Article original

L'apport de la géostatistique à l'étude des risques liés à la pollution atmosphérique

Jacques Deraisme[1]
Michel Bobbia[2]
1. Géovariances, 49 bis, av. Franklin Roosevelt, 77212 AVON
<deraisme@geovariances.fr>

2. Air Normand, 21, av. de la porte des champs, 76000 ROUEN

Tirés à part : J. Deraisme

Résumé. Les applications de la géostatistique ont longtemps été réservées aux estimations minières et pétrolières. Depuis une décennie, la généralisation à l'étude de tout phénomène physique se déployant dans l'espace a vu une multiplication des contributions géostatistiques, aussi bien sur les plans théorique que pratique. Dans cet article, nous nous attachons à faire le point sur les apports de la géostatistique à l'étude de la pollution atmosphérique. Son objet est donc essentiellement de présenter les concepts fondamentaux de la discipline sous un angle intuitif et de les illustrer sur un cas réel. Celui-ci provient d'une campagne d'échantillonnage par tubes passifs de la pollution par le dioxyde d'azote (NO_2) dans l'agglomération rouennaise. Deux types de problèmes sont posés par les organismes en charge de la surveillance de la qualité de l'air : l'établissement de cartes « justes » de différents polluants, qui puissent être communiquées au public, et la quantification de risques (dépassement de seuils, exposition, etc.) par rapport à une réglementation qui se révèle de plus en plus précise. La géostatistique en tant que branche de la théorie des probabilités est particulièrement bien placée pour répondre à ces deux questions, car elle fixe un cadre méthodologique rigoureux et parfaitement adapté pour développer des méthodes opérationnelles.

Mots clés : modèle statistique ; surveillance environnement ; évaluation risque ; polluants atmosphériques.

Summary. Geostatistics in the study of air pollution-related risks

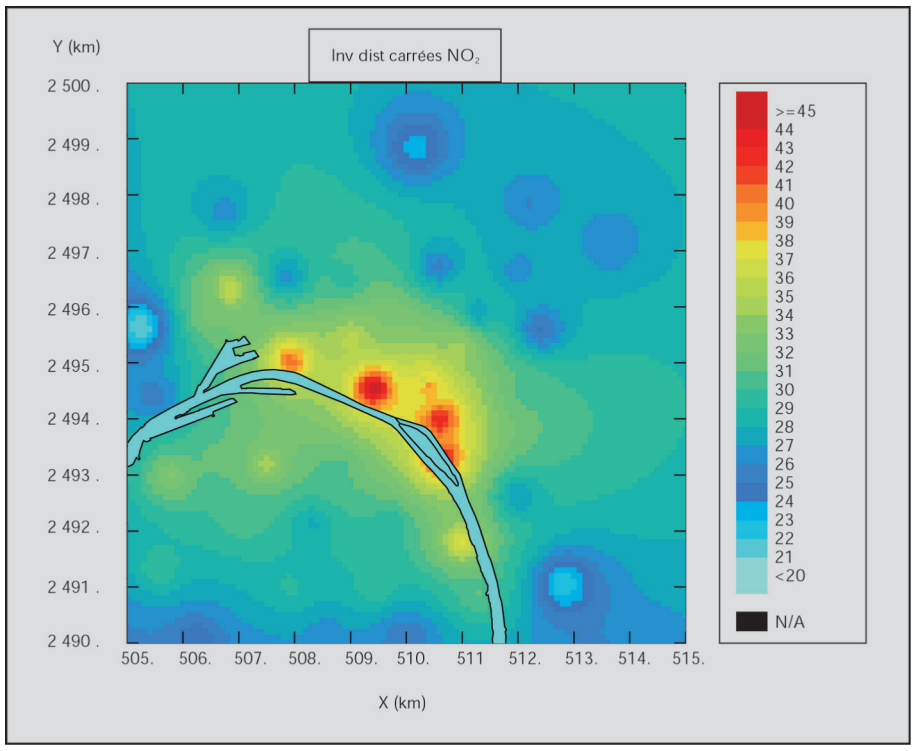
Geostatistics has long been applied in the mining and oil industries. In the past decade, its use has been extended to the study of all physical phenomena occurring in space, with important theoretical and practical results. This article, intended to present the contributions geostatistics can make to the study of air pollution, summarizes the basic concepts of this discipline from an intuitive perspective and illustrates them with a real case, based on a passive sampling campaign for nitrogen dioxide pollution in the Rouen area. This campaign had two separate goals: the provision of accurate maps of the different pollutants, to be made available to the public, and the quantification of risks (thresholds exceeded, exposure, etc.) in relation to ever more restrictive and precise regulations. Geostatistics, as a branch of probability theory, is particularly useful in dealing with these two issues, because its rigorous framework is especially appropriate for the development of operational methods.

