Modélisation spatio-temporelle de la pollution atmosphérique urbaine à partir d'un réseau de surveillance de la qualité de l'air

Adriana Coman

École Doctorale Sciences et Ingénierie : Matériaux, Modélisation et Environnement - SIMME
Centre d'Etudes et Recherches en Thermique, Environnement et Systèmes - CERTES

THÈSE

présentée pour l'obtention du

Doctorat de l'Université Paris Est
(Spécialité Sciences de l'Environnement)

par

Adriana COMAN

Modélisation spatio-temporelle de la pollution atmosphérique urbaine à partir des mesures d'un réseau de surveillance de la qualité de l'air

Soutenue publiquement le 24 septembre 2008

Composition du jury

Rapporteurs :	Mme. Sylvie THIRIA	Professeur, Université de Versailles
	M. Francis ALLARD	Professeur, Université de La Rochelle
Examinateurs :	M. Gilles BERGAMETTI	Directeur de recherche CNRS, Universités Paris 12 et 7
	M. Marc BOCQUET	Chercheur HDR, École des Ponts
	M. Gilles ROUSSEL	Maître de Conférences, Université du Littoral, Côté d'Opale
	M. Gilbert SAPORTA	Professeur, Conservatoire National des Arts et Métiers
Directeur de thèse :	M. Yves CANDAU	Professeur, Université Paris 12
Co-directeur de thèse :	Mlle. Anda IONESCU	Maître de Conférences, Université Paris 12

Centre d'Etudes et Recherches en Thermique, Environnement et Systèmes - CERTES
Université Paris XII – EA 3481

En matière de prévision, le jugement est supérieur a l'intelligence.
L'intelligence montre toutes les possibilités pouvant se produire.
Le jugement discerne parmi ces possibilités celles qui ont le plus de chance de se réaliser.
Gustave le Bon, Hier et Demain

Remerciements

Je tiens à remercier avant tout, les membres du jury qui me font l'honneur de participer à l'examen de ce travail.

J'exprime ma gratitude à Mme. Sylvie Thiria, Professeur à l'Université de Versailles, d'avoir bien voulu être rapporteur et pour toutes ses remarques constructives.

Je remercie M. Francis Allard, Professeur à l'Université de la Rochelle, d'avoir accepté de rapporter sur ce manuscrit et d'avoir examiné minutieusement ce travail.

J'adresse mes chaleureux remerciements à M. Marc Bocquet, Directeur Adjoint au CEREA, à l'École des Ponts, d'avoir accepté d'apporter son regard critique à ce travail. Son désir, de rédiger un rapport sur mon manuscrit, malgré le fait qu'on fait partie de la même École Doctorale de l'Université Paris Est, a été extrêmement touchant.

Je tiens à exprimer ma gratitude à M. Gilbert Saporta, Professeur au Conservatoire National des Arts et Métiers, pour sa participation au jury et d'avoir accepté la "lourde" tâche d'être le président.

Je remercie aussi M. Gilles Bergametti, Directeur de Recherche CNRS aux Universités Paris 12 et 7, et M. Gilles Roussel Maître de Conférences, à l'Université du Littoral, Côté d'Opale pour la gymnastique intellectuelle imposée par ma formation mathématique qui a influencé d'une façon déterminante la redaction de ce manuscrit.

Je remercie tout particulièrement mes deux co-encadrants de thèse : M. Yves Candau, Professeur à l'Université Paris 12, de m'avoir accueillie dans son équipe de recherche au sein du CERTES, de m'avoir communiqué sa rigueur scientifique, et de m'avoir toujours poussée à aller de l'avant et j'ai laissé à la fin Mlle. Anda Ionescu, Maître de Conférences à l'Université Paris 12. Pour tous ses bons conseils, pour son soutien dans les moments de doute, pour s'être montrée plutôt une sœur qu'une chef, toujours inquiète de mon bien être, je ne saurais jamais trop lui remercier.

Un petit mot pour M. Crisan Ionescu : sans lui, rien de tout cela ne serait arrivé. Je lui exprime ici toute ma reconnaissance.

Les années de gestation de cette thèse ne se sont pas déroulées sans peine. Durant des moments difficiles, je me suis largement appuyé sur le soutien, la compréhension et l'amitié de mon entourage. C'est donc le tour de mes amis roumains, français ou "assimilés" d'être mentionnés.

Je vais commencer d'abord par **les amis du CERTES** : merci à tous pour les moments inoubliables de travail et de franche rigolade, pour les encouragements et toutes les discussions que l'on a eues dans la "salle café". Qu'ils soient des titulaires ou des thésards comme moi (la liste serait trop longue pour mentionner tous les noms), les moments passés ensemble resteront gravés dans ma

mémoire, ainsi que ceux passés avec la petite Célia, qui a fait mon bonheur pendant plus d'une année.

Merci à Gilles Forêt, Maître de Conférences à l'Université Paris 12 au LISA, pour ses « coups de pouce », pour sa disponibilité quand j'ai eu le plus besoin, et à toute l'équipe "Modélisation" du LISA qui m'a aidée surtout en ce qui concerne le travail sur le modèle CHIMERE.

Merci aux amis roumains Cécilia et Radu : une fois qu'ils m'ont "adoptée", ils ne m'ont jamais abandonnée ; à Viviana qui a été témoin de "l'accouchement" plus ou moins douloureux de cette thèse ; elle n'a jamais cessé de m'encourager ; ainsi qu'aux "hollandais" Remus et Daniela. Sans Remus je n'aurais jamais fait connaissance avec le monde de l'assimilation de données. Tous les deux m'ont chaleureusement accueillie chez eux pendant mon stage hollandais. Je remercie aussi Manuela et Camelia qui, de près ou de loin, m'ont soutenue pour que ces années de travail soient à la fin couronnées de succès.

Je ne remercierais jamais assez ma famille pour leur soutien indéfectible et toute la patience et la compréhension dont ils ont fait preuve pendant ces années. Cette thèse leur est dédiée.

Table des matières

Introduction générale et objectifs

Contexte et motivations

La pollution atmosphérique est une atteinte à la pureté de l'air et à l'intégrité du fonctionnement de l'atmosphère ; elle correspond à la présence d'une ou de plusieurs substances à des concentrations et durant des périodes suffisantes pour créer un impact négatif sur les êtres vivants et/ou sur l'environnement.

En France, la loi sur l'air et l'utilisation rationnelle de l'énergie (Loi 96-1236 du 30 décembre 1996 intégrée au Code de l'Environnement, LAURE) a conduit à un fort développement de la surveillance de la qualité de l'air. Premièrement, elle a imposé l'élaboration des plans régionaux pour la qualité de l'air (PRQA) en s'appuyant sur une estimation de rejets de polluants dans l'atmosphère et sur une évaluation de la qualité de l'air et de ses effets sur la santé et sur l'environnement.

Il est important de noter que les connaissances en matière de pollution atmosphérique ont fortement progressé dernièrement. D'ailleurs, la loi sur l'air (LAURE) a permis le développement des réseaux dans de petites villes et même des régions rurales éloignées, et l'élaboration des PRQA, qui ont comme objectifs la prévention et la réduction de la pollution atmosphérique. Dans le cadre des PRQA, tous les réseaux doivent se doter d'outils de modélisation pour comprendre le phénomène de dispersion de la pollution atmosphérique, pour simuler des scénarios de pollution, pour estimer l'impact de la pollution sur la santé publique et même prévoir les pics de pollution ; le but est d'obtenir le maximum de renseignements nécessaires pour pouvoir prendre des mesures de réduction des émissions (trafic, industrie) et pour informer le public. Cette loi définit notamment les zones prioritaires à surveiller, et vise à terme une surveillance de l'ensemble du territoire.

Pour diagnostiquer l'état de la pollution atmosphérique, l'idéal serait d'installer autant de stations de mesure qu'il est nécessaire pour identifier les structures spatiales et/ou spatio-temporelles des champs de polluants. Ceci n'est pas concevable car les dispositifs sont assez coûteux, donc, les capteurs ont été placés à des endroits stratégiques : sous le vent des agglomérations (dans le panache urbain), près du trafic ou des aéroports, mais aussi dans la région rurale pour pouvoir étudier la pollution de fond. Ces capteurs fournissent des valeurs horaires de plusieurs polluants atmosphériques.

La question qu'on peut se poser maintenant est : de quel domaine la mesure effectuée est-elle représentative ? Est-ce que le nombre des mesures est suffisant pour réaliser une estimation sur toute une région ? Plusieurs décisions de modélisation sont possibles : premièrement, on peut effectuer une

analyse spatiale par interpolation des valeurs enregistrées, au même moment, par les stations de mesure disponibles. Nous disposons donc d'un certain nombre d'observations, distribuées sur la région d'étude, et on utilise une méthode d'**interpolation spatiale** (par exemple le krigeage) pour obtenir le champ de concentrations d'un polluant atmosphérique et, si possible une évaluation de la variance de l'erreur d'estimation.

Une deuxième possibilité serait d'effectuer une **interpolation spatio-temporelle**. Si les capteurs enregistrent des valeurs horaires, on peut disposer de séries temporelles assez longues pour essayer d'identifier une corrélation spatio-temporelle présente dans les données. En outre, prendre en compte simultanément la distribution spatiale et temporelle du phénomène de dispersion des polluants atmosphériques pourrait améliorer les résultats purement spatiaux.

Troisièmement, la modélisation numérique par des **modèles de chimie-transport** peut nous aider à obtenir une estimation de l'état de la pollution hors des zones couvertes par la mesure (extrapolation). Les modèles déterministes décrivent de manière fondamentale les processus qui interviennent dans l'évolution des concentrations des polluants : le transport, la chimie atmosphérique, la météorologie. Cependant, vue la complexité des phénomènes étudiés, la variabilité spatiale et temporelle des émissions de polluants, quel que soit le modèle, il ne pourra jamais reproduire parfaitement la réalité physique.

Pour améliorer les résultats d'un tel modèle déterministe on peut appliquer une technique appelée **assimilation de données**. L'objectif est de combiner les observations fournies par les stations de mesure et les résultats de simulations d'un modèle de chimie-transport pour estimer, de façon plus réaliste, la valeur moyenne par maille du modèle d'une variable. Il existe deux classes distinctes de méthodes d'assimilation de données : les méthodes séquentielles (basées sur le Filtre de Kalman) et les méthodes variationnelles (comme 4D-VAR). Ces méthodes différent essentiellement par leur coût numérique et la façon de les implémenter.

Mise à part l'estimation, l'autre but de toute étude de la pollution atmosphérique est la prévision. Elle permet de prendre, à court et à moyen terme, des mesures de réduction des émissions afin de limiter l'exposition de la population à des niveaux qui pourraient être nocifs pour la santé. La prévision de pics de pollution est essentiellement effectuée à l'aide soit des **modèles déterministes**, soit des **modèles statistiques** qui utilisent des méthodes de régression, de classification et des réseaux de neurones. Le principal inconvénient de ces dernières est que plus un épisode a été rare par le passé, plus un épisode de même type sera difficile à prévoir. De plus, ils ne permettent pas de comprendre l'origine des pics de pollution prévus, puisqu'ils s'appuient sur de variables corrélées avec la pollution mais pas forcément causales. Les modèles déterministes tridimensionnels de simulation de la qualité de l'air sont mieux adaptés à la compréhension de la pollution atmosphérique que les modèles statistiques, mais ils s'avèrent lourds au niveau du temps de calcul. En outre, les simulations ont mis en évidence une difficulté : le transport d'erreurs de prévision. Cependant, une façon d'éviter la propagation de l'erreur est l'assimilation de données qui corrige à chaque pas de temps le modèle pour qu'il s'approche des observations.

Objectifs

Il existe donc deux directions majeures de recherche concernant la modélisation de la pollution atmosphérique : la première est représentée par le besoin d'une amélioration de la **cartographie** des champs de concentration de polluants atmosphériques en utilisant des méthodes capables de prendre en compte la distribution spatiale et/ou spatio-temporelle des données fournies par les stations de surveillance, et la deuxième est la **prévision** des niveaux de polluants. Ces deux directions ont été abordées dans ce travail.

Cette étude concerne uniquement la pollution à l'échelle régionale et, en particulier, celle urbaine, car en milieu urbain il y a une concentration importante de sources de pollution et, de plus, une grande majorité de la population est exposée à ces conditions. La région choisie est celle d'Île-de-France et les principaux polluants ciblés sont le dioxyde d'azote et l'ozone, dont les mesures horaires sont fournies par AIRPARIF, le réseau de surveillance de la qualité de l'air de la région.

L'objectif principal est d'utiliser le pouvoir informationnel des bases de données fournies par un réseau de qualité de l'air, soit indépendamment, soit pour améliorer les sorties d'un modèle déterministe, dans un but de cartographie et prévision des concentrations des polluants.

Pour atteindre cet objectif, nous avons considéré les variables **espace** et **temps** d'une part *découplées*, et d'autre part comme un *continuum spatio-temporel*, suivant une démarche en quatre étapes, les trois premières visant essentiellement la cartographie, alors que la dernière vise la prévision :

- appliquer les méthodes d'interpolation spatiale sur des jeux de données réelles (concentrations de polluants atmosphériques) pour tester la capacité actuelle du réseau de surveillance de la qualité de l'air dans la région ; cela revient à trouver une bonne corrélation spatiale entre les données disponibles et la modéliser de façon à obtenir une image cohérente avec la réalité représentée par les mesures ;
- étendre la corrélation spatiale, en intégrant la dimension temporelle pour effectuer une interpolation spatio-temporelle des concentrations de polluants mesurées ;
- appliquer des méthodes d'assimilation de données pour corriger les simulations d'un modèle régional de chimie-transport (CHIMERE) en utilisant les observations disponibles ;
- construire des modèles de prévision statistiques de type "boîte noire" (réseaux de neurones) et les appliquer aux séries temporelles de concentrations de polluants pour ainsi quantifier les niveaux de pollution du jour suivant.

Plan du mémoire

Cette thèse est organisée en six chapitres. Le **premier chapitre** consiste en une introduction générale sur l'atmosphère, afin de mieux comprendre les phénomènes de dispersion des polluants atmosphériques, ainsi que sur des notions générales concernant la pollution atmosphérique (classification, sources), les principaux polluants, les indices et les normes existantes au niveau européen.

Le domaine d'étude et les données utilisées sont brièvement présentés à la fin de ce chapitre.

Le **deuxième chapitre** est dédié à la présentation des méthodes d'interpolation spatiale et à leur application sur des jeux de données réelles : des concentrations de polluants atmosphériques enregistrées par les stations de mesure du réseau AIRPARIF. Parmi ces méthodes, nous avons choisi d'appliquer le krigeage, la seule qui tient compte de la structure spatiale des données. Nous comparons trois variantes de krigeage : le krigeage ordinaire, universel et intrinsèque généralisé, et ensuite, nous présentons les résultats obtenus en utilisant ces techniques.

Le **troisième chapitre** concerne la description de la méthode spatio-temporelle d'interpolation employée : le krigeage spatio-temporel et les résultats obtenus en appliquant cette méthode, ainsi qu'une brève comparaison avec les résultats précédents qui prenaient en compte uniquement la distribution spatiale des données.

Le **quatrième chapitre** expose la mise en place d'un premier système d'assimilation séquentielle de données sur le modèle numérique de chimie-transport CHIMERE à l'échelle régionale (toujours en région francilienne). L'objectif est d'utiliser conjointement d'une part, les connaissances sur la physique et la chimie de l'atmosphère, intégrées dans le modèle CHIMERE, et d'autre part, les mesures disponibles de concentrations de polluants atmosphériques, pour améliorer la représentation spatiale des champs de concentrations des polluants. Nous utilisons pour cela un schéma sous-optimal ayant comme principe de fonctionnement le Filtre de Kalman, notamment le Filtre de Kalman d'Ensemble. Les cartes analysées, ainsi que celles originales produites par le modèle seront comparées pour des épisodes précis avec les résultats obtenus par simple interpolation spatiale et/ou spatio-temporelle.

Un modèle déterministe nous offre des prévisions sur les mailles du domaine, mais ces mailles ont des dimensions assez larges. C'est la raison pour laquelle, en complément, nous avons essayé d'obtenir une prévision localement plus précise, sur une seule station de mesure. Le **cinquième chapitre** présente l'application des méthodes statistiques à la prévision, plus précisément une comparaison entre deux types d'architectures neuronales appliquées à la prédiction de l'ozone sur deux sites, l'un situé dans la zone urbaine et l'autre dans une zone rurale, et où on dispose des séries temporelles enregistrées par les stations de mesure. L'horizon de prédiction choisi est de 24 heures. Nous présentons plusieurs indices de performance pour ainsi comparer le pouvoir prédictif des modèles appliqués.

On finit ce mémoire par une **conclusion générale**, mettant en évidence les diverses contributions personnelles ainsi qu'une synthèse de l'ensemble de résultats obtenus, et, enfin, on définit quelques perspectives pour ce travail.

Chapitre 1

Généralités sur la pollution atmosphérique et sur la zone d'étude

Ce premier chapitre présente d'abord quelques généralités concernant l'atmosphère, les principaux polluants, ainsi que leurs effets négatifs sur la santé, et quelques éléments sur la législation existante en France. À la suite de ces généralités, la zone d'étude ainsi que les données utilisées seront brièvement décrites.

1.1 Généralités sur l'atmosphère

1.1.1 Caractéristiques

L'atmosphère est la couche gazeuse qui constitue l'enveloppe de la Terre. Elle est composée de plusieurs couches superposées, mais son épaisseur totale est difficile à préciser car le nombre de molécules de gaz par mètre cube diminue progressivement avec l'altitude. On estime toutefois que 99% de la masse d'air atmosphérique se situe entre le niveau du sol et l'altitude de 30 km.

Les caractéristiques physiques de l'atmosphère telles que la pression et la température subissent des variations importantes lorsqu'on s'éloigne du sol terrestre. La pression atmosphérique diminue, en relation avec la raréfaction progressive en molécules de gaz. La température subit, quant à elle, des variations plus complexes, auxquelles sont associées les couches atmosphériques : troposphère, stratosphère, mésosphère et thermosphère qui définissent la structure verticale qui sera présentée dans la suite.

1.1.2 Structure verticale

La ***troposphère*** est la couche atmosphérique la plus proche du sol terrestre. Son épaisseur est variable : 7 km de hauteur au-dessus des pôles, 18 km au-dessus de l'équateur et environ 13 km, selon les saisons, dans la zone tempérée. C'est dans cette couche qu'on retrouve la plus grande partie des phénomènes météorologiques. Au fur et à mesure qu'on s'élève dans la troposphère, la température décroît de façon régulière d'environ 6 °C tous les 1 000 m pour atteindre −56 °C à la tropopause (zone séparant la troposphère de la stratosphère). Cette couche se divise en deux parties : la couche

libre située à la partie supérieure de la troposphère et la couche limite atmosphérique (CLA), qui est proche de la surface terrestre et dont la hauteur varie dans le temps et dans l'espace. C'est au travers de cette dernière que se font les échanges de masse, d'énergie et d'humidité entre le sol et l'atmosphère. Elle sera décrite plus en détail dans la section 1.2.5.

La ***stratosphère*** se situe au-dessus de la troposphère et c'est ici qu'on trouve la couche d'ozone. Cette dernière est essentielle à la vie sur Terre, car elle absorbe la majorité des rayons solaires ultraviolets qui sont extrêmement nocifs pour tout être vivant. Cette absorption provoque un dégagement d'énergie sous forme de chaleur. C'est pourquoi la température augmente lorsqu'on s'élève dans la stratosphère. Les mouvements de l'air y sont beaucoup moindres. Il s'agit d'un environnement beaucoup plus calme.

La ***mésosphère*** est au-dessus de la stratosphère. Dans cette couche, la température recommence à décroître avec l'altitude pour atteindre $-80\,^{\circ}\mathrm{C}$ à une altitude d'environ 80 km.

La couche la plus haute est la ***thermosphère***. Dans cette couche, la température augmente avec l'altitude et peut atteindre environ $100\,^{\circ}\mathrm{C}$. La thermosphère atteint des milliers de kilomètres d'altitude et disparaît graduellement dans l'espace. C'est la région où près des pôles se forment les aurores boréales et australes et où la pression devient presque nulle.

Pour illustrer ceci, la figure 1.1 présente les différentes couches atmosphériques ainsi que les variations de température et de pression que nous venons de détailler.

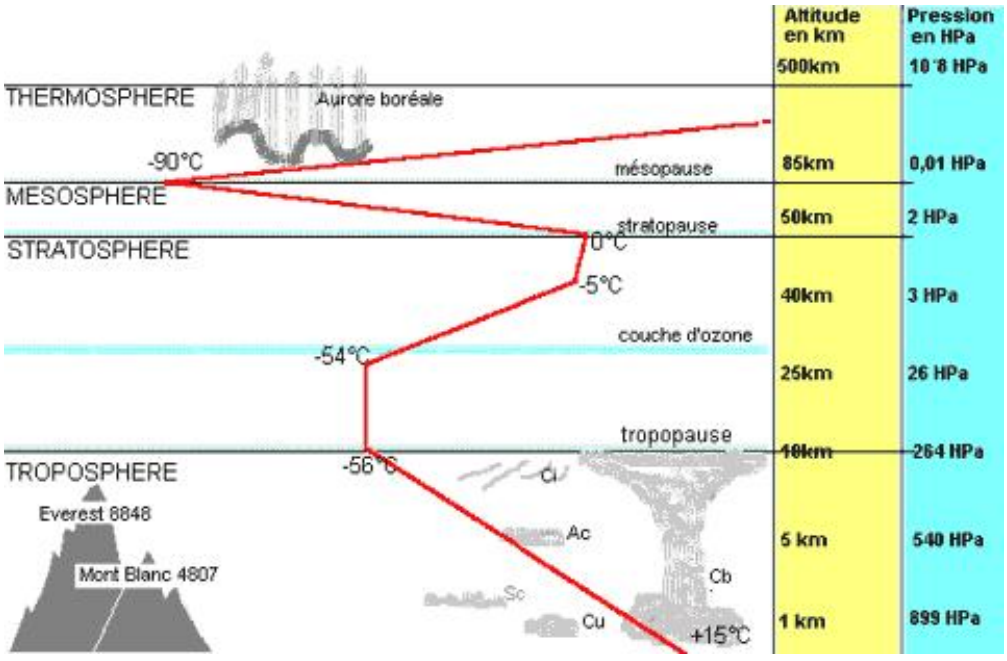


FIG. 1.1: Les couches de l'atmosphère d'après Atmosphère (2008).

1.1.3 Composition de l'atmosphère

L'air, principalement concentré dans la troposphère, est composé essentiellement d'azote et d'oxygène. En pourcentage, l'air sec est constitué de 78,08% d'azote, 20,94% d'oxygène, 0,93% d'argon, et de 0,038% de dioxyde de carbone. S'ajoutent à cela la vapeur d'eau et les aérosols. Les aérosols sont des particules liquides ou solides de taille 0,001 à 10 μm environ, d'origine et de nature variées. Les principales familles d'aérosols sont :

- les aérosols stratosphériques d'acide sulfurique liés en majeure partie aux éruptions volcaniques ;
- les aérosols troposphériques marins produits par les océans ;
- les aérosols désertiques constitués par des poussières minérales ;
- les aérosols anthropiques liés à la pollution urbaine ou aux feux de végétation ;
- les aérosols résultants de transformations chimiques. (IPCC, 2007)

1.2 Phénomènes physico-chimiques qui caractérisent l'atmosphère

Les espèces présentes dans l'atmosphère subissent des transformations liées à des processus dynamiques et chimiques qui seront décrits brièvement par la suite.

1.2.1 Les émissions surfaciques

La majeure partie des émissions est surfacique, c'est-à-dire que les espèces sont émises à une dizaine de mètres du sol. Néanmoins, les cheminées d'usines peuvent atteindre plusieurs dizaines de mètres de hauteur et même les avions volent à une altitude assez élevée. Toutes ces émissions constituent des sources. On reviendra sur les émissions à la fin de ce chapitre (section 1.7.3.2), pour présenter brièvement les axes principaux et les organismes chargés de réaliser l'inventaire des émissions sur le domaine qui nous intéresse.

1.2.2 Convection et advection

La division imaginaire d'un fluide tel que l'air en parcelles contiguës et non morcelées permet de décrire à un instant donné les propriétés de ce fluide sous la forme de grandeurs physiques ou chimiques attachées à chacune de ces parcelles, comme par exemple la masse de la parcelle, sa température, la proportion volumique de l'ozone qui s'y trouve contenu, etc. Or, lorsque le temps s'écoule, le fluide est généralement engagé dans un mouvement : les parcelles, en se déplaçant, effectuent alors un transport des propriétés du fluide, et c'est ce transport qui est appelé l'**advection**[1].

En météorologie, il est important de distinguer le mouvement horizontal du mouvement vertical de l'air, ce qui a conduit à une utilisation spécifique des termes "convection" et "advection" par rapport à leur définition classique.

Le transport horizontal par le vent des propriétés de l'air est appelé en météorologie, sauf

[1] http ://www.meteofrance.com

mention contraire, advection. Quant au transport par la vitesse verticale de l'air, son expression physique prend des formes différentes selon l'échelle temporelle à laquelle on se place : ainsi, aux échelles moyennes et aux petites échelles, l'hypothèse hydrostatique n'est plus observée et les mouvements verticaux sont désignés par le terme de convection.

On distingue deux types de convection : libre et forcée. La convection libre est un mouvement vertical ayant pour cause les différences de masse volumique (dues par exemple aux différences de température) à l'intérieur du fluide. La convection forcée est un mouvement vertical dont les causes sont mécaniques, par exemple l'influence du relief (montagnes).

Pour conclure, en météorologie on considère la convection comme l'expression d'un transport (masse, quantité de mouvement, chaleur et humidité) principalement sur la verticale ; dans la basse atmosphère les transports sont liés aux processus turbulents. L'advection sera le transport des mêmes quantités sur l'horizontale. Influencé par les gradients de pression et de température, ce transport sera décrit par le vent.

1.2.3 La turbulence

A des échelles spatio-temporelles inférieures à la dizaines de mètres et à la minute les mouvements de l'atmosphère ne peuvent plus être prévus de façon déterministe. L'écoulement atmosphérique à ces échelles est instable et son caractère chaotique fait qu'il ne peut être connu que par ses propriétés statistiques : il est turbulent. La turbulence favorise le mélange : elle augmente les transferts de quantité de mouvement, de chaleur ou de masse de plusieurs ordres de grandeur. Elle permet donc les transferts de chaleur du sol vers l'atmosphère et le mélange des polluants.

1.2.4 Le dépôt

Il existe deux types de dépôt, sec et humide, et dans les deux cas ce phénomène représente un puits par lequel les polluants quittent l'atmosphère pour la surface du sol, des bâtiments ou de la végétation. Le dépôt humide a lieu en présence de précipitations qui "lessivent" l'atmosphère en entraînant les polluants vers la surface. Le dépôt sec dépend principalement du degré de turbulence de l'atmosphère, des caractéristiques de la surface et des propriétés chimiques des polluants en contact avec celle-ci.

1.2.5 La couche limite atmosphérique (CLA)

La CLA est la basse couche atmosphérique dont l'écoulement est influencé par les effets thermiques et dynamiques dus à la surface. Son extension verticale dépend des conditions météorologiques (vent, température, humidité, rayonnement solaire) ainsi que de la topographie et du type d'environnement. Un rôle très important est joué par les profils verticaux de température et de vent qui déterminent la hauteur de la couche limite atmosphérique. Globalement, elle reste mince 1 à 2 km si on veut la comparer avec le reste de la troposphère. La partie inférieure de la couche de surface, située au voisinage immédiat du sol, est caractérisée par des champs de vitesses du vent fortement hétérogènes et non-stationnaires et des forces de frottement qui sont prépondérantes. L'intensité du rayonnement solaire parvenant dans la CLA dépend de la quantité de nuages et de leur nature. Elle détermine la quantité d'énergie lumineuse disponible pour les réactions photolytiques, comme celles

qui interviennent au cours du cycle de production/destruction de l'ozone.

Étudier la CLA revient avant toute chose à étudier la turbulence atmosphérique. Elle évolue avec une périodicité diurne, dont les grandes phases sont invariantes dans leur principe, mais dont l'amplitude dépend des processus dynamiques et thermiques liés au domaine étudié. L'évolution de la CLA, au cours de la journée, a donc une influence directe sur les concentrations de polluants atmosphériques.

1.3 Pollution atmosphérique

1.3.1 Définition

Selon la Loi sur l'air et l'utilisation rationnelle de l'énergie de 1996 la pollution atmosphérique est définie comme : "l'introduction par l'homme, directement ou indirectement, dans l'atmosphère et les espaces clos, de substances ayant des conséquences préjudiciables de nature à mettre en danger la santé humaine, à nuire aux ressources biologiques et aux écosystèmes, à influer sur les changements climatiques, à détériorer les biens matériels, à provoquer des nuisances olfactives".

Le terme pollution regroupe une multitude de mécanismes et d'actions dont la conséquence est une dégradation de notre environnement. Depuis le début du siècle, l'accroissement démographique et le développement industriel ont occasionné d'importantes émissions de gaz et d'aérosols. Les modifications de la constitution de l'atmosphère induites peuvent avoir des répercussions aussi bien à l'échelle locale qu'à l'échelle planétaire.

1.3.2 Les différentes échelles

Les problèmes liés à la pollution atmosphérique sont donc rencontrés à plusieurs échelles d'espace et se font ressentir à différentes échelles de temps. Une synthèse des diverses formes de pollution de l'air est présentée dans le tableau 1.1.

Echelle spatiale	**Echelle temporelle**	**Problèmes rencontrés**	**Polluants en cause**	**Principaux effets**
locale	heures	pollutions urbaines	SO_2, NO_x, COV, poussières etc.	santé, corrosion des matériaux
régionale (>100 km)	jours	pluies acides pollution photochimique	SO_2, NO_x COV, CO	dommage sur les milieux et les biens
planétaire	années	couche d'ozone effet de serre	CFC[1], NO_x CO_2, CH_4, N_2O O_3, CFC	santé, végétaux, modification du climat

TAB. 1.1: Les diverses formes de pollution de l'air (Elichegaray, 1997).

En détaillant les trois échelles mentionnées, on commence par le cas de la pollution de proximité (**micro-échelle** ou **échelle locale**), où il s'agit de problèmes rencontrés au voisinage des sources de rejets. On observe directement les effets du chauffage individuel, des fumées et des modes de transport. Les polluants incriminés sont des composés primaires : oxydes de soufre et d'azote, poussières, monoxyde de carbone, composés organiques volatiles (COV), métaux lourds. La surveillance de la pollution de proximité est réalisée par un réseau de surveillance de la qualité de l'air.

Au niveau **régional** (**méso-échelle**), des phénomènes physico-chimiques complexes et variés peuvent intervenir. La pollution à méso-échelle concerne les zones où des phénomènes secondaires apparaissent, comme la pluie acide qui affecte les forêts et les écosystèmes aquatiques, ou bien la production de l'ozone dans les basses couches de l'atmosphère.

Au niveau **planétaire** (**macro-échelle**), les études couvrent des vastes régions où les effets des polluants agissent sur l'ensemble de la planète : réduction de la couche de l'ozone à haute altitude ou encore augmentation de l'effet de serre qui pourrait provoquer d'importants changements climatiques .

1.4 Les principaux polluants atmosphériques

Un polluant est un corps d'origine anthropique ou non, à l'état solide, liquide ou gazeux, contenu dans l'atmosphère et qui ne fait pas partie de la composition normale de l'air ou qui est présent en quantité anormalement grande. Selon leur mode de production, on peut classer les polluants en 2 catégories :

- les polluants primaires, qui sont émis directement dans l'air par des sources identifiables naturelles ou anthropiques ;
- les polluants secondaires, qui sont produits dans l'air par l'interaction de deux ou plusieurs polluants primaires ou par réaction avec les constituants normaux de l'atmosphère, avec ou sans photoactivation.

Pour évaluer les effets de la pollution de l'air, il est nécessaire de prendre en compte trois facteurs : l'émission, le transport et la transformation chimique des polluants. Des paramètres relatifs à la source du polluant (hauteur de rejet, débit, température...), des paramètres météorologiques et climatiques (rayonnement solaire, température, turbulence, vitesse et direction du vent...) et des paramètres topographiques jouent un rôle prépondérant dans le transport et la transformation chimique des polluants.

Les polluants primaires présentent des fortes concentrations à proximité des sources, puis tendent à diminuer au fur et à mesure qu'on s'éloigne de celles-ci du fait de leur dilution dans l'air. Parmi les polluants primaires qui sont mesurés par les organismes de surveillance de la qualité de l'air, on trouve : le monoxyde de carbone (CO), le dioxyde de soufre (SO_2), les oxydes d'azote

[1] Chlorofluorocarbones

(NO_x), les particules en suspension (PS) et en particulier le plomb (Pb), ainsi que les composés organiques volatils (COV). Ces polluants primaires peuvent se transformer dans la basse atmosphère, sous l'action du rayonnement solaire et de la chaleur, en polluants dits secondaires tels que l'ozone et autres polluants photochimiques (PAN ou nitrates de peroxyacétyle, aldéhydes, cétones, etc.).

Une description des effets de ces polluants à court, moyen et long terme figure dans le rapport du Haut Comité de la Santé Publique (HCSP, 2000) ainsi que dans les études toxicologiques et épidémiologiques menées par l'Observatoire régional de la Santé d'Île-de-France (ERPURS, 1997). Quelques informations concernant les divers polluants atmosphériques, notamment sur : leur origine et leurs effets sont résumées dans les sous-sections qui suivent.

1.4.1 Les oxydants atmosphériques

Dans la troposphère, et particulièrement dans la phase gazeuse, la quasi-totalité des réactions chimiques procède par un mécanisme radicalaire. Les radicaux libres sont formés pour la plupart par photolyse de composés minoritaires tels que le dioxyde d'azote, l'ozone et le formaldéhyde (HCHO). Ces composés participent à l'initiation des processus de transformations chimiques de l'ensemble des composés réactifs, dont les COV. Ces réactions sont possibles grâce au rayonnement solaire. Nous trouvons ces oxydants dans l'atmosphère sous forme moléculaire (O_3, nitrate de peroxyacétyle) ou radicalaire (OH^*, HO_2^*, RO_2^*, NO_3^*) à très courte durée de vie. Ces derniers, notamment le radical hydroxyle (OH^*), sont les principaux agents à l'origine des transformations photochimiques dans l'atmosphère (Seinfeld et Pandis, 1998).

1.4.2 Le dioxyde de soufre (SO_2)

- **Origine**

Il provient essentiellement de la combustion de combustibles fossiles contenant du soufre : fuel et charbon (chaudières à bois, hauts fourneaux, moteurs thermiques). Compte tenu du développement du nucléaire, de l'utilisation de combustibles moins chargés en soufre et des systèmes de dépollution des cheminées d'évacuation des fumées, les concentrations ambiantes ont diminué de plus de 50% depuis 15 ans. A l'heure actuelle, les principaux sous-secteurs responsables pour des émissions de dioxyde de soufre sont : le raffinage du pétrole, la production d'électricité et le résidentiel.

- **Réactivité**

En présence d'humidité sous forme de radical OH^*, SO_2 se transforme en acide sulfurique (H_2SO_4) qui contribue au phénomène des pluies acides et à la dégradation de la pierre et des matériaux de certaines constructions.

- **Impact sur la santé**

Du point de vue de l'impact sur la santé, le dioxyde de soufre est un gaz irritant. Le mélange acido-particulaire peut déclencher, selon les concentrations des différents polluants, des effets broncho-spastiques chez l'asthmatique, augmenter les symptômes respiratoires aigus chez l'adulte

(toux, gêne respiratoire), altérer la fonction respiratoire chez l'enfant (baisse de la capacité respiratoire, excès de toux ou de crise d'asthme).

1.4.3 Les poussières ou particules en suspension (PS)

• Origine

Elles constituent un complexe de substances organiques ou minérales. Elles peuvent être d'origine naturelle (volcan) ou anthropique (combustion industrielle ou de chauffage, incinération, véhicules). On distingue les particules fines, provenant des fumées des moteurs diesel ou de vapeurs industrielles recondensées, et les grosses particules provenant des chaussées ou d'effluents industriels (combustion et procédés). Les dimensions des particules sont comprises entre 0,5 μm (considérée comme dimension minimale de division des particules) et 10 μm (limite au-dessus de laquelle les particules se déposent sous l'influence de la gravité). Parmi les particules en suspension, les plus connues et étudiées sont les PM_1, les $PM_{2.5}$ et les PM_{10}[2].

• Réactivité

Les particules les plus fines peuvent transporter des composés toxiques dans les voies respiratoires inférieures (sulfates, métaux lourds, hydrocarbures). Elles accentuent ainsi les effets des polluants acides, dioxyde de soufre et acide sulfurique notamment.

• Impact sur la santé

Les plus grosses sont retenues par les voies aériennes supérieures. Les plus fines, à des concentrations relativement basses, peuvent, surtout chez l'enfant, irriter les voies respiratoires ou altérer la fonction respiratoire. Certaines particules ont des propriétés mutagènes et cancérigènes : c'est le cas de certains hydrocarbures aromatiques polycycliques (HAP). Des recherches sont actuellement développées pour évaluer l'impact des composés émis par les véhicules diesel.

1.4.4 Les oxydes d'azote (NO_x)

• Origine

Les oxydes d'azote présents dans la troposphère sont principalement émis sous forme de monoxyde d'azote NO, lors de la combustion de combustibles fossiles ou de biomasse. Le NO se transforme rapidement en dioxyde d'azote (NO_2) et on peut évaluer alors le contenu en oxydes d'azote par la somme $NO + NO_2$. Ils proviennent surtout du transport routier et des installations de combustion (centrales énergétiques, ...). Le monoxyde d'azote et le dioxyde d'azote font l'objet d'une surveillance attentive dans les centres urbains. L'utilisation d'un pot catalytique permet une diminution des émissions de chaque véhicule. Néanmoins, les concentrations dans l'air ne diminuent guère compte tenu de l'âge et de l'augmentation forte du parc et du trafic automobiles.

• Réactivité

Les niveaux de NO_x induisent la formation ou la perte d'ozone dans la basse atmosphère. Ils

[2]PM = Particulate Matter, la dimension correspond au diamètre aérodynamique (aéraulique) maximal des particules

contribuent également au phénomène des pluies acides (l'acide nitrique HNO_3 en solution donne des ions H^+ responsables de l'acidité des pluies et des ions nitrates).

• **Impact sur la santé**

Le NO_2 pénètre dans les plus fines ramifications des voies respiratoires. Il peut, dès 200 $\mu g.m^{-3}$, entraîner une altération de la fonction respiratoire et une hyperréactivité bronchique chez l'asthmatique et chez les enfants, et augmenter la sensibilité des bronches aux infections microbiennes.

1.4.5 Les Composés Organiques Volatiles (COV)

• **Origine**

On entend par Composé Organique Volatil (COV)[3] tout composé organique à l'exclusion du méthane ayant une pression de vapeur de 0,01 kPa ou plus à une température de 20 °C. Leur origine est à la fois anthropique et biotique ; il s'agit d'**hydrocarbures** (émis par évaporation des bacs de stockage pétroliers, remplissage des réservoirs automobiles), de **composés organiques** (provenant des procédés industriels ou de la combustion incomplète des combustibles), de **solvants** (émis lors de l'application des peintures, ou d'encres, le nettoyage des surfaces métalliques et des vêtements), ou bien de **composés organiques** émis par l'agriculture et par le milieu naturel.

• **Réactivité**

Ils interviennent dans le processus de formation d'ozone dans la basse atmosphère.

• **Impact sur la santé**

Les effets sont très divers selon les polluants : ils vont de la simple gêne olfactive à des irritations cutanées et des réactions allergiques. En cas d'exposition aiguë (forte concentrations pendant une durée assez brève) les COV provoquent des irritations des voies respiratoires et digestives. Le benzène est reconnu comme cancérigène et en situation d'exposition chronique peut induire le développement de plusieurs types de cancers.

1.4.6 Le monoxyde de carbone (CO)

• **Origine**

Il provient de la combustion incomplète des combustibles et carburants. Des taux importants de CO peuvent être rencontrés quand le moteur tourne dans un espace clos (garage) ou quand il y a une concentration de véhicules qui roulent au ralenti dans des espaces couverts (tunnel, parking), ainsi qu'en cas de mauvais fonctionnement d'un appareil de chauffage.

• **Réactivité**

Il contribue à la formation de l'ozone.

• **Impact sur la santé**

Il se fixe à la place de l'oxygène sur l'hémoglobine du sang conduisant à un manque d'oxygénation du système nerveux, du cœur, des vaisseaux sanguins. A des taux importants, et à doses

[3] Arrêté du 2 février 1998

répétées, il peut être à l'origine d'intoxication chronique avec céphalées, vertiges, asthénie, vomissements. En cas d'exposition très élevée et prolongée, il peut être mortel ou laisser des séquelles neuropsychiques irréversibles.

1.4.7 Les métaux lourds : plomb, cadmium, vanadium, mercure (Pb, Cd, V, Hg)

• Origine

Le plomb a été employé dans l'essence du fait de ses propriétés antidétonnantes. Depuis quelques années, les essences sans plomb ou à teneurs réduites en plomb ont permis d'abaisser les teneurs dans l'air très en deçà des seuils de nuisance. Le cadmium a des origines très diverses, essentiellement industrielles. Le vanadium est un métal blanc, brillant qui possède une bonne résistance à la corrosion. Les émissions naturelles de vanadium sont dues principalement à l'activité volcanique, tandis que celles liées à l'activité humaine proviennent surtout de la métallurgie. Les principales sources d'émissions de mercure sont la combustion du charbon, les activités minières et les incinérateurs ainsi que les recyclage (thermomètres, lampes au mercure).

• Impact sur la santé

Ces métaux ont la propriété de s'accumuler dans l'organisme, engendrant ainsi un risque de toxicité à long terme impliquant d'éventuelles propriétés cancérigènes. Le plomb est un toxique neurologique, rénal et du sang. Le cadmium a un effet négatif sur l'appareil rénal, mais il est aussi un irritant respiratoire. Le vanadium est essentiellement un toxique respiratoire qui peut conduire, selon les concentrations, à une simple irritation ou à des lésions pulmonaires plus graves. En ce qui concerne le mercure, il est toxique sous toutes ses formes. D'abord, sous sa forme vapeur, il attaque les voies respiratoires, ensuite il se solubilise dans le plasma, le sang et l'hémoglobine. Quand il atteint la circulation sanguine, le mercure attaque les reins, le cerveau et le système nerveux.

1.4.8 L'ozone (O_3)

• Origine

Il existe un cas particulier de polluant atmosphérique : l'ozone. Il est essentiel de distinguer le "bon" ozone stratosphérique (90% de l'ozone) du "mauvais" ozone troposphérique (10%). Celui stratosphérique, présent à une altitude comprise entre 13 et 30 km, constitue la couche d'ozone. Il est utile dans le sens où il absorbe les rayons ultraviolets de longueurs d'onde comprises entre 230 et 300 nm, nocifs pour la matière vivante. Néanmoins, depuis quelques années, la couche d'ozone stratosphérique est en train de diminuer. Plusieurs campagnes scientifiques ont été lancées pour identifier les processus de destruction de l'ozone, cette diminution étant associée aux effets d'émissions anthropiques comme les chlorofluorocarbures (CFC).

La première idée sur l'origine de l'ozone troposphérique a été de dire que l'ozone stratosphérique, plus lourd que l'air, descendait et que cela constituait la principale source de l'ozone troposphérique. Mais si ce phénomène de descente est en partie responsable pour une quantité d'ozone troposphérique, la plus grande majorité peut et doit être attribuée aux réactions catalysées

par la lumière solaire entre plusieurs polluants précurseurs (principalement, les oxydes d'azote et les composés organiques volatils).

Les concentrations d'ozone sont particulièrement dépendantes de la lumière du soleil ; par conséquent, les épisodes de forte pollution sont toujours susceptibles de se développer après des périodes soutenues de chaleur et un temps calme. La concentration d'ozone fait partie d'un bilan permanent entre les mécanismes de destruction, la production photochimique et l'alimentation en fond permanent d'ozone dans la troposphère. Il faut noter aussi que les transformations photochimiques s'effectuent durant le transport par les vents. Sous l'influence des émissions des véhicules en milieu urbain, les polluants primaires favorisent la destruction de l'ozone par réaction avec le monoxyde d'azote. Les pics de pollution d'ozone se produisent généralement en période estivale et en zone périurbaine à distance des zones d'émission des polluants primaires. La lutte contre la pollution atmosphérique en milieu urbain passe donc aussi par une meilleure connaissance des phénomènes de production-destruction de l'ozone.

• L'équilibre photochimique

Les réactions chimiques présentant la production et la destruction de l'ozone dans la troposphère sont très nombreuses. Les principales étapes du cycle de l'ozone incluent quelques espèces chimiques notamment : le monoxyde d'azote (NO), le dioxyde d'azote (NO_2), le monoxyde de carbone (CO), les composés organiques volatils (COV), les radicaux hydroxyles (OH) et hydropéroxyles (HO_2) au cours d'un processus rapide, l'équilibre photochimique, et d'un processus plus lent, le cycle des radicaux. On détaille par la suite le premier processus, l'équilibre photochimique.

Une molecule d'ozone est formée par la combinaison d'un atome d'oxygène dans l'état fondamental ($O^{(3p)}$) avec une molécule de dioxygène et avec un troisième corps, M, permettant de stabiliser la réaction :

$$O_2 + O^{(3p)} + M \rightarrow O_3 + MM. \tag{1.1}$$

L'atome $O^{(3p)}$ provient souvent d'une molécule de NO_2, photodissociée par un rayonnement d'une longueur d'onde comprise entre 290 et 400 nanométres (visible et proche ultra-violet) :

$$NO_2 + h\nu \rightarrow NO + O^{(3p)}. \tag{1.2}$$

La troisième réaction est appelée **titration** de l'ozone par NO et implique une molécule de NO qui peut réagir avec une molécule d'ozone pour créer le NO_2 :

$$NO + O_3 \rightarrow NO_2 + O_2. \tag{1.3}$$

Ces trois équations ne peuvent pas expliquer la production nette d'ozone qui conduit à des concentrations parfois très élevées. Pour expliquer ces niveaux il faut faire appel au cycle de radicaux, trop long pour être expliqué dans ce paragraphe, mais qui peut être trouvé dans Seinfeld (1986).

• Impact sur la santé

L'ozone troposphérique est un polluant. Lors d'une exposition prolongée, il provoque des

irritations oculaires, de la toux et une altération pulmonaire, surtout chez les enfants et les asthmatiques. Il contribue aux pluies acides ainsi qu'à l'effet de serre.

Comme on peut voir dans cette section, les polluants sont plus ou moins nuisibles à la santé humaine, selon leur concentration dans l'air et la sensibilité de chaque individu. Par conséquent, les législations nationales et internationales (Programme Air pur pour l'Europe CAFE, la Directive 96/62/CE, ainsi que les Directives filles) ont été créées afin de contrôler la quantité de polluants émise dans l'atmosphère et de s'assurer que les objectifs pour améliorer la qualité ambiante soient atteints (voir section 1.6).

1.5 Mesure des polluants. Incertitudes de mesures

Dans un but de modélisation, il est très important de connaître la façon dont on effectue les mesures des polluants ainsi que leurs incertitudes associées. À titre d'exemple, on présente dans cette sous-section les incertitudes associées aux polluants étudiés, valeurs qui ont été rapportées par AIRPARIF[4], l'association chargée de surveiller la qualité de l'air sur l'ensemble de la région Île-de-France (ESQUIF, 2001).

Effectuée à l'aide des stations automatiques, la mesure d'ozone est effectuée par absorption dans l'ultraviolet (253,7 nm). L'air échantillonné est envoyé dans une chambre optique soit directement, soit après passage sur un filtre éliminant l'ozone. La mesure de l'absorption due à l'ozone est déterminée par différence entre l'absorption UV de l'échantillon et l'absorption UV de l'échantillon exempt d'ozone.

La mesure de NO se fait par chimiluminescence avec l'ozone. Un échantillon d'air est mélangé à l'ozone dans une chambre optique. L'ozone réagit avec le NO pour former le dioxyde d'azote sous une forme excitée. En se désexcitant, le dioxyde d'azote émet un rayonnement. Le signal émis est mesuré par un photomultiplicateur et permet de calculer le taux de NO présent.

Le dioxyde d'azote NO_2 n'est pas mesuré directement. On mesure d'abord le contenu en NO de l'échantillon et après le NO_2 est converti en NO grâce à un four de catalyse. Le monoxyde d'azote est alors mesuré une deuxième fois par chimiluminescence avec l'ozone. Cette mesure donnera la concentration totale en oxydes d'azote NO_x de l'échantillon initial. On retrouve ensuite la concentration de NO_2 par différence entre les mesures de NO_x et de NO. Pour les unités de mesure de la concentration il faut se reporter à l'annexe A.

L'incertitude de la mesure est présentée dans le tableau 1.2 :

[4] www.airparif.asso.fr

Concentration (ppb)	Incertitudes (ppb)	
	O_3	NO_2
10	$\pm 3,1$	$\pm 6,8$
30	$\pm 3,1$	$\pm 6,8$
50	$\pm 3,6$	$\pm 7,2$
70	$\pm 4,2$	$\pm 8,2$
90	$\pm 4,9$	$\pm 9,4$
100	$\pm 5,3$	$\pm 10,0$
150	$\pm 7,6$	$\pm 13,8$
200	$\pm 10,0$	$\pm 18,0$
250	$\pm 12,5$	$\pm 22,4$

TAB. 1.2: Incertitude de mesure pour différentes concentrations d'ozone et de NO_2 (AIRPARIF, 2007)

1.6 Législation sur la pollution atmosphérique et les réseaux de surveillance

1.6.1 Législation

En France, la loi sur l'air et l'utilisation rationnelle de l'énergie (Loi 96-1236 du 30 décembre 1996 intégrée au Code de l'Environnement) a conduit à un développement de la surveillance de la qualité de l'air. Premièrement, elle stipule que l'élaboration des plans régionaux pour la qualité de l'air doit s'appuyer sur une estimation des rejets de polluants dans l'atmosphère et sur une évaluation de la qualité de l'air et ses effets sur la santé et sur l'environnement.

Des organismes chargés de réaliser l'étude de la qualité de l'air sont présents dans toutes les agglomérations de plus de 100 000 habitants. Ces réseaux de surveillance sont placés sous la responsabilité des Associations Agréées de Surveillance de la Qualité de l'Air (AASQA). Les présidents de ces associations ont décidé de la création d'une association fédératrice, appelée ATMO[5], qui permet une meilleure homogénéisation et une certaine efficacité et cohérence au niveau de décisions. Un autre organisme très important, qui est chargé de réaliser et de diffuser des inventaires d'émissions polluantes de toutes sources, est le Centre Interprofessionel Technique de la Pollution Atmosphérique (CITEPA[6]) créé en 1961. Les inventaires sont réalisés à l'échelle départementale et régionale environ tous les cinq ans et sont appelés "cadastres d'émissions". Les deux autres organismes importants à mentionner sont l'Institut de l'Environnement Industriel et des Risques (INERIS[7]) qui a pour mission d'évaluer et de prévenir les risques accidentels ou chroniques pour l'homme et l'environnement liés aux installations industrielles, aux substances chimiques et aux exploitations souterraines, et l'Agence de l'Environnement et de la Maîtrise de l'Énergie (ADEME[8]).

[5] http ://atmo-france.org
[6] www.citepa.org
[7] www.ineris.fr
[8] www.ademe.fr

Il est important de noter que les connaissances en matière de pollution atmosphérique ont fortement progressé dernièrement. D'ailleurs, la loi sur l'air a permis le développement du réseau des petites villes et mêmes des régions rurales éloignées, et l'élaboration des Plans Régionaux pour la Qualité de l'Air (PRQA), qui ont comme objectifs la prévention et la réduction de la pollution atmosphérique. Dans le cadre des PRQA, tous les réseaux doivent se doter d'outils de modélisation pour comprendre le phénomène de dispersion de la pollution atmosphérique, simuler des scénarios de pollution, estimer l'impact de la pollution sur la santé publique et même prévoir les pics de pollution pour prendre des mesures de réduction des émissions (trafic, industrie) et pour informer le public. Cette loi définit notamment les zones prioritaires à surveiller, et vise à terme une surveillance de l'ensemble du territoire.

Les normes de la pollution atmosphérique ont été établies dans le cadre des directives relatives à la fixation des valeurs limites pour les polluants atmosphériques les plus connus. La Directive Cadre 96/62/EC sur la qualité de l'air ambiant et les Directives filles (1999/30/CE pour le NO_2 et 2002/3/CE relative à l'ozone) fixent un cadre réglementaire de surveillance de la qualité de l'air. Les exigences formulées dans ces directives sont de plusieurs ordres. Il existe trois niveaux pour déclencher l'alerte de la population en cas d'épisode de pollution par un polluant : le seuil de protection de la santé, le seuil d'information de la population et le seuil d'alerte de la population.

Le niveau d'information et de recommandation est déclenché lorsque le niveau de concentration de polluants dans l'atmosphère atteint un certain seuil qui risque d'avoir des effets limités et transitoires sur la santé des personnes particulièrement sensibles. Lors du déclenchement du seuil d'information, plusieurs mesures sont mises en place :

- le renforcement des contrôles antipollution,
- le renforcement des contrôles de vitesse,
- la vérification des contrôles techniques obligatoires,
- les recommandations faites aux automobilistes de réduire leur vitesse de 20 km/h sur certaines voies,
- l'instauration de mesures tarifaires pour le stationnement résidentiel.

Le niveau d'alerte est déclenché lorsque le niveau de concentration des polluants atteint un seuil au-delà duquel une exposition de courte durée présente un risque pour la santé ou pour l'environnement. Les recommandations sanitaires sont les mêmes que pour le niveau d'information et de recommandation. Cependant, selon la gravité de l'épisode, des mesures de restriction voire de suspension des activités, peuvent être prises, comme :

- la réduction des vitesses obligatoires sur certaines voies,
- l'immobilisation de 10% des véhicules du parc des administrations et des services publics,
- l'interdiction de la circulation de transit des poids lourds,
- la mise en place de la circulation alternée et gratuité des transports en commun.

Dans le tableau 1.3 on présente le récapitulatif des objectifs de qualité, les valeurs limites, les seuils de recommandation et d'information et les seuils d'alerte pour les principaux polluants mesurés par AIRPARIF (2007).

1.6.2 Les stations de mesure

Les stations de mesure sont, elles aussi, classées selon des critères européens qui tiennent compte de leur emplacement. Ainsi, on peut trouver six types de stations :

- urbaines et péri-urbaines (avec une densité de population minimum 4 000 habitants par km^2 dans un rayon de 1 km autour de la station) ;
- trafic (situées près des voies de circulation) ;
- industrielles (localisées au voisinage d'installations telles que les centrales thermiques ou les unités d'incinération d'ordures ménagères) ;
- rurales (installées à 50 km en moyenne des agglomérations et qui peuvent suivre les phénomènes de transfert de pollution par l'action du vent) ;
- stations d'observation qui ne répondent pas aux critères des stations précédentes, mais qui sont utilisées pour completer le réseau, car essentielles à la compréhension des phénomènes de pollution.

Après avoir vu le cadre législatif concernant la qualité de l'air et la classification des diverses stations de mesure, nous présentons dans la suite la zone d'étude à laquelle nous nous sommes intéressés, ainsi que les données que nous avons utilisées.

1.7 Zone d'étude et données disponibles

1.7.1 La topographie

Dans le creux central du Bassin parisien, la région Île-de-France se présente comme un immense carrefour de voies naturelles avec notamment de nombreuses convergences fluviales (voir figure 1.2). L'érosion climatique et fluviale a modelé le paysage de plateaux, plaines, buttes et vallées. D'origine sédimentaire, le relief de l'Île-de-France est caractérisé par la prédominance de surfaces quasi horizontales. Néanmoins, ces surfaces, plateaux et plaines, ne sont pas monotones : il est rare de rencontrer des paysages aussi variés sur de si courtes distances.

1.7.2 Le climat d'Île-de-France

Cette région est le théâtre de nombreux événements météorologiques qui sont plutôt favorables à la dispersion des polluants. Comparé aux autres types de climat français, celui de l'Île-de-France est caractérisé par une certaine modération, pratiquement dans tous les domaines. En effet, l'Île-de-France se trouve dans un bassin, aux limites des influences océaniques à l'ouest et continentales

Polluant	Type d'objectif	Variables utilisées	Période de référence	Valeur	Incertitude exigée
SO_2	Valeur limite : protection de la santé humaine	Percentile 99.7 de concentrations horaires	Année civile	$380\,\mu g.m^{-3}$	15%
	Valeur limite : protection des écosystèmes	Moyenne annuelle	Année civile	$20\,\mu g.m^{-3}$	15%
	Seuil de recommandation et d'information	Moyenne horaire		$300\,\mu g.m^{-3}$	15%
	Seuil d'alerte	Moyenne horaire	3 heures consécutives	$500\,\mu g.m^{-3}$	15%
NO_2	Valeur limite : protection de la santé humaine	Percentile 98 de concentrations horaires	Année civile	$200\,\mu g.m^{-3}$	15%
	Seuil de recommandation et d'information	Moyenne horaire		$200\,\mu g.m^{-3}$	15%
	Seuil d'alerte	Moyenne horaire		$400\,\mu g.m^{-3}$	15%
O_3	Seuil de recommandation et d'information	Moyenne horaire		$180\,\mu g.m^{-3}$	15%
	Seuil d'alerte	Moyenne horaire	3 seuils : 3 heures conséc 3 heures conséc 1 heure	$240\,\mu g.m^{-3}$ $300\,\mu g.m^{-3}$ $360\,\mu g.m^{-3}$	15%
CO	Valeur limite : protection de la santé humaine	Moyenne 8 heures	Année civile	$10\,mg.m^{-3}$	15%
Benzène	Valeur limite	Moyenne annuelle	Année civile	$5\,\mu g.m^{-3}$	25%
PM_{10}	Valeur limite	Moyenne annuelle	Année civile	$40\,\mu g.m^{-3}$	25%
	Valeur limite	Moyenne journalière	Année civile	$50\,\mu g.m^{-3}$	25%
Pb	Valeur limite	Moyenne annuelle	Année civile	$500\,ng.m^{-3}$	25%

TAB. 1.3: Les diverses valeurs seuil pour les polluants atmosphériques (AIRPARIF, 2007) dérivées de la Directive Cadre et de Directives filles CE.

à l'est. On rencontre donc les deux types de climat en alternance, mais l'influence océanique a tendance à prendre le dessus. Le climat océanique, venteux ou pluvieux, est favorable à la dispersion

FIG. 1.2: Le relief de la région Île-de-France

de la pollution par brassage et lessivage de l'atmosphère. Cependant, certaines situations météorologiques, anticyclones et absence du vent, inversion de température, bloquent les polluants sur place et peuvent conduire pour les mêmes émissions de l'agglomération, à des niveaux nettement supérieurs à ceux des jours les moins pollués. Ainsi, à partir d'émissions de polluants équivalentes en lieu et en intensité, les niveaux de polluants dans l'environnement peuvent varier d'un facteur vingt suivant les conditions météorologiques.

1.7.2.1 La température

Si on regarde les températures moyennes annuelles sur la zone d'étude, on remarque tout de suite l'îlot de chaleur produit par l'agglomération parisienne et provoqué par l'omniprésence des surfaces bétonnées, des chauffages urbains, et de l'asphalte. La différence entre le centre de Paris et la banlieue lointaine dépasse 2,5 °C en moyenne annuelle, ce qui est considérable. La différence de température avec la banlieue et surtout la campagne, est notamment plus sensible en fin de nuit. Lorsque le vent est faible et le ciel dégagé, elle peut atteindre 7° à 8 °C ! En revanche, l'après-midi, elle ne dépasse généralement pas 2° à 3 °C (PPA, 2007).

1.7.2.2 Le vent

La région d'Île-de-France n'est pas réputée pour être une région très venteuse. Toutefois, sa position assez proche des influences maritimes l'expose à un certain nombre de phénomènes violents. De fortes rafales de vent peuvent être observées en toutes saisons. Les vents dominants soufflent du sud-ouest (surtout en hiver et en automne). Les vents du nord-est (bise) sont également assez fréquents (notamment en hiver et en été). En revanche les vents ne viennent que très rarement du sud-est. Il ne s'agit bien souvent que de phases très temporaires.

1.7.2.3 Les précipitations

Ceci peut paraître assez paradoxale mais l'Île-de-France est l'une des régions les plus sèches de France ; du moins si l'on tient compte de la quantité de précipitations qui tombe sur l'ensemble d'une année (600 mm d'eau par an à Paris alors que la moyenne nationale est d'environ 750 mm). Le nombre moyen de jours de pluie ou de neige est en revanche beaucoup plus important et au-dessus de la moyenne nationale et oscille entre 160 et 170 par an, ce qui représente en moyenne un jour sur deux (PPA, 2007).

1.7.2.4 L'ensoleillement

Si on la compare à d'autres régions de France, l'Île-de-France n'est pas très ensoleillée avec environ 1700 h par an, pour une moyenne nationale d'environ 1850 h. Le minimum d'ensoleillement est observé en décembre, à la fois parce que les journées sont courtes mais également très grises - la part de l'ensoleillement n'est en effet que de 20% et le nombre de jours où le ciel reste totalement couvert s'élève à 13. Le mois le plus ensoleillé est août avec 51% de part de soleil en moyenne sur une journée et seulement un jour de ciel couvert en permanence (PPA, 2007).

1.7.3 Les données disponibles

1.7.3.1 Les mesures surfaciques

Dans cette étude, on a utilisé les mesures effectuées par le réseau de surveillance de la qualité de l'air de la zone d'Île-de-France : AIRPARIF[9]. Ce réseau possède actuellement 47 stations réparties en Île-de-France, dont 12 à Paris, qui mesurent des indicateurs caractéristiques des sources de pollution. Parmi ces 47 stations, on retrouve 40 stations de fond, dont 32 urbaines ou périurbaines et 8 rurales (régionales), 6 stations de proximité (ou trafic) qui mesurent ce que respirent les piétons, les cyclistes et les automobilistes dans le flux de circulation et une seule station d'observation au 3ème étage de la Tour Eiffel (voir les figures 1.3 et 1.4).

Deux laboratoires mobiles viennent compléter ce dispositif et permettent des campagnes périodiques de mesures là où une station permanente de mesure n'est pas justifiée (étude d'impact d'installations ou d'industries, mesures au milieu du trafic...). Ils donnent également la possibilité de vérifier si les stations fixes sont bien représentatives de la qualité de l'air étudié. Ils sont un

[9] www.airparif.asso.fr

précieux outil de recherche pour mieux connaître la nature et l'intensité de la pollution régionale. Ils permettent également de valider de futurs sites de stations fixes par campagnes de mesure de deux à trois semaines.

Les coordonnées des stations de mesure exprimées en latitude, longitude ont été obtenues par l'intermédiaire de la BDQA[10] (Base de Données de la Qualité d'Air) et converties avec le logiciel de l'IGN[11], CIRCE 2000, en coordonnées Lambert I Nord.

Les données utilisées sont des concentrations moyennes horaires, exprimées en $\mu g.m^{-3}$, de dioxyde d'azote (NO_2) et d'ozone (O_3), enregistrées par les stations automatiques d'AIRPARIF, couvrant diverses périodes : l'été 1999 et l'année 2000. Ces données, disponibles sur le site d'AIRPARIF, sont issues des stations automatiques situées dans la région d'Île-de-France.

Pour le dioxyde d'azote, la zone d'étude choisie correspond grossièrement à Paris et à la petite couronne, au cœur de l'agglomération, là où les concentrations sont les plus élevées et la répartition spatiale des stations plus uniforme. Il existe 15 stations, dont l'emplacement est représenté dans la figure 1.4.

FIG. 1.3: Les stations d'AIRPARIF situées sur la grande couronne

En revanche, pour l'ozone, la zone d'étude est plus large, étant donné qu'une observation d'ozone en milieu rural ou en altitude est souvent plus représentative des quelques kilomètres qui l'entourent qu'une observation en milieu urbain, influencée plus par les phénomènes locaux. Dans ce cas, les données correspondent uniquement aux mesures de pollution de fond qui mettent en évidence les caractéristiques générales et les tendances de la pollution atmosphérique à l'échelle urbaine. Il faut noter que certaines grandes zones ne possèdent pas ou possède très peu de stations

[10] www.atmonet.org
[11] www.ign.fr

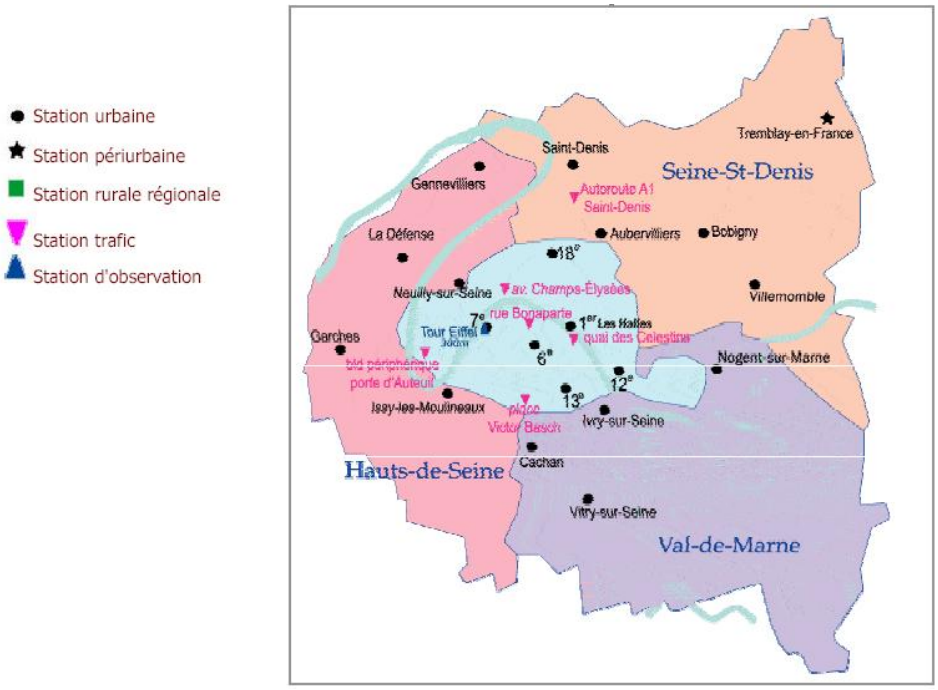


FIG. 1.4: Les stations d'AIRPARIF situées à Paris et la petite couronne

hors des zones urbaines ; dans ce cas, on a opté pour des stations péri-urbaines (voir la figure 1.3).

1.7.3.2 Les émissions de polluants atmosphériques

La deuxième source de données, pour la modélisation ultérieure de la pollution atmosphérique en Île-de-France, sont les émissions. Comme évoqué précédemment, c'est le CITEPA l'organisme chargé par la Direction de la prévention des pollutions et des risques (DPPR) pour effectuer l'inventaire des émissions de polluants atmosphériques. Notons qu'il existe deux sources principales d'émissions : anthropiques et biogéniques. Pour les premières, les totaux annuels d'émissions fournis par le CITEPA pour chaque polluant sont d'abord distribués spatialement par AIRPARIF, suivant le type de source et par secteur d'activité. Sur la région parisienne, les émissions dues au trafic sont prépondérantes avec des taux de 60% pour les NO_x et 55% pour le COV. Pour exemplifier, on présente les cadastres d'émissions de NO_x (figure 1.5) et COV (figure 1.6) pour l'année 2000 qui sont donnés par AIRPARIF.

Une fois les totaux annuels distribués dans l'espace, ils sont distribués temporellement en utilisant les profils temporels de GENEMIS[12] (diurnes, par semaine, et saisonniers).

En revanche, pour les émissions biogéniques, les seules sources disponibles, dues à Simpson et al. (1995), ont une résolution spatiale très faible et seul l'isoprène (COV) a été pris en compte.

[12]Generation and Evaluation of Emission Data, http ://www.ier.uni-stauttgart.de/forschung/projektwebsites/genemis/

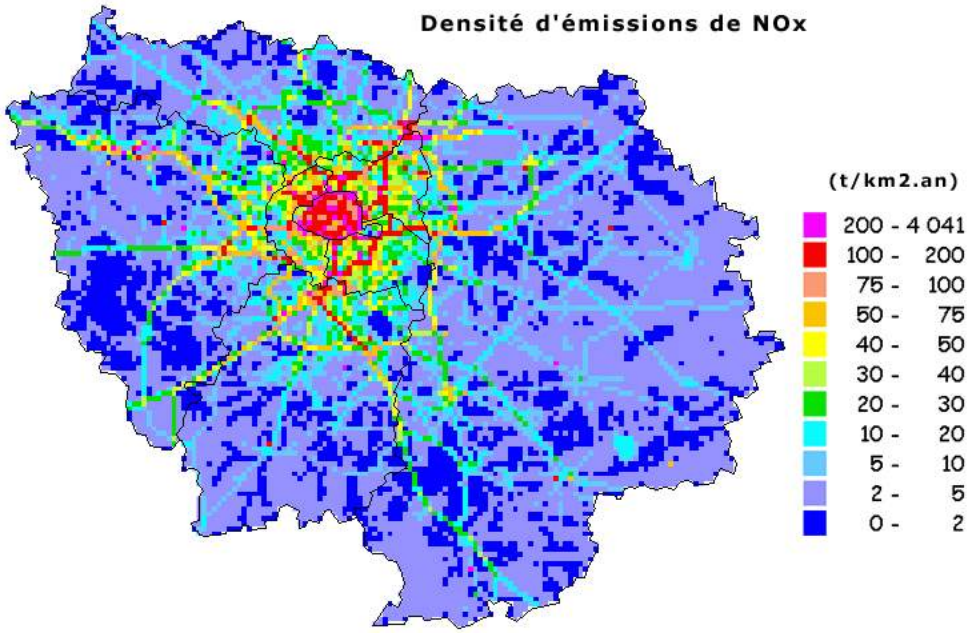


FIG. 1.5: Les émissions de NO_x sur la grande couronne présentées par AIRPARIF.

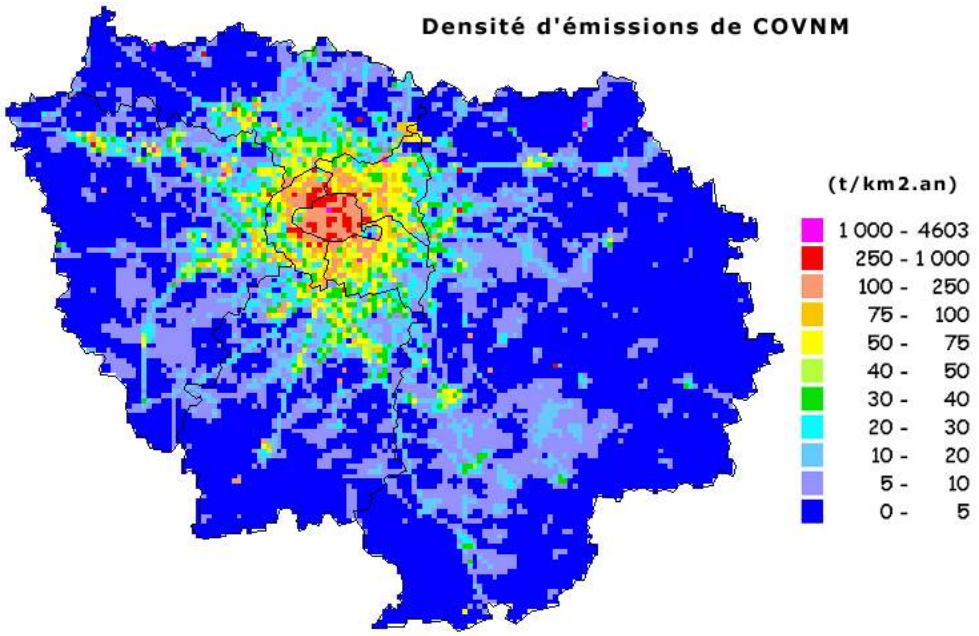


FIG. 1.6: Les émissions de COVNM (COV non-méthanique) sur la grande couronne présentées par AIRPARIF.

1.8 Conclusion du chapitre

L'objet de ce premier chapitre a été de donner d'abord une brève description de l'atmosphère afin de mieux comprendre les phénomènes de dispersion des polluants atmosphériques, qui ont été présentés par la suite (origine, impact sur la santé, législation appliquée en France).

La deuxième partie du chapitre a été consacrée à la description de la zone d'étude et des données utilisées. Ces données constituent la base de la modélisation spatiale ou spatio-temporelle des champs de concentrations de polluants, qui fera l'objet des chapitres suivants.

Chapitre 2

Interpolation spatiale

Ce chapitre présente d'abord une synthèse des principales méthodes d'interpolation spatiale : statistiques (plus précisément géostatistiques) et analytiques. L'objectif principal de cette présentation est de mettre en évidence les différences existantes entre les hypothèses nécessaires à l'application de chaque groupe de méthodes, ainsi que de tracer le cadre général de l'utilisation des modèles probabilistes dans l'analyse de la dispersion des polluants atmosphériques en zone urbaine (une introduction à la géostatistique linéaire). La deuxième partie sera consacrée à l'application des méthodes géostatistiques pour obtenir une cartographie des polluants atmosphériques sur la région d'Île-de-France.

2.1 Introduction

L'interpolation spatiale est un outil mathématique qui peut être utilisé lors de l'étude d'un phénomène naturel qui se déploie dans l'espace. La région de l'espace géographique concernée par cette étude (le domaine) sera notée par D. Le phénomène naturel examiné, nommé par Matheron (1962) **phénomène régionalisé**, est représenté par un certain nombre de mesures effectuées sur ce domaine.

"Le phénomène régionalisé n'est jamais accessible exhaustivement ; le plus souvent il passe par le filtre d'un échantillonnage. La connaissance, même exhaustive, du phénomène régionalisé ne suffit pas à résoudre les problèmes, car les valeurs numériques ne sont pas le réel, mais une première image (analytiquement très riche, structuralement très pauvre) de celui-ci" (Matheron, 1965).

On peut supposer que le phénomène régionalisé est bien caractérisé par une fonction numérique z définie sur le domaine borné D, appelée **Variable Régionalisée** (VR). En général, une variable régionalisée varie très irrégulièrement et ne peut pas être représentée par une fonction mathématique explicite ; néanmoins, elle reste 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 la Variable Régionalisée. On peut passer à un deuxième niveau d'abstraction si on considère la Variable Régionalisée comme la réalisation d'une **Fonction Aléatoire** $Z(s)$ (FA), en se plaçant ainsi dans un cadre stochastique. Contrairement aux méthodes déterministes, les méthodes **stochastiques** incorporent le concept de

hasard et, grâce à cette modélisation, les erreurs d'estimation peuvent être calculées (voir figure 2.1).

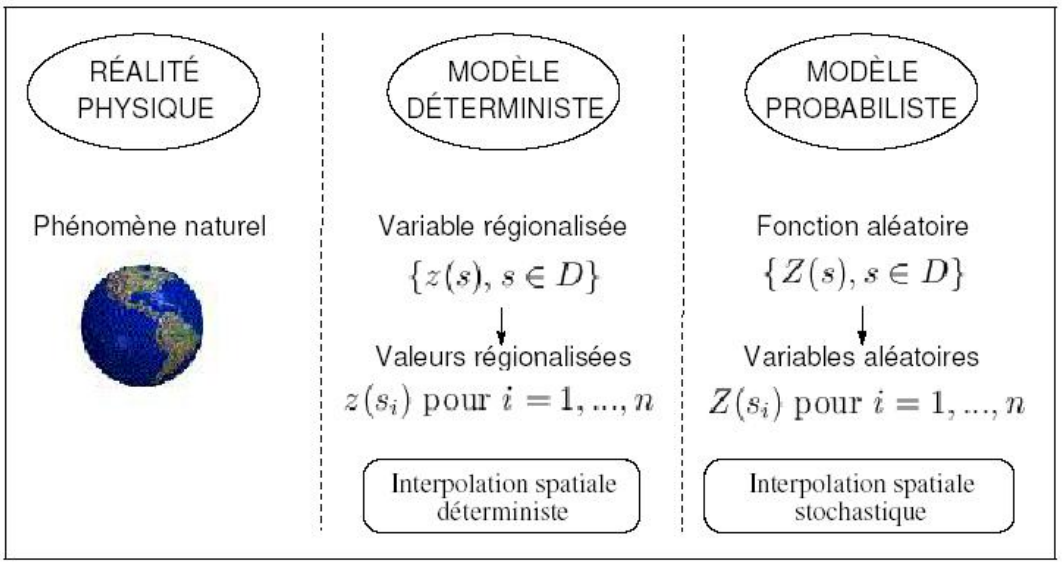


FIG. 2.1: Les deux niveaux d'abstraction nécessaires à la modélisation spatiale d'après Chauvet (1999).

L'interpolation spatiale permet, à l'aide de fonctions mathématiques, d'estimer une grandeur en tous les points ou nœuds d'une grille régulière superposée au domaine d'étude, en se basant sur des mesures ponctuelles et dispersées de la même grandeur.

2.1.0.3 Algorithmes. Classification

Les algorithmes d'interpolation spatiale couramment utilisés en matière de cartographie des champs de pollution concernent pour l'essentiel des méthodes de type **analytique** (ou déterministe) comme, par exemple, les méthodes de triangulation, les fonctions splines ou les méthodes de pondération par rapport aux inverses des distances. Chacun de ces algorithmes utilise sa propre technique de pondération entre les variables à cartographier, et de ce fait les résultats peuvent être différents. Ces techniques ont l'avantage de la simplicité de la mise en œuvre, mais leur inconvénient majeur provient du fait qu'elles ne prennent pas en compte les caractéristiques spatiales naturelles du phénomène étudié (la dispersion des polluants). Seul l'algorithme d'**interpolation géostatistique** du krigeage intègre une règle de pondération directement déduite de la variabilité spatiale des mesures de la concentration du polluant. On classera donc les méthodes d'interpolation spatiale dans deux groupes distincts : analytiques et stochastiques ; une brève description sera donnée en ce qui suit.

2.2 Méthodes analytiques

Cinq grandes classes de méthodes analytiques ont été proposées dans la littérature : les méthodes barycentriques (pondération par rapport aux distances), les méthodes d'ajustement par

morceaux, les fonctions de base radiales, la régression polynomiale et les fonctions splines.

2.2.0.4 Méthodes barycentriques : pondération par rapport aux distances

Les méthodes barycentriques prédisent la valeur d'une variable régionalisée en un point non échantillonné s_0, $\widehat{z}(s_0)$, par une moyenne pondérée des valeurs observées aux points s_i, $z(s_i)$, en affectant aux valeurs qui sont mesurées dans un endroit proche, des poids ω_i plus élevés qu'à celles mesurées à des endroits plus éloignés :

$$\widehat{z}(s_0) = \sum_{i=1}^{n} \omega_i z(s_i) \text{ avec } \sum_{i=1}^{n} \omega_i = 1. \tag{2.1}$$

Les poids ω_i, affectés aux différentes mesures, sont contraints à la somme de 1, afin d'éviter que l'estimation soit "biaisée". Ces poids sont fonction de la **distance euclidienne** $r_i = |s_i - s_0|$ 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. Parfois, seules les observations situées dans un certain voisinage de s_0, noté $V(s_0)$, sont prises en compte. Pour choisir le voisinage, il existe plusieurs méthodes : soit on considère les sites localisés à l'intérieur d'un cercle centré en s_0 et de rayon prédéterminé R, soit on prend les m_0 sites d'observation les plus proches de s_0, soit on peut combiner les deux méthodes décrites précédemment et prendre au plus m_1 sites voisins de s_0, mais situés à une distance inférieure à R.

Parmi les méthodes barycentriques, la plus connue est la méthode de pondération par rapport aux inverses des distances, où les poids ω_i sont estimés par une formule du type :

$$\omega_i(r_i) = \frac{1/r_i^d}{\sum_{i \in V(s_0)} 1/r_i^d}, \quad d > 0.$$

En pratique, on prend souvent $d = 1$ ou 2 ; toute autre fonction de distance peut être utilisée, pourvu qu'elle soit décroissante. L'exemple précédent fait partie de la classe des fonctions de distance qui interpolent sans lisser.

Une propriété souhaitable pour la formulation des poids est que la fonction des distances représentée soit différentiable aux points d'observation. Cette propriété est satisfaite, par exemple, par les modèles exponentiels $exp(-\beta r^2)$, avec $\beta > 0$.

Un exemple de fonction de distance qui interpole avec lissage est donné par le modèle de Cressman (1959), utilisé pour la première fois en météorologie :

$$\omega_i(r_i) = \begin{cases} \left(\frac{s^2 - r_i^2}{s^2 + r_i^2}\right)^{\alpha} & \text{pour } r_i < s \\ 0 & \text{pour } r_i \geq s \end{cases}$$

où le paramètre rayon s détermine le degré de lissage du champ estimé.

La méthode de pondération par rapport aux distances offre un outil simple et flexible, mais ses performances sont un peu limitées.

Une première propriété de cette méthode est que les valeurs interpolées sont toujours comprises

entre la valeur minimale et celle maximale mesurées, ce qui permet d'éviter les valeurs aberrantes, mais cela peut devenir un inconvénient dans certains cas.

Une deuxième propriété est la forme "d'œil-de-bœuf" des isolignes autour des sites d'observation, c'est-à-dire des cercles concentriques autour de chaque site, ce qui n'est pas forcément réaliste. De plus, étant donné que seule la distance par rapport au point à estimer compte, cette méthode a tendance à surpondérer les groupements de données, alors que celles-ci sont en partie redondantes (Arnaud et Emery, 2000).

2.2.0.5 Méthodes d'ajustement par morceaux

Alors que la plupart des méthodes d'interpolation ont comme but d'estimer les valeurs aux nœuds d'une grille régulière, parfois, on a besoin d'estimer la valeur moyenne des petites aires à l'intérieur d'un grand domaine. Une alternative est de partitioner le domaine ; il existe deux principaux types de partitionnement d'un champ D en régions disjointes à partir des sites d'observation : par polygones ou par triangles.

La première alternative est appelée méthode des **polygones de Thiessen** (Cressie, 1993) et consiste à définir, pour chaque site d'observation, un polygone d'influence, déterminé de manière à ce que chaque point du polygone soit plus proche d'un certain site que de tout autre site. Dans $\mathbb{R}^2$, les polygones sont obtenus en traçant les médiatrices des segments joignant s_0 aux autres sites ; ensuite on prend le polygone le plus petit qui contient s_0. Le champ D est alors partitionné en polygones convexes de petite surface quand les données sont groupées, et de plus grande surface quand les données sont isolées.

Le partitionnement par triangles, nommé **triangulation**, découpe le champ en triangles disjoints dont les sommets sont les sites d'observation. La méthode la plus connue est celle de Delaunay, qui conduit à une triangulation optimale (évitant les angles aigus). Elle est basée sur un partitionnement par polygones de Thiessen. Le critère appliqué est le suivant : les sommets de chaque triangle sont les sites de D tels que les polygones de Thiessen ainsi définis aient un côté commun.

Une fois le partitionnement effectué, on passe à l'interpolation. La première remarque concernant l'interpolation est qu'il existe plusieurs façons de l'effectuer. La plus simple est celle du plus proche voisin : la valeur mesurée en un site est attribuée à tous les points situés à l'intérieur du polygone de ce site. Cette méthode présente un inconvénient majeur : la discontinuité quand on passe d'un polygone à l'autre. Pour éviter cette discontinuité il existe comme alternative la méthode d'interpolation par voisinage naturel due à Sibson (1981). D'abord, les polygones de Thiessen associés aux sites d'observation sont tracés, puis ces polygones sont retracés en ajoutant le site s_0 pour lequel on désire faire une estimation. Ensuite on superpose le dernier partitionnement au premier (tracé sans le site s_0) et on construit le poids pour chaque observation s_i en considérant l'aire 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.

Pour la triangulation, on a aussi plusieurs possibilités d'effectuer l'interpolation : on peut faire appel à une interpolation linéaire ou à une interpolation plus lisse en utilisant l'algorithme d'Akima

(1978). Si on utilise l'interpolation linéaire, pour faire une estimation au point s_0 situé à l'intérieur d'un triangle, chaque site d'observation reçoit un poids égal à la proportion de surface occupée par le triangle opposé à ce site ; ainsi lorsque s_0 se rapproche d'un sommet, la pondération de ce sommet devient prépondérante et l'estimation se rapproche de la valeur observée. La surface obtenue est continue mais non différentiable et, comme on ne peut pas extrapoler au-delà de l'enveloppe convexe des sites d'observation, on ne peut pas faire l'estimation au bord du domaine. La méthode d'Akima consiste à ajuster à l'intérieur de chaque triangle de Delaunay une surface dont l'équation est un polynôme du cinquième degré, qui a un aspect très lisse.

L'expérience a démontré que les méthodes basées sur la triangulation donnent de bons résultats à partir d'un nombre assez élevé de points initiaux. L'avantage majeur de cette méthode est sa capacité d'exhiber les discontinuités, alors que la plupart des autres méthodes ont tendance à les lisser.

2.2.0.6 Les surfaces de tendance (trend surface)

Cette méthode consiste à trouver une surface polynomiale en x et y, coordonnées géographiques d'un site s (en considérant $s \in \mathbb{R}^2$), surface qui passe par les points initiaux et qui sera ajustée en utilisant la méthode des moindres carrés.

Mathématiquement on cherche un polynôme de la forme :

$$\widehat{z}(s) = \widehat{z}(x, y) = \sum_{j+k \leq p} \alpha_{jk} x^j y^k. \tag{2.3}$$

Les coefficients α_{jk} s'obtiennent en minimisant la somme : $\Sigma_{i=1}^{N}[\widehat{z}(s_i) - z(s_i)]^2$. Le degré p du polynôme est l'ordre de la surface ; en général, il est inférieur ou égal à 3. En général, on obtient une solution unique. Bien que dans certains logiciels on trouve des tests statistiques concernant la significativité de l'ordre de la surface, ils reposent sur des hypothèses d'indépendance des valeurs rarement vérifiées sur des données spatialisées (Arnaud et Emery, 2000).

2.2.0.7 Splines

L'idée de l'interpolation par fonctions splines est d'ajuster une surface d'énergie de flexion minimale sur le domaine D (Duchon, 1976; Wahba, 1990). Ceci revient à minimiser une intégrale d'espace, sous contrainte de passer par les points initiaux (interpolation exacte) ou dans leur voisinage (lissage) à proximité des points de données. (Il existe donc deux catégories de splines : les splines d'interpolation contraintes à passer par les points d'observation et les splines de lissage qui passent seulement à proximité de ces points.)

La méthode des **splines d'interpolation** consiste à choisir une fonction qui soit la plus lisse possible, tout en restituant les valeurs mesurées aux sites $s_i = (x_i, y_i)$, $i = 1, ..., n$. On cherche ainsi la fonction $\widehat{z}(s)$ qui représente la surface d'une plaque mince et flexible que l'on astreint à passer

par les points initiaux et qui minimise l'énergie de flexion :

$$\int\int\{(\frac{\partial^2\widehat{z}}{\partial x^2})^2+2(\frac{\partial^2\widehat{z}}{\partial x\partial y})^2+(\frac{\partial^2\widehat{z}}{\partial y^2})^2\}dxdy$$

sous la contrainte : $\widehat{z}(s_i)=z(s_i), \forall i=1...n$.

La solution s'écrit sous la forme :

$$\widehat{z}(s)=\widehat{z}(x,y)=a_0+a_1x+a_2y+\sum_{i=1}^{n}b_iK(s-s_i), \tag{2.5}$$

où $K(h)=|h|^2\ln(|h|), \forall h\in\mathbb{R}^2$, avec $h=s-s_i$.

Les **splines de lissage** s'obtiennent par généralisation dans $\mathbb{R}^2$ d'une fonction de la même forme que 2.5 qui minimise, cette fois-ci, l'expression :

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

où ρ est un paramètre de lissage fixé *a priori*, qui mesure le compromis entre la contrainte d'ajustement des données et la contrainte de lissage de la fonction spline.

2.2.0.8 Les fonctions de base radiales

Une autre approche analytique est celle basée sur une interpolation spatiale donnée sous la forme d'une somme pondérée de n fonctions de base radiales (en anglais Radial Basis Functions-RBF). Une formule générale pour ce type d'interpolation est :

$$\widehat{z}(s)=\sum_{i=1}^{n}\lambda_i\mathbf{\Phi}(\|s-s_i\|)+p(s),\ \ s\in\mathbb{R}^2, \tag{2.7}$$

où $\|.\|$ représente la norme euclidienne, $\mathbf{\Phi}$ sont les fonctions de base radiales, λ_i les poids et $p(s)$ est un polynôme d'ordre m.

Les types de RBF les plus fréquemment utilisés sont :

- $\mathbf{\Phi}(r)=e^{-c^2r^2}$ (exponentielle) ;
- $\mathbf{\Phi}(r)=\sqrt{c^2+r^2}$ (multiquadrique) ;
- $\mathbf{\Phi}(r)=\frac{1}{\sqrt{c^2+r^2}}$ (multiquadrique inverse) ;
- $\mathbf{\Phi}(r)=(r^2+c^2)^{\frac{3}{2}}$ (spline cubique) ;
- $\mathbf{\Phi}(r)=r^2\log r$ ou $\mathbf{\Phi}(r)=(r^2+c^2)\log(r^2+c^2)$ (spline plaque mince),

r étant la distance relative entre le centre (site d'observation) et le point de calcul ($\|s-s_i\|$) et c une constante positive qui joue le rôle d'un paramètre de lissage. Plus cette constante est grande, plus la surface est lisse.

Les fonctions RBF jouent le rôle du variogramme dans le cas de krigeage, c'est-à-dire qu'elles définissent l'ensemble optimal de poids nécessaire dans l'interpolation. L'expérience a démontré que l'interpolation basée sur le modèle multiquadrique ou spline donne presque toujours de bons résultats.

2.3 Méthodes stochastiques - Géostatistique

Le premier niveau d'abstraction, dans un but de modélisation spatiale, est d'utiliser des Variables Régionalisées, comme on l'a vu pour les méthodes analytiques. Avec le deuxième niveau d'abstraction effectué (le passage à la fonction aléatoire, voir figure 2.1), d'un point de vue mathématique, l'objet d'étude devient la fonction aléatoire (ou processus aléatoire) $Z(s)$, qui est définie à la fois sur l'espace "géographique" et sur un espace probabilisé $Z(s, \omega)$, où $s \in \mathbb{R}^n$ (dans notre cas $n = 2$) et ω est un événement de l'espace probabilisé. L'usage dans la littérature géostatistique est de négliger l'écriture de l'événement ω et d'écrire simplement $Z(s)$. Par rapport à l'approche de la statistique classique, la différence fondamentale est que les diverses variables aléatoires ne sont pas indépendantes. Ce sont les corrélations qui existent entre les variables aléatoires définies en plusieurs points de l'espace qui permettront de décrire la "structure" spatiale du phénomène régionalisé.

Les données fournies par les stations de mesure $z(s_i)$ sont des informations fragmentaires que l'on possède sur une réalisation particulière $z(s)$ de $Z(s)$. On peut décrire la fonction aléatoire par sa loi de probabilité entière, ou plus synthétiquement par ses moments du premier et du second ordre. En fait il existe deux raisons pour la simplification proposée : d'une part, la commodité des manipulations mathématiques, et d'autre part, la possibilité de réaliser l'inférence statistique (on a nécessairement un nombre limité de données et on ne peut pas obtenir la loi spatiale entière qui contient une infinité de paramètres).

2.3.0.9 Moment du premier ordre

Le moment du premier ordre d'une fonction aléatoire correspond à son *espérance* mathématique, qui, en toute généralité, dépend du point s :

$$E[Z(s)] = m(s).$$

Selon Saporta (1990), l'espérance en un point s donné représente la "moyenne" autour de laquelle les valeurs possibles de $Z(s)$ se distribuent. (Il s'agit d'une moyenne calculée sur les différentes réalisations de la fonction aléatoire, pas d'une moyenne dans l'espace.)

2.3.0.10 Moment du second ordre

Les moments du second ordre fournissent une description élémentaire de la loi bivariable de $Z(s)$, c'est-à-dire la loi de probabilité entre les valeurs prises en deux sites s_1 et s_2.

Dans ce document on utilisera les moments d'ordre deux suivants :

- la covariance entre $Z(s_1)$ et $Z(s_2)$ qui quantifie le degré de ressemblance entre les valeurs prises en s_1 et s_2 :

$$Cov[Z(s_1), Z(s_2)] = E[Z(s_1)Z(s_2)] - m(s_1)m(s_2) = C(s_1, s_2);$$

- la covariance entre la variable aléatoire $Z(s_1)$ et elle-même (variance ou variance *a priori*) :

$$Cov[Z(s_1), Z(s_1)] = var[Z(s_1)] = E[Z(s_1) - m(s_1)]^2;$$

- le semi-variogramme entre $Z(s_1)$ et $Z(s_2)$ qui mesure la dissemblance entre les valeurs prises en s_1 et s_2 :

$$\gamma(s_1, s_2) = \frac{1}{2} var[Z(s_1) - Z(s_2)].$$

L'inférence statistique concerne la restitution des caractéristiques de la fonction aléatoire à partir d'un ensemble de données expérimentales. Le problème d'inférence soulève une difficulté de taille : la variable régionalisée ne constitue qu'une seule réalisation de la fonction aléatoire ; par conséquent, la fonction aléatoire ne peut pas être définie sans ambiguïté à partir de la variable régionalisée. De façon plus générale, on peut se demander comment un événement unique (ici, la réalisation d'une fonction aléatoire) peut faire l'objet d'une approche scientifique, la condition de "répétabilité" qui, seule, fonde l'objectivité de nos disciplines scientifiques faisant ici défaut par définition (Matheron, 1965).

