UMR 7041
Archéologies et Sciences de l'Antiquité

Cahier des thèmes transversaux ArScAn (Vol. IX)

2007 - 2008

- EXTRAIT -

Nanterre, Novembre 2009

Méthodes de cartographie et approche géostatistique

La cartographie de la pollution au dioxyde d'Azote en Alsace.

Ophélie LEMARCHAND
(Géovariances)
(lemarchand@geovariances.com)
Nicolas JEANNÉE
(Géovariances)
(jeannee@geovariances.com)

Résumé

L'article présente l'intérêt des méthodes géostatistiques pour la cartographie et l'analyse du risque environnemental, à travers une étude de pollution en dioxyde d'azote (NO_2) sur la région Alsace. Les méthodes usuelles d'interpolation sont discutées, ainsi que la valeur ajoutée de l'approche géostatistique. A partir des mesures et des informations complémentaires disponibles, la géostatistique permet de cartographier les concentrations sur le domaine d'étude, tout en associant à cette cartographie une estimation de la confiance que l'on peut lui accorder, l'incertitude associée au processus de prédiction spatiale étant inéluctable.

CONTEXTE DE L'ÉTUDE

La législation en place et une forte demande du public imposent aujourd'hui aux organismes responsables du suivi de la qualité de l'air de produire des cartes précises de polluant à partir d'observations issues de stations de mesures isolées. Les techniques classiques de cartographie ne suffisent plus. Les cartes qui en résultent sont souvent irréalistes et le choix de la méthode retenue pour les établir est fréquemment fondé sur des critères subjectifs.

L'apport de la géostatistique pour la cartographie est illustré dans cet article à travers le traitement de mesures de dioxyde d'azote en Alsace. Ces concentrations moyennes annuelles ont été calculées à partir de mesures collectées en 2004 par l'Association régionale de Surveillance de la Qualité de l'Air (ASPA) et ont fait l'objet d'une étude[1].

Le dioxyde d'azote est un polluant atmosphérique essentiellement issu des sources de combustions automobiles, industrielles et thermiques (chauffage) ; il est donc principalement présent en milieu urbain et dans les zones de trafic routier.

La carte ci-après présente les différentes stations de mesure en Alsace constituant le jeu de

1 - ASPA 2005.

Cahier des thèmes transversaux ArScAn (Vol IX)
Thème VII - Méthodes de cartographie et approche géostatistique
Nanterre, 2009, p. 203-214

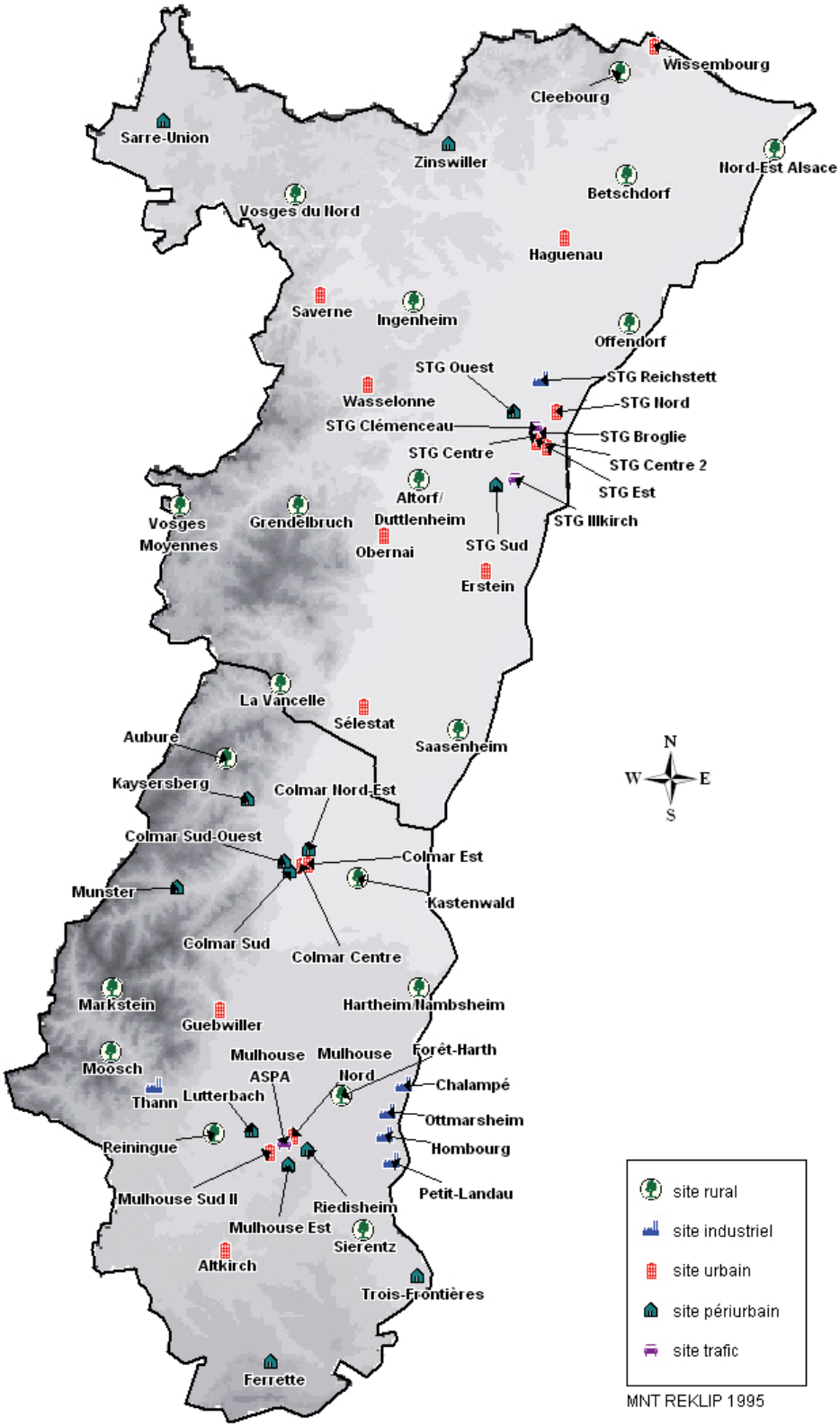

Figure 1 : Carte d'implantation des stations de mesure de NO_2 en Alsace (ASPA – Campagne régionale 2004, Source d'information ASPA 05020802-ID). (Les images seront disponibles en couleur sur le WEB)

données (fig. 1) établi par l'ASPA.

