Evaluation des incertitudes associées aux méthodes géostatistiques: ANNEXE B

Analyse d'un jeu des données d'ozone fourni par AIRNORMAND

Laboratoire Central de Surveillance de la Qualité de l'Air

Giovanni CARDENAS
Unité Informatique et instrumentation pour l'environnement (2IEN)

Laure MALHERBE
Unité Modélisation et analyse économique pour la gestion des risques (MECO)

Direction des Risques Chroniques (DRC)
Décembre 2003

TABLE DES MATIÈRES

LISTE DE FIGURES

LISTE DES TABLEAUX

1 Contexte de l'étude

Le but de ce travail est de montrer sur un cas réel comment on peut prendre en compte les différentes sources d'incertitude tout au long de l'analyse et de la modélisation géostatistiques et comment il est possible d'évaluer l'incertitude finale sur les estimations.

Dans un premier point sont montrées les applications de la géostatistique linéaire qui permet d'effectuer l'estimation d'une variable à l'aide de l'algorithme du krigeage. L'apport de la géostatistique non linéaire et des simulations conditionnelles dans l'évaluation d'intervalles de confiance fait l'objet d'un second point.

La zone d'étude s'étend sur une surface d'environ 56 000 Km^2. Elle se situe dans le Nord de la France et englobe les départements du Nord-Pas-De-Calais, de Picardie, de Normandie et d'Île de France.

La surveillance de la qualité de l'air dans cette région est assurée par cinq AASQA : AREMA LM, OPAL'AIR, AREMARTOIS, AIR NORMAND, ATMO PICARDIE et AIRPARIF.

Les données exploitées dans ce travail ont été fournies par l'association AIR NORMAND . Elles sont issues d'une campagne de mesure de l'ozone réalisée pendant l'été 2000 du 26 juin au 4 septembre (soit 10 semaines de mesure) et coordonnée par ATMO PICARDIE. La semaine du 26 juin au 3 juillet a été choisie à titre d'illustration.

Les détails concernant l'échantillonnage et le traitement des données sont consignés dans le rapport: « Campagne interrégional d'étude de l'ozone et du dioxyde d'azote par tubes à diffusion passive (26 juin – 04 septembre 2000 » de l'association Atmo Picardie.

La plupart des sites ont été implantés dans des zones rurales. Ils se répartissent dans l'espace selon un maillage prédéfini de pas égal à 25 km. Dans les zones urbaines, les mailles sont plus resserrées : le pas d'échantillonnage se réduit à 12.5 km pour l'Ile de France et à 6.25 km pour les agglomérations des autres régions.

Le nombre total de sites est ainsi de 229et chaque site étant équipé d'au moins deux échantillonneurs.

En plus des sites ruraux et urbains (ou périurbains) la zone d'étude comprend quelques points de mesure localisés sur le littoral nord. On verra par la suite que le type de site a une influence sur la variabilité spatiale de la concentration d'ozone.

De surcroît, des tubes passifs ont été installés à proximité de quelques stations fixes qui sont pour la plupart localisées dans des zones urbaines ou périurbaines. Ces sites sont appelés sites étalons car ils permettent d'établir une relation entre ces deux types de mesure.

Les échantillonneurs passifs mesurent l'absorbance du polluant (l'ozone dans ce cas) durant une période de temps donnée (une ou deux semaines), alors que les capteurs donnent la moyenne de la concentration tous les quarts d'heure. La méthode qui est employée pour passer des valeurs d'absorbance à des valeurs de concentration moyenne par semaine et pour calculer l'incertitude associée (ou variance de l'erreur de mesure) utilise les données des stations fixes. Elle fait appel à une méthode de régression orthogonale pondérée appelée régression de modèle II (par moindres carres généralisés) et à la loi de propagation de l'incertitude. Sur ce point on peut se référer au rapport mentionné précédemment. En appliquant cette méthodologie, les AASQA impliquées dans cette campagne de mesure ont obtenu pour chaque semaine et pour chaque site d'échantillonnage la concentration moyenne de l'ozone (C est) et la variance de l'erreur de mesure (VEM).

On remarque que les données ont été obtenues par des méthodes rigoureuses et qu'elles se répartissent dans l'espace de manière satisfaisante. Ces caractéristiques facilitent l'analyse spatiale de la concentration. En revanche l'absence de toute donnée de variable auxiliaire ne permet pas d'effectuer une analyse multivariable

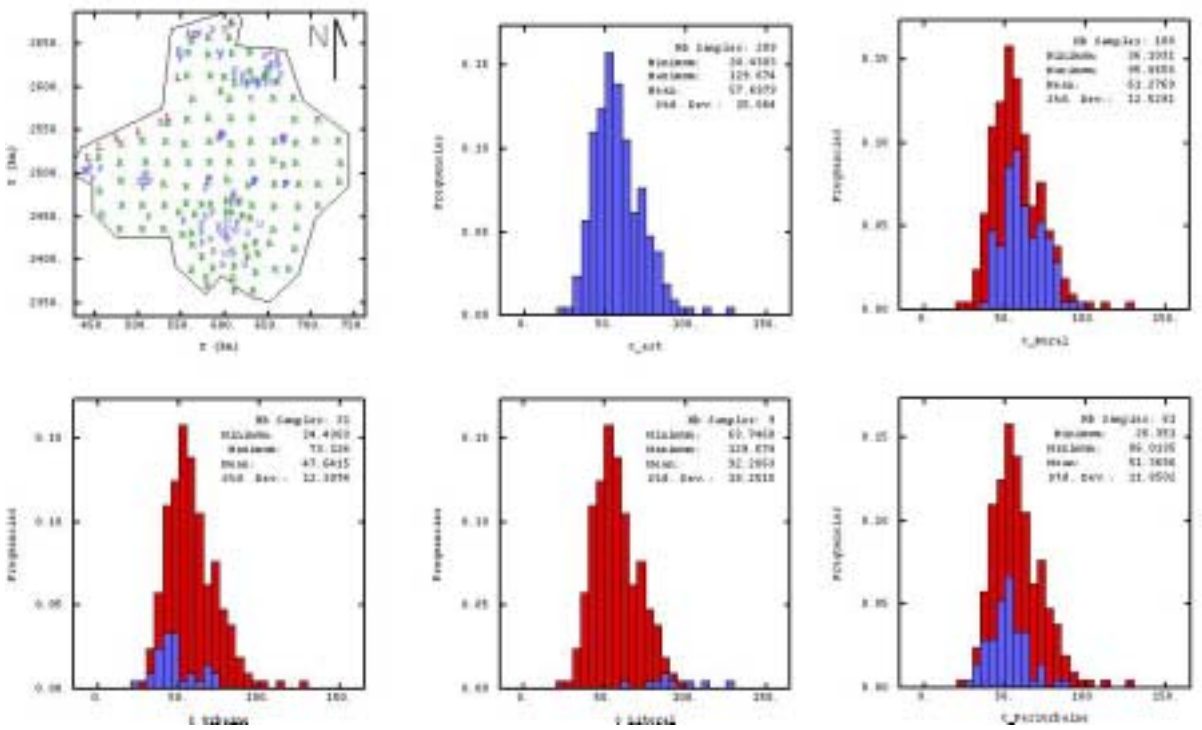
L'ensemble du travail présenté ci-après a été réalisé à l'aide du logiciel ISATIS, version 4.1.3 (Géovariances).

2 Analyse préliminaire des données

La première semaine (du 26 juin au 3 juillet) a été choisie pour la premier partie de cette étude. Les données correspondantes se composent de 209 mesures de tubes validées et de 32 mesures de stations fixes (voir Figure 1, Figure 2 et Figure 3).

L'implantation géographique des sites de mesure est présentée à la Figure 1. Les zones où la densité de tubes est plus grande correspond aux agglomérations (lettres P et U en bleu) - Paris, Lille Rouen, Le Havre , etc-. Les sites ruraux sont signalés par la lettre R (en vert), ceux du littoral par la lettre L e(n rouge).

Pour l'ensemble des sites, la moyenne des concentrations est de 57.7 µg/m^3. Ces concentrations sont comprises entre 24 à 130 µg/m^3, mais la moitié d'entre elles se trouvent dans un intervalle allant de 46 µg/m^3 (quantile 25%) à 66 µg/m^3 (quantile 75%).



Dans la carte :
L en rouge : Sites du littoral
U et P en bleu : Sites urbains et périurbains
R : Sites ruraux

Dans les histogrammes de types de sites (R,U,L,P):
En Rouge : allure de la distribution de fréquences de l'ensemble des sites.
En bleu : La partie correspondant à chaque type de site.

Figure 1: Implantation géographique des tubes passifs, statistiques et histogrammes des données par type de site.

Plus de la moitié des mesures (108 données) correspondent à des sites ruraux, avec une moyenne de 61 µg/m^3. Cette moyenne est plus élevée que celle des sites urbains et périurbains (47.6 µg/m^3 et 51.4 µg/m^3). Cette différence peut s'expliquer par les phénomènes de formation et de transport de l'ozone.

De même la moyenne élevée pour les 9 sites du littoral (92 µg/m^3) s'explique par la proximité de la mer, où les dépôts d'ozone sont moindres.

Les mesures des tubes situés à proximité des 32 analyseurs fixes sont comparées aux concentrations mesurées par ces analyseurs, ces derniers sont principalement implantés dans les agglomérations.

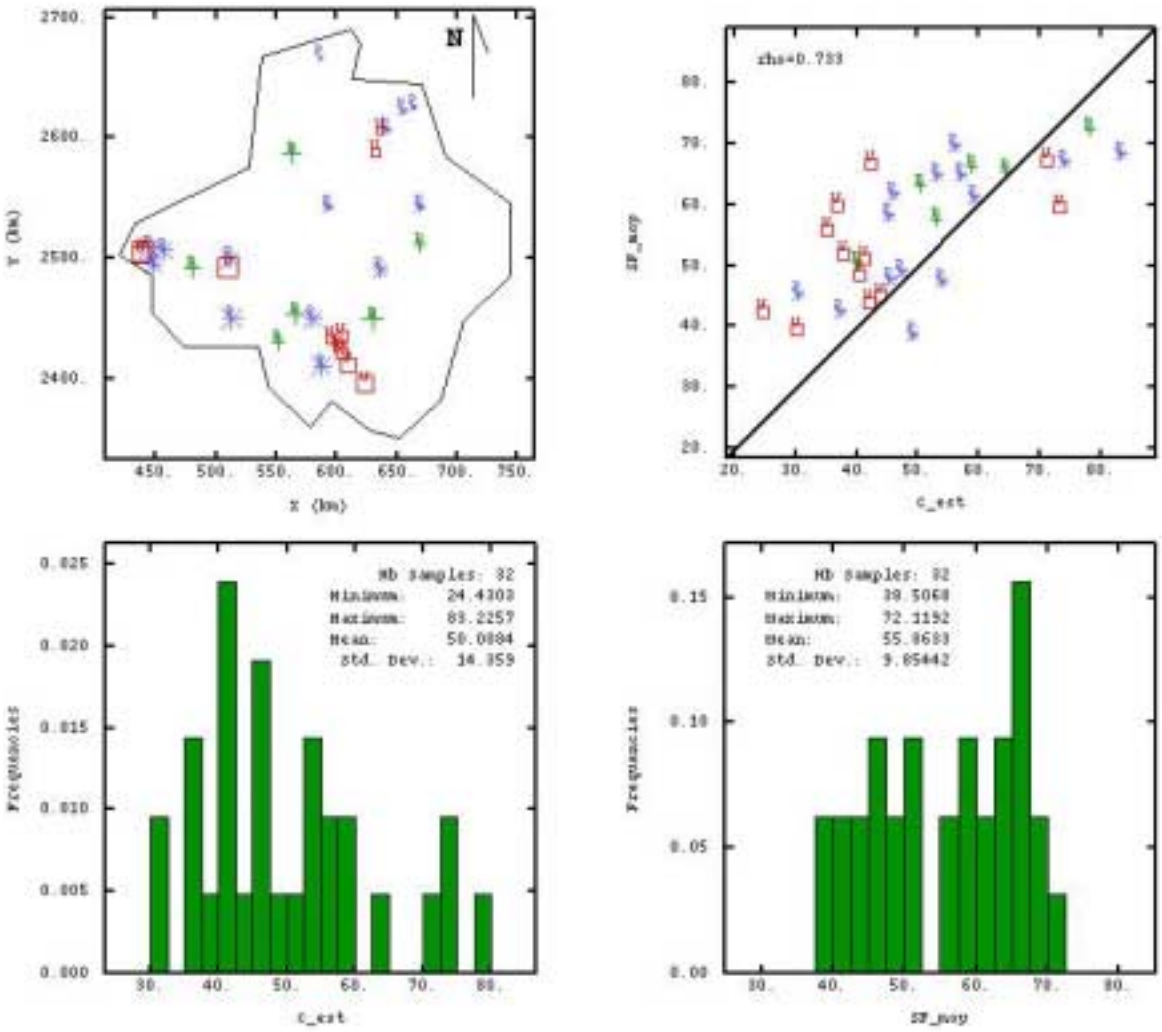


Figure 2: Comparaison entre concentration des stations fixes et des tubes passifs

L'écart observé entre les mesures obtenues par deux techniques différentes indique la présence d'incertitudes sur les données de concentration. La méthode employée par les AASQA pour quantifier l'ampleur de ces incertitudes est décrite dans le rapport de campagne mentionné plus haut. Une variance de l'erreur de mesure est ainsi associée à chaque donnée de tube. Les statistiques de cette variance sont présentées à la Figure 3.

L'indicateur d'incertitude est plus précisément l'écart-type de l'erreur de mesure mais pour les besoins de la modélisation géostatistique, on utilise plutôt la variance de l'erreur de mesure qui est égale au carré de l'écart.

Le nuage de corrélation entre l'écart-type de l'erreur de mesure et la concentration estimée « C est » montre une relation linéaire. Ainsi, plus la valeur d'ozone donnée par les tubes est élevée, plus cette valeur est incertaine.

Par suite, les observations faites pour la variable de concentration valent aussi pour la VEM, à savoir que la VEM est plus faible dans les agglomérations que dans les zones rurales et qu'elle prend ses valeurs les plus fortes sur le littoral nord.

En effet il faut rappeler que les deux variables (EEM et « C est ») sont calculées à l'aide des coefficients (b_0 et b_1) déduits de la droite d'étalonnage absorbance - lecture de capteurs, et que ces coefficients sont obtenus en utilisant la méthode d'ajustement orthogonale pondérée (moindres carres généralisés).

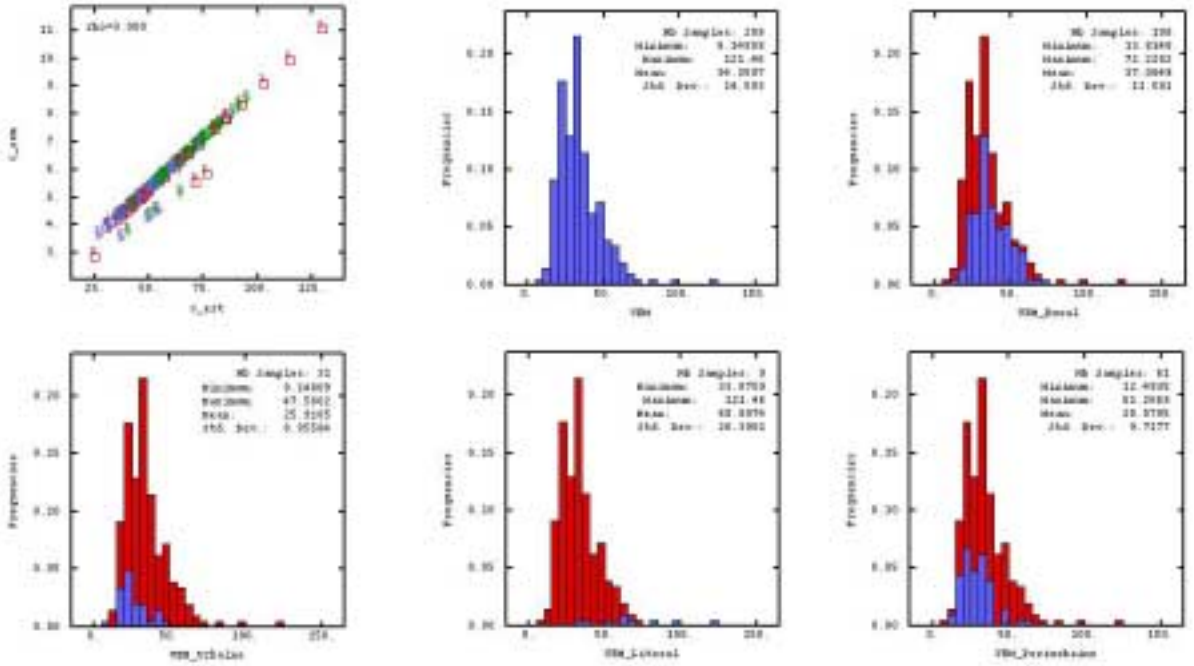


Figure 3: Statistiques et histogrammes de la VEM par type de site. Régression linéaire entre les données des tubes passifs et l'écart de l'erreur de mesure (EEM)

3 Analyse spatiale des données

Cette analyse a pour fin de cerner les caractéristiques spatiales du phénomène de pollution étudié et de décrire la structure de ce phénomène à l'aide du variogramme.

Nous insisterons dans notre propos sur la façon de prendre en compte les différentes sources d'incertitude (données de concentration, paramètres du modèle) tout au long de cette analyse.

La nuée variographique, qui est le nuage des écarts quadratiques $½[Z(x)-Z(x+h)]^2$ en fonction de la distance h, permet de visualiser l'ensemble des points qui contribuent au calcul du variogramme expérimental et les couples de sites (x, x+h) qui leur correspondent.

En théorie le nombre total de couples qui peuvent être formés avec n sites de mesure est de n*(n-1)/2. Dans notre cas, on obtiendrait pour les 209 sites un nuage de 736 points. En pratique, on ignore l'influence des paires de points séparées par une distance supérieure à la moitié du champ d'étude .En effet comme il est expliqué par la suite, c'est la structure à petite distance du variogramme qui importe le plus. D'autre part, plus la distance augmente, plus le nombre de couples qui peuvent être constitués diminue . Dans notre cas on étudiera les deux cents premiers kilomètres du variogramme.

Les variogrammes peuvent être calculés dans plusieurs directions afin de vérifier si la variabilité spatiale des concentrations a ou non un caractère isotrope. En pollution de l'air, on pense souvent à l'effet que peuvent avoir la direction et la vitesse du vent. Toutefois, l'influence des ces deux paramètres n'est pas toujours aisée à évaluer à l'échelle d'une semaine ou de quinze jours. Une rapide analyse du variogramme dans plusieurs directions n'a pas révélé d'anisotropie apparente ; en conséquence les analyses seront réalisées sur des variogrammes isotropes

3.1 Etude de la nuée variographique

Pour quelques couples de points l'écart quadratique entre les concentrations d'ozone est supérieur à 4000 $(\mu g/m^3)^2$ (Figure 4), alors que la variance des données est d'à peine 245 $(\mu g/m^3)^2$. Comme on pouvait s'y attendre, ces couples sont formés par la conjugaison des trois valeurs maximales du littoral avec des valeurs plus faibles mesurées dans les agglomérations (ils sont marqués en bleu dans la nuée variographique et dans la carte d'implantation des données). Pour la plupart des autres points, les écarts quadratiques n'excèdent pas 1500 $(\mu g/m^3)^2$ (écarts de concentration inférieurs à 40 $\mu g/m^3$).

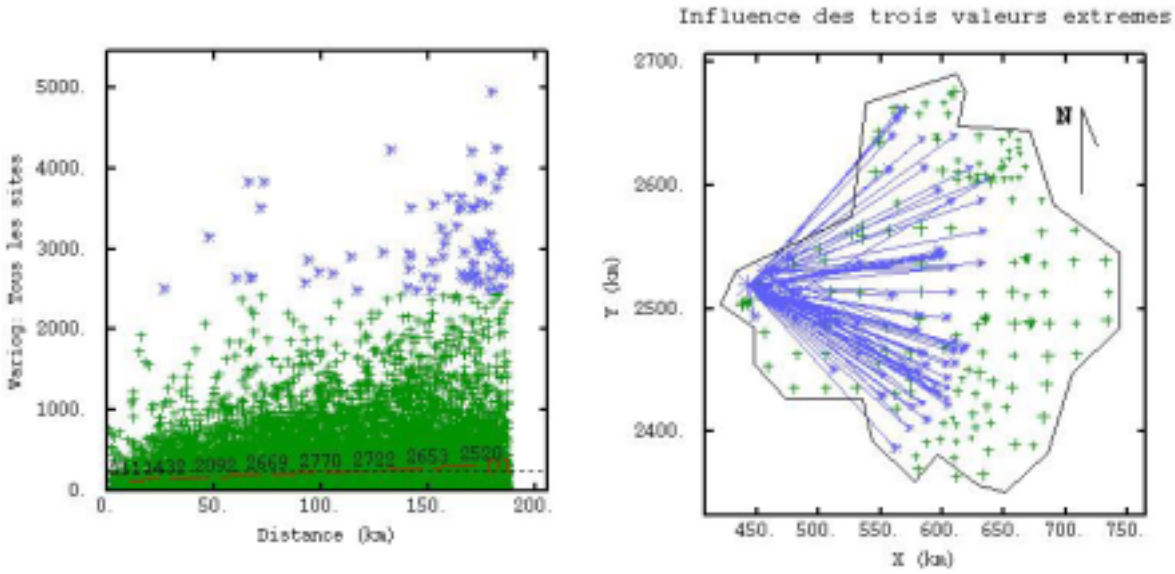


Figure 4: Nuage variographique et carte de localisation de l'ensemble des mesures, en bleu les couples formés avec les trois valeurs les plus fortes localisées dans le littoral.

D'après la Figure 4, il ne paraît pas judicieux de mêler les différents types de sites dans l'analyse spatiale. Aussi les données sont-elles divisées en trois groupes.

Les données urbaines et périurbaines, qui sont plus proches les unes des autres, prennent en moyenne les valeurs les plus faibles. Ces caractéristiques vont nous permettre d'évaluer la variabilité de l'ozone à petite distance (une trentaine de kilomètres) et de modéliser ce polluant dans les agglomérations.

Avec les données rurales, qui sont bien reparties dans le domaine d'étude, il sera possible de quantifier la variabilité du polluant à une plus grande échelle (quelques centaines de kilomètres). Ces données caractérisent la pollution dite de fond.

Les données du littoral, peu nombreuses, ne permettent pas d'analyser en détail la structure spatiale du phénomène de pollution dans cette région. Leurs valeurs numériques seront prises en compte au moment d'estimer les concentrations dans cette partie du domaine.

La nuée variographique a été calculée avec les données des 108 sites ruraux (108*107/2=5 778 couples possibles) jusqu'à une distance de 200 Km (Figure 5).

La variance des données rurales est de 157 $(\mu g/m^3)^2$ mais le variogramme peut atteindre des valeurs dix fois supérieures (1500 $(\mu g/m^3)^2$). Les sites responsables de ces valeurs sont localisés dans tout le domaine.

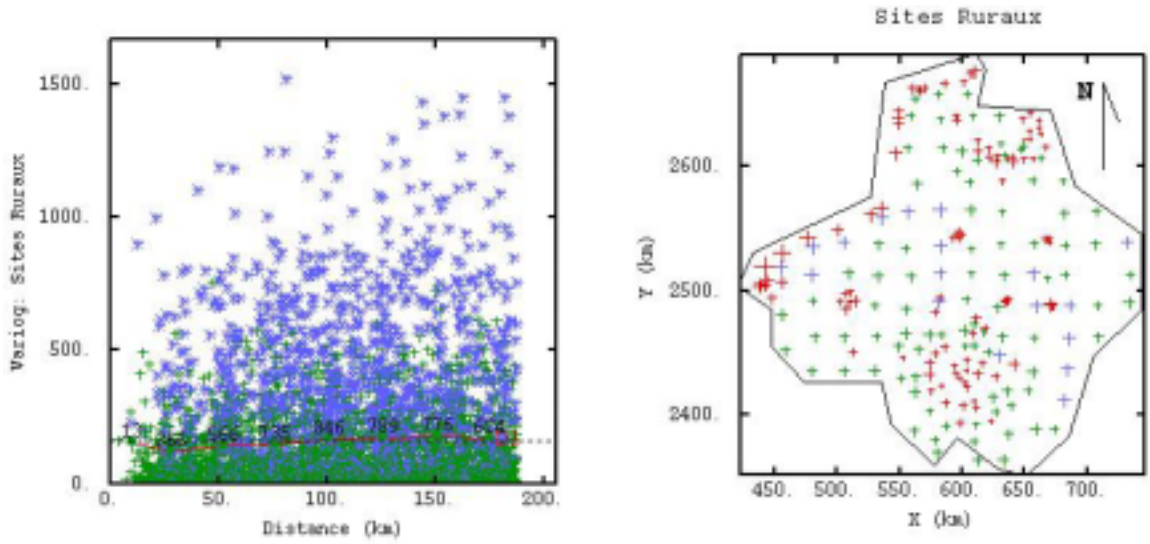


Figure 5: Nuage variographique de mesures des sites ruraux, et carte d'implantation : croix vertes et bleues (les points bleus correspondent aux valeurs des concentrations les plus fortes pour les mesures rurales)

3.2 Etude du variogramme à courte distance :

Les variogrammes expérimentaux sont obtenus en faisant la moyenne par classe de distance de la nuée variographique. Ils ont été calculés pour chaque type de site jusqu'à une distance de 30 km (Figure 6).

En théorie, la connaissance du variogramme à courte distance est importante du fait que l'allure du variogramme au voisinage de l'origine reflète la continuité et la régularité spatiales de la variable régionalisée. L'ajustement du modèle variographique à l'origine se révèle influent sur les résultats du krigeage et même, dans certains cas, sur la stabilité numérique du système de krigeage.

En pratique, les mesures disponibles ne sont souvent pas en nombre suffisant pour que l'on puisse caractériser le variogramme à ces faibles distances. Pour le variogramme rural, on ne dispose pas de données distantes de moins de 10 km. Pour celui des agglomérations, les données ne sont pas distantes de moins de 5 km.

D'autre part, les variogrammes des sites urbains et périurbains sont proches, indiquant par là une variabilité similaire au sein des agglomérations (entre 5 et 30 km). En revanche la variabilité en zone rurale est supérieure à celle qui est observée en zone urbaine, notamment pour des distances comprises entre 10 et 20 km.

Quand on prend en compte l'ensemble des mesures, la variabilité augmente logiquement, un certain nombre de couples de données étant obtenus par différents types de sites de mesure.

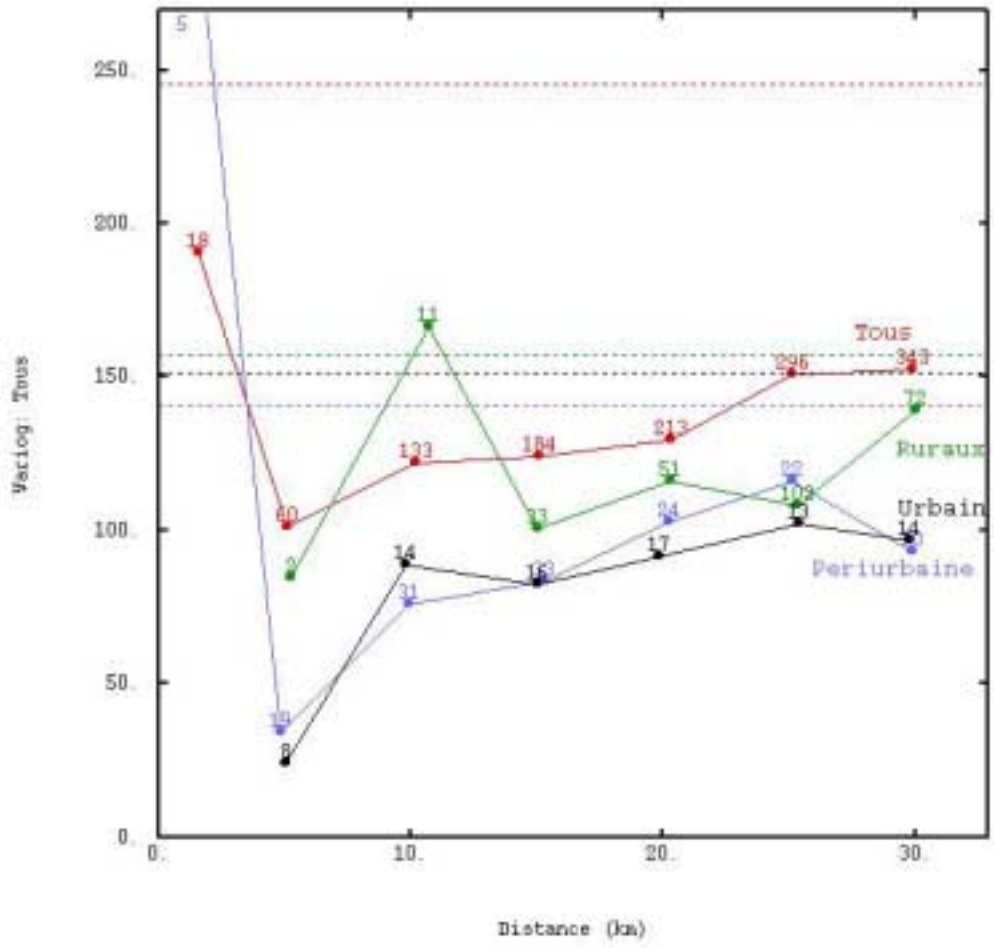


Figure 6: Variogrammes expérimentaux à petites distances (de 0 à 30 Km) pour les sites ruraux, périurbains et urbains et pour l'ensemble des sites (littoral inclus)

Le variogramme sert en outre à déterminer le degré de stationnarité de la variable régionalisée étudiée.

