Méthodes de représentation de la qualité de l'air

Guide d'utilisation des méthodes de la géostatistique linéaire

Laboratoire Central de Surveillance de la Qualité de l'Air

Convention 115/03

Laure MALHERBE, Laurence ROUÏL,
Unité Modélisation et Analyse Economique pour la gestion des Risques (MECO)

Direction des Risques Chroniques (DRC)

Décembre 2003

Méthodes de représentation de la qualité de l'air

Guide d'utilisation des méthodes de la géostatistique linéaire

Laboratoire Central de Surveillance de la Qualité de l'Air

Convention 115/03

financée par la Direction de la Prévention des Pollutions et des Risques (DPPR)

Décembre 2003

LAURE MALHERBE
LAURENCE ROUÏL

Personne ayant participé à l'étude : Giovanni CARDENAS

Ce document comporte 70 pages (hors couverture et annexes).

	Rédaction	Vérification	Approbation
NOM	Laure MALHERBE Laurence ROUÏL	Laurence ROUÏL	Martine RAMEL
Qualité	Ingénieurs Etudes et Recherches Direction des Risques Chroniques	Ingénieur Etudes et Recherches Direction des Risques Chroniques	Coordination LCSQA Direction des Risques Chroniques
Visa			

1. FORMULAIRE

La modélisation géostatistique repose sur les notions fondamentales de fonction aléatoire et de loi spatiale. Pour en faciliter la compréhension, un certain nombre de définitions s'impose.

Dans toute la suite, D désigne le domaine étudié et s les points de ce domaine.

Variable aléatoire : variable dont les réalisations possibles sont connues mais dont le résultat final ne peut être déterminé avant d'effectuer la mesure.

En un point s, la variable aléatoire considérée est notée Z(s).

z(s), valeur observée, est une réalisation de Z(s).

Fonction aléatoire : famille de toutes les variables aléatoires $\{Z(s) ; s \in D\}$. Elle est définie dans tout le domaine, y compris aux points où il n'y a pas de mesure disponible.

Variable régionalisée : réalisation $\{z(s) ; s \in D\}$ d'une fonction aléatoire.

Champ : domaine de l'espace où la variable régionalisée prend ses valeurs.

Une fonction aléatoire $\{Z(s) ; s \in D\}$ est caractérisée par sa **loi spatiale**.

Loi spatiale : donnée de toutes les lois de probabilité qui régissent les vecteurs aléatoires $\{Z(s_1)\ldots Z(s_n)\}$ que l'on peut extraire de $\{Z(s) ; s \in D\}$. En pratique la description de la loi spatiale se limite à ses deux premiers moments : moyenne et covariance.

Moyenne (ou espérance mathématique) : $E[Z(s)]=m(s)$

L'espérance est la moyenne autour de laquelle les valeurs possibles de Z(s) se distribuent.

Variance :

$Var[Z(s)]=E\{[Z(s)-m(s)]^2\}= E\{[Z(s)]^2\}-[m(s)]^2$

La variance mesure la dispersion de Z(s) autour de sa moyenne m(s). Elle est égale à la covariance de Z(s) avec elle-même.

Covariance :

Soient s_1 et s_2, deux points de D

$Cov[Z(s_1),Z(s_2)]=E\{[Z(s_1)-m(s_1)]\ [Z(s_2)-m(s_2)]\}= E[Z(s_1)Z(s_2)]-[m(s_1)][m(s_2)]$

La covariance quantifie le lien linéaire entre les valeurs prises en s_1 et s_2.

Corrélation spatiale *:*

$$\rho_{12} = \frac{\mathrm{cov}(Z(s_1), Z(s_2))}{\sqrt{Var(Z(s_1))}\sqrt{Z(s_2)}}$$

Elle a le même sens que la covariance mais elle est normalisée.

2. RESUME

L'élaboration de cartes de répartition des concentrations des polluants réglementés sur des bases temporelles appropriées est un bon moyen de répondre aux objectifs de surveillance de la qualité de l'air définis dans les textes réglementaires nationaux et européens. Ces cartes sont relativement aisées à construire dans les zones de petite surface, où la densité de points de mesure est suffisante (sites urbains et péri-urbains), en appliquant par exemple des règles simples d'interpolation. En revanche, là où peu de stations de mesure sont implantées, cartographier la qualité de l'air nécessite l'usage de techniques de traitement complémentaires et plus sophistiquées, souvent regroupées sous le terme de « modélisation » ou d'analyse objective. Elles regroupent :

- Des méthodes d'interpolation classiques ;
- Des approches statistiques ou géostatistiques ;
- Des approches purement déterministes.

Les deux premières catégories de méthodes reposent plus naturellement sur l'usage des données de concentrations mesurées. Elles correspondent à une vision stochastique des phénomènes qui induit des simplifications parfois abusives (mauvaise représentation des effets locaux ou d'une météorologie très complexe par exemple). Les méthodes déterministes incorporent les données « annexes » comme données d'entrée et calculent les concentrations par la résolution numérique d'équations aux dérivées partielles. Cette approche engendre également des simplifications dues au formalisme mathématique des équations choisies et des processus de calcul de moyenne dans le temps et l'espace induits par la discrétisation spatiale et temporelle du problème.

Opérer une distinction stricte entre ces catégories de méthodes n'est donc pas judicieux par rapport à la problématique de l'évaluation de la qualité de l'air, qui loin d'être une discipline complètement déterminée, répond tout de même à certaines règles.

Les approches les plus pertinentes résident donc dans les développements mixtes incluant toutes les informations disponibles (concentrations mesurées et données annexes). Celles-ci sont en plein essor actuellement.

Cette étude a pour objet de fournir aux AASQA un bilan sur les méthodes de modélisation disponibles pour réaliser des cartographies de la qualité de l'air. Les méthodes déterministes ayant été largement étudiées au cours des années antérieures (Rouïl et Wroblewski, 2002, Rouïl, 2001), l'essentiel du rapport est consacré aux méthodes d'interpolation et plus particulièrement aux techniques de la géostatistique linéaire.

Il propose une analyse de ces techniques, de leurs conditions de mise en œuvre, de la démarche à suivre pour en faire un usage intelligent et efficace, et de leurs limites. Des éléments sur les données d'entrée nécessaires sont également fournis.

3. INTRODUCTION

3.1 CONTEXTE ET OBJECTIFS

La Directive Cadre 96/62/EC sur la qualité de l'air ambiant, et les Directives filles qui en découlent (1999/30/EC, 2000/69/EC et 2002/3/CE) fixent un cadre réglementaire de surveillance des polluants atmosphériques. Outre la définition de valeurs cibles de concentrations, de seuils d'alerte et éventuellement de seuils d'information, ces textes prévoient l'évaluation de la qualité de l'air dans les états membres. L'obligation de faire un rapport de la situation observée dans chaque pays découpé en zones géographiques relève de cette mission. De même une information quotidienne, voire horaire (pour les oxydes d'azote, le dioxyde de soufre et l'ozone) doit être fournie au public et aux autorités compétentes.

L'élaboration de cartes de répartition des concentrations des polluants réglementés sur des bases temporelles appropriées est un bon moyen de répondre à ces objectifs. Ces cartes sont relativement aisées à construire dans les zones de petite surface, où la densité de points de mesure est suffisante (sites urbains et péri-urbains), en appliquant par exemple des règles simples d'interpolation. En revanche, là où peu de stations de mesure sont implantées, cartographier la qualité de l'air nécessite l'usage de techniques de traitement complémentaires et plus sophistiquées, souvent regroupées sous le terme de « modélisation » ou d'analyse objective.

L'application de méthodes numériques est précisément considérée dans les textes réglementaires (excepté pour l'ozone pour lequel des seuils d'évaluation ne sont pas définis) :

- Elles peuvent être utilisées en complément des mesures disponibles sur le réseau de stations fixes dans les zones où les concentrations varient entre un seuil d'évaluation inférieur et un seuil d'évaluation supérieur fixés pour chaque polluant.
- La qualité de l'air peut être évaluée à l'aide des seules techniques de modélisation, dans les zones où les concentrations sont toujours plus faibles que le seuil d'évaluation inférieur.

Suivant les polluants et suivant les indicateurs recherchés (moyenne horaire, sur 8 heures ou journalière) l'erreur commise en simulant les concentrations plutôt qu'en les mesurant doit varier entre 30 et 50%.

Cependant aucune recommandation sur le type de démarche et d'outils applicables afin de fournir l'information recherchée sur la qualité de l'air n'est disponible dans les directives.

En fait, les méthodes de modélisation s'appuient nécessairement sur les sources d'information disponibles, qui se répartissent en trois grandes classes :

- Les données de mesure fournies par un réseau de stations fixes dont le nombre est évidemment limité par des considérations de coût et de maintenance,
- Les données issues de campagnes de mesure spécifiques (tubes à diffusion passive, moyens de mesure mobiles), limitées dans le temps et dans l'espace. Le nombre de points de prélèvement avec les tubes passifs peut être important afin de mieux cerner l'information obtenue par les stations fixes,
- L'ensemble des données qui ne sont pas des mesures de concentration mais qui contribuent à expliquer la présence de tel ou tel polluant. Il s'agit des données d'activité et d'émission, des caractéristiques de site (bâti, couverture végétale, relief), et des données météorologiques.

Ces informations permettent de reconstituer des cartes de concentrations à l'aide de techniques mathématiques qui relèvent :

- Des méthodes d'interpolation classiques,
- Des approches statistiques ou géostatistiques
- Des approches purement déterministes

Les deux premières catégories de méthodes reposent plus naturellement sur l'usage des données de concentrations mesurées. Elles correspondent à une vision stochastique des phénomènes qui induit des simplifications parfois abusives (mauvaise représentation des effets locaux ou d'une météorologie très complexe par exemple). Les méthodes déterministes incorporent les données « annexes » comme données d'entrée et calculent les concentrations par la résolution numérique d'équations aux dérivées partielles. Cette approche entraîne également des simplifications dues au formalisme mathématique des équations choisies et des processus de calcul de moyenne dans le temps et l'espace engendrés par la discrétisation spatiale et temporelle du problème.

Considérer de façon indépendante ces catégories de méthodes n'est donc pas judicieux par rapport à la problématique de l'évaluation de la qualité de l'air, qui, sans être une discipline complètement déterminée, répond tout de même à certaines règles.

Les approches les plus pertinentes résident donc dans les développements mixtes incluant toutes les informations disponibles (concentrations mesurées et données annexes). Celles-ci sont en plein essor actuellement.

3.2 OBJET DU DOCUMENT

Ce guide a pour but de fournir aux AASQA un bilan sur les méthodes de modélisation disponibles pour réaliser des cartographies de la qualité de l'air. L'usage des modèles déterministes ayant fait l'objet de multiples travaux au sein du LCSQA (Rouïl et Wroblewski, 2002, Rouïl, 2002, Wroblewski, 2003), l'essentiel du document est consacré aux techniques d'interpolation et plus particulièrement à l'application de la géostatistique.

Une première partie recense les différentes approches possibles pour cartographier la qualité de l'air et en fournit une description très générale. Les perspectives ouvertes par le couplage des méthodes déterministes et géostatistiques sont évoquées à cette occasion.

Une analyse des techniques géostatistiques et des conditions de leur mise en œuvre est ensuite proposée (partie II). Elle s'appuie sur une synthèse bibliographique et sur les expériences des AASQA ou des centres de recherche spécialisés dans ce domaine. Elle intègre en particulier les résultats d'un travail effectué par le Centre de Géostatistique de l'Ecole des Mines de Paris sur la cartographie du NO_2 dans deux agglomérations et au voisinage d'une route (convention INERIS/ARMINES). Elle exploite également les conclusions d'une étude sur l'évaluation des incertitudes associées aux méthodes géostatistiques (étude LCSQA-INERIS, 2003).

Un chapitre complémentaire consacré aux données d'entrée (mesures et informations auxiliaires) fournit des recommandations sur la nature de l'information qu'il est nécessaire de recueillir pour utiliser efficacement ces méthodes.

En conclusion de ce rapport, des préconisations générales sur l'usage des modèles pour la cartographie de la qualité de l'air sont émises.

Certains aspects, tels l'échantillonnage et l'extrapolation dans le temps, n'ont été que partiellement abordés. La version actuelle de ce guide pourra être mise à jour en 2004, en tenant compte des remarques des AASQA et des résultats d'études consacrées au problème temporel (LCSQA, INERIS et EMD, 2004).

PREMIÈRE PARTIE : LES DIFFÉRENTS OUTILS DE CARTOGRAPHIE

1. L'INTERPOLATION

L'interpolation a pour but d'obtenir une information spatiale continue sur la qualité de l'air à partir de mesures ponctuelles de concentrations. Elle consiste à mettre en œuvre des algorithmes mathématiques ou probabilistes afin d'estimer les concentrations de polluants entre les points d'échantillonnage.

Il existe de nombreuses méthodes d'interpolation, parmi lesquelles il peut sembler délicat de faire un choix.

Nous distinguerons deux catégories de méthodes :

- Les méthodes d'interpolation classiques, qui utilisent des algorithmes purement mathématiques. Ne sont pas inclus dans cette catégorie les modèles physico-chimiques de la dispersion des polluants qui ne sont pas des méthodes d'interpolation.
- Les méthodes d'estimation géostatistiques**,** qui s'appuient sur une modélisation probabiliste du phénomène étudié. Ainsi elles sont fondées sur la reconstitution du phénomène à l'aide d'une fonction relationnelle établie sur des analyses statistiques.

Elles s'appliquent à des ***variables régionalisées***, c'est-à-dire à des fonctions numériques qui prennent leurs valeurs dans des régions délimitées de l'espace appelées *champs*. Dans notre cas, les variables régionalisées étudiées sont les concentrations de polluants.

1.1 LES METHODES D'INTERPOLATION CLASSIQUES

Ces méthodes se divisent en deux groupes, selon les données qu'elles requièrent.

1.1.1 Les méthodes qui traitent uniquement les données de la variable étudiée

Cette classe regroupe un grand nombre de techniques. ***La plupart définissent la valeur recherchée en un point comme une combinaison linéaire pondérée des mesures disponibles.*** Ce sont des méthodes utilisées dans de nombreux domaines de la physique, accessibles et implantées dans les logiciels de représentation graphique les plus connus.

Parmi les approches les plus usitées il faut noter :

- Les méthodes par partitionnement de l'espace, basées sur la définition de zones d'influence susceptibles de participer au calcul de la concentration au point cible (méthode des polygones de Thienssen par exemple). Seules les mesures les plus influentes sont donc incorporées.
- Les méthodes barycentriques qui considèrent un nombre plus important de données et qui, dans la formule d'interpolation, affectent un poids plus représentatif à celles qui sont les plus proches du point cible : méthode d'interpolation par l'inverse des distances ou par l'inverse des carrés des distances par exemple.
- Les méthodes de régression polynomiale qui recherchent une surface définie par une équation polynomiale, s'ajustant au mieux, au sens des moindres carrés, aux points de mesure disponibles.

La mise en œuvre de ces méthodes ne nécessite pas d'effort de modélisation et permet de réaliser sans difficulté une cartographie de la variable étudiée.

De tels exemples se trouvent en Belgique (www.irceline.be/~celinair), aux Pays-Bas (www.lml.rivm.nl) ou encore en Suède (www.ivl.se), pays qui présentent des réseaux de mesure relativement denses. Varns et al. (2001) interpolent par la méthode de l'inverse des distances les concentrations d'ozone entre 30 points de mesure situés dans l'agglomération de Dallas (mesures par tubes passifs, superficie de la zone d'étude : 24 000 km^2).

Ces techniques présentent néanmoins deux défauts majeurs :

- Elles ignorent la structure spatiale de la variable, en offrant des contours très lisses d'interpolation (effet « œil de bœuf » par exemple) et peuvent omettre de représenter des situations locales très spécifiques, d'où le risque d'aboutir à des cartes peu réalistes,
- Aucun critère permettant de juger de la précision de ces cartes n'est formulé.

Les fonctions *splines* peuvent être également employées pour l'estimation locale. L'interpolation par spline consiste à ajuster aux données une surface de courbure minimale. Cette méthode produit également des surfaces interpolées lisses. C'est pourquoi elle s'applique de préférence aux variables qui évoluent lentement dans l'espace. Si en revanche les données présentent de brusques variations, elle est inappropriée.

Coyle et al. (2002) utilisent ainsi un algorithme d'interpolation qui minimise la courbure pour estimer les concentrations journalières ([0_3] entre 12h et18h) moyennes d'ozone entre vingt stations de mesure réparties dans tout le Royaume-Uni.

La méthode des *splines* n'est pas plus détaillée, son formalisme équivalant celui du krigeage ; ce dernier a l'avantage de permettre le choix d'un modèle cohérent avec la structure spatiale du phénomène étudié (Arnaud et Emery, 2000).

1.1.1 Les méthodes qui nécessitent l'usage de variables secondaires

La concentration d'un polluant en un point est déduite d'une relation mathématique avec une ou plusieurs variables explicatives déterminées, relation établie généralement par régression multilinéaire.

Deletraz et Dabos (1999) ont recours à cette technique afin de cartographier les dépôts secs de NO_2 à proximité d'un axe routier. Cette méthode a servi aussi à exploiter les données du programme INTERREG II mené sur le Fossé Rhénan sous la coordination française de l'ASPA (2000) et à réaliser des cartographies thématiques. Une relation est établie entre la variable de concentration et chaque variable auxiliaire et permet de dresser une carte de concentration pour chacune de ces variables. Les variables auxiliaires sont classiquement puisées parmi les données d'émission, les modèles numériques de terrain ou l'occupation des sols. L'intégration de données satellitaires dans cette démarche tend à se développer (Ung et al, 2001, programme GMES du 6ieme PCRD).

La principale difficulté de l'interpolation par régression est la recherche de variables explicatives corrélées linéairement avec la variable étudiée et connues avec une résolution spatiale suffisante. Il peut s'agir de variables descriptives des sources (inventaire des émissions) ou du milieu (occupation du sol, altimétrie, population…), ou de fonctions mathématiques de ces variables. En outre, pour établir la régression sur un échantillon de données statistiquement important, vingt à trente points de mesure sont au minimum nécessaires. Si ces difficultés et ces contraintes peuvent être surmontées, l'emploi de méthodes de régression constitue une manière simple et efficace de réaliser une cartographie. On veillera cependant à ne pas appliquer les relations obtenues au-delà de leur domaine de validité (i.e. pour des valeurs de variables explicatives auxquelles ne correspondait aucune donnée de concentration). Dans une étude comparative consistant à interpoler les concentrations moyennes saisonnières d'ozone par la méthode du plus proche voisin, des moindres carrés, du krigeage et par régression, Nikiforov et al. (1998) montrent la bonne performance de cette dernière technique.

La cartographie par régression peut être en fait considérée comme un cas particulier du krigeage avec dérive externe, dans l'hypothèse où les résidus de la régression ne sont pas spatialement corrélés, et à ceci près qu'il ne s'agit pas d'un interpolateur exact. (Pour une description du krigeage avec dérive externe, voir le chapitre 1.2.). Si toutefois cette hypothèse n'est pas vérifiée, les résidus contiennent alors une information sur la variabilité spatiale du polluant. Se limiter à la régression conduit à négliger cette information. Afin d'en tenir compte, il est souhaitable de faire appel à la géostatistique, en complétant la régression par un krigeage des résidus ou en effectuant directement un krigeage avec dérive externe.

Les méthodologies officielles de cartographie de la qualité de l'air adoptées au Royaume Uni font largement appel à cette approche, notamment pour les polluants principalement influencés par des sources locales (trafic routier) tels que le benzène ou les oxydes d'azote (Stedman 1998). La méthode de régression adoptée dans ce contexte sera plus largement étudiée par l'Ecole des Mines de Douai dans son programme LCSQA de 2004.

1.2 LES METHODES D'ESTIMATION GEOSTATISTIQUES

Les techniques de la géostatistique linéaire consistent à estimer les concentrations par une combinaison linéaire des données expérimentales mais, à la différence des méthodes classiques d'interpolation, elles tiennent compte à la fois du caractère aléatoire du phénomène considéré et de l'existence d'une certaine structure spatiale. Cette propriété constitue un avantage majeur de ces techniques qui, en outre, permettent d'intégrer des informations auxiliaires dans l'estimation.

La mise en œuvre du krigeage passe ainsi par une étape d'analyse des données et de modélisation, destinée à décrire et à représenter mathématiquement la structure spatiale de la variable de pollution. La carte de concentration qui en résulte s'accompagne d'une carte de variance de l'erreur d'estimation, qui, si elle a surtout une valeur qualitative, fournit une indication intéressante sur la précision de l'estimation.

Ces notions et l'application des différentes méthodes de la géostatistique sont largement développées dans la seconde partie du document.

2. LES MODELES DETERMINISTES

2.1 DESCRIPTION DES METHODES

Les méthodes d'interpolation précédemment décrites sont basées sur le traitement mathématique de données de concentrations mesurées dans l'air ambiant. ***Les méthodes déterministes s'appuient sur l'analyse et la simulation mathématique, en fonction du site et de la météorologie, de la transformation chimique et du transport d'émissions polluantes d'origine anthropogénique et naturelle.***

Ces modèles reposent sur la résolution numérique des équations aux dérivées partielles qui régissent le comportement physico-chimique des polluants. On distingue différentes approches pour aborder cette question (Rouïl et Wroblewski, 2001). Parmi les plus courantes :

- Les méthodes analytiques qui proposent une solution exacte d'un problème simplifié. Dans le domaine de la dispersion atmosphérique les approches basées sur la représentation gaussienne du panache de polluant sont parmi les plus utilisées.
- Les méthodes lagrangiennes considèrent le panache de polluants comme une infinité de particules, l'évolution de chacune d'entre elle étant suivie dans un référentiel qui lui est propre. Les concentrations sont déduites par sommation du nombre de particules à un instant donné, dans un volume donné.
- Les méthodes eulériennes considèrent un référentiel fixe sur lequel est bâti un maillage tridimensionnel. Les concentrations en chaque nœud de cette grille sont alors calculées par intégration numérique des équations, via des processus itératifs souvent très élaborés.

