Algorithmes d'optimisation pour la surveillance et l'estimation de la pollution de l'air

Khaoula Karroum

ROYAUME DU MAROC
Université Mohammed V - RABAT - جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم

THÈSE EN COTUTELLE PRÉSENTÉE

En vue de l'obtention du :

Doctorat

DE L'UNIVERSITÉ MOHAMMED-V DE RABAT

FACULTÉ DES SCIENCES DE RABAT

CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

Structure de recherche : Laboratoire de Recherche Informatique et Télécommunications

(LRIT-CNRST, URAC 29)

Discipline : Sciences de l'ingénieur

Spécialité : Informatique et Télécommunications

ET

L'UNIVERSITÉ DU LITTORAL COTE D'OPALE

ÉCOLE DOCTORALE SCIENCES, TECHNOLOGIE, SANTÉ

Structure de recherche : Laboratoire de Physico-Chimie de l'Atmosphère (EA 4493)

Discipline : PHYSIQUE. Milieux dilués et optiques fondamentales

Soutenue le 22/02/2021

Par **Khaoula KARROUM**

Algorithmes d'optimisation pour la surveillance et l'estimation de la pollution de l'air

Devant le jury composé de :

Loubna ECHABBI	Professeur de l'enseignement supérieur, INPT, Rabat-Maroc	Rapportrice
Hervé RIVANO	Professeur des universités, INSA de Lyon-France	Rapporteur
Abderrahim EL QADI	Professeur de l'enseignement supérieur, ENSAM, Rabat-Maroc	Président/Rapporteur
Egor DMITRIEV	Expert, IMN-RAS, Moscou-Russie	Examinateur
Yann BEN MAISSA	Professeur Habilité, INPT, Rabat-Maroc	Encadrant
Anton SOKOLOV	Maître de conférences, ULCO, Dunkerque-France	Encadrant
Mohamed EL HAZITI	Professeur de l'enseignement supérieur, ESTS, Rabat-Maroc	Directeur de thèse
Hervé DELBARRE	Professeur des universités, ULCO, Dunkerque-France	Directeur de thèse

Remerciements

Cette thèse a été préparée dans le cadre d'une cotutelle entre le Laboratoire de Recherche en Informatique et Télécommunications (**LRIT**) à la Faculté des Sciences de Rabat (**FSR**) de l'université Mohamed V (**UMV**) sous la direction de Pr. ElHaziti et l'encadrement de Pr. Ben Maissa, et le Laboratoire de Physico-Chimie de l'Atmosphère (**LPCA**) à l'université du Littoral Côte d'Opale (**ULCO**) de Dunkerque sous la direction de Pr. Delabarre et l'encadrement de Dr. Sokolov.

Je remercie, tout d'abord, bien vivement mon directeur de thèse ; **Monsieur Hervé Delbarre** (Professeur des universités à l'ULCO, Dunkerque), de m'avoir proposé et initié à cette cotutelle, pour son accueil chaleureux au sein du LPCA en tant que directeur du laboratoire, de sa confiance en moi pour mener et réaliser ce travail et de m'introduire au monde de la physique atmosphérique. J'ai beaucoup appris à vos côtés, merci pour vos conseils et votre disponibilité tout au long de ces trois ans.

Je remercie vivement aussi **Monsieur Mohamed El Haziti** (Professeur de l'enseignement supérieur à l'ESTS), mon co-directeur de thèse, pour sa bonne humeur, pour toutes les discussions scientifiques, les conseils et surtout votre confiance et encouragements, sans votre aide ce travail n'aurait jamais été réalisé.

J'adresse également mes sincères remerciements à mon encadrant **Monsieur Yann Ben Maissa** (Professeur Habilité à l'INPT, Rabat), d'avoir accepté ma proposition de candidature comme doctorante au sein du LRIT et d'avoir confiance en moi et mes compétences tout au long de ce travail. Je vous remercie aussi pour votre disponibilité permanente, soutien continu et échanges scientifiques.

Je tiens à exprimer ma grande gratitude et précieux remerciements à mon encadrant **Monsieur Anton Sokolov** (Maître de conférences à l'ULCO, Dunkerque), de m'avoir accueilli la première fois quand je suis arrivée à Dunkerque. Ce travail n'aurait jamais été réalisé sans votre aide, patience, disponibilité permanente vos conseils scientifiques et votre ouverture d'esprit pour améliorer de ce travail.

Je tiens à exprimer tous mes remerciements et gratitude aux membres du jury qui ont accepté de juger ce travail. Je tiens à remercier **Madame Loubna Echabbi** (Professeur de

l'enseignement supérieur à l'INPT), **Monsieur Hervé Rivano** (Professeur des universités à l'INSA de Lyon) et **Monsieur Abderrahim EL QADI** (Professeur de l'enseignement supérieur à l'ESTS) pour avoir accepté d'être les rapporteurs de cette thèse. Je remercie également **Monsieur Egor DMITRIEV** (expert à l'IMN-RAS, Moscou-Russie) pour avoir accepté d'être l'examinateur de ce travail.

Je tiens également à remercier **ATMO Hauts-de-France**, pour la transmission des données utilisées dans ce travail.

Mes remerciements vont aussi aux **membres et collègues du LRIT** et **UMV**, parmi lesquels il y a ceux que je connais depuis ma première année universitaire. Je serai toujours heureuse et honorée d'avoir partagée ce long chemin avec vous et en votre présence, soyez surs que vous avez contribuez à la réalisation de ce travail et je vous dois tant. Je voudrai remercier aussi les **membres et collègues du LPCA** et de la **MREI2**, c'était un plaisir d'avoir partagé mon quotidien de travail avec vous lors de mes séjours à Dunkerque.

Je ne peux oublier **mes ami(e)s doctorant(e)s** de l'**ULCO**, qui ont été pour moi d'inoubliables compagnons de route, de vrais supporters et une deuxième famille en France. Je ne me vois pas m'adapter à la vie si facilement à Dunkerque sans votre aide et accueil chaleureux. Vos diverses nationalités et backgrounds me fascinaient chaque jour que je passais avec vous. Grâce à vous, j'ai appris beaucoup de choses. Vous avez contribué énormément à la réalisation de ce travail et je suis très contente d'avoir la chance que vous faites partie de ma vie.

A **ma famille**, grâce à qui ce travail ne pourrait s'accomplir. **Mes parents, mes deux sœurs** et **frère** je ne pourrai jamais trouver les bons mots pour vous remercier assez, pour votre soutien, amour et encouragements continus. Je vous aime beaucoup.

Ce travail a été financé par la **bourse d'excellence** du Centre National pour la Recherche Scientifique et Technique (**CNRST**) au Maroc et par l'Université du Littoral Côte d'Opale (**ULCO**) de Dunkerque, dans le cadre d'une cotutelle entre l'**UMV** (**LRIT**) de Rabat et l'**ULCO** (**LPCA**).

Valorisations scientifiques

1.Publications scientifiques

i. Khaoula Karroum, Yijun Lin, Yao-Yi Chiang, Yann Ben Maissa, Mohamed El Haziti, Anton Sokolov, and Hervé Delbarre. "A Review of Air Quality Modeling." MAPAN (2020) : 1-14. DOI :10.1007/s12647-020-00371-8

ii. Khaoula Karroum, Anton Sokolov, Yann Ben Maissa, Mohamed El Haziti, and Hervé Delbarre. "Spatial and temporal variability impact on air pollution interpolation : a case study on PM_{10} estimation in northern-france. " Pollution Research (2020) (Article accepté)

iii. Khaoula Karroum, Anton Sokolov, Yann Ben Maissa, Hervé Delbarre, and Mohamed El Haziti."A Topology Optimization approach for accurate air pollution interpolation : the case of Northern France air quality monitoring network" (en cours de soumission)

2. Communication orale

Khaoula Karroum , Anton Sokolov, Hervé Delbarre, Yann Ben Maissa, Mohamed El Haziti. " Optimization of sensor placement : A case study of PM_{10} network in Dunkirk. " Proceedings of the 6th World Congress on New Technologies (NewTech'20) Virtual Conference – August, 2020. Paper No. ICEPR 144. DOI : 10.11159/icepr20.144

3. Communication affichée

Khaoula Karroum, Anton Sokolov, Hervé Delbarre, Yann Ben Maissa, Mohamed El Haziti. " Spatial and temporal analysis of aerosol distribution by interpolation techniques on regional scale" 5ème journée scientifique CAPPA, France, Lille, 06 Mars 2019.

Résumé

La mise en œuvre d'un système de surveillance de la qualité de l'air nécessite la prise en considération de phénomènes météorologiques complexes, de sources d'émission variées et des limites induites par les équipements coûteux.
Les trois principales contributions de cette thèse concernant la surveillance et l'estimation de la pollution de l'air sont : une revue des techniques d'estimation de la qualité de l'air, une étude de l'influence de la variabilité spatiale et temporelle de la pollution de l'air sur la précision des méthodes d'interpolation, ainsi qu'une proposition de méthode d'optimisation d'un réseau de surveillance de la qualité de l'air.
Les données de mesures et de modélisation de la concentration des particules PM_{10} ont été fournies par ATMO Hauts-de-France. Dans un premier temps, nous avons fait une synthèse bibliographique sur les techniques de modélisation de la qualité de l'air, détaillant leurs avantages et leurs limites dans l'étude de la pollution de l'air. Ensuite, nous avons estimé la pollution de l'air dans la région des Hauts-de-France au moyen de méthodes d'interpolation spatiale. Nous avons ensuite proposé une optimisation de la technique d'interpolation de la pondération à distance inverse (IDW) qui permet d'améliorer le coefficient de détermination (R^2). La précision de l'interpolation se dégrade sur les sites proches des sources d'émission (par exemple en situation industrielle) et exposés à des phénomènes météorologiques locaux (par exemple en zone côtière). Le moyennage des données de PM_{10} à des échelles temporelles pertinentes a permis le filtrage de l'influence de ces phénomènes dans l'interpolation. Le meilleur R^2 obtenu correspond à la période de moyennage de 24 heures, similaire à la durée de périodicité de certains phénomènes météorologiques locaux tels que la brise de mer se produisant dans les zones côtières.
Par ailleurs, nous proposons une approche pour optimiser le réseau de stations de mesure dans l'agglomération de Dunkerque qui minimise l'erreur quadratique moyenne (RMS) de l'estimation de la pollution atmosphérique obtenue par interpolation IDW à l'aide des données d'ADMS (Atmospheric Dispersion Modeling System) et du modèle de panache gaussien. Il a été démontré que la configuration optimisée permet d'obtenir une meilleure estimation de concentration en PM_{10} par rapport au réseau réel des stations de mesure déployé par ATMO. Les stations d'ATMO sont situées à proximité des sources d'émission, tandis que pour la topologie résultante de l'optimisation appliquée à la pollution des sources diffuses (ADMS), les stations sont dispersées sur tout le site d'étude, et pour une pollution canalisée (modèle de panache gaussien), les stations entourent la source d'émission. Enfin, une approche fiable et efficace a été proposée pour améliorer la précision de l'estimation de la pollution atmosphérique dans une zone d'intérêt particulière, telle que les zones résidentielles ou industrielles.

Mots-clés : interpolation, pollution de l'air, météorologie locale, optimisation, réseau de surveillance.

Abstract

The implementation of an air quality monitoring system requires taking in consideration complex meteorological phenomena, sources of emission and limitations drawn by the costly equipment.
The three main contributions made by the present thesis regarding monitoring and estimation of air pollution are : a review of techniques for estimating air quality, influence of air pollution's spatial and temporal variability on precision of interpolation methods, and a suggestion for a possible optimization of Dunkirk air quality monitoring network.
Data of measurements and modeling of PM_{10} concentrations were provided by ATMO Hauts-de-France. Firstly, we did a bibliographic synthesis on Air Quality Modeling (AQM) techniques, detailing their advantages and limits in studying air pollution. Then, we estimated air pollution in the Hauts-de-France region by means of spatial interpolation methods. We proposed an optimization of Inverse distance Weighting (IDW) interpolation technique that allows improving the coefficient of determination (R^2). We noticed that the accuracy of the interpolation degrades at sites nearby emission sources (e.g., industries) and exposed to local meteorological phenomena (e.g., coastal zone). The influence of these phenomena was filtered by averaging the PM_{10} data at different time scales (ranging from one hour to 3 months). The best R^2 obtained corresponded to the 24 hours averaging period, similar to the periodicity of some local weather phenomena such as sea breezes occurring in coastal areas.
Furthermore, we suggest an approach to optimize the network of measurement stations in Dunkirk agglomeration that minimizes root-mean-square (RMS) error of air pollution estimation obtained by IDW interpolation using data of ADMS (Atmospheric Dispersion Modeling System) and the Gaussian plume model. It was shown that the optimized configuration allows obtaining better PM_{10} concentration estimations compared to the real deployed measuring stations network of ATMO.
Actual ATMO stations are located near the emission sources, while for the resulting topology of stations optimization on diffuse sources pollution (ADMS) stations were scattered throughout the region, and for point source pollution (Gaussian plume) stations surrounded the emission source. Finally, a reliable and efficient approach was proposed for improving the accuracy of estimation of air pollution in an area of special interest, such as residential or industrial areas.

Keywords : interpolation, atmospheric pollution, local meteorology, optimization, monitoring network.

Table des matières

Introduction générale

A l'aube du deuxième millénaire, près de la moitié de la population mondiale (48% en 2000) vivait dans les zones urbaines [1], avec une prévision de croissance du nombre de citadins de 2% par an au cours des trois prochaines décennies. La population urbaine, à l'échelle mondiale, a par ailleurs dépassé la population rurale pour la première fois en 2007 [2]. On voit donc que l'urbanisation a été l'un des phénomènes les plus marquants du XXème siècle. La croissance du tissu urbain peut avoir pour origine le mouvement d'une population rurale en quête d'emploi, de meilleures conditions d'éducation ou de la meilleure qualité de services que garantissent les centres urbanisés. Cependant, le principal facteur d'urbanisation est généralement la propre croissance de la population, en particulier dans les pays en voie de développement [3]. Le nombre et la taille des mégapoles (dont la population dépasse 10 millions d'habitants) se sont considérablement accrus au cours de la seconde moitié du XXème siècle, suite à la croissance démographique de plus en plus centrée dans les villes. En 1800, Londres était la seule ville au monde avec une population de plus d'un million d'habitants. Les villes d'au moins 1 million d'habitants sont passées à trois au début du XXème siècle, alors qu'on en compte aujourd'hui plus de 450 [4]. En 1970, il n'y avait que deux mégapoles (Tokyo et New York), alors qu'aujourd'hui il y en a 23 (Pékin, Paris, Mexico, Delhi, Los Angeles, etc.), et on en prévoit 37 à travers le monde d'ici 2025 [5].

Ces régions densément peuplées émettent d'importantes quantités de contaminants gazeux et aérosols dans l'atmosphère et il en résulte une altération de la qualité de l'air et des pics de pollution de plus en plus fréquents. Ces conséquences néfastes ne sont d'ailleurs pas seulement locales, puisque les masses d'air polluées peuvent également être transportées sur de très grandes distances depuis les sources d'émission (sur des milliers de kilomètres) et contribuer à

la pollution de fond hémisphérique globale, selon les conditions météorologiques rencontrées. La distance moyenne de transport du carbone et des autres composants primaires de particules fines (PM) peut atteindre 200 km pour la plupart des mégapoles [6]. Les distances maximales de transport sont nettement plus élevées, avec 25% de ces polluants étant transportées à plus de 2000 km.

Face à l'augmentation alarmante et continue de la pollution atmosphérique, il est devenu indispensable de surveiller et contrôler au quotidien la qualité de l'air que l'on respire, dans le but de générer un environnement plus sain et moins toxique pour les êtres humains, les animaux et les plantes. Dans ce domaine, la recherche actuelle vise à mieux comprendre les interactions entre les émissions, la physico-chimie atmosphérique et la météorologie, en prenant en compte des échelles extrêmement variées, locales, régionales et continentales. Ainsi, une approche expérimentale de la pollution de l'air nécessite de prendre en compte le caractère multi-échelle des phénomènes en jeu dans le déploiement et l'installation d'outils de surveillance de la qualité de l'air, dans le but d'améliorer la qualité de la mesure et une meilleure modélisation prédictive de la pollution de l'air.

Problématiques

La mise en œuvre d'un dispositif de surveillance de la qualité de l'air, dans un contexte urbain et/ou industriel donné, est une tâche ardue compte tenu de la complexité des phénomènes de pollution à court, moyen et long terme, s'agissant notamment de l'évolution des émissions anthropiques, de la variabilité météorologique et des réactions physico-chimiques complexes. A cela s'ajoute le coût de construction et de déploiement d'équipements onéreux limitant le nombre de stations de mesure. Cependant, ce constat doit être modéré, car les progrès métrologiques récents dans le domaine de la pollution de l'air ouvrent des possibilités inédites de déploiement de capteurs en un grand nombre.

Qu'il s'agisse de la mise en œuvre de stations classiques de la qualité de l'air ou de l'élaboration d'un réseau plus dense de capteurs, la complexité des phénomènes de pollution, à l'échelle d'un centre urbain et industrialisé, implique une mise en œuvre de méthodes d'optimisation de l'outil de surveillance, afin de rendre compte au mieux de la pollution de l'air aux meilleures

résolution spatiale et temporelle.

Dans ce but, considérant les moyens de surveillance de la qualité de l'air actuels et à venir dans les prochaines années, on se propose d'aborder, dans ce travail de thèse, les deux grandes problématiques suivantes :

(1) Quelle est l'influence de la variabilité temporelle et spatiale de la pollution de l'air d'un site donné sur la précision des techniques d'interpolation ?

(2) Quelle peut être la topologie optimale d'un réseau de surveillance de la qualité de l'air donné en tenant compte des caractéristiques du site ?

Pour ces 2 problématiques, on s'inspirera de l'exemple de la région très urbanisée des Hauts-de-France, avec une focalisation sur l'agglomération très industrialisée de Dunkerque.

Contributions

Dans ce travail, nous étudions la pollution de l'air via la concentration des particules PM_{10} (c'est-à-dire de diamètre aérodynamique inférieur à 10 µm), pour lesquelles les stations de mesures sont assez nombreuses, d'abord à l'échelle large de la région des Hauts-de-France dans une première partie, puis à l'échelle plus étroite de l'agglomération de Dunkerque. Les trois principales contributions apportées dans cette thèse sont les suivantes :

- **Contribution 1 : Revue sur les techniques d'estimation de la qualité de l'air**

Il existe, dans la littérature, une multitude de techniques de représentation de la pollution de l'air. Un premier travail de synthèse des travaux a été réalisé sur les modèles de la qualité de l'air a (Air Quality Model ; AQM), afin d'en préciser les avantages et les limites, tout en apportant des recommandations selon l'objectif visé et les données d'entrée disponibles. A l'intention des physico-chimistes, il est important de préciser que les travaux abordés ici excluent les modèles déterministes de prévision de la qualité de l'air, c'est-à-dire les modèles de chimie-transport.

- **Contribution 2 : L'influence de la variabilité spatiale et temporelle sur l'interpolation de la pollution atmosphérique : étude de cas de l'estimation des PM_{10} du Nord de la France**

Ce second travail vient mettre l'accent sur l'importance de la prise en compte de la variabilité

temporelle de la pollution de l'air dans la précision des techniques d'interpolation employées dans l'estimation de la concentration des PM_{10} en région des Hauts-de-France. La méthode d'interpolation a pu être adaptée en fonction du « comportement » de la pollution. L'effet d'un lissage des données à plusieurs échelles sur l'interpolation a pu être étudié à plusieurs échelles temporelles en lien avec les phénomènes météorologiques en jeu. Nous avons enfin analysé la sensibilité de cette interpolation à la densité et la perturbation des données traitées.

- **Contribution 3 : Une approche d'optimisation de la topologie de surveillance de la qualité de l'air, pour une interpolation précise de la pollution atmosphérique en Nord de la France**

Les résultats des deux premières contributions nous ont aidé dans le choix et l'élaboration d'une optimisation de la topologie d'un réseau de surveillance de la qualité de l'air réel de l'agglomération de Dunkerque. Deux modèles de dispersion de l'air - Atmosperic Dispersion Modeling System (ADMS) [7] et le modèle de panache gaussien [8], ont été utilisés pour générer les données de vérité de terrain de la concentration des PM_{10}. L'application d'algorithmes d'optimisation a permis de comparer le réseau de surveillance réel d'ATMO Hauts-de-France (https ://www.atmo-hdf.fr), avec les configurations obtenues à l'issue de cette optimisation. De plus, nous avons suggéré une approche pour améliorer la précision d'interpolation dans une région d'intérêt spécifique, qui s'est avérée être une solution simple et efficace.

Le manuscrit s'articule en quatre parties :

- Le premier chapitre présente une introduction générale sur la pollution atmosphérique ; ses sources, le rôle des phénomènes météorologiques dans sa dispersion, et son impact sur la santé et l'environnement. Ce chapitre décrit aussi les outils de surveillance de la qualité de l'air, ainsi que les zones d'étude et les données utilisées.
- Le deuxième chapitre est consacré à une étude comparative sur les techniques d'estimation de la qualité de l'air disponibles, prenant en compte les données disponibles et les mesures possibles pour valider cette estimation. En outre, des recommandations sont fournies pour construire un modèle de la qualité de l'air.
- Dans le chapitre 3, on expose l'impact de la variabilité temporelle de la pollution de l'air sur la précision de l'interpolation au niveau de la région des Hauts-de-France.

• Au chapitre 4, on analyse enfin l'optimisation appliquée aux données PM_{10} générées par des modèles de dispersion de l'air par rapport à la topologie réelle du réseau de mesure « ATMO Hauts-de-France ». On examine aussi une proposition d'optimisation du réseau pour améliorer la précision d'estimation au niveau d'une partie spécifique de la zone d'étude, comme la zone résidentielle de l'agglomération de Dunkerque.

Chapitre 1

Observation de la pollution de l'air

La qualité de l'air locale est liée à de nombreux facteurs tels que les conditions météorologiques responsables du transport et de la dispersion des polluants gazeux et aérosols, la nature et la variabilité des sources, ainsi que les réactions chimiques qu'ils subissent. La complexité de ces facteurs eux-mêmes et de leur couplage rend difficile la prévision de la pollution de l'air. Ce chapitre a pour but, dans un premier temps, d'introduire la notion de pollution de l'air et son origine, les phénomènes météorologiques pertinents, ainsi que son impact sur la santé et l'environnement. Cette introduction nous amènera ensuite à présenter les outils de surveillance de la qualité de l'air, en mettant l'accent sur leur résolution. La fin du chapitre est dédiée à la présentation de la base de données expérimentales utilisée dans le troisième et quatrième chapitre.

1.1 La pollution de l'air

Selon l'Organisation Mondiale de la Santé (OMS), la pollution de l'air représente la contamination de l'environnement intérieur ou extérieur par un agent chimique, physique ou biologique qui modifie les caractéristiques naturelles de l'atmosphère. Cette pollution se caractérise par un mélange complexe de polluants dans l'atmosphère, fortement influencée par des phénomènes météorologiques.

1.1.1 Matières gazeuses et particulaires PM_{10} et leurs sources

En raison des impacts négatifs de la pollution atmosphérique, plusieurs gouvernements ont élaboré leur propre réglementation en vue de contrôler l'émission dans l'atmosphère de polluants potentiellement nocifs. En France, la loi sur l'air et l'utilisation rationnelle de l'énergie de 1996 (LAURE, par son acronyme français) a rendu obligatoire, pour le gouvernement, la surveillance et l'information du public sur l'état de la qualité de l'air dans le pays. Dans ce but, un organisme de surveillance (ATMO) a la charge de mesurer la concentration des polluants suivants règlementés par le Ministère en charge de l'Environnement en France (`https://www.ecologie.gouv.fr/politiques-publiques-reduire-pollution-lair`) :

- Les oxydes d'azotes (NO_x) proviennent en grande partie du trafic routier. Principalement émis sous forme de NO, qui réagit avec l'oxygène, l'ozone ou des espèces radicalaires pour former NO_2 [9].

- Le dioxyde de soufre (SO_2) provient généralement de la combustion de combustibles fossiles contenant du soufre : fioul, charbon. Il se transforme en sulfate ou en acide sulfurique en milieu humide sous l'action de la lumière et des oxydants photochimiques [10].

- Les composants organiques volatiles (COV), notamment les BTEX (benzène, toluène, éthylbenzène et xylène) représentent un élément essentiel dans le cycle de l'ozone troposphérique et dans la formation des aérosols.

- L'Ozone (O_3) est formé dans l'atmosphère par des réactions photochimiques entre l'oxygène et les polluants primaires tels que les COV, les NO_x [11].

- Les métaux lourds (plomb : Pb, cadmium Cd, arsenic : As et nickel : Ni). -Le monoxyde de carbone (CO) contribue à la formation de l'ozone et provient de la combustion incomplète des combustibles et des carburants.

-Les hydrocarbures aromatiques polycycliques (HAP) sont produits par la combustion incomplète de matières organiques comme les carburants et le bois, et peuvent provenir aussi des constituants naturels du charbon et du pétrole.

- La matière particulaire (Particulate Matter en anglais ou PM) PM_{10} et $PM_{2.5}$, plus communément appelée aérosols , est un polluant atmosphérique très important pour ses effets sur la santé [12]. Les PM ne sont pas une entité chimique spécifique mais se réfèrent à une

suspension de solides, de liquides ou d'une combinaison de particules solides et liquides dans l'air [13]. Les PM sont considérés être parmi les six principaux polluants désignés par le « Clean Air Act » américain (1971), mesurés et examinés dans l'élaboration et l'ajustement des normes environnementales et sanitaires. Les travaux de cette thèse portent essentiellement sur des mesures de la concentration des PM_{10} (chapitre 3 et 4).

Les PM_{10} représentent une sous-catégorie des PM, avec un diamètre aérodynamique 10 µm, c'est-à-dire environ un cinquième du diamètre d'un cheveu humain. Le diamètre aérodynamique d'une particule d'aérosol est le diamètre d'une particule parfaitement sphérique de masse volumique égale à 1000 kg m-3, qui a une vitesse de sédimentation égale à celle de la particule d'aérosol mesurée.

Les PM proviennent à la fois de sources anthropiques (principalement liées à la combustion) et naturelles (par exemple les sels marins, la poussière. . .). Les particules primaires, grossières ou fines, se forment directement et sont le plus souvent associées aux sources de combustion, notamment le trafic, l'industrie et le chauffage domestique. Les PM secondaires sont plus fines et se forment dans l'atmosphère grâce à des conversions chimiques et physiques de précurseurs gazeux tels que les oxydes d'azote, les oxydes de soufre et les composés organiques volatils. Les particules primaires affectent généralement les échelles locales, tandis que les particules secondaires affectent les zones régionales [14].

Les sources d'émission des polluants peuvent être classifiées en deux grands types :

- Les émissions canalisées représentent les rejets des industries dans l'atmosphère par toute sorte de conduite, canalisation ou tuyauterie.
- Les émissions diffuses (appelées aussi émissions multi-sources) proviennent de plusieurs sources comme la pollution émise par un réseau routier en milieu urbain, le chauffage domestique, les opérations de déversement de déchets, les activités agricoles, etc.

La modélisation de la pollution issue des sources canalisées et diffuses nécessitent un inventaire des sources et l'établissement de clés temporelles de répartition, ce qui la rend particulièrement complexe.

1.1.2 Pollution de l'air et la météorologie à variabilité multi-échelles

L'évolution des polluants dans l'atmosphère dépend en partie des phénomènes météorologiques qui jouent notamment un rôle important dans le transport ou l'accumulation des polluants du fait de l'advection, transport horizontal des masses d'air, et de la convection, mouvement vertical des masses d'air [15]. Ainsi, les caractéristiques spatiales et temporelles temporelle d'un champ de pollution sont liées aux échelles spatiales et temporelles des phénomènes météorologiques.

Echelles spatiale et temporelles des phénomènes météorologiques

La figure 1.1 décrit les principaux phénomènes météorologiques en jeu, ainsi que leurs échelles spatiales et temporelles. De manière plus générale, suivant leur dimension (échelle spatiale) et leur durée de vie (échelle temporelle), les phénomènes météorologiques peuvent être classés en 3 catégories [16] : la micro-échelle, la méso-échelle et la macro-échelle.

Les phénomènes météorologiques à macro-échelle ont la plus grande taille (diamètre > 1000

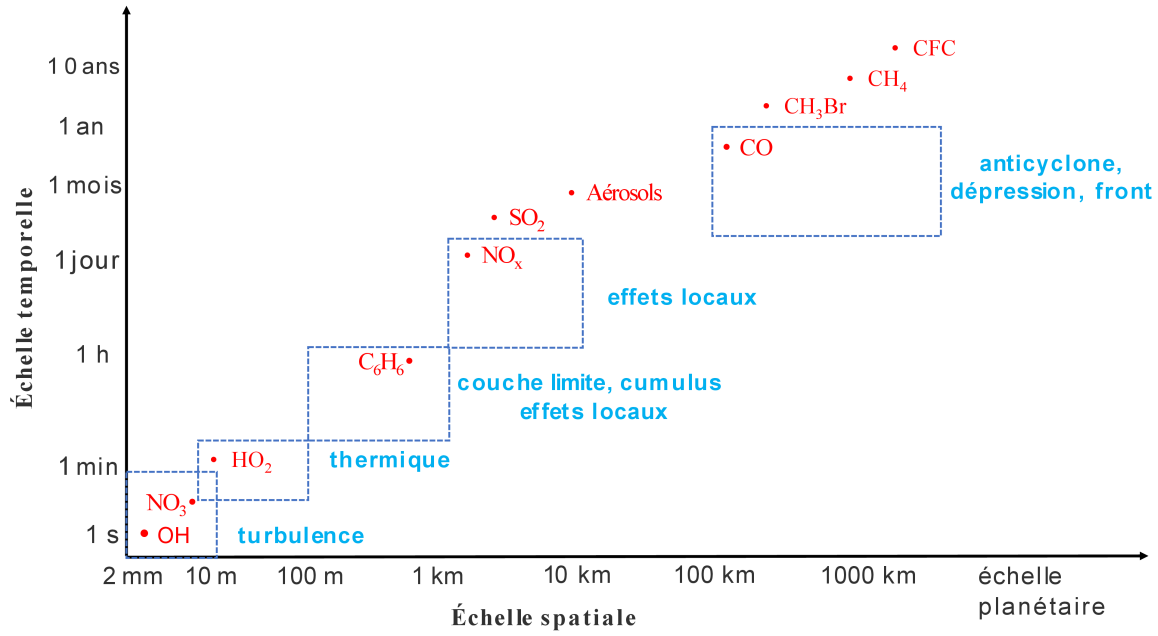

FIGURE 1.1 – Echelles spatio-temporelles de phénomènes météorologiques, associées à la durée de vie de certains polluants atmosphériques. (adapté de [17] et [18])

km) et la plus longue durée de vie (plusieurs jours ou semaines) des trois classes. Généralement, cette catégorie comporte les phénomènes à l'échelle synoptique continentale (environ entre 1 000 et 10 000 km) et planétaire. Les phénomènes météorologiques à macro-échelle englobent des phénomènes météorologiques très vastes dont dérivent les phénomènes à méso-échelle et à micro-échelle comme les cyclones de latitude moyenne, les anticyclones et les dépressions. Les cyclones tropicaux et les fronts sont également classés comme une perturbation régionale à échelle synoptique [16].

D'une durée de plusieurs heures à quelques jours, les phénomènes à méso-échelle ont une taille de 1 à 1000 km de diamètre. Ils résultent de mécanismes de forçage internes et externes. Ce forçage interne, dû à la libération d'énergie thermique latente et du mouvement dynamique résultant de la pression locale et des gradients de température, contribue à la génération de ces phénomènes ainsi qu'à leur circulation. Le forçage externe se traduit par des caractéristiques à méso-échelle créées par les caractéristiques et les processus de circulation à macro et échelle planétaire. Les circulations à méso-échelle peuvent être générées à partir de l'advection de température et de tourbillon à grande échelle, des transferts d'énergie thermodynamique des nuages et des perturbations atmosphériques dues aux inhomogénéités de surface, entre autres ([19] ; [20]). Par exemple, la brise de mer, qui crée un vent de la mer vers la terre, se produit à la suite du réchauffement inégal de la terre et de l'eau générant un gardient de pression de l'air [21]. Pendant la journée, sous l'effet du rayonnement solaire, la terre se réchauffe plus rapidement que l'eau de la mer, ce qui conduit à une différence de température et donc de pression de l'air au-dessus de la terre et de la mer. Cela se traduit par une advection d'air plus frais et plus dense au-dessus de l'eau vers la terre. Puisque les gradients de pression et de température les plus forts existent près de la limite terre-mer, les vents les plus forts se produisent normalement juste à côté de la plage et diminuent à l'intérieur des terres. En outre, du fait que le plus grand gradient de température entre la terre et l'eau de mer a normalement lieu pendant l'après-midi, les brises de mer sont les plus fortes à ce moment. L'inverse de ce phénomène a lieu la nuit, lorsque la brise de terre se forme en réponse au refroidissement de la surface terrestre par rapport à la mer. Ce type de phénomène se produit d'autres situations similaires par exemple les brises de pente à l'interface plaine-montagne.

La dernière catégorie représente les phénomènes à micro-échelle. Ils se produisent sur des échelles de temps très courtes, allant de quelques secondes à quelques minutes, avec un diamètre moyen compris entre moins d'un mètre et quelques kilomètres. Les systèmes météorologiques à l'échelle microscopique (par exemple les thermiques, tourbillons turbulents etc.) se produisent à des échelles de temps très courtes, allant de quelques secondes à quelques minutes, avec un diamètre moyen compris entre moins d'un mètre et quelques kilomètres. A échelle, les processus atmosphériques sont dominés par les conditions de surface terrestre et les échanges d'énergie dans la partie la plus basse (de moins de 1 km) de la troposphère appelée la couche limite atmosphérique. La turbulence est un exemple des phénomènes à micro-échelle. Elle est représentée par une augmentation dans la capacité de mélange des composants de l'air. La turbulence atmosphérique de type thermique est due au réchauffement de la terre par le rayonnement solaire qui mène à une convection de la couche limite (décrite ci-dessous).

Focus sur la couche limite atmosphérique

La couche limite atmosphérique est la partie de la troposphère en contact direct avec le sol [22] et son rôle est particulièrement important car les polluants sont émis en son sein. D'une épaisseur de quelques centaines de mètres à quelques kilomètres suivant l'ensoleillement et le vent, le comportement dynamique de la couche limite est complexe et la turbulence, d'origine thermique et mécanique, y favorise les échanges verticaux, jouant par conséquent un rôle majeur dans la dispersion des polluants.

Les trois principaux composants de la couche limite sont la couche convective mélangée (appelée aussi couche mixte), la couche résiduelle et la couche stable (figure 1.2). Dans la couche mixte, la turbulence est généralement causée par la convection thermique (transport vertical) et par le cisaillement du vent. La convection thermique se produit pendant la journée lorsque la surface chauffée du sol, sous l'effet du rayonnement solaire, crée des thermiques d'air chaud provenant du sol. Ainsi, une couche mélangée turbulente commence à se développer en profondeur environ une demi-heure après le lever du soleil et atteint sa profondeur maximale à la fin de l'après-midi. La turbulence qui en résulte tend à mélanger et diluer efficacement les polluants émis. De plus, la formation d'une couche stable au sommet de la couche mixte

(appelée zone d'entraînement) agit comme un couvercle pour les thermiques ascendants, limitant ainsi physiquement la dispersion verticale des polluants [22]. La couche de mélange subit généralement un cycle diurne s'élevant à une hauteur de 1 à 2 km dans la journée et s'abaissant à 100 à 300 m la nuit [23]. La nuit, lorsque les thermiques cessent de se former et que la turbulence diminue, le sommet de la couche de mélange s'abaisse pour former la base de la couche stable (nocturne). Un mélange vertical limité se produit dans la couche stable, ce qui entraîne une augmentation des concentrations de polluants influençant ainsi les conditions chimiques initiales pour la photochimie.

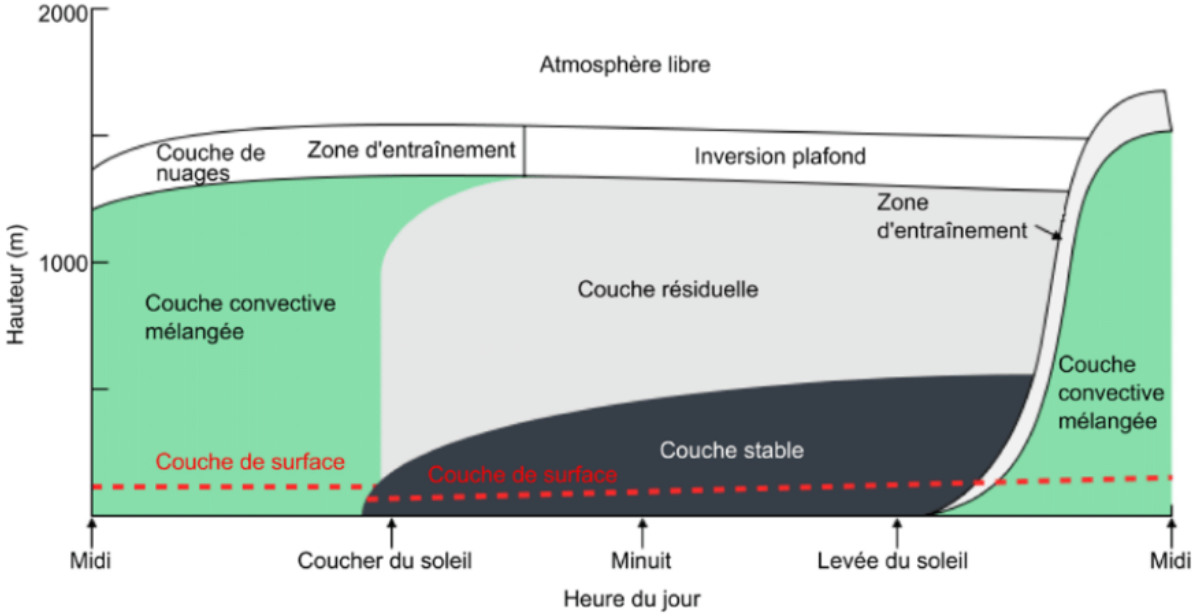

FIGURE 1.2 – Evolution de la couche limite atmosphérique [22].