Pour permettre l'inférence statistique, des hypothèses limitatives, traduisant une certaine homogénéité du phénomène dans l'espace ("stationnarité") seront nécessaires (détails dans la section 2.5.1), ainsi qu'une autre hypothèse appelée "ergodicité". Cette dernière est une propriété du modèle probabiliste. Il n'en existe pas d'équivalent empirique au niveau de la Variable Régionalisée. En fait nous sommes obligés de recourir à un modèle probabiliste ergodique : nous ne disposons que d'une seule réalisation du processus : comment savoir ce qu'aurait été la limite des moyennes spatiales sur une autre réalisation ? Nous sommes contraints de supposer que cette limite est l'espérance mathématique, c'est-à-dire qu'elle est la même pour toutes les réalisations. Ceci revient à choisir comme définition de l'espérance mathématique la limite de la suite des moyennes spatiales (Matheron, 1965). Nous remarquerons que la stationnarité n'entraîne pas l'ergodicité.

2.4 Le krigeage

Le krigeage est la première méthode d'interpolation spatiale qui tient compte de la structure spatiale des données. La première mention de cette méthode est due à l'ingénieur minier sud-africain Krige (1951), d'où vient le nom du *kriging* en anglais et *krigeage* en français. Le formalisme mathématique est dû au français Matheron (1962, 1963b), qui a aussi assuré son développement au Centre de Géostatistique de l'École des Mines de Paris.

L'idée de base du krigeage est de prévoir la valeur de la Variable Régionalisée étudiée (par exemple la concentration d'un polluant atmosphérique) en un site non échantillonné s_0 par une combinaison linéaire de données ponctuelles adjacentes :

$$\widehat{z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i z(s_i). \tag{2.12}$$

Les poids λ_i associés à chacune des valeurs régionalisées observées sont choisis de façon à obtenir une estimation non biaisée et une variance minimale. Ils dépendent de la localisation des

observations et de leur structure de dépendance spatiale. Le krigeage fait ainsi partie de la classe de meilleures estimations linéaires sans biais (BLUE en anglais - Best Linear Unbiased Estimator).

La technique de krigeage passe d'abord par le choix du voisinage : le domaine du champ qui contient le site à estimer et les données utilisées dans l'estimation. Dans certains cas il n'est pas raisonnable de conserver toutes les données, soit parce qu'elles sont trop nombreuses, soit parce qu'elles sont trop éloignées. Cela correspond à un voisinage glissant, par opposition à celle de voisinage unique où toutes les données sont prises en compte. On reviendra sur cette question dans le cadre de l'analyse variographique, car le choix du voisinage a une forte influence sur les résultats.

Il faut préciser que la quantité autour de laquelle gravite tout le formalisme du krigeage n'est pas l'estimateur lui-même, mais *l'erreur d'estimation*, c'est-à-dire l'écart entre la quantité à estimer et l'estimateur. Au niveau de la Variable Régionalisée, cet écart n'est autre que l'erreur commise faute de connaître la vraie valeur de la quantité à estimer. Transposée au niveau de la Fonction Aléatoire, cette erreur d'estimation devient, dans le modèle probabiliste, une variable aléatoire.

Pour arriver au formalisme du krigeage on doit d'abord présenter le **modèle de base** qui a la même forme que le modèle de régression classique ou locale, mais les erreurs sont supposées dépendantes spatialement. On peut décomposer la Fonction Aléatoire $Z(s)$:

$$Z(s) = \mu(s) + \delta(s), \quad s \in D, \tag{2.13}$$

où $\mu(.)$ est *la structure déterministe* pour l'espérance de $Z(.)$ et $\delta(.)$ une fonction aléatoire stationnaire, d'espérance nulle et de structure de dépendance connue, appelée aussi *fonction résiduelle* ou *résidu*.

On peut encore détailler cette décomposition en suivant Cressie (1993) de façon suivante :

$$Z(s) = \mu(s) + W(s) + \eta(s) + \epsilon(s), \quad s \in D, \tag{2.14}$$

où

- $\mu(.) = E(Z(.))$ représente la variation à grande échelle, structure déterministe pour l'espérance de $Z(.)$;
- $W(.)$ représente une variation lisse à petite échelle (structure stochastique de fluctuations autour de $\mu(.)$ dépendantes spatialement), continue, dans le sens que $E(W(s+h)-W(s))^2 \to 0$ quand $h \to 0$, intrinsèque (voir la définition dans la section 2.5.1) ; si elle existe, la portée est supérieure à la distance minimale entre deux sites d'échantillonnage ;
- $\eta(.)$ représente une variation micro-échelle (portée plus petite que la distance minimale entre deux sites d'échantillonnage), structure stochastique présentant une dépendance spatiale ;
- $\epsilon(.)$ représente une erreur de mesure, structure stochastique sans dépendance spatiale (bruit blanc).

La fonction aléatoire $\delta(.)$ du modèle 2.13 est formée par le regroupement des termes $W(.)$, $\eta(.)$, $\epsilon(.)$ du modèle précédent 2.14.

Pour formuler complètement le modèle, il faut spécifier la **forme de la tendance** $\mu(.)$. Finalement c'est cette tendance qui précise le type de krigeage effectué. Les trois types classiques de krigeage (Gratton, 2002) sont :

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

La structure de dépendance du résidu $\delta(.)$ doit elle aussi être précisée. Si elle n'est pas connue préalablement, ce qui est souvent le cas dans la pratique, elle est déterminée à partir des données lors de l'analyse variographique (cf. section 2.5).

Une fois le modèle complètement énoncé, le krigeage peut être effectué en un point s_0 quelconque du domaine D. Il s'agit maintenant de calculer les poids λ_i de la combinaison linéaire 2.12 qui respectent les contraintes déjà énoncées : non-biais, soit

$$E[\widehat{Z}(s_0) - Z(s_0] = 0, \tag{2.15}$$

tout en minimisant la variance de l'erreur de prévision

$$Var[\widehat{Z}(s_0) - Z(s_0]. \tag{2.16}$$

La terminologie est variée, mais la plus importante remarque qu'on peut faire ici est qu'"il n'existe qu'un seul krigeage et il n'y a qu'une seule démarche pour aborder un problème d'estimation au sens local" (Chauvet, 1999). Cette démarche doit s'adapter aux circonstances particulières (différentes hypothèses de stationnarité par exemple) et, par la suite, les formulations ultimes peuvent présenter des différences notables.

Ce qui distingue vraiment le krigeage des autres méthodes d'interpolation spatiale est qu'il est le seul à tenir compte de la structure de dépendance spatiale des données. De plus, puisqu'il s'agit d'une méthode stochastique, le krigeage permet d'estimer les erreurs de prévision. Étant un interpolateur exact, il restitue les valeurs régionalisées mesurées aux sites d'observation. Cependant, il est possible d'effectuer aussi un krigeage dit avec erreurs de mesure. Par ailleurs, il est possible d'ajouter des variables régionalisées auxiliaires à la tendance générale $\mu(.)$ du modèle ; le krigeage incorporant une telle tendance est nommé **krigeage avec dérive externe** (Goovaerts, 2000; Wackernagel, 2003).

Le krigeage possède également d'autres extensions multivariables, notamment le **cokrigeage**. Par cette méthode, $\widehat{z}(s_0)$ 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 $j = 1, ..., q$:

$$\widehat{z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_{i,0} z(s_i) + \sum_{j=1}^{q} \sum_{i \in V(s_0)} \lambda_{i,j} w_j(s_{i,j}).$$

Les poids $\lambda_{i,j}$ sont calculés en minimisant la variance de l'erreur de prévision sous la contrainte de non-biais, tout en tenant compte de la structure spatiale de la variable régionalisée et de celles des variables auxiliaires.

2.5 Analyse variographique

Le modèle (2.13) du krigeage suppose la connaissance de la structure de dépendance spatiale de la fonction aléatoire $\delta(.)$, le résidu. Cependant, en pratique, celle-ci est rarement connue. L'analyse variographique est une étape préalable au krigeage qui permet d'estimer cette structure. L'analyse spectrale peut aussi apporter beaucoup de renseignements sur cette structure spatiale. Rappelons que pour un processus stationnaire, la densité spectrale est la transformée de Fourier de la covariance. Comme cette dernière n'est pas vraiment connue, il reste la possibilité de l'estimer à partir des données, en intégrant éventuellement les informations partielles dont on dispose.

Dans cette section on présente d'abord le concept de stationnarité, ensuite on définit une fonction représentant la dépendance spatiale, le semi-variogramme, l'outil nécessaire pour estimer la structure spatiale de la fonction résiduelle, et en fin de section on présente la modélisation du variogramme à partir des valeurs expérimentales.

2.5.1 Hypothèse de stationnarité

Le processus naturel étudié étant unique, une seule réalisation de la fonction aléatoire $Z(.)$ est observable. Afin de rendre possible l'inférence statistique, une hypothèse de stationnarité est émise concernant la fonction aléatoire résiduelle $\delta(.)$. Au sens strict, la stationnarité signifie que la loi de probabilité de la fonction aléatoire est invariante par translation. En krigeage on recourt à une simplification : la stationnarité postulée est faible. Elle concerne uniquement les moments d'ordre 1 et 2 de la fonction aléatoire ou de ses accroissements, plutôt que sa distribution entière. On parle ainsi de stationnarité d'ordre 2 et de stationnarité intrinsèque. Les deux types de stationnarité sont définis dans la suite.

2.5.1.1 Stationnarité d'ordre 2

- $E[\delta(s)] = m = 0 \quad \forall\ s \in D$:
  L'espérance du résidu $\delta(.)$ existe et reste constante pour tous les sites d'observation.
- $Cov[\delta(s), \delta(s+h)] = C(h) \quad \forall\ s, s+h \in D$:
  La covariance de $\delta(.)$ entre toutes les paires de sites s et $s+h$ existe et dépend uniquement de h, le vecteur de translation entre les deux sites. Cette fonction de covariance est appelée **covariogramme**.

2.5.1.2 Stationnarité intrinsèque

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

Dans les deux cas de stationnarité présentés, l'espérance est une constante et elle ne permet donc pas d'apprécier la structuration des données ; par conséquent, sa connaissance n'est pas indispensable pour résoudre les problèmes d'estimation. En revanche, les moments d'ordre 2 (covariance

et variogramme) quantifient le degré de ressemblance ou de dissemblance entre les valeurs prises en deux sites en fonction de leur séparation, h, donc ils reflètent la structure de "régionalisation". "Leur inférence à partir des données expérimentales et leur modélisation est une étape cruciale dans une étude géostatistique et c'est elle qui permettra postérieurement d'estimer les valeurs inconnues de la variable régionalisée et d'assortir les estimations d'une mesure de leur précision" (Arnaud et Emery, 2000).

Dans le cadre stationnaire d'ordre 2, le covariogramme et le variogramme sont reliés par la relation :

$$2\gamma(h) = 2(C(0) - C(h)). \tag{2.18}$$

La relation 2.18 montre que l'existence du covariogramme implique l'existence du variogramme. Par contre, l'implication inverse n'est vraie que si $\gamma(.)$ est borné. Ainsi, il existe des fonctions aléatoires qui sont stationnaires intrinsèques, mais elles ne sont pas stationnaires d'ordre 2 (par exemple, le mouvement brownien fractionnaire). Il est possible que certains phénomènes régionalisés présentent une dispersion infinie. Pour cette raison, l'hypothèse de stationnarité intrinsèque sera privilégiée et la dépendance spatiale sera représentée par un variogramme plutôt que par un covariogramme.

Une observation très importante en ce qui concerne les hypothèses de stationnarité est que les caractéristiques d'un phénomène régionalisé peuvent changer radicalement selon l'échelle à laquelle on le regarde, et un modèle satisfaisant à une certaine échelle peut ne plus l'être à une échelle différente (Arnaud et Emery, 2000). Ainsi il n'est pas nécessaire que la fonction aléatoire qui modélise le phénomène soit stationnaire ou intrinsèque sur tout le champ. On peut restreindre les hypothèses stationnaire et intrinsèque à une échelle locale. Il est donc possible de trouver des zones localement stationnaires, mais il peut arriver qu'elles aient peu de données, rendant impossible l'inférence statistique. On a deux facteurs à concilier : la taille de la zone considérée "homogène" et le nombre de données disponibles.

2.5.2 Propriétés du semi-variogramme

La plupart des géostatisticiens travaillent avec la fonction $\gamma(.)$, la moitié du variogramme, appelée **semi-variogramme**. Cette fonction est de type négatif conditionnel (Matheron, 1965; Christakos, 1984), c'est-à-dire que :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(s_i - s_j) \leq 0 \tag{2.19}$$

pour n'importe quel ensemble fini de points $\{s_i : \ i = 1, ..., n\}$ et n'importe quels nombres réels $\{a_i : \ i = 1, ..., n\}$ tels que $\sum_{i=1}^{n} a_i = 0$. L'adjectif conditionnel se réfère au fait que l'inégalité n'est valable que sous la condition que la somme des poids a_i est nulle. Cette propriété assure la positivité de la variance de toute combinaison linéaire de variables aléatoires issues de $\delta(.)$ et représente l'extension aux semi-variogrammes du caractère semi-défini positif des fonctions de covariance (Christakos, 1984). Dans le cadre de l'hypothèse stationnaire d'ordre 2 toutes les combinaisons linéaires sont autorisées, car l'espérance et la covariance existent. Par contre, l'hypothèse intrinsèque ne garantit que l'existence des moments pour les accroissements de la fonction aléatoire, et par

conséquent, seules les combinaisons linéaires de poids total nul sont autorisées.

2.5.2.1 Isotropie

Une autre propriété importante du semi-variogramme concerne son isotropie. Le semi-variogramme ne dépend que de h, vecteur de translation entre les points s et $s+h$. Si cette dépendance ne concerne que la norme de ce vecteur, on dit que le semi-variogramme est isotrope. S'il dépend aussi de la direction de h, le variogramme est considéré alors comme anisotrope. Il existe deux types d'anisotropie : une géométrique et l'autre zonale. Pour détecter les anisotropies, on peut visualiser les valeurs du variogramme expérimental lorsque le vecteur de séparation h se déplace dans l'espace. Si on travaille dans un espace bidimensionnel on obtient une surface. Si les lignes d'isovaleurs de la carte variographique dessinent des cercles concentriques on dit que le variogramme est isotrope. Si on obtient des ellipses, il s'agit d'une anisotropie géométrique et si on obtient de bandes, d'une anisotropie zonale. Autrement dit (voir plus loin les définitions de la portée et du palier), si la portée change avec la direction et le palier reste constant, on a une anisotropie géométrique. En revanche, si le palier varie avec la direction et la portée reste la même, on a une anisotropie zonale.

2.5.2.2 Effet de pépite

Le comportement à l'origine du variogramme ou de la covariance reflète le degré de régularité spatiale de la variable régionalisée sous-jacente. On peut classer le comportement à l'origine en trois catégories, de la plus régulière à la plus erratique.

- Parabolique : comportement qui indique une très grande régularité spatiale du phénomène ; par conséquent la variable régionalisée est différentiable ou lisse.
- Linéaire : comportement moins régulier ; la variable est continue, mais non différentiable.
- Discontinu ("effet de pépite") : le variogramme présente un saut abrupt à l'origine, ce qui indique une faible ressemblance entre les valeurs régionalisées voisines (valeurs mesurées en deux sites très proches).

Alors que $\gamma(0) = 0$ toujours, l'effet de pépite est obtenu au voisinage de l'origine si :

$$\gamma(h) \to c_0 > 0 \text{ quand } h \to 0. \tag{2.20}$$

Le terme *pépite* est appliqué pour toutes les discontinuités à l'origine, même si leur cause est différente. Généralement, la discontinuité est due à n'importe quel terme parmi les trois derniers mentionnés dans la décomposition 2.14. Mathématiquement, la présence d'une telle discontinuité n'est pas possible pour les processus continus, donc il ne s'agit pas du processus $W(.)$. En se ramenant au modèle 2.14, c'est donc dire que l'effet de pépite est associé à la fonction aléatoire $\eta(.)$ at aux erreurs de mesure. En absence des points de mesure situés très proches, c'est pratiquement impossible d'identifier quelle est la partie responsable de l'apparition d'un tel *effet pépitique* (Chilès et Delfiner, 1999). Ainsi, si le processus est supposé continu à une micro-échelle, alors la seule raison possible pour que $c_0 > 0$ est l'erreur de mesure. Dans la pratique, on ne sait pas si la variance à micro-échelle est continue ou pas ; selon Matheron, le plus souvent on considère qu'elle ne l'est pas,

et elle est modélisable comme du bruit blanc.

Par conséquent, l'effet de pépite, c_0 peut être décomposé :

$$c_0 = c_{MS} + c_{ME}, \tag{2.21}$$

où c_{MS} représente la variance du bruit blanc traduisant la variation à micro-échelle, et c_{ME}, l'erreur de mesure.

Le plus souvent, les équations du krigeage supposent l'absence de l'erreur de mesure ($c_{ME} = 0$), faute d'avoir répliqué les mesures. Ignorer la décomposition 2.21 est la source de la controverse "le krigeage est/n'est pas un interpolateur exact" (Cressie, 1993). La conclusion de Cressie (1993) est que le krigeage est un interpolateur exact si on est prêt à assumer que $c_{ME} = 0$, mais dans les cas réels, cette hypothèse n'est pas vérifiée.

L'analyse de la régularité d'un variogramme à l'origine a été formalisée par Stein (1999) qui montre que, dans un cadre asymptotique, il est important d'estimer avec précision la régularité de la covariance. Dans ce sens, il plaide pour l'utilisation du modèle de Matérn qui sera décrit dans la section 2.5.4.

2.5.2.3 Portée et palier

La relation mathématique entre la covariance et le variogramme (2.18) montre que, à l'infini, $C(h)$ tend vers 0, tandis que $\gamma(h)$ tend vers un palier qui est égal à la variance *a priori* $C(0)$. La distance à partir de laquelle ce palier est atteint est appelée portée ; l'atteinte d'un tel plateau indique qu'à partir d'une certaine distance il n'y a plus de dépendance spatiale entre les données. Un palier ne peut 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 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, donc cette fonction n'est pas stationnaire d'ordre 2, mais stationnaire intrinsèque.

2.5.3 Inférence du variogramme et de la covariance

On cherche à estimer le variogramme à partir des données disponibles : les observations $z(x_i)$ pour $i = 1, ..., n$. Si on reprend la définition du variogramme sous l'hypothèse intrinsèque on obtient :

$$\begin{aligned}
\gamma(h) &= \frac{1}{2}Var(\delta(s) - \delta(s+h)) \\
&= \frac{1}{2}Var[(Z(s) - \mu(s)) - (Z(s+h) - \mu(s+h))] \\
&= \frac{1}{2}Var(Z(s) - Z(s+h)) \\
&= \frac{1}{2}E[\{Z(s) - Z(s+h)\}^2] - \frac{1}{2}\{\mu(s) - \mu(s+h)\}^2.
\end{aligned} \tag{2.22}$$

Si $\mu(.)$ est une fonction constante, le deuxième terme s'annule et le semi-variogramme est estimable directement à partir des $z(s_i)$ comme dans le krigeage simple et ordinaire. En revanche, quand

$\mu(.)$ n'est pas constante, comme c'est le cas du krigeage universel, l'estimation doit se baser sur la fonction $\delta(.)$. Le variogramme peut alors être estimé à partir des valeurs de l'écart $z(s_i) - \mu(s_i)$, les résidus. Pour l'instant, on se situe dans le cas classique d'un krigeage simple ou ordinaire, quand le dernier terme de la relation 2.22 disparaît et le variogramme est estimé à partir des données brutes $z(x_i)$.

Dans le cas réel, on n'a pas accès directement au variogramme, mais on peut, à partir des données dispersées, calculer le variogramme expérimental avec la formule :

$$2\widehat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} [z(s_i) - z(s_j)]^2, \tag{2.23}$$

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

Il existe un autre estimateur proposé par Cressie (1993), plus robuste, qui diminuerait les effets d'une valeur "atypique", de formule :

$$2\widehat{\gamma}(h) = \frac{\left[\frac{1}{|N(h)|}\sum_{N(h)}[z(s_i) - z(s_j)]^{1/2}\right]^4}{0.457 + 0.494/|N(h)|} \tag{2.24}$$

avec la même signification que précédemment pour $N(h)$.

Quand les données sont réparties irrégulièrement sur le domaine D, ce qui est toujours le cas en pratique, le variogramme expérimental est erratique. Afin de rendre $\gamma(.)$ plus robuste, des tolérances sont introduites sur la longueur et éventuellement l'angle du vecteur h, c'est-à-dire que l'on retient dans le calcul toutes les paires de données dont la séparation est "approximativement" égale à h. Pour des distances très grandes le semi-variogramme n'est pas fiable vu que la dispersion de l'estimateur $\widehat{\gamma}(.)$ autour de $\gamma(.)$ augmente. La plupart de géostatisticiens ne prennent en compte que les distances inférieures à la moitié du diamètre du domaine D (Arnaud et Emery, 2000; Chilès et Delfiner, 1999).

Pour étudier la stabilité numérique du variogramme expérimental, il est utile de visualiser la "nuée variographique", c'est-à-dire le nuage des carrés des différences $\{[z(s_i + h) - z(s_i)]^2\}$ en fonction du vecteur h ou de sa norme $|h|$.

L'examen de la nuée variographique permet de repérer les couples de données qui déstabilisent le variogramme expérimental ainsi que les mesures atypiques pour vérifier ensuite si elles sont aberrantes (Chilès et Delfiner, 1999).

Pour obtenir un variogramme expérimental on a besoin de spécifier plusieurs paramètres : la direction de calcul, le pas et la tolérance. Si l'un de ces paramètres n'est pas correctement spécifié cela peut conduire à des variogrammes expérimentaux impossibles à interpréter et modéliser.

En ce qui concerne la direction de calcul, en cas d'isotropie, toutes les directions sont équivalentes et on calcule ce qu'on appelle un variogramme "omnidirectionnel" qui ne dépend que de la

distance $|h|$ et pas de l'orientation. En revanche, en cas d'anisotropie géométrique ou zonale on ne calcule le variogramme expérimental que le long de certaines directions significatives qui permettront ensuite de reconstituer le variogramme dans toutes les directions (Journel et Huijbregts, 1978).

Le deuxième paramètre crucial pour le calcul du variogramme est le pas de calcul. En pratique, les données ne sont pas issues d'un échantillonnage régulier, il faut donc trouver un pas adéquat ; un pas trop petit conduira à un variogramme erratique, tandis qu'avec un pas trop grand, le variogramme présentera peu de détails.

En règle générale, la tolérance sur les distances a été introduite afin de rendre le variogramme expérimental plus stable. Elle est établie à la moitié du pas de part et d'autre de la distance considérée. Une tolérance trop grande a pour effet le lissage du variogramme et la surestimation de l'effet pépitique. Pour les tolérances angulaires on admet une tolérance de maximum 90°, lorsqu'on retrouve le cas particulier du variogramme omnidirectionnel. Une tolérance angulaire de 22.5° peut conduire à une certaine stabilité du variogramme directionnel.

2.5.4 Modélisation du variogramme

Une fois le variogramme expérimental calculé, il faut lui ajuster une courbe théorique, qui devra être définie pour toutes les distances et toutes les directions de l'espace. Cependant, on ne peut pas utiliser n'importe quelle fonction. Une condition essentielle est qu'elle soit du type conditionnel négatif. Comme la vérification de cette contrainte est une tâche assez difficile, on doit choisir un modèle de variogramme parmi un ensemble de fonctions couramment utilisées.

Six modèles isotropes ($r = |h|$) parmi les plus connus sont présentés dans la suite. Partout, dans ce qui suit, on impose les conditions : $a \geq 0$, $c_0 \geq 0$, $C \geq 0$.

1. *Modèles avec palier* :
   (a) modèle pépitique de palier C (fig.2.2) ;

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

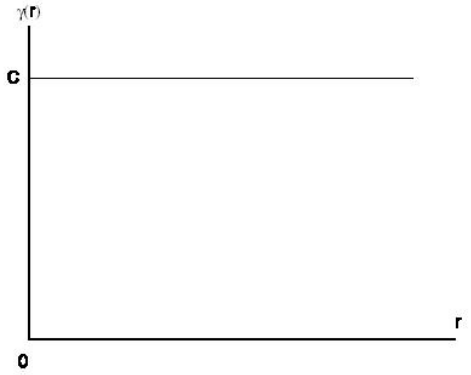


FIG. 2.2: Variogramme pépitique.

   (b) modèle sphérique de portée a et de palier C (fig. 2.3) ;

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 + C(\frac{3}{2}\frac{r}{a} - \frac{1}{2}\frac{r^3}{a^3}) & \text{pour } 0 < r \leq a \\ c_0 + C & \text{pour } r \geq a \end{cases}$$

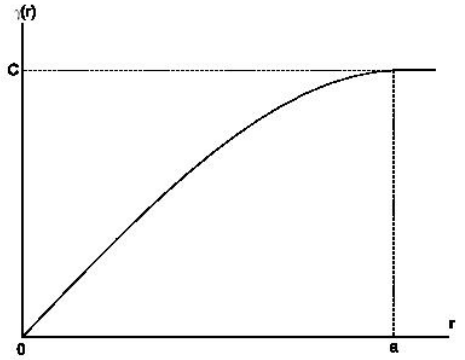


FIG. 2.3: Variogramme sphérique (pour $c_0 = 0$).

(c) modèle exponentiel de paramètre a et de palier C (fig. 2.4) ;

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 + C(1 - exp(-\frac{r}{a})) & \text{pour } r > 0 \end{cases}$$

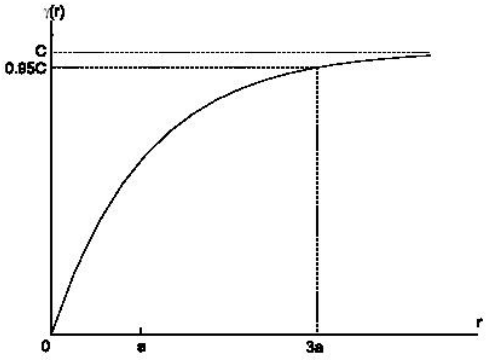


FIG. 2.4: Variogramme exponentiel (pour $c_0 = 0$).

(d) modèle cubique de portée a et de palier C (fig. 2.5) ;

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 + C(7\frac{r^2}{a^2} - \frac{35}{4}\frac{r^3}{a^3} + \frac{7}{2}\frac{r^5}{a^5} - \frac{3}{4}\frac{r^7}{a^7}) & \text{pour } 0 < r \\ c_0 + C & \text{pour } r \geq a \end{cases}$$

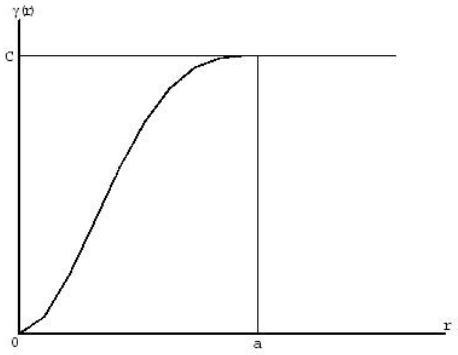


FIG. 2.5: Variogramme cubique (pour $c_0 = 0$).

(e) modèle gaussien de paramètre a et de palier C (fig. 2.6) ;

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 + C(1 - exp(-\frac{r^2}{a^2})) & \text{pour } r > 0 \end{cases}$$

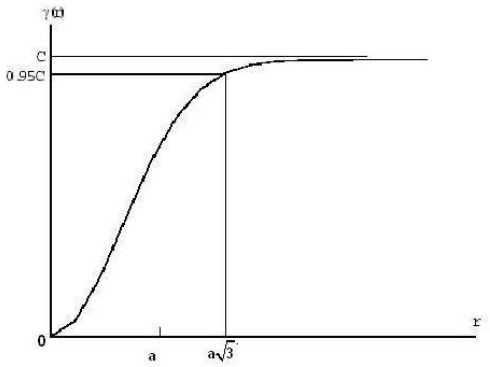


FIG. 2.6: Variogramme gaussien (pour $c_0 = 0$).

2. *Modèles à effet de trou*

On parle d'un effet de trou lorsque le variogramme n'est pas une fonction croissante de la distance. Un modèle adapté à ce cas est le sinus cardinal de paramètre a et de palier C (fig. 2.7).

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 + C(1 - \frac{sin(r/a)}{r/a}) & \text{pour } r > 0 \end{cases}$$

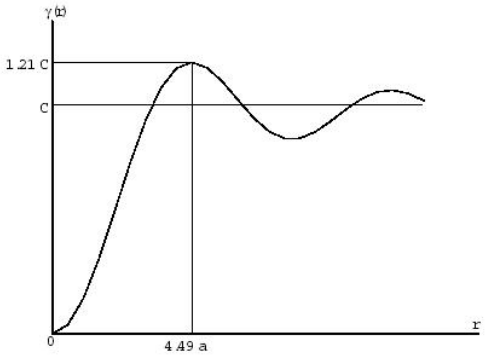


FIG. 2.7: Variogramme sinus cardinal (pour $c_0 = 0$).

3. *Modèle sans palier* :

(a) modèle puissance d'exposant θ et facteur d'échelle ω (fig. 2.8) ;

(b) modèle linéaire de pente ω (fig. 2.9).

On finit la liste des modèles de variogramme disponibles avec celui recommandé par Stein (1999), le modèle de Matérn, qui est un modèle isotrope admissible :

$$\gamma(r) = \begin{cases} c_0 + c_1(1 - \frac{1}{2^{\nu-1}}\Gamma(\nu)(\frac{r}{a})^\nu K_\nu(\frac{r}{a})) & \text{pour } r > 0 \\ 0 & \text{pour } r = 0 \end{cases}$$

où ν représente un paramètre de lissage, K_ν est la fonction Bessel modifiée du second degré et Γ est la fonction gamma. Le paramètre de lissage du modèle accorde plus de flexibilité pour modéliser

$$\gamma(r) = \begin{cases} 0 & \text{pour } r = 0 \\ c_0 + \omega r^\theta & \text{pour } r > 0 \end{cases} \quad \text{avec } 0 < \theta < 2$$

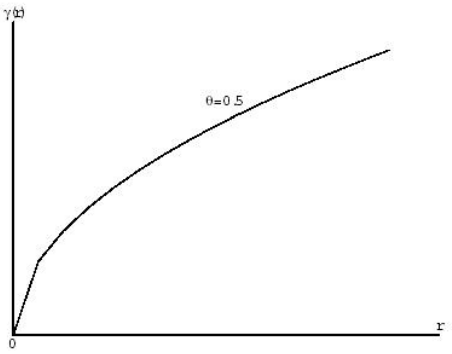


FIG. 2.8: Variogramme puissance (pour $c_0 = 0$).

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

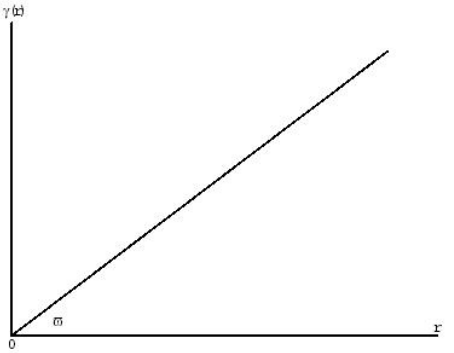


FIG. 2.9: Variogramme linéaire (pour $c_0 = 0$).

la covariance spatiale, en particulier pour r très petit. Pour diverses valeurs particulières de ν, on retrouve d'autres modèles admissibles, déjà présentés, comme par exemple le modèle exponentiel (Marchant et Lark, 2007).

Tous les modèles élémentaires proposés peuvent être combinés de manière à décrire des régionalisations plus complexes. Ainsi toute somme de modèles élémentaires est aussi un modèle admissible. Cette approche d'addition de modèles forme ce qu'on appelle un modèle linéaire de *régionalisation* ou *structure gigogne*, car $\gamma(.)$ apparaît comme la superposition de plusieurs modèles élémentaires hiérarchisés, chacun caractérisant une échelle spatiale particulière.

Après la sélection du modèle, celui-ci doit être ajusté au semi-variogramme expérimental à l'aide d'une méthode d'estimation de paramètres. Parmi ces méthodes on retrouve trois classes distinctes : les méthodes de **maximum de vraisemblance**, les approches **bayesiennes** et surtout les critères quadratiques, basés sur un estimateur des **moindres carrés** généralisés ou pondérés, qui sont d'ailleurs les plus utilisés. L'approche de maximum de vraisemblance consiste à chercher les valeurs des paramètres qui attribuent aux données observées la plus grande vraisemblance (meilleure

solution au sens des statistiques des erreurs), tandis que celle bayesienne maximise la probabilité du modèle étant données les mesures, par l'intermédiaire de la règle de Bayes. Quant à l'estimateur des moindres carrés, il est obtenu analytiquement à partir des mesures. En utilisant la méthode des moindres carrés, le vecteur des paramètres du modèle choisi, noté par θ, est estimé en minimisant :

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

où $\widehat{\gamma}(.)$ est le semi-variogramme empirique, $\gamma(., \theta)$ est le modèle variographique de paramètres θ, w_k est le poids associé à la donnée $\widehat{\gamma}(r_k)$, et r_1 à r_J sont les distances pour lesquelles une estimation du semi-variogramme a été calculée. Plusieurs possibilités pour le choix des poids ont été proposées dans la littérature. Dans cette étude, ils ont été choisis proportionnels avec le nombre de paires de données, $N(h)$, sur lequel on a calculé chaque point du variogramme expérimental. Si tous les poids sont fixés à 1 cela revient aux *moindres carrés ordinaires*. Une fois le modèle ajusté, on peut l'utiliser pour faire l'estimation souhaitée (section 2.6).

2.6 Estimation par krigeage. Types de krigeage

Dans cette section on présente d'abord la méthodologie générale qui permet de construire l'estimation par krigeage, ensuite on détaillera les principales variantes que l'on peut rencontrer dans les applications pratiques.

Rappelons d'abord que l'objectif du krigeage est de prévoir la valeur de la variable régionalisée à interpoler $z(.)$ en un site non échantillonné noté s_0. Avant de passer à l'estimation proprement dite, il faut choisir la taille m_0 du voisinage utilisé en krigeage en se basant sur une certaine connaissance de la structure de dépendance spatiale entre les observations.

Pour écrire les contraintes du krigeage, qui sont à la base de l'estimation, on fait appel au développement proposé par Chauvet (1999).

1. **Contrainte de linéarité**

   L'estimation $\widehat{Z}(s_0)$ est cherchée sous la forme d'une combinaison linéaire des données $Z(s_i)$; elle s'écrit :

   $$\widehat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i). \tag{2.26}$$

   Les paramètres à déterminer sont la constante a et les poids λ_i de chaque observation.

2. **Contrainte d'autorisation**

   Comme l'erreur d'estimation est une combinaison linéaire construite sur la fonction aléatoire, il faut s'assurer que les manipulations à l'ordre 2 sur cette erreur sont "autorisées". Cette contrainte intervient différemment en fonction du modèle de stationnarité adopté. Ainsi, dans le modèle stationnaire d'ordre 2, toutes les combinaisons linéaires sont autorisées. La contrainte

d'autorisation n'intervient que dans le cas où la fonction aléatoire $\delta(.)$ du modèle est supposée stationnaire intrinsèque. Dans ce cas, une combinaison linéaire est autorisée si et seulement si son poids total est nul.

3. **Contrainte de non-biais ou d'universalité**

   L'absence de biais se traduit par la relation : $E[\widehat{Z}(s_0) - Z(s_0)] = 0$ qui exprime le fait que l'erreur d'estimation est d'espérance nulle.

4. **Contrainte d'optimalité**

   La dernière contrainte à respecter est la minimisation de la variance de l'estimation

$$Var[\widehat{Z}(s_0) - Z(s_0)]. \tag{2.27}$$

NOTATIONS : On favorise une notation matricielle pour faciliter l'écriture des systèmes d'équations à résoudre ; voici quelques notations (où m_0 représente le nombre d'observations qui se trouvent dans le voisinage $V(s_0)$ considéré dans le processus d'estimation) :

- $\mathbf{Z}$ est le vecteur $m_0 \times 1$ des variables aléatoires qui intervient dans l'estimation ;
- $\boldsymbol{\lambda}$ est le vecteur $m_0 \times 1$ des poids associés aux variables aléatoires ci-dessus ;
- $\boldsymbol{\delta}$ est le vecteur $m_0 \times 1$ des erreurs associées aux mêmes variables aléatoires ;
- $\mathbf{1}_{m_0}$ est le vecteur $m_0 \times 1$ qui a 1 partout.

Dans le *cadre stationnaire d'ordre 2* :

- $\boldsymbol{\Sigma}$ est la matrice $m_0 \times m_0$ de variances-covariances de δ, dont la diagonale est composée uniquement de σ^2, la variance commune à toutes les erreurs $\delta(s)$ pour $x \in D$;
- c_0 est le vecteur $m_0 \times 1$ des covariances entre δ et $\delta(s_0)$.

Dans le *cadre stationnaire intrinsèque* :

- $\boldsymbol{\Gamma}$ est la matrice $m_0 \times m_0$ dont l'élément (i, j) est $\gamma(s_i - s_j)$, soit le semi-variogramme entre $\delta(s_i)$ et $\delta(s_j)$;
- γ_0 est le vecteur $m_0 \times 1$ dont l'élément i est $\gamma(s_i - s_0)$.

2.6.1 Le krigeage simple

On considère d'abord le cadre stationnaire d'ordre 2. Sous cette hypothèse, le krigeage le moins complexe est celui dans lequel l'espérance de la fonction aléatoire étudiée est supposée connue et constante sur tout le domaine D. Le modèle s'écrit :

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

avec m **constante connue** et $\delta(.)$ fonction aléatoire stationnaire de second ordre d'espérance nulle et de structure spatiale connue. Si on suit la démarche proposée par Chauvet on obtient les équations suivantes :

1. **Contrainte de linéarité**

   L'estimation de $Z(s_0)$ s'écrit de la manière suivante :

   $$\widehat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) = a + \boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{Z}. \tag{2.29}$$

2. **Contrainte d'autorisation**

   Cette contrainte est inactive dans le cadre stationnaire d'ordre 2.

3. **Contrainte de non-biais**

   La condition d'absence de biais est :

   $$E[\widehat{Z}(s_0) - Z(s_0)] = E[a + \boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{Z} - Z(s_0)] = a + \boldsymbol{\lambda}^{\boldsymbol{t}} m \mathbf{1}_{m_0} - m = 0. \tag{2.30}$$

   Il faut donc s'assurer que $a = m(1 - \boldsymbol{\lambda}^{\boldsymbol{t}} \mathbf{1}_{m_0})$. On obtient alors la formule :

   $$\widehat{Z}(s_0) = m(1 - \boldsymbol{\lambda}^{\boldsymbol{t}} \mathbf{1}_{m_0}) + \boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{Z} = m + \boldsymbol{\lambda}^{\boldsymbol{t}} (\boldsymbol{Z} - m \mathbf{1}_{m_0}). \tag{2.31}$$

4. **Contrainte d'optimalité**

   La dernière contrainte à respecter, la minimisation de l'erreur d'estimation, conduit à la relation suivante :

   $$\begin{aligned} Var[\widehat{Z}(s_0) - Z(s_0)] &= Var[m + \boldsymbol{\lambda}^{\boldsymbol{t}} (\boldsymbol{Z} - m \mathbf{1}_{m_0}) - Z(s_0)] \\ &= Var[m + \boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{\delta} - m - \delta(s_0)] \\ &= Var[\boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{\delta}] + Var[\delta(s_0)] - 2Cov[\boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{\delta}, \delta(s_0)] \\ &= \boldsymbol{\lambda}^{\boldsymbol{t}} Var[\delta] \boldsymbol{\lambda} + Var[\delta(s_0)] - 2\boldsymbol{\lambda}^{\boldsymbol{t}} Cov[\boldsymbol{\delta}, \delta(s_0)] \\ &= \boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{\Sigma} \boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}^{\boldsymbol{t}} \boldsymbol{c}_0 \end{aligned} \tag{2.32}$$

   On considère l'expression (2.32) comme une fonction $f(\boldsymbol{\lambda})$, dont le gradient nécessite le calcul de ses dérivées partielles :

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

   Si on calcule le hessien de $f(\boldsymbol{\lambda})$ on obtient $2\boldsymbol{\Sigma}$, matrice qui est semi-définie positive car il s'agit d'une matrice de variances-covariances multipliée par une constante positive. La fonction $f(\boldsymbol{\lambda})$ est donc convexe et le point critique est celui pour lequel le gradient s'annule, c'est-à-dire $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0$.

Par conséquent l'estimation par krigeage simple peut se faire en utilisant la formule :

$$\widehat{Z}(s_0) = m + \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - m\mathbf{1}_{m_0}). \tag{2.34}$$

Un avantage important du krigeage est qu'il permet le calcul de la variance de l'erreur d'estimation, appelée aussi "variance de krigeage". A partir des résultats précédents, la variance de krigeage simple σ^2_{KS} en un point s_0 peut être exprimée :

$$\begin{aligned} \sigma^2_{KS}(s_0) &= Var[\widehat{Z}(s_0) - Z(s_0)] \\ &= \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0 + \sigma^2 - 2\boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0 \\ &= \sigma^2 - \boldsymbol{c}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_0. \end{aligned} \tag{2.35}$$

Même si cette formule donne une idée sur la précision de l'estimation calculée, il ne faut pas oublier que cette variance ne tient pas compte de la variabilité due à l'estimation du semi-variogramme. On a donc une sous-estimation de la vraie variance d'estimation.

2.6.2 Le krigeage ordinaire

La méthode de krigeage simple qui part de la prémisse que l'espérance de la fonction aléatoire soit connue et constante a été généralisée au cas où **l'espérance est inconnue et constante localemen**t (sur le voisinage de krigeage $V(s_0)$). Ce type de krigeage ne requiert pas l'hypothèse de stationnarité d'ordre deux, mais l'hypothèse plus générale, celle de stationnarité intrinsèque. Le modèle sur lequel s'appuie le krigeage ordinaire est :

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

avec μ quasi-constante inconnue, pour la tendance, et $\delta(.)$ fonction aléatoire intrinsèque d'espérance nulle et de structure spatiale connue, pour la partie résiduelle. Pour l'estimation, on reprend les quatre étapes du krigeage qui correspondent aux quatre contraintes déjà énoncées.

1. **Contrainte de linéarité**

L'estimation doit être une combinaison linéaire des variables aléatoires impliquées dans le calcul :

$$\widehat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) = a + \boldsymbol{\lambda}^{\boldsymbol{t}}\boldsymbol{Z}. \tag{2.37}$$

2. **Contrainte d'autorisation**

Dans le modèle intrinsèque, une combinaison linéaire est autorisée si et seulement si la somme totale des poids est nulle (une combinaison linéaire d'accroissements est équivalente à une combinaison linéaire de variables aléatoires avec des poids de somme nulle). En écrivant l'erreur

d'estimation, on obtient :

$$\begin{aligned}\widehat{Z}(s_0) - Z(s_0) &= a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0) \\ &= a + \sum_{i \in V(s_0)} \lambda_i (\mu + \delta(s_i)) - \mu - \delta(s_0) \\ &= \underbrace{a + \mu \sum_{i \in V(s_0)} \lambda_i - \mu}_{\text{termes non aléatoires}} + \sum_{i \in V(s_0)} \lambda_i \delta(s_i) - \delta(s_0).\end{aligned} \tag{2.38}$$

Il faudra donc travailler avec la contrainte que la somme de poids λ_i vaut 1 pour s'assurer de l'existence des deux premiers moments de l'erreur d'estimation.

3. **Contrainte de non-biais**

Comme pour tout autre type de krigeage on souhaite une estimation sans biais, alors on impose la contrainte :

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

Comme précédemment, on a la contrainte $\sum_{i \in V(s_0)} \lambda_i = 1$ qui doit être respectée ; on obtient alors que a doit être fixé à zéro. Dans ces conditions, compte tenu de la première contrainte de linéarité, la formule d'estimation 2.37 peut être simplifiée à :

$$\widehat{Z}(s_0) = \sum_{i \in V(s_0)} \lambda_i Z(s_i) \quad \text{avec} \quad \sum_{i \in V(s_0)} \lambda_i = 1. \tag{2.40}$$

4. **Contrainte d'optimalité**

En utilisant les relations antérieures on impose maintenant la condition de minimalisation de la variance de l'erreur d'estimation en utilisant $\mathbf{\Gamma}$ et γ_0 :

$$\begin{aligned}&Var[\widehat{Z}(s_0) - Z(s_0)] \\ &= E\left[\left(\sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0)\right)^2\right] \\ &= E\left[\sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - 2\delta(s_0) \sum_{i \in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2\right] \\ &= E[\underbrace{\sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum_{i \in V(s_0)} \lambda_i \delta(s_i)^2}_{\text{terme 1}} + \\ &\quad \underbrace{\sum_{i \in V(s_0)} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i \in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2}_{\text{terme 2}}].\end{aligned} \tag{2.41}$$

Si on réécrit les deux termes on a :

$$\begin{aligned} & Var[\widehat{Z}(s_0) - Z(s_0)] \\ &= E\left[-\frac{1}{2}\sum_{i\in V(s_0)}\sum_{j\in V(s_0)}\lambda_i\lambda_j(\delta(s_i)-\delta(s_j))^2\right] + E\left[\sum_{i\in V(s_0)}\lambda_i(\delta(s_i)-\delta(s_0))^2\right] \\ &= -\sum_{i\in V(s_0)}\sum_{j\in V(s_0)}\lambda_i\lambda_j\gamma(s_i-s_j) + 2\sum_{i\in V(s_0)}\lambda_i\gamma(s_0-s_i) \\ &= -\boldsymbol{\lambda}^t\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t\gamma_0. \end{aligned} \tag{2.42}$$

Il nous reste donc à minimiser cette expression conditionnellement aux contraintes précédentes. Il faut donc utiliser la technique classique des multiplicateurs de Lagrange. Notons que l'annulation de la dérivée partielle par rapport à l (le Lagrangien) ne fait que restituer la première contrainte.
On a donc la fonction $f(\boldsymbol{\lambda}, l) = -\boldsymbol{\lambda}^t\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t\gamma_0 + 2l(\boldsymbol{\lambda}^t\mathbf{1}_{m_0} - 1)$. Le vecteur de ses dérivées partielles est :

$$\frac{\partial}{\partial\boldsymbol{\lambda}}f(\boldsymbol{\lambda}, l) = -2\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\gamma_0 + 2l\mathbf{1}_{m_0}$$

et le point critique de cette fonction est

$$\widehat{\boldsymbol{\lambda}} = \boldsymbol{\Gamma}^{-1}(\gamma_0 + l\mathbf{1}_{m_0}). \tag{2.44}$$

Le Lagrangien est estimé en utilisant la contrainte que la somme des poids vaille 1, donc

$$\widehat{l} = -\frac{1 - \mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\gamma_0}{\mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_{m_0}}. \tag{2.45}$$

Si on remplace l par $\widehat{l}$ (éq. 2.45) dans $\widehat{\boldsymbol{\lambda}}$ (éq. 2.44) on obtient :

$$\widehat{\boldsymbol{\lambda}}^t = (\gamma_0 + \mathbf{1}_{m_0}\frac{1 - \mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\gamma_0}{\mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_{m_0}})^t\boldsymbol{\Gamma}^{-1}. \tag{2.46}$$

En remplaçant l'expression (2.46) dans la formule d'estimation (2.40) on obtient :

$$\widehat{Z}(s_0) = (\gamma_0 + \frac{1 - \mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\gamma_0}{\mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_{m_0}}\mathbf{1}_{m_0})^t\boldsymbol{\Gamma}^{-1}Z \tag{2.47}$$

et la variance de l'estimation σ^2_{KS} s'écrit :

$$\sigma^2_{KS}(s_0) = Var[\widehat{Z}(s_0) - Z(s_0)] = \gamma_0^t\boldsymbol{\Gamma}^{-1}\gamma_0 - \frac{(1 - \mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\gamma_0)^2}{\mathbf{1}_{m_0}^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_{m_0}}. \tag{2.48}$$

2.6.3 Le krigeage universel

Postuler que l'espérance d'une fonction aléatoire reste constante ou quasi-constante sur le domaine D est souvent erroné. C'est la raison pour laquelle le krigeage ordinaire a été généralisé pour arriver à un modèle de tendance dans lequel soit l'espérance est une fonction de coordonnées spatiales, soit elle est une fonction de variables régionalisées auxiliaires connues exhaustivement. Dans le premier cas se situe le krigeage universel, et dans le deuxième, le krigeage avec dérive

externe.

Le modèle de krigeage universel est :

$$Z(s) = \sum_{j=0}^{p} f_j(s)\beta_j + \delta(s), \qquad s \in D, \tag{2.49}$$

avec $f_j(s)$ fonctions de la position (coordonnées géographiques), β_j paramètres inconnus et $\delta(.)$ fonction aléatoire stationnaire intrinsèque d'espérance nulle et de structure de dépendance spatiale connue.

Physiquement, la dichotomie postulée n'aura d'intérêt que si elle sépare des phénomènes d'échelles différentes. On espère donner une signification naturelle à cette dichotomie : d'un côté une composante régionale, une "tendance" correspondant aux "basses fréquences" du phénomène et, d'un autre côté, une composante résiduelle, erratique, correspondant aux "hautes fréquences" du phénomène. D'un point de vue mathématique, on espère que le résidu possédera les bonnes propriétés de stationnarité d'ordre 2, qui permettront de le traiter par les méthodes déjà connues.

On se rapproche ainsi des méthodes de type fréquentiel : ce que nous appelons "dérive" est alors associé aux basses fréquences, sans préjuger de leur origine, ni de leur signification physique (Chauvet, 1999). Il existe deux interprétations pour cette dérive. La plus intuitive consiste à dire que, à l'échelle de travail, la dérive a une forme analytique simple, en général porteuse d'aucun message naturaliste. La deuxième considère la dérive comme un phénomène suffisamment régulier pour pouvoir raisonnablement être approchée par les premiers termes d'un développement selon un certain jeu de fonctions simples.

En pratique, on choisit souvent une tendance linéaire ou quadratique. Par exemple, si l'utilisateur cherche une tendance linéaire alors les fonctions $f_j(s)$ choisies sont : $f_0(s) = 1$, $f_1(s) = x$ et $f_2(s) = y$. Par la suite nous introduisons des nouvelles notations :

- le vecteur des paramètres de la dérive

$$\boldsymbol{\beta} = (\beta_0,\ \beta_1, \ldots,\ \beta_p), \tag{2.50}$$

- le vecteur

$$x_0 = (f_0(s_0),\ f_1(s_0), \ldots,\ f_p(s_0)) \tag{2.51}$$

- la matrice X de dimension $m_0 \times (p+1)$ dont l'élément (i, j) est $f_j(s_i)$. Dans le cas particulier mentionné (tendance linéaire), la matrice X est :

$$X = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_{m_0} & y_{m_0} \end{pmatrix}$$

En suivant la même démarche, on obtient les équations du krigeage universel.

1. **Contrainte de linéarité**

Comme toujours, l'estimation est une combinaison linéaire des $Z(s_i)$:

$$\widehat{Z}(s_0) = a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) = a + \boldsymbol{\lambda}^t \boldsymbol{Z}. \tag{2.52}$$

2. **Contrainte d'autorisation**

Suite à l'hypothèse de stationnarité intrinsèque, il faut s'assurer que l'erreur soit une combinaison linéaire d'accroissements de $\delta(.)$. Donc pour l'erreur de prévision on a :

$$\begin{aligned} \widehat{Z}(s_0) - Z(s_0) &= a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0) \\ &= a + \sum_{i \in V(s_0)} \lambda_i (s_i \boldsymbol{\beta} + \delta(s_i)) - s_0 \boldsymbol{\beta} - \delta(s_0) \\ &= a + \sum_{i \in V(s_0)} \lambda_i s_i \boldsymbol{\beta} - s_0 \boldsymbol{\beta} + \sum_{i \in V(s_0)} \lambda_i \delta(s_i) - \delta(s_0) \end{aligned} \tag{2.53}$$

Il faut donc que $\sum_{i \in V(s_0)} \lambda_i = 1$ comme précédemment.

3. **Contrainte de non-biais**

En reprenant le calcul on a :

$$\begin{aligned} E[\widehat{Z}(s_0) - Z(s_0)] &= E[a + \sum_{i \in V(s_0)} \lambda_i Z(s_i) - Z(s_0)] \\ &= a + \sum_{j=0}^{p} (\sum_{i \in V(s_0)} \lambda_i f_j(s_i) - f_j(s_0)) \beta_j. \end{aligned} \tag{2.54}$$

Afin que cette espérance vaille zéro pour tout β_j, $j = 0, ...p$, il faut que $a = 0$ et

$$\sum_{i \in V(s_0)} \lambda_i f_j(s_i) = f_j(s_0) \text{ pour } j = 0, \ldots, p. \tag{2.55}$$

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. Sous forme matricielle ces contraintes s'écrivent :

$$\boldsymbol{\lambda}^t \boldsymbol{X} = \boldsymbol{x}_0^t. \tag{2.56}$$

Si pour simplifier les calculs, on suppose comme dans le cas d'une tendance linéaire que $f_0(.) = 1$ (postulat usuel en krigeage Cressie (1993)), cela nous permet d'éliminer une contrainte, notamment $\sum_{i \in V(s_0)} \lambda_i = 1$, écrite pour $j = 0$ qui coincide avec la contrainte d'autorisation.

4. **Contrainte d'optimalité**

La fonction à minimiser est :

$$f(\boldsymbol{\lambda}, l) = -\boldsymbol{\lambda}^t \boldsymbol{\Gamma} \boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t \gamma_0 + 2(\boldsymbol{\lambda}^t \boldsymbol{X} - \boldsymbol{x}_0^t) \boldsymbol{l}, \tag{2.57}$$

où le vecteur $\boldsymbol{l}$ représente les $p+1$ Lagrangiens à calculer. Le gradient de la fonction (2.57) vaut :

$$\frac{\partial}{\partial \boldsymbol{\lambda}} f(\boldsymbol{\lambda}, l) = -2\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\gamma_0 + 2\boldsymbol{X}\boldsymbol{l}. \tag{2.58}$$

Le vecteur de Lagrangiens est estimé par :

$$\widehat{l} = (\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\gamma_0). \tag{2.59}$$

Ensuite, si on remplace $\widehat{l}$ dans $\widehat{\boldsymbol{\lambda}}$ on obtient l'unique point de minimum de la fonction $f(\boldsymbol{\lambda}, l)$:

$$\widehat{\boldsymbol{\lambda}} = \boldsymbol{\Gamma}^{-1}(\gamma_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\gamma_0). \tag{2.60}$$

Ainsi, l'estimation en s_0 sera donnée par :

$$\widehat{Z}(s_0) = (\gamma_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\gamma_0))^t\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}. \tag{2.61}$$

La variance de krigeage σ^2_{KU} devient alors :

$$\sigma^2_{KU}(s_0) = Var[\widehat{Z}(s_0) - Z(s_0)] = \gamma_0^t\boldsymbol{\Gamma}^{-1}\gamma_0 - (\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\gamma_0)^t(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\gamma_0). \tag{2.62}$$

2.6.4 L'analyse variographique en krigeage avec modèle de tendance

En présence d'une tendance, l'analyse variographique se complique, car les structures systématiques peuvent contaminer le variogramme expérimental et conduire à des résultats inacceptables. Pour traiter des données en présence d'une tendance, on a à notre disposition quatre méthodes : le krigeage universel (décrit dans la section précédente 2.6.3), le krigeage intrinsèque généralisé (voir la section 2.6.6), le krigeage résiduel et le krigeage avec dérive externe. La différence principale entre ces méthodes est que celui universel et celui résiduel sont basés sur une estimation explicite de la dérive, tandis que le krigeage intrinsèque généralisé l'élimine par une procédure de filtrage.

Tel que mentionné dans la section 2.4, le semi-variogramme est estimé à partir des résidus $z(s_i) - \widehat{\mu}(s_i)$, obtenus suite à une estimation du vecteur des paramètres de la tendance β. En pratique, quand la dérive n'est pas connue, ni constante, la démarche classique consiste à l'estimer, et ensuite à la retirer des données en réduisant ainsi le problème à celui étudié auparavant.

Cette demarche a été largement appliqué aux séries temporelles, procédure connue sous le nom anglais "detrending". Même si cette approche est raisonnable, "il est très difficile d'analyser les propriétés théoriques d'une telle combinaison entre deux procédures d'estimation. Il faut toujours se demander dans quelle mesure les variations haute fréquence sont corrompues par les estimations basse fréquence de la moyenne et causer ainsi la violation de la condition d'une moyenne nulle" (Chilès et Delfiner, 1999).

En revenant à l'estimation du vecteur β, ceci requiert la connaissance de la structure de dépendance spatiale de la fonction aléatoire résiduelle $\delta(.)$. 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 obtenir d'abord un estimateur au sens des moindres carrés ordinaires de β, ensuite estimer un semi-variogramme sur les résidus, puis calculer un estimateur des moindres carrés généralisés de β, et ainsi de suite (Cressie, 1993). À mesure que les itérations avancent, 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.

Néanmoins, Cressie (1993) remarque également que, même si on applique la méthode itérative décrite, l'estimation du semi-variogramme reste biaisée (Matheron, 1965). Les conséquences de ce biais sont difficiles à évaluer. D'après Cressie (1993), la méthode du krigeage universel peut s'avérer efficace, mais il reste quand même une certaine insatisfaction liée au biais présent dans l'estimation du semi-variogramme.

2.6.5 Lien entre le krigeage universel et le krigeage résiduel (appliqué sur les résidus d'une régression)

Reprenons l'équation d'estimation du krigeage universel (2.61) et écrivons-la dans le cas stationnaire d'ordre deux (on utilise donc, la fonction de covariance à la place d'un semi-variogramme). Cela revient à écrire :

$$\widehat{Z}(s_0) = (c_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}c_0))^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \tag{2.63}$$

La même estimation peut être obtenue en effectuant un krigeage simple (avec $m = 0$) 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 estimés par une méthode des moindres carrés généralisés (GLS) pour obtenir $\widehat{\beta}_{gls}$, ensuite la prévision en s_0 est obtenue en additionnant la prévision de la tendance générale par régression à la prévision par krigeage simple (KS) des erreurs $e = Z - X\widehat{\beta}_{gls}$:

$$\begin{aligned}
\widehat{Z}(s_0) &= x_0^t\widehat{\beta}_{gls} + \widehat{e}_{KS}(s_0) \\
&= x_0^t\widehat{\beta}_{gls} + c_0^t\boldsymbol{\Sigma}^{-1}e \\
&= x_0^t\widehat{\beta}_{gls} + c_0^t\boldsymbol{\Sigma}^{-1}(Z - X\widehat{\beta}_{gls}) \\
&= (x_0^t - c_0^t\boldsymbol{\Sigma}^{-1}X)\widehat{\beta}_{gls} + c_0^t\boldsymbol{\Sigma}^{-1}Z \\
&= (x_0^t - c_0^t\boldsymbol{\Sigma}^{-1}X)(X^t\boldsymbol{\Sigma}^{-1}X)^{-1}X^t\boldsymbol{\Sigma}^{-1}Z + c_0^t\boldsymbol{\Sigma}^{-1}Z \\
&= (c_0 + \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}c_0))^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}.
\end{aligned} \tag{2.64}$$

Certains utilisateurs emploient cette approche en effectuent d'abord une régression qui ne tient pas compte de la structure spatiale de données, comme par exemple une régression de moindres carrés ordinaires ($\widehat{\beta}_{ols} = (X^tX)^{-1}X^tZ$, Cressie (1993)). Les prévisions obtenues ne sont pas les mêmes que celles obtenues par krigeage universel. Cependant, Kitanidis (1993) a démontré que l'utilisation d'une fonction de covariance/semi-variogramme, estimée sur les résidus obtenus par une régression des moindres carrés ordinaires (OLS), est satisfaisante, car elle n'est vraiment pas trop différente d'une fonction estimée sur les résidus obtenus après plusieurs itérations, et donc l'estimation finale n'est pas trop influencée.

2.6.6 Le krigeage intrinsèque généralisé

Estimer un processus aléatoire à moyenne inconnue nécessite, comme on l'a déjà vu, d'ajouter des contraintes sur l'estimateur. On cherche cette fois-ci une alternative au modèle (2.49) de la dichotomie entre la tendance et les résidus. Une façon très élégante de traiter ce problème d'indétermination consiste à transformer linéairement le processus aléatoire de manière à éliminer les termes inconnus. Le processus transformé est appelé processus aléatoire généralisé.

Le krigeage intrinsèque, qui sera appelé généralisé par la suite, étend et formalise le principe de prédiction par krigeage universel. Le point de départ de cette théorie est de traduire la contrainte d'autorisation en termes de relation d'orthogonalité. Le principe fondamental est donc le suivant : on impose que l'erreur de prédiction soit minimale au sens quadratique, tout en requérant qu'elle soit orthogonale en un certain sens à l'espace engendré par les termes paramétriques constituant la moyenne du processus. C'est le principe de la meilleure approximation qui justifie la condition d'orthogonalité, au sens où il n'est pas possible d'améliorer l'erreur de prédiction en ajoutant à l'estimateur toute combinaison linéaire de termes paramétriques. On introduit avec cette théorie la notion de covariance généralisée qui complète l'arsenal dont nous avons besoin pour travailler sur les processus aléatoires généralisés.

On commence par introduire la notion de fonctions aléatoires intrinsèques d'ordre k. Cette notion constitue une généralisation naturelle des fonctions aléatoires intrinsèques (à accroissements stationnaires) déjà définies, qui correspondent au cas particulier $k = 0$ et qui seront appelées FAI-0. Dans le cadre stationnaire, l'outil de travail essentiel était la covariance qui devait être de type positif. Dans le cadre intrinsèque, on gagne en généralité et on utilise le variogramme qui, contrairement à la covariance, n'est pas borné et permet de décrire des phénomènes présentant une capacité de dispersion potentiellement illimitée. De plus, on exige seulement que $-\gamma$ soit de type conditionnel positif.

L'extension de FAI-0 aux FAI-k est naturelle et nécessite un changement de notation. L'erreur d'estimation est donnée par la formule $Z(s_0) - \sum_{i=1}^{n} \lambda_i Z(s_i)$ où les s_i sont les points de mesure et s_0 est le point à estimer. Par la suite, en affectant à s_0 un poids λ_0 égal à -1, on obtient une combinaison linéaire de la forme $\sum_{\alpha=0}^{n} \lambda_\alpha Z(s_\alpha)$.

La condition d'universalité s'écrit :

$$\sum_{\alpha} \lambda_\alpha f_j(s_\alpha) = 0, \tag{2.65}$$

où les f_j sont des monômes de degré $\leq k$. On dira qu'une combinaison linéaire attribuant les poids λ_α aux points s_α est autorisée à l'ordre k si elle vérifie la condition (2.65), c'est-à-dire qu'elle annule les polynômes de degré inférieur ou égal à k. Maintenant, pour rendre possible l'inférence statistique, il suffit d'imposer la condition de stationnarité aux seules combinaisons linéaires autorisées. Mathématiquement, cela s'exprime en disant que, pour toute combinaison linéaire autorisée, la fonction aléatoire $x \to \sum_\alpha \lambda_\alpha Z(s + s_\alpha)$ est stationnaire en s. On montre qu'il existe alors une fonction appelée covariance généralisée (CG), qui n'est pas nécessairement une vraie covariance, $K(h)$, telle

que l'on ait :

$$E[(\sum_{\alpha} \lambda_\alpha Z(s_\alpha))^2] = \sum_{\alpha} \sum_{\beta} \lambda_\alpha \lambda_\beta K(s_\alpha - s_\beta) \tag{2.66}$$

pour toute combinaison linéaire autorisée. Dans le cas particulier d'une FAI-0 on obtient :

$$K(h) = -\gamma(h). \tag{2.67}$$

On remarque que ce n'est pas $Z(s)$ elle-même le véritable instrument de travail, mais la classe de toutes les fonctions aléatoires égales à $Z(s)$ à un polynôme près de degré $\leq k$ à coefficients aléatoires. D'ailleurs, on appellera cette classe d'équivalence Fonction Aléatoire Intrinsèque d'ordre k (FAI-k). Si $Y(s)$ est un élément quelconque de cette classe, alors tout autre élément $Z(s)$ sera de la forme :

$$Z(s) = Y(s) + \sum_{l} A_l f_l(s). \tag{2.68}$$

La notion de dérive est donc implicitement contenue dans celle de FAI-k, mais il n'est nullement nécessaire de supposer $Y(s)$ stationnaire. En général, une FAI-k n'admet pas de représentation stationnaire.

Comme on l'a vu, l'erreur d'estimation est par construction une combinaison linéaire autorisée, alors que la variance d'estimation ne dépend que de la CG $K(h)$. L'estimateur optimal s'obtient donc en résolvant le même système que dans le cas du krigeage universel, avec le seul changement consistant à remplacer le variogramme $\gamma(h)$ par la CG $K(h)$.

Une observation immédiate est que les FAI-k ont un degré de généralité beaucoup plus élevé que les FA stationnaires ordinaires. Par exemple si on prend $k+1$ coefficients positifs $a_0, ..., a_k$, il existe toujours une FAI-k admettant la covariance généralisée :

$$K(h) = \sum_{i=0}^{k} (-1)^{i+1} a_i |h|^{2i+1}, \tag{2.69}$$

alors que manifestement, cette fonction n'est pas une covariance stationnaire.

Le modèle polynômial (2.69) de covariance généralisée se prête bien à une procédure entièrement automatisée d'inférence statistique, car les coefficients a_i interviennent linéairement et sont donc faciles à estimer (Kitanidis, 1993). Christakos et Thesing (1993) démontrent qu'on peut obtenir un accord raisonnable avec les données expérimentales, sans dépasser l'ordre $k = 1$ ou 2, mais il existe quand même toujours des types de données qui se prêtent mal à une représentation par le modèle de covariance polynomiale.

Dans le cas du modèle (2.69), comme la covariance généralisée doit être une fonction de type conditionnel non négatif, on obtient, en appliquant les conditions d'admissibilité de Christakos (1984), les restrictions suivantes :

- si $k = 0$, alors

$$a_0 \geq 0;$$

- si $k = 1$, alors

$$a_1, a_0 \geq 0;$$

- si $k = 2$, alors

$$a_2, a_0 \geq 0 \qquad a_1 \geq -2\left[\frac{5}{3}\left(\frac{n+3}{n+1}\right)a_0 a_2\right]^{1/2},$$

avec $n = 1, 2, 3$.

Un modèle remarquable de CG utilisé pour les phénomènes non-stationnaires est :

$$K(h) = h^m \log(h), \text{ avec } m > 0; \tag{2.73}$$

en combinant les modèles (2.69) et (2.73) on obtient un modèle très souvent utilisé surtout dans le cas d'un processus atmosphérique (Christakos et Thesing, 1993) :

$$K(h) = c_0\delta(h) - a_0 h + a_1 h^2 \log(h) + a_2 h^3, \tag{2.74}$$

avec les restrictions imposées à ses coefficients (en $\mathbb{R}^2$) :

$$a_2, a_0, c_0 \geq 0 \text{ et } a_1 \geq -1.5(a_0 a_2)^{1/2}. \tag{2.75}$$

Le point encore délicat est celui de l'inférence statistique de la covariance généralisée. La technique décrite par Christakos et Thesing (1993) est basée sur l'identification automatique d'une covariance généralisée, en utilisant une régression par moindres carrés, à partir d'un certain nombre préétabli de modèles élémentaires de CG. Elle utilise un estimateur de type jackknife et teste les rangs accordés aux modèles de CG, pour trouver celui qui correspond le mieux aux critères de sélection. Cette procédure automatique est controversée premièrement à cause du rôle très limité du géostatisticien qui doit choisir parmi les modèles identifiés par l'ordinateur. Dans le cas d'une modélisation classique, avec des variogrammes expérimentaux, le rôle principal revenait à l'opérateur, qui identifie le modèle et qui effectue l'ajustement pour calculer ces paramètres. Un autre problème est dû à l'impossibilité d'inférer expérimentalement une CG. Cette technique sera décrite en détail par la suite et sera étendue dans le cas spatio-temporel (chapitre suivant), pour tenir compte de la dimension temporelle.

2.6.6.1 Description de la procédure automatique

Les principales étapes de la procédure automatique mentionnée auparavant, basée sur les travaux de Christakos (1992) et Christakos et Thesing (1993), seront détaillées par la suite, en faisant référence à la concentration de polluant comme variable à interpoler. L'ordre k de la FAI $- k$ sera noté dans la suite par ν et on l'appellera ordre de continuité (la notation k_x sera utilisée pour la fonction de covariance généralisée).

- **Première étape :** l'ensemble des données contenant les coordonnées spatiales suivies par les valeurs mesurées de la variable à interpoler est analysé : généralement, avoir des données très proches spatialement est à éviter à cause des instabilités numériques que cela pourrait générer.

- **Deuxième étape :** pour chaque nœud s_i du domaine D, un sous-domaine N_i, contenant un nombre fixe de données, est choisi (à l'intérieur d'un domaine de rayon fixé), données qui révèlent des caractéristiques importantes concernant la variabilité spatiale de la variable étudiée. Une covariance généralisée spatiale *initiale* est fixée (dans cette étude le modèle choisi est $k_x(r) = -r$, modèle valide, qui respecte les conditions d'admissibilité).
  Dans chaque sous-domaine, chaque donnée est retirée et l'estimation, ainsi que la variance de l'erreur d'estimation sont calculées en utilisant le reste des données :

$$\widehat{X}_\nu(s_i) = \sum_j \lambda_j X_\nu(s_j) \tag{2.76}$$

  et

$$\sigma^2_{x,\nu} = E\left[\widehat{X}_\nu(s_i) - X_\nu(s_i)\right]^2 = \sigma^2_{i,\nu} \tag{2.77}$$

  pour toutes les valeurs de l'ordre de continuité ν ($\nu = 0, 1, 2$), où les coefficients λ_j doivent être déterminés par la minimisation de $\sigma^2_{i,\nu}$.
  Ensuite, on classe les erreurs ($\sigma^2_{i,0}$, $\sigma^2_{i,1}$ et $\sigma^2_{i,2}$) en leur assignant des rangs (1, 2 ou 3) et on calcule la moyenne de ces rangs pour tous les points appartenant au sous-domaine :

$$\text{Avrank}(N_i, \nu) = \frac{\sum_{l \in N_i} \text{rang}(l, \nu)}{|N_i|}, \quad \text{pour } \nu = 0, 1, 2 \tag{2.78}$$

  Le ν qui conduit au rang le plus bas est celui qui sera choisi comme ordre de la FAI$-\nu$, appelé aussi l'ordre de continuité sur le sous-domaine N_i.

- **Troisième étape :** étant donné l'ordre de continuité ν, on détermine la covariance généralisée (CG $-\ \nu$). Il existe un ensemble de p candidates de la forme 2.69 (où $p = 3$ si $\nu = 0$, $p = 7$ si $\nu = 1$ et $p = 15$ si $\nu = 2$). Les coefficients de ces modèles ($c_{p,0}$ et $a_{p,l}$, avec $l = 0, 1, ..., \nu$) sont estimés en respectant deux conditions : le meilleur ajustement pour les données disponibles dans chaque sous-domaine et les conditions d'admissibilité. Pour accomplir cette tâche, on retire chaque point et on fait l'estimation en utilisant le reste des points. On écrit l'incrément spatial d'ordre ν :

$$Y(s_i) = \sum_{j \in N_i \cup i} q_{ji} X(s_j), \tag{2.79}$$

  où $q_{ii} = -1$ et $q_{ji} = \lambda_j (i \neq j)$, avec λ_j les coefficients d'estimation obtenus pour chaque point s_i en utilisant le voisinage N_i et en supposant $k^0_p(r) = -r$. On calcule l'espérance mathématique du carré de l'estimation :

$$A_i = E[Y(s_i)^2] = \sum_{j_a \in N_i \cup i} \sum_{j_b \in N_i \cup i} q_{j_a i} q_{j_b i} k_p(s_{j_a} - s_{j_b}), \tag{2.80}$$

  et on définit une fonction objectif :

$$F = \sum_{i \in N_i} [Y(s_i)^2 - A_i]^2. \tag{2.81}$$

  En introduisant $k_p(r)$ dans l'équation 2.80, on calcule les coefficients $c_{p,0}$ et $a_{p,l}$, avec $l = 0, 1, ..., \nu$ en minimisant l'équation 2.81 par rapport à ces coefficients, c'est-à-dire qu'on impose les conditions :

$$\frac{\partial F}{\partial c_{p,0}} = \frac{\partial F}{\partial a_{p,l}} = 0 \tag{2.82}$$

pour tout $l = 0, 1, ..., \nu$, et on résout les équations 2.82 pour $c_{p,0}^{(1)}$ et $a_{p,l}^{(1)}$ en trouvant ainsi $k_p^{(1)}(r)$ (les coefficients déterminés doivent vérifier les conditions d'admissibilité). On répète cette procédure mais on utilise $k_p^{(1)}(r)$ à la place de $k_p^{(0)}(r)$ jusqu'à la convergence, c'est-à-dire $c_{p,0}^{(m-1)} \approx c_{p,0}^{(m)}$ et $a_{p,l}^{(m-1)} \approx a_{p,l}^{(m)}$. Par conséquent, $k_p^{(m-1)}(r) \approx k_p^{(m)}(r)$. Si ces coefficients respectent les conditions d'admissibilité alors on retient cette candidate et on recommence la procédure itérative avec la nouvelle fonction de covariance jusqu'à la convergence de la solution. Si les coefficients ne respectent pas les conditions d'admissibilité on passe à la candidate suivante.

À la fin, parmi les candidates qui respectent les deux conditions antérieures, on doit trouver la meilleure. Une bonne mesure pour vérifier la qualité de cet ajustement est donnée par le calcul d'un indice :

$$\eta_p = \frac{\sum_{i \in N_i} Y(s_i)^2}{\sum_{i \in N_i} A_i}. \tag{2.83}$$

L'indice η est calculé pour chaque modèle de covariance retenu et le modèle dont l'indice η est le plus proche de 1 est sélectionné.

- **Quatrième étape :** l'ordre de continuité ν, ainsi que le modèle CG $-\ \nu$ sélectionnés sont utilisés pour calculer les poids en résolvant le système de krigeage. Ce système est le même que celui du krigeage universel, avec une tendance polynomiale de degré ν dépendant des coordonnées spatiales ; la seule différence par rapport au krigeage universel et à son système est que, dans celui intrinsèque généralisé, la matrice $\mathbf{\Gamma}$ de KU, dont les éléments $\gamma_{ij} = \gamma(s_i - s_j)$ représentaient le variogramme calculé pour la distance qui séparait les deux points s_i et s_j est remplacée par une autre K, dont les éléments k_{ij} sont calculés en utilisant la covariance généralisé (CG) $k_{ij} = k(s_i - s_j)$.
- **Cinquième étape :** les poids calculés sont utilisés pour cartographier la concentration d'un polluant sur tout le domaine d'étude, ainsi que la variance de cette estimation (l'écart-type).

Le modèle de FAI-k est très important surtout d'un point de vue théorique. Travailler sur des classes d'équivalence de fonctions aléatoires permet de bien différencier les propriétés intrinsèques, qui sont accessibles expérimentalement, de celles qui restent en général indéterminées. Cela nous permet aussi d'élargir le champ des modèles disponibles et de préciser ce qui doit et peut être identifié du variogramme ou de la covariance généralisée en présence d'une dérive. Le modèle de FAI-k confirme aussi en particulier l'identité pratique entre le modèle du krigeage universel et les FAI-k lorsque la covariance généralisée se ramène, au signe près, à un variogramme, ainsi que l'identité entre le modèle des splines plaque mince ou bien RBF (voir les sections 2.2.0.7 et 2.2.0.8) avec la FAI-1 quand on utilise une CG de type logarithmique (2.74).

Une première faiblesse de la méthode provient du fait que les covariances généralisées utilisées sont isotropes (sinon on se retrouve en présence d'un trop grand nombre de paramètres). Une autre faiblesse de la méthode vient de son manque de robustesse vis-à-vis des variables qui ne vérifient pas bien les hypothèses intrinsèques : présence de pics isolés, alternance de zones plates et de zones chahutées, données atypiques.

Fondée uniquement sur les accroissements généralisés, l'approche décrite ici ne nécessite plus la dichotomie entre dérive et résidus. Pour atteindre les mêmes buts que dans le modèle de krigeage

universel, on fait l'économie de nombreuses hypothèses et de nombreux paramètres. De plus, une nouvelle famille de covariances est apparue, les covariances polynomiales, qui dépendent linéairement d'un petit nombre de paramètres.

Pour conclure, le modèle FAI-k à covariances polynomiales n'est pas forcément toujours le meilleur modèle possible, mais il est caractérisé par une grande rapidité de réponse et généralement les cartes d'interpolation produites en utilisant cette méthode sont cohérentes.

2.7 Conclusion partielle du chapitre

En analysant les méthodes présentées dans ce chapitre, on peut dire que le krigeage semble être la méthode la plus intéressante, et cela, pour plusieurs raisons. Premièrement, de la même façon qu'avec les méthodes barycentriques, l'utilisateur du krigeage a le choix d'interpoler localement ou globalement. De plus, il s'agit d'une méthode stochastique, donc on peut calculer les erreurs de prévision. 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. 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. On finit cette présentation théorique par un récapitulatif des formulations pour les quatre types de krigeage (on garde les notations déjà annoncées en début de section 2.6, et pour $\boldsymbol{X}$ et $\boldsymbol{x}_0$ voir les notations de la section 2.6.3).

- le KS (Krigeage Simple) :
  - **Modèle** : $Z(s) = m + \delta(s)$ avec m constante connue
  - **Fonction à minimiser** : $f(\boldsymbol{\lambda}) = \boldsymbol{\lambda^t \Sigma \lambda} + \sigma^2 - 2\boldsymbol{\lambda^t c}_0$
  - **Estimation** : $\widehat{Z}(s_0) = m + \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - m\boldsymbol{1}_{m_0})$
  - **Calcul de la variance** : $\sigma^2_{KS}(s_0) = \sigma^2 - \boldsymbol{c}_0^t \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0$
- le KO (Krigeage Ordinaire) :
  - **Modèle** : $Z(s) = \mu + \delta(s)$ avec μ quasi-constante inconnue
  - **Fonction à minimiser** : $f(\boldsymbol{\lambda}, l) = -\boldsymbol{\lambda}^t \boldsymbol{\Gamma} \boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t \gamma_0 + 2l(\boldsymbol{\lambda}^t \boldsymbol{1}_{m_0} - 1)$
  - **Estimation** : $\widehat{Z}(s_0) = (\gamma_0 + \frac{1 - \boldsymbol{1}_{m_0}^t \boldsymbol{\Gamma}^{-1} \gamma_0}{\boldsymbol{1}_{m_0}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{1}_{m_0}} \boldsymbol{1}_{m_0})^t \boldsymbol{\Gamma}^{-1} Z$
  - **Calcul de la variance** : $\sigma^2_{KO}(s_0) = \gamma_0^t \boldsymbol{\Gamma}^{-1} \gamma_0 - \frac{(1 - \boldsymbol{1}_{m_0}^t \boldsymbol{\Gamma}^{-1} \gamma_0)^2}{\boldsymbol{1}_{m_0}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{1}_{m_0}}$
- le KU (Krigeage Universel) :
  - **Modèle** : $Z(s) = \sum_{i=0}^{p} f_j(s) \beta_j + \delta(s)$ avec $f_j(s)$ fonctions de la position (coordonnées géographiques)
  - **Fonction à minimiser** : $f(\boldsymbol{\lambda}, l) = -\boldsymbol{\lambda}^t \boldsymbol{\Gamma} \boldsymbol{\lambda} + 2\boldsymbol{\lambda}^t \gamma_0 + 2(\boldsymbol{\lambda}^t \boldsymbol{X} - \boldsymbol{x}_0^t)\boldsymbol{l}$
  - **Estimation** : $\widehat{Z}(s_0) = (\gamma_0 + \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^t \boldsymbol{\Gamma}^{-1} \gamma_0))^t \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}$
  - **Calcul de la variance** : $\sigma^2_{KU}(s_0) = \gamma_0^t \boldsymbol{\Gamma}^{-1} \gamma_0 - (\boldsymbol{x}_0 - \boldsymbol{X}^t \boldsymbol{\Gamma}^{-1} \gamma_0)^t (\boldsymbol{X}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{x}_0 - \boldsymbol{X}^t \boldsymbol{\Gamma}^{-1} \gamma_0)$
- le KI (Krigeage Intrinsèque Généralisé) :

- **Modèle** : classe d'équivalence Fonction Aléatoire Intrinsèque d'ordre k (FAI-k). Si $Y(s)$ est un élément quelconque de cette classe, alors tout autre élément $Z(s)$ sera de la forme : $Z(s) = Y(s) + \sum_l A_l f_l(s)$
- L'estimateur optimal s'obtient en résolvant le même système que dans le cas du KU, avec le seul changement consistant à remplacer le variogramme $\gamma(h)$ par la CG $K(h)$

2.8 Cartographie des polluants atmosphériques en Île-de-France

Dans le cadre général de la surveillance de la qualité de l'air, la cartographie de la pollution atmosphérique répond à des exigences réglementaires et à un besoin d'information de la population vis-à-vis de la répartition spatiale de la pollution. Une analyse de cette répartition, corrélée avec d'autres données de la région (émissions, topographie, mode d'occupation du sol etc.) est essentielle pour une meilleure compréhension de la dispersion des polluants atmosphériques en zone urbaine.

L'objectif principal de cette section est d'appliquer la méthode d'interpolation décrite antérieurement, le krigeage, sur des jeux de données réelles : des mesures effectuées par le réseau de surveillance AIRPARIF (données publiques), représentant des concentrations de dioxyde d'azote et d'ozone. Le but est d'obtenir des cartes d'isoconcentrations de polluants. Ces cartes seront comparées par la suite (chapitre suivant) à celles obtenues par une modélisation spatio-temporelle à partir des séries temporelles fournies par les mêmes stations de mesure que dans le cas purement spatial. La question à laquelle on veut donner une réponse est la suivante : peut-on tracer une carte correcte d'isoconcentrations de polluants atmosphériques en utilisant uniquement les mesures disponibles à un moment donné sur le domaine d'intérêt ? Le manque de données spatiales est-il tellement important que les cartes produites donnent une image très éloignée de la réalité de la pollution atmosphérique ? Est-ce que les mesures enregistrées sont elles-mêmes suffisantes, ou il est vraiment nécessaire de prendre en compte tous les phénomènes physico-chimiques qui conduisent à la production des certains polluants atmospériques ?

2.8.1 Méthodologie géostatistique-généralités

La mise en œuvre du krigeage s'effectue en plusieurs étapes. La première est *l'analyse exploratoire* qui permet d'étudier les données, d'identifier leur structure et de prendre les bonnes décisions en ce qui concerne leur modélisation. *La formulation du modèle* est la deuxième étape ; elle consiste d'abord à faire un choix sur la forme de la tendance déterministe (le trend) de la Variable Régionalisée étudiée, et d'effectuer ensuite *l'analyse variographique*. L'étape suivante est *l'estimation* proprement-dite et le calcul de la variance d'erreur de l'estimation. On recourt enfin à une *validation croisée* ; la validation permet de comparer les performances des différents modèles pour sélectionner celui qui conduit aux meilleures prévisions. Chacune de ces étapes est détaillée dans ce chapitre sur des exemples concrets.

2.8.1.1 Objectifs de l'analyse exploratoire

Le but d'une étude exploratoire est d'identifier la distribution spatiale des données, d'apprécier leur degré d'homogénéité, de rechercher et de visualiser les observations atypiques. Pour cela, le géostatisticien a à sa disposition plusieurs outils, parmi lesquels, les plus souvent utilisés sont les histogrammes et les nuages de valeurs.

Comme l'analyse exploratoire vise à donner une idée de la distribution spatiale des données, des statistiques descriptives telles que la moyenne, la médiane, les quantiles et l'histogramme peuvent être utilisés pour préciser cette distribution. Si cette distribution présente un important écart à la normalité (elle est loin d'une distribution gaussienne), une transformation de données peut être envisagée. Dans ce cas, le reste de l'analyse devrait être effectuée sur les données transformées. Trouver une transformation appropriée, dont le résultat s'approche vraiment d'une gaussienne, n'est pas une tâche facile, mais ce fait peut être accompli en utilisant un technique comme celle proposée par Box et Cox (1964). Dans ce cas il faut s'assurer d'être en mesure d'effectuer la transformation inverse après le krigeage.

Pour une description spatiale, on peut visualiser la répartition des valeurs enregistrées, en prenant comme axes les deux coordonnées spatiales x et y et en représentant les données par des cercles proportionnels aux valeurs mesurées. Cet outil graphique aide notamment à juger si l'espérance de la fonction aléatoire qui modélise la variable régionalisée peut être vue comme constante. La présence d'une variabilité très grande indiquerait l'absence d'une moyenne constante, et donc peut-être l'existence d'une dérive qu'il faut correctement identifier. Les caractéristiques d'une telle dérive peuvent être illustrées par les graphiques de la variable régionalisée en fonction des coordonnées. On représente les deux graphiques, l'un en fonction de la coordonnée x et l'autre par rapport à y, et on analyse les résultats ; si les graphiques montrent autre chose qu'une tendance linéaire de pente nulle on a affaire à une espérance de la fonction aléatoire qui n'est certainement pas stationnaire.

2.8.1.2 L'analyse variographique-démarche utilisée

Pour effectuer une interpolation spatiale par krigeage, l'analyse variographique est indispensable. L'utilisateur doit d'abord choisir l'estimateur qu'il va utiliser. On a toujours le choix entre un estimateur **classique** et un estimateur plus **robuste** (Cressie, 1993), qui diminuerait les effets d'une valeur atypique.

Ensuite, il faut choisir la **distance maximale** d'estimation du variogramme et la largeur de la fenêtre dans le cas des données réparties irrégulièrement. On rappelle que selon Journel et Huijbregts (1978) ou Cressie (1993), la distance maximale d'estimation ne doit jamais dépasser la moitié de la distance maximale entre les stations de mesure.

De plus, le dernier paramètre du variogramme, qui est **le nombre des classes de distances**, doit être choisi en faisant un compromis entre la quantité et la qualité des points composant le semi-variogramme expérimental. Il faut s'assurer que chaque point provienne d'un nombre suffisant de données. Lorsque le nombre total de données est très faible, le calcul d'un variogramme expérimental directionnel est impossible, car le nombre de représentants pour chaque classe de distance est

insignifiant.

Ainsi, dans notre cas d'étude, où il y a peu de stations de mesure, on a plutôt opté pour un variogramme expérimental **omnidirectionnel** (isotrope) calculé pour **la moitié de la distance maximale** entre les points de mesure et avec une **tolérance fixée à la moitié du pas** choisi pour le calcul des points du variogramme expérimental (voir section 2.5.3).

La **validation croisée** de type "leave-one-out" est fréquemment utilisée dans la géostatistique afin de comparer la qualité des prévisions provenant de divers modèles (Arnaud et Emery, 2000; Cressie, 1993). Elle consiste à retirer une à une les observations, pour ensuite les prévoir par krigeage, en utilisant les autres restantes. Cette méthode peut être utilisée non seulement pour comparer les résultats, mais aussi pour choisir parmi les modèles disponibles, c'est-à-dire choisir d'une part le modèle de variogramme et d'autre part, le type de krigeage à utiliser. On calcule les erreurs de prévision et, à partir de ces erreurs, des indices sont calculés comme par exemple la moyenne des carrés des erreurs, EQM (Erreur Quadratique Moyenne). D'autres indices qui peuvent être utilisés sont : la moyenne ou la médiane des erreurs en valeur absolue, la dernière étant favorisée en cas de présence des valeurs extrêmes. Ainsi, en appliquant plusieurs fois la validation croisée pour chaque modèle employé, ceux pour lesquels on obtient les moindres EQM sont jugés être les meilleurs.

Plusieurs auteurs dénoncent l'utilisation de la validation croisée pour sélectionner un modèle (Cressie, 1993; Isaaks et Srivastava, 1990). Dans leur opinion il s'agit d'un modèle de type "boîte noire" qu'il faut éviter, car le rôle de l'utilisateur est ainsi diminué. Néanmoins, cette technique possède un caractère automatique qui est parfois nécessaire, comme dans le cas de la théorie des FAI-k de (Matheron, 1973). On rappelle ici que le modèle qui est finalement choisi en utilisant la validation croisée est le meilleur parmi les modèles envisagés et selon l'indice calculé, mais il n'est pas, en tout cas, optimal (le meilleur parmi tous les modèles possibles) (Chilès et Delfiner, 1999). Enfin, pour des grands jeux de données, la sélection du modèle par validation croisée n'est pas envisageable. D'ailleurs, elle n'est même pas nécessaire, car le jeu initial de données peut être divisé en deux : une partie utilisée pour l'interpolation et l'autre pour le test. La variable régionalisée est interpolée aux sites d'observation de test à partir du premier groupe de données et on procède ensuite au calcul des erreurs de prévision et d'autres indices sur l'ensemble de test.

2.8.2 Présentation des données et de la zone d'étude

2.8.2.1 Choix des polluants analysés

Dans cette étude, on a choisi deux polluants atmosphériques, très présents dans l'air qui surplombe la région d'Île-de-France : le dioxyde d'azote (NO_2) et l'ozone (O_3). L'étude des autres polluants présentés dans le premier chapitre soit n'est plus justifiée par leur importance au niveau de l'impact (c'est le cas du SO_2 qui a beaucoup diminué ces dernières années), soit le nombre de stations de mesure est trop faible pour réaliser une cartographie à partir de leurs observations uniquement.

2.8.2.2 Zone d'étude

La zone d'étude n'est pas la même pour le dioxyde d'azote et pour l'ozone.

- **Zone d'étude du dioxyde d'azote**

Le choix de la zone d'étude a été en grande partie dicté par l'emplacement des stations de mesure.

Si on regarde la densité moyenne annuelle des émissions de NO_x, on retrouve bien des niveaux élevés dans Paris intra-muros, ensuite des valeurs aussi élevées, mais sur des zones plus restreintes, en proche banlieue, qui diminuent de façon nette en grande couronne.

Compte tenu de la durée de vie plutôt courte du dioxyde d'azote (de quelques secondes à quelques heures), la zone susceptible d'être fortement polluée par le NO_2 reste à l'intérieur de la petite couronne. C'est la raison pour laquelle les stations de mesure d'Airparif qui enregistrent le NO_2 se trouvent exclusivement dans Paris et sa proche banlieue (exceptées 2 stations rurales : Fontainebleau et Rambouillet).

Selon les fluctuations des émissions et les situations météorologiques, la distribution spatiale de la concentration de NO_2 pour cette même région sera différente. La zone d'étude pour le NO_2, présentée dans la figure 2.10, a été légèrement réduite par rapport à la petite couronne (figure 1.4) pour éviter les régions très peu couvertes de capteurs, où l'estimation correspondrait à une extrapolation.

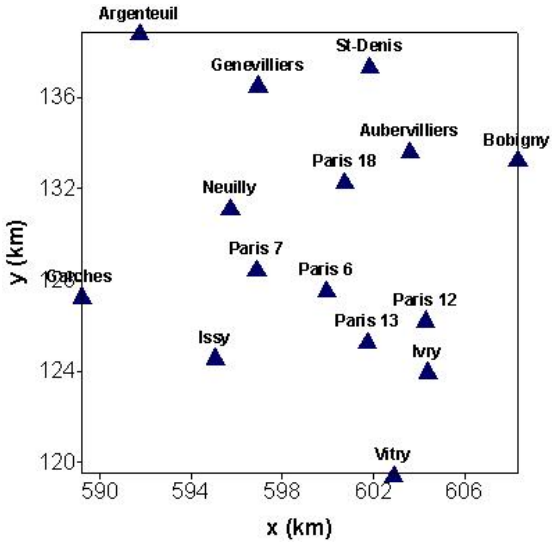


FIG. 2.10: Les stations mesurant le dioxyde d'azote utilisées dans cette étude.

- **Zone d'étude de l'ozone**

Le dioxyde d'azote est l'un des précurseurs de l'ozone. Les deux polluants sont donc en partie complémentaires. Par rapport au NO_2, dont la durée de vie est relativement courte, les panaches d'ozone peuvent se déplacer sur plusieurs dizaines de kilomètres. Sans généraliser, les forts niveaux de NO_2 sont rencontrés plus souvent en agglomération urbaine, tandis que ceux d'ozone le sont

majoritairement en zone rurale, voire grande banlieue.

La zone d'étude de l'ozone est plus vaste et englobe aussi la grande couronne (voir la figure 2.11)

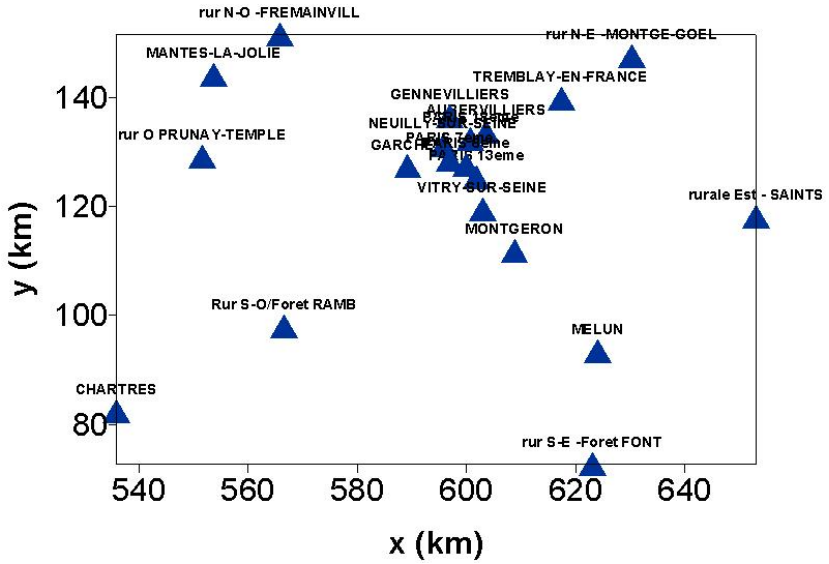


FIG. 2.11: Les stations AIRPARIF mesurant l'ozone utilisées dans cette étude.

