Estimation de la migration d'une pollution accidentelle dans un projet routier à l'aide des réseaux de neurones artificiels

E. EL TABACH
L. LANCELOT
I. SHAHROUR
H. MAILLOT

Laboratoire de Mécanique de Lille (UMR CNRS 8107) Université des Sciences et Technologies de Lille (USTL), Polytech'Lille avenue Paul-Langevin 59655 Villeneuve-d'Ascq
Eddy.El-Tabach@polytech-lille.fr
Laurent.Lancelot@polytech-lille.fr
Isam.Shahrour@polytech-lille.fr
Henri.maillot@polytech-lille.fr

Y. NAJJAR

Kansas State University, Dept. of Civil Engineering Manhattan, KS 66505 USA
ea4146@ksu.edu

Résumé

L'évaluation de la profondeur de la zone contaminée D en fonction du temps et de la quantité de polluant injectée dans un sol Q après une pollution routière accidentelle est essentielle pour étudier le risque de contamination de la ressource en eau souterraine et pour concevoir des plans d'intervention. Cet article présente une méthode pour estimer D et Q en utilisant les réseaux de neurones artificiels. Une base de données est produite à partir de cas simulés en utilisant un modèle par éléments finis. Plusieurs modèles de réseaux de neurones artificiels par rétropropagation de l'erreur sont évalués par leur capacité à généraliser la simulation sur des données indépendantes. Leur comportement est également comparé à un modèle plus classique de régression multilinéaire. Les réseaux de neurones montrent une très bonne aptitude à simuler les évolutions de D et Q.
La méthodologie proposée est appliquée à l'analyse du risque de pollution par le trichloréthylène des eaux souterraines le long de l'axe d'un projet routier dans le Nord de la France.

Mots-clés : eau, hydrocarbure, NAPL, non saturé, pollution accidentelle, réseaux de neurones artificiels, route, trichloréthylène.

Estimating the migration of an accidental pollution in a highway project using artificial neural networks

Abstract

Accurate estimation of depth of contaminated zone D and the quantity of pollutant injected into a soil Q after an accidental pollution occurred in road transport is essential to asses the risk of water resources contamination. This paper presents a method for estimating D and Q after an accidental pollutant discharge at the soil surface. First a database is generated from simulated cases using a finite element model. For each case, D and Q are computed as a function of the most related parameters. Different feedforward artificial neural networks with error backpropagation are trained and tested using subsets of the database, and the ability of these networks to generalize on independent simulated data are validated on another subset of the database. Their behavior is compared and analyzed with regard to more common multilinear regression approximation tool. The proposed method is used to analyze the risk for a DNAPL pollution of groundwater resources concerned by a road project in the north of France.

Key words : accidental pollution, artificial neural network, hydrocarbons, NAPL, numerical model, road, trichloroethylene, unsaturated, water.

NDLR : Les discussions sur cet article sont acceptées jusqu'au 1er mars 2006.

1

Introduction

Les accidents routiers impliquant le déversement de matières toxiques ou dangereuses au cours de leur transport peuvent poser de graves problèmes environnementaux. Les hydrocarbures et les solvants chlorés (qui sont des liquides non miscibles avec l'eau, appelés *non aqueous phase liquids*, ou NAPL, dans la littérature anglo-saxonne) sont parmi les matières transportées les plus dangereuses. Ces produits ont des effets variables selon la quantité et la nature du produit déversé et la sensibilité du milieu récepteur. Les produits plus denses que l'eau (DNAPL) posent des problèmes majeurs car ils migrent plus profondément sous l'effet de la gravité et le volume de produit non piégé dans la zone non saturée du sol peut atteindre la nappe phréatique et la contaminer. L'impact d'un projet routier sur l'environnement doit ainsi être étudié avec attention dans le but, d'une part, d'optimiser son tracé lors de la phase de conception en fonction de diverses contraintes, dont la vulnérabilité des zones traversées à la pollution accidentelle et, d'autre part, pour élaborer un plan d'intervention en cas d'accident mettant en jeu une pollution.

Les zones vulnérables à la pollution le long du tracé routier peuvent s'étendre sur plusieurs kilomètres. Sur cette distance, la topographie, la géologie et les propriétés du sol peuvent varier dans une large mesure. De nombreux modèles numériques permettant de simuler le transfert des NAPL dans les sols non saturés existent (par exemple Guarnaccia *et al.*, 1997 ; Katyal *et al.*, 1991). Basés sur une description précise des mécanismes régissant le transfert (écoulement multiphasique, échanges entre phases, transport du polluant dilué dans les phases liquide et gazeuse), ces logiciels supposent la connaissance de paramètres nombreux dans une zone hétérogène, et ils exigent des efforts de calcul importants.

L'approche proposée dans cet article vise à limiter le recours à ces modèles et repose sur l'estimation de l'évolution de la contamination du sol à l'aide des réseaux de neurones artificiels. Ces réseaux sont construits sur une base de données obtenue par simulations numériques. La capacité de différents réseaux à généraliser la simulation à l'ensemble de la zone d'étude a été évaluée, analysée et comparée à une technique plus classique, à savoir la régression linéaire multiple. Le réseau optimal a été employé pour prévoir la profondeur de la zone contaminée dans le cas étudié, concernant le déversement accidentel de trichloréthylène le long de l'axe d'un projet routier dans le Nord de la France.

2

Problème étudié – Méthodologie

2.1

Présentation du cas étudié

Le projet routier concerne le passage à 2 x 2 voies de la RN2 entre Avesnes-sur-Helpe et Maubeuge (environ 20 kilomètres) dans le département du Nord. Globalement orienté N-S, le projet traverse une vallée près de Bachant, où des captages importants d'eau potable sont localisés.