L'objectif de l'étude est de cartographier de manière continue la concentration en NO_2 à partir des mesures ponctuelles (fig. 1), sur l'ensemble de la région, dans le but de prédire la qualité de l'air aux endroits dépourvus de dispositif de mesure.

L'ensemble des modélisations géostatistiques a été réalisé avec le logiciel *Isatis*, version 8.4, développé par Géovariances.

1 - POURQUOI INTERPOLER ?

Un des objectifs de l'interpolation est de connaître la valeur la plus probable en chaque point du domaine d'étude à partir des données mesurées (en général l'interpolation est réalisée sur une grille régulière avec une certaine résolution) puis de représenter les principales tendances du phénomène via une carte. Cette carte doit être claire, simple et facilement interprétable par le destinataire, qu'il soit grand public ou scientifique.

Une dérive à éviter serait de vouloir atteindre une trop grande précision cartographique et de donner l'impression de connaître la valeur exacte en tout point. Une estimation sur une grille de résolution très fine (par rapport à l'échantillonnage des données) peut être ainsi mal interprétée. De même, étendre la zone au delà de la zone échantillonnée, c'est-à-dire travailler en extrapolation, peut poser des problèmes de communication. Il est donc important de garder à l'esprit une idée de la zone de représentativité des données.

2 - MÉTHODES USUELLES D'INTERPOLATION. APPROCHES DÉTERMINISTES

Il existe de multiples méthodes d'interpolation, l'objectif étant de choisir l'approche reflétant au mieux la physique du phénomène. Ce paragraphe vise à présenter certaines de ces approches déterministes et à en montrer les limites.

La majorité des algorithmes reposent sur le principe suivant : si l'on note Z la variable à prédire en un point x_0, l'interpolation (notée Z^*) a pour objectif d'estimer cette valeur par une combinaison linéaire des mesures disponibles aux points x_i, notées $Z(x_i)$:

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i . Z(x_i)$$

Connaissant les valeurs mesurées $Z(x_i)$, les algorithmes d'interpolation vont différer uniquement dans le choix des pondérateurs λ_i attribués à chaque mesure.

On retient que c'est la proximité spatiale entre les mesures et le point cible qui va déterminer la valeur interpolée. De nombreux phénomènes, s'ils impliquent une composante spatiale, ne peuvent cependant être simplement prédits sur base exclusive de cette composante spatiale, impliquant des mécanismes beaucoup plus complexes. A titre d'exemple, de nombreuses maladies découlent à la fois de facteurs environnementaux, géographiques (exposition au soleil) et socio-culturels (alimentation) ; le risque d'occurrence de telles maladies ne peut évidemment se prédire simplement par la proximité de cas déjà répertoriés et doit intégrer les autres facteurs de risque.

2.1. Interpolation par le plus proche voisin

La méthode du plus proche voisin (ou polygone d'influence) consiste à assigner à la valeur du point cible la valeur de la donnée mesurée la plus proche.

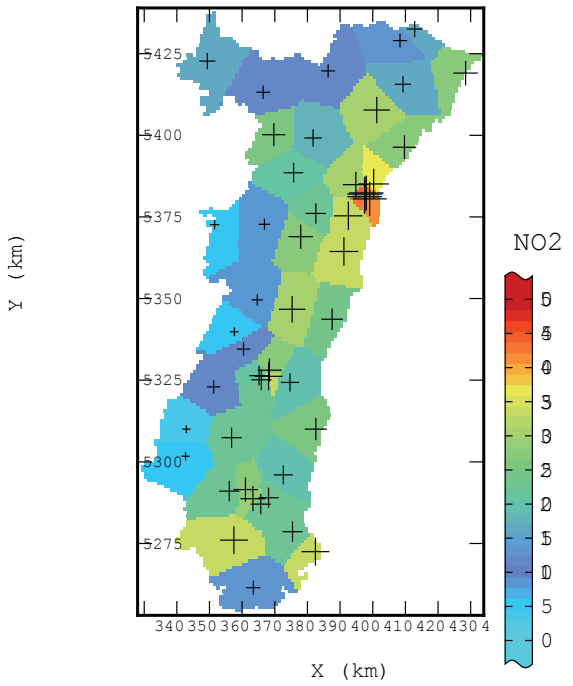

Figure 2 : Interpolation par plus proche voisin. La cartographie obtenue par cette méthode présente clairement des effets géométriques et ne semble pas du tout réaliste.

2.2. Interpolation par moyenne mobile

L'interpolation par moyenne mobile consiste à assigner à la valeur du point cible la moyenne des données situées dans son voisinage. La notion de voisinage est donc très importante et a une influence directe sur le résultat de la cartographie.

La carte produite par moyenne mobile présente toujours des effets géométriques qui tendent à être lissés avec l'augmentation du nombre de voisins considérés Plus ce nombre de voisins sera important, plus la carte se rapprochera de la moyenne des mesures.

2.3. Interpolation par distance inverse

L'interpolation distance inverse consiste à assigner à la valeur du point cible la moyenne pondérée des données situées dans son voisinage, le pondérateur affecté à chaque point de mesure étant inversement proportionnel à la distance entre