Les coordonnées géographiques utilisées sont de type coordonnées Lambert I Nord (voir la section 1.7.3).

2.8.2.3 Les cas d'étude choisis

Les cas les plus intéressants à étudier sont les situations de pic, quand, suite aux conditions météorologiques qui leur sont favorables (situation anticyclonique, absence de vent), les concentrations de polluants atmosphériques grimpent, atteignant des valeurs supérieures aux limites admises par la législation existante (voir la section 1.6). Les pics de dioxyde d'azote apparaissent souvent le matin, lorsque la couche limite n'est pas encore totalement développée. Pour l'ozone, les pics se produisent surtout aux heures les plus chaudes de la journée (14 ou 15 heures).

Plus précisément, dans ce travail on présente d'abord en détail l'étude d'un pic de pollution pour chacun des polluants : le 29 juillet 1999 à 8 heures pour le dioxyde d'azote, et le 30 juillet 1999 à 14 heures pour l'ozone. On se propose d'analyser la répartition spatiale de ces données mesurées et ensuite d'obtenir des cartes d'isoconcentrations et de variance de l'erreur d'estimation. L'étude d'un autre jour de pic sera ensuite présentée pour comparer l'efficacité du krigeage appliqué sur des épisodes de forte et de faible pollution (pendant le même jour) :

- **pour le** NO_2 : le 17 Juillet 1999 à 8 heures (forte pollution) et à 15 heures (faible pollution) ;
- **pour l'ozone** : le 17 Juillet 1999 à 6 heures (faible pollution) et à 15 heures (forte pollution).

2.8.2.4 Les statistiques descriptives des données étudiées

• Le dioxyde d'azote

On présente d'abord dans le tableau 2.1 les statistiques descriptives (spatiales) pour les mesures de dioxyde d'azote enregistrées par les stations automatiques d'AIRPARIF pour les cas choisis, en s'intéressant principalement à la date du **29 juillet 1999 à 8 heures** du matin (la dernière colonne du tableau), dont l'analyse sera présentée plus en détail ; pour les autres dates, uniquement les résultats principaux seront donnés.

Rappelons les deux définitions des quartiles : Q_1 est le premier quartile si au moins 25% des individus prennent une valeur inférieure ou égale à Q_1 et au moins 75% des individus prennent une valeur supérieure ou égale à Q_1 ; Q_3 est le troisième quartile si au moins 75% des individus prennent une valeur inférieure ou égale à Q_3 et au moins 25% des individus prennent une valeur supérieure ou égale à Q_3.

Statistiques (μg.m^{-3})	17/07/1999 8h	17/07/1999 15h	29/07/1999 8h
Minimum	41,00	11,00	58,00
Q1	58,00	16,50	67,25
Moyenne	73,00	21,93	96,64
Médiane	66,00	22,00	97,00
Q3	83,50	26,00	115,25
Maximum	124,00	44,00	140,00
Ecart-type	21,64	8,25	29,92

TAB. 2.1: Statistiques descriptives pour le NO_2 (pour les cas d'étude choisis).

En analysant le tableau 2.1, on remarque une grande valeur de l'écart-type, presqu'un tiers de la moyenne, excepté pour les données enregistrées le 17 Juillet 1999 à 15 heures ; l'après-midi les valeurs de dioxyde d'azote sont faibles, entre 11 et 44 μg/m^3 et par conséquent, leur variabilité réduite. En revanche, les données enregistrées le matin exhibent une variabilité assez importante qui peut être un facteur favorable dans le processus d'estimation.

• L'ozone

Dans le tableau 2.2 on présente les mêmes statistiques que précédemment, pour l'ozone. La principale remarque que l'on peut faire est que la dispersion des données est assez faible, surtout pour les données enregistrées le 30 Juillet 1999 à 14 heures (dernière colonne du tableau), la date qui nous intéresse le plus dans le cas de ce polluant, car elle correspond à un pic important d'ozone, et elle sera étudiée par d'autres méthodes d'analyse.

Statistiques ($\mu g.m^{-3}$)	17/07/1999 6h	17/07/1999 15h	30/07/1999 14h
Minimum	6,00	115,0	162,00
Q1	17,00	126,50	176,75
Moyenne	28,73	152,36	189,55
Médiane	29,00	140,00	190,50
Q3	38,50	159,50	202,50
Maximum	74,00	255,00	226,00
Ecart-type	17,30	37,38	18,22

TAB. 2.2: Statistiques descriptives pour le O_3 (pour les cas d'étude choisis).

2.8.3 Représentation des champs de concentration du dioxyde d'azote sur la région d'Île-de-France

On présente plusieurs cas d'étude pour le dioxyde d'azote : le 29 Juillet 1999 à 8 heures (analyse détaillée) et le 17 Juillet 1999 à 8 heures et à 15 heures.

2.8.3.1 Champs de NO_2 le 29 Juillet 1999 à 8 heures

On commence notre analyse avec les données enregistrées le 29 Juillet 1999 à 8 heures. On rappelle que l'heure est exprimée en temps civil. Dans la figure 2.12 on peut retrouver les stations automatiques appartenant à AIRPARIF qui mesurent le dioxyde d'azote et qui ont été utilisées dans cette étude.

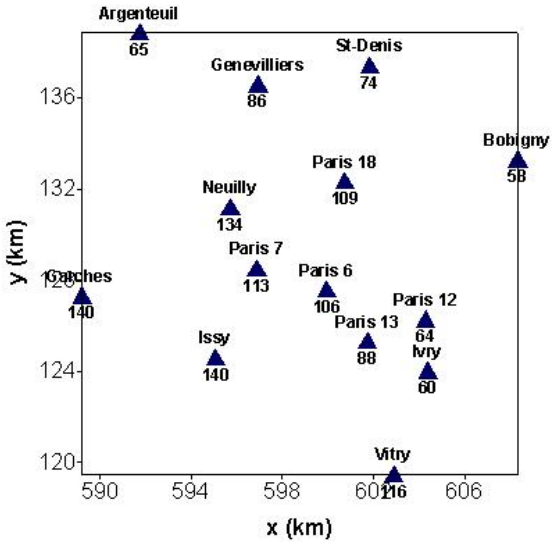
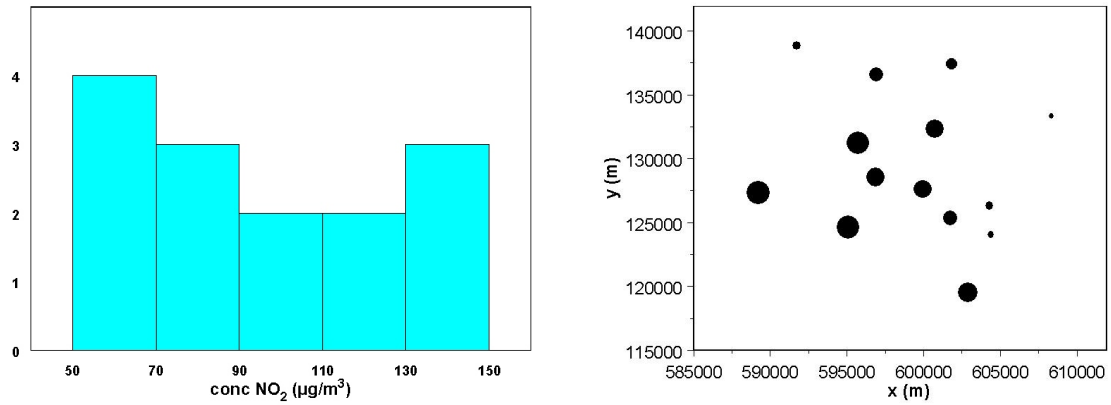


FIG. 2.12: Les mesures de NO_2 enregistrées le **29 Juillet 1999 à 8 heures** par les stations d'AIRPARIF.

Analyse statistique et exploratoire

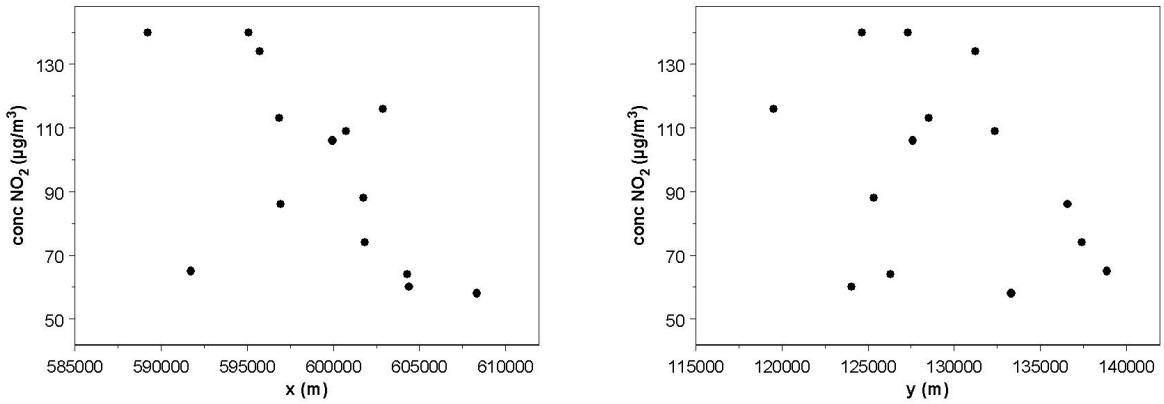
Dans la figure 2.13(a) on retrouve l'histogramme des mesures étudiées. Ce premier jeu de données présente une distribution asymétrique avec des fréquences plus élevées d'une part pour les valeurs faibles et d'autre part pour les valeurs fortes. La normalité peut être sérieusement mise en doute ; donc une transformation de données peut être envisagée.



(a) Histogramme de NO_2. (b) Mesures de NO_2.

FIG. 2.13: Distribution spatiale des données de NO_2 enregistrées le 29 Juillet 1999 à 8 heures.

Dans la figure 2.13(b) on représente les données enregistrées par les stations de mesure, la taille des symboles utilisés pour localiser les stations sur la carte étant proportionnelle aux mesures enregistrées. Cette figure fait ressortir une variabilité assez importante qui nous fait penser qu'on ne peut pas considérer l'espérance de la fonction aléatoire constante. Il faut donc identifier la forme de cette dérive.



(a) Concentrations de NO_2 fonction de x. (b) Concentrations de NO_2 fonction de y.

FIG. 2.14: Concentrations de NO_2 enregistrées le 29 Juillet 1999 à 8 heures comme fonction de coordonnées spatiales.

Pour cela, on a décidé de représenter les concentrations de NO_2 comme fonctions de leurs coordonnées spatiales $2D$, x et y. Dans le cas du dioxyde d'azote (voir figures 2.14(a), 2.14(b)), on constate une forte dépendance décroissante linéaire ou peut-être quadratique en fonction de x. Par contre, par rapport à y, il n'y a pas de tendance nette qui se dégage. Dans ce cas, l'application du krigeage universel avec une tendance dépendant de la première coordonnée spatiale, x, est donc envisageable.

En regardant les figures déjà analysées, il est difficile de tirer des conclusions concernant la stationnarité de la Variable Régionalisée à interpoler. De plus, le nombre réduit de stations rend inutile l'étude de la stationnarité en employant une fenêtre mobile. Par conséquent, on va considerer la variable aléatoire (la concentration de NO_2) comme intrinsèque, et on va appliquer deux types de krigeage : ordinaire et universel. Ensuite, on va comparer les résultats avec ceux obtenus en utilisant la théorie de Matheron de FAI-k décrite dans la section 2.6.6.

Analyse variographique

La nuée variographique de NO_2 présentée sur la figure 2.15, montre qu'il existe au moins quatre couples de stations de mesure qui paraissent atypiques. Cela correspond aux couples de stations Issy-Ivry, Issy-Paris12, Neuilly-Paris12 et Neuilly-Argenteuil, c'est-à-dire les paires qui opposent les valeurs les plus fortes, enregistrées à l'Ouest du domaine d'étude, aux valeurs les plus faibles, caractéristiques pour la zone Est du domaine. Dans ce genre de situations, d'habitude, il est conseillé d'enlever les mesures atypiques. Dans ce cas précis, enlever ces mesures n'est pas recommandable, car le contraste entre la partie Ouest et Est est bien défini et, en plus, le nombre de stations disponibles est trop petit pour pouvoir se permettre une telle approche. On décide, donc, de garder toutes les données initiales. De plus, aucune transformation appliquée n'a conduit à une distribution plus symétrique de données. On n'a pas d'autre choix que de travailler sur les données originelles.

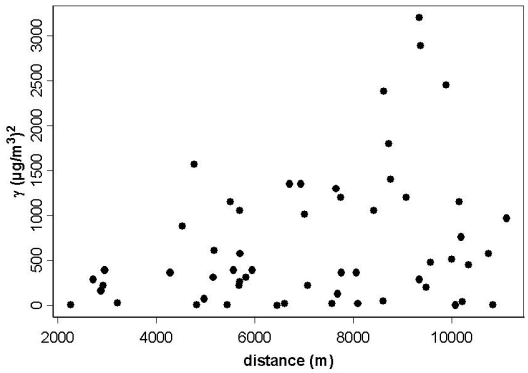


FIG. 2.15: Nuée variographique de NO_2.

• **Variogramme obtenu à partir des données brutes**

Les paramètres du modèle théorique ajusté au variogramme expérimental ont été estimés en fonction de la distribution spatiale de données **brutes**. Pour une distance maximale de 12 km et 5

classes de distance, on a obtenu un pas de 2 400 m (avec une tolérance de 1 200 m) ; en utilisant la méthode des moindres carrés pondérés (les poids étant proportionnels au nombre de paires de points situées dans chaque classe de distance), on ajuste au variogramme expérimental un modèle gaussien (voir la section 2.5.4) avec les caractéristiques suivantes : une portée de 10 km, un palier de 1 676 $(\mu g.m^{-3})^2$ et une valeur pépitique de 65 $(\mu g.m^{-3})^2$ (voir la figure 2.16).

On signale, par ailleurs, que les autres essais pour trouver un ajustement correct n'ont pas

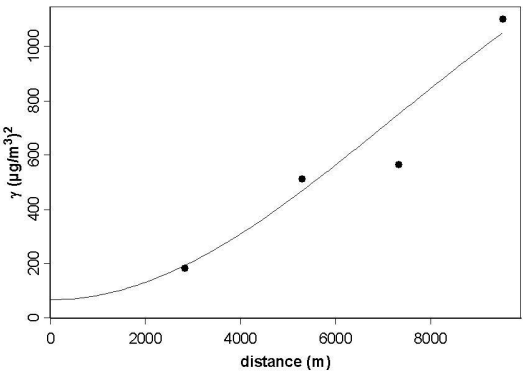


FIG. 2.16: Variogramme expérimental sur les données brutes de NO_2, **le 29 Juillet 1999, à 8 heures**, avec l'ajustement d'un modèle gaussien.

abouti, faute d'une interprétation physique cohérente. On remarque également que, même pour cet ajustement, le palier obtenu est très grand.

- **Variogramme obtenu à partir des résidus**

Pour la technique de krigeage universel (KU), le variogramme utilisé doit être obtenu à partir des résidus, ce qui correspond aux données brutes auxquelles on a enlevé la tendance. Sur les données brutes, on a effectué d'abord une régression qui ne tient pas compte de la structure spatiale des données (voir la section 2.6.4) pour identifier **une tendance linéaire** par rapport aux coordonnées spatiales x et y. Ensuite, à partir des résidus, on a déterminé un variogramme expérimental, auquel on a ajusté un modèle gaussien, dont les caractéristiques sont : une portée de 10,5 km, un palier de 779 $(\mu g.m^{-3})^2$ et une valeur pépitique de 78 $(\mu g.m^{-3})^2$ (voir la figure 2.17). On remarque que le palier a été vraiment diminué en éliminant la tendance. En revanche, l'allure générale du variogramme reste la même. Par rapport à la démarche expliquée dans la section 2.6.4, on a décidé de faire une seule itération, car Kitanidis (1987) et Hengl et al. (2003) soutiennent qu'une seule itération est souvent suffisante. Le krigeage universel a donc été appliqué sur les résidus des moindres carrés ordinaires.

- **Covariance généralisée $K(h)$ à partir des données brutes**

Comme le nombre de stations disponibles est assez réduit, (on rappelle que pour la date analysée on dispose de 14 mesures de NO_2), notre choix s'est porté sur un voisinage global (incluant

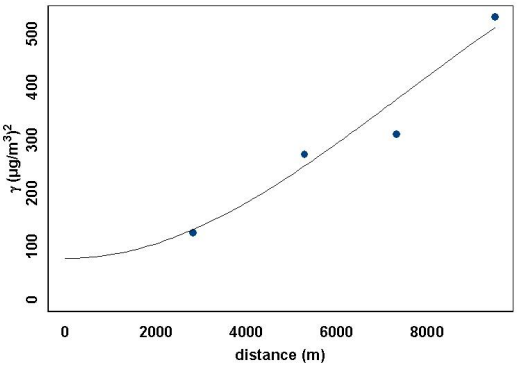


FIG. 2.17: Variogramme expérimental sur les résidus obtenus après une régression de moindres carrés ordinaires de NO_2, le **29 Juillet 1999 à 8 heures**, avec l'ajustement d'un modèle gaussien.

toutes les mesures disponibles) pour essayer d'identifier, par la procédure automatique mentionnée dans la section 2.6.6, un ordre de continuité et une covariance généralisée de type polynomial, qui seront ensuite utilisés pour résoudre le système de krigeage et donner l'estimation sur l'ensemble du domaine. Dans ce cas précis, l'ordre identifié a été $\nu = 1$ et la covariance polynomiale est de type linéaire.

2.8.3.2 KO, KU et KI appliqués sur les données de NO_2 le 29 Juillet 1999 à 8 heures

Une fois le variogramme ajusté et la covariance généralisée estimée, on peut maintenant appliquer une méthode d'interpolation. On appliquera les trois variantes de krigeage : ordinaire sur les données brutes, universel sur les résidus des moindres carrés ordinaires, et intrinsèque généralisé pour le même jeu de données brutes. Les performances des méthodes seront évaluées par des tests de validation croisée qui ont été appliqués pour chaque méthode.

Tout d'abord, on remarque le fait que les cartes obtenues en appliquant les trois méthodes de krigeage sont très semblables (voir les figures 2.18(a), 2.18(c), et 2.18(e)), avec un niveau de pollution qui décroît du Sud-Ouest vers le Nord-Est. On peut remarquer, également, que pour les deux premières méthodes, trois stations qui ont enregistré des concentrations élevées : Neuilly avec 134 $\mu g/m^3$, Paris 7 avec 113 $\mu g/m^3$ et Issy avec 140 $\mu g/m^3$ n'ont pas été bien "classées" par rapport aux isolignes. Ce n'est pas le cas pour la méthode du krigeage intrinsèque généralisé, où toutes les mesures sont classées entre les bonnes isolignes. De plus, pour cette dernière méthode, on peut remarquer aussi les valeurs un peu plus faibles pour l'écart-type de l'erreur d'estimation.

Obtenir des valeurs "mal classées" par rapport aux lignes d'isoconcentration n'est pas en désaccord avec le fait que le krigeage est un interpolateur exact. L'explication réside dans la discontinuité de l'estimateur basé sur un variogramme discontinu. En effet, le variogramme utilisé est discontinu à l'origine avec un effet de pépite qui influence la continuité du champ estimé (voir la section 2.5.2.2).

L'effet de pépite identifié dans les deux variogrammes expérimentaux influencent :

- le champ estimé : les valeurs mesurées sont mal classées entre les isolignes dans la limite de la valeur pépitique ;
- la carte de variance de l'estimation : dans le cas du KI le champ estimé est continu au voisinage des mesures et la variance de l'estimation est nulle aux stations de mesure ; lors du KO ou KU la variance de l'estimation ne descend pas en-dessous de la valeur pépitique. Ceci fait que la carte de variance de l'estimation obtenue pour le KO et KU présente une zone uniforme correspondant à la valeur pépitique, tandis que celle du KI présente plus de variabilité autour des points de mesure (isolignes concentriques).

Les effets de bords sont présents dans tous les résultats et ils sont comparables comme ordre de grandeur pour les trois méthodes.

Les différences entre les champs estimés sont essentiellement dans la partie Sud-Ouest (la variance d'estimation est maximale aussi) ; alors qu'avec le KO on ne dépasse pas de beaucoup la valeur maximale mesurée ($140\,\mu g/m^3$), les deux autres méthodes (KU, KI) gardent une tendance croissante vers l'extérieur du domaine, conduisant à des valeurs estimées plus importantes.

Si on regarde les résultats de la validation croisée (tableau 2.3), on ne peut pas dire qu'il existe une méthode, parmi celles appliquées, qui ressort comme la "meilleure". C'est d'ailleurs le cas si on regarde le tableau avec les statistiques globales sur les tests "Leave-One-Out" (tableau 2.4). On peut constater que parmi les 14 stations disponibles au moment de l'analyse, il existe au moins 4 stations pour lesquelles les tests "Leave-One-Out" donnent des mauvais résultats, c'est-à-dire une sous-estimation ou une sur-estimation au site de mesure de plus de $15\,\mu g.m^{-3}$. C'est le cas des stations Argenteuil et Vitry qui sont situées aux bords du domaine, mais aussi de deux autres, situées à l'intérieur du domaine : Neuilly et Paris 7, proches l'une de l'autre, mais enregistrant des mesures assez différentes. Pour les deux premières stations, la mauvaise estimation peut être mise sur le compte du manque du voisinage, les deux étant assez éloignées de leurs voisine ; en particulier, le site de Vitry peut être influencé par une pollution locale, provenant de la grande zone industrielle aux alentours. Pour les deux dernières, Neuilly et Paris 7 cela peut s'expliquer par le fait que les deux stations forment un couple qui marque le bord du panache estimé au Sud-Ouest du domaine ; si on retire l'une de deux, alors l'estimation faite en gardant l'autre mesure sera soit une sur-estimation dans le cas de Paris 7, qui enregistre une concentration inférieure, et donc le bord du panache sera poussé vers le centre du domaine, soit une sous-estimation dans le cas de Neuilly, quand le panache est estimé plus près de l'extrémité. Néanmoins, les quatre stations sont assez représentatives sur l'ensemble de 14 donc, leur présence est vraiment nécessaire pour avoir une image plus précise sur l'épisode de pollution analysé.

Les tests de validation croisée indiquent une erreur d'estimation de l'ordre de $10\,\mu g/m^3$ en moyenne pour les stations à l'intérieur du domaine, et plus importante aux bords ($30 - 40\,\mu g/m^3$).

Station	Valeur mesurée ($\mu g.m^{-3}$)	**KO** ($\mu g.m^{-3}$)	**KU** ($\mu g.m^{-3}$)	**KI** ($\mu g.m^{-3}$)
ARGENTEUIL	65,00	96,45	**82,90**	101,83
BOBIGNY	58,00	97,93	42,90	**47,24**
GARCHES	140,00	112,21	152,74	**148,47**
GENEVILLIERS	86,00	102,30	**92,69**	103,63
ISSY	140,00	97,12	**132,36**	129,16
IVRY	60,00	**69,26**	76,46	78,36
NEUILLY	134,00	104,04	**113,95**	110,10
PARIS 12	64,00	**65,42**	69,38	69,39
PARIS 13	88,00	89,31	**87,57**	**87,80**
PARIS 18	109,00	**106,90**	95,24	93,40
PARIS 6	106,00	98,17	**102,19**	101,37
PARIS 7	113,00	130,60	**128,32**	**128,30**
ST-DENIS	74,00	84,99	**75,59**	70,62
VITRY	116,00	84,05	82,75	**89,82**

TAB. 2.3: Validation croisée pour le NO_2 (**29/07/1999 8h**) pour les trois méthodes d'interpolation appliquées. Les meilleures estimations ont été mises en gras.

Méthode	**MIN** ($\mu g.m^{-3}$)	**MAX** ($\mu g.m^{-3}$)	**MAE** ($\mu g.m^{-3}$)	**RMSE** ($\mu g.m^{-3}$)
KO	1,31	42,88	19,34	17,47
KU	0,43	33,25	12,15	14,79
KI	0,20	36,83	14,10	17,12

TAB. 2.4: Statistiques globales pour les tests "Leave-One-Out" obtenus pour les trois types de krigeage spatial appliqués pour le NO_2 (**29/07/1999 8h**).

2.8.3.3 Champs de NO_2 le 17 Juillet 1999 à 8 heures et à 15 heures

On continue l'analyse des champs de concentrations de dioxyde d'azote obtenus en utilisant une méthode d'interpolation spatiale avec deux cas : valeurs maximales enregistrées à 8 heures du matin et valeurs minimales à 15 heures de l'après-midi, le **17 Juillet 1999**.

On réduira l'analyse à la présentation du variogramme et des champs obtenus par interpolation. Tout d'abord un commentaire visant les statistiques descriptives de ces deux jeux de données (les deux premières colonnes du tableau 2.1) : en ce qui concerne les données mesurées à 8 heures, on observe une grande variabilité (minimum de $41\,\mu g.m^{-3}$, maximum de $124\,\mu g.m^{-3}$), tandis que pour les données enregistrées à 15 heures, la plage de valeurs est plus réduite et l'écart-type est moindre.

2.8.3.4 Analyse variographique - 17 Juillet 1999 à 8 heures

- **Variogramme obtenu à partir des données brutes**

Dans la figure 2.19(a) on présente le modèle gaussien qui a été ajusté au variogramme expérimental pour le jeu de données brutes. On remarque une bonne qualité de l'ajustement, avec les paramètres suivants : un palier de $524\,(\mu\text{g.m}^{-3})^2$, une valeur pépitique de $49\,(\mu\text{g.m}^{-3})^2$ et une portée de 5,6 km qui se traduira après par une représentation correcte du champ de concentrations de dioxyde d'azote en utilisant le krigeage ordinaire (voir la figure 2.21(a)).

Alors que l'effet de pépite est tout à fait comparable à celui obtenu dans le cas précédent, le palier et la portée sont sensiblement inférieurs, laissant présager une distribution spatiale plus "concentrée" autour du maximum, que dans le cas précédent.

- **Variogramme obtenu à partir des résidus**

Pour le krigeage universel (KU), on a besoin d'un variogramme ajusté sur les résidus. Pour cet exemple, on a effectué une itération complète pour arriver à une régression des moindres carrés généralisés. Pour cela, on ne s'est plus arrêté après une régression des moindres carrés ordinaires, comme dans le cas du 29 Juillet 1999. On a commencé par une régression des moindres carrés ordinaires appliquée aux données brutes pour identifier une tendance linéaire ; ensuite, sur les résidus obtenus on a ajusté un premier modèle de variogramme de type gaussien, présenté dans la figure 2.20(a). Ultérieurement, en utilisant la matrice $\mathbf{\Gamma}$ obtenue (voir les notations de la section 2.6), on procède à une régression des moindres carrés généralisés et, sur les nouveaux résidus, on ajuste un deuxième modèle de variogramme gaussien (voir la figure 2.20(b)), qui sera utilisé, enfin, dans le krigeage universel. Il faut, quand même, remarquer que les différences entre les paramètres de l'ajustement sur les résidus ordinaires et ceux généralisés sont insignifiantes, comme le montre d'ailleurs la figure 2.20. Donc, sur les données analysées, on pouvait effectuer le krigeage universel après une seule régression des moindres carrés ordinaires, sans que cela affecte l'estimation.

- **Covariance généralisée $K(h)$ à partir des données brutes**

La procédure automatique d'identification de l'ordre de continuité et de la covariance généralisée (CG) conduit à une covariance linéaire, donc le terme contenant le logarithme (voir la section 2.6.6) reçoit un coefficient nul. Le seul coefficient non-nul dans la formule 2.74 est celui du h.

2.8.3.5 KO, KU et KI appliqués sur les données de NO_2 le 17 Juillet 1999 à 8 heures

La carte présentée dans la figure 2.21(c) obtenue par KU est très semblable à celle obtenue par krigeage intrinsèque généralisé (figure 2.21(e)). Les différences par rapport au krigeage ordinaire (figure 2.21(a)) sont minimales. Si on compare ces cartes à celles obtenues le 29 Juillet 1999 à 8 heures, on peut remarquer que le panache est plus "dilué" pour le deuxième cas, ce que la différence de portée des deux variogrammes laissait entrevoir.

Par ailleurs, on présente chaque fois les cartes de l'écart-type de l'erreur d'estimation obtenues. Les différences entre les trois représentations spatiales sont très faibles, ainsi qu'entre les cartes de l'écart-type associées. On observe quand même des valeurs un peu plus faibles pour l'écart-type

de l'erreur estimé par krigeage intrinsèque généralisé.

Par rapport à l'effet de pépite, on peut faire les mêmes remarques que dans le cas précèdent concernant les valeurs "mal classées" et la typologie de la carte de variance de l'estimation. Si on compare encore une fois les deux exemples, on remarque que la typologie est différente : le 29 Juillet 1999 on avait une pollution décroissante Sud-Ouest Nord-Est, et le 17 Juillet 1999 à 8 heures, elle est décroissante Nord-Ouest Sud-Est.

2.8.3.6 Analyse variographique et krigeage appliqué sur les données de NO_2 le 17 Juillet 1999 à 15 heures

Si on regarde les valeurs enregistrées à 15 heures de l'après-midi (toujours le 17 Juillet 1999) la situation est un peu plus compliquée, car le modèle ajusté au variogramme expérimental calculé sur les données brutes, qui est le meilleur parmi ceux testés, présente une portée d'environ 27 km (voir la figure 2.19(b)) qui dépasse largement la distance maximale entre deux sites de mesure, ainsi que la zone d'étude ; par conséquent, les estimations faites en utilisant ce modèle sont très lisses (figures 2.22(a)) et ne correspondent pas à la situation réelle décrite par les valeurs enregistrées. Cette situation est similaire à l'identification d'une structure à une échelle différente, plus grande, ou elle correspond à un manque total de corrélation spatiale entre les mesures analysées ; cela peut être dû à une absence de variabilité dans les données. On remarque également que l'essai d'ajuster un modèle théorique de variogramme sur les résidus des moindres carrés ordinaires (plusieurs modèles de tendance ont été testés linéaires ainsi que quadratiques) n'ont pas abouti à des résultats cohérents. Pour le cas où on enregistre des valeurs faibles, il paraît difficile de capter une quelconque structure spatiale par l'intermédiaire d'un variogramme, car toutes les variances expérimentales obtenues sont de l'ordre de la pépite. Le variogramme obtenu, presque plat à l'échelle d'étude, conduit à un champ uniforme, sans structure spatiale bien définie.

Seule la méthode du krigeage intrinsèque généralisé produit une carte raisonnable exhibant une typologie de répartition spatiale légèrement différente de la précédente, avec deux petits pics : le premier autour de Neuilly (une cause possible étant le trafic), et un deuxième, moins important, autour d'Ivry (zone industrielle). On retrouve une certaine cohérence entre le champ estimé à 8 heures et celui de 15 heures, avec une diminution générale des niveaux à 15 heures, et légèrement plus prononcée pour Paris intra-muros (entre les deux pics). La carte de variance a la même allure pour 8 heures et 15 heures, avec des valeurs plus faibles pour le deuxième cas.

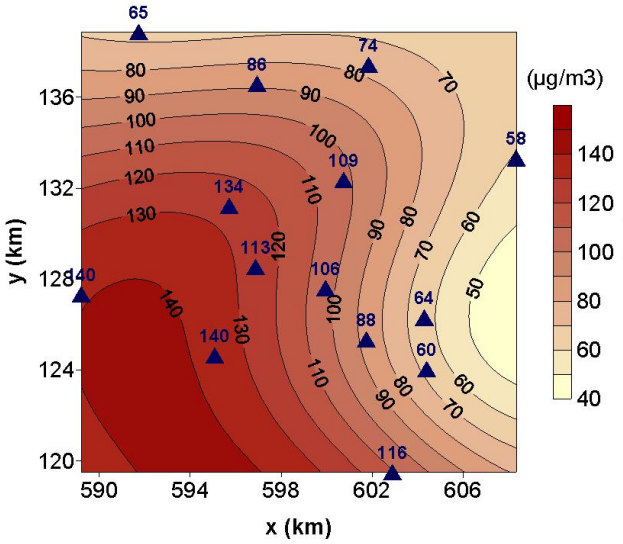
Récapitulatif sur l'analyse des données de NO_2

On présente un tableau récapitulatif (voir tableau 2.5) des paramètres obtenus pour chaque modèle ajusté aux variogrammes expérimentaux : la valeur pépitique, le palier et la portée, ainsi qu'un qualificatif pour les cartes obtenues en utilisant ces paramètres.

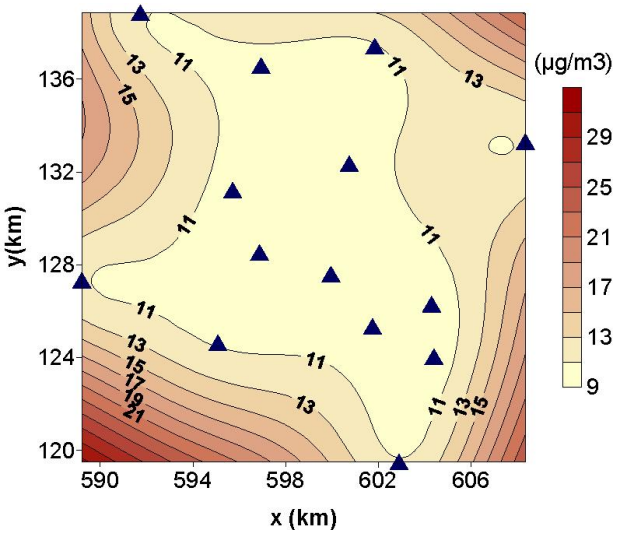
Date	**Type de données**	**Modèle**	**Valeur pépitique** $(\mu g.m^{-3})^2$	**Palier** $(\mu g.m^{-3})^2$	**Portée** (mètres)	**Qualificatif**
17/07/99 h8	**brutes**	Gaussien	49	525	5660	bonne
17/07/99 h8	**résidus-ols**	Gaussien	20	383	5465	bonne
17/07/99 h8	**résidus-gls**	Gaussien	23	385	5465	bonne
17/07/99 h15	**brutes**	Gaussien	50	346	27670	très mauvaise
29/07/99 h8	**brutes**	Gaussien	65	1676	10000	bonne
29/07/99 h8	**résidus-ols**	Gaussien	78	779	10500	bonne

TAB. 2.5: Tableau récapitulatif avec les paramètres des modèles ajustés aux variogrammes expérimentaux pour les cas étudiés. Les résidus-ols correspondent à ceux obtenus par les moindres carrés ordinaires, et les résidus-gls, par des moindres carrés généralisés.

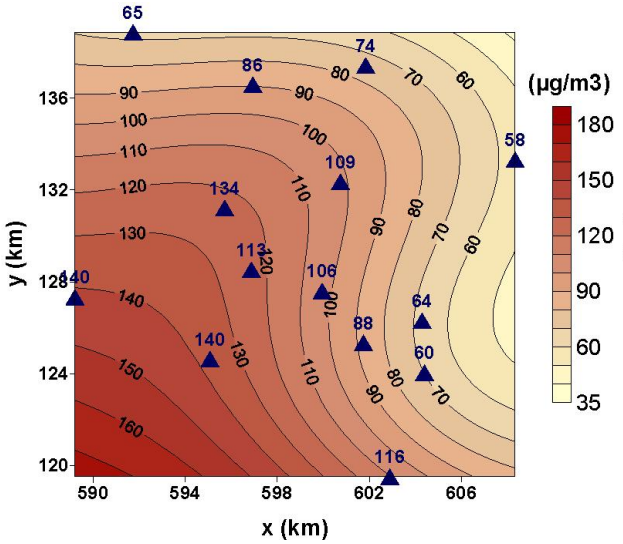
Pour conclure, sur ce polluant on a analysé trois cas différents (trois jeux de données) et pour chacun on a ajusté un modèle gaussien au variogramme expérimental calculé soit sur les données brutes, soit sur les résidus, ou bien on a utilisé la théorie des FAI-k et le krigeage intrinsèque généralisé. On constate que parfois il est très difficile de trouver un bon ajustement pour le variogramme expérimental et donc, si les paramètres obtenus ne sont pas cohérents avec la réalité physique (comme une portée qui dépasse la taille du domaine) alors la qualité de l'estimation est très mauvaise. Par contre, avec le krigeage intrinsèque la qualité de l'estimation est toujours correcte et les cartes obtenues sont cohérentes avec les valeurs enregistrées par les stations de mesure.



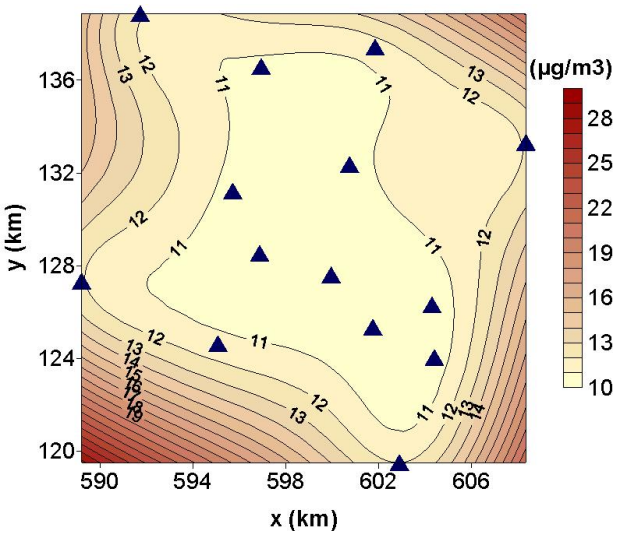
(a) Krigeage Ordinaire (KO).



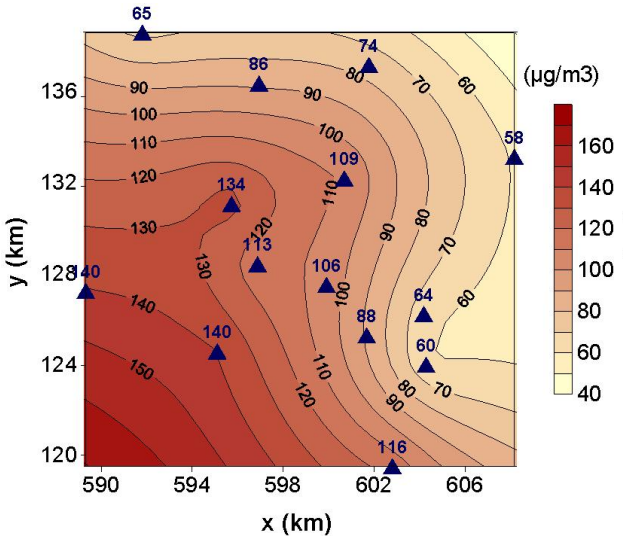
(b) Écart-type de l'erreur d'estimation par KO.



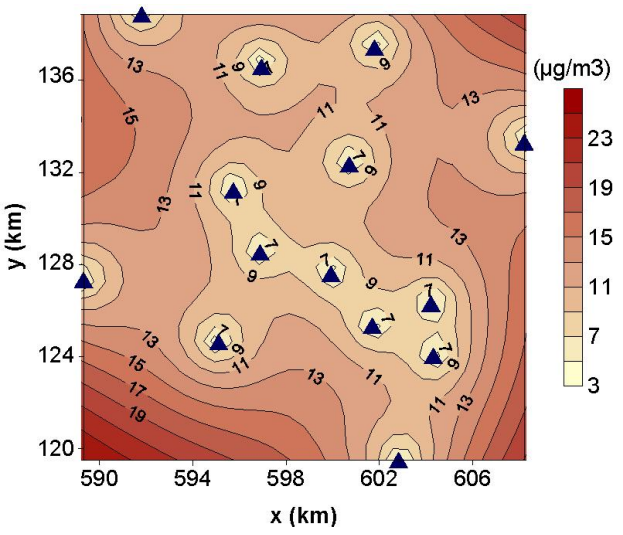
(c) Krigeage Universel (KU).



(d) Écart-type de l'erreur d'estimation par KU.

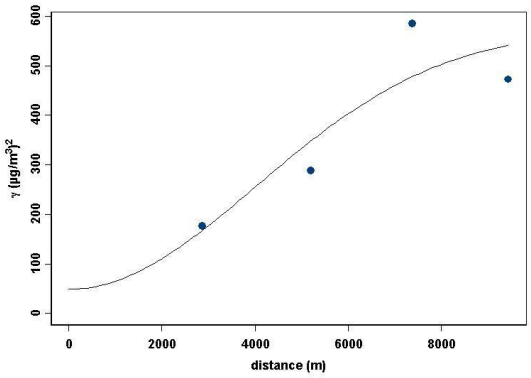


(e) Krigeage Intrinsèque Généralisé (KI).

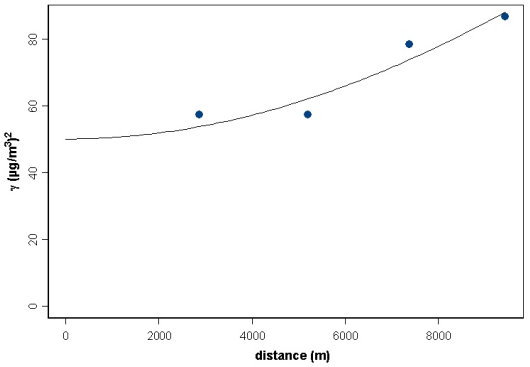


(f) Écart-type de l'erreur d'estimation par KI.

FIG. 2.18: Estimations des champs de concentrations de NO_2 le **29 Juillet 1999 à 8 heures** en appliquant les trois variantes de krigeage KO, KU et KI avec les cartes de l'écart-type de l'erreur associées. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

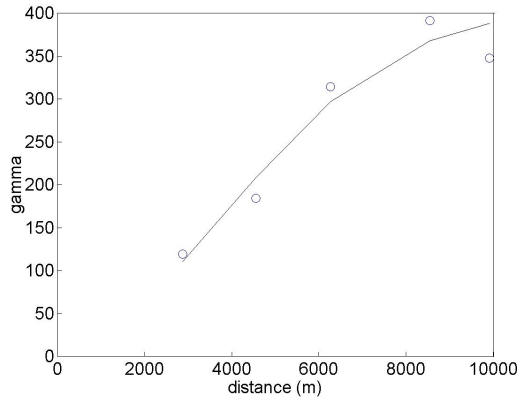


(a) Modèle gaussien ajusté au variogramme expérimental pour les données brutes de NO_2 du **17 Juillet 1999 à 8 heures** (forte pollution).

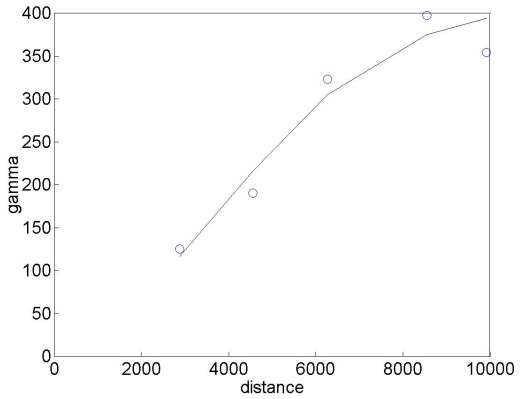


(b) Modèle gaussien ajusté au variogramme expérimental de NO_2 (données brutes) le **17 Juillet 1999 à 15 heures** (faible pollution).

FIG. 2.19: Les deux modèles gaussiens ajustés aux variogrammes expérimentaux pour le NO_2 mesuré le **17 Juillet 1999** à 8 heures et 15 heures.

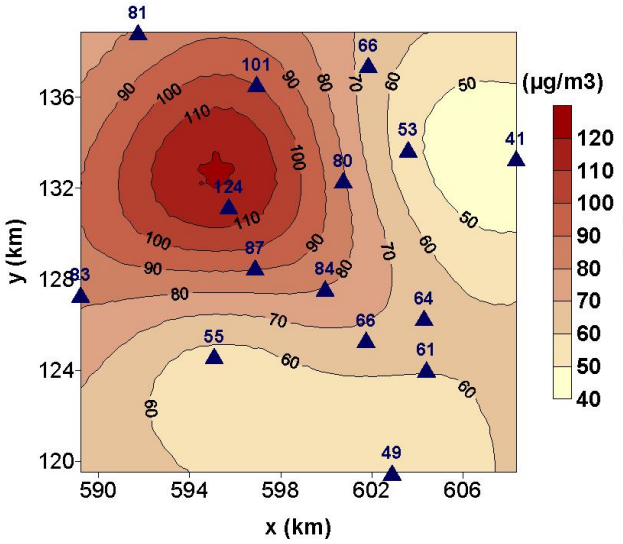


(a) Variogramme expérimental sur les résidus obtenus après une régression de moindres carrés **ordinaires** de NO_2, le **17 Juillet 1999 à 8 heures**, avec l'ajustement d'un modèle gaussien.

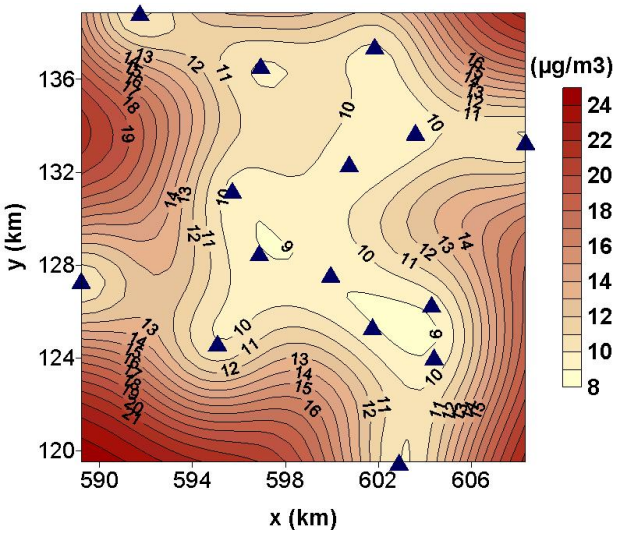


(b) Variogramme expérimental sur les résidus obtenus après une régression de moindres carrés **généralisés** de NO_2, le **17 Juillet 1999 à 8 heures**, avec l'ajustement d'un modèle gaussien.

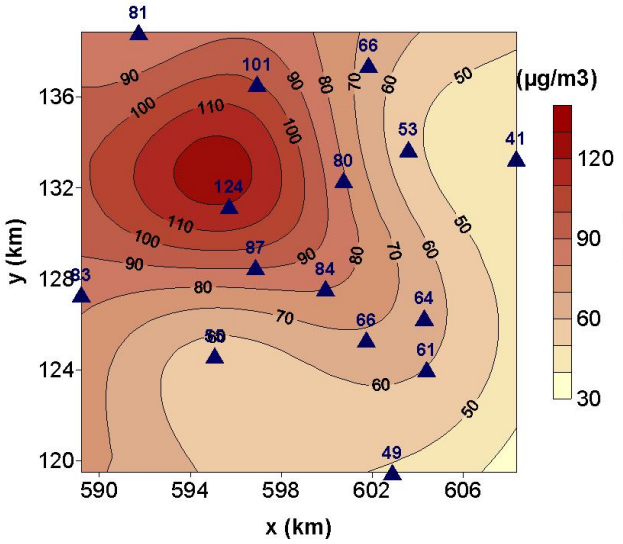
FIG. 2.20: Variogrammes sur les résidus de NO_2, ordinaires ou généralisés, obtenus le 17 Juillet 1999 à 8 heures.



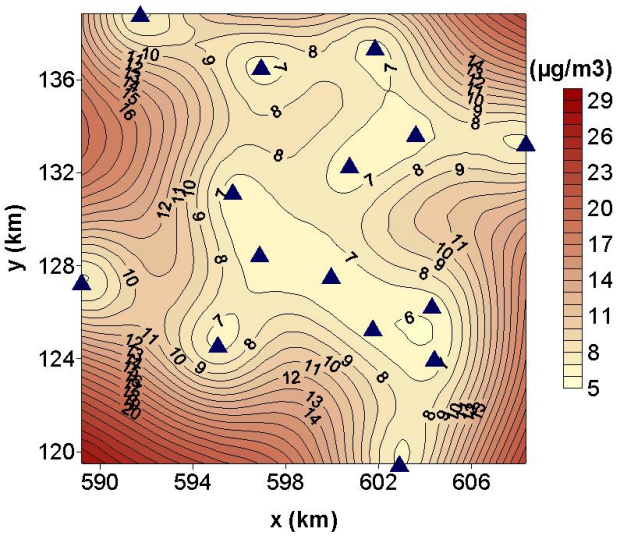
(a) Krigeage Ordinaire (KO).



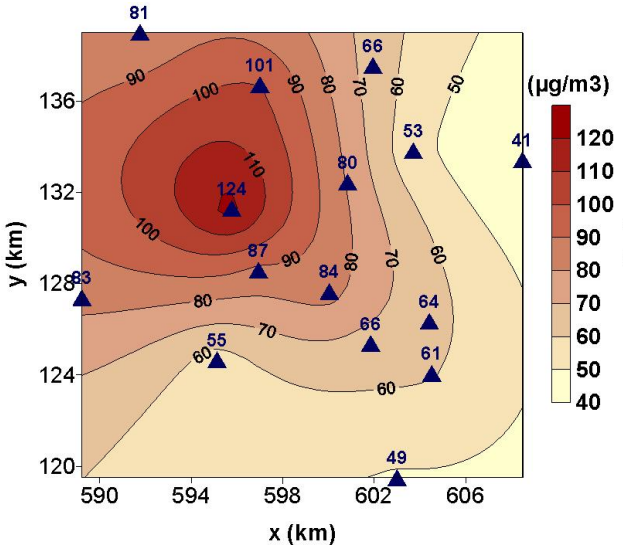
(b) Écart-type de l'erreur d'estimation par KO.



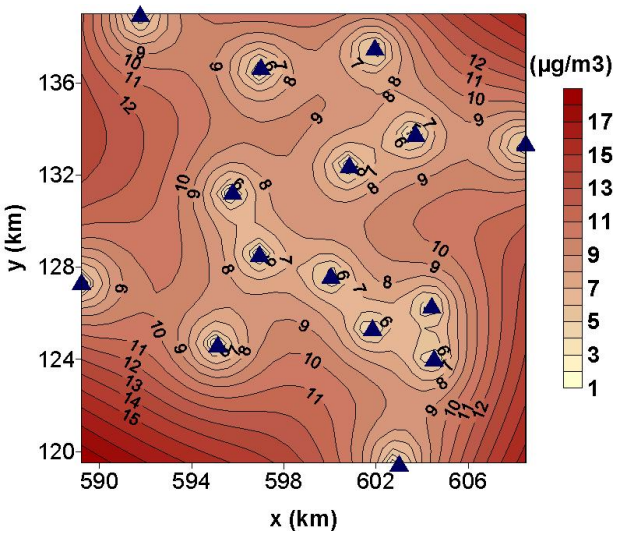
(c) Krigeage Universel (KU).



(d) Écart-type de l'erreur d'estimation par KU.

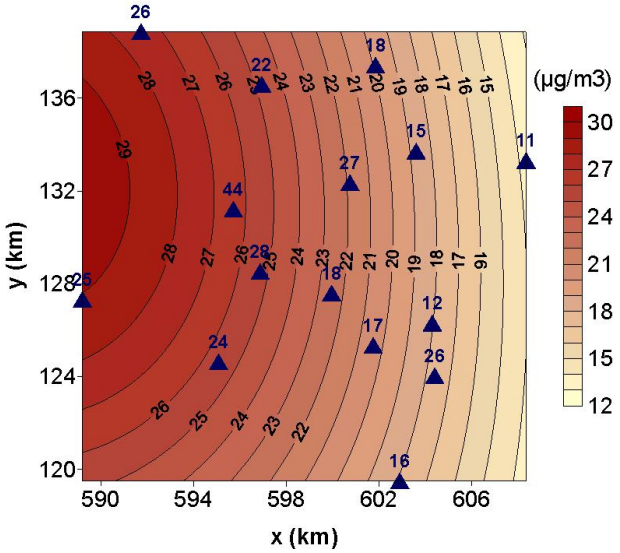


(e) Krigeage Intrinsèque Généralisé (KI).

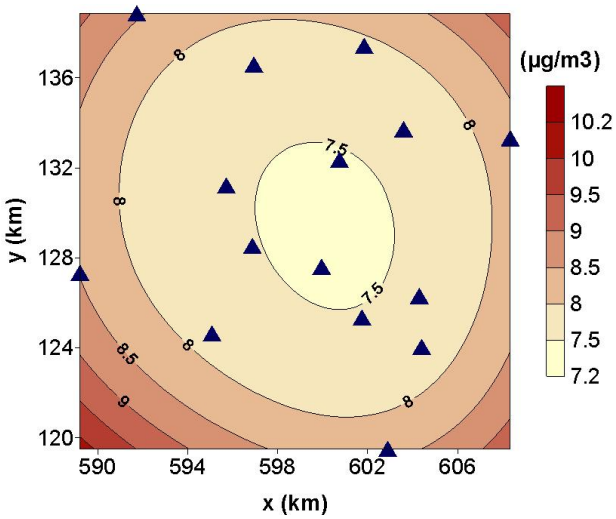


(f) Écart-type de l'erreur d'estimation par KI.

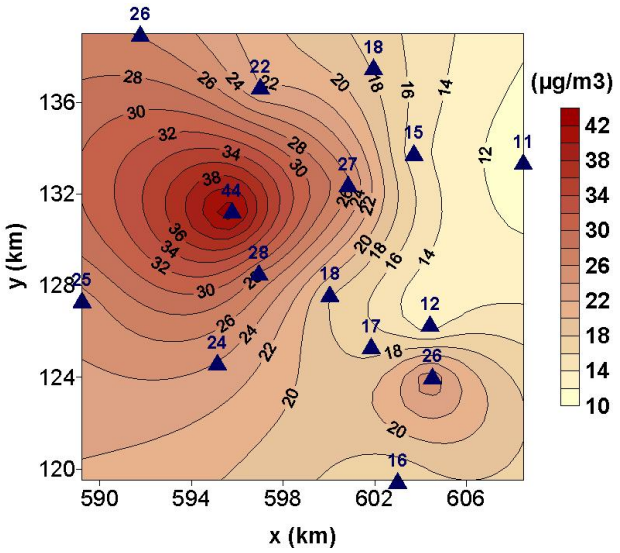
FIG. 2.21: Estimations des champs de concentrations de NO_2 le **17 Juillet 1999 à 8 heures** en appliquant les trois variantes de krigeage KO, KU et KI avec les cartes de l'écart-type de l'erreur associées. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.



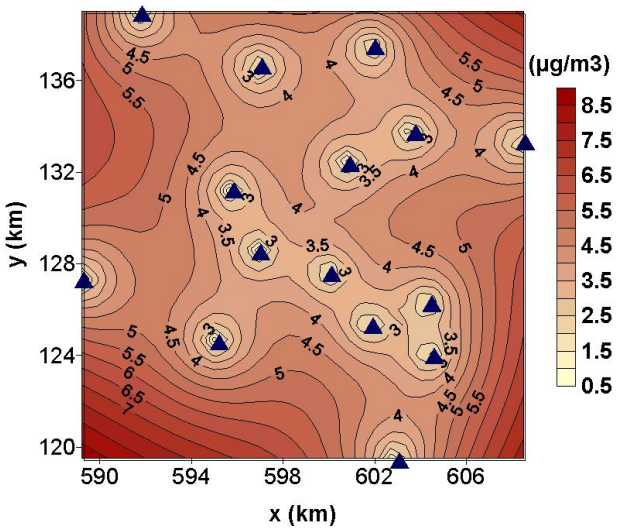
(a) Krigeage Ordinaire (KO).



(b) Écart-type de l'erreur d'estimation par KO.



(c) Krigeage Intrinsèque Généralisé (KI).



(d) Écart-type de l'erreur d'estimation par KI.

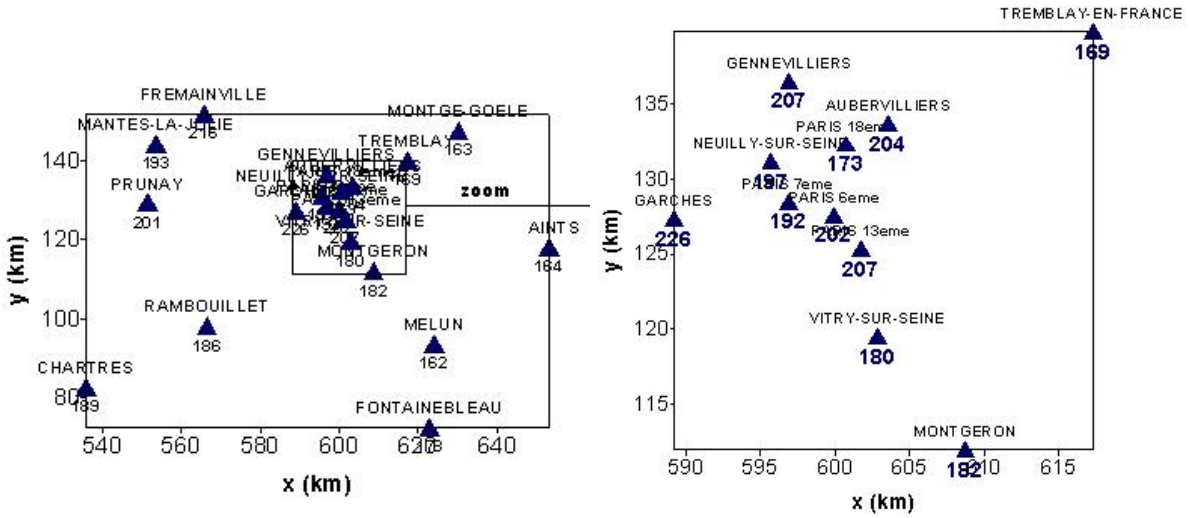
FIG. 2.22: Estimations des champs de concentrations de NO_2 le **17 Juillet 1999 à 15 heures** en appliquant les deux variantes de krigeage KO et KI avec les cartes de l'écart-type de l'erreur associées. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

2.8.4 Représentation des champs de concentrations d'ozone sur l'Île-de-France

En passant au deuxième polluant analysé, l'ozone, il faut se rappeler que la taille du domaine est supérieure à celle utilisée auparavant. Les mesures effectuées en zone rurale sont parfois plus représentatives de quelques kilomètres que celles effectuées en zone urbaine.

2.8.4.1 Champs d'ozone le 30 Juillet 1999 à 14 heures

Comme pour le dioxyde d'azote, on commence par présenter les mesures incluses dans l'étude (voir la figure 2.23(a)), avec un zoom effectué sur l'agglomération parisienne (figure 2.23(b)), où les valeurs enregistrées ont été assez importantes.



(a) Les mesures d'ozone enregistrées le **30/07/1999 14h** sur la grande couronne.

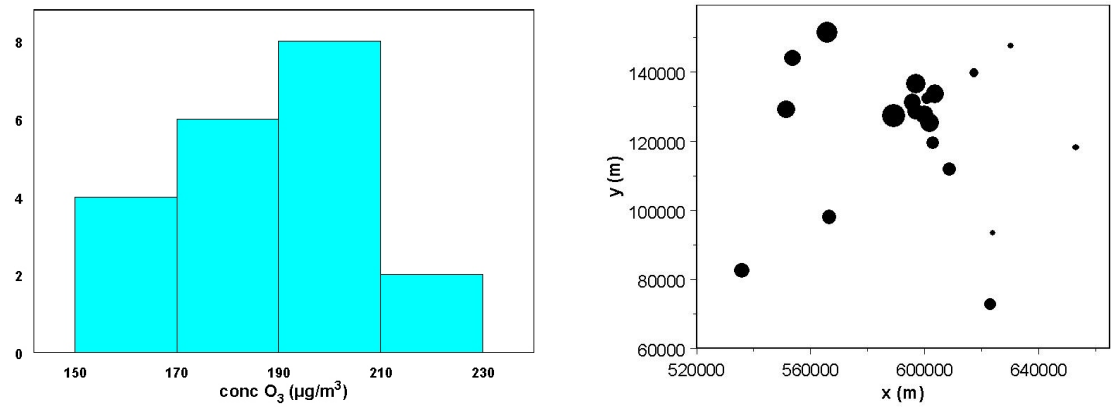
(b) Zoom sur les stations de l'agglomération parisienne.

FIG. 2.23: Toutes les mesures d'ozone enregistrées le **30/07/1999 14h** et utilisées dans l'interpolation.

Analyse statistique et exploratoire

Ce premier jeu de données analysé, enregistré par les stations de mesure le **30 Juillet 1999 à 14 heures**, présente une distribution plus symétrique que celle obtenue pour le dioxyde d'azote, mais, néanmoins, pas normale (voir la figure 2.24(a)). Étant donné que le krigeage est spécialement efficace pour des données présentant une distribution normale, on pourra s'attendre que la méthode de krigeage fonctionne mieux pour l'ozone que pour le dioxyde d'azote ; reste à voir si la couverture spatiale du domaine assurée par les stations automatiques est suffisante pour permettre une bonne estimation du champ d'ozone.

Dans la figure 2.24(b) on peut visualiser la représentation spatiale des mesures disponibles, avec la même convention que celle utilisée dans le cas du dioxyde d'azote : la taille des symboles est proportionnelle aux valeurs enregistrées. On peut remarquer que le centre de Paris est bien couvert

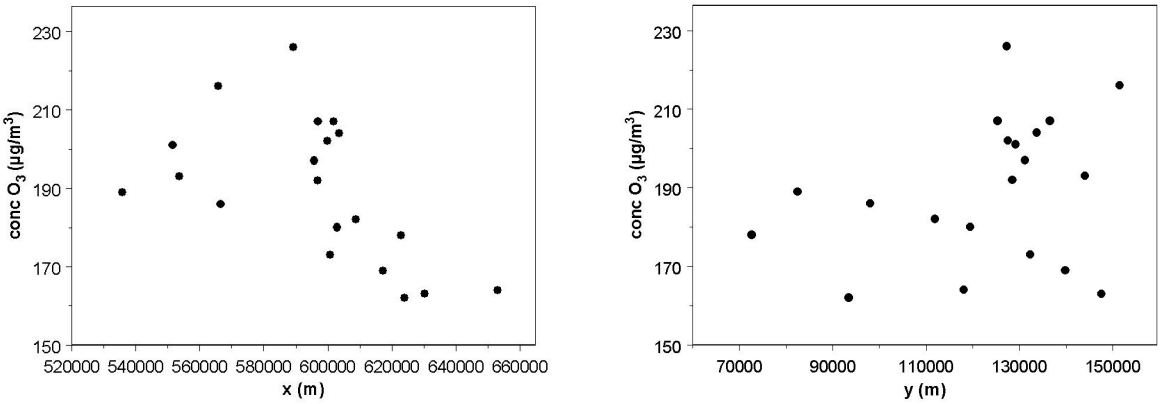


(a) L'histogramme de O_3. (b) Les mesures de O_3.

FIG. 2.24: Distribution spatiale de données de NO_2 le **30 Juillet 1999 à 14 heures**.

par une agglomération de stations automatiques (voir la figure 2.11), tandis que, pour le reste du domaine, la couverture spatiale est assez faible. Il reste des zones où on ne dispose pas de mesures et où on ne peut pas vérifier la qualité de la représentation spatiale obtenue par une méthode d'interpolation. Une autre remarque que l'on peut faire est le fait que pour ce premier cas analysé, aux mesures fournies par les stations automatiques d'AIRPARIF on a rajouté une station située à l'extrémité Sud-Ouest du domaine, appelée CHARTRES-FULBERT, dont les mesures sont disponibles par le réseau LIG'AIR[1], une association régionale créée en 1996 pour assurer la surveillance de la qualité de l'air en région Centre.

Dans le cas analysé, il paraît assez clair que les figures représentant les concentrations comme



(a) Concentrations de O_3 comme fonction de x. (b) Concentrations de O_3 comme fonction de y.

FIG. 2.25: Concentrations de O_3 enregistrées le **30 Juillet 1999 à 14 heures** comme fonction de coordonnées spatiales.

fonctions de coordonnées spatiales (2.25(a), 2.25(b)) montrent une dépendance linéaire plus forte pour la première coordonnée x, mais, néanmoins, présente pour les deux coordonnées spatiales. En

[1] http ://www.ligair.fr

regardant toutes les quatre figures déjà mentionnées, il est difficile de tirer des conclusions concernant la stationnarité de la variable régionalisée à interpoler. De plus, le petit nombre de stations rend inutile l'étude de la stationnarité en employant une fenêtre mobile. Par conséquent, on va considérer la variable aléatoire associée aux concentrations d'ozone comme intrinsèque, et on va appliquer les deux types de krigeage, celui ordinaire et celui universel. Ensuite, on va comparer les résultats avec ceux obtenus en utilisant la théorie de Matheron des FAI-k, décrite dans la section 2.6.6.

Analyse variographique

La nuée variographique de l'ozone est présentée dans la figure (2.26). En suivant la même démarche que celle employée pour le dioxyde d'azote, on peut identifier deux points atypiques. On retrouve ainsi deux paires de stations concernées : Rambouillet-Paris 6 et Rambouillet-Paris 18, qui correspondent, en gros, aux valeurs les plus fortes, respectivement les plus faibles de ce jeu de données. En suivant le même raisonnement que précédemment (pour le dioxyde d'azote), on décide de garder toutes les données initiales et de construire le variogramme expérimental associé.

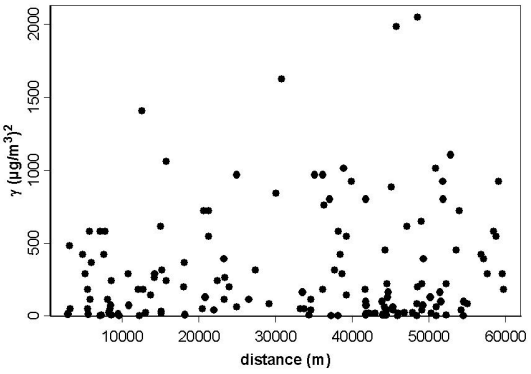
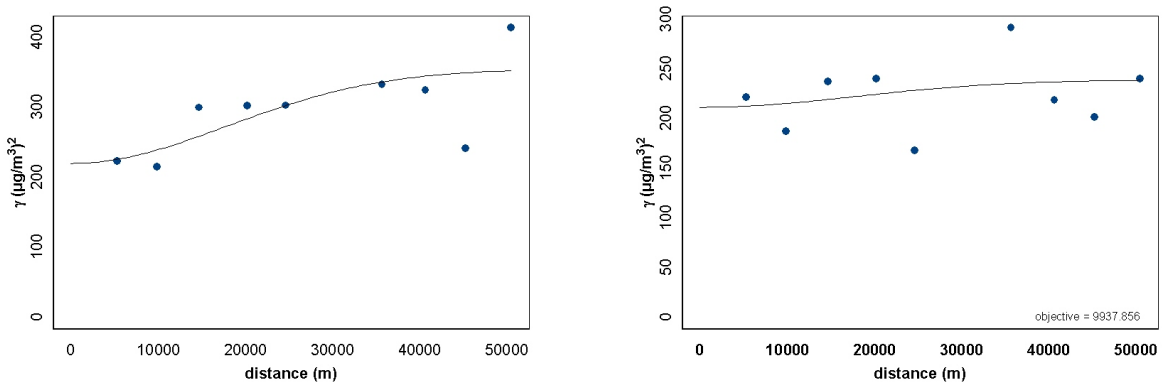


FIG. 2.26: Nuée variographique de O_3 pour les données enregistrées le **30 Juillet 1999 à 14 heures**.

• Variogramme obtenu à partir des données brutes

Dans le cas de l'ozone, vue la taille du domaine, supérieure à celle du domaine de NO_2, le variogramme expérimental a été construit en prenant comme distance maximale 51 km et 10 classes de distances. Le variogramme ajusté est un modèle gaussien (figure 2.27(a)) avec les caractéristiques suivantes : une valeur pépitique de 220 $(\mu g.m^{-3})^2$, la portée de 25 km et le palier à 135 $(\mu g.m^{-3})^2$. Dans le cas de l'ozone, pour ce jeu de données, il faut préciser qu'on a au moins trois modèles qui sont appropriés pour l'ajustement au variogramme expérimental : le modèle gaussien, déjà décrit, le modèle exponentiel et celui sphérique, avec des paramètres du même ordre de grandeur. Le seul critère appliqué dans ce travail pour départager les trois modèles possibles est la validation croisée. Source de critiques de la part de beaucoup de géostatisticiens, cette technique nous mène à la conclusion que le modèle de variogramme qui conduit à la RMSE la plus réduite est le modèle gaussien et c'est celui-ci qu'on a utilisé par la suite.



(a) Le modèle *gaussien* ajusté pour le variogramme expérimental sur les données brutes d'ozone le **30 Juillet 1999 à 14 heures**.

(b) Le modèle *gaussien* ajusté pour le variogramme expérimental sur les résidus des moindres carrés ordinaires d'ozone le **30 Juillet 1999 à 14 heures**.

FIG. 2.27: Exemples de modèles ajustés au variogramme expérimental (données brutes et résidus) pour l'ozone le 30 Juillet 1999 à 14 heures.

• **Variogramme obtenu à partir des résidus**

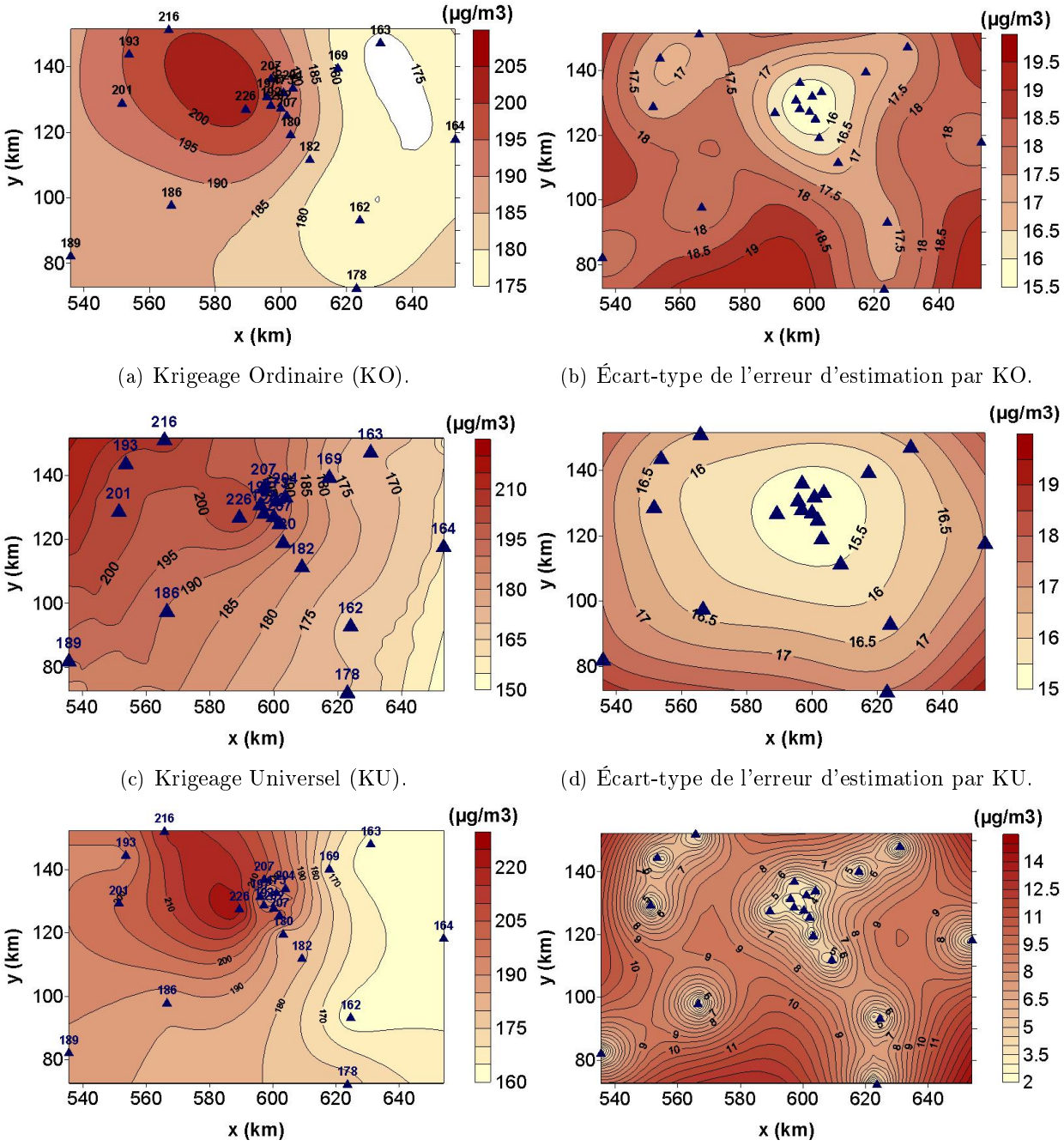
En procédant de la même manière que pour le dioxyde d'azote, on se résume à une seule régression des moindres carrés ordinaires pour éliminer une tendance linéaire et obtenir des résidus, sur lesquels on a ajusté un modèle gaussien de variogramme avec des paramètres très semblables à ceux obtenus antérieurement. Les valeurs de la portée et de l'effet pépitique sont les mêmes. En revanche, le palier descend de $135\,(\mu\text{g.m}^{-3})^2$ à $28\,(\mu\text{g.m}^{-3})^2$. De cette façon, le variogramme devient presque plat. Par conséquent, le champ estimé sera lisse.

• **Covariance généralisée $K(h)$ à partir des données brutes**

L'application du krigeage intrinsèque généralisé passe par l'identification de l'ordre de continuité et des coefficients de la covariance généralisée. Pour le premier jeu de données, on a obtenu $\nu = 0$ et, comme dans le cas du dioxyde d'azote, la même covariance généralisé linéaire.

2.8.4.2 KO, KU et KI appliqués sur les données de O_3 le 30 Juillet 1999 à 14 heures

Une fois le modèle de variogramme choisi on passe à l'estimation. Comme précédemment, on a choisi de présenter les cartes obtenues par interpolation en utilisant les trois méthodes déjà présentées : le krigeage ordinaire, le krigeage universel et celui intrinsèque généralisé pour comparer ainsi les résultats. Les différences entre les trois cartes sont assez évidentes. Tandis que pour le KO et le KI le pic d'ozone est très localisé, le KU produit une carte très lisse, car le variogramme ajusté est presque plat. Une autre différence concerne les lignes d'isoconcentration. Il existe plusieurs stations qui ne sont pas bien classées par le KO et le KU ; ce n'est pas le cas pour la méthode du krigeage intrinsèque généralisé, où toutes les mesures sont classées entre les bonnes isolignes. Ceci est dû à



(a) Krigeage Ordinaire (KO).

(b) Écart-type de l'erreur d'estimation par KO.

(c) Krigeage Universel (KU).

(d) Écart-type de l'erreur d'estimation par KU.

(e) Krigeage Intrinsèque Généralisé (KI).

(f) Écart-type de l'erreur d'estimation par KI.

FIG. 2.28: Estimations des champs de concentrations d'ozone le **30 Juillet 1999 à 14 heures** en appliquant les trois variantes de krigeage KO, KU et KI avec les cartes de l'écart-type de l'erreur associées. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

l'effet de pépite, et cela a déjà été commenté dans le cas du NO_2. Le manque de stations qui caractérise l'extrémité Sud-Ouest du domaine, ainsi que la zone Nord-Est conduit à des représentations assez différentes en utilisant les trois types de krigeage mentionnés.

Pour mieux comparer la qualité de ces cartes on a effectué une validation croisée et les résultats sont présentés dans le tableau 2.6.

Station	Valeur mesurée (μg.m^{-3})	**KO** (μg.m^{-3})	**KU** (μg.m^{-3})	**KI** (μg.m^{-3})
AUBERVILLIERS	204,00	**190,32**	189,89	178,01
GARCHES	226,00	195,83	193,39	**198,79**
GENEVILLIERS	207,00	**195,30**	193,80	193,45
MANTES	193,00	**197,74**	211,59	209,58
MELUN	162,00	181,82	**174,20**	175,77
MONTGERON	182,00	184,61	**182,61**	174,57
NEUILLY	197,00	198,16	194,85	**196,87**
PARIS 13	207,00	191,32	189,24	**194,49**
PARIS 18	173,00	**196,85**	194,47	202,16
PARIS 6	202,00	193,93	191,36	**195,26**
PARIS 7	192,00	197,62	**194,12**	200,29
RAMBOUILLET	186,00	**189,49**	193,78	198,23
TREMBLAY	169,00	**183,33**	186,41	184,58
VITRY	180,00	192,44	**189,13**	199,53
MONTGE	163,00	180,75	182,61	**169,28**
FREMAINVILLE	216,00	191,46	**201,34**	199,52
PRUNAY	201,00	191,41	**204,55**	196,15
FONTAINEBLEAU	178,00	**180,22**	159,49	152,89
SAINTS	164,00	182,95	**167,71**	159,22
CHARTRES	189,00	**184,84**	208,28	150,76

TAB. 2.6: Validation croisée pour l'ozone (**30/07/1999 14h**) pour les trois types de krigeage appliqués. Les meilleures estimations ont été mises en gras.

Le tableau 2.6 nous aide à séparer les stations en deux catégories. La première contient les stations sur lesquelles les estimations par validation croisée sont toujours de mauvaise qualité. C'est le cas premièrement de Garches, station qui enregistre la valeur la plus élevée de l'ensemble ; en conséquence, si on retire cette valeur, et on effectue l'estimation, on obtient forcément une concentration plus faible que celle réelle. Ensuite, on retrouve Paris 13 et Paris 6 avec des valeurs plus élevées que le reste de stations situées en pleine ville. Dernièrement, on retrouve les stations situées en bord du domaine : Fremainville, Montge-en-Goele, Saints et Melun (voir la figure 2.23(a)).

Le deuxième groupe est constitué de stations dont les valeurs sont bien reproduites par la procédure de krigeage.

On remarque la station Chartres-Fulbert qui a été incluse dans l'étude pour couvrir la partie Sud-Ouest du domaine, où la mesure enregistrée a été bien estimée uniquement par krigeage ordinaire, les deux autres méthodes estimant des valeurs bien différentes de celle réelle.

Les statistiques globales pour les tests "Leave-One-Out" effectués montrent que le krigeage intrinsèque n'est pas la méthode la plus efficace parmi les trois méthodes appliquées, même si elle reste la seule à avoir bien classé toutes les stations et malgré les valeurs les plus faibles sur la carte de l'écart-type associée.

Méthode	MIN (μg.m^{-3})	MAX (μg.m^{-3})	MAE (μg.m^{-3})	RMSE (μg.m^{-3})
KO	1,16	30,17	12,23	14,72
KU	0,61	32,61	12,95	15,19
KI	0,13	38,24	15,22	17,98

TAB. 2.7: Statistiques globales pour les tests "Leave-One-Out" obtenus pour les trois types de krigeage appliqués sur les données d'ozone enregistrées le **30 Juillet 1999 à 14 heures**.

2.8.4.3 Champs d'ozone le 17 Juillet 1999 à 6 heures et à 15 heures

On continue l'analyse des champs de concentrations d'ozone par la présentation de deux cas différents choisis de la même façon que ceux pour le dioxyde d'azote, c'est-à-dire les valeurs maximales et respectivement minimales enregistrées le **17 Juillet 1999**, jour de pic d'ozone dans la région parisienne.

Champs d'ozone le 17 Juillet 1999 à 6 heures

On commence par l'analyse des données de faible pollution, enregistrées à **6 heures** du matin.

• Variogramme obtenu à partir des données brutes

L'analyse variographique fournit un modèle gaussien, caractérisé par une valeur pépitique très faible de 20 $(\mu g.m^{-3})^2$, une portée d'environ 21 km et un palier de 537 $(\mu g.m^{-3})^2$. C'est surtout le palier qui est très grand.

• Variogramme obtenu à partir des résidus

Sur les données brutes, on a appliqué une itération complète pour arriver aux résidus des moindres carrés généralisés, sur lesquels on a ajusté un deuxième modèle gaussien de paramètres : valeur pépitique très faible de 32 $(\mu g.m^{-3})^2$, une portée d'environ 27 km et un palier de 581 $(\mu g.m^{-3})^2$ (figure 2.29).

• Covariance généralisée $K(h)$ à partir des données brutes

Cette fois-ci, l'ordre de continuité estimé est $\nu = 2$ et la covariance généralisée toujours linéaire.

2.8.4.4 KO, KU et KI appliqués sur les données de O_3 le 30 Juillet 1999 à 6 heures

Les cartes obtenues en utilisant le modèle de variogramme gaussien, avec le KO et le KU sont présentées dans les figures 2.30(a), 2.30(c). On les compare avec la carte obtenue en utilisant la théorie de Matheron, FAI-k, pour le même jeu de données et l'écart-type associé. Les différences

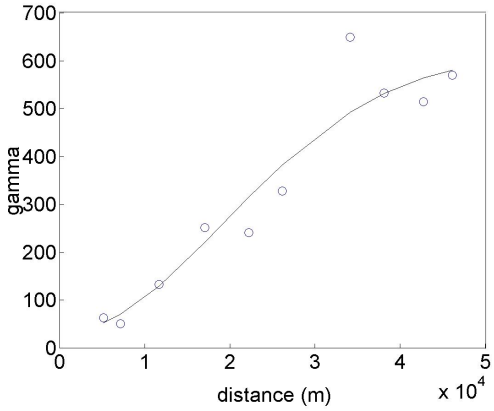


FIG. 2.29: Le modèle *gaussien* ajusté pour le variogramme expérimental sur les résidus des moindres carrés généralisés d'ozone le **17 Juillet 1999 à 6 heures**.

sont évidentes. Premièrement, on remarque la zone bizarre estimée par le krigeage intrinsèque à droite de l'agglomération de stations urbaines qui caractérise le centre du domaine. Deuxièmement, il y a un petit pic de $80\,\mu\text{g.m}^{-3}$ estimé dans une zone qui n'est pas couverte par les mesures. Ces deux anomalies nous confirment que, sur ces jeux de données, le krigeage intrinsèque n'est pas du tout efficace, car il s'agit de zones importantes d'extrapolation, où les résultats peuvent être surprenants. Quant aux deux autres méthodes, on constate des différences notables uniquement sur la partie Sud-Ouest, là où on ne dispose pas des mesures.

Champs d'ozone le 17 Juillet 1999 à 15 heures

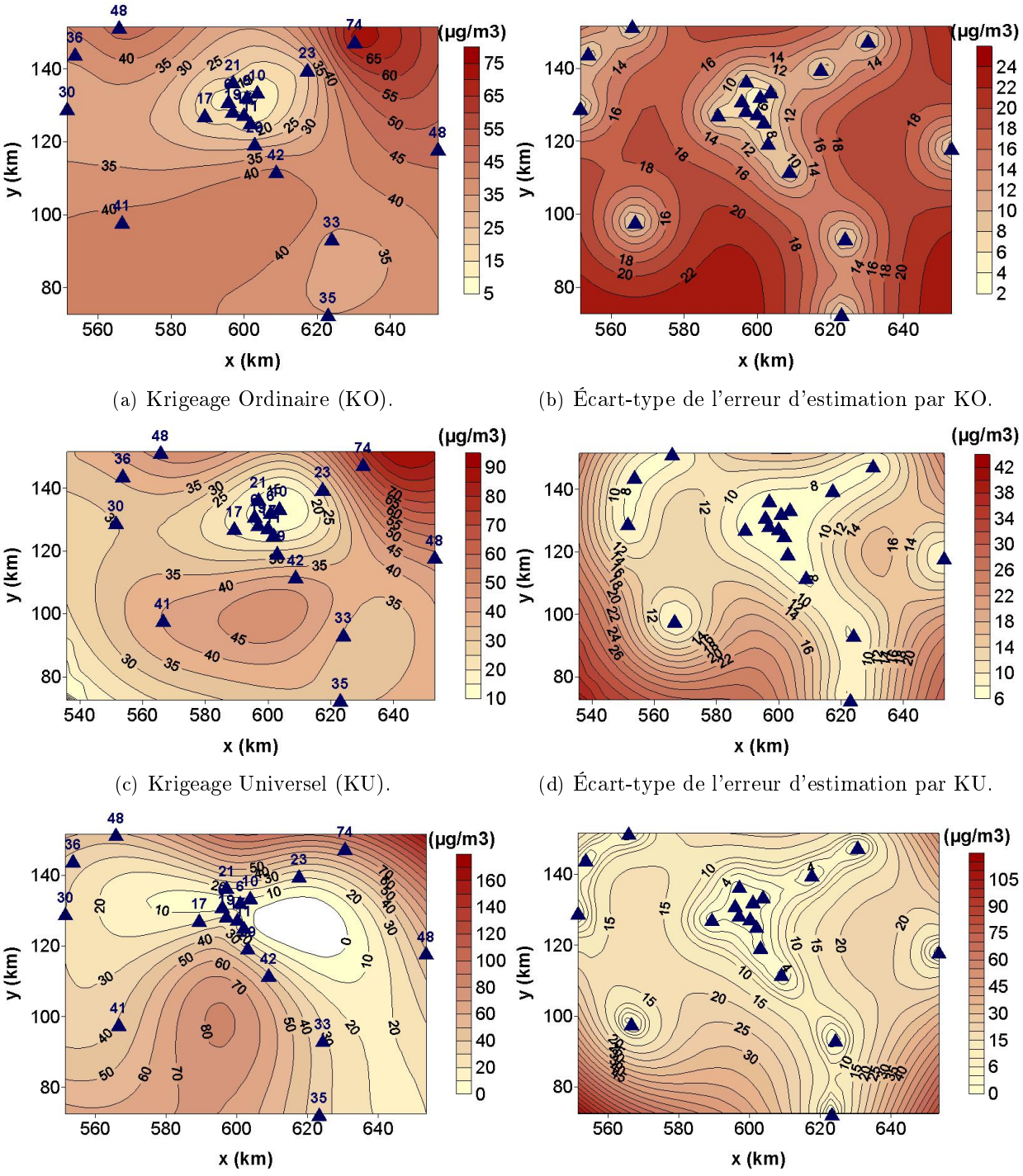
On continue par les données correspondantes à un épisode de forte pollution enregistré à **15 heures**.

- **Variogramme obtenu à partir des données brutes**

Le deuxième cas analysé, le **17 Juillet 1999 à 15 heures**, est très important, dans la mesure où il représente le moment d'un pic d'ozone qui a caractérisé la région parisienne pendant la deuxième décade de Juillet 1999. On va comparer d'ailleurs ce cas avec les résultats obtenus en utilisant un modèle déterministe corrigé par une méthode séquentielle d'assimilation de données. Le problème avec ces données est que leur variabilité est très importante (voir le tableau 2.2) et tous les essais pour ajuster un modèle admissible au variogramme expérimental ont échoué, faute d'une interpretation physique cohérente des paramètres résultants. En revanche, on peut effectuer une régression des moindres carrés généralisés.

- **Variogramme obtenu à partir des résidus**

La technique utilisée a été déjà décrite. On ajuste d'abord sur les résidus des moindres carrés ordinaires un modèle gaussien de paramètres : valeur pépitique très faible de $35\,(\mu\text{g.m}^{-3})^2$, une portée d'environ 21,5 km et un palier de $471\,(\mu\text{g.m}^{-3})^2$. Ensuite, sur les résidus généralisés (obtenus



(a) Krigeage Ordinaire (KO).

(b) Écart-type de l'erreur d'estimation par KO.

(c) Krigeage Universel (KU).

(d) Écart-type de l'erreur d'estimation par KU.

(e) Krigeage Intrinsèque Généralisé (KI).

(f) Écart-type de l'erreur d'estimation par KI.

FIG. 2.30: Estimations des champs de concentrations d'ozone le **17 Juillet 1999 à 6 heures** en appliquant les trois variantes de krigeage KO, KU et KI avec les cartes de l'écart-type de l'erreur associées. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

avec la matrice $\mathbf{\Gamma}$ du premier modèle gaussien) on ajuste un nouveau modèle gaussien (voir la figure 2.31). Les nouveaux paramètres sont eux aussi assez proches de ceux obtenus précédemment : valeur pépitique très faible de $34\,(\mu\text{g.m}^{-3})^2$, une portée d'environ 22,5 km et un palier de $574\,(\mu\text{g.m}^{-3})^2$.

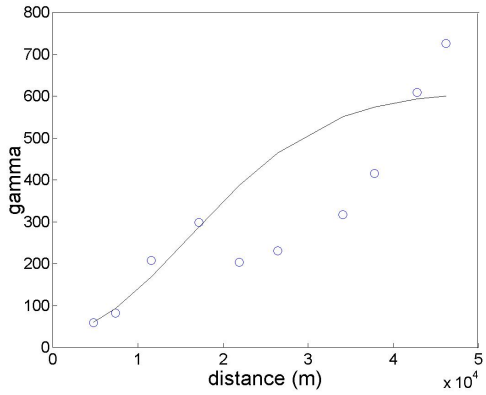


FIG. 2.31: Le modèle *gaussien* ajusté pour le variogramme expérimental sur les résidus des moindres carrés généralisés d'ozone le **17 Juillet 1999 à 15 heures**.

• **Covariance généralisée $K(h)$ à partir des données brutes**

En ce qui concerne le troisième type de krigeage, intrinsèque généralisé, l'ordre de continuité estimé est $\nu = 1$ et la covariance généralisée toujours linéaire.

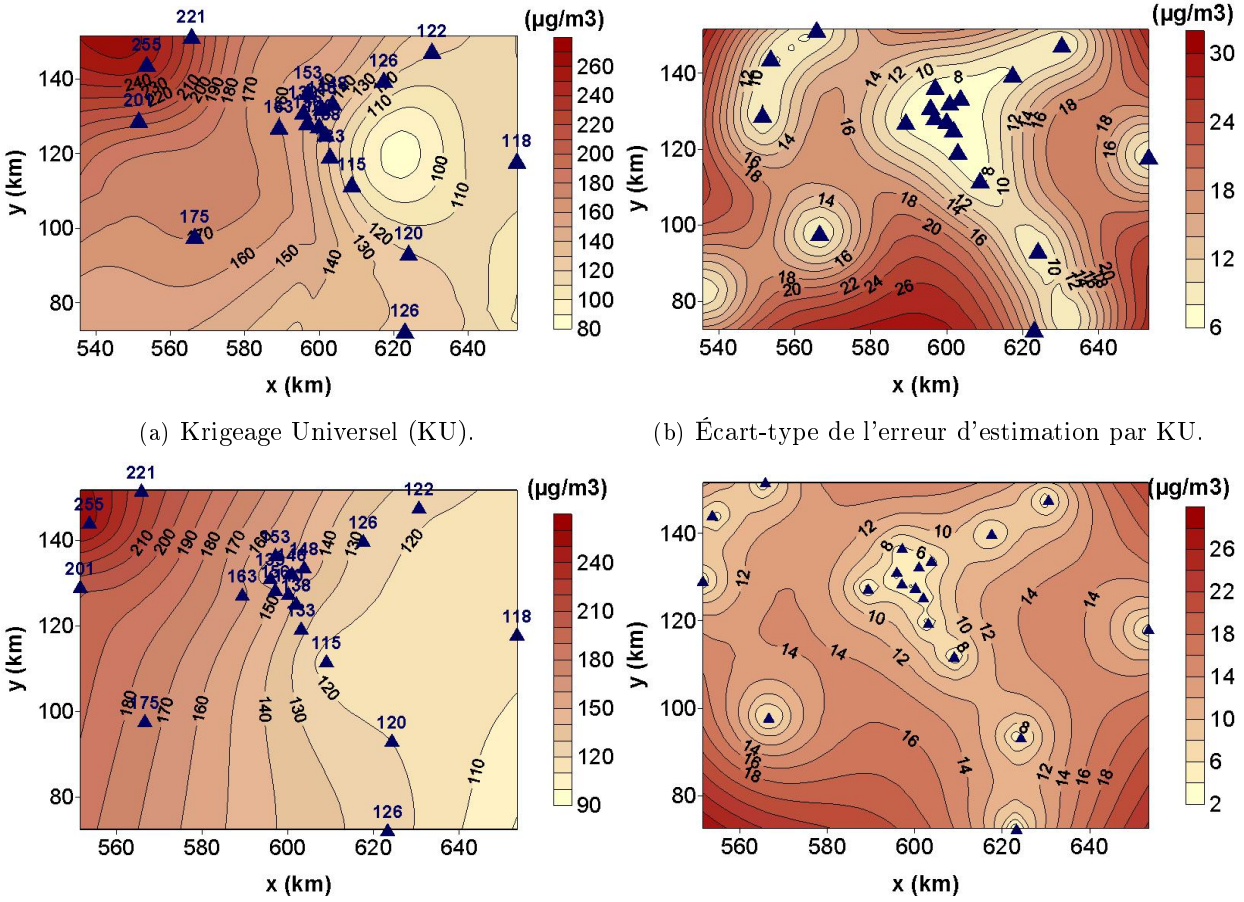
2.8.4.5 KO, KU et KI appliqués sur les données de O_3 le 17 Juillet 1999 à 15 heures

La carte obtenue en appliquant le krigeage universel est présentée dans la figure 2.32(a), accompagnée de la carte de l'écart-type de l'erreur associée. En employant la méthode de krigeage intrinsèque on obtient la carte présentée dans la figure 2.32(c) avec la carte de l'écart-type associé (voir la figure 2.32(d)). Les différences sont encore une fois évidentes. La carte obtenue par KI est plus lisse que celle obtenue par KU. La zone dépourvue de mesures située à droite de l'agglomération urbaine est estimée de façons différentes, ainsi que celle située au Sud-Ouest du domaine. Les pics simulés au coin Nord-Ouest du domaine ont la même allure, mais, pour bien préciser la forme du panache, on aurait besoin des mesures en dehors de ce domaine, plus au nord et à l'ouest. De plus, ne pas intégrer dans l'analyse spatiale des contraintes liées aux conditions météorologiques, ou aux émissions, nous empêche d'obtenir des cartes convenables pour cette zone d'étude.

Récapitulatif sur l'analyse des données de O_3

Dans le tableau 2.8 on présente un récapitulatif des paramètres obtenus pour les modèles ajustés aux variogrammes expérimentales utilisés dans le trois types de krigeage.