Dans ce secteur, des études géologiques et hydrogéologiques ont été réalisées, qui ont montré que le tracé envisagé repose sur une couche de calcaires carbonifères fracturés protégée par une épaisseur de 5 à 15 mètres de limons, sables limoneux et argiles limoneuses, et que la nappe d'eau exploitée se trouve à une profondeur variant de 5 à 25 mètres, à l'intérieur des calcaires fracturés (Fig. 1). Une série de sondages et d'essais a été effectuée : 26 sondages à la tarière, 5 sondages carottés, 20 piézomètres et 6 essais au pressiomètre. Ces sondages ont donné des indications sur la profondeur de la couche de couverture limoneuse. Les piézomètres ont permis de suivre les variations de profondeur de la nappe sur une période d'environ deux ans (septembre 2000 à avril 2002). En outre, des essais de perméabilité *in situ* et au laboratoire ont été effectués, et la perméabilité mesurée varie entre 8×10^{-10} et 5×10^{-7} m/s selon la teneur en argile. Les courbes de rétention de la couche de couverture limoneuse ont été également mesurées sur les cinq échantillons carottés en utilisant la méthode de l'extracteur à plaque, et les paramètres de la relation de Van Genuchten (1980) ont été déterminés en laboratoire (voir Tableau I).

La figure 1 montre également les zones en déblai et en remblai. Dans les zones en déblai le terrassement de la couche protectrice de couverture limoneuse peut augmenter la vulnérabilité de l'aquifère sous jacent pendant la phase de construction, mais ces zones seront protégées en phase d'exploitation (plate-forme étanche, bassins de rétention imperméables, dispositifs de collecte des effluents pollués). Par contre, dans les tronçons en remblai, il y a risque pour les véhicules de quitter accidentellement la route et de se retrouver au niveau du terrain naturel, menaçant ainsi les ressources d'eaux souterraines dans les zones vulnérables. La migration des polluants dans les sols limoneux de la couche de couverture de l'aquifère doit par conséquent être soigneusement analysée dans les zones où le profil est en remblai.

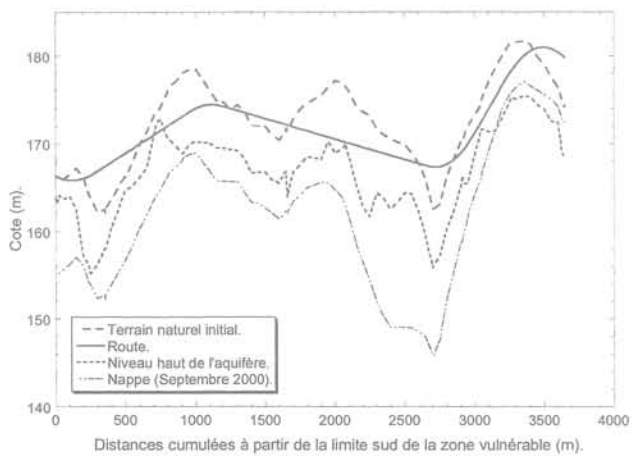


FIG. 1 **Coupe verticale de l'axe de la route.**
Cross-section along the axis of the highway project.

TABLEAU I **Propriétés des fluides et des sols.**
Fluid and porous media properties

		Eau	TCE	Air
Propriétés des fluides	ρ (Kg/m^3)	998,2	1456	1,17
	μ (Kg/(m.s))	0,001	0,000556	0,00002
	$\sigma_{\alpha\beta}$ (dynes/m)	7275	3174	4750
Propriétés des sols	Type du sol	ϕ	k (m/s)	ρ (kg/m^3)
	Limon	0,36	10^{-6}, 10^{-7}, 10^{-8}	1760
	Craie	0,42	1×10^{-5}	1600
Relation pression capillaire-saturation	Saturations résiduelles (tous les types de sol)			
	S_{WR}	S_{GR}	[a]S_{NNWR}	[a]S_{NWR}
	0,068	0,02	0,16	0,12
	Paramètres du modèle de Van Genucten (1980)			
	Type du sol	[b]a_D (cm^{-1})	[b]a_I (cm^{-1})	n
	Limons	0,004	0,008	1,25
	Craie	0,02	0,04	1,3

[a] S_{NNWR} and S_{NWR} sont, respectivement, la saturation résiduelle en polluant comme phase non mouillante (avec l'eau) ou comme phase mouillante (avec l'air).
[b] *D* pour drainage et *I* pour imbibition.

2.2

Méthodologie adoptée

Une étude numérique préliminaire par éléments finis a été entreprise pour simuler le transfert vertical d'un polluant de type NAPL. Cette étude a permis de dégager les paramètres ayant une influence sensible sur la réponse du modèle en terme de profondeur de sol contaminé à l'issue d'une certaine durée de mise en contact du polluant à sa surface. Parmi ceux-ci, certains ayant une importante variabilité naturelle (épaisseur de la couche de couverture, profondeur de la nappe, perméabilité de la couche de couverture) ont été retenus comme paramètres d'entrée pour les outils de prédiction de la contamination, de même que le paramètre temps de contact sol-polluant, qui est déterminant.

Compte tenu des résultats des différents sondages effectués le long du tracé de la route, et des temps d'intervention sur site prévisibles en cas d'accident, des intervalles représentatifs pour les quatre paramètres retenus ont été définis. Une base de données a été construite à partir des calculs par éléments finis effectués pour des combinaisons des paramètres pris dans leurs intervalles représentatifs. Cette base de données a ensuite servi à construire des modèles de prédiction de D et Q pour toute combinaison de paramètres d'entrée non comprise dans la base de données, afin d'estimer la migration de la pollution sur l'ensemble de la zone d'étude.

Différents modèles de prédiction de type réseau de neurones artificiels ont été construits en utilisant cette base, et leur aptitude à généraliser la simulation à des cas non utilisés pour leur construction a été évaluée. Le modèle le plus performant a été utilisé pour déterminer le profil de pollution le long de l'axe du projet routier.

