

**Ecole des Mines  
de Paris**

---

**Contrat INERIS/ARMINES no. CNS 003 27 84**

*Etude sur la réalisation de cartographies de la qualité de l'air dans les zones peu/pas  
couvertes par les réseaux de stations fixes à l'aide de méthodes géostatistiques*

Rapport d'avancement no. 3

**MÉTHODOLOGIE DE CARTOGRAPHIE DE LA  
CONCENTRATION ANNUELLE DE NO<sub>2</sub> SUR  
L'AGGLOMÉRATION DE MULHOUSE**

**Chantal de FOUQUET**

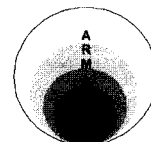
-----

**Rapport N-6/03/G**

**Avril 2003**



**Ecole des Mines de Paris**  
**CENTRE DE GÉOSTATISTIQUE**  
35 rue Saint-Honoré  
77305 FONTAINEBLEAU (France)  
Tél. : 33-1 64 69 47 81 Fax : 33-1 64 69 47 05  
<http://cg.ensmp.fr>



---

1. Récapitulatif des études précédentes .....	1
1.1. Campagnes de mesures .....	1
1.2. Résumé des études antérieures .....	2
2. Finalisation de l'étude variographique exploratoire .....	6
2.1. Rappels sur les données .....	6
2.2. Statistiques élémentaires de NO <sub>2</sub> mesuré aux tubes .....	9
2.3. Examen des tubes rapprochés .....	14
2.4. Corrélations entre NO <sub>2</sub> – tubes et les variables auxiliaires .....	15
2.5. Modélisation .....	26
3. Comparaison d'estimateurs par validation croisée. ....	29
3.1. Modèles de référence .....	29
3.2. Sensibilité aux variables auxiliaires .....	33
3.3. Précision des estimations .....	35
3.4. Sensibilité au cadastre des émissions .....	38
4. Conclusions .....	39

## **1. Récapitulatif des études précédentes**

L'étude des relations entre les mesures par échantillonneurs passifs de la concentration en NO<sub>2</sub> sur l'agglomération de Mulhouse et différents "paramètres explicatifs" a été abordée dans plusieurs travaux antérieurs. Pour la présente étude, les documents suivants ont été consultés :

- "Diagnostic de la qualité de l'air sur l'agglomération de Mulhouse ; répartition spatiale de la pollution atmosphérique. Rapport final" de mars 2002 (source d'information ASPA 02031901-I-D) et son annexe "Méthodes d'interpolation spatiale", transmis par l'ASPA ou disponible sur son site Internet.
- les transparents de la présentation du rapport de DESS de M. Ahmed Hamadouche, intitulé "Automatisation du logiciel ISATIS", datés de septembre 2002 et disponibles sur ce même site.

Dans la suite, ces études sont désignées comme "étude ASPA" ou DESS pour la partie d'analyse des corrélations entre variables.

- un "examen préliminaire des données de Mulhouse" a été mené par M. Pierre Chauvet, dans le cadre de la convention INERIS/ARMINES CNS 0032784 (rapport N-15/02/G d'octobre 2002), qui constitue la première partie du présent travail. L'étude présentée dans le premier rapport d'avancement a été effectuée indépendamment des études menées par l'ASPA. Dans la suite, ce travail est désigné comme "étude CG".

Nous mentionnons enfin le rapport de stage d'option (Ecole des Mines de Paris, juin 2002) effectué par Guillaume L'Hégaret auprès d'ASCOPARG, traitant de la même problématique, pour NO<sub>2</sub> et pour le Benzène :

- "Mise en place d'une méthodologie pour la cartographie du Benzène et du dioxyde d'azote à l'échelle des départements du Rhône et de l'Isère".

Nous commençons par rappeler les résultats acquis concernant le NO<sub>2</sub> sur Mulhouse, objet de la présente étude.

### **1.1. Campagnes de mesures**

Les campagnes de mesures exploitées dans l'étude ASPA et l'étude CG sont présentées dans le document "source d'information ASPA 02031901-I-D". Les principales caractéristiques sont les suivantes :

- 75 sites répartis en zone urbaine ont été instrumentés en tubes PASSAM, en vue de "qualifier les niveaux de pollution atmosphérique mesurés, en comparaison avec les résultats obtenus sur d'autres sites", et également pour établir des cartographies des concentrations fournissant une représentation spatiale de la pollution. Les trois stations permanentes du réseau de surveillance (Mulhouse Nord, Est et Sud) ont été équipées chacune d'un ensemble de trois tubes passifs, pour la comparaison des deux types de mesures (stations fixes, tubes).

- la campagne comporte deux phases, chacune composée de trois quinzaines successives. La séquence hivernale couvre la période du 7 février au 21 mars 2001, et la séquence estivale la période du 22 mai au 3 juillet 2001.

Les roses des vents indiquent une prédominance annuelle SO et N-NE, analogue à la moyenne durant les deux phases, avec cependant une moindre fréquence du secteur N-NE durant la phase hivernale.

Enfin, le décret du 15 février 2002 fixe pour NO<sub>2</sub> un niveau de recommandation de 200  $\text{ng}/\text{m}^3$  en concentration horaire et un niveau d'alerte de 400  $\text{ng}/\text{m}^3$ , en concentration horaire également.

## **1.2. Résumé des études antérieures**

### **1.2.1. Méthode d'estimation**

Les deux études visent au krigeage avec dérive externe de la concentration moyenne en NO<sub>2</sub>, annuelle ou saisonnière. Dans le cas particulier d'une covariance pépétique et d'un voisinage unique, le krigeage en dérive externe est équivalent à une régression linéaire multiple.

L'étude ASPA considère la "moyenne annuelle", calculée comme la moyenne des deux phases de trois quinzaines chacune. Sont conservées les stations comportant au moins deux quinzaines pour chaque phase.

L'étude CG traite les moyennes saisonnières calculées sur trois quinzaines, hivernale et estivale. L'effet de prise de moyenne temporelle est moins marqué, ce qui explique la corrélation plus médiocre avec les variables explicatives.

Dans les deux cas, le logarithme translaté des variables explicatives est utilisé, les variables étant divisées par leur moyenne avant translation pour l'étude CG ( $\log\left(1 + \frac{Y}{m_Y}\right)$  au lieu de  $\log(1 + Y)$ ,  $m_Y$  désignant la moyenne expérimentale de Y).

L'estimation présentée dans le rapport DESS retient pour la région de Mulhouse 4 variables en dérive externe :

- une variable liée au bâti, obtenue par régression linéaire multiple sur NO<sub>2</sub> (aux tubes) des trois variables occupation des sols suivantes : bâti dense, bâti lâche et industrie
- une variable liée aux émissions, obtenue par régression linéaire multiple sur NO<sub>2</sub> des données du cadastre des émissions (C<sub>6</sub>H<sub>6</sub>, CO, BAP, NO<sub>x</sub>, particules) ;
- la population et l'altitude.

Comme ISATIS n'accepte que trois variables en dérive externe, plusieurs choix de dérives sont testés, puis les résultats sont comparés par "validation croisée".

Le rapport ASPA présente les cartes obtenues avec les émissions (NO<sub>x</sub> et benzène) et la densité de population en dérive externe (après transformation logarithmique de ces variables auxiliaires, cf. annexe p9).

Pour l'étude CG, les variables auxiliaires finalement retenues sont les trois variables décrivant le bâti (dense, lâche ou occupation industrielle), ainsi que la densité de population (là encore après transformation logarithmique).

### 1.2.2. Sélection de variables explicatives

Comment identifier, parmi les nombreuses variables auxiliaires proposées (bâti en milieu urbanisé, végétation dominante, densité de population, altitude ...) celles présentant un fort pouvoir explicatif de la concentration en NO<sub>2</sub>, permettant d'améliorer l'estimation? Les deux études utilisent une Analyse en Composantes Principales (ACP) pour mettre en évidence les relations entre variables explicatives, et rechercher notamment des familles de variables ou une redondance d'information, ou des oppositions. Dans le cas de Mulhouse, les variables auxiliaires sont informées aux noeuds d'une grille, mais leur valeur "exacte" aux stations de mesure est inconnue. Deux questions se posent:

1. Sur quel ensemble examiner les relations entre variables : sur la grille, restreinte au domaine à estimer, ou aux stations de mesures ?

Idéalement, les stations de mesures (analyseurs et tubes) devraient être "représentatives" de l'ensemble de la zone à estimer. En pratique, c'est rarement le cas : les tubes ou les analyseurs sont implantés préférentiellement, par exemple dans les agglomérations principales, c'est-à-dire dans les zones de forte densité de population et de forte concentration en NO<sub>2</sub>; dans d'autres cas les fonds de vallées ou les zones de faible altitude sont surreprésentés par rapport aux versants d'altitude forte ou moyenne.

Pour résumer l'information apportée par des variables auxiliaires, c'est l'ACP des valeurs aux noeuds de la grille d'estimation qui devrait être effectuée, puisque ces variables synthétiques seront ensuite utilisées dans le krigeage en dérive externe ou dans une régression linéaire multiple. Les facteurs ainsi obtenus doivent alors être interpolés aux tubes. Ces facteurs, en particulier leur valeur aux stations de tubes, dépendent alors de la zone à estimer; ce qui introduit une part d'arbitraire : en effet, les relations entre variables auxiliaires peuvent évoluer notablement entre les périmètres au centre d'une agglomération ou en périphérie plus ou moins éloignée. Par ailleurs, les variables auxiliaires aux noeuds d'une grille présentent une forte corrélation spatiale, et le résultat de l'ACP n'est qu'indicatif, puisque l'hypothèse d'indépendance (ou d'absence de corrélation) des tirages n'est pas vérifiée. Cela peut se traduire par une corrélation spatiale des différents facteurs, bien que point à point non corrélés.

Les relations entre paramètres explicatifs et variable d'intérêt (la concentration en NO<sub>2</sub>) sont nécessairement établies à partir des mesures aux stations, et mettent en jeu les relations entre paramètres explicatifs en ces points. Pour l'estimation, on est toujours amené à extrapoler le modèle multivariable, calé aux points de mesure, à l'ensemble du domaine à estimer. C'est particulièrement évident dans le cas d'une décomposition de type [concentration]=[régression linéaire multivariable]+ [résidu], puisque les paramètres de la régression sont calés aux mesures de la concentration. Si les relations entre les paramètres explicatifs, ou entre ces paramètres et les concentrations, diffèrent fortement entre les points expérimentaux et le domaine à estimer, alors le modèle calé aux points expérimentaux peut devenir inadapté sur certaines parties du domaine. L'estimation devient alors très imprécise, fournissant éventuellement des résultats aberrants, sans que cela apparaisse nécessairement sur la variance d'estimation, calculée dans le modèle calé aux points expérimentaux. C'est en particulier le cas si les stations sont implantées de façon préférentielle par rapport à la variable à estimer.

La comparaison des ACP des variables auxiliaires, aux “noeuds de grille” d’une part, et aux stations d’autre part (cf. figures 11 et 22 du rapport CG, avec cependant des variables différentes entre les deux ACP) est utile pour détecter une implantation préférentielle des stations dans l’espace des variables auxiliaires ou des paramètres explicatifs. On peut également représenter les stations dans le nuage des individus des noeuds de grille (en identifiant le plus proche voisin de chaque station, par exemple).

2. Pour effectuer l’ACP aux stations, ou pour le krigeage en dérive externe, il faut interpoler les variables explicatives aux stations : comment ?

Les propriétés d’un estimateur diffèrent de celles des “valeurs réelles” ; par exemple, valeurs estimées et “réelles” n’ont pas la même variance (de dispersion), ni la même corrélation . L’ACP étant fondée sur l’analyse des relations de variance-covariance des différentes variables, les résultats obtenus sur les variables auxiliaires interpolées aux stations sont nécessairement différents de ceux que l’on obtiendrait sur les valeurs “exactes” de ces variables auxiliaires en ces points.

P. Chauvet a examiné deux procédés d’interpolation à partir des données auxiliaires à maille kilométrique : le krigeage, qui nécessite une étude variographique préalable des variables auxiliaires aux noeuds de grille, et la “migration”, c’est-à-dire l’estimation par la valeur au noeud de grille le plus proche. La figure 20 du rapport CG montre, pour les émissions de NOx provenant du cadastre, l’effet de lissage du krigeage par rapport à la migration, et la dispersion assez importante du nuage de corrélation de ces deux estimateurs. Les nuages de corrélation entre la moyenne saisonnière en NO<sub>2</sub>, et les émissions de NOx diffèrent sensiblement selon l’interpolateur retenu, les coefficients de corrélation étant supérieurs pour le krigeage (0.52 contre 0.45 pour l’été, 0.51 contre 0.40 pour l’hiver).

La figure 3 de l’annexe ASPA indique qu’une interpolation de type inverse des distances a été utilisée, pour informer les stations à partir des noeuds de grille.

**Remarque :** L’étude ASPA a été menée à maille 500m (annexe, p9), l’étude CG à maille kilométrique. Les maillages informés transmis au CG sont les suivants :

- 250m et 1000m pour l’altitude ;
- 1000m pour les émissions et pour la densité de population ;
- 200m pour l’occupation des sols.

G. PERRON (ASPA) a précisé qu’une interpolation en  $1/d^4$  a été utilisée pour informer la grille des résultats au pas 500m, à partir de valeurs fournies à maille kilométrique.

Pour le DESS, le nombre d’observations varie suivant les familles de variables auxiliaires : 11984 pour l’altitude, la densité de population et les 8 variables d’occupation des sols ; 462 observations pour les émissions. Une ACP est donc effectuée par famille de variables, pour les données “grille”, ce qui conduit à sélectionner un nombre réduit de facteurs (combinaisons linéaires des variables explicatives), qui permettent de restituer la majeure partie de la variance. Une régression linéaire multiple de ces facteurs sur la

concentration mesurée permet de sélectionner les quatre contributions les plus fortes, jugées significatives au sens du test de Student. Comme on l'a vu, parmi ces quatre facteurs (bâti dense, lâche et industriel ; cinq émissions dont NO<sub>x</sub> ; population ; altitude), trois sont ensuite retenus suivant un critère de validation croisée dans le krigeage en dérive externe.

Le rapport ASPA présente les cartes obtenues à l'aide des "variables auxiliaires" des émissions à la maille de 500 m<sup>2</sup>, et de la densité de population (s'agit-il des transformées logarithmique de ces émissions, ou bien d'une combinaison linéaire de ces variables, obtenue par exemple par régression de la concentration en NO<sub>2</sub> sur ces deux variables explicatives ?). Le facteur ou les variables associés au bâti, ainsi que l'altitude ont donc été écartés.

Pour l'étude CG, l'ensemble des variables auxiliaire est disponible sur le domaine comportant 870 noeuds de grille, délimité autour de l'agglomération de Mulhouse. Une première ACP (figure 11) fait apparaître un regroupement de variables indiquant une anthropisation du milieu (densité de population, et bâti dense, lâche ou industrie) et les émissions, l'altitude et la végétation décrivant le second axe. Ces dernières variables, moins liées à la pollution, sont alors écartées, et une deuxième ACP cette fois-ci aux stations (avec migration), est menée sur les variables du premier groupe et les deux moyennes saisonnières (figure 22). Les trois variables bâti lâche, dense et industrie, ainsi que Zpop (variable la plus proche du premier facteur) sont alors sélectionnées, et une régression linéaire multiple de la concentration sur ces trois ou quatre variables auxiliaires est tentée. Mais la corrélation entre le résultat de la régression linéaire multiple et la moyenne saisonnière (figure 26 pour l'été, figure 28 pour l'hiver), reste assez médiocre, avec des coefficients de corrélation respectifs de 0.49 et 0.55).

**Remarque :** le pas de la grille ne coïncide pas nécessairement avec le support de la variable auxiliaire ; par exemple, la densité de population, exprimée en nombre d'habitants par hectare, déduite des données de recensement, peut être calculée par hectare (maille 100m), pour 6,25 ha (maille 250m), ou par km<sup>2</sup> ; les valeurs pour ces trois différents supports peuvent être fournies à maille kilométrique. Par ailleurs, le support auquel les variables sont disponibles n'est pas nécessairement le "meilleur" au sens des corrélations avec la concentration, qui est la variable d'intérêt ; le support "optimal" au sens de la maximisation de la corrélation peut varier selon le polluant considéré. Il n'est pas nécessairement identique pour NO<sub>2</sub> et O<sub>3</sub>, par exemple. La recherche du support "optimal" est laissée en suspens dans la présente étude.