Quelle que soit la démarche retenue, lorsque les composants étudiés sont considérés comme inertes chimiquement, seule la partie transport est traitée (équation de transport-diffusion). En revanche le traitement des polluants secondaires issus de la transformation chimique de composés primaires (ozone, aérosols), est assuré par un module spécifique, qui gère un nombre souvent réduit de réactions chimiques (par souci de simplicité et de performance numérique) agissant sur des espèces modèles. En effet la chimie de l'ozone met en jeu des centaines de Composés Organiques Volatils (COV) qu'il est illusoire de prendre en compte de manière individuelle. Aussi est-il d'usage de « réduire » les mécanismes chimiques en raisonnant sur un nombre limité de classes de composés jugées représentatives (espèces modèles) et de réactions. Les phénomènes de dépôt sec et humide, déterminants dans le cas des particules, et de certains polluants gazeux (ozone, dioxyde de soufre) sont également modélisés par des modules spécifiques.

2.2 L'ECHELLE SPATIALE

Ces méthodes sont par nature bien adaptées à la représentation cartographique de la qualité de l'air. En effet elles permettent de calculer les concentrations en chaque point d'une grille tridimensionnelle entièrement définie par l'utilisateur.

Différentes hypothèses de modélisation sont applicables suivant l'échelle considérée, mais il n'y a pas de contraintes strictes sur l'étendue du domaine de calcul : il pourra aussi bien se limiter à la rue (à condition que le modèle puisse reproduire les phénomènes de turbulence caractérisant cette échelle), que s'étendre au niveau du continent (ce qui autorise quelques simplifications).

Il s'agit certainement de l'un des principaux atouts de la modélisation déterministe. Cette technique reposant sur une conceptualisation mathématique des phénomènes physiques, elle est théoriquement adaptée là où peu de données de mesures sont disponibles - typiquement en zone rurale -, pourvu que les données d'entrée existent.

De plus si cela est cohérent avec la précision des données d'entrée et s'il n'existe pas de contraintes de temps de calcul ou de stockage informatique (ce qui est parfaitement illusoire en pratique) il est théoriquement possible de représenter les phénomènes sur des mailles de petite taille et de les cartographier avec une bonne précision.

Cependant la part d'aléatoire propre aux phénomènes de dispersion atmosphérique ne peut être ainsi représentée, ce qui induit par défaut une incertitude naturelle dans les résultats. On parle souvent d'incertitude inhérente aux phénomènes.

2.3 L'ECHELLE TEMPORELLE

La cartographie de la qualité de l'air comme outil d'aide à la surveillance implique de disposer d'informations sur des bases temporelles relativement longues, de l'ordre de la saison ou de l'année. Cela pose un certain nombre de difficultés par rapport à l'usage de certains modèles déterministes. En effet, le temps nécessaire à la réalisation des calculs constitue indéniablement une limitation à leur usage en conditions opérationnelles.

Cependant, cette constatation ne vaut pas pour les modèles analytiques de type gaussien qui reposent sur le calcul en tous points et pour toute échéance temporelle des concentrations par une simple formule mathématique. Le domaine de validité de ces méthodes est cependant assez limité (de quelques centaines de mètres à une dizaine de kilomètres de la source), et leur mise en œuvre se cantonne à des problèmes spécifiques tels que l'évaluation de l'impact des rejets d'une installation industrielle ou d'un noyau urbain. Les modèles tridimensionnels adaptables à tous types de situation spatiale, sont en revanche nettement plus coûteux à utiliser. D'immenses efforts se concentrent aujourd'hui afin de réduire notablement les coûts en temps de calcul de ces méthodes.

Le modèle CHIMERE, développé par l'IPSL a été conçu sur la base de simplifications importantes (Schmidt et al 2001) qui permettent de réduire considérablement les coûts en temps de calcul par rapport à d'autres modèles de la même gamme. Par ailleurs le développement de versions parallélisées de codes de chimie transport des polluants atmosphériques tend à se développer (Tremback et al, 2000, Wolke et al, 2001). Le principe est de permettre l'usage de ces modèles en répartissant différentes tâches sur des machines multi-processeurs, afin de diminuer les temps de calculs. Cependant il est encore admis que les coûts prohibitifs liés à l'intégration numérique des équations qui régissent ces phénomènes reste un facteur limitant pour la réalisation de simulations sur de longues périodes.

L'usage des modèles déterministes pour cartographier la qualité de l'air étant donc suspendu à ces contraintes, différentes solutions sont envisagées par les modélisateurs pour réaliser à l'aide de ces outils des simulations annuelles. Si le modèle est suffisamment optimisé et supporté par un système informatique performant, il est possible de tirer des indicateurs annuels par le traitement de résultats de simulations obtenus pour chaque jour et chaque heure de l'année. Il est par exemple possible de réaliser une année de simulation avec le modèle CHIMERE Continental en 2 jours de calcul.

D'autres méthodes s'appuient sur des analyses climatologiques et l'exploitation d'un nombre restreint de situations météorologiques, représentatives du site, auxquelles sont affectées des fréquences d'apparition (Moussiopoulos, 1998).

La question relative à la réalisation de simulations déterministes (à l'aide de modèles de type eulérien ou lagrangien) de la qualité de l'air sur de longue périodes reste néanmoins complètement d'actualité.

2.4 PRECISION DES MODELES DETERMINISTES ET DONNEES D'ENTREE

La qualité et l'incertitude des résultats fournis par les modèles déterministes sont dépendantes de plusieurs facteurs (Rouïl, 2001):

- La résolution du modèle, dans les directions verticale et horizontale. Cela concerne concrètement le nombre de couches utilisées pour stratifier la troposphère libre (zone dans laquelle évolue le panache de polluants). Des couches de faible épaisseur près du sol sont nécessaires pour appréhender correctement les mécanismes thermique et mécanique qui conditionnent le transport des polluants. Le pas de maillage dans les directions horizontales détermine la finesse avec laquelle l'impact des sources de pollution pourra être modélisé. Généralement plus le maillage est fin et plus les résultats sont précis à condition que cette précision soit cohérente avec celle des données d'entrée.
- La qualité des schémas et approximations numériques utilisés dans le modèle. Sur ce point, l'utilisateur du code non spécialiste des développements numériques, n'a que très peu de marge de manœuvre.
- La précision des données d'entrée.

En effet les modèles déterministes sont exigeants en terme de données d'entrée : caractéristiques de site, données météorologiques, données d'émission et conditions de concentration et de flux aux limites du domaine considéré sont indispensables pour mettre en œuvre ces méthodes. Si la résolution spatiale et temporelle de ces informations est l'un des facteurs conditionnant la qualité des résultats, il faut néanmoins en apprécier l'impact véritable, en regard du coût que représente leur acquisition, et de l'incertitude qui accompagne le traitement d'une information de plus en plus fine et donc dépendante d'un plus grand nombre de facteurs (localisation, clefs de répartition temporelle, régimes de fonctionnement des usines…). Ces considérations amènent à repositionner les exigences des modèles déterministes sur les données d'entrée par rapport aux objectifs.

2.5 EVALUATION ET VALIDATION

Les modèles déterministes s'appuient sur la résolution approchée d'équations mathématiques censées décrire la physique du problème. Comme cela est évoqué dans le paragraphe précédent, disposer d'une formulation satisfaisante et de données d'entrée adaptées ne suffit pourtant pas à assurer la qualité parfaite des résultats. En effet, le nombre de paramètres susceptibles d'influencer le résultat final est tel qu'il est indispensable de procéder à une phase de « calage » afin d'optimiser les valeurs qui leur sont affectées. Généralement cette étape s'effectue sur des épisodes de courte durée (quelques jours) pour lesquels l'on dispose de suffisamment de données de mesure validées. Elle est incontournable, et ne dispense en aucun cas le modélisateur de remettre régulièrement an cause les résultats qu'il obtient par rapport aux informations dont il dispose.

Cette constatation conduit petit à petit à privilégier le développement d'approche « hybrides » liant modélisation déterministe et exploitation de mesures ponctuelles afin d'affiner les diagnostics.

3. LES METHODES HYBRIDES

La caractéristique majeure des questions liées à la qualité de l'air, et qui la distingue d'autres disciplines, est que les phénomènes mis en jeu ne sont ni complètement déterminés ni totalement stochastiques. Une part d'aléatoire, représentée souvent par la notion de turbulence, régit l'ensemble de ces comportements.

Ainsi une vision déterministe du problème induit des simplifications dues au formalisme mathématique et à la manipulation de moyennes en temps et en espace. De même une vision purement statistique s'avère également restrictive par le manque de physique prise en compte dans la modélisation.

Ainsi l'on s'accorde de plus en plus à considérer que la solution réside certainement dans l'exploitation conjointe de ces méthodes, afin d'affiner les résultats issus de la modélisation de la qualité de l'air. Elles supposent principalement des traitements mathématiques complémentaires, appliqués aux méthodes déterministes, afin d'introduire dans la simulation physique et mathématique des phénomènes, les données mesurées (et supposées donc réelles), généralement des concentrations. L'idée est donc de mieux coller à la réalité du terrain qui comprend une part stochastique non considérée par les modèles déterministes.

Quelques exemples d'approches hybrides sont donnés ci-dessous :

- ***L'assimilation de données*** qualifie le principe de base qui vise à assurer une cohérence entre calcul déterministe et mesures. Très développées dans le domaine de la météorologie ou de l'océanographie, ces approches tendent désormais à largement se répandre pour la qualité de l'air. Plusieurs méthodologies sont proposées dans ce cadre. Globalement elles consistent à réduire, par une procédure spécifique mise en œuvre au cours du processus itératif propre au modèle déterministe, l'écart entre la valeur numérique calculée, et la valeur exacte de la concentration. Cette procédure donne lieu au calcul d'une valeur « analysée » supposée plus « juste », qui est introduite dans le calcul déterministe à la place de la valeur initialement calculée. Plusieurs philosophies sont développées :
    - L'analyse peut être « séquentielle ». Dans ce cas, à chaque fois qu'une mesure est disponible, une nouvelle valeur du champ analysé est calculée. Les méthodes d'interpolation optimale sont parmi les plus anciennes (1960). Elles se basent sur la décomposition de la variable analysée en combinaison linéaire de la valeur initialement calculée, et des erreurs, « correctement » pondérées. Ces poids sont déterminés afin de minimiser une mesure de la différence entre la valeur réelle et la valeur analysée. Les méthodes basées sur les filtres de Kalman entrent également dans cette catégorie (Van Loon et al, 1997). Très complexes à mettre en œuvre, elles permettent d'introduire une composante d'évolution temporelle au système. Ces méthodes sont généralement bien appropriées pour ajuster des prévisions effectuées par un modèle.

- L'autre méthode repose sur une formulation « variationnelle » du système (Le Dimet et al, 2002, Daescu et al, 2003). Dans ce cas, les résultats du modèle sont calculés sur une période donnée (par exemple 24h), durant laquelle les mesures disponibles sont également stockées. La fonction mesurant la différence entre mesure et calcul est construite pour cette période. Le problème d'assimilation revient à rechercher les valeurs de la variable analysée qui minimisent cette fonction de coût. Pour se faire, suivant les techniques classiques d'optimisation sans contraintes, un modèle adjoint est construit, dont est déduit le gradient de la fonction de coût, qui doit être nul en situation optimale. Les méthodes dites « 3D-VAR » ou plus récemment « 4D-VVAR » sont des exemples de ces techniques. Elles sont généralement coûteuses et difficiles à mettre en œuvre de manière opérationnelle.

- ***L'adaptation statistique*** (Blond et al, 2002): les résultats issus de modèles déterministes peuvent être analysés en regard de ceux déduits d'analyses statistiques des observations aux stations disponibles. Généralement le but de cette démarche est d'améliorer la précision de la simulation déterministe localement, dans les zones où ces mesures sont localisées. En effet le modèle déterministe calcule les concentrations en chaque nœud d'un maillage tridimensionnel plus ou moins raffiné. Aussi l'idée de les corriger en fonction des données obtenues sur un réseau de stations de mesures, conduit à des résultats optimisés sur une surface relativement réduite. Ceci vaut en particulier lorsque le modèle est utilisé en mode prédictif.

  Des procédures d'adaptation statistique ont été développées par le Laboratoire de Statistique d'Orsay et évaluées sur la région Grand Ouest (Grancher et al., 2003, *Cartographie et prévision des champs de pollution à l'échelle locale, à partir des résultats de simulation d'un modèle continental*). Elles sont en cours d'implémentation dans certaines AASQA. Certaines d'entre elles ont été utilisées dans le cadre du projet PREV'AIR afin de fournir quotidiennement des cartographies de concentrations d'ozone de la veille ajustées grâce aux observations disponibles (Honoré et Malherbe, 2003).

- ***L'exploitation de cartographies d'erreurs*** établies à partir de l'interpolation (par exemple par la géostatistique) de la différence entre mesure et calcul aux stations du domaine permet de développer des méthodologies de réduction de l'incertitude sur les cartographies.
- ***L'utilisation des résultats issus d'un modèle déterministe comme dérive externe dans une procédure d'interpolation géostatistique (***Lajaunie et al, 2001, Wackernagel, 2002) : des expériences de ce type ont été menées, par AIRPARIF, de manière très récente sur l'Ile de France afin d'affiner les résultats issus du modèle CHIMERE régional

Même si elles font encore l'objet de nombreuses recherches menées actuellement, les méthodes hybrides se développent largement dans le domaine de la qualité de l'air avec des résultats prometteurs. Il est cependant important de rappeler qu'il s'agit de développements purement mathématiques qui peuvent s'avérer très complexes, et qui restent donc réservés aux spécialistes.

SECONDE PARTIE : LES MÉTHODES DE LA GÉOSTATISTIQUE LINÉAIRE

1. INTRODUCTION ET DEFINITIONS

La géostatistique est une application des méthodes d'analyse probabiliste à l'étude de phénomènes corrélés dans l'espace appelés ***phénomènes régionalisés***. A ce titre, elle fournit différents outils pour répondre au problème posé par la cartographie de la qualité de l'air.

On suppose que le phénomène régionalisé peut être décrit par la donnée d'une fonction numérique Z définie dans un domaine circonscrit de l'espace (le ***champ***) et désignée sous le terme de ***variable régionalisée***. Cette fonction n'est connue que partiellement, par l'intermédiaire d'un échantillonnage. Pour la qualité de l'air, cete variable est la concentration d'un polluant.

A partir des mesures disponibles et d'une **information relative à la position géographique de ces données**, les techniques de la géostatistique permettent de représenter la structure spatiale du phénomène considéré et d'estimer la répartition de la variable régionalisée dans la zone d'étude. Elles permettent en outre, du fait de leur caractère probabiliste, de prendre en compte la **part d'aléatoire propre à l'évaluation de la qualité de l'air.**

Elles introduisent la notion de fonction aléatoire afin de traduire d'une part l'aspect erratique de la variable régionalisée étudiée, qui empêche de prédire avec certitude les valeurs prises en différents points, et d'autre part, l'existence d'une certaine structure spatiale.

Les notions indispensables à la compréhension de ces techniques sont définies ci-après.

1.1 STATIONNARITE

L'interprétation des caractéristiques de la fonction aléatoire, en l'occurrence de ses deux premiers moments : espérance et covariance, impose de formuler une hypothèse sur ses propriétés de stationnarité. **La stationnarité est l'invariance par translation de la loi spatiale du processus étudié.**

- Une fonction aléatoire est ***stationnaire d'ordre deux*** si l'espérance et la covariance existent et sont stationnaires :

$$\forall\ s,\ E[Z(s)]=m$$

$$\forall\ s,\ \forall\ h,\ E[Z(s)Z(s+h)]-m^2=C(h)$$

- Une fonction aléatoire est ***strictement intrinsèque*** si ses accroissements Z(s+h)-Z(s) sont stationnaires d'ordre deux.

$$\forall\ s,\ E[Z(s+h)-Z(s)]=m$$

$$\forall\ s,\ \forall\ h,\ E\{[Z(s)-Z(s+h)]^2\}=2\gamma(h)$$

$\gamma(h)$ est appelé *semi-variogramme* ou plus couramment ***variogramme***.

Dans l'hypothèse de stationnarité d'ordre deux, covariance et variogramme existent et sont liés par la relation $\gamma(h) = C(0)-C(h)$. Dans l'hypothèse intrinsèque, seul le variogramme existe. C'est pourquoi il est généralement préféré à la covariance pour décrire et interpréter la structure spatiale du phénomène étudié.

Il faut noter **que la stationnarité dépend de l'échelle de travail**. Selon cette échelle, un même phénomène peut être considéré ou non comme stationnaire.

1.2 VARIOGRAMME

C'est l'outil fondamental en géostatistique pour analyser et modéliser la structure spatiale de la variable régionalisée. Il est défini aussi bien dans le cadre stationnaire d'ordre 2 que dans le cadre strictement intrinsèque. Il représente la variabilité moyenne des concentrations entre deux points en fonction de la distance *h* qui les sépare.

Comme il a été établi ci-dessus, le ***variogramme théorique*** d'une fonction aléatoire a pour expression :

$\gamma(h) = \frac{1}{2} E\left\{ [Z(s+h) - Z(s)]^2 \right\}$, où Z(s) est la concentration du polluant au point de l'espace *s*.

Le variogramme réel d'une fonction aléatoire est inconnu mais il peut être évalué à partir des données d'échantillonnage. On obtient ainsi le ***variogramme expérimental*** :

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{N(h)} [z(s_i) - z(s_j)]^2$$

N(h) : nombre de couples de points de mesure distants de h

$z(s_i)$: concentration au point de mesure s_i.

Ce variogramme expérimental n'est pas défini partout, notamment aux distances h pour lesquelles il n'existe aucune paire de points de mesure. Aussi lui est-il ajusté un **modèle de variogramme,** qui est la somme d'une ou plusieurs fonctions mathématiques.

2. RECOMMANDATIONS DE MISE EN ŒUVRE

Ce chapitre décrit et illustre la démarche qui permet de bien comprendre la nature des données et d'éviter une modélisation aveugle, susceptible d'aboutir à des cartes peu représentatives du phénomène en jeu. Les choix à opérer à chaque étape doivent être souvent considérés au cas par cas, en s'aidant de l'expérience. L'important est d'appuyer sa décision sur une observation rigoureuse et approfondie des données.

Pour des exemples détaillés d'analyse exploratoire, avec l'interprétation qui s'en suit, on peut se reporter aux études conduites sur Mulhouse et Montpellier (Fouquet, 2003, Le Loc'h, 2003).

La présentation théorique des différentes méthodes d'estimation est joint en annexe (Annexe A).

2.1 ETAPES D'UNE ETUDE GEOSTATISTIQUE LINEAIRE

2.1.1 Démarche générale

La mise en œuvre des techniques d'estimation géostatistiques exige une analyse préalable approfondie des données expérimentales, afin de bien cerner les caractéristiques de ces données et de décrire et modéliser la structure spatiale de la variable de pollution. Une étude géostatistique comporte généralement les étapes suivantes :

- 1. L'analyse structurale

Cette étape a pour fin de dégager les principales caractéristiques de la variable de pollution, notamment ses propriétés de stationnarité, de régularité et d'isotropie, et de construire un modèle géostatistique. L'examen des relations entre cette variable et des variables supplémentaires - qui sont soit des variables de concentration d'autres polluants, soit des variables auxiliaires, susceptibles d'apporter des informations de nature à améliorer les estimations -, est également inclus dans cette analyse.

L'analyse structurale constitue la partie essentielle de l'étude géostatistique. Le soin apporté à sa réalisation détermine la qualité de l'étape suivante.

- 2. L'estimation

Une fois le modèle géostatistique défini, l'estimation par krigeage peut être mise en œuvre.

Cette démarche, bien entendu, n'est pas figée et selon les résultats obtenus, on peut être ramené à la première étape.

L'usage de la géostatistique en pollution de l'air a pour fonction principale de fournir des cartes de concentration. Aussi est-ce aux **méthodes d'estimation locale de la géostatistique linéaire** que l'on s'intéresse en premier lieu. Toutefois ces méthodes n'offrent qu'une réponse partielle aux exigences réglementaires concernant l'évaluation des incertitudes d'estimation et des dépassements de seuil. On peut donc être conduit, dans une seconde approche, à recourir aux techniques de la géostatistique non linéaire.

2.1.2 Analyse exploratoire des données

La méthodologie décrite dans les paragraphes suivants s'applique à tout jeu de données de concentration, que celles-ci soient issues d'un réseau dense d'analyseurs ou d'un ensemble de tubes à échantillonnage passif.

Dans ce dernier cas, qui est le plus fréquent, les données recueillies sont généralement des concentrations hebdomadaires ou bihebdomadaires mesurées pendant plusieurs semaines consécutives et à des saisons contrastées. Selon l'objectif de son travail, l'utilisateur a le choix d'exploiter séparément chaque séquence de mesure ou de traiter les moyennes de ces séquences sur chaque saison ou sur l'année. Dans cette seconde option (cartographie sur le long terme), une analyse des relations statistiques et structurales entre les moyennes des semaines, des saisons et de l'année peut être utile à la bonne compréhension du phénomène de pollution et contribuer à une estimation plus précise et plus cohérente des concentrations.