L'application des algorithmes géostatistiques exige en effet que soit vérifiée une hypothèse de stationnarité relative à cette variable.

Les variogrammes expérimentaux associés à chaque type de site ont été calculés jusqu'à une distance de 200 km (Figure 7).

Les valeurs extrêmes influent grandement sur le variogramme expérimental de l'ensemble de mesures. Ce dernier est presque linéaire, et par suite non stationnaire. Cependant il suffit de retirer trois des plus fortes valeurs pour qu'il se stabilise à un palier (ce qui correspond à une portée d'à peu près 150 km).

Les variogrammes des sites urbains et périurbains présentent un décrochement entre 50 et 150 km. Ce brusque changement peut s'expliquer par le fait qu'entre 50 et 150 km les couples de données sont formés par des sites de mesure implantés dans des villes différents, alors qu'à des distances inférieures, la plupart des couples sont

constitués par des sites de mesure d' une même agglomération. Ce résultat montre que les concentrations sont plus ou moins variables selon les agglomérations.

Le premier point significatif du variogramme expérimental des sites ruraux est associé à des distances entre données d'environ 20 à 30 km. Ce variogramme augmente ensuite lentement. Il devient égal à la variance des données (157 $(\mu g/m^3)^2$) autour de 100 km et poursuit sa croissance jusqu'à un palier d'environ 175 $(\mu g/m^3)^2$ qu'il atteint à une distance 125 km., Le nombre de couples de données par classe de distance augmente aussi jusqu'à une distance de 100 km (846), puis se met à diminuer.

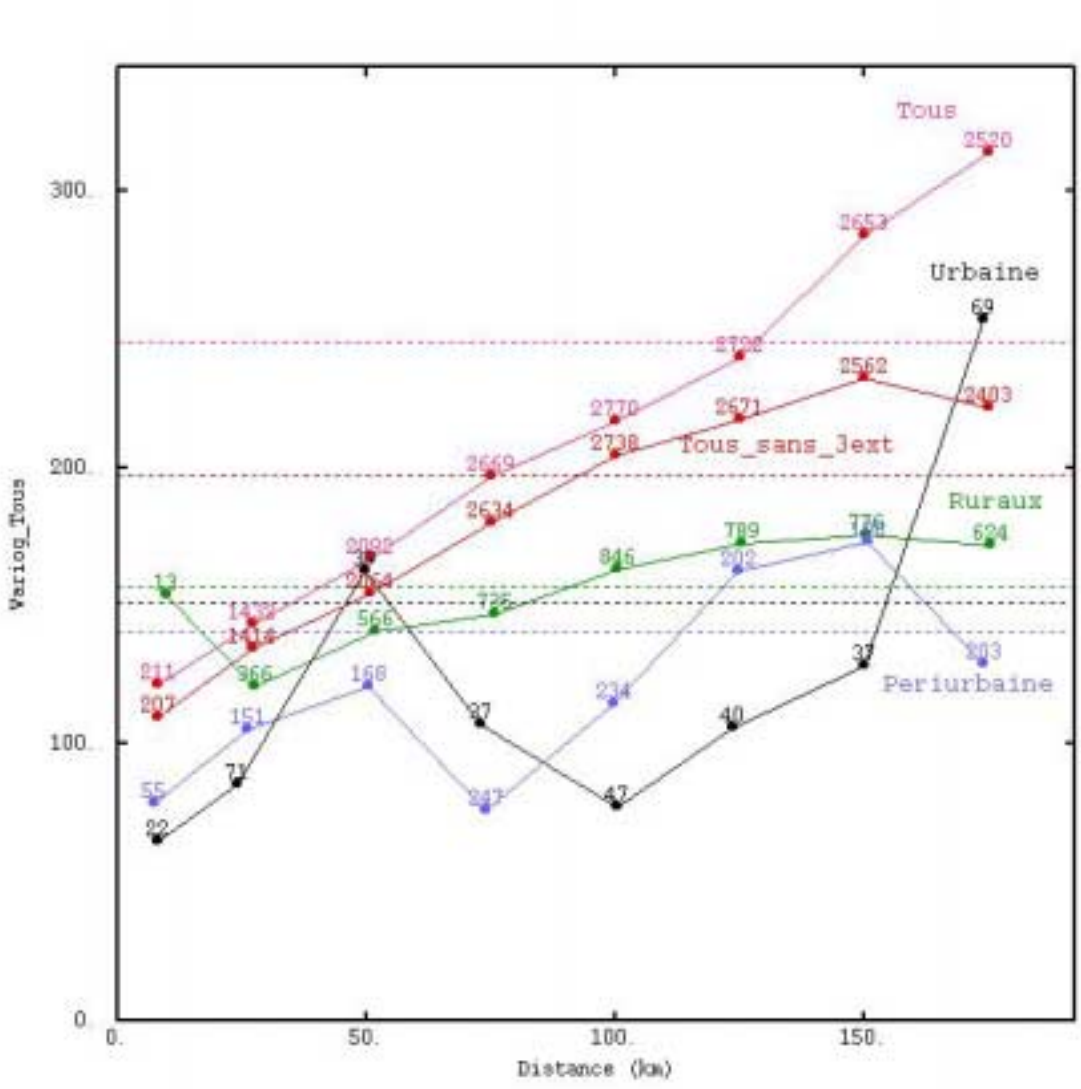


Figure 7: Variogrammes expérimentaux à grandes distances (de 0 à 200 Km) pour les sites ruraux, périurbains et urbains et pour l'ensemble des sites (littoral inclus) avec et sans les trois valeurs extrêmes

Théoriquement la valeur du variogramme à l'origine doit être zéro mais si on extrapole l'ensemble des variogrammes expérimentales de la Figure 6 ou de la Figure 7 à l'axe Y on obtient des valeurs différentes de zéro, ce saut à l'origine représente « l'effet de pépite ».

Dans notre cas on verra que comme on dispose des variances d'erreur de mesure non constantes, l'effet de pépite sera égal à la moyenne de ces variances (35).

L'étude conduite dans les parties 2 et 3 a permis de dégager les principales caractéristiques des données de concentration disponibles :

- différences entre les données de type urbain, périurbain, rural et littoral
- stationnarité
- régularité et variabilité à diverses échelles
- zone d'influence des échantillons

Ces caractéristiques – que la qualité de l'échantillonnage et le grand nombre de données ont rendu évidentes- permettront de sélectionner la méthode la plus appropriée pour la modélisation géostatistique.

4 Modélisation du variogramme

Le choix du modèle de variogramme est très important car le résultat obtenu dépend de la valeur de ce modèle. Seule l'expérience peut nous indiquer les modèles mathématiques les plus adaptés à tel ou tel type de régionalisation. Nous verrons que le facteur déterminant est le comportement du variogramme à petite distance,. Or, dans la plupart des cas, il s'agit justement de la zone inaccessible expérimentalement. En définitive, la modélisation de cette dans cette zone (liée à l'effet de pépite) sera l'élément décisif.

Le modèle qui a été ajusté est une superposition de structures de portées différentes. Ce type de variogramme est appelé « ***variogramme gigogne*** ».

Comme il a été mentionné, l'effet de pépite est supposé égal à la variance moyenne de l'erreur de mesure. On sait que la pollution par l'ozone est un phénomène de grande échelle, peu variable sur de courtes distances. Ces caractéristiques sont souvent modélisées par des schémas paraboliques à l'origine, tel un modèle cubique ou un modèle gaussien. Dans le cas présent, un modèle composé de deux structures gaussiennes de portées égales à 26 km et 136 km a été ajusté.

En géostatistique stationnaire, le variogramme ($\gamma(h)$) est directement lié à la covariance (C(h)).

$$\gamma(h)_{théorique} = Modèle \quad Gaussien$$

$$\gamma(h)_{théorique} = palier * \left\{1 - \exp\left[-\left(\frac{1.73 * h}{portee}\right)^2\right]\right\}$$

$$\gamma'(h) = Modèle \quad gigogne \quad ajusté = \sum [effet \quad de \quad pepite + \gamma(h)]$$

$$\gamma'(h) = 35 + 81 * \left\{1 - \exp\left[-\left(\frac{1.73 * h}{26}\right)^2\right]\right\} + 59 * \left\{1 - \exp\left[-\left(\frac{1.73 * h}{136}\right)^2\right]\right\}$$

$$\gamma(0) = 0, \quad alors : \quad \gamma'(0) = \sum [effet \quad de \quad pepite + \gamma(0)] = 35$$

$$C(h) = C(0) - \gamma(h) \quad avec : C(0) = Palier, \quad alors :$$

$$C(h)_{théorique} = palier * \exp\left[-\left(\frac{1.73 * h}{portee}\right)^2\right]$$

$$C'(h) = \text{cov}\, ariance\ gigogne = \sum [C(h)]$$

$$C'(h) = 81 * \exp\left[-\left(\frac{1.73 * h}{26}\right)^2\right] + 59 * \exp\left[-\left(\frac{1.73 * h}{136}\right)^2\right]$$

$$C'(0) = \sum [effet\ de\ pepite + C(0)] = 35 + 81 + 59 = 175$$

Ces expressions signifient que la courbe de la covariance peut être obtenue en retournant le variogramme (sans prendre en compte l'effet de pépite). Dans toute la suite, les termes variogramme et covariance seront employés sans distinction. La Figure 9 montre cette équivalence.

Les modèles ont été définis avec le module de base du logiciel ISATIS,. Ils s'accordent bien avec les données expérimentales des agglomérations dans les vingt premiers kilomètres et avec les données rurales au-delà (Figure 8).

A titre indicatif, un modèle a été construit pour l'ensemble des mesures, tous types de site confondus (seules les trois valeurs extrêmes mesurées sur le littoral ont été enlevées). Les paramètres de ce modèle sont proches de ceux du modèle rural, à l'exception du palier de la seconde structure, qui est deux fois plus grand (Figure 8) . Comme le voisinage d'estimation qui sera utilisé a un rayon de 50 km, les estimations devraient peu changer si l'on choisissait ce modèle. En revanche, la variance d'estimation devrait être plus grande.

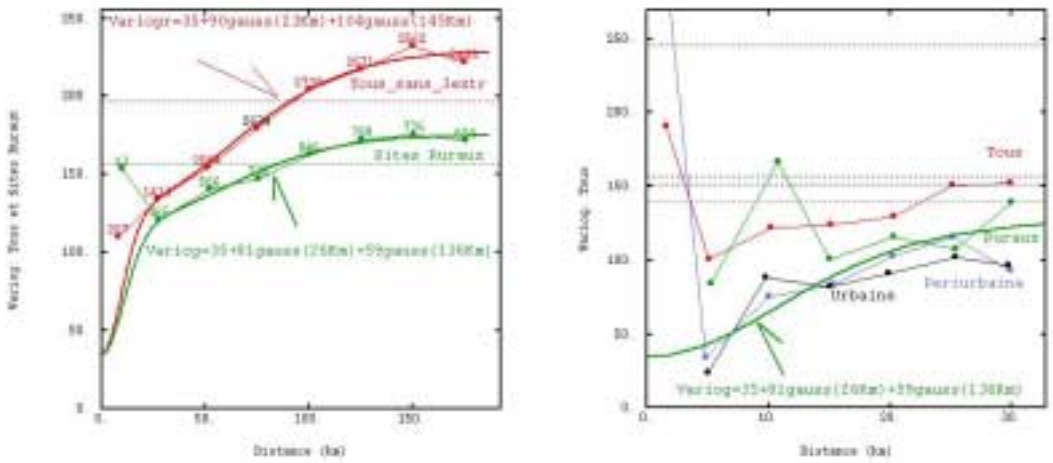


Figure 8 : Modèles ajustés sur les variogrammes expérimentaux pour les sites ruraux et pour l'ensemble de mesures en enlevant les 3 valeurs extrêmes du littoral (à droite : 30 premiers Km).

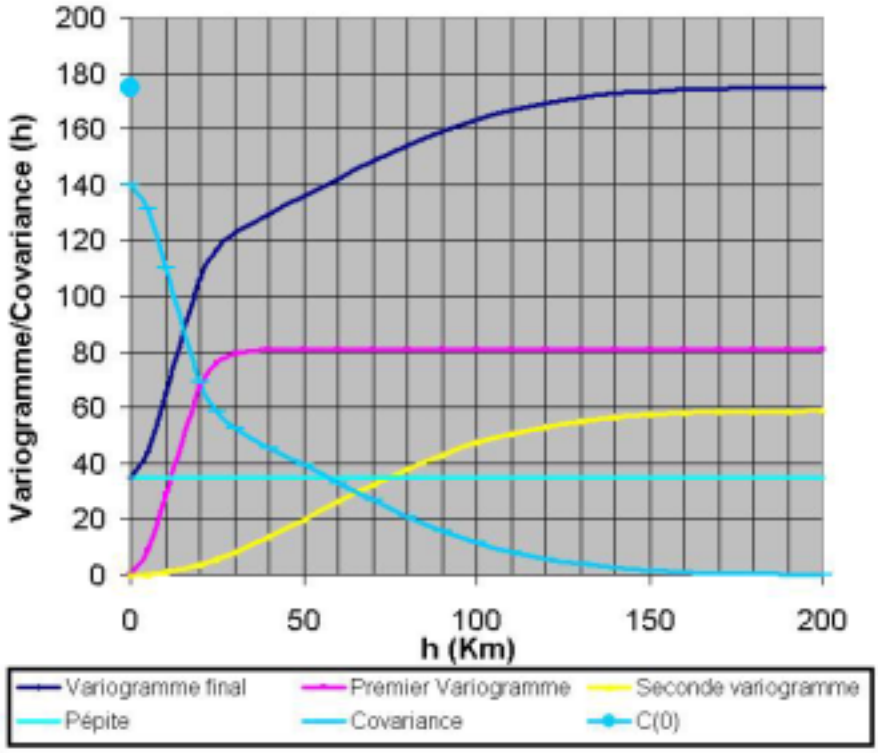


Figure 9 : Composition du variogramme et équivalence avec la covariance

La Figure 9 présente séparément les trois structures qui composent le modèle de variogramme (en turquoise : effet de pépite égal à 35, en rose : première structure de courte portée -36 km-, en jaune : seconde structure de grande portée –136 km-), ainsi que le variogramme final, somme de ces structures (en bleu foncé) et la covariance associée (point et courbe bleu ciel).

5 Validation du Modèle

La méthode employée le plus couramment pour vérifier la qualité d'un modèle est la validation croisée. Elle sert surtout à choisir entre plusieurs modèles, puisqu'en comparant les statistiques qui en résultent, elle permet de distinguer le modèle qui donne en moyenne les estimations les plus précises.

La validation croisée consiste à éliminer temporairement un point de l'ensemble des données puis à l'estimer par krigeage à l'aide du modèle ajusté et des données restantes. Ce processus est répété pour chaque donnée.

En tout point de mesure, on peut donc calculer l'erreur d'estimation (écart entre mesure et estimation). Plus cette erreur est grande, moins le modèle est adéquat.

A titre d'exemple, trois points de mesure situés dans des environnements différents ont été pris pour points cibles. Les résultats de la validation croisée sont présentés dans la Figure 10, le Tableau 1 et le Tableau 2. Les cercles matérialisent le voisinage d'estimation de chacun des points choisis. Ils ont pour rayon 50 km.

Le point 117 est de type rural. La concentration qui y est mesurée est de 59 µg/m^3. Les sites voisins, distants d'environ 25 km, sont du même type. Le point 117 illustre donc la validation du modèle pour les sites ruraux.

Le point 127 est également un site rural, avec une concentration mesurée égale à 83 µg/m^3. A la différence de la configuration d'échantillonnage précédente, le voisinage d'estimation contient deux sites urbains très proches qui influenceront l'estimation.

Le point 165 est de type urbain, avec une concentration mesurée égale à 24 µg/m^3. Son voisinage est constitué principalement de sites urbains très proches et de quelques sites ruraux.

Les points 127 et 165 illustrent donc la validation du modèle pour les régions de transition entre les zones rurales et les agglomérations et pour les agglomérations.

6 Le krigeage

Avant d'analyser les résultats de la validation croisée pour les trois sites choisis et pour l'ensemble des mesures nous rappelons brièvement les formules et les équations de l'algorithme d'estimation appelé krigeage.

On souhaite estimer la variable Z, dans notre cas la concentration d'ozone, par une combinaison linéaire des mesures. La formule 1 fournit l'expression de l'estimateur, qui est une moyenne pondérée des données. Par convention, on utilisera l'astérisque pour caractériser la valeur estimée, afin de la distinguer de la valeur réelle qui reste inconnue. Les facteurs de pondération (λ) doivent minimiser la variance de l'erreur d'estimation indiquée dans la formule 2. Cette variance est appelée variance de krigeage. En 1 et 2 ***«v»*** représente le volume à estimer, lequel peut être un bloc ou un point.

La formule 3 est déduite de la formule 2, en utilisant la définition de variance ($Var(x)=E(x^2)-E(x)^2$) et le fait que l'espérance de l'erreur d'estimation doit être égale à 0 pour obtenir un estimateur sans biais.

$$1 \cdots Z_v^* = \sum_{i=1}^{N} (\lambda_i Z_i)$$

$$2 \cdots \sigma^2 = Var\left[Z_v^* - Z_v \right]_{\min}$$

$$3 \cdots \sigma^2 = Var[Z_v] + Var\left[Z_v^* \right] - 2Cov\left[Z_v, Z_v^* \right]$$

$Z_v^* = Estimateur\ de\ Z_v$

$\lambda = Ponderateurs$

$\sigma^2 = Variance\ d'estimation$

$Var\left[Z_v^* \right] = Variance \quad de \quad Z^* (\text{Z estimée})$

$Cov\left[Z_v, Z_v^* \right] = \text{Co}variance \quad entre\ Z_v^*\ et\ Z_v$

La variance de krigeage dépend de la variance du point ou bloc à estimer, de la variance de la combinaison linéaire des données et de la covariance entre les échantillons et le point ou bloc à estimer (voir la formule 3). Les formules 4, 5 et 6 donnent les expressions de ces trois variances.

$$4 \cdots Var\left[Z_v \right] = C_{vv}$$

$$5 \cdots Var\left[Z_v^* \right] = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\lambda_i \lambda_j C_{ij} \right)$$

$$6 \cdots Cov\left[Z_v, Z_v^* \right] = \sum_{i=1}^{N} \left(\lambda_j C_{iv} \right)$$

$C_{vv} = \text{Co}variance \quad du\ point\ ou\ block\ à\ estimer$

$C_{ij} = \text{Co}variance\ entre\ les\ echantillons$

$C_{iv} = \text{Co}variance\ entre\ les\ echantillons\ et\ le\ point\ ou\ block\ à\ estimer$

$C_h = (C_{h=0}) - \gamma_h : Relation \quad liant\ ces\ \text{cov}ariances\ au\ \text{var}iogramme$

Si le support est un point, l'expression 4 est égale à ***C'(0),*** c'est-à-dire ***175.*** S'il est un bloc (une surface carrée dans notre cas), le calcul de la covariance de ce bloc implique d'en effectuer une discrétisation, comme il est expliqué plus loin.

Toutes ces covariances sont fonction de la corrélation spatiale des données c'est-à-dire du modèle de covariance ou du variogramme qui a été ajusté aux données expérimentales.

Il existe une relation inverse entre les trois types de variances et le variogramme, comme le variogramme est lié directement aux distances entre les données ***(h),*** cela se traduit par une relation inverse entre ces variances et la distance entre échantillons.

En remplaçant les expressions 4, 5 et 6 dans l'expression 3, on obtient bien le fait que la variance de l'erreur d'estimation est une fonction de la covariance (ou du variogramme) ajustée sur les données. C'est la raison pour laquelle il faut avoir recours à des modèles autorisés ou définis positifs.

On démontre aussi que la variance d'estimation dépend en grande partie du plan d'échantillonnage. Plus les points de mesure sont rapprochés, plus la variance d'estimation est faible. Plus la distance entre les mesures et le point ou bloc à estimer est grande, plus la variance d'estimation augmente.

Ce résultat s'observe en pratique. On sait par exemple que des mesures régulièrement espacées offrent une meilleure couverture que des mesures regroupées en « grappes ». On sait également que la meilleure précision est obtenue au voisinage des points de mesure et qu'elle se dégrade loin de ces derniers.

L'expression 7 est le système de krigeage qui fait intervenir le modèle de variogramme. Ce système d'équations nous permet de calculer les pondérateurs ***(λ)*** optimaux qui minimisent la variance de l'erreur d'estimation . Ce système s'écrit sous la forme matricielle suivante ***AX=B*** où

- ***A*** est la matrice des valeurs de covariance (ou de variogramme) calculées entre les points expérimentaux ;
- ***B*** est la matrice des valeurs de covariance calculées entre chaque point de mesure et le bloc ou point à estimer (ce terme est donc différent selon s'il s'agit d'une estimation ponctuelle ou de bloc) ;
- ***X*** est la solution du système ou matrice de coefficients ***λ.***

On appelle ce système, système du ***krigeage ordinaire (KO)*** afin de le différencier des autres types de krigeage moins communs, comme le krigeage à moyenne connue ou « krigeage simple (KS) » (cf. paragraphes suivants).

7. Système de krigeage

$$\begin{cases} \sum_{j=1}^{N} \left(\lambda_j^{KO} \gamma_{ij} \right) - \mu = \bar{\gamma}_{iv} \\ \sum_{i=1}^{N} \lambda_i^{KO} = 1 \end{cases}$$

8.

$$\sigma_{KO}^2 = \sum_{i=1}^{N} \left(\lambda_i^{KO} \bar{\gamma}_{iv} \right) - \bar{\gamma}_{vv} - \mu$$

Notons qu'il y a une condition supplémentaire pour les pondérateurs λ doivent respecter une condition supplémentaire. En effet on démontre que pour que l'estimateur soit sans biais, il faut que la somme de ces pondérateurs soit égale à 1. Cette condition exige d'introduire la valeur ***μ,*** qui est appelée coefficient de Lagrange.

L'expression 8 est une autre formule de calcul de la variance de l'erreur d'estimation. Elle inclut le coefficient de Lagrange (***μ***) qui est issu de la résolution du système de krigeage.

En résumé, la démarche à adopter pour mettre en œuvre le krigeage consiste à

- déterminer le modèle (variogramme ou covariance) à partir des données expérimentales,
- sélectionner le voisinage d'estimation, c'est-à-dire les données qui contribueront à estimer le point ou bloc,
- résoudre enfin le système de krigeage $X=B(A)^{-1}$.
- Une fois que les pondérateurs et le coefficient de Lagrange ont été calculés, il suffit d'appliquer la formule 1 pour calculer la valeur estimée et la formule 3 ou 8 pour connaître la variance de l'erreur d'estimation.

Dans le krigeage ordinaire le nombre d'équations du système est égal au nombre des données du voisinage de krigeage plus 1. Théoriquement, plus on dispose des données dans le voisinage, plus la qualité du krigeage s'améliore. Toutefois, dans la pratique, on s'aperçoit qu'une ou deux couronnes de points de mesure autour du point cible suffisent et qu'elles masquent presque totalement l'influence des données les plus lointaines (***effet d'écran***).

Cet effet d'écran nous permettrent de réduire considérablement le nombre des équations du système linéaire.

Dans notre cas, au lieu d'employer un voisinage unique qui engloberait la totalité des données (209 équations à résoudre !), nous avons choisi un voisinage circulaire glissant de 50 km de rayon. Dans les zones rurales, un tel voisinage correspond à deux auréoles de points de mesure (maille d'échantillonnage de 25 Km).Par exemple le voisinage d'estimation du point 117 contient 20 données; le nombre d'équations à résoudre est donc de 21.

La Figure 10, le Tableau 1 et le Tableau 2 présentent les résultats de la validation croisée par krigeage ordinaire pour les points 117, 127 et 165. Ces résultats montrent l'effet d'écran et l'influence des mesures les plus proches des points à estimer.

Le voisinage du point 117 se caractérise par une bonne symétrie : on distingue deux auréoles de tubes, la première à une distance d'environ 25 km et la seconde à environ 50 Km. La valeur numérique des pondérateurs ainsi calculés varie entre 15.5% (pour un des points les plus proches) à 1.84% (pour un des points se trouvant à la limite du voisinage). La valeur estimée est de 63 µg/m^3 et l'erreur d'estimation n'est que de 3,8 µg/m^3.

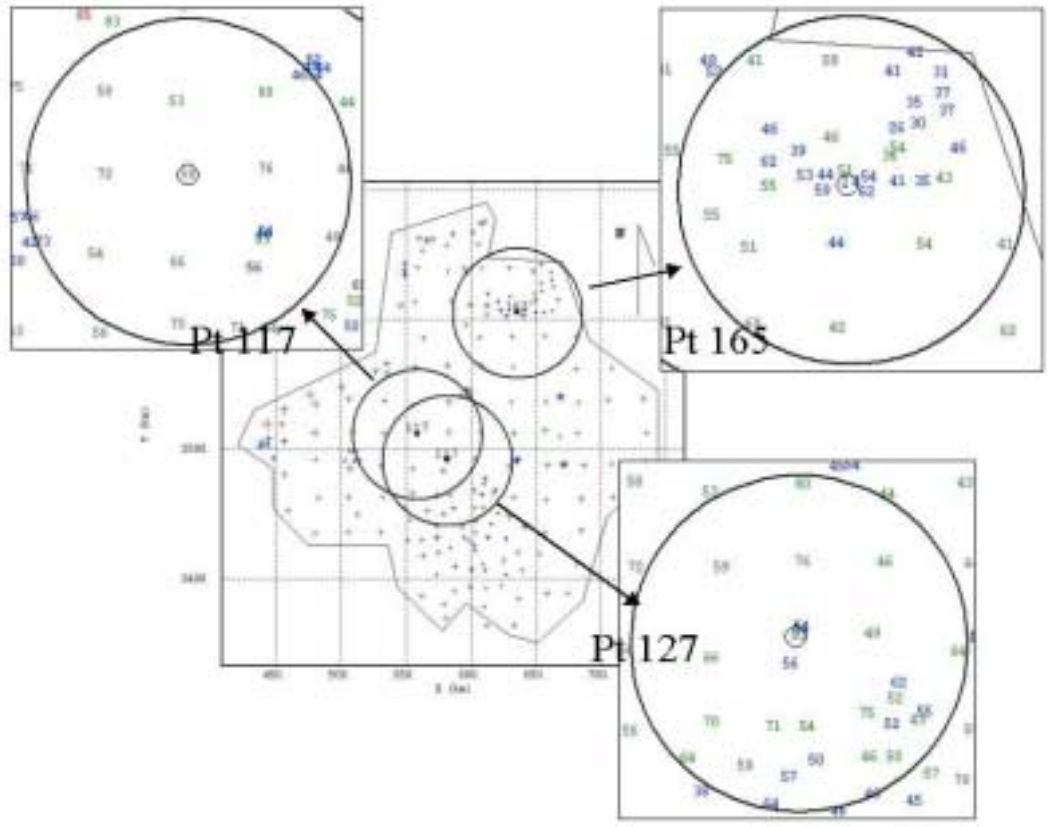
En revanche pour les voisinages des points 127 et 165 la configuration des échantillons est plus asymétrique à cause de la densité des données des agglomérations.

