



Application de modèles grande échelle à la problématique régionale : cas de l'ozone

***Laboratoire Central de Surveillance
de la Qualité de l'Air***

Convention n°115/03

*Cécile HONORÉ, Laure MALHERBE
Unité Modélisation et Analyse Economique
pour la gestion des Risques (MECO)*

Direction des Risques Chroniques (DRC)

Décembre 2003



Application de modèles grande échelle à la problématique régionale : cas de l’ozone

Laboratoire Central de Surveillance de la Qualité de l’Air

Convention n°115/03

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

Décembre 2003

Cécile HONORÉ
Laure MALHERBE

Ce document comporte 30 pages (hors couverture et annexes).

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

TABLE DES MATIÈRES

1	Résumé.....	4
2	Introduction.....	5
3	Présentation generale des méthodes de krigeage.....	8
3.1	Krigeage des observations.....	8
3.1.1	Krigeage ordinaire.....	8
3.1.2	Krigeage lognormal.....	10
3.2	Krigeage avec utilisation de Chimère.....	10
3.2.1	Krigeage des innovations.....	10
3.2.2	Krigeage avec dérive externe.....	11
3.3	Ajout d’informations secondaires.....	11
3.3.1	Ajout d’informations dans le krigeage des innovations.....	11
3.3.2	Ajout d ‘information dans le krigeage avec dérive externe.....	13
4	Gestion informatique des méthodes de krigeage.....	14
4.1	Architecture des programmes de krigeage.....	17
4.1.1	Gestion des données d'entrée.....	17
4.1.2	Krigeage.....	20
4.1.3	Gestion des données de sorties.....	21
4.1.4	Validation de la méthode de krigeage.....	22
4.2	Format des données d'entrée.....	23
4.2.1	Observations.....	23
4.2.2	Concentrations simulées.....	24
4.2.3	Données d’émissions spatialisées.....	25
4.2.4	Données d'urbanisation spatialisées.....	25
4.2.5	Observations utilisées pour la validation.....	26
4.2.6	Observations d'une espèce additionnelle utilisées pour le cokrigeage.....	26
4.3	Gestion des données de sortie: création de cartes avec GMT.....	27
5	liste des annexes.....	30

1 RESUME

La représentation cartographiée des champs de pollution atmosphérique est une priorité en terme de surveillance de la qualité de l'air, celle-ci étant exigée sur l'ensemble du territoire par les Directives Européennes.

La dernière décennie a vu le développement de modèles de simulation de la qualité de l'air - statistiques ou déterministes -, qui permettent une estimation spatialisée des champs de polluants sur l'ensemble des régions. Néanmoins, le développement et la mise en œuvre de tels modèles est coûteuse en moyens humains (capitalisation des compétences).

Depuis le printemps 2003, le système PREV'AIR¹ de modélisation déterministe de la qualité de l'air délivre quotidiennement des prévisions - à trois jours d'échéance - et des simulations - pour la veille - des concentrations de polluants atmosphériques à grande échelle sur l'Europe de l'Ouest. Les prévisions sont présentées au grand public sous forme de cartes sur le site <http://prevair.ineris.fr>; et sous forme de fichiers de données numériques aux organismes en charge de la surveillance de la qualité de l'air qui en font la demande. Le système met en œuvre le modèle CHIMERE-Continental développé depuis 1996 par l'IPSL/CNRS. La faible résolution du modèle CHIMERE-Continental, 50km*50km, met cependant hors de portée du système PREV'AIR la simulation des champs de polluants à l'échelle locale.

Une approche pragmatique consiste à utiliser l'ensemble des informations disponibles - observations de routine des AASQA et résultats de simulation grande échelle tels que ceux issus de PREV'AIR - pour cartographier les champs de pollution atmosphérique avec une résolution relativement fine. Une façon de la mener à bien est de mettre en œuvre des méthodes statistiques, dites de « krigeage ». Ces dernières corrigent les sorties du modèle déterministe de telle façon qu'aux points précis des observations, la concentration calculée est égale à la concentration observée.

Le présent rapport a pour l'objectif d'aider à la prise en main par les AASQA de ces outils, afin qu'elles puissent disposer d'une méthodologie de traitement des données issues de PREV'AIR pour réaliser des cartographies régionales. Il est organisé de la manière suivante: dans une première partie, les différentes méthodes de krigeage sont brièvement présentées de façon théorique. La deuxième partie vise à décrire de façon pratique l'architecture de programmes informatiques mis à la disposition des AASQA et incorporant les procédures d'adaptation statistique. Le format des données d'entrée requises par les différentes méthodes est également décrit de façon détaillée.

¹ mis en place à l'INERIS en collaboration avec le Ministère de l'Ecologie et du Développement Durable, l'Institut Pierre-Simon Laplace (IPSL) du CNRS et l'ADEME

2 INTRODUCTION

La représentation cartographiée des champs de pollution atmosphérique est une priorité en terme de surveillance de la qualité de l'air, celle-ci étant exigée sur l'ensemble du territoire par les Directives Européennes. Or, la surveillance par la mesure est coûteuse en équipement et en maintenance, de sorte qu'il existe en France, selon les polluants, de grandes régions peu couvertes par les observations de routine des Associations Agréées de Surveillance de la Qualité de l'Air (AASQA)

Par ailleurs, la dernière décennie a vu le développement de modèles de simulation de la qualité de l'air - statistiques ou déterministes -, qui permettent une estimation spatialisée des champs de polluants sur l'ensemble des régions. Néanmoins, le développement et la mise en œuvre de tels modèles est coûteuse en moyens humains (capitalisation des compétences).

Depuis le printemps 2003, le système PREV'AIR² de modélisation déterministe de la qualité de l'air délivre quotidiennement des prévisions - à trois jours d'échéance - et des simulations - pour la veille - des concentrations de polluants atmosphériques à grande échelle sur l'Europe de l'Ouest. Les prévisions sont présentées au grand public sous forme de cartes sur le site <http://prevair.ineris.fr>; sous forme de fichiers de données numériques aux organismes en charge de la surveillance de la qualité de l'air qui en font la demande (voir l'Annexe 1 pour la liste de ces organismes au 22/12/2003). Le système met en œuvre le modèle CHIMERE-Continental développé depuis 1996 par l'IPSL/CNRS. La faible résolution du modèle CHIMERE-Continental met cependant hors de portée du système PREV'AIR la simulation des champs de polluants à l'échelle locale.

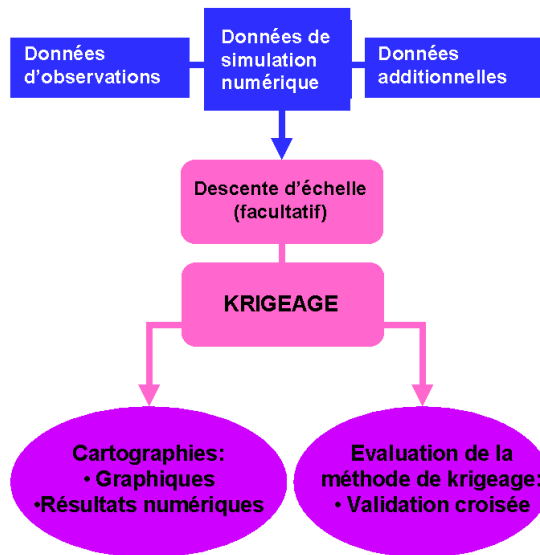
Une approche pragmatique consiste à utiliser l'ensemble des informations disponibles - observations de routine des AASQA et résultats de simulation grande échelle tels que ceux issus de PREV'AIR - pour cartographier les champs de pollution atmosphérique avec une résolution relativement fine. Dans le cadre du Laboratoire Central de Surveillance de la Qualité de l'Air, une étude de faisabilité a été menée³, ayant pour objectif de répondre à la question suivante: **moyennant une adaptation statistique basée sur l'utilisation des observations de routine des AASQA, est-il possible d'utiliser les sorties d'un modèle de simulation de la qualité de l'air à basse résolution pour cartographier la pollution atmosphérique à l'échelle locale?**

L'approche envisagée par l'INERIS est basée sur des méthodes de «krigeage», qui permettent d'exploiter les sorties du modèle CHIMERE-Continental à une échelle de quelques kilomètres en introduisant une correction basée sur les observations réelles. Le principe en est décrit schématiquement par la figure ci-dessous: à partir d'observations de la concentration d'ozone sur le domaine considéré, et de concentrations simulées de cette même espèce et éventuellement de données additionnelles, une cartographie de la concentration atmosphérique de ce polluant est établie. Notons que le krigage est un interpolateur exact : aux points précis des stations qui délivrent les observations, la concentration calculée est égale à la concentrations observée.

² mis en place à l'INERIS en collaboration avec le Ministère de l'Ecologie et du Développement Durable, l'Institut Pierre-Simon Laplace (IPSL) du CNRS et l'ADEME

³ par le laboratoire de statistiques de l'Université d'Orsay en collaboration avec le Laboratoire de Météorologie Dynamique du CNRS. Rapport d'étude disponible sur le site du LCSQA : www.lcsqa.org

L'étape de krigeage est précédée d'une étape de "descente d'échelle", c'est-à-dire de passage de la grille du modèle CHIMERE-Continental à une grille "fine" sur laquelle on souhaite établir la cartographie de la pollution atmosphérique.