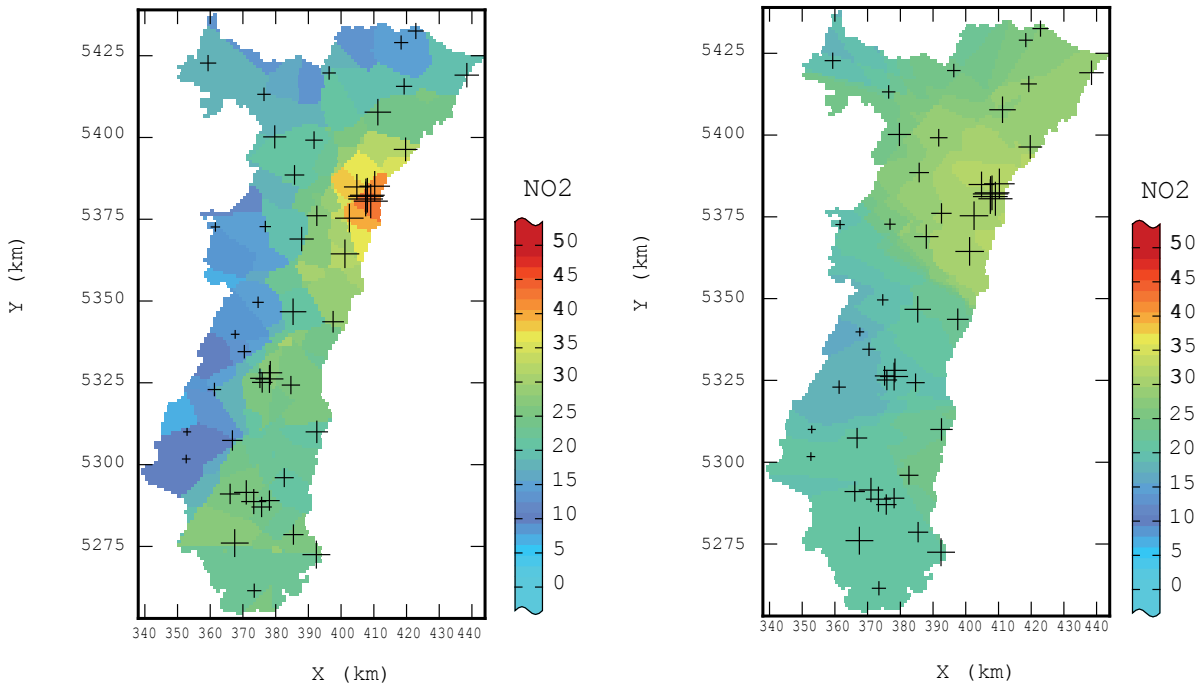

Figure 3 : Interpolation par moyenne mobile pour 3 voisins (à gauche) et pour 15 voisins (à droite).

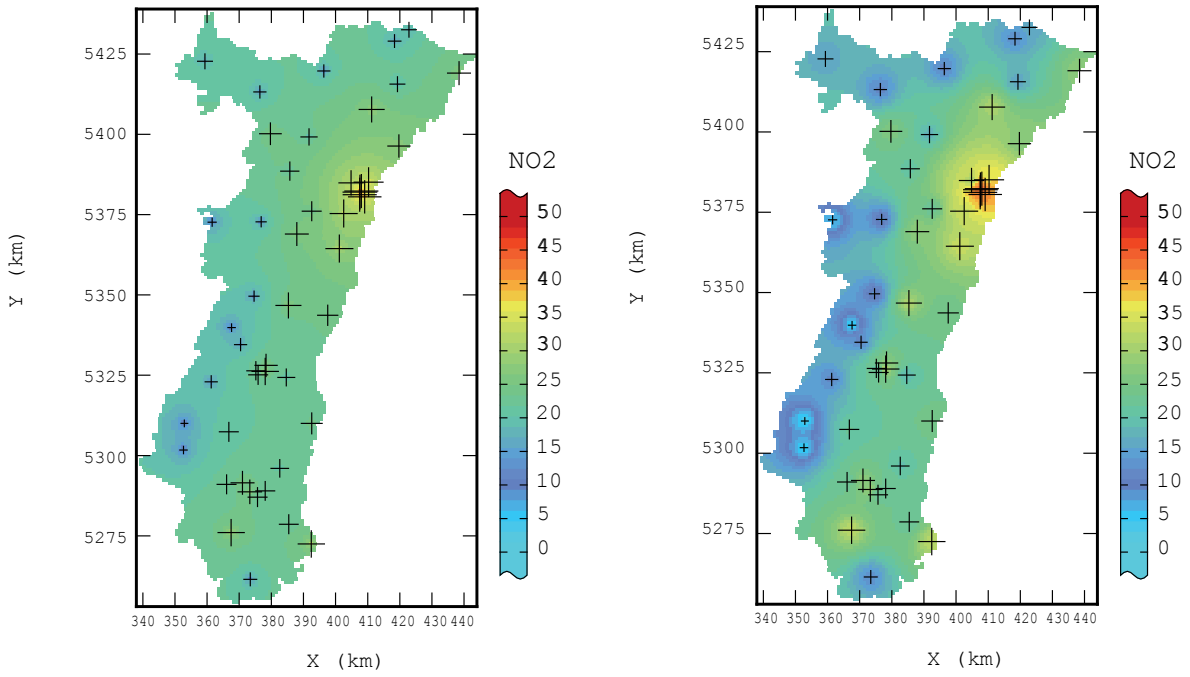

Figure 4 : Interpolation par inverse des distances (à gauche) et inverse des distances au carré (à droite).

la cible et le donnée et la somme des pondérateurs étant égal à 1 :

$$\lambda_i = \frac{1}{d(x_0, x_i)}, \; i = 1,...,n$$

Il est également possible de pondérer les mesures non pas par l'inverse des distances mais par l'inverse des distances au carré.

La carte produite par distance inverse présente moins d'artefacts, la pondération lissant les variations de concentrations.

À travers cet exemple, on peut voir qu'il est possible d'obtenir autant de cartes qu'il existe de méthodes. Le choix de la méthode est basé le plus souvent sur l'allure de la carte obtenue selon un critère subjectif (expérience, esthétique).

Dans le calcul de ces approches déterministes, seule la configuration géométrique des données intervient dans le calcul des pondérateurs. Pour une configuration de données on obtient la même pondération, que le phénomène soit régulier ou erratique. C'est sur ce point que la géostatistique se distingue des autres méthodes.

Pour une présentation plus exhaustive des méthodes d'interpolation déterministes, se reporter à la bibliographie[2].

3 - ANALYSE DE DONNÉES. STATISTIQUE OU GÉOSTATISTIQUE ?