3

Modélisation numérique

3.1

Modèle NAPL-Simulator

Le logiciel NAPL-Simulator (Guarnaccia *et al.*, 1997) a été développé pour simuler la migration des NAPL dans les sols non saturés. Pour un sol indéformable et isotrope, l'équation d'écoulement pour une phase donnée peut être écrite sous la forme suivante (Abriola et Pinder, 1985) :

$$\phi \frac{\partial (S_\alpha \rho_\alpha)}{\partial t} = \nabla \cdot [(\rho_\alpha k k_{r\alpha} / \mu_\alpha)(\nabla P_\alpha + \rho_\alpha g \nabla z)] \quad (1)$$

L'indice α représente ici la phase fluide (eau, air ou polluant), ϕ est la porosité du milieu poreux, S_α et P_α sont respectivement la saturation et la pression de la phase α, K est la perméabilité intrinsèque, $K_{r\alpha}$ est la perméabilité relative à la phase α, μ_α et ρ_α sont respectivement la viscosité et la masse volumique de la phase α, g est l'accélération gravitationnelle et z est la profondeur. Le logiciel comprend une description de la relation perméabilité relative-saturation-pression dans les milieux poreux biphasiques ou triphasiques et prend en compte les hystérésis et le piégeage des fluides. La technique de résolution utilisée dans ce logiciel est basée sur la méthode des éléments finis avec un schéma implicite en temps.

3.2

Cas de référence

La colonne de sol de référence est représentative du profil géologique rencontré dans le Nord de la France (Fig. 2). Une couche constituée de limons ou d'alluvions, ayant une épaisseur H_c de 4,5 mètres et une perméabilité moyenne $K = 10^{-7}$ m/s, recouvre une couche aquifère de craie d'approximativement 40 mètres d'épaisseur, renfermant la nappe exploitée, de profondeur $H_W = 2$ mètres. Le modèle numérique utilisé est unidimensionnel. Les propriétés des sols, des fluides et les paramètres de la relation de saturation-pression de Van Genuchten (1980) sont données dans le tableau I.

Les simulations sont menées en considérant comme polluant le trichloréthylène (ou TCE), qui appartient à la famille des solvants chlorés. Le TCE compte parmi les produits les plus dangereux, en raison de sa basse viscosité et de sa densité relative élevée, et les cas de pollution de sols ou de nappes par ce produit font l'objet d'une littérature abondante (Pankow *et al.*, 1996). Les simulations sont effectuées en deux phases (Fig. 2). La première phase consiste à déterminer le profil de saturation initiale en eau qui est conditionné par le niveau de la nappe (H_W) et les propriétés de rétention des sols. La seconde phase concerne le rejet du polluant, qui est simulé par l'application d'une charge constante e = 5 cm à la surface du sol pendant un laps de temps t_c. Pour chaque simulation le profil de saturation ainsi que la quantité de polluant injecté dans le sol sont calculés en fonction du temps.

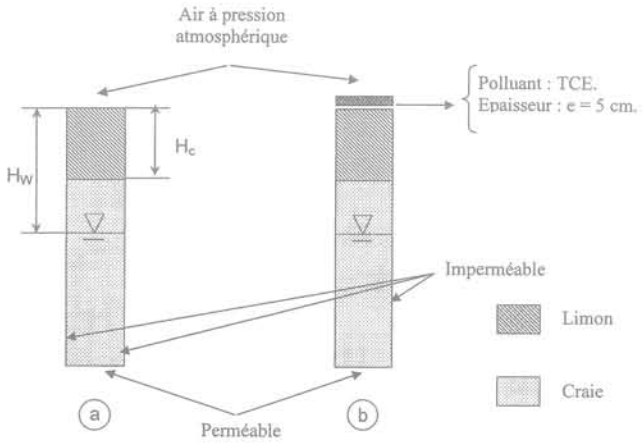


FIG. 2 **Conditions initiales et aux limites pour les phases : (a) de drainage de la colonne de sol, (b) de simulation du déversement accidentel.**
Initial and boundary conditions for (a) drainage phase of the soil column, and (b) simulation of the accidental spill.

4

Choix des variables d'entrées et constructions de la base de données

4.1

Étude paramétrique

Une étude préliminaire a étudié l'influence des paramètres principaux du sol et du polluant et de la géométrie du problème sur la profondeur de la zone contaminée et la quantité de polluant infiltré, estimées à l'aide du logiciel NAPL-Simulator (Guarnaccia *et al.*, 1997) pour la colonne de sol de référence (Lancelot *et al.*, 2003). L'objectif poursuivi était double : identifier les paramètres prépondérants de la migration et estimer l'impact des erreurs et incertitudes liées à leur détermination. Il a été montré que la perméabilité K du sol de couverture a un impact majeur sur les simulations : elle contrôle à la fois la profondeur et la vitesse de migration du polluant. L'effet de la variabilité spatiale des paramètres géométriques comme l'épaisseur de la couche de couverture H_c et la profondeur de la nappe phréatique H_w a aussi été examiné. Si la couche de couverture, dont la perméabilité est nettement plus faible que celle de l'aquifère sous jacent, est plus épaisse, il est clair que la protection de cette dernière contre la pollution sera plus efficace, tant en termes de profondeur contaminée que de quantité de polluant infiltrée. Par ailleurs, un niveau de nappe plus élevé se traduit par une migration du polluant plus rapide, car les forces de rétention dans la zone non saturée sont plus faibles. Enfin, les circonstances du déversement de polluant à la surface du sol ont également été prises en compte, et il a été montré que la durée t_c du contact entre le sol et le polluant était le facteur déterminant.

En conclusion de cette étude paramétrique, quatre paramètre d'entrée pour le modèle de prévision de la migration du polluant dans le sol ont été retenus : l'épaisseur H_c et la perméabilité K de la couche de couverture, la profondeur H_w de la nappe et le temps de contact t_c entre le polluant et la surface du sol.