Le voisinage d'estimation pour le point 127 compte 25 données. Trois d'entre elles, cependant, sont très proches du point à estimer. La somme des pondérateurs qui leur sont attribués par le système de krigeage est de 92.5%. Cet effet d'écran et ce voisinage asymétrique sont aussi responsables de pondérateurs négatifs affectés à trois points de mesure éloignés. La valeur estimée au point 127 est de 51.5 µg/m^3 (valeur proche de la moyenne des trois données : 46 µg/m^3, 54 µg/m^3, 56 µg/m^3, moyenne=52 µg/m^3) et l'erreur d'estimation est de 31.7 µg/m^3 (environ dix fois supérieure à celle du point 117).

La situation n'est pas plus favorable pour le point 165. Le voisinage composé de 30 données montre quelques regroupements dans la partie supérieure et des points de mesure très proches du point à estimer.

Le système de krigeage a privilégié les 5 données qui entourent le point 165 (62 µg/m^3, 54 µg/m^3, 51 µg/m^3, 44 µg/m^3 et 59 µg/m^3),puisqu'à eux seuls, les pondérateurs totalisent un pourcentage de 100% Les pondérateurs associés aux données restantes sont proches de 0 ou négatifs. La valeur estimée est 55 µg/m^3 et l'erreur d'estimation est de 30.5 µg/m^3 .

Ces résultats montrent que pour juger la pertinence d'un modèle par validation croisée, il vaut mieux que les données se répartissent régulièrement dans l'espace. En effet, par suite des effets d'écran et d'asymétrie, les regroupements de données affectent les résultats de l'estimation.



Les valeurs numériques correspondent à la concentration de l'ozone.
La couleur verte représente les sites ruraux et la couleur bleue les sites urbains et périurbains.

Figure 10 : Validation croisée pour les points :117, 127 et 165

Comme il a été signalé, il existe d'autres types de krigeage, qui se déduisent du krigeage ordinaire. Nous allons notamment analyser ***le krigeage simple ou krigeage à moyenne connue***. Même si dans notre cas la moyenne est inconnue, la comparaison des résultats du krigeage simple avec ceux du krigeage ordinaire apporte des informations supplémentaires pour tester la qualité de l'estimation.

L'expression 9 qui suit, est l'estimateur du krigeage simple. Le signe «'» est utilisé pour distinguer les poids du krigeage simple de ceux du krigeage ordinaire. On constate qu'un facteur s'ajoute dans le krigeage simple. Celui-ci est la moyenne connue, multipliée par son pondérateur appelé ***« le poids de la moyenne ».***

L'estimateur du krigeage simple est automatiquement sans biais, aussi n'est-il pas nécessaire que la somme des poids vaille 1. En conséquence, le système de krigeage (expression 10) ne fait pas intervenir le coefficient de Lagrange et le nombre d'équations est égal aux nombre des données du voisinage d'estimation. La variance d'estimation du krigeage simple est donné par l'expression 11.

L'expression 9 montre que lorsqu'en krigeage simple, le poids de la moyenne est faible (voisin de 0), alors la teneur estimée Z^* dépend principalement des valeurs locales de Z, c'est-à-dire des points qui appartiennent à la configuration de krigeage, et non pas de la teneur moyenne dans le domaine. Le degré de stationnarité requis est donc moins important, de plus les estimateurs du ***KO*** (expression 1) et du ***KS*** (expression 9) sont proches, ainsi que les variances d'estimation (expression 8 et 11).

$$9 \cdots Z_V^* = \sum_{i=1}^{N} \lambda_i' Z_i + m\left(1 - \sum \lambda_i'\right)$$

$$10 \cdots \sum \lambda_j' \gamma_{ij} = \bar{\gamma}_{iv}$$

$$11 \cdots \sigma_{KS}^2 = \sum \lambda_i' \bar{\gamma}_{iv} - \bar{\gamma}_{vv}$$

$Z_V^* = Estimateur \cdots de \cdots Z_V$

$\lambda' = Ponderateurs \cdots du \cdots krigeage \cdots simple$

$m = moyenne \cdots connue$

$\sigma_{KS}^2 = Variance \cdots d'estimation \cdots du \cdots krigeage \cdots simple$

$\left(1 - \sum \lambda_i'\right) = Le \cdots poids \cdots de \cdots la \cdots moyenne$

Moins il y a d'information disponible dans le voisinage de krigeage, plus l'importance de la moyenne est grande. De ce fait le poids de la moyenne donne une idée du nombre de données disponibles dans le voisinage d'estimation et indique dans quelle mesure l'hypothèse de stationnarité est valable. Afin d'illustrer ce propos, nous avons imposé une valeur quelconque à la moyenne (0 dans notre exemple) puis nous avons utilisé le système de krigeage simple de l'expression 10 pour connaître les pondérateurs λ'.

Le Tableau 1 met en parallèle, pour les points 117, 127 et 165, les résultats des krigeage ordinaire et du krigeage simple. Les poids assignés à la moyenne (ligne 3 du tableau) sont faibles, indiquant par là que la quantité des données disponibles dans ces trois voisinages est appropriée (on constate que plus il y a des échantillons dans le voisinage plus ce poids est petit).

No.	*Item*	*Expression*	*Point 117 = 59 µg/m³*		*Point 127 = 83 µg/m³*		*Point 165 = 24 µg/m³*	
			KO	***KS***	***KO***	***KS***	***KO***	***KS***
1	*Nb des données dans le voisinage*		20		25		30	
2	*Moyenne connue*		N/A	0	N/A	0	N/A	0
3	*Poids assigné à la moyenne (%)*	*10*		3.1		0.5		0,1
4	*Variance de Z**	*5*	53,4	50,6	127,1	126,6	129,6	129,5
5	*Covariance entre Z et Z**	*6*	52,0		126,8		129,5	
6	*Paramètre de Lagrange*	*7*	-1,4	N/A	-0,3	N/A	-0.1	N/A
7	*Variance d'estimation*	*3 ou KO :8 et KS :11*	124,4	124,4	48,4	48,4	45,5	45,5
8	*Ecart-type d'estimation*		11,2	11,2	7,0	7,0	6,7	6,7
9	*Valeur Estimée*	*KO :1 et KS :9*	63,2	61,3	51,5	51,2	55,0	54,9

Tableau 1 : Résultats de la validation croisée des points 117, 127 et 165

Les résultats du krigeage simple approchent logiquement ceux du krigeage ordinaire (valeurs estimées très proches, mêmes variances d'estimation).

Comme le support d'estimation dans une validation croisée est ponctuel et que la variance d'un point est égale à la somme des paliers du modèle, on obtient ***Var[Z_v]=Var Z=175.***

Les valeurs de la variance d'estimation présentées dans ce tableau ont été vérifiées pour le point 117, en krigeage ordinaire, et pour le point 165, en krigeage simple (lignes 4, 5 et 6 du Tableau 1) :

Calcul de la variance d'estimation pour le point 117 en KO :

$$\sigma^2_{KO} = Var[Z_V] + Var\left[Z_V^*\right] - 2Cov\left[Z_V, Z_V^*\right]$$

$$\sigma^2_{KO} = 175 + 53.5 - 2(52) = 175 + 53.4 - 104 = 228.4 - 104$$

$$\underline{\sigma^2_{KO} = 124.4} \qquad \underline{\sigma_{KO} = \sqrt{124.4} = 11.2}$$

$$\sigma^2_{KO} = \sum_{i=1}^{N} (\lambda_i^{KO}\gamma_{iv}) - \gamma_{vv} - \mu = -Cov\left[Z_V, Z_V^*\right] + Var[Z_V] - \mu$$

$$\sigma^2_{KO} = -52 + 175 - (-1.4) \qquad \underline{\sigma^2_{KO} = 124.4}$$

Calcul de la variance d'estimation pour le point 165 en KS:

$$\sigma^2_{KS} = Var[Z_V] + Var\left[Z_V^*\right] - 2Cov\left[Z_V, Z_V^*\right]$$

$$\sigma^2_{KS} = 175 + 129.5 - 2(129.5) = 175 + 129.5 - 259 = 304.5 - 259$$

$$\underline{\sigma^2_{KS} = 45.5} \qquad \underline{\sigma_{KS} = \sqrt{45.5} = 6.74}$$

$$\sigma^2_{KS} = \sum_{i=1}^{N} (\lambda'_i\gamma_{iv}) - \gamma_{vv} = -Cov\left[Z_V, Z_V^*\right] + Var[Z_V]$$

$$\sigma^2_{KS} = -129.5 + 175 \qquad \underline{\sigma^2_{KS} = 45.5}$$

Vérification du système de krigeage ordinaire pour le point 117

$$\sum_{j=1}^{N} (\lambda_j^{KO}\gamma_{ij}) - \mu = \gamma_{iv}$$

$$-Var\left[Z_V^*\right] - \mu = -Cov\left[Z_V, Z_V^*\right]$$

$$-53.4 - (-1.4) = -52$$

$$-53.4 + 1.4 = -52: \quad -52 = -52$$

Vérification du système de krigeage simple pour le point 165:

$$\sum_{j=1}^{N} (\lambda'_j\gamma_{ij}) = \gamma_{iv}$$

$$Var\left[Z_V^*\right] = Cov\left[Z_V, Z_V^*\right]$$

$$129.5 = 129.5$$

7 Statistiques de la Validation Croisée

La validation croisée permet de comparer aux points de mesure la valeur réelle avec la valeur estimée grâce aux données voisines et au modèle ajusté. Les expressions 12, 13, 14 et 15 fournissent quatre statistiques qui peuvent être utilisées comme critères de sélection d'un modèle.

12 $Erreur = Z_V - Z_V^*$ 13 $Variance\ de\ l'Erreur = \left(Z_V - Z_V^*\right)^2$

14 $Erreur\ reduite = \frac{Z_V - Z_V^*}{\sigma *}$ 15 $Variance\ de\ l'Erreur\ reduite = \frac{\left(Z_V - Z_V^*\right)^2}{\sigma_*^2}$

L'expression 12 est l'erreur d'estimation. Plus elle est faible (proche de 0) plus le modèle est adéquat. L'expression 13 donne la variance de l'erreur. Plus cette variance est faible, plus stable est l'estimateur. Si on divise ces statistiques par la variance d'estimation on trouve l'erreur réduite et la variance réduite (expression 14 et 15).

Ces statistiques ont été calculées pour les trois points 117, 127 et 165 (Tableau 2).; on observe que le modèle est plus précis dans les zones rurales que dans les zones de transition ou dans les agglomérations. Ces résultats, cependant, sont affectés par l'effet d'écran et la géométrie du voisinage de krigeage.

Le Tableau 2 donne ces statistiques pour les trois points 117, 127 et 165 ; on observe que le modèle est plus précis quand on estime dans les zones rurales que quand on estime dans les zones de transition ou les zones des agglomérations ; cependant on a vu que ces résultats sont affectés par l'effet d'écran et la géométrie symétrique des échantillons dans le voisinage de krigeage.

Item	***Point 117***	***Point 127***	***Point 165***
Erreur	3.78	31.68	30.53
Variance de l'erreur	14.26	10003.47	932.34
Erreur réduite	0.34	4.55	4.53
Variance de l'erreur réduite	0.11	20.73	20.48

Tableau 2 : Statistiques individuelles pour la validation croisée des trois points

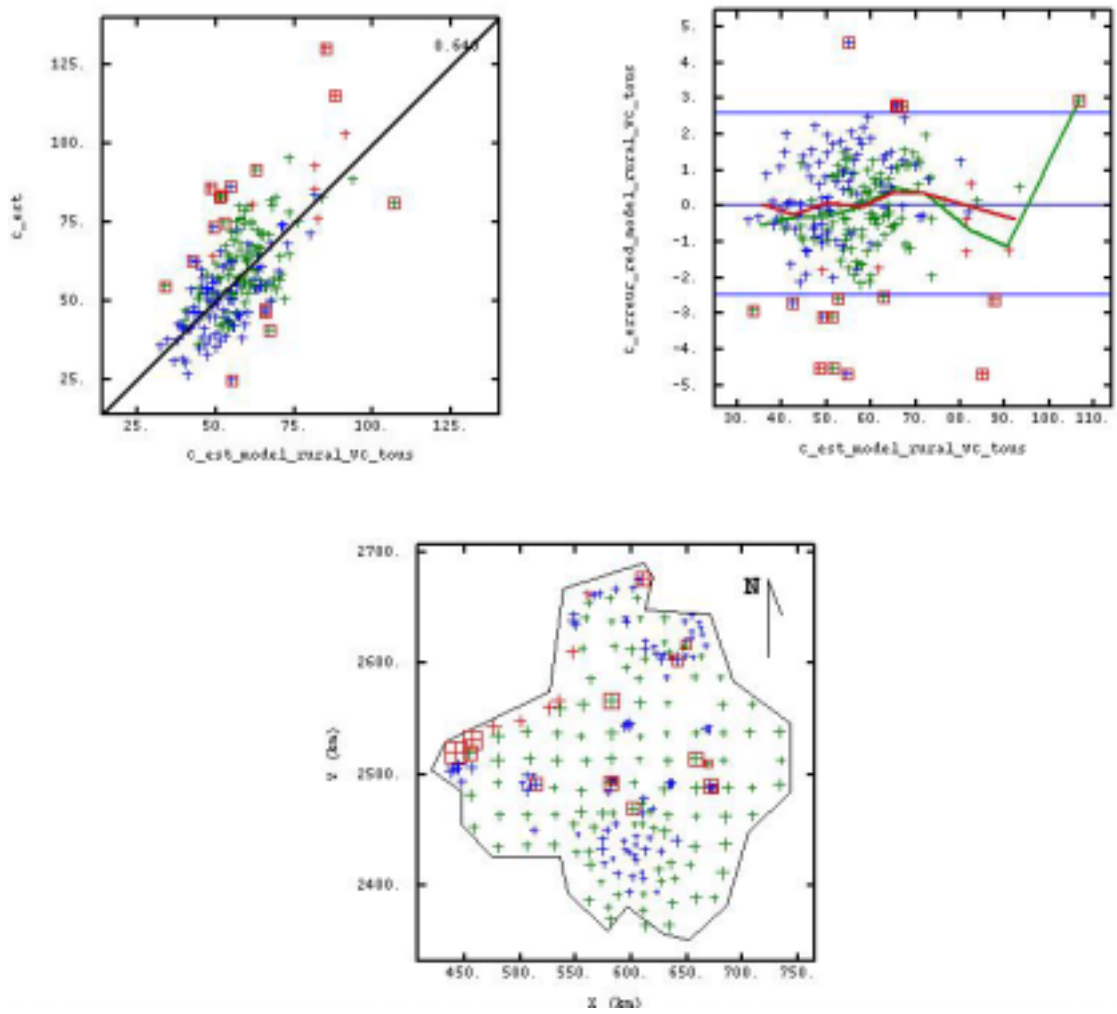
La validation croisée, effectuée sur la totalité des mesures, conduit aux statistiques d'erreur suivantes :

16 $Moyenne\ de\ l'Erreur = \frac{1}{N}\sum_{i=1}^{N}\left(Z_i - Z_i^*\right)$ 17 $Moyenne\ de\ la\ Variance\ de\ l'Erreur = \frac{1}{N}\sum_{i=1}^{N}\left(Z_i - Z_i^*\right)^2$

18 $Moyenne\ de\ l'Erreur\ reduite = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Z_i - Z_i^*}{\sigma_i^*}\right)$ 19 $Moyenne\ de\ la\ Variance\ de\ l'Erreur\ reduite = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Z_i - Z_i^*}{\sigma_i^*}\right)^2$

Les expressions 15 et 19 font intervenir le rapport entre la variance expérimentale $(Z\text{-}Z^*)^2$ et la variance théorique (variance de krigeage qui est fonction des paliers du modèle). Plus ces rapports sont proches de 1, plus le modèle est approprié.

La Figure 11 montre de façon graphique les résultats de quelques-unes de ces statistiques pour tous les 209 mesures.



Croix vertes : Sites ruraux.
Croix Rouges : Sites du littoral.
Croix bleues : Sites périurbains et urbains.
Carres rouges : Les sites dont l'erreur réduite est plus grand que +/- 2.5.

Figure 11 : Validation croisée du modèle ajusté sur les données rurales. La validation a utilisé l'ensemble des données.

La plupart des mesures réelles sont comprises entre 30 µg/m^3 et 80 µg/m^3 tandis que la plupart des valeurs estimées varient entre 40 µg/m^{3}et 70 µg/m^3. Ce resserrement de la distribution des concentrations est ce qu'on appelle ***l'effet de lissage du krigeage***. Il est dû à ce que les valeurs estimées sont une moyenne pondérée des données du voisinage d'estimation.

Si globalement, les points sont proches de la bissectrice, l'estimateur est sans biais conditionnel (pente=1). Quelques valeurs fortes sont ici sous-estimées (carrés rouges au-dessus de la bissectrice).

Les carrés rouges identifient les échantillons dont l'erreur réduite est supérieure à +/-2.5 (ceci correspondrait au 1.24% des valeurs si on suppose que la distribution des erreurs autour de la moyenne 0 obéit à la loi de Gauss). On observe dans la carte d'implantation qu'il s'agit de quelques sites urbains, de quelques sites littoraux et de quelques sites ruraux proches des agglomérations

Dans le nuage entre erreur réduite d'estimation et valeurs estimées les lignes bleues et vertes donnent la moyenne des erreurs réduites d'estimation par classe de valeur estimée. Elles oscillent autour de la valeur 0. La ligne rouge ne prend pas en compte les carres rouges c'est-à-dire les échantillons avec une erreur réduite d'estimation plus grand que +/- 2.5.

Les statistiques qui excluent ces mesures sont appelées « ***statistiques robustes*** », le tableau de la Figure 12 montre la moyenne des statistiques de la validation croisée pour toutes les données ainsi que pour les données

robustes. Dans notre cas, seul un modèle a été ajusté mais dans le cas où plusieurs modèles seraient en compétition, on choisirait de préférence celui qui présente les meilleures statistiques (sauf si l'expérience et la connaissance du terrain suggèrent plutôt d'opter pour un autre modèle).

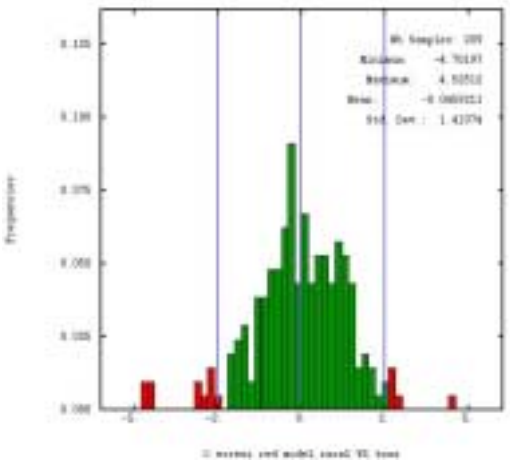
Si on multipliait chacun des paliers du modèle par la variance moyenne de l'erreur réduite d'estimation (1.99 dans notre cas), alors en moyenne, la variance de krigeage obtenue par validation croisée serait égale à la moyenne des variances expérimentales (c'est-à-dire à la variance de l'erreur 146.42).

Cette possibilité est offerte par le logiciel ISATIS. Ce résultat signifie que lorsqu'on applique un même facteur multiplicatif à chacun des paliers d'un modèle les valeurs estimées ne varient pas mais la variance d'estimation est multipliée par ce facteur.

Le modèle initial ***[γ(h)=35+81gauss(26Km)+59gauss(136Km)]*** a donné une variance d'erreur égale à 146.42 (moyenne des erreurs quadratiques expérimentales) et une variance moyenne de krigeage égale à 73.6.

Quand on le multiplie par 1.99, le modèle devient ***[γ(h)=70+162gauss(26Km)+118gauss(136Km)]***. La variance de l'erreur reste inchangée (146.42) mais la variance moyenne de krigeage passe à 146.42 (deux fois 73.6). Comme ces deux variances sont maintenant égales on obtient une moyenne des variances de l'erreur réduite d'estimation égale à 1.

Cette option n'a pas été retenue pour réajuster les paliers des composantes du modèle. En effet, il nous a semblé préférable de conserver le modèle initial, qui décrit correctement la structure spatiale du phénomène, et de ne réadapter localement que la valeur de la variabilité à grâce à la variance de l'erreur de mesure (cf. paragraphes qui suivent).



Item	*Nb*	*Moyenne*	*Variance*
Erreurs	209	-0.66367	146.42
Erreurs Réduites	209	-0.04592	1.99

Erreurs: Sites robustes	193	0.30768	99.22
Erreurs Réduites: Sites robustes	193	0.06674	1.15

Figure 12 : Histogramme de l'erreur réduite et statistiques des résultats de la validation croisée

Quand on divise les erreurs par la valeur vraie des mesures, on obtient les erreurs relatives La Figure 13 montre les statistiques de ces erreurs, lesquelles, pour la plupart, ne dépassent pas 50% de la valeur vraie. Ce type de calcul sera employé dans les chapitres ultérieurs afin de déterminer le degré d'incertitude des cartes obtenues.

La validation croisée a été mise en œuvre sur l'ensemble des données. Une autre possibilité consiste à sélectionner aléatoirement 10 jeux de données contenant chacun 10% des échantillons (Figure 14), puis à effectuer séparément la validation croisée pour chaque jeu (Jeannée et al., 2003). Les statistiques calculées sont les mêmes mais elles portent seulement sur 10% des données.

Le fait de restreindre le nombre de données de validation et de les sélectionner aléatoirement permet d'atténuer l'influence des zones sur-échantillonnées (telles que les agglomérations). Les statistiques issues d'une telle procédure sont présentées dans le Tableau 3. Chaque jeu contient 19 données sélectionnées de façon aléatoire.

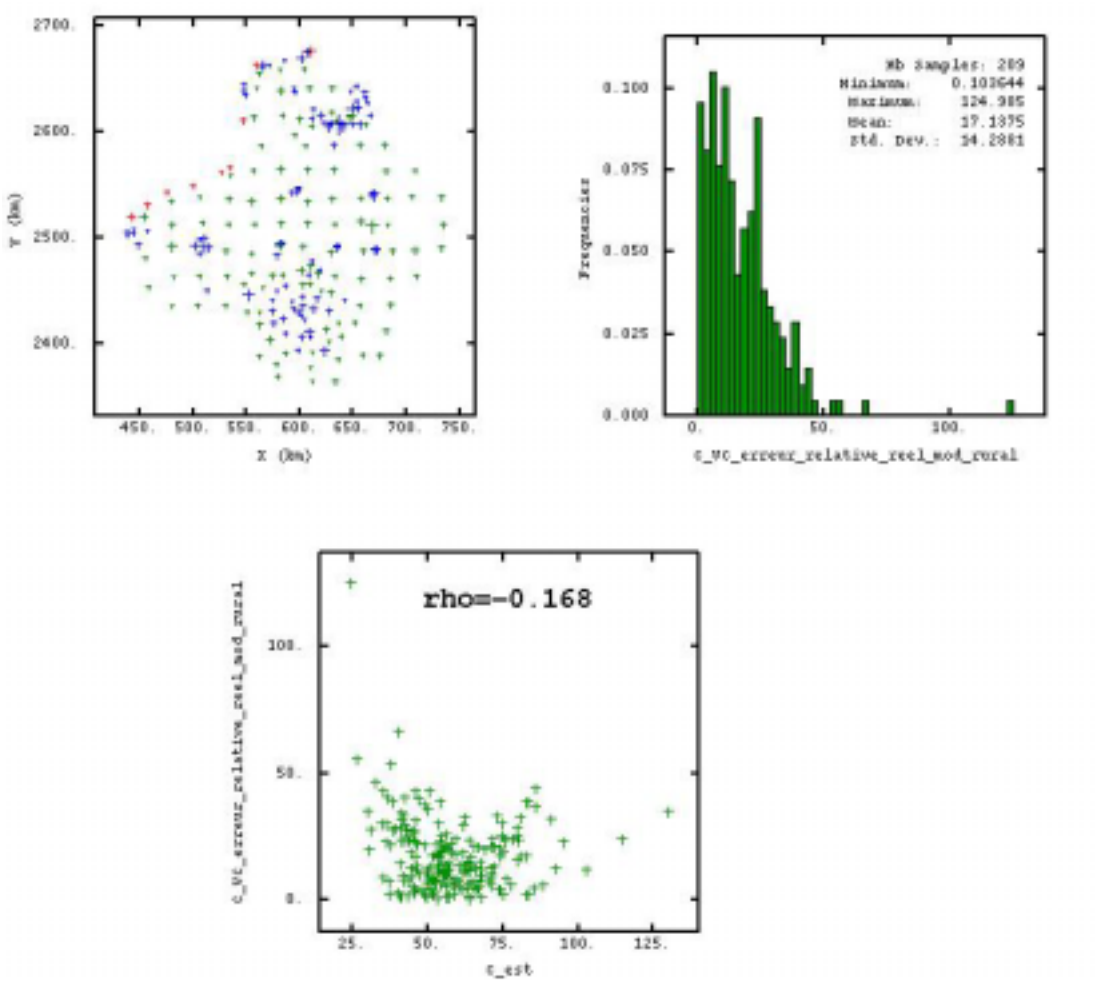


Figure 13 : Carte et Statistiques des erreurs relatives

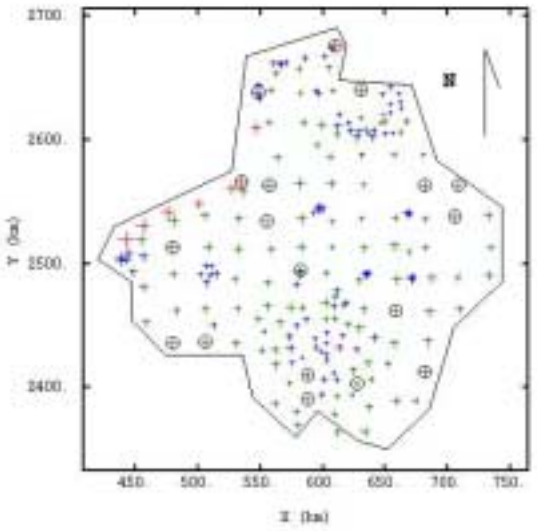
Moyenne des erreurs	*Moyenne des erreurs quadratiques*
-4,61	54,88
1,98	73,19
1,32	78,77
-1,42	83,17
1,67	91,60
-0,38	117,56
-4,40	121,16
-3,81	121,73
0,77	138,84
-5,10	140,89

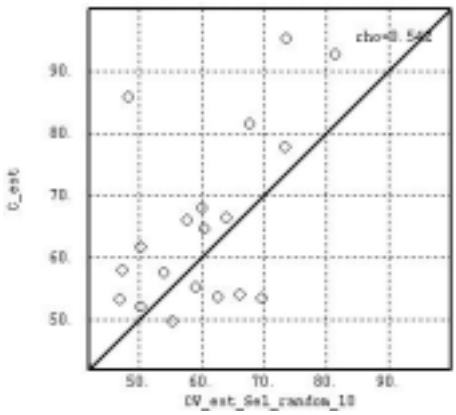
	Moyenne des erreurs	*Moyenne des erreurs quadratiques*
Moyenne	-1,40	102,18
Minimum	-5,10	54,88
Maximum	1,98	140,89
Ecart-type	2,84	29,67

Tableau 3 : Statistiques de la validation des 10 jeux de données différents

Les résultats des statistiques varient beaucoup d'un jeu de données à un autre. Ainsi la plus petite moyenne des erreurs quadratiques est de 55, elle est presque trois fois inférieure à la moyenne calculée pour l'ensemble des données (146).

On voit donc que les résultats numériques de la validation croisée sont très sensibles au voisinage d'estimation. Aussi doivent-ils être plutôt interprétés qualitativement , par exemple pour comparer deux modèles. En revanche, ils ne permettent pas de mesurer les incertitudes de façon précise.