1.1.3 Impacts de la pollution de l'air

L'OMS estime que l'exposition à la pollution de l'air ambiant cause chaque année environ 3,7 millions de décès prématurés dans le monde [24]. Une majeure partie de la population mondiale (91 %) est exposée à une pollution atmosphérique qui dépasse les directives de l'OMS [25]. Suite au taux d'exposition à la pollution de l'air, les effets vont des impacts subtils légers non apparents jusqu'au décès prématuré (figure 1.3) [24]. L'exposition à long terme aux polluants atmosphériques comme l'ozone (O_3) et les particules (PM) conduit à une augmentation du risque de décès dus à des causes cardiopulmonaires [26], et à d'autres

impacts négatifs comme la santé mentale par des démences [27] ou le cancer des poumons [28], etc. Par voie de conséquence, la détérioration de la qualité de l'air est à l'origine de la réduction de l'espérance de vie et à une augmentation considérable des coûts financiers et qualité de vie aussi [29].

La qualité de l'air a aussi une influence considérable sur l'Environnement. Les concentrations

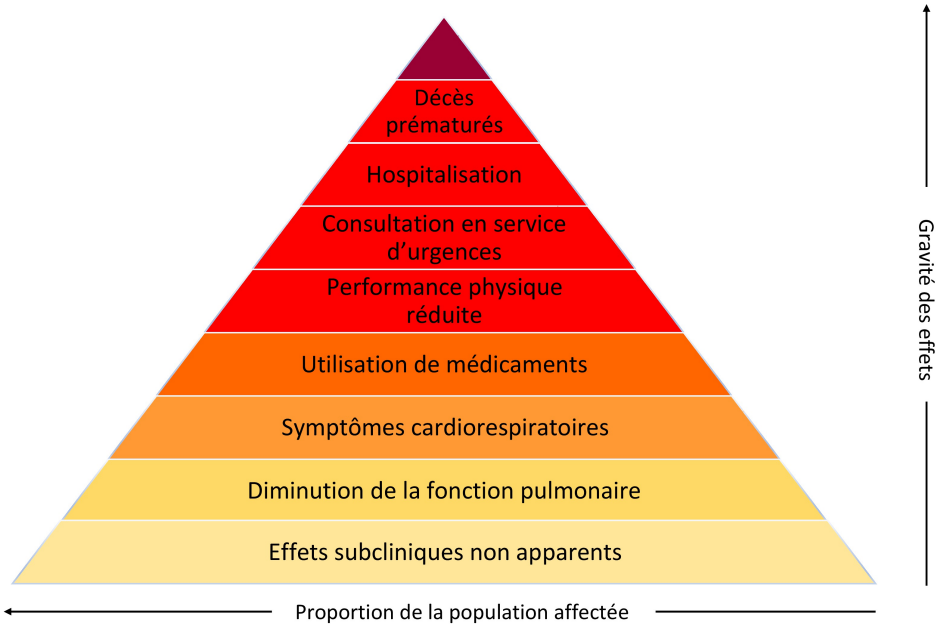

FIGURE 1.3 – Pyramide des effets de la pollution de l'air sur la santé [30].

élevées d'ozone endommagent les plantes et réduisent le rendement des cultures ([31] ; [32] ; [33]). Les aérosols de sulfate et de nitrate contribuent aux précipitations acides, ce qui peut rendre les lacs et les ruisseaux inhabitables pour les poissons. Les précipitations acides peuvent également éroder les bâtiments, les monuments et la peinture des voitures.

1.1.4 Description de la zone d'étude 1 : Région Hauts-de-France

La première partie de ce travail est consacrée à l'étude de la pollution de l'air en région Hauts-de-France. L'application de la loi du 16 janvier 2015 relative à la délimitation des régions de la France, a conduit à la fusion des deux régions Nord-Pas-de-Calais et Picardie en une seule région dite Hauts-de-France le 1er Janvier 2016 (figure 1.4). Cette région compte plus de 6 millions d'habitants pour une densité de 189 habitants/km^2 au 1er janvier 2015.

C'est la troisième région la plus peuplée de l'hexagone, et la deuxième la plus densément peuplée de France métropolitaine après l'Île-de-France, sa voisine au sud. Elle occupe 5.7 % de la superficie de la France métropolitaine avec une superficie de 32000 km², en se situant au cœur de l'Europe du Nord et du triangle Paris-Bruxelles-Londres. La région abrite de nombreuses activités industrielles et agricoles et connaît un important trafic routier et maritime, tant au niveau de ses ports que du transport de passagers. Elle est délimitée au nord par la mer du Nord sur une distance de 45 km, et à l'ouest, par la Manche sur une distance de 120 km, d'où son climat tempéré de type océanique (`https://www.prefectures-regions.gouv.fr/hauts-de-france`). La région de Hauts-de-France contribue à 12.2% aux émissions de

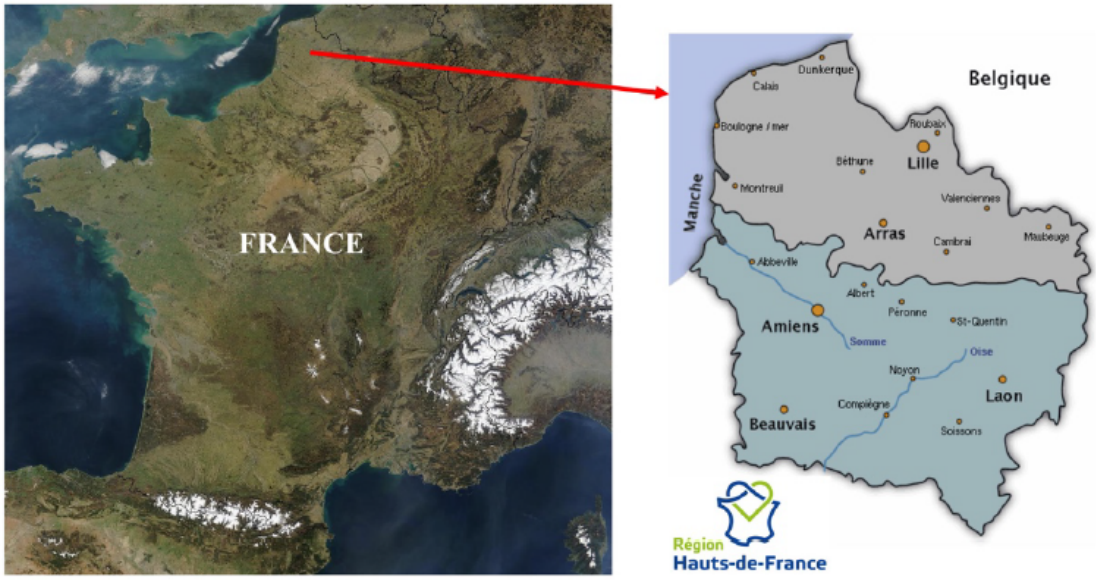

FIGURE 1.4 – Localisation de la région Hauts-de-France.

PM_{10} en France (5.2Kg/hab. en Hauts-de-France contre 4kg/hab. en France). Ce chiffre élevé est justifié par l'importance des secteurs agricole, résidentiel, du transport routier et de l'industrie manufacturière et leur rôle dans l'émission de PM_{10} dans la région (figure 1.5). Le principal secteur d'émission pour l'année de 2015 est le secteur agricole par 35% des émissions après avoir été le deuxième en 2012, suite à une réduction dans les émissions industrielles de 11.2 kt en 2012 (30% des émissions PM_{10}), à 6.2 kt en 2015 (20% des émissions PM_{10}). Le deuxième secteur d'émission avec un taux presque stable est le secteur résidentiel avec 8.6 kt en 2015 et 8.3 kt en 2012. Et en dernier le secteur du transport routier avec une valeur de 5.6

kt. La région de Hauts-de-France a dépassé la valeur indicative en moyenne annuelle et journalière des concentrations de PM_{10} indiqué par l'OMS en 2018.(`https://www.atmo-hdf.fr/joomlatools-files/docman-files/Bilan_annuel/Bilan_QA_2018.pdf`)
Cette zone d'étude représente un bon exemple d'une zone à pollution atmosphérique complexe, du fait de la diversité des sources de pollution de l'air et de la complexité des phénomènes météorologiques qui rendent difficile l'étude et l'interprétation de la qualité de l'air.

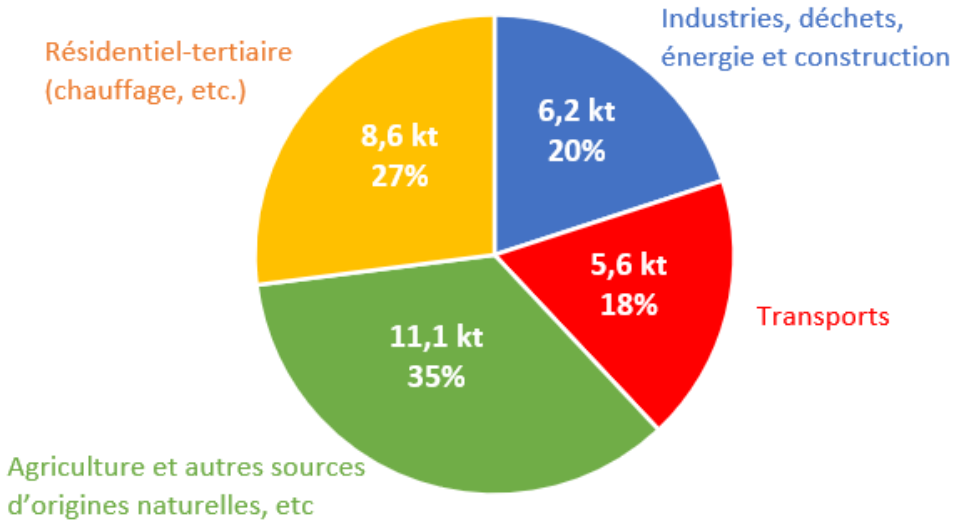

FIGURE 1.5 – Les émissions régionales les PM_{10} par secteur d'activité en 2015 (Adapté de ATMO Hauts-de-France, inventaire des émissions, 2015).

Description de la zone d'étude 2 : Agglomération Dunkerquoise

Dunkerque est la cinquième ville la plus peuplée de la région Hauts-de-France, avec presque 91 412 d'habitants le 1er janvier 2016 (figure 1.6). Grâce à la centrale nucléaire de Gravelines et le terminal méthanier de Loon-Plage, Dunkerque est la première plateforme énergétique de la région. A proximité de cette centrale se trouve le grand port maritime de Dunkerque qui s'étend sur une distance de 17 km. Ce port est classé troisième de France, et par sa proximité géographique avec la mer du Nord, il est la voie maritime la plus empruntée au monde avec un trafic moyen de 600 navires par jour. Dunkerque est une des 17 communes de la communauté urbaine de Dunkerque Grand Littoral, qui compte environ 202000 habitants en 2014. Cette agglomération dunkerquoise comporte un parc industriel dense, comprenant notamment des

installations pétrochimiques et sidérurgiques et une usine de production d'aluminium. Et la zone dunkerquoise abrite trois de ces établissements ; Arcelor Mittal, Aluminium Dunkerque et la Société de la Raffinerie de Dunkerque. Les activités urbaines et industrielles de cette ville font d'elle un site attractif aux études liées à la pollution de l'air et la dynamique atmosphérique. (`https://www.ville-dunkerque.fr`) Atmo Hauts-de-France a estimé dans

FIGURE 1.6 – La localisation de la ville de dunkerque au sein de la région Hauts-de-France.

son bilan de la qualité de l'air de 2018, que 13.8 kg de particules PM_{10} sont produites par habitant au niveau de la Communauté Urbaine de Dunkerque en 2015, contre 3.8 kg par habitant du Nord. Ceci s'explique par la forte activité industrielle qui contribue majoritairement aux émissions PM_{10} allant jusqu'à 87% contre 20% au niveau régional (figure 1.7) et la participation des secteurs agricole, résidentiel et du transport routier par un taux de 5% chacun. Les concentrations de PM_{10} les plus élevées sont présentes au niveau de la zone industrialo-portuaire au nord de Dunkerque, de Gravelines à l'ouest du territoire et le long de l'autoroute A16 au sud. Des dépassements de la valeur limite en moyenne sont observées à la Communauté Urbaine de Dunkerque (`https://www.atmo-hdf.fr/joomlatools-files/docman-files/Bilans-territoriaux/2019/BT07-CU_dunkerque-2018.pdf`).

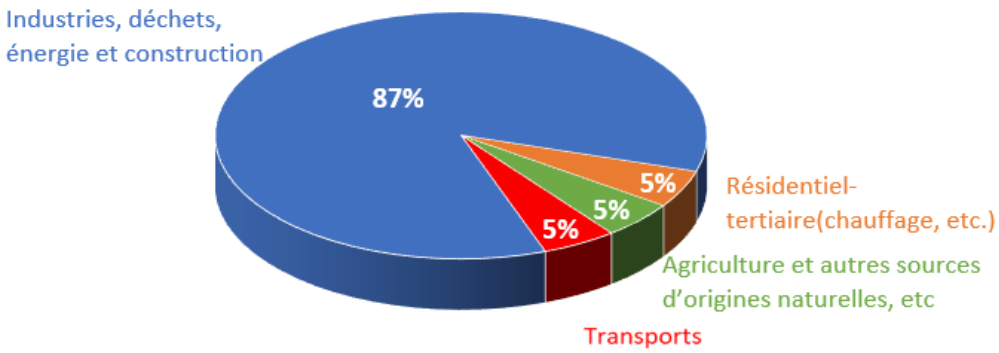

FIGURE 1.7 – Les émissions de la communauté urbaine de Dunkerque les PM_{10} par secteur d'activité en 2015. (Adapté d'ATMO Hauts-de-France, inventaire des émissions, 2015)

1.1.5 Surveillance de la qualité de l'air

En application du Code de l'Environnement en France, la surveillance de la qualité de l'air et l'information du public sont réalisées par des Associations Agréées pour la Surveillance de la Qualité de l'Air (AASQA) [34], regroupées au sein d'une fédération, ATMO France, qui les représente au niveau national et partage leur expertise et les moyens. Les réseaux de surveillance de la qualité de l'air et de mesures de la pollution atmosphérique sont implantés dans toutes les régions du pays (`http://www.developpement-durable.gouv.fr`).

Cette surveillance fait appel à différents moyens, parmi lesquels figure en premier lieu le déploiement de stations fixes équipées d'un ou plusieurs instruments de mesures, qui relèvent la pollution de l'air en continu 24 heures sur 24 (lorsqu'il s'agit de mesures automatiques), ou à une fréquence régulière [34]. L'avantage principal de ces stations est la disponibilité des mesures pour une variété de polluants et la fiabilité de ces mesures pour le suivi de la pollution à long terme. Ces stations de mesure nécessitent du personnel qualifié (techniciens, informaticiens, ingénieurs, etc.) pour assurer le bon fonctionnement des équipements, l'étalonnage, la récupération et le traitement des données. Ce personnel doit également avoir des compétences en maintenance électrique et électronique des appareils. Outre le coût des ressources humaines, le budget de construction et d'équipement d'une seule station est de plusieurs dizaines de milliers d'euros. Cela limite fortement le nombre des installations de surveillance, notamment à proximité des points chauds comme les sites industriels, et par conséquent la résolution spatiale des cartes de pollution.

La surveillance de la qualité de l'air peut également être effectuée par des modèles physiques appelées les modèles de dispersion atmosphérique, qui simulent le transport et la dispersion de la pollution atmosphérique [35]. En prenant comme entrée la position et le taux d'émission des sources de pollution identifiées et les données météorologiques, ces modèles sont en mesure d'estimer la pollution atmosphérique à un endroit donné.

Les limitations de la surveillance par les stations de mesures fixes poussent maintenant les acteurs dans le domaine de qualité de l'air à développer une solution plus flexible grâce aux capteurs à bas coût, moins précis mais plus petits que les stations et surtout beaucoup moins chers. Ces capteurs sont conçus pour être déployés en grand nombre dans une ville donnée et communiquent généralement sans fil en formant, appelé un réseau de capteurs sans fil. Ces appareils fixes ou mobiles, permettent de mesurer la pollution à une échelle urbaine, par exemple en équipant les bus de la ville de Paris [36]. En mobilité, ils offrent une résolution spatiale et temporelle élevée, bien que concentrés sur des itinéraires spécifiques et des périodes spécifiques (par exemple à l'heure de pointe). Ces données sont donc plus denses que les données de surveillance classiques et les complètent avantageusement. Grâce aux capteurs à faible coût, la surveillance de la qualité de l'air n'est plus limitée aux grandes villes [34].

Bien que les nœuds de capteurs soient peu coûteux et puissent être déployés à grande échelle, la densité de capteurs d'un réseau est forcément limitée par des contraintes d'installation et de maintenance du réseau. Par conséquence, la conception du réseau nécessite l'emploi de méthodes d'optimisation du nombre et de l'emplacement des nœuds de capteurs pour garantir une surveillance fiable [37]. Nous abordons cette problématique dans le dernier chapitre pour le cas de l'agglomération de Dunkerque.

Des méthodes variées ont été proposées dans la littérature pour estimer les concentrations de polluants par les techniques de surveillance précédentes et font l'objet d'une discussion dans le chapitre 2 de ce manuscrit [38]. Une des approches couramment utilisée est l'interpolation spatiale, notamment les méthodes géostatistiques [39]. Ces méthodes permettent de déterminer la qualité de l'air aux endroits non surveillés, en utilisant les données d'un réseau de surveillance de la qualité de l'air existant, pour estimer la distribution spatiale des polluants. Le chapitre 3 est consacré à la mise en œuvre de ces méthodes pour l'estimation des

concentrations PM_{10} de la région de Hauts-de-France.

Données de la qualité de l'air en Hauts-de-France : ATMO Hauts-de-France

ATMO Hauts-de-France est une des 18 associations de surveillance la qualité de l'air en France (`https://www.atmo-hdf.fr`). Sa tâche est d'observer l'air de la région d'Hauts-de-France, de notifier au quotidien et d'alerter en cas de pollution ainsi que d'accompagner les projets liés à la qualité de l'air. Les données traitées dans cette thèse, en première et deuxième partie, ont été fournies par ATMO Hauts-de-France.

- **Réseau de stations de mesure de la région Hauts-de-France**

La première partie de ce travail repose sur des données de concentrations PM_{10} mesurées en 2013 par le réseau de surveillance d'ATMO Hauts-de-France, qui surveille la qualité de l'air, grâce à des stations de mesure réparties en région Hauts-de-France et classées selon la nature des émissions en stations urbaines, péri-urbaines, à proximité industrielle ou automobile et rurales [40].

- **Modélisation des concentrations PM_{10} : ADMS**

Les données de la deuxième partie de cette thèse, ont été modélisées par le système ADMS. ADMS (Atmospheric Dispersion Modelling System) est une famille de systèmes conçus par Cambridge Environmental Research Consultants (CERC) du Royaume-Uni pour la modélisation de la qualité de l'air en différentes situations (à titre d'exemple, ADMS airport : algorithmes pour modéliser la dispersion des moteurs d'avion et ses émission, ADMS-Puff : méthodologie pour modéliser le devenir des rejets gazeux denses et passifs, etc.). Les données traitées dans la deuxième partie de cette thèse ont été calculées par ATMO Hauts-de-France en utilisant ADMS-Urban (version 3.1.6.0) sur la base de l'inventaire des émissions réalisé en région Hauts-de-France.

La prévision des concentrations de polluants dans une zone urbaine est un problème de modélisation d'une morphologie complexe, d'où la création d'ADMS-Urban qui est un système de modélisation de la qualité de l'air pour des villages, villes ou même de grandes zones urbaines. ADMS-Urban est le seul modèle pratique qui prend en compte tous les types de sources d'émission (trafic, industriel, commercial, domestique et d'autres sources moins bien définies) avec

leurs différents niveaux de complexité (y compris les canyons de rue), tout en étant capable de fournir une sortie bien détaillée allant de l'échelle d'une rue jusqu'à toute une zone urbaine (`https://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html`).
ADMS-Urban peut être utilisé dans des objectifs variés, comme les études d'exposition à la pollution atmosphérique, les évaluations de la qualité de l'air et ses effets sur la santé, la fourniture de prévisions détaillées de la qualité de l'air à l'échelle d'une rue, etc.

1.2 Conclusions

Dans ce chapitre, nous avons mis l'accent sur l'intérêt sociétal et la complexité de l'étude de la pollution atmosphérique. Les phénomènes météorologiques, responsables du transport et de la dispersion des polluants atmosphériques, jouent un rôle majeur dans la variabilité spatiale et temporelle de la concentration des polluants de l'air, de même que la distribution spatiale et temporelle des sources de pollution. La région Hauts-de-France, étudiée ici, est un objet d'étude adéquat pour aborder, par des méthodes d'optimisation, la conception d'un réseau de mesures dans un environnement complexe, par la variété de son tissu urbain et industriel et sa position géographique littorale, générant des phénomènes météorologiques locaux comme la brise de mer.

Chapitre 2

Revue sur les techniques d'estimation de la qualité de l'air

Dans ce chapitre, nous présentons notre étude comparative sur les techniques d'estimation de la qualité de l'air. Plus de détails techniques sont présentés dans notre premier article sous forme d'une revue publié dans le journal MAPAN (ISSN : 0970-3950)[38]. L'objectif est de fournir un résumé des différentes techniques utilisées pour effectuer une estimation de la pollution atmosphérique, en prenant en compte les données disponibles et les mesures possibles pour valider cette estimation. En outre, des recommandations sont fournies pour construire un modèle de la qualité de l'air.

Ce chapitre fournit tout d'abord un état de l'art sur les revues de la modélisation de la qualité de l'air (MQA, en anglais Air Quality Modeling) en section 2.1, ensuite les contributions apportées sont indiquées en section 2.2. La section 2.3 est dédiée à une revue bibliographique qui détaille les différentes composantes d'un modèle de la qualité de l'air, et la dernière section est consacrée aux conclusions déduites de ce travail.

2.1 Etat de l'art

La modélisation de la qualité de l'air (MQA) peut avoir différentes interprétations selon son domaine d'utilisation. Dans le domaine informatique, une représentation du niveau de la qualité de l'air par des méthodes numériques (pour estimation, prédiction, analyse, etc.) est

vue comme une modélisation de cette qualité, tandis que dans le domaine de physico-chimie de l'atmosphère, la modélisation de la qualité de l'air concerne les modèles de chimie-transport (Chemistry Transport Model, CMT). Ces modèles décrivent le processus physique et chimique de la dispersion de la qualité de l'air, par des équations mathématiques ou des relations empiriques.

Des revues antérieures ont essayé de synthétiser les travaux menés sur la MQA. Chacun traite ce sujet depuis un point de vue différent. Les revues sur la MQA existantes dans la littérature, couvrent la plupart du temps un des éléments de la MQA. Parmi ces dernières, on trouve Ryan et al. [41] qui synthétisent les différentes études n'ayant utilisé que la technique de la régression de l'utilisation des terrains (RUT, en anglais Land Use Regression). Cette méthode prend en compte le lien existant entre les caractéristiques topographiques et environnementales de la zone étudiée, pour estimer la dispersion des polluants dans l'air.

RUT est largement employée pour caractériser l'exposition à la pollution atmosphérique intra-urbaine. Ryan et al. [41] discutent dans leur travail, les similitudes et les différences entre les variables d'entrée employées dans les six études sur lesquelles ils se sont basés dans cet article de revue. Une étude comparative entre RUT et d'autres techniques d'estimation de la qualité de l'air a été menée par Hoek et al. [42]. Pour 25 cas d'études examinés, la méthode de RUT s'est avérée être meilleure que d'autres techniques testées, comme le krigeage (appelé aussi Gaussian Process Regression, GPR) et les modèles de dispersion. Les auteurs ont également proposé une amélioration possible à la RUT, en ajoutant d'autres variables d'entrée qui améliorent davantage la précision de cette dernière. Zhang et al. [43] [44], se sont concentrés dans leur deux revues, uniquement sur la prévision de la qualité de l'air en temps réel (RT-AQF) . Dans ces deux papiers, les auteurs discutent de la possibilité d'avoir accès aux principales techniques existantes de RT-AQF, l'historique d'utilisation de ces dernières et les moyens possibles pour améliorer la précision des modèles RT-AQF dans la partie I [43]. En outre, les auteurs présentent les défis qui limitent la performance des RT-AQF ainsi que leurs perspectives dans une deuxième partie du même travail [44].

Notre revue bibliographique vient compléter les travaux antérieurs, en détaillant l'importance de chaque élément dans la conception et le développement d'un MQA, et en listant les diffé-

rentes techniques et outils disponibles pour la mise en œuvre de ce dernier.

2.2 Contributions

Ce chapitre se résume à la présentation des modèles de la qualité de l'air (MQA), permettant d'étudier et d'analyser le comportement et les différentes sources contributrices dans la pollution atmosphérique. Les MQAs existants ont été employés pour une variété d'objectifs et dans de nombreuses situations, à titre d'exemple, pour mesurer le taux d'exposition à des polluants spécifiques, ou pour analyser la contribution de ces derniers dans l'aggravation d'une maladie respiratoire, etc. Les articles de revue des MQAs précédentes couvrent généralement l'un des éléments constitutifs des MQAs, dans ce chapitre, nous tenterons plutôt d'identifier le rôle et la pertinence de chaque composant pour la construction des MQAs, y compris les techniques existantes pour construire des MQAs, l'influence de la disponibilité de différents types de données dans la performance d'un MQA, et les mesures de validation des MQAs les plus utilisées. Nous présentons des recommandations pour la construction d'un MQA en fonction de son objectif et des jeux de données disponibles, en indiquant ses limites et ses avantages.

Ce travail représente un ensemble de recommandations utiles pour les personnes souhaitant construire un MQA, compte tenu des contraintes qui s'imposent, telles que la disponibilité des données et des ressources techniques / informatiques.

2.3 Revue bibliographique sur la modélisation de la qualité de l'air (MQA)

La synthèse bibliographique que nous proposons vise à fournir une analyse approfondie basée sur une multitude de travaux portant sur la prévision de la qualité de l'air, en soulignant les divers éléments de la modélisation de la qualité de l'air. Cette synthèse s'articule sur la présentation des performances et des points forts des méthodes existantes de prévision/estimation et d'évaluation de la qualité de l'air, ainsi que sur l'examen du rôle des données d'entrée de

ces modèles, y compris leur accessibilité (gratuites, payantes ou limitées), leur types (directement mesurés ou substitués) et leur sources (par exemple, données sur le trafic et l'utilisation des terrains/Land use data), étant donné que la disponibilité des données est généralement l'aspect le plus critique lors de la sélection d'un modèle de qualité de l'air. Notre synthèse bibliographique introduit ensuite un résumé des mesures d'évaluation et de validation des modèles de qualité de l'air, et enfin des recommandations sur le développement des MQAs en fonction des données disponibles et des objectifs visés.

2.3.1 Techniques d'estimation/prévision de la qualité de l'air des MQAs

Avoir une idée claire sur les techniques de modélisation et d'analyse des polluants atmosphériques peut conduire à une meilleure interprétation des résultats de la MQA, une meilleure sélection du modèle de la qualité de l'air qui soit approprié, et l'identification des contributions primaires qui intensifient la proportion de la pollution atmosphérique.

Bien que diverses techniques aient été mises au point pour prédire ou prévoir les polluants atmosphériques, il est essentiel de choisir une méthode appropriée en fonction de l'objectif de l'application (par exemple, prédire/prévenir la pollution atmosphérique, révéler les facteurs contribuant à la pollution atmosphérique, etc.) et les données d'entrée (la densité du réseau de stations de surveillance et les informations sur l'utilisation des sols). Par exemple, si nous disposons d'un réseau dense ou si nous voulons ajouter un site de surveillance au réseau, la modélisation de la dispersion (comme CALINE : California Line Source Dispersion Model) est la bonne solution, car elle s'adapte facilement à de nouveaux polluants et/ou à de nouvelles zones géographiques sans l'ajout de sites de surveillance supplémentaires dans la zone d'étude.

Dans cette section on classe les techniques de MQA en quatre catégories :

- La régression de l'utilisation des terrains (RUT).
- L'apprentissage automatique.
- Les techniques hybrides.
- Autres techniques moins fréquemment utilisées.

Régression de l'utilisation des terrains (RUT)

La régression de l'utilisation des terrains (RUT, en anglais Land Use Regression) est une technique qui développe des modèles stochastiques pour prédire les concentrations de polluants atmosphériques sur un site donné en utilisant les variables prédictives (par exemple, l'utilisation des terrains environnants, le réseau routier, le trafic, l'environnement physique et la population) basées sur les systèmes d'information géographique (SIG) et le suivi des mesures de polluants atmosphériques. Le projet SAVIAH (Small Area Variations in Air quality and Health) est la première étude à modéliser les variations à petite échelle des polluants atmosphériques par RUT [45] et démontre que la cartographie de régression basée sur le SIG est un outil robuste pour prédire la qualité de l'air à une résolution spatiale fine avec des données de surveillance limitées.

Combinée à des stratégies efficaces, la RUT peut être utilisée pour expliquer les conditions de qualité de l'air (par exemple, les variations saisonnières des activités humaines où la densité de population n'est une source importante de pollution de l'air qu'en hiver, et inversement en été, les indicateurs industriels ont une plus grande influence) et pour révéler les principales causes de cette dernière [46]. Outre la prévision des concentrations de polluants atmosphériques, la RUT peut également déduire l'influence sur le milieu environnant. Chen et al. (2016) révèlent que l'exposition au trafic routier intense peut affecter la cognition humaine, en utilisant l'image satellite des mesures de $PM_{2.5}$ avec d'autres variantes (par exemple, la longueur de la route, l'âge et le sexe) comme entrées du modèle RUT [47]. Les auteurs ont constaté que le fait de vivre à proximité d'une route principale pourrait entraîner une augmentation de l'incidence de la démence.

Le "bon choix" des variables joue un rôle important dans l'élaboration des modèles RUT. Ross et al. (2007) affirment [48] qu'avec les variables du trafic et d'utilisation des terrains (Land use data), le modèle RUT arrive à estimer plus de 60% des concentrations observées de $PM_{2.5}$ sur une large zone pour trois différentes périodes de l'année et des zones. En outre, l'ajout de variables plus pertinentes dans le modèle RUT aurait un effet positif sur les performances. De plus, Ross et al. (2006) font améliorer la précision d'estimation de la pollution atmosphérique de 54 à 79% [49] en ajoutant en entrée une variable de trafic (trafic dans un rayon de 300m

autour des stations de surveillance).

ADDRESS (A Distance Decay Regression Selection Strategy) est une stratégie de sélection de distances tampons (buffer) optimisées pour les prédicteurs potentiels afin de maximiser les performances du modèle. ADDRESS modélise la pollution atmosphérique liée au trafic à Los Angeles, un paysage urbain extraordinairement complexe. En calculant les coefficients de corrélation des co-variables spatiales (données commerciales, résidentielles, industrielles et d'utilisation des terrains à l'air libre) avec les résidus des concentrations d'exposition, le modèle crée une série de décroissance de la distance. Cette stratégie améliore la RUT traditionnelle avec une prédiction normalement distribuée et une précision variant entre 87 et 91% [50].

La performance de la RUT peut également être améliorée en la couplant avec certaines approches développées pour la pollution de l'air. Lee et al. (2014) utilisent l'approche de modélisation ESCAPE (European Study of Cohorts for Air Pollution Effects) avec la RUT et montrent son efficacité dans l'explication des variations spatiales, même lorsqu'il s'agit d'une forte densité de routes et d'une zone peuplée comme la ville de Taipei.

ESCAPE est une étude des expositions à long terme à la pollution atmosphérique ayant des effets sur la santé humaine pour 15 pays européens. Les auteurs appliquent cette approche en développant des modèles RUT de Taipei qui sélectionnent les principales entrées pertinentes à prendre en compte dans le modèle final, en effectuant de multiples analyses sur les prédicteurs prévus pour obtenir un modèle avec une meilleure précision. En outre, les expositions estimées par LURESCAPE (qui combine RUT et ESCAPE) ont une portée plus large de la variabilité des polluants et fournissent des résultats avec une meilleure résolution spatiale par rapport aux deux algorithmes classiques d'interpolation spatiale, c'est-à-dire le krigeage ordinaire et les méthodes du voisin le plus proche (par exemple, le site de mesure). Les résultats pourraient être utiles pour évaluer l'influence de l'exposition à long terme au dioxyde d'azote (NO_2) et aux oxydes d'azote (NO_x) sur les cohortes épidémiologiques de la métropole de Taipei.

La méthode RUT repose fortement sur la disponibilité de données spatiales (utilisation des terrains/Land use data) sans tenir compte des effets spatiaux, tels que la non-stationnarité spatiale et l'autocorrélation spatiale, qui limitent les performances de la RUT en réduisant la précision des prévisions et en augmentant l'incertitude. Bertazzon et al. (2015) développent

un modèle "vent – RUT" [46], qui est une variante du modèle RUT incluant le vent comme variable météorologique pertinente, qui pourrait atténuer les problèmes de non-stationnarité spatiale et d'autocorrélation spatiale. Dans ce travail, les auteurs ont présenté un modèle alternatif de deux modèles, c'est-à-dire un modèle autorégressif spatial (ARS) qui résout la non-stationnarité spatiale et un modèle de régression pondéré géographiquement (RPG) qui traite de l'autocorrélation spatiale. Ils remplacent le ARS et le RPG par un seul modèle de RUT qui est mathématiquement plus simple et qui surpasse la RUT traditionnelle avec une amélioration de 10 à 20% de la moyenne R^2 (coefficient de détermination, définie plus loin dans ce chapitre).

Jusqu'à ce jour, le modèle RUT montre encore sa puissante capacité de prédiction de la qualité de l'air. En prenant l'exemple de la région métropolitaine de Houston, aux États-Unis, Zhai et al. (2016) montrent comment le défi de la précision de l'échelle spatiale devrait être étudié en profondeur. L'impact d'un prédicteur dans une distance/un rayon spécifique au sein d'une zone d'étude n'a pas le même effet dans une zone d'étude différente (non-stationnarité spatiale) [51]. Les auteurs affirment que la nécessité d'une compréhension claire des mécanismes de dispersion physico-chimique n'est pas toujours une obligation dans tout développement de MQA. Ce modèle RUT atteint un taux d'erreur moyen (TEM) inférieur à 20% avec le meilleur $R^2 = 0{,}78$, le plus petit TEM = 11,84%, et la plus petite erreur quadratique moyenne RMSE = 1,43 et surpasse le krigeage ordinaire en utilisant des variables aux échelles spatiales optimisées.

Apprentissage automatique

L'apprentissage automatique est une technique puissante pour prédire/estimer la qualité de l'air, via des modèles basés sur l'informatique et les techniques statistiques. Par exemple, Basu et al. (2018) développent un algorithme qui identifie les interactions dans un système en utilisant des forêts aléatoires itératives qui pourraient être appliquées à de nombreux domaines scientifiques [52]. Plusieurs études ont utilisé divers algorithmes d'apprentissage automatique (par exemple, réseaux de neurones, forêt d'arbres décisionnels (en anglais Random Forest, RF) et régression) pour modéliser la qualité de l'air en raison de leurs performances prometteuses

depuis plus d'une décennie. Par exemple, l'approche des réseaux de neurones artificiels (RNA) a été utilisée dès 2003 pour la prévision des particules ($PM_{2.5}$) [53].

En comparant les avantages et les limites de trois algorithmes de réseaux de neurones pour la prévision des polluants atmosphériques : le réseau de neurones perceptron multicouche (MLP), le perceptron multicouche carré (en anglais square multilayer perceptron) et le réseau de fonctions de base radiales (FBR, radial basis function), Ordieres et al. (2005) concluent que le RBF est le meilleur prédicteur parmi ces trois algorithmes, surpassant les deux autres par des temps d'apprentissage plus courts et une meilleure stabilité (l'indépendance de la variabilité de l'estimation sur les données d'entraînement utilisées). En général, le RNA reste une bonne option pour la prévision de la qualité de l'air qui donne de meilleurs résultats que les autres modèles classiques, comme la persistance (qui est un modèle simple supposant que le niveau de concentration de polluant à un moment donné correspond à la valeur qui s'est produite à la même heure la veille de $y_t = x_t$) et les modèles de régression linéaire, comme l'indiquent Ordieres et al. (2005) [53].