4.2

Base de données pour les modèles de prédiction

La base de données est employée pour créer les modèles de prévision de D et de Q. Elle a été construite à partir de simulations par éléments finis pour les combinaisons suivantes des paramètres d'entrée : 1×10^{-6} m/s, 1×10^{-7} m/s et 1×10^{-8} m/s pour K, 0,5, 1, 3, 5 et 7 jours pour t_c, 15 valeurs entre 0 et 20 mètres pour H_c et 9 valeurs entre 0 et 45 mètres pour H_W.

Au total, la base de données inclut plus de 2 000 cas, qui sont divisés en 3 sous-ensembles. Le premier sous-ensemble, portant sur environ la moitié de la base de données, est employé pour créer les modèles par apprentissage. Le deuxième sous-ensemble (25 % de la base de données) est utilisé pour tester la performance des modèles en cours d'apprentissage. Ce processus permet en particulier d'optimiser la durée de cette phase, et de choisir l'architecture optimale (nombre de

neurones « cachés »). Le troisième sous-ensemble est enfin utilisé pour valider les modèles de prévision sur des cas « vierges ».

On notera que K intervient par son logarithme dans la base de données de manière à ne pas faire intervenir des ordres de grandeur trop hétérogènes dans les données. De même chaque paramètre d'entrée et de sortie a été normalisé relativement à ses valeurs minimum et maximum, ce qui permet, en disposant d'un ensemble de données plus homogène, un meilleur apprentissage du réseau.

5

Modèles de prédiction de la migration de la contamination

5.1

Modèles de régression linéaire multiple (RLM)

A partir de la base de données de cas de migration de polluant, il s'agit de généraliser l'estimation de cette migration pour tout point de la zone d'étude, donc pour toute combinaison des paramètres d'entrée. Le moyen le plus simple d'effectuer cette estimation est de procéder par régression linéaire sur les 4 paramètres d'entrée.

On recherchera donc une approximation g dépendant linéairement des 4 variables d'entrée X_j :

$$Y = g(X) = g(X_1, ..., X_4) = a_1X_1 + ... a_4X_4 \qquad (2)$$

Ainsi D (modèle 1) et Q (modèle 2) sont exprimés en fonction de log (K), H_c, H_w et t_c. Les coefficients du modèle sont déterminés par la méthode des moindres carrés à partir du premier sous-ensemble de la base de données, correspondant à la phase d'apprentissage. La qualité de prédiction du modèle peut être estimée en calculant le coefficient de détermination R^2 (équation 10) sur les données du sous ensemble d'apprentissage, mais également sur le reste de la base de données, pour tester l'aptitude du modèle à généraliser sa prédiction.

Le tableau II résume les valeurs du coefficient de détermination obtenues par régression multilinéaire pour la profondeur contaminée D et la quantité de polluant injectée Q, sur l'ensemble de données correspondant à l'apprentissage (donc au calcul des coefficients de la régression, équation 2) et sur des données n'ayant pas servi au cours de la phase d'apprentissage. On trouve des valeurs de R^2 assez faibles (de l'ordre de 0,5 à 0,6), notamment pour Q, en particulier pour les données de validation. Un modèle linéaire sur les paramètres d'entrée choisis ne constitue donc pas un outil de simulation satisfaisant.

5.2

Réseaux de neurones artificiels

5.2.1

Construction des réseaux

La structure et le fonctionnement des réseaux de neurones artificiels sont très documentés (par exemple Fausett, 1994 ; Ripley, 1996 ; Najjar *et al.*, 1997 ; Maier et Dandy, 2000). On se contentera ici d'une brève description des RNA et de l'implémentation utilisée dans cette étude.

Les réseaux de neurones artificiels sont constitués de cellules élémentaires de calcul (nœuds ou neurones) interconnectés. Les réseaux les plus répandus sont les perceptrons multicouches à rétropropagation de l'erreur. L'architecture d'un tel réseau est schématisée sur la figure 3. Elle comporte 3 couches. La couche d'entrée transmet les variations du phénomène modélisé au réseau, dont les réponses sont matérialisées dans la couche de sortie. Une ou plusieurs couches intermédiaires (ou cachées) sont interconnectées aux couches d'entrée et de sortie. Le rôle de ces couches cachées est de permettre au réseau d'associer les entrées données aux sorties également connues, lors d'un processus d'apprentissage.

Mathématiquement, un réseau avec trois couches, où n est le nombre de nœuds d'entrées, m le nombre de nœuds cachés et k le nombre de nœuds de sortie, est basé sur l'équation suivante :

$$O_k = S\left(\sum_{j=1}^{m} W_{jk} \times S\left(\sum_{i=1}^{n} W_{ij} X_i\right)\right) \qquad (3)$$

où S est une fonction de transfert, les facteurs W_{jk} sont les poids des connexions entre les neurones de la couche cachée et de la couche de sortie, les W_{ij} les poids des connexions entre les neurones de la couche d'entrée et ceux de la couche cachée, les O_k sont les

TABLEAU II **Coefficients R^2 entre les valeurs cibles et les valeurs estimées pour les différents modèles.**
R^2 values between target and predicted outputs for all models.

