



Analyse et développement de modèles statistiques pour l'estimation et la prédiction spatiale et temporelle de la pollution atmosphérique à partir de données issues de capteurs mobiles

Mohamed Yacine Idir

► To cite this version:

Mohamed Yacine Idir. Analyse et développement de modèles statistiques pour l'estimation et la prédiction spatiale et temporelle de la pollution atmosphérique à partir de données issues de capteurs mobiles. Statistiques [math.ST]. Université Paris-Saclay, 2023. Français. <NNT : 2023UPASG107>. <tel-04518362>

HAL Id: tel-04518362

<https://theses.hal.science/tel-04518362v1>

Submitted on 23 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Analyse et développement de modèles statistiques pour l'estimation et la prédiction spatiale et temporelle de la pollution atmosphérique à partir de données issues de capteurs mobiles

*Analysis and development of statistical models for the
spatial and temporal estimation and prediction of air
pollution using mobile sensor data*

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, sciences et technologies de l'information et de la
communication (STIC)

Spécialité de doctorat : Informatique mathématique

Graduate School : Informatique et sciences du numérique. Référent : Université de
Versailles-Saint-Quentin-en-Yvelines

Thèse préparée dans les unités de recherche PICS-L (Université Gustave Eiffel),
ESTACA'Lab (ESTACA), sous la direction de **Olivier ORFILA**, CR, HDR, la codirection de
Patrice CHATELLIER, Directeur de recherche, le coencadrement de **Vincent JUDALET**,
Maître de conférences

Thèse soutenue à Montigny-le-Bretonneux, le 21 décembre 2023, par

Yacine Mohamed IDIR

Composition du jury

Membres du jury avec voix délibérative

Karine Zeitouni
Professeure des universités, UVSQ
Denis Allard
Directeur de recherche, INRAE
Hervé Rivano
Directeur de recherche, INRIA/INSA Lyon
Hanene Azzag
Maître de conférence, Université Paris 13

Présidente
Rapporteur et Examinateur
Rapporteur et Examinateur
Examinatrice

Titre : Analyse et développement de modèles statistiques pour l'estimation et la prédiction spatiale et temporelle de la pollution atmosphérique à partir de données issues de capteurs mobiles

Mots clés : Qualité de l'air, Capteurs mobiles, modélisation statistique, Géostatistique.

Résumé : La pollution atmosphérique urbaine, fléau mondial causant des millions de décès par an, rend les cartographies précises de ce phénomène non seulement pertinente mais vitale pour la santé publique. Actuellement, la surveillance de la qualité de l'air est assurée par des stations fixes de surveillance de la qualité de l'air. Ces stations de référence fournissent une mesure très précise de la qualité de l'air au prix d'une couverture spatiale limitée.

L'idée d'utiliser de nouveaux capteurs à faible coût développés à partir de récentes avancées technologiques, plus petits, intégrant un système de positionnement global (GPS) a rapidement émergée. Les scientifiques disposent ainsi d'outils supplémentaires pour affiner les cartes spatio-temporelles de la pollution atmosphérique et créer de nouveaux ensembles de données fournissant des informations sur la qualité de l'air qui n'étaient pas dis-

ponibles auparavant.

La génération de cartographies précises de la qualité de l'air à l'aide de ces capteurs bas coût présente plusieurs défis majeurs. Ces défis sont principalement liés à la nature du phénomène étudié, à la précision et au volume des données.

Compte tenu de ces difficultés, il est essentiel de savoir comment combiner toutes ces sources de données floues pour obtenir une image claire de la pollution urbaine. C'est dans ce contexte que s'inscrit cette thèse dont le but est l'analyse et le développement de modèles statistiques qui exploitent les données acquises par des capteurs mobiles bas coût. Elle contribue à l'objectif de fournir des estimations et des prévisions spatiales et temporelles plus précises de la pollution de l'air urbain, en combinant modèles mathématiques et avancées technologiques.

Title : Analysis and development of statistical models for the spatial and temporal estimation and prediction of air pollution using mobile sensor data

Keywords : Air quality, Mobile sensors, Statistical modeling, Geostatistics.

Abstract : Urban air pollution, a global health crisis causing millions of deaths every year, makes accurate mapping of this phenomenon not only relevant, but vital to public health. Currently, air quality is measured by fixed air quality monitoring stations. These reference stations provide a highly accurate measure of air quality, at the cost of limited spatial coverage.

The idea of using new low-cost sensors developed from recent technological advances, smaller in size and incorporating a global positioning system (GPS), quickly emerged. This gives scientists additional tools to refine spatio-temporal maps of air pollution and create new datasets providing information on air quality that was previously unavailable.

Generating precise air quality maps using these low-cost sensors presents several major challenges. These challenges are mainly related to the nature of the phenomenon being studied, and to the accuracy and volume of the data.

Given these difficulties, it is important to know how to combine all these fuzzy data sources to obtain a clear picture of urban pollution. The aim of this thesis is to analyze and develop statistical models that exploit data acquired by low-cost mobile sensors. It contributes to the objective of providing more accurate spatial and temporal estimates and forecasts of urban air pollution, by combining mathematical models and technological advances.

Remerciements

En réfléchissant au parcours qui a abouti à cette thèse, je suis submergé de gratitude pour les nombreuses personnes qui m'ont soutenu, guidé et inspiré tout au long de ce parcours.

Avant tout, je souhaite exprimer mes remerciements aux membres du jury pour avoir consenti à évaluer les travaux présentés dans ce manuscrit de thèse : Je tiens à exprimer ma profonde gratitude au Dr Denis Allard et au Dr Hervé Rivano pour avoir accepté la tâche importante d'être rapporteurs de ce manuscrit. Leur engagement et leur savoir dans mon domaine de recherche me sont d'ores et déjà une source d'encouragement considérable. Je tiens également à remercier chaleureusement Dr. Karine Zeitouni et Dr. Hanene Azzag pour leur rôle d'examinatrices. Je suis également reconnaissant envers Laure Malherbe. Leur participation à l'évaluation de ce travail est grandement appréciée.

Je tiens à témoigner ma sincère reconnaissance envers mon directeur de thèse le Dr. Olivier Orfila, dont l'expertise, la patience et les conseils ont façonné bien plus que ce travail. Ta confiance en moi, manifestée bien avant l'entame de cette thèse, conjuguée à ta constante bienveillance et à l'absence de toute parole négative, ont constitué le pilier solide sur lequel j'ai pu m'appuyer tout au long de cette aventure intellectuelle.

Je remercie beaucoup mes deux encadrants, le Dr Patrice Chatellier et le Dr. Vincent Judalet. Leur profonde connaissance du domaine, associée à leurs orientations judicieuses et à leurs relectures minutieuses, ont été déterminantes durant ces trois années de recherche.

Au cours de cette thèse, réalisée en deux lieux distincts, j'ai eu le privilège de côtoyer des personnes incroyablement enrichissantes. Tout d'abord, je tiens à exprimer ma gratitude à mes collègues de l'Université Gustave Eiffel : un grand merci à Rémi, Jérémy et Ziri pour nos innombrables discussions lors des déjeuners.

Je tiens également à remercier chaleureusement Pascale Dyperat, auprès de qui j'ai appris l'humilité à travers mes nombreuses défaites au ping-pong durant nos pauses.

Enfin, un remerciement sincère à Damien Ingrosso, dont la bonne humeur et la joie de vivre ont apporté un rayon de soleil à mes journées.

À mes amis, doctorants et collègues de bureau à l'ESTACA : une profonde gratitude est adressée à Mokhtar, Aude, Youssef et Mathieu. Votre soutien et nos innombrables

débats philosophiques au bureau ont grandement enrichi mon parcours.

Je tiens également à exprimer ma reconnaissance envers Antoine, Khaled et Victor. Sans leur précieuse collaboration, ce manuscrit aurait sans doute vu le jour bien plus tôt.

Sur une note plus personnelle, je souhaite exprimer ma profonde gratitude envers ma chère Alice. Sa présence, son soutien moral sans faille et ses relectures minutieuses ont été essentiels durant la dernière étape de cette thèse.

Je tiens également à remercier ma famille. Mes deux sœurs, Lilia et Inès, ont été constamment à mes côtés, m'apportant leur soutien sans relâche. Quant à Badis, mon frère, il m'a grandement éclairé sur l'univers de la recherche. Enfin, à mes parents, dont le dévouement et l'assistance n'ont jamais tari depuis le jour de ma naissance.

Introduction

Nous sommes nombreux à consulter les prévisions météorologiques avant de sortir de chez nous. Nous consultons ces prévisions pour prendre des décisions quotidiennes : « Un parapluie s'impose-t-il ce matin ? » ou « Serait-il plus sage de reporter ma promenade ? » Ces considérations, apparemment banales, jouent un rôle important pour notre confort et notre préparation au quotidien. Pourtant, si la pluie peut nous incommoder, elle constitue rarement une menace directe pour notre santé. En revanche, l'air que nous respirons, chargé de polluants potentiels, peut avoir des conséquences immédiates et à long terme sur notre santé. Si nous donnons la priorité aux prévisions météorologiques pour ne pas être surpris par une averse, ne devrions-nous pas être plus vigilants quant à la qualité de l'air que nous respirons ?

La pollution de l'air, menace silencieuse mais omniprésente, affecte des millions de personnes dans le monde. Dans son rapport datant d'avril 2022, l'organisation mondiale de la santé établit que 99 % de la population mondiale respire un air dont la concentration de polluants pose des problèmes pour notre santé. Au niveau mondial, le nombre de décès lié à la pollution atmosphérique est estimée à 7 millions par an dont 4,2 millions liés à la pollution de l'air extérieur. En France cette dernière est responsable de 48 000 décès prématurés, et a un coût annuel pour le système de santé de 100 milliards d'euros.

Les difficultés à prévoir avec précision la qualité de l'air entravent notre capacité à adapter notre comportement en fonction des niveaux de pollution. Souvent, ces prévisions fournissent un indice général pour l'ensemble d'une région plutôt qu'un aperçu spécifique des variations locales de la qualité de l'air. Cette approche générale ne permet pas de saisir les nuances complexes de la pollution atmosphérique urbaine qui peuvent varier de manière significative même sur de courtes distances. En effet, la pollution de l'air est loin d'être uniformément répartie dans une ville. Une carte détaillée de la pollution atmosphérique urbaine nous permettrait de nous poser de nouvelles questions avant de sortir : « Dois-je prendre mon masque ? », « Un autre itinéraire pourrait-il offrir un air plus pur ? » ou encore « Dois-je adapter mon effort ou décaler mon horaire de sortie ».

Une partie des difficultés à prévoir précisément cette pollution est liée au coût important des capteurs précis et certifiés qui en limite le nombre et qui constituent donc un maillage insuffisamment dense. Avec les progrès technologiques et l'avènement des capteurs de pollution miniaturisés et bas coût, nous pourrions disposer d'outils pour surveiller, analyser et prédire la qualité de l'air avec une précision sans précédent. Ces capteurs, de par leur petite taille et prix modique, peuvent être déployés en grand

nombre et à bord de divers véhicules et mobiles. Plusieurs cas d'usage ont déjà été imaginé avec des flottes de capteurs mobiles embarqués sur des tramways, voitures ou vélos sillonnant la ville et mesurant les variations de la pollution atmosphérique au plus près. Dans un monde où la qualité de l'air peut fluctuer de façon spectaculaire sur de courtes distances et dans de courts laps de temps, l'accès à des données au caractère prédictif peut avoir de fortes conséquences en matière de santé publique.

En 2017, issues de la collaboration entre l'université Gustave Eiffel et l'ESTACA, des technologies de boîtiers mobiles de pollution, équipés entre autres de divers capteurs de pollution bas coût et un logiciel d'affichage des données brutes ont vues le jour. Rapidement, le sujet de l'exploitation des grandes quantités de données s'est orienté vers l'utilisation de méthodes issues des mathématiques appliquées pour accroître le pouvoir descriptif des données récoltées. Ainsi, un plan d'expérimentation pour déployer nos capteurs sur des vélos et mesurer la qualité de l'air à l'échelle d'un quartier a été mis en œuvre afin de collecter une première base générique de données de pollution. C'est dans ce contexte qu'est née cette thèse dont le but est l'analyse et le développement de modèles statistiques qui exploitent les données acquises par des capteurs mobiles et peu onéreux. Elle contribue à l'objectif de fournir des estimations et des prévisions spatiales et temporelles plus précises de la pollution de l'air.

Le point de départ de cette thèse était de comparer les méthodes de cartographie de la qualité de l'air présentes dans la littérature et historiquement utilisées avec des stations fixes de référence. Pour traiter les mesures et en extraire des informations pertinentes, il est possible de différencier deux types d'approches : la modélisation déterministe (physique) et la modélisation statistique (empirique). La première approche repose sur la résolution numérique des équations, basées sur les lois physiques (lois de la thermodynamique, mécanique des fluides, etc.). À l'échelle continentale ou nationale, ces modèles sont intensivement utilisés. Cependant, à l'échelle régionale ou urbaine, bien qu'ils demeurent utilisés, ils présentent des difficultés d'utilisation pour une cartographie détaillée de la qualité de l'air. Cette limitation est attribuable à la complexité inhérente de la modélisation à petite échelle, où certaines variables peuvent être inconnues ou partiellement occultées. La seconde approche, étudiée dans cette thèse, emploie un panel de données décrivant les variables déterminantes du système (mesures des concentrations, observations météorologiques...) et cherche à trouver des équations empiriques qui relient ces données.

Ces modèles statistiques, qui serviront à approximer le phénomène déterministe qu'est la dispersion des polluants, devront être descriptifs pour nous aider à mieux comprendre ce processus complexe à petite échelle. Ils devront par ailleurs fournir l'erreur associée à la prédiction, afin de mieux juger de la qualité et pertinence de cette dernière.

De plus, ces modèles devront être à même de gérer d'énormes quantités de données. En effet, les capteurs bas coûts peuvent produire jusqu'à une mesure par seconde, obtenant ainsi des jeux de données dépassant les dizaines voire centaines de milliers d'observations. Ils offriront ainsi aux responsables politiques de nouveaux outils pour lutter contre la pollution atmosphérique.

En explorant les modèles issus de la géostatistique spatio-temporelle et en les comparant à d'autres familles de modèles, ces modèles se sont imposés en répondant à plusieurs de nos critères de recherche énoncés dans le paragraphe précédent. Ils surpassent les techniques d'interpolations déterministes en termes de précision de prédiction et fournissent des mesures d'incertitude. Leur caractère descriptif les rend plus performants que les modèles d'apprentissage qui sont plutôt considérés comme des boîtes noires.

Cependant, deux limitations majeures dans l'approche globale de la thèse ont été détectées :

- Les modèles issus de la géostatistique spatio-temporelle ont une complexité cubique par rapport au nombre d'observations. Ceci restreint leur utilisation, surtout pour des applications en temps réel.
- Les données issues de capteurs bas coût sont bruitées et biaisées. En effet, en comparant ces données aux données des stations fixes de la qualité de l'air on remarque un écart significatif.

Les travaux de cette thèse se sont concentrés sur la résolution de ces deux problèmes. En explorant les méthodes d'approximations des processus gaussiens et plus précisément l'approximation de Vecchia. Un modèle hiérarchique à trois niveaux a été proposé, capable de mieux compenser la nature bruitée et peu précise des données issues de capteurs bas coût par rapport aux modèles précédemment étudiés.

Le dernier pan de cette recherche s'est concentré sur l'utilisation de capteurs considérés plus fiables conjointement avec les capteurs bas coûts. Cette approche à double typologie de capteurs peut offrir un équilibre optimal entre la quantité et la qualité des données ce qui pourrait révolutionner la façon dont nous surveillons et interprétons les phénomènes spatiaux.

En outre, si les efforts déployés dans le cadre de cette thèse ont permis de relever certains des défis inhérents aux capteurs peu coûteux, il est impératif que les efforts futurs mettent l'accent sur l'amélioration des capteurs eux-mêmes. Les progrès dans la technologie des capteurs, les techniques d'étalonnage et les mécanismes de contrôle

de la qualité peuvent combler le fossé entre les capteurs bon marché et les capteurs de haute qualité. De tels développements permettront non seulement d'améliorer la fiabilité des données, mais aussi d'étendre l'applicabilité des méthodologies proposées dans cette thèse.

Table des matières

Remerciements	iii
Introduction	vi
Table des matières	x
Table des figures	xiii
Liste des tableaux	xvii
Liste des Acronymes	xviii
Liste des symboles	xx
Summary of the thesis in English - Résumé de la thèse en anglais	xxii
Air under surveillance : atmospheric quality, sensors and mapping . . .	xxiii
Towards better pollution mapping : exploring existing models and techniques	xxvii
From fundamentals to advanced techniques : a comparison of pollution models	xxxii
Geostatistical model enhancement : autocorrelation, approximation and application	xxxvi
Measure synergy : Practical strategies for better air quality forecasting .	xlili
Conclusions and outlook	xlvi
1 Air sous surveillance : cartographie, capteurs et qualité atmosphérique	2
1.1 La pollution atmosphérique : définition et méthodes de mesure . .	5
1.1.1 Définitions et origines de la pollution atmosphérique	5
1.1.2 Principaux polluants	7
1.1.3 Impacts sur la santé	10
1.1.4 Surveillance de l'air : étape clé contre la pollution atmosphérique	12
1.1.5 Méthodes actuelles d'évaluation de la qualité de l'air	13
1.1.6 Émergence des capteurs bas coût	16

1.1.7	Technologie des capteurs bas coût	18
1.1.8	Spatialisation des données : de la mesure à la cartographie	22
1.2	Stratégies de recherche	24
1.2.1	Défis de la thèse	24
1.2.2	Questions de recherche	26
1.2.3	Méthodologie	27
1.2.4	Jeux de données utilisés durant la thèse	28
1.3	Conclusion	38
	Références	39
2	Vers une meilleure cartographie de la pollution : exploration des modèles et techniques existantes	48
2.1	Concepts fondamentaux des méthodes de spatialisation	50
2.2	Approches traditionnelles avec les stations fixes	51
2.2.1	Techniques d'interpolation déterministes	54
2.2.2	Méthodes géostatistiques	55
2.2.3	Modèles d'utilisation des sols	56
2.2.4	Modèles de dispersion	60
2.3	Spatialisation via les capteurs mobiles bas coût	63
2.4	Conclusion	66
	Références	67
3	Des fondamentaux aux techniques avancées : une comparaison des modèles de pollution	74
3.1	Méthodes utilisant uniquement des données de pollution atmosphérique	76
3.1.1	Techniques déterministes	79
3.1.2	Méthodes stochastiques	88
3.2	Intégration et comparaison de modèles exploitant des données externes	109
3.2.1	Méthodologie	109
3.2.2	Résultats	122
3.2.3	Discussion	131
3.3	Conclusion	134
	Références	136
4	Affinement du modèle géostatistique : autocorrélation, approximation et application	138
4.1	Définition du modèle géostatistique à partir de l'approche hiérarchique	141
4.2	Prise en compte et modélisation de l'autocorrélation des capteurs bas coût	144
4.3	Méthodes d'approximation du processus gaussien	149

4.4	Étude et application de l'approximation de Vecchia au nouveau modèle proposé	150
4.4.1	Méthodologie	151
4.4.2	Résultats d'évaluation	160
4.4.3	Application numérique	163
4.4.4	Discussion	168
4.5	Conclusion	169
	Références	171
5	Synergie des mesures : stratégies pratiques pour une meilleure pré- vision de la qualité de l'air	174
5.1	Extension pratique du modèle théorique	176
5.1.1	Reconceptualisation du modèle	176
5.1.2	Méthodologie expérimentale	178
5.2	Évaluation globale : résultats et questions de recherche	185
5.2.1	Création de cartographies	186
5.2.2	Pouvoir descriptif du modèle	188
5.2.3	Complexité	191
5.2.4	Erreur associée à la prédiction	192
5.2.5	Étude du nombre de capteurs	193
5.3	Conclusion	196
	Références	198
6	Conclusion générale	199
6.1	Synthèse des travaux de recherche	199
6.2	Conclusions et perspectives	205

Table des figures

1	Methodology used to answer the research questions.	xxvi
2	RMSE as a function of the percentage of data used in the three scenarios studied.	xxxiii
3	Comparative mapping of PM ₁₀ concentrations according to different models	xxxvi
4	Configurations studied : the column indicates the order from left to right : Spatial, Temporal, Random, Middle Out, MMD, Sensor. The first two rows indicate spatial distance only, the third and fourth rows indicate temporal distance only, and the last two rows indicate spatio-temporal distance. Even rows indicate conditioning based primarily on observations from the same sensors. Uneven rows indicate conditions regardless of sensor. The black circles represent observations ordered from 1 to 25, the triangular dot corresponds to number 26, and the squared dot represents the conditioning set for observation 26.	xli
5	PM ₁₀ Maps resulting from application of the proposed model using Vecchia's approximation to estimate parameters.	xlii
6	Study of the impact of the number and type of sensors on prediction variance.	xlvi
1.1	Un policier londonien, avec Big Ben en arrière-plan, dirige la circulation pendant le pic de pollution de 1952.	6
1.2	Exemple d'une station de mesure fixe de qualité de l'air gérée par Airparif	14
1.3	Exemple de boîtier mobile de pollution contenant un capteur de pollution bas coût utilisé dans l'étude [31]	20
1.4	Méthodologie retenue pour répondre aux questions de recherche.	27
1.5	Emplacement de la station fixe de référence et des lignes de tramway dans la ville de Zurich.	30
1.6	Boîtier mobile d'Atmotrack incluant un capteur de pollution bas coût, installé sur le toit d'un véhicule de la Poste.	31

1.7	Emplacement des capteurs fixes d'Atmotrack (LCS) et des stations de surveillance à Nantes	32
1.8	Trajectoires des vélos équipés de capteurs mobiles dans le quartier des Halles	35
1.9	Représentation des trajectoires de capteurs mobiles réelles pour la simulation d'un jeu de données.	37
2.1	Les différentes échelles de modélisation de la qualité de l'air. Du global au local (source : Michelot et al. 2015).	52
2.2	Cartographie des niveaux d'ozone dans la région d'Île de France produite par Airparif. Source : "Airparif.fr" [3].	53
2.3	Principaux modèles d'utilisation des sols présents dans la littérature, allant d'une complexité croissante.	57
2.4	Sorties du modèle CHIMERE pour la concentration du NO ₂ avec une résolution spatiale de 2 km utilisant 2000 CPUs d'un supercalculateur.	62
3.1	Représentation des données d'OpenSense dans un espace spatio-temporel.	77
3.2	Représentation d'un espace bidimensionnel divisé en cellules par la méthode de Voronoï	80
3.3	Application de la méthode de diagrammes de Voronoï à une série de données issue d'une ligne de tramway	82
3.4	Application de la méthode de TPS à une série de données issue d'une ligne de Tramway	84
3.5	Illustration de l'application de la méthode IDW sur des données provenant d'une ligne de tramway.	86
3.6	Illustration de l'application des trois méthodes déterministes.	87
3.7	Variogrammes empiriques spatio-temporels directionnels.	96
3.8	Variogrammes spatio-temporels associés au modèle de krigeage ordinaire.	97
3.9	Variogrammes spatio-temporels associés au modèle de krigeage simple.	98
3.10	Variogrammes spatio-temporels associés au modèle de krigeage avec dérive externe.	99
3.11	Comparaison entre les prévisions et les valeurs réelles de quatre lignes de tramway le 4 mars 2016 dans le premier scénario.	100
3.12	Comparaison entre les prévisions et les valeurs réelles de quatre lignes de tramway le 4 mars 2016 dans le deuxième scénario.	100
3.13	Comparaison entre les prévisions et les valeurs réelles de quatre lignes de tramway le 4 mars 2016 dans le troisième scénario.	101

TABLE DES FIGURES

3.14	RMSE en fonction du pourcentage de données utilisé dans les trois scénarios étudiés.	103
3.15	Biais en fonction du pourcentage de données utilisé dans les trois scénarios étudiés.	103
3.16	Corrélations en fonction du pourcentage de données utilisé dans les trois scénarios étudiés.	104
3.17	Cartes des concentrations d'ozone résultant de l'estimateur KED à Zurich, ici pour le 4 mars 2016. De 5 heures à 22 heures.	105
3.18	Cartes de l'estimateur de variance résultant de l'estimateur KED à Zurich, présentées ici pour le 4 mars 2016. De 5 heures du matin à 10 heures du soir.	106
3.19	Données d'humidité dans la ville de Nantes, du 20 au 30 novembre 2018.	111
3.20	Illustration de certaines covariables spatiales utilisées dans cette étude.	113
3.21	Sorties PM ₁₀ de prévisions et d'analyses du modèle CHIMERE du 20 au 30 novembre 2020 pour la ville de Nantes	114
3.22	Résultats des RMSE en validant sur les capteurs bas coût dans le premier scénario.	123
3.23	Résultats des RMSE en validant sur les capteurs bas coût dans le deuxième scénario.	124
3.24	Résultats des corrélation en validant sur les stations de surveillance fixes dans le premier scénario.	126
3.25	Résultats des corrélation en validant sur les stations de surveillance fixes dans le deuxième scénario.	127
3.26	Cartographie comparative des concentrations PM ₁₀ selon différents modèles	130
4.1	Représentation des erreurs classiques des capteurs bas coût. Tiré de l'article : Maag et. al (2018). A survey on sensor calibration in air pollution monitoring deployments	145

4.2	Configurations étudiées : la colonne indique l'ordre de gauche à droite : Spatial, Temporel, Aléatoire, Middle Out, MMD, Capteur. Les deux premières lignes indiquent la distance spatiale uniquement, les troisième et quatrième lignes indiquent la distance temporelle uniquement, et les deux dernières lignes indiquent la distance spatio-temporelle. Les lignes paires indiquent que le conditionnement repose principalement sur les observations des mêmes capteurs. Les rangées impaires indiquent des conditionnements quel que soit le capteur. Les cercles noirs représentent les observations ordonnées de 1 à 25, le point triangulaire correspond au numéro 26 et les carrés représentent l'ensemble de conditionnement de l'observation 26.	157
4.3	Kullback-Leibler divergence	161
4.4	Log de l'efficacité relative	162
4.5	Comparaison du modèle proposé avec le modèle classique	166
4.6	Cartographie résultant de l'application du modèle proposé en utilisant l'approximation de Vecchia pour estimer les paramètres . .	167
5.1	Exemple de données récoltées avec le capteur mobile 4 vs données issues du boîtier de référence durant le tour 4.	182
5.2	Emplacement des mesures de trafic routier ainsi que des mesure à l'aide de nos boîtier mobiles.	185
5.3	Production d'une cartographie spatio-temporelle des concentrations de PM dans le quartier des Halles en utilisant le modèle hiérarchique (5.3) combinant données issues de capteurs bas coût et de capteur de référence.	187
5.4	Production d'une cartographie spatio-temporelle des variances des prédictions de PM dans le quartier des Halles en utilisant le modèle hiérarchique (5.3) combinant données issues de capteurs bas coût et de capteur de référence.	193
5.5	Étude de l'impact du nombre et du type de capteurs sur la variance de la prédiction.	197

Liste des tableaux

2	Air quality mapping studies using mobile sensors.	xxxi
1.1	Tableau des normes et réglementations des principaux polluants .	17
1.2	Tableau récapitulatif des principaux capteurs bas coût de PM, issu de la publication de Rai et. al[68]	21
2.1	Études de cartographie de la qualité de l'air utilisant des capteurs mobiles.	64
3.1	Différents modèles et leurs paramètres respectifs.	97
3.2	Table récapitulative des covariables utilisées.	115
3.3	Tableau récapitulatif des résultats des différents indicateurs de performance sur les deux modes de validation et dans les deux scénarios, pour les dix modèles étudiés. L'écart type des logarithmes des concentrations de PM ₁₀ étudiés est de 1.03.	129
4.1	Résultat de la procédure d'estimation des paramètres pour le mo- dèle de krigeage avec dérives externes 4.9 et le modèle hiérar- chique à 3 niveaux proposé 4.10.	165
5.1	Résumé des covariables utilisées dans cette étude.	183
5.2	Estimation des paramètres de la moyenne pour le jeu de données d'Atmotrack utilisant le modèle hiérarchique (4.10).	190
5.3	Estimation des paramètres de la fonction de covariance exponen- tielle spatio-temporelle.	190

Liste des Acronymes

- AASQA** Associations Agréées de Surveillance de la Qualité de l’Air. 51
- ANN** réseaux neuronaux artificiels. 59
- ARE** Efficacité relative asymptotique. 158
- CAMS** Service de surveillance de l’atmosphère de Copernicus. 113
- CO** monoxyde de carbone. 5
- CO₂** dioxyde de carbone. 5
- COV** Composés Organiques Volatils. 7, 8
- D KL** Divergence de Kullback-Leibler. 158
- EQM** Erreur Quadratique Moyenne. 194
- Fr** France. 15
- GAM** Modèle Additif Généralisé. 118
- GP_ML** Processus Gaussien avec l’algorithme de maximum de vraisemblance.
119
- GP_VG** Processus Gaussien avec l’algorithme de variogramme spatio-temporel.
119
- GPS** Global Positioning System. 20
- HAP** Hydrocarbures Aromatiques Polycycliques. 9
- IDW** Pondération inverse à la distance. 54
- KED** Krigeage avec Dérives Externes. 89
- LR** Régression Linéaire. 117
- LUR** Land-Use-Regression. 57

- MAE** erreur absolue maximale. 116
- MMD** Distance Maximale Minimale. 153
- NMDA** N-méthyl-d-aspartate. 11
- NOx** oxydes d'azote. 5
- NR** Régression en Réseau. 117
- O₃** Ozone. 7, 18
- OCDE** Organisation de Coopération et de Développement Économiques. 6
- OK** Krigeage Ordinaire. 55
- OMS** Organisation Mondiale de la Santé. 5
- PCB** polychlorobiphényles. 8
- PCDD** dibenzo-dioxines polychlorées. 8
- PCDF** dibenzo-furanes polychlorés. 8
- PM** Particules de Matière. 9
- RF** Forêts Aléatoires. 119
- RMSE** Racine de l'erreur quadratique moyenne. 81
- SK** Krigeage Simple. 89
- SO₂** dioxyde de soufre. 5
- SVM** Machine à Vecteurs de Support. 121
- TPS** Thin Plate Spline. 79
- UE** Union Européenne. 15

Liste des symboles

Symbole	Description
β	Paramètres des covariables
γ	Autocorrelation des capteurs
λ	Poids des observations
μ	Moyenne du champ spatio temporelle
ν	Champ aléatoire de moyenne nulle
ϕ	Paramètre de régularisation
σ^2	Variance
τ^2	Pépite/ variance du capteur
τ_F^2	Pépite/ variance du capteur de référence
θ_1	Paramètre spatial de la fonctions de covariance
θ_1	Paramètre temporel de la fonctions de covariance
Θ	Paramètres de la fonctions de covariance
a	Biais additif
b	Biais multiplicatif
C	Constante spécifique
$C()$	Fonction de covariance
d	Distance
E	Graphe des rues connectées
F	Mesures issues des stations fixes de références
F_m	Mesures issues du capteur mobile de références
I	Matrice identité
K	Matrice binaire d'appartenance de l'observation au capteurs
L	Vraisemblance
M	Taille de l'ensemble de conditionnement
N	Distribution gaussienne
n	Dimension du jeu de données
p	Dimension de la matrice de covariable

Suite à la page suivante

Symbole	Description
$p(\cdot)$	Densité de probabilité
$p'(\cdot)$	Approximation de la densité
R	Matrice de corrélation
t	Coordonnées temporelle
$v(i)$	Sous vecteur conditionnant de la i ème observation
(x, y)	Coordonnées spatiales
X	Matrice des covariables
Y	Fonction aléatoire des vraies concentration des polluants
Y^*	Valeur du vrai signal à l'endroit non échantillonné
y	Réalisation de la fonction aléatoire des vraies concentration des polluants
Z	Fonction aléatoire des concentration des polluants mesurées par les capteurs bas coût
z	Mesures des capteurs bas coût

Summary of the thesis in English - Résumé de la thèse en anglais

Table of contents

Air under surveillance : atmospheric quality, sensors and mapping . . .	xxiii
Towards better pollution mapping : exploring existing models and techniques	xxvii
From fundamentals to advanced techniques : a comparison of pollution models	xxxii
Geostatistical model enhancement : autocorrelation, approximation and application	xxxvi
Measure synergy : Practical strategies for better air quality forecasting .	xlili
Conclusions and outlook	xlvi

This document is structured around 6 chapters. Each chapter presents a stage in the research work carried out during this thesis. The focus of this thesis is on the creation and analysis of statistical models that use data from low-cost mobile sensors to generate accurate predictions and estimates of air pollution maps. This project emerged from a collaboration between ESTACA Engineering School and Gustave Eiffel University (formerly known as IFSTTAR).

Air under surveillance : atmospheric quality, sensors and mapping

The harmful effects of air pollution are well known and the World Health Organisation (WHO) is constantly warning of its harmful effects on human health. According to the WHO database released in April 2022, 99% of the world's population lives in places where air pollution levels exceed WHO recommended limits¹.

The organization has also declared that air pollution is the leading environmental cause of death, responsible for 7 million premature deaths worldwide in 2018, with 4 million due to outdoor air pollution.

Addressing urban air pollution involves understanding a variety of pollutants. **Gaseous pollutants** like sulfur dioxide, nitrogen oxides, carbon monoxide, ozone, and volatile organic compounds primarily result from fossil fuel combustion., affecting the atmosphere's composition and posing health risks, including respiratory issues and blood complications [39]. **Persistent organic pollutants** are toxic chemicals, including dioxins and polychlorobiphenyls, which originate from incomplete combustion and accumulate in the food chain [72]. **Heavy metals**, such as lead and mercury, naturally occur but can enter our food and water supply, accumulating in the human body and posing health risks [38]. Lastly, **Particulate Matter (PM)**, varying in size, originates from both natural and human activities, with finer particles posing greater health risks [63].

Scientific literature is abundant with studies proving the relationship between air quality and its negative effect on the human body. Numerous studies, for example, have found a link between air quality deterioration and respiratory disorders such as asthma, bronchitis and lung cancer [88, 17, 67, 13, 15, 47]. Other studies investigate the negative impact on the cardiovascular system [2, 60, 20, 9] and the nervous system [45, 46]. Air pollution can also have a negative impact on other parts of the human body, such as the skin, reproductive system or the immune system [68, 12, 25].

A lengthy, multistep procedure involving scientists and politicians is required to minimize air pollution in major centers and its detrimental impact on people's health and the environment. The intricate nature of air pollution means there's no singular, swift solution. Air quality is affected by various pollutants from different sources, demanding thorough analysis and tailored measures.

1. Source : <https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data>

While scientists play a crucial role in measuring, understanding, and proposing solutions for pollution, politicians are needed to implement policies based on these findings. However, challenges arise when combating pollution conflicts with goals like economic growth. Continuous air quality monitoring is vital for effective pollution reduction. This thesis contributes to evaluating and mapping atmospheric pollution using innovative measurement methods, combining mathematical models and technological advancements.

Currently, air quality monitoring is carried out using fixed air quality monitoring stations. These stations are managed by national environmental protection agencies. These reference stations provide a very precise measurement of air quality at the cost of limited spatial coverage. The stations can generate detailed time series data, but only at limited locations. This makes it difficult to compile reliable and representative information for a city or a region as a whole, and therefore a more macroscopic view of trends in pollution fields is provided.

The idea of using new Low Cost Sensors (LCS) developed from recent technological advances, smaller, integrating Global Positioning System (GPS) technology has quickly emerged. This offers scientists additional tools to refine spatial-temporal maps of air pollution, and create new data sets that provide air quality information that was not previously available from the limited number of stationary monitors in the city.

Numerous instruments from commercial entities, non-profits, and startups have entered the market to date [5]. While many studies have used mobile monitoring to assess atmospheric pollution exposure, few have focused on mapping urban air pollution using low-cost sensors at a detailed spatio-temporal scale. Transitioning from low-cost sensor readings to detailed urban pollutant mapping requires statistical or physical models. This complex process, central to our thesis, aims to inform the public about potential exposure levels, identify pollution sources, and refine statistical models for precise urban air quality mapping using affordable mobile sensors.

Generating precise air quality maps using low-cost sensors presents several significant challenges. These challenges stem from the nature of the phenomenon under study, the accuracy of the data, and the sheer volume of the data. The primary challenges are :

- **Complexity at a micro scale** : Air quality is influenced by numerous factors like wind flow, traffic, and local weather conditions, which can vary significantly over short distances, making neighborhood-scale mapping intricate and often influenced by site-specific factors.

- **Challenges with Mobile Sensor Data :** The primary limitation of low-cost sensors is their accuracy. Regardless of the statistical model used, if the data is biased, the mapping outputs will reflect this limitation, and under-sampling remains a significant issue.
- **Data Quantity Challenges :** The third challenge arises from the complexity and time required to process the massive volume of data. Low-cost sensors can produce data every second, leading to potentially redundant vast amounts of data.

In view of the difficulties outlined above, it is vital to know how to combine all these fuzzy data sources to build up a clear picture of urban pollution. The developed statistical model will aim to include all specifics of the problem to deliver increasingly accurate pollution maps.

More specifically, this research will aim to answer the following questions :

- What are the best statistical models for mapping urban air quality using low-cost fixed and mobile sensors : in terms of prediction accuracy (interpolation and temporal extrapolation), descriptive power, complexity and prediction uncertainty?
- What influence does the ratio between low-cost, mobile and fixed sensors have on the quality of prediction : is it preferable to deploy low-cost fixed or mobile sensors?
- Can a low-cost sensor system eventually replace current air monitoring methods? Can we obtain the same assessment accuracy as a fixed reference monitoring station?

To address these research questions, a two-part methodology is proposed (Figure 1). The first part involved identifying all model families from existing literature (Chapter 2). This phase entailed studying these families, such as geostatistics, and selecting models best suited to our application, which led to a first publication. Furthermore, we implemented and compared these models using real data from low-cost mobile and fixed sensors. After defining performance indicators, we compared existing models in various scenarios, culminating in a second publication, detailed in Chapter 3.

The second part of the methodology focuses on developing new models to surpass existing ones. We began by creating models to address the shortcomings of previously examined ones. This deep dive into the statistical foundations led to the proposal of novel model, resulting in a third publication. This new model, tailored to handle noisy and abundant data, aims to map urban pollution with optimal spatial and temporal precision in real-time, with findings presented in Chapters 4 and 5.

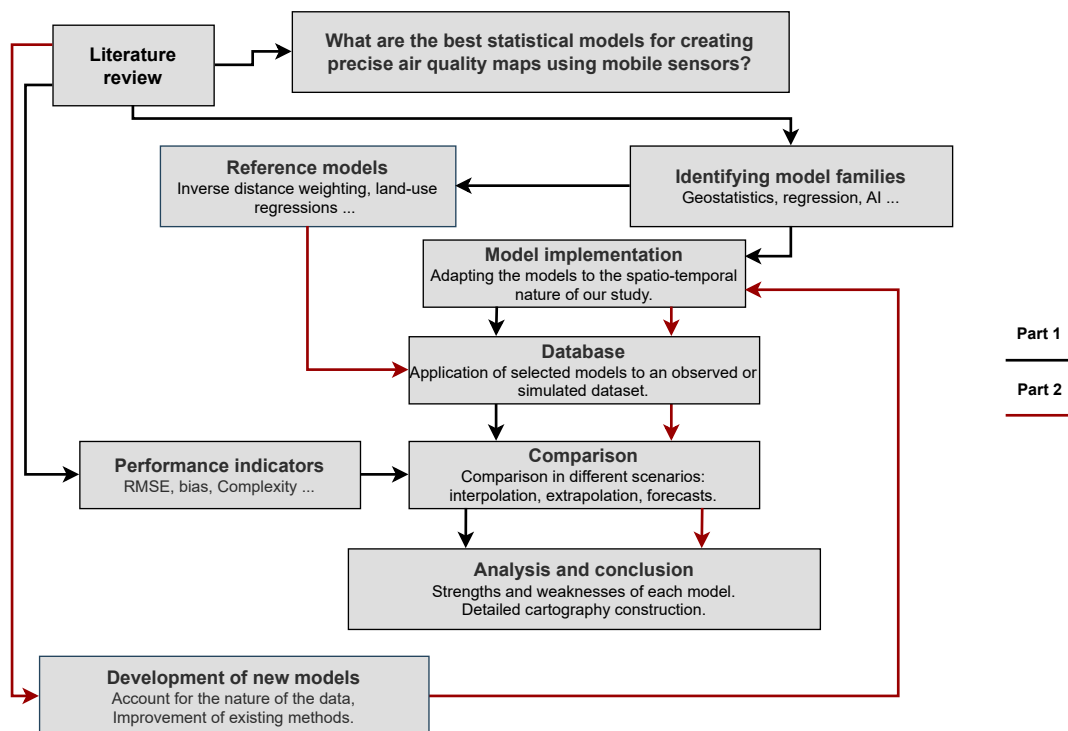


Figure 1 – Methodology used to answer the research questions.

Towards better pollution mapping : exploring existing models and techniques

The first step in our methodology is to carry out an exhaustive review of the methods and models available in the literature for air quality mapping. This chapter is devoted to this objective.

Air quality mapping currently relies on data from strategically placed fixed monitoring stations, representing both urban and rural areas. These stations provide continuous and reliable measurements of key atmospheric pollutants. Satellite data offers broader pollutant concentration views, albeit with lesser spatial resolution. Numerical dispersion models, integrating meteorological parameters and emission data, are commonly used to predict pollutant spatial distribution, often validated or adjusted based on ground observations. At a continental scale, such as Europe, these models are predominant. Systems like PREV'AIR, led by INERIS and supported by entities like Météo-France, combine ground station data, satellite observations, and numerical models to produce air quality maps. Airparif, monitoring air quality in Île-de-France, uses a similar approach. While these maps provide valuable general insights, increased spatial precision is needed for specific needs, as current models may not detail street-level variations. Alternative methods, categorized into four distinct families, are explored in academic literature to address this issue :

Deterministic interpolation techniques

Deterministic interpolation methods generally share the same mathematical foundation, estimating values at unmeasured locations as a weighted average of measurements from surrounding monitoring stations. They differ in sampling weight choices and station selection. Three commonly used interpolation techniques for air quality mapping are spatial averaging, nearest neighbor, and inverse distance weighting (IDW).

Spatial averaging calculates the mean pollutant measurements from nearby monitoring stations within a predefined grid or region, assuming equal influence from all stations regardless of their distance to the unmeasured location. This approach is unrealistic for air pollution estimation as it overlooks spatial variability [74, 75, 44]. The nearest neighbor method assigns pollutant measurements from the closest monitoring station to the unmeasured site, regardless of the distance [6, 59]. Inverse distance weighting, a popular deterministic interpolation technique, uses a weighted average of data from deployed sensors to determine

values at unsampled locations, giving more weight to closer locations. Due to its ease of use, IDW is often used as a benchmark [57, 86].

The limitation of deterministic interpolation techniques is their low extrapolation accuracy. They aren't considered models as they don't describe data or indicate prediction uncertainty. Due to these limitations, they've become less popular for atmospheric pollution estimation but still appear in comparative studies of spatial interpolation methods [16, 78].

Geostatistics

Geostatistical methods, like deterministic interpolation methods, operate on the principle that nearby measurements have more influence. However, in geostatistics, measurements are treated as random variables. Geostatistics predicts values at unsampled locations using a weighted combination of observations, with weights determined by the data's covariance structure. Two main forms of kriging are commonly used : ordinary kriging and kriging with external drifts. Ordinary kriging assumes an unknown constant mean in a local neighborhood, while kriging with external drifts a describable polynomial general trend model.

Scientific literature is abundant with studies where geostatistical models have been successfully applied for air quality mapping. For instance, Kim et al.[43] developed an ordinary kriging prediction model for long-term particle concentrations in seven major Korean cities. Whitworth et al.[84] modeled benzene levels in ambient air using geostatistics. Schneider et al.[73] modeled nitrogen dioxide concentration in Oslo using dispersion model output as a covariate in a geostatistical model. Liu et al.[51] studied spatial ozone variation over 12 hours using measurements from 19 monitoring stations in Toronto, Canada, finding kriging more accurate than the nearest neighbor method. Ferreira et al.[22] measured atmospheric pollution concentrations in central Lisbon using ordinary kriging from nine monitoring stations, correlating road traffic with pollutant levels. Janssen et al.[37] applied a geostatistical model to estimate pollutant concentration levels on a 4x4 km² grid in Belgium, outperforming other interpolation techniques.

More sophisticated than deterministic interpolation approaches, geostatistics also provides associated prediction uncertainty. However, these techniques come with a relatively high computational cost.

Land use models

Land-use models assume that local environmental factors, such as land use, meteorological factors, building density, and traffic density, solely influence air quality in a specific location. These models establish a link between selected environmental predictive factors and available air quality measurements. Although many covariables can be integrated, the term "land-use models" remains predominant.

Land-use models have two sub-families : regression models and machine learning.

A - Regression Models :

Land-Use-Regression (LUR) models assume a simple relationship, often linear, between predictive variables and pollutant measurements. The SAVIAH project [7] was the first to model small-scale atmospheric pollutant variations using LUR, proving its robustness. The choice of variables is crucial in LUR models. Ross et al. [69, 70] demonstrated the importance of traffic variables in their models. Zhai et al.[89] emphasized challenges related to spatial scale, noting the phenomenon of spatial non-stationarity. Despite the popularity of LUR methods, they assume a unrealistic relationship between the variable of interest and predictive variables.

B - Machine Learning :

When it comes to obtaining new information from large, varied and dynamic data sets, machine learning proves particularly useful. A machine learning algorithm analyzes training data (the same data used by LURs) and produces an inferred function that can be used for prediction in unsampled locations. In general, this function is quite complex and cannot be interpreted directly by the user.

Several studies have applied machine learning to predict air pollution levels, such as [76, 11, 10, 87]. Brokamp et al. [8] highlighted that combining random forests with significant variables can surpass LUR limitations. However, machine learning algorithms are often seen as "black boxes" with low descriptive capabilities.

In general, land-use models heavily rely on the availability of other variables. They overlook spatial effects like non-stationarity and spatial and temporal autocorrelation.

Dispersion model

Dispersion models, often used on a large scale like a region or country, can also be adapted to smaller scales, such as a city. These models replicate how physical and chemical processes lead to the production of atmospheric pollutants. They incorporate environmental factors, like those used in LUR models, and are frequently applied to predict traffic-related pollution. Implementing these models requires a deep understanding of the involved processes. For pollutants like particles, especially in environments with pronounced topography and different sources, the physical mechanisms are both extremely complex and not fully known.

Deterministic dispersion models are categorized into three approaches based on their calculation method : Eulerian, Lagrangian, and Gaussian models. Simplified, these models require different input datasets (pollution sources, geography, meteorology, etc.) [77]. They solve equations governing atmospheric phenomena, providing pollutant field maps and forecasts.

Vardoulakis et al. [82] studied four Gaussian dispersion models used in street canyons. Hamer et al. [29] described the Eulerian dispersion model EPISODE and its application to modeling nitrogen dioxide concentration. Fallah et al. [21] used a Gaussian dispersion model to enhance the characterization of atmospheric pollution near roads. Gibson et al. [24] also utilized the Gaussian dispersion model to assess particle emissions, NO_x, and nitrogen oxides.

The Eulerian deterministic chemistry-transport model CHIMERE [53], widely recognized for its scientific quality, is frequently used in France, especially by PREV'AIR. It calculates hourly concentrations of several pollutants. These calculations are performed on domains ranging from regional to continental scales.

The main advantage of dispersion models is that they don't require a dense network of monitoring stations. However, their limitations mainly stem from input data and model postulates. Acquiring a variety of input data can be costly and sometimes demand significant computational resources. Moreover, some assumptions these models rely on might be unsuitable or specific to local contexts, making modeling small-scale random fluctuations challenging.

Mapping urban air quality using mobile sensors

The use of fixed stations has been the traditional method for spatialization techniques. With the emergence of low-cost sensors, there was an inclination to adapt these established methods to the new data. However, the challenge lies in the inherently different nature of spatio-temporal observations from mobile low cost sensors. Unlike fixed stations, mobile sensors don't consistently cover all areas and lack continuous data for specific locations, especially during low traffic or nighttime. Zwack et al. combined mobile surveillance data-based regression models with a dispersion model called "Quick Urban Industrial Complex" (QUIC) for air pollution assessment [92]. Adams and Kanaroglou integrated LUR modeling with neural networks to combine mobile and stationary surveillance data for air pollution forecasting [1]. Gressent et al. mapped Nantes city using Atmotrack data, incorporating dispersion model outputs into a geostatistical model [26]. Hasenfratz et al. employed non-linear LUR models to develop a high-resolution air pollution map for Zurich [32]. Table 2 lists published studies using low-cost mobile sensors for air quality mapping.

Article	Method	Region	Pollutants	Vehicle
Marjovi et al. [55]	LUR	Lausanne, Switzerland	UFP	Bus
Hart et al. [31]	LUR	Texas, USA	PM2.5	Bicycles
Apte et al. [3]	Deterministic Methods	Oakland, USA	NO, NO2, BC	Cars
Hasenfratz et al. [33]	LUR	Zurich, Switzerland	UFP	Trams
Hasenfratz et al. [32]	LUR	Zurich, Switzerland	UFP	Trams
Marjovi et al. [56]	LUR	Lausanne, Switzerland	UFP	Bus
Li et al. [48]	Geostatistics	Zurich, Switzerland	UFP	Trams
Lim et al. [49]	Machine Learning	Seoul, South Korea	PM2.5	Pedestrians
Adams et al. [1]	Machine Learning	Hamilton, Canada	NO2	Vans
Hankey et al. [30]	LUR	Minneapolis, USA	BC PM2.5	Bicycles
Gressent et al. [26]	Geostatistics	Nantes, France	PM10	Cars
Do et al. [18]	Machine Learning	Antwerp, Belgium	Various pollutants	Bicycles
Zhang et al. [90]	Machine Learning	Songdo, South Korea	CO2, PM2.5, PM10	Cars
Song et al. [79]	Machine Learning	Beijing, China	PM2.5	Cars
Van et al. [81]	LUR	Ghent, Belgium	BC	Bicycles
Guan et al. [27]	Geostatistics	Oakland, USA	NO2	Cars
Mariano et al. [54]	Machine Learning	Zurich, Switzerland	UFP	Trams
Ma et al. [52]	Machine Learning	China	PM2.5	Cars
Idir et al. [36]	Geostatistics	Zurich, Switzerland	O3	Tram
Zhao et al. [91]	Machine Learning	China	PM2.5	Cars
Qin et al. [65]	Machine Learning	China	PM2.5	Cars
Qin et al. [64]	Machine Learning	Antwerp, Belgium	NO2	Vans
Messier et al. [58]	Geostatistics	Oakland, USA	NO2 BC	Cars

Table 2 – Air quality mapping studies using mobile sensors.

From fundamentals to advanced techniques : a comparison of pollution models

This chapter is devoted to the evaluation and implementation of statistical models for estimating and predicting air pollution. We adopt a progressive methodological approach, starting with simple models and gradually increasing their complexity to better address our problem.

Methods using only air pollution data

The first part of this chapter focuses on exploring methods that exploit only air pollution data. This includes data from both low-cost mobile sensors and fixed air quality stations, which serve as a reference. The choice of this approach lies in its simplicity and ease of implementation, providing an initial view of the structure and dynamics of air pollution.

Three geostatistical approaches were studied. Apart from mobile sensors data, two of them used fixed station data to predict air quality. Each of these three methods make different assumptions :

- Simple Kriging (SK) with a varying known mean : the time series of the fixed monitoring station was chosen to be the overall mean.
- Ordinary Kriging (OK) with a constant piecewise mean, but unknown.
- Kriging with external drift (KED) : the data from the fixed monitoring station was used to estimate the underlying mean.

The three approaches were compared to one deterministic interpolation technique, here considered as the reference (IDW), in three scenarios. The evaluation of the result of each of them used the following three performance indicators : RMSE, Bias and correlation.

Regarding cross-validation, three distinct scenarios for random sample selection were defined. These scenarios, related to segmenting the data into training and validation sets, aim to determine the ideal adaptation conditions for each model :

- Randomized : Points picked without considering location or time, aiming

for data reconstruction.

- Random paths : Varied-length paths chosen, mimicking mobile sensor data and emphasizing nearby extrapolation.
- Tramway paths : Uses data from specific tram routes, targeting extrapolation to farther areas.

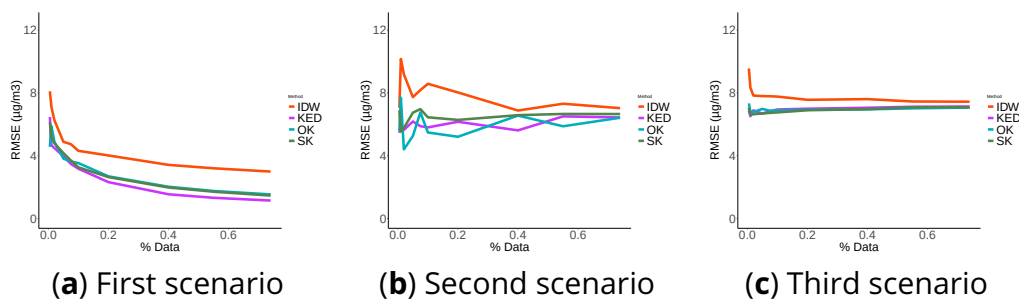


Figure 2 – RMSE as a function of the percentage of data used in the three scenarios studied.

RMSE results are presented in figure 2. Kriging with external drift, which appears to be the optimal choice after this first stage of comparison. It offers versatility and the ability to integrate other data sources in the form of drifts. Moreover, it provides uncertainty estimates that can guide sensor placement, and its descriptive nature promotes a better understanding of the phenomenon under study.

Although univariate geostatistics has its advantages, the integration of multivariate geostatistics could offer interesting prospects for future research. This will require a careful comparison of several methods in terms of complexity, prediction error, data used. The next logical step would be to make the model more complex by incorporating covariates, in order to further optimize air quality prediction.

Integration and comparison of models using external data

In this part, we extend our exploration of mathematical models dedicated to urban air quality mapping by integrating external data. The aim is to assess the potential positive impact of this additional data on prediction accuracy.

We integrate thirteen auxiliary variables into ten separate models drawn from various approaches in the literature. The effectiveness of each model in incorporating these variables to refine predictions of urban air quality will be assessed.

Different scenarios are studied :

- Temporal Interpolation : Use day D's data to estimate pollution for day D.
- Forecasting : Use day D's data to predict pollution for day D+1.

In each of the two scenarios, two types of assessment will be carried out :

- Cross-validation on low-cost sensors.
- Validation on fixed stations.

The ten compared models are :

- Inverse distance weighting (IDW).
- Linear regression (LR).
- Linear regression on network (NR).
- General additive model (GAM).
- Geostatistics using spatio-temporal variograms (GP_VG).
- Geostatistics using maximum likelihood estimation (GP_ML).
- Random Forests (RF).
- Artificial Neural Networks (ANN).
- Gradient boosting (XGboost).
- Support vector machine (SVM).

The results show that in the case of cross-validation on low-cost sensors in the interpolation scenario, all models are roughly equivalent and offer excellent results, with equivalent RMSEs, and equivalent correlations ranging from 0.75 to 0.86. Furthermore, in this case, neither the bias nor the MAE indicates a faulty model.

In the case of low-cost sensor forecasting, RMSE scores decrease for all models, as do correlations, which now range from 0.15 to 0.43. Examination of the bias and MAE concludes that the ANN, GAM and GP_VG models are unsuitable for forecasting purposes.

Concerning validation on fixed air quality stations, the RF, XGBOOST and ML_GP models offer the best results on all performance indicators, with correlations above 0.4 for all three models. We also note that the GAM and ANN models perform significantly worse.

Finally, with regard to air quality station forecasts, the study of performance indicators shows that all models are inadequate.

Overall, the machine learning models, with the exception of ANN (XGBOOST RF and SVM), appear to be better suited to the forecasting scenario. As expected, we make more errors in the forecast scenario, whether validated on LCS or fixed stations.

The errors are much greater when we validate on fixed stations, which may indicate a difference in the nature of the data collected by the LCS and those from fixed stations.

Figure 3 displays the maps generated by the ten different models at 12 :00 on 26/11/2018, using the data from Atmotrack, deploying a dozen fixed and mobile low-cost sensors in the city of Nantes, more information on the dataset is available in the study by Gressent et al. [26].

Upon completion of this comparative analysis between models, the geostatistical model (kriging with external drift) with maximum likelihood estimation approach was selected for the continuation of this thesis, as it addresses all the research questions posed in Chapter 1. However, its computational complexity represents a limitation for its use with mobile low cost sensors.

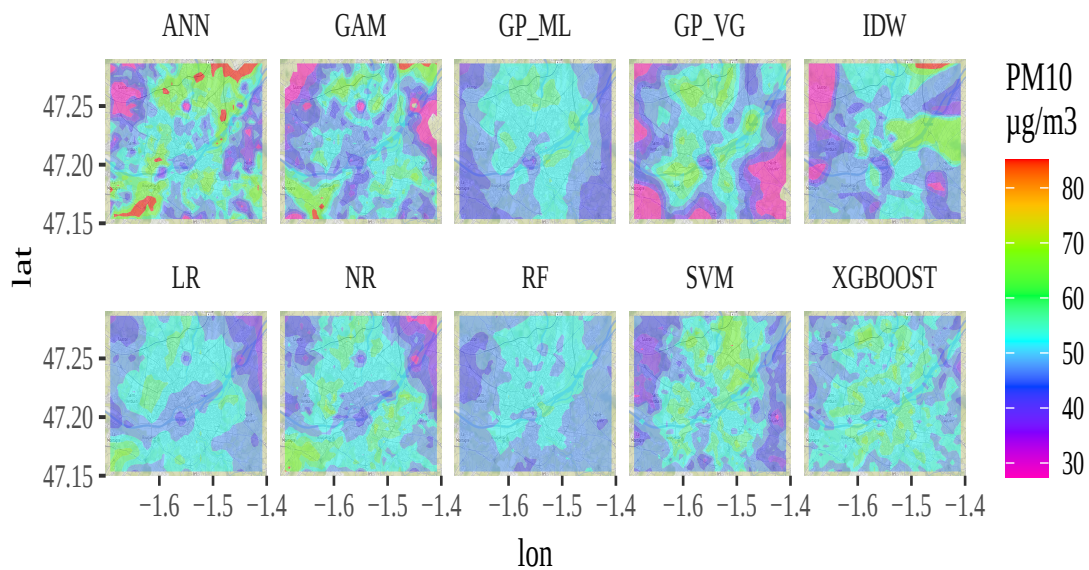


Figure 3 – Comparative mapping of PM10 concentrations according to different models

Geostatistical model enhancement: autocorrelation, approximation and application

The Geostatistical model using gaussian process with maximum likelihood was deemed relevant for predicting air quality and addressing atmospheric pollution mapping challenges. However, its computational complexity and the need to adapt it to low-cost sensor data were identified as significant hurdles. This chapter addresses these two challenges. Firstly, the geostatistical model is re-defined using a hierarchical approach, this approach will be useful for modelling the autocorrelation of the low cost sensors, enhancing prediction accuracy. To reduce the computational cost of the Gaussian process model, approximation techniques for maximum likelihood are explored. The popular Vecchia approximation method [83] is adapted and tested on real trajectory-simulated data. Finally, this approximation method is applied to Atmotrack data, offering a detailed atmospheric pollution map.

Modeling sensors autocorrelation using hierarchical approach

One way of defining the geostatistical model other than by solving the kriging equation system is via hierarchical modeling. Let's put $Z(x) = (z_1(x_1), z_2(x_2), \dots, z_n(x_n))$

the data observed by our low-cost sensors, indexed over the time space x (we use a single variable instead of three (x, y, t) to lighten the notation). The vector of pollutant concentrations (called the true signal) at the same locations is $Y = (y_1(x_1), y_2(x_2), \dots, y_n(x_n))$. The difference between Y and Z is the sensor error.

Assuming the process of interest (the true signal) follows a multivariate normal distribution with mean βX and covariance matrix $\sigma^2 R$, where R is a matrix $n \times n$ with elements $r_{i,j} = \rho(x_i, x_j)$. This gives :

$$\begin{aligned} Z/Y &\sim N(Y, \tau^2 \cdot I) \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \end{aligned} \tag{1}$$

The mean is modeled as a linear combination of p covariates. Thus, the matrix X comprises $p+1$ columns and n rows, with an additional column consisting of 1's to represent the constant. The β vector contains $p+1$ parameters, which must be estimated with the covariance function parameters.

This is a two-level hierarchical model :

- The first level encodes the behavior of the sensors (error, autocorrelation).
- The second level encodes the behavior of the pollutants, where the average depends linearly on several covariates (weather, proximity of a green space, etc.). The covariance function encodes pollutant diffusion.

By using the expected mean, we derive the explicit form of the minimum mean square error estimator for $\hat{Y}^*$:

$$\hat{Y}^* = \beta X^* + r^* \cdot (R + \tau^2/\sigma^2 \cdot I)^{-1}(Z - \beta X) \tag{2}$$

In this approach, the fact that the estimator is linear in Z is a consequence of the Gaussian modeling assumption. In classical geostatistics, the starting point is to restrict attention to estimators that linear, and to look for the one that is optimal, in the sense of mean-squared error. The resulting expressions for the simple and ordinary kriging estimators can also be derived as examples of the much older statistical ideas of best linear prediction, and best unbiased linear prediction, respectively. See, for example, Chapter 1 of Stein (1999) [80].

The classical geostatistical model assumes that sensor errors are decorrelated in space and time and between sensors, and that they are centered on the true process Y . It becomes clear that the assumptions underlying the data

model might not suit data collection from multiple low cost sensors.

The low cost sensors are likely to be affected by different measurement errors, potentially correlated spatially and temporally.

When mobile sensors are deployed, some of them may malfunction for a variety of reasons, such as technical problems or blocked air inlets. These malfunctions can lead to errors in the overall measurements for a specific sensor, while the data from the other sensors remains relatively reliable. In such scenarios, the erroneous measurements of one malfunctioning sensor can lead to an overestimation of τ^2 , thus diminishing the relevance of the measurements of all the sensors.

To account for this type of bias, the mixed model framework was used to propose the 3-level hierarchical model. This allows the model to capture the variation and bias of errors for each sensor, for each sensor j :

$$\begin{aligned} Z_i/(Y, a_j) &\sim N(a_j + Y, \tau^2 \cdot I) \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \\ a_j &\sim N(0, \gamma^2) \quad \text{For } j : 1 \dots J \end{aligned} \tag{3}$$

The result is a compact form of the model (3) :

$$Z \sim N(\beta X, \tau^2 \cdot I + \sigma^2 \cdot R(\Theta) + K \cdot \gamma^2 \cdot K') \tag{4}$$

Where K is the binary matrix $n \times J$ which codes the belonging of observation i to sensor j , and K' its transpose. J is the number of sensors.

Study of the Vecchia approximation

The complexity of Gaussian Processes (GP) represents a major challenge, especially when dealing with large data sets, often numbering in the tens of thousands or more. Relative to the number of observations, GP complexity evolves quadratically for memory and cubically for time. This limits their direct application to today's large datasets, such as those derived from high-frequency sampling low-cost sensors.