Key words: models; statistical; environmental monitoring; risk assessment; air pollutants.

La diffusion dans un espace géographique donné d'un polluant, quel qu'il soit, est un phénomène dont les caractéristiques peuvent se résumer par les traits suivants : il est régi par des lois physico-chimiques complexes dépendant d'une multitude de paramètres locaux d'environnement ; il se déploie dans un espace à trois dimensions et il est par nature dynamique. Par opposition à la modélisation déterministe (très lourde à mettre en œuvre), la géostatistique[1] choisit délibérément un cadre probabiliste en proposant un modèle que l'on peut résumer de la façon suivante : le phénomène d'observation est considéré comme une

[1]Application des méthodes probabilistes à l'étude de variables régionalisées dans l'espace.

Figure 1. Carte de la concentration en NO_2 interpolée par inverse des carrés des distances.



N/A : valeurs manquantes.

réalisation d'un processus aléatoire. Ce processus est représenté par une fonction aléatoire Z(x, y, z, t) où x, y et z sont les coordonnées dans un espace géographique à trois dimensions et où t est le temps. Afin de simplifier l'exposé et de l'illustrer par des résultats concrets découlant de la théorie géostatistique, on se contentera ici d'une fonction aléatoire Z(x, y). Cela signifie que l'on s'intéresse à un niveau constant au-dessus du sol (zone dans laquelle nous respirons) et que l'on fera de la photo avant de faire du cinéma. Lorsque l'on voudra réintroduire la troisième dimension de l'espace ainsi que l'aspect temporel, des problèmes nouveaux et difficiles vont évidemment se poser, toujours objets de recherches.

La fonction aléatoire Z(x, y) doit être davantage spécifiée si on veut à la fois la caractériser à partir des observations (mesures) et l'utiliser pour répondre à des questions comme : quelle est la valeur probable de la concentration en polluant en un point de l'espace dépourvu de mesure ? Quelle est la probabilité qu'un seuil de pollution soit dépassé ? Parmi l'ensemble des fonctions aléatoires possibles, on se restreindra à des fonctions qui sont stationnaires, au moins pour ce qui concerne leurs moyenne et variance. Ce qui fait l'originalité du modèle géostatistique par rapport à d'autres modèles, en particulier statistiques, c'est le fait que l'on ne suppose pas *a priori* que les mesures en deux points sont non corrélées. C'est la caractérisation de cette corrélation par une fonction de la distance qui est ainsi au cœur de la démarche géostatistique.

Exemple de la pollution par le dioxyde d'azote

Afin d'illustrer notre exposé, nous avons utilisé quelques éléments d'une étude effectuée sur la base de campagnes de mesures par tubes à diffusion sur l'agglomération rouennaise. Six campagnes d'une durée de 15 jours chacune (en février, avril, mai, août, octobre et décembre 2000) ont permis d'échantillonner la concentration en NO_2 sur à peu près 80 sites dont la localisation a été choisie pour représenter la pollution dite de fond. On s'intéresse ici à la moyenne des six campagnes en chaque site, considérée comme représentative de la moyenne annuelle. Le premier problème qui se pose est de fournir une carte donnant le niveau de pollution en tout point. Une telle carte est généralement établie en deux temps : une interpolation sur un maillage régulier de l'espace suivie d'une restitution graphique. La question importante est donc de choisir la méthode d'interpolation adéquate. Il en existe un nombre pratiquement infini, parmi lesquelles la méthode dite d'inverse des carrés des distances *(figure 1)*. Cette méthode consiste à calculer en chaque nœud

du maillage une moyenne des mesures pondérée par des coefficients inversement proportionnels au carré de la distance entre le point interpolé et chaque point de mesure.

Réalisation d'une carte avec la géostatistique

Les méthodes classiques d'interpolation [1] présentent deux inconvénients majeurs. Le premier est le caractère arbitraire de la pondération effectuée : pourquoi telle méthode plutôt qu'une autre, pourquoi le carré de la distance et pas le cube ou la puissance 2,85 ? Le second est que les caractéristiques intrinsèques du phénomène interpolé n'entrent pas en ligne de compte : que la diffusion dans l'espace de la pollution soit très régulière ou plus chaotique, la pondération sera la même, puisque ne dépendant que de la configuration géométrique des données. Intuitivement, on voudrait que pour une distance donnée entre la mesure et le point interpolé, on donne à la mesure plus de poids dans le premier cas (diffusion régulière) que dans le second (diffusion chaotique).

À ces deux inconvénients se rajoute le fait que l'on ne sait pas qualifier la confiance que l'on peut accorder à la carte obtenue, alors que l'on sait qu'il existe une erreur d'estimation.

Le modèle probabiliste proposé par la géostatistique remédie à ces faiblesses dans la mesure où la façon d'interpoler va dépendre de la structure de variabilité du polluant dans l'espace, telle qu'elle peut être observée sur les mesures [2]. Par ailleurs, l'outil utilisé, le variogramme[2], permet de calculer la variance de l'erreur commise lors de l'estimation.