En outre, Xu et al. (2014) [54] proposent un cadre d'exploration bidimensionnel basé sur la régression vectorielle de soutien (RVS) pour prédire les $PM_{2.5}$ à Pékin en 2014. Ce modèle prend en compte les séries chronologiques de $PM_{2.5}$ des stations de surveillance environnantes, pour montrer comment les concentrations de $PM_{2.5}$ se dispersent dans l'espace et dans le temps. Cette étude déploie le SVM (support vectoriel machine) de Weka (Waikato Environment for Knowledge Analysis ; un outil gratuit et facile d'emploi pour l'apprentissage automatique et l'exploration de données) et découvre qu'avec l'augmentation de la portée géographique et du décalage temporel les erreurs de prévision diminuent, mais l'amélioration des performances diminue également. Malgré cette contrainte, le modèle a atteint un bon équilibre entre la performance et le coût de modélisation.

Jiang et al. (2015) [55] rapportent que les arbres de régression sont également utiles pour prédire la qualité de l'air. Ils détectent la pollution de l'air extérieur en se basant sur les messages partagés sur les médias sociaux, Weibo (Twitter chinois) et utilisent un classificateur pour distinguer les niveaux de pollution de l'air à Pékin selon les messages des internautes de la ville, partagés sur Weibo. Les auteurs prévoient l'indice de la qualité de l'air (IQA) à Pékin

en utilisant le gradient tree boost (GTB), qui construit itérativement un arbre de régression à partir des résidus et de la somme pondérée des arbres de régression. Grâce à la fonction multi-additive du GTB, ils développent un modèle de surveillance et de suivi efficace pour la prévision des polluants atmosphériques, en utilisant la classification des données des médias (pour classer si un message posté par un internaute est négatif ou positif, et obtenir à la fin une classification de tous les messages des internautes de la catégorie "excellente" à la catégorie "pollution grave"). De cette façon, les données des médias sociaux pourraient être comparées aux mesures réelles de l'IQA apèrs avoir effectué un filtrage à plusieurs niveaux pour ne prendre en compte que les données des médias sociaux sur la pollution de l'air extérieur dans la région de Pékin (rejeter les données des médias sociaux qui ne concernent pas la pollution de l'air extérieur, les messages publicitaires, et celles qui sont rédigées par des internautes en dehors de la région étudiée).

Certaines études visent à améliorer les performances d'un modèle existant en ajoutant d'autres variables pertinentes ou en proposant une version améliorée du modèle pour le rendre plus précis. En utilisant les résultats du modèle WRF-Chem (Weather Research and Forecasting, WRF) couplé avec la chimie [56] comme entrées en plus des mesures de polluants atmosphériques, Xi et al. (2015) [57] conçoivent un framework d'évaluation complet pour améliorer la performance des prédictions. Ils testent cinq différents algorithmes d'apprentissage automatique sur quatre groupes différents d'ensembles de données, où chaque groupe comprend différentes caractéristiques des données d'entrée (par exemple, l'observation de la pollution, les prévisions météorologiques, la vitesse du vent, etc.), et la technique de RF s'avère être le meilleur prédicteur pour la plupart des groupes de données. Cette stratégie de combinaison conduit à une amélioration de 3 % du modèle de départ, qui n'est pas incorporé aux données du WRF-Chem. En outre, les auteurs concluent que la disponibilité de plus d'informations augmente la possibilité d'améliorer la précision du modèle.

RF est une méthode d'apprentissage automatique pour la classification (et la régression) réalisée par la génération de classificateurs sous forme d'arbres aléatoires et leur assemblage par un agrégateur. Grâce à l'agrégation Bootstrap (expliquée en section 3.2.4 de ce manuscrit), RF construit un ensemble d'arbres de décision qui contribuent à prédire l'indice de la qualité

de l'air (IQA) pour la ville de Shenyang, et l'agrégation de tous les résultats de ces arbres fournit la classification de l'IQA [58]. Dans cet article, RF montre de bons résultats par rapport à trois autres algorithmes de détection de la qualité de l'air urbain. En conséquence, ce modèle produit une précision globale de 81% pour la prévision de l'IQA, et comme toutes les données utilisées proviennent d'Internet, il est possible d'appliquer cette méthode à d'autres villes également.

Brokamp et al. (2017) [59] rapportent que lorsque l'on combine la technique de forêt aléatoire avec des variables qui ont un impact significatif sur la pollution de l'air comme les variables d'utilisation des terrains (land use data), il devient possible de pallier les limites de la RUT en capturant les relations non linéaires et les interactions complexes entre les prédicteurs et le résultat, en utilisant un ensemble de données d'apprentissage de petite taille. Ce MQA donne de meilleurs résultats que la RUT avec une diminution de l'erreur prédictive fractionnaire d'au moins 5% pour la plupart des éléments polluants étudiés (tels que l'aluminium, le cuivre et le fer) et une erreur prédictive fractionnaire évaluée par la validation croisée de moins de 30%, à l'aide des diverses données d'entrée [59].

Hybride

Dans cette revue, nous avons choisi de qualifier de "technique hybride" tout travail qui adopte plus d'une catégorie d'algorithmes (par exemple, l'apprentissage automatique, la géostatistique et la régression de l'utilisation des terrains) pour développer un modèle de qualité de l'air. C'est le cas de la plupart des MQAs spatiotemporels qui étudient non seulement l'aspect spatial de la pollution de l'air mais aussi l'aspect temporel, chacun via une méthode différente (comme expliqué ci-dessous).

Les modèles hybrides se basent généralement sur deux algorithmes ou plus, qui donnent un modèle de la qualité de l'air plus solide, offrant de meilleurs résultats qu'une seule méthode. Par exemple, Wilton et al. (2010) [60] utilisent le modèle de dispersion CALINE3 [61] pour la prévision de la pollution des routes à partir du modèle de dispersion météorologique ; ils ont ensuite réalisé un modèle RUT en utilisant des mesures spatiales concernant le site d'étude pour améliorer les estimations des concentrations spatiales. Ce modèle hybride permet d'améliorer

les valeurs de R^2 pour les villes Seattle et LA. En outre, les auteurs n'ont réussi à améliorer la prédiction des polluants que dans les modèles RUT traditionnels, puisqu'ils incluent la longueur des routes et la densité du trafic comme prédicteurs.

La plupart des techniques hybrides modélisent les aspects spatiaux et temporels séparément, ce qui permet un traitement spécifique pour chacun, puis combinent les résultats des deux. Zheng et al. (2013) [62] tentent d'estimer la qualité de l'air en temps réel en définissant deux classificateurs séparés : les classificateurs spatial et temporel. Le classificateur spatial utilise un RNA (réseau de neurones artificiels) pour modéliser la corrélation spatiale entre la qualité de l'air de différents endroits, en prenant comme données d'entrée les caractéristiques liées à l'espace (par exemple, la densité des points d'intérêt et la longueur des autoroutes). Le classificateur temporel utilise un champ aléatoire conditionnel à chaîne linéaire (en anglais a linear-chain conditional random field) pour représenter la dépendance temporelle de la qualité de l'air sur un site en prenant les caractéristiques temporelles de ses variables (par exemple, le trafic et la météorologie) comme prédicteurs. Ce modèle surpasse quatre autres méthodes standard bien connues sur cinq ensembles de données de la Chine [62].

Un autre exemple d'utilisation des réseaux de neurones est présenté par Zheng et al. (2015) [63] qui prévoient les concentrations de $PM_{2.5}$ dans les 48 heures à venir sur un site de surveillance en combinant le modèle spatial (basé sur RNA) et les classificateurs temporels (basés sur la régression linéaire) avec un agrégateur dynamique qui intègre les résultats spatiaux et temporels d'une manière qui utilise les données météorologiques comme référence de conformité. Pour chaque station, les informations météorologiques seront prises en compte telles que la vitesse et la direction du vent, l'état du temps (brume/soleil/etc.), etc. afin de déterminer un poids pour chaque classificateur. De plus, les auteurs créent un prédicteur qui détecte les changements brusques dans l'air. Ils évaluent la performance du modèle en utilisant 43 villes en Chine, et leurs résultats ont surpassé tous les autres modèles de base représentés par la moyenne mobile autorégressive (en anglais autoregressive moving average), la régression linéaire et l'arbre de régression. La précision était de 75% dans les six premières heures, et elle reste bonne même en cas de changements brusques de la qualité de l'air.

Pour étudier indépendamment chacune des variabilités spatiales et temporelles et exploiter

les variables spatiotemporelles, Li et al. (2017) [63] traitent plus d'une caractéristique de la pollution atmosphérique (par exemple, caractéristique spatiale, caractéristique temporelle, caractéristique locale, etc.) afin de développer un modèle spatiotemporel pour prédire les oxydes d'azote à une haute résolution spatiotemporelle, en incorporant les caractéristques non linéaires et spatiales. Leur approche intègre les relations non linéaires, les comportements constants et aléatoires des prédicteurs en exprimant la variabilité spatiotemporelle des concentrations avec des modèles à effets mixtes. Ensuite, ils effectuent un apprentissage de tous ces modèles et réalisent une optimisation sous conditions de tenir compte de la contrainte des endroits où les données sont incomplètes dans le temps, en minimisant la différence entre les concentrations afin d'ajuster la sortie de prédiction correspondante. En outre, Li et al. (2017) utilisent le modèle de dispersion CALINE4 pour estimer la moyenne (moyenne temporelle) des polluants atmosphériques sur les routes. Cette approche réduit la variance et améliore la fiabilité des prévisions, en améliorant les résultats des caractéristiques mixtes initiales (qui n'incorporent pas l'apprentissage de l'ensemble des données et l'optimisation sous contrainte proposée) avec des valeurs R^2 égales à 85 et 86% pour le dioxyde d'azote (NO_2) et les oxydes d'azote (NO_x), respectivement.

Autres

Toutes les autres techniques de la MQA, qui sont notamment importantes mais moins fréquemment utilisées que les catégories discutées auparavant sont discutées ci-dessous.
Le premier exemple est celui des techniques géostatistiques, qui intègrent des statistiques pour analyser la variance spatiotemporelle des polluants atmosphériques. Fontes et al. (2010) [64] ont réalisé une interpolation de la qualité de l'air pour la zone urbaine de la région de Porto/Asprela et ont montré que la pondération de distance inverse (IDW) est meilleure pour interpoler que le krigeage pour cette région. Ramos et al. (2016) ont développé [65] une technique combinant l'IDW et le krigeage utilisant un ensemble de variables pertinentes à la qualité de l'air. Les auteurs montrent que le modèle hybride est plus performant que l'utilisation de chaque méthode séparément. Les méthodes géostatistiques s'avèrent être meilleures que d'autres techniques pour quelques cas d'étude, comme l'indiquent Rivera- Gonzàlez et al.

(2015) [66], en déduisant que le krigeage ordinaire est la méthode la plus performante parmi les six méthodes testées. Le krigeage ordinaire se montre puissant non seulement en raison de sa bonne précision d'interpolation pour prédire les concentrations de polluants atmosphériques, mais aussi parce qu'il calcule l'erreur standard correspondante (variance d'estimation) de la prédiction.

Une autre technique géostatistique utile est la méthode de régression basée sur la télédétection par satellite telle qu'appliquée par Guo et al. (2014) [67], où ils prédisent les $PM_{2.5}$ au niveau du sol en fonction uniquement de PARASOL niveau 2 AOD (nom du satellite responsable des observations) en utilisant quatre modèles empiriques différents : le modèle de régression linéaire, le modèle de régression quadratique, le modèle de régression de puissance et le modèle de régression logarithmique. Tous les modèles produisent des résultats raisonnablement bons, mais ils sous-estiment les valeurs des concentrations de $PM_{2.5}$ par rapport à celles mesurées au niveau du sol.

Les modèles stochastiques représentent également un moyen fiable de l'analyse de la qualité de l'air. Sun et al (2013) [68] utilisent les Modèles de Markov Cachés (MMC, en anglais Hidden Markov Models) qui sont rarement employés, et essaient de représenter la couche cachée du modèle par une distribution non gaussienne. Dans le but d'améliorer les MMC, les auteurs mettent en œuvre trois différentes fonctions de distribution des émissions : log-normale, gamma et valeur extrême généralisée (VEG) pour prédire les concentrations de $PM_{2.5}$. Par conséquent, le taux de prédiction réel pourrait être amélioré de 150% grâce à ces trois distributions non gaussiennes et au bon contrôle de la génération/arrêt de fausses alertes (qui sont responsables de prévenir lorsque la qualité de l'air dépasse un seuil d'indice de pollution défini) qui pourraient être réduites de 78%.

Un autre modèle statistique qui donne de bons résultats dans l'ajustement de l'erreur de la MQA [69] est l'approche prédictive du filtre de Kalman (KF). Le KF sert d'outil d'ajustement de l'erreur, en augmentant la valeur de R^2 de 43%, pour les prévisions du modèle brut, à 90% pour les prévisions corrigées par le biais du KF sur plus de 90% des sites étudiés. Ce modèle permet d'augmenter les coefficients de corrélation avec les mesures, en outre la méthodologie est facilement adaptable pour des applications en temps réel.

Certaines études de la modélisation de la qualité de l'air calibrent les MQAs juste avant d'effectuer la prédiction pour améliorer leur performance, et cela se fait en appliquant certaines conditions lors du développement du modèle. L'efficacité de cette calibration est ensuite validée par l'un des paramètres mentionnés plus loin en section 2.3.3. Par exemple, Li et al. (2017) [63] imposent des contraintes pour les variables explicatives afin de réduire le surapprentissage du modèle et d'obtenir de meilleurs résultats, tandis que dans d'autres études ([50], [70]) ne sélectionnent que les prédicteurs ayant une valeur de p (mesure de correlation) supérieure à 0,1. Brokamp et al. (2017) [59] effectuent la même sélection pour leurs données d'entrée, en plus d'un paramètre de facteurs de variance et d'inflation (FVI) qui devrait être inférieur à trois pour conserver une variable comme prédicteur dans le modèle afin d'améliorer le R^2 du modèle.

Plusieurs MQA de RUT examinés dans ce chapitre adoptent une approche progressive pour déterminer les données d'entrée les plus pertinentes à considérer dans le modèle (par exemple [48], [70], [71], [72], etc.) et pour se débarrasser de tous les prédicteurs inutiles.

2.3.2 Analyse des données d'entrée des MQAs

La collecte de toutes les données disponibles est la première étape dans la sélection d'une technique appropriée pour la construction d'un MQA, après avoir défini l'objectif de de ce dernier (par exemple l'analyse de l'effet de la pollution atmosphérique sur la santé de l'humain [47], ou la détermination de la qualité de l'air dans les espaces verts [73], etc.) Dans cette section, nous décrivons les données couramment utilisées dans les articles examinés dans les sections ci-dessus en indiquant, si possible, les références aux jeux de données gratuits et payants qui pourraient être utilisés pour développer un MQA.

Accessibilité

L'accessibilté est l'un des facteurs les plus critiques pour l'élaboration d'un modèle de qualité de l'air d'une bonne précision ([54], [57] et [74]) car l'intégration de plus d'informations (variables) peut généralement améliorer la précision des MQAs. Dans cette section, nous classons les données des articles examinés selon leur accessibilité en trois catégories : gratuites

(accessibles au public), payantes et non autorisées (confidentielles/difficiles à obtenir).

Données gratuites, payantes et non autorisées

De nombreuses études utilisent des ensembles de données gratuits disponibles en ligne. Dans le cas de la région métropolitaine de Houston, au Texas, aux États-Unis, Zhai et al. (2016) [51] utilisent des données publiées en ligne (accessibles au public comme : concentrations de polluants, utilisation/couverture des sols, réseau routier et données de recensement) et fournissent dans leur article des liens permettant d'accéder à ces données. Yu et al. (2016) [58] prévoient les IQA pour la ville de Shenyang, en Chine, en utilisant des données disponibles en ligne du trafic et des routes de Baidu et de Google maps. Avec les données disponibles sur la qualité de l'air de Ciudad Juarez et El Paso-Mexique, Ordieres et al. (2005) déterminent les concentrations de $PM_{2.5}$ pour les 16 heures restantes de la journée [53]. Lin et al. (2017) extraient des caractéristiques géographiques d'OpenStreetMap (cartographie en ligne du monde) pour construire leur modèle d'estimation [75] et un autre de prévision de la qualité de l'air [76]. L'exploitation de données open source permet de généraliser les modèles à d'autres domaines d'étude.

De plus, les données sont parfois disponibles même pour un grand nombre de villes comme dans l'étude de Xu et al. (2014) [54] ; les auteurs prévoient la pollution de l'air pour 190 villes chinoises en se basant sur des données disponibles en ligne. Le cas de Vienne, en Autriche [77], représente un cas similaire, où les informations géographiques volontaires (IGV) servent à générer des modèles d'utilisation des terrains/land use data, sans aucune technique de télédétection ni données observées supplémentaires.

Cependant, certaines données sont déclarées être gratuites, mais nous n'avons pas pu y accéder en raison de dysfonctionnement au niveau des liens fournis ([66], [71]) (vérifié le 14/03/2018). Les ensembles de données de modélisation de la qualité de l'air ne sont pas toujours accessibles en raison de confidentialité (droits sur les données), seuls les membres d'une organisation (par exemple, un laboratoire de recherche ou une université) peuvent avoir accès à ce type de données. Ramos et al. 2016 [65] travaillent sur la qualité de l'air de la ville canadienne de Calgary, l'ensemble de données de cette ville contient des informations qui sont soit : publiques (données sur le volume du trafic, recensement de la population, et informations sur les em-

placements géographiques des industries) ; destinées aux membres de l'Université de Waterloo uniquement (trafic routier et utilisation des sols) ; et des informations qu'ils ont obtenues de la Surveillance nationale de la pollution atmosphérique du Canada (données sur la qualité de l'air, vitesse et direction du vent). Dans le cas de la prévision de la qualité de l'air en temps réel à Pékin et Shanghai, Zheng et al. (2013) utilisent les trajectoires GPS d'un grand nombre de taxis qu'ils rassemblent eux-mêmes [62]. Dans certaines régions, même les concentrations de polluants ne sont pas disponibles. Fontes et Barros ont mené une campagne pour mesurer eux-mêmes les concentrations de polluants dans la région urbanisée d'Asprela à Porto, et ces mesures sont toujours confidentielles [64].

Le dernier type de données est celui des données payantes. Par exemple, TeleAtlas est une société qui fournit des informations cartographiques numériques comme les données d'un réseau routier ([60], [78]). Ces données payantes contribuent à améliorer les performances des MQAs en les intégrant aux autres données existantes. Yang et al. (2017) combinent des images de la télédétection par satellite provenant de sources payantes avec des mesures observées, et leur modèle s'avère être précis pour le cas du polluant NO_2 [71]. Nous résumons les liens utiles vers les données gratuites et payantes dans le tableau 2.1.

Disponibilité et qualité des données

À ce jour, le manque de données entrave le développement de modèles de qualité de l'air car il influence le choix de variables significatives pour arriver à une estimation précise de la pollution de l'air. Par exemple, les données géographiques de la zone étudiée sont des variables nécessaires dans le modèle car elles décrivent la topologie de la région (par exemple, les espaves verts auraient une valeur de concentration de polluants totalement différente de celle de la route ou du bâtiment recouvert de goudron) ([50], [70], [71]). En outre, les informations relatives au trafic [79], les données météorologiques [49], le nombre des stations de mesure du réseau de surveillance (connu pour être le principal obstacle en raison de sa faible densité dans presque tous les travaux AQM) ([48], [70], [51], [62], [64], [65], [71]), et d'autres informations ([80], [66], [74]) sont nécessaires pour modéliser la qualité de l'air. Le manque de données peut conduire à des résultats inexacts, à titre d'exemple, lorsque le nombre de stations de

Type	Références	Données	Liens et sources
Payant	[50]	Trafic, Réseau routier	TeleAtlas :tomtommaps.com
	[67], [69]	PM_2, au niveau du sol	MODIS : https://modis.gsfc.nasa.gov
	[78]	Réseau routier	TeleAtlas
		Trafic	ESRI : https://www.esri.com/fr-fr/home
	[45]	Toutes les données	ARC/INFO
	[63]	Densité de trafic	ESRI, ArcGIS : https://www.esri.com/fr-fr/arcgis/products/arcgis-pro
		Distance de la route	ESRI
		Densité de population	ArcGIS
Gratuit	[48]	Données sur l'utilisation des terrains	USGS : https://www.usgs.gov
		Mesures des polluants	EPA : https://www.epa.gov/outdoor-air-quality-data
		Densité de population	U.S. Census Bureau :http://www.census.gov/main/www/access.html
		Réseau routier	ESRI : http://www.openstreetmap.org/#map=5/51.500/-0.100
	[58]	Données météorologiques	https://rp5.ru/Météo_Monde :Weather for 243 Countries of the World
		POI (point d'intérêt)	Google maps : https://www.google.com/maps
		Données sur le trafic et les routes	Baidu maps : https://map.baidu.com
	[65]	Données horaires des PM_2	NAPS CANADA : http://maps-cartes.ec.gc.ca/rnspa-naps/data.aspx
		Mesures météorologiques	National Climatic Data and Information Archive of Environment Canada : http://climate.weather.gc.ca/index_e.html
		Informations sur le brouillard	Environnement et Changement climatique Canada : http://www.ec.gc.ca/infosmog/default.asp?lang%80=%80En&n=669E620B-1
		Densité de population	Statistics Canada : http://www12.statcan.ca/censusrecensement/2006/ref/dict/geo021-eng.cfm
	[75]	Données géographiques	OSM : https://www.openstreetmap.org/
	[79]	Données relatives aux véhicules	https://pubs.acs.org

TABLE 2.1 – Références utiles aux ensembles de données (vérifié le 03/04/2018)

surveillance est faible, les méthodes d'interpolation produisent des résultats de faible précision dû au petit nombre des sites de mesure, contrairement aux réseaux avec des stations de surveillance qui sont plus denses ([76], [81] et [82]). Certains MQAs utilisent les variables complémentaires/de substitution ou supplémentaires pour pallier le manque des données.

Par exemple, l'inclusion d'un indicateur significatif de la pollution atmosphérique comme les informations sur le vent ([70], [79]) produit une bonne amélioration par rapport au modèle traditionnel et considère que la variable météorologique est l'information manquante et nécessaire pour améliorer le modèle. Les variables de substitution aident dans de nombreux cas à pallier le manque de données en estimant les variables réelles à partir d'autres données d'entrée comme les données issues de satellites.

En raison du manque de variables géographiques, Su et al. (2017) utilisent les données de télédétection ETM+ pour couvrir le manque de données et obtenir des résultats de prédiction plus précis avec des données de verdure ou de luminosité du sol. Les données issues de la télédétection par satellite montrent la capacité à expliquer la variation spatiale du NO_2 dans la région du delta de la rivière des Perles en Chine en remplaçant les données industrielles, géographiques et socio-économiques manquantes [71]. Les images satellitaires peuvent être une alternative à d'autres données essentielles, comme lorsque Yu et al. (2016) tentent d'obtenir la longueur des routes et l'état des embouteillages en les étudiant et en les induisant à partir d'images de services cartographiques publics [58].

Cependant, parfois, même les variables de substitution doivent encore être ajustées lorsqu'elles semblent être bruitées [71] ; sinon, il serait inutile de les utiliser. Les variables météorologiques (telles que la vitesse et la direction du vent, et la température) sont également des indicateurs de la qualité de l'air. Leur rôle significatif a été discuté dans plusieurs études, sachant que les éléments météorologiques sont des facteurs clés qui affectent le comportement de la pollution de l'air tels que les émissions de polluants, le transport et la transformation. Par exemple, Seo et al. (2018) [83] examinent l'influence de la météorologie sur les mesures à long terme de la qualité de l'air et les changements observés dans les PM_{10} et l'O_3 qui sont liés aux tendances météorologiques. Ils induisent que l'augmentation à long terme de la vitesse du vent sur la décennie 2002-2012 entraîne une amélioration de la qualité de l'air (en plus des politiques de

contrôle des émissions appliquées), provoquant une ventilation des polluants.

De plus, Kaminska et al. (2018) [84] ont analysé les effets de la pollution atmosphérique en utilisant les conditions météorologiques ainsi que les informations sur le trafic de la ville de Wroclaw en Pologne, et ils ont trouvé que les paramètres météorologiques tels que la vitesse du vent sont les impacteurs les plus importants dans la modélisation des $PM_{2.5}$, en plus du flux de trafic pour les NO_x.

Dans une autre étude sur l'interpolation de la qualité de l'air, les données météorologiques sont intégrées comme prédicteurs en plus d'un ensemble d'informations, où Le et al. (2019) utilisent un modèle d'apprentissage approfondi [85] pour estimer et prédire la pollution de l'air au niveau de la ville de Séoul. Les auteurs examinent la performance du modèle avec différentes combinaisons d'entrées (avec uniquement les données sur la pollution de l'air, les données sur la pollution de l'air et la météorologie, les données sur la pollution de l'air et le volume du trafic, les données sur la pollution de l'air et la vitesse moyenne des véhicules, les données sur la pollution de l'air et la pollution de l'air externe, et la pollution de l'air et tous les facteurs connexes). La meilleure RMSE de l'interpolation et la prévision obtenue est celle de l'ensemble de données incluant les mesures la pollution de l'air et les données météorologiques, encore meilleure que celle introduisant toutes les données d'entrée disponibles, ce qui prouve que les paramètres météorologiques ont le rôle le plus important par rapport aux autres.

À force d'avoir un impact pertinent sur la qualité de l'air, les facteurs météorologiques qui influencent la qualité de l'air sont maintenant même étudiés de façon plus détaillée. Xie et al. (2019) [86] étudient le rôle de différents types de champs de vent sur différents polluants dans la région du delta de la rivière des Perles. Les auteurs ont découvert que les distributions spatiales des $PM_{2.5}$, des PM_{10} et du NO_2 dépendent des modèles de champs de vent. La qualité de l'air changeait en fonction des caractéristiques de chaque type de champ de vent. Par conséquent, le rôle des paramètres météorologiques dans l'étude et l'analyse (par exemple, [87]) de la qualité de l'air s'avère être très important, ainsi que dans le contrôle et l'amélioration de cette dernière (par exemple, [88]).

2.3.3 Validation des MQAs

Il existe plusieurs méthodes de validation pour évaluer la performance d'un MQA. Dans cette section, nous choisissons d'aborder les méthodes les plus courantes, en identifiant les paramètres/mesures utilisés dans les travaux discutés auparavant dans ce chapitre (tableau 2.2). Chaque système, service ou modèle développé doit être validé pour s'assurer qu'il répond à l'objectif visé. L'étape de validation joue un rôle important dans l'évaluation des performances des MQAs, qui ne peut en aucun cas être idéal pour ces deux raisons [89] :

- Les observations expriment des mesures uniques pour des polluants sous des conditions météorologiques, topographiques, etc., propres à une zone géographique, tandis que les modèles de qualité de l'air estiment des moyennes de l'ensemble des données.
- De différentes sources sont responsables des incertitudes des prédictions du modèle, comme la turbulence de la couche atmosphérique, le bruit inclus dans les données d'entrée ou les incertitudes dans la physico-chimie sur laquelle les modèles de la qualité de l'air se basent ([90]-[91]).

L'évaluation fiable des modèles de qualité de l'air consiste à effectuer une analyse statistique des performances, qui fait appel à des mesures spécifiques. La raison pour laquelle on construit un modèle de qualité de l'air définit les paramètres de validation à utiliser, en fonction des circonstances et des conditions du cas étudié. Lorsque le MQA est destiné à évaluer l'état de santé, le paramètre de validation à utiliser doit rechercher les causes les plus corrélées (dans les données d'entrée du modèle) avec la détérioration de la santé, ce qui est différent de l'évaluation de la qualité d'une prédiction où nous recherchons le modèle le plus précis possible.

Les différentes applications/technologies des MQAs nécessitent des paramètres de validation différents pour évaluer les performances par diverses manières, et c'est la raison pour laquelle il n'existe pas de mesure générale adaptée à tous les cas et dans toutes les conditions de la modélisation de la qualité de l'air [92].

Mesures utilisées pour la validation

Nous ne donnons pas une description exhaustive de tous les paramètres possibles de l'évaluation de la qualité de l'air dans cette étude, mais nous introduisons les plus courants dans les études discutées précédemment, et qui sont :

- L'erreur quadratique moyenne (en anglais Root Mean Square Error, RMSE) est adoptée comme critère décisif de la performance du modèle dans de nombreuses études sur la pollution de l'air, RMSE exprime la différence entre les valeurs prédites par un modèle et les valeurs réellement observées par la formule détaillée en chapitre 3 section 3.2.4.

- R^2, désigné par R^2 ou r^2, est le coefficient de détermination représentant la proportion de la variance de la variable expliqué qui est prévisible à partir des variables explicatives, additionnée de la variation expliquée/variation totale (détaillé plus en section 3.2.4).

- R, le coefficient de corrélation désigné par R ou r, est utilisé dans de nombreux travaux pour mesurer la corrélation entre les variables (ensembles de données d'entrée) et entre les valeurs observées et modélisées également, pour voir dans quelle mesure elles sont corrélées.

- La valeur p est la corrélation entre les variables, pouvant être évaluées par la valeur p, que nous comparons au niveau de signification et de contribution dans l'estimation/prévision (généralement représenté par $\alpha = 0{,}05$). Dans le cas p la corrélation est donc différente de 0. Sinon, il est impossible de conclure que la corrélation est différente de 0.

- La validation croisée est l'une des techniques les plus courantes pour évaluer la variation de la performance de prédiction d'un modèle. En divisant l'ensemble de données de manière aléatoire en X sous-groupes, le modèle serait réajusté X fois, chaque sous-groupe étant retiré à son tour de l'ensemble d'apprentissage, sachant qu'une partie des données sert à ajuster les différents modèles (apprentissage) et le reste à mesurer la performance prédictive du modèle par les erreurs de validation (ce qui pourrait être fait par l'une des mesures discutées ci-dessus). À la fin, le modèle le plus performant est adopté [93]. La validation croisée est également une bonne solution pour détecter les problèmes de surapprentissage.

Le tableau 2 présente les paramètres/mesures de validation utilisés dans les articles discutés précédemment.

Les MQAs peuvent également être évalués par rapport à la couverture spatiale ou temporelle

des données de validation. Par exemple, les modèles RUT sont surtout utilisés pour l'analyse spatiale plutôt que pour la dimension temporelle. Les échelles spatiales peuvent être différentes les unes des autres, par exemple, il peut s'agir d'une ville ([46], [49], [70]), d'un certain nombre de villes ([48] et [69]), ou même d'un État entier [51]. En outre, certaines études qui traitent de la prédiction spatiotemporelle peuvent avoir des échelles spatiales et temporelles différentes. Par exemple, des prévisions bihebdomadaires pour différents sous-comtés de Californie [63], des prévisions quotidiennes pour la région frontalière du Mexique [53], 74 villes de Chine [57], Montréal [65] et Mexico [66], des prévisions horaires pour [54], [58], [67] et [74], et même une prévision en temps réel pour [62].

Références	Paramètres de validation
[50]	-Facteurs de variance et d'inflation (VIF) -Distances de Cook -Statistique I de Moran -Test de Chow -Biais moyen normalisé (NMB) -Erreur moyenne normalisé (NME) et validation croisée (CV)
[72]	Le meilleur fitness et le R^2 ajusté le plus élevé
[70]	Validation croisée de l'exclusion (LOOCV), R^2, et R^2
[71]	R^2, RMSE, LOOCV et validation croisée régionale (RCV)
[48]	CV, RMSE et pourcentage d'erreur moyen absolu (MAPE)
[51]	CV, R^2, taux d'erreur moyen (MER) et RMSE
[59]	LOOCV
[53]	RMSE, R^2, et erreur moyenne absolue (MAE)
[55]	Corrélation avec les mesures réelles
[58]	Précision, recall, score F, erreur relative absolue (ERA) et CV
[60]	CV et R^2
[63]	LOOCV
[62]	Recall et précision
[69]	R^2, RMSE, NME, biais moyen (MB), NMB et R
[74]	CV, RMSE, et R
[65]	LOOCV, RMSE, et R^2
[66]	CV et RMSE
[94]	R^2

TABLE 2.2 – Exemples de paramètres de validation utilisés pour la pollution atmosphérique

2.3.4 Recommandations pour développement des MQAs

L'objectif de la modélisation de la qualité de l'air est non seulement d'obtenir les résultats les plus précis en termes de prédiction/prévision de la pollution atmosphérique, mais aussi

de détecter les principaux facteurs déterminant la qualité de l'air. Certaines techniques de la gestion de la qualité de l'air qui étudient l'aspect de cause-effet entre la pollution de l'air et l'environnement permettent de tirer des conclusions sur les principales raisons/sources de la pollution de l'air. Les deux modèles de cause- effet que nous avons rencontrés (qui donnent une explication des effets existants en trouvant les causes) sont le RUT et les forêts aléatoires, qui permettent d'accéder aux principaux facteurs contribuant à la qualité de l'air. Dans cette section, nous recommandons le MQA à adopter en fonction des données disponibles, en d'autres termes, les techniques possibles à employer selon les variables disponibles dans l'ensemble de données.

Recommandations pour la mise en place des MQAs

Sur la base de plusieurs travaux sur la modélisation de la qualité de l'air, nous élaborons un ensemble de recommandations qui aident, lors de l'élaboration d'un MQA, à choisir les données d'entrée appropriées pour une certaine méthode et vice versa. Les principales recommandations tirées sont :

I) L'utilisation des indicateurs de trafic avec la RUT en même temps que les concentrations de polluants, en raison des bonnes performances qu'elle montre dans diverses études. Pour les modèles de RUT, dans la plupart des cas, les prédicteurs liés au trafic sont les données d'entrée les plus influentes parmi les variables données, en plus des informations liées aux variables d'utilisation des terrains/land use data [71]. Dans le cas de l'étude de Ross et al. [49], seules les informations relatives au trafic représentent plus de 54 % de la variation des polluants étudiés. En outre, Lee et al. (2014) constatent que la longueur des routes principales, les zones vertes urbaines, les zones semi-naturelles et les zones boisées sont les indicateurs les plus significatifs [70]. Même en utilisant trois modèles différents de RUT comme dans l'article [48], où les mesures de la période d'entrée sont différentes ainsi que les variables utilisées (un modèle avec 28 comtés pour la période de (1999-2001), le second pour la même période mais seulement pour 9 comtés, et le dernier pour l'hiver 2000 pour 28 comtés), le trafic explique la plus grande partie de la variance de la pollution atmosphérique (37-44%) dans tous les modèles, suivi par l'indicateur de densité de population. De plus, lorsqu'ils prédisent la pollution

atmosphérique à Los Angeles [50], Su et al. (2009) utilisent plusieurs variables comme entrées du modèle RUT, et comme la zone étudiée est proche de la route, l'impact du trafic local est évidemment le plus significatif, en ignorant tous les autres contributeurs. Compte tenu des critères saisonniers, le trafic est identifié comme étant le facteur le plus significatif en été, et la densité de population en hiver pour le NO_2 [70].

De plus, Moore et al. (2007) prouvent que le volume du trafic est l'un des facteurs les plus importants de plus des zones industrielles et gouvernementales [78]. Zhai et al. (2016) confirment que le trafic reste le facteur le plus important avec les variables d'utilisation des sols/land use data à comparaison à la donnée de répartition de la population et à la distance par rapport à la côte, et cela est dû à la grande urbanisation et au trafic routier intensif à Houston, aux États-Unis [51].

II) La technique de forêt aléatoire (Random Forest) représente une bonne option pour prédire la qualité de l'air lorsqu'on dispose de données de détection urbaine, telles que les points d'intérêt (POI), les données de substitution de la part des fournisseurs de cartes publiques et d'autres informations pertinentes, comme indiqué dans [58]. Les indicateurs d'utilisation des terrains/land use data constituent un autre cas d'utilisation de cette technique. Brokamp et al. (2017) emploient les indicateurs d'utilisation des terrains comme prédicteurs à travers la technique des forêts aléatoires pour modéliser la qualité de l'air produisant une bonne précision en se basant sur ces données [59].

III) Les données des médias sociaux peuvent fournir des informations utiles pour l'estimation de la qualité de l'air pour le cas des citoyens qui expriment leur opinion sur la pollution de l'air dans leur ville par le biais des médias sociaux. L'apprentissage automatique est recommandé dans ce cas, par exemple le gradient tree boosting (GTB) [55] qui résout les problèmes de classification et de régression. L'avantage des données des médias sociaux est qu'elles permettent de connaître le niveau de pollution de l'air dans des lieux non surveillés, en particulier dans les grandes villes, comme le montre le travail de [95], qui utilisent l'apprentissage automatique pour mesurer la pollution de l'air. Fontes et al. n'incluent pas dans le travail de [64] des prédicteurs tels que les données météorologiques, les informations sur le trafic, la distance par rapport à l'océan et les émissions industrielles, les seules données disponibles dont

ils disposent sont les concentrations de polluants. Dans ce cas, les techniques de krigeage et d'IDW permettent d'estimer les polluants atmosphériques en interpolant les mesures des stations de surveillance. Ainsi, même dans la situation où nous n'avons que les concentrations de polluants, l'estimation de la pollution de l'air est toujours possible grâce aux méthodes de krigeage et d'IDW, avec toujours une bonne précision comme dans les travaux de [96] et [97]. Les données de télédétection par satellite peuvent prédire l'indice de la qualité de l'air ; ainsi, dans [74] Guo et al. (2016) sont capables de donner des résultats raisonnablement bons en appliquant des algorithmes de régression à l'aide de données satellitaires. Ce type de données vient pallier les limites des outils de surveillance traditionnelle en termes de couverture et résolution spatiales et donne des résultats de bonne précision lorsqu'elles sont réalisées par des modèles de régression ([98]-[99]).