L'étude de faisabilité menée par le Laboratoire de Statistiques de l'Université d'Orsay s'est formalisée par une analyse menée sur un cas concret, celui de la cartographie des concentrations d'ozone sur la région "Grand Ouest"⁴ pendant l'été 2001. Elle a montré, sur le cas étudié, l'efficacité des méthodes géostatistiques de krigeage afin de cartographier à l'échelle locale la pollution atmosphérique due à l'ozone, en "mixant" sorties à basse résolution d'un modèle déterministe de simulation numérique de la qualité de l'air et observations, complétées éventuellement de données additionnelles (émissions, taux d'urbanisation, observations de dioxyde d'azote...). Plusieurs approches avaient été étudiées durant cette évaluation. Chaque méthode a ses spécificités en terme de temps de calcul et de difficulté de mise en œuvre; les cartes réalisées sont différentes dans le détail, mais les critères statistiques ne permettent pas de privilégier une méthode par rapport aux autres.

Différentes méthodes de krigeage ont ainsi été évaluées :

- Krigeage des observations; lognormal;
- Krigeage avec le modèle en dérive externe;
- Krigeage avec le modèle et le taux d'urbanisation en dérive externe;
- Krigeage des innovations; avec les émissions en dérive externe; avec le taux d'urbanisation en dérive externe; avec double dérive externe (émissions, taux d'urbanisation);
- Cokrigeage avec les concentrations de dioxyde d'azote

Chacune d'elle a été implémentée dans des programmes informatiques spécifiques. Ces derniers, ainsi que les données utilisées, sont à disposition des AASQA sous forme d'un CD Rom (ou d'une archive informatique). Les données d'entrée sont fournies pour que les

⁴ i.e. la Vendée, le Poitou, les Pays de la Loire, la Basse-Normandie et l'est de la région Centre.

programmes puissent être facilement testés; l'utilisateur devra bien sûr avoir recours à ces propres données pour mettre en œuvre les méthodes de krigeage sur le cas précis qui l'intéresse.

Le présent rapport, dont l'objectif est d'aider à la prise en main de ces outils, est organisé de la manière suivante: dans une première partie, les différentes méthodes de krigeage sont brièvement présentées de façon théorique. La deuxième partie vise à décrire de façon pratique l'architecture des programmes; ceci est fait sur l'exemple de la méthode de krigeage des innovations avec les émissions et le taux d'urbanisation en dérive externe. On décrit également de façon détaillée le format des données d'entrée requises par les différentes méthodes.

3 PRESENTATION GENERALE DES METHODES DE KRIGEAGE

Plusieurs méthodes de krigeage sont proposées pour corriger les sorties du modèle CHIMERE et intégrer les observations dans la cartographie (krigeage des innovations, krigeage avec dérive externe). Elles offrent également la possibilité d'introduire des variables secondaires, comme les émissions ou le pourcentage d'urbanisation, ou de prendre en compte les corrélations spatiales avec d'autres polluants. Regroupées sous le terme de krigeage des observations, les deux premières méthodes ne reposent que sur l'exploitation des données de mesure. Elles sont présentées dans ce chapitre et fournies dans les programmes à titre de complément.

3.1 KRIGEAGE DES OBSERVATIONS

3.1.1 Krigeage ordinaire

Cette méthode consiste à interpoler les concentrations aux points de la grille à **partir des seules données d'observation**. La sortie du modèle CHIMERE n'est pas utilisée.

En un point s de la grille, la concentration, considérée comme une variable aléatoire, est estimée par une combinaison linéaire des observations :

$$\hat{Z}(s) = \sum_i \lambda_i Z(s_i)$$

Les pondérateurs λ_i sont calculés de telle façon que cet estimateur soit sans biais et optimal, c'est-à-dire de variance d'erreur minimale. En un point s donné, les deux conditions à respecter sont donc :

$$\begin{cases} E[\hat{Z}(s) - Z(s)] = E[\sum_i \lambda_i Z(s_i) - Z(s)] = 0 & (1) \\ Var[\hat{Z}(s) - Z(s)] = Var[\sum_i \lambda_i Z(s_i) - Z(s)] \text{ minimale} & (2) \end{cases}$$

L'hypothèse requise pour résoudre ce système est que Z est stationnaire d'ordre deux, c'est-à-dire que sa moyenne et sa covariance existent et sont stationnaires (indépendants de la position géographique du point considéré), ou que les accroissements de Z sont stationnaires d'ordre deux.

- La condition (1) s'écrit :

$$m(\sum_i \lambda_i - 1) = 0$$

où m est la moyenne de Z en tout point du domaine considéré (ou de façon moins contraignante, en tout point de la zone qui entoure s et à l'intérieur de laquelle les données $Z(s_i)$ sont utilisées.)

m étant inconnue, cette condition implique que $\sum_i \lambda_i = 1$

- Dans la condition (2), la variance de l'erreur d'estimation se développe de la façon suivante :

$$\text{Var}\left[\sum_i \lambda_i Z(s_i) - Z(s)\right] = \sum_i \sum_j \lambda_i \lambda_j C(s_i - s_j) - 2 \sum_i \lambda_i C(s_i - s) + C(s - s)$$

où $C(h) = E[Z(s)Z(s+h)] - m^2$ est la covariance de la fonction Z .

- Si Z est stationnaire d'ordre deux, cette covariance est liée directement à la fonction appelée *variogramme* (notée γ) par la relation : $\gamma(h) = C(0) - C(h)$

$$\text{où } \gamma(h) = \frac{1}{2} \cdot \text{Var}[Z(s+h) - Z(s)]$$

$\gamma(h)$ mesure la variabilité des concentrations entre deux points en fonction de la distance h qui sépare ces points.

- Si seuls les accroissements de Z sont stationnaires d'ordre deux, la covariance n'est pas définie et on ne considère que le variogramme.

Dans le cas le plus général, la variance de l'erreur d'estimation s'écrit donc :

$$\text{Var}\left[\sum_i \lambda_i Z(s_i) - Z(s)\right] = 2 \sum_i \lambda_i \gamma(s - s_i) - \sum_{i,j} \lambda_i \lambda_j \gamma(s_i - s_j)$$

La minimisation de cette variance sous la contrainte de non biais exige d'introduire un paramètre supplémentaire, μ , appelé multiplicateur de Lagrange. Le système d'équations à résoudre devient :

$$\begin{cases} \sum_j \lambda_j \gamma(s_i - s_j) - \mu = \gamma(s_i - s) \\ \sum_i \lambda_i = 1 \end{cases} \quad (\text{S})$$

Cela nécessite de définir un modèle de variogramme. En pratique, ce modèle est ajusté sur le variogramme expérimental calculé à partir des données de mesure :

$$\gamma_{\text{exp}}(h) = \frac{1}{2n(h)} \sum_{n(h)} [Z(s_i) - Z(s_j)]^2 \quad (\text{variogramme expérimental})$$

$n(h)$: nombre de couples de points (s_i, s_j) distants de $h \pm \Delta h$

Δh est un paramètre de tolérance sur la distance. Il sert à atteindre un nombre suffisant de couples pour calculer le variogramme à la distance h .

La résolution numérique du système (S) permet d'obtenir les valeurs des pondérateurs λ_i .

La variance de l'erreur d'estimation associée $\hat{Z}(s)$ est appelée variance de krigeage :

$$\sigma^2(s) = \text{Var}[\hat{Z}(s) - Z(s)] = \sum \lambda_i \gamma(s_i - s) - \mu$$

Elle fournit une indication sur la précision de l'estimation.

3.1.2 Krigeage lognormal

Certains phénomènes environnementaux mesurent des distributions de valeurs très dissymétriques. Lorsque le variogramme correspondant ne montre aucune structure de corrélation spatiale, une solution est de symétriser la distribution en appliquant une transformation logarithmique aux données initiales : $Y = \log(Z)$

Si le variogramme de la variable transformée Y est plus structuré, alors la modélisation et le krigeage ordinaire sont réalisés pour cette nouvelle variable. L'estimation de Z en un point s s'obtient par une transformation inverse :

$$\hat{Z}(s) = \exp(\hat{Y}(s) + \sigma_Y^2 / 2 - \mu_Y)$$

où σ_Y^2 et μ_Y sont la variance de l'erreur de krigeage et le paramètre de Lagrange associés à la variable Y .

Ce type d'estimation, peu pratiqué pour la pollution de l'air du fait qu'en général les distributions des concentrations n'en justifient pas l'usage, est appelé *krigeage lognormal*.

3.2 KRIGEAGE AVEC UTILISATION DE CHIMERE

Deux types d'approche sont considérés :

- dans le premier (krigeage des innovations), le krigeage porte sur la variable de différence entre concentration observée et modèle ;
- dans le second (krigeage avec dérive externe), le krigeage porte directement sur la variable de concentration observée aux stations de mesure.