Analyse spatiale des données

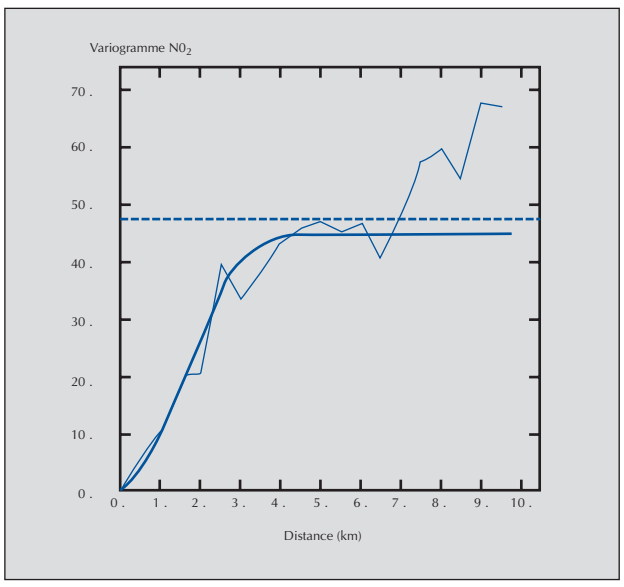
L'objet de cette analyse, fondement de la géostatistique, est de caractériser le degré de corrélation entre deux points de l'espace formant un vecteur $\vec{h}$. Pour un vecteur donné, on recherche tous les couples de mesures pour lesquels on calcule leur différence et leur variance. La variance des écarts peut ainsi être obtenue pour un certain nombre de vecteurs particuliers qui dépendent de la configuration géométrique de l'échantillonnage. C'est ce que l'on appelle un variogramme expérimental, qui est interprété comme un estimateur du variogramme sous-jacent de la fonction aléatoire. Après avoir « ajusté » les variogrammes expérimentaux par des fonctions mathématiques, on obtient un modèle de variogramme $\gamma(\vec{h})$ qui caractérise le degré de corrélation du polluant entre deux points quelconques de l'espace. Le variogramme est alors défini ainsi :

$$\gamma(\vec{h}) = \frac{1}{2} Var\,[Z(x+\vec{h}) - Z(x)],$$

où x est la position dans l'espace (vecteur des coordonnées) et $\vec{h}$ le vecteur distance.

Par la simple définition du variogramme, on comprend son utilisation pour caractériser la variance d'erreurs d'interpolation :

Figure 2. Variogramme expérimental et son ajustement par un modèle.



on peut en effet considérer le variogramme comme la variance de l'erreur commise lorsque l'on estime la valeur en un point par la valeur en un autre point distant de $\vec{h}$. La *figure 2* montre, sur l'exemple de la moyenne annuelle du NO_2 à Rouen, le variogramme omnidirectionnel et son ajustement.

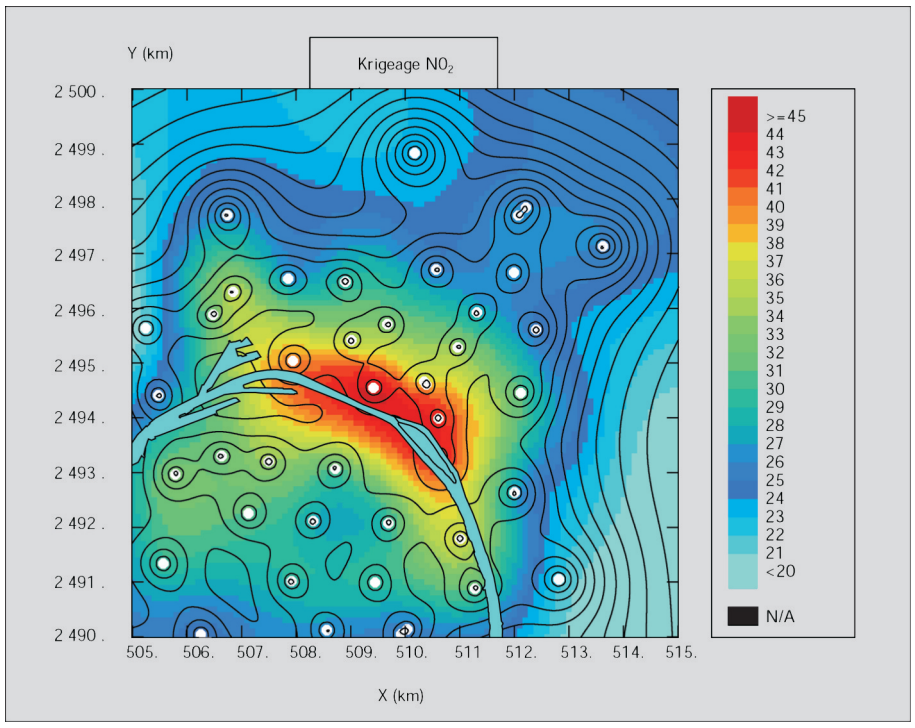
Le variogramme est ainsi une fonction croissante de la distance jusqu'à une distance critique, appelée portée. Au-delà de la portée, le variogramme se stabilise autour d'un palier, ce qui signifie que les valeurs sont alors indépendantes. Ainsi, plus la portée est grande, plus les valeurs des mesures auront une influence à grande distance. Le comportement du variogramme peut être différent selon les directions ; on aura alors un modèle de variogramme anisotrope, qui pourrait, dans certains cas, manifester l'influence de vents dominants dans la diffusion de la pollution : plus grande corrélation dans la direction des vents et, *a contrario*, augmentation plus rapide de la variabilité dans la direction perpendiculaire.