To solve this "big n problem", a number of approaches have been proposed in the literature. From sparse covariance matrices [23, 42, 19], to sparse inverse covariance matrices (i.e precision matrix) [71, 50, 61], and low-rank matrices [35, 85, 66, 4, 14, 40], the sparse Cholesky decomposition of the precision matrix (introduced by Vecchia) [83, 41] will be the focus of this study. A full review and

comparison of several of these methods can be found in Heaton et al.[34].

Vecchia's approximation replaces the full conditioning vectors $v(i) = 1, \dots, i-1$ with a sub-vector $v'(i)$ (which we call the neighbors of observation i). This sub-vector is often chosen to contain indices corresponding to observations close in distance to the i th observation. The result is a valid joint distribution :

$$p'(Z) = p(z_1) \prod_{i=2}^n p(z_i / z_{v'(i)}) \quad (5)$$

Unlike the true join distribution, $p'(Z)$ now depends on the order in which the observations are classified in addition to the size of the sub-vector v' , which we'll note M .

This research aims to assess the efficacy of various configurations for a spatio-temporal gaussian process integrated with a mixed-effects model parameter. The configurations are strategic permutations of two key elements : the order of observations and the selection of the conditional set of neighbors.

For ordering, six methods are explored :

1. Random ordering.
2. Sorting by temporal coordinates.
3. Sorting by spatial coordinates.
4. Sorting by minimum-maximum distance (MMD).
5. Ordering from the center outwards.
6. Sorting by sensor, grouping observations temporally per sensor.

For the conditional set of neighbors, which refers to the selection of data points considered in the conditional probability calculation, three distance definitions are used :

1. Spatial distance.
2. Spatio-temporal distance.
3. Purely temporal distance.

Additionally, the belonging of observations to a sensor's conditioning set can be either sensor-dependent or independent.

By combining the six ordering strategies with the six neighbor selection criteria, 36 configurations are derived for the Vecchia approximation. Each configuration's performance is evaluated by progressively increasing the size of the conditioning set and calculating selected performance indicators. The illustrative figure 4 showcases a fictional example in a spatio-temporal space with data

from six sensors.

The results show the superiority of the min-max distance algorithm for the ordering choice, supporting the findings of other studies in the literature [28]. We also prove the effectiveness of the random ordering algorithm, which doesn't require the prior definition of a spatio-temporal distance. These two ordering configurations scored, on average, 102% better results in the log Kullback-Leibler divergence compared with the four other ordering schemes studied. These results are supplemented with a study of asymptotic relative efficiency, providing guidance for optimizing the quality of parameter estimation.

The last part of this chapter is dedicated to a successful numerical analysis, where the proposed model and the chosen Vecchia approximation are applied to a real dataset of air quality obtained from low-cost mobile and fixed sensors for pollution mapping and prediction, demonstrating its practical applicability (Figure 5).

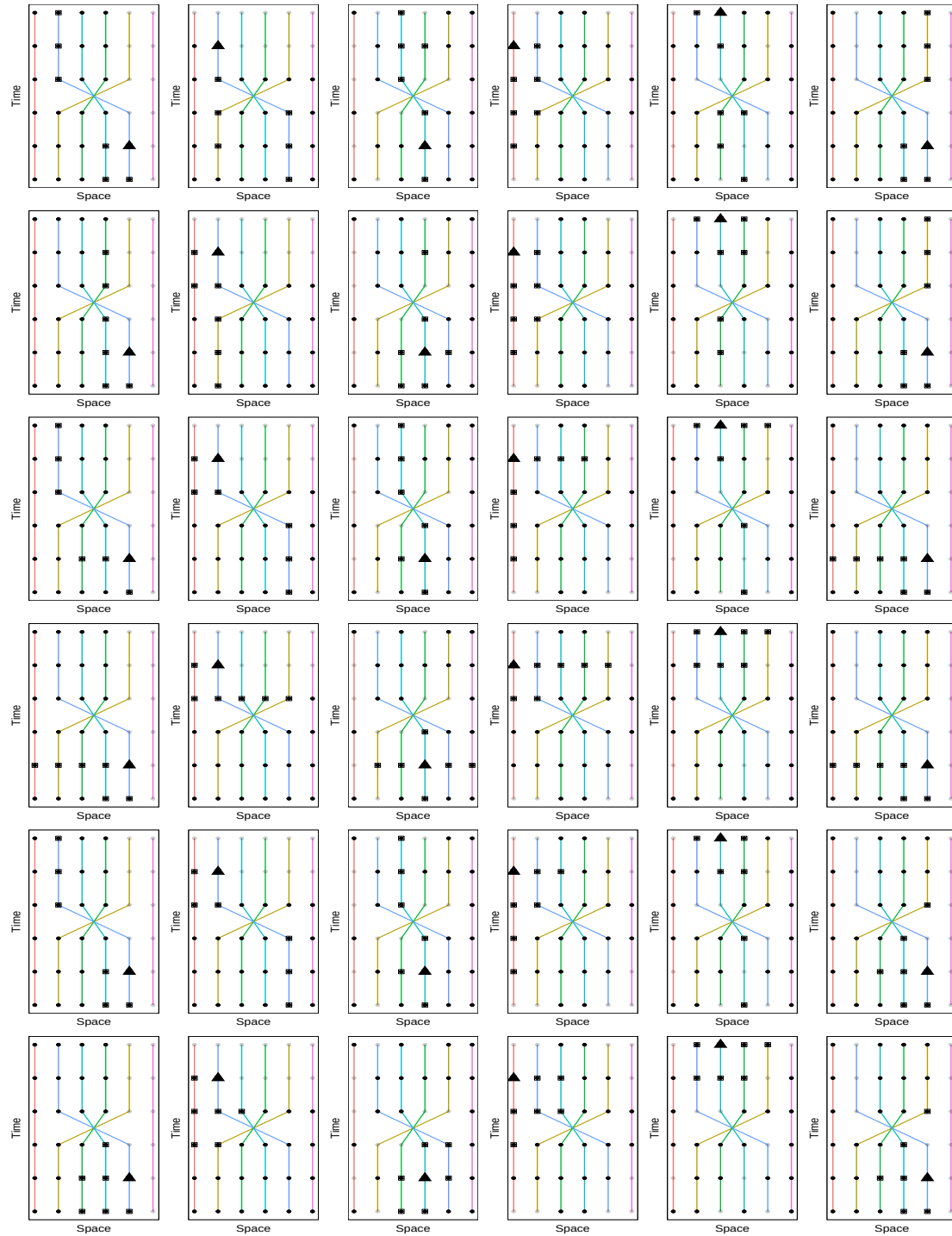


Figure 4 – Configurations studied : the column indicates the order from left to right : Spatial, Temporal, Random, Middle Out, MMD, Sensor. The first two rows indicate spatial distance only, the third and fourth rows indicate temporal distance only, and the last two rows indicate spatio-temporal distance. Even rows indicate conditioning based primarily on observations from the same sensors. Uneven rows indicate conditions regardless of sensor. The black circles represent observations ordered from 1 to 25, the triangular dot corresponds to number 26, and the squared dot represents the conditioning set for observation 26.

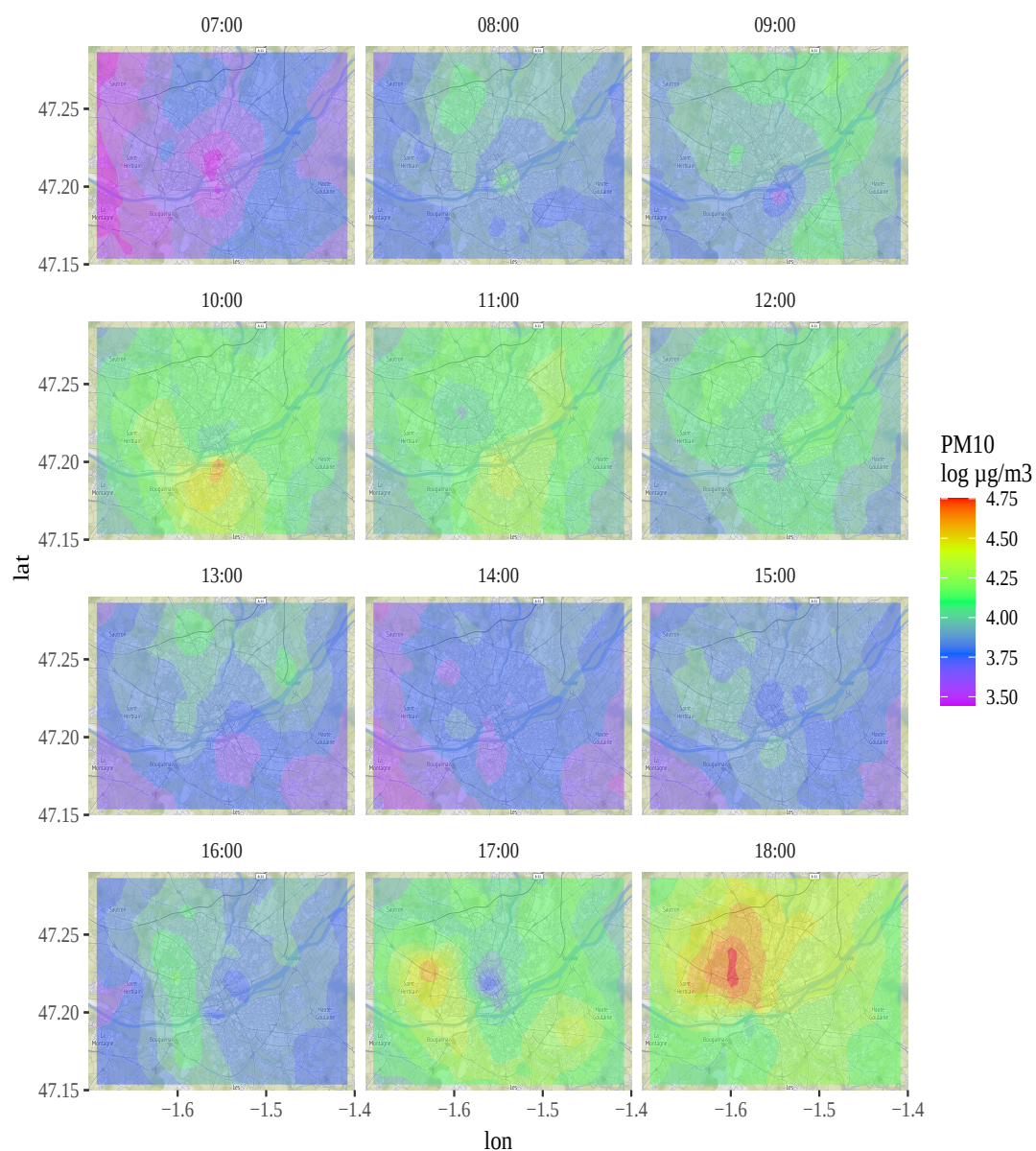


Figure 5 – PM10 Maps resulting from application of the proposed model using Vecchia's approximation to estimate parameters.

Measure synergy : practical strategies for better air quality forecasting

In the previous chapter, the strengths of a new model were emphasized, specifically designed to address the challenges associated with low-cost sensors, which frequently produce data with inconsistent accuracy.

The focus of this chapter is on further optimization, aiming to develop new methods to better handle this imprecise data. Initially, the theoretical limitations of the prior model will be discussed, followed by the introduction of a practical experiment devised to address these challenges.

Secondly, The benefits of mobile versus fixed sensors will be examined in light of the findings from the latest experiments.

Practical extension of the theoretical model

The earlier model 3 was efficient in managing errors arising from using low-cost sensors, outperforming the traditional geostatistical model. However, it has a significant limitation : it requires the biases from the sensors to follow a centered normal distribution, i.e. the sum of the biases is equal to zero. This condition, while crucial for model identifiability, is too restrictive in the context of low-cost sensors.

A more pragmatic approach is proposed through a hierarchical model 6. This model integrates both multiplicative and additive biases of the low-cost sensors. In the absence of additional data, these biases can not be estimated.

$$\begin{aligned} Z(x_i, t_i)/Y &\sim N(a \cdot Y(x_i, t_i) + b, \tau^2 \cdot I) \quad i = 1 : n \\ F(x_i, t_i)/Y &\sim N(Y(x_i, t_i), \tau_F^2 \cdot I) \quad i = 1 : n_F \quad \text{with} \quad n_F < n \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \end{aligned} \quad (6)$$

Yet, when more accurate data F becomes available, which offers better insights to the actual signal Y at the same spatio-temporal points sampled by the mobile sensors, it can be integrated into the model.

This accurate data comes from a mobile device equipped with the Fidas Frog [62], a highly precise portable pollution sensor with GPS functionality. It operates in tandem with a low-cost sensor, enabling calibration and parameter estimation

during deployment.

The Fidas Frog is a high-resolution optical spectrometer capable of rapidly and reliably determining fine dust concentrations with an 8-hour battery life. It simultaneously measures concentrations of various particulate matters and determines particle size distribution ranging from 0.18 to 40 μm (256 channels).

Given the high costs associated with such advanced sensors, it's unfeasible to deploy them as widely as the cheaper ones, reaffirming the indispensability of the low-cost sensors.

This method is analogous to in-situ calibration with reference stations, often referred to as "Rendezvous based calibration". However, in this context, the interval between each meeting is notably long. Another perspective is comparing this to a pre-deployment calibration step, but conducted during deployment to ensure identical conditions, which also allows the measurement of the impact of various environmental variables on the sensors.

In order to calibrate all low-cost mobile sensors with a limited number of reference sensors, an experimental plan has been devised, which will be elaborated upon in the subsequent section.

Experimental plan

Aiming for data with a very fine spatial resolution, five mobile sensors were deployed alongside a reference sensor in the district of "Les Halles" in Paris for one day. This location was chosen due to the presence of a fixed reference station by Air Parif, simplifying data comparison and validation. Additionally, This district combines dense urban spaces, parks and high-traffic avenues, offering a wide variety of configurations in a small area.

The reference sensor must be close to the mobile sensors for a long enough period to precisely estimate the parameters a and b in model 6. Hence, five rounds were planned, each lasting about 45 minutes. In every round, the Fidas Frog (the reference sensor) was deployed next to a different mobile sensor, ensuring a robust parameter estimation for each mobile sensor.

Planning these experiment, special attention was given to the mobile sensors' trajectories. The aim was to maximize the distance between them while ensuring complete spatial coverage of the neighborhood streets. This issue re-

sembled the well-known "Chinese postman" problem, which involves finding the shortest path that crosses every street in a neighborhood at least once.

Subsequently, the estimation of multiplicative and additive biases incorporated into the model was carried out. Calibration was done separately for each sensor to achieve optimal performance across different sensors without altering their respective correlations.

Let $Z'_j(x, t)$ be the raw data from sensor j sampled at location x and time t , $F(x, t)$ the data from the Fidas Frog at the same spatiotemporal locations x, t , and $Z_j(x, t)$ the calibrated data from sensor j .

A linear calibration of the raw data, meant to correct any potential bias, is described as :

$$Z_j(x, t) = a_j + b_j \cdot Z'_j(x, t) \quad (7)$$

In the above equation, the only known term is $Z'_i(x, t)$. Estimating a_j (additive bias) and b_j (multiplicative bias) is essential to obtain calibrated data. This estimation involves $F(x, t)$ as :

$$F(t) = a_j + b_j \cdot Z'_j(x, t) + \epsilon \quad (8)$$

a_i and b_i estimation from the above equation was done using the ordinary least squares method, which minimizes ϵ . This calibration was conducted for each sensor individually, using all data from sensor i from the measurement round when it was in contact with the reference data F from the Fidas® Frog. There were as many a_j and b_j estimations as there were sensors.

The model, along with the practical approach to parameter estimation, can shed light on the questions we formulated at the start of this thesis. It also meets the requirements of complexity and describability, provides information on the errors associated with prediction, and generates accurate air quality maps. In addition, it offers us the opportunity to study the number and type (mobile or fixed) of sensors required to achieve a defined accuracy.

Study on the Number of Sensors

The research investigates the number of vehicles with mobile sensors needed for accurate urban air pollution monitoring. An important question is how a fleet of mobile sensors compares to fixed monitors in terms of spatial prediction

and interpolation.

The mean squared error (MSE) was computed considering different number of both sensor types in two scenarios : spatial interpolation and short-term temporal forecasting.

Since kriging is an unbiased method, MSE is simply the predictive variance. The latter depends only on the position of the observations and the model parameters. It can therefore be obtained without the need for pollution data.

To simulate data similar to fixed sensors, certain locations were chosen as monitoring sites, with the same sampling frequency as mobile sensors for accurate comparison.

Two prediction types, spatial interpolation and a 45-minute temporal forecast, were compared. Results displayed in Figure 6 indicated that prediction variance remains consistent across different sensor types and numbers. However, a slight advantage was observed in favor of mobile sensors in the interpolation scenario. This suggests that more mobile sensors lead to broader spatial coverage, minimizing uncertainty.

Conclusions and outlook

Air pollution is a significant public health issue, and accurate mapping and identification of areas exceeding regulatory thresholds in France and the EU are crucial for protecting citizens' health. This thesis explored atmospheric pollution modeling, including prediction uncertainties and the potential contributions of mobile sensors, both methodologically and practically. Accurate atmospheric prediction is complex, especially for forecasting, even with known variables like weather predictions, due to inherent uncertainties. Machine learning and AI show promise but do not always outperform other methods.

Chemistry-transport models struggle with fine-scale concentration variations due to computational power and data resolution limitations. Geostatistics, particularly kriging with external drift, has proven robust for mapping using in-situ measurements, considering local concentration as functions of chemistry-transport model outputs and environmental variables.

Low-cost mobile sensors, despite their spatial coverage and increased in-situ measurements, face challenges like high malfunction probability and data reliability.

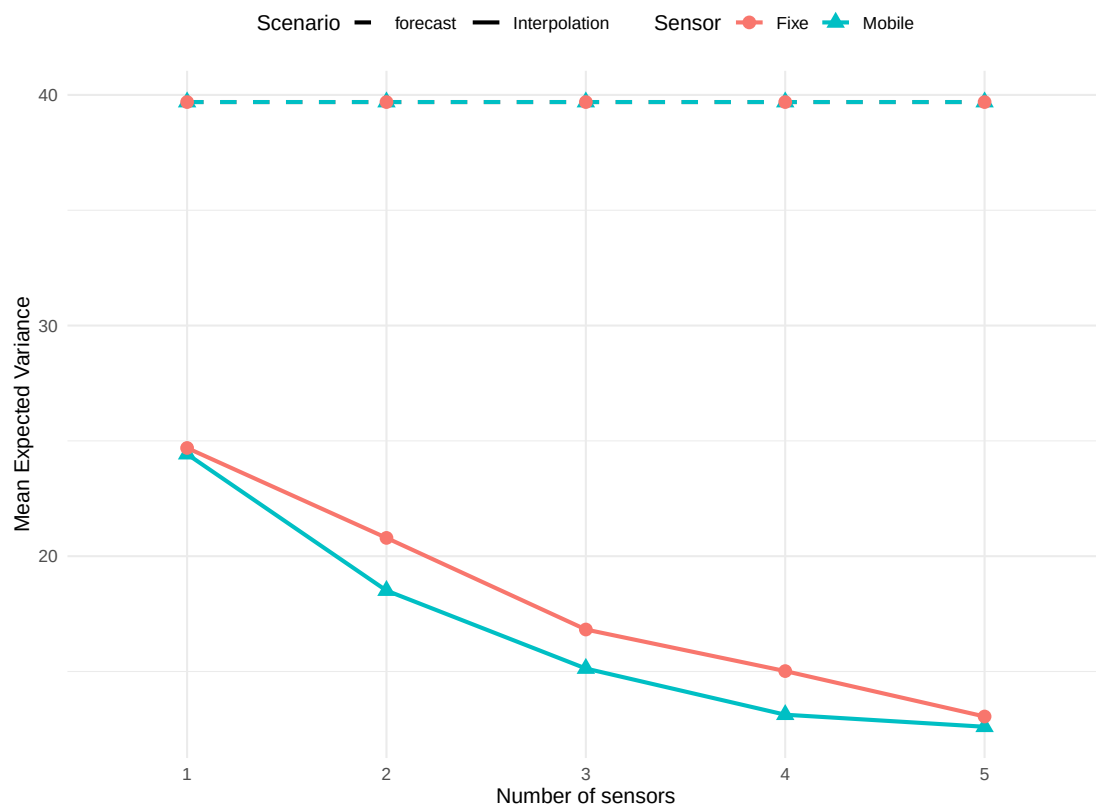


Figure 6 – Study of the impact of the number and type of sensors on prediction variance.

bility. This work focused on developing innovative models that account for the intrinsic error of low-cost sensors, as data reliability remains a barrier to their widespread adoption.

Future challenges include creating hybrid models that innovatively combine physical and geostatistical models, beyond using physical model outputs as inputs for geostatistical models. A stochastic chemistry-transport model that adheres to the behavior and statistics of prior simulations could offer precise mapping with lower computational costs. Advances in solving stochastic partial differential equations with non-separable spatio-temporal covariance functions could theoretically account for pollutant dispersion advection, but the challenge lies in effectively modeling, simplifying, and parameterizing the physics in these simulations.

In addition to modeling efforts, improving the sensors themselves, particularly through calibration, could be key to enhancing the accuracy of low-cost

sensors, as biased data significantly compromises utility.

Chapitre 1

Air sous surveillance : cartographie, capteurs et qualité atmosphérique

Points clés du chapitre

- Constat de 7 millions de morts prématurées dues à la pollution.
- Émergence des technologies bas coût encore peu exploitées pour l'estimation spatio-temporelle précise de la pollution.
- Définition de la question de recherche : quels sont les meilleurs modèles statistiques pour cartographier la qualité de l'air à l'aide de capteurs bas coût ?
- Définition de la méthodologie : comparaisons de différentes familles de modèles sur des jeux de données réels et comparaisons de leurs performance puis proposition d'un nouveau modèle.

Contenu du chapitre

1.1	La pollution atmosphérique : définition et méthodes de mesure . . .	5
1.1.1	Définitions et origines de la pollution atmosphérique	5

1.1.2	Principaux polluants	7
1.1.3	Impacts sur la santé	10
1.1.4	Surveillance de l'air : étape clé contre la pollution atmosphérique	12
1.1.5	Méthodes actuelles d'évaluation de la qualité de l'air	13
1.1.6	Émergence des capteurs bas coût	16
1.1.7	Technologie des capteurs bas coût	18
1.1.8	Spatialisation des données : de la mesure à la cartographie	22
1.2	Stratégies de recherche	24
1.2.1	Défis de la thèse	24
1.2.2	Questions de recherche	26
1.2.3	Méthodologie	27
1.2.4	Jeux de données utilisés durant la thèse	28
1.3	Conclusion	38
	Références	39

Ce premier chapitre décrit le contexte dans lequel s'inscrit cette thèse. Il offre une vue d'ensemble du sujet et introduit la problématique de recherche. On débutera par une définition de la pollution atmosphérique, identifiant les principaux polluants et leurs effets néfastes sur la santé. L'analyse portera également sur l'importance de l'évaluation et des mesures dans le processus de réduction de ce phénomène nocif.

Ensuite les méthodes d'évaluation et de mesure de la qualité de l'air seront présentées, en insistant sur les approches actuelles ainsi que sur celles émergentes liées à l'usage de capteurs bas coût.

Enfin, les défis et les questions de recherche qui orienteront ce manuscrit seront clairement définis. La méthodologie adoptée ainsi que les jeux de données sélectionnés pour atteindre ces objectifs seront détaillés.

1.1 La pollution atmosphérique : définition et méthodes de mesure

1.1.1 Définitions et origines de la pollution atmosphérique

La pollution de l'air ou pollution atmosphérique est une modification de la composition de l'air par des polluants nuisibles à la santé et à l'environnement.

Au cours des quatre-vingt dernières années, la consommation mondiale d'énergie a augmenté rapidement en raison de la croissance économique, de la croissance démographique, et de l'industrialisation.

Cette augmentation de la consommation d'énergie, principalement due à une utilisation accrue des combustibles fossiles, a entraîné la libération d'importantes quantités de dioxyde de carbone (CO₂), gaz à effet de serre contribuant au réchauffement de la planète. De plus, elle émet un certain nombre de polluants atmosphériques tels que le monoxyde de carbone (CO), le dioxyde de soufre (SO₂), les oxydes d'azote (NO_x) et des particules de matières.

La pollution atmosphérique est devenue une menace mondiale majeure pour la santé humaine. Par le passé, plusieurs épisodes majeurs de pollution atmosphérique survenus dans le monde au début du vingtième siècle ont eu de graves conséquences sur la santé. Un exemple marquant et tragique en occident est le brouillard mortel qui s'est abattu sur Londres en 1952 et qui a provoqué 12 000 décès et de graves effets à long terme sur la santé de la population.¹

Les effets néfastes de la pollution atmosphérique sont bien connus et l'Organisation Mondiale de la Santé (OMS) ne cesse de mettre en garde sur la gravité de la situation.

Selon la base de données de l'OMS publiée en avril 2022, 99 % de la population mondiale vivait dans des endroits où les niveaux de pollution atmosphérique dépassent les limites recommandées par l'OMS.

L'organisation a également déclaré que la pollution de l'air est la principale cause

1. Source : www.britannica.com/event/Great-Smog-of-London

environnementale de décès, responsable de 7 millions de décès prématurés dans le monde en 2018, dont 4 millions dus à la pollution de l'air extérieur.



Figure 1.1 – Un policier londonien, avec Big Ben en arrière-plan, dirige la circulation pendant le pic de pollution de 1952.

À mesure que la population s'installe dans les villes du monde entier, le nombre de décès dus à la pollution de l'air en milieu urbain augmente considérablement. D'ici 2050, la population mondiale devrait dépasser les 9 milliards d'habitants, et la part de la population vivant dans les villes devrait passer de 50 à 70 %, soit 6,3 milliards de personnes [48]. Avec la croissance rapide de la population urbaine, la demande d'énergie et de transport augmentera. En conséquence, l'Organisation de Coopération et de Développement Économiques (OCDE) prévoit qu'en l'absence de changement de politique, le nombre de décès dus à la pollution de l'air extérieur doublera en 2050 par rapport aux niveaux actuels.

La croissance urbaine est aussi à l'origine de nombreux problèmes environnementaux : en plus d'augmenter la consommation d'énergie comme on l'a vu plus haut, elle entraîne la perte des espaces verts et la destruction de la biodiversité [1]. Ces modifications sont indirectement liées à l'augmentation de la

pollution atmosphérique [33, 64].

Cette pollution de l'air provient en grande partie du processus de combustion. C'est un mélange complexe contenant de nombreux composants toxiques qui dépendent de la source d'émission. Plusieurs sources d'émissions de polluants existent, on peut citer les émissions anthropiques comme les rejets industriels, le trafic routier, le chauffage résidentiel, le transport maritime, le trafic aérien, la construction, les activités agricoles, les guerres et les incendies. Des sources naturelles peuvent aussi être à l'origine d'une dégradation de la qualité de l'air tels que les tremblements de terre, les éruptions volcaniques, les incendies de forêt spontanés et les températures extrêmes. La liste des composants chimiques de la pollution atmosphérique en milieu urbain a été fournie dans de nombreuses études [52, 8, 4, 17, 53, 50].

1.1.2 Principaux polluants

Les zones urbaines produisent une variété de gaz et de particules dont les caractéristiques dépendent de facteurs tels que la météo ou la source d'émissions. On dénote principalement quatre variétés de polluants :

1 - Les polluants gazeux : Les polluants gazeux tels que le SO₂, NO_x, CO, Ozone (O₃) ou Composés Organiques Volatils (COV) contribuent dans une large mesure à la modification de la composition de l'atmosphère. Ils sont dus à la combustion de combustibles fossiles dans la majorité des cas [41].

Les oxydes d'azote sont émis sous forme de NO qui réagit rapidement avec l'ozone ou les particules présentes dans l'atmosphère pour former du NO₂. Les principales sources anthropiques sont les sources de combustion mobiles et fixes. En outre, l'ozone dans les basses couches de l'atmosphère est formé par une série de réactions impliquant le NO₂, des composés organiques volatils et la lumière du soleil.

Le CO, quant à lui, est un produit de combustion incomplète. Sa principale source est le transport routier. Le SO₂ anthropique résulte lui de la combustion de combustibles fossiles contenant du soufre (principalement le charbon et les huiles lourdes) et de la fonte de minerais contenant du soufre. Les sources naturelles

telles que les volcans et les océans peuvent aussi contribuer aux émissions, à hauteur d'environ 2 % des émissions totales. Enfin, les COV constituent une classe de composés qui comprend des espèces chimiques de nature organique telles que le benzène.

Bien que la plupart des gaz polluants soient respirés et touchent surtout les voies respiratoires, ils peuvent aussi entraîner des complications sanguines (comme avec le CO et le benzène) et des formes de cancer.

2 - Les polluants organiques persistants : Les polluants organiques persistants constituent un groupe de produits chimiques toxiques. Ils persistent dans l'environnement pendant de longues périodes, et leurs effets sont amplifiés lorsqu'ils remontent la chaîne alimentaire, ce qu'on appelle la bioamplification. Ils comprennent les pesticides, ainsi que les dioxines, les furannes et les polychlorobiphényles. En général, le terme générique "dioxines" est utilisé pour désigner les dibenzo-dioxines polychlorées (PCDD) et les dibenzo-furanes polychlorés (PCDF). En revanche, les polychlorobiphényles (PCB) sont appelés "composés de type dioxine" et peuvent agir de manière similaire aux dioxines en termes de toxicité [72].

Les dioxines se forment lors d'une combustion incomplète, chaque fois que des matériaux contenant du chlore, comme par exemple des matières plastiques, sont brûlés.

Émises dans l'atmosphère, les dioxines ont tendance à se déposer sur le sol et dans l'eau. La plupart des dioxines présentes dans les plantes proviennent des pesticides et entrent dans la chaîne alimentaire, où elles s'accumulent en raison de leur capacité à se lier de manière stable aux lipides.

3 - Les métaux lourds : Les métaux lourds, tels que le plomb, le mercure, le cadmium, l'argent, le nickel, le vanadium, le chrome et le manganèse, sont des composants naturels de la croûte terrestre. Ces substances peuvent contaminer l'air que nous respirons en raison de leur dispersion atmosphérique, provenant de diverses activités industrielles, y compris l'incinération et la production d'énergie. Ces polluants aériens retombent ensuite dans nos écosystèmes, infiltrant

notre chaîne alimentaire et nos ressources en eau. Les émissions industrielles et les effluents des eaux usées sont parmi les voies principales par lesquelles ces toxines entrent dans notre environnement.

Bien que ces métaux soient essentiels en faibles concentrations pour certaines fonctions biologiques, leur présence excessive dans l'environnement peut conduire à une bioaccumulation préjudiciable. Ce phénomène se produit lorsque les organismes absorbent ces métaux plus rapidement qu'ils ne peuvent les éliminer, entraînant une accumulation toxique qui peut avoir des conséquences graves sur la santé humaine, comme l'ont souligné diverses études [37].

4 - Les particules en suspension ou particules de matière :

Les Particules de Matière (PM) sont des polluants atmosphériques constitués de divers mélanges de particules en suspension dans l'air respirable.

Ces particules peuvent être de source naturelles ou anthropiques, telles que les usines, les centrales électriques, les véhicules, les incendies, les travaux de construction et la poussière transportée par le vent [67].

La taille des particules varie et elles peuvent être classées en trois catégories : les particules ultra-fines (inférieure à $0,1 \mu m$), les particules fines (inférieure à $1 \mu m$) et les particules grossières (plus grandes que $1 \mu m$). La taille des particules détermine l'endroit où elles se déposeront dans les voies respiratoires. Les particules fines et ultra-fines peuvent atteindre les alvéoles pulmonaires, tandis que les particules grossières se déposent principalement dans les voies respiratoires supérieures. Les particules sont composées de métaux, de composés organiques, de matières biologiques et d'ions.

La composition des particules peut varier en fonction des polluants qu'elles ont absorbés. Les particules ultrafines et fines sont plus dangereuses que les particules plus grosses, principalement en raison de leur teneur en métaux, en Hydrocarbures Aromatiques Polycycliques (HAP) et autres composants organiques. Ces particules peuvent avoir des effets morbides sur les voies respiratoires et sur le système cardiovasculaire.

1.1.3 Impacts sur la santé

Comme vu précédemment, la pollution atmosphérique est l'une des principales causes de mortalité en région urbaine. Elle a de nombreux effets sur plusieurs parties du corps humain :

- **Système respiratoire** : Dans les grandes villes, la pollution extérieure est liée à une augmentation des infections respiratoires aiguës, qui sont à l'origine de près d'un tiers des décès chez les enfants de moins de cinq ans [14, 18, 46]. L'exposition au NO₂ et aux particules de matière dans cinq villes allemandes a été associée à une augmentation des laryngo-trachéo-bronchites, principalement dues à des maladies grippales virales [73]. Une étude menée dans trois villes finlandaises aboutit à des conclusions similaires [36]. L'impact négatif des polluants atmosphériques est particulièrement important chez les personnes souffrant d'infections pulmonaires préexistantes ou d'autres maladies pulmonaires, car elles sont plus sensibles. Les enfants sont également vulnérables, peut-être parce qu'ils ont une surface pulmonaire relativement plus grande et qu'ils pratiquent davantage d'activités physiques en plein air, ce qui les expose d'avantage à la pollution de l'air.

Si le tabagisme est souvent associé au développement du cancer du poumon [54], d'autres facteurs de risque doivent être pris en compte [15]. La pollution atmosphérique intérieure et extérieure contribue de manière significative à l'apparition de cette maladie, en provoquant des lésions de l'ADN, des aberrations chromosomiques et d'autres biomarqueurs qui font actuellement l'objet de recherches. De nombreuses études ont établi une corrélation directe entre les polluants atmosphériques et les taux élevés de cancer du poumon [90, 19].

-- **Système cardiovasculaire** : De nombreux rapports ont établi une corrélation entre l'exposition aux PM et les maladies cardiovasculaires [3, 59, 21].

Gold et al.[28] ont découvert que chaque augmentation de 5 à 6 µg/m³ de PM_{2.5} (particules de matière avec un diamètre inférieur à 2.5 µm) entraînait une

augmentation significative des maladies cardiovasculaires, allant de 0,5 % à 1,5 %. Ils ont également constaté qu'une exposition aiguë à la pollution atmosphérique particulaire entraînait une augmentation de 69 % des décès d'origine cardiovasculaire. Il est intéressant de noter que le taux de mortalité dû aux maladies cardiovasculaires était plus élevé (69 %) que celui des maladies respiratoires (28 %) après une exposition aux PM_{2,5}.

Des études sur les PM₁₀ ont également montré un lien entre la pollution atmosphérique et les maladies cardiovasculaires [86, 66], telles que les cardiopathies. Ces études ont révélé que lorsque les niveaux de PM₁₀ augmentaient de 10 µg/m³, il y avait une augmentation correspondante des taux d'hospitalisation pour cardiopathie ischémique (0,7 %) et insuffisance cardiaque congestive (0,8 %) [57].

Dans de nombreuses autres études [9, 83, 62, 63] les auteurs ont examiné la relation entre la pollution de l'air en général et les maladies cardiovasculaires tout en explorant les changements de la pression artérielle. Dans l'une de ces études, les chercheurs ont constaté une forte association entre la teneur en carbone des particules de la pollution atmosphérique et l'augmentation de la pression artérielle.

- Système nerveux : Le système nerveux peut être affecté par l'exposition à des métaux lourds tels que le plomb, le mercure et l'arsenic, ainsi qu'aux dioxines. Cette exposition peut être neurotoxique avec des symptômes tels que des troubles de la mémoire, des troubles du sommeil, de l'irritabilité, de la fatigue, des tremblements des mains, une vision floue et des troubles de l'élocution [23, 69].

D'autres études ont montré que l'exposition au plomb, au mercure et à l'arsenic peut provoquer des neuropathies. L'exposition au plomb en particulier peut entraîner des lésions dans le système dopaminergique, le système glutamate et le complexe des récepteurs N-méthyl-D-aspartate (NMDA), qui sont tous importants pour les fonctions de la mémoire [44, 45].

L'exposition au mercure a également été liée à des cancers neurologiques. En outre, on a constaté que les dioxines réduisaient la vitesse de conduction ner-

veuse et entravaient le développement mental des enfants [81, 84].

La pollution de l'air par les PM a aussi été associée à un risque accru d'accident vasculaire cérébral, comme l'a montré une étude menée au Danemark [20].

La pollution de l'air peut aussi avoir des impact négatifs sur d'autre parties du corps humain, tel que la peau, le système reproductif ou le système immunitaire [70, 10, 27].

Au vu de toutes les preuves scientifiques sur la dangerosité de la pollution atmosphérique, la lutte contre cette dernière est devenue un enjeu majeur.

1.1.4 Surveillance de l'air : étape clé contre la pollution atmosphérique

La lutte contre la pollution de l'air est une entreprise complexe qui exige la collaboration de différents acteurs, notamment des scientifiques et des politiciens.

C'est une procédure longue et à plusieurs étapes : la complexité de la pollution de l'air fait qu'il n'existe pas de solution unique ou rapide. La qualité de l'air est influencée par une variété de polluants provenant de diverses sources, nécessitant des analyses approfondies et des mesures adaptées à chaque situation. De plus, la lutte contre la pollution en général peut être en concurrence avec d'autres objectifs divergents tels que la croissance économique ou l'acceptabilité politique. C'est l'une des raisons pour lesquelles nos politiques et décideurs sont moins enclins à s'y attaquer.

Rôle des scientifiques : Les scientifiques jouent un rôle essentiel dans la mesure et l'évaluation de l'exposition, la compréhension des sources de pollution, des types de polluants, et des impacts sur la santé humaine et sur l'environnement. Ils sont responsables de la réalisation d'études et de recherches, pour identifier les causes précises de la pollution, évaluer ses effets, et proposer des solutions adaptées.

Rôle des politiques : Les politiciens, quant à eux, sont en mesure de mettre en place des mesures et des réglementations basées sur les découvertes scientifiques. Leur engagement est crucial pour la mise en œuvre de mesures efficaces, la mobilisation des ressources nécessaires, et l'assurance du respect des réglementations par les industries et le public.

Rôle des industriels : Les industriels occupent une double position face à la pollution atmosphérique. D'un côté, certains, en tant que contributeurs majeurs à la pollution, sont incités par les découvertes scientifiques et les pressions politiques à réorienter leurs pratiques, réduisant ainsi leurs émissions tout en valorisant leur image. De l'autre, des acteurs industriels innovants exploitent le défi de la pollution comme une opportunité, développant des services et technologies pour la surveillance et l'information sur la qualité de l'air. Ces dynamiques illustrent la capacité de l'industrie à être à la fois partie du problème et de la solution.

Une étape cruciale dans ce processus en plusieurs étapes est la surveillance de la qualité de l'air : Pour que les efforts de réduction de la pollution de l'air soient efficaces, il est indispensable de surveiller constamment la qualité de l'air. Cela permet de suivre les niveaux de pollution, d'évaluer l'efficacité des mesures prises, et d'ajuster les stratégies en conséquence. Cette surveillance implique l'installation de stations de mesure, la collecte de données, et l'analyse régulière de ces données pour prendre des décisions éclairées.

C'est dans ce contexte que s'écrit cette thèse. Elle a pour but de rajouter une pierre à l'édifice de l'évaluation et de la cartographie de la pollution atmosphérique en utilisant des nouvelles méthodes de mesures, en combinant modèles mathématiques et avancées technologiques.

1.1.5 Méthodes actuelles d'évaluation de la qualité de l'air

Actuellement, la surveillance de la qualité de l'air est effectuée à l'aide de stations fixes de référence [40, 38]. Ces stations sont gérées par les agences de

protection de l'environnement agréées par le gouvernement, telles que Airparif en Île-de-France 1.2. Ces stations de référence fournissent une mesure précise de la qualité de l'air mais leur couverture spatiale est limitée, par exemple seulement 22 stations couvrant 50×50 km² (113 km² par station) à Pékin, 14 stations couvrant 1572 km² (112 km² par station) à Londres, et 61 stations (221 km² par station) en Flandre, en Belgique.

Les stations peuvent seulement générer des séries temporelles détaillées à des endroits précis. Il est donc difficile d'avoir des informations fiables et représentatives pour une ville ou une région. Par conséquent, seule une vue plus macroscopique des tendances des champs de pollution est fournie.



Figure 1.2 – Exemple d'une station de mesure fixe de qualité de l'air gérée par Airparif

Prenant l'exemple des particules de matière : Airparif dispose de cinq stations capables de mesurer les PM_{2.5} à Paris intra-muros et d'une seule station dans tout le département du Val-de-marne. Sachant que les émissions des PM sont majoritairement dues au trafic routier et au chauffage des résidences, ces deux sources pouvant changer d'un quartier à l'autre, il en résulte une concentration de PM très hétérogène spatialement. Il est donc très difficile d'avoir une représentation spatiale précise des concentrations de PM sur toute la région ur-

baine de l'Île-de-France.

Les chercheurs ont également placé des capteurs de pollution pour évaluer l'exposition individuelle à la pollution atmosphérique dans des lieux d'intérêt [74, 75, 26, 61], tels que les grands axes routiers à forte circulation et les sites industriels. Les mesures obtenues dans ces stations ou à ces endroits ne peuvent refléter que le niveau de pollution de l'air dans de petites zones. De même, les campagnes de terrain spécialisées peuvent utiliser des dizaines de dispositifs de mesure et obtenir une couverture spatiale très dense, mais leur déploiement est généralement limité à de courtes périodes ne dépassant pas quelques semaines.

En approfondissant les aspects complexes de la pollution atmosphérique, il devient impératif d'ancrer les discussions dans un cadre de normes établies. Les seuils réglementaires, fixés par différents pays et organismes internationaux, précisent ce qui est considéré comme « sûr » ou « acceptable » en matière de qualité de l'air. Le tableau 1.1 présente les principales réglementations et normes en matière de qualité de l'air émanant de l'Union Européenne (UE) et la France (Fr).

L'harmonisation des normes de qualité de l'air à l'échelle mondiale reste un défi majeur, compte tenu de la diversité des seuils réglementaires adoptés par différents pays [16]. Par exemple, les États-Unis, à travers l'agence de protection de l'environnement, et la Chine ont établi leurs propres normes, qui peuvent varier significativement en termes de limites admissibles pour les polluants tels que les particules PM_{2.5} et PM₁₀, le NO₂, l'ozone et le SO₂. Ces différences reflètent non seulement les priorités environnementales et les capacités de surveillance de chaque pays, mais aussi les défis socio-économiques et les niveaux de développement industriel. L'absence d'un consensus global sur les valeurs limites conduit à une hétérogénéité des réponses politiques et réglementaires face à la pollution atmosphérique. Cela peut entraîner des inégalités en matière de santé publique, car les populations de certains pays sont plus exposées à des niveaux de pollution potentiellement dangereux. De plus, cette disparité complique la tâche des entreprises multinationales qui doivent naviguer à travers

un paysage réglementaire fragmenté et parfois contradictoire.

1.1.6 Émergence des capteurs bas coût

Les progrès des technologies des capteurs et de la communication ont permis l'émergence de capteurs peu coûteux, ce qui a ouvert de nouvelles possibilités d'évaluation de la pollution atmosphérique et de l'exposition.

Bien qu'il n'existe pas de définition standard d'un capteur « bas coût », dans cette thèse, cette expression fait référence à des dispositifs dont le prix se situe généralement entre 20 et 500 euros, contrairement aux systèmes de surveillance traditionnels qui peuvent coûter plusieurs milliers à centaines de milliers d'euros. Toutefois, le coût doit être aussi bas que possible pour atteindre les objectifs d'un système de surveillance de la pollution de l'air basé sur ces capteurs, de manière à ce qu'un déploiement à grande échelle soit commercialement possible.

Plusieurs études portant sur le sujet innovant de la surveillance de la qualité de l'air à l'aide de capteurs bas coût ont été publiées [87, 11, 43]. Elles se sont concentrées sur les avantages, les défis et leur potentiel pour diverses applications [12, 49, 56, 76]. D'autres études explorent les nouvelles technologies de capteurs capables de surveiller les polluants atmosphériques gazeux et/ou particulaires [2, 7, 42].

Bien que présentant de nombreux avantages, l'utilisation de capteurs à bas coût pour évaluer la pollution atmosphérique urbaine présente également des inconvénients. Prises séparément, les données de ces capteurs sont souvent bruitées et peu précises.

Des études ont analysé les performances de capteurs de qualité de l'air peu coûteux ainsi que leurs avantages et leur viabilité pour la surveillance des niveaux de pollution de l'air dans les zones urbaines [39, 58]. Aucun des capteurs testés n'a montré une bonne corrélation avec les données de référence dans les faibles concentrations ambiantes (de 0 à 15 $\mu\text{g}/\text{m}^3$).

D'autres études [47, 5] menées dans des environnements de laboratoire contrô-

Table 1.1 – Tableau des normes et réglementations des principaux polluants

Substance	Type de réglementation	Valeur	Durée
DIOXYDE d'AZOTE (NO ₂)	Objectif de qualité	40 µg/m ³ (Fr)	Moyenne annuelle
	Valeurs limites	40 µg/m ³ (UE)	Moyenne annuelle
	Seuil d'information	200 µg/m ³ (Fr)	Moyenne horaire
	Seuils d'alerte	400 µg/m ³ (UE)	Moyenne horaire
PARTICULES (PM ₁₀)	Objectif de qualité	30 µg/m ³ (Fr)	Moyenne annuelle
	Valeurs limites	40 µg/m ³ (UE)	Moyenne annuelle
	Seuil d'information	50 µg/m ³ (Fr)	Moyenne journalière
	Seuils d'alerte	80 µg/m ³ (Fr)	Moyenne journalière
PARTICULES (PM _{2.5})	Objectif de qualité	10 µg/m ³ (Fr)	Moyenne annuelle
	Valeur cible	20 µg/m ³ (Fr)	Moyenne annuelle
	Valeurs limites	25 µg/m ³ (Fr)	Moyenne annuelle
DIOXYDE de SOUFRE (SO ₂)	Objectif de qualité	50 µg/m ³ (Fr)	Moyenne annuelle
	Valeurs limites	125 µg/m ³ (UE)	Moyenne journalière
	Seuil d'information	300 µg/m ³ (Fr)	Moyenne horaire
	Seuils d'alerte	500 µg/m ³	Moyenne horaire
OZONE (O ₃)	Objectif de qualité	120 µg/m ³ (Fr)	Moyenne sur 8h
	Valeur cible	120 µg/m ³ (Fr)	Moyenne sur 8h
	Seuil d'information	180 µg/m ³ (Fr)	Moyenne horaire
	Seuils d'alerte	240 µg/m ³	Moyenne horaire
MONOXYDE de CARBONE (CO)	Valeurs limites	10 000 µg/m ³ (Fr)	Moyenne sur 8h
BENZÈNE (C ₆ H ₆)	Objectif de qualité	2 µg/m ³ (Fr)	Moyenne annuelle
	Valeurs limites	5 µg/m ³ (UE)	Moyenne annuelle
Plomb (Pb)	Objectif de qualité	0.25 µg/m ³ (Fr)	Moyenne annuelle
	Valeurs limites	0.5 µg/m ³ (UE)	Moyenne annuelle

lés, ont montré qu'un processus d'étalonnage est nécessaire pour garantir la fiabilité des mesures.

Bien que les déploiements sur le terrain de réseaux de capteurs de qualité de l'air à faible coût se soient révélés prometteurs pour mesurer les niveaux ambiants d'O₃ et les niveaux de CO, NO et NO₂ [55, 65], les processus d'étalonnage n'ont pas été suffisants pour atteindre le niveau de qualité des stations de gestion de la qualité de l'air standard, même après un étalonnage initial sur le terrain [88]. Des approches d'étalonnage sur le terrain ont été proposées [82], mais elles présentent des limites et n'ont pas encore été entièrement testées.

À ce jour, plusieurs études scientifiques ont démontré le manque de fiabilité des données issues des capteurs bas coût [24]. Contradictoirement, d'autres études scientifiques ont démontré la faisabilité et fiabilité d'un déploiement d'un réseau de capteurs bas coût de pollution sur le terrain [58]. Cela dit, nous n'avons trouvé aucune analyse détaillée de la capacité du réseau à capturer la variabilité spatio-temporelle des polluants à petite échelle et à suppléer les stations de référence.

Il est important de poursuivre les recherches et les essais sur ces réseaux afin d'améliorer leur fiabilité et leur précision dans l'évaluation et de l'exposition à la pollution atmosphérique.

1.1.7 Technologie des capteurs bas coût

Est exposé ici le principe des capteurs à faible coût pour la détection des polluants gazeux et des particules en suspension, ces deux groupes étant au centre des travaux de cette thèse.

De nombreuses descriptions détaillées des méthodes de détection des capteurs sont disponibles dans la littérature. Un bref aperçu est donné ici.

Capteurs à faible coût pour la surveillance des polluants gazeux

Pour mesurer les polluants atmosphériques gazeux, deux types de capteurs à faible coût sont actuellement disponibles sur le marché : les capteurs métal-

oxyde-semi-conducteur et les capteurs électrochimiques.

Lorsqu'ils sont exposés au gaz recherché, les capteurs de type métal-oxyde-semi-conducteur utilisent un oxyde métallique qui modifie ses propriétés électriques. Cette variation peut être aisément quantifiée et est en corrélation avec la concentration gazeuse [25]. Ce sont des capteurs de petite taille et légers, avec des temps de réponse rapides, des limites de détection basses et une faible consommation d'énergie [2, 65]. Cependant, leur courbe de réponse n'est pas linéaire et ils sont sensibles aux changements des conditions environnementales [78].

Un capteur de gaz électrochimique réagit avec le gaz mesuré et produit un signal électrique proportionnel à la concentration du gaz.

Un capteur électrochimique standard se compose d'une électrode de détection et d'une contre-électrode, séparées par une fine couche d'électrolyte. Grâce à de minuscules ouvertures capillaires et à des couches barrières hydrophobes, une quantité donnée de gaz peut réagir avec l'électrode de détection pour former un signal électrique suffisant tout en évitant les fuites d'électrolyte. La concentration de gaz est déterminée en mesurant le courant proportionnel à la concentration du gaz mesuré [79].

Capteurs à faible coût pour la détection des particules de matière

Les capteurs bas coût de PM reposent sur le principe de diffraction de la lumière. Les capteurs basés sur ce principe sont peu coûteux à fabriquer, ont une faible consommation d'énergie et des temps de réponse rapides [85]. Dans cette méthode, une source lumineuse éclaire les particules, puis la lumière diffractée est mesurée par un photomètre.

Pour les particules dont le diamètre est supérieur à 0,3 μm , la quantité de lumière diffractée est à peu près proportionnelle à leur concentration mais les particules les plus petites peuvent ne pas être détectées par cette méthode [42, 80]. Les particules peuvent être séparées selon leur taille en utilisant un algorithme de traitement du signal [60] ou en fixant un filtre optique à l'entrée [77].

Le tableau 1.2 référence les principaux capteurs bas coût PM du marché (extrait de la publication de Rai et. al[68]).

L'utilisation de flottes de ces capteurs peu coûteux à bord de véhicules et mobiles (voitures, bus, tramways, piétons, etc.) circulant dans une zone urbaine afin d'obtenir une meilleure représentation des polluants sont de plus en plus utilisés. Ils sont installés dans ce qu'on appelle des boîtiers mobiles de pollution. Ces boîtiers peuvent regrouper un ou plusieurs capteurs (de polluants, de température et d'humidité), un microprocesseur, une antenne GPS (Global Positioning System), une carte mémoire, une batterie, un ventilateur et même un écran. La figure 1.3 montre un boîtier mobile utilisé dans l'étude de Hasenfratz et al. [31]

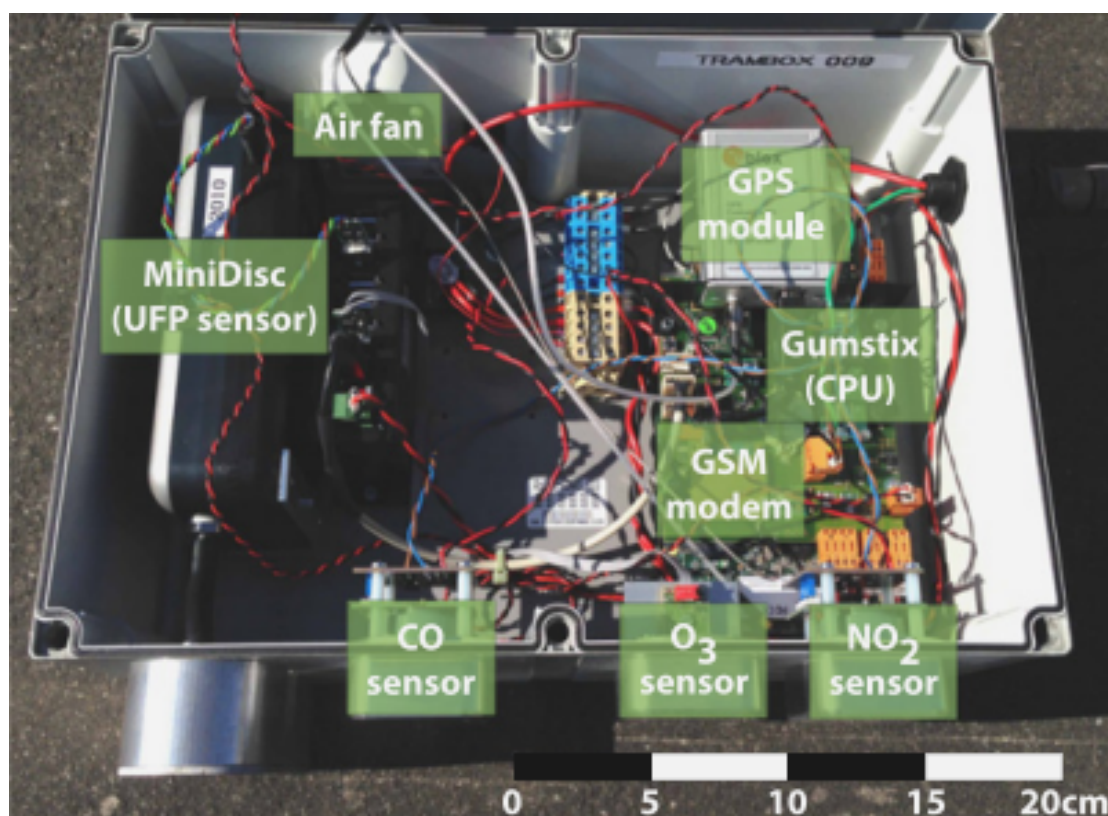


Figure 1.3 – Exemple de boîtier mobile de pollution contenant un capteur de pollution bas coût utilisé dans l'étude [31]

Modèle	Taille (mm)	Poids (g)	Alimentation électrique	Courant max(mA)	Coût (dollar)	Taille des particules décelables	Plage de concentration	Tests documentés
Alphasense OPC-N2	75×64×60	105	5 V DC	175	500	0.38–17 µm	0.1–1,500,000 µg/m ³	Oui
Dylos DC 1100 Pro	178×114×76	544	110 V AC	NA	300	0.5–10 µm	0–106 particules/cm ³	Oui
Dylos DC 1700	178×114×76	544	110 V AC ou batterie	NA	400	0.5–10 µm	0–106 particules/cm ³	Oui
Novafitness SDL301	204×100×36	580	5 V DC	NA	250	0.3–10 µm	0–1000 µg/m ³	Non
Novafitness SDL607	73×73×20	120	5 V DC	NA	120	0.3–10 µm	0–1000 µg/m ³	Non
Novafitness SDS011	71×70×23	NA	5V DC	80	35	0.3–10 µm	0–1000 µg/m ³	Non
Novafitness SDS018	59×45×20	NA	5 V DC	70	40	0.3–10 µm	0–1000 µg/m ³	Non
Novafitness SDS021	42×32×24	NA	5 V DC	70	35	0.3–10 µm	0–1000 µg/m ³	Non
Novafitness SDS198	71×70×23	NA	5 V DC	80	80	1–100 µm	0–20000 µg/m ³	Non
Plantower PMS 1003	65×42×23	NA	5 V DC	120	20	0.3–10 µm	0–500 µg/m ³	Oui
Plantower PMS 3003	65×42×23	NA	5 V DC	120	20	0.3–10 µm	NA	Oui
Samyoung DSM501A	59×45×20	25	5 V DC	90	15	> 1.0 µm	0–1400 µg/m ³	Oui
Sharp DN7C3CA006	50×44×20	52	5 V DC	180	20	0.5–2.5µm	25–500 µg/m ³	Oui
Sharp GP2Y1010AU0F	46×30×18	15	5 V DC	20	10	> 0.5µm	0–600 µg/m ³	Oui
Shinyei PPD42NS	59×45×22	24	5 V DC	90	15	> 1.0 µm	0–28 particules/cm ³	Oui
Shinyei PPD60PV	88×60×20	36	5 V DC	NA	250	> 0.5 µm	0–70 particules/cm ³	Oui

Table 1.2 – Tableau récapitulatif des principaux capteurs bas coût de PM, issu de la publication de Rai et. al[68]

Contrairement aux stations traditionnelles de surveillance de la qualité de l'air, l'utilisation d'un réseau de capteurs mobiles peu coûteux, capables de se déplacer dynamiquement dans l'environnement, fournira des données d'une résolution sans précédent. Voici quelques exemples notables de projets de recherche utilisant des capteurs bas coût mobiles pour surveiller la pollution de l'air : les projets « OpenSense » en Suisse [31], « Array of Things » à Chicago, aux États-Unis [13], l'Imperial County Community Air Monitoring Network [22] en Californie, aux États-Unis, « Gotcha » II à Shenzhen, en Chine [89], et le « Air Map Korea Project » dans les grandes villes de Corée du Sud.

Dans ce contexte, un capteur mobile pourrait être un bon compromis entre précision des données et la résolution spatiale, permettant une couverture spatiale élevée sur de grandes zones sans utiliser un grand nombre de capteurs fixes. Cependant, en raison de la résolution temporelle réduite de tout lieu échantillonné, il est difficile de générer des cartes de pollution avec une résolution temporelle élevée à l'échelle quotidienne ou horaire.

1.1.8 Spatialisation des données : de la mesure à la cartographie

La surveillance de la qualité de l'air à l'aide de capteurs mobiles suscite un intérêt croissant. Plusieurs dispositifs ont été mis au point pour surveiller, en temps réel, la variabilité spatiale et temporelle de la qualité de l'air à l'aide de différents instruments. Gozzi et al.[29] ont fait le point sur l'utilisation des capteurs mobiles pour la surveillance des particules.

La plupart des études scientifiques ont utilisé la surveillance mobile pour évaluer l'exposition à la pollution atmosphérique ou pour étudier les caractéristiques spatiales et temporelles de la pollution atmosphérique à un certain endroit. Seules quelques études se sont intéressées à la production, à une échelle spatio-temporelle fine, de cartographies de la pollution atmosphérique urbaine à l'aide de capteurs mobiles bas coût.

La transition des mesures de la qualité de l'air provenant de capteurs à faible coût vers la cartographie des polluants sur une zone urbaine étendue, avec une haute précision spatiale et temporelle, requiert l'utilisation de modèles statistiques ou physiques de spatialisation. Ce processus complexe s'appuie sur des modèles mathématiques et peut nécessiter des données auxiliaires. La réalisation de cette cartographie est essentielle pour informer le public sur les niveaux potentiels d'exposition, en identifiant par exemple les zones sensibles à éviter ou les itinéraires optimaux entre deux points. De plus, elle permet d'identifier les sources principales de pollution, ce qui offre la possibilité aux responsables de prendre des mesures correctives.

La spatialisation, ou plus précisément, la production de cartographies spatio-temporelles à partir des données brutes constitue le cœur de cette thèse. Dans ce qui suit, ont été explorés les défis majeurs et les questions de recherche qui en découlent.

Ainsi, il est impératif d'adapter les modèles statistiques appropriés à cette nouvelle méthode de mesure, tout en garantissant une utilisation optimale des dispositifs de surveillance disponibles et une interprétation précise des données obtenues. Par conséquent, l'objectif de cette recherche est d'identifier et de perfectionner les modèles statistiques pour générer une cartographie précise de la qualité de l'air en milieu urbain grâce à des capteurs mobiles à faible coût.

1.2 Stratégies de recherche

1.2.1 Défis de la thèse

La génération de cartographies précises de la qualité de l'air à l'aide de capteurs bas coût présente plusieurs défis majeurs. Ces défis sont principalement liés à la nature du phénomène étudié, à la précision et au volume des données. On pourra citer les défis suivants :

Difficulté de la cartographie du phénomène à petite échelle (aléas, données manquantes) :

La qualité de l'air est influencée par une multitude de facteurs, tels que la vitesse du vent, le trafic routier, la production industrielle, les catastrophes naturelles comme les incendies et les conditions météorologiques locales. Ces facteurs peuvent varier considérablement sur de courtes distances, rendant la cartographie à l'échelle d'un quartier particulièrement complexe.

Par exemple, après avoir comparé les données de surveillance de deux rues de Copenhague (Jagtvej et Bredgade), Berkowicz et al. [6] ont soutenu que les mesures effectuées en bord de route dépendaient du site et n'étaient pas représentatives d'une zone urbaine plus vaste. Ils ont démontré que les concentrations mesurées pouvaient être très différentes sur ces deux sites. Une autre étude [71] a montré que les mesures de la qualité de l'air effectuées à l'intersection de deux rues du centre de Londres dépendaient fortement de la vitesse locale du vent et de la géométrie des rues et des bâtiments entourant le capteur.

Les phénomènes météorologiques tels que le vent, la température et l'humidité peuvent entraîner des variations rapides et imprévisibles de la qualité de l'air. Ces variables diffèrent souvent d'un quartier à l'autre à la manière des îlots de chaleur au sein d'une ville. Malheureusement, les mesures à grande échelle sont souvent les seules dont nous disposons.

Par Conséquent, le phénomène qui nous intéresse est extrêmement complexe à une échelle locale, il est influencé par une multitude de facteurs, dont certains sont connus tandis que d'autres restent inconnus. Les données relatives à ces facteurs sont souvent incomplètes et peu représentatives.

Difficulté liée à la nature des données issues des capteurs mobiles (données peu précises)

Comme observé précédemment, la précision des capteurs à faible coût demeure leur principal défaut. Peu importe le modèle statistique utilisé, si les données sont biaisées et ne fournissent aucune information sur l'état réel du signal, les sorties de cartographie refléteront également cette limitation.

Enfin, même si le coût abordable de ces capteurs permet un déploiement à grande échelle, le sous-échantillonnage reste une problématique avérée. Il est souvent impossible d'échantillonner simultanément tous les quartiers d'une ville, ce qui nous confronte aux mêmes défis que ceux rencontrés lors de l'échantillonnage à l'aide de stations de mesure classiques, à une échelle réduite.

Lorsque les capteurs à bas coût sont déployés en grandes quantités et en utilisant les bons modèles d'étalonnage et de reconstitution, ils sont néanmoins en mesure de fournir des informations complexes et complémentaires aux stations de surveillance fixes.

Difficulté liée à la quantité de données (complexité, problème du « big n »)

Le troisième défi est dû à la complexité et au temps nécessaire pour traiter les grandes quantités de données générées.

Contrairement aux deux premières difficultés où le manque de données sur les facteurs régissant le phénomène à petite échelle et sur la pollution à des

endroits précis est manifeste, le surabondance de données de pollution peut également poser un défi délicat à gérer.