3.2.1 Krigeage des innovations

On suppose que la concentration d'ozone en un point s est la somme de la concentration fournie par le modèle en ce point et d'un résidu aléatoire :

$$Z(s) = \text{mod}(s) + \tilde{Z}(s)$$

Aux stations de mesure, ce résidu a pour valeur la différence entre concentration mesurée et concentration simulée. Cette différence est désignée sous le terme d'*innovation*.

Le krigeage des innovations consiste à interpoler par krigeage ordinaire la variable \tilde{Z} aux points de la grille. En un point s de cette grille, l'estimation de Z s'obtient donc en ajoutant au modèle la valeur interpolée de \tilde{Z} :

$$\hat{Z}(s) = \text{mod}(s) + \sum_i \lambda_i \tilde{Z}(s_i)$$

- $\text{mod}(s)$ est la valeur du modèle Chimère au point s .
- $\tilde{Z}(s_k)$ est l'innovation à la station de mesure s_k , calculée par la relation :

$$\tilde{Z}(s_k) = Z(s_k) - \text{mod}(s_k)$$

$Z(s_k)$: observation au point s_k

$\text{mod}(s_k)$: valeur du modèle interpolée au point s_k

- Les poids λ_i sont solutions du système de krigeage associé à la variable \tilde{Z} .

Le modèle de variogramme doit être ajusté sur le variogramme expérimental des innovations.

Comparé au krigeage des observations, le krigeage des innovations permet d'apporter une information sur la structure du panache, introduite par le modèle CHIMERE. Le rôle des observations est de corriger les erreurs modèle.

3.2.2 Krigeage avec dérive externe

Cette méthode, proche de la précédente, revient à considérer la concentration en un point comme la somme d'une fonction linéaire du modèle (la dérive externe) et d'un résidu aléatoire : $Z(s) = \alpha \text{mod}(s) + \beta + R(s)$

La principale différence réside dans l'expression de l'estimateur. En effet, comme dans le krigeage ordinaire des observations, celui-ci s'obtient par une combinaison linéaire des concentrations mesurées. L'existence d'une dérive déterminée par le modèle CHIMERE est prise en compte implicitement dans les pondérateurs λ_i :

$$\hat{Z}(s) = \sum_i \lambda_i Z(s_i)$$

Les poids λ_i doivent satisfaire non seulement aux conditions du krigeage ordinaire (biais d'estimation nul et variance d'erreur minimale) mais également à la condition supplémentaire de dérive externe :

$$\sum_i \lambda_i \text{mod}(s_i) = \text{mod}(s)$$

Les constantes α et β n'apparaissent pas explicitement dans le système de krigeage. Toutefois, le krigeage impose de définir un modèle variographique pour la variable résiduelle sous-jacente. Plusieurs procédures sont possibles à cette fin. La méthode ici proposée consiste :

- à estimer par moindres carrés les constantes α et β sur l'ensemble des stations de mesure
- à calculer les résidus $Z(s_i) - (\alpha \text{mod}(s_i) + \beta)$ en chaque station
- à calculer le variogramme expérimental de ces résidus

3.3 AJOUT D'INFORMATIONS SECONDAIRES

Les méthodes présentées dans toute la suite reposent sur l'hypothèse que la variable d'intérêt (les innovations en 3.3.1, les concentrations d'ozone en 3.3.2) est corrélée avec les variables secondaires. Cette hypothèse doit être vérifiée préalablement.

3.3.1 Ajout d'informations dans le krigeage des innovations

a) Prise en compte du NO₂ par un cokrigeage

Le principe de la méthode est identique à celui du krigeage des innovations mais, en plus des données d'innovation, l'estimation de la variable \tilde{Z} utilise les données de NO₂. Elle tient compte à cette fin de la structure de covariance de \tilde{Z} , de celle du NO₂ et de la covariance croisée entre \tilde{Z} et NO₂.

Cette dernière se définit comme : $C_{\tilde{Z}Z_{NO_2}}(h) = \text{cov}[\tilde{Z}(s+h), Z_{NO_2}(s)]$

Le variogramme correspondant est: $\gamma_{\tilde{Z}Z_{NO_2}}(h) = \frac{1}{2} E\{[\tilde{Z}(s+h) - \tilde{Z}(s)][Z_{NO_2}(s+h) - Z_{NO_2}(s)]\}$

Notons Z_1 la variable innovation pour le polluant d'intérêt, ici l'ozone. ($Z_1 = \tilde{Z}$)

Z_2 la concentration du second polluant, ici le NO_2

L'estimateur de la concentration en un point s devient :

$$\hat{Z}(s) = \text{mod}(s) + \sum_i \sum_{j=1}^2 \lambda_{ij} Z_j(s_i)$$

Les poids doivent satisfaire aux conditions suivantes :

$$\left\{ \begin{array}{l} \sum_i \lambda_{ij} = 1 \quad \text{si } j = 1 \text{ (variable d'intérêt)} \\ \sum_i \lambda_{ij} = 0 \quad \text{si } j = 2 \text{ (variable secondaire) si } j=1 \text{ (variable d'intérêt)} \\ \text{minimisation de la variance d'estimation} \end{array} \right.$$

Le calcul des poids impose de modéliser le variogramme des innovations, le variogramme du NO_2 et le variogramme croisé de ces deux variables. Celui-ci est calculé expérimentalement de la façon suivante :

$$\gamma_{12\text{exp}}(h) = \frac{1}{2n(h)} \sum_{N(h)} [z_1(s_i) - z_1(s_j)][z_2(s_i) - z_2(s_j)]$$

La technique du cokrigeage se révèle intéressante lorsque les données de la variable secondaire ne sont disponibles qu'en un nombre limité de stations.

b) Prise en compte des émissions et du pourcentage d'urbanisation comme dérivés externes

De façon générale, notons Φ_1, \dots, Φ_N les variables secondaires, connues en tout point de la grille de calcul.

La technique de la dérive externe, présentée au paragraphe 1.2.2, est un moyen efficace d'introduire ces variables dans la cartographie.

On suppose ici que ces variables expliquent les innovations :

$$\tilde{Z}(s) = Z(s) - \text{mod}(s) = \sum_{k=1}^N \alpha_k \Phi_k(s) + \beta + R(s)$$

L'estimation de \tilde{Z} en un point s s'écrit de la même manière que dans le krigeage des innovations :

$$\hat{Z}(s) = \text{mod}(s) + \sum_i \lambda_i \tilde{Z}(s_i)$$

mais les poids λ_i doivent remplir la condition supplémentaire de dérive externe :

$$\sum_i \lambda_i \Phi_k(s_i) = \Phi_k(s) \quad \forall k \in [1 \dots N]$$

Le variogramme utilisé dans la résolution du système de krigeage est modélisé sur les résidus. $\tilde{Z} - (\sum_{k=1}^N \alpha_k \Phi_k + \beta)$, où les coefficients α_k et la constante β sont estimés par moindres carrés.

3.3.2 Ajout d'information dans le krigeage avec dérive externe

On suppose que non seulement le modèle CHIMERE mais également les variables secondaires Φ_1, \dots, Φ_N expliquent les concentrations :

$$Z(s) = \alpha \text{mod}(s) + \sum_{k=1}^N \alpha_k \Phi_k(s) + \beta + R(s)$$

Ces variables s'ajoutent donc au modèle en tant que dérivées externes.

L'estimateur ne change pas :

$$\hat{Z}(s) = \sum_i \lambda_i Z(s_i)$$

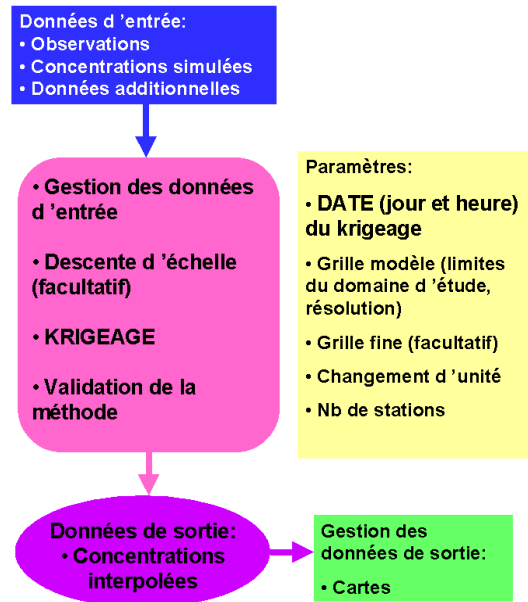
Mais les poids λ_i doivent vérifier la condition de dérive externe pour l'ensemble des variables :

$$\left\{ \begin{array}{l} \sum_i \lambda_i \text{mod}(s_i) = \text{mod}(s) \\ \sum_i \lambda_i \Phi_k(s_i) = \Phi_k(s) \quad \forall k \in [1 \dots N] \end{array} \right.$$

Le variogramme utilisé dans la résolution du système de krigeage est modélisé sur les résidus $Z - (\alpha \text{mod} + \sum_{k=1}^N \alpha_k \Phi_k + \beta)$, où les coefficients α_k et la constante β sont estimés par moindres carrés.

