SOPHIE BAILLARGEON

Le krigeage :
revue de la théorie et application à l'interpolation spatiale de données de précipitations

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

Faculté des sciences et de génie
UNIVERSITÉ LAVAL
QUÉBEC

Avril 2005

©Sophie Baillargeon, 2005

Résumé

Le krigeage est une méthode stochastique d'interpolation spatiale qui prévoit la valeur d'un phénomène naturel en des sites non échantillonnés par une combinaison linéaire sans biais et à variance minimale des observations du phénomène en des sites voisins. Ce mémoire se consacre à l'étude de cette méthode. Elle y est d'abord comparée à d'autres méthodes d'interpolation spatiale et ses fondements mathématiques sont examinés. La résolution des équations du krigeage est donc détaillée et commentée. L'analyse variographique, étape préalable au krigeage, est aussi présentée. En plus d'avoir pour objectif l'approfondissement de la théorie du krigeage, ce mémoire vise à expliciter son utilisation. Ainsi, une méthodologie de mise en oeuvre du krigeage est proposée et illustrée. Les performances du krigeage sont ensuite comparées à celle d'autres méthodes, et ce, pour résoudre une problématique d'interpolation spatiale multivariable de données de précipitations dans un cadre de modélisation hydrologique.

Avant-propos

Je tiens tout d'abord à remercier les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire. Premièrement, merci à mon directeur de recherche, monsieur Louis-Paul Rivest, professeur au Département de mathématiques et statistique de l'Université Laval. Il s'est toujours montré disponible et intéressé par mon projet. Il a su me diriger tout en me laissant une grande liberté dans mes travaux, ce que j'ai beaucoup apprécié. Merci aussi à ma co-directrice de recherche, madame Jacynthe Pouliot, professeure au Département des sciences géomatiques de l'Université Laval. Elle a suscité mon intérêt pour la géostatistique et m'a proposé ce projet. Ses précieux conseils de rédaction m'ont également permis d'écrire un document qui, je l'espère, pourrait intéresser autant un géomaticien qu'un statisticien. De plus, je remercie monsieur Vincent Fortin, chercheur à l'Institut de recherche d'Hydro-Québec. C'est grâce à lui que la problématique d'interpolation de données de précipitations dans un cadre de modélisation hydrologique a pu être traitée. Il a collaboré activement à mon mémoire en fournissant temps, conseils, données et en acceptant d'en être un examinateur.

D'autres personnes ont été impliquées dans ce projet, notamment madame Josée Fitzback, étudiante au doctorat en sciences géomatiques de l'Université Laval. Je la remercie d'avoir établi les contacts qui ont rendu le projet possible. Travailler avec elle fut toujours un plaisir. Merci également à messieurs Jean-François Mahfouf et Pierre Pellerin, météorologues à Environnement Canada, pour l'intérêt qu'ils ont porté envers le projet. J'exprime de plus ma gratitude envers le CRSNG, qui a financé mes travaux de recherche en m'octroyant une bourse d'étude supérieure.

En finissant, je désire bien sûr remercier les membres de ma famille, mes parents et mes soeurs, qui ont toujours encouragé mes études et grâce à qui j'ai grandi dans un milieu épanouissant. Mes derniers remerciements s'adressent à mon amoureux, Mathieu, qui est toujours un stimulant à aller de l'avant. Il m'a écouté avec intérêt parler des hauts et des bas de ma rédaction, m'a aidé à garder le sourire en temps de stress et a répondu à mes nombreuses questions LATEX. Je le remercie simplement d'avoir été là.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

Des données à référence spatiale sont de plus en plus exploitées, et ce, dans divers domaines de recherche. Qu'il s'agisse de précipitations mesurées à des stations météorologiques, de la densité d'un minerai dans des échantillons de sol ou de la concentration de gaz carbonique dans l'air en certains sites, ces données possèdent toutes un point en commun : elles sont localisées dans l'espace géographique. Le traitement statistique de ce type de données demande une attention particulière car l'hypothèse classique selon laquelle les observations sont indépendantes et identiquement distribuées est rarement vérifiée. Des méthodes statistiques adaptées à l'analyse de données à référence spatiale ont été développées (Ripley, 1981; Cressie, 1993). Ce mémoire porte sur l'une de ces méthodes, le krigeage, développée par Matheron (1962, 1963a,b). Le krigeage sert à effectuer de l'interpolation spatiale, c'est-à dire qu'il permet de prévoir la valeur prise par un phénomène naturel un site à partir d'observations ponctuelles de ce phénomène en des sites voisins.

1.1 Objectifs du mémoire

L'objectif principal de ce mémoire est l'étude du krigeage. La première question de recherche qui a motivé ce travail est donc simplement : qu'est-ce que le krigeage ? La réponse à une question en amenant une autre, il faut d'abord se demander en quoi consiste l'interpolation spatiale. Évidemment, ces travaux de recherche ne visent pas seulement à définir le krigeage. Ils ont pour but de comprendre la méthode et d'apprendre comment l'utiliser. Ainsi, des sous-objectifs du mémoire sont d'approfondir les fondements mathématiques du krigeage et d'examiner la méthodologie générale de

mise en oeuvre du krigeage. Les fondements du krigeage se basent sur une connaissance de la structure de dépendance spatiale des données. Cependant, en pratique, cette structure n'est pas connue. Elle est estimée lors de l'analyse variographique. Cette étape préalable au krigeage se doit donc aussi d'être exposée. Pour compléter cette étude du krigeage, nous nous sommes demandé si le krigeage permettait de répondre efficacement à une problématique d'interpolation spatiale. La problématique traitée ici concerne l'interpolation de données de précipitations dans le but de fournir ces données en entrée à un modèle de simulation hydrologique.

1.2 Structure du mémoire

Ce mémoire est divisé en fonction des objectifs énoncés ci-dessus. Le chapitre 2 donne une première idée générale de ce qu'est le krigeage. L'interpolation spatiale y est d'abord définie, puis les principales méthodes d'interpolation spatiale, dont le krigeage, sont décrites sommairement. Ce chapitre présente donc le krigeage en le situant par rapport à ses méthodes rivales. Ensuite, les chapitres 3 et 4 approfondissent les fondements théoriques de l'analyse variographique et du krigeage respectivement. Ainsi, les trois chapitres suivant l'introduction permettent de connaître et comprendre le krigeage. Ensuite, le chapitre 5 jette les ponts entre la théorie et la pratique en proposant une démarche d'utilisation du krigeage. Cette démarche y est illustrée avec des données de précipitations. Ces données sont en fait une tranche d'un plus gros jeu de données exploité dans un article rédigé pour les actes du colloque *Géomatique 2004* par Baillargeon *et al.* (2004). Cet article est inséré dans ce mémoire ; il constitue le chapitre 6. Le but de l'article est de comparer la performance de certaines méthodes d'interpolation pour répondre à la problématique de préparation d'intrants pour un modèle hydrologique. Les méthodes comparées sont quatres techniques de krigeage, soit le krigeage ordinaire, le krigeage universel, le krigeage avec dérive externe et le cokrigeage ordinaire, deux techniques de régression locale et une méthode témoin déterministe. Dans la conclusion, des parallèles sont établis entre les travaux du chapitre 6 et d'autres études semblables dans la littérature.

Chapitre 2

Interpolation spatiale

Ce chapitre présente une revue des principales méthodes d'interpolation spatiale afin de comprendre leurs principes de base et de comparer le krigeage aux autres méthodes. Il ressort alors que malgré les nombreux points communs entre les méthodes, le krigeage se distingue par sa prise en compte de la structure de dépendance spatiale des données.

2.1 Définition et notation

L'interpolation spatiale est un traitement mathématique parfois utile lors de l'étude d'un phénomène naturel qui se déploie continûment sur le territoire. La région de l'espace géographique concernée par cette étude est ici appelée « champ » et notée D. Le phénomène naturel examiné est représenté par une certaine mesure localisée sur le territoire. Par exemple, pour étudier un gisement d'or, on peut utiliser la mesure de la densité du minerai dans le sol. Tel que l'a initié Matheron (1962), une telle mesure est nommée « variable régionalisée » et elle est vue comme une fonction numérique définie sur le champ D. Elle sera notée $\{z(s), s \in D\}$ où $s = (x, y)$ représente un point du champ identifié par ses coordonnées géographiques. La valeur de cette fonction en un point particulier s_i, notée $z(s_i)$, porte le nom de « valeur régionalisée » (Wackernagel, 2003, p.41).

En pratique, on n'a pas une valeur régionalisée en tout point s du champ D. Par exemple, pour un gisement d'or, la densité du minerai dans le sol n'est mesurée qu'en quelques sites où une carotte de sol est prélevée. Peu importe l'application en question, notons ces sites d'observation s_i avec $i = 1, ..., n$. L'interpolation spatiale répond au be-

soin de connaître la valeur d'une variable régionalisée en un site s_0 du champ D autre qu'un des sites d'observation. Elle se définit par la prévision de la valeur d'une variable régionalisée en un site où elle n'a pas été mesurée à partir des valeurs régionalisées observées $z(s_1)$ à $z(s_n)$ (Arnaud et Emery, 2000, p.20 ; Cressie, 1993, p.105). La valeur prédite en s_0 sera notée $\hat{z}(s_0)$. La définition énoncée ci-dessus est en fait celle de l'interpolation spatiale « point à point » : elle se base sur des données associées à des points de l'espace géographique pour effectuer une prévision elle aussi ponctuelle. Les autres types d'interpolation spatiale, tel que « point à surface » ou « point à ligne » (e.g. le tracé de courbes de niveaux), ne seront pas étudiés dans ce mémoire.

La variable régionalisée est l'outil de base en interpolation spatiale. Les méthodes d'interpolation s'appuyant uniquement sur cette entité mathématique sont dites déterministes car aucune notion probabiliste n'intervient dans la définition de variable régionalisée. L'interpolation spatiale peut aussi s'effectuer par une méthode stochastique. Dans ce cas, un deuxième niveau d'abstraction est effectué dans la modélisation du phénomène naturel. La variable régionalisée est vue comme une réalisation d'une fonction aléatoire $\{Z(s), s \in D\}$, aussi appelée processus stochastique, et toute valeur régionalisée $z(s_i)$ est considérée comme une réalisation d'une variable aléatoire $Z(s_i)$.

La figure 2.1 résume la notation présentée dans cette section. Elle représente du même coup la modélisation d'un phénomène naturel selon les deux niveaux d'abstraction décrits précédemment.

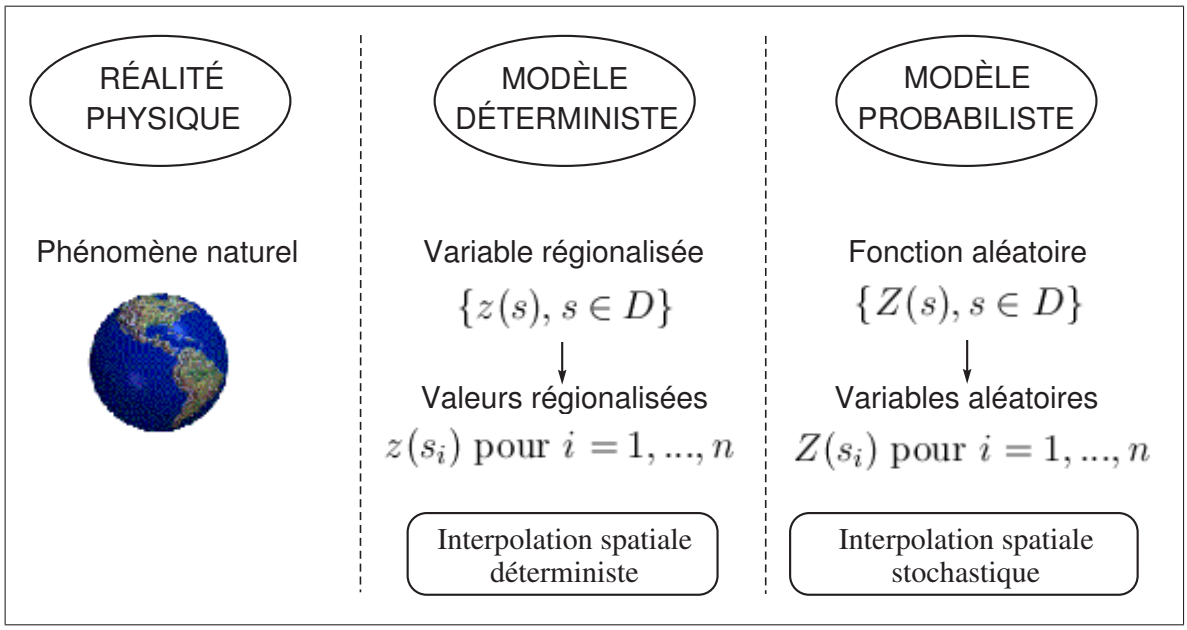

FIG. 2.1 – *Les deux niveaux d'abstraction en interpolation spatiale*

2.2 Revue des méthodes d'interpolation spatiale

Dans cette section, les principales méthodes d'interpolation spatiale sont recensées. Le livre de Arnaud et Emery (2000) a été le point de départ de cet inventaire. La terminologie proposée dans ce livre est employée ici. Trois grandes classes de méthodes déterministes ont été dénombrées : les méthodes barycentriques, les méthodes d'interpolation par partitionnement de l'espace et les splines. Du côté des méthodes stochastiques, les techniques de régression classique, de régression locale et de krigeage ont été listées. Contrairement aux méthodes déterministes, les méthodes stochastiques incorporent le concept de hasard. Elles proposent toutes un modèle probabiliste incluant un ou des termes d'erreurs aléatoires pour formaliser le comportement du phénomène naturel à l'étude. Grâce à cette modélisation, des erreurs de prévision peuvent être calculées. Les paragraphes suivants proposent une brève introduction à chacune de ces six classes de méthodes.

2.2.1 Méthodes barycentriques

Les méthodes d'interpolation déterministes de type barycentrique (Arnaud et Emery, 2000, p.67), aussi appelées « moyennes mobiles » (Ripley, 1981, p.36) ou « approximation de Kernel » (Myers, 1994), sont très intuitives. Elles prévoient la valeur d'une variable régionalisée en un point non échantillonné s_0 par une moyenne pondérée des valeurs régionalisées observées :

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i z(s_i) \text{ avec } \sum_{i=1}^{n} \lambda_i = 1.$$

Les poids λ_i sont contraints de sommer à la valeur 1 afin que la prévision ne présente pas de distorsion par rapport à la valeur réelle. Ces poids sont fonction de la distance euclidienne $|s_i - s_0|$ entre le site d'observation s_i et le site de prévision s_0 de façon à ce que les sites les plus proches aient plus d'influence dans l'interpolation. Souvent, un poids nul est accordé aux observations les plus éloignées. Ainsi, seules les observations localisées dans un certain voisinage de s_0, noté $V(s_0)$, sont prises en compte. La façon la plus simple de déterminer un voisinage est de prendre les n_0 sites d'observation les plus proches ou les sites tombant à l'intérieur d'un cercle centré en s_0 de rayon prédéterminé. Le champ peut aussi être divisé en quadrants ou en octants dont l'origine est s_0. Le voisinage peut alors comprendre les n_0^* sites d'observation le plus proche de s_0 pour chaque secteur.

Un exemple populaire de méthode barycentrique est la méthode de l'inverse de la

distance à une certaine puissance d. Par cette méthode, la prévision en un point s_0 prend la forme :

$$\hat{z}(s_0) = \sum_{i \in V(s_0)} \frac{1/|s_i - s_0|^d}{\sum_{i \in V(s_0)} 1/|s_i - s_0|^d} z(s_i) \quad , \quad d > 0.$$

2.2.2 Méthodes d'interpolation par partitionnement de l'espace

Les méthodes d'interpolation par partitionnement de l'espace (Arnaud et Emery, 2000, sections 2.1 et 2.2 ; Ripley, 1981, p.38) forment en fait un sous-ensemble des méthodes barycentriques. Ces méthodes se distinguent par l'utilisation d'un partitionnement du champ d'étude (Okabe *et al.*, 1992) afin de déterminer les poids des observations et le voisinage du point de prévision. Cette section présente donc d'abord des techniques de partitionnement de l'espace, puis des méthodes d'interpolation se basant sur ces techniques.

Il existe deux principaux types de partitionnement d'un champ D en régions disjointes à partir des sites d'observation : par polygones et par triangles. Le partitionnement par polygones porte plusieurs noms. Les principaux sont : « polygones de Thiessen » , « diagramme de Voronoi » , et « mosaïque ou tesselation de Dirichlet ». Ce partitionnement est formé en définissant, pour chaque site d'observation s_i, un polygone d'influence de sorte que chaque point du polygone soit plus proche de s_i que de tout autre site d'observation. Ce partitionnement est illustré par les lignes pleines de la figure 2.2. De son côté, le partitionnement par triangles, nommé triangulation, découpe le champ en triangles disjoints dont les sommets sont les sites d'observation. Différents critères existent pour sélectionner les sommets appartenant à un même triangle. Le plus connu de ces critères est celui de Delaunay . Il se base sur un partitionnement par polygones de Thiessen. Les sites d'observation ayant un côté de leurs polygones de Thiessen en commun sont reliés par une droite, formant ainsi la triangulation. La figure 2.2 illustre en lignes pointillées la triangulation de Delaunay correspondant aux polygones de Thiessen des sept sites d'observation présentés en exemple.

Plusieurs méthodes d'interpolation se basent sur un partitionnement de l'espace. La plus simple de ces méthodes est celle du **plus proche voisin**. La valeur régionalisée mesurée en un site d'observation est attribuée à tous les points localisés dans le polygone de ce site. Les polygones de Thiessen sont aussi à la base de l'**interpolation par voisinage naturel**, une méthode due à Sibson (1981). Par cette méthode, la prévision de la valeur régionalisée en s_0 prend la forme d'une moyenne pondérée des

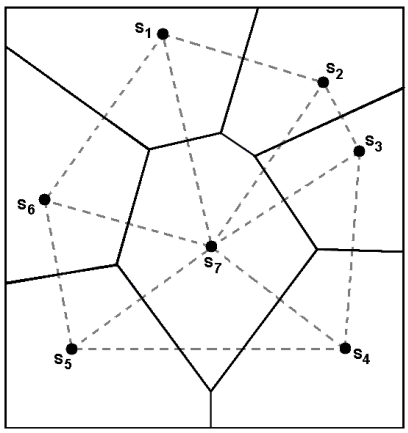

FIG. 2.2 – *Polygones de Thiessen (lignes pleines) accompagnés de la triangulation de Delaunay associée (lignes pointillées)*

valeurs régionalisées observées. La figure 2.3 illustre comment le poids de chacune des observations est déterminé. Préalablement, les polygones de Thiessen associés aux sites d'observation sont tracés (e.g. mosaïque de droite de la figure 2.3). Ensuite, les polygones de Thiessen sont reformés en ajoutant au champ le site s_0 pour lequel une prévision est voulue (e.g. mosaïque centrale de la figure 2.3). Cette nouvelle mosaïque est finalement superposée aux polygones de Thiessen initiaux sans le site s_0 (e.g. mosaïque de gauche de la figure 2.3). Le poids de l'observation en s_i est alors l'aire A_i de l'intersection entre le polygone de s_0 et le polygone initial de s_i divisée par l'aire totale du polygone de s_0. Dans l'exemple de la figure 2.3, la prévision est donc $\hat{z}(s_0) = \sum_{i=1}^{7} A_i z(s_i) / \sum_{i=1}^{7} A_i$ avec $A_3, A_4 = 0$.

À partir d'une triangulation, une interpolation s'effectue en ajustant une surface, souvent un polynôme, dans chaque triangle. Par exemple, en **interpolation linéaire**, des plans sont ajustés dans les triangles de Delaunay. La façon géométrique d'effectuer cette interpolation est illustrée par la figure 2.4. Il suffit de tracer le point de prévision s_0 dans le champ et de le relier aux trois sommets du triangle à l'intérieur duquel il est localisé. Ainsi, ce triangle est divisé en trois petits triangles. Seules les valeurs régionalisées des sites formant les sommets du grand triangle sont incluses dans l'interpolation. De plus, le poids de chacune de ces valeurs est égal à la portion de surface du grand triangle occupée par le petit triangle opposé au site. Par cette méthode, la prévision en s_0 pour l'exemple de la figure 2.4 est donc $\hat{z}(s_0) = \frac{A_1 z(s_1) + A_6 z(s_6) + A_7 z(s_7)}{A_1 + A_6 + A_7}$.

En général, les méthodes d'interpolation par partitionnement de l'espace possèdent les propriétés d'être locales et exactes. C'est donc dire qu'elles n'utilisent dans l'interpolation que les observations localisées assez près du point de prévision selon un certain

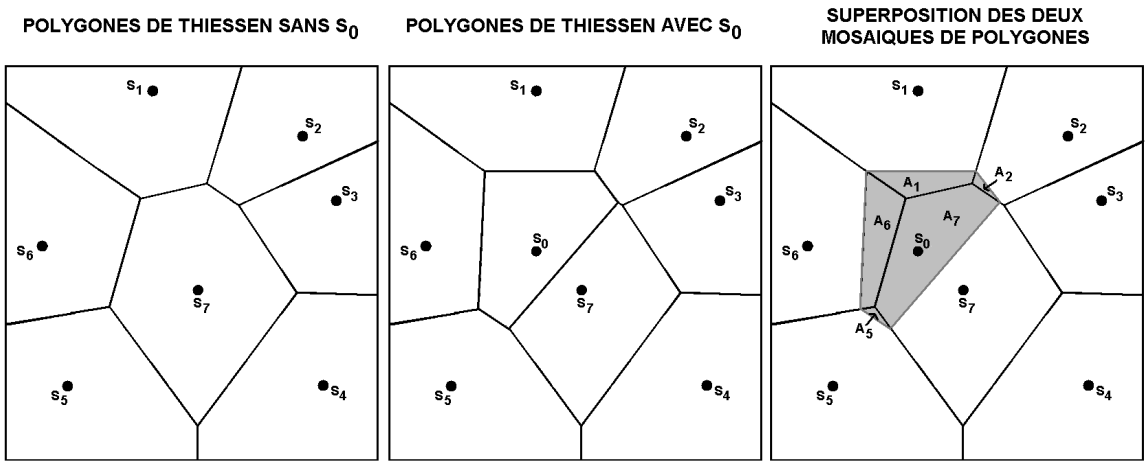

FIG. 2.3 – *Exemple de partitionnement à l'origine d'une interpolation par voisinage naturel*

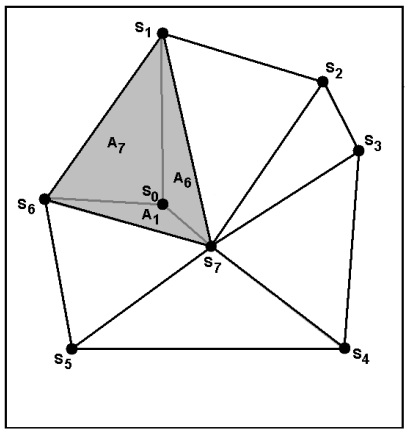

FIG. 2.4 – *Exemple de partitionnement à l'origine d'une interpolation linéaire*

critère de voisinage et que les prévisions qu'elles fournissent aux points d'observation sont égales aux observations. Ces méthodes sont déterministes car elles ne modélisent pas la variable régionalisée par une fonction aléatoire. Notons que les méthodes d'interpolation par partitionnement fournies en exemple ici sont les plus simplistes. Elles produisent des surfaces d'interpolation présentant, sauf exception, des changements abruptes. En outre, l'interpolation linéaire ne permet pas d'extrapoler en dehors de l'enveloppe convexe des sites. Plusieurs autres méthodes d'interpolation par partitionnement ne présentent pas ces inconvénients.

2.2.3 Splines

Pour clore la présentation des méthodes déterministes, mentionnons l'interpolation à l'aide de splines. Ce type d'interpolation ne s'effectue pas point par point comme avec les méthodes barycentriques. L'idée est plutôt d'ajuster une surface sur tout le champ D. Une spline est en fait une famille de fonctions régulières de courbure minimale. Il existe deux catégories de splines : les splines d'interpolation contraintes de passer par les points d'observation et les splines de lissage qui passent seulement à proximité de ces points. La spline de lissage du type « plaque mince » (Duchon, 1975, 1976), ou « thin plate spline » en anglais (Wahba, 1990, p.30 ; Cressie, 1993, p.181), est présentée ici en exemple. Il s'agit de la généralisation dans l'espace à deux dimensions d'une spline cubique de lissage (Hastie et Tibshirani, 1990, p.9). Cette spline est une fonction $\hat{z}$ de la forme :

$$\hat{z}(s) = \hat{z}(x, y) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{n} b_i e(s, s_i), \qquad s \in D$$

avec $e(s, s_i) = |s - s_i|^2 ln(|s - s_i|)$, qui minimise :

$$\sum_{i=1}^{n} [\hat{z}(s_i) - z(s_i)]^2 + \rho \int\int \{(\frac{\partial^2 \hat{z}}{\partial x^2})^2 + 2(\frac{\partial^2 \hat{z}}{\partial x \partial y})^2 + (\frac{\partial^2 \hat{z}}{\partial y^2})^2\} dxdy$$

où ρ est un paramètre de lissage fixé à priori. L'intégrale incluse dans ce critère à minimiser est appelée « énergie de flexion » . Elle réfère au concept physique d'énergie de flexion d'une mince plaque de métal. Les surfaces générées par ces splines peuvent donc être comparées à des plaques de métal flexible ajustées de façon à minimiser leurs degrés de courbure tout en passant à proximité des points d'observation. D'autres types de splines permettent d'effectuer de l'interpolation spatiale, notamment les fonctions multiquadratiques (Hardy, 1971), aussi appelées fonctions radiales de base (Myers, 1994), et les splines régularisées avec tension (Mitášová et Mitáš, 1993).

De plus, certaines splines permettent d'effectuer de l'interpolation spatiale multivariable, c'est-à-dire de prévoir la valeur d'une variable régionalisée en un site non échantillonné à partir des observations de cette variable, mais aussi à partir des observations d'autres variables régionalisées pertinentes. C'est le cas des splines laplaciennes dites partielles qui incorporent des sous-modèles linéaires fonctions de variables régionalisées auxiliaires. Ces splines sont mieux connues sous leur nom anglais « partial thin plate splines » (Hutchinson, 1991; Wahba, 1990, p.73). Pour leur part, les méthodes d'interpolation par partitionnement de l'espace et les méthodes barycentriques ne sont pas adaptées à une telle intégration de variables régionalisées auxiliaires.

2.2.4 Régression classique

Tout comme les splines, la régression classique permet d'effectuer une interpolation en ajustant une surface aux valeurs régionalisées observées. Cependant, la régression est une méthode stochastique. Elle suppose que la variable régionalisée à l'étude est une fonction aléatoire qui se décompose comme suit :

$$Z(s) = \mu(s) + \epsilon(s), \qquad s \in D \tag{2.1}$$

où $\mu(\cdot)$ est la structure déterministe pour l'espérance fonction de la localisation des observations et $\epsilon(\cdot)$ est une fonction aléatoire normale d'espérance nulle, de variance homogène et ne présentant pas de structure de dépendance spatiale. Ainsi, $\epsilon(\cdot)$ est un processus gaussien de bruit blanc représentant des erreurs de mesure indépendantes. La tendance $\mu(s)$ peut prendre plusieurs formes. La plus utilisée est un polynôme de degré d des coordonnées :

$$\mu(s) = \mu(x, y) = \sum_{l+k \leq d} \beta_{lk} x^l y^k. \tag{2.2}$$

En général, le degré du polynôme est inférieur ou égale à trois. En interpolation spatiale, ce modèle de régression est souvent appelé « surface de tendance » (Ripley, 1981, p.29). Les coefficients du modèle sont ajustés par une méthode quelconque d'estimation. Par exemple, la méthode du maximum de vraisemblance ou celle des moindres carrés peut être employée. Par les moindres carrés, les paramètres β_{lk} sont choisis de façon à minimiser :

$$\sum_{i=1}^{n} [\hat{z}(s_i) - z(s_i)]^2$$

où $\hat{z}(s_i) = \sum_{l+k \leq d} \hat{\beta}_{lk} x^l y^k$ est la valeur prédite par le modèle de régression au point d'observation s_i et $z(s_i)$ est la valeur régionalisée observée en ce point. D'autres formes pour la structure $\mu(\cdot)$ ont été suggérées. Par exemple, des fonctions trigonométriques pourraient être utilisées pour faire de l'interpolation spatiale par « séries de Fourier » (Helson et Lowdenslager, 1958, 1961). De plus, des variables régionalisées auxiliaires peuvent être ajoutées dans la tendance $\mu(s)$. Ainsi, de l'interpolation spatiale multivariable peut être effectuée en exploitant la corrélation entre les variables auxiliaires et la variable à interpoler.

Notons que l'interpolation par régression classique est considérée déterministe par certains auteurs (Arnaud et Emery, 2000, p.73 ; Burrough et McDonnell, 1998, p.158). D'ailleurs, l'emploi de la méthode des moindres carrés rend superflue la modélisation stochastique de la variable aléatoire. Par contre, cette modélisation, accompagnée des hypothèses de normalité, d'indépendance et d'homoscédasticité des erreurs du modèle, permet de calculer des tests de significativité de la surface et des erreurs de prévision.

Cependant, ces statistiques ne sont presque jamais fiables en interpolation spatiale car le postulat d'indépendance est rarement vérifié sur des données spatiales. Malgré tout, l'interpolation par régression classique est ici considérée stochastique, comme le font Klinkenberg et Waters (1990), car cette technique fut à l'origine développée dans un cadre statistique.

De plus, l'interpolation par régression classique possède les propriétés d'être approximative et globale. La surface d'interpolation qu'elle génère ne passe donc pas nécessairement par les points d'observation et toutes les observations, mêmes celles éloignées, vont influencer avec le même poids la prévision en n'importe quel point du champ. En raison de ces caractéristiques, la méthode produit des surfaces plutôt lisses.

2.2.5 Régression locale

Une extension de la régression classique permet de diminuer l'impact sur l'interpolation en un point s_0 des observations éloignées de ce point. Il s'agit de la « régression locale » (Cleveland et Devlin, 1988), aussi nommée « régression kernel » (Wand et Jones, 1995, p.114) ou plus spécifiquement « régression pondérée géographiquement » (Fotheringham *et al.*, 2002). Cette méthode postule le même modèle qu'en régression classique (équation 2.1). Cependant, les coefficients de la tendance $\mu(\cdot)$ sont maintenant estimés par une méthode locale d'estimation telle les moindres carrés pondérés ou la méthode du maximum de vraisemblance pondérée. Par exemple, par les moindres carrés pondérés, la surface estimée en un point s_0 minimise :

$$\sum_{i=1}^{n} \lambda_i(s_0) \left[\hat{z}(s_i)|_{s_0} - z(s_i)\right]^2$$

où $\hat{z}(s_i)|_{s_0}$ est la valeur prédite de $z(\cdot)$ au point d'observation s_i à partir de la surface ajustée en s_0 et $z(s_i)$ est la valeur régionalisée observée en s_i. Le poids de l'observation au point s_i prend typiquement la forme :

$$\lambda_i(s_0) = K\left(\frac{|s_0 - s_i|}{h(s_0)}\right)$$

où $K(\cdot)$ est une fonction de poids, $|s_0 - s_i|$ la distance euclidienne entre le point de prévision s_0 et le point d'observation s_i et $h(\cdot)$ est la taille du voisinage. Plusieurs fonctions de poids sont proposées dans la littérature (Loader, 1999, p.48 ; Wand et Jones, 1995, p.31). Par exemple, une de ces fonctions est la fonction Epanechnikov définie par $K(u) = 1 - u^2$ pour $0 \leq u < 1$, et 0 sinon. En outre, le paramètre $h(s_0)$ est habituellement déterminé en choisissant d'abord la fraction α des points d'observation à inclure dans l'ajustement du modèle. Ensuite, on attribue à $h(s_0)$ la valeur de la distance

entre s_0 et le $[n\alpha]^{\text{ième}}$ plus proche point d'observation de s_0 (Loader, 1999, p.20), où $[n\alpha]$ représente la valeur entière du produit $n\alpha$. Ainsi, plus une valeur régionalisée à été mesurée loin du point de prévision s_0, moins elle a d'impact sur l'interpolation en s_0. Au-delà d'une certaine distance elle n'a plus d'impact du tout.

En pratique, l'utilisateur de la régression locale a quatre choix à faire avant d'ajuster une surface. Il doit d'abord déterminer la forme de la tendance $\mu(\cdot)$. Si une forme polynômiale est choisie, le degré du polynôme doit être spécifié. Ensuite, la fonction de poids et la taille du voisinage doivent être sélectionnées. Le paramètre le plus difficile à déterminer est la taille du voisinage. Il a une grande influence sur la surface obtenue : plus sa valeur est grande, plus la surface d'interpolation est lisse. Quelques procédures de sélection automatique de la taille de voisinage ont été proposées dans la littérature (Cleveland et Loader, 1996). Finalement, l'utilisateur choisit la procédure d'estimation des paramètres qu'il emploiera. Les moindres carrés pondérés est certainement la procédure la plus simple. Si une procédure dépendante de la distribution des erreurs $\epsilon(s_i)$ est choisie, telle qu'une méthode de maximum de vraisemblance, les hypothèses émises concernant la distribution des erreurs auront de l'influence sur les prévisions.

Afin de retomber sur une régression classique en partant d'une régression locale, il suffit de prendre une fonction de poids uniforme et une fraction de voisinage égale 1. De plus, une regression locale avec une tendance polynomiale de degré 0 ajustée par la méthode des moindres carrés pondérés revient à une méthode barycentrique (Loader, 2004). Dans ce cas, la seule distinction entre les deux méthodes est que la modélisation stochastique de la régression locale permet le calcul d'erreurs de prévision. Cependant, pour la même raison qu'en régression classique (voir section 2.2.4), il faut douter de la fiabilité de ces erreurs.

2.2.6 Krigeage

Les deux méthodes stochastiques présentées dans les sections précédentes ne prennent pas en considération la structure de dépendance spatiale des données. Le krigeage est la première méthode d'interpolation spatiale à en avoir tenu compte. Les travaux de l'ingénieur minier sud-africain Krige (1951) sont précurseurs de la méthode. Cependant, le terme *krigeage* et le formalisme de cette méthode sont dus au français Matheron (1962, 1963a,b), qui en a aussi assuré le développement au Centre de Géostatistique de l'École des Mines de Paris. En fait, les fondements de la méthode ont été développés parallèlement par d'autres chercheurs, notamment le météorologue Gandin (1963) de l'ex-URSS, mais c'est aujourd'hui sous la terminologie proposée par Matheron qu'elle

est la plus connue (Cressie, 1990).

L'idée de base du krigeage est de prévoir la valeur de la variable régionalisée étudiée en un site non échantillonné s_0 par une combinaison linéaire de données ponctuelles adjacentes :

$$\hat{z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i z(s_i). \tag{2.3}$$

Les poids λ_i associés à chacune des valeurs régionalisées observées sont choisis de façon à obtenir une prévision non biaisée et de variance minimale. Ces poids dépendent de la localisation des observations et de leur structure de dépendance spatiale. En fait, le krigeage est le nom donné à la meilleure prévision linéaire sans biais, en anglais « best linear unbiased predictor » ou « BLUP », dans un cadre spatial.

Le modèle de base du krigeage a la même forme que le modèle de régression classique ou locale, mais les erreurs sont maintenant supposées dépendantes spatialement. Il s'énonce comme suit :

$$Z(s) = \mu(s) + \delta(s), \qquad s \in D \tag{2.4}$$

où $\mu(\cdot)$ est la structure déterministe pour l'espérance de $Z(\cdot)$ et $\delta(\cdot)$ une fonction aléatoire stationnaire, d'espérance nulle et de structure de dépendance connue. Pour formuler complètement le modèle, il faut spécifier la forme de la tendance $\mu(\cdot)$. C'est en fait cette tendance qui précise le type de krigeage effectué. Les trois types classiques de krigeage sont :

- Le krigeage simple : $\mu(s) = m$ est une constante connue.
- Le krigeage ordinaire : $\mu(s) = \mu$ est une constante inconnue.
- Le krigeage universel : $\mu(s) = \sum_{j=0}^{p} f_j(s)\beta_j$ est une combinaison linéaire de fonctions de la position s.

De plus, la structure de dépendance de la fonction aléatoire $\delta(\cdot)$ doit être précisée. Si elle n'est pas connue préalablement, ce qui est presque toujours le cas en pratique, elle est déterminée à partir des données lors de l' « analyse variographique » que nous verrons en détails au chapitre 3. Cette étape permet de décrire la variabilité spatiale de phénomènes régionalisés. Ensuite, le modèle étant complètement énoncé, le krigeage peut être effectuée en un point s_0 quelconque du champ D. En fait, il s'agit simplement de déterminer la valeur des poids λ_i de la combinaison linéaire 2.3 qui respectent la contrainte de non-biais $E[\hat{Z}(s_0) - Z(s_0)] = 0$ tout en minimisant la variance de l'erreur de prévision $Var[\hat{Z}(s_0) - Z(s_0)]$ (voir chapitre 4).

Le krigeage est une méthode d'interpolation très souple. Il peut être global ou local dépendamment du voisinage choisi. De plus, selon le développement classique du krigeage qui sera suivi dans ce mémoire, il s'agit d'une méthode d'interpolation exacte.

Il restitue donc les valeurs régionalisées mesurées aux sites d'observation. Cependant, il est aussi possible d'effectuer un krigeage dit avec erreurs de mesure (Cressie, 1993, p.128) qui est approximatif. Finalement, comme en régression classique ou locale, des variables régionalisées auxiliaires peuvent être intégrées au krigeage en les ajoutant à la tendance générale $\mu(\cdot)$ du modèle. Le krigeage incorporant une telle tendance est nommé « krigeage avec dérive externe » (Goovaerts, 1997). Le krigeage possède aussi d'autres extensions multivariables, notamment le « cokrigeage » (Wackernagel, 2003). Par cette méthode, $\hat{z}(s_0)$ prend la forme d'une combinaison linéaire pondérée des observations de la variable régionalisée à interpoler et des variables régionalisées auxiliaires notées $\{w_j(s), s \in D\}$ avec $j = 1, ..., q$:

$$\hat{z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_{i,0} z(s_i) + \sum_{j=i}^{q} \sum_{i \in V(s_0)} \lambda_{i,j} w_j(s_{i,j}).$$

Le cokrigeage prend en considération les structures de dépendance spatiale des variables auxiliaires afin de trouver les poids $\lambda_{i,j}$ minimisant la variance de l'erreur de prévision sous la contrainte de non-biais.

2.2.7 Autres méthodes

Les sections précédentes ont introduit les méthodes d'interpolation les plus connues et utilisées. Mentionnons tout de même ici quelques-unes des autres méthodes existantes. Certaines faiblesses du krigeage ont mené au développement de méthodes bayésiennes (Gaudard *et al.*, 1999; Banerjee *et al.*, 2004). Ces faiblesses sont principalement liées à la présence de biais dans l'estimation du variogramme en krigeage universel (Cressie, 1993, p.165) et au fait que l'incertitude d'une prévision calculée en krigeage n'incorpore pas l'incertitude associée à l'estimation du variogramme. Le modèle bayésien typique en interpolation spatiale effectue un krigeage avec une fonction de covariance de paramètres inconnus. Ces modèles peuvent aussi permettre d'intégrer à l'interpolation de l'information à priori concernant la tendance $\mu(\cdot)$. Un des grands avantages de l'approche bayésienne est la perspective conditionnelle qui permet un raisonnement probabiliste, après observation des données, sans recourir à l'idée d'une répétition de l'expérience, condition souvent difficile à concevoir en géostatistique. Certains utilisent aussi les réseaux de neurones artificiels pour interpoler des données (Bryan et Adams, 2002; Rigol *et al.*, 2001). Les réseaux de neurones agissent alors comme une méthode de régression non paramétrique. Dans ce mémoire, aucune de ces méthodes ne sera approfondie.

2.3 Conclusion du chapitre

De la revue présentée dans ce chapitre, le krigeage semble être la méthode d'interpolation la plus intéressante, et ce, pour plusieurs raisons. Premièrement, comme avec les méthodes barycentriques et la régression locale, l'utilisateur du krigeage a le choix d'interpoler localement ou globalement. De plus, puisqu'il s'agit d'une méthode stochastique, le krigeage permet d'estimer des erreurs de prévisions. Le krigeage possède également plus d'extensions multivariables que les autres catégories de méthodes. Cependant, ce qui distingue vraiment le krigeage des autres méthodes introduites précédemment est qu'il est le seul à tenir compte de la structure de dépendance spatiale des données. Ainsi, on peut s'attendre à ce que le krigeage génère les prévisions spatiales les plus justes. De plus, l'estimation des erreurs qu'il produit est plus fiable que celles produites par les autres méthodes stochastiques, car les postulats de base du krigeage modélisent mieux la réalité pour des données à référence spatiale. Le krigeage ressort donc gagnant de la comparaison théorique avec les autres méthodes d'interpolation. Ses fondements théoriques sont approfondis dans les deux prochains chapitres.

Chapitre 3

Analyse variographique

Le modèle 2.4 sur lequel se base le krigeage suppose la connaissance de la structure de dépendance spatiale de la fonction aléatoire $\delta(\cdot)$. Cependant, en pratique, celle-ci est rarement connue. L'analyse variographique est une étape préalable au krigeage qui permet de l'estimer. Cette analyse est en fait l'étude du comportement spatial de la variable régionalisée examinée. Dans ce chapitre, le concept de stationnarité est d'abord défini, puis le modèle de base du krigeage est exposé en termes de variations spatiales à différentes échelles. De la définition de stationnarité est tirée une fonction représentant la dépendance spatiale : le semi-variogramme. La fin du chapitre se consacre à l'estimation et à la modélisation de cette fonction.

3.1 Hypothèse de stationnarité

Le processus naturel étudié étant unique, une seule réalisation de la fonction aléatoire $Z(\cdot)$ est observable. Afin de rendre possible l'inférence statistique malgré l'unicité de la réalisation disponible, une hypothèse de stationnarité est émise concernant la fonction aléatoire $\delta(\cdot)$. Au sens strict, la stationnarité signifie que la loi de probabilité de la fonction aléatoire est invariante par translation, c'est-à-dire qu'elle ne dépend pas de l'origine du champ. Par contre, en krigeage, la stationnarité postulée est faible : elle ne concernent que les moments d'ordre 1 et 2 de la fonction aléatoire ou de ses accroissements plutôt que sa distribution entière. La théorie du krigeage peut être développée en postulant la *stationnarité de second ordre* ou la stationnarité de second ordre intrinsèque (plus simplement nommée *stationnarité intrinsèque*) de $\delta(\cdot)$ (Cressie, 1993, p.40 et 53 ; Arnaud et Emery, 2000, p.106). Ces deux types de stationnarité se définissent ainsi :

Stationnarité de second ordre :

1. $E(\delta(s)) = m = 0 \qquad \forall \quad s \in D \quad :$
   L'espérance de $\delta(\cdot)$ existe et est la même en tout site. Dans le modèle du krigeage, cette espérance est même supposée nulle.
2. $Cov(\delta(s), \delta(s+h)) = C(h) \qquad \forall \quad s, s+h \in D \quad :$
   La covariance de $\delta(\cdot)$ entre toute paire de sites s et $s+h$ existe et dépend uniquement de h, le vecteur de translation entre ces points. Cette fonction de covariance est appelée « covariogramme ».

Stationnarité intrinsèque :

1. $E(\delta(s+h) - \delta(s)) = 0 \qquad \forall \quad s \in D \quad :$
   L'espérance de tout accroissement $\delta(s+h) - \delta(s)$ est nulle.
2. $Var(\delta(s+h) - \delta(s)) = 2\gamma(h) \qquad \forall \quad s, s+h \in D \quad :$
   La variance de tout accroissement $\delta(s+h) - \delta(s)$ existe et dépend uniquement de h. Cette fonction de variance est appelée « variogramme ».