IV) Les $PM_{2.5}$ et les NO_x sont souvent sélectionnés pour l'étude de la qualité de l'air (les polluants les plus examinés dans les articles discutés auparavant), car les $PM_{2.5}$ sont l'un des polluants les plus dangereux pour la santé humaine et l'environnement, et les NO_x sont un bon traceur de la pollution liée au trafic. Nous conseillons d'utiliser des MQAs basés sur l'apprentissage automatique pour analyser les $PM_{2.5}$ et les PM_{10}, tandis que les modèles RUT (par exemple [51], [71], [72]) et les modèles hybrides (par exemple [60], [63] et [80]]) sont généralement utilisés pour modéliser les concentrations de $PM_{2.5}$ et de NO_x. Pour étudier les polluants liés au trafic tels que les gaz de CO_2 [94] et les NO_x [79] et les particules telles que les hydrocarbures aromatiques polycycliques liés aux particules (PB-PAH), le nombre de particules (PNC) [79] et les particules ultrafines (UFP) [100], il est préférable d'utiliser des modèles mathématiques pour prédire les concentrations de ces polluants sur les routes.

2.4 Conclusions

En se basant sur plus de 40 travaux sur la qualité de l'air, nous avons conclu que les principales méthodes utilisées dans l'estimation de la pollution atmosphérique sont la méthode de RUT, l'apprentissage automatique et les méthodes hybrides. De plus, l'insertion des variables de trafic dans la méthode de RUT, améliore la performance de cette dernière. Par ailleurs,

l'utilisation des techniques de krigeage et de pondération de distance inverse est recommandée, quand le jeu de données d'entrée n'est pas dense.

Bien que de nombreuses limitations puissent affecter les performances du modèle (par exemple, le manque ou la mauvaise qualité des ensembles de données), il existe des alternatives et des moyens efficaces pour construire un modèle plus précis (par exemple, des données complémentaires ou des techniques d'hybridation). Étant donné qu'il existe plusieurs options possibles pour construire des MQAs, cette étude pourrait servir d'ébauche introductive pour la sélection des ensembles de données et des techniques, qui couvre les éléments essentiels des MQAs : techniques existantes, types d'ensembles de données et méthodes de validation, en mettant l'accent sur les limites et les points forts des MQAs. La modélisation de la qualité de l'air pourrait être réalisée par une multitude de méthodes disponibles, mais revenir à l'objectif de la création d'un MQA aide et détermine les techniques à utiliser. A travers cette revue, nous présentons une sorte de ligne directrice à toute personne intéressée par l'élaboration d'un MQA en détaillant chaque élément de ce dernier.

Chapitre 3

Influence de la variabilité spatiale et temporelle de la pollution atmosphérique sur la précision des méthodes d'interpolation

Dans ce chapitre, nous décrivons d'abord les méthodes d'interpolation utilisées pour estimer la pollution de l'air de la région Hauts-de-France, et ensuite nous introduisons les résultats de l'application de ces dernières qui ont fait l'objet d'un l'article, publié dans le journal Pollution Research (ISSN : 0257-8050, Ref. No. PR-F-2249).

Nous comparons dans ce présent chapitre les performances d'un ensemble de techniques d'interpolation spatiale pour estimer les concentrations de PM_{10} de la région des Hauts-de-France. Cette estimation a été évaluée par le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE) et les intervalles de confiance à 95% correspondants. Ces mesures sont utilisées pour examiner la performance des méthodes de triangulation, la version classique et optimisée de la méthode de pondération inverse de la distance, ainsi que la régression de processus gaussien avec deux noyaux différents. La distribution spatiale de l'erreur de l'estimation est étudiée pour analyser la dépendance des concentrations PM_{10} vis-à-vis des zones industrielles et des phénomènes atmosphériques côtiers.

Pour évaluer l'influence de la météorologie locale de la région Hauts-de-France sur la dispersion de la pollution de l'air, nous avons estimé le coefficient de détermination de l'interpolation des données PM_{10} moyennées sur différentes échelles temporelles. Nous avons moyenné les données sur des périodes, puis évalué le coefficient de détermination pour chaque période.
La sensibilité des techniques d'interpolation utilisées au bruit introduit dans les données traitées ainsi qu'à la densité de ces dernières a été vérifiée. En outre, le rôle de la densité des données dans la précision de performance de l'interpolation dans l'estimation des concentrations PM_{10} a été mis en évidence.
Ce chapitre est subdivisé en 6 parties : la section 3.1 aborde les notions fondamentales de l'interpolation, suivi d'un état de l'art dans la deuxième section. La section 3.3 introduit un descriptif de la pollution particulaire en région Hauts-de-France et les données utilisées. La section 3.4 décrit les contributions apportées du travail de ce chapitre Ensuite, la section 3.5 vient présenter et discuter les résultats obtenus de l'application de l'interpolation spatiale. Et la dernière section est consacrée aux conclusions et perspectives du présent chapitre.

3.1 Notions fondamentales : méthodes d'interpolation spatiale

Nous présentons ici une explication des méthodes d'interpolation utilisées pour l'estimation des concentrations PM_{10} de la région Hauts-de-France.

3.1.1 Méthodes basées triangulation

Les quatre premières méthodes que nous avons appliquées sont des méthodes basées sur la triangulation. Une explication brève de cette dernière est introduite avant de passer aux méthodes d'interpolation.

- Triangulation de Delaunay

  Etant donnée un ensemble $P = \{p_1, ..., p_n\}$ de sites d'un plan, la triangulation de P est sa subdivision en triangles dont les sommets sont les sites de P. Une triangulation est dite de Delaunay si le cercle circonscrit à chaque triangle ne contient aucun site de

P en son intérieur [101].

- Diagramme de Voronoi

  La triangulation de Delaunay est le duel du diagramme de Voronoi. Pour un ensemble $P = \{p_1, ..., p_n\}$ de sites d'un plan, le diagramme de Voronoi est la partition de ce plan en n cellules (polygones convexes), une pour chaque site de P, avec la propriété qu'un point q se trouve dans la cellule correspondante à un site pi si et seulement si : distance $(p_i, q) < d(p_j, q)$, pour i distinct de j [102]. En sachant que la distance est définie comme étant la distance qui sépare les deux points : $||p_i - q||$ qui désigne la distance euclidienne entre p_i et q (distance (p_i, q)).

Voisin le plus proche

L'interpolation est une approximation de la valeur d'une fonction pour un certain point à valeur non connue dans un espace, par le biais des autres points (voisins), où la valeur de cette fonction est donnée. Pour notre cas d'étude, nous avons sélectionné un ensemble de méthodes d'interpolation, les plus courantes dans la littérature, tout en s'intéressant à l'optimisation de la méthode de pondération inverse à la distance. L'algorithme du voisin le plus proche sélectionne (en se basant sur la décomposition par le diagramme de Voronoi) la valeur du point le plus proche en ignorant les autres valeurs des points voisins, ce qui donne un interpolant constant par morceaux. Cette méthode est une des techniques les plus basiques, stables et moins coûteuses en termes de calcul, parmi tous les algorithmes d'interpolation.

Interpolation linéaire

L'interpolation linéaire est un algorithme qui construit une triangulation de Delaunay en connectant les points donnés pour former un ensemble de triangles, de telle manière qu'aucune arête de triangle ne soit intersectée par d'autres triangles. Le résultat est un maillage triangulaire étendu sur la grille des points d'entrée. Chaque triangle définit un plan au-dessus des points de la grille se trouvant dans le triangle, avec l'inclinaison et l'élévation du triangle déterminés par les trois points de données d'origine définissant le triangle. Tous les noeuds de la grille dans un triangle sont définis par la surface triangulaire. Pour l'estimation de la valeur

d'un point s, situé au sein du triangle formé par les sommets s1, s2 et s3, dans un espace 3D, et en sachant que $s1 = (x1, y1, z1), s2 = (x2, y2, z2)$ et $s3 = (x3, y3, z3)$, on applique l'équation suivante :

$$Z(s) = Z(x, y) = \alpha x + \beta y + \gamma \tag{3.1}$$

Avec, x et y sont les coordonnées de l'emplacement du point s. La solution est obtenue à partir d'une combinaison linéaire des observations aux sommets du triangle en résolvant le système :

$$\begin{cases} x_1 + \beta y_1 + \gamma = z_1 \\ x_2 + \beta y_2 + \gamma = z_2 \\ x_3 + \beta y_3 + \gamma = z_3 \end{cases} \tag{3.2}$$

Avec z_1, z_2 et z_3 sont les valeurs au niveau des points s_1, s_2 et s_3. La résolution de ce système fournit les coefficients nécessaires (α, β, et γ) au calcul de tout point inclus dans le triangle.

Voisin naturel

Le voisin naturel fournit une estimation qui est plus lisse que celle du voisin le plus proche. Cette technique est basée sur la pondération par surface au lieu de la distance, elle construit la triangulation de Delaunay des points d'entrée et sélectionne la cellule (polygone) la plus proche qui forme l'enveloppe convexe autour du point à interpoler et utilise la zone proportionnelle comme poids [103]. La première étape consiste à construire un premier diagramme de Voronoi à partir des points d'entrée xi à valeurs connues, puis un nouveau diagramme de Voronoi autour du nouveau point ajouté x, dont nous voulons estimer la valeur. Le calcul du poids se fait par la part du volume du nouveau polygone crée par l'insertion de x, vis-à-vis le diagramme initial. Chaque voisin naturel aura son propre poids en divisant le volume du nouveau polygone sur le volume de l'intersection entre le nouveau polygone de x et les anciens polygones de ces voisins naturels. La valeur de x est ainsi calculée par cette équation, dans un espace 2D :

$$g(x) = \sum_{i=1}^{n} w_i(x) a_i \tag{3.3}$$

Avec g(x) est l'estimation au point x, par l'ensemble des poids w_i calculés des points d'entrée à valeurs connues ai. Et le poids w_i est calculé par la formule suivante :

$$w_i(x) = \frac{A(x_i)}{A(x)} \tag{3.4}$$

Où A(x) est le volume du nouveau polygone centré en x, et $A(x_i)$ est le volume de l'intersection entre le nouveau polygone centrée en x et le diagramme initial des x_i [103].

Splines cubiques

La méthode de Spline cubique utilise des polynômes d'ordre 3 par morceau pour joindre chaque paire des n points d'entrée consécutifs, pour une courbe qui soit continue. Une Spline cubique $S = \{f(t) | t \in [0,n]\}$, est une fonction de classe C^2, ayant une dérivée seconde continue et non nulle. Pour n points d'entrée, nous avons n-1 intervalles et n-1 polynômes ; $f_1, ..., f_{n-1}$. Dans un espace 2D pour un point aux coordonnées (x, y), chaque polynôme f, a une équation de la forme :

$$\begin{aligned} x(t) &= a_x t^3 + b_x t^2 + c_x t + d_x \\ y(t) &= a_y t^3 + b_y t^2 + c_y t + d_y \end{aligned} \tag{3.5}$$

a, b, c et d sont déterminés par la résolution d'un système qui provient des conditions d'égalité des dérivées secondes à tous les points interpolés. Ces coefficients sont différents pour chaque intervalle. Spline crée par la fin, une surface qui passe par les points d'entrée fournis, en présentant le moins de changements de pente possible à tous les points par l'ensemble des polynômes [104].

3.1.2 Pondération inverse à la distance (Inverse Distance Weighting, IDW)

Nous avons utilisé deux versions de la méthode de pondération inverse à la distance. Dont la première est la version classique qui applique la formule générale d'IDW, et la deuxième est avec une optimisation au niveau du paramètre de la puissance de distance.

Version classique

L'estimation de la valeur d'intérêt au niveau du site à interpoler est calculée par la pondération inverse de la distance séparant ce site des autres sites à valeurs connues [105]. Cette méthode admet que l'influence d'un site par rapport à un autre soit proportionnelle à la distance qui les séparent. Autrement dit, les points xi les plus proches de la position cible x à estimer, sont supposés avoir une plus grande influence sur la valeur interpolée que ceux plus éloignés. La fonction d'interpolation d'IDW est définie par, dans un espace 2d :

$$y(x) = \frac{\sum_{i=0}^{N} w_i(x) y_i}{\sum_{i=0}^{N} w_i(x)} \tag{3.6}$$

Où :

$$w_i(x) = \frac{1}{d(x,x_i)^p} \tag{3.7}$$

Avec y(x) est la valeur interpolée à la position x à partir des valeurs connues y_i des N points x_i, et dont chacun à un poids w_1 représenté par la distance entre x et xi avec un exposant p. La valeur de p pour cette version d'IDW est 1.

Version optimisée

Dans la plupart des travaux qui utilisent la pondération inverse à la distance, la puissance p est fixée à une valeur 2 par défaut ([106] ; [107] ; [108]), comme mentionné dans la partie contexte de ce chapitre. Nous avons opté à chercher la valeur optimale de cette puissance pour chacune des situations étudiées de la pollution de l'air (pour chaque moment t). Cette optimisation vise la minimisation de la différence entre les valeurs observées (les mesures PM_{10}) et les valeurs estimées, c'est-à-dire elle minimise l'erreur de l'interpolation. Cette erreur est calculée par une validation croisée (Leave-one-out cross-validation, [109]), qui consiste à retirer une des valeurs observées (les mesures) et l'estimer, dans notre cas par IDW, puis calculer la différence entre les valeurs observées et leur estimation. Ensuite, reproduire ce processus pour chaque mesure de l'ensemble.

Nous avons choisi d'optimiser la puissance de la distance de la pondération inverse (IDW) par la méthode recherche de section dorée. Cet algorithme est utilisé pour trouver le minimum

local (ou le maximum) dans un espace 1D, d'une fonction unimodale (une fonction unimodale a un seul minimum ou maximum dans un intervalle [a, b]) f dans un intervalle donné [110]. La définition générale de cette technique se résume dans la réduction de l'intervalle de recherche qui contient le minimum, en chaque nouvelle itération de l'algorithme. Pour un intervalle donné [a,b], il existe un point $c \in [a,b]$ tel que $f(c) < f(a)$ et $f(c) < f(b)$, donc [a,b] délimites le minimum de f, et c est une valeur approximative au minimum de f. La réduction de la taille de l'intervalle se fait par l'introduction de nouveau point dans cet intervalle. A titre d'exemple, si on prend un point $d \in [a,b]$ que si $f(d) > f(c)$, d devient la nouvelle borne de l'intervalle et remplace soit a ou b (selon la position de c), pendant que c représente toujours la meilleure estimation trouvée du minimum. Et si le contraire, $f(d) < f(c)$ c sera la nouvelle borne, tandis que d devient la nouvelle meilleure estimation du minimum de f, et dans les deux cas on obtient un intervalle de recherche qui est plus petit que l'initial. Ce processus de recherche continue jusqu'à ce qu'une condition d'arrêt soit rencontrée. Cette dernière décide si la terminaison de l'algorithme aura lieu ou non, selon les critères d'arrêts donnés. Par exemple, imposer un certain nombre d'itérations que l'algorithme ne doit pas dépasser, fixer une valeur pour la distance séparant a de b et qu'une fois aboutit la recherche s'arrêtera, etc. Enfin, la dernière valeur de f(x) obtenue représentera le minimum global de cette fonction f.

3.1.3 Régression des processus Gaussiens

Dans la régression des processus Gaussiens [111], la sortie y de la fonction f en un point x est définie par :

$$y = f(x) + \varepsilon \tag{3.8}$$

Avec x, y et ε sont des vecteurs aléatoires d'une dimension n. ε est l'erreur de la représentation en (3.8) que nous supposons qu'elle suit une distribution normale, avec une espérance $E[\varepsilon] = 0$. Cette méthode suppose que la fonction f(x) est décrite par une distribution Gaussienne (Gaussian Process, GP) [112] :

$$f(x) \sim GP(m(x), k(x,x')) \tag{3.9}$$

Un processus Gaussien (GP) est une distribution gaussienne sur des fonctions définie par une moyenne et une fonction de covariance. L'espérance de la moyenne m(x) reflète les valeurs espérées au point x :

$$m(x) = E[f(x)] \tag{3.10}$$

m(x) est souvent fixée à 0 (m (x) = 0), afin d'éviter des calculs postérieurs coûteux et de faire l'inférence à partir de la fonction de covariance seulement. La fonction de covariance k(x, x') modélise la dépendance entre les valeurs de la fonction (f(x)) à différents points d'entrée x et x' par :

$$k(x,x') = E[(f(x) - m(x))(f(x') - m(x'))] \tag{3.11}$$

Cette fonction de covariance est aussi appelée noyau (ou kernel en anglais). Il est choisi en se basant sur des hypothèses, telles que la régularité et les modèles probables espérés des données d'entrée. L'hypothèse générale assumée est que la corrélation entre deux points diminue avec la distance qui les sépare. Les points les plus proches devraient se comporter de manière plus similaire que les points qui sont plus éloignés les uns des autres. Le noyau le plus utilisé et qui prend en considération cette hypothèse est le noyau d'exponentielle au carré (voir ci-dessous).

Noyau exponentiel au carré

Le noyau exponentiel au carré est défini par :

$$k(x,x') = \sigma_f^2 exp[-\frac{1}{2}\frac{(x-x')^2}{\sigma_l^2}] \tag{3.12}$$

Où (x-x') représente la distance euclidienne qui sépare x et x'. Les deux paramètres σ_f (l'écart type de f(x)) et σ_l (la longueur caractéristique) peuvent être modifiés, pour augmenter ou réduire la corrélation a priori entre les points x et x', et par conséquent ajuster la variabilité de la fonction résultante.

Noyau d'exponentielle au carré avec une détermination automatique du paramètre de longueur caractéristique

Nous avons utilisé un deuxième noyau qui diffère du premier par la possibilité d'adopter σ_l qui s'ajuste automatiquement (par la méthode de quasi-newton [113], proposé par Matlab) selon les différents points de x d'entrée en question, tandis que pour le premier noyau ce paramètre reste fixe pour les différents points traités.

$$k(x,x') = \sigma_f^2 exp[-\frac{1}{2}\sum_{m=1}^{d}\frac{(x_m - x'_m)^2}{\sigma_m^2}] \tag{3.13}$$

Avec, σ_m ,$m \in [1,..,d]$ est la longueur caractéristique avec une dimension d. σ_m variera selon chaque x qui sera défini pour chacun des prédicteurs de dimension d, contrairement au noyau précédent où nous trouvons une échelle de longueur correspondante pour l'ensemble des prédicteurs.

Une fois qu'une fonction moyenne et un noyau sont choisis, nous utilisons le processus Gaussien pour modéliser les valeurs de fonction a priori, et suite à cette modélisation nous pouvons ainsi estimer les valeurs postérieures de la fonction d'intérêt.

3.1.4 Procédure et mesures d'évaluation de la précision

Afin de pouvoir comparer et juger la précision des mesures estimées à l'aide des différentes méthodes d'interpolation spatiale, nous avons utilisé les techniques suivantes :

Validation croisée d'un contre tous

La validation croisée d'un contre tous (en anglais Leave One Out Cross Validation, LOOCV), est un moyen d'apprentissage automatique permettant l'estimation de la fiabilité d'un modèle ou d'une méthode. Le nombre de données disponibles étant limité, cette technique permet de généraliser les estimations à un ensemble de données quasi indépendant. LOOCV consiste à retirer l'une des valeurs de données observées de l'ensemble des points d'entrée, et à essayer de l'estimer par une méthode, puis à calculer la différence entre les valeurs observées et estimées (erreur d'estimation). Ensuite, reproduire ce processus pour chaque

mesure de l'ensemble. Nous avons appliqué LOOCV pour évaluer deux mesures statistiques : l'erreur quadratique moyenne (RMSE, [114]) et le coefficient de détermination R^2 [115]. Le premier permet d'évaluer la précision du modèle et le second mesure la performance par le pourcentage de variance expliquée des données.

Nous calculons le RMSE pour chaque station à part (RMSE station), de plus du RMSE pour l'ensemble des données de toutes les stations (RMSE total) comme suit :

$$RMSE(station j) = \sqrt{\frac{1}{n_j}\sum_{i=1}^{n}(X_{obs,i,j} - X_{estim,i,j})^2} \tag{3.14}$$

où $X_{obs,i,j}$ sont les valeurs mesurées et $X_{estim,i,j}$ sont les valeurs estimées des concentrations de PM_{10} à la station j au temps i, et n_j est le nombre de mesures disponibles en j.

$$RMSE(total) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_j}\sum_{j=1}^{n}(X_{obs,i,j} - X_{estim,i,j})^2} \tag{3.15}$$

où m est le nombre de stations.

Le coefficient de détermination R^2 est défini comme suit :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_{obs,i} - X_{estim,i})^2}{(X_{obs,i} - \overline{X_i})^2} \tag{3.16}$$

Le R^2 quantifie la préicision des valeurs estimées par le modèle/la méthode en question par rapport aux valeurs observées.

Technique de Bootstrap

Pour évaluer les intervalles de confiance [45] des estimations statistiques que nous avons obtenues, nous avons utilisé la méthode de rééchantillonnage "bootstrap". La technique du bootstrap [116] consiste à rééchantillonner les données observées que nous utilisons pour évaluer les performances d'un algorithme, pour calculer l'incertitude du paramètre de validation (dans notre cas c'est le RMSE et le R^2). Cette technique sélectionne à plusieurs reprises et avec remplacement des échantillons aléatoires de même taille à partir de l'ensemble de données d'origine. Ainsi, chaque donnée de la base de données d'origine peut être sélectionné aucune,

une ou plusieurs fois dans chaque échantillon bootstrap. Ensuite, nous calculons la valeur du paramètre d'évaluation pour chaque échantillon bootstrap plusieurs fois, nous classons ces valeurs et prenons l'intervalle de confiance (IC) à 95% ([45] ; [117]).

3.2 Etat de l'art

Comme indiqué dans le chapitre 1, il existe une multitude de méthodes qui permettent l'estimation de la qualité de l'air, desquelles nous avons sélectionné les plus utilisées et que nous détaillons et comparons dans ce chapitre. Parmi ces techniques il y a la pondération inverse à la distance (Inverse Distance Weighting, IDW) avec une version optimisée que nous avons évaluée. La principale difficulté rencontrée lors de l'application de l'IDW est la définition de la valeur du paramètre de puissance [118]. Cela se fait généralement avant l'application de l'algorithme. L'approche habituelle pour trouver la valeur optimale du paramètre de puissance est par recherche exhaustive, en échantillonnant toutes les valeurs possibles dans un intervalle donné, à une taille de pas choisie et déterminée en entrée [119]. Les résultats de cette méthode dépendent de l'intervalle de recherche, comme dans les travaux de Li et al. ([120] ; [121]), où ce dernier a été fixé entre les valeurs de 1 à 5, avec un pas de 0.5. D'autres auteurs adoptent une valeur de p donnée qui est le plus souvent égale à 2, car c'est la valeur la plus utilisée par défaut dans la littérature, mais sans avoir une justification théorique ou expérimentale, ([106] ; [107] ; [108]).

Bien que la meilleure valeur d'exposant varie en fonction de la nature, les données étudiées diffèrent d'un domaine d'étude à un autre, et d'une situation à une autre comme démontré par Pasini et al., 2015. Selon les périodes d'échantillonnage le meilleur exposant (4.0, 3.0 et 6.0) changeait de valeur [122], où l'exposant p continuait à changer de valeur pour les différentes périodes d'échantillonnage traitées. De Mesnard (2013) a montré que l'exposant p variait selon la nature physique des données étudiées, et devrait être déduit de la forme de pollution rencontrée avec un raisonnement élémentaire pour chaque situation [118]. Pour cela, nous avons opté de faire varier la valeur de p pour chaque moment t évalué pour les données que nous étudions. Cette évaluation a été effectuée par l'algorithme de recherche de

section dorée (recherche d'un minimum), d'ores et déjà décrit dans la partie 3.1.2 du présent chapitre.

Le deuxième point examiné est le lien entre la qualité de l'air et la météorologie. L'un des éléments qui rendent la pollution de l'air difficile à étudier est le caractère imprévisible des phénomènes météorologiques, qui ont une grande influence sur la dispersion des polluants dans l'air [123]. Plusieurs travaux se sont intéressés à l'étude de la variabilité de la pollution de l'air vis-à-vis de ces phénomènes météorologiques ([124] ; [125] ; [126]), où ils ont montré la forte incidence de ces derniers dans le comportement des polluants dans l'air. Cette influence aboutit à une périodicité observée dans la pollution de l'air, qui ressemble à celle de ces phénomènes. D'ailleurs, cela a été repéré lors d'une estimation par interpolation pour des données moyennées avec des résolutions temporelles différentes ([127] ; [128]). Cet impact de variabilité temporelle de la pollution de l'air sur la précision des méthodes d'interpolation spatiale est examiné dans ce chapitre menant à des résultats similaires à ceux trouvés par [127] et [128]. De même, dans ce chapitre nous examinons cet impact pour les concentrations PM_{10} de la région des Hauts-de-France.

En proposant de travailler avec des échelles temporelles différentes des données moyennées de ces polluants, nous étudierons leur variabilité temporelle ainsi que l'influence de cette dernière sur la précision des méthodes d'interpolation.

3.3 Pollution particulaire en région Hauts-de-France et données utilisées

Dans cette section, nous proposons de présenter et d'analyser la pollution atmosphérique en région Hauts-de-France (décrite en chapitre 1 section 1.1.4), à travers les mesures du réseau de surveillance fournies par ATMO Hauts-de-France, pour une meilleure interprétation des résultats de l'interpolation spatiale en section 3.5. La carte présentée en figure 3.1, montre l'emplacement des 37 stations ayant pour rôle de mesurer de différents polluants présents dans la région en question.

Nous remarquons que la partie nord du réseau de stations de surveillance est plus dense dans

les zones industrielles et urbaines que dans la zone rurale au sud de la région. Certaines stations se trouvent à proximité des industries, comme dans l'agglomération dunkerquoise (DKI, DKC, et DKG), tandis que d'autres sont déployées à côté du port de pêche (tel que les stations BO de Boulogne). Alors que d'autres stations sont plutôt destinées à la surveillance de la qualité de l'air au niveau des zones urbaines ayant un trafic routier et une densité de population importants, comme à Lille (MN1 et MC7), Valenciennes (les stations VA), et tout au long de la côte d'Opale (les stations BO et CA).

L'influence de la nature de ces émissions sur le comportement de la pollution atmosphérique est discutée ultérieurement dans la section 3.5.

Les données utilisées sont des mesures quarts horaires des concentrations PM_{10} du premier janvier 2016 jusqu'au 31 décembre de la même année. Le cadre blanc sur la figure 3.1 délimite le site d'étude où l'interpolation spatiale a été appliquée et représentée sur les différentes cartes introduites dans ce chapitre.

Le prétraitement des données d'entrée a consisté d'abord à éliminer les valeurs aberrantes (valeurs supérieures aux valeurs réelles possibles des concentrations de PM_{10}), puis à supprimer toutes les valeurs négatives et les mesures invalides représentées par NAN (Not A Number). Dans la plupart des cas, ces valeurs résultent d'un dysfonctionnement des stations/capteurs de mesure.

Pour une meilleure interprétation des résultats de l'interpolation spatiale des concentrations de PM_{10}, nous analysons d'abord le comportement de ces dernières auprès de certaines stations par rapport à la moyenne régionale (tableau 3.1 et figures de 3.2 à 3.5), dans l'objectif d'étudier leur variabilité à l'échelle régionale. Pour cela, sur les 37 stations disponibles, nous en avons sélectionné 12 pour mettre en évidence le comportement des stations que nous avons classées en quatre catégories : côtières, urbaines, rurales et industrielles, en fonction de la nature des émissions auxquelles elles sont exposées. Pour chaque catégorie, nous avons tracé les mesures obtenues à partir de 3 stations, ainsi que la moyenne régionale pour un mois, à savoir, août 2016, compte tenu de la variabilité des concentrations de PM_{10} qui devient plus forte en raison des phénomènes météorologiques qui se produisent en cette période de l'année, comme la brise de mer [129].

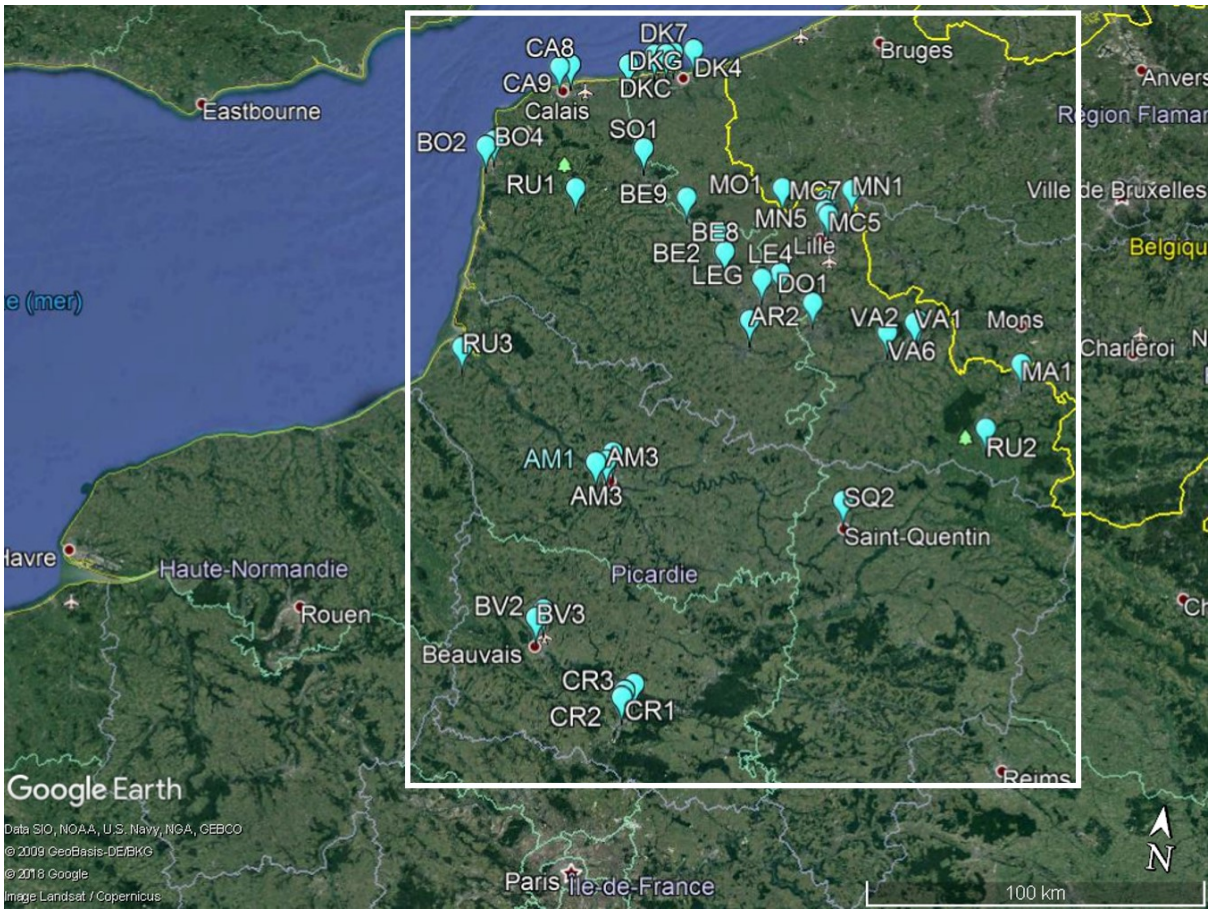

FIGURE 3.1 – Emplacements des 37 stations de mesures des concentrations PM_{10} en région Hauts-de-France.

Le tableau 3.1 indique la variance des stations sélectionnées par rapport à la moyenne régionale des concentrations de PM_{10}.

Station	RU1	RU2	RU3	CA9	CA8	BO4	MC7	VA1	MN1	DK4	DKC	DKG
$(\mu g/m^3)$	1.97	1.54	1.57	1.20	1.33	1.4	3.52	2.20	2.15	3.30	8.76	5.96

TABLE 3.1 – Variance des stations par rapport à la moyenne régionale des concentrations de PM_{10}

La variance la plus élevée correspond aux stations DKG et DKC, qui sont exposées aux activités industrielles au niveau de la ville de Dunkerque et influencées par les phénomènes météorologiques côtiers, suivies de la station MC7 située en milieu urbain. Juste après, vient la station DK4 qui est un peu plus éloignée de la zone industrielle par rapport à DKC et DKG, mais qui pourrait également être exposée à cette source d'émission. Ensuite, viennent les stations VA1 et MN1 situées sur les communes de Valenciennes et Roubaix (qui connaissent

d'importantes activités humaines et une forte population), et enfin nous avons les stations rurales et côtières (loin de la zone industrielle) qui ont la plus petite variance parmi les douzes stations en question. Les valeurs du tableau 3.1 se confirment par les figures de 3.2 à 3.5.

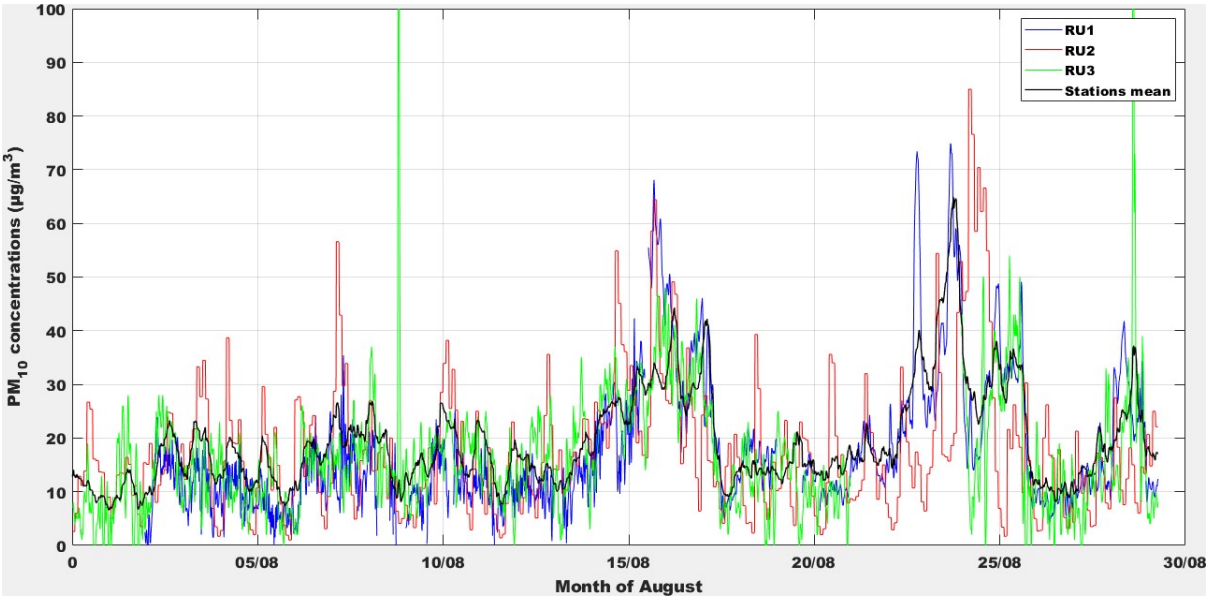

FIGURE 3.2 – Concentrations de PM_{10} des stations rurales et de moyenne régionale pour le mois d'août de 2016.

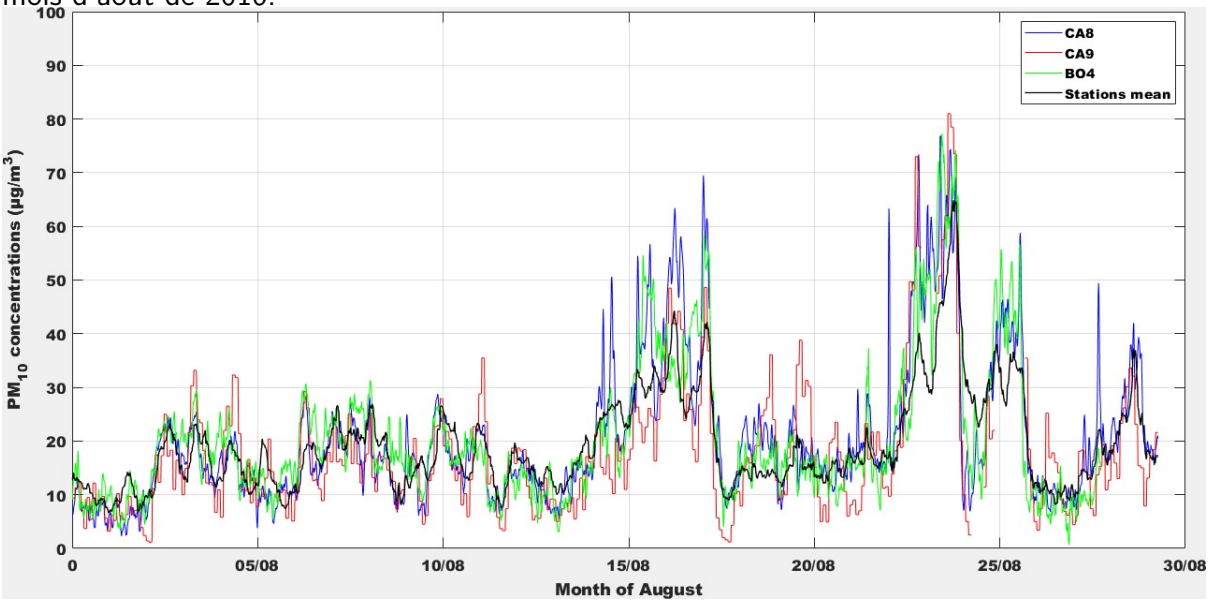

FIGURE 3.3 – Concentrations de PM_{10} des stations côtières et de moyenne régionale pour le mois d'août de 2016.