4 GESTION INFORMATIQUE DES METHODES DE KRIGEAGE

Au cours de l'étude, chacune des méthodes de krigeage étudiées a été implémentée dans un programme, développé à l'aide du logiciel R, logiciel en accès libre et disponible pour les systèmes d'exploitation les plus courants⁵. Tous les programmes ont la même structure, qui est décrite au paragraphe 4.1 ci-dessous, et schématisée dans la figure suivante.



Toutes les données d'entrée requises par les programmes sont fournies sous forme de fichiers ASCII. Le tableau ci-dessous précise de façon synthétique les données d'entrée nécessaires pour la mise en œuvre de chaque méthode.

⁵ Les modalités pratiques de téléchargement du logiciel figurent en Annexe 2.

	Concentrations (moyennes horaires) observées de l'espèce cartographiée	Concentrations (moyennes horaires) simulées de l'espèce cartographiée	Concentrations (moyennes horaires) observées d'une espèce additionnelle	Données d' émissions spatialisées	Données d' urbanisation spatialisées
Echelle	Stations de mesure	Grille modèle	Stations de mesure	Grille spécifique	Grille spécifique
Krigeage des observations R/krigeageobs.R	X				
Krigeage lognormal des observations R/krigeagelognormal.R	X				
Krigeage des observations avec le modèle en DE ⁶ R/derive.R	X	X			
Krigeage des observations avec le modèle et le taux d'urbanisation en DE R/derivemodpurb.R	X	X			X
Krigeage des innovations R/krigeageinnov.R	X	X			
Krigeage des innovations avec les émissions en DE R/innovderiveemi.R	X	X		X	

⁶ DE = dérive externe

Krigeage des innovations avec le taux d'urbanisation en DE R/innovderiveurb.R	X	X			X
Krigeage des innovations avec les émissions et le taux d'urbanisation en DE R/innovderivedouble.R	X	X		X	X
Cokrigeage R/cokrigeage.R	X	X	X		
Données d'entrée	donnees/obs_gr2.dat donnees/coord_gr2.dat	donnees/CONT20021 1J-AIRPL13	donnees/Moyenne_horaire_2001_NO2.txt donnees/stationscokri.dat	donnees/CITEPA_GRIILLE_SGO.txt	donnees/propurb2.dat
Validation	donnees/Moyenne_horaire_2001_O3.txt donnees/stationsverif.dat				

On présente de façon détaillée au paragraphe 4.2 le format des données d'entrée requises par chacune des méthodes de krigeage. Enfin, au paragraphe 4.3, on décrit brièvement la procédure de création des cartes de concentrations calculées à l'aide des méthodes de krigeage.

L'ensemble des programmes, données d'entrée et programmes de création de cartes est disponible pour les utilisateurs sous forme d'un CD Rom (ou d'une archive informatique)⁷. Les données d'entrée sont fournies pour que les programmes puissent être facilement testés; l'utilisateur devra bien sûr avoir recours à ces propres données pour mettre en oeuvre les méthodes de krigeage sur le cas précis qui l'intéresse.

4.1 ARCHITECTURE DES PROGRAMMES DE KRIGEAGE

Les programmes réalisés sont adaptés au logiciel R version 6.0 (Attention : la syntaxe de certaines fonctions peut changer avec les versions ultérieures). Pour les méthodes de krigeage des innovations et avec dérive externe, le package geoR a été utilisé; pour la méthode de cokrigeage, le package GSTAT a été utilisé.

Tous les programmes ont la même structure. A titre d'exemple, on reprend ci-dessous dans ses grandes lignes la structure du programme de krigeage des innovations avec en dérive externe les émissions de NOx et le taux d'urbanisation.

4.1.1 Gestion des données d'entrée

On commence par détailler la partie relative à la gestion des données d'entrée et à la définition des différents paramètres de krigeage (modèle, observations, données additionnelles, date de krigeage...). Les commentaires en rouge ne figurent pas dans les programmes. Les différents paramètres à adapter par l'utilisateur figurent en bleu, ainsi que les noms des fichiers contenant les données d'entrée. Enfin, les lignes en vert sont relatives à la gestion des dates dans les fichiers de données modèles et d'observations (en particulier passage à l'heure TU pour toutes les données d'observations); c'est un point auquel l'utilisateur devra être particulièrement attentif.

#	le 07/04/2003	# chargement des packages utilises
#	Delphine GRANCHER	#-----
#on fait un krigeage ordinaire sur les innovations		<i>Il s'agit aussi bien des packages R que de fonctions personnelles dont on a besoin dans le programme</i>
# avec pour derive externe les emissions et l'urb		library(chron)
		library(akima)
		library(geoR)
		library(splancs)
<i>Commande à copier sans le dièse dans la fenêtre de commande pour exécuter le programme</i>		source("R/chargefonction.R")
#source("R/innovderivedouble.R")		#-----
<i>Effacement de toutes les données existant en mémoire</i>		# limites du domaine d'etude
rm(list=ls())		#-----
#-----		sud<-45.5
		nord<-49.5

⁷ Les programmes de krigeage figurent dans le répertoire "R"; les données d'entrée dans le répertoire "donnees"; les scripts de création de cartes dans le répertoire "gmt".

```

ouest<- -4.5
est<-1.5

#-----
#definition de la grille du modele
#-----
Pas en latitude et en longitude de la grille
modèle

paslat<-0.25
paslong<-0.5

#-----
# autres parametres
#-----
Paramètre de conversion des concentrations
modèles de ppb en µg/m3
unite = 1.997
Flag sur le vent (vaut 1 ou 0 selon que l'on
souhaite raffiner ou pas la prise en compte de
l' « effet de côte » (cf. ci-dessous))
vent=0
Pas en latitude et en longitude de la grille fine
sur laquelle on veut établir une cartographie
de la pollution atmosphérique
petitpas<-0.08333/2 # pas de la petite grille
idem solPL
Nombre d'observations utilisées
nbstations=30
Gestion de l'effet de côte8
cotes=matrix(scan("donnees/cotes.dat"),ncol=
2,byrow=TRUE)
#-----
#grille d'interpolation
#-----
Définition de la grille fine
grx <-seq( ouest,est,by=petitpas)
gry <-seq(sud,nord,by=petitpas)

xi2<-
ouest+(as.numeric(gl(length(grx),length(gry),
ordered=FALSE))-1)*petitpas

```

⁸ Données à fournir en entrée aux programmes de krigeage, sous forme d'un fichier texte à deux colonnes contenant les coordonnées (longitude,latitude) des côtes sur le domaine d'étude.

```

yi2<-rep(gry,length(grx))

#-----
#interpolation des emissions
#-----
Gestion des données d'entrée (1): données
additionnelles (émissions)

Chargement des données d'émissions
emissions<-
matrix(scan("donnees/CITEPA_GRILLE_SG
O.txt",skip=1),ncol=7,byrow=TRUE)

#le polluant est NOX
emissions=cbind(emissions[,1],emissions[,2],
emissions[,7])
write(t(emissions),file="gmt/emissions.xyz",n
columns=3,append=FALSE)

Interpolation linéaire des données d'émissions
sur la grille fine
emissions2=interp(emissions[,1],emissions[,
2],emissions[,3],xi2,yi2)

Filtrage des données d'émissions: on ne
conserve que les points de la grille fine pour
lesquels on a des données d'émissions
xi2=xi2[!is.na(emissions2$z)]
yi2=yi2[!is.na(emissions2$z)]

#calcul de la propurb a la meme echelle que le
modele
#-----
-
Gestion des données d'entrée (2): données
additionnelles (taux d'urbanisation)
Chargement des données de taux
d'urbanisation
Propurb<-
matrix(scan("donnees/propurb2.dat"),ncol=14
4,byrow=TRUE)
propurb<-Propurb
dim(propurb)<-c(1,144*96)
Définition de la grille associée aux données
de taux d'urbanisation
xurb=seq(ouest,est,by=(6/143))
yurb=seq(nord,sud,by=-4/95)
Yurb<-rep(yurb,length(xurb))

```

```
Xurb<-
ouest+(as.numeric(gl(length(xurb),length(yurb),ordered=FALSE))-1)*(6/143)
```

Interpolation linéaire des données de taux d'urbanisation sur la grille fine

```
purb=interp(Yurb,Xurb,propurb,gry,grx,ncp=0)
```

```
purb<-matrix(purb$z,nrow=1)
```

```
purb<-as.vector(purb)
```

Filtrage des données de taux d'urbanisation: on ne conserve les données que pour les points de grille pour lesquels on a des données d'émissions

```
purb=purb[!is.na(emissions2$z)]
```

```
#-----
#chargement des donnees en fonction du jour etudie
#-----
```

Gestion des données d'entrée (3): données modèles

Paramètre de définition du fichier des données modèles

```
deb<-dates("04/30/01") # jour de reference des donnees
```

Paramètre de définition de la date de krigeage (jour et heure)

```
jour <- dates("07/28/01")
```

```
heure <- 15
```

Définition de la grille modèle

```
grx2 <-seq( ouest,est,by=paslong)
```

```
gry2 <-seq(sud,nord,by=paslat)
```

```
x2<-rep(grx2,length(gry2))
```

```
y2<-
```

```
sud+(as.numeric(gl(length(gry2),length(grx2),ordered=FALSE))-1)*paslat
```

