

RAPPORT DE STAGE en ENTREPRISE

présenté par

Racha Amina DJAGHLOUL

Master Data science

Mission effectuée du 25 septembre 2023 au 25 septembre 2024

Sujet de la mission :

**Analyse des Données de Microcapteurs pour la
Surveillance de la Qualité de l'Air en Région Sud PACA**

Référent de stage : Monsieur Pierre PUDLO, Enseignant chercheur

Maitre de stage : Monsieur Morgan JACQUINOT, Ingénieur d'études et modélisation

Établissement de formation : Faculté des sciences - Université d'Aix Marseille

Entreprise d'accueil : ATMOSUD, 146 Rue Paradis, 13006 Marseille

Remerciements

Je tiens à exprimer ma profonde gratitude et mes sincères remerciements à toutes les personnes qui m'ont soutenu et aidé tout au long de mon alternance chez AtmoSud.

Je souhaite tout d'abord exprimer ma reconnaissance envers **M.Morgan Jacquinot** pour m'avoir accordé cette opportunité d'effectuer mon alternance au sein de leur entreprise. Son encadrement, ses conseils et son soutien constant ont grandement contribué à enrichir mon expérience professionnelle.

Ma reconnaissance s'adresse plus particulièrement à **M. Alexis STEPANIAN**, responsable du pôle Appui Technique à AtmoSud, pour avoir été mon référent tout au long de l'alternance.

Je tiens également à remercier chaleureusement toute l'équipe de "Appui Technique- AtmoSud" pour leur accueil chaleureux, leur disponibilité et leur collaboration.

Je tiens à remercier **Monsieur Pierre Pudlo**, Enseignant chercheur de l'Université d'Aix Marseille et **Monsieur Grégory MAILLARD**, Professeur à l'Université d'Aix-Marseille et responsable du Master de Data science, pour m'avoir accompagné.

Mes remerciements vont également à mes collègues de travail, avec qui j'ai eu le plaisir de collaborer tout au long de mon alternance. Leur convivialité, leur esprit d'équipe et leurs encouragements ont créé une ambiance de travail agréable et stimulante.

Enfin, je tiens à exprimer ma gratitude envers mes parents, ma famille et mes amis pour leur soutien inconditionnel, leurs encouragements et leur compréhension durant cette période d'alternance.

Table des matières

Remerciements	1
Introduction	4
Cadre général	5
Présentation de l'association AtmoSud	5
Les particules fines (PM ₁₀ ,PM _{2,5})	7
1 Extraction des données	10
1.1 Microcapteurs	10
1.2 La station de mesure	12
1.3 La différence entre les deux moyens de mesure	13
1.4 Répartition des sites de mesure de l'étude	13
1.5 Méthodologie d'extraction des données	14
1.5.1 Extraction via API	14
1.5.2 Extraction via des requêtes SQL sur R	15
2 Correction des données des microcapteurs	16
2.1 Méthodologie de correction	16
2.1.1 Objectif	16
2.1.2 Approches de Machine Learning pour la correction initiale	17
2.1.3 Correction en temps réel	19
2.2 Analyse des résultats des corrections	20
2.2.1 Première correction	20
2.2.2 Evaluation d'une autre période d'entraînement des modèles	27
2.2.3 Synthèse des résultats de la deuxième correction - en temps réel	28
2.3 Discussion : Exploration des données	37
3 Méthodes d'Interpolation pour la Cartographie en Temps Réel	39
3.1 Interpolation avec Inverse Distance Weighting (IDW)	39
3.2 Méthode d'Interpolation 4 : K-Nearest Neighbors (kNN)	42
3.2.1 Principe de Fonctionnement	42
3.2.2 Résultats des Simulations	43

3.3	Interpolation avec la Fonction <code>interp</code>	46
3.4	Comparaison des Méthodes	49
3.4.1	Analyse Comparative des Résultats	49
3.4.2	Évaluation des Avantages et Inconvénients	49
3.5	Conclusion	49
Conclusion		51
Annexe A		52
Annexe B -		52
Annexe C - Résultats sur l'interface <code>pspot</code>		53
Annexe D - Bulle d'air		54

Introduction

Dans le cadre de mon alternance de Master 2 chez AtmoSud, j'ai eu l'opportunité de travailler sur un projet innovant visant à déployer plusieurs microcapteurs autonomes pour la surveillance de la qualité de l'air dans la région Sud Provence-Alpes-Côte d'Azur (PACA). Ces dispositifs sont installés en divers points du territoire dans le cadre de plusieurs projets menés par AtmoSud. Le nombre de ces capteurs est prévu d'augmenter significativement dans les prochaines années grâce aux avancées technologiques, à leur accessibilité croissante et à l'intérêt grandissant des citoyens pour la surveillance de leur environnement immédiat.

AtmoSud s'engage dans la science participative pour soutenir ces efforts et améliorer son expertise en développant des méthodes d'analyse pour gérer les gros volumes de données générés par ces dispositifs de mesure en continu. Mon alternance avait pour principaux objectifs de :

1. Mettre en place des outils de modélisation numérique pour optimiser la correction des données issues des microcapteurs, notamment par :
 - L'optimisation des méthodes utilisées pour l'Assurance Qualité / Contrôle Qualité (QA/QC) des microcapteurs ;
 - La prise en compte de co-variables environnementales pour expliquer les données générées par les capteurs ;
 - Le déploiement d'outils de Machine Learning pour améliorer l'évaluation des données.
2. Participer à l'affichage des cartographies afin d'améliorer les outils de prévision de la qualité de l'air, générées quotidiennement à l'échelle horaire, et traduites en cartes pour l'ensemble de la région Sud PACA avec une résolution de 25 mètres.

Dans ce rapport, je vais présenter le contexte du projet, les travaux que j'ai réalisés, ainsi que les principales contributions que j'ai apportées. Je vais également discuter de la méthodologie que j'ai utilisée, des résultats obtenus et des perspectives d'amélioration pour l'avenir. Enfin, je conclurai en soulignant l'importance de ces méthodes pour la réalisation de cartes précises de la qualité de l'air, essentielles pour protéger notre environnement et la santé publique.

Cadre général

Dans le cadre de la préparation de mon projet de Master 2 en data science, j'ai réalisé une alternance de 12 mois au sein d'AtmoSud à Marseille. Elle s'est déroulée du 25 septembre 2023 au 27 septembre 2024.

L'objectif est d'effectuer une correction sur les données collectées par des microcapteurs pour la surveillance de la qualité de l'air. Ce travail vise à développer et appliquer des méthodes de correction pour éliminer les erreurs et biais des données, assurant ainsi leur fiabilité pour l'analyse environnementale. La méthodologie comprendra l'analyse des données brutes, le développement de méthodes de correction adaptées, leur application aux données de microcapteurs, et l'analyse des résultats.

Après avoir corrigé les données brutes recueillies par une dizaine de microcapteurs en les ajustant avec les mesures issues de stations fixes, l'étape suivante a consisté à cartographier les résultats. Cette cartographie a permis d'obtenir une vue globale et précise de la qualité de l'air. La création de cartes horaires détaillées s'est avérée indispensable pour une gestion optimale des polluants atmosphériques, en particulier des particules fines **PM_{2.5}**, dont les effets nocifs sur la santé sont bien établis.