Krigeage des données

La méthode géostatistique pour interpoler sur un maillage à partir des mesures s'appelle le krigeage, nom donné par le professeur Matheron, fondateur de la discipline [3], en hommage au docteur Krige pour ses travaux sur les gisements d'or sud-africains. Il se définit comme le meilleur estimateur linéaire sans biais et consiste donc en un lissage aréal des mesures faites en différents points de l'espace à l'aide d'une pondération tenant compte de l'éloignement des sites de mesure. Le critère d'optimalité est que la variance de l'erreur, dite variance de krigeage, est minimale. Cette variance est calculable à partir du variogramme : par conséquent, les pondérateurs de krigeage dépendent du type de variogramme que l'on a ajusté. Le krigeage ne dépend pas simplement de la distance avec les sites de mesure mais aussi de la corrélation spatiale du polluant entre ces mêmes points. En outre, à la différence d'autres interpolateurs classiques,

[2]Fonction caractérisant le degré de corrélation spatiale des données.

Figure 3. Carte du NO_2 krigé, avec la localisation des sites de mesures et les courbes d'isovaleurs de l'écart type de krigeage.



le recours à un modèle probabiliste permet d'associer à la valeur interpolée sa variance de krigeage, qui est un indicateur de la qualité de l'interpolation. Sur la *figure 3*, on a représenté, en mode rastérisé[3], la valeur du NO_2 krigée aux nœuds d'une grille de maille carrée de 100 mètres. On a superposé sur la même figure les sites de mesures par des symboles (gros points blancs) et les courbes d'isovaleurs de l'écart type de krigeage. Il n'est pas surprenant de voir les courbes d'écart type se resserrer autour des points de mesure ; ce qui est intéressant, c'est d'observer la croissance de l'erreur lorsque l'on s'écarte des points de données.

Les mesures dont on dispose sont par ailleurs entachées d'incertitude, ce qui introduit une variabilité à petite échelle que l'on peut interpréter comme un bruit aléatoire dont l'intensité n'est pas nécessairement constante mais peut dépendre de la concentration mesurée en polluant. Le krigeage peut alors tenir compte de cette incertitude en affectant un pondérateur plus fort à une donnée certaine qu'à une donnée incertaine.

On peut généraliser le krigeage au cas où l'on disposerait sur les sites de mesure de valeurs de plusieurs polluants, corrélés entre eux au sens géostatistique, c'est-à-dire que le polluant 1 au point x est corrélé au polluant 2 au point $x + h$. On peut ainsi effectuer le *cokrigeage* du polluant 1 (ainsi que du polluant 2) à l'aide d'une combinaison linéaire des mesures du polluant 1 et du polluant 2. Les pondérateurs de cette combinaison linéaire sont alors calculés à partir des variogrammes des deux polluants. L'intérêt est d'améliorer la précision de l'estimation.

[3]Mode point par point.

Utilisation de co-facteurs

Il est certain que l'établissement d'une carte à partir des seules mesures du polluant se prive d'informations capitales pour la connaissance de la pollution. Le mécanisme de génération de la pollution par le dioxyde d'azote met en cause les émissions d'oxydes d'azote par les combustions à hautes températures, en particulier les émissions par les automobiles, les centrales thermiques, les usines d'incinération et les installations de chauffage. La corrélation statistique et géostatistique peut d'ailleurs être vérifiée et quantifiée à partir de l'analyse des concentrations mesurées et des variables, appelées co-facteurs, qui ont un lien avec la pollution (émissions, densité de population, topographie, occupation du sol, facteurs environnementaux et météorologiques).

