LCSQA

Laboratoire Central
de Surveillance de la Qualité de l'Air

Etude n°10 – Assistance en modélisation (Rapport 3/3)

Utilisation du module « Geostatistical Analyst » d'ARCVIEW dans le cadre de la qualité de l'air

Novembre 2004
Convention : 04000087

Giovanni CARDENAS

INERIS

MINISTÈRE
DE L'ÉCOLOGIE
ET DU DÉVELOPPEMENT
DURABLE

Utilisation du module « Geostatistical Analyst » d'ARCVIEW dans le cadre de la qualité de l'air

Laboratoire Central de Surveillance de la Qualité de l'Air

Convention n°04000087

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

Etude n° 10 : Assistance en modélisation
Rapport N° 3/3

NOVEMBRE 2004

Giovanni CARDENAS

Ce document comporte 89 pages

	Rédaction	Vérification	Approbation
NOM	Giovanni CARDENAS	Olivier SAINT-JEAN Laurence ROUIL	Martine RAMEL
Qualité	Etudiant en Stage Ingénieur Etudes et Recherches Direction des Risques Chroniques	Responsable unité 2IEN Ingénieur Etudes et Recherches Direction des Risques Chroniques	Coordination LCSQA Direction des Risques Chroniques
Visa			

TABLE DES MATIERES

LISTE DE FIGURES

LISTE DE TABLEAUX

1. INTRODUCTION

Les objectifs visés dans ce document sont les suivants :

- Expliquer et détailler les ajustements de paramètres et les méthodes de calcul effectués par le module « Geostatistical Analyst » («GA»), disponible dans le Système d'Information Géographique Arcwiew

- Voir les interactions avec les données qui sont autorisées par «GA» et détailler leur fonctionnement.

- Application concrète à un jeu des données d'ozone correspondant à une campagne menée sur la partie nord de la France (régions Ile de France, Normandie, Picardie et Nord Pas-de-Calais), la semaine prise en considération dans cette étude est comprise entre le 26 juin et le 3 juillet 2000.

  Les données sur la pollution de l'air par le NO_2 dans la région de Bourg-en-Bresse seront aussi utilisées. Elles correspondent à une campagne de mesure menée par l'association Air de l'Ain et des Pays de Savoie sur l'année 2001.

- Nous nous appuierons sur les résultats de l'étude LCSQA (Laboratoire Central pour la Surveillance de la Qualité de l'Air) n°15, menée en 2003 par Laure Malherbe et Giovanni Cardenas.

2. EXPLORATION DES DONNEES (EXPLORATORY SPATIAL DATA ANALYSIS : ESDA)

2.1. REMARQUES GENERALES SUR L'UTILISATION

Les remarques générales portent sur des point particuliers, mais sont relatives à l'ensemble du module, ou au moins à plusieurs outils différents. De cette manière les redondances entre les étapes sont limitées.

- On peut remarquer dans un premier temps que «GA» ajuste automatiquement l'échelle d'affichage des graphiques, sans que l'on puisse la modifier. Il donne alors comme légende : {nom de la variable} * 10^x, c'est l'expression de l'échelle qui s'affiche. C'est-à-dire que la valeur que l'on va lire est exactement l'expression de la légende, ce qui donne mathématiquement :

  $$j = i * 10^x$$

  si i est la valeur lue sur les axes et j la valeur non modifiée de la variable.

- Une sélection multiple se fait simplement en pressant les touche maj ou ctrl et en cliquant sur les classes désirées. La sélection est conservée quant on passe d'une variable à une autre et également lorsqu'on sort de la fenêtre de dialogue.

- Lorsque plusieurs points sont superposés, mais présentent des valeurs différentes,

  c'est à dire lorsque l'on a des valeurs doublon, une fenêtre de dialogue apparaît. ««GA»» propose à l'utilisateur plusieurs méthodes pour intégrer les données : moyenne, maximum ou minimum des valeurs, retirer les données doubles ou les intégrer.

- On remarque que les fenêtres d'exploration des données sont indépendantes les une des autres. On peut donc naviguer de l'une à l'autre et aussi travailler sur ArcMap lorsqu'elles sont ouvertes.

Dans cette partie, les outils qui sont proposés dans la commande « Explore Data » sont examinés. Nous allons expliquer le fonctionnement et voir l'application au cas concret.

2.2. HISTOGRAMME

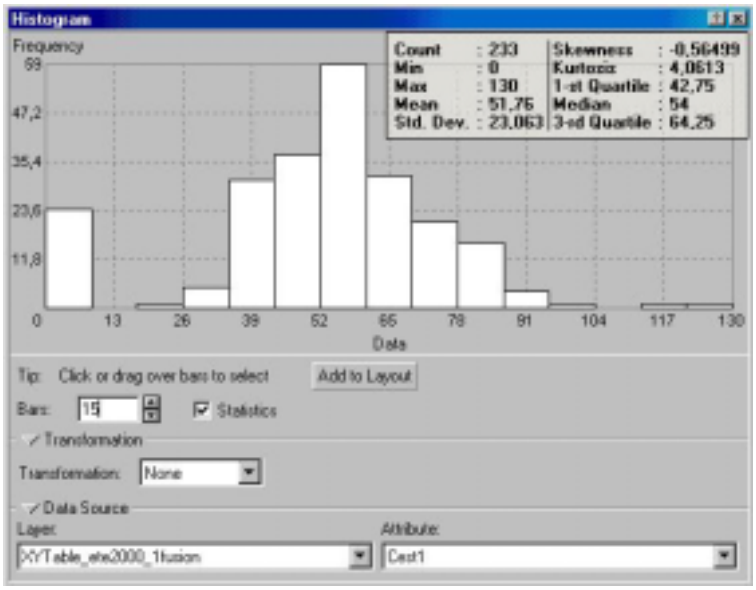


Figure 1 : Fenêtre de dialogue permettant de visualiser la répartition de la variable

Ce premier outil affiche l'histogramme de la variable sélectionnée au moyen des listes déroulantes « Layer » et « Attribute », dans la rubrique « Data Source » :

- « Layer » permet de sélectionner la couche où se trouve la variable

- « Attribute » répertorie les noms de champs de la table attributaire de la couche sélectionnée, cette liste permet de sélectionner la variable que l'on souhaite visualiser.

Cet outil permet donc de visualiser la distribution de la variable. Dans ««GA»», on cherche à obtenir une variable qui ressemble à une loi normale. Si on observe une asymétrie dans la distribution de cette variable, alors le module nous propose un jeu de transformations à appliquer à celle-ci.

Cependant, l'hypothèse de normalité n'est pas requise pour effectuer une carte d'estimation par krigeage.

La rubrique « Transformation » permet d'appliquer une transformation à la variable pour adapter son histogramme à celui d'une loi normale. Les transformations disponibles sont : Box-Cox, Log et Arcsin.

- la transformation Box-Cox se définit mathématiquement par :

  $Y(s) = \frac{Z(s)^{\lambda} - 1}{\lambda}$, où $\lambda \neq 0$ est un paramètre fixé par l'utilisateur.

  Cette transformation permet de déplacer la moyenne de la distribution et de disperser ou de contracter cette distribution en jouant sur la valeur du paramètre. Le cas particulier λ=½ permet à une variable de comptage d'avoir une variance qui dépende moins du nombre d'évènements comptés. Ce cas particulier permet aussi de faire approcher la distribution d'une distribution normale.

- la transformation logarithmique :

  $Y(s) = \ln(Z(s))$, pour Z(s) > 0

  Cette transformation s'utilise en général lorsque la variable présente une queue dans les valeurs forte et que le pic de la distribution est décalé vers la gauche. En effet, d'après l'allure de la courbe du logarithme népérien, les fortes valeurs sont plus fortement atténuées que les valeurs faibles.

- la transformation Arcsinus :

  $Y(s) = \sin^{-1}(Z(s))$

  dans le sens où $\sin^{-1}(.)$ est la fonction réciproque de sin(.) elle est définie pour $Z(s) \in [0;1]$ et renvoie une valeur sur l'intervalle [0 ; π/2]. Sa forme permet une transformation qui conserve les valeurs basses et moyennes et qui amplifie les valeurs fortes de l'intervalle de définition.

  Cette transformation s'utilise donc si la variable est une proportion ou une probabilité (on a $Z(s) \in [0;1]$) pour laquelle on souhaite « gonfler » les valeurs fortes.

La case à cocher « Statistics » permet d'afficher ou non les statistiques de la variable. « Bars » permet, quant à lui, d'adapter les classes de l'histogramme. On peut remarquer que les valeurs de la grille sont calculées sur les valeurs des classes par défaut (10 classes), mais ne sont pas actualisées avec le changement du nombre de classes. Ce problème constitue un obstacle à une lecture directe et rapide des informations.

La répartition de classe utilisée se fait par classes d'amplitude égales, ensuite « GA » affecte une fréquence à chaque classe.

Le bouton « Add to Layout » permet de récupérer l'histogramme et de l'afficher dans la mise en page du document.

Dans notre cas, la variable considérée est la concentration estimée en ozone, à chaque point de mesure. Elle est désignée par « Cest », nous voyons que son histogramme est presque gaussien mais présente une forte classe correspondant aux valeurs [0 ; 8,67]. Il est à noter que lors de l'intégration des données de la table excel, les valeurs manquantes marquées par « N/A » ont été remplacées par 0. Par conséquent on sait que cette classe est sur-représentée. Cette partie de l'explorateur de données montre clairement que le remplacement des valeurs a faussé la distribution des données. Il faudra donc en tenir compte et retirer les points qui ont été corrigés pour ne pas fausser le variogramme.

Il est possible de faire une sélection par classe depuis l'histogramme en cliquant simplement sur la classe ou le groupe de classes visé. Il est ainsi possible, par exemple, d'identifier seulement les entités qui présentent des valeurs extrêmes.

Si on souhaite voir l'histogramme d'un type de données par rapport à l'histogramme de l'ensemble, on doit faire une sélection par attributs dans ArcMap et ensuite celle-ci est visible dans la fenêtre de l'outil.

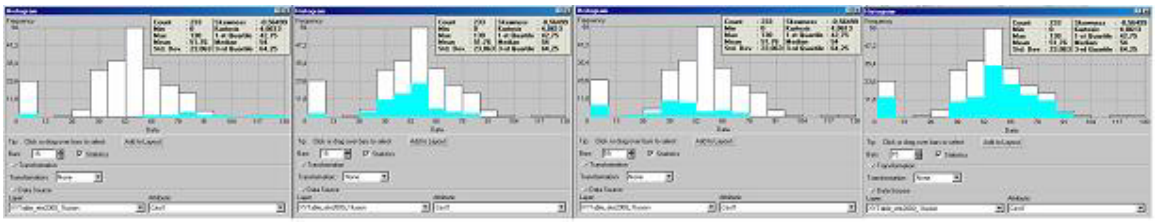


Figure 2 : Sélection des données littorales, puis périurbaines, puis urbaines, et enfin rurales

2.3. NORMAL QQPLOT

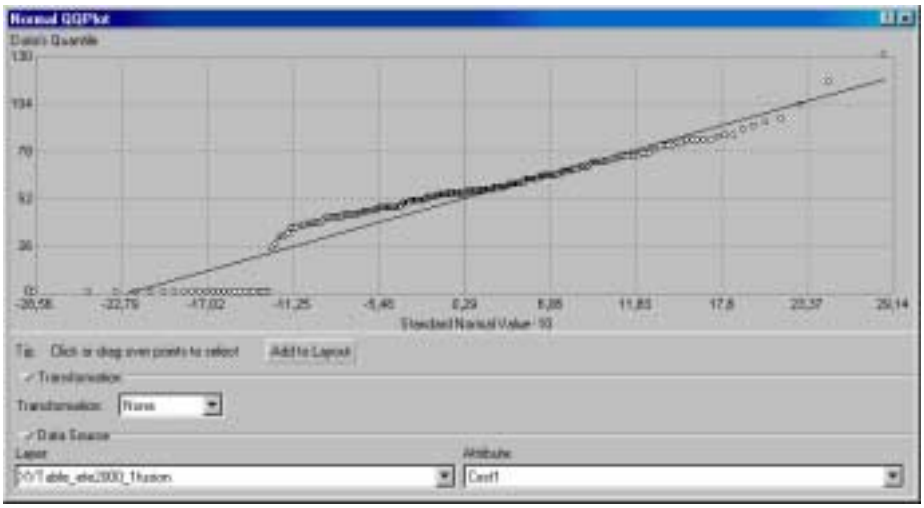


Figure 3 : Fenêtre de dialogue de Normal QQPlot

Normal QQPlot met en relation les données de la variable en ordonnées, avec les valeurs de la loi normale, en abscisses. L'objectif est là aussi de voir si la distribution de la variable se rapproche d'une distribution normale centrée réduite.

Pour construire ce graphe, «GA» calcule en premier lieu la valeur du quantile correspondant à chacune des données. Puis, il calcule en sens inverse la valeur que l'on aurait dû rentrer dans une loi gaussienne centrée réduite pour obtenir ce même quantile. De cette manière nous avons une correspondance entre la loi de distribution du jeu de données et la loi normale centrée réduite. Il suffit ensuite de mettre en relation les valeurs de chacune des distributions correspondant à un même quantile.

En effet si les points s'alignent le long de la droite y = x, alors les valeurs correspondant aux mêmes quantiles sont identiques pour les deux distributions, et on peut assimiler la distribution de la variable à une loi normale.

Le graphique place la droite d'équation y = x et les points avec les coordonnées :

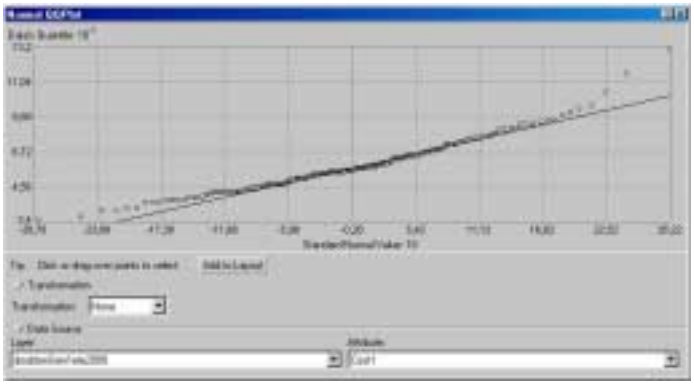
$x = x_{in}$

$y = x_{id}$

Où x_{in} et x_{id} qui sont respectivement les valeurs de la loi normale et des données correspondant à des quantiles équivalents.

Les rubriques « Data Source » et « Transformation » ont la même utilité et le même fonctionnement que dans la fenêtre Histogramme, nous ne re-détaillerons donc pas leur utilisation. Il en va de même pour le bouton « Add to Layout ».

Dans notre cas, la majorité des points se rapprochent de la droite, c'est-à-dire que les valeurs des données tendent à être les mêmes que ceux d'une loi normale pour les mêmes quantiles. Ce graphique conforte l'observation de l'histogramme (figure ci-dessus). On voit en effet que les premières valeurs des données sont nulles, forçant les points à ne pas suivre la droite d'équivalence. Ces points qui n'ont pas de valeur légitime influencent la distribution des données, il faudra donc les supprimer.



Après suppression des valeurs nulles, les données extrêmes s'éloignent d'une loi normale. Mais globalement la loi de distribution des données peut être considérée comme étant normale. En effectuant une sélection des trois valeurs les plus fortes, qui s'éloignent d'une distribution gaussienne, on voit que ce sont des valeurs littorales.

Une sélection attributaire selon le critère : Cest ≠ 0 permet de supprimer les valeurs nulles. Il suffit ensuite de créer une nouvelle couche à partir de la sélection nommée « doublonSem1ete2000 ».

Jusqu'à présent, les deux études de données successives ont permis d'identifier les données extrêmes et d'avoir une idée sur leur répartition. Mais aussi de remarquer le « vice de procédure » lorsque nous avons affecté arbitrairement la valeur nulle aux valeurs inconnues, et de supprimer ces valeurs.

2.4. TREND ANALYSIS

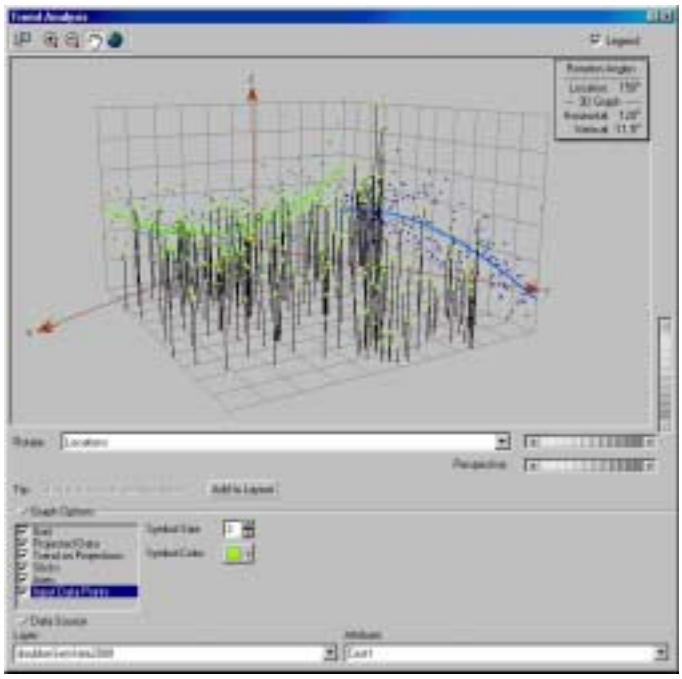


Figure 4 : Fenêtre de dialogue pour visualiser la dérive des données

Une dérive (trend ou drift) affecte tout le jeu de données, en effet, on peut essayer de la décrire comme étant un processus physique qui influence la variable. Elle est considérée alors comme la tendance moyenne de la variable. La variable aléatoire se décompose selon la formule :

$$Z(s) = \mu(s) + R(s)$$

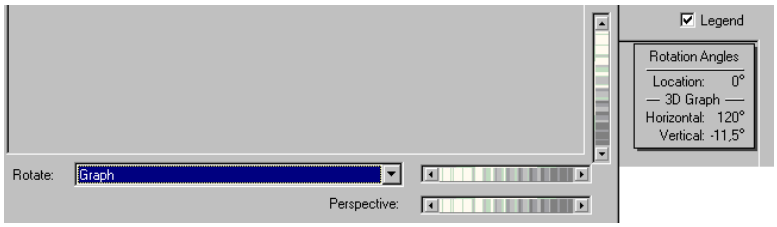
Où μ(s) qui est une fonction déterministe qui représente la dérive (tendance de la variable) et R(s) représente le résidu.

Par exemple, on remarque que la concentration en NO_2 est plus forte dans les zones de forte densité de population que dans les zones rurales. La densité de population est une dérive possible lors de l'étude de la concentration en NO_2.

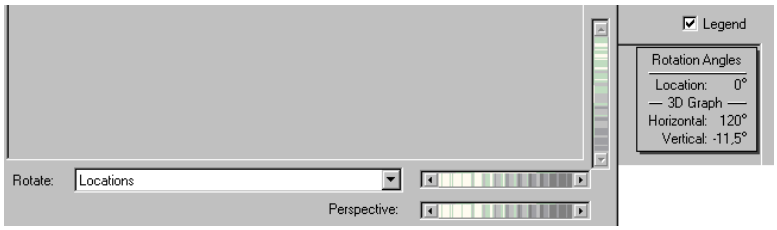
La dérive peut être modélisée mathématiquement sous la forme d'un polynôme. «GA» propose de modéliser la dérive par un polynôme de degré inférieur ou égal à 3, il est cependant conseillé de garder le degré de la dérive aussi bas que possible pour simplifier les équations de calcul. «GA» doit calculer les coefficients du polynôme, plus le degré est bas et moins de coefficients seront calculés par le module.

2.4.1. CE QUE FAIT L'OUTIL :

Pour identifier une dérive, l'outil élève les point de la carte suivant la direction « Z » et selon leurs valeurs pour la variable choisie. Cette dernière est spécifiée comme pour les autres outils dans la liste « Attribute ». Ensuite les points élevés sont projetés sur les plans (X,Z) et (Y,Z) qui sont perpendiculaires au plan de la carte. Ces plans sont par défaut orientés selon les directions Nord et Sud. Cependant il est possible de corriger cette projection pour l'adapter à son propre cas de figure. En effet on peut orienter les plans de projection en sélectionnant « Graph » dans le champ « Rotate » et en cliquant sur les molettes. On voit les changements dans la fenêtre et ils sont répertoriés dans l'onglet « 3D Graph » de la légende.



Pour changer l'angle de projection de la carte sur les axes il faut maintenant choisir « Locations » dans le champ « Rotate ». On fait maintenant tourner la carte par rapport aux axes de projections. Les changements sont là aussi visibles dans la fenêtre de l'outil, mais aussi dans l'onglet « Rotation Angles » de la légende.



Ces modifications permettent de trouver la vue la plus adaptée pour identifier la direction et l'ordre de la dérive. En effet, l'outil adapte automatiquement une courbe polynomiale aux données projetées. Une courbe prononcée indiquera la présence d'une dérive dans la variable. Cependant il est difficile de réellement identifier une distribution avec dérive par rapport à une autre distribution sans dérive. Il est donc conseillé de ne supprimer la dérive que lorsqu'on a identifié le phénomène physique qui en est à l'origine. En effet ajouter une dérive implique le calcul de nouveaux paramètres, ce qui nuit à la précision du modèle : plus il y a de paramètres à estimer, moins le modèle est précis.

2.4.2. LES FONCTIONS DE L'OUTIL :

Cet outil pour l'analyse des données fournit différentes fonctions disponibles dans la rubrique « Graph Options » de la fenêtre de dialogue.

- En premier lieu, il est possible de choisir les éléments que l'on souhaite afficher dans la fenêtre en cochant ou décochant les cases correspondantes. Il est aussi possible, pour chacun des éléments, de choisir la taille d'affichage.

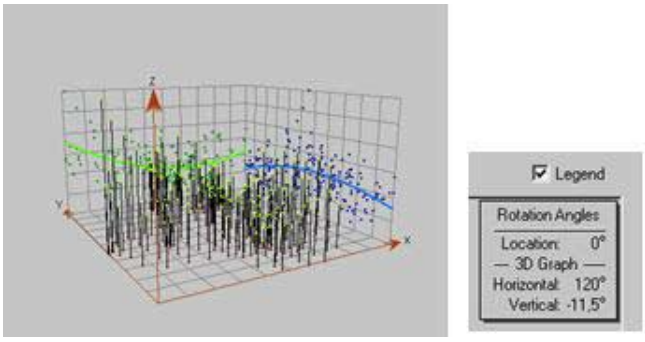
- Pour la grille, il est possible de choisir le resserrement de la maille suivant chacune des directions.

- Pour la projection des données sur les trois plans, les bâtons, les axes et les données, on peut choisir la couleur et la taille de représentation.

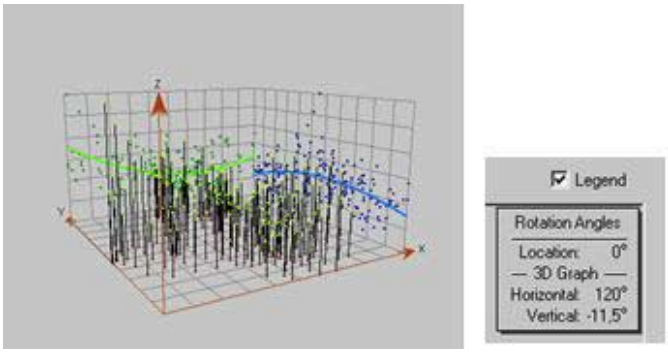
- Pour les courbes des points projetés, on peut choisir l'ordre de la courbe, ce qui permet d'ajuster l'ordre de la dérive. Ensuite, on peut choisir les options de représentation.

Ces options sont donc principalement liées à la lisibilité du graphe. Mais pour ce qui concerne l'ordre de la courbe d'ajustement, l'utilisateur peut manipuler cette option pour voir de quel ordre est la dérive des données. Le manuel d'utilisation de «GA» ne donne pas d'indication quant à la méthode d'ajustement des coefficients pour la courbe représentant la dérive. Il est probable que la méthode utilisée soit la méthode des moindres carrés, mais aucun indice ne permet de le vérifier.

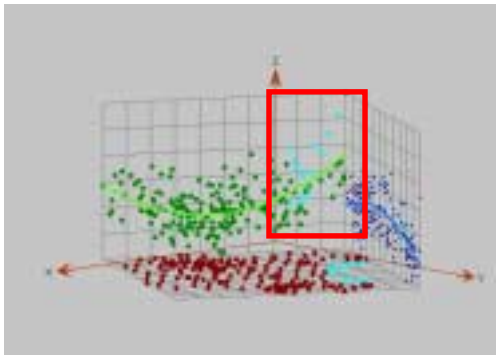
2.4.3. CAS CONCRET



Dans notre cas sur l'étude de la pollution en ozone, l'outil utilisé directement donne le résultat ci dessus. On peut donc déjà s'attendre à une dérive.