Ainsi, les capteurs bas coût sont capables de produire des données chaque seconde, générant ainsi une très grande quantité de données qui peuvent être parfois redondante.

Ce problème est répandu dans tous les domaines de la surveillance environnementale, où les avancées technologiques ont généré une abondance de données. Plusieurs études se sont penchées sur cette question, dont certaines portent spécifiquement sur la qualité de l'air [91, 34].

1.2.2 Questions de recherche

Au vu des difficultés énoncées précédemment, il est primordial de savoir comment combiner toutes ces sources de données floues pour construire une image nette de la pollution urbaine. Le modèle statistique recherché aura comme objectif d'inclure toutes les spécificités du problème pour offrir des cartes de pollution toujours plus précises.

Plus spécifiquement, notre recherche aura pour but de répondre aux questions suivantes :

- Quels sont les modèles statistiques les plus performants en matière de précision de la prédiction (interpolation ou extrapolation temporelle), de pouvoir descriptif, de complexité et d'incertitude associée à la prédiction pour réaliser une cartographie de la qualité de l'air urbain à l'aide de capteurs bas coût fixes et mobiles ?
- Quelle est l'influence du ratio entre les capteurs à faible coût, mobiles et fixes sur la qualité de la prédiction ?
- Est-ce qu'un système de capteur bas coût peut remplacer, à terme, les méthodes actuelles de surveillance de l'air ? Peut-on obtenir la même précision d'évaluation qu'avec une station de surveillance fixe de référence ?

1.2.3 Méthodologie

Afin de répondre aux questions de recherche posées, une méthodologie en deux parties a été élaborée, illustrée dans la Figure 1.4.

En premier lieu, il convient de repérer toutes les familles de modèles présentes dans la littérature. Durant cette étape, le travail a consisté à étudier ces grandes familles (telles que la géostatistique) et à choisir les modèles s'adaptant le mieux à notre problématique. Cette phase a fait l'objet d'une première publication [35].

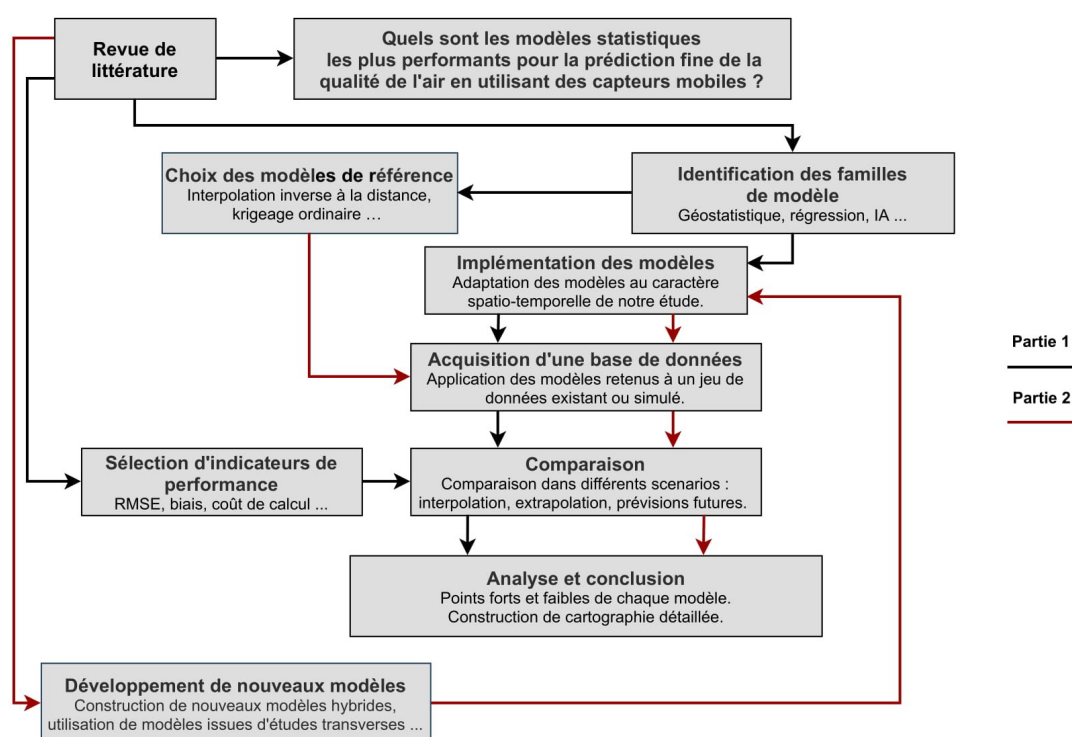


Figure 1.4 – Méthodologie retenue pour répondre aux questions de recherche.

Toujours dans la première partie, la seconde étape a consisté en la mise en œuvre et la comparaison de divers modèles préalablement appliqués à des problématiques semblables. Cette implantation a été réalisée sur un jeu de données réel issu d'un échantillonnage avec des capteurs mobiles et fixes à faible coût. Après avoir défini les indicateurs de performance utilisés, une comparaison des modèles déjà présents dans la littérature a été effectuée dans différents

scénarios. Les résultats ont fait l'objet d'une deuxième publication. Les résultats de ces deux études, formant la première partie de notre méthodologie seront vu au chapitre 3.

La deuxième partie se concentrera sur le développement de nouveaux modèles visant à obtenir de meilleurs résultats que ceux déjà étudiés. Cela commencera par la création de modèles pour résoudre les problèmes des méthodes précédemment examinées. Nous approfondirons les bases statistiques de notre étude, conduisant à la proposition de modèles inédits. Une troisième publication présentant ce nouveau modèle a été produite sur cette seconde partie. Ce modèle, conçu pour améliorer l'existant et adapté à la nature des données (bruitées et en grande quantité), vise à produire une cartographie de la pollution urbaine avec une précision spatiale et temporelle maximale et en temps réel. Les résultats de cette étude seront présentés au chapitre 4 et 5.

1.2.4 Jeux de données utilisés durant la thèse

La mise en application de notre méthodologie, destinée à la comparaison ainsi qu'à l'élaboration de nouveaux modèles statistiques, s'effectuera en s'appuyant sur des jeux de données issus de capteurs à faible coût acquis tout au long de cette thèse.

Étant donné la rareté des études réalisées pour la production de cartes sur la pollution atmosphérique urbaine à l'aide de capteurs bas coût, les jeux de données publics disponibles se trouvent en quantité limitée. A cette contrainte s'ajoute la nécessité de l'acquisition des données provenant des stations fixes des organismes de surveillance gouvernementales, qui sont considérées comme une référence.

Cette sous-section apporte des précisions sur les jeux de données utilisés durant la thèse, leur provenance, mais également leurs atouts et leurs limites. Chaque source de données offre ses propres défis et opportunités.

Ces jeux de données, de par leur nature diversifiée et complémentaire, offrent un terrain d'expérimentation robuste pour tester, valider et améliorer les modèles. Ils serviront de fil conducteur tout au long des chapitres suivants, permettant ainsi au lecteur de mieux comprendre la portée et l'applicabilité de travaux présentés.

- **Jeu de données d'OpenSense**

Le projet OpenSense² est un projet suisse visant à intégrer les mesures de la qualité de l'air provenant de sources de données mobiles hétérogènes afin de comprendre les effets sur la santé de l'exposition à la pollution atmosphérique et de fournir des cartes de la qualité de l'air urbain à haute résolution. Ce projet a déployé plusieurs capteurs mobiles de qualité de l'air sur les toits des tramways de la ville suisse de Zurich et des bus de Lausanne, recueillant la mesure des concentrations d'ozone et des particules ultra-fines (UFP). De plus amples informations sur les données ainsi que sur la méthodologie de collecte des données peuvent être trouvées dans ces études [51, 32].

Dans cette thèse, sont exploitées les données de concentrations d'ozone mesurées par les capteurs mobiles déployés sur le toit des tramways de Zurich. La figure 1.5 montre l'emplacement de la station fixe de référence ainsi que les trajectoires des tramways.

L'objectif étant de prédire la concentration sur une résolution spatio-temporelle très détaillée, une seule semaine de données a été utilisée (du 28 février au 5 mars 2016).

Ces données contiennent des mesures de cinq capteurs sur les lignes numérotées 4, 7, 8, 12 et 13. Elles fournissent ainsi un jeu de données dépassant 40 000 observations.

Les données d'OpenSense procurent la concentration d'ozone en parties par milliard (ppb) dans un volume donné (volume de polluant gazeux par 10^9 volumes d'air ambiant). Afin de le convertir en $\mu g/m^3$ pour corres-

2. Source : <https://tec.ee.ethz.ch/research/networked-embedded-systems/opensense.html>

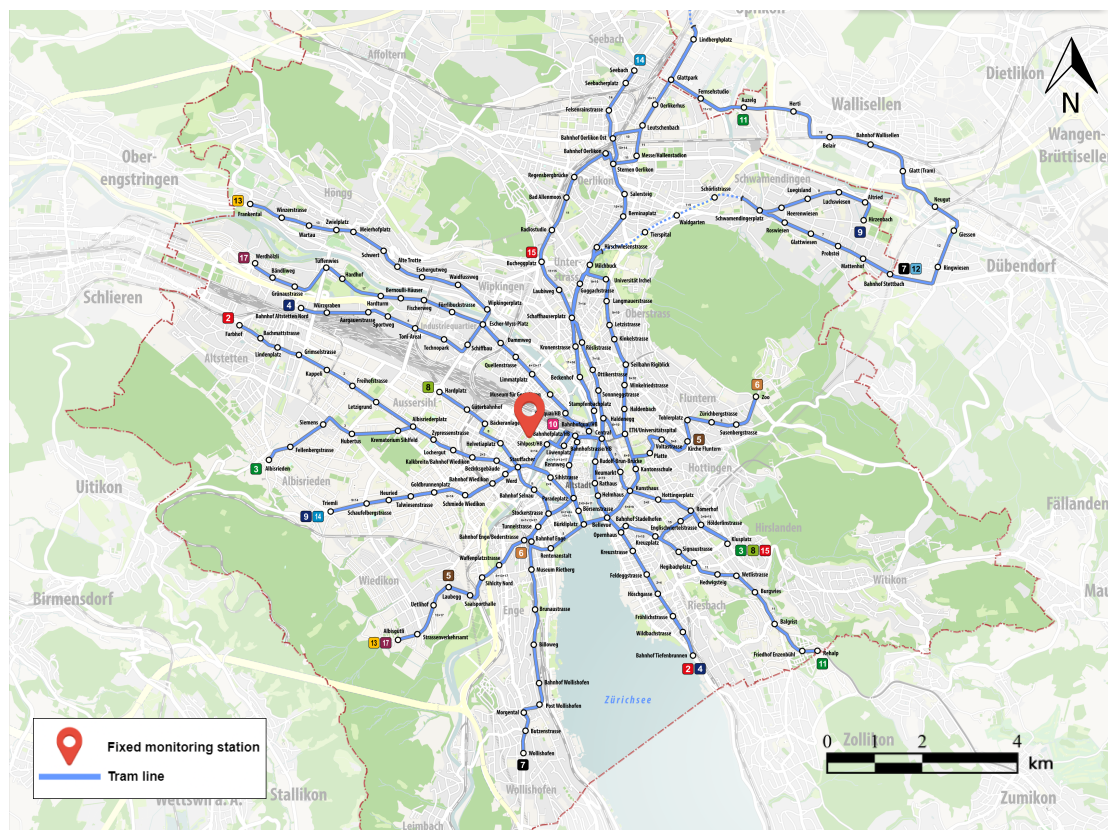


Figure 1.5 – Emplacement de la station fixe de référence et des lignes de tramway dans la ville de Zurich.

pondre à l'unité des données provenant de la station de surveillance fixe, la formule suivante a été appliquée : $\mu\text{g}/\text{m}^3 = (\text{ppb}) \cdot (12.187) \cdot (M)/293$ où M est la masse moléculaire du polluant ozone ($M(\text{O}_3) = 48$). On suppose une pression atmosphérique de 1 atmosphère et une température de 20 °C.

Même si ces données présentent des inconvénients, notamment l'échantillonnage uniquement sur des trajectoires statiques de la ville (lignes de trams et de bus), elles n'en demeurent pas moins très précieuses pour l'application et l'évaluation de nouvelles approches de modélisation de la variabilité spatio-temporelle de la pollution en milieu urbain.

- **Jeu de données d'Atmotrack**

Atmotrack (atmotrack.fr) est une société française spécialisée dans la mesure de la qualité de l'air basée à Nantes. Elle dispose d'une flotte de capteurs mobiles bas coût placés sur des véhicules utilitaires tels que ceux utilisés par la Poste (Figure 1.6), ainsi que des capteurs fixes à bas coût placés à des endroits stratégiques de la ville de Nantes, notamment à proximité des stations fixes de mesure de la qualité de l'air.

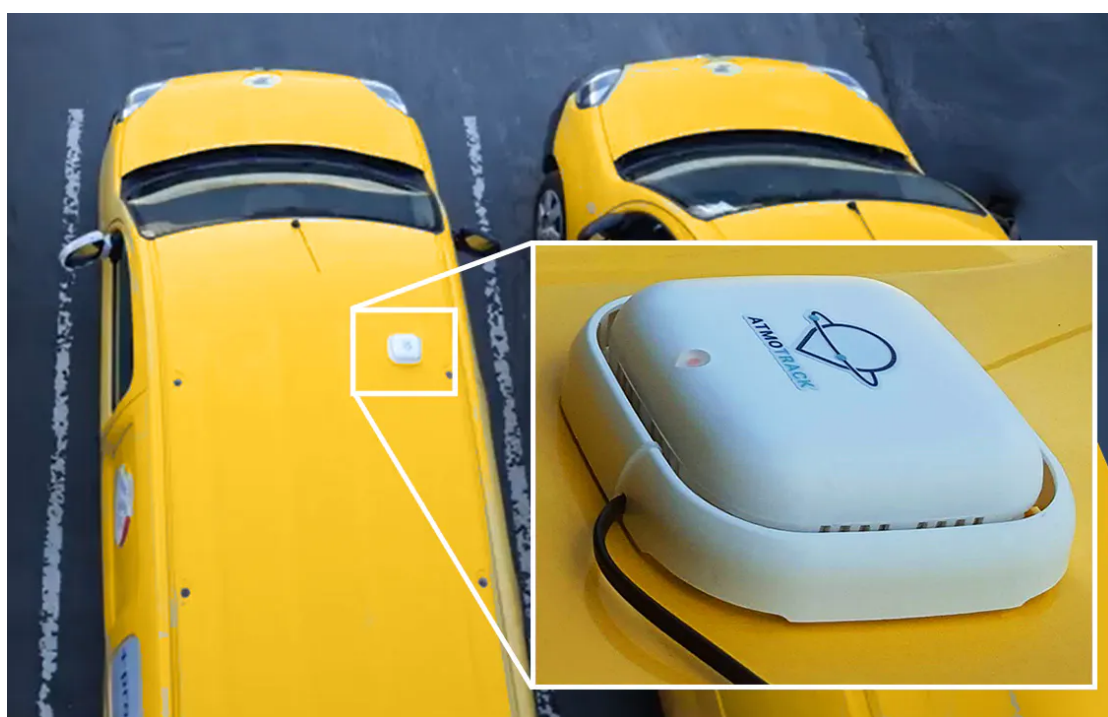


Figure 1.6 – Boîtier mobile d'Atmotrack incluant un capteur de pollution bas coût, installé sur le toit d'un véhicule de la Poste.

L'ensemble des données fournies par Atmotrack est diversifié et contient différents types de mesures de PM_{2,5} et de PM₁₀ collectées tout au long du mois de novembre 2018. Il convient de noter que les observations de PM₁₀ sont plus fréquentes dans les stations fixes de mesure de la qualité de l'air. Par conséquent, la concentration de PM₁₀ a été choisie comme polluant initial à étudier dans cette étude.

Les résultats de l'ensemble des données sur les vingt premiers jours sont principalement ceux obtenus à partir de capteurs fixes. Afin de garantir l'homogénéité de notre ensemble de données tout en conservant une quantité suffisante de données pour entraîner nos modèles, nous nous sommes limités aux dix derniers jours de novembre 2018, plus précisément du 20 au 30 novembre 2018.

La majorité des observations sont concentrées dans et autour de la ville de Nantes. Cependant, une petite partie des données obtenues à partir des capteurs bas coût, qui n'étaient pas présents dans cette zone, a été exclue de l'analyse. La figure 1.7 représente la zone d'étude qui a été retenue.

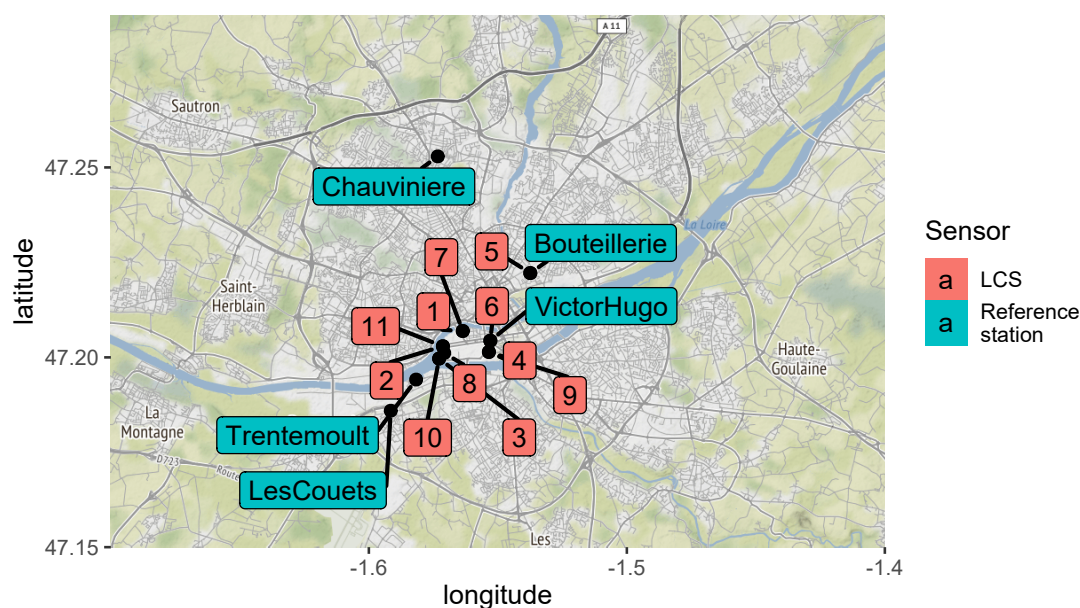


Figure 1.7 – Emplacement des capteurs fixes d'Atmotrack (LCS) et des stations de surveillance à Nantes

Étant donné que les capteurs mobiles sont installés sur le toit des véhicules utilitaires avec un échantillonnage opportuniste, la quantité de données collectées par ces capteurs tend à diminuer les soirs de semaine et

à cesser le dimanche. Pour maintenir la cohérence entre les données collectées par les capteurs fixes et mobiles, notre étude n'utilisera que les données collectées entre 4 heures et 19 heures du lundi au samedi inclus.

Sur les 15 capteurs fixes à faible coût, trois étaient situés à proximité de la station de surveillance fixe « Bouteillerie », tandis que trois autres étaient situés près de la station « Victor Hugo ». La figure 1.7 fournit une représentation visuelle de l'emplacement des stations fixes de mesure de la qualité de l'air et des capteurs fixes à faible coût. Afin d'éliminer la redondance des données, un seul capteur fixe par emplacement a été pris en compte.

Les capteurs bas coût ont été pré-calibrés avec les stations fixes par Atmotrack. Afin de préserver l'intégrité des données, une procédure mineure de post-traitement a été mise en œuvre. Plus précisément, les cinq premières minutes des parcours mobiles ont été exclues de l'ensemble des données, car il a été observé que pendant cette période, les capteurs mobiles produisaient des données déformées. Ensuite, une médiane de 15 observations par capteur a été appliquée pour lisser les erreurs des capteurs.

L'ensemble des données comprend un total de 132 891 observations, dont 81 461 observations collectées par des capteurs mobiles et 51 430 observations par des capteurs fixes. De plus amples informations concernant l'ensemble de données d'Atmotrack sont disponibles dans l'étude de Gressent et al.[30].

Ces données acquises présentent des avantages substantiels en comparaison avec celles obtenues via le réseau Opensense, notamment grâce à l'emploi d'une quantité supérieure de capteurs économiques. L'approche hybride adoptée, combinant des capteurs à bas coût tant fixes que mobiles, fournit une plateforme robuste pour étudier l'impact de leur répartition sur la qualité de la prédiction. Les capteurs mobiles, en particulier, offrent une flexibilité sans précédent puisqu'ils ne sont pas contraints par des itinéraires prédéfinis, permettant ainsi une collecte de données plus hétérogène et potentiellement plus représentative.

- **Jeu de données de l'expérimentation dans le quartier les Halles**

Le quartier des Halles, situé au cœur de Paris, est un carrefour majeur de la vie urbaine, combinant une activité commerciale intense, des zones piétonnes animées et une circulation routière dense. Sa position stratégique en fait un lieu idéal pour étudier la qualité de l'air, car il reflète les défis environnementaux auxquels sont confrontées de nombreuses métropoles modernes. De plus, la présence d'une station AirParif à proximité offre une opportunité unique de comparer les données recueillies avec des mesures de référence établies. Dans ce contexte, nous avons entrepris une série d'expérimentations visant à mesurer la qualité de l'air à l'aide de capteurs mobiles de pollution. Si les capteurs mobiles bas coût présentent l'avantage d'être économiquement accessibles, ils souffrent souvent de limitations en termes de précision et de fiabilité. Cette étude vise donc à pallier, même partiellement, ces défauts en combinant l'utilisation de ces capteurs avec des méthodologies rigoureuses et des outils de référence. L'objectif est de fournir un ensemble de données hétérogènes, permettant ainsi le développement de modèles mathématiques capables d'exploiter au maximum les informations recueillies pour une meilleure compréhension de la qualité de l'air.

Deux types de capteurs ont été déployés : cinq boîtiers mobiles, conçus en interne et équipés de capteurs bas coût, et un capteur de référence, le Fidas Frog.

Lors de la conception du plan d'expérimentation, un objectif clé était de maximiser les distances inter-capteurs. Pour ce faire, nous avons utilisé une modification du problème du postier chinois pour créer une trajectoire qui passe par toutes les rues. Cette approche nous a permis d'optimiser la couverture de la zone d'étude et d'assurer une collecte de données complète.

La trajectoire empruntée par les vélos équipés de capteurs est un élément essentiel de notre étude (Figure 1.8.) Chaque vélo était équipé d'un

capteur et suivait une trajectoire préétablie, passant par toutes les rues du quartier. Le circuit a été réalisé plusieurs fois pour collecter des données sur une plage temporelle plus importante tout en passant plus de fois au même endroit.

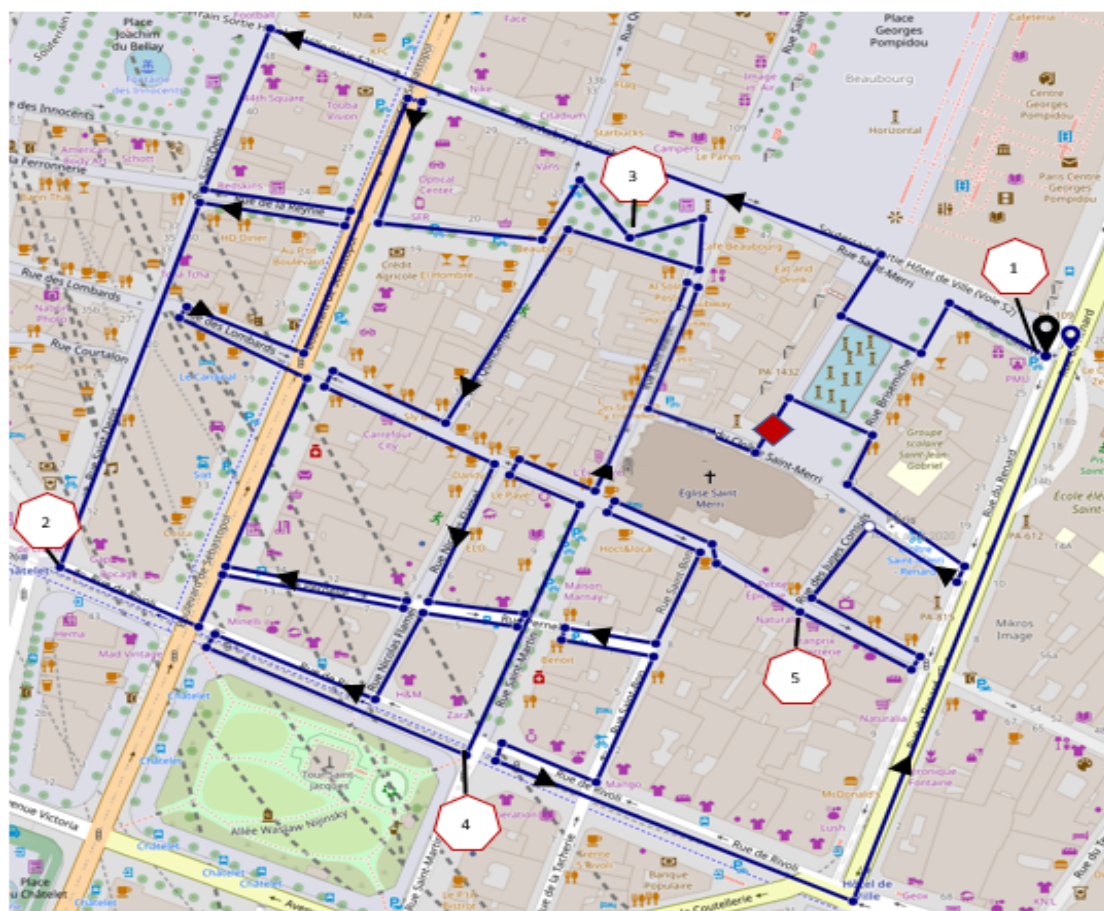


Figure 1.8 – Trajectoires des vélos équipés de capteurs mobiles dans le quartier des Halles

Pour optimiser la collecte de données, les cinq capteurs bas coût ont été initialement répartis à équidistance le long de la trajectoire. Les boîtiers 1, 2, 4 et 5 progressaient à une vitesse uniforme pour éviter les croisements. Le capteur mobile de référence, Fidas Frog, alternait avec chaque capteur bas coût pendant les rotations. Plusieurs rotations ont été effectuées pour une collecte de données étendue. La trajectoire englobait diverses conditions de mesure, y compris des rues à forte circulation, des zones piétonnes, des

parcs et des grandes avenues. Le capteur numéro 3 empruntait une trajectoire inverse, croisant ainsi les autres capteurs à chaque rotation.

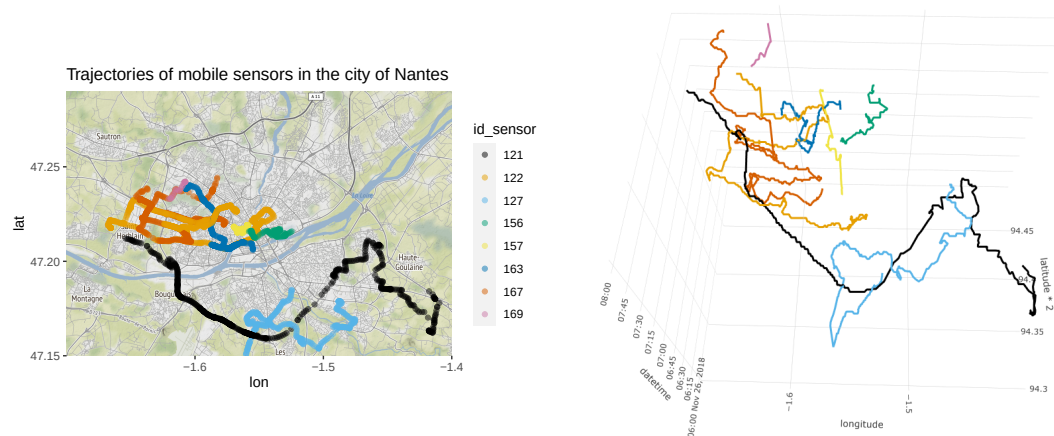
L'ensemble des données totalise 57 300 observations, dont 49 500 provenant des capteurs mobiles bas coût et 7 800 du capteur de référence Fidas Frog. Ces données ont été collectées lors de 5 tours, pendant lesquels le capteur de référence a alterné avec chaque capteur bas coût. Les mesures ont été prises entre 10h30 et 16h15 le 14 septembre 2020.

- **Simulation d'un jeu de données**

Les précédents jeux de données présentés sont idéaux pour appliquer des méthodes d'interpolation ou comparer la précision de la prédiction de ces dernières entre elles. Pour cela on peut utiliser une procédure dite de validation croisée. Cela dit, si l'on s'intéresse à l'optimisation de la phase d'estimation des paramètres d'un modèle bien précis, on doit connaître les vraies valeurs des paramètres qui sont à l'origine du jeu de données s'ils existent. Or, il n'est pas possible de les connaître dans les cas précédents. Pour remédier à ce problème, on a recours à la simulation de données selon un modèle avec des paramètres connus au préalable. L'objectif recherché ensuite est de comparer plusieurs méthodes ou configuration d'estimation des paramètres pour un même modèle afin de parfaire cette étape.

Il est important que la simulation soit représentative d'un vrai jeu de données pour que les résultats trouvés puissent être extrapolés et exploités avec des données réelles issues d'expérimentations in situ.

Les données proviennent de capteurs mobiles qui enregistrent les coordonnées GPS et l'heure de leurs mesures physiques respectives sont utilisées. Il convient de noter que les données de mesures de pollution ne constituent pas l'objet principal de ce jeu de données. Nous utilisons plutôt les données spatio-temporelles (coordonnées et horodatages) pour simuler le processus gaussien sur la base des paramètres estimés à partir des données de mesures réelles. Cette méthodologie garantit que nos données synthétisées reflètent fidèlement le modèle statistique que nous sou-



(a) Représentation spatiale des trajectoires dans la ville de Nantes. (b) Représentation spatio-temporelle des trajectoires.

Figure 1.9 – Représentation des trajectoires de capteurs mobiles réelles pour la simulation d'un jeu de données.

haitons étudier. Elle fournit donc un environnement expérimental contrôlé dans lequel nous connaissons précisément le véritable modèle statistique sous-jacent.

La représentation visuelle de ce processus de simulation, y compris les données de trajectoires utilisées, se trouve dans la figure 1.9. Cette figure donne un aperçu spatial des trajectoires des capteurs et de leurs progressions temporelles, ce qui permet de visualiser la base sur laquelle un processus gaussien est simulé.

L'ensemble des données simulées contient 3000 observations spatio-temporelles, provenant de 8 capteurs, sur une période de deux heures.

1.3 Conclusion

Ce premier chapitre de la thèse a éclairé le contexte complexe et les défis inhérents à la cartographie de la qualité de l'air en utilisant des capteurs mobiles, soulignant l'importance et la difficulté de cette tâche. Il a également présenté la méthodologie et les ensembles de données envisagés pour relever ces défis, jetant les bases nécessaires pour les développements ultérieurs de cette recherche.

Le prochain chapitre se consacrera à fournir un état de l'art complet des méthodes de spatialisation. Il examinera les diverses études ayant fait usage de capteurs mobiles à bas coût pour créer des cartographies précises de la qualité de l'air. Cette analyse critique vise à souligner les contributions importantes et à identifier les lacunes dans le domaine, posant ainsi les fondations pour les travaux de recherche à venir.

Références

- [1] Evgenios Agathokleous et al. "Ozone affects plant, insect, and soil microbial communities : A threat to terrestrial ecosystems and biodiversity". In : *Science Advances* 6.33 (2020), eabc1176.
- [2] Manuel Aleixandre, Michel Gerboles et al. "Review of small commercial sensors for indicative monitoring of ambient gas". In : *Chem. Eng. Trans* 30 (2012).
- [3] Endawoke Amsalu et al. "Acute effects of fine particulate matter (PM 2.5) on hospital admissions for cardiovascular disease in Beijing, China : a time-series study". In : *Environmental Health* 18 (2019), p. 1-12.
- [4] Komalkirti Apte et Sundeep Salvi. "Household air pollution and its effects on health". In : *F1000Research* 5 (2016).
- [5] Th Becker et al. "Air pollution monitoring using tin-oxide-based microreactor systems". In : *Sensors and Actuators B : Chemical* 69.1-2 (2000), p. 108-119.
- [6] Ruwim Berkowicz et al. "Using measurements of air pollution in streets for evaluation of urban air quality—meterological analysis and model calculations". In : *Science of the total environment* 189 (1996), p. 259-265.
- [7] MK Bhanarkar et PM Korake. "Soil salinity and moisture measurement system for grapes field by wireless sensor network". In : *Cogent Engineering* 3.1 (2016), p. 1164021.
- [8] Lizan D Bloemsma, Gerard Hoek et Lidwien AM Smit. "Panel studies of air pollution in patients with COPD : systematic review and meta-analysis". In : *Environmental research* 151 (2016), p. 458-468.
- [9] Robert D Brook. "Why physicians who treat hypertension should know more about air pollution". In : *The Journal of Clinical Hypertension* 9.8 (2007), p. 629-635.
- [10] Julie Carré et al. "Does air pollution play a role in infertility? : a systematic review". In : *Environmental Health* 16 (2017), p. 1-16.

- [11] Nuria Castell et al. "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" In : *Environment international* 99 (2017), p. 293-302.
- [12] Núria Castell et al. "Real-world application of new sensor technologies for air quality monitoring". In : *ETC/ACM Technical Paper* 16 (2013), p. 34.
- [13] Charles E Catlett et al. "Array of things : a scientific research instrument in the public way : platform design and early lessons learned". In : *Proceedings of the 2nd international workshop on science of smart city operations and platforms engineering*. 2017, p. 26-33.
- [14] Anoop J Chauhan et Sebastian L Johnston. "Air pollution and infection in respiratory illness". In : *British medical bulletin* 68.1 (2003), p. 95-112.
- [15] Aaron J Cohen et al. "The global burden of disease due to outdoor air pollution". In : *Journal of Toxicology and Environmental Health, Part A* 68.13-14 (2005), p. 1301-1307.
- [16] Université d'Aarhus. *comparaison des normes de qualité de l'air*. Consulté le 1 octobre 2023. 2023. url : https://www2.dmu.dk/AtmosphericEnvironment/Expost/database/docs/AQ_limit_values.pdf.
- [17] Gennaro D'Amato et al. "Urban air pollution and climate change as environmental risk factors of respiratory allergy : an update". In : *Journal of Investigational Allergology and Clinical Immunology* 20.2 (2010), p. 95-102.
- [18] Lyndsey A Darrow et al. "Air pollution and acute respiratory infections among children 0-4 years of age : an 18-year time-series study". In : *American journal of epidemiology* 180.10 (2014), p. 968-977.
- [19] Christiana A Demetriou et al. "Biomarkers of ambient air pollution and lung cancer : a systematic review". In : *Occupational and environmental medicine* 69.9 (2012), p. 619-627.
- [20] Francesca Dominici et al. "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases". In : *Jama* 295.10 (2006), p. 1127-1134.

- [21] Yixing Du et al. "Air particulate matter and cardiovascular disease : the epidemiological, biomedical and clinical evidence". In : *Journal of thoracic disease* 8.1 (2016), E8.
- [22] Paul B English et al. "The Imperial County Community Air Monitoring Network : a model for community-based environmental monitoring for public health action". In : *Environmental health perspectives* 125.7 (2017), p. 074501.
- [23] KB Ewan et Roger Pamphlett. "Increased inorganic mercury in spinal motor neurons following chelating agents." In : *Neurotoxicology* 17.2 (1996), p. 343-349.
- [24] Stephen Feinberg et al. "Long-term evaluation of air sensor technology under ambient conditions in Denver, Colorado". In : *Atmospheric measurement techniques* 11.8 (2018), p. 4605.
- [25] George F Fine et al. "Metal oxide semi-conductor gas sensors in environmental monitoring". In : *sensors* 10.6 (2010), p. 5469-5502.
- [26] Dominique Gillis, Ivana Semanjski et Dirk Lauwers. "How to monitor sustainable mobility in cities? Literature review in the frame of creating a set of sustainable mobility indicators". In : *Sustainability* 8.1 (2015), p. 29.
- [27] Drew A Glencross et al. "Air pollution and its effects on the immune system". In : *Free Radical Biology and Medicine* 151 (2020), p. 56-68.
- [28] Diane R Gold et al. "Ambient pollution and heart rate variability". In : *Circulation* 101.11 (2000), p. 1267-1273.
- [29] Fernando Gozzi, Giancarlo Della Ventura et Augusto Marcelli. "Mobile monitoring of particulate matter : State of art and perspectives". In : *Atmospheric Pollution Research* 7.2 (2016), p. 228-234.
- [30] Alicia Gressent et al. "Data fusion for air quality mapping using low-cost sensor observations : Feasibility and added-value". In : *Environment international* 143 (2020), p. 105965.
- [31] David Hasenfratz et al. "Deriving high-resolution urban air pollution maps using mobile sensor nodes". In : *Pervasive and Mobile Computing* 16 (2015), p. 268-285.

- [32] David Hasenfratz et al. "Pushing the spatio-temporal resolution limit of urban air pollution maps". In : *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE. 2014, p. 69-77.
- [33] Enzhu Hu et al. "Effect of elevated tropospheric ozone on soil carbon and nitrogen : a meta-analysis". In : *Environmental Research Letters* 17.4 (2022), p. 043001.
- [34] Wei Huang et al. "An overview of air quality analysis by big data techniques : Monitoring, forecasting, and traceability". In : *Information Fusion* 75 (2021), p. 28-40.
- [35] Yacine Mohamed Idir et al. "Mapping urban air quality from mobile sensors using spatio-temporal geostatistics". In : *Sensors* 21.14 (2021), p. 4717.
- [36] JJ Jaakkola et al. "Low-level air pollution and upper respiratory infections in children." In : *American Journal of Public Health* 81.8 (1991), p. 1060-1063.
- [37] Lars Järup. "Hazards of heavy metal contamination". In : *British medical bulletin* 68.1 (2003), p. 167-182.
- [38] Michael Jerrett et al. "Particulate air pollution, social confounders, and mortality in small areas of an industrial city". In : *Social science & medicine* 60.12 (2005), p. 2845-2863.
- [39] Karoline K Johnson et al. "Field test of several low-cost particulate matter sensors in high and low concentration urban environments". In : *Aerosol and Air Quality Research* 18.3 (2018), p. 565.
- [40] Pavlos S Kanaroglou et al. "Establishing an air pollution monitoring network for intra-urban population exposure assessment : A location-allocation approach". In : *Atmospheric Environment* 39.13 (2005), p. 2399-2409.
- [41] Klea Katsouyanni. "Ambient air pollution and health". In : *British medical bulletin* 68.1 (2003), p. 143-156.
- [42] Kirsten A Koehler et Thomas M Peters. "New methods for personal exposure monitoring for airborne particles". In : *Current environmental health reports* 2 (2015), p. 399-411.

- [43] Prashant Kumar et al. "The rise of low-cost sensing for managing air pollution in cities". In : *Environment international* 75 (2015), p. 199-205.
- [44] Stephen M Lasley et Mary E Gilbert. "Glutamatergic components underlying lead-induced impairments in hippocampal synaptic plasticity." In : *Neurotoxicology* 21.6 (2000), p. 1057-1068.
- [45] Stephen M Lasley, Mary C Green et Mary E Gilbert. "Rat hippocampal NMDA receptor binding as a function of chronic lead exposure level". In : *Neurotoxicology and teratology* 23.2 (2001), p. 185-189.
- [46] Truong Giang Le et al. "Effects of short-term exposure to air pollution on hospital admissions of young children for acute lower respiratory infections in Ho Chi Minh City, Vietnam." In : *Research Report (Health Effects Institute)* 169 (2012), p. 5-72.
- [47] Duk-Dong Lee et Dae-Sik Lee. "Environmental gas sensors". In : *IEEE sensors journal* 1.3 (2001), p. 214-224.
- [48] Jos Lelieveld et al. "The contribution of outdoor air pollution sources to premature mortality on a global scale". In : *Nature* 525.7569 (2015), p. 367-371.
- [49] Alastair Lewis et Peter Edwards. "Validate personal air-pollution sensors". In : *Nature* 535.7610 (2016), p. 29-31.
- [50] Chengqian Li et al. "Mechanisms in endocrinology : main air pollutants and diabetes-associated mortality : a systematic review and meta-analysis". In : *European journal of endocrinology* 171.5 (2014), R183-R190.
- [51] Jason Jingshi Li et al. "Sensing the air we breathe-the opensense zurich dataset". In : *Proceedings of the national conference on artificial intelligence*. T. 1. CONF. 2012, p. 323-325.
- [52] Jinhui Li et al. "Major air pollutants and risk of COPD exacerbations : a systematic review and meta-analysis". In : *International journal of chronic obstructive pulmonary disease* (2016), p. 3079-3091.

- [53] Qing Li et al. "Impacts of household coal and biomass combustion on indoor and ambient air quality in China : Current status and implication". In : *Science of the Total Environment* 576 (2017), p. 347-361.
- [54] Colin D Mathers, Ties Boerma et Doris Ma Fat. "Global and regional causes of death". In : *British medical bulletin* 92.1 (2009), p. 7-32.
- [55] MI Mead et al. "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks". In : *Atmospheric Environment* 70 (2013), p. 186-203.
- [56] Lidia Morawska et al. "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment : How far have they gone?" In : *Environment international* 116 (2018), p. 286-299.
- [57] Robert D Morris. "Airborne particulates and hospital admissions for cardiovascular disease : a quantitative review of the evidence." In : *Environmental health perspectives* 109.suppl 4 (2001), p. 495-500.
- [58] Said Munir et al. "Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—A case study in Sheffield". In : *Environmental monitoring and assessment* 191 (2019), p. 1-22.
- [59] Zeina Nasser et al. "Outdoor particulate matter (PM) and associated cardiovascular diseases in the Middle East". In : *International Journal of Occupational Medicine and Environmental Health* 28.4 (2015), p. 641-661.
- [60] Amanda L Northcross et al. "A low-cost particle counter as a realtime fine-particle mass monitor". In : *Environmental Science : Processes & Impacts* 15.2 (2013), p. 433-439.
- [61] Bart Ostro et al. "Air pollution and mortality : results from a study of Santiago, Chile." In : *Journal of exposure analysis and environmental epidemiology* 6.1 (1996), p. 97-114.
- [62] Ubiratan de Paula Santos et al. "Effects of air pollution on blood pressure and heart rate variability : a panel study of vehicular traffic controllers in the city of Sao Paulo, Brazil". In : *European heart journal* 26.2 (2005), p. 193-200.

- [63] Jennifer L Peel et al. "Ambient air pollution and cardiovascular emergency department visits in potentially sensitive groups". In : *American Journal of Epidemiology* 165.6 (2007), p. 625-633.
- [64] Michael P Perring et al. "Out of sight, Out of mind—but not Out of scope : the need to consider ozone (O₃) in restoration science, policy, and practice". In : *Restoration Ecology* 30.7 (2022), e13622.
- [65] Ricardo Piedrahita et al. "The next generation of low-cost personal air quality sensors for quantitative exposure monitoring". In : *Atmospheric Measurement Techniques* 7.10 (2014), p. 3325-3336.
- [66] C Arden Pope III et al. "Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution". In : *Circulation* 114.23 (2006), p. 2443-2448.
- [67] Ulrich Pöschl. "Atmospheric aerosols : composition, transformation, climate and health effects". In : *Angewandte Chemie International Edition* 44.46 (2005), p. 7520-7540.
- [68] Aakash C Rai et al. "End-user perspective of low-cost sensors for outdoor air pollution monitoring". In : *Science of The Total Environment* 607 (2017), p. 691-705.
- [69] Ranjit Nihal Ratnaike. "Acute and chronic arsenic toxicity". In : *Postgraduate medical journal* 79.933 (2003), p. 391-396.
- [70] Wendy Roberts. "Air pollution and skin disorders". In : *International Journal of Women's Dermatology* 7.1 (2021), p. 91-97.
- [71] A Scaperdas et RN Colvile. "Assessing the representativeness of monitoring data from an urban intersection site in central London, UK". In : *Atmospheric Environment* 33.4 (1999), p. 661-674.
- [72] Arnold Schecter et al. "Dioxins : an overview". In : *Environmental research* 101.3 (2006), p. 419-428.
- [73] J Schwartz et al. "Air pollution and acute respiratory illness in five German communities". In : *Environmental Research* 56.1 (1991), p. 1-14.

- [74] Ivana Semanjski et al. "Integrating big data into a sustainable mobility policy 2.0 planning support system". In : *Sustainability* 8.11 (2016), p. 1142.
- [75] Ivana Semanjski et al. "Policy 2.0 platform for mobile sensing and incentivized targeted shifts in mobility behavior". In : *Sensors* 16.7 (2016), p. 1035.
- [76] Emily G Snyder et al. "The changing paradigm of air pollution monitoring". In : *Environmental science & technology* 47.20 (2013), p. 11369-11377.
- [77] Sinan Sousan et al. "Evaluation of the Alphasense optical particle counter (OPC-N2) and the Grimm portable aerosol spectrometer (PAS-1.108)". In : *Aerosol Science and Technology* 50.12 (2016), p. 1352-1365.
- [78] Laurent Spinelle et al. "Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B : NO, CO and CO₂". In : *Sensors and Actuators B : Chemical* 238 (2017), p. 706-715.
- [79] Joseph R Stetter et Jing Li. "Amperometric gas sensors a review". In : *Chemical reviews* 108.2 (2008), p. 352-366.
- [80] A Thomas et J Gebhart. "Correlations between gravimetry and light scattering photometry for atmospheric aerosols". In : *Atmospheric Environment* 28.5 (1994), p. 935-938.
- [81] Frank Thömke et Hanns Christian Hopf. "Pontine lesions mimicking acute peripheral vestibulopathy". In : *Journal of Neurology, Neurosurgery & Psychiatry* 66.3 (1999), p. 340-349.
- [82] Wataru Tsujita et al. "Gas sensor network for air-pollution monitoring". In : *Sensors and Actuators B : Chemical* 110.2 (2005), p. 304-311.
- [83] Bruce Urch et al. "Acute blood pressure responses in healthy adults during controlled air pollution exposures". In : *Environmental health perspectives* 113.8 (2005), p. 1052-1055.
- [84] Jens Walkowiak et al. "Environmental exposure to polychlorinated biphenyls and quality of the home environment : effects on psychodevelopment in early childhood". In : *The Lancet* 358.9293 (2001), p. 1602-1607.

- [85] Yang Wang et al. "Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement". In : *Aerosol science and technology* 49.11 (2015), p. 1063-1077.
- [86] Gregory A Wellenius, Joel Schwartz et Murray A Mittleman. "Particulate air pollution and hospital admissions for congestive heart failure in seven United States cities". In : *The American journal of cardiology* 97.3 (2006), p. 404-408.
- [87] Jeffrey W White et Matthew M Conley. "A flexible, low-cost cart for proximal sensing". In : *Crop Science* 53.4 (2013), p. 1646-1649.
- [88] David E Williams et al. "Validation of low-cost ozone measurement instruments suitable for use in an air-quality monitoring network". In : *Measurement Science and Technology* 24.6 (2013), p. 065803.
- [89] Xiangxiang Xu et al. "Gotcha ii : Deployment of a vehicle-based environmental sensing system". In : *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. 2016, p. 376-377.
- [90] Xian-Jun Yu et al. "Characterization of somatic mutations in air pollution-related lung cancer". In : *EBioMedicine* 2.6 (2015), p. 583-590.
- [91] Yu Zheng, Furui Liu et Hsun-Ping Hsieh. "U-air : When urban air quality inference meets big data". In : *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, p. 1436-1444.

Chapitre 2

Vers une meilleure cartographie de la pollution : exploration des modèles et techniques existantes

Points clés du chapitre

- Présentation des quatre grandes familles de spatialisation de données utilisées avec les stations fixes : interpolation déterministes, géostatistique, modèles d'utilisation du sol, modèles de dispersion physique.
- Constat que les études qui cartographient la qualité de l'air en utilisant des capteurs mobiles bas coûts sont peu nombreuses.
- Constat que les études utilisent principalement des modèles d'utilisation du sol

Contenu du chapitre

2.1	Concepts fondamentaux des méthodes de spatialisation	50
2.2	Approches traditionnelles avec les stations fixes	51

2.2.1	Techniques d'interpolation déterministes	54
2.2.2	Méthodes géostatistiques	55
2.2.3	Modèles d'utilisation des sols	56
2.2.4	Modèles de dispersion	60
2.3	Spatialisation via les capteurs mobiles bas coût	63
2.4	Conclusion	66
	Références	67

La première phase de la méthodologie adoptée consiste en une analyse approfondie des méthodes et modèles décrits dans la littérature scientifique pour la cartographie de la qualité de l'air. Ce présent chapitre est dédié à cet objectif.

Tout d'abord, une clarification des définitions relatives aux méthodes et modèles de cartographie, ainsi qu'à leurs différentes échelles d'application est proposée. Par la suite, nous présenterons les diverses familles de méthodes utilisées pour la cartographie de la qualité de l'air. Historiquement, ces méthodes ont d'abord été appliquées en utilisant les données issues des stations fixes de référence.

L'avènement des capteurs bas coût étant une innovation relativement récente, ce type de capteur a suscité un moins grand nombre d'études (en comparaison avec les stations fixes) axées sur la cartographie de la qualité de l'air. Nous présenterons dans ce qui suit les études majeures qui ont réussi à exploiter efficacement ce type de données.

Nous concluons ce chapitre par une synthèse des points abordés et établissons les directions qui seront suivies par les chapitres ultérieurs du manuscrit.

2.1 Concepts fondamentaux des méthodes de spatialisation

Les termes « méthode », « modèle » et « technique » sont souvent utilisés de manière interchangeable dans le domaine de la pollution de l'air, mais ils possèdent des nuances. Une méthode fait généralement référence à une approche générale ou à une démarche à utiliser pour traiter un problème. Un modèle est un cadre théorique ou une représentation mathématique utilisé pour décrire le phénomène étudié. La technique se réfère à la mise en œuvre pratique du modèle à des endroits non échantillonnés.

De ce fait, un modèle est plus généraliste qu'une technique. Il décrit la dispersion des polluants et permet de simuler leur concentration en des points non échantillonnés afin de créer une cartographie. Une technique d'interpolation peut également produire des cartographies; cependant, elle ne fournit pas d'informations supplémentaires sur le processus sous-jacent.

Les méthodes de spatialisation décrivent les outils, techniques et modèles statistiques ou mathématiques qui permettent de transformer des observations ponctuelles en un panorama spatio-temporel continu. L'intérêt majeur de ces méthodes réside dans leur capacité à fournir une vision globale et complète des phénomènes étudiés, ce qui est essentiel pour la prise de décision et la planification stratégique, notamment en matière de santé publique et de politiques environnementales.

Avant d'explorer ce sujet en profondeur, clarifions deux concepts clés : interpolation et extrapolation. L'interpolation se concentre sur l'estimation ou la prédiction des valeurs à des emplacements non mesurés, mais entourés par des points de mesure connus, permettant ainsi une reconstitution de la distribution spatio-temporelle d'une variable. Plus délicate, l'extrapolation vise à prédire des valeurs en dehors des zones mesurées, en s'appuyant sur les tendances identifiées à partir des données existantes. Par conséquent, la prévision des données de pollution futures constitue une forme d'extrapolation, car elle anticipe les

données au-delà de l'intervalle temporel de collecte initiale.

Selon l'échelle de cartographie, des outils de modélisation adaptés sont requis. Par exemple, pour simuler des polluants à l'échelle européenne, une résolution de 50 km est adoptée, tandis qu'en France, cette résolution peut atteindre 5 km. Les cartographies et sorties des modèles régionaux peuvent ensuite être utilisées par les Associations Agréées de Surveillance de la Qualité de l'Air (AASQA) au niveau local comme conditions limites, enrichissant ainsi les modèles locaux de qualité de l'air en milieu urbain (Figure 2.1).

La section suivante décrit des différentes familles de méthodes utilisées pour la cartographie de la qualité de l'air en utilisant des stations fixes, posant ainsi les bases pour une compréhension approfondie et une discussion ultérieure sur leur pertinence et adaptabilité aux capteurs mobiles bas coût.

2.2 Approches traditionnelles avec les stations fixes

Actuellement, les cartographies de la qualité de l'air s'appuient principalement sur des données provenant de stations fixes de surveillance, stratégiquement positionnées pour représenter tant les zones urbaines que rurales. Ces stations fournissent des mesures continues et fiables des principaux polluants atmosphériques. De plus, des données satellites offrent des vues d'ensemble des concentrations de polluants à grande échelle, bien qu'avec une résolution spatiale moindre. Les modèles numériques de dispersion, qui intègrent des paramètres météorologiques et des données sur les émissions, sont couramment utilisés pour prédire la répartition spatiale des polluants. Ces prévisions sont souvent validées ou ajustées en fonction des observations au sol.

À grande échelle, notamment au niveau du continent européen, ce sont ces modèles numériques qui sont les plus utilisés. Ils reposent sur des outils numériques qui simulent les processus chimiques et physiques responsables de l'évolution des concentrations des polluants dans l'air. Ces modèles doivent être

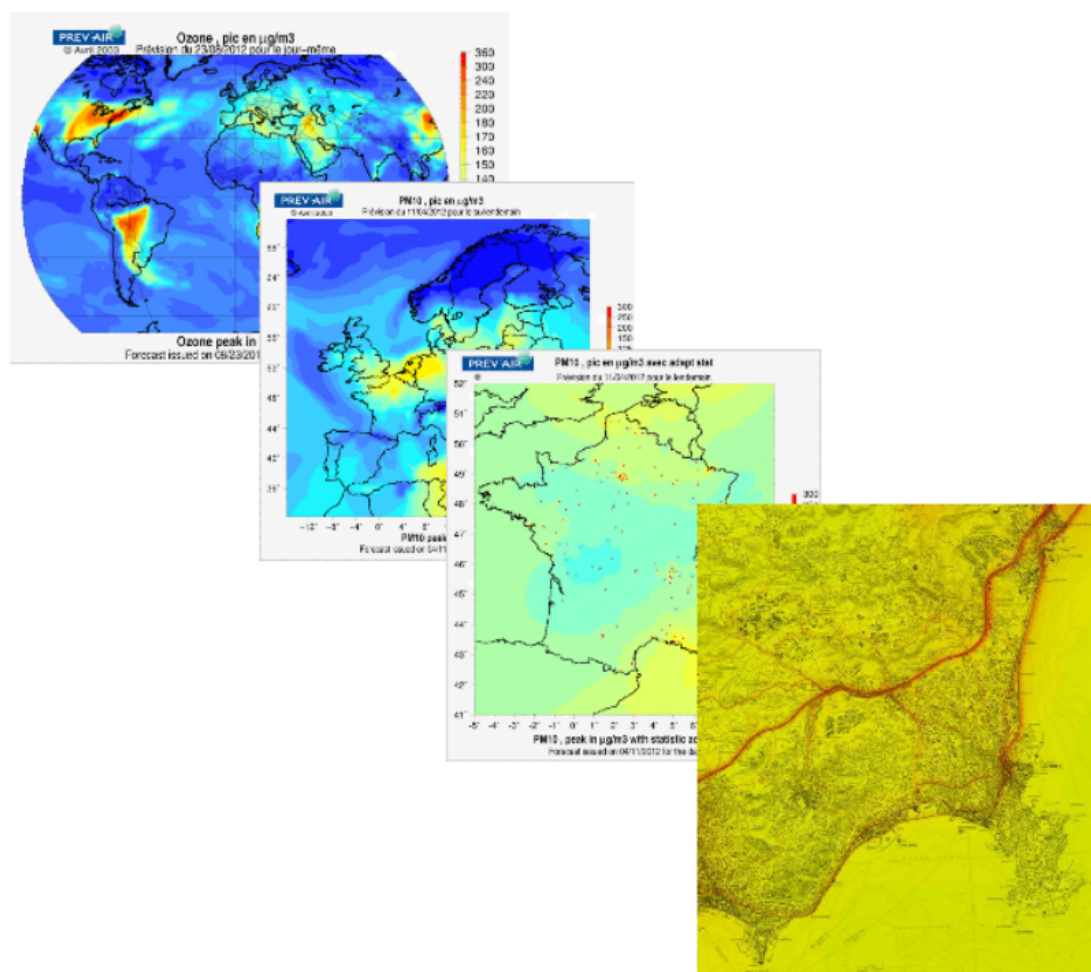


Figure 2.1 – Les différentes échelles de modélisation de la qualité de l'air. Du global au local (source : Michelot et al. 2015).

alimentés par plusieurs sources de données, dont les émissions de polluants issues de plusieurs secteurs d'activité (trafic routier, activités industrielles et agricoles, chauffage résidentiel, transport maritime...) et les conditions météorologiques qui influent sur la dispersion des polluants et sur l'intensité des processus chimiques.

PREV'AIR est un système collaboratif piloté par l'INERIS et soutenu par des entités telles que Météo-France et ATMO France. Il génère des cartographies de la qualité de l'air en combinant les données provenant des stations de mesure

au sol, des observations satellitaires et des modèles numériques. Ces cartographies offrent des prévisions quotidiennes et des évaluations des concentrations de divers polluants atmosphériques à l'échelle nationale et européenne.

Airparif, l'organisme de surveillance de la qualité de l'air en Île-de-France, utilise un réseau de stations de mesure pour collecter des données sur les polluants atmosphériques. Ces données, combinées à des modèles numériques, permettent de créer des cartographies représentant les concentrations de polluants à l'échelle régionale (Figure 2.2).

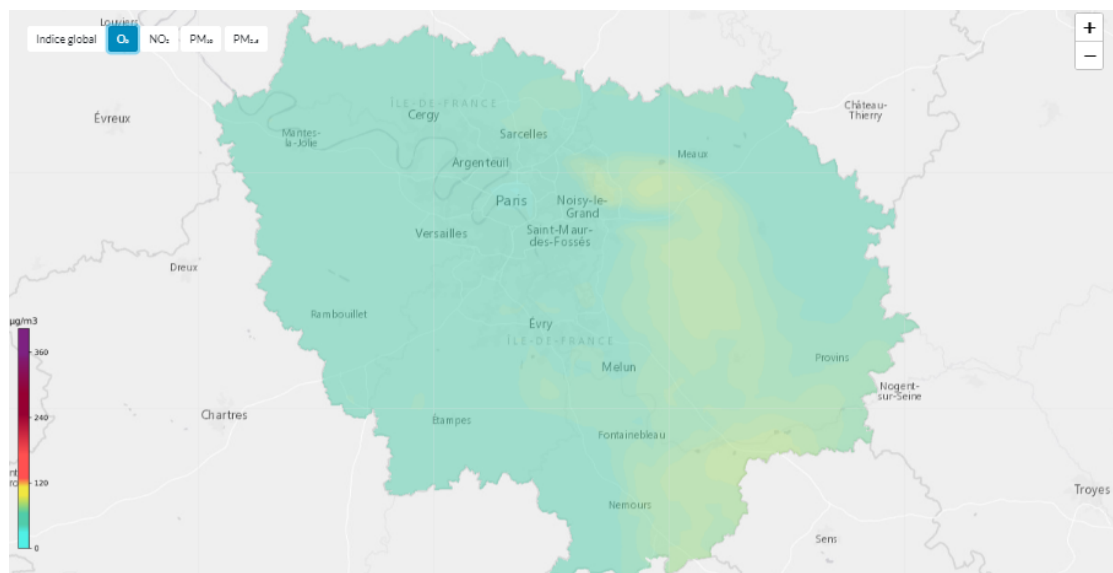


Figure 2.2 – Cartographie des niveaux d'ozone dans la région d'Île de France produite par Airparif. Source : "Airparif.fr" [3].

Il est clair que ces cartographies offrent des informations utiles pour une analyse générale. Cependant, pour répondre à certains besoins spécifiques, une précision spatiale accrue est nécessaire. Bien que les modèles numériques actuels fournissent des indications pertinentes, ils ne parviennent pas toujours à détailler les variations à l'échelle de la rue. Face à ce constat, d'autres méthodes, explorées intensément dans la littérature académique, méritent notre attention. Ces méthodes, détaillées par la suite, sont répertoriées en quatre familles distinctes : les techniques d'interpolation déterministes, la géostatistique, les mo-

dèles d'utilisation du sol et enfin les modèles de dispersions urbains.

2.2.1 Techniques d'interpolation déterministes

Toutes les techniques d'interpolation déterministes partagent en général la même base mathématique. Elles estiment toutes la valeur à un endroit non mesuré comme une moyenne pondérée des mesures effectuées dans les stations de surveillance environnantes. Elles diffèrent par le choix des poids d'échantillonnage et du choix des stations prises en compte. Trois techniques d'interpolation sont couramment utilisées dans le contexte de la cartographie de la qualité de l'air : la moyenne spatiale, le plus proche voisin, et la Pondération inverse à la distance (IDW) pour « Inverse Distance Weighting »

Le calcul de la moyenne spatiale consiste simplement à calculer la moyenne des mesures de polluants provenant des stations de surveillance proches situées à l'intérieur d'une grille prédéfinie, d'une région ou même d'une ville. Cette technique suppose que les mesures des différentes stations de surveillance ont une influence égale, quelle que soit leur distance par rapport au lieu non mesuré. Cette hypothèse n'est pas réaliste dans l'estimation de la pollution de l'air car elle omet toute variabilité spatiale [49, 50, 28].

La technique d'interpolation du plus proche voisin attribue les mesures de polluants de la station de surveillance la plus proche au site non mesuré, quelle que soit la distance qui les sépare [6, 42]. Le principal inconvénient de cette technique est que les mesures des autres points voisins sont ignorées.

La pondération inverse à la distance est l'une des techniques d'interpolation déterministe les plus couramment utilisées. La moyenne pondérée des données collectées par les capteurs déployés est utilisée pour déterminer la valeur à l'endroit non échantillonné. Cette stratégie accorde plus de poids aux endroits les plus proches, car on suppose que les mesures les plus proches ont un impact plus important sur la valeur que celles qui sont plus éloignées. En raison de sa facilité d'utilisation, cette approche est fréquemment appliquée comme référence. Elle a été utilisée par Marshall et al.[39] à Vancouver, au Canada, pour comparer

la variabilité urbaine des concentrations de NO et de NO₂ avec un modèle de régression et un modèle numérique. Wong et al.[60] ont comparé différentes méthodes d'interpolation, dont l'IDW, pour estimer les concentrations d'ozone et de particules.

La limite des techniques d'interpolation déterministe réside dans leur faible précision d'extrapolation. Ces techniques ne sont pas considérées comme des modèles car elles ne décrivent pas les données et n'indiquent pas le degré d'incertitude associé à la prédiction. En raison de ces limitations, elles ne sont plus couramment utilisées dans l'estimation de la pollution atmosphérique. Cependant, elles apparaissent généralement dans les études comparatives des méthodes d'interpolation spatiale [11, 53].

2.2.2 Méthodes géostatistiques

À l'instar des techniques d'interpolation déterministes, les méthodes géostatistiques s'appuient sur l'idée que les mesures à proximité ont plus d'influence. Toutefois, dans ce contexte, les mesures sont traitées comme des variables aléatoires.