Les campagnes d'échantillonnage s'attachent généralement à la mesure d'un polluant. Si toutefois l'étude a pour objet de cartographier dans une même zone et à une même période plusieurs composés, alors une analyse structurale conjointe des données de ces polluants est recommandée.

a) Histogramme

L'étude de l'histogramme permet :

- d'apprécier la variabilité des données et de détecter l'éventuelle coexistence de plusieurs populations de données. Ainsi, si l'histogramme présente plusieurs modes, il est intéressant d'observer à quel ensemble de données correspond chacun d'eux, en particulier si chaque mode est lié à une zone géographique différente ou si au contraire les données associées à chacun sont mélangées dans l'espace. Il importe aussi de voir si ces modes sont créés par des types de sites différents. Pour ce faire, le logiciel ISATIS a l'avantage de mettre en correspondance l'histogramme et la carte d'implantation des données (Figure 1). Selon les résultats de cette analyse et selon son expérience du terrain, l'utilisateur pourra être conduit à exploiter simultanément l'ensemble des données, à traiter séparément différents secteurs géographiques ou encore à ne conserver dans la modélisation qu'une partie des données, celle qui décrit le mieux le phénomène étudié (une telle décision mérite toutefois d'être confirmée par l'analyse variographique) .

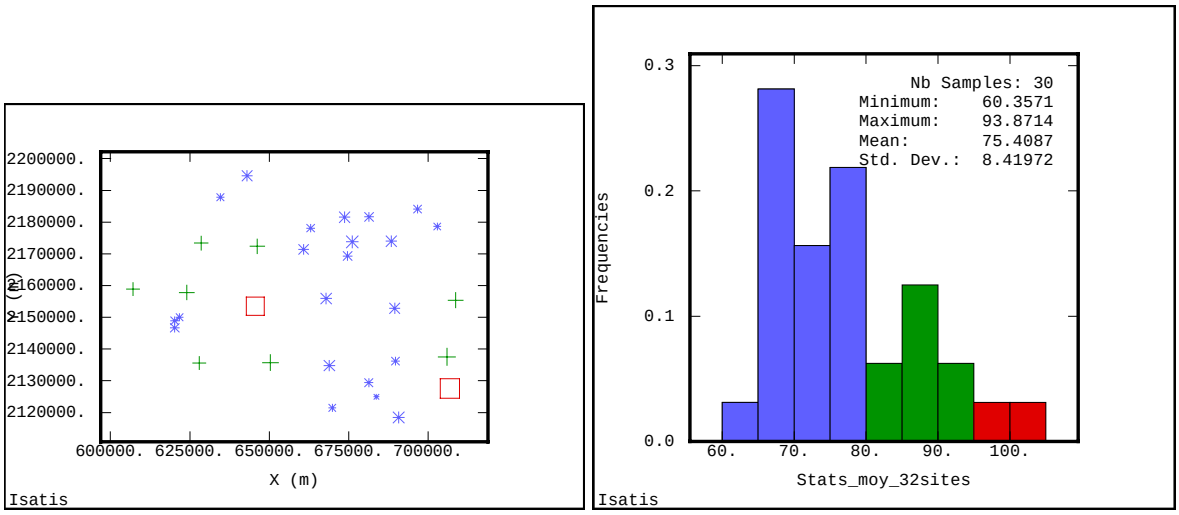


Figure 1 –Implantation des données et histogramme saisonnier (ozone, Allier, moyenne de 14 semaines de mesure, été 2001, données ATMO Auvergne).

Les concentrations les plus élevées correspondent aux zones de plus haute altitude. Les concentrations les plus basses correspondent aux zones de plus basse altitude et aux villes.

- de mettre en évidence des observations atypiques qui risquent d'influencer les résultats de l'analyse. Il n'existe pas de règle générale sur la façon de prendre en compte ces observations. La décision relève plutôt de l'expérience de l'utilisateur et du jugement qu'il porte sur la validité et la représentativité de ces données. Si celles-ci décrivent le phénomène au même titre que les autres mesures, il n'y a pas de raison pour les supprimer.

b) Corrélations entre périodes de mesure

Cette étape est recommandée lorsque les données sont issues de périodes d'échantillonnage différentes, typiquement de plusieurs semaines ou quinzaines de mesure en été et/ou en hiver.

L'étude des corrélations entre ces périodes, au sein d'une même saison d'une part, et entre deux saisons différentes d'autre part, permet d'examiner si les situations de pollution décrites par chaque séquence de mesure ont des caractéristiques similaires ou contrastées.

Une analyse en composantes principales (cf. description en annexe A) effectuée sur les concentrations moyennes hebdomadaires ou bihebdomadaires peut contribuer efficacement à cette étude. Elle indique en effet les regroupements ou les oppositions entre les périodes et les saisons.

Il est intéressant de compléter ces résultats en analysant les corrélations entre les séquences de mesure et les moyennes de ces dernières prises sur une saison ou sur l'année. Cet examen permet d'évaluer en première approche la représentativité temporelle d'une semaine ou d'une quinzaine par rapport à la saison et de la saison par rapport à l'année.

c) Corrélations entre polluants

L'étude des corrélations entre polluants est recommandée lorsque les données de deux ou plusieurs polluants sont disponibles pour la même période.

En effet, si, pour cette période, les concentrations des divers polluants ont été mesurées sur des ensembles distincts de points et qu'elles se révèlent fortement corrélées entre elles, alors un cokrigeage de ces polluants est préconisé.

d) Relations avec les coordonnées de l'espace

Le tracé de l'évolution des concentrations en fonction de x (axe EO) et de y (axe NS) est un moyen de détecter la présence d'une dérive spatiale qui, si elle existe, devrait se traduire graphiquement par une croissance ou une décroissance du nuage de points (Figure 2).

Dans ce même but, il peut être intéressant d'introduire les coordonnées x et y dans l'analyse multivariable décrite ci-après (§ e).

S'il existe un lien marqué entre les concentrations et les coordonnées de l'espace, il faut s'assurer que ce lien n'est pas créé artificiellement par une implantation préférentielle des points de mesure.

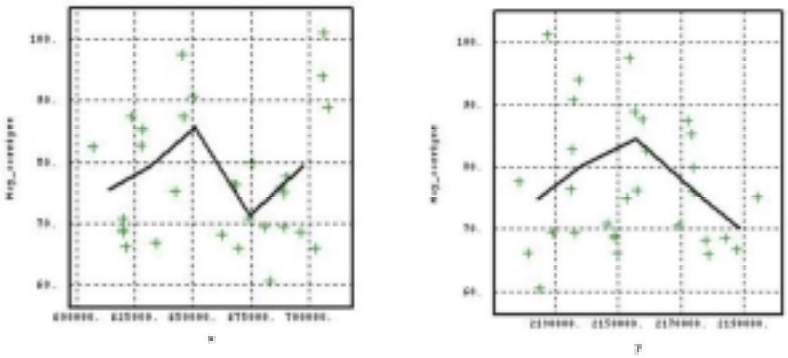


Figure 3 – Concentration moyenne estivale d'ozone en fonction des coordonnées de l'espace (Allier, 2001, données ATMO Auvergne). Aucune relation des concentrations avec x et y n'est clairement mise en évidence.

e) Variables auxiliaires et relations entre ces variables et les concentrations

L'information apportée par des variables complémentaires, liées directement ou indirectement aux concentrations et dont les valeurs sont connues en de nombreux points de l'espace, est susceptible d'améliorer la précision des cartes de pollution.

La sélection des variables auxiliaires les plus pertinentes, c'est-à-dire des variables qui expliquent le mieux les concentrations, ne peut se faire sans un examen détaillé :

- des relations entre les différentes variables auxiliaires ;
- des relations entre les variables auxiliaires et les concentrations.

Les tableaux ci-après, établis d'après les études conduites en 2002 par l'INERIS et le Centre de Géostatistique de l'Ecole des Mines de Paris, synthétisent la méthodologie proposée pour cette analyse.

Calculs réalisés	Objectifs
Calcul des statistiques des variables auxiliaires : statistiques élémentaires, histogrammes, variogrammes, corrélations entre variables. Étude de la stabilité de ces statistiques lorsque l'on modifie l'ensemble de points sélectionnés pour le calcul. Dans certains cas, une transformation logarithmique des variables initiales peut se révéler utile car elle a un effet stabilisateur. Exemple de transformation : $\boldsymbol{Y} = \log(1 + \boldsymbol{Z} / \overline{\boldsymbol{Z}})$ $\boldsymbol{Z}$: var***iable d'origine*** $\overline{\boldsymbol{Z}}$: ***moyenne de Z***	Ces calculs permettent d'apprécier la cohérence de l'ensemble des variables auxiliaires, en vue de réduire le nombre de ces variables et d'étudier les relations de ces variables avec les concentrations.
Analyse en composantes principales (ACP) Cette technique mathématique d'analyse multivariable permet de réduire un système complexe de corrélations en un plus petit nombre de dimensions. Elle remplace l'ensemble initial de variables par de nouvelles variables de variance maximale (les facteurs de l'ACP), non corrélées deux à deux et qui sont des combinaisons linéaires des variables d'origine. Réalisée sur la totalité des variables auxiliaires, l'ACP permet d'observer les corrélations, similarités ou oppositions entre ces variables et de mettre en évidence des familles de variables auxiliaires. Pour une analyse plus fine des corrélations, une ACP peut être ensuite effectuée au sein de chaque famille.	Ce travail a pour fin de réduire le nombre initial de variables auxiliaires. Il permet d'éliminer les informations redondantes pour ne conserver que certaines variables représentatives de l'ensemble. Il fournit en outre des variables synthétiques (les facteurs) qui peuvent être également employées comme variables auxiliaires.

Remarque technique :

Lorsque les variables auxiliaires se présentent sous la forme d'une grille de valeurs, l'utilisateur a deux possibilités pour conduire l'ACP :

1. L'ACP est effectuée aux nœuds de la grille. Ainsi obtient-on les valeurs des facteurs en tout point de cette grille, en vue de l'utilisation éventuelle de ces variables synthétiques comme dérives externes.

Pour disposer des valeurs des facteurs aux points de mesure on procèdera à une interpolation classique, à un krigeage ou à une **migration**. Cette dernière solution, d'usage courant en géostatistique, consiste à affecter aux points expérimentaux les valeurs des facteurs calculées aux nœuds les plus proches.

2. Une valeur des variables auxiliaires est attribuée préalablement à chaque point d'échantillonnage (par interpolation linéaire classique, krigeage ou migration) puis l'ACP est réalisée sur cet ensemble plus restreint de points.

   On suppose que les valeurs des facteurs fournies aux points de mesure par l'ACP se calculent aux nœuds de la grille par la même combinaison linéaire des variables initiales. Cette hypothèse est admissible **si les points de mesure sont représentatifs de l'ensemble de la zone à estimer**.

   Cette seconde méthode a l'avantage de préparer l'étape suivante c'est-à-dire l'étude des corrélations entre variables auxiliaires et concentrations aux points de mesure.

Une comparaison aux points de mesure des facteurs des ACP issus de ces deux méthodes permet de vérifier si les relations entre variables auxiliaires diffèrent sensiblement ou non entre les tubes et l'ensemble de la zone à estimer.

Calculs réalisés (suite)	**Objectifs**
Relations entre variables auxiliaires et concentrations Tracé du nuage de corrélation et calcul du coefficient de corrélation entre les variables auxiliaires (brutes, transformées ou synthétiques) précédemment sélectionnées et les concentrations Analyse en composantes principales effectuée sur les variables auxiliaires présélectionnées et les concentrations. Il convient là encore de vérifier si les corrélations et relations entre variables sont sensibles ou non à une modification de l'ensemble des points étudiés. Remarque : cette étape requiert comme données les valeurs des variables auxiliaires et des facteurs aux points de mesure (cf. remarque précédente).	Sélection des variables auxiliaires les plus corrélées avec les concentrations.

f) Etude du variogramme expérimental

Nuée variographique

La nuée variographique est le nuage de corrélation, au facteur 1/2 près, du carré des accroissements de la variable en fonction de la distance entre points de mesure. C'est donc le nuage des points $(½)[Z(s+h)-Z(s)]^2$.

Couplé dans ISATIS avec la carte d'implantation des données, cet outil permet de vérifier la présence de données spatialement resserrées -données nécessaires à la modélisation du variogramme à courte distance- et de repérer les couples de points qui contribuent le plus à la variabilité moyenne (Figure 4).

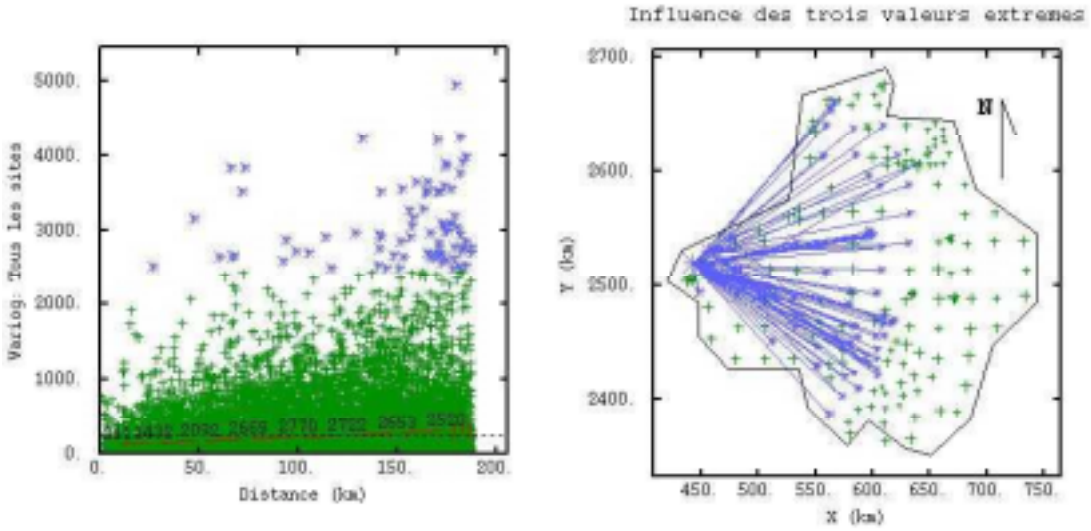


Figure 4 – Exemple de nuée variographique et carte d'implantation des mesures (données d'ozone, été 2000, première semaine de mesure, Nord de la France, données ATMO Pïcardie, AIR NORMAND, AIRPARIF, OPAL'AIR, AREMA LM, AREMARTOIS)

On examinera alors la répartition de ces couples dans l'espace.

- S'ils se répartissent de façon quelconque, la variabilité dont ils sont responsables peut être considérée comme une caractéristique du phénomène et il n'y a pas lieu de supprimer des points.
- S'ils sont créés par un ou quelques points en particulier, on vérifiera si ces points mesurent ou non le même phénomène.
  Dans le cas illustré ci-dessus, la nuée variographique est grandement influencée par une valeur extrême d'ozone, mesurée sur le littoral. Cette valeur, due à un phénomène local, induit des différences de concentration qui ne sont pas représentatives des écarts observés en moyenne sur le domaine d'étude. Aussi est-il pertinent de l'ignorer dans la modélisation variographique.

Calcul du variogramme expérimental

Le variogramme expérimental est calculé pour des distances allant jusqu'à la moitié du champ ainsi qu'il est généralement admis. Au-delà, le nombre de couples de points intervenant dans son calcul décroît, lui conférant un aspect plus erratique, et il perd en robustesse. Ce calcul implique que l'on définisse plusieurs paramètres :

- les directions de calcul, c'est à dire les directions de l'espace selon lesquelles le variogramme est calculé.

Le choix de ce paramètre est indissociable **de l'analyse des anisotropies**. En calculant le variogramme dans différentes directions de l'espace, on cherchera à voir s'il se différencie ou non selon ces directions. S'il est identique quelle que soit la direction considérée, on est dans le cas isotrope et le variogramme peut être calculé simultanément dans toutes les directions. Dans le cas contraire, on est en présence d'une anisotropie dont il convient d'identifier les directions principales (directions de continuité maximale et minimale). On peut s'aider à cette fin de la **carte variographique**, qui est la représentation en deux dimensions du variogramme (Figure 5).

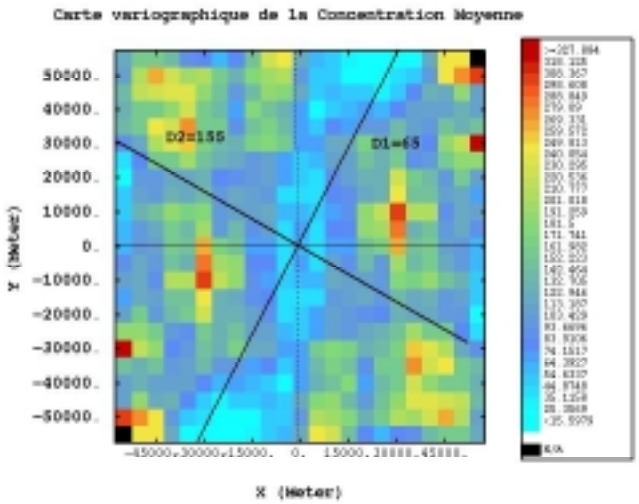


Figure 5 – Exemple de carte variographique (ozone, Allier, , moyenne de 14 semaines de mesure, été 2001, données ATMO Auvergne). Il s'agit de la carte des isovaleurs du variogramme expérimental dans toutes les directions de l'espace. Cette carte fait apparaître une anisotropie de 65° environ par rapport à l'axe est-ouest.

- le pas de calcul, c'est-à-dire la distance minimale entre deux points du variogramme expérimental.

  Généralement, ce paramètre est pris égal à la maille d'échantillonnage dans le cas d'un échantillonnage régulier. Pour un échantillonnage irrégulier, et si le nombre de données le permet, on lui attribue une valeur proche de la distance minimale entre points de mesure.

- la tolérance sur les distances et la tolérance angulaire (si anisotropie)

  En pratique, afin de disposer d'un nombre suffisant de données, le calcul du variogramme pour une distance h fait intervenir des couples de points distants de $h \pm \Delta h$ (la plupart du temps $\Delta h = h/2$). De même, le calcul du variogramme dans une direction θ fait intervenir des couples de points situés dans le secteur angulaire $\theta \pm \Delta\theta$ ($0 < \Delta\theta \leq 45°$). Les paramètres de tolérance Δh et $\Delta\theta$ permettent d'atténuer le caractère erratique du variogramme et de rendre ce dernier plus robuste lorsque les données expérimentales employées dans son calcul sont peu nombreuses ou réparties irrégulièrement dans l'espace.

  Les valeurs de pas et tolérance doivent être ajustés de façon que la structure du variogramme devienne apparente.

Analyse du variogramme expérimental

Un examen approfondi du variogramme expérimental est indispensable pour appréhender la structure spatiale du phénomène et déterminer le type de modélisation le plus approprié (isotrope, anisotrope, stationnaire, non stationnaire). On considèrera les caractéristiques suivantes :

1. Croissance du variogramme aux grandes distances et stationnarité. La forme du variogramme aux grandes distances nous renseigne sur la stationnarité de la fonction aléatoire sous-jacente.

   Si le variogramme est borné, alors, en théorie, la fonction aléatoire sous-jacente est stationnaire d'ordre 2. Dans ce cas il convient d'examiner la ***portée*** du variogramme, *i.e.* la distance à partir de laquelle celui-ci se stabilise (autrement dit la distance à partir de laquelle deux échantillons ne sont plus corrélés spatialement), et son ***palier***, *i.e.* sa valeur de stabilisation.

   Dans un champ très grand, cette valeur est théoriquement égale à la variance de la variable aléatoire mais à l'échelle des domaines considérés, il n'est pas rare que le palier et la variance des données soient différents. Dans ce cas on ajustera le palier sur la valeur de stabilisation du variogramme expérimental.

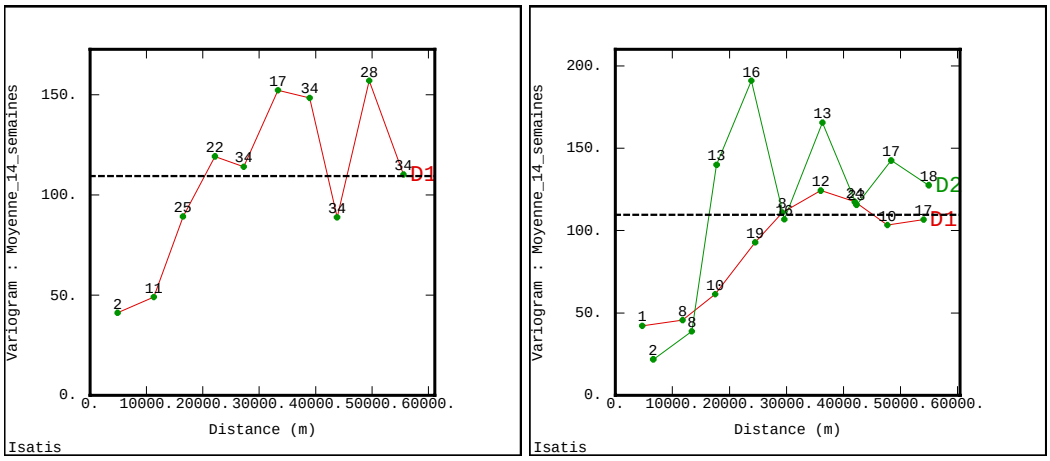
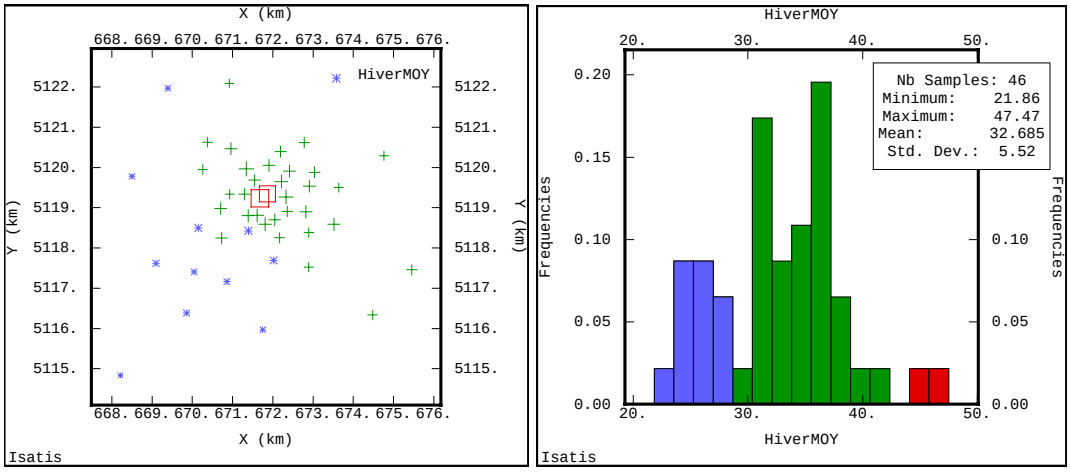


Figure 6 – Exemple de variogramme expérimental (ozone, Allier, moyenne de 14 semaines de mesure, été 2001). A gauche : variogramme calculé dans toutes les directions de l'espace. A droite : variogramme calculé selon les directions de continuité maximale et minimale mises en évidence dans la carte variographique

   Si le variogramme croît indéfiniment mais que cette croissance soit moins rapide que h^2, alors la fonction aléatoire vérifie l'hypothèse de stationnarité intrinsèque. Les outils de la géostatistique stationnaire restent applicables.