L'analyse exploratoire des données est au cœur de la démarche géostatistique. En effet, c'est au cours de cette phase que se fait l'essentiel du travail. A travers le calcul et la visualisation d'outils de statistiques élémentaires, il est possible de mettre en évidence certaines caractéristiques statistiques de la variable étudiée.

L'*histogramme* est un moyen simple et rapide de représenter la *distribution* d'une variable c'est-à-dire la répartition des fréquences d'apparition de celle-ci. Il peut permettre de détecter d'éventuels anomalies et d'étudier la dispersion de la variable : valeurs extrêmes, mise en évidence de sous-

2 - Arnaud, Emery 2000.

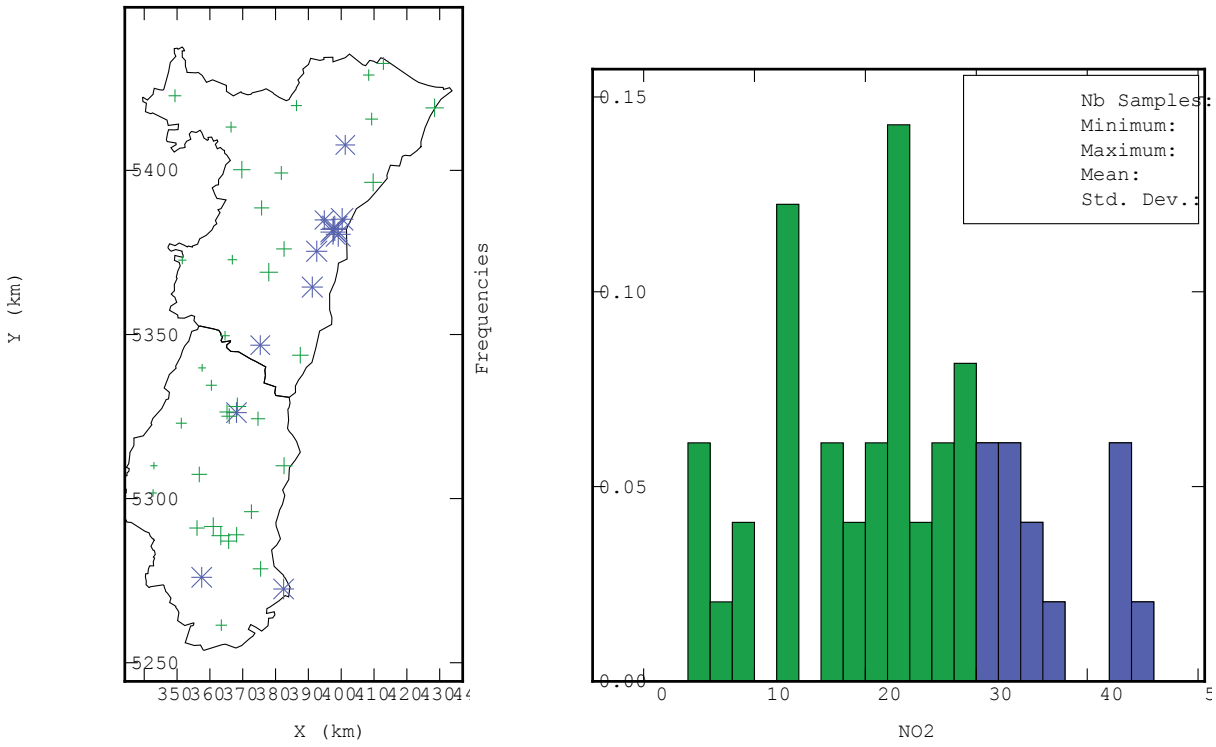

Figure 5 : Histogrammes des concentrations en NO_2 et localisation correspondante (indication en bleu des valeurs supérieures à 30 µg/m³).

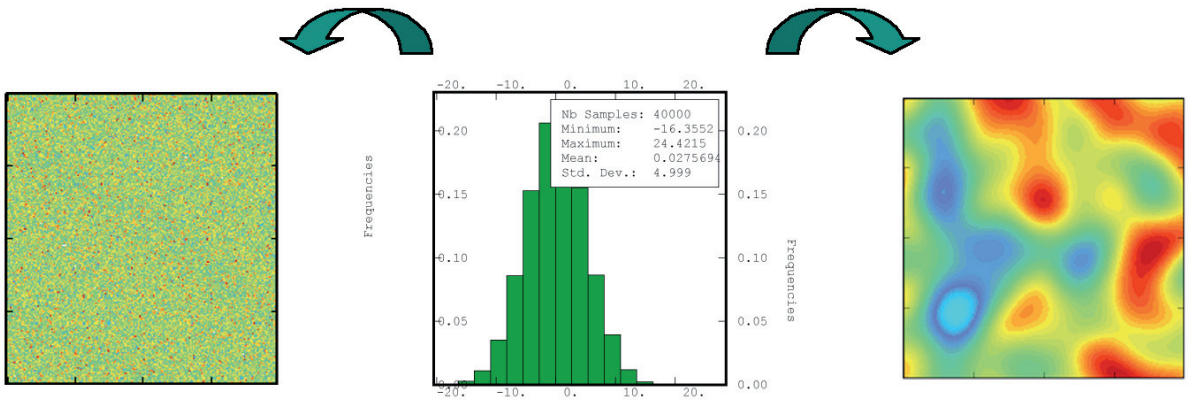

Figure 6 : Illustration de la limite des statistiques classiques.