La géostatistique prédit des valeurs à des endroits non échantillonnés en utilisant une combinaison pondérée d'observations. Les poids sont déterminés par la structure de covariance des données déduite de la structure réelle des données. En géostatistique, deux formes principales de krigeage sont couramment utilisées : le krigeage ordinaire et le krigeage universel. Le krigeage ordinaire suppose une moyenne constante inconnue dans un voisinage local, tandis que le krigeage universel suppose un modèle de tendance général polynomial descriptible.

La littérature présente de nombreuses études où des modèles géostatistiques ont été appliqués avec succès pour créer des cartographies de la qualité de l'air :

Kim et al.[27] ont développé un modèle de prédiction par Krigeage Ordinaire (OK) pour prédire les concentrations de particules à long terme dans sept

grandes villes coréennes. Whitworth et al.[57] ont modélisé les niveaux de benzène dans l'air ambiant dans un environnement urbain à l'aide de la géostatistique. Schneider et al.[48] quant à eux ont réussi à modéliser la concentration de dioxyde d'azote à Oslo, en utilisant la sortie d'un modèle de dispersion comme covariable dans un modèle géostatistique.

Liu et al. [31] ont étudié la variation spatiale de l'ozone pendant 12 heures en utilisant des mesures obtenues à partir de 19 stations de surveillance à Toronto, au Canada. Un modèle de krigeage ordinaire a été utilisé pour estimer les niveaux de concentration d'ozone et les résultats ont montré que la technique de krigeage était plus précise que la méthode du plus proche voisin. Ferreira et al. [14] ont mesuré les concentrations de pollution atmosphérique dans la partie centrale de Lisbonne en utilisant un modèle de krigeage ordinaire à partir de neuf stations de surveillance. Ils ont également calculé les corrélations entre le trafic routier et les niveaux de trois polluants (CO, NO et NO₂). Janssen et al. [25] ont appliqué un modèle géostatistique qui a permis d'estimer les niveaux de concentration des polluants sur une grille de 4×4 km² en Belgique. Une procédure de validation croisée a démontré la supériorité de ce modèle sur les autres techniques d'interpolation.

Plus sophistiquée que les technique d'interpolation déterministes, la géostatistique fournit également l'incertitude associée à la prédiction. Cependant, ces techniques souffrent d'un coût de calcul relativement élevé.

2.2.3 Modèles d'utilisation des sols

Ces modèles partent du principe que les facteurs environnementaux locaux, tels que l'utilisation des sols, les facteurs météorologiques, la densité des bâtiments et la densité du trafic, sont les seuls facteurs qui influencent la qualité de l'air dans un lieu spécifique.

Ils établissent un lien entre les facteurs prédictifs environnementaux sélectionnés et les mesures de la qualité de l'air disponibles. Connaissant la valeur de ces facteurs environnementaux locaux à des endroits non échantillonnés, on peut

reconstruire la concentration du polluant étudié à ces endroits.

Bien qu'il soit possible d'intégrer de nombreuses covariables autres que l'utilisation des sols, telles que des variables météorologiques (température, humidité, etc.) ou de trafic routier, le terme "land-use models" (modèles d'utilisation des sols) est resté prédominant pour désigner ce type de modèles.

Les modèles d'utilisation des sols ont deux sous-familles : les modèles de régression et l'apprentissage automatique. La distinction des deux sous familles dépend de la complexité du lien entre la mesure de polluant et les covariables gardées, faisant le compromis entre complexité et pouvoir descriptif (Figure 2.3).

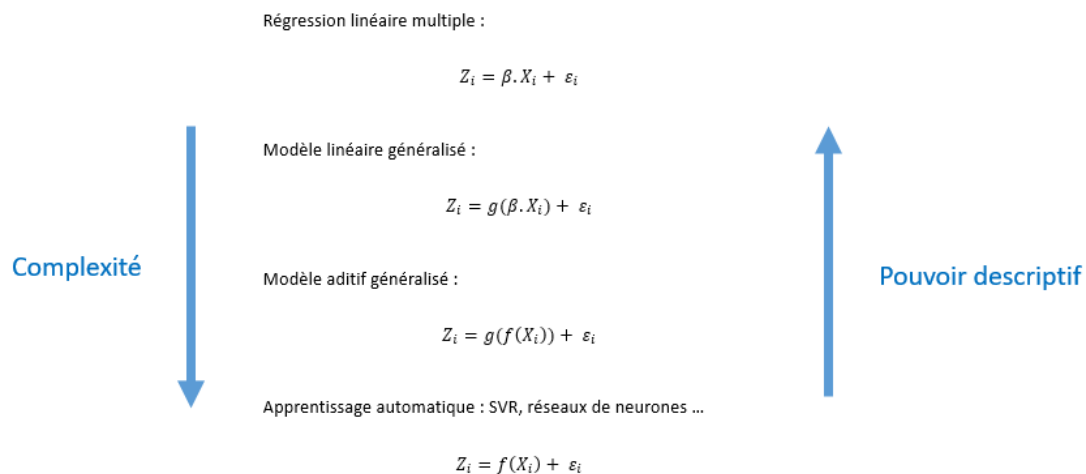


Figure 2.3 – Principaux modèles d'utilisation des sols présents dans la littérature, allant d'une complexité croissante.

A - Modèles de régressions

Communément appelés LUR pour "Land-Use-Regression", ces modèles supposent généralement un lien linéaire entre les variables prédictives et les mesures de polluants.

Le projet SAVIAH (Small Area Variations in Air quality and Health) est la première étude à modéliser les variations à petite échelle des polluants atmosphé-

riques par LUR [7] et démontre que les LUR sont un outil robuste pour prédire la qualité de l'air à une résolution spatiale fine avec des données de surveillance limitées.

Le bon choix des variables joue un rôle important dans la construction des modèles LUR. Ross et al. [45] montrent qu'avec les variables sur le trafic et les variables d'occupation des sols, leur modèle LUR explique plus de 60 % des variations des concentrations de PM_{2,5} sur une vaste zone. Dans une autre étude, les auteurs ont réussi à augmenter la variation expliquée de la pollution atmosphérique de 54 à 79 % en ajoutant une variable relative au trafic (trafic dans un rayon de 300 m autour des stations de surveillance) [46].

En utilisant la région métropolitaine de Houston (États-Unis) comme étude de cas, Zhai et al. [62] ont souligné la nécessité d'approfondir les défis liés à l'échelle spatiale. Ils soulignent que l'influence d'un prédicteur à une certaine distance ou dans un certain rayon dans une zone d'étude peut ne pas être cohérente dans une autre. De manière impressionnante, leur modèle LUR enregistre un taux d'erreur moyen inférieur à 20 % et affiche une valeur R^2 de 0,78. Cette performance dépasse celle du krigeage ordinaire lorsque des variables d'échelles spatiales optimisées sont utilisées.

Bien que les chercheurs aient choisi différentes formes de modélisation dans leurs modèles LUR, la plupart d'entre eux ont utilisé la régression linéaire pour modéliser la relation entre le niveau de polluants et les variables prédictives [32, 59, 35, 40, 26, 1, 5].

Ryan et al [47] résument l'histoire et les applications des modèles de régression de l'utilisation des sols, qui ont été largement utilisés pour caractériser l'exposition à la pollution atmosphérique intra-urbaine. Hoek et al. [23] analysent 25 cas de LUR et concluent que les LUR sont plus performants que d'autres techniques, telles que le krigeage et les modèles de dispersion.

Les modèles LUR sont donc extrêmement populaires pour la cartographie

fine de la qualité de l'air ; néanmoins, elles supposent un lien « simple » entre la variable d'intérêt (la pollution atmosphérique) et les variables prédictives, ce qui n'est pas forcément le cas.

B - Apprentissage automatique

Lorsqu'il s'agit d'obtenir de nouvelles informations à partir d'ensembles de données vastes, variés et dynamiques, l'apprentissage automatique s'avère particulièrement utile. Un algorithme d'apprentissage automatique analyse les données d'apprentissage (les mêmes données qu'utilisent les LUR) et produit une fonction déduite qui peut être utilisée pour la prédiction dans des endroits non échantillonnés. En général, cette fonction est assez complexe et ne peut pas être interprétée directement par l'utilisateur.

De nombreuses études ont appliqué cette méthode pour prédire les niveaux de pollution de l'air : Singh et al. [51] ont identifié les sources de pollution et prédit la qualité de l'air urbain à l'aide de méthodes d'apprentissage d'ensemble. Certains algorithmes d'apprentissage automatique ont été combinés avec des modèles flous pour prédire les niveaux de pollution de l'air [10].

Cabaneros et al. [9] ont fourni un inventaire des modèles de réseaux neuronaux artificiels (ANN) pour la prévision de la pollution de l'air.

En utilisant les résultats du modèle numérique comme données d'entrée en plus des mesures de polluants atmosphériques, Xi et al. [61] ont réussi à développer un modèle d'apprentissage automatique pour améliorer les performances de prédiction. Ils ont évalué cinq algorithmes d'apprentissage automatique différents sur quatre groupes différents d'ensembles de données où chaque groupe comprend différentes caractéristiques d'entrée (par exemple, observation de la pollution, prévisions météorologiques, vitesse du vent, etc). Le modèle des forêts aléatoires a été le plus performant dans la plupart des groupes. Les auteurs concluent que la disponibilité de plus d'informations augmente la possibilité d'améliorer la précision du modèle.

Brokamp et al. [8] affirment qu'en combinant les forêts aléatoires avec des variables qui ont un impact significatif sur la pollution de l'air, comme les variables d'utilisation des sols, il devient possible de dépasser les limites des LUR

dans la capture des relations non linéaires et des interactions complexes entre les prédicteurs et la pollution de l'air, même avec des données d'entraînement de petite taille.

L'apprentissage automatique, avec sa capacité à modéliser des relations complexes et non linéaires, offre une alternative prometteuse aux modèles de régression traditionnels. Cependant, ces algorithmes sont souvent perçus comme des « boîtes noires », ce qui signifie qu'ils peuvent produire des prédictions précises sans nécessairement fournir une explication claire de la manière dont ils sont arrivés à ces prédictions. Cette caractéristique peut poser problème, en particulier dans des domaines tels que la qualité de l'air où la compréhension des mécanismes sous-jacents est cruciale pour élaborer des politiques et des interventions efficaces.

Plus généralement, les modèles d'utilisation des sols s'appuient fortement sur la disponibilité d'autres variables, sans prendre en compte les effets spatiaux tels que l'autocorrélation spatiale et temporelle des mesures.

2.2.4 Modèles de dispersion

Fréquemment employés à une large échelle, telle qu'une région ou un pays, les modèles de dispersion sont également adaptés à des échelles plus restreintes, comme celle d'une ville.

Ces modèles sont basés sur la description des mécanismes physiques et chimiques, ainsi que sur la résolution numérique d'équations fondées sur les principes physiques tels que la thermodynamique et la mécanique des fluides. La mise en œuvre de ces modèles exige une compréhension approfondie des processus en jeu. Pour un polluant comme les particules, notamment dans un environnement à topographie prononcée et près des sources, les mécanismes physiques concernés sont à la fois extrêmement complexes et non entièrement connus.

Les modèles de dispersion reproduisent la manière dont les processus phy-

siques et chimiques conduisent aux concentrations de polluants atmosphériques. Ils font intervenir des facteurs environnementaux, tels que ceux utilisés dans les modèles LUR, et ont été fréquemment appliqués à la prévision de la pollution liée au trafic.

Les modèles déterministes se classent eux aussi en trois catégories selon leur approche de calcul : les modèles eulériens, lagrangiens, et gaussiens. De manière simplifiée, ces modèles demandent différents jeux de données en entrée (sources de pollution, géographie, météorologie, etc.) [52]. Ils résolvent les équations gouvernant les phénomènes atmosphériques, restituent une cartographie des champs de polluants et établissent des prévisions.

Vardoulakis et al. [56] ont étudié quatre modèles de dispersion gaussiens utilisés dans les canyons de rue. Hamer et al. [18] ont décrit le modèle eulérien de dispersion EPISODE et son application à la modélisation de la concentration du NO_2 . Fallah et al. [13] ont utilisé quant à eux un modèle de dispersion gaussien pour améliorer la caractérisation de la pollution atmosphérique à proximité des routes. Gibson et al. [15] ont eux aussi utilisé le modèle de dispersion gaussien pour évaluer les émissions de particules, de NO_x et d'oxydes d'azote.

Le modèle eulérien déterministe de chimie-transport CHIMERE [34] dont la haute qualité scientifique est largement reconnue, est très utilisé en France, notamment par PREV'AIR. Il calcule les concentrations horaires de plusieurs dizaines de polluants (gazeux et particulaires). Les simulations peuvent couvrir des zones allant d'une échelle régionale, comme l'Ile de France, à une échelle continentale comme dans la figure 2.4. Le maillage des sorties est ajustable en fonction du coût de calcul engagé. Il peut être aussi précis que 2 km, grâce à l'utilisation de supercalculateurs, ou s'étendre jusqu'à 100 km pour des zones continentales.

Le principal avantage des modèles de dispersion est qu'ils ne nécessitent pas un réseau dense de stations de surveillance.

Les limites des modèles de dispersion sont principalement dues aux données d'entrée et aux postulats liés à ces modèles. L'obtention d'une variété de

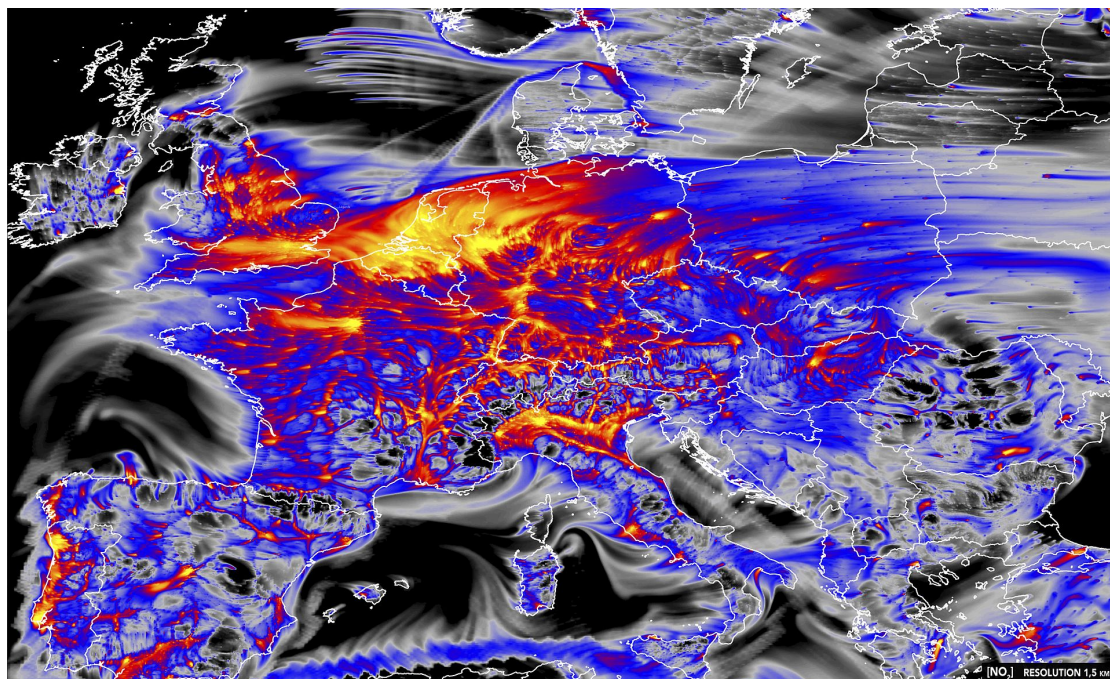


Figure 2.4 – Sorties du modèles CHIMERE pour la concentration du NO₂ avec une résolution spatiale de 2 km utilisant 2000 CPUs d'un super-calculateur.

données d'entrée peut s'avérer coûteuse et parfois nécessiter des ressources de calcul conséquentes. De plus, certaines des hypothèses sur lesquelles reposent ces modèles peuvent être inadaptées ou spécifiques à des contextes locaux, rendant complexe la modélisation des fluctuations aléatoires à petite échelle.

Méthodes hybrides

Au-delà de ces quatre familles de spatialisation, une autre catégorie se distingue : les méthodes « hybrides » qui incarnent la convergence entre différentes familles de méthodes que nous avons mentionnées.

Les modèles hybrides se composent généralement de deux algorithmes ou plus, qui aboutissent à un modèle de qualité de l'air plus solide, fournissant de meilleurs résultats qu'en utilisant une seule méthode. Par exemple, Wilton et al [58] utilisent les sorties du modèles de dispersion numérique comme nouvelle variable dans un modèle LUR. Ce modèle hybride permet d'améliorer les valeurs

de prédiction pour les deux villes étudiées (Seattle et Los Angeles). Schiender et al. [48] quand à eux ont utilisé les sorties du modèle de dispersion EPISODE comme entrée dans un modèle de géostatistique.

2.3 Spatialisation via les capteurs mobiles bas coût

Historiquement, les stations fixes ont servi de référence pour développer les approches de spatialisation. Ainsi, lors de l'émergence des capteurs bas coût, la première intuition a été d'appliquer ces approches éprouvées à ces nouvelles données.

Le principal défi de la jonction entre les méthodes conçues pour les réseaux de capteurs fixes et les données mobiles des capteurs bas coût réside dans la nature intrinsèquement différente des observations spatio-temporelles.

De plus, les capteurs mobiles ne couvrent pas constamment toutes les zones et ne fournissent pas de données continues pour des lieux spécifiques, comme dans des conditions de faible trafic ou pendant la nuit. Ainsi, une haute résolution spatiale est obtenue au détriment de la résolution temporelle pour ces données.

Zwack et al. ont combiné des modèles de régression basés sur des données de surveillance mobile avec un modèle de dispersion nommé « Quick Urban Industrial Complex » (QUIC) pour l'évaluation de la pollution de l'air [65]. Adams et Kanaroglou ont utilisé des réseaux neuronaux pour combiner les données de surveillance mobiles et fixes pour la prévision de la pollution de l'air [2].

Gressent et al. [16] ont proposé une cartographie de la ville de Nantes en utilisant des capteurs bas coût mobiles et fixes. Ils ont utilisé les sorties du modèle de dispersion (ADMS-Urban) comme entrée dans un modèle géostatistique. Hasenfratz et al. ont utilisé des modèles LUR non linéaires pour développer une carte de la pollution de l'air avec une résolution spatiale élevée de 100×100 m² pour Zurich, en Suisse [21].

Le tableau 2.1 énumère les études publiées utilisant des capteurs mobiles à faible coût pour la cartographie de la qualité de l'air. Comme le montre ce

Vers une meilleure cartographie de la pollution : exploration des modèles et techniques existantes

Article	Méthode	Région	Polluants	Véhicule
Marjovi et al. [37]	LUR	Lausanne, Suisse	UFP	Bus
Hart et al. [20]	LUR	Texas, USA	PM2.5	Vélos
Apte et al. [4]	Techniques déterministes	Oakland, USA	NO, NO ₂ , BC	Voitures
Hasenfratz et al. [22]	LUR	Zurich, Suisse	UFP	Trams
Hasenfratz et al. [21]	LUR	Zurich, Suisse	UFP	Trams
Marjovi et al. [38]	LUR	Lausanne, Suisse	UFP	Bus
Li et al. [29]	Géostatistiques	Zurich, Suisse	UFP	Trams
Lim et al. [30]	Apprentissage automatique	Séoul, Corée du Sud	PM2.5	Piétons
Adams et al. [2]	Apprentissage automatique	Hamilton, Canada	NO ₂	Fourgonnettes
Hankey et al. [19]	LUR	Minneapolis, USA	BC PM2.5	Vélos
Gressent et al. [16]	Géostatistiques	Nantes, France	PM10	Voitures
Do et al. [12]	Apprentissage automatique	Anvers, Belgique	Divers polluants	Vélos
Zhang et al. [63]	Apprentissage automatique	Songdo, Corée du Sud	CO ₂ , PM2.5, PM10	Voitures
Song et al. [54]	Apprentissage automatique	Pékin, Chine	PM2.5	Voitures
Van et al. [55]	LUR	Gand, Belgique	BC	Vélos
Guan et al. [17]	Géostatistiques	Oakland, USA	NO ₂	Voitures
Mariano et al. [36]	Apprentissage automatique	Zurich, Suisse	UFP	Trams
Ma et al. [33]	Apprentissage automatique	Chine	PM2.5	Voitures
Idir et al. [24]	Géostatistiques	Zurich, Suisse	O ₃	Tram
Zhao et al. [64]	Apprentissage automatique	Chine	PM2.5	Voitures
Qin et al. [44]	Apprentissage automatique	Chine	PM2.5	Voitures
Qin et al. [43]	Apprentissage automatique	Anvers, Belgique	NO ₂	Fourgonnettes
Messier et al. [41]	Géostatistiques	Oakland, USA	NO ₂ BC	Voitures

Table 2.1 – Études de cartographie de la qualité de l'air utilisant des capteurs mobiles.

tableau, différents modèles de différentes familles ont été utilisés pour relever le défi de la cartographie de la qualité de l'air urbain à l'aide de capteurs mobiles à faible coût.

On peut noter que les données issues des capteurs mobiles bas coût sont rarement exploitées en conjonction avec des modèles de dispersion numérique. Au lieu de cela, elles sont fréquemment associées à des modèles d'utilisation des sols. Cette préférence est intrinsèquement liée à la nature des données recueillies. Étant donné que ces données sont mobiles, elles capturent de manière plus exhaustive les caractéristiques spatiales de la zone étudiée. Par conséquent, elles peuvent établir avec une précision accrue le lien entre ces caractéristiques spatiales et les niveaux de pollution. Néanmoins, il convient de souligner que les

sorties des modèles de dispersion numérique peuvent servir d'entrée précieuse pour d'autres modèles, enrichissant ainsi leur pertinence et leur précision.

2.4 Conclusion

Dans la littérature scientifique, diverses méthodes de cartographie de la qualité de l'air ont été présentées. Les modèles de dispersion physico-chimique dominant particulièrement à l'échelle nationale ou continentale. Toutefois, à l'échelle régionale ou urbaine, bien qu'ils demeurent utilisés, ces modèles montrent certaines limites qui restreignent la possibilité de réaliser une cartographie détaillée de la qualité de l'air. Ces limites est attribuable à la complexité inhérente à la modélisation à petite échelle, où certaines variables peuvent être inconnues ou partiellement occultées.

En revanche, les modèles empiriques, qui établissent des liens entre les concentrations de polluants et d'autres variables, sont davantage privilégiés. Ils sont largement exploités pour la cartographie de la qualité de l'air à l'échelle urbaine, car ils sont mieux adaptés pour prendre en compte la nature volumineuse et bruitée des données produites par les capteurs mobiles bas coût.

Pour cette raison, l'attention a été portée sur les trois autres familles de modèles dans la suite de ce manuscrit. Bien que les résultats des modèles de dispersion seront utilisés, leur développement ne sera pas approfondi. L'étape suivante de la méthodologie consistera à comparer ces méthodes de cartographie à l'aide d'un ensemble de données recueillies via des capteurs bas coût. Aucune étude scientifique n'a encore effectué une comparaison de plusieurs familles de méthodes de cartographie sur un jeu de données issu de capteurs mobiles. Cette comparaison constituera le cœur du chapitre suivant.

Références

- [1] Ariane Adam-Poupart et al. "Spatiotemporal modeling of ozone levels in Quebec (Canada) : a comparison of kriging, land-use regression (LUR), and combined Bayesian maximum entropy-LUR approaches". In : *Environmental health perspectives* 122.9 (2014), p. 970-976.
- [2] Matthew D Adams et Pavlos S Kanaroglou. "Mapping real-time air pollution health risk for environmental management : Combining mobile and stationary air pollution monitoring with neural network models". In : *Journal of environmental management* 168 (2016), p. 133-141.
- [3] Airparif. *Cartographie des niveaux d'ozone dans la région d'île de France*. Consulté le 24 juin 2023. 2023. url : <https://www.airparif.asso.fr/surveiller-la-pollution/la-pollution-en-direct-en-ile-de-france>.
- [4] Joshua S Apte et al. "High-resolution air pollution mapping with Google street view cars : exploiting big data". In : *Environmental science & technology* 51.12 (2017), p. 6999-7008.
- [5] Bernardo S Beckerman et al. "Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California". In : *Atmospheric environment* 77 (2013), p. 172-177.
- [6] Michael Brauer et al. "A cohort study of traffic-related air pollution impacts on birth outcomes". In : *Environmental health perspectives* 116.5 (2008), p. 680-686.
- [7] David J Briggs et al. "Mapping urban air pollution using GIS : a regression-based approach". In : *International Journal of Geographical Information Science* 11.7 (1997), p. 699-718.
- [8] Cole Brokamp et al. "Exposure assessment models for elemental components of particulate matter in an urban environment : A comparison of regression and random forest approaches". In : *Atmospheric Environment* 151 (2017), p. 1-11.

- [9] Sheen Mclean Cabaneros, John Kaiser Calautit et Ben Richard Hughes. "A review of artificial neural network models for ambient air pollution prediction". In : *Environmental Modelling & Software* 119 (2019), p. 285-304.
- [10] M Cacciola et al. "Aspects about air pollution prediction on urban environment". In : *Proceedings of the 4th IMEKO TC19 Symposium on Environmental Instrumentation and Measurements*. 2013.
- [11] Despina Deligiorgi et Kostas Philippopoulos. "Spatial interpolation methodologies in urban air pollution modeling : application for the greater area of metropolitan Athens, Greece". In : *Advanced air pollution* 17 (2011), p. 341-362.
- [12] Tien Huu Do et al. "Graph-Deep-Learning-Based Inference of Fine-Grained Air Quality from Mobile IoT Sensors". In : *IEEE Internet of Things Journal* 7.9 (2020), p. 8943-8955.
- [13] Masoud Fallah-Shorshani, Maryam Shekarzifard et Marianne Hatzopoulou. "Integrating a street-canyon model with a regional Gaussian dispersion model for improved characterisation of near-road air pollution". In : *Atmospheric Environment* 153 (2017), p. 21-31.
- [14] Francisco Ferreira et al. "Air quality monitoring and management in Lisbon". In : *Environmental Monitoring and Assessment* 65 (2000), p. 443-450.
- [15] Mark D Gibson, Soumita Kundu et Mysore Satish. "Dispersion model evaluation of PM_{2.5}, NO_x and SO₂ from point and major line sources in Nova Scotia, Canada using AERMOD Gaussian plume air dispersion model". In : *Atmospheric Pollution Research* 4.2 (2013), p. 157-167.
- [16] Alicia Gressent et al. "Data fusion for air quality mapping using low-cost sensor observations : Feasibility and added-value". In : *Environment international* 143 (2020), p. 105965.
- [17] Yawen Guan et al. "Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles". In : *Journal of the American Statistical Association* 115.531 (2020), p. 1111-1124.

- [18] Paul D Hamer et al. "The urban dispersion model EPISODE. Part 1 : A Eulerian and subgrid-scale air quality model and its application in Nordic winter conditions". In : *Geoscientific Model Development Discussions* (2019), p. 1-57.
- [19] Steve Hankey et Julian D Marshall. "Land use regression models of on-road particulate air pollution (particle number, black carbon, PM_{2.5}, particle size) using mobile monitoring". In : *Environmental science & technology* 49.15 (2015), p. 9194-9202.
- [20] Ronan Hart, Lu Liang et Pinliang Dong. "Monitoring, mapping, and modeling spatial-temporal patterns of PM_{2.5} for improved understanding of air pollution dynamics using portable sensing technologies". In : *International journal of environmental research and public health* 17.14 (2020), p. 4914.
- [21] David Hasenfratz et al. "Deriving high-resolution urban air pollution maps using mobile sensor nodes". In : *Pervasive and Mobile Computing* 16 (2015), p. 268-285.
- [22] David Hasenfratz et al. "Pushing the spatio-temporal resolution limit of urban air pollution maps". In : *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE. 2014, p. 69-77.
- [23] Gerard Hoek et al. "A review of land-use regression models to assess spatial variation of outdoor air pollution". In : *Atmospheric environment* 42.33 (2008), p. 7561-7578.
- [24] Yacine Mohamed Idir et al. "Mapping urban air quality from mobile sensors using spatio-temporal geostatistics". In : *Sensors* 21.14 (2021), p. 4717.
- [25] Stijn Janssen et al. "Spatial interpolation of air pollution measurements using CORINE land cover data". In : *Atmospheric Environment* 42.20 (2008), p. 4884-4903.
- [26] Jules Kerckhoffs et al. "A national fine spatial scale land-use regression model for ozone". In : *Environmental research* 140 (2015), p. 440-448.
- [27] Sun-Young Kim et al. "Ordinary kriging approach to predicting long-term particulate matter concentrations in seven major Korean cities". In : *Environmental health and toxicology* 29 (2014).

- [28] PL Kinney et al. "Methods development for epidemiologic investigations of the health effects of prolonged ozone exposure. Part III. An approach to retrospective estimation of lifetime ozone exposure using a questionnaire and ambient monitoring data (US sites)." In : *Research Report (Health Effects Institute)* 81 (1998), p. 79-108.
- [29] Jason Jingshi Li et al. "Estimating urban ultrafine particle distributions with gaussian process models". In : *Research@ Locate14* (2014), p. 145-153.
- [30] Chris C Lim et al. "Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea". In : *Environment international* 131 (2019), p. 105022.
- [31] L-J Sally Liu et AJ Rossini. "Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto—a pilot study". In : *Environment International* 22.6 (1996), p. 677-692.
- [32] Wu Liu et al. "Land use regression models coupled with meteorology to model spatial and temporal variability of NO₂ and PM₁₀ in Changsha, China". In : *Atmospheric Environment* 116 (2015), p. 272-280.
- [33] Rui Ma et al. "Fine-Grained Air Pollution Inference with Mobile Sensing Systems : A Weather-Related Deep Autoencoder Model". In : *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.2 (2020), p. 1-21.
- [34] Sylvain Mailler et al. "CHIMERE-2017 : From urban to hemispheric chemistry-transport modeling". In : *Geoscientific Model Development* 10.6 (2017), p. 2397-2423.
- [35] Alessandro Marcon et al. "Development and transferability of a nitrogen dioxide land use regression model within the Veneto region of Italy". In : *Atmospheric Environment* 122 (2015), p. 696-704.
- [36] Pedro Mariano, Susana Marta Almeida et Pedro Santana. "Pollution Prediction Model Using Data Collected by a Mobile Sensor Network". In : *2020 5th International Conference on Smart and Sustainable Technologies (SpliTech)*. IEEE. 2020, p. 1-6.

- [37] Ali Marjovi, Adrian Arfire et Alcherio Martinoli. "Extending urban air quality maps beyond the coverage of a mobile sensor network : data sources, methods, and performance evaluation". In : *Proceedings of the International Conference on Embedded Wireless Systems and Networks*. CONF. 2017.
- [38] Ali Marjovi, Adrian Arfire et Alcherio Martinoli. "High resolution air pollution maps in urban environments using mobile sensor networks". In : *2015 International Conference on Distributed Computing in Sensor Systems*. IEEE. 2015, p. 11-20.
- [39] Julian D Marshall, Elizabeth Nethery et Michael Brauer. "Within-urban variability in ambient air pollution : comparison of estimation methods". In : *Atmospheric Environment* 42.6 (2008), p. 1359-1369.
- [40] Xia Meng et al. "A land use regression model for estimating the NO₂ concentration in Shanghai, China". In : *Environmental research* 137 (2015), p. 308-315.
- [41] Kyle P Messier et al. "Mapping air pollution with Google Street View cars : Efficient approaches with mobile monitoring and land use regression". In : *Environmental science & technology* 52.21 (2018), p. 12563-12572.
- [42] Kristin A Miller et al. "Long-term exposure to air pollution and incidence of cardiovascular events in women". In : *New England Journal of Medicine* 356.5 (2007), p. 447-458.
- [43] Xuening Qin et al. "Fine-Grained Urban Air Quality Mapping from Sparse Mobile Air Pollution Measurements and Dense Traffic Density". In : *Remote Sensing* 14.11 (2022), p. 2613.
- [44] Xuening Qin et al. "Street-level Air Quality Inference Based on Geographically Context-aware Random Forest Using Opportunistic Mobile Sensor Network". In : *2021 the 5th International Conference on Innovation in Artificial Intelligence*. 2021, p. 221-227.
- [45] Zev Ross et al. "A land use regression for predicting fine particulate matter concentrations in the New York City region". In : *Atmospheric Environment* 41.11 (2007), p. 2255-2269.

- [46] Zev Ross et al. "Nitrogen dioxide prediction in Southern California using land use regression modeling : potential for environmental health analyses". In : *Journal of exposure science & environmental epidemiology* 16.2 (2006), p. 106-114.
- [47] Patrick H Ryan et Grace K LeMasters. "A review of land-use regression models for characterizing intraurban air pollution exposure". In : *Inhalation toxicology* 19.sup1 (2007), p. 127-133.
- [48] Philipp Schneider et al. "Mapping urban air quality in near real-time using observations from low-cost sensors and model information". In : *Environment international* 106 (2017), p. 234-247.
- [49] Joel Schwartz. "Air pollution and hospital admissions for the elderly in Birmingham, Alabama". In : *American journal of epidemiology* 139.6 (1994), p. 589-598.
- [50] Joel Schwartz. "Lung function and chronic exposure to air pollution : a cross-sectional analysis of NHANES II". In : *Environmental research* 50.2 (1989), p. 309-321.
- [51] Kunwar P Singh, Shikha Gupta et Premanjali Rai. "Identifying pollution sources and predicting urban air quality using ensemble learning methods". In : *Atmospheric Environment* 80 (2013), p. 426-437.
- [52] R Sivacoumar et K Thanasekaran. "Comparison and performance evaluation of models used for vehicular pollution prediction". In : *Journal of Environmental Engineering* 127.6 (2001), p. 524-530.
- [53] Ji-Young Son, Michelle L Bell et Jong-Tae Lee. "Individual exposure to air pollution and lung function in Korea : spatial analysis using multiple exposure approaches". In : *Environmental research* 110.8 (2010), p. 739-749.
- [54] Jun Song, Ke Han et Marc Stettler. "Deep-MAPS : Machine Learning based Mobile Air Pollution Sensing". In : *IEEE Internet of Things Journal* (2020).
- [55] Annelies Van den Hove et al. "Development of a land use regression model for black carbon using mobile monitoring data and its application to pollution-avoiding routing". In : *Environmental research* 183 (2020), p. 108619.

- [56] Sotiris Vardoulakis et al. "Modelling air quality in street canyons : a review". In : *Atmospheric environment* 37.2 (2003), p. 155-182.
- [57] Kristina W Whitworth et al. "Kriged and modeled ambient air levels of benzene in an urban environment : an exposure assessment study". In : *Environmental Health* 10.1 (2011), p. 1-10.
- [58] Darren Wilton et al. "Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA". In : *Science of the total environment* 408.5 (2010), p. 1120-1130.
- [59] Kathrin Wolf et al. "Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany". In : *Science of the Total Environment* 579 (2017), p. 1531-1540.
- [60] David W Wong, Lester Yuan et Susan A Perlin. "Comparison of spatial interpolation methods for the estimation of air quality data". In : *Journal of Exposure Science & Environmental Epidemiology* 14.5 (2004), p. 404-415.
- [61] Xia Xi et al. "A comprehensive evaluation of air pollution prediction improvement by a machine learning method". In : *2015 IEEE international conference on service operations and logistics, and informatics (SOLI)*. IEEE. 2015, p. 176-181.
- [62] Liang Zhai et al. "Land use regression modeling of PM_{2.5} concentrations at optimized spatial scales". In : *Atmosphere* 8.1 (2016), p. 1.
- [63] Dan Zhang et Simon S Woo. "Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network". In : *IEEE Access* 8 (2020), p. 89584-89594.
- [64] Bu Zhao et al. "Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning". In : *Environmental Science & Technology* 55.8 (2021), p. 5579-5588.
- [65] Leonard M Zwack et al. "Modeling spatial patterns of traffic-related air pollutants in complex urban terrain". In : *Environmental Health Perspectives* 119.6 (2011), p. 852-859.

Chapitre 3

Des fondamentaux aux techniques avancées : une comparaison des modèles de pollution

Points clés du chapitre

- Preuve de la supériorité de la géostatistique spatio-temporelle comparé aux techniques d'interpolation déterministes.
- Comparaison de 10 modèles issus de trois familles de méthodes distinctes en utilisant des variables auxiliaires.
- Mise en évidence de différences significatives entre les données issues des capteurs à bas coût et celles issues des stations de références.
- Choix du modèle géostatistique en utilisant le maximum de vraisemblance pour la suite des travaux de recherche.

Contenu du chapitre

3.1 Méthodes utilisant uniquement des données de pollution atmosphérique	76
3.1.1 Techniques déterministes	79
3.1.2 Méthodes stochastiques	88
3.2 Intégration et comparaison de modèles exploitant des données externes	109
3.2.1 Méthodologie	109
3.2.2 Résultats	122
3.2.3 Discussion	131
3.3 Conclusion	134
Références	136

Le présent chapitre traite de l'évaluation et de l'implémentation de modèles statistiques pour estimer et prédire la pollution atmosphérique. Nous suivons une démarche méthodologique progressive, partant de modèles simples pour graduellement accroître leur complexité et mieux répondre à notre problématique. Dans ce qui suit nous présenterons des modèles de plus en plus élaborés. Chaque modèle est décrit en détail : ses bases mathématiques, son contexte d'application et son adaptation à notre problématique. Puis, il est mis en pratique sur un jeu de données, évalué et confronté à d'autres techniques.

La première partie de ce chapitre se focalise sur l'exploration de méthodes qui exploitent uniquement des données de pollution atmosphérique. Cela comprend des données issues tant de capteurs mobiles à faible coût que de stations fixes de qualité de l'air, qui servent de référence. Le choix de cette approche se justifie par sa simplicité et sa facilité de mise en œuvre et offre une première

vue de la structure et de la dynamique de la pollution atmosphérique.

Dans la seconde partie, l'analyse sera étendue par la comparaison de modèles issus de différentes familles, qui intègrent des variables supplémentaires. L'enrichissement du modèle avec ces données additionnelles devrait mener à des résultats plus précis et favoriser une compréhension approfondie du phénomène de la pollution atmosphérique.

Cette démarche nous permettra de conclure ce chapitre en ayant une vision claire des avantages et des limites de chaque méthode, et de choisir la plus adaptée à notre problématique.

3.1 Méthodes utilisant uniquement des données de pollution atmosphérique

Ce sous-chapitre se penche sur la comparaison entre les méthodes d'interpolation déterministes et stochastiques provenant de la géostatistique. L'objectif est de déduire des valeurs non mesurées à partir des données de pollution atmosphérique recueillies par les capteurs à bas coût et les stations fixes de référence. Pour l'instant, l'attention se limite aux données de pollution, sans prendre en compte d'autres informations supplémentaires. Chaque méthode sera décrite, ajustée au contexte spécifique, appliquée sur le jeu de données OpenSense présenté auparavant, puis évaluée et comparée aux autres afin de déterminer la plus efficace.

Cette tâche de prédiction de la qualité de l'air peut être considérée comme un problème d'interpolation d'une fonction dans un espace tridimensionnel, plus précisément un espace spatio-temporel. Cette perspective est illustrée dans la figure 3.1, qui met en scène le jeu de données OpenSense.

Dans cet espace tridimensionnel, chaque coordonnée (x, y, t) est associée à une mesure précise, indicative de la qualité atmosphérique. Ces mesures, dis-

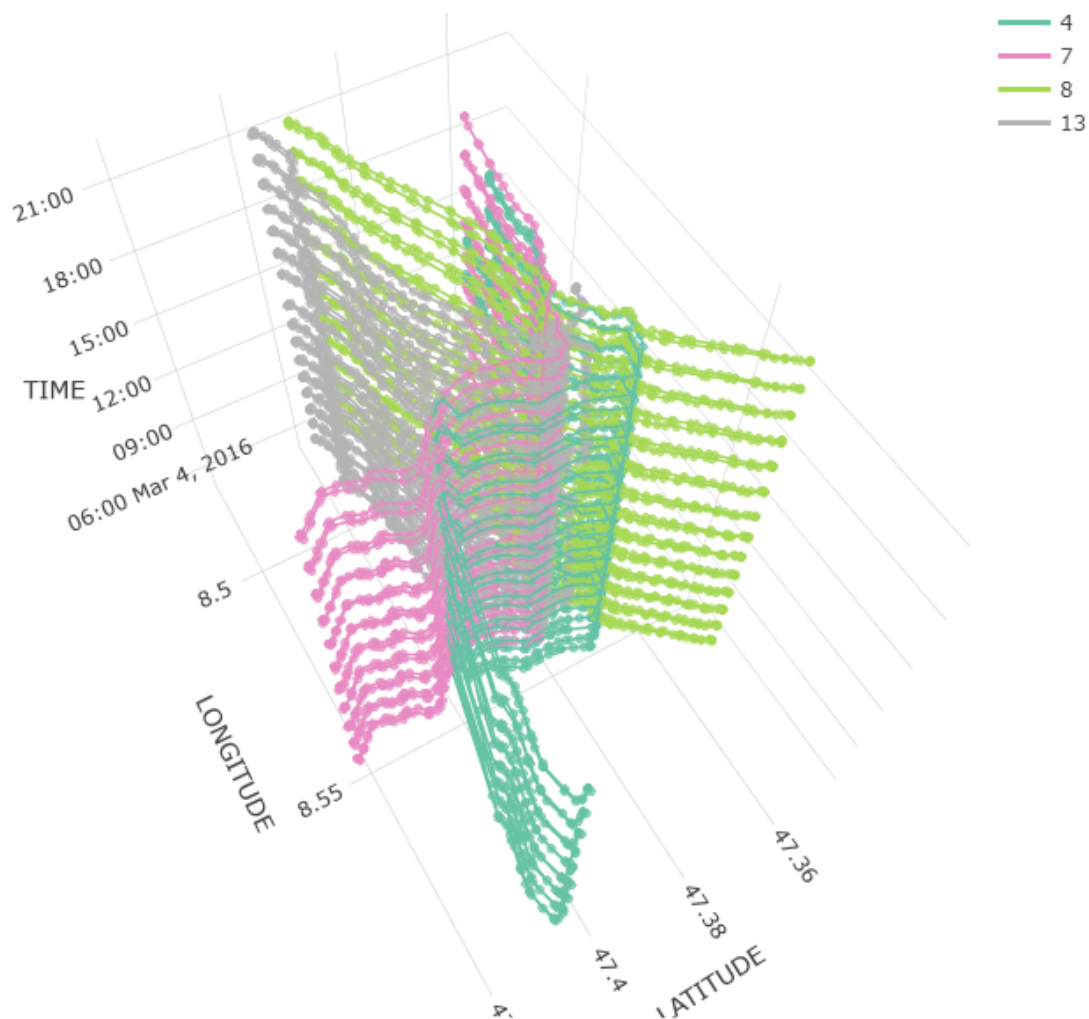


Figure 3.1 – Représentation des données d'OpenSense dans un espace spatio-temporel.

tribuées de manière ordonnée, sont le reflet des trajets empruntés par les lignes de tramway pourvues de capteurs atmosphériques. L'impératif scientifique consiste à prédire la qualité de l'air en des lieux et instants non directement observés et en se basant sur l'information disponible des points environnants. Ainsi, l'approche adoptée se situe dans le cadre d'une problématique d'interpolation spatio-temporelle, visant à élaborer une fonction qui, en restant fidèle aux mesures existantes, facilite l'extrapolation de la qualité de l'air en différents points spatio-temporels.

Les méthodes présentées s'appuient sur la première loi de la géographie : "Eve-

rything is related to everything else, but near things are more related than distant things” (Tobler, 1970). Cela signifie que le modèle proposé doit avoir la propriété : une observation est davantage corrélée avec une observation proche de l’espace-temps qu’avec une autre observation qui a été collectée plus loin.

Dans le cadre de l’analyse comparative des modèles menée, les données du 4 mars ont été choisies pour deux raisons principales. Premièrement, cette date présente le volume d’observations le plus élevé parmi l’ensemble des dates disponibles dans la base de données d’OpenSense. Deuxièmement, elle caractérise de manière représentative le profil diurne des fluctuations de l’ozone, avec un pic notable aux environs de 14 heures.

Pour la validation croisée, trois scénarios distincts de sélection aléatoire des échantillons ont été définis. Ces scénarios, relatifs à la segmentation des données en ensembles d’apprentissage et de validation, visent à déterminer les conditions d’adaptation idéales pour chaque modèle :

- Le premier scénario (SC1) sélectionne une proportion de points au hasard, indépendamment de leur position spatiale ou du moment de leur collecte. Cela simule une situation où nous cherchons à reconstruire les données entre les emplacements échantillonnés.
- Le second scénario (SC2) vise à refléter une situation plus réaliste. Des parcours de longueurs variées sont sélectionnés, tout en conservant le même pourcentage de données. Cette approche cherche à reproduire la collecte de données réelles à partir d’un capteur mobile, c’est-à-dire à extrapoler les données vers des lieux proches des points d’échantillonnage
- Le dernier scénario (SC3) n’utilise que les données résultant de la trajectoire de tramways spécifiques. Cela correspond à une extrapolation pour des endroits « éloignés » des points d’échantillonnage, ce qui sera souvent le cas dans la pratique.

Deux familles de méthodes seront comparées avec ces données et ces scénarios : les techniques déterministes et les méthodes géostatistiques.

3.1.1 Techniques déterministes

Les techniques d'interpolation déterministes sont utilisées pour estimer des valeurs d'une fonction en des points non connus à partir de valeurs observées. Ces méthodes ne tiennent pas compte d'une incertitude ou d'une variabilité aléatoire, mais utilisent des formules ou des procédures déterminées pour obtenir une estimation. Bien qu'elles présentent certaines lacunes, notamment l'inaptitude à fournir une évaluation de l'incertitude, leur attrait réside dans leur facilité d'implémentation et leur transparence interprétative. Dans cette étude, l'attention sera portée sur l'application et l'analyse de trois méthodes déterministes prédominantes : le diagramme de Voronoï, les splines à plaques minces (TPS, de l'anglais "Thin Plate Splines") et la pondération inverse à la distance. L'exploitation optimale de ces techniques nécessitera la définition d'une métrique de distance adéquate.

Diagramme de Voronoï

Le diagramme de Voronoï, également connu sous le nom de tessellation de Voronoï, est une méthode d'interpolation spatiale qui subdivise l'espace en régions ou cellules à partir d'un ensemble de points [3]. Chaque cellule correspond à une région de l'espace qui est plus proche de son point de données associé que de tout autre point de données.

Dans le contexte de la prédiction de la qualité de l'air, les données de chaque capteur peuvent être utilisées pour créer une cellule de Voronoï, et l'estimation de la qualité de l'air en un point quelconque est celle du capteur le plus proche. En cas d'équidistance entre plusieurs points (frontière des cellules), c'est la moyenne des observations des capteurs concernés qui est prise en compte.

Si le diagramme de Voronoï est couramment utilisée dans un espace bidimensionnel, comme illustré dans la figure 3.2, sa généralisation à un espace tridimensionnel est assez directe. Cependant, l'extension à un espace spatio-temporel comporte des défis plus subtils.

Pour mieux comprendre cette difficulté, considérons l'exemple suivant : nous souhaitons prédire la concentration d'un polluant atmosphérique à un rond-

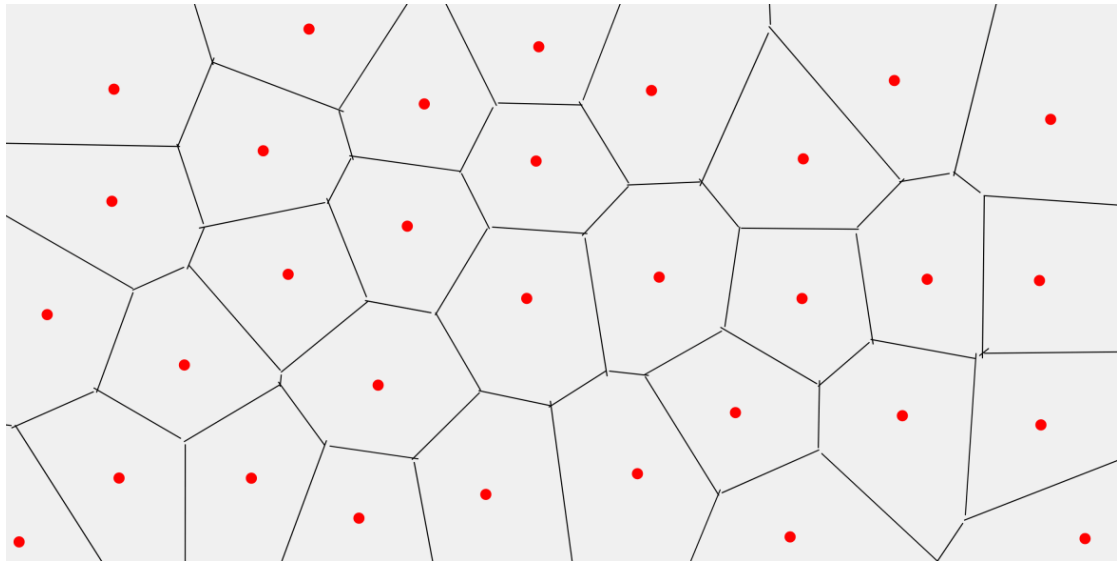


Figure 3.2 – Représentation d'un espace bidimensionnel divisé en cellules par la méthode de Voronoï

point spécifique d'une ville à 12h00 précises. Les capteurs de pollution sont mobiles et sillonnent la ville. Le diagramme de Voronoï suggérerait de choisir l'observation la plus proche. Supposons que nous ayons deux observations à notre disposition : la première est une mesure prise au rond-point en question à 12h10 et la seconde est une mesure effectuée dans une rue adjacente, à une distance de 100 mètres, à 12h00. Laquelle de ces deux observations est la plus « proche » ? Devons-nous privilégier le décalage de 10 minutes ou la distance de 100 mètres ?

Pour résoudre ce dilemme, une distance spatio-temporelle a été définie et sera utilisée tout au long de ce sous-chapitre. :

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2 + C^2 \cdot (t_i - t)^2} \quad (3.1)$$

Le paramètre C , dont l'unité est exprimée en termes distance par un temps, est déterminé par une méthode de validation croisée. Pour ce faire, l'ensemble de données est divisé en deux groupes : un groupe d'apprentissage et un groupe de validation. Par la suite, le groupe d'apprentissage est utilisé pour prédire les mesures de qualité de l'air aux emplacements spatio-temporels correspondant au groupe de validation. En modifiant le paramètre C , son effet sur la précision

de la prédiction est observé. La valeur de C qui réduit au minimum la Racine de l'erreur quadratique moyenne (RMSE) est alors choisie. Dans cette étude menée avec les données d'OpenSense, il a été constaté que les meilleurs résultats étaient obtenus avec un C égal à 1.07 m/s .

Une valeur de ce paramètre qui est supérieure à 1 comme dans cette étude, suggère que le processus de dispersion varie plus dans le temps (en seconde) que dans l'espace (en mètre).

L'estimation de ce paramètre répond à la question soulevée précédemment dans notre exemple : il serait donc plus judicieux de considérer l'observation prise à 100 mètres mais au même moment comme étant « plus proche » et par conséquent, plus susceptible de fournir une prédiction fiable de la concentration de polluants au rond-point à 12h00. C'est une conclusion qui se verra confirmée dans la seconde partie de ce sous-chapitre, où nous observerons que la variation de la concentration de polluants est plus marquée dans le temps (à l'échelle d'une journée) que dans l'espace (à l'échelle d'une ville).

La Figure 3.3 illustre l'application de cette méthode aux données de pollution récoltées par une unique ligne de tram (la ligne 4). Le résultat est une prédiction en « escaliers » qui suit assez fidèlement la tendance globale, sous réserve qu'il y ait suffisamment de points d'échantillonnage.

Bien que les tramways suivent des trajectoires prédéfinies et reviennent régulièrement sur leurs pas comme illustré dans la figure 3.1, l'algorithme tend systématiquement à prédire la valeur à un endroit donné en se basant sur l'observation la plus récente. Cela est vrai même si le capteur est déjà passé exactement à cet endroit quelques heures auparavant. L'algorithme ne choisira pas cette observation plus ancienne car elle est « lointaine » en termes de distance spatio-temporelle. Cette particularité est attribuable au fait que la variation de la concentration de polluants est plus grande dans le temps qu'elle ne l'est dans l'espace, comme nous l'avons évoqué plus tôt. De plus, la fréquence d'échantillonnage dans cette étude est supérieure au temps nécessaire pour que le tram effectue un aller-retour sur sa ligne. Par conséquent, l'observation la plus

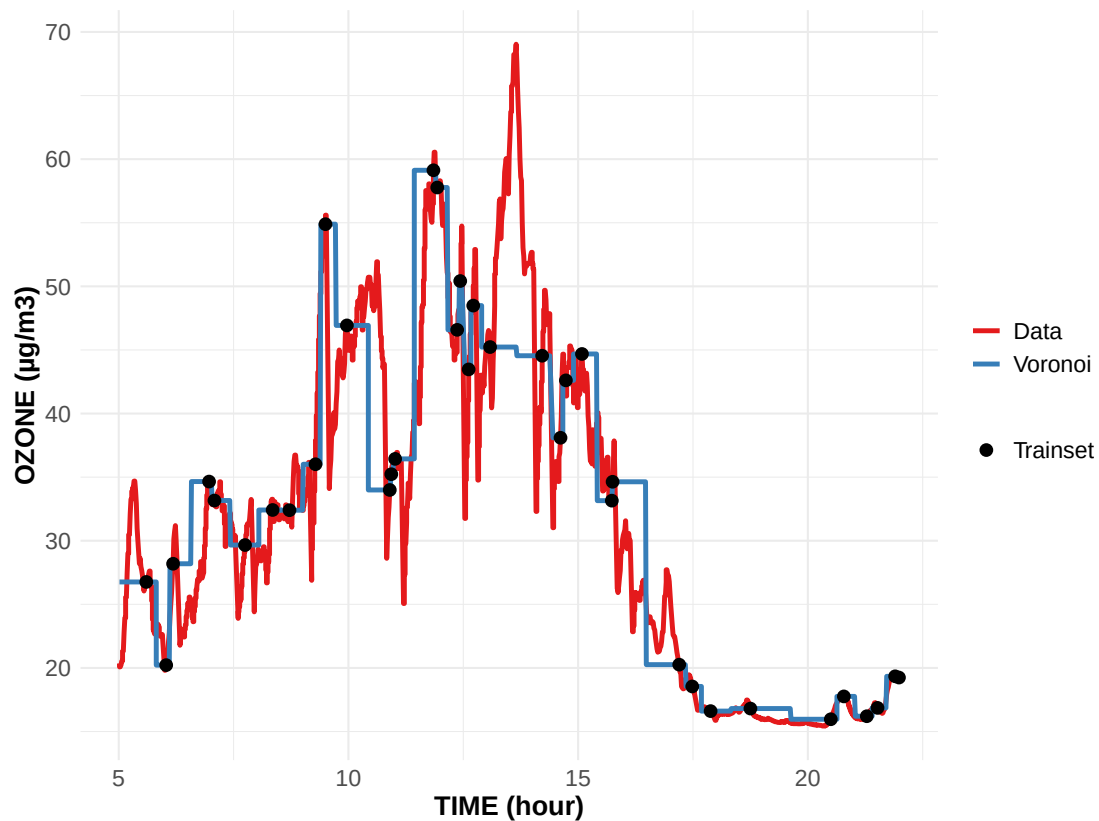


Figure 3.3 – Application de la méthode de diagrammes de Voronoï à une série de données issue d'une ligne de tramway

récente est généralement plus représentative de la situation actuelle qu'une observation plus ancienne prise exactement au même endroit.

La simplicité de mise en œuvre du diagramme de Voronoi et le fait qu'il n'exige pas d'hypothèse sur la distribution des données en font une méthode attrayante. Cependant, elle montre ses limites lorsque l'échantillonnage est peu dense. En effet, en l'absence d'une densité d'échantillonnage adéquate, cette méthode peut avoir du mal à produire des résultats satisfaisants. Un autre inconvénient notable de cette méthode est l'absence d'une estimation de l'incertitude associée aux prédictions, ce qui limite son utilité pour évaluer la fiabilité des résultats obtenus.

Splines à plaques minces

Les splines à plaques minces représentent une technique d'interpolation basée sur les splines. Cette méthode vise à minimiser la courbure de la surface interpolée, aboutissant ainsi à une surface lisse qui s'approche de tous les points de données. Cette approche est caractérisée par la minimisation d'une fonctionnelle énergétique, minimisant la distance de la courbe générée aux points de données plus l'intégrale du carré de la courbure de la fonction interpolatrice qui caractérise la courbure de la fonction.

Dans le cadre de la modélisation de la qualité de l'air, les TPS offrent un moyen efficace de générer une surface continue qui décrit les niveaux de pollution atmosphérique à travers l'espace et le temps.

Formellement, la méthode TPS ajuste une fonction spatio-temporelle $f(x, y, t)$ qui approche les observations Z des capteurs. Cette fonction est obtenue en minimisant la fonction énergétique suivante :

$$E_{\text{tps}}(f) = \sum_{i=1}^n \|Z_i - f(x_i, y_i, t_i)\|^2 + \phi D \quad (3.2)$$

Avec

$$D = \int \int \int \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial t} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial y \partial t} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 + \left(\frac{\partial^2 f}{\partial t^2} \right)^2 \right] dx dy dt \quad (3.3)$$

Le paramètre de régularisation ϕ est introduit pour contrôler la rigidité de la surface interpolée.

La Figure 3.4 montre l'application de la méthode de TPS à la série de données vu précédemment. Il est à noter que les prédictions en début de matinée sont influencées par les observations réalisées entre 10h et 15h, correspondant à l'aller-retour du tramway.

Bien que les TPS soient capables de modéliser des relations spatio-temporelles complexes, offrant ainsi une surface continue et lisse, il convient de mentionner

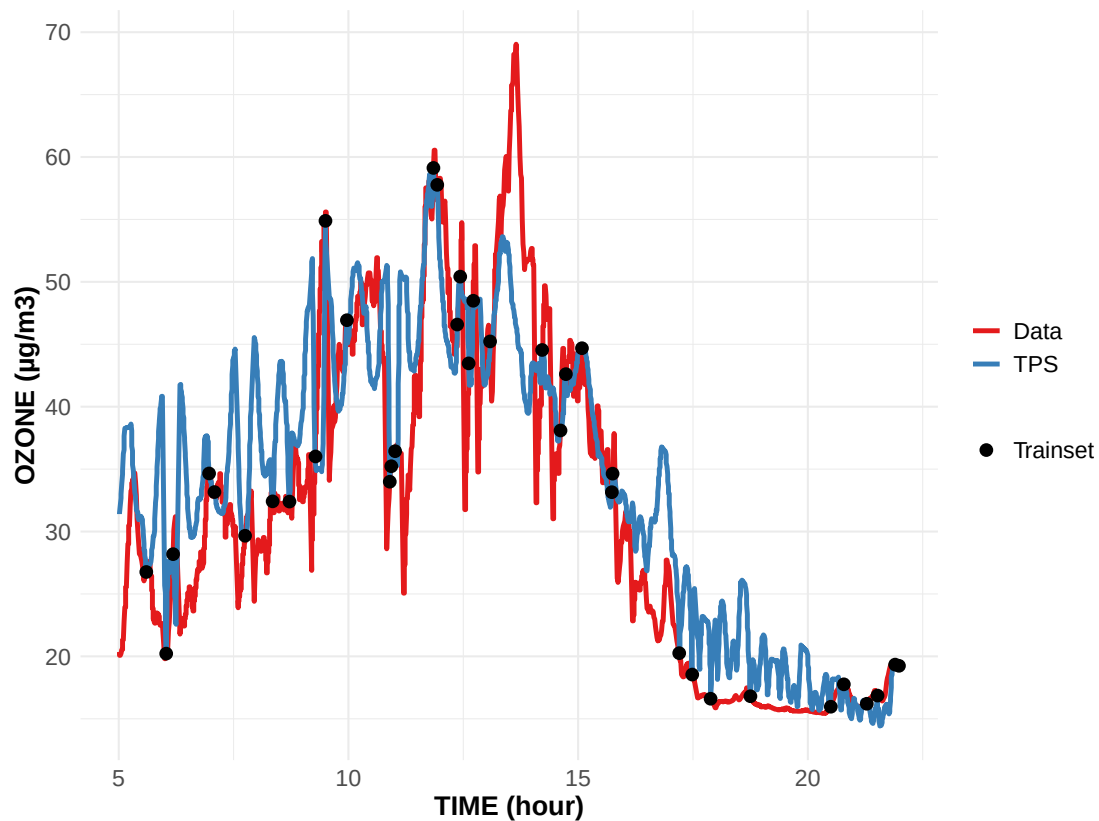


Figure 3.4 – Application de la méthode de TPS à une série de données issue d'une ligne de Tramway

que cette méthode peut être sensible aux valeurs aberrantes et aux effets de bord.

Pondération inverse à la distance

La pondération inverse à la distance (IDW, de l'anglais "Inverse Distance Weighting") est une méthode déterministe d'interpolation. Elle est conçue pour estimer des valeurs en des points non échantillonnés en se basant sur une combinaison linéaire des valeurs des points échantillonnés, ces dernières étant pondérées par l'inverse de leur distance respective [20].

La formulation mathématique de l'IDW est exprimée comme suit :

$$Z^*(x, t) = \sum_{i=1}^n \lambda_i Z(x_i, t_i) \quad (3.4)$$

où les poids λ_i sont définis par :

$$\lambda_i = \frac{1/d_i^p}{\sum_{i=1}^n 1/d_i^p} \quad (3.5)$$

Ici d_i représente la distance entre $Z^*(x, t)$ et $Z(x_i, t_i)$. $Z^*(x, t)$ étant le point où l'on veut prédire, $Z(x_i, t_i)$ les données à notre disposition. Cette distance est calculée selon l'équation 3.1. Les poids diminuent avec l'augmentation de la distance, et cette décroissance est accentuée lorsque la puissance p augmente. Dans cette étude, la puissance p a été fixée à 2, ce qui signifie que les points proches ont une influence prépondérante sur l'estimation, traduisant ainsi une interpolation spatio-temporelle de nature locale.

La Figure 3.5 dépeint l'efficacité de l'IDW lors de son application sur les données collectées par la ligne de tramway. Un avantage notable de l'IDW par rapport à d'autres méthodes déterministes est sa robustesse face aux valeurs aberrantes. Les capteurs à bas coût, utilisés dans cette étude, peuvent produire des mesures erronées, dues à des erreurs instrumentales ou à des conditions environnementales extrêmes. Toutefois, l'IDW minimise l'impact de ces anomalies en considérant l'ensemble des observations, et non pas seulement les plus proches. Cette caractéristique renforce la fiabilité des prédictions obtenues par cette méthode.