Item	*Nb*	*Moyenne*	*Variance*
Erreurs	19	-5.10	140.89
Erreurs Réduites		-0.52	1.71

Figure 14 : Exemple d'un jeu de données choisi aléatoirement (Cercles noirs) et résultats de la validation croisée

8 Résultats de l'estimation par krigeage sans VEM

L'estimation a été réalisée par krigeage ordinaire à moyenne inconnue. Le modèle utilisé est celui qui a été précédemment validé et le voisinage d'estimation est un cercle de 50 km de rayon.

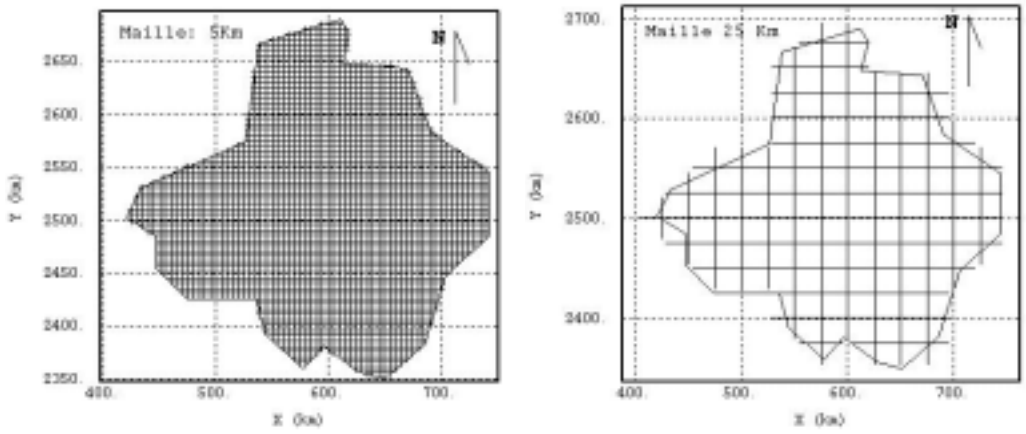


Figure 15 : Deux mailles d'estimation

Afin de mettre en évidence l'influence de la maille de calcul sur les incertitudes d'estimation, deux tailles de maille, 5 et 25 km, sont considérées.

La première grille de calcul (pas de 5km) comprend 2510 nœuds , la seconde (pas de 25 km) en compte 98. (Figure 15).

Pour chaque grille, deux types d'estimation ont été réalisés : dans un premier calcul, les concentrations ont été estimées ponctuellement (support de prélèvement des échantillons), au centre de chaque maille ; dans un second calcul, des concentrations moyennes par maille ont été estimées.

Avec un support d'estimation ponctuel, les variances et covariances se déduisent directement du variogramme. Avec un support de dimensions non nulles, la variance du bloc et la covariance entre ce bloc et les points de mesure sont calculées par discrétisation.

Cette méthode consiste à superposer une grille d'une dimension donnée à un bloc de la maille, puis à calculer la valeur de la covariance pour plusieurs distances entre les nœuds de cette grille. La moyenne de toutes ces covariances est alors calculée et on obtient la variance du bloc. Cette méthode est aussi appliquée pour le calcul de la variance entre les échantillons (points) et le bloc, qui est le terme de droite du système de krigeage de bloc (C_{iv} ou γ_{iv}).

Pour une taille de bloc de 5 km, la variance de bloc (expression 4) est de 137.1. Pour une taille de bloc de 25 km, elle est de 98 (plus petite). Quant à la variance ponctuelle C(0), elle est égale à la somme des paliers du variogramme (175), valeur qui sera toujours supérieure aux variances moyennes de bloc.

Les statistiques associées aux quatre estimations réalisées sont indiquées dans le Tableau 4 et le Tableau 5.

VARIABLE	Maille (Km)	Point/Bloc	Nb	Minimum	Maximum	Moyenne	Variance	CV
Estimation	5	P	2510	33.56	117.36	61.38	132.80	0.19
		B		33.74	116.89	61.37	131.64	0.19
	25	P	98	34.09	109.36	61.15	130.81	0.19
		B		38.21	104.26	61.30	116.12	0.18

Tableau 4 : Statistiques des valeurs estimées d'ozone en µg/m^3 (semaine du 26 juin au 3 juillet 2000)

Le Tableau 5 affiche pour la pollution de l'ozone les statistiques de l'écart-type de l'erreur d'estimation par les quatre estimateurs indiqués.

VARIABLE	Maille (Km)	Point/Bloc	Nb	Minimum	Maximum	Moyenne	Variance	CV
Ecart-type d'estimation	5	P	2510	6.52	12.24	8.88	1.20	0.12
		B		2.67	10.58	6.39	2.21	0.23
	25	P	98	6.8	11.22	9.29	1.23	0.12
		B		2.06	7.13	4.50	1.33	0.26

Tableau 5 : Statistiques des valeurs estimées d'ozone en µg/m^3 (semaine du 26 juin au 3 juillet 2000)

Les deux estimations ponctuelles fournissent des statistiques similaires, quelle que soit la taille de la maille, car elles font appel au même type de support. Les estimations par bloc, en revanche, révèlent entre elles des différences, du fait que les supports sont différents.

Elles sont plus lissées lorsque les blocs sont plus larges : avec une maille de 25 km, les valeurs maximales sont plus faibles, et les valeurs minimales plus fortes, qu'avec une maille de 5 km. De plus la variance et le coefficient de variation sont moins élevés. On appelle cela l'***«effet de support ».***

On constate que l'écart-type de l'erreur est affecté par la valeur C(0) qui dépend du type de krigeage. Comme il a été indiqué, la variance ponctuelle est toujours supérieure à la variance de bloc. On obtient de la même façon des variances d'estimation ponctuelle supérieures à celles de l'estimation de bloc.

A cause des valeurs de la covariance moyenne de bloc on voit aussi que les estimations de la moyenne de bloc de 5 Km sont plus incertaines que les estimations de la moyenne de bloc de 25 Km.

Comme pour la validation croisée, les résultats de l'estimation sont présentés pour quelques points choisis : il s'agit des nœuds les plus proches du point 117.

No.	*Item*	*Expression*	*Bloc ou point les plus près du TP No. 117(Voir Figure 10, Figure 21)*			
			Maille : 5 Km		*Maille : 25 Km*	
			Krigeage Ordinaire (KO)			
			Bloc	*Ponctuel*	*Bloc*	*Ponctuel*
1	*Nb des données dans le voisinage*		*19*		*24*	
4	*Covariance moyenne de bloc ou de point C(0)*		137.09	175	98.03	175
5	*Variance de Z* (Estimée Z)*	*5*	112.56	114.5	71.75	72.52
6	*Covariance entre Z et Z**	*6*	111.98	113.94	71.34	72.18
7	*Paramètre de Lagrange*	*7*	-0.57	-0.56	-0.41	-0.33
8	*Variance d'estimation*	*3 ou KO :8 et KS :11*	25.68	61.63	27.01	103.15
9	*écart-type d'estimation*		5.07	7.85	5.2	10.16
10	*Valeur Estimée*	*KO :1 et KS :9*	60.8	60.73	62.7	62.66
11	*Pente de la régression Z|Z**	*20*	0.99	0.99	0.99	0.99
12	*Corrélation entre Z et Z**	*21*	0.9	0.8	0.85	0.64

Tableau 6 : Comparaison de Statistiques des estimations pour une valeur estimée choisie

Dans la Figure 21 on voit que la localisation géographique et donc le voisinage des deux nœuds sont différents (19 échantillons pour le nœud de maille 5 Km et 24 échantillons pour le nœud de maille 25 Km). En conséquence les résultats des estimations sont légèrement différents selon de maille (autour de 61 µg/m^3 pour le nœud de la maille de 5Km et autour de 63 pour le nœud de la maille de 25 Km).

Les résultats de l'estimation (Tableau 6) corroborent les observations faites précédemment, à savoir que les valeurs estimées varient peu selon les supports mais que les variances des erreurs d'estimation ponctuelles sont supérieures à celles des estimations de bloc.

$$20 \quad p = \frac{Cov\left[Z_v, Z_v^*\right]}{Var\left[Z_v^*\right]} = \frac{Cov\left[Z_v, Z_v^*\right]}{Cov\left[Z_v, Z_v^*\right] - \mu}$$

$$21 \quad \rho = \frac{Cov\left[Z_v, Z_v^*\right]}{\sqrt{Var\left[Z_v^*\right] * Var\left[Z_v\right]}}$$

Les lignes 11 et 12 du Tableau 6 ont été calculées à l'aide des expressions 20 et 21. La pente de la régression linéaire (expression 20) indique si l'estimateur présente ou non un biais conditionnel. En effet si la régression est linéaire et de pente 1, alors l'espérance de valeurs réelles conditionnées par les valeurs estimées est égale aux valeurs estimées (***E[*Z_v*|*Z_v**]=*Z_v**,*** voir par exemple le nuage de corrélation de la Figure 11).

En pratique, les distributions de Z_v et Z_v^* sont rarement connues, la véritable forme de l'espérance conditionnelle Z_v, considérée comme une fonction de Z_v^*, est donc inconnue. On se contente de calculer la pente de façon théorique (expression 20) et d'observer si cette pente est proche de 1.

Le coefficient de corrélation (expression 21) est fonction des variances. Aussi est-il influencé par le support considéré (point ou bloc) et par les caractéristiques du voisinage d'estimation, notamment par les distances entre points de mesure d'une part, entre points de mesure et bloc ou point cible d'autre part. Ainsi les coefficients de corrélation diminuent quand la taille du support et la dimension de la maille sont réduites.

Si le coefficient de corrélation (expression 21) est égal à 0 alors les variables sont non corrélées, s'il est égal à 1 elles sont corrélées linéairement et s'il est égal à –1, elles sont corrélées linéairement mais au sens inverse. On cherchera donc à ce que cette valeur soit la plus proche possible de 1.

9 Résultats de l'estimation par krigeage avec VEM

La variance de l'erreur de mesure apporte une information sur l'incertitude des données expérimentales et sur la variabilité à l'origine. Nous présentons la façon d'en tirer parti lors du krigeage.

Le krigeage avec erreur de mesure consiste à estimer la concentration non bruitée à l'aide des données de mesure, qui, elles, sont entachées d'erreur. L'expression 22 indique le nouveau système de krigeage.

$$22.\cdots Z_v^* = \sum_{i=1}^{N}\left[\lambda_i^{VEM}\left(Z_i+\varepsilon_i\right)\right]$$

$$23\cdots \sigma^2 = Var\left[Z_v^* - Z_v\right]_{\min} = Var\left[\sum_{i=1}^{N}\left[\lambda_i^{VEM}\left(Z_i+\varepsilon_i\right)\right] - Z_v\right]_{\min}$$

$$24\cdots \sigma^2 = Var\left[Z_v\right] + Var\left[Z_v^*\right] - 2Cov\left[Z_v, Z_v^*\right] + Var\left[\varepsilon_i\right]$$

25. Système de krigeage

$$\begin{cases} \sum_{j=1}^{N}\left[\lambda_j^{VEM}\left(\gamma_{ij} - VEM_i\right)\right] - \mu = \bar{\gamma}_{iv} \\ \sum_{i=1}^{N}\lambda_i^{VEM} = 1 \end{cases}$$

26.

$$\sigma_{KO}^2 = \sum_{i=1}^{N}\left(\lambda_i^{VEM}\bar{\gamma}_{iv}\right) - \bar{\gamma}_{vv} - \mu$$

Le système de krigeage est égal à la moyenne pondérée de la variable bruitée plus l'erreur de mesure. Les pondérateurs ($\boldsymbol{\lambda^{VEM}}$) sont les nouveaux poids qui minimisent la nouvelle variance d'estimation des expressions 23 et 24, où un nouveau terme est rajouté qui est la variance de l'erreur de mesure.

On suppose que les erreurs de mesure ne sont ni systématiques, ni corrélées avec la variable mesurée ni corrélées entre elles (cependant on a vu dans la Figure 3 une corrélation entre la VEM et la concentration qui provient de la méthode de calcul de ces deux valeurs).

Ces hypothèses sont nécessaires pour que les expressions 25 et 26 équivaillent aux expressions 7, 8 (KO) et 10, 11 (KS). La seule différence réside dans la diagonale de la matrice de covariance (membre de gauche du système de krigeage). Les termes de cette diagonale ne sont plus $\boldsymbol{C_{ii}}$ ***ou*** $\boldsymbol{\gamma_{ii}}$ mais $\boldsymbol{C_{ii}}$ ***+VEM***$_j$ ***ou*** $\boldsymbol{\gamma_{ii}}$ ***-VEM***$_j$. Les pondérateurs associés à ce nouveau système sont notés $\boldsymbol{\lambda^{VEM}}$.

Le modèle variographique utilisé pour l'estimation de l'ozone (en rouge dans la Figure 16) est identique au modèle ajusté précédemment, moins l'effet de pépite. On suppose en effet que la discontinuité du variogramme à l'origine est due en totalité à l'erreur de mesure (pas de microstructure prise en compte et modélisée par un effet de pépite, par exemple). Cette variabilité à l'origine n'est plus introduite sous la forme d'un effet de pépite mais point par point, par l'intermédiaire de la VEM.. La covariance correspondant à ce modèle de variogramme reste la même que précédemment, à l'exception de sa valeur à l'origine qui est réduite de l'effet de pépite. On obtient ***CELA (0)=140***.

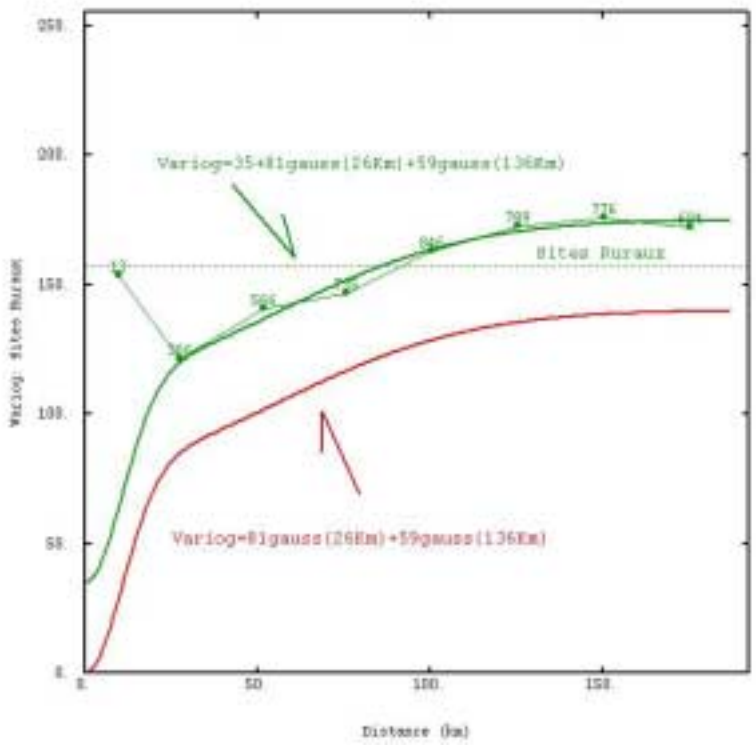


Figure 16 : Modèle pour l'estimation avec variance de l'erreur de mesure

En vert modèle ajusté, en rouge modèle sans effet de pépite

Le Tableau 7 montre les résultats des estimations avec VEM.

VARIABLE	Maille (Km)	Point/Bloc	Nb	Minimum	Maximum	Moyenne	Variance	CV
Estimation	5	P	2510	31.41	101.80	60.10	118.90	0.18
		B		31.70	101.55	60.10	117.87	0.18
	25	P	98	31.41	98.78	59.91	117.43	0.18
		B		36.57	95.20	60.01	104.53	0.17

Tableau 7 : Résultats de l'estimation avec VEM

Les concentrations maximales (autour de 100 µg/m^3) sont légèrement plus faibles dans le krigeage avec VEM. La variance et le coefficient de variation sont également moins élevés.

VARIABLE	Maille (Km)	P/B	Nb	Minimum	Maximum	Moyenne	Variance	CV
Ecart-type d'estimation	5	P	2510	2.06	10.75	6.54	2.59	0.25
		B		1.98	10.61	6.39	2.53	0.25
	25	P	98	2.68	9.54	7.06	2.61	0.23
		B		1.66	7.30	4.50	1.53	0.27

Tableau 8 : Résultats de l'écart-type de l'erreur estimation avec VEM

En ce qui concerne l'écart-type de l'erreur d'estimation, les statistiques sont approximativement identiques avec ou sans VEM, s'il s'agit d'un krigeage de bloc (à l'exception de quelques valeurs minimales légèrement plus faibles dans le krigeage avec VEM). En revanche, s'il s'agit d'estimations ponctuelles, les écart-types d'estimation sont beaucoup plus faibles avec VEM. Cela est dû au fait qu'en un point donné, le krigeage avec VEM cherche à estimer la concentration non bruitée et que l'effet de pépite ne s'ajoute pas à la variance de krigeage.

Le krigeage de bloc avec VEM a été réalisé sur des blocs de 5 km (cartes de la Figure 17 et histogrammes de la Figure 18). Un voisinage circulaire glissant de 50 km de rayon a été conservé.

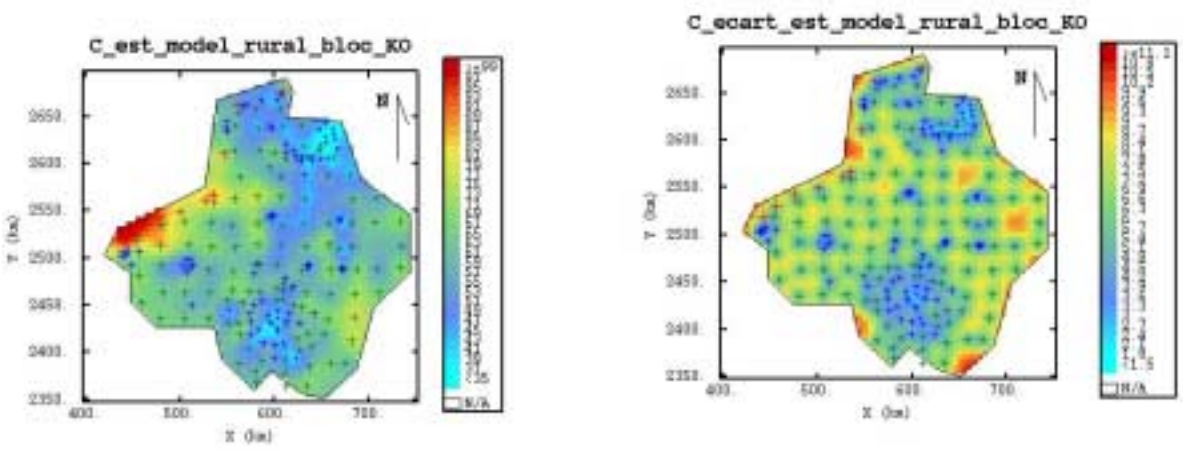


Figure 17 : Cartes d'estimation et de l'écart-type d'estimation de l'ozone pour la semaine du 26 juin au 3 juillet (Maille de 5 Km)

Des concentrations d'ozone sensiblement plus faibles (indiquées en bleu clair) sont visibles dans les agglomérations (surtout en Île de France et dans l'agglomération lilloise).

Les valeurs les plus fortes sont localisées dans la zone ouest du littoral (autour de 100 µg/m^3, couleur rouge), les valeurs intermédiaires dans les zones rurales (autour de 70 µg/m^3, couleur verte à jaune,) et dans les zones de transition entre les zones rurales et les agglomérations (autour de 60 µg/m^3, couleur vert claire).

L'écart-type d'estimation est compris entre 2 et 10.6 avec une moyenne de 6.4. Les zones urbaines, où la densité de mesures est la plus grande, se caractérisent par les écarts-types d'estimation les plus faibles (voisines de 3, couleur bleu ciel). Dans les zones rurales , les écarts-types d'estimation s'étendent entre 6 et 8 (couleur de verte à jaune). Les écarts les plus forts (voisins de 10, couleur orange) se trouvent en périphérie du domaine et dans quelques zones où manquaient des données.

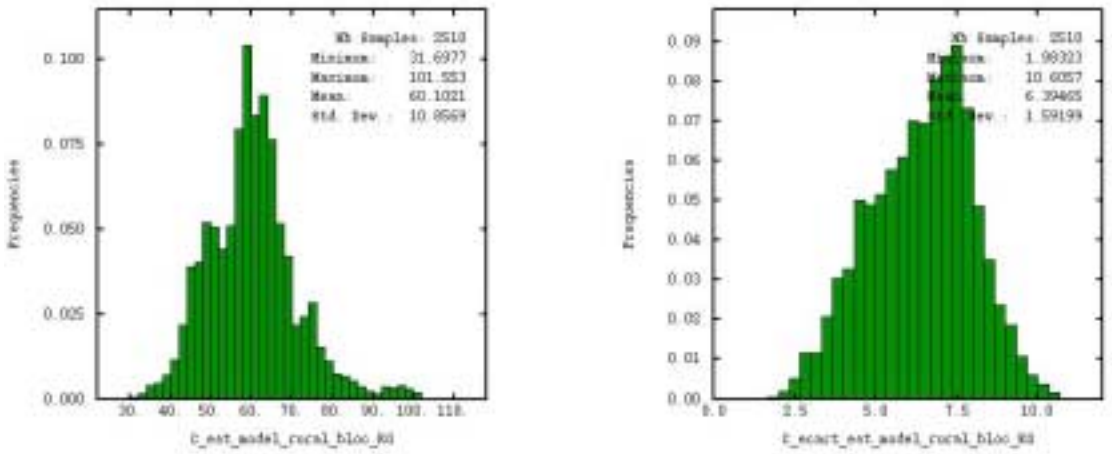


Figure 18 : Histogrammes de l'estimation et de l'écart d'estimation de bloc (Maille de 5 Km)

L'estimation de bloc a été également réalisée pour la maille de 25 km (Figure 19) -même voisinage d'estimation et même échelle de couleur que précédemment. La nouvelle carte fait voir de la même manière des concentrations plus faibles dans les agglomérations des valeurs fortes dans la zone ouest du littoral, et des valeurs intermédiaires dans les zones rurales.

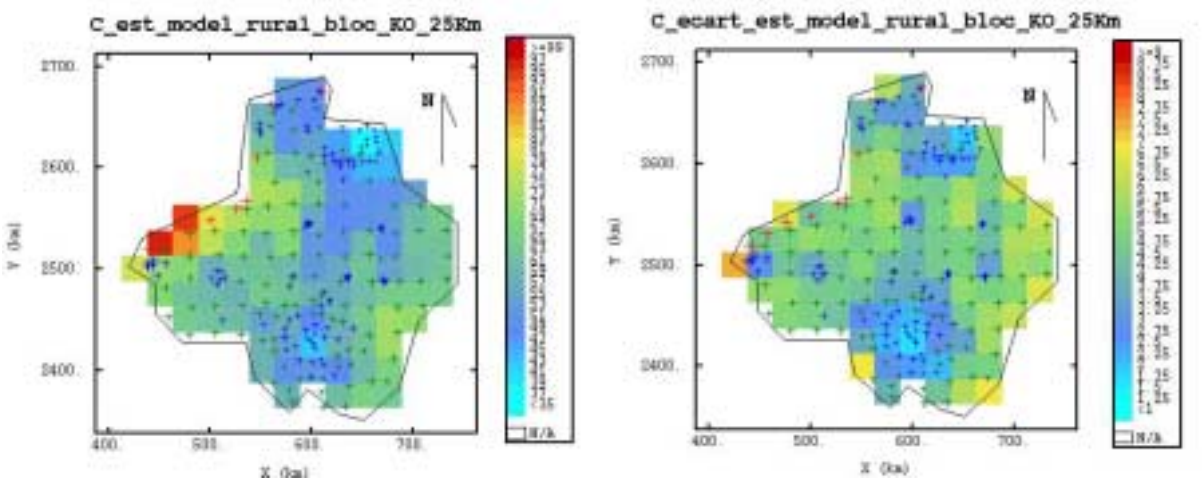


Figure 19 : Cartes d'estimation et de l'écart-type d'estimation de l'ozone pour la semaine du 26 juin au 3 juillet (Maille de 25 Km)

Les résultats de l'estimation ponctuelle (maille de 25 km) sont comparés à ceux de l'estimation de bloc (Figure 20).

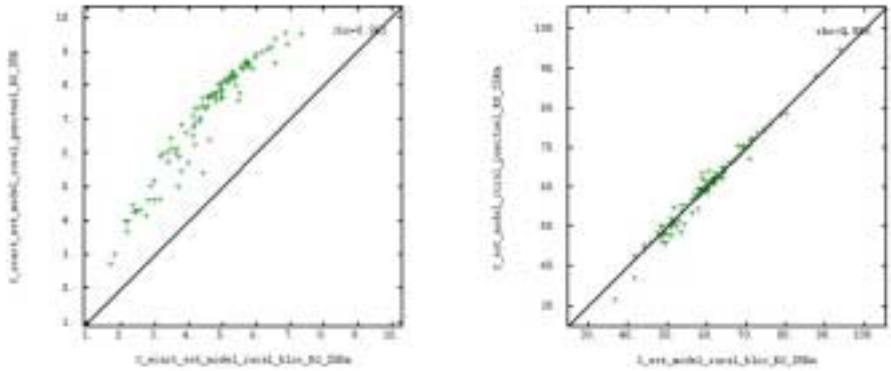


Figure 20 : Comparaison entre krigeage du bloc et krigeage ponctuel (estimation et écart-type d'estimation de l'ozone) pour la semaine du 26 juin au 3 juillet (Maille de 25 Km)

Les valeurs estimées sont très proches mais l'écart-type d'estimation associé au krigeage ponctuel est plus grand. Ainsi le type de support (point ou bloc) et la taille de la maille d'estimation sont des paramètres importants de la cartographie. C'est pourquoi, dans la présentation d'une carte de pollution, il est conseillé de préciser le support d'estimation utilisé.

Le Tableau 9 présente les résultats de l'estimation avec VEM aux nœuds des maillages de 5 km et 25 km les plus proches du point 117 (mêmes nœuds que dans le Tableau 6).

No.	Item	Exp. ($\lambda=\lambda^{VEM}$)	Bloc ou point les plus près du TP No. 117(Voir Figure 10, Figure 21)							
			Maille : 5 Km				Maille : 25 Km			
			Bloc		Ponctuel		Bloc		Ponctuel	
			KO	KS	KO	KS	KO	KS	KO	KS
1	Nb des données dans le voisinage		19				24			
2	Moyenne connue		N/A	0	N/A	0	N/A	0	N/A	0
3	Poids assigné à la moy. (%)	10		1.3		1.25		1.07		0.9
4	Covariance moyenne de bloc ou du point C(0)		137,1		140		98,0		140	
5	Variance de Z* (Estimée Z)	5	112,7	111,5	114,6	113,4	71,5	70,5	72,3	71,4
6	Covariance entre Z et Z*	6	112,1		114,0		71,0		71,9	
7	Paramètre de Lagrange	7	-0,6	N/A	-0,6	N/A	-0,5	N/A	-0,4	N/A
8	Variance d'estimation	3 ou KO :8 et KS :11	25,6	25,6	26,6	26,6	27,6	27,6	68,6	68,6
9	écart-type d'estimation		5,1	5,1	5,2	5,2	5,2	5,2	8,3	8,3
10	Valeur Estimée	KO :1 et KS :9	60,4	59,6	60,3	59,6	62,2	61,5	62,1	61,6
11	Pente de la régression Z|Z*	25	1,0	N/A	1,0	N/A	1,0	N/A	1,0	N/A
12	Corrélation entre Z et Z*	26	0,90	0,90	0,90	0,90	0,85	0,85	0,71	0,71

Tableau 9 : Comparaison de Statistiques des estimations pour une valeur estimée choisie

La comparaison entre le Tableau 6 et le Tableau 9 met clairement en évidence l'effet de la covariance à l'origine sur la variance d'estimation du krigeage ponctuel: la différence entre ces deux variances est strictement égale la valeur de l'effet de pépite: 35.

Ci-après sont vérifiés certains des résultats numériques fournis par Isatis (Tableau 9). La vérification porte sur les résultats du krigeage ordinaire de bloc (maille de 5 km) et du krigeage ponctuel à moyenne connue (maille de 25 Km).