Comme c'était d'ailleurs le cas pour le dioxyde d'azote aussi, dans le cas du deuxième polluant choisi, l'ozone, les trois méthodes d'interpolation spatiale décrites et appliquées : KO, KU et KI ne donnent pas entièrement satisfaction. On peut conclure qu'il n'existe pas une "meilleure" méthode parmi celles appliquées et que la qualité de la représentation spatiale dépend en grande partie de la variance présente dans les données. Le point clé de cette application est le modèle qu'on peut ajuster au variogramme expérimental. Malgré le fait que l'ajustement peut être correct, pendant les épisodes de forte pollution, on constate que le nombre de mesures n'est pas suffisant pour pouvoir reconstituer, à partir de ces mesures, le panache formé sur la zone d'étude ou en dehors de celle-ci.



(a) Krigeage Universel (KU).

(b) Écart-type de l'erreur d'estimation par KU.

(c) Krigeage Intrinsèque Généralisé (KI).

(d) Écart-type de l'erreur d'estimation par KI.

FIG. 2.32: Estimations des champs de concentrations d'ozone le **17 Juillet 1999 à 15 heures** en appliquant le KI avec la carte de l'écart-type de l'erreur associée. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

Date	**Type de données**	**Modèle**	**Valeur pépitique** $(\mu g.m^{-3})^2$	**Palier** $(\mu g.m^{-3})^2$	**Portée** (mètres)	**Qualificatif**
17/07/99 h6	**brutes**	Gaussien	20	537	21000	assez bonne
17/07/99 h6	**résidus-ols**	Gaussien	27	647	29617	–
17/07/99 h6	**résidus-gls**	Gaussien	32	581	27199	bonne
17/07/99 h15	**résidus-ols**	Gaussien	35	471	21550	–
17/07/99 h15	**résidus-gls**	Gaussien	34	574	22493	assez bonne
30/07/99 h14	**brutes**	Gaussien	220	135	25000	assez bonne
30/07/99 h14	**résidus-ols**	Gaussien	220	28	25000	assez bonne

TAB. 2.8: Tableau récapitulatif avec les paramètres des modèles ajustés aux variogrammes expérimentaux pour les cas étudiés d'ozone. Les résidus-ols correspondent à ceux obtenus par les moindres carrés ordinaires, et les résidus-gls, par des moindres carrés généralisés.

Étendre la zone d'étude ne servira à rien, car il n'existe pas de mesures en dehors de celles utilisées

ici ; on sera donc forcé à faire de l'extrapolation. Par rapport aux résultats obtenus pour le NO_2, ponctuellement ils peuvent être meilleurs pour l'ozone, mais globalement les cartes obtenues sont moins représentatives, car la couverture du domaine d'étude par les stations de mesure s'est révélée moins bonne du point de vue pouvoir informationnel que pour le NO_2 ; pour l'ozone il y a des zones assez étendues qui ne sont pas couvertes de stations et où on se retrouve en situation d'extrapolation.

2.8.5 Conclusion partielle sur les résultats obtenus par interpolation spatiale

Les résultats présentés dans ce chapitre visaient la représentation spatiale des champs de concentrations de polluants atmosphériques, notamment le dioxyde d'azote et l'ozone. Les cartes obtenues par interpolation spatiale en appliquant trois variantes de krigeage ordinaire, universel et intrinsèque généralisé, ont été comparées pour tirer des conclusions concernant la question : quelle est, parmi les trois, la méthode la plus appropriée et la plus efficace en termes de validation croisée pour représenter spatialement les données disponibles. La réponse n'est pas évidente. Une partie des conclusions sont communes aux deux cas analysés (NO_2 et O_3), une autre est spécifique.

- Pour le dioxyde d'azote

Premièrement, on peut remarquer que la zone d'étude est bien couverte par les stations de mesure pour éviter l'extrapolation ; d'ailleurs, c'est dans ce but que la taille de cette zone a été un peu diminuée par rapport à celle couverte par les stations automatiques d'AIRPARIF.

Un point clé du problème d'interpolation par krigeage est l'ajustement d'un modèle théorique au variogramme expérimental. Déjà, la détermination du variogramme expérimental est basée sur un certain nombre de choix empiriques, mais heureusement, l'estimation n'est pas très sensible aux paramètres du variogramme. Ceci fait que si les paramètres estimés sont, en toute simplicité, cohérents avec la réalité physique, alors on a toutes les chances d'obtenir une estimation *correcte*.

Parmi les trois cas analysés, on a deux cas de fortes épisodes de pollution et un seul de faible pollution. Dans les deux cas présentant des concentrations élevées, l'ajustement du variogramme est acceptable et on constate que les cartes d'isoconcentrations obtenues par les trois types de krigeage sont très semblables ; les cartes de l'écart-type de l'erreur associées sont elles aussi proches, avec des valeurs un peu plus faibles pour le krigeage intrinsèque généralisé. En revanche, pour le cas de faible pollution, la situation est délicate à cause d'un mauvais ajustement d'un modèle théorique au variogramme expérimental. Ne pouvant pas identifier la structure spatiale des données par un semi-variogramme, on est obligé à recourir à la méthode de covariance généralisée, pour pouvoir tenir compte de cette structure dans l'estimation.

Les statistiques moyennes effectuées sur les tests "Leave-One-Out", montrent qu'aucune des trois méthodes n'est privilégiée par rapport aux autres. Elles indiquent une erreur d'estimation de l'ordre de $10\,\mu g/m^3$ en moyenne pour les stations à l'intérieur du domaine, et plus importante aux bords ($30 - 40\,\mu g/m^3$). L'estimation peut être jugée comme satisfaisante.

Par ailleurs, on a pu constater l'influence de l'effet pépitique sur le champ estimé (conduisant à des valeurs mal classées, voire des discontinuités dans le champ estimé au niveau des points de mesure), ainsi que sur la carte de variance de l'erreur (lors du KO ou KU la variance de l'estimation ne descend pas en-dessous de la valeur pépitique, mais elle est nulle aux points de mesure lors de l'application du KI). Les effets de bords sont présents dans tous les résultats et ils sont comparables comme ordre de grandeur pour les trois méthodes.

- Pour l'ozone

Pour ce deuxième polluant, la situation est plus difficile ; on s'aperçoit dès le debut que la répartition des stations de mesure laisse des zones creuses importantes. D'ailleurs, l'impression générale, en analysant les deux épisodes de forte pollution présentés, est que le nombre de mesures n'est pas suffisant pour pouvoir reconstituer, à partir de ces mesures, la forme du panache. Étendre la zone d'étude ne servira à rien, car il n'existe pas de mesures en dehors de celles utilisées ici ; on sera donc forcé à faire de l'extrapolation. L'analyse effectuée conduit donc à la conclusion que le nombre de stations disponibles et leur répartition spatiale sont nettement insuffisants pour obtenir des cartes de pollution d'ozone convenables.

Alors que dans le cas du NO_2, on n'a pas pu ajuster un variogramme dans le cas d'une faible variabilité spatiale, dans le cas de l'ozone, on n'arrive pas à trouver une structure spatiale (ajuster un variogramme cohérent), dans le cas d'une forte variabilité. Le manque de structure peut être rencontré donc dans plusieurs situations différentes.

Malgré le fait que l'aspect des cartes n'est pas très convaincant, les tests "Leave-One-Out" ont conduit à des erreurs de l'ordre $12 - 15\,\mu g/m^3$, qui peuvent être jugées comme satisfaisantes. Comme c'était le cas pour le dioxyde d'azote, on remarque aussi que, pour l'ozone, la représentativité des stations situées en bord du domaine est importante, tandis que, dans l'agglomération, cette représentativité diminue, à cause d'une certaine redondance. Bien que, dans le cas du NO_2, le KI donnait toujours des résultats convenables (sans être forcément les meilleurs), ceci n'est plus le cas pour la configuration des stations d'ozone. On peut penser que cette méthode se comporte moins bien que les autres en situation d'extrapolation, pouvant générer des artefacts.

2.8.6 Conclusion du chapitre

Le but de ce chapitre a été de décrire les méthodes d'interpolation spatiales qui tiennent compte de la structure spatiale des données, en particulier le krigeage, méthodes appliquées par la suite sur des jeux de données réelles pour ainsi répondre à la question principale qu'on s'est posée : est-ce que les mesures enregistrées par les stations automatiques sont suffisantes pour obtenir une image réaliste de la situation concernant la pollution atmosphérique dans la région d'Île-de-France ? Pour l'instant la réponse à cette question est plutôt mitigée ; d'un côté, on reconnaît la facilité de la mise en œuvre d'une *simple* interpolation spatiale (même si, parmi les méthodes d'interpolation spatiale, le krigeage n'est pas la plus simple), mais d'un autre côté, l'interprétation du variogramme expérimental (choix des paramètres : distance maximale, tolérance, pas de calcul), le choix d'un

modèle d'ajustement et le type de krigeage sont des décisions assez difficiles à prendre, basées surtout sur l'expérience de l'utilisateur. Il faut remarquer également que le manque d'une couverture spatiale correcte, par des stations de mesure, est un facteur limitatif dans le processus d'estimation, le deuxième étant la qualité des mesures prises en compte dans le krigeage. Reste l'espoir que, en rajoutant la dimension temporelle, on peut améliorer la qualité, voire la confiance, de nos représentations spatiales.

Chapitre 3

Interpolation spatio-temporelle

Le début du troisième chapitre est dédié à la présentation du contexte et de la continuité spatio-temporelle, éléments incontournables pour une modélisation spatio-temporelle. La question principale à laquelle on veut trouver une réponse est la suivante : est-ce que l'utilisation des séries temporelles disponibles aux emplacements des stations de mesure (valeurs antérieures ou postérieures au moment de l'estimation) peut améliorer la précision de l'estimation, faite à un certain moment ? Est-ce que la richesse temporelle peut combler le manque spatial déjà évoqué dans le chapitre antérieur ? Pour essayer de répondre à cette question, on reprendra les mêmes configurations analysées antérieurement (dans le cas de l'analyse spatiale) pour le dioxyde d'azote et l'ozone, mais tout d'abord on présentera les bases théoriques de l'approche.

3.1 Approche géostatistique pour l'analyse spatio-temporelle

Les processus spatio-temporels sont présents dans un grand nombre de domaines des sciences de la terre et de l'ingénierie. Parmi ceux-ci, les processus atmosphériques peuvent être considérés comme des fonctions spatio-temporelles présentant des fluctuations continues complexes. Cependant, la plupart des outils théoriques et techniques disponibles pour traiter les données dans l'espace et dans le temps ont été établis indépendamment pour chaque variable ; l'importance de la variabilité spatio-temporelle n'était pas appréciée jusqu'il y a quelques années auparavant.

La géostatistique offre une grande variété de méthodes pour modéliser les processus spatio-temporels comme des réalisations de fonctions aléatoires. Kyriakidis et Journel (1999) présentent dans leur article des modèles qui tentent de relier les deux types de coordonnées, spatiales et temporelles, pour décrire la variabilité spatio-temporelle, plus complexe que les deux types de variabilités prises chacune séparément.

Les données spatio-temporelles sont analysées traditionnellement par des modèles développés soit pour les distributions spatiales, soit pour celles temporelles. Une solution évidente pour le problème posé est de considérer le phénomène spatio-temporel comme la réalisation d'une fonction aléatoire en $n+1$ dimensions (n pour l'espace physique plus 1 pour la dimension temporelle). Cette approche demande l'extension des techniques spatiales existantes pour le domaine spatio-temporel. Malgré l'apparente facilité théorique de cette extension, il existe un certain nombre de problèmes

pratiques et théoriques qui devraient être résolus avant toute application sur des données spatio-temporelles. Ces problèmes incluent les différences qualitatives existantes entre les informations spatiales et celles temporelles, ainsi que la présence de la périodicité temporelle et la non-stationnarité spatiale.

Il existe des différences majeures entre les données spatiales et celles temporelles. Premièrement, les données temporelles sont ordonnées : passé, présent, futur, tandis que les données spatiales bi- ou tri-dimensionnelles n'exhibent pas un tel ordre. De plus, les phénomènes spatio-temporels présentent des échelles différentes au niveau spatial et temporel et ils ne sont pas comparables physiquement. Pour prendre en compte cette différence, une solution opérationnelle est de séparer la corrélation spatio-temporelle dans deux termes distincts, soit comme produit (Mejia et Rodriguez-Iturbe, 1974), soit comme somme (Bilonick, 1985) de termes spatiaux et temporels. Les études citées montrent que, malgré la simplification énoncée, cette solution conduit à des résultats fiables.

Un autre problème très important est la présence dans la plupart des cas d'une périodicité temporelle et d'une non-homogénéité spatiale. On peut observer une grande variété de périodicités : saisonnières, cycliques ou pseudo-cycliques, journalières. En ce qui concerne la composante périodique, on a deux possibilités : la première est de la traiter comme si elle faisait partie de la tendance et de la filtrer (Brockwell et Davis, 1987; Séguret, 1989) ; la deuxième possibilité est de considérer les séries comme stationnaires, avec une composante temporelle inclue dans le variogramme ou le correlogramme (Rouhani et Wackernagel, 1990).

3.2 Champs d'application des modèles spatio-temporels

Il existe un très vaste champ d'applications et d'approches pour les modèles spatio-temporels ; parmi les plus connues on peut citer :

- la détermination de la dérive spatio-temporelle dans le processus de dépôt de polluants atmosphériques (Bilonick, 1985; Rouhani et al., 1992; Vyas et Christakos, 1997) ;
- la caractérisation de la variabilité spatio-temporelle des paramètres géophysiques de la Terre (Bogaert et Christakos, 1997) ;
- la modélisation de l'évolution temporelle de la teneur en eau du sol (Goovaerts et Sonnet, 1993) ;
- l'estimation des précipitations (Rouhani et Wackernagel, 1990) ;
- l'utilisation des modèles spatio-temporels de maladies pour estimer l'exposition des êtres humains aux polluants atmosphériques (Christakos et Hristopulos, 1998) ;
- la conception des réseaux pour améliorer la surveillance des processus atmosphériques spatio-temporels (Mejia et Rodriguez-Iturbe, 1974; Mardia et Goodall, 1993).

3.3 Cadre spatio-temporel pour les processus stochastiques

Les modèles spatio-temporels géostatistiques fournissent le cadre probabiliste pour analyser les données et effectuer l'estimation. Les modèles stochastiques de ce genre sont basés sur un petit

nombre de paramètres qui peuvent être inférés à partir des données et qui ne prennent pas en compte directement les équations différentielles sous-jacentes qui définissent le processus étudié. Le bruit ajouté dans ces modèles est, la plupart du temps, une mesure de notre ignorance sur les paramètres inconnus, en attendant qu'un modèle déterministe plus élaboré soit utilisé.

Généralement, les modèles stochastiques utilisent les données qui sont placées dans un espace spatio-temporel $\mathbf{D} \times \mathbf{T}$, où $\mathbf{D} \subseteq \mathbb{R}^2$ et $\mathbf{T} \subseteq \mathbb{R}^1$. Une variable aléatoire spatio-temporelle $Z(s,t)$ est donc une variable à laquelle on attribue une valeur en tout point de l'espace $\mathbf{D}$ et à n'importe quel moment dans le temps $\mathbf{T}$, suivant une certaine distribution spatio-temporelle.

L'analyse quantitative rigoureuse d'un processus naturel spatio-temporel repose sur le concept de continuum spatio-temporel. Il existe plusieurs interpretations de ce concept. Premièrement, on peut regarder ce continuum comme une extension quadri-dimensionnelle qui permet d'effectuer le calcul différentiel et toute autre opération mathématique, mais qui n'a pas de signification physique. Il constitue juste le cadre pour arranger les événements physiques qui nous intéressent. Deuxièmement, on peut considérer ce concept comme étant un système de coordonnées qu'on peut utiliser pour modéliser nos processus, en passant d'un système spatial bi- ou tri-dimensionnel à un autre, dans lequel on inclut aussi l'axe du temps.

D'un point de vue qualitatif, il existe quand même des différences remarquables entre les axes spatial et temporel. Les données temporelles peuvent être ordonnées : passé, présent et futur, tandis que, d'un point de vue spatial, on ne peut pas définir une telle succession. Le concept d'isotropie est bien défini dans le cadre spatial, mais il n'a aucun sens dans le cadre spatio-temporel. Les échelles sont elles aussi différentes.

Le concept de "champ" est parfaitement adapté à la représentation des phénomènes naturels et environnementaux continus. Un champ continu repose notamment sur des points-échantillons et la subdivision de l'espace en cellules. Il peut être représenté par un ensemble de cellules contenant des points-échantillons mesurés explicitement et, par des méthodes d'interpolation, on peut estimer des valeurs sur les points non échantillonnés.

Dans cette étude on travaillera sur un espace tri-dimensionnel $\mathbf{D} \times \mathbf{T}$ discrètisé en $N \times T$ points, où N et T représentent le nombre de points dans les domaines spatial et respectivement temporel. Pour chaque ensemble de NT points qui correspond au vecteur aléatoire $\{z(s_1,t_1),...,z(s_N,t_T)\}$ on peut définir une fonction aléatoire $Z(s,t)$ caractérisée par sa loi spatio-temporelle. L'inférence de cette loi repose sur des réalisations successives dudit vecteur aléatoire. Malheureusement, celles-ci ne sont pas disponibles pour chaque location espace-temps $(s,t) \in \mathbf{D} \times \mathbf{T}$, donc l'inférence classique consiste à considerer uniquement les paires séparées par le même vecteur espace-temps $(h,\tau) \in \mathbf{D} \times \mathbf{T}$. Un vecteur aléatoire s'appelle strictement stationnaire sur le domaine $\mathbf{D} \times \mathbf{T}$ si sa loi spatio-temporelle est invariante à la translation. Une hypothèse souvent faite dans les modélisations spatio-temporelles est que la fonction aléatoire $Z(s,t)$ est stationnaire d'ordre 2, ce qui concerne uniquement les moments d'ordre 1 et 2 : la moyenne est considérée constante sur tout le domaine spatio-temporel, tandis que la fonction de covariance dépend uniquement des vecteurs de

translation dans l'espace, $h = s - s'$, et dans le temps, $\tau = t - t'$.

Dans les sciences environnementales il existe deux approches de modélisation :

- La première consiste à décomposer la fonction aléatoire en deux parties : une tendance qui tient compte de la variabilité lisse du processus spatio-temporel, et une composante résiduelle qui modélise les fluctuations autour de cette tendance dans l'espace et dans le temps.
- La deuxième consiste à traiter le processus soit comme un ensemble de vecteurs aléatoires spatiaux, soit comme un ensemble de séries temporelles. On a donc affaire soit à une collection finie (T membres) de fonctions aléatoires spatiales $Z(s)$ corrélées dans le temps, soit à une collection finie (N membres) de séries temporelles corrélées spatialement.

3.4 L'état de l'art

En ingénierie environmentale, la plupart de recherches sont focalisées sur l'estimation spatio-temporelle de la concentration d'un polluant. Pour obtenir des estimations correctes, on est amené à utiliser l'une de techniques suivantes : l'estimation spatiale basée sur les modèles de covariance moyennées dans le temps, le cokrigeage comme méthode multivariable pour interpoler les données manquantes, ainsi que l'estimation basée sur des modèles de covariance spatio-temporels qui respectent certaines conditions de permissibilité ou d'admissibilité.

Il existe six modèles stochastiques qui ont été utilisés précédemment dans diverses études pour modéliser les processus spatio-temporels, modèles qui seront présentés par la suite.

1. Le premier modèle, le plus simple, est un modèle de type **moyenne temporelle**, utilisé par Bilonick (1983) et Switzer (1988). On considère les données enregistrées dans l'espace spatio-temporel $\mathbb{R}^2 \times \mathbf{T}$ et pour un instant t_k on utilise la fonction de covariance spatiale suivante :

$$C(h, t_k) = \frac{1}{N(h, t_k)} \sum_{i=1}^{N(h,t_k)} [x(s_i, t_k) - m_k][x(s_i + h, t_k) - m_k], \tag{3.1}$$

où $N(h, t_k)$ désigne le nombre des paires de données séparées par un vecteur spatial de dimension h au moment t_k, et m_k est la moyenne des données disponibles au moment t_k. On suppose que cette fonction de covariance ne varie pas dans le temps et alors on prend, comme fonction de covariance générale, la moyenne spatiale expérimentale par rapport aux séries temporelles. Donc, la principale hypothèse de ce modèle est la stationnarité temporelle et, implicitement, l'absence d'une quelconque continuité spatio-temporelle.

2. Le deuxième modèle, appliqué par Berkowitz (1992) pour estimer des données hydrologiques, est appelé tendance mobile (moving trend). Il consiste en une séparation entre la tendance (ou la dérive) et le bruit à l'aide de l'équation :

$$H(x, y, t) = \phi(x, t|\Theta_t) + \varepsilon(x, y, t), \tag{3.2}$$

où $\varepsilon(x, y, t)$ représente le bruit blanc, tandis que $\phi(x, t|\Theta_t)$ désigne la dérive, dérive qui dépend des coordonnées spatiales et temporelles. L'estimation, à chaque pas de temps, est réalisée en utilisant uniquement des données contemporaines, indiquant ainsi l'absence d'une relation entre les informations disponibles à chaque instant, ainsi que l'absence d'une continuité spatio-temporelle.

3. Le troisième modèle appliqué par Rouhani et Wackernagel (1990) et par Rouhani et al. (1992) fait appel au **cokrigeage**. Le variogramme croisé entre les locations i et j est exprimé comme une combinaison entre les variogrammes construits pour des échelles de temps différentes, tandis que le variogramme à une échelle temporelle spécifique est donné par un modèle élémentaire, avec un coefficient qui sera déterminé à l'aide d'une analyse en composantes principales. Dans ce type de modèles, la continuité spatio-temporelle n'est que partiellement prise en compte.
   Les variations spatio-temporelles complexes peuvent être simplifiées en les considérant comme produit ou somme de variations spatiales et/ou temporelles, ce qui constitue la base de trois autres approches qui seront présentées en fin de cette section. Tout d'abord, on présente deux types de simplifications possibles.
   Ainsi, dans une étude réalisée par Mejia et Rodriguez-Iturbe (1974), le champ aléatoire considéré comme homogène et stationnaire peut être décrit sous la forme :

$$Z(s, t) = \sum_{i=1}^{N} X_i(s) Y_i(t), \tag{3.3}$$

   où $X_i(s)$ représentent des fonctions aléatoires spatiales et $Y_i(t)$ des fonctions aléatoires temporelles. La fonction de covariance correspondante devient alors un modèle séparable :

$$C(h, \tau) = C_s(h) C_t(\tau), \tag{3.4}$$

   où h désigne le vecteur de la distance spatiale, et τ l'intervalle de temps.
   Une deuxième possibilité est de représenter la fonction de covariance spatio-temporelle par une somme de deux termes : le premier spatial et le deuxième temporel Bilonick (1985).

4. Cette simplification conduit à un quatrième modèle, dans lequel **les variables spatiale et temporelle sont considérées séparément**, donc il n'existe aucune interaction entre les deux structures de corrélation (temporelle et spatiale) ; les deux sont prises en compte, mais la continuité spatio-temporelle est simplifiée. Dimitrakopoulos et Luo (1994) ont montré que, dans certains cas, le modèle de type somme peut conduire à un système de krigeage singulier.

5. Le cinquième modèle, dans lequel la continuité spatio-temporelle est limitée, est appelé **voisinage glissant** et il est appliqué uniquement quand les variations sont uni-directionnelles et de vitesse constante.

6. Enfin, le dernier modèle, le S/TRF (en anglais-SpatioTemporal Random Field), le plus complexe, est celui qui prend en compte totalement la continuité spatio-temporelle. Il a été décrit rigoureusement par Christakos (1992).

Mis à part les modèles géostatistiques énoncés antérieurement, à partir de l'an 2000 ont été développés et appliqués des modèles essayant d'approcher les processus spatio-temporels non-stationnaires complexes comme :

- l'approche de type maximum de vraisemblance ;
- l'estimation hierarchique bayesienne ;
- l'application du Filtre de Kalman.

3.5 Le modèle spatio-temporel S/TRF

Dans cette étude on utilise le modèle spatio-temporel aléatoire (S/TRF-en anglais Spatio-Temporal Random Field) décrit et développé par Christakos (1992). La caractérisation de la continuité dans l'espace et dans le temps est l'un des aspects les plus importants de la modélisation S/TRF ; cette continuité est approchée par le variogramme spatio-temporel expérimental. Les modèles permissibles de covariance/variogramme spatio-temporels sont approchés par des critères de permissibilité appropriés au processus spatio-temporel.

L'estimation d'un processus spatio-temporel est développée en termes de krigeage spatio-temporel. Le point clé est l'analyse de la singularité du système de krigeage spatio-temporel. L'influence de la fonction de covariance, des tendances et des configurations des données sur la singularité du système sera discutée dans cette section. De plus, l'invariance tensorielle du système de krigeage spatio-temporel est analysée en termes de tendances dans l'espace et dans le temps.

Il existe quelques conditions pour pouvoir construire un modèle spatio-temporel cohérent avec la réalité physique (Christakos, 1992).

- Premièrement, le modèle doit refléter d'une façon adéquate les évolutions macroscopiques et microscopiques qui caractérisent le processus étudié. De la même façon, les échelles de variabilité spatiale et temporelle doivent être correctement représentées.
- Deuxièmement, dû au caractère aléatoire de la variabilité des données au niveau microscopique, ces processus doivent être considérés comme stochastiques, l'incertitude étant intrinsèquement liée à l'evolution spatio-temporelle et pas à la description statistique des réalisations possibles.
- Troisièmement, le modèle doit être capable d'estimer quantitativement n'importe quel processus en termes de variabilité spatiale non-homogène et de variabilité temporelle non-stationnaire.

D'un point de vue mathématique, un S/TRF est une collection de réalisations d'un processus naturel respectant une certaine distribution spatio-temporelle. Dans ce modèle, l'espace et le temps forment une combinaison complexe avec des effets simultanés et liés à l'evolution spatio-temporelle. Le domaine spatio-temporel est construit comme le produit cartésien $\mathbb{R}^2 \times \mathbf{T}$, avec $(s,t) \in \mathbb{R}^2 \times \mathbf{T}$ les coordonnées spatio-temporelles, et $Z(s,t)$ un S/TRF défini sur $\mathbb{R}^2 \times \mathbf{T}$ avec des valeurs réelles positives.

On présente d'abord quelques notions fondamentales sur la **métrique spatio-temporelle** utilisée dans cette étude. Parmi les caractéristiques quantitatives de la géométrie spatio-temporelle, une place centrale est occupée par la **métrique** de cette géométrie, c'est-à-dire l'ensemble de relations mathématiques définissant les distances. Une métrique spatio-temporelle est donc une fonction définie pour un système de coordonnées spatio-temporelles, telle que la distance spatio-temporelle entre deux points du système est déterminée par les coordonnées de ces deux points. Il existe au moins deux structures possibles : une *séparée* et l'autre *composée*. La première traite les concepts de distances séparément en espace et en temps $dp = (|ds|, dt)$, où, par exemple, la distance spatiale peut prendre une forme euclidienne $|ds| = \sqrt{\sum_{i=1}^{n} ds_i^2}$, et dt est l'intervalle de temps entre les deux points considérés. La métrique composée est plus compliquée, dans le sens que les paramètres spatiaux et temporels sont liés par des expressions analytiques, comme par exemple la métrique Riemanienne spatio-temporelle définie par la formule :

$$|dp| = \sqrt{\sum_{i,j=1}^{n} g_{ij} ds_i ds_j + g_{00} dt^2 + 2dt \sum_{i=1}^{n} g_{0i} ds_i} \tag{3.5}$$

où les coefficients g_{ij} sont des fonctions qui dépendent de l'espace et du temps. Dans ce travail, on a utilisé uniquement la **métrique séparée**, dans laquelle la distance spatiale est **euclidienne**.

On décrit la continuité spatio-temporelle par une covariance expérimentale spatio-temporelle et on fait des hypothèses qui reflètent les caractéristiques des données spatio-temporelles. On définit un S/TRF ordinaire (OS/TRF) $X(s,t)$ comme une application du domaine spatio-temporel $\mathbb{R}^2 \times \mathbf{T}$ avec des valeurs dans un espace Hilbert noté $L_2(\Omega, F, P)$:

$$X(s,t) : \mathbb{R}^2 \times \mathbf{T} \to L_2(\Omega, F, P).$$

Une caractérisation incomplète mais satisfaisante d'un S/TRF $X(s,t)$ est donnée par l'intermédiaire de deux premiers moments :

- l'espérance mathématique comme fonction d'espace et de temps :

$$m_x(s,t) = E[X(s,t] = \int_{-\infty}^{\infty} \chi dF(\chi),$$

  où $F(\chi)$ représente la fonction de probabilité de $X(s,t)$ telle que : $F(\chi) = P[\chi \leq X(s,t)]$;

- la covariance spatio-temporelle :

$$\begin{aligned} C_x(s,t,s',t') &= E[X(s,t) - m_x(s,t)][X(s',t') - m_x(s',t')] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\chi - m)(\chi' - m') dF(\chi, \chi'); \end{aligned}$$

- le variogramme spatio-temporel défini comme :

$$\begin{aligned} \gamma_x(s,t,s',t') &= \frac{1}{2} E[X(s,t) - X(s',t')]^2 \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\chi - \chi')^2 dF(\chi, \chi'). \end{aligned}$$

Dans le cas le plus général, un (OS/TRF) représente un processus naturel non-stationnaire et non-homogène. Un exemple très simple est donné par les processus de la forme :

$$X(s,t) = Y(s,t) + p_{\nu,\mu}(s,t), \tag{3.8}$$

où $Y(s,t)$ représente un (OS/TRF) qui est stationnaire et homogène, et $p_{\nu,\mu}(s,t)$ est un polynôme de degré ν en s et μ en t.

La moyenne exprime la tendance spatio-temporelle existante dans les données, et la fonction de covariance exprime les corrélations et les dépendances entre les points/moments dans l'espace/temps. Un cas particulier, très favorable, est celui où on considère le champ comme étant homogène spatialement et stationnaire d'un point de vue temporel. Cela veut dire que la moyenne est constante et la covariance est calculée avec la formule :

$$C_x(s,t;s',t') = C_x(h,\tau),$$

où $h = s - s'$ et $\tau = t - t'$.

Comme dans le cas spatial, il existe une relation mathématique entre les deux fonctions covariance et variogramme dans le cas de stationnarité/homogénéité mentionné auparavant :

$$C_x(h,\tau) = C_x(\mathbf{0},0) - \gamma_x(h,\tau).$$

3.6 Caractérisation de la continuité spatio-temporelle

La continuité spatio-temporelle est une propriété qui caractérise la relation entre les données situées en différents endroits du domaine spatio-temporel.

La caractérisation de la continuité spatio-temporelle doit inclure :

- La description de la continuité spatio-temporelle par l'intermédiaire de la covariance ou du variogramme expérimental(e) pour obtenir les bases de test des hypothèses sur la continuité spatio-temporelle et pour ajuster les modèles de covariance/variogramme.

- L'investigation des hypothèses effectuées dans la modélisation de la continuité spatio-temporelle qui peuvent refléter et résumer les caractéristiques des covariances/variogrammes expérimentaux.

- L'ajustement d'un modèle permissible de covariance/variogramme pour caractériser la continuité spatio-temporelle d'un processus naturel.

3.6.1 Description de la continuité spatio-temporelle

La continuité spatio-temporelle peut être décrite par l'intermédiaire de la covariance/variogramme spatio-temporel(le) expérimental(e). La formule de calcul pour la covariance expérimentale est :

$$C^*(h,\tau) = \frac{1}{N_{(h,\tau)}} \sum_{i=1}^{N_{(h,\tau)}} [x(s_i,t_i) - m^*][x(s_i+h,t_i+\tau) - m^*] \tag{3.11}$$

et la formule du variogramme expérimental :

$$\gamma^*(h,\tau) = \frac{1}{2N_{(h,\tau)}} \sum_{i=1}^{N_{(h,\tau)}} [x(s_i,t_i) - x(s_i+h,t_i+\tau)]^2 \tag{3.12}$$

où m^* représente la moyenne des données et $N_{(h,\tau)}$ le nombre de paires séparées par le vecteur de distance h et l'intervalle de temps τ. Une première description est donc réalisée en utilisant le variogramme basé sur le vecteur spatio-temporel de séparation (h,τ). Le comportement du variogramme au voisinage de l'origine renseigne sur les caractéristiques de régularité, tandis que celui à des très larges distances renseigne sur l'homogénéité par rapport à la continuité spatio-temporelle.

3.6.2 Les hypothèses de la continuité spatio-temporelle

Pour caractériser la continuité spatio-temporelle d'un processus naturel, il est nécessaire de faire des hypothèses sur les caractéristiques du variogramme expérimental. L'étude de la continuité spatio-temporelle peut être vue comme un pont entre les variogrammes expérimentaux et les modèles existants de variogrammes. D'un côté, ceci peut résumer les caractéristiques d'un variogramme expérimental et, d'un autre côté, peut produire les bases de sélection pour un modèle approprié d'un variogramme.

Les hypothèses sur la continuité spatio-temporelle impliquent :

- **L'homogénéité**. La portée du variogramme détermine la zone d'influence d'une valeur d'un processus naturel. Le comportement du variogramme aux larges distances montre le degré d'homogénéité du processus.
- **L'anisotropie**. On parle d'anisotropie quand les variogrammes expérimentaux calculés présentent des différences marquantes selon la direction de calcul. Une anisotropie qui peut être réduite à l'isotropie par une simple transformation linéaire des coordonnées spatiales s'appelle anisotropie géométrique. Autrement, on a une anisotropie zonale qui est plus compliquée à modéliser.
- **La régularité** est déterminée par le comportement du variogramme à l'origine. Comme dans le cas spatial, on peut rencontrer trois types de comportement à l'origine : parabolique, linéaire et l'effet de pépite. Ce dernier caractérise le saut du variogramme à l'origine (discontinuité aux petites distances) et ceci peut être le résultat d'une erreur de mesure ou de représentativité.
- **La séparabilité**. Si les allures des variogrammes calculés pour des valeurs distinctes de τ (temps fixé) ne présentent pas des différences notables, on peut envisager de considérer un modèle de variogramme séparable, comme par exemple un modèle de type produit ou une combinaison somme-produit entre deux composantes l'une spatiale et l'autre temporelle.

3.6.3 Les fonctions spatio-temporelles de covariance/variogramme

L'étude des modèles pour les fonctions de covariance/variogramme clôt le volet dédié à la continuité spatio-temporelle. Dans la plupart des cas, cette étape est d'ailleurs la plus importante pour pouvoir effectuer une estimation stochastique spatio-temporelle correcte.

3.6.3.1 Les critères de permissibilité

La condition principale d'une estimation stochastique spatio-temporelle est de s'assurer que la variance estimée est toujours non-négative. Pour cela, on étudie les critères de permissibilité qui garantissent que cette condition est vérifiée.

Définition Une fonction continue $C(s,t;s',t')$ est une fonction de covariance pour un OS/TRF $X(s,t)$ si et seulement si elle est de **type non-négatif**, c'est-à-dire :

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j C(s_i,t_i;s'_j,t'_j) \geq 0 \tag{3.13}$$

pour tous les entiers $N \geq 1$, tous les (s_i,t_i) et $(s'_j,t'_j) \in \mathbb{R}^2 \times \mathbf{T}$ et n'importe quels réels a_i, a_j.

De la même façon, une fonction $\gamma(s,t;s',t')$ représente une fonction de variogramme pour un OS/TRF $X(s,t)$ si et seulement si elle est une fonction de **type négatif conditionnel** :

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \gamma(s_i,t_i;s'_j,t'_j) \leq 0, \;\; \forall \sum_{i=1}^{N} a_i = 0. \tag{3.14}$$

Par définition, $-\gamma(s_i,t_i;s'_j,t'_j)$ est appelée fonction de type positif conditionnel.

Une remarque très importante concerne la validité d'une fonction de covariance/variogramme dans un espace d'une certaine dimension. Si la fonction est admissible dans un espace d'une certaine dimension, alors elle est admissible aussi dans un espace de dimension inférieure (Christakos, 1992). La réciproque est fausse.

L'éventail des modèles de covariance admissibles peut être élargi grâces aux propriétés suivantes :

- n'importe quelle combinaison de fonctions de type positif avec des coefficients positifs est, à son tour, une fonction de type positif, donc éligible ;
- tout produit de fonctions de type positif est aussi de type positif.

Ces propriétés restent valables aussi pour les fonctions de type négatif.

Pour vérifier la permissibilité d'une fonction candidate à une fonction de covariance/variogramme, on a deux critères.

Premier critère : Si $\gamma(h,\tau)$ est un variogramme spatio-temporel permissible, alors il doit satisfaire :

$$\lim \frac{\gamma(h,\tau)}{|h|^2+\tau^2} = 0 \tag{3.15}$$

quand $|h| \to \infty$ et $\tau \to \infty$.

Deuxième critère : Si $\gamma(h,\tau)$ est un variogramme spatio-temporel permissible, alors la fonction $\exp[-\alpha\gamma(h,\tau)]$ doit être de type positif pour toutes les valeurs de α.

Exemple : Un exemple de fonction de variogramme admissible en $\mathbb{R}^2 \times \mathbf{T}$ est :

$$\gamma(h,\tau) = \sqrt{|h|^2+\tau^2}, \tag{3.16}$$

car la fonction $\exp[-\alpha\gamma(h,\tau)]$ est de type positif pour tout α.

Définition Une fonction de covariance $C(h,\tau)$ est de **type strictement positif** si elle vérifie la formule :

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j C(h_{ij},\tau_{ij}) > 0 \tag{3.17}$$

pour toutes les valeurs réelles a_i non-nulles simultanément. De la même façon, le modèle de variogramme est de **type strictement négatif conditionnel** si :

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \gamma(h_{ij},\tau_{ij}) < 0, \quad \forall \sum_{i=1}^{N} a_i = 0. \tag{3.18}$$

Cette propriété, d'être de type strictement positif (respectivement strictement négatif) conditionnel, est très importante dans le contexte du système de krigeage spatio-temporel, car elle conduit à une solution unique, ce qui est d'une grande importance pratique dans le cas d'une estimation spatio-temporelle.

3.6.3.2 Les modèles spatio-temporels de covariance/variogramme

A. Une fonction de covariance $C(h,\tau)$ est séparable si elle vérifie l'équation 3.4. Par extension, pour une fonction de variogramme, un modèle séparable respecte la formule suivante :

$$\gamma(h,\tau) = C_s(\mathbf{0})\gamma_t(\tau) + \gamma_s(h)C_t(0) - \gamma_s(h)\gamma_t(\tau), \tag{3.19}$$

où $\gamma(h,\tau)$ représente le variogramme spatio-temporel, tandis que γ_t et γ_s désignent le variogramme temporel, respectivement spatial (De Cesare et al., 2001).

Les propriétés d'un tel modèle sont :

- La fonction $C(h,0)$, appelée fonction spatiale limite du modèle de covariance $C(h,\tau)$, reflète la structure spatiale marginale de cette fonction de covariance. D'une façon similaire, $C(0,\tau)$ est appelée fonction temporelle limite et correspond à la structure temporelle marginale. Les deux structures peuvent être complètement différentes.

- La relation entre les covariances avec des distances spatiales fixées ou des intervalles de temps fixés est une relation de proportionnalité :

$$\frac{C(h_i,\tau)}{C(h_j,\tau)} = \frac{C_s(h_i)C_t(\tau)}{C_s(h_j)C_t(\tau)} = \frac{C_s(h_i)}{C_s(h_j)} = \text{constant}, \quad \forall \tau \geq 0, \tag{3.20}$$

où h_i et h_j sont des distances spatiales fixées, ou bien :

$$\frac{C(h,\tau_i)}{C(h,\tau_j)} = \frac{C_s(h)C_t(\tau_i)}{C_s(h)C_t(\tau_j)} = \frac{C_t(\tau_i)}{C_t(\tau_j)} = \text{constant}, \quad \forall \mathbf{h} \geq 0 \tag{3.21}$$

où τ_i et τ_j sont des intervalles de temps fixés. Ceci indique que la structure spatiale de la covariance spatio-temporelle est invariante par rapport aux intervalles de temps, et que la structure temporelle ne varie pas par rapport aux distances spatiales.

Les modèles séparables correspondent à un cas particulier de S/TRF $X(s,t)$ qui est un produit d'un SRF $X_s(s)$ par un TRF $X_t(t)$ avec les deux composantes indépendantes :

$$X(s,t) = X_s(s)X_t(t). \tag{3.22}$$

B. Un deuxième modèle, plus général, de la fonction de covariance est de type non séparable :

$$C(h,\tau) = C[(a^2|h|^2 + b^2\tau^2)^{1/2}], \tag{3.23}$$

où $|h|^2 = h^T h$ et a et b sont des coefficients réels. Par conséquent, le modèle de variogramme correspondant est lui aussi de type non séparable :

$$\gamma(h,\tau) = C(0,0) - C[(a^2|h|^2 + b^2\tau^2)^{1/2}] = \gamma[(a^2|h|^2 + b^2\tau^2)^{1/2}]. \tag{3.24}$$

Un tel modèle correspond aux hypothèses suivantes : la covariance spatio-temporelle présente une structure uniforme sur le domaine spatio-temporel et, par conséquent, le type de structure spatiale marginale coïncide avec le type de structure temporelle marginale, même si les deux coefficients mentionnés a et b qui apparaissent dans la formule sont différents.

3.7 Le krigeage spatio-temporel

L'idée de base du krigeage spatio-temporel est la même que celle du krigeage spatial, c'est-à-dire qu'on veut prévoir la valeur de la variable régionalisée étudiée en un site non échantillonné $X(s_0,t_0)$ par une combinaison linéaire de données ponctuelles adjacentes disponibles (sauf que cette fois-ci on dispose aussi de données temporelles antérieures ou postérieures au moment de l'estimation) :

$$\widehat{X}(s_0,t_0) = \sum_{i=1}^{N} \lambda_i X(s_i,t_i). \tag{3.25}$$

Les poids λ_i associés à chacune des valeurs régionalisées observées sont choisis de façon à obtenir une estimation non biaisée :

$$E\left[\widehat{X}(s_0,t_0) - X(s_0,t_0)\right] = 0 \tag{3.26}$$

et une variance minimale :

$$\sigma_k^2 = \min\ E\left[\widehat{X}(s_0,t_0) - X(s_0,t_0)\right]^2 \tag{3.27}$$

Ils dépendent de l'emplacement des observations et de leur structure de dépendance spatiale. Par la suite, on détaillera, comme dans le cas spatial, les principales méthodes de krigeage en fonction des hypothèses de stationnarité faites sur la variable régionalisée à estimer.

3.7.1 Le krigeage simple spatio-temporel (KSS/T)

Dans ce cas, on considère le champ aléatoire $X(s,t)$ comme homogène et stationnaire avec une **moyenne constante connue** m (pas forcément nulle), comme dans le cas purement spatial. On travaille avec l'estimateur suivant :

$$\widehat{X}(s_0,t_0) = m + \sum_{i=1}^{N} \lambda_i [X(s_i,t_i) - m]. \tag{3.28}$$

La variance de l'estimation peut être calculée en utilisant la formule :

$$E[\widehat{X}(s_0,t_0) - X(s_0,t_0)]^2 = C(\mathbf{0},0) - 2\sum_{i=1}^{N} \lambda_i C(h_{0i},\tau_{0i}) + \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j C(h_{ij},\tau_{ij}) \tag{3.29}$$

et sa minimisation par la méthode des multiplicateurs de Lagrange conduit au système de KSS/T suivant :

$$\sum_{i=1}^{N} \lambda_i C(h_{ij},\tau_{ij}) = C(h_{0j},\tau_{0j}), \;\; j = 1,..,N. \tag{3.30}$$

Ce système s'écrit sous forme matricielle :

$$\mathbf{C}\lambda = \theta \tag{3.31}$$

où $\mathbf{C}$ représente la matrice de covariance, λ est le vecteur de poids et θ représente le vecteur de covariance qui se trouve dans le membre droit de l'équation 3.30.

La solution de ce système est :

$$\lambda = \mathbf{C}^{-1}\theta. \tag{3.32}$$

La variance de krigeage simple s'écrit :

$$\sigma^2_{SK}(s_0,t_0) = C(h_{00},\tau_{00}) - \sum_{i=1}^{N} \lambda_i C(h_{0i},\tau_{0i}). \tag{3.33}$$

Une première remarque évidente qu'on peut faire par rapport à l'équation 3.32 est que le système de krigeage a une solution unique si et seulement si la matrice $\mathbf{C}$ n'est pas singulière, ce qui implique en fait que la fonction de covariance doit être de type strictement positif.

3.7.2 Le krigeage ordinaire spatio-temporel (KOS/T)

Pour effectuer un krigeage ordinaire spatio-temporel on considère la variable régionalisée comme homogène et stationnaire présentant une **moyenne constante inconnue** notée m. L'estimateur KOS/T s'écrit alors :

$$\widehat{X}(s_0,t_0) = \sum_{i=1}^{N} \lambda_i X(s_i,t_i). \tag{3.34}$$

La condition d'absence de biais devient : $\sum_{i=1}^{N} \lambda_i = 1$ et le système de krigeage ordinaire s'écrit sous forme détaillée :

$$\begin{cases} \sum_{i=1}^{N} \lambda_i C(h_{ij}, \tau_{ij}) - \mu = C(h_{0j}, \tau_{0j}), \;\; j = 1, .., N \\ \sum_{i=1}^{N} \lambda_i = 1 \end{cases} \tag{3.35}$$

ou bien, sous forme matricielle :

$$\mathbf{K}\beta = \Theta, \tag{3.36}$$

où β représente le vecteur de poids : $\beta^T = (\lambda^T, \mu)$, Θ est le vecteur représentant le membre droit de l'équation 3.35 : $\Theta^T = (\theta^T, 1)$, et $\mathbf{K}$ est la matrice de krigeage ordinaire :

$$\mathbf{K} = \begin{pmatrix} \mathbf{C} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \tag{3.37}$$

Dans cette formule, $\mathbf{1}$ désigne le vecteur dont tous les éléments valent 1, tandis que $\mathbf{C}$ est la même matrice de covariance que celle mentionnée dans le cas du krigeage spatio-temporel simple. Le vecteur β s'obtient facilement par :

$$\beta = \mathbf{K}^{-1}\Theta. \tag{3.38}$$

La variance du krigeage ordinaire peut s'écrire :

$$\sigma^2_{OK}(s_0, t_0) = C(\mathbf{0}, 0) + \mu - \sum_{i=1}^{N} \lambda_i C(h_{0i}, \;\; \tau_{0i}). \tag{3.39}$$

Comme dans le cas du krigeage simple, on peut faire le même genre de remarque en regardant l'équation 3.38, c'est-à-dire que le système de krigeage a une solution unique si et seulement si la matrice $\mathbf{K}$ du système est non-singulière, et cela peut se traduire mathématiquement par la relation :

$$|\mathbf{K}| = |\mathbf{C}| \cdot |0 - \mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}| \neq 0 \tag{3.40}$$

où $|\;|$ représente le déterminant. Par conséquent, le système de krigeage ordinaire a une solution unique si la fonction de covariance est de type strictement positif.

3.7.3 Le krigeage universel spatio-temporel (KUS/T)

Le krigeage spatio-temporel universel est une technique d'estimation d'un S/TRF $X(s,t)$ non-homogène et non-stationnaire avec une dérive de forme bien déterminée. En supposant que la dérive est donnée par le modèle :

$$m(s,t) = \sum_{j=1}^{L} \alpha_j f_j(s,t) = \mathbf{f}^T\alpha, \tag{3.41}$$

où $\mathbf{f}^T = [f_1(s,t), ..., f_L(s,t)]$ est un vecteur dont les composantes sont des fonctions connues, tandis que $\alpha = [\alpha_1, ..., \alpha_L]$ représente un vecteur de coefficients inconnus.

L'estimateur KUS/T est utilisé donc pour estimer une valeur en (s_0, t_0), en utilisant une expression linéaire :

$$\widehat{X}(s_0, t_0) = \sum_{i=1}^{N} \lambda_i X(s_i, t_i). \tag{3.42}$$

On impose les contraintes habituelles, comme le non-biais :

$$\sum_{i=1}^{L} \lambda_i f_j(s_i, t_i) = f_j(s_0, t_0), \quad j = 1, ..., L. \tag{3.43}$$

Dans ces conditions, le système de krigeage universel s'écrit :

$$\begin{cases} \mathbf{C}\lambda - \mathbf{F}\mu = \mathbf{C}_0, \\ \mathbf{F}^T\lambda = \mathbf{F}_0 \end{cases} \tag{3.44}$$

avec les vecteurs : $\lambda^T = [\lambda_1, ..., \lambda_N]$, $\mu^T = [\mu_1, ..., \mu_L]$, $C_0^T = [C(h_{01}, \tau_{01}), ..., C(h_{0N}, \tau_{0N})]$, $F_0^T = [f_1(s_0, \tau_0), ..., f_L(h_0, \tau_0)]$ et la matrice $\mathbf{F} = [(F_{ij})] = [f_j(s_i, t_i)]$.

Sous forme matricielle, le système du krigeage devient :

$$\mathbf{K}\beta = \mathbf{\Theta}, \tag{3.45}$$

où β est le vecteur des poids $\beta_T = (\lambda^T, \mu^T)$, $\mathbf{\Theta}$ est le vecteur du membre droit de l'équation 3.44 ($\Theta = (\theta_T, F_0^T)$) et $\mathbf{K}$ la matrice de krigeage universel :

$$\mathbf{K} = \begin{pmatrix} \mathbf{C} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{pmatrix}, \tag{3.46}$$

où $\mathbf{C}$ représente toujours la matrice de covariance.

Le vecteur des poids s'obtient à partir de 3.46 :

$$\beta = \mathbf{K}^{-1}\Theta \tag{3.47}$$

et la variance du krigeage universel s'exprime sous la forme :

$$\sigma_{UK}^2(s_0, t_0) = C(\mathbf{h}_{00}, \tau_{00}) - \mathbf{C}_0^T\lambda + \mathbf{F}_0^T\mu. \tag{3.48}$$

3.7.3.1 La dérive spatio-temporelle

Le vecteur f^T, qui représente la dérive, inclut un ensemble de fonctions caractérisées par deux propriétés : la première vise l'indépendance linéaire pour s'assurer qu'on obtient une solution unique du système du krigeage universel, tandis que la deuxième concerne l'invariance tensorielle par rapport au changement de l'origine/unité du système de coordonnées, afin d'assurer la même condition d'unicité.

Il existe plusieurs formes analytiques pour la dérive, parmi lesquelles on peut citer : des fonctions polynomiales, des expressions de type Fourier, ou bien une combinaison des deux. Par un souci de clarté, le cas particulier de dérive dans un espace $\mathbb{R}^2 \times \mathbf{T}$ sera présenté, mais ce cas peut se généraliser facilement à d'autres dimensions.

Voici quelques exemples pour les dérives énumérées précédemment.

Premièrement, la forme polynomiale de la dérive est donnée sous forme condensée :

$$f^T = [1, x, y, t, ..., x^\xi, y^\xi, t^\zeta], \tag{3.49}$$

où ξ et ζ sont les ordres (degrés) dans l'espace et dans le temps.

Deuxièmement, la forme de type Fourier utilisée pour la dérive est :

$$f^T = [1, \sin\omega_x x \cdot \sin\omega_y y \cdot \sin\omega_t t, \sin\omega_x x \cdot \sin\omega_y y \cdot \cos\omega_t t, ... \\ \cos\omega_x nx \cdot \cos\omega_y ny \cdot \sin\omega_t nt, \cos\omega_x nx \cdot \cos\omega_y ny \cdot \cos\omega_t nt] \tag{3.50}$$

où ω_x, ω_y et ω_t sont les fréquences et n représente l'ordre de la série de Fourier.

La dernière forme, mixte, peut être générée par un modèle séparable du genre

$$m(s,t) = m(s)m(t), \tag{3.51}$$

où $m(s)$ représente la tendance spatiale, tandis que $m(t)$ est la tendance temporelle. Un exemple de dérive mixte est donné par la formule suivante :

$$f^T = [1, \sin\omega_t t, \cos\omega_t t, x\sin\omega_t t, x\cos\omega_t t, y\sin\omega_t t, y\cos\omega_t t, ...]. \tag{3.52}$$

Le seul modèle de tendance qui sera utilisé est celui polynômial. Pour l'utiliser on a besoin de connaître et d'appliquer un certain critère d'admissibilité :

Critère. Une dérive d'une forme polynomiale d'ordre ξ/ζ remplit la condition d'invariance tensorielle si et seulement si tous les termes d'ordre inférieur ou égal à $(\xi-1)/(\zeta-1)$ sont présents.

Exemple. La dérive ayant la formule $f^T = [1, x, y, x^2, y^2, t^2, xy, xt, yt]$ n'est pas acceptée, tandis que $f^T = [1, x, y, t, xy]$ respecte le critère d'invariance tensorielle et donc ce dernier modèle est admissible.

3.7.3.2 La singularité du système de krigeage spatio-temporel

Les problèmes de singularité apparaissent dans le krigeage universel suite à une éventuelle singularité de la matrice $\mathbf{K}$ du système :

$$\mathbf{K} = \begin{pmatrix} \mathbf{C} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{pmatrix}, \tag{3.53}$$

où $\mathbf{F}$ représente la matrice de tendance :

$$\mathbf{F} = \begin{pmatrix} f_1(s_1,t_1) & f_2(s_1,t_1) & ... & f_L(s_1,t_1) \\ f_1(s_2,t_2) & f_2(s_2,t_2) & ... & f_L(s_2,t_2) \\ ... & ... & ... & ... \\ f_1(s_N,t_N) & f_2(s_N,t_N) & ... & f_L(s_N,t_N) \end{pmatrix}. \tag{3.54}$$

Selon 3.53, la singularité de la matrice du krigeage $\mathbf{K}$ dépend de deux matrices : $\mathbf{C}$ et $\mathbf{F}$. Parfois, la singularité de $\mathbf{K}$ peut être due à l'impact de la matrice de tendance $\mathbf{F}$ sur la matrice $\mathbf{C}$. Quand la matrice $\mathbf{C}$ est définie positive, on peut voir que la singularité de $\mathbf{K}$ dépend uniquement

de la singularité de $\mathbf{F}$:

$$|\mathbf{K}| = |\mathbf{C}||0 - \mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}| = (-1)^L|\mathbf{C}||\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}|. \tag{3.55}$$

La matrice $\mathbf{F}$ devient singulière quand le nombre de points N est inférieur au nombre de termes de la dérive L (les dimensions de cette matrice sont exactement N et L). D'habitude, dans les cas pratiques, on a la relation $N \geq L$. Souvent, on cherche une dérive d'ordre pas trop élevé (inférieur à 2) pour les deux composantes spatiale et temporelle.

3.7.4 Le krigeage intrinsèque spatio-temporel (KIS/T)

La dernière méthode de krigeage présentée dans ce chapitre est le krigeage intrinsèque spatio-temporel, qui représente l'extension naturelle du krigeage intrinsèque spatial décrit dans le chapitre antérieur.

L'objet fondamental sur lequel on applique la théorie de krigeage spatio-temporel est un modèle très général de champ aléatoire spatio-temporel S/TRF. Intuitivement, l'idée de base est que la variabilité d'un S/TRF peut être représentée par l'intermédiaire des degrés d'écartement par rapport à l'homogénéité et à la stationnarité (Christakos et Vyas, 1998). La variation d'un champ modélisé par un S/TRF est caractérisée par la tendance spatio-temporelle et par des fluctuations irrégulières.

Cette théorie est basée sur un opérateur mathématique, noté Q, qui transforme le processus initial non-homogène et non-stationnaire $X(s,t)$ en un processus homogène et stationnaire, $Y(s,t)$.

$$X(s,t) \xrightarrow{Q} Y(s,t). \tag{3.56}$$

L'opérateur Q est construit de façon à posséder certaines propriétés très utiles, comme par exemple celle de filtrer la tendance spatio-temporelle ou celle purement spatiale, soit en contenant des composantes polynomiales spatio-temporelles de degrés ν en espace et μ en temps.

Cet opérateur est construit discrètement, à l'aide des coefficients $q(s_i,t_j)$ associés à chaque point s_i de l'espace et à chaque instant t_j, quand $i = 1,\ldots,m$ et $j = 1,\ldots,p_i$, où p_i désigne la taille de la séquence temporelle disponible à chaque emplacement s_i. L'équation 3.56 devient ainsi :

$$\sum_{i=1}^{m}\sum_{j=1}^{p_i} q(s_i,t_j)X(s_i,t_j) = Y(s,t), \tag{3.57}$$

où les coefficients $q(s_i,t_j)$ sont choisis tels que la condition suivante (3.58) soit satisfaite :

$$\sum_{i=1}^{m}\sum_{j=1}^{p_i} q(s_i,t_j)s_i^{\rho}t_j^{\zeta} = 0 \tag{3.58}$$

pour tout $\rho = |\rho| \leq \nu$ et $\zeta \leq \mu$, où $s_i^{\rho} = s_1^{\rho_1}\cdots s_n^{\rho_n}$.

Le champ original $X(s,t)$ est appelé un S/TRF d'ordres de continuité ν en espace et μ dans le temps et il est noté $S/TRF - \nu/\mu$. En pratique, l'opérateur Q agit comme une échelle spécifique

(voisinage) qui peut changer à l'intérieur du domaine spatio-temporel. L'équation 3.57 peut être vue comme une approximation avec des différences finies d'une équation différentielle, avec des dérivées partielles, qui décrit le modèle à l'intérieur du voisinage considéré. Cet opérateur $Q[X]$ peut être vu comme un filtre passe-haut qui augmente les détails et essaie d'annihiler les tendances, tandis que $X-Q[X]$ peut jouer le rôle d'un filtre passe-bas qui contient les tendances et les effets saisonniers.

Les deux ordres ν et μ s'appellent *ordres de continuité* spatio-temporelle car ils représentent, en moyenne, cette continuité dans le sens où ils caractérisent la complexité des fonctions décrivant la dérive dans la distribution spatio-temporelle. La distribution de la différence d'ordres de continuité $\nu - \mu$ donne des informations relatives aux tendances spatiale et/ou temporelle. Une différence positive implique une structure spatiale plus complexe et donc une dominante spatiale, tandis qu'une différence négative relève une plus grande complexité temporelle. Les deux ordres de continuité offrent aussi des informations sur le modèle stochastique sous-jacent. Ces paramètres déterminent l'ordre de l'opérateur Q, dans le sens où il faut savoir combien de données on peut utiliser dans l'espace et combien il faut retourner dans le temps pour chercher les informations nécessaires. La taille de ce voisinage donne des indications sur les rayons d'influence de chaque donnée utilisée dans l'estimation.

La théorie de S/TRF est basée sur la décomposition de la covariance générale qui est non-homogène et non-stationnaire, comme une somme de deux composantes :

$$c_x(s,t;s',t') = k_x(\mathbf{h},\tau) + p_{\nu,\mu}(s,t;s',t'), \tag{3.59}$$

où le premier terme, $k_x(h,\tau)$, désigne la partie homogène et stationnaire qui est appelé covariance spatio-temporelle généralisée (CGST) ($h = s - s'$ et $\tau = t - t'$), tandis que le deuxième est un polynôme de degrés ν en s et s' et μ en t et t'.

Une fonction de covariance généralisée spatio-temporelle isotrope du point de vue spatial et qui est très utilisée dans les applications atmosphériques est la suivante :

$$k_x(r,\tau) = c\delta(r)\delta(\tau) + \sum_{l=0}^{\mu}(-1)^{l+1}a_l\tau^{2l+1}\delta(r) + \sum_{i=0}^{\nu}(-1)^{i+1}b_i r^{2i+1}\delta(\tau) + \sum_{i=0}^{\nu}\sum_{l=0}^{\mu}(-1)^{i+l}a_{il}r^{2i+1}\tau^{2l+1} \tag{3.60}$$

où les coefficients impliqués dans cette équation respectent certaines conditions d'admisibilité (SANLIB, 1995).

Comme dans le cas spatial, le problème principal est l'inférence des coefficients mentionnés. Le modèle polynômial (3.60) de covariance généralisée se prête bien à une procedure entièrement automatisée d'inférence statistique, car les coefficients interviennent linéairement et sont donc faciles à estimer.

Le problème qu'on veut résoudre est le même : compte tenu des valeurs mesurées disponibles $X(s_i,t_j)$, avec $i = 1,2,\ldots,L$ et $j = 1,2,\ldots,M$, prévoir la valeur de la variable régionalisée à interpoler $X(.,.)$ en un site non échantillonné noté s_k ($k \neq i$) à l'instant t_l ($l \neq j$). Pour obtenir les équations du krigeage on fait appel à la démarche proposée par Christakos (1992), basée sur les

contraintes de linéarité, non-biais et optimalité.

1. **Contrainte de linéarité**
   La contrainte de base est que l'estimation prenne la forme d'une combinaison linéaire des données $X(s_i, t_j)$; elle s'écrit :

$$\widehat{X}(s_k, t_l) = \sum_{i=1}^{m} \sum_{j=1}^{p_i} \xi_{ij} X(s_i, t_j).$$

2. **Contrainte de non-biais**
   L'absence de biais se traduit par la relation : $E[\widehat{X}(s_k, t_l) - X(s_k, t_l)] = 0$ qui exprime le fait que l'erreur d'estimation est d'espérance nulle.

3. **Contrainte d'optimalité**
   La dernière contrainte à respecter vise la minimisation de la variance $Var[\widehat{X}(s_k, t_l) - X(s_k, t_l)]$.

   Finalement, la variance de l'estimation est calculée en utilisant la formule :

$$\sigma_x^2(s_k, t_l) = \sum_{i=1}^{m} \sum_{j=1}^{p_i} \sum_{i'=1}^{m} \sum_{j'=1}^{p_{i'}} \xi_{ij} \xi_{i'j'} k_x(h_{ii'}, \tau_{jj'}) - 2 \sum_{i=1}^{m} \sum_{j=1}^{p_i} \xi_{ij} k_x(h_{ki}, \tau_{lj}) + k_x(\mathbf{0}, 0). \tag{3.62}$$

3.7.4.1 Description de la procédure automatique

Les principales étapes de la procédure automatique mentionnée auparavant, basée sur les travaux de Christakos et Bogaert (1996), seront détaillées par la suite, en faisant référence à la concentration de polluant comme variable à interpoler.

- **Première étape :** l'ensemble des données contenant les coordonnées spatiales et respectivement temporelles, suivies par les valeurs mesurées de la variable à interpoler est analysé : généralement, les données très proches spatialement sont évitées à cause des instabilités numériques qu'elles peuvent générer.
- **Deuxième étape :** pour chaque nœud du domaine spatio-temporel, un sous-domaine contenant un nombre fixe de données est choisi (à l'intérieur d'un domaine de rayon fixé), données qui révèlent des caractéristiques importantes concernant la variabilité spatio-temporelle de la variable étudiée. Une covariance généralisée spatio-temporelle initiale est fixée (dans cette étude le modèle choisi est $k_x(r, \tau) = r\tau$, modèle valide, qui respecte les conditions d'admissibilité).
  Dans chaque sous-domaine, chaque donnée est retirée et l'estimation est calculée en utilisant le reste des données $(\widehat{X}(s, t)_l)$ pour toutes les combinaisons possibles des deux ordres de continuité ν et μ ($\nu = 0, 1, 2$ et $\mu = 0, 1, 2$).
  Ensuite, on calcule l'erreur pour chaque point d'estimation $(s, t)_l$:

$$e_{l,\nu/\mu} = |\widehat{X}(s, t)_l - X(s, t)_l|_{\nu/\mu} \tag{3.63}$$

  et après on classe les erreurs en leur assignant des rangs. Ensuite, la moyenne de ces rangs est calculée pour tous les points en fonction de toutes les combinaisons possibles (ν, μ). Pour chaque sous-domaine, la combinaison (ν, μ) qui conduit au rang le moins élevé est sélectionnée

comme ordre de continuité pour le sous-domaine en cause.
L'estimation de la covariance généralisée est le pas le plus important qui suit et, en plus, le plus coûteux du point de vue calcul. Il existe un ensemble de candidates comme $CGST$ de la forme 3.60. Les coefficients de ces modèles sont estimés en respectant deux conditions : le meilleur ajustement pour les données disponibles dans chaque sous-domaine et les conditions d'admissibilité (SANLIB, 1995). Pour accomplir cette tâche, on retire chaque point et on fait l'estimation en utilisant le reste des points. Ceci produit un incrément spatio-temporel d'ordre ν/μ :

$$Y_q(s_a, t_b) = \sum_{i=1}^{m} \sum_{j=1}^{p_i} \xi_{ij,ab} X(s_i, t_j), \tag{3.64}$$

où p_i désigne la taille de la séquence temporelle disponible à chaque emplacement s_i, alors que les coefficients $\xi_{ij,ab}$ ont été déterminés antérieurement avec les ordres de continuité ν, μ et le modèle de $CGST$ choisi. On calcule l'espérance mathématique du carré de l'erreur d'estimation :

$$A_{ab} = E[Y_q(s_a, t_b)^2] = \sum_{i=1}^{m} \sum_{j=1}^{p_i} \sum_{i'=1}^{m} \sum_{j'=1}^{p_{i'}} \xi_{ij,ab} \xi_{i'j',ab} k_x(r_{ii'}, \tau_{jj'}) \tag{3.65}$$

et on définit une fonction objectif :

$$F = \sum_{a=1}^{m} \sum_{b=1}^{p_a} [Y_q(s_a, t_b)^2 - A_{ab}^2]^2. \tag{3.66}$$

Les coefficients recherchés sont présents dans F par l'intermédiaire de A_{ab}. On minimise F par rapport à ces coefficients et on résout le problème d'optimisation. Si ces coefficients respectent les conditions d'admissibilité alors on retient cette candidate et on recommence la procédure annoncée (itérative) avec la nouvelle fonction de covariance jusqu'à la convergence de la solution (la dépendance de F des coefficients n'est pas linéaire). Si les coefficients ne respectent pas les conditions d'admissibilité on passe à la candidate suivante.
À la fin, parmi les candidates qui respectent les deux conditions antérieures, on doit trouver la meilleure. Une bonne mesure pour vérifier la qualité de cet ajustement est donnée par la formule :

$$\eta = \frac{\sum_{N_k} Y_q(s_a, t_b)^2}{\sum_{N_k} A_{ab}} \tag{3.67}$$

L'indice η est calculé pour chaque modèle de covariance retenu et le modèle dont l'indice η est le plus proche de 1 est sélectionné.

- **Troisième étape :** les ordres de continuité ν/μ, ainsi que le modèle $CGST$ sélectionnés sont utilisés dans le système d'estimation spatio-temporelle pour calculer les poids.
- **Quatrième étape :** les poids calculés sont utilisés pour cartographier la concentration d'un polluant sur tout le domaine d'étude, ainsi que la variance de cette estimation (l'écart-type).

3.8 Conclusion partielle du chapitre

Dans la première partie du chapitre on a présenté brièvement les principales techniques d'interpolation basées sur le krigeage, applicables dans le cadre *spatio-temporel*, pour obtenir des représentations spatiales des champs de concentrations de polluants.

Dans cette étude, on a appliqué la dernière méthode décrite, le krigeage intrinsèque généralisé spatio-temporel (KIS/T) (voir la section 3.7.4), pour essayer d'améliorer la qualité des cartes montrant quelques épisodes particuliers de pollution atmosphérique (par dioxyde d'azote et par ozone) dans la région d'Île-de-France. Le choix de la méthode est justifié par la facilité de cette approche automatique qui limite l'intervention de l'utilisateur et qui évite l'analyse variographique. Une remarque utile pour ceux qui veulent utiliser cette méthode : l'algorithme est disponible sur le web[1]. Dans la suite, on présentera les résultats obtenus par KIS/T pour deux cas de forte pollution, un pour chaque polluant mentionné.

3.9 Champs de concentrations de polluants obtenus par interpolation spatio-temporelle

On commence la présentation des résultats obtenus utilisant l'interpolation spatio-temporelle, par un court rappel sur les hypothèses de travail. On n'est plus dans le cas purement spatial ; on dispose, cette fois-ci, des *séries temporelles* enregistrées par les mêmes stations de mesure utilisées précédemment pour effectuer le krigeage spatial (voir les figures 2.10 et 2.11). On a la possibilité d'utiliser ces données sans restriction, dans le but d'améliorer la représentation spatiale de deux polluants atmosphériques, notamment le dioxyde d'azote et l'ozone, mais sans intégrer dans cette approche des variables exogènes (météorologiques, topographiques...).

La question principale à laquelle on voulait une réponse, était de savoir si, en utilisant plus de mesures, du même polluant, et enregistrées aux mêmes sites de mesure, mais avant et/ou après le moment d'estimation choisi, on peut reconstituer la corrélation spatio-temporelle nécessaire dans l'estimation spatio-temporelle et améliorer ainsi nos connaissances sur la cartographie des polluants atmosphériques. Il faut souligner encore une fois que les nombres de stations et les zones d'étude respectives restent les mêmes que dans le cas spatial pour les deux polluants (voir la section 2.8.2.2).

Pour essayer de répondre à la question précédente, on a appliqué le **krigeage intrinsèque généralisé spatio-temporel** (KIS/T), décrit dans la section 3.7.4, et implémenté par la procédure automatique développée par Christakos et Bogaert (1996), sur les données de NO_2 et d'ozone. Les deux cas principaux de forte pollution analysés spatialement sont : pour le dioxyde d'azote, le 29 Juillet 1999 à 8 heures, et pour l'ozone, le 30 Juillet 1999 à 14 heures.

Les variations spatio-temporelles des processus présentés dans cette étude exhibent des hétérogénéités qui ne sont pas compatibles avec les hypothèses d'un S/TRF stationnaire et homogène. À la place de celui-ci, on peut utiliser donc un modèle plus général qui sera capable de prendre en compte les hétérogénéités mentionnées. Par un choix approprié de la fonction Q (décrite dans la section 3.7.4), ce modèle pourra être construit sur les bases de la théorie S/TRF présentée ; en particulier, la fonction Q est choisie telle que :

- elle élimine les tendances spatio-temporelles représentées par des fonctions polynomiales de degré ν en espace et μ en temps ;

[1] http ://www.unc.edu/depts/case/SANLIB/

- le S/TRF initial est transformé par Q dans un $S/TRF - \nu/\mu$.

Le cas particulier $\nu = 0$, $\mu = 0$ correspond à un S/TRF avec des **incréments homogènes/stationnaires**. Cette classe de modèles peut gérer des variabilités assez compliqués en termes mathématiques rigoureux (Christakos, 1992). Par exemple, un champ aléatoire de ce type peut exhiber des tendances linéaires en espace et en temps dues au premier moment statistique (la moyenne) de l'incrément. (Par convention, les ordres $\nu = -1$, $\mu = -1$ correspondent à un champ homogène/stationnaire, sans aucune tendance.)

L'estimation effectuée en utilisant l'algorithme KIS/T se fait point par point, en parcourant tous les nœuds du maillage. Un voisinage doit être défini pour chaque point d'estimation. On a choisi de prendre en compte toutes les données disponibles, c'est-à-dire qu'on ne travaille pas avec des voisinages glissants.

L'algorithme doit fournir les deux ordres de continuité, celui spatial et celui temporel, (ν, μ), décrits dans la section 3.7.4, ainsi que les coefficients de la covariance généralisée spatio-temporelle (CGST) cherchée parmi les modèles isotropes de type 3.60. Pour réduire considérablement le temps de calcul, les deux ordres de continuité mentionnés ont été fixés à 0, c'est-à-dire qu'on considère les incréments spatio-temporels homogènes/stationnaires ; il reste à identifier le meilleur ajustement pour nos données d'une fonction de covariance généralisée spatio-temporelle (CGST). On cherche une CGST de type polynomial, ce qui correspond, pour $\nu = \mu = 0$, à la formulation :

$$k_x(r, \tau) = c_1 \delta_r \delta_t - c_2 r \delta_t - c_3 \tau \delta_r + c_4 r \tau. \tag{3.68}$$

où k_x désigne la CGST isotrope, r la distance dans l'espace, τ l'intervalle de temps et δ_r, δ_t prennent la valeur 1 si r, respectivement τ s'annule et 0 autrement. Les trois premiers termes représentent des effets de pépite spatiale et/ou temporelle, tandis que le dernier terme est purement polynomial. Les 4 coefficients sont déterminés par une méthode itérative, présentée dans la section 3.7.4, basée sur un algorithme d'optimisation avec contraintes afin d'assurer la permissibilité de coefficients obtenus. ces conditions sont nécessaires pour que la matrice du système de krigeage ne soit pas singulière. La solution de ce système représente les poids des données utilisées dans l'estimation dans chaque nœud du maillage.

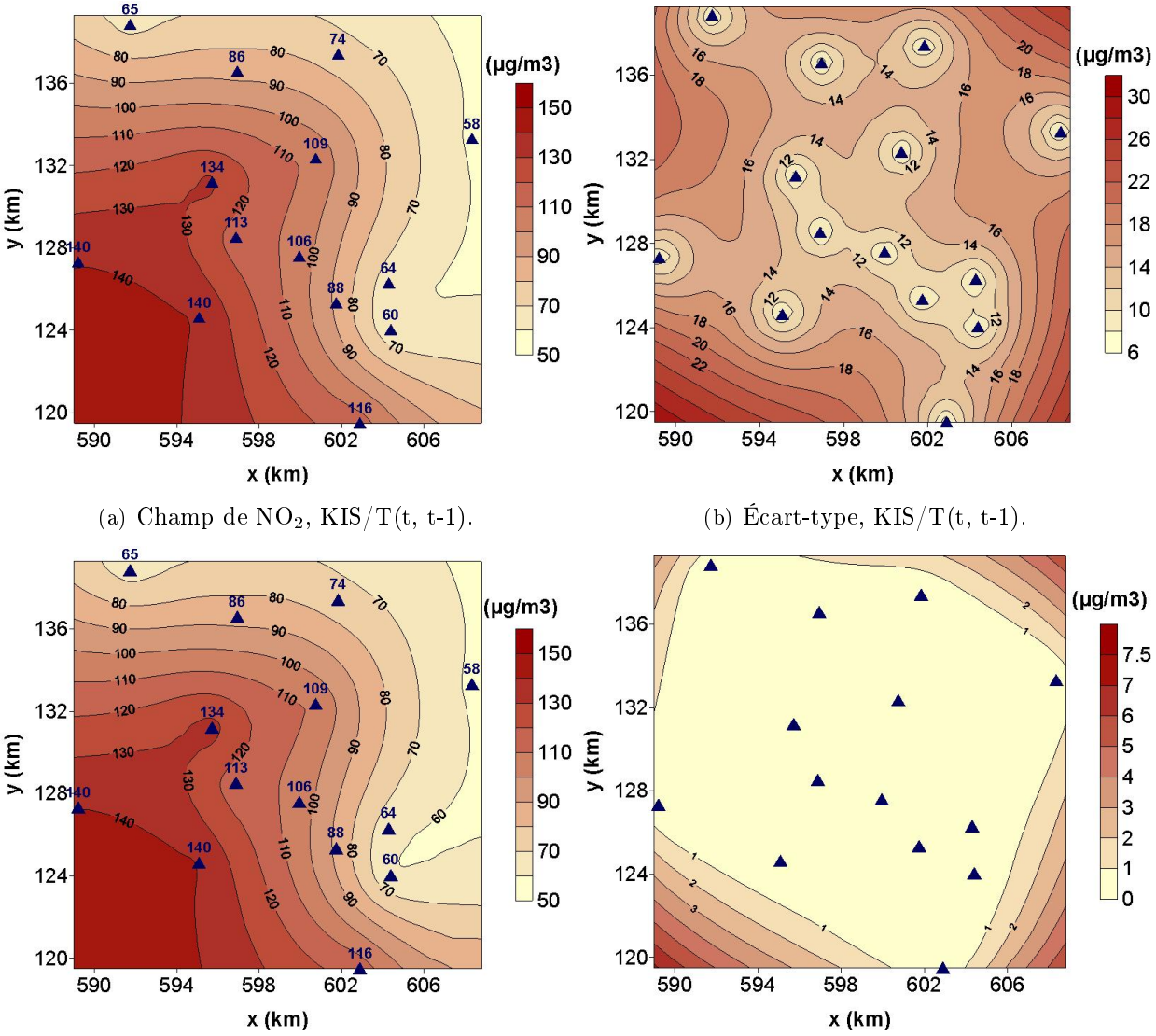
Une fois l'estimation calculée sur toute la grille, on présente les résultats sous forme de cartes d'isoconcentrations de polluants, complétés par la carte de l'écart-type d'erreur associée (calculé d'après la formule 3.7.4) et par des tests de validation croisée (de type Leave-One-Out). Des résultats succincts seront présentés pour chacun de deux jeux de données, et on commence par le dioxyde d'azote.

3.9.1 Champs de NO_2 le 29 Juillet 1999 à 8 heures

Les données enregistrées le 29 Juillet 1999 à 8 heures ont déjà été présentées (voir la section 2.8.2.4). Dans les figures 3.1 et 3.2, on présente les cartes obtenues, ainsi que celles de l'écart-type de l'erreur associées.

Les tests effectués ont montré une sensibilité de l'algorithme d'optimisation, appliqué pour

estimer les coefficients de la CGST, aux unités de mesure des distances, plus précisément au rapport entre la distance spatiale et celle temporelle. La distance maximale entre deux stations de mesure de NO_2 ne dépasse pas 23 km. L'intervalle de temps pour les données considérées dans l'estimation varie entre 1 heure et 5 heures. Pour cette raison, comme on travaille avec des données horaires, tous les calculs ont été effectués sur des voisinage de données où les distances spatiales et/ou temporelles sont mesurées en km/heures.



(a) Champ de NO_2, KIS/T(t, t-1).

(b) Écart-type, KIS/T(t, t-1).

(c) Champ de NO_2, KIS/T(t, t-1, t-2).

(d) Écart-type, KIS/T(t, t-1, t-2).

FIG. 3.1: Estimations des champs de concentrations de NO_2 le **29 Juillet 1999 à 8 heures** en appliquant le Krigeage Intrinsèque Spatio-Temporel KIS/T (t=8) avec les cartes de l'écart-type de l'erreur associées : a),b) à partir des mesures enregistrées à 7 et 8 heures ; c),d) à partir des mesures enregistrées à 6, 7 et 8 heures. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

On présente l'évolution de l'estimation du champ de concentrations de dioxyde d'azote, au fur et à mesure qu'on augmente le nombre de mesures disponibles. On remarque qu'au début, la carte obtenue (figure 3.1(a)) en utilisant les données enregistrées une heure avant le moment de

l'estimation (on utilise donc les mesures enregistrées au moment t et $t-1$, où t=8), resemble très bien au krigeage intrinsèque généralisé spatial (figure 2.18(e)). Ce qu'on peut remarquer pour les coefficients de la CGST est que, pour cette première estimation, on obtient une covariance qui dépend uniquement de l'espace : les deux premiers coefficients sont non-nuls, alors que les 2 derniers qui mettent en avant la partie temporelle, sont nuls.

Si on rajoute progressivement les données enregistrées 2, 3 ou 4 et respectivement 5 heures avant le moment de l'estimation (les cartes 3.1(c), 3.2(a), 3.2(c) et 3.2(e)), on constate qu'il apparaît un phénomène de lissage visible surtout sur les lignes d'isoconcentration marginales. Par exemple, celle qui marque la valeur de $60\,\mu\text{g.m}^{-3}$ présente à l'est du domaine sur les cartes 3.1(a), 3.1(c) et qui est très faible sur 3.2(a) disparaît complètement sur les deux dernières cartes. Le même comportement est visible pour les lignes de $130\,\mu\text{g.m}^{-3}$ et $140\,\mu\text{g.m}^{-3}$ au coin sud-ouest du domaine. Même si les différences entre les cartes obtenues sont visibles, l'allure générale est la même, avec un panache au coin sud-ouest du domaine et des valeurs plus faibles pour le coin opposé. On a pris la décision de s'arrêter avec l'utilisation de données enregistrées avant le moment de l'estimation, à 5 heures. Les tests effectués nous ont paru suffisants pour dire que si on continue à rajouter des données, ceci conduira à un lissage encore plus accentué, principalement parce que les niveaux enregistrés à un moment antérieur celui de l'estimation diminuent (on rappelle qu'il s'agit de 8 heures du matin quand les stations de mesure de la zone parisienne détectent un pic de dioxyde d'azote dû au trafic). Les différences entre les coefficients de la CGST obtenus sont remarquables, dans le sens que, à partir de 2 heures, les deux derniers coefficients de la formule 3.68 sont non-nuls, donc il résulte une dominante temporelle, ainsi qu'une combinée, spatio-temporelle.

Dans le même panel de figures (à droite) on peut suivre l'évolution des cartes de l'écart-type de l'erreur d'estimation. Pour l'estimation faite en utilisant les données enregistrées une heure avant, on rappelle que les coefficients obtenus pour la CGST montrent une dominante spatiale, les niveaux sont légèrement supérieurs à ceux obtenus en utilisant le krigeage intrinsèque spatial (voir la carte 2.18(f)), mais au-delà de 2 heures, quand les coefficients des termes qui contiennent le temps sont non-nuls, les niveaux de l'écart-type diminuent sensiblement.

Pour les cinq estimations présentées, dans le but de vérifier l'efficacité de la méthode appliquée, on a effectué également des tests *Leave-One-Out*, et dans ce cas, pour chaque station, on a retiré toute la série temporelle des données enregistrées par celle-ci, et on a effectué l'estimation avec les données restantes. Dans le tableau 3.1 on présente les résultats détaillés sur les stations. La remarque générale que l'on peut faire en analysant ces résultats est que les meilleurs résultats (plus proches de mesures) sont obtenus pour les deux premières estimations. Mais il faut souligner les faibles différences existantes entre les cinq séries de résultats pour les situations analysées.

Une autre remarque vise les stations Argenteuil et Vitry, situées toutes les deux aux bords, et où l'estimation reste assez loin (d'un ordre de grandeur en moyenne de $40\,\mu\text{g.m}^{-3}$ et respectivement $20\,\mu\text{g.m}^{-3}$) de la valeur mesurée. La même observation peut être faite, tout comme dans le cas spatial, pour la station de Neuilly, qui est très représentative pour le centre du domaine, et où les différences dépassent en moyenne $25\,\mu\text{g.m}^{-3}$. Pour conclure cette analyse individuelle des tests *Leave-One-Out* (LOO) sur les stations, on peut dire que des améliorations par rapport au krigeage

Station	Mesure	**KI**	**KIST1h**	**KIST2h**	**KIST3h**	**KIST4h**	**KIST5h**
ARGENTEUIL	65,00	101,83	**102,92**	107,72	111,74	108,50	117,00
BOBIGNY	58,00	47,24	**66,57**	69,22	70,08	75,38	76,70
GARCHES	140,00	148,47	130,19	131,17	131,50	**131,89**	130,05
GENEVILLIERS	86,00	103,63	90,57	90,40	90,38	90,31	**90,00**
ISSY	140,00	129,16	125,21	**125,49**	125,29	122,77	122,50
IVRY	60,00	78,36	**79,22**	79,35	81,38	85,05	85,53
NEUILLY	134,00	110,10	**110,03**	109,70	109,49	107,66	106,91
PARIS 12	64,00	69,39	73,84	**69,49**	76,00	80,21	85,89
PARIS 13	88,00	87,80	88,70	**87,44**	**88,32**	91,25	89,96
PARIS 18	109,00	93,40	**91,36**	91,17	91,09	89,56	87,32
PARIS 6	106,00	101,37	**101,48**	**101,25**	101,26	100,51	101,08
PARIS 7	113,00	128,30	129,00	128,85	126,42	120,23	**117,14**
ST-DENIS	74,00	70,62	**78,32**	**78,72**	78,81	80,42	79,55
VITRY	116,00	89,82	83,18	85,31	86,00	89,07	**96,27**

TAB. 3.1: Validation croisée pour le NO_2 (**29/07/1999 8h**) pour les cinq cas d'interpolation spatio-temporelle analysés, ainsi que pour celui spatial. Les meilleures estimations mesurées en $\mu g.m^{-3}$ ont été mises en gras.

intrinsèque spatial sont obtenues uniquement pour trois stations : Genevilliers, Paris7, et Vitry, même si l'estimation faite à cette dernière station reste assez éloignée de la mesure effectuée.

Enfin, dans le tableau 3.2 on présente les statistiques effectuées sur les tests (LOO) précé-

Méthode	**MIN** ($\mu g.m^{-3}$)	**MAX** ($\mu g.m^{-3}$)	**MAE** ($\mu g.m^{-3}$)	**RMSE** ($\mu g.m^{-3}$)
KI	0,20	36,83	14,10	17,12
KIST1h	0,7	37,92	14,62	18,07
KIST2h	0,56	42,72	14,66	18,54
KIST3h	0,32	46,74	15,39	19,43
KIST4h	3,25	43,5	16,21	19,63
KIST5h	1,96	52,0	16,76	21,17

TAB. 3.2: Statistiques globales pour les tests "Leave-One-Out" obtenus pour les cinq cas de krigeage spatio-temporel appliqués pour le NO_2, ainsi que pour le krigeage spatial intrinsèque pour comparaison (**29/07/1999 8h**).

dents. Le phénomène de lissage mentionné conduit donc à une faible dégradation progressive de la RMSE (qui croît de $18\,\mu g.m^{-3}$ à $21\,\mu g.m^{-3}$), au fur et à mesure qu'on rajoute des données dans le processus d'estimation. La même observation peut être faite concernant la MAE qui augmente de $14,5\,\mu g.m^{-3}$ à $16,5\,\mu g.m^{-3}$.

Globalement, pour le NO_2, on peut dire que, le fait d'avoir utilisé plus de données sur chaque site de mesure n'améliore pas l'estimation spatiale effectuée auparavant. Quelles sont les explications possibles ? Premièrement, le nombre de données utilisées peut être une cause. La séquence temporelle utilisée est peut être trop courte pour que l'algorithme puisse reconstituer une bonne

corrélation spatio-temporelle. Mais l'utilisation d'une séquence temporelle plus longue conduit à des lissages de l'estimation, donc à une dégradation de la qualité de la représentation spatiale. Deuxièmement, le choix d'une covariance généralisée spatio-temporelle de type isotrope peut être aussi mise en cause. Il est possible que les données incluses dans l'étude exhibent une variabilité plus complexe qui n'est peut pas être exprimée par l'intermédiaire d'un modèle isotrope. En troisième position, et pas le moins important, on peut mentionner la limitation qu'on a imposée en considérant que les incréments spatio-temporels sont homogènes/stationnaires ($\nu = \mu = 0$).

Pour conclure, pour le dioxyde d'azote, sur le domaine choisi, on avait obtenu des résultats assez corrects en utilisant l'interpolation spatiale et on obtient le même type de résultats en utilisant le krigeage intrinsèque spatio-temporel, mais les améliorations "espérées" n'ont pas été obtenues.

3.9.2 Champs de O_3 le 30 Juillet 1999 à 14 heures

Le deuxième polluant atmosphérique analysé est l'ozone. Rappelons que l'impression générale après l'analyse spatiale était que le nombre de mesures réparties dans l'espace n'était pas suffisant pour pouvoir reconstituer la forme du panache.

On a appliqué le même algorithme en utilisant les mêmes limitations que dans le cas de dioxyde d'azote, c'est-à-dire : les ordres de continuité $\nu = \mu = 0$, une covariance généralisée CGST d'équation 3.68 et on a rajouté aux données correspondantes au moment de l'estimation les données mesurées 1 heure, 2 heures, etc. avant ou après le moment d'estimation souhaité, dans le cas de l'ozone : 14 heures, le 30 Juillet 1999.

On commence par rappeler que le domaine d'étude pour ce deuxième polluant est plus étendu que celui pour le dioxyde d'azote. Ceci n'est pas anodin, dans la mesure où, on a déjà fait la remarque concernant la sensibilité du modèle aux unités de mesure pour les distances spatiales/temporelles. Compte tenu du fait que la distance maximale entre deux stations mesurant l'ozone dépasse 120 km, et qu'on travaille, comme dans le cas précédent, avec des mesures horaires on s'attendait à avoir des problèmes avec l'algorithme d'optimisation quand on veut utiliser les données enregistrées une heure et même deux heures avant le moment de l'estimation. Parmi les modèles de CGST fournis par l'algorithme, aucun ne remplissait la condition d'admissibilité et cela conduisait chaque fois à des systèmes de krigeage singuliers, donc pas d'estimation possible. On a essayé de palier à cet inconvénient de deux façons différentes : le premier était de rajouter également dans l'algorithme les données mesurées 1 heure (respectivement 2 heures) après le moment de l'estimation, et le deuxième était de modifier l'unité de mesure pour les distances temporelles (on transforme le temps, exprimé en heures, en minutes). Les deux alternatives proposées ont conduit à l'échec.

En revanche, à partir de 3 heures de données rajoutées avant et/ou après le moment de l'estimation, même en gardant l'unité de mesure pour l'espace en km et celle pour les intervalles de temps en heures, on arrive à des covariances généralisées qui respectent les condition d'admissibilité et donc à des estimations correctes du champ de concentrations d'ozone.

Dans la figure 3.3 on présente les cartes obtenues utilisant 3, 4 et respectivement 5 heures **avant et après** le moment de l'estimation qui est le 30 Juillet 1999 à 14 heures. Donc pour la première carte on a utilisé les données mesurées entre 11 heures et 17 heures le même jour, pour la deuxième entre 10 heures et 18 heures et pour la dernière entre 9 heures du matin et 19 heures du soir. Ces cartes sont à comparer avec celle obtenue par krigeage intrinsèque spatial (figure 2.28(e)).

L'évolution observée sur ces cartes n'est plus du tout la même que celle remarquée dans le cas du dioxyde d'azote. Si en passant de 3 heures à 4 heures, on remarque un effet de lissage, l'utilisation des données de 9 heures à 19 heures produit une carte moins lisse que la précédente. En revanche, les cartes de l'écart-type de l'erreur associées, présentées pour chaque estimation effectuée, montrent des niveaux similaires (très faibles). Néanmoins, les cartes exhibent le même panache au coin nord-ouest du domaine, sauf que ce panache a une forme différente, fonction de données utilisées. Concernant les coefficients de la CGST estimés, on peut faire la même remarque que pour le dioxyde d'azote, notamment que les deux derniers sont non-nuls, ce qui met en avant la partie temporelle rajoutée dans le processus d'estimation.

Dans le tableau 3.3 on présente les tests "Leave-One-Out" effectués pour les trois situations présentées. On peut constater que, ponctuellement, le fait d'utiliser les données antérieures et postérieures au moment de l'estimation améliore les résultats, même si, en moyenne, cette amélioration n'est pas significative. Au moins pour trois stations : Aubervilliers, Fontainebleau et Chartres, **la correction par rapport au krigeage spatial** est marquante : d'environ $14\,\mu\text{g.m}^{-3}$ pour la première, $21\,\mu\text{g.m}^{-3}$ pour la deuxième et $36\,\mu\text{g.m}^{-3}$ pour la dernière.

Les statistiques moyennes sur les tests LOO montrent que, à l'exception de l'estimation faite en utilisant les données enregistrées entre 11 heures et 17 heures (3 heures avant et après) qui révèle une certaine amélioration par rapport à l'estimation spatiale, les deux autres sont très similaires à ce qu'on avait obtenu auparavant. Globalement, le gain obtenu en utilisant des mesures enregistrées à différents moments de temps, autres que celui de l'estimation est faible, même si, ponctuellement, ce gain peut être non-négligeable.

Quelles sont les raisons d'un tel comportement ? Pouvait-on espérer plus de cette extension spatio-temporelle ? Comme déjà évoqué dans le chapitre consacré à l'interpolation spatiale, la technique de krigeage intrinsèque s'est montrée relativement impuissante pour reconstituer le champ d'ozone à partir des mesures disponibles (peu nombreuses et d'une répartition spatiale creuse). On rappelle que la répartition spatiale n'a pas été modifiée. Nous avons juste enrichi la base de données utilisée dans l'estimation (par des courtes séries temporelles précédant et suivant le moment de l'estimation) en essayant de combler ainsi le manque spatial. La covariance généralisée spatio-temporelle s'avère, dans certains cas, elle-aussi impuissante pour prendre en compte la variabilité spatio-temporelle exhibée par les données. La limitation de modélisation qu'on a imposée (les deux ordres de continuité nuls) pour alléger le volume de calcul et ainsi réduire le temps nécessaire pour obtenir l'estimation, peut s'avérer très restrictive. D'un autre côté, il faut bien mettre en évidence le fait que les essais de modélisation présentés étaient basés uniquement sur des mesures enregistrées

Station	Mesure	**KI**	**KIST3h**	**KIST4h**	**KIST5h**
AUBERVILLIERS	204,00	178,01	178,00	**191,99**	178,00
GARCHES	226,00	198,79	**198,56**	195,04	198,31
GENEVILLIERS	207,00	193,45	193,48	194,68	**195,85**
MANTES	193,00	209,58	**210,30**	210,75	229,26
MELUN	162,00	175,77	175,40	175,07	**174,93**
MONTGERON	182,00	174,57	178,14	188,80	**184,84**
NEUILLY	197,00	196,87	196,88	**197,26**	202,68
PARIS 13	207,00	194,49	194,47	190,95	**194,45**
PARIS 18	173,00	202,16	202,16	**196,99**	209,78
PARIS 6	202,00	195,26	195,26	195,26	195,26
PARIS 7	192,00	200,29	200,29	**197,52**	200,29
RAMBOUILLET	186,00	198,23	197,61	**194,47**	196,86
TREMBLAY	169,00	184,58	**179,08**	186,42	184,21
VITRY	180,00	199,53	199,48	**194,02**	199,42
MONTGE	163,00	169,28	**170,77**	208,46	183,02
FREMAINVILLE	216,00	199,52	200,05	**201,87**	201,03
PRUNAY	201,00	196,15	196,1	**201,48**	196,04
FONTAINEBLEAU	178,00	152,89	165,18	198,22	**181,88**
SAINTS	164,00	159,22	161,90	180,65	**164,81**
CHARTRES	189,00	150,76	**191,09**	195,94	211,55

TAB. 3.3: Validation croisée pour l'ozone (**30/07/1999 14h**) pour les trois situations analysées. Les meilleures estimations, mesurées en $\mu g.m^{-3}$, ont été mises en gras.