Comparaison et conclusions relatives aux méthodes déterministes

La Figure 3.6 illustre l'application des trois méthodes déterministes - le diagramme de Voronoi, les TPS et IDW - sur un ensemble de données recueillies durant une journée entière, dans le contexte du scénario d'extrapolation (scénario SC3). Il est à noter que les résultats issus des deux premiers scénarios ne montrent pas de variations significatives, rendant leur discussion moins perti-

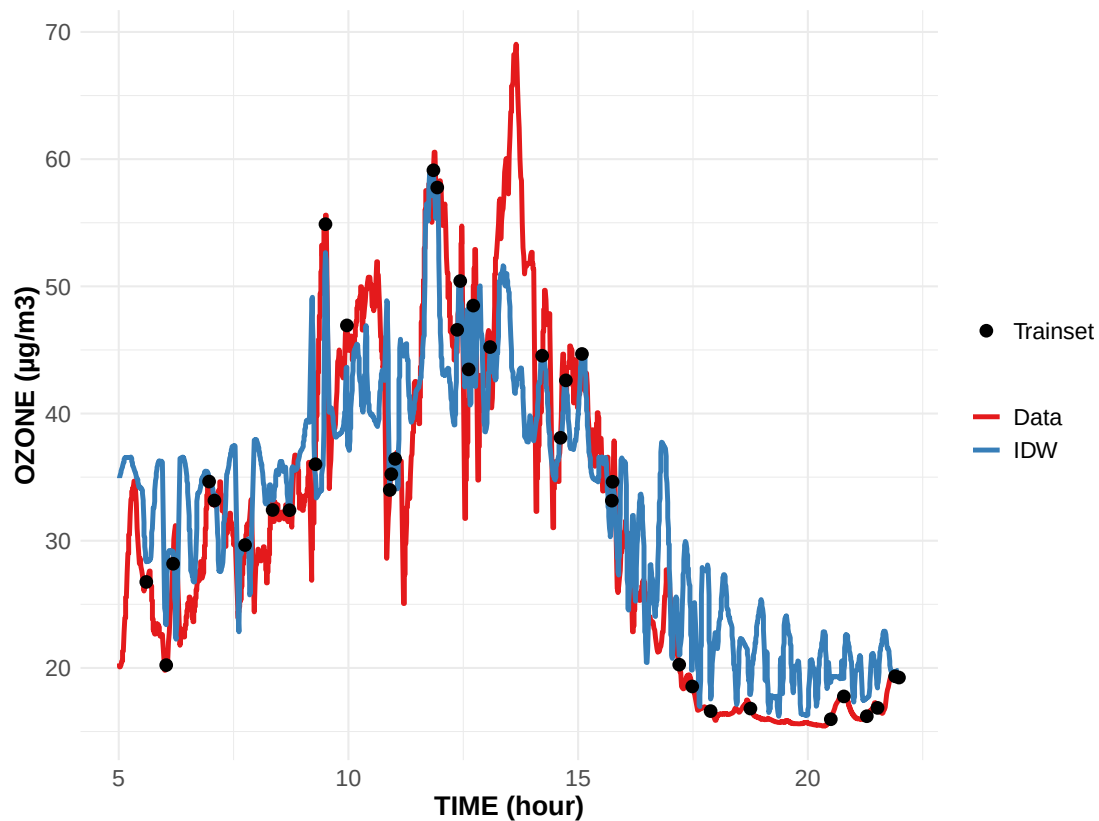


Figure 3.5 – Illustration de l'application de la méthode IDW sur des données provenant d'une ligne de tramway.

nente pour notre analyse.

La performance en termes de calcul constitue un critère essentiel dans la sélection d'une méthode d'interpolation pour prédire la qualité de l'air à partir de données issues de capteurs mobiles. De ce point de vue, le diagramme de Voronoï et l'IDW se distinguent, affichant une complexité de requête de $O(\log(n))$ une fois la structure de données nécessaire en place. En revanche, les TPS requièrent une inversion de matrice qui induit une complexité de $O(n^3)$, ce qui peut se révéler problématique pour les grands ensembles de données.

Sur le plan des performances prédictives, les trois méthodes présentent des résultats comparables. Toutefois, l'IDW se distingue légèrement avec un RMSE

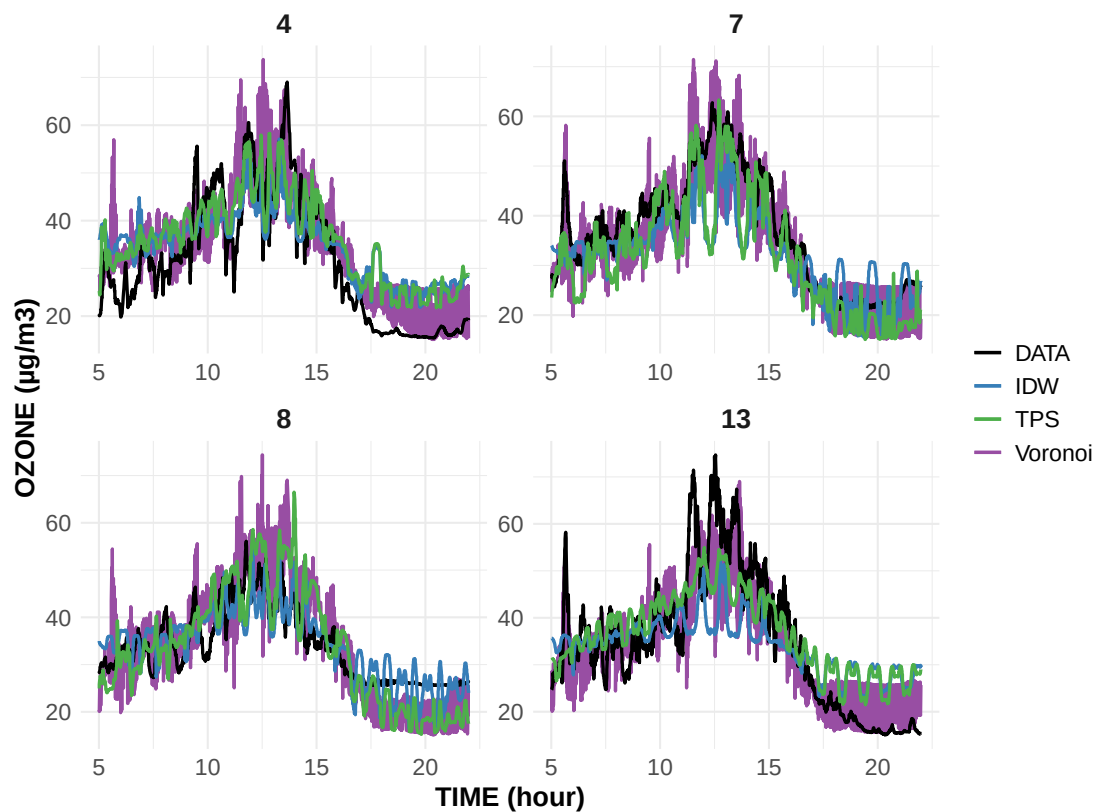


Figure 3.6 – Illustration de l'application des trois méthodes déterministes.

de 7.67, contre 7.82 pour le TPS et 8.05 pour le diagramme de Voronoï. Ces chiffres suggèrent qu'il n'est pas indispensable d'augmenter la complexité au profit de la précision prédictive dans notre application.

Une perspective intéressante serait d'intégrer les données temporelles des stations fixes de référence. Pour l'IDW, cela pourrait se traduire par une augmentation du poids des observations issues de ces stations, généralement perçues comme plus fiables. Cette modification réduirait artificiellement la distance entre ces observations et le point de prédiction, renforçant ainsi leur influence. Cependant, les méthodes déterministes ne semblent pas être les plus adaptées à cette démarche.

Le manque de flexibilité des techniques déterministes pour intégrer d'autres

types de données, leur capacité limitée à modéliser les processus non stationnaires, et surtout leur incapacité à fournir des estimations d'incertitude et leur manque de pouvoir descriptif, constituent des lacunes majeures. De plus, leur efficacité dépend beaucoup de la densité de l'échantillonnage. En effet, lorsqu'un échantillonnage est dense, ces méthodes offrent une bonne prédiction, mais elles deviennent moins fiables lorsque les observations sont distantes ou rares. De plus, elles supposent que les observations sont précises et ne prennent pas en compte la nature des capteurs bas coût. Ces observations suggèrent que des méthodes alternatives pourraient être plus appropriées pour cette étude. Néanmoins, l'IDW servira de technique de référence pour les comparaisons avec d'autres modèles qui seront examinés plus tard dans ce manuscrit.

3.1.2 Méthodes stochastiques

Après l'analyse des méthodes déterministes, l'étude se tourne vers les méthodes stochastiques, en se concentrant particulièrement sur celles dérivées de la géostatistique. Ces techniques traitent les données comme des variables aléatoires, permettant ainsi de modéliser la pollution atmosphérique comme une fonction aléatoire indexée spatialement et temporellement.

Les méthodes stochastiques de la géostatistique, en modélisant les observations comme des variables aléatoires, offrent une prise en compte de l'incertitude des mesures. Elles permettent également d'intégrer diverses sources de données de pollution atmosphérique, qu'il s'agisse de capteurs à bas coût, de stations fixes ou d'autres sources.

Le krigeage¹ se rapporte spécifiquement à la phase de prédiction au sein de la modélisation. Il procède à l'évaluation de la valeur en un site non échantillonné en adoptant une combinaison linéaire des mesures, où les coefficients sont déterminés en fonction de la dispersion des données, cela au travers notamment

1. Ce nom reconnaît l'influence de D. G. Krige, qui a été le précurseur de l'utilisation des méthodes statistiques dans l'industrie minière sud-africaine [12]

d'un variogramme. Les modèles tels que le Krigeage avec Dérives Externes (KED) et le Cokrigeage offrent la possibilité d'incorporer des covariables exogènes, tandis que le Krigeage Simple (SK) s'exécute en l'absence de ces dernières.

Le but est de déterminer les modèles de géostatistique spatio-temporels les plus adaptés pour prédire la pollution atmosphérique urbaine à partir de capteurs mobiles et d'évaluer leurs avantages par rapport aux méthodes déterministes.

La section suivante détaille la méthodologie adoptée, fournissant une vue d'ensemble de l'approche comparative entre les différentes méthodes.

Méthodologie

La méthodologie adoptée pour évaluer la pertinence des techniques géostatistiques spatio-temporelles est présentée ci-après. Elle met en lumière la comparaison entre différents modèles géostatistiques et l'apport potentiel par rapport à la méthode IDW standard.

1. Sélection des modèles :

Trois des plus populaires approches géostatistiques sont explorées. Deux d'entre elles ont intégré les données des stations fixes pour la prédiction de la qualité de l'air. Ces méthodes, basées sur des postulats distincts, sont :

- Krigeage simple avec une moyenne connue variable : la série temporelle de la station fixe est utilisée comme moyenne générale.
- Krigeage ordinaire avec une moyenne constante par segments, mais inconnue.
- Krigeage avec dérive externe : les données de la station fixe servent à estimer la moyenne sous-jacente.

2. Étude variographique : Préalablement à l'application du krigeage qui constitue la phase de prédiction, une estimation de la variabilité inhérente aux données s'avère cruciale. L'étude s'est concentrée sur l'estimation du variogramme,

évitant des hypothèses strictes sur la stationnarité du champ aléatoire. Suivant les recommandations d'Arnaud et al. [2], le variogramme expérimental a été limité à la moitié de la distance maximale entre deux points. Sachant que la distance maximale entre deux points dans cette étude était de 12,8 km, les variogrammes ont donc été limités à 6 km. La limite temporelle a été définie manuellement, en l'augmentant progressivement jusqu'à l'identification d'un seuil dans le variogramme. Une semaine de données a été exploitée pour estimer les variogrammes empiriques. Pour étudier une éventuelle anisotropie des données liée à des facteurs externes (comme le vent), deux variogrammes empiriques spatio-temporels dans les deux directions statiques (nord-sud et est-ouest) ont été calculés.

3. Procédure d'évaluation des modèles :

Afin d'évaluer les différents modèles, une procédure de validation croisée a été effectuée et les moyennes des indicateurs de performance sélectionnés ont été calculées. En variant la taille de l'ensemble de données d'entraînement, des conclusions sur l'efficacité des modèles dans différentes conditions sont présentées. Seules les données du 04/03, jour où nous disposons du plus grand volume de données, ont été utilisées dans cette procédure de validation croisée à des fins de prédiction/interpolation suivant les trois scénarios décrits en début de chapitre.

Les trois méthodologies ont été comparées avec une méthode d'interpolation déterministe, choisie dans ce contexte comme étalon (IDW) et ce pour l'ensemble des trois situations envisagées. Pour juger de la pertinence des résultats issus de chaque technique, trois critères de performance, décrits ci-dessous, ont été utilisés pour l'évaluation. :

- La racine de l'écart quadratique moyen (RMSE) a été choisie comme principal indicateur de performance pour mesurer l'erreur car c'est la mesure la plus fréquemment utilisée pour évaluer les différences entre les valeurs prédites par un modèle ou un estimateur et les valeurs observées. Les trois modèles géostatistiques présentés dans cet article ont été construits pour

minimiser cette erreur.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (Z_i^* - Z_i)^2}{n}} \quad (3.6)$$

- L'indicateur de performance de biais a été choisi pour contrôler l'absence de biais des estimateurs. Les trois estimateurs géostatistiques sont théoriquement sans biais. Cet indicateur de performance est utilisé pour vérifier cela.

$$\text{BIAIS} = 1/n \sum_{i=1}^n (Z_i^* - Z_i) \quad (3.7)$$

- L'indicateur de performance de corrélation a été sélectionné pour traiter le caractère économique des capteurs. En cas de biais, il est nécessaire de mesurer la performance de corrélation et de la comparer à la RMSE.

$$\text{CORR} = \frac{\sum_{i=1}^n (Z_i^* - \bar{Z}^*)(Z_i - \bar{Z})}{\sqrt{\sum_{i=1}^n (Z_i^* - \bar{Z}^*)^2 \sum_{i=1}^n (Z_i - \bar{Z})^2}} \quad (3.8)$$

Théorie des modèles géostatistiques

Il existe deux façons d'intégrer le temps dans la géostatistique spatiale. La première est sous la forme de cokrigage, et la seconde, plus naturelle, consiste à considérer le temps comme une dimension séparée, ce qui sera le cas dans cette étude. Ce qui a été considéré ici comme réalisation, est un échantillon unique mesuré dans un volume d'air.

Étant donné un support D dans $\mathbb{R}^n$ et un espace probabiliste $(\Omega, \mathcal{A}, \mathbb{P})$, une fonction aléatoire est une fonction de deux variables $Z(x, w)$ telle que, pour chaque x dans D , la section $Z(x, \cdot)$ est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$.

Dans ce cas, $D = \mathbb{R}^2 \times \mathbb{R}_+$ où $\mathbb{R}^2$ représente l'espace et $\mathbb{R}_+$ le temps, la fonction aléatoire est simplement notée $Z(x, t)$, et une réalisation de cette fonction aléatoire est représentée par $z(x, t), Z_i(x, t)$ ou $Z(x_i, t_i)$, avec $x \in \mathbb{R}^2$ et $t \in \mathbb{R}_+$.

Les méthodes présentées dans cette section ont été théoriquement définies dans des travaux antérieurs [5]. Dans la section suivante, nous introduisons les détails théoriques nécessaires pour comprendre les modèles.

- Krigeage simple avec une moyenne variable :

L'application du krigeage simple nécessite deux hypothèses : la stationnarité d'ordre deux du champ aléatoire, et la connaissance de la moyenne sur l'ensemble du domaine D . Supposer que les données collectées par les capteurs mobiles proviennent d'un champ aléatoire stationnaire d'ordre deux est une hypothèse forte qui n'est pas forcément réaliste. Dans ce modèle, les données fournies par la station de surveillance fixe $F(t)$ sont supposées être la moyenne globale. Soustraire la valeur de la station fixe des données fournies par les capteurs mobiles est supposé être un champ stationnaire d'ordre deux avec une moyenne nulle.

L'estimateur de krigeage simple est :

$$Z^*(x, t) = \mu + \sum_{i=1}^n \lambda_i (Z(x_i, t_i) - \mu) = \sum_{i=1}^n \lambda_i Z(x_i, t_i) \quad (3.9)$$

où μ est la moyenne du champ aléatoire et est égal à zéro. Pour produire le meilleur estimateur linéaire, nous devons nous assurer que la variance de l'estimation est minimale et que l'estimateur est sans biais.

La condition sans biais est automatiquement vérifiée, et n'implique aucune contrainte supplémentaire car :

$$\mathbb{E}[Z^*(x, t) - Z(x, t)] = \sum_{i=1}^n \lambda_i \mathbb{E}Z(x_i, t_i) = 0 \quad (3.10)$$

Cela conduit aux équations de krigeage simple :

$$\sum_{j=1}^n \lambda_j \gamma(x_i - x_j, t_i - t_j) = \gamma(x_i - x, t_i - t) \quad i = 1, \dots, n \quad (3.11)$$

La résolution du système d'équations (3.11) donne les différents lambda dans la combinaison linéaire (3.9).

- Krigeage ordinaire :

L'application du krigeage ordinaire fait des hypothèses moins restrictives, à savoir une moyenne constante mais inconnue. L'estimateur linéaire du krigeage

ordinaire s'écrit :

$$Z^*(x, t) = \sum_{i=1}^n \lambda_i Z(x_i, t_i) \quad (3.12)$$

Pour garantir la condition de non-biais :

$$\mathbb{E}[Z^*(x, t)] = \mathbb{E}\left[\sum_{i=1}^n \lambda_i Z(x_i, t_i)\right] = \mu \sum_{i=1}^n \lambda_i \quad (3.13)$$

$$\sum_{i=1}^n \lambda_i = 1 \quad (3.14)$$

L'objectif est de minimiser l'erreur, caractérisée par sa moyenne quadratique $E(Z^* - Z)^2$ sous la condition de non-biais (3.14) en utilisant le multiplicateur de Lagrange μ . Les poids qui minimisent l'erreur sont la solution de :

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(x_i - x_j, t_i - t_j) + \mu &= \gamma(x_i - x, t_i - t) \quad i = 1, \dots, n \\ \sum_{i=1}^n \lambda_i &= 1 \end{aligned} \quad (3.15)$$

Le système d'équations (3.15) est appelé le système de krigeage ordinaire, et sa résolution donne les poids λ_i pour l'estimateur linéaire (3.12).

Krigeage avec dérive externe :

Le krigeage avec dérive externe ou le krigeage de régression suppose que $Z(x, t)$ peut être décomposé en deux parties, une déterministe $\mu(x, t)$ et l'autre stochastique $Y(x, t)$:

$$Z(x, t) = \mu(x, t) + Y(x, t) \quad (3.16)$$

avec Y étant intrinsèquement stationnaire avec une moyenne zéro. f_0, f_1, f_L sont des fonctions déterministes avec $f : D \rightarrow \mathbb{R}$, et $\mu(x, t)$ est une combinaison linéaire de ces fonctions évaluées en (x, t) :

$$\mu(x, t) = \sum_{l=0}^L a_l f_l(x, t) \quad (3.17)$$

avec $f_0(x, t) = 1$

$$Z(x_i, t_i) = \mu(x_i, t_i) + Y(x_i, t_i) = \sum_{l=0}^L a_l f_l(x_i, t_i) + Y(x_i, t_i) \quad (3.18)$$

Les différentes fonctions $f_l(x, t)$ représentent les covariables dites « dérivées externes » utilisées pour estimer la moyenne sous-jacente; dans cette étude, seule une fonction $f_1(x, t) = F(t)$, qui représente les données de la station fixe, a été utilisée.

L'estimateur de krigeage linéaire avec dérivées externes est donc écrit :

$$Z^*(x, t) = \sum_{i=1}^N w_i Z(x_i, t_i) = \sum_{i=1}^N w_i \left(\sum_{l=0}^1 a_l f_l(x_i, t_i) + Y(x_i, t_i) \right) \quad (3.19)$$

La condition de non-biais est satisfaite si et seulement si :

$$\sum_{i=1}^n w_i f_l(x_i, t_i) = f_l(x, t) \quad l = 0, 1 \quad (3.20)$$

Couplée à la condition de variance minimale, cela donne le système de krigeage (3.21) :

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(x_i - x_j, t_i - t_j) + \sum_{l=0}^1 a_l f_l(x_i, t_i) &= \gamma(x_i - x, t_i - t) \quad i = 1, \dots, n \\ \sum_{i=1}^n w_i f_l(x_i, t_i) &= f_l(x, t) \quad l = 0, 1 \end{aligned} \quad (3.21)$$

Résultats

Dans cette section, les résultats obtenus suite à l'application de la méthodologie décrite précédemment sur l'ensemble de données sont présentés. L'étude variographique est abordée en premier, avec la présentation des deux variogrammes directionnels distincts, ainsi que les variogrammes expérimentaux respectifs et leurs correspondants théoriques pour chaque modèle en lice. S'ensuit une prédiction effectuée pour la journée du 4 mars, de 5h à 22h, en utilisant

chacun des trois modèles.

Les résultats de la procédure de validation croisée pour chaque scénario, mettant en avant trois indicateurs de performance. Ces indicateurs sont évalués pour les trois modèles géostatistiques spatio-temporels, ainsi que pour la méthode IDW.

Enfin, Une prédiction de la concentration d'ozone obtenue via le modèle KED ainsi que l'incertitude associée seront présentées. Ces cartographies sont calculées en exploitant l'ensemble des informations disponibles pour la journée considérée.

- Etude Variographique :

- Anisotropie :

La dispersion de la pollution atmosphérique urbaine peut présenter des caractéristiques isotropes ou anisotropes. Dans un contexte isotrope, la concentration de polluants se disperse de manière uniforme, sans privilégier aucune direction spécifique. Cela signifie que la pollution se propage de manière égale dans toutes les directions. À l'inverse, un comportement anisotrope de la pollution atmosphérique indique que la dispersion des polluants varie selon la direction. Par exemple, la pollution pourrait se propager plus rapidement le long des axes routiers principaux qu'entre les bâtiments ou dans les zones moins ventilées. L'anisotropie de la dispersion de la pollution peut être mise en évidence sur les variogrammes expérimentaux, où les caractéristiques du variogramme (seuil, portée) varient selon la direction. En général, on observe que les directions des portées les plus longues et les plus courtes sont orthogonales.

Deux variogrammes directionnels spatio-temporels ont été calculés en utilisant les paires de points dans l'axe nord-sud et dans l'axe est-ouest. La figure 3.7 montre qu'il n'y a pas de différences significatives entre ces deux variogrammes.

- Variance spatio-temporelle :

L'étude de la variabilité spatio-temporelle des données a montré une nette différence entre la variabilité spatiale et la variabilité temporelle. Les différents variogrammes ont montré qu'en moyenne, il y avait une plus grande différence entre deux mesures échantillonnées à quelques heures d'intervalle au même endroit, qu'entre deux mesures échantillonnées au même moment n'importe

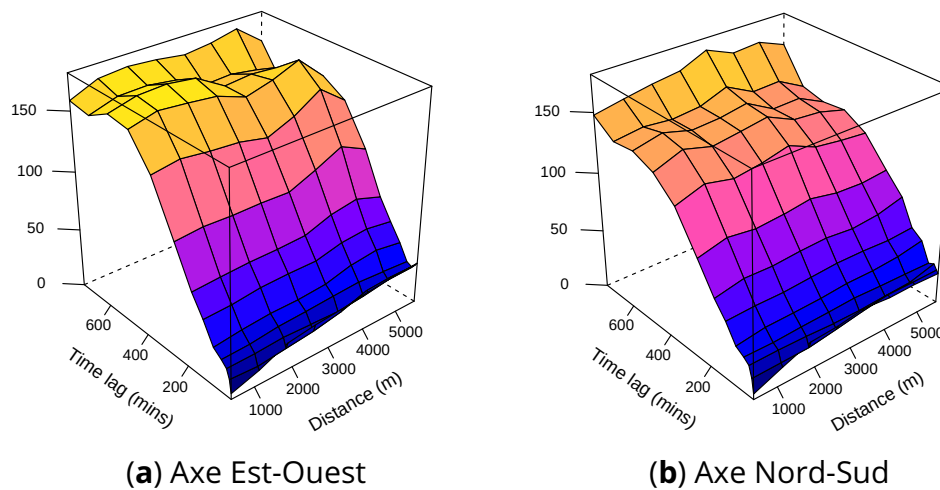


Figure 3.7 – Variogrammes empiriques spatio-temporels directionnels.

où dans l'espace (à l'échelle d'une ville), ce qui justifie l'approche traditionnelle utilisant les stations fixes pour la surveillance de la qualité de l'air.

Les trois variogrammes (données ajustées dans la Figure 3.8, résidus dans la Figure 3.9, et résidus estimés dans la Figure 3.10) montrent exactement la même variabilité purement spatiale. En effet, pour les variogrammes résiduels, la soustraction s'est limitée à la composante temporelle fournie par la station de surveillance fixe, laissant la variabilité spatiale inchangée.

Les trois variogrammes empiriques calculés montrent de petits effets de pépité; cependant, il n'y a pas de données au même moment et au même endroit simultanément, car aucun des tramways ne se croise. Par conséquent, les points de données collectés à proximité proviennent nécessairement du même capteur, et ces mesures ne sont pas indépendantes de la concentration d'ozone (autocorrelation des capteurs). C'est pourquoi ces variogrammes montrent une faible variabilité à l'origine, qui ne reflète pas nécessairement la variabilité réelle du phénomène étudié.

- *Modélisation :*

Le variogramme théorique est construit en intégrant l'anisotropie spatio-temporelle

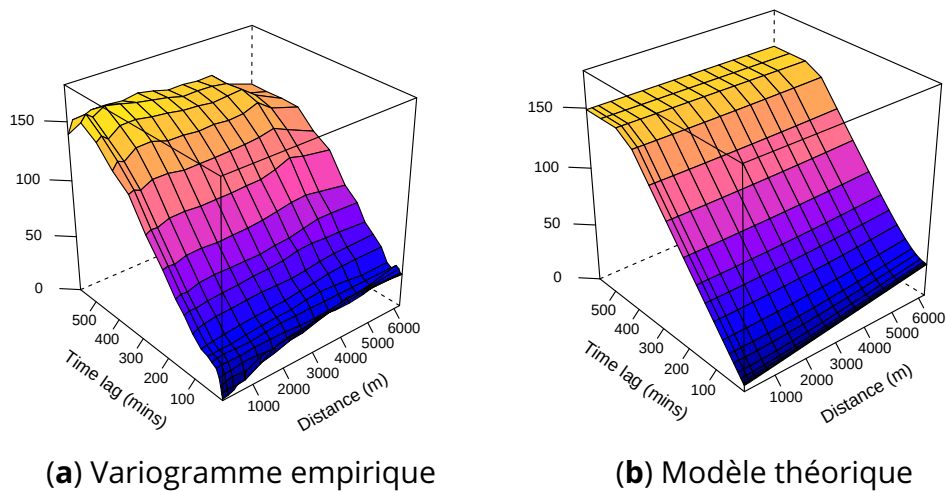


Figure 3.8 – Variogrammes spatio-temporels associés au modèle de krigeage ordinaire.

dans un modèle métrique non séparable :

$$\gamma(x, t) = \gamma_{joint}(\sqrt{x^2 + (C.t)^2}) \quad (3.22)$$

où γ_{joint} est un variogramme connu qui peut inclure un effet de pépité, et C est un paramètre d'anisotropie spatio-temporelle défini comme le nombre d'unités spatiales équivalent à une unité de temps. L'estimation de C a été effectuée en même temps que tous les autres paramètres du modèle théorique (c'est-à-dire le seuil, la pépité, et la portée) en minimisant la moyenne des écarts quadratiques entre l'échantillon et la surface du variogramme ajusté [18]. L'algorithme d'optimisation utilisé est L-BFGS-B [21], qui est la variante à contraintes de l'algorithme d'optimisation Broyden-Fletcher-Goldfarb-Shanno à mémoire limitée. Les différents modèles et leurs paramètres respectifs peuvent être trouvés dans le Tableau 3.1 pour les trois modèles.

Table 3.1 – Différents modèles et leurs paramètres respectifs.

Méthode	Modèle S-P	C(m/s)	Modèle Conjoint	Seuil(($\mu\text{g}/\text{m}^3$) ²)	Pépité(($\mu\text{g}/\text{m}^3$) ²)	Portée(m)
Krigeage simple	Métrique	105.16	Sphérique	82.30	5.00	30415.43
Krigeage ordinaire	Métrique	91.18	Linéaire	148.8	5.00	38073.4
Krigeage avec dérive externe	Métrique	83.03	Exponentiel	59.86	2.00	9872.405

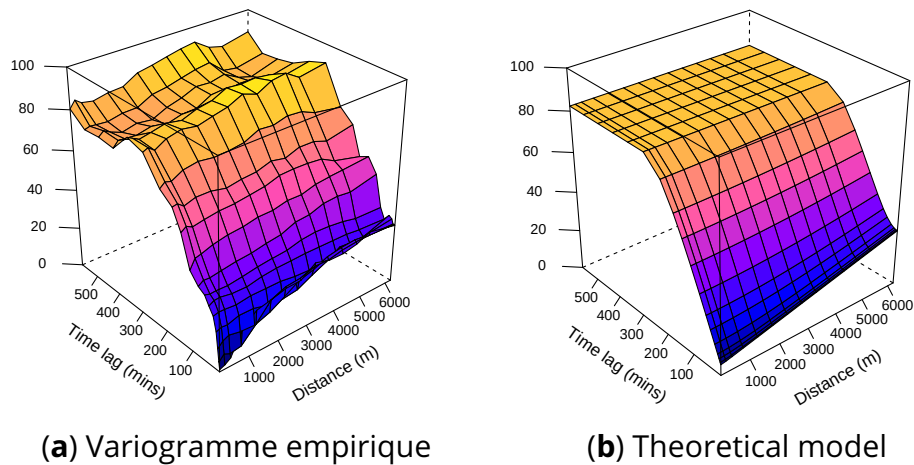


Figure 3.9 – Variogrammes spatio-temporels associés au modèle de krigeage simple.

Comme prévu, le modèle de variogramme associé au krigeage ordinaire a montré la portée et le seuil les plus élevés, contrairement aux deux autres modèles, où les données de la station fixe ont partiellement expliqué la variance, ce qui a entraîné une portée et un seuil plus faibles.

- Séries spatio-temporelles :

Les figures 3.11, 3.12 et 3.13 montrent les prévisions pour les différentes trajectoires des lignes de tramway, dans le premier, deuxième et troisième scénarios. Seuls quatre capteurs fonctionnaient ce jour-là : les capteurs des lignes 4, 7, 8 et 13.

Les prédictions des trois modèles montrent une similitude notable, attribuable à une variabilité spatiale commune aux trois variogrammes. Cette variabilité spatiale, étant moins prononcée que la variabilité temporelle, a conduit les estimateurs à privilégier des données spatialement proches. Étant donné que cette variabilité spatiale reste constante entre les modèles, les prédictions sont naturellement similaires.

Il est important de souligner que les estimateurs n'ont pas interpolé les données

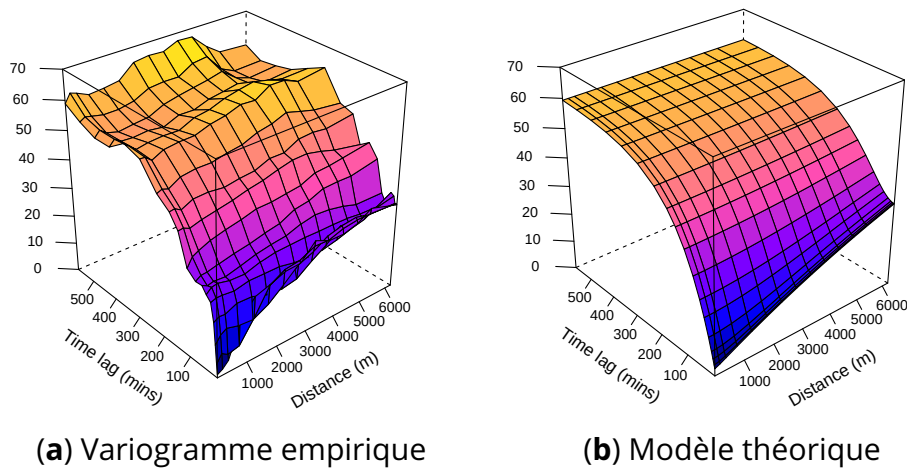


Figure 3.10 – Variogrammes spatio-temporels associés au modèle de krigeage avec dérive externe.

exactement aux emplacements échantillonnés. Cela est dû à l'effet de pépité, représentant les erreurs de mesure, rendant ces estimateurs inexacts. Par conséquent, ils ont tendance à filtrer ces erreurs.

Dans le scénario 3 (Figure 3.13), en l'absence de données pour une ligne de tramway, les prédictions s'alignent sur celles des lignes voisines. Par exemple, les prédictions pour la ligne 2 étaient analogues aux données de la ligne 17, et inversement.

Les incohérences dans les prédictions à certains emplacements sont principalement dues à l'absence de données proximales, un phénomène plus marqué dans le troisième scénario. Cette inadéquation est exacerbée en fin de journée, où on observe une divergence significative dans les mesures.

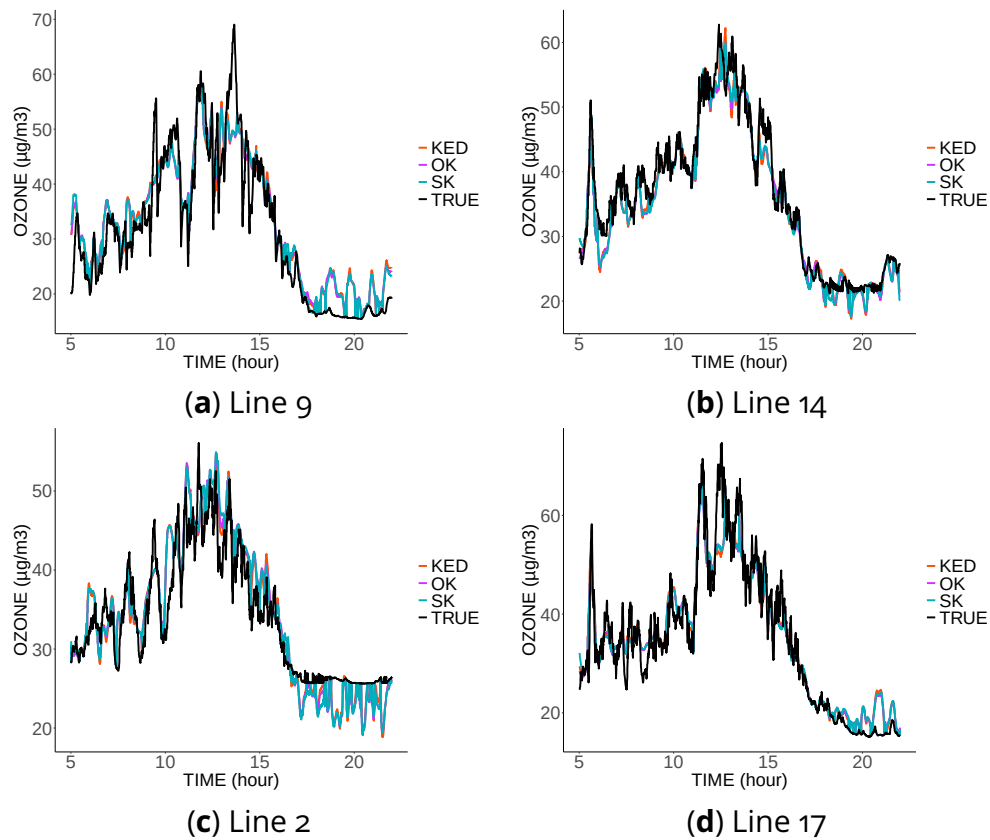


Figure 3.11 – Comparaison entre les prévisions et les valeurs réelles de quatre lignes de tramway le 4 mars 2016 dans le premier scénario.

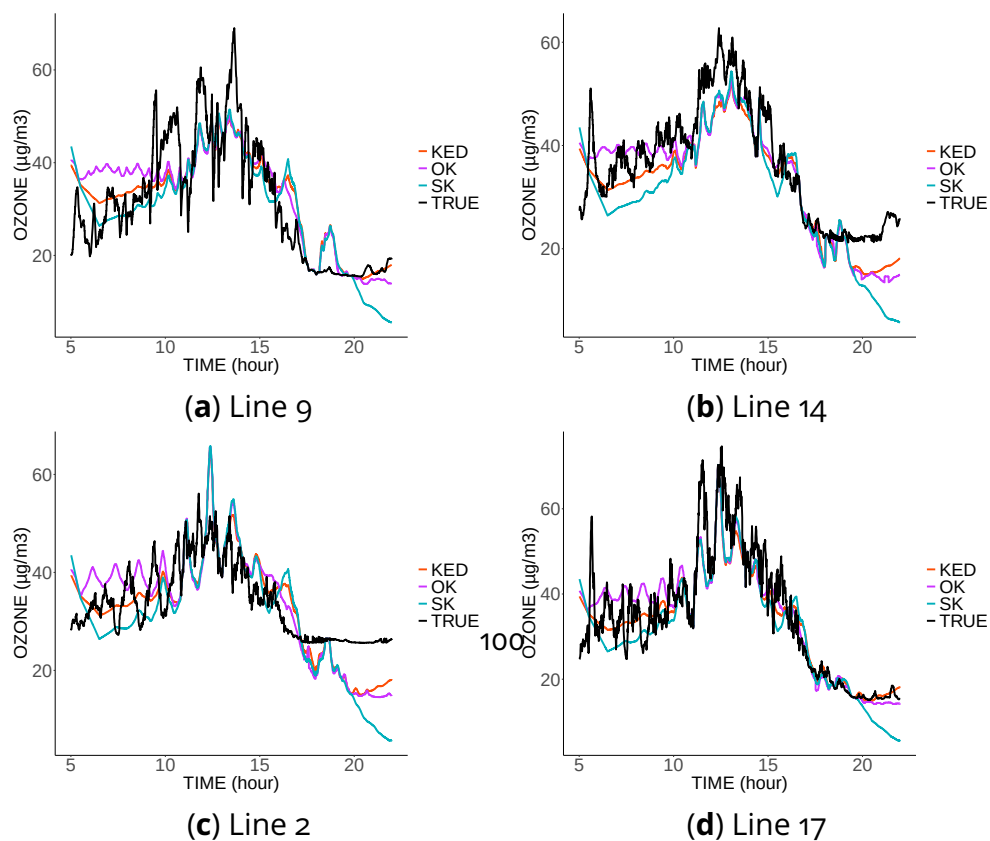


Figure 3.12 – Comparaison entre les prévisions et les valeurs réelles de quatre lignes de

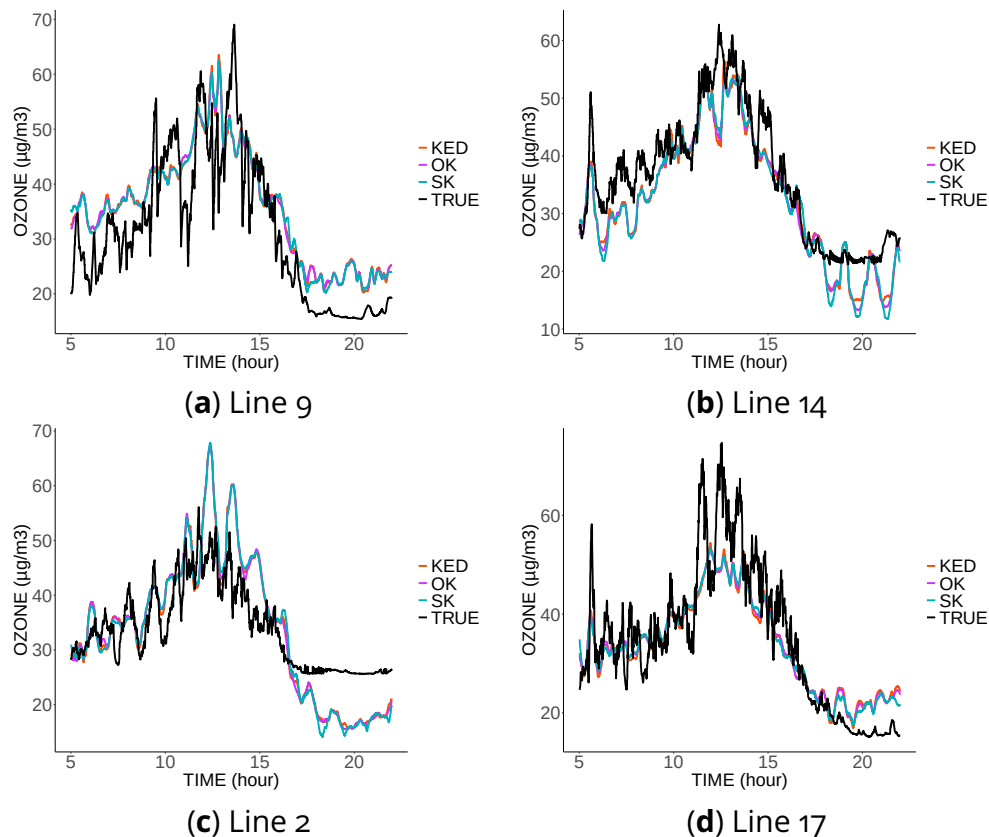


Figure 3.13 – Comparaison entre les prévisions et les valeurs réelles de quatre lignes de tramway le 4 mars 2016 dans le troisième scénario.

- Résultats des indicateurs de performance :

La première chose à remarquer dans le RMSE (Figure 3.14) est que les trois méthodes probabilistes sont nettement plus performantes que l'interpolation déterministe dans chaque scénario. Comme on pouvait s'y attendre, dans le premier scénario, plus le nombre de données utilisées était élevé, moins il y avait d'erreurs. Ce n'est pas le cas dans le troisième scénario où nous avons remarqué que l'erreur atteignait un minimum. Quelle que soit la quantité de données utilisées, le RMSE n'est pas descendu en dessous de $6 \mu\text{g}/\text{m}^3$.

L'estimateur KED a montré les erreurs les plus faibles dans le cas de la reconstruction des données à des endroits proches des données échantillonnées (scé-

nario 1) suivi par SK et OK. Nous avons conclu que la contribution des données des stations fixes dans un tel environnement était utile et que le KED optimisait son utilisation. Dans les deux autres scénarios, ces conclusions sont moins pertinentes.

Les biais des quatre méthodes tendent vers zéro dans le premier scénario, tandis que dans le troisième scénario, la prédiction semble systématiquement biaisée (Figure 3.15). Bien que les méthodes stochastiques soient systématiquement plus performantes que la méthode IDW, cette supériorité n'est pas aussi manifeste dans le deuxième scénario. La figure de corrélation 3.16 consolide l'idée que le KED semble le mieux adapté dans le premier scénario, alors que les résultats sont moins affirmatifs de corrélation dans les deuxième et troisième scénarios.

Globalement, le premier scénario montre des indicateurs de performance stables, avec une amélioration des prédictions à mesure que davantage de données sont intégrées. Cette tendance n'est pas observée dans le troisième scénario, où un plateau est atteint indépendamment du volume de données. Le deuxième scénario présente des caractéristiques intermédiaires entre les deux premiers.

Il a été conclu dans cette étude que le krigeage avec dérive externe, utilisant les données de la station de mesure fixe comme variable externe, s'avérerait le plus approprié pour l'interpolation des données. Pour l'extrapolation loin des lieux d'échantillonnage, le krigeage ordinaire semble tout aussi performant.

Comme prévu, et à l'instar de la majorité des méthodes de reconstruction des données, la géostatistique est plus performante dans le cas de l'interpolation que dans celui de l'extrapolation, quel que soit le critère de performance considéré.

- Production de cartographies :

Pour répondre à l'objectif de création de cartes de pollution, l'algorithme KED

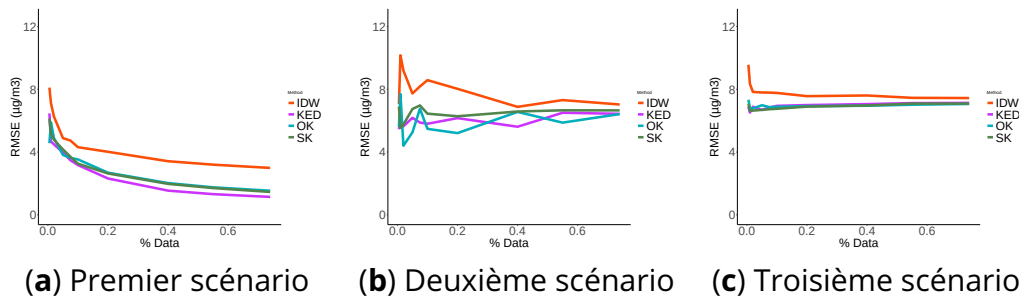


Figure 3.14 – RMSE en fonction du pourcentage de données utilisé dans les trois scénarios étudiés.

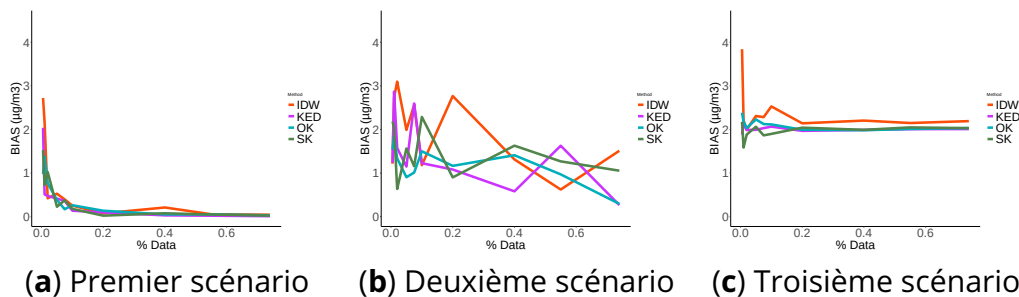


Figure 3.15 – Biais en fonction du pourcentage de données utilisé dans les trois scénarios étudiés.

a été déployé en utilisant l'ensemble des données des capteurs mobiles et de la station de surveillance fixe sur une journée. La Figure 3.17 illustre les cartographies obtenues pour le 4 mars 2016, couvrant la plage horaire de 5 heures à 22 heures, période d'activité des quatre capteurs mobiles.

Ces cartographies ont permis d'identifier les zones de forte concentration d'ozone à Zurich, malgré l'utilisation de seulement quatre capteurs mobiles. La Figure 3.17 met en évidence un pic de concentration d'ozone en milieu de journée, suivi d'une diminution significative en soirée et la nuit.

Les concentrations augmentent à partir de 6 heures du matin, atteignant un maximum entre 12 heures et 13 heures, avec des valeurs dépassant $60 \mu\text{g}/\text{m}^3$ au nord-ouest de la ville. Vers 19 heures, les concentrations reviennent à des niveaux similaires à ceux de la nuit précédente, approximativement $20 \mu\text{g}/\text{m}^3$.

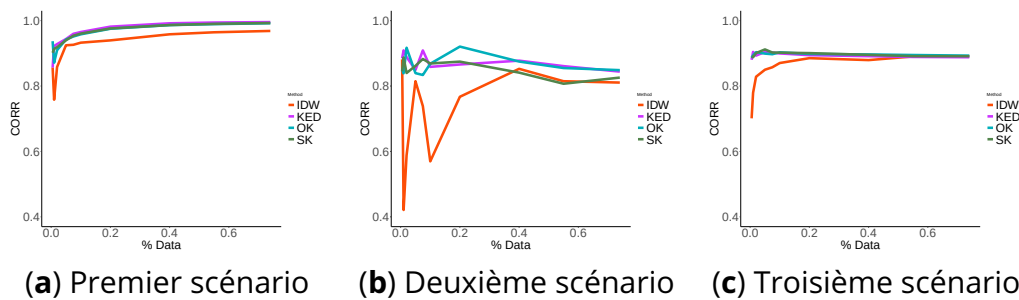


Figure 3.16 – Corrélations en fonction du pourcentage de données utilisé dans les trois scénarios étudiés.

Comme mentionné auparavant, l'un des avantages des modèles géostatistiques est de fournir une incertitude de prédiction. La Figure 3.18 montre la variance associée à l'estimateur KED.

La variance de krigeage n'est pas liée aux valeurs des données, mais uniquement à leur emplacement; c'est pourquoi il n'y a pas de corrélation entre les figures 3.17 et 3.18. Plus l'emplacement des données collectées est éloigné, plus la variance est élevée, et inversement. La relation entre la variance et la distance par rapport aux données est directement influencée par le variogramme spatio-temporel.

Comme prévu, la variance était minimale dans le centre de la ville, où le nombre de données collectées était le plus élevé. Les emplacements des quatre capteurs sont facilement visibles à certains moments de la journée (11 heures ou 17 heures).

La variance maximale peut être observée au bord des cartes montrées à 5 heures du matin et à 22 heures. Ces deux cartes ont la particularité d'avoir des données collectées d'un seul côté du temps, ce qui entraîne une grande distance temporelle (début et fin de la journée) en plus d'une grande distance spatiale (bord de la carte) par rapport aux données collectées, ce qui, comme nous l'avons dit plus haut, implique une plus grande incertitude.

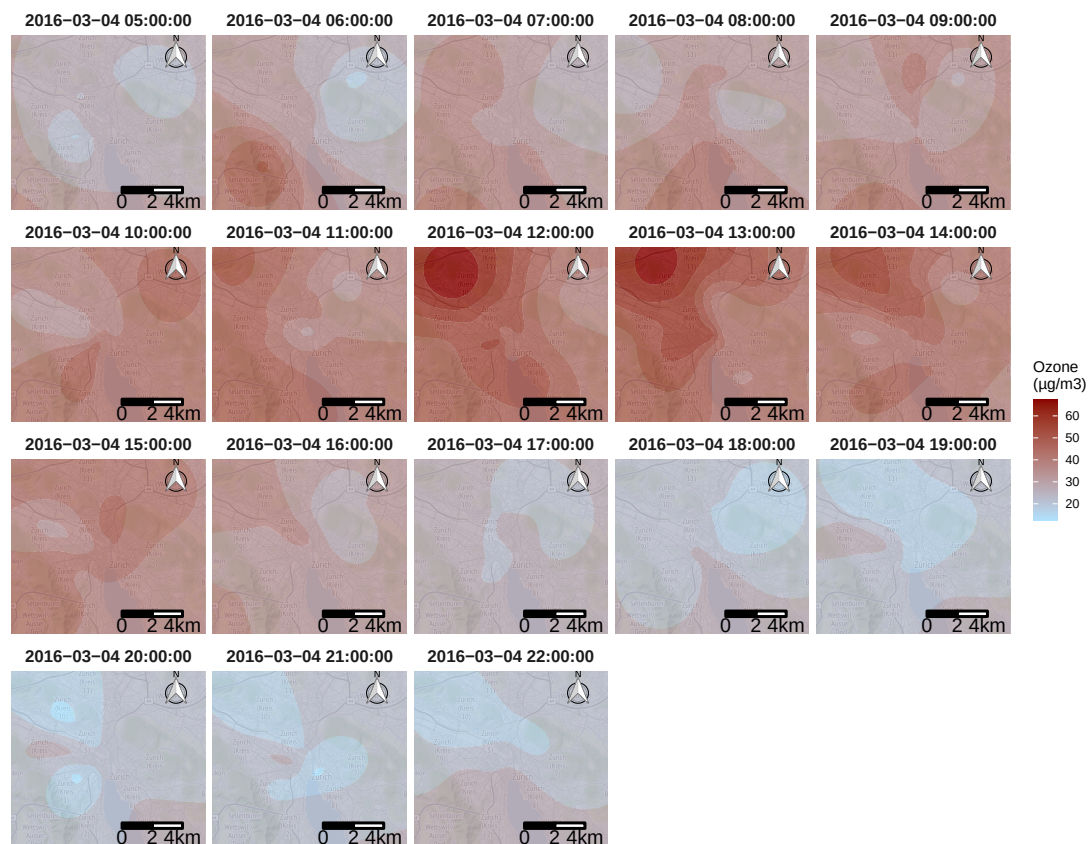


Figure 3.17 – Cartes des concentrations d’ozone résultant de l’estimateur KED à Zurich, ici pour le 4 mars 2016. De 5 heures à 22 heures.

Discussion

La mise en œuvre de la géostatistique spatio-temporelle dans cette étude a révélé son potentiel en tant qu’outil efficace pour traiter les défis posés par les capteurs de surveillance mobiles. Là où de nombreuses études dépendent de diverses covariables pour estimer la qualité de l’air, cette approche se distingue en générant des cartographies de la pollution atmosphérique en se basant exclusivement sur les données des capteurs.

Malgré la soustraction des données des stations fixes, une grande variabilité spatio-temporelle persiste. Cela démontre que les capteurs mobiles sont particulièrement aptes à capter cette variabilité.

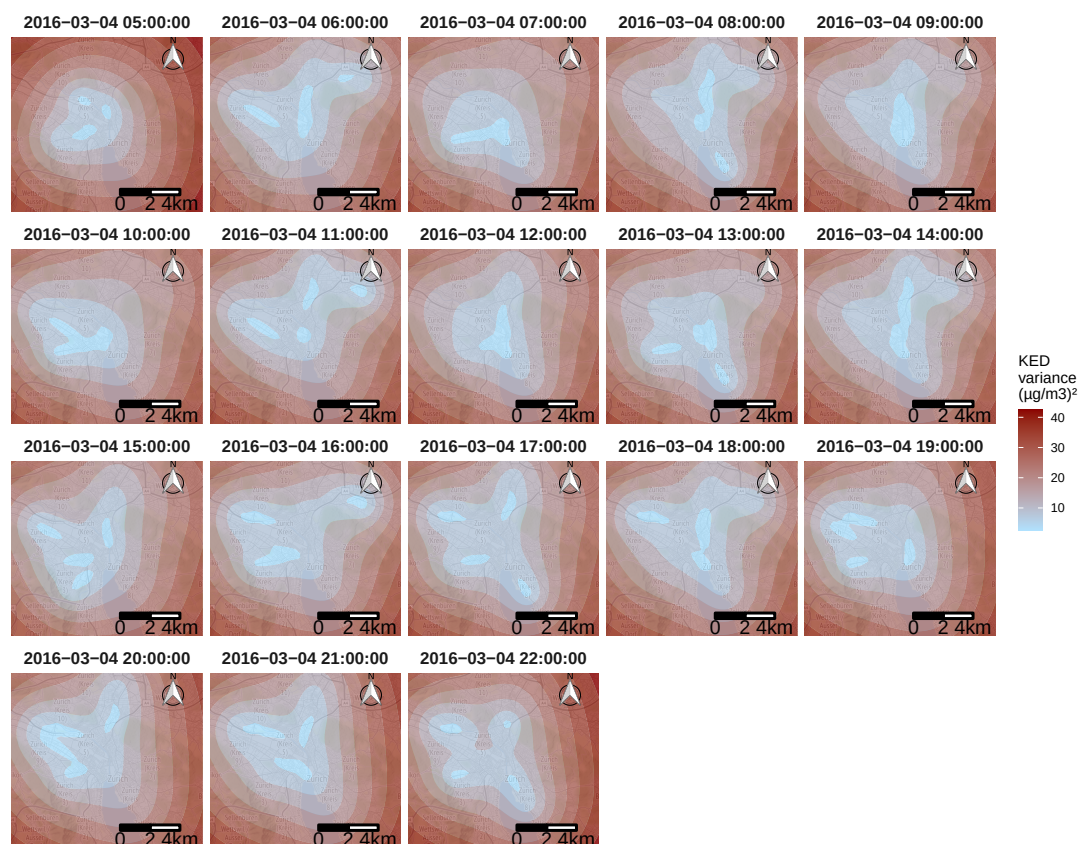


Figure 3.18 – Cartes de l'estimateur de variance résultant de l'estimateur KED à Zurich, présentées ici pour le 4 mars 2016. De 5 heures du matin à 10 heures du soir.

Toutefois, l'étude présente plusieurs limites qui doivent être soulignées :

- Les tramways n'ont pas parcouru tous les types de voies et ont donc enregistré un type spécifique de pollution urbaine.
- La méthodologie n'a pas été employée pour déterminer le meilleur modèle d'estimation de la concentration d'ozone, mais plutôt la concentration mesurée par des capteurs analogues à ceux utilisés dans cette étude.
- Cette étude n'a pas pris en compte l'autocorrélation des erreurs des données provenant d'un même tramway, puisque les données des capteurs mobiles ont été traitées comme indépendantes, conditionnellement aux concentrations d'ozone.

Le krigeage ordinaire, bien qu'il n'utilise pas directement les données des stations fixes, est indirectement influencé par ces dernières, notamment lors

de l'étalonnage. Ainsi, sa moyenne est ajustée en fonction de ces stations.

L'absence d'une relation observable entre les coordonnées spatiales et la variable d'intérêt (concentration d'ozone) a rendu impossible l'utilisation du krigeage universel². Le manque de variables auxiliaires rend également les prédictions en dehors des zones de collecte hautement incertaines. Aucun des modèles employés n'a pu extrapoler au-delà de la plage de données existantes, ce qui limite les prédictions aux valeurs extrêmes des sous-ensembles de données.

Il convient de souligner que l'ajout d'anisotropie, augmentant la complexité du modèle, ne garantit pas une amélioration significative des prédictions. Ainsi, une voie plus fructueuse pourrait être d'enrichir le modèle avec des données additionnelles, telles que des informations météorologiques ou de trafic, en s'appuyant sur des travaux antérieurs.

Conclusion

Dans cette étude, l'exploration et la comparaison des performances de trois méthodes géostatistiques spatio-temporelles ont été réalisées face à une approche d'interpolation déterministe. Il est clair que les méthodes probabilistes surpassent la méthode déterministe. La géostatistique univariée s'est avérée particulièrement efficace pour l'interpolation des sites proches des zones d'échantillonnage.

Lorsqu'il s'agit de prédictions au-delà de ces zones, l'incorporation de variables auxiliaires, sous la forme de cokrigeage ou de krigeage avec dérive externe se révèle nécessaire.

Le krigeage avec dérive externe, qui se présente comme le choix optimal à l'issue de cette première étape de comparaison, offre une polyvalence et une capacité à intégrer d'autres sources de données sous forme de dérives. De plus, il fournit des estimations d'incertitude qui peuvent guider le placement des capteurs et son caractère descriptif favorise une meilleure compréhension du phé-

2. Équivalent mathématiquement au KED, le krigeage universel utilise les coordonnées spatio-temporelle comme dérives.

nomène étudié.

En conclusion, bien que la géostatistique univariée ait ses avantages, l'intégration de la géostatistique multivariée pourrait offrir des perspectives intéressantes pour les recherches futures. Cela nécessitera une comparaison minutieuse de plusieurs méthodes en termes de complexité, d'erreur de prédiction, de données utilisées. L'étape logique suivante serait de complexifier le modèle en incorporant des covariables, afin d'optimiser davantage la prédiction de la qualité de l'air, qui sera l'objet du prochain sous-chapitre.

3.2 Intégration et comparaison de modèles exploitant des données externes

Dans cette section, nous élargissons notre exploration des modèles mathématiques dédiés à la cartographie de la qualité de l'air urbain en intégrant des données externes. L'objectif est d'évaluer l'impact positif potentiel de ces données additionnelles sur la précision des prédictions.

Le sous-chapitre 3.1 a mis en évidence les contraintes des méthodes se basant exclusivement sur l'utilisation des données de pollution. Même avec l'adoption de modèles probabilistes avancés, l'omission de variables auxiliaires peut rendre les prédictions hors des zones d'échantillonnage incertaines.

Dans cette phase de recherche, l'étude se fait sur un ensemble de données fourni par AtmoTrack. La pertinence de cet ensemble réside dans la fusion des observations de pollution (PM₁₀) issues tant de capteurs mobiles que de capteurs fixes à bas coût, ce qui enrichit la dimension d'analyse.

Treize variables auxiliaires sont intégrées dans dix modèles distincts issus de diverses familles de méthodes de la littérature. L'efficacité de chaque modèle à incorporer ces variables pour affiner les prédictions de la qualité de l'air urbain sera évaluée. Les spécificités des données et des modèles seront détaillées dans la section méthodologique qui suit.

En confrontant ces modèles, notre ambition est de distinguer les approches les plus efficaces pour une représentation précise et en temps réel de la pollution urbaine. Les conclusions tirées orienteront les choix méthodologiques pour les phases ultérieures de cette thèse.

3.2.1 Méthodologie

La méthodologie adoptée dans cette étude est structurée en trois étapes principales : la présentation des données utilisées, l'approche d'évaluation des

modèles, et enfin, la description des modèles appliqués à notre ensemble de données.

A - Données

Un ensemble diversifié de données pour cette étude a été exploité, englobant des informations sur la pollution atmosphérique provenant de capteurs à bas coût et de stations fixes de surveillance. De plus, des données contextuelles liées à l'environnement des observations ont été intégrées. Cette partie vise à détailler ces variables, à préciser leurs origines et à illustrer leur pertinence.

- Données sur la pollution atmosphérique :

Les données de pollution utilisées proviennent de trois sources distinctes :

- 1- Les données primaires sur la qualité de l'air, fournies par Atmotrack, sont issues de capteurs à bas coût.
- 2- Les mesures de référence sur la qualité de l'air sont extraites des stations fixes situées à Nantes.
- 3- Les informations sur la qualité de l'air dérivées du modèle de chimie-transport CHIMERE servent de covariables.

Pour rappel, le jeu de données fourni par Atmotrack contient à la fois des données de capteurs bas coût mobiles embarqués sur des véhicules utilitaires, ainsi que des données de capteurs fixes intelligemment placés, notamment à proximité des stations fixes de la qualité de l'air.

En plus des données à faible coût provenant d'Atmotrack, nous avons accès aux données de la station de surveillance fixe gérée par Air Pays-de-la-Loire et disponibles sur le portail dédié de « Data Air Pays de la Loire » [13]. Il s'agit d'une série de mesures provenant de cinq stations de référence pour tout le mois de novembre à une fréquence d'une observation toutes les 15 minutes. L'emplacement des capteurs fixes à faible coût d'Atmotrack et de la station de surveillance fixe peut être trouvé dans la Figure 1.7.

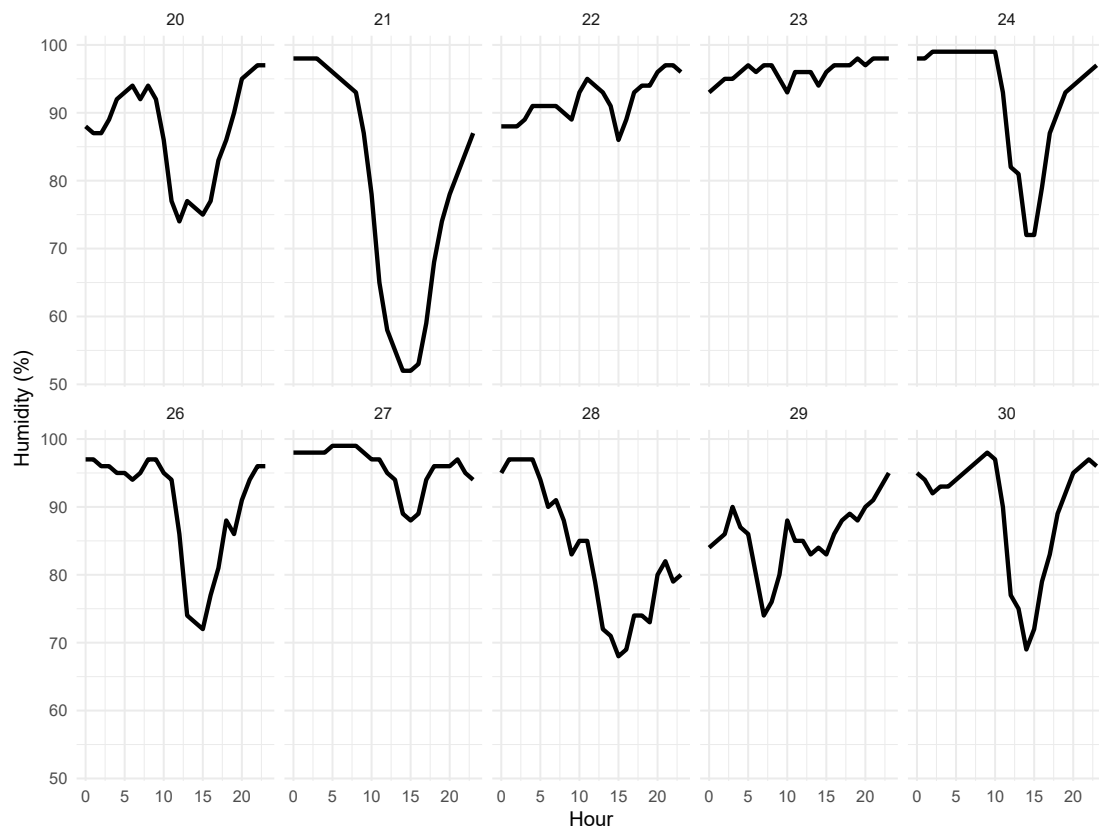


Figure 3.19 – Données d'humidité dans la ville de Nantes, du 20 au 30 novembre 2018.