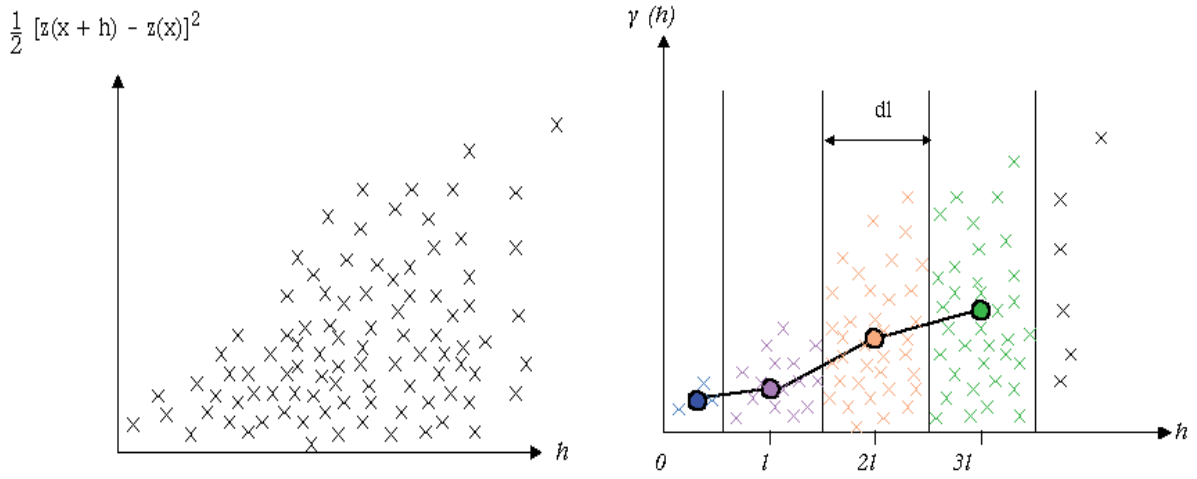

Figure 7 : Principe de calcul du variogramme expérimental.
Construire ensuite un modèle de variogramme, c'est-à-dire ajuster une fonction mathématique sur le variogramme expérimental, permet de connaître la valeur du variogramme pour toute distance h et permet d'intégrer la connaissance a priori sur la continuité spatiale du phénomène.

populations distinctes, etc.

La figure 6 illustre le fait que pour un ensemble de données montrant des statistiques identiques (même histogramme c'est-à-dire même moyenne, minimum, maximum, variance), la représentation cartographique est pourtant très différente. La variabilité spatiale de la carte de gauche est en effet beaucoup plus importante que celle de droite. Ceci illustre les limites des statistiques classiques et met en évidence le besoin d'outils supplémentaires permettant de quantifier la variabilité spatiale des données.

4 - LA CARTOGRAPHIE GÉOSTATISTIQUE

Note (Laurent Aubry)

Le krigeage est une méthode d'interpolation spatiale, considérée comme la plus juste d'un point de vue statistique, qui permet une estimation linéaire basée sur des grandeurs statistiques de la donnée spatialisée. Cette méthode se base sur le calcul, l'interprétation et la modélisation de la variance en fonction de la distance entre les données (variogramme). Contrairement aux autres méthodes, le krigeage s'accompagne d'un calcul d'erreur de l'estimation associée.

Pour un exemple d'application voir la figure 9.

4.1. Estimation monovariable

Tout comme pour la plupart des interpolations classiques, l'estimation par krigeage au point x_0 est obtenue par combinaison linaire des n mesures aux points x_i.

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i . Z(x_i)$$

Le krigeage se différencie uniquement par le choix des pondérateurs λ_i . Le pondérateur affecté à chaque donnée est déterminé par :

– la distance entre les données x_i et le point cible x_0 (tout comme les interpolateurs classiques) ;

– la distance séparant les données entre elles (clusterisation, existence de regroupements) ;

– le comportement spatial du phénomène.

La configuration des données et la position du point cible étant connues, il suffit pour réaliser

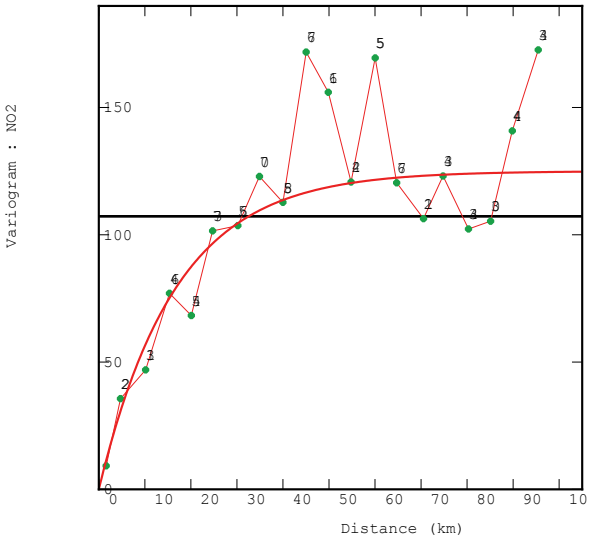

Figure 8 : Variogramme expérimental (trait fin rouge) et modèle de variogramme ajusté (trait épais rouge).

l'estimation de calculer expérimentalement la structure spatiale puis de la modéliser via une fonction, le variogramme. Le variogramme représente l'évolution de la variabilité du phénomène en fonction de la distance entre deux points. L'idée sous-jacente est qu'en moyenne, l'écart entre deux mesures proches est petit et l'écart entre deux mesures éloignées est grand. Ainsi le variogramme expérimental est déduit en calculant la quantité suivante pour chaque couple de points distants de *h* :

$$\gamma(h) = \frac{1}{2}(Z(x_i + h) - Z(x_i))^2$$

Les valeurs obtenues sont alors moyennées par classe de distance.