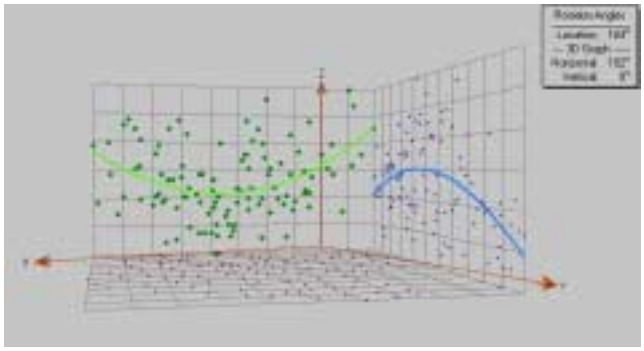
Après ajustement de la direction, on voit clairement une dérive sur la courbe verte, selon la direction 155°.



Après un examen plus minutieux, on s'aperçoit que les hautes valeurs littorales tirent la courbe verte vers le haut. Nous avons donc identifié le phénomène correspondant à la dérive.

Divisons maintenant les données en trois groupes, comme décrit dans l'étude LCSQA n°15, données urbaines et périurbaines pour la variabilité à petite distances, données rurales pour la pollutions de fond (à grandes distances) et données littorales.

Si on s'attache à la représentation des données rurales, on crée une nouvelle couche ne contenant que ce type de données, et on obtient un graphe ne traitant que ce type de données.



On voit que l'on a toujours une courbe qui s'ajuste aux projections. Mais n'ayant pas de phénomène physique pour interpréter cette courbe, on ne considère pas de dérive sur les données rurales.

On peut remarquer par ailleurs, qu'il n'y a pas d'échelle sur le graphique 3D qui est tracé, ceci amoindrit la lisibilité au graphique.

2.5. VORONOI MAP

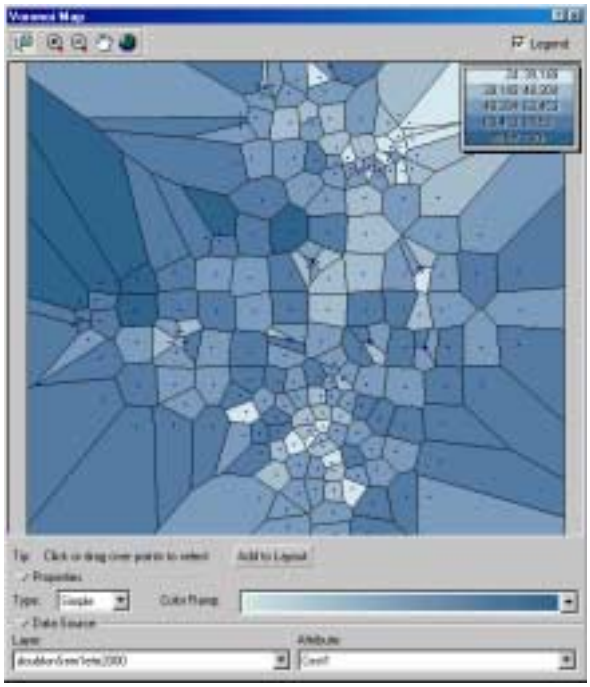


Figure 5 : Fenêtre de dialogue de l'outil Voronoi Map

La carte de Voronoi représente les polygones formés en considérant les points les plus proches du point origine du polygone (i.e. le capteur)que d'un autre point origine. Ainsi la zone -le polygone- entourant chacun des points origine -les capteurs- regroupe les points les plus proches de ceux-ci. Les polygones voisins sont ceux qui partagent un côté de cette zone.

2.5.1. DESCRIPTION DES METHODES DE REPRESENTATION

Différentes méthodes statistiques pour calculer les valeurs prises par les polygones peuvent remplir différents objectifs :

- lissage local :
  - la moyenne (mean) : la valeur que prend le polygone est la moyenne du point et de ses voisins
  - le mode : les polygone sont répartis en cinq classes, la valeur que prend une cellule est la valeur de la classe la plus fréquente entre le polygone et ses voisins (le mode)

- la médiane (median) : la valeur du polygone est la médiane de la distribution de fréquence de la cellule et de ses voisins

- variation locale :
  - l'écart type (StDev -standard deviation-): la valeur de la cellule est l'écart type entre la cellule et ses voisins
  - l'écart interquartiles (IQR -interquartiles range-): la valeur de la cellule est la différence entre le premier et le troisième quartile de la distribution du polygone et de ses voisins
  - l'entropie (entropy): les polygones sont répartis en cinq classes selon la méthode des écarts naturels. C'est-à-dire que les classes sont délimitées par les sauts observés dans les valeurs des données, c'est un compromis pour obtenir des classes qui ne soient pas trop grandes aux extrémités et mettre en valeur les changements dans les données. L'entropie est calculée selon la formule :

    $$E = -\sum p_i * \log_2(p_i)$$

    où p_i est la proportion de polygones assignés à chaque classe, on démontre que :

    $E_{\min} = 0$, quand tous les polygones sont dans la même classe

    et $E_{\max} = -\sum\left(\frac{1}{n_i} * \log_2\left(\frac{1}{n_i}\right)\right)$

    où n_i est le nombre des effectifs

- valeurs locales extrêmes :
  - cluster : les polygones sont répartis en cinq classes, lorsque qu'un polygone n'appartient pas aux mêmes classes que ses voisins, alors il est coloré en gris pour le différencier des ses voisins

- influence locale :
  - simple : la valeur du polygone est simplement la valeur de la donnée en son centre

2.5.2. APPLICATION AU CAS DE LA POLLUTION PAR L'OZONE

On s'intéresse à la carte de tous les types de sites. Parmi toutes les méthodes proposées, on peut en regrouper certaines dont les cartes sont très similaires, plusieurs méthodes pouvant se rapporter au même objectif.

- Cluster : on remarque sur la carte correspondante que les valeurs extrêmes sont généralement situées au niveau des espaces urbanisés. On voit aussi que les valeurs fortes sont les valeurs littorales et qu'elles correspondent aux points que l'on a déjà identifiés.

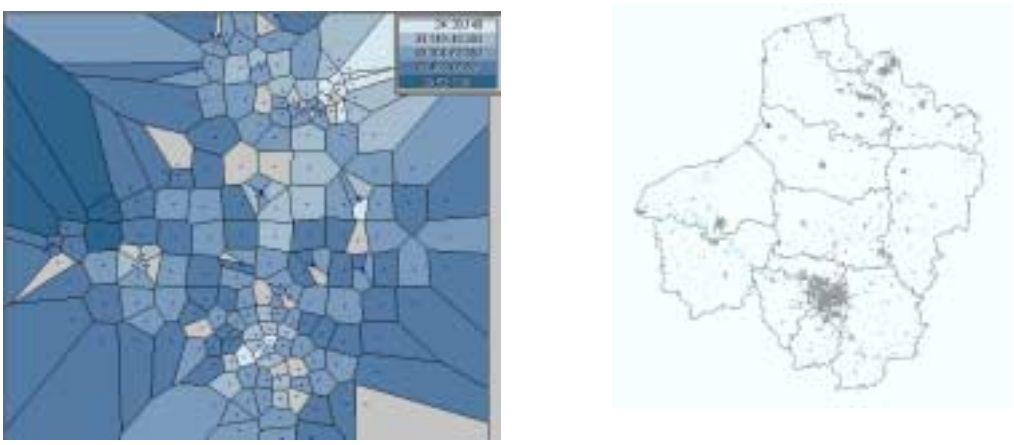


Figure 6 : Carte de Voronoi selon la méthode Cluster et carte représentant les zones urbanisées

- Ecart type (StDev) : cette représentation, qui ressemble beaucoup à celle des écarts interquartiles, met en avant les zones de forte variabilité, c'est-à-dire que la variable y est moins stable que dans une zone de faible variabilité. Ces zones correspondent aux zones frontières entre valeurs fortes et faibles, où les données varient fortement d'un polygone à l'autre. Celles-ci sont principalement situées dans la Somme et dans l'Aisne. Cette carte met aussi en valeur les valeurs extrêmes du littoral (ouest de la Seine Maritime). Bien que ces derniers points soient entourés de valeurs fortes, les écarts types y sont aussi très forts, ceci confirme leur « extrémisme » par rapport aux valeurs prises dans le reste de la zone d'étude.

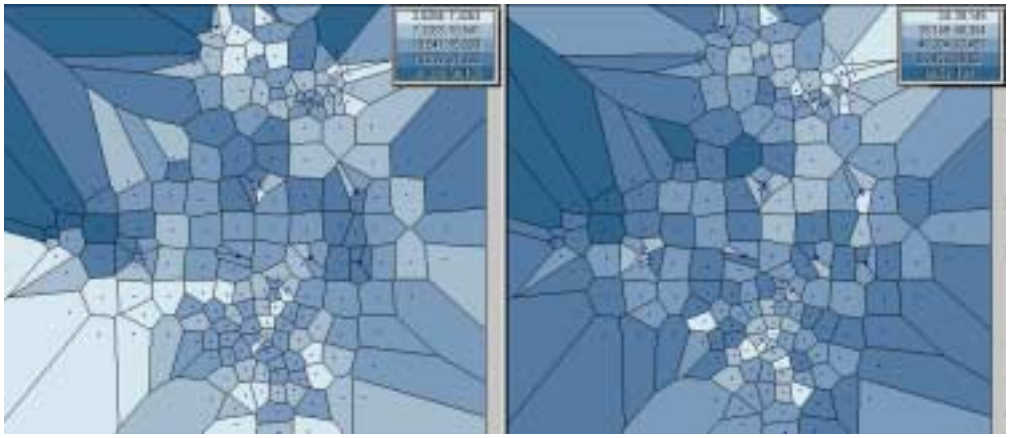


Figure 7 : Cartes de Voronoi selon les méthodes de l'écart type (à gauche) et simple (à droite), on se sert de la méthode simple pour étudier la carte de gauche

- Entropie (entropy) : comme nous l'avons vu dans la partie théorique, l'entropie augmente en même temps que le nombre de classes dans le voisinage. Elle donne donc une information sur la variation autour du point considéré, mais pas sur la carte entière. Dans notre cas, on voit que les zones de forte entropie sont dispersées sur le territoire. On identifie les zones de forte et moyenne variation sur le pourtour des agglomérations et aux zones frontières entre forte et faible concentration en ozone.

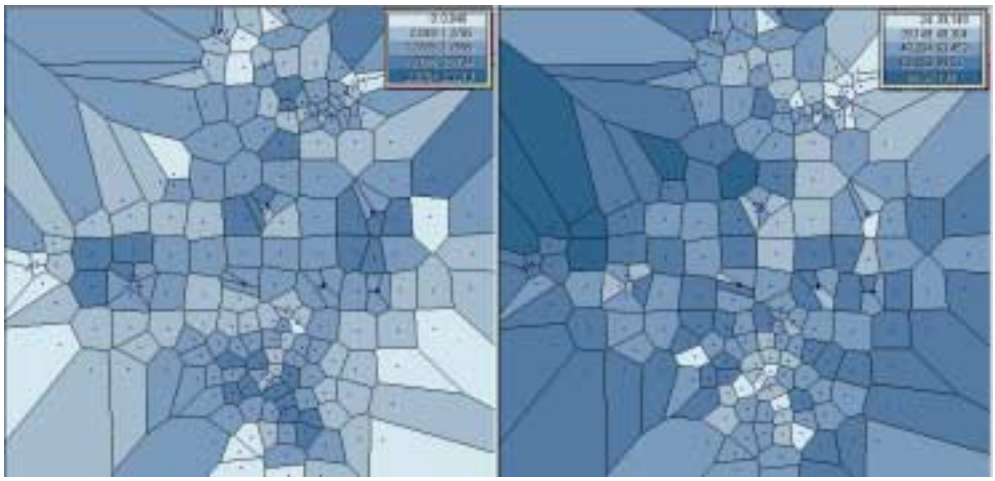


Figure 8 : Comparaison de cartes de l'entropie et selon la méthode simple

- Moyenne (mean), Médiane (median) et Simple : maintenant les représentations que nous allons utiliser ont tendance à lisser la carte. Nous n'allons plus mettre en valeur les fortes différences comme précédemment, mais plutôt voir les principales tendances que prennent les valeurs sur l'espace étudié. On voit sur ces cartes les zones qui prennent globalement de fortes ou de faibles valeurs. On peut noter que les informations sont redondantes d'une carte à l'autre. On voit donc que les grandes agglomérations prennent globalement des valeurs basses, comme nous l'avons déjà vu. Mais surtout les grandes valeurs littorales sont difficilement lissées par la carte de la moyenne, ceci confirme là encore leurs valeurs véritablement extrêmes. De plus l'Est de l'Aisne et de l'Ouest de la Somme et de la Seine Maritime prennent aussi en générale des valeurs relativement fortes.

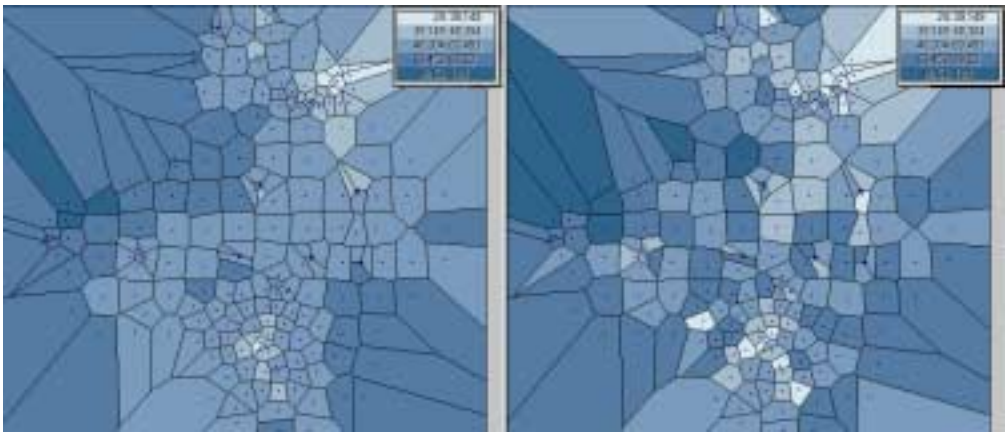


Figure 9 : Cartes de Voronoi selon la méthode de la moyenne (à gauche) et la méthode simple (à droite)

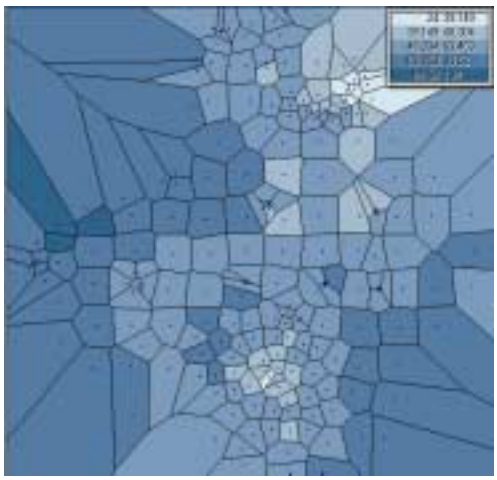


Figure 10 : Carte de la médiane

La carte du mode peut aussi renseigner sur les tendances que peuvent prendre les cellules, et fournit les mêmes informations.

2.6. NUAGE VARIOGRAPHIQUE

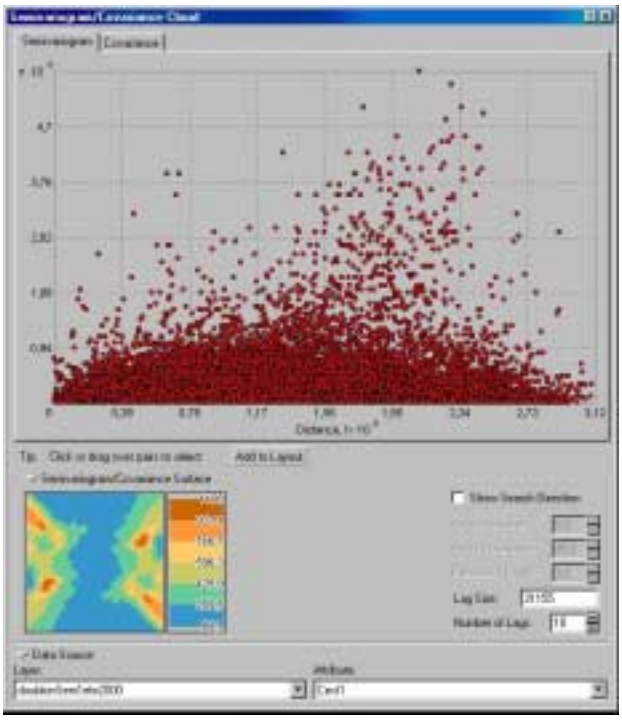


Figure 11 : Nuage variographique de la variable « Cest » dans la couche doublonSem1ete2000

2.6.1. DESCRIPTION DE L'OUTIL

Le nuage variographique est une représentation particulière du variogramme expérimental : à un point du nuage correspond un couple de données. «GA» affiche avec le nuage variographique la surface variographique, qui est une autre représentation du variogramme expérimental.

2.6.1.1 LE NUAGE VARIOGRAPHIQUE

Si on note $z(s_i)$ la valeur de la fonction aléatoire au $i^{ème}$ point du jeu de données, alors le semivariogramme expérimental (appelé couramment variogramme expérimental) s'exprime selon l'expression :

$$\hat{\gamma}(i,j) = \frac{\left(z(s_i) - z(s_j)\right)^2}{2}$$

où $\hat{\gamma}(i,j)$ est le variogramme expérimental pour le couple (i,j) séparé par une distance h.

La covariance expérimentale s'exprime quant à elle de la manière suivante :

$$\hat{C}_{ij}(h) = Cov(z(s_i), z(s_j)) = (z(s_i) - \bar{z}) * (z(s_j) - \bar{z})$$

Où $\bar{z}$ est la moyenne empirique des données.

Dans un cas plus général, lorsque l'on regroupe les données selon la distance et l'angle du segment du couple, on a :

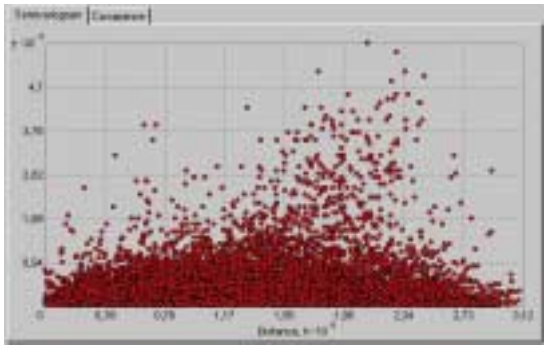
$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2 \quad \text{et} \quad \hat{C}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - \bar{Z}) * (Z(s_j) - \bar{Z})$$

où h varie selon la classe considérée,

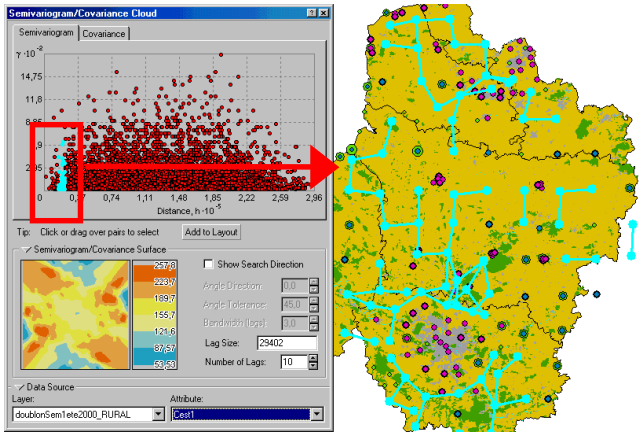
i et j varient pour une classe,

N(h) représente donc le nombre de paires (i,j) dans une classe.

La distinction dans la notation entre variogrammes et variogrammes expérimentaux se fait en ajoutant un « ^ » au dessus de leur symbole.



Le nuage variographique trace donc le variogramme expérimental pour chaque couple en fonction de la distance qui sépare les points de ce couple. Chaque point correspond à un couple de données, sélectionner les points permet de voir sur ArcMap le couple considéré.



Ici les couples qui sont séparés par la même distance sont sélectionnés, mais leurs variogrammes varient. Il est possible de même, de ne sélectionner que les couples ayant des variogrammes similaires, mais sur des distances variables.

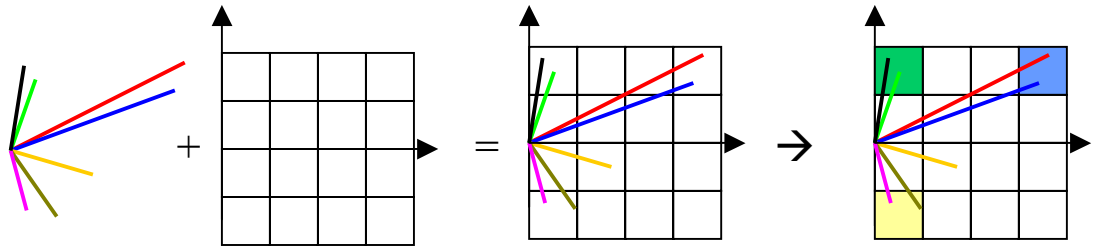
2.6.1.2 LA SURFACE VARIOGRAPHIQUE

La surface variographique est une autre représentation du variogramme expérimental qui renseigne sur la valeur moyenne de ce dernier pour une direction et une distance donnée. En effet ce mode de représentation se justifie pour identifier les anisotropies. Il débute par le regroupement des données en classes (binning) selon les distances et les orientations communes des couples de points. Ce regroupement se fait sous «GA» sur la base d'une grille.

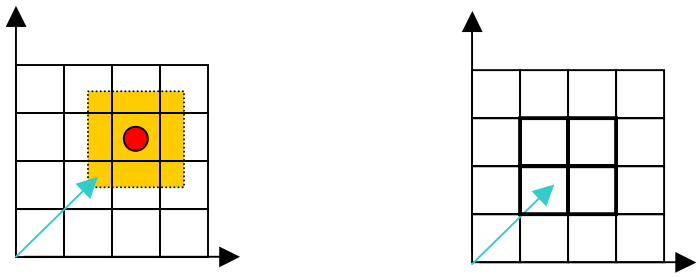
Pour ce faire, chaque vecteur formé par un couple de points est identifié et ils sont placés avec une origine commune, et orientés vers la droite de l'axe des ordonnées.

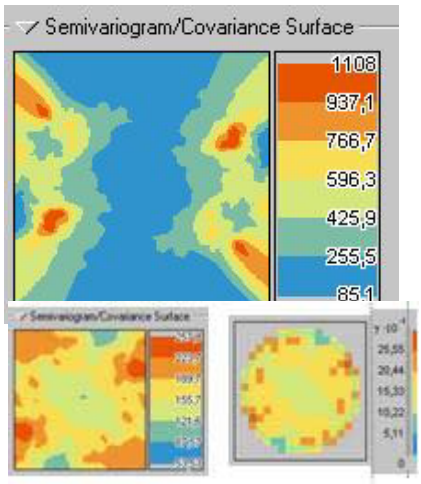
Ensuite une grille est superposée à cet ensemble de vecteurs dont l'utilisateur peut modifier le pas. Ceci permet d'ajuster et de choisir le nombre de classes.

Après cette étape, les vecteurs se trouvant dans la même case sont considérés comme faisant partie de la même classe. «GA» va alors calculer la moyenne des semivariogrammes pour chaque classe.



De plus, «GA» lisse la surface variographique en prenant en compte dans le calcul d'une cellule, à la fois les vecteur proches et contenus par celle-ci. Ce lissage est surtout utile dans le cas où la donnée est régulièrement espacée et où l'on risque d'avoir des sauts de valeur. Pour chaque cellule, «GA» superpose un carré de côté le double du pas de la grille. Ensuite, on affecte un poids aux vecteurs compris dans ce carré, ce poids croit linéairement du le côté vers le centre de la cellule. C'est donc avec tous ces vecteurs pondérés que «GA» calcule la valeur de la cellule. Il est à noter que chaque vecteur contribue à quatre cellules et que la somme des poids qui lui sont associés est égale à un.





On obtient donc pour chaque case une valeur qui est reportée sur une échelle, les couleurs froides représentant les faibles valeurs et les couleurs chaudes, les fortes.

La surface dessinée dans cet outil est différente de celle dessinée dans la commande Geostatistical Wizard. Celle de l'outil est lissée quelque soit le pas, alors que l'autre nous laisse voir les classes (crénelage).

Cette remarque s'illustre avec les deux exemples ci contre.

Il est aussi possible de ne dessiner que les points orientés dans une direction spécifique. Ceci, dès que la case « Show Search Direction » est cochée. Cette option permet d'examiner la nuée variographique selon différentes directions. Il est possible de voir si les valeurs subissent une influence dans une direction privilégiée. C'est ce que l'on appelle une anisotropie. Différents paramètre interviennent : la largeur de la bande (bandwidth), l'angle de tolérance et l'angle de direction.