- Données auxiliaires :

Différentes études ont utilisé diverses catégories de covariables, comme les données des stations de surveillance fixes dans la première étude de ce chapitre [11], les données de pollution moyenne annuelle dans Gressent et al.[8], ou les covariables d'utilisation du sol dans Hesenfrats et al.[10].

Outre les mesures de pollution provenant de capteurs à bas coût et de stations de surveillance, l'intégration de covariables supplémentaires est cruciale pour établir des relations entre les mesures de pollution et divers facteurs environnementaux. Une fois cette relation établie, le modèle peut prédire les concentrations de polluants dans des zones non couvertes par les mesures, en se basant sur ces informations. Il est donc essentiel que ces covariables soient disponibles sur l'ensemble de la zone d'étude, y compris dans les régions non couvertes par les capteurs. Plusieurs types de covariables ont été utilisés dans cette

étude, divisés en trois groupes :

Covariables temporelles

Les covariables temporelles sont des variables qui varient uniquement avec le temps et non dans l'espace. En d'autres termes, si deux observations sont prises à des endroits différents de la ville mais au même moment, elles auraient la même valeur pour ces covariables. Dans cette étude, les données météorologiques, plus précisément la température et l'humidité, ont été utilisées comme covariables temporelles.

Bien que ces variables ne soient pas uniformément réparties dans toute la ville, comprendre leur tendance temporelle globale peut aider à prévoir les tendances des concentrations de polluants. Les données météorologiques ont été obtenues à partir du site « public.opendatasoft.com » [16].

La figure 3.19 présente les données d'humidité pour la période d'étude.

Covariables spatiales

Contrairement aux covariables temporelles qui varient uniquement dans le temps, les covariables spatiales varient uniquement dans l'espace. Cela signifie que si deux observations sont prises au même endroit, quel que soit le moment où elles ont été collectées, elles auront la même valeur pour ces covariables. Ces covariables restent donc inchangées pour les capteurs fixes à faible coût et les stations de surveillance de la qualité de l'air. La plupart de ces covariables ont été obtenues à partir site « Open Street Map » [17].

Le trafic routier étant un facteur important affectant la concentration de PM et en l'absence de données directes sur le trafic routier, l'attention s'est portée sur des covariables spatiales telles que la proximité des différents types de routes présentes dans la zone d'étude. De plus, le nombre de bâtiments dans un rayon de 500 mètres autour de chaque observation a été quantifié pour compléter l'analyse. Le modèle numérique de terrain de Nantes a été utilisé pour déterminer l'altitude ou l'élévation de chaque observation. La figure 3.20 illustre certaines de covariables spatiales.

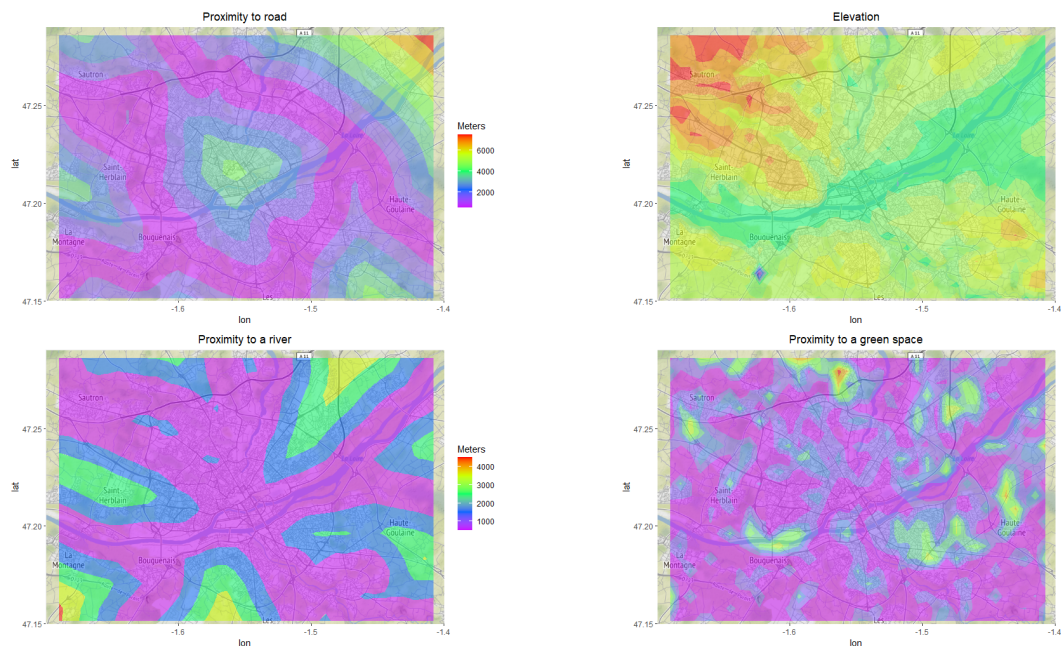


Figure 3.20 – Illustration de certaines covariables spatiales utilisées dans cette étude.

Covariables spatio-temporelles

CHIMERE est un modèle de transport chimique tridimensionnel, basé sur l'intégration de l'équation de continuité de masse pour les concentrations de plusieurs espèces chimiques. Ce modèle est utilisé par un grand nombre d'agences de surveillance de la qualité de l'air en Europe et est le plus utilisé en France. Le Service de surveillance de l'atmosphère de Copernicus (CAMS) [6] fournit des prévisions quotidiennes issues du modèle CHIMERE sur 4 jours des concentrations des principaux polluants atmosphériques, dans les couches les plus basses de l'atmosphère pour le domaine européen. Leur couverture horizontale est de $0,1^\circ$ (approximativement 10 à 20 km). Parallèlement aux prévisions, les modèles effectuent des analyses quotidiennes des polluants près de la surface en assimilant, c'est-à-dire en fusionnant, des observations vieilles de un jour avec les modèles. L'analyse et la prévision à un jour ont été sélectionnées comme covariables pertinentes. Ces deux covariables spatio-temporelles sont disponibles à des intervalles de temps horaires. Il n'y a pas de grande variance spatiale à l'échelle de notre zone d'étude. La figure 3.21 montre la variation tempo-

relle de ces deux covariables.

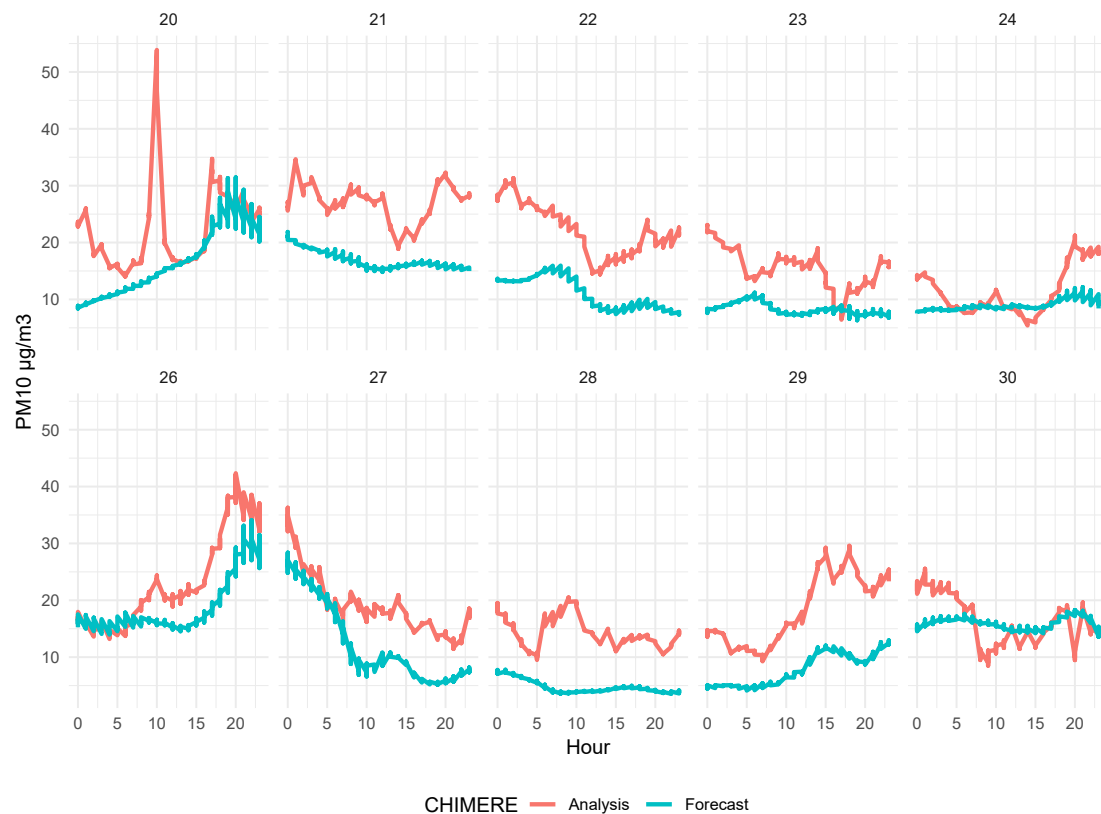


Figure 3.21 – Sorties PM₁₀ de prévisions et d’analyses du modèle CHIMERE du 20 au 30 novembre 2020 pour la ville de Nantes

Afin d’assurer que chaque mesure issue des capteurs à bas coût soit correctement associée à ses covariables respectives, une interpolation est nécessaire. En effet, certains échantillons de capteurs à bas coût sont enregistrés à une fréquence supérieure à 1Hz, tandis que certaines covariables ne sont disponibles qu’à des intervalles de 15 minutes ou d’une heure. Le tableau 3.2 récapitule l’ensemble des covariables employées dans cette recherche, en précisant leur source et leur dimension.

Variable	Origine	Dimension
Température	public.opendatasoft	Temporelle
Humidité	public.opendatasoft	Temporelle
Altitude	data.nantesmetropole	Spatiale
Proximité de : autoroute	Open Street Map	Spatiale
Proximité de : voie principale	Open Street Map	Spatiale
Proximité de : voie primaire	Open Street Map	Spatiale
Proximité de : voie secondaire	Open Street Map	Spatiale
Proximité de : voie tertiaire	Open Street Map	Spatiale
Proximité de : voie résidentielle	Open Street Map	Spatiale
Proximité d'un espace vert	Open Street Map	Spatiale
Proximité d'un site industriel	Open Street Map	Spatiale
Nb de bâtiments dans 500m	Open Street Map	Spatiale
Proximité d'une rivière	Open Street Map	Spatiale
Analyse CHIMERE	CAMS	Spatio-temporelle
Prévision CHIMERE	CAMS	Spatio-temporelle

Table 3.2 – Table récapitulative des covariables utilisées.

B - Procédure d'évaluation et de comparaison des modèles

Afin d'adresser les questions de recherche, un ensemble de modèles issus de la littérature a été sélectionné, représentant une diversité de principes de modélisation.

Les performances des modèles sont évaluées à travers deux scénarios distincts :

- **Interpolation temporelle** : L'objectif est de reconstituer une représentation de la pollution pour une période où des observations sont disponibles. Concrètement, les données du jour J servent à estimer la pollution du même jour J.
- **Prévision** : Il s'agit d'estimer les niveaux de pollution pour une période sans observations disponibles, en se basant sur des observations antérieures. Ainsi, les données du jour J servent à prédire la pollution du jour J+1.

Dans chacun des deux scénarios, deux types d'évaluations seront menés :

- **Validation croisée sur les capteurs à faible coût** : Dans le premier scénario, une validation croisée est effectuée en excluant un capteur à chaque itération. Les modèles sont entraînés sur N-1 capteurs et validés sur le cap-

teur restant. Pour le second scénario, les capteurs du jour J servent à l'entraînement du modèle et la validation sera effectuée sur les données des capteurs pour le jour $J+1$.

- **Validation sur les stations fixes de surveillance de la qualité de l'air :**

Les mesures des stations fixes sont considérées comme des mesures de référence. Même s'il n'existe pas de correspondance directe entre leurs mesures et celles des capteurs à faible coût, les deux types de dispositifs mesurent les mêmes polluants. Cette évaluation permet d'apprécier la pertinence des capteurs à faible coût, indépendamment des modèles, de leur calibration, et de leur potentiel à compléter ou remplacer le réseau traditionnel de stations fixes. Dans le premier scénario, les modèles entraînés sur les données des capteurs bas coût du jour J sont utilisés pour estimer les valeurs des stations fixes pour le même jour. Dans le second, les données des capteurs bas coût du jour J servent à prédire les valeurs des stations fixes pour le jour $J+1$.

Pour chaque scénario et méthode de validation, et pour chaque modèle, quatre indicateurs de performance sont calculés : la racine de l'erreur quadratique moyenne, le biais, le coefficient de corrélation, et l'erreur absolue maximale (MAE).

En conclusion, une comparaison des cartes produites par tous les modèles, en utilisant les données du jour le plus fréquemment échantillonné, sera présentée.

C - Modèles

Dans cette partie, le développement détaillé des 10 modèles utilisés et des paramètres appliqués sera présenté. Les mesures de concentration des PM ont été préalablement transformées en logarithme. Soit Z le vecteur des observations issues des capteurs bas coût (log-transformées), Z_i la i ème observation, X_i la covariable associée à l'observation i , P la dimension de X et n la taille de Z . Les analyses ont été réalisées avec R.

Pondération inverse de la distance

Vu au sous chapitre précédent et décrite par l'équation 3.4, cette méthode déterministe sera la seule des 10 méthodes à ne pas utiliser de covariables.

On évaluera ainsi l'apport potentiel des méthodes utilisant des covariable à une méthode déterministe de référence.

Régression linéaire

La Régression Linéaire (LR) est un outil statistique largement adopté en raison de sa simplicité, de sa faible complexité et de sa capacité descriptive. Contrairement à la méthode IDW, elle nécessite un ensemble conséquent de covariables. La formule de la régression linéaire suppose une relation linéaire entre les observations et les covariables :

$$Z_i = \beta X_i + \epsilon_i \quad (3.23)$$

Avec $\epsilon_i \sim N(0, \sigma^2)$. Les coefficients β sont estimés soit par la méthode des moindres carrés ordinaires, soit par le maximum de vraisemblance, en supposant que les erreurs ϵ sont indépendantes et normalement distribuées. Ainsi, Z est également supposé suivre une distribution normale. Le terme βX représente la tendance générale de Z .

La régression linéaire suggère que la concentration d'un polluant est directement et linéairement associée à ses covariables. Bien que l'on présume que les observations soient indépendantes, cette présomption est fréquemment remise en question à cause des processus de dispersion en matière de pollution. Cependant, un modèle de régression linéaire, malgré sa simplicité, peut capturer une part significative de la variance des données.

Régression en réseau

Se basant sur marjovi et al. [14], la Régression en Réseau (NR pour Network Regression) présente de nombreuses similitudes avec la régression linéaire, postulant que les covariables influencent linéairement la concentration des polluants.

Pour surmonter le défi de l'indépendance des observations, une covariable supplémentaire, représentant la concentration moyenne de PM des rues similaires, est intégrée à l'équation :

$$Z_i = \beta X_i + \epsilon_i + \beta_{p+1} \sum_{[i-j] \in E}^{n'} Z_j / n' \quad (3.24)$$

Le dernier terme de l'équation évoque l'utilisation des observations moyennes des segments de rue connectés (dans le réseau E) au segment d'observation i pour estimer la concentration de pollution de ce segment. La principale complexité réside dans l'identification des rues corrélées, c'est-à-dire dans la construction du réseau E . Marjovi et al.[14] ont conceptualisé la distance entre les rues corrélées en termes de similarité des covariables.

$$d(i, j) = \|X_i - X_j\| \quad (3.25)$$

Afin d'établir la connectivité entre les segments de rue, leurs distances sont calculées deux à deux, et chaque segment est relié à son voisin le plus proche en termes de covariables. Ce processus est répété jusqu'à ce que le graphe E soit entièrement connecté.

Modèle Additif Généralisé

Utilisé par Hasenfratz et al.[10], le Modèle Additif Généralisé (GAM) postule que les observations sont indépendantes et qu'une relation non linéaire existe entre ces observations et les covariables :

$$Z_i = S(X_i) + \epsilon_i \quad (3.26)$$

où S représente une fonction non linéaire lisse des covariables X . Dans cet étude, S est définie comme une spline de régression lisse avec un maximum de 9 degrés de liberté.

Les modèles additifs généralisés sont à la frontière entre les modèles de régressions et les modèles d'apprentissage automatique.

Géostatistique

L'introduction des covariables dans le modèle géostatistique s'effectue naturellement par le biais du krigeage avec dérivées externes. Sa formulation mathématique est équivalente à l'équation de krigeage avec dérivée externe que nous avons présentée précédemment. Cependant, dans ce contexte, plutôt que de se baser sur une unique fonction comme dérive, les 13 covariables sont utilisées.

Les observations Z des capteurs bas coût sont modélisées comme une réalisation bruitée d'un processus stochastique, il est décomposé en deux parties, une déterministe, l'autre aléatoire :

$$Z_i(x_i, t_i) = \beta X_i + \nu_i(x_i, t_i) \quad (3.27)$$

Ainsi, le modèle permet de capturer la tendance globale par la partie déterministe en utilisant une régression linéaire, tandis que la variabilité restante est capturée par le processus aléatoire $\nu(x, t)$.

Le processus aléatoire $\nu(x, t)$ suit un processus gaussien avec une moyenne nulle et une fonction de covariance spatio-temporelle isotrope $R(\Theta)$.

Deux manières d'estimer les paramètres du modèle seront comparées :

- Maximum de vraisemblance : Le calcul de la vraisemblance du modèle géostatistique est complexe pour de grands ensembles de données, comme celui de cette étude. Ici les paramètres β et Θ sont estimés simultanément, conduisant à l'algorithme noté Processus Gaussien avec l'algorithme de Maximum de Vraisemblance (GP_ML).
- Variogrammes spatio-temporels : Une autre façon d'estimer les paramètres est au moyen de variogrammes spatio-temporels (GP_VG) , comme ceux utilisés dans le sous chapitre précédent [11]. Les paramètres de la moyenne β sont estimés par moindres carrés ordinaires; puis, les paramètres de la fonction de covariance(Θ) sont estimés par les moindres carrés pondérés sur le variogramme expérimental [18].

Forêts aléatoires

L'algorithme des Forêts Aléatoires (RF) est une méthode d'ensemble sophis-

tiquée qui exploite les principes des arbres de décision. Il génère un ensemble d'arbres de décision à partir de sous-ensembles aléatoirement échantillonnés des données d'entraînement. La prédiction finale est obtenue en agrégeant les prédictions individuelles de ces arbres. Cette stratégie d'agrégation conduit fréquemment à des prédictions robustes et précises. Les RF sont particulièrement efficaces pour traiter des caractéristiques ayant des relations non linéaires. Toutefois, elles présentent certains inconvénients : une absence d'interprétabilité directe du modèle, un risque potentiel de sur-apprentissage, et la nécessité de déterminer préalablement le nombre d'arbres à intégrer dans l'ensemble.

L'algorithme principal pour construire des arbres de décision appelé ID3 par Quinlan et al. [19], est un algorithme fondamental pour la construction de forêts aléatoires. Il opère une recherche avide, de haut en bas, à travers les branches possibles sans revenir en arrière. Dans le contexte des arbres de décision pour la régression, l'ID3 peut opter pour la réduction de l'écart-type plutôt que le gain d'information. Cette mesure évalue la diminution de la variabilité suite à une séparation des données selon un certain attribut. L'objectif devient alors de privilégier les attributs engendrant la réduction la plus significative de cet écart-type.

XGBoost

L'algorithme XGBoost est une méthode d'optimisation basée sur le gradient qui assemble un modèle en succession, en adaptant des apprenants faibles, tels que les arbres de décision, au gradient négatif d'une fonction de perte spécifique [4]. Cette fonction mesure la divergence entre les prédictions produites et les observations réelles. À chaque itération, un nouvel apprenant est calibré sur l'erreur résiduelle de la prédiction précédente, avec pour objectif la minimisation de la perte globale. Le modèle final résulte de l'agrégation pondérée de ces apprenants, les coefficients étant établis via le processus d'optimisation.

Disponible en tant que bibliothèque open source, XGBoost fournit une plateforme pour le boosting de gradient compatible avec plusieurs langages de programmation, tels que C++, Java, Python, R et Julia. Sa conception, axée sur l'efficacité et les performances, le rend particulièrement prisé dans les compéti-

tions d'apprentissage automatique.

Machine à vecteurs de support

Cortes et al.[7] ont proposé la méthode Machine à Vecteurs de Support (SVM) pour la classification et la régression. La régression utilisée dans cette méthode peut être considérée comme une technique non paramétrique car elle repose sur des fonctions de noyau.

Dans cette étude, l'utilisation du noyau radial est mise en œuvre, et la régression SVM linéaire epsilon-insensible est appliquée. (ϵ -SVM). Dans la régression ϵ -SVM, l'objectif est de trouver une fonction $l(X)$ qui dévie de Z par une valeur pas plus grande que ϵ pour chaque point d'apprentissage X_i , et en même temps est aussi plate que possible :

$$Z_i = l(X_i) + \epsilon \quad (3.28)$$

Il faut donc trouver la fonction linéaire $l(Z_i) = \beta X_i + b$ et s'assurer qu'elle est aussi plate que possible, cela revient au problème d'optimisation convexe pour minimiser la valeur de la norme pour le vecteur de paramètre β [7].

Une constante ($C > 0$) détermine le compromis entre la planéité de l et la quantité jusqu'à laquelle des écarts supérieurs à ϵ sont tolérés.

Pour rendre l'algorithme SVM non linéaire, les données d'apprentissage X sont prétraitées en utilisant une transformation $X \rightarrow F$ dans un espace de caractéristiques F , comme détaillé dans Aizerman et al. [1] et Nilsson et al. [15]. L'algorithme standard de régression SVM est ensuite appliqué.

Réseau de neurones artificiels

Dans un réseau neuronal artificiel (ANN), une fonction de transfert traduit les signaux d'entrée en signaux de sortie. Quatre types de fonctions de transfert sont couramment utilisés : unit step (seuil), sigmoid, linéaire par morceaux, et gaussienne.

Un réseau neuronal à k couches est une fonction mathématique f , qui est une composition de fonctions multivariées : $f_1, f_2, \dots, f_k$, et g , définies comme

suit :

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^p \quad (3.29)$$

$$f(X) = g \circ f_k \circ \dots \circ f_1(X) \quad (3.30)$$

$$f_i(X) = a(w_i x + b_i) \quad (3.31)$$

Ici, n représente la dimension de l'entrée X . Chaque fonction f_i est une combinaison linéaire de l'entrée X avec ses coefficients w_i et un biais b_i

a est appelée une fonction d'activation (elle est la même pour chaque fonction intermédiaire f_i) La fonction d'activation peut prendre différentes formes, ici la fonction sigmoïde est utilisée.

3.2.2 Résultats

Cette section détaille les résultats issus de l'application de la méthodologie sur l'ensemble de données. La discussion débutera par l'examen des résultats de quelques indicateurs de performance pour les scénarios de prédiction et d'interpolation, validés à la fois sur les capteurs bas coût et les stations fixes de surveillance de la qualité de l'air. Ensuite, un tableau synthétisera les performances des différents modèles selon tous les indicateurs et scénarios. Enfin, une analyse comparative des cartes de concentration de PM générées par les dix modèles sera présentée pour le jour le plus échantillonné.

Résultats des indicateurs performances

- Validation croisée sur les données de capteurs bas coût

L'analyse des RMSE des dix modèles pour l'interpolation temporelle (figure 3.22) montre une performance globalement homogène, avec une RMSE moyenne d'environ $0,37 \log(\mu\text{g}/\text{m}^3)$. Le modèle géostatistique, basé sur les variogrammes spatio-temporels, se distingue avec la RMSE la plus faible, soit $0,087 \log(\mu\text{g}/\text{m}^3)$, tandis que le modèle SVM affiche la valeur la plus élevée à $0,733 \log(\mu\text{g}/\text{m}^3)$.

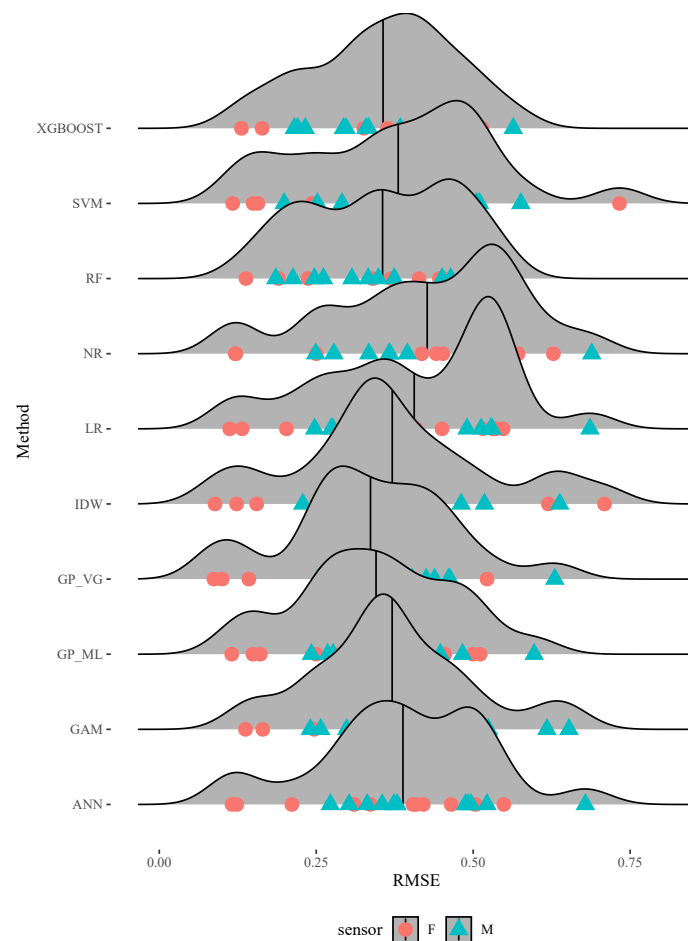


Figure 3.22 – Résultats des RMSE en validant sur les capteurs bas coût dans le premier scénario.

Il est également notable que les distributions des erreurs quadratiques moyennes pour le modèle géostatistique, estimées à la fois par la méthode du variogramme et celle du maximum de vraisemblance, présentent des similitudes notables. Cette observation s'étend à la comparaison entre le modèle de régression linéaire multiple et celui de la régression en réseau : malgré les ressemblances entre les deux approches, le modèle de régression linéaire multiple affiche des performances supérieures

En raison de la proximité de certains capteurs fixes entre eux, la prédiction

pour ces capteurs devient facile pour tous les modèles étudiés.

En outre, ces résultats confortent les conclusions tirées du chapitre 3.1 : même dans un jeu de données différent, les modèles géostatistiques semblent surpasser une simple interpolation inverse à la distance. Cette observation renforce l'intérêt des modèles géostatistiques pour nos recherches futures.

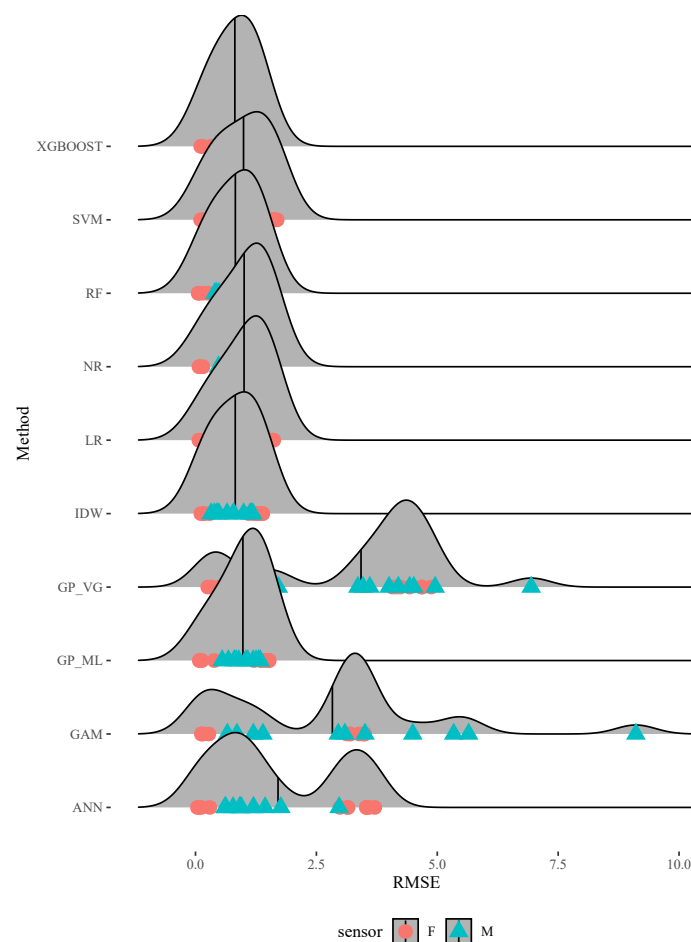


Figure 3.23 – Résultats des RMSE en validant sur les capteurs bas coût dans le deuxième scénario.

La figure 3.23 révèle une augmentation significative de l'erreur lors de la prédiction pour une journée sans données d'observation. Plus précisément, la

RMSE moyenne des dix modèles atteint $1,44 \log(\mu\text{g}/\text{m}^3)$. L'algorithme XGBOOST se distingue par la plus faible erreur, avec une RMSE minimale de $0,87 \log(\mu\text{g}/\text{m}^3)$, tandis que le maximum est attribué au modèle géostatistique avec un modèle de variogrammes spatio-temporels, affichant une RMSE de $3,42 \log(\mu\text{g}/\text{m}^3)$.

Dans le scénario de prévision, contrairement à celui de l'interpolation temporelle, les covariables jouent un rôle crucial. En effet, face à des observations distantes dans le temps, et en raison de la variabilité plus élevée des covariables associées aux capteurs mobiles, ces dernières permettent d'expliquer une plus grande part de la variance.

Contrairement au modèle de régression Linéaire et de régression en réseaux qui maintiennent une distribution d'erreur similaire, une différence significative est remarquée entre les deux méthodes d'estimation des paramètres du krigeage avec dérives externes. L'approche du maximum de vraisemblance démontre une meilleure performance, attribuable en partie à une estimation plus précise des paramètres des covariables, qui tient compte de la dépendance des observations. L'autre facteur de cette supériorité réside dans la manière dont les paramètres de la fonction de corrélation sont estimés dans le modèle du variogramme : un poids plus élevé est accordé aux mesures proches de l'origine du variogramme, ce qui améliore la prédiction des observations proches.

Il est également intéressant de noter que les modèles complexes tels que les réseaux neuronaux artificiels et les modèles additifs généralisés ne sont pas efficaces pour prévoir des valeurs pour des jours non échantillonnés ici.

- Validation sur des stations de surveillance fixes :

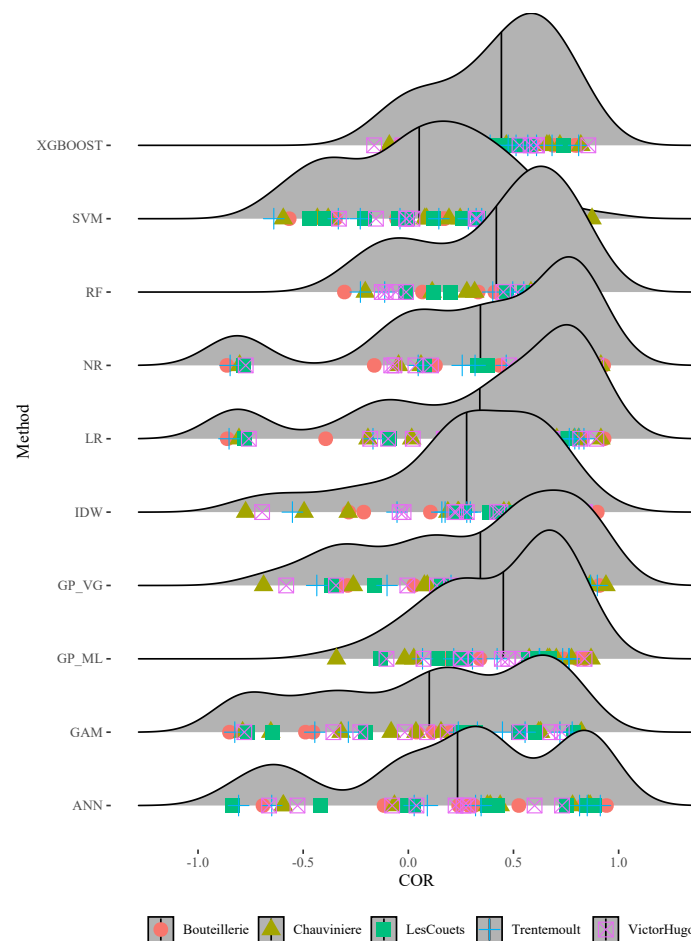


Figure 3.24 – Résultats des corrélations en validant sur les stations de surveillance fixes dans le premier scénario.

La Figure 3.24, illustrant les résultats du scénario d'interpolation temporelle validé sur stations fixes de qualité de l'air. Cela révèle une corrélation négative sur certaines stations pour l'ensemble des modèles testés. En moyenne, tous les modèles produisent un coefficient de corrélation de 0,3. Le modèle géostatistique avec estimation par maximum de vraisemblance obtient la meilleure performance, affichant une corrélation moyenne de 0,45 sur l'ensemble des stations. À l'opposé, le modèle SVM affiche la corrélation la plus faible, avec un score de 0,05.

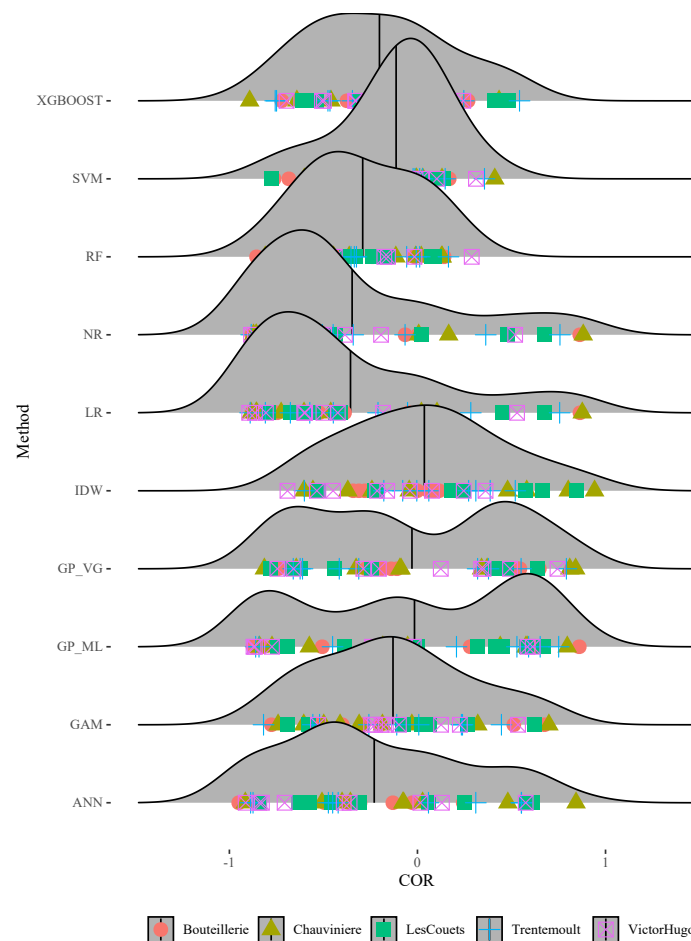


Figure 3.25 – Résultats des corrélations en validant sur les stations de surveillance fixes dans le deuxième scénario.

La même observation que pour la validation croisée sur les capteurs à faible coût est valable ici. Dès que nous nous intéressons à une prédiction, la plupart des modèles ne parviennent pas à donner de meilleurs résultats. C'est encore pire dans le cas d'une validation sur des stations fixes de qualité de l'air, où nous pouvons obtenir une corrélation moyenne négative pour tous les modèles à l'exception d'IDW. La corrélation moyenne pour tous les modèles est de -0,16.

Ces résultats, illustrés dans la figure 3.25, mettent en évidence l'incapacité des modèles à prédire efficacement les mesures collectées par les stations de référence à partir de données antérieures provenant de capteurs bas coût.

- Tableau récapitulatif :

Le tableau 3.3 présente tous les résultats des indicateurs de performance pour tous les modèles dans les deux scénarios et les deux méthodes de validation. Dans le cas de la validation croisée sur des capteurs à faible coût, tous les modèles sont à peu près équivalents et offrent d'excellents résultats, avec des RMSE équivalents, comme le montre la figure 3.22, et des corrélations équivalentes allant de 0,75 à 0,86. En outre, dans ce cas, ni le biais ni la MAE n'indiquent un modèle défectueux.

Dans le cas de la prévision par capteur à faible coût, les scores RMSE se détériore pour tous les modèles, comme le montre la figure 3.23, ainsi que pour les corrélations, qui varient désormais de 0,15 à 0,43. L'examen du biais et du MAE conclut à l'inadaptation des modèles ANN, GAM et GP_VG dans le cas d'une prévision.

Concernant la validation sur des stations fixes de qualité de l'air, les modèles RF, XGBOOST, et ML_GP offrent les meilleurs résultats sur l'ensemble des indicateurs de performance, avec des corrélations supérieures à 0,4 pour ces trois modèles. Nous notons également que les modèles GAM et ANN sont nettement moins performants.

Enfin, en ce qui concerne les prévisions des stations de qualité de l'air, l'étude des indicateurs de performance indique que tous les modèles sont inadéquats.

Dans l'ensemble, les modèles d'apprentissage automatique (XGBOOST RF et SVM), à l'exception du modèle ANN, semblent mieux adaptés au scénario de prévision.

Comme prévu, davantage d'erreurs sont commises dans le scénario de prévision, que la validation soit effectuée sur des capteurs bas coût ou des stations fixes.

Les erreurs s'avèrent nettement plus marquées lors de la validation sur les stations fixes, suggérant une possible disparité dans la nature des données recueillies par les capteurs bas coût comparativement à celles des stations fixes.

Des fondamentaux aux techniques avancées : une comparaison des modèles de pollution

Modèle	Scénario	Validation	RMSE ($\log(\mu\text{g}/\text{m}^3)$)	COR	Biais ($\log(\mu\text{g}/\text{m}^3)$)	MAE ($\log(\mu\text{g}/\text{m}^3)$)
IDW	Interpolation	Capteurs bas coût	0.371	0.826	-0.005	2.29
		Stations fixe de référence	0.760	0.277	0.0403	1.56
	Prévision	Capteurs bas coût	0.822	0.353	-0.454	3.37
		Stations fixe de référence	0.892	0.0366	-0.434	1.53
LR	Interpolation	Capteurs bas coût	0.406	0.766	0.018	2.25
		Stations fixe de référence	0.915	0.340	-0.204	1.71
	Prévision	Capteurs bas coût	1.00	0.409	-0.219	3.78
		Stations fixe de référence	1.67	-0.356	-0.513	2.70
NR	Interpolation	Capteurs bas coût	0.427	0.759	0.023	2.32
		Stations fixe de référence	0.990	0.343	-0.239	1.73
	Prévision	Capteurs bas coût	1.01	0.411	-0.229	3.77
		Stations fixe de référence	1.69	-0.347	-0.502	2.75
GAM	Interpolation	Capteurs bas coût	0.371	0.840	-0.066	2.22
		Stations fixe de référence	4.58	0.0996	-2.72	8.74
	Prévision	Capteurs bas coût	2.83	0.243	0.886	11.9
		Stations fixe de référence	3.43	-0.130	-1.126	6.39
GSTAT	Interpolation	Capteurs bas coût	0.336	0.840	0.007	2.18
		Stations fixe de référence	1.18	0.343	-0.456	2.58
	Prévision	Capteurs bas coût	3.42	0.201	0.0242	11.8
		Stations fixe de référence	3.62	-0.0293	-0.518	6.71
GPGP	Interpolation	Capteurs bas coût	0.345	0.810	0.013	2.13
		Stations fixe de référence	0.863	0.452	-0.193	1.60
	Prévision	Capteurs bas coût	0.980	0.368	-0.356	3.08
		Stations fixe de référence	1.17	-0.015	-0.642	1.95
RF	Interpolation	Capteurs bas coût	0.356	0.852	-0.097	2.22
		Stations fixe de référence	0.677	0.419	-0.103	1.31
	Prévision	Capteurs bas coût	0.824	0.409	-0.362	3.30
		Stations fixe de référence	0.873	-0.291	-0.299	1.55
ANN	Interpolation	Capteurs bas coût	0.388	0.756	-0.0574	2.32
		Stations fixe de référence	1.65	0.234	-0.634	2.84
	Prévision	Capteurs bas coût	1.71	0.152	-0.626	9.11
		Stations fixe de référence	2.01	-0.230	-1.33	4.06
SVM	Interpolation	Capteurs bas coût	0.381	0.802	-0.058	2.26
		Stations fixe de référence	0.750	0.0518	-0.125	1.38
	Prévision	Capteurs bas coût	0.993	0.351	-0.555	3.65
		Stations fixe de référence	0.817	-0.113	-0.339	1.53
XGBOOST	Interpolation	Capteurs bas coût	0.356	0.861	-0.0894	2.27
		Stations fixe de référence	0.709	0.443	-0.113	1.47
	Prévision	Capteurs bas coût	0.814	0.429	-0.357	3.34
		Stations fixe de référence	0.876	-0.202	-0.290	1.60

Table 3.3 – Tableau récapitulatif des résultats des différents indicateurs de performance sur les deux modes de validation et dans les deux scénarios, pour les dix modèles étudiés. L'écart type des logarithmes des concentrations de PM₁₀ étudiés est de 1.03.

Cartographies

La figure 3.26 affiche les cartes générées par les dix modèles différents à 12h00 le 26/11/2018, en utilisant toutes les données de ce jour. Ce jour a été choisi car il contenait la plus grande quantité de données, avoisinant les 20 000 observations.

Comme prévu, les cartes produites par les modèles de régression linéaire et de régression en réseau sont presque indiscernables. Nous remarquons également des similitudes entre les deux modèles géostatistiques, mais les différences deviennent plus évidentes vers les bords de la carte à mesure que l'on s'éloigne des observations.

Pour l'ensemble des modèles, à l'exception de l'IDW, nous remarquons une tendance de concentration des polluants autour du centre et de la périphérie routière de la ville de Nantes. En parallèle, des concentrations plus faibles sont observées sur l'île de Nantes, où sont localisés la plupart des capteurs bas coût fixes.

Il est crucial de souligner que ces cartographies illustrent uniquement la dispersion spatiale des concentrations à un instant donné de la journée et ne capturent pas leur variation temporelle. En dépit de cela, ces visualisations restent précieuses pour appréhender les comportements généraux et les tendances spatiales des concentrations de polluants.

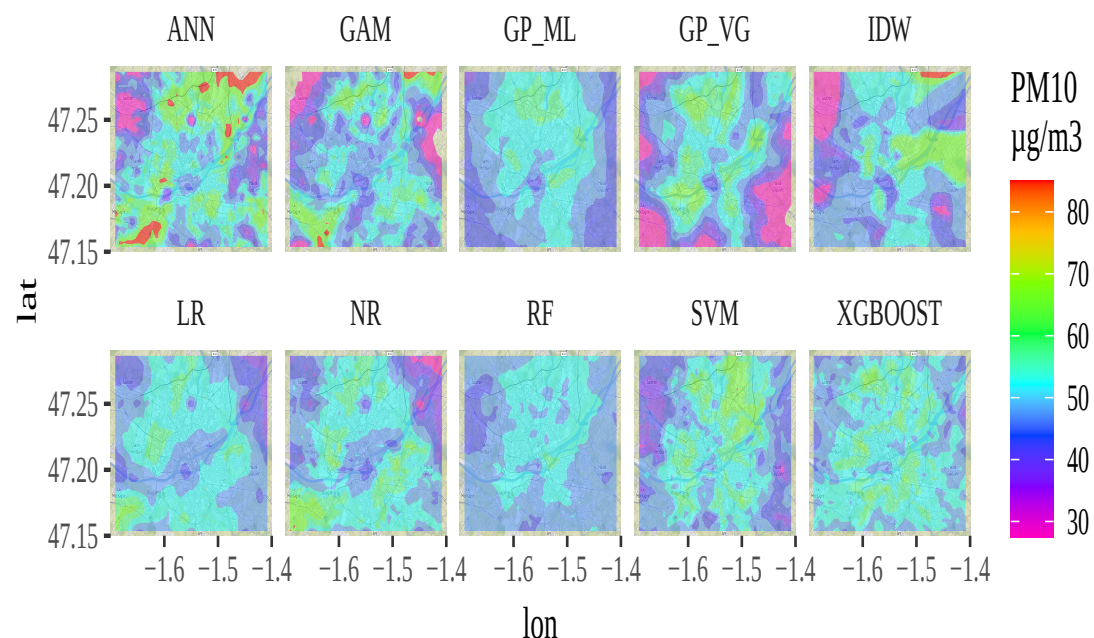


Figure 3.26 – Cartographie comparative des concentrations PM10 selon différents modèles

3.2.3 Discussion

D'après les résultats de cette étude, il est clair que la qualité des données provenant des capteurs bas coût est relativement faible et ne reflète pas nécessairement les tendances captées par les stations fixes de surveillance de la qualité de l'air. À l'avenir, nous devons soit adapter nos modèles pour tenir compte de la nature à bas coût de ces capteurs, soit ajuster notre stratégie d'échantillonnage en combinant l'utilisation de capteurs à bas coût avec celle de capteurs plus performants et fiables.

L'approche adoptée pour estimer les paramètres du même modèle peut avoir un impact considérable sur les résultats obtenus. En particulier, le modèle de géostatistique basé sur le variogramme spatio-temporel se concentre sur l'estimation des paramètres à l'origine du variogramme, privilégiant ainsi les prédictions pour des lieux proches des observations en termes d'espace et de temps.

Dans le scénario de prévision étudié, les méthodes explorant des relations complexes et non linéaires avec les covariables sont à privilégier. Lorsqu'on tente de prédire loin des observations, notamment dans un contexte où la variabilité temporelle prédomine sur la variabilité spatiale, l'influence des observations avoisinantes devient négligeable. Par conséquent, la prédiction repose principalement sur les covariables. C'est un domaine où les modèles d'apprentissage automatique se montrent particulièrement performants.

Limites de l'étude :

-La recherche a débuté par l'exploitation de versions simplifiées des divers modèles, avec la conscience du potentiel d'accroître leur complexité. Par exemple, le modèle GAM pourrait être enrichi par l'intégration de relations plus élaborées, le modèle géostatistique pourrait incorporer l'anisotropie, et les réseaux neuronaux pourraient bénéficier de l'ajout de couches additionnelles. Cette approche épurée visait à établir un étalon de performance pour les différents modèles.

-Le type de capteur bas coût utilisé, mobile ou fixe, a indéniablement influencé la couverture des différentes formes de pollution urbaine. Du fait de leur caractère d'échantillonnage opportuniste, les capteurs mobiles tendent à échantillonner principalement les axes de circulation majeurs, capturant de ce fait surtout la pollution liée au trafic. Par conséquent, ils peuvent passer à côté de la pollution de fond habituellement détectée par certains capteurs fixes, qu'ils soient de bas coût ou des stations de référence.

-Les données issues des capteurs bas coût ont été traitées comme étant indépendantes, conditionnellement à la concentration de PM. Cela signifie que l'étude n'a pas pris en compte l'autocorrélation des erreurs des données issues d'un même capteur.

Capteur bas coût fixe vs mobile :

-L'exploration de la relation entre le ratio de capteurs bas coût fixes et mobiles utilisés et la précision de la prédiction représente une tâche complexe. Bien que les capteurs mobiles fournissent des données plus abondantes, ces derniers sont plus susceptibles de subir des pannes, ce qui rend difficile une comparaison numérique objective.

L'estimation des paramètres de la tendance globale des modèles, c'est-à-dire l'impact des différentes covariables sur la concentration des polluants, est optimisée par l'utilisation de capteurs mobiles. Par exemple, si tous les capteurs fixes avaient été positionnés sur des voies similaires à des autoroutes, l'effet de la catégorie de rue sur la concentration des polluants n'aurait pas été discernable. Plus largement, étant donné que les covariables spatiales des capteurs fixes demeurent constantes, l'absence de capteurs mobiles pourrait nous confronter à des défis d'identifiabilité des paramètres relatifs à ces covariables spatiales [9].

Il est nécessaire de différencier deux aspects : pour l'estimation des paramètres, l'utilisation de capteurs mobiles est privilégiée, car ils offrent une cou-

verture spatiale plus vaste sans nécessairement réduire la couverture temporelle par rapport aux capteurs fixes. En ce qui concerne la précision de la prédiction, c'est-à-dire la minimisation des erreurs de prédiction après la phase d'estimation des paramètres, la réponse reste incertaine, même si les capteurs fixes semblent avoir l'avantage, car ils demandent moins d'efforts pour fonctionner en continu.

Dans l'ensemble, le choix entre un capteur mobile ou fixe est complexe et dépend grandement à la fois du capteur et du modèle utilisés, ainsi que de leur application spécifique.

3.3 Conclusion

Au cours de cette étude, Différents modèles de prédiction de la qualité de l'air ont été explorés et comparés, allant des méthodes de régression et d'interpolation spatiale les plus simples à des techniques plus avancées telles que les modèles géostatistiques et les méthodes d'apprentissage automatique.

Cette étude met en évidence un défi majeur : la qualité des données provenant des capteurs à bas coût. Ces données se sont révélées moins fiables que celles des stations fixes de surveillance de la qualité de l'air.

Cette observation permet de répondre, au moins partiellement, à notre question de recherche initiale. À la lumière des résultats actuels, il semblerait que les capteurs mobiles bas coût ne soient pas encore suffisamment précis pour remplacer les stations fixes de surveillance de l'air.

Cela suggère qu'il est nécessaire d'adapter notre modèle à la nature de ces données de basse qualité. Cette adaptation pourrait impliquer la création d'un nouveau modèle qui prendrait en compte les spécificités de ces capteurs à bas coût, une direction que nous explorerons dans le chapitre suivant.

Dans le contexte de prévisions à long terme, il est recommandé d'adopter des méthodes d'apprentissage automatique qui se penchent sur l'exploration de relations complexes et non linéaires avec les covariables. L'essor de ces modèles avancés et le développement continu de l'intelligence artificielle suggèrent un avenir prometteur pour l'amélioration de la précision prédictive dans de tels scénarios.

L'analyse comparative durant ce chapitre révèle que le modèle géostatistique utilisant le maximum de vraisemblance se démarque des autres. Non seulement il produit des résultats comparables aux méthodes d'apprentissage automatique tout en offrant un aspect descriptif que ces dernières ne fournissent généralement pas, mais il surpasse aussi les modèles de régression et d'interpolation spatiale plus simples.

En plus de son caractère descriptif, le modèle de krigeage avec dérives externes fournit une mesure d'incertitude associée à chaque prédiction, un atout

précieux pour évaluer la fiabilité des prévisions et qui peut être utilisé pour optimiser le placement des capteurs afin de minimiser l'erreur de prédiction. De plus, il peut facilement intégrer des données auxiliaires, améliorant ainsi encore la précision des prédictions. Il répond ainsi à de nombreuses questions liées à notre application.

Cependant, ce modèle a un inconvénient majeur : sa complexité qui est de l'ordre de $O(n^3)$. Ce niveau de sophistication rend le modèle plus difficile à mettre en œuvre, une problématique que nous explorerons plus en détail dans le chapitre suivant de cette thèse.

En conclusion, le chapitre présent a démontré avec succès le développement de modèles qui intègrent les mesures de qualité de l'air obtenues à l'aide de capteurs à bas coût déployés à travers la ville. Ces modèles ont permis de générer des cartographies de la pollution atmosphérique d'une précision supérieure à celles actuellement disponibles, réalisant ainsi l'objectif principal de cette thèse. Il est particulièrement remarquable que ces avancées aient été accomplies sans accès à des données cruciales telles que le trafic routier, souvent considérées comme déterminantes dans l'évaluation de la qualité de l'air.

Références

- [1] MA Aizerman, E Braverman et LI Rozonoer. "Probability problem of pattern recognition learning and potential functions method". In : *Avtomat. i Telemekh* 25.9 (1964), p. 1307-1323.
- [2] Michel Arnaud et Xavier Emery. *Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatistiques*. Hermès, 2000.
- [3] Franz Aurenhammer. "Voronoi diagrams—a survey of a fundamental geometric data structure". In : *ACM Computing Surveys (CSUR)* 23.3 (1991), p. 345-405.
- [4] Tianqi Chen et Carlos Guestrin. "Xgboost : A scalable tree boosting system". In : *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, p. 785-794.
- [5] Jean-Paul Chiles et Pierre Delfiner. *Geostatistics : modeling spatial uncertainty*. T. 497. John Wiley & Sons, 2009.
- [6] Copernicus. *Analyse et provision du modele CHIMERE*. Consulté le 1 Janvier 2022. 2022. url : <https://www.copernicus.eu/fr/acces-aux-donnees/plateformes-classiques-dacces-aux-donnees>.
- [7] Corinna Cortes et Vladimir Vapnik. "Support-vector networks". In : *Machine learning* 20.3 (1995), p. 273-297.
- [8] Alicia Gressent et al. "Data fusion for air quality mapping using low-cost sensor observations : Feasibility and added-value". In : *Environment international* 143 (2020), p. 105965.
- [9] Joseph HA Guillaume et al. "Introductory overview of identifiability analysis : A guide to evaluating whether you have the right type of data for your modeling purpose". In : *Environmental Modelling & Software* 119 (2019), p. 418-432.
- [10] David Hasenfratz et al. "Deriving high-resolution urban air pollution maps using mobile sensor nodes". In : *Pervasive and Mobile Computing* 16 (2015), p. 268-285.

- [11] Yacine Mohamed Idir et al. "Mapping urban air quality from mobile sensors using spatio-temporal geostatistics". In : *Sensors* 21.14 (2021), p. 4717.
- [12] Daniel G Krige. "A statistical approach to some basic mine valuation problems on the Witwatersrand". In : *Journal of the Southern African Institute of Mining and Metallurgy* 52.6 (1951), p. 119-139.
- [13] Data Air Pays de la Loire. *Mesures des concentrations dans l'air de PM₁₀*. Consulté le 1 janvier 2023. 2023. url : <https://data.airpl.org/dataset/mesures>.
- [14] Ali Marjovi, Adrian Arfire et Alcherio Martinoli. "High resolution air pollution maps in urban environments using mobile sensor networks". In : *2015 International Conference on Distributed Computing in Sensor Systems*. IEEE. 2015, p. 11-20.
- [15] Nils J Nilsson. "Learning machines." In : (1965).
- [16] Opendatasoft. *Mesures des de la temperature et de l'humidité dans la ville de Nantes*. Consulté le 1 janvier 2022. 2022. url : <https://public.opendatasoft.com>.
- [17] OpenStreetMap. *Covaraibles spatiales extraites pour la ville de Nantes*. Consulté le 1 janvier 2022. 2022. url : <https://www.openstreetmap.fr/>.
- [18] Edzer Pebesma et Gerard Heuvelink. "Spatio-temporal interpolation using gstat". In : *RFID Journal* 8.1 (2016), p. 204-218.
- [19] J. Ross Quinlan. "Induction of decision trees". In : *Machine learning* 1 (1986), p. 81-106.
- [20] Donald Shepard. "A two-dimensional interpolation function for irregularly-spaced data". In : *Proceedings of the 1968 23rd ACM national conference*. 1968, p. 517-524.
- [21] Ciyou Zhu et al. "Algorithm 778 : L-BFGS-B : Fortran subroutines for large-scale bound-constrained optimization". In : *ACM Transactions on mathematical software (TOMS)* 23.4 (1997), p. 550-560.

Chapitre 4

Affinement du modèle géostatistique : autocorrélation, approximation et application

Points clés du chapitre

- Proposition d'un nouveau modèle hiérarchique à trois niveaux.
- Étude de l'approximation de Vecchia à ce nouveau modèle proposé : supériorité de l'arrangement utilisant l'algorithme min-max distance et l'arrangement aléatoire.
- Prise en compte de l'autocorrélation des capteurs comme source de biais.

Contenu du chapitre

4.1	Définition du modèle géostatistique à partir de l'approche hiérarchique	141
-----	---	-----

4.2	Prise en compte et modélisation de l'autocorrélation des capteurs bas coût	144
4.3	Méthodes d'approximation du processus gaussien	149
4.4	Étude et application de l'approximation de Vecchia au nouveau modèle proposé	150
4.4.1	Méthodologie	151
4.4.2	Résultats d'évaluation	160
4.4.3	Application numérique	163
4.4.4	Discussion	168
4.5	Conclusion	169
	Références	171

Au chapitre précédent, la conclusion a porté sur la pertinence du modèle géostatistique utilisant le maximum de vraisemblance en comparaison avec d'autres méthodes de prédiction de la qualité de l'air, en réponse aux enjeux de la cartographie de la pollution atmosphérique. Cependant, une faiblesse clé dans ce modèle a également été identifiée : sa complexité de calcul. De plus, nous avons mis en évidence la nécessité d'adapter notre modèle à la nature des données provenant des capteurs bas coût. Dans ce chapitre, ces deux défis sont abordés.

Tout d'abord, une nouvelle définition du modèle géostatistique par le biais d'une approche hiérarchique est introduite. Cette approche sera féconde afin d'adapter ce dernier à notre problématique. L'idée est de construire un modèle plus robuste, capable de mieux gérer les données de faible qualité issues des capteurs à bas coût.

L'exploration débutera par la question de l'autocorrélation des capteurs à bas coût. En prenant en compte ces effets, l'objectif est d'améliorer encore la précision des prédictions.

Ensuite, face à la complexité de calcul du krigeage avec dérivées externes, une solution possible réside dans l'application de techniques d'approximation

du maximum de vraisemblance. L'étude détaille l'une des méthodes les plus populaires, l'approximation de Vecchia, se fera en l'adaptant à la problématique spécifique et en évaluant ses performances sur un jeu de données simulé à partir de trajectoires réelles.

Enfin, l'application de cette méthode d'approximation au nouveau modèle sur le jeu de données d'Atmotrack permettra de proposer une cartographie détaillée de la pollution atmosphérique.

Commençons donc par redéfinir notre modèle géostatistique dans une perspective hiérarchique.

4.1 Définition du modèle géostatistique à partir de l'approche hiérarchique

Soit $Z(x) = (z_1(x_1), z_2(x_2), \dots, z_n(x_n))$ les données observées par nos capteurs bas coût, indexées sur l'espace temps x (une seule variable est utilisée au lieu de trois (x, y, t) pour alléger la notation). Le vecteur des concentrations de polluants (appelé vrai signal), aux mêmes endroits est $Y = (y_1(x_1), y_2(x_2), \dots, y_n(x_n))$. La différence entre Y et Z correspond à l'erreur des capteurs.

Supposons que le processus d'intérêt (le vrai signal) suit une loi normale multivariée avec une moyenne μ et une matrice de covariance $\sigma^2 R$, où R est une matrice $n \times n$ avec les éléments $r_{i,j} = \rho(x_i, x_j)$. Cela donne :

$$Y \sim N(\mu, \sigma^2 R) \quad (4.1)$$

La moyenne μ peut être connue à l'avance, fixe dans l'espace-temps mais inconnue, ou même varier dans l'espace-temps.

Ici, $\sigma^2 R$ est la matrice de covariance qui capture la corrélation spatio-temporelle de la pollution atmosphérique (qui peut être isotrope ou non). C'est elle qui « encode » la dispersion des polluants. La matrice de covariance peut être paramétrée à l'aide d'une fonction telle que la fonction de covariance de Matérn ou exponentielle.

Dans le modèle géostatistique classique, l'erreur du capteur (ou le modèle de données) est généralement modélisée par un bruit blanc décorrélé. Cela signifie que les erreurs des capteurs sont supposées être indépendantes et identiquement distribuées entre les lieux, les points temporels et les capteurs. De plus, elles sont centrées sur le vrai signal. Il en résulte le modèle largement étudié [3] (4.2) :

$$\begin{aligned} Z/Y &\sim N(Y, \tau^2 \cdot I) \\ Y &\sim N(\mu, \sigma^2 R) \end{aligned} \quad (4.2)$$

τ^2 représente l'erreur de mesure (ou pépité) et I est la matrice identité.

Finalement, nous pouvons obtenir la distribution marginale de la variable de données Z :

$$p(Z) = \int p(Z/Y) \cdot p(Y) dY \quad (4.3)$$

Cela résulte en une forme compacte du modèle (4.4) :

$$Z \sim N(\mu, \tau^2 \cdot I + \sigma^2 R) \quad (4.4)$$

Supposant que l'objectif réside dans la prédiction de la valeur des polluants en un lieu non échantillonné Y^* à partir des données connues Z .

Dans ce cas, le vecteur (Y^*, Z) suit également une loi normale multivariée, et l'estimateur d'erreur quadratique moyenne minimale $\hat{Y}^*$ est obtenu en utilisant le résultat standard de la distribution gaussienne multivariée.

THEOREME : Soit $X = (X1, X2)$ une variable multivariée gaussienne, avec un vecteur de moyenne $\mu = (\mu_1, \mu_2)$ et une matrice de covariance

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

C'est-à-dire, $X \sim N(\mu, \Sigma)$. Alors, la distribution conditionnelle de $X1$ connaissant $X2$ est également multivariée normale, $X1 | X2 \sim N(\mu_{1|2}, \Sigma_{1|2})$, où

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (X2 - \mu_2) \quad (4.5)$$

et

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \quad (4.6)$$

Pour appliquer ce théorème à notre problème de prédiction, notons que (Y^*, Z) est multivariée gaussienne avec le vecteur de moyenne $\mu = (\mu_Y, \mu_Z)$ et la matrice de covariance :

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2 r \\ \sigma^2 r' & \sigma^2 R + \tau^2 I \end{pmatrix}$$

où r est le vecteur avec les éléments $r_i = \rho(x^*, x_i)$ pour $i = 1 \dots n$ codant la

covariance entre le point non échantillonné et les données.