Tout processus stationnaire de deuxième ordre est stationnaire intrinsèque. En effet, dans le cadre stationnaire de second ordre, l'équation $2\gamma(h) = 2(C(0) - C(h))$ relie le covariogramme et le variogramme. L'existence du covariogramme implique donc l'existence du variogramme. Par contre, l'implication inverse n'est vraie que si $\gamma(\cdot)$ est bornée. Il se peut donc que le covariogramme d'une fonction aléatoire intrinsèque ne soit pas défini. Par exemple, le mouvement brownien fractionnaire (Hida, 1980, p.298) est une fonction aléatoire $\{B(s), s \in \Re^2\}$ gausienne d'espérance nulle et de covariance :

$$Cov(B(s), B(s+h)) = (|s|^\nu + |s+h|^\nu - |h|^\nu)\sigma^2, \quad 0 < \nu < 2.$$

Cette covariance ne constitue pas un covariogramme car elle ne dépend pas uniquement de h. Elle dépend aussi de l'emplacement du point s par rapport à l'origine du plan. Par contre, le variogramme de $B(\cdot)$ existe puisque $Var(B(s+h) - B(s)) = 2|h|^\nu\sigma^2$ est une fonction de h uniquement. Ainsi, certaines fonctions aléatoires sont stationnaires intrinsèques mais non stationnaires de deuxième ordre. L'hypothèse intrinsèque est donc plus générale.

Cette généralité est utile en krigeage car certains phénomènes régionalisés présentent une dispersion infinie. C'est le cas des gisements d'or selon Krige (1951). Postuler l'existence de la variance de la fonction aléatoire $\delta(\cdot)$ est donc parfois inadéquat. Pour cette raison, l'hypothèse de stationnarité intrinsèque sera privilégiée dans ce mémoire, tel que le font la majorité des auteurs sur le sujet. Ainsi, une dépendance spatiale sera représentée par un variogramme plutôt qu'un covariogramme. D'ailleurs, le variogramme est plus facile à estimer car il ne requiert pas d'estimation de l'espérance de $\delta(\cdot)$. Pour ces raisons, le variogramme est l'outil privilégié en analyse variographique. En fait, la plupart des auteurs travaillent avec $\gamma(\cdot)$, la demie du variogramme, appelée semi-variogramme. C'est aussi ce qui sera fait dans le reste de ce mémoire.

3.2 Décomposition de la variation spatiale

Afin d'examiner plus à fond le semi-variogramme, il faut d'abord revenir sur le modèle de base du krigeage énoncé au chapitre précédent par l'équation 2.4. Ce modèle peut en fait se détailler de la façon suivante (Cressie, 1993, p.112) :

$$Z(s) = \mu(s) + \omega(s) + \eta(s) + \epsilon(s), \qquad s \in D \tag{3.1}$$

où

- $\mu(\cdot) = E(Z(\cdot))$: variation à grande échelle, structure déterministe pour l'espérance de $Z(\cdot)$.
- $\omega(\cdot)$: variation lisse à petite échelle (échelle plus grande que la distance minimale entre deux sites d'échantillonnage), structure stochastique de fluctuations autour de $\mu(\cdot)$ dépendantes spatialement.
- $\eta(\cdot)$: variation micro-échelle (échelle plus petite que la distance minimale entre deux sites d'échantillonnage), structure stochastique présentant une dépendance spatiale.
- $\epsilon(\cdot)$: erreur de mesure, structure stochastique sans dépendance spatiale (bruit blanc).

Ainsi, la variable régionalisée $z(\cdot)$ est supposée être composée de variations à différentes échelles emboîtées les une dans les autres (Oliver, 2001). La fonction aléatoire $\delta(\cdot)$ du modèle 2.4 est formée par le regroupement des termes $\omega(\cdot)$, $\eta(\cdot)$ et $\epsilon(\cdot)$ du modèle précédent.

La figure 3.1 permet d'illustrer l'équation 3.1. Un exemple de variable régionalisée $z(\cdot)$ a été créé en sommant des fonctions $\mu(\cdot)$, $\omega(\cdot)$, $\eta(\cdot)$ et $\epsilon(\cdot)$ simulées avec le logiciel S-Plus. En observant $z(\cdot)$ sur tout son domaine x et $y \in \{0, ..., 100\}$ dans le graphique supérieur gauche, on note une tendance générale d'augmentation de la valeur de z pour une augmentation de la coordonnée en x peu importe la coordonnée en y. Cette tendance générale, représentée par le terme $\mu(\cdot)$, pourrait être modélisée par un plan. En agrandissant la région délimitée par le cube bleu du graphique à grande échelle, on obtient le graphique à moyenne échelle au centre. Sur ce graphique, des fluctuations lisses autour de la tendance générale d'augmentation par rapport à l'axe des x ressortent clairement. Ces fluctuations sont modélisées par $\omega(\cdot)$ dans l'équation 3.1. Finalement, regardons $z(\cdot)$ d'encore plus près sur le graphique de droite, qui est un agrandissement du cube rouge du graphique à moyenne échelle. La surface représentée dans ce graphique n'est pas tout à fait lisse. Elle présente de petites irrégularités qui n'étaient pratiquement pas visibles sur les deux graphiques à plus grandes échelles. Les termes $\eta(\cdot)$ et $\epsilon(\cdot)$ sont attribuées à ces variations difficilement distinguables entre elles.

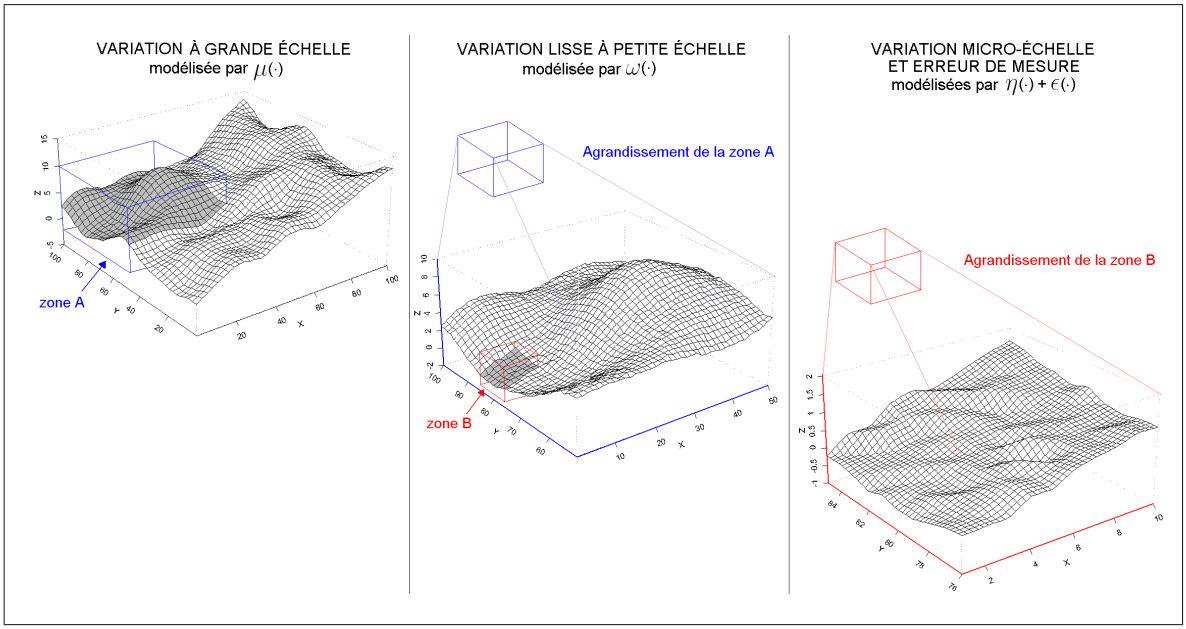

FIG. 3.1 – *Exemple de variable régionalisée illustrant un emboîtement de variations à différentes échelles*

La décortication de la fonction aléatoire résiduelle $\delta(\cdot)$ présentée dans cette section sera maintenant utile pour expliquer les propriétés du semi-variogramme.

3.3 Propriétés du semi-variogramme

Tel que défini à la section 3.1, le semi-variogramme est une fonction paire, i.e. $\gamma(h) = \gamma(-h)$, nulle en $h = 0$ et positive partout ailleurs. Toutes ces caractéristiques découlent en fait d'une propriété générale des semi-variogrammes : ce sont des fonctions de type négatif conditionnel (Christakos, 1984), c'est-à-dire que :

$$\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j \gamma(s_i - s_j) \leq 0$$

pour n'importe quel ensemble fini de points $\{s_i : i = 1, ..., l\}$ et n'importe quels nombres réels $\{a_i : i = 1, ..., l\}$ tel que $\sum_{i=1}^{l} a_i = 0$. Ainsi, une fonction continue peut représenter un semi-variogramme si et seulement si elle respecte cette propriété qui assure la positivité de la variance de toute combinaison linéaire de variables aléatoires issues de $\delta(\cdot)$. Cette propriété est en fait l'extension aux semi-variogrammes du caractère semi-défini positif bien connu des fonctions de covariance.

Avant de décrire certains attributs des semi-variogrammes, voici deux exemples de fonctions pouvant faire état de semi-variogramme :

$$\gamma(h) = \begin{cases} 0.2 + 0.8\left(\frac{3}{2}\frac{|h|}{6} - \frac{1}{2}\frac{|h|^3}{6^3}\right), & 0 \leq |h| \leq 6 \\ 0.2 + 0.8, & |h| > 6 \end{cases} \qquad \gamma(h) = 0.2 + 0.8\left(1 - exp\left(-\frac{|h|}{2}\right)\right), \quad |h| \geq 0$$

où $|h|$ est la norme du vecteur de h. Ces fonctions sont tracées dans les graphiques de la figure 3.2. Il s'agit de semi-variogrammes isotropes, d'effet de pépite valant 0.2, de palier exact ou asymptotique valant 1 et de portée exacte ou pratique valant 6. Les paragraphes suivants définissent ces caractéristiques des semi-variogrammes.

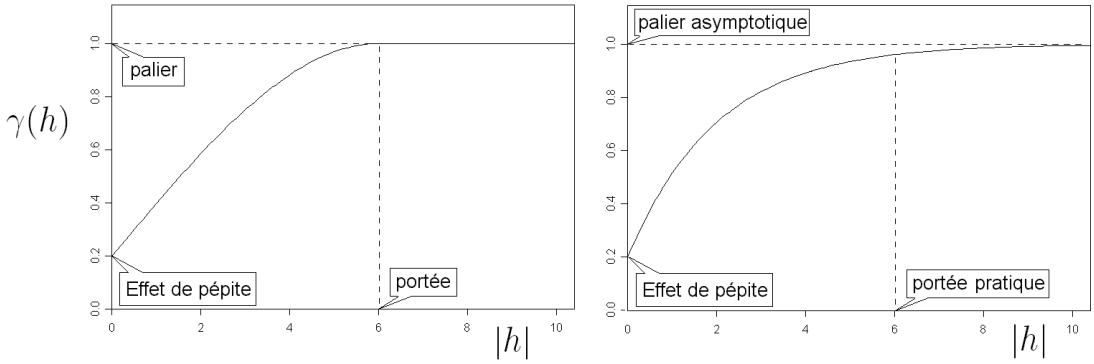

FIG. 3.2 – *Exemple de semi-variogrammes*

Isotropie

Le semi-variogramme ne dépend que de h, le vecteur de translation entre les points s et $s+h$. Ce vecteur contient de l'information sur la distance entre ces deux points, par l'intermédiaire de sa norme, ainsi que sur l'orientation de h. Si le semi-variogramme ne dépend en fait que de la norme de h, il est dit isotrope. S'il dépend aussi de la direction du vecteur de translation, il est alors anisotrope. Dans ce mémoire, les semi-variogrammes seront toujours supposés isotropes. Pour un traitement de l'anisotropie, le lecteur est référé à Goovaerts (1997). Rappelons que la norme euclidienne d'un vecteur $h = (x_h, y_h)$ est $r = |h| = \sqrt{x_h^2 + y_h^2}$. Le semi-variogramme sera dorénavant noté :

$$\gamma(r) = \frac{1}{2}Var(\delta(s) - \delta(s+h)) \qquad \forall\ h \text{ de norme } r \text{ et } \forall\ s, s+h \in D \qquad (3.2)$$

Effet de pépite

Au voisinage de l'origine, un semi-variogramme peut être continu ou discontinu. Si $\lim_{r\to 0^+} \gamma(r) = c_o > 0$, alors c_o est appelé effet de pépite. Un saut abrupt à l'origine dénote une faible ressemblance entre les valeurs régionalisées très voisines. Un effet de pépite s'explique par des variations non-détectées à une micro-échelle. En se ramenant au modèle 3.1, c'est donc dire que l'effet de pépite est associé à la fonction aléatoire $\eta(\cdot)$. L'adaptation du krigeage pour le cas où des erreurs de mesure sont présentes associera aussi l'effet de pépite à la fonction aléatoire de bruit blanc $\epsilon(\cdot)$ (Cressie, 1993, p.128).

Portée et palier

Les autres caractéristiques du semi-variogramme sont plutôt dues aux variations observables $\omega(\cdot)$ de l'équation 3.1. C'est le cas du comportement du semi-variogramme lorsque r augmente. Le semi-variogramme peut ou non atteindre un plateau. L'atteinte d'un plateau indique qu'à partir d'une certaine distance il n'y a plus de dépendance spatiale entre les données. Cette distance est nommée portée et le terme palier dénote la variance à laquelle le plateau se présente. Il s'agit en fait de la variance commune aux variables aléatoires $\delta(s)$ pour tout $s \in D$. Un palier peut n'être atteint qu'asymptotiquement. Dans ce cas, la portée réelle est infinie, mais une portée pratique est définie par la distance à laquelle le semi-variogramme atteint 95% de la valeur de son palier.

Si un semi-variogramme est non borné, il ne possède ni portée, ni palier. La variance de la fonction aléatoire n'est pas définie pour un tel semi-variogramme. Cette fonction aléatoire n'est donc pas stationnaire de deuxième ordre, mais seulement stationnaire intrinsèque. Rappelons que c'est le cas du mouvement Brownien mentionné en exemple à la section 3.1.

3.4 Estimation du semi-variogramme

On cherche à estimer le semi-variogramme à partir des données disponibles, c'est-à-dire les $z(s_i)$ pour $i = 1, ..., n$. Pour ce faire, notons d'abord que :

$$\gamma(r) = \frac{1}{2} Var(\delta(s) - \delta(s+h)) = \frac{1}{2} E[\{\delta(s) - \delta(s+h)\}^2]$$

car $E[\delta(s)] = 0 \quad \forall\ s \in D$. Cependant, le semi-variogramme peut aussi se développer de la façon suivante :

$$\begin{aligned}
\gamma(r) &= \frac{1}{2}Var(\delta(s) - \delta(s+h)) \\
&= \frac{1}{2}Var[(Z(s) - \mu(s)) - (Z(s+h) - \mu(s+h))] \\
&= \frac{1}{2}Var(Z(s) - Z(s+h)) \qquad \text{car } \mu(\cdot) \text{ n'est pas une fonction aléatoire} \\
&= \frac{1}{2}E[\{Z(s) - Z(s+h)\}^2] - \frac{1}{2}E[Z(s) - Z(s+h)]^2 \\
&= \frac{1}{2}E[\{Z(s) - Z(s+h)\}^2] - \frac{1}{2}\{\mu(s) - \mu(s+h)\}^2
\end{aligned}$$

Ainsi, si $\mu(\cdot)$ est une fonction constante, le deuxième terme s'annule et le semi-variogramme est estimable directement à partir des $z(s_i)$. C'est le cas en krigeage simple et ordinaire.

Par contre, lorsque $\mu(\cdot)$ n'est pas une fonction constante tel qu'en krigeage universel ou en krigeage avec dérive externe, l'estimation doit se baser sur la fonction $\delta(\cdot)$. Des observations de $\delta(\cdot)$ sont disponibles si la fonction $\mu(\cdot)$ est connue. Le semi-variogramme peut alors être estimé à partir des valeurs $z(s_i) - \mu(s_i)$. Cependant, en pratique, $\mu(\cdot)$ n'est pas connue. Cette fonction est donc estimée, puis le semi-variogramme expérimental est calculé à partir des résidus $e(s_i) = z(s_i) - \hat{\mu}(s_i)$. Toutefois, Matheron (1970, p.155) a lui même soulevé que cette estimation est biaisée. Une discussion plus approfondie de cette problématique se trouve à la section 4.4.

L'estimateur du semi-variogramme le plus commun est celui des moments (Cressie, 1993, p.69) :

$$\hat{\gamma}(r) = \frac{1}{2|N(r)|}\sum_{N(r)}(z(s_i) - z(s_j))^2 \quad \text{ou} \quad \frac{1}{2|N(r)|}\sum_{N(r)}(e(s_i) - e(s_j))^2 \tag{3.3}$$

où $N(r) = \{(i,j) \text{ tel que } |s_i - s_j| = r\}$ et $|N(r)|$ est le nombre de paires distinctes de l'ensemble $N(r)$.

Si les données disponibles sont réparties sur une grille régulière couvrant le champ D, $\gamma(\cdot)$ est estimable pour un petit nombre de distances, chacune associée à plusieurs couples de données. Toutefois, les données sont plus souvent réparties irrégulièrement sur D. Dans ce cas, plus de distances sont représentées, mais avec un très petit nombre de paires de données. Afin de rendre $\hat{\gamma}(\cdot)$ plus robuste, des tolérances sont introduites sur les distances. Le semi-variogramme expérimental est donc calculé seulement pour certaines distances r_k, avec $k = 1, ..., K$, en considérant les couples de données éloignées d'une

distance à peu près égale à r_k. L'ensemble $N(r_k)$ est ainsi redéfini par $\{(i,j)$ tel que $|s_i - s_j| = r_k \pm b_k\}$ où b_k est une tolérance déterminée par l'utilisateur.

En outre, Arnaud et Emery (2000, p.126) affirment que le semi-variogramme expérimental n'est pas fiable pour de trop grandes distances, car la dispersion de $\hat{\gamma}(\cdot)$ autour de $\gamma(\cdot)$ augmente lorsque r devient grand. Ils conseillent donc de calculer le semi-variogramme expérimental seulement pour les distances inférieures à la moitié de la distance maximale entre deux points d'observation.

D'autres estimateurs de semi-variogramme sont proposés dans la littérature. Cressie (1993, p.74) présente notamment des estimateurs plus robustes que l'estimateur classique des moments qui contient une puissance au carré très sensible aux valeurs extrêmes. En outre, si plusieurs observations de la variable régionalisées sont disponibles dans le temps aux mêmes sites d'observation, un estimateur possible est (Abtew *et al.*, 1985) :

$$\hat{\gamma}(s_i - s_j) = \frac{\sum_{k=1}^{m}(z_k(s_i) - z_k(s_j))^2}{2m} \quad \text{ou} \quad \frac{\sum_{k=1}^{m}(e_k(s_i) - e_k(s_j))^2}{2m} \tag{3.4}$$

pour tout couple d'observations (i,j) où k est un indice référant à la période de temps, $z_k(s_i)$ est donc l'observation de la variable régionalisée z au point s_i au pas de temps k et m est le nombre total de périodes entrant dans le calcul. Cette idée de sommer sur les pas de temps pour estimer un semi-variogramme vient de l'« interpolation optimale » de Gandin (1963). Cette méthode est en fait l'équivalent d'un krigeage dont les équations sont développées à partir d'un corrélogramme, soit un covariogramme $C(h)$ divisé par la variance $C(0)$, plutôt qu'un semi-variogramme. En interpolation optimale, ce corrélogramme est toujours estimé à partir d'une série temporelle de données à la manière du semi-variogramme expérimental 3.4. Une limite de l'estimateur 3.4 est qu'il n'est pertient que si les caractéristiques de la structure de dépendance spatiale du phénomène naturel à l'étude varient peu ou pas dans le temps.

3.5 Modélisation du semi-variogramme

Le semi-variogramme expérimental présenté dans la section précédente estime le semi-variogramme théorique pour un nombre fini de distances seulement. De plus, il ne forme pas nécessairement un semi-variogramme valide, c'est-à-dire qu'il ne s'agit peut-être pas d'une fonction conditionnellement négative définie. Le semi-variogramme expérimental est donc modélisé par une fonction de type négatif conditionnel, définie pour toutes les distances $r \in \Re^+$. Cette modélisation rendra possible le krigeage.

Tester le caractère négatif conditionnel d'une fonction est une tâche assez difficile. Christakos (1984) explique une procédure pour y arriver. Cependant, en pratique, il est plus simple d'utiliser un modèle variographique classique, qui a été démontré valide. Voici donc certains modèles variographiques couramment utilisés (Arnaud et Emery, 2000, p.133), qui seront d'ailleurs employés dans les chapitres 5 et 6 :

Modèles avec palier : Le semi-variogramme est borné, la fonction aléatoire associée est donc stationnaire de second ordre.

Portée exacte :

- Modèle pépitique de palier c_0 :

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 & \text{pour } r > 0 \end{cases}$$

  Ce modèle représente l'absence de dépendance spatiale. Un krigeage avec ce modèle revient à une régression classique (Marcotte, 1988).

- Modèle linéaire avec palier d'effet de pépite c_0, de palier $c_0 + c$ et de portée a :

$$\gamma(r) = \begin{cases} c_0 + \frac{c}{a} r & \text{pour } 0 \leq r \leq a \\ c_0 + c & \text{pour } r > a \end{cases}$$

- Modèle sphérique d'effet de pépite c_0, de palier $c_0 + c$ et de portée a :

$$\gamma(r) = \begin{cases} c_0 + c\left(\frac{3}{2}\frac{r}{a} - \frac{1}{2}\frac{r^3}{a^3}\right) & \text{pour } 0 \leq r \leq a \\ c_0 + c & \text{pour } r > a \end{cases}$$

Portée asymptotique :

- Modèle exponentiel d'effet de pépite c_0, de palier $c_0 + c$ et de portée pratique égale à $3a$:

$$\gamma(r) = c_0 + c\left(1 - exp\left(-\frac{r}{a}\right)\right) \qquad \text{pour } r \geq 0$$

- Modèle gaussien d'effet de pépite c_0, de palier $c_0 + c$ et de portée pratique égale à $a\sqrt{3}$:

$$\gamma(r) = c_0 + c\left(1 - exp\left(-\frac{r^2}{a^2}\right)\right) \qquad \text{pour } r \geq 0$$

Modèles sans palier : Le semi-variogramme n'est pas borné, la fonction aléatoire associée est donc seulement stationnaire intrinsèque.

– Modèle linéaire sans palier d'effet de pépite c_0 et de pente m :

$$\gamma(r) = c_0 + mr \qquad \text{pour } r \geq 0$$

Ce modèle est associé au mouvement brownien fractionnaire de paramètre $\lambda = 1$ présenté en exemple à la section 3.1.

– Modèle puissance d'effet de pépite c_0, d'exposant ν et de facteur d'échelle m :

$$\gamma(r) = c_0 + mr^{\nu} \qquad \text{pour } r \geq 0, \quad 0 < \nu < 2$$

Il s'agit d'une généralisation du modèle linéaire. Le mouvement brownien fractionnaire de paramètre ν quelconque possède un semi-variogramme de cette forme.

La figure 3.3 comporte un graphique pour chacun de ces modèles variographiques. Notons que toute somme de modèles élémentaires est aussi un modèle admissible. Cette approche d'addition de modèles forme ce qui est appelé un modèle linéaire de régionalisation (Arnaud et Emery, 2000, p.139).

En plus de sélectionner le modèle, celui-ci doit être ajusté au semi-variogramme expérimental ; c'est donc dire que les paramètres du modèle doivent être estimés. Cet ajustement peut se faire à l'oeil, mais il s'effectue habituellement à l'aide d'une méthode d'estimation. La méthode la plus couramment utilisée est celle des moindres carrés pondérés. Le vecteur des paramètres du modèle choisi, noté θ, est donc souvent estimé en minimisant :

$$\sum_{k=1}^{K} w_k[\hat{\gamma}(r_k) - \gamma(r_k, \theta)]^2$$

où $\hat{\gamma}(\cdot)$ est le semi-variogramme empirique, $\gamma(\cdot, \theta)$ est le modèle variographique de paramètres θ, w_k est le poids associé à la donnée $\hat{\gamma}(r_k)$, et r_1 à r_K sont les distances pour lesquelles une estimation du semi-variogramme a été calculée. Plusieurs poids w_k sont proposés dans la littérature. Ils peuvent évidemment tous être fixés à un, ce qui revient aux moindres carrés ordinaires. Il est aussi intuitif d'envisager les poids $w_k = 1/|N(r_k)|$, soit l'inverse du nombre de paires de données ayant entré dans le calcul de $\hat{\gamma}(r_k)$. Fuentes (2000) discute d'autres poids pertinents et présente différentes méthodes d'estimation, telles la méthode du maximum de vraisemblance ainsi que des approches bayésiennes.

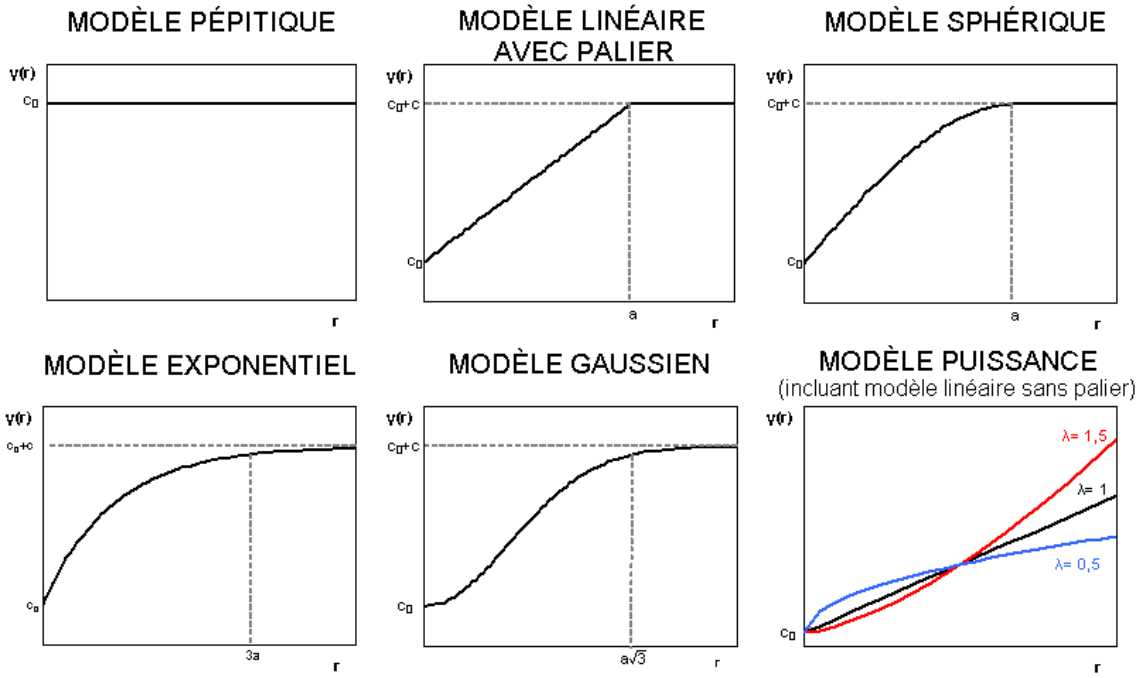

FIG. 3.3 – *Modèles de semi-variogrammes les plus communs*

3.6 Conclusion du chapitre

Les rudiments de l'analyse variographique ont été présentés dans ce chapitre. Nous avons défini un outil, le semi-variogramme, pour représenter une dépendance spatiale et nous avons montré comment estimer et modéliser cette fonction. Le chapitre 5 reviendra sur l'analyse variographique pour expliquer comment elle s'effectue en pratique. Notamment, le chapitre 5 traitera de comment choisir l'estimateur du semi-variogramme et le modèle variographique. Mais tout d'abord, passons à l'élaboration des équations du krigeage.

Chapitre 4

Théorie du krigeage

Ce chapitre présente le détail mathématique de l'obtention de la forme d'une prévision en krigeage. Tout d'abord, une démarche générale pour obtenir cette prévision est explicitée. Ensuite, cette démarche est suivie pour les trois grands types de krigeage présentés à la section 2.2.6, soit le krigeage simple, le krigeage ordinaire et le krigeage universel qui se généralise par le krigeage avec modèle de tendance. Des discussions sur certains aspects théoriques du krigeage d'un de ces types et sur d'autres types de krigeage closent le chapitre. Les références principales de ce chapitre sont : *Statistics for Spatial Data* de Cressie (1993) et *La théorie des variables régionalisées, et ses applications* de Matheron (1970).

4.1 Démarche générale de résolution des équations du krigeage

Rappelons d'abord que l'objectif du krigeage est de prévoir la valeur de la variable régionalisée à interpoler $z(\cdot)$ en un site non échantillonné noté s_0. La première étape pour atteindre ce but consiste à déterminer le « voisinage de krigeage » (Arnaud et Emery, 2000, p.180). Ce voisinage se définie par le domaine du champ D contenant s_0 ainsi que les sites $s_{[1]}$ à $s_{[n_0]}$ associés aux observations utilisées dans la prévision de $z(s_0)$. Ces sites doivent former un sous-ensemble de l'ensemble des sites d'observation, donc $\{s_{[1]}, s_{[2]}, ..., s_{[n_0]}\} \subseteq \{s_1, s_2, ..., s_n\}$. Ainsi, le voisinage de krigeage est en fait le même voisinage que celui introduit à la section 2.2.1. Il sera aussi noté $V(s_0)$. Le choix du voisinage de krigeage se base sur une certaine connaissance de la structure de dépendance spatiale entre les observations. La taille n_0 de ce voisinage doit cependant être assez

grande pour mener à une prévision précise.

Par la suite, la formule de prévision par krigeage peut être trouvée. Pour effectuer cette étape, la démarche proposée par Chauvet (1999, p.117) sera suivie dans ce mémoire. Chauvet conseille de procéder par l'écriture des contraintes du krigeage que voici :

1. **Contrainte de linéarité**
   La contrainte de base du krigeage est que la prévision prenne la forme d'une combinaison linéaire des données. Elle doit donc s'écrire ainsi :
   $$\hat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i)$$
   Les poids λ_i et la constante a sont les inconnus du problème.
2. **Contrainte d'autorisation**
   Il faut s'assurer que l'espérance et la variance de l'erreur de prévision $\hat{Z}(s_0) - Z(s_0)$ existent. Cette contrainte n'intervient que dans le cas où la fonction aléatoire $\delta(\cdot)$ du modèle de base 2.4 est supposée stationnaire intrinsèque.
3. **Contrainte de non-biais**
   La prévision par krigeage doit posséder la propriété d'absence de biais. Il faut donc que $E[\hat{Z}(s_0) - Z(s_0)] = 0$.
4. **Contrainte d'optimalité**
   Les poids λ_i et la constante a sont déterminés de façon à minimiser $Var[\hat{Z}(s_0) - Z(s_0)]$ sous les contraintes précédentes.

Cette démarche mènera à la résolution d'équations. Pour faciliter l'écriture de ces équations qui s'avèrent parfois longues, une notation matricielle sera employée. Voici donc tout d'abord quelques remarques sur cette notation :

- $\boldsymbol{Z}$ est le vecteur n_0x1 des variables aléatoires $Z(s_{[1]})$ à $Z(s_{[n_0]})$ intervenant dans la prévision de $Z(s_0)$
- $\boldsymbol{\lambda}$ est le vecteur n_0x1 des poids associés aux variables aléatoires $Z(s_{[1]})$ à $Z(s_{[n_0]})$
- $\boldsymbol{\delta}$ est le vecteur n_0x1 des erreurs associées aux variables aléatoires $Z(s_{[1]})$ à $Z(s_{[n_0]})$
- Dans le cadre stationnaire de second ordre :
  - $\boldsymbol{\Sigma}$ est la matrice n_0xn_0 de variances-covariances de $\boldsymbol{\delta}$, dont la diagonale est composée uniquement de σ^2, la variance commune à toutes les erreurs $\delta(s)$ pour $s \in D$
  - $\boldsymbol{c}_0$ est le vecteur n_0x1 des covariances entre $\boldsymbol{\delta}$ et $\delta(s_0)$

– Dans le cadre stationnaire intrinsèque :

  – $\mathbf{\Gamma}$ est la matrice n_0xn_0 dont l'élément (i, j) est $\gamma(s_{[i]} - s_{[j]})$, soit le semi-variogramme entre $\delta(s_{[i]})$ et $\delta(s_{[j]})$, les éléments i et j de $\boldsymbol{\delta}$

  – $\boldsymbol{\gamma}_0$ est le vecteur n_0x1 dont l'élément i est $\gamma(s_{[i]} - s_0)$, le semi-variogramme entre $\delta(s_{[i]})$ et $\delta(s_0)$

Notons que la prévision $\hat{Z}(s_0)$ obtenue est une variable aléatoire. Sa valeur numérique est calculée en remplaçant les variables aléatoires $Z(s_i)$ par les valeurs régionalisées observées $z(s_i)$.

4.2 Krigeage simple

La théorie du krigeage a d'abord été développée dans un cadre stationnaire de second ordre. Sous cette hypothèse, le krigeage le moins complexe est celui dans lequel la stationnarité postulée est de deuxième ordre et l'espérance de la fonction aléatoire étudiée est supposée connue et constante sur tout le champ. Il s'agit du krigeage simple (Matheron, 1970, p.122). Ce krigeage repose sur la modélisation suivante de la fonction aléatoire d'intérêt :

$$Z(s) = m + \delta(s), \qquad s \in D$$

avec m constante connue et $\delta(\cdot)$ fonction aléatoire stationnaire de second ordre d'espérance nulle et de structure de dépendance connue. La stationnarité de second ordre implique que le semi-variogramme de $\delta(\cdot)$ atteigne un palier.

Ce modèle peut être réécrit sous forme vectorielle en fonction de la variable aléatoire à prévoir $Z(s_0)$ et de celles qui serviront à faire la prévision $\boldsymbol{Z} = (Z(s_{[1]}), Z(s_{[2]}), ..., Z(s_{[n_0]}))^t$, soient les variables aléatoires associées aux données. Notons $\boldsymbol{Z}^* = (Z(s_0), \boldsymbol{Z}^t)$ et $\boldsymbol{\delta}^* = (\delta(s_0), \boldsymbol{\delta}^t)$, alors le modèle s'écrit :

$$\boldsymbol{Z}^* = m\mathbf{1}_{(n_0+1)} + \boldsymbol{\delta}^* \qquad \text{avec} \qquad \begin{cases} m \text{ constante connue} \\ E[\boldsymbol{\delta}^*] = \mathbf{0} \\ Var[\boldsymbol{\delta}^*] = \begin{pmatrix} \sigma^2 & \boldsymbol{c}_0^t \\ \boldsymbol{c}_0 & \boldsymbol{\Sigma} \end{pmatrix} \text{ connus} \end{cases} \tag{4.1}$$

Toutes les informations nécessaires à l'écriture des contraintes sont comprises dans ce modèle. Suivons donc la démarche proposée afin de trouver la formule de prévision en krigeage simple.

1. Contrainte de linéarité

La prévision de $Z(s_0)$ doit être de la forme :

$$\hat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) = a + \boldsymbol{\lambda}^t \boldsymbol{Z}$$

2. Contrainte d'autorisation

Cette contrainte n'intervient pas en krigeage simple car l'hypothèse de stationnarité de second ordre de la fonction aléatoire résiduelle implique l'existence de l'espérance et de la variance de toute variable aléatoire $Z(s)$ pour $s \in D$.

3. Contrainte de non-biais

La propriété d'absence de biais de la prévision doit être vérifiée. Il faut donc s'assurer du respect de l'égalité suivante :

$$E[\hat{Z}(s_0) - Z(s_0)] = E[a + \boldsymbol{\lambda}^t \boldsymbol{Z} - Z(s_0)] = a + \boldsymbol{\lambda}^t m \mathbf{1}_{n_0} - m = 0$$

Cette égalité implique que $a = m(1 - \boldsymbol{\lambda}^t \mathbf{1}_{n_0})$. La prévision peut donc être réécrite sous la forme :

$$\hat{Z}(s_0) = m(1 - \boldsymbol{\lambda}^t \mathbf{1}_{n_0}) + \boldsymbol{\lambda}^t \boldsymbol{Z} = m + \boldsymbol{\lambda}^t (\boldsymbol{Z} - m\mathbf{1}_{n_0})$$

4. Contrainte d'optimalité

Finalement, les inconnus de la formule précédente, c'est-à-dire les poids λ_i, sont trouvés en minimisant la variance de l'erreur de prévision. Cette variance s'écrit :

$$\begin{aligned}
Var[\hat{Z}(s_0) - Z(s_0)] &= Var[m + \boldsymbol{\lambda}^t(\boldsymbol{Z} - m\mathbf{1}_{n_0}) - Z(s_0)] \\
&= Var[m + \boldsymbol{\lambda}^t \boldsymbol{\delta} - m - \delta(s_0)] \\
&= Var[\boldsymbol{\lambda}^t \boldsymbol{\delta}] + Var[\delta(s_0)] - 2Cov[\boldsymbol{\lambda}^t \boldsymbol{\delta}, \delta(s_0)] \\
&= \boldsymbol{\lambda}^t Var[\boldsymbol{\delta}]\boldsymbol{\lambda} + Var[\delta(s_0)] - 2\boldsymbol{\lambda}^t Cov[\boldsymbol{\delta}, \delta(s_0)] \\
&= \boldsymbol{\lambda}^t \boldsymbol{\Sigma} \boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}^t \boldsymbol{c}_0
\end{aligned}$$

Cette expression doit être vue comme une fonction des λ_i, qui peut être notée $f(\lambda_1, ..., \lambda_{n_0}) = f(\boldsymbol{\lambda})$. Le gradient de cette fonction, c'est-à-dire le vecteur de ses dérivées

partielles, s'écrit :

$$\frac{\partial}{\partial \boldsymbol{\lambda}} f(\boldsymbol{\lambda}) = \frac{\partial}{\partial \boldsymbol{\lambda}} \left(\boldsymbol{\lambda}^t \boldsymbol{\Sigma} \boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}^t \boldsymbol{c}_0 \right) = 2\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{c}_0$$

Toutes les composantes de ce vecteur sont nulles au point $\hat{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0$. De plus, la matrice $n_0 \times n_0$ dont l'élément (i, j) est $\frac{\partial^2}{\partial \lambda_j \partial \lambda_i} f(\boldsymbol{\lambda})$, appelée hessien, est $2\boldsymbol{\Sigma}$. Cette matrice est semi-définie positive car il s'agit d'une matrice de variances-covariances multipliée par une constante positive. Ainsi, la fonction $f(\boldsymbol{\lambda})$ est convexe et le point critique $\hat{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0$ est un minimum global (Khuri, 1993, p.282).

Ainsi, la prévision de $Z(s_0)$ par le krigeage simple se formule :

$$\hat{Z}(s_0) = m + \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - m\mathbf{1}_{n_0}).$$

Cette expression est équivalente à $E[\hat{Z}(s_0)|\boldsymbol{Z}]$ si $Z(\cdot)$ est supposé de loi normale multidimentionnelle. Cette équivalence est discutée à la section 4.5.1. Une autre statistique d'intérêt est la valeur minimale de la variance de l'erreur de prévision, appelée « variance de krigeage ». En krigeage simple elle vaut :

$$\begin{aligned} \sigma^2(s_0) = Var[\hat{Z}(s_0) - Z(s_0)] &= \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0 + \sigma^2 - 2\boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0 \\ &= \sigma^2 - \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0 \end{aligned}$$

Cette statistique donne une idée de la précision de la prévision obtenue. Cependant, il ne faut pas oublier que cette variance ne tient pas compte de la variabilité due à l'estimation du semi-variogramme. Elle sous-estime donc la véritable variance de prévision.

4.3 Krigeage ordinaire

L'hypothèse du krigeage simple voulant que l'espérance de la fonction aléatoire $Z(\cdot)$ soit connue est rarement vérifiée. Cette méthode a donc été généralisée au cas où l'espérance est inconnue et constante localement, c'est-à-dire sur le voisinage de krigeage. Il s'agit du krigeage ordinaire (Matheron, 1970, p.125), la technique de krigeage la plus fréquemment utilisée selon Gratton (2002). Ce type de krigeage ne requière pas une hypothèse de stationnarité d'ordre deux. Ainsi, il sera développé ici sous l'hypothèse plus générale de stationnarité intrinsèque. Le modèle de base de cette méthode s'énonce comme suit :

$$Z(s) = \mu + \delta(s), \qquad s \in D \tag{4.2}$$

avec μ quasi-constante inconnue et $\delta(\cdot)$ fonction aléatoire stationnaire intrinsèque d'espérance nulle et de structure de dépendance connue. Le caractère *quasi-constant* de μ

signifie que l'espérance n'est pas contrainte à demeurer la même partout dans le champ D. Elle doit par contre rester constante à l'intérieur de chaque voisinage de krigeage (Arnaud et Emery, 2000, p.185). Ainsi, une prévision au point s_0 se base sur le modèle vectoriel suivant :

$$\boldsymbol{Z}^* = \mu \mathbf{1}_{(n_0+1)} + \boldsymbol{\delta}^* \qquad \text{avec} \qquad \begin{cases} \mu \text{ constante inconnue} \\ E[\boldsymbol{\delta}^*] = \mathbf{0} \\ \boldsymbol{\Gamma}, \boldsymbol{\gamma}_0 \quad \text{connus} \end{cases}$$

Voici, étape par étape, la résolution des équations du krigeage ordinaire se basant sur Cressie (1993, p.119).

1. Contrainte de linéarité

La prévision $\hat{Z}(s_0)$ doit être une combinaison linéaire des variables aléatoires $Z(s_{[1]})$ à $Z(s_{[n_0]})$. Elle s'exprime donc sous la forme :

$$\hat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) = a + \boldsymbol{\lambda}^t \boldsymbol{Z}$$

2. Contrainte d'autorisation

À cause de l'hypothèse de stationnarité intrinsèque de la fonction aléatoire $\delta(\cdot)$, seules les combinaisons linéaires d'accroissements telles :

$$\sum_{i=1}^{l} w_i(\delta(s_i) - \delta(s_i + h_i)) \quad \text{avec } s_i, s_i + h_i \in D \text{ pour tout } i = 1, ..., l$$

possèdent nécessairement des moments de premier et de deuxième ordre. Ainsi, l'erreur de prévision doit prendre la forme d'une somme pondérée d'accroissements afin d'assurer l'existence de son espérance et sa variance. Cependant, toute combinaison linéaire de l accroissements est aussi une combinaison linéaire de $2l$ variables aléatoires, avec la particularité que la somme des pondérateurs soit nulle. Cette affirmation se justifie par l'égalité $\sum_{i=1}^{l} w_i(\delta(s_i) - \delta(s_i + h_i)) = \sum_{i=1}^{l} w_i\delta(s_i) - \sum_{i=1}^{l} w_i\delta(s_i + h_i)$ où les $2l$ pondérateurs $w_1, ..., w_l, -w_1, ..., -w_l$ s'annulent. L'implication inverse est aussi vraie : toute somme pondérée de $\delta(s_i)$ dont l'addition des poids donne zéro est une combinaison linéaire d'accroissements. En effet, $\sum_{i=1}^{l} w_i\delta(s_i) = \sum_{i=1}^{l} w_i(\delta(s_i) - \delta(s_0))$ lorsque $\sum_{i=1}^{l} w_i = 0$. Ainsi, une combinaison linéaire d'accroissements est équivalent à une combinaison linéaire de variables aléatoires avec des poids de somme nulle. En réécrivant

l'erreur de prévision de la façon suivante :

$$\begin{aligned}\hat{Z}(s_0) - Z(s_0) &= a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0) \\ &= a + \sum_{i \in V(s_0)} \lambda_i (\mu + \delta(s_i)) - \mu - \delta(s_0) \\ &= \underbrace{a + \mu \sum_{i \in V(s_0)} \lambda_i - \mu}_{\text{termes non aléatoires}} + \sum_{i \in V(s_0)} \lambda_i \delta(s_i) - \delta(s_0)\end{aligned}$$

il ressort que cette erreur est une combinaison linéaire d'accroissements de $\delta(\cdot)$ si et seulement si $\sum_{i \in V(s_0)} \lambda_i - 1 = 0$. Il faudra donc travailler par la suite avec la contrainte que la somme des poids λ_i vaut un pour s'assurer de l'existence des deux premiers moments de l'erreur de prévision.