## **2. Finalisation de l'étude variographique exploratoire**

Nous complétons l'étude CG en détaillant certains points : précisions sur les coordonnées des variables auxiliaires, calcul de régressions empiriques notamment.

### **2.1. Rappels sur les données**

#### **2.1.1. Coordonnées**

##### **Occupation des sols**

L'écriture des coordonnées des valeurs d'occupation des sols, à maille 200m, n'est pas systématique, comme le montrent les coordonnées y de points successifs le long d'une colonne (d'abscisse  $x=379900.00m$ ) :

5295500.00, 5295699.99, 5295900.00, 5296100.00, 5296300.00, 5296500.00, 5296699.99, 5296900.00

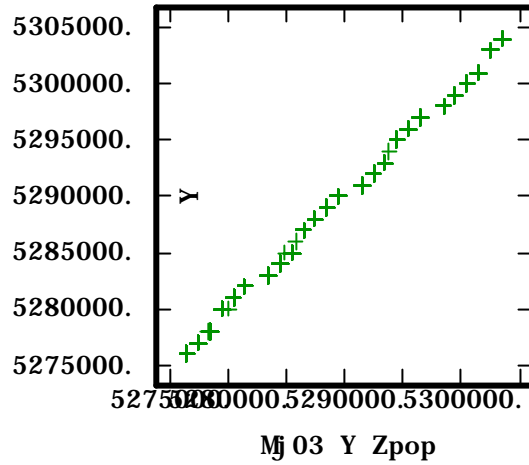
Ces différences d'arrondi sont sans incidence sur les calculs de variogramme et sur le krigeage, mais peuvent provoquer des décalages d'une maille lorsque les données sont transposées sans précaution sur une grille. Or certaines corrélations entre concentrations et variables auxiliaires se révèlent ici sensibles au décalage d'une maille des variables sur la grille.

Remarque pratique dans ISATIS : pour le calcul d'un histogramme, les classes sont définies par leur borne inférieure ; pour la construction d'une grille, les "blocs" sont centrés aux noeuds de la grille.

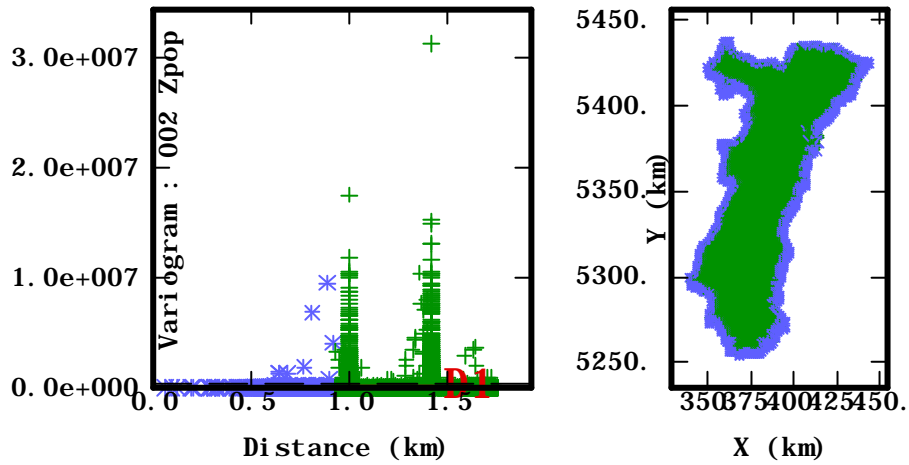
##### **Densité de population : irrégularités locales de la "grille"**

La densité de population est fournie à maille kilométrique, mais avec quelques irrégularités de coordonnées, principalement sur l'ordonnée, mises en évidence par la nuée variographique à petite distance. La nuée variographique est le nuage de corrélation, au facteur 1/2 près, du carré des accroissements de la variable en fonction de l'interdistance des points de mesure. C'est donc un outil commode pour vérifier la présence de données rapprochées (figure 1.) ; les lacunes ne sont pas détectables de cette façon, mais par exemple par les histogrammes locaux sur les coordonnées. Les irrégularités de localisation de ces données se trouvent ici principalement sur le pourtour de la zone informée, ainsi que localement, à l'Est.

Pour la région de Mulhouse notamment, les coordonnées présentent de légères irrégularités qui peuvent provoquer des artefacts : si la "migration" de la densité de population pour informer les tubes utilise sans précaution une grille intermédiaire, des décalages sensibles apparaissent. On observe que l'incidence de ces approximations concernant les coordonnées sur les corrélations avec la concentration en NO<sub>2</sub> n'est pas négligeable.

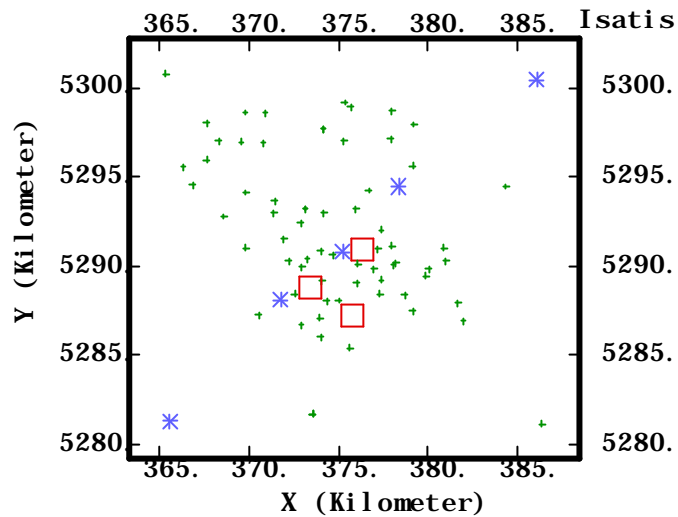


a) densité de population : nuage de corrélation des ordonnées initiales et après migration sur une grille régulière.



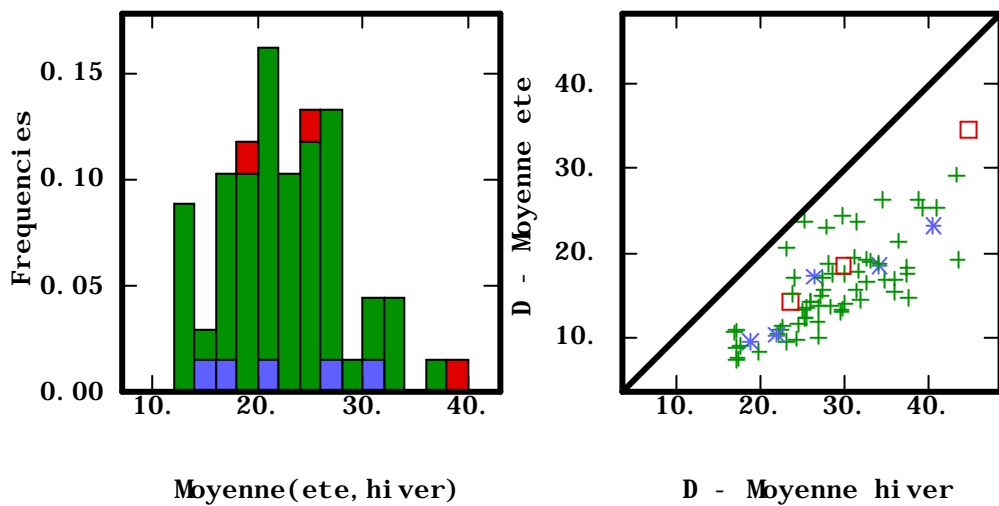
b) Densité de population : nuée variographique à petites distances et implantation des données resserrées.

**Figure 1. Coordonnées des variables explicatives**



+ 1 tube (67 stations), \* 2 tubes (5 stations), h 3 tubes (stations à proximité des analyseurs) ; total : 75 stations.

a) implantation des tubes



b) Histogramme des moyennes “annuelles” et nuage de corrélation des moyennes saisonnières, avec indication des stations multi-tubes.

**Figure 2. Mulhouse. Nombre de tubes par station**

### **2.1.2. Dates de mesures**

75 “stations” sont équipées de tubes, parmi lesquelles 67 comportent 1 tube, 5 en comportent deux et trois (à proximité des analyseurs) en comportent 3. Dans le cas de plusieurs tubes, seule la moyenne de leurs mesures a été fournie. Les cinq stations à deux tubes sont disposées le long d’une ligne d’orientation sud-ouest, nord-est ; les trois stations à trois tubes sont implantées au centre. Dans l’étude CG, seules étaient retenues les stations à un tube (voir les commentaires en début du rapport CG). Pour établir les régressions sur un plus grand nombre de points, les régressions correspondant à des moyennes par classes, nous tenons compte ici de tous les tubes. Les résultats diffèrent donc de ceux présentés dans “*l’examen préliminaire des données de Mulhouse*”. On vérifie que les stations comportant plusieurs tubes recouvrent l’ensemble des classes de concentration.

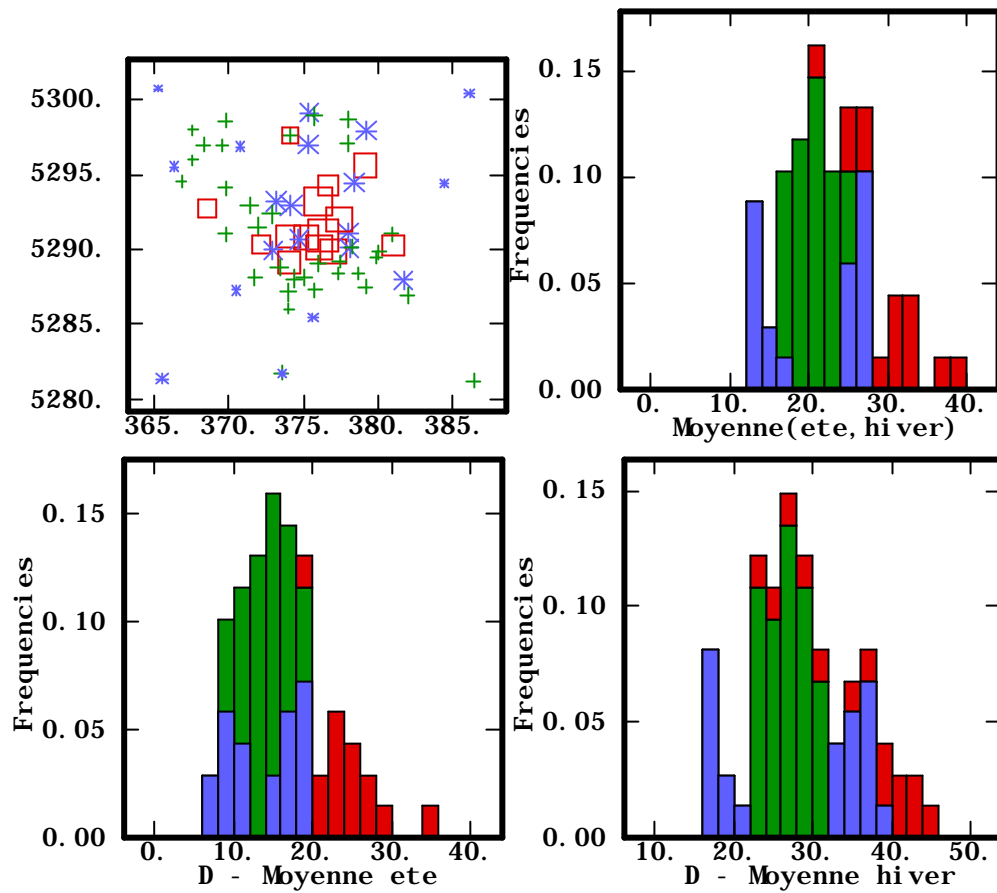
Les mesures hivernales sont disponibles en 71 stations pour chacune des trois quinzaines, mais seulement 64 stations sont communes à ces trois quinzaines ; les mesures estivales sont disponibles respectivement en 69, 69 et 68 stations, 62 stations étant communes aux trois quinzaines. 52 stations disposent des 6 mesures (trois hivernales, trois estivales), parmi lesquelles 44 stations comportent un seul tube.

Les moyennes saisonnières fournies correspondent à une moyenne sur au moins deux quinzaines, et la moyenne annuelle, aux 68 stations ainsi informées simultanément en hiver et en été.

Pour les calculs de variabilité, travailler avec une population homogène nécessiterait de retenir les 44 stations monotube informées pour les six quinzaines. Le nombre de données est alors réduit, et les évaluations de variance deviennent moins fiables. L’alternative est la suivante : soit la population est homogène, mais avec un effectif réduit ; soit l’effectif augmente, la population devenant hétérogène. Il est donc important de vérifier la stabilité des résultats par rapport à la population retenue. Nous conservons dans la suite les stations avec plusieurs tubes, et comparons les résultats aux moyennes saisonnières calculées sur trois quinzaines d’une part, et sur deux quinzaines ou plus, d’autre part. Finalement, les essais de validation croisée sont effectués sur les moyennes sur trois quinzaines, mesurées sur un ou plusieurs tubes.

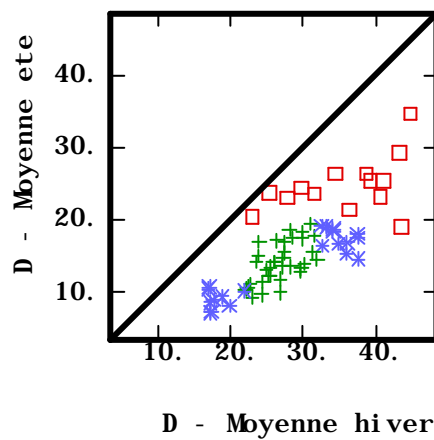
### **2.2. Statistiques élémentaires de NO<sub>2</sub> mesuré aux tubes**

Les moyennes hivernales sont supérieures aux moyennes estivales, avec 29 mg/m<sup>3</sup> et 16 mg/m<sup>3</sup> respectivement, le coefficient de variation étant inférieur en hiver. Les histogrammes, sensibles à l’irrégularité de l’implantation des stations, présentent plusieurs modes : deux pour la moyenne estivale, et trois pour la moyenne hivernale (figure 3.). Les valeurs fortes se situent plutôt au nord de la partie centrale, et les valeurs plus faibles en périphérie. Si les valeurs faibles correspondent aux mêmes stations en été et en hiver, les fortes concentrations estivales correspondent aux fortes concentrations hivernales, mais aussi à des valeurs moyennes. Réciproquement, le mode des fortes concentrations hivernales correspond aux plus fortes concentrations estivales (mode rouge et valeurs intermédiaires fortes en bleu). En résumé : si une concentration est forte en hiver, il en est de même en été ; mais certaines fortes concentrations estivales coïncident avec des concentrations hivernales intermédiaires.



Rouge : mode des fortes concentrations estivales. Bleu : mode des faibles concentrations hivernales et valeurs basses du mode des fortes concentrations hivernales (les deux ensembles se correspondent sur les différents histogrammes et la dimension du symbole \* permet le repérage sur la carte d'implantation). Vert : concentrations intermédiaires.

Noter les deux couples de tubes proches, correspondant à la superposition de deux symboles (carré et croix au Nord, étoile et croix au Sud).