L'utilisation de ces co-facteurs dans des procédures de krigeage permet d'améliorer l'interpolation lorsque les corrélations sont significatives. Deux techniques sont à disposition : le krigeage avec dérive externe, ou le *cokrigeage* colocalisé [4]. Ces deux techniques mettent en œuvre deux modèles différents de représentation de la réalité. Le krigeage avec dérive externe considère les co-facteurs comme autant de paramètres qui représentent la tendance à grande échelle (ou dérive) du phénomène de pollution. Il s'agit donc d'une approche non stationnaire, à la différence du *cokrigeage* colocalisé qui considère l'ensemble des variables, polluant et co-facteurs, sur le même plan et utilise leurs corrélations à toute échelle. Quel que soit le point de vue adopté et la technique qui en découle, on réalise ainsi une intégration de quelques mesures, peu nombreuses, et d'une information exhaustive des paramètres qui, indirectement, influent sur la

concentration en polluant. Le résultat est que l'on obtient une carte plus réaliste dans la mesure où elle porte l'empreinte des co-facteurs. Dans l'exemple de la cartographie de la moyenne annuelle du NO_2 à Rouen, on a utilisé la carte des émissions et de la densité de population. Une combinaison de ces deux facteurs est montrée sur la *figure 4*.

La carte du NO_2 interpolé par *cokrigeage* colocalisé avec ce co-facteur *(figure 5)*, reste sans biais, c'est-à-dire qu'en moyenne il y a autant de points estimés par défaut que par excès. Elle se distingue de la précédente, lorsque à la fois le réseau des sites de mesure est plus lâche et que la trace des infrastructures routières est apparente.

Concernant la variance de l'erreur, la modification par rapport à la carte précédente n'est pas spectaculaire, ce qui explique qu'elle ne soit pas montrée dans cet article. On note toutefois que la variance est sensiblement inférieure lorsque l'on est à proximité d'un site de mesure mais qu'elle reste du même ordre de grandeur en extrapolation.

Analyse du risque

Le krigeage et ses dérivés fournissent ce que l'on peut appeler, par abus de langage, la valeur la plus probable de la concentration en polluant en tout point de l'espace, associée à la variance de l'erreur commise. Cela a deux conséquences. La première est que la carte gomme les « pics » et les « creux » de pollution et est « attirée » vers la moyenne de la pollution sur la zone d'intérêt. C'est la propriété de lissage du krigeage : la variabilité réelle dans l'espace de la pollution n'est pas reproduite quand on interpole les données.

La deuxième conséquence est que l'on n'a pas accès à la distribution complète de l'erreur ; on n'en connaît que la moyenne (nulle par construction) et la variance.

Or les questions qui se posent en termes de risque nécessitent de raisonner sur des modèles numériques qui, à la différence du krigeage, restituent toute la gamme des valeurs pouvant être atteintes par le polluant.

Exemples de questions pratiques

Nous ne reprenons pas ici le catalogue complet des critères de surveillance de la qualité de l'air définis dans les législations nationale et communautaire. De façon synthétique, nous pouvons classer ces critères en plusieurs catégories. D'abord, il y a ceux qui correspondent à des seuils sur la concentration de tel et tel polluant. Toutefois, pour un polluant donné, il n'y a pas un mais plusieurs seuils selon des bases de calcul différentes : moyenne horaire, journalière ou annuelle.

Par rapport à ces seuils, on cherchera à déterminer leur probabilité de dépassement compte tenu de la méthode d'estimation choisie.

Il y a ensuite ce qui concerne l'évaluation du risque qu'il y a pour la population d'être exposée à un niveau donné de pollution. Par exemple, quel est le nombre de personnes qui ont une probabilité de 95 % d'être exposées à une pollution au dioxyde d'azote supérieure à l'objectif de qualité, soit 40 µg/m^3 ? Il est clair que les solutions aux problèmes liés aux dépassements forment un élément de réponse, mais il faut également préciser comment on va définir la population [5].

Il y a enfin des critères qui lient une valeur limite au nombre d'occurrences dans une période donnée. Par exemple, la

Figure 4. Carte du logarithme de la somme des émissions et de la densité de population.