3. Contrainte de non-biais

Comme en krigeage simple, la prévision doit être non biaisée, il faut donc que :

$$E[\hat{Z}(s_0) - Z(s_0)] = E[a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0)] = a + \mu(\sum_{i \in V(s_0)} \lambda_i - 1) = 0$$

De l'étape précédente, la contrainte $\sum_{i \in V(s_0)} \lambda_i = 1$ doit être respectée. Ainsi, a doit être fixé à zéro. La forme de la prévision se simplifie donc à :

$$\hat{Z}(s_0) = \sum_{i \in V(s_0)} \lambda_i Z(s_i) \qquad \text{avec} \qquad \sum_{i \in V(s_0)} \lambda_i = 1$$

4. Contrainte d'optimalité

Trouvons maintenant les poids λ_i qui minimisent la variance de l'erreur de prévision. Pour ce faire, exprimons d'abord cette variance en fonction de $\mathbf{\Gamma}$ et $\boldsymbol{\gamma}_0$, car ces statistiques sont connues.

$$
\begin{aligned}
&Var[\hat{Z}(s_0) - Z(s_0)] \\
&= E[(\hat{Z}(s_0) - Z(s_0))^2] \\
&= E\left[\left(\sum_{i\in V(s_0)} \lambda_i Z(s_i) - Z(s_0)\right)^2\right] \\
&= E\left[\left(\sum_{i\in V(s_0)} \lambda_i \mu + \sum_{i\in V(s_0)} \lambda_i \delta(s_i) - \mu - \delta(s_0)\right)^2\right] \\
&= E\left[\left(\sum_{i\in V(s_0)} \lambda_i \delta(s_i) - \delta(s_0)\right)^2\right] \\
&= E\left[\left(\sum_{i\in V(s_0)} \lambda_i \delta(s_i)\right)^2 - 2\delta(s_0) \sum_{i\in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2\right] \\
&= E\left[\sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - 2\delta(s_0) \sum_{i\in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2\right] \\
&= E\Bigg[\underbrace{\sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum_{i\in V(s_0)} \lambda_i \delta(s_i)^2}_{1} + \\
&\qquad\qquad \underbrace{\sum_{i\in V(s_0)} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i\in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2}_{2}\Bigg]
\end{aligned}
$$

Le terme 1 peut être réécrit de la façon suivante :

$$
\begin{aligned}
&\sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum_{i\in V(s_0)} \lambda_i \delta(s_i)^2 \\
&= \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum_{j\in V(s_0)} \lambda_j \sum_{i\in V(s_0)} \lambda_i \delta(s_i)^2 \\
&= \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \frac{1}{2} \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_i)^2 - \frac{1}{2} \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \delta(s_j)^2 \\
&= -\frac{1}{2} \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \left(-2\delta(s_i)\delta(s_j) + \delta(s_i)^2 + \delta(s_j)^2\right) \\
&= -\frac{1}{2} \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i \lambda_j \left(\delta(s_i) - \delta(s_j)\right)^2
\end{aligned}
$$

Le terme 2 peut être réécrit de la façon suivante :

$$\begin{aligned}
&\sum_{i\in V(s_0)} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i\in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2 \\
&= \sum_{i\in V(s_0)} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i\in V(s_0)} \lambda_i \delta(s_i) + \sum_{i\in V(s_0)} \lambda_i \delta(s_0)^2 \\
&= \sum_{i\in V(s_0)} \lambda_i \left(\delta(s_i)^2 - 2\delta(s_0)\delta(s_i) + \delta(s_0)^2\right) \\
&= \sum_{i\in V(s_0)} \lambda_i \left(\delta(s_0) - \delta(s_i)\right)^2
\end{aligned}$$

Donc

$$\begin{aligned}
&Var[\hat{Z}(s_0) - Z(s_0)] \\
&= E\left[-\frac{1}{2} \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i\lambda_j \left(\delta(s_i) - \delta(s_j)\right)^2\right] + E\left[\sum_{i\in V(s_0)} \lambda_i \left(\delta(s_0) - \delta(s_i)\right)^2\right] \\
&= -\frac{1}{2} \sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i\lambda_j E\left[\left(\delta(s_i) - \delta(s_j)\right)^2\right] + \sum_{i\in V(s_0)} \lambda_i E\left[\left(\delta(s_0) - \delta(s_i)\right)^2\right] \\
&= -\sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i\lambda_j \frac{1}{2} Var\left[\delta(s_i) - \delta(s_j)\right] + 2\sum_{i\in V(s_0)} \lambda_i \frac{1}{2} Var\left[\delta(s_0) - \delta(s_i)\right] \\
&= -\sum_{i\in V(s_0)} \sum_{j\in V(s_0)} \lambda_i\lambda_j \gamma\left(s_i - s_j\right) + 2\sum_{i\in V(s_0)} \lambda_i \gamma\left(s_0 - s_i\right) \\
&= -\boldsymbol{\lambda}^t \boldsymbol{\Gamma} \boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t \boldsymbol{\gamma}_0
\end{aligned}$$

Il faut minimiser cette expression sous la contrainte $\boldsymbol{\lambda}^t \mathbf{1}_{n_0} = 1$. Cette étape sera effectuée à l'aide d'un Lagrangien noté ℓ. La fonction à minimiser est : $f(\boldsymbol{\lambda}, \ell) = -\boldsymbol{\lambda}^t \boldsymbol{\Gamma} \boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t \boldsymbol{\gamma}_0 + 2\ell(\boldsymbol{\lambda}^t \mathbf{1}_{n_0} - 1)$. Le vecteur de ses dérivées partielles par rapport aux λ_i est :

$$\frac{\partial}{\partial \boldsymbol{\lambda}} f(\boldsymbol{\lambda}, \ell) = -2\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\gamma}_0 + 2\ell \mathbf{1}_{n_0}$$

La fonction $f(\boldsymbol{\lambda}, \ell)$ admet un point critique lorsque $\frac{\partial}{\partial \boldsymbol{\lambda}} f(\boldsymbol{\lambda}, \ell) = \mathbf{0}$, c'est-à-dire au point $\hat{\boldsymbol{\lambda}} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \ell \mathbf{1}_{n_0})$.

Le Lagrangien ℓ est estimé en utilisant la contrainte que la somme des poids vaille 1, donc que $\mathbf{1}_{n_0}^t \hat{\boldsymbol{\lambda}} = \mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \ell \mathbf{1}_{n_0}) = 1$. On obtient :

$$\hat{\ell} = \frac{1 - \mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}_0}{\mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1} \mathbf{1}_{n_0}}$$

Ensuite, ℓ est remplacé par $\hat{\ell}$ dans $\hat{\boldsymbol{\lambda}}$.

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \frac{1 - \mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{\mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}\mathbf{1}_{n_0}}\mathbf{1}_{n_0})$$

Matheron (1969, p.35) a prouvé qu'il s'agit bien de l'unique minimum de $f(\boldsymbol{\lambda}, \ell)$.

Ainsi, $Z(s_0)$ est prévu en krigeage ordinaire par l'expression :

$$\hat{Z}(s_0) = (\boldsymbol{\gamma}_0 + \frac{1 - \mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{\mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}\mathbf{1}_{n_0}}\mathbf{1}_{n_0})^t \boldsymbol{\Gamma}^{-1} Z$$

La variance de krigeage s'écrit alors :

$$\sigma^2(s_0) = Var[\hat{Z}(s_0) - Z(s_0)] = \boldsymbol{\gamma}_0^t \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - \frac{(1 - \mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0)^2}{\mathbf{1}_{n_0}^t \boldsymbol{\Gamma}^{-1}\mathbf{1}_{n_0}} \quad .$$

4.4 Krigeage avec modèle de tendance

L'hypothèse de stationnarité sur laquelle repose les deux types de krigeage présentés précédemment peut souvent être mise en doute. En particulier, il semble souvent erroné de postuler que l'espérance de la fonction aléatoire étudiée reste constante ou quasi-constante sur le champ D. En krigeage avec un modèle de tendance, aussi appelé krigeage en présence d'une dérive, l'espérance est une fonction des coordonnées spatiales ou de variables régionalisées auxiliaires connues exhaustivement. La majorité des auteurs en géostatistique parlent alors de krigeage universel (Matheron, 1969 ; Cressie, 1993, p.151) et de krigeage avec dérive externe (Goovaerts, 1997, p.194 ; Wackernagel, 2003, p.283) respectivement.

Krigeage universel : Le modèle de base du krigeage universel est :

$$Z(s) = \sum_{j=0}^{p} f_j(s)\beta_j + \delta(s), \qquad s \in D \tag{4.3}$$

avec $f_j(s)$ fonctions de la position $s = (x, y)$, β_j paramètres inconnus et $\delta(\cdot)$ fonction aléatoire stationnaire intrinsèque d'espérance nulle et de structure de dépendance connue. Les $f_j(s)$ sont déterminés par l'utilisateur. Les graphiques des $z(s_i)$ en fonction des coordonnées x_i et y_i sont utilisés pour guider le choix de ces fonctions. Souvent, une tendance linéaire ou quadratique est choisie. Par exemple, dans le cas d'une tendance linéaire, les $f_j(s)$ sont : $f_0(s) = 1$, $f_1(s) = x$ et $f_2(s) = y$.

Krigeage avec dérive externe : En krigeage avec dérive externe, le modèle s'écrit :

$$Z(s) = \sum_{j=0}^{p} f_j(\boldsymbol{w})\beta_j + \delta(s), \qquad s \in D$$

où $\boldsymbol{w} = (w_1(s), ..., w_q(s))$ est le vecteur des valeurs prises par les q variables régionalisées auxiliaires au point s. Si une tendance linéaire est choisie, les $f_j(\boldsymbol{w})$ seront $f_0(\boldsymbol{w}) = 1$, $f_1(\boldsymbol{w}) = w_1$, ... , $f_q(\boldsymbol{w}) = w_q$.

La résolution des équations du krigeage est la même pour ces deux types de krigeage. Dans les paragraphes qui suivent, nous utilisons la notation du krigeage universel, mais la démarche est exactement la même en krigeage avec dérive externe en changeant les $f_j(s)$ pour des $f_j(\boldsymbol{w})$. En fait, le modèle de tendance peut aussi être composé de fonctions des coordonnées spatiales et de variables régionalisées auxiliaires simultanément.

Sous forme matricielle, le modèle utilisé pour prévoir $Z(s_0)$ s'énonce comme suit :

$$\boldsymbol{Z}^* = \begin{pmatrix} \boldsymbol{x}_0\boldsymbol{\beta} \\ \boldsymbol{X}\boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\delta}^* \qquad \text{avec} \qquad \begin{cases} E[\boldsymbol{\delta}^*] = \boldsymbol{0} \\ \boldsymbol{\Gamma}, \boldsymbol{\gamma}_0 \quad \text{connus} \end{cases} \tag{4.4}$$

où $\boldsymbol{Z}^* = (Z(s_0), \boldsymbol{Z})$, $\boldsymbol{\delta}^* = (\delta(s_0), \boldsymbol{\delta})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$, $\boldsymbol{x}_0 = (f_0(s_0), f_1(s_0), ..., f_p(s_0))$ et $\boldsymbol{X}$ est une matrice $n_0 \times (p+1)$ dont l'élément (i, j) est $f_j(s_i)$. Suivons la démarche des contraintes afin de prévoir $Z(s_0)$ par krigeage universel.

1. Contrainte de linéarité

Encore une fois, $\hat{Z}(s_0)$ doit être une combinaison linéaire des $Z(s_i)$. Il s'écrit donc sous la forme :

$$\hat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) = a + \boldsymbol{\lambda}^t \boldsymbol{Z}$$

2. Contrainte d'autorisation

Comme en krigeage ordinaire, à cause de l'hypothèse de stationnarité intrinsèque, il faut s'assurer que l'erreur $\hat{Z}(s_0) - Z(s_0)$ soit une combinaison linéaire d'accroissements de $\delta(\cdot)$. Pour en être une, on a vu à la section précédente que la somme des poids des

termes aléatoires doit valoir zéro. L'erreur de prévision s'écrit :

$$\begin{aligned}\hat{Z}(s_0) - Z(s_0) &= a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0) \\ &= a + \sum_{i \in V(s_0)} \lambda_i (\boldsymbol{x}_i \boldsymbol{\beta} + \delta(s_i)) - \boldsymbol{x}_0 \boldsymbol{\beta} - \delta(s_0) \\ &= \underbrace{a + \sum_{i \in V(s_0)} \lambda_i \boldsymbol{x}_i \boldsymbol{\beta} - \boldsymbol{x}_0 \boldsymbol{\beta}}_{\text{termes non aléatoires}} + \sum_{i \in V(s_0)} \lambda_i \delta(s_i) - \delta(s_0)\end{aligned}$$

Il faut donc encore que $\sum_{i \in V(s_0)} \lambda_i = 1$ pour assurer l'existence de l'espérance et de la variance de $\hat{Z}(s_0) - Z(s_0)$.

3. Contrainte de non-biais

L'espérance de l'erreur de prévision s'écrit :

$$\begin{aligned}E[\hat{Z}(s_0) - Z(s_0)] &= E[a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0)] \\ &= a + \sum_{i \in V(s_0)} \lambda_i \boldsymbol{x}_i \boldsymbol{\beta} - \boldsymbol{x}_0 \boldsymbol{\beta} \\ &= a + \sum_{i \in V(s_0)} \lambda_i \sum_{j=0}^{p} f_j(s_i) \beta_j - \sum_{j=0}^{p} f_j(s_0) \beta_j \\ &= a + \sum_{j=0}^{p} \left(\sum_{i \in V(s_0)} \lambda_i f_j(s_i) - f_j(s_0) \right) \beta_j\end{aligned}$$

Afin que cette espérance vaille zéro pour tout β_j, $j = 0, ..., p$, il faut que $a = 0$ et $\sum_{i \in V(s_0)} \lambda_i f_j(s_i) - f_j(s_0) = 0$ pour $j = 0, ..., p$. Ces contraintes s'écrivent $\boldsymbol{\lambda}^t \boldsymbol{X} = \boldsymbol{x}_0^t$ sous forme matricielle.

Ainsi, sans oublier la contrainte d'autorisation $\sum_{i \in V(s_0)} \lambda_i = 1$, il y a au total $p + 2$ contraintes sur les poids $\boldsymbol{\lambda}$ en krigeage universel. Toutefois, afin de simplifier les calculs, on suppose que $f_0(\cdot) = 1$. Ainsi, $\sum_{i \in V(s_0)} \lambda_i f_j(s_i) = f_j(s_0)$ pour $j = 0$ revient à $\sum_{i \in V(s_0)} \lambda_i = 1$. Ce postulat est usuel en krigeage (Cressie, 1993, p.152 ; Goovaerts, 1997, p.140). Il permet d'éliminer une contrainte. L'estimateur devient donc :

$$\hat{Z}(s_0) = \sum_{i \in V(s_0)} \lambda_i Z(s_i) \quad \text{avec} \quad \sum_{i \in V(s_0)} \lambda_i f_j(s_i) = f_j(s_0) \text{ pour } j = 0, ..., p$$

4. Contrainte d'optimalité

Finalisons la démarche en minimisant $Var[\hat{Z}(s_0) - Z(s_0)]$. Comme pour le krigeage ordinaire, cette variance vaut $-\boldsymbol{\lambda}^t\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t\boldsymbol{\gamma}_0$. Cette expression doit être minimisée en respectant les contraintes d'autorisation et de non-biais $\boldsymbol{\lambda}^t\boldsymbol{X} = \boldsymbol{x}_0^t$. Cette étape sera effectuée à l'aide d'un vecteur, noté $\boldsymbol{\ell}$, de $p+1$ Lagrangiens. La fonction à minimiser est donc : $f(\boldsymbol{\lambda}, \boldsymbol{\ell}) = -\boldsymbol{\lambda}^t\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t\boldsymbol{\gamma}_0 + 2(\boldsymbol{\lambda}^t\boldsymbol{X} - \boldsymbol{x}_0^t)\boldsymbol{\ell}$. Son gradient est :

$$\frac{\partial}{\partial \boldsymbol{\lambda}} f(\boldsymbol{\lambda}, \boldsymbol{\ell}) = -2\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\gamma}_0 + 2\boldsymbol{X}\boldsymbol{\ell}$$

Il vaut zéro au point $\hat{\boldsymbol{\lambda}} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \boldsymbol{X}\boldsymbol{\ell})$. En utilisant la contrainte $\boldsymbol{X}^t\hat{\boldsymbol{\lambda}} = \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \boldsymbol{X}\boldsymbol{\ell}) = \boldsymbol{x}_0$, le vecteur de Lagrangiens $\boldsymbol{\ell}$ est estimé par :

$$\hat{\boldsymbol{\ell}} = (\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0)$$

Ensuite, $\boldsymbol{\ell}$ est remplacé par $\hat{\boldsymbol{\ell}}$ dans $\hat{\boldsymbol{\lambda}}$. L'unique point minimum de la fonction $f(\boldsymbol{\lambda}, \boldsymbol{\ell})$ (Matheron, 1969, p.35) est :

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0))$$

Ainsi, $Z(s_0)$est prévu par

$$\hat{Z}(s_0) = (\boldsymbol{\gamma}_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0))^t\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}$$

En outre, la variance de krigeage devient :

$$\begin{aligned} \sigma^2(s_0) &= Var[\hat{Z}(s_0) - Z(s_0)] \\ &= \boldsymbol{\gamma}_0^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - (\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0)^t(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0) \end{aligned}$$

Notons que le krigeage ordinaire peut être vu comme un cas particulier du krigeage universel avec $p = 0$ et $f_0(\cdot)$=1. Ainsi, la prévision en krigeage ordinaire, ainsi que sa variance, peuvent être retrouvées en remplaçant $\boldsymbol{X}$ par $\mathbf{1}_{n_0}$ et $\boldsymbol{x}_0$ par 1 dans les deux expressions précédentes.

4.4.1 Lien entre le krigeage avec modèle de tendance et le krigeage sur les résidus d'une régression

Dans un cas stationnaire d'ordre deux, une prévision par krigeage avec modèle de tendance prend la forme :

$$\hat{Z}(s_0) = (\boldsymbol{c}_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0))^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}$$

La même prévision peut être obtenue en effectuant un krigeage simple avec espérance nulle ($m = 0$) sur les résidus d'une régression linéaire qui tient compte de la dépendance spatiale des erreurs (Hengl *et al.*, 2003). Les paramètres du modèle sont donc estimés par une méthode des moindres carrés généralisés ou par une méthode du maximum de vraisemblance avec une matrice de variances-covariances non diagonale. Dans ce cas, une prévision de $Z(s_0)$ est formée en additionnant la prévision de la tendance générale par régression à la prévision par krigeage simple des erreurs $\boldsymbol{e} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{gls}$:

$$\begin{aligned}
\hat{Z}(s_0) &= \boldsymbol{x}_0^t\hat{\boldsymbol{\beta}}_{gls} + \hat{e}_{KS}(s_0) \\
&= \boldsymbol{x}_0^t\hat{\boldsymbol{\beta}}_{gls} + \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{e} \\
&= \boldsymbol{x}_0^t\hat{\boldsymbol{\beta}}_{gls} + \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{gls}) \\
&= (\boldsymbol{x}_0^t - \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\hat{\boldsymbol{\beta}}_{gls} + \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \\
&= (\boldsymbol{x}_0^t - \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \\
&= (\boldsymbol{c}_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0))^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}
\end{aligned}$$

Cependant, certains utilisent cette approche en effectuant une régression qui ne tient pas compte de la dépendance spatiale des données, par exemple une régression des moindres carrés ordinaires ($\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Z}$). Dans ce cas, les prévisions ne sont pas tout à fait les mêmes que celles obtenues par krigeage avec modèle de tendance. Ce type de krigeage est parfois appelé en anglais « detrended kriging » (Nalder et Wein, 1998; Phillips *et al.*, 1992). Enfin, une autre approche parfois utilisée consiste à simplement effectuer un krigeage ordinaire sur des erreurs $z(s_i) - w(s_i)$. Ce type de krigeage porte parfois le nom de « krigeage résiduel » (Hessami, 2002, p.29). Ce krigeage est en fait un krigeage avec dérive externe avec une seule variable régionalisée auxiliaire, donc $q = 1$ et $\boldsymbol{w} = w(s)$, en fixant $f_0(\boldsymbol{w}) = 1$, $f_1(\boldsymbol{w}) = w(s)$ et $\beta_1 = 1$. Ainsi, β_0 est le seul paramètre inconnu dans la dérive. Le modèle s'écrit donc $Z(s) = \beta_0 + w(s) + \delta(s)$, ou $Z(s) - w(s) = \beta_0 + \delta(s)$ pour $s \in D$. Ce modèle est particulièrement intéressant lorsque la variable régionalisée auxiliaire mesure le même phénomène naturel que la variable régionalisée à interpoler, mais avec moins de précision.

4.4.2 Problème de l'analyse variographique en krigeage avec modèle de tendance

L'analyse variographique est une étape problématique en krigeage avec un modèle de tendance. Tel que mentionné à la section 3.2.3, le semi-variogramme est estimé à partir des résidus $e(s_i) = z(s_i) - \hat{\mu}(s_i)$ obtenus suite à une estimation de $\boldsymbol{\beta}$. Il s'agit d'un problème de régression qui peut être résolu par une méthode des moindres carrés. Cependant, le modèle de régression est particulier : les erreurs ne sont pas indépendantes.

Une estimation correcte de $\boldsymbol{\beta}$ requiert la connaissance de la structure de dépendance spatiale de la fonction aléatoire résiduelle $\delta(\cdot)$. Toutefois, cette structure est inconnue. L'analyse variographique vise justement à l'estimer. Le problème revient donc à son point de départ.

Afin de sortir de ce cercle vicieux, une solution consiste à d'abord obtenir un estimateur des moindres carrés ordinaires de $\boldsymbol{\beta}$, estimer ensuite un semi-variogramme sur les résidus, puis calculer un estimateur des moindres carrés généralisés de $\boldsymbol{\beta}$, et ainsi de suite (Cressie, 1993, p.166). Au fil des itérations, les résidus des moindres carrés ordinaires s'approchent de plus en plus des résidus des moindres carrés généralisés. Hengl *et al.* (2003) affirment qu'en pratique une seule itération suffit pour obtenir des résultats de krigeage satisfaisants.

Cependant, Cressie (1993, p.167) soulève que même en appliquant cette méthode itérative, l'estimation du semi-variogramme reste biaisée (Matheron, 1969, p.37, 1970, p.155). Les conséquences de ce biais sont cependant difficiles à évaluer. Ainsi, le krigeage avec un modèle de tendance reste une méthode qui peut s'avérer bonne, mais la prudence est de mise lors de son utilisation. Le biais dans l'estimation du semi-variogramme en krigeage avec modèle de tendance est une des motivations au développement de méthodes bayésiennes.

4.5 Discussions

Les trois sections précédentes ont présenté l'essentiel de la théorie du krigeage formalisée par Matheron (1962, 1963a,b, 1965, 1969, 1970). Ces sections devraient donc suffirent à une compréhension de la méthode dans le but d'une utilisation de base. Cependant, avant de passer à la mise en oeuvre du krigeage, il est intéressant de s'attarder sur certains aspects théoriques plus poussés ainsi que sur d'autres types de krigeage développés plus récemment. Les remarques théoriques traitées ici touchent la normalité des données, le travail sur des données transformées, notamment par la fonction logarithmique, et le cokrigeage, une extension du krigeage dans le cas multivariable.

4.5.1 Normalité des données

En statistique classique, les méthodes reposent souvent sur un postulat de normalité des erreurs du modèle. En krigeage, c'est différent. Des hypothèses sont émises seulement

sur les moments de la distribution de $Z(\cdot)$ par l'intermédiaire de $\delta(\cdot)$ et non sur la loi de la distribution. L'utilisateur du krigeage peut donc croire que la performance de la méthode est indépendante de la loi de distribution des données. Malheureusement, ce n'est pas le cas. Le krigeage a tendance à fournir de meilleures prévisions lorsque les données suivent une loi normale.

Cette remarque peut être vérifiée en pratique, mais elle s'explique aussi théoriquement. Elle est attribuable à la contrainte en apparence anodine de linéarité de la prévision. Cette hypothèse est typiquement inadéquate dans un contexte non normal (Cressie, 1993, p.110). Cependant, si $Z(\cdot)$ est une fonction aléatoire gaussienne, cette linéarité est appropriée.

Pour justifier cette affirmation, rappelons d'abord que la prévision par krigeage est linéaire, sans biais et qu'elle minimise la variance de l'erreur de prévision. Elle minimise donc l'erreur quadratique moyenne :

$$E[(\hat{Z}(s_0) - Z(s_0))^2] = Var[\hat{Z}(s_0) - Z(s_0)] + \underbrace{E[\hat{Z}(s_0) - Z(s_0)]^2}_{\text{vaut 0 par absence de biais}}$$

Une prévision non contrainte à être linéaire et minimisant l'erreur quadratique moyenne conditionnelle $E[(\hat{Z}(s_0) - Z(s_0))^2|\boldsymbol{Z}]$ serait $E[Z(s_0)|\boldsymbol{Z}]$ (Cressie, 1993, p.108). Théoriquement, cette prévision spatiale est équivalente ou meilleure à celle obtenue par krigeage. Cependant, elle dépend de la loi de distribution de $Z(\cdot)$. Supposons donc que $Z(\cdot)$ est une fonction aléatoire gaussienne telle que :

$$\begin{pmatrix} Z(s_0) \\ \boldsymbol{Z} \end{pmatrix} \sim N_{(n+1)} \left(\begin{pmatrix} \boldsymbol{x_0}^t\boldsymbol{\beta} \\ \boldsymbol{X\beta} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \boldsymbol{c}_0^t \\ \boldsymbol{c}_0 & \boldsymbol{\Sigma} \end{pmatrix} \right)$$

Alors, de la formule de l'espérance conditionnelle dans un cas normal multivarié (Rencher, 1995, p.99), nous obtenons :

$$E(Z(s_0)|\boldsymbol{Z}) = \boldsymbol{x_0}^t\boldsymbol{\beta} + (\boldsymbol{Z} - \boldsymbol{X\beta})^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0.$$

Cette expression est bien une combinaison linéaire des $Z(s_i)$. L'hypothèse de linéarité est par conséquent appropriée dans le cas normal. De plus, cette prévision est en fait la même que celle du krigeage universel dans un cadre stationnaire de second ordre si $\boldsymbol{\beta}$ est remplacée par son estimation du maximum de vraisemblance $\hat{\boldsymbol{\beta}}_{\text{MV}} = (\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}$. En effet, nous obtenons alors :

$$\begin{aligned} \hat{E}_{\text{MV}}(Z(s_0)|\boldsymbol{Z}) &= \boldsymbol{x_0}^t\hat{\boldsymbol{\beta}}_{\text{MV}} + (\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{MV}})^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0 \\ &= \boldsymbol{c_0}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + (\boldsymbol{x_0} - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c_0})^t\hat{\boldsymbol{\beta}}_{\text{MV}} \\ &= (\boldsymbol{c}_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c_0}))^t\boldsymbol{\Sigma}^{-1}Z \end{aligned}$$

Il en est de même en krigeage simple et ordinaire. Dans le cas du krigeage simple, l'espérance de $Z(\cdot)$ est connue, donc aucun paramètre n'a a être estimé dans la prévision optimale $E(Z(s_0)|\boldsymbol{Z}) = \mu + (\boldsymbol{Z} - \mu \mathbf{1}_{n_0})^t \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0$.

Il faut donc garder en tête que la prévision spatiale par krigeage peut mal performer lorsque $Z(\cdot)$ est loin d'être gaussienne. Dans un tel cas, une transformation des données ou encore l'utilisation de méthodes de la géostatistique non-linéaire peuvent être envisagées. Ces deux solutions sont détaillées dans les prochains points de discussion.

4.5.2 Transformation de données

Il peut parfois être avantageux de travailler sur une transformation des données plutôt que sur les données brutes. Par transformation de données, nous entendons ici une fonction mathématique (par exemple le logarithme, la racine, le carré, etc.) appliquée aux données. Une transformation peut aider à rendre la distribution des données plus près d'une distribution gaussienne. Elle peut aussi être employée afin de s'assurer que les prévisions par krigeage prennent des valeurs à l'intérieur d'un certain intervalle. Par exemple, si une variable régionalisée telle la concentration d'un polluant dans l'air est étudiée, ses valeurs régionalisées seront obligatoirement comprises entre 0 et 100. Cependant, rien n'empêche une prévision par krigeage de cette variable de prendre une valeur à l'extérieur de cet intervalle. Le même problème se présente pour des variables régionalisées positives telle une quantité de précipitations tombées. Afin de régler ce problème, une solution possible est d'ajouter des contraintes aux poids de krigeage. Par exemple, en forçant les poids du krigeage à être positifs, les prévisions seront assurément positives si les valeurs régionalisées sont toutes positives. Une autre solution est d'appliquer une transformation aux données.

Lorsqu'une transformation de données est utilisée afin de corriger un écart à la normalité, une transformation de la famille proposée par Box et Cox (1964) est suggérée. Pour des données de type pourcentage, une transformation logit ou probit serait appropriée. Rappelons que $logit(Z(s)) = \ln\left(\frac{Z(s)}{1-Z(s)}\right)$ et $probit(Z(s)) = \Phi^{-1}(Z(s))$ où $\Phi^{-1}(\cdot)$ est l'inverse de la fonction de répartition normale standard. Finalement, dans le cas de variables régionalisées positives, extraire la racine ou appliquer la fonction logarithmique sont envisageables. Pour cette dernière option, si les valeurs régionalisées peuvent être nulles, elles doivent être additionnées d'une constante positive avant de les log-transformer.

En krigeage, l'emploi d'une transformation de données s'effectue en trois étapes :

1. la transformation des données,
2. la mise en oeuvre du krigeage (incluant l'analyse variographique),
3. la transformation inverse des prévisions obtenues.

Ce dernier point est l'étape délicate car il ne suffit pas d'utiliser la fonction inverse de celle utilisée pour transformer les données. Afin de s'assurer du non-bais des prévisions, un facteur correctif doit être intégré à la transformation inverse. Ce facteur dépend notamment de la transformation choisie et du type de krigeage employé. Cependant, la littérature sur les transformations de données en krigeage avec modèle de tendance semble inexistante. Seul le cas du krigeage ordinaire, et parfois celui du krigeage simple, sont traités. La détermination du facteur correctif est une étape théorique qui n'est pas détaillée dans ce mémoire. Seuls les formules finales de prévision sont présentées.

Krigeage lognormal

La transformation de données la plus usuelle en géostatistique est la transformation logarithmique. Le krigeage avec cette transformation a été baptisé, il s'agit du « krigeage lognormal ». Matheron (1962, p.257) a lui même abordé le sujet dans sa première publication sur le krigeage. Un excellent résumé des principales publications sur le krigeage lognormal a été écrit par Rivoirard (1990). Il met l'accent sur le fait que le terme krigeage lognormal ne fait pas référence à une unique formule de prévision. Dans ce mémoire, le krigeage lognormal mentionné par Cressie (1993, p.135) est présenté.

Notons $\{Y(s) = \ln(Z(s)), s \in D\}$ la nouvelle fonction aléatoire associée à la variable régionalisée transformée. Une prévision spatiale de $Z(s_0)$ s'obtient par krigeage lognormal avec la formule :

$$\hat{Z}(s_0) = \exp\left(\hat{Y}(s_0) + \sigma_Y^2(s_0)/2 - \ell_Y\right)$$

où $\hat{Y}(s_0)$ est la prévision de $Y(s_0)$ par krigeage ordinaire, $\sigma_Y^2(s_0)$ est la variance de cette prévision et ℓ_Y est le lagrangien intervenant dans la résolution des équations du krigeage ordinaire sur les $Y(s_i)$, $i \in V(s_0)$. Le facteur $\exp\left(\sigma_Y^2(s_0)/2 - \ell_Y\right)$ sert donc à corriger pour le biais de la prévision. La variance de cet estimateur est donnée par Cressie (1993, p.136).

Il faut cependant mentionner que Roth (1998) s'interroge sur la pertinence du krigeage lognormal pour la prévision spatiale. Il soutient que la propriété d'absence de biais de sa prévision provient d'une surestimation de la majorité des valeurs régionalisées compensée par une sous-estimation de quelques valeurs plus grandes. Il conseille donc d'utiliser avec précaution cette méthode.

Krigeage trans-gaussien

Plus généralement, Cressie (1993, p.137) propose une formule de prévision qui suppose que la fonction aléatoire transformée est $Y(s) = \phi^{-1}(Z(s))$ pour tout $s \in D$ où ϕ est une fonction deux fois différentiable choisie telle que $Y(\cdot)$ soit gausienne. La prévision s'écrit :

$$\hat{Z}(s_0) = \phi(\hat{Y}(s_0)) + \phi''(\hat{\mu}_Y)\{\sigma_Y^2(s_0)/2 - \ell_Y\}$$

où $\hat{Y}(s_0)$ est la prévision de $Y(s_0)$ par krigeage ordinaire, $\sigma_Y^2(s_0)$ est la variance de cette prévision, ℓ_Y est le lagrangien intervenant dans la résolution des équations du krigeage ordinaire sur les $Y(s_i) = \phi^{-1}(Z(s_i))$, $i \in V(s_0)$ et $\hat{\mu}_Y = \mathbf{1}_{n_0}^t \mathbf{\Sigma}_Y^{-1} \boldsymbol{Y} / (\mathbf{1}_{n_0}^t \mathbf{\Sigma}_Y^{-1} \mathbf{1}_{n_0})$. Il est bien mis en évidence par cette formule que l'analyse variographique doit être effectuée sur les données transformées. Cette prévision se base sur un développement en série Taylor de $\phi(Y(s))$ autour de μ_Y, l'espérance de $Y(s)$. Elle approximativement sans biais pour $\sigma_Y^2(s_0)$ petit. Les détails de la démarche menant à cette prévision peuvent être trouvés dans Kozintseva (1999, p.17). Notons que pour $\phi(\cdot) = \exp(\cdot)$, les prévisions des krigeages trans-gaussien et lognormal sont approximativement les mêmes.

Dans la même ligne d'idées, De Olivieira *et al.* (1997) ont proposé un modèle bayésien trans-gaussien. Ce modèle ne requiert pas la détermination préalable de la transformation, mais plutôt d'une famille paramétrique de transformations.

4.5.3 Géostatistique multivariable

Jusqu'à maintenant, dans ce chapitre, l'accent a été mis sur l'aspect stationnaire ou non des différents types de krigeage. Deux branches de la géostatistique se sont dessinées : la géostatistique stationnaire et la géostatistique non-stationnaire. Le krigeage simple et le krigeage ordinaire (sections 4.2 et 4.3) se classent dans la division stationnaire et le krigeage universel et le krigeage avec dérive externe (section 4.4) dans la division non-stationnaire. Un autre classement possible des techniques géostatistiques concerne le nombre de variables régionalisées incluses dans l'analyse. Cet aspect a d'ailleurs été abordé à la section 2.2.6. En géostatistique univariable, une seule variable régionalisée intervient dans les calculs tandis qu'en géostatistique multivariable, plus d'une variable régionalisées est exploitée. Ainsi, le krigeage simple, le krigeage ordinaire et le krigeage universel se classent dans la catégorie univariable et le krigeage avec dérive externe dans la branche multivariable.

En plus du krigeage avec dérive externe, la géostatistique multivariable comporte le

cokrigeage, brièvement présenté à la section 2.2.6. Dans cette section, nous revenons sur cette méthode pour l'approfondir davantage. Mais avant, mentionnons qu'une interpolation multivariable peut parfois être effectuée en utilisant une technique univariable. Par exemple, des variables auxiliaires catégoriques, telles qu'un type de roche, peuvent permettre de partitionner le champ D en strates distinctes. Dans ce cas, s'il y a suffisamment de points d'observation, le krigeage peut s'effectuer à l'intérieur des strates (Goovaerts, 1997, p.187) plutôt que sur tout le champ. Ainsi, un semi-variogramme distinct est calculé par strate et une prévision en un point est effectuée uniquement à partir des points d'observation appartenant à la même strate. En ce qui concerne les variables auxiliaires continues, une façon simple de les exploiter est d'effectuer un krigeage simple avec une espérance variant localement (Goovaerts, 1997, p.190). Cette espérance, supposée connue, est alors estimée en chaque point de prévision par un modèle de régression incluant les variables régionalisées auxiliaires.

L'objectif de la géostatistique multivariable est d'améliorer les prévisions en exploitant le lien entre des variables régionalisées. Les méthodes de la géostatistique multivariable peuvent être particulièrement utiles lorsque peu d'observations de la variable régionalisée à interpoler sont disponibles. En pratique, il faut cependant savoir que l'utilisation d'une de ces techniques semble pouvoir améliorer la qualité des prévisions obtenues seulement si les corrélations entre les variables régionalisées auxiliaires et la variable régionalisée à interpoler sont supérieures à 0.4 en valeur absolue (Asli et Marcotte, 1995).

Cokrigeage

Le cokrigeage prévoit la valeur de la variable régionalisée $z(\cdot)$ en un point s_0 par une combinaison linéaire de la forme :

$$\hat{z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_{i,0} z(s_i) + \sum_{j=i}^{q} \sum_{i \in V(s_0)} \lambda_{i,j} w_j(s_{i,j}).$$

Les poids de cette combinaison linéaire sont choisis de façon à minimiser la variance de l'erreur de prévision sous une contrainte de non-biais, comme en krigeage. Pour ce faire, toutes les variables régionalisées sont modélisées par des fonctions aléatoires, même les variables régionalisées auxiliaires. C'est d'ailleurs ce qui distingue le cokrigeage des autres techniques de la géostatistique multivariable. Cette modélisation stochastique permet d'exploiter, en plus des relations entre les variables régionalisées, la comportement spatial de chacune de ces variables.

Le modèle de base est maintenant (Goovaerts, 1997, p.203) :

$$Z(s) = \mu_0(s) + \delta_0(s) \quad \text{et} \quad W_j(s) = \mu_j(s) + \delta_j(s) \text{ pour tout } j \in \{1, ..., q\}, \quad s \in D$$

où les $\mu_j(\cdot)$ sont des structures déterministes pour les espérances et les $\delta_j(\cdot)$ sont des fonctions aléatoires, chacune d'espérance nulle et conjointement stationnaires (de deuxième ordre ou intrinsèquement) de structure de dépendance connue. La stationnarité conjointe (Arnaud et Emery, 2000, p.157) implique que :

Dans le cadre stationnaire d'ordre deux : $Cov(\delta_j(s), \delta_k(s+h))$ soit uniquement fonction de h, le vecteur de translation entre les points s et $s+h$, pour tout couple $j, k \in \{0, ..., q\}$ et toute paire de points $s, s+h \in D$. Ces fonctions de covariance, notées $C_{jk}(h)$, sont nommées covariogrammes simples ($j = k$) et croisés ($j \neq k$) ;

Dans le cadre stationnaire intrinsèque : $Cov[\delta_j(s+h)-\delta_j(s), \delta_k(s+h)-\delta_k(s)]$ soit uniquement fonction de h, le vecteur de translation entre les points s et $s+h$, pour tout couple $j, k \in \{0, ..., q\}$ et toute paire de points $s, s+h \in D$. Ces fonctions de covariance, notées $2\gamma_{jk}(h)$, sont nommés variogrammes simples ($j = k$) et croisés ($j \neq k$).

La structure de dépendance spatiale de l'ensemble des fonctions aléatoires résiduelles est donc représentée par les $C_{jk}(h)$ ou par les $\gamma_{jk}(h)$. Cressie (1993, p.66) note qu'il existe en fait deux généralisations du semi-variogramme au cas multivariable. Au lieu de la fonction $\gamma_{jk}(\cdot)$ introduite précédemment, certains utilisent la fonction $\pi_{jk}(h) = \frac{1}{2}Var(\delta_j(s+h) - \delta_k(s))$, que Wackernagel (2003, p.149) nomme pseudo semi-variogramme croisé. L'analyse variographique précédant un cokrigeage doit inclure l'estimation et la modélisation des covariogrammes, ou des (pseudo) semi-variogrammes, pour tout $j, k \in \{0, ..., q\}$. Cette tâche s'avère souvent complexe. Comme dans le cas univariable, seuls les modèles assurant la positivité de la variance de toute combinaison linéaire de variables aléatoires sont admissibles. Les « modèles linéaires de corégionalisation » garantissent cette positivité. Ils sont donc fréquemment utilisés dans la pratique du cokrigeage. Le lecteur est référé à Goovaerts (1997, p.109), Wackernagel (2003, p.145) ou Arnaud et Emery (2000, p.157) pour plus d'informations sur ces modèles ou sur tout autre aspect de l'analyse variographique menant au cokrigeage.

Comme pour le krigeage, les versions simple, ordinaire et avec modèle de tendance du cokrigeage ont été développées. Le type de cokrigeage dépend de la forme des fonctions $\mu_j(\cdot)$ pour $j \in \{0, ..., q\}$ de la même façon qu'en krigeage. La résolution des équations de cokrigeage ne sera pas présentée dans ce mémoire. Hoef et Cressie (1993) ainsi que Matheron (1970, p.187) le font pour le cokrigeage avec modèle de tendance. Goovaerts (1997, p.204) soulève cependant que ce type de cokrigeage est très difficile à implanter, particulièrement en raison du biais dans l'estimation des covariogrammes ou des semi-variogrammes sur les résidus. Ce problème, traité dans la section 4.4 pour le krigeage, devient encore plus sévère dans le cas multivariable. L'utilisation du cokrigeage

simple ou ordinaire est donc plus recommandé. Goovaerts (1997, p.203), Wackernagel (2003, p.159) ainsi que Arnaud et Emery (2000, p.182) présentent ces types de cokrigeage. Notons que plusieurs variantes de formules de prévision peuvent être obtenues en cokrigeage ordinaire dépendamment des contraintes de non-biais choisies (Isaaks et Srivastava, 1989, p.403).

En plus d'être la seule méthode géostatistique multivariable à tenir compte du comportement spatial des variables régionalisées auxiliaires, le cokrigeage est la seule méthode qui n'exige pas une connaissance exhaustive de ces variables auxiliaires. En effet, une technique telle que le krigeage avec dérive externe a besoin, pour chaque variable régionalisée auxiliaire, de données aux points d'observation s_1 à s_n de la variable régionalisée à interpoler et en tous les points pour lesquels une prévision est désirée. Le cokrigeage n'a pas de telles contraintes. Les sites d'observation de chaque variable régionalisée auxiliaire peuvent être localisés n'importe où sur le champ D. Cependant, le cokrigeage a plus de chance de bien performer comparativement au krigeage s'il y a plus d'observations des variables régionalisées auxiliaires que de la variable régionalisée à interpoler et que ces observations ne sont pas toutes mesurées aux mêmes sites. De plus, dans le cas où les variables auxiliaires son connues exhaustivement, le cokrigeage semble interpoler mieux que le krigeage avec dérive externe seulement lorsque les corrélations entre les variables auxiliaires et la variable à interpoler sont supérieures à 0.75 (Asli et Marcotte, 1995).