R^2	Modèle RLM		Modèles MLP			
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	
			MLP (4, 29, 1)	*MLP (4, 32, 1)*	*MLP (4, 27, 2)*	
	D	*Q*	*D*	*Q*	*D*	*Q*
Phase d'apprentissage	0,639	0,527	0,978	0,989	0,976	0,993
Phase de test	–	–	0,978	0,980	0,974	0,984
Phase de validation	0,577	0,350	0,950	0,981	0,947	0,984

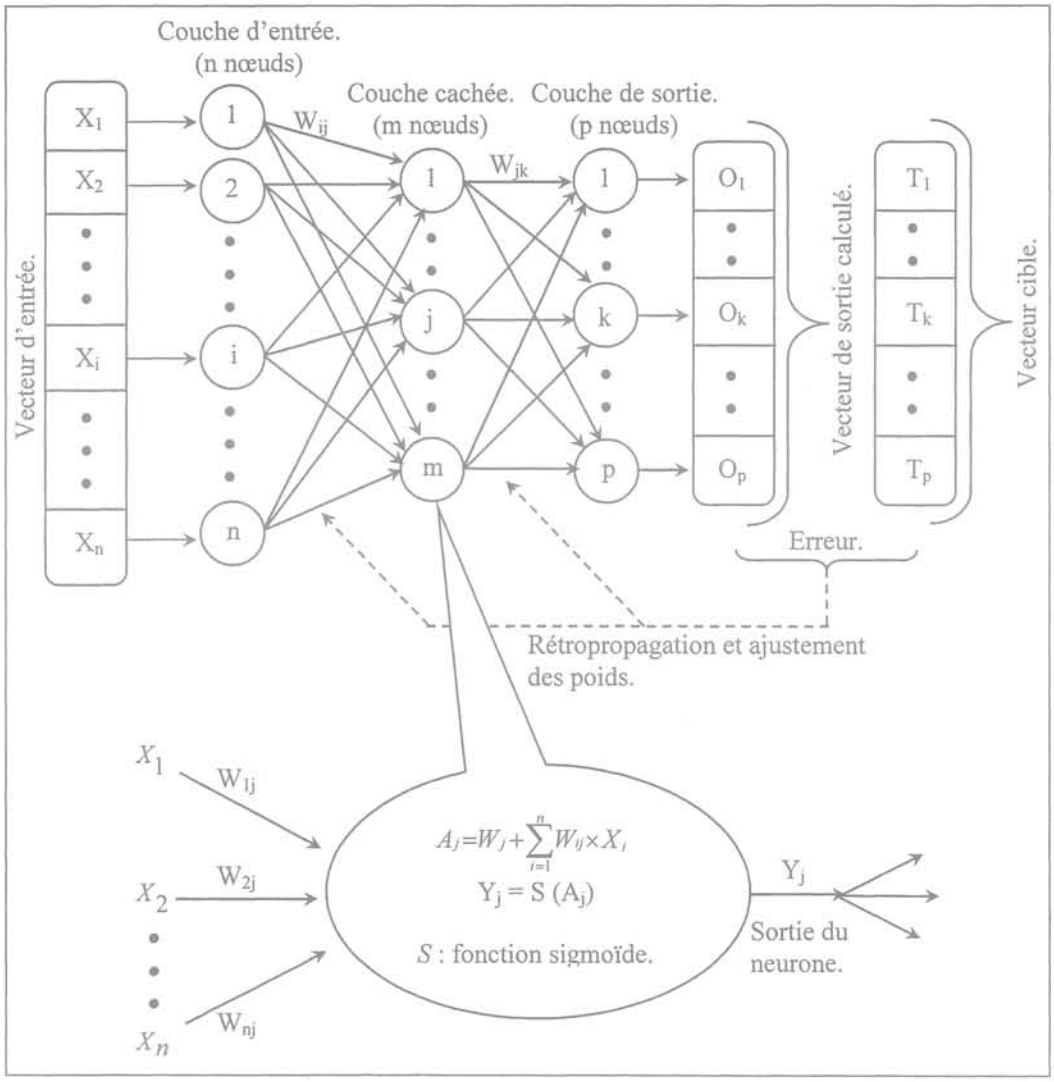


FIG. 3 **Architecture d'un réseau à 3 couches avec rétropropagation de l'erreur.**
Architecture of a typical multilayer backpropagation artificial neural network.

valeurs de sortie du réseau tandis que les X_i symbolisent les entrées.

Dans la plupart des applications de ces réseaux la fonction de transfert utilisée est la fonction sigmoïdale (éq. 4). Elle est continue et différentiable, qualités requises dans le processus d'apprentissage des perceptrons multicouches.

$$S(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Le processus d'apprentissage consiste en l'ajustement des poids entre les couches afin de minimiser l'écart entre les sorties calculées par le réseau et les sorties cibles, connues. Il repose sur une technique de recherche de minimum dans l'espace de l'erreur globale en fonction des poids constituant les paramètres du réseau. Les données dédiées à cette phase d'apprentissage sont propagées de l'entrée à la sortie du réseau, et les erreurs entre les valeurs en sortie et les valeurs cibles sont « rétropropagées » en ajustant les poids de chaque connection d'après une règle d'apprentissage *(delta-rule)* afin de réduire l'erreur globale. Cet aller-retour est effectué pour tout cas de la base d'apprentissage, et répété encore jusqu'à ce que les sorties simulées et les valeurs cibles du réseau coïncident, à une certaine tolérance près.

L'erreur globale à minimiser fréquemment utilisée est l'erreur quadratique moyenne (*average squared error*, ASE), définie dans l'équation 5 :

$$ASE = \frac{1}{p} \times \frac{1}{s} \times \sum_{q=1}^{s} \sum_{k=1}^{p} \left(T_{qk} - O_{qk}\right)^2 \tag{5}$$

où O_{qk} et T_{qk} sont respectivement les valeurs simulées et réelles du nœud de sortie k pour le cas q, s est le nombre de cas, p est le nombre de nœuds de sortie.

Si on peut en théorie approcher d'aussi près qu'on veut les valeurs cibles à condition d'augmenter suffisamment le nombre de cycles d'apprentissage, il faut noter que ceci se fait, à partir d'un certain nombre de cycles, au détriment de l'aptitude du réseau à généraliser ses prédictions à des cas non utilisés au cours de cette phase d'apprentissage, ce qui est pourtant le but poursuivi. La technique de validation croisée *(cross-validation)* consiste alors à calculer l'erreur globale ASE simultanément sur les données d'apprentissage et sur des données de test, indépendantes, jusqu'à atteindre le minimum de l'erreur sur ces données de test (Fig. 4). L'apprentissage est alors considéré comme terminé.