Présentation de l'association AtmoSud

Le système de surveillance de la qualité de l'air

L'État a confié à des organismes agréés par le ministère la mission de surveiller la qualité de l'air en France. L'ensemble de ces associations forme la fédération nationale Atmo France, et chaque région française dispose d'un observatoire. La surveillance est effectuée sur la base de la loi sur l'Air et l'Utilisation rationnelle de l'Énergie du 30 décembre 1996 (LAURE).

LAURE : "L'objectif est la mise en œuvre du droit reconnu à chacun de respirer un air qui ne nuise pas à sa santé. Cette action d'intérêt général consiste à prévenir, à surveiller, à réduire ou à supprimer les pollutions atmosphériques, à préserver la qualité de l'air et, à ces fins, à économiser et à utiliser rationnellement l'énergie."

Loi n° 96-1236 du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie.

AtmoSud est l'observatoire de la qualité de l'air de la région Sud Provence-Alpes-Côte d'Azur, qui voit le jour en 1972. Elle est membre de la Fédération française ATMO, qui regroupe toutes les AASQA du pays.

Historique d'Atmosud

AtmoSud (le nouveau nom d'Air PACA) est né le 10 janvier 2012 de la fusion des associations Atmo PACA et AIRFOBEP. Ce regroupement, ou application de la loi Grenelle 2, préserve le patrimoine de nos monuments historiques et permet un pool d'outils et de savoir-faire pour répondre aux multiples enjeux de notre territoire. Le [tableau 1](#) ci-dessous présente l'historique de l'association.

**Fig. 1** – Zoom sur la Fig

La carte présentée ci-contre montre l'ensemble des AASQA qui composent le réseau ATMO France. La fédération compte 18 AASQA, 1 par région administrative de métropole et d'outre-mer. AtmoSud possède 3 antennes : Marseille (siège social), Martigues et Nice. Les territoires d'action ainsi que les différentes missions qui animent la structure sont répartis entre ces 3 antennes. Ces antennes travaillent conjointement à l'amélioration du réseau, des techniques, de la recherche et des connaissances

Tab. 1 – Historique d'AtmoSud

Année	Association	Description
1972	AIRFOBEP	Association chargée de surveiller la qualité de l'air de l'Ouest des Bouches-du-Rhône
1982	AIRMARAIX	Association chargée de surveiller la qualité de l'air de l'Est des Bouches-du-Rhône, du Var et du Vaucluse
1989	Qualit'Air	Association chargée de surveiller la qualité de l'air des Alpes-de-Haute-Provence, des Hautes-Alpes et des Alpes-Maritimes
2006	Atmo PACA	Fusion de Qualit'Air et AIRMARAIX, association chargée de surveiller la qualité de l'air des Alpes-de-Haute-Provence, des Hautes-Alpes, des Alpes-Maritimes, de l'Est des Bouches-du-Rhône, du Var et du Vaucluse
2012	Air PACA	Fusion d'AIRFOBEP et Atmo PACA, l'observatoire de la qualité de l'air de la région PACA
2018	AtmoSud	Évolution du nom Air PACA en AtmoSud

Missions et moyens

Moyens et acteurs

Afin de remplir sa mission, AtmoSud compte une soixantaine de salariés, et dispose d'un budget annuel de plus d'un million d'euros. Selon le site d'Atmosud ([1](#)).

L'impartialité de ses actions est assurée par la composition quadripartite de son assemblée générale regroupant quatre collèges :

- Collectivités territoriales
- Services de l'État et établissements publics
- Industriels
- Associations de protection de l'environnement et de consommateurs, des personnalités qualifiées et/ou professionnels de la santé

Cette collégialité équilibrée lui confère une certaine indépendance.

Les missions principales des AASQA

AtmoSud est un acteur majeur dans le domaine de la protection de l'environnement. Elle exerce ses activités à l'échelle nationale/internationale. AtmoSud s'est forgé une solide réputation en tant qu'expert en surveillance et amélioration de la qualité de l'air,

Les deux piliers d'AtmoSud sont :

1. **Garantir un observatoire de référence :** Mise en place de surveillance régulière (mesures, émissions, modélisation), traitement des données, diffusion d'informations (alertes, bulletins quotidiens, indice Atmo), et service d'intervention post-accidentel (QAPA, circulaire Lubrizol).
2. **Favoriser l'engagement :** AtmoSud surveille et informe pour préserver la qualité de l'air et lutter contre le dérèglement climatique. Avec son observatoire et l'engagement des acteurs territoriaux, elle inspire un air meilleur. En accompagnant les acteurs dans l'objectivation de la situation, la planification et le suivi des activités du territoire, AtmoSud informe, sensibilise et forme à la qualité de l'air. Elle innove également par des actions de recherche et développement dans le cadre de projet de recherche appliquée et de projet Européen, tout en jouant un rôle d'expert et d'intermédiation, avec un statut collégial, transparent et indépendant.

Missions principales confiées par l'État à AtmoSud

L'État confie à chaque AASQA, dans sa région de compétence, les missions suivantes :

- **La surveillance de la qualité de l'air :** En réponse aux contrats de surveillance de la qualité de l'air, trois champs de compétences sont nécessaires pour AtmoSud :
 - La mesure (terrain) incluant développement d'un réseau de mesure avec capteurs, enregistrement des données en temps réel
 - La modélisation à l'aide de modèles à différentes échelles spatiales et temporelles dans le but de prédire les concentrations de polluants
 - Les inventaires visent à déterminer les émissions de tous les polluants atmosphériques et gaz à effet de serre sur une période et une zone bien définies.

- **La diffusion des résultats et des prévisions :** La diffusion des résultats et des prévisions peut se traduire par l'information et par la sensibilisation du public.

AtmoSud dispose également d'un outil qui répertorie des nuisances tels que l'odeur, le bruit et même la fumée peuvent causer de l'inconfort aux personnes

“AtmoSud surveille, évalue l'exposition des populations à la pollution pour permettre à chacun d'agir et informer/alérer la population, les acteurs, les décideurs, les autorités,” a déclaré le président de l'association AtmoSud, *M. Jean Pierre Maria*.

Les particules fines (**PM₁₀** ,**PM_{2,5}**)

“La pollution atmosphérique par les particules raccourcit la vie dans le monde, encore plus que les cigarettes. Il n'y a pas de plus grand risque actuel pour la santé humaine.”

*Michael Greenstone,
Milton Friedman Distinguished Service Professor in Economics, University of Chicago*

Origine et dynamique

Les particules fines (PM_{10} et $PM_{2.5}$) sont des polluants atmosphériques provenant principalement des activités humaines telles que le trafic routier, les activités industrielles, le chauffage domestique, et les feux de biomasse. Elles peuvent également avoir des sources naturelles comme les tempêtes de sable, les feux de forêt et les éruptions volcaniques.

Les PM_{10} (particules dont le diamètre est inférieur à 10 micromètres) et les $PM_{2.5}$ (particules dont le diamètre est inférieur à 2.5 micromètres) diffèrent par leur taille, ce qui influence leur comportement dans l'atmosphère et leurs effets sur la santé. Les PM_{10} peuvent rester en suspension dans l'air pendant des heures voire des jours et se déplacer sur de longues distances, tandis que les $PM_{2.5}$ peuvent rester en suspension pendant des semaines.

Dans la [figure 2](#), vous pouvez observer la taille relative des particules fines par rapport à un cheveu humain.