4.5.4 Autres types de krigeage

En plus du krigeage simple, du krigeage ordinaire et du krigeage avec modèle de tendance, plusieurs autres types de krigeage ont été développés dans la littérature. Ils ont comme motivation l'amélioration des faiblesses du krigeage classique. Certaines de ces méthodes sont mentionnées ici.

Méthodes résistantes aux données extrêmes

Le krigeage classique étant assez sensible aux données extrêmes, des versions robustes du krigeage ordinaire et du krigeage universel ont été développées.

Krigeage robuste : Proposé par Hawkins et Cressie (1984), le krigeage robuste représente une solution de rechange au krigeage ordinaire. L'idée derrière cette méthode

est de diminuer le poids des données extrêmes attribuables à une distribution non gaussienne des données. Cressie (1993, p.144) présente cette technique.

Krigeage « median polish » : Le krigeage universel possède aussi sa version robuste : le krigeage « median polish » Cressie (1984, 1986). Cette méthode est spécialement adaptée aux données se présentant sur une grille dans le plan, qui sont alors vues comme un tableau de fréquences.

Géostatistique non stationnaire

Le krigeage avec modèle de tendance n'est pas le seul à entrer dans la catégorie non stationnaire. Des généralisations de ce krigeage ont fait place à d'autres techniques.

Krigeage bayésien : Une modélisation autre que celle du krigeage avec modèle de tendance est de considérer la fonction de variation à grande échelle $\mu(\cdot)$ aléatoire plutôt que déterministe. Les paramètres de cette fonction deviennent alors des variables aléatoires qui sont associées à des loi à priori. C'est l'idée de base du krigeage bayésien, proposé par Omre (1987) afin de permettre l'introduction de connaissances à priori dans la prévision par krigeage.

Krigeage intrinsèque généralisé : Le biais dans l'estimation du semi-variogramme en krigeage avec modèle de tendance a mené Matheron (1973) à proposer le krigeage intrinsèque généralisé. Chauvet (1999, p.177) présente cette méthode, qui fait intervenir des fonctions aléatoires plus complexes que celles stationnaires de second ordre ou intrinsèque ainsi qu'un outil structural autre que le semi-variogramme appelé covariance généralisée.

Géostatistique non linéaire

Tel qu'expliqué à la section 4.5.1, la linéarité des prévisions par krigeage implique que celui-ci accomplisse parfois de mauvaises performances lorsque les données sont loin d'être normales. La géostatistique non linéaire (Cressie, 1993, p.278) propose des solutions à ce problème.

Krigeage d'indicatrices : Le krigeage d'indicatrices, suggéré par Journel (1983), permet d'estimer la probabilité que la fonction aléatoire $Z(\cdot)$ dépasse un certain seuil, et ce pour plusieurs seuils prédéterminés. Ainsi, une estimation de la fonction de répartition de la variable aléatoire à interpolée $Z(s_0)$ est obtenue. L'espérance

de la variable aléatoire peut ensuite être estimée à partir de cette distribution empirique.

Krigeage disjonctif : Le krigeage disjonctif cherche à trouver la prévision optimale de la forme $\sum_{i=1}^{n_0} f_i(Z(s_i))$, soit une combinaison linéaire de fonctions des données plutôt que des données elles-mêmes. Pour ce faire, une connaissance de la distribution bivariée des données est requise. Ce type de krigeage, d'un niveau mathématique élevé, est dû à Matheron (1976). Rivoirard (1994) explicite la méthode, qui est très exigeante en terme de conditions de stationnarité et de ressources informatiques.

4.6 Conclusion du chapitre

Ainsi, le krigeage consiste simplement en une prévision linéaire sans biais à variance minimale. La résolution d'un système d'équations conforme aux hypothèses du modèle permet d'obtenir une prévision par krigeage. Cependant, cette prévision a tendance a moins bien performer dans certaines circonstances telles la non normalité des données ou lorsque les données sont peu nombreuses.

Seule l'interpolation spatiale ponctuelle a été traitée dans ce chapitre. Notons que la théorie pour effectuer par krigeage une prévision sur une sous-région du champ D est développée dans la littérature. Il s'agit du krigeage par bloc (Goovaerts, 1997, p.152).

Passons maintenant de la théorie à la pratique. Une méthodologie de mise en oeuvre du krigeage est explicitée et illustrée dans le prochain chapitre.

Chapitre 5

Mise en oeuvre du krigeage

Maintenant que le krigeage a été défini, penchons-nous sur la question : comment effecter un krigeage en pratique ? Ce chapitre présente une démarche d'utilisation du krigeage et aborde la question du support informatique à employer. Ensuite, un jeu de données à interpoler est décrit et une partie de ce jeu de données est traitée afin d'illustrer la démarche proposée.

5.1 Méthodologie géostatistique

Tel que l'indique Cressie (1993, p.289), la mise en oeuvre du krigeage s'effectue en suivant certaines étapes. La première est l'analyse exploratoire qui permet de se familiariser avec les données et d'éclairer le choix du modèle. Ensuite, le modèle peut être énoncé. Cette deuxième étape inclue le choix de la forme de la tendance déterministe pour l'espérance de $Z(\cdot)$, l'analyse variographique, puis une validation croisée si désirée. Cette dernière sous-étape permet de comparer la performance de différents modèles afin de sélectionner celui susceptible de mener aux meilleures prévisions. Finalement, l'interpolation est effectuée par krigeage. La figure 5.1 schématise cette démarche. La flèche courbe reliant la validation croisée à la détermination de $\mu(\cdot)$ illustre le fait que les étapes de la formulation du modèle peuvent être reprises afin de comparer plusieurs modèles. Chacune des étapes sont vues plus en détails dans les paragraphes suivants.

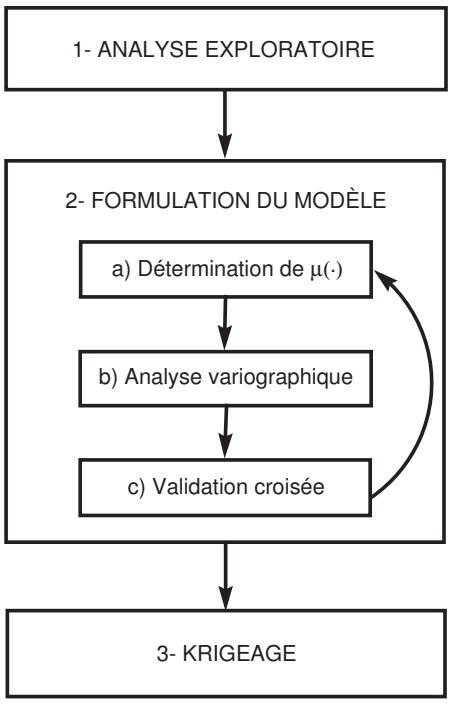

FIG. 5.1 – *Méthodologie géostatistique*

5.1.1 Analyse exploratoire

Comme dans toute analyse statistique, il est bon de débuter nos travaux par une analyse exploratoire des données. Isaaks et Srivastava soulignent l'importance de cette étape en lui consacrant pratiquement le tiers de leur livre *Applied Geostatistics* (1989, p.10 à p.183). Tel que le font Isaaks et Srivastava, cette étape préliminaire sera présentée ici en la divisant en trois volets : un volet univarié, un volet bivarié et un volet spatial.

Description univariée :

L'analyse exploratoire vise à donner une idée de la distribution des données. Rappelons qu'en géostatistique, les données sont considérées comme une réalisation partielle d'une fonction aléatoire. Ainsi, chaque donnée serait une observation d'une variable aléatoire différente et chaque variable aléatoire aurait sa propre distribution. L'échantillon de données serait donc multivarié. Toutefois, l'analyse exploratoire est habituellement débutée en oubliant cette hypothèse de base et en considérant les données comme plusieurs réalisations d'une même variable aléatoire. Des statistiques descriptives univariées telles la moyenne et la variance expérimentale ainsi que des outils graphiques

tels l'histogramme et le diagramme en boîte peuvent alors être utilisés afin de se familiariser avec la distribution de la variable aléatoire. Si cette distribution ne semble vraiment pas être gausienne, une transformation de variable peut être envisagée. La transformation la plus adéquate sera donc recherchée. Cette tâche peut être accomplie par une méthode d'essais/erreurs ou encore en utilisant des techniques proposées par Box et Cox (1964). Si une transformation est choisie, le reste de l'analyse exploratoire devrait être effectuée sur les données transformées.

Description bivariée :

Lorsqu'une variable régionalisée auxiliaire est disponible, il est conseillé de l'étudier avec les outils énumérés dans le paragraphe précédent, mais il est particulièrement important d'examiner sa relation avec la variable régionalisée d'intérêt. Pour ce faire, il est aussi supposé que cette variable régionalisée auxiliaire est une réalisation d'une variable aléatoire plutôt que d'une fonction aléatoire. Dans ce cas, le lien entre la variable aléatoire auxiliaire et la variable aléatoire à l'étude peut être décrit par un diagramme de dispersion, par des mesures d'association entre deux variables telles une corrélation, ainsi que par régression simple classique. Dans le cas où plusieurs variables régionalisées auxiliaires sont disponibles, des corrélations conditionnelles et la régression multiple peuvent être employées.

Toutes ces techniques requièrent des couples de données aux points d'observation s_1 à s_n. Si certaines variables régionalisées auxiliaires n'ont pas été observées en ces points, une interpolation spatiale est préalablement effectuée afin de prévoir ces valeurs régionalisées. Une simple technique déterministe, telle une méthode barycentrique, est habituellement employée pour effectuer cette tâche. Les données interpolées obtenues seront ensuite nécessaires à la résolution des équations du krigeage si le krigeage avec dérive externe est choisi.

Description spatiale :

En se rappelant maintenant que les observations sont en fait multivariées, les variables régionalisées peuvent être explorées spatialement. Cette analyse descriptive est particulièrement importante pour la variable régionalisée à interpoler. Les données peuvent d'abord être visualisées à l'aide de graphiques 3D ou de cartes de courbes de niveaux. Ces outils graphiques aident notamment à juger si l'espérance de la fonction aléatoire modélisant la variable régionalisée peut être vue comme une fonction

constante. La présence de valeurs fortement différentes les unes des autres indiquerait que ce n'est pas le cas. Ainsi, la stationnarité du premier moment est étudiée par ces graphiques.

Cette stationnarité peut aussi être examinée à l'aide d'une fenêtre mobile s'il y a assez de données (Isaaks et Srivastava, 1989, p.46). Cette fenêtre mobile est simplement un outil qui permet de délimiter des sous-régions du champ qui peuvent se chevaucher. Pour chacune de ces sous-régions, des statistiques descriptives de base sont calculées, telles la moyenne et la variance expérimentale. Ainsi, la stationnarité du deuxième moment peut aussi être étudiée avec une fenêtre mobile. Par exemple, une variance expérimentale qui augmenterait toujours en agrandissant la fenêtre suggérerait une variance infinie, donc une hypothèse de stationnarité intrinsèque serait plus appropriée qu'une hypothèse de stationnarité de deuxième d'ordre.

En outre, le comportement directionnel d'une variable régionalisée s'illustre par un graphique de la variable régionalisée en fonction des coordonnées le long d'une direction (Arnaud et Emery, 2000, p.115). Le plus souvent, deux de ces graphiques sont faits, un pour les coordonnées en x et l'autre pour les coordonnées en y. Si les points d'un de ces graphiques montrent une tendance autre que linéaire de pente nulle, alors l'espérance de la fonction aléatoire à interpoler n'est certainement pas stationnaire.

La structure de dépendance spatiale que présente les données sera étudiée en profondeur à l'étape de l'analyse variographique. Elle peut aussi être explorée préliminairement à l'aide de *h-scatterplots* (Isaaks et Srivastava, 1989, p.52 ; Arnaud et Emery, 2000, p.117), nommés *nuages de corrélation différée* en français. Il s'agit de nuages de points formés par les couples $(z(s_i), z(s_i + h))$. Ils indiquent, pour un vecteur de translation donné h, si les valeurs prises aux sites séparés par h sont semblables. Lorsque les sites d'observation ne sont pas répartis de façon régulière sur le territoire et que la dépendance spatiale est supposée isotrope, ces graphiques de dispersion peuvent être faits pour les couples $(z(s_i), z(s_j))$ avec $|s_i - s_j|$ approximativement égale à certaines distances prédéterminées.

5.1.2 Formulation du modèle

Une bonne exploration des données est essentielle au choix du modèle. En se basant sur les résultats de cette analyse descriptive, le type de krigeage à employer peut être choisi judicieusement et l'analyse variographique menée de façon plus éclairée. Si plusieurs modèles sont envisagés, la validation croisée peut être utilisée afin d'en sélectionner un. Mais avant tout, il faut déterminer si l'interpolation sera faite directe-

ment sur la fonction aléatoire d'intérêt ou sur une transformation de celle-ci. Le plus simple est bien sûr de travailler sur les données non transformées. Cette option est toujours privilégiée. Cependant, un utilisateur peut choisir de travailler sur une transformation des données pour améliorer la qualité de l'interpolation dans le cas d'un important écart à la normalité. Dans ce cas, il doit toutefois s'assurer d'être en mesure d'effectuer la transformation inverse après le krigeage.

a) Détermination de $\mu(\cdot)$:

Le choix de la forme de $\mu(\cdot)$ implique l'adoption d'un type de krigeage (simple, ordinaire, universel ou avec dérive externe) ainsi que l'écriture de la dérive lors de la sélection d'un krigeage avec modèle de tendance. Afin d'accomplir cette tâche, il faut d'abord déterminer si des variables régionalisées auxiliaires seront ajoutées dans la dérive du modèle. Si oui, un krigeage avec dérive externe sera effectué. Il semble potentiellement profitable d'employer ce type de krigeage seulement si des liens assez forts sont trouvés lors de la description bivariée (voir section 4.5.3). Dans ce cas, il faut définir le modèle de tendance, c'est-à-dire sélectionner les variables auxiliaires à inclure dans le modèle et déterminer les fonctions $f_i(\cdot)$ pour $i = 0, ..., p$. Ces choix se baseront sur les analyses de régression faites lors de l'exploration des données.

Dans le cas où aucune variable régionalisée auxiliaire n'est disponible, ou que celles-ci ne sont pas assez reliées à la variable régionalisée d'intérêt, un krigeage univariable est sélectionné. Si l'analyse descriptive a clairement mis en évidence une non stationnarité de l'espérance de la fonction aléatoire à interpoler, le krigeage universel est choisi. Les fonctions $f_i(\cdot)$ composant le modèle de tendance sont alors déterminées à l'aide des graphiques de comportement directionnel. Par contre, lorsque la stationnarité du premier moment peut être acceptée, un krigeage stationnaire est sélectionné. Si l'espérance de la fonction aléatoire à interpoler est connue, le krigeage simple est employé, sinon le krigeage ordinaire est utilisé.

b) Analyse variographique :

Pour compléter la formulation du modèle, l'analyse variographique est menée afin d'estimer le semi-variogramme de la fonction aléatoire résiduelle. L'analyse variographique est réalisée après la détermination de $\mu(\cdot)$ afin de savoir si elle se basera sur les observations $z(s_i)$ ou sur les résidus $e(s_i) = z(s_i) - \hat{\mu}(s_i)$. Ainsi, si un krigeage avec modèle de tendance a été choisi, il faut en premier lieu obtenir les résidus $e(s_i)$. Nous

proposons d'effectuer cette tâche en suivant la démarche itérative présentée à la section 4.4. Ensuite, le semi-variogramme peut être estimé puis modélisé.

Estimation du semi-variogramme : L'utilisateur doit d'abord choisir l'estimateur qu'il emploiera, la distance maximale d'estimation et la largeur de la fenêtre dans le cas de données réparties irrégulièrement. Ces choix n'auront pas un très grand impact sur l'interpolation finale. L'estimateur le plus commun est celui des moments. Un estimateur robuste pourrait être envisagé s'il y a des données extrêmes. En outre, il est conseillé de calculer le semi-variogramme expérimental seulement pour des distances plus petites que la demie de la distance maximale entre deux points d'observation. Finalement, la largeur de la fenêtre est choisie en faisant un compromis entre la quantité et la qualité des points composant le semi-variogramme expérimental. En effet, plus la largeur de la fenêtre est petite, plus le semi-variogramme expérimental comporte de points, mais plus ces points sont imprécis car ils sont calculés à partir de peu de données. Idéalement, il faut s'assurer que chaque point provienne d'un nombre suffisant de données. En conséquence, l'analyse variographique est difficile et imprécise lorsqu'il y a peu de données.

Modélisation du semi-variogramme : L'étape de l'analyse variographique qui influence le plus le résultat final de l'interpolation est la sélection du modèle variographique. Ce choix doit idéalement tenir compte des analyses variographiques antérieures effectuées sur des variables régionalisées modélisant le même phénomène naturel, de l'analyse exploratoire et de l'allure du semi-variogramme expérimental. Par exemple, une apparente discontinuité à l'origine du semi-variogramme expérimental ou une représentation 3D de la variable régionalisée présentant un comportement erratique suggère l'ajout d'un effet de pépite dans le modèle. En outre, si l'analyse descriptive à l'aide de fenêtre mobile a indiqué que la variance de la fonction aléatoire est infinie, les modèles sans palier sont favorisés. Finalement, les paramètres du modèle variographique choisi doivent être ajustés au semi-variogramme empirique. Si l'utilisateur désire automatiser cet ajustement, il doit sélectionner une procédure d'estimation. Les différentes procédures devraient mener à des estimations assez semblables. Il est bon de toujours vérifier après coups le bon sens des paramètres estimés obtenus. Notamment, les estimations d'un effet de pépite, d'une portée ou d'un palier doivent toujours être positives.

c) Validation croisée :

La validation croisée de type « leave-one-out » est fréquemment utilisée dans la pratique géostatistique afin de comparer la qualité des prévisions provenant de divers

modèles (Isaaks et Srivastava, 1989, p.351 ; Arnaud et Emery, 2000, p.152 ; Cressie, 1993, p.101 ; Wackernagel, 2003, p.87). Cette technique peut être utilisée non seulement pour comparer les modèles, mais pour en choisir un, c'est-à-dire choisir le type de krigeage et le modèle variographique employé. La validation croisée de type « leave-one-out » consiste à retirer une à une les observations pour ensuite les prévoir par krigeage à partir des autres données. La prévision par validation croisée d'une observation $Z(s_i)$ sera notée $\hat{Z}_{-i}(s_i)$. Des erreurs de validation croisée sont calculées en soustrayant la valeur observée à la la valeur estimée : $e_{-i}(s_i) = \hat{Z}_{-i}(s_i) - Z(s_i)$. À partir de ces erreurs, un ou des indices sont calculés (voir Myers, 1993, pour une liste d'indices possibles). Souvent, l'indice de la moyenne des erreurs au carré $\frac{1}{n}\sum_{i=1}^{n} e_{-i}(s_i)^2$, nommé EQM (Erreur Quadratique Moyenne), est employé. Les avantages de cet indice sont qu'il incorpore le biais et la variance de la distribution des erreurs et que l'absence de division par l'écart type de krigeage évite la dépendance des résultats par rapport au type de krigeage effectué. Un indice équivalent mais plus robuste est la moyenne ou la médiane des erreurs en valeur absolue. Cet indice peut donc être préféré à l'EQM lors de la présence de valeurs extrêmes. Un indice tel que l'EQM est vu comme une erreur de prévision. On cherche donc à le minimiser. Ainsi, les étapes a) à c) de la formulation du modèle sont effectuées pour plusieurs modèles et ceux obtenant les plus petits EQM sont jugés meilleurs. Une façon d'automatiser le choix du modèle est donc de retenir celui qui minimise l'EQM.

Plusieurs auteurs dénoncent l'utilisation de la validation croisée pour sélectionner un modèle (e.g. Isaaks et Srivastava, 1989, p.514 ; Cressie, 1993, p.104 ; Goovaerts, 1997, p.105). Ils soutiennent qu'un modèle devrait être choisi en se basant sur l'expérience de l'utilisateur et sur les informations disponibles concernant le phénomène régionalisé étudié. Ils n'aiment donc pas l'aspect « boîte noire » de la sélection d'un modèle par validation croisée. Cependant, le caractère automatique de cette technique est parfois requis pour certaines applications, telle que celle traitée au chapitre 6. Davis (1987) propose une très intéressante discussion sur l'utilisation de la validation croisée en géostatistique. Il ne rejette pas son utilisation pour automatiser la sélection d'un modèle, mais il rappelle que le modèle choisi est simplement le meilleur parmi les modèles envisagés et selon l'indice de validation croisée utilisé. Il n'est donc pas optimal, c'est-à-dire le meilleur parmi tous les modèles possibles. Notons aussi que cette technique requiert de grands temps de calculs, car pour chaque modèle, les paramètres du semi-variogramme doivent être ajustés et une prévision par krigeage doit être effectuée en chaque point d'observation. Ainsi, pour de très grands jeux de données, la sélection de modèle par validation croisée n'est pas envisageable. Marcotte (1995) pousse encore plus loin l'utilisation de la validation croisée dans la formulation du modèle. Il propose une approche d'estimation des paramètres du modèle variographique basée sur la validation croisée. Dans ce mémoire, cette technique n'est pas approfondie. Nous proposons plutôt

d'estimer les paramètres du semi-variogramme par une méthode des moindres carrés (voir section 3.5).

En fait, lorsqu'un grand nombre de données sont disponibles, la validation croisée n'est pas nécessaire. Dans ce cas, le jeu de données peut être divisé en un jeu de données pour l'interpolation et un jeu de données de test. La variable régionalisée est ensuite interpolée aux sites d'observation des données de test à partir des données du premier jeu de données. Des erreurs de prévision sont obtenues en ces sites et des indices de validation peuvent encore être calculés. Cette technique est théoriquement mieux que la validation croisée car les données utilisées pour la validation sont indépendantes des données servant à effectuer l'interpolation.

5.1.3 Krigeage

Après avoir sélectionné un modèle, l'interpolation peut être effectuée. Cette étape en est une de calculs seulement. Les prévisions et les variances de krigeage sont évaluées pour tous les points désirés. Pour ce faire, il suffit de d'abord calculer $\boldsymbol{\Gamma}$ et $\boldsymbol{\gamma}_0$, ou $\boldsymbol{\Sigma}$ et $\boldsymbol{c}_0$, à partir du modèle variographique, puis de remplacer ces valeurs ainsi que les observations de la variable régionalisée d'intérêt et des variables régionalisées auxiliaires, s'il y a lieu, dans les expressions trouvées au chapitre 4 pour le type de krigeage choisi.

5.2 Logiciels informatiques

La pratique du krigeage n'est pas envisageable sans ordinateur. Certains choisissent de programmer la méthode, car les calculs intervenant dans la démarche sont relativement simples. Cependant, une bonne analyse variographique requiert un support graphique assez puissant. Il est donc certainement plus simple d'utiliser un logiciel déjà sur le marché. Afin de trouver un logiciel géostatistique adapté à ses besoins, le site web `http://www.ai-geostats.org/` de la Commission européenne est une excellente référence. Une grande quantité de logiciels y sont listés, décrits et comparés. Le lecteur cherchant de l'information sur les différents logiciels géostatistiques disponibles est donc référé à ce site web.

Un des logiciels les plus connus en géostatistique est GSLIB de Deutsch et Journel (1998). Le code source de ce logiciel peut être téléchargé gratuitement sur le site web `http://www.gslib.com/`. Cependant, le livre *GSLIB : Geostatistical Software Li-*

brary and User's Guide est indispensable pour son utilisation. Ce logiciel permet d'effectuer plusieurs types de krigeage. Il est fréquemment utilisé par les chercheurs en géostatistique. Par contre, dans ce mémoire, les calculs géostatistiques ont plutôt été effectués avec le logiciel S-Plus (Insightful, 2004) bien connu des statisticiens. Ce logiciel propose un module géostatistique, S+SpatialStats, qui se base sur le livre de Cressie (1993). Ce module s'utilise facilement, mais il est assez limité, ajustant uniquement des modèles variographiques bornés et effectuant seulement des prévisions par krigeage ordinaire ou universel. Le logiciel Gstat (Pebesma et Wesseling, 1998), offert en librairie S-Plus (Pebesma, 2004a), est beaucoup plus complet. Il propose plusieurs modèles variographiques bornés ou non et effectue du krigeage ou du cokrigeage simple, ordinaire ou avec modèle de tendance. Il permet aussi de faire des simulations géostatistiques et de la validation croisée. La librairie Gstat se télécharge gratuitement à partir du site web `http://www.gstat.org`. Elle est aussi offerte en extension pour le logiciel R (R project, 2004), qui peut être vu comme une version gratuite de S-Plus. De plus, la documentation très complète de Gstat (Pebesma, 2004b, 1999) rend facile son utilisation même si celle-ci s'effectue par la saisie de commandes. Gstat est donc un très bon logiciel géostatistique accessible à tous. C'est ce logiciel, sous sa forme de librairie S-Plus, qui a été utilisé pour effectuer les analyses de la section suivante et du chapitre 6. Ce mémoire contient d'ailleurs à l'annexe B le programme S-Plus qui a permis d'effectuer les interpolations au chapitre 6.

5.3 Application de l'interpolation spatiale : présentation des données

À ce point-ci du mémoire, assez d'informations ont été fournies pour savoir comment effectuer un krigeage. Le chapitre suivant traitera de l'utilisation du krigeage afin de résoudre une problématique d'interpolation spatiale. Celle-ci concerne la préparation d'intrants pour des modèles de simulation hydrologique. Ces modèles, qui produisent des prévisions de débit des cours d'eau, basent leurs simulations sur les quantités de pluie tombées. Ces données observées doivent être fournies aux modèles sur une certaine grille. Par contre, la principale source de données de précipitations, soit les stations météorologiques, fournie des données ponctuelles irrégulièrement réparties sur le territoire. Une interpolation spatiale doit donc être effectuée afin de d'estimer les précipitations en chaque noeud de la grille.

Les données traitées au chapitre 6 ont été fournies par Hydro-Québec et Environnement Canada. Il s'agit de mesures de quantités de pluie tombées provenant de 16 sta-

tions météorologiques couvrant le bassin versant de la rivière Gatineau dans la région de l'Outaouais au Québec. La carte des bassins versants prioritaires du Québec[1] (figure 5.2) permet de situer ce bassin versant dans la province. La période de test utilisée est le mois d'août 2003. Les données représentent toutes des précipitations cumulées sur une période de 6 heures. Ainsi, le jeu de données comprend 124 observations par stations pour les 124 périodes de 6 heures du mois d'août 2003. Pour chacune de ces périodes, une interpolation spatiale est requise. En termes géostatistiques, la variable régionalisée à interpoler par période de 6 heures est donc la pluie tombée au sol, le champ d'étude est le bassin versant de la Gatineau et 16 valeurs régionalisées ont été mesurées. Des observations d'une variable régionalisée auxiliaire sont aussi disponible. Il s'agit des prévisions de pluie pour les mêmes périodes de temps fournies par un modèle atmosphérique d'Environnement Canada nommé GEM. Pour cette variable régionalisée auxiliaire, 234 valeurs régionalisées réparties sur une grille aux 10 km couvrant le bassin versant de la Gatineau sont disponibles pour chaque période. L'interpolation spatiale sera d'ailleurs effectuée en chacun des 234 points de cette grille. Cette section a pour but de décrire les deux sources des données qui seront analysées au chapitre 6.

Notons que dans ce mémoire, la localisation d'un site est toujours exprimée en coordonnées UTM (Tranverse de Mercator Universelle) plutôt qu'en latitude-longitude. Les coordonnées latitude-longitude sont des mesures d'angle se basant sur la forme sphérique de la Terre, tandis que les coordonnées UTM sont obtenues en projetant la surface sphérique du globe sur un plan. En conséquence, l'emploi d'une distance euclidienne est correct pour des coordonnées UTM, mais n'est pas du tout recommandé pour des coordonnées latitude-longitude.

5.3.1 Stations météorologiques

Les précipitations observées proviennent des 16 stations localisées sur la figure 5.3. Bien que certaines de ces stations se retrouvent légèrement à l'extérieur du champ d'étude du bassin de la rivière Gatineau, elles sont incluses dans l'interpolation pour augmenter le nombre de stations et parce que leur proximité par rapport au champ d'étude signifie qu'elles permettront très probablement d'améliorer la qualité de l'interpolation.

À chaque station météorologique, les précipitations sont mesurées par un précipitomètre. Cet appareil de mesure n'est bien sûr pas parfait. Les précipitomètres qui ont

[1] Source : Site internet du Ministère de l'environnement du Québec
http://www.menv.gouv.qc.ca/eau/bassinversant/bassins/carte.htm

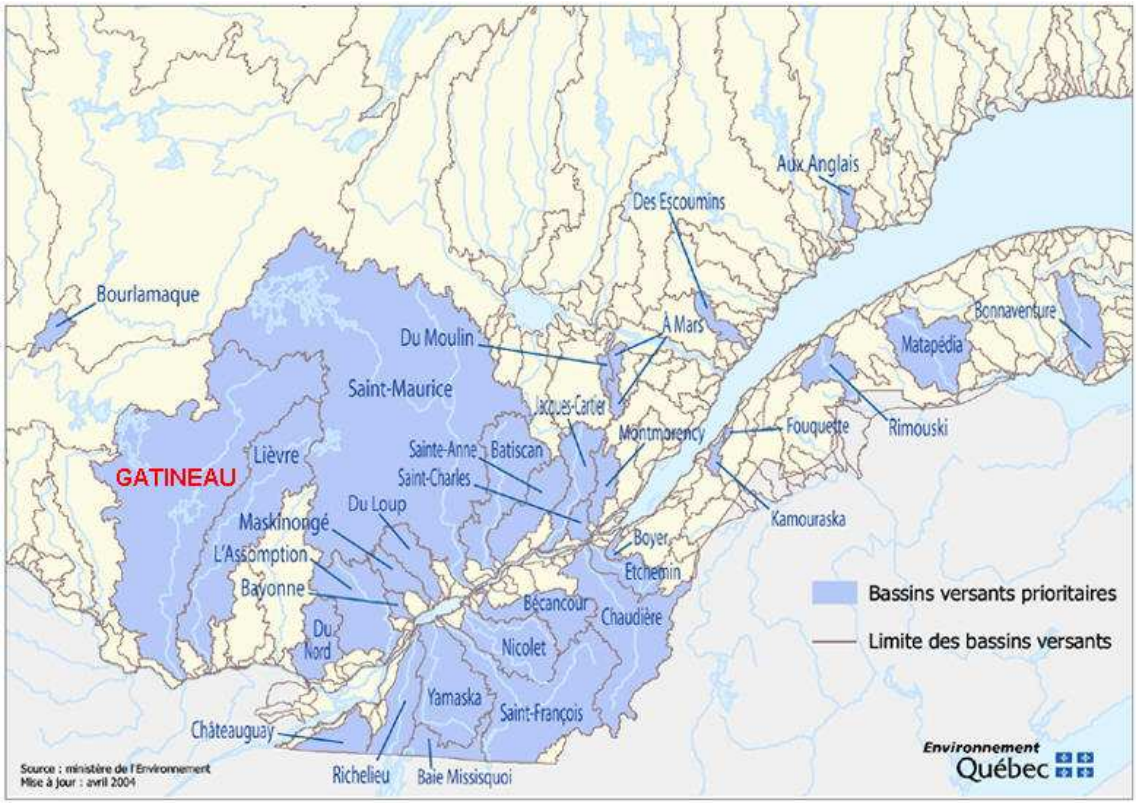

FIG. 5.2 – *Carte des bassins versants prioritaires du Québec*

fourni les observations analysées dans ce mémoire ont généralement une précision de lecture de cumul de précipitations égale à 0.2 mm. Seuls 4 précipitomètre font exception, ceux des stations Mercier Amont, Barrière Amont, Cabonga Amont et Dozois Est, avec une précision de lecture de cumul de 3.75 mm. Ces précipitomètres sont donc beaucoup moins précis que les autres. De plus, Yang *et al.* (1999) ont mis en évidence que le vent cause une sous-estimation des précipitations par les précipitomètres. Cette sous-estimation, qui augmente en fonction de la force des vents, est évidemment plus importante pour des précipitomètres très exposés aux vents. Ainsi, la qualité des observations dépend notamment du type de précipitomètre, de son exposition, de son entretien et des conditions de vent et de température au moment de la mesure. Malheureusement, cette variabilité dans la qualité des observations n'est pas prise en compte lors des interpolations spatiales effectuées dans ce mémoire. L'emploi de techniques de krigeage avec erreurs de mesures permettrait par contre de le faire (Cressie, 1993, p.128).

Notons que pour certaines périodes de 6 heures, au moins une des 16 stations n'a pas rapporté d'observations pour une raison quelconque. En fait, jusqu'à 4 données sur

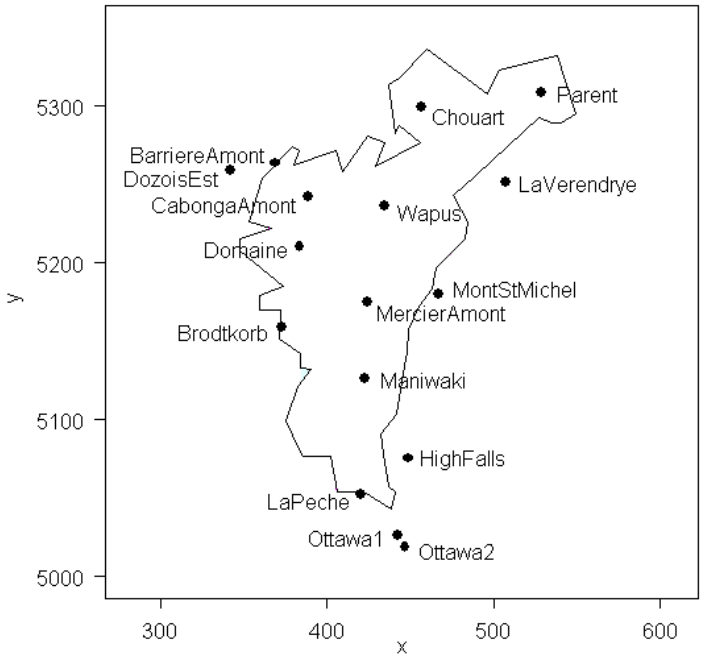

FIG. 5.3 – *Localisation des 16 stations météorologiques par rapport aux limites du bassin versant de la rivière Gatineau*

16 sont manquantes pour quelques périodes. Les interpolations sont donc effectuées à partir de 12 à 16 mesures de la variable régionalisée à l'étude. Webster et Oliver (1993) affirment que 150 données sont nécessaires à une estimation vraiment fiable du semi-variogramme et plus encore pour décrire l'anisotropie. Pour les données utilisées ici, il est donc impensable d'étudier l'anisotropie de la variable régionalisée et les estimations de semi-variogrammes sont peu précises. En conséquence, il est difficile de modéliser visuellement les semi-variogrammes expérimentaux et l'emploi de la procédure de sélection automatique du modèle variographique par validation croisée est justifiable.

5.3.2 Modèle atmosphérique GEM

La source de données auxiliaires aux stations météorologiques est le modèle numérique de prévision météorologique GEM (Global Environnemental Multi-échelles) (Côté *et al.*, 1998). Ce modèle atmosphérique est utilisé à Environnement Canada pour effectuer des prévisions météorologiques à court, moyen et long terme. Les données utilisées ici proviennent du modèle GEM régional de prévision à court terme. La version de ce modèle employée pour générer les données analysées ici travaille sur une grille de

24 km par 24 km. Pour obtenir les données sur la grille de 10 km par 10 km mentionnée à la section 5.4, Environnement Canada a interpolé les sorties du modèle.

Le modèle GEM utilise comme système de référence temporel le temps UTC (Temps Universel Coordonné). Le temps UTC est la base légale de l'heure dans le monde. Il s'agit en fait de l'heure du méridien d'origine de Greenwich. Ce système de référence facilite le travail sur de grands territoires traversant plusieurs fuseaux horaire. Dans la vie de tous les jours au Québec, c'est l'heure normale de l'est (HNE) en hiver ou l'heure avancée de l'est (HAE) en été qui est utilisé. Pour se ramener à l'heure HNE à partir de l'heure UTC il suffit de soustraire 5 heures et pour se rammener à l'heure HAE il faut soustraire 4 heures.

Le modèle GEM fournit des prévisions deux fois par jour : à 00 UTC et à 12 UTC. Les prévisions du modèle GEM régional concernent les 48 prochaines heures. Ces 48 heures sont divisées en 8 périodes de 6 heures. Les variables météorologiques prévues par GEM sont nombreuses. Une seule de ces variable est utilisée ici : les précipitations cumulées sur des périodes de 6 heures. Étant donné que le modèle GEM fournit à toutes les 12 heures des prévisions pour un intervalle de temps de 48 heures, des prévisions GEM se chevauchent dans le temps. Pour chaque période de 6 heures, 4 prévisions GEM sont disponibles. Pour illustrer cet énoncé, considérons la période entre 18 UTC le 4 août 2003 et 00 UTC le 5 août 2003. Cette période est localisée par le rectangle hachuré sur la figure 5.4. Quatre prévisions GEM concernent cette période : la prévision émise à 00 UTC le 3 août 2003 pour la huitième période de 6 heures plus tard (P8), la prévision émise à 12 UTC le 3 août 2003 pour la sixième période de 6 heures plus tard (P6), la prévision émise à 00 UTC le 4 août 2003 pour la quatrième période de 6 heures plus tard (P4), la prévision émise à 00 UTC le 4 août 2003 pour la deuxième période de 6 heures plus tard (P2). Ainsi, un choix doit être fait parmi ces prévisions. Des spécialistes en prévision numérique du temps avec le modèle GEM ont jugé que les prévisions les plus exactes étaient celles effectuées pour la deuxième et la troisième période de 6 heures suivant le moment de l'émission des prévisions. Les prévisions exploitées dans ce mémoire sont donc celles associées aux intervalles P2 et P3 de chaque émission de prévisions, soit les segments bleus de la figure 5.4. Un météorologue dirait donc que les données auxiliaires employées dans le chapitre 6 sont les prévisions de précipitations du modèle GEM régional 24 km 6h-18h ramenées à 10 km. Des informations sur le fonctionnement du modèle atmosphérique GEM peuvent être trouvées sur le site internet du Service météorologique du Canada (SMC, 2002), une division d'Environnement Canada.

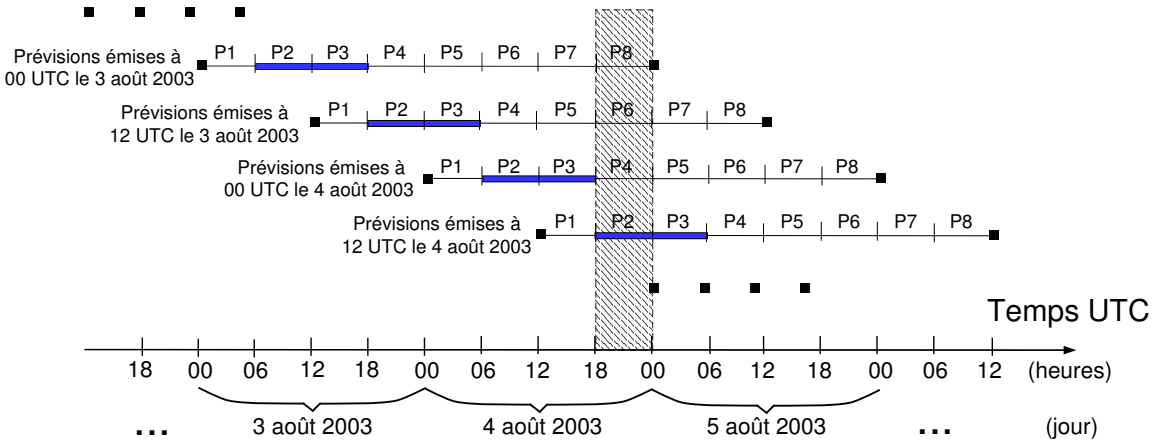

FIG. 5.4 – *Schéma temporel de l'émission de prévisions par le modèle atmosphérique GEM*

5.4 Illustration de la méthodologie géostatistique

Prenons maintenant une tranche du jeu de données décrit à la section précédente afin d'illustrer la méthodologie géostatistique présentée à la section 5.1. La période de 6 heures choisie est celle utilisée en exemple à la section précédente entre 18 UTC le 4 août 2003 et 00 UTC le 5 août 2003. En heure avancée de l'est (HAE), il s'agit de la période du 4 août 2003 entre 14 heures et 20 heures. Voici donc, en suivant les étapes de la figure 5.1, une analyse exploratoire de ces données, la formulation des modèles de krigeage, puis les résultats de krigeage. Afin de comparer le mise en oeuvre de différents types de krigeage, un krigeage ordinaire, un krigeage universel, un krigeage avec modèle de tendance et un cokrigeage ordinaire seront effectués.

5.4.1 Analyse exploratoire

Pendant la période du 4 août 2003 entre 14 et 20 heures, le bassin versant de la Gatineau a connu des pluies nulles à fortes dépendamment du secteur. Le tableau 5.1 présente des statistiques de base sur les données de stations et celles de GEM pour cette période. Le modèle GEM peut parfois produire des prévisions légèrement négatives, mêmes si ces valeurs sont impossibles en réalité. C'est ce qui est arrivé pour la période du 4 août 2003 entre 14 et 20 heures, la valeur minimale des données GEM étant de

-0.03. De plus, les prévisions GEM sont ici nettement supérieures aux observations des stations avec des précipitations moyennes de 5.40 mm contrairement à 1.53 pour les stations. Malgré cette surestimation de GEM par rapport aux stations, les données des deux sources sont relativement reliées, avec une corrélation de 0.35. Notons cependant que cette corrélation est plus ou moins précise car elle est calculée à partir de 16 couples de données seulement. Pour obtenir ces couples de données, les valeurs générées par le modèle GEM ont dû être interpolées aux sites où sont localisées les stations météorologiques. Cette interpolation a été effectuée par la méthode de l'inverse de la distance au carré, en considérant pour chaque site d'observation un voisinage d'interpolation constitué des quatre points de la grille du modèle GEM formant le carré dans lequel tombe le site. Les couples de données stations-GEM seront aussi nécessaires pour déterminer les poids de la prévision par krigeage avec dérive externe.

Source de données	**Min**	**Max**	**Moyenne**	**Écart-type**	**Corrélation**
Stations météorologiques	0.00	9.00	1.53	2.66	0.35
Modèle atmosphérique GEM	-0.03	21.13	5.40	5.56	

TAB. 5.1 – *Statistiques descriptives sur les données*

La figure 5.5 présente les histogrammes des données. Les données des deux sources présentent des distributions asymétriques avec des fréquences plus élevées pour les valeurs faibles. Cette observation est particulièrement critique pour les observations de stations. La normalité de ces données peut être sérieusement mise en doute. Une transformation logarithmique aiderait certainement à rendre plus symétrique la distribution des données, mais cette transformation est déconseillée par plusieurs (Roth, 1998). De plus, la librairie gstat utilisée ici pour mettre en oeuvre l'interpolation par krigeage ne permet pas d'effectuer la transformation inverse des données après l'interpolation. Dans l'étude présentée ici, les calculs seront donc effectués sur les données non transformées. Cependant, étant donné que le krigeage est spécialement adéquat pour des données normales (voir section 4.5.1), il est possible que ses performances laisse ici à désirer. Notons que la distribution de données de précipitations cumulées sur un petit intervalle de temps est souvent asymétrique en raison de la masse de données à zéro. Pour des pas de temps annuel ou mensuel un tel problème se pose rarement, mais plus le pas de temps est petit, plus ce problème est prononcé. Pour l'interpolation de données de précipitations à une fine résolution temporelle comme celle de 6 heures, il serait intéressant d'étudier la modélisation proposée par Velarde *et al.* (2004) qui est spécialement adaptée aux distributions de variables non négatives avec une masse de probabilité à zéro.