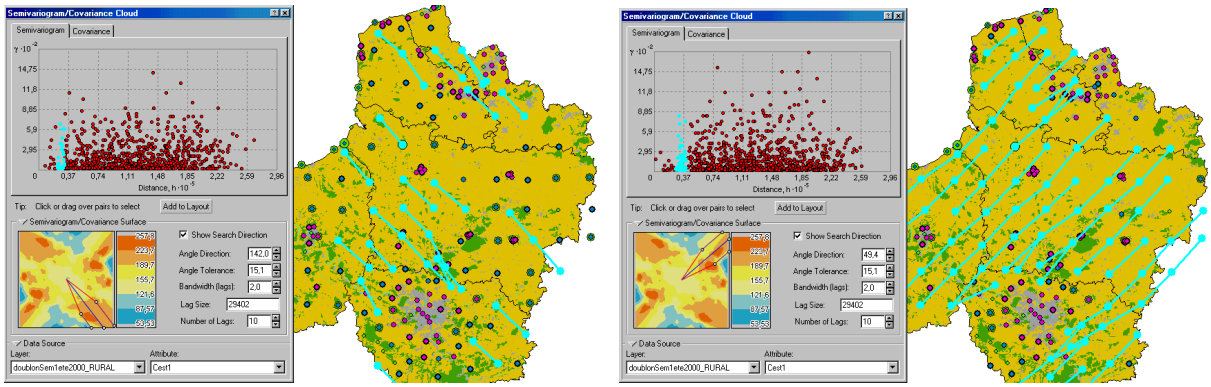


Figure 12 : Exploration de la nuée variographique selon les différents directions

Ces exemples montrent bien que les variations du semivariogramme ont lieu en fonction de la direction. L'orientation des couples sélectionnés suit la direction choisie, avec la tolérance précisée à l'écran.

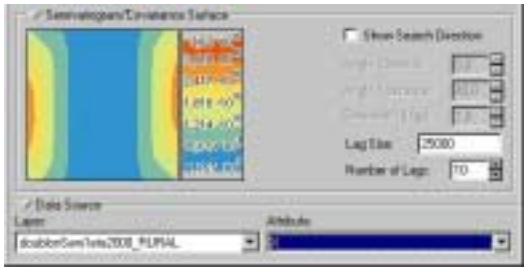
2.6.1.3 LA PRISE EN COMPTE DES ANISOTROPIES

L'anisotropie est une simple transformation des données selon l'équation :

$$s^+ = \begin{pmatrix} \sqrt{r} & 0 \\ 0 & \sqrt{r} \end{pmatrix} \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} * s$$

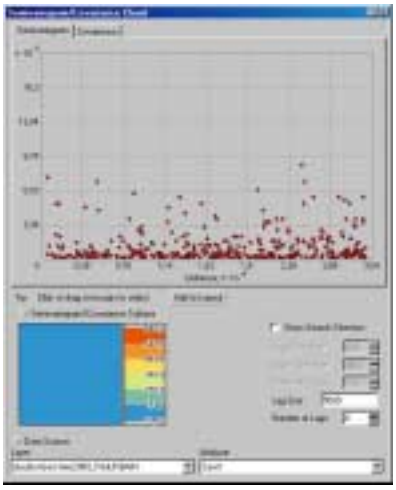
où θ est l'angle de rotation, r le rapport entre le grand axe et le petit axe de l'ellipse.

Après cette transformation, on calcule les distances $\left\| s_i^+ - s_o^+ \right\|$ pour le variogramme et la covariance. Dans le cas d'un variogramme gigogne, il est possible d'associer l'anisotropie à chacun des modèles qui composent ce dernier.



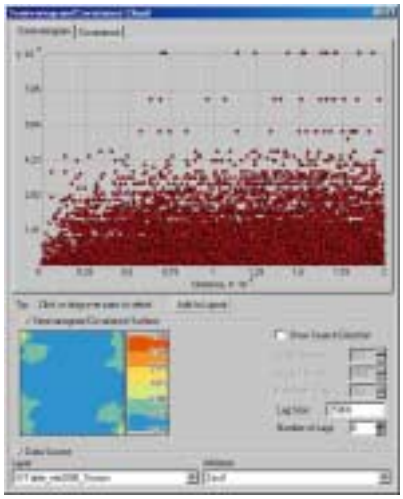
Par comparaison voici un exemple de très forte anisotropie : en prenant simplement la variable X pour la même couche de données. C'est-à-dire que maintenant nous dessinons la surface variographique des coordonnées des points de relevé selon la direction x.

Les valeurs du semivariogramme augmentent avec la distance horizontale qui sépare les points, mais pas avec la distance verticale. En effet X varie selon la direction horizontale tandis que c'est Y qui varie verticalement.



Pour l'étude du variogramme à faible distance, les valeurs rurales ne permettent pas un ajustement suffisant. Il faut se baser sur les données urbaines et périurbaines. L'analyse du nuage variographique correspondant est assez difficile, car il est très homogène sur toute la distance choisie : 0 à 30 km. On le voit bien sur la surface variographique qui est totalement uniforme.

2.6.2. APPLICATION AU CAS DE LA POLLUTION PAR L'OZONE



Dans notre cas le variogramme expérimental que l'on obtient sans écarter aucune donnée ne ressemble en rien à celui présenté dans l'Etude n°15. En effet les valeurs maximales sont de 8460(µg/m^3)² alors que le nuage de l'étude monte jusqu'à 5000(µg/m^3)². Rappelons que lors de la transformation du tableau de valeur au format excel vers un fichier de forme sur ArcMap, nous avons dû modifier les valeurs inconnues « N/A » en autant de « 0 ». C'est cette étape qui a faussé ce premier variogramme. Maintenant, réexaminons les valeurs dont nous disposons et les dispositions prises dans l'étude.

D'après l'examen des données fait par l'étude, il ne faut s'intéresser qu'aux données rurales qui sont le plus uniformément réparties. Les nuages de points obtenus sont similaires à ceux de l'Etude n°15. C'est-à-dire que lorsqu'on prend en compte les données de tous les types de sites (en ayant écarté les valeurs nulles), le couple de plus haute valeur est à environ 5000(µg/m^3)² alors que la variance calculée sous ArcMap est de 247 (µg/m^3)². Les plus fortes valeurs du semivariogramme sont dues aux trois points littoraux de forte valeur.

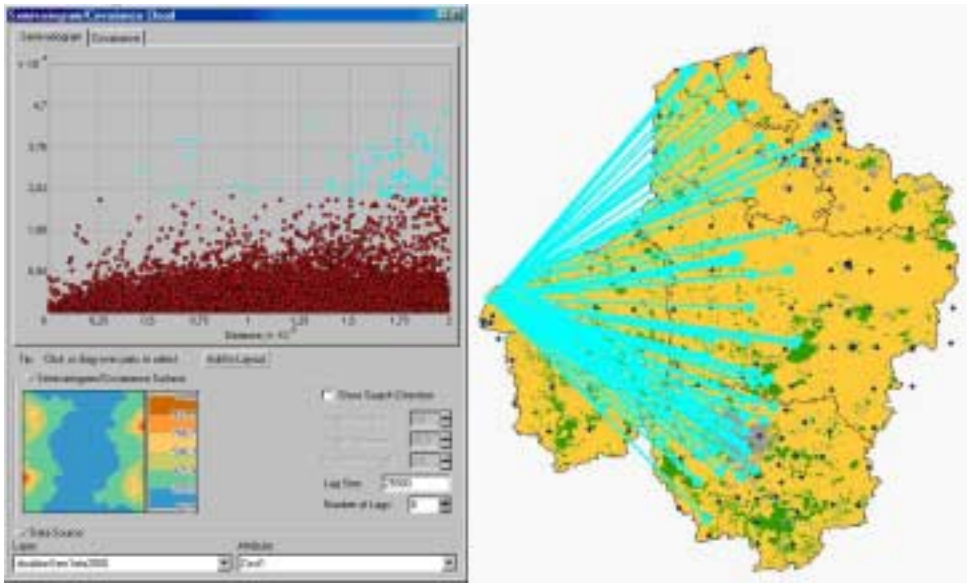


Figure 13 : Nuage variographique de tous les types de données et correspondance entre les valeurs fortes et leur localisation sur la carte (couples mis en surbrillance)

Ces valeurs littorales influent fortement sur le variogramme, elles tendent à créer une dérive comme nous l'avions vu dans le module « Trend Analysis ».

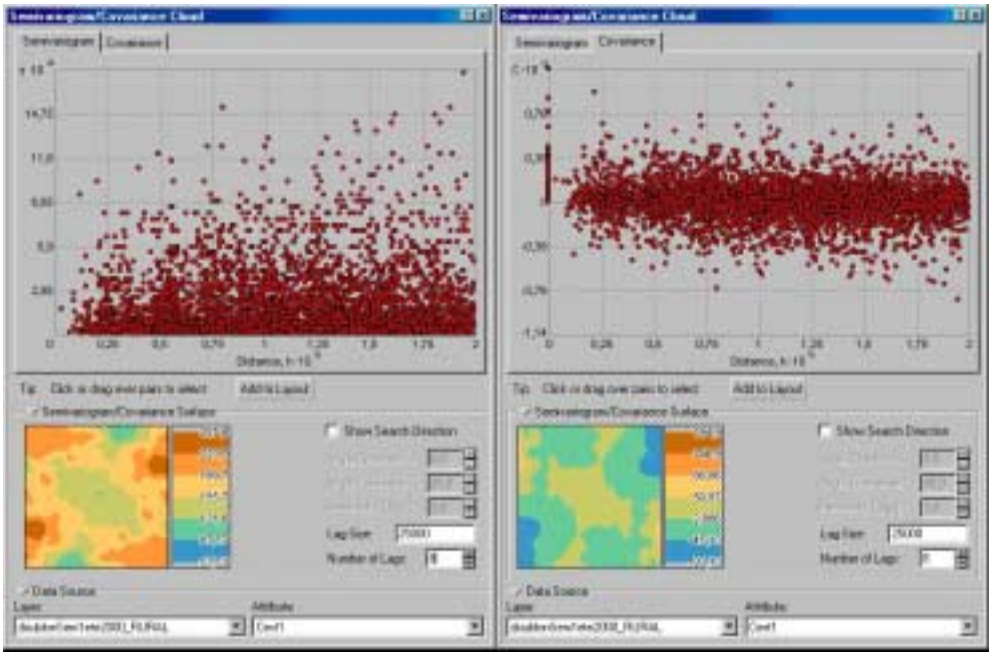
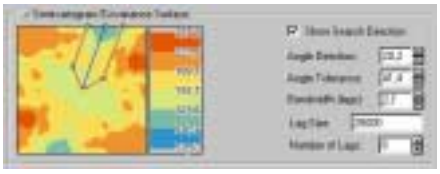


Figure 14 : Le nuage variographique et covariographique avec les paramètres ainsi que le jeu de données ajustés

L'espacement et la dispersion sur tout le territoire des données rurales, permet d'accéder à la pollution de fond. Dans ce cas, les valeurs du nuage de points sont mieux adaptées car ce dernier est moins écrasé par les fortes valeurs des couples extrêmes ($Z_{littoral} - Z_i$). On remarque par ailleurs que la surface du covariogramme est inversée par rapport à la surface du variogramme.



La surface variographique présente une très légère anisotropie dans la direction 20°. Elle peut se justifier par le fait que dans le cas d'une isotropie, la surface variographique devrait être constituée de cercles concentriques pour les différentes valeurs.

Or le pourtour de cette surface présente une majorité de valeurs relativement fortes (on ne descend pas au dessous du jaune), sauf un « trou » relativement marqué dans la direction 20°. Cependant la tendance marquée par ce « trou » n'est pas observable dans cette direction pour toutes les distances. L'expérience pourra nous dire si on peut considérer cette variation comme une réelle anisotropie. Pour s'en assurer on pourra ajuster deux modèles, l'un isotrope, l'autre anisotrope, puis comparer leurs validations croisées.

Cet outil permet d'obtenir un nouveau point de vue sur les données que l'on manipule. Il permet en effet d'identifier les couples de forte variabilité et de confirmer les observations précédentes. En faisant varier pas et nombre de pas, on peut se focaliser sur une partie ou bien sur toutes les données. On fait ainsi un tri dans les propriétés des données. En effet, plus le pas est grand, et moins on a de couples, et plus le nombre de pas est élevé, plus on va chercher des couples de points éloignés.

Quelques limites se présentent dans l'utilisation de cet outil. La première est qu'il n'est pas possible d'obtenir les statistiques de la sélection directement depuis le module. En effet, nous avons dû repasser sous ArcMap pour obtenir la variance. Cette manipulation ne pose pas de problème lorsque l'on travaille avec les outils d'exploration. Mais plus tard sous Geostatistical Wizard, on ne peut plus passer sous ArcMap : on ne pourra plus faire d'analyse rapide des données à ce stade.

La seconde limite observée est qu'il n'est pas possible de tracer le variogramme expérimental. Cela peut constituer une gêne car pour l'analyse à petite distance, il est difficile de tirer des observations et des remarques à partir d'un nuage uniforme. Un variogramme expérimental pourrait permettre de mieux identifier les variations.

3. CONSTRUCTION DE CARTES PAR KRIGEAGE

Dans cette partie, nous utilisons désormais la commande Geostatistical Wizard du module «GA». C'est cette commande qui va permettre de construire différents types de cartes contrairement aux outils détaillés précédemment qui ne permettaient qu'une exploration des données.

Cette partie suit les étapes de la création d'une carte. C'est à dire que les fenêtres de progression seront expliquées et commentées les unes à la suite des autres.

3.1. PREMIÈRE FENÊTRE DE PROGRESSION : CHOOSE INPUT DATA AND METHOD

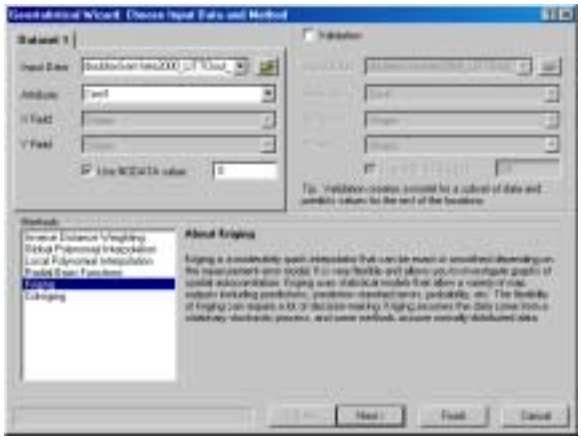


Figure 15 : Choix de la variable et de la méthode

Cette fenêtre permet de rentrer les données que l'on souhaite analyser. Pour cela on sélectionne la couche puis le champs de la variable qui nous intéresse. En l'occurrence la couche qui ne contient que les données rurales, urbaines et périurbaines (« doublonSem1ete2000_LITTOout ») est sélectionnée, et les valeurs utilisées sont celles de la première semaine de la campagne d'étude de l'ozone (« Cest1 »).

Ce choix est dicté à la fois par l'étude des données menée dans la première partie, mais nous suivons aussi les choix qui ont été faits lors de l'étude des données sous Isatis®. Rappelons en effet que les données de type littoral sont très fortes et induisent une dérive dans les données, nous les écartons donc pour ajuster le modèle.

De plus les données rurales ont une bonne dispersion sur l'ensemble du territoire, elles servent donc à caractériser le modèle variographique à grande échelle. Quant aux données urbaines et périurbaines, elles servent à ajuster le modèle pour une petite échelle.

Contrairement aux outils où nous avons dû manipuler les données pour supprimer les données affectées arbitrairement à 0, ici il est possible de préciser une valeur qui ne soit pas comptée dans le variogramme. Il suffit en effet de cocher la case « Use NODATA value » et de préciser la valeur qui ne doit pas être prise en compte.

La méthode qui nous intéresse ici est le Krigeage linéaire et monovariable.

3.2. DEUXIEME FENETRE : GEOSTATISTICAL METHOD SELECTION

Figure 16 : Choix de la méthode Géostatistique et des options à appliquer sur la variable

En se basant sur l'étude, on sélectionne un krigeage ordinaire. Le principe du krigeage ordinaire est expliqué brièvement dans un premier temps. Puis les résultats par rapport aux données sont montrés.

3.2.1. PRINCIPE DU KRIGEAGE

Dans cette partie notre intérêt ne se porte que sur les méthodes linéaires de krigeage. Ces méthodes sont le krigeage ordinaire, le krigeage simple et le krigeage universel.

Le principe commun pour toutes les méthodes de krigeage, est de trouver un estimateur de la variable qui soit sans biais (espérance nulle de l'erreur d'estimation), et qui minimise la variance d'erreur d'estimation.

L'estimateur utilisé est une combinaison linéaire des données :

$$Z^*(s) = \sum_i \lambda_i Z(s_i)$$

Où les $Z(s_i)$ sont les variables aux points échantillons s_i et les λ_i les poids associés à chacun des $Z(s_i)$.

3.2.2. LE KRIGEAGE ORDINAIRE

C'est selon cette méthode retenue dans l'étude LCSQA que nous allons étudier les données dans un premier temps.

Dans le krigeage ordinaire la moyenne est supposée inconnue. Le système de krigeage est :

Système de krigeage

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

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

Notons qu'il y a une condition supplémentaire pour les pondérateurs $\boldsymbol{\lambda}$. 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 $\boldsymbol{\mu}$, qui est appelée coefficient de Lagrange.

3.2.3. LE KRIGEAGE SIMPLE

Dans ce cas, la valeur moyenne de la variable que l'on doit estimer est supposée connue. La moyenne va avoir une influence dans les équations de krigeage, et notamment dans l'expression de l'estimateur :

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

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

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

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

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

$m = moyenne \cdots connue$

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

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

L'expression $\left(1 + \sum_i \lambda_i'\right)$ est le poids de la moyenne, moins il y a d'information dans le voisinage du krigeage et plus celui-ci sera important. Il en découle que le poids de la moyenne donne une indication sur le nombre de voisins et leur importance dans l'estimation.

3.2.4. LE KRIGEAGE UNIVERSEL

Cette méthode ressemble au krigeage ordinaire, mais suppose que la moyenne de la variable varie selon la position dans l'espace. Cette méthode permet donc de prendre en compte ce que l'on nomme la dérive de la variable, c'est la tendance moyenne de variation qu'à celle-ci. Il y a donc une dichotomie entre la tendance moyenne, qui est déterministe, et la variable résiduelle:

$$Z(s) = m(s) + \varepsilon(s)$$

La dérive peut être modélisée par une fonction polynomiale dont les paramètres sont estimés grâce aux données (krigeage universel) ou en s'appuyant sur la connaissance d'une variable auxiliaire mieux échantillonnée mais moins précise (dérive externe).

La seconde option, qui est une méthode de la géostatistique non-stationnaire, n'est pas disponible dans « GA ».

En krigeage universel, le résidu a une moyenne théorique nulle. Le choix de l'ordre de la dérive peut se faire en ajustant un polynôme d'ordre n.

Dans « GA», la dérive est modélisée par un polynôme qui peut être de degré trois au maximum. Par exemple, si on prend une dérive de degré deux, sur une variable bidimensionnelle, elle s'exprime :

$$m(s) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 xy + \beta_4 x^2 + \beta_5 y^2$$

Il faut remarquer que la non- stationnarité d'un phénomène est fonction de l'échelle de travail, par exemple, un phénomène non- stationnaire peut être décelé par un variogramme calculé à grandes distances mais si l'on réduit considérablement l'échelle il est possible de masquer la non –stationnarité et de se retrouver avec un variogramme avec palier.

3.2.5. LES OPTIONS DISPONIBLES DANS LE MODULE

Après cette étape théorique, le traitement numériques des données disponibles est exposé.

Dans cette fenêtre, en plus du choix de la méthode de krigeage et du type de carte à réaliser, on peut choisir deux options sur la variable sélectionnée : doit-on la transformer et a-t-elle une dérive ? Suivant la méthode choisie, plusieurs possibilités sont disponibles, nous les récapitulons dans le tableau ci-dessous:

		Type de krigeage	**Transformations Box-Cox, arcsinus et logarithme (BAL)**	**Transformation Normal Score (NST)**	**Suppression de la dérive**
Géostatistique Linéaire	Géostatistique stationnaire	Ordinaire	Disponible	Non disponible	suppression de la dérive, après la transformation
		Simple	Disponible	Disponible	non disponible
	Géostatistique non stationnaire	Universel	Disponible	Non disponible	modélisation de la dérive, après la transformation
Géostatistique non Linéaire	Géostatistique stationnaire	d'Indicatrice	non disponible	Non disponible	non disponible
		de Probabilité	non disponible	Non disponible	non disponible
		Disjonctif	Disponible	Disponible	modélisation de la dérive, avant la transformation NST, après la transformation BAL

Tableau 1 : Récapitulatif des méthodes et des options disponibles dans le module

Les transformations de type « BAL » ont pour but de modifier la variable pour la faire ressembler à une variable gaussienne. La transformation « normal score » peut transformer effectivement la variable en variable gaussienne.

3.2.5.1 UTILISATION DE LA DERIVE

Le tableau montre que l'on peut utiliser la dérive avec le krigeage ordinaire. Le traitement de la dérive dans ce cas se fait selon les étapes :

- Le module «GA» va demander l'ordre de la dérive pour estimer les paramètres du polynôme de la dérive.
- Il va ensuite considérer le résidu comme la différence entre la valeur réelle et la dérive. En d'autres termes, «GA» a supprimé la dérive de la variable brute pour ne garder que le résidu.
- Le variogramme dessiné est celui des résidus, c'est donc sur les résidus que l'on ajuste le modèle.
- Ensuite on va estimer les résidus selon le krigeage ordinaire.

- A la fin, on va rajouter la valeur correspondante de la dérive au résidu estimé.

C'est le krigeage du résidu. Etant donné que «GA» ne propose pas de modèle d'ajustement non stationnaire (modèle linéaire par exemple), c'est pour pouvoir ajuster les données que cette option est disponible. En supprimant la dérive, on rend la variable stationnaire.

Dans l'option « krigeage universel », le module estime là aussi les coefficients de la dérive, mais celle-ci est prise en compte dans le calcul du variogramme.

Sa prise en compte libère une fenêtre de dialogue. Il est possible d'y choisir l'échelle sur laquelle on adapte la dérive. Si on considère une dérive globale (sur la zone entière), alors on tend vers l'extrême « global ». Si au contraire, on a une dérive locale qui se situe sur de petites portions de la zone, on tend vers « Local ». Il faut que l'utilisateur choisisse donc le rayon d'action de sa dérive. Cette fenêtre affiche la surface correspondante aux valeurs que prend la dérive dans la zone d'étude.

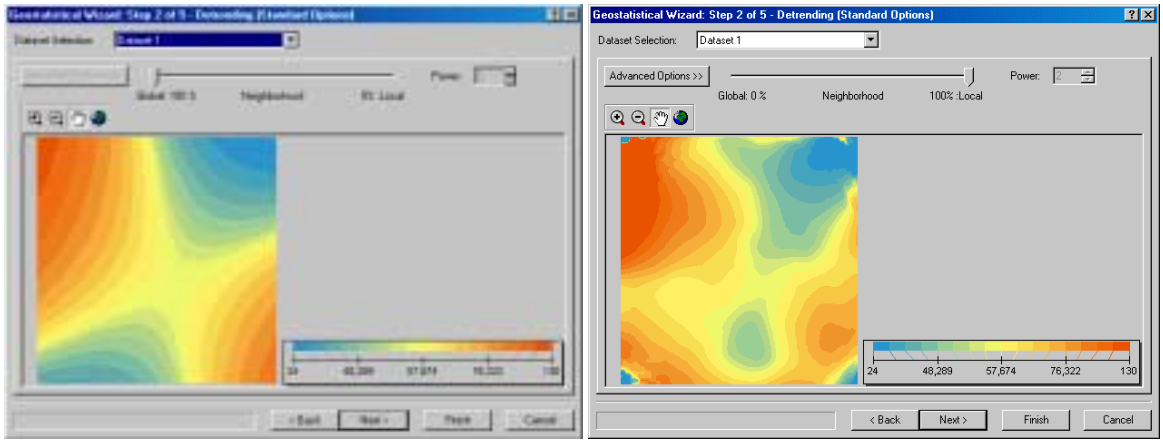


Figure 17 : Fenêtre pour effectuer le choix de la dérive globale ou locale, illustré avec l'exemple de l'ozone

3.2.5.2 DESCRIPTION DES TYPE DE CARTES DISPONIBLES

Les types de cartes qui sont disponibles avec les méthodes de krigeage :

- Prediction Map :

  C'est le type de carte qui permet de construire une carte d'estimation de la variable choisie sur toute l'aire d'étude. Les calculs effectués sont les calculs d'estimation de la variable selon la méthode de krigeage choisie. Cependant ce type de carte est disponible pour toutes les méthodes de krigeage, excepté le krigeage par indicatrice et le krigeage de probabilité. C'est ce type de carte qui permet une analyse directe et quantitative de la variable.

- Quantile Map :

  Ce type de carte est disponible pour les méthodes ordinaire, simple et universelle de krigeage. Dans la seconde fenêtre de dialogue de Geostatistical Wizard, l'utilisateur peut préciser le quantile qu'il désire prendre en compte. Cette carte ne peut être calculée que si la variable suit une loi normale et est stationnaire.

  Un quantile p est la valeur pour laquelle la proportion p de valeurs de la variable lui sont inférieures. On a donc une probabilité p d'avoir la réalisation de la variable qui soit inférieure au quantile.

  Nous avons vu que la variable doit tendre vers une loi normale, de cette manière, la différence entre l'estimateur et la réalisation tend elle aussi vers une loi normale. Les quantiles calculés pour la variable sont donc les quantiles de la loi normale. La carte nous montre les valeurs limites pour lesquelles on est sûr à p% (pour un quantile p) que la valeur de l'estimation y est égale ou inférieure.

- Probability Map :

  Ce type de carte est disponible pour toutes les méthodes krigeage. Le principe de ce type de carte est très proche du type précédent. En effet, le module va calculer la probabilité que l'on dépasse ou que l'on ne dépasse pas un certain seuil. L'utilisateur va donc préciser dans la seconde fenêtre de dialogue à la fois le seuil en unité de la variable, puis si on calcule la probabilité de se trouver au dessus ou en dessous du seuil. Il est possible d'ajuster le seuil en cliquant sur le bouton « Set … », une nouvelle fenêtre s'ouvre dans laquelle on peut modifier les paramètres de choix.