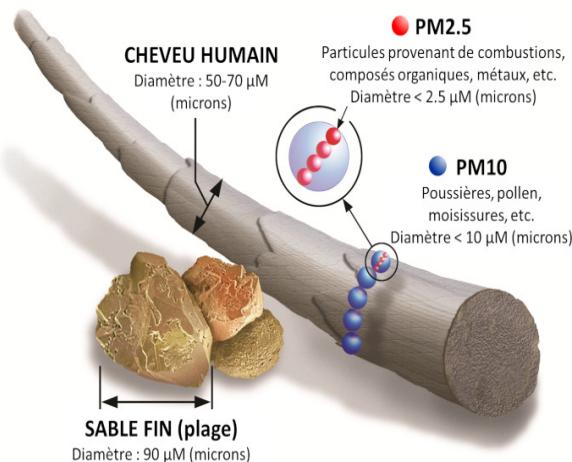


Fig. 2 – Comparaison de diamètres entre un cheveu, un grain de sable et des particules $PM_{2.5}$ et PM_{10} . [Source : <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>]

Sources et impact des PM_{10} et $PM_{2.5}$

Les principales sources de PM_{10} et $PM_{2.5}$ comprennent les émissions des véhicules, les installations industrielles, le chauffage résidentiel, les activités agricoles, et la combustion de biomasse. Les particules fines peuvent également se former par des réactions chimiques complexes impliquant des précurseurs tels que le dioxyde de soufre (SO_2), les oxydes d'azote (NO_x), l'ammoniac (NH_3) et les composés organiques volatils (COV).

L'impact des PM_{10} et $PM_{2.5}$ sur la santé

L'exposition aux particules fines, en particulier les $PM_{2.5}$, est associée à divers effets néfastes sur la santé. Les particules fines peuvent pénétrer profondément dans les poumons et

même entrer dans la circulation sanguine, ce qui peut provoquer des problèmes respiratoires et cardiovasculaires.

Une exposition à court terme aux particules fines peut entraîner des irritations des yeux, du nez, de la gorge et des poumons, ainsi qu'une aggravation des maladies respiratoires telles que l'asthme et la bronchite. À long terme, une exposition prolongée peut provoquer des maladies chroniques comme le cancer du poumon, les maladies cardiaques, et une diminution de la fonction pulmonaire.

L'impact des PM₁₀ et PM_{2.5} sur l'environnement

Les particules fines affectent également l'environnement de plusieurs manières :

- Réduction de la visibilité : Les particules fines peuvent provoquer un voile atmosphérique qui réduit la visibilité, un phénomène souvent observé sous forme de smog.
- Dépôts acides : Les particules acides peuvent se déposer sur le sol et l'eau, contribuant ainsi à l'acidification des sols et des cours d'eau, ce qui peut nuire aux écosystèmes aquatiques et terrestres.
- Impact sur la végétation : Les particules peuvent se déposer sur les feuilles des plantes, réduisant leur capacité à photosynthétiser et affectant leur croissance.

Les particules fines, en interagissant avec d'autres polluants atmosphériques, contribuent également à la formation d'ozone troposphérique et au réchauffement climatique.

Valeurs de référence

Les seuils réglementaires pour le dioxyde d'azote dans l'air sont définis par (3) comme ceci :

	Seuil de recommandation et d'information	Seuil d'alerte	Valeur limite
PM _{2.5}	25 µg/m ³ en moyenne annuelle	50 µg/m ³ en moyenne annuelle	Moyenne annuelle : 25 µg/m³
PM ₁₀	20 µg/m ³ en moyenne annuelle	50 µg/m ³ en moyenne annuelle	Moyenne annuelle : 40 µg/m³

Chapitre 1

Extraction des données

Ce chapitre se concentre sur l'extraction et le prétraitement des données de PM_{2.5} et PM₁₀ issues de deux sources principales : les stations de mesure via la base de données 'MESMOD' et les microcapteurs via l'API "µspot" d'AtmoSud. L'objectif est de garantir la qualité et la cohérence des données pour les analyses ultérieures, dans le cadre de la surveillance de la qualité de l'air. Nous commencerons par explorer les caractéristiques et rôles des microcapteurs et des stations de mesure, puis nous détaillerons les méthodes d'extraction des données, en utilisant des API pour les microcapteurs et des requêtes SQL pour les stations. Cette démarche est essentielle pour assurer une intégration précise des données provenant de diverses sources.

1.1 Microcapteurs

Définition des microcapteurs

Un microcapteur est un dispositif miniaturisé conçu pour mesurer ou représenter des composés présents dans l'air. Ces capteurs utilisent des phénomènes physiques, chimiques ou biologiques pour transformer une grandeur physico-chimique en un signal électrique. Ils sont adaptés à différents environnements, qu'il s'agisse de l'air intérieur, de l'air extérieur ou de la mobilité.



Fig. 1.1 – Un microcapteur installé à l’extérieur loin de la station.

Applications et avantages des microcapteurs

Bien que la technologie des microcapteurs soit encore émergente et que leur fiabilité ne soit pas toujours garantie, ils offrent des avantages significatifs :

- **Accessibilité et coût réduit** : Les microcapteurs sont souvent moins coûteux que les stations de mesure traditionnelles, ce qui permet une surveillance plus étendue et accessible de la qualité de l’air.
- **Portabilité et flexibilité** : Grâce à leur petite taille, les microcapteurs peuvent être facilement installés dans diverses infrastructures, y compris les bâtiments, les véhicules, ou être portés par des individus.
- **Sensibilisation du public** : Les microcapteurs permettent de sensibiliser davantage le public aux problèmes de qualité de l’air en fournissant des données accessibles en temps réel.

Limites et défis des microcapteurs

Malgré leurs nombreux avantages, les microcapteurs présentent également des limitations :

- **Fiabilité et précision** : Les microcapteurs peuvent présenter des biais importants par rapport aux mesures de référence, surtout en fonction des milieux où ils sont utilisés.
- **Calibration nécessaire** : Pour garantir des données fiables, les microcapteurs doivent être régulièrement calibrés, souvent en utilisant des stations de mesure de référence.
- **Durabilité** : La durabilité et la stabilité des microcapteurs peuvent varier, nécessitant des remplacements ou des ajustements fréquents.

1.2 La station de mesure

Contrairement aux microcapteurs, une station de mesure est un laboratoire équipé de matériel de référence pour mesurer avec haute précision différents polluants au même endroit, qu'ils soient réglementés ou non. Elles suivent des protocoles stricts de maintenance et de calibration pour garantir la qualité des données produites.

Les stations de mesure sont utilisées pour tester et calibrer les microcapteurs. Par exemple, les stations d'AtmoSud sont régulièrement utilisées pour vérifier la précision des microcapteurs avant leur déploiement sur le terrain. Elles fournissent des données fiables et sont essentielles pour les études environnementales, la recherche scientifique et les politiques de gestion de la qualité de l'air.



Fig. 1.2 – Une station fixe avec les microcapteurs à l'extérieur.

1.3 La différence entre les deux moyens de mesure

- **Taille et portabilité :** Les microcapteurs sont compacts et portables, tandis que les stations de mesure sont de grands laboratoires fixes.
- **Coût :** Les microcapteurs sont généralement moins coûteux, facilitant leur déploiement à grande échelle.
- **Précision :** Les stations de mesure offrent une précision supérieure grâce à des équipements calibrés et des protocoles stricts.
- **Applications :** Les microcapteurs sont idéaux pour des mesures locales et en temps réel, tandis que les stations de mesure sont utilisées pour des analyses détaillées et de longue durée.