Méthode	**MIN** ($\mu g.m^{-3}$)	**MAX** ($\mu g.m^{-3}$)	**MAE** ($\mu g.m^{-3}$)	**RMSE** ($\mu g.m^{-3}$)
KI	0,13	38,24	15,22	17,98
KIST3h	0,12	29,16	12,26	14,76
KIST4h	0,26	45,46	14,46	17,70
KIST5h	0,81	36,78	14,98	18,19

TAB. 3.4: Statistiques globales pour les tests "Leave-One-Out" obtenus pour les trois estimations par krigeage intrinsèque spatio-temporel appliqués sur les données d'ozone enregistrées le **30 Juillet 1999 à 14 heures**.

de l'ozone. Aucune variable exogène n'a été incluse jusque là dans l'étude. Or, la production de l'ozone est très influencée par ses précurseurs, et la météorologie joue un rôle prépondérant dans sa dispersion. Ce sont des facteurs qui ont une contribution décisive si on veut modéliser le comportement de ce polluant.

La conclusion qui s'impose est que seules les mesures d'ozone ne peuvent pas contribuer d'une manière significative à reconstituer, d'une façon satisfaisante, un épisode de pollution atmosphérique.

3.10 Conclusion du chapitre

Dans ce troisième chapitre on a décrit brièvement les lignes directrices de la géostatistique spatio-temporelle traditionnelle (Kyriakidis et Journel, 1999) et on a appliqué la méthode de krigeage intrisèque spatio-temporel (KIS/T) sur des séquences très courtes de données horaires (des concentrations de polluants atmosphériques) de dioxyde d'azote et d'ozone, mesurées par les stations qui ont été utilisées dans le cadre spatial, en ayant comme objectif déclaré l'amélioration, si possible, de la représentation spatiale des champs de polluants atmosphériques. En théorie, le fait d'incorporer des mesures enregistrées à différents moments de temps, autres que celui de l'estimation, peut apporter une certaine quantité d'information, qui pourrait améliorer nos connaissances sur la situation étudiée.

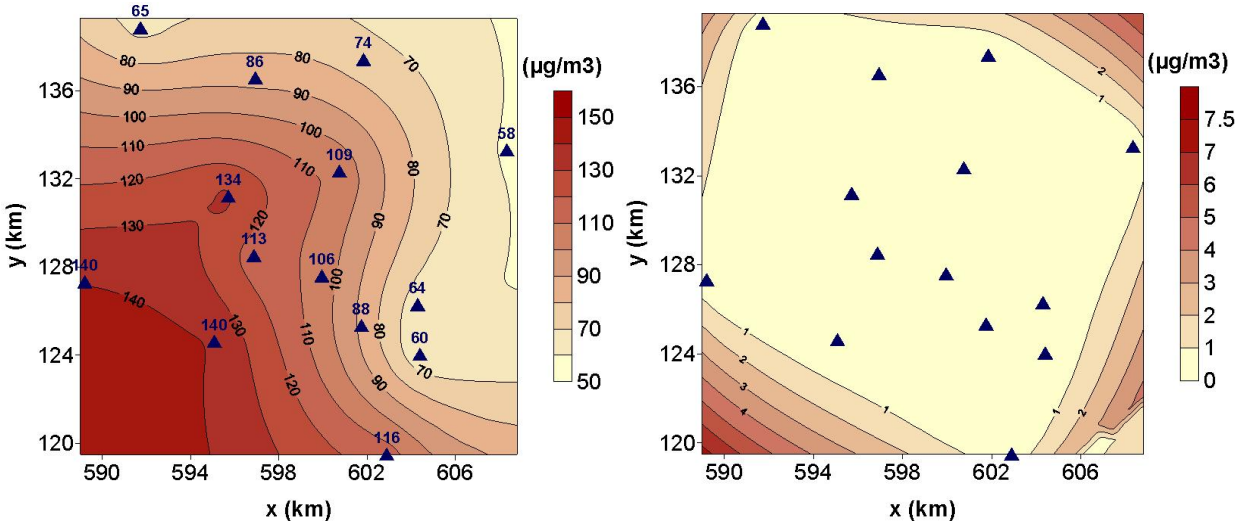
Compte tenu du nombre et de la qualité des données dont nous disposions, les résultats obtenus ne sont pas à la hauteur de ce que les premisses théoriques laissaient présager. Comme observation générale concernant la méthode appliquée, on peut constater que la structure de corrélation spatio-temporelle des données incluses dans l'étude est peut-être trop complexe pour qu'une covariance généralisée spatio-temporelle de type polynômial puisse la reproduire correctement. Comme limitation dans la procédure appliquée, il faut rappeler que les ordres de continuité n'ont pas été estimés, mais fixés à zéro, ce qui correspond à l'hypothèse que les incréments spatio-temporels sont homogènes/stationnaires. De plus, comme on avait déjà remarqué dans le cas spatial, la répartition spatiale des stations mesurant l'ozone est creuse et cela influence d'une manière négative l'estimation. Les résultats faibles peuvent être causés également par l'échelle temporelle à laquelle on a travaillé. Il est possible que pour des données plus lisses, genre moyennes journalières, la méthode s'avère plus efficace, mais pour les épisodes étudiés, cela n'a pas été le cas.

Pour les deux polluant étudiés, le dioxyde d'azote et l'ozone, on a présenté pour chacun un épisode de forte pollution. Ce qu'on peut remarquer pour les deux cas analysés est que, généralement, cela n'améliore pas d'une manière significative l'estimation spatiale déjà présentée. Néanmoins, il existe des stations où les tests de validation croisée ont mis en avant des meilleurs résultats avec l'estimation spatio-temporelle. Étant donnée la complexité des calculs, l'utilisation de cette méthode n'est pas justifiée en termes de gain de précision par rapport à celle purement spatiale. Elle a pourtant le mérite d'appuyer les résultats obtenus avec la méthodologie spatiale, car les champs estimés dans les deux cas sont assez similaires, et avec une variance d'estimation moindre dans le cas spatio-temporel. Indirectement, on montre la robustesse des résultats obtenus dans la méthodologie spatiale. Ces remarques sont valables et utiles essentiellement pour le cas du NO_2, où l'interpolation spatiale s'était avérée plus pertinente que pour l'ozone, ceci étant dû en majeure partie à la couverture spatiale de la région par des stations de mesure.

Il faut souligner également que, dans cette application, on n'a pas utilisé d'autres variables, comme les données météorologiques ou les émissions, qui ont une grande influence sur la production et le transport de polluants, notamment sur l'ozone.

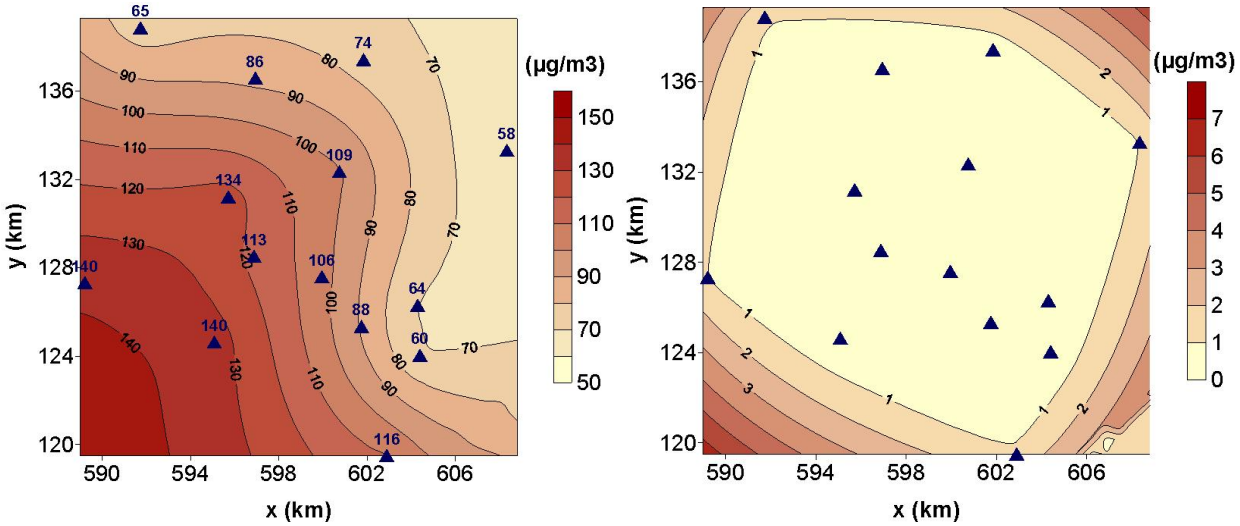
Pour conclure, la géostatistique spatio-temporelle monovariable n'est pas en mesure de décrire adéquatement la continuité spatio-temporelle des processus étudiés. Pour cela, des modèles

de transport physico-chimiques sont en général mieux adaptés. C'est pour cette raison que notre attention sera orientée dans le chapitre suivant vers les modèles de chimie-transport.



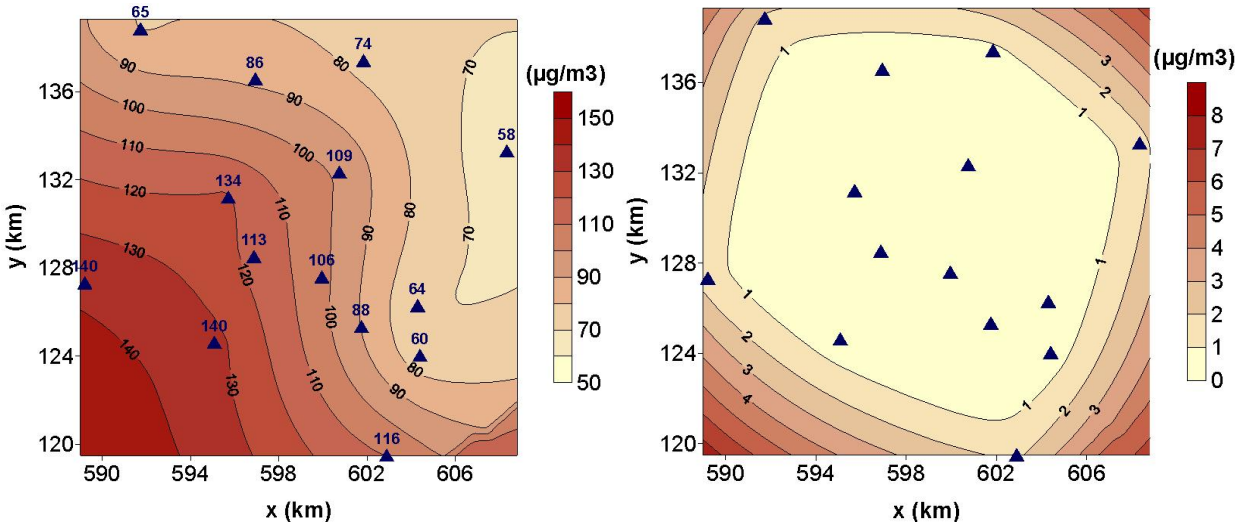
(a) Champ de NO_2, KIS/T(t,t-1,t-2,t-3).

(b) Écart-type, KIS/T(t,t-1,t-2,t-3).



(c) Champ de NO_2, KIS/T(t,t-1,t-2,t-3,t-4).

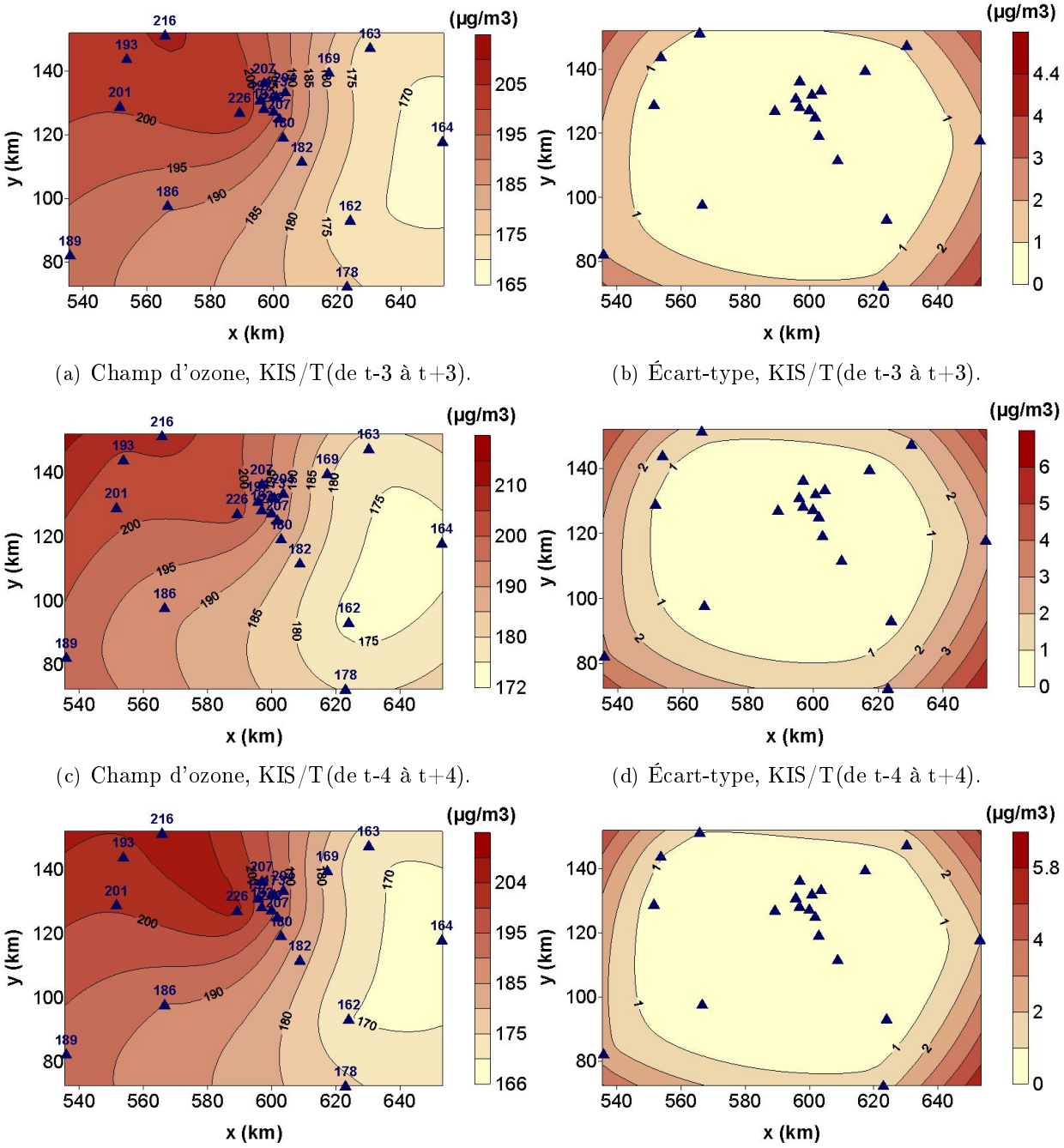
(d) Écart-type, KIS/T(t,t-1,t-2,t-3,t-4).



(e) Champ de NO_2, KIS/T(t,t-1,t-2,t-3,t-4,t-5).

(f) Écart-type, KIS/T(t,t-1,t-2,t-3,t-4,t-5).

FIG. 3.2: Estimations des champs de concentrations de NO_2 le **29 Juillet 1999 à 8 heures** en appliquant le Krigeage Intrinsèque Spatio-Temporel KIS/T (t=8) avec les cartes de l'écart-type de l'erreur associées : a),b) à partir des mesures enregistrées de 5 heures à 8 heures ; c), d) à partir des mesures enregistrées de 4 heures à 8 heures ; e), f) à partir des mesures enregistrées de 3 heures à 8 heures. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.



(a) Champ d'ozone, KIS/T(de t-3 à t+3).

(b) Écart-type, KIS/T(de t-3 à t+3).

(c) Champ d'ozone, KIS/T(de t-4 à t+4).

(d) Écart-type, KIS/T(de t-4 à t+4).

(e) Champ d'ozone, KIS/T(de t-5 à t+5).

(f) Écart-type, KIS/T(de t-5 à t+5).

FIG. 3.3: Estimations des champs de concentrations de O_3 le **30 Juillet 1999 à 14 heures** en appliquant le Krigeage Intrinsèque Spatio-Temporel KIS/T (t=14) avec les cartes de l'écart-type de l'erreur associées : a), b) à partir des mesures enregistrées de 11 heures à 17 heures ; c), d) à partir des mesures enregistrées de 10 heures à 18 heures ; e), f) à partir des mesures enregistrées de 9 heures à 19 heures. Les triangles correspondent à l'emplacement des stations de mesure et la valeur au-dessus du symbole, à la mesure.

Chapitre 4

Assimilation de données sur des modèles de chimie-transport

La modélisation de la pollution atmosphérique s'appuie sur des modèles de chimie-transport (CTM), qui sont devenus indispensables pour l'évaluation, la prévision et le contrôle de la qualité de l'air. Les CTM répondent aussi à un besoin d'estimation de l'état de la pollution hors des zones couvertes par la mesure. Cependant, vue la complexité des phénomènes, des lois d'évolution du système atmosphérique, ainsi que des incertitudes sur les différentes entrées du modèle, les CTM sont uniquement des représentations approximatives de la réalité, donc entachés d'erreurs. Afin d'obtenir des meilleures représentations spatiales des champs de concentration de polluants, on peut introduire des observations de surface dans le modèle, pour le contraindre à suivre les mesures. C'est précisément ce que proposent les méthodes d'assimilation de données.

Cette partie de l'étude s'articule autour de la technique d'assimilation de données. On commence d'abord par une courte présentation des modèles de chimie-transport, ainsi que de la méthode d'assimilation de données, avec l'accent mis sur les méthodes séquentielles, notamment sur le Filtre de Kalman d'Ensemble. Cette méthode a été choisie pour être implémentée sur un modèle de chimie-transport simulant la qualité de l'air sur la région d'Île-de-France, CHIMERE, dans le but d'améliorer la cartographie de deux polluants atmosphériques, l'ozone et le dioxyde d'azote, sur cette région. Les résultats obtenus en appliquant la méthode mentionnée seront présentés dans la deuxième partie de ce chapitre.

4.1 Modèles de Chimie-Transport. Généralités.

Les CTM sont des implémentations numériques d'un ensemble d'équations décrivant une réalité physique ; ils permettent aujourd'hui de simuler, de façon assez réaliste, les phénomènes très complexes qui se déroulent dans les basses couches de l'atmosphère. L'objectif principal d'un tel modèle est de calculer les concentrations de polluants sur un domaine tridimensionnel discrétisé. Ce domaine peut couvrir une zone de quelques dizaines de kilomètres (échelle régionale), jusqu'à plusieurs milliers de kilomètres (échelle continentale) et même l'échelle planétaire.

Il existe deux systèmes de référence par rapport auxquels il est possible d'étudier la dynamique des polluants atmosphériques : le système Lagrangien, qui considère l'objet étudié et le suit dans son déplacement (donc le référentiel se déplace avec l'écoulement d'air considéré) et le système Eulerien qui est fixé naturellement à la surface terrestre (Hanna et al., 1982). Par conséquent, il existe plusieurs types de modèles déterministes ; parmi eux on distingue : les modèles Lagrangiens, Gaussiens, Euleriens et ceux qui s'occupent de la modélisation chimique des composants réactifs.

À l'échelle régionale, celle qui nous intéresse dans cette étude, les modèles de chimie-transport le plus souvent utilisés sont de type eulérien : un maillage est superposé sur le domaine et la quantité de polluant qui transite dans chaque boîte ainsi délimitée est calculée.

L'évolution des concentrations de polluants est régie principalement par les **conditions météorologiques** et les **émissions** de polluants, en reliant leurs variations temporelles aux **phénomènes de transport**, ainsi qu'aux processus physico-chimiques de **production** et de **perte**, tout en tenant compte des concentrations des composés chimiques aux limites du domaine (**conditions aux limites**) et des concentrations des composés chimiques au début de la période de temps considérée (**concentrations initiales**).

La modélisation de cet ensemble de phénomènes est effectuée par un système d'équations aux dérivées partielles, résolu à l'aide des divers schémas numériques. Cette résolution numérique implique la **discrétisation spatiale** du domaine tridimensionnel considéré, discrétisation qui dépend de plusieurs facteurs : la taille du domaine, la durée de vie des polluants, ainsi que du cadastre d'émissions. Il faut mentionner également que la résolution numérique implique une **discrétisation temporelle** de la période à laquelle on s'intéresse. On parle d'une résolution temporelle qui dépend du maillage spatial, de la durée de la période considérée, et bien sûr, des propriétés des polluants.

Les processus physico-chimiques pris en compte généralement dans un modèle de chimie-transport sont très complexes. Parmi eux, on peut mentionner :

- la **dispersion horizontale** liée à la composante horizontale du vent (cela peut être très important, pour une espèce comme l'ozone qui peut être transportée sur plusieurs dizaines et même centaines de kilomètres) ;
- la **dispersion verticale** liée à la composante verticale du vent (facteur dynamique) et à la convection turbulente qui se produit sous l'effet du chauffage du sol par le soleil (facteur thermique) ;
- le **dépôt sec** qui dépend du type de sol, des espèces chimiques considérées et des conditions météorologiques (il s'agit d'un processus de perte qui peut être très important à grande échelle pour une espèce comme l'ozone) ;
- les **réactions chimiques**, surtout celles de photolyse qui transforment les précurseurs comme les oxydes d'azotes et les Composés Organiques Volatiles (COV) en ozone. Leur nombre est très élevé ; par conséquent, des hypothèses simplificatrices sont faites, visant à réduire le nombre d'espèces et de réactions chimiques ;
- la **microphysique et chimie des aérosols** : nucléation, condensation, aggrégation.

Vue la complexité des phénomènes traités par les modèles de chimie-transport, on s'attend à avoir besoin d'un certain nombre de données d'entrée, qui sont obtenues auprès de plusieurs organismes impliqués dans le domaine de la qualité de l'air. Premièrement, les CTM ont besoin des **émissions** des précurseurs issues de sources ponctuelles, linéaires, surfaciques, d'origine anthropique ou naturelle. Deuxièmement, ils demandent des **données météorologiques**, qui ne sont pas produites par le modèle lui-même, mais elles proviennent des modèles météorologiques et de diverses bases de données pour le domaine d'intérêt. Troisièmement, les **concentrations de polluants aux limites** sont très importantes, car le domaine d'intérêt n'est pas complètement isolé, donc des espèces chimiques sont transportées de l'extérieur vers l'intérieur du domaine. Également, les **concentrations initiales** sont indispensables pour initialiser le modèle et, pour finir, il y a le **mode d'occupation du sol** et en particulier la **végétation** qui jouent un rôle très important surtout pour le processus de dépôt.

Au-delà des aspects méthodologiques, visant la mise en œuvre d'un tel modèle de chimie-transport, il existe des avantages, mais aussi des limitations qu'on essaiera de découvrir par la suite. Parmi les avantages on peut énumérer le fait que, contrairement aux modèles statistiques présentés antérieurement ou de type "boîte noire", les CTM intègrent l'ensemble des connaissances sur les phénomènes étudiés, pour fournir une information continue et *potentiellement* très complète. Ils permettent ainsi de quantifier, même d'une façon simplifiée, les interactions entre les différents processus.

En revanche, tous ces modèles ne peuvent délivrer que l'état des connaissances disponibles à un certain moment. De plus, ils ont besoin d'une quantité très importante d'informations de plusieurs types : météorologique, géophysique ou chimique. L'évaluation faite par des experts sur les entrées révèle des fortes incertitudes qui touchent l'ensemble de données injectées dans le système : les émissions, les champs météorologiques, les réactions chimiques, les processus de dépôt, les conditions initiales ou bien les conditions aux limites. Dans ces conditions, les concentrations simulées par ces modèles sont, bien sûr, incertaines. De plus, les décisions de modélisation prises ne sont que des simplifications de la réalité (des paramétrisations), donc elles ne sont pas en mesure de décrire avec une extrême précision les phénomènes atmosphériques très complexes. Une autre limitation importante est liée aux coûts numériques des schémas, utilisés dans les diverses étapes de la modélisation, notamment l'intégration dans le temps d'un nombre très élevé des variables. Vu le nombre assez élevé de sources d'incertitude et des limitations dans les modèles de chimie-transport, on s'attend à obtenir des résultats (concentrations simulées) entachés d'erreurs.

L'une des stratégies le plus utilisée dernièrement pour améliorer la qualité des simulations numériques des modèles de chimie-transport est l'utilisation d'une technique appelée assimilation de données, qui sera décrite par la suite.

4.2 Assimilation de données : méthodes et coût numérique

L'assimilation de données (AD par la suite) vise à prendre la meilleure partie de deux informations dont on dispose : d'une part, un modèle numérique décrivant par l'intermédiaire des équations différentielles, la physique et la chimie du système, et d'autre part, les observations *in situ*, utilisées

via une procédure d'assimilation de données, afin de corriger les prédictions du modèle, d'estimer l'état du système et d'évaluer la précision de cette estimation. Il s'agit donc, d'exploiter **conjointement** la modélisation et l'information disponible (les observations) pour identifier au mieux les caractéristiques physiques et l'état du système concerné, ce qui permet éventuellement aussi de surveiller, voire de prévoir, ses évolutions.

La géostatistique se retrouve impuissante devant la nécessité de fournir des lois spatiales pour des variables soumises à des contraintes physiques et elle se cantonne, en général, aux systèmes statiques (Bertino, 2001). L'AD apporte quelque chose de plus, en introduisant le modèle *physique* dans un cadre stochastique. Les connaissances des processus dynamiques complexes et non-linéaires sont combinées aux observations directes. En contraignant l'évolution d'un système stochastique par un système d'équations dynamiques, l'AD fournit *a posteriori* les moments des variables aléatoires étudiées et, en particulier, les covariances spatiales non stationnaires qui vérifient les contraintes physiques du système.

L'aspect le plus important de la technique d'assimilation de données est le fait qu'il s'agit d'un problème inverse. Cette procédure comporte deux étapes : d'abord, on doit indiquer les incertitudes du modèle et par la suite on utilise les données mesurées (observations) pour estimer les incertitudes le plus précisément possible. Par conséquent, les problèmes d'assimilation de données sont des problèmes inverses : les incertitudes indiquées du modèle, qui sont les entrées du système, doivent être reconstruites en utilisant les mesures, qui sont les sorties du système.

Des problème similaires sont rencontrés dans plusieurs domaines : la navigation, le traitement du signal, la physique des solides, la physique du plasma. Malgré la diversité des problèmes et des systèmes physiques étudiés, les méthodes utilisées pour résoudre ces problèmes sont très similaires, même s'ils ont été souvent développés indépendamment. Ils peuvent être formulés de façon probabiliste et leur but reste le même : déterminer une approximation raisonnable de la fonction de distribution de probabilité conditionnelle de l'état du système, compte tenu des informations disponibles. Quant aux algorithmes numériques utilisés, ils sont souvent très similaires et fondamentalement indépendants des équations qui régissent la système physique étudié. Même si la procédure d'estimation est indépendante du système physique considéré, des différences significatives existent.

On distingue classiquement deux familles de méthodes utilisées pour l'implémentation numérique des algorithmes d'AD : d'une part, les méthodes variationnelles (Le Dimet et Talagrand, 1986), fondées sur une procédure de minimisation (on parle de 3D-Var quand l'évolution en temps n'est pas prise en compte, de 4D-Var quand elle l'est), et d'autre part, les méthodes séquentielles, liées à la théorie de l'estimation et au Filtre de Kalman (Kalman, 1960).

4.2.1 AD séquentielle

Cette catégorie de méthodes est dérivée de la théorie de l'estimation statistique : l'Interpolation Optimale (Daley, 1991), qui a été la méthode la plus utilisée dans les centres de prévision météorologiques, et le Filtre de Kalman (KF) (Kalman, 1960).

La grandeur fondamentale qui nous intéresse est le vecteur d'état du système dynamique utilisé. Nous devons effectuer l'estimation de ce vecteur. Cette estimée est de nature conditionnelle : il s'agit de la moyenne du vecteur d'état, conditionné sur les observations successives. Il faudra également un moyen pour juger la qualité de cette estimation ; ce moyen sera constitué par la covariance de l'erreur correspondante à l'estimation effectuée.

D'un point de vue algorithmique, dans l'assimilation séquentielle, le modèle est intégré sur un intervalle de temps sur lequel les observations disponibles sont distribuées. Quand le modèle arrive à un moment où on dispose d'une observation, l'état prédit par le modèle est corrigé avec cette nouvelle observation et l'intégration est ré-initialisée avec l'état analysé précédent. La procédure est répétée pour prendre ainsi en compte toutes les observations. Par conséquent, l'assimilation séquentielle est une alternance des analyses calculées au moment de l'observation et des intégrations du modèle. Ce caractère séquentiel était recherché surtout dans la prévision météorologique, mais cela peut constituer un inconvénient si on veut effectuer une réassimilation *a posteriori* des observations, car la propagation de l'information est faite dans un sens unique : du passé vers le futur, et jamais dans le sens inverse.

Conçu pour les systèmes linéaires, le Filtre de Kalman ne pouvait pas être appliqué directement sur les systèmes non-linéaires. Une solution consiste en un développement en série de Taylor des opérateurs non-linéaires qui interviennent dans l'algorithme autour des moyennes conditionnelles, développement sur lequel on retient les termes jusqu'à un certain ordre préétabli. Cette version du filtre a été appelée Filtre de Kalman Étendu. Une description très détaillée de ce développement se trouve dans Jaswinski (1970). Par contre, son coût numérique pour des grands systèmes atmosphériques est très élevé et c'est pour cela qu'un grand ensemble de schémas sous-optimaux ont été créés afin de pouvoir appliquer cette méthode pour la prédiction opérationnelle. Parmi ceux-ci, les plus connus sont : le RRSQRT[1] de Verlaan et Heemink (1997), les filtres SEEK[2] et SEIK[3] de Pham et al. (1997). Un cas spécial de schéma sous-optimal est le Filtre de Kalman d'Ensemble (EnKF, par la suite) développé par Evensen (1994).

4.2.2 AD variationnelle

Cette deuxième catégorie de méthodes part du principe qu'il faut ajuster une solution pour toutes les observations disponibles, sur toute la période d'assimilation. Les états estimés sont ainsi corrigés en utilisant toute l'information exploitable. D'un point de vue algorithmique, on cherche la solution qui minimise une fonction de coût prédéfinie : typiquement, cette fonction comporte une somme pondérée de carrés de différences entre les observations et les sorties du modèle, les poids reflétant nos connaissances sur la précision des observations et du modèle (Talagrand, 1997). Le principe de cette méthode est basé sur des évaluations successives du gradient de la fonction objectif (par l'intermédiaire des équations adjointes), en utilisant un algorithme de descente pour approcher

[1] Reduced Rank Square Root
[2] Singular Extended Evolutive Kalman Filter
[3] Singular Extended Interpolated Kalman Filter

un minimum. On minimise ainsi l'écart entre le modèle et les observations sur une fenêtre d'assimilation, et cela est particulièrement utile quand on veut ré-analyser les épisodes passés (Bouttier et Courtier, 1999). Une brève description du problème variationnel sera donnée dans la section 4.4.4.

4.3 Types d'applications de l'assimilation de données

Il existe plusieurs types d'applications de l'assimilation de données en fonction de l'objectif proposé :

- si l'objectif, en utilisant les CTMs, est de faire de la **prévision** (notamment pour la pollution atmosphérique), le recours aux méthodes d'assimilation de données est incontournable (même si les derniers résultats présentés par les équipes impliquées dans cette direction de recherche ne sont pas très encourageants sur l'amélioration de la prévision) ;
- si l'objectif visé n'est pas la prévision d'un état, mais plutôt une **estimation plus précise** de cet état en utilisant le modèle, ainsi que les mesures enregistrées jusqu'au moment présent, on appelle cela "now-casting" ou **filtrage** ; l'accent est mis sur une correction très fréquente du modèle en utilisant les mesures dès qu'elles deviennent disponibles ;
- un troisième type d'application est représenté par ce qu'on appelle un **lisseur** (en anglais "smoothing" ou "hind-casting") ; le comportement dynamique du modèle est reconstruit à l'aide de toutes les mesures disponibles (même postérieures au moment souhaité d'estimation), ainsi que du modèle ;
- dans un registre différent, on rencontre l'**estimation de paramètres** (ou **calibration**) ; les paramètres à estimer peuvent être liés à la dynamique du système ou au processus d'observation. À chaque pas de temps, le filtre améliore le système en utilisant les connaissances acquises sur les paramètres. Autrement dit, le système apprend sur lui-même ;
- il existe des méthodes d'assimilation de données qui peuvent aider à la **conception des réseaux** de mesure, en permettant le calcul des effets obtenus lors du fait de rajouter ou d'éliminer certaines stations de mesure.

4.4 Les concepts de base dans l'assimilation de données

L'analyse est la production d'une image précise de l'état vrai du système, à un moment donné (Bouttier et Courtier, 1999). L'information objective qui peut être utilisée pour produire l'analyse est représentée par les observations fournies par les stations de mesure. Si l'état du modèle est sur-déterminé par les observations, alors l'analyse est réduite à une interpolation. Dans la plupart des cas, le problème d'analyse est sous-déterminé parce que les observations ne sont pas assez nombreuses et car elles sont reliées indirectement aux variables du modèle. Pour obtenir un problème bien posé, il est nécessaire d'avoir une information sous forme d'une *ébauche*-première estimation de l'état du système. Cette *ébauche* peut être générée comme une climatologie ou comme une sortie d'une analyse précédente en utilisant quelques hypothèses de consistance dans le temps, de l'état du modèle, comme la stationnarité, ou l'évolution prédite par le modèle. Cette information est accumulée dans l'état du système et elle est propagée à toutes les variables du modèle. Donc,

l'**assimilation de données** peut être définie comme une technique d'analyse dans laquelle l'information est accumulée dans l'état du modèle, en profitant des contraintes de consistance avec les lois d'évolution temporelle et les propriétés physiques.

4.4.1 Vecteur d'état, l'espace de contrôle et les observations

Pour introduire les concepts de base utilisés dans l'assimilation de données, on fait appel à la demarche proposée par Bouttier et Courtier (1999).

• Vecteur d'état

Le système de notations adopté est celui utilisé dans le domaine de l'assimilation de données. Un modèle est toujours limité par sa résolution, donc il ne peut pas simuler l'état *vrai* du système. On peut juste approcher cet état par l'état réel moyen par maille du modèle, représenté par le vecteur x^t (où l'exposant t signifie *true* en anglais). Les composantes du vecteur x^t discrétisé dans l'espace sont les valeurs des différentes variables d'état en chaque nœud de la grille du modèle.

Un deuxième vecteur, très important, est le vecteur d'ébauche ou en anglais *forecast*, noté x^f ; il s'agit d'une première estimation de l'état vrai, avant que toute analyse soit effectuée. Cette notation est utilisée quand on traite des systèmes dynamiques. Pour ceux statiques on préfère la notation x^b avec b qui signifie *background*. Celle-ci est la notation utilisée d'ailleurs dans l'interpolation statistique.

Le dernier vecteur est celui d'analyse x^a, celui qui nous intéresse le plus et qu'on veut estimer.

• L'espace de contrôle

Parfois, dans le cas d'un système complexe, il est préférable de ne pas résoudre le problème d'analyse pour toutes les variables présentes dans le système, soit parce que on ne connaît pas leur interaction avec celles qui nous intéressent, soit parce que on est limité en pouvoir computationnel. Dans cette situation, l'espace du modèle ne coïncide plus avec l'espace *de travail* ; ce dernier est appelé espace des variables de contrôle. Autrement dit, on doit chercher une correction δx telle que :

$$x^a = x^b + \delta x \tag{4.1}$$

soit le plus proche possible de x^t. D'un point de vue mathématique, cela revient à contraindre x^a à appartenir à l'espace engendré par x^b plus le sous-espace du vecteur des variables de contrôle.

• Les observations

Pour effectuer l'analyse, on a besoin d'un certain nombre de mesures sur un certain nombre de variables ; ces mesures sont groupées dans un **vecteur d'observation** noté y. Pour utiliser l'information contenue dans ces mesures dispersées sur le domaine, il faut qu'on soit capable de la comparer au vecteur d'état. En réalité, on a besoin d'une fonction liant les deux vecteurs mentionnés. Il faut remarquer tout de suite le problème de support, car les mesures prises *in situ* ne correspondent que très rarement à un point de la grille du modèle. Ce problème est traité en considérant un opérateur H, appelé opérateur de projection des observations, reliant les observations y à x. La plupart du temps, cet opérateur est considéré linéaire et noté par $\mathbf{H}$ (la notation en gras est spécifique au cas linéaire). Le nombre de mesures disponibles peut varier d'un instant à l'autre, par conséquent, la

dimension du vecteur y, ainsi que le nombre de lignes de H peuvent varier aussi.

• **Les innovations et les résidus d'analyse**

L'élément clé d'une analyse est l'utilisation des différences entre le vecteur d'état et les mesures disponibles aux points de mesure, c'est-à-dire :

$$y - H(x). \tag{4.2}$$

Quand on utilise le vecteur d'ébauche, x^b, les différences sont appelées **innovations**, tandis que lorsqu'on utilise celui d'analyse, x^a, ces différences sont appelées **résidus d'analyse**.

4.4.2 Les erreurs

L'information cruciale que le système d'assimilation doit recevoir est une décision de modélisation des erreurs présentes dans le système. On parle premièrement d'*erreurs de background* :

$$\varepsilon^b = x^b - x^t, \tag{4.3}$$

de moyenne $\overline{\varepsilon^b}$ et de covariance :

$$P^b = \overline{(\varepsilon^b - \overline{\varepsilon^b})(\varepsilon^b - \overline{\varepsilon^b})^T}. \tag{4.4}$$

Le deuxième type d'erreur est celui *d'observation* :

$$\varepsilon^0 = y - H(x^t) \tag{4.5}$$

de moyenne $\overline{\varepsilon^0}$ et de covariance :

$$R = \overline{(\varepsilon^0 - \overline{\varepsilon^0})(\varepsilon^0 - \overline{\varepsilon^0})^T}. \tag{4.6}$$

Le dernier type d'erreur est l'*erreur d'analyse* définie par :

$$\varepsilon^a = x^a - x^t, \tag{4.7}$$

de moyenne $\overline{\varepsilon^a}$. Une mesure de l'erreur effectuée dans l'analyse est la trace de la matrice de covariance de l'erreur d'analyse, notée P^a :

$$Tr(P^a) = \overline{\|\varepsilon^a - \overline{\varepsilon^a}\|^2}. \tag{4.8}$$

Les moyennes calculées sur les erreurs sont appelées biais et elles font partie d'un problème systématique dans l'assimilation : une dérive dans le modèle, un biais dans les observations ou une erreur systématique dans la façon dont elles sont utilisées.

4.4.3 Formulation du problème

L'objet de notre étude est donc le vecteur d'état x^t. On suppose qu'on obtient une première idée sur ce vecteur par x^b résultant d'une analyse antérieure ou d'une estimation issue des principes généraux. Cela constitue la meilleure estimation de l'état du système en absence de toute autre

information. Les observations effectuées sur le système, rendues sous la forme du vecteur des observations y, à travers l'opérateur de projection noté H, renseignent sur cet état. De plus, on suppose que, idéalement, on connaît les statistiques jusqu'à l'ordre deux de l'erreur d'observation, ainsi que de l'erreur d'ébauche. On cherche alors, à l'aide des observations effectuées, à améliorer l'estimation de l'état du système x^a par rapport à la connaissance *a priori* de celui-ci, x^b. On cherche également à connaître l'erreur commise au cours de cette analyse.

4.4.4 Interpolation statistique

Quand on veut combiner les deux sources d'information dont nous disposons, on écrit x^a comme une combinaison linéaire entre le vecteur englobant les connaissances *a priori*, x^b, et le vecteur des observations y, donc sous la forme :

$$x^a = x^b + K(y - H(x^b)), \tag{4.9}$$

en intégrant ainsi dans cette équation les innovations. L'opérateur linéaire K qui permet cette approche s'appelle gain ou matrice de poids et il est obtenu en minimisant l'erreur scalaire commise dans l'analyse, c'est-à-dire $Tr(P_a)$, voir l'équation 4.8.

L'équation fondamentale d'une analyse linéaire est donnée sous la forme d'une estimation de moindres carrés, appelée aussi BLUE (en anglais-Best Linear Unbiased Estimation). Pour obtenir la formule de K on présentera d'abord quelques notations et hypothèses.

• **Notations** utilisées :

- x^t : état réel du système ("true" en anglais) (dim n) ;
- x^b : état prédit du système ("background" en anglais) (dim n) ;
- x^a : état analysé du système (dim n) ;
- y : vecteur d'observations (dim m) ;
- $\mathbf{H}$: opérateur d'observation ($m \leq \dim \leq n$) ;
- $\mathbf{P^b}$: matrice de covariance de l'erreur de prédiction ($x^b - x^t$) (dim $n \times n$) ;
- $\mathbf{P^a}$: matrice de covariance de l'erreur d'analyse ($x^a - x^t$) (dim $n \times n$) ;
- $\mathbf{R}$: matrice de covariance de l'erreur d'observation ($y - \mathbf{H}(x^t)$) (dim $m \times m$) ;

• **Les hypothèses :**

- l'opérateur d'observation $\mathbf{H}$ est linéaire ;
- Q et R sont des matrices positives définies ;
- les erreurs sont non-biaisées dans le sens où $\overline{x^b - x^t} = \overline{y - H(x^t)} = 0$;
- les erreurs ne sont pas corrélées $\overline{(x^b - x^t)(y - H(x^t))^T} = 0$;
- l'analyse est linéaire, comme décrite antérieurement ;
- l'analyse est optimale dans le sens où on impose la condition que l'état analysé soit le plus proche possible de l'état vrai, conformément au critère de variance minimale ;

• **Estimation optimale : analyse BLUE**

Pour trouver l'expression de la matrice de gain on écrit d'abord la relation mathématique qui conduit au calcul de la matrice de covariance de l'erreur d'analyse $\mathbf{P^a}$ à l'aide des vecteurs d'erreur introduits auparavant :

$$\varepsilon^a = \varepsilon^b + K(\varepsilon^0 - H(\varepsilon^b)). \tag{4.10}$$

Par conséquent, on obtient pour la matrice P^a l'expression suivante :

$$\begin{aligned} P^a &= E\left[(\varepsilon^a)(\varepsilon^a)^T\right] \\ &= E\left[(\varepsilon^b + K(\varepsilon^0 - H(\varepsilon^b)))(\varepsilon^b + K(\varepsilon^0 - H(\varepsilon^b)))^T\right] \\ &= E\left[((I-KH)\varepsilon^b + K\varepsilon^0)((I-KH)\varepsilon^b + K\varepsilon^0)^T\right] \\ &= E\left[(I-KH)\varepsilon^b(\varepsilon^b)^T(I-KH)^T\right] + E\left[K\varepsilon^0(\varepsilon^0)^T K^T\right] \\ &= (I-KH)P^b(I-KH)^T + KRK^T, \end{aligned} \tag{4.11}$$

où on a utilisé le fait que les erreurs ne sont pas corrélées et que l'opérateur K est linéaire. Par la suite on fait varier K de δK et on étudie la variation de $Tr(P^a)$. En effectuant le calcul (voir Bouttier et Courtier (1999) page 14 ou Bocquet (2004) page 20) on trouve pour K, en imposant la condition d'optimalité, la formule :

$$K = P^b H^T (R + HP^b H^T)^{-1}. \tag{4.12}$$

• **Equivalence avec le problème d'optimisation variationnelle**

L'analyse BLUE est équivalente à la solution du problème d'optimisation variationnelle suivant : $x^a = Arg\ min J$, avec

$$J(x) = (x - x^b)^T (P^b)^{-1}(x - x^b) + (y - H(x))^T R^{-1}(y - H(x)) \tag{4.13}$$

$$= J_b(x) + J_0(x). \tag{4.14}$$

J est appelée fonction de coût (ou de pénalité), J_b est le terme de background et J_0 celui d'observation. On peut démontrer que l'analyse x^a est optimale dans le sens où elle est la plus proche, dans le sens de moindres carrés, de l'état vrai x^t. Si, en plus, les distributions de probabilité des erreurs de background et d'observation sont gaussiennes, alors on obtient que x^a est également la solution de maximum de vraisemblance de x^t. Pour obtenir la solution du problème, on a besoin d'évaluer le gradient de la fonction J, qui peut être calculé en utilisant l'expression :

$$\nabla J(x) = 2(P^b)^{-1}(x - x^b) - 2H^T(P^b)^{-1}(y - H(x)) \tag{4.15}$$

pour approcher de cette façon le minimum désiré.

Revenant à l'interpolation statistique, on a construit donc les bases nécessaires pour définir et résoudre le problème d'une assimilation séquentielle qui tiendra compte en plus de **la dimension temporelle**. Il ne s'agit plus d'effectuer une seule analyse d'un vecteur d'état, compte tenu d'une ébauche et d'un ensemble d'observations ; on veut intégrer maintenant, outre le temps, un modèle d'évolution de l'état du système, dans lequel chaque *background* est créé par une ébauche calculée à partir d'une ancienne analyse.

4.4.5 Modèle stochastique

Une première information, dont nous disposons sur l'état du système, est le modèle dynamique M qui assure la transition temporelle entre deux états successifs du système. Dans le cadre de la problématique du transport réactif d'espèces chimiques, c'est le modèle de chimie-transport (CTM) qui assume ce rôle. Ce modèle regroupe toutes les informations sur la dynamique et les lois d'évolution du système, informations qui sont traduites dans un code de calcul. Bien sûr, ce modèle est entaché d'erreurs liées, soit à la description des processus invoqués, soit à la résolution numérique du problème. C'est la raison pour laquelle on effectue une extension de ce modèle déterministe à un modèle stochastique, en rajoutant une erreur (perturbation stochastique) supposée avoir une distribution gaussienne centrée, de moyenne nulle et matrice de covariance $Q(t_k)$:

$$x^t(t_{k+1}) = M(x^t(t_k)) + w(t_k), \tag{4.16}$$

où w est connu comme l'erreur du modèle et sa distribution est donnée par la formule suivante :

$$w \to N(0, Q(t_k)) \sim \exp\left[-\frac{1}{2} w^T Q(t_k)^{-1} w\right]. \tag{4.17}$$

Les observations, disponibles uniquement aux points de mesure sont, quant à elles, représentées par un vecteur $y(t_k)$ de dimension m, correspondant aux temps discrets du modèle. Comme il a déjà été évoqué, il faut remarquer le problème de support, car les mesures prises *in situ* ne correspondent que très rarement à un point de grille du modèle. Ce problème est traité en considérant un opérateur H, appelé opérateur de projection des observations, reliant les observations y à x. L'erreur qu'on rajoute, notée $v(t_k)$, comporte deux composantes : les erreurs de mesure dues à l'instrumentation, ainsi que les erreurs de représentativité. Leur somme, appelée erreur d'observation, affecte additivement le vecteur y :

$$y(t_k) = H(t_k) x^t(t_k) + v(t_k), \tag{4.18}$$

où le vecteur $v(t_k)$ est supposé avoir une distribution gaussienne centrée de moyenne nulle et de matrice de covariance $R(t_k)$, et il est supposé non-corrélé avec le vecteur d'erreur du modèle $w(t_k)$:

$$v(t_k) \to N(0, R(t_k)) \sim \exp\left[-\frac{1}{2} v(t_k)^T R(t_k)^{-1} v(t_k)\right]. \tag{4.19}$$

Généralement, on considère la matrice $R(t_k)$ connue, mais en pratique c'est un enjeu important d'avoir une estimation correcte de cette erreur, car l'analyse effectuée dépend de cette estimation, comme on verra dans la section 4.4.6. En général, on considère que $R(t_k)$ est diagonale, ce qui revient à supposer que les observations sont indépendantes les unes des autres.

On dispose maintenant de tous les éléments nécessaires pour présenter le filtre de Kalman, qui reste l'outil de base pour une assimilation séquentielle.

4.4.6 Le filtre de Kalman

Cet algorithme, introduit par Kalman (1960), a été construit initialement pour les systèmes linéaires : c'est-à-dire que le modèle M, dont on a fait référence, est linéaire (voir l'équation 4.16),

ainsi que l'opérateur des observations H (voir l'équation 4.18). L'idée principale était de réaliser des interpolations optimales successives en propageant entre deux échéances d'observation le vecteur d'état et la covariance d'erreur selon le modèle (voir la figure 4.1). Il faut évaluer l'estimée du vecteur d'état à l'instant t_{k+1} par récurrence, en utilisant pour cela l'estimée de ce même vecteur obtenue à l'instant précédent, t_k, ainsi que l'information acquise depuis cet instant. On n'utilise donc jamais la totalité des observations, les calculs sont donc simplifiés et rapides.

L'une des hypothèses clé de travail est la gaussiannité des distributions de probabilités pour

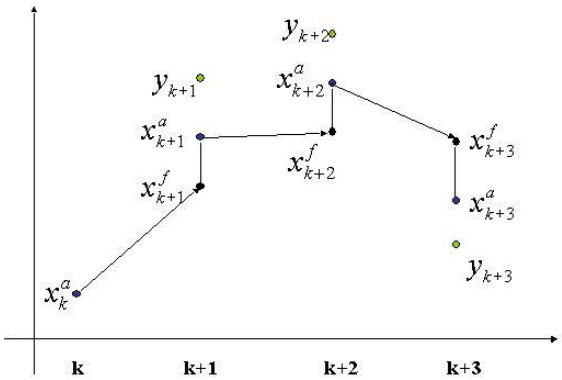


FIG. 4.1: Le schéma d'une procédure séquentielle d'assimilation de données. D'après Bocquet (2004).

les deux types d'erreur mentionnées : du modèle et des observations, qui sont d'ailleurs maintenues, car les opérateurs appliqués sont linéaires ; l'état du système devient ainsi lui-même gaussien, plus précisément le vecteur d'état est un processus de Gauss-Markov. Dans ce cas, les deux premiers moments statistiques (moyenne et matrice de covariance de l'erreur) qui caractérisent entièrement la distribution gaussienne sont les seuls à être calculés. L'autre hypothèse clé est l'indépendance entre les erreurs d'observation et celles du modèle. En utilisant cette approche classique, sur des systèmes linéaires, on obtient une estimation optimale au sens de variance minimale et de maximum de vraisemblance. Concernant les notations, on garde les mêmes que dans le cas de l'interpolation optimale, avec un petit changement pour la notion du *background* qui sera remplacée par *forecast* (en anglais) (faisant foi de la dimension temporelle) ; le vecteur x^b sera donc remplacé par x^f, l'erreur de l'ébauche sera notée par ε^f et la matrice de covariance de l'erreur d'ébauche P^f.

Après une phase d'initialisation (l'état initial étant considéré gaussien), le filtre de Kalman opère en deux étapes d'une façon récursive : une étape de prévision par le modèle, suivie par une étape de correction de la prévision lorsqu'une nouvelle observation devient disponible. Les équations du filtre de Kalman sont les suivantes :

Initialisation

L'état du système, ainsi que la matrice de covariance de l'erreur sont initialisés par :

$$x^a(t_0) = m_0, \tag{4.20}$$

$$P^a(t_0) = P_0. \tag{4.21}$$

1. **Étape de prévision (ébauche)** qui définit l'évolution de l'état :

$$x^f(t_{k+1}) = M(t_k, t_{k+1})x^a(t_k), \tag{4.22}$$

et celle de la matrice de covariance :

$$P^f(t_{k+1}) = M(t_k, t_{k+1})P^a(t_k)M(t_k, t_{k+1})^T + Q(t_k). \tag{4.23}$$

2. **Étape de correction (analyse)** qui consiste à remplacer la moyenne et la covariance prédites pendant l'étape de propagation, par des équivalents, étant donnée la nouvelle information devenue disponible (les mesures enregistrées au moment de l'analyse) :

$$x^a(t_{k+1}) = x^f(t_{k+1}) + K(t_{k+1})(y^0(t_{k+1}) - H(t_{k+1})x^f(t_{k+1})), \tag{4.24}$$

$$P^a(t_{k+1}) = P^f(t_{k+1}) - K(t_{k+1})H(t_{k+1})P^f(t_{k+1}) = (I - K(t_{k+1})H(t_{k+1}))P^f(t_{k+1}), \tag{4.25}$$

où la matrice de gain K est calculée en utilisant l'équation déjà présentée dans le cas de l'interpolation optimale (éq 4.12) :

$$K(t_{k+1}) = P^f(t_{k+1})H(t_{k+1})^T[H(t_{k+1})P^f(t_{k+1})H(t_{k+1})^T + R(t_{k+1})]^{-1}. \tag{4.26}$$

Ensuite, avec le vecteur d'état et la matrice de covariance de l'erreur analysés ($x^a(t_{k+1})$,$P^a(t_{k+1})$) on revient à l'étape d'ébauche.

En pratique, l'implémentation de ce filtre est loin d'être triviale. Si, théoriquement, ce filtre est stable, il peut diverger lors de sa mise en œuvre numérique à cause du caractère non-symétrique de l'équation 4.25. De plus, la sensibilité du filtre aux erreurs numériques est très connue. Plusieurs algorithmes ont été développés pour palier à cet inconvénient, notamment les algorithmes de type *racine carrée*.

En regardant attentivement les équations qui définissent le filtre de Kalman, on peut observer que la matrice de gain K ne dépend pas des observations. Donc, cette matrice peut être calculée en avance, car les observations ne sont pas nécessaires. Si le modèle est autonome (i.e. l'opérateur de transition M et celui d'observation H, ainsi que les matrices R et Q ne dépendent pas du temps), observable et contrôlable, on peut démontrer que le filtre de Kalman possède un régime permanent unique (Maybeck, 1979). Les matrices de gain K et de covariance de l'erreur d'analyse P_a convergent vers des valeurs limite K_∞ et P_∞.

En pratique, il est vraiment très difficile de montrer la convergence du filtre vers le régime permanent par avance. Néanmoins, comme les conditions mentionnées sont seulement suffisantes et pas nécessaires, elles n'impliquent pas que le filtre de Kalman ne convergera pas vers un régime permanent. Toutefois, il existe plusieurs algorithmes qui ont été proposés pour calculer la matrice de gain en régime permanent. Un inconvénient de cette approche asymptotique est qu'elle ne peut pas être utilisée dans les systèmes caractérisés par une forte non-linéarité ou quand la distribution spatio-temporelle des observations est irrégulière.

4.4.7 Le filtre de Kalman étendu (EKF)

Si le modèle n'est pas linéaire, l'estimation optimale de variance minimale ne peut pas être obtenue en utilisant uniquement les deux premiers moments, mais en utilisant un nombre infini de moments d'ordre supérieur (la distribution gaussienne ne reste pas gaussienne quand on lui applique un opérateur non-linéaire).

L'idée la plus simple pour palier à cet inconvénient est de contourner la non-linéarité du modèle dynamique en effectuant un développement en série de Taylor de l'opérateur non-linéaire qui intervient dans l'algorithme et en retenant par exemple uniquement le terme du premier ordre (linéarisation du premier ordre de l'opérateur de transition F_k non-linéaire) :

$$x(t_{k+1}) = F_k(x(t_k)), \tag{4.27}$$

alors dans les équations 4.22 et 4.23 on utilise comme opérateur M, l'opérateur F_k linéarisé au premier ordre :

$$M(t_k, t_{k+1}) = \frac{\partial F_k}{\partial x(t_k)}. \tag{4.28}$$

Cette extension du Filtre de Kalman est appelée Filtre de Kalman Étendu (en anglais EKF-**E**xtended **K**alman **F**ilter). Il existe un Filtre de Kalman Étendu du premier ordre, mais aussi du deuxième ordre et ainsi de suite ; on peut trouver une description détaillée de ces algorithmes, ainsi que quelques exemples dans Jaswinski (1970).

Avec cette linéarisation, le filtre de Kalman perd son caractère optimal. Toutefois, il a été appliqué avec succès dans des modèles faiblement non-linéaires.

Négliger les termes d'ordre supérieur dans la linéarisation est un inconvénient majeur de cet algorithme, car ceci peut introduire des instabilités dans le filtre lorsqu'il est appliqué à des modèles fortement non-linéaires (Evensen, 1994). Un deuxième inconvénient est que, dans le cas d'une application à grande échelle, quand la dimension du vecteur d'état est très grande, le stockage de la matrice de covariance d'erreur peut poser des problèmes. Une troisième difficulté est causée par la nécessité de développer un modèle linéaire tangent, difficile à construire et à maintenir.

Pour tenter d'améliorer les performances des filtres sur des systèmes dynamiques fortement non-linéaires, il est nécessaire de définir ce que devrait être le système d'assimilation qui rende compte de tous les moments des distributions statistiques. En effet, une dynamique non-linéaire propage les moments de façon non-triviale, à la différence d'une dynamique linéaire. Pour cette raison, on présentera dans la suite l'interprétation probabiliste du problème d'estimation par AD.

4.4.8 Interprétation probabiliste

On rappelle le problème initial à résoudre : compte tenu d'une connaissance *a priori* mais imparfaite du système, et d'un jeu plus récent d'observations, quel est l'état du système le plus probable, et quelle est la variance de l'erreur commise ?

En utilisant une approche probabiliste, le problème devient : connaissant la densité de probabilité de l'ébauche (background) ainsi que celle des observations, quelle est la densité de probabilité du système d'être dans état ou un autre ? L'instrument principal de l'approche probabiliste est le théorème de Bayes, qui fournit la densité de probabilité du vecteur x connaissant l'observation y (ou densité conditionnelle) :

$$p_{X|Y}(x|y) = p_{Y|X}(y,x)\frac{p_X(x)}{p_Y(y)}. \tag{4.29}$$

À partir de la distribution de probabilité conditionnelle $p_{X|Y}$, il faut définir un estimateur, dont la qualité principale souhaitée est d'être non-biaisé. Cet estimateur peut être celui de **variance minimum** ou celui **a posteriori de probabilité maximum** (en anglais-**maximum likelihood estimate** MLE). Dans le cas gaussien, les deux estimateurs coïncident. Une remarque concernant les notations de cette section : les distributions de probabilité ont été notées avec p et les densités qui correspondent aux distributions avec ϕ.

Dans le cadre d'une assimilation séquentielle sur un modèle non-linéaire, le problème d'estimation d'état est résolu formellement si on détermine **le maximum de vraisemblance de la densité conditionnelle aux données passées** $\phi(x_k|y_{1:k})$ connaissant la densité de probabilité de l'erreur du modèle (ou d'évolution) $\phi(x_{k+1}|x_k)$ et la densité de probabilité des observations $\phi(y_k|x_k)$, où $y_{1:k} = \{y_1, y_2 ... y_k\}$.

Rigoureusement, le fait d'ajouter une perturbation aléatoire (du bruit) aux équations différentielles déterministes qui régissent l'évolution de l'état du système (l'équation 4.16), les transforme en équations différentielles stochastiques d'Itô. La résolution théorique de telles équations est très difficile. Ce qui nous intéresse pour résoudre le problème de filtrage est de **déterminer l'équation d'évolution de la fonction de densité de probabilité conditionnelle**.

Si cette densité de probabilité existe, (elle sera notée $\phi(x_k|y_{1:k})$), elle obéira à l'équation de Fokker-Planck, appelée aussi équation directe de Kolmogorov :

$$\frac{\partial \phi}{\partial t} = L(\phi), \tag{4.30}$$

où L représente l'opérateur de diffusion, qui est défini par la formule :

$$L(\cdot) = -\sum_{i=1}^{n} \frac{\partial(\cdot F_i)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2(\cdot Q_{ij})}{\partial x_i \partial x_j} \tag{4.31}$$

dans laquelle Q représente la matrice de covariance de l'erreur du modèle et F_i est la i_{me} composante de l'opérateur de transition non-linéaire F.

L'équation de Fokker-Planck peut être résolue comme une équation aux dérivées partielles pour des petites dimensions de l'état ($N \leq 4$). Dans le cas d'une application avec des modèles de chimie-transport complexes et des dimensions du vecteur d'état très importantes, la discrétisation devient pratiquement impossible. C'est pour cette raison que les méthodes d'AD appliquées sur des cas réels évitent l'intégration explicite de l'équation Fokker-Planck et préfèrent par exemple les méthodes de type Monte Carlo.

Si on connaît la densité $\phi(x_{k-1}|y_{1:k-1})$, l'équation 4.30 permet de calculer $\phi(x_k|y_{1:k-1})$. En appliquant la règle de Bayes, utilisant les notations précédentes, on obtient la densité de probabilité *a posteriori* connaissant une nouvelle donnée y_k, ce qui nous permet de l'assimiler dans le système :

$$\phi(x_k|y_{1:k}) = \frac{\phi(x_k|y_{1:k-1})\phi(y_k|x_k)}{\int_{R^n} \phi(x|y_{1:k-1})\phi(y_k|x)dx}, \tag{4.32}$$

où au numérateur, la première densité de probabilité (*a priori*) est le résultat de l'intégration du modèle de $k-1$ à k, et la seconde est la densité des observations (Miller et al., 1999). Le dénominateur est un terme de régularisation pour que l'intégrale de la densité qui se trouve à gauche vaille 1.

Il reste pourtant très difficile de simuler la densité *a posteriori* (voir l'équation 4.32), principalement à cause de la complexité des évaluations du dénominateur qui requièrent l'évaluation des intégrales dans $\mathbb{R}^n$ (Bertino, 2001). Par conséquent, travailler avec *tous les moments statistiques* n'est pas envisageable. On se résume alors à utiliser des hypothèses et des schémas qui permettront le calcul des moments statistiques jusqu'à l'ordre 2 si possible, sinon on recourt à une linéarisation. Le prix à payer sera l'optimalité de l'estimation.

4.4.9 Nécessité d'appliquer des schémas sous-optimaux pour les systèmes réels

Le filtre de Kalman fournit l'estimateur optimal de l'état du système connaissant les observations ; cette optimalité, dans le sens du minimum de variance et de maximum de vraisemblance (on est dans le cas gaussien) est assurée uniquement dans le cas d'un système **linéaire**. La non-linéarité des systèmes atmosphériques est le principal inconvénient lorsqu'on souhaite appliquer le filtre de Kalman pour assimiler les observations. C'est la raison de l'extension faite avec le filtre étendu, où l'opérateur de transition est linéarisé au premier ordre. En revanche, *la stabilité* du filtre est perdue pour les systèmes non-linéaires.

D'autres inconvénients proviennent du fait que la matrice de covariance de l'erreur n'est plus qu'une *approximation* (due à la linéarisation), donc une source d'erreur difficile à quantifier. Un autre inconvénient est *la taille* de la matrice de covariance de l'erreur. La mise en œuvre du filtre exige la manipulation des matrices de covariance de l'erreur d'analyse ($P^a(t_k)$) et de l'ébauche ($P^f(t_k)$), de très grande taille : $n \times n$, où n représente la dimension du vecteur d'état, pouvant arriver jusqu'à l'ordre 10^6. Les matrices de covariance des erreurs du modèle, $Q(t_k)$, et des observations, $R(t_k)$, ne sont pas connues en pratique. La deuxième est plus accessible, même si elle n'est pas connue avec précision. La première, par contre, est loin d'être estimée avec précision car les sources des erreurs dans le modèle sont multiples et difficiles à quantifier.

Le seul moyen pour implémenter le filtre de Kalman sur des systèmes réels est d'utiliser des approximations pour **réduire la taille** et **simplifier l'évolution** de ces matrices, de façon à éviter, si possible, la linéarisation de l'opérateur d'évolution. En procédant de cette façon, on perd l'optimalité du filtre, mais malheureusement c'est le prix à payer pour appliquer cette technique sur les systèmes réels non-linéaires. Dans ce but, plusieurs formes dégradées du filtre ont été conçues (voir la section 4.2.1), et parmi celles-ci le Filtre de Kalman d'Ensemble. Introduit par Evensen (1994), cet algorithme a déjà été utilisé dans plusieurs domaines de recherche : océanographie, pollution

atmosphérique, météorologie ; il sera décrit en détail dans la section suivante.

4.4.10 Le Filtre de Kalman d'Ensemble (EnKF) : méthodologie

L'EnKF utilise une méthode Monte Carlo pour estimer la matrice de covariance de l'erreur en intégrant un ensemble d'états entre deux temps d'observations. Dans le cas d'un modèle non linéaire ou avec des matrices de covariance approximées, même avec des statistiques d'erreur gausiennes, les estimateurs ne sont plus gaussiens et ne coïncident plus. Les deux premiers moments ne suffisent plus à decrire les différentes densités de probabilité. Or, en pratique, dans la plupart des applications les moments d'ordre élevé ne sont pas calculés et donc négligés afin d'avoir autant d'inconnues que d'équations (fermeture du problème). L'approximation de fermeture étant inconsistante avec les dynamiques fortement non linéaires, une alternative a été introduite par l'EnKF. Cette approche est basée sur la représentation de la distribution de probabilité de l'estimée du vecteur d'état par un ensemble de N états possibles, notés ψ_1, ψ_2,... ψ_N, chaque membre de cet ensemble étant considéré comme un échantillon de la distribution de l'état réel ; l'idée derrière cet algorithme est le théorème de convergence de grands nombres. Les moments statistiques sont approximés par les statistiques des échantillons : on calcule les deux premiers moments empiriques (moyenne et covariance) sur les échantillons. L'EnKF gère les statistiques de l'erreur jusqu'à l'ordre 2, donc c'est un filtre gaussien. On doit supposer que les lois conditionnelles de l'état et des observations sont gaussiennes (elles seront notées $\phi(x_k^t|x_k^f)$ et $\phi(y_k|x_k^f)$) pour pouvoir appliquer une étape de correction linéaire et pour représenter la distribution de probabilité en utilisant uniquement la moyenne et la covariance comme dans le filtre classique.

L'idée principale est donc d'utiliser un ensemble d'estimations d'état (un nuage de points), obtenues en perturbant le vecteur d'état, à la place d'une seule, pour représenter la densité de probabilité du vecteur d'état du système, et de calculer la matrice de covariance de l'erreur sur cet ensemble. On obtient, de cette façon, un **ensemble d'états** du modèle qui évoluent dans l'espace de l'état sans aucune linéarisation. La moyenne de cet ensemble représente la *meilleure* estimation, et la variance de l'ensemble, la variance de l'erreur d'estimation. À chaque pas de temps, chaque observation est représentée par un autre ensemble, dont la moyenne est la mesure enregistrée, et la variance de l'ensemble représente l'erreur de mesure (Burgers et al., 1998) (justification et détails dans la section 4.4.10.2). Ainsi, on combine une étape de *prédiction stochastique* avec une étape d'*analyse stochastique* pour obtenir les statistiques correctes pour l'ensemble analysé. L'algorithme est plutôt intuitif ; ces principales étapes sont détaillées par la suite.

Principales étapes de l'algorithme

La première étape de l'algorithme, celle d'**initialisation**, consiste à générer N états du système $\psi^a(t_0)$ pour représenter l'incertitude dans l'état initial du système.

Dans la deuxième étape, de **propagation** ou **ébauche**, le modèle stochastique propage la distribution de l'état :

$$\psi_i^f(t_k) = F(\psi_i^a(t_{k-1}), w_i(t_k)) \text{ avec } i \in 1, \ldots, N \tag{4.33}$$

et calcule la meilleure estimation comme la moyenne sur l'ensemble :

$$x^f(t_k) = \frac{1}{N}\sum_{i=1}^{N} \psi_i^f(t_k), \tag{4.34}$$

tandis que la matrice de covariance de l'erreur est elle aussi calculée sur l'ensemble :

$$P^f(t_k) = \frac{1}{N-1} E^f(t_k) E^f(t_k)^T, \tag{4.35}$$

où

$$E^f(t_k) = [\psi_1^f(t_k) - x^f(t_k), \psi_2^f(t_k) - x^f(t_k), ..., \psi_N^f(t_k) - x^f(t_k)]. \tag{4.36}$$

La troisième étape est celle de **correction** ou **analyse**. Chaque fois que les mesures deviennent disponibles, elles sont utilisées pour actualiser la moyenne de l'ensemble en utilisant la formule :

$$\psi_i^a(t_k) = \psi_i^f(t_k) + K(t_k)[y(t_k) - H(t_k)\psi_i^f(t_k)], \tag{4.37}$$

où la matrice de gain K a pratiquement la même expression qu'avant (voir les équations 4.12 et 4.26) :

$$K(t_k) = P^f(t_k)H(t_k)^T[H(t_k)P^f(t_k)H(t_k)^T + R(t_k)]^{-1}. \tag{4.38}$$

Dans les équations précédentes, on a gardé les mêmes notations pour l'ébauche et la matrice de covariance de l'erreur associée à cette ébauche (x^f, P^f). La seule différence, par rapport au KF traditionnel, est que dans l'EnKF la connexion entre les deux est encore plus évidente : dans l'équation 4.35 le vecteur x^f intervient directement par l'intermédiaire de l'équation 4.36. Pour la *meilleure* estimation du vecteur d'état x^a, on utilise une moyenne calculée sur les différentes réalisations du vecteur d'état analysé ψ_i^a :

$$x^a(t_k) = \frac{1}{N}\sum_{i=1}^{N} \psi_i^a(t_k), \tag{4.39}$$

tandis que la matrice de covariance de l'erreur analysée est calculée exactement comme celle de l'ébauche sur l'ensemble, en utilisant la formule :

$$P^f(t_k) = \frac{1}{N-1} E^a(t_k) E^a(t_k)^T, \tag{4.40}$$

où

$$E^a(t_k) = [\psi_1^a(t_k) - x^a(t_k), \psi_2^a(t_k) - x^a(t_k), ..., \psi_N^a(t_k) - x^a(t_k)]. \tag{4.41}$$

Avantages et limitations

Un premier avantage du filtre d'ensemble est que la matrice de covariance de l'ébauche P^f est par construction positive définie ; deuxièmement, le modèle linéaire tangent n'est pas nécessaire, car les états du système sont propagés en utilisant l'opérateur original, sans aucune linéarisation (Evensen, 2004).

Parmi les inconvénients de l'EnKF, on peut remarquer que l'écart type des erreurs dans l'état du système, qui est de nature statistique, converge lentement car il dépend directement de N, le nombre de membres d'ensemble utilisés pour représenter les états possibles du système. Un autre inconvénient est la possible divergence du filtre, qui arrive quand le filtre commence à accorder trop

de confiance au modèle, en ignorant les observations. On peut éviter ce problème, en augmentant la covariance de l'ensemble. Une méthode pour résoudre cet inconvenient est d'utiliser une inflation soit additive (Corazza et al., 2002), soit multiplicative (Anderson, 2001) (voir la section 4.4.10.5). Une deuxième méthode est la localisation de la covariance qui consiste à limiter le rayon d'influence qu'une observation de surface peut avoir sur les mailles du modèle pour corriger les concentrations simulées par CHIMERE.

4.4.10.1 La représentation des statistiques des erreurs

Dans le filtre de Kalman, les matrices de covariance de l'erreur pour l'ébauche P^f et pour l'état analysé P^a sont représentées à l'aide de l'état vrai du système par les formules :

$$P^f = \overline{(x^f - x^t)(x^f - x^t)^T}, \tag{4.42}$$

$$P^a = \overline{(x^a - x^t)(x^a - x^t)^T}, \tag{4.43}$$

Toutefois, on ne connaît jamais l'état vrai du système ; dans le filtre d'ensemble on utilise alors les formules :

$$P_e^f = \overline{(x^f - \overline{x^f})(x^f - \overline{x^f})^T} \tag{4.44}$$

$$P_e^a = \overline{(x^a - \overline{x^a})(x^a - \overline{x^a})^T} \tag{4.45}$$

où le tiré ($\overline{x^a}$ ou $\overline{x^f}$) symbolise la moyenne sur l'ensemble. Par conséquent, dans la représentation utilisée, c'est la moyenne de l'ensemble qui est considérée comme la meilleure estimation, et dans ce cas, la covariance de l'ensemble peut être interprétée comme la covariance de l'erreur de la meilleure estimée.

4.4.10.2 Le schéma d'analyse

L'étape d'analyse de l'algorithme KF utilise les formules (4.42, 4.43) pour calculer les deux matrices de covariance de l'erreur. Pour dériver les équations correspondantes dans l'algorithme EnKF, celles qui utilisent les formules 4.44 et 4.45, Burgers et al. (1998) ont montré qu'il est cohérent de traiter les observations comme des variables aléatoires et d'effectuer l'analyse sur chaque membre de l'ensemble en utilisant des observations différentes (perturbées) pour éviter ainsi la réduction trop importante de la matrice de covariance d'analyse. Quelle est l'explication d'un tel procédé ? Si on regarde l'équation d'analyse pour chaque membre de l'ensemble (on simplifie la notation en supprimant le t_k) :

$$\psi_i^a = \psi_i^f + K_e[y - H\psi_i^f], \;\; i = 1, ...N \tag{4.46}$$

où la matrice de gain est calculée comme :

$$K_e = P_e^f H^T [H P_e^f H^T + R]^{-1}, \tag{4.47}$$

et on écrit l'équation d'analyse pour la moyenne de l'ensemble, cela revient à écrire :

$$\overline{\psi^a} = \overline{\psi^f} + K_e[y - H\overline{\psi^f}]. \tag{4.48}$$

Par conséquent, si on considère la matrice de perturbations de l'ensemble qui sera notée X', avec ses deux variantes X'^f et X'^a pour l'ébauche et l'analyse, celles-ci satisferont la relation :

$$X'^a = X'^f - K_e H X'^f = (I - K_e H) X'^f; \tag{4.49}$$

ainsi, la covariance de l'ensemble sera calculée d'après la formule :

$$P_e^a = (I - K_e H) P_e^f (I - K_e H)^T. \tag{4.50}$$

En comparant cette équation à celle présentée dans le filtre KF original, (éq 4.25 et 4.11), on observe qu'elle est sous-estimée. Pour palier à cet inconvénient, Burgers et al. (1998) ont proposé de perturber chaque observation, donc de créer un ensemble de perturbations :

$$y_i = y + \varepsilon_i, \text{ avec } \varepsilon_i \sim N(0, R), \tag{4.51}$$

dont l'écriture matricielle utilisée pour ce nouveau ensemble sera Y, tandis que la matrice de perturbations des observations sera notée Y'. La matrice de covariance de l'ensemble d'observations, notée R_e est utilisée dans l'étape de correction par la matrice de gain :

$$K_e = P_e^f H^T [H P_e^f H^T + R_e]^{-1}. \tag{4.52}$$

L'équation d'analyse écrite pour chaque membre de l'ensemble

$$\psi_i^a = \psi_i^f + K_e [y_i - H \psi_i^f], \tag{4.53}$$

devient pour la moyenne de l'ensemble :

$$\overline{\psi^a} = \overline{\psi^f} + K_e [\overline{y} - H \overline{\psi^f}], \tag{4.54}$$

où $\overline{y}$ représente la moyenne synthétique de l'ensemble d'observations. On compare cette équation avec celle du KF original (4.48) et on remarque que la seule différence consiste dans le fait que y a été remplacé par $\overline{y}$. Ce problème est résolu si on impose la contrainte $\overline{\varepsilon} = 0$ pour ainsi s'assurer que $\overline{y} = y$.

La matrice de perturbation de l'ensemble devient :

$$X'^a = (I - K_e H) X'^f + K_e Y', \tag{4.55}$$

et la matrice de covariance de l'erreur aura (après toutes les réductions - voir le calcul dans Van Leeuwen (1998)) l'expression :

$$P_e^a = (I - K_e H) P_e^f + (I - K_e H) X'^f Y'^T K_e^T + K_e Y' X'^{fT} (I - K_e H)^T. \tag{4.56}$$

Si la condition $\overline{(\psi^f - \overline{\psi^f})\varepsilon^T} = 0$ est vérifiée, les deux derniers termes de l'expression 4.56 s'annulent et l'expression de la matrice de covariance d'erreur est la même que celle du filtre de Kalman original.

4.4.10.3 Le schema d'analyse de type racine carrée

Pour diminuer l'effort computationnel effectué pendant l'étape d'analyse, plusieurs auteurs ont proposé d'éliminer la **perturbation des observations** qui peut introduire une erreur d'échantillonnage supplémentaire dans le système. Quelques méthodes pour calculer l'analyse, sans introduction des perturbations dans les mesures, ont été développées et présentées par Anderson (2001), Whitaker et Hamill (2002) et Tippett et al. (2003). Dans cette étude, on présente un schéma direct et très simple, appelé "schéma de type racine carrée" (Evensen, 2004), qui est utilisé pour obtenir l'état analysé du système, qui évite de faire des hypothèses supplémentaires sur le manque des corrélations entre les erreurs de mesure.

En utilisant les notations originales d'Evensen (2004), on définit d'abord la **matrice d'ensemble**, celle qui contient tous les N membres d'ensemble (n reste la dimension du vecteur d'état) :

$$A = (\psi_1, \psi_2, ..., \psi_N) \in \Re^{n \times N}. \tag{4.57}$$

Pour stocker la **moyenne de l'ensemble** dans chaque colonne de la matrice $\overline{A}$, on peut écrire :

$$\overline{A} = A \times \mathbf{1}_N, \tag{4.58}$$

où $\mathbf{1}_N \in \Re^{N \times N}$ est la matrice avec le même et unique élément sur chaque ligne $1/N$, et pour la **matrice de perturbations de l'ensemble** on utilise la notation A', où

$$A' = A - \overline{A} = A(I - \mathbf{1}_N). \tag{4.59}$$

On peut définir alors la **matrice de covariance d'erreur de l'ensemble** comme :

$$P_e = \frac{A'(A')^T}{N-1}. \tag{4.60}$$

On introduit maintenant la **matrice de mesure** S pour les perturbations de l'ensemble :

$$S = HA' \in \mathbb{R}^{m \times N}, \tag{4.61}$$

où m représente le nombre d'observations utilisées à un moment donné, dans l'algorithme d'assimilation, qui permet de calculer la matrice $C \in \mathbb{R}^{m \times m}$:

$$C = SS^T + (N-1)R. \tag{4.62}$$

Si la matrice C est inversible, la décomposition en valeurs propres s'écrit sous la forme :

$$C^{-1} = Z\Lambda^{-1}Z^T, \tag{4.63}$$

où les matrices Z et Λ ont la même dimension, $m \times m$. Si m est grand, cela peut devenir la partie la plus coûteuse du point de vue temps de calcul.

L'équation de l'analyse écrite en termes de matrice de covariance de l'erreur s'écrit :

$$A^{a\prime}A^{a\prime T} = A'(I - S^T C^{-1} S)A'^T, \tag{4.64}$$

où A^a représente la matrice de l'ensemble analysé.

De plus, si on tient compte de la formule précédente (4.63) on obtient :

$$\begin{aligned} A^{a\prime}A^{a\prime T} &= A'(I - S^T Z\Lambda^{-1}Z^T S)A'^T && (4.65)\\ &= A'(I - (\Lambda^{-\frac{1}{2}}Z^T S)^T(\Lambda^{-\frac{1}{2}}Z^T S)A'^T && (4.66)\\ &= A'(I - X_2^T X_2)A'^T, && (4.67)\end{aligned}$$

où $X_2 \in \mathbb{R}^{m\times N}$ a été défini comme le produit $X_2 = \Lambda^{-\frac{1}{2}}Z^T S$, et dont le rang est $rang(X_2) = min(m, N-1)$.

En décomposant maintenant en valeurs singulières la nouvelle matrice X_2 :

$$U_2\Sigma_2 V_2^T = X_2 \tag{4.68}$$

avec $U_2 \in \mathbb{R}^{m\times m}$, $\Sigma_2 \in \mathbb{R}^{m\times N}$ et $V_2 \in \mathbb{R}^{N\times N}$, l'équation 4.67 permettant d'actualiser la matrice de covariance devient :

$$\begin{aligned} A^{a\prime}A^{a\prime T} &= A'(I - [U_2\Sigma_2 V_2^T]^T[U_2\Sigma_2 V_2^T])A'^T && (4.69)\\ &= A'(I - V_2\Sigma_2^T\Sigma_2 V_2^T)A'^T && (4.70)\\ &= A'V_2(I - \Sigma_2^T\Sigma_2)V_2^T A'^T && (4.71)\\ &= \left(A'V_2\sqrt{I - \Sigma_2^T\Sigma_2}\right)\left(A'V_2\sqrt{I - \Sigma_2^T\Sigma_2}\right)^T. && (4.72)\end{aligned}$$

Finalement, la solution pour analyser les perturbations de l'ensemble s'exprime :

$$A^{a\prime} = A'V_2\sqrt{I - \Sigma_2^T\Sigma_2}, \tag{4.73}$$

forme considérée comme standard pour le schéma d'analyse de type *racine carrée* et qui produit un ensemble de perturbations de variance correcte.

Pour l'algorithme de l'EnKF les pas à suivre dans l'implémentation sont :

1. Génération de l'ensemble (la matrice A)
2. Calcul de $\overline{A}$ et de A'
3. Calcul de la matrice C
4. Décomposition de la matrice C calculée comme : $Z\Lambda Z^T$

5. Actualisation de la moyenne en utilisant la formule :

$$\overline{\psi^a} = \overline{\psi^f} + A'S^T Z\Lambda^{-1}Z^T(d - H\overline{\psi^f}), \tag{4.74}$$

en suivant la séquence de transformations vectorielles :

- $y_1 = Z^T(d - H\overline{\psi^f})$
- $y_2 = \Lambda^{-1}y_1$

- $y_3 = Zy_2$
- $y_4 = S^T y_3$
- $\overline{\psi^a} = \overline{\psi^f} + A' y_4$

6. Calcul de la matrice intermédiaire : $X_2 = \Lambda^{-\frac{1}{2}} Z^T S$
7. Décomposition $X_2 = U_2 \Sigma_2 V_2^T$
8. Évaluation de l'analyse d'ensemble de perturbations $A^{a\prime} = A' V_2 \sqrt{I - \Sigma_2^T \Sigma_2}$ qu'on rajoute à la moyenne prédite d'ensemble $\overline{\psi^f}$.

4.4.10.4 Construction de champs pseudo-aléatoires

Pour effectuer la perturbation des champs de concentrations de polluants à chaque pas de temps, on utilise dans cette étude la méthode de *génération de champs pseudo-aléatoires* décrite en détail par Evensen (1994). Elle consiste à produire des champs 2D caractérisés statistiquement par une fonction de covariance bien précise, (dans notre cas elle est gaussienne, isotrope), de moyenne nulle et variance égale à 1.

Une méthode extrêmement efficace pour générer les champs pseudo-aléatoires est la transformée de Fourier rapide (FFT- en anglais Fast Fourier Transform). Le soft utilisé dans ce travail, pour produire ces champs est celui d'Evensen, détaillé dans son livre Evensen (2006) et basé sur le software FFTW[4] (Fastest Fourier Transform in the West) développé par Matteo Frigo et Steven G. Johnson.

Revenant au filtre d'ensemble, on continue par la présentation succincte des quelques remarques concernant un problème très important : la divergence du filtre.

4.4.10.5 Divergence du filtre

Le problème principal dans les applications réelles est la divergence du filtre d'ensemble, quand les observations reçoivent des poids négligeables et le filtre suit le modèle sans le corriger.

Pour éviter la tendance du filtre à diverger, une première démarche, proposée par Houtekamer et Mitchell (1998) a été de diviser les membres d'ensemble en deux groupes distincts ; ainsi on peut utiliser la covariance de l'erreur d'ébauche calculée sur le premier groupe, pour effectuer l'analyse sur le deuxième, et l'analyse du deuxième, pour calculer l'ébauche sur le premier groupe. Ceci peut prévenir la diminution de la covariance d'erreur de la prédiction, donc la divergence du filtre.

[4] http ://www.fftw.org/

Deux autres solutions ont été proposées pour prévenir la divergence du filtre, conduisant à deux variantes de l'EnKF : ETKF - Ensemble Transform Kalman Filter (Bishop et al., 2001) et EAKF - Ensemble Adjustement Kalman Filter (Anderson, 2001). Dans ces cas, l'analyse est calculée sans rajouter des perturbations aux observations. Les deux derniers algorithmes sont basés sur l'augmentation de la covariance d'erreur par une opération d'addition ou multiplication des différences entre les membres de l'ensemble et la moyenne, par un facteur r, dont le but est de supprimer la sous-représentation de la variance, due à l'utilisation d'un ensemble d'une dimension pas assez grande :

$$\psi_i^f \longleftarrow r(\psi_i^f - x^f) + x^f. \tag{4.75}$$

Une autre solution, proposée par la même équipe (Houtekamer et Mitchell, 2001), est de localiser la covariance d'erreur par l'application d'un produit Shur avec une fonction de corrélation, démarche effectuée par Houtekamer et Mitchell (2001) et reprise par Constantinescu et al. (2007) et qui apparemment, en pratique, donne de bons résultats.

4.4.11 Avantages et inconvénients. Justification du choix

En théorie, dans le cadre linéaire, les deux approches variationnelle et séquentielle sont équivalentes, elles doivent conduire au même résultat. En revanche, dans les cas réels qui sont souvent non linéaires, les deux techniques se compliquent considérablement et l'équivalence n'est plus démontrée.

Dans cette étude, on a choisi d'appliquer une méthode séquentielle d'AD, notamment le Filtre de Kalman d'Ensemble. La principale raison de notre choix est le fait que la méthode variationnelle nécessite l'écriture d'un code adjoint en relation étroite avec le modèle direct, ce qui n'est pas le cas du filtre de Kalman qui permet une implémentation modulaire indépendante du code numérique du modèle. De plus, en parlant d'assimilation séquentielle, on peut décrire cette technique comme un algorithme permettant de corriger l'état du modèle, au fur et à mesure que des nouvelles données arrivent ; c'est le point de vue du concepteur de modèles numériques déterministes. Le deuxième est celui d'un statisticien ; pour lui, il est important d'améliorer la prédiction opérationnelle en profitant des relations non linéaires entre les différentes sources de données.

En ce qui concerne le choix précis de l'EnKF parmi les techniques d'assimilation séquentielle, ceci est dû au fait que, premièrement, parmi les schémas sous-optimaux mentionnés précédemment, celui-ci est plus facile à implémenter. Deuxièmement, les équipes qui ont déjà comparé les performances de ces techniques séquentielles sur les systèmes non-linéaires, recommandent l'utilisation de l'EnKF.

4.5 Le modèle de chimie-transport CHIMERE

Cette section est dédiée à la description du modèle CHIMERE multi-échelle et, plus précisément, à la version V200211K[5] utilisée dans le cadre de cette étude. Ce modèle a été développé suite

[5] http ://euler.lmd.polytechnique.fr/chimere/

à une collaboration entre les laboratoires LMD[6] de l'École polytechnique, LISA[7] de l'Université Paris XII, IPSL[8] et l'INERIS[9].

L'objectif de ce modèle est de simuler les variations spatio-temporelles des concentrations d'une quarantaine d'espèces gazeuses liées à la photochimie de la basse troposphère. L'approche, de type Eulérien, est basée sur l'équation de continuité de concentration des différentes espèces chimiques. Si **c** désigne la concentration d'une espèce chimique dans la phase gazeuse, son évolution est régie par l'équation d'advection-diffusion-réaction suivante :

$$(\frac{\partial c}{\partial t}) + \nabla(uc) = \nabla(K\nabla c) + P - L. \tag{4.76}$$

Dans cette équation on peut donc considérer **c** comme le vecteur contenant les concentrations de toutes les espèces chimiques dans tous les points de la grille du modèle, **u** est le vecteur vitesse du vent, et **K** le tenseur des diffusivités turbulentes. **P** et **L** représentent respectivement les termes de production et de pertes (dues aux réactions chimiques, aux émissions et au dépôt sec) (Blond, 2002).

La version continentale de CHIMERE couvre l'Europe de l'Ouest, mais avec une résolution spatiale assez faible (environ 50 km), de sorte que ce modèle ne peut pas être utilisé pour simuler les panaches dans les agglomérations urbaines. En revanche, les valeurs des concentrations simulées par cette version peuvent être utilisées comme conditions aux limites pour la version régionale du modèle. Cette dernière version, utilisée dans ce travail, présente une résolution spatiale de 6×6 km avec un maillage constitué de 25×25 cellules (voir la figure 4.2). Verticalement, le modèle décrit les concentrations dans 8 couches verticales couvrant l'ensemble de la couche limite, de la surface jusqu'à 500 hPa.

Comme **données météorologiques** d'entrée, le modèle nécessite les variables standard des modèles numériques de prévision en chacun des nœuds du maillage : vent horizontal, température, humidité spécifique, pression de surface, couverture nuageuse de trois types différents : les nuages bas, de moyenne altitude et hauts (Blond, 2002). Toutes les entrées météorologiques sont fournies par ECMWF[10]. Elles correspondent aux prévisions à 6 heures et présentent une résolution horizontale de 0,5 degré. À la première vue, cette résolution n'est pas assez fine par rapport à la résolution du modèle, mais dans ce cas méthodologique, on peut considérer que le champ météorologique utilisé est suffisamment représentatif et que les erreurs induites ne sont pas aussi importantes.

Comme **mécanisme chimique**, CHIMERE utilise une version réduite du MELCHIOR (Modèle d'Etude Lagrangienne de la CHImie de l'Ozone à l'échelle Régionale), développé au sein du service d'Aéronomie du CNRS[11]. La version réduite de ce mécanisme contient 44 espèces chimiques condensées dans 116 réactions chimiques.

Le modèle avec son mécanisme réduit requiert en entrée les **émissions** de 16 espèces chimiques essentielles dans la photochimie. Les émissions anthropiques visent 5 espèces : le NO_x, le SO_2, le CO,

[6] Laboratoire de Météoroologie Dynamique, Palaiseau
[7] Laboratoire Inter-Universitaire des Systèmes Atmosphériques, Créteil
[8] Institut Pierre-Simon Laplace / Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette
[9] Institut National de l'Environnement Industriel et des Risques
[10] European Center for Medium-range Weather Forecasts
[11] http ://www.aero.jussieu.fr

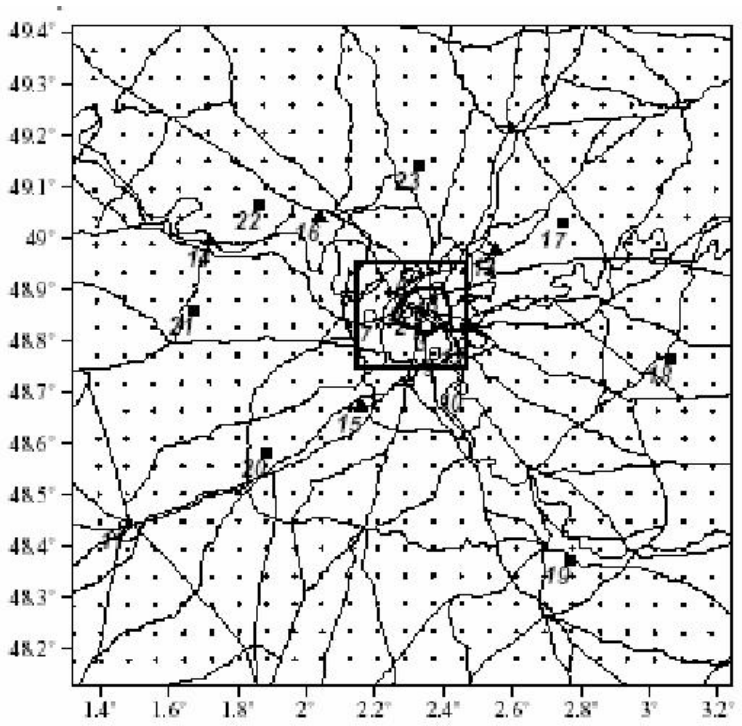


FIG. 4.2: Le maillage du domaine pour la version régionale du modèle CHIMERE. Le carré noir, situé au centre, correspond à la région parisienne. Les coordonnées sont exprimées en latitude, longitude.

le CH_4 et les Composées Organiques Volatiles Non Méthaniques. Les émissions biogéniques de COV portent sur l'isoprène et le terpène. Toutes les émissions, exprimées en termes de flux surfaciques sont fournies pour la région d'Île-de-France par AIRPARIF, et elles sont données pour un "jour moyen" défini selon le jour de la semaine et de l'heure de la journée. Elles sont après modifiées de sorte que leurs totaux annuels correspondent à ceux fournis par le CITEPA[12] par grands secteurs d'activité.

La méthode numérique pour résoudre le système d'équations différentielles (4.76) est une adaptation de l'algorithme du second ordre TWOSTEP proposé par Verwer (1994). Le système non-linéaire obtenu par ce schéma est résolu par la méthode Gauss-Seidel à chaque itération, le temps d'intégration étant fixé à $2,5$ minutes.

Enfin, il faut souligner encore une fois que le modèle régional prend comme **conditions aux limites** les concentrations de 14 espèces, simulées par la version continentale, interpolées sur les bords du domaine. En fonction de la direction du vent, ces concentrations, notamment les espèces lentes, aux bornes du domaines sont advectées à l'intérieur par le schéma numérique de transport PPM (Parabolic Piecewise Method) développé par Colella et Woodward (1984). Pour les concentrations initiales, la simulation est lancée dix jours avant la période d'intérêt, afin de laisser le temps au modèle d'arriver à l'équilibre.

La prévision est souvent considérée comme l'aboutissement du développement et de l'utilisa-

[12]Centre Interprofessionel Technique d'Etude de la Pollution Atmosphérique

tion d'un modèle. Mais c'est aussi le meilleur moyen de juger objectivement des capacités d'un outil. Dès le début du développement de CHIMERE, une version de prévision expérimentale a été utilisée quotidiennement. Le meilleur moyen d'améliorer le système, d'apprécier ses forces et ses faiblesses pour les corriger ensuite, a été la comparaison des mesures avec ce qui était prévu la veille par le modèle. Chaque année, une nouvelle version est proposée, version qui intègre des enseignements importants qui ont permis au modèle d'évoluer et de s'enrichir.

4.6 Évaluation du modèle CHIMERE à l'aide des observations. Etudes de sensibilité.