**Figure 3. Histogramme et corrélation des concentrations saisonnières en NO2.**

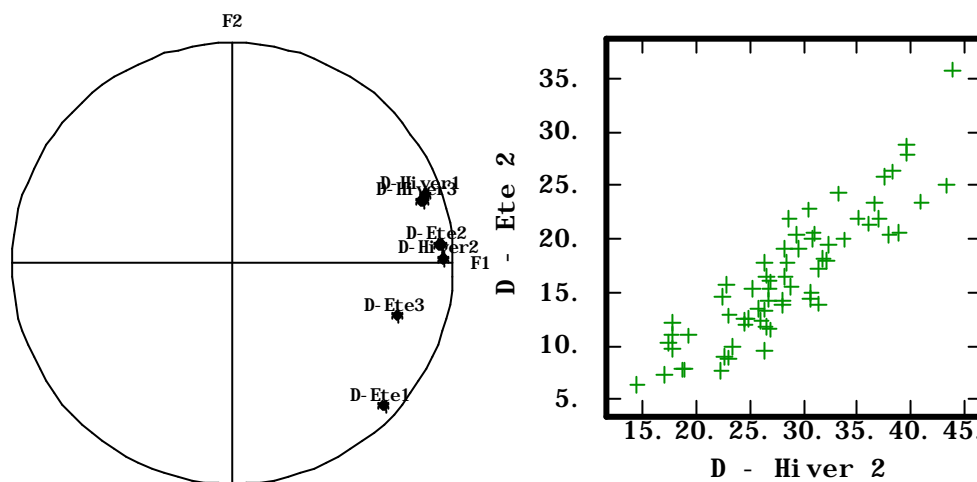
moyenne	nombre	minimum	maximum	moyenne	variance	coefficient de variation
hivernale	74	16.98	44.70	28.97	46.04	0.23
	68	16.98	44.70	28.86	47.29	0.24
	64	16.98	44.70	28.68	48.62	0.24
estivale	69	7.25	34.64	16.18	30.76	0.34
	68	7.25	34.64	16.21	31.14	0.34
	62	7.25	34.64	16.13	30.66	0.34
annuelle	68	12.23	39.67	22.53	34.49	0.26
	52	12.23	39.67	22.26	38.19	0.28

a) Sensibilité des statistiques aux ensembles de tubes retenus : deux lignes supérieures, au moins deux quinzaines informées par saison ; ligne inférieure : trois quinzaines informées.

plus de 2 quinzaines : 68 stations 0.77	3 quinzaines : 52 stations 0.80
--	------------------------------------

b) Coefficients de corrélation des moyennes hivernales et estivales

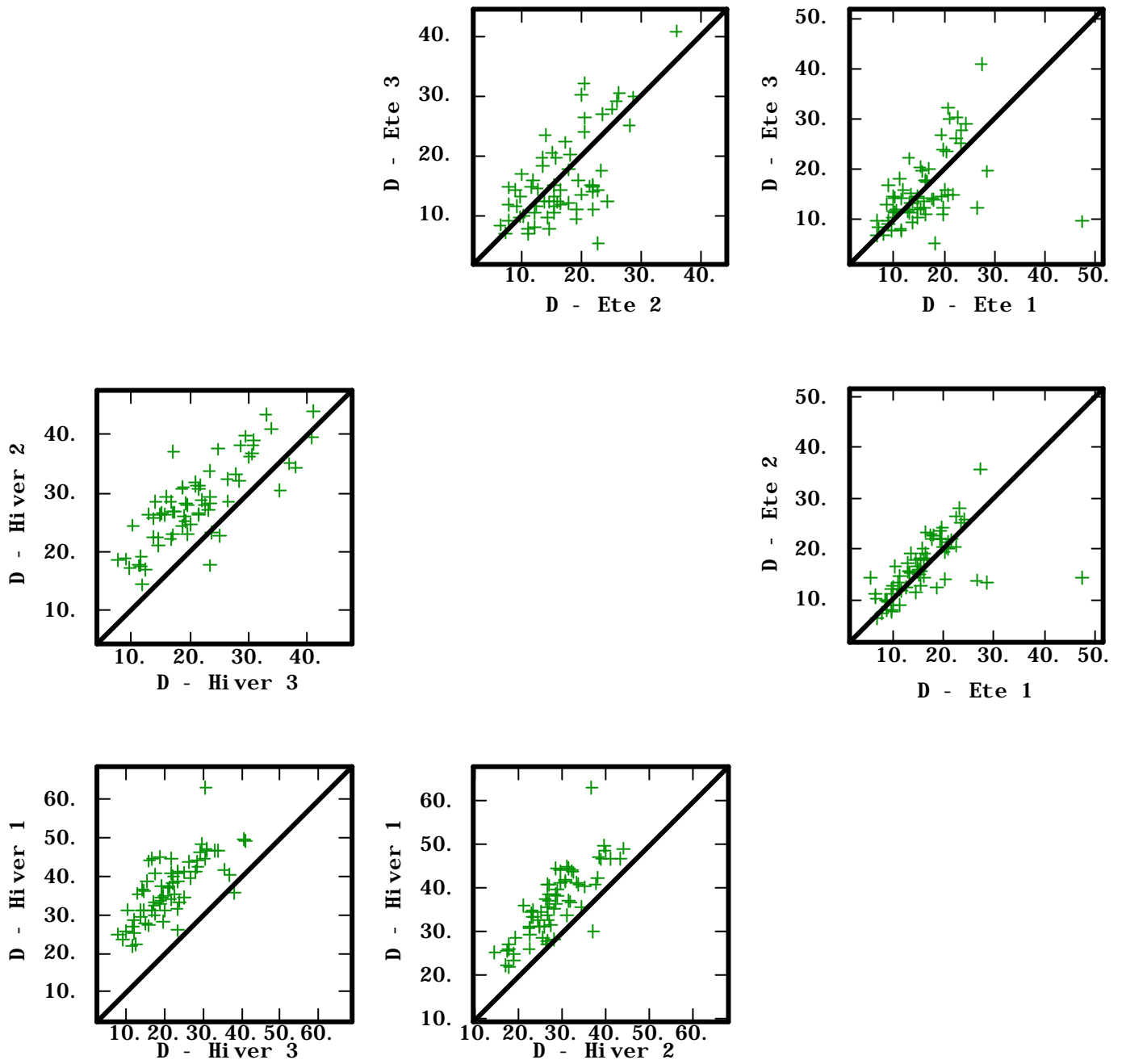
**Tableau 1. Statistiques élémentaires des concentrations en NO<sub>2</sub> (tubes)**



a) cercle de corrélation des facteurs 1 et 2

b) nuage de corrélation des moyennes hivernale-2 et estivale-2

**Figure 4. ACP des moyennes par quinzaines**



La droite indique la première bissectrice.

**Figure 5. Nuage de corrélation entre les moyennes par quinzaine, respectivement hivernales et estivales**

	hiver 1	hiver 2	hiver 3	été 1	été 2
hiver 2	<b>.81</b> (68)				
hiver 3	.73 (67)	<b>.80</b> (67)			
été 1	.40 (65)	.62 (65)	.48 (65)		
été 2	.78 (65)	<b>.89</b> (65)	<b>.82</b> (66)	.58 (66)	
été 3	.56 (64)	.67 (64)	.48 (64)	.46 (64)	.65 (63)

Le chiffre du bas indique l'effectif. En gras, les coefficients supérieurs à 0.8

**Tableau 2. Coefficient de corrélation des moyennes NO2 tubes par quinzaine**

Les différents ensembles apparaissant sur le nuage de corrélation des moyennes saisonnières correspondent aux modes des histogrammes. Globalement, ce nuage est linéaire, avec un coefficient de corrélation de 0.77 .

Les statistiques élémentaires semblent peu sensibles au nombre de stations, selon que trois quinzaines ou seulement deux sont informées (figure 3. et tableau 1.).

L'Analyse en Composantes Principales, effectuée sur les moyennes par quinzaine, centrées et normées, indique un regroupement des quinzaines hivernales, contrairement aux estivales. La deuxième quinzaine hivernale diffère un peu des deux autres, de même que la deuxième quinzaine d'été, plus proche de la deuxième quinzaine hivernale que des deux autres quinzaines estivales. Ces résultats sont confirmés par les coefficients de corrélation, assez élevés entre quinzaines hivernales, médiocres entre quinzaines estivales, et curieusement assez élevés entre la deuxième quinzaine estivale et les valeurs hivernales (tableau 2. et figures 4. et 5.).

A posteriori, ces résultats montrent l'intérêt de mesures sur plusieurs quinzaines successives, en vue de couvrir différentes situations. Il serait intéressant de relier ces situations aux conditions météorologiques par exemple, pour expliquer la typologie des quinzaines.

Ces ACP sont effectuées à titre exploratoire, les concentrations présentant une certaine corrélation spatiale et les stations étant implantées de façon irrégulière. On s'écarte par conséquent de l'hypothèse classique d'indépendance des échantillons.

### 2.3. Examen des tubes rapprochés

La carte d'implantation des stations indique deux tubes distants de 3.1m, mesurés durant les six quinzaines, ainsi que deux tubes distants de 3.7m, mesurés durant les trois quinzaines estivales.

Les coordonnées, ainsi que les concentrations mesurées, sont reportées au tableau 3.

Le couple mesuré durant l'été seulement présente, à une distance inférieure à 4m, une variation du simple au double pour la moyenne sur trois quinzaine. Par quinzaine, l'écart entre les deux mesures reste important. Ce couple de stations a-t-il été implanté précisément pour mesurer un contraste présumé important ?

Le couple situé au Nord présente des écarts importants, mais opposés entre été et hiver, fournissant des moyennes annuelles proches. L'amplitude des écarts varie beaucoup selon les quinzaines. Là encore, l'implantation de ce couple de tubes est-elle préférentielle ou non? Ces écarts sont-ils représentatifs d'erreurs de mesures, ou bien de variations à très petites distances ?

<b>interdistance</b>	<b>3.1m</b>		<b>3.7m</b>	
<b>coordonnées</b>	x=374 127.9 y=5 297 667.3	x=374 127.8 y=5 297 664.2	x=373 518.6 y= 5 281 734.1	x=373 516.5 y= 5 281 731.0
<b>NO2 hivernal</b>	23.06	29.69		
<b>NO2 estival</b>	20.52	13.34	7.25	15.12
<b>NO2 annuel</b>	21.79	21.51		

**Tableau 3. Concentrations mesurées sur des tubes proches.**

## 2.4. Corrélations entre NO<sub>2</sub>- tubes et les variables auxiliaires

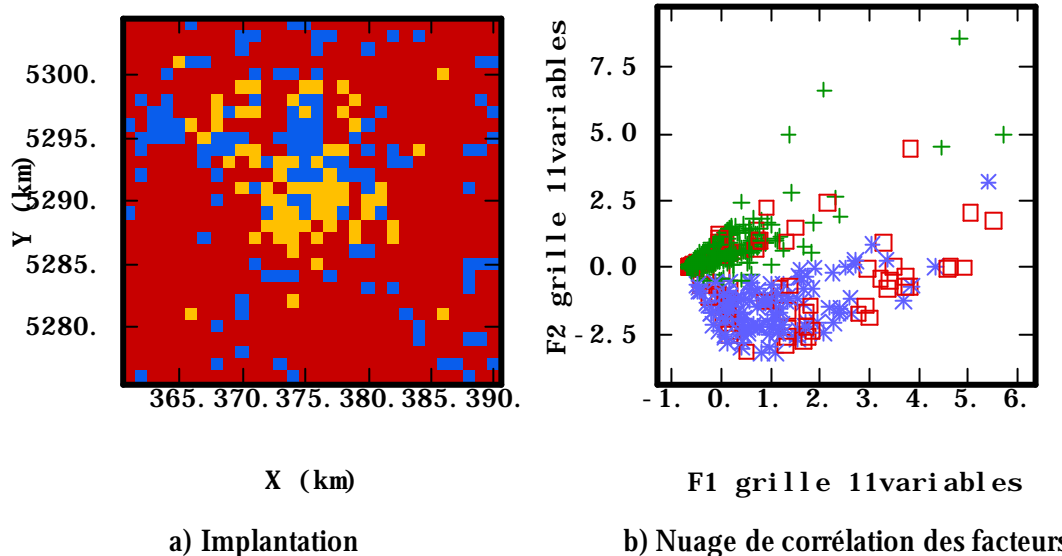
La stabilité des facteurs obtenus par les ACP, selon que celle-ci est effectuée sur les stations ou sur la grille à estimer, a été vérifiée pour la grille kilométrique, de la façon suivante :

- calcul du logarithme translaté des variables sur la grille, le facteur étant choisi identique à celui retenu pour les stations ; ACP sur ces variables
- repérage des noeuds les plus proches des stations (69 noeuds) ; ACP des variables pour ces seuls noeuds.

Même si les représentations des variables dans les cercles de corrélation diffèrent un peu entre grille et station, les nuages de corrélation des facteurs homologues, calculés sur la grille ou pour les seules stations, présentent le plus souvent une forte corrélation linéaire, une station, située au Nord-Ouest, faisant parfois exception.

L'ACP aux stations est utilisée dans la suite pour examiner les relations avec les concentrations en NO<sub>2</sub>. **La comparaison, aux stations, des facteurs des deux ACP permet de vérifier si les relations entre variables auxiliaires diffèrent sensiblement ou non entre les tubes et l'ensemble de la zone à estimer.** Si ces relations sont différentes, la validité du modèle calé aux tubes n'est nullement garantie pour l'ensemble de la zone à estimer.

On vérifie par ailleurs que les valeurs aux noeuds de grille proches des tubes recouvrent assez largement l'ensemble du nuage de corrélation des facteurs, mais ne constituent pas une discrétisation de ce nuage.



11 variables : logarithmes translatsés de la densité de population, de 3 occupations des sols (bâti lâche, bâti dense et industrie, et de 7 émissions). Les noeuds proches des stations sont en orange sur la carte d'implantation, et en rouge (carrés) sur le nuage de corrélation.

Le nuage, qui n'a pas été étudié de façon détaillée, indique différentes typologies des relations entre variables auxiliaires. L'ensemble repéré en bleu sur le nuage de corrélation est reporté, en bleu également, sur la carte d'implantation.

**Figure 6. Nuage de corrélation des deux premiers facteurs d'ACP sur les données de la grille kilométrique autour de Mulhouse.**

### 2.4.1. Occupation des sols

Les études antérieures ont montré que seules les trois variables : bâti dense, bâti lâche et industrie présentent des corrélations utilisables avec la concentration en NO<sub>2</sub> (cf. figures 11 et 22 du rapport CG, par exemple). L'ACP de ces trois variables, informées aux tubes par migration par le plus proche voisin, montre un premier facteur (48% de la variance) lié au bâti dense et à l'industrie, et un deuxième facteur, lié au bâti lâche (36% de la variance). L'ajout des moyennes saisonnières (figure 7.) montre que le bâti dense est corrélé aux moyennes saisonnières en NO<sub>2</sub>, cette corrélation étant plus marquée pour la moyenne estivale qu'hivernale (tableau 4.).

Le premier facteur de l'ACP des trois occupations de sols améliore légèrement les corrélations avec NO<sub>2</sub> par rapport à la variable "bâti dense", de même que la transformation logarithmique (il s'agit ici du logarithme après division par la moyenne, calculée sur une grille au voisinage de Mulhouse, puis translation de 1).

Les facteurs de l'ACP sont modifiés si celle-ci est effectuée sur le logarithme des variables du bâti, le premier facteur contribuant alors pour 64% à la variance totale. Utiliser le facteur de l'ACP des transformées logarithmiques comme variable "explicative" améliore peu les résultats par rapport au logarithme lui-même (tableau 4., voir aussi la figure 22 du rapport CG). Par la suite, on retient donc la transformée logarithmique.

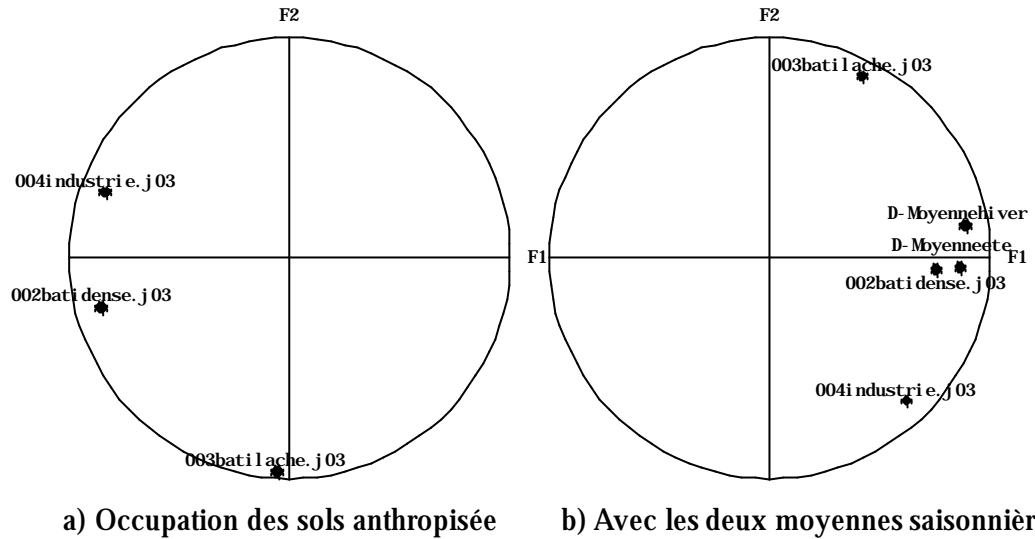
La corrélation entre NO<sub>2</sub> et l'ensemble des variables auxiliaires (bâti, logarithme, facteurs) est sensible aux stations retenues. Elle s'améliore lorsque les moyennes saisonnières sont calculées sur exactement trois quinzaines.