Chargement des données modèles

```
CONT200211JAIRPL13<-
matrix(scan("donnees/CONT200211JAIRPL13"),ncol=(1+length(grx2)*length(gry2)),byrow=TRUE)
```

```
nojour<-as.numeric(jour-deb)
```

```
ligne<-(nojour-1)*24+heure
```

```
determ<-
CONT200211JAIRPL13[ligne,2:length(CONT200211JAIRPL13[1,])]*unite
```

Interpolation des données modèles sur la grille fine

```
determ2<-
```

```
interpmodel(determ,sud,petitpas,grx,gry,grx2,gry2,x2,y2,vent,cotes)
```

Filtrage des données modèles: on ne conserve les données que pour les points de grille pour lesquels on a des données d'émissions

```
determ2<-determ2[!is.na(emissions2$z)]
```

```
emissions2=emissions2$z[!is.na(emissions2$z)]
```

```
transit<-ll2mtu(sud,ouest,yi2,xi2)
```

```
xii<-(transit$x)/1000
```

```
yii<-(transit$y)/1000
```

```
#-----
#chargement des observations
#-----
```

Gestion des données d'entrée (4): données d'observations

```
obs<-
matrix(scan("donnees/obs_gr2.dat"),ncol=nbstations,byrow=TRUE)
```

```
ligne<-(as.numeric(jour-deb)-1)*24+heure
obsj <-
c(obs[ligne+2,1:8],obs[ligne,9:nbstations])
```

```
obsj<-recodnan(obsj,-999,NA)
```

Suppression des données d'observations manquantes

```
indnan<-which(!is.na(obsj)) #indice des stations renseignees.
```

```
z<-obsj[indnan]
```

Définition de la localisation des stations de mesure

```
coordstat<-
matrix(scan("donnees/coord_gr2.dat"),ncol=2,byrow=TRUE)
```

```
lambda<-coordstat[,1]
```

```
phi<-coordstat[,2]
```

```
#-----
#chargement des donnees d urbanisation
```

```

#-----
Pour chaque station, on calcule le taux d'urbanisation, la concentration simulée et les émissions par interpolation aux points de mesure. Les valeurs manquantes sont traitées.
propstat=rep(0,nbstations)
i1=rep(0,nbstations)
i2=rep(0,nbstations)
ggry=seq(nord,sud,by=-petitpas)
for (i in 1:nbstations){
  i1[i]=min(which(ggry<lambda[i]))-1

  i2[i]=max(which(grx<phi[i]))
propstat[i]=Propurb[i1[i],i2[i]]
}
propstat=propstat[indnan]

zmod<-
interpstations(determ,sud,petitpas,grx,gry,grx
2,gry2,x2,y2,phi,lambda,vent,cotes)
zmod<-zmod[indnan]
transit<-ll2mtu(sud,ouest,lambda,phi)
xs<-transit$x[indnan]/1000
ys<-transit$y[indnan]/1000

emistat<-
interp(emissions[,1],emissions[,2],emissions
[,3],phi[indnan],lambda[indnan])
emistat<-emistat$z
indnane<-which(!is.na(emistat))
emistat<-emistat[indnane]

xs<-xs[indnane]
ys<-ys[indnane]
zmod<-zmod[indnane]
z<-z[indnane]
propstat=propstat[indnane]

```

4.1.2 Krigeage

Les commentaires ci-dessous sont spécifiques à la méthode de krigeage des innovations avec les émissions et le taux d'urbanisation en dérive externe. On donne en Annexe 2 des détails sur l'estimation du variogramme dans le cas de la méthode de cokrigeage.

```

#-----
#Estimation des parametres par moindres
carres ordinaires
#-----
#Estimation des parametres par moindres
carres ordinaires
#-----
Création d'une matrice contenant les données
additionnelles et les valeurs modèle
Phi <-
cbind(matrix(emistat,ncol=1),matrix(propstat,
ncol=1),matrix(
rep(1,times=length(zmod)),ncol=1))
mu de la régression linéaire des innovations
sur les variables additionnelles
beta <- solve(Phi,z-zmod)
alpha<-beta[1]
gamma<-beta[2]
mu<-beta[3]

#Calcul du variogramme sur les residus
#-----
Calcul des résidus de la régression
Z<- (z -zmod)- mu -alpha*emistat-
gamma*propstat

s.out <-
output.control(messages.screen=FALSE)
geodata<-cbind(xs,ys,Z)
geodata<-as.geodata(geodata)

```

Estimation par moindres carrés des coefficients alpha et gamma et de la constante

Calcul du variogramme empirique

Calcul du variogramme expérimental de ces résidus

```
vario<-
variog(messages.screen=FALSE,geodata,max
.dist=250,pairs.min=1,uvec=seq(0,250,by=30)
)
```

geodata : données sur lesquelles est calculé le variogramme, ici les résidus

max.dist est la distance maximale entre deux points au-delà de laquelle le variogramme n'est plus calculé (250 km dans l'exemple). Cette distance ne dépasse généralement pas le tiers ou la moitié du champ total. Au-delà, le comportement du variogramme devient plus erratique et risque de fausser l'ajustement automatique.

pairs.min est le nombre minimal de couples de stations nécessaire pour calculer chaque point du variogramme expérimental. Idéalement, une trentaine de couples serait nécessaire mais ce nombre est à ajuster en fonction des mesures disponibles.

uvec=seq fournit les distances minimale et maximale de calcul du variogramme (0 et 250 km dans l'exemple) ainsi que le pas de calcul (ici 30 km).

La valeur de ce pas doit être choisie de façon qu'un nombre suffisant de couples intervienne dans le calcul de chaque point du variogramme expérimental et que la structure de ce variogramme devienne apparente. Elle doit être ajustée par des tests préliminaires sur des données de la zone considérée.

#approximation du vario

#-----

Ajustement automatique d'un modèle de variogramme sur le variogramme expérimental

```
varioest<-
variofit(messages.screen=FALSE,minimisatio
n.function
="nlm",vario2,ini.cov.pars=c(250,5),fix.nugget
=TRUE,fix.kappa=TRUE,max.dist=vario2$ma
x.dist)
#plot(vario)
#lines(varioest)
```

minimisation.function : fonction de minimisation utilisée pour ajuster les paramètres. Deux possibilités : optim ou nlm
ini.cov.pars : initialisation des valeurs pour le palier et pour la portée (ici palier_{ini}=250 et portée_{ini}=5 km)

fix.nugget : si fix.nugget vaut TRUE, l'effet de pépite prend une valeur fixe égale à 0 (valeur par défaut) ou à δ (il convient alors d'ajouter nugget= δ en argument dans la fonction). Si fix.nugget vaut FALSE, l'effet de pépite est estimé.

Par défaut, le modèle choisi est de type exponentiel. Pour ajuster d'autres types de modèle, il faut ajouter l'argument supplémentaire : cov.model= « spherical » (modèle sphérique), cov.model= « linear » (modèle linéaire), cov.model= « cubic » (modèle cubique) ou cov.model= « gaussian » (modèle gaussien).

max.dist : distance maximale jusqu'à laquelle le variogramme est ajusté. Dans cet exemple, elle est prise égale à la distance maximale de calcul du variogramme expérimental.

4.1.3 Gestion des données de sorties

La partie relative à la gestion des données de sortie est succincte. Les valeurs du modèle sur la grille fine sont écrites dans des fichiers avec les coordonnées des points correspondants. De même, l'estimation des concentrations par krigeage sur la grille fine est écrite dans un autre fichier. Ces fichiers (*.xyz) sont destinés à être traités par des Système d'Information Géographique. Les noms des fichiers contenant les données de sortie du programme sont à adapter éventuellement par l'utilisateur.

#-----

#donnees envoyees pour realiser des cartes avec GMT

#-----

```
#write(rbind(xi2,yi2,determ2),file="gmt/modelemi.xyz",ncolumns=3,append=FALSE)
```

```
#write(rbind(xi2,yi2,Kinnovderivedouble$predict+determ2),file="gmt/innovderivedouble_pre
dict.xyz",ncolumns=3,append=FALSE)
#write(rbind(xi2,yi2,Kinnovderivedouble$krige.var),file="gmt/innovderivedouble_erreur.xyz
",ncolumns=3,append=FALSE)
```

4.1.4 Validation de la méthode de krigeage

La validation de la méthode de krigeage est faite de deux manières différentes:

- Validation croisée: les observations disponibles pour toutes les stations sauf une sont utilisées pour construire le variogramme. Sur la station laissée "de côté", la valeur de la concentration obtenue par krigeage à partir de toutes les autres observations est comparée à la valeur effectivement observée. On effectue cette comparaison pour chacune des stations disponibles.
- Validation sur un jeu de données indépendantes: certaines stations ne sont pas utilisées pour construire le variogramme. En ces points de mesure, on compare la valeur des concentrations obtenues par krigeage aux valeurs effectivement observées.

Dans les deux cas, la comparaison se fait de manière statistique à l'aide de la racine de l'erreur quadratique moyenne.