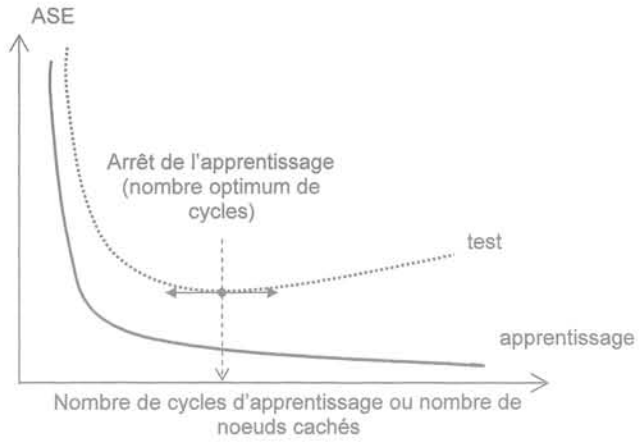


FIG. 4 **Critère de convergence et architecture optimale du réseau.**
Convergence criterion and optimum network architecture.

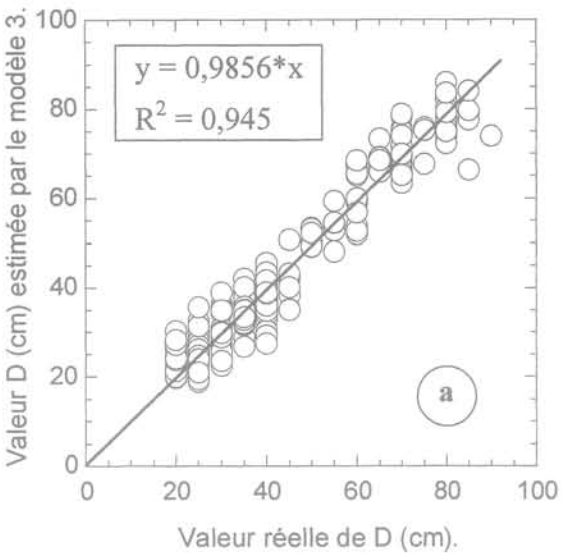
5.2.2

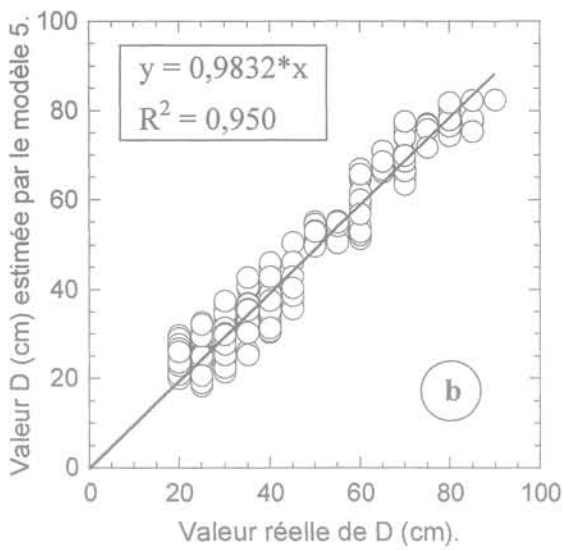
Architecture du réseau optimal

La performance globale d'un réseau dépend du nombre de couches cachées et du nombre de nœuds dans chaque couche cachée. Dans notre cas le réseau comporte 3 couches (soit une couche cachée). Pour déterminer le nombre optimal de neurones dans la couche cachée, on peut, de la même façon que pour la détermination du nombre optimal de cycles d'apprentissage, procéder par validation croisée. On augmente progressivement le nombre de nœuds dans la couche cachée en estimant à chaque fois l'erreur globale ASE (éq. 5) calculée pour le sous-ensemble de test de la base de données. Le nombre de nœuds cachés pour lequel cette erreur globale ASE commence à croître est pris comme optimum (Fig. 4).

Plusieurs formes d'architectures de réseau ont été tentées dans cette étude. Pour un réseau reliant les variables d'entrées $\{X_1, X_2..., X_n\}$ aux variables de sorties $\{O_1, O_2..., O_p\}$ et contenant une couche cachée avec m nœuds, on note :

$$\{O_1, O_2, ..., O_p\} = RNA_{n\text{-}m\text{-}p}\{X_1, X_2, ..., X_n\} \quad (6)$$





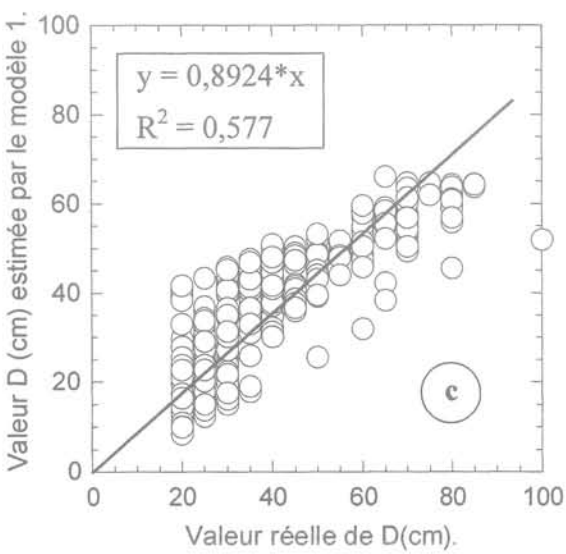


FIG. 5 **Comparaison entre les values cibles et les valeurs estimées de D pour les modèles 3 (RNA à paramètre de sortie unique, graphe a), 5 (RNA à 2 paramètres de sortie, graphe b) et 1 (régression multilinéaire, graphe c), phase de validation.**
Comparison between target and predicted values for D using single output BPNN model 1 (a), combined outputs BPNN model 3 (b) and multi-linear regression model 4 (c) for validation phase.