En résumé, les microcapteurs et les stations de mesure jouent des rôles complémentaires dans la surveillance de la qualité de l'air. Les microcapteurs offrent flexibilité et accessibilité, tandis que les stations de mesure garantissent une haute précision et une fiabilité des données, nécessaire pour la validation et la calibration des microcapteurs.

1.4 Répartition des sites de mesure de l'étude

Nous avons installé un total de 10 microcapteurs devant différentes stations de référence pour évaluer la performance et la transférabilité des méthodes de correction entre les différents sites. Les stations de l'étude sont les suivantes :

Tab. 1.1 – Répartition des sites de mesure

Ville	Fond	Trafic
Aix	Aix Art	Aix Roy René
Nice	Nice Arson	Nice Magnan
Marseille	Marseille Cinq	Marseille Kaddouz
Toulon	Toulon Claret	Toulon Foch
Marignane	X	
Salon	X	

Une carte ci-dessous illustre la répartition des microcapteurs installés devant des stations de référence dans la région PACA. Cette visualisation géographique permet de mieux comprendre la répartition spatiale des sites de mesure et leur proximité avec les stations de référence.

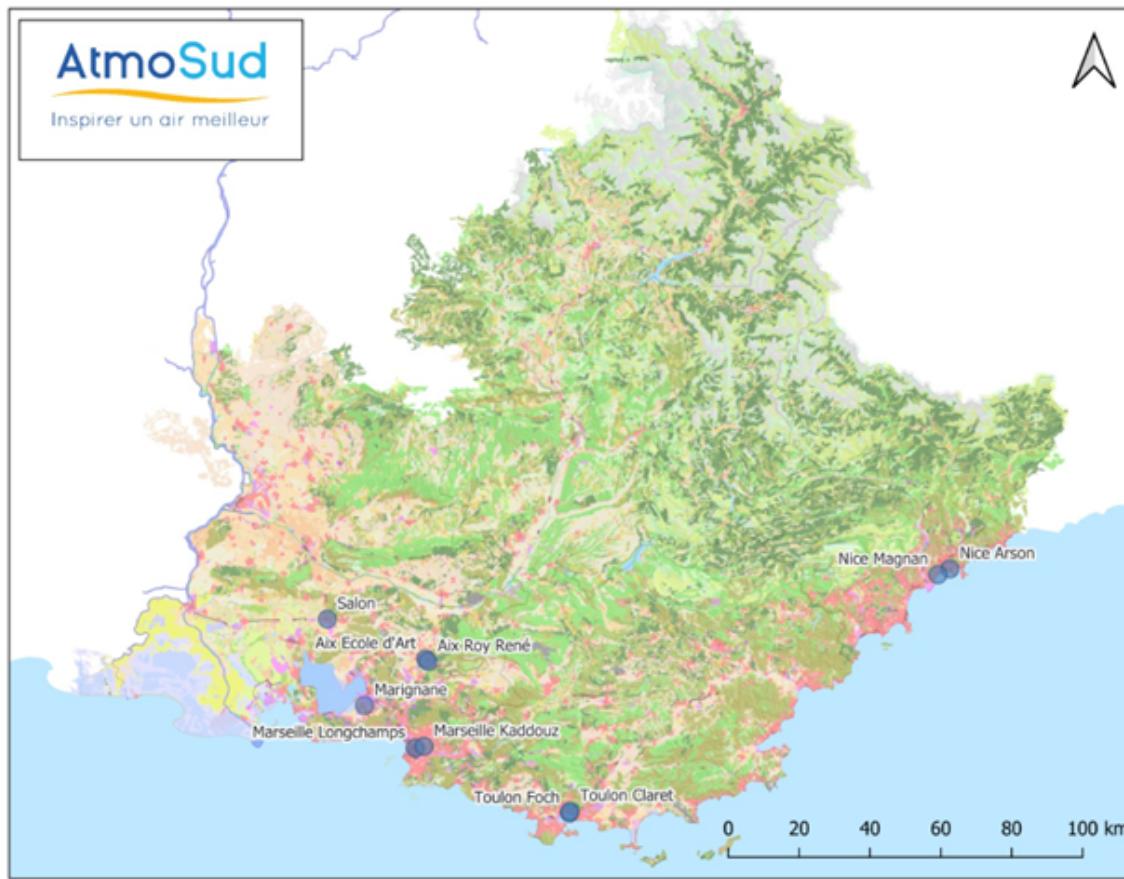


Fig. 1.3 – Localisation des microcapteurs de l'étude à coté des stations fixes.

1.5 Méthodologie d'extraction des données

Pour extraire les données de qualité de l'air relatives aux particules fines $\text{PM}_{2.5}$ et PM_{10} , deux méthodes principales ont été utilisées : l'extraction via API et l'extraction via des requêtes SQL sur R.

1.5.1 Extraction via API

Qu'est qu'un API ?

Une API (Application Programming Interface) est un ensemble de règles et de protocoles qui permettent à différentes applications de communiquer entre elles. En d'autres termes, une API définit les méthodes et les formats que les applications peuvent utiliser pour demander et échanger des données ou des services. Les API sont essentielles pour intégrer diverses applications et services, permettant ainsi aux développeurs d'accéder à des fonctionnalités et des données sans avoir à les recréer.

Les principales caractéristiques des API incluent :

- **Standardisation** : Les API utilisent des standards pour garantir que les différents systèmes puissent comprendre et traiter les requêtes et les réponses.
- **Modularité** : Les API permettent d'ajouter de nouvelles fonctionnalités à une application existante sans modifier son code de base.

- **Sécurité** : Les API peuvent contrôler l'accès aux données et aux services, offrant ainsi une couche de sécurité.

Comment utiliser une API dans RStudio ?

L'extraction des données des microcapteurs via l'API a été réalisée en utilisant RStudio et des requêtes HTTP. Elle s'effectue en envoyant une requête POST à l'URL spécifiée avec les paramètres appropriés. L'URL utilisée est : "https://spot.atmo-france.org/export-api/observations". Cette requête nécessite également des en-têtes spécifiques pour l'authentification et le type de contenu.

Voici les étapes détaillées du processus, ainsi que le code utilisé :

1. **Installation des Packages Nécessaires** Avant de commencer, il est important d'installer les packages R nécessaires pour travailler avec les API. Les packages couramment utilisés incluent `httr` et `jsonlite`.

```
install.packages("httr")
install.packages("jsonlite")
```

2. **Définition des paramètres de l'API** : Les dates de début et de fin, ainsi que les codes des types d'observations, ont été spécifiés.
3. **Envoi de la requête API pour les ID internes des capteurs** : Une requête HTTP POST a été envoyée pour récupérer les ID des capteurs.
4. **Extraction des mesures des capteurs** : Une deuxième requête POST a été envoyée pour récupérer les mesures des capteurs sur la période spécifiée.
5. **Mise en forme des données** : Si la requête est réussie, nous devons convertir les données JSON reçues en un format R utilisable, comme un data frame.

```
# Convertir la réponse JSON en data frame
data <- fromJSON(content(response, "text"))
```

1.5.2 Extraction via des requêtes SQL sur R

L'extraction des données des stations de surveillance via des requêtes SQL a été réalisée en utilisant RStudio. Voici les étapes détaillées du processus, ainsi que le code utilisé :

1. **Connexion à la base de données** : En utilisant la fonction `dbConnect()` du package `RMySQL`, nous avons établi une connexion à la base de données MySQL.
2. **Définition des paramètres** : Les paramètres de la requête, tels que l'hôte, l'utilisateur, le mot de passe, le nom de la base de données, les stations sélectionnées, et les dates de début et de fin, ont été définis.
3. **Exécution de la requête SQL** : Une requête SQL a été exécutée pour extraire les données correspondant aux stations et à la période spécifiée.
4. **Mise en forme des données** : Les données extraites ont été transformées et mises en forme pour être prêtes pour le prétraitement.