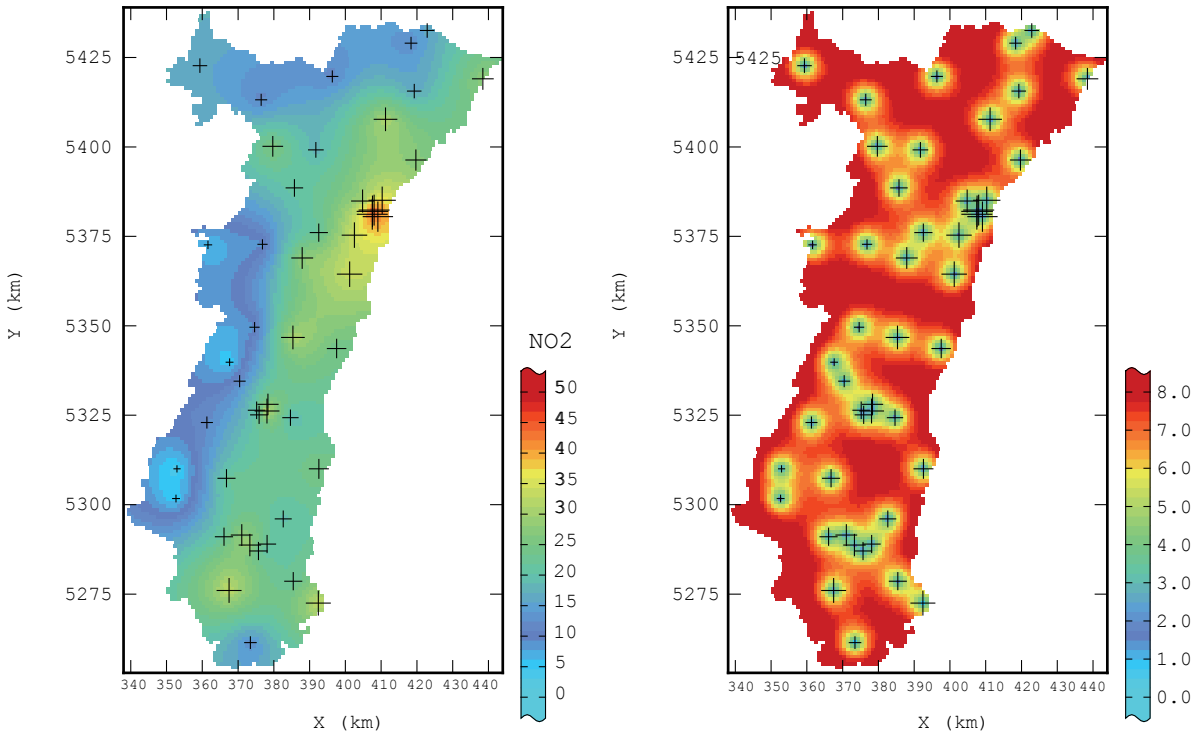

Figure 9 : Estimation de la concentration en dioxyde d'azote par krigeage ordinaire (à gauche) et écarts-types de l'erreur de krigeage associée (à droite).

La figure 9 illustre le résultat de l'estimation des concentrations en dioxyde d'azote par krigeage ordinaire – toutes les estimations sont données en µg/m^3. Cette carte laisse apparaître les grandes tendances des concentrations sur l'ensemble de la zone d'étude.

Un avantage essentiel du krigeage par rapport aux interpolateurs classiques réside dans la quantification de l'incertitude associée à l'estimation (qui existe toujours), rendue possible par la modélisation de la structure spatiale. Cette incertitude est usuellement représentée par la carte d'écarts-types (de l'erreur) de krigeage.

L'écart-type de krigeage :

– prend des valeurs minimales à proximité des points de données, la confiance que nous pouvons avoir en l'estimation étant alors grande,

– augmente lorsque l'on s'écarte des données ; cette augmentation est fonction de la structure spatiale choisie.

La carte d'écarts-types de krigeage peut alors être considérée comme un indicateur qualitatif de la précision de l'estimation. De ce point de vue, le figure 8 permet de bien identifier les zones mal reconnues, où la qualité de l'estimation est la moins bonne. L'utilisation quantitative des écarts-types de krigeage comme un outil permettant d'obtenir un intervalle de confiance nécessite des hypothèses supplémentaires et doit se faire prudemment.

4.2. Intégration de co-variables

Note (Laurent Aubry)

Le co-krigeage permet d'améliorer les estimations obtenues par krigeage en utilisant l'information fournie par d'autres variables (variables secondaires) que la variable principale.

Cependant, l'allure très lisse de l'estimation par krigeage, due à un échantillonnage localement pauvre, en limite le réalisme. Il est par conséquent intéressant de chercher à prendre en compte d'autres variables liées au dioxyde d'azote et qui pourraient en affiner l'estimation. Supposons en effet que nous connaissions, outre les valeurs de concentration en NO_2, une variable auxiliaire Z_2 en certains points potentiellement différents des stations de mesure. En cas de corrélation entre les deux variables, l'estimation de la concentration en dioxyde d'azote au point x_0 gagnerait à être obtenue

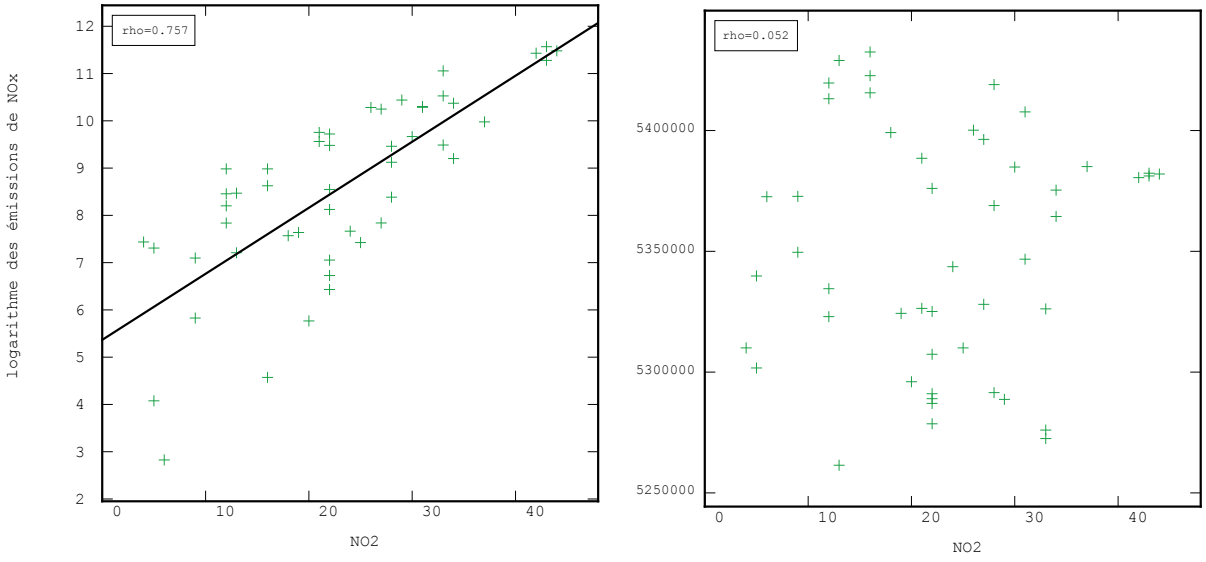

Figure 10 : Exemple de nuage de corrélation - Bonne corrélation entre les concentrations de NO_2 et le logarithme des émissions de NOx (à gauche) et absence de corrélation avec la coordonnée Y (à droite).