L'évaluation d'un modèle, faite en comparant ses sorties à des mesures effectuées par les stations du réseau de surveillance, est une étape indispensable qui permet d'évaluer les marges d'incertitude à prendre en compte sur tout résultat fourni par le modèle.

Sur CHIMERE, une première évaluation a été faite dans le cadre de la thèse de Nadège Blond (Blond, 2002). Les périodes testées sont les étés 1999, 2000 et 2001. Les comparaisons des sorties du modèle CHIMERE à des observations de surface d'**ozone**, ainsi qu'à des mesures aéroportées, suggèrent un comportement assez "réaliste" du modèle en ce qui concerne ce polluant. Par contre, les concentrations de **dioxyde d'azote** sont moins bien reproduites. Le nombre d'erreurs supérieures à 15 ppb commises par le modèle régional sur les concentrations d'ozone est de 10% en Île-de-France. De plus, 40% des dépassements du seuil d'information et de recommandation concernant l'ozone (90 ppb) ont été mal simulés (erreurs supérieures à 15 ppb) sur les stations parisiennes. Pour le dioxyde d'azote, 30% des erreurs commises sont supérieures à 15 ppb. Les dépassements du seuil de 50 ppb sont systématiquement mal représentés, dû aux faibles résolutions horizontale et verticale du modèle ainsi que aux incertitudes liées aux émissions.

La période qui nous intéresse le plus est l'été 1999, plus précisément le mois de Juillet avec deux épisodes de forte pollution par l'ozone : le 17 et le 18 Juillet. D'une manière générale, pour les épisodes mentionnés, le modèle régional sous-estime les niveaux d'ozone : pour le premier épisode uniquement dans le panache, tandis que, pour le deuxième, sur tout le domaine (voir les cartes 4.20(a) et 4.22(a)). On remarque également un léger décalage dans le positionnement du panache du 18 Juillet 1999, dû au fait que ce jour-là, le vent horizontal était faible (inférieur à $3\,\mathrm{m.s^{-1}}$) et le modèle est particulièrement sensible aux données météorologiques pendant les épisodes stagnantes (Vautard et al., 2001). La comparaison pour cette période a été effectuée en utilisant les vols des avions DIMONA ou ARAT de la campagne ESQUIF en 1999, qui ont montré, par ailleurs, que la version régionale de CHIMERE reproduit assez correctement les concentrations d'ozone en altitude (la remarque vise les niveaux 2 et 3 du modèle). En revanche, le modèle régional simule moins bien les concentrations de dioxyde d'azote, et les raisons sont multiples ; premièrement, il s'agit d'un polluant lié directement aux émissions qui sont fortement incertaines et, deuxièmement, les pics apparaissent lors de la progression de la couche de mélange, quand le modèle est plus sensible aux paramètres (Blond, 2002). L'évaluation des performances du modèle par rapport aux mesures a mis en évidence des erreurs systématiques, notamment la nuit et pendant les épisodes de forte pollution,

mais aussi des erreurs aléatoires.

C'est la principale raison pour laquelle, dans la seconde partie de sa thèse, N. Blond a essayé de corriger l'erreur globale commise sur les simulations des champs de concentrations de polluants, à l'aide des observations de surface, en utilisant deux **méthodes séquentielles d'assimilation de données**, notamment l'**interpolation optimale** (Daley, 1991) et le **krigeage intrinsèque sur les observations** (OBK) et **sur les innovations** (INK) (Blond, 2002). La conclusion de son étude montre que la combinaison modèle numérique-observations de surface permet d'améliorer la représentation des champs des concentrations de polluants : pour l'ozone il est possible de représenter les pics à moins de 14 μg.m^{-3} près sur l'Île-de-France, le nombre d'erreurs supérieures étant limité à 2%, tandis que, pour le dioxyde d'azote, le système d'assimilation est capable de représenter les pics de concentration à moins de 14 μg.m^{-3}, avec un nombre d'erreurs supérieures commises inférieur à 5%.

D'un autre côté, le développement du modèle adjoint de CHIMERE a permis plusieurs études de sensibilité des pics de pollution à différents paramètres, comme les émissions ou bien les constantes de réaction chimique (Menut et al., 2000). Par ailleurs, plusieurs études ont été menées à l'aide de CHIMERE afin de développer une méthodologie originale d'estimation de l'incertitude sur les résultats d'une simulation de scénarios de réduction d'émissions, voir par exemple Beekmann et Derognat (2003). La méthode est basée sur des simulations de Monte-Carlo contraintes par les observations : on cherche les combinaisons de paramètres, dans une certaine gamme d'incertitude, produisant des simulations proches des observations. Ces combinaisons de paramètres sont ensuite utilisées pour effectuer des scénarios de réduction d'émissions et évaluer l'éventail d'incertitude. Il est à noter que seules les incertitudes liées aux processus déroulés sur le domaine ont été prises en compte, par conséquent, les conditions aux limites ont été gardées fixes. En revanche, l'étude vise l'évaluation de l'incertitude sur plusieurs paramètres d'entrée, comme : les émissions, les paramètres météorologiques, les taux chimiques, les fréquences de photolyse.

4.7 La mise en œuvre du Filtre de Kalman d'Ensemble

Une fois le modèle de chimie-transport utilisé dans cette étude, CHIMERE, ayant été décrit, cette section sera consacrée à la description détaillée du système d'assimilation implémenté sur CHIMERE, en utilisant la méthode mentionnée et décrite elle-aussi auparavant, l'EnKF (voir la section 4.4.10).

4.7.1 Vecteur d'état

On commence la présentation de notre système d'assimilation par l'objet principal de notre étude, notamment le vecteur d'état. Il est composé de toutes les concentrations, dans tous les nœuds du maillage qui couvre le domaine présenté (voir la figure 4.2), en considérant aussi l'étendue sur la verticale (les 8 niveaux du modèle) pour toutes les espèces chimiques prises en compte dans les calculs (elles sont en nombre de 44) ; cela nous conduit à une dimension très grande pour notre vecteur d'état : 220 000 éléments.

4.7.2 Période d'étude

La période d'étude choisie dans cette partie du travail est la deuxième décade du mois de Juillet 1999, plus précisément entre le 11 et le 20 Juillet 1999, quand l'atmosphère de la région parisienne a été caractérisée par deux épisodes de forte pollution par l'ozone, notamment le 17 et le 18. D'ailleurs, ces deux pics ont été représentés assez bien par le modèle, mais on remarque quand même des différences importantes entre les concentrations simulées par CHIMERE et celles enregistrées aux points de mesure pour les dix jours mentionnés. Rappelons que ceci a été le point de départ de la thèse de N. Blond qui a implémenté sur CHIMERE un système d'assimilation de données, basé sur le krigeage intrinsèque, appliqué une fois directement sur les observations (notée OBK), et une fois, sur les innovations (INK), c'est-à-dire, sur les différences entre les observations et les concentrations simulées par le modèle aux points de mesure.

Toute implémentation d'un schéma d'assimilation par EnKF sur un modèle de chimie-transport (et pas uniquement ce type de modèle) passe par ce qu'on appelle une période de "**spin-up**", période pendant laquelle, une fois créé l'ensemble, on laisse les membres évoluer pour une courte période de temps, pour qu'ils puissent arriver à un équilibre dynamique. On a choisi une période de "spin-up" de 24 heures, toute la journée du 10 Juillet 1999, et on a commencé l'assimilation le 11 Juillet 1999 à 0 h TC[13], procédure appliquée avec un pas de temps horaire pour les dix jours suivants, jusqu'au 20 Juillet 1999 à 23h.

4.7.3 Les données disponibles

Les données utilisées dans la procédure d'assimilation ont été fournies par le réseau d'AIRPARIF. Il s'agit des données horaires de concentrations d'ozone, mesurées à la surface, par 17 stations situées dans Paris intra-muros et sa grande couronne, qu'on peut visualiser sur la carte (voir la figure 4.3). À ces 17 stations on a rajouté une autre : Chartres-Fulbert, déjà mentionnée dans le chapitre 2, comme une station faisant partie de l'association LIG'AIR.

4.7.3.1 Choix des stations. Répartition dans deux groupes.

Un problème très important est le nombre réduit de stations disponibles par rapport à la dimension du vecteur d'état. On a divisé les stations en deux groupes distincts : un premier groupe, qui a été utilisé dans le processus d'assimilation pour contraindre le modèle, et le deuxième, gardé comme témoin pour valider les résultats obtenus par assimilation. Vue la répartition spatiale de nos 18 stations, la séparation a été effectuée en tenant compte du fait que les stations rurales sont très représentatives pour les zones voisines, donc elles ont été presque toutes utilisées pour assimiler l'ozone, sauf Mantes qui, située à une distance relativement réduite par rapport à ses deux voisines, Prunay et Fremainville, et très corrélée avec leurs mesures, a été maintenue comme témoin pour la région rurale de Nord-Ouest. Pour la petite couronne, on a décidé d'utiliser une seule station dans

[13] temps civil

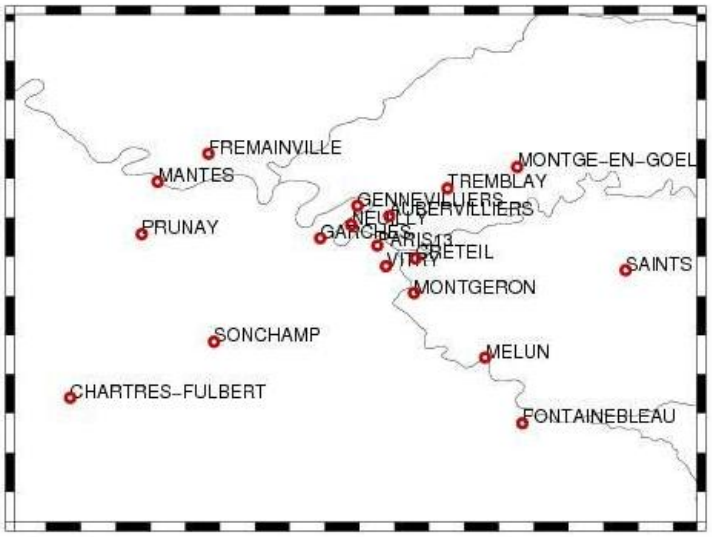


FIG. 4.3: Les stations de mesure appartenant à AIRPARIF utilisées dans l'assimilation séquentielle par EnKF.

l'assimilation (Montgeron) et on a gardé les 7 autres comme témoins.

4.7.4 L'ensemble initial

L'influence de l'initialisation dans une application d'assimilation de données peut être plus ou moins importante suivant la nature du modèle. On dit qu'on espère que le filtre corrige de lui-même l'imprécision de l'initialisation. Pour obtenir un ensemble initial, une possibilité sera, par exemple, de partir avec plusieurs états générés par le modèle et donner une approximation sur un sous-espace vectoriel en utilisant l'Analyse en Composantes Principales (pour une description de l'ACP voir Saporta (1990)). Cette méthode a été appliquée par exemple par Brasseur et al. (1999) en océanographie. La critique apportée à cette méthode provient du fait qu'on utilise pour initialiser le filtre uniquement le modèle et pas les observations. Cela peut introduire un déséquilibre en faveur du modèle, car les phénomènes absents du modèle mais présents dans les données ne seront donc pas représentés dans les covariances successives, ce qui limite les possibilités de correction (Bertino, 2001).

Dans cette étude, l'ensemble initial a été créé par l'addition des champs "pseudo-aléatoires" d'ozone (voir section 4.4.10.4) lisses et corrélés verticalement, à un champ de référence produit par le modèle. Ces champs ont été tirés de la même distribution, caractérisée par une moyenne nulle et une variance unitaire. Ils ont été multipliés par un certain pourcentage de la concentration simulée par le modèle et additionnés aux champs originaux produits par le même modèle. Les pourcentages testés varient entre 20 et 40 et les résultats obtenus, pour les différents cas, sont présentés dans la section 4.8.4.3. Chacun de ces champs est obtenu en utilisant la technique décrite antérieurement, la FFT (voir la section 4.4.10.4). La longueur de décorrélation, qui caractérise la fonction de covariance gaussienne testée, a été variée elle aussi sur une échelle assez grande, de 24 km jusqu'à 70 km ; l'influence de cette variation sur les résultats sera analysée dans la section 4.8.4.2.

4.7.5 Perturbations effectuées

4.7.5.1 Perturbations sur le modèle

Une fois créé l'ensemble et après avoir effectué une première analyse, compte tenu de la première série d'observations, pour chacun de ces membres, un **nouveau champ de concentrations d'ozone** est généré, par l'addition d'un champ pseudo-aléatoire au champ **analysé** produit à l'instant précédent. Les caractéristiques de ces champs pseudo-aléatoires ont été décrites antérieurement : il s'agit d'une fonction de covariance gaussienne, isotrope, d'une longueur de décorrélation qui varie entre 24 et 70 km. Les champs aléatoires sont indépendants pour chaque membre de l'ensemble. Ces nouveaux champs seront propagés directement par le modèle, analysés (ou corrigés) en utilisant les observations surfaciques d'ozone et ensuite re-perturbés pour les introduire, encore une fois, dans le modèle et ainsi de suite.

4.7.5.2 Perturbations sur les observations

Dans cette étude, on a considéré la matrice R, représentant la matrice de covariance de l'erreur sur les observations, comme diagonale (ce qui revient à supposer l'indépendance des observations). Pour chaque mesure d'ozone, il a été supposé que l'écart-type de l'erreur est égal à $10\,\mu\text{g.m}^{-3}$. Cette erreur représente d'un côté l'incertitude sur la mesure effectuée par la station et d'un autre côté l'erreur de représentativité liée au fait que l'opérateur **H** associe la concentration d'ozone mesurée dans un endroit à toute une cellule de la grille. Ceci est concordant avec la valeur de 5 ppb, obtenue dans une étude effectuée par Flemming et al. (2003), basée sur la méthode observationnelle de Hollingsworth et Lönnberg (1986), où la variance de l'erreur d'observation est obtenue par extrapolation d'une covariance sphérique modélisant la covariance issue des stations voisines.

En utilisant l'algorithme de type racine carrée, décrit dans la section 4.4.10.3, on évite de perturber les observations. Toutefois, pour comparaison, on a effectué aussi des perturbations sur les observations et on a appliqué un algorithme de correction, mais en remplaçant R par la matrice de covariance des perturbations R_e (voir la section 4.4.10.2) ; la conclusion qui s'impose est que, dans notre cas, il n'existe pas une différence évidente entre les résultats obtenus en employant les deux types d'algorithme.

4.7.6 L'intervalle de temps entre deux assimilations

Dans les simulations effectuées en utilisant le système d'assimilation décrit, des perturbations sont rajoutées toutes les heures aux champs analysés d'ozone, et l'étape de correction est appliquée à chaque heure. L'assimilation est donc effectuée très souvent, pour obtenir un maximum de gain. On a effectué aussi des tests en corrigeant le champ d'ozone toutes les trois heures, mais on a constaté que les statistiques sur les résultats se dégradent, au fur et à mesure que l'intervalle de temps entre deux corrections augmente.

4.7.7 L'opérateur de projection des observations

Pour finir avec la mise en œuvre du système d'assimilation on décrit brièvement l'opérateur *linéaire* **H**, une matrice de dimension $m \times N$, qui fait le lien entre les observations prises aux sites de mesure et les mailles du modèle. La projection de chaque site sur le maillage du domaine fournit la position d'une valeur unitaire dans la matrice d'observation ; le reste des valeurs de cette matrice sont nulles.

Pour résumer, on veut effectuer un filtrage sur un modèle de chimie-transport pour améliorer les estimations successives de quelques espèces chimiques (l'accent est mis sur l'ozone), et la seule espèce dont les mesures sont utilisées dans le processus de correction, avec un pas temporel d'une heure, est l'ozone. Toutes les autres espèces sont, elles-aussi *impactées*, car elles font partie du vecteur d'état. L'assimilation est effectuée en additionnant des champs "pseudo-aléatoires" aux champs analysés **d'ozone**, en les propageant par le modèle, et en les corrigeant ensuite, à l'aide des observations, par un algorithme de type *racine carrée* (4.4.10.3). Parmi les 18 stations de mesure disponibles, on a utilisé 10 dans le processus d'assimilation, toutes étant situées plutôt dans la zone rurale, exceptée une seule station urbaine.

4.8 Résultats et discussion

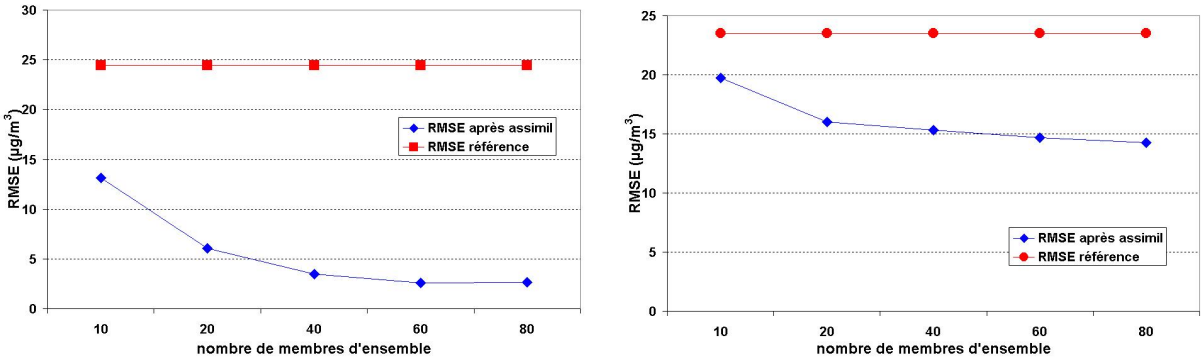
Rappelons d'abord l'objectif principal de l'application de l'EnKF sur le modèle CHIMERE : l'amélioration de la cartographie des deux polluants atmosphériques l'ozone et le dioxyde d'azote sur la région d'Île-de-France. C'est la raison pour laquelle on a choisi de perturber uniquement les champs d'ozone simulés par le modèle et pas les paramètres incertains du modèle.

Pour étudier l'impact de l'assimilation sur les champs de concentration d'ozone, on a effectué un grand nombre de simulations en faisant varier les paramètres clé de cet algorithme : le nombre de membres d'ensemble, la longueur de décorrélation qui caractérise la fonction de covariance gaussienne utilisée pour produire les champs pseudo-aléatoires, ainsi que la variance introduite dans l'ensemble par cette procédure.

Afin de contrôler la qualité des analyses obtenues, on a gardé des stations-témoin, qui n'ont pas été utilisées dans le processus d'assimilation, et sur lesquelles on vérifiera la qualité des analyses.

4.8.1 Statistiques moyennes sur les résultats

Les résultats d'une comparaison entre les analyses produites à l'aide de l'EnKF, les simulations brutes du modèle, et les observations de surface sont présentés en termes de RMSE et MAE calculés sur les différences analyses-observations et les différences simulations brutes-observations pour chaque groupe de stations : d'assimilation ou de validation, en fonction du nombre de membres d'ensemble, appelé **nrens**. On commence par la présentation des statistiques **moyennes** obtenues sur les deux groupes de stations mentionnés. Les statistiques ont été calculées pour chaque station sur toute la période analysée (la deuxième décade du mois de juillet 1999), ensuite une valeur moyenne a été déterminée pour chaque groupe de stations. Ces statistiques concernent premièrement la RMSE et la MAE moyennes (voir l'annexe B) sur chaque groupe de stations.



(a) RMSE moyenne calculée sur le groupe de stations d'assimilation pour la deuxième décade du mois de juillet 1999.

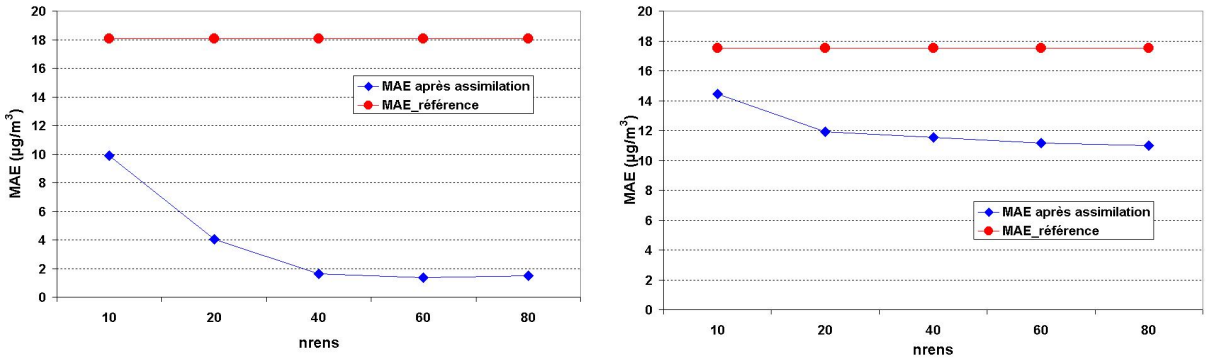
(b) RMSE moyenne calculée sur le groupe de stations de validation pour la deuxième décade du mois de juillet 1999.

FIG. 4.4: RMSE moyenne calculée sur les deux types de stations pour divers nombres de membres d'ensemble pour (a) les stations d'assimilation et (b) les stations de validation.

Dans la figure 4.4 on peut visualiser la variation de la RMSE moyenne avec la taille de l'ensemble. Cinq tailles différentes de l'ensemble ont été testées : 10, 20, 40, 60 et 80 membres. On peut remarquer que le passage de 10 membres à 20 est crucial : on obtient une diminution moyenne de la RMSE d'environ $7-8\,\mu g/m^3$ pour les stations d'assimilation et de $4-5\,\mu g/m^3$ sur les stations de validation ; ensuite, on peut noter que l'augmentation de la taille de l'ensemble n'est pas essentielle, car la réduction de la RMSE obtenue est très faible pour les deux types de stations. Il apparaît ainsi qu'avec 20 ou 40 membres d'ensemble on arrive à bien représenter l'espace des erreurs ; reste à comparer les performances obtenues sur les champs 2D de concentrations de polluant. On remarque également la valeur de la RMSE, notée référence, qui a été calculée comme moyenne, pour les stations de chaque groupe, sur les différences entre les simulations brutes du modèle et les observations et qui, dans les deux cas, enregistre une valeur d'environ $24\,\mu g/m^3$.

La description des résultats est complétée par les statistiques moyennes sur les MAE obtenues sur les mêmes résultats (voir la figure 4.5). On retrouve la même situation qu'avant (pour la RMSE), c'est-à-dire une diminution importante de la MAE quand on passe de 10 membres d'ensemble à 20 et ensuite une variation très lente avec le nombre d'ensemble jusqu'à 80 membres.

Sur les deux figures présentées précédemment, on constate la différence entre les statistiques calculées sur les deux types de stations. Cette différence est normale. La réduction des erreurs sur les stations utilisées dans l'assimilation est, bien évidemment, plus importante que celle obtenue sur les stations de validation. Un point significatif sur lequel il faut insister est la répartition spatiale des stations de validation, concentrées plutôt dans la zone urbaine, exceptée Mantes, qui est située en zone rurale. Dans la suite, on détaillera un peu plus cette analyse, en considérant chaque station séparément sur les deux groupes définis antérieurement : d'assimilation et de validation.



(a) MAE moyenne calculée sur les stations d'assimilation pour la deuxième décade du mois de juillet 1999.

(b) MAE moyenne calculée sur les stations de validation pour la deuxième décade du mois de juillet 1999.

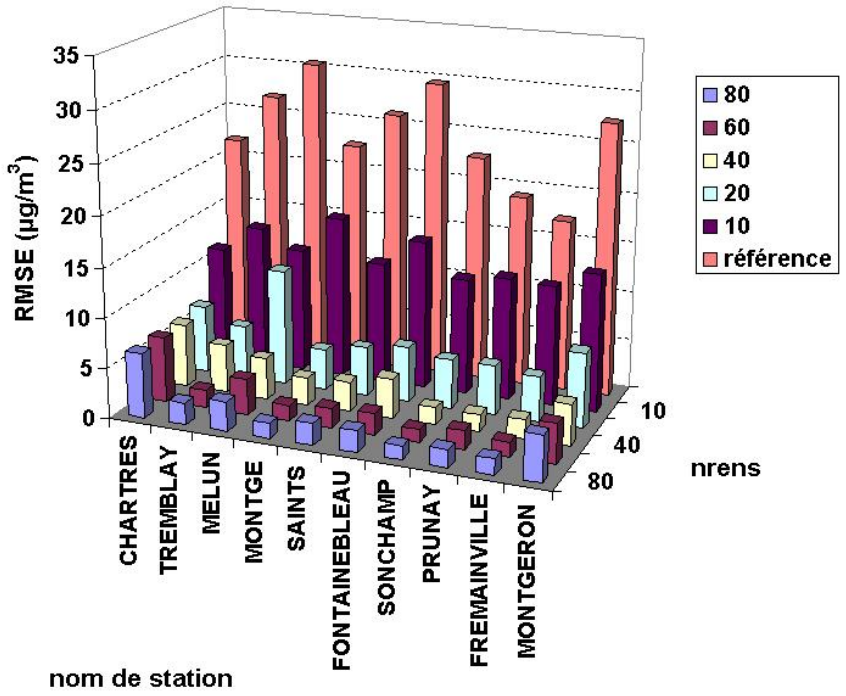
FIG. 4.5: MAE moyenne calculée sur les deux types de stations pour divers nombres de membres d'ensemble (nrens), pour (a) les stations d'assimilation et (b) les stations de validation.

4.8.2 Statistiques : RMSE et MAE sur les séries temporelles des stations

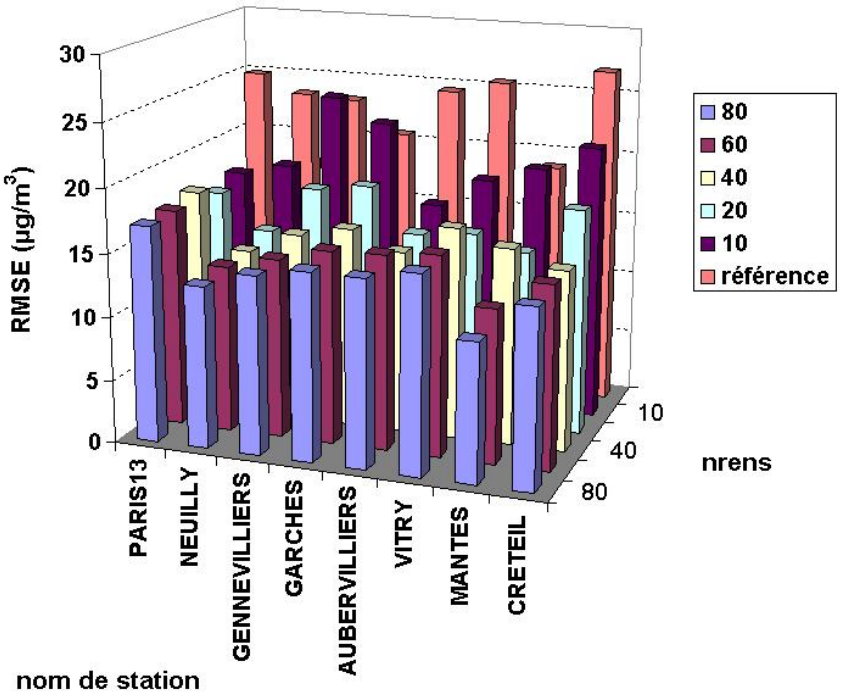
En regardant les figures (4.6(a) et 4.6(b)) qui présentent les deux mêmes statistiques défalquées par stations, il apparaît clairement que la RMSE des différences analyses-observations est plus faible que celle entre les simulations et les observations (notée sur le graphique comme "référence"). On garde la même notation qu'avant pour la taille de chaque ensemble : **nrens**. Pour les stations d'assimilation, le gain est notable, en revanche, pour les stations de validation on peut remarquer que la simulation qui utilise 10 membres d'ensemble conduit à des mauvais résultats. On observe clairement dans la figure 4.6(b) qu'il existe au moins trois stations : Gennevilliers, Garches et Mantes, où on n'obtient pas d'amélioration en termes de RMSE, et ceci est dû à une sous-représentation de l'espace des erreurs quand on utilise 10 membres d'ensemble. Pour chacune de trois stations on obtient soit des surestimations (comme c'est le cas pour Garches et Gennevilliers le 13 et respectivement le 16 Juillet), soit des sous-estimations (comme à Mantes pour le 17 Juillet à 13 heures). Un biais important qui peut être introduit dans les champs analysés est dû aux contraintes imposées aux concentrations : elles doivent être positives après analyse. Si ce n'est pas le cas on exige qu'elles soient nulles et ceci introduit généralement un biais très important. Cette situation défavorable ne se retrouve plus sur les quatre simulations qui suivent, pour lesquelles la réduction de la RMSE est nette sur toutes les stations de validation.

On analyse également la figure 4.7, où on a représenté les statistiques sur les résultats en termes de MAE. On retrouve la même situation que précédemment, c'est-à-dire que, pour les stations d'assimilation l'amélioration est évidente, tandis que pour celles de validation on obtient encore la même mauvaise simulation qui utilise 10 membres d'ensemble, pour laquelle deux stations, notamment Gennevilliers et Garches gâchent les statistiques.

En revenant aux figures 4.7(a) et 4.7(b), on peut remarquer qu'il existe des stations sur lesquelles la réduction des erreurs est plus manifeste que pour d'autres. Pour les cinq tailles différentes d'ensemble, on a dans le groupe des stations d'assimilation des exemples comme Sonchamp, Fontainebleau, Fremainville, Tremblay ou Montge-en-Goele qui sont toujours en tête de liste pour



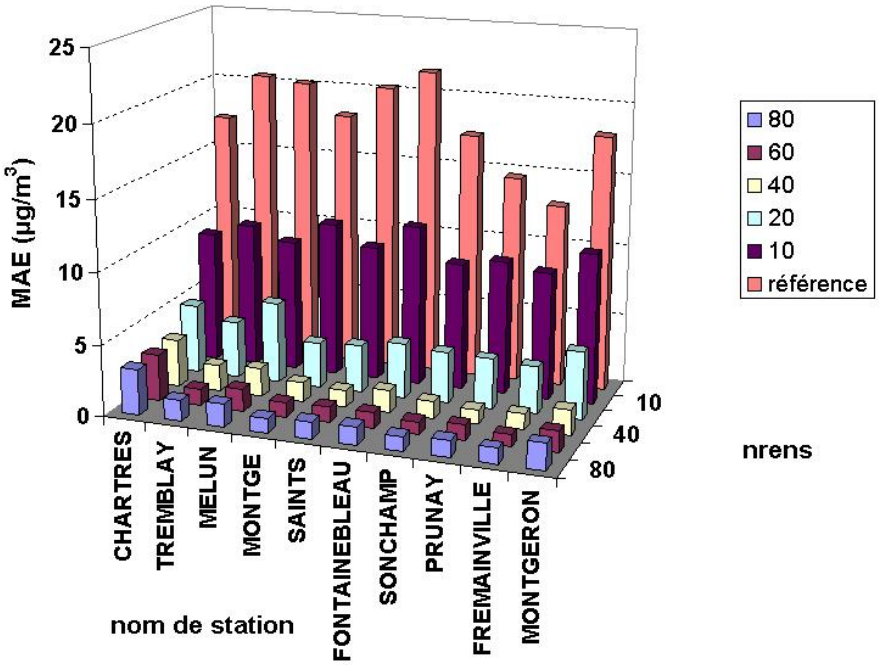
(a) RMSE calculée pour chaque station d'assimilation comme fonction du nombre de membres d'ensemble.



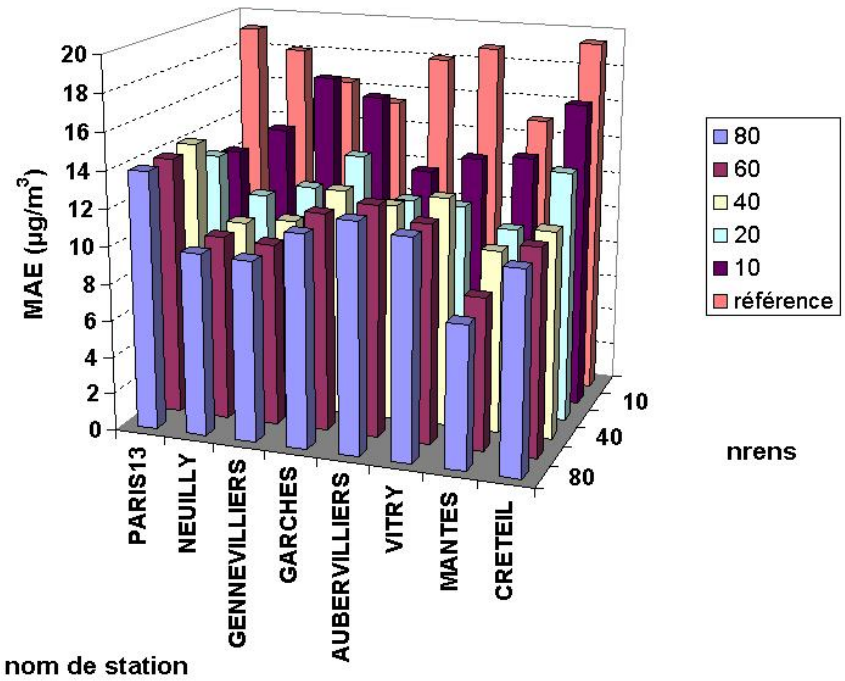
(b) RMSE calculée pour chaque station de validation comme fonction du nombre de membres d'ensemble.

FIG. 4.6: RMSE calculée par station pour divers nombres de membres d'ensemble pour (a) les stations d'assimilation et (b) les stations de validation.

prouver l'efficacité de la technique d'assimilation. De l'autre côté, dans le groupe des stations de validation, on retrouve Créteil, Aubervilliers, Vitry ou Neuilly, stations où la performance du sys-



(a) MAE calculée pour chaque station d'assimilation comme fonction du nombre de membres d'ensemble.



(b) MAE calculée pour chaque station de validation comme fonction du nombre de membres d'ensemble.

FIG. 4.7: MAE calculée par station pour divers nombres de membres d'ensemble pour (a) les stations d'assimilation et (b) les stations de validation.

tème est toujours meilleure par rapport aux autres (voir l'emplacement des stations sur la carte 4.3).

4.8.3 Séries temporelles : exemples et analyses

• Stations d'assimilation

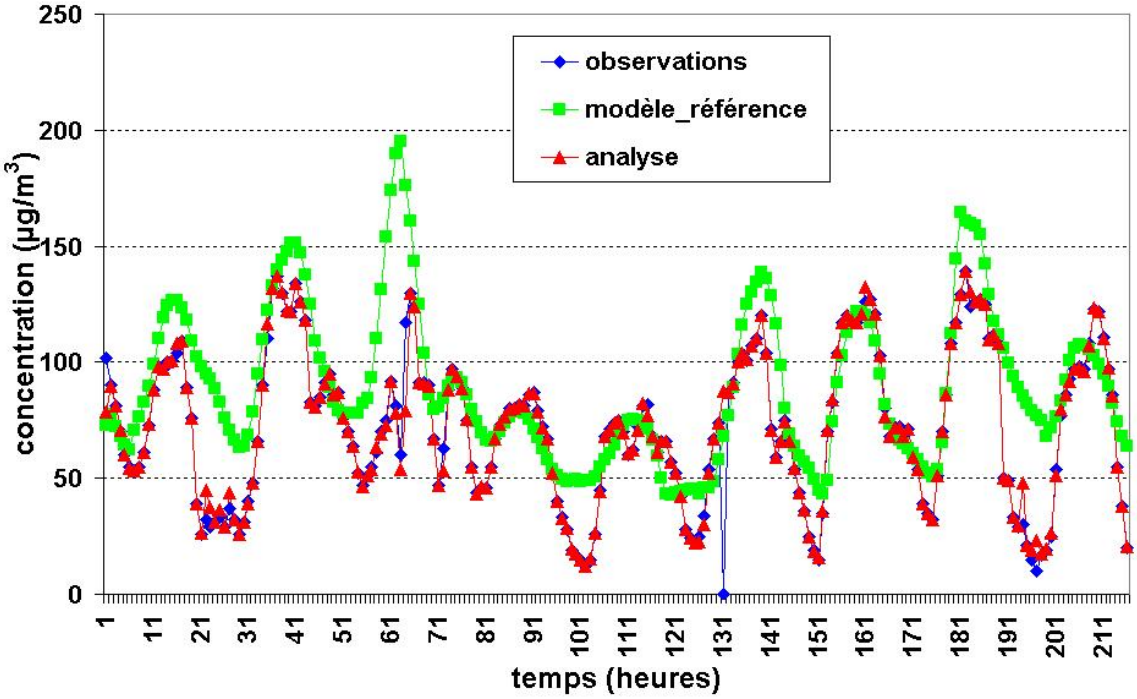
Prenons comme exemples deux figures représentant les profils temporels des concentrations d'ozone observées, simulées par le modèle CHIMERE et analysées par EnKF sur deux stations d'assimilation : Fontainebleau et Sonchamp (voir les figures 4.8(a) et 4.8(b)), en utilisant 40 membres d'ensemble. Nous remarquons que, globalement, le biais assez important simulé par le modèle aux sites de mesure est corrigé chaque fois par l'EnKF. La simple visualisation de ces séries temporelles nous montre que les concentrations analysées sur les stations d'assimilation reproduisent bien les concentrations mesurées par les stations. Il n'y a rien d'étonnant parce que, en utilisant la technique d'assimilation de données, on veut contraindre le modèle à s'approcher des observations là où l'écart entre les deux est important. Comme l'analyse représente une combinaison entre deux sources d'information : la prédiction (ou l'ébauche) et les observations, la composante ayant moins d'incertitude recevra plus de poids.

Par ailleurs, on remarque le contraste entre les deux stations rurales choisies comme exemples : Fontainebleau, située en zone rurale Sud-Est, où les concentrations d'ozone ne sont pas très élevées par rapport à Sonchamp, située à l'Ouest, où on retrouve les panaches d'ozone formés pendant cette deuxième décade de Juillet 1999.

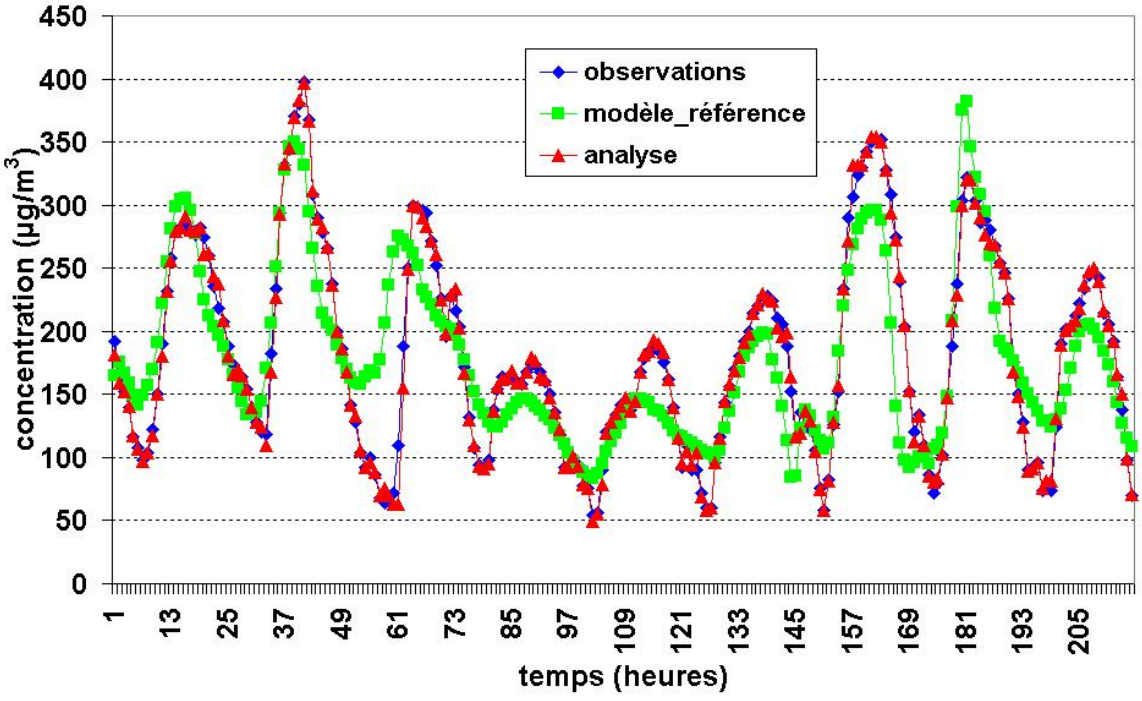
Pour une meilleure visualisation des améliorations obtenues avec l'EnKF, on a décidé de représenter les erreurs commises sur la période d'assimilation (11-19 Juillet 1999) en calculant les différences entre les observations et le modèle ou entre les observations et l'analyse pour les deux stations mentionnées : Fontainebleau et Sonchamp. Comme on peut voir dans la figure 4.9, l'amplitude des erreurs décroît progressivement avec l'augmentation de la taille de l'ensemble. Le gros écart du modèle, pour les deux séries analysées, est caractéristique pour la date de 13 Juillet, situation qui sera décrite plus en détail à la fin de cette section. Une autre remarque qu'on peut faire ici vise le fait que les séries temporelles obtenues sur les stations d'assimilation montrent un rapprochement très fort par rapport aux observations. Ceci est dû au fait qu'on a introduit une très forte perturbation dans le champ simulé par le modèle (voir la section 4.7.5) et on a considéré que l'incertitude dans les observations est très faible : régime d'assimilation particulier. Entre les deux informations, le filtre fera plus de confiance à la partie caractérisée par une incertitude faible, donc il va s'approcher d'observations, ce qu'on souhaite, car l'intérêt réside principalement dans la comparaison avec l'interpolation spatiale.

• Stations de validation

On garde les mêmes conventions que pour les stations d'assimilation quand on représente les profils temporels obtenus aux stations de validation. Remarquons d'abord la faible représentativité spatiale de ces stations, due au fait que les stations urbaines sont nombreuses et concentrées à l'intérieur de l'agglomération parisienne. L'utilisation d'une seule station urbaine dans le processus d'assimilation est suffisante pour reconstruire le champs de concentrations d'ozone dans la zone cible. On a effectué plusieurs simulations en remplaçant chaque fois la station urbaine utilisée par

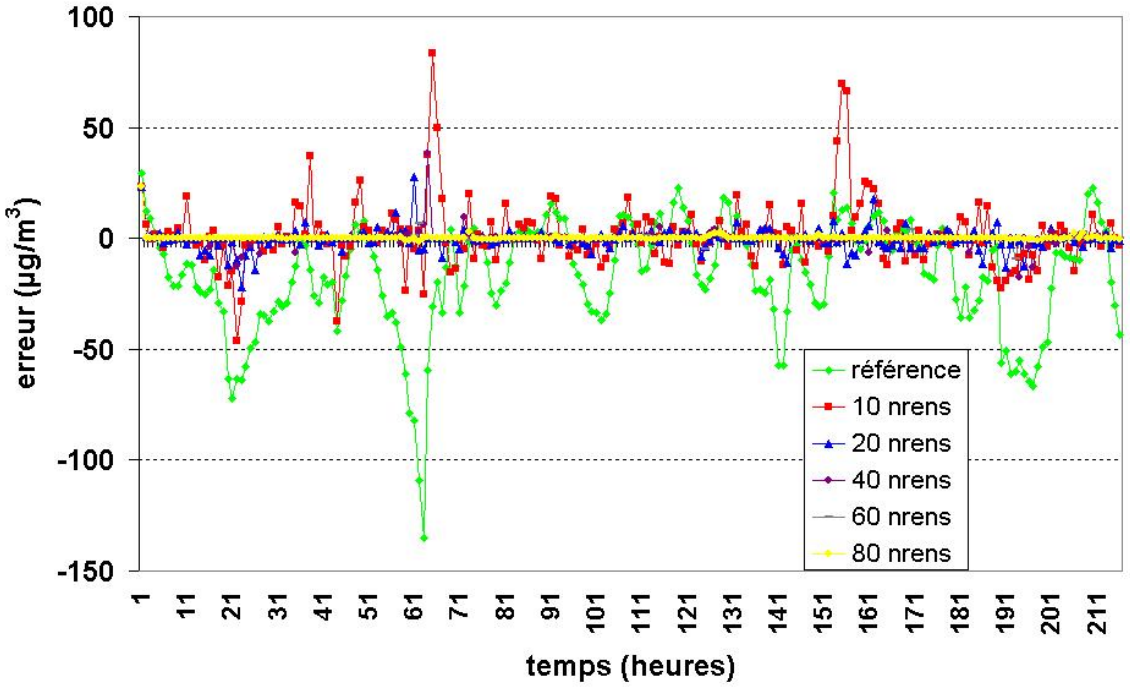


(a) Séries temporelles des concentrations d'ozone : observée (en bleu), simulée par CHIMERE (en vert) et analysée (en rouge), à Fontainebleau.

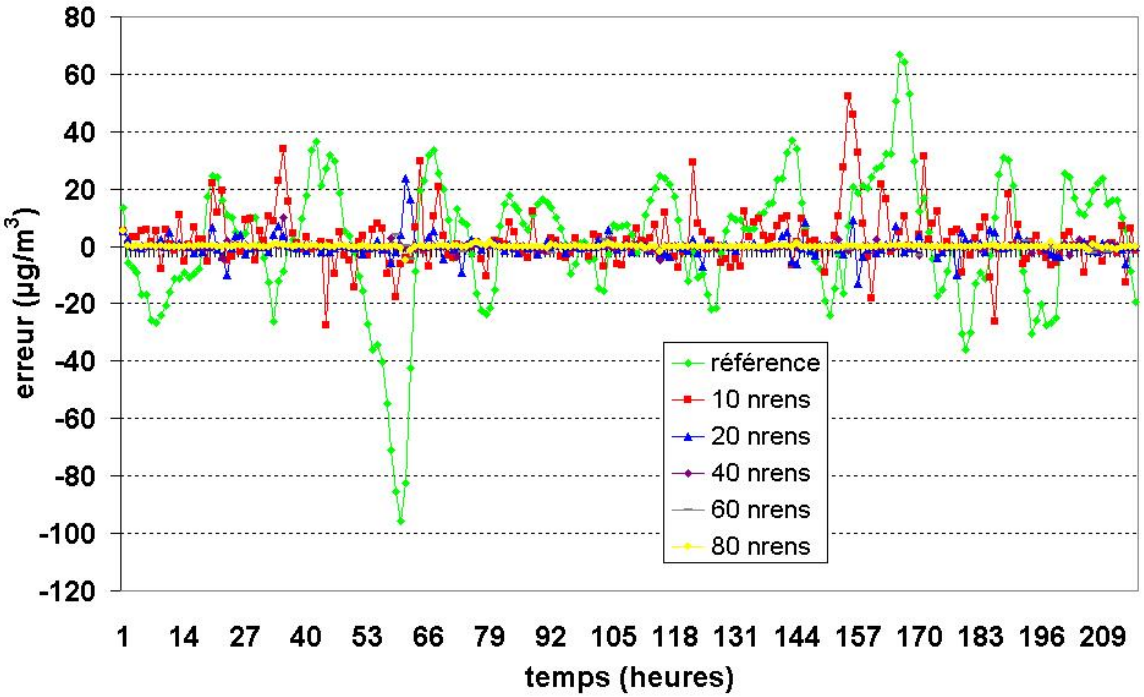


(b) Séries temporelles des concentrations d'ozone : observée (en bleu), simulée par CHIMERE (en vert) et analysée (en rouge), à Sonchamp.

FIG. 4.8: Exemples de profils temporels obtenus sur deux stations d'assimilation : Fontainebleau et Sonchamp.



(a) Séries temporelles des erreurs sur les concentrations d'ozone : simulée par CHIMERE (en vert) et analysée en utilisant l'EnKF avec des différentes tailles de l'ensemble, à Fontainebleau.



(b) Séries temporelles des erreurs sur les concentrations d'ozone : simulée par CHIMERE (en vert) et analysée en utilisant l'EnKF avec des différentes tailles de l'ensemble, à Sonchamp.

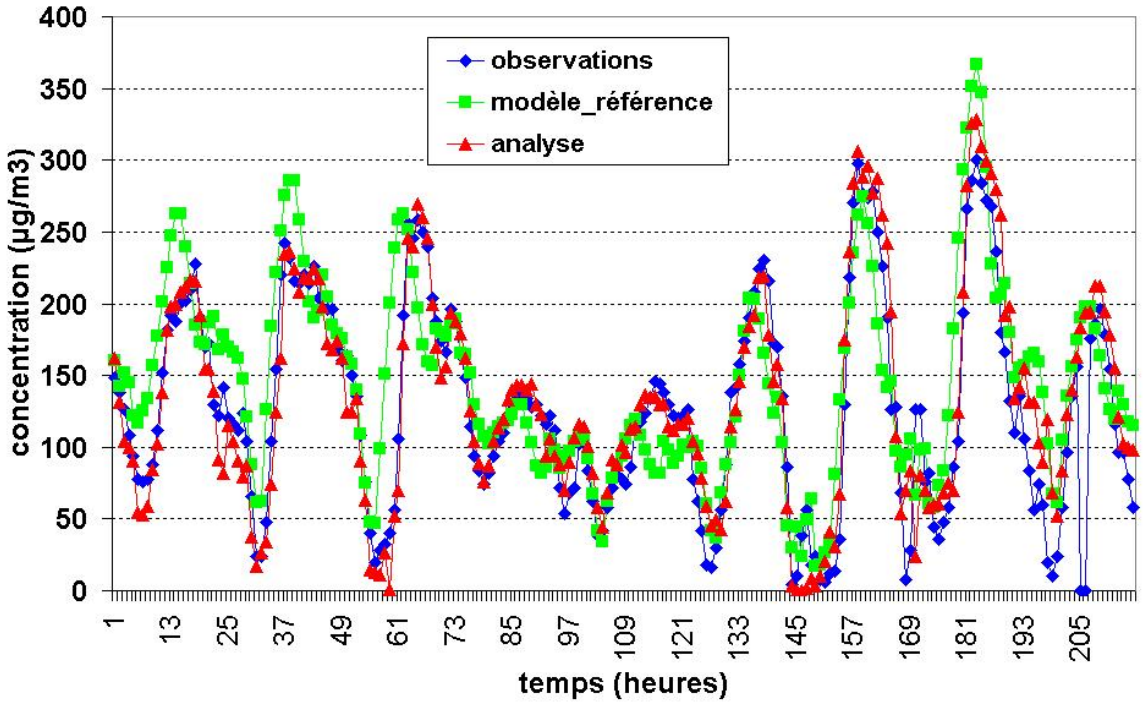
FIG. 4.9: Exemples de profils temporels obtenus pour les erreurs commises sur deux stations d'assimilation : Fontainebleau et Sonchamp.

une autre de la même nature, et les résultats obtenus en termes de RMSE et MAE suivent une variation assez faible, démontrant que l'analyse effectuée à partir des autres observations proches reconstruit l'information manquante. En revanche, pour les stations rurales la situation est un peu plus compliquée. Comme leur répartition spatiale est creuse, si nous retirons une station rurale qui est plus éloignée des autres points de mesure, l'analyse se rapproche du modèle, qui devient la seule information disponible. C'est la raison pour laquelle, finalement, on a décidé de garder comme seule station-témoin Mantes, qui est située dans le voisinage proche de deux autres : Prunay et Fraimainville, donc on peut se dispenser de l'information donnée par cette station.

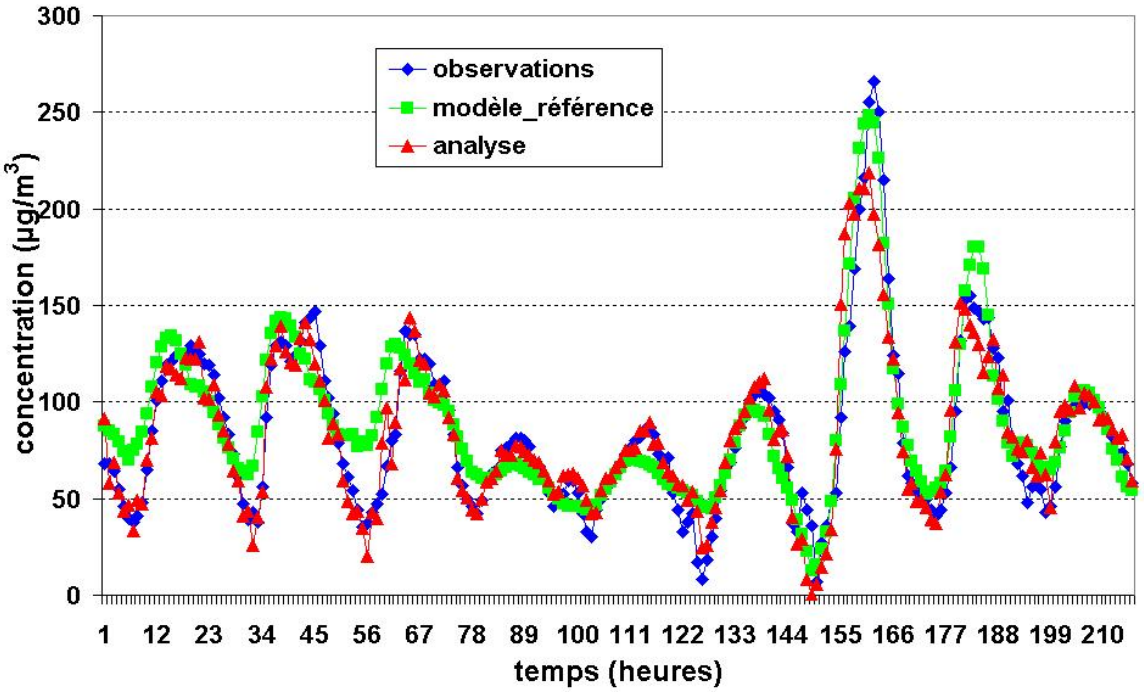
On a choisi de représenter deux stations de validation, notamment Neuilly et Mantes (figure 4.10). Tout d'abord, on remarque que les séries temporelles n'ont pas la même allure que précédemment (pour les stations utilisées dans l'assimilation). Comme on pouvait s'y attendre, l'amélioration introduite par l'analyse est moindre par rapport aux mesures enregistrées. En ce qui concerne Neuilly (figure 4.10(a)), on remarque d'abord une sous-estimation par l'assimilation le 12 Juillet le matin (le modèle sur-estime) et ensuite une bonne correction sur le pic de l'après-midi ; cette correction continue le 13 Juillet, quand on arrive à réduire le grand écart estimé par le modèle, et également le 15 Juillet. En revanche, le pic du 18 Juillet est moins bien corrigé : en fait, l'écart par rapport aux observations est réduit à la moitié. Pour la deuxième station, Mantes (figure 4.10(b)), on remarque d'abord que l'écart du modèle est, d'une manière générale, réduit par rapport à l'autre station analysée (les concentrations sont, elles aussi, moins élevées), et cet écart caractérise plutôt les heures très matinales. Par contre, pour le pic enregistré le 17 Juillet, on constate que c'est bien le modèle qui simule correctement la concentration d'ozone et pas l'analyse.

En regardant les séries temporelles des erreurs commises sur ces deux stations (la figure 4.11), on constate que la décroissance de l'amplitude avec la taille de l'ensemble n'a plus du tout la même allure. Dans le cas des stations d'assimilation cette décroissance était nette pour 80 membres d'ensemble. Pour les stations de validation la courbe qui correspond à 80 membres d'ensemble, montre une amplitude d'erreur équivalente à celle obtenue pour 60 et même 40 membres d'ensemble. Cela renforce la conclusion partielle déjà exprimée, selon laquelle 20 ou 40 membres d'ensemble suffisent largement pour représenter l'espace des erreurs.

Ce qu'on peut remarquer sur les quatre séries temporelles présentées (les figures 4.8 et 4.10) est que, en moyenne, le cycle diurne d'ozone est bien reproduit par le modèle CHIMERE, avec un petit bémol pour les heures matinales quand, d'habitude, le modèle surestime les concentrations réelles d'ozone. Une deuxième observation qu'on peut faire est que le modèle simule mal les concentrations du 13 Juillet 1999 (le troisième pic de chaque série temporelle) ; on remarque un écart très important sur tous les profils temporels présentées, dû probablement aux conditions aux limites. Cette surestimation doit être corrigée par l'EnKF. Comme les contraintes introduites dans le champ d'ozone sont plutôt réduites comme nombre (on rappelle que l'assimilation est faite sur 10 stations), et comme les conditions aux limites sont les mêmes qu'avant, ainsi que les émissions, (on a choisi de perturber uniquement les champs de concentrations d'ozone), il faut s'attendre quand même à avoir des explosions numériques surtout aux bords du domaine et sur les zones qui ne sont pas couvertes par des stations. On reviendra sur ces explosions dans la section suivante, quand on présentera les

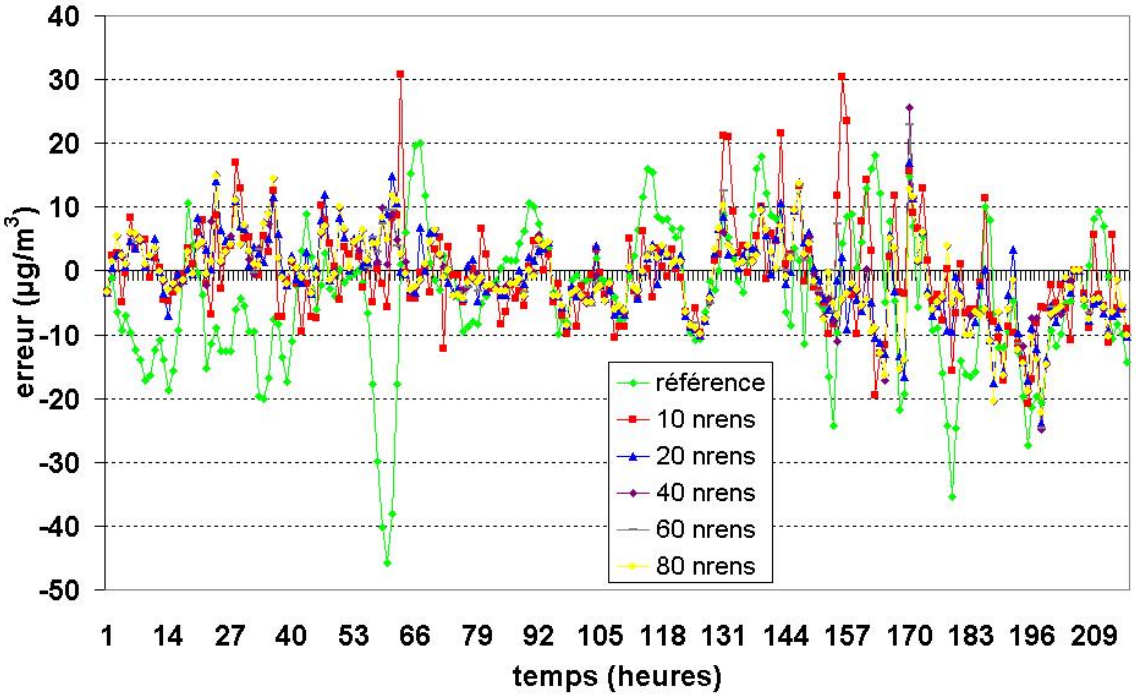


(a) Séries temporelles des concentrations d'ozone : observée (en bleu), simulée par CHIMERE (en vert) et analysée (en rouge), à Neuilly.

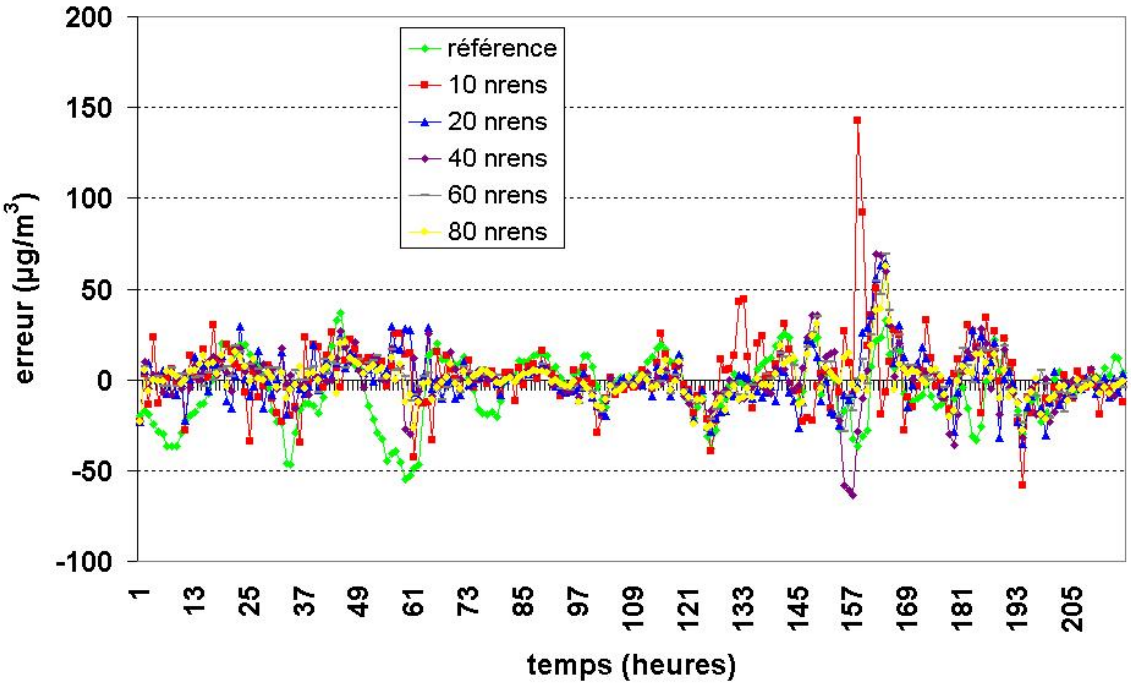


(b) Séries temporelles des concentrations d'ozone : observée (en bleu), simulée par CHIMERE (en vert) et analysée (en rouge), à Mantes.

FIG. 4.10: Exemples de profils temporels obtenus sur deux stations de validation : Neuilly et Mantes.



(a) Séries temporelles des erreurs sur les concentrations d'ozone : simulée par CHIMERE (en vert) et analysée en utilisant l'EnKF avec des différentes tailles de l'ensemble, à Neuilly.



(b) Séries temporelles des erreurs sur les concentrations d'ozone : simulée par CHIMERE (en vert) et analysée en utilisant l'EnKF avec des différentes tailles de l'ensemble, à Mantes.

FIG. 4.11: Exemples de profils temporels obtenus pour les erreurs commises sur deux stations d'assimilation : Neuilly et Mantes.

cartes 2D obtenues.

4.8.4 Sensibilité du système d'assimilation aux paramètres

Une fois le système d'assimilation construit, on teste sa sensibilité aux paramètres en effectuant des simulations dans lesquelles on fait varier séparément ses paramètres : la taille de l'ensemble, la longueur de décorrélation qui caractérise la fonction de covariance gaussienne du champ pseudo-aléatoire utilisé comme perturbation du champ simulé par CHIMERE, ainsi que la variance introduite par ces perturbations, et qui était proportionnelle à la concentration simulée par le modèle.

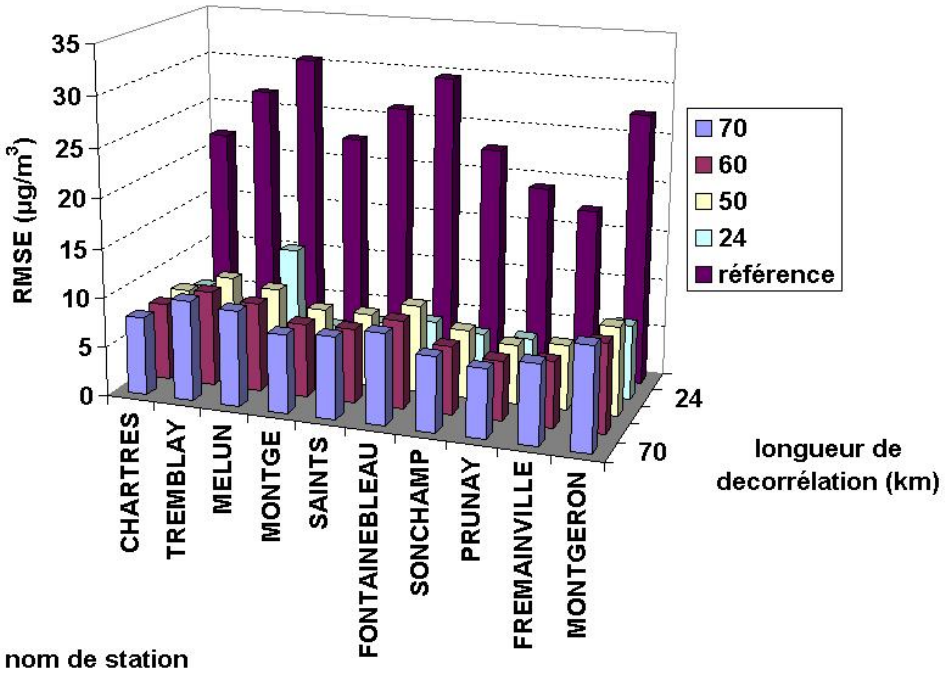
4.8.4.1 Le nombre de membres d'ensemble

La sensibilité du système d'assimilation par rapport à la taille de l'ensemble a été déjà présentée tout au long de cette discussion. On a appliqué l'EnKF avec cinq tailles différentes : 10, 20, 40, 60 et 80 membres d'ensemble. Pour résumer, les résultats présentés antérieurement montrent que, certes, pour les stations d'assimilation, il existe une nette diminution en termes de RMSE ou MAE avec la taille de l'ensemble et que cette amélioration diminue fortement pour les stations de validation (un bémol pour la représentativité spatiale de ces dernières - situées plutôt dans la région urbaine). La conclusion apparente est que, avec 20, maximum 40 membres d'ensemble, on arrive à bien représenter l'espace des erreurs caractéristique au système. Si on compare le taux de décroissance de la RMSE ou de la MAE avec le taux théorique $O(1/\sqrt{nrens})$, on voit que sur un nombre inférieur de membres d'ensemble la décroissance est plus rapide pour la RMSE, mais à la fin, pour 60 et 80 membres d'ensemble, la pente devient plus faible que celle théorique, presque plate.

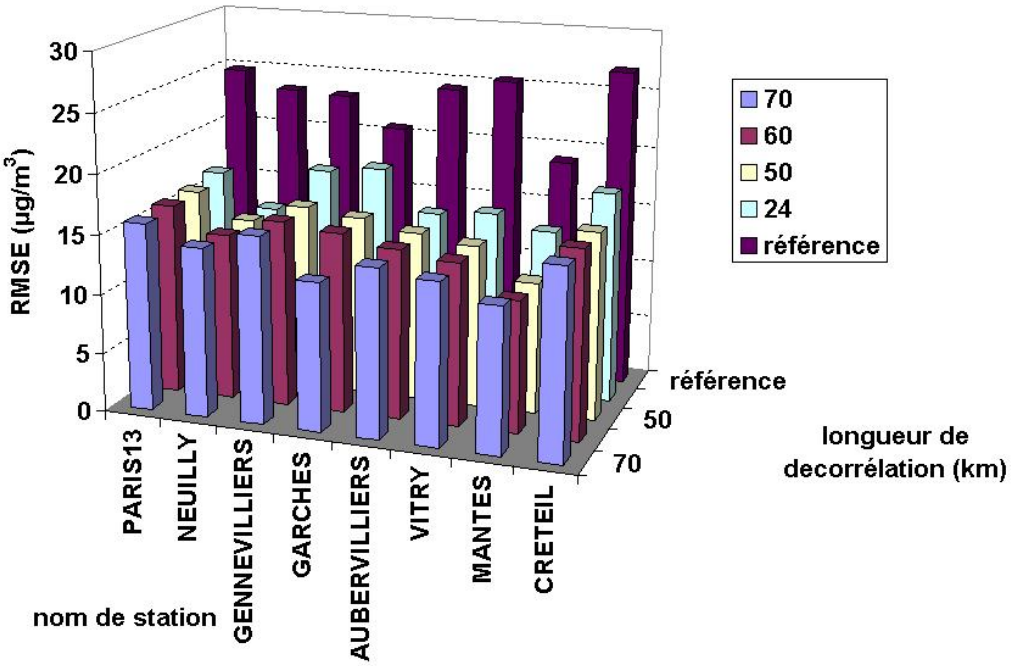
4.8.4.2 La longueur de décorrélation

L'évaluation de la performance du système par rapport à la longueur de décorrélation r_x et r_y (pour les deux directions de l'espace) a été effectuée en comparant 4 dimensions différentes, en commençant avec 24 km (4 cellules de la grille), en continuant avec 50 km, ensuite 60 km, pour finir avec 70 km, mais en gardant les deux autres paramètres constants : la variance introduite dans le système reste proportionnelle à 20% de la concentration mesurée et la taille de l'ensemble de 20 membres. On a considéré, dans toutes les simulations, les deux distances, r_x et r_y, égales. La variation de la RMSE défalquée sur chaque station montre qu'il n'y a pas une grande variabilité. Toutefois, on peut remarquer que pour les stations d'assimilation, il apparaît que, en moyenne, la simulation avec une décorrélation de 24 km est la plus efficace (sauf pour la station Melun), tandis que, pour les stations de validation on obtient une très faible amélioration pour la simulation qui utilise une décorrélation de 50 km.

Pourquoi le système d'assimilation implémenté n'est pas très sensible à la variation de ce paramètre ? On a effectué une autre simulation en diminuant cette décorrélation jusqu'à 10 km. Les champs 2D de concentrations de polluants ont complètement perdu leur continuité affichant des



(a) RMSE calculée pour chaque station d'assimilation comme fonction de la longueur de décorrélation.



(b) RMSE calculée pour chaque station de validation comme fonction de la longueur de décorrélation.

FIG. 4.12: RMSE calculée par station pour diverses longueurs de décorrélation, (a) les stations d'assimilation et (b) les stations de validation.

structures bizarres, très fragmentées. En revanche, si on effectue une simulation avec une décorrélation de 120 km, les résultats statistiques ne s'améliorent pas, même si les champs 2D obtenus deviennent très lisses.

4.8.4.3 La variance introduite dans le modèle par les perturbations pseudo-aléatoires

Le dernier paramètre testé est la variance introduite dans le modèle par les champs pseudo-aléatoires additionnés aux champs simulés par CHIMERE. Elle reste chaque fois proportionnelle aux concentrations simulées par le modèle. Rappelons que les champs *pseudo-aléatoires* de moyenne nulle et variance unitaire ont été multipliés par un pourcentage de la concentration analysée précédemment et ensuite additionnés au même champ de concentrations analysé obtenu à l'étape précédente ; ce pourcentage varie d'abord de 10%, en passant par 20%, 25%, 30% pour finir avec 50%. Toutes les simulations effectuées présentent une longueur de décorrélation de 50 km dans les deux directions de l'espace et une taille de 20 membres d'ensemble. L'analyse de la figure 4.13 montre que, pour les deux groupes de stations, la simulation la plus efficace est celle avec 20% de variance dans l'ensemble. Néanmoins, on remarque la faible variabilité du système par rapport à ce paramètre, signe que peut-être le seuil de 10% pour la variance introduite dans l'ensemble était déjà trop grande.

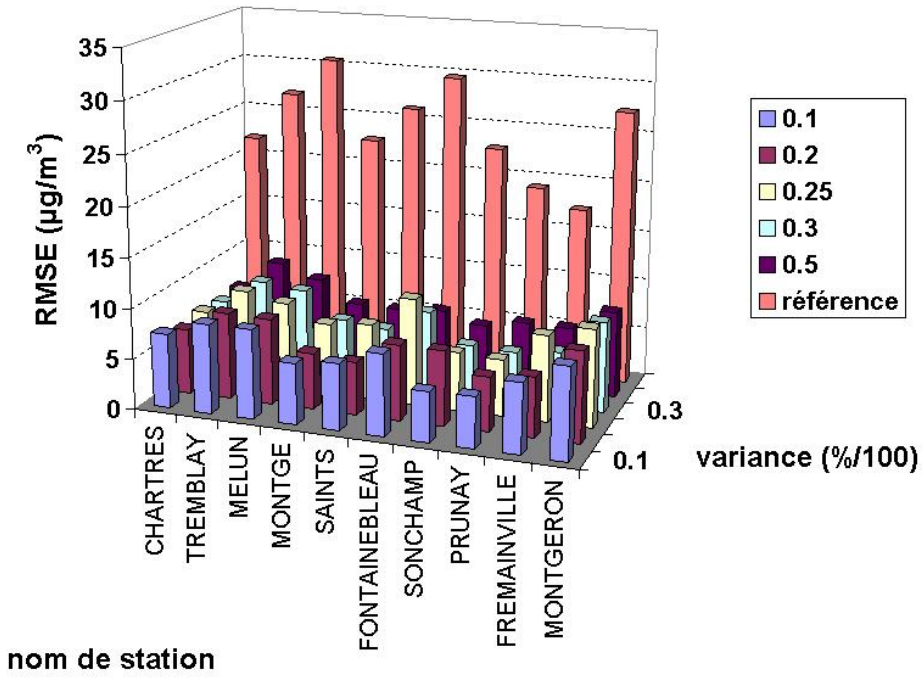
On constate, en effectuant des tests de sensibilité aux paramètres, que le système conçu varie lentement par rapport à chacun de trois paramètres testés. Cette variation a été discutée en termes d'amélioration de la RMSE pour les deux groupes de stations. Reste à savoir si en comparant les champs 2D, obtenus avec ce système, les différences sont plus ou moins flagrantes.

4.8.5 Champs 2D d'ozone simulés en utilisant l'EnKF

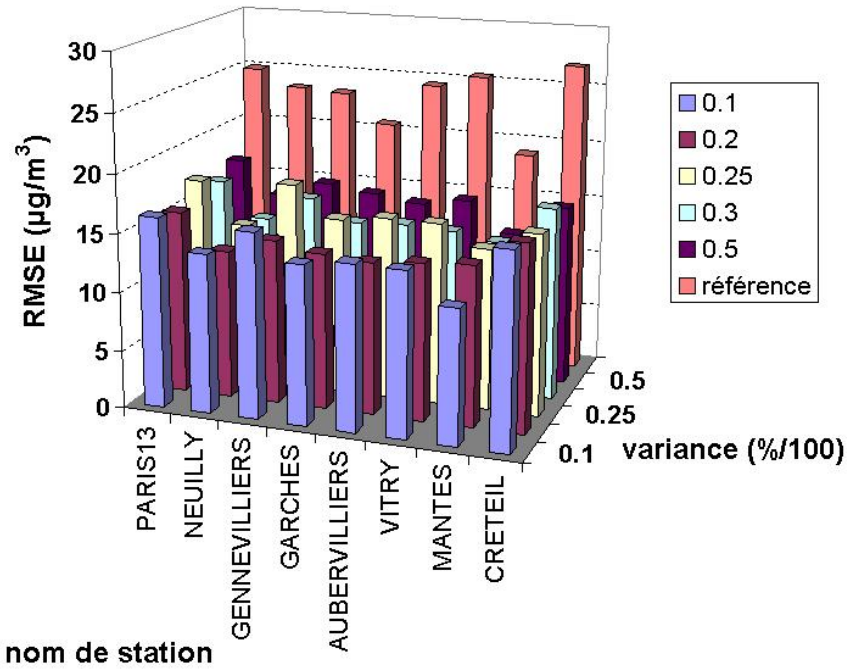
D'une manière générale, on peut constater que l'EnKF corrige les champs produits par le modèle, en modifiant leur gradient, de façon à prendre en compte les mesures disponibles. Contraindre le champ en utilisant uniquement 10 mesures d'ozone et quelques membres d'ensemble conduit parfois à des explosions numériques visibles surtout aux bords et sur les zones dépourvues de stations. Cette situation défavorable est due, probablement, aux conditions aux limites qui n'ont pas été perturbées. Dans l'approche classique de l'EnKF, les paramètres incertains du modèle devraient être perturbés. Parmi ceux-ci, les conditions aux limites et les émissions de précurseurs sont en tête de liste. En voulant garder une approche simpliste du phénomène (on a perturbé uniquement les concentrations d'ozone), on peut perdre la continuité des champs simulés. Toutes les analyses qui seront présentées par la suite, ont été obtenues en utilisant les paramètres suivants : 80 membres d'ensemble, la longueur de décorrélation de 24 km et une perturbation de 20% de la concentration de champ simulé.

- **Le champ d'ozone obtenu pour le 12 Juillet 1999 à 15 heures**

Un premier exemple de champ corrigé par l'EnKF concerne la date du 12 Juillet 1999 à 15 heures (figure 4.14). On remarque que le modèle simule un panache allongé au sud-ouest du domaine, autour de la station Sonchamp ; la forme de ce panache a été changée après l'assimilation. Le champ analysé présente une augmentation de la concentration d'ozone de 18 $\mu g/m^3$ (9 ppb) sur la station mentionnée ; par ailleurs, toute la zone rurale à l'est du domaine, surestimée par CHIMERE d'une moyenne d'environ 30 $\mu g/m^3$, soit 15 ppb (il s'agit des stations comme : Saints, Fontainebleau,



(a) RMSE calculée pour chaque station d'assimilation comme fonction de la variance introduite avec le champ pseudo-aléatoire additionné au champ initial produit par le modèle.

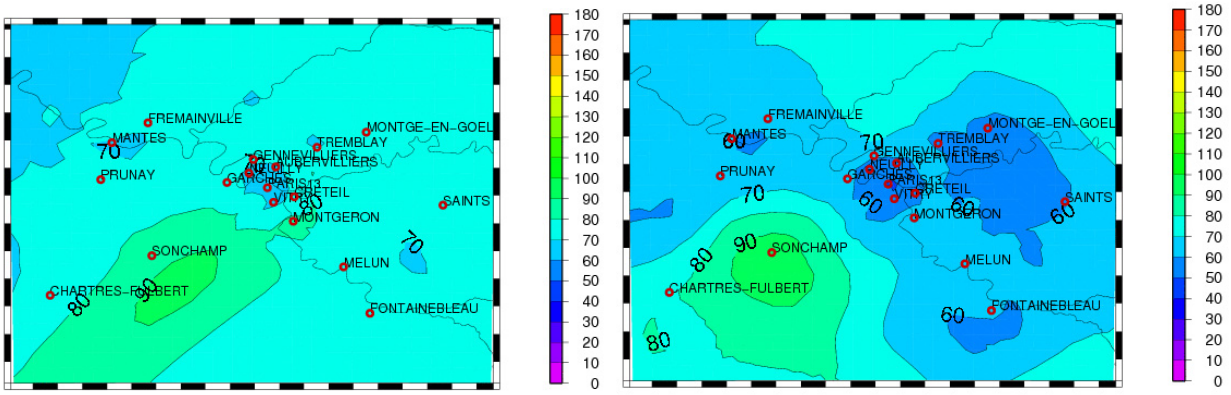


(b) RMSE calculée pour chaque station de validation comme fonction de la variance introduite avec le champ pseudo-aléatoire additionné au champ initial produit par le modèle.

FIG. 4.13: RMSE calculée par station pour diverses variances introduites avec le champ pseudo-aléatoire, (a) les stations d'assimilation et (b) les stations de validation.

Montge-en-Goele ou Melun) a été corrigée par l'EnKF. La zone rurale ouest est moins concernée, car la surestimation du modèle est moindre (environ $15-17\,\mu g/m^3$, soit 8 ppb), donc la correction éga-

lement. En revanche, pour les stations urbaines, celles qui ont été gardées comme stations-témoin, l'analyse sous-estime d'une moyenne de $20 - 25\,\mu g/m^3$ (10-12 ppb) les concentrations enregistrées. Par conséquent, pour cette date on a plutôt bien *corrigé* la zone rurale, mais l'analyse est moins précise sur la zone urbaine, où on a utilisé une seule station pour l'assimilation.



(a) Champ d'ozone (en ppb) simulé le 12 Juillet 1999 à 15 heures (TC) par le modèle.

(b) Champ d'ozone (en ppb) corrigé par l'EnKF le 12 Juillet 1999 à 15 heures (TC).

FIG. 4.14: Comparaison entre le champ initial d'ozone simulé par le modèle (a) et celui corrigé par l'EnKF (b) le 12 Juillet 1999 à 15 heures (TC).

Dans la figure 4.15 on présente également la carte de l'écart-type de l'estimation effectuée le 12 Juillet 1999 à 15 heures utilisant l'EnKF. On remarque la présence des valeurs très faibles sur les zones couvertes par les stations et qui augmentent en s'approchant des bords du domaine vers le sud, mais qui diminuent sensiblement vers le nord, signe que tous les membres d'ensemble prédisent des valeurs (analyses) semblables et qui sont très proches de celles prédites par le modèle de référence.

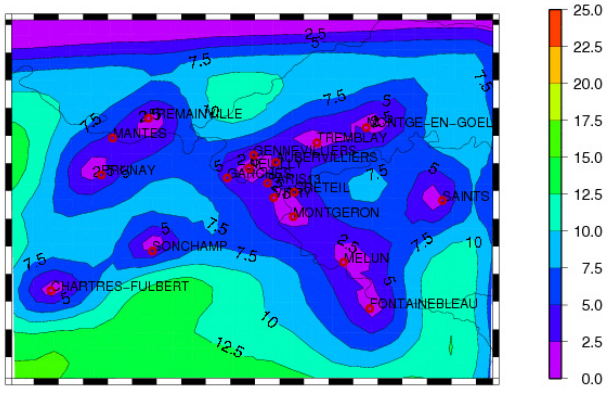
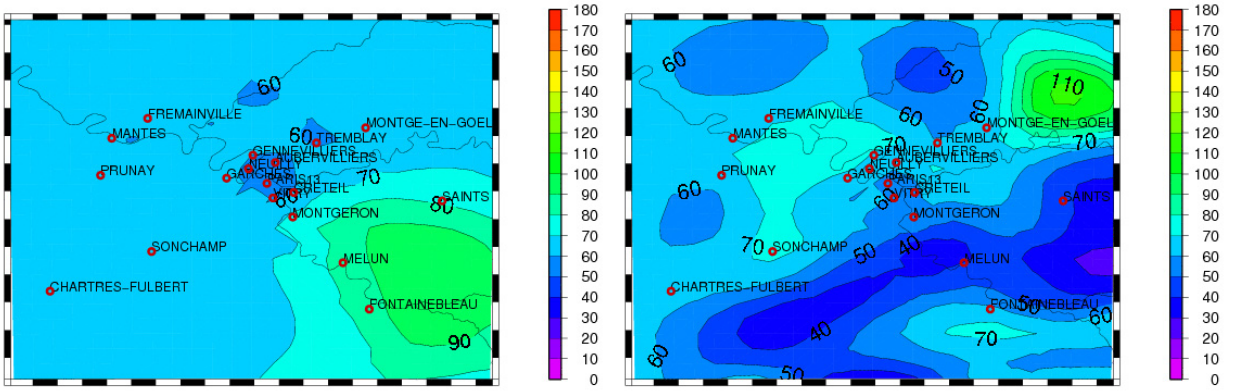


FIG. 4.15: Carte de l'écart-type de l'erreur de l'estimation par EnKF du champ d'ozone (en ppb) le 12 Juillet 1999 à 15 heures (TC).

- **Le champ d'ozone obtenu pour le 13 Juillet 1999 à 15 heures**

Une situation particulière a été enregistrée le 13 Juillet 1999 toute la journée. À cette date, le modèle présente un écart très élevé par rapport aux mesures disponibles, visible sur toutes les

séries temporelles déjà présentées, écart qui est dû probablement aux conditions limites et aux émissions. Les essais pour corriger cet écart ont donné plus ou moins satisfaction, dans la mesure où plusieurs simulations (surtout avec un nombre de membres d'ensemble inférieur à 60) ont conduit aux explosions numériques proches des bords du domaine. À 15 heures **le modèle** simule un faux panache au sud-est du domaine (figure 4.16(a)). Les concentrations mesurées aux stations situées au voisinage de ce panache montrent qu'il s'agit d'une surestimation majeure qui varie de $60-65\,\mu g/m^3$ (30 ppb) pour les stations Fontainebleau et Saints jusqu'à $92\,\mu g/m^3$ (46 ppb) à Melun.



(a) Champ d'ozone (en ppb) simulé le 13 Juillet 1999 à 15 heures (TC) par le modèle.

(b) Champ d'ozone (en ppb) corrigé par l'EnKF le 13 Juillet 1999 à 15 heures (TC).

FIG. 4.16: Comparaison entre le champ initial d'ozone simulé par le modèle (a) et celui corrigé par l'EnKF (b) le 13 Juillet 1999 à 15 heures (TC).

Cette surestimation est corrigée par le filtre, ainsi que la légère sous-estimation de $20\,\mu g/m^3$ de Sonchamp. Par contre, **l'analyse** révèle un autre panache, d'un niveau plus élevé, au coin nord-est du domaine, panache non-justifié par les mesures voisines, et en plus, non-vérifiable car il nous manque l'information nécessaire (figure 4.16(b)). Il peut être le résultat d'un artefact numérique suite aux perturbations introduites dans chaque membre de l'ensemble.

La carte de l'écart-type de l'erreur d'estimation associée exhibe des valeurs faibles pour la zone couverte par les stations de mesure. En revanche, sur la zone du panache simulé par l'EnKF, ainsi que le coin sud-est du domaine, la variance calculée atteint des niveaux assez importants.

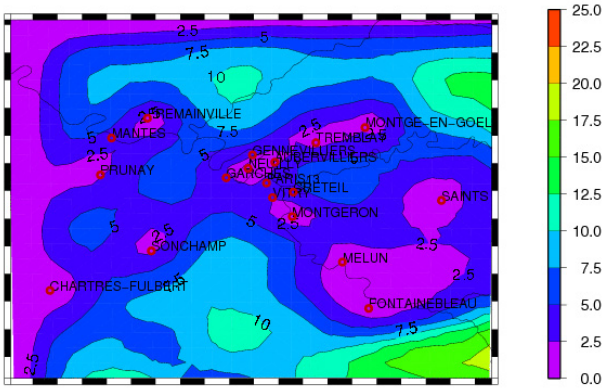
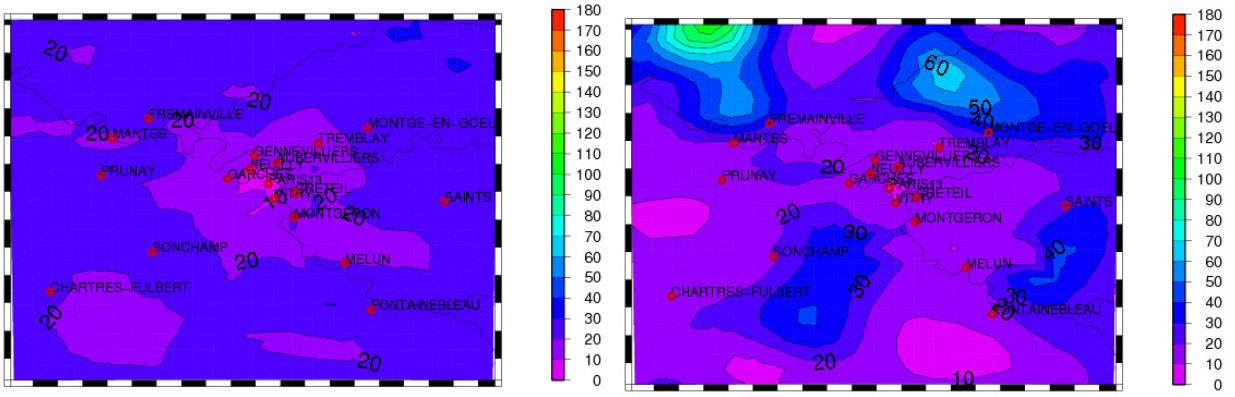


FIG. 4.17: Carte de l'écart-type de l'erreur de l'estimation par EnKF du champ d'ozone (en ppb) le 13 Juillet 1999 à 15 heures (TC).

• **Les champs d'ozone obtenus pour le 17 Juillet 1999 à 7 heures et à 16 heures**

La situation enregistrée le 17 Juillet 1999 à 7 heures est un très bon exemple de situation dans laquelle c'est le filtre qui produit une mauvaise carte, tandis que le modèle respecte la continuité du champ. À gauche, dans la figure 4.18, CHIMERE présente une très légère surestimation à côté de Sonchamp et Prunay. Partout ailleurs, on peut remarquer une bonne concordance entre les valeurs simulées et celles mesurées. À droite, sur la carte analysée, on aperçoit une explosion au coin nord-ouest du domaine et une autre, moins importante, dans la partie nord, dues probablement aux artefacts numériques introduits avec les perturbations du champ d'ozone. On peut conclure sur cet exemple que le modèle produit une bonne carte, tandis que le filtre conduit à une dégradation de celle-ci.



(a) Champ d'ozone (en ppb) simulé le 17 Juillet 1999 à 7 heures (TC) par le modèle.

(b) Champ d'ozone (en ppb) corrigé par l'EnKF le 17 Juillet 1999 à 7 heures (TC).

FIG. 4.18: Comparaison entre le champ initial d'ozone simulé par le modèle (a) et celui corrigé par l'EnKF (b) le 17 Juillet 1999 à 7 heures (TC).

Sur la carte 4.19 présentant l'écart-type de l'erreur, on retrouve des niveaux très importants au coin nord-ouest du domaine, là où l'explosion a fait son apparition, signe que l'explosion est due aux perturbations introduites dans le champ d'ozone.

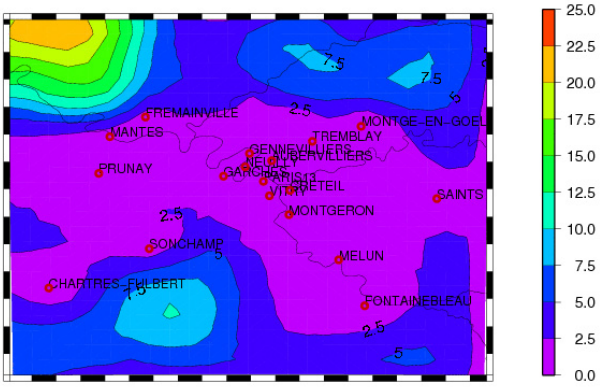
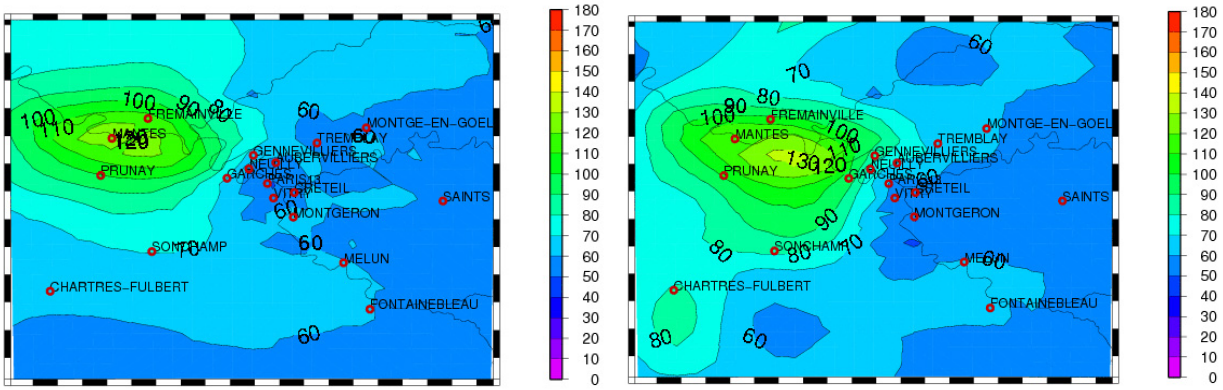


FIG. 4.19: Carte de l'écart-type de l'erreur de l'estimation par EnKF du champ d'ozone (en ppb) le 17 Juillet 1999 à 7 heures (TC).

En analysant la situation à 16 heures, tout d'abord, on remarque que le panache simulé par le modèle est un peu poussé vers l'intérieur du domaine, allongé vers le coin sud-ouest, et accentué par l'analyse (figure 4.20). Tous ces changements sont justifiés si on regarde les concentrations mesurées : à Chartres et Sonchamp les sous-estimations d'environ $32\,\mu g/m^3$ (16 ppb) et respectivement $24\,\mu g/m^3$ (12 ppb) sont réduites par le filtre ; en revanche, à Mantes le modèle simulait une légère sous-estimation qui n'a pas été corrigée par l'EnKF. Le reste du domaine était déjà en bon accord avec les mesures enregistrées, par conséquent, les corrections du champ sont mineures.



(a) Champ d'ozone (en ppb) simulé le 17 Juillet 1999 à 16 heures (TC) par le modèle.

(b) Champ d'ozone (en ppb) corrigé par l'EnKF le 17 Juillet 1999 à 16 heures (TC).

FIG. 4.20: Comparaison entre le champ initial d'ozone simulé par le modèle (a) et celui corrigé par l'EnKF (b) le 17 Juillet 1999 à 16 heures (TC).

Les variances calculées sur la zone du panache montrent des niveaux élevés, niveaux qui caractérisent d'ailleurs toute la moitié gauche du domaine à l'exception de la proximité des stations de mesure.