Chapitre 2

Correction des données des microcapteurs

Depuis octobre 2023, Atmosud a installé plusieurs microcapteurs dans différents sites de la région PACA (Marseille, Aix, Nice, Toulon, etc.). Chacun de ces microcapteurs est situé à proximité d'une station de référence pour permettre une comparaison précise des données. L'objectif principal de cette installation est d'évaluer et de corriger les mesures des microcapteurs afin de les rapprocher des mesures de référence obtenues par les stations fixes.

2.1 Méthodologie de correction

Pour assurer la fiabilité des données, nous avons défini une période d'entraînement QA/QC (Quality Assurance/Quality Control) durant laquelle les microcapteurs ont été comparés aux stations de référence de manière intensive, avec une installation côte à côte pendant 13 jours pour collecter des données horaires précises. Nous développons et sélectionnons le meilleur modèle de correction à partir de ces données d'entraînement en utilisant des techniques de machine learning, puis appliquons ce modèle pour corriger les données. Enfin, nous effectuons une seconde correction en temps réel sur les données corrigées pour garantir une méthode de correction transférable entre les sites.

2.1.1 Objectif

L'objectif principal est de corriger les données horaires en deux étapes successives :

- **Correction 1 :** Pendant la période QA/QC, nous utilisons les données collectées pour développer et ajuster des modèles de correction. Ces modèles incluent la régression simple, la régression multiple prenant en compte l'humidité relative, et les Support Vector Machines (SVM). La proximité constante des microcapteurs avec les stations de référence durant cette période permet de créer des modèles de correction robustes et précis.
- **Correction 2 :** Une fois la période QA/QC terminée, nous appliquons les modèles de correction développés en temps réel sur les données des microcapteurs. Cette correction en temps réel est possible grâce à la conservation du positionnement des microcapteurs à proximité des stations, ce qui garantit une validation et une mise à jour continue des données corrigées.

Cette méthodologie permet de maintenir une précision optimale des mesures des microcapteurs en assurant que les corrections soient ajustées et appliquées en temps réel, en fonction des données obtenues des stations fixes.

2.1.2 Approches de Machine Learning pour la correction initiale

La régression simple avec exposant

La première méthode utilisée est une régression simple de la forme :

$$y = a \cdot x^{\text{exposant}} + b \quad (2.1)$$

Où :

- x représente les données collectées par les microcapteurs,
- y représente les données de référence collectées par la station fixe,
- a et b sont des coefficients déterminés par la régression,
- **exposant** est un paramètre déterminé pour mieux ajuster le modèle aux données.

Méthodologie

Nous avons implémenté la méthodologie suivante pour ajuster et sélectionner les meilleurs modèles de régression simple :

1. Initialisation et préparation des données :
 - Nettoyage des données pour éliminer les valeurs aberrantes et les valeurs manquantes.
 - Synchronisation temporelle des séries des microcapteurs et des stations fixes.
2. Boucle de calcul des modèles :
 - Pour chaque paire de station et microcapteur :
 - Itération sur différents exposants de régression (de 0.5 à 1.3 par pas de 0.1).
 - Filtrage des données en fonction d'un seuil de validation.
 - Ajustement d'un modèle de régression simple
3. Sélection du meilleur modèle :
 - Pour chaque paire station-microcapteur, sélection du modèle avec le RMSE (Root Mean Squared Error) le plus bas comme meilleur modèle.
4. Application du modèle :
 - Application du modèle sélectionné à toutes les données pour obtenir les mesures corrigées du microcapteur.
5. Évaluation des modèles :
 - Calcul du RMSE brut (avant correction) et du RMSE après la première correction pour chaque modèle sélectionné.
 - Calcul du pourcentage d'amélioration du RMSE pour évaluer l'efficacité de la correction.

La régression linéaire multiple

Pour améliorer l'ajustement, nous avons introduit l'humidité relative comme variable explicative supplémentaire dans un modèle de régression multiple de la forme $y = ax + bz + c$

$$y = a \cdot x + b \cdot z + c \quad (2.2)$$

Où :

- x représente les données des microcapteurs,
- z représente l'humidité relative,
- y représente les données de la station fixe,
- a , b , et c sont des coefficients déterminés par la régression.

Méthodologie

1. Collecte des données supplémentaires :

- *Données d'humidité relative* : Intégration des mesures d'humidité relative aux données de concentration de PM_{2.5} et PM₁₀.
- *Synchronisation temporelle* : Assurer que toutes les données sont synchronisées temporellement.

2. Ajustement du modèle :

- Utilisation de la méthode des moindres carrés pour estimer les paramètres a , b , et c .

3. Évaluation du modèle :

- Calcul du coefficient de détermination (R^2).
- Calcul de l'erreur quadratique moyenne (MSE).
- Comparaison des performances avec celles des autres modèles.

Support Vector Machines (SVM)

Pour explorer des méthodes plus avancées, nous avons appliqué un modèle de Support Vector Machines (SVM) avec l'humidité relative comme variable supplémentaire. Les SVM sont connus pour leur capacité à capturer des relations non linéaires entre les variables.

Méthodologie

1. Sélection du modèle :

- *Filtrage des données* : Filtrage des données pour éliminer les valeurs manquantes et aberrantes, ainsi que les valeurs des capteurs dépassant un seuil défini.
- *Entraînement* : Les données de concentration de PM_{2.5}, PM₁₀ et d'humidité relative sont utilisées pour entraîner un modèle de Support Vector Regression (SVR).

2. Ajustement du modèle :

- *Modèle SVR* : Application de la régression epsilon-insensitive pour ajuster le modèle et minimiser les erreurs significatives entre les prédictions et les valeurs réelles.

3. Évaluation du modèle :

- *Test sur un ensemble de données supplémentaires* : Calcul du RMSE pour évaluer les performances du modèle sur des données non vues durant l'entraînement.
- *Amélioration du modèle* : Mesure de l'amélioration par rapport aux données brutes.

2.1.3 Correction en temps réel

L'objectif est de pouvoir corriger au fil de l'eau les mesures des microcapteurs à partir d'un couple microcapteur/station de référence. Il est donc nécessaire d'évaluer si cette correction basée sur une comparaison d'un couple microcapteur/station est applicable sur les microcapteurs déployés.

Méthodologie

- Calculer le ratio moyen sur 3h, 6h, 12h et 24 heures entre les mesures du couple microcapteur/station de référence.
- Appliquer ce ratio aux données des autres microcapteurs déployés.
- Les calculs suivants sont effectués :
 - **rmse_ratio (3h, 6h, 12h et 24h)** : À chaque itération, un couple microcapteur/station est choisi comme couple de référence pour le calcul du ratio (ou (**3h, 6h, 12h, 24h**)), qui sera appliqué par la suite sur les données des autres microcapteurs (à côté d'autres stations).
 - Le RMSE est calculé entre les données avant l'application du ratio (c-à-dire avec corr1) et après l'application des différents ratios (avec corr1 + corr2 avec ratio_12h, avec corr1 + corr2 avec ratio_24h...etc).
 - La formule du calcul de l'amélioration est :

$$\text{amélioration} = \frac{\text{rmse_cor1} - \text{rmse_cor2}}{\text{rmse_cor1}} \times 100$$

- La formule de l'incertitude élargie est :

$$\text{Incertitude_élargie} = \frac{\text{Écart_type}}{\text{seuil_horraire}}$$

avec seuil horaire = $25 \mu\text{g}/\text{m}^3$ pour les PM_{2.5}.