   Une croissance plus rapide que la fonction h^2 est révélatrice de la présence d'une *dérive*, *i.e.* d'une non stationnarité, et imposent de recourir aux outils de la géostatistique non stationnaire.



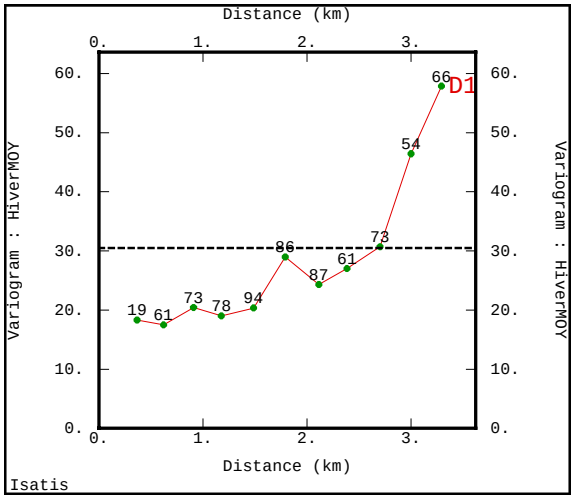


Figure 7 – Variogramme expérimental du dioxyde d'azote (Bourg-en-Bresse, moyenne de quatre quinzaines de mesure, hiver 2002, données Air de l'Ain et des Pays de Savoie). Mise en évidence d'un comportement non stationnaire. L'usage de variables auxiliaires dans un krigeage avec dérive externe se révèle particulièrement pertinent dans ce cas.

Remarque : un variogramme apparemment borné n'est toutefois pas incompatible avec l'existence d'une non stationnarité dans le domaine d'étude. Ce peut être le cas dans la cartographie d'une pollution en agglomération, lorsque les concentrations décroissent avec la distance mesurée à partir du centre ville puis se stabilisent à des valeurs quasi-nulles (Roth, 1999).

2. Isotropie et anisotropie. La notion d'isotropie a été introduite au paragraphe précédent. Si le variogramme ne dépend pas de la direction, il est dit ***isotrope***. S'il révèle des différences selon les directions de l'espace, il est dit ***anisotrope***. On distingue couramment deux types d'anisotropie :

- les *anisotropies géométriques* : les portées différent selon les directions mais le palier reste identique. Ce type d'anisotropie, dû par exemple à des directions de vent préférentielles, peut s'observer en pollution de l'air.

- les *anisotropies zonales*: les paliers changent selon les directions. Ces situations plus complexes peuvent être observées en présence d'une stratification de l'espace : par exemple en géologie, la teneur d'un élément sera peu variable dans une couche (faible palier) mais très variable perpendiculairement à celle-ci (palier plus élevé).

  Le risque de rencontrer ce type d'anisotropie en pollution de l'air est moindre.

3. Régularité. Le degré de régularité de la variable régionalisée est donné par la continuité et la dérivabilité en moyenne quadratique du variogramme $\hat{\gamma}$ (h) lorsque h tend vers 0.
   - Un comportement à l'origine parabolique de $\hat{\gamma}$(h) ($\hat{\gamma}$(h)~h^2) indique une grande régularité de la variable régionalisée qui est continue et différentiable.
   - Un comportement à l'origine linéaire ($\hat{\gamma}$(h)~h) montre que la variable régionalisée est moins régulière (elle est continue mais non différentiable).
   - Une discontinuité à l'origine, appelée ***effet de pépite***, signale une plus grande irrégularité de la variable régionalisée. Cet effet de pépite est dû à l'existence d'une microstructure spatiale non détectée par l'échantillonnage et/ou à la présence d'une erreur de mesure. Il y a effet de pépite pur si $\hat{\gamma}$ (h) est constant pour tout h strictement positif.

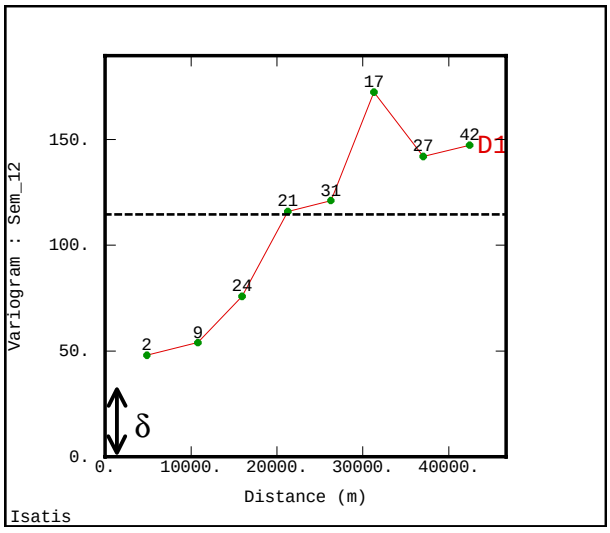


Figure 8 – Exemple de variogramme expérimental (ozone, Allier, une semaine de mesure, été 2001, données ATMO Auvergne. A très faible distance, le variogramme expérimental semble tendre vers une valeur non nulle δ. Cette discontinuité à l'origine constitue l'effet de pépite.

Cas multivariable : calcul des variogrammes croisés

Le calcul des variogrammes simples et croisés peut être réalisé dans l'optique d'un cokrigeage lorsqu'on dispose d'une ou plusieurs variables auxiliaires et que la variable à estimer se révèle fortement corrélée avec ces dernières (ρ>0,7) (Marcotte, 2003).

Ce calcul est également valable lorsqu'on veut estimer simultanément plusieurs variables de concentration corrélées entre elles.

Le variogramme expérimental croisé de deux variables Z_1 et Z_2 s'obtient par la formule :

$$\hat{\gamma}_{12}(h) = \frac{1}{2N(h)} \sum_{N(h)} [z_1(s_i) - z_1(s_j)][z_2(s_i) - z_2(s_j)]$$

Il est calculé sur les points pour lesquels les valeurs des deux variables sont disponibles. Si ces variables sont connues sur des ensembles distincts, voire disjoints, de points, il convient de procéder à une migration. Le principe de la migration a été donné en e) : aux points de la première variable, on affecte, par exemple, les valeurs de la seconde variable mesurées aux sites les plus proches.

L'étude des variogrammes expérimentaux simples et croisés permet d'examiner :

- si Z_2 vérifie aussi l'hypothèse de stationnarité d'ordre 2 ou de stationnarité intrinsèque;
- si l'on se trouve dans le cas d'un modèle à résidu (modèle de Markov), à savoir si la structure croisée de Z_1 et de Z_2 est proportionnelle à la structure simple de Z_2 : $\gamma_{12}(h)=a\gamma_2(h)$. Si tel est le cas, Z_1 est liée à Z_2 par la relation : $Z_1=aZ_2+b+R$, avec R : résidu aléatoire.

  Il est intéressant de contrôler cette hypothèse lorsque Z_2 est connue de façon dense et que l'on envisage de réaliser un cokrigeage colocalisé.

Remarque sur le mélange des sites

Dans les cartographies de grandes dimensions, l'échantillonnage mêle généralement des sites de types différents : rural, périurbain, urbain. Le variogramme expérimental doit-il prendre en compte indistinctement l'ensemble des données ?

L'utilisateur est amené à faire un choix, en fonction du type de phénomène qu'il souhaite représenter et du nombre de données disponibles.

Dans l'exemple qui suit, une analyse exploratoire a été conduite séparément pour chaque type de site. Dans la mesure où l'on souhaitait cartographier l'ozone à l'échelle de la région, seuls les sites ruraux ont été finalement conservés dans le calcul du variogramme. Pour vérifier la pertinence de ce choix, on s'est assuré qu'aux courtes distances, le modèle de variogramme ajusté sur ces sites était également cohérent avec les points expérimentaux urbains et périurbains. **Dans l'estimation, la totalité des sites (ruraux, urbains, périurbains et littoraux) a été conservée.**

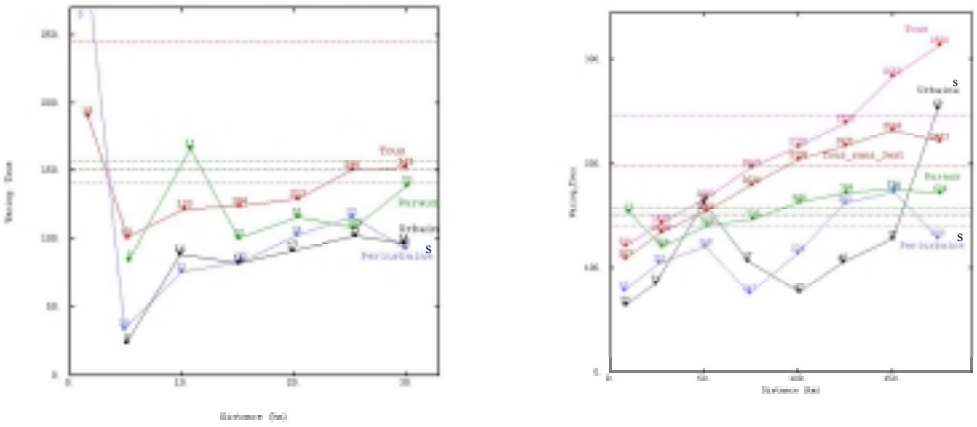


Figure 9 – Calcul des variogramme aux courtes distances (graphique de gauche) et à plus grande échelle (graphique de droite) pour chaque type de site et pour l'ensemble des sites –ozone, été 2000, première semaine de mesure, Nord de la France-

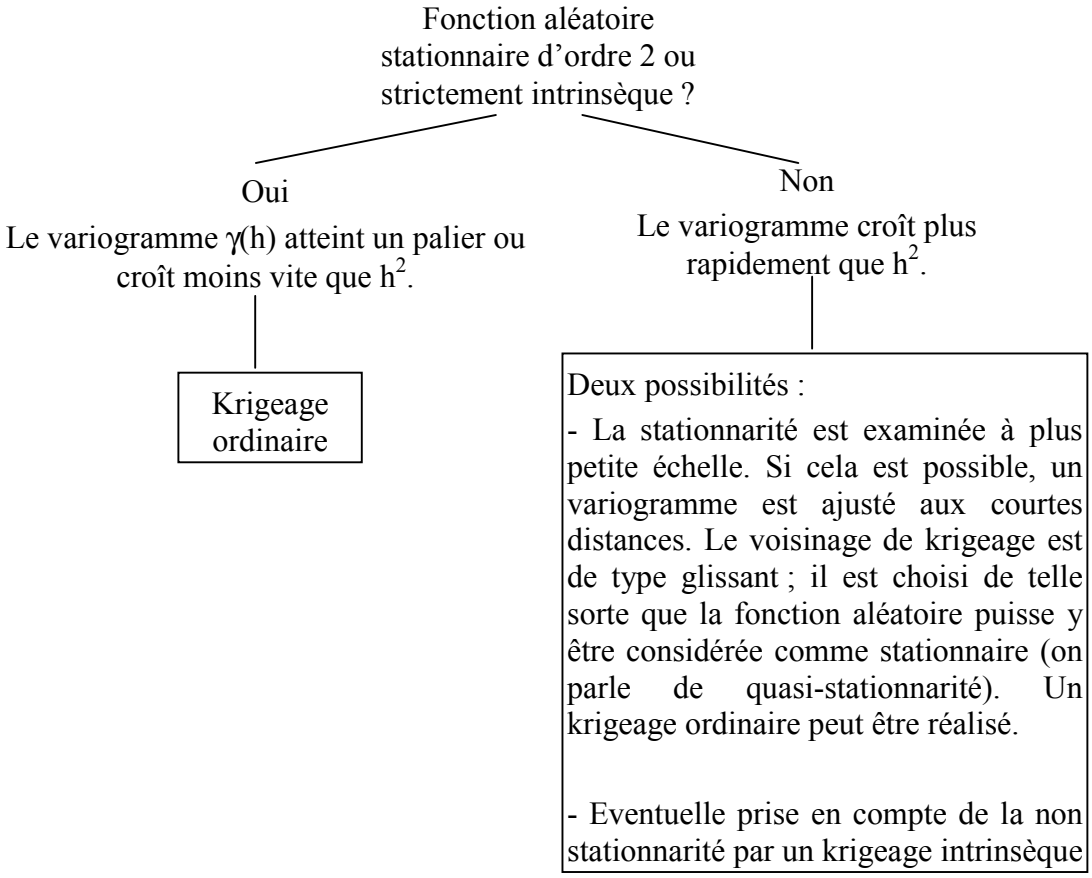
La question du choix des sites est toutefois à traiter au cas par cas. Le retrait de certaines données demeure un point délicat, et cela, d'autant plus que les mesures sont en nombre restreint. De plus, certaines stations bien que référencées comme urbaines peuvent être impliquées dans le cadre d'études de pollution de fond. Si des données présentent des caractéristiques différentes des autres, sans pour autant perturber les statistiques ni le variogramme expérimental, elles peuvent être conservées dans l'analyse. Dans les autres situations, la connaissance du terrain est décisive.

f) Choix de la méthode d'estimation et modélisation

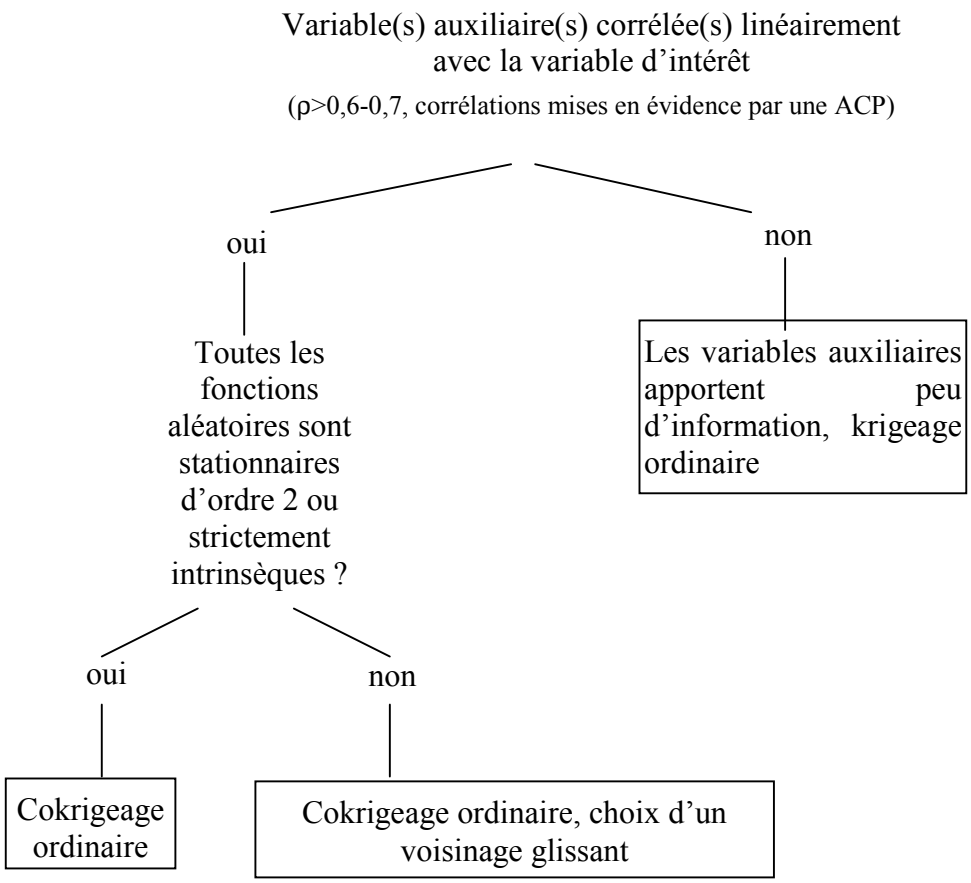
Ainsi mises en évidence dans les étapes qui précèdent, les propriétés structurales de la variable étudiée et la nature des relations entre cette variable et d'éventuelles variables complémentaires guident l'utilisateur dans le choix d'une méthode.

Les organigrammes présentés ci-après ne prétendent pas saisir toute la gamme des situations possibles mais ils illustrent de manière schématique la façon dont ce choix peut être orienté. Un modèle variographique est ensuite défini. Cette modélisation dépend de la méthode sélectionnée.

Cas monovariable. Aucune variable auxiliaire disponible



Cas multivariable. La ou les variables auxiliaires[1] sont disponibles en un nombre limité de points, qui sont communs ou non aux points de mesure de la variable principale.

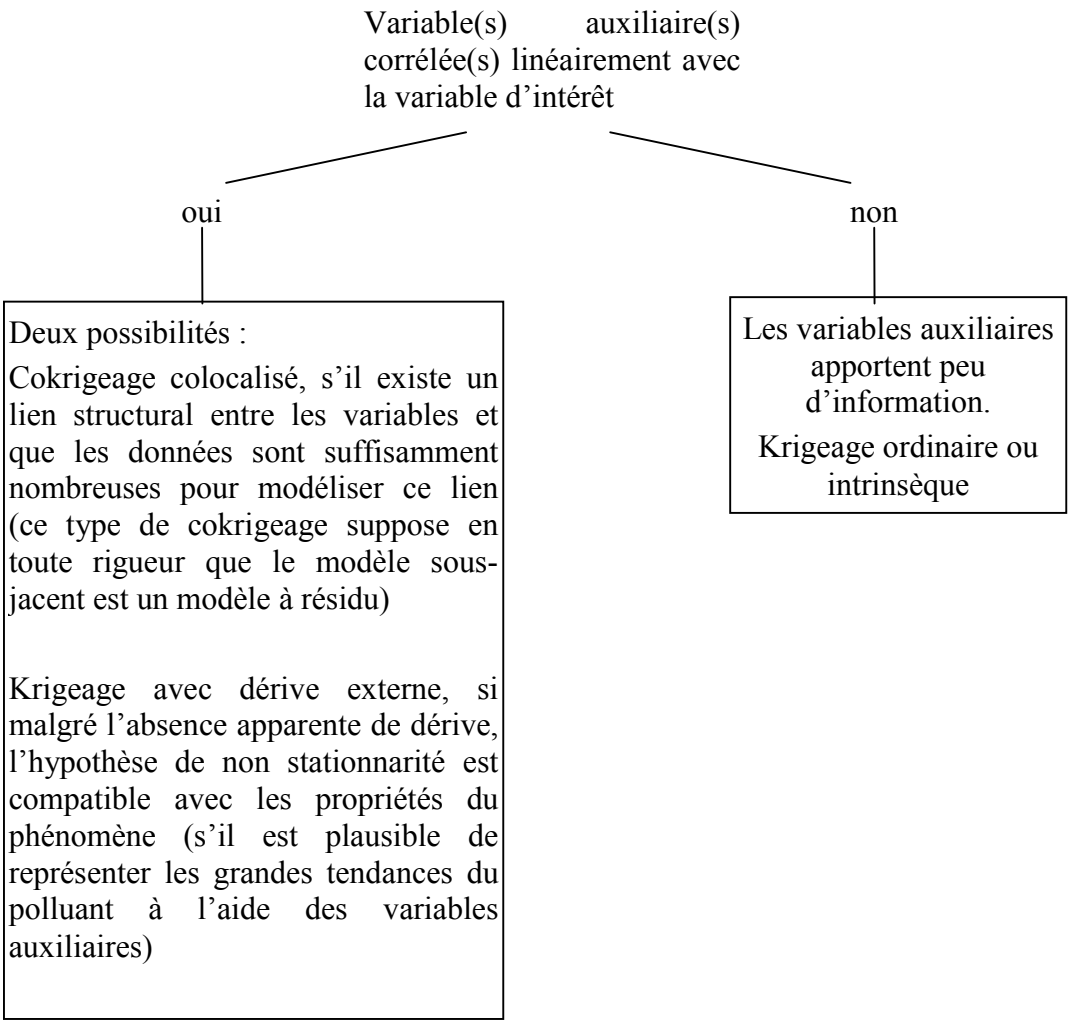


Remarque :

Si les valeurs des variables auxiliaires ne sont connues qu'aux points de mesure des concentrations (cas isotopique), l'intérêt d'un cokrigeage est réduit.

[1] Il s'agit des variables auxiliaires brutes ou après éventuelle transformation.

Cas multivariable. La ou les variables auxiliaires sont disponibles sur une grille.



Le terme *variables auxiliaires* est entendu au sens large. Il regroupe les variables auxiliaires d'origine (y compris les concentrations d'autres polluants) et les variables auxiliaires créées éventuellement par l'utilisateur à partir des variables initiales (variables ayant subi une transformation mathématique, facteurs de l'analyse en composantes principales, combinaisons linéaires ou non linéaires de plusieurs variables).