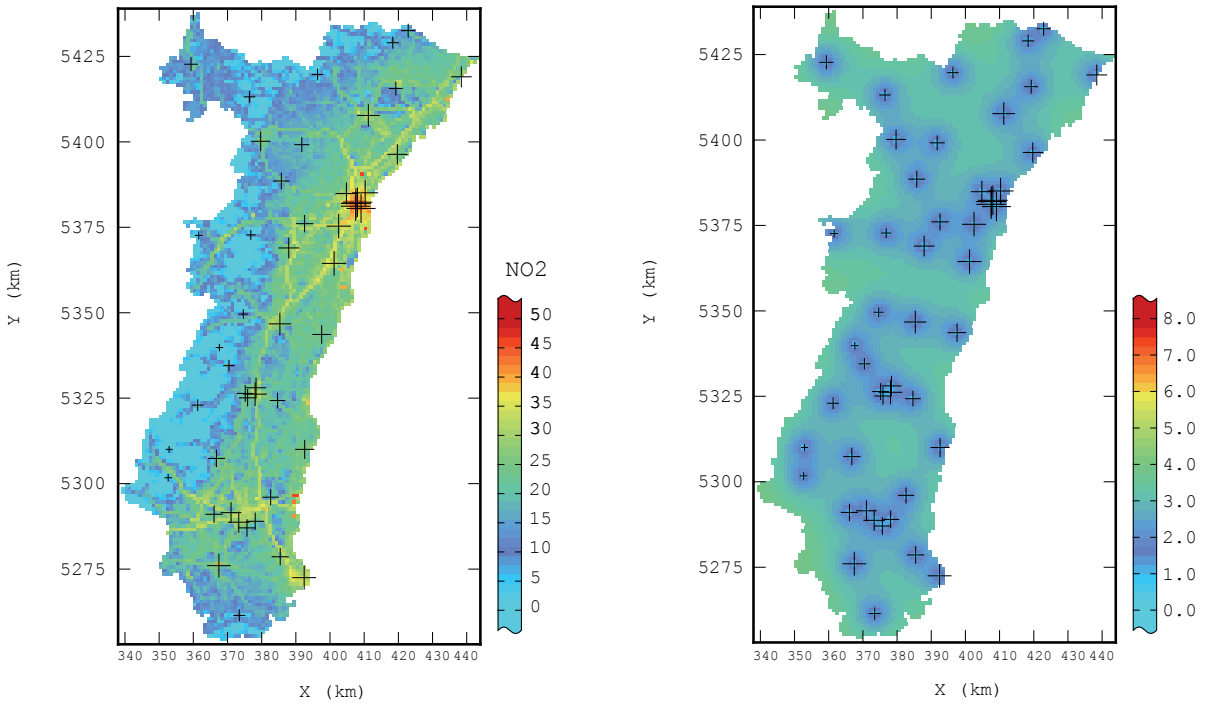

Figure 11 : Estimation de la concentration en dioxyde d'azote par co-krigeage colocalisé, intégration de l'altitude et du cadastre d'émissions (à gauche), et écarts-types de l'erreur de krigeage associée (à droite).

par combinaison linéaire des n concentrations aux stations de mesure x_i et des p valeurs de la co-variable aux points x_j.

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_{1i}.Z_1(x_i) + \sum_{j=1}^{p} \lambda_{2j}.Z_2(x_j)$$

La prise en compte de variables auxiliaires, corrélées à la variable d'intérêt, afin d'améliorer l'interpolation est alors appelée *co-krigeage*.

Le nuage de corrélation permet d'étudier le lien entre deux variables mesurées sur un même échantillon. Il consiste à représenter conjointement, sur un même graphique, les n couples de mesures $(x_1,y_1),\ldots,(x_n,y_n)$, en présentant en abscisse la 1ère variable et en ordonnée la 2nde.

Dans notre cas, la variable auxiliaire (correspondant à une combinaison entre altitude, le dioxyde d'azote ayant tendance à diminuer avec l'altitude, et émissions d'oxyde d'azote faisant ressortir le trafic routiers et les agglomérations) est connue en chaque nœud de la grille d'estimation. Prendre en compte l'ensemble des informations disponibles (49 points de mesure et 8302 nœuds de grille) conduirait à résoudre un système d'équations très lourd, alors qu'il apparaît intuitivement inutile d'utiliser toutes les valeurs de la co-variable pour estimer la concentration en dioxyde d'azote en un nœud donné. Le co-krigeage colocalisé est alors une « simplification » du co-krigeage qui consiste à ne retenir dans le système de co-krigeage que les points de mesure en NO_2, les valeurs de la co-variable en ces points de mesure, plus la valeur de la co-variable au point où l'on procède à l'estimation (donc 49*2+1 données en voisinage unique, et en voisinage glissant seulement les données situées dans le voisinage).

La figure 10 montre l'influence de l'intégration de la variable auxiliaire sur l'estimation des concentrations en dioxyde d'azote, rendant cette dernière nettement plus réaliste que les résultats

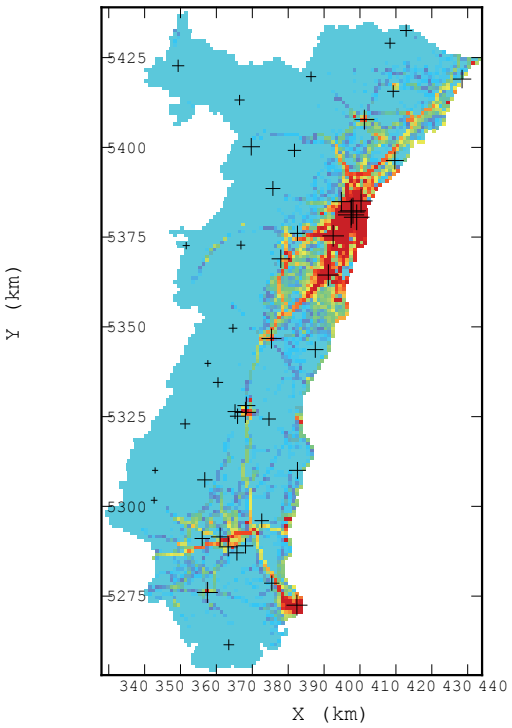

Figure 12 : Probabilité de dépasser un seuil de 30µg/m³.