- Cette valeur représente la plage autour de la valeur mesurée où l'on s'attend à ce que la vraie valeur se trouve avec le niveau de confiance choisi.
- L'incertitude élargie est une estimation de l'intervalle dans lequel la vraie valeur est probablement contenue. Par exemple, si on mesure une valeur en PM_{2.5} de $15 \mu\text{g}/\text{m}^3$ avec une incertitude étendue de $\pm 2 \mu\text{g}/\text{m}^3$ (à k=2), cela signifie qu'on peut être à peu près sûr (à environ 95% de confiance) que la vraie mesure est entre 13 et $17 \mu\text{g}/\text{m}^3$.

Remarque : La 2ème correction est appliquée sur les données corrigées avec la correction 1.

2.2 Analyse des résultats des corrections

2.2.1 Première correction

Coefficients des modèles de la régression avec exposant sur le polluant PM_{2,5}

Station	Capteur	Ville	a	b	exposant
2041	BD6BAF	Salon	6,31	-6,22	0,5
2043	BD6AA4	Marignane	0,93	1,16	1
3029	BD6BAA	Aix_art	0,89	0,82	1
3043	BD6ABC	Marseille_cinq_av	0,48	2,06	1,1
3071	BD6B8B	Toulon_claret	0,76	0,51	1,1
24035	C19D41	Nice_magnan	5,27	-4,45	0,5
24036	C19D40	Nice_arson	3,43	-3,07	0,7

Tab. 2.1 – Tableau des coefficients de régression avec exposant pour différentes villes et capteurs.

Coefficients des modèles de la régression avec l'humidité relative sur le polluant PM_{2,5}

L'introduction de l'humidité relative a amélioré légèrement la précision du modèle. Les résultats montrent une réduction légère de l'erreur quadratique moyenne pour certains des microcapteurs, indiquant que le modèle explique mieux la variance des données de la station fixe.

Ville	Station	Capteur	a_{multi}	b_{multi}	c_{multi}
Salon	2041	BD6BAF	0,83	-0,01	4,1
Marignane	2043	BD6AA4	0,93	-0,01	2,18
Aix Art	3029	BD6BAA	0,89	-0,03	3,35
Marseille 5 Av	3043	BD6ABC	0,75	-0,05	5,04
Toulon Claret	3071	BD6B8B	1,12	-0,02	1,05
Nice Magnan	24035	C19D41	0,89	0,08	-3,5
Nice Arson	24036	C19D40	0,94	-0,04	6,62

Tab. 2.2 – Tableau des coefficients de régression multiple pour différentes villes et capteurs.

Graphiques des résultats

Les graphiques suivants illustrent les résultats des corrections apportées par les modèles sur les mesures des capteurs par rapport aux références. Chaque graphique représente une station de mesure spécifique, montrant les mesures brutes en gris, les corrections appliquées sauf pour la période QA/QC en rouge, et les corrections spécifiquement pour la période QA/QC en bleu. Les lignes vertes représentent les seuils de qualité de l'air. Les informations sur les performances des modèles, telles que le RMSE avant et après la correction, sont également fournies en bas de chaque graphique.

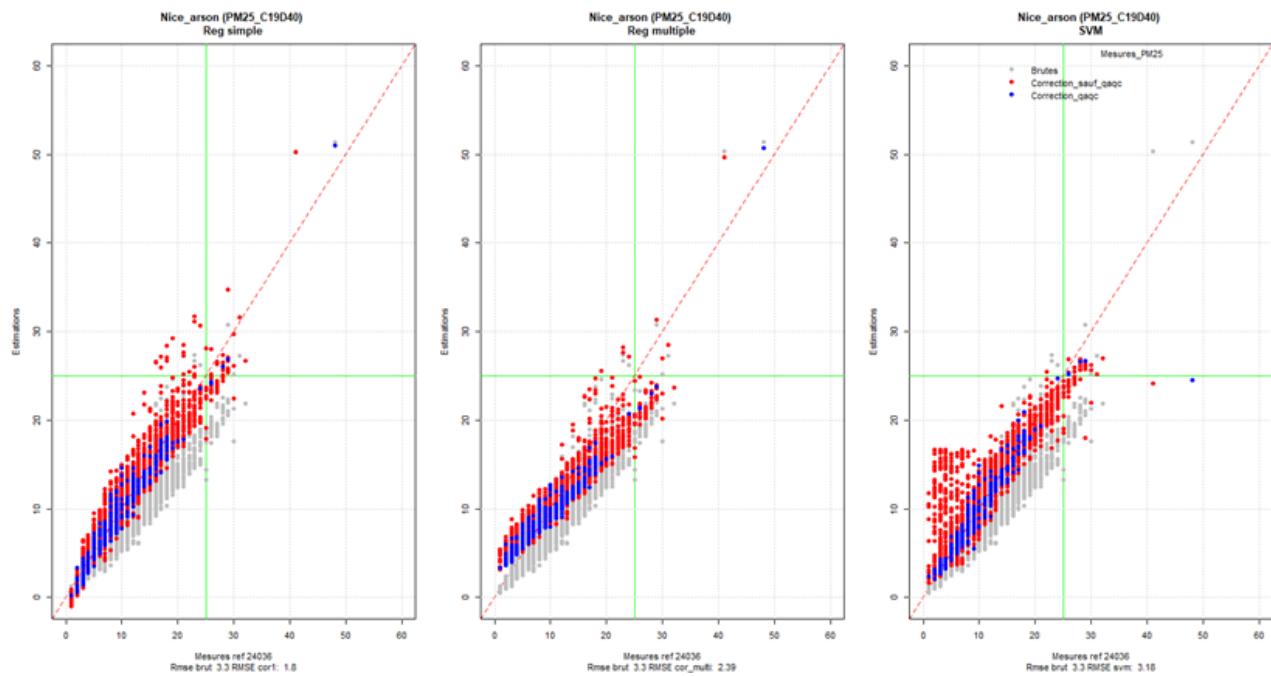


Fig. 2.1 – Données corrigées et brutes des microcapteurs vs les mesures de la station de Nice Arson en $\mu\text{g}/\text{m}^3$