Les commentaires en rouge ne figurent pas dans les programmes. Les différents paramètres à adapter par l'utilisateur figurent en bleu, ainsi que les noms des fichiers contenant les données d'entrée. Enfin, les lignes en vert sont relatives à la gestion des dates dans les fichiers de données d'observations utilisées pour la validation croisée.

<pre>#----- #Validation croisee #----- <i>En chaque station, on estime la valeur de la concentration obtenue par krigeage à partir des valeurs aux autres stations et on la compare avec la vraie valeur La RMSE sert de critère pour valider la méthode de krigeage.</i> RMSE<-0 for(no in 1:length(xs)){no xstemp<-xs[-no] ystemp<-ys[-no] ztemp<-z[-no] Ztemp<-z[-no] zmodtemp<-zmod[-no] emistattemp<-emistat[-no] propstattemp<-propstat[-no] geodatatemp<- as.geodata(cbind(xstemp,ystemp,Ztemp)) Geodatatemp<- as.geodata(cbind(xstemp,ystemp,ztemp- zmodtemp)) vario2temp<- variog(messages.screen=FALSE,geodatatemp</pre>	<pre>,max.dist=250,pairs.min=2,uvec=seq(0,250,by =30)) varioesttemp<- variofit(messages.screen=FALSE,minimisation n.function ="nlm",vario2temp,ini.cov.pars=c(250,5),fix.nu g get=TRUE,fix.kappa=TRUE,max.dist=vario2 \$max.dist) Ktemp<-krige.conv(Geodatatemp, loc=cbind(xs[no],ys[no]),krige=krige.control(t ype.krige="ok",trend.d=~(emistattemp+p ropstattemp),trend.l=~(emistat[no]+propstat[n o]),obj.model=varioesttemp),output=s.out) #print(c(Ktemp\$predict,(z[no]-zmod[no]))) RMSE<-c(mean((Ktemp\$predict-(z[no]- zmod[no]))^2),RMSE) } RMSE<-sqrt(sum(RMSE)/(length(xs)-1)) print('RMSE=') print(RMSE) (...) #----- #validation grace a des stations supplementaires</pre>
--	---


```

#-----
Pour un certain nombre de stations choisies,
on compare la valeur obtenue par krigeage
avec la vraie valeur. Ces stations sont
différentes de celles utilisées pour le krigeage.
coordstat2<-
matrix(scan("donnees/stationsverif.dat"),ncol
=2,byrow=TRUE)
lambda2<-coordstat2[,1]
phi2<-coordstat2[,2]

debvalid<-dates("01/01/01")
ttvalid<-
matrix(scan("donnees/Moyenne_hora_2001_
O3.txt",skip=1),ncol=23,byrow=TRUE)
lignevalid<-(as.numeric(jour-
debvalid))*24+heure

valid<-ttvalid[lignevalid,2:23]
valid<-valid[-c(4,10,12,17,18,20,22)]
valid<-recodnan(valid,-999,NA)
indnan2<-which(!is.na(valid))

transit2<-ll2mtu(sud,ouest,lambda2,phi2)
xs2<-transit2$x[indnan2]/1000
ys2<-transit2$y[indnan2]/1000

emistat2<-
interpp(emissions[,1],emissions[,2],emissions
[,3],phi2,lambda2)
emistat2<-emistat2$z[indnan2]
indnane2<-which(!is.na(emistat2))
emistat2<-emistat2[indnane2]

zmod2<-
interpstations(determ,sud,petitpas,grx,gry,grx
2,gry2,x2,y2,phi2,lambda2,vent,cotes)

zmod2<-zmod2[indnan2]
#zmod2<-
interpp(y2,x2,determ,lambda2,phi2,ncp=0)
#zmod2<-zmod2$z[indnan2]

propstat2=rep(0,15)
i1=rep(0,15)
i2=rep(0,15)
for (i in 1:15){
  i1[i]=max(which(grx<phi2[i]))
  i2[i]=min(which(ggry<lambda2[i]))-1
  propstat2[i]=Propurb[i2[i],i1[i]]
}

propstat2=propstat2[indnan2]
propstat2=propstat2[indnane2]
xs2<-xs2[indnane2]
ys2<-ys2[indnane2]
zmod2<-zmod2[indnane2]
valid=valid[indnan2]
valid<-valid[indnane2]
lambda2=lambda2[indnan2]
phi2=phi2[indnan2]

est<-krige.conv(output=s.out
,Geodata,loc=cbind(xs2,ys2),krige=krige.cont
rol(type.krige="ok",trend.d=~emistat+propsta
t,tren
d.l=~emistat2+propstat2,obj.model=varioest))
est<-est$predict

print(cbind(lambda2[indnane2],phi2[indnane2]
),est+zmod2,valid))

crit=sqrt(sum((est+zmod2-
valid)^2)/length(xs2))
print(crit)

```

4.2 FORMAT DES DONNEES D'ENTREE

4.2.1 Observations

Le nombre de stations de mesure considérées pour le krigeage est spécifié en dur dans les programmes de krigeage (paramètre nbstations).

Chaque ligne du fichier donnees/obs_gr2.dat contient les concentrations moyennes horaires observées sur l'ensemble des stations de mesure considérées pour le krigeage. Les données manquantes sont identifiées par le champ "-999".

Extrait du fichier:

5.9000000e+01	4.6000000e+01	7.0000000e+01	6.2000000e+01	5.3000000e+01
6.7000000e+01	5.3000000e+01	7.2000000e+01	8.1000000e+01	6.1000000e+01
6.2000000e+01	6.3000000e+01	5.6000000e+01	5.6000000e+01	5.6000000e+01
5.3000000e+01	-9.9900000e+02	9.5000000e+01	8.6000000e+01	8.4000000e+01
7.8000000e+01	6.4000000e+01	6.8000000e+01	6.7000000e+01	8.3000000e+01
9.1000000e+01	7.6000000e+01	8.3000000e+01	4.8000000e+01	5.4000000e+01
7.1000000e+01	5.0000000e+01	7.6000000e+01	6.0000000e+01	5.0000000e+01
6.7000000e+01	5.3000000e+01	-9.9900000e+02	8.0000000e+01	5.8000000e+01
6.0000000e+01	6.8000000e+01	6.4000000e+01	7.1000000e+01	5.5000000e+01
5.0000000e+01	-9.9900000e+02	9.5000000e+01	8.4000000e+01	8.3000000e+01
7.4000000e+01	6.0000000e+01	5.7000000e+01	6.4000000e+01	7.9000000e+01
9.3000000e+01	7.7000000e+01	8.4000000e+01	4.8000000e+01	6.3000000e+01

Attention: les concentrations observées sont supposées être données en $\mu\text{g}/\text{m}^3$.

La localisation géographique des stations de mesure considérées pour le krigeage est "renseignée" dans le fichier donnees/coord_gr2.dat: la ième ligne contient la latitude et la longitude d'une station de mesure. Attention à la cohérence entre fichiers d'observations et fichier de coordonnées des stations: la ième colonne du fichier d'observations donnees/obs_gr2.dat contient la valeur de la concentration moyenne horaire mesurée par la station dont les coordonnées géographiques figurent sur la ième ligne du fichier de coordonnées donnees/coord_gr2.dat.

Il faut spécifier en dur dans les programmes de krigeage le jour de référence des données d'observations (paramètre deb, i.e. la date de la veille du premier jour pour lequel on dispose de données d'observations), ainsi que le jour et l'heure pour lesquels on veut effectuer le krigeage (paramètres jour et heure).

Enfin, le fichier de données d'observations doit être complet (i.e. "sans trou" dans les dates).

4.2.2 Concentrations simulées

Les informations sur la grille modèle sur laquelle sont simulées les concentrations de l'espèce cartographiée sont spécifiées en dur dans les programmes de krigeage (paramètres relatifs aux limites du domaine d'étude: sud, nord, ouest, est; paramètres relatifs au pas de la grille modèle: paslat, paslong).

Attention: les concentrations simulées sont supposées être données en ppb. Le paramètre de conversion (unite) est défini en dur dans les programmes de krigeage: unite = 1.997 pour l'ozone. Si les sorties modèles sont effectuées directement en $\mu\text{g}/\text{m}^3$, il faut changer la valeur de ce paramètre: unite = 1.000.

Chaque ligne du fichier donnees/CONT200211J-AIRPL13 contient la date (codée sous la forme AAAAMMJJHH; l'heure est donnée en heure TU, de 00 à 23 heures) en premier champ; les concentrations moyennes horaires simulées sur le reste de la ligne.