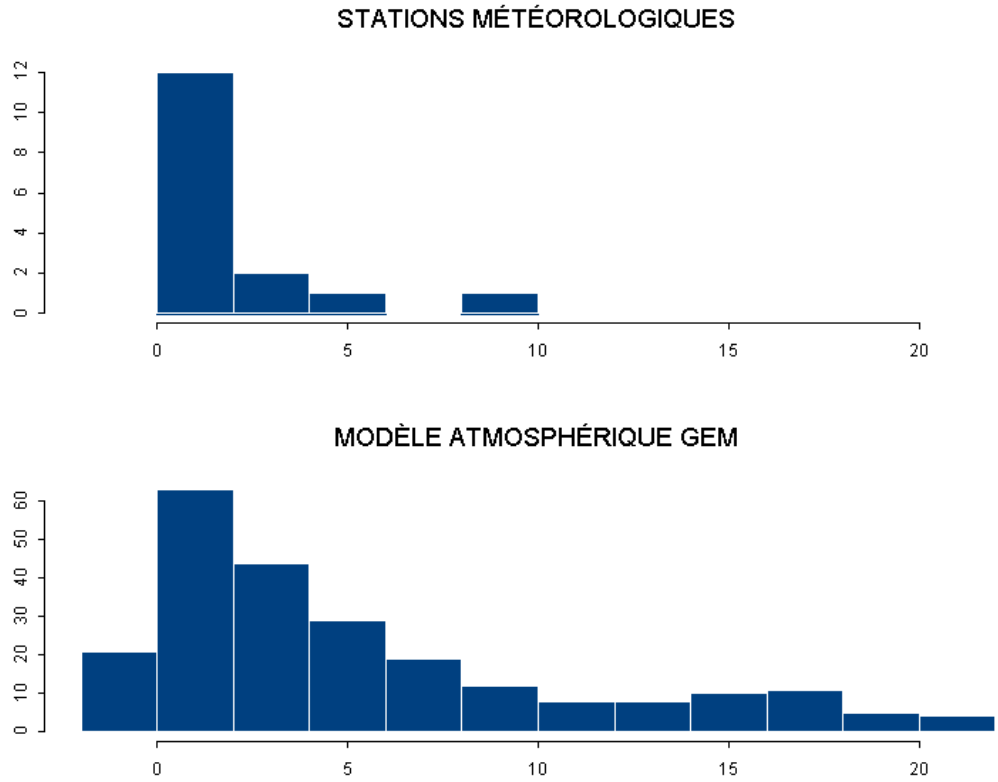

FIG. 5.5 – *Histogrammes des données*

Voyons maintenant comment les données se répartissent dans l'espace. Les figures 5.6 et 5.7 présentent cette répartition spatiale en 3 dimensions et en 2 dimensions respectivement. Afin de tracer les courbes de niveaux de la figure 5.7, les observations des stations ont d'abord été interpolées sur les points de la grille GEM par la méthode de l'inverse de la distance. Le graphique de gauche de la figure 5.6 contient les valeurs des observations de stations. Parmi les 16 stations, six ont mesuré plus de 0.2 mm de pluie. Les précipitations sont donc très localisées. Les deux plus fortes observations de stations sont celles des stations Maniwaki et La Vérendrye. De ces figures, il ressort que les observations de pluie les plus élevées se situent au nord-est du bassin versant. Les deux sources de données sont en accord sur ce point. Par contre, l'autre pic de précipitations n'est pas tout à fait situé au même endroit par les deux sources. Les stations le situe plutôt au sud tandis que le modèle GEM le situe plutôt à l'ouest du bassin. À partir des graphiques de gauche des figures 5.6 et 5.7, il est difficile de tirer des conclusions sur la stationnarité de la variable régionalisée à interpoler. De plus, le petit nombre de stations rend inutile l'étude de la stationnarité au moyen d'une fenêtre mobile. Des graphiques de comportement directionnel pourraient peut-

être aider. La figure 5.8 contient ces graphiques sur lesquels une droite de régression a été tracée. Comme il avait été observé sur les figures 5.6 et 5.7, les précipitations ont tendance à augmenter en se déplaçant vers le nord et l'est. Cependant, en moyenne cette augmentation n'est pas si forte. À la limite, on pourrait donc juger plausible le postulat de stationnarité de ces données.

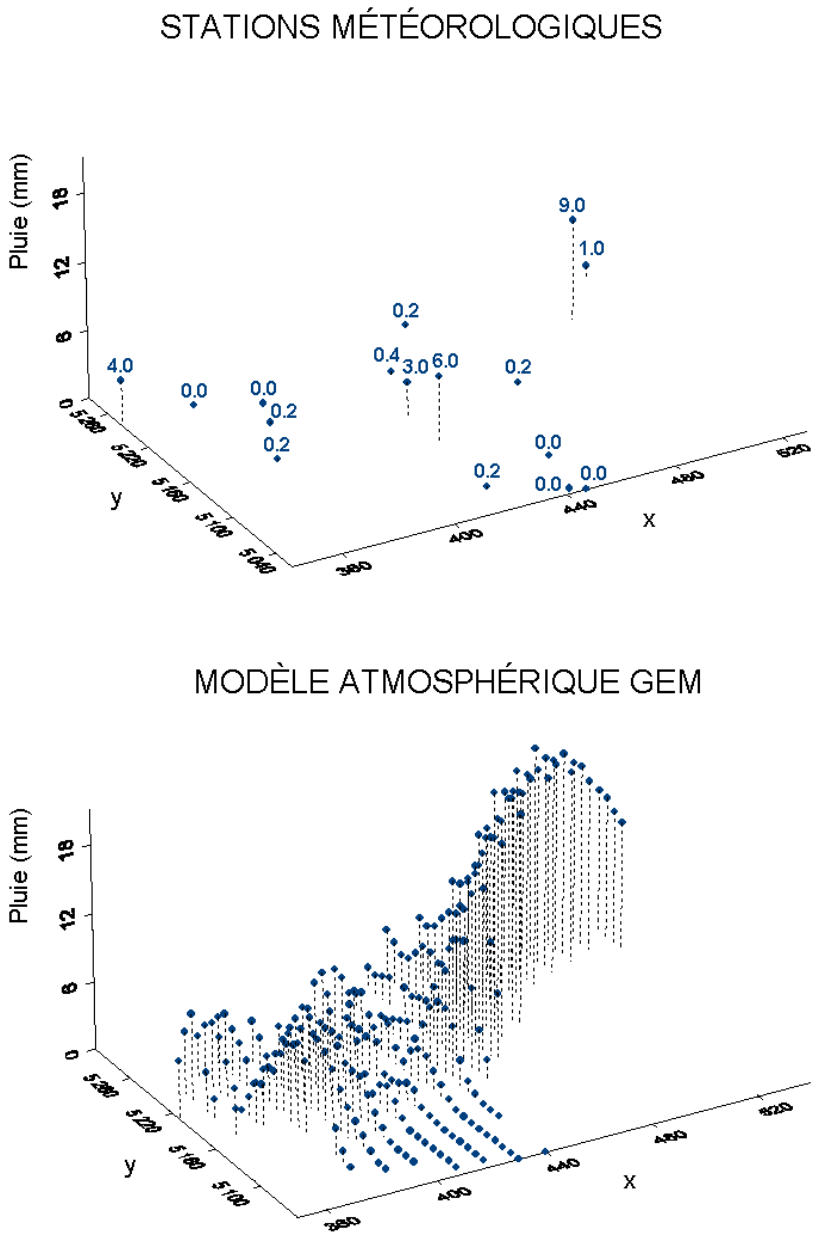

FIG. 5.6 – *Diagrammes de dispersion 3D des données*

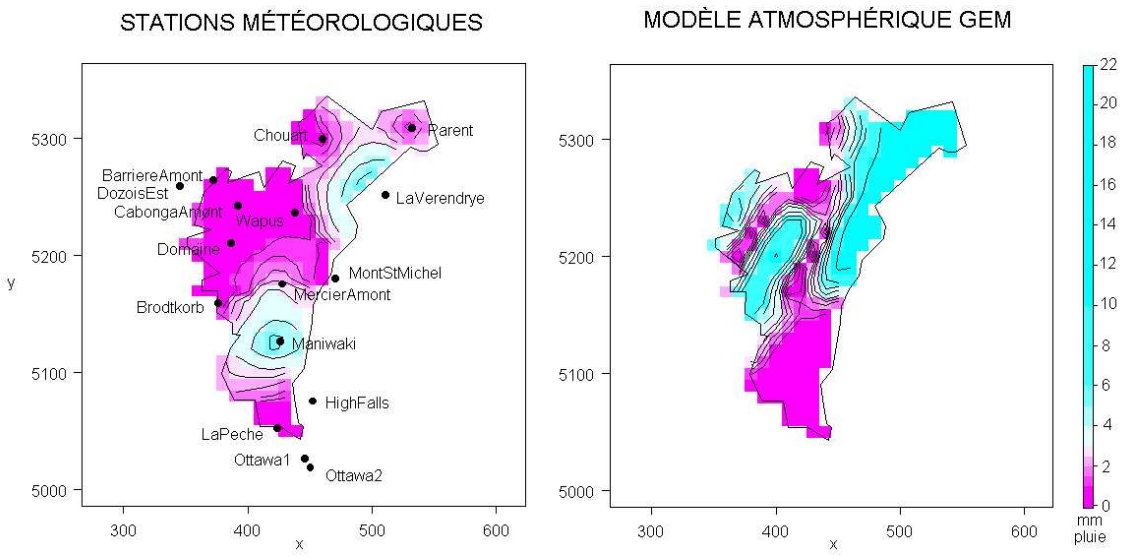

FIG. 5.7 – *Courbes de niveaux des données*

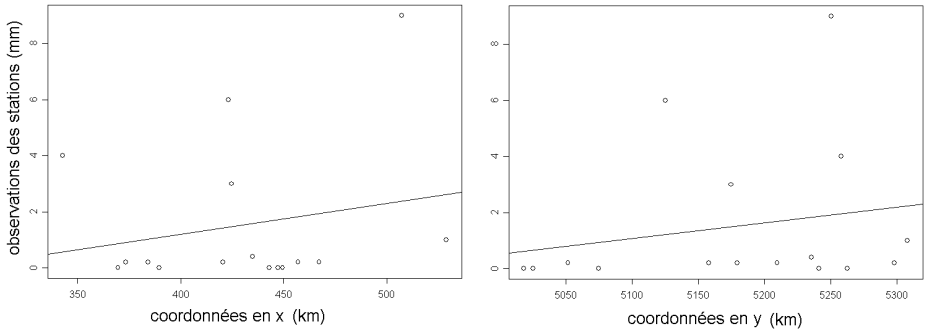

FIG. 5.8 – *Graphiques de comportement directionnel des données*

5.4.2 Formulation du modèle

Quatre types de krigeages sont effectués ici : le krigeage ordinaire, le krigeage universel, le krigeage avec dérive externe et le cokrigeage ordinaire. Tous ces krigeages ont été effectués avec un voisinage composé de toutes les données en raison de leur faible nombre. Pour le krigeage et le cokrigeage ordinaire, aucun choix ne doit être fait concernant les tendances des modèles. Il s'agit toujours de constantes inconnues. Pour les krigeages avec modèle de tendance, des tendances linéaires ont été choisies pour minimiser le nombre de paramètres du modèle. Ainsi, la tendance est $\mu(s) = \beta_0 + \beta_1 x + \beta_2 y$ en krigeage universel et $\mu(s) = \beta_0 + \beta_1 w$ en krigeage avec dérive externe.

Ensuite, l'analyse variographique est effectuée. La figure 5.9 présente les semi-variogrammes expérimentaux pour le krigeage ordinaire, le krigeage universel et le krigeage avec dérive externe. Ces variogrammes sont très semblables. Ainsi, il semble qu'estimer le semi-variogramme sur les résidus ou sur les données brutes n'influence pas beaucoup l'allure du semi-variogramme. En cokrigeage, en plus des semi-variogrammes simples pour les données de stations et les données du modèle GEM, le semi-variogramme croisé entre ces données doit être calculé. La figure 5.10 présente ces semi-variogrammes. Celui des données GEM est beaucoup plus structuré que celui des stations. Cette observation est peu étonnante car le modèle GEM est déterministe. La figure 5.10 contient aussi les modèles variographiques ajustés aux semi-variogrammes expérimentaux. Ces modèles sont de la forme linéaire sans palier et sans effet de pépite car des modèles plus complexes sont difficiles à ajuster avec la librairie gstat de S-Plus. Sur le semi-variogramme croisé de la figure 5.10, il ressort cependant qu'un modèle avec effet de pépite aurait été plus juste. Les trois modèles forment un modèle linéaire de corégionalisation, ce qui assure la positivité des variances tel que mentionné à la section 4.5.3.

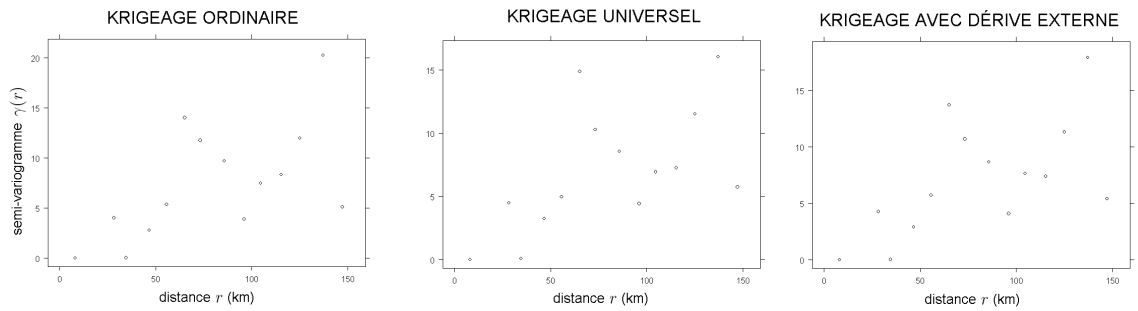

FIG. 5.9 – *Semi-variogrammes expérimentaux estimés en vue du krigeage ordinaire, du krigeage universel et du krigeage avec dérive externe*

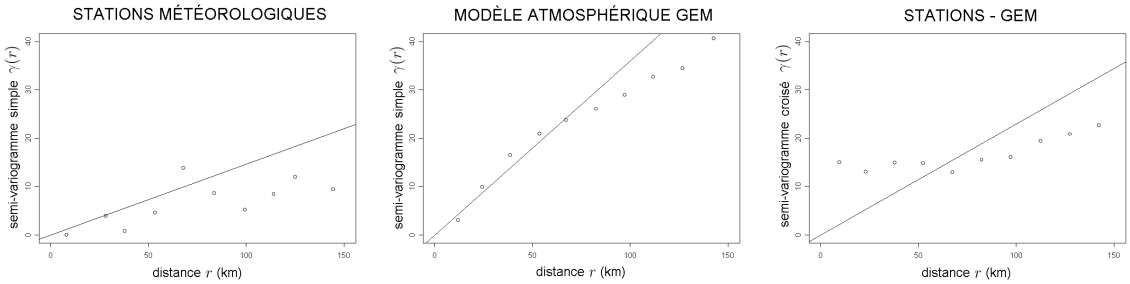

FIG. 5.10 – *Semi-variogrammes simples et croisé estimés en vue du cokrigeage ordinaire accompagnés des droites les modélisant*

Pour les autres types de krigeage, le modèle variographique est choisi par validation croisée. Ainsi, plusieurs modèles sont ajustés et ensuite comparés. Pour illustrer cette étape, la figure 5.11 présente les ajustements des sept modèles testés pour le krigeage ordinaire. Ces sept modèles sont ceux présentés à la section 3.5. Il est difficile de sélectionner à l'oeil le meilleur modèle. La validation croisée facilite le travail. Elle a ici sélectionné le modèle sphérique.

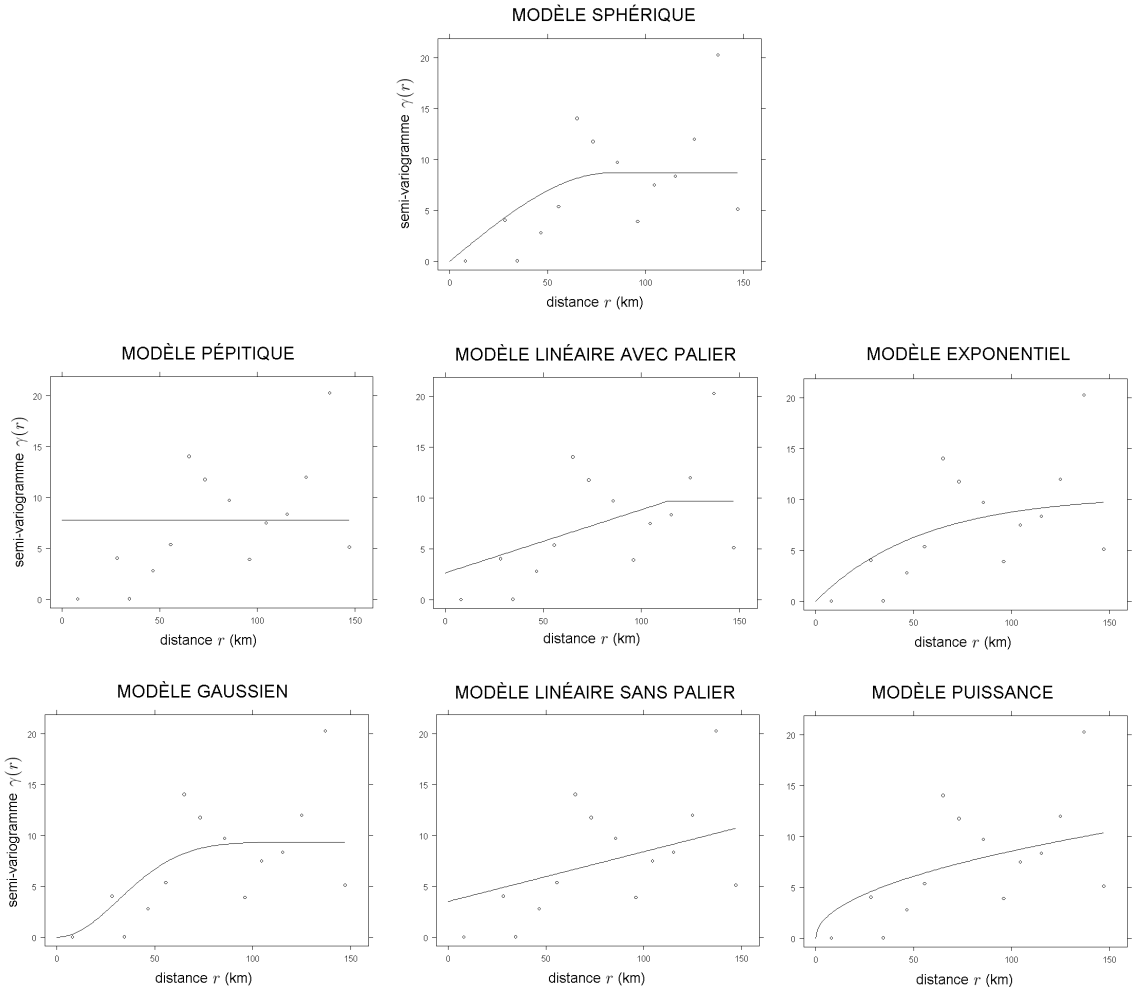

FIG. 5.11 – *Modélisation du semi-variogramme estimé en vue du krigeage ordinaire*

La figure 5.12 représente les surfaces d'interpolation qui sont obtenues avec les différents modèles. Le modèle pépitique donne une surface plane en krigeage ordinaire car il revient à une régression classique et le modèle de tendance est ici une constante. Les modèles sphérique, exponentiel et pépitique ont une allure relativement semblable dans la figure 5.11. En conséquence, ils produisent des surfaces d'interpolations qui se

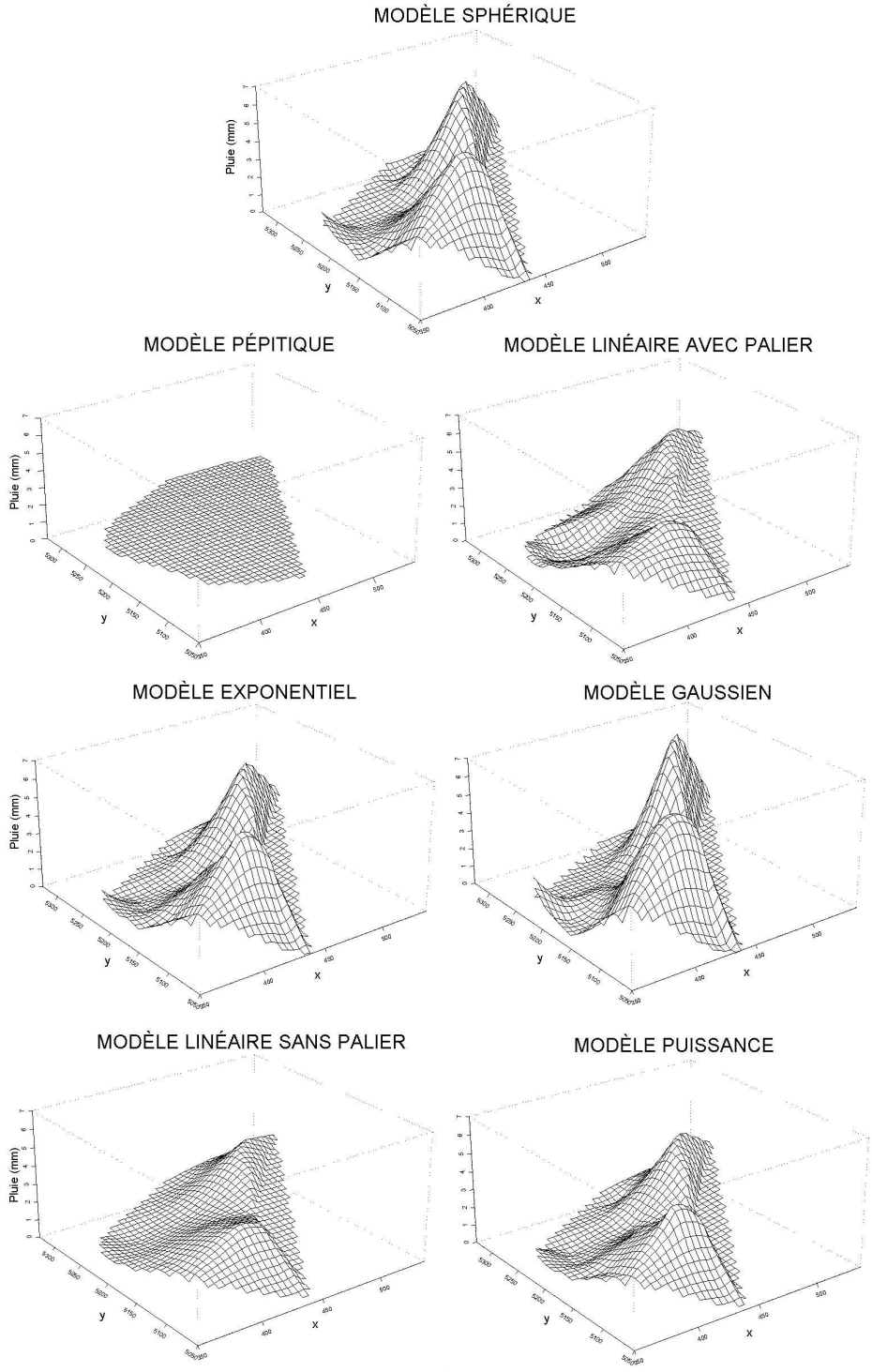

FIG. 5.12 – *Comparaison des surfaces d'interpolation obtenues avec les différents modèles variographiques*

ressemblent. La surface produite avec le modèle gaussien ressemble aussi à ces surfaces, mais elle comporte des pics plus élevés. Finalement, l'utilisation d'un modèle linéaire sans palier plutôt qu'avec palier a pour effet de lisser davantage la surface d'interpolation.

5.4.3 Krigeage

Une fois le modèle spécifié, l'interpolation peut être effectuée. Les surfaces d'interpolation obtenues des différents types de krigeage effectués sont présentées dans la figure 5.13 et le tableau 5.2 contient des statistiques descriptives sur ces valeurs interpolées. On remarque tout de suite que les méthodes multivariables produisent de plus grandes valeurs interpolées, ce à quoi on s'attendait étant donné que les données GEM ont tendance à être plus élevées que les données de stations. Le cokrigeage accorde plus d'importance aux données GEM que le krigeage avec dérive externe. En effet, la surface d'interpolation qu'il produit ressemble beaucoup aux données GEM tandis que la surface du krigeage avec dérive externe se rapproche plus des données de stations. En outre, le krigeage ordinaire et le krigeage universel produisent des valeurs interpolées très semblables.

Méthodes d'interpolation	**Min**	**Max**	**Moyenne**	**Écart-type**
Krigeage ordinaire	0.00	5.99	1.82	1.48
Krigeage universel	0.00	6.00	1.79	1.42
Krigeage avec dérive externe	0.00	7.67	1.94	1.81
Cokrigeage ordinaire	0.00	11.84	2.24	3.19

TAB. 5.2 – *Statistiques descriptives sur les valeurs interpolées*

On peut se demander maintenant laquelle de ces méthodes produit les meilleures prévisions. Il n'y a pas de réponse absolue à cette question. Pour y répondre, il faut d'abord déterminer un critère de performance. Si on choisit comme critère la minimisation de l'EQM de validation croisée, alors la meilleure méthode est ici le krigeage ordinaire avec un EQM de 7.40, comparé à 9.86 pour le krigeage universel, 10.59 pour le krigeage avec modèle de tendance et 9.43 pour le cokrigeage ordinaire.

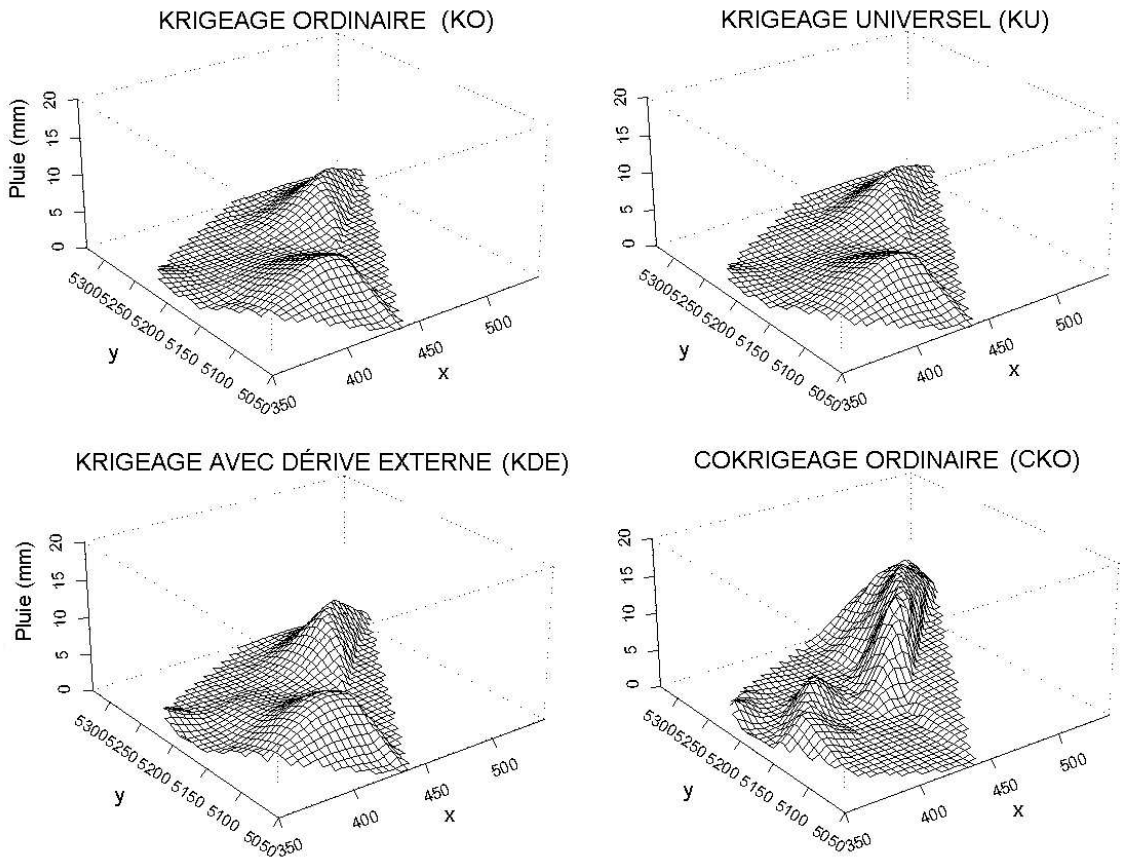

FIG. 5.13 – *Comparaison des surfaces d'interpolation obtenues des différents types de krigeage*

5.5 Conclusion du chapitre

Ce chapitre a présenté une méthodologie géostatistique qui fut illustrée à l'aide de données de précipitations cumulées sur une certaine période de temps. Le chapitre suivant s'intéresse non plus à l'interpolation des données pour une seule période, mais à l'interpolation automatisé des données pour toute période de temps. Le chapitre 6 constitue en fait l'intégrale d'un article rédigé pour les actes du colloque *Géomatique 2004* par Baillargeon *et al.* (2004). On y compare l'efficacité de quatre types de krigeage à deux techniques de régression locale (voir section 2.2.5) et à la méthode de l'inverse de la distance (voir section 2.2.1) afin d'interpoler les données de précipitations. Pour assurer son autosuffisance, cet article contient une brève description théorique du krigeage et de la régression locale. Ces descriptions peuvent être vues comme des résumés des sections 2.2.6 et 2.2.5. De plus, l'article étant contraint de respecter certaines limites

de longueur, sa présentation est la plus synthétique possible. Dans ce mémoire, certains aspects de l'article sont approfondis davantage. C'est le cas de la description des données qui fut détaillée à la section 5.3. La conclusion du mémoire contient également une discussion qui situe l'article du chapitre 6 par rapport à la littérature en interpolation de données de précipitations. Finalement, on retrouve en annexe un complément d'analyse des résultats du chapitre 6 et une copie du programme S-Plus qui a permis d'effectuer l'interpolation.

Chapitre 6

Interpolation statistique multivariable de données de précipitations dans un cadre de modélisation hydrologique

Un article de Sophie BAILLARGEON, Jacynthe POULIOT, Louis-Paul RIVEST, Vincent FORTIN et Josée FITZBACK présenté dans le cadre du colloque *Géomatique 2004 : un choix stratégique* qui s'est tenu à Montréal les 27 et 28 octobre 2004.

Résumé

Divers organismes québécois responsables de la gestion de l'eau se servent de modèles hydrologiques pour simuler et prévoir le débit des rivières. Les champs de précipitations observés donnés en entrée à ces modèles sont obtenus par interpolation spatiale des données de stations météorologiques. Cependant, la faible densité du réseau de stations rend les valeurs interpolées incertaines. Pour améliorer la qualité de cet intrant, l'intégration de nouvelles sources de données est envisagée. Dans le cadre d'un projet de maîtrise, des méthodes statistiques d'interpolation spatiale multivariable ont été étudiées. Une première comparaison théorique a permis d'établir que les techniques de régression locale et de krigeage étaient les plus prometteuses pour produire l'intrant des précipitations observées. Des variantes de ces techniques ont été testées par validation croisée et à l'aide de simulations hydrologiques. Les données de test provenaient de 16 stations météorologiques localisées sur le bassin versant de la rivière Gatineau pour

le mois d'août 2003. Les précipitations prévues par le modèle numérique de prévision météorologique GEM (Environnement Canada) ont été utilisées comme variable auxiliaire. Finalement, l'approche proposée consiste à interpoler les données, pour un temps donné, par la technique parmi celles examinées qui produit les plus petites erreurs de validation croisée. Les tests indiquent que cette approche accomplit d'aussi bonnes performances que la méthode de l'inverse de la distance

Abstract

Various organizations in charge of water management in Quebec use hydrological models to simulate and forecast rivers discharge. The observed precipitation fields given as input for those models are obtained by spatial interpolation of weather stations data. However, the low density of the station's network causes noisy interpolated values. To improve the quality of this input, the integration of new data sources is considered. For the purposes of a master's degree, multivariable statistical methods for spatial interpolation were studied. A preliminary theoretical comparison led to the conclusion that the local regression techniques and the kriging techniques were the most promising to produce the input of observed precipitation fields. Consequently, variants of these techniques were tested by cross-validation and hydrological simulations. The test data came from 16 weather stations located in the Gatineau watershed for August 2003. Precipitation forecasts generated with the numerical weather prediction model GEM (Environment Canada) were used as an auxiliary variable. Finally, the proposed approach consists in interpolating data, for a given period, by the technique among the examined ones which produces the smallest cross-validation errors. The tests suggest that this approach performs as well as the inverse distance method.

6.1 Introduction de l'article

En hydrologie, les précipitations tombées représentent une information indispensable car elles sont la source de l'apport en eaux aux rivières. Des prévisions de débit de cours d'eau se basent donc sur ces données. Ainsi, les précipitations observées constituent l'un des intrants aux modèles hydrologiques de prévision et de simulation (CEHQ, 2004). De tels modèles sont employés par des organismes comme Hydro-Québec et le Ministère de l'environnement du Québec afin de maximiser la production hydroélectrique québécoise tout en minimisant les risques de catastrophes hydrologiques telles que des inondations.

Actuellement, les données de précipitations observées fournies aux modèles proviennent uniquement des stations météorologiques situées sur les bassins versants à l'étude. Toutefois, la densité du réseau de stations sur le territoire du Québec est faible. Par exemple, le bassin versant examiné dans cet article, celui de la rivière Gatineau, est étudié à partir de 16 stations pour une superficie de 21400 km^2. De plus, le mauvais fonctionnement d'un certain nombre de précipitomètres est fréquent. La difficulté d'effectuer la maintenance en hiver en est une cause. Le nombre de stations fournissant des données est donc souvent en deçà du nombre total de stations disponibles sur le territoire. Il faut également souligner que la qualité des mesures prises par certains de ces précipitomètres peut être mise en doute. Notamment, la localisation de certaines stations a été choisie pour leur accessibilité et non pour leur exposition adéquate. En conséquence, ces stations ont tendance à sous-estimer les précipitations à cause du vent. Ainsi, les précipitations observées données en intrant aux modèles hydrologiques proviennent d'un petit nombre de mesures ponctuelles de fiabilité variable. En outre, les modèles ne traitent pas directement les observations ponctuelles des stations. Ces observations doivent d'abord être interpolées sur une grille couvrant le bassin versant. De cette interpolation découle donc une incertitude supplémentaire.

Des intrants de précipitations observées plus fiables et précis permettraient d'obtenir des prévisions plus justes des modèles hydrologiques, d'où une meilleure gestion de l'eau. Cette amélioration serait certainement obtenue en augmentant la densité et la fiabilité du réseau de stations météorologiques au sol. Cependant, cette solution s'avère très coûteuse. Il est plus réaliste de miser sur le perfectionnement de la méthode employée pour effectuer l'interpolation et l'intégration de nouvelles sources de données. Les sources de données envisagées sont les modèles numériques de prévision du temps, les radars météorologiques au sol et la télédétection.

Dans le cadre d'un projet de maîtrise réalisé à l'Université Laval, des méthodes statistiques d'interpolation spatiale multivariable pouvant résoudre cette problématique ont été étudiées. Ce projet visait d'abord à recenser et comparer théoriquement les méthodes. Les techniques de régression locale et de krigeage sont ressorties comme étant prometteuses et assez simples à implanter. La section 2 de cet article comporte une brève introduction à ces méthodes. Ensuite, le projet avait pour but de comparer la performance des méthodes choisies pour interpoler des données de précipitations. Les travaux visant l'atteinte de cet objectif ont finalement mené au développement d'une démarche qui semble fournir des prévisions de précipitations un peu plus justes que celles obtenues de la méthode témoin de l'inverse de la distance. De plus, cette approche permet d'intégrer à l'interpolation des données provenant d'autres sources que les stations météorologiques. Les données sur lesquelles cette étude a été menée sont décrites à la section 3. Ensuite, la méthodologie suivie est détaillée à la section 4,

puis les résultats sont présentés et analysés à la section 5.

6.2 Méthodes statistiques d'interpolation spatiale

L'interpolation spatiale se définie par la prévision de la valeur d'une variable en un site à partir de valeurs mesurées en des sites voisins. Elle peut s'effectuer par une méthode déterministe ou stochastique. Les polygones de Thiessen, la méthode de l'inverse de la distance et les splines sont des exemples de méthodes déterministes (Arnaud et Emery, 2000). Dans ce projet, ces méthodes ont été mises de côté au profit des méthodes stochastiques, qui proposent toutes un modèle probabiliste pour formaliser le comportement du phénomène physique à l'étude. Les surfaces de tendance, la régression locale, le krigeage et les méthodes bayésiennes ont été recensés comme méthodes stochastiques. Cependant, les surfaces de tendance ont été considérées trop simplistes pour être approfondies. Cette méthode est en fait une régression classique reposant sur une hypothèse d'indépendance des observations qui est rarement vérifiée avec des données spatialisées. Les erreurs de prévision ou les tests sur les paramètres de la tendance que la méthode permet de calculer ne sont donc pas fiables (Ripley, 1981, p.29-35). D'autre part, bien que les méthodes bayésiennes semblent très prometteuses (Gaudard *et al.*, 1999), elles n'ont pas été étudiées en raison de leur complexité et des moyens à disposition. De plus, ces méthodes sont relativement nouvelles et peu de logiciels sur le marché permettent de les utiliser. Ainsi, ce projet se concentre sur la régression locale et le krigeage, qui sont ici décrits brièvement.

6.2.1 Régression locale

La régression locale est une méthode de lissage qui permet d'ajuster une surface de régression (Cleveland et Devlin, 1988). Pour répondre à une problématique d'interpolation spatiale, la variable d'intérêt est modélisée par une fonction linéaire ou quadratique des coordonnées spatiales x et y. Cette fonction est ajustée par la méthode des moindres carrés pondérés. Ce qui différencie la régression locale des surfaces de tendance est cette pondération des données, qui est fonction de la distance géographique entre les sites d'observation et le site pour lequel une prévision est voulue. Certains nomment « régression pondérée géographiquement » ce type de régression locale permettant de faire de l'interpolation spatiale (Fotheringham *et al.*, 2002).

Pour utiliser cette technique, il faut préalablement choisir une fonction de poids et

spécifier la taille du voisinage. La fonction de poids détermine dans quelle mesure les observations les plus proches ont plus d'importance dans l'interpolation. Dans ce projet, la fonction Epanechnikov est toujours employée. Ainsi, pour effectuer une prévision au point $s_0 = (x_0, y_0)$, le poids de l'observation prise au point $s_i = (x_i, y_i)$ est $1 - \frac{|s_0 - s_i|}{h(s_0)}^2$ pour $0 \leq |s_0 - s_i|/h(s_0) < 1$, et 0 sinon. L'expression $|s_0 - s_i|$ représente la distance euclidienne entre les points s_0 et s_i, et $h(s_0)$ est la distance au-delà de laquelle les observations se voient accorder un poids nul. Ainsi, $h(s_0)$ constitue la limite du voisinage de s_0. Une pratique courante en régression locale est de spécifier ce voisinage par la fraction des points d'observation que l'utilisateur désire inclure dans la prévision. Plus cette fraction est grande, plus la surface ajustée est lisse.

En interpolation spatiale par régression locale, l'intégration de données autres que celles de la variable d'intérêt se fait par l'ajout de variables auxiliaires dans la tendance. Les poids des données restent cependant fonction des coordonnées spatiales seulement.

6.2.2 Krigeage

En krigeage, la valeur de la variable d'intérêt est prévue en un point par une somme pondérée des observations ponctuelles disponibles. Les poids des données sont choisis de façon à ce que l'interpolation soit sans biais et à variance minimale. Cette technique a été introduite par le Français G. Matheron en 1962 (Matheron, 1962, 1963b). Il s'agit de la première méthode d'interpolation à tenir compte de la structure de dépendance spatiale des données. Notons que le krigeage repose sur les mêmes bases théoriques que l'« interpolation optimale » employée en météorologie (Gandin, 1963).

Il existe plusieurs types de krigeage, qui diffèrent selon la forme postulée pour l'espérance de la variable d'intérêt. Par exemple, lorsqu'il est supposé que l'espérance soit constante et connue, on parle de krigeage simple. S'il est postulé que l'espérance soit constante mais inconnue, il s'agit de krigeage ordinaire. Enfin, le krigeage universel repose sur l'hypothèse que cette espérance soit une fonction des coordonnées spatiales. Ainsi, ce dernier type de krigeage n'est pas stationnaire par rapport à l'espérance contrairement aux deux autres.

La stationnarité se définie ici par la constance de l'espérance, mais aussi par la covariance entre deux observations qui dépend uniquement de la distance entre ces observations. Tous les types de krigeage postulent la stationnarité de la covariance, ou, plus généralement, du semi-variogramme. Cette fonction, qui représente la structure de dépendance spatiale des données, doit être estimée et modélisée avant d'effectuer l'interpolation. La modélisation de cette fonction cause parfois des problèmes en pratique

car cette étape demeure difficilement automatisable.

Le krigeage propose principalement deux façons d'intégrer des variables auxiliaires : le krigeage avec dérive externe et le cokrigeage. En krigeage avec dérive externe, il est supposé que l'espérance de la variable d'intérêt dépende des variables auxiliaires. La théorie de ce krigeage est en fait la même que la théorie du krigeage universel, qui comporte aussi une espérance non constante. Pour sa part, le cokrigeage suggère de prévoir la variable d'intérêt par une combinaison linéaire pondérée de ses observations et des observations des variables auxiliaires. Cette technique requiert l'étude de la dépendance spatiale entre les variables en plus de l'étude des dépendances spatiales simples. Pour plus de détails sur la théorie du krigeage, le lecteur est référé au livre de (Cressie, 1993).

6.3 Données de test et site d'étude

Suite à l'étude théorique, des techniques de régression locale et de krigeage ont été testées en pratique. Les données employées pour cette expérimentation couvrent le bassin versant de la rivière Gatineau pendant le mois d'août 2003. Tel que mentionné précédemment, 16 stations météorologiques se situent à l'intérieur ou à proximité de ce bassin versant d'une superficie de 21400 km^2. Ces stations appartiennent au Ministère de l'environnement du Québec, à Environnement Canada, à Hydro-Québec ou à la Société de protection des forêts contre le feu (SOPFEU). Les données de base de ce projet proviennent de ces stations. Il s'agit de mesures ponctuelles de précipitations cumulées sur 6 heures. Le graphique de gauche de la figure 6.1 permet de localiser les stations et présente le champ moyen de précipitations observées pour une période de 6 heures en août 2003.

En plus des stations météorologiques, une autre source de données a été considérée : le modèle numérique de prévision météorologique, ou modèle atmosphérique, GEM (Global Environnemental Multi-échelles) d'Environnement Canada. Les données générées par ce modèle qui ont été employées dans l'interpolation sont des prévisions de précipitations cumulées sur 6 heures. Ces données se présentent sur une grille de 10km par 10km dont les points sont représentés sur le graphique de droite de la figure 6.1. Ce graphique présente aussi le champ moyen des précipitations 6h prévues par GEM en août 2003. Il indique que le modèle GEM a tendance à surestimer les précipitations comparativement aux stations.