A titre d'exemple, la figure 3.2 montre que les stations rurales RU1 et RU3 suivent assez bien la moyenne régionale avec quelques pics, tandis que RU2 connait des oscillations plus fortes autour de cette moyenne, probablement dues à des phénomènes météorologiques locaux

caractérisés par les vents légers et une inversion de température vu qu'il s'agit d'une région vallonnée. Les données des stations côtières sont présentées dans la figure 3.3, où nous observons que les trois stations ont un comportement qui suit plutôt bien la moyenne régionale plus que les stations rurales. S'agissant des stations côtières (zone terre-mer), où des phénomènes météorologiques (comme la brise de mer) peuvent avoir lieu et peuvent provoquer une forte variation des concentrations de PM_{10} autour de la moyenne régionale, tel est le cas en période du 15 au 18 août sur la figure 3.4, où les concentrations de CA8 et BO2 deviennent beaucoup plus élevées que la moyenne régionale.

Les oscillations des stations urbaines se révèlent être plus fortes que celles des stations rurales et côtières autour de la moyenne régionale des concentrations de PM_{10} (figure 3.4), confirmant les résultats du tableau 3.1. Les stations MC7 et VA1 ont un comportement similaire qui suit la moyenne régionale pendant la majorité de la période du mois d'août, tandis que MN1 a une plus forte variation autour de cette moyenne pendant tout le mois. Cela pourrait être dû à une exposition plus importante de MN1 à une source de pollution urbaine (trafic, combustion, chauffage, etc.), que les deux autres stations de MC7 et VA1. Quant aux stations à proximité industrielle (figure 3.5), on constate que les concentrations de PM_{10} des stations DK et DKC présentent les plus fortes oscillations autour de la moyenne régionale parmi toutes les stations étudiées, alors que DK4 tend davantage à un comportement plus fort que les stations côtières et plus faible que celle à proximité industrielle. Les raisons de ces fluctuations sont les phénomènes météorologiques (puisqu'il s'agit également de stations côtières), ainsi que les activités industrielles dans cette zone (partie côtière de la ville de Dunkerque).

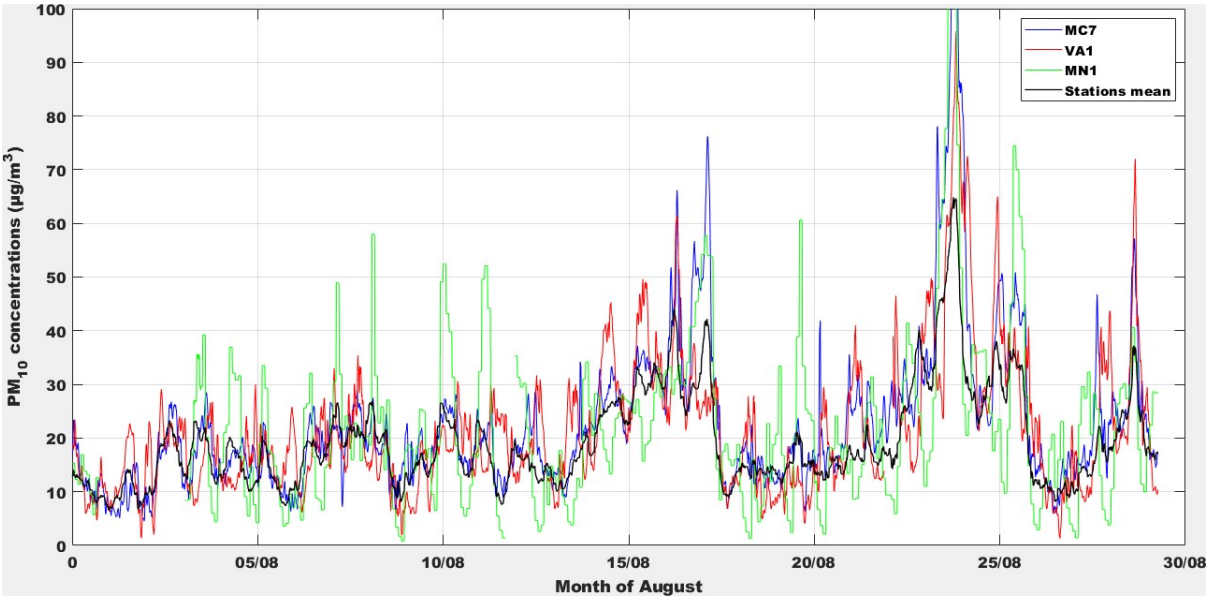

FIGURE 3.4 – Concentrations de PM_{10} des stations urbaines et de moyenne régionale pour le mois d'août de 2016.

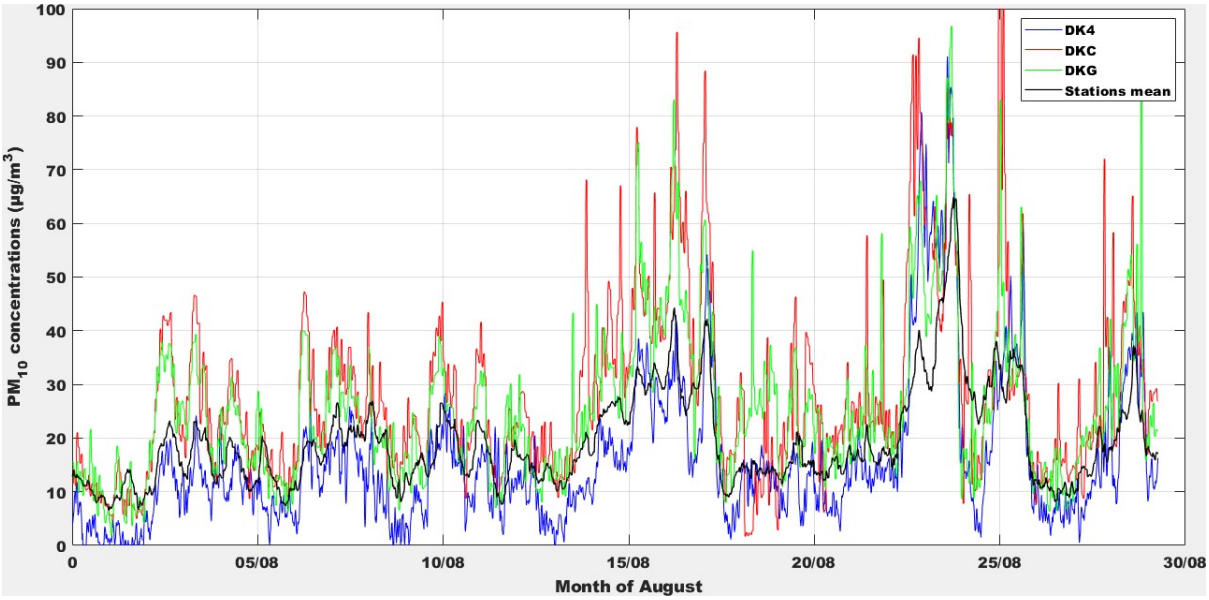

FIGURE 3.5 – Concentrations de PM_{10} des stations à proximité de la zone industrielle de Dunkerque et de moyenne régionale pour le mois d'août de 2016.

Par conséquent, chaque station ayant un comportement qui diffère de la moyenne régionale des concentrations de PM_{10} dans la région des Hauts-De-France, est exposée aux émissions anthropiques locales et aux phénomènes météorologiques spécifiques. L'influence de cette exposition sur la performance de l'interpolation spatiale sera notamment discutée dans la section 3.5.

3.4 Contributions

La pollution de l'air au niveau de la région de Hauts-de-France est une pollution issue de différentes sources d'émission (urbaine, industrielle, résidentielle, etc.), et exposée à différents phénomènes météorologiques (brouillard, brise de mer, etc.). Par conséquent, le caractère complexe de cette pollution rend son estimation assez difficile pour les méthodes d'interpolation classiques. Pour cela, nous proposons d'employer une interpolation adaptative aux différentes situations rencontrées de cette pollution de l'air par la méthode de pondération inverse à la distance avec une puissance optimisée de cette dernière.

Nous étudions l'impact des phénomènes météorologiques sur la performance des techniques d'interpolation dans l'estimation des concentrations des polluants PM_{10} dans la région, par le recours à un moyennage de ces derniers sur différentes périodes de temps.

Nous examinons également la pertinence et l'influence du bruit et de la densité des stations de mesure sur la précision de l'interpolation.

Le présent travail est la première étude menée pour l'estimation des concentrations de PM_{10} de la région des Hauts-de-France qui couple l'interpolation spatiale à la variabilité temporelle.

3.5 Résultats et discussion

La méthode que nous choisissons pour examiner l'effet des données moyennées temporellement sur les performances des méthodes d'interpolation (section 3.2.2) ainsi que pour générer toutes les cartes de ce chapitre est l'IDW, car elle a un faible RMSE et un R^2 élevé (tableau 3.2). En outre, l'IDW est rapide et facile à calculer, et elle est physiquement plus précise que les techniques de triangulation (voir annexe).

S'agissant de toutes les cartes de ce chapitre, la ligne noire en gras reliant la France et la Belgique représente la côte qui s'étend de la France à la Belgique, tandis que la ligne rouge en gras représente les frontières entre les deux pays.

La figure 6 présente la moyenne annuelle des concentrations de PM_{10} en 2016 dans la région des Hauts-de-France, illustrant le niveau de pollution de l'air de cette dernière. Les stations isolées sur la figure sont représentées par une zone concentrique de même niveau de PM_{10}, ou

ce qu'on appelle "the bull eye effect". Il s'agit d'une limitation de l'algorithme IDW, lorsque les points de données d'entrée disponibles sont d'une faible densité avec une distribution spatiale ayant des sites isolés. En effet, cela conduit à une diminution du R^2 dans certaines méthodes et à un problème de surapprentissage dans d'autres, comme confirmé par les résultats exposés plus loin dans cette section. La figure 6 vient confirmer l'hypothèse faite en section 3.1, discutant l'influence de proximité des stations aux sources d'émission. Les concentrations de PM_{10} les plus élevées sont remarquées au niveau des stations à : Dunkerque (DKC, DKI et DKG) en raison de l'activité industrielle dans cette partie de la ville ; Boulogne là où il y a un port de pêche ; Lille et la région de Roubaix étant des zone urbaines à forte densité de population avec un trafic routier intense, de même que pour Valenciennes tandis que pour Beauvais l'aéroport international pourrait être la principale source de la pollution de l'air.

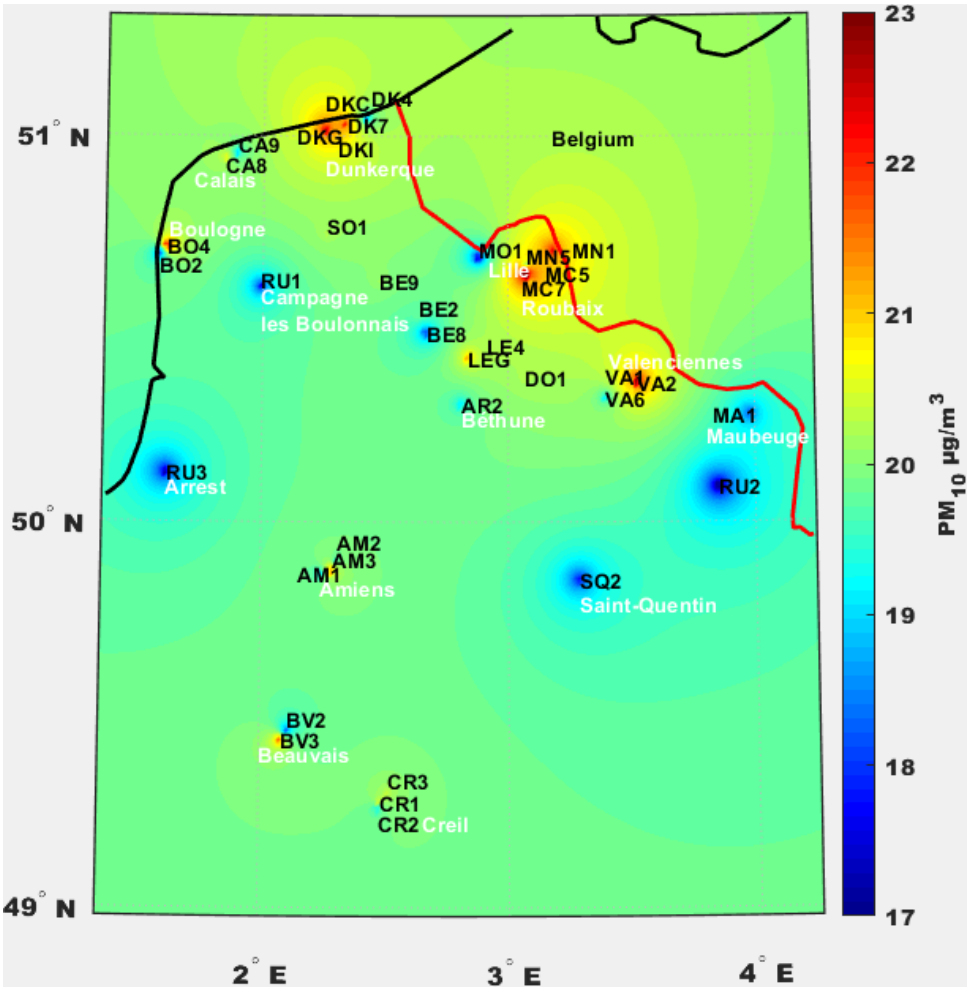

FIGURE 3.6 – Moyenne annuelle de l'année 2016 des concentrations de PM_{10} pour la région Hauts-de-France.

La figure 3.7 représentant l'écart type annuel de la distribution spatiale des concentrations de PM_{10} de la région des Hauts-de-France, permet d'observer les niveaux de variation et le comportement de ces dernières pour une meilleure interprétation des résultats d'interpolation spatiale. Cette carte montre de fortes variations de PM_{10} dans la partie industrielle de la ville de Dunkerque (DKG, DKC et DK7), dans les stations urbaines des villes de Lille et de Roubaix (stations MN et MC5). Une forte variation est également constatée au niveau de Creil (stations CR), ce que nous supposons être dû à la pollution de l'air venant de Paris compte tenu de la proximité de la ville de Creil à l'agglomération parisienne.

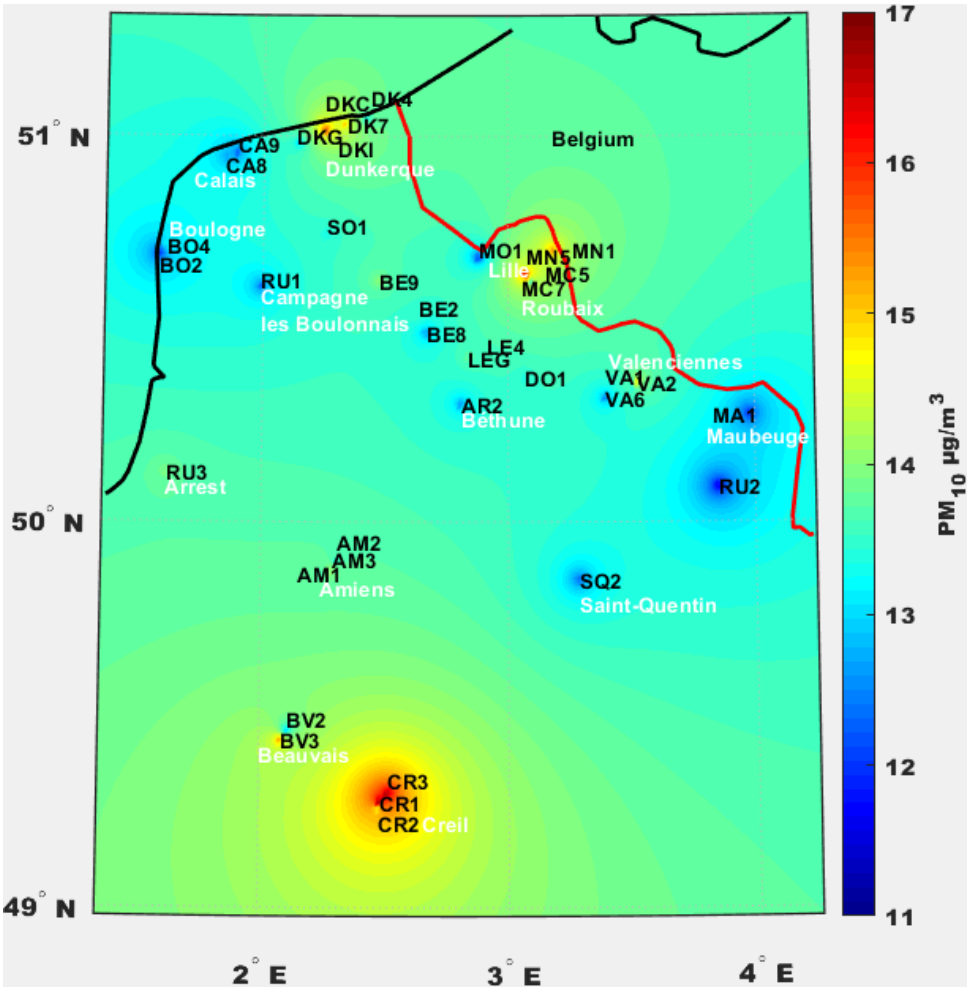

FIGURE 3.7 – Ecart type annuel de l'année 2016 des concentrations de PM_{10} pour la région Hauts-de-France.

3.5.1 Résultats de l'interpolation spatiale

S'agissant des méthodes d'interpolation spatiale, nous voulions voir l'efficacité de différents types de techniques d'interpolation dans l'estimation de la pollution de l'air, c'est pourquoi nous avons sélectionné un ensemble de méthodes de complexité d'algorithme différente (méthodes figurant dans le tableau 3.2). Il convient de rappeler que les résultats de l'application des techniques d'interpolation dans le tableau 3.2 sont obtenus après le prétraitement indiqué en section 3.1.

Méthode d'interpolation	RMSE ($\mu g/m^3$)	Intervalle de confiance de RMSE ($\mu g/m^3$)	R^2
Voisin plus proche	9.55	9.41-9.71	0.52
Interpolation linéaire[0]	8.84	8.70-8.99	0.58
Voisin naturel[0]	8.73	8.59-8.88	0.59
Spline[0]	9.61	9.44-9.77	0.51
IDW	7.74	7.63-7.84	0.68
IDW version optimisée	7.45	7.33-7.53	0.70
GPR (avec le 1^{er} noeud)	7.75	7.46-8.02	0.68
GPR (avec le 2^{me} noeud)	7.81	7.5-8.08	0.67

TABLE 3.2 – Résultats des méthodes d'interpolation

Les résultats du tableau 3.2 montrent que la technique d'IDW en version optimisée donne le RMSE le plus petit avec 7,45 μg/m^3 et le R² le plus élevé avec 70%, tandis que l'intervalle de confiance à 95% démontre que les méthodes IDW, IDW en version optimisée, GPR avec les 2 noeuds atteignent un RMSE comparable variant entre 7,33 à 8,08 μg/m^3. Toutes les méthodes basées sur la triangulation (plus proche voisin, interpolation linéaire, voisin naturel et spline) ont produit un intervalle de confiance de RMSE à 95% de 8,59 à 9,77 μg/m^3. La différence de précision entre toutes ces techniques reste faible. Compte tenu de la faible densité des sites de mesures disponibles de plus de la distribution spatiale de ces derniers sur l'ensemble de la zone d'étude, la distinction d'un meilleur interpolateur pour les concentrations de PM_{10} de la région des Hauts-de-France devient difficile. Cela prouve le rôle important que la densité des données et leur répartition géographique jouent dans la précision des méthodes de l'interpolation.

0. En plus de la méthode d'interpolation utilisée, nous avons utilisé le voisin le plus proche pour l'extrapolation car c'est la technique la plus stable parmi celles appliquées pour cette fin.

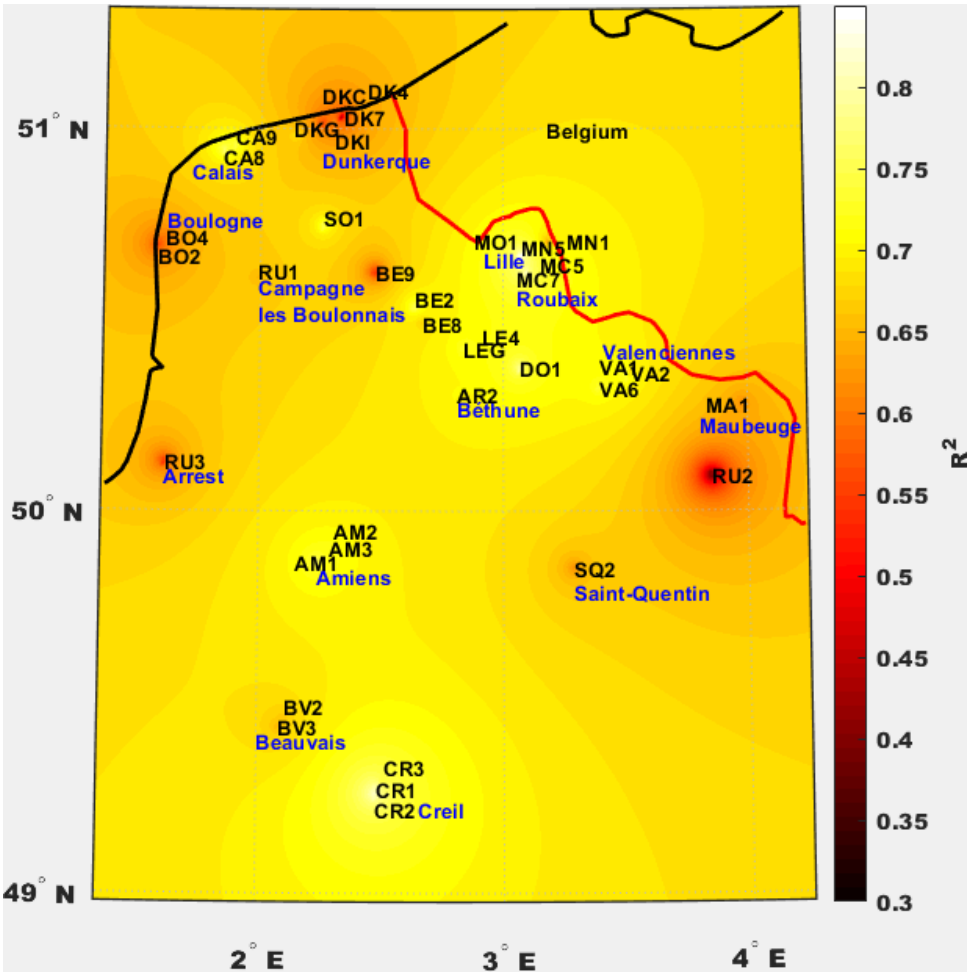

FIGURE 3.8 – R^2 de l'interpolation d'IDW des concentrations de PM_{10}.

Pour analyser le comportement de l'interpolation spatiale appliqué dans les différentes zones de la région étudiée, nous estimons le R^2 de chaque station par IDW, puis interpolons ces valeurs par cette même technique (figure 3.8). On remarque que R^2 varie entre 30% et plus de 80% et que les zones à proximité industrielle et exposées aux phénomènes météorologiques locaux, comme dans la ville de Dunkerque, ont des valeurs et une variance de pollution de l'air élevées, de plus du petit nombre de stations de mesure déployées au niveau de ces mêmes zones, tous ces facteurs ont engendrés des valeurs élevées de RMSE. A titre d'exemple, les zones à petites valeurs de R^2 à Dunkerque comme au niveau des stations DK4 et DKC qui sont situés à proximité des sources d'émission industrielles de la ville, et au niveau de la zone de RU2 où nous ne disposons que d'une seule station, de plus, l'emplacement géographique de cette station favorise la circulation de la pollution de l'air vu qu'il s'agit d'une zone montagneuse.

Nous n'avons pas assez d'information sur la totalité de la région étudiée pour interpréter le résultat de l'interpolation spatiale obtenu dans la figure 3.8.

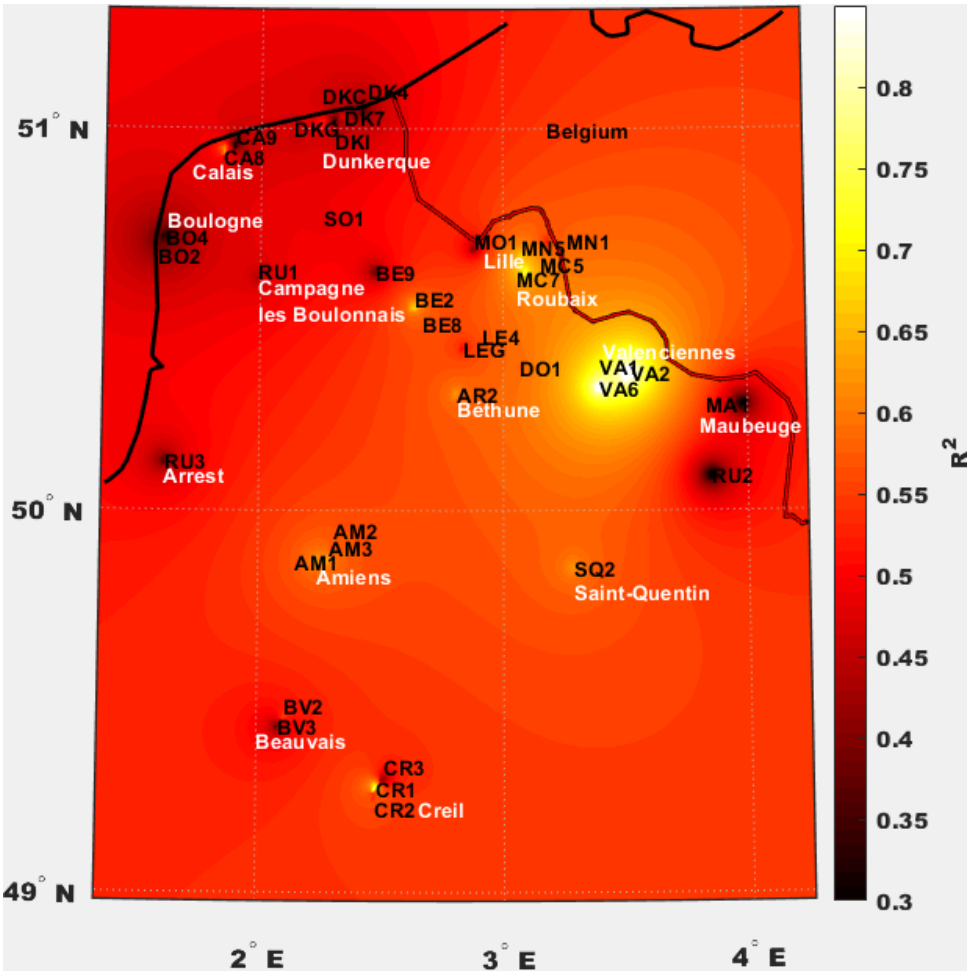

FIGURE 3.9 – R^2 de l'interpolation d'IDW des concentrations de PM_{10}.

L'effet des phénomènes météorologiques ainsi que l'exposition aux sources d'émission sur les concentrations de PM_{10} est observé notamment dans l'interpolation spatiale effectué en figure 3.9, où nous avons calculé le R^2 des concentrations de PM_{10} par l'interpolation linéaire pour chaque station, puis nous avons interpolé ces valeurs de R^2 par IDW. On remarque qu'il y a une dégradation importante de R^2 dans presque toute la région étudiée par rapport à la carte de la figure 3.8, en ayant toujours les valeurs les plus petites de R^2 aux zones qui sont à proximité des sources d'émission et qui sont sous influence des phénomènes météorologiques locaux comme le vent ou la turbulence (comme mentionné dans la section 2.1), ce qui produit une variation élevée de la variance de la pollution atmosphérique et conduit par conséquent à

une erreur plus élevée dans les performances d'interpolation, en s'agissant toujours des mêmes zones discutées en figure 3.8. La diminution du R^2 de l'interpolation linéaire en comparaison avec l'IDW est expliquée par la simplicité et la limitation du principe de la triangulation aux trois voisins les plus proches sur lequel l'interpolation linéaire se base.

3.5.2 Résultats du moyennage temporel

Le littoral de la région des Hauts-de-France est exposé à l'influence des phénomènes météorologiques locaux conduisant à une forte variation des concentrations de PM_{10}. Afin d'étudier la correspondance entre la précision de l'interpolation et la météorologie locale de la région, nous avons suggéré un lissage des fluctuations de PM_{10}. Après avoir proposé une amélioration de l'interpolation spatiale des concentrations de PM_{10} dans la région des Hauts-de-France en optimisant les paramètres des techniques d'interpolation sur des données quart horaires fournies par ATMO, nous avons créé de nouveaux jeux de données à travers un moyennage temporel à différentes périodes allant jusqu'à 3 mois . Ensuite, nous avons appliqué l'IDW pour estimation de RMSE et R^2.

Les résultats présentés au tableau 3.3 indiquent que le R^2 atteint sa valeur maximale pour le moyennage temporel correspondant à la période d'un jour, cela est confirmé par les intervalles de confiance de 95% qui révèlent le grand R^2 pour cette même période. Ce résultat démontre que la précision de l'interpolation spatiale est sensible aux effets météorologiques locaux et peut être dégradée par l'influence de ces derniers, tel que la brise de mer. Ainsi, les fluctuations non corrélées observées sur la figure 3.3 des stations côtières autour de la moyenne régionale pourraient entraîner une dégradation de la précision de l'interpolation spatiale.

Période d'échelle temporelle des données moyennées	R^2	Intervalle de confiance
15 minutes	0.68	0.66-0.69
1 heure	0.69	0.68-0.7
3 heures	0.72	0.71-0.73
6 heures	0.74	0.73-0.75
1 jour	0.80	0.79-0.81
1 semaine	0.77	0.76-0.78
2 semaines	0.72	0.71-0.73
1 mois	0.62	0.61-0.63
3 mois	0.45	0.44-0.46

TABLE 3.3 – Résultats d'interpolation d'IDW interpolation pour différentes échelles temporelles

Pour analyser l'influence de la variation des périodes de moyennage temporel sur les différentes stations du réseau de surveillance en question, une illustration du R^2 de l'IDW est présentée dans la figure 3.10 pour les données moyennées sur une période d'un jour.
L'amélioration de R^2 est remarqué au niveau de toute la région d'étude par rapport à la figure 3.8 (données quart horaires), cependant nous repérons toujours les valeurs les plus petites de R^2 dans les stations qui sont exposées à certaines sources d'émission de pollution atmosphérique comme les stations DK et RU2 (la même échelle de R^2 est conservé pour pouvoir comparer ce résultat avec la carte de la figure 3.8). Par conséquent, le filtrage des effets des phénomènes météorologiques locaux par le lissage proposé a abouti à des valeurs de R plus grande dans l'estimation des PM_{10}, mais avec une distribution spatiale qui varie sur la région étudiée. Cependant, la petite valeur observée de R^2 dans la zone RU2, pourrait être due à l'exposition à une source d'émission de pollution atmosphérique (comme discuté en section 3.5), de plus de la faible densité de sites de mesure dans cette zone.
La cartographie des autres périodes de moyennage temporel indiquées sur le tableau 3.3 est présentée en annexe (voir figures de 2 à 6 en annexe).

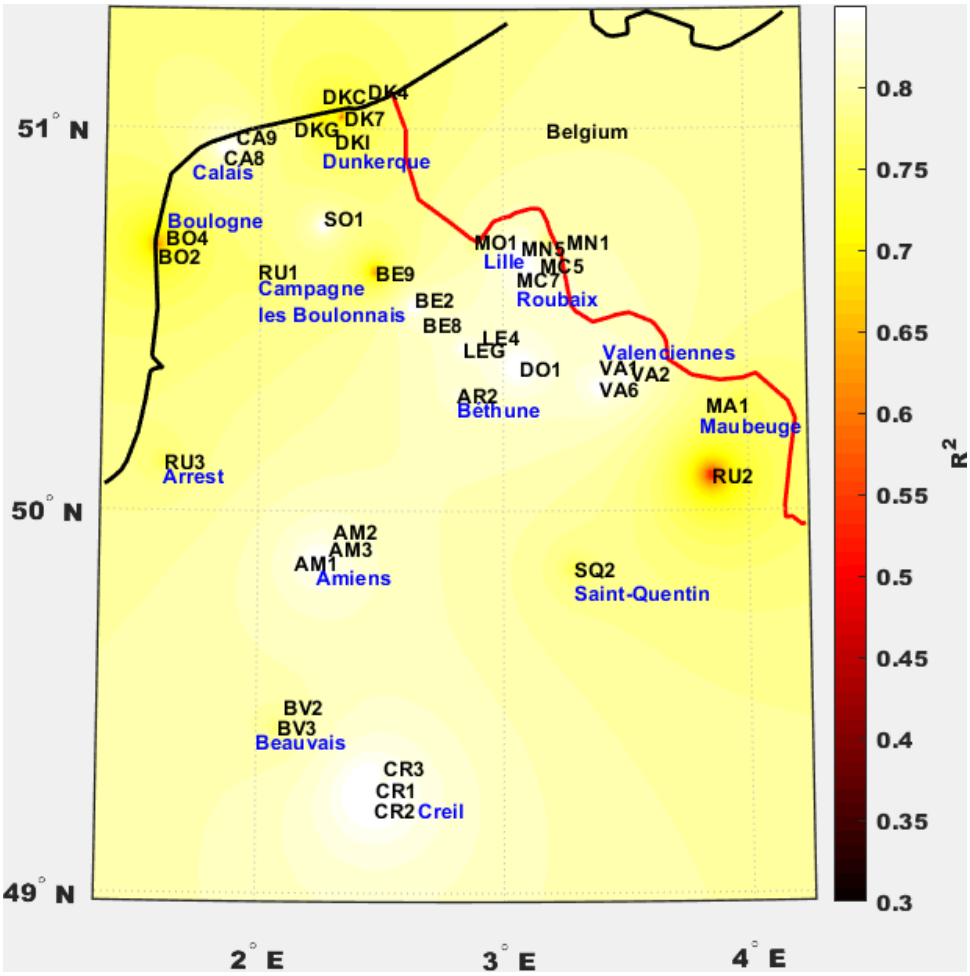

FIGURE 3.10 – R^2 d'IDW pour les données des concentrations de PM_{10} moyennées d'une période d'un jour .

3.5.3 Sensibilité des méthodes aux perturbations et à la densité des données d'entrée

Dans cette partie, nous examinons la sensibilité des méthodes d'interpolation spatiale aux perturbations et analysons le rôle que joue la densité des données dans la précision de performance de cette interpolation.

Résultats de la sensibilité des méthodes aux perturbations

Le but de perturber les mesures de concentrations de PM_{10} dont nous disposons, est d'étudier l'impact des perturbations réelles qui ont lieu par exemple lors d'un dysfonctionnement dans les stations de mesure sur l'interpolation spatiale. Les perturbations sont introduites dans

les concentrations quart horaires fournies par ATMO. Pour vérifier la sensibilité des techniques d'interpolation appliquées, nous perturbons les mesures de notre base de données d'entrée en ajoutant un bruit gaussien non corrélé d'un écart type de $\sigma = 5\ \mu g/m^3$. Cette valeur a été choisie comme approximation à la variance des mesures réelles. Ensuite, nous appliquons les mêmes techniques de l'interpolation utilisées avant et estimons la valeur de RMSE de chacune pour la nouvelle base de données bruitées créée (tableau 3.4).

Nous définissons les données perturbées par l'équation (3.17), en sachant que la base de données observée est la base de données fournie par ATMO des concentrations de PM_{10} :

$$\text{Base de données d'entrée perturbée} = \text{ base de données observée} + \text{ bruit gaussien}(\sigma = 5\mu g/m^3) \tag{3.17}$$

Méthode d'interpolation	RMSE (données observées)	RMSE (données perturbées)	$\delta RMSE$(différence)
Voisin plus proche	9.55	11.70	2.15
Interpolation linéaire	8.84	10.81	1.97
Voisin naturel	8.73	10.67	1.94
Spline	9.61	11.82	2.21
IDW	7.74	9.26	1.52
IDW version optimisée	7.48	8.93	1.45
GPR (avec le 1^{er} noeud)	7.75	9.45	1.70
GPR (avec le 2^{me} noeud)	7.81	9.52	1.71

TABLE 3.4 – Résultats des méthodes d'interpolation sur des données perturbées par le bruit gaussien

Le tableau 3.4 présente le RMSE des données de mesure (du tableau 3.2) et des données perturbées (en troisième colonne du tableau 3.4) ainsi que la différence entre ces deux derniers (en quatrième colonne). Le tableau 3.4 montre que δRMSE reste comparable entre toutes les méthodes, où la différence d'erreur la plus faible concerne l'IDW en version optimisée est de 1,45 $\mu g/m^3$, tandis que la différence la plus élevée est celle de Spline de 2,21 $\mu g/m^3$. On remarque que δRMSE d'IDW et GPR sont plus petits que les δRMSE des méthodes basées sur la triangulation, et ceci dû au fait que ces dernières se basent sur les trois points voisins pour interpoler, par conséquent, une fois que l'un de ces trois points est perturbé, cela affecte la précision l'interpolation beaucoup plus par rapport à l'interpolation par IDW ou GPR, qui utilisent tous les sites de mesure disponible pour effectuer l'estimation. Etant donné que

Spline est une technique qui essaye d'ajuster l'interpolation pour que son estimation soit la plus proche possible des points de données d'entrée, ce qui pose un problème de surapprentissage menant Spline à avoir la plus grande δRMSE parmi toutes les méthodes.