Calcul de la variance d'estimation pour le bloc 5Km en KO :

$$\sigma^2_{KO} = Var[Z_V] + Var\left[Z_v^*\right] - 2Cov\left[Z_V, Z_v^*\right]$$

$$\sigma^2_{KO} = 137.1 + 112.7 - 2(112.1) = 287.7 - 224.2$$

$$\underline{\sigma^2_{KO} = 25.6} \qquad \underline{\sigma_{KO} = \sqrt{25.6} = 5.06}$$

$$\sigma^2_{KO} = \sum_{i=1}^{N} (\lambda_i^{VEM}\gamma_{iV}) - \gamma_{VV} - \mu = -Cov\left[Z_V, Z_v^*\right] + Var[Z_V] - \mu$$

$$\sigma^2_{KO} = -112.1 + 137.1 - (-0.6) \quad \underline{\sigma^2_{KO} = 25.6}$$

Calcul de la variance d'estimation ponctuelle pour la maille de 25 Km en KS:

$$\sigma^2_{KS} = Var[Z_V] + Var\left[Z_v^*\right] - 2Cov\left[Z_V, Z_v^*\right]$$

$$\sigma^2_{KS} = 140 + 71.4 - 2(71.4) = 140 + -71.4$$

$$\underline{\sigma^2_{KS} = 68.6} \qquad \underline{\sigma_{KS} = \sqrt{68.6} = 8.28}$$

$$\sigma^2_{KS} = \sum_{i=1}^{N} (\lambda_i^{VEM}\gamma_{iV}) - \gamma_{VV} = -Cov\left[Z_V, Z_v^*\right] + Var[Z_V]$$

$$\sigma^2_{KS} = -71.4 + 140 \qquad \underline{\sigma^2_{KS} = 68.6}$$

Vérification du système de krigeage ordinaire pour le block de 5 Km:

$$\sum_{j=1}^{N} (\lambda_j^{VEM}\gamma_{ij}) - \mu = \gamma_{iV}$$

$$-Var\left[Z_v^*\right] - \mu = -Cov\left[Z_V, Z_v^*\right]$$

$$-112.7 - (-0.6) = -112.1$$

$$-112.1 = -112.1$$

Vérification du système de krigeage simple ponctuelle pour la maille de 25 Km :

$$\sum_{j=1}^{N} (\lambda_j^{VEM}\gamma_{ij}) = \gamma_{iV}$$

$$Var\left[Z_v^*\right] = Cov\left[Z_V, Z_v^*\right]$$

$$71.4 = 71.4$$

La Figure 21 représente deux exemples de voisinage d'estimation (krigeage ponctuel, maille de 25 km, krigeage de bloc, maille de 5 km). Les valeurs numériques attribuées à chaque donnée sont les pondérateurs (en %), solutions du système de krigeage.

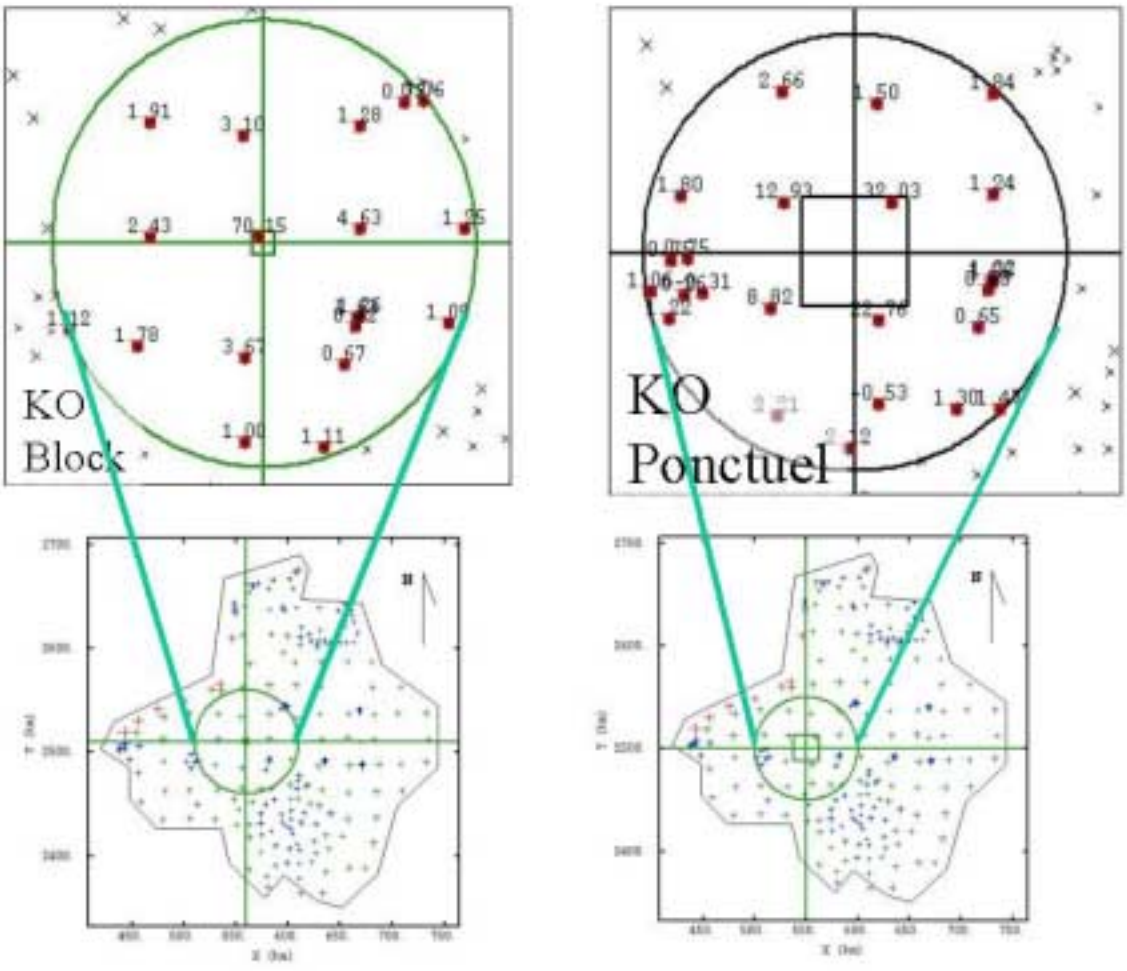


Figure 21 : Exemple de poids de krigeage avec VEM, maille de 5KM

Pour le nœud de la maille de 5 Km, on voit que l'effet d'écran est produit par un seul échantillon (le point 117), car le poids qui lui a été attribué est de l'ordre de 70%. Pour le nœud de la maille de 25 Km, l'effet d'écran est formé par 4 échantillons dont la somme des pondérateurs est de 77%. Ces observations confirment que la modélisation à proximité de l'origine est la plus importante (comportement à petites distances).

A titre de comparaison, la variance d'interpolation (Yamamoto, 20002), dont le concept est présenté dans le corps du rapport, a été calculée en krigeage ponctuel et en krigeage de bloc pour les points ou blocs les plus proches des tubes 117, 127 et 165.

Soit une concentration estimée Z* : $Z^* = \sum_i \lambda_i Z_i$. On rappelle que la variance d'interpolation est obtenue par une combinaison linéaire des écarts entre Z* et les Z_i : $\sigma^2_{\text{int}\,erp} = \sum_i \lambda_i (Z^* - Z_i)^2$

Les pondérateurs sont égaux aux poids de krigeage, après une éventuelle correction si ces derniers sont négatifs. Deux types de correction ont été effectués (on note λ'_i les poids corrigés):

Premier type de correction :

$$\lambda_i' = 0 \; si \lambda_i < 0$$

$$\lambda_i' = \frac{\lambda_i}{\sum_{i,\lambda_i>0} \lambda_i} \; si \lambda_i \geq 0$$

Deuxième type de correction :

Soit c défini par :

$c = -\min\{\lambda_i, i = 1...n\}$, si $\min\{\lambda_i\}<0$

c=0, si $\min\{\lambda_i\}\geq 0$

$$\lambda_i' = \frac{\lambda_i + c}{\sum_i (\lambda_i' + c)}$$

Point ou bloc cible		Krigeage ponctuel avec effet de pépite	Krigeage ponctuel avec VEM	Krigeage de bloc avec effet de pépite	Krigeage de bloc avec VEM
Point ou bloc le plus proche du tube 117	Variance d'interpolation (sans correction, tous les poids étaient positifs)	40,4	41,9	37,9	39,3
	Variance de krigeage	61,6	25,7	26,6	25,6
	Variance des données dans le voisinage de krigeage	145,6	145,6	145,6	145,6
Point ou bloc le plus proche du tube 127	Variance d'interpolation Correction de type 1	204,9	202,9	151,7	150,2
	Variance d'interpolation Correction de type 2	191,8	191,2	144,7	144,3
	Variance de krigeage	46,4	10,8	11,2	10,6
	Variance des données dans le voisinage de krigeage	120,3	120,3	120,3	120,3
Point ou bloc le plus proche du tube 165	Variance d'interpolation Correction de type 1	176,6	174,3	237,2	233,4
	Variance d'interpolation Correction de type 2	160,9	159,7	217,7	215,2
	Variance de krigeage	43,0	7,46	4,26	3,93
	Variance des données dans le voisinage de krigeage	134,8	134,8	134,8	134,8

Tableau 10 – Calcul de la « variance d'interpolation» et comparaison avec la variance de krigeage

A la différence de la variance de krigeage, la variance d'interpolation dépend de la valeur numérique des données à l'intérieur du voisinage. Près des points 127 et 165 où, sur des distances relativement courtes, des concentrations assez contrastées ont été mesurées, la variance d'interpolation augmente effectivement. Elle est nettement supérieure à la variance de krigeage. En revanche, elle ne tient pas compte de l'effet du support de l'estimation (point ou bloc). Quelques différences apparaissent selon la méthode utilisée pour corriger les poids négatifs. Le lien avec la variance des données dans le voisinage est peu évident. Les valeurs les plus proches du bloc ou du point cible semblent influer le plus sur cet indicateur.

La variance d'interpolation n'apporte pas les mêmes informations que la variance de krigeage, à laquelle elle ne saurait donc se substituer. Toutefois, elle semble fournir une information complémentaire intéressante sur l'incertitude d'estimation liée à la variabilité locale des données.

10 Analyses des incertitudes

Après avoir décrit les méthodes de krigeage qui permettent d'effectuer des estimations et recensé les facteurs qui influent sur la valeur numérique de l'écart-type de l'erreur d'estimation (désormais appelé écart-type de krigeage), nous essaierons d'interpréter les résultats obtenus et de leur donner un sens pratique.

La carte des valeurs estimées est une moyenne pondérée des mesures qui se trouvent dans un voisinage d'estimation. Lorsqu'on réalise un krigeage de bloc, la valeur estimée représente la moyenne dans tout le bloc. En revanche, dans un krigeage ponctuel, la valeur de concentration est estimée au centre du bloc. Comme on l'a vu, ces deux valeurs sont très proches mais les écart-types de krigeage peuvent être très différents.

Les estimations ponctuelles présentent toujours des écart-types de krigeage supérieurs à ceux des estimations par bloc. Cela signifie que l'erreur commise en estimant une concentration en un point est plus importante que l'erreur commise en estimant une concentration représentative d'une surface.

La concentration estimée en un point de mesure (estimation ponctuelle), avec un modèle sans effet de pépite est égale à la valeur mesurée, puisque le krigeage est un interpolateur exact, et l'écart-type d'estimation vaut 0. Au fur et à mesure que l'on s'éloigne de ce point, l'influence de l'échantillon sur la valeur estimée s'atténue et l'écart-type d'estimation augmente.

Les cartes de l'écart-type d'estimation de la Figure 17 et de la Figure 19 montrent ainsi que les valeurs les plus fortes sont localisées dans des zones sans données tandis que les plus faibles se trouvent dans des zones pourvues d'une grande densité de mesures, comme c'est le cas des agglomérations.

L'utilisation conjointe de la carte d'estimation et de la carte des écarts-types d'estimation permet d'avoir une idée de ***la valeur moyenne de la concentration d'ozone en un point ou sur une surface*** et, en même temps, de repérer ***qualitativement*** les zones où cette valeur moyenne est ***plus ou moins*** proche de la valeur vraie inconnue.

11 Critère d'incertitude des Directives Européennes

Les directives européennes demandent que l'incertitude des résultats des modèles par rapport à la valeur réelle ne dépasse pas 30 à 50% selon le polluant et l'échelle temporelle considérés. Aucun critère n'est fourni pour les concentrations hebdomadaires d'ozone. Nous retenons ici une valeur de 50% qui correspond à l'objectif fixé pour la concentration horaire d'ozone et la moyenne journalière de NO_2. Ce critère est peu aisé à interpréter car la valeur réelle reste inconnue, cependant le sens que nous lui donnons est que la valeur estimée de la concentration doit être comprise dans un intervalle allant de 0.5 fois à 1.5 fois la valeur réelle:

$$(Z_v - 0.5 * Z_v) \leq Z_v^* \leq (Z_v + 0.5 * Z_v) \quad \textit{Intervalle de 50\% d'incertitud e sur la valeur reélle}$$

$$Z_v : \textit{Valeur reelle} \qquad Z_v^* : \textit{Valeur estimée}$$

On troue ainsi les conditions qui doivent être respectées pour avoir une incertitude inférieure ou égale à 50% de la valeur réelle:

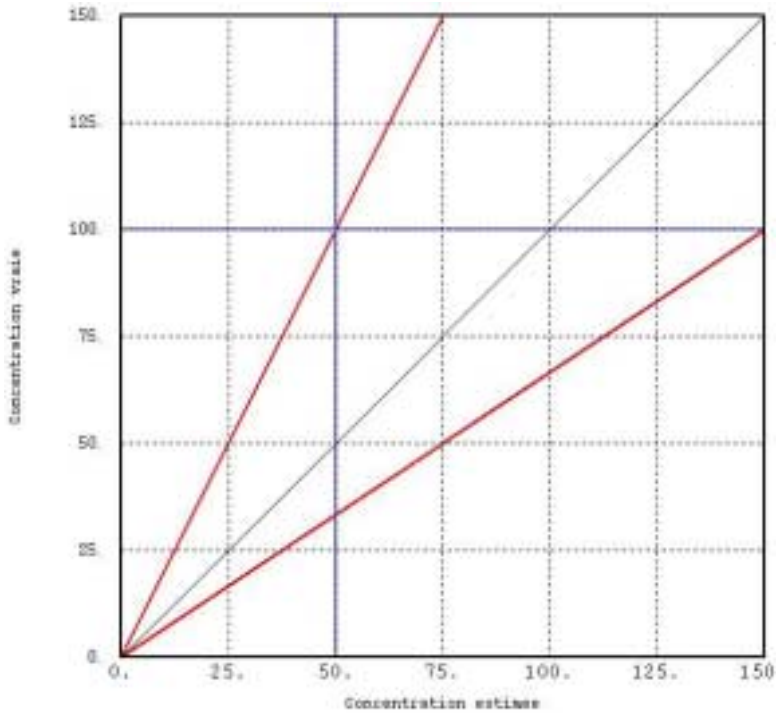
Borne Supérieure pour Z_v :

$$(Z_v - 0.5 * Z_v) \leq Z_v^*$$
$$Z_v^* \geq 0.5 * Z_v$$
$$Z_v \leq 2 * Z_v^*$$

Borne Inférieure pour Z_v :

$$Z_v^* \leq (Z_v + 0.5 * Z_v)$$
$$Z_v^* \leq 1.5 * Z_v$$
$$Z_v \geq \frac{2}{3} * Z_v^*$$

Si ces inégalités sont respectées alors l'incertitude sera inférieure à 50% de la valeur réelle dans les deux sens (positif et négatif). Toutefois la valeur réelle reste inconnue. En faisant certaines hypothèses sur les lois de distribution on peut néanmoins calculer l'intervalle des valeurs de concentration dans lequel peut se trouver la valeur réelle inconnue.



Lignes épaisses rouges : Bornes inférieure (resp. supérieure) et supérieure (resp. inférieure) des valeurs vraies (resp. estimées) pour avoir une incertitude maximale de 50% de la valeur vraie, la borne inférieure (resp. supérieure) corresponde à 0.5 fois la valeur vraie (resp. 2 fois la valeur estimée) et la borne supérieure (resp. inférieure) à 1.5 fois la valeur vraie (resp. 2/3 fois la valeur estimée).

Figure 22 : Critère d'incertitude des directives européennes

Les lignes rouges de la Figure 22 représentent les limites entre lesquelles la valeur estimée doit être comprise pour avoir une incertitude inférieure à 50% de la valeur réelle. Par exemple, pour une valeur de concentration réelle de 100 µg/m^3, la valeur de l'estimation, et donc les bornes de l'intervalle de confiance, doivent être comprises entre 50 µg/m^3 et 150 µg/m^3. Inversement, si la valeur estimée est de 50 µg/m^3, la valeur réelle qu'elle estime doit être comprise entre 33 µg/m^3 et 100 µg/m^3.

Ces limites ne sont pas symétriques, elles sont directement corrélées à la valeur de la concentration vraie. En effet, plus la valeur réelle est grande, plus l'intervalle d'incertitude maximale de 50% est large. Cela signifie que pour les fortes concentrations, on s'autorise à avoir un intervalle de confiance plus large autour de la concentration estimée.

Par exemple pour une valeur réelle de 100 µg/m^3 les valeurs estimées doivent être comprises entre 50 µg/m^3 et 150 µg/m^3 (intervalle de 100 µg/m^3), tandis que pour une valeur de concentration réelle de 25 µg/m^3 les valeurs estimées doivent être comprises entre 12.5 µg/m^3 et 37.5 µg/m^3 (intervalle de 25 µg/m^3).

On va présenter trois méthodes grâce auxquelles on peut déterminer des intervalles de confiance, la première méthode consiste à prendre en compte les résultats de l'estimation par krigeage ordinaire et l'écart-type de l'erreur d'estimation. La deuxième méthode est fondée sur l'espérance conditionnelle et la troisième utilise les simulations conditionnelles.

Une fois les intervalles de confiance calculés, on pourra les comparer avec les bornes inférieures et supérieures correspondant à une incertitude de 50%, afin de déterminer si l'on respecte on non la réglementation européenne.

12 Intervalles de confiance : critère de l'écart-type de l'erreur d'estimation

Si on suppose, que la distribution de fréquences des erreurs d'estimations autour de leur moyenne 0 obéit à la loi de Gauss, on peut donner avec exactitude la marge d'erreur de la valeur estimée.

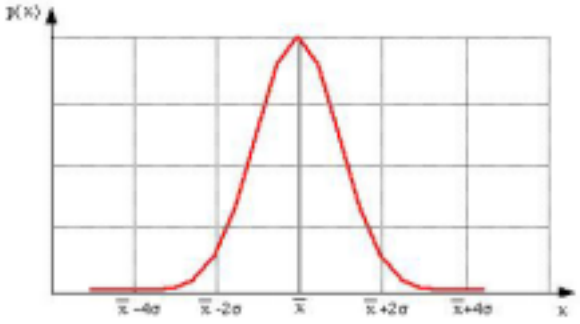


Figure 23 : La loi de Gauss

Une distribution gaussienne est entièrement déterminée par sa moyenne et son écart-type. La moyenne des erreurs d'estimations étant nulle par définition, il s'ensuit que l'écart-type suffit à donner une connaissance exhaustive de leur distribution.

Sous cette hypothèse l'erreur d'estimation est donc parfaitement caractérisée par son écart-type et il est évident que cette connaissance statistique est la seule à laquelle on puisse accéder. Si en effet l'erreur était exactement déterminée elle cesserait par là même de mériter ce nom puisqu'il suffirait de corriger en conséquence l'estimation qu'elle affecte.

Si on caractérise la précision d'une estimation par l'erreur qui a moins de 5 chances sur 100 d'être dépassée dans un sens ou dans l'autre, on obtient l'intervalle de confiance à 95%. La marge d'erreur ainsi définie est donc celle qui est obtenue en reportant, de part et autre de l'estimation, deux fois (très précisément 1,96 fois) la valeur de l'écart-type.

Si on suppose une distribution de l'erreur non gaussienne, continue et uni modale, la marge de l'erreur pour obtenir un intervalle de confiance à 95% est de +/- 3 fois l'écart-type (Chiles, 1999, page 177). L'ignorance de la distribution de l'erreur d'estimation conduit donc à des intervalles de confiance plus large (de largeur de 6σ, au lieu de 4σ si la distribution est supposée gaussienne).

$$(Z_v^* - 2*\sigma_K) \leq Z_v \leq (Z_v^* + 2*\sigma_K) \quad \textit{Intervalle de confiance à 95\% avec distribution de l'erreur gaussienne}$$

$$(Z_v^* - 3*\sigma_K) \leq Z_v \leq (Z_v^* + 3*\sigma_K) \quad \textit{Intervalle de confiance à 95\% sans restriction sur la distribution des erreurs}$$

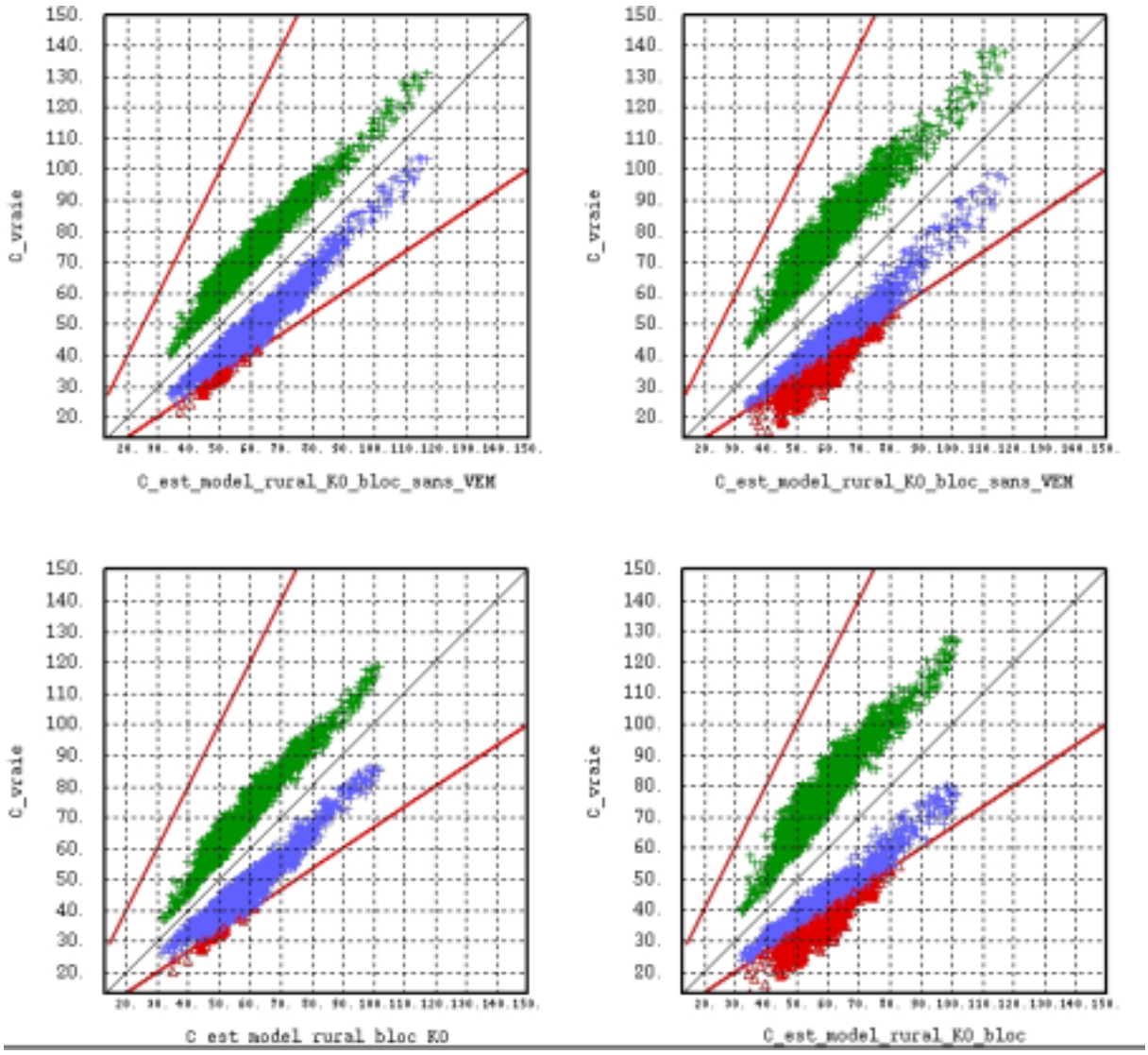
La Figure 24 et la Figure 25 montrent, pour les quatre estimateurs étudiés (krigeage ponctuel et de bloc, avec et sans VEM), les nuages de corrélation entre les valeurs estimées et les deux intervalles de confiance à 95% (+/- deux fois et +/- 3 fois l'écart-type).

Dans ces figures, l'axe X correspond aux valeurs estimées. Les croix vertes indiquent la limite supérieure et les croix bleues la limite inférieure de l'intervalle de confiance (axe Y).

La distance que sépare les croix vertes et les croix bleues représente donc l'intervalle de valeurs probables de la concentration réelle pour une valeur estimée donnée, en prenant en compte les hypothèses faites sur la distribution de l'erreur d'estimation; on observe donc que ces intervalles sont symétriques par rapport à la valeur estimée.

Par ailleurs, les lignes épaisses rouges représentent les bornes inférieure et supérieure entre lesquelles les valeurs estimées doivent se trouver pour que l'incertitude de l'estimation relativement à la valeur vraie soit inférieure à 50%.

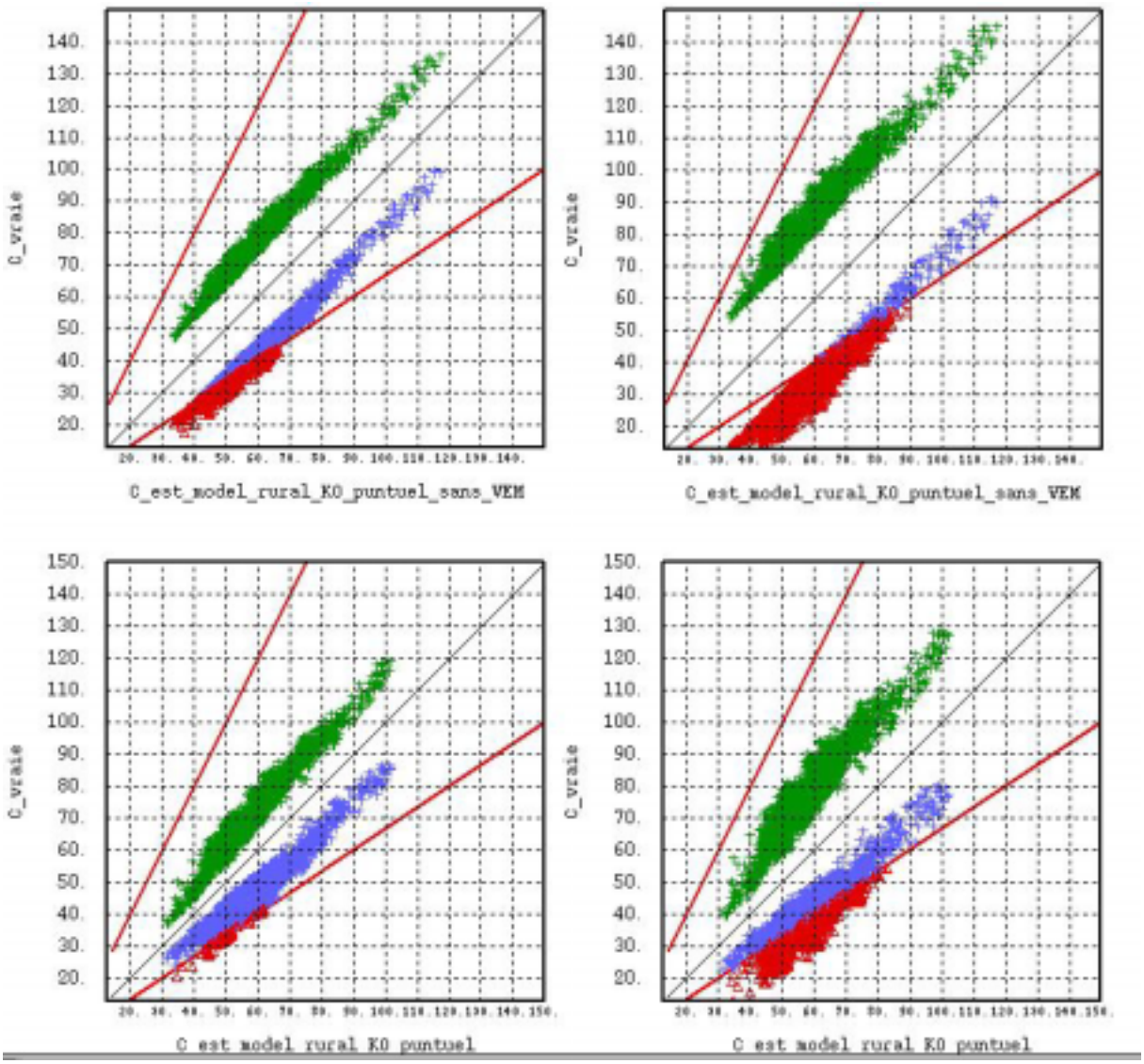
Dans ces figures on a repéré avec des triangles rouges les valeurs qui ne respectent pas ces limites.