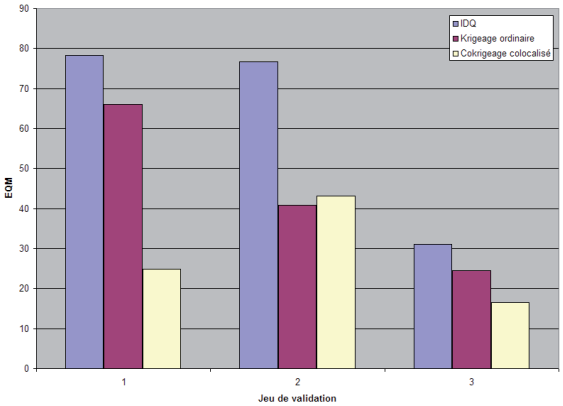

Figure 13 : Erreurs Quadratiques Moyennes (EQM) obtenues pour chaque méthode d'estimation sur les différents jeux de validation.

précédents. Cette estimation multivariable conduit également à une diminution de l'incertitude associée à l'estimation, les écarts-types obtenus étant inférieurs à ceux obtenus par krigeage ordinaire.

4.3. Dépassement de seuils de pollution

Par construction, les différentes techniques d'interpolation gomment les valeurs extrêmes, peu probables, et sont attirées vers la moyenne de la variable sur la zone d'intérêt. C'est la propriété de lissage du krigeage : la variabilité réelle du phénomène dans l'espace n'est pas reproduite quand on interpole les données. Pour fournir une réponse à la problématique fréquente de dépassement de seuils, comparer les valeurs de la carte produite par co-krigeage à un seuil et obtenir ainsi une réponse du type au-dessus/en dessous du seuil n'est donc pas correct. En effet, il est important de garder à l'esprit qu'à chaque carte produite par co-krigeage, il est possible de fournir une carte renseignant sur l'incertitude associée à l'interpolation. Ces incertitudes vont permettre de considérer la variabilité réelle du phénomène. Leur prise en compte est donc indispensable pour répondre à la question de dépassement de seuil, le résultat fourni n'étant alors pas une réponse binaire, dépassement/non dépassement, mais une probabilité de dépasser ce seuil. La méthode utilisée pour le calcul de cette probabilité est appelé espérance conditionnelle. Elle ne repose que sur un simple calcul de probabilité à partir des valeurs estimées et de l'incertitude associée mais elle nécessite une hypothèse de normalité de la variable (pouvant s'obtenir par transformation).

La carte suivante illustre la probabilité de dépasser un seuil de 30µg/m³ avec intégration des variables auxiliaires précédemment identifiées.

VALIDATION

Il est courant de conserver un certain pourcentage des données disponibles afin de tester, en ces points, la qualité de prédiction des différentes méthodes d'interpolation envisagées. Dans le cas présent, trois jeux constitués chacun

de 10% des données ont été sélectionnés de façon aléatoire parmi les données disponibles. Pour chaque jeu, les données de validation ont été retirées et les valeurs de concentrations ré-estimées en ces points à partir des 90% de données restantes. Différentes statistiques d'erreur peuvent alors être calculées pour chaque méthode d'interpolation à partir de trois jeux de validation. En particulier, l'erreur quadratique moyenne (EQM), i.e. la moyenne des écarts au carré entre valeur vraie (non prise en compte) et valeur prédite, est usuellement calculée. La figure ci-dessous montre ainsi, sur les trois jeux de données, la valeur ajoutée du krigeage, qui prend en compte la structure spatiale du phénomène, et surtout du cokrigeage qui intègre en plus la corrélation avec les variables auxiliaires.

CONCLUSION

Le recours à des méthodes d'interpolation déterministe, qui plus est de façon automatique, peut donner lieu à des résultats désastreux, notamment en raison de l'absence de tout moyen de contrôle sur le modèle construit. Il faut en effet accepter que toute estimation présente un risque inévitable d'erreur radicale dans les zones non échantillonnées. Il est par conséquent fondamental de tenir compte, non seulement des données numériques, mais également de toutes les autres sources d'information dont on dispose, connaissances générales sur le phénomène, expérience des praticiens[3]. Toute estimation doit être raisonnée : elle doit, autant que faire se peut, chercher à assurer que le résultat sera physiquement plausible et réaliste.

Lorsqu'ils sont mis en œuvre dans cet esprit, les outils géostatistiques apportent une valeur ajoutée indéniable pour la cartographie et l'analyse de risque de variables environnementales. Cette valeur ajoutée réside dans :

- l'utilisation de la structure spatiale intrinsèque du phénomène pour son estimation ;
- l'intégration rigoureuse de variables auxiliaires liées au phénomène d'intérêt, en améliorant ainsi l'estimation ;
- la quantification de l'erreur associée à chaque estimation.

3 - Matheron 1978.

REMERCIEMENTS

Les auteurs remercient l'ASPA, particulièrement Gilles Perron et Emmanuel Rivière, pour la mise à disposition de ces données.

Pour plus d'informations, www.atmo-alsace.net

BIBLIOGRAPHIE

ARNAUD M., EMERY X. 2000. *Estimation et interpolation spatiale. Méthodes déterministes et méthodes géostatistiques*. Paris : Hermès Science Publications.

ASPA 2005. *Cartographie régionale NO_2, C_6H_6, O_3 en Alsace*, ASPA 05020802-ID, ASPA.

MATHERON G. 1978. *Estimer et choisir. Cahiers du Centre de Morphologie Mathématique de Fontainebleau* 7, Ecole des Mines de Paris.

GEOVARIANCES 2008. *Technical References*, Geovariances, Ecole des Mines de Paris.