Dans le krigeage avec dérive externe, la relation entre la variable principale et les dérivées est supposée (localement) linéaire ; la transformée logarithmique est alors préférable à la variable bâti dense initiale. En cas de non linéarité marquée entre les deux variables, on améliore le caractère "explicatif" de la variable auxiliaire V en lui substituant une modélisation de la régression expérimentale  $f(V) = E\{Y|V\}$ . La fonction f est calculée par ajustement de la moyenne de la variable principale par classes de valeurs de la variable explicative.

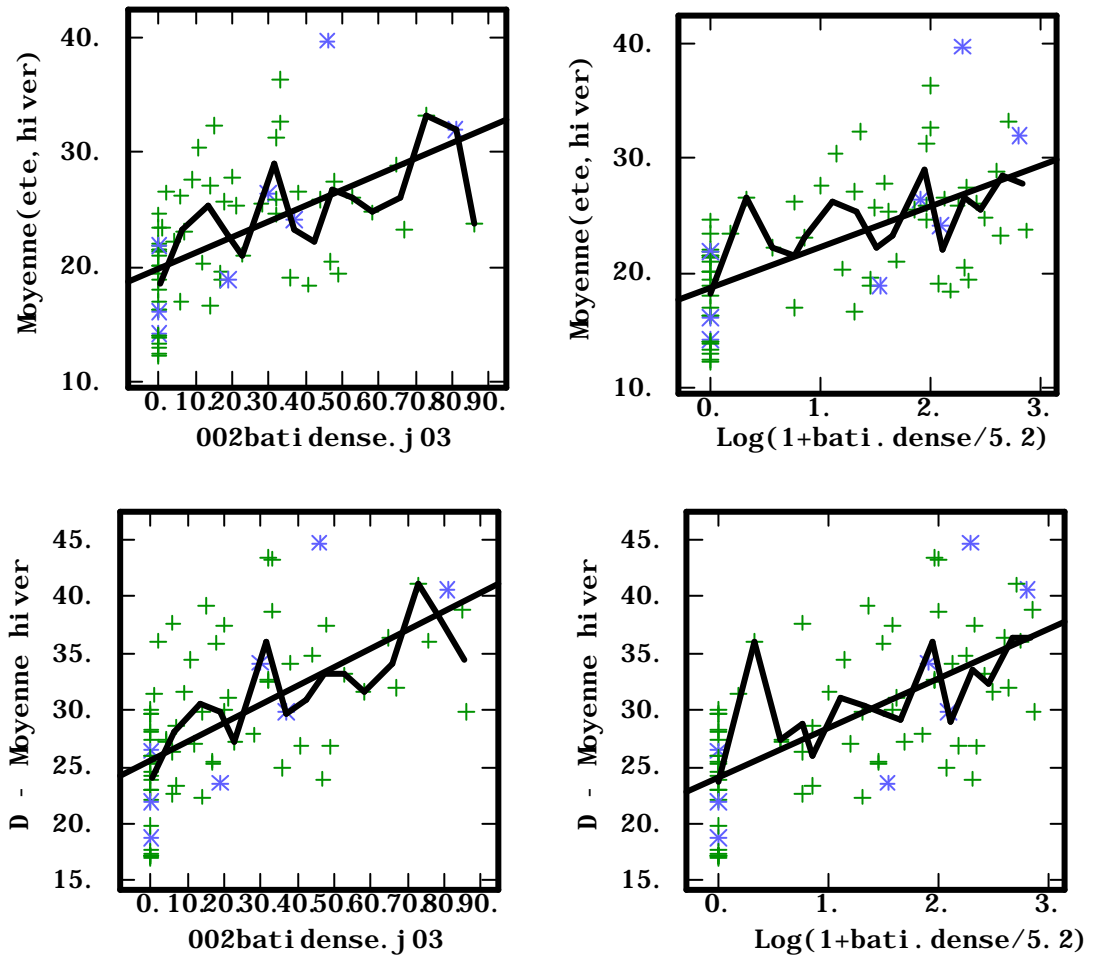
La transformation logarithmique a ici pour effet de rendre plus linéaire la régression entre le bâti lâche et les moyennes saisonnières, en particulier pour les faibles valeurs du bâti (figure 8.). La moyenne par classe, calculée sur les nuages de corrélation, est raisonnablement ajustée par la droite de la régression linéaire des concentrations en NO<sub>2</sub> sur la transformée logarithmique du bâti, sauf peut-être encore aux faibles valeurs.

De façon générale, la démarche est la suivante :

- vérifier la linéarité, ou la non linéarité, de la régression des concentrations sur les variables auxiliaires ou les facteurs de l'ACP.
- si la régression expérimentale est effectivement linéaire, on conserve la ou les variables explicatives (ou les facteurs) les mieux corrélées à la concentration.



**Figure 7. ACP de l'occupation des sols migrée aux tubes (migration par le plus proche voisin). 68 stations.**



Les \* indiquent les stations multi-tubes.

**Figure 8. Corrélation entre la concentration NO2 tubes et le bâti dense.**

moyenneNO2 tubes	nombre	industrie	bâtidense	Log 1 + bâti.dense/5.2	Facteur 1 3occupations de sols	Facteur 1 Log 3 occupations de sols
hiver	68	.36	.56	.62	-.55	.63
	64	.41	.63	.67	-.63	.66
été	68	.49	.44	.51	-.56	.56
	62	.51	.46	.53	-.58	.58
annuel	68	.44	.54	.61	-.59	.64
	52	.50	.60	.66	-.65	.67

*Première ligne, 68 stations : mesure sur deux ou trois quinzaines par saison. seconde ligne : mesures sur trois quinzaines par saison.*

**Tableau 4. Coefficient de corrélation entre NO2 annuel tubes et l'occupation du sol (migration par le plus proche voisin).**

- sinon, on modélise la régression empirique, calculée si nécessaire sur des classes de même effectif (cf. l'étude de G. L'Hégaret pour l'ASCOPARG; la variable utilisée en dérive externe est alors  $f(V)$ , la fonction  $f$  étant ajustée pour chaque concentration, si plusieurs polluants sont étudiés; ou alors, on cherche une transformation de la variable auxiliaire (par exemple, le logarithme translaté), rendant linéaire la régression empirique.

En pratique, le passage en logarithme "étable" les faibles valeurs de la variable auxiliaire, ce qui se révèle utile si les effectifs correspondants sont importants et les non linéarités marquées pour ces classes de valeurs ; dans ce cas, travailler sur la variable auxiliaire "initiale" par classes iso-effectifs est souvent préférable au calcul par classes de largeur constante.

On montre en statistiques que l'espérance conditionnelle correspond au "meilleur estimateur" point par point, d'où l'intérêt de la modélisation de la régression empirique.

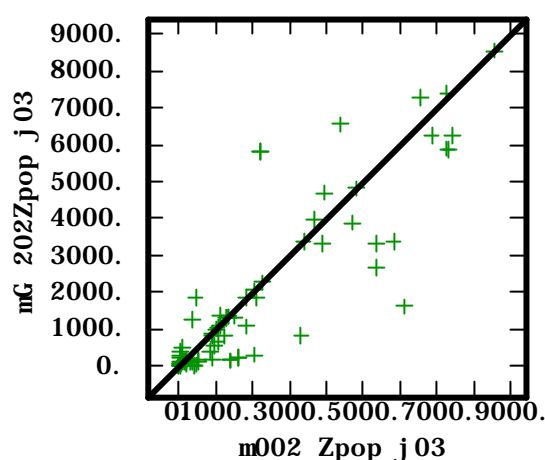
Une "bonne" variable auxiliaire (modélisation de la régression empirique calculée sur des classes de largeur variable, ou facteur de l'ACP des variables transformées) est à rechercher au cas par cas, en tenant compte de la variabilité des concentrations autour des moyennes par classes.

## 2.4.2. Densité de population

La densité de population est fournie sur un réseau quasi-régulier, à maille kilométrique. Les irrégularités de coordonnées pour la densité de population ont une influence sensible sur les corrélations, selon que les valeurs aux tubes sont informées à partir des valeurs initiales de la densité de population, ou après calage sur une grille intermédiaire de maille kilométrique (voir figure 9. le nuage de corrélation de la migration de densité de population aux tubes, avec ou sans passage par une grille intermédiaire). Pour éviter de cumuler des approximations successives, les résultats sont présentés sans passage par une grille intermédiaire, contrairement au rapport CG.

La densité de population étant fournie à maille kilométrique, la méthode retenue pour informer les tubes a une forte influence sur les corrélations. Nous examinons les cas suivants :

- migration aux tubes de la densité ou de son logarithme translaté  $\text{Log}(1 + Z_{\text{pop}}/375)$ , ce qui correspond à l'estimation par le noeud de "grille" le plus proche - la valeur 375 étant la moyenne de la densité de population sur une zone autour de Mulhouse ;
- krigeage ponctuel du logarithme translaté, avec un variogramme isotrope  $(0.05+0.13 \text{ Sphérique}(a=2500\text{m})+1.41 \text{ Sphérique}(a=15000\text{m}))$ , en voisinage glissant de rayon 2500m ;
- par référence à la méthode retenue par l'ASPA pour informer une grille à maille 500m pour la densité de population, estimation en  $1/d^4$  du logarithme translaté, avec un voisinage choisi à 1300m.



N02/Mulhouse tubes

- Variable #1 : m002 Zpop j 03
- Variable #2 : mG 202Zpop j 03

Nb. samples :	75	75
Minimum :	0	0
Maximum :	8544	8544
Mean :	1999.73	1761.73
Std. Dev. :	2312.2	2239.21
Coef of Var. :	1.15626	1.27102

**Figure 9. Nuage de corrélation de la densité de population migrée aux tubes, avec (ordonnée) ou sans (abscisse) passage par une grille intermédiaire.**

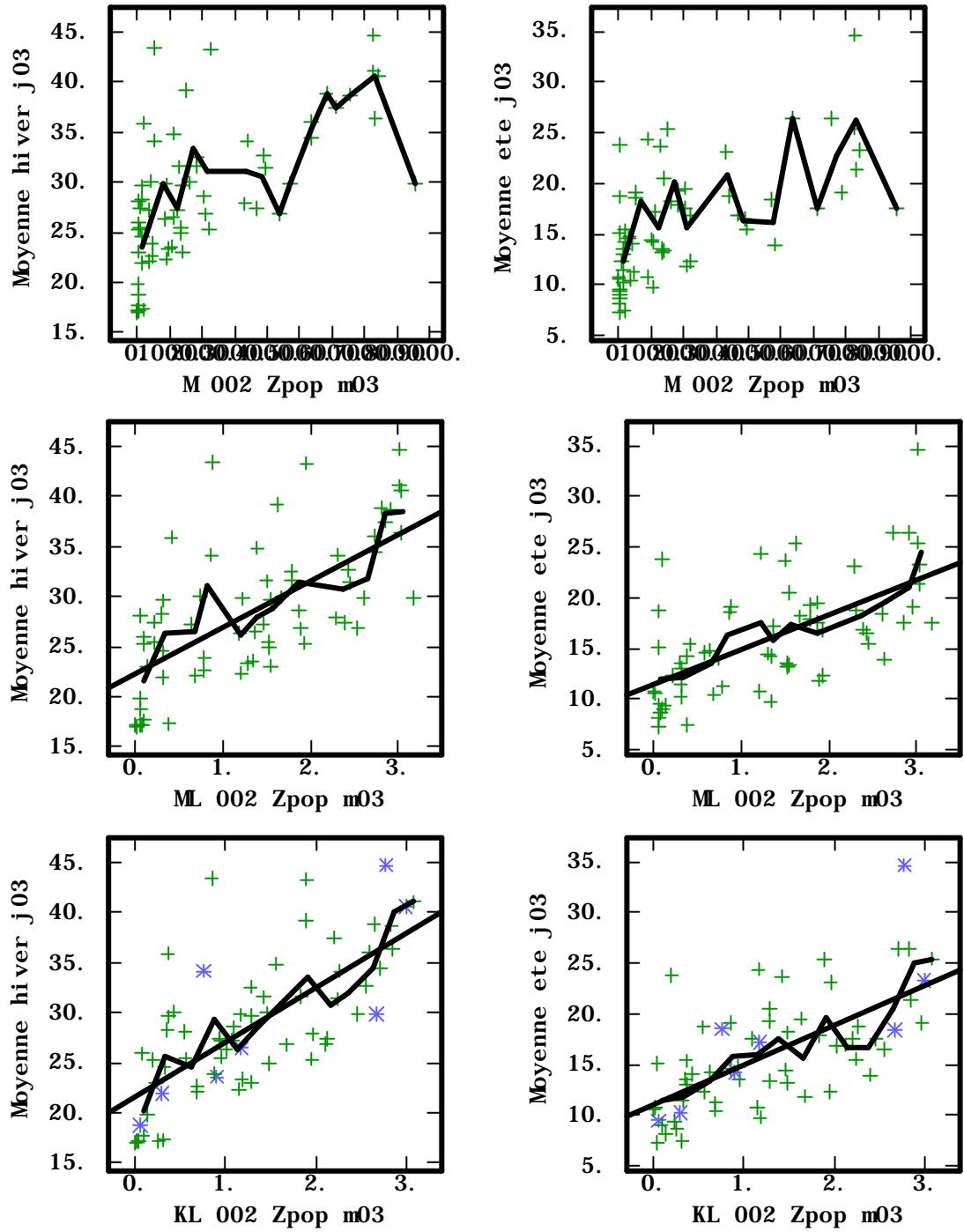
Zpop	moyenne hivernale		moyenne estivale		moyenne annuelle	
	effectif	corrélation	effectif	corrélation	effectif	corrélation
migration Zpop	68	.57	68	.57	68	.61
	64	.64	62	.61	52	.68
migration Log Zpop	68	.63	68	.60	68	.66
	64	.67	62	.63	52	.71
krigeage Log Zpop	68	.53	68	.52	68	.56
	64	.59	62	.55	52	.63
1/d4 Log Zpop	68	.66	68	.62	68	.68
	64	.70	62	.65	52	.74

*Première ligne, 68 stations : mesure sur deux ou trois quinzaines par saison. Seconde ligne : mesures sur trois quinzaines par saison.*

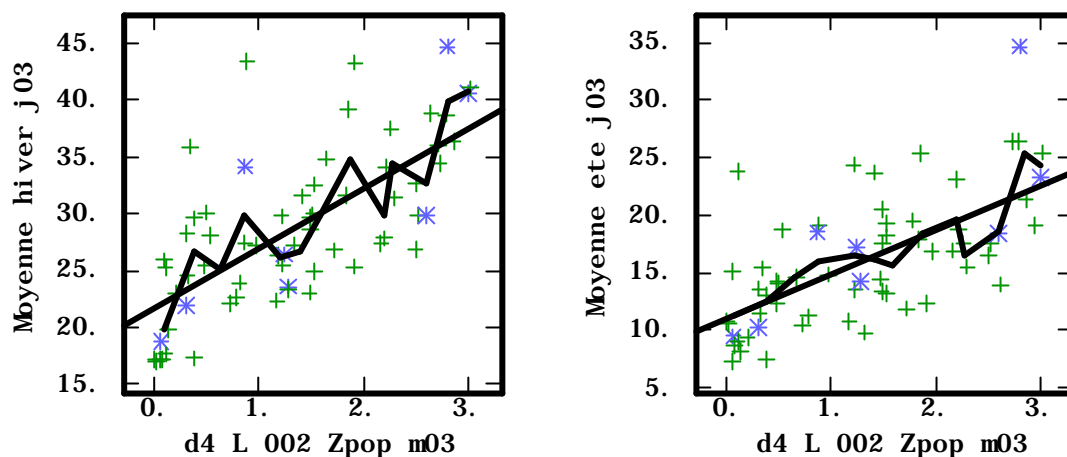
**Tableau 5. Corrélation entre NO<sub>2</sub> et la densité de population.**

Les corrélations sont reportées au tableau 5. Comme pour l'occupation des sols, du fait de la forte dissymétrie de l'histogramme de ces variables auxiliaires, le passage en logarithme améliore les corrélations, et fournit une courbe de régression proche d'une droite (figure 10.). Les coefficients de corrélation sont légèrement meilleurs pour les moyennes calculées sur exactement trois quinzaines, ainsi que pour la moyenne annuelle par rapport aux moyennes saisonnières.