Figures en haut: Intervalles de confiance pour une estimation par krigeage ordinaire sans VEM
Figures en bas: Intervalles de confiance pour une estimation par krigeage ordinaire avec VEM
Figures à gauche : Intervalles de confiance en prenant en compte deux fois l'écart-type
Figures à droite : Intervalles de confiance en prenant en compte trois fois l'écart-type
Croix bleues : Limite inférieure de l'intervalle de confiance à 95%.
Croix vertes: Limite supérieure de l'intervalle de confiance à 95%.
Lignes épaisses rouges : Bornes inférieure (resp. supérieure) et supérieure (resp. inférieure) des valeurs vraies (resp. estimées) pour avoir une incertitude maximale de 50% relativement à la valeur vraie. La borne inférieure (resp. supérieure) correspond à 0.5 fois la valeur vraie (resp. 2 fois la valeur estimée) et la borne supérieure (resp. inférieure) à 1.5 fois la valeur vraie (resp. 2/3 fois la valeur estimée).
Triangles rouges : Valeurs à l'extérieur des bornes

Nb Nœuds	Critère	Intervalle	VEM	Minimum	Maximum	Moyenne	écart-type
2510	+/-2 σ	Minimale	Sans	21,4	103,8	48,6	11,0
			Avec	20,0	86,2	47,3	9,8
		Maximale	Sans	40,0	130,7	74,2	12,6
			Avec	36,9	118,6	72,9	12,7
	+/-3 σ	Minimale	Sans	13,5	98,3	42,2	11,1
			Avec	12,6	80,1	40,9	9,6
		Maximale	Sans	43,2	138,6	80,6	13,4
			Avec	39,3	127,8	79,3	13,8

Figure 24 : Statistiques et nuages de corrélation entre la valeur estimée (axe X) et les bornes des intervalles de confiance (axe Y) pour les estimations de bloc par krigeage ordinaire sur la maille de 5Km



Figures en haut: Intervalles de confiance pour une estimation par krigeage ordinaire sans VEM
Figures en bas: Intervalles de confiance pour une estimation par krigeage ordinaire avec VEM
Figures à gauche : Intervalles de confiance en prenant en compte deux fois l'écart-type
Figures à droite : Intervalles de confiance en prenant en compte trois fois l'écart-type
Croix bleues : Limite inférieure de l'intervalle de confiance à 95%.
Croix vertes: Limite supérieure de l'intervalle de confiance à 95%.
Lignes épaisses rouges : Bornes inférieure (resp. supérieure) et supérieure (resp. inférieure) des valeurs vraies (resp. estimées) pour avoir une incertitude maximale de 50% relativement à la valeur vraie. La borne inférieure (resp. supérieure) correspond à 0.5 fois la valeur vraie (resp. 2 fois la valeur estimée) et la borne supérieure (resp. inférieure) à 1.5 fois la valeur vraie (resp. 2/3 fois la valeur estimée).
Triangles rouges : Valeurs à l'extérieur des bornes

Nb Nœuds	Support	Intervalle	VEM	Minimum	Maximum	Moyenne	écart-type
2510	+/-2 σ	Minimale	Sans	17,2	99,1	43,6	11,1
			Avec	19,6	86,2	47,0	9,8
		Maximale	Sans	47,1	135,7	79,1	12,3
			Avec	36,9	119,0	73,2	12,7
	+/-3 σ	Minimale	Sans	7,2	91,0	34,7	11,1
			Avec	12,1	80,0	40,5	9,7
		Maximale	Sans	53,8	144,9	88,0	12,8
			Avec	39,4	128,4	79,7	13,8

Figure 25 : Statistiques et nuages de corrélation entre la valeur estimée (axe X) et les bornes des intervalles de confiance (axe Y) pour les estimations ponctuelles par krigeage ordinaire sur la maille de 5Km

Les statistiques des intervalles de confiance des estimations de bloc calculées avec et sans la VEM sont similaires. Dans le cas d'une distribution gaussienne de l'erreur, la borne inférieure moyenne de l'intervalle oscille entre 47 et 48 µg/m^3 et la borne supérieure entre 73 et 74 µg/m^3. Dans le cas d'une distribution non gaussienne de l'erreur, la borne inférieure moyenne de l'intervalle oscille entre 41 et 42 µg/m^3 et la borne supérieure entre 79 et 80 µg/m^3.

Pour les estimations ponctuelles le fait de prendre en compte on non la VEM est plus important car les statistiques sont plus différents que dans le cas des estimations de bloc. Dans les nuages de la partie supérieure de la Figure 25 qui correspondent aux estimations sans VEM on observe que la distance entre les intervalles est plus grande que dans les figures de la partie inférieur correspondant aux estimations avec VEM.

Par conséquent, les estimations les plus sensibles sont les estimations ponctuelles sans VEM : dans ces estimations l'écart-type d'estimation est influencé par une forte valeur de la variance du point à estimer (laquelle est égale à 175 car on prend en compte l'effet de pépite). Comme les écart-types sont plus élevés dans ce type d'estimation, on rencontre nécessairement des intervalles de confiance plus larges.

Quand on compare les intervalles de confiance avec le critère d'incertitude ,on peut conclure que :

- Si on suppose que les erreurs ont une distribution gaussienne, l'ensemble des estimations respecte l'objectif de qualité de 50% d'incertitude (2 et 3% seulement des estimations sont en dehors des limites, voir le Tableau 11). En revanche, dans l'estimation ponctuelle sans VEM, le pourcentage des valeurs à l'extérieur des limites atteint 22.7%.

- Sans hypothèse sur la distribution des erreurs, près de la moitié des estimations de bloc et des estimations ponctuelles avec VEM sont susceptibles de ne pas satisfaire à l'objectif de qualité (entre 41 et 46% des estimations, voir le Tableau 12). L'estimation ponctuelle sans VEM présente les résultats les moins satisfaisants (seulement 6.5% des valeurs remplissent de façon sûre la condition d'incertitude).

Nb Nœuds	Support	VEM	Nb valeurs	%
2510	Bloc	Sans	51	2
		Avec	50	2
	Ponctuelle	Sans	569	22.7
		Avec	66	2.6

Tableau 11 : Nombre de valeurs estimées dont l'incertitude peut dépasser 50%, avec distribution de l'erreur gaussienne (KO, maille de 5 Km)

Nb Nœuds	Support	VEM	Nb valeurs	%
2510	Bloc	Sans	1028	41
		Avec	1067	42.5
	Ponctuelle	Sans	2347	93.5
		Avec	1158	46

Tableau 12 : Nombre de valeurs estimées dont l'incertitude peut dépasser 50%, sans contrainte d'une distribution gaussienne de l'erreur (KO, maille de 5 Km)

Une autre façon de présenter les résultats est de comparer les deux intervalles de confiance à 95% avec les intervalles d'incertitude de 50% comme suit :

Condition pour la borne inférieure, pour un intervalle de confiance à 95%, avec distribution gaussienne de l'erreur :

$$\begin{aligned}(Z_v^* - 2*\sigma_K) &\geq (\tfrac{2}{3}*Z_v^*)\\ -2*\sigma_K &\geq (\tfrac{2}{3}*Z_v^* - Z_v^*)\\ -2*\sigma_K &\geq -\tfrac{1}{3}*Z_v^*\\ -2*\frac{\sigma_K}{Z_v^*} &\geq -\frac{1}{3}\\ \frac{\sigma_K}{Z_v^*} &\leq \frac{1}{6} \leq 0.167\end{aligned}$$

Condition pour la borne supérieure, pour un intervalle de confiance à 95%, avec distribution gaussienne de l'erreur :

$$\begin{aligned}(Z_v^* + 2*\sigma_K) &\leq (2*Z_v^*)\\ 2*\sigma_K &\leq (2*Z_v^* - Z_v^*)\\ \sigma_K &\leq \frac{1}{2}*Z_v^*\\ \frac{\sigma_K}{Z_v^*} &\leq \frac{1}{2} \leq 0.5\end{aligned}$$

Condition pour la borne inférieure, pour un intervalle de confiance à 95%, sans l'hypothèse d'une distribution gaussienne de l'erreur:

$$\begin{aligned}(Z_v^* - 3*\sigma_K) &\geq (\tfrac{2}{3}*Z_v^*)\\ -3*\sigma_K &\geq (\tfrac{2}{3}*Z_v^* - Z_v^*)\\ -3*\sigma_K &\geq -\tfrac{1}{3}*Z_v^*\\ -3*\frac{\sigma_K}{Z_v^*} &\geq -\frac{1}{3}\\ \frac{\sigma_K}{Z_v^*} &\leq \frac{1}{9} \leq 0.111\end{aligned}$$

Condition pour la borne supérieure, pour un intervalle de confiance à 95% sans restriction sur la distribution de l'erreur:

$$\begin{aligned}(Z_v^* + 3*\sigma_K) &\leq (2*Z_v^*)\\ 3*\sigma_K &\leq (2*Z_v^* - Z_v^*)\\ \sigma_K &\leq \frac{1}{3}*Z_v^*\\ \frac{\sigma_K}{Z_v^*} &\leq \frac{1}{3} \leq 0.33\end{aligned}$$

Ces résultats signifient que :

- Si le rapport entre l'écart-type d'estimation et la valeur estimée est inférieur ou égal à 0.11, alors l'incertitude d'estimation est inférieure ou égale à 50%.
- Si ce rapport est supérieur à 0.11 et inférieur ou égal à 0.17, il faudrait supposer que les erreurs soient gaussiennes pour affirmer que cette incertitude est inférieure ou égale à 50%.
- Si ce rapport est supérieur à 0.17, alors le respect de l'objectif de qualité n'est pas garanti ; l'incertitude peut dépasser 50%.

Pour montrer l'application de cette méthodologie, on a calculé ce rapport pour les quatre types de krigeage. La Figure 26 montre les délimitations des zones dans lesquelles l'incertitude est inférieure à 50% de la valeur réelle, en faisant les deux hypothèses sur la distribution de l'erreur d'estimation.

Dans les cartes de cette figure, les zones en jaune représentent l'hypothèse la plus contraignante, c'est-à-dire sans restriction sur la distribution de l'erreur (cette distribution est simplement supposée continue et uni modale). Les zones vertes viennent se rajouter si on suppose une erreur gaussienne. Les zones rouges sont celles où

l'incertitude peut excéder 50%, quelle que soit l'hypothèse retenue. Comme on pouvait s'y attendre, ces zones correspondent aux régions où la densité des mesures est faible.

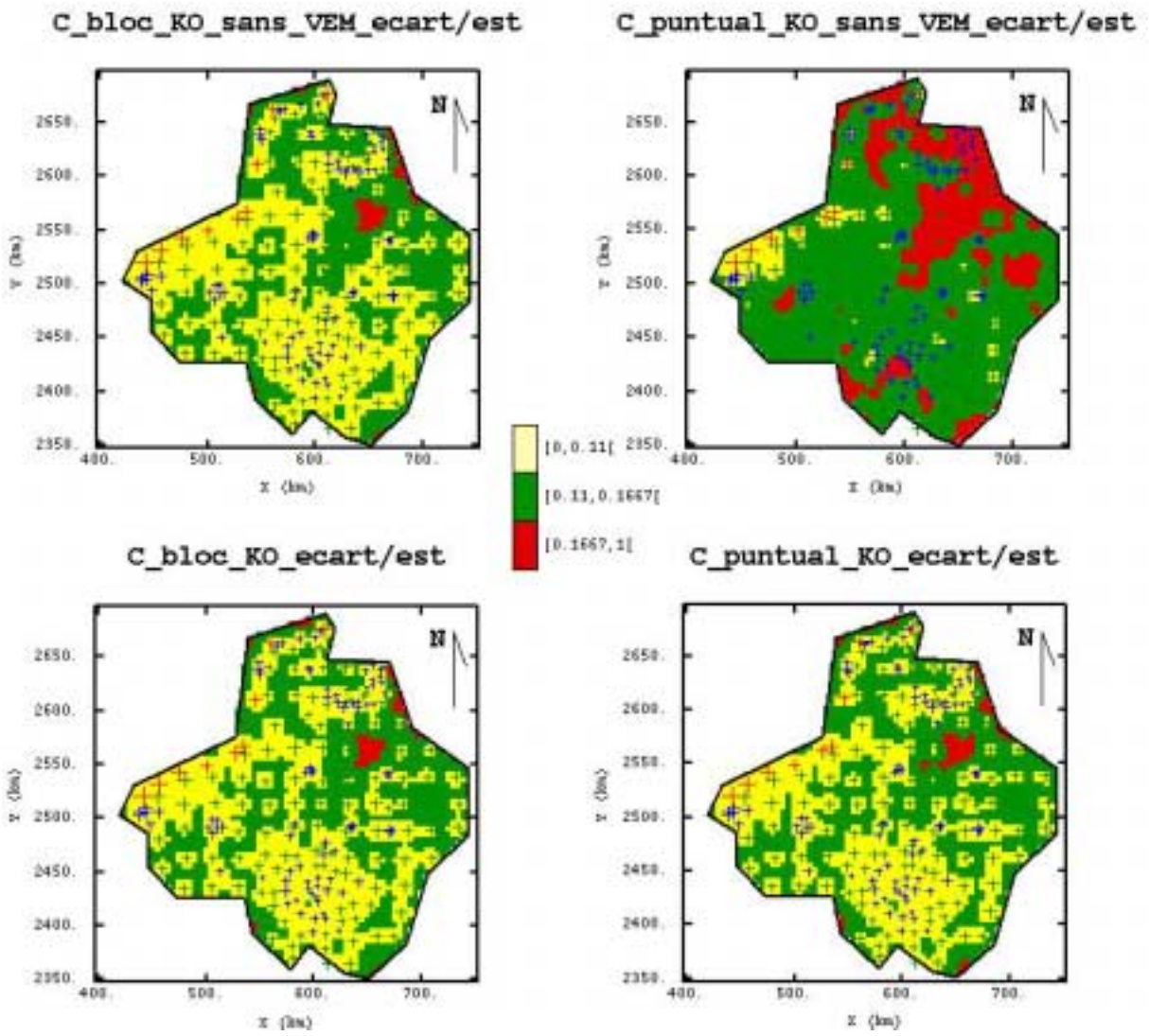


Figure 26 : Délimitation des zones d'incertitude en prenant en compte la relation écart/estimation, pour 4 types de krigeage, maille de 5 km

13 Intervalles de confiance : critère de l'espérance conditionnelle

Limites de la géostatistique linéaire

Avec la géostatistique « linéaire » on a estimé la valeur de la concentration de l'ozone sur le domaine d'étude à partir d'une combinaison linéaire des mesures des échantillonneurs passifs (le krigeage). On a également obtenu la variance (ou écart-type) d'estimation (variance de la différence entre valeurs vraie et estimée), et on a vu qu'une variance d'estimation faible indique une bonne précision.

Prenons au hasard la valeur d'un des blocs estimée par krigeage ordinaire avec VEM, par exemple la valeur 75.54 µg/m^3 et son écart-type de l'erreur d'estimation : 3.25. Comme cet écart est faible, la valeur vraie a de grandes chances d'être par exemple inférieure à 80 µg/m^3. Prenons en revanche une valeur estimée de 71.9 µg/m^3 avec un écart-type de 9.2. Comme cet écart-type est élevé, la valeur estimée, dans ce cas, ne représente qu'une moyenne de tendances et la valeur vraie peut être inférieure ou supérieure à 80 µg/m^3. Le fait que la valeur estimée soit inférieure à 80 µg/m^3 ne signifie donc pas que la vraie valeur le soit aussi.

Cette distinction est essentielle si on cherche à estimer pour la concentration d'un polluant (de l'ozone, par exemple) le dépassement d'un seuil réglementé, évènement représenté par l'indicatrice: $I_{Z(x)>seuil}$.

Par conséquent, faire une estimation de la concentration ne répond pas au problème de l'estimation de fonctions ***f[Z(x)]*** de ***Z(x)*** comme les indicatrices. On doit alors recourir aux méthodes « non linéaires », comme l'espérance conditionnelle, le krigeage disjonctif ou les simulations conditionnelles.

Les indicatrices

L'intérêt des indicatrices provient de ce que toute fonction ***f[Z(x)]*** peut s'exprimer à l'aide des indicatrices de ***Z(x)***. Considérons en effet une fonction ***f[Z(x)]*** prenant les valeurs f_0, f_1, f_2, f_3 quand ***Z(x)*** vaut ***0, 1, 2, 3***. Une telle fonction peut s'écrire :

$$\sum_i (f_i \, I_{Z(x)=i}) = f_0 \, I_{Z(x)=0} + f_1 \, I_{Z(x)=1} + \ldots$$

Une telle somme est bien égale à f_0 si ***Z(x)=0***, à f_1 si ***Z(x) = 1***, etc…

En géostatistique on utilise souvent des indicatrices cumulées au-dessus d'un seuil, par exemple : $I_{Z(x)>2}$ fonction de ***Z(x)*** égale à ***1*** si ***Z(x)*** est ***>= 2*** (donc égale à 2 ou à 3), à 0 sinon. On a :

$$I_{Z(x)>=2} = I_{Z(x)=2} + I_{Z(x)=3}$$

L'Espérance Conditionnelle

Le meilleur estimateur de l'indicatrice ***I***(Z(x) >= seuil) conditionnée aux données Z(x_i) est l'espérance conditionnelle :

$$E[I_{Z(x)\geq seuil}] = P[Z(x) \geq s \,/\, Z(x_i) = z_i, i = 1,\ldots, N]$$

Cette espérance conditionnelle n'est en pratique calculable que dans le cas d'une fonction aléatoire multigaussienne, dans ce cas la loi (multivariable) de variables telles que Z(x), Z(x_1), …, Z(x_N) est multigaussienne (i.e. toute combinaison linéaire de ces variables est encore gaussienne).

La conséquence est que la loi de Z(x) conditionnellement aux données Z(x_i), est une loi gaussienne ayant pour moyenne le krigeage à moyenne connue de Z(x) (la moyenne d'une variable gaussienne réduite est zéro) et pour variance, la variance de krigeage simple de Z(x):

$$[Z(x)/\, Z(x_i) = z_i, \quad i = 1,\ldots,N] \quad = Loi \quad gaussienne \quad [Z(x)^{KS}, \sigma_{KS}]$$

$$U = \frac{[Z(x)/\, Z(x_i) = z_i, \quad i = 1,\ldots,N] - Z(x)^{KS}}{\sigma_{KS}} \quad : Variable \quad gaussienne \quad réduite \quad correspondant$$

$$G(x) = P[U \leq x] = \frac{1}{\sqrt{2*\pi}} \int_{-\infty}^{x} e^{-\frac{U^2}{2}} * dU \qquad : Fonction \quad de \quad répartition \quad d'une \quad gaussienne \quad réduite$$

Si on remplace la gaussienne réduite dans la définition de l'espérance conditionnelle, on a :

$$[Z(x)/\ Z(x_i) = z_i, i = 1,...,N] = U * \sigma_{KS} + Z(x)^{KS}$$

$$P[Z(x) \geq seuil\ /\ Z(x_i) = z_i, i = 1,...,N] = P[\langle Z(x)\ /\ Z(x_i) = z_i \rangle \geq seuil, i = 1,...,N]$$

$$E[\mathrm{I}_{Z(x) \geq seuil}] = P[U * \sigma_{KS} + Z(x)^{KS} \geq seuil]$$

$$E[\mathrm{I}_{Z(x) \geq seuil}] = P\left[U \geq \left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right)\right]$$

$$E[\mathrm{I}_{Z(x) \geq seuil}] = 1 - G\left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right)$$

$$E[\mathrm{I}_{Z(x) \leq seuil}] = G\left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right)$$

On obtient d'une façon très simple la probabilité de dépassement de seuil ponctuel. En résumé il suffit de calculer le krigeage simple et l'écart-type de krigeage de la concentration du polluant, puis de calculer la relation $[(seuil\text{-}Z^{KS})/\sigma^{KS}]$ et enfin de calculer directement la probabilité pour des valeurs normales [G(U)].

Pour être assuré que la variable Z(x) a une distribution gaussienne, il faut appliquer auparavant une transformation gaussienne à la variable brute. De la même façon, le seuil ponctuel à appliquer doit correspondre à la transformée gaussienne du seuil ponctuel brut. La méthode choisie pour ces transformations est ***l'anamorphose gaussienne*** qui sera expliquée plus bas.

Dans le cas des estimations de bloc, il est nécessaire de transformer la variable brute dont le support est ponctuel en une variable gaussienne de support de bloc. Il faut utiliser à cette fin un modèle de changement de support comme cela sera aussi expliqué.

Intervalles de confiance par l'espérance conditionnelle :

Pour calculer un intervalle de confiance on procède de la façon inverse car il ne s'agit plus de calculer la probabilité de dépassement à partir d'un seuil donné mais de calculer le seuil à partir d'une probabilité de dépassement donnée :

$$E[\mathrm{I}_{Z(x) \leq seuil}] = P[Z(x) \leq seuil] = G\left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right)$$

$$\left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right) = G^{-1}\{P[Z(x) \leq seuil]\}$$

$$seuil = G^{-1}\{P[Z(x) \leq seuil]\} * \sigma_{KS} + Z(x)^{KS}$$

$$E[\mathrm{I}_{Z(x) \geq seuil}] = P[Z(x) \geq seuil] = 1 - G\left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right)$$

$$\left(\frac{seuil - Z(x)^{KS}}{\sigma_{KS}}\right) = G^{-1}\{1 - P[Z(x) \geq seuil]\}$$

$$seuil = G^{-1}\{1 - P[Z(x) \geq seuil]\} * \sigma_{KS} + Z(x)^{KS}$$

On obtient ainsi de façon très simple tout percentile et en conséquence tout intervalle de confiance. La Figure 27 montre quelques intervalles de confiance qui peuvent être calculés avec une distribution gaussienne. La valeur inférieure d'un intervalle de confiance de 95% est donc le seuil dont la probabilité d'être dépassé est de 97.5% et la valeur supérieure est le seuil dont la probabilité d'être dépassé est de 2.5%:

$$seuil_{\inf} = G^{-1}\{1 - P[Z(x) \geq seuil]_{\inf}\} * \sigma_{KS} + Z(x)^{KS}$$

$$seuil_{\inf} = G^{-1}\{1 - 0.975\} * \sigma_{KS} + Z(x)^{KS}$$

$$seuil_{\inf} = -1.96 * \sigma_{KS} + Z(x)^{KS}$$

$$seuil_{\sup} = G^{-1}\{1 - P[Z(x) \geq seuil]_{\sup}\} * \sigma_{KS} + Z(x)^{KS}$$

$$seuil_{\sup} = G^{-1}\{1 - 0.025\} * \sigma_{KS} + Z(x)^{KS}$$

$$seuil_{\sup} = 1.96 * \sigma_{KS} + Z(x)^{KS}$$

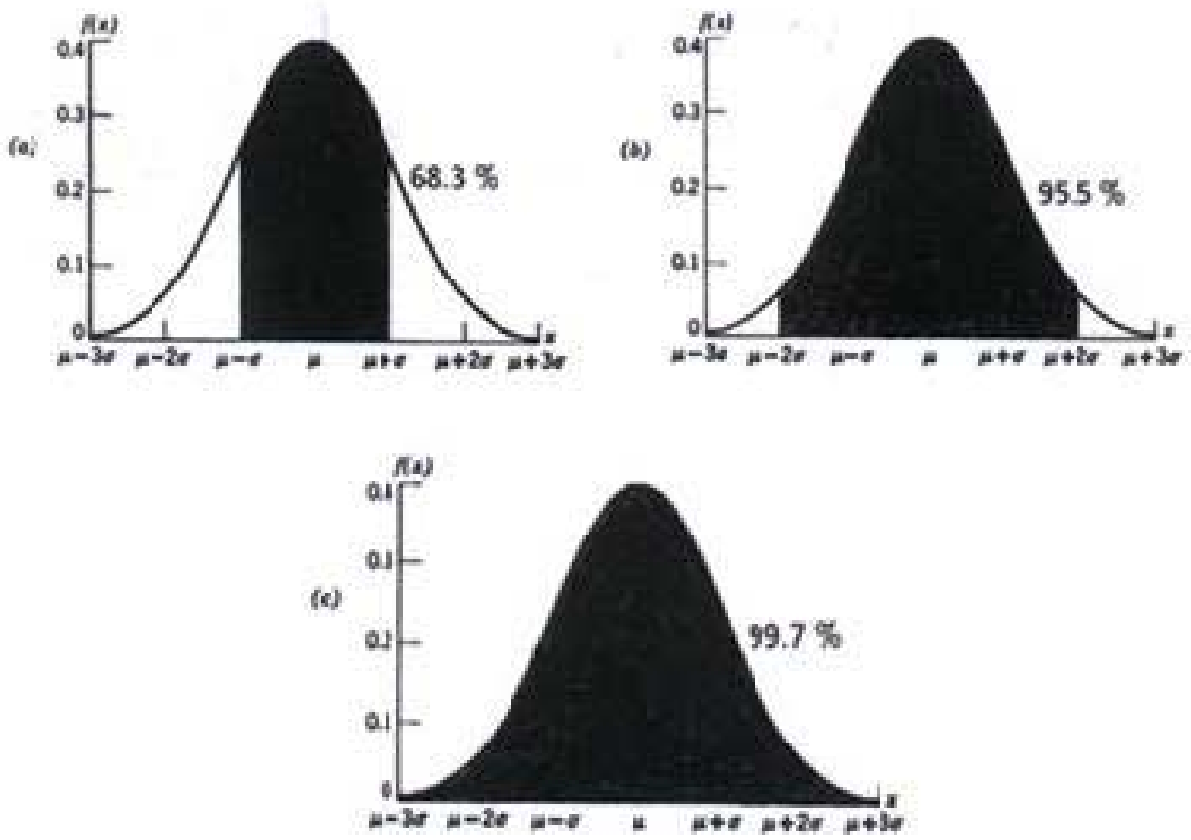


Figure 27 : Quelques Intervalles de confiance pour la distribution gaussienne

Le résultat est proche de celui de la section précédente, lorsqu'on avait calculé des intervalles de confiance pour la variable brute Z^{KO}. La différence est qu'à présent, le calcul se fait directement sur une variable gaussienne. Il n'est donc pas nécessaire de faire des hypothèses sur la distribution de l'erreur comme c'était le cas auparavant.

Il ne faut pas oublier que les seuils trouvés sont des valeurs gaussiennes. Pour les convertir en valeurs brutes on utilise l'inverse de la fonction d'anamorphose, comme indiqué plus bas.

La fonction d'anamorphose :

L'espérance conditionnelle suppose multigaussiennes les lois multivariables des $Z(x_i)$, et a fortiori gaussienne la loi monovariable de Z(x). En pratique, la variable étudiée Z(x) est rarement gaussienne, de sorte qu'une transformation est nécessaire pour se ramener à une variable gaussienne.

L'anamorphoses gaussienne est une transformation consistant à déformer l'histogramme de la variable étudiée Z(x) pour se ramener à un histogramme gaussien réduit. En supposant que sa distribution n'est pas gaussienne, on considère la fonction aléatoire Z(x) comme une fonction de la gaussienne centrée réduite Y(x):

$$Z(x) = \Phi[Y(x)]$$

Connaître la fonction d'anamorphose de Z(x) est équivalent à connaître sa distribution ; pour une variable gaussienne, l'anamorphose est linéaire.