Les données du modèle atmosphérique GEM ont été choisies parce qu'elles étaient facilement accessibles et qu'elles présentaient un bon potentiel. Notons cependant que la

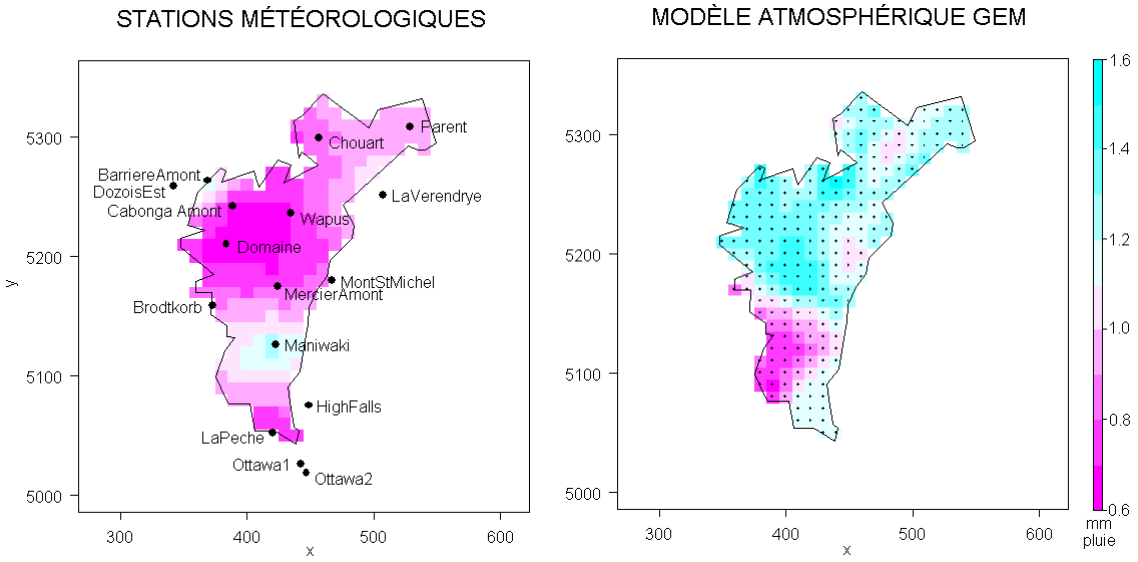

FIG. 6.1 – *Emplacement des stations météorologiques ainsi que des points de la grille du modèle GEM sur le bassin versant de la rivière Gatineau et champs moyens de précipitations observées et prévues pour une période de 6 heures en août 2003 (coordonnées spatiales UTM sur la zone 18T en km)*

variable auxiliaire aurait pu provenir d'autres sources, telles que des radars météorologiques au sol ou des capteurs satellitaires. Bien sûr, des traitements préalables adéquats auraient dû être appliqués à ces données avant de les combiner aux observations des stations.

6.4 Méthodologie

Les données décrites ci-dessus ont permis de comparer la performance de techniques de régression locale et de krigeage pour produire l'intrant des précipitations observées pour un modèle hydrologique. Dans ce projet, la qualité des données interpolées par les différentes méthodes a été testée par validation croisée de type « leave-on-out » (Isaaks et Srivastava, 1989, p.351-368). Cette technique consiste à enlever une à une les observations des stations pour ensuite les prévoir à partir des autres données. Des erreurs de validation croisée sont ensuite obtenues en soustrayant les valeurs prédites aux valeurs observées. Cette procédure est fréquemment utilisée dans le domaine de l'hydrologie pour tester la performance prédictive de méthodes d'interpolation (Goovaerts, 2000; Haberlandt et Kite, 1998). Deux techniques de régression locale et quatre techniques de

krigeage ont été appliquées aux données. La moitié de ces techniques sont univariables, elles n'utilisent donc que les données des stations. Les autres sont multivariables ; elles intègrent les données du modèle atmosphérique GEM à l'interpolation. L'utilisation de ces deux groupes de techniques a permis d'évaluer l'apport de l'intégration des données GEM à l'interpolation. Le tableau suivant décrit les méthodes employées :

Identifiant	**Méthode**	**Type**	**Paramètre** (krigeage : modèle variographique, régression locale : fraction voisinage)
RLxy	Régression locale par rapport aux coordonnées spatiales x et y	univariable	Sélection par validation croisée
RLxyw	Régression locale par rapport à x et y, ainsi qu'à la variable auxiliaire	multivariable	Sélection par validation croisée
KO	Krigeage ordinaire	univariable	Sélection par validation croisée
KU	Krigeage universel	univariable	Sélection par validation croisée
KDE	Krigeage avec dérive externe	multivariable	Sélection par validation croisée
CKO	Cokrigeage ordinaire	multivariable	Linéaire sans palier

TAB. 6.1 – *Description des méthodes d'interpolation employées*

Dans ce projet, la sélection des paramètres de l'interpolation s'est toujours effectuée sans intervention humaine. Ainsi, une utilisation en temps réel des méthodes proposées pour produire les intrants à un modèle hydrologique serait envisageable. La solution adoptée pour le cokrigeage (CKO) était de toujours ajuster le même modèle aux semi-variogrammes expérimentaux. Le modèle linéaire sans palier a été choisi car il est le plus simple à ajuster. Pour toutes les autres méthodes, la validation croisée était utilisée afin d'automatiser l'étape de la sélection du modèle variographique ou de la fraction de voisinage. En krigeage, les modèles variographiques pépitique, sphérique, exponentiel, gaussien, linéaire avec et sans palier et puissance étaient en compétition. Pour la régression locale, les fractions de voisinage possibles étaient 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 et 1. À chaque temps, le modèle ou la fraction choisi était celui qui minimisait la moyenne des erreurs de validation croisée au carré (Erreur Quadratique Moyenne). L'indice plus robuste de la médiane des valeurs absolues des erreurs a aussi été testé, mais il donnait de moins bons résultats. Notons que l'utilisation de la validation croisée pour choisir la fraction de voisinage en régression locale est répandue et approuvée (Cleveland et Loader, 1996). Cependant, on ne peut en dire autant concernant l'utilisation de la validation croiée en géostatistique. Notamment, Isaaks et Srivastava (1989, p.514) ainsi que Cressie (1993, p.104) soutiennent que l'expertise humaine est le meilleur outil pour la

modélisation du semi-variogramme et ils déconseillent l'emploi de la validation croisée à cette étape. Pour sa part, Davis (1987) ne rejette pas l'approche. Il rappelle toutefois que le modèle variographique choisi par validation croisée n'est pas optimal, il est simplement le meilleur parmi les modèles envisagés selon le critère utilisé.

Les six méthodes du tableau 6.1 ont été employées pour interpoler les données à chaque période de temps pour laquelle il a plu. En août 2003, cela représente 92 périodes de 6 heures sur 124. En fait, le krigeage avec dérive externe (KDE) et la régression locale multivariable (RLxyw) n'ont pu être appliqués lorsque les données GEM étaient trop peu variables. Ainsi, ces méthodes n'ont été utilisées que pour 63 des 92 périodes de pluie du mois. En outre, les techniques de régression locale et de krigeage examinées dans ce projet peuvent produire des prévisions inférieures à zéro, ce qui est inapproprié pour une variable non négative telle qu'une quantité de précipitations. Les prévisions négatives ont donc toujours été ramenées à zéro.

Plutôt que d'utiliser la même méthode d'interpolation à toutes les périodes de temps, l'option de sélectionner la méthode la plus appropriée à chaque temps a été examinée. Une septième approche a donc été testée. Celle-ci consiste à effectuer les prévisions à un temps donné par la méthode parmi les six du tableau 1 qui minimise la moyenne des erreurs de validation croisée au carré. Cette procédure d'interpolation sera ici nommée « Sélection » . Haberlandt et Kite (1998) ont aussi exploité, dans un contexte de modélisation hydrologique, l'idée d'appliquer à chaque période de temps la méthode la plus appropriée selon un certain critère. Ils ont effectué un krigeage avec dérive externe conditionnellement à une corrélation entre les données de stations et la variable auxiliaire dépassant un certain seuil. Pour une corrélation inférieure au seuil, le krigeage ordinaire était employé. Cette procédure est très logique, mais elle permet seulement de choisir entre deux méthodes : une univariable et une multivariable. L'approche proposée ici est plus souple, permettant de choisir entre autant de méthodes que le désire l'utilisateur. D'ailleurs, aucune procédure de ce genre ne semble avoir été suggérée dans la littérature.

Le modèle hydrologique HYDROTEL (INRS-ETE, 1998) a été employé pour tester l'approche proposée. Ce modèle comporte un module d'interpolation de données par la méthode de l'inverse de la distance. Cette méthode d'interpolation a donc été choisie comme base de comparaison. Une prévision effectuée par cette méthode en un point de l'espace est donnée par la moyenne des observations des trois stations les plus proches avec une pondération inversement proportionnelle à la distance entre le point d'observation et le point de prévision.

Les calculs ont tous été effectués avec le logiciel S-Plus. Les fonctions de la librairie

Gstat (Pebesma, 2004a) ont permis d'exécuter tous les types de krigeage.

6.5 Résultats

La figure 6.2 permet de comparer les erreurs de validation croisée des différentes procédures d'interpolation. L'identifiant « InvDist » désigne la méthode témoin de l'inverse de la distance. Les autres méthodes sont identifiées comme dans la section précédente. Pour les diagrammes en boîte, les moustaches représentent le premier et le neuvième décile, les contours de la boîte le premier et le troisième quartile, la ligne centrale la médiane et le cercle la moyenne. L'indice EQM (erreur quadratique moyenne) a aussi été ajouté à la figure 6.2. Toutes ces statistiques sont calculées à partir des erreurs de validation croisée de toutes les stations et de toutes les périodes de temps d'utilisation des méthodes, ce qui représente entre 900 et 1400 données environ.

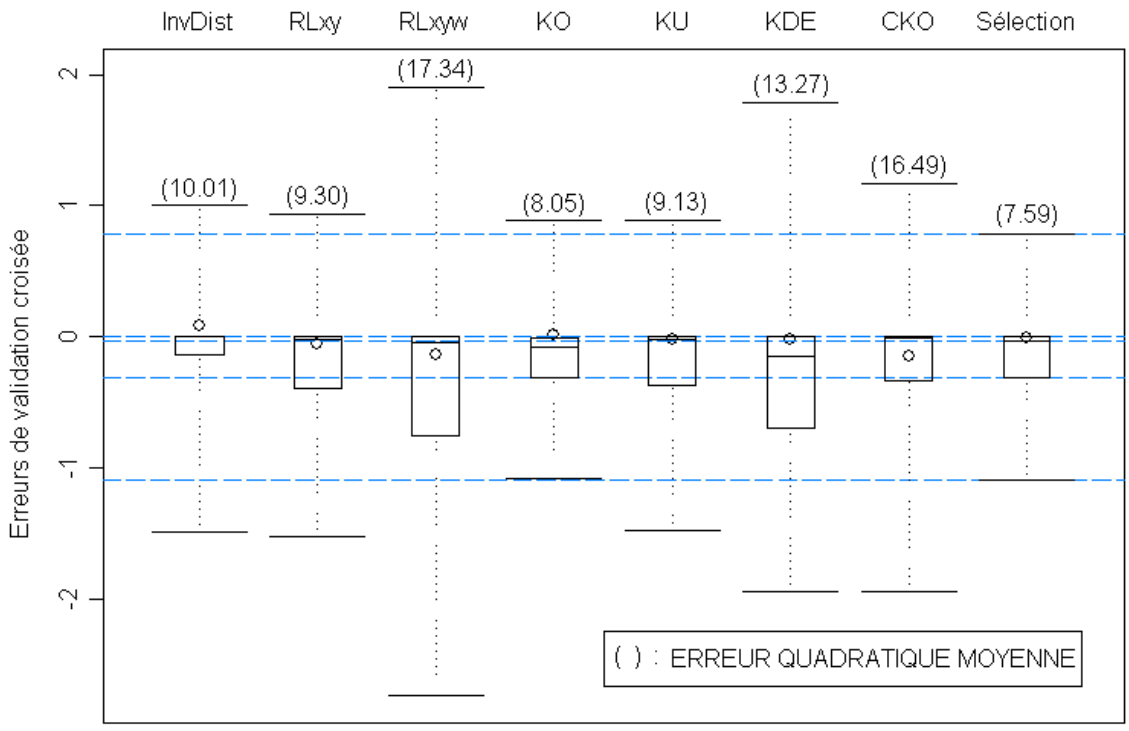

FIG. 6.2 – *Diagrammes en boîte des erreurs de validation croisée pour les techniques d'interpolation spatiale examinées*

Les lignes pointillées de la figure 6.2 mettent en évidence que les résidus de l'approche Sélection ont tendance à être moins extrêmes que ceux de toute autre méthode. En effet,

le diagramme en boîte de Sélection possède les plus courtes moustaches. De plus, son erreur quadratique moyenne est inférieure à toutes les autres. Par contre, la boîte du diagramme en boîte de la méthode de l'inverse de la distance est plus petite que celle de l'approche Sélection. Les erreurs de validation croisée de la méthode témoin sont donc plus concentrées autour de zéro. En fait, ce grand nombre d'erreurs presque nulles est associé aux précipitations observées de zéro millimètre de pluie. La méthode de l'inverse de la distance reproduit bien ces observations, possiblement à cause du petit voisinage de prévision qu'elle emploie. Dans ce projet, en krigeage, toutes les données étaient incluses dans l'interpolation. Il serait intéressant de tester l'utilisation des techniques de krigeage avec un voisinage restreint.

L'impact sur des simulations hydrologiques de l'emploi des méthodes Sélection et Inverse de la distance pour interpoler les données fournies en entrée au modèle a été évalué avec le modèle HYDROTEL. La figure 6.3 présente les débits simulés accompagnés des observations de débit prises à la station hydrométrique des rapides Ceizur sur la rivière Gatineau.

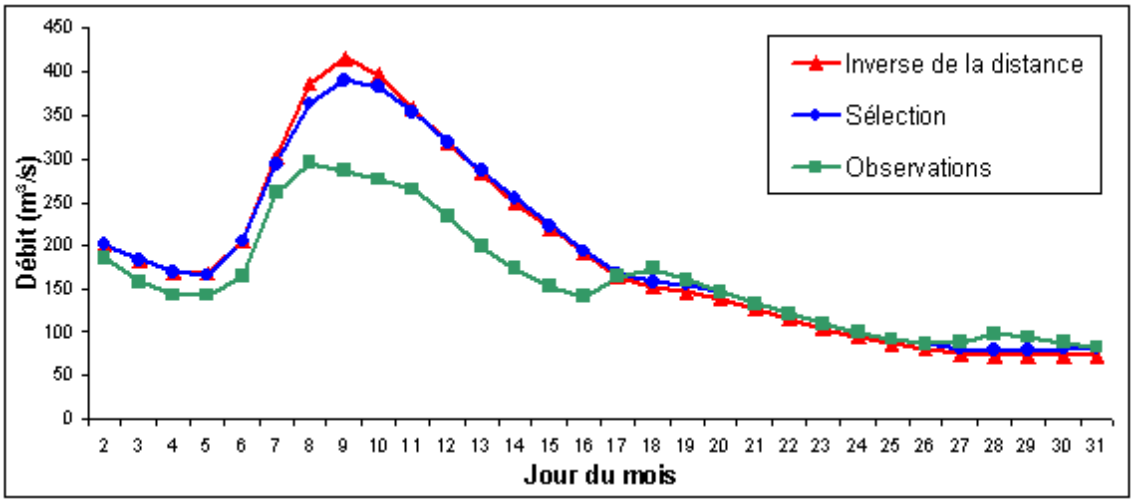

FIG. 6.3 – *Débit de la rivière Gatineau aux Rapides Ceizur en août 2003*

Cette figure indique que les méthodes Sélection et Inverse de la distance accomplissent des performances semblables. Les simulations qu'elles permettent d'obtenir d'HYDROTEL sont toujours très rapprochées. Ainsi, pour les données de test employées, la méthode Sélection proposée ne semble pas donner de résultats significativement meilleurs que la méthode de l'inverse de la distance programmée dans HYDROTEL. Par contre, étant donné que les tests ont été effectués sur une courte période et un seul bassin versant, les résultats ne sont pas généralisables à l'ensemble du territoire du Québec. Il serait donc intéressant de refaire ces tests pour d'autres périodes présentant

des événements de pluie différents et d'autres bassins versants.

Approfondissons maintenant les résultats obtenus de l'approche Sélection. Comme expliqué précédemment, à chaque période de temps cette procédure choisie une méthode à appliquer aux données. La figure 6.4 représente les fréquences de sélection des méthodes par l'approche Sélection.

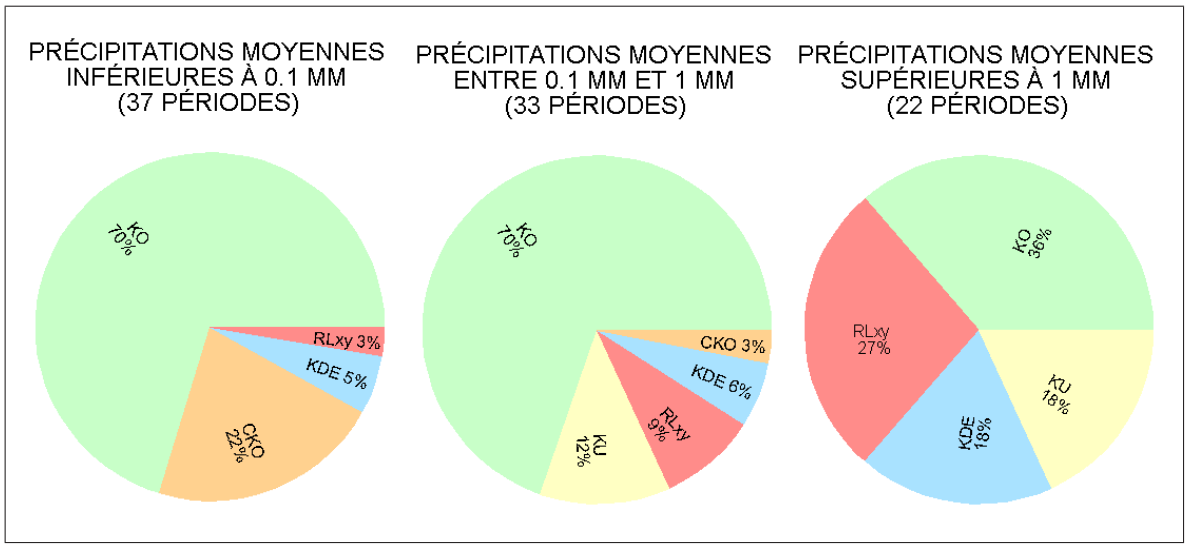

FIG. 6.4 – *Fréquences de sélection des méthodes par l'approche Sélection pour les 92 périodes de pluie d'août 2003 catégorisées selon l'intensité des précipitations*

Pour la majorité des périodes (au total 57 sur 92), l'approche Sélection a choisi d'appliquer le krigeage ordinaire (KO) aux données. Il n'est donc pas étonnant que les erreurs de validation croisée de la procédure Sélection se distinguent peu de celles du krigeage ordinaire dans la figure 6.2. Ces deux méthodes présentent des diagrammes en boîte très semblables. Elles sont notamment les deux méthodes qui surestiment le moins les précipitations, présentant les résidus négatifs les moins gros. En outre, il est pertinent de se demander si l'erreur quadratique moyenne de 7.59 pour la méthode Sélection est significativement plus petite que celle de 8.05 du krigeage ordinaire. En fait, il n'est pas étonnant que l'approche Sélection ait un indice EQM plus petit que toutes les méthodes du tableau 1 considérant le fait qu'à chaque période de temps la procédure Sélection applique justement la méthode qui minimise le EQM. Ainsi, il ne serait pas vraiment exact de conclure ici que l'approche Sélection accomplit de meilleures performance que le krigeage ordinaire employé systématique. Cependant, la procédure Sélection est définitivement plus souple, permettant d'inclure des variables auxiliaires dans l'interpolation. Elle offre donc un bon potentiel de performance, notamment pour des bassins versants contenant moins de stations que celui de la rivière Gatineau.

Concernant les deux autres méthodes univariables d'interpolation, soient la régression locale par rapport aux coordonnées spatiales (RLxy) et le krigeage universel (KU), elles donnes des résultats relativement bons, mais elles ne ressortent pas comme étant meilleures que la méthode de l'inverse de la distance. En effet, les moustaches de leurs diagrammes en boîte sont aussi longues que celles de la méthode témoin, mais leurs boîtes sont plus larges (figure 6.2). Malgré tout, une utilisation sporadique de ces méthodes pourrait amener de bons résultats, surtout pour des périodes de pluies assez fortes. En effet, la figure 6.4 indique que RLxy et KU sont utilisés dans l'approche Sélection respectivement 27% et 18% des périodes de précipitations moyennes supérieures à 1 millimètre.

Comme pour les méthodes RLxy et KU, l'emploi du krigeage avec dérive externe (KDE) ou du cokrigeage ordinaire (CKO) pour interpoler les données n'est avantageux que s'il est effectué au bon moment. Le cokrigeage ordinaire se distingue surtout pour les périodes de pluie très faible (précipitations moyennes inférieures à 0.1 mm). L'approche Sélection l'utilise 22% des périodes de cette classe (figure 6.4). À l'opposé, le krigeage avec dérive externe semble mieux performer pour des périodes de pluies fortes. La figure 4 indique que cette méthode produit les plus petites erreurs de validation croisée selon l'indice EQM pour 18% des périodes de précipitations moyennes supérieures à 1 mm. Cependant, il est étonnant de remarquer que les périodes de temps pour lesquelles une de ces méthodes multivariables est choisie ne présentent pas nécessairement une forte corrélation entre les données de stations et les données GEM. En fait, ces corrélations sont inférieures à 0.5 pour 86% des 92 périodes de pluie. Les grandes étendues des diagrammes en boîte de CKO et de KDE sur la figure 6.2 s'expliquent sûrement par ces faibles corrélations. L'intégration à l'interpolation de variables auxiliaires plus corrélées aux observations de stations donnerait possiblement de meilleurs résultats.

Finalement, la méthode multivariable de régression locale RLxyw ne fournit jamais de bons résultats, qu'elle soit utilisée constamment ou sporadiquement. Elle n'est choisie pour aucune des périodes en août 2003 par la procédure Sélection. De plus, ses erreurs de validation croisée sont celles qui s'éloignent le plus de zéro.

6.6 Conclusion de l'article

Dans ce projet, certaines techniques de régression locale et de krigeage ont été employées afin d'interpoler des données de précipitations. Dans une perspective de préparation d'intrants pour un modèle hydrologique utilisé en temps réel, une procédure a été proposée. Celle-ci consiste à sélectionner une méthode statistique de prévision

spatiale par validation croisée à chaque période de temps. Cette approche est plus intéressante que de toujours employer la même technique, car elle combine les forces des différentes méthodes. En particulier, les données de variables auxiliaires peuvent être intégrées à la prévision par le biais d'une méthode multivariable d'interpolation si la validation croisée indique que ce serait avantageux de le faire. Pour les données de test du bassin versant de la Gatineau en août 2003, la méthode proposée accomplit d'aussi bons résultats que la méthode de l'inverse de la distance. Ces méthodes ont été comparées par validation croisée et à l'aide du modèle hydrologique de prévision et de simulation HYDROTEL.

Rappelons qu'une seule variable auxiliaire a été employée ici : les prévisions générées par le modèle atmosphérique GEM. Il serait intéressant d'étudier l'intégration d'autres variables auxiliaires, qui fourniraient possiblement des résultats plus notables. L'utilisation d'une technique permettant d'exploiter la dépendance temporelle entre les périodes de temps serait aussi une avenue intéressante, qui n'a pas été explorée dans ce projet. Finalement, il faut noter que les méthodes examinées ici ne sont pas vraiment adaptées à l'interpolation de variables non négatives prenant souvent des valeurs nulles, comme les données de précipitations aux 6 heures. Une version lognormale du krigeage assure des prévisions positives, mais cette technique est déconseillée en raison des difficultés que la transformation inverse des prévisions présente (Roth, 1998). L'emploi d'autres approches, notamment des méthodes bayésiennes (Banerjee *et al.*, 2004; Velarde *et al.*, 2004), serait sûrement indiqué pour des données de précipitations.

Chapitre 7

Conclusion

L'article du chapitre précédent illustre une application courante du krigeage : l'interpolation de données de précipitations. Avant de clore ce mémoire, une revue de la littérature concernant ce domaine d'application est présenté dans la section suivante. Cette revue de littérature permet de comparer les résultats obtenus dans notre article ainsi que la méthodologie suivie pour obtenir ces résultats à ce que d'autres chercheurs ont fait. Rappelons qu'en plus de l'examen des résultats inclus dans le chapitre 6, l'annexe A présente une analyse complémentaire des résultats et l'annexe B contient le programmes S-Plus qui a permis de faire les interpolations.

7.1 Littérature en interpolation de données de précipitations

L'interpolation spatiale de données de précipitations est une problématique sur laquelle un bon nombre de chercheurs se sont penchés. Les méthodes utilisées pour interpoler ces données sont variées comme le démontre le projet « Spatial Interpolation Comparison 97 » constituant un numéro de la revue en ligne *Journal of Geographic Information and Decision Analysis* (http://www.geodec.org/). Les chercheurs ayant participé à ce projet devaient interpoler des précipitations journalières mesurées par 100 stations météorologiques couvrant le territoire de la Suisse. Ils ont employé des méthodes barycentriques (Ali, 1998), des splines (Hutchinson, 1998a,b; Saveliev *et al.*, 1998), la régression locale (Rajagopalan et Lall, 1998), des techniques de krigeage (Atkinson et Lloyd, 1998; Bruno et Capicotto, 1998) et des réseaux de neurones (Demyanov *et al.*, 1998; Lee *et al.*, 1998). En plus de ces méthodes, d'autres chercheurs emploient

aussi pour interpoler des données de précipitations des techniques simples de partitionnement de l'espace (e.g. Dirks *et al.*, 1998; Haberlandt et Kite, 1998; Goovaerts, 2000) et de régression classique (e.g. Kruizinga et Yperlaan, 1978; Tabios et Salas, 1985; Abtew *et al.*, 1985), ainsi que des méthodes bayésiennes plus complexes (e.g. Gaudard *et al.*, 1999; Johns *et al.*, 2001; Velarde *et al.*, 2004). Bref, toutes les classes de méthodes présentées au chapitre 2 se retrouvent dans la littérature sur l'interpolation de précipitations. Les techniques qui semblent les plus fréquemment utilisées en pratique sont les méthodes barycentriques d'inverse de la distance et les techniques classiques de krigeage tel que le krigeage ordinaire.

Comme dans le chapitre 6 de ce mémoire, les publications concernant l'interpolation de précipitations ont souvent pour but de comparer des méthodes (Kruizinga et Yperlaan, 1978; Creutin et Obled, 1982; Tabios et Salas, 1985; Phillips *et al.*, 1992; Abtew *et al.*, 1985; Dirks *et al.*, 1998; Haberlandt et Kite, 1998; Nalder et Wein, 1998; Goovaerts, 2000; Chegini *et al.*, 2001; Syed *et al.*, 2003). Lorsqu'un grand nombre de données sont disponibles, ces comparaisons sont effectuées en divisant le jeu de données en deux : un jeu de données pour l'interpolation et l'autre pour la validation. Cette méthode est idéale car la validation est complètement indépendante de la formulation des modèles. Cependant, souvent les données sont trop peu nombreuses et la comparaison des méthodes est plutôt faite par validation croisée comme dans le chapitre 6. Peu importe que la validation soit indépendante ou croisée, elle permet d'obtenir des erreurs de prévisions. Afin de cerner la meilleure méthode, c'est-à-dire celle qui produit les plus petites erreurs, la pratique usuelle est de calculer des indices à minimiser tel l'EQM. Il serait intéressant de pouvoir faire des tests formels sur ces indices pour comparer les méthodes de façon rigoureuse. Par exemple, pourquoi ne pas comparer les EQM de deux méthodes en faisant un test t pairé sur les erreurs au carré ? Malheureusement un tel test serait incorrect en raison de la dépendance spatiale des erreurs. À cause de cette dépendance, aucun test simple n'est approprié. Des tests formels sont possibles seulement dans le cas où plusieurs jeux de données pouvant être considérés indépendants servent à tester les méthodes d'interpolation. Par exemple, Nalder et Wein (1998) travaille sur des précipitations mensuelles et ils possèdent des données pour 12 mois consécutifs. Supposant l'indépendance temporelle entre les mois et la normalité de leurs données, ils comparent deux méthodes d'interpolation par un test t pairé sur les 12 couples d'indices de validation croisée. À l'annexe A, un test non-paramétrique équivalent est effectué sur nos données pour comparer la méthode de l'inverse de la distance à la méthode Sélection en supposant l'indépendance entre certaines périodes de 6 heures du mois d'août 2003.

Dans la littérature, les résultats des comparaisons diffèrent d'une étude à l'autre. Les performances des méthodes dépendent de plusieurs facteurs, notamment des résolutions

temporelle et spatiale des données, ainsi que des paramètres des modèles comme le semi-variogramme en krigeage. Les articles répertoriés ici traitent de l'analyse de précipitations annuelles, mensuelles, journalières, horaires ou totales pour des événements de précipitations de durées quelconques. Les densités des réseaux de stations utilisés varient d'environ une station pour 3 km^2 à une station pour 50 000 km^2. Il est donc impossible de tirer une conclusion générale et aucune méthode ne ressort comme étant universellement la meilleure. D'ailleurs, même à l'intérieur d'une étude, les performances d'une méthode se démarquent rarement nettement des autres. Par exemple, Kruizinga et Yperlaan (1978) ne préconisent l'emploi d'aucune méthode spécifique car ils n'observent pas de différences marquées entre les méthodes. D'autres auteurs recommandent la méthode parmi les meilleures qu'ils jugent la plus pratique. Par exemple, Tabios et Salas (1985), Abtew *et al.* (1985) et Syed *et al.* (2003) notent tous des performances relativement équivalentes entre le krigeage ordinaire et les fonction multiquadratiques (type de splines). Cependant, Tabios et Salas (1985) ainsi que Abtew *et al.* (1985) recommandent l'utilisation du krigeage ordinaire car il permet de calculer des erreurs de prévision, tandis que Syed *et al.* (2003) choisissent d'employer les fonctions multiquadratiques car ils les considèrent plus faciles d'utilisation. De même, sur un réseau dense de stations, Dirks *et al.* (1998) n'obtiennent pas d'améliorations significatives des résultats en employant un krigeage ordinaire plutôt qu'une méthode de l'inverse de la distance. Ils recommandent donc la méthode plus simple de l'inverse de la distance. Ainsi, des résultats comme ceux obtenus au chapitre 6 qui ne mettent pas en évidence la supériorité de méthodes élaborées comme le krigeage sur des méthodes simples tel l'inverse de la distance sont usuels avec des données de précipitations. En fait, cette remarque semble se généraliser aussi à d'autres types de données. Dans une étude expérimentale sur des données simulées, Zimmerman *et al.* (1999) obtiennent une meilleure interpolation avec le krigeage ordinaire ou universel qu'avec la méthode de l'inverse de la distance seulement lorsque le plan d'échantillonnage des données est régulier, le bruit est faible et la corrélation spatiale est forte. Dans l'étude du chapitre 6, aucune de ces caractéristiques n'est rencontrée.

Certaines études examinent des méthodes d'interpolation multivariable. Notamment, Phillips *et al.* (1992); Nalder et Wein (1998); Goovaerts (2000) et Chegini *et al.* (2001) incorporent l'élévation à l'interpolation des précipitations. Ils le font principalement par splines partielles, krigeage simple avec une espérance variant localement, krigeage ordinaire sur les résidus d'une régression des moindres carrés ordinaire (detrended kriging), krigeage avec dérive externe et cokrigeage. Ces méthodes multivariables semblent donner de meilleurs résultats que les méthodes univariables sur des régions montagneuses à l'échelle de dizaines de milliers de km^2 (Phillips *et al.*, 1992) ou lorsque la corrélation entre les données de stations et l'élévation est supérieure à 0.75 (Goovaerts, 2000; Chegini *et al.*, 2001). Notons que toutes ces études ont été menées sur des

précipitations annuelles ou mensuelles. Pour de plus fines résolutions temporelles, telle qu'une résolution journalière, une forte relation entre l'élévation et les précipitations peut être mise en doute selon Haberlandt et Kite (1998). Ces chercheurs ont analysés des données journalières de 1986 à 1990 mesurées par 81 stations sur le bassin versant de la rivière Mackenzie dans le nord-ouest du Canada d'une superficie d'environ 1 800 000 km^2 . Même s'ils observent une corrélation moyenne de 0.52 entre l'élévation et le cumul annuel des précipitations, cette corrélation tombe à 0.06 pour les observations journalières. Ils misent donc davantage sur l'intégration à l'interpolation d'une autre variable auxiliaire : les précipitations prévues par un modèle atmosphérique. Leurs travaux se rapprochent donc beaucoup des nôtres. De plus, ils ont eux aussi étudié l'interpolation de précipitations dans un contexte de modélisation hydrologique et ils ont utilisé des simulations hydrologiques en plus de la validation croisée pour comparer les méthodes testées. La seule méthode multivariable qu'ils ont examinée est le krigeage avec dérive externe. Ils appliquent cette méthode soit à tous les pas de temps de leur période test, soit uniquement lorsque la corrélation entre les observations des stations et les données auxiliaires dépasse un certain seuil (0.5 ou 0.3 dépendamment de la variable auxiliaire en question). Si la corrélation est trop faible, le krigeage ordinaire est employé. Haberlandt et Kite (1998) obtiennent de meilleurs résultats en appliquant conditionnellement le krigeage avec dérive externe plutôt qu'en l'utilisant à chaque pas de temps. Par contre, les indices de validation croisée obtenus pour le krigeage avec dérive externe conditionnel ne sont que très légèrement meilleurs que ceux du krigeage ordinaire. De plus, pour les simulations hydrologiques, c'est le krigeage ordinaire qui donne les meilleurs résultats. Leurs résultats sont donc semblables à ceux du chapitre 6 : selon la validation croisée, les méthodes multivariables apportent une amélioration à la qualité de l'interpolation seulement si elles sont utilisées au bon moment, mais cette amélioration ne semble pas avoir d'impact en bout de ligne sur la qualité des prévisions hydrologiques.

7.2 Synthèse du mémoire

Après la lecture de ce mémoire, le lecteur devrait savoir ce qu'est le krigeage, comprendre la méthode et être capable de l'utiliser. Le krigeage a ici été présenté d'un point de vue théorique et appliqué. Le choix de ce sujet a d'abord été justifié en comparant le krigeage aux autres méthodes d'interpolation spatiale. Ensuite, les fondements mathématiques du krigeage et de son étape préalable, l'analyse variographique, ont été examinés et discutés. La pratique du krigeage a finalement été traitée en détaillant une méthodologie de mise en oeuvre du krigeage et en utilisant le krigeage pour résoudre une problématique d'interpolation spatiale. Au fil des chapitres, le krigeage a donc été

défini, approfondi, puis illustré. C'est ainsi que les objectifs du mémoire de comprendre et de savoir utiliser le krigeage ont été rencontrés.

Il ressort de ce mémoire que le krigeage repose sur des bases mathématiques solides. Ces bases sont plus rigoureuses que celles de la plupart des autres méthodes d'interpolation. La mise en oeuvre du krigeage est cependant complexifiée par l'étape de l'analyse variographique. Cette étape, lorsqu'effectuée tel que recommandé, requiert l'intervention humaine. Dans une problématique telle que celle traitée au chapitre 6 où l'interpolation doit être complètement automatisée, le krigeage pose donc un problème méthodologique. La solution proposée ici pour résoudre ce problème est d'utiliser la validation croisée pour sélectionner le modèle variographique. Par contre, les résultats du chapitre 6 montrent que la complexité du krigeage n'est pas nécessairement garante de meilleures performances en pratique par rapport à des méthodes d'interpolation plus simples.

7.3 Travaux futurs

Afin de poursuivre les travaux de recherche présentés dans ce mémoire, les méthodes bayésiennes d'interpolation spatiale ainsi que les différentes méthodes d'interpolation spatio-temporelle pourraient être étudiées. Toutes ces méthodes seraient particulièrement intéressante à utiliser pour résoudre la problématique d'interpolation de données de précipitations traitée au chapitre 6. L'interpolation spatio-temporelle surpasserait possiblement l'interpolation spatiale pour de fines résolutions temporelles. En effet, dans l'analyse d'observations d'un phénomène naturel évoluant dans le temps, plus les intervalles de temps entre les collectes de données sont courts, plus la dépendance temporelle entre les données a tendance à être forte. Étant donné que les outils de collecte de données travaillent à des résolutions temporelles de plus en plus fines, la dépendance temporelle entre les observations est de plus en plus présente dans les données. Kyriakidis et Journel (1999) proposent une revue très complète des modèles géostatistiques permettant d'exploiter la dimension temporelle d'un jeu de données.

Bibliographie

Abtew, W., Obeysekera, J. et Shih, G. (1985). Spatial analysis for monthly rainfall in south florida. *Water Resources Bulletin*, 29(2):179–188.

Ali, A. (1998). Nonparametric spatial rainfall characterization using adaptative kernel estimator. *Journal of Geographic Information and Decision Analysis*, 2(2):34–43. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Arnaud, M. et Emery, X. (2000). *Estimation et interpolation spatiale*. Hermes Science Publications, Paris.

Asli, M. et Marcotte, D. (1995). Comparison of approaches to spatial estimation in a bivariate context. *Mathematical Geology*, 27(5):641–658.

Atkinson, P. M. et Lloyd, C. D. (1998). Mapping precipitation in switzerland with ordinary and indicator kriging. *Journal of Geographic Information and Decision Analysis*, 2(2):65–76. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Baillargeon, S., Pouliot, J., Rivest, L.-P., Fortin, V. et Fitzback, J. (2004). Interpolation statistique multivariable de données de précipitations dans un cadre de modélisation hydrologique. *Dans les actes du colloque national Géomatique 2004 de l'Association canadienne des sciences géomatiques*, Montréal, 27 et 28 octobre 2004. Disponible en ligne : http://www.geomatique2004.com/fr/cnt/conferences/popup/10327.pdf (Page consultée le 28 avril 2005).

Banerjee, S., Carlin, B. et Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, New York.

Box, G. E. P. et Cox, D. R. (1964). An analysis of transformations (avec discussions). *Journal of the Royal Statistical Society*, 26:211–252.

Bruno, R. et Capicotto, B. M. (1998). Geostatistical analysis of pluviometric data : Irf-k approach. *Journal of Geographic Information and Decision Analysis*, 2(2):127–

138. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Bryan, B. et Adams, J. (2002). Three-dimensional neurointerpolation of annual mean precipitation and temperature surfaces for china. *Geographical Analysis*, 34(3):94–111. Disponible en ligne : http://www.gisca.adelaide.edu.au/~bbryan/papers/bryan&adams_2002_geog_anal_nnets.pdf (Page consultée le 28 avril 2005).

Burrough, P. et McDonnell, R. (1998). *Principles of Geographical Information Systems*. Oxford University Press, New York.

CEHQ (2004). Centre d'expertise hydrique du Québec : Prévisions hydrologiques et hydrauliques. [En ligne], http://www.cehq.gouv.qc.ca/hydrometrie/prevision/index.htm (Page consultée le 28 avril 2005).

Chauvet, P. (1999). *Aide mémoire de la géostatistique linéaire*. Cahiers de Géostatistique, Fascicule 2. Ecole Nationale Supérieur des Mines de Paris, Centre de Géostatistique, Fontainebleau.

Chegini, E., Mahdian, M., Bandarabadi, S. et Mahdavi, M. (2001). Survey on application of geostatistical methods for estimation of rainfall in arid and semiarid regions in south west of iran. *Dans Proceedings of the 6th International Conference on GeoComputation*, University of Queensland, Brisbane, Australia.

Christakos, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research*, 20(2):251–265.

Cleveland, W. et Devlin, S. (1988). Locally weighted regression : An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cleveland, W. S. et Loader, C. (1996). Smoothing by local regression : Principles and methods. *Dans* Härdle, W. et Schimek, M. G., éditeurs : *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49. Physica Verlag, Heidelberg. Disponible en ligne : http://cm.bell-labs.com/cm/ms/departments/sia/doc/smoothing.springer.pdf (Page consultée le 28 avril 2005).

Cressie, N. (1984). Towards resistant geostatistics. *Dans* Verly, G., David, M., Journel, A. et Marechal, A., éditeurs : *Geostatistics for Natural Resources Characterization, Part 1*, pages 21–44. Reidel, Dordrecht.

Cressie, N. (1986). Kriging nonstationary data. *Journal of the American Statistical Association*, 81(395):625–634.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22(3):239–252.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Creutin, J. et Obled, C. (1982). Objective analysis and mapping techniques for rainfall fields : An objective comparison. *Water Resources Research*, 18(2):413–431.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M. et Staniforth, A. (1998). The operational cmc-mrb global environmental multiscale (gem) model : Part i - design considerations and formulation. *Monthly Weather Review*, 126:1373–1395.

Davis, B. (1987). Uses and abuses of cross-validation in geostatistics. *Mathematical Geology*, 19(3):241–248.

De Olivieira, V., Kedem, B. et Short, D. (1997). Baysien prediction of transformed gaussian random fields. *Journal of the American Statistical Association*, 92:1422–1433.

Demyanov, V., Kanevski, M., Chernov, S., Savelieva, E. et Timonin, V. (1998). Neural network residual kriging application for climatic data. *Journal of Geographic Information and Decision Analysis*, 2(2):215–232. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Deutsch, C. et Journel, A. (1998). *GSLIB : Geostatistical Software Library and User's Guide, 2nd edition*. Applied Geostatistics Series. Oxford University Press, New York. second edition.

Dirks, K. N., Hay, J. E., Stow, C. D. et Harris, D. (1998). High-resolution studies of rainfall on norfolk island. part ii : The interpolation of high-spatial-resolution rainfall data on norfolk island. *Journal of Hydrology*, 208:187–193.

Duchon, J. (1975). Fonctions-spline du type plaque mince en dimension 2. Rapport technique 231, Université de Grenoble.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *Revue Française d'Automatique et de Recherche Opérationnelle (R.A.I.R.O.) Analyse numérique*, 10(R-3):5–12.

Fotheringham, A., Brunsdon, C. et Charlton, M. (2002). *Geographically Weighted Regression : the analysis of spatially varying relationships*. John Wiley & Sons Ltd, Chichester.

Fuentes, M. (2000). Fitting algorithms for the semivariogram, and examples. Notes du cours ST 810M : Spatial Statistics, Lectures 4 à 8, North Carolina State University, Raleigh. Disponibles en ligne : http://www4.stat.ncsu.edu/~fuentes/st810/lectures/lectures.html (Page consultée le 28 avril 2005).

GANDIN, L. (1963). *Objective analysis of meteorological fields*. Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad. (traduit en anglais par Israel Program for Scientific Translations, Jerusalem, 1965).

GAUDARD, M., KARSON, M., LINDER, E. et SINHA, D. (1999). Bayesian spatial prediction. *Environmental and Ecological Statistics*, 6:147–171.

GOOVAERTS, P. (1997). *Geostatistics for Natural Ressources Evaluation*. Applied Geostatistics Series. Oxford University Press, New York.

GOOVAERTS, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129.

GRATTON, Y. (2002). Le krigeage : La méthode optimale d'interpolation spatiale. *Les Articles de l'Institut d'Analyse Géographique*. Disponible en ligne : http://www.iag.asso.fr/pdf/krigeage_juillet2002.pdf (Page consultée le 28 avril 2005).

HABERLANDT, H. et KITE, G. (1998). Estimation of daily space-time precipitation series for macroscale hydrological modelling. *Hydrological Processes*, 12:1419–1432.