Extrait du fichier:

2001050101 43.8 43.6 43.7 43.7 45.5 40.6 35.8 38.8 42.0 44.9 46.5 49.0 47.2 43.8
 44.0 43.7 42.5 42.8 33.7 33.4 36.3 40.1 43.1 44.5 47.1 45.6 45.1 44.6 43.8 41.6 39.8
 30.7 29.7 34.3 36.2 40.3 42.4 44.5 43.8 45.3 44.2 42.8 40.4 36.6 26.0 24.1 30.4 32.3
 35.7 38.7 42.3 40.7 44.5 43.5 42.5 38.8 32.0 27.8 26.0 26.3 28.9 33.4 37.5 38.5 38.9
 44.2 43.9 42.8 36.9 32.2 31.6 32.4 28.5 31.6 33.6 37.6 38.2 41.3 45.2 45.6 43.8 37.0
 34.1 33.9 33.8 37.2 36.0 36.5 37.3 39.0 42.4 46.2 46.2 43.9 38.3 35.2 34.9 35.0 40.5
 42.6 42.4 42.1 45.3 46.1 46.2 45.8 43.0 39.2 36.4 35.5 35.6 40.1 42.1 44.3 46.7 48.3
 44.7 45.3 44.6 43.1 39.9 37.6 36.9 36.3 40.1 40.9 41.2 43.1 45.7 42.6 45.1 44.2 42.8
 41.6 38.8 38.5 38.1 39.5 40.9 40.1 37.9 40.8 39.2 45.0 44.6 44.9 43.4 41.2 40.2 39.6
 40.0 40.5 40.5 36.7 36.0 37.8 45.7 45.6 47.0 46.6 43.3 42.3 40.8 41.1 42.4 42.7 37.9
 35.4 37.1 45.7 46.0 47.9 48.1 46.0 44.0 41.2 39.9 41.3 40.6 39.4 32.8 31.1 46.1 46.2
 47.7 48.3 47.0 45.2 42.5 40.1 39.5 39.5 36.3 33.3 31.8 44.7 44.5 46.4 48.0 47.7 47.4
 45.3 42.7 39.1 37.3 36.0 36.0 35.6 43.1 42.7 44.3 47.0 48.5 49.5 48.6 45.5 40.8 36.4
 34.9 33.3 30.5

Le jour de référence des données modèles est spécifié en dur dans les programmes de krigeage (paramètre deb, i.e. la date de la veille du premier jour pour lequel on dispose de données modèles (et d'observations)), ainsi que le jour et l'heure pour lesquels on veut effectuer le krigeage (paramètres jour et heure).

Enfin, le fichier de données modèles doit être complet (i.e. "sans trou" dans les dates).

4.2.3 Données d'émissions spatialisées

Chaque ligne du fichier donnees/CITEPA_GRILLE_SGO.txt contient les coordonnées (longitude, latitude) du domaine sur lequel sont disponibles les données d'émissions et les émissions des espèces SO₂, CO, COV, NH₃ et NO_x sur le reste de la ligne. La première ligne du fichier est réservée pour la définition des données qui figurent dans le fichier.

Extrait du fichier:

Longi	Lati	SO2	CO	COV	NH3	NOX
-2.5	47.3	2542.9	4786.2	2660.5	26348.2	1853.7
-2.5	47.5	12.1	328.7	116.2	1580.0	86.9
-2.3	46.7	5.7	190.2	63.6	1127.0	47.3
-2.3	46.9	4.6	154.1	51.5	913.1	38.3
-2.3	47.1	6.6	204.0	69.6	1123.3	51.9
-2.3	47.3	4189.6	8104.9	4457.6	44453.0	3110.2
-2.3	47.5	55.1	1502.2	531.1	7220.0	397.3
-2.1	46.7	6.1	205.1	68.5	1215.2	51.0

Sous réserve de modifier les programmes de krigeage, il est possible d'utiliser d'autres données d'émissions.

4.2.4 Données d'urbanisation spatialisées

Les informations (nombre de points de longitude et latitude, résolution, limites sud, nord, ouest, est) sur le domaine sur lequel sont disponibles les données de taux d'urbanisation sont spécifiées en dur dans les programmes de krigeage. Chaque ligne du fichier donnees/propurb2.dat contient le taux d'urbanisation par point de grille.

4.2.5 Observations utilisées pour la validation

Chaque ligne du fichier donnees/Moyenne_hora_2001_O3.dat contient les concentrations moyennes horaires observées sur l'ensemble des stations de mesure considérées pour la validation. Les données manquantes sont identifiées par le champ "-999". La première ligne du fichier est réservée pour la définition des données qui figurent dans le fichier.

Extrait du fichier:

Identite	BASTO3	CHAUO3	GIROO3	JAPLO3	JPERO3	MONCO3	PROVO3
RICAO3	SLUCO3	BLUMO3	LESSO3	ALARO3	BARTO3	MONPO3	VEGAO3
LACMO3	EXUPO3	BAIRO3	VICTO3	LAENO3	FOLIO3	MAZAO3	
010120010100	46	38	49	41	-999	60	44
46	50	-999	-999	53	56	48	39
010120010200	45	37	48	37	-999	60	43
43	47	-999	-999	52	56	48	37
010120010300	44	40	50	43	-999	59	44
46	48	-999	-999	52	56	49	37
010120010400	45	38	50	42	-999	58	44
46	46	-999	-999	52	54	47	39
010120010500	44	36	48	40	-999	57	43
45	45	-999	-999	52	54	46	41

Attention: les concentrations observées sont supposées être données en µg/m3.

La localisation géographique des stations de mesure considérées pour la validation est indiquée dans le fichier donnees/stationsverif.dat: la ième ligne contient la latitude et la longitude d'une station de mesure. Attention à la cohérence entre fichiers d'observations et fichier de coordonnées des stations: la (i+1)ème colonne du fichier d'observations donnees/Moyenne_hora_2001_O3.dat contient la valeur de la concentration moyenne horaire observée en la station dont les coordonnées géographiques figurent sur la ième ligne du fichier de coordonnées donnees/stationsverif.dat.

Il faut spécifier en dur dans les programmes de krigage le jour de référence des données d'observations utilisées pour la validation (paramètre debvalid, i.e. la date du premier jour pour lequel on dispose de données d'observations).

Enfin, le fichier de données d'observations utilisées pour la validation doit être complet (i.e. "sans trou" dans les dates).

4.2.6 Observations d'une espèce additionnelle utilisées pour le cokrigage

Chaque ligne du fichier donnees/Moyenne_hora_2001_NO2.dat contient les concentrations moyennes horaires observées sur l'ensemble des stations de mesure considérées pour le cokrigage. Les données manquantes sont identifiées par le champ "-999". La première ligne du fichier est réservée pour la définition des données qui figurent dans le fichier.

Extrait du fichier:

Identite	CHAUN2	EAUXN2	GIRON2	JAPLN2	JPERN2	MONCN2	PROVN2							
RICAN2	SLUCN2	BLUMN2	LESSN2	ALARN2	BARTN2	MONPN2	EXUPN2							
BAIRN2	VICTN2	LAENN2	MAZAN2											
010120010100	11	-999	7	9	-999	-999	7	-999	4	4	6	6	13	
9 84	9	10	-999											
010120010200	11	-999	6	13	-999	-999	7	-999	4	4	6	6		
14 11 85	8	9	-999											
010120010300	10	-999	5	7	-999	-999	6	-999	4	3	6	5	13	
9 85	7	7	-999											
010120010400	10	-999	5	8	-999	-999	6	-999	4	3	7	5	11	
11 84	6	6	-999											
010120010500	12	-999	6	9	-999	-999	6	-999	6	8	8	6	12	
10 74	7	6	-999											

Attention: les concentrations observées sont supposées être données en µg/m3.

La localisation géographique des stations de mesure considérées pour la validation est "renseignée" dans le fichier donnees/stationscokri.dat: la ième ligne contient la latitude et la longitude d'une station de mesure. Attention à la cohérence entre fichiers d'observations et fichier de coordonnées des stations: la (i+1)ème colonne du fichier d'observations donnees/Moyenne_hora_2001_NO2.dat contient la valeur de la concentration moyenne horaire observée en la station dont les coordonnées géographiques figurent sur la ième ligne du fichier de coordonnées donnees/stationscokri.dat.

Il faut spécifier en dur dans les programmes de krigeage le jour de référence des données d'observations utilisées pour la validation (paramètre debvalid, i.e. la date du premier jour pour lequel on dispose de données d'observations).

Enfin, le fichier de données d'observations utilisées pour la validation doit être complet (i.e. "sans trou" dans les dates).

4.3 GESTION DES DONNEES DE SORTIE: CREATION DE CARTES AVEC GMT

A chaque exécution d'un programme de krigeage, on récupère en sortie des fichiers de suffixe "xyz" contenant:

- Les concentrations simulées par le modèle et interpolées sur la grille fine
- Les concentrations estimées par krigeage des concentrations sur la grille fine
- La variance de l'erreur de krigeage
- Les observations disponibles.

Le format de ces fichiers est le suivant: longitude, latitude, champ. Les concentrations sont en µg/m3.

Le tableau ci-dessous contient, pour chaque méthode de krigeage, la liste des fichiers "xyz" créés en vue de la réalisation de cartes.

Le script "gmt/cartes.sh" permet la réalisation de cartes avec le logiciel GMT (Generic Mapping Tool), logiciel en accès libre et disponible pour les systèmes d'exploitation les plus courants.