Les valeurs maximales du coefficient de corrélation avec les moyennes saisonnières de NO<sub>2</sub> sont obtenues pour une interpolation aux tubes en  $1/d^4$  du logarithme de la densité de population, le krigeage fournissant les corrélations les plus faibles. Les plus fortes corrélations, par migration ou interpolation en  $1/d^4$  du logarithme de la densité de population, sont analogues à celles vues pour le logarithme du bâti ou pour le premier facteur de l'ACP des logarithmes d'occupation des sols.



**Figure 9 . Nuages de corrélation entre la densité de population et les moyennes saisonnières de NO2 tubes.**



**Figure 10. Nuages de corrélation entre la densité de population, interpolée en  $1/d^4$  aux tubes, et les moyennes saisonnières de NO<sub>2</sub> tubes.**

### 2.4.3. Cadastre des émissions

Les corrélations entre variables du cadastre des émissions et les mesures de NO<sub>2</sub> aux tubes sont sensibles aux écarts de coordonnées, selon que la migration des valeurs aux tubes utilise une grille intermédiaire ou non. Les résultats suivants, obtenus sans grille intermédiaire, diffèrent de ceux présentés dans le rapport CG.

Les histogrammes des sept variables du cadastre des émissions étant très fortement dissymétriques, on en prend le logarithme translaté ; le facteur de “normation” est ici choisi proche de la moyenne de la variable migrée sur les 75 stations (et non plus comme la moyenne sur une zone autour de Mulhouse). Nous n’avons pas cherché à optimiser ce facteur.

Le facteur est le suivant :

NOx	C6H6	COV.NM	particules	SO <sub>2</sub>	CO	BAP
20 000.	500.	45 000.	2 500.	10 000.	125 000.	700.

Une forte corrélation entre les émissions a été constatée dans l’étude CG (Figure 17 du rapport CG). L’ACP avait mis en évidence une corrélation entre ces variables et les moyennes saisonnières de NO<sub>2</sub> (figure 22). Travaillant sur les variables migrées aux stations, nous examinons si les facteurs de l’ACP améliorent les corrélations par rapport aux variables cadastrales. L’ACP est effectuée après migration des variables aux tubes, c’est-à-dire sur 75 données seulement. Les facteurs, et donc les corrélations avec le NO<sub>2</sub> aux tubes, différeraient si cette ACP était menée sur la “grille” du cadastre autour de Mulhouse.

Par rapport à l’étude CG, les différences constatées proviennent d’une part de la migration (sans grille intermédiaire), et d’autre part du nombre de stations, incluant les stations multitubes.

Parmi les émissions, la moins corrélée aux concentrations en NO<sub>2</sub> est SO<sub>2</sub>, alors que le benzène présente des corrélations analogues à NOx. La corrélation avec les mesures de NO<sub>2</sub> aux tubes n’est pas améliorée (ni détériorée) si l’on remplace les émissions les mieux corrélées par le premier facteur de l’ACP. Les résultats obtenus en supprimant SO<sub>2</sub> de l’ACP,

puis en rajoutant la densité de population ne montrent aucune amélioration notable (tableau 6.).

La linéarité des régressions entre les mesures aux tubes et les transformées logarithmiques des variables du cadastre des émissions (ou des facteurs de l'ACP de ces transformées) est satisfaisante (figure 13.).

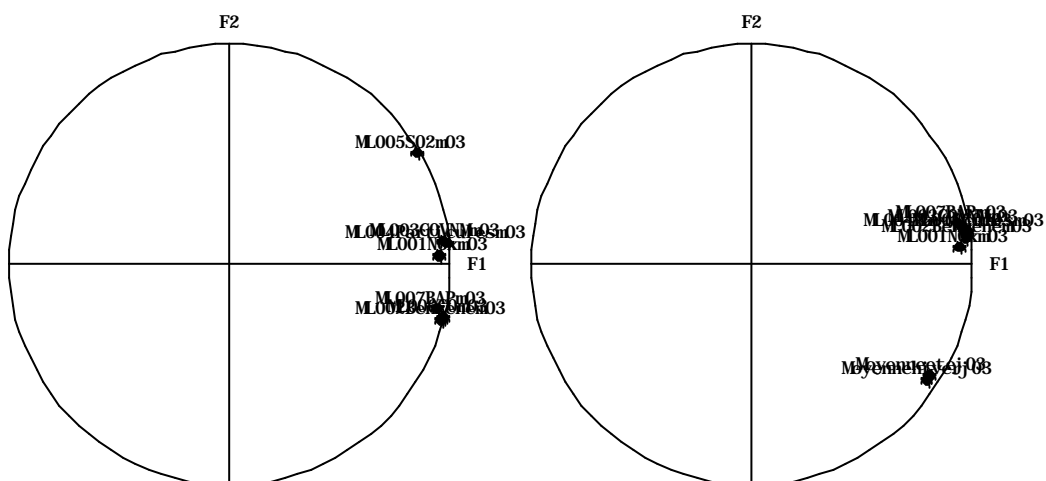
Pour NOx, le krigeage (variogramme :  $0,0715+0.11\text{Sphérique}(ax=10\text{km},ay=6.5\text{km})$ , voisinage restreint à 2500m) améliore peu les résultats (tableau 6.). Ce résultat concorde avec le rapport CG. L'amélioration étant minimale, les calculs ne sont donc pas repris exhaustivement.

Dans la suite, les valeurs des variables auxiliaires aux stations sont obtenues par migration.

	effectif	Log NOx		Log C6H6	F1 Log émissions	F1 Log émission sans SO2
		krigeage	migration	migration	migration	migration
hiver	68	.65	.62	.65	.63	.65
	64	.70	.68	.70	.68	.70
été	68	.69	.69	.68	.67	.68
	62	.69	.69	.68	.68	.69
annuel	68	.71	.69	.70	.69	.70
	52	.75	.74	.75	.74	.75

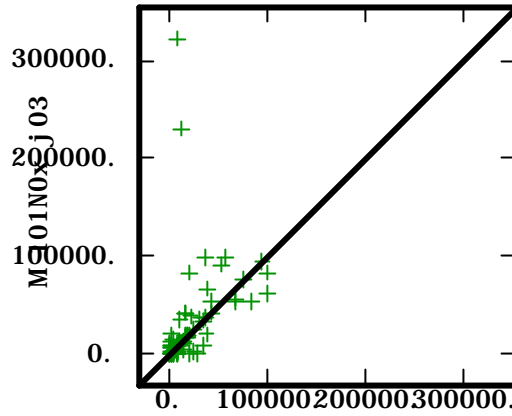
*Première ligne, 68 stations : mesure sur deux ou trois quinzaines par saison. Seconde ligne : mesures sur trois quinzaines par saison.*

**Tableau 6. Coefficients de corrélation des moyennes saisonnières et des logarithmes translatsés des variables du cadastre des émissions.**



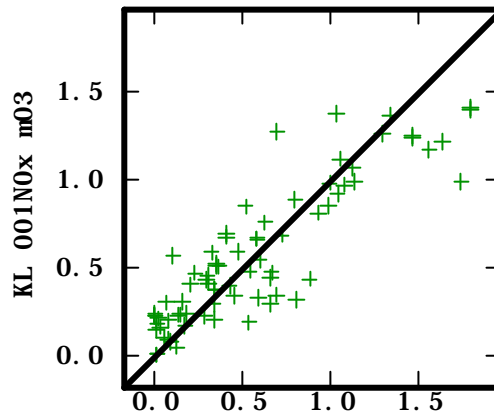
a) ACP des sept émissions      b) émissions hors SO2 et avec moyennes saisonnières

**Figure 11. ACP de NO2- tubes et du cadastre des émissions**



M 001 NOx m03

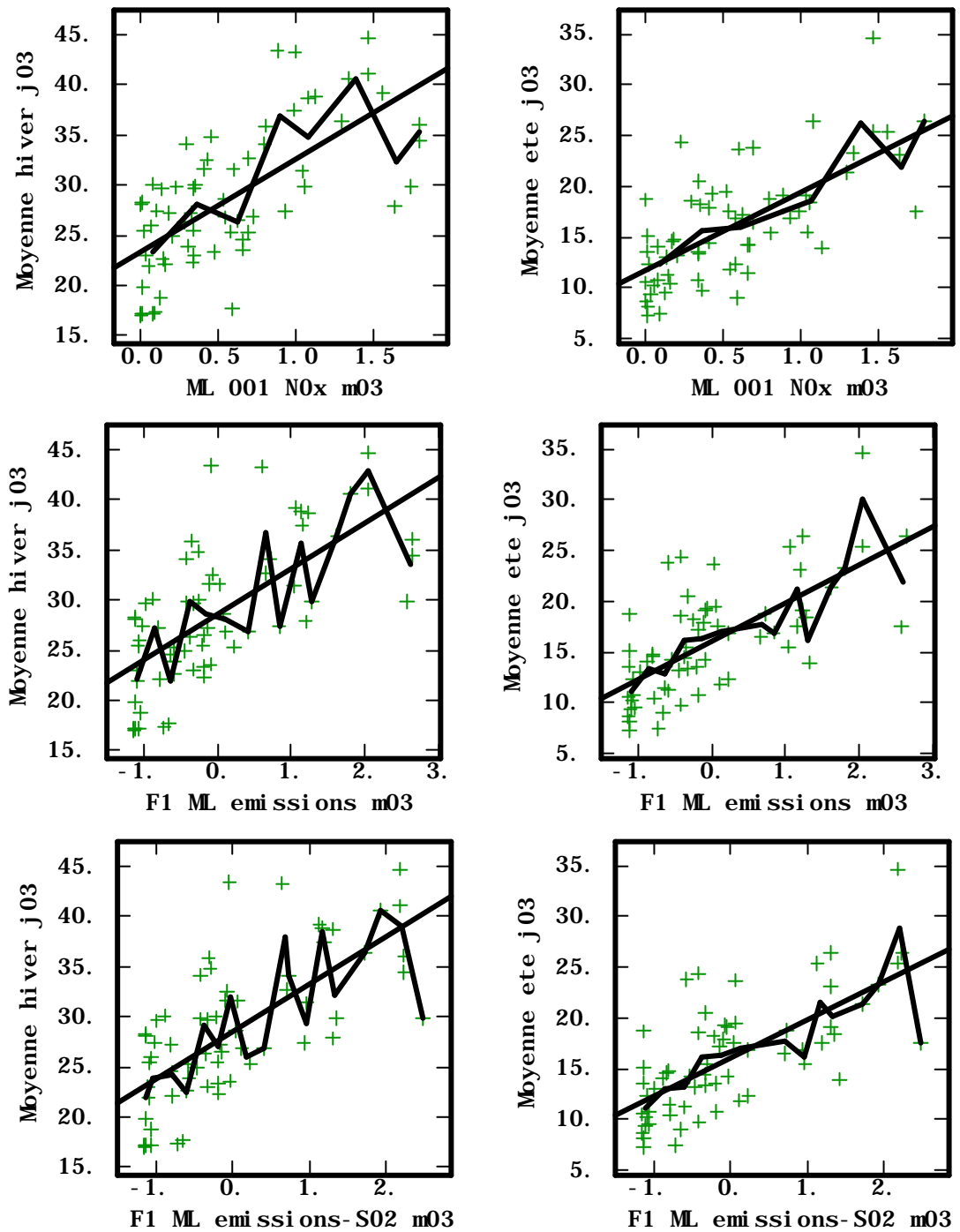
a) migration sans (abscisse) et avec (ordonnée) grille intermédiaire



ML 001 NOx m03

b) Logarithme translaté : migration et krigeage (sans grille intermédiaire)

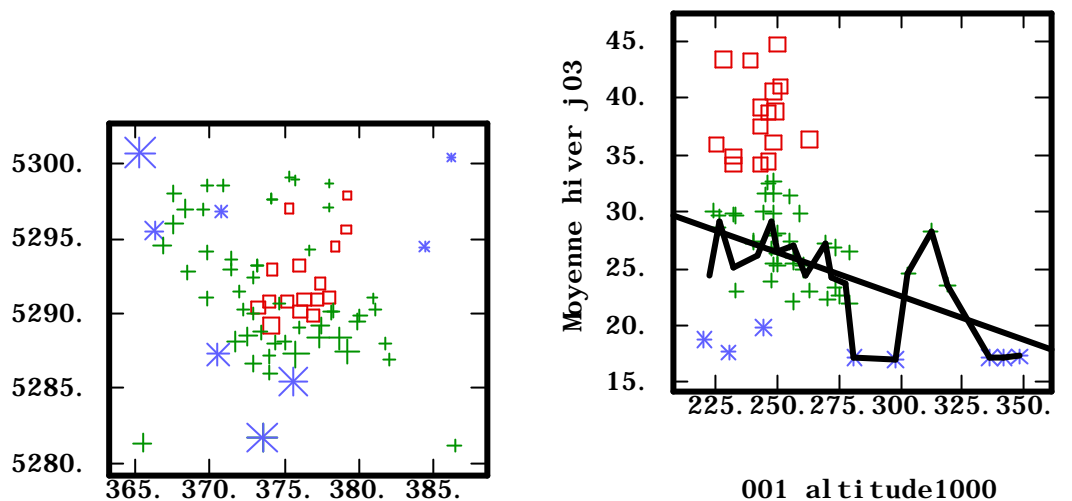
**Figure 12. Nuages de corrélation de NOx interpolé aux tubes**



**Figure 13. Nuage de corrélation des moyennes saisonnières et du cadastre des émissions**

## 2.4.4. Altitude

L'altitude à maille 1000m semble être une régularisée de celle à maille 250m. Les coefficients de corrélation avec le NO<sub>2</sub> mesuré aux tubes, légèrement supérieurs en valeur absolue pour l'altitude à maille 1000m qu'à maille 250m, restent médiocres (-0.5 pour la moyenne hivernale, valeur la plus marquée).



a) Implantation et altitude des tubes

b) Nuage de corrélation altitude-NO<sub>2</sub>

Carrés rouges : fortes valeurs de NO<sub>2</sub>, étoiles bleues : faibles valeurs de NO<sub>2</sub>.

Sur la carte a) la taille des symboles augmente avec l'altitude, qui varie entre 200 et 350m environ.

**Figure 14. Nuage de corrélation de NO<sub>2</sub> mesuré aux tubes et de l'altitude.**

## 2.5. Modélisation

Les ACP, effectuées en première approche aux tubes et non sur l'ensemble de la zone à estimer, ainsi que l'examen des corrélations entre variables auxiliaires (dont les facteurs d'ACP) et les concentrations en NO<sub>2</sub>, indiquent les principales variables explicatives suivantes, précédemment mises en évidence par ASPA (rapport ASPA ou DESS).

- logarithme translaté du bâti, ou facteur associé
- logarithme translaté de la densité de population
- logarithme translaté de NO<sub>x</sub>, ou facteur associé (facteur calculé sur les émissions sans SO<sub>2</sub>)
- altitude.

Ces variables sont partiellement redondantes, comme le montre leur ACP (toujours aux 75 tubes, figure 15.)