Modélisation

- ***Cas monovariable***

- Cas stationnaire d'ordre 2 ou strictement intrinsèque

Un modèle variographique est une expression analytique que l'on ajuste sur le variogramme expérimental. Toute fonction mathématique ne peut être utilisée comme modèle. Les modèles élémentaires autorisés les plus courants sont :

Modèles avec palier (compatibles avec l'hypothèse de stationnarité d'ordre 2)	
- le modèle pépitique (de palier C)	$\gamma(h) = 0$ si $h=0$ $\gamma(h) = C$ si $\|h\|>0$
- le modèle sphérique (de palier C et de portée a)	$\gamma(h) = C[(3/2).(\|h\|/a)-(1/2).(\|h\|^3/a^3)]$ si $\|h\|<a$ $\gamma(h) = C$ si $\|h\|>a$
- le modèle exponentiel (de palier C et de portée a)	$\gamma(h) = C[1-\exp(-\|h\|/a)]$
- le modèle cubique (de palier C et de portée a)	$\gamma(h) = C[7(\|h\|^2/a^2)-(35/4).(\|h\|^3/a^3)+(7/2).(\|h\|^5/a^5)-(3/4).(\|h\|^7/a^7)]$ si $\|h\|<a$ $\gamma(h) = C$ si $\|h\|>a$
- le modèle gaussien (de palier C et de portée a)	$\gamma(h) = C[1-\exp(-\|h\|^2/a^2)]$
Modèles sans palier (compatibles avec l'hypothèse de stationnarité intrnsèque stricte)	
- le modèle linéaire de facteur d'échelle ω	$\gamma(h) = \omega.\|h\|$
- le modèle puissance d'exposant α et de facteur d'échelle ω	$\gamma(h) = \omega.\|h\|^\alpha$ avec $0\leq\alpha<2$

Il est possible de combiner ces modèles en les additionnant. Ainsi un modèle se compose presque toujours d'un effet de pépite et d'une ou plusieurs structures élémentaires (rarement plus de deux ou trois : il est inutile en effet de chercher à en multiplier le nombre).

Un modèle dit « gigogne » s'écrit ainsi :
$\gamma(h) = \delta+\gamma_1(h) +\gamma_2(h)+\ldots..+\gamma_N(h)$

δ est l'éventuel effet de pépite.

- Il peut être obtenu par extrapolation à l'origine du variogramme.
- Il peut être aussi déterminé par la variance moyenne de l'erreur de mesure si celle-ci est une donnée disponible et que la variabilité des concentrations à courte distance puisse être jugée négligeable (du fait des caractéristiques du polluant) ou modélisée par une structure à très courte portée.

Remarque : une autre méthode, détaillée en annexe (cf annexe B), permet de prendre en compte la variabilité liée à l'erreur de mesure. Elle consiste à soustraire au modèle de variogramme l'effet de pépite dû à la variance de l'erreur de mesure (VEM) et à introduire dans le système de krigeage la VEM associée à chaque point de donnée.
Cette méthode a l'avantage d'accorder un poids plus important aux données les plus précises, limitant l'influence de l'erreur de mesure sur les résultats de l'estimation. En revanche, elle présente le risque de lisser excessivement la carte d'estimation, quand les fortes valeurs de la VEM sont associées aux fortes valeurs de concentration.

Quel que soit le mode d'ajustement retenu, la modélisation du variogramme aux courtes distances est particulièrement importante.

D'autre part, il existe un lien étroit entre les propriétés de la variable étudiée et le type de modèle susceptible de s'ajuster au données : ainsi le modèle cubique ou gaussien exprime une continuité caractéristique de polluants comme l'ozone (Figure 10). Notons cependant qu'en l'absence d'effet de pépite, le modèle gaussien peut conduire à des instabilités numériques (problème d'inversion de matrice lors de la résolution du système de krigeage). L'usage d'un tel modèle est peu recommandé dans ces circonstances.

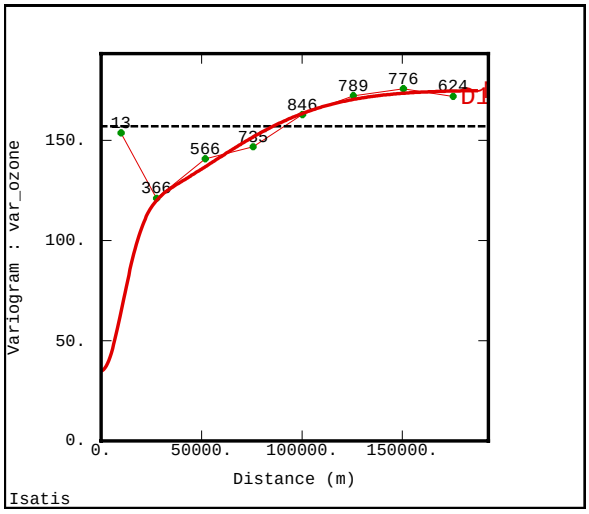


Figure 10 – Modélisation du variogramme expérimental par un effet de pépite et une somme de deux gaussiennes (données d'ozone du Nord de la France, première semaine de mesure, été 2001).

Le phénomène est supposé très régulier. On considère que la discontinuité à l'origine est due entièrement à l'erreur de mesure. L'effet de pépite est pris égal à la variance moyenne de l'erreur de mesure.

Si des anisotropies géométriques ou zonales ont été détectées, il convient d'ajuster le variogramme dans les différentes directions d'anisotropie :

- Dans une anisotropie géométrique, les composantes pépitiques et les paliers sont identiques quelle que soit la direction de l'espace considérée. En revanche, les portées doivent être ajustées en fonction de la direction. Cet ajustement est réalisé dans les directions de portée minimale et maximale (Figure 11).

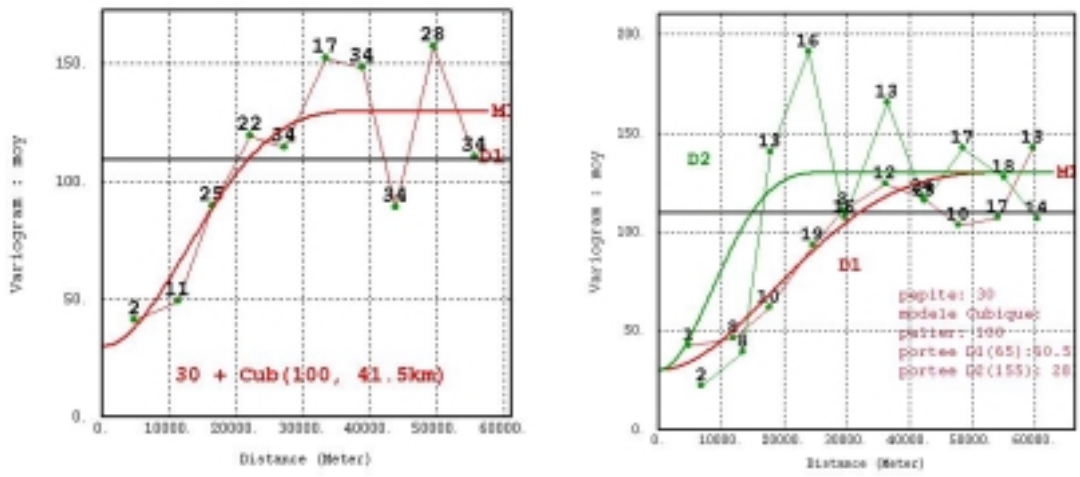


Figure 11 – Modélisation du variogramme omnidirectionnel et bidirectionnel (ozone, Allier, 2001).

Un modèle a été ajusté sur le variogramme omnidirectionnel et sur le variogramme bidirectionnel. En raison du nombre limité de données, la modélisation anisotrope est plus délicate, en particulier dans la direction de variabilité maximale (en vert sur le graphique). La variable altitude semblant expliquer en grande partie l'anisotropie, on voit l'intérêt que pourra présenter l'introduction de cette variable dans le modèle.

- La modélisation d'une anisotropie zonale est plus délicate. Le modèle d'anisotropie zonale le plus simple s'obtient par la somme d'une ou plusieurs composantes isotropes et d'une composante présentant une anisotropie géométrique. La portée de cette dernière composante est rejetée à l'infini dans la direction où le palier est le plus faible.

- Cas intrinsèque généralisé

Il n'est pas possible, comme dans le cas stationnaire, d'ajuster graphiquement un modèle de covariance généralisée. La modélisation, telle qu'elle est notamment réalisée dans l'option de modélisation non stationnaire d'Isatis, implique la mise en œuvre de processus d'optimisation. Elle procède comme suit :

- identification du degré de la dérive : plusieurs degrés sont successivement testés. Les résultats de la validation croisée associés à chaque essai fournissent un critère de sélection du degré optimal ;
- détermination par ajustements successifs de la fonction de covariance généralisée optimale, compatible avec le degré de la dérive. Cette covariance est constituée d'une ou plusieurs structures de base qui ont été présélectionnées par l'utilisateur.

• ***Cas multivariable***

- Cokrigeage

La modélisation procède de la même manière que précédemment, à la différence qu'il faut ajuster simultanément :

- les variogrammes simples associés à chaque variable
- le variogrammes croisé des deux variables

Cet ajustement est contraint par des conditions de positivité.

- Krigeage avec dérive externe

Plusieurs possibilités se présentent :

- on suppose que la non stationnarité est totalement prise en compte dans la ou les dérive externe(s). Il convient alors de définir un modèle de covariance (de variogramme) pour la fonction aléatoire stationnaire sous-jacente (c'est-à-dire $Z - \sum_i a_i \Phi_i$).

  Vu que les coefficients a_i sont estimés implicitement au moment du krigeage, l'ajustement d'un tel modèle est assez arbitraire. Une pratique est d'estimer le variogramme uniquement à partir des paires de points peu ou pas affectés par la dérive (c'est-à-dire à partir des paires de points perpendiculaires à la dérive) (Saito et Goovaerts, 2001, Wackernagel, 2002). Une autre méthode est de se placer dans le cadre non stationnaire en prenant pour dérives les dérives externes.
- on procède à un ajustement non stationnaire, la ou les dérives externes s'ajoutant à d'éventuelles dérives polynomiales.

  Remarque : l'utilisateur est contraint d'initialiser la procédure d'ajustement non stationnaire, en proposant une liste de modèles mathématiques élémentaires avec leurs portées. Afin de suggérer des modèles appropriés à la structure du phénomène, il peut être intéressant d'effectuer au préalable une régression linéaire multiple sur les variables en dérive puis d'examiner la structure variographique des résidus de cette régression.

g) Contrôle du modèle variographique

Avant de passer à l'étape d'estimation, il est nécessaire de contrôler la qualité du modèle qui a été ajusté. On effectue à cette fin une ***validation croisée***. Si plusieurs modèles semblent possibles la validation croisée permet également de trancher en faveur d'un modèle.

La validation croisée impose que l'on définisse le **voisinage de krigeage** autour de chaque point cible, c'est-à-dire que l'on sélectionne les points expérimentaux à prendre en compte dans l'estimation.

Ce voisinage est dit *unique* si tous les sites de mesure interviennent dans l'estimation en un point. Il est dit *glissant* s'il se réduit à une portion du domaine d'étude. Un voisinage

unique est adapté aux situations dans lesquelles les données sont en effectif limité ou à celles dans lesquelles le variogramme est peu structuré (forte composante pépitique).

Dans les autres cas, un voisinage glissant circulaire ou elliptique suffit. Un tel voisinage permet de réduire la dimension du système de krigeage. Il se caractérise par sa forme et sa taille, qui déterminent la zone de recherche au-delà de laquelle les données ne sont plus considérées, et, à l'intérieur de cette zone, par le nombre de points de mesure effectivement utilisés dans l'estimation. Ceux-ci sont souvent recherchés par quadrant afin d'être répartis aussi uniformément que possible dans l'espace. Selon les auteurs, un minimum de 10 à 15 points est recommandé, soit 3 à 4 points par quadrant.

Par ailleurs, la géométrie et les dimensions du voisinage doivent être cohérentes avec le domaine de validité du modèle de variogramme et avec les caractéristiques structurales de la variable régionalisée, en particulier la portée et les anisotropies. Il est notamment conseillé de ne pas limiter la taille du voisinage à la portée de la première structure, parce que des points situés au-delà de cette dernière peuvent jouer dans l'estimation. S'il existe une anisotropie, on peut adopter une zone de recherche elliptique parallèle à la direction de plus grande continuité. Toutefois, cette préconisation n'est pas systématiquement judicieuse et une zone de recherche circulaire convient si l'on augmente le nombre de points à prendre en compte.

Le principe et les résultats la validation croisée sont décrits plus en détail dans le rapport sur les incertitudes (LCSQA-INERIS, 2003). Nous en donnons ici une description synthétique.

Principe de la validation croisée :

La validation croisée consiste à éliminer temporairement un point de l'ensemble des données puis à estimer sa valeur par krigeage à l'aide des données voisines (*i.e.* incluses dans le voisinage de krigeage de ce point) et du modèle de variogramme qui a été ajusté. Cette opération est répétée pour tous les points.

En tout point d'échantillonnage, on obtient donc une concentration estimée Z^*, accompagnée de son écart-type de krigeage, noté σ_K : $\sigma_K = \sqrt{Var(Z^* - Z)}$. Cet indicateur n'est fonction que du variogramme et de l'implantation des données dans le voisinage de krigeage. Il quantifie la dispersion possible de la valeur vraie (connue dans le cas d'une validation croisée) autour de la valeur estimée.

L'estimation peut être alors comparée avec la concentration réelle Z.

On désigne ainsi par erreur d'estimation la variable Z^*-Z, par erreur réduite d'estimation la variable $(Z^*-Z)/\sigma_K$ et par erreur relative la variable $100|Z^*-Z|/Z$.

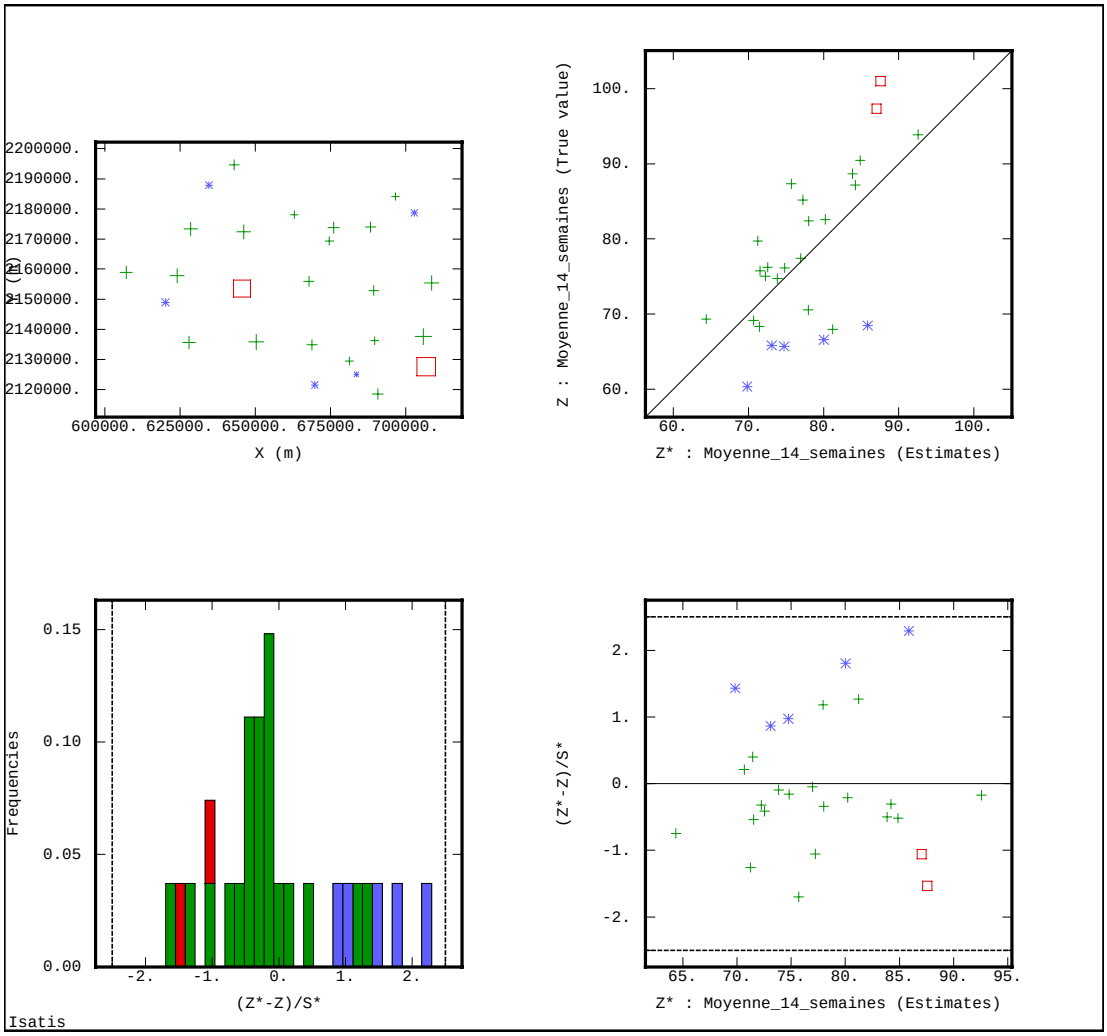
Plusieurs statistiques peuvent être calculées :

- moyenne et variance de l'erreur
- moyenne et variance de l'erreur réduite
- moyenne et variance de l'erreur relative
- corrélation entre Z et Z*

La qualité d'un modèle est d'autant meilleure que :

- la moyenne de l'erreur et de l'erreur réduite est plus proche de 0 : ce critère traduit l'absence de biais;

- la variance de l'erreur est plus faible : ce critère traduit la robustesse de l'estimateur ;
- la variance de l'erreur réduite est plus proche de 1 : ce critère indique que l'écart-type de krigeage reflète correctement la précision de l'estimation ;
- la moyenne de l'erreur relative est plus proche de 0 : ce critère traduit la bonne précision de l'estimateur;
- la corrélation entre Z et Z* est plus élevée et le nuage de corrélation plus resserré.
- Le nombre de données robustes, c'est-à-dire de données dont l'erreur réduite est inférieure à 2,5 en valeur absolue, est plus faible.



Moyenne des erreurs : -0,35 Variance des erreurs : 61,46

Moyenne des erreurs standardisées : -0,02 Variance des erreurs standardisées : 0,98

Erreurs relatives : min=0,58 max=25,39 moy=8,49 var=40,28

Figure 12 – Exemple de validation croisée (ozone, Allier, été 2001, modèle bidirectionnel monovariable)

Globalement, le modèle s'accorde avec les données expérimentales. En tout point, l'erreur réduite est inférieure à 2,5 en valeur absolue et la variance de cette erreur est proche de 1. Les erreurs relatives n'excèdent pas 26%. En revanche, la variance de l'erreur est assez élevée et le nuage de corrélation assez dispersé (ρ=0,664). Quelques valeurs fortes sont sous-estimées et quelques valeurs faibles surestimées.

Rôle de la validation croisée et interprétation des résultats :

Le rôle principal de la validation croisée est de fournir des critères statistiques de sélection entre plusieurs modèles possibles et de vérifier l'adéquation du modèle sélectionné avec les données expérimentales. Plus précisément, la validation croisée aide à choisir un modèle et un voisinage.

Les statistiques d'erreur relative donnent une indication sur l'incertitude des estimations (au sens des directives européennes[2]) dans l'enveloppe convexe des points de mesure.

Les résultats de la validation croisée n'ont cependant qu'une valeur qualitative :

- d'une part, la validation n'est réalisée que sur les points qui ont servi à construire le modèle ;
- d'autre part, ces résultats sont influencés par la configuration de l'échantillonnage et ils reflètent la qualité du modèle dans les zones les plus riches en données.

Ils sont insuffisants pour quantifier l'incertitude dans le champ.

Autre méthode

Une autre procédure consiste à supprimer les points non pas un par un mais par groupe. Pour chaque ensemble de points extrait des données initiales, les concentrations sont réestimées à l'aide des données restantes et les statistiques d'erreur sont calculées. Cette procédure complète efficacement la validation croisée. Comparée à cette dernière, elle permet de limiter l'influence de la configuration des données sur les statistiques d'erreur mais elle demeure insuffisante pour quantifier l'incertitude dans tout le domaine. Elle n'a, elle aussi, qu'une valeur indicative dans la zone d'échantillonnage.

2.1.3 Estimation

Cette étape est relativement rapide si l'analyse structurale a été menée à bien correctement.

Il convient, pour la mettre en œuvre :

- de définir la grille d'estimation
- de préciser le voisinage de krigeage.

a) Choix d'une grille d'estimation

Pour obtenir une estimation des concentrations dans un certain domaine, le krigeage est réalisé sur une grille régulière de points ou de blocs qui couvre ce domaine.

Dans un krigeage ponctuel, la taille de la maille de calcul est définie selon le degré de détail souhaité. La grille doit être suffisamment resserrée pour que la carte obtenue représente effectivement les estimations par krigeage et non la méthode d'interpolation employée par le logiciel pour tracer les isocontours. Toutefois, il n'est pas nécessairement utile ni pertinent de choisir une maille d'estimation très fine lorsque l'échantillonnage est peu dense.

[2] Ecart relatif entre valeurs modélisée et mesurée

Dans un krigeage de blocs, l'estimation est réalisée en moyenne sur chaque maille. La taille de ces mailles doit être un compromis entre le détail de représentation recherché et la précision d'estimation à atteindre. En effet, plus la taille des blocs augmente, plus l'écart-type de krigeage diminue : hormis dans les zones dépourvues de sites de mesure, le risque de commettre une erreur est moindre lorsqu'on estime une concentration moyenne sur un support relativement large que lorsqu'on estime une concentration sur un support ponctuel ou quasi-ponctuel. En contrepartie, plus la taille des blocs augmente, plus la carte présente des discontinuités marquées, avec, à l'intérieur de chaque maille, une représentation uniforme des concentrations.

