

# UNE APPROCHE TRANSDUCTIVE POUR LA CARTOGRAPHIE DE RISQUE ENVIRONNEMENTAL

Nicolas Gilardi<sup>1</sup>

Tom Melluish

*Institut de Minéralogie et Géo chimie  
Université de Lausanne  
1015 Lausanne  
Suisse  
(e-mail: gilardi@idiap.ch)*

*Department of Computer Science  
Royal Holloway University of London  
Egham, Surrey TW20 0EX  
England  
(e-mail: tom@dcs.rhbnc.ac.uk)*

Le domaine des études environnementales est de plus en plus demandeur d'outils d'aide à la décision pouvant s'intégrer dans des systèmes d'information géographique. Dans le cas de problèmes de pollution, par exemple, il est utile de cartographier une estimation de l'incertitude de la prédiction faite sur la teneur en polluant. Une autre information utile est celle qui consiste à cartographier directement une estimation du risque de dépasser un seuil de pollution (généralement un seuil légal).

Ce type de carte est très utile dans la mesure où il permet de visualiser rapidement les zones en danger, les zones peu touchées, mais aussi les zones où il convient de mieux étudier le phénomène.

Un certain nombre de méthodes existe pour résoudre le problème du dépassement de seuil, en particulier en Géostatistiques. La plus simple d'entre elles est le Krigeage des Indicatrices (Goovaerts, 1997). Cette méthode ne requiert aucune autre contrainte sur les données que celles exigées par le Krigeage Ordinaire (Mathéron, 1962). En revanche, elle possède le défaut de "lisser" les résultats. Une autre approche consiste à utiliser les méthodes de simulation. Cependant, bien qu'elles fournissent d'excellents résultats, elles sont lourdes à mettre en oeuvre et exigent généralement de fortes contraintes sur la distribution de probabilité des données.

La méthode présentée dans cette publication utilise une approche "transductive" (Gamerman et al., 1998) pour évaluer la crédibilité de l'estimation d'un modèle lorsque celui-ci prédit, par exemple, que l'on va dépasser un seuil fixé. La seule hypothèse faite sur les données est qu'elles soient indépendamment et identiquement distribuées (iid) (Saunders et al., 1999). Néanmoins, l'algorithme présenté ici est une version "simplifiée" d'une méthode similaire mais plus complexe à mettre en oeuvre.

L'idée générale est la suivante: considérant une séquence  $z_1, \dots, z_l$  où  $z_i = (x_i, y_i)$ ,  $x_i \in \mathbb{R}^n$ ,  $y_i \in \mathbb{R}$ , et un élément  $z_{l+1} = (x_{l+1}, y_{l+1})$ , on construit un modèle sous-jacent permettant de calculer  $\hat{y}_{l+1}$ , estimation de  $y_{l+1}$ . Afin d'obtenir une information sur la crédibilité de cette estimation, on va

---

<sup>1</sup>IDIAP, C.P. 592, 1920 Martigny, Suisse

définir une fonction de *conformité*  $t : \mathbb{R}^{l+1} \rightarrow \mathbb{R}$  ayant la propriété suivante: la *probabilité* que la séquence  $(z_1, \dots, z_l, z_{l+1})$  soit telle que  $t(z_1, \dots, z_l, z_{l+1}) \leq r$  est inférieure ou égale à  $r$ .

La fonction

$$t(z_1, \dots, z_{l+1}) = \frac{\#\{i : \alpha_i \geq \alpha_{l+1}\}}{l+1}$$

où les  $\alpha_i = F(z_i, \{z_1, \dots, z_l, z_{l+1}\})$  (avec  $i = 1, \dots, l, l+1$ ) représentent l'*étrangeté* de l'élément  $z_i$  par rapport à la séquence  $(z_1, \dots, z_l, z_{l+1})$ , répond à cette condition.

Grâce à ces coefficients d'étrangeté et à cette fonction de conformité, il est possible, dans le cas d'un problème de régression, d'évaluer la confiance que l'on peut avoir que  $y_{l+1}$  dépasse la valeur seuil  $y^*$ . Pour cela, on peut procéder comme suit:

1. on construit un modèle de prédiction des  $y_i$  de  $(z_1, \dots, z_l, (x_{l+1}, y^*))$ <sup>1</sup> à partir d'un algorithme de régression donné.
2. on définit l'étrangeté de  $z_i$ , pour  $i = 1, \dots, l$ , par rapport à  $(z_1, \dots, z_l, z_{l+1})$  par  $\alpha_i = y_i - \hat{y}_i$ , où  $\hat{y}_i$  est l'estimation de  $y_i$  fournie par le modèle créé précédemment.
3. on définit l'étrangeté de  $z_{l+1}$  par rapport à  $(z_1, \dots, z_l, z_{l+1})$  par  $\alpha_{l+1} = y^* - \hat{y}_{l+1}$ .
4. on évalue la confiance que l'on a (en pourcent) que  $y_{l+1}$  dépasse  $y^*$  comme étant  $100(1 - \frac{\#\{i:\alpha_i \geq \alpha_{l+1}\}}{l+1})$ .