De préférence aux facteurs des ACP, nous utilisons les logarithmes translattés des principales variables explicatives, qui restent inchangées si l'on modifie l'ensemble de référence (tubes ou grille d'estimation). Les coefficients de corrélation sont récapitulés au tableau suivant. Si l'on considère le premier facteur de l'ACP (aux tubes) des cinq variables explicatives

(facteur des Log de l'occupation des sols, Log traduit de la densité de population, Log traduit de NOx, facteur des émissions hors SO2, altitude maille 1000m.), les corrélations ont effectivement augmenté, et les nuages indiquent une relation approximativement linéaire entre la concentration NO2 mesurée aux tubes, et ce premier facteur de l'ACP.

A cette étape, nous disposons donc d'une variable auxiliaire synthétique (le facteur de l'ACP des cinq variables explicatives), de trois variables jugées explicatives, assez bien corrélées à la concentration en NO2 (les logarithmes traduits, migrés aux tubes, pour le bâti dense, la densité de population, et le cadastre d'émission de NOx), et de l'altitude, faiblement corrélée. Le facteur synthétique fournit les meilleures corrélations ; cependant, pour l'estimation, comment le recalculer aux noeuds de grille ? Deux solutions :

- effectuer la même combinaison linéaire qu'aux points expérimentaux ; le résultat obtenu ne coïncide donc pas avec le premier facteur d'ACP tel qu'il serait obtenu à partir des noeuds de grille ;

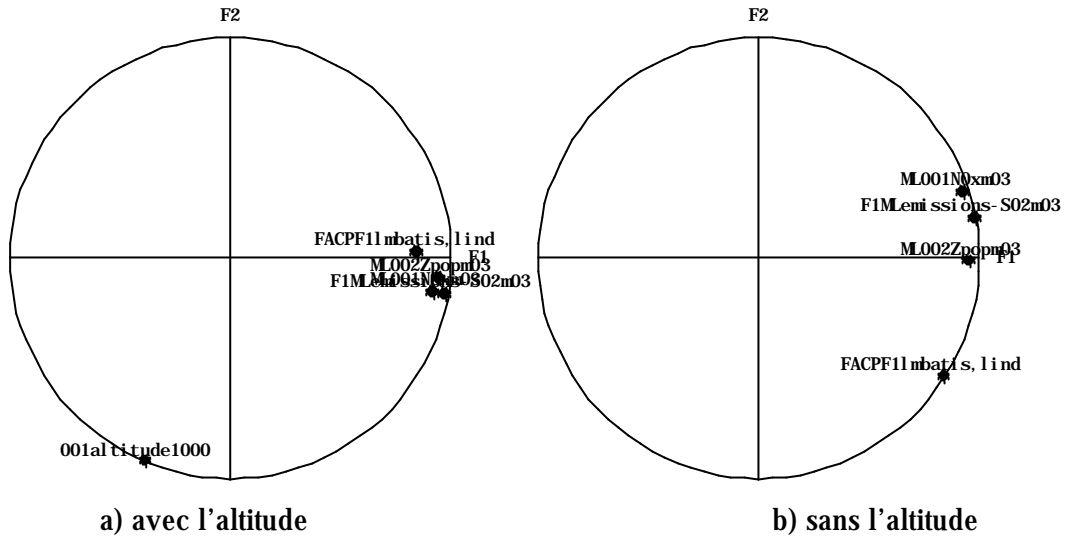
- calculer l'ACP aux noeuds de grille, en déduire (par exemple par migration par le plus proche voisin) la valeur aux points expérimentaux, et reprendre l'étude de corrélation entre le facteur d'ACP et la concentration en NO3 en ces points. Comme pour toute ACP effectuée sur une grille d'estimation, le facteur d'ACP ainsi obtenu dépend de l'extension de la grille.

Une solution est apportée par la méthode de dérive externe (ou son approximation par régression linéaire multiple) dans laquelle les variables explicatives pressenties sont utilisées en dérive, le krigeage se chargeant "d'optimiser localement" la combinaison linéaire des variables explicatives. Il reste alors à choisir le voisinage de krigeage, les variables mises en dérive, et le modèle variographique.

moyenne NO2 tubes	effectif	Facteur 1 5 variables explicatives	Log Zpop	Log Nox	Log bâti dense	altitude maille 1000m
hiver	68	.72	.63	.62	.62	- .51
	64	.76	.67	.68	.67	- .50
été	68	.70	.60	.69	.51	- .39
	62	.72	.63	.69	.53	- .39
annuel	68	.75	.66	.69	.61	- .48
	52	.80	.71	.74	.66	- .48

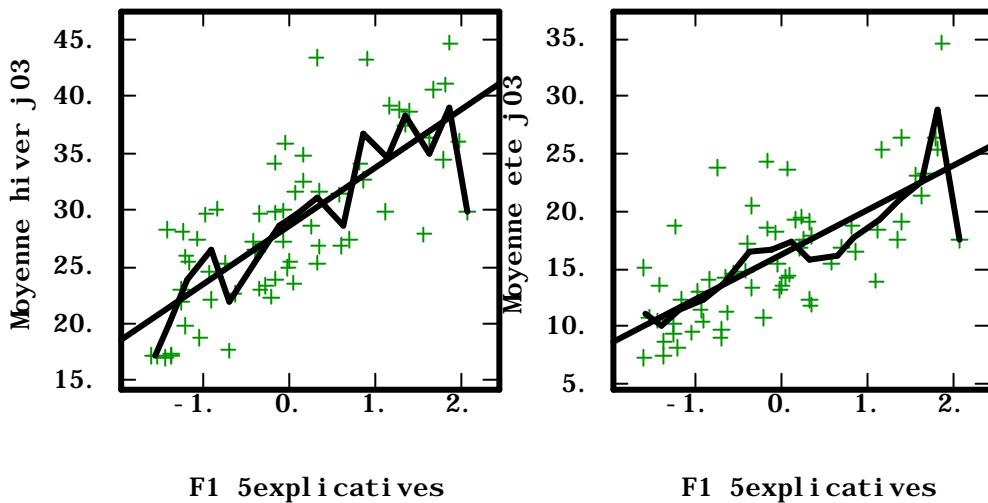
*Première ligne, 68 stations: mesur sur deux ou trois quinzaines par saison. Deuxième ligne: mesures sur trois quinzaines par saison.*

**Tableau 7. Coefficients de corrélation des variables "explicatives" et des moyennes saisonnières.**



Variables (migration par le plus proche voisin, sans grille intermédiaire) : facteur des Log de l'occupation des sols, Log translaté de la densité de population, Log translaté de NOx, facteur des émissions hors SO2, altitude maille 1000m.

**Figure 15. ACP des variables explicative aux tubes, avec et sans l'altitude.**



Variables (migration par le plus proche voisin, sans grille intermédiaire) : facteur des Log de l'occupation des sols, Log translaté de la densité de population, Log translaté de NOx, facteur des émissions hors SO2, altitude maille 1000m.

**Figure 16. Nuage de corrélation des moyennes saisonnières et du premier facteur d'ACP (aux tubes) des cinq variables explicatives.**

### **3. Comparaison d'estimateurs par validation croisée.**

La validation croisée est utilisée pour comparer le krigeage avec différentes variables en dérive (externe). Le calcul est classique : suppression de chaque tube successivement, puis estimation par les mesures des moyennes saisonnières sur les autres tubes, utilisant les variables auxiliaires pour les stations à réestimer.

Les mêmes données sont utilisées pour l'inférence du modèle et pour la validation croisée. Cette inférence est effectuée de manière indirecte, en utilisant les critères proposés dans ISATIS.

#### **3.1. Modèles de référence**

Par essais successifs, nous retenons comme référence un modèle donnant des résultats non optimaux mais satisfaisants pour les trois moyennes (hiver, été, "annuelle"), les critères proposés dans ISATIS pour le choix des dérivées (rang moyen, variance d'erreur d'un ajustement par moindres carrés) divergeant sur le "meilleur" modèle.

La pente du variogramme linéaire est déduite du calage automatique (option d'ajustement non stationnaire, dans ISATIS ; ce facteur intervient sur les statistiques des erreurs "standardisées").

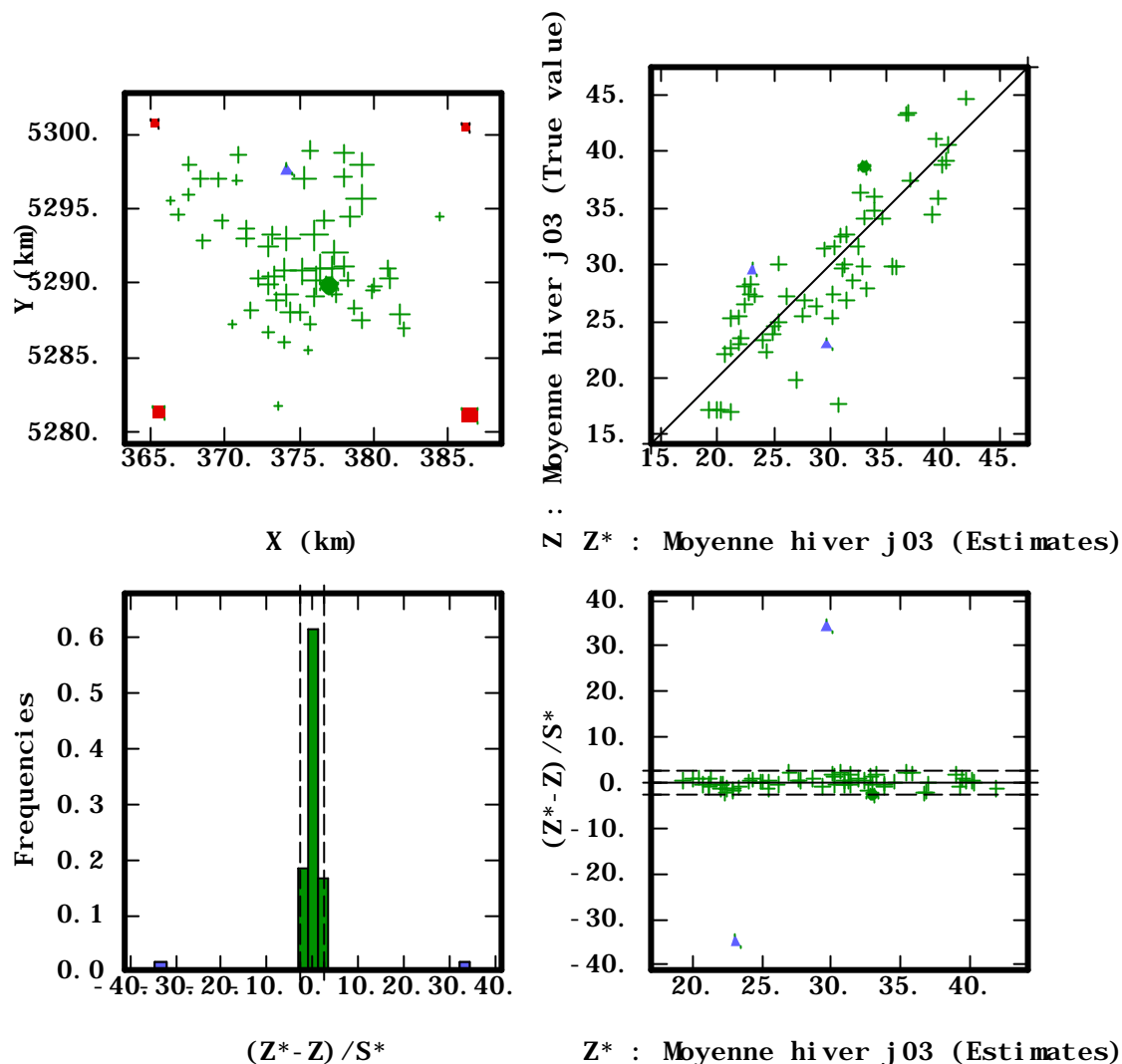
Nous n'avons pas examiné l'amélioration éventuelle des résultats dans le cas de variogrammes gigognes ou anisotropes.

Le modèle de référence est le suivant :

- voisinage : rayon de 10 km (4 secteurs angulaires, 4 points au minimum pour l'estimation, optimum : 12 points par secteur, ce qui correspond à 3 auréoles complètes pour une grille régulière).
- 3 variables en dérive externe, obtenues par migration aux tubes : logarithme translaté de l'émission de NO<sub>x</sub>, de la densité de population, et du bâti dense.
- variogramme linéaire (pentes : hiver 0.006, été et annuel : 0.005).

Pour les tubes distants de 3 mètres environ, donc quasiment confondus à l'échelle de l'agglomération, le poids de krigeage affecté au tube voisin du tube à estimer devient très proche de 1, et la variance de krigeage est très faible (les variogrammes retenus ne comportant pas d'effet de pépite, sauf pour ce seul modèle). Les erreurs d'estimations pour chacun des deux tubes sont alors opposées, car proches de  $Z_1 - Z_2$  et  $Z_2 - Z_1$  respectivement, comme on l'observe sur les figures 17. et suivantes. La variance de krigeage étant très faibles, les erreurs standardisées sont fortes ; ceci est d'autant plus marqué lorsque l'écart entre les deux mesures est important.

Dans la suite, les statistiques pour la moyenne hivernale portent sur 60 tubes, celles sur la moyenne estivale sur 59, et sur 49 tubes pour la moyenne "annuelle", au lieu de 64, 62 et 52 tubes précédemment. Les tubes manquant n'ont pas été réestimés, par manque de données en nombre suffisant dans leur voisinage.



N02/Milhouse tubes  
- Variable #1 : Moyenne hiver j03  
Standard Parameter File for Model: DE. hiver  
Standard Parameter File for Neighborhood: Isotrope rayon 10000  
Cross validation statistics based on 60 test data

	Mean	Variance
Error	0.19361	15.08894
Std. Error	0.01788	40.95236

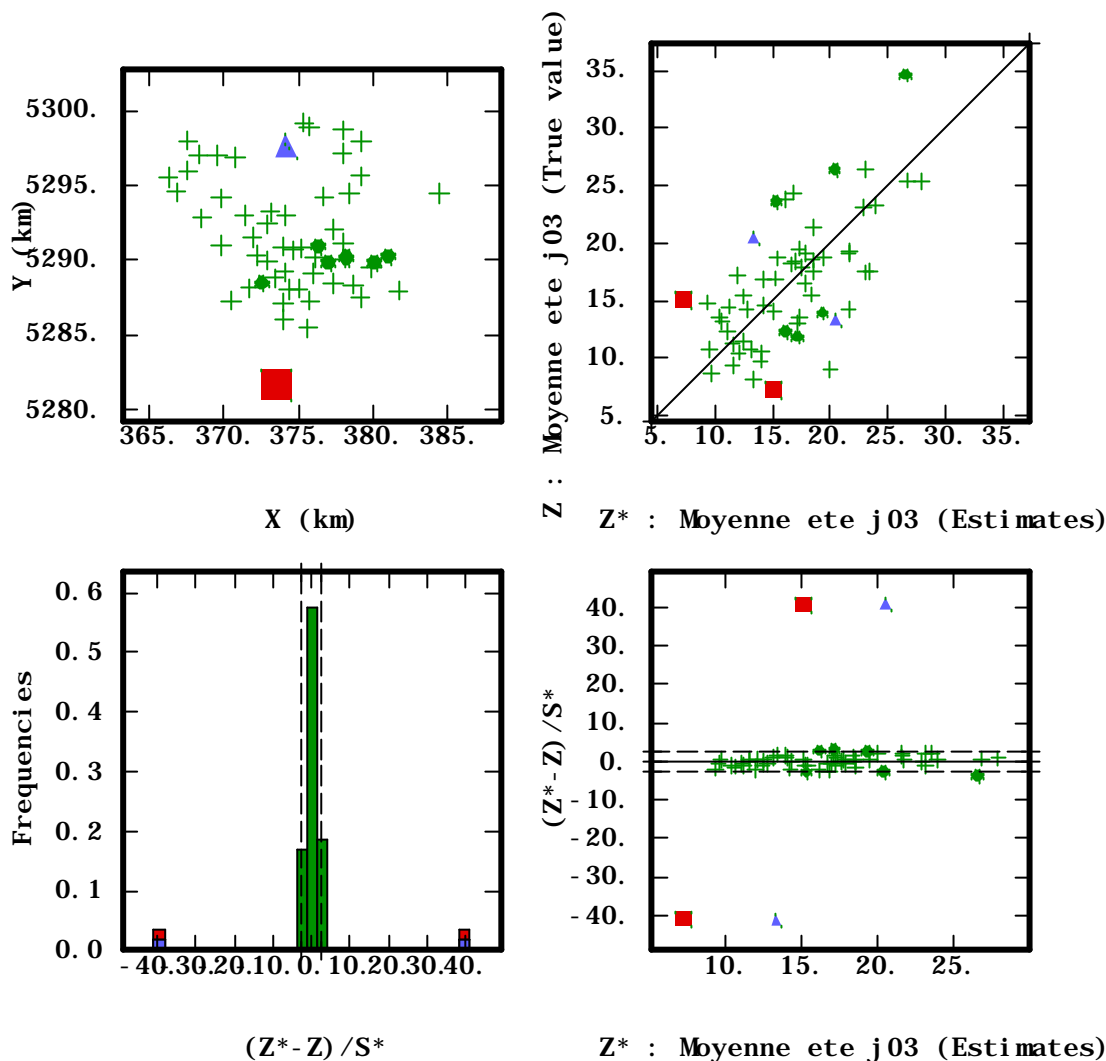
Cross validation statistics based on 57 robust data

	Mean	Variance
Error	0.30373	13.72420
Std. Error	0.06538	1.35057

A data is robust when its Standardized Error lies between -2.500000 and 2.500000

Les tubes en rouge ne sont pas réestimés, par manque de données en nombre suffisant dans le voisinage. Le couple de tubes distants de 3m au Nord est indiqué en bleu.