Dans ce cas, le théorème donne l'estimation avec l'erreur quadratique moyenne minimale :

$$\hat{Y}^* = \mu_Y + r \cdot (R + \tau^2 / \sigma^2 \cdot I)^{-1} (Z - \mu_Z) \quad (4.7)$$

Dans la terminologie géostatistique traditionnelle, la prédiction de la valeur à des endroits non échantillonnés, donnée par 4.7, est appelée le krigeage. Dans l'hypothèse prise où les paramètres sont connus, l'estimateur est linéaire par rapport aux données. Pour utiliser l'estimateur en pratique, nous devons insérer des valeurs estimées pour les paramètres du modèle.

Si nous connaissons la valeur de μ à l'avance sur tout le domaine, on parle de krigeage simple. La distinction entre le krigeage simple et le krigeage avec dérivées externes réside dans le fait que, dans ce dernier, la valeur moyenne est considérée comme une combinaison linéaire de certaines covariables X . Cela conduit à un estimateur similaire à 4.7 à l'exception que μ est remplacé par son estimateur des moindres carrés généralisés :

$$\hat{\mu} = \hat{\beta}X = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{Y} \quad (4.8)$$

Dans notre dérivation du krigeage simple et avec dérivées externes, nous commençons par un modèle stochastique pour le processus d'intérêt, Y et les données Z puis nous dérivons la forme explicite de l'estimateur d'erreur quadratique moyenne minimale pour $\hat{Y}^*$. Dans cette approche, le fait que l'estimateur soit linéaire en Z est une conséquence de l'hypothèse de modélisation gaussienne. En géostatistique classique, le point de départ est de restreindre l'attention aux estimateurs qui sont linéaires en Z , et de chercher celui qui est optimal, au sens de l'erreur quadratique moyenne. Les expressions résultantes pour les estimateurs de krigeage simple et avec dérivées externes vu au chapitre précédent (équations :3.9,3.19) peuvent également être dérivées comme des exemples des idées statistiques beaucoup plus anciennes de meilleure prédiction linéaire, et de meilleure prédiction linéaire non biaisée, respectivement. Ceci est le cas, par exemple, dans le chapitre 1 de l'ouvrage de Stein (1999) [25].

4.2 Prise en compte et modélisation de l'autocorrélation des capteurs bas coût

Il est rappelé le modèle vu précédemment, cette fois-ci avec une moyenne qui varie dans l'espace-temps :

$$\begin{aligned} Z/Y &\sim N(Y, \tau^2 \cdot I) \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \end{aligned} \tag{4.9}$$

La moyenne ici est modélisée comme une combinaison linéaire de p covariables. Ainsi, la matrice X est composée de $p + 1$ colonnes et n lignes, avec une colonne supplémentaire constituée uniquement de 1 pour représenter la constante. Le vecteur β renferme $p + 1$ paramètres, qui doivent être estimés conjointement avec les paramètres de la fonction de covariance.

C'est un modèle hiérarchique à deux niveaux :

- Le premier niveau encode le comportement des capteurs (l'erreur, l'autocorrélation).
- Le deuxième encode le comportement des polluants, où la moyenne dépend linéairement de plusieurs covariables (météo, proximité d'un espace vert...). La diffusion des polluant est codée elle par la fonction de covariance.

Dans cette configuration hiérarchique, on constate que les hypothèses sur le modèle de données peuvent ne pas être appropriées pour des données collectées par plusieurs capteurs à faible coût. Ces derniers sont susceptibles d'être affectés par différentes erreurs de mesure, potentiellement corrélées spatialement et temporellement.

Les capteurs à faible coût utilisés pour mesurer la pollution atmosphérique posent des problèmes spécifiques en matière de qualité des données. L'étude de Maag et al. [20] met en évidence plusieurs erreurs typiques pouvant compromettre la précision et la fiabilité des mesures (Figure 4.1). De plus, notre analyse du chapitre précédent souligne une divergence entre ces données et celles ob-

tenues par les stations fixes dédiées à la surveillance de la qualité de l'air.

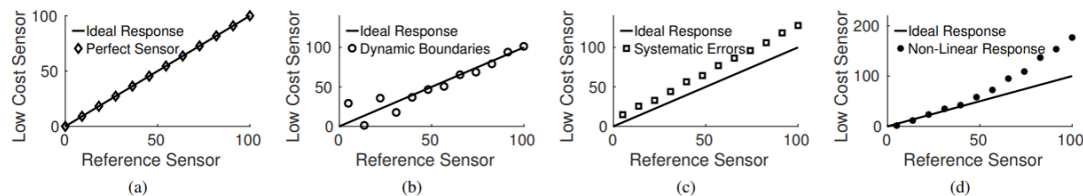


Figure 4.1 – Représentation des erreurs classiques des capteurs bas coût. Tiré de l'article : Maag et. al (2018). A survey on sensor calibration in air pollution monitoring deployments

D'autres types d'erreurs peuvent être présents dans les capteurs bas coût ou peuvent se révéler après quelques heures d'utilisation :

- **Différences intra-modèles** : Il n'est pas rare d'observer des différences de performance entre des unités d'un même modèle de capteur. Ceci peut être dû à des facteurs tels que la qualité de fabrication ou des variations mineures dans les composants utilisés.
- **Drift** : Le phénomène de « Drift » se réfère à l'incapacité des capteurs à faible coût à maintenir une performance de mesure stable sur une longue période. Ce facteur peut conduire à des erreurs croissantes avec le temps.
- **Effets environnementaux** : Certaines propriétés physiques ambiantes, notamment les conditions de température et d'humidité, peuvent avoir un effet significatif sur la réponse d'un capteur. Les capteurs peuvent ne pas réagir de la même manière dans des conditions environnementales différentes, ce qui peut biaiser les mesures.

Il est essentiel de prendre en compte ces erreurs lors de l'élaboration de modèles pour modéliser au mieux la pollution atmosphérique. Le modèle hiérarchique introduit précédemment offre un cadre idéal pour modéliser et gérer ces sources d'erreurs. Par exemple, il serait possible de modéliser des limites dynamiques, comme présentées dans la figure 4.1, en intégrant dans la variance du modèle des données une composante proportionnelle à la variable d'intérêt, Y .

Toutefois, il convient de mettre l'accent sur un point crucial : bien que le modèle puisse traiter efficacement certains types d'erreurs de mesure, la calibra-

tion des capteurs reste un passage incontournable. Il existe des erreurs que la modélisation ne peut pas couvrir sans rendre les paramètres du modèle non identifiables. Un ensemble de données biaisé à cause d'une calibration inadéquate engendrera inévitablement des prédictions faussées, quelle que soit la sophistication du modèle.

Autocorrélation des capteurs

Lors de la mise en route des capteurs mobiles, il peut arriver que certains d'entre eux dysfonctionnent pour diverses raisons, telles que des problèmes techniques ou un blocage de l'entrée d'air. Ces dysfonctionnements peuvent engendrer des erreurs dans l'ensemble des mesures de l'un de ces capteurs, alors que les données des autres capteurs demeurent relativement fiables. Dans de tels scénarios, les mesures erronées d'un capteur défaillant peuvent conduire à une surestimation de τ^2 , diminuant ainsi la pertinence des mesures de l'ensemble des capteurs.

Pour tenir compte de ce type de biais, nous utilisons le cadre des modèles mixtes [8]. Cela permet au modèle de capturer la variation et le biais des erreurs à travers les points spatio-temporels pour chaque capteur, qui peuvent être affectés par divers facteurs tels que l'étalonnage du capteur, la durée de vie de la batterie et les conditions environnementales. Le cadre du modèle mixte permet également d'incorporer des covariables supplémentaires susceptibles d'affecter les capteurs, telles que la température, l'humidité ou l'emplacement du capteur.

Les modèles mixtes, également connus sous le nom de modèles à effets mixtes ou de modèles multiniveaux, se sont révélés très utiles dans divers domaines scientifiques. Leur principale application réside dans l'analyse de données présentant des structures hiérarchiques ou emboîtées complexes, pour lesquelles les méthodes statistiques traditionnelles peuvent ne pas être efficaces.

Ainsi Pour chaque capteur $j : 1, \dots, J$:

$$\begin{aligned} Z_i / (Y, a_j) &\sim N(a_j + Y, \tau^2 \cdot I) \\ Y &\sim N(\beta X, \sigma^2 R) \\ a_j &\sim N(0, \gamma^2) \quad \text{Pour } j : 1 \dots J \end{aligned} \quad (4.10)$$

Le modèle précédent ne prédit qu'un biais constant par capteur (a_j pour le capteur j). La pente ne variant pas pour chaque capteur (le biais reste constant dans le temps). Ici, la moyenne des décalages (ou biais) des capteurs est nulle et leur variance est γ^2 . Bien que plus réaliste, l'hypothèse d'un biais moyen non nul conduit à un modèle non identifiable.

Enfin, nous pouvons obtenir la distribution marginale de la variable de données Z :

$$p(Z) = \int \int p(Z / (a_j, Y)) \cdot p(Y) \cdot p(a_j) dY da_j \quad (4.11)$$

Il en résulte une forme compacte du modèle (4.10) :

$$Z \sim N(\beta X, \tau^2 \cdot I + \sigma^2 \cdot R(\Theta) + K \cdot \gamma^2 \cdot K') \quad (4.12)$$

où K est la matrice binaire $n \times J$ qui code l'appartenance de l'observation i au capteur j , et K' sa transposée.

Après intégration, le décalage entre les capteurs se traduit dans le modèle final par une « corrélation intra-capteur » : toutes les paires d'observation provenant des mêmes capteurs auront un paramètre de covariance supplémentaire γ dans la matrice de variance-covariance, quelle que soit leur distance spatio-temporelle.

Si un biais moyen non nul b est introduit (ou un biais qui dépend de certaines covariables), nous aurions obtenu le modèle suivant :

$$Z \sim N(\beta X + b, \tau^2 \cdot I + \sigma^2 \cdot R(\Theta) + K \cdot \gamma^2 \cdot K') \quad (4.13)$$

Le modèle (4.13) aura la même vraisemblance pour chaque constante $\beta_0 + b$.

Cela valide la définition de la non-identifiabilité statistique des paramètres du modèle [9].

Cependant, le modèle (4.10) est identifiable et plus réaliste que le modèle géostatistique classique. Ses paramètres peuvent être estimés par maximum de vraisemblance :

$$L(\Theta, \beta, \tau, \gamma|Z) = \operatorname{argmax}(z, \Theta, \tau, \gamma) \quad (4.14)$$

$$L(\Theta, \beta, \tau, \gamma|Z) = \arg \max [(2\pi)^{-\frac{n}{2}} |\tau I + \sigma^2 \cdot R(\Theta) + K\gamma^2 K'|^{-\frac{1}{2}} \\ \times \exp \left(-\frac{1}{2} (Z - \beta X)' (\tau^2 I + \sigma^2 \cdot R(\Theta) + K\gamma^2 K')^{-1} (Z - \beta X) \right)]$$

Après la formulation du modèle, l'étape suivante consiste à estimer ses paramètres. Cela permettra d'utiliser l'espérance conditionnelle pour prédire les niveaux de pollution atmosphérique dans des zones non observées, en se référant aux données observées par les capteurs.

Bien que l'estimation des paramètres du modèle proposé (4.10) soit théoriquement possible par maximum de vraisemblance, comme mentionné ci-dessus, cette approche devient très difficile à mettre en œuvre lorsqu'il s'agit de grands ensembles de données atteignant plusieurs milliers ou centaines de milliers d'observations, d'autant plus lors de l'ajout d'un nouveau paramètre supplémentaire dans le modèle.

4.3 Méthodes d'approximation du processus gaussien

Les processus gaussiens (GP) sont devenus un outil largement utilisé pour modéliser une variété de phénomènes. Offrant une flexibilité remarquable [24], les GP offrent un cadre idéal pour la modélisation des données avec des non-linéarités arbitraires et des dépendances complexes. Grâce à une structure de covariance dédiée, ils capturent l'autocorrélation des observations, allant des champs spatiaux et spatio-temporels [5], jusqu'aux modèles psychologiques [27].

En conséquence, ils trouvent de nombreuses applications dans les domaines de la géostatistique, de l'apprentissage automatique et des modèles mixtes [2, 29, 24]. Ils présentent de nombreux avantages, tels que leur capacité à fournir des distributions prédictives complètes ainsi que des estimations ponctuelles.

La complexité des processus gaussiens représente un défi majeur. Par rapport au nombre d'observations, la complexité des GP évolue de manière quadratique pour la mémoire et de manière cubique pour le temps. Cela limite leur application directe aux grands ensembles de données actuels, tels que ceux dérivés de l'échantillonnage à haute fréquence de capteurs à faible coût.

Pour résoudre ce problème « big n problem », de nombreuses approches ont été proposées dans la littérature. Ces études portent sur les matrices de covariance creuse [7, 18, 6], à des matrices de covariance inverse creuse (i.e matrice de précision) [23, 19, 21], et les matrices de rang inférieur [13, 28, 22, 1, 4, 15]. La décomposition de Cholesky creuse de la matrice de précision [26, 16] sera l'objet de cette étude. Un examen complet et une comparaison de plusieurs de ces méthodes sont présentés dans l'étude de Heaton et al.[12].

Il est également important de noter que, bien que les méthodes d'approximation puissent réduire la complexité computationnelle, elles introduisent des erreurs d'approximation. Par conséquent, il est essentiel d'évaluer soigneuse-

ment l'efficacité et la précision de ces méthodes dans ce scénario d'application spécifique.

En conclusion, bien que les processus gaussiens constituent un outil analytique robuste pour la modélisation de données complexes, leur application à des ensembles de données volumineux reste un défi qui exige des méthodes d'approximation performantes. L'avancement de la recherche dans ce domaine est essentiel et se poursuit avec l'objectif de perfectionner les techniques d'approximation, afin de rendre les processus gaussiens encore plus efficaces et précis pour l'analyse de grandes quantités de données.

4.4 Étude et application de l'approximation de Vecchia au nouveau modèle proposé

Cette section reprend le modèle hiérarchique à trois niveaux développé pour modéliser avec précision le phénomène physique, tout en intégrant les erreurs associées à l'utilisation de capteurs économiques. La recherche présentée explore l'application de l'approximation de Vecchia à la fonction de vraisemblance pour ce nouveau modèle spatio-temporel.

Diverses configurations de l'approximation de Vecchia sont étudiées, notamment en termes d'ordonnancement, de définitions de distance et de conditionnement basé sur le type de capteur. Ces configurations sont évaluées pour l'approximation de la vraisemblance d'un processus gaussien spatiotemporel, utilisant des données simulées basées sur des trajectoires réelles de capteurs du jeu de données d'Atmotrack.

La dernière partie de cette étude sera consacrée à une analyse numérique réussie, dans laquelle le modèle proposé et l'approximation de Vecchia choisie sont appliqués aux données de pollution d'Atmotrack pour la cartographie et la prévision de la pollution, démontrant ainsi son applicabilité pratique. Enfin, le gain potentiel par rapport à un modèle géostatistique classique sera également

montré.

Les contributions de cette étude sont les suivantes :

- Modélisation de l'incertitude des capteurs à faible coût en plus des phénomènes environnementaux.
- Étude de l'approche de Vecchia sur un ensemble de données de trajectoires réelles.
- Étudier plusieurs configurations, y compris de nouvelles configurations non testées telles que la modélisation spatio-temporelle, l'ordre des capteurs et les capteurs voisins.
- Application l'approche proposée à un ensemble de données sur la pollution dans le monde réel pour la cartographie et la prévision de la pollution.

Le contenu est structuré en trois sections principales :

- Présentation de la méthodologie, incluant une définition de l'approximation de Vecchia et des configurations étudiées ainsi que des indicateurs de performance utilisés.
- Exposition des résultats obtenus pour les différentes configurations.
- Application numérique du modèle et de l'approximation choisie.

Une discussion sur les implications de nos résultats et les perspectives futures de cette ligne de recherche conclut cette partie.

4.4.1 Méthodologie

L'approximation de Vecchia est une méthode qui réduit la complexité des modèles gaussiens en estimant la vraisemblance à partir d'un sous-ensemble de points. Elle a été proposée par Vecchia en 1988 [26], et depuis lors, elle est devenue une méthode largement utilisée pour gérer efficacement les modèles spatiaux complexes avec de grandes quantités de données [14, 16, 17].

L'enjeu principal des modèles spatiaux réside dans l'estimation des paramètres et des prédictions, qui nécessitent souvent la manipulation de matrices de covariance. Pour des dimensions spatiales élevées, cette tâche devient lourde en temps de calcul et exigeante en termes de stockage. L'approximation de Vecchia répond à ce défi en sélectionnant les points de données les plus pertinents et en se concentrant sur les interactions locales pour l'estimation.

L'idée sur laquelle repose cette approximation est que les données spatiales sont principalement influencées par les observations proches, plutôt que par l'ensemble des données disponibles. Cette simplification réduit drastiquement la complexité des calculs, rendant possible le traitement de bases de données de grande envergure.

Plus précisément, l'approximation de Vecchia est basée sur le remplacement de la densité du modèle 4.12, qui peut être écrite comme une densité conditionnelle :

$$p(Z) = p(z_1) \prod_{i=2}^n p(z_i/z_1, \dots, z_{i-1}) = p(z_1) \prod_{i=2}^n p(z_i/z_{v(i)}) \quad (4.15)$$

L'approximation de Vecchia remplace les vecteurs de conditionnement complets $v(i) = 1, \dots, i - 1$ par un sous-vecteur $v'(i)$ (appelé vecteur conditionnant de l'observation i). Ce sous-vecteur est souvent choisi pour contenir les indices correspondant aux observations proches en distance de la $i^{\text{ème}}$ observation. On obtient une distribution conjointe :

$$p'(Z) = p(z_1) \prod_{i=2}^n p(z_i/z_{v'(i)}) \quad (4.16)$$

Contrairement à $p(Z)$, $p'(Z)$ dépend de l'ordre dans lequel les observations sont classées et de la longueur du sous-vecteur v' notée M . Le choix des individus qui constitue ce vecteur est une conséquence de ces deux caractéristiques. L'emploi du vecteur entier ramènerait à la fonction de vraisemblance initiale. Néanmoins, l'utilisation d'un sous-vecteur réduit permet de diminuer considérablement la complexité de calcul de la fonction de vraisemblance. L'utilisation de densités conditionnelles permet également la parallélisation des calculs, ce

qui rend l'approximation de Vecchia relativement peu coûteuse en termes de calcul. Toutefois, l'approximation de Vecchia peut ne pas capturer efficacement les structures de dépendance complexes, en particulier dans les cas de dépendance à longue portée, ce qui limite son utilité pour certains processus spatiaux. Le choix judicieux des points de conditionnement est crucial, car il influence grandement la précision de l'approximation [17].

En somme, l'approximation de la vraisemblance de Vecchia offre un équilibre entre la facilité de calcul et la précision, ce qui en fait un outil puissant pour la modélisation des processus gaussiens dans les champs spatiotemporels de grandes dimensions.

A- Configurations étudiées

L'objectif premier de cette recherche est d'étudier l'efficacité de différentes configurations de Vecchia pour un processus gaussien spatio-temporel doté d'un paramètre de modèle à effets mixtes. Dans ce contexte, les configurations sont des permutations stratégiques de deux éléments cruciaux : l'ordre des observations et la sélection de l'ensemble conditionnel de voisins.

Six façons de ranger les observations sont étudiées ici : arrangement aléatoire, arrangement par coordonnées (spatiale et temporelle), arrangement suivant l'algorithme min-max distance (Distance Maximale Minimale (MMD)), arrangement à partir du centre (middle out), arrangement suivant les capteurs. Pour chacune des six configurations envisagées, six critères différents ont été appliqués pour sélectionner l'ensemble conditionnel des voisins. Ces critères sont basés sur trois définitions distinctes de la distance, tout en prenant en compte si l'observation appartient ou non au même capteur. Cela conduit à un total de trente-six configurations possibles.

Ordre ou arrangement des observations : essentiellement, l'ordre détermine la séquence dans laquelle les points de données sont traités :

- **Arrangement aléatoire** : Les observations sont rangées dans une séquence complètement aléatoire.
- **Arrangement par coordonnées** : Ici, les observations sont ordonnées en fonction de leurs coordonnées géographiques spatiales ou temporelles. Cette méthode est alignée sur la progression spatiale ou temporelle naturelle, garantissant que les observations voisines sont traitées plus ou moins successivement.
- **Arrangement par distance min-max (MMD)** : Cet ordre sélectionne d'abord un point au centre désigné par $\bar{x}$, puis choisit séquentiellement le point suivant qui a la distance minimale maximale par rapport à tous les points sélectionnés précédemment :

$$\begin{aligned}\tau(1) &= \arg \min_{i \in 1, \dots, n} \|z_i - \bar{z}\|, \\ \tau(j) &= \arg \max_{i \notin \tau(1), \dots, \tau(j-1)} \min_{k \in 1, \dots, j-1} \|z_i - z_{\tau(k)}\|, \quad j > 1. \\ \|z_i - z_j\| &= \sqrt{\frac{\|x_i - x_j\|^2}{(\theta_1)^2} + \frac{|t_i - t_j|^2}{(\theta_2)^2}}\end{aligned}\tag{4.17}$$

Les paramètres (θ_1, θ_2) sont estimés sur un sous échantillon du jeu des données. Le résultat est que pour chaque $k = 1, \dots, n$ les premiers k points forment un ensemble couvrant l'espace, aucun d'entre eux n'étant trop proche les uns des autres. L'ordre MMD est basé donc sur l'idée que chaque observation est entourée par les observations précédentes dans l'ordre.

- **Arrangement à partir du centre "Middle-out"** : L'arrangement à partir du centre commence l'ordonnancement à partir d'une observation centrale et se déplace vers l'extérieur dans toutes les directions.
- **Arrangement par capteur** : Dans ce cas, les observations sont ordonnées selon leur appartenance aux capteurs, regroupant toutes les observations du capteur 1 dans l'ordre temporel, avant de procéder de la même manière pour le capteur 2, et ainsi de suite.

Ensemble de voisins conditionnant : l'ensemble de voisins conditionnant fait référence à la sélection des points de données qui sont pris en compte dans

le calcul de la probabilité conditionnelle pour un point de données spécifique. En général, dans la littérature, nous choisissons simplement les M observations qui viennent avant dans l'ordre. Ici sont pris les M plus proche voisins de l'observation qui vienne avant dans l'ordre. Pour calculer les M plus proches voisins, trois définitions de la distance sont utilisées :

- **Voisins spatiaux** : Ici la distance est définie uniquement en fonction de la dimension spatiale. Le choix de cette approche est de favoriser l'inclusion d'observations provenant de différents moments temporels dans l'ensemble des voisins, de sorte que les vraisemblances conditionnelles contiennent des informations une plage temporelle plus grande.
- **Voisins temporels** : Cette approche est analogue à la précédente, mais utilise uniquement la distance dans le temps pour définir la proximité des données.
- **Voisins spatio-temporels** : Combinaisons des deux distances précédentes, ici sont pris les M plus proche voisins spatio-temporels.

Outre le choix de la métrique de distance pour la sélection des voisins les plus proches, il est également possible d'influer sur l'inclusion des observations conditionnelles appartenant au même capteur que l'observation i :

- **Dépendant du capteur** : Il est exigé que les M observations de l'ensemble de conditionnement de l'observation i soient issues du même capteur que l'observation i , même si cela implique d'inclure des observations plus éloignées. Cette méthode est choisie pour permettre l'étude de la corrélation potentielle entre Θ et γ^2 .
- **Indépendant du capteur** : l'ensemble des voisins est sélectionné indépendamment de leur appartenance au capteur.

En combinant les six stratégies d'ordonnancement et les six critères de sélection des voisins (3×2), un ensemble complet de 36 configurations pour l'approximation de Vecchia est obtenu. Ces configurations sont étudiées et évaluées afin de déterminer la meilleure approximation. Pour chaque configuration, la taille M de $v'(i)$ est augmentée progressivement et les indicateurs de performance choisis sont calculés. La figure 4.2 illustre ces différentes configurations

sur un exemple fictif dans un espace spatio-temporel provenant de 6 capteurs. Elle illustre le vecteur conditionnant de la 26^{ème} observation selon les 36 configurations étudiées.

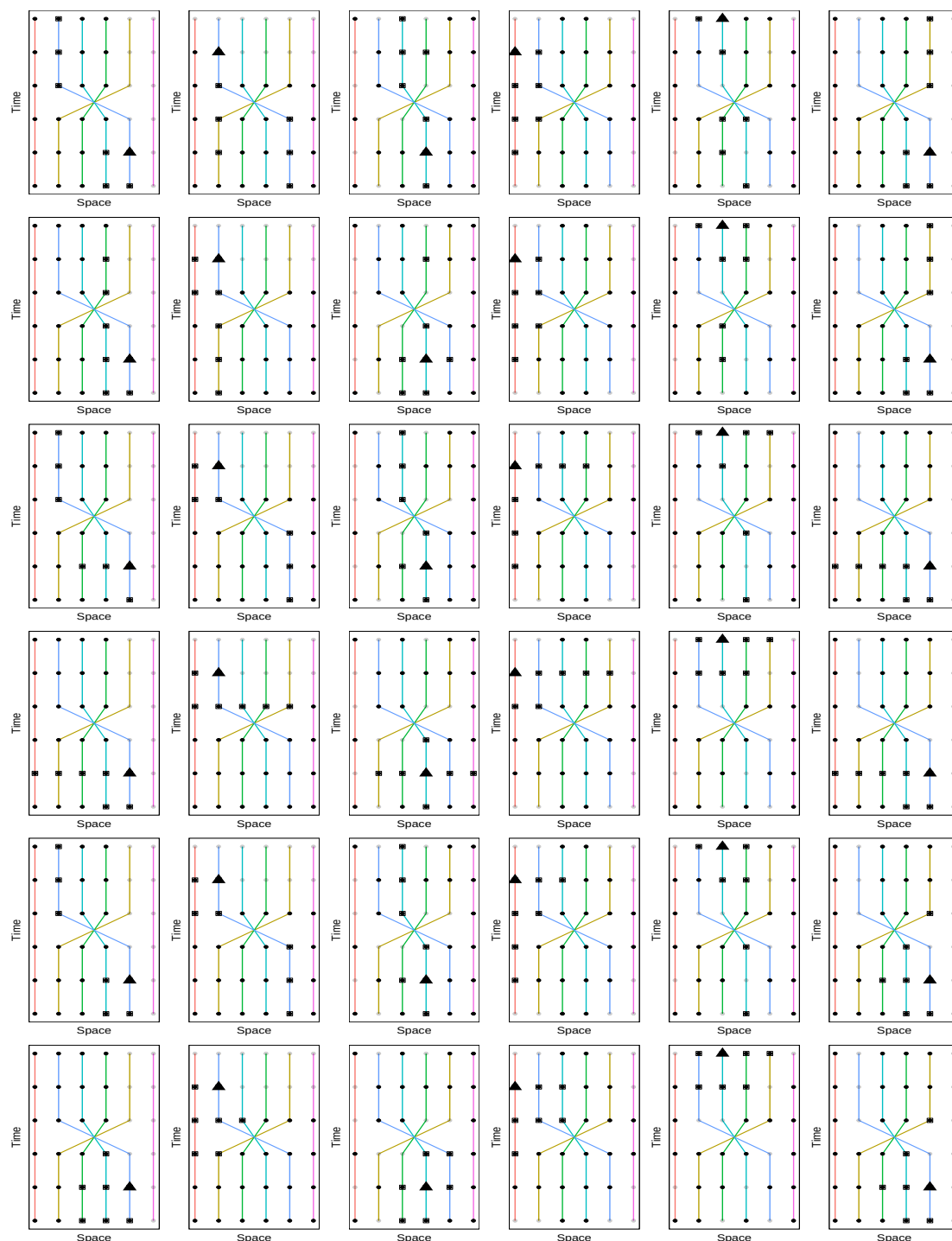


Figure 4.2 – Configurations étudiées : la colonne indique l'ordre de gauche à droite : Spatial, Temporel, Aléatoire, Middle Out, MMD, Capteur. Les deux premières lignes indiquent la distance spatiale uniquement, les troisième et quatrième lignes indiquent la distance temporelle uniquement, et les deux dernières lignes indiquent la distance spatio-temporelle. Les lignes paires indiquent que le conditionnement repose principalement sur les observations des mêmes capteurs. Les rangées impaires indiquent des conditionnements quel que soit le capteur. Les cercles noirs représentent les observations ordonnées de 1 à 25, le point triangulaire correspond au numéro 26 et les carrés représentent l'ensemble de conditionnement de l'observation 26.

B- Indicateurs de performance

Les performances de chaque configuration sont évaluées sur la base de deux indicateurs principaux :

- Divergence de Kullback-Leibler (D KL) : Cette mesure quantifie la différence entre le modèle approximé dérivé de chaque configuration et le modèle original. La divergence de Kullback-Leibler, ou entropie relative, fournit une mesure de l'information perdue lorsque le modèle approximé est utilisé à la place du véritable modèle. Une divergence KL plus faible indique une approximation plus proche du modèle ciblé, ce qui implique une configuration supérieure.

$$KL[P||Q] = \frac{1}{2} \left[\text{tr} (\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \quad (4.18)$$

- Efficacité relative asymptotique (ARE) : L'ARE fournit une mesure de l'efficacité de l'estimation des paramètres de covariance, calculée à l'aide de l'information de Fisher généralisée pour les vraisemblances mal spécifiées. La variance du paramètre dans l'approximation est obtenue à partir des éléments diagonaux de l'inverse de la matrice hessienne de la vraisemblance de Vecchia. En revanche, la variance réelle des paramètres est dérivée de l'information de Fisher.[10]. Essentiellement, elle capture la précision des estimations des paramètres de covariance. Des valeurs ARE plus faibles indiquent des estimations plus efficaces, et donc une configuration plus efficace.

$$ARE(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\hat{\theta})}{\text{Var}(\tilde{\theta})} \quad (4.19)$$

L'exploration de ces diverses configurations, combinée aux critères d'évaluation des performances, permet de mettre en lumière les stratégies optimales pour l'approximation du processus gaussien spatio-temporel, en particulier dans le contexte des données de capteurs à faible coût pour la cartographie et la prévision de variables environnementales.

C- Données sur les trajectoires et paramètres de simulation

La pierre angulaire de l'approche méthodologique pour comparer différentes approximations de Vecchia au modèle réel réside dans la simulation du processus gaussien au moyen de trajectoires issues de capteurs mobiles. Cette étape fournit un environnement expérimental contrôlé dans lequel nous connaissons précisément le véritable modèle statistique sous-jacent, permettant ainsi une évaluation précise des configurations de Vecchia grâce aux indicateurs de performance.

$$C(\mathbf{x}, t; \mathbf{x}', t') = \sigma^2 \exp \left(-\sqrt{\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{(\theta_1)^2} + \frac{|t - t'|^2}{(\theta_2)^2}} \right) \quad (4.20)$$

Pour simuler le processus gaussien, une matrice de variance-covariance dérivée des distances spatiales et temporelles entre les observations à l'aide de l'équation 4.20 est construite. Les paramètres de la fonction de covariance sélectionnée servent d'entrées à cette matrice. Le paramètre τ^2 est ensuite ajouté à la diagonale de cette matrice. La méthode de décomposition de Cholesky est ensuite utilisée pour générer un ensemble de variables aléatoires corrélées.

L'intégration du bruit de mesure propre à chaque capteur représente une étape essentielle de la simulation afin de refléter l'incertitude typique des relevés issus de capteurs peu coûteux. Du bruit est ajouté à chaque observation associée à un capteur spécifique, tiré au hasard d'une distribution normale avec une moyenne nulle et une variance γ^2 . Ce bruit introduit un élément réaliste de variation aléatoire, reflétant les incertitudes de mesure intrinsèques des capteurs bas coût.

Les paramètres sont sélectionnés de manière à ce que les résultats obtenus soient extrapolables et applicables à des données réelles provenant d'expérimentations sur le terrain. À cette fin, les paramètres du modèle sont estimés par la méthode du maximum de vraisemblance classique sur un sous-ensemble du jeu de données. Les valeurs estimées sont ensuite utilisées pour générer un processus gaussien simulé.

4.4.2 Résultats d'évaluation

Dans cette section, les résultats de l'enquête sur les performances des différentes configurations sont présentés, mettant en lumière l'impact de l'ordonnement des observations et de la taille du vecteur conditionnant sur la divergence de Kullback-Leibler et l'efficacité relative asymptotique pour l'estimation des paramètres de covariance.

- Divergence Kullback-Leibler

Les résultats de l'évaluation de la divergence de Kullback-Leibler sont présentés dans la figure 4.3. Des scores plus faibles indiquent une approximation plus proche du modèle réel, ce qui suggère que ces configurations capturent les caractéristiques statistiques sous-jacentes de manière plus efficace.

Conformément aux attentes, on observe une diminution des scores de divergence à mesure que le nombre de voisins inclus dans l'approximation augmente.

L'efficacité de l'ordre MMD en contexte spatio-temporel ressort nettement, comme l'ont démontré également Guinness et al.[10] dans le contexte purement spatial. Étonnamment, l'ordre aléatoire affiche des performances comparables à celles de l'ordre MMD, le surpassant même dans certaines configurations.

Il a été observé que contraindre le vecteur conditionnant à ne contenir que des observations du même capteur n'améliore pas les résultats. La configuration où les capteurs ne sont pas contraints à l'appartenance aux capteurs spécifiques fournit de meilleurs résultats pour tous les ordres étudiés, en raison de l'implication de voisinages plus proches.

Le choix de la distance utilisée est important, en particulier pour les petites valeurs de M : la distance spatiale seule est nettement moins performante que les distances temporelle et spatio-temporelle pour les petites valeurs de M , même

si la différence diminue au fur et à mesure que l'on augmente M .

Globalement, la distance spatio-temporelle est à privilégier quel que soit le nombre de voisins utilisés; cependant, l'utilisation de cette distance nécessite un paramètre supplémentaire à estimer au préalable. L'utilisation de la distance purement temporelle demeure un excellent choix si l'on souhaite éviter cette étape.

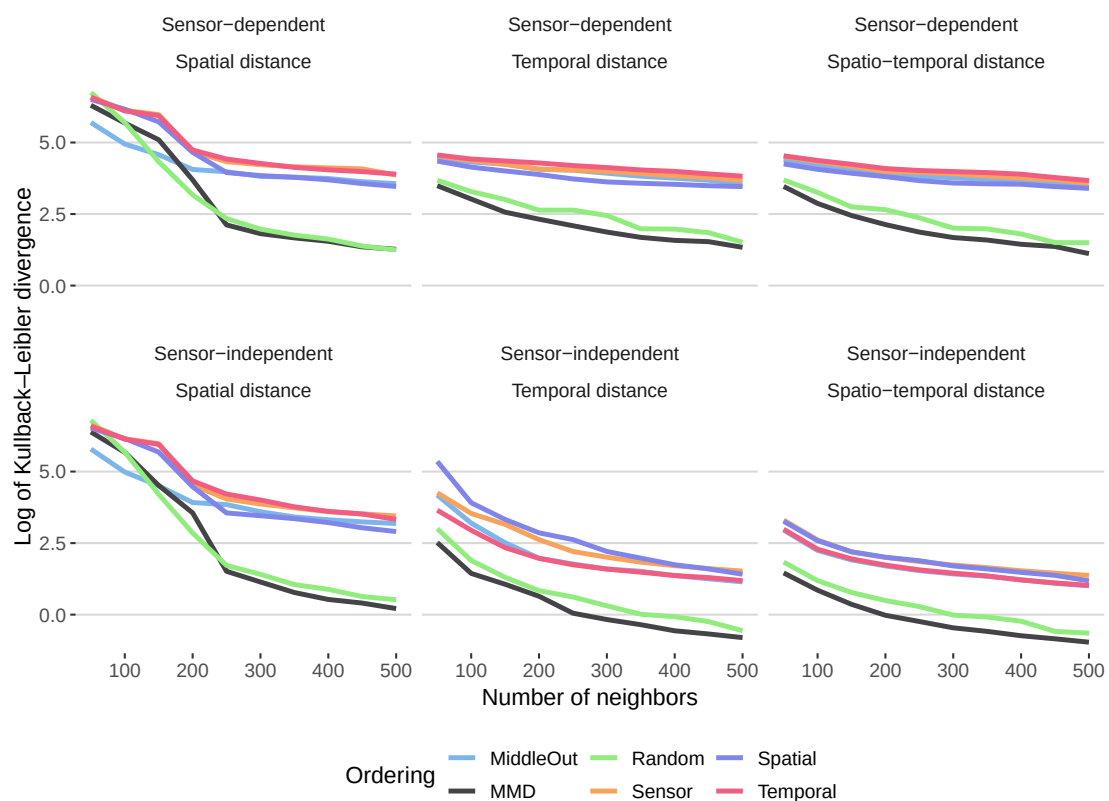


Figure 4.3 – Kullback–Leibler divergence

Efficacité relative asymptotique

Les résultats de l'efficacité relative asymptotique sont présentés dans la figure 4.4, les résultats étant sur une échelle logarithmique pour faciliter l'analyse.

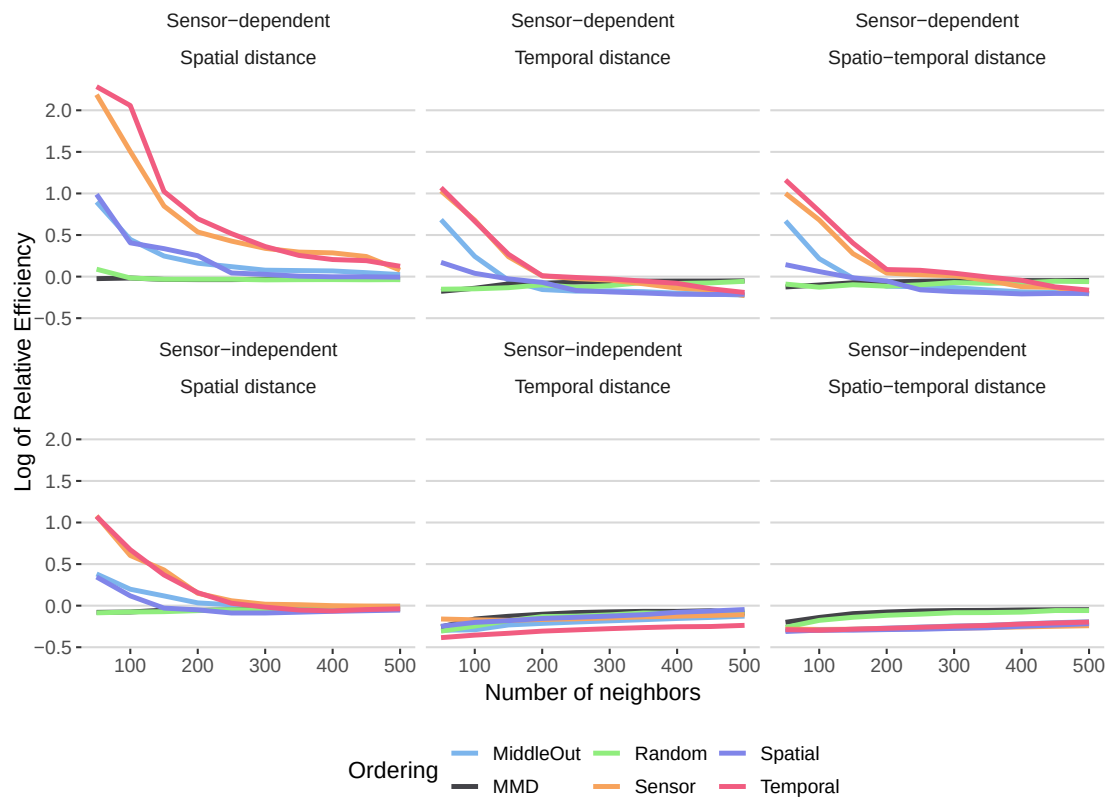


Figure 4.4 – Log de l'efficacité relative

Un score de 0 indique que le rapport entre la variance des paramètres estimés dans l'approximation est égal à la variance des vrais paramètres du modèle.

Concordant avec les données de la divergence de Kullback-Leibler, les ordres MMD et aléatoires se rapprochent le plus du modèle réel, reflétant des ARE approchant 1.

Une augmentation du nombre de voisins se traduit par une convergence vers le modèle réel et, par conséquent, une ARE se rapprochant de 1, indépendamment de l'ordre ou de la configuration du voisinage.

Il est intéressant de noter que, dans certaines configurations, l'ARE peut être inférieure à 1, indiquant que la variance des paramètres approximés est inférieure à celle du vrai modèle. Ceci est particulièrement évident lorsqu'aucune contrainte n'est imposée sur le choix du voisinage par rapport à la source du

capteur et lors de l'utilisation de distances temporelles ou spatio-temporelles pour des valeurs faibles de M .

Dans le cas de petites valeurs de M , la situation est comparable à l'examen d'un modèle simplifié où les paramètres dévient de ceux du modèle réel. La figure correspondante de la divergence Kullback-Leibler montre une divergence importante pour un faible M , indiquant que l'estimation des paramètres se concentre davantage sur ces valeurs biaisées, mais avec une précision accrue. Par conséquent, il est intéressant d'étudier le compromis entre la variance et le biais de l'estimateur en modulant la valeur de M .

L'analyse ARE renforce l'argument selon lequel les contraintes basées sur l'appartenance de l'observation au capteur sont contre-productives : non seulement elles s'écartent du vrai modèle, mais elles introduisent également une plus grande variance dans l'estimation des paramètres.

Concernant l'utilisation de la distance temporelle ou spatio-temporelle, sans contrainte sur l'appartenance aux capteurs, l'ordre temporel offre une variance plus faible sur les paramètres estimés. Cette variance augmente jusqu'à atteindre celle du vrai modèle au fur et à mesure que M augmente.

En conclusion, la configuration optimale pour l'estimation des paramètres utilise les métriques de distance temporelle ou spatiotemporelle, sans contraindre les observations selon leur appartenance aux différents capteurs, et en adoptant un ordre suivant l'algorithme MMD ou un ordre aléatoire. Si l'objectif est de minimiser la variance d'estimation de certains paramètres spécifiques, l'exploration d'ordres alternatifs, spécifiquement conçus pour minimiser cette dernière, peut être envisagée.

4.4.3 Application numérique

Dans cette section est présentée l'application numérique dans laquelle est utilisée la configuration de Vecchia optimale (distance minimale maximale avec distance spatio-temporelle) pour estimer les paramètres de la vraisemblance du modèle décrit dans la section 2. Cette application emploie l'ensemble des don-

nées sur la qualité de l'air recueillies par une flotte de capteurs fixes et mobiles produits par Atmotrack à Nantes, utilisées dans la seconde partie du chapitre 3.

Comme dans le chapitre précédent, l'utilisation des données d'Atmotrack nécessite une transformation logarithmique au préalable, en raison de leur distribution asymétrique, contrairement aux données plus uniformément réparties d'OpenSense. Et ce afin de mieux adhérer à l'hypothèse de normalité. Cette approche s'appuie sur des recherches antérieures, qui ont souvent révélé que les niveaux de pollution suivent une distribution log-normale, confirmant ainsi la pertinence de cette méthode pour analyser ce type de données.[11].

Comparaison avec le modèle géostatistique classique

Les résultats de l'estimation des paramètres pour les modèles du krigeage avec dérives externes 4.9 et le modèle précédemment proposé 4.10 sont présentés dans le tableau 4.1. Comme prévu, nous observons une plus grande vraisemblance dans le modèle 4.10, puisque le modèle 4.9 est imbriqué dans le modèle 4.10. Ceci implique que la vraisemblance du modèle 4.10 est au moins égale à celle du modèle 4.9 lorsque $\gamma = 0$. Dans le cas d'un paramètre excédentaire γ , une estimation de 0 pour γ entraînerait un retour au premier modèle avec la même vraisemblance.

En outre, il est évident qu'une part importante de la variance du modèle 4.9 a été capturée par γ . Il est difficile d'effectuer une comparaison plus complète entre les deux modèles, car même au sein d'un modèle unique, il peut exister plusieurs maxima locaux dans la fonction de vraisemblance, ce qui conduit à diverses interprétations des données.

L'analyse comparative entre notre modèle et le modèle géostatistique classique révèle des améliorations notables. En termes de corrélation, le modèle proposé présente un gain passant de 0,3 à 0,42. De plus, la valeur de l'erreur quadratique moyenne (RMSE) s'améliore, passant de 1 à 0,7.

Paramètre	Modèle géostatistique classique	Modèle proposé
Variance $\sigma^2(\log(\mu g/m^3))$	$1,9 \cdot 10^{-1}$	$2,6 \cdot 10^{-2}$
Dimension spatiale θ_1 1/m	$1,38 \cdot 10^4$	$4 \cdot 10^3$
Dimension temporelle θ_2 1/s	$8,1 \cdot 10^3$	10^3
Pépité $\tau^2(\log(\mu g/m^3))$	$5 \cdot 10^{-4}$	$2,9 \cdot 10^{-2}$
Autocorrélation du capteur $\gamma^2(\log(\mu g/m^3))$	-	$2,6 \cdot 10^{-2}$
Log-vraisemblance	35112	37924

Table 4.1 – Résultat de la procédure d'estimation des paramètres pour le modèle de krigeage avec dérives externes 4.9 et le modèle hiérarchique à 3 niveaux proposé 4.10.

Il est important de souligner que, bien que les résultats soient modestes, ils sont obtenus en estimant à partir de données à faible coût, sans l'aide de covariables pertinentes telles que le trafic routier. De plus, la comparaison est effectuée par rapport à des données de référence fixes (Figure 4.5). Dans de telles conditions, la plupart des modèles éprouvent des difficultés à offrir des performances optimales. Néanmoins, le modèle proposé parvient à capturer une tendance globale satisfaisante. Il convient de mentionner que le jour sélectionné pour cette analyse n'était pas nécessairement le plus propice pour la prédiction, mais il a été choisi en raison de la quantité de données disponibles.

Une divergence est observée avec les données issues des stations fixes. L'hypothèse est que les capteurs sont principalement axés sur la mesure de la pollution issue du trafic, alors que les stations fixes sont orientées vers la surveillance de la pollution de fond. Une exception notable est la station Victor Hugo, qui semble mesurer la pollution liée au trafic. Cela pourrait expliquer pourquoi la différence entre les prédictions des modèles et les données de cette station est moins prononcée.

Génération de cartographies

L'estimation des paramètres du modèle proposé a permis d'utiliser la formule de l'espérance conditionnelle pour générer des cartographies de PM₁₀. Les prédictions ont été calculées sur une grille spatiale régulièrement espacée de

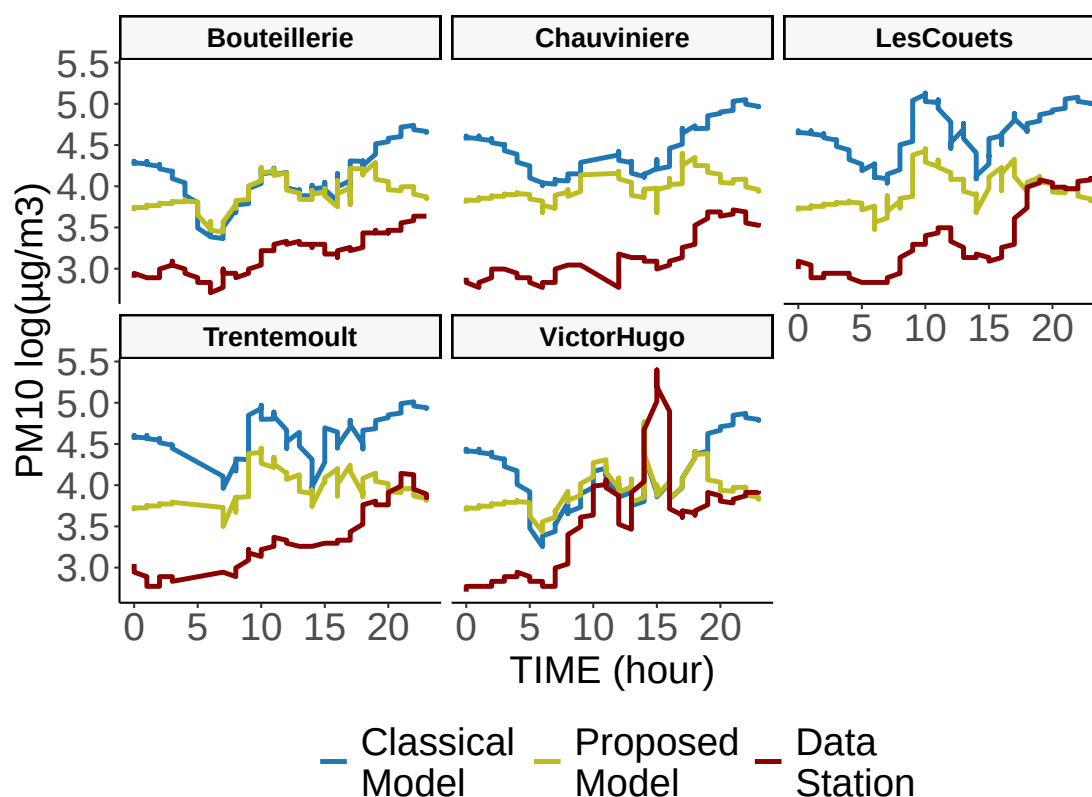


Figure 4.5 – Comparaison du modèle proposé avec le modèle classique

taille 100x100 points, couvrant la zone d'étude de la ville de Nantes, sur douze périodes de temps. Cela donne 100x100x12 points de 7h à 18h, soit 10 000 points par heure. Pour chaque point de cette grille, l'ensemble des covariables ont été calculées.

La figure 4.6 démontre l'application réussie de cette méthodologie à l'ensemble de données d'Atmotrack dans la ville de Nantes. Des pics quotidiens ont été détectés spécifiquement à 9 heures et à 18 heures dans certains endroits de la ville. Il a également été possible de noter les variations des tendances temporelles qui représentent la moyenne journalière habituelle, marquée par des concentrations de particules plus importantes aux heures de début et de fin de journée. Ces augmentations coïncident avec les heures de forte affluence du trafic routier.

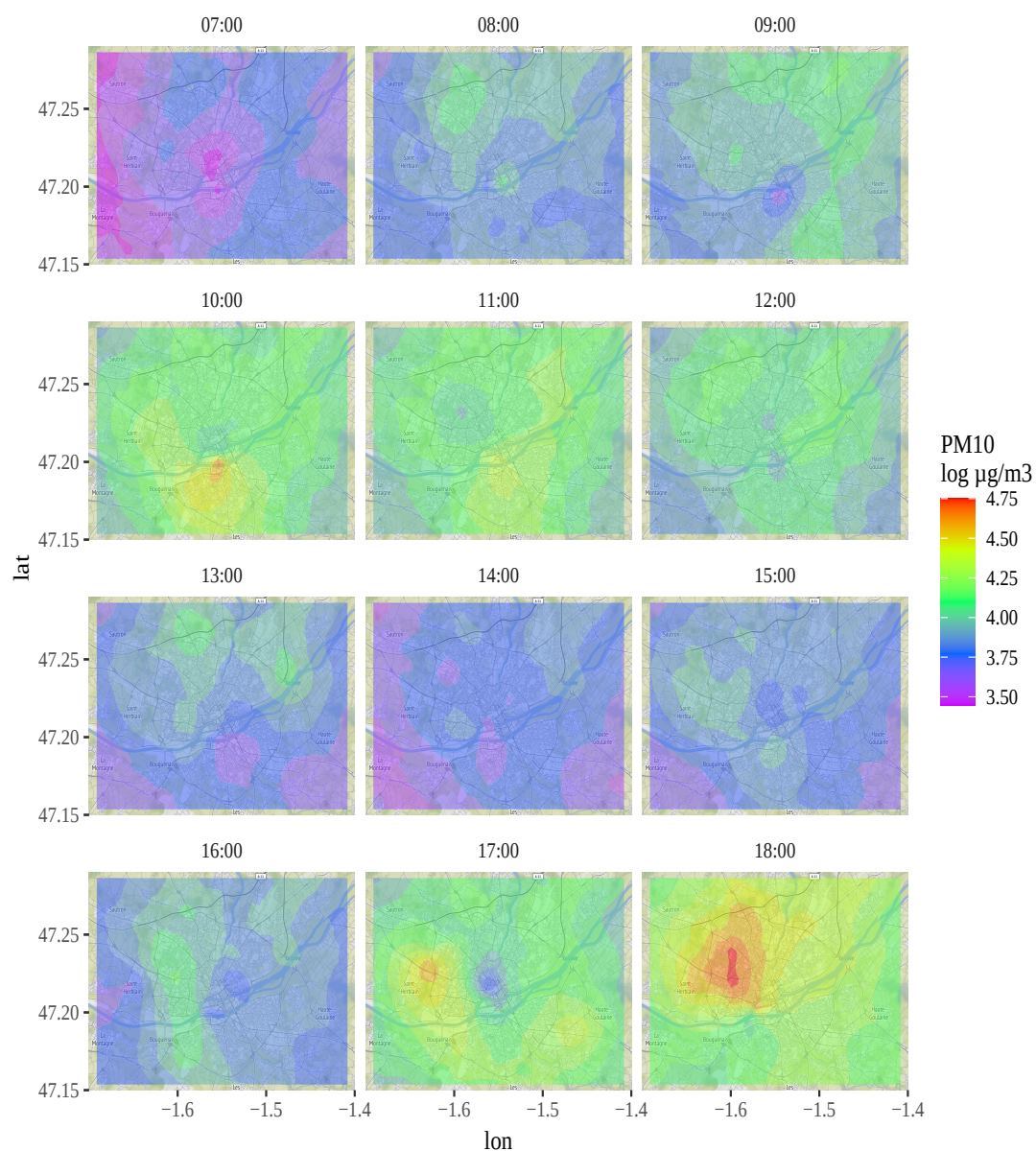


Figure 4.6 – Cartographie résultant de l'application du modèle proposé en utilisant l'approximation de Vecchia pour estimer les paramètres

4.4.4 Discussion

Les résultats de l'efficacité relative englobent la variabilité de tous les paramètres et sont calculés en faisant la moyenne de la diagonale de la matrice d'information de Fisher. Cependant, si l'on se concentre sur un seul paramètre, tel que γ , l'efficacité relative de ce paramètre particulier peut être meilleure dans une configuration spécifique. Lorsque l'on considère tous les paramètres simultanément, le graphique d'efficacité relative permet d'identifier la configuration optimale.

Dans certains cas, il peut être donc plus avantageux de sélectionner une configuration présentant une divergence de Kullback-Leibler légèrement supérieure afin d'obtenir une meilleure estimation de certains paramètres.

Si l'on s'intéresse principalement à l'estimation de γ , la variance intra-capteur des capteurs à faible coût, il pourrait être intéressant d'inclure uniquement les observations du même capteur dans l'ensemble de conditionnement. Cette approche permettrait d'obtenir une estimation plus ciblée du paramètre souhaité.

En plus de fournir les variances des paramètres, la matrice d'information de Fisher donne également un aperçu de la covariance entre les paramètres. Une covariance significative a été observée entre les paramètres de variance σ^2 et γ^2 . Cette covariance est due à la formulation du modèle et au fait que la plupart des observations de l'ensemble de conditionnement proviennent du même capteur en raison de leur proximité spatiale et temporelle.

Plusieurs alternatives ont été explorées mais n'ont pas été présentées, notamment l'utilisation d'observations provenant uniquement d'autres capteurs dans l'ensemble de conditionnement. Conformément à nos anticipations, ces résultats étaient inférieurs, suggérant une plus grande efficacité en se basant exclusivement sur les données indépendamment de leur appartenance au capteurs.

4.5 Conclusion

Cette recherche a illustré l'efficacité de l'approximation de Vecchia appliquée à un processus gaussien spatio-temporel doté d'un paramètre d'autocorrélation des capteurs. La conception de ce modèle s'est faite dans l'optique d'intégrer les erreurs intrinsèques et les corrélations relatives aux mesures issues de capteurs économiques.

L'exploration de l'approximation de Vecchia sur ledit modèle met en exergue la prééminence de l'ordre suivant l'algorithme MMD et le tri aléatoire. Cette découverte est importante : elle présente une technique robuste pour l'approximation, facilitant ainsi l'estimation des paramètres du modèle face à d'importants volumes de données.

L'application réussie de ce modèle à un ensemble de données de mesures de la qualité de l'air souligne son potentiel pour une application plus large dans le domaine de la surveillance de l'environnement. Les capteurs à faible coût sont de plus en plus répandus en raison de leur prix abordable et de leur portabilité, et cette étude fournit une orientation prometteuse pour la recherche future dans ce domaine.

En substance, ces travaux s'inscrivent dans la littérature grandissante sur l'application de l'approximation de Vecchia et la modélisation via des capteurs environnementaux bas coût. Des recherches ultérieures pourraient explorer l'application de ce modèle à d'autres types de phénomènes physiques spatio-temporels et dans différents environnements urbains. D'autres études pourraient également tirer parti du cadre hiérarchique pour affiner la modélisation des erreurs mentionnées précédemment dans ce chapitre. Une telle démarche renforcerait la validation de nos conclusions et étendrait la portée du modèle.

Une voie prometteuse pour améliorer significativement la précision des prédictions de la pollution atmosphérique consisterait à disposer de covariables plus pertinentes. Il est recommandé d'exploiter les données de trafic, qui, mal-

heureusement, n'ont pas été accessibles pour la ville de Nantes dans le cadre de cette étude. De plus, l'incorporation de résultats issus de modèles de dispersion urbains pourrait également contribuer à affiner la cartographie des polluants.

Enfin, bien que le modèle hiérarchique à trois niveaux proposé affiche une amélioration vis-à-vis du modèle géostatistique traditionnel, ses prédictions demeurent relativement éloignées des observations enregistrées par les stations fixes de référence. La sophistication du modèle théorique trouve ses limites. Il est impératif d'enrichir notre approche par des moyens pratiques pour mieux répondre à nos enjeux. Ce sera l'objet du prochain chapitre.

Références

- [1] Onureena Banerjee, Laurent El Ghaoui et Alexandre d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data". In : *The Journal of Machine Learning Research* 9 (2008), p. 485-516.
- [2] Sudipto Banerjee, Bradley P Carlin et Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.
- [3] Jean-Paul Chiles et Pierre Delfiner. *Geostatistics : modeling spatial uncertainty*. T. 497. John Wiley & Sons, 2009.
- [4] Noel Cressie et Gardar Johannesson. "Fixed rank kriging for very large spatial data sets". In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 70.1 (2008), p. 209-226.
- [5] Noel Cressie et Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.
- [6] Juan Du, Hao Zhang et VS2549562 Mandrekar. "Fixed-domain asymptotic properties of tapered maximum likelihood estimators". In : (2009).
- [7] Reinhard Furrer, Marc G Genton et Douglas Nychka. "Covariance tapering for interpolation of large spatial datasets". In : *Journal of Computational and Graphical Statistics* 15.3 (2006), p. 502-523.
- [8] Andrew Gelman et Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.
- [9] Joseph HA Guillaume et al. "Introductory overview of identifiability analysis : A guide to evaluating whether you have the right type of data for your modeling purpose". In : *Environmental Modelling & Software* 119 (2019), p. 418-432.
- [10] Joseph Guinness. "Permutation and grouping methods for sharpening Gaussian process approximations". In : *Technometrics* 60.4 (2018), p. 415-429.

- [11] David Hasenfratz et al. "Deriving high-resolution urban air pollution maps using mobile sensor nodes". In : *Pervasive and Mobile Computing* 16 (2015), p. 268-285.
- [12] Matthew J Heaton et al. "A case study competition among methods for analyzing large spatial data". In : *Journal of Agricultural, Biological and Environmental Statistics* 24 (2019), p. 398-425.
- [13] David Higdon. "A process-convolution approach to modelling temperatures in the North Atlantic Ocean". In : *Environmental and Ecological Statistics* 5 (1998), p. 173-190.
- [14] Raphaël Huser, Michael L Stein et Peng Zhong. "Vecchia Likelihood Approximation for Accurate and Fast Inference in Intractable Spatial Extremes Models". In : *arXiv preprint arXiv :2203.05626* (2022).
- [15] Matthias Katzfuss et Noel Cressie. "Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets". In : *Journal of Time Series Analysis* 32.4 (2011), p. 430-446.
- [16] Matthias Katzfuss et Joseph Guinness. "A general framework for Vecchia approximations of Gaussian processes". In : (2021).
- [17] Matthias Katzfuss et al. "Vecchia approximations of Gaussian-process predictions". In : *Journal of Agricultural, Biological and Environmental Statistics* 25 (2020), p. 383-414.
- [18] Cari G Kaufman, Mark J Schervish et Douglas W Nychka. "Covariance tapering for likelihood-based estimation in large spatial data sets". In : *Journal of the American Statistical Association* 103.484 (2008), p. 1545-1555.
- [19] Finn Lindgren, Håvard Rue et Johan Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields : the stochastic partial differential equation approach". In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 73.4 (2011), p. 423-498.
- [20] Balz Maag, Zimu Zhou et Lothar Thiele. "A survey on sensor calibration in air pollution monitoring deployments". In : *IEEE Internet of Things Journal* 5.6 (2018), p. 4857-4870.

- [21] Douglas Nychka et al. "A multiresolution Gaussian process model for the analysis of large spatial datasets". In : *Journal of Computational and Graphical Statistics* 24.2 (2015), p. 579-599.
- [22] Joaquin Quinonero-Candela et Carl Edward Rasmussen. "A unifying view of sparse approximate Gaussian process regression". In : *The Journal of Machine Learning Research* 6 (2005), p. 1939-1959.
- [23] Havard Rue et Leonhard Held. *Gaussian Markov random fields : theory and applications*. CRC press, 2005.
- [24] Eric Schulz, Maarten Speekenbrink et Andreas Krause. "A tutorial on Gaussian process regression : Modelling, exploring, and exploiting functions". In : *Journal of Mathematical Psychology* 85 (2018), p. 1-16.
- [25] Michael L Stein. *Interpolation of spatial data : some theory for kriging*. Springer Science & Business Media, 1999.
- [26] Aldo V Vecchia. "Estimation and model identification for continuous spatial processes". In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 50.2 (1988), p. 297-312.
- [27] Ruud Wetzels et al. "Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task". In : *Journal of Mathematical Psychology* 54.1 (2010), p. 14-27.
- [28] Christopher K Wikle et Noel Cressie. "A dimension-reduced approach to space-time Kalman filtering". In : *Biometrika* 86.4 (1999), p. 815-829.
- [29] Christopher KI Williams et Carl Edward Rasmussen. *Gaussian processes for machine learning*. T. 2. 3. MIT press Cambridge, MA, 2006.

Chapitre 5

Synergie des mesures : stratégies pratiques pour une meilleure prévision de la qualité de l'air

Points clés du chapitre

- Proposition d'un modèle innovant prenant en compte les biais multiplicatifs et additifs des capteurs bas coût.
- Mise en œuvre d'une expérimentation dans le quartier des Halles basée sur ce nouveau modèle pour estimer ses paramètres.
- Démonstration de la réduction significative de l'erreur quadratique moyenne grâce aux capteurs mobiles.