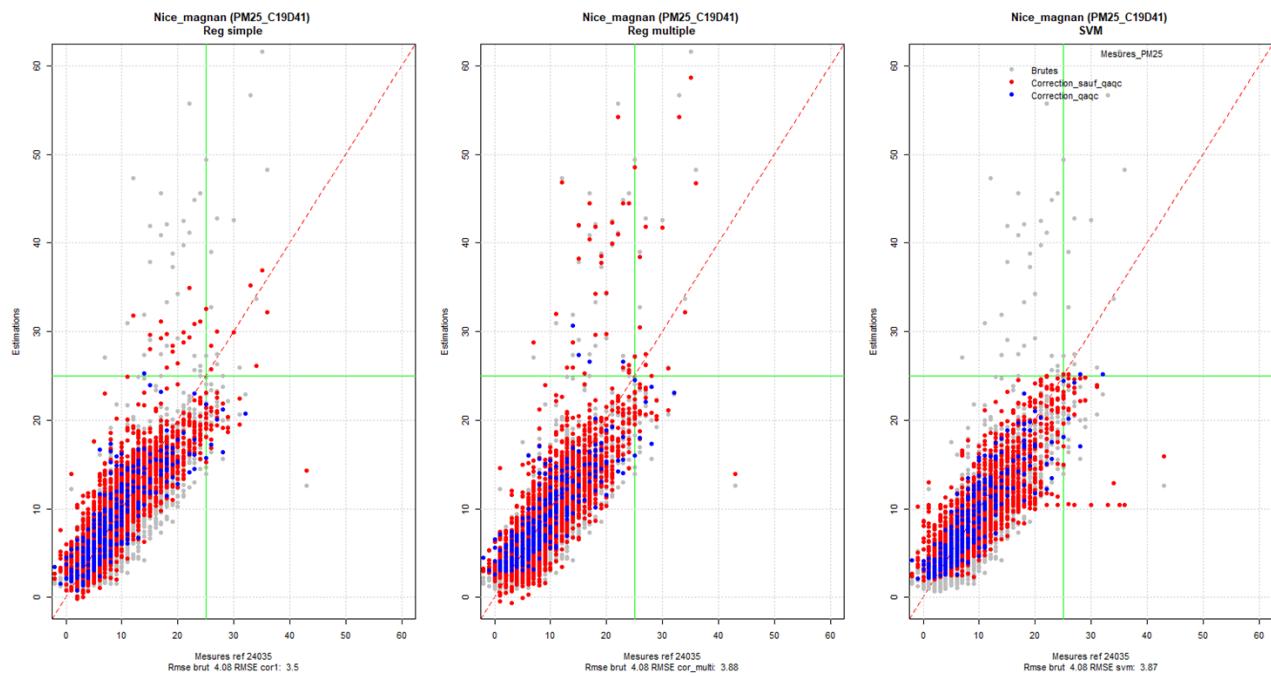


Fig. 2.2 – Données corrigées et brutes des microcapteurs vs les mesures de la station de Nice Magnan en $\mu\text{g}/\text{m}^3$

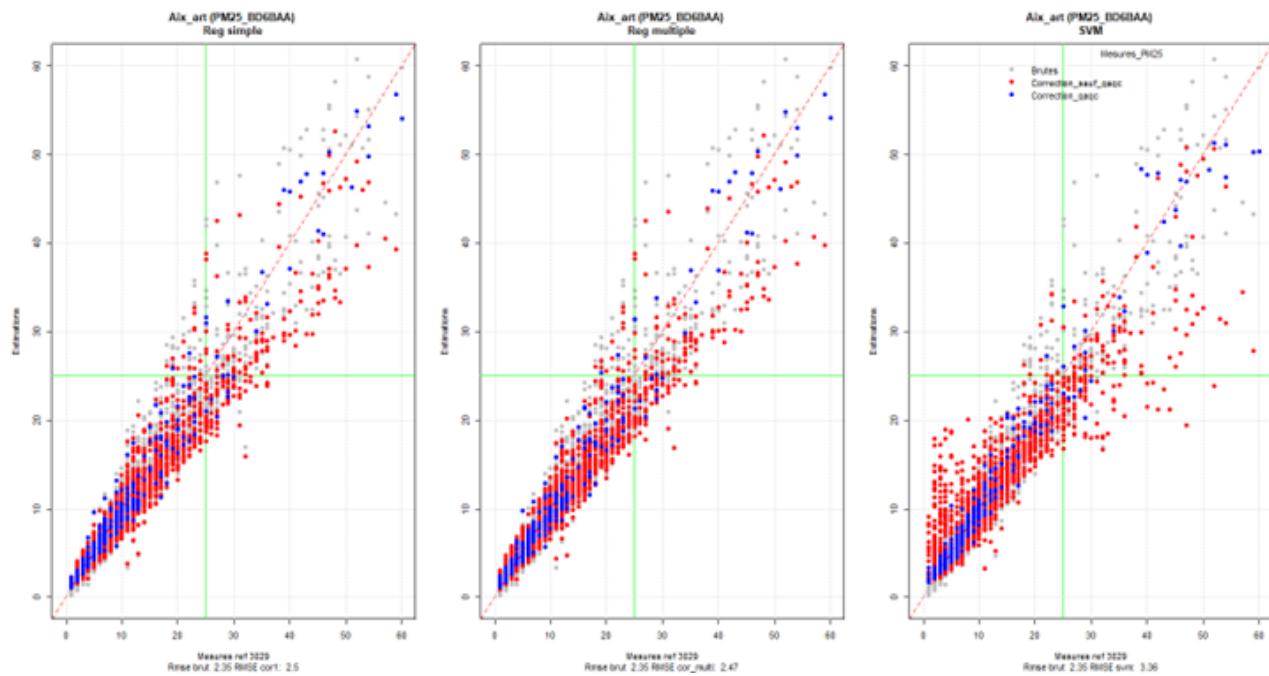


Fig. 2.3 – Données corrigées et brutes des microcapteurs vs les mesures de la station d’Aix Art en $\mu\text{g}/\text{m}^3$

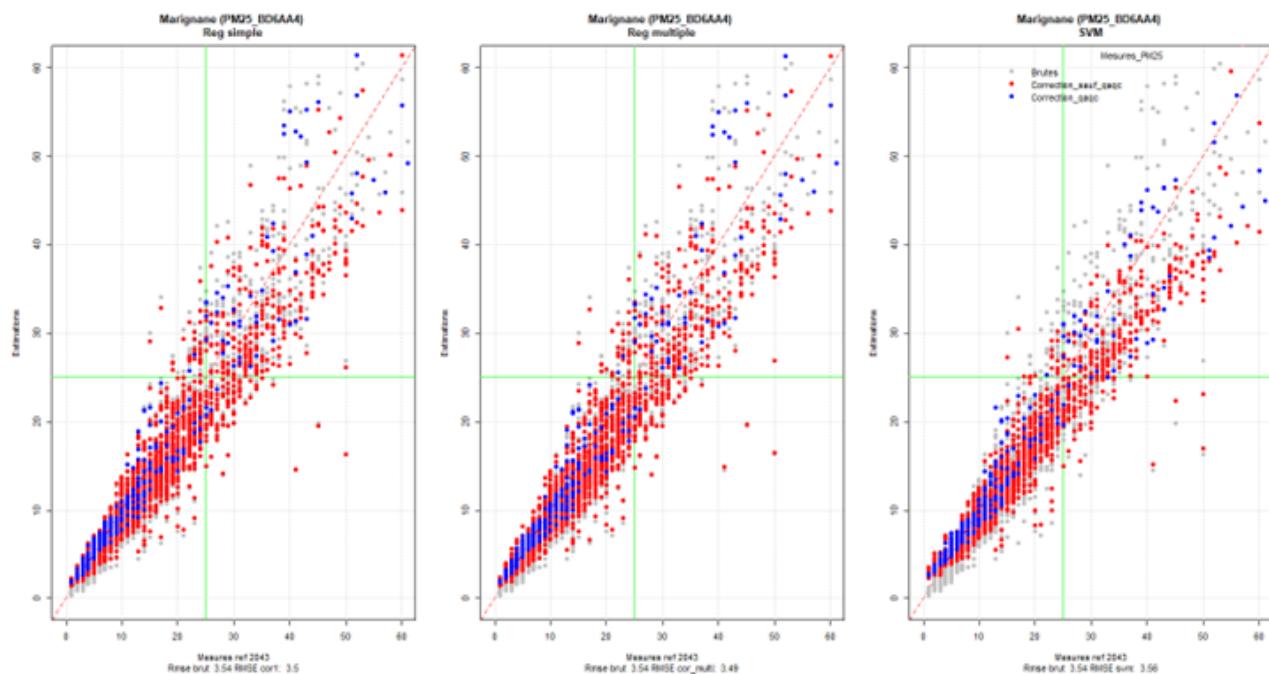


Fig. 2.4 – Données corrigées et brutes des microcapteurs vs les mesures de la station de Marignane en $\mu\text{g}/\text{m}^3$

