10] J. Oosterbaan, P. Jamet, Caractérisation de pollutions en HAP : contribution de l'analyse exploratoire de données d'audit internationales, Les Techniques de l'Industrie Minérale 3 (3e trimestre) (1999) 57–63.

[11] P.J. Wilson, J.H. Wells, Charbon, coke et sous-produits, Librarie polytechnique Charles-Béranger, Paris, 1953, pp. 209–251.

[12] M.B. Yunker, L.R. Snowdon, et al., Polycyclic Aromatic Hydrocarbon composition and potential sources for sediment samples from the Beaufort and Barents Seas, Environ. Sci. Technol. 30 (1996) 1310–1320.