  Une fois ces paramètres réglés, la surface représente la probabilité que l'estimateur soit inférieur ou supérieur au seuil fixé. La carte va donc quantifier directement un risque de dépassement (en trop ou en moins) du seuil. Elle ne renseigne sur les valeurs de la variable qu'indirectement. En effet l'utilisateur ne va avoir une idée de la valeur au point « s » que parce qu'il connaît le seuil et la probabilité de dépassement. Par contre elle la qualifie qualitativement car nous montre un risque sur cette variable.

- Prediction Standard Error Map :

  Cette carte représente l'écart type de l'estimateur qui est la racine carrée de la variance d'estimation. Elle est disponible pour toutes les méthodes de krigeage, sauf pour le krigeage d'indicatrice et de probabilité elle est remplacée par la Standard Error of Indicators Map.

  Cette carte ne nous renseigne pas, elle non plus, quantitativement sur la variable. En effet, la surface représentée est l'écart type de l'estimation. C'est une carte qui quantifie la dispersion possible de la valeur vraie autour de la valeur estimée. Donc plus l'écart type est grand en un point, et moins précise est l'estimation de ce point. Dans le cas gaussien, 95% du temps, la vraie valeur est comprise dans l'intervalle formé par la valeur estimée à laquelle on ajoute ou on retranche le double de l'écart type. Plus on s'éloignera des points de mesure, et plus les valeurs de la carte seront fortes.

- Standard Error of Indicators Map :

  Ce type de carte n'est disponible que pour un krigeage par indicatrice, de probabilité ou disjonctif. C'est la représentation de l'écart type de la fonction indicatrice calculée à partir du jeu de données. Les remarques sur le type de carte précédent sont donc là aussi valables. Celle-ci va nous indiquer la précision de l'estimation en un point.

Les cartes de probabilité et de quantile sont des types de carte de probabilité et de risque, avec des variables qui se rapprochent de variables gaussiennes. Or ces cartes de probabilité sont prises en compte par la géostatistique non linéaire. Dans ce cas, la variable est transformée pour être effectivement gaussienne. Le résultat obtenu sera donc plus rigoureux que des cartes faites à partir des méthodes de la géostatistique linéaire où l'hypothèse que la variable est gaussienne n'est pas respectée.

Une carte de probabilité ou de quantile construite à partir d'une méthode géostatistique linéaire est un raccourci dans l'exploitation des données. En effet il est possible d'avoir tous les types de cartes avec une seule méthode donc une seule analyse des données. Mais les résultats obtenus ne seront que des approximations.

En ce qui concerne l'aspect de la fenêtre, le coin supérieur gauche la rubrique « Selection » récapitule les choix effectués dans la rubrique « Geostatistical Methods ». Les autres rubriques, qui sont indisponibles avec les choix que nous avons faits, sont relatives aux autres choix de carte et aussi si on dispose de plusieurs jeux de données.

3.2.6. LES CHOIX EFFECTUES POUR LE CAS DE LA POLLUTION PAR L'OZONE

Dans notre cas, et d'après les analyses de l'étude LCSQA n°15, nous ne sommes pas sûr d'avoir une dérive pour les données rurales. On choisit dans un premier temps un krigeage ordinaire sans dérive. De plus on n'applique aucune transformation aux données rurales, car comme vu dans l'exploration de données, la distribution des données n'est pas asymétrique. Nous satisfaisons donc les hypothèses préalables à l'analyse.

Par ailleurs on ne s'intéresse dans un premier temps qu'à faire une carte d'estimation, on sélectionne donc dans la fenêtre « Prediction Map ».

Dans la phase d'exploration des données, l'outil Trend Analysis présentait une courbe ajustée à la projection des données selon une direction d'environ 160°. On peut donc se demander si cette courbe ne représente pas une dérive. Sur ce jeu de données nous n'avons pas de phénomène physique qui puisse expliquer une dérive.

Donc si on considére qu'il y a effectivement une dérive, celle-ci ne sera pas une dérive externe. On choisira plutôt un krigeage universel où la dérive est un polynôme dont on estime les coefficients.

Dans la partie suivante, nous allons traiter la variable selon le krigeage universel pour ensuite comparer les résultats des deux méthodes.

3.3. TROISIEME FENETRE : SEMIVARIOGRAM/COVARIANCE MODELING

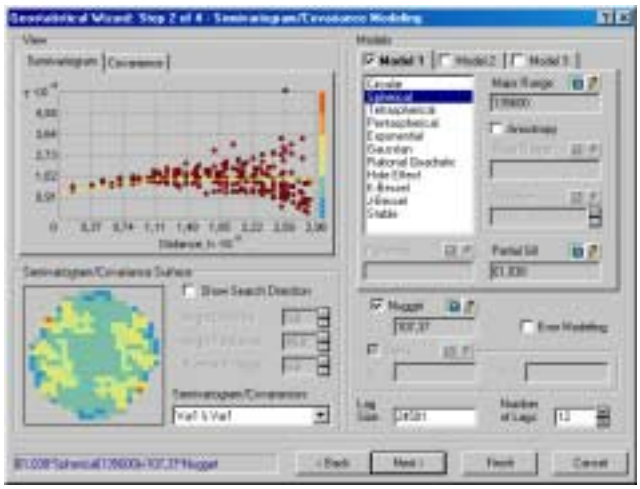
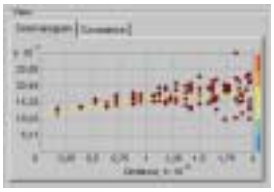


Figure 18 : Fenêtre d'ajustement des modèles variographiques, avec les paramètres par défaut

Cette fenêtre permet d'ajuster un modèle de variogramme après avoir choisi la méthode d'interpolation et la variable à interpoler.

3.3.1. DESCRIPTION DES RUBRIQUES DE LA FENETRE



La rubrique « View » affiche la nuée de points du variogramme ou du covariogramme, il suffit de cliquer sur l'onglet correspondant. On y voit aussi la courbe du modèle sélectionné qui est représentée par la ligne jaune. L'échelle de la surface variographique y est aussi jointe sur le bord droit.

Dans la rubrique « Models » on choisit en premier lieu les modèles à ajuster à la nuée de points. Les modèles 1, 2 puis 3, dès que l'on coche leur case, vont s'ajouter les uns aux autres. De cette manière, «GA» permet de construire ce que l'on appelle une « structure gigogne » constituée de trois modèles différents en plus d'un effet de pépite. Même si c'est un modèle à part entière, ce dernier est n'est pas traité comme les autres modèles. Sur chacun des onglets, on peut choisir un des modèles qui sont disponibles, ceux-ci ont tous un palier, il est par conséquent impossible de modéliser une variable qui ne soit pas stationnaire ou intrinsèquement stationnaire.

Ensuite, les paramètres liés à chacun des onglets peuvent être réglé par l'utilisateur ou bien mis à une valeur par défaut. Ce choix s'effectue en poussant les petits boutons en haut à droite de chaque des zones d'affichage des paramètres. Le crayon correspond au choix de l'utilisateur (la zone est blanche), la calculatrice correspond à une attribution par défaut du paramètre.

- « Major Range » correspond à la portée du modèle dans le cas isotrope, et à la grande portée dans le cas anisotrope. Dans le cas où le modèle atteint asymptotiquement son palier, major range et minor range valent 95% de cette valeur.

- Cocher la sous rubrique « Anisotropy » provoque la prise en compte de nouveaux paramètres :

- « Minor Range » qui est la petite portée,
- et « Direction » qui est l'angle positif entre le Nord et le grand axe de l'ellipse des portées.

- « Partial Sill » est le palier partiel, en effet, le palier total est la somme de l'effet de pépite et du palier partiel, il correspond au palier du modèle. Il ne peut pas varier avec la direction, il est donc définit pour le modèle et est le même dans toutes les directions. «GA» ne prend donc pas en compte les anisotropies zonales (palier et portée varient), mais seulement les anisotropies géométriques (seule la portée varie).

- « Parameter » est un paramètre relatif aux modèles K-Bessel, J-Bessel et Stable.

Maintenant, les paramètres que nous allons voir sont plus généraux et ne s'appliquent pas uniquement au modèle qui lui correspond.

- « Nugget » est l'effet de pépite, c'est un modèle à part entière, comme nous l'avons vu. Mais sa particularité est qu'il atteint tout de suite son palier, sans avoir de portée, il est donc discontinu. Il sert à modéliser les variations brusques de la variance, c'est donc grâce à ce modèle que l'on quantifie la structure (ou la non structure) de la variable.

- « Lag Size » est la taille du pas (lag en anglais) de la grille de la surface variographique. Comme nous l'avions vu, c'est en le modifiant que l'on modifie aussi le nombre de couples pris en compte dans le nuage variographique.

- « Number of Lags » est le nombre de pas, on va donc là aussi pouvoir influer sur le nombre de couples que l'on va prendre en compte.

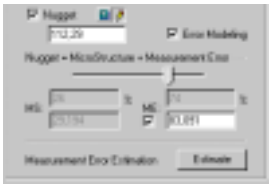
Plus généralement, la combinaison de ces deux paramètres va nous permettre de « zoomer » sur les données. En les modifiant, on joue sur le nuage variographique, et indirectement sur le modèle qui y est ajusté.

Avec les choix que nous avons fait, la sous-rubrique « Shift » reste inaccessible. Aucun paramètres n'est attribué par défaut dans les zones d'affichage « X » et « Y ». En effet ces paramètres servent à faire correspondre les variables dans le cas du cokrigeage

Lorsque l'on coche la case « Error Modeling », la rubrique change pour faire apparaître de nouveaux paramètres. L'effet de pépite peut se décomposer en variation à très petite échelle (microscale variation ou microstructure, $\eta(s_i)$) et en erreur de mesure (measurement error, $\delta_t(s_i)$).

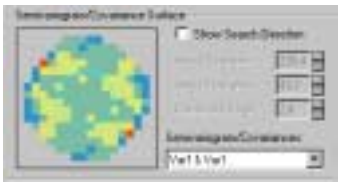
«GA» permet de choisir le poids de chacun dans l'effet de pépite et rappelle dans les zones d'affichage les valeurs de chacun de ces paramètres. De façon logique, lorsqu'on décoche la case de l'effet de pépite, ces paramètres ne sont plus accessibles.

Au cours des calculs de l'estimation, le module ne prend pas en compte l'erreur de mesure.



Lorsque l'on a des valeurs triple ou multiple dans l'étude, et que l'on coche la case de l'erreur de mesure située juste en dessous de « ME », il est possible d'estimer l'erreur de mesure en poussant le bouton « Estimate » ou bien de la renseigner soi même.

Cette option est disponible lorsque l'on n'a pas retiré dès le début les valeurs multiples (cf 2.1Remarques générales).



Intéressons nous maintenant à la dernière rubrique : « Semivariogram/Covariance Surface ». La surface variographique (ou de la covariance) y est affichée selon le procédé décrit dans le descriptif du nuage variogaphique.

On distingue en effet les cases, ainsi que la forme circulaire qui découle de ce procédé. La case « Show Search Direction » qui est la même que dans l'outil nuage variographique a un fonctionnement et un intérêt identiques.

La distinction réside dans l'intéraction avec la nuée de points. En effet, nous avons vu que le modèle est superposé à la nuée de points, en cas d'anisotropie, c'est le modèle qui correspond à la direction montrée qui est affiché. De plus, si on compare les deux surfaces, on s'aperçoit que leurs échelles ne sont pas tout à fait identiques. La première a un pas de 34,04 $(\mu g/m_3)^2$ et la première classe débute à 53,53 $(\mu g/m_3)^2$ tandis qu'avec des paramètres identiques (8 pas de 25000m), la seconde surface présente une première classe qui débute à 0 et s'incrémente de 51,1 $(\mu g/m_3)^2$. Malgré ces différences et un resserrement de la maille de la surface différent, on voit que les deux surfaces sont très semblables.

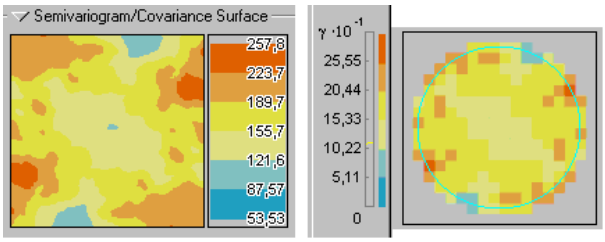


Figure 19 : Comparaison des surfaces variographiques de l'outil d'exploration de données et de la commande d'interpolation « Geostatistical Wizard », les paramètres sont : 8 pas de 25000m

En dernier lieu, nous remarquons dans le coin inférieur gauche de la fenêtre une petite zone d'affichage. C'est là que sont récapitulées les informations des rubriques que nous venons de décrire. C'est en effet la formule finale du modèle de variogramme, ajusté, c'est-à-dire le variogramme théorique qui va être pris en compte lors du krigeage qui apparaît.

$$\gamma_{théorique}(h) = palier * f(distance, portée)$$

Le variogramme gigogne étant la somme de plusieurs variogrammes théoriques

3.3.2. APPLICATION AU CAS DE LA POLLUTION PAR L'OZONE

Dans le cas de la variable qui nous intéresse : les relevés de concentration d'ozone dans les sites ruraux, nous avons réutilisé le modèle ajusté de l'étude LCSQA n°15, c'est-à-dire :

$$\gamma(h) = 35 + 81 * \left(1 - e^{-\left(\frac{1,73*h}{26000}\right)^2}\right) + 59 * \left(1 - e^{-\left(\frac{1,73*h}{136000}\right)^2}\right)$$

Ce qui correspond à :

Variogramme ajusté = effet de pépite + modèle gaussien à faible portée
+ modèle gaussien à longue portée

La formule générale d'un modèle gaussien étant :

$$\gamma(h) = C * \left(1 - e^{-\left(\frac{1,73*h}{a}\right)}\right) \quad , \text{ pour tout h}$$

La formule générale de l'effet de pépite est :

$\gamma(h) = 0$ si h = 0
C si h ≠ 0

où C est le palier et a la portée.

Dans l'ajustement effectué, l'effet de pépite est égal à la variance moyenne de l'erreur de mesure. Le choix du modèle gaussien s'est fait en relation avec la nature de la pollution. La pollution par l'ozone est un phénomène à grande échelle, donc peu variable sur de courtes distances, cette caractéristique est souvent rendue avec un modèle ajusté parabolique pour des h faibles. Notre cas semble se mieux s'adapter à une combinaison de deux modèles gaussiens.

Trois poids différents de l'erreur de mesure dans l'effet de pépite sont considérés (0%, 50% et 100%). Les trois estimations faites sont identiques, cependant l'écart type correspondant varie en fonction de la répartition de l'effet de pépite. En effet, celui-ci est composé de l'erreur de mesure et de la variation à très petite échelle, ces deux parties ayant chacune un poids dans l'effet de pépite. Plus l'erreur de mesure est grande dans l'effet de pépite, et plus l'écart type correspondant est faible.

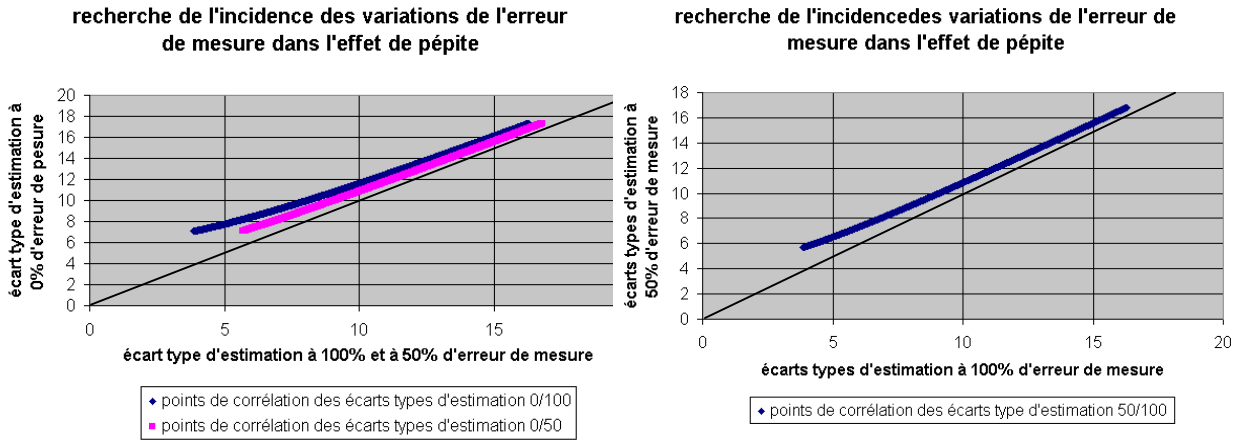


Figure 20 : Nuages de corrélation des écarts types de krigeage pour les différents poids de l'erreur de mesure dans l'effet de pépite

Cet exemple illustre donc bien que le module ne prend pas en compte l'erreur de mesure lors de cette étape.

3.4. QUATRIEME FENETRE : « SEARCHING NEIGHBOURHOOD »

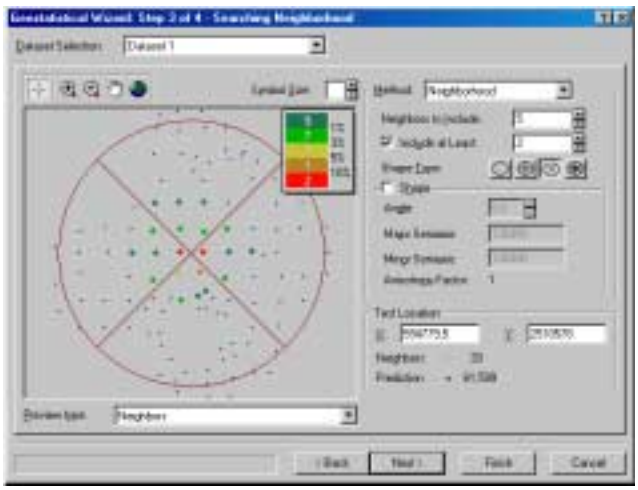


Figure 21 : Fenêtre de dialogue de la quatrième étape : le choix du voisinage

Dans cette étape de la méthode d'interpolation, les paramètres qui régissent le choix du voisinage peuvent être réglés et ajustés par l'utilisateur.

Le choix du voisinage permet de trier parmi les points échantillons. En effet certains points échantillons peuvent être situés si loin du point à estimer qu'il n'existe aucune corrélation entre eux. Prendre en compte ce point dans l'estimation peut alors l'influencer négativement. Limiter le nombre de voisins à prendre en compte permet aussi d'avoir une meilleure vitesse d'exécution.

3.4.1. DESCRIPTION DES RUBRIQUES DE LA FENETRE

Le premier choix qui est donné à l'utilisateur, est le choix du jeu de données à représenter. Ce choix n'est réel que dans le cas du cokrigeage où l'on prend en compte deux jeux de données. Pour les autres méthodes, la seule possibilité est « Dataset 1 ».

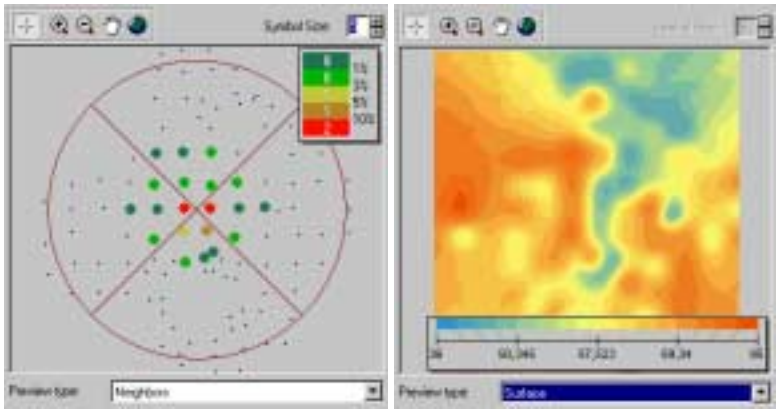


Figure 22 : Les deux possibilités de la zone d'affichage : « voisinage » et « surface »

La commande « Preview type » permet de choisir d'afficher dans la zone d'affichage la limite du voisinage ou bien la surface d'interpolation.

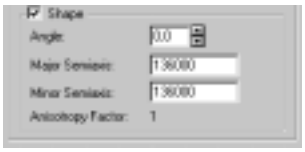
Lorsque la zone d'affichage nous montre la localisation des points échantillons et surligne les points qui font partie du voisinage, les couleurs de surlignage correspondent à l'importance de la valeur absolue du poids accordé au point. Plus celle-ci est forte, plus le point aura d'importance dans l'estimation. La légende correspondante est affichée dans le coin supérieur droit de la zone, celle-ci affiche de plus les effectifs de chaque classe dans les rectangles de couleur.

La zone nous montre aussi la limite et la forme du voisinage. Il est possible de modifier la taille des points surlignés dans « Symbol Size ». On peut ensuite naviguer dans la zone avec les boutons de contrôles situés juste au dessus. Le viseur permet de changer l'origine de la limite de voisinage. On peut ainsi simuler le voisinage des points d'interpolation : on voit quels points échantillons sont pris en compte et leurs poids.

Dans la zone la surface d'interpolation, on peut observer les variations de la surface en fonction des ajustements de paramètres. La légende apparaît au bas de la surface.

Dans la rubrique « Method », on choisit en premier lieu la méthode : voisinage (neighborhood en anglais), la liste ne laissant aucun choix…

Viennent ensuite les principaux paramètres de voisinage, c'est-à-dire le nombre maximal et le nombre minimal de voisins à inclure, puis la répartition des secteurs de la zone de voisinage. «GA» ve donc aller chercher le nombre de voisins spécifié pour chacun des secteurs de la zone de voisinage. S'il n'y a pas assez de points échantillon dans la limite de voisinage, «GA» va aller chercher les points extérieurs les plus proches. Notons que l'on peut supprimer le nombre minimum de voisins en décochant la case correspondante. De cette manière, on ne sort pas des limites de voisinage. Diviser la zone de voisinage en secteurs peut servir pour s'assurer que les points échantillon pris en compte proviennent de toutes les directions.



Par défaut, la zone de voisinage est calquée sur la plus grande portée du modèle ajusté. Ce choix peut se comprendre, en effet, la portée marque la distance maximale de corrélation entre les données.

Au-delà, les données n'étant plus corrélées, il n'est plus nécessaire de la prendre en compte pour l'estimation. Par défaut, «GA» considère toutes les directions comme aussi importantes les unes que les autres, la zone de voisinage est donc un cercle.

Cependant l'utilisateur peut modifier la forme de cette zone en cochant la case de la sous-rubrique « Shape » puis en spécifiant ses dimensions. De plus, lorsque l'on considère une anisotropie, la zone le prend aussi en compte par défaut. En effet, si les données sont mieux corrélées dans une direction que dans une autre, alors il est naturel d'étendre la zone de voisinage dans la direction du grand axe de l'anisotropie. La limite de voisinage dessine alors une ellipse.

La dernière rubrique : « Test Location » renseigne l'utilisateur sur l'estimation en un point de la zone d'estimation.

L'utilisateur n'a qu'à rentrer les coordonnées du point à estimer, soit avec le viseur de la zone d'affichage, soit en remplissant les champs de X et de Y. «GA» affiche alors le nombre de voisins pris en compte (« Neighbors ») puis l'estimation qui est faite pour ce point (« Prediction »).

Le choix du voisinage va influer sur le calcul des valeurs estimées. Car ce choix est en effet le choix des $Z(s_i)$, qui sont les points échantillons qui seront pris en compte dans l'expression de l'estimateur. C'est pour cette raison que le choix du voisinage est important, il ne faut pas sous estimer le nombre de voisins à prendre en compte, mais il ne faut pas non plus le surestimer.

L'utilisateur ne peut pas influer sur la taille la maille pour le calcul des points d'estimation. Or les résultats du krigeage changent avec la taille des blocs à estimer. L'utilisateur n'ayant aucun pouvoir sur celle-ci, les équations de krigeage sont résolues pour un krigeage ponctuel.

3.4.2. APPLICATION AU CAS DE LA POLLUTION PAR L'OZONE

Dans la pratique, on observe un effet d'écran. C'est-à-dire qu'une ou deux couronnes de points autour du point à estimer masquent par leur poids prépondérant l'influence des autres points plus éloignés. C'est ce que l'on remarque sur l'affichage du voisinage : les points les plus extérieurs, qui ne sont situés qu'à deux pas de la maille du point à estimer, n'ont un poids maximum que de trois pourcent. Cet effet permet de réduire le voisinage et par conséquent le nombre d'équations à résoudre.

Dans l'étude LCSQA n°15, le choix du voisinage est un voisinage dit glissant de 50km de rayon. Ceci correspond à ce que l'on vient d'observer : on se contente d'une estimation avec les deux premières auréoles de points (la maille en zone rurale étant de 25km). Pour respecter cette option, nous devons modifier la forme par défaut du voisinage. La limite étant de 136km (portée la plus longue), nous la modifions à 50km, en ce qui concerne le nombre de voisins à prendre en compte, on fixe le maximum à 30 et le minimum à 10 pour un secteur unique. On conserve une limite de voisinage circulaire car on ne prend pas en compte d'anisotropie.