**Figure 17. NO2 hivernal. Validation croisée pour les modèles de référence : population, bâti dense et NOx en dérive, variogramme linéaire.**



Isatis

N02/Mulhouse tubes  
 - Variable #1 : Moyenne ete j03  
 Standard Parameter File for Model: DE.ete  
 Standard Parameter File for Neighborhood: Isotrope rayon 10000  
 Cross validation statistics based on 59 test data

	Mean	Variance
Error	0.08402	18.90193
Std. Error	0.00563	115.16656

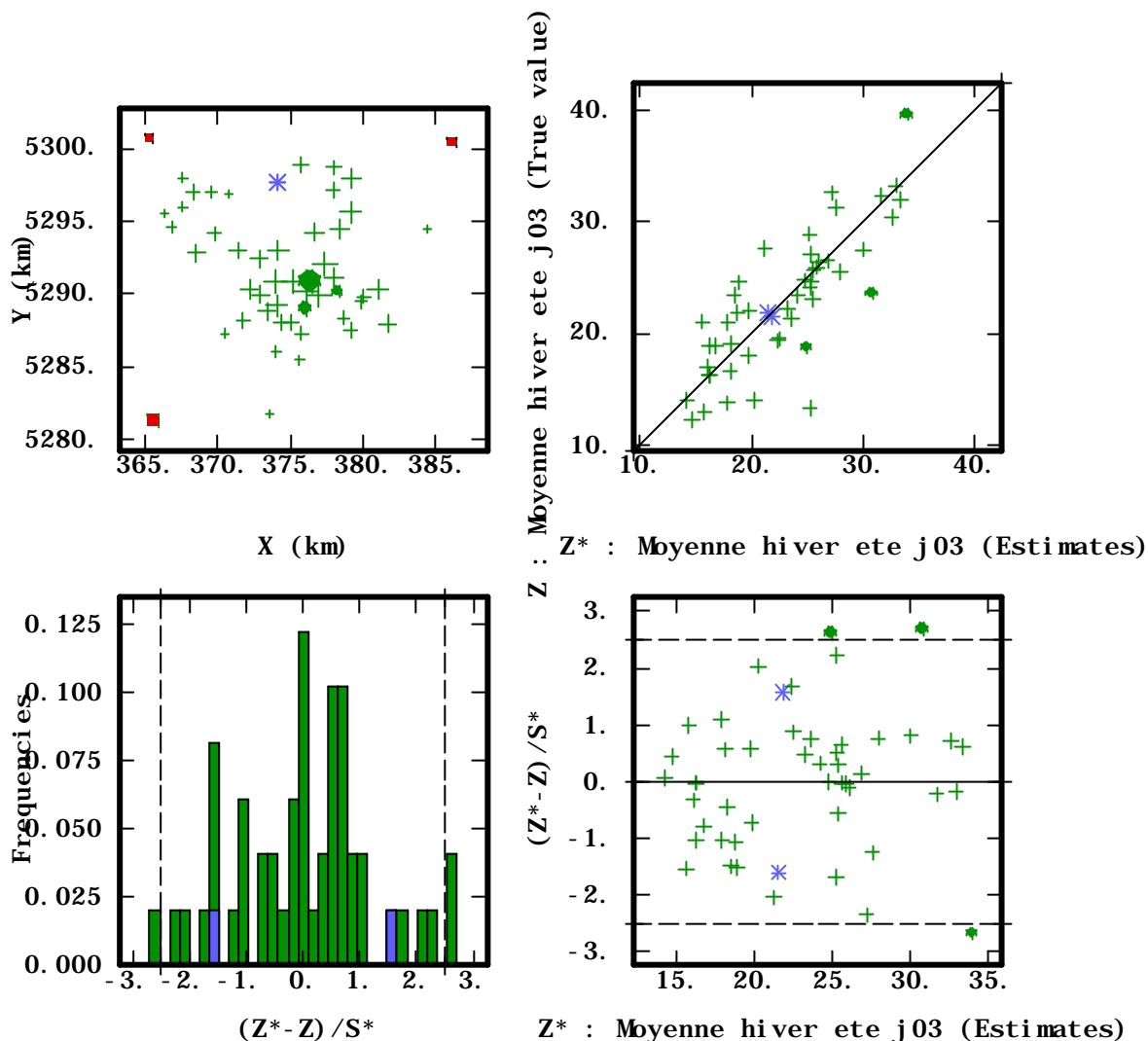
Cross validation statistics based on 49 robust data

	Mean	Variance
Error	0.26365	13.10691
Std. Error	0.02858	1.50188

A data is robust when its Standardized Error lies between -2.500000 and 2.500000

Le couple de tubes rapprochés au Sud est reporté en rouge, celui au Nord est reporté en bleu.

**Figure 18. NO2 estival. Validation croisée pour les modèles de référence : population, bâti dense et NOx en dérive, variogramme linéaire.**



N02/Milhouse tubes		
- Variable #1 : Moyenne hiver ete j03		
Standard Parameter File for Model: DE. etehiver		
Standard Parameter File for Neighborhood: Isotrope rayon 10000		
Cross validation statistics based on 49 test data		
	Mean	Variance
Error	0.15471	12.50872
Std. Error	0.01910	1.46625
Cross validation statistics based on 46 robust data		
	Mean	Variance
Error	0.00724	10.76155
Std. Error	-0.03778	1.09855
A data is robust when its Standardized Error lies between -2.500000 and 2.500000		

Les tubes en rouge ne sont pas réestimés, par manque de données en nombre suffisant dans le voisinage. Le couple de tubes distants de 3m au Nord est indiqué en bleu.

**Figure 19. NO2 "annuel". Validation croisée pour les modèles de référence : population, bâti dense et NOx en dérive, variogramme linéaire.**

Nous ne traçons pas de carte d'estimation de la moyenne saisonnière ou annuelle de NO<sub>2</sub>, les modèles retenus étant assez voisins de ceux utilisés par ASPA ; nous n'avons pas non plus comparé les cartes obtenues suivant les variantes.

La moyenne hivernale présente les meilleurs résultats (coefficient de corrélation  $(Z, Z^*)$  supérieur à 0.8), et la moyenne estivale les moins précis, ce coefficient étant inférieur à 0.7; la moyenne annuelle, calculée directement sur les moyennes des six quinzaines, indique une qualité intermédiaire (avec un coefficient de corrélation supérieur à 0.8 pour le modèle de référence). Si la moyenne et la variance de NO<sub>2</sub> sont supérieures en hiver, le coefficient de variation est plus important en été (rappel, tableau 1).

Les résultats de la validation croisée, ainsi que le coefficient de corrélation entre les moyennes saisonnières et leurs estimations, sont reportés au tableau suivant (et cf. figure 19.). Les nuages de corrélation entre les moyennes saisonnières  $Z$  et leur ré-estimation par validation croisée  $Z^*$  sont linéaires : ces coefficients décrivent donc de façon pertinente les liaisons entre ces variables.

### **3.2. Sensibilité aux variables auxiliaires**

Les écarts observés entre les différentes variantes restent souvent faibles ; les conclusions sont donc à valider sur d'autres campagnes.

Plusieurs analyses de sensibilité des résultats d'estimation ont été menées par validation croisée, en modifiant le variogramme ou les fonctions mises en dérive, ainsi que le maillage du cadastre des émissions en NO<sub>x</sub> (cf. paragraphe suivant). Dans ces différents cas, les nuages de corrélation  $(Z, Z^*)$  restent linéaires, avec au plus un à deux points s'écartant de l'ensemble du nuage, et indiquant la mauvaise précision locale de l'estimation.

Le coefficient de corrélation  $(Z, Z^*)$  et la variance de l'erreur de ré-estimation sont ici deux critères presque toujours concordants pour comparer les différents modèles: par variable, le coefficient de corrélation croît lorsque la variance de l'erreur diminue.

Pour les variantes examinées, le modèle de référence fournit les "meilleures" estimations pour la moyenne hivernale (avec des résultats analogues lorsque seuls la densité de population et le bâti dense sont utilisées comme dérive), et des résultats intermédiaires pour les valeurs estivales ou annuelles.

Pour les trois moyennes temporelles, le variogramme pépitiq ue fournit des résultats médiocres: le krigeage en dérive externe est donc préférable à une régression linéaire multiple, d'autant plus que celle-ci serait effectuée classiquement en voisinage unique. En voisinage glissant, le krigeage en dérive externe avec un variogramme pépitiq ue revient à ajuster localement une régression linéaire multiple.

Pour la moyenne estivale, la meilleure variante testée conserve les trois variables (bâti dense, densité de population et émissions de NO<sub>x</sub>) en dérive, avec un variogramme sphérique de portée 3km. ce variogramme est donc conservé pour les études de sensibilité ultérieures.

Pour cette moyenne estivale, la densité de population et le bâti, seules ou ensemble, donnent des résultats médiocres. Au contraire, NO<sub>x</sub> comme seule dérive donne des résultats très

voisins de ceux obtenus en rajoutant la densité de bâti dense et la densité de population, ceci pour le variogramme linéaire comme pour le variogramme sphérique "optimal" de portée 3km. Lorsque NOx est disponible, ces deux variables apportent peu. Ces résultats sont cohérents avec la forte corrélation notée précédemment entre NOx et la moyenne estivale.

Des résultats analogues s'observent pour la moyenne annuelle : pour un variogramme linéaire, NOx comme seule dérive semble équivalente aux trois variables en dérive. Pour la moyenne annuelle, la meilleure variante s'obtient avec comme dérive premier facteur de l'ACP des 5 variables explicatives. Or NOx est alors redondant, puisque ce facteur est calculé à la fois sur NOx et sur le premier facteur d'ACP de l'occupation anthropisée des sols (bâti dense, bâti lâche et industrie).

validation croisée NO2		hiver		été		"annuel"	
effectif		60		59		49	
critère		r(Z,Z*)	erreur	r(Z,Z*)	erreur	r(Z,Z*)	erreur
<b>référence</b>	3 variables linéaire	<b>0.83</b>	<b>m=0.19</b> <b>s2=15.1</b>	<b>0.64</b>	<b>m=0.08</b> <b>s2=18.9</b>	<b>0.81</b>	<b>m=0.15</b> <b>s2=12.5</b>
(3 variables) variogramme	pépitique	0.71	m=-0.07 s2=23.5	0.61	m=-.11 s2=18.8	0.68	m=-.12 s2=20.2
	sphérique3km	0.81	m=0.05 s2=16.2	0.67	m=-.13 s2=17.0	0.79	m=-.13 s2=13.6
	sphérique5km	0.83	m=0.21 s2=14.9	0.63	m=0.04 s2=19.4	0.80	m=0.05 s2=13.0
(variogramme linéaire) dérives	population	0.81	m=.11 s2=16.4	0.58	m=0.03 s2=21.0	0.81	m=.17 s2=12.5
	bâti dense	0.81	m=.20 s2=16.6	0.57	m=0.01 s2=21.4	0.80	m=.11 s2=13.2
	NOx	0.80	m=.22 s2=16.9	0.64	m=0.01 s2=18.6	0.82	m=.21 s2=11.8
	population & bâti	0.83	m=.12 s2=14.4	0.56	m=0.04 s2=22.5	0.81	m=.05 s2=12.5
	F1 de 5 variables	0.82	m=.16 s2=15.4	0.60	m=0.01 s2=20.5	0.83	m=.18 s2=11.6
sphérique 3km	NOx			0.66	m=-0.08 s2=17.5		

**Tableau 8. Résultats de la validation croisée pour les modèles de référence.**

Pour la moyenne hivernale, les deux variables bâti dense et densité de population donnent à l'inverse une estimation aussi bonne avec ou sans ajout de NOx. Si une seule de ces trois variables est mise en dérive, les résultats se détériorent légèrement, NOx correspondant à l'estimation la moins précise.

En résumé, densité de population et bâti “expliquent” pour partie les moyennes hivernales, et NOx les moyennes estivales (et annuelles). La variance la plus forte s’observe pour les moyennes estivales. Si l’on s’intéresse à cette moyenne, un approfondissement de l’étude variographique, pour rechercher d’autres paramètres explicatifs ou améliorer le variogramme, pourrait être utile.

Pour les différentes variantes, les coefficients de corrélation des résultats de validation croisée sont souvent forts, généralement supérieurs à 0.90, voire à 0.95, à l’exception notable du variogramme pépitiq.

Remarque : dans le krigeage avec dérive externe, les dérives sont supposées parfaitement connues aux noeuds de la grille d’estimation. Lorsque ce n’est pas le cas (variables explicatives connues à maille kilométrique, pour une grille d’estimation à maille 500m\*500m), la variance de l’erreur d’interpolation des variables auxiliaires est souvent négligée.

### **3.3. Précision des estimations**

Les statistiques des erreurs relatives de validation croisée  $\frac{Z - Z^*}{Z}$  sont utilisées pour compléter la comparaison entre variantes, et pour évaluer la précision des estimations (tableau 9.).

Pour le modèle de référence, la majorité des erreurs relatives sont inférieures à 20% pour la moyenne hivernale, la seule erreur supérieure à 50% correspondant à un tube isolé (figure 20., voir également les résultats de validation croisée, figure 19. et suivantes).

Comme indiqué, l’estimation estivale est beaucoup moins précise, avec une proportion importante d’erreurs relatives supérieures à 30%, qui ne correspondent pas nécessairement à de fortes variances de krigeage (voir les erreurs standardisées, la pente du variogramme étant ajustée pour le modèle de référence). Comparativement à ces erreurs, les écarts entre variantes restent peu importants.

Il paraît donc plus important de chercher à gagner en précision en augmentant le nombre de stations, ou en étudiant la précision effective des mesures, plutôt qu’en “optimisant” le modèle variographique (fonctions en dérives, variogramme).

Le couple de stations très rapprochées au sud, correspondant aux tubes situés à moins de trois mètres, montre une réestimation médiocre de la moyenne estivale (figure 20.) ; l’autre couple (au Nord) indique également une réestimation peu précise de la moyenne estivale.