Notons que dans un krigeage avec dérive externe ou dans un cokrigeage colocalisé, la grille d'estimation coïncide avec la grille des variables auxiliaires. Si celle-ci est suffisamment dense, elle peut être conservée pour l'estimation. Si elle est jugée trop lâche, une procédure consiste à interpoler préalablement les variables auxiliaires sur un maillage plus resserré. On cherchera à conserver au maximum la structure de ces variables. Une interpolation classique (inverse d'une puissance élevée des distances) est généralement suffisante.

Avec une grille d'estimation relativement fine, le krigeage ponctuel et le krigeage de blocs conduisent à des cartes d'estimation très similaires. La réalisation d'un krigeage de blocs a l'avantage de couvrir la totalité du domaine d'étude, fournissant à l'intérieur de chaque maille une concentration moyenne et sa variance d'estimation associée.

b) Choix d'un voisinage de krigeage

La définition d'un voisinage de krigeage n'est pas immédiate. Elle requiert au contraire une grande attention et doit s'adapter à la situation étudiée.

La validation croisée, dont il a été question précédemment, est un moyen de comparer différents voisinages. Une autre façon de contrôler la qualité d'un voisinage est de se placer en un point cible et d'observer si l'écart-type de krigeage diminue lorsque la taille de ce voisinage et le nombre de points utilisés dans l'estimation augmentent (Marcotte, 2003). On retiendra de préférence le voisinage conduisant à la variance d'estimation la plus faible (Ce test peut s'effectuer dans Isatis, dans la fenêtre de krigeage).

2.1.4 Interprétation de la carte de variance de krigeage

Comparées aux méthodes d'interpolation classiques, les méthodes de krigeage présentent l'intérêt d'associer aux valeurs estimées un indicateur de la précision de l'estimation. Cet indicateur est la variance de l'erreur de krigeage, encore appelée *variance de krigeage* (σ_K^2), ou sa racine carrée, l'*écart-type de krigeage* (σ_K). Son mode de calcul dans les différents types de krigeage est fourni en annexe A.

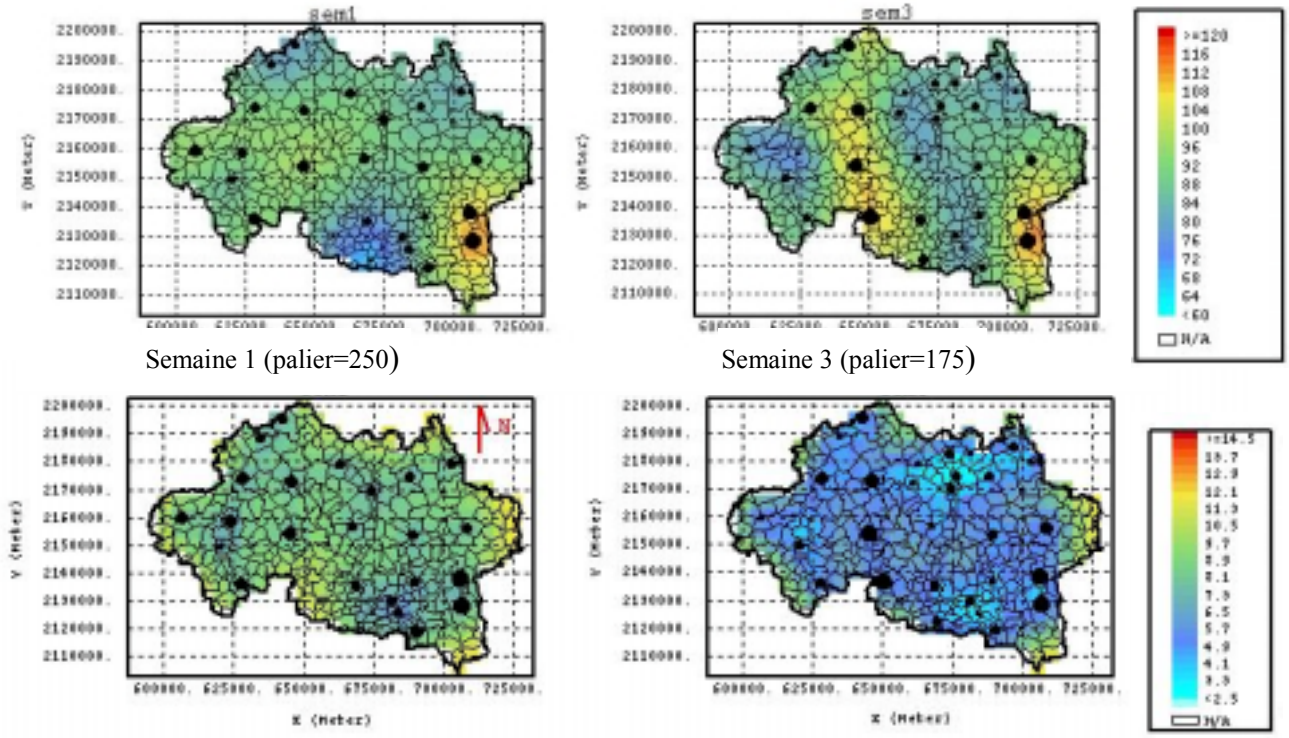
Quelle information peut-on tirer de ce paramètre ? Il faut à cette fin revenir à la définition de la variance de krigeage.

La variance de krigeage quantifie la dispersion possible de la valeur vraie, mais inconnue, autour de la valeur estimée.

Elle dépend uniquement du modèle de variogramme et de la configuration des données dans le voisinage de krigeage.

⇨ L'importance de l'écart-type de krigeage relativement aux concentrations estimées reflète la plus ou moins grande facilité avec laquelle le krigeage peut fournir des estimations précises, compte tenu de la variabilité du phénomène. Il s'agit d'une information moyenne sur le domaine d'étude.

Ainsi, soient deux semaines de mesure présentant des configurations d'échantillonnage et des niveaux de concentration similaires mais des variabilités différentes. Les écarts-types de krigeage associés à la variabilité la plus forte (palier total élevé) sont les plus grands, traduisant une moindre précision (Figure 13).



Ligne du haut : cartes d'estimation (µg/m³). Ligne du bas : écart-type de krigeage.

⇨ Les variations spatiales de l'écart-type de krigeage indiquent la perte de précision lorsqu'on s'éloigne des points de mesure. Elles renseignent sur les zones où l'échantillonnage est suffisamment dense (faible écart-type, bonne précision) et sur celles où il est trop espacé (grand écart-type, précision médiocre ou mauvaise).

En revanche, il faut avoir à l'esprit que si la variance de krigeage est fonction du variogramme, qui représente une **moyenne spatiale** sur un ensemble de couples de points, **elle ne dépend pas des valeurs numériques des données à l'intérieur du voisinage**.

⇨ L'écart-type de krigeage ne reflète donc pas la variabilité locale des concentrations et ne constitue pas une vraie mesure de l'incertitude locale.

Si, moyennant certaines hypothèses, l'écart-type de krigeage permet d'estimer des intervalles de confiance autour des concentrations estimées et en déduire une information sur l'incertitude, il demeure insuffisant pour quantifier parfaitement cette dernière. Ce point est plus largement discuté dans le rapport LCSQA (2003) *Evaluation des incertitudes associées aux méthodes géostatistiques*. On accordera surtout à l'écart-type de krigeage une valeur qualitative.

En particulier, le tracé de la carte du rapport σ_K/Z^* est un moyen visuel efficace de repérer les zones de bonne, médiocre ou mauvaise précision.

D'autre part, comme on l'a déjà précisé, l'écart-type d'un krigeage de blocs est fonction de la taille de la maille d'estimation (il diminue quand la dimension des blocs augmente). C'est pourquoi, en toute rigueur, une carte de krigeage et sa carte d'écart-type associée devraient faire mention du support spatial d'estimation.

2.1.5 Evaluation des incertitudes

Plusieurs sources contribuent à l'incertitude de la carte finale :

- L'incertitude de mesure et d'échantillonnage
- L'incertitude sur les paramètres du modèle
- L'incertitude sur les paramètres de l'estimation (voisinage de krigeage, support de l'estimation)
- L'incertitude due au caractère aléatoire de la variable de pollution.

Aussi voit-on l'importance de conduire rigoureusement les campagnes de mesure et l'analyse géostatistique afin d'en réduire l'impact.

Si des outils de contrôle sont disponibles (validation croisée, carte de l'écart-type de krigeage), la géostatistique linéaire révèle ses limites lorsqu'il s'agit d'évaluer l'incertitude au sens des directives européennes. (Celle-ce s'exprime comme l'écart relatif entre modélisation et observation : $100.|Z^*-Z|/Z$).

Un travail supplémentaire d'analyse, avec le recours aux techniques de la géostatistique non linéaire, notamment l'espérance conditionnelle, se révèle nécessaire.

2.2 ECHELLE SPATIALE ET TEMPORELLE

2.2.1 Echelle spatiale

Le souci de cartographier la pollution aussi bien dans les zones densément peuplées (agglomérations) que dans les régions dépourvues de stations fixes (zones rurales) impose de considérer des domaines de dimensions variables, qui vont de quelques kilomètres à quelques centaines de kilomètres.

On peut ainsi chercher à cartographier la pollution atmosphérique sur des départements ou des régions entières, ce qui soulève quelques questions. A cause de l'échantillonnage nécessairement limité, l'imprécision des estimations ne risque-t-elle pas d'être importante et de rendre peu fiables et peu significatives les cartographies produites ? Une même carte peut-elle englober des types d'environnement très différents ?

Diverses études attestent la pertinence des techniques géostatistiques dans ces diverses situations, **pourvu que l'échantillonnage et l'analyse de données soient conduits de façon rigoureuse et que l'on dispose de variables auxiliaires appropriées**. Cette dernière condition vaut spécialement pour les cartographies de grande échelle, susceptibles d'englober des zones très contrastées (du fait du relief, de l'occupation des sols, de la nature des sources polluantes).

Si l'estimation par krigeage de la qualité de l'air dans des agglomérations est désormais pratique courante (des cartes sont ainsi disponibles sur les sites Internet des AASQA), la réalisation de cartographies régionales est encore peu répandue en France. **La coexistence dans un même domaine d'étude de zones où le polluant agit différemment constitue l'une des principales difficultés de la cartographie à l'échelle régionale.**

Toutefois, l'expérience récente de plusieurs AASQA est riche d'enseignement sur ce sujet. L'ASCOPARG a étudié en profondeur l'élaboration d'une cartographie de l'ozone dans le département de l'Isère (125 km x 125 km), qui se caractérise par une occupation du sol et un relief très variés. Une cartographie à si grande échelle se révèle faisable, avec un schéma d'échantillonnage judicieusement défini et des variables auxiliaires soigneusement choisies. Les conclusions d'une étude réalisée par l'INERIS en 2002 sur la cartographie de l'ozone dans le département de l'Allier (125 km x 100 km, été 2001) vont dans le même sens. Notons que Casado et al. (1994) cartographient les concentrations horaires moyennes d'ozone au sud-ouest des Etats-Unis dans un domaine de superficie comparable (160 km x 160 km).

Pendant les étés 2000 et 2001, six associations (ATMO Picardie, AIR NORMAND, AIRPARIF, OPAL'AIR, AREMA Lille Métropole, AREMARTOIS) ont coopéré pour mesurer l'ozone et le dioxyde d'azote dans le Nord de la France. Le domaine d'étude couvre une superficie de 56000 km^2. Les représentations cartographiques de ces polluants ont été obtenues par la géostatistique. Cette entreprise de grande ampleur confirme la possibilité d'utiliser les méthodes géostatistiques à l'échelle régionale. Mais comme il a été mentionné, **il convient d'être vigilant sur l'échantillonnage spatial qui doit s'adapter aux gradients présumés de concentration, et sur le choix des sites de mesure utilisés dans l'étude variographique**. Dans cette étude seuls les sites ruraux, considérés comme les plus représentatifs du phénomène global, interviennent dans le calcul du variogramme. Les autres sites (urbains, périurbains) ne sont employés qu'au moment de l'estimation.

Enfin, confirmant l'ensemble de ces remarques, une cartographie estivale de l'ozone dans toute la région Rhône-Alpes a été établie par les AASQA du groupe GIERSA à partir de six semaines de mesure (une semaine par mois).

Remarquons qu'Anglais et Américains n'hésitent pas à élaborer des cartographies sur de très vastes territoires, uniquement à partir des données de stations fixes (Coyle et al., 2002, Purushothaman et al., 1999, Nikiforov et al., 1998, Lefohn et al., 1988). La recherche de la précision ne constitue pas nécessairement une exigence. Leur objectif est de fournir à grande échelle une image possible des concentrations auxquelles les populations ou les milieux naturels sont en moyenne exposés. Les conclusions que l'on peut tirer localement de telles représentations sont de nature qualitative et peuvent guider la réalisation de campagnes de mesure complémentaires.

2.2.2 Echelle temporelle

Selon l'objectif recherché, on distingue différentes échelles de temps :

- Des cartes établies avec un pas horaire ou journalier délivrent une information sur l'évolution spatiale et temporelle du polluant considéré.
- La réalisation de cartes bilans sur quelques semaines, sur une saison ou sur une année, permet de représenter a posteriori une situation moyenne.

Dans le premier cas, des données de stations fixes qui seules, fournissent une information à ce pas de temps, sont nécessaires (Georgopoulos et al., 1997, Liu et al., 1996). Le réseau de stations doit donc être suffisamment dense.

La réalisation de cartes sur de courts pas de temps, grâce à l'exploitation conjointe de données de tubes et de données d'analyseurs – ces dernières permettant d'établir un modèle de changement de support temporel- (Roth, 2000) n'est pas une méthode opérationnelle.

Dans le second cas, des cartographies multihebdomadaires, saisonnières ou annuelles peuent être obtenues :

- à partir d'un réseau de stations fixes, si celui-ci est suffisamment dense (Nikiforov et al., 1998, Casado et Vyas, 1994, Lefohn et al.,1987, 1988)
- à partir de campagnes d'échantillonnage par tubes à diffusion.

Jusqu'à ce jour, les cartes fondées sur l'exploitation des données de tubes décrivent une situation moyenne sur la durée des campagnes, soit quelques semaines en été et/ou en hiver. La dimension temporelle n'est pas encore prise en compte au moment du krigeage afin d'extrapoler ces cartes à des périodes plus longues, typiquement la saison ou l'année comme le demandent les directives européennes.

Une exploitation plus judicieuse des mesures estivales et hivernales grâce à la technique du cokrigeage a été suggérée par le Centre de Géostatistique (Fouquet, 2003). Les logiciels disponibles sur le marché ne permettent pas encore sa mise en œuvre.

3. LES DONNEES

3.1 DONNEES DE CONCENTRATION

Elles constituent la donnée d'entrée des méthodes d'interpolation et de géostatistique mais elles peuvent être aussi utilisées pour l'analyse des résultats issus de modèles déterministes afin d'en améliorer la qualité. L'objet de ce paragraphe est d'émettre un certain nombre de recommandations pour le recueil et l'utilisation de ces informations afin de cartographier la qualité de l'air.

3.1.1 Types de données

L'évaluation de la qualité de l'air d'une région peut s'appuyer sur deux catégories de mesures :

- des **mesures en continu** faites aux stations fixes. Le nombre minimal de stations est défini par la réglementation en fonction de l'environnement (urbain, périurbain, rural) et de l'importance de la population.
- des campagnes de mesures réalisées pendant des périodes de temps limitées. Les directives européennes les désignent sous l'expression de **mesures indicatives**. Ces campagnes peuvent être conduites, soit à l'aide de moyens mobiles, c'est-à-dire d'analyseurs installés dans un camion laboratoire ou une remorque, soit à l'aide de tubes à échantillonnage passif.

La nature de l'information obtenue dépend du mode d'échantillonnage choisi :

Source de données	Types de données	Information apportée par ces données pour la représentation de la qualité de l'air	Remarques
Stations fixes	Concentrations quart-horaires fournies en continu	Les données de stations fixes peuvent servir : - à établir des cartographies si elles sont suffisamment nombreuses dans le domaine étudié - à extrapoler à de plus longues périodes les informations qui résultent des campagnes de mesure, sous réserve d'un traitement statistique ou géostatistique adapté - à évaluer l'incertitude de mesure des échantillonneurs passifs	
Tubes à échantillonnage passif	Concentrations moyennes sur une à plusieurs semaines en plusieurs points de l'espace	- Méthode actuellement la plus employée pour disposer à moindre coût et avec une résolution spatiale suffisante de données moyennes sur la qualité de l'air - Possibilité d'interpoler entre les	

		données mesurées afin de cartographier la pollution sur de vastes domaines	
Moyens mobiles	Concentrations quart-horaires fournies en continu pendant quelques jours à quelques semaines, en un point donné de l'espace	- Moyen de disposer d'une information sur la qualité de l'air dans des communes non équipées de stations fixes - Même rôle que les stations fixes : couplage avec les échantillonneurs passifs afin de quantifier leur incertitude de mesure - Estimation de moyennes saisonnière et annuelles et de centiles, reconstitution possible de séries chronologiques, après un traitement statistique approprié	Nous ne considérons pas ici les études locales (étude d'impact, validation d'une station fixe, choix d'une future station fixe) qui ne concernent pas la réalisation d'une cartographie régionale

Actuellement, les cartographies par interpolation reposent de façon quasi exclusive sur l'exploitation des données de tubes à diffusion. En effet, seul ce mode d'échantillonnage permet de collecter à des coûts acceptables un grand nombre de données dans l'espace, afin d'évaluer la qualité de l'air d'une agglomération, d'un département ou d'une région entière.

La réalisation de cartes à partir des seules données de stations fixes n'est possible que dans les zones dotées d'un réseau de mesure suffisamment dense (ex : région de Rouen-Le Havre, ou de Fos-Marseille-Etang de Berre)

Comme on l'a évoqué, les données d'analyseurs (stations fixes et de moyens mobiles) n'en constituent pas moins une aide précieuse pour exploiter les données de tubes et quantifier l'incertitude de ces dernières. Des techniques de calcul de l'incertitude, étudiées notamment par AIR NORMAND et fondées sur la norme NF/ISO 13752 ou en ce qui concerne l'ozone sur la norme NF ENV 13005, ont été appliquées par des AASQA à l'occasion de campagnes de mesure (ATMO Picardie, 2000, ASCOPARG, 2001). Cette incertitude, exprimée comme une variance de l'erreur de mesure, peut être incorporée dans le modèle en tant qu'effet de pépite ou introduite point par point dans l'algorithme de krigeage (annexe A).

Dans tous les cas, on s'assurera de ne pas mêler dans une même carte, à moins d'appliquer un traitement préliminaire, des données dont les supports spatial et/ou temporel sont différents.

3.1.2 Collecte des données

3.1.2.1 Echantillonnage dans l'espace

Les données d'un analyseur peuvent être jugées représentatives de toute une zone de caractéristiques relativement homogènes (émissions, occupation du sol, relief, météorologie). Dans la plupart des cas cependant, le nombre de stations fixes dans le domaine étudié est trop faible pour permettre une analyse variographique fiable (problème d'instabilité des statistiques et du variogramme). La recherche d'une information spatiale plus détaillée au moyen de campagnes de mesure est alors indispensable à la réalisation d'une cartographie.

Plusieurs considérations interviennent dans la définition d'un schéma spatial d'échantillonnage, en particulier :

- la précision recherchée ;
- la densité de l'information auxiliaire disponible ;
- la taille du domaine que l'on souhaite cartographier et les moyens de mesure que l'on peut mettre en œuvre ;
- l'échelle spatiale du phénomène étudié (polluant transporté ou non sur de grandes distances).

Schéma d'échantillonnage

Il existe différents modes d'échantillonnage :

- dans un échantillonnage aléatoire uniforme, les positions des données sont tirées au hasard dans le champ ;
- dans un échantillonnage aléatoire stratifié, le champ est préalablement divisé en sous-domaines (les strates), à l'intérieur desquelles les positions des données sont tirées aléatoirement ;
- dans un échantillonnage systématique, les points de mesure sont régulièrement répartis dans l'espace.

La répartition des points de mesure dépend des possibilités matérielles d'installer un capteur ou un analyseur, et du type de site (urbain de fond, périurbain, rural) auquel on s'intéresse. On cherchera cependant, dans la mesure du possible, à disposer les échantillons selon un **maillage régulier**. Outre sa simplicité, une telle configuration a l'avantage de l'efficacité. A nombre d'échantillon égal, la régularité de la maille facilite l'inférence de la structure spatiale et garantit une meilleure précision. Il est en effet prouvé que la variance de l'erreur d'estimation d'un échantillonnage aléatoire stratifié est inférieure à celle d'un échantillonnage aléatoire uniforme. Dans le cas d'un échantillonnage régulier, cette variance est généralement plus faible encore (Arnaud et Emery, 2000, Fouquet C. (de), 1997).

Ces recommandations, qui ont été également émises par le Groupe de Travail *Echantillonneurs Passifs* s'appliquent en premier lieu aux campagnes par tubes passifs.

Densité spatiale d'échantillonnage

Peu de recommandations sur la densité d'échantillonnage apparaissent dans la littérature scientifique.

Pour cartographier les concentrations de NO_2 dans une agglomération, le Groupe de Travail *Echantillonneur Passifs* suggère de subdiviser préalablement le domaine étudié en fonction de l'occupation du sol, et d'adapter la densité de l'échantillonnage à chaque subdivision. Ces conseils peuvent tout aussi bien s'appliquer à la cartographie de vastes régions, dans lesquelles le degré d'urbanisation est très variable.