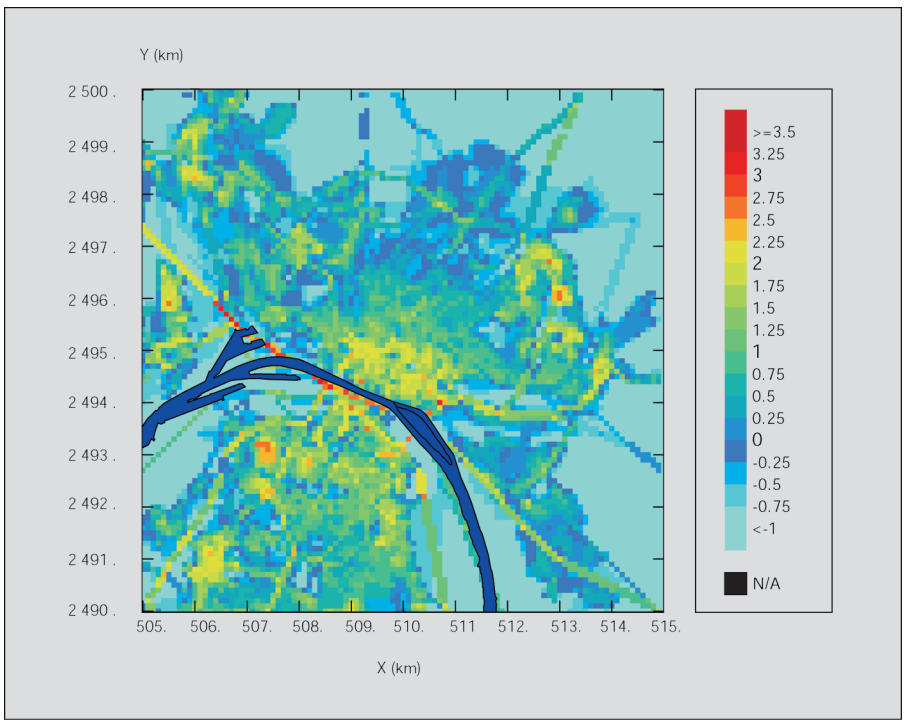
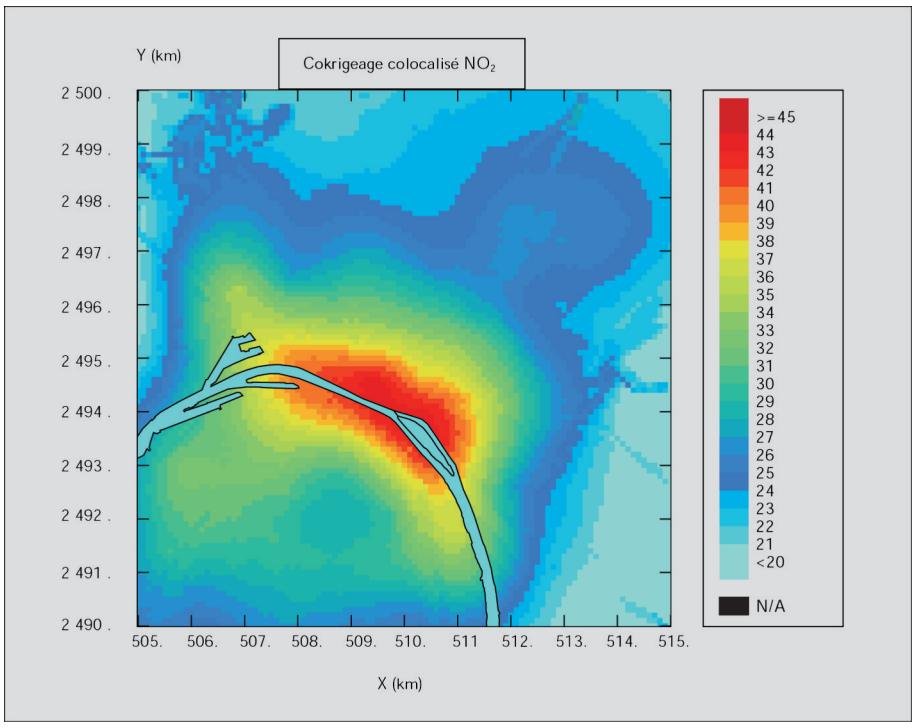


Figure 5. Carte du *cokrigeage* du NO_2 avec les émissions et la densite de population comme co-facteurs.



concentration quotidienne en PM_{10} ne doit pas dépasser 50 µg/m^3 plus de 35 jours par an. On voudra ainsi calculer le percentile pour la fréquence correspondant au nombre d'occurrences admises et le comparer au seuil de concentration. Dans l'exemple donné, on comparera le percentile 90,4 % (100 × (1-35/365)) à 50 µg/m^3.

Réponses simples mais erronées

Comme il a été dit dans l'introduction à ce chapitre, le krigeage ne fournit pas une réponse correcte. Ainsi, rien ne permet d'affirmer que la variance de krigeage est suffisante pour caractériser l'intervalle de confiance de l'estimation. Il n'est pas correct, par exemple, d'affirmer qu'il y a une probabilité de 97,5 % que la pollution soit supérieure au krigeage moins 2 fois l'écart type de krigeage. Cela supposerait, en fait, que l'erreur de krigeage a une distribution gaussienne indépendante d'un point à un autre : or une telle hypothèse est loin d'être vérifiée dans la pratique.

Il faut utiliser une représentation du phénomène qui possède les mêmes caractéristiques de variabilité spatiale que la réalité, donc le même variogramme. Il s'ensuit que les techniques d'analyse de risques fondées sur des simulations de type Monte Carlo, par tirage selon une loi *a priori* et indépendamment d'un point à un autre, ne sont pas adaptées. En effet, la valeur possible en un point est corrélée à la valeur en un point voisin. Il faut donc se tourner vers des méthodes moins simplificatrices : ce sont les méthodes géostatistiques non linéaires et en particulier les simulations conditionnelles.