Les polynômes d'Hermite:

La fonction d'anamorphose peut être déterminée par les coefficient de son développement en polynômes d'Hermite :

$$\Phi[Y(x)] = f_0 + f_1 * H_1[Y] + f_2 * H_2[Y] + \ldots = \sum_{n=0}^{+\infty} \{ f_n * H_n[Y] \}$$

Le nombre de coefficients calculées (*n*) indique le degré n=0,1,2,… du polynôme d'Hermite ajusté. De cette façon, si on change le degré du polynôme ajusté, le coefficient f_n de chacun de polynômes ne bougera pas. On peut ainsi stocker une centaine de coefficients, et choisir ensuite le degré 10, 20 ou 30 de l'approximation polynomiale.

Ces polynômes sont définis à partir de la densité de probabilité gaussienne réduite:

$$g(Y) = \frac{1}{\sqrt{2 * \pi}} e^{-\frac{Y^2}{2}}$$

par la formule de Rodrigues:

$$H_n(Y) = \frac{1}{\sqrt{n!} * g(Y)} \frac{d^n g(Y)}{dY^n} \qquad où \qquad \sqrt{n!} = facteur \quad de \quad normation$$

On a par exemple :

$$H_0(Y) = 1$$
$$H_1(Y) = -Y$$
$$H_2(Y) = \frac{(Y^2 - 1)}{\sqrt{2}}$$

Ils sont ensuite calculables par récurrence grâce à la relation :

$$H_{n+1}(Y) = -\frac{1}{\sqrt{n+1}} * Y * H_n(Y) - \sqrt{\frac{n}{n+1}} * H_{n-1}(Y) \qquad où \qquad n > 0$$

Il est donc facile, connaissant un valeur gaussienne Y(x), d'en déduire les polynômes d'Hermite correspondants $H_n[Y(x)]$. En pratique on se contente d'en calculer au maximum quelques dizaines. En dehors de H_0 (constant) ils vérifient, pour Y gaussienne réduite :

$$E[H_n(Y)] = 0$$
$$\text{var}[H_n(Y)] = 1$$
$$\text{cov}[H_p(Y), H_n(Y)] = 0 \qquad pour \qquad n \neq p$$

Pris au même point x, ces polynômes sont donc orthogonaux, dès lors que Y(x) est gaussienne réduite.

Pratiquement toute fonction peut se développer en polynômes d'Hermite, par exemple l'anamorphose gaussienne peut se décomposer ainsi :

$$\Phi[Y(x)] = f_0 + f_1 * H_1[Y] + f_2 * H_2[Y] + \ldots = \sum_{n=0}^{+\infty} \{f_n * H_n[Y]\}$$

Du fait de l'orthogonalité des polynômes, on a:

$$E\{\Phi[Y(x)]\ H_n[Y]\} = E\left\langle \left[\sum_{p=0}^{+\infty} \{f_p * H_p[Y]\} \right] * H_n[Y] \right\rangle$$

$$E\{\Phi[Y(x)]\ H_n[Y]\} = \sum_{p=0}^{+\infty} f_p E\langle H_p[Y] * H_n[Y] \rangle = f_n$$

Ceci permet en pratique de calculer ces coefficients, pour toute fonction donnée. Noter en particulier que :

$$f_0 = E[\Phi[Y(x)] \qquad et \qquad Var[\Phi[Y(x)] = \sum_{1}^{\infty} (f_n)^2$$

En pratique, l'anamorphoses étant connue, une telle relation permet de calculer la structure de la gaussienne à partir de la structure brute, ou inversement.

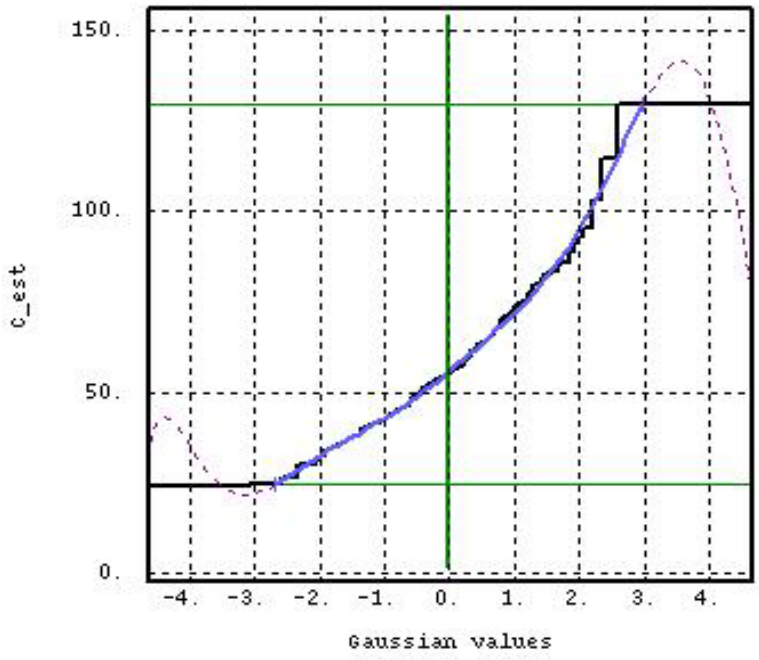


Figure 28 : Fonction d'anamorphose ponctuelle

Dans notre cas on a pris la variable brute C_est qui est la concentration de l'ozone en µg/m^3, on a calculé l'anamorphose gaussienne associée, puis on a ajusté un polynôme d 'Hermite de degré 10, ce qui veut dire qu'on a calculé 10 coefficients f_n par l'expression donnée précédemment.

La Figure 28 montre en noir la fonction de répartition de l'anamorphose calculée expérimentalement et en bleu la fonction d'anamorphose correspondante (modèle théorique) ajustée par les 10 polynôme d'Hermite.

Les valeurs de ces 10 coefficients sont données dans le tableau qui suit :

F_0	F_1	F_2	F_3	F_4
57.7	-15.3	2.9	-0.79	0.16

F_5	F_6	F_7	F_8	F_9
0.33	-0.31	0.35	-0.06	-0.39

Tableau 13 : Coefficients de polynômes d'Hermite ponctuels

Dans la Figure 28 on observe que la variable brute est comprise entre 24.4 µg/m^3 et 129.7 µg/m^3 et la variable gaussienne correspondant entre –2.7 et 2.97. Dans le Tableau 13 on démontre conformément à la théorie que le premier coefficient f_0 = 57.7 correspond à la moyenne de l'ozone.

La somme de tous les coefficients, sauf le premier, élevés au carré donne 244.01, valeur proche de la variance de l'ozone 245.36. En remplaçant les coefficients dans la formule de l'anamorphose, on obtient l'expression qui permet passer des valeurs brutes aux valeurs gaussiennes et vice-versa :

$$Z(x) = \Phi[Y(x)] = \sum_{n=0}^{+\infty} \{ f_n * H_n[Y] \}$$

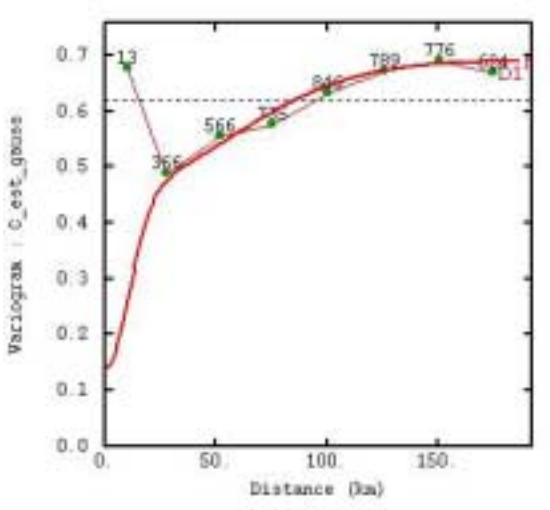
$$Z(x) = 57.7 + [(-15.3) * (-Y)] + \left[(2.9) * \frac{(Y^2 - 1)}{\sqrt{2}} \right] + \ldots$$

$$Z(x) = 55.65 + 15.3 * Y + 2.05 * Y^2 + \ldots$$

L'étape suivante est de modéliser le variogramme de la variable gaussienne. Pour cela il faut faire les mêmes hypothèses et les mêmes choix que dans la modélisation de la variable brute. En particulier on ne modélisera le variogramme expérimental qu'à partir des données rurales. Pour simplifier les choses, on a gardé les mêmes proportions entre les paliers du modèle de façon à obtenir un variogramme ayant la même allure que celui de la variable brute.

Pour le moment en effectuera le calcul du seuil de dépassement sans prendre en compte la VEM, car cette variance n'est connue que pour les mesures brutes de l'ozone et il faudrait la recalculer pour les données gaussiennes.

La Figure 29 montre le variogramme expérimental et le modèle ajusté sur les données rurales de la variable gaussienne réduite ; elle montre aussi les paramètres du modèle de la variable brute et ceux de la nouvelle variable gaussienne.



Variable	Effet de Pépite	Palier 1° structure	Portée 1° structure	Palier 2° structure	Portée 2° structure
Brute	35	81		59	
Gaussienne	0.14	0.32	26	0.23	136

Figure 29 : Modèle de variogramme de la variable gaussienne

Changement de support:

Les formules du calcul des intervalles de confiance fondées sur l'espérance conditionnelle présentées sont applicables seulement dans le cas d'estimations ponctuelles, car on transforme les mesures de l'ozone prises sur un support supposé ponctuel et donc on calcule le seuil à partir d'une probabilité de dépassement ponctuel.

En géostatistique non linéaire pour calculer le seuil pour un support supérieur, il est nécessaire de prendre en compte l'effet de support en envisageant un modèle de changement de support. Ce modèle est appelé « modèle gaussien discret ».

Ce modèle de changement de support est nécessaire pour passer de la distribution de mesures (connue) à celle inconnue de blocs. La variable Z_v peut s'exprimer comme la transformée d'une variable gaussienne réduite, notée Y_v :

$$Z(v) = \Phi[Y(v)]$$

Connaître la fonction d'anamorphose est exactement équivalent à connaître la distribution de Z_v, ce qui est le but poursuivi.

Le modèle gaussien discret repose sur l'hypothèse que le couple (Y_x, Y_v), suit une loi bigaussienne. La fonction d'anamorphose de bloc se décompose également en polynômes d'Hermite :

$$\Phi[Y(v)] = f_0 + \left\{ f_1 * r^1 * H_1[Y_v] \right\} + \left\{ f_2 * r^2 * H_2[Y_v] \right\} + \ldots = \sum_{n=0}^{+\infty} \left\{ f_n * r^n * H_n[Y_v] \right\}$$

R étant le coefficient de corrélation point - bloc entre Y_x et Y_v. Comme dans le cas de la gaussienne ponctuelle on vérifie les expressions suivants:

$$f_0 = E[\Phi[Y(x)] = E[\Phi[Y(v)]$$

$$Var[Z(x)] = Var[\Phi[Y(x)] = \sum_{1}^{\infty} (f_n)^2$$

$$Var[Z(v)] = Var[\Phi[Y(v)] = \sum_{1}^{\infty} (f_n * r^n)^2$$

Dans la pratique, le coefficient de changement de support, qui est compris entre 0 et 1, est calculé à partir de la variance de dispersion de blocs, ainsi :

$$Var[Z(v)] = Var[Z(x)] * r^2$$

$$Var[Z(v)] = Var[Z(x)] - \dddot{\gamma}(v,v)$$

$$r = \sqrt{\frac{Var[Z(v)]}{Var[Z(x)]}}$$

$$r = \sqrt{\frac{Var[Z(x)] - \dddot{\gamma}(v,v)}{Var[Z(x)]}}$$

Le variogramme moyen des blocs est calculé à partir de la méthode de discrétisation expliquée dans le paragraphe 8. Quand la somme des paliers du variogramme ponctuel diffère de la variance des données, le variogramme moyen de bloc est normalisé par la relation variance des données /somme des paliers.

Dans notre cas le coefficient de changement de support (point – bloc de 5 Km de côté) correspondant aux données d'ozone a été calculé ainsi:

$$Var[Z(x)] = 244.01$$
$$\dddot{\gamma}(v,v) = 38.1$$
$$\sum(paliers) = 175$$

$$\sum(paliers) \neq Var[Z(x)], \quad alors:$$
$$\frac{Var[Z(x)]}{\sum(paliers)} = \frac{244.01}{175} = 1.39$$
$$\dddot{\gamma}(v,v)_{normalise} = 38.1*1.39 = 52.97$$

$$r = \sqrt{\frac{Var[Z(x)] - \dddot{\gamma}(v,v)}{Var[Z(x)]}}$$
$$r = \sqrt{\frac{244.01 - 52.97}{244.01}} = 0.885$$

Une fois le coefficient de changement de support (point - bloc de 5 Km de côté) déterminé, on calcule aisément les coefficients des polynômes d'Hermite et, par suite, la transformée gaussienne des blocs.

Les coefficients des polynômes d'Hermite sont calculés en multipliant les coefficients des polynômes d'Hermite ponctuels par le coefficient de changement de support élevé à la puissance n, n étant égal au degré du polynôme.

La Figure 30 montre en noir la fonction de répartition de l'anamorphose calculée expérimentalement, en bleu la fonction d'anamorphose ponctuelle et en rouge la fonction d'anamorphose des blocs de 5 Km.

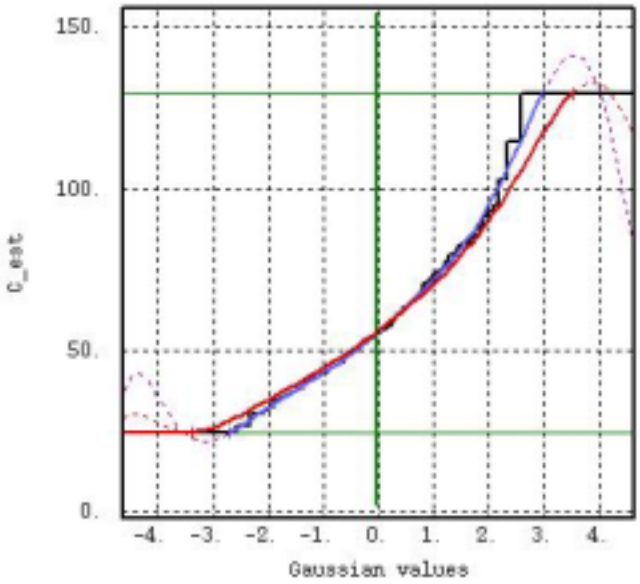


Figure 30 : Fonction d'anamorphose de bloc

Dans la Figure 30 on observe que la variable gaussienne des blocs est comprise entre –3.4 et 3.9.

Les valeurs de 10 coefficients de polynômes d'Hermite de l'anamorphose des blocs sont données dans le tableau qui suit :

F_0	F_1	F_2	F_3	F_4
57.7	-13.6	2.29	-0.56	0.1

F_5	F_6	F_7	F_8	F_9
0.18	-0.15	0.15	-0.024	-0.14

Tableau 14 : Coefficients de polynômes d'Hermite des blocs

La somme des coefficients de polynômes d'Hermite des blocs, sauf le premier, élevés au carré donne la variance des blocs de l'ozone: 191.04. En remplaçant les coefficients dans la formule de l'anamorphose on obtient l'expression qui permet passer des valeurs brutes de bloc aux valeurs gaussiennes et vice-versa :

$$Z(v) = \Phi[Y(v)] = \sum_{n=0}^{+\infty} \left\{ f_n * r^n * H_n[Y_v] \right\}$$

$$Z(v) = 57.7 + \left[(-13.6) * (-Y_v)\right] + \left[(2.29) * \frac{({Y_v}^2 - 1)}{\sqrt{2}} \right] + \ldots$$

$$Z(v) = 56.08 + 13.6 * Y_v + 1.61 * {Y_v}^2 + \ldots$$

L'espérance conditionnelle s'obtient de la même façon que dans le cas ponctuel, en remplaçant le krigeage simple de la gaussienne par celui de la gaussienne des blocs.

Calcul de la probabilité de dépassement de seuil de bloc:

$$E[1_{Z(v) \geq seuil}] = P[Z(v) \geq seuil] = 1 - G\left(\frac{seuil - Z(v)^{KS}}{\sigma_{KS}} \right)$$

$$\left(\frac{seuil - Z(v)^{KS}}{\sigma_{KS}} \right) = G^{-1}\{1 - P[Z(v) \geq seuil]\}$$

$$seuil = G^{-1}\{1 - P[Z(v) \geq seuil]\} * \sigma_{KS} + Z(v)^{KS}$$

Calcul du seuil de dépassement de bloc à partir de la probabilité:

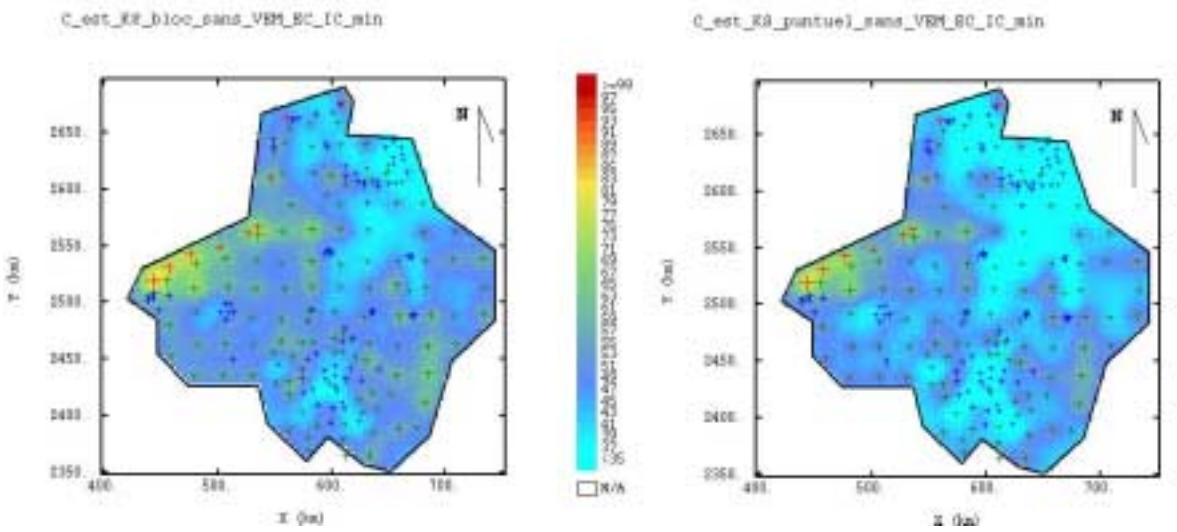
$$seuil = G^{-1}\{1 - P[Z(v) \geq seuil]\} * \sigma_{KS} + Z(v)^{KS}$$

Résultats du calcul par espérance conditionnelle:

La Figure 31 et la Figure 32 présentent les statistiques et les cartes des limites inférieures et supérieures des intervalles de confiance calculées par l'espérance conditionnelle.

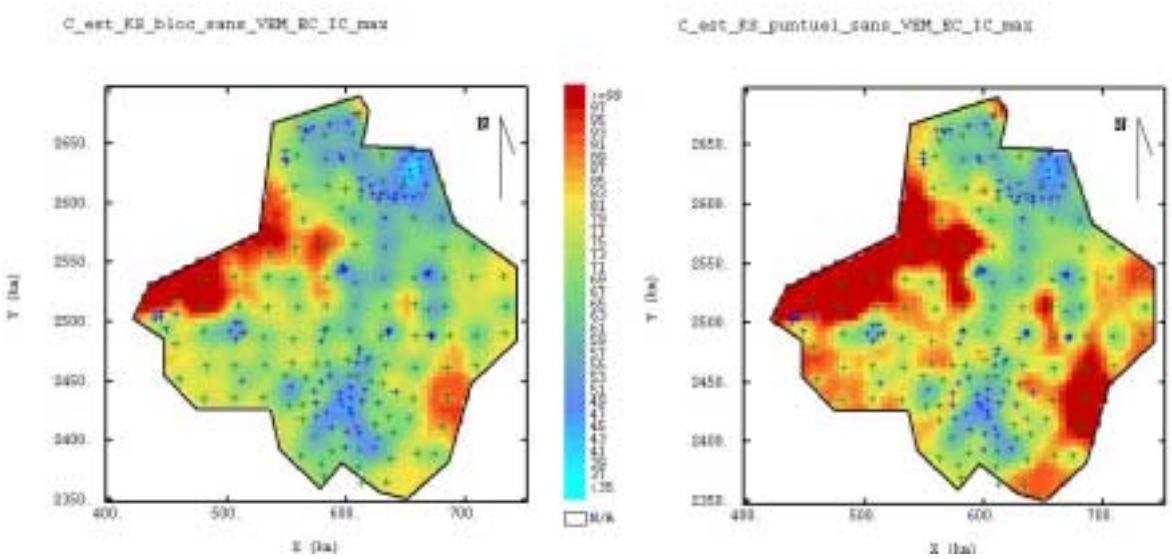
Il est intéressant de comparer ces cartes avec celles des estimations : elles présentent à peu près la même allure, avec des valeurs plus fortes sur le littoral et des valeurs plus faibles dans les zones urbaines. Dans les cartes des limites inférieure et supérieure d'intervalles, les zones de valeurs fortes et de valeurs faibles sont néanmoins plus étalées.

L'utilisation conjointe de ces cartes délivre une information plus aisée (plus immédiate) à interpréter que les cartes d'estimation et de l'écart-type d'estimation produites séparément.



Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	32,3	84,3	49,8	7,5
	Ponctuel		24,2	79,3	44,9	7,8

Figure 31 : Statistiques et cartes de la limité inférieure de l'intervalle de confiance à 95% calculée par l'espérance conditionnelle (maille de 5Km)

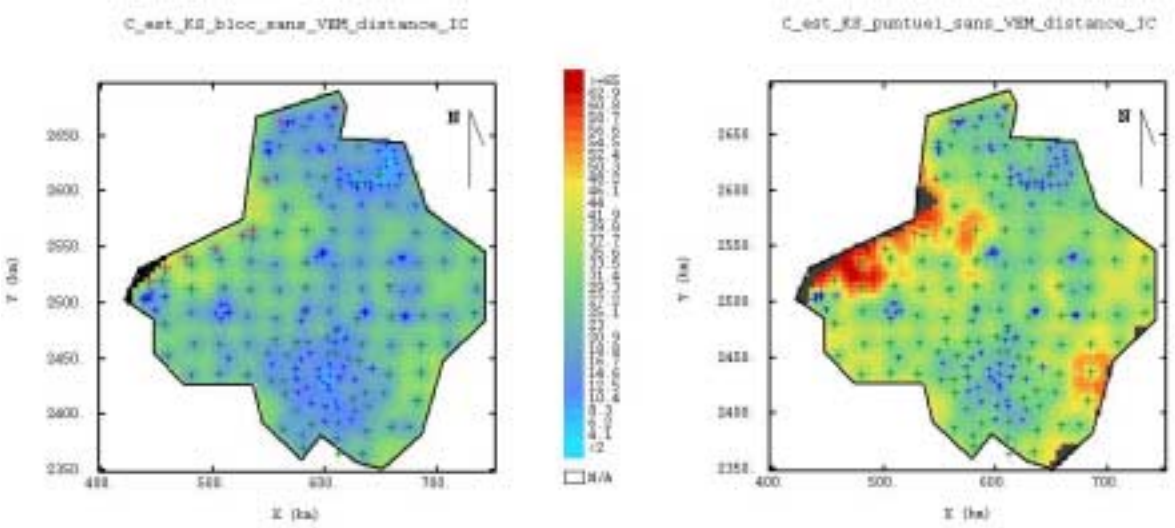


Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	39.9	121	72,7	12.7
	Ponctuel		42,7	140,7	81,3	16,3

Figure 32 : Statistiques et cartes de la limite supérieure de l'intervalle de confiance à 95% calculée par l'espérance conditionnelle (maille de 5Km)

Les cartes de la Figure 33 représentent la largeur des largeurs d'intervalles de confiance calculés par l'espérance conditionnelle. Celles-ci sont plus petites dans les zones urbaines (valeur minimale de 7.25 µg/m^3 pour l'estimation de bloc sans VEM) et plus grandes dans le littorale (valeur maximale de 73.4 µg/m^3 pour l'estimation ponctuelle sans VEM).

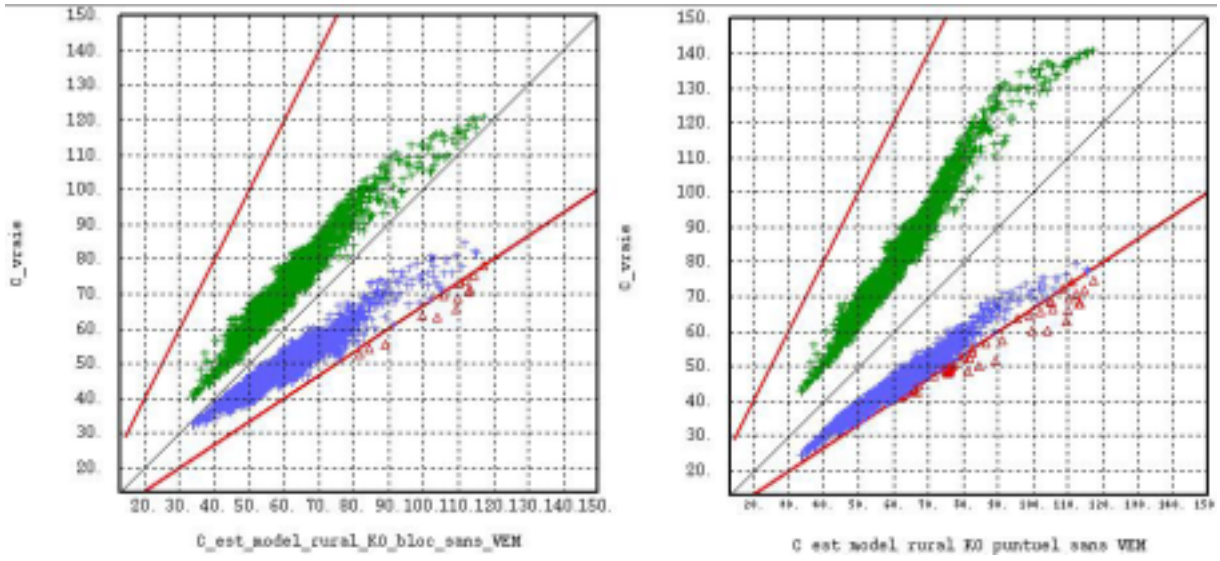
Pour les estimations de bloc la largeur moyenne est 23 µg/m^3, pour l'estimation ponctuelle sans VEM (avec effet de pépite) elle est de 36.5 µg/m^3.



En noir : zones où l'incertitude est susceptible d'excéder 50% de la valeur réelle

Support	Nb Nœuds	VEM	Minimum	Maximum	Moyenne	écart-type
Bloc	2510	Sans	7.25	50.4	23	7.17
Ponctuel			18.4	73.4	36.5	9.5

Figure 33 : Cartes de la largeur des intervalles de confiance calculés par espérance conditionnelle pour les estimations par krigeage ordinaire sur la maille de 5Km



Figures à gauche : Intervalles de confiance pour une estimation par krigeage ordinaire de bloc
Figures à droite : Intervalles de confiance pour une estimation par krigeage ordinaire ponctuelle
Croix bleues : Limite inférieure de l'intervalle de confiance à 95%. (Quantile 2.5%)
Croix vertes: Limite supérieure de l'intervalle de confiance à 95%. (Quantile 97.5%)
Lignes épaisses rouges : Bornes inférieure (resp. supérieure) et supérieure (resp. inférieure) des valeurs vraies (resp. estimées) pour avoir une incertitude maximale de 50% relativement à la valeur vraie. La borne inférieure (resp. supérieure) correspond à 0.5 fois la valeur vraie (resp. 2 fois la valeur estimée) et la borne supérieure (resp. inférieure) à 1.5 fois la valeur vraie (resp. 2/3 fois la valeur estimée).
Triangles rouges : Valeurs à l'extérieur des limites.

Figure 34 : Nuages de corrélation entre la valeur estimée (axe X) et les bornes des intervalles de confiance calculées par espérance conditionnelle (axe Y) pour les estimations par krigeage ordinaire sur la maille de 5Km

La Figure 34 montre les nuages de corrélation entre les limites d'intervalles de confiance et les estimations par krigeage ordinaire. Les intervalles de confiance ne sont pas totalement symétriques comme ceux qui ont été calculés par le critère de l'écart-type dans la section précédente. Cela est la conséquence de la transformation gaussienne. En effet, les intervalles sont symétriques dans l'espace gaussien mais au moment de transformer les seuils de la variable gaussienne pour obtenir ceux de la variable brute, la symétrie est partiellement perdue.