HARDY, R. L. (1971). Multiquadratic equations of topography and other irregular surfaces. *Journal of Geophysical Research*, 76:1905–1915.

HASTIE, T. J. et TIBSHIRANI, R. J. (1990). *Generalized additive models*, volume 43 de *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

HAWKINS, D. M. et CRESSIE, N. (1984). Robust kriging—a proposal. *J. Internat. Assoc. Math. Geol.*, 16(1):3–18.

HELSON, H. et LOWDENSLAGER, D. (1958). Prediction theory and Fourier series in several variables. *Acta Mathematica*, 99:165–202.

HELSON, H. et LOWDENSLAGER, D. (1961). Prediction theory and Fourier series in several variables. II. *Acta Mathematica*, 106:175–213.

HENGL, T., GEUVELINK, G. et STEIN, A. (2003). Comparison of kriging with external drift and regression-kriging. *Technical note, ITC*. Disponible en ligne : http://www.itc.nl/library/Papers_2003/misca/hengl_comparison.pdf (Page consultée le 28 avril 2005).

HESSAMI, M. (2002). *Comparison of soft computing systems for the post-calibration of weather radar*. Thèse de doctorat, Université Laval, Québec.

HIDA, T. (1980). *Brownian Motion*. Sringer-Verlgas, New York.

Hoef, J. et Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology*, 25(2):219–240.

Hutchinson, M. F. (1991). The application of thin plate smoothing splines to continent-wide data assimilation. *Dans* Jasper, J., éditeur : *Data Assimilation Systems*, pages 104–113. Bureau of Meteorology Research Report No. 27, Bureau of Meteorology, Melbourne.

Hutchinson, M. F. (1998a). Interpolation of rainfall data with thin plate smoothing splines : I. two dimensional smoothing of data with short range correlation. *Journal of Geographic Information and Decision Analysis*, 2(2):139–151. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Hutchinson, M. F. (1998b). Interpolation of rainfall data with thin plate smoothing splines : Ii. analysis of topographic dependence. *Journal of Geographic Information and Decision Analysis*, 2(2):152–167. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

INRS-ETE (1998). Institut national de la recherche scientifique - Eau, Terre et Environnement : Le modèle hydrologique Hydrotel. [En ligne], http://www.inrs-ete.uquebec.ca/activites/modeles/hydrotel/fr/accueil.htm (Page consultée le 28 avril 2005).

Insightful (2004). Statistics software and data mining software. [En ligne], http://www.insightful.com/ (Page consultée le 28 avril 2005).

Isaaks, E. et Srivastava, R. (1989). *Applied Geostatistics*. Oxford University Press, New York.

Johns, C., Nychka, D., Kittel, T. et Daly, C. (2001). Infilling sparse records of precipitation fields. *Journal of the American Statistical Association*, 98(464):796–806.

Journel, A. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15(3):445–469.

Khuri, A. (1993). *Advanced Calculus with Applications in Statistics*. John Wiley & Sons, Inc., New York.

Klinkenberg, B. et Waters, N. (1990). Spatial interpolation I. *Dans* Goodchild, M. et Kemp, K., éditeurs : *NCGIA Core Curriculum in GIS*. National Center for Geographic Information and Analysis, University of California, Santa Barbara CA. [En ligne], http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u40.html (Page consultée le 28 avril 2005).

KOZINTSEVA, A. (1999). Comparison of three methods of spatial prediction. Mémoire de maîtrise, University of Maryland, College Park. Disponible en ligne : http://techreports.isr.umd.edu/reports/1999/MS_99-12.pdf (Page consultée le 28 avril 2005).

KRIGE, D. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society*, 52:119–139.

KRUIZINGA, S. et YPERLAAN, G. (1978). Spatial interpolation of daily totals of rainfall. *Journal of Hydrology*, 36:65–73.

KYRIAKIDIS, P. C. et JOURNEL, A. G. (1999). Geostatistical space-time models : a review. *Mathematical Geology*, 31(6):651–684.

LEE, S., CHO, S. et WONG, P. M. (1998). Rainfall prediction using artificial neural networks. *Journal of Geographic Information and Decision Analysis*, 2(2):233–242. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

LOADER, C. (1999). *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York.

LOADER, C. (2004). Smoothing : Local regression techniques. *Dans* GENTLE, J., HÄRDLE, W. et MORI, Y., éditeurs : *Handbook of Computational Statistics*, pages 539–564. Springer-Verlag, New York. Disponible en ligne : http://www.herine.net/stat/papers/hbcs.pdf (Page consultée le 28 avril 2005).

MARCOTTE, D. (1988). Trend surface analysis as a special case of irf-k kriging. *Mathematical Geology*, 20(7):821–824.

MARCOTTE, D. (1995). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, 27(5):659–672.

MATHERON, G. (1962). *Traité de géostatistique appliquée, Tome I*. Mémoires du Bureau de Recherches Géologiques et Minières, No.14. Editions Technip, Paris.

MATHERON, G. (1963a). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

MATHERON, G. (1963b). *Traité de géostatistique appliquée, II : Le Krigeage*. Mémoires du Bureau de Recherches Géologiques et Minières, No.24. Editions B. R. G. M., Paris.

MATHERON, G. (1965). *Les variables régionalisées et leur estimation*. Masson, Paris.

MATHERON, G. (1969). *Le krigeage universel*. Les cahiers du Centre de morphologie mathématique de Fontainebleau, Fascicule 1. École de Mines de Paris, Fontainebleau.

MATHERON, G. (1970). *La théorie des variables régionalisées, et ses applications*. Les cahiers du Centre de morphologie mathématique de Fontainebleau, Fascicule 5. École des Mines de Paris, Fontainebleau.

MATHERON, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Pobability*, 5:439–468.

MATHERON, G. (1976). A simple substitute for conditional expectation : The disjunctive kriging. *Dans* GUARASCIO, M., DAVID, M. et HUIJBREGTS, C., éditeurs : *Advanced Goestatistics in the Mining Industry*, pages 221–236. Reidel, Dordrecht.

MITÁŠOVÁ, H. et MITÁŠ, L. (1993). Interpolation by regularized spline with tension : I. theory and implementation. *Mathematical Geology*, 25(6):641–655.

MYERS, D. (1993). Cross-validation and variogram estimation. *Theory of Probability and its Applications*, 37:345–347.

MYERS, D. (1994). Spatial interpolation : an overview. *Geoderma*, 62:17–28.

NALDER, I. A. et WEIN, R. W. (1998). Spatial interpolation of climatic normals : test of a new method in the canadien boreal forest. *Agricultural and Forest Meteorology*, 92:211–225.

OKABE, A., BOOTS, B. et SUGIHARA, K. (1992). *Spatial tessellations : concepts and applications of Voronoĭ diagrams*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester. With a foreword by D. G. Kendall.

OLIVER, M. (2001). Determining the spatial scale of variation in environmental properties using the variogram. *Dans* TATE, N. J. et ATKINSON, P. M., éditeurs : *Modeling Scale in Geographical Information Science*, pages 193–219. Jonh Wiley & Sons, Inc., New York.

OMRE, H. (1987). Bayesian kriging - merging observations and qualified guesses in kriging. *Mathematical Geology*, 19(1):25–39.

PEBESMA, E. (2004a). Multivariable geostatistics in s : the gstat package. *Computers & Geosciences*, 30:683–691. Disponible en ligne : http://www.geog.uu.nl/~pebesma/publ/candg2004.pdf (Page consultée le 28 avril 2005).

PEBESMA, E. J. (1999). Gstat user's manual. [En ligne], http://www.gstat.org/gstat.pdf (Page consultée le 28 avril 2005).

PEBESMA, E. J. (2004b). Package 'gstat'. [En ligne], http://www.gstat.org/bin/gstat.pdf (Page consultée le 28 avril 2005).

Pebesma, E. J. et Wesseling, C. G. (1998). Gstat : a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences*, 24(1):17–31. Disponible en ligne : http://www.geog.uu.nl/~pebesma/publ/gstatcg.pdf (Page consultée le 28 avril 2005).

Phillips, D. L., Dolph, J. et Marks, D. (1992). A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain. *Agricultural and Forest Meteorology*, 58:119–141.

R project (2004). The r project for statistical computing. [En ligne], http://www.r-project.org/ (Page consultée le 28 avril 2005).

Rajagopalan, B. et Lall, U. (1998). Locally weighted polynomial estimation of spatial precipitation. *Journal of Geographic Information and Decision Analysis*, 2(2): 44–51. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Rencher, A. C. (1995). *Methods of multivariate analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York.

Rigol, J. P., Jarvis, C. H. et Stuart, N. (2001). Artificial neural networks as a tool for spatial interpolation. *International Journal of GIS*, 15(4):323–343.

Ripley, B. D. (1981). *Spatial statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Rivoirard, J. (1990). Review of lognormal estimation for in situ reserves. *Mathematical Geology*, 22(2):213–221.

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

Roth, C. (1998). Is lognormal kriging suitable for local estimation ? *Mathematical Geology*, 30(8):999–1009.

Saveliev, A. A., Mucharamova, S. S. et Piliugin, G. A. (1998). Modeling of the daily rainfall values using surface under tension and kriging. *Journal of Geographic Information and Decision Analysis*, 2(2):52–64. Disponible en ligne : http://www.geodec.org/gida_4.htm (Page consultée le 28 avril 2005).

Sibson, R. (1981). A brief description of natural neighbour interpolation. *Dans* Barnett, V., éditeur : *Interpreting Multivariate Data*, Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics, pages 21–36. John Wiley & Sons Inc., New York.

SMC (2002). Service Météorologique du Canada. [En ligne], version actuelle de GEM http://www.msc-smc.ec.gc.ca/cmc/op_systems/regional_forecast_f.html, ancienne version de GEM http://www.msc-smc.ec.gc.ca/cmc/GEWEX/regional_forecast_f.html , présentation de GEM http://www.cmc.ec.gc.ca/rpn/gef_html_public/ (Pages consultées le 28 avril 2005).

Syed, K., Goodrich, D., Myers, D. et Sorooshian, S. (2003). Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. *Journal of Hydrology*, 271:1–21.

Tabios, G. Q. et Salas, J. D. (1985). A comparative analysis of techniques for spatial interpolation of precipitation. *Water Resources Bulletin*, 21(3):365–380.

Velarde, L., Migon, H. et Pereira, B. (2004). Space-time modeling of rainfall data. *Environmetrics*, 15(6):561–576.

Wackernagel, H. (2003). *Multivariate Geostatistics : an Introduction with Applications*. Springer-Verlag, Berlin. Third completely revised edition.

Wahba, G. (1990). *Spline models for observational data*, volume 59 de *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wand, M. P. et Jones, M. C. (1995). *Kernel smoothing*, volume 60 de *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Webster, R. et Oliver, M. (1993). How large a sample is needed to estimate the regional variogram adequately ? *Dans* Soares, A., éditeur : *Geostatistics Tròia '92*, pages 155–166, Dordrecht. Kluwer Academic Publishers.

Yang, D., Goodison, B., Metcalfe, J., Louie, P., Leavesley, G., Emerson, D., Golubev, V., Elomaa, E., Gunther, T., Hanson, C., Pangburn, T., Kang, E. et Milkovic, J. (1999). Quantification of precipitation measurement discontinuity induced by wind shields on national gauge. *Water Resources Research*, 35(2):491–508. Disponible en ligne : http://www.uaf.edu/water/faculty/yang/1998WR900042.pdf (Page consultée le 28 avril 2005).

Zimmerman, D., Pavlik, C., Ruggles, A. et Armstrong, M. P. (1999). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, 31(4):375–390.

Annexe A

Complément d'analyse des résultats du chapitre 6

Dans cette annexe, quelques résultats supplémentaires intéressants relatifs aux travaux effectués dans le chapitre 6 sont présentés. Tout d'abord, le tableau A.1 indique le nombre de fois que chaque fraction de voisinage a été choisie par la validation croisée pour les deux types de régression locale effectués. Pour la régression locale incluant la variable auxiliaire, seules les fractions 0.8, 0.9 et 1 ont été considérées car le modèle contient un paramètre de plus qu'en régression locale univariée. Un plus grand nombre de données sont donc requises dans l'ajustement du modèle. Probablement en raison du petit nombre d'observations analysées, la fraction choisie le plus fréquemment est 1 pour les deux types de régression locale.

Variables explicatives de la régression locale	**Fraction de voisinage**							**Total**
	0.4	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
coordonnées x et y	16 0.17	8 0.09	3 0.03	2 0.02	3 0.03	10 0.11	50 0.54	92
coordonnées x, y et variable auxiliaire w	-	-	-	-	11 0.17	13 0.21	39 0.62	63

TAB. A.1 – *Fréquences de sélection des fractions de voisinage en régression locale*

Les fréquences de sélection des modèles variographiques par la validation croisée sont présentées dans le tableau A.2 en fonction du type de krigeage. Le modèle pépitique est celui le plus souvent choisi pour les trois types de krigeage. Afin de mieux comprendre ce résultat, regardons les fréquences de sélection des modèles en fonction de l'intensité

Type de krigeage	**Modèle variographique**							**Total**
	Pépitique	**Linéaire avec palier**	**Sphérique**	**Exponentiel**	**Gaussien**	**Linéaire sans palier**	**Puissance**	
ordinaire	36 0.39	20 0.22	8 0.09	4 0.04	15 0.16	6 0.07	3 0.03	92
universel	36 0.39	23 0.25	6 0.07	8 0.09	8 0.09	9 0.10	2 0.02	92
avec dérive externe	22 0.35	16 0.25	2 0.03	3 0.05	15 0.25	4 0.06	1 0.02	63

TAB. A.2 – *Fréquences de sélection des modèles variographiques en krigeage ordinaire, universel et avec modèle de tendance*

des précipitations. Le tableau A.3 présente ces fréquences pour le krigeage ordinaire. On remarque que le modèle pépitique est surtout retenu lorsque les précipitations sont de faible ou moyenne intensité. Lorsque les précipitations moyennes sur le bassin versant sont supérieures à 1 mm, les modèles les plus souvent sélectionnés sont plutôt le modèle linéaire avec palier et le modèle gaussien.

Classe de périodes	**Modèle variographique**							**Total**
	Pépitique	**Linéaire avec palier**	**Sphérique**	**Exponentiel**	**Gaussien**	**Linéaire sans palier**	**Puissance**	
Précipitations moyennes inférieures à 0.1 mm	24 0.65	5 0.14	2 0.05	0 0.00	4 0.11	2 0.05	0 0	37
Précipitations moyennes entre 0.1 mm et 1 mm	11 0.33	7 0.21	5 0.15	2 0.06	5 0.15	1 0.03	2 0.06	33
Précipitations moyennes supérieures à 1 mm	1 0.05	8 0.36	1 0.05	2 0.09	6 0.27	3 0.14	1 0.05	22
Total	36	20	8	4	15	6	3	92

TAB. A.3 – *Fréquences de sélection des modèles variographiques en krigeage ordinaire en fonction de l'intensité des précipitations*

L'analyse des résultats en fonction de l'intensité des précipitations apporte beaucoup d'informations. La figure A.1 illustre sous forme de diagrammes en boîte la distribution

des erreurs de validation croisée par intensité de précipitations pour la méthode témoin de l'inverse de la distance et pour la méthode Sélection qui a été développée. Bien que la méthode Sélection ne semble pas donner des prévisions vraiment plus justes que celles générées par la méthode de l'inverse de la distance lorsque les précipitations sont de faible ou moyenne intensité, on note une amélioration des prévisions par la méthode Sélection dans le cas de précipitations plus fortes. En effet, pour des précipitations moyennes sur le bassin versant supérieures à 1 mm, le diagramme en boîte des erreurs de validation croisée de la méthode Sélection a des moustaches plus courtes et une boîte moins longue que le diagramme en boîte pour la méthode de l'inverse de la distance. Dans une perspective de prévision hydrologique, ce sont ces périodes de fortes précipitations qui ont le plus d'importance. Ainsi, la méthode Sélection proposée améliore l'interpolation spatiale comparativement à la méthode de l'inverse de la distance lorsque ça importe le plus.

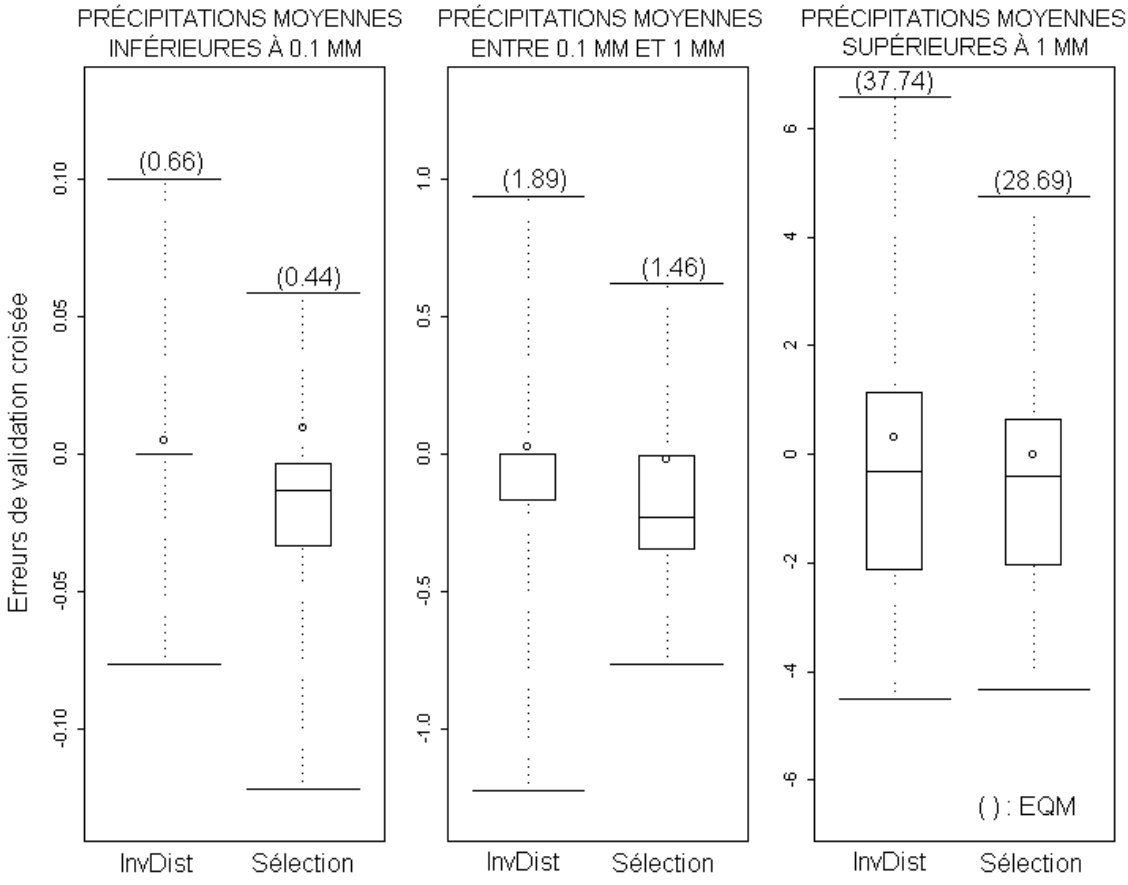

FIG. A.1 – *Diagrammes en boîte des erreurs de validation croisée catégorisées selon l'intensité des précipitations pour la méthode de l'inverse de la distance et la méthode Sélection*

En supposant l'indépendance temporelle entre les précipitations aux différentes périodes, il est possible d'effectuer un test statistique de comparaison des erreurs quadratiques moyennes (EQM) de validation croisée entre la méthode témoin de l'inverse de la distance et la méthode Sélection. Étant donné que la normalité de ces EQM est douteuse, des tests non paramétriques sont plus appropriés. Voici donc, dans le tableau A.4, des tests de rangs signés de Wilcoxon par station pour comparer les deux méthodes selon l'intensité des précipitations. Une valeur positive de la statistique de Wilcoxon

Station	Périodes de précipitations moyennes inférieures à 0.1 mm		Périodes de précipitations moyennes entre 0.1 mm et 1 mm		Périodes de précipitations moyennes supérieures à 1 mm		Toutes les périodes	
	Stat	**P-valeur**	**Stat**	**P-valeur**	**Stat**	**P-valeur**	**Stat**	**P-valeur**
Mercier Amont	-3,11	0,00**	-1,88	0,06	1,77	0,08	-1,54	0,12
Barrière Amont	-1,57	0,12	1,84	0,07	0,35	0,73	0,63	0,53
Cabonga Amont	-1,68	0,09	-0,17	0,87	1,08	0,28	0,25	0,80
Dozois Est	-2,01	0,04**	2,88	0,00**	2,05	0,04**	2,26	0,02**
Domaine	-3,76	0,00**	0,12	0,91	0,62	0,54	-0,74	0,46
Brodtkorb	-3,83	0,00**	-0,37	0,71	1,23	0,22	-1,08	0,28
Chouart	-0,94	0,35	-0,96	0,34	-0,37	0,71	-1,69	0,09
La Vérendrye	1,27	0,20	-0,59	0,56	-0,60	0,55	-0,87	0,38
Wapus	-3,76	0,00**	1,08	0,28	1,90	0,06	0,71	0,48
Mont St-Michel	-3,09	0,00**	0,40	0,69	0,38	0,70	-1,41	0,16
La Pêche	0,02	0,99	-3,14	0,00**	1,76	0,08	-0,97	0,33
Maniwaki	-3,18	0,00**	0,98	0,33	1,28	0,20	-0,02	0,99
Ottawa 1	0,26	0,80	-1,65	0,10	-0,28	0,78	-0,97	0,33
Ottawa 2	-0,58	0,56	-2,62	0,01**	0,65	0,52	-1,52	0,13
High Falls	0,25	0,80	-1,70	0,09	1,22	0,22	-0,39	0,70
Parent	0,92	0,36	-1,22	0,22	0,60	0,55	0,00	1,00

** test significatif au seuil de 5%

TAB. A.4 – *Tests des rangs signés de Wilcoxon pour comparer les EQM de validation croisée par station en fonction de l'intensité des précipitations*

signifie que les EQM pour la méthode Sélection sont plus petits que pour la méthode de l'inverse de la distance. Peu de tests sont significatifs au seuil de 5%. Les tests significatifs pour des précipitations de faible intensité indiquent tous que la méthode de l'inverse de la distance performe significativement mieux que la méthode Sélection, tandis que le test significatif pour des précipitations de forte intensité indique le contraire.

Annexe B

Programme S-Plus

Voici le programme S-Plus qui a permis d'interpoler les données pour chaque période de 6 heures d'août 2003 par la méthode de l'inverse de la distance, la régression locale par rapport à x et y, la régression locale par rapport à x, y et les données GEM, le krigeage ordinaire, le krigeage universel, le krigeage avec dérive externe et le cokrigeage ordinaire. Ce programme appelle les fichiers *Donnees* et *Grille*. Le premier fichier contient les valeurs régionalisées mesurées aux stations et générées par le modèle GEM pour toutes les périodes du mois d'août 2003 et tous les sites d'observation. Les coordonnées spatiales UTM sur la zone 18T en km des sites d'observation sont aussi incluses dans le fichier. Le deuxième fichier comprend plutôt les coordonnées spatiales UTM des sites de prévision ainsi que les données fournies par GEM en ces sites. Afin de décrire la forme de ces fichiers, les tableaux B.1 et B.2 en contiennent des extraits.

Notons que ce programme SPLUS est fonctionnel, mais non optimisé. Il est donc assez long et ce principalement en raison de la technique de sélection des modèles variographiques par validation croisée utilisée.

Periode	**x**	**y**	**Stations**	**GEM**
1	384.024	5209.559	0.00	0.00
1	373.293	5158.051	0.00	0.00
1	456.994	5298.090	0.00	0.00
1	507.430	5250.337	0.00	0.00
⋮	⋮	⋮	⋮	⋮

Tab. B.1 – *Extrait du fichier nommé Donnees*

Periode	x	y	GEM
1	430.000	5050.000	0.00
1	440.000	5050.000	0.00
1	410.000	5060.000	0.00
1	420.000	5060.000	0.00
⋮	⋮	⋮	⋮