Simulations géostatistiques

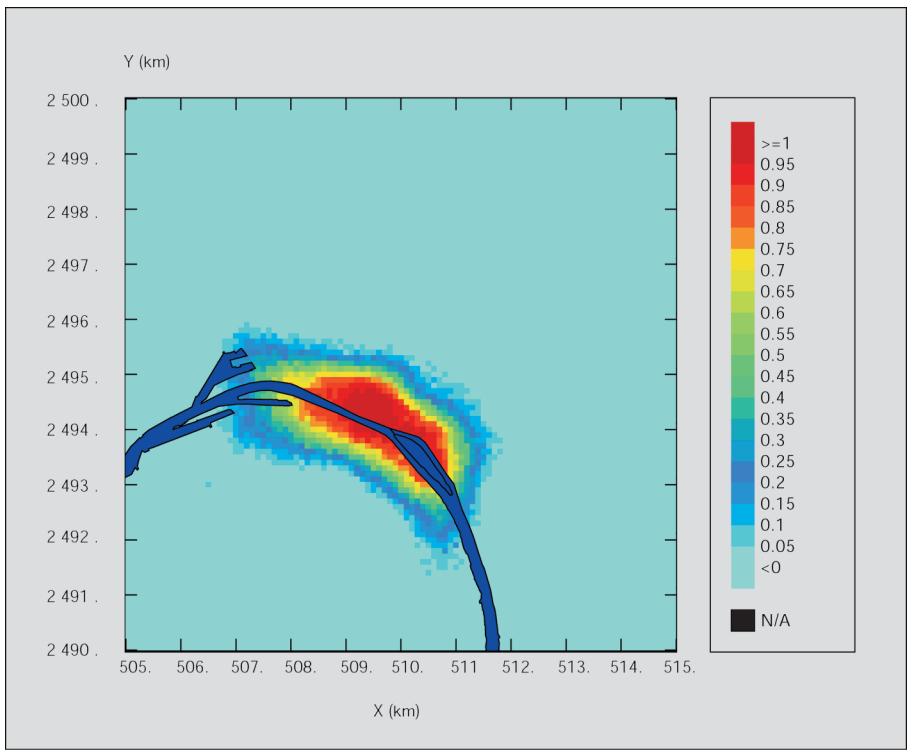
La géostatistique fournit une réponse tout à fait cohérente aux questions qui font appel à des opérateurs non linéaires comme l'application d'un seuil. Parmi les méthodes utilisables, les simulations sont les plus souples. Elles donnent la possibilité de simuler en chaque point de l'espace une réalisation, ou valeur possible, de la variable étudiée de telle façon que les caractéristiques de la variabilité spatiale, prises en charge par le variogramme, soient reproduites. La simulation est, par ailleurs, dite conditionnelle car elle est calée aux données et retrouve les valeurs des mesures en ces points. Tout l'intérêt consiste à calculer un grand nombre de simulations (autres représentations possibles du phénomène, compte tenu de ce qu'on en sait), permettant ainsi de faire des raisonnements en probabilité. En chaque point du maillage, on a ainsi un histogramme des valeurs possibles, dont la moyenne converge vers le krigeage.

En appliquant en chaque point le seuil de concentration en NO_2 de 40 µg/m^3 sur 100 simulations, on a obtenu *(figure 6)* une carte de la probabilité de dépassement de ce seuil [6].

À partir de là, on peut établir des stratégies de surveillance du territoire et mieux dimensionner les réseaux de capteurs.

La traduction au niveau de la population concernée du risque d'exposition nécessite de prendre quelques précautions. Le premier point consiste à définir la population exposée, ce qui est en soi une question difficile vu la mobilité quotidienne de la population. Ce n'est pas l'objet de cet article que de répondre à cette question ; aussi nous contenterons-nous, modestement, de considérer la population au sens du recensement (nous avons conscience que cela ne clôt pas le débat).