Le tableau 3.4 révèle une petite différence de δRMSE entre les différentes méthodes appliquées même si certaines sont plus précises que d'autres, nous supposons que c'est en raison du jeu de données d'entrée qui n'est pas assez dense pour que les méthodes qui sont plus avancés (IDW et GPR) puissent l'exploiter plus et améliorer leur précision d'interpolation en comparaison avec des techniques plus simple (méthodes basées triangulation).

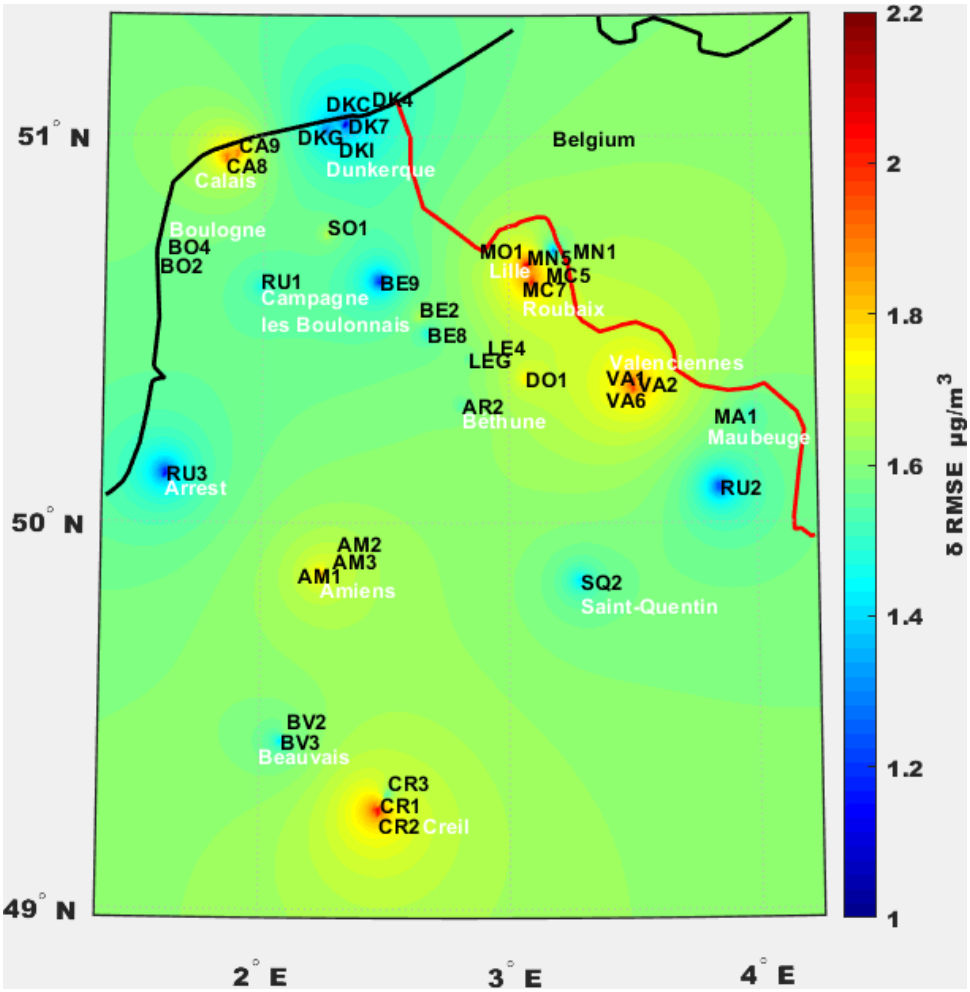

FIGURE 3.11 – différence entre RMSE des données observées et RMSE des données perturbées par l'interpolation d'IDW des concentrations de PM_{10}.

Dans le but d'examiner l'effet de ces perturbations sur les différentes stations du réseau de surveillance, la figure 3.11 montre le δRMSE calculé à partir des données perturbées pour chaque station par la méthode d'IDW puis interpolé par cette même technique.Les valeurs

de δRMSE sont faibles au niveau des stations qui sont exposées à des sources d'émission de pollution atmosphérique (stations discutées dans la section 3.5), contrairement aux stations qui ont un R^2 plus élevé (figure 3.8), dont le δRMSE est plus grand. Cela pourrait être expliqué par la proximité des stations de mesure aux sources d'émission locales (les stations industrielles de la ville de Dunkerque) qui rend leur comportements décorrélées, et ne pas être affecté par la suite après avoir ajouté du bruit à leurs mesures, ce qui est indiqué par la faible différence δRMSE au niveau de ces stations (comme les stations DK, RU2 et MC7).En revanche, les autres stations qui sont loin de toute activité produisant de la pollution de l'air, et qui ont un comportement corrélé des concentrations de PM_{10}, quand une station est exposée à des perturbations cela affecte le comportement de ses stations voisines et produit une augmentation dans l'erreur d'interpolation comme présenté dans la figure 3.11.

Par ailleurs, nous avons appliqué de différentes valeurs de perturbations sur certaines des techniques d'interpolation utilisées (figure 3.12), pour examiner la précision de ces dernières par rapport à des perturbations plus grandes. La croissance des valeurs de perturbations appliquées (de 0 à 3 σ) s'accompagne par une augmentation des valeurs de RMSE, ce qui était attendu. De plus, nous avons remarqué qu'IDW et GPR continuent à se comporter de manière similaire en ayant des valeurs de RMSE comparables, tandis que la méthode d'interpolation linéaire a un RMSE plus grand qu'IDW et GPR.

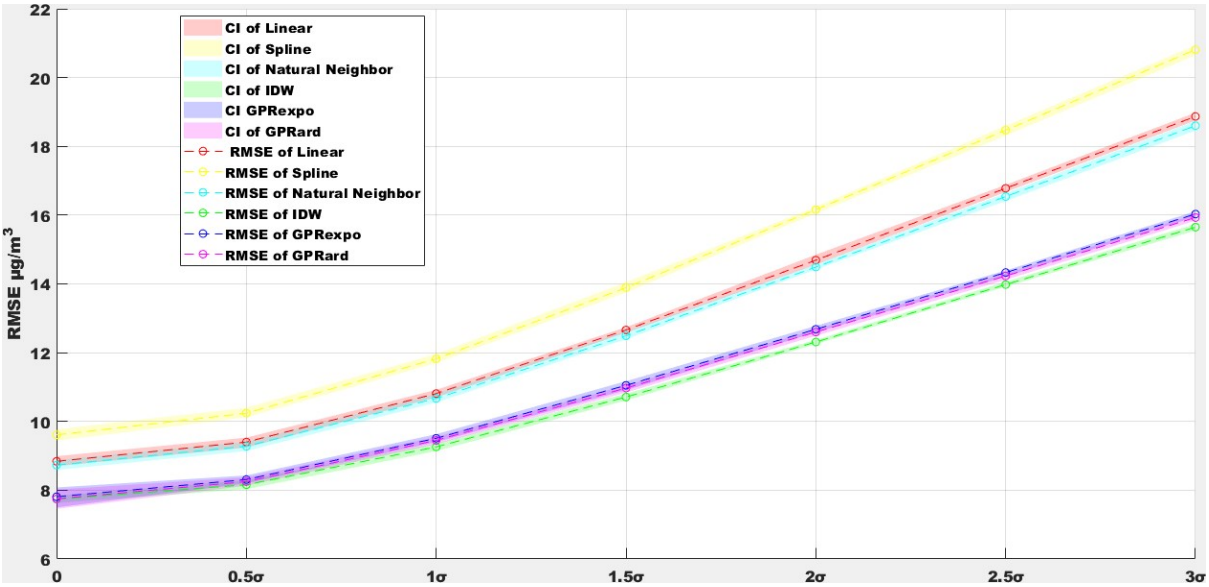

FIGURE 3.12 – Sensibilité des méthodes aux perturbations (bruit gaussien).

Résultats de la sensibilité des méthodes à la densité des données d'entrée

Le dernier point évalué dans ce chapitre est l'impact de la densité des données sur la précision des techniques d'interpolation spatiale. Pour ce faire, nous avons appliqué les techniques d'interpolation précédentes sur les mesures de concentrations de PM_{10}, en éliminant à chaque fois et au hasard une des stations de mesure en entrée, et en estimant le RMSE correspondant au nouveau nombre des stations qui restent. La figure 13 illustre le RMSE de chaque méthode avec un intervalle de confiance de 95% en fonction du nombre de stations utilisées en interpolation. Nous avons choisi six des techniques utilisées auparavant, à savoir l'interpolation linéaire, la Spline, le voisin naturel, l'IDW et le GPR avec deux noyaux.
Comme prévu, la diminution du nombre de stations altère la précison des méthodes d'interpolation (figure 3.13), indiqué par le RMSE qui continue à s'accroître avec chaque station enlevée, au niveau des différentes méthodes d'interpolation utilisées, mettant en évidence le rôle crucial que la densité des données joue dans l'amélioration de la précision des techniques de l'interpolation (comme discuté dans le chapitre 2).

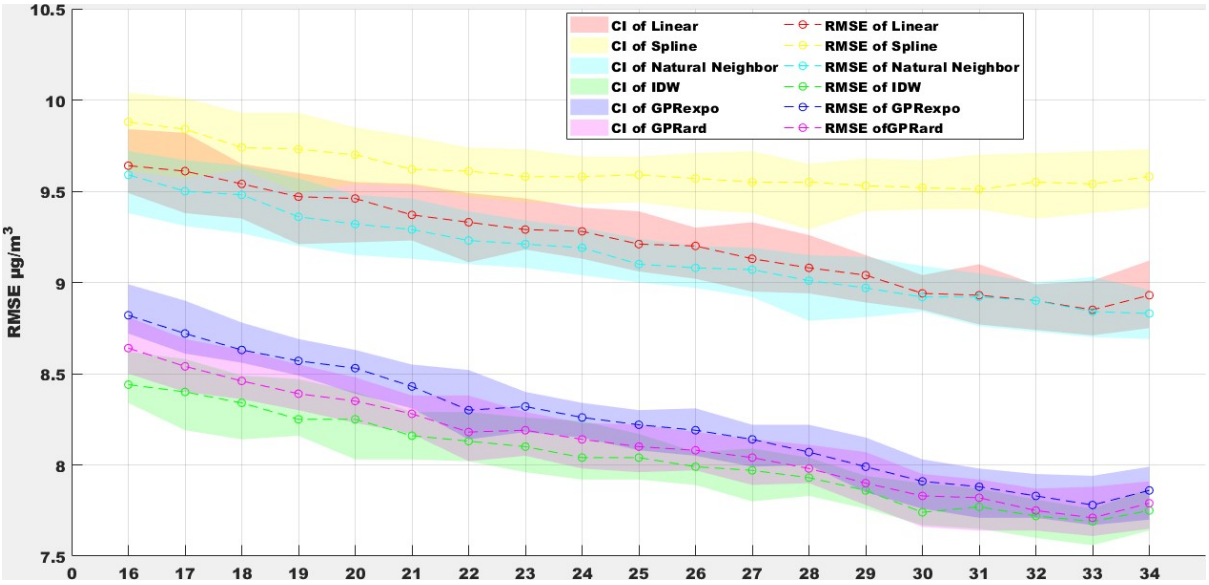

FIGURE 3.13 – Sensibilité des méthodes de l'interpolation à la densité des données (nombre de stations).

3.6 Conclusions

L'étude menée dans ce chapitre a montré que la variabilité spatiale et temporelle de la pollution de l'air de la région des Hauts-de-France, joue un rôle important dans l'estimation de cette dernière par des techniques d'interpolation. Cette région nous a permis d'observer l'influence de diversité des sources d'émission ainsi que l'exposition aux phénomènes météorologiques sur la performance des techniques d'interpolation. Les méthodes d'interpolation de pondération inverse à la distance avec ses deux versions et la régression par les processus Gaussiens ont donné des résultats similaires dans l'estimation des concentrations de PM_{10}. Pour toutes les techniques d'interpolation appliquées, l'erreur d'estimation était plus grande dans les zones proches des sources d'émission et dans les zones soumises à des phénomènes atmosphériques intensifs (zones côtières et vallonnées).

L'impact des phénomènes atmosphériques locaux sur la pollution de l'air a été filtré par un moyennage de temps appliqué sur les mesures d'entrée, produisant un maximum de coefficient de détermination (R^2) pour la période d'un jour. Cela a montré l'influence des phénomènes météorologiques locaux de périodicité d'un jour sur la pollution de l'air de la région de Hauts-de-France comme la brise de mer. Et enfin la forte sensibilité des méthodes au bruit et à la densité des mesures d'entrée a été montrée.

Nous suggérons comme continuité à ce travail, l'inclusion de la composante météorologique ainsi que temporelle de la pollution de l'air étudiée qui pourraient apporter plus d'informations sur la dispersion locale de la pollution et sur la précision de l'interpolation. Et aussi, proposer une optimisation du réseau de la surveillance de la qualité de l'air de la région, spécifique à la pollution de l'air de cette dernière.

Chapitre 4

Evaluation et optimisation du réseau de surveillance de la qualité de l'air de Dunkerque

Dans ce chapitre, on se donne pour objectif de fournir une technique d'estimation des positions optimales des stations de mesure, minimisant l'erreur d'interpolation. Cette technique est appliquée sur deux jeux de données PM_{10} générés par deux modèles de qualité de l'air : Atmospheric Dispersion Modeling System (ADMS) [7], et le modèle de panache gaussien [8]. Le résultat de ces modélisations sert à fournir les valeurs de vérité de terrain de la pollution de l'air. Les valeurs de cette pollution sont estimées par l'interpolation de pondération de distance inverse dans les deux situations, qu'on évalue ensuite par l'erreur (RMSE) qui représente la différence entre les valeurs de vérité de terrain et les valeurs interpolées. L'estimation de l'erreur d'interpolation est présentée en fonction du nombre de stations optimisées permettant d'évaluer le nombre de stations de surveillance nécessaires.

Ensuite, nous étudions et analysons la configuration des stations obtenue par l'algorithme génétique et l'optimisation des essaims de particules, avec celle du réseau de surveillance de la qualité de l'air déployée par ATMO Hauts-de-France.

Et en dernier lieu, nous proposons une approche permettant d'augmenter la précision d'interpolation dans une région d'intérêt spécifique, à titre d'exemple, dans la zone résidentielle de

Dunkerque.

Ce chapitre est divisé en sept sections : la première est une présentation des notions de l'optimisation, suivie d'une deuxième dédiée à un état d'art du travail effectué. La section 4.3 aborde les outils de modélisation de la dispersion atmosphérique : ADMS et le modèle de panache gaussien, en section 4.4 nous précisons le prétraitement effectué sur les données d'entrée. La section 4.5 détaille les contributions apportées dans cette partie, et la section 4.6 est consacrée aux résultats du développement et analyse des contributions. Enfin, la section 4.7 présente les conclusions retenues de l'ensemble du chapitre.

4.1 Notions fondamentales : Optimisation

L'optimisation est le processus de minimisation ou de maximisation d'une fonction de coût, afin de trouver la meilleure solution à des problèmes dits « complexes/difficiles » à résoudre, quand il s'agit des réseaux de capteurs sans fils [130]. L'optimisation pour ce genre de problèmes, ne peut pas se baser sur des méthodes déterministes, qui sont brièvement définies par les méthodes exactes produisant toujours la même sortie pour une même entrée [131], avec un délai de temps d'exécution prédéfini. Les métaheuristiques sont des algorithmes très répandus pour la problématique d'optimisation, où méta signifie « haut niveau » et heuristique signifie « solution ». Ce terme désigne donc une procédure de haut niveau affectée aux problèmes sans solution déterministe [131]. Ces dernières peuvent être soit mono-objectif, soit multi-objectifs, l'objectif ici représente la variable à optimiser.

L'algorithme métaheuristique vise à trouver une solution dans un espace de recherche. Certains algorithmes se focalisent sur la recherche locale tandis que d'autres se focalisent sur la recherche globale. La fonction de coût est formulée en considérant les différentes métriques de l'objectif visé. Il existe deux types de métaheuristiques : métaheuristique à solution unique et métaheuristique à base de population [131]. Les algorithmes à solution unique, comme le recuit simulé et la recherche tabou, visent à améliorer la solution obtenue par une recherche locale ; leur tâche se limite à l'exploitation, tandis que les métaheuristiques à base de population sont dédiés à l'exploration, autrement dit à la recherche globale ayant pour but d'aboutir

à une nouvelle solution plus fiable. Il y a deux familles de techniques à base de population : les algorithmes évolutifs et l'intelligence en essaim. Dans les algorithmes métaheuristiques évolutifs, chaque solution à un problème d'optimisation est définie comme étant un vecteur de variables d'optimisation comme suit :

$$X = (x_1, x_2, ..., x_i, ..., x_N) \tag{4.1}$$

Où X est une solution du problème d'optimisation, x_i la ième variable d'optimisation du vecteur de solution X, et N est le nombre de variables d'optimisation. Le vecteur X est estimé en minimisant la différence entre les simulations et les mesures. Cette erreur de simulation est quantifiée par une fonction de coût $f(X)$ (qui représente généralement la norme euclidienne de la différence entre les simulations et les mesures). Par conséquent, le problème d'optimisation avce contraintes peut être formulé comme suit :

$$Etant\ donne\acute{}: X = (x_1, x_2, ..., x_i, ..., x_N),$$
$$Minimiser f(X),$$
$$En\ satisfaisant : bi_i \leq x_i \leq bs_i\ ,\ \ i = 1...N$$

Avec bi_i et bs_i sont la borne inférieure et borne supérieure qui délimitent la valeur de la variable x_i. Les algorithmes évolutifs incluent l'algorithme génétique et l'évolution différentielle. L'intelligence des essaims englobe : l'optimisation des essaims de particules, l'optimisation des colonies de fourmis, l'optimisation des colonies d'abeilles artificielles, l'optimisation de la recherche de nourriture bactérienne, la recherche de coucou et l'algorithme de luciole [131].

Méthodes et validation

Dans cette section, nous expliquons les méthodes et mesures utilisées dans notre travail pour réaliser et évaluer le processus d'optimisation appliqué.

Algorithme génétique

L'algorithme génétique (en anglais genetic algorithm : GA) se base sur le principe de l'évolution naturelle des espèces [132]. Les principales étapes pourraient être résumées par la reproduction, le croisement chromosomique, la mutation des gènes et la sélection. Première-

ment, une population initiale de n individus (chromosomes qui représentent dans notre cas les positions des stations) est générée de manière heuristique ou aléatoire. Cette population est censée évoluer au fil des générations, pour produire de nouveaux individus qui atteindraient une meilleure fitness. Cette fitness est déterminée par une fonction de fitness (expliquée ci-dessous), qui est développée selon le problème d'optimisation en question. L'étape de sélection implique d'évaluer chaque individu par cette fonction de fitness de manière à ne retenir que les individus les plus aptes (se rapprochant de l'objectif d'optimisation), pour être les parents de la génération suivante. Le processus de reproduction de nouveaux individus (enfants) est appliqué via deux techniques : le croisement qui est une combinaison d'un couple de parents, et la mutation consistant à introduire un changement aléatoire chez un parent.

Optimisation de l'essaim de particules

L'optimisation des essaims de particules (en anglais Particle Swarm Optimization : PSO) est un algorithme qui imite le processus de recherche de nourriture du comportement de vol des oiseaux dans la nature. Il utilise une «population» de solutions candidates pour déterminer une solution optimale pour le problème à résoudre, en explorant et en exploitant l'espace de recherche. Le degré d'optimalité est mesuré par une fonction d'évaluation/fitness définie par l'utilisateur. Les membres de la population sont appelés «particules», ils sont dispersés dans l'espace du problème, où chaque particule a :

- Une position, exprimée par des coordonnées dans l'espace de recherche.
- Une vitesse, qui permet à la particule de se déplacer, lors des itérations et de changer sa position. Il évolue en fonction de son meilleur voisin, de sa meilleure position et de sa position antérieure. Cette évolution permet de s'approcher de la position optimale des particules.
- Un voisinage, qui est un ensemble de particules qui interagissent directement avec la particule en question, en particulier celui qui a la meilleure performance.

Pour chaque instant t, chaque particule met à jour ses deux composantes :

- Sa meilleure position visitée (X_{Pbest})

- La position du meilleur voisin dans l'essaim (X_{Pbest}), qui correspond à la valeur optimale atteinte de la fonction de fitness.

À chaque itération, la vitesse $v_i(t+1)$ et la position $x_i(t+1)$ doivent être mises à jour par les équations suivantes :

$$v_i(t+1) = w * v_i(t) + c_1 * r_1 * (X_{Pbest,j} - x_j(t)) + c_2 * r_2 * (X_{Gbest} - x_j(t)) \quad (4.2)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (4.3)$$

Avec :

— w est le poids d'inertie,

— $v_i(t)$ est le vecteur de vélocité de la particule j à l'instant t,

— $x_i(t)$ est le vecteur de position de la particule j à l'instant t,

— $X_{Pbest,j}$ est le vecteur de la meilleure situation personnelle observée par la particule j jusqu'à l'itération courante,

— X_{Gbest} est le vecteur de la meilleure position globale trouvée par l'essaim jusqu'à l'itération courante,

— c_1 et c_2 sont les coefficients d'accélération. Ils représentent le degré de «confiance» dans la meilleure solution trouvée par la particule individuelle (c_1 : paramètre cognitif) et par l'ensemble de l'essaim (c_2 : paramètre social) [133],

— r_1 et r_2 sont des vecteurs contenant des nombres aléatoires avec une distribution uniforme dans l'intervalle de $[0,1]$.

Kennedy et Eberhart détaillent plus la procédure de PSO en [134].

Fonction de fitness

La fonction de fitness a pour rôle de mesurer le degré de précision d'un algorithme d'optimisation. D'après les résultats obtenus en chapitre 3 de l'interpolation spatiale, nous avons choisi l'IDW (décrite en section 3.2.2) comme fonction de fitness étant donné qu'elle représente le meilleur compromis de précision/temps de calcul parmi l'ensemble des techniques examinées en chapitre 3.

Validation

La performance des méthodes est évaluée par une mesure qui calcule l'erreur d'estimation. Dans les processus d'optimisation, la fonction de coût est utilisée pour trouver la solution la plus appropriée, qui donne la meilleure précision.

Fonction de coût

La fonction de coût que nous avons déterminé est le RMSE décrit en section 3.2.4. Nous avons aussi calculé le R^2 (section 3.2.4) correspondant à chaque méthode d'optimisation appliquée.

4.2 Etat de l'art

La rentabilité des capteurs par rapport à celles des stations de mesure de la qualité de l'air traditionnelles permet des déploiements importants et améliore ainsi la résolution spatiale des méthodes de surveillance actuelles [135]. Mais bien que les capteurs à faible coût (low-cost sensors) soient peu coûteux et puissent être déployés à grande échelle, le budget de déploiement est généralement limité, ce qui signifie que seul le nombre nécessaire de nœuds doit être déployé. En conséquence, une approche de déploiement doit être utilisée pour optimiser le nombre et les emplacements des nœuds de capteurs tout en assurant une surveillance fiable [37].

Cette optimisation dépendra de l'objectif visé pouvant être soit : la bonne couverture de l'air mesurée par des capteurs montés sur des bus [136] ; ou la prise en considération des caractéristiques de terrain d'une région complexe à surveiller pour l'estimation de sa qualité de l'air, comme pour une région résidentielle connaissant un trafic routier très important [137] ; ou l'aboutissement du chemin optimal pour déployer des capteurs mobiles, tout en réduisant le temps de calcul [138] etc.

Lancia et al., [136] ont conçu un modèle basé sur l'algorithme génétique pour évaluer et améliorer la qualité de l'air dans la ville de Rome. Leur projet a visé un déploiement de capteurs installés dans les bus de la ville. Leur contribution réside dans la détermination du nombre de capteurs permettant l'obtention d'une bonne couverture de l'air mesuré. Cependant, Van Nguyen et al. (2013) [138] ont développé un algorithme qui minimise l'erreur de la prédiction

de la qualité de l'air. Cet algorithme a permis de trouver le chemin optimal pour les capteurs mobiles à déployer, les auteurs ont pu déterminer le chemin le plus informatif avec un temps de calcul réduit. Cependant, que Lerner et al. (2019) [137], ont proposé un algorithme qui cherche les meilleurs emplacements pour des micro capteurs pour une région résidentielle mixte (avec un réseau routier important), en vue de mesurer la qualité de l'air. Ils ont obtenu des résultats meilleurs que les précédentes études, étant donné qu'ils ont tenu compte des caractéristiques de ces micro capteurs dans leur algorithme. Par conséquent, ce dernier représente une solution extrêmement rapide et flexible, et facilement implémentable dans différentes régions pour les micro capteurs.

Concernant notre travail, le but principal est de minimiser l'erreur d'interpolation pour les emplacements optimaux des stations/capteurs de mesure des concentrations PM_{10} à Dunkerque, pour deux types de bases de données. Notre étude permet également de voir la différence entre les configurations des stations obtenues après optimisation sur deux types de donnée : les données issues d'ADMS, et les données générées par le modèle de panache gaussien, vu que dans la littérature, les deux types de modélisation sont utilisés pour obtenir les valeurs de vérité de terrain. En outre, nous avons réussi à proposer une approche simple et facile à appliquer, afin d'avoir une estimation plus précise des concentrations PM_{10} pour une zone déterminée, comme la zone résidentielle de la ville.

4.3 Modélisation de la dispersion atmosphérique et données utilisées

Pour pouvoir évaluer notre estimation des concentrations PM_{10}, nous avons besoin des données représentant la vérité de terrain (comme approximation aux valeurs réelles) sur l'ensemble du site étudié, même aux endroits où nous ne disposons pas de stations de mesure. Pour ce faire, on fait appel à des systèmes de modélisation atmosphérique, dans notre cas nous avons utilisé ADMS, vu qu'il est le seul choix qui nous a été fourni par ATMO Hauts-de-France et qu'elle utilise pour sa modélisation de la qualité de l'air.

Par ailleurs, nous avons également utilisé le modèle de panache gaussien pour générer les

données de vérité de terrain, dans le but de comparer les topologies de stations de mesure obtenues par l'optimisation sur un jeux de données représentant une pollution de l'air provenant des sources diffuses (ADMS), et un autre d'une source canalisée (modèle de panache gaussien).

4.3.1 Modélisation par le modèle de panache gaussien

Le modèle de panache gaussien (en anglais Gaussian Plume Model, GPM) est l'un des modèles de dispersion de l'air les plus anciens, les plus simples et les plus couramment utilisés. Il suppose que la dispersion de l'air est constante, signifiant une turbulence atmosphérique stationnaire et homogène, qui n'est pas toujours le cas en réalité. La concentration de polluant $C(x,y,z)(g/m^3)$ en tout point de la zone étudiée est exprimée par l'équation suivante :

$$C(x,y,z) = \frac{Q}{2\pi u \sigma_y \sigma_z} exp(-\frac{y^2}{2\sigma_y^2})[exp(-\frac{(z-H)^2}{2\sigma_z^2}) + exp(-\frac{(z+H)^2}{2\sigma_z^2})] \quad (4.4)$$

Avec :

- Q (g/s) est le taux d'émission de polluants,
- u (m/s) est la vitesse du vent,
- σ_y (m) et σ_z (m) sont des coefficients de dispersion horizontale et verticale définis par les classes de stabilité atmosphérique et
- H (m) est la hauteur hypothétique du rejet effectif du polluant, qui est égale à la somme de la hauteur de la source de polluant plus l'élévation du panache.

Nous avons déterminé les coefficients σ_y (m) et σ_z (m) en se basant sur les données de stabilité (exprimée par la hauteur de la couche limite divisée par la longueur de Monin-Obukhov) générées par le modèle ADMS. Le manuel d'ADMS (`https://www.cerc.co.uk/environmental-software/assets/data/doc_userguides/CERC_ADMS_5_2_User_Guide.pdf`), détaille les étapes de calcul des valeurs de ces coefficients en se basant sur les données de stabilité, selon le classement de catégories de stabilité de Pasquill-Gifford [139].

Plusieurs systèmes de modélisation atmosphérique avancés se basent sur le GPM comme ADMS-Urban.

4.3.2 Modélisation par le modèle ADMS-Urban

Atmospheric Dispersion Modeling System (ADMS), est une famille de systèmes conçue pour modéliser la qualité de l'air. Les données utilisées dans ce chapitre sont des données modélisées par ADMS-Urban, fournies de la part d'ATMO Hauts-de-France.

ADMS-Urban est un outil avancé de modélisation de la pollution atmosphérique, qui a été développé dans le but de fournir des calculs à haute résolution des concentrations de pollution compatibles avec toutes les zones d'étude de l'environnement urbain, quelle que soit leur taille. Ce modèle multi-échelle peut être utilisé pour examiner les concentrations à proximité d'un seul carrefour routier ou sur une zone s'étendant sur l'ensemble d'une grande ville. ADMS-Urban a été largement utilisé lors des évaluations de la qualité de l'air effectuées par les autorités locales au Royaume-Uni et au cours d'une large gamme d'études de planification et de politique à travers le monde (`www.cerc.co.uk`).

ADMS-Urban est une des versions du système de modélisation de la dispersion atmosphérique (ADMS), qui a été conçu dans l'intention d'étudier les impacts des émissions des installations industrielles. ADMS-Urban permet une caractérisation complète de la grande variété d'émissions observée dans les zones urbaines, y compris un modèle d'émissions du trafic routier. Il comprend également d'autres fonctionnalités, notamment :

- Les effets du mouvement des véhicules sur la dispersion des émissions du trafic,
- Les effets d'un terrain complexe sur la dispersion des polluants, et
- Les effets d'un bâtiment sur la dispersion des polluants émis à proximité.

ADMS-Urban utilise un préprocesseur météorologique qui calcule les paramètres de la couche limite à partir d'une variété de données d'entrée, notamment la date et l'heure, la vitesse et la direction du vent, la température de surface et la couverture nuageuse. Les données météorologiques peuvent être des données brutes, moyennées toutes les heures ou analysées statistiquement.

Les émissions dans l'atmosphère d'une zone urbaine proviennent généralement d'une grande variété de sources. Dans une seule et même zone urbaine étudiée, il y aura probablement des émissions industrielles provenant des cheminées, ainsi que des émissions provenant de la circulation routière et des systèmes de chauffage domestique. Les types de sources explicites

disponibles dans ADMS-Urban pour représenter les différentes configurations d'émissions sont :

- Les routes : pour lesquelles les émissions sont spécifiées en termes de flux de véhicules, et de la dispersion initiale supplémentaire causée par les véhicules en mouvement.
- Les points industriels : pour lesquels la montée du panache de la cheminée est incluse dans la modélisation.
- Les zones : où une ou plusieurs sources sont le mieux représentées comme étant uniformément réparties sur une zone.
- Les volumes : où une source ou des sources sont mieux représentées comme étant uniformément réparties dans un volume.

ADMS-Urban intègre des modules dans son système pour tenir compte des effets complexes des différents éléments agissant sur l'atmosphère, et qui sont :

— un module de chimie « Generic Reaction Set (GRS) » [140], qui permet d'introduire des réactions complexes dues aux différentes composées chimiques présentes dans l'atmosphère,

— un module de terrain, pour prendre en considération l'influence de la topographie sur la qualité de l'air modélisé en utilisant FLOWSTAR [141],

— et un module pour les canyons de rues, pour inclure l'effet des canyons de rue sur la dispersion des émissions du trafic routier, en se basant sur Operational Street Pollution Model (OSPM) [142].

Pour mettre en œuvre cette modélisation, ADMS implique d'avoir recours principalement aux données du cadastre des émissions de polluants (qui recense la nature et la quantité de ces polluants selon leur localisation), ainsi que les données météorologiques et topographiques.

4.4 Prétraitement des données

Les entrées utilisées d'ADMS sont des données horaires du 01-01-2013 au 30-03-2013 concernant les données de : mesures de particules de PM_{10}, longitude, latitude, altitude, direction du vent, vitesse du vent et stabilité de l'atmosphère (exprimée par la hauteur de la couche limite divisée par la longueur de Monin-Obukhov).

Les données des concentrations de PM_{10}, la longitude et la latitude ont été utilisées pour exprimer la pollution atmosphérique modélisée par ADMS. Pour générer les données d'entrée du GPM, la longitude, la latitude, l'altitude, la direction du vent, la vitesse du vent et la stabilité de l'atmosphère ont été employées. Les données brutes qui nous ont été fournies sont 14 fichiers générés par ADMS, ayant chacun une résolution de maillage de points couvrant une partie de l'agglomération de Dunkerque. Ces points indiquent où les concentrations de PM_{10} ont été modélisées (longitude, latitude et altitude). Les données de sortie générées et stockées dans ces fichiers sont les concentrations de polluants, le temps de mesure, la vitesse et la direction du vent et la stabilité.

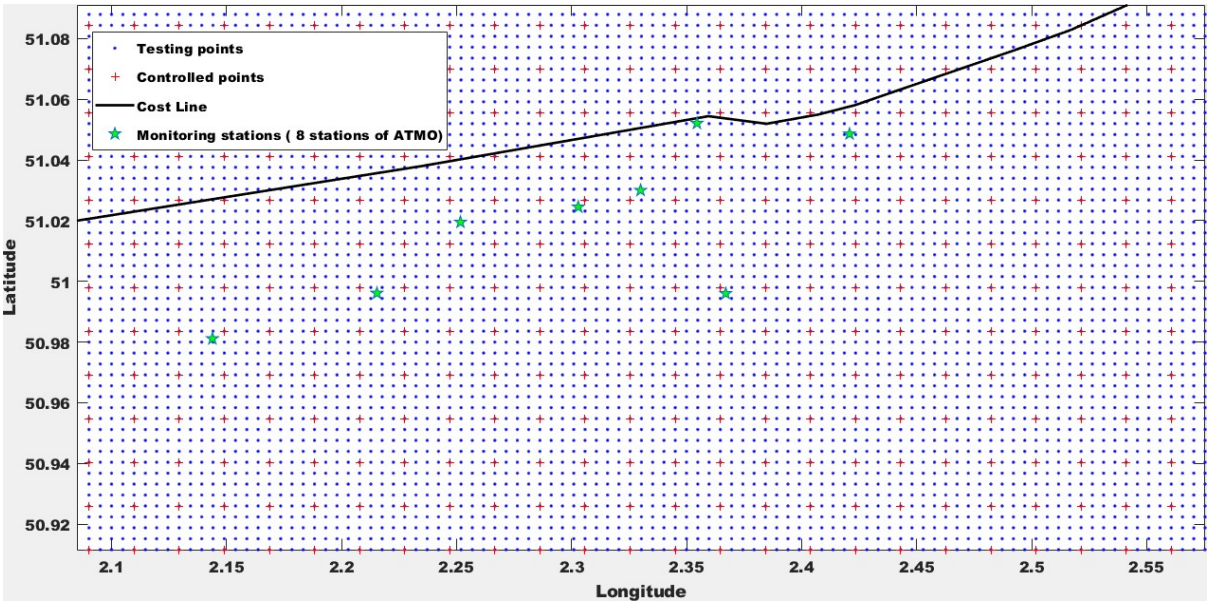

FIGURE 4.1 – Les deux maillages réguliers adoptés : points de test en bleu (valeurs ADMS) et points de contrôle en rouge (où la précision d'optimisation est calculée via RMSE).

Après avoir filtré les valeurs aberrantes des 14 fichiers que nous avons combiné en un seul, nous avons dû préparer les données sur lesquelles nous allons travailler en définissant deux maillages réguliers comme illustré dans la figure 4.1. Le premier maillage représente les points de test avec 50x100 points, pour avoir des données régulièrement distribuées sur toute la zone d'étude, et qui représentent les données brutes générées par ADMS extraites des fichiers combinés par la méthode IDW. Le deuxième maillage illustre les points de contrôle, ils sont utilisés pour calculer l'erreur de configuration des stations obtenues via l'optimisation, par la différence entre les valeurs ADMS et les valeurs estimées par la fonction de fitness IDW.

4.5 Contributions

Dans ce chapitre, on s'intéresse à deux types de modélisation de la pollution atmosphérique : ADMS et le modèle de panache gaussien. Ces derniers sont utilisés pour la génération des données de vérité de terrain des concentrations de PM_{10}, sur lesquels nous proposons d'appliquer une optimisation des positions des stations de mesure pour estimer ces concentrations. Les résultats de cette optimisation nous permettront d'analyser et d'étudier les configurations obtenues des stations de mesures optimisées pour les données des deux modèles de dispersion, et de les comparer au réseau de surveillance de la qualité de l'air réel d'ATMO. Deux algorithmes des plus utilisés dans la littérature [143] : l'algorithme génétique [132] et l'optimisation des essaims de particules (P.S.O) [134] sont utilisés pour trouver les positions optimales des stations de mesure. Nous proposons d'utiliser l'interpolation par pondération de distance inverse [105] afin d'estimer les valeurs de pollution, tout en minimisant l'erreur de cette interpolation. Une estimation de l'erreur d'interpolation en fonction du nombre de stations optimisées permettra d'évaluer le nombre de stations de surveillance nécessaires.
Et enfin, nous suggérons une approche permettant d'améliorer la précision de l'interpolation dans une région d'intérêt spécifique, par exemple la zone résidentielle de Dunkerque.

4.6 Résultats et discussion

Pour les données GPM, nous avons pris comme exemple une position réelle d'une source d'émission de pollution dans l'agglomération de Dunkerque qui représente l'une des sources industrielles (une des cheminées de l'aciérie Arcelor Mittal de Dunkerque : `https://fce.arcelormittal.com/dunkerque`). Nous nous sommes servis des autres sorties fournies par ADMS comme la vitesse et la direction du vent, la stabilité et l'altitude pour générer les données de GPM. Le taux d'émission a été estimée valeur par le cadastre d'émission ATMO, $Q = 19g/s$.
Le tableau 4.1 montre les données utilisées pour générer des données par GPM.