Méthode de krigeage Nom du programme	Concentrations simulées par le modèle	Concentrations estimées par krigeage	Variance de l'erreur de krigeage	Concentrations observées
Krigeage des observations R/krigeageobs.R	gmt/model.xyz	gmt/Kobs_predict.xyz	gmt/Kobs_erreur.xyz	gmt/obs.xyz
Krigeage lognormal des observations R/krigeagelognormal.R	gmt/model.xyz	gmt/Klog_predict.xyz	gmt/Klog_erreur.xyz	
Krigeage des observations avec le modèle en DE ⁹ R/derive.R	gmt/model.xyz	gmt/derive_predict.xyz	gmt/derive_erreur.xyz	
Krigeage des observations avec le modèle et le taux d'urbanisation en DE R/derivemodpurb.R	gmt/model.xyz	gmt/derivemodpurb_predict.xyz	gmt/derivemodpurb_erreur.xyz	
Krigeage des innovations R/krigeageinnov.R	gmt/model.xyz	gmt/Kinnov_predict.xyz	gmt/Kinnov_erreur.xyz	
Krigeage des innovations avec les émissions en DE R/innovderiveemi.R	gmt/modelemi.xyz	gmt/innovderiveemi_predict.xyz	gmt/innovderiveemi_erreur.xyz	
Krigeage des innovations avec le taux d'urbanisation en DE R/innovderiveurb.R	gmt/model.xyz	innovderiveurb_predict.xyz	innovderiveurb_erreur.xyz	

⁹ DE = dérive externe

Krigeage des innovations avec les émissions et le taux d'urbanisation en DE R/innovderivedouble.R	gmt/modelemi.xyz	gmt/innovderive_double_predict.xyz	gmt/innovderive_double_erreur.xyz	
Cokrigeage R/cokrigeage.R	gmt/modelcokri.xyz	gmt/cokrigeage.xyz gmt/cokrigeageN02.xyz	gmt/cokrigeage-erreurs.xyz	

D'autres logiciels permettant la réalisation de cartes peuvent être utilisés, à condition d'adapter le format des fichiers "xyz".

5 LISTE DES ANNEXES

Repère	Désignation précise	Nb/N° pages
1	Annexe 1: Organismes "clients" du système PREV'AIR au 22/12/2003	2
2	Annexe 2: Procédure de chargement de R et des modules complémentaires (chron, akima, geoR, gstat et splancs)	1
3	Annexe 3: Estimation du variogramme dans le cas du cokrigage	1

Annexe 1: Organismes "clients" du système PREV'AIR au 22/12/2003

A la date de production du présent rapport, 17 organismes – dont 14 AASQA et 2 organismes italiens en charge de la surveillance de la qualité de l'air à l'échelle locale - disposent d'un compte sur le serveur PREV'AIR sur lequel sont déposés quotidiennement les fichiers de données numériques de prévision de la qualité de l'air.

Pour chaque organisme, les données de prévision sont extraites quotidiennement sur un domaine spécifié au départ par le demandeur, de sorte que les fichiers de données numériques de prévision de la qualité de l'air sont effectivement propres à chaque utilisateur.

Ces fichiers sont de deux types :

- Données 2D surfaciques pour un nombre restreint de polluants – typiquement ceux pour lesquels les cartes de prévision PREV'AIR sont diffusées, soit, à la date de production du présent rapport, ozone, dioxyde d'azote et particules de diamètre inférieur à 2.5µm.
- Données 3D pour un nombre plus important (soixante dix dans la version actuelle du modèle) d'espèces chimiques modèles prises en compte dans le mécanisme chimique MELCHIOR.

Les utilisations possibles de ces données à l'échelle locale sont:

- Cartographie
- Prévision / simulation déterministe: définition de conditions aux limites (CL) des domaines "utilisateur" pour la mise en œuvre à l'échelle locale de modèles numériques de simulation de la qualité de l'air
- Prévision / simulation statistique

Le tableau ci-dessous dresse la liste des organismes utilisateurs du système PREV'AIR au 22/12/2003; pour chaque organisme est précisée l'utilisation envisagée des données, telle qu'elle nous a été mentionnée dans le formulaire d'inscription au service de mise à disposition des données numériques PREV'AIR.

Nom de l'organisme	Domaine de compétence	Utilisation envisagée des données
AIRCOM	Basse-Normandie	Pas d'utilisation précisée
AIRLOR	Lorraine	Aide à la prévision; CL pour la modélisation
AIRLR	Languedoc-Roussillon	Prévision experte [O3] J+1
AIRMARAIX	Provence-Alpes-Côte d'Azur	CL pour modèle régional PACA méso-échelle.
AIRNORMAND	Haute-Normandie	Pas d'utilisation précisée
AIRPARIF	Ile-de-France	CL
AIRPL	Pays de la Loire	Modélisation déterministe SAMAA (ACRI); cartographie de l'ozone à l'échelle régionale
ARPA-EMR (EmilieRomagne)	Emilie Romagne (Italie)	Aide à la prévision de la qualité de l'air; CL du modèle régional de simulation de la qualité de l'air
ARPA-Piemonte	Piémont (Italie)	Aide à la prévision de la qualité de l'air; CL du modèle régional de simulation de la qualité de l'air

ASPA	Alsace	CL du modèle de prévision CHIMERE-Régional utilisé quotidiennement
ATMO AUVERGNE	Auvergne	Modélisation (code CALGRID)
ATMO PC	Poitou-Charentes	Adaptation locale de CHIMERE (cartographie)
GIERSA	Rhône-Alpes	Modélisation
LIGAIR	Centre	Cartographie Prévision
LIMAIR	Limousin	Prévision
Ecole des Mines de Douai	-	Recherche
ORAMIP	Midi-Pyrénées	Couplage avec CHIMERE- Régional

Annexe 2: Procédure de chargement du logiciel R et des modules complémentaires (chron, akima, geoR, gstat et splancs)

1) Dans Windows

Aller sur la page:

<http://mirrors.toulouse.inra.fr/R/>

Chargement de R

Dans la rubrique *precompiled binary distributions*, sélectionner :
Windows (95 and later)

Cliquer sur :
base

puis sur:
rw1081.exe

Chargement et utilisation des packages dans R

Cliquer sur *Package sources* dans la rubrique *software* de la page d'accueil

Cliquer sur le nom du package souhaité

Cliquer à nouveau sur le nom du package et l'enregistrer en tant que fichier .zip (ex : akima.zip)

Pour charger le package, ouvrir R
puis installer le package par le menu "Packages/ Install package from local zip file"

Pour activer le package, taper dans R la commande :
Library(nom_du_package) (exemple : library(akima))

2) Sous UNIX

Aller sur la page:

<http://mirrors.toulouse.inra.fr/R/>

Chargement de R et des packages

Dans la rubrique *Source code for all platforms* :

Cliquer sur *R-1.8.1.tgz* (ou autre version) pour récupérer le code source de R

Cliquer sur *contributed packages* puis sur le nom du package souhaité

Installation et décompression

Installer : rpm -ivh "nom du package"

Vérification : rpm -Vp "nom du package"

Décompresser les archives

Compiler le programme : ./configure, make, make install, make clean...

Annexe 3: Estimation du variogramme dans le cas du cokrigeage

Le calcul et la modélisation des variogrammes simple et du variogramme croisé est réalisée en plusieurs étapes, à l'aide des fonctions *gstat*, *variogram* et *fit.lmc* du package *gstat*.

La fonction *gstat* permet de créer progressivement un objet contenant toutes les informations nécessaires au calcul et à la modélisation du variogramme :

a) On fournit les données de NO₂ à l'objet *geodata* :

```
geodata<-gstat(id="NO2"),formula=NO2~1,locations=~xs+ys,data=geodNO2)
```

b) On lui ajoute les données d'innovations :

```
geodata<-gstat(geodata,id="ozone"),formula=ozone~1,locations=~xs+ys,data=geod)
```

c) On entre un modèle de variogramme par défaut pour le NO₂ (éventuellement ajusté à l'œil) :

```
geodata<-gstat(geodata,id="NO2"),model=vgm(200, "Exp",150,nugget=100))
```

200 : palier du variogramme

« Exp » : modèle mathématique utilisé (ici exponentiel)

150 : portée du variogramme

nugget=100 : effet de pépite

d) On entre un modèle de variogramme par défaut pour les innovations :

```
geodata<-gstat(geodata,id="ozone"),model=vgm(200, "Exp",150,nugget=100))
```

e) On entre un modèle de variogramme par défaut pour le NO₂ et les innovations (modèle de corégionalisation) :

```
geodata<-gstat(geodata,id=c("NO2","ozone"),model=vgm(200, "Exp",150,nugget=100))
```

Remarque :liste des principaux modèles mathématiques

Exp : exponentiel

Sph : sphérique

Gau : gaussien

Lin : linéaire

f) On calcule les variogrammes simples et croisé associé à l'objet *geodata* précédemment créé :

```
vario<-variogram(geodata,width=30)
```

width : largeur des classes de distance (un point du variogramme associé à la distance h_1 est calculé à partir des couples de stations distantes de $h_1 \pm \text{width}$)

g) On ajuste sur ces variogrammes un modèle linéaire de corégionalisation

```
varioest<-fit.lmc(vario,geodata)
```