TAB. B.2 – *Extrait du fichier nommé Grille*

Programme :

```
# INITIALISATION des jeux de données qui contiendront les résultats :

# Résidus de validation croisée :
Err=data.frame(matrix(rep(NA,dim(Donnees)[1]*32),nrow=dim(Donnees)[1]))
dimnames(Err)<-list(1:dim(Err)[1],c("Periode","x","y","InvDist","KO1","KO2","KO3","KO4","KO5",
               "KO6","KO7","KO","KU1","KU2","KU3","KU4","KU5","KU6","KU7","KU","KDE1","KDE2",
               "KDE3","KDE4","KDE5","KDE6","KDE7","KDE","CKO","RLxy","RLxyw","Selection"))
Err[,1:3]=Donnees[,1:3]

# Prévisions :
Prev=data.frame(matrix(rep(NA,dim(Grille)[1]*32),nrow=dim(Grille)[1]))
dimnames(Prev)<-list(1:dim(Prev)[1],c("Periode","x","y","InvDist","KO1","KO2","KO3","KO4","KO5",
                "KO6","KO7","KO","KU1","KU2","KU3","KU4","KU5","KU6","KU7","KU","KDE1","KDE2",
                "KDE3","KDE4","KDE5","KDE6","KDE7","KDE","CKO","RLxy","RLxyw","Selection"))
Prev[,1:3]=Grille[,1:3]

# Variances de prévision :
Var=data.frame(matrix(rep(NA,dim(Grille)[1]*32),nrow=dim(Grille)[1]))
dimnames(Var)<-list(1:dim(Var)[1],c("Periode","x","y","InvDist","KO1","KO2","KO3","KO4","KO5",
               "KO6","KO7","KO","KU1","KU2","KU3","KU4","KU5","KU6","KU7","KU","KDE1","KDE2",
               "KDE3","KDE4","KDE5","KDE6","KDE7","KDE","CKO","RLxy","RLxyw","Selection"))
Var[,1:3]=Grille[,1:3]

# Informations Diverses :
Info=data.frame(matrix(rep(NA,123*137),nrow=123))
dimnames(Info)<-list(1:dim(Info)[1],c("NbStations","Nbplus0","Classe","MoyStations",
  "EtypeStations","MoyInvDist","EtypeInvDist","MoyGEM","EtypeGEM","InvDistEQM","InvDistEAM",
  "KO1pepite","KO1EQM","KO1EAM","KO2pepite","KO2ppalier","KO2portee","KO2EQM","KO2EAM",
  "KO3pepite","KO3ppalier","KO3portee","KO3EQM","KO3EAM","KO4pepite","KO4ppalier","KO4porteep",
  "KO4EQM","KO4EAM","KO5pepite","KO5ppalier","KO5porteep","KO5EQM","KO5EAM","KO6pepite",
  "KO6pente","KO6EQM","KO6EAM","KO7pepite","KO7echelle","KO7puissance","KO7EQM","KO7EAM",
  "KOmodele","KOEQM","KOEAM","KUcorrX","KUpvalueX","KUcorrY","KUpvalueY","KU1pepite","KU1EQM",
  "KU1EAM","KU2pepite","KU2ppalier","KU2portee","KU2EQM","KU2EAM","KU3pepite","KU3ppalier",
  "KU3portee","KU3EQM","KU3EAM","KU4pepite","KU4ppalier","KU4porteep","KU4EQM","KU4EAM",
  "KU5pepite","KU5ppalier","KU5porteep","KU5EQM","KU5EAM","KU6pepite","KU6pente","KU6EQM",
  "KU6EAM","KU7pepite","KU7echelle","KU7puissance","KU7EQM","KU7EAM","KUmodele","KUEQM","KUEAM",
  "KDEpossible","KDEcorrGEM","KDEpvalueGEM","KDE1pepite","KDE1EQM","KDE1EAM","KDE2pepite",
  "KDE2ppalier","KDE2portee","KDE2EQM","KDE2EAM","KDE3pepite","KDE3ppalier","KDE3portee",
  "KDE3EQM","KDE3EAM","KDE4pepite","KDE4ppalier","KDE4porteep","KDE4EQM","KDE4EAM","KDE5pepite",
  "KDE5ppalier","KDE5porteep","KDE5EQM","KDE5EAM","KDE6pepite","KDE6pente","KDE6EQM","KDE6EAM",
```

```
  "KDE7pepite","KDE7echelle","KDE7puissance","KDE7EQM","KDE7EAM","KDEmodele","KDEEQM","KDEEAM",
  "CKOpenteStations","CKOpenteGEM","CKOpenteStaGEM","CKOEQM","CKOEAM","RLxyfen","RLxyEQM",
  "RLxyEAM","RLxywfen","RLxywEQM","RLxywEAM","Selection","SelectionEQM","SelectionEAM"))

# Ajout d'informations dans le jeu de données Info :
for(t in 1:123){
  Info$NbStations[t]=dim(Donnees[Donnees$Periode==t,])[1]
  Info$Nbplus0[t]=sum(ifelse(Donnees$Stations[Donnees$Periode==t]>0,1,0),na.rm=T)
  Info$MoyStations[t]=mean(Donnees$Stations[Donnees$Periode==t],na.rm=T)
  Info$EtypeStations[t]=stdev(Donnees$Stations[Donnees$Periode==t],na.rm=T)
  Info$MoyGEM[t]=mean(Grille$GEM[Grille$Periode==t],na.rm=T)
  Info$EtypeGEM[t]=stdev(Grille$GEM[Grille$Periode==t],na.rm=T)
  Info$Classe[t]=ifelse(Info$MoyStations[t]==0,0,NA)
  }

library(gstat)


# FILTRE : Mise de côté des périodes de temps pour lesquelles aucune station n'a mesuré de pluie
Pluie=na.omit(ifelse(Info$Classe=="NA",row(Info),NA))


# INVERSE DE LA DISTANCE

for (t in Pluie) {

  Periodet=na.omit(ifelse(Donnees$Periode==t,row(Donnees),NA))

  # Validation croisée :
  for (i in 1:length(Periodet)){
    distances=sqrt((Donnees$x[Periodet[i]]-Donnees$x[Periodet[-i]])**2
              +(Donnees$y[Periodet[i]]-Donnees$y[Periodet[-i]])**2)
    voisinage=na.omit(ifelse(rank(distances)<=3,Periodet[-i],NA))
    distances=na.omit(ifelse(rank(distances)<=3,distances,NA))
    poids=as.vector((1/distances)/sum(1/distances))
    valid=sum(poids*Donnees$Stations[voisinage],na.rm=T)
    Err$InvDist[Periodet[i]]=Donnees$Stations[Periodet[i]]-valid
    }

  # Prévision sur la grille GEM :
  GrillePeriodet=na.omit(ifelse(Grille$Periode==t,row(Grille),NA))
  for (j in GrillePeriodet){
    distances=sqrt((Grille$x[j]-Donnees$x[Periodet])**2+(Grille$y[j]-Donnees$y[Periodet])**2)
    voisinage=na.omit(ifelse(rank(distances)<=3,Periodet,NA))
    distances=na.omit(ifelse(rank(distances)<=3,distances,NA))
    poids=as.vector((1/distances)/sum(1/distances))
    Prev$InvDist[j]=sum(poids*Donnees$Stations[voisinage],na.rm=T)
    }
  Info$InvDistEQM[t]=mean((Err$InvDist[Err$Periode==t])**2,na.rm=T)
  Info$InvDistEAM[t]=median(abs(Err$InvDist[Err$Periode==t]),na.rm=T)
  Info$MoyInvDist[t]=mean(Prev$InvDist[Prev$Periode==t],na.rm=T)
  Info$EtypeInvDist[t]=stdev(Prev$InvDist[Prev$Periode==t],na.rm=T)
  Info$Classe[t]=ifelse(Info$MoyInvDist[t]>1,">1",
                 ifelse(Info$MoyInvDist[t]<=0.1,"(0,0.1]","(0.1,1]"))
  }
```

```
# RÉGRESSION LOCALE PAR RAPPORT À X ET Y

ErrValid=data.frame(matrix(rep(NA,dim(Donnees)[1]*7),nrow=dim(Donnees)[1]))
for (t in Pluie) {

  # Sélection de la fraction du voisinage dans la fenêtre par validation croisée :
  Periodet=na.omit(ifelse(Donnees$Periode==t,row(Donnees),NA))
  j=1
  for(a in c(1,0.9,0.8,0.7,0.6,0.5,0.4)){
    k=trunc(a*Info$NbStations[t])
    for (i in 1:length(Periodet)){
      distances=sqrt((Donnees$x[Periodet[i]]-Donnees$x[Periodet[-i]])**2
                     +(Donnees$y[Periodet[i]]-Donnees$y[Periodet[-i]])**2)
      voisinage=na.omit(ifelse(rank(distances)<=k,Periodet[-i],NA))
      distances=na.omit(ifelse(rank(distances)<=k,distances,NA))
      poids=as.vector(1-(distances/ceiling(max(distances)))**2)
      reg=lm(Stations~1+x+y,data=Donnees[voisinage,],weights=poids)
      valid=predict(reg,Donnees[Periodet[i],])
      ErrValid[Periodet[i],j]=Donnees$Stations[Periodet[i]]-ifelse(valid<0,0,valid)
      }
    j=j+1
    }
  indice=data.frame(apply(ErrValid[Periodet,]**2,2,mean))
  minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
  Info$RLxyfen[t]= if (minimums[1,1]==1) 1 else if (minimums[1,1]==2) 0.9 else
                   if (minimums[1,1]==3) 0.8 else if (minimums[1,1]==4) 0.7 else
                   if (minimums[1,1]==5) 0.6 else if (minimums[1,1]==6) 0.5 else 0.4
  Err$RLxy[Periodet]=ErrValid[Periodet,minimums[1,1]]
  Info$RLxyEQM[t]=mean((Err$RLxy[Err$Periode==t])**2,na.rm=T)
  Info$RLxyEAM[t]=median(abs(Err$RLxy[Err$Periode==t]),na.rm=T)

  # Prévision grille GEM :
  GrillePeriodet=na.omit(ifelse(Grille$Periode==t,row(Grille),NA))
  k=trunc(Info$RLxyfen[t]*Info$NbStations[t])
  for (j in GrillePeriodet){
    distances=sqrt((Grille$x[j]-Donnees$x[Periodet])**2+(Grille$y[j]-Donnees$y[Periodet])**2)
    voisinage=na.omit(ifelse(rank(distances)<=k,Periodet,NA))
    distances=na.omit(ifelse(rank(distances)<=k,distances,NA))
    poids=as.vector(1-(distances/ceiling(max(distances)))**2)
    reg=lm(Stations~1+x+y,data=Donnees[voisinage,],weights=poids)
    Interpo=predict(reg,Grille[j,],se.fit=T)
    Prev$RLxy[j]=ifelse(Interpo$fit<0,0,Interpo$fit)
    Var$RLxy[j]=Interpo$se.fit
    }
  }


# RÉGRESSION LOCALE PAR RAPPORT À X, Y ET AUX DONNÉES GEM

ErrValid=data.frame(matrix(rep(NA,dim(Donnees)[1]*3),nrow=dim(Donnees)[1]))
for (t in KDEpossible) {

  # Sélection de la fraction du voisinage dans la fenêtre par validation croisée :
```

```
  Periodet=na.omit(ifelse(Donnees$Periode==t,row(Donnees),NA))
  j=1
  for(a in c(1,0.9,0.8)){
    k=trunc(a*Info$NbStations[t])
    for (i in 1:length(Periodet)){
      distances=sqrt((Donnees$x[Periodet[i]]-Donnees$x[Periodet[-i]])**2+(Donnees$y[Periodet[i]]
                     -Donnees$y[Periodet[-i]])**2)
      voisinage=na.omit(ifelse(rank(distances)<=k,Periodet[-i],NA))
      distances=na.omit(ifelse(rank(distances)<=k,distances,NA))
      poids=as.vector(1-(distances/ceiling(max(distances)))**2)
      reg=lm(Stations~1+x+y+GEM,data=Donnees[voisinage,],weights=poids)
      valid=predict(reg,Donnees[Periodet[i],])
      ErrValid[Periodet[i],j]=Donnees$Stations[Periodet[i]]-ifelse(valid<0,0,valid)
      }
    j=j+1
    }
  indice=data.frame(apply(ErrValid[Periodet,]**2,2,mean))
  minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
  Info$RLxywfen[t]=if(minimums[1,1]==1)1 else if(minimums[1,1]==2)0.9 else if(minimums[1,1]==3)0.8
  Err$RLxyw[Periodet]=ErrValid[Periodet,minimums[1,1]]
  Info$RLxywEQM[t]=mean((Err$RLxyw[Err$Periode==t])**2,na.rm=T)
  Info$RLxywEAM[t]=median(abs(Err$RLxyw[Err$Periode==t]),na.rm=T)

  # Prévision grille GEM :
  GrillePeriodet=na.omit(ifelse(Grille$Periode==t,row(Grille),NA))
  k=trunc(Info$RLxywfen[t]*Info$NbStations[t])
  for (j in GrillePeriodet){
    distances=sqrt((Grille$x[j]-Donnees$x[Periodet])**2+(Grille$y[j]-Donnees$y[Periodet])**2)
    voisinage=na.omit(ifelse(rank(distances)<=k,Periodet,NA))
    distances=na.omit(ifelse(rank(distances)<=k,distances,NA))
    poids=as.vector(1-(distances/ceiling(max(distances)))**2)
    reg=lm(Stations~1+x+y+GEM,data=Donnees[voisinage,],weights=poids)
    Interpo=predict(reg,Grille[j,],se.fit=T)
    Prev$RLxyw[j]=ifelse(Interpo$fit<0,0,Interpo$fit)
    Var$RLxyw[j]=Interpo$se.fit
    }
  }


# KRIGEAGE ORDINAIRE

for(t in Pluie){

  # Pour estimer le variogramme expérimental :
  vario=variogram(Stations~1,~x+y,Donnees[Donnees$Periode==t,],cutoff=150,width=10)

  # 1- Modèle pépitique :
  # Pour ajuster le modèle et noter les info par rapport à ce modèle :
  modele1=fit.variogram(vario,vgm(max(vario[,3]),"Nug"),fit.method=1)
  Info$KO1pepite[t]= modele1$psill[1]
  # Pour faire validation croisée, interpolation, sauver résultats et noter informations :
  if (modele1[dim(modele1)[1],2]>0) {
    valid=krige.cv(Stations~1,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,])
    Err$KO1[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
    Interpo=krige(Stations~1,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,],
```

```
                newdata=Grille[Grille$Periode==t,])
  Prev$KO1[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO1[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO1[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO1[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO1[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KO1EQM[t]=mean((Err$KO1[Err$Periode==t])**2,na.rm=T)
Info$KO1EAM[t]=median(abs(Err$KO1[Err$Periode==t]),na.rm=T)

# 2- Modèle linéaire avec palier :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele2=fit.variogram(vario,vgm(max(vario[,3]),"Lin",140,0.1*max(vario[,3])),fit.method=1)
modele2=if (modele2[1,2]<0 || modele2[2,2]<0 || modele2[2,3]<0)
          fit.variogram(vario,vgm(max(vario[,3]),"Lin",140),fit.method=1) else modele2
Info$KO2pepite[t]=if(dim(modele2)[1]==2) modele2$psill[1] else 0
Info$KO2ppalier[t]=if(dim(modele2)[1]==2) modele2$psill[2] else modele2$psill[1]
Info$KO2portee[t]=if(dim(modele2)[1]==2) modele2$range[2] else modele2$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele2[dim(modele2)[1],2]>0 && modele2[dim(modele2)[1],3]>0) {
  valid=krige.cv(Stations~1,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,])
  Err$KO2[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,],
                newdata=Grille[Grille$Periode==t,])
  Prev$KO2[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO2[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO2[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO2[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO2[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KO2EQM[t]=mean((Err$KO2[Err$Periode==t])**2,na.rm=T)
Info$KO2EAM[t]=median(abs(Err$KO2[Err$Periode==t]),na.rm=T)

# 3- Modèle sphérique :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele3=fit.variogram(vario,vgm(max(vario[,3]),"Sph",140,0.1*max(vario[,3])),fit.method=1)
modele3=if (modele3[1,2]<0 || modele3[2,2]<0 || modele3[2,3]<0)
          fit.variogram(vario,vgm(max(vario[,3]),"Sph",140),fit.method=1) else modele3
Info$KO3pepite[t]=if(dim(modele3)[1]==2) modele3$psill[1] else 0
Info$KO3ppalier[t]=if(dim(modele3)[1]==2) modele3$psill[2] else modele3$psill[1]
Info$KO3portee[t]=if(dim(modele3)[1]==2) modele3$range[2] else modele3$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele3[dim(modele3)[1],2]>0 && modele3[dim(modele3)[1],3]>0) {
valid=krige.cv(Stations~1,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,])
  Err$KO3[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,],
                newdata=Grille[Grille$Periode==t,])
  Prev$KO3[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO3[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO3[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO3[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO3[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
```

```
  }
Info$KO3EQM[t]=mean((Err$KO3[Err$Periode==t])**2,na.rm=T)
Info$KO3EAM[t]=median(abs(Err$KO3[Err$Periode==t]),na.rm=T)

# 4- Modèle exponentiel :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele4=fit.variogram(vario,vgm(max(vario[,3]),"Exp",140,0.1*max(vario[,3])),fit.method=1)
modele4=if (modele4[1,2]<0 || modele4[2,2]<0 || modele4[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Exp",140),fit.method=1) else modele4
Info$KO4pepite[t]=if(dim(modele4)[1]==2) modele4$psill[1] else 0
Info$KO4ppalier[t]=if(dim(modele4)[1]==2) modele4$psill[2] else modele4$psill[1]
Info$KO4porteep[t]=if(dim(modele4)[1]==2) 3*modele4$range[2] else 3*modele4$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele4[dim(modele4)[1],2]>0 && modele4[dim(modele4)[1],3]>0) {
  valid=krige.cv(Stations~1,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,])
  Err$KO4[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KO4[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO4[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO4[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO4[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO4[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KO4EQM[t]=mean((Err$KO4[Err$Periode==t])**2,na.rm=T)
Info$KO4EAM[t]=median(abs(Err$KO4[Err$Periode==t]),na.rm=T)

# 5- Modèle gaussien :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele5=fit.variogram(vario,vgm(max(vario[,3]),"Gau",140,0.1*max(vario[,3])),fit.method=1)
modele5=if (modele5[1,2]<0 || modele5[2,2]<0 || modele5[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Gau",140),fit.method=1) else modele5
Info$KO5pepite[t]=if(dim(modele5)[1]==2) modele5$psill[1] else 0
Info$KO5ppalier[t]=if(dim(modele5)[1]==2) modele5$psill[2] else modele5$psill[1]
Info$KO5porteep[t]=if(dim(modele5)[1]==2) sqrt(3)*modele5$range[2] else sqrt(3)*modele5$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele5[dim(modele5)[1],2]>0 && modele5[dim(modele5)[1],3]>0) {
  valid=krige.cv(Stations~1,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,])
  Err$KO5[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KO5[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO5[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO5[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO5[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO5[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KO5EQM[t]=mean((Err$KO5[Err$Periode==t])**2,na.rm=T)
Info$KO5EAM[t]=median(abs(Err$KO5[Err$Periode==t]),na.rm=T)

# 6- Modèle linéaire sans palier :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele6=fit.variogram(vario,vgm(max(vario[,3])/150,"Lin",,0.1*max(vario[,3])),fit.method=1)
```

```
modele6=if (modele6[1,2]<0 || modele6[2,2]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Lin"),fit.method=1) else modele6
Info$KO6pepite[t]=if(dim(modele6)[1]==2) modele6$psill[1] else 0
Info$KO6pente[t]=if(dim(modele6)[1]==2) modele6$psill[2] else modele6$psill[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele6[dim(modele6)[1],2]>0) {
  valid=krige.cv(Stations~1,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,])
  Err$KO6[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KO6[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO6[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO6[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO6[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO6[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KO6EQM[t]=mean((Err$KO6[Err$Periode==t])**2,na.rm=T)
Info$KO6EAM[t]=median(abs(Err$KO6[Err$Periode==t]),na.rm=T)

# 7- Modèle puissance :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele7=fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1,0.1*max(vario[,3])),fit.method=1)
modele7=if (modele7[1,2]<0 || modele7[2,2]<0 || modele7[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1),fit.method=1) else modele7
Info$KO7pepite[t]=if(dim(modele7)[1]==2) modele7$psill[1] else 0
Info$KO7echelle[t]=if(dim(modele7)[1]==2) modele7$psill[2] else modele7$psill[1]
Info$KO7puissance[t]=if(dim(modele7)[1]==2) modele7$range[2] else modele7$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if(modele7[dim(modele7)[1],2]>0&&modele7[dim(modele7)[1],3]>0&&modele7[dim(modele7)[1],3]<2){
  valid=krige.cv(Stations~1,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,])
  Err$KO7[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KO7[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KO7[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KO7[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KO7[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KO7[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KO7EQM[t]=mean((Err$KO7[Err$Periode==t])**2,na.rm=T)
Info$KO7EAM[t]=median(abs(Err$KO7[Err$Periode==t]),na.rm=T)

# Choix du modèle :
indice=data.frame(c(Info$KO1EQM[t],Info$KO2EQM[t],Info$KO3EQM[t],Info$KO4EQM[t],
                    Info$KO5EQM[t],Info$KO6EQM[t],Info$KO7EQM[t]))
minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
Info$KOmodele[t]=minimums[1,1]
if (Info$KOmodele[t]==1) {
  Err$KO[Err$Periode==t]=Err$KO1[Err$Periode==t]
  Prev$KO[Prev$Periode==t]=Prev$KO1[Prev$Periode==t]
  Var$KO[Var$Periode==t]=Var$KO1[Var$Periode==t]
  } else
if (Info$KOmodele[t]==2) {
```

```
    Err$KO[Err$Periode==t]=Err$KO2[Err$Periode==t]
    Prev$KO[Prev$Periode==t]=Prev$KO2[Prev$Periode==t]
    Var$KO[Var$Periode==t]=Var$KO2[Var$Periode==t]
    } else
  if (Info$KOmodele[t]==3) {
    Err$KO[Err$Periode==t]=Err$KO3[Err$Periode==t]
    Prev$KO[Prev$Periode==t]=Prev$KO3[Prev$Periode==t]
    Var$KO[Var$Periode==t]=Var$KO3[Var$Periode==t]
    } else
  if (Info$KOmodele[t]==4) {
    Err$KO[Err$Periode==t]=Err$KO4[Err$Periode==t]
    Prev$KO[Prev$Periode==t]=Prev$KO4[Prev$Periode==t]
    Var$KO[Var$Periode==t]=Var$KO4[Var$Periode==t]
    } else
  if (Info$KOmodele[t]==5) {
    Err$KO[Err$Periode==t]=Err$KO5[Err$Periode==t]
    Prev$KO[Prev$Periode==t]=Prev$KO5[Prev$Periode==t]
    Var$KO[Var$Periode==t]=Var$KO5[Var$Periode==t]
    } else
  if (Info$KOmodele[t]==6) {
    Err$KO[Err$Periode==t]=Err$KO6[Err$Periode==t]
    Prev$KO[Prev$Periode==t]=Prev$KO6[Prev$Periode==t]
    Var$KO[Var$Periode==t]=Var$KO6[Var$Periode==t]
    } else {
    Err$KO[Err$Periode==t]=Err$KO7[Err$Periode==t]
    Prev$KO[Prev$Periode==t]=Prev$KO7[Prev$Periode==t]
    Var$KO[Var$Periode==t]=Var$KO7[Var$Periode==t]
    }
  Info$KOEQM[t]=mean((Err$KO[Err$Periode==t])**2,na.rm=T)
  Info$KOEAM[t]=median(abs(Err$KO[Err$Periode==t]),na.rm=T)
  }


# KRIGEAGE UNIVERSEL :

for(t in Pluie){

  corr=cor.test(Donnees$Stations[Donnees$Periode==t],Donnees$x[Donnees$Periode==t],
                alternative="two.sided",method="pearson")
  Info$KUcorrX[t]=corr$estimate
  Info$KUpvalueX[t]=corr$p.value
  corr=cor.test(Donnees$Stations[Donnees$Periode==t],Donnees$y[Donnees$Periode==t],
                alternative="two.sided",method="pearson")
  Info$KUcorrY[t]=corr$estimate
  Info$KUpvalueY[t]=corr$p.value

# Première itération :

  # Estimation semi-variogramme sur les résidus ols :
  vario=variogram(Stations~1+x+y,loc=~x+y,Donnees[Donnees$Periode==t,],cutoff=150,width=10)

  # 1- Modèle pépitique :
  modele1=fit.variogram(vario,vgm(max(vario[,3]),"Nug"),fit.method=1)
  if (modele1[dim(modele1)[1],2]>0) {
    valid=krige.cv(Stations~1+x+y,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,])
```

```
  Err$KU1[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KU1[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KU1EQM[t]=mean((Err$KU1[Err$Periode==t])**2,na.rm=T)
Info$KU1EAM[t]=median(abs(Err$KU1[Err$Periode==t]),na.rm=T)

# 2- Modèle linéaire avec palier :
modele2=fit.variogram(vario,vgm(max(vario[,3]),"Lin",140,0.1*max(vario[,3])),fit.method=1)
modele2=if (modele2[1,2]<0 || modele2[2,2]<0 || modele2[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Lin",140),fit.method=1) else modele2
if (modele2[dim(modele2)[1],2]>0 && modele2[dim(modele2)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,])
  Err$KU2[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KU2[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KU2EQM[t]=mean((Err$KU2[Err$Periode==t])**2,na.rm=T)
Info$KU2EAM[t]=median(abs(Err$KU2[Err$Periode==t]),na.rm=T)

# 3- Modèle sphérique :
modele3=fit.variogram(vario,vgm(max(vario[,3]),"Sph",140,0.1*max(vario[,3])),fit.method=1)
modele3=if (modele3[1,2]<0 || modele3[2,2]<0 || modele3[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Sph",140),fit.method=1) else modele3
if (modele3[dim(modele3)[1],2]>0 && modele3[dim(modele3)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,])
  Err$KU3[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KU3[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KU3EQM[t]=mean((Err$KU3[Err$Periode==t])**2,na.rm=T)
Info$KU3EAM[t]=median(abs(Err$KU3[Err$Periode==t]),na.rm=T)

# 4- Modèle exponentiel :
modele4=fit.variogram(vario,vgm(max(vario[,3]),"Exp",140,0.1*max(vario[,3])),fit.method=1)
modele4=if (modele4[1,2]<0 || modele4[2,2]<0 || modele4[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Exp",140),fit.method=1) else modele4
if (modele4[dim(modele4)[1],2]>0 && modele4[dim(modele4)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,])
  Err$KU4[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KU4[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KU4EQM[t]=mean((Err$KU4[Err$Periode==t])**2,na.rm=T)
Info$KU4EAM[t]=median(abs(Err$KU4[Err$Periode==t]),na.rm=T)

# 5- Modèle gaussien :
modele5=fit.variogram(vario,vgm(max(vario[,3]),"Gau",140,0.1*max(vario[,3])),fit.method=1)
modele5=if (modele5[1,2]<0 || modele5[2,2]<0 || modele5[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Gau",140),fit.method=1) else modele5
if (modele5[dim(modele5)[1],2]>0 && modele5[dim(modele5)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,])
  Err$KU5[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KU5[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KU5EQM[t]=mean((Err$KU5[Err$Periode==t])**2,na.rm=T)
Info$KU5EAM[t]=median(abs(Err$KU5[Err$Periode==t]),na.rm=T)

# 6- Modèle linéaire sans palier :
modele6=fit.variogram(vario,vgm(max(vario[,3])/150,"Lin",,0.1*max(vario[,3])),fit.method=1)
modele6=if (modele6[1,2]<0 || modele6[2,2]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Lin"),fit.method=1) else modele6
if (modele6[dim(modele6)[1],2]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,])
```

```
    Err$KU6[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
    } else Err$KU6[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Info$KU6EQM[t]=mean((Err$KU6[Err$Periode==t])**2,na.rm=T)
  Info$KU6EAM[t]=median(abs(Err$KU6[Err$Periode==t]),na.rm=T)

  # 7- Modèle puissance :
  modele7=fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1,0.1*max(vario[,3])),fit.method=1)
  modele7=if (modele7[1,2]<0 || modele7[2,2]<0 || modele7[2,3]<0)
            fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1),fit.method=1) else modele7
  if(modele7[dim(modele7)[1],2]>0&&modele7[dim(modele7)[1],3]>0&&modele7[dim(modele7)[1],3]<2){
    valid=krige.cv(Stations~1+x+y,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,])
    Err$KU7[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
    } else Err$KU7[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Info$KU7EQM[t]=mean((Err$KU7[Err$Periode==t])**2,na.rm=T)
  Info$KU7EAM[t]=median(abs(Err$KU7[Err$Periode==t]),na.rm=T)

  # Choix du modèle :
  indice=data.frame(c(Info$KU1EQM[t],Info$KU2EQM[t],Info$KU3EQM[t],Info$KU4EQM[t],
                      Info$KU5EQM[t],Info$KU6EQM[t],Info$KU7EQM[t]))
  minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
  Info$KUmodele[t]=minimums[1,1]
  modele = if (Info$KUmodele[t]==1) modele1 else if (Info$KUmodele[t]==2) modele2 else
           if (Info$KUmodele[t]==3) modele3 else if (Info$KUmodele[t]==4) modele4 else
           if (Info$KUmodele[t]==5) modele5 else if (Info$KUmodele[t]==6) modele6 else modele7

  # Calcul des résidus gls :
  g.gls=gstat(id="Stations",form=Stations~1+x+y,loc=~x+y,model=modele,
              data=Donnees[Donnees$Periode==t,])
  prev.gls=predict(g.gls,Donnees[Donnees$Periode==t,],BLUE=T)
  res.gls=Donnees$Stations[Donnees$Periode==t]-prev.gls$Stations.pred

  # Estimation semi-variogramme sur les résidus gls :
  vario=variogram(res.gls~1,loc=~x+y,Donnees[Donnees$Periode==t,],cutoff=150,width=10)

# Deuxième itération :

  # 1- Modèle pépitique :
  # Pour ajuster le modèle et noter les info par rapport à ce modèle :
  modele1=fit.variogram(vario,vgm(max(vario[,3]),"Nug"),fit.method=1)
  Info$KU1pepite[t]= modele1$psill[1]
  # Pour faire validation croisée, interpolation, sauver résultats et noter informations :
  if (modele1[dim(modele1)[1],2]>0) {
    valid=krige.cv(Stations~1+x+y,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,])
    Err$KU1[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
    Interpo=krige(Stations~1+x+y,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,],
                  newdata=Grille[Grille$Periode==t,])
    Prev$KU1[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
    Var$KU1[Var$Periode==t]=Interpo[,4]
    } else {
    Err$KU1[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
    Prev$KU1[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
    Var$KU1[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
    }
  Info$KU1EQM[t]=mean((Err$KU1[Err$Periode==t])**2,na.rm=T)
  Info$KU1EAM[t]=median(abs(Err$KU1[Err$Periode==t]),na.rm=T)
```

```
# 2- Modèle linéaire avec palier :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele2=fit.variogram(vario,vgm(max(vario[,3]),"Lin",140,0.1*max(vario[,3])),fit.method=1)
modele2=if (modele2[1,2]<0 || modele2[2,2]<0 || modele2[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Lin",140),fit.method=1) else modele2
Info$KU2pepite[t]=if(dim(modele2)[1]==2) modele2$psill[1] else 0
Info$KU2ppalier[t]=if(dim(modele2)[1]==2) modele2$psill[2] else modele2$psill[1]
Info$KU2portee[t]=if(dim(modele2)[1]==2) modele2$range[2] else modele2$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele2[dim(modele2)[1],2]>0 && modele2[dim(modele2)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,])
  Err$KU2[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+x+y,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KU2[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KU2[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KU2[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KU2[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KU2[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KU2EQM[t]=mean((Err$KU2[Err$Periode==t])**2,na.rm=T)
Info$KU2EAM[t]=median(abs(Err$KU2[Err$Periode==t]),na.rm=T)

# 3- Modèle sphérique :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele3=fit.variogram(vario,vgm(max(vario[,3]),"Sph",140,0.1*max(vario[,3])),fit.method=1)
modele3=if (modele3[1,2]<0 || modele3[2,2]<0 || modele3[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Sph",140),fit.method=1) else modele3
Info$KU3pepite[t]=if(dim(modele3)[1]==2) modele3$psill[1] else 0
Info$KU3ppalier[t]=if(dim(modele3)[1]==2) modele3$psill[2] else modele3$psill[1]
Info$KU3portee[t]=if(dim(modele3)[1]==2) modele3$range[2] else modele3$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele3[dim(modele3)[1],2]>0 && modele3[dim(modele3)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,])
  Err$KU3[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+x+y,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KU3[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KU3[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KU3[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KU3[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KU3[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KU3EQM[t]=mean((Err$KU3[Err$Periode==t])**2,na.rm=T)
Info$KU3EAM[t]=median(abs(Err$KU3[Err$Periode==t]),na.rm=T)

# 4- Modèle exponentiel :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele4=fit.variogram(vario,vgm(max(vario[,3]),"Exp",140,0.1*max(vario[,3])),fit.method=1)
modele4=if (modele4[1,2]<0 || modele4[2,2]<0 || modele4[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Exp",140),fit.method=1) else modele4
Info$KU4pepite[t]=if(dim(modele4)[1]==2) modele4$psill[1] else 0
```

```
Info$KU4ppalier[t]=if(dim(modele4)[1]==2) modele4$psill[2] else modele4$psill[1]
Info$KU4porteep[t]=if(dim(modele4)[1]==2) 3*modele4$range[2] else 3*modele4$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele4[dim(modele4)[1],2]>0 && modele4[dim(modele4)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,])
  Err$KU4[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+x+y,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,],
                newdata=Grille[Grille$Periode==t,])
  Prev$KU4[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KU4[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KU4[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KU4[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KU4[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KU4EQM[t]=mean((Err$KU4[Err$Periode==t])**2,na.rm=T)
Info$KU4EAM[t]=median(abs(Err$KU4[Err$Periode==t]),na.rm=T)

# 5- Modèle gaussien :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele5=fit.variogram(vario,vgm(max(vario[,3]),"Gau",140,0.1*max(vario[,3])),fit.method=1)
modele5=if (modele5[1,2]<0 || modele5[2,2]<0 || modele5[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Gau",140),fit.method=1) else modele5
Info$KU5pepite[t]=if(dim(modele5)[1]==2) modele5$psill[1] else 0
Info$KU5ppalier[t]=if(dim(modele5)[1]==2) modele5$psill[2] else modele5$psill[1]
Info$KU5porteep[t]=if(dim(modele5)[1]==2) sqrt(3)*modele5$range[2] else sqrt(3)*modele5$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele5[dim(modele5)[1],2]>0 && modele5[dim(modele5)[1],3]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,])
  Err$KU5[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+x+y,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,],
                newdata=Grille[Grille$Periode==t,])
  Prev$KU5[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KU5[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KU5[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KU5[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KU5[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KU5EQM[t]=mean((Err$KU5[Err$Periode==t])**2,na.rm=T)
Info$KU5EAM[t]=median(abs(Err$KU5[Err$Periode==t]),na.rm=T)

# 6- Modèle linéaire sans palier :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele6=fit.variogram(vario,vgm(max(vario[,3])/150,"Lin",,0.1*max(vario[,3])),fit.method=1)
modele6=if (modele6[1,2]<0 || modele6[2,2]<0)
            fit.variogram(vario,vgm(max(vario[,3])/150,"Lin"),fit.method=1) else modele6
Info$KU6pepite[t]=if(dim(modele6)[1]==2) modele6$psill[1] else 0
Info$KU6pente[t]=if(dim(modele6)[1]==2) modele6$psill[2] else modele6$psill[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele6[dim(modele6)[1],2]>0) {
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,])
  Err$KU6[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+x+y,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,],
                newdata=Grille[Grille$Periode==t,])
```

```
  Prev$KU6[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KU6[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KU6[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KU6[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KU6[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KU6EQM[t]=mean((Err$KU6[Err$Periode==t])**2,na.rm=T)
Info$KU6EAM[t]=median(abs(Err$KU6[Err$Periode==t]),na.rm=T)

# 7- Modèle puissance :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele7=fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1,0.1*max(vario[,3])),fit.method=1)
modele7=if (modele7[1,2]<0 || modele7[2,2]<0 || modele7[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1),fit.method=1) else modele7
Info$KU7pepite[t]=if(dim(modele7)[1]==2) modele7$psill[1] else 0
Info$KU7echelle[t]=if(dim(modele7)[1]==2) modele7$psill[2] else modele7$psill[1]
Info$KU7puissance[t]=if(dim(modele7)[1]==2) modele7$range[2] else modele7$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if(modele7[dim(modele7)[1],2]>0&&modele7[dim(modele7)[1],3]>0&&modele7[dim(modele7)[1],3]<2){
  valid=krige.cv(Stations~1+x+y,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,])
  Err$KU7[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+x+y,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KU7[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KU7[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KU7[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KU7[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KU7[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KU7EQM[t]=mean((Err$KU7[Err$Periode==t])**2,na.rm=T)
Info$KU7EAM[t]=median(abs(Err$KU7[Err$Periode==t]),na.rm=T)

# Choix du modèle :
indice=data.frame(c(Info$KU1EQM[t],Info$KU2EQM[t],Info$KU3EQM[t],Info$KU4EQM[t],
                    Info$KU5EQM[t],Info$KU6EQM[t],Info$KU7EQM[t]))
minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
Info$KUmodele[t]=minimums[1,1]
if (Info$KUmodele[t]==1) {
  Err$KU[Err$Periode==t]=Err$KU1[Err$Periode==t]
  Prev$KU[Prev$Periode==t]=Prev$KU1[Prev$Periode==t]
  Var$KU[Var$Periode==t]=Var$KU1[Var$Periode==t]
  } else
if (Info$KUmodele[t]==2) {
  Err$KU[Err$Periode==t]=Err$KU2[Err$Periode==t]
  Prev$KU[Prev$Periode==t]=Prev$KU2[Prev$Periode==t]
  Var$KU[Var$Periode==t]=Var$KU2[Var$Periode==t]
  } else
if (Info$KUmodele[t]==3) {
  Err$KU[Err$Periode==t]=Err$KU3[Err$Periode==t]
  Prev$KU[Prev$Periode==t]=Prev$KU3[Prev$Periode==t]
  Var$KU[Var$Periode==t]=Var$KU3[Var$Periode==t]
  } else
if (Info$KUmodele[t]==4) {
```

```
    Err$KU[Err$Periode==t]=Err$KU4[Err$Periode==t]
    Prev$KU[Prev$Periode==t]=Prev$KU4[Prev$Periode==t]
    Var$KU[Var$Periode==t]=Var$KU4[Var$Periode==t]
    } else
  if (Info$KUmodele[t]==5) {
    Err$KU[Err$Periode==t]=Err$KU5[Err$Periode==t]
    Prev$KU[Prev$Periode==t]=Prev$KU5[Prev$Periode==t]
    Var$KU[Var$Periode==t]=Var$KU5[Var$Periode==t]
    } else
  if (Info$KUmodele[t]==6) {
    Err$KU[Err$Periode==t]=Err$KU6[Err$Periode==t]
    Prev$KU[Prev$Periode==t]=Prev$KU6[Prev$Periode==t]
    Var$KU[Var$Periode==t]=Var$KU6[Var$Periode==t]
    } else {
    Err$KU[Err$Periode==t]=Err$KU7[Err$Periode==t]
    Prev$KU[Prev$Periode==t]=Prev$KU7[Prev$Periode==t]
    Var$KU[Var$Periode==t]=Var$KU7[Var$Periode==t]
    }
  Info$KUEQM[t]=mean((Err$KU[Err$Periode==t])**2,na.rm=T)
  Info$KUEAM[t]=median(abs(Err$KU[Err$Periode==t]),na.rm=T)
  }


# FILTRE : Mise de côté des périodes pour lesquelles les variances leave-one-out
# des prévisions GEM ne sont pas assez grandes pour KDE
for(t in Pluie){
  IDlignes=na.omit(ifelse(Donnees$Periode==t,row(Donnees),NA))
  PetiteVar=rep(NA,length(IDlignes))
  for (i in 1:length(IDlignes)){
    PetiteVar[i]=ifelse(var(Donnees$GEM[IDlignes[-i]])<0.0001,1,0)
    }
  Info$KDEpossible[t]= if (sum(PetiteVar)>0) "non" else "oui"
  }
KDEpossible=na.omit(ifelse(Info$KDEpossible=="oui",row(Info),NA))


# KRIGEAGE AVEC DÉRIVE EXTERNE :

for(t in KDEpossible){

  corr=cor.test(Donnees$Stations[Donnees$Periode==t],Donnees$GEM[Donnees$Periode==t],
                alternative="two.sided", method="pearson")
  Info$KDEcorrGEM[t]=corr$estimate
  Info$KDEpvalueGEM[t]=corr$p.value

# Première itération :

  # Estimation semi-variogramme sur les résidus ols :
  vario=variogram(Stations~1+GEM,loc=~x+y,Donnees[Donnees$Periode==t,],cutoff=150,width=10)

  # 1- Modèle pépitique :
  modele1=fit.variogram(vario,vgm(max(vario[,3]),"Nug"),fit.method=1)
  if (modele1[dim(modele1)[1],2]>0) {
    valid=krige.cv(Stations~1+GEM,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,])
    Err$KDE1[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
```

```
  } else Err$KDE1[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KDE1EQM[t]=mean((Err$KDE1[Err$Periode==t])**2,na.rm=T)
Info$KDE1EAM[t]=median(abs(Err$KDE1[Err$Periode==t]),na.rm=T)

# 2- Modèle linéaire avec palier :
modele2=fit.variogram(vario,vgm(max(vario[,3]),"Lin",140,0.1*max(vario[,3])),fit.method=1)
modele2=if (modele2[1,2]<0 || modele2[2,2]<0 || modele2[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Lin",140),fit.method=1) else modele2
if (modele2[dim(modele2)[1],2]>0 && modele2[dim(modele2)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,])
  Err$KDE2[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KDE2[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KDE2EQM[t]=mean((Err$KDE2[Err$Periode==t])**2,na.rm=T)
Info$KDE2EAM[t]=median(abs(Err$KDE2[Err$Periode==t]),na.rm=T)

# 3- Modèle sphérique :
modele3=fit.variogram(vario,vgm(max(vario[,3]),"Sph",140,0.1*max(vario[,3])),fit.method=1)
modele3=if (modele3[1,2]<0 || modele3[2,2]<0 || modele3[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Sph",140),fit.method=1) else modele3
if (modele3[dim(modele3)[1],2]>0 && modele3[dim(modele3)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,])
  Err$KDE3[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KDE3[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KDE3EQM[t]=mean((Err$KDE3[Err$Periode==t])**2,na.rm=T)
Info$KDE3EAM[t]=median(abs(Err$KDE3[Err$Periode==t]),na.rm=T)

# 4- Modèle exponentiel :
modele4=fit.variogram(vario,vgm(max(vario[,3]),"Exp",140,0.1*max(vario[,3])),fit.method=1)
modele4=if (modele4[1,2]<0 || modele4[2,2]<0 || modele4[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Exp",140),fit.method=1) else modele4
if (modele4[dim(modele4)[1],2]>0 && modele4[dim(modele4)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,])
  Err$KDE4[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KDE4[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KDE4EQM[t]=mean((Err$KDE4[Err$Periode==t])**2,na.rm=T)
Info$KDE4EAM[t]=median(abs(Err$KDE4[Err$Periode==t]),na.rm=T)

# 5- Modèle gaussien :
modele5=fit.variogram(vario,vgm(max(vario[,3]),"Gau",140,0.1*max(vario[,3])),fit.method=1)
modele5=if (modele5[1,2]<0 || modele5[2,2]<0 || modele5[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Gau",140),fit.method=1) else modele5
if (modele5[dim(modele5)[1],2]>0 && modele5[dim(modele5)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,])
  Err$KDE5[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  } else Err$KDE5[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
Info$KDE5EQM[t]=mean((Err$KDE5[Err$Periode==t])**2,na.rm=T)
Info$KDE5EAM[t]=median(abs(Err$KDE5[Err$Periode==t]),na.rm=T)

# 6- Modèle linéaire sans palier :
modele6=fit.variogram(vario,vgm(max(vario[,3])/150,"Lin",,0.1*max(vario[,3])),fit.method=1)
modele6=if (modele6[1,2]<0 || modele6[2,2]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Lin"),fit.method=1) else modele6
if (modele6[dim(modele6)[1],2]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,])
  Err$KDE6[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
```

```
    } else Err$KDE6[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Info$KDE6EQM[t]=mean((Err$KDE6[Err$Periode==t])**2,na.rm=T)
  Info$KDE6EAM[t]=median(abs(Err$KDE6[Err$Periode==t]),na.rm=T)

  # 7- Modèle puissance :
  modele7=fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1,0.1*max(vario[,3])),fit.method=1)
  modele7=if (modele7[1,2]<0 || modele7[2,2]<0 || modele7[2,3]<0)
            fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1),fit.method=1) else modele7
  if(modele7[dim(modele7)[1],2]>0&&modele7[dim(modele7)[1],3]>0&&modele7[dim(modele7)[1],3]<2){
    valid=krige.cv(Stations~1+GEM,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,])
    Err$KDE7[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
    } else Err$KDE7[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Info$KDE7EQM[t]=mean((Err$KDE7[Err$Periode==t])**2,na.rm=T)
  Info$KDE7EAM[t]=median(abs(Err$KDE7[Err$Periode==t]),na.rm=T)

  # Choix du modèle :
  indice=data.frame(c(Info$KDE1EQM[t],Info$KDE2EQM[t],Info$KDE3EQM[t],Info$KDE4EQM[t],
                      Info$KDE5EQM[t],Info$KDE6EQM[t],Info$KDE7EQM[t]))
  minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
  Info$KDEmodele[t]=minimums[1,1]
  modele = if (Info$KDEmodele[t]==1) modele1 else if (Info$KDEmodele[t]==2) modele2 else
           if (Info$KDEmodele[t]==3) modele3 else if (Info$KDEmodele[t]==4) modele4 else
           if (Info$KDEmodele[t]==5) modele5 else if (Info$KDEmodele[t]==6) modele6 else modele7

  # Calcul des résidus gls :
  g.gls=gstat(id="Stations",form=Stations~1+GEM,loc=~x+y,model=modele,
              data=Donnees[Donnees$Periode==t,])
  prev.gls=predict(g.gls,Donnees[Donnees$Periode==t,],BLUE=T)
  res.gls=Donnees$Stations[Donnees$Periode==t]-prev.gls$Stations.pred

  # Estimation semi-variogramme sur les résidus gls :
  vario=variogram(res.gls~1,loc=~x+y,Donnees[Donnees$Periode==t,],cutoff=150,width=10)

# Deuxième itération :

  # 1- Modèle pépitique :
  # Pour ajuster le modèle et noter les info par rapport à ce modèle :
  modele1=fit.variogram(vario,vgm(max(vario[,3]),"Nug"),fit.method=1)
  Info$KDE1pepite[t]= modele1$psill[1]
  # Pour faire validation croisée, interpolation, sauver résultats et noter informations :
  if (modele1[dim(modele1)[1],2]>0) {
    valid=krige.cv(Stations~1+GEM,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,])
    Err$KDE1[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
    Interpo=krige(Stations~1+GEM,~x+y,model=modele1,data=Donnees[Donnees$Periode==t,],
                  newdata=Grille[Grille$Periode==t,])
    Prev$KDE1[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
    Var$KDE1[Var$Periode==t]=Interpo[,4]
    } else {
    Err$KDE1[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
    Prev$KDE1[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
    Var$KDE1[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
    }
    Info$KDE1EQM[t]=mean((Err$KDE1[Err$Periode==t])**2,na.rm=T)
    Info$KDE1EAM[t]=median(abs(Err$KDE1[Err$Periode==t]),na.rm=T)
```

```
# 2- Modèle linéaire avec palier :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele2=fit.variogram(vario,vgm(max(vario[,3]),"Lin",140,0.1*max(vario[,3])),fit.method=1)
modele2=if (modele2[1,2]<0 || modele2[2,2]<0 || modele2[2,3]<0)
          fit.variogram(vario,vgm(max(vario[,3]),"Lin",140),fit.method=1) else modele2
Info$KDE2pepite[t]=if(dim(modele2)[1]==2) modele2$psill[1] else 0
Info$KDE2ppalier[t]=if(dim(modele2)[1]==2) modele2$psill[2] else modele2$psill[1]
Info$KDE2portee[t]=if(dim(modele2)[1]==2) modele2$range[2] else modele2$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele2[dim(modele2)[1],2]>0 && modele2[dim(modele2)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,])
  Err$KDE2[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+GEM,~x+y,model=modele2,data=Donnees[Donnees$Periode==t,],
              newdata=Grille[Grille$Periode==t,])
  Prev$KDE2[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KDE2[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KDE2[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KDE2[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KDE2[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KDE2EQM[t]=mean((Err$KDE2[Err$Periode==t])**2,na.rm=T)
Info$KDE2EAM[t]=median(abs(Err$KDE2[Err$Periode==t]),na.rm=T)

# 3- Modèle sphérique :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele3=fit.variogram(vario,vgm(max(vario[,3]),"Sph",140,0.1*max(vario[,3])),fit.method=1)
modele3=if (modele3[1,2]<0 || modele3[2,2]<0 || modele3[2,3]<0)
          fit.variogram(vario,vgm(max(vario[,3]),"Sph",140),fit.method=1) else modele3
Info$KDE3pepite[t]=if(dim(modele3)[1]==2) modele3$psill[1] else 0
Info$KDE3ppalier[t]=if(dim(modele3)[1]==2) modele3$psill[2] else modele3$psill[1]
Info$KDE3portee[t]=if(dim(modele3)[1]==2) modele3$range[2] else modele3$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele3[dim(modele3)[1],2]>0 && modele3[dim(modele3)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,])
  Err$KDE3[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+GEM,~x+y,model=modele3,data=Donnees[Donnees$Periode==t,],
              newdata=Grille[Grille$Periode==t,])
  Prev$KDE3[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KDE3[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KDE3[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KDE3[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KDE3[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KDE3EQM[t]=mean((Err$KDE3[Err$Periode==t])**2,na.rm=T)
Info$KDE3EAM[t]=median(abs(Err$KDE3[Err$Periode==t]),na.rm=T)

# 4- Modèle exponentiel :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele4=fit.variogram(vario,vgm(max(vario[,3]),"Exp",140,0.1*max(vario[,3])),fit.method=1)
modele4=if (modele4[1,2]<0 || modele4[2,2]<0 || modele4[2,3]<0)
          fit.variogram(vario,vgm(max(vario[,3]),"Exp",140),fit.method=1) else modele4
Info$KDE4pepite[t]=if(dim(modele4)[1]==2) modele4$psill[1] else 0
Info$KDE4ppalier[t]=if(dim(modele4)[1]==2) modele4$psill[2] else modele4$psill[1]
```

```
Info$KDE4porteep[t]=if(dim(modele4)[1]==2) 3*modele4$range[2] else 3*modele4$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele4[dim(modele4)[1],2]>0 && modele4[dim(modele4)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,])
  Err$KDE4[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+GEM,~x+y,model=modele4,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KDE4[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KDE4[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KDE4[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KDE4[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KDE4[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KDE4EQM[t]=mean((Err$KDE4[Err$Periode==t])**2,na.rm=T)
Info$KDE4EAM[t]=median(abs(Err$KDE4[Err$Periode==t]),na.rm=T)

# 5- Modèle gaussien :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele5=fit.variogram(vario,vgm(max(vario[,3]),"Gau",140,0.1*max(vario[,3])),fit.method=1)
modele5=if (modele5[1,2]<0 || modele5[2,2]<0 || modele5[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3]),"Gau",140),fit.method=1) else modele5
Info$KDE5pepite[t]=if(dim(modele5)[1]==2) modele5$psill[1] else 0
Info$KDE5ppalier[t]=if(dim(modele5)[1]==2) modele5$psill[2] else modele5$psill[1]
Info$KDE5porteep[t]=if(dim(modele5)[1]==2) sqrt(3)*modele5$range[2] else sqrt(3)*modele5$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele5[dim(modele5)[1],2]>0 && modele5[dim(modele5)[1],3]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,])
  Err$KDE5[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+GEM,~x+y,model=modele5,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KDE5[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KDE5[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KDE5[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KDE5[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KDE5[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KDE5EQM[t]=mean((Err$KDE5[Err$Periode==t])**2,na.rm=T)
Info$KDE5EAM[t]=median(abs(Err$KDE5[Err$Periode==t]),na.rm=T)

# 6- Modèle linéaire sans palier :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele6=fit.variogram(vario,vgm(max(vario[,3])/150,"Lin",,0.1*max(vario[,3])),fit.method=1)
modele6=if (modele6[1,2]<0 || modele6[2,2]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Lin"),fit.method=1) else modele6
Info$KDE6pepite[t]=if(dim(modele6)[1]==2) modele6$psill[1] else 0
Info$KDE6pente[t]=if(dim(modele6)[1]==2) modele6$psill[2] else modele6$psill[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if (modele6[dim(modele6)[1],2]>0) {
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,])
  Err$KDE6[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+GEM,~x+y,model=modele6,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KDE6[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
```

```
  Var$KDE6[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KDE6[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KDE6[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KDE6[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KDE6EQM[t]=mean((Err$KDE6[Err$Periode==t])**2,na.rm=T)
Info$KDE6EAM[t]=median(abs(Err$KDE6[Err$Periode==t]),na.rm=T)

# 7- Modèle puissance :
# Pour ajuster le modèle et noter les info par rapport à ce modèle :
modele7=fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1,0.1*max(vario[,3])),fit.method=1)
modele7=if (modele7[1,2]<0 || modele7[2,2]<0 || modele7[2,3]<0)
           fit.variogram(vario,vgm(max(vario[,3])/150,"Pow",1),fit.method=1) else modele7
Info$KDE7pepite[t]=if(dim(modele7)[1]==2) modele7$psill[1] else 0
Info$KDE7echelle[t]=if(dim(modele7)[1]==2) modele7$psill[2] else modele7$psill[1]
Info$KDE7puissance[t]=if(dim(modele7)[1]==2) modele7$range[2] else modele7$range[1]
# Pour faire validation croisée, interpolation, sauver résultats et noter informations :
if(modele7[dim(modele7)[1],2]>0&&modele7[dim(modele7)[1],3]>0&&modele7[dim(modele7)[1],3]<2){
  valid=krige.cv(Stations~1+GEM,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,])
  Err$KDE7[Err$Periode==t]=valid[,5]-ifelse(valid[,3]<0,0,valid[,3])
  Interpo=krige(Stations~1+GEM,~x+y,model=modele7,data=Donnees[Donnees$Periode==t,],
               newdata=Grille[Grille$Periode==t,])
  Prev$KDE7[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$KDE7[Var$Periode==t]=Interpo[,4]
  } else {
  Err$KDE7[Err$Periode==t]=rep(NA,dim(Donnees[Donnees$Periode==t,])[1])
  Prev$KDE7[Prev$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  Var$KDE7[Var$Periode==t]=rep(NA,dim(Grille[Grille$Periode==t,])[1])
  }
Info$KDE7EQM[t]=mean((Err$KDE7[Err$Periode==t])**2,na.rm=T)
Info$KDE7EAM[t]=median(abs(Err$KDE7[Err$Periode==t]),na.rm=T)

# Choix du modèle :
indice=data.frame(c(Info$KDE1EQM[t],Info$KDE2EQM[t],Info$KDE3EQM[t],Info$KDE4EQM[t],
                    Info$KDE5EQM[t],Info$KDE6EQM[t],Info$KDE7EQM[t]))
minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
Info$KDEmodele[t]=minimums[1,1]
if (Info$KDEmodele[t]==1) {
  Err$KDE[Err$Periode==t]=Err$KDE1[Err$Periode==t]
  Prev$KDE[Prev$Periode==t]=Prev$KDE1[Prev$Periode==t]
  Var$KDE[Var$Periode==t]=Var$KDE1[Var$Periode==t]
  } else
if (Info$KDEmodele[t]==2) {
  Err$KDE[Err$Periode==t]=Err$KDE2[Err$Periode==t]
  Prev$KDE[Prev$Periode==t]=Prev$KDE2[Prev$Periode==t]
  Var$KDE[Var$Periode==t]=Var$KDE2[Var$Periode==t]
  } else
if (Info$KDEmodele[t]==3) {
  Err$KDE[Err$Periode==t]=Err$KDE3[Err$Periode==t]
  Prev$KDE[Prev$Periode==t]=Prev$KDE3[Prev$Periode==t]
  Var$KDE[Var$Periode==t]=Var$KDE3[Var$Periode==t]
  } else
if (Info$KDEmodele[t]==4) {
  Err$KDE[Err$Periode==t]=Err$KDE4[Err$Periode==t]
```

```
    Prev$KDE[Prev$Periode==t]=Prev$KDE4[Prev$Periode==t]
    Var$KDE[Var$Periode==t]=Var$KDE4[Var$Periode==t]
    } else
  if (Info$KDEmodele[t]==5) {
    Err$KDE[Err$Periode==t]=Err$KDE5[Err$Periode==t]
    Prev$KDE[Prev$Periode==t]=Prev$KDE5[Prev$Periode==t]
    Var$KDE[Var$Periode==t]=Var$KDE5[Var$Periode==t]
    } else
  if (Info$KDEmodele[t]==6) {
    Err$KDE[Err$Periode==t]=Err$KDE6[Err$Periode==t]
    Prev$KDE[Prev$Periode==t]=Prev$KDE6[Prev$Periode==t]
    Var$KDE[Var$Periode==t]=Var$KDE6[Var$Periode==t]
    } else {
    Err$KDE[Err$Periode==t]=Err$KDE7[Err$Periode==t]
    Prev$KDE[Prev$Periode==t]=Prev$KDE7[Prev$Periode==t]
    Var$KDE[Var$Periode==t]=Var$KDE7[Var$Periode==t]
    }
  Info$KDEEQM[t]=mean((Err$KDE[Err$Periode==t])**2,na.rm=T)
  Info$KDEEAM[t]=median(abs(Err$KDE[Err$Periode==t]),na.rm=T)
  }


# COKRIGEAGE ORDINAIRE

for (t in Pluie) {

  # Estimation du variogramme expérimental :
  g.interpo=gstat(id="Stations",formula=Stations~1,loc=~x+y,data=Donnees[Donnees$Periode==t,])
  g.interpo=gstat(g.interpo,id="GEM",formula=GEM~1,loc=~x+y,data=Grille[Grille$Periode==t,])
  vario=variogram(g.interpo,cutoff=150,width=15)

  # Ajustement modèle linéaire de corégionalisation (linéaire sans effet de pépite)
  g.interpo=gstat(g.interpo,model=vgm(max(vario[,3])/150,"Lin"),fill.all=T)
  g.interpo=fit.lmc(vario,g.interpo)
  Info$CKOpenteStations[t]=g.interpo$model$Stations[1,2]
  Info$CKOpenteGEM[t]=g.interpo$model$GEM[1,2]
  Info$CKOpenteStaGEM[t]=g.interpo$model$Stations.GEM[1,2]

  # Validation croisée
  Periodet=na.omit(ifelse(Donnees$Periode==t,row(Donnees),NA))
  for (i in 1:length(Periodet)){
    g.valid=gstat(id="Stations",formula=Stations~1,loc=~x+y,data=Donnees[Periodet[-i],],
                  model=g.interpo$model$Stations)
    g.valid=gstat(g.valid,id="GEM",formula=GEM~1,locations=~x+y,data=Grille[Grille$Periode==t,],
                  model=g.interpo$model$GEM)
    g.valid=gstat(g.valid,id=c("Stations","GEM"),model=g.interpo$model$Stations.GEM)
    valid=predict(g.valid,Donnees[Periodet[i],])
    Err$CKO[Periodet[i]]=Donnees$Stations[Periodet[i]]
                         -ifelse(valid$Stations.pred<0,0,valid$Stations.pred)
    }
  Info$CKOEQM[t]=mean((Err$CKO[Err$Periode==t])**2,na.rm=T)
  Info$CKOEAM[t]=median(abs(Err$CKO[Err$Periode==t]),na.rm=T)

  # Prévision grille GEM :
  Interpo = predict(g.interpo,Grille[Grille$Periode==t,])
```

```
  Prev$CKO[Prev$Periode==t]=ifelse(Interpo[,3]<0,0,Interpo[,3])
  Var$CKO[Var$Periode==t]=Interpo[,4]
  }


# MÉTHODE SÉLECTION : choix d'une méthode par validation croisée

for( t in Pluie) {
  indice=data.frame(c(Info$KOEQM[t],Info$KUEQM[t],Info$KDEEQM[t],
                      Info$CKOEQM[t],Info$RLxyEQM[t],Info$RLxywEQM[t]))
  minimums=na.omit(ifelse(indice==min(indice,na.rm=T),row(indice),NA))
  if (minimums[1,1]==1) {
    Info$Selection[t]="KO"
    Err$Selection[Err$Periode==t]=Err$KO[Err$Periode==t]
    Prev$Selection[Prev$Periode==t]=Prev$KO[Prev$Periode==t]
    Var$Selection[Var$Periode==t]=Var$KO[Var$Periode==t]
    } else
  if (minimums[1,1]==2) {
    Info$Selection[t]="KU"
    Err$Selection[Err$Periode==t]=Err$KU[Err$Periode==t]
    Prev$Selection[Prev$Periode==t]=Prev$KU[Prev$Periode==t]
    Var$Selection[Var$Periode==t]=Var$KU[Var$Periode==t]
    } else
  if (minimums[1,1]==3) {
    Info$Selection[t]="KDE"
    Err$Selection[Err$Periode==t]=Err$KDE[Err$Periode==t]
    Prev$Selection[Prev$Periode==t]=Prev$KDE[Prev$Periode==t]
    Var$Selection[Var$Periode==t]=Var$KDE[Var$Periode==t]
    } else
  if (minimums[1,1]==4) {
    Info$Selection[t]="CKO"
    Err$Selection[Err$Periode==t]=Err$CKO[Err$Periode==t]
    Prev$Selection[Prev$Periode==t]=Prev$CKO[Prev$Periode==t]
    Var$Selection[Var$Periode==t]=Var$CKO[Var$Periode==t]
    } else
  if (minimums[1,1]==5) {
    Info$Selection[t]="RLxy"
    Err$Selection[Err$Periode==t]=Err$RLxy[Err$Periode==t]
    Prev$Selection[Prev$Periode==t]=Prev$RLxy[Prev$Periode==t]
    Var$Selection[Var$Periode==t]=Var$RLxy[Var$Periode==t]
    } else {
    Info$Selection[t]="RLxyw"
    Err$Selection[Err$Periode==t]=Err$RLxyw[Err$Periode==t]
    Prev$Selection[Prev$Periode==t]=Prev$RLxyw[Prev$Periode==t]
    Var$Selection[Var$Periode==t]=Var$RLxyw[Var$Periode==t]
    }
  Info$SelectionEQM[t]=mean((Err$Selection[Err$Periode==t])**2,na.rm=T)
  Info$SelectionEAM[t]=median(abs(Err$Selection[Err$Periode==t]),na.rm=T)
  }
```