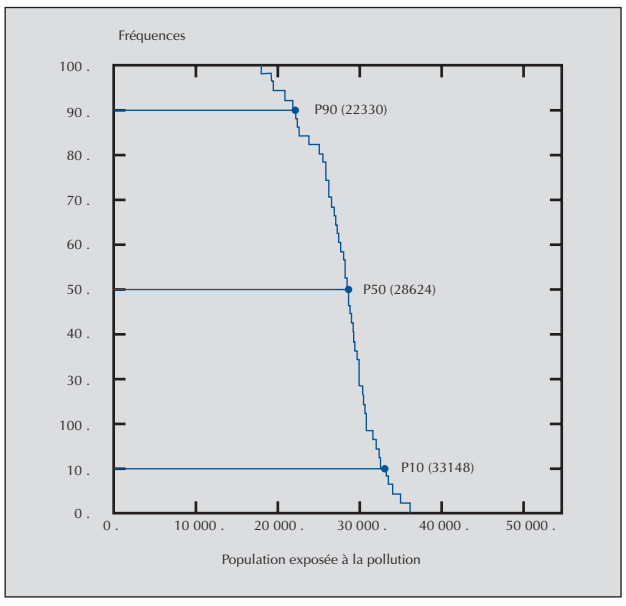
Figure 6. Carte de probabilité que le NO_2 dépasse le seuil de 40 µg/m^3.



Pour chaque simulation, on calcule la population exposée en sommant les habitants pour les points où la valeur simulée est supérieure au seuil. On obtient ainsi grâce aux 100 simulations un histogramme de la population exposée *(figure 7)*.

Sur cet histogramme, on peut donc appliquer un niveau de risque et déterminer ainsi que 6,3 % de la population (22 300 personnes) a 90 % de probabilité d'être exposée à une pollution supérieure à 40 µg/m^3. Le même calcul effectué en utilisant le krigeage aurait sous-estimé d'un facteur 2 la population exposée (sous-estimation due à l'effet de lissage).

Figure 7. Histogramme de la population exposée à une pollution annuelle moyenne supérieure à 40 µg/m^3.



Conclusion

Peu après la publication de la loi sur l'air du 30 décembre 1996, l'émergence de l'utilisation des tubes à diffusion a contraint les organismes de surveillance de la qualité de l'air à se pencher sur les questions de représentation des résultats. Le nombre élevé de points de mesure a conduit à utiliser des méthodes d'interpolation pour présenter des cartes. Les connaissances de l'époque ne permettaient que l'utilisation de méthodes purement déterministes comme les *splines* ou l'inverse des distances (au carré). Le problème de la subjectivité de la formulation mathématique a alors été soulevé et a conduit naturellement à s'intéresser aux techniques géostatistiques.

La multiplication récente des possibilités spatiales de mesure a renforcé l'intérêt pour la géostatistique jusqu'alors très peu utilisée dans le domaine de la pollution atmosphérique.

Outre que le krigeage est la méthode optimale d'interpolation linéaire, il présente de nombreux avantages : l'incertitude de l'interpolation peut être évaluée, les modèles de variogramme peuvent être anisotropes, des informations complémentaires peuvent être intégrées, etc. Par ailleurs, si le krigeage ordinaire suffit largement à l'établissement de cartes présentables au public, l'analyse du risque nécessite un peu plus d'attention. En

effet, l'application la plus directe dans le domaine de la surveillance de la qualité de l'air, serait de couper simplement une carte krigée à un seuil réglementaire afin d'en déduire les zones de dépassement, ou encore de croiser cette même carte avec des données telles que la répartition de la population pour en calculer l'exposition. Malheureusement, cela conduit à une sous-estimation. Une réflexion théorique sur le cadre probabiliste de la géostatistique conduit, dans un souci de rigueur et de cohérence, à proposer les simulations comme réponse au problème de l'analyse du risque.

Par ailleurs, la nécessité de disposer de nombreux points de mesure est une difficulté incontournable qui pose la question de l'optimisation du réseau de mesures, encore dans l'attente de réponses claires. Des recherches font l'objet d'axes de travail au Laboratoire central de surveillance de la qualité de l'air avec l'appui de Géovariances, société de services en géostatistique qui commercialise le logiciel ISATIS [7]. ■

Références

1. Arnaud M, Emery X. *Estimation et interpolation spatiale*. Paris : Hermès Science, 2000 ; 217 p.

2. Chilès JP, Delfiner P. *Geostatistics Modeling Spatial Uncertainty*. New York : Wiley & Sons, 1999 ; 695 p.

3. Matheron G. *Estimer et choisir*. Fascicule 7. Paris : Centre de Géostatistique ; École des Mines de Paris, 1978 ; 175 p.

4. Bobbia M, Pernelet V, Roth C. L'intégration des informations indirectes à la cartographie géostatistique des polluants. *Poll Atmos* 2001 ; 170 : 251-62.

5. Deraisme J, Jaquet O, Jeannée N. Uncertainty management for environmental risk assessment using geostatistical simulations. In : *GeoENV IV – Geostatistics for Environmental Applications*. Barcelone : Kluwer Academic, 2002 (à paraître).

6. Bobbia M. *Investigations en vue d'une approche géostatistique de la qualité de l'air - Analyse du risque : les simulations conditionnelles*. Rouen : Air Normand, 2001 ; 48 p.

7. Bleines B, Deraisme J, Geffroy F, *et al*. *Isatis software Manual*. Paris : Géovariances, 2002 ; 579 p.