- **Le champ d'ozone obtenu pour le 18 Juillet 1999 à 15 heures**

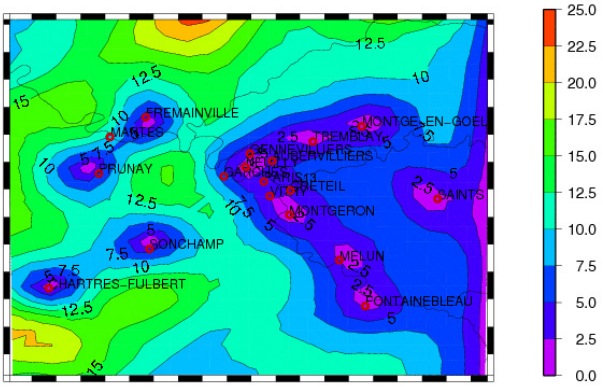
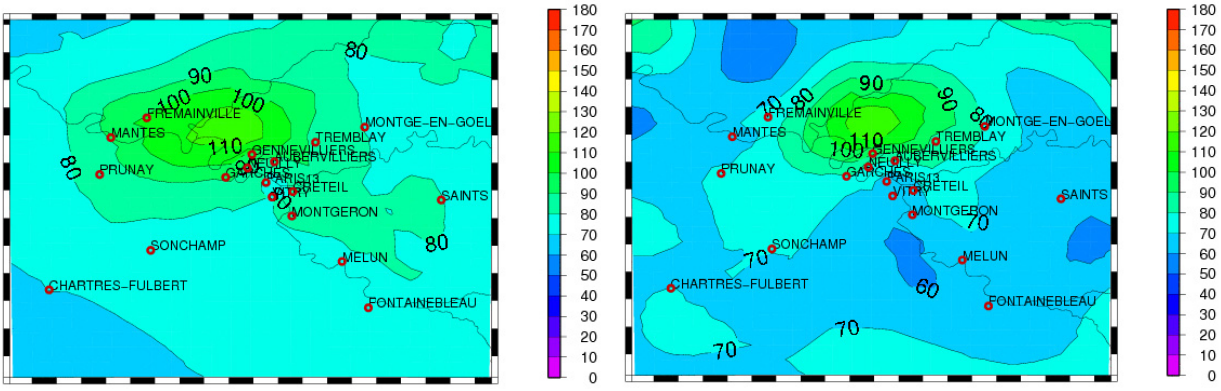


FIG. 4.21: Carte de l'écart-type de l'erreur de l'estimation par EnKF du champ d'ozone (en ppb) le 17 Juillet 1999 à 16 heures (TC).

L'épisode d'ozone produit le 17 Juillet 1999 sur la grande couronne a continué le jour suivant, le 18 Juillet 1999, mais avec une intensité un peu réduite. On présente une comparaison entre les deux champs 2D de concentrations d'ozone obtenus le même jour, à 15 heures (voir la figure 4.22) d'une part par le modèle et d'autre part corrigé par le filtre. Cette fois-ci, les différences entre les deux cartes sont nettes. Si le jour d'avant, le 17 Juillet, le modèle simulait correctement le panache, le filtre ayant corrigé uniquement son ampleur et sa forme, alors, pour le 18 Juillet, on voit distinctement une surestimation simulée par CHIMERE, sur presque toute la surface du domaine. La surestimation la plus importante ($54\,\mu g/m^3$ ou 27 ppb) est celle de Fremainville, signe que le panache n'était pas positionné correctement. Cette prémisse est renforcée par les concentrations mesurées à Prunay et Mantes. Par ailleurs, on peut constater que les concentrations d'ozone sur toute une pléiade de stations, comme : Montgeron, Saints, Fontainebleau, Melun, situées dans la zone sud-est du domaine, avec des différences moyennes de $30\,\mu g/m^3$ (15 ppb) ont été corrigées par le filtre. En revanche, pour les stations de la région parisienne qui n'ont pas été utilisées dans l'analyse, comme Vitry, Paris13, Créteil, on remarque encore une sous-estimation de l'EnKF.



(a) Champ d'ozone (en ppb) simulé le 18 Juillet 1999 à 15 heures (TC) par le modèle.

(b) Champ d'ozone (en ppb) corrigé par l'EnKF le 18 Juillet 1999 à 15 heures (TC).

FIG. 4.22: Comparaison entre le champ initial d'ozone simulé par le modèle (a) et celui corrigé par l'EnKF (b) le 18 Juillet 1999 à 15 heures (TC).

À l'image des corrections faites par le filtre sur le champ d'ozone simulé le 18 Juillet à 15 heures, la carte de l'écart-type de l'erreur (figure 4.23) exhibe des valeurs assez homogènes qui ne dépassent pas 15 ppb sur tout le domaine.

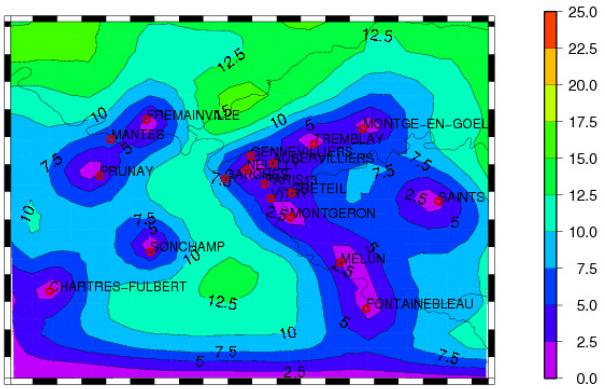


FIG. 4.23: Carte de l'écart-type de l'erreur de l'estimation par EnKF du champ d'ozone (en ppb) le 18 Juillet 1999 à 15 heures (TC).

4.8.6 L'impact de l'assimilation de l'ozone sur les estimations de dioxyde d'azote

Rappelons que les seules mesures utilisées dans le processus d'assimilation de données sont celles d'ozone. Toutefois, on a étudié également l'impact de la procédure d'assimilation sur les concentrations des autres espèces, comme le dioxyde d'azote. Le NO_2 n'est pas uniquement un précurseur de l'ozone ; comme évoqué dans le chapitre 1, par l'équation appelée de **titration** (1.3), l'ozone peut réagir avec le NO pour former du NO_2. Cet équilibre photochimique est pris en compte dans la modélisation par CHIMERE. Le fait d'assimiler l'ozone, donc de corriger les concentrations de cette espèce, peut exercer une influence sur les concentrations des autres espèces impliquées dans la destruction de l'ozone, notamment sur le NO_2.

Pour tester l'impact de l'assimilation de l'ozone sur les estimations de dioxyde d'azote, on a calculé, pour cette dernière espèce, le même genre de statistiques, c'est-à-dire la RMSE des différences analyses-observations et simulations-observations pour toute la période d'étude et pour chaque groupe de stations. On constate que, sur toutes les stations, la correction effectuée sur l'ozone n'engage pas de changements importants dans les concentrations de dioxyde d'azote : les différences sont mineures, d'environ 1-2 ppb en termes de RMSE ou MAE.

4.8.7 Des effets numériques non-souhaités

Pour tester la sensibilité du système d'assimilation par rapport à ses paramètres, on a effectué un très grand nombre de simulations. Parfois, même si les statistiques effectuées sur toute la période d'assimilation montrent une amélioration ponctuelle sur la zone couverte par des stations, on arrive à obtenir des champs de concentrations d'ozone irréguliers, très fragmentés. La perte de continuité est rencontrée souvent quand le nombre de membres d'ensemble est réduit, et diminue au fur et à mesure que ce dernier augmente. On a pu constater que même en utilisant un ensemble de 80 membres, les cartes d'ozone présentées pour le 13 Juillet 1999 à 15 heures et le 17 Juillet 1999

à 7 heures ne sont pas très continues et exhibent des faux pics, surtout sur la zone qui n'est pas couverte par les stations de mesure. Une cause possible pour la perte de continuité est la décision de perturber uniquement les champs d'ozone : on modifie de cette façon le gradient de notre champ, et on le corrige là où on bénéficie d'informations, mais on garde les conditions aux limites et les émissions sans les perturber.

Peut-on éliminer complètement les artefacts numériques ? La réponse est plutôt négative si on garde le système d'assimilation actuel. En revanche, si on choisit de perturber les conditions aux limites et les émissions, qui sont les entrées le plus incertaines du système, on aura plus de chances d'obtenir des champs plus précis et moins irréguliers. Toutefois, il faut souligner l'existence d'un certain scepticisme quant aux améliorations spectaculaires qui peuvent être obtenues, compte tenu de la quantité insuffisante de contraintes par rapport à la dimension du problème.

4.8.8 Conclusion partielle du chapitre

L'objectif de ce chapitre a été d'utiliser *conjointement* d'une part, les connaissances sur la physique et la chimie de l'atmosphère, et d'autre part, les mesures disponibles de concentrations de polluants atmosphériques pour améliorer la représentation spatiale des champs de concentrations de polluants. Pour cela, on a implémenté sur un modèle de chimie-transport, CHIMERE, une méthode séquentielle d'assimilation de données, notamment le Filtre de Kalman d'Ensemble, choisie principalement pour sa facilité d'implémentation.

Les données utilisées pour corriger les champs de polluants produits par le modèle sont des mesures horaires d'ozone, fournies par AIRPARIF, sur 17 stations situées à Paris et sur la grande couronne, auxquelles on a rajouté la station Chartres-Fulbert faisant partie de l'organisation LIG'AIR. Parmi ces 18 stations, on a utilisé 10 pour assimiler l'ozone (9 rurales et une urbaine), et les 8 autres ont été gardées comme stations-témoin, pour la validation. La période d'étude choisie est la deuxième décade du mois de juillet 1999.

L'objet de travail dans le cadre de cette procédure d'assimilation est le vecteur d'état ; dans notre cas d'étude, il a été composé des concentrations de toutes les espèces chimiques prises en compte par le modèle (44), dans tous les nœuds du maillage qui couvre le domaine (25×25), en considérant aussi l'étendue sur la verticale (8 niveaux), soit au total 220 000 valeurs.

L'idée principale de cet algorithme est d'utiliser un ensemble d'estimations d'état, à la place d'une seule, pour représenter la densité de probabilité de l'estimation du vecteur d'état et de calculer la matrice de covariance de l'erreur sur cet ensemble. On obtient, de cette façon, un **ensemble d'états** du modèle qui évoluent dans l'espace de l'état sans aucune linéarisation. La moyenne de cet ensemble représente la *meilleure* estimation, et la variance d'ensemble, la variance de l'erreur de l'estimation.

L'approche classique d'un filtre d'ensemble est de créer un ensemble initial d'états du modèle, de dimension raisonnable, (habituellement inférieure à 100 membres), et d'appliquer les deux étapes du filtre : la propagation et la correction. Le but est d'estimer, à chaque pas de temps et pour

chaque étape, le vecteur d'état et la matrice de covariance de l'erreur (x^f, P^f pour la propagation, respectivement x^a, P^a pour l'analyse), en perturbant les paramètres incertains du modèle. Parmi les paramètres incertains, inconturnables quand il s'agit de construire un modèle de chimie-transport qui doit tenir compte de la réalité physique, on peut mentionner les émissions et les conditions aux limites. Dans ce travail, on a choisi de perturber uniquement les champs de concentrations d'ozone ; la correction est faite avec un pas de temps horaire.

Globalement, on peut dire que, par assimilation des données, les différences entre les observations et les concentrations analysées diminuent beaucoup aux sites de mesure utilisées pour corriger les champs d'ozone, tandis que les corrections effectuées sur les stations-témoin sont moins importantes. En termes de RMSE moyenne calculée sur chaque groupe de stations : d'assimilation et de validation, on obtient une diminution d'environ $20\,\mu g/m^3$ sur le premier groupe et respectivement d'approximativement $10\,\mu g/m^3$ pour le groupe de validation, quand on utilise 80 membres d'ensemble.

La sensibilité du système d'assimilation par rapport à ses paramètres a été testée : le nombre de membres d'ensemble, la longueur de décorrélation qui caractérise la fonction de covariance gaussienne utilisée pour créer la perturbation, et la variance introduite dans le système par cette perturbation. Pour résumer, sur les statistiques ponctuelles effectuées aux sites de mesure, le gain obtenu en passant de 10 à 20 membres d'ensemble est significatif, par contre, la variation ultérieure est très lente. Il apparaît que 20 ou 40 membres d'ensemble reproduisent bien l'espace des erreurs ; en revanche, les champs 2D produits présentent souvent des explosions numériques sur les bords et sur les zones qui ne sont pas couvertes par les observations. En augmentant la taille de l'ensemble jusqu'à 80 membres, on arrive à obtenir des meilleures représentations, même si parfois, comme évoqué précédemment pour des cas précis, les gradients introduits dans le champ d'ozone par l'intermédiaire des perturbations, conduisent à des mauvaises représentations spatiales. Le système d'assimilation n'apparaît pas très sensible aux deux autres paramètres mentionnés.

L'impact de l'assimilation séquentielle des mesures surfaciques d'ozone sur les champs de concentrations de dioxyde d'azote a été également étudié. En utilisant uniquement les mesures d'ozone, on ne s'attend pas à obtenir des améliorations spectaculaires sur le NO_2. En effet, en termes de statistiques RMSE moyenne, sur les deux groupes de stations on constate qu'on n'obtient pas de changements majeurs dans les concentrations de NO_2. Malheureusement, les cas particuliers d'épisodes de pollution analysés montrent que, pour obtenir une meilleure représentation spatiale du NO_2, ce système d'assimilation est loin d'être suffisant.

Pour conclure, en analysant les résultats obtenus en utilisant le système d'assimilation conçu, on constate des améliorations nettes sur les zones couvertes par les stations de mesure, surtout quand ces stations ont été utilisées pour corriger le champ initial produit par CHIMERE. Par ailleurs, on peut voir qu'il y a des limitations quand on veut contraindre le modèle (et son vecteur d'état de 220000 éléments) en utilisant uniquement 10 stations de mesure : vrai problème de dimensions. On a effectué des simulations dans lesquelles on a utilisé toutes les mesures disponibles (sans garder des stations de validation), mais le gain n'est pas trop important, d'où l'idée que, c'est probablement,

la façon de perturber le champ de concentrations d'ozone qui est responsable de cette limitation. Pour palier à cet inconvénient, il faut peut être perturber les entrées incertaines du modèle comme par exemple, les émissions et les conditions aux limites. Une autre alternative, pour arriver à notre but, l'amélioration de la représentation spatiale des champs de polluants, est de construire un lisseur qui tiendra compte des toutes les informations, même postérieures au moment souhaité d'estimation.

L'assimilation de données sur un modèle de chimie-transport est la troisième méthode d'analyse spatio-temporelle appliquée dans cette étude. Pour conclure, on présente dans la section suivante une courte comparaison entre les résultats obtenus par les différentes méthodes utilisées.

4.8.9 Comparaison de distributions spatiales de polluants obtenues par les différentes méthodes appliquées

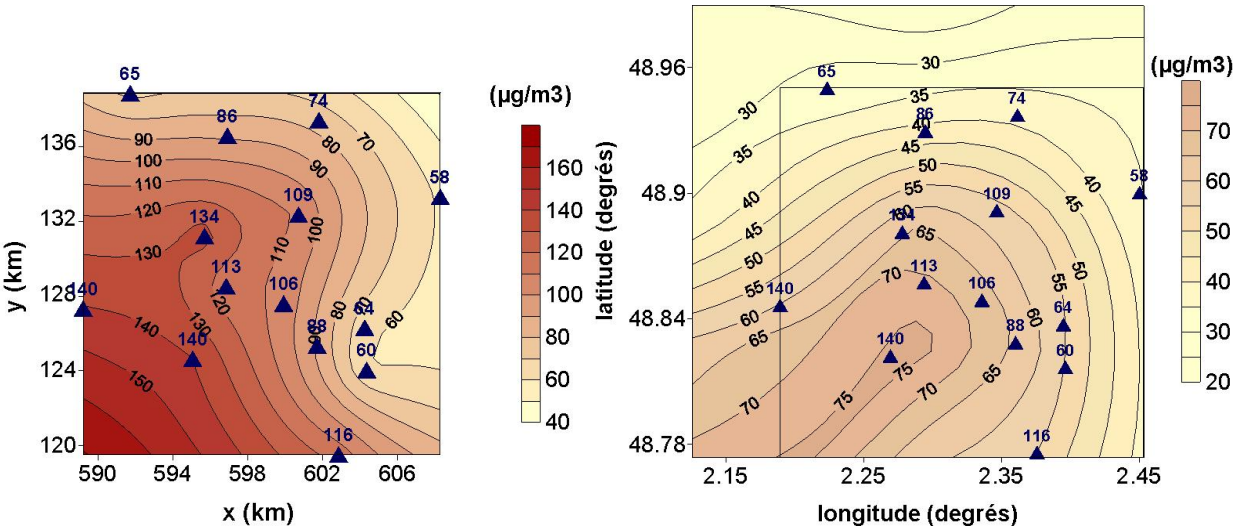
Les différentes méthodes d'analyse spatiale ou spatio-temporelle appliquées dans ce travail pour exploiter le pouvoir informationnel des mesures effectuées par un réseau de surveillance de la qualité de l'air sont : le krigeage spatial, le krigeage spatio-temporel et l'assimilation de données. Par la suite, on va analyser pour chacun de deux polluants : le dioxyde d'azote et l'ozone, les résultats obtenus en comparant les trois méthodes. Rappelons que l'assimilation a été effectuée uniquement pour la deuxième décade du mois de Juillet, et que l'interpolation spatio-temporelle appliquée a conduit à des résultats presque similaires avec ceux spatiaux avec un lissage au fur et à mesure qu'on rajoutait des données (raison pour laquelle les champs obtenus par krigeage intrinsèque spatio-temporel ne seront pas présentés dans cette section). Une autre observation très importante concerne les domaines de travail. Ceux utilisés pour effectuer l'interpolation spatiale ont déjà été décrits (voir section 2.8.2.2) ; rappelons que les coordonnées des stations utilisées étaient les transformées des coordonnées géographiques (latitude-longitude fournies par la BDQA) en coordonnées planes, Lambert I Nord. On a pu observer quelques non-concordances entre ces coordonnées et celles géographiques des stations impliquées dans l'assimilation (en degrés décimaux), raison pour laquelle, dans un souci d'homogénéité, nous avons utilisé uniquement les coordonnées des stations données par la BDQA.

4.8.9.1 Champs de NO_2

Pour le dioxyde d'azote, on commence par le cas de forte pollution enregistré le **29 Juillet 1999 à 9 heures (TC)**. Avec les données disponibles et sur le domaine d'étude choisi, qui couvrait Paris et la petite couronne, on avait obtenu une estimation spatiale assez correcte (figure 4.24(a)), dont la carte montrait un pic situé dans la zone sud-ouest du domaine, là où les mesures dépassaient les $130\,\mu g.m^{-3}$.

La seule comparaison que l'on peut effectuer pour ce cas est entre la carte produite avec l'interpolation spatiale (on a choisi le krigeage intrinsèque généralisé, mais les trois cartes obtenues étaient semblables), et celle produite par le modèle CHIMERE, sans assimilation. Ce qu'on peut remarquer faisant un zoom sur l'agglomération parisienne et découpant la carte qui lui corréspond (figure 4.24(b)), est que le modèle simule un panache orienté nord-est sud-ouest, ce qui est cohérent avec le vent prédominant à 6 heures du matin (qui soufflait du nord-est avec une vitesse de 4 m/s, et

à 9 heures il gardait encore la même direction, mais en diminuant en intensité), mais l'amplitude du panache est sous-estimée par CHIMERE d'environ 70 μg.m^{-3}. Les difficultés du modèle à reproduire convenablement les champs de concentrations de dioxyde d'azote sur la zone urbaine sont donc évidentes si on compare les deux cartes.



(a) Champ de NO_2 obtenu par interpolation spatiale le 29 Juillet 1999 à 9 heures (TC).

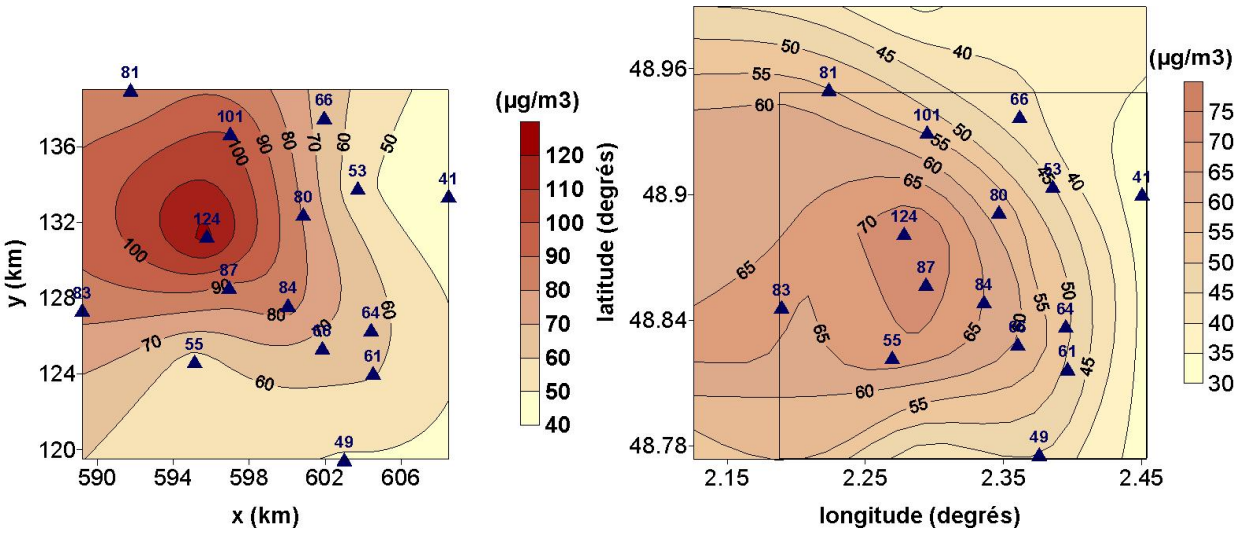
(b) Champ de NO_2 simulé par le modèle le 29 Juillet 1999 à 9 heures (TC).

FIG. 4.24: Comparaison entre le champ de dioxyde d'azote a) obtenu par interpolation spatiale et b) simulé par le modèle le 29 Juillet 1999 à 9 heures (TC).

Le deuxième cas analysé est celui du **17 Juillet 1999 à 9 heures (TC)** (quand le vent arrive de l'est avec une vitesse de 2 m/s). Comme déjà remarqué, l'assimilation des mesures surfaciques d'ozone n'exerce pas une grande influence sur les estimations d'une autre espèce chimique comme le dioxyde d'azote. Pour cette raison, on compare uniquement l'estimation obtenue par interpolation spatiale (krigeage intrinsèque généralisé, voir la figure 4.25(a)) avec le champ produit par CHIMERE (figure 4.25(b)). Comme dans le cas précèdent, de forte pollution du dioxyde d'azote, on remarque la formation d'un panache dont l'amplitude est sous-estimés. La différence (sous-estimation) atteint facilement 40 μg.m^{-3}.

Le troisième épisode, de faible pollution, concerne la même date du **17 Juillet 1999** mais l'après-midi, à **16 heures (TC)** (le vent était encore faible, 2 m/s et arrivait toujours de l'est). La même comparaison est effectuée (figure 4.26) : le champ obtenu par interpolation spatiale avec le champ produit par le modèle. Cette fois-ci, on tombe dans l'autre extrême : la surestimation par CHIMERE. Le petit panache simulé par le modèle dépasse d'environ 25 μg.m^{-3} les valeurs interpolées qui ont conduit à l'obtention du champ situé à gauche dans la figure 4.26. Les causes possibles d'une surestimation des concentrations de NO_2 sont : d'une part une surestimation des émissions et/ou un mauvais diagnostic de la hauteur de mélange ; si le mélange vertical est trop faible, cela favorise l'accumulation de NO_2 dans la première couche du modèle.

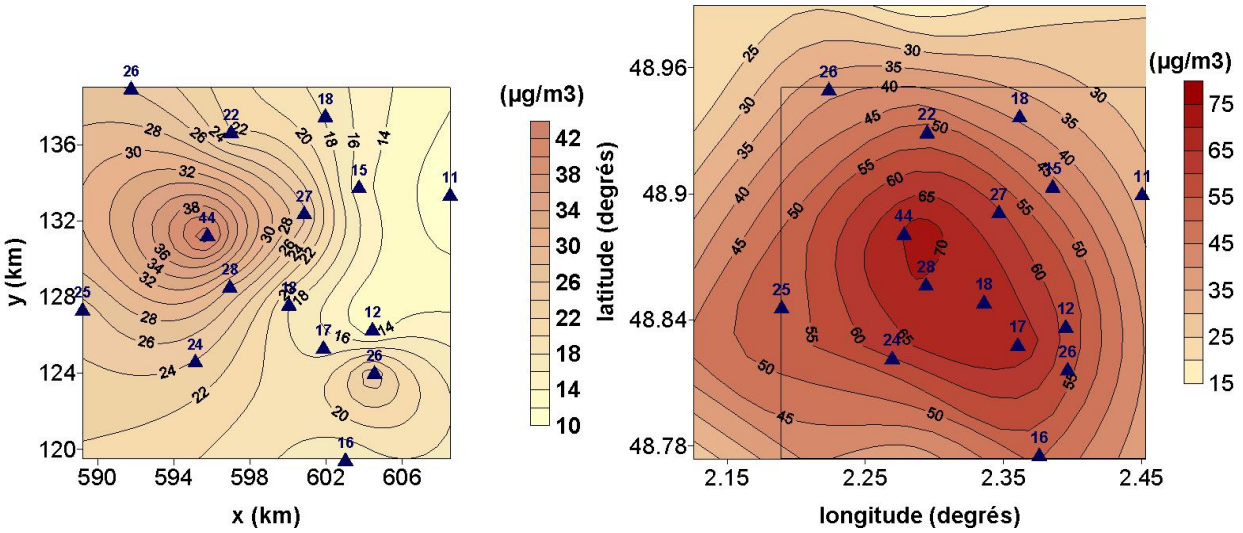
L'analyse effectuée pour les trois cas de pollution par dioxyde d'azote présentés, confirme qu'en ce qui concerne ce polluant, pour la cartographie, c'est plutôt l'interpolation spatiale qui



(a) Champ de NO_2 obtenu par interpolation spatiale le 17 Juillet 1999 à 9 heures (TC).

(b) Champ de NO_2 simulé par le modèle le 17 Juillet 1999 à 9 heures (TC).

FIG. 4.25: Comparaison entre le champ de dioxyde d'azote a) obtenu par interpolation spatiale et b) simulé par le modèle le 17 Juillet 1999 à 9 heures (TC).



(a) Champ de NO_2 obtenu par interpolation spatiale le 17 Juillet 1999 à 16 heures (TC).

(b) Champ de NO_2 simulé par le modèle le 17 Juillet 1999 à 16 heures (TC).

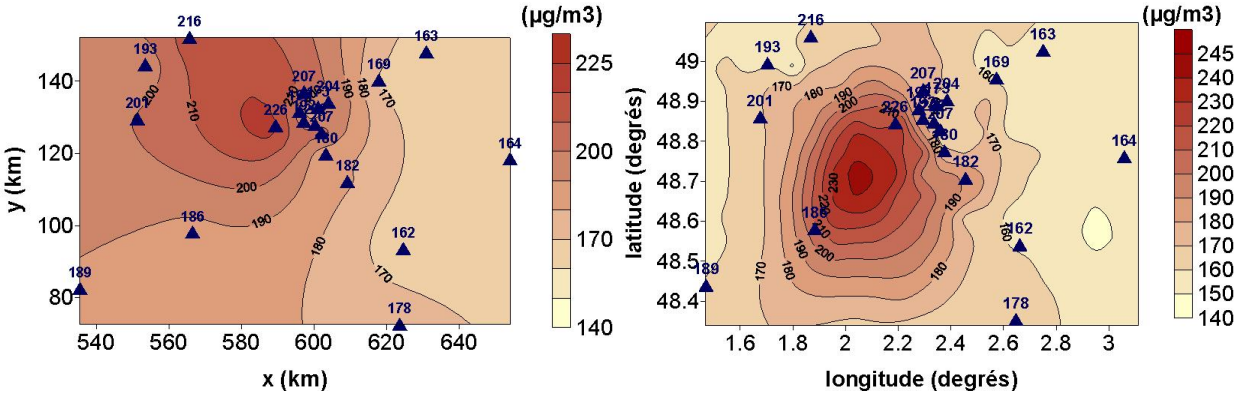
FIG. 4.26: Comparaison entre le champ de dioxyde d'azote a) obtenu par interpolation spatiale et b) simulé par le modèle le 17 Juillet 1999 à 16 heures (TC).

devrait être utilisée, car le modèle de chimie-transport simule grossièrement le NO_2.

4.8.9.2 Champs d'ozone

En ce qui concerne l'ozone, on présente aussi une brève comparaison entre les méthodes appliquées pour obtenir les champs de concentrations pour deux épisodes de forte pollution (le 17 Juillet 1999 à 16 heures et le 30 Juillet 1999 à 15 heures), et un troisième, de faible pollution (le 17 Juillet 1999 à 7 heures).

On commence par le cas principal d'ozone analysé par krigeage spatial, et repris dans le chapitre dédié à l'estimation spatio-temporelle, celui du **30 Juillet 1999 à 15 heures (TC)**. D'un point de vue météorologique, la situation à 12 heures sur Paris était caractérisée par un vent faible (2 m/s) arrivant (sur la zone centrale du domaine) du nord-est, tandis qu'à 15 heures il change un peu de direction (il arrive du nord) avec la même vitesse. Comme évoqué antérieurement, l'assimilation a été effectuée pour la deuxième décade de Juillet 1999, par conséquent on peut comparer uniquement les résultats obtenus par interpolation avec ceux simulés par CHIMERE. La carte d'interpolation spatiale présentée (figure 4.27(a)) est celle obtenue par krigeage intrinsèque généralisé qui était la plus cohérente parmi les trois types de krigeage appliqués. Elle exhibe un pic d'ozone autour de la station Garches qui présentait une concentration de 226 μg.m^{-3} ; le panache formé est orienté vers le coin nord-ouest du domaine, là où les mesures dépassent les 200 μg.m^{-3}. Rappelons que cette carte est obtenue utilisant uniquement les mesures enregistrées par les stations de mesure d'AIRPARIF. La deuxième carte présentée est découpée du domaine initial du modèle pour correspondre en gros au domaine sur lequel on a interpolé les mesures. Sur cette partie du domaine, CHIMERE simule un panache (figure 4.27(b)) situé plutôt au sud-ouest de Paris, sur la zone dépourvue de stations.

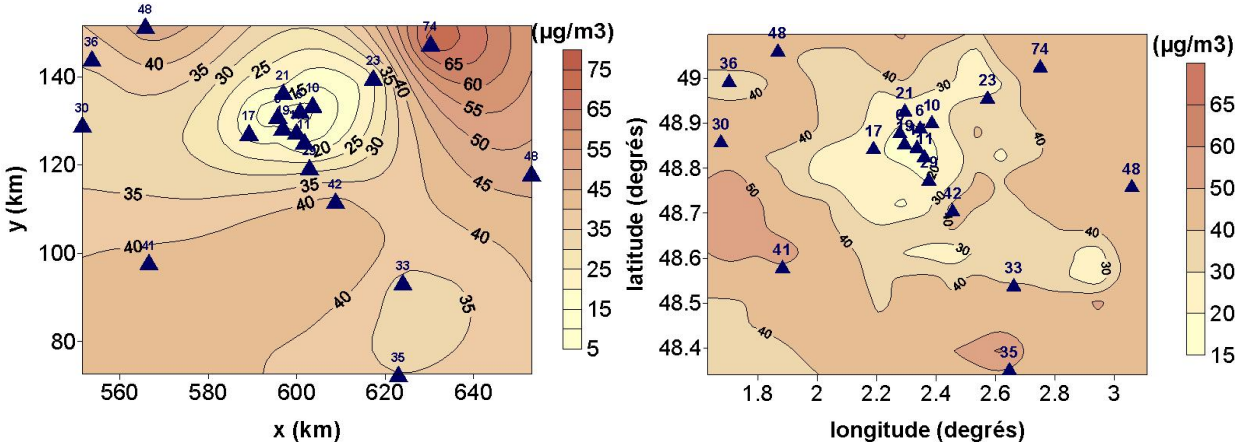


(a) Estimation spatiale d'ozone le 30 Juillet 1999 à 15 heures (TC).

(b) Estimation obtenue par CHIMERE le 30 Juillet 1999 à 15 heures (TC).

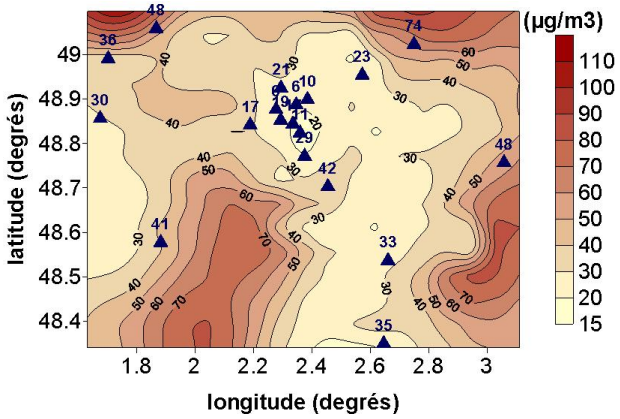
FIG. 4.27: Comparaison des champs de concentrations d'ozone le 30 Juillet 1999 à 15 heures (TC) obtenus en utilisant a) le krigeage spatial intrinsèque généralisé et b) le modèle de chimie-transport CHIMERE.

Le deuxième cas analysé est celui enregistré le **17 Juillet 1999 à 7 heures (TC)**. Cette situation est caractérisée météorologiquement par un vent très faible 2 m/s qui arrive de l'est. La première carte présentée (figure 4.28(a)) est obtenue par krigeage ordinaire. Elle exhibe des niveaux assez faibles de pollution pour cette heure matinale, excepté le coin nord-est du domaine. Par comparaison, le champ simulé par CHIMERE (figure 4.28(b)) est du même ordre de grandeur, mais il est plus fragmenté, sans accentuer les niveaux des concentrations du nord-est, où la différence par rapport à la mesure disponible est de $30 - 35\,\mu$g.m^{-3}. En revanche, le champ corrigé par le filtre d'ensemble (voir la figure 4.28(c)) simule des concentrations assez élevées au coin nord-est, mais aussi au coin nord-ouest (rappelons que le filtre présentait une explosion à cet endroit précis) ainsi que dans la zone sud, sud-ouest du domaine.



(a) Estimation spatiale d'ozone le 17 Juillet 1999 à 7 heures (TC).

(b) Estimation obtenue par CHIMERE le 17 Juillet 1999 à 7 heures (TC).

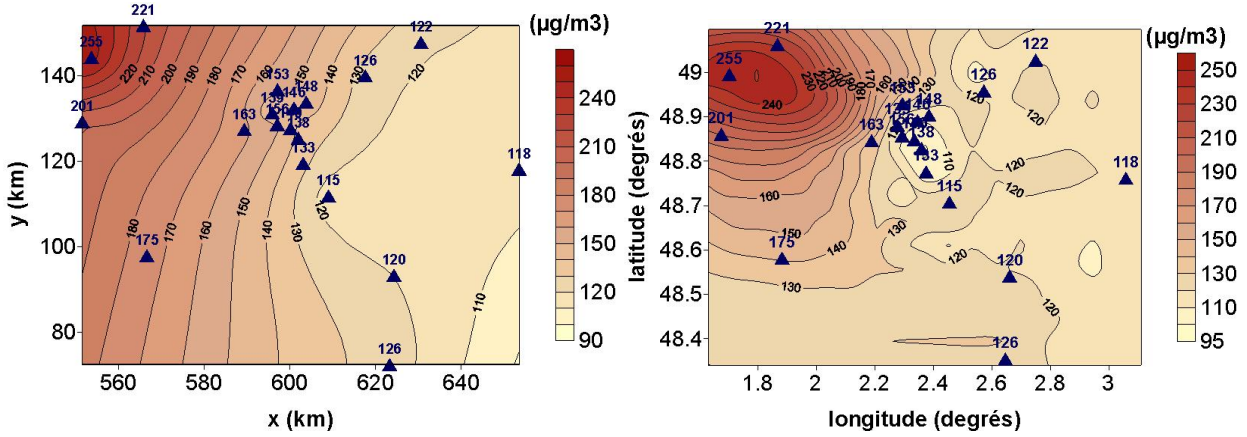


(c) Estimation obtenue par l'EnKF le 17 Juillet 1999 à 7 heures (TC).

FIG. 4.28: Comparaison des champs de concentrations d'ozone le 17 Juillet 1999 à 7 heures (TC) obtenus en utilisant a) le krigeage spatial ordinaire et b) le modèle de chimie-transport CHIMERE et c) le filtre d'ensemble.

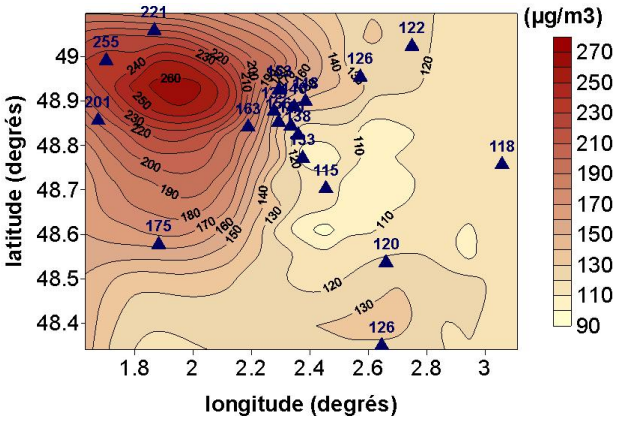
Le dernier épisode d'ozone analysé concerne la même date, **17 Juillet 1999**, mais à **16 heures (TC)**(vent faible 2 m/s qui arrive de l'est, sud-est). L'interpolation spatiale produit une carte (figure 4.29(a)) extrêmement lisse, dont les lignes d'isoconcentration sont presque parallèles. Il est évident que si on utilise uniquement les mesures disponibles on n'arrive pas à reconstituer le panache, qui est d'ailleurs bien simulé par CHIMERE (figure 4.29(b)) et bien en concordance avec le vent qui se dirige vers l'ouest. On remarque la zone où est située la station Sonchamp, station qui restitue une mesure de 175 μg.m^{-3}, valeur largement sous-estimé par le modèle. En revanche, le filtre produit un panache d'une amplitude un peu plus élevée, situé un peu plus vers l'intérieur du domaine et plus étendu vers la station Sonchamp qui est au bord du panache. La moitié droite du domaine n'exhibe pas de différences importantes pour ces deux dernières cartes analysées.

Pour conclure, après les trois situations décrites, il devient évident que les mesures seules fournies par les stations de mesure semblent ne pas être suffisantes pour reconstituer un épisode de forte pollution par l'ozone. Ceci est dû principalement au fait que ce polluant a une durée de vie assez longue, qui varie de quelques secondes à quelques heures et donc qui peut être transporté sur plusieurs kilomètres (d'où l'importance du vent). La production de l'ozone est aussi très influencée



(a) Estimation spatiale d'ozone le 17 Juillet 1999 à 16 heures (TC).

(b) Estimation obtenue par CHIMERE le 17 Juillet 1999 à 16 heures (TC).



(c) Estimation obtenue par l'EnKF le 17 Juillet 1999 à 16 heures (TC).

FIG. 4.29: Comparaison des champs de concentrations d'ozone le 17 Juillet 1999 à 16 heures (TC) obtenus en utilisant a) le krigeage spatial ordinaire, b) le modèle de chimie-transport CHIMERE et c) le filtre d'ensemble.

par les émissions, d'où l'importance de leur prise en compte dans le processus du modélisation. Statistiquement, sur la période testée (deuxième décade du mois de juillet 1999) et sur l'ozone, le filtre d'ensemble arrive à produire des analyses assez précises sur les zones couvertes par des stations de mesure, mais il y a encore de la marge pour perfectionner ce système. Généralement, la qualité des estimations produites par le modèle reste correcte et avec un système d'assimilation encore plus performant que celui décrit et implémenté actuellement sur CHIMERE, cette qualité peut encore être améliorée.

Chapitre 5

Prédiction de l'ozone par des réseaux neuronaux

Ce chapitre présente une application complémentaire à l'interpolation spatiale ou spatio-temporelle et à l'assimilation de données présentées précédemment. L'objectif de ces trois méthodes mentionnées était de produire des cartes de concentrations de polluants atmosphériques. Le but de ce chapitre est de développer une méthode de prédiction de la concentration d'un polluant, sur un horizon de 24 heures, prédiction *localisée* au niveau d'une seule station de mesure, et basée sur la série temporelle enregistrée par la station. Cette approche consiste en une méthode de type "boîte noire", plus simple du point de vue méthodologique et données nécessaires, qu'un modèle déterministe. On présente d'abord les réseaux neuronaux : l'architecture, les éléments constitutifs, les principes de fonctionnement ; ensuite, sur un cas réel, toujours celui de la région d'Île-de-France, on compare le pouvoir prédictif d'une architecture complexe, contenant 24 perceptrons multi couches (PMC), rangés en cascade, à celui d'une architecture classique, constitué d'un seul PMC, avec une couche de sortie de 24 neurones, un pour chaque heure de l'horizon de prédiction.

5.1 Contexte

Modéliser les fluctuations d'ozone et effectuer une bonne prédiction sont deux tâches très importantes dans le domaine de la qualité de l'air. Pour cela, on peut utiliser deux types de modèles : déterministes ou statistiques (boîtes noires). L'utilisation des équations différentielles pour construire un modèle déterministe pour prédire l'ozone sur un domaine bien précis est un processus très complexe qui doit tenir compte de plusieurs interactions physiques et chimiques entre les prédicteurs et qui nécessite des entrées assez précises et très nombreuses (emissions, données météorologiques, mode d'occupation du sol, conditions aux limites et initiales-voir le chapitre 4).

Vautard et al. (2001) ont développé un modèle hybride statistique-déterministe de chimie-transport pour prédire l'ozone dans la région parisienne pour l'été 1999. Le modèle utilise des prédictions météorologiques en temps réel pour prédire l'ozone sur un horizon de trois jours. La partie déterministe est représentée par le modèle de chimie-transport, CHIMERE. Pour estimer les concentrations de fond d'ozone, les auteurs utilisent un modèle de regression basé sur trois pré-

curseurs : la somme des émissions de NO_x sur les sept derniers jours, un indice de température concernant les deux derniers jours et un autre indice pour la quantité de radiations de deux derniers jours. Dans la zone urbaine, les concentrations d'ozone sont bien prédites avec un coefficient de corrélation entre les valeurs prédites et les observations de $0,7-0,8$, et une erreur quadratique moyenne qui varie entre $15-20\,\mu g.m^{-3}$ pour un horizon de 24 heures, et entre $20-30\,\mu g.m^{-3}$ pour 72 heures.

Par comparaison avec les modèles déterministes, les modèles statistiques sont plus faciles à implémenter et utiliser. Il existe plusieurs approches qui essaient d'établir une relation mathématique entre les prédicteurs et les predictants. La plupart de ces modèles nécessitent une analyse statistique préliminaire pour déterminer lesquels de ces variables sont les plus importantes pour la prédiction. En général, ces modèles sont basés sur la détection des "structures" qui sont utilisées ensuite pour prédire les concentrations de polluant désirées.

Parmi les modèles statistiques, les réseaux neuronaux artificiels, et en particulier le perceptron multi couches (PMC), ont été largement appliqués pendant la dernière décennie pour prédire les concentrations de gaz ou celles de particules. Les principes fondamentaux des réseaux de neurones (RN) permettent de les situer dans la perspective des méthodes classiques de traitement statistique de données ; la technique des réseaux de neurones formels doit être considérée comme une extension puissante des techniques bien connues, telles que la régression (Dreyfus et al., 2002). Dans la suite, on présente quelques généralités sur les RN (en anglais-ANN Artificial Neural Network).

5.2 Généralités sur les réseaux neuronaux

Selon Fausett (1994), un réseau de neurones artificiel est un système de traitement de l'information qui a quelques caractéristiques communes avec les réseaux de neurones biologiques. Un réseau de neurones consiste en un grand nombre d'éléments simples de traitement, appelés neurones. Chaque neurone est connecté aux autres par l'intermédiaire des liaisons de communication orientées, chacune avec son poids associé. Les poids représentent l'information à utiliser par le réseau pour résoudre un problème (les paramètres du modèle à ajuster).

Chaque neurone est caractérisé par un état interne, appelé activation. En réalité, un neurone est une fonction non-linéaire, paramétrée, à valeurs bornées, de ses variables d'entrée. Typiquement, un neurone envoie son activation comme un signal à quelques autres neurones. Il est à noter qu'un neurone ne peut envoyer qu'un seul signal à la fois, bien que le signal soit transmis à plusieurs neurones.

La figure 5.1 présente un réseau très simple dans lequel on a trois neurones en entrée, un dans une couche appelée cachée et deux en sortie. Le neurone caché Y, reçoit la somme des signaux pondérés (poids w_i) des neurones X_1, X_2 et X_3 ($Y = w_1X_1 + w_2X_2 + w_3X_3$) et le transforme par sa fonction d'activation, notée ici par g, qui peut être **non-linéaire**, en signal de sortie transmis

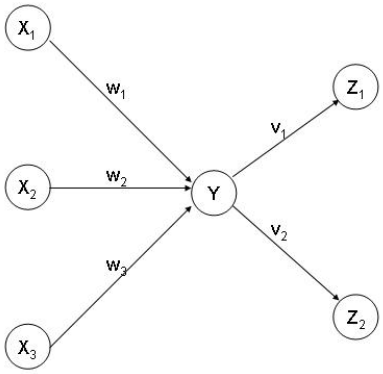


FIG. 5.1: Structure d'un réseau très simple.

après pondération (poids v_j) comme entrée à chacun de deux neurones Z_1 et Z_2 :

$$Z_1 = v_1 g(w_1 X_1 + w_2 X_2 + w_3 X_3)$$
$$Z_2 = v_2 g(w_1 X_1 + w_2 X_2 + w_3 X_3).$$

Selon le mode de propagation du signal, il existe des réseaux non-bouclés ("feed-forward") ou bouclés (récursifs). Dans le premier cas, le signal est transmis à partir des entrées vers les sorties, de sorte qu'une couche ne peut utiliser que les sorties des couches précédentes. Pour le deuxième, le graphe des connexions est cyclique : lorsqu'on se déplace dans le réseau en suivant le sens des connexions, il est possible de trouver au moins un chemin qui revient à son point de départ. L'arrangement des neurones en couches et le modèle de connexions intra- et inter-couches déterminent l'architecture du réseau.

Chaque couche reçoit un vecteur d'entrée et le transforme en vecteur de sortie. L'intérêt majeur réside dans les propriétés qui résultent de leur association en réseaux, c'est-à-dire de la composition des fonctions **non-linéaires** réalisées par chacun des neurones qui font partie d'une couche.

L'exemple le plus connu et le plus simple d'un réseau neuronal est le Perceptron Multi-Couches (PMC). Dans une telle structure, les neurones qui font partie de la même couche n'ont pas de connexion entre eux (voir un exemple dans la figure 5.2). Pour avoir un modèle non-linéaire, une couche "cachée" est nécessaire en plus de la couche de sortie. Tous les neurones faisant partie de la même couche ont la même fonction d'activation et de plus, il existe des connexions complètes entre une couche et la suivante (pour un PMC "classique"). Les choix traditionnels pour la fonction d'activation sont de type sigmoïde :

- la tangente hyperbolique $g(x) = \tanh(x)$ ou
- la fonction logistique $g(x) = \frac{1}{1+\exp(-x)}$.

Le mode de calcul d'un PMC est le suivant : la première couche reçoit le vecteur d'entrée, lui applique la fonction d'activation, et le résultat est transmis à la deuxième couche ; et la même pro-

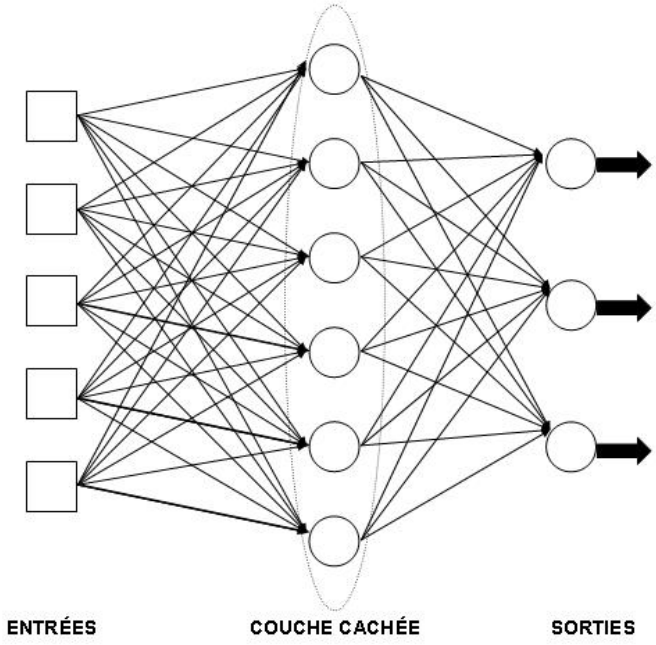


FIG. 5.2: Structure d'un PMC

cédure se répète en propageant le signal.

Les poids du réseau sont affectés aux connexions entre les couples de neurones de deux couches successives représentées par des flèches dans la figure 5.2. La méthode utilisée pour la détermination des poids est une caractéristique très importante du réseau, de même que son architecture. La technique permettant leur obtention est appelée apprentissage. Mathématiquement, on peut formuler l'objectif de l'apprentissage comme suit : trouver les poids représentés par le vecteur w tel que $x \to f(x, w)$ soit un bon modèle des données $(x_i, y_i)_{1 \leq i \leq N}$. Il existe deux types d'apprentissage : supervisé et non-supervisé. Le cas le plus courant d'apprentissage, celui supervisé, consiste à présenter une séquence x_i de vecteurs d'apprentissage en entrée, chacun associé à un vecteur de sortie y_i désirée. Les poids du réseau sont ensuite ajustés selon l'algorithme d'apprentissage, de façon à trouver pour les entrées fixées, les sorties désirées, avec un minimum d'erreur. Soit E la fonction d'erreur à minimiser :

$$E(w) = \sum_{i=1}^{N} d(f(x_i, w), y_i), \tag{5.1}$$

où d peut être la distance euclidienne. Il faut donc trouver w pour bien approcher les données, c'est-à-dire pour minimiser la fonction d'erreur qui constitue une mesure de l'écart entre les réponses réelles du réseau et les réponses désirées. Le problème à résoudre est donc un problème d'optimisation non-linéaire, qui est souvent résolu par des algorithmes d'optimisation locale comme les algorithmes de descente, où la direction de descente est choisie en fonction du gradient. Il existe plusieurs algorithmes plus performants d'un point de vue vitesse de convergence ; parmi ceux-ci on retrouve les méthodes dites de gradients conjugués : le gradient conjugué-formule BFGS (l'algorithme de **B**royden, **F**letcher, **G**oldfarb and **S**hanno), "**s**caled **c**onjugate **g**radient" (SCG), le

quasi-Newton BFGS (l'une des meilleures méthodes, mais coûteuse d'un point de vue mémoire requise) et Levenberg-Marquardt, toutes décrites exhaustivement dans Bishop (1995) par exemple.

Un état de l'art des applications de PMC dans les sciences atmosphériques a été écrit par Gardner et Dorling (1998). Les auteurs ont conclu que l'apprentissage supervisé d'un PMC peut donner une approximation correcte pour n'importe quelle fonction lisse, sans faire aucune hypothèse sur la distribution des données. Maier et Dandy (2000) ont établi le contexte général et ont développé quelques lignes directrices pour l'application de ces modèles dans d'autres domaines, comme l'hydrologie. Les auteurs ont analysé 43 articles utilisant les réseaux neuronaux et ils ont remarqué "la tendance parmi les chercheurs d'appliquer l'ANN pour des problèmes pour lesquels les autres méthodes n'ont pas été efficaces".

5.3 Applications des réseaux neuronaux pour la prédiction de l'ozone

Depuis une dizaine d'années, beaucoup de recherches se sont concentrées sur l'amélioration de la qualité de la prédiction des concentrations d'ozone sur un horizon qui peut varier de 1 heure à 24 heures.

Viotti et al. (2002) ont appliqué les réseaux neuronaux pour prédire plusieurs polluants : l'ozone, le dioxyde d'azote, le monoxyde de carbone, le benzène et les particules, en utilisant une forme spéciale de la fonction *logistique* d'activation avec trois paramètres ajustables et l'algorithme d'apprentissage appelé rétro-propagation (en anglais back-propagation). Mises à part les espèces mentionnées, d'autres entrées sélectionnées sont des données météorologiques et de trafic. Leurs résultats ne sont pas facile à interpreter, car les indices de performance calculés ne font pas partie de la liste des indices classiques.

Abdul-Wahab et Al-Alawi (2002) ont concentré leur attention sur l'identification des facteurs qui régularisent les niveaux des concentrations d'ozone pendant la journée et ils ont trouvé une contribution de la météorologie de l'ordre de $33 - 41\,\%$. Ils ont mis en évidence une importante contribution de la température, du monoxyde d'azote, du sulphate, de l'humidité relative et du dioxyde d'azote, mais une contribution moins importante qu'attendue du rayonnement solaire.

Balaguer Ballester et al. (2002) présentent une comparaison entre plusieurs modèles appliqués pour prédire, 24 heures en avance, les concentrations d'ozone : des modèles auto-régressifs à moyenne mobile avec entrées exogènes (ARMAX), le PMC et un modèle à réponse impulsionelle finie FIR (en anglais-finite impulse response). Leur attention s'est concentrée sur les pics d'ozone enregistrés en été entre 1996 et 1999 par trois stations situées en zone rurale ou urbaine en Espagne. Ils utilisent comme entrées des concentrations d'ozone et de dioxyde d'azote, ainsi que des données météorologiques comme : la vitesse du vent, sa direction et la température. Les cinq critères de performance calculés donnent de bons résultats et montrent que le PMC est un modèle plus performant que le modèle linéaire ARMAX et le modèle dynamique neuronal FIR.

Dans une étude exhaustive, Schlink et al. (2003) ont testé quinze modèles statistiques : la persistance, la régression linéaire multiple, l'autorégression, les réseaux neuronaux et les modéles additifs généralisés pour prédire l'ozone. Tous ces modèles ont été appliqués sur dix ensembles de données représentant des émissions et des données météorologiques partout en Europe. La comparaison a été faite aussi avec un modèle déterministe. Les résultats obtenus ont conduit les auteurs à rejeter l'hypothèse de la présence d'une non-linéarité dynamique dans la série temporelle d'ozone et, en même temps, ils ont affirmé l'existence d'une non-linéarité statique entre l'ozone et les données météorologiques et/ou les autres concentrations de polluants (ses précurseurs). Pour la comparison des différents modèles, ils utilisent l'indice de succès (success index -SI) et l'indice de concordance (d_2), dont les définitions sont données dans la section (5.6.4). Leurs résultats montrent un succès assez limité pour les techniques linéaires, mais une meilleure performance pour les réseaux neuronaux et les modèles additifs généralisés qui peuvent gérer mieux les non-linéarités statiques.

Kukkonen et al. (2003) effectuent une comparaison entre cinq modèles neuronaux, un modèle linéaire statistique et un autre déterministe, pour prédire les concentrations de dioxyde d'azote et de particules enregistrées par deux stations situées au centre de Helsinki. Les entrées utilisées, mises à part les concentrations de polluants mentionnés, sont des données de trafic ainsi que des données météorologiques. L'horizon de prédiction choisi est de 24 heures. Pour éviter le sur-apprentissage ils utilisent une technique de régularisation Bayésienne (Foxall et al., 2002). Les résultats obtenus montrent que les modèles non-linéaires sont plus performants que celui déterministe ou celui linéaire statistique.

Agirre-Basurko et al. (2006) présentent une prédiction en temps réel de l'ozone et du dioxyde d'azote pour un horizon de 8 heures en utilisant des données de trafic et météorologiques. Les modèles testés : le PMC, la régression linéaire et la persistance ont été comparés par l'intermédiaire des critères qui font partie de ce qu'on appelle le Model Validation Kit[1], et plus précisément : le coefficient de corrélation (R), l'Erreur Quadratique Moyenne Normalisé (NMSE), le facteur FA_2, le Biais Fractionnel (FB) et la Variance Fractionnelle (FV) présentés dans la section 5.6.4. Les résultats montrent que c'est le PMC qui est le modèle le plus performant, exhibant une bonne précision à prédire les concentrations de polluants, exceptés, paradoxalement, les cas de deux et trois heures après le moment choisi comme début de la prédiction.

Sousa et al. (2007) appliquent les réseaux neuronaux pour prédire l'ozone en utilisant l'analyse en components principales (ACP) pour réduire le nombre des entrées du modèle ainsi que leur redondance. Les résultats obtenus en utilisant les données originales sont comparées avec ceux obtenus en utilisant l'ACP. La conclusion est en faveur de l'ACP avec un coefficient de corrélation de 0.73 et une erreur quadratique moyenne de $21,78\,\mu g.m^{-3}$.

Finalement, Brunelli et al. (2007) présentent le réseau neuronal récurrent de type Elman pour prédire deux jours à l'avance la concentration diurne maximale des polluants suivants : SO_2, NO_2, O_3, CO, PM_{10} à Palerme (Italie) en utilisant comme prédicteurs météorologiques la vitesse du vent et sa direction, la pression et la température. Le coefficient de corrélation obtenu varie entre 0,72

[1] www.harmo.org/kit/default.asp

et 0,97 pour les divers polluants testés.

Ce bref état de l'art montre un certain consensus en faveur des modèles non-linéaires par rapport à ceux linéaires. Pourtant il y a un aspect qui semble être ignoré : certains auteurs utilisent les modèles statiques, d'autres des modèles dynamiques, mais le profit obtenu en utilisant une architecture dynamique, plus complexe, n'a pas été démontré. Ce choix est lié à la présence ou l'absence d'une non-linéarité dynamique dans la série temporelle de l'ozone.

Certains chercheurs ont essayé, par des méthodes variées, de vérifier la présence d'une non-linéarité dynamique dans les séries temporelles d'ozone. Parmi eux, Paluš et al. (2001) ont utilisé la technique des données de substitution (surrogate) uni- et multi-variées ainsi qu'une fonctionnelle informative théorique appelée redondance. Leur conclusion est qu'il n'y a pas de preuve sur l'existence d'une dynamique non-linéaire dans la série temporelle d'ozone de leur étude en République Tchèque ; de plus, ils ont trouvé que la série temporelle d'ozone était liée aux séries temporelles météorologiques par une relation de dépendance faiblement décroissante linéaire à long terme, et dans certains cas, renforcée par une non-linéarité à court terme. Les études effectuées par Haase et Schlink (2001) et Schlink et al. (2003) mènent vers la même conclusion que Paluš et al. (2001). Schlink et al. (2001) détectent une très faible non-linéarité dynamique dans une série temporelle d'ozone enregistrée dans la zone urbaine de Berlin, en Allemagne.

5.4 Conclusions des précédentes études et l'objectif visé

Dans ce chapitre, nous concentrons notre attention sur la prédiction de concentrations d'ozone en utilisant les réseaux neuronaux. Cette méthode, très simple, peut apporter des informations supplémentaires pour les modèles de chimie-transport appliqués sur des grilles d'une grande dimension. La sélection des modèles neuronaux parmi ceux statistiques est basée sur l'expérience acquise lors des études citées auparavant, qui mettent en évidence la supériorité des modèles non-linéaires et en particulier les réseaux neuronaux, par rapport à ceux linéaires.

Vues les discussions précédentes concernant la non-linéarité dynamique dans la série temporelle d'ozone, dans cette étude on propose une comparaison entre deux architectures neuronales : une "dynamique" et l'autre "statique". L'architecture "dynamique" est représentée par une cascade de 24 PMC rangés de telle façon que la sortie d'un perceptron constitue une entrée pour le suivant. L'architecture "statique" est représentée par un seul PMC avec une couche de sortie constituée de 24 neurones. Par convention, à partir de maintenant on appellera structure "dynamique" la structure qui tient compte des valeurs prédites précédemment (24 PMC) et "statique" le modèle constitué d'un seul PMC.

On a sélectionné une base de données qui couvre une année entière et pas seulement la période d'été pour ses éventuels pics d'ozone ; ce choix est justifié par le fait que les concentrations moyennes affectent également la santé de la population. Pour la même raison, nous nous sommes concentrés sur tout l'horizon de 24 heures et pas seulement sur la valeur diurne maximale.

5.5 Description du site, description des données

5.5.1 Zone d'étude

Les séries temporelles d'ozone et de dioxyde d'azote ont été mesurées par AIRPARIF, l'organisation responsable de la surveillance de la qualité de l'air dans la région parisienne, présentée dans le chapitre 1. Pour cette étude, on a sélectionné deux stations mesurant l'ozone, l'une située dans la zone urbaine, l'autre dans la zone rurale. La première station, urbaine, se trouve à Aubervilliers (2.3855°N, 48.9039°E-voir la figure 1.4). C'est une station où les influences dues au trafic sont très importantes et où AIRPARIF mesure les concentrations d'ozone et de dioxyde d'azote. Pour la deuxième station, rurale, localisée à Prunay (1.6749°N, 48.8580°E-voir la figure 1.3) uniquement les mesures d'ozone étaient disponibles (le NO_2 ne présente pas beaucoup d'intérêt en zone rurale pour qu'il soit mesuré). Les données couvrent la période d'un an, commençant en Août 2000 jusqu'en Juillet 2001. Les deux séries temporelles présentent des discontinuités. Comme la plupart des méthodes statistiques, la modélisation neuronale demande des échantillons complets. Pour cela, on a remplacé les valeurs absentes qui ne concernaient pas plus de 4 heures consécutives par des valeurs obtenues par interpolation linéaire entre la valeur précédente et celles qui suivaient ; sinon le jour entier a été éliminé de la base de données. Ceci a été d'ailleurs le cas pour plusieurs jours à la station d'Aubervilliers.

Les données météorologiques ont été mesurées par la station de Météo France située au Parc Montsouris à Paris. Cette station mesure en continu la température (T), l'humidité relative (RH), la durée d'ensoleillement (SD), le rayonnement global (SR), la vitesse du vent (VV) et sa direction (DV), et fournit des moyennes horaires pour chacune des variables énumérées.

5.5.2 Statistiques préliminaires

Étant donnée la complexité du processus de formation de l'ozone, il est naturel d'évaluer l'effet des variables météorologiques sur cette estimation : est-ce que leur utilisation permettra une meilleure prédiction des concentrations d'ozone ? La première analyse statistique montre une forte corrélation croisée entre l'ozone et ces variables.

Viotti et al. (2002) notent que "le choix correct du nombre d'heures qui précèdent l'heure de la prédiction pour chaque paramètre d'entrée" doit être choisi "par l'intermédiaire d'une analyse statistique précise d'autocorrélations et de corrélations croisées parmi les variables mesurées".

On a choisi pour notre étude un laps de 24 heures entre l'ozone et les données météorologiques. La raison principale pour ce choix est le fait qu'on voulait construire un modèle purement prédictif, en utilisant uniquement des données mesurées. Les corrélations croisées entre l'ozone et les données météorologiques (décalées de 24 heures) sont présentées dans les tableaux 5.1 et 5.2 pour les deux stations : Aubervilliers et Prunay et la fonction d'autocorrélation est représentée dans la figure 5.3.

Corrélations croisées	O_3	NO_2
Température (T)	0,51	-0,22
Humidité relative (HR)	-0,51	0,27
Rayonnement global (SR)	0,50	-0,25
Durée d'ensoleillement (SD)	0,31	-0,13
Direction du vent (WD)	-0,01	-0,01
Vitesse du vent (WS)	0,10	-0,38
O_3	1,00	-0,37
NO_2	-0,37	1,00

TAB. 5.1: Corrélations croisées entre la série temporelle d'ozone et celles de NO_2 et des données météorologiques (décalées de 24 heures) à Aubervilliers.

Corrélations croisées	O_3
Température (T)	0,44
Humidité relative (HR)	-0,49
Rayonnement global (SR)	0,52
Durée d'ensoleillement (SD)	0,35
Direction du vent (WD)	0,03
Vitesse du vent (WS)	0,17

TAB. 5.2: Corrélations croisées entre la série temporelle d'ozone et les données météorologiques (décalées de 24 heures) à Prunay.

Station	**Polluant**	**Valeurs manquantes** (heures)	**Moyenne** ($\mu g.m^{-3}$)	**Maximum** ($\mu g.m^{-3}$)	**Médiane** ($\mu g.m^{-3}$)	**Écart-type** ($\mu g.m^{-3}$)
Prunay	O_3	102	53,47	189,00	53,00	28,65
Aubervilliers	O_3	354	41,69	153,00	42,00	25,44
	NO_2	728	34,88	195,00	32,50	24,66

TAB. 5.3: Statistiques descriptives des données pour les concentrations d'ozone mesurées à Prunay, ainsi que les concentrations d'ozone et de NO_2 mesurées à Aubervilliers.

Le tableau 5.3 présente les statistiques descriptives pour les deux polluants. On remarque que la moyenne des mesures d'ozone à Prunay est plus grande que la moyenne à Aubervilliers, ce qui est normal, car on sait que l'ozone se forme plus en zone rurale qu'en zone urbaine. Cette remarque est valable aussi pour la valeur maximale, mais aussi pour l'écart-type.

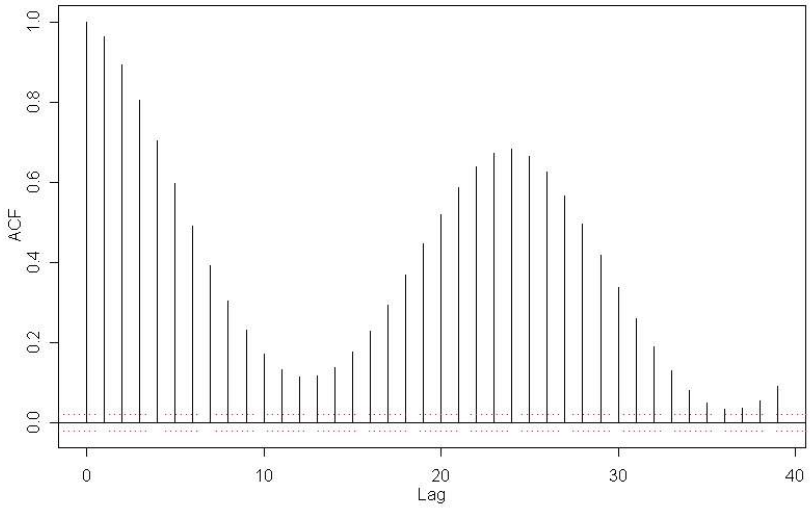


FIG. 5.3: La fonction d'autocorrélation de l'ozone (ACF) et lag en heures.

Une autre caractéristique importante pour notre base de données est le grand nombre de valeurs manquantes de NO_2, fait qui a déterminé le découpage de la base de données d'Aubervilliers en cinq blocs qui ont été traités séparément et ensuite concaténés.

5.6 Modèles neuronaux de prédiction

5.6.1 Les architectures neuronales

Le perceptron multi-couches PMC est le modèle de réseau neuronal le plus souvent utilisé dans le cas de la prédiction de la qualité de l'air. On présente ici les résultats obtenus en utilisant d'un côté une architecture classique et d'un autre côté, une architecture plus complexe, spécialement conçue, qui sera décrite.

L'architecture simple utilisée est parfaitement adaptée pour notre but. Il s'agit d'un seul PMC (voir la figure 5.4) avec une seule couche cachée qui contient 20 neurones et une couche de sortie avec 24 neurones, un pour chaque heure de l'horizon de prédiction.

L'architecture complexe est constituée d'une cascade de 24 PMC. Chaque PMC contient une couche cachée avec 3 neurones et un seul neurone dans la couche de sortie, couche qui constitue l'entrée pour le PMC suivant (voir figure 5.5). Chaque bloc d'entrée est composé de 24 valeurs horaires consécutives de concentrations d'ozone plus les données météorologiques à un seul moment de temps. Le premier bloc d'entrée contient uniquement des concentrations d'ozone mesurées, tandis que les autres blocs contiennent aussi des concentrations d'ozone antérieurement prédites par le réseau neuronal. La sélection des paramètres météorologiques a été faite en se baissant sur les corrélations croisées présentées dans la section (5.5.2). Gardner et Dorling (1998) ont proposé d'utiliser aussi comme entrées deux variables périodiques représentant le cycle diurne : $sin(2\pi h/24)$ et $cos(2\pi h/24)$, où h représente l'heure de la prédiction. Elles ont été testées, mais leur utilité dans le

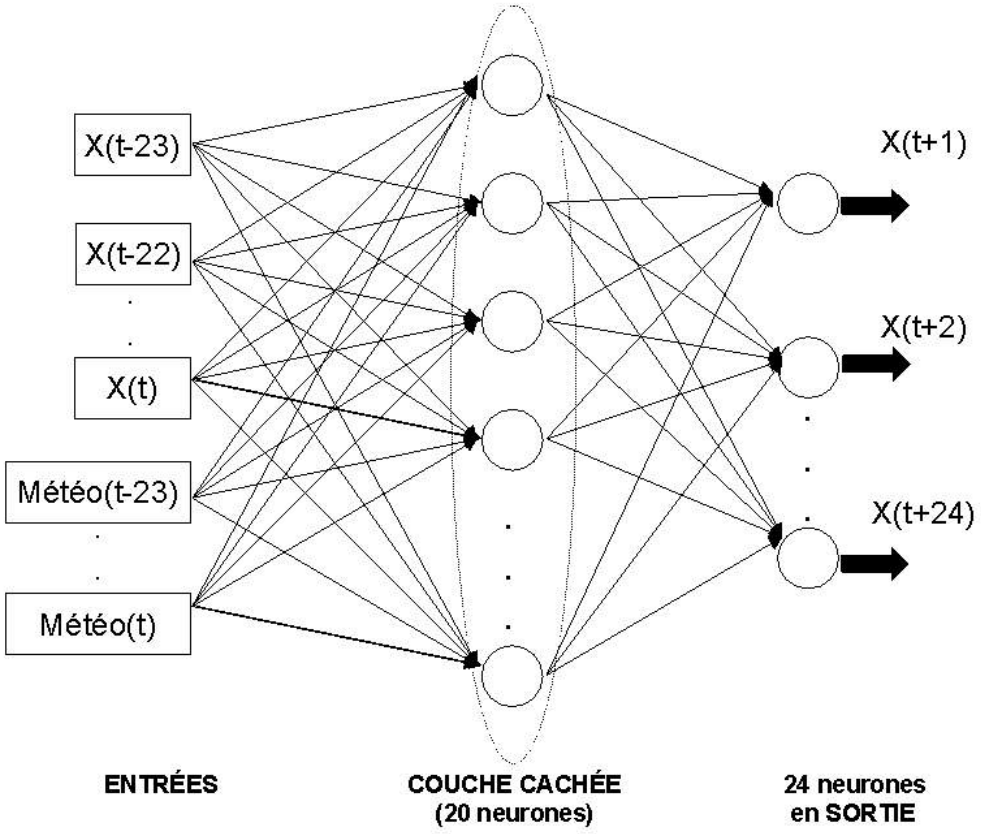


FIG. 5.4: L'architecture du réseau (1 PMC) utilisé pour la prédiction de l'ozone sur un horizon de 24 heures.

cas précis décrit dans cette étude n'a pas été démontrée.

La relation entre le nombre d'échantillons d'apprentissage et le nombre de poids ajustables est crucial (Maier et Dandy, 2000). Le deuxième ne doit pas dépasser le premier (en réalité, il est préférable d'avoir un ratio assez important entre les deux). Dans notre cas, avec une architecture complexe, on prend la décision de "superposer" les données. En négligeant l'heure de la journée à laquelle on commence la prédiction, nous élargissons la base de données. Pour chaque échantillon d'entrée, nous déplaçons les données d'une heure de telle façon que, à chaque déplacement, la valeur précédente d'ozone est éliminée et la suivante est ajoutée. Par exemple, si le premier échantillon commence à 1 heure du matin et finit à minuit, l'échantillon suivant commencera à 2 heures du matin le premier jour et finira à 1 heure du matin le jour suivant, etc. De cette manière, on est capable de générer approximativement 8000 cas d'apprentissage.

Les données d'entrée ont été d'abord standardisées, multipliées par un facteur de 0.7 (Ionescu et Candau, 2007) et après randomisées. Ce traitement est justifié dans la section suivante.

Les résultats obtenus avec les 2 architectures neuronales seront d'une part intercomparés, d'autre part, évalués par rapport au modèle le plus simple, qui est la persistance. Le modèle de

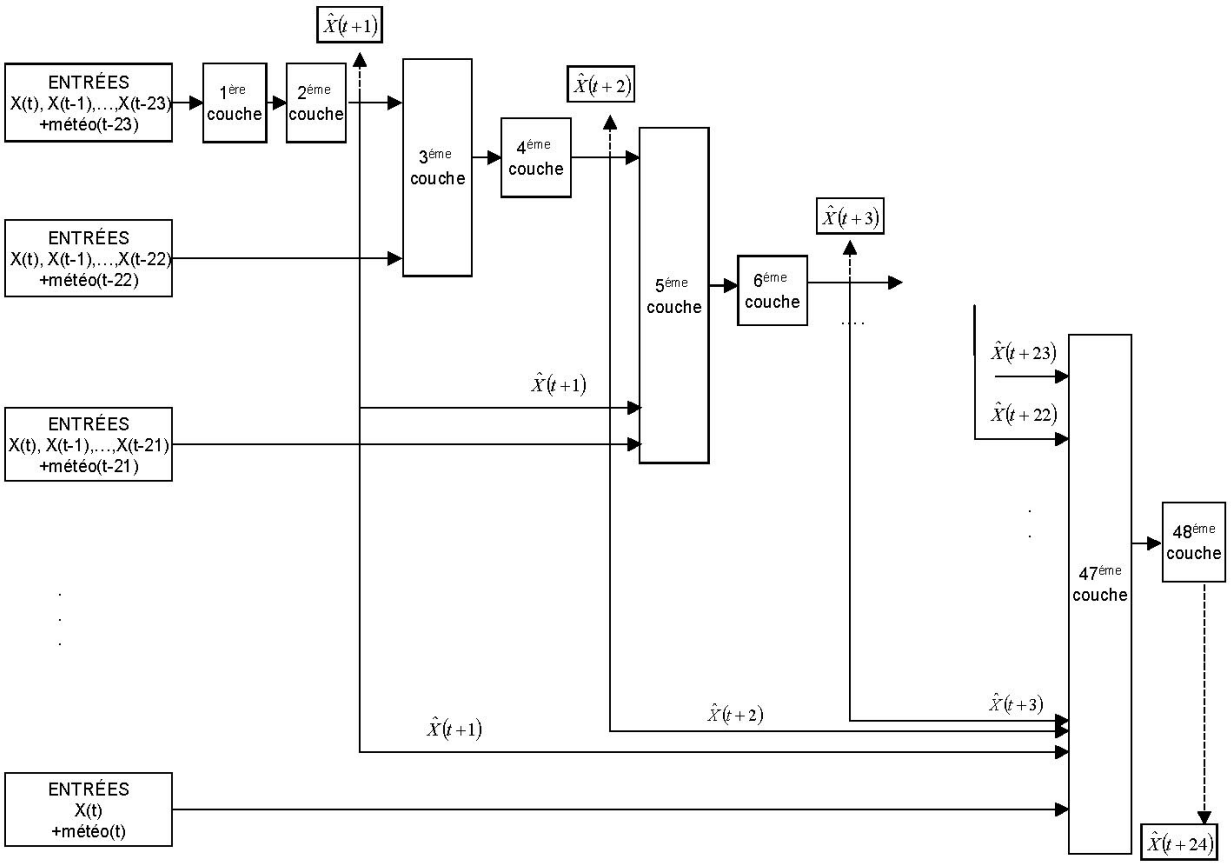


FIG. 5.5: L'architecture du réseau (24 PMC) spécialement conçu pour la prédiction de l'ozone sur un horizon de 24 heures.

persistance consiste à attribuer à chaque concentration de l'horizon de prédiction, la valeur mesurée le jour précédent à la même heure.

5.6.2 Fonctions d'activation et algorithmes d'apprentissage

Le problème d'apprentissage dans les réseaux neuronaux peut être formulé en termes d'optimisation : il s'agit de minimiser la fonction d'erreur. La fonction d'erreur utilisée dans cette étude est l'erreur quadratique moyenne et elle constitue une mesure de l'écart entre les réponses réelles du réseau et les réponses désirées. Cette optimisation se fait de manière itérative, en modifiant les poids en fonction du gradient de la fonction de coût : le gradient est estimé par une méthode spécifique aux réseaux de neurones, dite méthode de rétro-propagation (Rumelhart et al., 1986), puis il est utilisé pour l'algorithme d'optimisation proprement dit.

On a sélectionné deux méthodes d'optimisation du gradient conjugué : SCG (scaled conjugate gradient, Möller (1993)) et celle de quasi-Newton BFGS (Shanno, 1978; Bishop, 1995). Ce sont des méthodes d'optimisation locale, basées sur des techniques itératives de résolution, qui ne conduisent pas à un optimum global, mais généralement les solutions obtenues sont considérées comme satisfaisantes. La sélection de ces méthodes d'optimisation est justifiée principalement par la demande

réduite de mémoire active et pas par le taux de convergence. Cependant, la méthode SCG offre parfois une amélioration importante de la vitesse de convergence comparée aux autres algorithmes de type gradient conjugué, plus conventionnels (Bishop, 1995).

Comme les deux méthodes sont sensibles à l'échelle des données, nous les avons standardisées. De plus, la fonction d'activation utilisée pour les neurones appartenant aux couches cachées est une sigmoïde en particulier la fonction tangente hyperbolique, qui est comprise entre -1 et 1. C'est la deuxième raison qui nous a conduit à standardiser les données initiales (Fausett, 1994).

5.6.3 Généralisation

Pour la prédiction, la propriété la plus importante d'un algorithme est son pouvoir de généraliser et de filtrer le bruit. Le pouvoir de généraliser se traduit par la capacité du modèle de faire des prédictions correctes quand on l'applique sur des ensembles qui n'ont pas été utilisés pour le calibrer. Parfois, il se peut que le réseau apprenne "par cœur" les exemples d'apprentissage et qu'il ne puisse pas généraliser par la suite, face aux nouveaux exemples (Schlink et al., 2003).

Pour éviter le sur-apprentissage il existe des techniques de régularisation qui peuvent être utilisées. Parmi celles-ci on peut mentionner celle de l'"arrêt prématuré" (en anglais, *Early Stopping*). Pour l'appliquer, il faut d'abord diviser l'ensemble initial de données en trois sous-ensembles disjoints appelés : ensemble d'apprentissage, de validation et de test (Bishop, 1995; Ripley, 1996). Le premier ensemble est utilisé pour calculer le gradient et actualiser les poids ajustables du réseau. L'erreur calculée sur le deuxième ensemble est surveillée pendant l'apprentissage et quand le réseau arrive au sur-apprentissage cette erreur commence à croître. À ce moment, l'apprentissage est arrêté et les derniers poids obtenus sont retournés comme solution du problème. Le troisième ensemble est utilisé pour évaluer la performance du modèle (Nunnari et al., 2004). Dans notre étude, l'ensemble d'apprentissage représente 60% de la base initiale, tandis que les deux autres 20% chacun, la division étant faite après randomisation de la base initiale.

5.6.4 Indices de performance

Pour évaluer la performance d'un modèle de prédiction on calcule souvent des indices statistiques, soit pour chaque heure de l'horizon de prédiction de l'ensemble de test, soit comme des indices globaux pour tout cet ensemble de test.

Parmi les indices statistiques classiques, on a décidé de calculer l'erreur quadratique moyenne (RMSE), la moyenne de l'erreur absolue (MAE), le biais moyen (MBE) et le coefficient de détermination (R^2) (voir l'annexe B).

De plus, on détermine l'indice de concordance d_2 (Willmott, 1982) qui est une mesure capable d'exprimer le degré d'absence d'erreur (Gardner et Dorling, 2000) et qui permet la comparison entre

différents modèles appliqués sur différents ensembles de données (Schlink et al., 2003) :

$$d_2 = 1 - \frac{\sum_{i=1}^{n} |P_i - O_i|^2}{\sum_{i=1}^{n} \left(|P_i - \overline{O}| + |O_i - \overline{O}|\right)^2}. \tag{5.2}$$

Les autres indices mentionnés dans cette étude sont l'erreur quadratique moyenne normalisée (NMSE) :

$$NMSE = \frac{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}{\overline{O} - \overline{P}}, \tag{5.3}$$

le facteur FA_2 qui donne le pourcentage des prédictions pour lesquelles la valeur du rapport O/P appartient à l'interval $[0.5, 2]$, le biais fractionnel (FB) :

$$FB = 2\frac{\overline{O} - \overline{P}}{\overline{O} + \overline{P}} \tag{5.4}$$

et la variance fractionnelle (FV) :

$$FV = 2\frac{\sigma_O^2 - \sigma_P^2}{\sigma_O^2 + \sigma_P^2}, \tag{5.5}$$

avec P_i et O_i les concentrations d'ozone prédites et respectivement mesurées et $\overline{O}$ la moyenne des observations.

Pour prédire les dépassements des différents seuils pré-établis (comme le seuil d'alerte), Schlink et al. (2006) proposent trois indices spécifiques. Le premier, le taux vrai positif (TPR) correspond à la fraction de dépassements correctement prédits :

$$TPR = A/M, \tag{5.6}$$

où A représente les dépassements correctement prédits et M tous les dépassements observés. Le second est le taux "faux positif" (FPR) calculé avec la formule :

$$FPR = \frac{F - A}{N - M}, \tag{5.7}$$

où N représente le nombre d'événements considérés et F tous les dépassements prédits. Le dernier est l'indice de succès (SI) qui combine les deux précédents :

$$SI = TPR - FPR. \tag{5.8}$$

5.7 Résultats et discussion

5.7.1 Résultats à Aubervilliers

5.7.1.1 Les résultats du modèle 24 PMC

On commence la présentation des résultats à Aubervilliers par ceux obtenus en utilisant le modèle 24 PMC. Il s'agit de huit séries de résultats statistiques, pour chaque simulation effectuée, en termes d'indices globaux calculés sur les ensembles de test (voir tableau 5.4) afin de sélectionner

Entrées (Nombre) : Description	**Algorithme d'apprentissage**	**Moyenne** (μg.m^{-3})	**Écart-type** (μg.m^{-3})	**RMSE** (μg.m^{-3})	**MAE** (μg.m^{-3})	**MBE** (μg.m^{-3})	$\mathbf{d_2}$	$\mathbf{R^2}$
(24) : O_3	BFGS	32,05	28,55	19,55	14,87	0,46	0,83	0,54
	SCG	31,41	27,80	19,01	14,51	0,17	0,83	0,53
(26) : O_3, T, NO_2	BFGS	29,85	27,43	18,48	14,07	-0,05	0,84	0,55
	SCG	30,84	27,71	18,49	14,14	0,42	0,84	0,55
(28) : O_3, T, RH, SR, NO_2	BFGS	30,13	27,24	17,87	13,66	0,08	0,85	0,57
	SCG	30,81	27,84	18,32	13,90	0,25	0,85	0,57
(30) : O_3, T, RH, SR, SD, WS, NO_2	BFGS	29,83	26,74	17,62	13,54	-0,01	0,85	0,57
	SCG	30,50	26,90	18,00	13,87	0,03	0,85	0,55

TAB. 5.4: Résultats obtenus à Aubervilliers. Tous les indices de performance (RMSE, MAE, MBE, d_2 et R^2) sont calculés sur l'ensemble de test et sur tout l'horizon de prédiction.

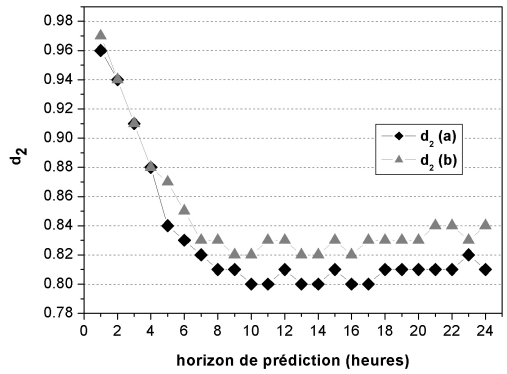
la "meilleure" configuration possible pour cette architecture assez complexe.

Pour ce site on a eu un problème majeur avec les données manquantes de NO_2, raison pour laquelle on a éliminé plusieurs jours pour obtenir une base commune pour toutes les données qui ont été utilisées dans le processus de prédiction. Ensuite, chaque simulation a été effectuée deux fois, avec deux algorithmes d'apprentissage différents (BFGS et SCG). Les entrées pour chaque simulation suivent un chemin ascendent (voir le tableau 5.4). On commence avec 24 entrées : uniquement des concentrations d'ozone mesurées ; ensuite on rajoute deux prédicteurs : la température (T) et le NO_2, au total 26 entrées. Pour la simulation qui exhibe 28 entrées on a rajouté deux autres : l'humidité relative (RH) et le rayonnement global (SR), tandis que pour la quatrième les deux derniers prédicteurs rajoutés sont la durée d'ensoleillement (SD) et la vitesse du vent (WS), au total 30 entrées.

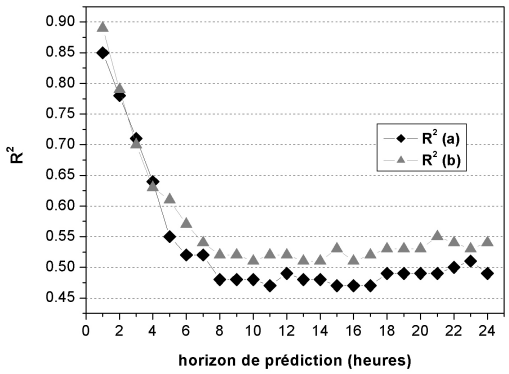
Un bon compromis entre la dimension du réseau et le nombre de prédicteurs utilisés montre que les meilleurs résultats en termes d'indices de concordance d_2 et de coefficient de corrélation sont obtenus en utilisant l'algorithme BFGS et les entrées suivantes : les 24 heures précédentes de concentrations d'ozone, trois paramètres météorologiques : température, humidité relative et rayonnement global, ainsi que les mesures de NO_2 (voir le tableau 5.4).

Les meilleurs résultats varient de $R^2 = 0,89$, RMSE =$9,03\,\mu$g.m^{-3}, $d_2 = 0,97$ et MAE =$6,51\,\mu$g.m^{-3} pour la première heure de l'horizon de prédiction, jusqu'à $R^2 = 0,54$, RMSE =$18,95\,\mu$g.m^{-3}, $d_2 = 0,84$ et MAE =$14,88\,\mu$g.m^{-3} à la fin de l'horizon de prédiction. Ces valeurs peuvent être retrouvées dans la figure 5.6, respectivement pour chaque indice, sur la courbe notée (b) dans chacune de quatre figures, courbe qui représente l'évolution de l'indice de performance sur l'horizon de prédiction pour la simulation la plus performante effectuée.

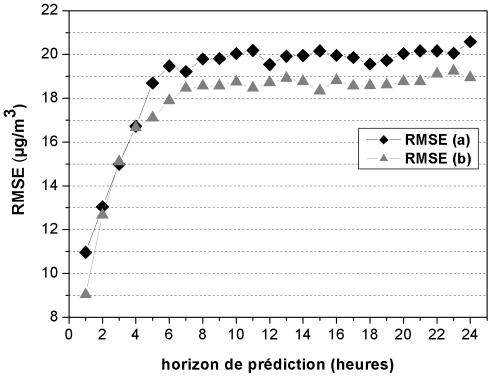
Toutefois, on peut remarquer que les modèles appliqués exhibent des niveaux similaires de performance. Ceci est evident si on visualise la figure 5.6, où les quatres indices de performance (d_2, R^2, RMSE, MAE) ont été représentés pour la meilleure simulation et respectivement la moins



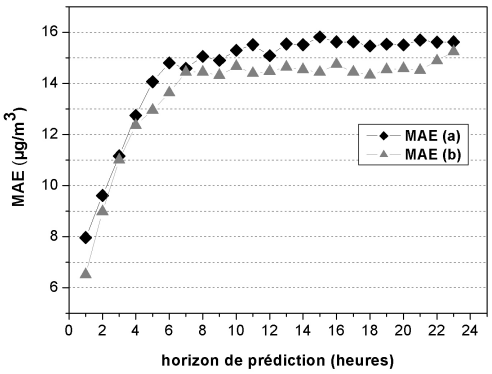
(a) d_2



(b) R^2



(c) RMSE



(d) MAE

FIG. 5.6: Comparaison des quatre indices de performance (d_2, R^2, RMSE et MAE) sur deux simulations effectuées à Prunay : la moins performante (en noir) et la plus efficace (en gris) sur tout l'horizon de prédiction.

performante effectuées à cette station. On peut remarquer également une même evolution : graduellement décroissante pour les deux premiers indices, et croissante pour les deux derniers, avec une tendance de stabilisation après le premier quart ou tiers de l'horizon de prédiction, période sur laquelle les deux courbes d'erreur sont pratiquement superposées.

Une autre remarque qu'on peut faire est que les variables exogènes, i.e. météorologiques et les concentrations de NO_2 n'ont pas augmenté sensiblement la performance du modèle. Il est probable que leur influence ne soit pas cruciale pour la prédiction de l'ozone, s'il n'y a pas une variation très forte d'une journée à l'autre. Des telles variations apparaissent uniquement lors des pics de pollution, mais, dans notre base de données, les exemples de pics ne sont pas très nombreux. Une autre explication possible est le fait que les paramètres météorologiques ont été enregistrés dans une station située au centre de Paris, tandis qu'Aubervilliers est situé loin, en pleine banlieue ; leur représentativité peut donc sérieusement être mise en doute.

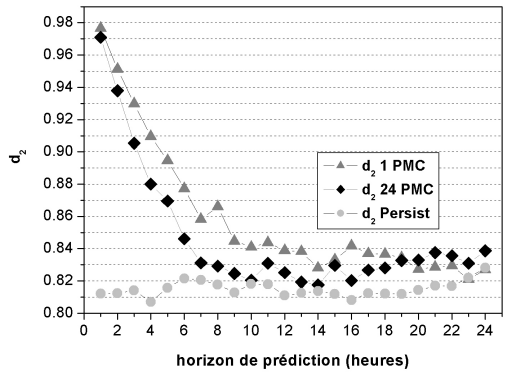
5.7.1.2 Comparaison effectuée à Aubervilliers entre les modèles : 24 PMC, 1 PMC et la persistance

On a comparé les meilleurs résultats obtenus en utilisant le modèle 24 PMC avec les résultats obtenus en utilisant d'un côté une architecture plus simple, codée 1 PMC, avec les mêmes entrées, et d'un autre côté le modèle de persistance (voir le tableau 5.5).

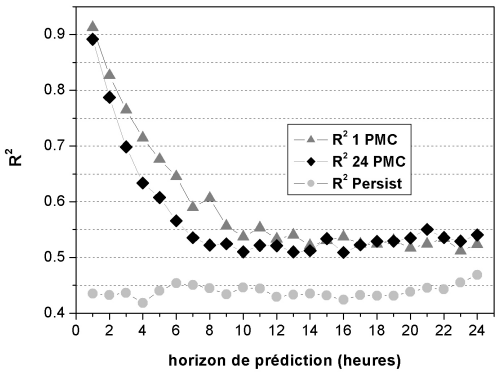
Modèle	**Moyenne** (μg.m^{-3})	**Écart-type** (μg.m^{-3})	**RMSE** (μg.m^{-3})	**MAE** (μg.m^{-3})	**MBE** (μg.m^{-3})	$\mathbf{d_2}$	$\mathbf{R^2}$
24 PMC	30,13	27,24	17,87	13,66	0,08	0,85	0,57
1 PMC	30,51	27,71	17,70	13,36	-0,34	0,86	0,59
Persistance	30,13	27,24	22,31	16,37	-0,21	0,82	0,44

TAB. 5.5: Résultats comparatifs obtenus à Aubervilliers. Tous les indices de performance (RMSE, MAE, MBE, d_2 et R^2) sont calculés sur l'ensemble de test et sur tout l'horizon de prédiction.

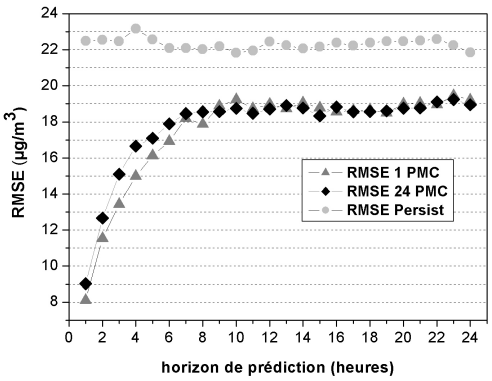
De cette comparaison, il ressort que l'architecture "statique" 1 PMC semble être légèrement plus performante que celle "dynamique", la cascade 24 PMC. Si on regarde les quatre statistiques sur tout l'horizon de prédiction (voir figure 5.7) on peut remarquer que, au moins pour les huit premières heures, le modèle 1 PMC dépasse en performance le modèle 24 PMC. Ce résultat, plutôt surprenant, peut témoigner de l'absence d'une certaine non-linéarité dynamique dans la série temporelle d'ozone, et/ou dans la relation d'entre ozone et les autres variables météorologiques. Rappelons que les études publiées par Paluš et al. (2001), Haase et Schlink (2001) et Schlink et al. (2003) conduisaient à la même conclusion *via* différentes méthodes que celle présentée ici.



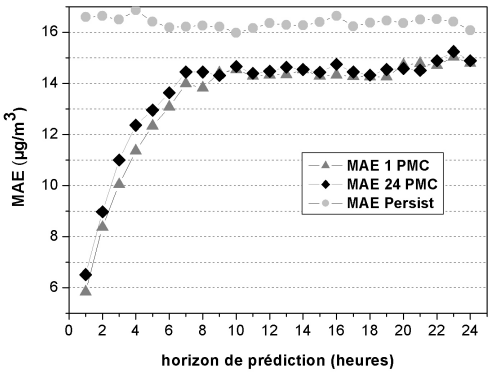
(a) d_2



(b) R^2



(c) RMSE



(d) MAE

FIG. 5.7: Indices de performance obtenus à Aubervilliers pour les trois modèles appliqués (1 PMC, 24 PMC, et la Persistance).