Cependant, dans l'étude LCSQA n°15, c'est l'ajustement du variogramme qui a lieu sur les échantillons de type rural, mais l'estimation se fait en prenant en compte tous les types de sites. De cette manière on peut prendre en compte les valeurs très fortes littorales et les valeurs basses des agglomérations pour les estimations. Ce changement d'échantillon n'est pas possible directement sur «GA». On ne pourra donc pas faire les mêmes estimations que dans l'étude.

Dans cette partie, nous conservons l'échantillon rural pour garder une continuité dans les explications. Les résultats de cette partie seront donc différents de ceux de l'étude LCSQA n°15. Les valeurs extrêmes que l'on peut estimer seront moins fortes que celles de l'étude LCSQA n°15 étant donné qu'on estime sans les valeurs échantillon extrêmes, l'intervalle d'estimation que l'on trouve pour ce cas est [36 ; 95].

Si on fait l'estimation sur tous les points échantillons, avec le modèle ajusté sur les points ruraux, on obtient un intervalle d'estimation de [24 ; 130]. La différence des intervalles d'estimation est flagrante pour les résultats obtenus avec le module. Cependant on remarque que l'intervalle de l'étude LCSQA est [35 ; 99]. Il nous faudra donc valider nos choix par la validation croisée qui est la dernière étape du module.

3.5. CINQUIEME FENETRE : CROSS VALIDATION

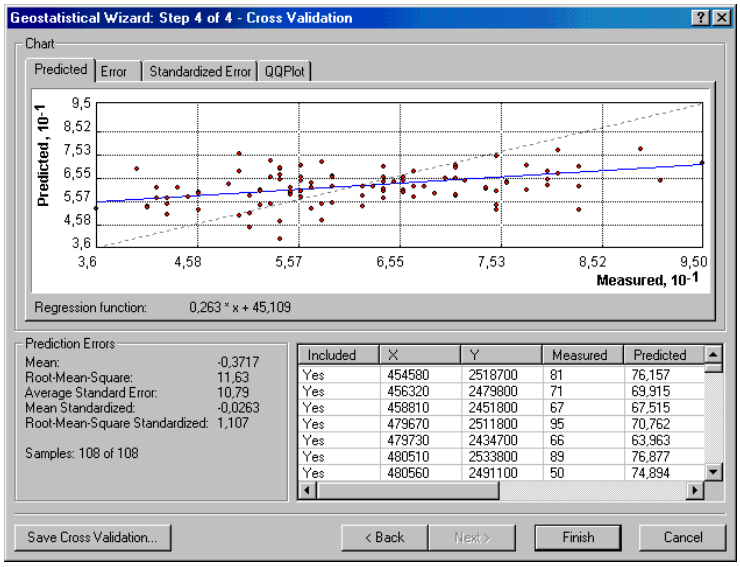


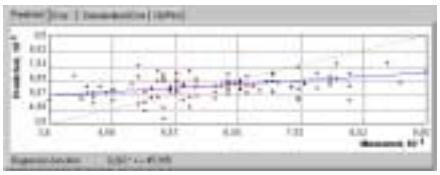
Figure 23 : Fenêtre de récapitulation de la validation croisée

Cette dernière étape sert à valider les choix qui ont été faits dans les étapes précédentes. Il est possible de revenir dans ces étapes avec le bouton « Back », pour changer certaines options et voir si l'estimation est meilleure.

Le principe de la validation croisée est d'estimer aux points exacts des échantillons, sans prendre en compte la valeur de celui sur lequel on se trouve. De cette manière, on peut voir la différence entre la valeur vraie et l'estimation qui en a été faite.

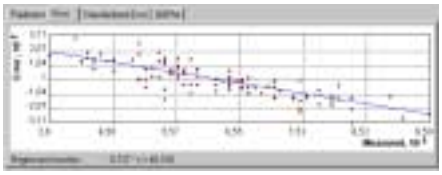
3.5.1. DESCRIPTION DES ELEMENTS DE LA FENETRE

La première rubrique « Chart » récapitule les graphes de différents outils statistiques qui permettent de vérifier la concordance entre valeur vraie et valeur estimée.



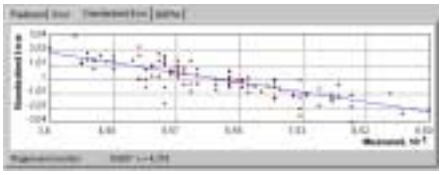
En premier lieu, on a le graphe qui confronte les valeurs vraies des échantillons (en ordonnée) et les valeurs estimées (en abscisse). Il affiche les points de chaque emplacement, la droite d'équation y=x en pointillé et la droite de régression des points en bleu.

Il est facile de voir que meilleure est l'estimation, plus les points sont proches de la droite y=x. Un indicateur de la pertinence du modèle est la forme elliptique du nuage de point et sa forte densité près de cette droite, il faut aussi que le nuage soit dense.



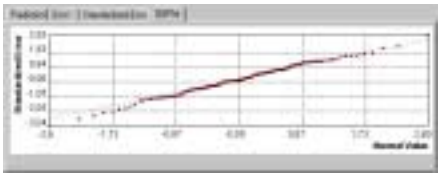
Ensuite, on a le graphique des erreurs. L'erreur commise lors de l'estimation est l'écart entre la valeur estimée et la valeur réelle : $Z^*(s_i) - Z(s_i)$.

Le graphe affiche l'erreur en fonction de la valeur mesurée et la droite de régression correspondante. L'estimation la meilleure sera celle qui minimise les erreurs, on aura donc une espérance conditionnelle la plus proche possible de y=0. on veut donc que les erreurs soient en moyenne proches de zéro.



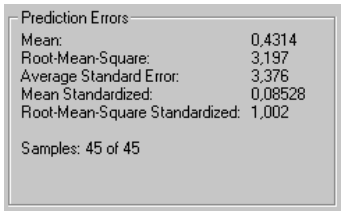
Il y a aussi le graphique de l'erreur réduite, l'erreur est divisée par l'écart type du krigeage : $\left(Z^*(s_i) - Z(s_i)\right) / \sigma_K(s)$.

Le graphe se présente de manière similaire au précédent, et une bonne estimation aura une espérance conditionnelle proche de y=0 en moyenne.



Le dernier graphique de cette rubrique est le graphe « QQPlot » qui met en relation les quantiles de l'erreur réduite (en ordonnée) avec le quantiles correspondants d'une loi normale.

C'est-à-dire que l'on calcule les quantiles de la distribution de l'erreur réduite que l'on compare au quantile correspondant d'une loi normale. Cette opération se fait selon le même principe que pour l'outil d'exploration de données « QQPlot ». Ce graphique nous renseigne donc sur la similitude entre la distribution de l'erreur réduite et une distribution normale. Si en effet, la distribution de l'erreur réduite se rapproche de la distribution d'une loi normale, alors on peut utiliser des méthodes qui requièrent l'hypothèse de la normalité sur l'erreur.



La seconde rubrique « Prediction Errors »est un récapitulatif des statistiques sur l'erreur d'estimation (prediction error dans le module).

La première ligne « Mean » est l'erreur moyenne, en d'autres termes c'est la moyenne des erreurs d'estimation :

$\frac{\sum_i \left(Z^*(s_i) - Z(s_i)\right)}{n}$, une bonne estimation fait tendre ce terme vers 0.

La seconde ligne « Root Mean Square » est l'écart type des erreurs d'estimation :

$\sqrt{\frac{\sum_i \left(Z^*(s_i) - Z(s_i)\right)^2}{n}}$, est à 1 pour une bonne estimation.

Plus ce terme est faible et plus stable est l'estimateur.

La troisième ligne « Average Standard Error » est la moyenne de l'écart type de krigeage, on le traduit en termes mathématiques :

$\frac{\sum_i \sigma_K(s_i)}{n}$, plus l'estimation est précise, et plus l'écart type σ_K est petit, une bonne estimation va donc tendre à minimiser ce terme.

La quatrième ligne « Mean Standardized » est la moyenne de l'erreur réduite :

$\frac{\sum_i \left(Z^*(s_i) - Z(s)\right)/\sigma_K(s)_i}{n}$, comme dans le graphe de l'erreur réduite, une bonne estimation approche ce terme de 0.

La dernière ligne « Root Mean Square Standardized » est l'écart type de l'erreur réduite, ce que l'on écrit :

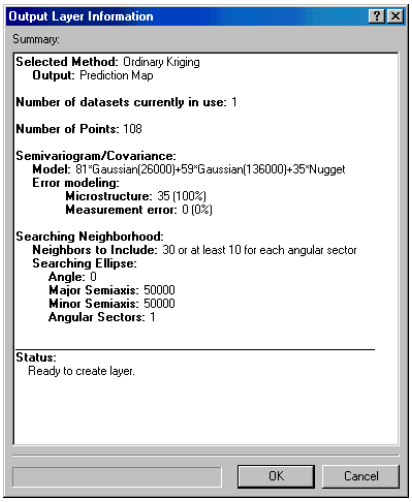
$\sqrt{\frac{\sum_i \left[\left(Z^*(s_i) - Z(s_i)\right)/\sigma_K(s_i)\right]^2}{n}}$, pour une bonne estimation, ce terme tend vers 1

Ces deux derniers termes caractérisent l'adaptation des erreurs expérimentales aux erreurs théoriques.

La dernière partie de la fenêtre est la table de comparaison entre valeur vraie et valeur estimée pour les points échantillons. Elle affiche pour chaque point s'il est inclus dans l'échantillon d'estimation, ses coordonnées, la valeur estimée et la valeur réelle, l'erreur et l'erreur réduite, l'écart type de l'erreur et enfin le quantile normal correspondant.

C'est donc la table récapitulative sur laquelle sont calculées les statistiques, et construits les graphes. Lorsqu'on sélectionne une ligne de cette table, le point correspondant est surligné dans les graphes.

La prédiction a été faite dans les équations de krigeage sans tenir compte de l'erreur de mesure. Dans l'étape de la validation croisée, «GA» va estimer la valeur en tenant compte de l'erreur de mesure. On n'a plus $\sigma^2=0$ dans la variance de krigeage. De cette manière, la variance d'estimation reflète au mieux la variance moyenne de l'erreur.



Cette dernière table peut être sauvegardée en un fichier *.dbf avec le bouton « Save Cross Validation ». Cette étape de la validation croisée est la dernière étape avant la création de la surface interpolée qui sera intégrée à la carte. Le module se termine en effet en cliquant sur « Finish », juste avant la création de la surface, une fenêtre de récapitulation des choix effectués apparaît. En la validant, on lance la génération de la surface d'interpolation.

La surface créée à l'aide du module est intégrée à la carte sous la forme d'une nouvelle couche. On pourra modifier chacune des options excepté la première fenêtre du module. En effet si on clique droit sur la couche dans la table des matières, et que l'on choisit « Method Properties », on relance le module à partir de la seconde fenêtre de dialogue.

3.5.2. APPLICATION AU CAS DE LA POLLUTION PAR L'OZONE

Comme nous l'avons vu dans la description, cette étape ne requiert aucun choix, mais nous donne des indications sur les choix effectués lors des étapes précédentes. Nous allons donc voir les statistiques pour les choix qui ont été faits.

Dans cette première partie, nous nous sommes contentés de reproduire les choix de l'étude LCSQA n°15. Voyons les statistiques de la validation croisée, puis la surface créée :

Statistiques sur l'ensemble de l'échantillon :

- Moyenne de l'erreur (Mean) : -0,3717
- Ecart type de l'erreur (Root-Mean-Square) : 11,63
- Moyenne de l'écart type de krigeage (Average Standard Error) : 10,79
- Moyenne de l'erreur réduite (Mean Standardized) : -0,0263
- Ecart type de l'erreur réduite (Root-Mean-Square Standardized) : 1,107

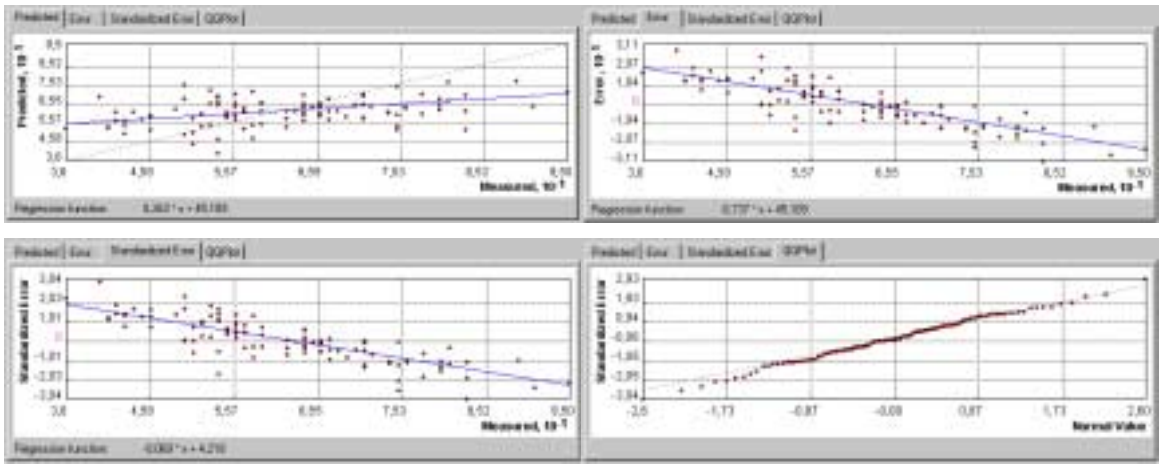


Figure 24 : Présentation des graphiques de la validation croisée pour le cas de l'ozone

On voit donc globalement que l'écart type de l'erreur réduite se rapproche bien de 1 la précision de l'estimateur est donc assez bonne. Notre estimateur est également peu biaisé car les moyennes de l'erreur réduite et de l'erreur sont faibles.

On présente ci-dessous la surface résultante des étapes et des choix de paramètres que nous avons fait tout au long de cette partie. Rappelons nous que cette carte est une carte des valeurs estimées. Ces données ont la variance de l'erreur d'estimation minimale.

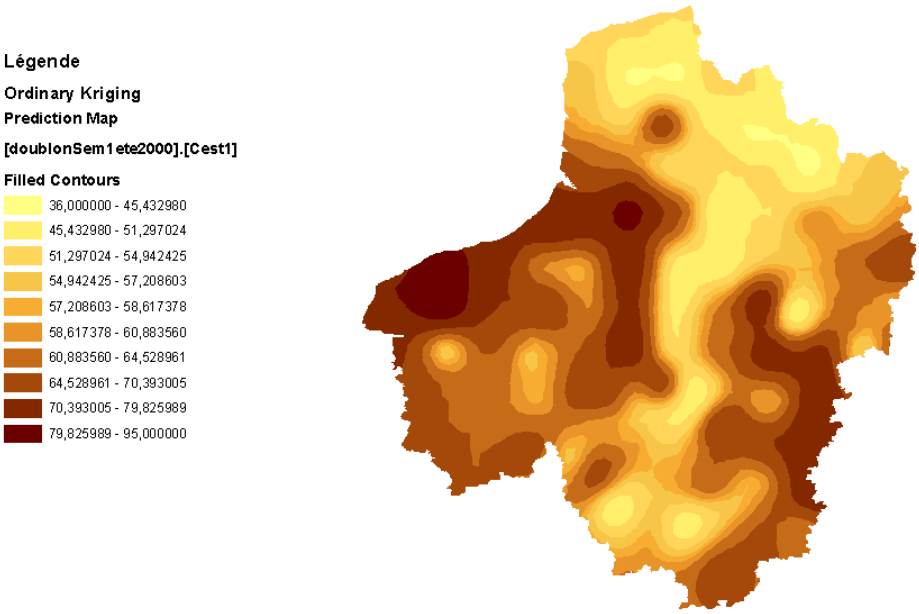


Figure 25 : Présentation de la surface créée avec les données sur la pollution de l'air par l'ozone pour la première semaine de la campagne de mesure de l'année 2000, estimation sur l'échantillon rural

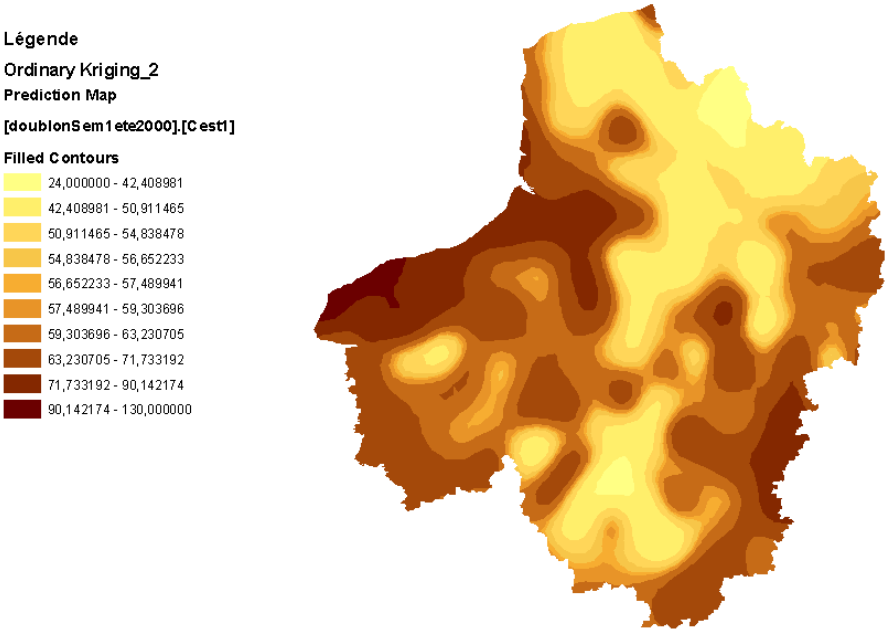


Figure 26 : La même étude, avec une estimation sur l'échantillon complet

Dans la partie suivante, nous allons modifier les ajustements des paramètres, puis nous comparerons les validations croisées afin de choisir la meilleure interpolation.

4. COMPARAISON DES METHODES DE KRIGEAGE

Cette partie traite de l'incidence des changements de paramètres pour tenter de trouver l'ajustement le plus juste et cohérent possible. Nous allons tout de même conserver certains paramètres de l'étude LCSQA n°15 qui gardent toute leur cohérence malgré les changements opérés.

On conserve les modèles paraboliques à l'origine (cubique et gaussien) car ils ont été adaptés en relation avec la nature de la pollution. En effet ils ont été choisis car la pollution par l'ozone est un phénomène à grande échelle qui varie peu à petite distance. Cette caractéristique étant propre au polluant, nous ne changerons pas de modèle.

On garde aussi la taille de la zone d'ajustement. Là aussi c'est un paramètre lié à la maille et à la taille du champ d'étude. Cette dernière étant d'environ 315km, on se limite à un ajustement sur les 200 premiers kilomètres.

On n'utilise pas de transformation Box-Cox, Arcsin ou logarithme, car nous n'allons pas utiliser d'option qui nécessite l'hypothèse de la distribution gaussienne de la variable.

L'effet de pépite sera toujours égal à 35 $(\mu g/m^3)^2$, étant donné que c'est la variance de l'erreur de mesure et que celle-ci est propre à l'échantillonnage.

Par contre, en relation avec les observations faites au long des étapes, nous allons explorer anisotropie et dérive, et donc en conséquence, les possibilités offertes par le krigeage universel. Et nous verrons aussi les différences d'échantillons à prendre en compte dans l'estimation.

Les tableaux suivants récapitulent les modèles ajustés pour effectuer nos comparaisons :

Krigeage Ordinaire (KO)						
échantillon rural		échantillon complet sans les données littorales			échantillon complet	
isotrope	Anisotrope	isotrope	isotrope	anisotrope	dérive externe	isotrope
modèle étude LCSQA	Modèle ajusté	modèle étude LCSQA	modèle ajusté	modèle ajusté	modèle ajusté	modèle étude LCSQA
cas n°1	cas n°2	cas n°3a	cas n°3b	cas n°4	cas n°5	cas n°6

Krigeage Simple (KS)						
échantillon rural			échantillon complet sans les données littorales			échantillon complet
isotrope	isotrope	anisotrope	isotrope	isotrope	anisotrope	isotrope
modèle étude LCSQA	modèle ajusté	modèle ajusté	modèle étude LCSQA	modèle ajusté	modèle ajusté	modèle étude LCSQA
cas n°7a	cas n°7b	cas n°8	cas n°9a	cas n°9b	cas n°10	cas n°11

Krigeage Universel (KU)			
échantillon rural	échantillon complet sans les données littorales		échantillon complet
isotrope	isotrope	isotrope	isotrope
modèle ajusté	modèle ajusté	modèle ajusté	modèle ajusté
cas n°12	cas n°13a	cas n°13b	cas n°14

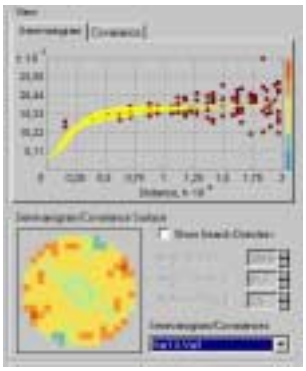
Tableau 2 : Récapitulatif des cas traités pour comparer les validations croisées

4.1. KRIGEAGE ORDINAIRE

4.1.1. CAS N°1 : KO, ECHANTILLON RURAL, ISOTROPE

C'est le cas qui suit les indications de l'étude LCSQA n°15. Ses résultats et sa description figurent dans la partie précédente.

4.1.2. CAS N°2 : KO, ECHANTILLON RURAL, ANISOTROPE



Pour une taille de pas et un nombre de pas identiques au cas n°1, nous avons réajusté le modèle :

81 * Gaussien (45000;27000;304,0) + 59 * Gaussien (136000; 62000;19,1) + 35 * Pépite

Nous avons conservé les paliers du cas n°1, mais en modélisant l'anisotropie, nous avons modifié les portées.

On vérifie si ce modèle est bien adapté en cochant la case « Show Searh Direction », puis pour chaque direction, nous regardons si le modèle s'ajuste bien.

On examine la surface variographique pour ajuster au mieux l'anisotropie. L'anisotropie de faible portée suit la direction du premier ovale des cases vertes, tandis que la seconde anisotropie (de grande portée) est dirigée vers la « faille » de la couronne extérieure formée de cases vertes et bleues.

On observe dans la direction du petit axe de la seconde anisotropie que le modèle s'ajuste moins bien, on modifie donc notre modèle pour obtenir :

81*Gaussien(50000;27000;304,0)+59*Gaussien(136000;87000;22,3)+35*Pépite

En ce qui concerne le choix du voisinage, à cause de l'effet d'écran, on va garder une limite de l'ordre de 50km pour le grand axe et le petit axe de la zone de voisinage. On va prendre en compte la petite anisotropie en choisissant une forme elliptique pour le voisinage. On ne peut pas prendre en compte la grande anisotropie à cause de l'effet d'écran. On va utiliser le rapport « r » de la petite anisotropie pour la limite de voisinage : 1,85. On prend finalement un voisinage (64750 ; 35000) à secteur unique avec au maximum 20 voisins et 7 au minimum.

On répertorie les statistiques de la validation croisée dans un tableau de comparaison.

4.1.3. CAS N°3 : KO, ECHANTILLON SANS LES DONNEES LITTORALES, ISOTROPE

Dans ce cas là, nous considérons un échantillon de tous les types de données, mis à part les données littorales. L'intérêt principal de cette configuration de paramètres est d'avoir une estimation qui prenne en compte les valeurs des agglomérations. Nous distinguerons deux cas :

a) nous reprenons le cas n°1, avec 50 voisins au maximum.

b) nous allons ensuite ajuster notre propre modèle :

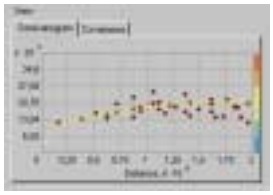
95*Gaussien(26000)+85*Gaussien(156000)+35*Pépite

On a ajusté les paliers et agrandi de 20km la grande portée. Pour le voisinage, on garde la même configuration étant donné que l'effet d'écran reste le même en zone rurale. Pour l'ajustement, nous nous sommes appuyés sur la correspondance entre le modèle théorique et le nuage variographique.

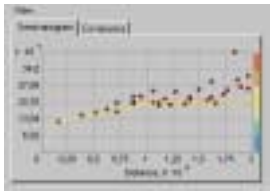
Nous verrons ainsi si l'estimation en prenant en compte les données urbaines et périurbaines est plus précise dans les « trous » des données rurales (qui sont les agglomérations).

4.1.4. CAS N°4 : KO, ECHANTILLON SANS LES DONNEES LITTORALES, ANISOTROPE

Nous allons étudier les effets de l'anisotropie pour un autre échantillon que les données rurales, le nouvel échantillon est composé des données rurales, urbaines et péri urbaines.



Nous pouvons cependant observer que l'anisotropie que l'on tente de modéliser se rapporte plus à une anisotropie zonale (changement de palier), comme illustré ci contre, qu'à une anisotropie géométrique.



Cette dernière n'est cependant pas absente, et c'est pourquoi on la prend en compte. Ces observations nous mènent à l'ajustement du modèle :

95*Gaussien(26000;18000;337,0)+85*Gaussien(160000;120000;18,0)+35*Pépite

Le voisinage choisi, comme dans le cas n°2, va reprendre l'anisotropie à petite portée, c'est-à-dire que nous allons respecter le rapport des axes : 1,44, avec des axes de longueur environ 50km. Nous avons donc une limite de voisinage à secteur unique et d'axes (56000 ; 38500) dans laquelle nous incluons 50 voisins au maximum et 20 au minimum. Nous déterminons les quotas de voisins de manière à ce que tous les voisins situés dans la zone soient pris en compte.