Paramètres	Valeurs
Longitude	2.2974
Latitude	51.0406
u (vitesse de vent)	Extraite d'ADMS
Direction de vent	Extraite d'ADMS
Statbilité (représentée par σ_y et σ_z)	Extraites d'ADMS
z (altitude)	Extraite d'ADMS
Q (taux d'émission)	19 g/s
H (height)	50 m

TABLE 4.1 – Paramètres utilisés pour générer les données GPM

Le tableau 4.2 présente les résultats de performance de RMSE de l'optimisation, des configurations obtenues par le GA et le PSO, ainsi que celle d'ATMO. Ces résultats concernent les deux ensembles de données du modèle GPM et ADMS. Les deux algorithmes ont réalisé une amélioration assez comparable par rapport aux stations d'ATMO pour les deux ensembles de données, avec PSO prenant plus de temps (environ 3 fois plus longtemps) que GA. Pour trouver le minimum global, les deux algorithmes doivent être exécutés plusieurs fois afin d'éviter que la sortie soit un des minimum locaux. Le coefficient de détermination (R^2) des données ADMS est beaucoup plus grand que celui des données GPM, qui ont pris plus de temps d'exécution que les données ADMS.

Par ailleurs, on observe que le R^2 pour les données GPM est très faible, cela pourrait être interprété par l'interpolation qui présente de meilleures performances lorsqu'il s'agit d'une pollution de l'air qui est largement dispersée, ce qui est le cas pour les sources diffuses (ex., chauffage domestique, aviation, etc.) modélisées par ADMS, contrairement aux données étroites du panache gaussien GPM (source de pollution canalisée), générées selon les conditions météorologiques propres à l'agglomération de Dunkerque, tel est le cas pour les sources d'émission industrielles.

L'interpolation a une mauvaise précision d'estimation pour les pics de pollution à des valeurs élevées avec une dispersion atmosphérique étroite (données GPM), ce qui produit des valeurs d'erreur RMSE importantes et de R^2 faibles. L'augmentation des coefficients de dispersion dans l'air (σ_y et σ_z dans l'équation (4.4)) donnerait des données GPM avec une dispersion

plus large, ce qui améliorerait la précision (RMSE et R^2) de l'interpolation.

	Données ADMS			Données modèle de panache gaussien		
	RMSE ($\mu g/m^3$)	R^2	Temps d'execution	RMSE ($\mu g/m^3$)	R^2	Temps d'execution
Algorithme génétique	4.15	0.94	1h30min [1]	6.83	0.09	3h24min[1]
P.S.O	4.10	0.95	4h19min[1]	6.89	0.07	11h30min[1]
ATMO	4.46	0.93	–	7.08	0.02	–

TABLE 4.2 – Résultats d'optimisation des données d'ADMS et du modèle de panache gaussien par rapport à ATMO

Dans le but de comparer l'emplacement des stations des configurations issues de l'optimisation, nous avons tracé les cartes montrant les positions des stations obtenues par les deux algorithmes (GA et PSO) respectivement dans les figures 4.2 et 4.3 sur les données de la moyenne des trois mois des concentrations de PM_{10}.

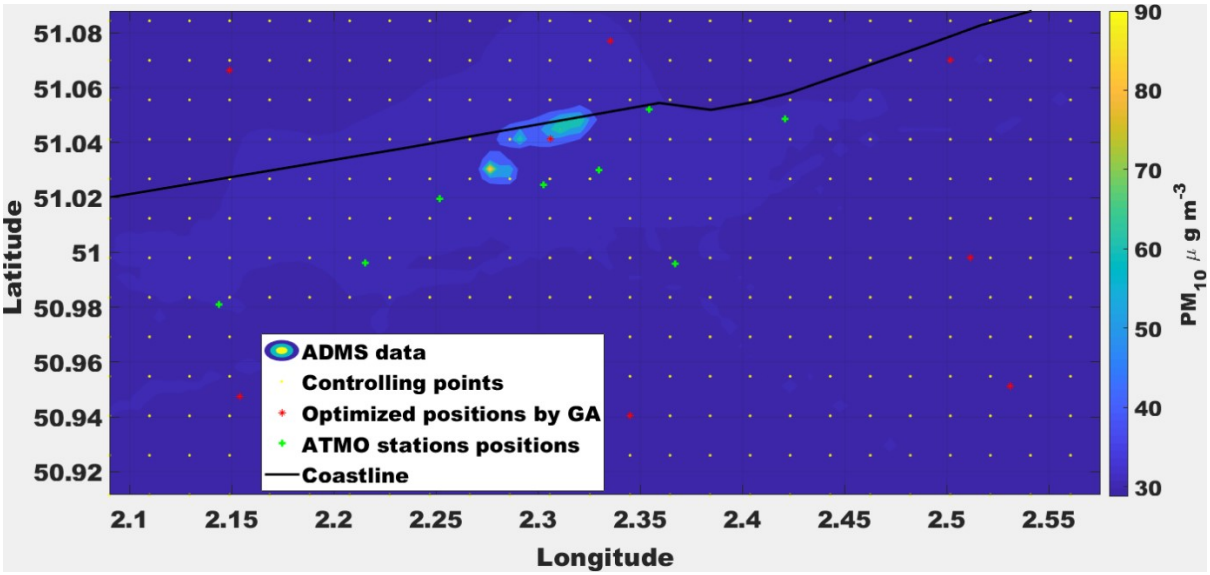

FIGURE 4.2 – Configuration des stations obtenue par GA sur les données ADMS de la moyenne des trois des concentrations de PM_{10}.

Les deux techniques d'optimisation GA et PSO ont produit des configurations où les stations de mesure sont réparties sur toute la zone d'étude, essayant ainsi de maintenir une valeur d'erreur la plus faible possible dans tous les points de contrôle et en particulier aux alentours des positions des stations résultantes. Contrairement aux stations ATMO, qui sont situées à côté des sources d'émission de l'agglomération. En outre, les deux stations qui sont

1. Ces calculs ont été réalisés sur un cluster de 36 coeurs du serveur de l'université de l'ULCO

en mer peuvent jouer un rôle pertinent en nous informant de la pollution de l'air circulant au niveau de la mer. En effet, en raison des phénomènes météorologiques locaux qui se produisent dans la zone côtière, la pollution de l'air émise sur la côte est en circulation continue entre terre et mer, permettant ainsi de surveiller la qualité de l'air au niveau de la mer.

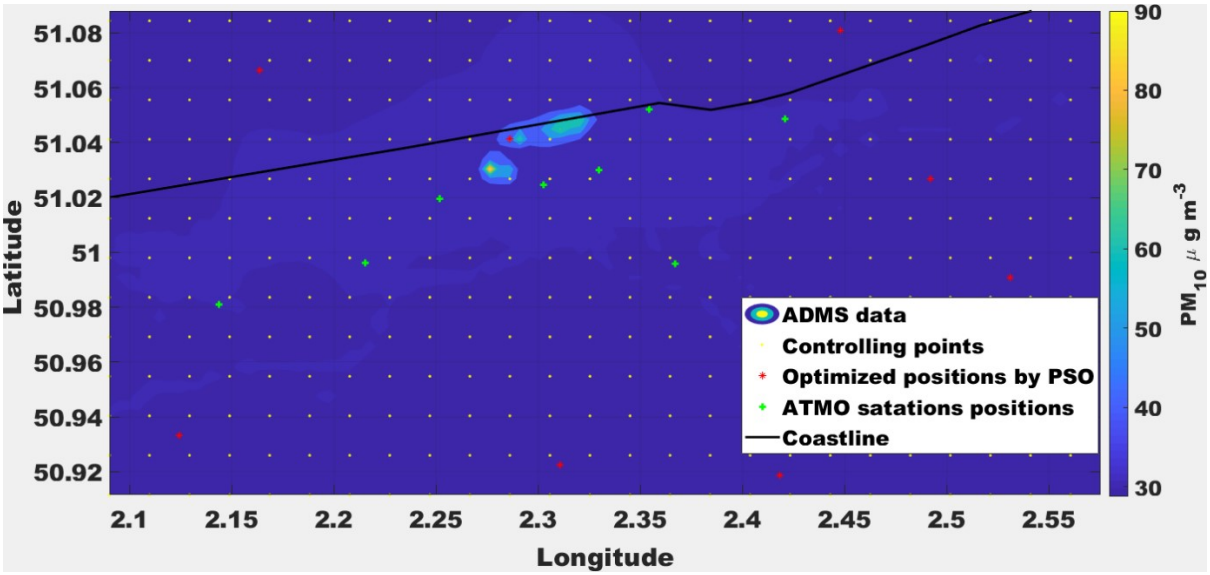

FIGURE 4.3 – Configuration des stations obtenue par PSO sur les données ADMS de la moyenne des trois des concentrations de PM_{10}.

Il convient de rappeler qu'ADMS modélise la pollution de l'air qui représente différentes sources d'émission (trafic, urbain, industriel, domestique, etc.) dispersée sur le domaine de la simulation, tandis que GPM représente certaines de ces sources (allant d'une à plusieurs).
La carte de la figure 4.4, illustre la configuration des stations de GA obtenue sur les données générées par GPM (en utilisant les paramètres présentés dans le tableau 4.1). On observe que la moyenne des trois mois des concentrations de PM_{10} de ce modèle est plus faible autour de la source d'émission par rapport à son environnement. Ceci est dû à la hauteur des cheminées, qui représente la hauteur à laquelle la source d'émission rejette les polluants, et également aux phénomènes météorologiques locaux, comme la turbulence, les polluants sont transportés de l'endroit où ils ont été émis vers les environs de cette dernière [144].
Les stations ont tendance à se placer à proximité de la source d'émission des données GPM que nous avons créées, contrairement aux résultats obtenus à partir des données ADMS (figures 4.2 et 4.3), puisque la méthode d'optimisation vise à placer les stations de manière à rendre le RMSE aussi petit que possible, où la variance de concentration est élevée (à côté de la

source d'émission). La modélisation ADMS prend en compte plusieurs sources d'émission (par exemple la pollution atmosphérique de fond (régionale), la pollution atmosphérique industrielle (locale), etc.), en plus de certains facteurs d'influence complexes tels que les caractéristiques topographiques de la zone étudiée du polluant et les phénomènes météorologiques qui font varier sa variance de concentration sur la zone d'étude. Alors que les valeurs de concentration de polluants simulées par GPM ne sont importantes qu'à proximité des sources d'émission, là où les stations de mesure finissent par être placées.

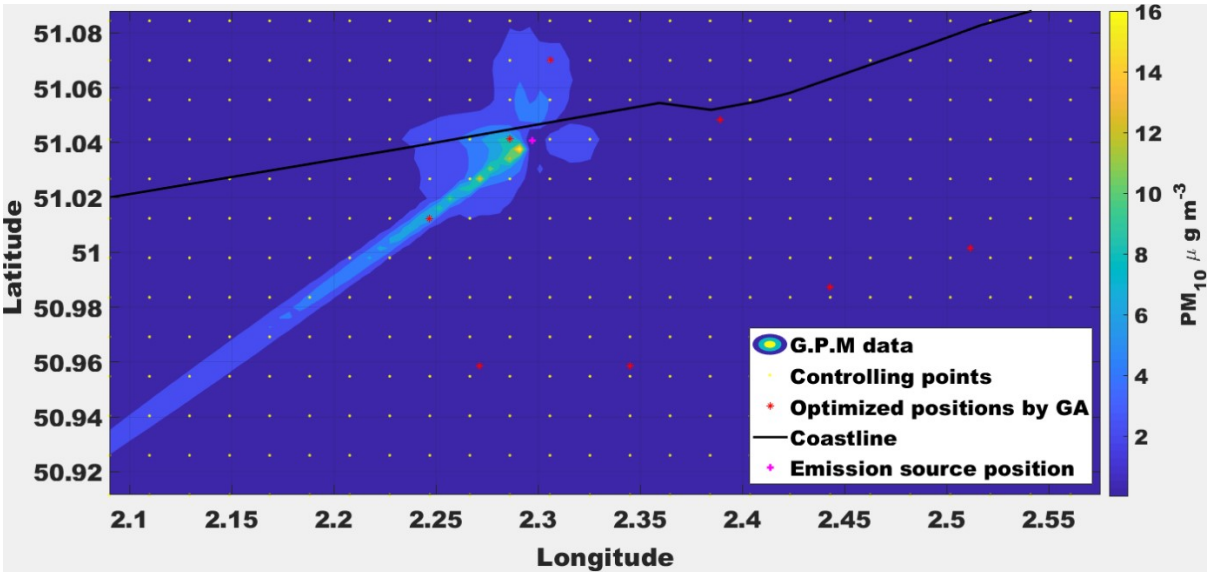

FIGURE 4.4 – Configuration des stations obtenue par PSO sur les données ADMS de la moyenne des trois des concentrations de PM_{10}.

Dans les figures 4.2 et 4.3 les deux algorithmes (GA et PSO) ont attribué uniformément des stations de mesure au niveau de l'agglomération dans ses différentes zones, en répartissant les stations de manière homogène sur toute la surface.

Nous venons proposer dans ce présent chapitre une approche pour l'amélioration de la précision de surveillance de la qualité de l'air dans une zone précise, et nous prenons à titre d'exemple la zone urbaine / industrielle de l'agglomération de Dunkerque, puisqu'elle représente à la fois une zone résidentielle et la principale zone émettrice de la pollution de l'air dans l'agglomération. Cette approche consiste en l'augmentation des points de contrôle dans la zone d'intérêt pour y obtenir une meilleure précision de l'interpolation. La figure 4.5 illustre la topologie résultante des stations de mesure, où nous avons rendu la résolution des points de contrôle huit fois plus importante dans la zone d'intérêt (la zone urbaine / industrielle) que dans le reste de toute

la zone d'étude (l'agglomération de Dunkerque).

Cette approche a permis d'augmenter le nombre de stations en zone d'intérêt en 4 stations (la moitié du nombre des stations disponibles), contre 4 stations restantes réparties uniformément dans le reste de la zone d'étude, ayant des points de contrôle moins denses. Par conséquent, le RMSE en zone urbaine / industrielle s'est réduit de 2,90 $\mu g/m^3$ avant densification des points de contrôle, à 2,32 $\mu g/m^3$ après application de cette dernière, ce qui représente environ 20% de diminution du RMSE et cela confirme l'efficacité de notre proposition comme moyen d'attribution à une zone spécifique plus de priorité par rapport aux autres.

L'utilisation de la norme p au lieu de la norme 2 comme fonction de coût est aussi une alternative efficace pour allouer plus de priorité aux zones avec de grandes valeurs d'erreur, en termes de précision de surveillance de la qualité de l'air, à chaque fois que la valeur de p devient plus importante.

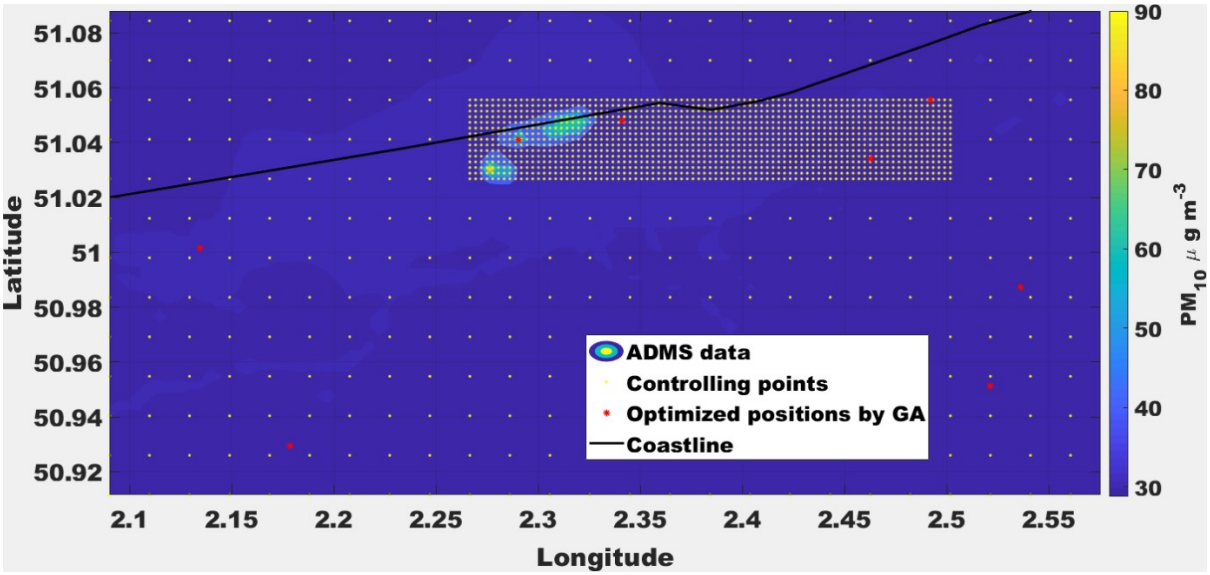

FIGURE 4.5 – Configuration des stations obtenue par GA pour une résolution 8 fois plus dense pour la zone urbaine / industrielle de l'agglomération de Dunkerque, sur la moyenne des données de PM_{10} d'ADMS de trois mois.

Le dernier point que nous avons examiné dans ce chapitre est l'influence du nombre des stations de mesure sur la précision de l'optimisation. Pour ce faire, nous avons tracé le RMSE d'optimisation GA obtenu pour analyser la sensibilité au nombre de stations de mesure optimisées (en faisant varier à chaque fois le nombre de stations de mesure), nous pouvons observer que de 3 à 20 stations le RMSE continue de diminuer légèrement à chaque fois qu'on

augmente le nombre de stations. Il convient également de noter qu'il est important d'exécuter l'algorithme plusieurs fois afin de vérifier que l'algorithme d'optimisation fournit l'optimum global comme sortie finale et non pas le minimum local.

Cette étude permet de voir le gain de précision apporté pour les différents nombres de stations déployées dans le réseau de surveillance, par exemple, pour une erreur d'interpolation inférieure à 4 $\mu g/m^3$ au moins 13 stations de surveillance doivent être déployées ; ce résultat semble être intéressant pour les associations de surveillance de la qualité de l'air comme ATMO.

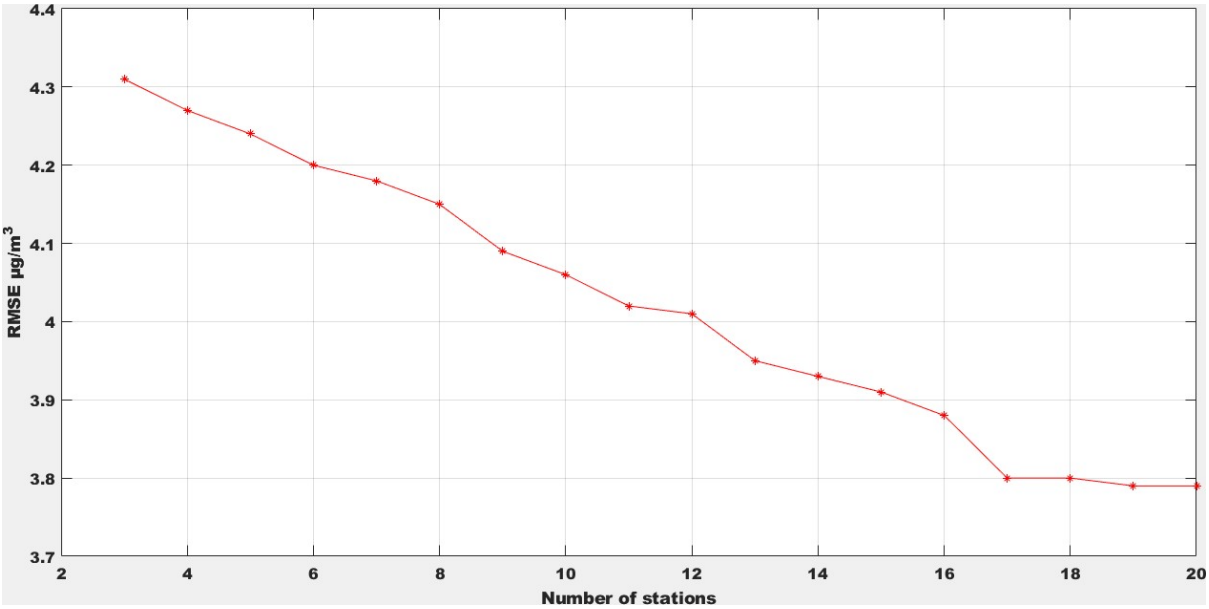

FIGURE 4.6 – RMSE de trois mois des concentrations de PM_{10} en fonction du nombre de stations utilisées par l'optimisation de GA.

4.7 Conclusions

Dans ce chapitre, nous avons analysé les configurations obtenues des stations de mesure par deux algorithmes d'optimisation : l'algorithme génétique et l'optimisation des essaims de particules, pour deux types de données de vérité de terrain (ground truth) : les données ADMS et les données du modèle de panache gaussien. Nous avons pu améliorer la précision d'estimation des concentrations PM_{10} par rapport à la topologie d'ATMO. L'erreur représentée par la différence entre ces valeurs de vérité de terrain et celles estimées par IDW représente la fonction de coût, utilisée dans les deux méthodes d'optimisation. Les positions optimisées des stations de mesure par ces deux dernières sont réparties sur toute la zone d'étude (agglomé-

ration de Dunkerque) lorsqu'elles sont appliquées sur les données ADMS, tandis que pour les données de panache gaussien, nous avons remarqué que les stations de mesure ont tendance à se positionner autour de la source d'émission.

D'autre part, nous avons pu obtenir une meilleure précision d'interpolation dans la zone urbaine/ industrielle de Dunkerque, en augmentant le nombre de points de contrôle (là où nous évaluons la performance de l'estimation des concentrations de PM_{10} par la fonction de coût) dans cette zone. Ceci a fait que le nombre des stations de mesure qui se situent dans la zone d'intérêt sélectionnée soit égal aux nombre de stations présentes dans le reste de la zone d'étude, menant par conséquent à une meilleure précision d'estimation dans la zone urbaine/industrielle.

La sensibilité de la précision d'interpolation au nombre de stations utilisées dans l'estimation a été étudiée. Le rôle important de la densité de données dans l'amélioration de la performance de l'interpolation a été montré, ainsi que le gain de précision apporté lors de chaque nouvel ajout de stations.

Pour pallier l'incapacité de l'interpolation à estimer, avec une bonne précision (R^2), les données générées par GPM pour un réseau de surveillance doté d'un nombre réduit de stations de mesure, les méthodes d'assimilation des données devront être appliquées en couplage avec les modèles de chimies-transport comme WRF-Chem (`https://www2.acom.ucar.edu/wrf-chem`).

Une dernière perspective de ce travail est de modifier l'algorithme dans le but de concevoir un réseau de capteurs sans fil pouvant remplacer les stations de mesures fixes. Dans ce cas, d'autres contraintes doivent être introduites comme la communication entre capteurs, la consommation d'énergie, et l'instauration d'un système de sauvegarde (backup system), etc.

Conclusion générale

Compte tenu des menaces qu'elle présente à la survie de notre planète, plusieurs efforts ont été déployés depuis des décennies pour étudier et comprendre la pollution atmosphérique, conditionnée par les sources d'émission exposées à une météorologie à variabilité spatio-temporelle complexe. La région des Hauts-de-France et l'agglomération de Dunkerque se sont avérées être des sites d'étude pertinents, de par la diversité des sources d'émission et de la complexité des phénomènes météorologiques qu'on y rencontre. Cette thèse représente le premier travail qui s'est intéressé à l'étude de la pollution atmosphérique au niveau de ces deux zones, en couplant les deux approches ; physico-chimie de l'atmosphère et informatique.

1. Contributions

La première tâche qu'on s'est donnée était d'effectuer une synthèse bibliographique sur les différents outils et techniques utilisées dans les travaux menés sur la pollution atmosphérique. Dans ce cadre, et d'un point de vue informatique, nous avons proposé une revue qui peut servir de guide pour chaque personne ayant l'intention de construire un Air Quality Model (AQM). Cette étude donne des recommandations concernant le choix de méthode possible pour développer un AQM, selon les données d'entrée disponibles (polluants, météorologie, trafic, densité, etc.).

Pour l'estimation de la pollution de l'air au niveau de la région de Hauts-de-France, un ensemble de techniques d'interpolation ont été appliquées sur les concentrations de PM_{10}, fournit par ATMO Hauts-de-France. Les méthodes de pondération inverse à la distance avec ses deux versions et la régression par les processus Gaussiens ont donné des résultats similaires dans l'estimation des concentrations de PM_{10}. La pondération inverse à la distance avec une puissance de distance optimisée a été ajustée pour s'adapter aux différentes situations de pollution

traitées. Pour toutes les techniques d'interpolation appliquées, l'erreur d'estimation était plus grande dans les zones proches des sources d'émission et dans les zones soumises à des phénomènes atmosphériques intensifs (zones côtières et vallonnées). L'impact de la variabilité temporelle de la pollution de l'air sur la précision de l'interpolation a été examiné en chapitre 3, pour pouvoir montrer le rôle que les phénomènes météorologiques locaux jouent dans la dispersion des polluants. Il s'est avéré que ces phénomènes ont une contribution importante sur la variabilité de la pollution de l'air, ce qu'on a pu montrer en moyennant les données d'entrée par des échelles temporelles qui ont aidé à filtrer l'influence des phénomènes météorologiques locaux sur la qualité de l'air. En outre, la sensibilité des méthodes d'interpolation à la densité des données ainsi que les perturbations dans les mesures ont été examiné. Cette étude a mis en évidence la nécessité de considérer les éléments qui contribuent à la dispersion de la pollution de l'air, tout comme la météorologie, dans la tâche de mesure/ modélisation de la qualité de l'air.

Le dernier chapitre a été consacré à l'optimisation du réseau de surveillance de la qualité de l'air de l'agglomération de Dunkerque. Cette optimisation a été appliquée sur deux types de données de vérité de terrain (ground truth) : les données ADMS et les données du modèle de panache gaussien ont été générées. L'erreur représentée par la différence entre ces valeurs de vérité de terrain et celles estimées par la méthode de pondération de distance inverse, représente la fonction de coût dans le processus d'optimisation. Les configurations obtenues des stations de mesure par deux algorithmes d'optimisation : l'algorithme génétique et l'optimisation des essaims de particules, donnent un réseau de stations étalées sur toute l'agglomération dunkerquoise, tandis que pour les données de panache gaussien elles se sont localisées autour de la source d'émission. Cela nous a permis de comparer ces topologies avec celle du réseau réel d'ATMO, qui positionne ses stations juste à côté des sources d'émission. Une approche d'amélioration de précision d'estimation de la pollution de l'air a été proposée enfin de ce travail. En augmentant le nombre de points de contrôle (là où nous évaluons la performance de l'estimation des concentrations de PM_{10} par la fonction de coût) dans notre zone d'intérêt, la précision dans cette dernière est devenue meilleure en raison du nombre important de stations de mesure résultant dans cette même zone.

Ce travail a donc permis de mettre l'accent sur le rôle de la météorologie et des sources d'émission dans la détermination de la qualité de l'air, en prenant pour exemple concret la région Hauts-de-France. Par ailleurs, on peut penser à proposer une optimisation de la topologie de surveillance de la qualité de l'air de l'agglomération Dunkerquoise.

2. Perspectives

En guise de continuité à ce travail, il serait important de mener le même processus d'optimisation sur des données de modèles 3D de dispersion de la pollution atmosphérique comme CMAQ [145] ou Meso-Nh / Chem [146], qui prennent en considération la dynamique atmosphérique à multi-échelles. Cela permettra l'étude de l'influence de la variabilité régionale et locale des phénomènes météorologiques sur la pollution de l'air.

Les méthodes d'assimilation de données sont un outil utile pour l'estimation de la pollution qui décrivent le comportement dynamique de l'atmosphère, comme WRF-Chem (`https://www2.acom.ucar.edu/wrf-chem`). Ces méthodes peuvent être considérées comme une approche alternative plus précise et avancée à l'interpolation, pour pallier sa faible précision pour les sources de pollution canalisées quand le nombre de stations de mesure est limité. En outre, nous envisageons une adaptation de l'optimisation effectuée en chapitre 4, pour qu'elle soit valide sur un réseau de capteurs sans fil au lieu de stations de mesures fixes. Dans ce cas, d'autres contraintes devraient être prises en compte, comme la communication entre capteurs, la consommation d'énergie, la prévoyance/mise en place d'un système de sauvegarde (backup system), etc.

Bien que notre proposition d'optimisation soit principalement conçue pour la surveillance de la pollution de l'air, elle peut également être utilisé pour d'autres problématiques environnementales, comme la pollution de l'eau au niveau de la même région étudiée.

Annexe

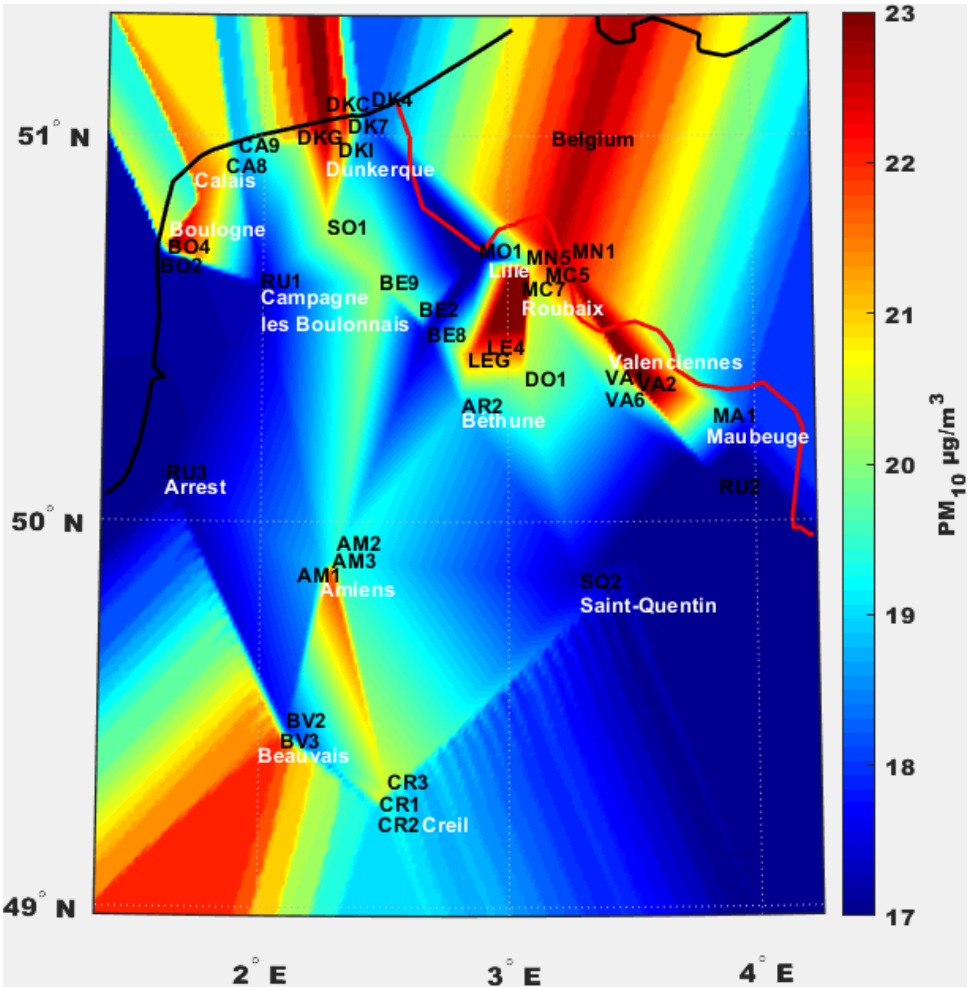

FIGURE 4.1 – Moyenne annuelle de l'année 2016 des concentrations de PM_{10} pour la région Hauts-de-France (utilisant l'interpolation linéaire de plus du voisin le plus proche pour extrapolation).

La figure en dessus est la carte de la moyenne annuelle de l'année 2016 des concentrations de PM_{10} pour la région Hauts-de-France, par la méthode basée-triagulation : l'interpolation linéaire. Cette carte donne une représentation incorrecte d'un point de vue physico-chimie de l'atmosphère de la distribution spatiale de la pollution de l'air, raison pour laquelle nous avons

choisi IDW en tant que moyen d'interpolation de toutes les cartes en chapitre 3.

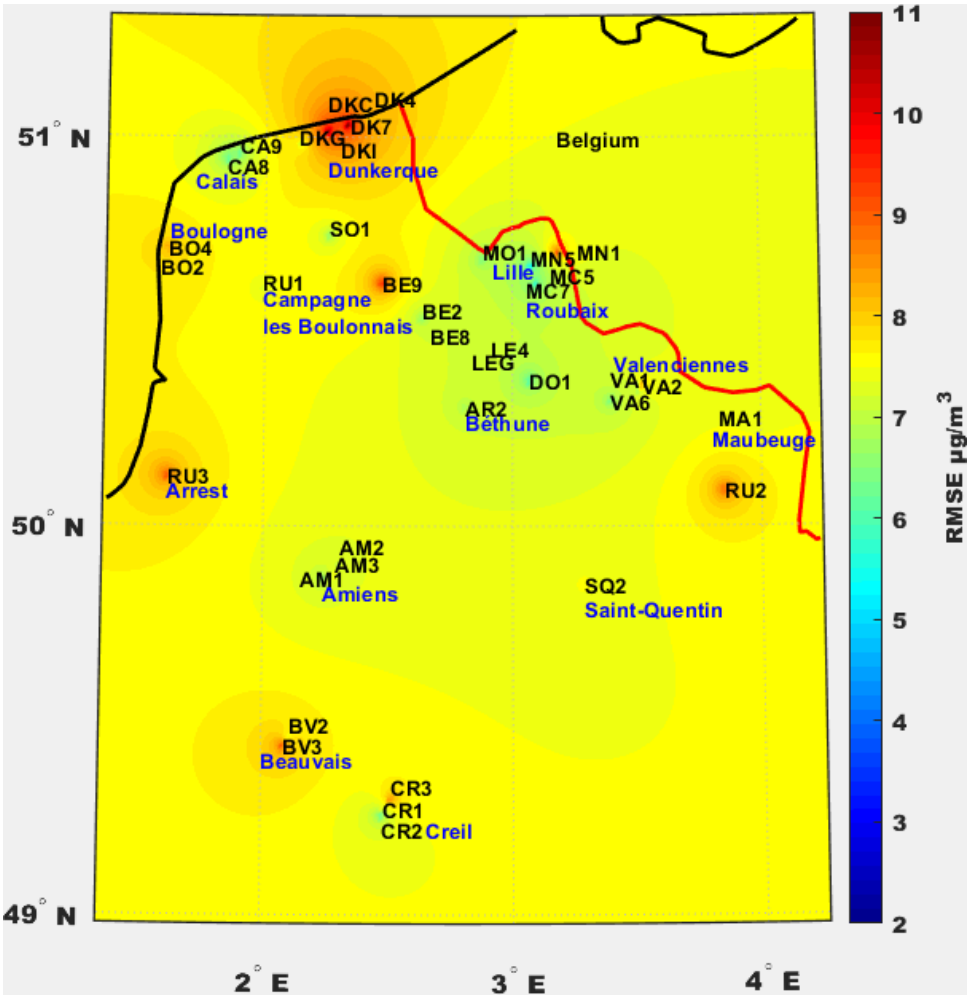

FIGURE 4.2 – RMSE de l'interpolation par IDW des concentrations de PM_{10} pour la région Hauts-de-France (pour les données quart horaires).

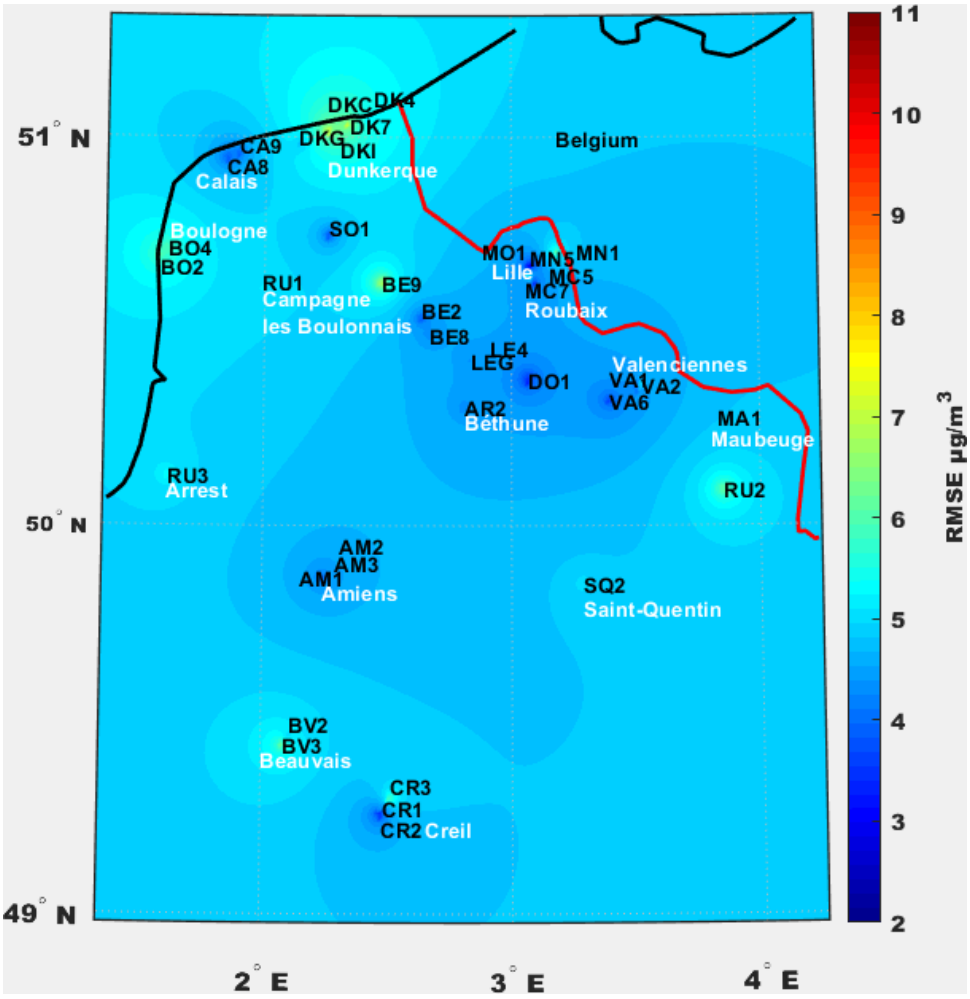

FIGURE 4.3 – RMSE de l'interpolation par IDW des concentrations de PM_{10} pour la région Hauts-de-France (pour les données moyennées journalières).