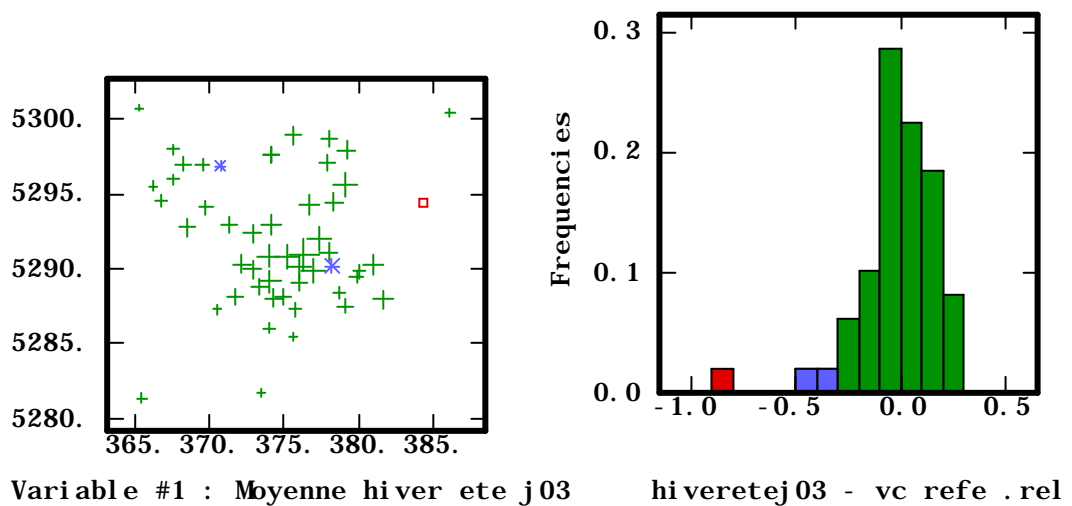
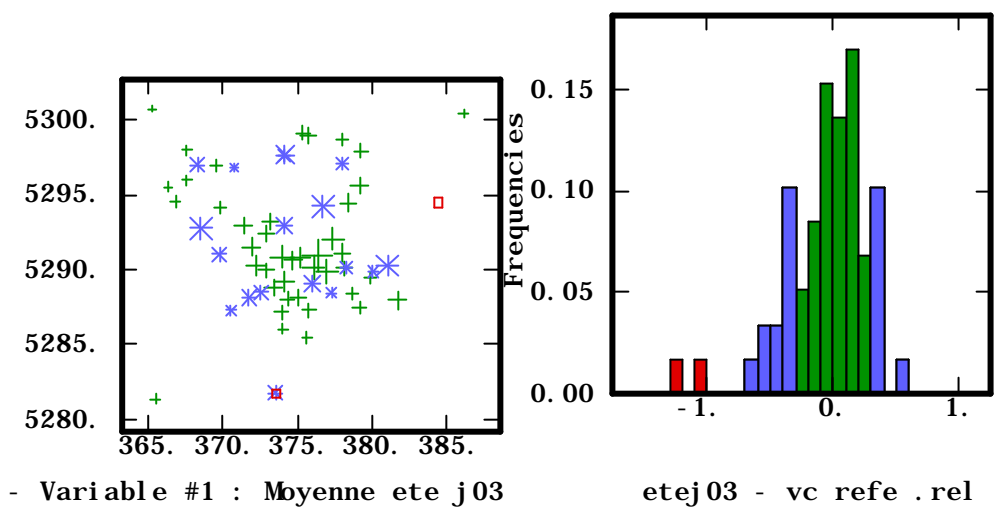
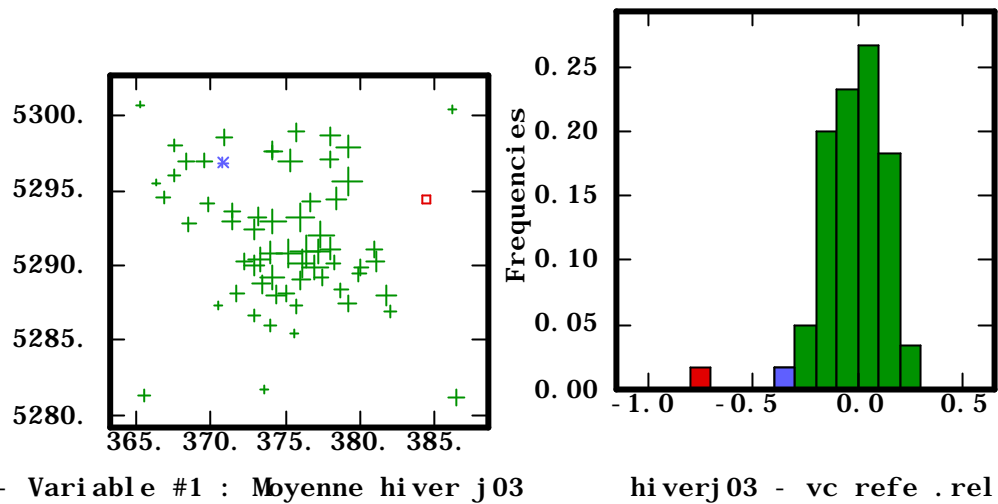
Ces estimations peu précises à petite distance sont dues à la forte variabilité de NO2 à l’échelle métrique : s’agit-il d’une variabilité réelle, ou de mesures imprécises par les tubes ?

Erreurs relatives - hiver : 60 tubes				
	minimum (%)	maximum (%)	moyenne (%)	écart-type (%)
référence : linéaire, 3 dérives	- 74.	22.	- 2.	16.
sphérique, 3 dérives	- 70.	25.	- 2.	16.
linéaire, population+bâti	- 60.	22.	- 2.	15.
linéaire, NOx	- 80.	27.	- 3.	17.

Erreurs relatives - été : 59 tubes				
	minimum (%)	maximum (%)	moyenne (%)	écart-type (%)
référence : linéaire, 3 dérives	- 123.	52.	- 6.	32.
sphérique, 3 dérives	- 108.	52.	- 5.	30.
linéaire, population+bâti	- 124.	52.	- 7.	34.
linéaire, NOx	- 108.	52.	- 6.	30.

Erreurs relatives - "annuel" : 49 tubes				
	minimum (%)	maximum (%)	moyenne (%)	écart-type (%)
référence : linéaire, 3 dérives	- 90.	26.	- 3.	19.
sphérique, 3 dérives	- 77.	30.	- 2.	19.
linéaire, population+bâti	- 77.	30.	- 3.	18.
linéaire, NOx	- 89.	25.	- 3.	19.

**Tableau 9. Statistiques élémentaires des erreurs relatives  $\frac{Z - Z^*}{Z}$  de validation croisée.**



Pour le repérage des couples de tubes rapprochés à 3 mètres, voir les figures présentant les résultats de validation croisée.

**Figure 20. Erreurs relatives  $\frac{Z - Z^*}{Z}$  de validation croisée pour le modèle de référence**

### 3.4. Sensibilité au cadastre des émissions

L'influence de la maille du cadastre des émissions figurait explicitement parmi les questions à examiner. Comme les corrélations sont sensibles au décalage d'une demi-maille de ce cadastre, nous effectuons la comparaison entre une maille kilométrique, et une moyenne sur 3km\*3km, soit 9 km<sup>2</sup>. Cette moyenne mobile est ensuite migrée aux tubes, la distance maximale entre le centre d'une maille de 9 km<sup>2</sup> et un tube étant de 1.414km (autrement dit, la maille de 3km\*3km est ici informée sur une grille de 1km\*1km et non de 3km\*3km). Le facteur de "normation" est identique pour les deux supports, égal à 20000.

La comparaison, pour le modèle de référence et pour deux variantes, est présentée au tableau 10. Numériquement, les écarts restent assez faibles, et les conclusions mériteraient d'être validées.

Pour la moyenne hivernale, les résultats sont analogues pour les deux mailles, légèrement meilleurs ou moins bons selon les cas. Rappelons que NOx ne constituait pas la "meilleure dérive" de cette moyenne hivernale. A maille 3km\*3km, les critères (coefficient de corrélation (Z,Z\*) et variance de l'erreur d'estimation) ne concordent plus.

Pour la moyenne estivale, le passage de la maille 1km\*1km à la maille 3km\*3km détériore systématiquement les résultats ; cet effet serait sans doutes plus marqué pour une grille cadastrale informée à maille 3km au lieu de 1km.

De façon surprenante, les résultats s'améliorent pour la moyenne annuelle.

maille du cadastre	validation croisée NO2		hiver		été		"annuel"	
	effectif		60		59		49	
	critère		r(Z,Z*)	erreur	r(Z,Z*)	erreur	r(Z,Z*)	erreur
<b>3*3</b>	variogramme linéaire	3dérives	<b>0.82</b>	<b>m=0.29</b> <b>s2=12.7</b>	<b>0.62</b>	<b>m=0.04</b> <b>s2=20.2</b>	<b>0.85</b>	<b>m=0.15</b> <b>s2=12.3</b>
1*1			0.83	m=0.19 s2=15.1	0.64	m=0.08 s2=18.9	0.81	m=0.15 s2=12.5
<b>3*3</b>	variogramme sphérique 3km	3dérives	<b>0.83</b>	<b>m=0.08</b> <b>s2=14.9</b>	<b>0.62</b>	<b>m=-.21</b> <b>s2=19.2</b>	<b>0.79</b>	<b>m=-.23</b> <b>s2=13.6</b>
1*1			0.81	m=0.05 s2=16.2	0.67	m=-.13 s2=17.0	0.79	m=-.13 s2=13.6
<b>3*3</b>	variogramme linéaire	NOx	<b>0.81</b>	<b>m=0.26</b> <b>s2=16.4</b>	<b>0.63</b>	<b>m=0.00</b> <b>s2=19.0</b>	<b>0.82</b>	<b>m=0.12</b> <b>s2=10.1</b>
1*1			0.80	m=.22 s2=16.9	0.64	m=0.01 s2=18.6	0.82	m=.21 s2=11.8

Ligne supérieure, en gras : cadastre 3km\*3km ; ligne inférieure : cadastre 1km\*1km.

Les 3 fonctions en dérive sont les logarithmes translétés du cadastres des émissions pour NOx, de la densité de population, et du bâti dense.

**Tableau 10. Validation croisée. Comparaison pour le cadastre des émissions à maille 3km\*3km et 1km\*1km (référence).**

## 4. Conclusions

Plusieurs points ont été laissés en suspens, notamment le choix du support optimal pour certaines variables auxiliaires (seule la régularisation du NOx “cadastral” a été examinée), ainsi que l’influence du facteur m dans le calcul du logarithme translaté  $\text{Log}(1+Z/m)$ .

Les principaux résultats mise en évidence sont les suivants :

- Les rares couple de tubes distants de moins de cinq mètres indiquent de forts écarts possibles sur les mesures par quinzaines. L’implantation de ces tubes est-elle préférentielle, précisément pour quantifier un contraste présumé important, ou ces écarts traduisent-ils l’amplitude des erreurs de mesures par quinzaine, ou alors la présence d’une très forte variabilité à petite distance ?
- Les corrélations entre NO<sub>2</sub> et les variables auxiliaires sont sensibles au décalage d’une demi-maille de la grille des paramètres explicatifs. Or certains de ces paramètres sont fournis suivant des grilles légèrement irrégulières. Ceci peut d’une part provoquer des artefacts si des grilles intermédiaires sont utilisées sans précautions ; d’autre part, il serait souhaitable de supprimer les incertitudes sur les coordonnées pour rechercher les corrélations avec NO<sub>2</sub>.
- Les corrélations entre les moyennes saisonnières, ou entre ces moyennes et les paramètres explicatifs sont sensibles aux ensembles de tubes retenus. Instrumenter des sites plus nombreux, au moins pour une campagne sur deux saisons (un hiver, un été), permettrait de valider certains résultats.
- Une nette différence de variabilité apparaît suivant les saisons, et pour chacune des saisons, une quinzaine diffère sensiblement des deux autres, sans explication par des congés ou par le calendrier scolaire. Il serait très utile de vérifier les conditions météorologiques durant ces quinzaines, ou de trouver une autre explication. Quoiqu’il en soit, réduire la durée de la campagne risquerait de mettre en cause son caractère représentatif. La représentativité saisonnière des trois quinzaines n’a d’ailleurs pas été démontrée. **Ceci pourrait être étudié en examinant les mesures des analyseurs, et en comparant pour ces analyseurs les résultats de ces trois quinzaines aux mesures durant toute la saison.**
- Les relations entre NO<sub>2</sub> et les variables “explicatives” mériteraient d’être validées sur des ensembles plus vastes de données. Le logarithme translaté de la densité de population et du bâti dense semblent adaptés en dérive externe pour la moyenne hivernale, tandis que le cadastre de NOx à maille kilométrique serait préférable pour la moyenne estivale. Cependant, les écarts entre les différentes variantes restent faibles.

Les résultats observés persisteraient-ils si l’on augmente le nombre de tubes?

- Ceci montre la nécessité d’approfondir la question du critère de comparaison des différents modèles variographiques proposés (variogramme et dérivées) . En particulier, dans une perspective d’application systématique d’une méthode de cartographie, la “robustesse” du modèle peut être préférée à son “optimalité” en termes de précision. Les variantes examinées ici sont-elles significativement différentes ?

Nous n’avons pas examiné les différences de comportement entre les tubes situés au centre ville ou en périphérie. Quelques réflexions sur ce sujet, ainsi que sur l’implantation des fortes valeurs aux tubes et pour les variables explicatives, sont données dans le premier rapport CG.

Ce travail pourrait être poursuivi, y compris pour le choix de cartes de NO<sub>2</sub> “les mieux appropriées”.

- Le cadastre des émissions, à maille kilométrique, améliore l'estimation pour la moyenne estivale principalement, ainsi que pour la moyenne annuelle. Son apport semble mineur pour la moyenne hivernales, lorsque la densité de population et le pourcentage d'occupation des sols (bâti dense) sont disponibles.
- La forte variance des erreurs de réestimation indique qu'une part importante de la variabilité n'est pas “expliquée” par les variables auxiliaires. On pense en particulier à la circulation routière, qui n'apparaît que très indirectement, via la densité de population ou le bâti. L'étude sur la variabilité de NO<sub>2</sub> dans la vallée de la Thure (S. Séguret, 2003; deuxième rapport d'avancement de la présente convention) montre, pour une route nationale, le très fort gradient de concentration en fonction de la distance à la route, en particulier pour les premiers mètres. Dans ce cas, la densité de population ou le bâti, même à maille 250m, ne peuvent “expliquer” la variabilité à distance métrique à décamétrique.
- Les relations entre moyennes estivales et hivernales n'ont pas été examinées de façon précise. Il serait intéressant de repérer ou de préciser des changements de typologie de la concentration en NO<sub>2</sub>, en particulier aux intersaisons. Les analyseurs, peu nombreux, ne permettront sans doute pas un tel examen.
- L'étude de mesures durant quelques années sur des analyseurs est indispensable pour évaluer la variabilité temporelle des concentrations. Un modèle variographique spatio-temporel est nécessaire pour évaluer les performances de différentes stratégies d'échantillonnage, par exemple en examinant la relation coût de mesure/précision sur la concentration moyenne annuelle (en distinguant éventuellement selon les zones de concentrations plus ou moins élevées).
- L'intérêt éventuel du cadastre des émissions devrait être examiné pour d'autres polluants : Benzène, Ozone, .... ainsi que pour les particules fines par exemple.

Les validations croisées montrent que les erreurs relatives supérieures à 30% sont rares pour la moyenne hivernale, ne sont pas exceptionnelles pour la moyenne annuelle, et sont fréquentes pour la moyenne estivale. Par comparaison, les écarts entre les variantes étudiées (variables en dérive, variogramme) restent faibles. A cette étape, il semble plus important **de chercher à améliorer la précision en augmentant le nombre de tubes ou en étudiant la précision des erreurs de mesures par les tubes, voire en examinant l'intérêt éventuel d'une variable auxiliaire liée à la circulation routière, plutôt que de ”raffiner” le modèle variographique.**

Enfin, deux autres pistes de travaux présentent une réelle utilité dans la problématique de la cartographie des polluants :

- **la comparaison des mesures aux stations fixes (analyseurs) et aux tubes.**

Une telle comparaison a été tentée par G. L'Hégaret pour des mesures de NO<sub>2</sub> et de Benzène (cf. rapport de stage d'option auprès d'ASCOPARG-COPARLY, 2002). Le résultat mériterait d'être validé.

**- l'évaluation de la variance de l'erreur de mesure aux tubes.**

Cette évaluation est facilitée lorsque plusieurs tubes sont implantés au même endroit (dans une même boîte) ; il serait très utile de disposer de données en provenance de différents sites.

On peut également tenter d'utiliser les mesures successives aux mêmes stations (tubes), et examiner l'évolution de l'effet de pépité en fonction de la moyenne ainsi que ses variations par régularisation temporelle.