4.1.5. CAS N°5 : KO, ECHANTILLON COMPLET, DERIVE EXTERNE

Dans la partie d'exploration des données, et plus particulièrement dans l'outil d'analyse de la dérive, nous avions remarqué que les valeurs littorales avaient une influence sur la répartition de données. De plus l'étude LCSQA n°15 a mis en évidence leur influence sur le nuage variographique, nous avons pu le vérifier. Dans ce cas, nous allons continuer à tenter de voir si la prise en compte de toutes les données pour l'estimation peut l'améliorer.

Pour ce faire, nous allons adapter un modèle et prendre en compte la dérive due aux données littorales. On sélectionne une dérive du second ordre.

Les modèles dont nous disposons nécessitent une variable stationnaire. Or nous avons vu que la nuée variographique n'atteint pas de seuil si on ne prend pas en compte la dérive. C'est pourquoi nous allons prendre en compte la dérive. Or nous avons expliqué que cette méthode n'est, a priori, pas correcte. Avec ce cas, nous allons voir si les résultats obtenus s'éloignent beaucoup des autres méthodes.

Nous choisissons une dérive globale car nous avons vu dans l'exploration de données que l'influence des fortes valeurs se fait ressentir sur toute la zone d'étude. C'est par ce choix que «GA» peut ensuite calculer les coefficients du polynôme de la dérive.

Nous ajustons ensuite nos modèles variographiques :

90*Gaussien(25000)+50*Gaussien(134000)+35*Pépite

Les paramètres du voisinage sont toujours une limite circulaire de rayon 50km à secteur unique, avec 50 voisins au maximum et 15 au minimum.

On reporte les statistiques de la validation croisée dans le tableau comparatif.

4.1.6. CAS N°6 : KO, ECHANTILLON COMPLET, ISOTROPE

Nous allons conserver le modèle choisi dans l'étude LCSQA n°15, le variogramme théorique sera identique au cas n°1, nous allons juste voir si la prise en compte de l'échantillon au complet permet d'obtenir de meilleurs résultats pour l'estimation.

Nous n'allons pas modéliser cette situation avec une anisotropie car la nuée de points avec l'échantillon complet ne se stabilise pas à un palier.

4.2. KRIGEAGE SIMPLE

4.2.1. CAS N°7 : KS, ECHANTILLON RURAL, ISOTROPE

Dans la fenêtre du choix de la méthode de krigeage et du type de carte à effectuer, nous choisissons krigeage simple et nous attribuons à la variable une moyenne nulle comme dans l'étude LCSQA. La valeur proposée par défaut par le module est la moyenne de l'échantillon. L'étude LCSQA montre que les résultats obtenus avec les deux méthodes krigeage sont très semblables.

a) Dans le cas présent, nous allons utiliser le modèle ajusté lors de l'étude LCSQA, avec un voisinage similaire à celui utilisé dans les cas précédents, 25 voisins au maximum et 10 au minimum. On reporte les résultats de la validation croisée dans le tableau final.
b) Ce cas est identique au précédent, excepté pour la moyenne qui celle de l'échantillon : 61,278 µg/m^3. Les statistiques de la validation croisée sont bien meilleures que le cas précédent.

On remarque que la distinction entre les cas de krigeage simple et de krigeage ordinaire se fait sur le krigeage, non sur le variogramme, nous allons donc pouvoir réutiliser certains variogrammes pour tester l'influence de la méthode de krigeage.

4.2.2. CAS N°8 : KS, ECHANTILLON RURAL, ANISOTROPE

Nous récupérons le modèle du cas n°2, avec la moyenne de la variable identique à celle de l'échantillon. Les paramètres de voisinage sont aussi les mêmes. Nous conservons la moyenne de l'échantillon comme valeur moyenne de la variable.

4.2.3. CAS N°9 : KS, ECHANTILLON SANS LES DONNEES LITTORALES, ISOTROPE

La séparation du cas n°3 est ici aussi répétée, nous respectons à nouveau les modèles et le voisinage.

a) modèle de l'étude LCSQA
b) modèle ajusté sous «GA»

Nous considérons toujours la moyenne de la variable comme la moyenne des données : 56,14µg/m^3.

4.2.4. CAS N°10 : KS, ECHANTILLON SANS LES DONNEES LITTORALES, ANISOTROPE

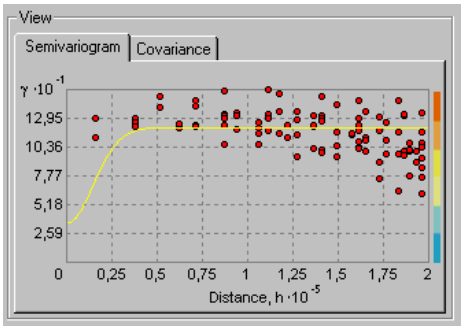
Dans ce cas de figure également, on ne modifie pas les paramètres du cas n°4, si ce n'est la méthode krigeage et la moyenne de la variable qui est la moyenne de l'échantillon.

4.2.5. CAS N°11 : KS, ECHANTILLON COMPLET, ISOTROPE

Nous procédons comme dans le cas n°6. Nous ne pouvons pas prendre en compte de dérive externe car nous sommes en krigeage simple. La moyenne est maintenant de 57,703µg/m^3.

4.3. KRIGEAGE UNIVERSEL

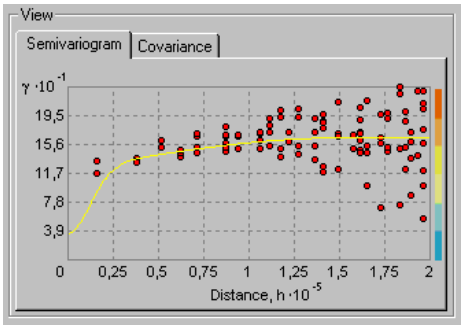
4.3.1. CAS N°12 : KU, ECHANTILLON RURAL, ISOTROPE



Nous avions remarqué qu'une courbe de degré deux s'adaptait encore aux données rurales dans la phase d'exploration des données. Cependant, si on sélectionne une valeur moyenne de degré deux à prendre en compte dans le krigeage universel, le variogramme résultant est difficilement ajustable car la variance décroît avec la distance.

Il devient difficile d'y ajuster un modèle avec l'effet de pépite et un modèle gaussien. Nous choisissons alors une fonction de degré un pour modéliser cette dérive.

Nous avons choisi une échelle globale pour la dérive, comme pour les mêmes raisons que dans le cas n°5.



On ne modélise pas d'anisotropie pour ce cas là car on voit que le nuage variographique s'évase vers les grandes distances, signe d'une anisotropie zonale que l'on ne peut prendre en compte dans le module.

On adapte alors le variogramme :

86*Gaussien(30000)+44*Sphérique(150000)+35*Pépite

Pour le voisinage, on reprend les paramètres usuels, voisinage unique de limite circulaire, de rayon 50km avec 30 voisins au maximum et 10 au minimum.

On reporte les statistiques dans le tableau.

4.3.2. CAS N°13 : KU, ECHANTILLON COMPLET SANS LES DONNEES LITTORALES, ISOTROPE

Nous allons maintenant prendre en compte en plus des données rurales, les données urbaines et périurbaines. Nous allons donc réajuster notre modèle de krigeage universel :

Nous verrons deux cas différents :

a) avec une dérive de degré un, nous ajustons : 98*Gaussien(28000)+35*Pépite

b) avec une dérive de degré deux, nous ajustons :

88*Gaussien(28000)+59*Sphérique(129900)+35*Pépite

On reprend le voisinage standard avec un maximum de voisins à 50 et un minimum à 20.

4.3.3. CAS N°14 : KU, ECHANTILLON COMPLET, ISOTROPE

Nous prenons maintenant toutes les données en compte. Et nous ajustons une nouvelle fois notre modèle de variogramme :

95*Gaussien(33000)+25*Gaussien(120000)+35*Pépite

Nous prenons le même voisinage que dans le cas précédent et on reporte les statistiques de la validation croisée dans le tableau final.

4.4. REMARQUES SUR LA VALIDATION CROISEE DES RESULTATS

Dans cette partie, nous allons rapidement discuter les résultats des validations croisées, qui sont récapitulés dans le tableau final fourni en fin du chapitre.

Dans les statistiques présentées à l'étape de la validation croisée, on s'intéresse principalement à la moyenne des erreurs, à l'écart type des erreurs, à la moyenne des écarts types de krigeage, et à l'écart type de l'erreur réduite. On distingue deux objectifs dans ces statistiques : caractériser la précision de l'estimateur et voir son adaptation à la réalité.

La précision de l'estimateur est caractérisée par la moyenne des erreurs qui indique s'il est bien ciblé. En effet, plus la moyenne tend vers zéro et plus les erreurs sont faibles, donc l'estimateur bien ciblé.

L'écart type de l'erreur et la moyenne de l'écart type de krigeage renseignent sur la plage de variation de l'estimateur. On cherche à minimiser ces valeurs pour que l'estimateur varie le moins possible, de cette manière, on cherche à brider les erreurs.

En effet, si on compare la variable à une série de tirs sur une cible. La moyenne indique si les tirs sont globalement biens dirigés vers la cible. Pour une bonne moyenne, on aura donc des tirs tout autour de la cible.

L'écart type indique si les tirs sont groupés. Plus il est petit, et plus les tirs sont groupés. Une bonne série de tirs sera groupée et bien centrée sur la cible. On cherche à faire de même avec l'estimateur.

Les estimateurs qui vérifient le mieux ces caractéristiques sont ceux des cas n°12 et 13a, c'est-à-dire des estimateurs issus du krigeage universel avec une dérive de degré 1. Le premier est mieux centré tandis que le second est mieux groupé.

L'écart type de l'erreur réduite nous renseigne quant à lui sur la correspondance entre l'écart type de l'erreur expérimentale, et l'écart type de l'erreur théorique (écart type de krigeage). Cette statistique est en effet le rapport entre les deux. Lorsque cette statistique tend vers 1, alors les types d'erreurs correspondent bien. L'erreur estimée correspond alors à l'erreur observée.

Cependant, si l'écart type de l'erreur réduite est inférieur à 1, l'erreur estimée est supérieure à l'erreur observée. C'est-à-dire que l'estimateur surestime l'erreur. Dans le cas contraire, on la sous-estime.

Les estimateurs qui font correspondre le mieux les erreurs sont ceux des cas n°1, 7b et 12. Ils correspondent tous trois à trois krigeages différents. Et tous trois sont des estimateurs de l'échantillon rural. C'est donc pour cet échantillon que l'on a le mieux réussi à faire correspondre les erreurs observées et estimées.

Maintenant, si on regarde quelles sont les meilleures statistiques, c'est le modèle du cas n°12 qui les obtient. C'est donc lui qui est le mieux ajusté et le mieux ciblé par rapport à l'échantillon.

Dans les cas présentés, nous avons aussi tenté de voir l'influence de l'échantillon. En effet, nous avons estimé selon les paramètres de l'étude LCSQA n°15 l'échantillon rural seulement, puis l'échantillon des données rurales, urbaines et périurbaines, et enfin l'échantillon complet (cas n°1, 3 et 6).

La moyenne de l'erreur est croissante, avec la taille de l'échantillon, en revanche, la variance de krigeage diminue. La moyenne de l'erreur réduite augmente elle aussi avec la taille de l'échantillon, mais de manière moins forte.

La variance de l'erreur réduite augmente également. L'estimateur s'adapte donc moins bien à l'échantillon complet qu'à l'échantillon rural.

Observons les nuages de corrélation :

Les valeurs estimées sont en abscisse et les valeurs mesurées en ordonnées.

Le nuage augmente de surface avec la taille de l'échantillon. C'est-à-dire que l'on retrouve l'observation faite sur les statistiques : l'erreur moyenne augmente. Cependant, on voit que les valeurs extrêmes sont mieux estimées lorsque l'échantillon prend en compte toutes les données.

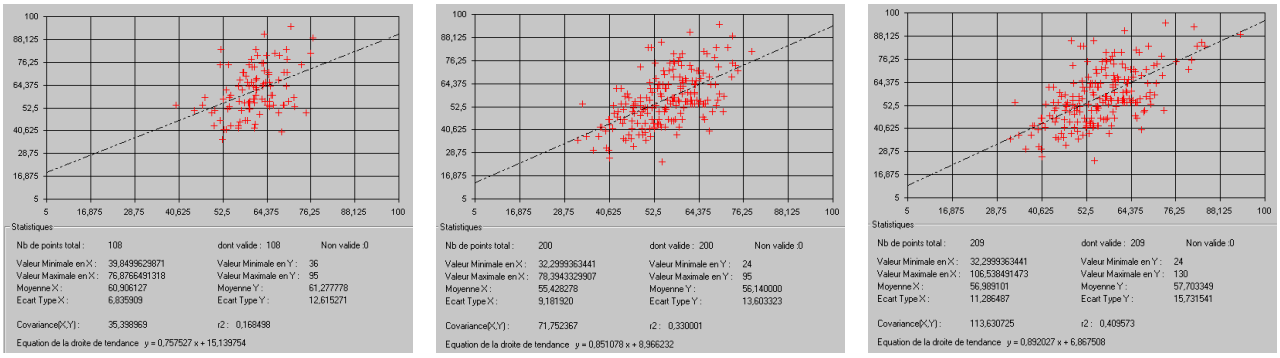


Figure 27 : Nuages de corrélation entre valeur estimée et valeur vraie pour l'échantillon rural, l'échantillon complet sans les données littorales et l'échantillon complet respectivement

La validation croisée nous a donc permis de sélectionner parmi un ensemble d'estimateurs, l'estimateur qui est le plus adapté à notre cas. Malgré nos observations dans l'outil « trend analysis », la meilleure estimation prend en compte une dérive de degré 1. Et ne prend en compte que l'échantillon rural, en effet, c'est l'échantillon le plus homogène des trois considérés.

n° du cas	méthode utilisée	statistiques sur l'erreur				
		erreur moyenne (0)	écart type de l'erreur (1)	moyenne de l'écart type de krigeage (min)	moyenne de l'erreur réduite (0)	écart type de l'erreur réduite (1)
cas n°1	KO, éch rural, modèle étude LCSQA	-0,3717	11,63	10,79	-0,0263	1,107
cas n°2	KO, éch rural, anisotrope	-0,4296	12,5	10,03	-0,03002	1,32
cas n°3a	KO, éch tout sauf littoral, modèle étude LCSQA	-0,643	11,23	9,251	-0,04247	1,313
cas n°3b	KO, éch tout sauf littoral, isotrope	-0,6566	11,28	9,624	-0,04137	1,289
cas n°4	KO, éch tout sauf littoral, anisotrope	-0,5831	11,45	10,08	-0,03472	1,256
cas n°5	KO, éch complet, dérive externe	-0,7098	12,18	9,499	-0,04836	1,385
cas n°6	KO, éch complet, modèle étude LCSQA	-0,7142	12,17	9,173	-0,05028	1,417
cas n°7a	KS, éch rural, modèle étude LCSQA	-6,507	14,43	10,75	-0,5779	1,338
cas n°7b	KS, éch rural, isotrope	-0,4363	11,65	10,75	-0,03171	1,111
cas n°8	KS, éch rural, anisotrope	-0,5125	12,46	9,962	-0,03665	1,32
cas n°9a	KS, éch tout sauf littoral, modèle étude LCSQA	-0,9171	11,35	9,229	-0,06838	1,323
cas n°9b	KS, éch tout sauf littoral, isotrope	-0,8814	11,38	9,607	-0,06195	1,297
cas n°10	KS, éch tout sauf littoral, anisotrope	-0,6847	11,47	10,07	-0,0432	1,258
cas n°11	KS, éch complet, modèle étude LCSQA	-0,9287	12,29	9,168	-0,07177	1,428
cas n°12	KU, éch rural, isotrope, dérive degré 1	0,1326	11,82	11,23	0,01783	1,107
Cas n°13a	KU, éch tout sauf littoral, isotrope, dérive degré 1	-0,2884	11,4	9,733	-0,01041	1,285
Cas n°13b	KU, éch tout sauf littoral, isotrope, dérive degré 2	-0,2515	12,15	10,12	-0,00818	1,336
cas n°14	KU, éch complet, isotrope, dérive degré 2	-0,4733	12,79	9,326	-0,02369	1,472

Tableau 3 : récapitulatif des changements de paramètres :

5. UTILISATION DU COKRIGEAGE :

Dans cette partie, nous allons appliquer la méthode du cokrigeage aux données de la pollution de l'air par le NO_2 dans les environs de Bourg-en-Bresse en 2001. Ces données sont fournies par l'association Air2savoie. Ces données correspondent à des relevés pendant quatre semaines en hiver puis quatre nouvelles semaines en été.

Les ajustements déjà effectués dans le rapport LCSQA 2003 (analyse effectué sous le logiciel Isatis®, specialisé en géostatistique) sont à nouveau réutilisés pour comparer les résultats obtenus et les possibilités offertes par «GA».

Cependant, l'étude sous Isatis® a pu considérer la densité de population comme dérive externe. Cette dérive est prise en compte avec une fonction auxiliaire qui peut être précisée. Or le module «GA» ne laisse pas cette possibilité. Nous avons vu en effet que la dérive prise en compte par le module est une dérive dont la forme est un polynôme et dont les coefficients doivent être estimés.

Après un rapide examen des données, nous voyons que nous pouvons considérer une dérive parabolique pour chacune des variables de pollution. Cependant une étude comparative des nuages de points de la validation croisée nous indique plutôt de suivre les ajustements effectués par défaut par le module. Un cokrigeage ordinaire est réalisé sans considérer la dérive sur ces données.

Les données indisponibles ont une valeur nulle, elle ne sont pas considérées dans l'étude. Nous les faisons correspondre à « NODATA Value ».

Les paramètres choisis sont un modèle sphérique de portée 3607m, qui est plus variable à petite distance qu'un modèle gaussien. Son palier est de 26 pour la moyenne hivernale, de 5 pour la moyenne estivale et de 11 pour le covariogramme croisé. L'effet de pépite associé est de 11 pour la moyenne estivale et de 14 pour la moyenne hivernale. On remarque qu'il n'est pas possible d'affecter un effet de pépite au covariogramme croisé.

La taille du pas utilisée est de 615m et le nuage de points est dessiné sur une distance maximale de 10 pas. On présente ci-dessous l'allure des modèles ajustés.

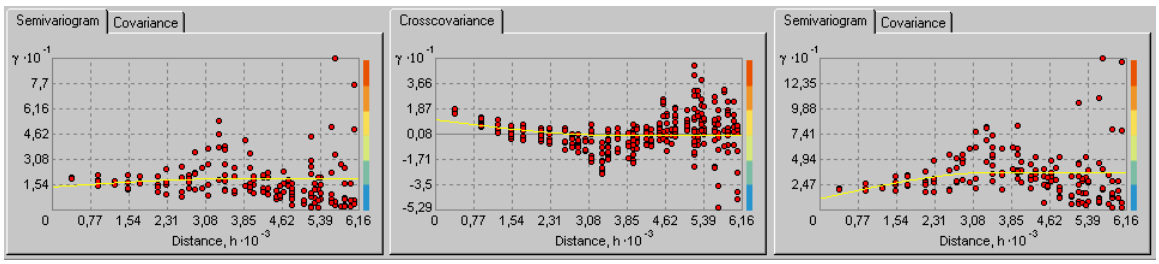
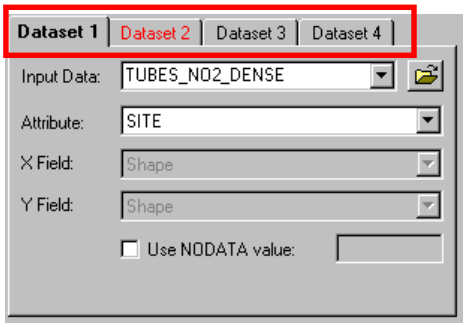


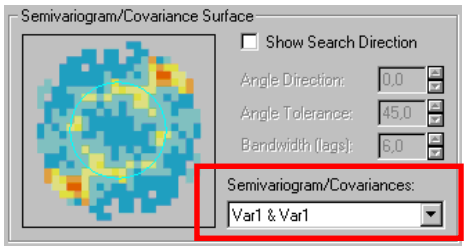
Figure 28 : Courbe des modèles de la moyenne estivale, du variogramme croisé et de la moyenne hivernale

On considère, pour réaliser les cartes, un voisinage unique que l'on traduit par une limite à 10000m et à secteur unique.

5.1. LES ETAPES A SUIVRE DANS LE MODULE :



Les étapes de la création d'une carte de cokrigeage sous «GA» sont les mêmes que celles pour créer une carte de krigeage. Il faut cependant préciser les autres variables à prendre en compte dans la fenêtre « Chose Input Data and Method ». Dans ce module, on peut choisir jusqu'à quatre jeux de données différents, mais seul le premier peut être considéré comme la variable principale de cokrigeage.



Il faudra ensuite préciser les paramètres de la matrice de variance/covariance dans la fenêtre « Semivariogram/Covariance Modeling ». Pour ce faire, on choisit le terme qui nous intéresse dans la matrice avec la liste « Semivariogram/Covariance ». C'est cette liste qui permet de passer à chacun des ajustements de modèle.

Le nombre de termes de la matrice de variance/covariance est égal au carré du nombre de variables prises en compte. Considérer trop de variables différentes peut donc rapidement conduire à un travail fastidieux d'ajustement et de calcul des nuées de points.

Pour ce qui concerne la limite de voisinage, il faudra préciser celle-ci pour chacun des jeux de données dans la fenêtre suivante. La suite, qui est l'étape de la validation croisée, se déroule comme pour le krigeage.

5.2. RESULTATS DU COKRIGEAGE SUR LES DONNEES DU NO_2 :

Nous allons examiner les statistiques de la validation croisée du cokrigeage et les comparer à celles du krigeage des moyennes estivales et hivernales.

variable	méthode	statistiques sur l'erreur			
		erreur moyenne (0)	variance de l'erreur (1)	variance de l'erreur réduite (1)	variance de l'erreur relative
Hmoy	krigeage ordinaire	0,1623	11,1798	0,5704	0,0111
	cokrigeage ordinaire	0,0223	8,0261	0,4492	0,0072
Emoy	krigeage ordinaire	0,0775	10,52	0,6365	0,0451
	cokrigeage ordinaire	0,0208	7,7184	0,5067	0,0283

Tableau 4:Récapitulatif des statistiques de krigeage et de cokrigeage des données de la pollution

Les statistiques du cokrigeage sont meilleures aussi bien pour la moyenne que pour la variance. Ceci confirme que l'estimateur par cokrigeage est plus précis et moins biaisé.

Cependant la variance de l'erreur réduite est moins forte pour le cokrigeage que pour le krigeage. Ceci nous montre que dans le cas du krigeage, les erreurs théoriques sont plus proches des erreurs expérimentales.

Lorsque l'on observe les nuages de points correspondants, on observe une meilleure corrélation pour le cokrigeage. Cette observation est moins nette que dans l'étude effectuée sous Isatis® avec la prise en compte de la dérive externe. Mais on voit quand même un nuage plus groupé et une meilleure estimation des valeurs extrêmes dans le cas du cokrigeage.

Nous avons placé en ordonnée les valeurs vraies et en abscisse les valeurs estimées.