Dans le cadre des expériences présentées ici, l'algorithme de régression permettant de construire le modèle sous-jacent est la *Kernel Ridge Regression* (Saunders et al., 1998), utilisant une fonction noyau de type Gaussien. Les résultats ont été comparés au Krigeage des Indicatrices. Cette méthode est identique au Krigeage Ordinaire à ceci près qu'elle s'appuie non pas sur des données continues mais sur des *indicatrices*:  $I(z_i, y^*) = 1$  si  $y_i \geq y^*$ ,  $I(z_i, y^*) = 0$  si  $y_i < y^*$ . Le calcul du semi-variogramme devient  $\gamma(z_i, h) = \frac{1}{2}\text{VAR}\{I(z_i, y^*) - I(z_i + h, y^*)\}$ , où  $h$  est le pas vectoriel du semi-variogramme sur lequel on ajustera le modèle  $\hat{\gamma}(z_i, h)$ . La formule du Krigeage devient quant à elle:

$$F(z_{l+1}) = \sum_{i=1}^l \lambda_i I(z_{l+1}, y^*)$$

---

<sup>1</sup>C'est là que se situe l'aspect "transductif" de la méthode: le modèle est construit pour répondre spécifiquement au problème posé, à savoir évaluer l'écart de  $y_{l+1}$  par rapport au seuil  $y^*$

où les  $\lambda_i$  sont solutions du système:

$$\begin{cases} \sum_{j=1}^l \lambda_j \hat{\gamma}_{i,j} + \mu = \hat{\gamma}_{i,l+1} & i = 1, \dots, l \\ \sum_{j=1}^l \lambda_j = 1 \end{cases}$$

avec  $\hat{\gamma}_{i,j} = \hat{\gamma}(z_i - z_j, h)$ . Le résultat de ce calcul est une valeur comprise entre 0 et 1 interprétée comme une estimation de la probabilité de dépasser la valeur seuil.

Le résultat donné par la p-Value Transduction et le Krigeage des Indicatrices ne représente donc pas la même chose. La première méthode donne un pourcentage de *confiance* dans le fait que l'on dépasse ou non le seuil, tandis que la seconde donne une estimation de la *probabilité* de dépasser ce seuil. Cependant, ces deux fonctions sont corrélées positivement. La comparaison de leurs résultats respectifs est donc intéressante pour comprendre cette relation.

Les données utilisées pour cette comparaison ont été fournies par la Commission Internationale pour la Protection des Eaux du Léman (CIPEL). Il s'agit de l'analyse physico-chimique d'un ensemble de 294 échantillons de sédiments prélevés en 1983 sur l'ensemble du Lac Léman. Seules les analyses de Cadmium et les coordonnées X et Y des points de mesure ont été utilisées dans cette étude. 196 points ont été utilisés pour construire les modèles (Krigeage et Ridge Regression), et les deux méthodes ont donné une mesure de la probabilité, pour les 96 points restants, de dépasser une valeur seuil de concentration en Cadmium fixée à 0.8 ppm. Ce seuil a été choisi selon des critères purement statistiques<sup>2</sup>, et ne correspond en rien à des seuils légaux.

La méthode de comparaison principale consiste à classer les 98 points de test en fonction de leur Indicatrice, telle qu'elle a été définie pour le Krigeage des Indicatrices, puis à comparer cette classe avec l'estimation du risque "*R*" d'appartenir à la classe 1 donné par les modèles. Des comparaisons plus qualitatives ont également été étudiées (histogramme des valeurs estimées, corrélation spatiale, corrélations entre les deux méthodes, etc...).

	Indicator Kriging	pVT on KRR
Wrong 0 ( $R > 50\%$ )	15	15
Wrong 1 ( $R < 50\%$ )	7	9

Table 1: Comparaison du nombre de points mal classés par le Krigeage des Indicatrices et la p-Value Transduction appliquée à la Kernel Ridge Regression (pVT on KRR) pour les 98 points de test

---

<sup>2</sup>Il s'agit d'une valeur légèrement supérieure à la moyenne assurant ainsi une bonne représentation des deux classes de concentration.

Statistics	Indicator Kriging	pVT on KRR
minimum	0.000	0.005
first quartil	0.121	0.046
median	0.345	0.188
third quartil	0.478	0.457
maximum	1.000	0.995

Table 2: Comparaison de la distribution des valeurs des résultats donnés par le Krigeage des Indicatrices et la p-Value Transduction (pVT) appliquée à la Kernel Ridge Regression (KRR) pour les 98 points de test.

L'analyse des résultats d'erreur de classification montrent bien la forte corrélation entre les résultats de la p-Value Transduction appliquée à la Kernel Ridge regression, et ceux du Krigeage des Indicatrices. La comparaison des distributions des valeurs, en revanche, montre un net décalage de la médiane. Une étude plus approfondie des résultats de la pVT comparée, par exemple, aux simulations de données, permettra une meilleure interprétation des résultats. Des essais avec d'autres modèles sous-jacents et éventuellement d'autres coefficients d'étrangeté sont également envisagés.

## Bibliographie

[Deutsch et al., 1997] C.V. Deutsch and A.G. Journel (1997). Geostatistical Software Library (GSLIB) and User's Guide, Oxford University Press, New York.

[Gammerman et al., 1998] A. Gammerman, V. Vovk and V. Vapnik (1998). Learning by Transduction, Neural Information of Processing Systems.

[Goovaerts, 1997] P. Goovaerts (1997). Geostatistics for Natural Resources Evaluation, Oxford University Press, New York.

[Mathéron, 1962] G. Mathéron (1962). Traité de Géostatistique Appliquée, .

[Saunders et al., 1998] C. Saunders, A. Gammerman and V. Vovk (1998). Ridge Regression Learning Algorithm in Dual Variables, ICML98.

[Saunders et al., 1999] C. Saunders, A. Gammerman and V. Vovk (1999). Transduction with Confidence and Credibility, IJCAI99.