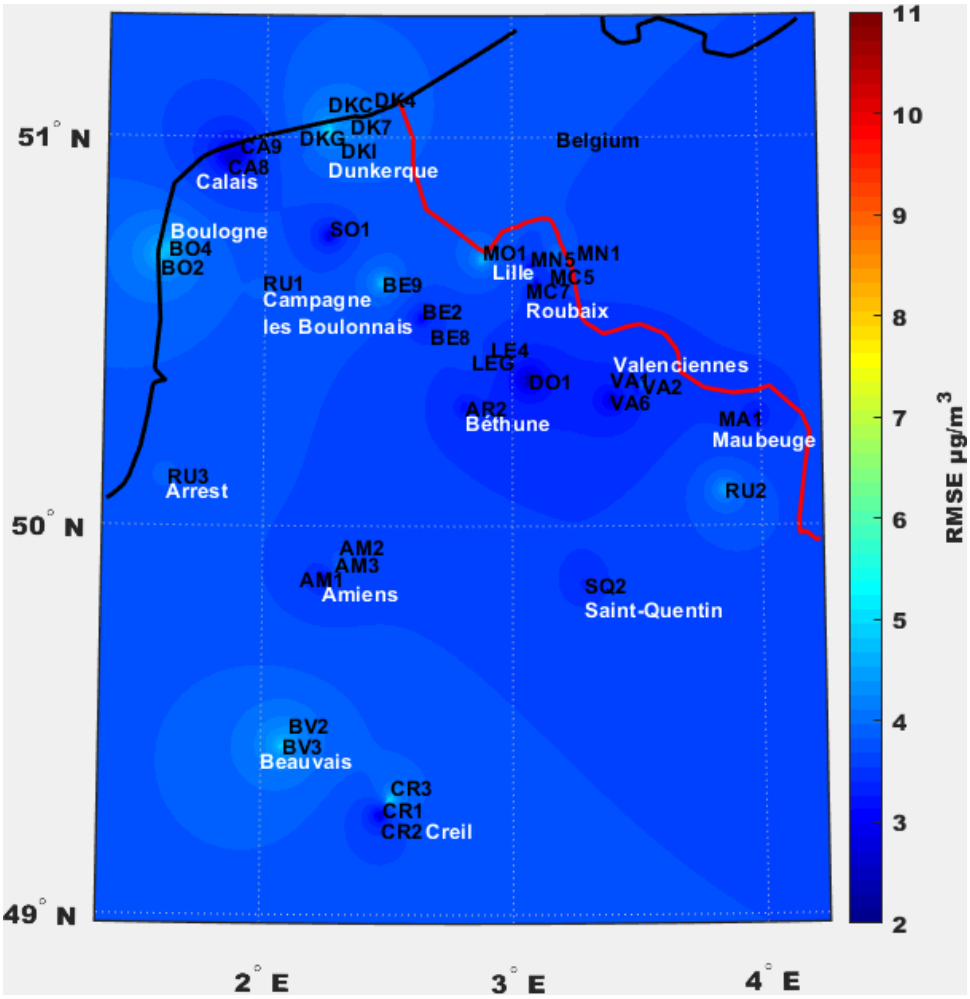

FIGURE 4.4 – RMSE de l'interpolation par IDW des concentrations de PM_{10} pour la région Hauts-de-France (pour les données moyennées hebdomadaires).

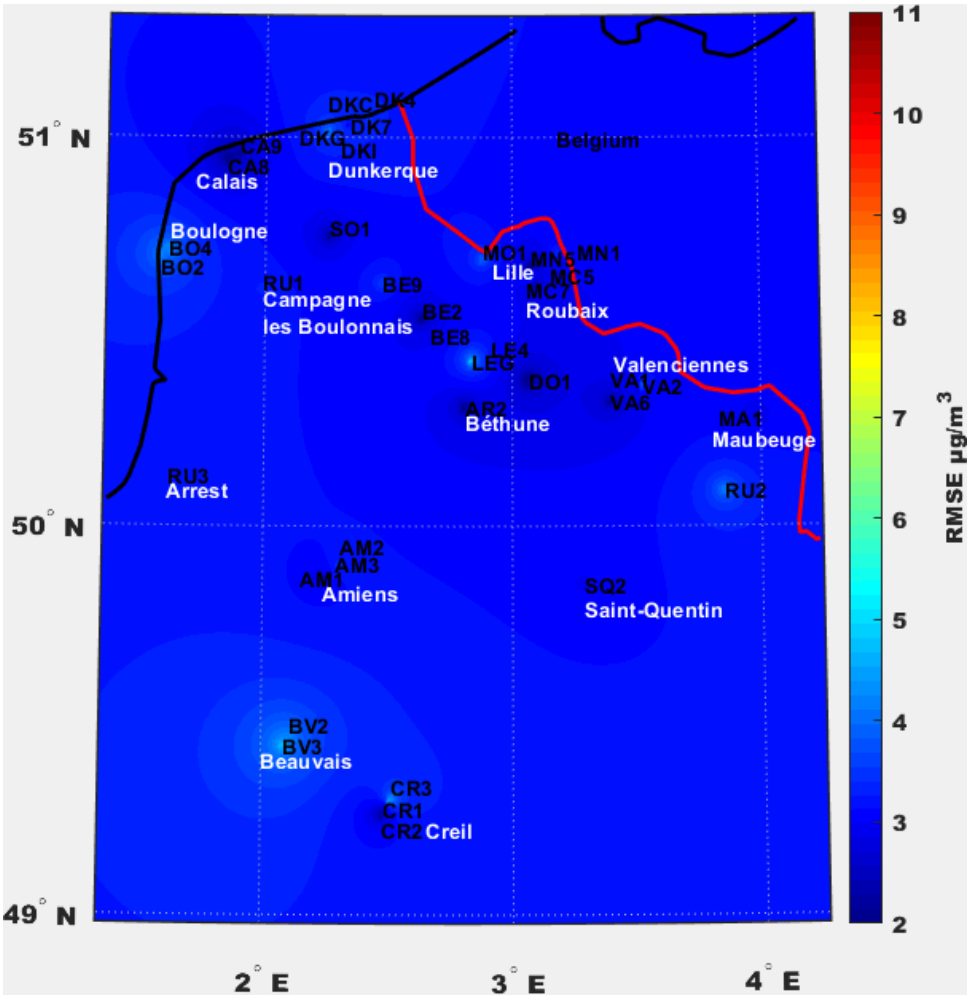

FIGURE 4.5 – RMSE de l'interpolation par IDW des concentrations de PM_{10} pour la région Hauts-de-France (pour les données moyennées mensuelles).

Les figures ci-dessus montrent la distribution spatiale du RMSE pour de différentes échelles temporelles de moyennage des concentrations de PM_{10}. Nous avons appliqué la méthode d'interpolation d'IDW pour sur les données moyennées de : 15 minutes, un jour, une semaine et un mois, pour pouvoir comparer la variation de RMSE par rapport aux périodes temporelles de moyennage. La figure 2 correspond au RMSE des concentrations de PM_{10} quart horaires fournies par ATMO Hauts-de-France, les grandes valeurs de RMSE correspondent aux zones proches géographiquement des sources d'émission (comme indiqué en section 3.5.1). En passant à une interpolation spatiale pour la période de moyennage d'un jour (figure 3), nous observons une diminution des valeurs d'erreur RMSE dans toute la carte, mais qui restent toujours élevées aux mêmes zones indiquées dans la carte précédente (figure 2). Ceci grâce au filtrage de l'influence des phénomènes météorologiques avec une périodicité d'un jour sur

les concentrations de PM_{10}. Ce RMSE continue de diminuer avec la croissance de la période de moyennage temporel, tel que les figures 4 et 5 le montre, où l'influence des phénomènes météorologiques à échelle temporelle d'une semaine (figure 4) et un mois (figure 5) a été filtré.

Références bibliographiques

[1] United Nations Population Division. *World Urbanization Prospects : The 2003 Revision*. 2004. URL : http://www.un.org/esa/population/publications/wup2003/2003WUPHighlights.pdf (visité le 01/06/2020).

[2] Alexander Baklanov, Luisa T Molina et Michael Gauss. "Megacities, air quality and climate". In : *Atmospheric Environment* 126 (2016), p. 235-249.

[3] Mark R Montgomery. "The urban transformation of the developing world". In : *science* 319.5864 (2008), p. 761-764.

[4] Demographia. *World Urban Areas (World Agglomerations) : 10th Annual Edition*. Mai 2014. URL : http://www.demographia.com/db-worldua.pdf (visité le 02/05/2020).

[5] United Nations. *World Urbanization Prospects : the 2011 Revision*. Août 2012. URL : http://esa.un.org/unpd/wup/Documentation/final-report.htm (visité le 06/01/2020).

[6] Timothy M Butler et Mark G Lawrence. "The influence of megacities on global atmospheric chemistry : a modelling study". In : *Environmental Chemistry* 6.3 (2009), p. 219-225.

[7] CERC. *ADMS 5.2. User Guide Cambridge Environmental Research Consultants Ltd*. Nov. 2016.

[8] Royal Netherlands Meteorological Institute (KNMI). *Luchtverontreiniging en weer. Staatsuitgeverij, Den Haag*. 1979.

[9] Ghassan B HAMRA et al. "Lung cancer and exposure to nitrogen dioxide and traffic : a systematic review and meta-analysis". In : *Environmental health perspectives* 123.11 (2015), p. 1107-1112.

[10] Charbel AFIF et al. "SO2 in Beirut : air quality implication and effects of local emissions and long-range transport". In : *Air Quality, Atmosphere & Health* 1.3 (2008), p. 167-178.

[11] Xinrong REN et al. "Ozone Production and Its Sensitivity to NO x and VOCs : Results from the DISCOVER-AQ Field Experiment, Houston 2013". In : *AGUFM* 2016 (2016), A14A-01.

[12] Bert BRUNEKREEF et Stephen T HOLGATE. "Air pollution and health". In : *The lancet* 360.9341 (2002), p. 1233-1242.

[13] William C HINDS. *Aerosol technology : properties, behavior, and measurement of airborne particles*. John Wiley & Sons, 1999.

[14] Kazuhiko ITO, Nan XUE et George THURSTON. "Spatial variation of $PM_{2.5}$ chemical species and source-apportioned mass concentrations in New York City". In : *Atmospheric Environment* 38.31 (2004), p. 5269-5282.

[15] C KETTERER et al. "Investigation of the planetary boundary layer in the Swiss Alps using remote sensing and in situ measurements". In : *Boundary-layer meteorology* 151.2 (2014), p. 317-334.

[16] P THUNIS et R BORNSTEIN. "Hierarchy of mesoscale flow assumptions and equations". In : *Journal of the atmospheric sciences* 53.3 (1996), p. 380-397.

[17] Isidoro ORLANSKI. "A rational subdivision of scales for atmospheric processes". In : *Bulletin of the American Meteorological Society* (1975), p. 527-530.

[18] John H SEINFELD et Spyros N PANDIS. *Atmospheric chemistry and physics : from air pollution to climate change*. John Wiley & Sons, 2016.

[19] Richard A ANTHES. "The general question of predictability". In : *Mesoscale Meteorology and Forecasting*. Springer, 1986, p. 636-656.

[20] Yuh-Lang LIN. *Mesoscale dynamics*. T. 630. Cambridge University Press Cambridge, 2007.

[21] Margarita G EVTYUGINA et al. "Photochemical pollution under sea breeze conditions, during summer, at the Portuguese West Coast". In : *Atmospheric Environment* 40.33 (2006), p. 6277-6293.

[22] Roland B STULL. "Mean boundary layer characteristics". In : *An Introduction to Boundary Layer Meteorology*. Springer, 1988, p. 1-27.

[23] National Research COUNCIL et al. *Acid deposition : atmospheric processes in eastern North America*. National Academy Press, 1983.

[24] WHO Regional Office for EUROPE et al. *Review of evidence on health aspects of air pollution–REVIHAAP Project*. WHO Regional Office for Europe, 2013.

[25] OMS. *Ambient air pollution - a major threat to health and climate*. 2018. URL : http://www.who.int/airpollution/ambient/en/ (visité le 12/07/2020).

[26] Michael JERRETT et al. "Long-term ozone exposure and mortality". In : *New England Journal of Medicine* 360.11 (2009), p. 1085-1095.

[27] Hong CHEN et al. "Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis : a population-based cohort study". In : *The Lancet* 389.10070 (2017), p. 718-726.

[28] Rob BEELEN et al. "Effects of long-term exposure to air pollution on natural-cause mortality : an analysis of 22 European cohorts within the multicentre ESCAPE project". In : *The Lancet* 383.9919 (2014), p. 785-795.

[29] Michel VEDRENNE et al. "An integrated assessment of two decades of air pollution policy making in Spain : Impacts, costs and improvements". In : *Science of the Total Environment* 527 (2015), p. 351-361.

[30] World Health ORGANIZATION. *Air quality guidelines : global update 2005 : particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. World Health Organization, 2006.

[31] Fitzgerald BOOKER et al. "The ozone component of global change : potential effects on agricultural and horticultural plant yield, product quality and interactions with invasive species". In : *Journal of Integrative Plant Biology* 51.4 (2009), p. 337-351.

[32] Jack FISHMAN et al. "An investigation of widespread ozone damage to the soybean crop in the upper Midwest determined from ground-based and satellite measurements". In : *Atmospheric Environment* 44.18 (2010), p. 2248-2256.

[33] GE SANDERS, JJ COLLS et AG CLARK. "Physiological changes in Phaseolus vulgaris in response to long-term ozone exposure". In : *Annals of Botany* 69.2 (1992), p. 123-133.

[34] Frédéric BOUVIER. "Le dispositif français de surveillance de la qualité de l'air". In : *POLLUTION ATMOSPHÉRIQUE* 1 (2012).

[35] Michael JERRETT et al. "A review and evaluation of intraurban air pollution exposure models". In : *Journal of Exposure Science & Environmental Epidemiology* 15.2 (2005), p. 185-204.

[36] Debanshee SAHA, Manasi SHINDE et Shail THADESHWAR. "IoT based air quality monitoring system using wireless sensors deployed in public bus services". In : *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing*. 2017, p. 1-6.

[37] M Emre KESKIN et al. "Wireless sensor network lifetime maximization by optimal sensor deployment, activity scheduling, data routing and sink mobility". In : *Ad Hoc Networks* 17 (2014), p. 18-36.

[38] Khaoula KARROUM et al. "A Review of Air Quality Modeling". In : *MAPAN* (2020), p. 1-14.

[39] Xingzhe XIE et al. "A review of urban air pollution monitoring and exposure assessment methods". In : *ISPRS International Journal of Geo-Information* 6.12 (2017), p. 389.

[40] François MATHE. *Evolution de la classification et des criteres d'implantation des stations de mesure de la qualite de l'air-participation a la reactualisation du guide de classification des stations*. 2010.

[41] 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.

[42] 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.

[43] Yang ZHANG et al. "Real-time air quality forecasting, part I : History, techniques, and current status". In : *Atmospheric Environment* 60 (2012), p. 632-655.

[44] Yang ZHANG et al. "Real-time air quality forecasting, part II : State of the science, current research needs, and future prospects". In : *Atmospheric Environment* 60 (2012), p. 656-676.

[45] 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.

[46] Stefania BERTAZZON et al. "Accounting for spatial effects in land use regression for urban air pollution modeling". In : *Spatial and spatio-temporal epidemiology* 14 (2015), p. 9-21.

[47] Hong CHEN et al. "Living near major roads and the incidence of dementia, Parkinson's disease and multiple sclerosis in Ontario, Canada : population-based study". In : *ISEE conference abstracts*. 2016.

[48] 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.

[49] 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.

[50] Jason G SU et al. "Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy". In : *Environmental research* 109.6 (2009), p. 657-670.

[51] Liang ZHAI et al. "Land use regression modeling of $PM_{2.5}$ concentrations at optimized spatial scales". In : *Atmosphere* 8.1 (2017), p. 1.

[52] Sumanta BASU et al. "Iterative random forests to discover predictive and stable high-order interactions". In : *Proceedings of the National Academy of Sciences* 115.8 (2018), p. 1943-1948.

[53] JB ORDIERES et al. "Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua)". In : *Environmental Modelling & Software* 20.5 (2005), p. 547-559.

[54] Wei XU et al. "$PM_{2.5}$ Air Quality Index Prediction Using an Ensemble Learning Model". In : *International Conference on Web-Age Information Management*. Springer. 2014, p. 119-129.

[55] Wei JIANG et al. "Using social media to detect outdoor air pollution and monitor air quality index (AQI) : a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter)". In : *PloS one* 10.10 (2015), e0141185.

[56] Georg A GRELL et al. "Fully coupled "online" chemistry within the WRF model". In : *Atmospheric Environment* 39.37 (2005), p. 6957-6975.

[57] 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.

[58] Ruiyun YU et al. "RAQ–A random forest approach for predicting air quality in urban sensing systems". In : *Sensors* 16.1 (2016), p. 86.

[59] 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.

[60] 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.

[61] Paul E BENSON. "A review of the development and application of the CALINE3 and 4 models". In : *Atmospheric Environment. Part B. Urban Atmosphere* 26.3 (1992), p. 379-390.

[62] 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.

[63] Yu ZHENG et al. "Forecasting fine-grained air quality based on big data". In : *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, p. 2267-2276.

[64] Tânia FONTES et Nelson BARROS. *Interpolation of air quality monitoring data in an urban sensitive area : the Oporto/Asprela case*. Edições Universidade Fernando Pessoa, 2010.

[65] Yuddy RAMOS et al. "Spatio-temporal models to estimate daily concentrations of fine particulate matter in Montreal : Kriging with external drift and inverse distance-weighted approaches". In : *Journal of exposure science & environmental epidemiology* 26.4 (2016), p. 405-414.

[66] Luis O RIVERA-GONZÁLEZ et al. "An assessment of air pollutant exposure methods in Mexico City, Mexico". In : *Journal of the air & waste management association* 65.5 (2015), p. 581-591.

[67] Yangjie GUO et al. "Satellite remote sensing of fine particulate matter ($PM_{2.5}$) air quality over Beijing using MODIS". In : *International Journal of Remote Sensing* 35.17 (2014), p. 6522-6544.

[68] Wei SUN et al. "Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California". In : *Science of the total environment* 443 (2013), p. 93-103.

[69] D KANG, R MATHUR et S Trivikrama RAO. "Assessment of bias-adjusted $PM_{2.5}$ air quality forecasts over the continental United States during 2007". In : *Geoscientific Model Development* 3.1 (2010), p. 309.

[70] Jui-Huan LEE et al. "Land use regression models for estimating individual NOx and NO_2 exposures in a metropolis with a high density of traffic roads and population". In : *Science of the total environment* 472 (2014), p. 1163-1171.

[71] Xiaofan YANG et al. "Development of $PM_{2.5}$ and NO_2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China". In : *Environmental pollution* 226 (2017), p. 143-153.

[72] Chao LIU et al. "A land use regression application into assessing spatial variation of intra-urban fine particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations in City of Shanghai, China". In : *Science of The Total Environment* 565 (2016), p. 607-615.

[73] Kin-Che LAM et al. "Environmental quality of urban parks and open spaces in Hong Kong". In : *Environmental monitoring and assessment* 111.1-3 (2005), p. 55-73.

[74] Hong GUO et al. "Comparison of four ground-level $PM_{2.5}$ estimation models using PARASOL aerosol optical depth data from China". In : *International journal of environmental research and public health* 13.2 (2016), p. 180.

[75] Yijun LIN et al. "Mining public datasets for modeling intra-city $PM_{2.5}$ concentrations at a fine spatial resolution". In : *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*. 2017, p. 1-10.

[76] Yijun LIN et al. "Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning". In : *Proceedings of the 26th ACM SIGSPATIAL international conference on advances in geographic information systems*. 2018, p. 359-368.

[77] Jamal JOKAR ARSANJANI et al. "Toward mapping land-use patterns from volunteered geographic information". In : *International Journal of Geographical Information Science* 27.12 (2013), p. 2264-2278.

[78] DK MOORE et al. "A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA". In : *Journal of Environmental Monitoring* 9.3 (2007), p. 246-252.

[79] Lianfa LI et al. "Modeling the concentrations of on-road air pollutants in southern California". In : *Environmental science & technology* 47.16 (2013), p. 9291-9299.

[80] Lianfa LI et al. "Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution". In : *Environmental science & technology* 51.17 (2017), p. 9920-9929.

[81] Sun-Young KIM, Lianne SHEPPARD et Ho KIM. "Health effects of long-term air pollution : influence of exposure prediction methods". In : *Epidemiology* (2009), p. 442-450.

[82] Adam A SZPIRO, Christopher J PACIOREK et Lianne SHEPPARD. "Does more accurate exposure prediction necessarily improve health effect estimates ?" In : *Epidemiology (Cambridge, Mass.)* 22.5 (2011), p. 680.

[83] Jihoon SEO et al. "Effects of meteorology and emissions on urban air quality : a quantitative statistical approach to long-term records (1999–2016) in Seoul, South Korea". In : *Atmospheric Chemistry and Physics* 18.21 (2018), p. 16121-16137.

[84] Joanna A KAMIŃSKA. "The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions : a case study in Wrocław". In : *Journal of environmental management* 217 (2018), p. 164-174.

[85] Van-Duc LE, Tien-Cuong BUI et Sang-Kyun CHA. "Spatiotemporal deep learning model for citywide air pollution interpolation and prediction". In : *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2020, p. 55-62.

[86] Jielan XIE et al. "The characteristics of hourly wind field and its impacts on air quality in the Pearl River Delta region during 2013–2017". In : *Atmospheric Research* 227 (2019), p. 112-124.

[87] Elżbieta RADZKA. "The Effect of Meteorological Conditions on Air Pollution in Siedlce". In : *Journal of Ecological Engineering* 21.1 (2020).

[88] Pengfei WANG et al. "Responses of $PM_{2.5}$ and O_3 concentrations to changes of meteorology and emissions in China". In : *Science of the Total Environment* 662 (2019), p. 297-306.

[89] John S IRWIN. "Statistical evaluation of centreline concentration estimates by atmospheric dispersion models". In : *International Journal of Environment and Pollution* 14.1-6 (2000), p. 28-38.

[90] Douglas G FOX. "Uncertainty in air quality modeling : a summary of the AMS workshop on quantifying and communicating model uncertainty, Woods Hole, Mass., September 1982". In : *Bulletin of the American Meteorological Society* 65.1 (1984), p. 27-36.

[91] MB BECK et al. "On the problem of model validation for predictive exposure assessments". In : *Stochastic Hydrology and Hydraulics* 11.3 (1997), p. 229-254.

[92] Joseph C CHANG et Steven R HANNA. "Air quality model performance evaluation". In : *Meteorology and Atmospheric Physics* 87.1-3 (2004), p. 167-196.

[93] Yongli ZHANG et Yuhong YANG. "Cross-validation for selecting a model selection procedure". In : *Journal of Econometrics* 187.1 (2015), p. 95-112.

[94] Scott A FRUIN et al. "Predictive model for vehicle air exchange rates based on a large, representative sample". In : *Environmental science & technology* 45.8 (2011), p. 3569-3575.

[95] Shike MEI et al. "Inferring air pollution by sniffing social media". In : *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE. 2014, p. 534-539.

[96] Awkash KUMAR et al. "Air quality mapping using GIS and economic evaluation of health impact for Mumbai city, India". In : *Journal of the Air & Waste Management Association* 66.5 (2016), p. 470-481.

[97] Mohammad Hassan EHRAMPOUSH et al. "A Comparison on Function of Kriging and Inverse Distance Weighting Models in PM10 Zoning in Urban Area". In : *Journal of Environmental Health and Sustainable Development* 2.4 (2017), p. 379-387.

[98] Changqing LIN et al. "Using satellite remote sensing data to estimate the high-resolution distribution of ground-level $PM_{2.5}$". In : *Remote Sensing of Environment* 156 (2015), p. 117-128.

[99] Xia MENG et al. "Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model". In : *Environmental Pollution* 208 (2016), p. 177-184.

[100] Neelakshi HUDDA et al. "Linking in-vehicle ultrafine particle exposures to on-road concentrations". In : *Atmospheric Environment* 59 (2012), p. 578-586.

[101] Der-Tsai LEE et Bruce J SCHACHTER. "Two algorithms for constructing a Delaunay triangulation". In : *International Journal of Computer & Information Sciences* 9.3 (1980), p. 219-242.

[102] Rolf KLEIN et Derick WOOD. "Voronoi diagrams based on general metrics in the plane". In : *Annual Symposium on Theoretical Aspects of Computer Science*. Springer. 1988, p. 281-291.

[103] Robin SIBSON. "A brief description of natural neighbour interpolation". In : *Interpreting multivariate data* (1981).

[104] PA LONGLEY et al. *Geographic Information Systems and Science. John Wiley (Second*. 2005.

[105] David F WATSON et GM PHILIP. "A refinement of inverse distance weighted interpolation". In : *Geo-processing* 2.4 (1985), p. 315-327.

[106] Qinghang HE, Zhenxi ZHANG et Chao YI. "3D fluorescence spectral data interpolation by using IDW". In : *Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy* 71.3 (2008), p. 743-745.

[107] Maduako Nnamdi IKECHUKWU et al. "Accuracy assessment and comparative analysis of IDW, spline and kriging in spatial interpolation of landform (Topography) : An experimental study". In : *Journal of Geographic Information System* 9.3 (2017), p. 354-371.

[108] Arilson José de OLIVEIRA JÚNIOR et al. "Aurora : Mobile application for analysis of spatial variability of thermal comfort indexes of animals and people, using IDW interpolation". In : *Computers and Electronics in Agriculture* 157 (2019), p. 98-101.

[109] Gail GONG. "Cross-validation, the jackknife, and the bootstrap : excess error estimation in forward logistic regression". In : *Journal of the American Statistical Association* 81.393 (1986), p. 108-113.

[110] George Elmer FORSYTHE. "Computer methods for mathematical computations." In : *Prentice-Hall series in automatic computation* 259 (1977).

[111] Carl Edward RASMUSSEN et Christopher KI WILLIAMS. "Gaussian Processes for Machine Learning the MIT Press". In : *Cambridge, Mass* (2006).

[112] 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.

[113] Jorge NOCEDAL et Stephen WRIGHT. *Numerical optimization*. Springer Science & Business Media, 2006.

[114] Cort J WILLMOTT. "Some comments on the evaluation of model performance". In : *Bulletin of the American Meteorological Society* 63.11 (1982), p. 1309-1313.

[115] Cort J WILLMOTT et Kenji MATSUURA. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". In : *Climate research* 30.1 (2005), p. 79-82.

[116] ST RAO et al. "Resampling and extreme value statistics in air quality model performance evaluation". In : *Atmospheric Environment (1967)* 19.9 (1985), p. 1503-1518.

[117] Bradley EFRON. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[118] Louis DE MESNARD. "Pollution models and inverse distance weighting : Some critical remarks". In : *Computers & Geosciences* 52 (2013), p. 459-469.

[119] Jin LI et al. "Application of machine learning methods to spatial interpolation of environmental variables". In : *Environmental Modelling & Software* 26.12 (2011), p. 1647-1659.

[120] Lixin Li et al. "Spatiotemporal interpolation methods for the application of estimating population exposure to fine particulate matter in the contiguous US and a real-time web application". In : *International journal of environmental research and public health* 13.8 (2016), p. 749.

[121] Lixin Li et al. "Fast inverse distance weighting-based spatiotemporal interpolation : a web-based application of interpolating daily fine particulate matter $PM_{2.5}$ in the contiguous US using parallel programming and kd tree". In : *International journal of environmental research and public health* 11.9 (2014), p. 9101-9141.

[122] MPB Pasini et al. "Ordinary Kriging and Inverse Distance Weighted applied in population spatialization of fig fly." In : *Revista Brasileira de Ciências Agrárias (Agrária)* 10.3 (2015), p. 452-459.

[123] Gea Oliveri Conti et al. "A review of AirQ Models and their applications for forecasting the air pollution health outcomes". In : *Environmental Science and Pollution Research* 24.7 (2017), p. 6426-6445.

[124] M Alifa et al. "The influence of meteorology and emissions on the spatio-temporal variability of PM10 in Malaysia". In : *Atmospheric Research* 246 (2020), p. 105107.

[125] Han Han et al. "Local and synoptic meteorological influences on daily variability in summertime surface ozone in eastern China". In : *Atmospheric Chemistry and Physics* 20.1 (2020), p. 203-222.

[126] Xuewei Hou et al. "Inter-annual variability in fine particulate matter pollution over China during 2013–2018 : Role of meteorology". In : *Atmospheric Environment* 214 (2019), p. 116842.

[127] Gongbo Chen et al. "Estimating spatiotemporal distribution of PM1 concentrations in China with satellite remote sensing, meteorology, and land use information". In : *Environmental pollution* 233 (2018), p. 1086-1094.

[128] C Berndt et Uwe Haberlandt. "Spatial interpolation of climate variables in Northern Germany—Influence of temporal resolution and network density". In : *Journal of Hydrology : Regional Studies* 15 (2018), p. 184-202.

[129] 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.

[130] Pritee PARWEKAR, Sireesha RODDA et Neeharika KALLA. "A study of the optimization techniques for wireless sensor networks (WSNs)". In : *Information systems design and intelligent applications*. Springer, 2018, p. 909-915.

[131] Ilhem BOUSSAID, Julien LEPAGNOT et Patrick SIARRY. "A survey on optimization metaheuristics". In : *Information sciences* 237 (2013), p. 82-117.

[132] DE GOLDBERG. *Genetic algorithms in search, machine learning and optimisation*. 1989.

[133] Shao-rong HUANG. "Survey of particle swarm optimization algorithm". In : *Computer Engineering and Design* 8 (2009), p. 39-42.

[134] James KENNEDY et Russell EBERHART. "Particle swarm optimization". In : *Proceedings of ICNN'95-International Conference on Neural Networks*. T. 4. IEEE. 1995, p. 1942-1948.

[135] Sharon MOLTCHANOV et al. "On the feasibility of measuring urban air pollution by wireless distributed sensor networks". In : *Science of The Total Environment* 502 (2015), p. 537-547.

[136] G LANCIA, F RINALDI et P SERAFINI. "A Facility location model for air pollution detection". In : *Mathematical Problems in Engineering* 2018 (2018).

[137] Uri LERNER, Or HIRSHFELD et B FISHBASINL. "Optimal deployment of a heterogeneous air quality sensor network". In : *Journal of Environmental Informatics* 34.2 (2019), p. 99-107.

[138] Linh VAN NGUYEN et al. "Locational optimization based sensor placement for monitoring gaussian processes modeled spatial phenomena". In : *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE. 2013, p. 1706-1711.

[139] Christian SEIGNEUR. *Air Pollution : Concepts, Theory, and Applications*. Cambridge University Press, 2019.

[140] Akula VENKATRAM et al. "The development and application of a simplified ozone modeling system (SOMS)". In : *Atmospheric Environment* 28.22 (1994), p. 3665-3678.

[141] DJ CARRUTHERS, JCR HUNT et WS WENG. "A computational model of stratified turbulent airflow over hills–FLOWSTAR I". In : *Proceedings of ENVIROSOFT : computer techniques in environmental studies, Springer-Verlag* (1988), p. 481-492.

[142] Ole HERTEL, Ruwim BERKOWICZ et Steinar LARSSEN. "The operational street pollution model (OSPM)". In : *Air Pollution Modeling and Its Application VIII*. Springer, 1991, p. 741-750.

[143] Satinder Singh MOHAR, Sonia GOYAL et Ranjit KAUR. "A Survey of Localization in Wireless Sensor Network Using Optimization Techniques". In : *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE. 2018, p. 1-6.

[144] Wu YEE-LIN et al. "Effects of local circulations, turbulent internal boundary layers, and elevated industrial plumes on coastal ozone pollution in the downwind Kaohsiung urban-industrial complex". In : *TAO : Terrestrial, Atmospheric and Oceanic Sciences* 21.2 (2010), p. 3.

[145] Jason CHING et Daewon BYUN. "Introduction to the Models-3 framework and the Community Multiscale Air Quality model (CMAQ)". In : *Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System* (1999).

[146] Jean Philippe LAFORE et al. "The Meso-NH atmospheric simulation system. Part I : Adiabatic formulation and control simulations". In : *Annales geophysicae*. T. 16. 1. Copernicus GmbH. 1998, p. 90-109.

Résumé

La mise en œuvre d'un système de surveillance de la qualité de l'air nécessite la prise en considération de phénomènes météorologiques complexes, de sources d'émission variées et des limites induites par les équipements coûteux. Les trois principales contributions de cette thèse concernant la surveillance et l'estimation de la pollution de l'air sont : une revue des techniques d'estimation de la qualité de l'air, une étude de l'influence de la variabilité spatiale et temporelle de la pollution de l'air sur la précision des méthodes d'interpolation, ainsi qu'une proposition de méthode d'optimisation d'un réseau de surveillance de la qualité de l'air. Les données de mesures et de modélisation de la concentration des particules PM_{10} ont été fournies par ATMO Hauts-de-France. Dans un premier temps, nous avons fait une synthèse bibliographique sur les techniques de modélisation de la qualité de l'air, détaillant leurs avantages et leurs limites dans l'étude de la pollution de l'air. Ensuite, nous avons estimé la pollution de l'air dans la région des Hauts-de-France au moyen de méthodes d'interpolation spatiale. Nous avons ensuite proposé une optimisation de la technique d'interpolation de la pondération à distance inverse (IDW) qui permet d'améliorer le coefficient de détermination (R^2). Le moyennage des données de PM_{10} à des échelles temporelles pertinentes a permis le filtrage de l'influence de ces phénomènes dans l'interpolation. Le meilleur R^2 obtenu correspond à la période de moyennage de 24 heures, similaire à la durée de périodicité de certains phénomènes météorologiques locaux tels que la brise de mer se produisant dans les zones côtières. Par ailleurs, nous proposons une approche pour optimiser le réseau de stations de mesure dans l'agglomération de Dunkerque qui minimise l'erreur quadratique moyenne (RMS) de l'estimation de la pollution atmosphérique obtenue par interpolation IDW à l'aide des données d'ADMS (Atmospheric Dispersion Modeling System) et du modèle de panache gaussien. Il a été démontré que la configuration optimisée permet d'obtenir une meilleure estimation de concentration en PM_{10} par rapport au réseau réel des stations de mesure déployé par ATMO. Enfin, une approche fiable et efficace a été proposée pour améliorer la précision de l'estimation de la pollution atmosphérique dans une zone d'intérêt particulière.

Mots-clés : interpolation, pollution de l'air, météorologie locale, optimisation, réseau de surveillance.

Abstract

The implementation of an air quality monitoring system requires taking in consideration complex meteorological phenomena, sources of emission and limitations drawn by the costly equipment. The three main contributions made by the present thesis regarding monitoring and estimation of air pollution are : a review of techniques for estimating air quality, influence of air pollution's spatial and temporal variability on precision of interpolation methods, and a suggestion for a possible optimization of Dunkirk air quality monitoring network. Data of measurements and modeling of PM_{10} concentrations were provided by ATMO Hauts-de-France. Firstly, we did a bibliographic synthesis on Air Quality Modeling (AQM) techniques, detailing their advantages and limits in studying air pollution. Then, we estimated air pollution in the Hauts-de-France region by means of spatial interpolation methods. We proposed an optimization of Inverse distance Weighting (IDW) interpolation technique that allows improving the coefficient of determination (R^2). The influence of these phenomena was filtered by averaging the PM_{10} data at different time scales (ranging from one hour to 3 months). The best R^2 obtained corresponded to the 24 hours averaging period, similar to the periodicity of some local weather phenomena such as sea breezes occurring in coastal areas. Furthermore, we suggest an approach to optimize the network of measurement stations in Dunkirk agglomeration that minimizes root-mean-square (RMS) error of air pollution estimation obtained by IDW interpolation using data of ADMS (Atmospheric Dispersion Modeling System) and the Gaussian plume model. It was shown that the optimized configuration allows obtaining better PM_{10} concentration estimations compared to the real deployed measuring stations network of ATMO. Finally, a reliable and efficient approach was proposed for improving the accuracy of estimation of air pollution in an area of special interest, such as residential or industrial areas.

Keywords : interpolation, atmospheric pollution, local meteorology, optimization, monitoring network.