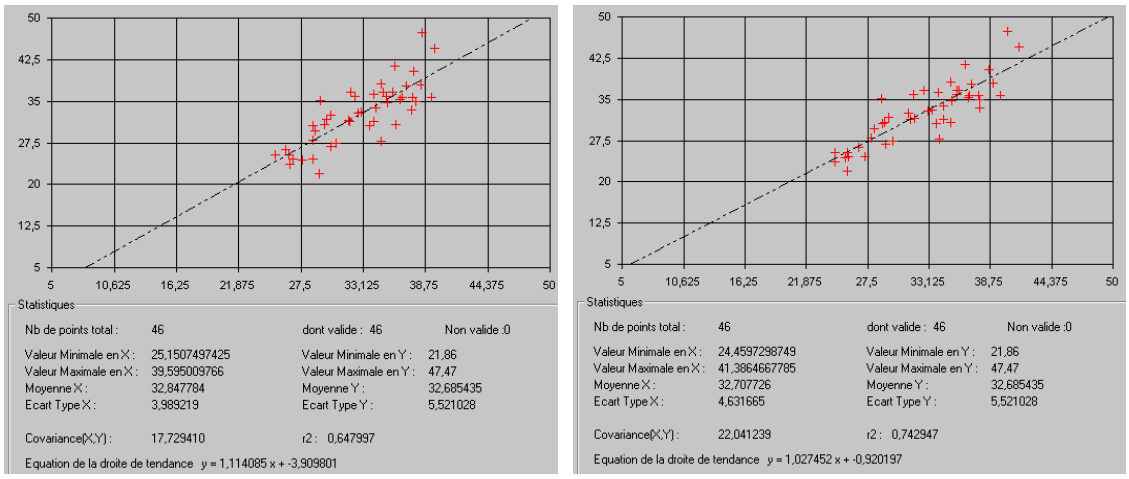


Figure 29 : Nuage de corrélation du krigeage et du cokrigeage sur la moyenne hivernale

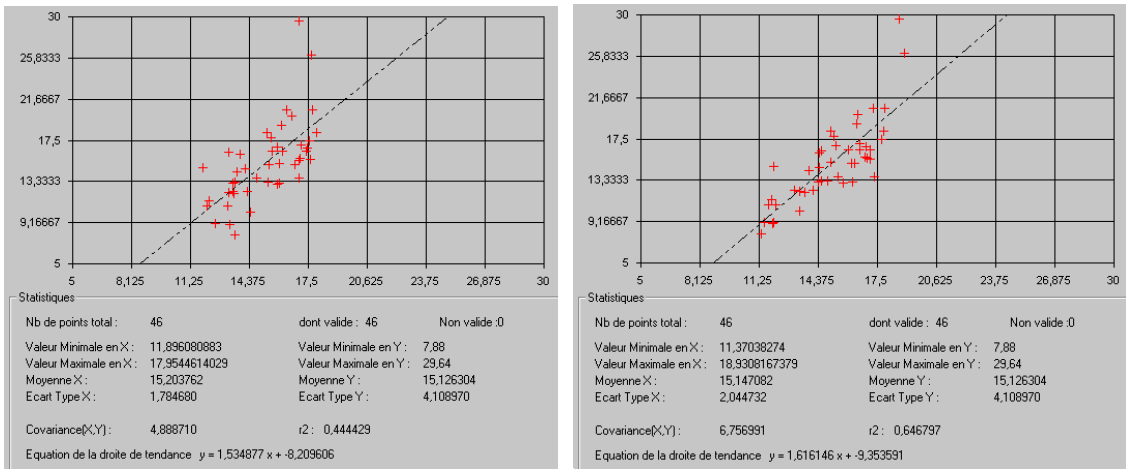


Figure 30 : Nuage de corrélation du krigeage et du cokrigeage sur la moyenne estivale

Les cartes du cokrigeage des moyennes hivernales et estivales, puis les cartes des krigeages sont présentées ci-après :

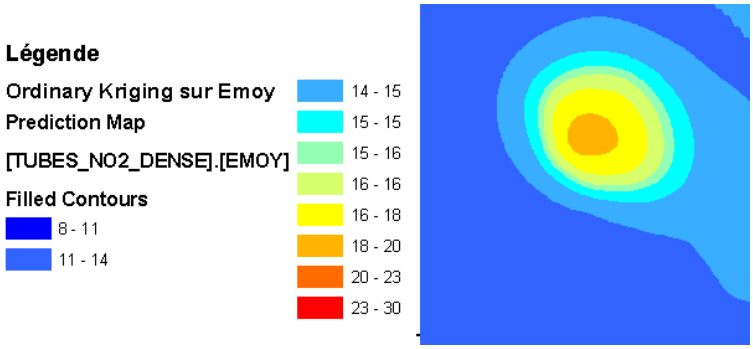


Figure 31 : Krigeage Ordinaire de la moyenne estivale

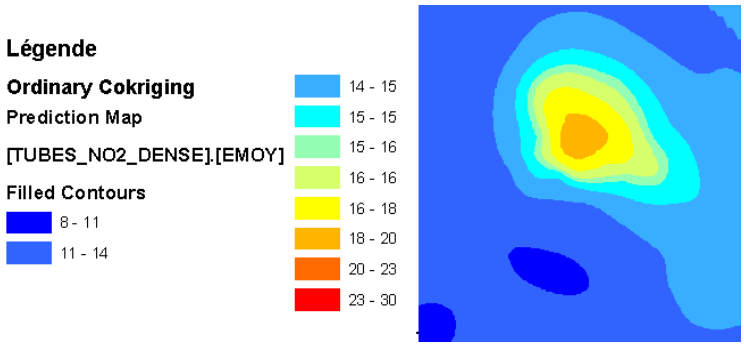


Figure 32 : Cokrigeage Ordinaire de la moyenne estivale avec la moyenne hivernale

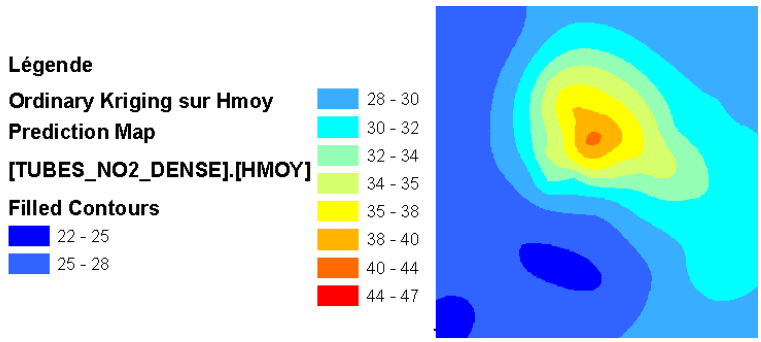


Figure 33 : Krigeage Ordinaire de la moyenne hivernale

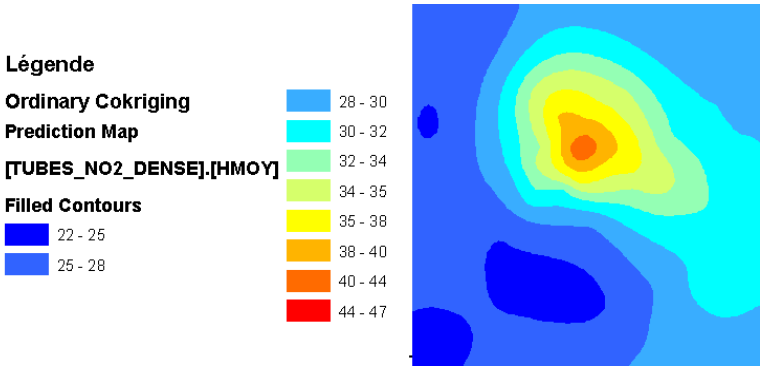
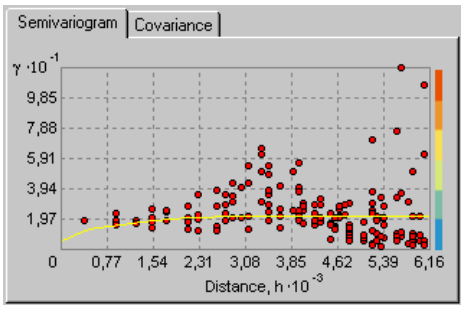


Figure 34 : Cokrigeage Ordinaire de la moyenne hivernale avec la moyenne estivale

5.3. KRIGEAGE DE LA MOYENNE ANNUELLE :

On s'intéresse maintenant aux résultats de la moyenne annuelle. Les résultats de l'estimation annuelle entre la moyenne des cokrigeages et le krigeage de la moyenne annuelle seront comparés.

Etant donnée que nous ne pouvons pas utiliser de dérive externe, nous examinons les nuages de corrélation de plusieurs adaptations, on retient un krigeage ordinaire de la moyenne annuelle. L'effet de pépite est de 15, le modèle choisi est un modèle gaussien de portée 3800m et de palier partiel 14. La taille du pas est toujours de 615m et on dessine le variogramme sur une distance de 10 pas.



Les paramètres de ce modèle sont ajustés automatiquement par le module. Nous avions débuté la modélisation avec un modèle gaussien, mais l'observation des nuages de corrélation et les statistiques de la validation croisée nous ont fait préférer un modèle sphérique. On présente ci contre la forme de modèle ajusté.

Nous avons reporté les résultats de la validation croisée des deux méthodes dans un tableau.

Modèle	**Erreur**		**Erreur relative**	**Coefficient de corrélation**
	moyenne	variance	Variance	
estimation directe de la moyenne annuelle expérimentale par krigeage universel	0,18	8,89	0,018	0,636
moyenne des estimations saisonnières par cokrigeage universel	-0,02	5,91	0,009	0,742

Tableau 5 : Récapitulatif des statistiques de la validation croisée des deux méthodes d'estimation de la pollution moyenne annuelle

Là aussi les statistiques relatives au cokrigeage sont meilleures que celles du krigeage. L'estimateur est moins biaisé et plus précis car la moyenne des erreurs et la variance sont moindres. Mais il est aussi mieux corrélé, on le voit avec le coefficient de corrélation.

On présente les nuages de corrélation pour les deux méthodes :

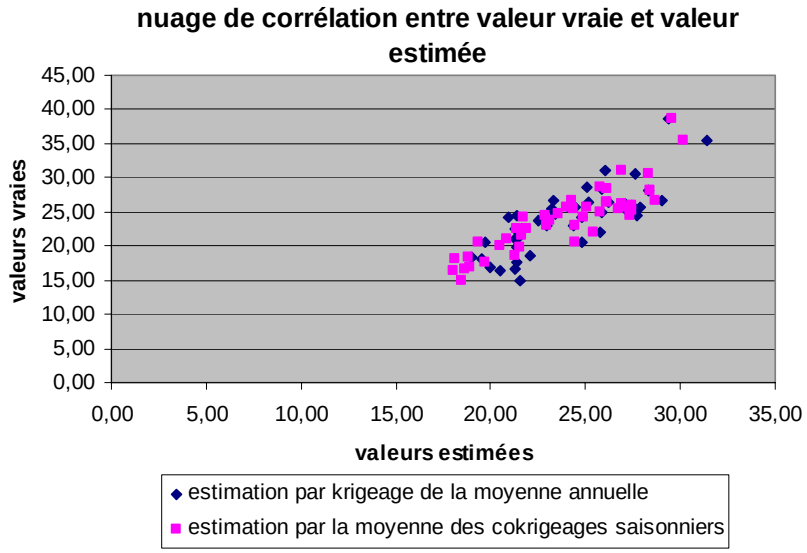


Figure 35 : Présentation des nuages de corrélation pour les deux méthodes de prise en compte de la moyenne

Des deux nuages superposés, celui de l'estimation par la moyenne des cokrigeage est légèrement mieux groupé. Il présente en effet moins de valeurs éloignées. Ces nuages illustrent donc bien les observations faites sur le tableau des statistiques.

Ces deux méthodes nous mènent à la production des cartes présentées ci-dessous :

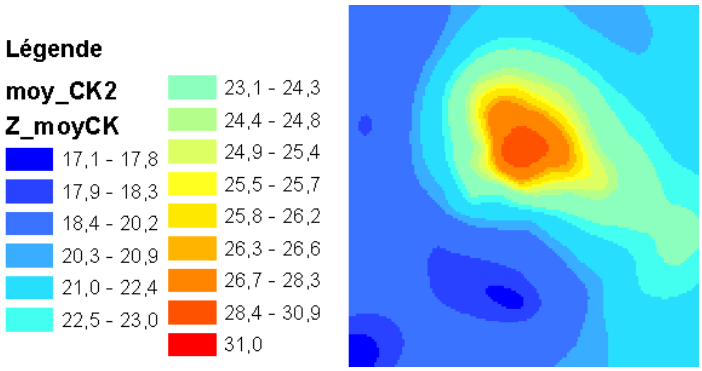


Figure 36 : Carte de la moyenne des cokrigeages saisonniers

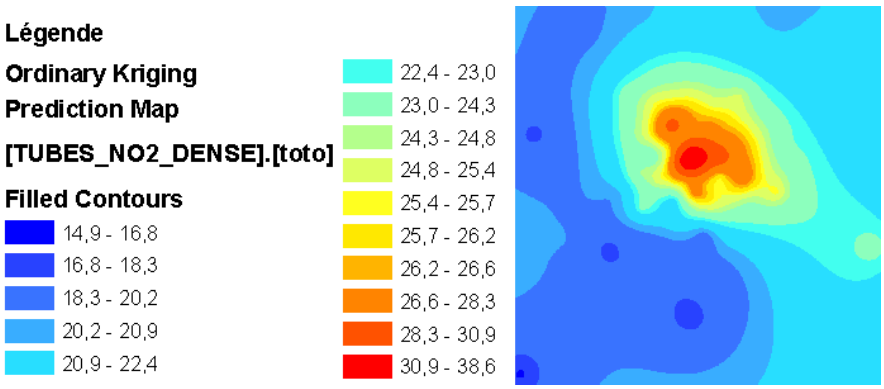


Figure 37 : Cartes du krigeage de la moyenne annuelle expérimentale

Au final, l'étude des données sous «GA» confirme bien les résultats obtenus avec Isatis®, avec cependant moins de netteté. On voit bien une démarcation entre les deux méthodes d'estimation, mais l'étude sous Isatis® nous montre des nuages et des statistiques mieux différenciés.

5.4. REMARQUES SUR L'UTILISATION DU MODULE :

Dans cette étude, nous avons utilisé le module Geostatistical Analyst, mais aussi quelques autres applications qui ont permis de faire toutes les opérations requises.

Certaines limites dans le traitement des données ont été mises à jour. En effet, il est impossible d'effectuer directement des opérations sur les résultats du module et de les cartographier. Pour pouvoir afficher la carte de la moyenne des cokrigeages, nous avons du passer par plusieurs étapes différentes. Le traitement des couches de cokrigeage pour aboutir à la couche de la moyenne des cokrigeages saisonniers est détaillé ci-dessous:

- Il faut en premier lieu récupérer les valeurs des estimations, qui ne sont disponibles sur aucune table attributaire, puisque ces dernières n'existent pas. Pour cela une autre application auxiliaire, « Raster to XYZ » est utilisée, pour convertir le raster de la couche d'estimation en une table des coordonnées des points d'estimation. Dans les champs de l'application, on peut choisir l'image raster à convertir en tableau (Single Band Raster Layer/Theme), le dossier d'enregistrement (Output File), le séparateur de colonnes (Delimiter), si les points ont un identifiant (Include ID column) et leur type (entier ou réel, Output Type) et si le tableau dispose d'un en-tête (Header).

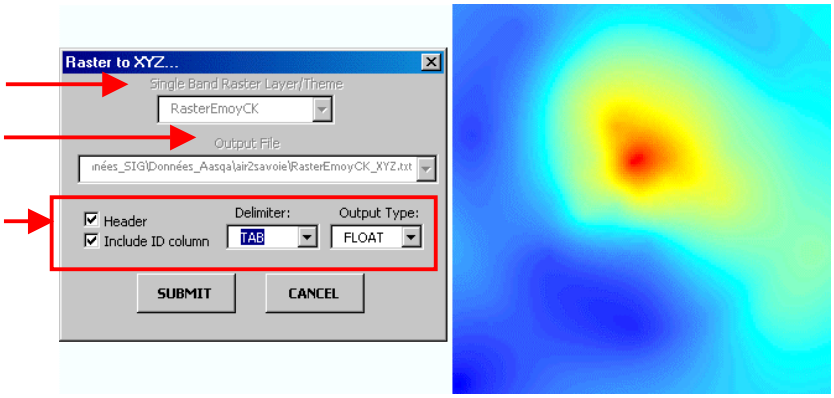
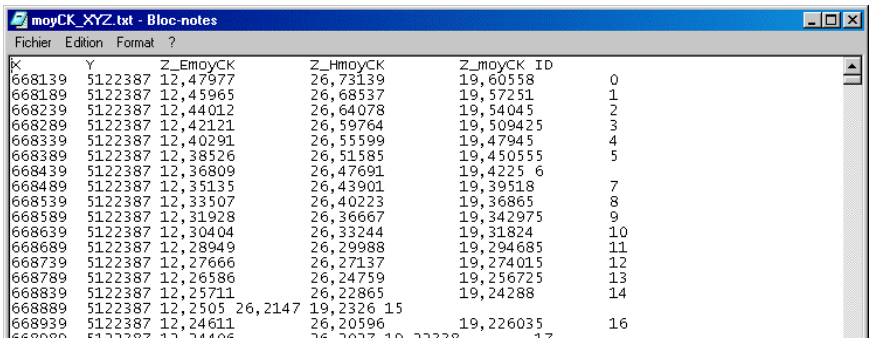


Figure 38 : Présentation de l'outil « Raster to XYZ » et de l'image correspondante à convertir

Grâce à cet outil, on obtient une table en format *.txt qui comporte quatre colonnes : identifiant, X, Y et Z. On fait cette opération pour le cokrigeage estival et le cokrigeage hivernal.

- On peut ouvrir les tables *.txt sous excel. La colonne Z représente la valeur estimée de la variable, on va donc rassembler les deux colonnes Z dans une troisième table et créer une sixième colonne qui est la moyenne de celles-ci. Pour joindre les deux tables et faire la moyenne sur les points que se correspondent, on se base sur l'identifiant. De cette manière nous avons créé la table de la moyenne des cokrigeages.


moyCK_XYZ.txt - Bloc-notes

Fichier Edition Format ?

X	Y	Z_EmoyCK	Z_HmoyCK	Z_moyCK	ID
668139	5122387	12,47977	26,73139	19,60558	0
668189	5122387	12,45965	26,68537	19,57251	1
668239	5122387	12,44012	26,64078	19,54045	2
668289	5122387	12,42121	26,59764	19,509425	3
668339	5122387	12,40291	26,55599	19,47945	4
668389	5122387	12,38526	26,51585	19,450555	5
668439	5122387	12,36809	26,47691	19,4225	6
668489	5122387	12,35135	26,43901	19,39518	7
668539	5122387	12,33507	26,40223	19,36865	8
668589	5122387	12,31928	26,36667	19,342975	9
668639	5122387	12,30404	26,33244	19,31824	10
668689	5122387	12,28949	26,29988	19,294685	11
668739	5122387	12,27666	26,27137	19,274015	12
668789	5122387	12,26586	26,24759	19,256725	13
668839	5122387	12,25711	26,22865	19,24288	14
668889	5122387	12,2505	26,2147	19,2326	15
668939	5122387	12,24611	26,20596	19,226035	16

*Figure 39 : Présentration de la table de la moyenne en format *.txt, on arenommé les colonnes Z pour ne pas les confondre*

- On importe la table de la moyenne dans le bloc de données qui nous intéresse.

  A partir de celle-ci, on crée la couche des points d'estimation de la moyenne (commande « Afficher les données XY… »). A ce stade, nous disposons d'une couche de points où est enregistrée la moyenne des cokrigeages. Or pour une représentation graphique, nous devons avoir des entités surfaciques.

En effet, la conversion raster vers table nous à fait basculer des données surfaciques à des données ponctuelles. Les points étant les centres des carrés de la maille raster. L'étape suivante permet de revenir à des données surfaciques.

- Disposant déjà de la maille correspondante, nous n'avons pas eu à la créer, mais dans le cas contraire il faudrait le faire avec l'outil « Analyse par Maille ». Nous avons joint les tables de la maille et de la couche de points. Nous avons ainsi une maille avec les valeurs des estimations, c'est-à-dire que les valeurs de la moyenne sont associées à des entités surfaciques.

- Il ne reste plus qu'à appliquer une symbologie en dégradé de couleur. Cependant, aucune méthode de classification n'étant commune à une « couche d'interpolation » (créée avec le krigeage de la moyenne annuelle) et à un « shapefile » (que nous venons de créer), nous avons dû faire correspondre manuellement les classes.

C'est seulement en suivant ces étapes que nous avons pu créer une nouvelle carte de la moyenne des cokrigeages saisonniers.

On peut voir à travers ces remarques que ce module de géostatistique est principalement destiné à un usage relativement superficiel. C'est-à-dire qu'une analyse poussée des données, et traitement rigoureux ne sont pas facilités.

6. UTILISATION DES METHODES DE KRIGEAGE NON LINEAIRE :

Nous avons déjà exploré les méthodes de la géostatistique linéaire du module et vu que ces méthodes permettent d'estimer la variable sur tout le domaine d'étude, mais aussi de donner la variance de l'estimation, en d'autres termes, sa précision. Cependant ces méthodes ne donnent pas d'indication précise sur le risque de dépassement d'une certaine valeur.

En effet, les cartes de risque disponibles avec les méthodes linéaires ne sont valables que pour les hypothèses de normalité de la distribution de la variable et de l'erreur. Les méthodes de la géostatistique non linéaire permettent de répondre efficacement à cette question sans effectuer d'hypothèse sur la variable autre que la stationnarité.

Dans cette partie les possibilités offertes par le module à propos des méthodes non linéaire sont explorées. Puis ces méthodes sont appliquées au cas concret de la pollution de l'air par le NO_2 à Bourg-en-Bresse.

6.1. PRINCIPES DE KRIGEAGE NON LINEAIRE DANS LE MODULE :

6.1.1. KRIGEAGE D'INDICATRICE

Si on prend le même modèle de la variable que pour le krigeage ordinaire, on a de même :

$$Z_t(s_i) = m + \varepsilon_t(s_i)$$

Seulement, on va maintenant transformer la variable avec la fonction indicatrice. Cette fonction fait correspondre 1 à la variable si celle-ci vérifie la condition de l'indicatrice. La condition est en général une condition de dépassement de seuil.

On transforme de cette manière les valeurs en échantillon binaire qui prend les valeurs 0 ou 1 :

$$\begin{aligned} I_{Z_t(s_i)\geq c_1} &= 1 \quad \text{si } Z_t(s_i) \geq c_1 \\ &= 0 \quad \text{sinon} \end{aligned}$$, les valeurs échantillon sont comparées au seuil c_1

On procéde ensuite de même que dans le cas du krigeage ordinaire, mais sur la variable transformée :

$$Z_t^1(s_i) = I_{Z_t(s_i)\geq c_1} = m_1 + \varepsilon_t^1(s_i)$$

Il est aussi possible de choisir plusieurs seuils différents, la variable d'étude sera alors transformée en plusieurs variables binaires distinctes. Un cokrigeage sur ces variables est ensuite effectué.

Les valeurs d'estimation prises par la carte donnent la tendance du point à prendre la valeur 0 ou 1. C'est donc la carte de probabilité que l'on a de dépasser le seuil de l'indicatrice.

6.1.2. KRIGEAGE DE PROBABILITE

Le modèle utilisé par cette méthode est double. Cette méthode utilise en effet à la fois le modèle du krigeage d'indicatrice une fois la variable transformée, et le modèle du krigeage ordinaire. Le modèle s'écrit:

$$Z_t^1(s_i) = I_{Z_t(s_i)\geq c_1} = m_1 + \varepsilon_t^1(s_i)$$
$$Z_t(s_i) = m + \varepsilon_t(s_i)$$

On va ensuite effectuer un cokrigeage de ces deux variables avec comme variable principale la variable indicatrice et le variable non transformée comme variable secondaire.

L'objectif est ainsi d'affiner la carte des tendances pour avoir des estimations plus précises. Bien que l'on prenne en compte la variable d'origine, la carte donne des valeurs sur la tendance qu'a la valeur de se rapprocher de 1 ou de 0. On a donc là aussi une carte de probabilité de dépasser le seuil c_1.

6.1.3. KRIGEAGE DISJONCTIF

Dans les deux cas précédents, l'estimateur était la somme pondérée des indicatrices des valeurs échantillons. Ce n'est pas une combinaison linéaire des valeurs échantillon (on a transformé la variable avec l'indicatrice), et c'est pour cette raison que l'on nomme ces méthodes non linéaires.

Dans le krigeage disjonctif, l'estimateur est la somme de fonctions des valeurs échantillon :

$$g^*(Z(s)) = \sum_i f_i(Z(s_i))$$

et considère le modèle de la variable : $f(Z(s_i)) = m + \varepsilon(s_i)$

Il est possible de décrire la distribution expérimentale de la variable à l'aide de fonctions indicatrices. En effet, la variable s'exprime comme la somme des indicatrices pondérées par la valeur prise par la variable. Cette expression est discontinue (d'où le nom disjonctif) car on exprime chaque indicatrice en un point que l'on connaît.

De plus, les indicatrices sont corrélées entre elles car elles reflètent la distribution de la variable. Si on effectue le krigeage de la somme, alors il faut effectuer un cokrigeage sur toutes les variables indicatrices. Il faudrait donc ajuster de très nombreux variogrammes si on veut décrire précisément la variable.

Pour éviter ce travail long et fastidieux, on peut transformer la variable pour obtenir des fonctions non corrélées entre elles et qui décrivent la variable. Cette transformation de la variable se fait grâce aux polynômes d'Hermite, on peut même exprimer n'importe quelle fonction de la variable en fonction de la somme de ces polynômes. De cette manière, le krigeage disjonctif de la fonction de la variable se résume au krigeage des différents polynômes d'Hermite qui la composent.

Cependant, l'utilisation des polynômes d'Hermite requiert une hypothèse de b-inormalité sur le couple (Y(s) ; Y(s+h)), Y(s) étant la variable que l'on cherche à estimer. Or nous avons vu que malgré l'hypothèse de bi-normalité sur (Y(s) ; Y(s+h)), la décomposition en polynômes d'Hermite est possible pour toute fonction de Y(s).

L'hypothèse de bi-normalité étant rarement vérifiée sur la variable brute (Z(s)), on la transforme par une fonction d'anamorphose pour obtenir une variable qui vérifie l'hypothèse : $Z(s) = \Phi(Y(s))$.

Au final on a donc le krigeage disjonctif sur Z(s) qui se résume à un série de krigeages sur les polynômes d'Hermite.

6.2. LA PROCEDURE DE KRIGEAGE DANS LE MODULE :

Après avoir expliqué le principe des méthodes, nous allons voir dans cette partie comment les appliquer.