Il est possible d'estimer D et Q à partir de deux réseaux séparés (modèles notés 3 et 4 respectivement, voir équations 7 et 8), mais il est également possible d'estimer D et Q au sein du même réseau (modèle 5) en considérant deux nœuds dans la couche de sortie (éq. 9). Le calcul du nombre optimal de nœuds cachés est de 29, 32 et 27 pour les modèles 3, 4 et 5 respectivement.

$$(D) = RNA_{4\text{-}29\text{-}1}\ (K, H_c, HW, t_c)\ \text{(modèle 3)} \qquad (7)$$

$$(Q) = RNA_{4\text{-}32\text{-}1}\ (K, H_c, HW, t_c)\ \text{(modèle 4)} \qquad (8)$$

$$(D, Q) = RNA_{4\text{-}27\text{-}2}\ (K, H_c, HW, t_c)\ \text{(modèle 5)} \qquad (9)$$

Il faut noter que la prévision simultanée de plusieurs paramètres de sortie dans un même réseau n'exige pas nécessairement une plus grande complexité par rapport à des réseaux séparés. Dans notre exemple, le modèle 5 contient en effet moins de nœuds dans la couche cachée que les autres modèles pour une erreur globale comparable. Physiquement, la profondeur D de la zone contaminée et la quantité Q de polluant injecté étant liées, cette dépendance a été identifiée et prise en compte par le réseau de neurones artificiels.

5.3

Discussion sur la performance des modèles

La performance des modèles peut être visualisée en représentant graphiquement les valeurs simulées en fonction des valeurs cibles. La distance des points ainsi obtenus à la première bissectrice donne une indication sur la façon dont le modèle se comporte. Ainsi, la figure 5 montre la comparaison entre les valeurs prévues par réseaux de neurones et les valeurs cibles pour D, dans le cas du sous-ensemble de la base de données destiné à la validation, quand D est estimé en utilisant le modèle 3 (réseau avec D paramètre de sortie unique, figure 5a), ou le modèle 5 (réseau à sortie combinée D et Q, figure 5b). Sur les mêmes graphiques la droite de régression passant par l'origine est également tracée et le coefficient de détermination R^2 pour cette ligne est calculé selon l'équation 10 :

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(T_i - Y_i\right)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(T_i - \overline{T_i}\right)^2} \qquad (10)$$

où : N est le nombre total de cas considérés ;
Y_i est la valeur de sortie calculée par le modèle ;
T_i est la valeur cible connue ;
$\overline{T}$ est la moyenne de l'ensemble des valeurs cibles pour les N cas considérés.

Un coefficient de détermination R^2 proche de l'unité indique une forte corrélation entre les valeurs simulées et les valeurs cibles. Si la pente de la droite de tendance est proche de 1, le modèle constitue une bonne approximation des données cibles.

Comme montré sur la figure 5, les modèles 3 et 5 (réseaux de neurones) donnent de très bons résultats pour la simulation de D. Les coefficients de détermination pour les modèles 1 à 5 pour toutes les phases d'apprentissage, de test et de validation sont indiqués dans le tableau II. Par contre, comme signalé au paragraphe 6.1, le modèle de régression multilinéaire ne parvient pas à prévoir les variations de D et, surtout, de Q de façon satisfaisante. La faible performance du modèle linéaire ne doit pas étonner : les phénomènes physiques représentés dans la base de données sont complexes et fortement non linéaires. Dans les réseaux de neurones artificiels, la non-linéarité est prise en compte par l'utilisation de fonctions de transfert non linéaires (éq. 5), et le degré de complexité peut être contrôlé en variant le nombre de nœuds dans la couche cachée. Les réseaux de neurones artificiels apparaissent donc comme un outil valable de prévision de la migration de la pollution dans le cas étudié.

6

Application : analyse du risque pour un déversement accidentel de NAPL dans un projet routier

Dans cette section, on présente une application du réseau de neurones artificiel à l'estimation du risque de pollution de la nappe par déversement accidentel dans le cadre du projet routier présenté dans le § 2.1.

6.1

Simulation de la profondeur de contamination le long de l'axe du projet

La figure 6 donne la profondeur de pénétration de polluant D estimée en utilisant le modèle 5. Les variations de D sont données pour une position particulière de la nappe phréatique (septembre 2000, basses eaux) et 5 valeurs du temps de contact t_c (de 0,5 à 7 jours). La profondeur de pénétration du trichloréthylène D montre un profil assez uniforme, variant par exemple

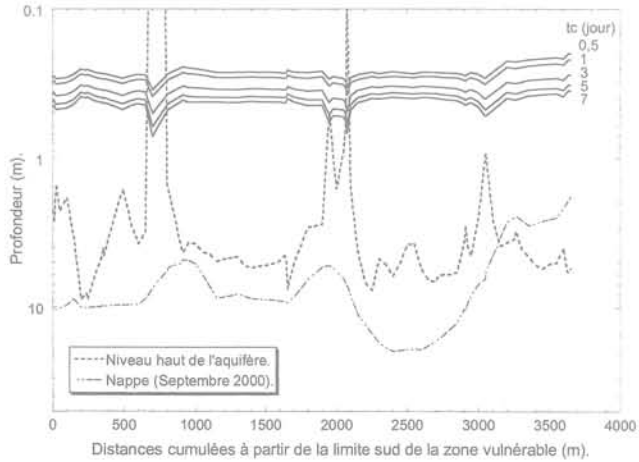


FIG. 6 **Évolution de D le long de l'axe de la route pour différentes valeurs du temps de contact t_c.**
Evolution of D along the highway project axis for different values of contact time t_c.

entre 0,2 à 0,3 mètre pour une période de contact de 12 heures, qui est le délai moyen d'intervention estimé pour un déversement routier accidentel (Martin et Roux, 1999). Des profondeurs plus importantes sont évidemment atteintes pour des temps de contact du polluant à la surface du sol plus élevés (environ 0,5 m pour t_c = 7 jours).