Dans son étude sur l'ozone dans le département de l'Isère, l'ASCOPARG analyse *a posteriori* l'efficacité du maillage choisi pour sa campagne par tubes (double maille de 20 km et 5km de large) et tire quelques préconisations de cette analyse.

Le Tableau 1 et le Tableau 2 synthétisent ces diverses recommandations.

Tableau 1 - Recommandations sur le maillage émises par les AASQA

Recommandations sur le maillage	Remarques
La taille maximale de la maille d'échantillonnage est celle au-delà de laquelle l'inférence d'une structure spatiale devient peu aisée, la corrélation spatiale étant très faible ou nulle.	Cette valeur est inconnue *a priori*. Toutefois, elle peut être évaluée à partir de campagnes de mesure antérieures ou d'études réalisées dans des régions de caractéristiques comparables (météorologie, topographie, émissions). Elle est généralement estimée à 20-25 km pour l'ozone.
Là où la corrélation avec la variable auxiliaire est plus faible, il est conseillé de diminuer la taille de la maille.	La corrélation avec les variables auxiliaires peut être estimée d'après le comportement dispersif des polluants. Ainsi, lorsque la variable auxiliaire *topographie* a vraisemblablement peu d'influence sur les concentrations (par exemple dans les régions de plaine où la migration des polluants est importante), il est suggéré de resserrer le maillage.
Adapter le maillage en fonction de la population présente. Dans les zones étendues, un maillage très resserré induit des coûts d'échantillonnage élevés qui ne sont pas nécessairement justifiés.	Dans les zones peuplées et hautement fréquentées, il convient d'échantillonner selon un maillage suffisamment fin (5 km par exemple). Dans les zones peu peuplées, la perte de précision due à l'élargissement de la maille (10 à 15 km par exemple) peut être jugée acceptable.
Ajuster le maillage en fonction de l'occupation du sol et des sources de pollution présentes	Resserrer le maillage dans les zones où les gradients de concentration risquent d'être élevés (ex : dans l'étude de la pollution de NO_2 et d'ozone dans le Nord de la France, la maille de 25 km est divisée par deux et par quatre dans ces zones)

Dans tous les cas, il est préconisé **de resserrer la maille d'échantillonnage en quelques endroits**, afin de faciliter la modélisation du variogramme aux petites distances (Fouquet, 1997). Lorsqu'une région présente des zones de pollution contrastées, il est conseillé d'effectuer ce resserrement local dans un secteur de fortes concentrations et dans un secteur de plus faibles concentrations, afin d'éviter l'introduction de biais.

De même, si des variables auxiliaires sont disponibles, on veillera à répartir les points de mesure dans les zones associées aux différentes classes de valeurs de ces variables (par exemple, dans les secteurs de faible, moyenne ou forte densité de population, ou encore à faible, moyenne ou grande altitude, si le domaine est montagneux).

Tableau 2 - Quelques tailles de maille recommandées (ordres de grandeur)

Dimension de la maille	Cas dans lesquels cette recommandation peut s'appliquer	Type de cartographie
20 à 25 km	Zone étendue peu peuplée. Faibles gradients de concentration. Concentration fortement corrélée avec une variable auxiliaire	Cartographie régionale
10 à 15 km	Gradients de concentration plus élevés. Concentration moins corrélée avec les variables auxiliaires	Cartographie régionale
5 km	Forts gradients de concentration. Zone densément peuplée	Cartographie régionale
1 à 2 km	Zone rurale	Cartographie d'une agglomération
250 m à 1 km	Zone urbaine ou industrielle, zone à points chauds	Cartographie d'une agglomération

Il s'agit de valeurs indicatives à ajuster selon les moyens disponibles et les caractéristiques du domaine et du polluant étudiés. En particulier, il est indispensable que la maille d'échantillonnage s'ajuste à l'échelle de transport du polluant, de façon à en saisir la structure spatiale.

Nombre minimal de points

Conjointement avec la maille d'échantillonnage, il importe de définir un nombre adéquat de points de mesure. Le nombre choisi résulte généralement d'un compromis entre la précision recherchée et les contraintes de coût. Dans ce compromis, intervient également la question de l'évaluation de la variabilité à petite distance (en resserrant localement la maille d'échantillonnage) et de l'erreur de mesure (en installant des tubes multiples). Disposant d'un nombre maximal de tubes à installer, on doit en effet choisir entre les options suivantes : répartir régulièrement les tubes dans l'espace, avec quelques resserrements locaux, afin de couvrir la zone d'étude le plus densément possible ; préférer un maillage plus lâche, donc un nombre de points de mesure plus faible, mais installer de nombreux doublets ou triplets de tubes.

Les préconisations sur le nombre minimal de points, là encore, sont peu nombreuses.

Diem (2003) a recensé plusieurs études géostatistiques sur la pollution de l'air. Celles-ci s'appuient sur des échantillons de taille extrêmement variable (entre 10 et 235 points). Comme le fait remarquer Diem, un nombre de points trop faible rend le variogramme instable et les résultats de l'estimation peu fiables. Selon Burroughs et Mc Donnel (1998), entre 50 et 100 points seraient nécessaires à la construction du variogramme. Le nombre de couples de données par classe de distances doit être en effet suffisant pour que le point du variogramme expérimental qui représente cette classe soit significatif. Cressie (1991) recommande un minimum de 30 paires par classe de distance.

Les résultats d'une analyse de la sensibilité du modèle au nombre de points, analyse réalisée par l'INERIS sur un jeu de 209 données d'ozone du Nord de la France, rejoignent les résultats de Burroughs et Mc Donnel (cf. Rapport LCSQA, 2003, Evaluation des incertitudes associées aux méthodes d'estimation géostatistiques).

En deçà de 95 points de mesure, le variogramme expérimental se dégrade et ne restitue pas correctement la structure de corrélation spatiale de l'ozone.

En revanche, il se révèle qu'avec un modèle bien ajusté (calé sur les 209 données initiales), un nombre relativement restreint de points est suffisant pour que la carte estimée reproduise les grandes tendances des concentrations. Ainsi, si de nombreuses mesures permettent une année de caractériser finement la structure spatiale d'un polluant dans une certaine région, est-il envisageable de réduire le nombre de données l'année suivante (sous réserve que les évolutions des émissions et de la météorologie n'induisent pas de modification de structure). Cette question relève aussi du problème de l'extrapolation temporelle et n'a pas encore trouvé de réponse. Une analyse plus approfondie est nécessaire.

3.1.2.2 Echantillonnage dans le temps

La réalisation de campagnes de mesure par tubes à diffusion a pour principal objectif de fournir une représentation moyenne sur le long terme de la qualité de l'air d'une région. Les AASQA étendent aujourd'hui leurs campagnes sur plusieurs semaines (consécuties ou réparties sur plusieurs mois), durant la saison estivale pour l'ozone, en été et en hiver pour le NO_2. Mais quelle fréquence et quelle durée d'échantillonnage permettent de s'assurer que la cartographie obtenue représente une situation annuelle ou saisonnière avec une précision acceptable ?

Dans le cas de mesures indicatives, les directives européennes fournissent à titre d'orientation des objectifs de qualité en ce qui concerne la période minimale à prendre en compte. Cette période, exprimée en % de temps, devrait être :

- supérieure à 10% en été pour l'ozone, le NO et le NO_2 (directive 2002/3/CE)
- de 14% pour le SO_2, le NO_2, les NO_x, les particules, le plomb, le benzène et le monoxyde de carbone (une mesure par semaine, au hasard, également répartie sur l'année ou huit semaines, également réparties sur l'année) (directives 2000/69/CE et 1999/30/CE)

Le respect de ces objectifs est demandé par l'arrêté du 17 mars 2003, à moins que l'échantillonnage choisi permette d'effectuer des estimations avec la précision requise.

Un guide technique européen (European Environment Agency, 1998. Guidance report on preliminary assessment under EC air quality directives. Technical report) préconise quant à lui une période d'échantillonnage égale à 20% de la période de référence (soit par exemple 2 fois 5 semaines ou 5 fois 2 semaines si la période de référence est l'année). Cette recommandation donne des résultats satisfaisants, comme l'indique la campagne de mesure de l'ozone réalisée dans la région Rhone-Alpes durant l'été 2002.

Les méthodes statistiques étudiées au sein du groupe de travail Etudes Mobiles, en particulier celle des plans de sondage (développée par ATMO Poitou-Charentes et évaluée par ATMO Poitou-Charentes et l'INERIS) et celle de l'échantillonnage stratifié (Ecole des Mines de Douai) devraient aider à formuler des recommandations. Toutefois, afin d'améliorer la précision des estimations, il apparaît aujourd'hui indispensable de coupler la réflexion sur l'échantillonnage temporel avec l'analyse du problème spatial.

3.1.2.3 Traitement temporel des données

L'utilisateur a le choix :

- d'établir directement des cartes à partir des données de mesure et de considérer ces cartes comme représentatives de la période de référence ;
- de traiter préalablement les données afin de les extrapoler sur de plus longues périodes.

L'estimation de concentrations moyennes annuelles par simple corrélation avec une station fixe est une méthode répandue (LCSQA-INERIS, Rapport *Représentativité des mesures et méthodes statistiques*, 2001), mais elle ne fournit pas d'intervalle de confiance autour de la moyenne estimée. Or les directives européennes définissent des objectifs de qualité en terme d'exactitude, ce qui suppose que l'on puisse évaluer les incertitudes sur les concentrations estimées.

Des méthodes plus poussées, que l'INERIS a recensé en 2002 dans sa mission d'assistance au groupe de travail Moyens Mobiles, pourraient être employées pour estimer une moyenne et l'incertitude sur cette moyenne[3].

[3] L'intérêt de ces différentes méthodes pour la réalisation de cartes moyennes sera étudié en 2004 (Fiche LCSQA-INERIS, *Géostatistique et prise en compte de l'aspect temporel*) .

- Dans une étude sur le benzène en Île de France, AIRPARIF a développé une approche par modélisation empirique, afin de réaliser des cartographies moyennes annuelles de ce polluant à partir de campagnes passives (Roth et Dégardin, 2001). Un modèle a été préalablement construit grâce à des données de stations fixes. Il exprime les concentrations journalières de benzène comme le produit de trois facteurs explicatifs : l'accumulation journalière (représentant principalement l'effet du trafic automobile), la dispersion journalière (effet du vent), et le facteur saisonnier (effet des conditions météorologiques). Ce modèle peut être calé localement par régression sur les concentrations mesurées aux points d'échantillonnage. Selon les tests effectués par AIRPARIF sur le benzène, il permettrait d'estimer la moyenne annuelle avec une incertitude de 5%. Rappelons qu'il s'agit là encore, d'une zone dense en nombre de stations de mesures, et que l'extrapolation de ce résultat en zone rurale n'est pas triviale.

- La méthode étudiée à l'Ecole des Mines de Douai est fondée sur la norme « ISO 9359 – Qualité de l'air – Echantillonnage aléatoire stratifié pour l'évaluation de la qualité de l'air ambiant ». Elle consiste à regrouper les mesures par classe météorologique, et à utiliser cette stratification pour estimer une concentration sur le long terme et l'incertitude sur cette moyenne.

- La théorie des sondages précédemment citée permet d'estimer une moyenne avec un intervalle de confiance, en s'appuyant si possible sur les données d'un site fixe auxiliaire afin d'améliorer l'estimation. L'avantage de cette méthode est de s'appliquer à tous les types de sites et de polluants, alors que ceux-ci ne se prêtent pas tous nécessairement à une modélisation, et de fournir un intervalle de confiance estimé selon une théorie rigoureuse (Tillé, 2001). En outre, elle ne requiert pas de variables externes telles que le trafic ou la météorologie. La seule contrainte d'utilisation est que l'échantillonnage compte un nombre suffisant de périodes de mesure et que celles-ci soient tirées de façon aléatoire ou aléatoire stratifiée. L'évaluation de cette méthode par ATMO Poitou-Charentes et l'INERIS est en cours de finalisation.

Une autre solution, vers laquelle on tend à s'orienter, est **de ne plus dissocier extrapolation temporelle et interpolation spatiale mais de coupler ces deux aspects au moment de l'estimation.** Le Centre de Géostatistique de l'Ecole des Mines de Paris (Fouquet, 2003) a montré en particulier l'intérêt d'un **cokrigeage temporel** (entre saisons et année). Ce point, là encore, sera approfondi en 2004.

3.2 DONNEES D'EMISSION

3.2.1 Nature de l'information

Les émissions représentent une donnée d'entrée fondamentale des modèles déterministes. Dans bon nombre de situations, elles constituent aussi une information secondaire pertinente pour le krigeage. Du fait qu'elles expliquent directement ou indirectement les concentrations (selon qu'il s'agit d'un polluant primaire ou secondaire), elles peuvent servir de variables auxiliaires dans l'interpolation par cokrigeage (multi)colocalisé, krigeage avec dérive externe ou régression.

Les difficultés liées à l'élaboration d'un inventaire d'émission résident essentiellement dans le recensement des sources de toute nature (industrie, trafic, chauffage, végétaux…) sur l'ensemble du domaine qui peut s'étendre sur plusieurs dizaines, voire centaines, de kilomètres. Aussi le travail minutieux d'inventaire peut-il être très coûteux en temps et en moyens humains. Cela est dû à l'étendue du domaine de calcul, à la variété des sources à considérer, aux lourdes incertitudes qui entachent certaines données (chauffage, végétaux), mais aussi à la bonne volonté et à la disponibilité des organismes qui détiennent ces informations.

Notons qu'à ce jour, l'élaboration d'un cadastre des émissions est guidée par les nécessités de la modélisation déterministe. C'est donc à l'utilisateur en charge de la cartographie de sélectionner, parmi les données d'émission disponibles, celles qui conviennent le mieux à la mise en œuvre du krigeage, voire même d'adapter le support spatial de ces données à ses propres besoins .

Ainsi, pour réaliser des cartes de veille en zone rurale sur de vastes domaines, il n'est nécessairement opportun de procéder à des estimations très fines, au km^2 par exemple. Dans ce cas, des inventaires suivant une maille de 5 à 10 km^2 sont des choix acceptables. En zone urbaine en revanche, à l'intérieur de domaines plus restreints, une maille resserrée (entre 200m et 1 km) est préconisée.

Le problème est le même pour la précision temporelle. Si l'objectif est d'établir des cartes de veille, un inventaire moyen dans le temps suffit. En général, seul un cadastre annuel est disponible. Toutefois, les grandes différences entre saisons, mises en évidence dans plusieurs études géostatistiques (Centre de Géostatistique, 2003), montrent l'intérêt que pourraient présenter des cadastres saisonniers.

3.2.2 Acquisition des données

Des inventaires d'émission à une échelle globale, existent et sont disponibles:

- inventaire EMEP (www.emep.int dans le cadre de UN-ECE[4])
- inventaire PRQA 1994 (France, inventaire CITEPA)
- inventaire GENEMIS (IER Stuttgart, www.uni-stuttgart.de/genemis programme EUROTRAC2)
- inventaire GEIA (Global Emission Inventory Activities, http//weather.engin.umich.edu/geia)

Des inventaires locaux de résolution plus fine, plus appropriés à la cartographie à l'échelle urbaine ou régionale, ont été élaborés dans certaines régions.

[4] UN-ECE : United-Nations, Commission Economique Européenne

3.3 DONNEES DE SITE ET DONNEES METEOROLOGIQUES

Cette catégorie regroupe les données annexes (hors émissions) susceptibles de fournir une information complémentaire lors de la mise en œuvre de modèles d'interpolation et de géostatistique (cartographie par régression, dérive externe, cokrigeage). Les variables concernées sont :

- ***La topographie,*** décrite par un modèle numérique de terrain (MNT).

  Des MNT sont fournis par l'IGN, avec un pas de discrétisation pouvant descendre jusqu'à 50 m.

  GTOPO30 est un MNT élaboré par l'USGS (U.S. Geological Survey). Il couvre le monde entier avec une résolution de 30 secondes d'arc (environ 1 km).

  Comme en témoignent plusieurs études (Jeannée, 2003, INERIS, 2002, ASCOPARG, 2001), la variable altitude sert plus particulièrement à la cartographie de l'ozone dans les régions de relief contrasté.

- ***L'occupation du sol***

  Cette variable se décompose en plusieurs catégories qui décrivent les surfaces bâties, industrielles, agricoles, forestières…

  L'inventaire Corine Land Cover de l'IFEN fournit à l'échelle 1 :100000 une description très détaillée de l'occupation du territoire, selon une nomenclature composée de 44 postes.

  Notons qu'un inventaire kilométrique est accessible gratuitement sur Internet à l'adresse suivante :
  http://gclf.umiacs.umd.edu/data/landcover/1km.shtml (site de l'université de Maryland). Quatorze classes d'occupation du sol sont considérées.

  Les variables auxiliaires recherchées sont en fait les densités de surface d'une ou plusieurs catégories (ex : densité de bâti). Ces données, en général, ne sont pas directement disponibles. Pour les obtenir, un traitement numérique de l'inventaire est nécessaire. On peut s'aider à cette fin d'un système d'information géographique.

- ***La densité de population***

  L'INSEE délivre deux types d'information :

  - la population par îlot, à l'échelle de l'agglomération;
  - la population par commune, à l'échelle régionale.

  Là encore, la densité de population n'est pas une donnée directement disponible. L'obtention d'une grille de densité de population nécessite un traitement numérique approprié, qui tienne compte si possible de la densité de bâti. Cette information est un moyen de ventiler la population dans l'espace de façon plus réaliste avant de calculer le nombre d'habitants par maille.

- ***Des variables météorologiques***

- température
- vitesse de vent
- direction de vent
- humidité

- couverture nuageuse
- …

On distingue différents modes d'acquisition de ces variables :

Elles peuvent être obtenues **en quelques points du domaine d'étude**, grâce à des stations de mesure qui sont représentatives à une altitude de dix mètres, afin de se dégager des perturbations du sol.

Les données issues de ces stations se présentent sous forme séquentielle, avec un pas de temps tri-horaire, horaire ou quart-horaire. Elles sont également traitées statistiquement, ce qui permet de disposer d'informations moyennes. Il s'agit des roses des vents qui fournissent suivant chaque direction (par classe de 20°) la fréquence et l'intensité du vent. Ces roses sont établies mensuellement, par saison ou annuellement.

Les sorties numériques issues des modèles de prévision météorologique (ARPEGE ou ALADIN pour la France) offrent également une **discrétisation sur le domaine d'étude des variables météorologiques.**

A ce jour, les variables météorologiques ne sont pas employées comme variables auxiliaires dans l'application des techniques géostatistiques.

Les roses des vents représentent cependant une donnée utile à la recherche et à l'interprétation des anisotropies.

D'autre part, il serait envisageable d'utiliser les champs météorologiques comme variable secondaire dans un krigeage avec dérive externe ou dans un cokrigeage.

4. CONCLUSIONS - RECOMMANDATIONS

L'objet de ce document est de fournir aux AASQA une synthèse et une analyse critique des méthodes numériques permettant d'élaborer des cartographies pertinentes de la qualité de l'air à l'aide des données disponibles (concentrations atmosphériques, émissions, météorologie, données de site).

Si les méthodes déterministes et les méthodes classiques d'interpolation ont été évoquées au début du rapport (partie 1), celui-ci s'est principalement attaché à l'étude des **méthodes de la géostatistique linéaire** (partie 2).

Le principe de ces techniques et la démarche à suivre dans leur mise en œuvre font l'objet du premier chapitre de la partie 2. Le second chapitre est consacré à la description des données d'entrée.

Un certain nombre de recommandations sur l'usage de la modélisation appliquée à la représentation cartographique de la qualité de l'air se dégage de cette analyse.

4.1 METHODES D'INTERPOLATION CLASSIQUES

- ✓ Elles reposent sur des concepts simples d'interpolation, accessibles à tous, et sont disponibles dans bon nombre de logiciels de représentation graphique.

- ✓ Pour limiter les incertitudes, elles nécessitent, par définition, de disposer d'un réseau assez dense de mesures. Elles peuvent également intégrer par des procédures de régression des variables externes telles que les émissions, les caractéristiques de site. Cette approche est pertinente lorsque les concentrations observées sont très corrélées aux émissions (polluants passifs émis au niveau du sol où les phénomènes de dispersion sont limités, par exemple).

- ✓ Ces méthodes ne peuvent généralement pas reproduire la véritable structure spatiale des concentrations, ce qui peut aboutir à des résultats peu réalistes.

- ✓ L'élaboration de cartographies sur de longues périodes ne peut se faire qu'à partir des données issues du réseau de mesure fixe ce qui limite l'accès à ces méthodes pour les zones rurales.

4.2 METHODES DE GEOSTATISTIQUE

- ✓ Les méthodes géostatistiques supposent que les valeurs de concentration mesurées sont la réalisation d'un processus aléatoire dont est modélisée la fonction de covariance ou le variogramme. A la différence des méthodes d'interpolation déterministes, elles permettent de prendre en compte la structure spatiale du phénomène étudié, par l'intermédiaire de ce variogramme, et de joindre à la carte d'estimation la carte de la variance de l'erreur d'estimation. Elles regroupent différents algorithmes d'estimation dont le choix dépend des caractéristiques du phénomène mises en évidence et des informations disponibles.

- ✓ Les méthodes d'estimation géostatique se révèlent adaptées à la représentation de la qualité de l'air à l'échelle urbaine et régionale, **pourvu que l'on dispose de données appropriées et suffisamment nombreuses pour décrire le phénomène en jeu.**

Cela suppose,

1. en amont :

▪ de définir un **échantillonnage correctement réparti dans l'espace et adapté aux gradients de concentration**, de façon à pouvoir identifier la structure de corrélation spatiale.

L'échantillonnage à l'aide de tubes à diffusion passive est la technique de mesure la plus employée pour couvrir l'ensemble d'une zone.