6.2.1. KRIGEAGE D'INDICATRICE

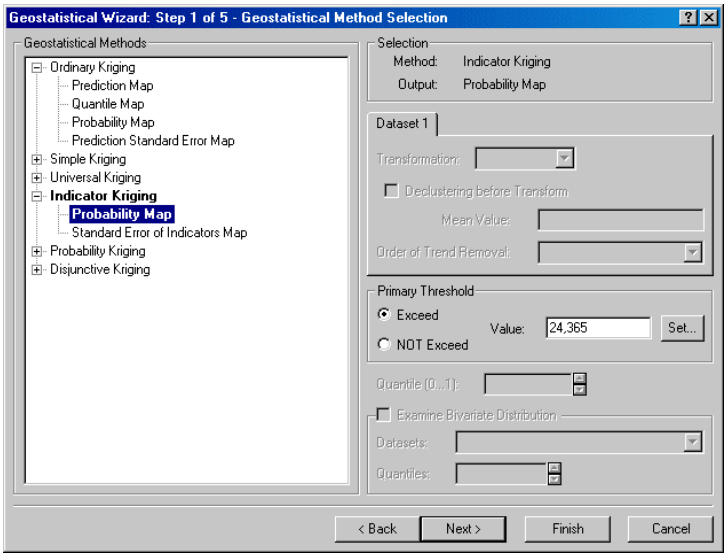
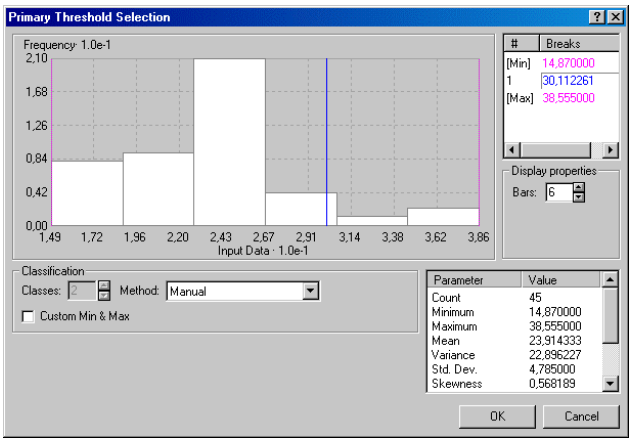


Figure 40 : Fenêtre de dialogue pour le krigeage d'indicatrice

6.2.1.1 LA PROCEDURE DANS LE MODULE

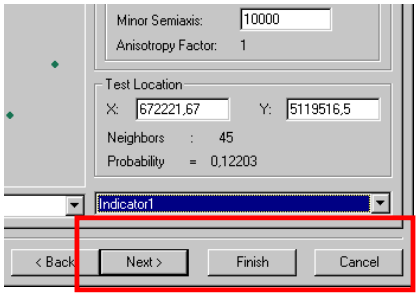
La première fenêtre donne le choix du seuil et permet de préciser si on cherche le risque de dépassement ou de non dépassement du seuil. La valeur par défaut est la moyenne des valeurs de l'échantillon, et il est possible d'ajuster le seuil avec différentes méthodes proposées lorsque l'on clique sur le bouton « Set… ».



Il est possible dans cette fenêtre d'ajuster soi même le seuil (manual), en prenant la médiane (quantile) ou avec la méthode smart quantile. Cette fenêtre auxiliaire aide au choix du seuil en présentant l'histogramme de la distribution et les principales statistiques de la variable.

La seconde fenêtre de progression est construite sur le même modèle que la fenêtre auxiliaire. On dispose de plus d'options. Il est ainsi possible de définir le nombre de seuils auxiliaires et donc le nombre de variables secondaires à utiliser. On peut les délimiter avec les mêmes méthodes que dans la fenêtre auxiliaire.

Une fois les seuils secondaires choisis, on retrouve la fenêtre d'ajustement du variogramme. Pour le choix des paramètres, on procède de même que pour le krigeage linéaire. Si on a choisi des seuils auxiliaires, et donc des variables indicatrices secondaires, on se retrouve dans le cas du cokrigeage, et on sélectionne les variogrammes à ajuster dans



la liste déroulante « Semivariogram/Covariances ».

Une fois les choix des paramètres effectués, on passe au choix du voisinage. Les paramètres principaux ne changent pas, mais on voit apparaître une nouvelle liste déroulante sous la rubrique « test location ». Cette liste permet de choisir la « variable » de cokrigeage.En effet, vu que l'on a créé ces variables à partir de la variable originale, ce ne sont pas des jeux de données.

On doit donc faire le choix de la variable dans une autre liste que celle où on fait le choix du jeu de données.

L'étape de la validation croisée présente les statistiques que nous avons vues précédemment, et un seul graphe. Ce dernier représente la corrélation entre les valeurs mesurées et les valeurs estimées. Etant donné que les valeurs estimées sont des probabilités, on n'a pas de droite de régression, mais on indique le seuil de l'indicatrice. On peut ainsi voir quelle est la probabilité estimée de dépassement du seuil pour les points expérimentaux. La table de validation croisée présente une colonne supplémentaire qui indique la valeur de l'indicatrice principale.

6.2.1.2 CAS CONCRET

Nous allons estimer le risque de dépasser un seuil de concentration en NO_2 de 30µg/m^3 sur la commune de Bourg-en-Bresse. Cette concentration est la concentration moyenne sur l'année, la variable origine que nous avons est donc la moyenne des moyennes saisonnières.

On ne prend pas en compte de seuil secondaire. On ne crée donc pas de variable indicatrice secondaire. On va ajuster seulement le variogramme de la variable indicatrice principale et aucun variogramme croisé. L'ajustement que nous effectuons est proche de celui proposé automatiquement par le module. Pour un pas de 500m on a le modèle :

0,01*Gaussien(3800) + 0,08*Pépite

On obtient un risque compris dans l'intervalle [0,019 ; 0,156]. On obtient donc un risque globalement faible dont on présente la carte ci-dessous. Cependant, même pour les points qui ont effectivement dépassé ce seuil, on n'estime qu'un risque très faible.

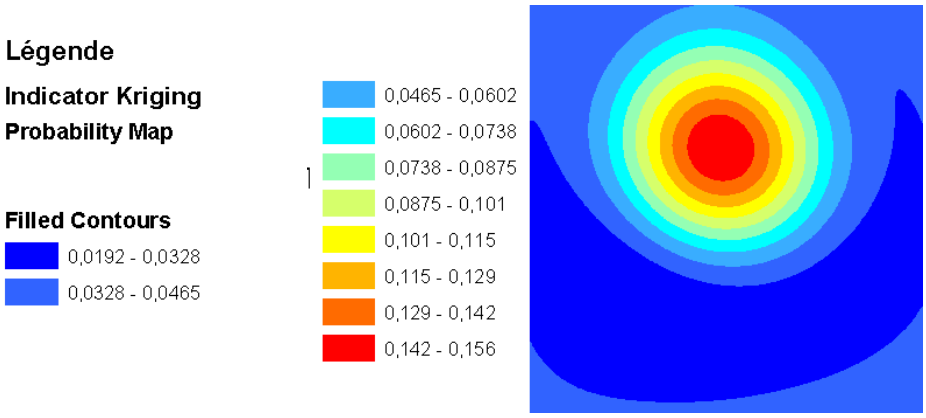


Figure 41 : Cartographie du risque de dépassement de la concentration en NO_2 pour un seuil de 30µg/m3 par krigeage d'indicatrice

6.2.2. KRIGEAGE DE PROBABILITE

Cette méthode prenant en compte à la fois la variable origine et la variable indicatrice due au seuil, nous allons devoir ajuster le variogramme de chacune des variables, puis leur covariance croisée. La procédure à suivre pour créer la carte est donc la même que pour un krigeage d'indicatrice.

On cartographie le même risque que pour le krigeage d'indicatrice, nous pourrons ainsi comparer les résultats.

On ajuste le variogramme de la variable indicatrice selon le même modèle que pour le krigeage d'indicatrice. Pour le variogramme de la variable origine (moyenne annuelle), on reprend le modèle utilisé dans la partie du cokrigeage avec un modèle gaussien, on a donc :

Variable indicatrice : 0,01*Gaussien(3800) + 0,08*Pépite

Variable origine : 14*Gaussien(3800) + 15*Pépite

Croisé : 0,37*Gaussien(3800) + 0,5*Pépite

L'intervalle d'estimation du risque est plus large que précédemment : [0 ; 0,203]. Mais on a là aussi un risque estimé faible pour des points qui dépassent le seuil, on présente la carte correspondante ci-dessous.

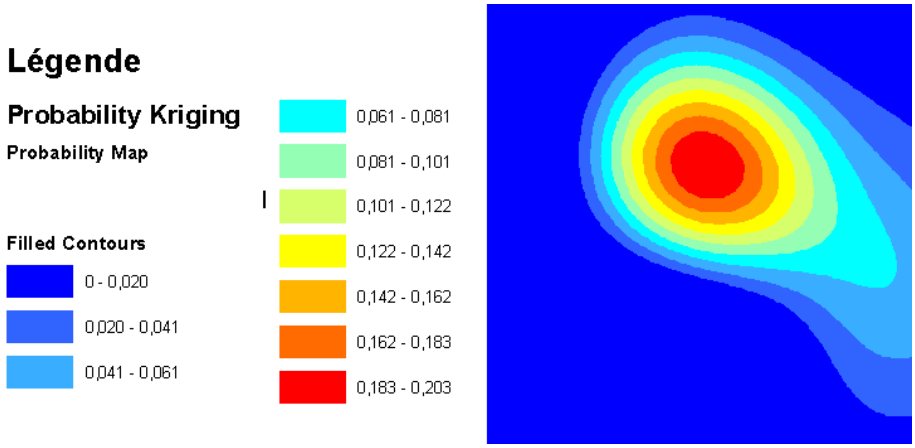


Figure 42 : Cartographie du risque de dépassement de la concentration en NO_2 pour un seuil de 30µg/m^3 par krigeage de probabilité

6.2.3. KRIGEAGE DISJONCTIF

6.2.3.1 LES OUTILS MIS A DISPOSITION

Ce type de krigeage autorise plus de possibilités que les autres méthodes. On peut en effet voir dès la fenêtre du choix de la méthode que l'on peut créer une carte de probabilité ou bien d'estimation. C'est-à-dire que suivant le type de carte, on peut estimer la variable brute ou bien un risque par rapport à un seuil fixé.

Dans la rubrique « dataset », il est possible de transformer la variable avec toutes les méthodes disponibles, et il est aussi possible de supprimer la dérive. Ces options sont disponibles pour pouvoir ajuster la variable et lui faire suivre au plus près une loi normale. De cette manière, on tente de remplir l'hypothèse de bi-normalité. En d'autres termes, ces options servent à ajuster la fonction d'anamorphose.

Cependant, ces options se basent sur une adaptation à une loi gaussienne approximative. Pour avoir une bonne anamorphose, il est plus judicieux de passer par la transformation par équivalent gaussien (normal score).

En effet, cette transformation fait correspondre chacun des quantiles de la distribution de la variable au quantile correspondant d'une variable normale selon trois méthodes différentes. Cette option déclenche l'ouverture d'une nouvelle boite de dialogue :

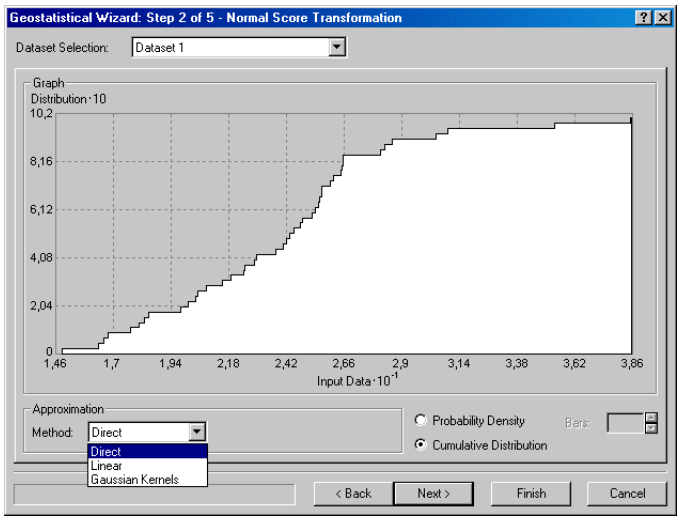


Figure 43 : Fenêtre de dialogue de la transformation par les équivalents gaussiens

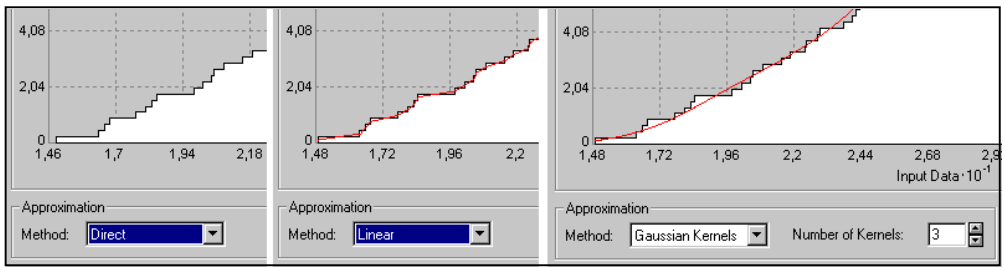


Figure 44 : Illustration des trois méthodes de transformation par les équivalents gaussiens

Les trois méthodes ont le même but, cependant les méthodes linéaires et les kernels gaussiens permettent d'avoir une progression des approximations plus douce. En effet, elles permettent de lisser l'approximation en « gommant » les marches de la distribution expérimentale. Cependant, ces deux méthodes induisent des hypothèses sur la distribution de la variable.

Uniformiser la répartition spatiale des données est possible à l'aide de la fonction « declustering ». C'est-à-dire que l'on assigne un poids plus faible aux données qui se trouvent dans des zones sur-échantillonnées et un poids plus fort pour celles dans les zones sous-échantillonnées.

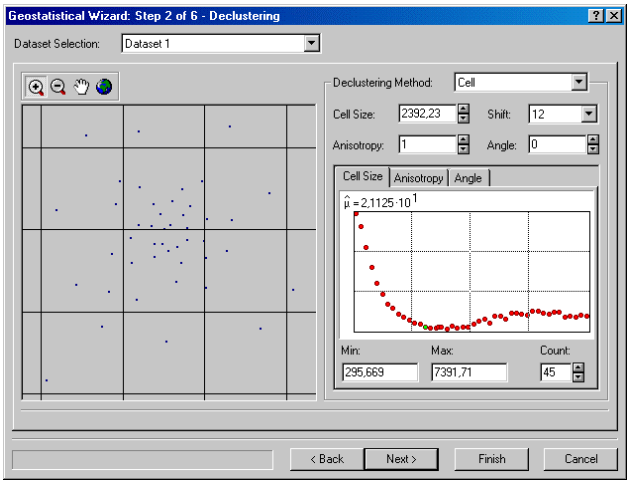


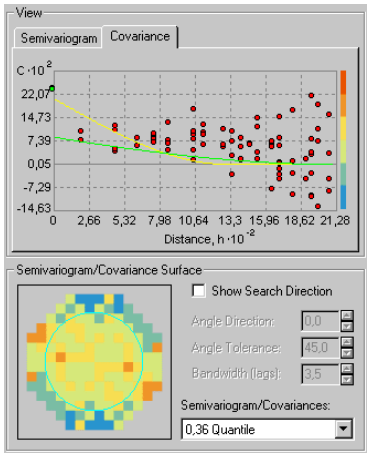
Figure 45 : Fenêtre de dialogue pour réuniformiser la répartition des données

Cet ajustement se fait notamment avec une grille de délimitation dont on peut choisir la taille. Dans chaque cellule de la grille, les données sont pondérées. On choisit la taille de maille de la grille avec la moyenne pondérée des données.

En effet, si les zones de forte densité correspondent à des zones de fortes valeurs (ce qui est notre cas), pour rétablir l'équilibre, on prend une taille de maille qui correspond à une faible moyenne pondérée. Au contraire, si les valeurs faibles sont les plus denses, alors on prend une taille de maille qui correspond à une forte moyenne pondérée.

Il existe aussi la méthode des polygones pour la même tâche. On raisonne cette fois sur la taille des polygones qui entourent chaque point échantillon. Cependant on voit que cette méthode rencontre de fortes difficultés pour les points proches de la frontière de la zone d'étude. C'est pourquoi nous ne considérons pas cette méthode dans notre étude.

On peut aussi vérifier la validité de l'hypothèse de bi-normalité avec l'option « Examine Bivariate Distribution ». Cette dernière option va libérer une fenêtre de dialogue après l'ajustement du modèle.



Cet examen permet de faire la comparaison entre la variable réelle et une variable gaussienne. Pour cela on compare les distributions des deux variables, le module permettant de dessiner pour chaque quantile deux courbes de comparaison.

La probabilité que $Z(s_i)$ et $Z(s_i+h)$ soient inférieures à un quantile z_p varie avec h et z_p. La courbe verte représente cette fonction sous l'hypothèse que la bi-normalité de $(Z(s_i)$; $Z(s_i+h))$ est vérifiée.

La courbe jaune représente le modèle ajusté à la variable.

Plus les deux courbes sont proches et plus on tend à vérifier l'hypothèse pour la variable que l'on considère.

Dans l'option « Examine Bivariate Distribution », on peut choisir le nombre de quantiles que l'on désire examiner. Puis dans la fenêtre de dialogue, on peut réajuster le modèle pour vérifier au mieux l'hypothèse de bi-normalité.

La suite des étapes de la création de carte (voisinage et validation croisée) suit le même schéma que pour les autres types de krigeage.

6.2.3.2 CAS CONCRET

Nous choisissons de transformer la variable suivant les équivalent gaussiens (normal score transformation) avec la méthode directe. De cette manière, nous faisons correspondre à la variable une variable gaussienne, tout faisant le moins d'hypothèses possible dessus.

On choisit de ne pas uniformiser la répartition des données avec « declustering » après un examen rapide des statistiques de la validation croisée.

On ajuste deux modèles différents, mais proches sont ajustés pour les deux types de carte que l'on obtient :

Estimation : 1*Gaussien(4000)+0,4*Pépite

Probabilité : 0,73*Gaussien(4000)+0,5*Pépite

On obtient la carte présentée ci-après :

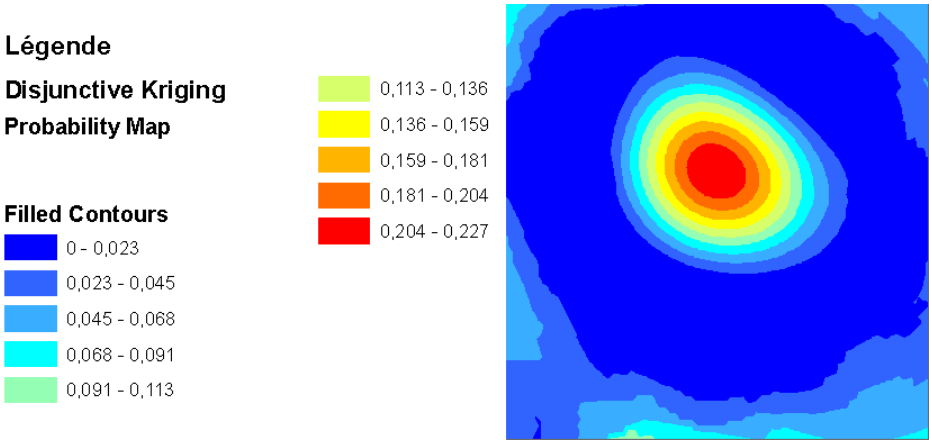


Figure 46 : Présentation de la carte de probabilité de dépasser une concentration de 30µg/m^3 de NO_2 faite avec la méthode du krigeage disjonctif

L'intervalle de l'estimation est encore plus grand que dans les cas précédents. Mais cette carte ne permet toujours pas de voir les zones où le seuil a été effectivement dépassé.

On compare ensuite les statistiques obtenues pour les validations croisées de chacunes des méthodes de krigeage non linéaire dans le tableau suivant :

type de krigeage	**statistiques sur l'erreur**		
	erreur moyenne (0)	variance de l'erreur (1)	variance de l'erreur réduite (1)
Krigeage d'indicatrice	0,0018	0,0797	0,9514
Krigeage de probabilité	0,0094	0,0746	0,9028
Krigeage disjonctif	0,0120	0,0765	1,4232

Tableau 6 : Récapitulatif des validations croisées des méthodes de krigeage non linéaire

On voit donc que c'est le krigeage d'indicatrice qui présente en moyenne des erreurs plus faibles. C'est également cette méthode qui présente les erreurs théoriques les plus proches des erreurs expérimentales, car sa variance de l'erreur réduite se rapproche le plus de 1.

C'est par contre le krigeage de probabilité qui présente l'estimateur le plus précis (variance de l'erreur la plus faible), mais il estime moins bien l'erreur que le krigeage de probabilité. De plus on remarque que les valeurs de la variance de l'erreur ne changent pas autant que les autres statistiques, d'une méthode à l'autre.

Globalement, on voit donc que c'est le krigeage de probabilité qui présente le meilleur estimateur de la pollution moyenne en NO_2.

De plus, si on regarde les statistiques de la validation croisée d'estimation, on voit que le krigeage disjonctif n'est pas très performant sur ce jeu de données.

erreur moyenne **(0)**	**variance de l'erreur** **(1)**	**variance de l'erreur réduite** **(1)**
0,386	10,17	1

Tableau 7 : Statistiques de la validation croisée de l'estimation par krigeage disjonctif

On voit que les erreurs sont très précisément estimées, mais que l'estimateur est moins ciblé et moins centré que les autres estimateurs de la partie à propos du cokrigeage.

7. CONCLUSION

Cette étude du module de géostatistique disponible sur ArcView nous a permis de décrire son application à des cas concrets sur la pollution de l'air par l'ozone et le NO_2.

A propos du module lui-même, nous avons constaté certaines limitations qui ne permettent pas d'effectuer une étude optimale. En particulier, la dérive externe avec une fonction auxiliaire ne peut être prise en compte.

D'autre part, nous avons remarqué que le krigeage ordinaire avec dérive, suit une méthode d'analyse qui ne semble pas rigoureuse. En effet, il retranche la valeur estimée de la dérive à la valeur brute et effectue une estimation sur le résidu. Puis à la dernière étape le module rajoute la valeur de la dérive au résidu estimé. La méthode appliquée est donc un krigeage des résidus.

Or cette méthode suppose que l'on puisse définir précisément la dérive (la part déterministe de la variable). Alors que le module ne peut prendre en compte qu'une estimation de cette dérive. C'est pour cela que l'on peut penser que cette méthode manque de rigueur et ce qui explique qu'elle n'ait pas été utilisée dans notre étude.

Nous avons aussi remarqué que l'usage du module se fait en suivant une progression linéaire. Si l'on souhaite comparer plusieurs méthodes, l'utilisation du module devient rapidement fastidieuse. Il faut sans cesse revenir dans les fenêtres de progression précédentes, modifier les paramètres, continuer la nouvelle étude, enregistrer la validation croisée, et ainsi de suite. Ensuite, on si on veut comparer plus de deux méthodes, il faut rassembler les statistiques de validation croisée dans un tableau à part.

Par ailleurs, dans l'utilisation pratique du logiciel, des limites dans l'affichage des données ont été données. En effet, le variogramme expérimental n'est pas disponible, l'ajustement du modèle théorique se faisant uniquement avec la nuée de points. L'ajustement se fait en changeant les paramètres à la main, on ne peut pas déformer sur le variogramme le modèle pour l'ajuster.

L'échelle d'affichage du variogramme peut aisément prêter à confusion pour un utilisateur peu habitué au module. D'autre part, dans la fenêtre de validation croisée, les nuages de corrélation ne sont pas toujours affichés de manière lisible. Nous avons en effet choisi d'utiliser une application auxiliaire, « Scatter », disponible sur le site de support d'ESRI pour afficher nos nuages de points.

De plus, pour pouvoir mener à bien l'étude, nous avons fait appel à plusieurs applications annexes. Celles-ci ne font pas partie des commandes de base et sont téléchargeables sur le site du support technique de Esri.

De façon générale, ce module est principalement destiné à des utilisateurs non spécialistes de la géostatistique et qui ne souhaitent pas une analyse très poussée des données. Il permet une analyse qui soit à la fois facile et qui puisse créer rapidement des cartes de première approche.

8. REFERENCES

« Using ArcGIS «GA» », (2001)

Cressie, N. « Statistics for Spatial Data, Revised Edition », (1993), John Wiley & Sons

Armstrong, M. et Carignan, J. "Géostatistique Linéaire, Application au domaine minier », (1997), Presses de l'Ecole des Mines de Paris

Deutsch, C. et Journel, A. « GSLIB : Geostatistical Software Library ans User's Guide », (1998), seconde édition, Oxford University Press

Cardenas G., Malherbe L., Evaluation des incertitudes associées aux méthodes géostatistiques, Convention N° 115/03, INERIS, Décembre 2003. Téléchargeable sur l'adresse :

http://www.lcsqa.org/rapport/rap/prog2003/ineris/Etude15_rapport_incertitude_avril2004.pdf

Rivoirard, J. Introduction to Disjunctive Kriging and Non-Linear Geostatistics, (1994), Oxford University Press,

LISTE DE DIFFUSION

Nom	Adresse/Service	Nb
	Dossier maître	1
G.Cardenas	2IEN	1
J. Lefèvre	Secrétariat 2IEN	1
L. Rouil	MECO	1
M. Ramel	DRCG	1
D. Rouyer	DRCG	1
MEDD		…

TOTAL **6**

PERSONNES AYANT PARTICIPE A L'ETUDE

Travail	Nom	Qualité	Date	Visa
Rédacteur	Giovanni Cardenas	Ingénieur Etudes	31/12/2004	
Responsable d'affaire				
Relecteur				
Vérificateur				
Approbateur				

Fin du Complément non destiné au client