6.2

Risque de contamination des eaux souterraines

Le profil de la profondeur de pénétration de polluant le long de l'axe de la route est utile parce qu'il donne une estimation de l'épaisseur de sol devant être traité ou enlevé après déversement du polluant, pour supprimer toute menace envers les eaux souterraines. Cependant une meilleure quantification du risque de pollution des eaux souterraines est obtenue si la profondeur de contamination D est représentée rapportée à l'épaisseur de la couche de couverture limoneuse H_c. La figure 7 donne la variation du rapport $R = D/H_c$ pour t_c = 0,5 jour. On retrouve un risque de contamination très élevé dans les zones en déblai, où l'épaisseur du sol de couverture a été réduite, voire complètement enlevée. La menace est cependant limitée à la durée de la phase de construction, durant laquelle toute mesure doit être prise pour éviter un déversement de polluant. Mais les secteurs critiques sont les zones hachurées sur la figure 7, qui correspondent aux zones en remblai, où le risque de mise en contact du polluant avec le milieu naturel persiste sur toute la durée de vie de l'ouvrage. On note que dans ces zones en remblai, le facteur R atteint 20 % au sud, contre environ 5 % ailleurs.

Cette étude simplifiée d'analyse du risque aide à identifier les zones les plus vulnérables du projet routier. Une analyse plus fine peut alors être conduite dans les zones affichant un facteur de risque R élevé, comme l'étude de l'influence du type de polluant ou de celle de facteurs météorologiques, et peut servir de base à des dispositions préventives ou des mesures curatives en cas d'accident.

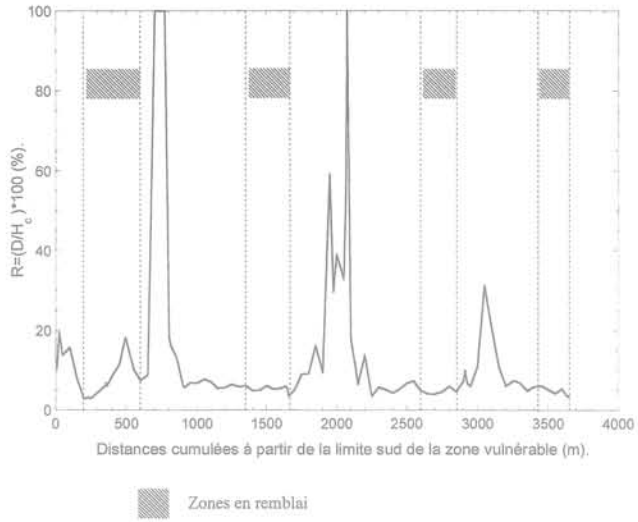


FIG. 7 **Évolution du facteur de risque R le long de l'axe du projet routier pour t_c = 0,5 jour.**
Evolution of risk factor R along the highway project axis for t_c = 0.5 day.

7

Conclusion

Cet article propose une approche basée sur les réseaux de neurones artificiels pour l'évaluation de la contamination d'un sol non saturé par déversement accidentel de polluant. L'étude vise l'analyse du risque de contamination des ressources en eau par déversement de polluants suite à un accident routier.

En l'absence des données de terrains sur la migration de polluants dans les sols non saturés, une base de données a été construite en utilisant une modélisation par éléments finis. Cette base de données a été ensuite utilisée pour calibrer le modèle « réseaux de neurones », qui a servi à établir le risque de pollution de la nappe dans une zone concernée par un projet routier.

L'étude réalisée montre que le modèle « réseaux de neurones » constitue un outil fort intéressant pour la modélisation du transfert de polluants dans les sols non saturés et pour l'étude d'impact de la construction des routes sur le sol et les ressources en eau.

Bibliographie

Abriola L.M., Pinder G.F. – A multiphase approach to the modeling of porous media contamination by organic compounds 1. Equation development. *Water Resources Research* 21, 1985, p. 19-26.

Fausett L.V. – *Fundamentals neural networks: Architecture, algorithms, and applications*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1994.

Guarnaccia J., Pinder G., Fishman M. – *NAPL : Simulator documentation*. EPA/600/R-97/102, Environmental Protection Agency, United States, 1997.

Katyal A.K., Kaluarachchi, Parker J.C. – *Mofat : a two-dimensional finite element program for multiphase flow and multicomponent transport, program documentation and user's guide*. EPA/600/2-91/020, Environmental Protection Agency, USA, 1991.

Lancelot L., El Tabach E., Shahrour I. – *Étude d'impact de la mise a 2 x 2 voies de la RN 2 entre Avesnes-sur-Helpes et Maubeuge sur les champs captants du synclinal de Bachant : modélisation des transferts de polluants*. Rapport final à la DDE de Nord, 2003.

Maier H.R., Dandy G.C. – Neural networks for the prediction and forecasting of water resources variables : a review of modeling issues and applications. *Environmental Modelling & Software* 15, 2000, p. 101-124.

Martin J.-C., Roux J.-C. – *Pollutions accidentelles routières et autoroutières*. Manuels et Méthodes 36, BRGM, France, 1999.

Najjar Y.M., Basheer I.A., Hajmeer M.N. – Computational neural networks for predictive microbiology : i. Methodology. *International Journal of Food Microbiology* 34, 1997, p. 27-49.

Pankow J.F., Stan Feenstra, Cherry J.A., Ryan M.C. – *Dense Chlorinated Solvents in Groundwater: Background and History of the Problem, Dense Chlorinated Solvents and other DNAPLs in Groundwater*. J.F. Pankow and J.A. Cherry (eds), Waterloo Press, Ontario, 1996.

Ripley B.D. – Pattern recognition and neural networks, Cambridge University Press, 1996, 403 p.

Van Genuchten M.T. – A closed-form equation for the hydraulic conductivity of unsaturated soils, *Soil Science Society of America Journal* 44, 1980, p. 892-898.

Zurada J.M. – Introduction to artificial neural systems, West Publishing Company, St. Paul, 1992.