On réalisera de préférence un échantillonnage systématique régulier et on veillera à resserrer la maille d'échantillonnage en plusieurs endroits, afin de caractériser la variabilité à courte distance. L'évaluation de la variance de l'erreur de mesure par un échantillonnage multiple est utile à l'ajustement du variogramme à l'origine.

Des calculs de sensibilité sur un cas test à l'échelle régionale indiquent qu'avec moins de 95 données, le variogramme expérimental s'écarte du modèle variographique supposé décrire correctement la réalité (car construit avec plus de 200 données), et que la concentration moyenne estimée dans le domaine d'étude devient instable. La dégradation du variogramme expérimental est plus particulièrement marquée en deçà de 40 données.

▪ de **rechercher des variables auxiliaires susceptibles d'expliquer les concentrations mesurées**. En particulier, l'usage de telles variables est efficace lorsqu'on souhaite établir des cartes sur de grandes étendues qui couvrent des environnements contrastés (ville/campagne, montagne/plaine…). Les variables auxiliaires permettent ainsi de restituer plus finement les variations locales de la pollution.

L'optimisation de l'échantillonnage et un choix judicieux de variables auxiliaires sont souvent préférables à la recherche d'un modèle optimal.

■ d'étendre l'échantillonnage sur plusieurs semaines, consécutives ou non, en hiver et/ou en été, afin de prendre en compte les évolutions de concentration et de variabilité spatiale observées dans le temps. Des recommandations plus précises sur ce point seront fournies en 2004.

2. dans la mise en œuvre des méthodes :

■ de conduire une analyse exploratoire des données rigoureuse et approfondie.

Cette étape ne se restreint pas au seul calcul du variogramme. On accordera une attention spéciale à **l'étude des statistiques et du variogramme en fonction de l'implantation des sites et du type de donnée**.

De même les **relations entre variables de concentration et variables auxiliaires** requièrent une analyse poussée (étude des corrélations, analyse en composantes principales, recherche de transformations permettant de linéariser ces relations). Dans une zone encore inexplorée, on ne peut en effet présumer du pouvoir explicatif des variables auxiliaires. Celui-ci dépend du site, de la saison, de l'implantation des mesures, de la résolution spatiale des données auxiliaires.

Pour la cartographie de l'ozone dans un domaine de relief contrasté, l'altitude semble toutefois une variable pertinente.

Pour la cartographie du dioxyde d'azote, il est plus délicat de formuler des recommandations. Notons que les émissions de NOx ne présentent pas toujours le pouvoir explicatif attendu.

Les propriétés de stationnarité de la variable régionalisée étudiée doivent être également considérées afin de choisir la méthode d'estimation la plus adéquate (avec ou sans dérive). En particulier, le risque de rencontrer des processus non stationnaires s'accroît avec la dimension du domaine.

■ de contrôler la qualité du modèle ajusté et du voisinage d'estimation à l'aide de la validation croisée.

On examinera les différentes statistiques d'erreur (moyenne et variance de l'erreur ou de l'erreur standardisée, erreurs relatives d'estimation) et le nuage de corrélation entre les valeurs mesurées et estimées

Ces statistiques d'erreur indiquent l'adéquation du modèle dans les zones les plus riches en points de mesure. Aussi, pour une meilleure appréciation de la qualité du modèle, peut-il être intéressant d'effectuer la validation croisée sur différentes sélections de points de mesure.

■ de définir une maille d'estimation adaptée à l'échelle de la représentation.

Dans une cartographie régionale, une maille de 5 km de côté semble être un choix pertinent. Dans une cartographie à l'échelle urbaine, une maille de 500 m est satisfaisante.

3. dans l'interprétation des résultats :

■ d'examiner la carte de **l'écart-type de krigeage, qui est un indicateur de la précision de l'estimation**.

Cet indicateur dépend uniquement du modèle de variogramme et de la configuration spatiale des points de mesure. Rapportée à la carte d'estimation, la carte d'écart-type de krigeage renseigne donc, en moyenne dans le domaine d'étude, sur la précision de l'estimation. Elle fait également apparaître les variations spatiales de cette précision en fonction de la densité d'échantillonnage mais elle ne tient pas compte des variations de précision dues à la variabilité locale des concentrations.

Elle permet en première approche et moyennant certaines hypothèses d'estimer un intervalle de confiance autour de l'estimation mais elle ne suffit pas à quantifier rigoureusement l'incertitude. Le recours aux outils plus complexes de la géostatistique non linéaire se révèle nécessaire à cette fin. Il permet en outre de compléter l'information sur les concentrations par une estimation de la probabilité de dépassement de seuil.

■ d'avoir à l'esprit que le krigeage ne reproduit pas la situation réelle mais qu'il en offre une représentation lissée. L'usage de variables explicatives connues de façon dense, au moyen d'un cokrigeage ou un krigeage avec dérive externe, permet cependant de mieux restituer des variations locales de concentration et d'aboutir à des cartes plus détaillées.

Les méthodes de krigeage ordinaire, de cokrigeage et de krigeage avec dérive externe ont fait leur preuve dans la cartographie de la pollution d'agglomérations urbaines pour des polluants tels que le dioxyde d'azote, le benzène ou l'ozone. Elles tendent aujourd'hui à s'élargir à des domaines de grande superficie, larges de plusieurs dizaines voire centaines de kilomètres.

La limite de ces méthodes pour la surveillance de la qualité de l'air réside dans leur exigence en données d'entrée. Dans la plupart des régions, les réseaux de stations fixes ne sont pas suffisamment denses pour qu'à partir des mesures fournies en continu, on puisse établir des cartes moyennes sur des périodes déterminées (mois, saisons, années). Ainsi les cartes sont-elles le plus souvent fondées sur des données d'échantillonnage (tubes), donc sur une information partielle du point de vue temporel. En outre, les outils actuels de la géostatistique ne permettent pas d'intégrer des données complémentaires de moyens mobiles ou de stations fixes dans le modèle, afin d'enrichir cette information.

Outre l'usage de la géostatistique non linéaire, le principal axe de recherche est donc le couplage des aspects temporels et spatiaux,

- afin de définir des stratégies d'échantillonnage appropriées à la mise en œuvre des méthodes géostatistiques ;
- afin d'estimer, à partir des données disponibles, des cartes de concentrations moyennes avec une incertitude mieux maîtrisée.

4.3 METHODES DETERMINISTES

- ✓ Les méthodes déterministes permettent de reconstruire des champs de concentrations en tous points d'une grille de discrétisation par la résolution numérique des équations régissant les phénomènes physico-chimiques mis en jeu. Elles s'appuient pour cela sur les caractéristiques du site étudié, les données d'émission et les données météorologiques qui constituent les points d'entrée du modèle.
- ✓ Ces méthodes fournissent toujours une cartographie quel que soit le nombre de points de mesure disponible. Pour cela elles sont une alternative idéale dans les zones peu/pas couvertes par le réseau de mesure. Il n'y a pas de limitation dans l'extension du domaine d'étude. Par nature plus le pas de discrétisation est fin, meilleure est la résolution du modèle, à condition de disposer du même niveau de qualité pour les données d'entrée. Cela induit une réflexion sur la recherche du meilleur compromis entre précision et coût de mise en œuvre. Cette question n'est pas encore résolue.
- ✓ La principale limitation de ces méthodes est l'aspect temporel. En effet l'usage des modèles les plus élaborés (eulériens ou lagrangiens), induit un coût en temps de calcul encore prohibitif. Ainsi ces méthodes s'adaptent bien à la cartographie d'épisodes de pollution (quelques jours) mais demeurent moins appropriées pour le traitement de longues périodes. Cette application, outre un matériel informatique performant, nécessite de formuler un certain nombre d'hypothèses simplificatrices pour limiter les calculs ou d'avoir accès à des techniques de parallélisation. Des progrès récents sont néanmoins constatés dans ces voies.
- ✓ L'évaluation de l'incertitude des résultats fournis par ces modèles demeure également une question en suspens. Les sources d'incertitude sont nombreuses et les processus complexes mis en jeu induisent parfois des compensations d'erreur qu'il est délicat d'évaluer (pour l'ozone en particulier).
- ✓ L'incorporation de données de mesures pour analyser et retraiter, a posteriori ou durant les processus itératifs les résultats fournis par les modèles déterministes de qualité de l'air est une voie d'avenir pour améliorer la qualité des cartes produites par les modèles déterministes. Des méthodologies sont élaborées, mais elles demeurent encore délicates à utiliser en conditions opérationnelles. En revanche il faut s'attendre dans le futur à l'adoption de cette démarche de manière quasi-systématique.

5. REFERENCES

Arrêté du 17 mars 2003 relatif aux modalités de surveillance de la qualité de l'air et à l'information du public.

ARNAUD M. et EMERY X. , 2000. Estimation et interpolation spatiale - *Méthodes déterministes et méthodes géostatistiques*. Hermès Science Publication.

ASCOPARG, 2001. Mise en place d'une méthodologie pour la cartographie de l'ozone à l'échelle du département de l'Isère. Rapport d'étude (http://www.atmo-rhonealpes.org/ascoparg)

ASPA, UMEG, 2000. Analyse transfrontière de la qualité de l'air dans le Rhin supérieur. Programme INTERREG II.

ASPA, 2002. Diagnostic de la qualité de l'air sur l'agglomération de Mulhouse. Annexe au rapport final sur la répartition spatiale de la pollution atmosphérique (ASPA 02031901-I-D)- Méthodes d'interpolation spatiale.

ATMO PICARDIE, AIRPARIF, AIR NORMAND, OPAL'AIR, AREMA LM, AREMARTOIS, 2000. Campagne interrégionale d'étude de l'ozone et du dioxyde d'azote par tubes à diffusion passive, 26 juin-04 septembre 2000.

BISHOP T.F.A., McBRATNEY A.B., 2001. *A comparison of prediction methods for the creation of field-extent soil property maps*. Geoderma, 103. 149-160.

BLOND N., BEL L., VAUTARD R., 2002, Three dimensional ozone data analysis with an air quality model over Paris area. Article soumis à Atmospheric Environment

BOBBIA M., PERNELET V., ROTH C., 2001. L'intégration des informations indirectes à la cartographie géostatistique des polluants. Pollution Atmosphérique n° 170 - Avril-Juin.

BOBBIA M., MIETLICKI F., ROTH C., 2000. Surveillance de la qualité de l'air par cartographie : l'approche géostatistique. Poster INRETS 2000, 5-8 juin, Avignon, France

BOURENNANE H., KING D., COUTURIER A., 2000. *Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities*. Geoderma, 97, 255-271.

CASADO L.S., ROUHANI S., CARDELINO C.A., FERRIER A.J., 1994. Geostatistical analysis and visualization of hourly ozone data. Atmospheric Environment 28, n°12, 2105-2118.

CHRISTAKOS G., VYAS V.M., 1998. A composite space/time approach to studying ozone distribution over Eastern United States. Atmospheric Environment, Vol. 32, n°16, 2845-2857.

COYLE M., SMITH R.I., STEDMAN J.R., WESTON K.J., FOWLER D., 2002. Quantifying the spatial distribution of surface ozone concentration in the UK. Atmospheric Environment, vol. 36, 1013-1024.

DAESCU D.N., CHARMICAEL G.R., An adjoint sensitivity method for the adaptative location of the observations in air quality modeling, J. Of the Atmospheric Sciences, Vol. 60, n°2, pp 434-450, 2003

DELETRAZ G., DARBOS P., 2001. Modélisation statistique de la pollution azotée à proximité d'un axe routier et évaluation des incidences sur l'environnement. Application au site de Buriatou. Colloque Risques, octobre 2001, Besançon.

DERAISME J., BOBBIA M., 2003. L'apport de la géostatistique à l'étude des risques liés à la pollution atmosphérique. Environnement, Risque et Santé, Vol. 2 , n°3, mai-juin 2003

DIEM J., 2003. A critical examination of ozone mapping from a spatial-scale perspective. Environmental Pollution, 125, 369-383

Directive du Conseil n° 1999/30/CE du 22 avril 1999 relative à la fixation de valeurs limites pour le dioxyde de soufre, le dioxyde d'azote, les oxydes d'azote, les particules et le plomb dans l'air ambiant

Directive 2000/69/CE du Parlement et du Conseil du 16 novembre 2000 concernant les valeurs limites pour le benzène et le monoxyde de carbone

Directive 2002/3/CE du Parlement européen et du Conseil du 12 février 2002 relative à l'ozone dans l'air ambiant

FOUQUET C.(de), Etude sur la réalisation de cartographies de la qualité de l'air dans les zones peu/pas couvertes par les réseaux de stations fixes à l'aide de méthodes géostatistiques (complément d'étude et synthèse), Ecole des Mines de Paris, Rapport N-9/03/G, juin 2003

FOUQUET C.(de), Méthodologie de cartographie de la concentration annuelle de NO_2 sur l'agglomération de Mulhouse, Rapport d'avancement N-6/03/G, Ecole des Mines de Paris, avril 2003

FOUQUET C. (de), 1997. Influence de la méthode d'estimation et de la maille de reconnaissance sur la quantification des pollutions : étude méthodologique à 2D. In Nicolas (ed.), Echantillonnage et environnement. Liège : Cebedoc, pp. 39-63.

GEORGOPOULOS P.G., PURUSHOTAMAN V., CHIOU R., 1997. Comparative evaluation of methods for estimating potential human exposure to ozone : photochemical modeling and ambient monitoring. Journal of Exposure and Environmental Epidemiology, Vol. 7, N°2

GRANCHER D., BEL L., VAUTARD R., 2003. Cartographie et prévision des champs de Laboratoire de statistique d'Orsay- CNRS- LCSQA.

Groupe de Travail des AASQA "Echantillonneurs passifs pour le dioxyde d'azote". Guide version 1, octobre 2001

HONORÉ C., MALHERBE L., 2003. Application des modèles grande échelle à la problématique régionale: cas de l'ozone. Rapport LCSQA.

INERIS (CARDENAS G., MALHERBE L.), 2003. Evaluation des incertitudes associées aux méthodes géostatistiques. Rapport LCSQA

INERIS (CARDENAS G., MALHERBE L.), 2002. Représentation de la qualité de l'air dans les zones peu ou pas couvertes par les stations de mesure fixes: partie II, application à la problématique d'une association. Rapport LCSQA

JEANNÉE N., FANGEAT E., BA M., 2003. Contributions pratiques d'une géostatistique raisonnée en environnement: méthodes et application à la cartographie nationale de la pollution par l'ozone en France, Géo-Evénement, 3-5 mars 2003, Paris

KYRIAKIDIS P.C., KIM J., MILLER N.L., 2001. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. J. American Meteorlogical Society, 40, 1855-1877.

LAJAUNIE C, WACKERNAGEL H, BERTINO L, 2001. Geostatistical Normalization: Case Studies, Technical Report N-31/01/G, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau.

LE DIMET F.X., NAVON I.M., DAESCU D.N., *Second order information in data assimilation*, Monthly Weather Review, Vol 130, n°3, pp 629-648, 2002

LEFOHN A.S., KNUDSEN H.P., LOGAN J.A., SIMPSON J., BHUMRALKAR C., 1987. An evaluation of the kriging method to predict 7-h seasonal mean ozone concentrations for estimating crop losses. J. Air Pollut. Control. Assn., 37(5), 595-602.

LEFOHN A.S., KNUDSEN H.P., McEVOY L.R., 1988. The use of kriging to estimate monthly ozone exposure parameters for the Southern United States. Environmental pollution, 53, 27-42.

LE LOC'H G., Etude exploratoire du dioxyde d'azote sur l'agglomération de Montpellier, Ecole des Mines de Paris, Rapport N-8/03/G, juin 2003

LIU S. L.-J., ROSSINI A.J., 1996. Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto - A pilot study. Environment International, N° 6, 667-692.

MARCOTTE D., 2003. Cours GLQ3401 : Géologie et géostatistique minières (partie géostatistique). http://geo.polymtl.ca/~marcotte/

MOUSSIOPOULOS N., 1998, Air quality in Athens : long term trend and expected evolution in 2004, Air Pollution VI, WIT Press, Computational Mechanics Publications, pp 425-434

NF ENV 13005. Normes fondamentales. Guide pour l'expression de l'incertitude de mesure, AFNOR, 1999.

NF ISO 13752 (novembre 1998). Qualité de l'air - Évaluation de l'incertitude d'une méthode de mesurage sur site en utilisant une seconde méthode comme référence.

NIKIFOROV S.V., AGGARWAL M., NADAS A., KINNEY P.L., 1998. *Methods for spatial interpolation of long-term ozone concentrations*. Journal of Exposure Analysis and Environmental Epidemiology, vol. 8, n°4, 465-481.

PHILLIPS D.L., LEE E. H., HERSTOM A. A., HOGSETT W. E., TINGEY D.T., 1997. *Use of auxiliary data for spatial interpolation of ozone exposure in southearn forests*. Environmetrics, 8, 43-61.

PURUSHOTHAMAN V., GEORGOPOULOS P.G., 1999. Integrating physico-chemical modeling, geostatistical techniques et geographical information systems for ozone exposure assessment. Ozone Research Center, Technic Report, ORC-TR99-01, May 99.

RIVOIRARD J., 2001. Which models for collocated cokriging ? Mathematical Geology, vol 33, n°2, pp 117-131.

ROTH C., DÉGARDIN D., 2001. Interpreting the results of diffusive sampling campaigns with respect to yearly standards. International Conference Measuring Air Pollutants by Diffusive Sampling, Montpellier, France, 26-28 septembre 2001.

ROTH C., 1999. Etude géostatistique des données de pollution des agglomérations du Havre et de Rouen. Phase 1 : traitement des données des tubes à diffusion ?

ROTH C., 2000. Etude géostatistique des données de pollution des agglomérations du Havre et de Rouen. Phase 2 : traitement des données de capteurs et étude de faisabilité sur la représentativité des stations ?

ROUÏL L., WROBLEWSKI A., 2002. Guide méthodologique en modélisation déterministe. Rapport LCSQA 2001.

ROUÏL L., 2001. Méthodologies d'évaluation des modèles et de l'incertitude. Rapport LCSQA, convention 42/2000.

SAITO H., GOOVAERTS P., 2001. Accoutnig for source location and transport direction into geostatistical prediction of contaminants. Environmental Science and Technology, 35, 4823-4829.

SCHAUG J., IVERSEN T., PEDERSEN U., 1993. Comparision of measurements and model results for airborne sulfur and nitrogen components with kriging. Atmospheric Environment, 27A, N°6, 831-844.

Table Ronde des Utilisateurs Isatis, 11 et 12 mars 2002. Présentations techniques, Géovariances?

STEDMAN J. Revised high resolution maps of background levels of air pollutants, report AEA Technologies, 1998, www.aeat.co.uk/netcen/airqual/reports/jsmaps/mphead.html

TAYANÇ M., 2000. *An assessment of spatial and temporal variation of sulfur dioxide levels over Istanbul, Turkey*. Environmental pollution, 107, 61-69.

TILLÉ Y., 2001. Théorie des Sondages - *Echantillonnage et estimation en populations finies*. Dunod?

TREMBACK C.J, WALKO R.L. 2000, The regional Atmospheric Modeling System (RAMS) : development for parallel Processing computer architectures, http:// atmet.com/html/papers/parallel.pdf

UNG A., WEBER C., PERRON G., HIRSCH J., KLEINPETER J., WALD L., MANDIN T., *Air pollution mapping over a city-virtual stations and morphological indicators,* 10th Int. Symposium "Transport and Air Pollution", September 2001, Boulder, Colorado

VAN LOON M., HEEMICK A.W, 1997, Kalman filtering for non linear atmospheric chemistry models : first experiences, CWI report

VARNS J.L., MULIK J.D., SATHER M.E., GLEN G., SMITH L., STALLINGS C., 2001. Passive sampling of ambient, gaseous air pollutants : an assessment from an ecological perspective. Environmental Pollution, Vol.107, 31-45.

WACKERNAGEL H., 1992. Cours de géostatistique multivariable. Ecole des Mines de Paris, 62 p.

WACKERNAGEL H., BERTINO L., SIERRA J.P., GONZÁLEZ DEL RÍO, 2002. *Multivariate kringing for interpolating with data from different sources*. In Quantitative Methods for Current Environmental Issues, Anderson et al. (eds), pp 57-75, Springer-Verlag, Londres

WACKERNAGEL H., 2002. Geostatistical normalization of air pollution transport model output and station data using ISATIS, IMPACT project report n°4 (ref N-20/02/G), http://cg.ensmp.fr

WOLKE R., KNOTH O., HELLMUTH O. , SHRODER W., WEICKERT J., 2000, load balancing in the parallel model system LM-MUSCAT for Multiscale Chemistry Transport simulations, contribution au projet GLOREAM (EUROTRAC 2)

6. LISTE DES ANNEXES

Repère	Désignation précise	Nb/N°pages
A	Présentation théorique des différentes méthodes d'estimation de la géostatistique linéaire	8
B	Ajustement du variogramme à l'origine et prise en compte de la variance de l'erreur de mesure	17
C	Principe de l'analyse en composantes principales et illustration	2

Annexe A

Présentation théorique des différentes méthodes d'estimation de la géostatistique linéaire

Annexe B

Ajustement du variogramme à l'origine et prise en compte de la variance de l'erreur de mesure

Annexe C

Principe de l'analyse en composantes principales et illustration

JP Chollet (LEGI)

LISTE DE DIFFUSION

Nom	Adresse/Service	Nb
BIRCK	Dossier maître	1
DOC		1
ROUÏL		1
MALHERBE		1
MATE		5
ADEME		2
EMD		1
LNE		1

TOTAL **13**

PERSONNES AYANT PARTICIPE A L'ETUDE

Travail	Nom	Qualité	Date	Visa

Fin du Complément non destiné au client