5.7.2 Résultats à Prunay

5.7.2.1 Les résultats du modèle 24 PMC

Une caractéristique importante des données enregistrées à Prunay est que les séries de valeurs manquantes ne sont pas très étendues, exceptée la fin du mois de juillet 2001. À cette station, on a effectué, comme précédemment, quatre types de simulations en utilisant, au début, seulement les mesures d'ozone (24 entrées) ; ensuite on a rajouté les deux prédicteurs périodiques ($sin(2\pi h/24)$ et $cos(2\pi h/24)$, où h représente l'heure de la prédiction (26 entrées). Pour les simulations avec 28 entrées, aux concentrations d'ozone on a rajouté quatre variables météorologiques : la température (T), l'humidité relative (RH) et le rayonnement global (SR) pour finir avec la durée d'ensoleillement (SD). Enfin, toutes les données disponibles (30 entrées) ont été prises en compte dans la dernière simulation. Chaque simulation a été effectuée en utilisant les deux algorithmes d'apprentissage mentionnés : BFGS et SCG. Les indices globaux obtenus sont présentés dans le tableau 5.6.

Encore une fois, il faut remarquer la faible différence entre les résultats obtenus avec ces différentes entrées. Comparés aux résultats obtenus à Aubervilliers, l'utilisation de toutes les données disponibles a été la plus efficace.

Entrées (Nombre) : Description	**Algorithme d'apprentissage**	**Moyenne** ($\mu g.m^{-3}$)	**Écart-type** ($\mu g.m^{-3}$)	**RMSE** ($\mu g.m^{-3}$)	**MAE** ($\mu g.m^{-3}$)	**MBE** ($\mu g.m^{-3}$)	$\mathbf{d_2}$	$\mathbf{R^2}$	**SI**
(24) : O_3	BFGS	53,86	28,00	18,09	13,89	0,45	0,86	0,58	0,52
	SCG	53,70	28,36	18,62	14,15	0,14	0,85	0,57	0,58
(26) : O_3, sin(t), cos(t)	BFGS	53,69	28,66	17,99	13,76	0,04	0,87	0,61	0,59
	SCG	54,16	28,38	17,91	13,75	0,31	0,86	0,60	0,74
(28) : O_3, T, RH, SR, SD	BFGS	53,84	29,11	17,91	13,71	0,15	0,87	0,62	0,71
	SCG	53,25	29,20	17,88	13,79	-0,28	0,87	0,63	0,81
(30) : O_3, T, RH, SR, SD, sin(t), cos(t)	BFGS	52,96	28,88	17,61	13,63	-0,38	0,88	0,63	0,66
	SCG	53,56	28,90	17,57	13,49	0,04	0,88	0,63	0,64

TAB. 5.6: Résultats obtenus en utilisant le modèle 24 PMC à Prunay. Tous les indices de performance (RMSE, MAE, MBE, d_2, R^2 et SI) sont calculés sur l'ensemble de test et sur tout l'horizon de prédiction.

Les indices de performance varient entre $R^2 = 0,92$, RMSE =$8,2\,\mu g.m^{-3}$, $d_2 = 0,98$ et MAE =$6,13\,\mu g.m^{-3}$ pour la première heure de l'horizon de prédiction, et $R^2 = 0,56$, RMSE =$19,55\,\mu g.m^{-3}$, $d_2 = 0,84$ et MAE =$15,56\,\mu g.m^{-3}$ à la fin de l'horizon de prédiction (voir la figure 5.8).

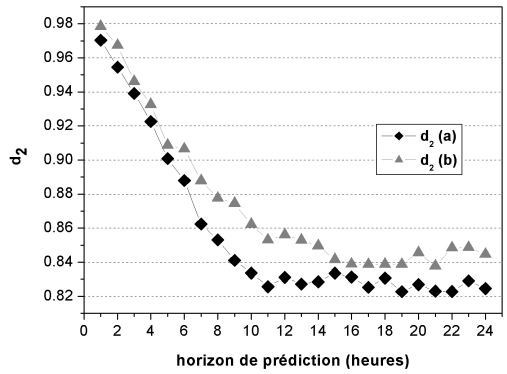
Globalement, les performances obtenues à Prunay sont supérieures à celles d'Aubervilliers. Ceci pourrait être expliqué d'une part, par la base de données plus grande à Prunay et, d'une autre part, par la complexité moindre de l'environment rural par rapport à celui urbain.

Les quatre indices de performance pour la meilleure et respectivement la moins performante simulation effectuées à cette station sont présentés dans la figure 5.8. Le comportement de ces quatre courbes reste le même que dans le cas précédent (la station urbaine).

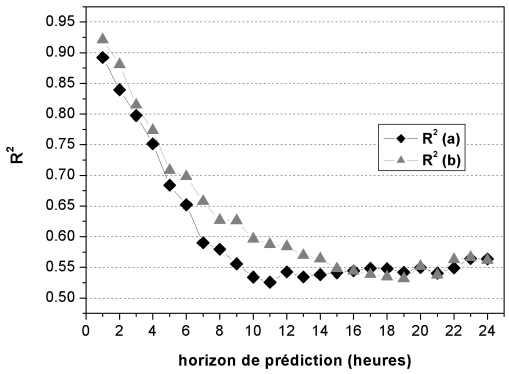
5.7.2.2 Comparaison effectuée à Prunay entre les modèles : 24 PMC, 1 PMC et la persistance

Model	**Moyenne** ($\mu g.m^{-3}$)	**Écart-type** ($\mu g.m^{-3}$)	**RMSE** ($\mu g.m^{-3}$)	**MAE** ($\mu g.m^{-3}$)	**MBE** ($\mu g.m^{-3}$)	$\mathbf{d_2}$	$\mathbf{R^2}$	**SI**
24 PMC	52.96	28.88	17.61	13.63	-0.38	0.88	0.63	0.66
1 PMC	53.17	29.41	17.61	13.38	-0.22	0.88	0.64	0.70
Persistance	53.56	28.90	23.16	17.72	0.58	0.82	0.46	0.51

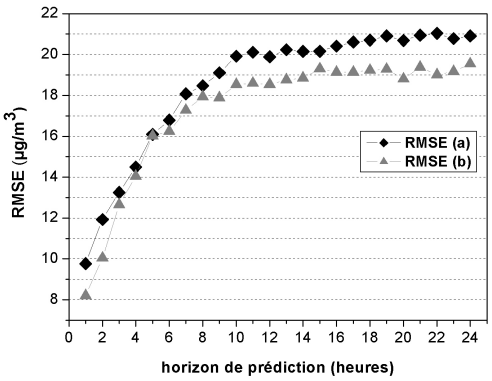
TAB. 5.7: Résultats comparatifs obtenus à Prunay. Tous les indices de performance (RMSE, MAE, MBE, d_2, R^2 et SI) sont calculés sur l'ensemble de test et sur tout l'horizon de prédiction.



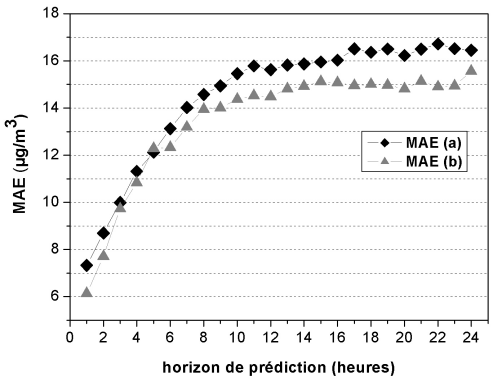
(a) d_2



(b) R^2



(c) RMSE



(d) MAE

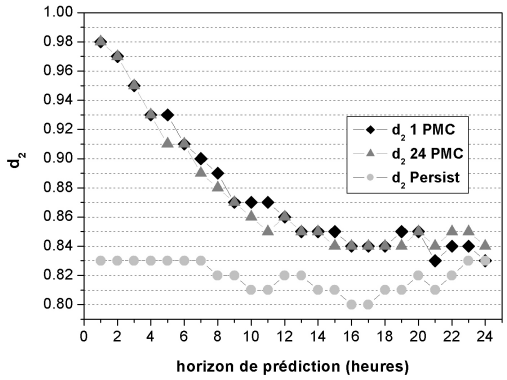
FIG. 5.8: Comparaison des quatre indices de performance (d_2, R^2, RMSE et MAE) sur deux simulations effectuées à Prunay : la moins performante (en noir) et la plus efficace (en gris) sur tout l'horizon de prédiction.

L'évaluation comparative de la performance des deux modèles neuronaux (24 PMC et 1 PMC) et du modèle de persistance est présentée dans le tableau 5.7. Encore une fois, on remarque une très faible amélioration en utilisant le modèle "statique" 1 PMC comparé au modèle "dynamique" 24 PMC, mais une autre, beaucoup plus importante, quand on les compare au modèle de persistance. Les résultats sont détaillés sur tout l'horizon de prédiction dans la figure 5.9. Par ailleurs, les résultats obtenus en utilisant les deux architectures neuronales sont très proches, surtout pour les indicateurs RMSE et MAE.

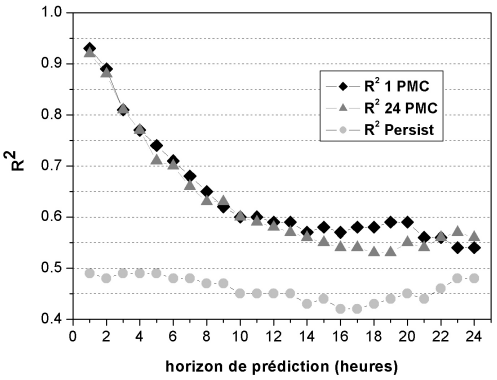
5.7.2.3 Indice pour le dépassement du seuil d'alerte

À Prunay, le nombre des dépassements (concentrations qui dépassent un seuil d'alerte) présents dans notre base de données nous permet de calculer l'indice de succès (voir section 5.6.4). Ceci n'est pas le cas à Aubervilliers où le nombre de dépassements est très faible.

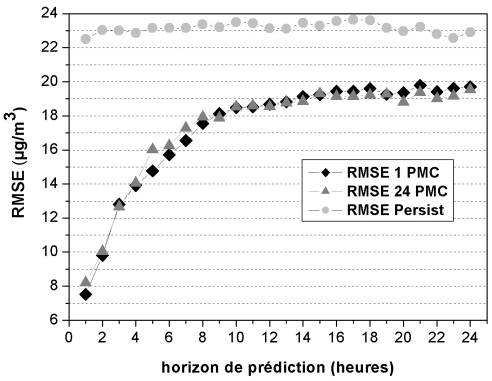
Le seuil d'alerte pour la protection de la santé est de $120\,\mu g.m^{-3}$ en moyenne sur huit heures consécutives, conformément à l'Organisation Mondiale de la Santé (WHO, 2001). En effectuant le



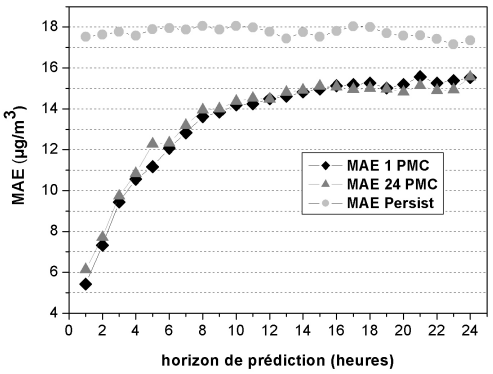
(a) d_2



(b) R^2



(c) RMSE



(d) MAE

FIG. 5.9: Indices de performance obtenus à Prunay pour les trois modèles appliqués (1 PMC, 24 PMC, et la Persistance).

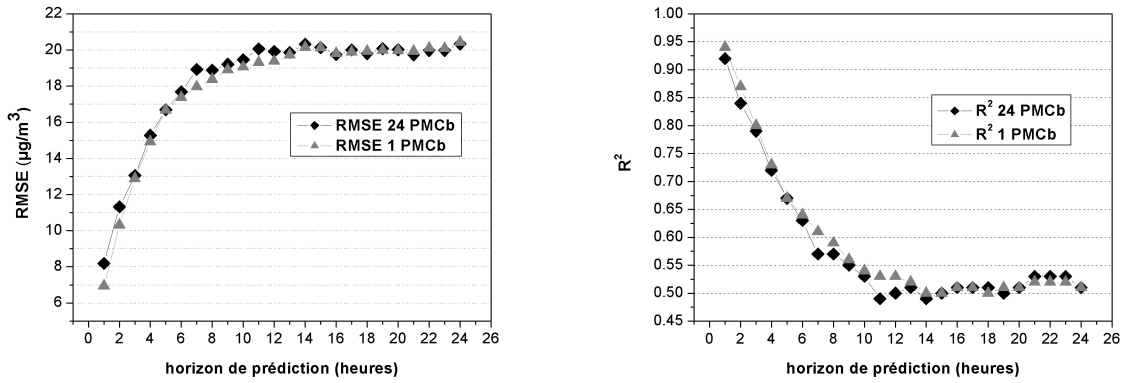
calcul à Prunay, on obtient donc l'indice de succès, présenté sur la dernière colonne du tableau 5.6. Pour les huit simulations effectuées avec l'architecture "dynamique", l'indice de succès SI varie de 0,52 à 0,81, donc, en utilisant ce type de modèle, on peut obtenir des valeurs assez bonnes. Les mêmes résultats sont présentés d'ailleurs pour comparer les trois modèles appliqués dans le tableau 5.7 dernière colonne. On voit encore une fois une très faible supériorité de l'architecture "statique" par rapport à celle "dynamique".

5.7.2.4 Comparaison sur les deux architectures utilisées

La question qu'on se pose maintenant est : peut-on tirer une conclusion après les simulations faites, concernant la nécessité d'utilisation d'une architecture complexe avec un lien récursif entre les prédictions effectuées ? Pour l'instant la réponse n'est pas si évidente. Compte tenu du fait que le nombre d'entrées dans chacun des 24 PMC de ce modèle n'est pas constant, et qu'on n'a pas les mêmes entrées dans le deux modèles, on ne peut pas donner une réponse précise. On se propose de modifier les entrées dans nos deux modèles et de renoncer à l'utilisation des entrées exogènes (météorologiques ou concentrations de dioxyde d'azote) pour mettre les deux modèles sur un pied d'égalité concernant les entrées.

L'architecture "statique", utilisant comme entrées uniquement des concentrations d'ozone, sera appelée par la suite 1 PMCb. La cascade devient une structure répétitive du modèle 1 PMC : dans chacun de 24 PMC on met le même bloc d'entrée de 24 heures de concentrations d'ozone, exempté des entrées météorologiques. Cette architecture sera appelée par la suite 24 PMCb. La seule différence entre les deux architectures est que la cascade utilise les valeurs prédites auparavant.

Les simulations effectuées conduisent aux graphiques présentés dans la figure 5.10. En détaillant sur tout l'horizon de prédiction, on observe un comportement similaire pour les deux indices de performance présentés : le RMSE et le R_2, d'où la conclusion partielle que l'architecture complexe n'est pas vraiment efficace, car elle ne montre pas son utilité. Une possible explication est la grande complexité de cette dernière et le ratio très faible entre le nombre d'échantillons d'apprentissage et le nombre très élevé de paramètres du modèle.



(a) d_2 (b) R^2

FIG. 5.10: Indices de performance comparatifs obtenus à Prunay pour les deux modèles neuronaux appliqués sur une base de données plus étendue

Un deuxième commentaire vise la complexité des structures présentées. Dans le modèle 24 PMCb les données météorologiques ont été ignorées car la structure deviendrait trop complexe ; en diminuant de 3 à 1 le nombre de neurones dans la couche cachée, on peut utiliser en plus les données météorologiques. La comparaison effectuée avec le modèle 1 PMC conduit aux résultats similaires.

5.7.2.5 Résultats obtenus en utilisant les mesures météorologiques comme valeurs prédites

D'autres simulations ont été effectuées à Prunay pour tester l'amélioration des résultats si on utilisait les données météorologiques prédites à l'avance par un service de météorologie pour le même moment que la prédiction. Comme nous ne disposons pas d'une telle prédiction, on utilise à la place, les valeurs mesurées, certes plus précises. Ceci nous permettra d'évaluer un gain maximum apporté par la météorologie. Évidemment, les corrélations croisées entre l'ozone et ses prédicteurs

sont plus importantes pour les données enregistrées à la même heure (voir le tableau 5.8) que pour celles décalées de 24 heures (voir le tableau 5.2). Comme attendu, la prédiction est plus précise dans ce cas que dans celui où on utilisait les données passées. Les indices de performance présentés pour tout l'horizon de prédiction (voir la figure 5.11) montrent des valeurs plus importantes que celles de la figure 5.9. Le gain total obtenu dans ce cas, comparé aux résultats présentés dans la figure 5.9, est d'approximativement $2\,\mu g.m^{-3}$ pour les indicateurs RMSE et MAE et de $0,05$ pour le d_2 et le R^2.

Corrélations croisées	O_3
Température (T)	0.49
Humidité relative (HR)	-0.58
Rayonnement global (SR)	0.57
Durée d'ensoleillement (SD)	0.40
Direction du vent (WD)	-0.23
Vitesse du vent (WS)	-0.34

TAB. 5.8: Corrélations croisées entre ozone et les données météorologiques enregistrées à la même heure à Prunay.

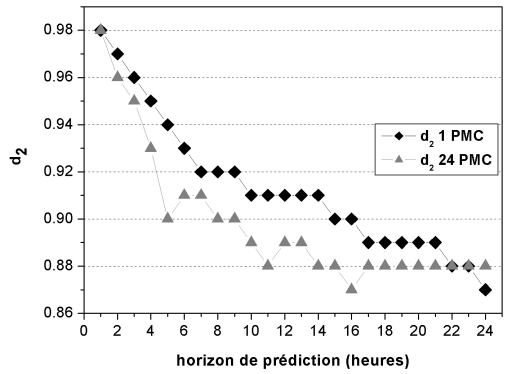
On remarque dans la figure 5.11 que les différences entre les performances des deux modèles appliqués (24 PMC et 1 PMC) sont plus nettes que celles présentées précédemment et cela en faveur de l'architecture "statique".

Malheureusement, en pratique les résultats seront moins précis parce qu'on aurait dû utiliser des vraies prédictions météorologiques et pas de valeurs mesurées. Ces simulations ont été effectuées juste pour trouver une amélioration maximale, mais en pratique cette amélioration ne sera jamais atteinte. Pour une application en temps réel, il serait possible uniquement d'utiliser les prévisions fournies par un service de prévision météorologique.

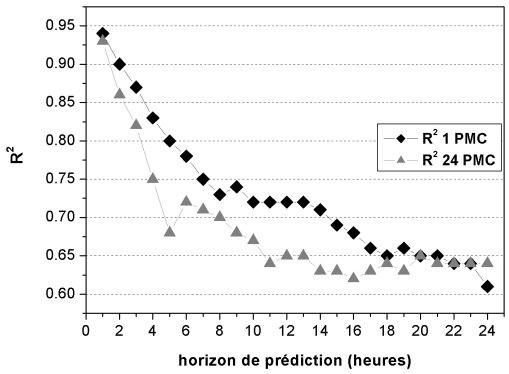
Les résultats présentés montrent que, même en utilisant des données météorologiques mesurées à la place de celles prédites, le gain obtenu est très faible. Il y a deux possibilités : soit on est dans le cas d'une faible dépendance de données météorologiques, soit ces dernières présentent une persistance très forte ; (par conséquent, on ne peut pas obtenir une grande amélioration en les utilisant, car elles n'apportent rien de nouveau par rapport à celles utilisées précédemment.)

5.7.2.6 Résultats sur une base de données plus étendue

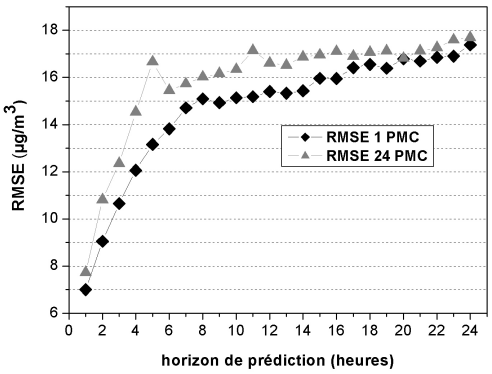
Comme évoqué précédemment dans la section 5.7.2.4, le ratio entre le nombre d'échantillons d'apprentissage et le nombre de paramètres du modèle est très faible. Pour palier à cet inconvénient, on a sélectionné une base de données plus étendue, qui couvre trois ans, de 2000 jusqu'à 2002, à Prunay, et on a appliqué les deux modèles neuronaux 24 PMCb et 1 PMCb (donc sans données météorologiques). L'ensemble de test a été créé uniquement sur les mesures enregistrées pendant la



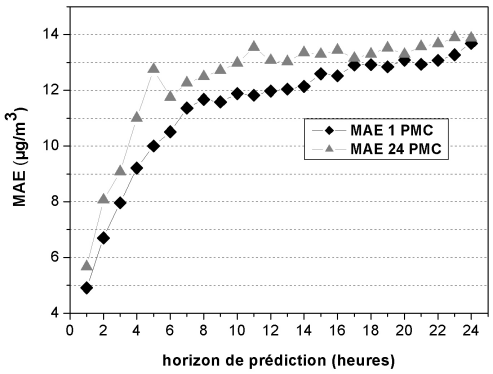
(a) d_2



(b) R^2



(c) RMSE



(d) MAE

FIG. 5.11: Indices de performance comparatifs obtenus à Prunay pour les deux modèles neuronaux appliqués en utilisant des vraies mesures météorologiques à la place des prévisions météorologiques.

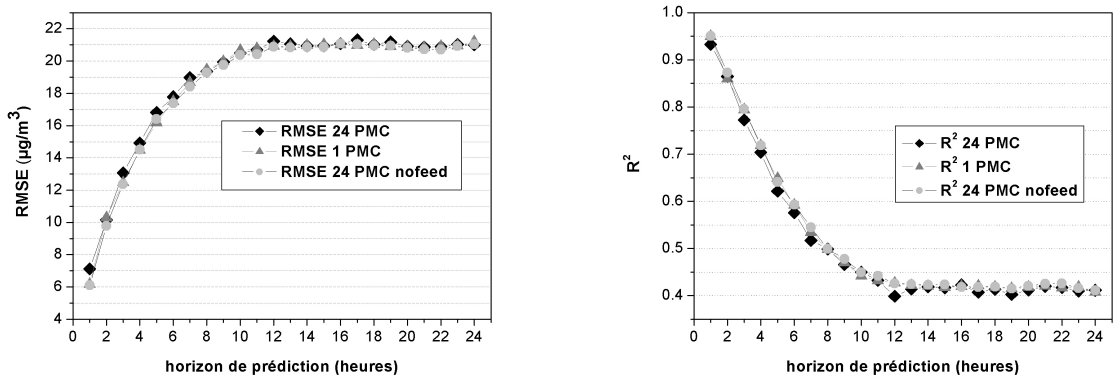
troisième année pour qu'il soit complètement indépendant et ceci pour toutes les simulations qui suivent.

En appliquant les modèles présentés dans la section 5.7.2 et, en plus, le même modèle 24 PMCb dans lequel on a enlevé toutes les connexions entre les divers PMC (24 PMCc), on obtient les résultats présentés dans la figure 5.12 qui montrent une performance quasiment similaire pour tous les modèles appliqués.

Les résultats obtenus sur cette base élargie en utilisant soit une architecture "dynamique", soit une "statique", sont très similaires et montrent qu'il n'y a pas une "meilleure architecture". De plus, encore une fois, on n'obtient pas d'amélioration si on utilise l'architecture plus complexe, "dynamique".

5.7.3 La sensibilité du modèle

Les modèles développés et appliqués dans ce chapitre sont très complexes du point de vue nombre de paramètres, donc il existe la possibilité d'arriver à un sur-apprentissage. Pour tester le



(a) d_2 (b) R^2

FIG. 5.12: Indices de performance comparatifs obtenus à Prunay sur une base de données plus large pour les trois modèles neuronaux appliqués 1 PMC, 24 PMC et 24 PMCc.

pouvoir de généralisation du modèle il faut aussi regarder sa sensibilité par rapport aux perturbations des entrées.

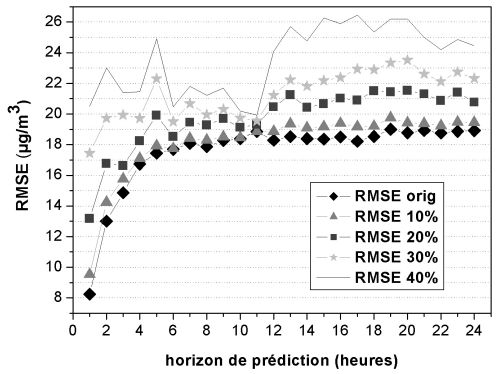
5.7.3.1 Premier test de sensibilité du modèle

En fixant les paramètres du modèle obtenus après l'apprentissage, on a **perturbé les valeurs d'entrée de l'ensemble de test** de 10% jusqu'à 40%. On remarque dans la figure 5.13 que les changements mineurs dans l'entrée conduisent aux réponses peu différentes du modèle. On remarque aussi une dégradation progressive de cette réponse quand les perturbations introduites dans l'ensemble augmentent. Ce test a été effectué sur quatre cas de figure : les modèles initiaux à Aubervilliers et Prunay (sur les bases initiales) et sur la base étendue à Prunay pour les deux architectures mentionnées : 1 PMC et 24 PMC.

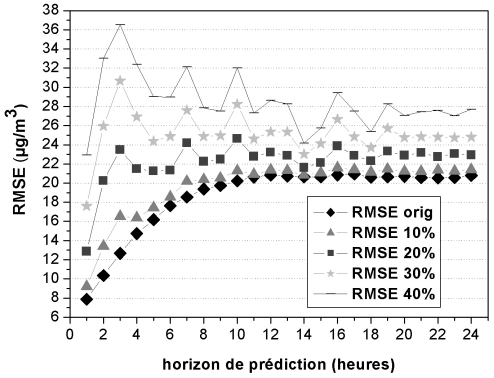
L'effet de chaque perturbation sur l'ensemble de test est différent. Par exemple, une perturbation aléatoire de 10% de l'ensemble de test induit une variation de jusqu'à 30% (pour une valeur de RMSE moyen de 7%) pour le modèle 24 PMC et jusqu'à 40% (pour un RMSE moyen de 5%) pour l'architecture "statique" 1 PMC quand on travaille sur la base de données étendue (voir figure 5.13(b), 5.13(d)). Une autre observation vise le fait que la variation de l'indice RMSE est plus importante pour les premières heures de l'horizon de prédiction, tandis que l'amplitude des variations devient négligeable pour la deuxième partie de l'horizon. Les figures 5.13(a) et 5.13(c) montrent le même comportement face aux perturbations sur les bases initiales à Aubervilliers et respectivement à Prunay.

5.7.3.2 Deuxième test de sensibilité du modèle

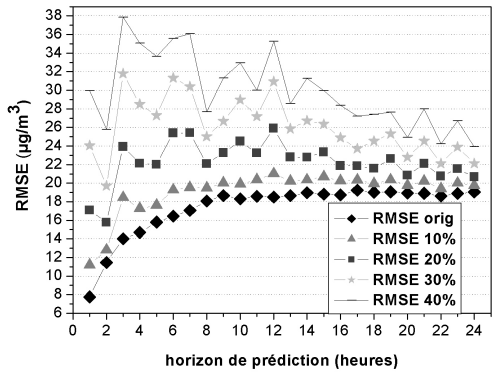
On a effectué un deuxième test en appliquant des **perturbations sur l'ensemble d'apprentissage**. Le modèle obtenu pour chaque ensemble d'apprentissage perturbé est évalué à la fin sur le même ensemble de test non-perturbé. Les résultats présentés dans la figure 5.14 pour les deux



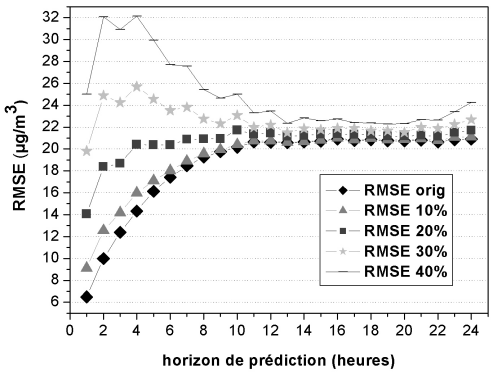
(a) RMSE à Aubervilliers en utilisant le modèle 24 PMC



(b) RMSE à Prunay sur la base étendue en utilisant le modèle 24 PMC



(c) RMSE à Prunay sur la base initiale en utilisant le modèle 24 PMC



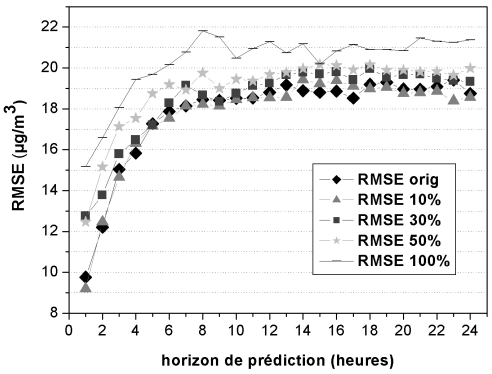
(d) RMSE à Prunay sur la base étendue en utilisant le modèle 1 PMC

FIG. 5.13: Indices de performance sur le même ensemble de test après une perturbation de ce même ensemble de test entre 10% et 40% à Aubervilliers et à Prunay sur les bases initiales, d'une année, en utilisant le modèle 24 PMC (a et c) et à Prunay sur la base de données élargie (3 ans) en utilisant les deux modèles neuronaux 24 PMC et 1 PMC (b et d).

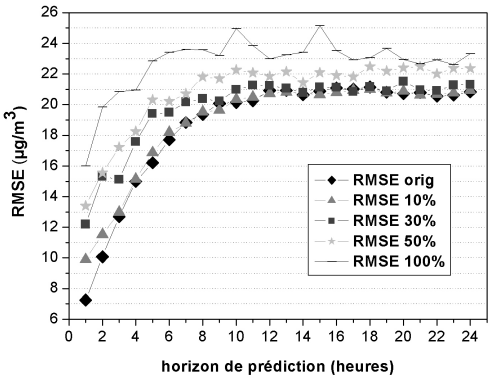
types d'architecture utilisées montrent des variations assez faibles pour des perturbations arrivant à un niveau de 50%.

Globalement, les tests effectués démontrent que, si on applique des petites perturbations aux entrées, les réponses exhibés ne changent pas beaucoup, mais au fur et à mesure que les perturbations augmentent, la qualité des réponses se dégrade et cela peut être considérée comme une bonne aptitude de généraliser de nos modèles. Néanmoins, il faut souligner une anomalie rencontrée dans le cas de l'architecture "statique", quand les réponses du modèle sont plutôt constantes pour la deuxième moitié de l'horizon de prédiction (figure 5.13(d)). Généralement, les résultats et le pouvoir de généraliser sont meilleurs pour la première moitié de l'horizon, plus précisément, pour les premières 8 heures.

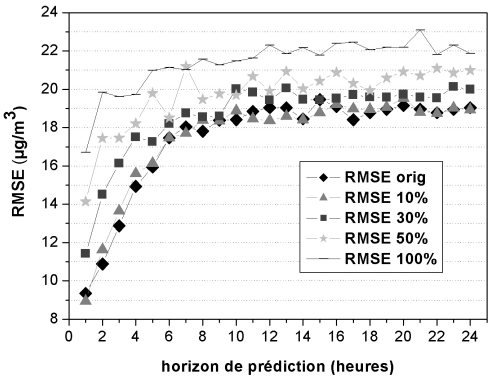
De plus, on a effectué encore 5 simulations, en partant des poids initiaux différents, et les



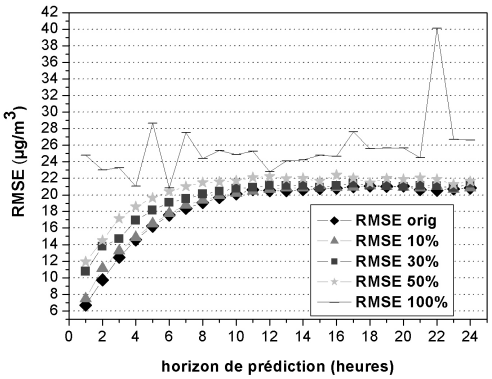
(a) RMSE à Aubervilliers en utilisant le modèle 24 PMC



(b) RMSE à Prunay sur la base étendue en utilisant le modèle 24 PMC



(c) RMSE à Prunay sur la base initiale en utilisant le modèle 24 PMC



(d) RMSE à Prunay sur la base étendue en utilisant le modèle 1 PMC

FIG. 5.14: Indices de performance sur le même ensemble de test originel après la perturbation de l'ensemble d'apprentissage, perturbation qui varie entre 10% et 100% : à Aubervilliers et à Prunay sur les bases initiales en utilisant le modèle 24 PMC (a et c) et à Prunay sur la base de données élargie (3 ans) en utilisant les deux modèles neuronaux 24 PMC et 1 PMC (b et d).

résultats obtenus ont été semblables, soutenant ainsi la stabilité des modèles appliqués.

5.8 Discussions et conclusion de chapitre

Le but de cette étude a été le développement d'un modèle statistique pour la prédiction de l'ozone, basé sur les données publiques fournies par un réseau de surveillance de la qualité de l'air, dans la région parisienne. Utilisé en complément du modèle de prédiction de chimie-transport, appliqué sur une grille avec des cellules de dimensions assez grandes (6×6 km), notre modèle permet d'obtenir localement (sur une station de mesure) une prédiction plus précise. Cette étude a été menée sur Paris et sa grande couronne en utilisant les données enregistrées pendant une année par deux stations de mesure, une urbaine et l'autre rurale.

Comme Paris est une ville qui souffre beaucoup à cause de la pollution, on a décidé de se concentrer non seulement sur les périodes de pic, mais sur toute une année et de prédire les concentrations d'ozone sur un horizon de 24 heures, contrairement aux autres études qui fournissent des prédictions de la valeur maximale diurne uniquement. Cette stratégie nous a permis d'étudier les dépassements du seuil de référence de $120\,\mu g.m^{-3}$ comme moyenne sur huit heures consécutives.

Les études précédentes ont démontré la supériorité des modèles non-linéaires et en particulier les réseaux neuronaux par rapport aux modèles linéaires, mais la question principale soulevée est : quelle serait la "meilleure" architecture dans le but de prédiction de l'ozone sur un horizon de 24 heures ? Est-ce qu'il est nécessaire d'avoir un modèle "dynamique" à architecture complexe ou est-il suffisant d'utiliser un modèle "statique" ? Beaucoup d'études ont été menées pour comparer les performances, révélant ainsi la supériorité des modèles non-linéaires par rapport à ceux linéaires, entre des modèles dynamiques ou entre des modèles statiques. Mais on ne trouve pas beaucoup d'études qui comparent les modèles statiques à ceux dynamiques. D'après nos connaissances, celle-ci serait la première étude qui, tout en utilisant le même type de modèle, le PMC, compare deux architectures, une "statique" contre l'autre "dynamique".

L'architecture "dynamique" est représentée par une cascade de 24 PMC avec un seul neurone en sortie, rangés de façon à ce que chaque sortie d'un PMC soit une entrée du PMC suivant. L'architecture "statique" consiste en un seul PMC avec une couche de sortie constituée de 24 neurones. La performance de ces deux modèles a été testée en calculant six indices de performance.

Les deux modèles neuronaux ont montré des bonnes performances avec une précision correcte pour des concentrations moyennes et élevées de l'ozone. On remarque une certaine amélioration quand on utilise dans le processus de prédiction des données météorologiques, ainsi qu'une décroissance naturelle de la performance sur l'horizon de prédiction. Même si, sur notre base de données, la persistance est moins importante que celle citée par d'autres études, nos résultats sont comparables avec ceux obtenus par les autres. L'indice de succès obtenu à Prunay varie entre 0,52 et 0,81 pour l'architecture "dynamique" et atteint la valeur de 0,70 pour l'architecture "statique".

Si on regarde la comparaison entre les deux modèles, on constate que pour la station urbaine la différence entre les deux performances est légèrement en faveur du modèle "statique". Donc, on peut conclure que ce modèle est plus performant au moins pour les premières huit heures de l'horizon de prédiction. En ce qui concerne la station rurale, les résultats sont semblables. De plus, les deux modèles dépassent largement la performance du troisième, la persistance.

En conclusion, sur notre base de données, les deux architectures "statique" et "dynamique" conduisent à des résultats équivalents ; on peut donc conclure que notre série temporelle d'ozone ne présente pas une non-linéarité dynamique, mais plutôt statique. D'autres chercheurs soutiennent ce même point de vue, mais en utilisant des méthodes différentes. Évidemment, ces conclusions sont issues des résultats obtenus sur une base de données précise et il serait difficile de prétendre leur généralité. Nous rappelons aussi le fait que la prédiction a été faite sur un horizon de 24 heures et que, par conséquent, ces résultats ne peuvent pas être extrapolés à d'autres échelles temporelles.

Il serait intéressant d'appliquer cette démarche à d'autres cas d'étude afin de tester la généralité des résultats obtenus. Mais sur ces deux séries temporelles nos résultats montrent l'absence d'une non-linéarité dynamique dans les séries temporelles d'ozone à l'échelle d'un jour et cette discussion peut être étendue sur l'interpretation physico-chimique de ces résultats.

Les modèles présentés étant très complexes, avec un grand nombre de paramètres, il existe le danger du sur-apprentissage. Pour tester le pouvoir de généraliser, on a étudié la sensibilité des paramètres face aux perturbations des entrées. Les simulations effectuées montrent que, si on applique des petites perturbations aux entrées, les réponses exhibés ne changent pas beaucoup, mais au fur et à mesure que les perturbations augmentent, la qualité des réponses se dégrade et cela peut être considérée comme une bonne aptitude de généraliser des modèles. Néanmoins, il faut souligner le cas particulier rencontré pour l'architecture "statique", quand les réponses du modèle sont plutôt constantes pour la deuxième moitié de l'horizon de prédiction. Généralement, les performances du modèle sont meilleures pour la première moitié de l'horizon, plus précisément, pour les 8 premières heures.

Chapitre 6

Conclusions et Perspectives

6.1 Objectif du travail

L'objectif de ce travail a été d'exploiter le pouvoir informationnel des mesures effectuées par un réseau de surveillance de la qualité de l'air (AIRPARIF) en explorant les différentes facettes de **la modélisation spatio-temporelle** de la pollution atmosphérique urbaine.

Toutes les méthodes et techniques présentées ont été appliquées sur la région d'Île-de-France, l'accent étant mis sur la modélisation inverse, basée essentiellement sur les concentrations mesurées dans la région francilienne, par AIRPARIF, qui fournit des mesures horaires, accessibles sans frais et quasiment en temps réel. La modélisation vise deux polluants, notamment le dioxyde d'azote et l'ozone, et en particulier les forts épisodes de pollution enregistrés pendant l'été 1999 pour ces deux polluants.

Le dioxyde d'azote (NO_2) est principalement produit par l'oxydation du monoxyde d'azote (NO) émis par les véhicules, qui représentent la principale source de pollution dans l'agglomération urbaine. Comme cette oxydation n'est pas instantanée, les concentrations en NO près des axes routiers sont plus élevées que celles de NO_2. En revanche, sur les sites éloignés des voies de circulation, la pollution par dioxyde d'azote est plus forte que celle par NO. Le NO_2 est caractérisé par une courte durée de vie (de quelques secondes à quelques heures) et il reste localisé autour des sources de monoxyde d'azote. En milieu urbain, deux épisodes de pollution par les oxydes d'azote sont généralement observés, un le matin et le deuxième le soir, fortement corrélés avec les variations du trafic routier. Les niveaux les plus forts sont enregistrés le matin et ils restent fortement liés au trafic et à la faible épaisseur de la couche atmosphérique de mélange. Compte tenu de ces informations, la zone susceptible d'être fortement polluée par le dioxyde d'azote reste Paris et sa petite couronne. D'ailleurs, dans la zone rurale, il y a peu de stations qui mesurent le NO_2.

Quant au deuxième polluant mentionné, l'ozone, les réactions chimiques conduisant à sa production sont très complexes. Il est considéré comme un polluant secondaire résultant principalement d'une transformation chimique des oxydes d'azote et de Composés Organiques Volatils (COV). Les pics d'ozone sont fréquents par fort ensoleillement et forte chaleur (en début d'après midi) et restent localisés, la plupart du temps, en dehors des zones urbaines. La durée de vie de l'ozone dépend de son altitude ; elle varie entre quelques secondes et quelques jours ; si l'altitude est assez importante, il

peut être transporté assez loin. Pour étudier cette espèce, la zone rurale de la grande couronne a été dotée de 8 stations automatiques, qui ont été prises en compte dans les diverses étapes de cette étude.

Nous avons abordé la modélisation en considérant les deux variables principales **l'espace** et **le temps** soit *découplées* soit en tant que *continuum spatio-temporel*.

6.2 Bilan de travaux

6.2.1 Interpolation spatiale

Le **premier volet** de cette étude visait d'abord la variable **espace**. Le phénomène naturel examiné, nommé par Matheron (1962) **phénomène régionalisé**, est représenté par un certain nombre de mesures de concentrations de polluants effectuées *au même moment*, sur ce domaine, mesures qui sont par la suite interpolées, pour obtenir des cartes d'isoconcentrations, ainsi que des cartes de variance de l'erreur d'estimation. Parmi les méthodes d'interpolation spatiale décrites, on a choisi d'appliquer **le krigeage**, la seule méthode qui tient compte de la *structure spatiale* des données.

La question principale qu'on s'est posée est : peut-on obtenir une image réaliste de la situation concernant la pollution atmosphérique dans la région d'Île-de-France en utilisant uniquement les mesures enregistrées par les stations automatiques ? À la fin du volet *spatial*, la réponse à cette question était plutôt mitigée ; d'un côté, on reconnaît la facilité de la mise en œuvre d'une *simple* interpolation spatiale (même si, parmi les méthodes d'interpolation spatiale, le krigeage n'est pas la plus simple), mais d'un autre côté, l'interprétation du variogramme expérimental (choix des paramètres : distance maximale, tolérance, pas de calcul), le choix d'un modèle d'ajustement et le type de krigeage sont des décisions assez difficiles à prendre, basées surtout sur l'expérience de l'utilisateur. Il faut remarquer également que le manque d'une couverture spatiale correcte par des stations de mesure, est un facteur limitatif dans le processus d'estimation, le deuxième étant la qualité des mesures prises en compte dans le krigeage.

Pour les deux polluants analysés, le dioxyde d'azote et l'ozone, on a appliqué sur des domaines différents (Paris et petite couronne pour la première espèce et la grande couronne pour la deuxième), trois variantes de krigeage : Ordinaire (KO), Universel (KU) et Intrinsèque Généralisé (KI) pour trois scénarios différents : deux épisodes de forte pollution et un de faible pollution, et on a comparé les résultats.

Pour **le dioxyde d'azote**, dans les deux cas présentant des concentrations élevées, l'ajustement du variogramme est acceptable et on constate que les cartes de concentrations obtenues par les trois types de krigeage sont très semblables ; les cartes de l'écart-type de l'erreur associées sont elles aussi proches, avec des valeurs un peu plus faibles pour le krigeage intrinsèque généralisé. En revanche, pour le cas de faible pollution, la situation est délicate à cause d'un mauvais ajustement d'un modèle théorique au variogramme expérimental. Ne pouvant pas identifier la structure spatiale des données par un variogramme, on est obligé de recourir à la méthode de covariance généralisée

et au KI, pour pouvoir tenir compte de cette structure dans l'estimation. Les statistiques moyennes effectuées sur les tests "Leave-One-Out" de validation croisée, montrent qu'aucune des trois méthodes n'est privilégiée par rapport aux autres. Elles indiquent une erreur d'estimation de l'ordre de $10\,\mu g/m^3$ en moyenne pour les stations à l'intérieur du domaine, et plus importante aux bords ($30 - 40\,\mu g/m^3$). L'estimation peut être jugée comme satisfaisante.

Alors que dans le cas du NO_2, on n'a pas pu ajuster un variogramme dans le cas d'une faible variabilité spatiale, dans le cas de l'ozone, on n'arrive pas à trouver une structure spatiale (ajuster un variogramme cohérent), dans le cas d'une forte variabilité. Le manque de structure peut être donc rencontré dans plusieurs situations différentes. Bien que l'aspect des cartes n'est pas très convaincant, les tests "Leave-One-Out" ont conduit à des erreurs de l'ordre $12 - 15\,\mu g/m^3$, ce qui peut être jugé comme satisfaisant. Comme c'était le cas pour le dioxyde d'azote, on remarque aussi que, pour l'ozone, la représentativité des stations situées en bord du domaine est importante, tandis que, dans l'agglomération, cette représentativité diminue. Bien que, dans le cas du NO_2, le KI donnait toujours des résultats convenables (sans être forcément les meilleurs), ceci n'est plus le cas pour la configuration des stations d'ozone. On peut penser que cette méthode se comporte moins bien que les autres en situation d'extrapolation, pouvant générer des artefacts.

La conclusion de ce premier volet était que, pour le dioxyde d'azote, l'estimation obtenue est plutôt correcte, car la couverture spatiale du domaine par des stations automatiques est bonne, tandis que, pour l'ozone, en analysant les deux épisodes de forte pollution présentés, on s'aperçoit que le nombre de mesures n'est pas suffisant pour pouvoir reconstituer la forme du panache. Étendre la zone d'étude ne servira à rien, car il n'existe pas de mesures en dehors de celles utilisées ici et on serait forcé à faire encore de l'extrapolation. L'analyse effectuée conduit donc à la conclusion que le nombre de stations disponibles et leur répartition spatiale sont plutôt insuffisants pour obtenir des cartes de pollution d'ozone convenables.

6.2.2 Interpolation spatio-temporelle

Le **deuxième volet** de cette étude rajoute à **l'espace** la deuxième variable : **le temps**. Deux décisions de modélisation sont possibles :

- on peut considérer les deux variables *découplées* et construire un modèle de corrélation spatio-temporelle séparable (une simplification de la réalité physique) ;
- on peut considérer que les deux variables forment un *continuum spatio-temporel*, donc on essaiera de reconstruire la structure de corrélation spatio-temporelle qui caractérise le phénomène analysé.

En choisissant la deuxième voie, on a appliqué une méthode très simple, le krigeage intrinsèque spatio-temporel (KIS/T) sur des séquences très courtes de données horaires (des concentrations de polluants atmosphériques) de dioxyde d'azote et d'ozone, mesurées par les stations qui ont été utilisées dans le cadre spatial, en ayant comme objectif déclaré l'amélioration, si possible, de la représentation spatiale des champs de polluants atmosphériques. En théorie, le fait d'incorporer des mesures enregistrées à différents moments de temps, autres que celui de l'estimation, peut apporter

une certaine quantité d'information, qui pourrait rendre plus robuste l'estimation au moment étudié.

Compte tenu du nombre et de la qualité des données dont nous disposions, les résultats obtenus ne sont pas à la hauteur de ce que les premisses théoriques laissaient présager. Comme observation générale concernant la méthode appliquée, on peut constater que la structure de corrélation spatio-temporelle des données incluses dans l'étude est peut être trop complexe pour qu'une covariance généralisée spatio-temporelle de type polynômial, comme celle utilisée par le KIS/T, puisse la reproduire correctement. Comme limitation dans la procédure appliquée, il faut rappeler que les ordres de continuité n'ont pas été estimés, mais fixés à zéro, ce qui correspond à l'hypothèse que les incréments spatio-temporels sont homogènes/stationnaires. De plus, comme on l'avait déjà remarqué dans le cas spatial, la répartition spatiale des stations mesurant l'ozone est creuse et cela influence d'une manière négative l'estimation. Ces résultats peu satisfaisants peuvent être causés également par l'échelle temporelle choisie : il est possible que pour des données plus lisses, du genre moyennes journalières, la méthode s'avère plus efficace, mais pour les épisodes précis étudiés, cela n'a pas été le cas.

Pour les deux polluants étudiés, le dioxyde d'azote et l'ozone, on a présenté pour chacun un épisode de forte pollution. Ce qu'on peut remarquer pour les deux cas analysés est que, généralement, le fait d'incorporer des mesures enregistrées à différents moments n'améliore pas d'une manière significative l'estimation spatiale déjà présentée. Néanmoins, il existe des stations où les tests de validation croisée ont mis en avant des meilleurs résultats avec l'estimation spatio-temporelle. Étant donnée la complexité des calculs, l'utilisation de cette méthode n'est pas justifiée en terme de gain de précision par rapport à celle purement spatiale. Elle a pourtant le mérite d'appuyer les résultats obtenus avec la méthodologie spatiale, car les champs estimés dans les deux cas sont assez similaires, et avec une variance d'estimation moindre dans le cas spatio-temporel. Indirectement, on montre la robustesse des résultats obtenus dans la méthodologie spatiale. Ces remarques sont valables et utiles essentiellement pour le cas du NO_2, où l'interpolation spatiale s'était avérée plus pertinente que pour l'ozone, ceci étant dû en majeure partie à la bonne couverture spatiale de la région par des stations de mesure.

Pour conclure, la géostatistique spatio-temporelle monovariable ne s'est pas avèré en mesure de décrire adéquatement la continuité spatio-temporelle des processus qu'on a étudiés. Pour cela, des modèles de transport physico-chimiques sont en général plus performants, mais nécessitent beaucoup plus d'information en entrée.

6.2.3 Assimilation de données

Dans le **troisième volet**, toujours dans le même contexte, celui de l'utilisation du pouvoir informationnel des mesures et du concept de *continuum spatio-temporel* on décide de corriger les sorties d'un modèle déterministe par les mesures du réseau de surveillance ; ceci revient donc à utiliser conjointement d'une part, les connaissances sur la physique et la chimie de l'atmosphère, et d'autre part, les mesures disponibles de concentrations de polluants atmosphériques, dans le but d'améliorer la représentation spatiale des champs de concentrations des polluants. Pour cela, on a

implémenté, sur la version régionale d'un modèle de chimie-transport, appelé CHIMERE, une méthode séquentielle d'assimilation de données, notamment le Filtre de Kalman d'Ensemble, choisie principalement pour sa facilité d'implémentation. On garde toujours la même zone d'étude, la région francilienne.

L'idée principale de cet algorithme est d'utiliser un ensemble d'estimations d'état (obtenues en perturbant le vecteur d'état), à la place d'une seule, pour représenter la densité de probabilité du vecteur d'état du système et de calculer la matrice de covariance de l'erreur sur cet ensemble. On obtient, de cette façon, un **ensemble d'états** du modèle qui évoluent dans l'espace de l'état. La moyenne de cet ensemble représente la *meilleure* estimation, et la variance d'ensemble, la variance de l'erreur d'estimation.

Les données utilisées pour corriger les champs de polluants produits par le modèle sont des mesures horaires d'**ozone** (la seule espèce assimilée), sur 18 stations situées à Paris et sur la grande couronne. Parmi ces 18 stations, on a utilisé 10, pour assimiler l'ozone (9 étant rurales et une urbaine) et les 8 autres ont été gardées comme stations-témoin. La période d'étude choisie est la deuxième décade du mois de juillet 1999.

Globalement, on peut dire que, par assimilation des données, les différences entre les observations et les concentrations analysées diminuent beaucoup aux sites de mesure utilisées pour corriger les champs d'ozone, tandis que les corrections effectuées sur les stations-témoin sont moins importantes.

En outre, la sensibilité du système d'assimilation par rapport à ses paramètres a été testée : le nombre de membres d'ensemble, la longueur de décorrélation qui caractérise la fonction de covariance gaussienne utilisée pour créer la perturbation, ainsi que la variance introduite dans le système par la perturbation. Pour résumer, sur les statistiques ponctuelles effectuées aux sites de mesure, le gain obtenu en passant de 10 à 20 membres d'ensemble est significatif, par contre, la variation ultérieure est très lente. Il apparaît que 20 ou 40 membres d'ensemble reproduisent bien l'espace des erreurs ; en revanche, les champs 2D produits présentent souvent des explosions numériques sur les bords et sur les zones qui ne sont pas couvertes par les observations. En augmentant la taille de l'ensemble jusqu'à 80 membres, on arrive à obtenir de meilleures représentations, même si parfois, comme évoqué précédemment pour des cas précis, les gradients introduits dans le champ d'ozone par l'intermédiaire des perturbations, conduisent à des mauvaises représentations. Le système d'assimilation n'apparaît pas très sensible aux deux autres paramètres mentionnés.

L'impact de l'assimilation séquentielle des mesures surfaciques d'ozone sur les champs de concentrations de dioxyde d'azote a été également vérifié. En utilisant uniquement les mesures d'ozone, on ne s'attend pas à obtenir des améliorations spectaculaires sur le NO_2. En effet, en termes de statistiques RMSE moyenne, sur les deux groupes de stations on constate qu'on n'obtient pas de changements majeurs dans les concentrations de NO_2. Malheureusement, les cas particuliers d'épisodes de pollution analysés montrent que, pour obtenir une meilleure représentation spatiale du NO_2, ce système d'assimilation n'est pas suffisant.

6.2.4 Prédiction de l'ozone par une approche neuronale

Le quatrième et dernier **volet** de ce travail vise la prévision locale, où uniquement la variable **temps** intervient. L'objectif de cette étude a été le développement d'un modèle **statistique** pour la prédiction de l'ozone, sur un horizon de 24 heures, basé sur les données fournies par AIRPARIF. Utilisé en complément du modèle de prédiction de chimie-transport, appliqué sur une grille avec des cellules de dimensions assez grandes ($6 \times 6\,\mathrm{km}$), notre modèle permet d'obtenir localement (sur une station de mesure) une prédiction plus précise. Cette étude a été menée toujours sur Paris et sa grande couronne en utilisant les données enregistrées pendant une année par deux stations de mesure, une urbaine (Aubervilliers) et l'autre rurale (Prunay).

On a décidé d'appliquer deux architectures neuronales différentes, une "dynamique" et l'autre "statique", pour comparer leur pouvoir prédictif. Dans les deux cas, on utilise en entrée les mesures des 24 heures précédentes, la différence entre les deux architectures provenant du fait que dans le cas de la structure "dynamique", la prévision, à chaque heure, fournie par le modèle est utilisée en entrée pour la prédiction à l'heure suivante. Ainsi, l'architecture "dynamique" est représentée par une cascade de 24 PMC (perceptron multi couches) avec un seul neurone en sortie, rangés de façon à ce que chaque sortie d'un PMC soit une entrée du PMC suivant. L'architecture "statique" consiste en un seul PMC avec une couche de sortie constituée de 24 neurones. La performance de ces deux modèles a été testée en calculant six indices de performance.

Les deux modèles neuronaux ont montré de bonnes performances, avec une précision correcte pour des concentrations moyennes et élevées de l'ozone. On remarque une certaine amélioration quand on utilise dans le processus de prédiction des données météorologiques, ainsi qu'une décroissance naturelle de la performance sur l'horizon de prédiction. Même si, sur notre base de données, la persistance est plus faible que celle citée par d'autres études, nos résultats sont comparables avec ceux obtenus par les autres. L'indice de succès obtenu à Prunay varie entre 0,52 et 0,81 pour l'architecture "dynamique" et atteint la valeur de 0,70 pour l'architecture "statique".

En conclusion, sur notre base de données, les deux architectures, "statique" et "dynamique", conduisent à des résultats équivalents ; on peut donc conclure que notre série temporelle d'ozone ne présente pas une non-linéarité dynamique, mais plutôt statique. D'autres chercheurs soutiennent ce même point de vue, mais en utilisant des méthodes différentes. Évidemment, ces conclusions sont issues des résultats obtenus sur une base de données précise et il serait difficile de prétendre leur généralité. Nous rappelons aussi le fait que la prédiction a été faite sur un horizon de 24 heures et que, par conséquent, ces résultats ne peuvent pas être extrapolés à d'autres échelles temporelles. Il serait intéressant d'appliquer cette démarche à d'autres cas d'étude afin de tester la généralité des résultats obtenus. Mais sur ces deux séries temporelles nos résultats montrent l'absence d'une non-linéarité dynamique dans les séries temporelles d'ozone à l'échelle d'un jour.

Généralement, les performances du modèle sont meilleures pour la première moitié de l'horizon, plus précisément, pour les 8 premières heures, ce qui doit être probablement en liaison avec

les transformations physico-chimiques subies par l'ozone sur les sites d'étude. Un autre point qui mérite d'être remarqué est le fait que nous n'avons pas étudié uniquement la prédiction des maxima d'ozone, mais de toutes les valeurs, dans le contexte de la moyenne glissante sur 8 heures, intervenant dans le standard de santé publique.

6.3 Conclusions générales

L'ensemble des résultats obtenus au cours de notre étude montre que la **cartographie** des champs de polluants atmosphériques sur la région d'Île-de-France est encore un sujet sensible. Malgré la diversité des méthodes appliquées, on ne peut pas tirer une conclusion en faveur d'une seule méthode, car chacune a ses avantages et ses inconvénients. Par exemple, la cartographie basée uniquement sur l'interpolation spatiale est méthodologiquement simple et ne nécessite pas une grande diversité de données. Cependant, elle repose entièrement sur la qualité de ces données et sur la couverture spatiale du domaine par des stations de mesure. On a vu que pour le dioxyde d'azote, analysé sur un domaine restreint : Paris et sa petite couronne, les résultats semblent plutôt satisfaisants. En revanche, pour l'ozone, la situation est plus complexe, car le nombre de stations et leur emplacement (sur la grande couronne) ne nous permettent pas de reconstituer la forme du panache apparu pendant les épisodes de forte pollution enregistrés en juillet 1999.

L'interpolation spatio-temporelle n'apportant pas une amélioration significative, on a donc décidé de recourir à la modélisation numérique par des modèles de chimie-transport, modélisation qui pourra nous aider à obtenir une estimation de l'état de la pollution hors des zones couvertes par les mesures (extrapolation). Les modèles déterministes tridimensionnels de simulation de la qualité de l'air sont mieux adaptés à la compréhension de la pollution atmosphérique que les modèles statistiques, mais ils s'avèrent lourds au niveau du temps de calcul et demandent beaucoup d'information en entrée difficile à obtenir. En outre, les simulations ont mis en évidence une difficulté : le transport d'erreurs de prévision. Cependant, une façon d'éviter la propagation de l'erreur est l'assimilation de données qui corrige à chaque pas de temps le modèle pour qu'il s'approche des observations (les seules mesures utilisées ont été celles d'ozone). De cette manière, on revient au problème initial. On veut corriger les sorties du modèle déterministe par les mesures disponibles, mais si leur nombre n'est pas assez élevé et leur qualité est faible, la reconstruction du champ d'ozone devient problématique. Toutefois, on peut remarquer que cette dernière méthode, l'assimilation de données, est plus appropriée que les autres, décrites précédemment, dans le cas de cette espèce. En revanche, le filtre d'ensemble, dans l'implémentation actuelle, ne peut pas améliorer l'estimation de champs de NO_2. En outre, la mauvaise modulation temporelle des émissions est en grande partie responsable de la faible qualité des estimations de dioxyde d'azote. Par conséquent, pour améliorer la représentation spatiale pour cette espèce, il faut combiner les efforts sur deux plans : d'une part, la modélisation par l'intermédiaire d'une meilleure prise en compte des variations des émissions et du processus de titration, et d'autre part, l'assimilation des mesures de NO_2 et en plus la perturbation des émissions, l'un des paramètres les plus incertains du modèle.

En ce qui concerne l'autre objectif proposé, la **prévision** sur un horizon de 24 heures des

concentrations d'ozone, les résultats obtenus en appliquant deux types d'architectures neuronales, une très simple, constituée d'un seul PMC (perceptron multi couches) avec une sortie de 24 neurones, et l'autre plus complexe, constituée d'une cascade de 24 PMC, montrent que, localement (sur deux stations de mesure), on peut obtenir des prévisions d'une bonne qualité ; généralement, on constate que sur les 8 premières heures de l'horizon de prédiction, les six indices de performance calculés indiquent une précision correcte pour des concentrations moyennes et élevées d'ozone. Sur notre base de données, les deux architectures mentionnées conduisent à des résultats équivalents ; cela nous a conduit à la conclusion que notre série temporelle d'ozone ne présente pas une non-linéarité dynamique, mais plutôt statique.

Globalement, ce travail met en valeur **le pouvoir informationnel des mesures de pollution**, dont l'utilisation **indépendante** ou **conjointement avec des modèles déterministes**, contribue, grâce aux techniques de modélisation inverse, à une meilleure connaissance de la pollution atmosphérique à l'échelle régionale.

6.4 Perspectives

Les perspectives de ce travail sont nombreuses car chaque partie de l'étude, à part les réponses qu'elle donne, soulève encore d'autres questions. Ces questions ayant été détaillées dans la présentation de chaque chapitre, nous n'allons pas y revenir ici, on se concentrera sur la principale direction dans laquelle ce travail pourrait évoluer. Elle vise essentiellement l'assimilation de données sur des modèles de chimie-transport, dans un but d'analyse ou de prévision de plusieurs espèces, dont les principales visées sont l'ozone et les oxydes d'azotes.

6.4.1 Échelle régionale

Dans un premier temps, le système actuel implémenté à l'échelle régionale, basé sur le modèle CHIMERE et les observations au sol, mesurées par AIRPARIF pourrait être amélioré :

- par l'assimilation des autres espèces chimiques présentes dans le modèle (notamment les oxydes d'azote et le CO) ;
- en perturbant des paramètres incertains du modèle (émissions, conditions aux limites,...) ;
- en estimant des paramètres du biais corrélés dans le temps (estimation et correction on-line du biais par l'augmentation du vecteur d'état - Dee et Da Silva (1998)).

L'assimilation de données a été appliquée dans cette étude à des problèmes de filtrage, mais le système sera facilement transposable au lissage, qui offre un cadre plus général et qui permettra d'améliorer encore les représentations spatiales des champs de polluants atmosphériques.

6.4.2 Échelle continentale

Un système d'assimilation, similaire à celui présenté dans cette étude, pourra être développé sur la version continentale de CHIMERE (et amélioré). Ce système pourra utiliser les mesures au sol (d'ozone) fournies par les réseaux européens de surveillance de la qualité de l'air.

À part les mesures au sol (les observations conventionnelles), il y a une grande ouverture pour l'assimilation de données satellitaires qui fournissent des informations exhaustives spatialement (radiances), ainsi que des profils verticaux pour différentes espèces, notamment pour l'ozone. Dans le cas de cette espèce, la majorité de sondeurs spatiaux étaient dédiés à l'étude de l'ozone stratosphérique. Néanmoins, depuis quelques années, l'intérêt s'est concentré sur les instruments avec une sensibilité suffisante pour mesurer les concentrations d'ozone dans la troposphère. Certains instruments (GOME, SCIAMACHY, IMG) étaient déjà capables d'observer les colonnes d'ozone troposphérique, mais le plus grand progrès a été atteint par les derniers instruments mis en orbite (IASI, TES). Leur sensibilité aux concentrations d'ozone de surface a été améliorée, ainsi que la fréquence temporelle de passage (environ 2 passages par jour). Tous ces instruments délivrent des radiances dans différentes longueurs d'ondes correspondant à la bande d'absorption de différents polluants. Le calcul des profils de concentrations à partir de ces informations est réalisé par modélisation inverse.

La première question qu'on peut se poser est quel type de données faut-il assimiler ? Des données brutes (radiances) ou des données inversées (colonnes, profils verticaux) ? La deuxième question est sur quels paramètres du modèle ces données satellitaires renseignent le plus (les émissions, la météo, les conditions aux limites) ?

Si on veut effectuer une assimilation de données brutes on est confronté au problème du choix du meilleur modèle radiatif. On peut recourir soit à un modèle couplé CTM+radiatif, soit à un modèle radiatif considéré comme opérateur d'observation. Dans les deux cas, on profite du fait qu'on n'a pas besoin de faire l'inversion, par contre le coût de calcul est assez élevé. Il y a quand même un petit inconvénient pour le modèle couplé, car l'évaluation de la radiance là où/quand il n'y pas de mesure est inutile.

Si on veut effectuer une assimilation des données inversées, la première étape consiste en obtenir des profils des concentrations d'ozone en utilisant, par exemple, un algorithme du genre Ozone ProfilE Retrieval Algorithm (OPERA, Van Oss (2002)), et ensuite le point encore délicat de l'assimilation sera la construction de l'opérateur d'observation qui tiendra compte de ce profil vertical. L'inconvénient principal de cette méthode et le manque de contrôle sur les erreurs d'inversion quand on veut obtenir les profils verticaux.

Bien qu'on a déjà souligné le fait qu'il existe des points forts et des faiblesses dans cette petite incursion parmi les données satellitaires, il est important d'insister sur l'énorme potentiel contenu dans ces données, qui reste encore à exploiter, pour ainsi contribuer à l'amélioration de nos connaissances sur les processus qui se déroulent dans l'atmosphère.

Annexe A

Unités de mesure de la concentration d'un polluant atmosphérique gazeux

Une concentration représente une quantité de matière (ou sa masse) exprimée pour un volume donné et divisée par ce volume (c=m/v). On utilise deux types d'unités pour quantifier les espèces chimiques dans l'air :

• **Abondance de l'espèce** : Les [ppb] (parties par billion) ou [ppm] (parties par millions, avec 1ppm=1 000ppb) ne sont pas des concentrations mais un rapport de mélange (*mixing ratio*), correspondant aux nombres de moles de l'espèce par rapport au nombre de moles du volume d'air considéré.

• **Concentration de l'espèce** : Les [μg.m^{-3}] représentent la masse de l'espèce considérée par rapport au volume d'air considéré.

On convertit une unité vers l'autre par la relation :

$$C_{\mu gm^{-3}} = C_{ppb}\frac{R^*T}{p \cdot M} \tag{A.1}$$

où T représente la temperature absolue en Kelvin, $R^* = 0,08314\,\text{hPa.m}^3\text{.K}^{-1}\text{.mol}^{-1}$ est la constante universelle des gaz parfaits, p signifie la pression (en hPa) et M la masse molaire de l'espèce. Pour les deux polluants analysés dans cette étude on utilise les approximations suivantes pour convertir l'unité [ppb] vers l'autre :

- NO_2 : $C_{\mu gm^{-3}} = C_{ppb} \cdot 1,90$

- O_3 : $C_{\mu gm^{-3}} = C_{ppb} \cdot 1,99$

Annexe B

Indices de performance

Au cours de ce travail on est amené plusieurs fois à calculer des diverses statistiques sur les résultats obtenus. Les plus connus parmi les indices de performance calculés habituellement sont présentés dans la suite :

• **Le coefficient de corrélation (R) :**

$$R = \sqrt{\frac{\sum_{i=1}^{n}(O_i - \overline{O})^2 - \sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}} \tag{B.1}$$

• **La racine de la moyenne de l'erreur quadratique-en anglais root mean squared error (RMSE) :**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{B.2}$$

• **L'erreur absolue moyenne (MAE) :**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \tag{B.3}$$

• **Le biais moyen (MBE) :**

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{B.4}$$

où P_i et O_i représentent les variables prédites et respectivement observées, tandis que $\overline{O}$ signifie la moyenne des observations.

Liste des figures

Liste des tableaux

Bibliographie

Abdul-Wahab, S.A., Al-Alawi, S.M., (2002). *Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks*, Environmental Modelling & Software 17, 219-228. 193

Agirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., (2006). *Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area*, Environmental Modelling & Software 21, 430-446. 194

AIRPARIF, (2007). *Les normes en vigueur*, www.airparif.asso.fr/airparif/p-nor.htm. 17, 19, 20, 239

Akima, H., (1978). *A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points*, ACM Transactions on Mathematical Software, **4**, No. 2, 148-159. 30

Anderson, J.L., (2001). *An ensemble adjustement Kalman filter for data assimilation*, Mon. Weather Rev., **129**, 2884-2903. 147, 149, 152

Arnaud, M., Emery, X., (2000). *Estimation et interpolation spatiale*, Hermes Science Publications, Paris. 30, 31, 38, 41, 64

Atmosphère, (2008). *http ://www.ffme.fr/technique/meteorologie/theorie/atmosphere/composition.htm*, site internet. 6, 233

Balaguer Ballester, E., Camps i Valls, G., Carrasco-Rodriguez, J.L., Soria Olivas, E., del Valle-Tascon, S., (2002). *Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks*, Ecological Modelling 156, 27-41. 193

Beekmann, M., Derognat, C., (2003). *Monte Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the atmospheric Pollution Over the Paris area (ESQUIF) campaign*, J. Geophys. Res., **108**, D17, 8559, doi :10.1029/2003JD003391. 156

Berkowitz, B., (1985). *A spatial, time-dependent approach to estimation of hydrologic data*, J. Hydrology, **135**, 133-142. 100

Bertino, L., (2001). *Assimilation de données pour la prédiction de paramètres hydrodynamiques et écologiques : cas de la lagune de l'Oder*, Thèse de doctorat, Ecole de Mines de Paris, France. 132, 144, 158

Bilonick, R.A., (1983). *Risk-qualified maps of hydrogen ion concentration for the New York state area for 1966-1987*, Atmospheric Environment, **17**, No. 12, 2513-2524. 100

Bilonick, R.A., (1985). *The space-time distribution of sulfate deposition in the northeastern U.S.A*, Atmospheric Environment, **19**, No. 11, 1829-1845. 98, 101

Bishop, C., Etherton, B., Majumdar, S., (2001). *Adaptive sampling with the ensemble transform kalman filter Part I : Theoretical aspects*, Mon. Weather Rev., **129**, 420-436. 152

Bishop, C.M., (1995). *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford. 193, 200, 201

Blond, N., (2002). *Assimilation de données photochimiques et prévision de la pollution troposphérique*, Thèse de doctorat, Ecole polytechnique, Paris, France. 153, 155, 156

Bocquet, M., (2004). *Introduction aux principes et méthodes de l'assimilation de données en géophysique*, Notes de cours de l'ecole Nationale Supérieure des Techniques Avancées. 138, 140, 235

Bogaert, P. et Christakos, G., (1997). *Spatiotemporal analysis and processing of thermomentric data over Belgium*, J. Geophysical Res., **102**, No. D22, 25831-25846. 98

Bouttier, F., Courtier, P., (1999). *Data assimilation concepts and methods*, Meteorological Training Course Lecture Series, ECMWF European Center for Medium-range Weather Forecasts, Reading, UK., **53**. 134, 135, 138

Box, G.E.P., Cox, D.R., (1964). *An analysis of transformations*, Journal of the Royal Statistical Society, Series **B26**, 211-246. 63

Brasseur, P., Ballabrera-Poy, J., Verron, J., (1999). *Assimilation of altimetric data in the mid-latitude oceans using the Singular Evolutive Extended Kalman Filter with an eddy-resolving, primitive equation model*, Journal of Marine Systems, **22**, 269-294. 158

Brockwell, P.J. et Davis, R.A., (1987). *Time Series : Theory and Methods*, Springer Verlag, New York Inc. 98

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., Vitabile, S., (2007). *Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy*, Atmospheric Environment 41, 2967-2995. 194

Burgers, G., van Loon, P. J., Evensen, G., (1998). *On the analysis scheme in the ensemble kalman filter*, Mon. Weather Rev., **126**, 1719-1724. 145, 147, 148

Chauvet, P., (1999). *Aide mémoire de la géostatistique linéaire*, Cahiers de Géostatistique, Fascicule 2, Ecole Nationale Supérieure des Mines de Paris, Centre de Géostatistique, Fontainebleau. 28, 36, 46, 52, 233

Chilès, J.P., Delfiner, P., (1999). *Geostatistics Modeling Spatial Uncertainty*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York. 39, 41, 54, 64

Christakos, G., (1984). *On the problem of permissible covariance and variogram models*, Water Resources Research, **20**, N.2, 251-265. 38, 57

Christakos, G., (1992) *Random field models in earth sciences*, Academic Press, San Diego, CA. 58, 101, 102, 106, 114, 118

Christakos, G. et Bogaert, P., (1996). *Spatiotemporal analysis of spring water ion process derived from measurements at the Dyle basin in Belgium*, IEEE Transactions on Geosciences and Remote Sensing, **34**, No.3, 626-642.

Christakos, G. et Hristopulos, D.T., (1998). *Spatiotemporal environmental health modellig : A tractatus stochasticus*, Kluwer Academic publ., Boston. 98

Christakos, G. et Vyas, V., (1998). *A composite space/time approach to studying ozone distribution over Eastern United States*, Atmospheric Environment, **32**, No.16, 2845-2857. 113

Christakos, G. et Raghu, R., (1996). *Dynamic stochastic estimation of physical variables*, Mathematical Geology, **28**, No.3, 341-365.

Christakos, G. et Bogaert, P., (1996). *Spatiotemporal analysis of springwater ion processes derived from measurements at the Dyle basin in Belgium*, IEEE Transactions Geosciences and remote Sensing, **34**, No.3, 626-642. 115, 117

Christakos, G., Thesing, A., (1993). *The intrinsic Random Field Model in the study of Sulfate Deposition Processes*, Atmospheric Environment, **27A**, No.10, 1521-1540. 57, 58

Colella, P., Woodward, P.R., (1984). *The picewise parabolic method (PPM) for gas-dynamical simulations*, Journal of Computational Physics, textbf11, 38-39. 154

Constantinescu, E. M., Sandu, A., Chai, T., Carmichael, G.R., (2007). *Ensemble-based chemical data assimilation. II : Covariance localization*, Q. J.R. Meteorol. Soc, DOI : 10.1002/qj.77. 152

Corazza, M., Kalany, E., Patil, D., (2002). *Use of the breeding technique to estimate the shape of the analysis 'errors of the day'*, J. Geophys. Res., **10**, 233-243. 147

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. 30, 35, 40, 41, 53, 55, 63, 64

Cressman, G.P., (1959). *An operational objective analysis system*, Mon. Weather Rev., **87**, 367-374. 29

Daley, R., (1991). *Atmospheric Data Analysis*, Cambridge Atmospheric and Space Science Series, Cambridge University Press. ISBN 0-521-38215-7. 132, 156

Dee, D.P., Da Silva, A.M., (1998). *Data assimilation in the presence of forecast bias*, Q.J.R. Meteorol. Soc., **124**, 269-295. 226

De Cesare, L., Myers, D.E., et Posa, D., (1997). *Spatio-temporal modeling of SO2 in Milan district*, in Baafi, E., and Schofield, N., eds., Geostatistical Wollongong '96, **2**, Kluwer Academic Publ., Dordrecht, 1310-1042.

De Cesare, L., Myers, D.E., et Posa, D., (2001). *Product-sum covariance for space-time modeling : an environmental application*, Environmetrics, **12**, 11-23. 107

Dimitrakopoulos, R., et Luo, X., (1994). *Spatiotemporal modelling : covariances and ordinary kriging systems*, Geostatistics for the Next Century, Kluwer Academic Publishers, Dordrecht, 88-93. 101

Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M.B., Badran, F., Thiria, S., Hérault, L., (2002). *Réseaux de neurones. Méthodologie et applications*, Eyrolles, Paris. 190

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**, N.R-3, 5-12. 31

Elichegaray, C., (1997). *Les pollutions de l'air : du local au global*,ADEME, 1997. 9, 239

ERPURS, (1997). *Signification et limites des indicateurs de pollution atmosphérique en milieu urbain. Impact sur la santé selon différents scenarios d'évolution de la pollution atmosphérique en agglomération parisienne. Episodes de pollution et santé en agglomération parisienne*, ORS Île-de-France, Décembre 1997. 11

ESQUIF, (2001). *Étude et Simulation de la QUalité de l'air en Île-de-France*, Institut Pierre Simon Laplace, Laboratoire Interuniversitaire des Systèmes Atmosphériques, Météo-France, Laboratoire d'Aérologie, Forschungszentrum Jülich, AIRPARIF, Rapport final, 2001. 16

Evensen, G., (1994). *Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics*, J. Geophys. Res., **99**, 143-162. 133, 142, 144, 151

Evensen, G., (2003). *The Ensemble Kalman Filter : Theoretical Formulation and Practical Implementation*, Ocean Dynamics, **53**, 343-367.

Evensen, G., (2004). *Sampling strategies and square root analysis schemes for the EnKF*, Ocean Dynamics, **54**, 539-560.

Evensen, G., (2006). *Data assimilation*, Springer Verlag. 146, 149

Fausett, L., (1994) *of Neural Networks. Architectures, Algorithms and Applications*, Prentice Hall, Englewood Cliffs, NJ 07632. 151

Flemming, J., M. van Loon, Stern, R., (2003). *Data assimilation for CTM based on optimum interpolatio and Kalman filter*, paper presented at 26th NATO/CCMS International Technical Meeting on Air Pollution Modeling and Its Application, NATO Comm. on the Challenges of the Mod. Soc., Istanbul. 190, 201

Foxall, R.J., Cawley, G.C., Dorling, S.R., Mandic, D.P., (2002). *Error functions for prediction of episodes of poor air quality*, Proceedings of the International Conference on Artificial Neural Networks (ICANN-2002), Springer Lecture Notes on Computer Science, Vol.2415, Madrid, Spain, 1031-1036. 159

Gardner, M.W., Dorling, S.R., (1998). *Artificial neural network (the multilayer perceptron)-a review of applications in the atmospheric sciences*, Atmospheric Environment 32, 2627-2636. 194

Gardner, M.W., Dorling, S.R., (2000). *Statistical surface ozone models : an improved methodology to account for non-linear behaviour*, Atmospheric Environment 34, 21-34. 193, 198

Gill, P., Murray, W., Wright, M., (1981). *Practical Optimisation*, Academic Press, New York. 201

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

Goovaerts, P. et Sonnet, P., (1993). *Study of spatial and temporal variations of hydrogeochemical variables using factorial kriging analysis*, in A. Soares, ed., Geostatistics Troia '92, **2**, Kluwer Academic Publ., Dordrecht, 745-756. 36

Gratton, Y., (2002). *Le krigeage : La méthode optimale d'interpolation spatiale*, Les Articles de l'Institut d'Analyse Géographique. 98

Grimfeld, A., (2005). *www.debatdeplacements.paris.fr* 36

Haas, T.C., (1995). *Local prediction of a spatio-temporal process with an application to wet sulfate deposition*, Jour. Am. Statistical Assoc., **90**, No. 432, 1189-1199.

Haas, T.C., (1998). *Statistical assessment of spatio-temporal pollutant trends and meteorological transport models*, Atmospheric Environment, **32**, No. 11, 1865-1879.

Haase, P., Schlink, U., (2001). *Non-parametric short-term prediction of ozone concentration in Berlin*, Proceedings Air Pollution Modelling and Simulation Conference, Paris.

Hanea, R. G. (2006). *Error Subspaces Filtering for Atmospheric Chemistry Data Assimilation Modeling*, PhD Thesis, Delft University of Technology, Netherlands. 195, 205

Hanea, R. G., Velders, G., Heemink, A. (2004). *Data assimilation of ground level ozone in Europe with a kalman filter and chemistry transport model*, J. Geophysical. Res., **109**, 5183-5198.

Hanna, S.R., Briggs, G.A., Hosker, R.P.Jr. (1982). *Handbook on atmospheric diffusion*, Technical Information Center, U.S. Departement of Energy.

Haut Comité de la Santé Publique, (2000). *Rapport du Comité sur la santé publique* 130

Heemink, A., Verlaan, M., Segers, A., (2001). *Variance reduced ensemble kalman filtering*, Mon. Weather Rev., **129**, 1718-1728. 11

Hengl, T., Geuvelink, G., Stein, A., (2003). *Comparison of kriging with external drift and regression-kriging*, Technical note, ITC.

Hollingsworth, A., Lönnberg, P., (1986) *The statistical structure of short-range forecast errors as determined from radiosonde data, Part I : The wind fields*, TellusA, **38**, 111-136. 55, 71

Houtekamer, P., Mitchell, H. L., (1998). *Data assimilation using an ensemble kalman filter technique*, Mon. Weather Rev., **126**, 796-811. 159

Houtekamer, P., Mitchell, H. L., (2001). *A sequential ensemble kalman filter for atmospheric data assimilation*, Mon. Weather Rev., **129**, 123-137. 151

Huang, H.C., et Hsu, N.J., (2004). *Modeling transport effects on ground-level ozone using a non-stationary space-time model*, Environmentrics, **15**, 251-268. 152

Ionescu A., Candau Y., (2007). *Air pollutant emissions prediction by process modelling - Application in the iron and steel industry in the case of a re-heating furnace*, Environmental Modelling & Software, **22**, 1362-1371.

International Panel for Climate Change, (2007). *Climate Change 2007 - The Physical Science*, Contribution of Working Group I to the Fourth Assessment Report of the IPCC. 199

Isaaks, E.H., Srivastava, R.H., (1990). *Applied Geostatistics*, Oxford University Press, New York. 7

Jaswinski, A. (1970). *Stochastic Process and Filtering Theory*, Mathematics in Science and Engineering, vol. **64**, Academic Press, New York. 64

Journel, A.G., Huijbregts, Ch., (1978). *Mining geostatistics*, Academic Press, New York. 133, 142

Kalman, R.E., (1960). *A New Approach to Linear Filtering and Prediction Problems*, Transactions of the ASME–Journal of Basic Engineering, **82** (Series D), 35-45. 42, 63

Kitanidis, P., (1987). *Parametric estimation of covariances of regionalized variables*, Water Resources Bulletin, **23**, 557-567. 132, 139

Kitanidis, P., (1993). *Generalized Covariance Functions in Estimation*, Mathematical Geology, **25**, No.5, 525-540. 71

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., (2003). *Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modeling system and measurements in central Helsinki*, Atmospheric Environment 37, 4539-4550. 55, 57

Kyriakidis, P.C. et Journel, A.G., (1999). *Geostatistical Space-Time Models : A Review*, Mathematical Geology, **31**, No.6, 651-684. 194

Lajaunie, C. et Wackernagel, H., (2000). *Geostatistical Approach to Change of Support Problems*, Technical Report N-30/01/G. 97, 125

Le Dimet, F. X., Talagrand, O., (1986). *Variational algorithms for analysis and assimilation of meteorological observations : theoretical aspects*, Tellus (**38A**), 97-110.

Legras, B., Delaygue, G., (2001). *Profils de température et stabilité de l'atmosphère*, http ://planet-terre.ens-lyon.fr/planetterre/XML/db/planetterre/metadata/LOM-atmosphere-temperature.xml#id2451224. 132

Loi N.96-1236 du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie, Journal Officiel de la République Française.

Maier, H.R., Dandy, G.C., (2000). *Neural networks for the prediction and forecasting of water resources variables : a review of modelling issues and applications*, Environmental Modelling & Software 15, 101-124.

Marchant, B.P., Lark, R.M., (2007). *The Matérn variogram model : Implications for uncertainty propagation and sampling in geostatistical surveys*, Geoderma, **140**, 337-345. 193, 199

Mardia, K.V. et Goodall, C.R. (1993). *Spatial-temporal analysis of multivariate environmental data*, in Patil, G.P. et Rao, C.R., eds., Multivariate environmental statistics, Elsevier, Amsterdam, 347-386. 45

Mateu, J., Montes, F., et Fuentes, M., (2003). *Recents advances in space-time statistics with applications to environmental data : An overview*, J. Geophysical. Res., **108**. 98

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. 27, 34, 220

Matheron, G., (1963b). *Traité de géostatistique appliquée, Tome 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. 34

Matheron, G., (1973). *The intrinsic random functions and their applications*, Advances in Applied Probability, **5**, 439-468. 27, 34, 38, 55

Maybeck, P. (1979). *Sochastic models, estimation, and control*, Mathematics in Science and Engineering, vol. **141, 141-2**, Academic Press, New York. 64

Medina, S., Le Tertre, A., Quenel, P., Le Moullec, Y., Lameloise, P., Guzzo, J.C., Festy, B., Ferry, R., Dab, W., (1997). *Air pollution and doctor's house calls : results from the ERPURS system for monitoring the effects of air pollution on public health in greater Paris, France, 1991-1995*, Environmental Research 75, 73-84. 141

Mejia, J.M., et Rodriguez-Iturbe, I., (1974). *On the synthesis of random field sampling from the spectrum : an application to the generation of hydrologic spatial processes*, Water Resources Res., **10**, No. 1, 705-711.

Menut, L., Vautard, R., Beekmann, M., Honoré, C., (2000). *Sensitivity of photochemical pollution using the adjoint of a simplified chemistry-transport model*, J. Geophys. Res., **105**, No. D12, 15,379-15,402. 98, 101

Miller, R.N., Carter, E.F., Blue, S.T., (1999). *Data assimilation into nonlinear stochastic models*, Tellus A, **51**, No. 2, 167-194. 156

Möller, M.S., (1993). *A scaled conjugate gradient algorithm for fast supervised learning*, Neural Networks 6, 525-534. 144

Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R., Chatterton, T., (2004). *Modelling SO2 concentration at a point with statistical approaches*, Environmental Modelling & Software 19, 887-905. 200

Paluš, M., Pelikán, E., Eben, K., Krejčíř, P., Juruš, P., (2001). *Nonlinearity and prediction of air pollution. Artificial Neural Nets and Genetic Algorithms*, Proceedings of the International Conference, Wien, Springer 2001, 473-476. 201

Pham, D., Verron, J., Roubaud, M., (1997). *A singular evolutive extended Kalman Filter for data assimilation in oceanography*, Journal of Marine Systems, **16**(3-4), 1194-1207. 195, 205

Plan de Protection de l'Atmosphère, (2007). *http ://www.drire.gouv.fr/ile-de-france* 133

Ripley, B.D., (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press. 21, 22

Rouhani, S., Ebrahimpour, R.M., Yaqub, I. et Gianella, E., (1992). *Multivariate geostatistical trend detection and network evaluation of space-time acid deposition data - I. Methodology*, Atmospheric Environment, **26A**, No. 14, 2603-2614. 201

Rouhani, S., Ebrahimpour, R.M., Yaqub, I. et Gianella, E., (1992). *Multivariate geostatistical trend detection and network evaluation of space-time acid deposition data - II. Application to NADP/NTN data*, Atmospheric Environment, **26A**, No. 14, 2615-2626. 101

Rouhani, S., et Wackernagel, H., (1990). *Multivariate geostatistical approach to space-time data analysis*, Water Resources Res., **26**, No. 4, 585-591. 98

Rumelhart, D.E., Hinton, G.E., Williams, R.J., (1986). *Learning internal representations by error propagation*, Parallel Distributed Processing : Explorations in the Microstructure of Cognition, vol1, Cambridge MA : MIT Press, 318-362. 98, 101

SANLIB, (1995). *Stochastic analysis software library and user's guide*, Stochastic Res. Group, Dep. environ. Sci. Eng., Univ. North Carolina, Chapel Hill, 1995. 200

Saporta, G., (1990). *Probabilités, analyse des données et statistique*, Éditions Technip, Paris. 114, 116

Schlink, U., John, S., Herbarth, O., (2001). *Transfer-Function Models Predicting Ozone in Urban Air*, Contribution to the SATURN Project, Annual Report 2001. 33, 158

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., Doyle, M., (2003). *A rigorous inter-comparison of ground-level ozone predictions*, Atmospheric Environment 37, 3237-3253. 195

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E., (2006). *Statistical models to asses the health effects and to forecast ground-level ozone*, Environmental Modelling & Software 21, 547-558. 194, 195, 201, 202, 205

Segers, A., (2002). *Data assimilation in atmospheric chemistry models using Kalman filtering*, PhD Thesis, Delft University of Technology, Netherlands. 202

Séguret, S.A., (1989). *Filtering periodic noise by using trigonometric kriging*, in M.Armstrong, ed., Geostatistics, **1**, Kluwer Academic Publ., Dordrecht, 481-491.

Seinfeld, J.H., (1986). *Atmospheric Chemistry and Physics of Air Pollution*, Wiley, New York. 98

Seinfeld, J.H., Pandis, S.N., (1998). *Atmospheric Chemistry and Physics from Air Pollution to Climate Change*, Wiley, New York. 15

Sibson, R., (1981). *A brief description of natural neighbour interpolation*, Barnett V., Interpretating Multivariate Data, John Wiley & Sons Inc., New York, 21-36. 11

Simpson, D., Guenther, A., Hewitt, C.N., Steinbrecher, R., (1995). *Biogenic emissions in Europe, 1. Estimates and uncertainties*, J. Geophysical. Res., **100**, 22875-22890. 30

Shanno, D.F., (1978). *Conjugate gradient methods with inexact line searches*, Mathematics of Operations Research3, 244-256. 24

Smith, R.L., Kolenikov, S. et Cox, L., (2003). *Spatiotemporal modeling of PM2,5 data with missing values*, J. Geophysical. Res., **108**. 200

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C., (2007). *Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations*, Environmental Modelling & Software 22, 97-103.

Stein, M.L., (1999). *Interpolation of Spatial Data : Some Theory for Kriging*, Springer Verlag, New York. 194

Switzer, P., (1988). *Non-stationary spatial correlations estimated from monitoring data*, in M. Armstrong, editor, Geostatistics, Kluwer Academic Publishers, 1, 55-67. 40, 44

Talagrand, O., (1997). *Assimilation of Observations, an Introduction*, Journal of the Meteorological Society of Japan, **75**, 191-209. 100

Tarantola, A., (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, USA. 133

Tippett M.K., Anderson, J.L.,Bishop, C.H., Hamill, T.M., Whitaker, J.S., (2003). *Ensemble square-root filters*, Mon. Weather Rev., **131**, 1485-1490.

Todling, R., (1999). *Estimation Theory and Foundations of Atmospheric Data Assimilation*, DAO Office Note 1999-01, Goddard Space Flight Center. 149

Van Leeuwen, P., (1998). *Comment on "Data assimilation Using an Ensemble Kalman Filter Technique"*, Mon. Weather Rev., **127**, 1374-1377.

Van Oss, R.F., Voors, R.H.M., Spurr, R.D.J., (2002). *Ozone Profile Algorithm*, in Algorithm theoretical baseline document, II, OMI Ozone Products. Ed. P.K.Bhartia, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA. 148

Vautard, R., Beekmann, M., Roux, J., Gombert, D., (2001). *Validation of a hybrid forecasting system for the ozone concentrations over the Paris area*, Atmospheric Environment 35, 2449-2461. 227

Venkatram, A., (1988). *On the use of kriging in the spatial analysis of acid precipitation data*, Atmospheric Environnement, **22**, 1963-1975. 155, 189

Verlaan, M., Heemink, A., (1997). *Tidal flow forecasting using reduced-rank square root filters*, Stochastic Hydro. Hydraul, **11**, 349-368.

Verwer, J. G., (1994). *Gauss-Seidel iteration for stiff ODEs from chemical kinetics*, SIAM Journal on Scientific Computing, **15**, 1243-1250. 133

Viotti, P., Liuti, G., Di Genova, P., (2002). *Atmospheric urban pollution : applications of an artificial neural network (ANN) to the city of Perugia*, Ecological Modelling 148, 27-46. 154

Vyas, V.M., et Christakos, G., (1997). *Spatiotemporal analysis and mapping of sulfate deposition data over Eastern U.S.A*, Atmospheric Environment, **31**, No. 21, 3623-3633. 193, 196

Wackernagel, H., (2002). *Geostatistical normalization of air pollution transport model output and station data using ISATIS*, Technical Report N-20/02/G. 98

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

Wahba, G., (1990). *Spline models for observational data*, **59**, Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, PA. 36

Whitaker, J.S., Hamill T.M., (2002). *Ensemble data assimilation without perturbed observations*, Mon. Weather Rev., **130**, 1913-1924. 31

WHO, (2001). *Air Quality Guidelines for Europe*, Second Edition. 149

Willmott, C.J., (1982). *Some comments on the evaluation of model performance*, Bulletin of the American Meteorological Society 63 (11), 1309-1313. 208
201

Résumé

Cette étude est consacrée à la modélisation spatio-temporelle de la pollution atmosphérique urbaine en utilisant un ensemble de méthodes statistiques exploitant les mesures de concentrations de polluants (NO_2, O_3) fournies par un réseau de surveillance de la qualité de l'air (AIRPARIF).

Le principal objectif visé est l'amélioration de la cartographie des champs de concentration de polluants (le domaine d'intérêt étant la région d'Île-de-France) en utilisant, d'une part, des méthodes d'interpolation basées sur la structure spatiale ou spatio-temporelle des observations (krigeage spatial ou spatio-temporel), et d'autre part, des algorithmes, prenant en compte les mesures, pour corriger les sorties d'un modèle déterministe (Filtre de Kalman d'Ensemble).

Les résultats obtenus montrent que dans le cas du dioxyde d'azote la cartographie basée uniquement sur l'interpolation spatiale (le krigeage) conduit à des résultats satisfaisants, car la répartition spatiale des stations est bonne. En revanche, pour l'ozone, c'est l'assimilation séquentielle de données appliquée au modèle (CHIMERE) qui permet une meilleure reconstitution de la forme et de la position du panache pendant les épisodes de forte pollution analysés.

En complément de la cartographie, un autre but de ce travail est d'effectuer localement la prévision des niveaux d'ozone sur un horizon de 24 heures. L'approche choisie est celle mettant en œuvre des méthodes de type réseaux neuronaux. Les résultats obtenus en appliquant deux types d'architectures neuronales indiquent une précision correcte surtout pour les 8 premières heures de l'horizon de prédiction.

Mots clés : pollution atmosphérique, dioxyde d'azote, ozone, cartographie, prévision, géostatistique, assimilation séquentielle de données, réseaux neuronaux artificiels.

Abstract

This study is devoted to the spatio-temporal modelling of air pollution at a regional scale using a set of statistical methods in order to treat the measurements of pollutant concentrations (NO_2, O_3) provided by an air quality monitoring network (AIRPARIF).

The main objective is the improvement of the pollutant fields mapping using either interpolation methods based on the spatial or spatio-temporal structure of the data (spatial or spatio-temporal kriging) or some algorithms taking into account the observations, in order to correct the concentrations simulated by a deterministic model (Ensemble Kalman Filter).

The results show that nitrogen dioxide mapping based only on spatial interpolation (kriging) gives the best results, while the spatial repartition of the monitoring sites is good. For the ozone mapping it is the sequential data assimilation that leads us to a better reconstruction of the plume's form and position for the analyzed cases.

Complementary to the pollutant mapping, another objective was to perform a local prediction of ozone concentrations on a 24-hour horizon; this task was performed using Artificial Neural Networks. The performance indices obtained using two types of neural architectures indicate a fair accuracy especially for the first 8 hours of prediction horizon.

Keywords : atmospheric pollution, nitrogen dioxide, ozone, mapping, prediction, geostatistics, sequential data assimilation, artificial neural networks.