Comparés aux intervalles calculés par le critère de l'écart-type d'estimation, les intervalles de confiance calculés par l'espérance conditionnelle sont plus resserrés pour les petites valeurs et plus larges plus les valeurs fortes de concentration.

De ce fait, la très grande majorité des valeurs satisfont à la condition d'une incertitude maximale de 50% (voir la Figure 34). Moins de 1% des estimations de bloc sans VEM et à peine 2% des estimations ponctuelles sans VEM ne la remplissent pas nécessairement (Tableau 15). Ces pourcentages sont nettement inférieurs à ceux qui ont été calculés en géostatistique linéaire (Tableau 11 et Tableau 12).

Quelques-unes des valeurs les plus incertaines sont localisées en périphérie du domaine d'étude, à proximité du littoral nord (carrés noir dans les cartes de la Figure 33).

Nb Nœuds	Support	VEM	Nb valeurs	%
2510	Bloc	Sans	14	0.56
	Ponctuel		58	2.3

Tableau 15 : Nombre de valeurs estimées dont l'incertitude peut dépasser 50%. Intervalle de confiance calculé par espérance conditionnelle (KO, maille de 5 Km)

La Figure 35 et la Figure 36 présentent les nuages de corrélation entre les limites des intervalles de confiance calculées par l'espérance conditionnelle et les limites calculées par le critère de l'écart-type d'estimation (estimation +/- 2 fois l'écart-type).

Pour les limites inférieures, l'espérance conditionnelle donne des valeurs plus resserrées autour de la moyenne (valeurs maximales plus petites et valeurs minimales plus grandes).

Pour les limites supérieures, la différence entre les deux méthodes est moins marquée. L'espérance conditionnelle a toutefois tendance à fournir des valeurs maximales plus fortes et des valeurs minimales plus faibles.

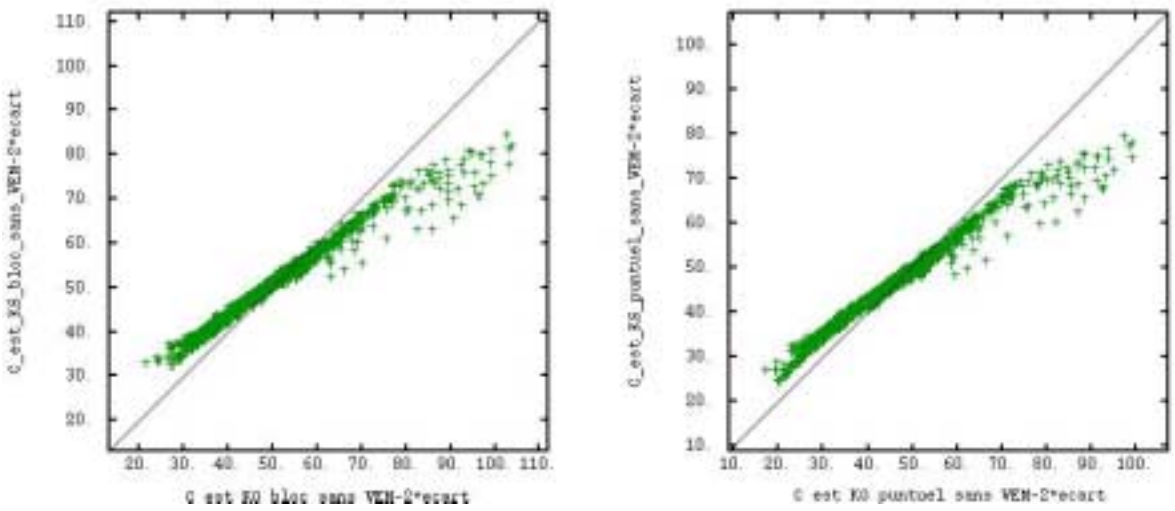


Figure 35 : Nuages de corrélation entre les limites inférieures des intervalles de confiance, calculées par l'espérance conditionnelle (axe Y), et par le critère de moins deux fois l'écart-type d'estimation (axe X), maille de 5Km

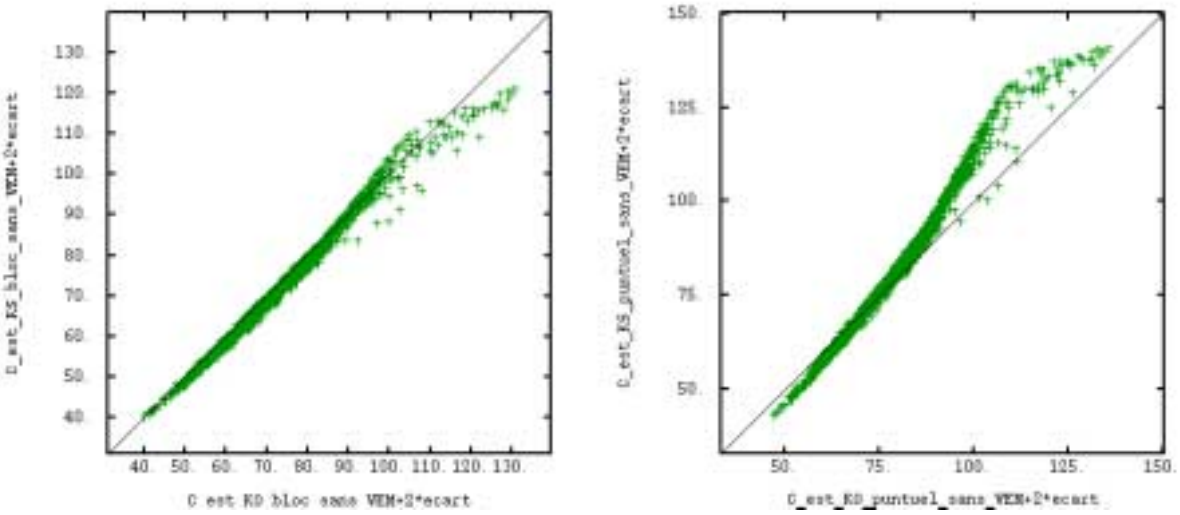


Figure 36 : Nuages de corrélation entre les limites supérieures des intervalles de confiance, calculées par l'espérance conditionnelle (axe Y), et par le critère de plus ou moins deux fois l'écart-type d'estimation (axe X), maille de 5Km

14 Intervalles de confiance : les simulations conditionnelles

Les but des simulations est de reproduire la variabilité spatiale de la variable régionalisée, dans notre cas la concentration de l'ozone. Chaque simulation est alors considérée comme une version (réalisation) possible de la réalité. De même une simulation conditionnelle est une simulation restituant aux points de mesures les valeurs qui y sont connues.

Les techniques de simulations engendrent des distributions asymptotiquement gaussiennes. Aussi, comme dans le cas de l'espérance conditionnelle, simule-t-on le plus souvent non pas directement la variable brute mais sa transformée gaussienne, avec sa propre structure, avant de repasser dans l'espace de travail initial.

Par ailleurs les fonctions aléatoires multigaussiennes, pour lesquelles toutes les combinaisons linéaires sont théoriquement gaussiennes, se prêtent bien au conditionnement. En effet, on a en tout point x :

$$Y(x) = Y(x)^{KS} + [Y(x) - Y(x)^{KS}] \qquad où : \qquad Y(x)^{KS} = Krigeage\ simple\ de\ la\ gaussienne$$

Dans le cas d'une fonction aléatoire multigaussienne, le résidu en tout point $[Y(x)-Y(x)^{KS}]$ est indépendant des valeurs aux points de données. L'idée est alors de substituer à ce résidu indépendant mais inconnu un résidu simulé ayant exactement la même structure.

Pour ce faire, on fabrique une simulation non conditionnelle de la variable, soit $Y(x)_S$, sur le domaine considéré, puis on calcule en tout point x, le résidu de son krigeage à partir des valeurs prises par Y_S aux points de données :

$$[Y(x)_S - Y(x)^{KS}{}_S]$$

La recombinaison :

$$Y(x)_{SC} = Y(x)^{KS} + [Y(x)_S - Y(x)^{KS}{}_S]$$

donne alors une autre simulation de la fonction aléatoire, mais qui est maintenant conditionnelle : en un point de donnée on retrouve bien la valeur continue :

$$Y(x_i)_{SC} = Y(x_i)^{KS} + [Y(x_i)_S - Y(x_i)^{KS}{}_S] = Y(x_i) \qquad car : \qquad Y(x_i)^{KS} = Y(x_i)$$

En résumé la méthode consiste à effectuer une simulation non conditionnelle de la variable gaussienne, puis à effectuer deux krigeages simples : le premier sur la variable gaussienne, le second sur la variable gaussienne simulée.

L'obtention de la variable gaussienne par anamorphose et du krigeage simple d'une variable gaussienne est réalisée de la même façon que pour l'espérance conditionnelle expliquée dans la section précédente.

Il existe de nombreuses méthodes et variantes pour simuler des fonctions aléatoires avec une covariance C(h). Celles-ci seront étudiées de façon plus approfondie en 2004.

Dans cette étude, la méthode retenue pour réaliser des simulations non conditionnelles [$\boldsymbol{Y(x)_S}$] de la variable gaussienne est la méthode des bandes tournantes. L'idée est de réaliser des simulations à une dimension, par des moyens s'apparentant aux séries chronologiques (par exemple les moyennes mobiles), puis de généraliser dans toutes les dimensions les résultats ainsi obtenus, par intégration sur toutes les directions de l'espace.

Résultats du calcul par simulations conditionnelles:

Les statistiques de 200 simulations obtenues sont contenues dans les tableaux qui suivent. Les simulations ponctuelles présentent une variabilité supérieure par rapport aux simulations de bloc. Cela apparaît dans la moyenne de l'écart-type des simulations qui est de 9.2 et dans le fait que les valeurs les plus extrêmes ont été obtenues par des simulations ponctuelles sans VEM (la valeur la plus petite: 16.9 µg/m^3 et la plus grande : 170.3 µg/m^3).

Les moyennes des simulations sont très proches de celles des estimations par krigeage ordinaire (comparer avec le Tableau 4 et le Tableau 7). Pour les estimations sans VEM, l'écart-type issu d'un krigeage ordinaire est légèrement inférieur (6.39 pour l'estimation de bloc et 8.88 pour l'estimation ponctuelle, voir le Tableau 5) à l'écart-type moyen des simulations (6.8 et 9.2 respectivement).

Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	22,8	83,6	45,2	8,5
	Ponctuel		16,9	80,4	40,2	8,1

Tableau 16 : Statistiques de la valeur minimale de 200 simulations conditionnelles (maille de 5 Km)

Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	39,3	157,4	83,1	19,0
	Ponctuel		43,4	170,3	92,8	21,2

Tableau 17 : Statistiques de la valeur maximale de 200 simulations conditionnelles (maille de 5 Km)

Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	33,9	113,5	61,3	11,4
	Ponctuel		33,4	112,6	61,3	11,5

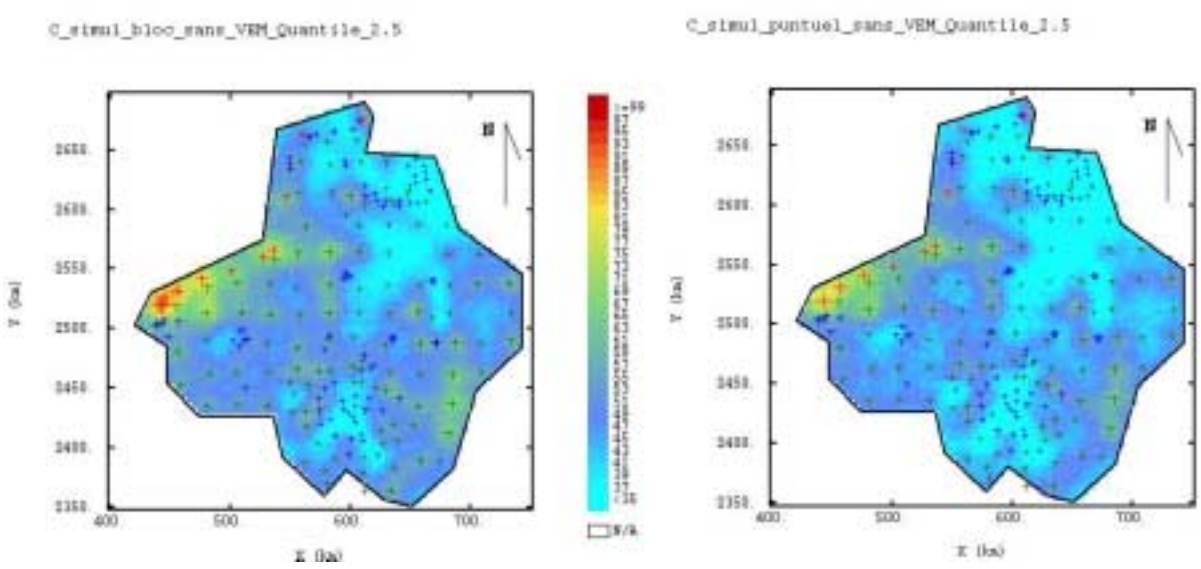
Tableau 18 : Statistiques de la valeur moyenne de 200 simulations conditionnelles (maille de 5 Km)

Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	2,1	17,2	6,8	2,3
	Ponctuel		4,1	20,4	9,2	2,6

Tableau 19 : Statistiques de l'écart-type de 200 simulations conditionnelles (maille de 5 Km)

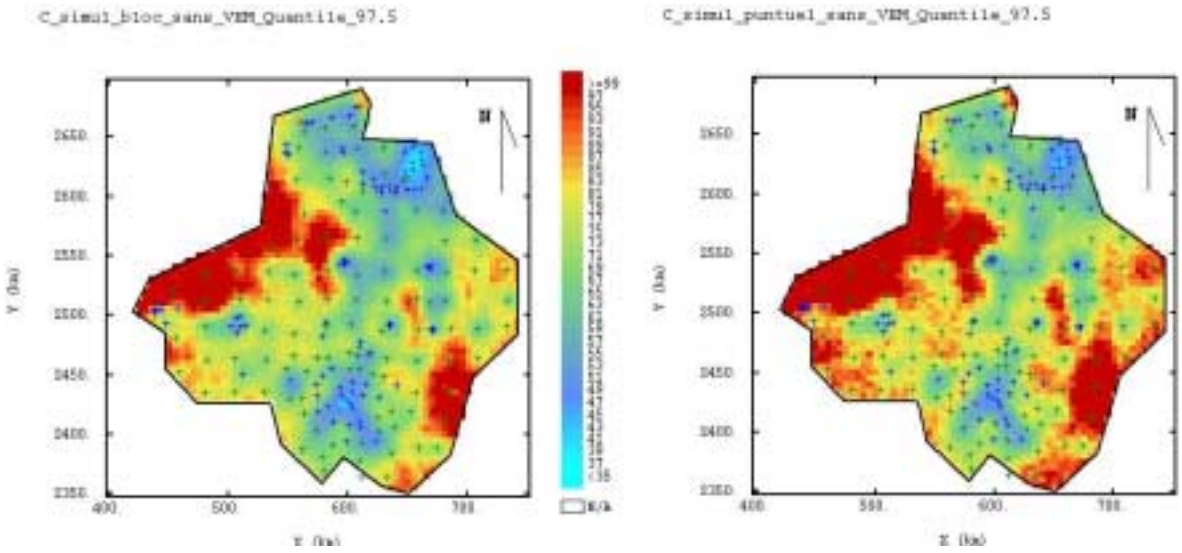
Les intervalles de confiance obtenus par simulations conditionnelles sont très proches des intervalles obtenus par espérance conditionnelle. Les cartes de la Figure 37 et de la Figure 38 confirment cette observation : elles diffèrent très peu des cartes de la Figure 31 et de la Figure 32 - les valeurs de plus forte concentration se retrouvent toujours à proximité du littoral et au sud-est du domaine, et les valeurs de plus faible concentration se situent près des zones urbaines.

En ce qui concerne les statistiques, les simulations présentent des valeurs plus extrêmes que l'espérance conditionnelle, de là des moyennes très proches mais des écart-types légèrement plus élevés.



Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	27,1	92,5	49,5	8,9
	Ponctuel		23,7	83,3	45,6	8,4

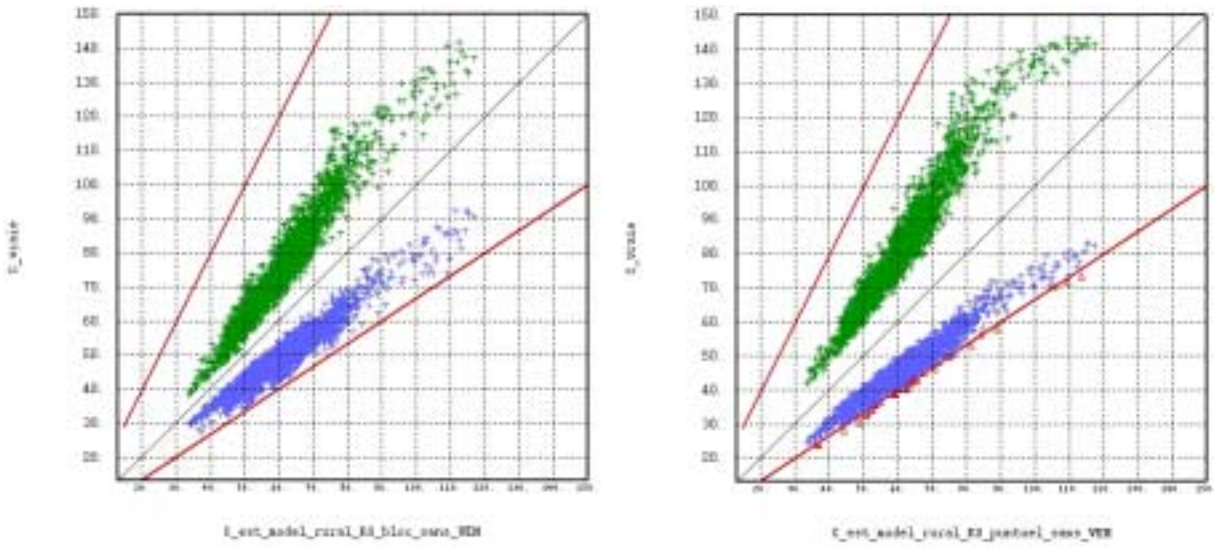
Figure 37 : Statistiques et cartes de la limite inférieure de l'intervalle de confiance à 95%, calculée par simulations conditionnelles selon la technique des bandes tournantes (maille de 5Km)



Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	38,2	141,6	76,1	16,1
	Ponctuel		41,8	143,3	82,1	17,5

Figure 38 : Statistiques et cartes de la limite supérieure de l'intervalle de confiance à 95%, calculée par simulations conditionnelles, selon la technique des bandes tournantes (maille de 5Km)

Seules 42 valeurs des intervalles de confiance calculés par simulations ponctuelles sans VEM ne remplissent pas nécessairement la condition d'incertitude. Pour la simulation de blocs, tous les intervalles de confiance respectent le critère d'incertitude de 50% de la valeur vraie (Figure 39 et Tableau 20).



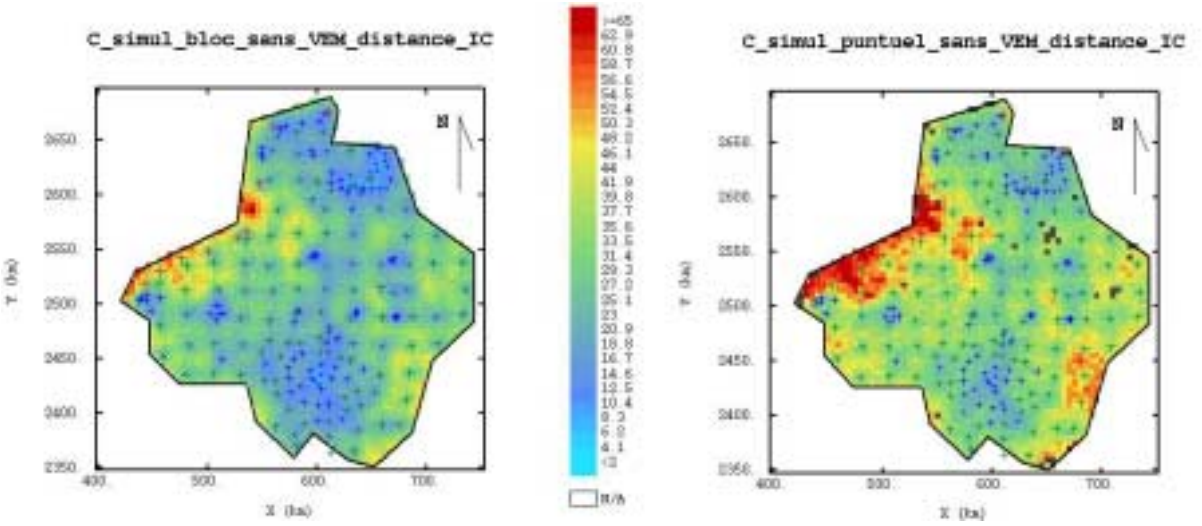
Figures à gauche : Intervalles de confiance pour une estimation par krigeage ordinaire de bloc
Figures à droite : Intervalles de confiance pour une estimation par krigeage ordinaire ponctuelle
Croix bleues : Limite inférieure de l'intervalle de confiance à 95%. (Quantile 2.5%)
Croix vertes: Limite supérieure de l'intervalle de confiance à 95%. (Quantile 97.5%)
Lignes épaisses rouges : Bornes inférieure (resp. supérieure) et supérieure (resp. inférieure) des valeurs vraies (resp. estimées) pour avoir une incertitude maximale de 50% relativement à la valeur vraie. La borne inférieure (resp. supérieure) correspond à 0.5 fois la valeur vraie (resp. 2 fois la valeur estimée) et la borne supérieure (resp. inférieure) à 1.5 fois la valeur vraie (resp. 2/3 fois la valeur estimée).
Triangles rouges : Valeurs à l'extérieur des limites.

Figure 39 : Nuages de corrélation entre la valeur estimée par krigeage ordinaire (axe X) et les bornes des intervalles de confiance calculées par simulations conditionnelles (axe Y), maille de 5Km

Nb Nœuds	Support	VEM	Nb valeurs	%
2510	Ponctuelle	Sans	42	1.7

Tableau 20 : Nombre de valeurs estimées dont l'incertitude peut dépasser 50%. Intervalle de confiance calculé par espérance conditionnelle (KO, maille de 5 Km)

Comme précédemment, les intervalles de confiance sont plus larges sur le littoral (valeur maximale de 76.5 µg/m^3 pour les simulations ponctuelles sans VEM), là où les concentrations sont les plus fortes, et plus resserrés dans les zones urbaines, là où les concentrations sont les plus faibles (valeurs minimale de 8.1 µg/m^3 pour les simulations de bloc sans VEM).



En noir : zones où l'incertitude est susceptible d'excéder 50% de la valeur réelle

Nb Nœuds	Support	VEM	Minimum	Maximum	Moyenne	écart-type
2510	Bloc	Sans	8,1	66,3	26,6	9,4
	Ponctuelle		16,7	76,5	36,4	10,4

Figure 40 : Cartes de la largeur des intervalles de confiance calculés par simulations conditionnelles (maille de 5Km)

Les nuages de corrélation entre les bornes des intervalles calculées par simulations conditionnelles et les bornes des intervalles calculées par espérance conditionnelle sont présentés dans la Figure 41 et la Figure 42.

Si les résultats des deux méthodes sont très similaires, on observe que la corrélation se détériore un peu pour les valeurs extrêmes. Cette observation confirme les résultats déjà mentionnés, à savoir que les valeurs simulées sont plus dispersées et que comparées aux valeurs calculées par l'espérance conditionnelle, elles atteignent des valeurs maximales plus fortes et des valeurs minimales plus faibles.

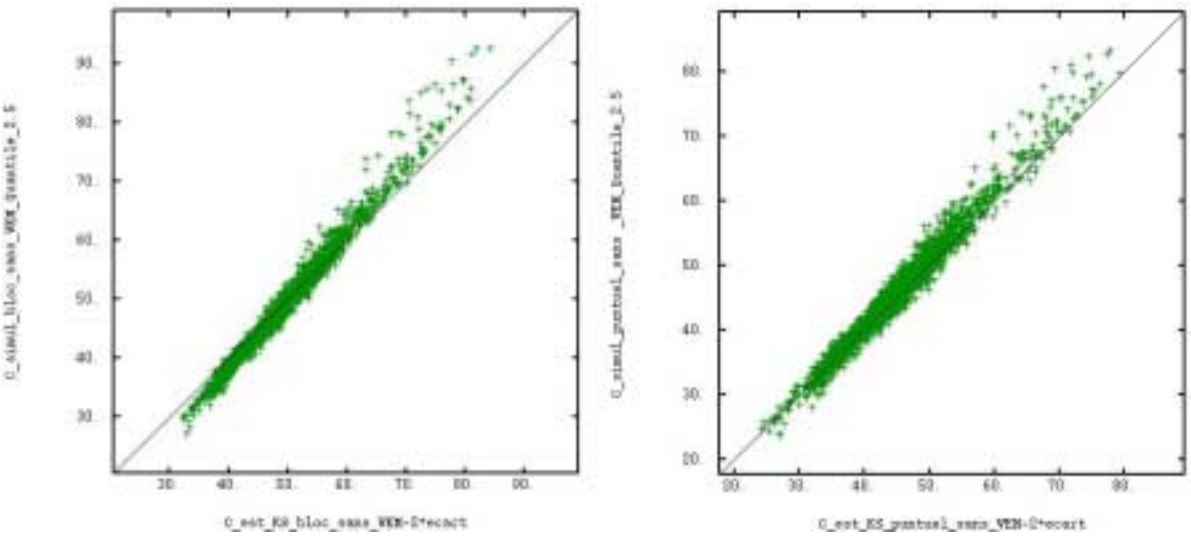


Figure 41 : Nuages de corrélation entre les limites inférieures des intervalles de confiance calculées par l'espérance conditionnelle (axe X), et par simulations conditionnelles (axe Y), maille de 5Km

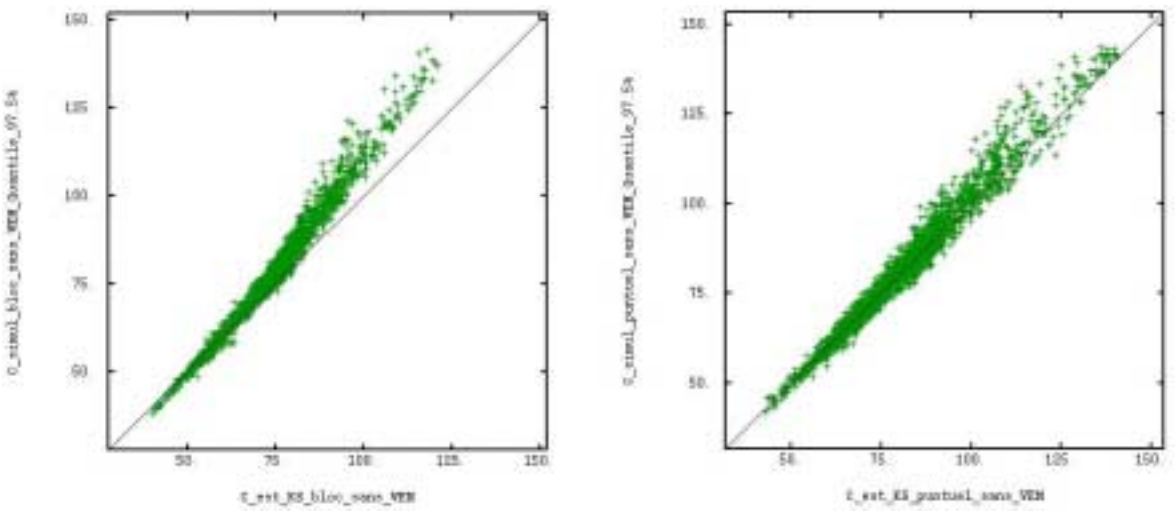


Figure 42 : Nuages de corrélation entre les limites supérieures des intervalles de confiance, calculées par l'espérance conditionnelle (axe X), et par simulations conditionnelles (axe Y), maille de 5Km