Contenu du chapitre

5.1	Extension pratique du modèle théorique	176
5.1.1	Reconceptualisation du modèle	176
5.1.2	Méthodologie expérimentale	178
5.2	Évaluation globale : résultats et questions de recherche	185

5.2.1	Création de cartographies	186
5.2.2	Pouvoir descriptif du modèle	188
5.2.3	Complexité	191
5.2.4	Erreur associée à la prédiction	192
5.2.5	Étude du nombre de capteurs	193
5.3	Conclusion	196
	Références	198

Au chapitre précédent, l'accent a été mis sur les atouts du modèle développé dans cette thèse, qui prend en considération les défis posés par les capteurs bon marché. Ces derniers, comme observé au chapitre 3, tendent souvent à fournir des données moins fiables que celles issues des stations fixes.

Ce chapitre s'appuiera sur les fondations établies précédemment et continuera la recherche d'amélioration des méthodes de création de cartographies de pollution spatio-temporelles de haute précision. L'objectif est de découvrir de nouvelles méthodes pour traiter plus efficacement les données imprécises. L'analyse débutera par une évaluation des limites théoriques du modèle décrit dans le chapitre précédent, suivie de la présentation d'une expérience pratique pour adresser ces problématiques.

Ensuite, les questions de recherche posées dans l'introduction de ce manuscrit seront reconsidérées et analysées à la lumière des résultats issus de la dernière série d'expérimentations.

Pour conclure, les principaux enseignements tirés de cette recherche seront synthétisés. Sur cette base, nous formulerons des recommandations et des perspectives pour les futures études et applications dans ce domaine.

5.1 Extension pratique du modèle théorique

Cette section développe les motivations, les données et le plan expérimental de l'expérimentation menée dans le quartier des Halles à Paris. Elle commence par surmonter la barrière théorique du modèle précédent, avant de présenter la méthodologie et les données. Les résultats de l'étude sont ensuite exposés.

5.1.1 Reconceptualisation du modèle

Rappelons le modèle 4.10 développé dans le chapitre précédent :

$$\begin{aligned} Z_i / (Y, a_j) &\sim N(a_j + Y, \tau^2 \cdot I) \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \\ a_j &\sim N(0, \gamma^2) \quad \text{Pour } j : 1 \dots J \end{aligned} \tag{5.1}$$

Il a été montré que ce modèle est mieux adapté pour gérer les erreurs associées à l'utilisation de capteurs à bas coût, offrant de meilleurs résultats que le modèle géostatistique classique qui part du principe que les données des capteurs décorréelées et sont centrées sur le phénomène. Toutefois, ce modèle présente une limitation majeure : il exige que les biais des capteurs suivent une distribution normale centrée, c'est-à-dire que la somme des biais soit égale à zéro. Cette condition est essentielle pour que le modèle soit indentifiable, mais en pratique elle est trop restrictive dans le contexte des capteurs à bas coût.

Une approche plus pragmatique serait d'utiliser le modèle hiérarchique suivant :

$$\begin{aligned} Z/Y &\sim N(a \cdot Y + b, \tau^2 \cdot I) \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \end{aligned} \tag{5.2}$$

Où a et b représentent respectivement le biais multiplicatif et le biais additif des capteurs à bas coût. En l'absence de données supplémentaires, ces biais ne peuvent être estimés.

Supposons maintenant que nous disposions de données plus précises sur les

valeurs de Y aux mêmes points spatio-temporels que ceux échantillonnés par les capteurs mobiles :

$$\begin{aligned} Z(x_i, t_i)/Y &\sim N(a.Y(x_i, t_i) + b, \tau^2 \cdot I) \quad i = 1 : n \\ F_m(x_i, t_i)/Y &\sim N(Y(x_i, t_i), \tau_F^2 \cdot I) \quad i = 1 : n_F \quad \text{avec } n_F < n \\ Y &\sim N(\beta X, \sigma^2 \cdot R) \end{aligned} \quad (5.3)$$

Nous n'utilisons pas directement les données de F car, compte tenu de leur coût d'acquisition, ces données sont rares. Ainsi, nous utiliserons un boîtier mobile équipé d'un capteur Fidas® Frog, un capteur de pollution portatif étalonné afin de nous servir de référence. Grâce à son GPS, ce capteur de référence sera déployé en binôme avec un capteur à bas coût, permettant une calibration et une estimation des paramètres a et b pendant le déploiement.

D'après le site du fabricant [5], « le Fidas Frog est un spectromètre optique de haute résolution permettant une détermination rapide, fiable et de qualité des poussières fines avec une autonomie de fonctionnement de 8h sur batterie » .

Il mesure simultanément les concentrations de particules PM₁, PM_{2,5}, PM₄, PM₁₀, en nombre et en masse et détermine une granulométrie des particules échantillonnées sur une gamme de taille de 0,18 à 40 µm (256 canaux).

Le coût de ce genre de capteurs reste trop élevé et leur volume trop important pour envisager un déploiement aussi massif que ceux des capteurs bas coûts. Cependant il permet à l'expérimentation proposée de bénéficier d'une solution de mesure de référence portative pouvant suivre les mobiles et ainsi pouvoir estimer les paramètres des différents biais.

Cette approche s'apparente à une calibration basée sur une rencontre in situ avec des stations de référence, souvent désignée sous le terme « calibration par rendez-vous » [1, 4, 7]. Toutefois, dans notre contexte, la zone tampon — espace où les deux capteurs sont supposés être dans les mêmes conditions et donc mesurer la même quantité de polluant — est très réduite. De plus, l'intervalle de temps de chaque rendez-vous est notablement long.

Une perspective alternative consiste à envisager cette démarche comme une phase de calibration préalable à l'utilisation des capteurs. Typiquement, cette étape implique la comparaison des mesures des capteurs avec des valeurs de référence et la prise en compte des divergences observées. Dans le contexte de cette étude, cette calibration s'effectue simultanément au déploiement, assurant ainsi l'uniformité des conditions environnementales. Cette méthode pourrait offrir l'avantage supplémentaire de quantifier l'influence des diverses variables environnementales sur la précision des capteurs à faible coût.

Afin de réaliser la calibration de tous les capteurs mobiles à bas coût avec un nombre limité de capteurs de référence, un protocole expérimental a été conçu. Les détails de cette méthodologie seront exposés dans la section suivante.

5.1.2 Méthodologie expérimentale

Plan d'expérimentation

Dans le but d'obtenir des données avec une résolution spatiale très fine, cinq capteurs mobiles ont été déployés, associés à un capteur de référence, dans le quartier des Halles à Paris, pour une durée d'une journée. La sélection de ce quartier n'était pas fortuite ; il abrite en effet une station fixe de référence de Air Parif, facilitant ainsi la comparaison et la validation des données collectées. De plus, il s'agit d'un quartier mêlant des espaces urbains denses, des parcs ainsi que des avenues à forte circulation représentant ainsi dans une aire restreinte une grande diversité de configurations.

La méthode proposée tire son efficacité de la proximité spatiale entre le capteur de référence et les capteurs mobiles. Il est donc crucial de maintenir le capteur de référence à une distance rapprochée des capteurs mobiles pendant une période suffisante pour permettre une estimation précise des paramètres a et b . À cette fin, un programme comprenant cinq cycles de mesure a été établi, avec une durée moyenne de 45 minutes par cycle. Au cours de chaque cycle, le capteur de référence Fidas Frog était positionné à proximité immédiate d'un

capteur mobile à bas coût différent, permettant ainsi la collecte de données comparatives.

Lors de la planification de ces déplacements, une attention particulière a été accordée aux trajectoires des capteurs mobiles. L'objectif était de maximiser la distance entre eux tout en garantissant une couverture spatiale complète des rues du quartier. Cette problématique est similaire au problème du "postier chinois", où il s'agit de trouver le plus court chemin permettant de traverser chaque rue d'un quartier au moins une fois.

Les mesures issues des capteurs mobiles étaient initialement brutes et non étalonnées. L'analyse préliminaire a révélé des fluctuations rapides et abruptes dans les données, se produisant en l'espace de quelques secondes. Afin d'atténuer ces variations soudaines, une correction préliminaire a été mise en œuvre sous forme de moyenne mobile sur une fenêtre de 30 secondes. Cette procédure de lissage a été systématiquement appliquée à l'ensemble des dispositifs de mesure, incluant à la fois les capteurs à bas coût et le capteur de référence Fidas Frog.

Ensuite, l'estimation des biais multiplicatifs et additifs introduits dans le modèle (5.3) en utilisant le capteur de référence a été réalisée. L'étalonnage a été effectué séparément pour chaque capteur afin d'obtenir les meilleures performances possibles pour les différents capteurs sans modifier leur corrélation respective.

Notons $Z'_i(x, t)$ les données brutes provenant du capteur i échantillonné à l'endroit x et au temps t , $F(x, t)$ les données provenant du Fidas® frog aux mêmes endroits spatio-temporels (x, t) et $Z_i(x, t)$ les données calibrées provenant du capteur i échantillonné à l'endroit x et au temps t .

Un étalonnage linéaire des données brutes, destiné à corriger un éventuel

biais, est défini comme suit :

$$Z_i(x, t) = a_i + b_i \cdot Z'_i(x, t) \quad (5.4)$$

Dans l'équation (5.4), le seul terme connu est $Z'_i(x, t)$. L'estimation de a_i (biais additif) et de b_i (biais multiplicatif) est nécessaire pour obtenir les données calibrées. L'estimation de a_i et b_i implique $F(x, t)$:

$$F_m(x, t) = a_i + b_i \cdot Z'_i(x, t) + \epsilon(x, t) \quad (5.5)$$

L'estimation de a_i et b_i à partir de l'équation (5.5) a été réalisée à l'aide de la méthode des moindres carrés ordinaires, qui minimise l'intégrale de ϵ^2 . Cette calibration a été effectuée pour chaque capteur individuellement, en utilisant toutes les données du capteur i issues du tour de mesure où il a été au contact des données de référence F du Fidas® frog. Il y a eu autant d'estimations de a_i et b_i qu'il y a de capteurs.

Donnée de pollution

Nos boîtiers mobiles dotés de capteurs à faible coût peuvent mesurer les concentrations de particules fines telles que PM₁₀, PM_{2.5} et PM₁. Le modèle Fidas® frog, quant à lui, est capable de détecter les PM₁₀, PM₄, PM_{2.5}, et PM₁. Le choix de l'étude s'est porté sur la concentration sur les PM_{2.5}, étant donné que la station d'Air Parif mesure exclusivement ce type de particules.

Il convient de mentionner que les capteurs mobiles à bas coût ainsi que le Fidas® Frog n'ont pas constamment fourni des performances optimales durant les essais, entraînant parfois des relevés de pollution inexacts ou manquants. Néanmoins, les informations relatives au positionnement spatio-temporel seront prises en compte dans la seconde phase de cette étude. Cette intégration permettra d'évaluer l'impact du nombre de capteurs, qu'ils soient fixes ou mobiles, sur la variabilité des prédictions de pollution. Il est important de noter que, contrairement à l'étude précédente, nous avons utilisé les données de concen-

tration des particules directement sans appliquer de transformation logarithmique.

L'expérience s'est déroulée sur 5 tours du quartier. La combinaison entre le capteur de référence et chaque capteur mobile a été déployée de manière séquentielle : le premier capteur lors du premier tour, le deuxième capteur lors du deuxième tour, et ainsi de suite :

1. **Premier Tour (10h30 - 11h15) :**
 - Durée : 45 minutes
 - Capteurs fonctionnels : 5 capteurs mobiles étaient opérationnels.
2. **Deuxième Tour (11h30 - 12h13) :**
 - Durée : 43 minutes
 - Capteurs fonctionnels : Seuls les capteurs 2, 4, et 5 étaient opérationnels.
3. **Troisième Tour (13h30 - 14h12) :**
 - Durée : 42 minutes
 - Capteurs fonctionnels : Les capteurs 2, 3, 4, et 5 étaient opérationnels.
4. **Quatrième Tour (14h30 - 15h11) :**
 - Durée : 41 minutes
 - Capteurs fonctionnels : Les capteurs 2, 3, 4, et 5 étaient opérationnels.
5. **Cinquième Tour (15h30 - 16h15) :**
 - Durée : 45 minutes
 - Capteurs fonctionnels : Les capteurs 2, 3, 4, et 5 étaient opérationnels.

Le Fidas® frog a rencontré des dysfonctionnements lors des premier et troisième tours. De ce fait, les données des capteurs 1 et 3 ont été jugées inutilisables pour la production de cartographies.

La Figure 5.1 illustre les données obtenues avec le capteur 4 comparées aux données de référence prises lors du quatrième tour. Bien qu'il y ait une corrélation notable entre les deux jeux de données, un biais d'amplitude ainsi qu'un biais constant ont été observés. En effet, le capteur mobile à faible coût n'a pas

détecté un pic majeur avec la même intensité que le capteur de référence.

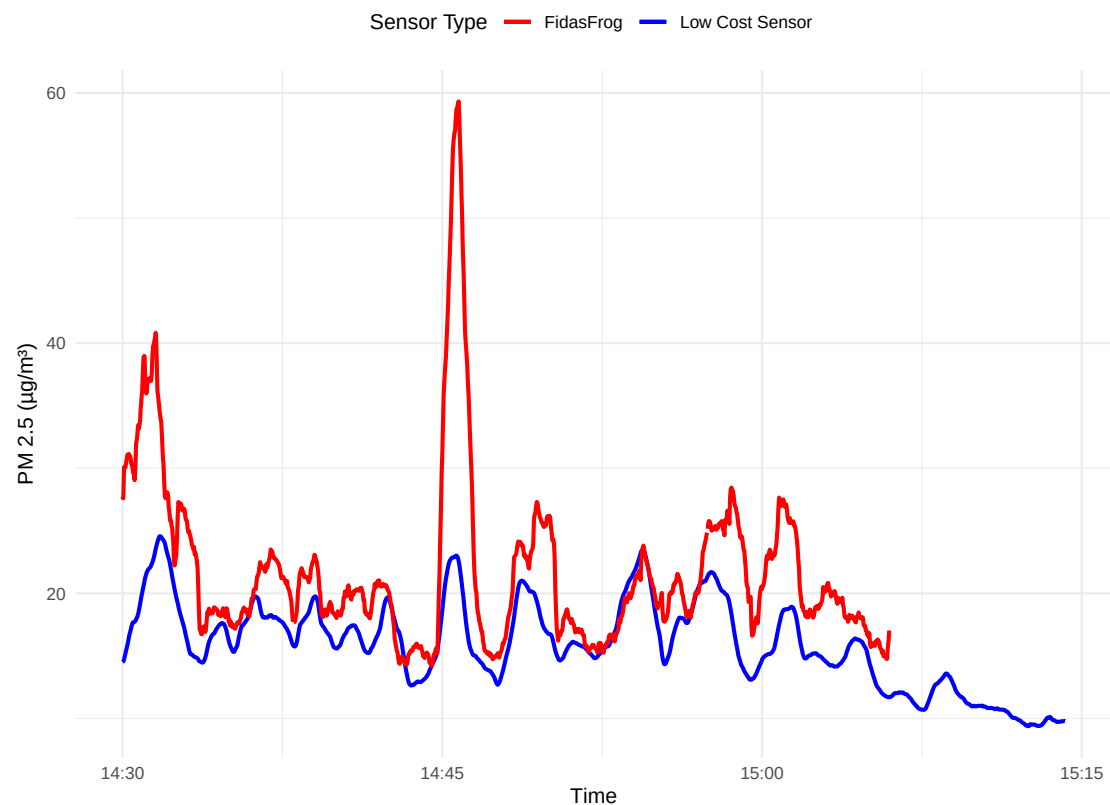


Figure 5.1 – Exemple de données récoltées avec le capteur mobile 4 vs données issues du boîtier de référence durant le tour 4.

Outre ces informations relatives à la pollution, d'autres données seront également intégrées afin d'alimenter le modèle 5.3. La section suivante détaillera ces données.

Données auxiliaires

À l'instar des précédents jeux de données, diverses covariables spatio-temporelles ont été recueillies. Étant donné que l'expérimentation actuelle se déroule à une échelle spatiale plus restreinte (un quartier spécifique plutôt qu'une ville entière, comme dans les études précédentes), l'importance de certaines covariables se

trouve diminuée en raison de leur moindre variabilité spatiale. Le tableau (5.1) récapitule les covariables utilisées au cours de cette recherche.

Variable	Origine	Dimension
Température	Meteofrance	Temporelle
Humidité	Meteofrance	Temporelle
Proximité de : voie primaire	Open Street Map	Spatiale
Proximité de : voie secondaire	Open Street Map	Spatiale
Proximité de : voie tertiaire	Open Street Map	Spatiale
Proximité de : voie résidentielle	Open Street Map	Spatiale
Proximité de : Voie piétonne	Open Street Map	Spatiale
Proximité d'un espace vert	Open Street Map	Spatiale
Nb de bâtiments dans 200m	Open Street Map	Spatiale
Taux d'occupation des rues adjacentes	opendata.paris	Temporelle
Débit trafic routier des rues adjacentes	opendata.paris	Temporelle

Table 5.1 – Résumé des covariables utilisées dans cette étude.

Une des données fondamentales acquises pour cette étude concerne le comptage routier réalisé au sein de la métropole parisienne. Elle apporte un éclairage essentiel sur la dynamique du trafic, un paramètre dont l'influence significative sur les émissions de polluants ne peut être négligée. Au sein du réseau routier de Paris, le suivi du trafic s'effectue principalement au moyen de boucles électromagnétiques intégrées dans la chaussée. Les archives de ces mesures sont accessibles via le portail de données ouvertes de la ville de Paris [6].

Ces données fournissent des informations sur le taux d'occupation et le flux de circulation pour plus de 3000 segments routiers. Toutefois elles ne suffisent pas à capturer entièrement la complexité du trafic parisien. De plus, plusieurs rues de la capitale, y compris celles impliquées dans l'expérimentation, ne figurent pas dans cette couverture.

Deux types de données sont ainsi récoltées :

- Le taux d'occupation, qui correspond au temps de présence de véhicules sur la boucle en pourcentage d'un intervalle de temps fixe (une heure pour

les données fournies). Ainsi, 25% de taux d'occupation sur une heure signifie que des véhicules ont été présents sur la boucle pendant 15 minutes. Le taux fournit une information sur la congestion routière. L'implantation des boucles est pensée de manière à pouvoir déduire, d'une mesure ponctuelle, l'état du trafic sur un arc.

- Le débit est le nombre de véhicules ayant passé le point de comptage pendant un intervalle de temps fixe (une heure pour ces données).

En prenant en compte les points de mesure situés dans le quartier des Halles, il est néanmoins possible de construire deux variables temporelles à partir des données sur les points de mesure adjacents à l'expérimentation. Les emplacements des points de mesure du trafic routier ainsi que les trajectoire des capteur mobiles sont présentés dans la Figure 5.2.

Avec l'ensemble des données rassemblées et les ajustements nécessaires réalisés, l'implémentation du modèle est désormais prête à être entreprise. Cette étape essentielle vise à valider et à évaluer l'efficacité des méthodes méthodologiques choisies. La prochaine section se consacrera à l'exposition détaillée de l'application du modèle ainsi qu'à l'analyse des résultats qui en découlent.

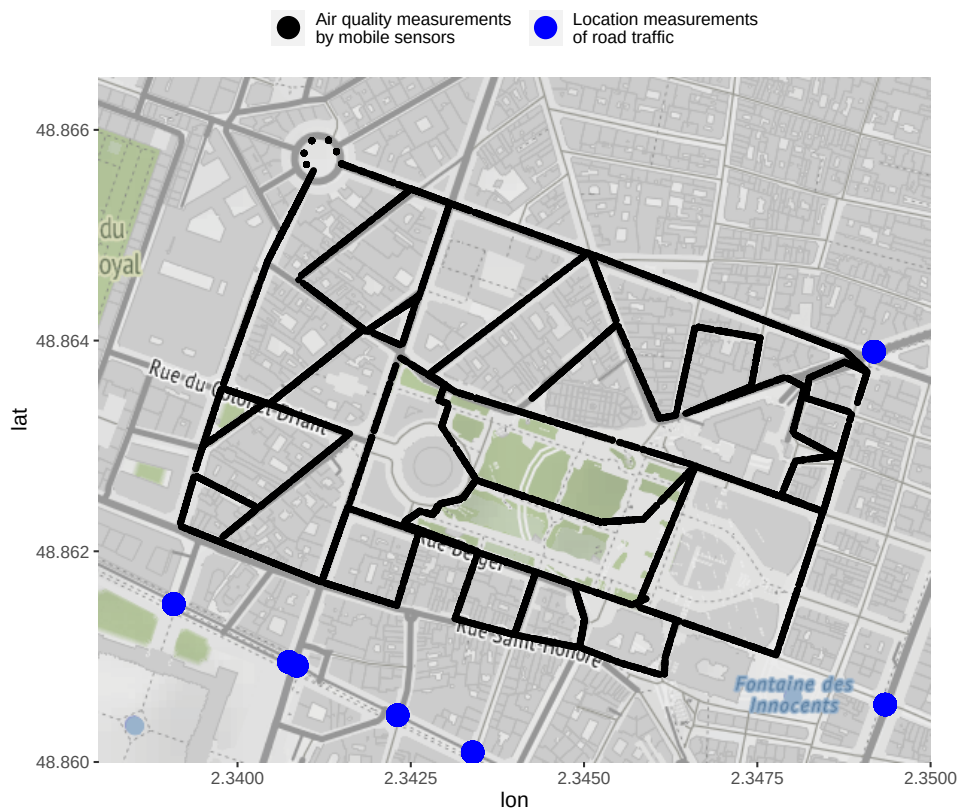


Figure 5.2 – Emplacement des mesures de trafic routier ainsi que des mesure à l'aide de nos boîtier mobiles.

5.2 Évaluation globale : résultats et questions de recherche

Cette section engage une étape importante : la mise en parallèle des résultats empiriques avec les hypothèses initiales. Cette analyse méthodique est cruciale pour juger de la cohérence du modèle par rapport aux hypothèses établies. L'analyse vise à évaluer dans quelle mesure les résultats répondent aux questions de recherche posées au départ de cette thèse.

5.2.1 Création de cartographies

Au cours de cette thèse, l'objectif et moteur principal a été le développement et l'exploration de méthodologies avancées pour créer des cartographies de la pollution atmosphérique détaillées. Cette ambition est née d'un besoin d'obtenir un niveau de détail et de précision supérieur à celui typiquement offert par les modèles conventionnels de dispersion de la pollution.

Pour parvenir à élaborer de telles cartographies précises, l'équation est double. D'une part, il est essentiel de disposer de données denses d'autre part le choix d'un modèle performant est une nécessité. La précision des cartographies dépend davantage de la densité des données que du modèle choisi en lui-même.

La quantité des capteurs dans les ensembles de données exploités pour cette thèse est restreinte. En effet, avec seulement quelques dizaines répartis à l'échelle urbaine, leur nombre dépasse à peine celui des stations de mesure fixes traditionnelles. Or, l'avantage principal des capteurs à bas coût réside dans la possibilité de les déployer massivement, exploitant ainsi leur potentiel pour une surveillance environnementale dense et étendue.

L'une des visions prospectives est de parvenir, à terme, au déploiement de centaines de capteurs dans une ville donnée. Dans cette perspective, l'expérimentation a été menée dans le quartier des Halles avec le déploiement de cinq capteurs, dans le but de simuler une densité accrue de données.

En ce qui concerne la sélection du modèle le plus performant, qui était la première question de recherche posée et introduite dès le chapitre 1, les enseignements issus de la revue bibliographique au chapitre 2 et de la comparaison des modèles traitée au chapitre 3 ont orienté vers l'adoption des modèles géostatistiques.

Ce choix n'est pas fortuit. En effet, le modèle de krigeage avec dérives externes, avec sa moyenne modélisée comme une fonction des covariables, s'ap-

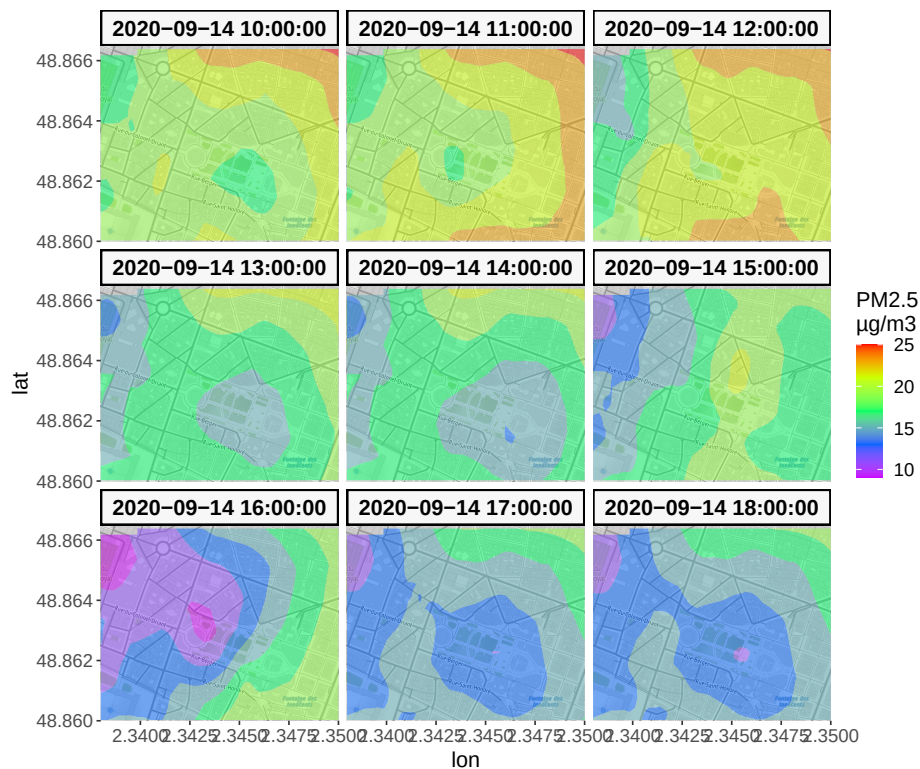


Figure 5.3 – Production d’une cartographie spatio-temporelle des concentrations de PM dans le quartier des Halles en utilisant le modèle hiérarchique (5.3) combinant données issues de capteurs bas coût et de capteur de référence.

puie fortement sur les modèles très en vogue d’utilisation du sol. Ces derniers se révèlent particulièrement aptes à capter les liens dynamiques existants entre la concentration de polluants et leur environnement immédiat. En prenant aussi en compte la corrélation des données entre elles, ce modèle est en mesure d’incorporer également les mesures de pollutions avoisinantes, afin de générer des cartographies détaillées de la pollution atmosphérique.

Les modèles (4.10) et (5.3), présentés respectivement dans les chapitres 4 et 5, approfondissent davantage cette démarche en optimisant l’utilisation des données provenant de capteurs à bas coût par ces types de modèles.

Ainsi, cette technique sophistiquée a été appliquée en exploitant chaque point de données récolté par les capteurs mobiles rigoureusement calibrés. La Figure 5.3 offre une visualisation détaillée, dévoilant la trame complexe des cartes

de la pollution atmosphérique, sur une période de 9 heures.

La méthode a démontré une capacité accrue à discerner et à identifier les zones présentant des concentrations élevées de matières particulaires, soulignant ainsi l'importance de telles cartographies haute résolution dans les considérations d'urbanisme et de santé publique.

Outre la cartographie des polluants, cette approche permet d'aborder les exigences des législations françaises et européennes concernant la délimitation des zones où les valeurs limites de pollution sont dépassées. En privilégiant la représentation sous forme de densité plutôt qu'une valeur ponctuelle, elle offre la possibilité de déterminer la probabilité de dépassement des seuils limites dans des zones spécifiques avec une précision accrue [2].

L'estimation des surfaces en dépassement sert aussi, dans un souci de santé publique, à évaluer le nombre d'habitants soumis à ces dépassements [3].

5.2.2 Pouvoir descriptif du modèle

La cartographie n'est pas le seul avantage de l'utilisation des capteurs mobiles; leur utilité s'étend à une compréhension approfondie de la pollution atmosphérique. En effet, la mesure précise de ce phénomène sera rendue possible grâce à l'augmentation du nombre de ces capteurs. Dans cette optique, le modèle proposé se distingue par sa capacité descriptive. Contrairement à certains modèles d'apprentissage automatique qui peuvent fonctionner comme des "boîtes noires", celui-ci offre une transparence sur les mécanismes sous-jacents. Les paramètres du modèle, tels que les paramètres de covariance, offrent des informations sur la processus de dispersion des polluants. De plus, les paramètres de la moyenne nous fournissent des renseignements précieux sur la manière dont différentes variables environnementales influencent la concentration des polluants dans l'air.

Pour illustrer le pouvoir descriptif de ce modèle, le jeu de données d'Atmo-track à Nantes a été utilisé. Grâce à la disponibilité de nombreux capteurs, les

paramètres de la moyenne et de la fonction de covariance ont été analysés pour mieux comprendre l'influence des variables environnementales sur la pollution.

Description des paramètres de la moyenne

La moyenne du modèle est essentielle pour déterminer les tendances centrales dans les observations. Les paramètres associés à cette moyenne peuvent indiquer l'état général de la pollution dans une région donnée, tout en mettant en évidence les variables qui contribuent le plus à ces niveaux. La compréhension de ces paramètres est donc cruciale pour cerner les principaux contributeurs à la pollution et envisager des solutions ciblées.

Le tableau 5.2 présente les covariables ainsi que l'estimation de leurs paramètres associés. Une première analyse révèle que les covariables ayant une dimension temporelle exercent la plus grande influence sur la tendance globale. Avec une plus grande variance temporelle du phénomène, ces variables aident à expliquer en partie cette variance. Quant aux covariables concernant la distance des rues, comme prévu, elle sont estimées avec des coefficients négatifs : l'éloignement d'une rue est généralement associé à une diminution de la concentration en polluants. Cependant, le type de rue ne semble pas avoir un impact significatif. Les paramètres des autres covariables ne reflètent pas nécessairement les attentes initiales. Pour avoir une compréhension plus approfondie, il serait judicieux de considérer des relations non linéaires dans les analyses futures.

Description des paramètres de covariances

La fonction de covariance traduit le comportement de dispersion des polluants dans l'atmosphère. Le tableau 5.3 détaille l'estimation des paramètres de la fonction de covariance exponentielle spatio-temporelle, incorporant un paramètre spécifique pour saisir l'autocorrélation intra-capteur.

La première observation notable est qu'une part significative de la variance globale est attribuée à la variance intra-capteur.

Variable (unité)	Estimation	Ecart type
Nombre de bâtiments dans un rayon de 500m	0.0031	0.0003
Distance de l'autoroute la plus proche (m)	-0.003	0.0005
Distance du périphérique le plus proche (m)	-0.0020	0.0005
Distance d'une rue primaire la plus proche (m)	-0.0039	0.0005
Distance d'une rue secondaire la plus proche (m)	-0.0038	0.0004
Distance d'une rue tertiaire la plus proche (m)	-0.0036	0.0004
Distance d'une rue résidentielle la plus proche (m)	-0.0031	0.0006
Distance d'un site industriel le plus proche (m)	0.0004	0.0005
Distance d'un feu de circulation le plus proche (m)	0.0016	0.0005
Distance d'un espace vert le plus proche (m)	-0.0041	0.0005
Température °C	-0.0339	1.7552
Humidité (%)	0.5967	0.2487
Sortie modèle CHIMERE PM10 (µg/m ³)	1.0725	1.5561
Altitude (m)	0.0375	0.0155

Table 5.2 – Estimation des paramètres de la moyenne pour le jeu de données d'Atmotrack utilisant le modèle hiérarchique (4.10).

Les valeurs des paramètres de portée spatiale et temporelle dépendent directement des unités utilisées; dans cette étude, ce sont les mètres et les secondes. Une analyse approfondie montre que la variance à l'échelle d'une journée est bien supérieure à celle observée à l'échelle d'une ville entière comme évoqué précédemment. Ainsi, deux observations espacées de 2 heures sont, en moyenne, aussi semblables que deux observations séparées par une distance de 16 km.

Variance $\sigma^2(\mu g/m^3)$	Portée spatiale $\theta_1(m)$	Portée Temporelle $\theta_2(s)$	Pépite $\tau^2(\mu g/m^3)$	Autocorrélation $\gamma^2(\mu g/m^3)$
66.2654	5394.9112	1220.6959	0.0034	35.9327

Table 5.3 – Estimation des paramètres de la fonction de covariance exponentielle spatio-temporelle.

5.2.3 Complexité

Le modèle hiérarchique utilisant un processus gaussien (5.3), malgré ses atouts remarquables, se heurtait à un enjeu majeur : sa complexité. Toutefois, cette difficulté a été traitée grâce à l'adoption de l'approche de Vecchia en matière d'approximation de vraisemblance.

L'approximation de Vecchia transforme les pratiques de calcul en modélisation statistique spatiale grâce à sa compatibilité avec la parallélisation. Cette méthode se distingue par sa capacité à segmenter la fonction de vraisemblance en composantes indépendantes, ce qui simplifie le traitement de grands ensembles de données spatiales. Cette propriété s'harmonise idéalement avec les infrastructures modernes de calcul parallèle. En tirant parti des ressources des architectures informatiques multicœurs et des systèmes distribués, les processus qui intègrent cette approximation peuvent être significativement accélérés, devenant ainsi cruciaux pour les applications exigeant des prédictions en temps réel.

L'exploration de différentes configurations de Vecchia dans le contexte spatio-temporel a confirmé sa capacité à générer des cartographies en temps réel de manière efficiente. Cette méthode ne brille pas seulement par sa rapidité mais aussi par sa précision, comblant ainsi l'écart entre performance et exactitude dans le champ des statistiques spatiales.

À titre illustratif, sur un ordinateur doté d'un processeur Intel Core i7-4810MQ 2.80 GHz comprenant 4 cœurs et 8 threads, l'évaluation de la fonction de vraisemblance maximale pour un ensemble de 50 000 observations est achevée en moins de dix minutes. Bien que des gains supplémentaires soient envisageables en adaptant le code R à C++ à l'aide de bibliothèques spécifiques, les performances actuelles répondent pleinement à nos exigences.

5.2.4 Erreur associée à la prédiction

Comme souligné précédemment, l'un des avantages principaux du modèle développé est sa capacité à fournir une mesure de l'incertitude de prédiction. La Figure 5.5 illustre la variance de cet estimateur. Il est essentiel de souligner que la variance est intrinsèquement liée à l'emplacement des points de données, et non à leurs valeurs. Cette distinction explique pourquoi il n'existe aucune corrélation entre les Figures 5.4 et 5.5.

La variance augmente en s'éloignant de l'endroit où les données ont été collectées et, inversement, diminue en s'en rapprochant. Les emplacements où les mesures ont été prises pendant les heures de pointe sont mis en évidence en rouge sur la carte. L'allure à laquelle la variance diminue et la distance par rapport aux données collectées dépend directement des paramètres de covariance estimés.

Comme prévu, la variance était minimale dans la région centrale du quartier où la majorité des données ont été mesurées. Les positions des trois capteurs sont particulièrement visibles à certains moments, tels que 11h ou 12h.

L'apogée de la variance peut être discernée aux périphéries des cartes, particulièrement visible à 10h et à partir de 17h. Ces cartes ont la particularité d'avoir des données collectées d'un seul côté du temps, ce qui entraîne une grande distance temporelle (début et fin de journée) aux données recueillies, ce qui implique une incertitude accrue.

Un point d'intérêt est la variance observée à 13h. Même s'il existe des points de données avant et après cette heure, leur distance temporelle est relativement grande (>30 mins). Le modèle identifie une variance temporelle assez grande, ce qui rend les observations à cette heure moins informatives concernant la qualité de l'air. Cela se traduit par une augmentation de la variance attendue.

Par conséquent, il est judicieux d'accorder moins d'importance aux cartes de

certaines heures (par exemple, 10h, 13h, 17h et 18h). Ce concept de variance servira à examiner l'impact du nombre de capteurs et de leur type (fixe ou mobile) sur le gain d'information potentiel et la diminution envisageable de la variance.

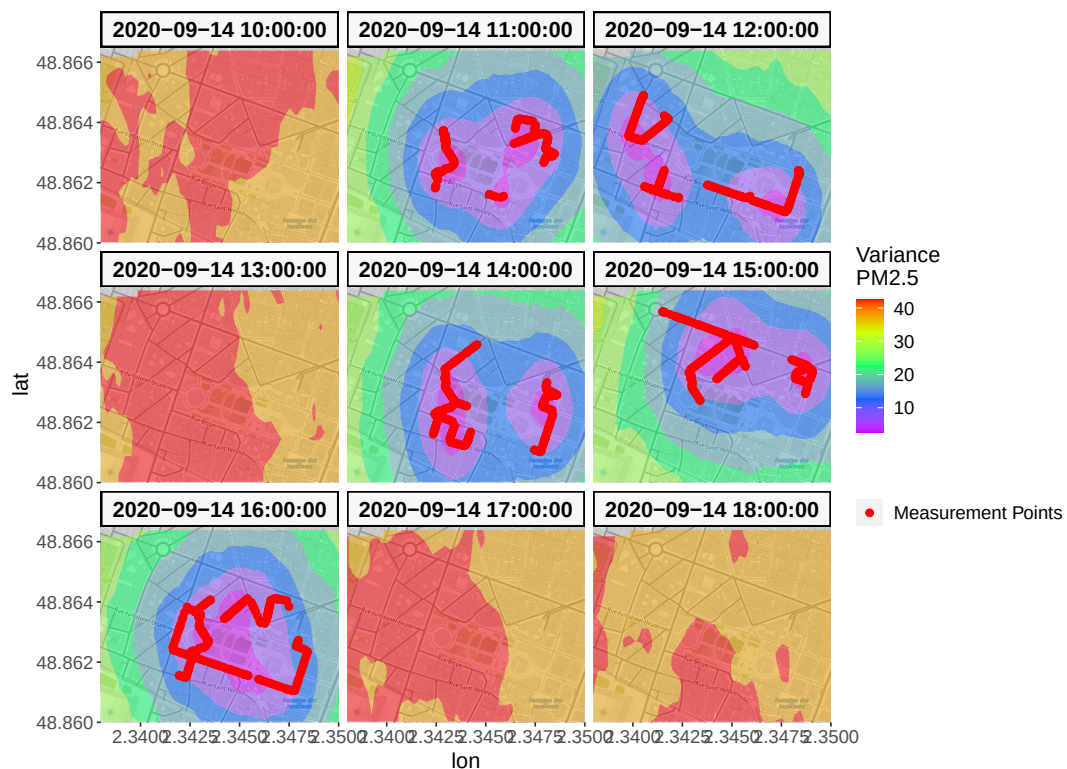


Figure 5.4 – Production d’une cartographie spatio-temporelle des variances des prédictions de PM dans le quartier des Halles en utilisant le modèle hiérarchique (5.3) combinant données issues de capteurs bas coût et de capteur de référence.

5.2.5 Étude du nombre de capteurs

L’utilisation de capteurs mobiles pour surveiller la pollution atmosphérique d’une ville soulève plusieurs questions pertinentes. Une question de recherche fondamentale soulevée en début de ce manuscrit concerne le nombre de capteurs requis pour parvenir à une précision prédictive souhaitée. Une autre question était d’évaluer et comparer une flotte de capteurs mobiles à un réseau de

capteurs fixes en termes d'amélioration des prédictions.

Pour aborder ces interrogations, une expérimentation a été menée afin d'examiner les bénéfices relatifs des capteurs mobiles par rapport aux capteurs fixes, ainsi que pour déterminer l'influence du nombre de capteurs sur la précision des prédictions. En variant le nombre de capteurs mobiles et fixes, l'Erreur Quadratique Moyenne (EQM) a été évaluée dans deux contextes différents : l'interpolation spatiale et la prédiction temporelle à court terme.

Le krigeage étant une méthode sans biais, l'EQM équivaut simplement à la variance prédictive. Cette dernière ne dépend que de la position des observations et des paramètres du modèle. Il est donc possible de l'obtenir sans nécessité de recourir à l'usage de données de pollution.

Durant la phase de collecte de données dans le quartier des Halles avec des capteurs mobiles, les relevés de pollution du premier cycle n'ont pas été inclus en raison d'une défaillance du capteur de référence. Toutefois, les données de trajectoire recueillies lors de cette première session restent précieuses. En effet, l'évaluation de l'erreur quadratique moyenne (EQM) repose sur les trajectoires, et le volume le plus conséquent de ces données (issues de 5 capteurs mobiles) a été enregistré durant ce cycle initial.

Afin de simuler des données comparables à celles issues d'un capteur fixe, des emplacements ont été choisis au hasard pour servir de points de surveillance. Bien que les capteurs fixes mesurent habituellement des moyennes horaires, ces points ont été considérés comme ayant une fréquence d'échantillonnage d'une seconde, alignée sur celle des capteurs mobiles, pour permettre une comparaison plus adéquate.

Posons $Y^* = Y^*(s, t) : s \in S^*, t \in T^*$ comme ensemble de prédiction, où T^* est l'heure de la prédiction et $S^* = s^*1, \dots, s^*n$ sont les $n = 2500$ points formant une grille couvrant le quartier des Halles.

Nous utilisons les paramètres de covariance spatiotemporelle estimés à partir des données analysées pour calculer la variance moyenne de prédiction :

$$EQM = 1/n \sum_i^n var[Y_{T^*}(s_i^*)|Y^J] = C_{ii} - 1/n \sum_i^n \Sigma_{iJ} \cdot \Sigma_J^{-1} \cdot \Sigma_{iJ}^t \quad (5.6)$$

J représente l'ensemble des points d'observation. Cet ensemble sera constitué de points issus d'une collecte à base soit de capteurs mobiles soit de capteurs fixes. C_{ii} représente la variance entre les points de prédiction. Σ_{iJ} la matrice de covariance entre les points de prédiction et les points d'observation. et Σ_J la matrice de covariance entre les points d'observation.

Dans l'équation 5.6, la magnitude de la seconde partie est inversement proportionnelle à la variance moyenne. En d'autres termes, une augmentation de la valeur de Σ_{iJ} ou une diminution de la valeur de Σ_J conduira à une réduction de la variance globale, résultant ainsi en une prédiction améliorée. Il est important de noter que la fonction de covariance est une fonction décroissante de la distance spatio-temporelle. Cela signifie que la proximité spatio-temporelle des points d'observation par rapport aux points de prédiction est cruciale pour la qualité de la prédiction. En effet, plus les points d'observation sont rapprochés spatio-temporellement des points de prédiction, plus précise sera la prédiction. De plus, plus les points d'observation sont éloignées les uns des autres, couvrant ainsi une plus grande étendue spatio-temporelle, plus la variance finale tend à être plus faible, ce qui améliore encore la qualité de la prédiction.

Rappelons que, pendant le premier tour, les observations couvraient la période de 10h30 à 11h15. Deux types de prédiction sont comparés :

- Une interpolation spatiale : 2500 points de prédiction à 11h ont été choisis.
- Une prévision temporelle à 45 min : 2500 points de prédiction à 12h ont été choisis.

Les résultats sont présentés dans la figure 5.5. La variance dans le scénario de prévision est identique, quels que soient le type de capteur et leur nombre. Cette observation s'explique par le fait que, même avec une différence temporelle de seulement 45 minutes, la variation des données sur l'axe temporel est si significative qu'il devient difficile voir impossible de tirer des informations pertinentes sur l'état de la qualité de l'air. En conséquence, la variance atteint sa valeur maximale.

En revanche, dans le scénario d'interpolation, on observe une tendance distincte : à mesure que le nombre de capteurs (qu'ils soient mobiles ou fixes) augmente, la variance moyenne diminue pour finalement se stabiliser. Il est intéressant de noter un léger avantage en faveur des capteurs mobiles. Cela est cohérent avec l'idée que le déploiement d'un nombre accru de capteurs mobiles offre une couverture spatiale plus étendue, réduisant ainsi l'incertitude associée.

5.3 Conclusion

La présente étude a mis en évidence l'efficacité du nouveau modèle. Cependant, pour en tirer pleinement parti, il est impératif de mettre en place le plan expérimental correspondant. Le jeu de données collecté aux Halles se révèle essentiel pour estimer les paramètres de ce modèle novateur.

Ce modèle répond à l'ensemble de nos questions de recherche. Il tient compte de la nature bruitée des données provenant de capteurs peu coûteux, utilise des données plus fiables pour corriger les biais, et offre ainsi des cartographies de plus en plus précises.

Il fournit un cadre descriptif permettant d'analyser l'impact des différentes variables environnementales sur le processus d'intérêt étudié. L'application de l'approximation Vecchia, étudiée précédemment, contribue à réduire la complexité du modèle. Enfin, cet outil se révèle essentiel pour examiner l'influence de l'emplacement et de la mobilité des capteurs sur la précision des prédictions.

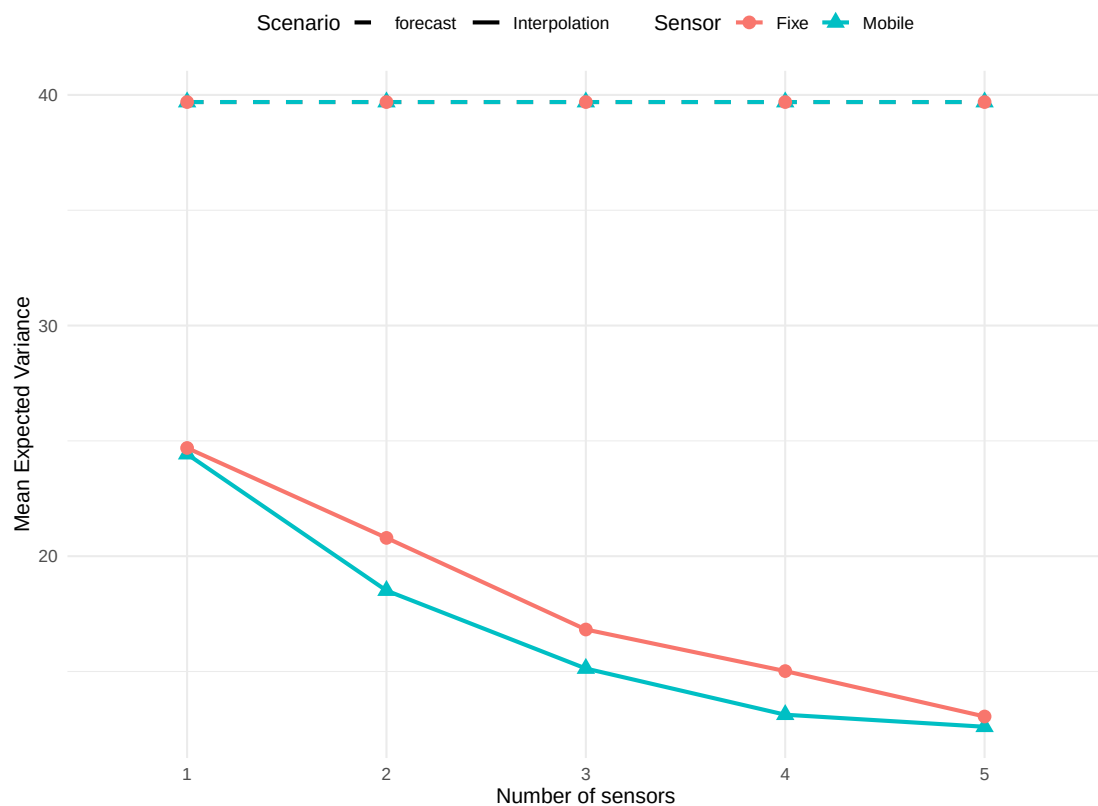


Figure 5.5 – Étude de l'impact du nombre et du type de capteurs sur la variance de la prédiction.

La formule d'incertitude des prédictions du modèle géostatistique offre un moyen d'examiner la manière dont l'incertitude peut être minimisée en fonction de la localisation et du ratio de capteurs fixes et mobiles à bas coût. Une fois l'estimation des paramètres terminée, cette méthode ne nécessite pas de données supplémentaires et permet une analyse théorique de l'influence du ratio fixe/mobile sur la précision prédictive. Toutefois, il est essentiel de noter que cette méthode a ses limites. Le modèle géostatistique est une approximation des dispersions réelles du processus physique, et s'appuyer sur un modèle simplifié pour déduire la configuration optimale des capteurs pourrait s'avérer trompeur. De plus, cette méthode ne considère pas la contribution des capteurs à l'estimation des paramètres du modèle, en particulier ceux des covariables.

Références

- [1] Adrian Arfire, Ali Marjovi et Alcherio Martinoli. "Model-based rendezvous calibration of mobile sensor networks for monitoring air quality". In : *2015 IEEE SENSORS*. IEEE. 2015, p. 1-4.
- [2] Maxime Beauchamp et al. "A necessary distinction between spatial representativeness of an air quality monitoring station and the delimitation of exceedance areas". In : *Environmental monitoring and assessment* 190 (2018), p. 1-27.
- [3] Claudio Gariazzo, Armando Pelliccioni et Andrea Bolignano. "A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic". In : *Atmospheric environment* 131 (2016), p. 289-300.
- [4] Balz Maag, Zimu Zhou et Lothar Thiele. "A survey on sensor calibration in air pollution monitoring deployments". In : *IEEE Internet of Things Journal* 5.6 (2018), p. 4857-4870.
- [5] Palas. *Fidas Frog*. Consulté le 1 octobre 2023. 2023. url : <https://www.palas.de/en/product/download/fidasfrog/datasheet/pdf>.
- [6] OpenData Paris. *Comptage routier - Données trafic issues des capteurs permanents*. Consulté le 1 octobre 2023. 2023. url : https://opendata.paris.fr/explore/dataset/comptages-routiers-permanents/export/?disjunctive.libelle&disjunctive.etat_trafic&disjunctive.libelle_nd_amont&disjunctive.libelle_nd_aval&sort=t_1h.
- [7] Jianbang Xiang et al. "Using Vehicles' Rendezvous for In Situ Calibration of Instruments in Fleet Vehicle-Based Air Pollution Mobile Monitoring". In : *Environmental Science & Technology* 54.7 (2020), p. 4286-4294.

Chapitre 6

Conclusion générale

6.1 Synthèse des travaux de recherche

Après une présentation détaillée de la modélisation statistique de la pollution atmosphérique à partir de mesures réalisées par des capteurs mobiles, un résumé des travaux est présenté, soulignant les points saillants de cette thèse.

Le sujet de la thèse a été introduit dans le chapitre 1, définissant la pollution atmosphérique, les principaux polluants, ainsi que leurs effets néfastes sur la santé. La surveillance de la qualité de l'air est une étape cruciale dans les efforts de réduction et de lutte contre ce fléau. Avec les avancées technologiques, plusieurs entreprises ont développé des capteurs miniaturisés de pollution, à la fois petits et abordables. Ces nouveaux outils ont suscité un intérêt croissant et sont devenus le centre de nombreuses recherches scientifiques. Ils offrent une nouvelle méthode pour cartographier la qualité de l'air avec une précision accrue, grâce à la possibilité de déployer un grand nombre de ces capteurs, qu'ils soient fixes ou mobiles.

Ensuite, le cadre de la problématique de recherche a été établi, soulignant les défis associés et définissant les questions de recherche. Cette étape a permis de clarifier les objectifs du modèle de reconstitution : il doit être performant pour la cartographie, descriptible, peu complexe et fournir des informations précises

sur la qualité de la prédiction. De plus, ce modèle doit permettre de répondre à des questions secondaires telles que l'emplacement optimal des capteurs, leur nombre ou leur caractère mobile.

Le chapitre 1 s'est clôturé par l'établissement d'une méthodologie rigoureuse destinée à répondre à ces interrogations. Cette méthodologie s'articule autour de trois étapes : une revue de la littérature, une analyse comparative des modèles présentés dans cette revue, et enfin l'amélioration des modèles existants.

Le chapitre 2 s'est centré sur les méthodes actuelles utilisées pour cartographier la qualité de l'air. Il a été observé que les modèles de dispersion physico-chimique prédominent, en particulier à grande échelle (niveau national ou continental). Ces modèles nécessitent des données précises et sont généralement utilisés parallèlement à des mesures de stations fixes gouvernementales de surveillance de la qualité de l'air.

La suite a impliqué le recensement des publications scientifiques utilisant des données provenant de capteurs bas coût, naturellement moins précises. Les cartographies générées à partir de ces données sont souvent à une échelle réduite et sont construites à l'aide d'approches empiriques. Contrairement aux modèles de dispersion, elles explorent les relations entre les variables environnementales plutôt que de les intégrer a priori dans des équations physico-chimiques.

Après cette revue de la littérature, trois familles de modèles susceptibles de répondre à la problématique de recherche ont été identifiées : les modèles qui utilisent des techniques d'interpolation déterministes, la géostatistiques et les modèles basés sur l'utilisation des sols. Ces derniers modèles englobent à la fois les modèles de régressions et les modèles d'apprentissage automatique.

Au cours du chapitre 3, une comparaison de ces trois familles de modèles a été réalisée en les appliquant à un jeu de données concret, examinant les forces et les faiblesses de chacune. Pour ce faire, une approche de comparaison graduelle a été adoptée, augmentant progressivement la complexité des modèles

afin d'évaluer le lien qui existe entre la complexité des modèles et la qualité de leurs prédictions.

Au cours de cette exploration graduelle, la première étape a consisté à comparer les méthodes déterministes aux méthodes géostatistiques, en se basant sur le jeu de données OpenSense à Zurich. Ces deux familles de méthode ont la spécificité de pouvoir produire des cartographies sans avoir recours à des variables externes (telles que la météo, le trafic ...). De cette étude, il ressort une supériorité de la géostatistique qui surpasse les méthodes d'interpolation dans presque tous les critères évoqués dans le premier chapitre. Elle a toutefois le désavantage d'être plus complexe.

La seconde étape de cette comparaison a mis en avant des modèles capables d'intégrer des données autres que celles liées à la pollution atmosphérique, ceci afin d'affiner et ajuster la prédiction. Un jeu de données complet d'AtmoTrack a été utilisé, incluant à la fois des capteurs mobiles et fixes bas coût. Pour évaluer les performances des modèles, une dizaine de covariables ont été recueillies (spatiales, temporelles ou spatio-temporelles) ainsi que les données issues des stations gouvernementales de surveillance de la qualité de l'air. Cette étude a permis de déterminer quel modèle, parmi les dix examinés issus des trois familles précédemment évoquées, était le plus apte à fournir les meilleures performances, et dans quelles circonstances.

De cette étude, plusieurs conclusions majeures ont été dégagées. Au sujet de la fiabilité des données issues des capteurs bas coût, une discordance significative a été observée entre les résultats des modèles et les données précises provenant des stations fixes. En outre, les défis liés à l'incertitude des cartographies dans des scénarios de prévisions ont été mis en évidence.

Les modèles géostatistiques spatio-temporels, estimés par la méthode du maximum de vraisemblance, ainsi que ceux basés sur l'apprentissage automatique, ont démontré une supériorité de performance par rapport aux autres modèles. Toutefois les méthodes d'apprentissage automatique ont donné des

résultats légèrement meilleurs que les autres méthodes dans les situations de prévision temporelle.

Les modèles géostatistiques se sont avérés particulièrement prometteurs pour répondre aux problématiques de recherche. Ils ont offert une qualité de prédiction généralement supérieure aux autres modèles, tout en fournissant une mesure d'incertitude pour la prédiction. De plus ils proposent un cadre descriptif, ce qui est un avantage par rapport aux méthodes d'apprentissage automatique. Malheureusement leur complexité constitue un inconvénient majeur car elle les rend moins adaptés pour traiter de grands ensembles de données.

À partir de ces constats, le choix s'est porté sur les modèles géostatistiques dans la suite de la thèse. L'ambition est de traiter simultanément les problématiques liées à la qualité des données de mesure et à la complexité inhérente à ces modèles.

Le chapitre 4 a abordé ces deux défis majeurs : la modélisation des erreurs associées aux capteurs bas coût et la gestion de la complexité croissante du modèle. Ce chapitre s'intègre dans l'étape d' « amélioration de l'existant » de la méthodologie.

Le modèle géostatistique traditionnel, basé sur un processus gaussien avec une moyenne définie comme une combinaison linéaire de certaines covariables (communément appelé krigeage avec dérives externes), a été modifié. Une approche hiérarchique inspirée des modèles mixtes a été adoptée pour introduire une composante modélisant les erreurs inhérentes à l'utilisation des capteurs à bas coût. Le résultat est un modèle hiérarchique à trois niveaux, capable non seulement de traiter le processus central de dispersion des polluants, mais également de traiter l'autocorrélation des capteurs bas coût en la considérant comme un biais constant propre à chaque capteur. Cette innovation est généralisable à toute collecte de données environnementales avec des capteurs potentiellement biaisés.

L'intégration de cette composante de biais rend le modèle plus complexe. Afin de limiter les répercussions de cette complexité accrue, des méthodes d'approximation de la fonction de vraisemblance ont été explorées, et l'approche de Vecchia a été choisie. Cette technique décompose la fonction de vraisemblance en éléments indépendants facilitant ainsi grandement son traitement. De plus cette décomposition s'adapte parfaitement aux architectures de calcul parallèle ce qui accélère sensiblement la rapidité des traitements.

Pour évaluer cette méthode, un jeu de données simulées a été généré, basé sur une collecte fictive de données biaisées, en utilisant des trajectoires spatio-temporelles réelles des capteurs mobiles d'Atmotrack. Cela a permis d'évaluer l'efficacité des différentes configurations de l'approximation de Vecchia, notamment en ce qui concerne l'ordre d'observation, le choix de distance et le nombre de voisins. Les résultats ont souligné la supériorité de l'arrangement basé sur l'algorithme de distance min-max, ainsi que la configuration aléatoire.

En conclusion du quatrième chapitre, une application numérique du modèle révisé a été réalisée, avec l'estimation de ses paramètres en utilisant les configurations optimales de l'approximation. Cet exemple a illustré la supériorité du modèle proposé par rapport au modèle géostatistique traditionnel.

Dans le chapitre 5, toujours dans une quête d'optimisation et la recherche du modèle idéal, une méthode pragmatique de collecte des données a été introduite, permettant de dépasser les contraintes théoriques du modèle exposé dans le chapitre précédent. Ce chapitre présente un modèle qui intègre un biais additif et multiplicatif, typiquement observés dans les capteurs à bas coût. Ce modèle ne se contente pas de représenter ; il fournit un cadre d'analyse optimal, mettant en lumière la forte variabilité du phénomène étudié, notamment sur le plan temporel. Il éclaire également les liens entre différentes variables environnementales et la dispersion des polluants. De plus il permet d'analyser l'impact du nombre et de la nature des capteurs (qu'ils soient mobiles ou fixes) sur la qualité des prédictions. Les résultats montrent que les capteurs mobiles sont plus performants pour minimiser l'erreur quadratique moyenne. Cepen-

nant, en matière de prévision, le phénomène se révèle extrêmement versatile, rendant sa prédiction complexe, que l'on utilise un ou cinq capteurs, mobiles ou fixes.

Ce présent et dernier chapitre a pour objectif de conclure cette thèse.

6.2 Conclusions et perspectives

La pollution atmosphérique est un fléau majeur pour la santé publique. La mise en place de cartographies précises et l'identification des zones dépassant les seuils réglementaires fixés par la France et l'Union Européenne sont essentielles pour protéger la santé des citoyens.

Au fil des chapitres de cette thèse, plusieurs aspects de la modélisation de la pollution atmosphérique ont été explorés : les incertitudes associées à la prédiction et les contributions potentielles des capteurs mobiles. L'étude a couvert des aspects tant méthodologiques que pratiques.

La réalisation de prédictions de pollution atmosphériques précises est une tâche complexe, particulièrement lorsqu'il s'agit de prévisions à court ou long terme. Même lorsque les variables temporelles sont connues à l'avance, comme les prévisions météorologiques, des incertitudes demeurent. Dans ce contexte, l'essor des algorithmes d'apprentissage automatique et de l'intelligence artificielle s'avère prometteur. Si leur puissance prédictive est indéniable, l'étude comparative n'a pas toujours démontré leur supériorité.

Les modèles de chimie-transport, en raison de contraintes de puissance de calcul et de résolution spatiale des données d'entrée, ne rendent pas compte des variations de concentration à petite échelle.

La géostatistique s'est avérée être une méthode robuste pour la production de cartographies à partir de mesures in-situ. Les processus de dispersion des polluants étant de nature complexe, l'estimation de la moyenne locale est un élément clé de la prédiction. Le krigeage avec dérives externes, qui considère la moyenne locale de la concentration comme une fonction des sorties d'un modèle de chimie-transport et d'autres variables environnementales, est particulièrement adapté aux données sur la qualité de l'air.

Les capteurs mobiles à bas coût, grâce à leur capacité à couvrir une vaste

étendue spatiale, se sont imposés comme des outils précieux pour estimer les paramètres des variables spatiales des modèles géostatistiques. Ils offrent également l'avantage de fournir des mesures in situ plus nombreuses et dispersées. Toutefois ils ne sont pas exempts de défauts. La probabilité importante de dysfonctionnement et l'omniprésente question de la fiabilité des données, demeurent des défis à surmonter.

Face à ces enjeux, ce travail s'est focalisé sur la création de modèles novateurs capables de tenir compte de l'erreur intrinsèque des capteurs bas coût. La fiabilité des données est, à ce jour, le principal obstacle à l'adoption massive de ces capteurs. Cela souligne l'importance des travaux menés dans cette thèse visant à élaborer des méthodes spécifiquement conçues pour adresser cette problématique.

Cette thèse a abouti à l'élaboration de deux modèles hiérarchiques conçus pour améliorer la gestion des données bruitées issues de capteurs mobiles à bas coût. Le premier modèle s'attache à la modélisation de l'autocorrélation des erreurs des données de capteurs, tandis que le second vise à affiner les prédictions en intégrant des observations de plus grande précision. Cependant, l'application efficace du second modèle nécessite un ensemble de données adéquat.

Pour les recherches futures, les efforts pourraient se concentrer sur l'amélioration du modèle développé au cours de cette thèse. Par exemple, en incorporant d'autres types d'erreurs de capteurs identifiées dans le modèle des données, ou en modélisant des relations non linéaires entre les covariables et la variable d'intérêt.

Une autre voie d'amélioration des prédictions réside dans l'acquisition de données plus pertinentes. Par exemple, l'utilisation des sorties issus de modèles physiques urbains, de données de trafic, etc. Le modèle proposé est entièrement compatible avec l'intégration et l'exploitation de ces types de données.

Pour l'avenir, un défi majeur est la création de nouveaux modèles hybrides

entre modèle physique et géostatistique, d'une manière plus innovante que de simplement utiliser les sorties des modèles physiques comme entrées dans un modèle géostatistique. L'élaboration d'un modèle de chimie-transport stochastique, qui respecte le comportement et les statistiques du modèle basé sur des simulations antérieures, serait une avancée notable. Ce méta-modèle pourrait ainsi générer des cartographie précise avec un coût de calcul moindre. Les avancées récentes dans la résolution d'équations aux dérivées partielles stochastique, dont la fonction de covariance spatio-temporelle est non séparable, peut en théorie tenir compte de l'advection sous-jacente liée au processus de dispersion des polluants. Toute la difficulté est de modéliser, simplifier et paramétrer efficacement la physique au sein de ces simulations.

Outre ces efforts de modélisation, l'accent devrait être mis sur l'amélioration des capteurs eux-mêmes. La calibration, notamment, pourrait se révéler être la clé pour augmenter la précision des capteurs à bas coût. Car, au-delà des modèles, si les données sont biaisées, leur utilité est largement compromise.