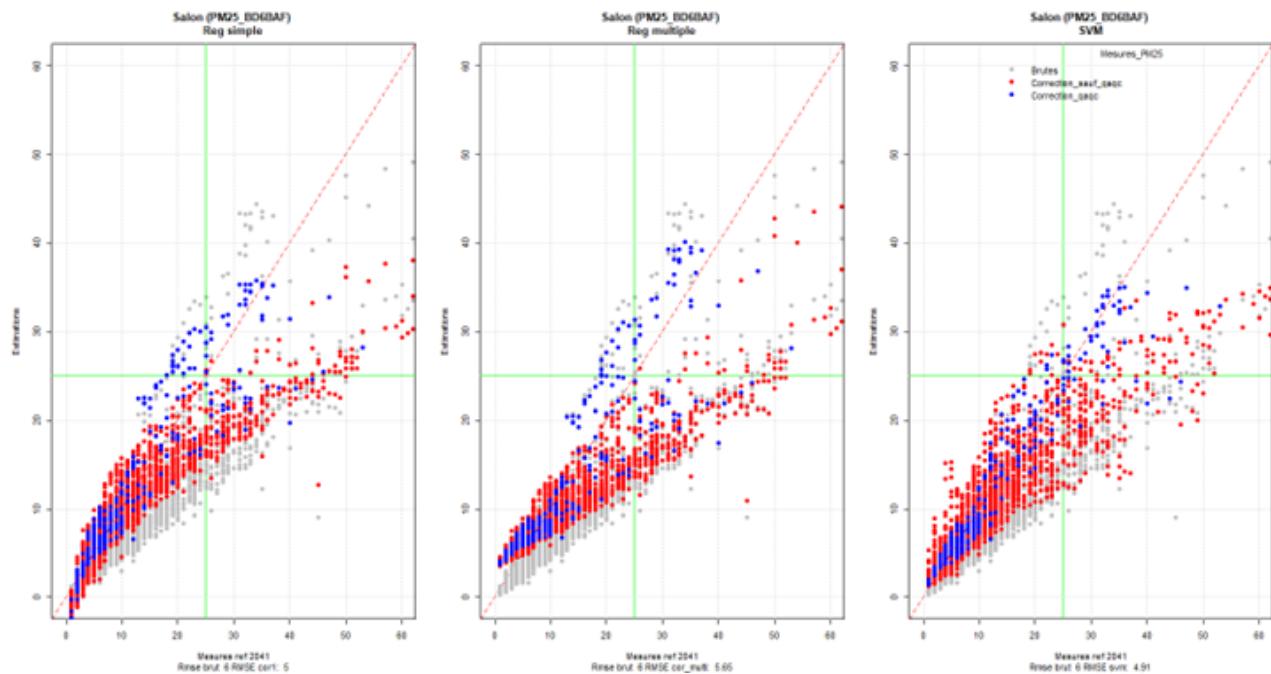


Fig. 2.5 – Données corrigées et brutes des microcapteurs vs les mesures de la station de Salon en $\mu\text{g}/\text{m}^3$

Table de comparaison des RMSE

Pour quantifier les performances et les améliorations des différentes méthodes, nous avons calculé les RMSE (Root Mean Square Error) sur les données de la période du 10/10/2023 au 10/02/2024 en utilisant les mesures brutes comme référence.

Tab. 2.3 – Les scores RMSE brut, régression simple, régression multiple, et SVM

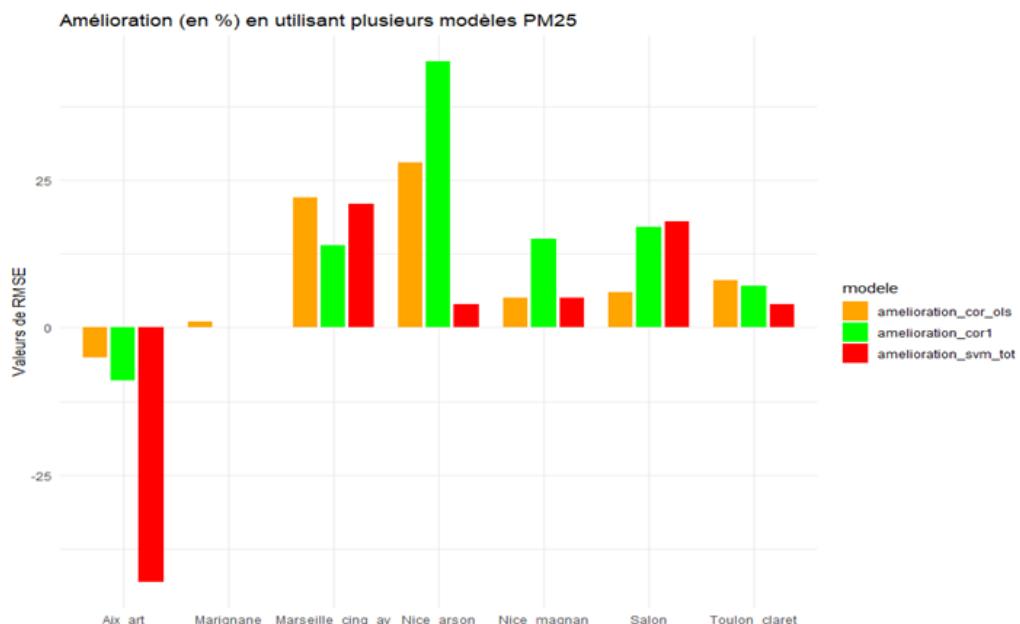
Ville	RMSE brut	RMSE régression simple	RMSE régression multiple	RMSE SVM
Salon	6	5	5,65	4,91
Marignane	3,54	3,5	3,49	3,56
Aix Art	2,35	2,5	2,47	3,37
Marseille_5av	3,63	3,1	2,84	2,88
Toulon_claret	4,14	3,8	3,8	3,98
Nice_magnan	4,08	3,5	3,88	3,87
Nice_arson	3,3	1,8	2,39	3,18
Moyenne	4,12	3,45	3,68	3,73

Tab. 2.4 – Améliorations pour régression simple, régression multiple, et SVM

Ville	Amélioration régression simple	Amélioration régression multiple	Amélioration SVM
Salon	17%	6%	18%
Marignane	0%	1%	0%
Aix Art	-9%	-5%	-43%
Marseille_5av	14%	22%	21%
Toulon_claret	7%	8%	4%
Nice_magnan	15%	5%	5%
Nice_arson	45%	28%	4%
Moyenne	16%	12%	9%

On constate que la régression avec exposant donne la meilleure amélioration en moyenne, et en deuxième position : régression multiple.

Les deux figures ci-dessous présentent les scores brutes, régression simple, régression multiple et SVM dans chaque station :

**Fig. 2.6** – Visualisations des améliorations en pourcentage des trois modèles utilisées

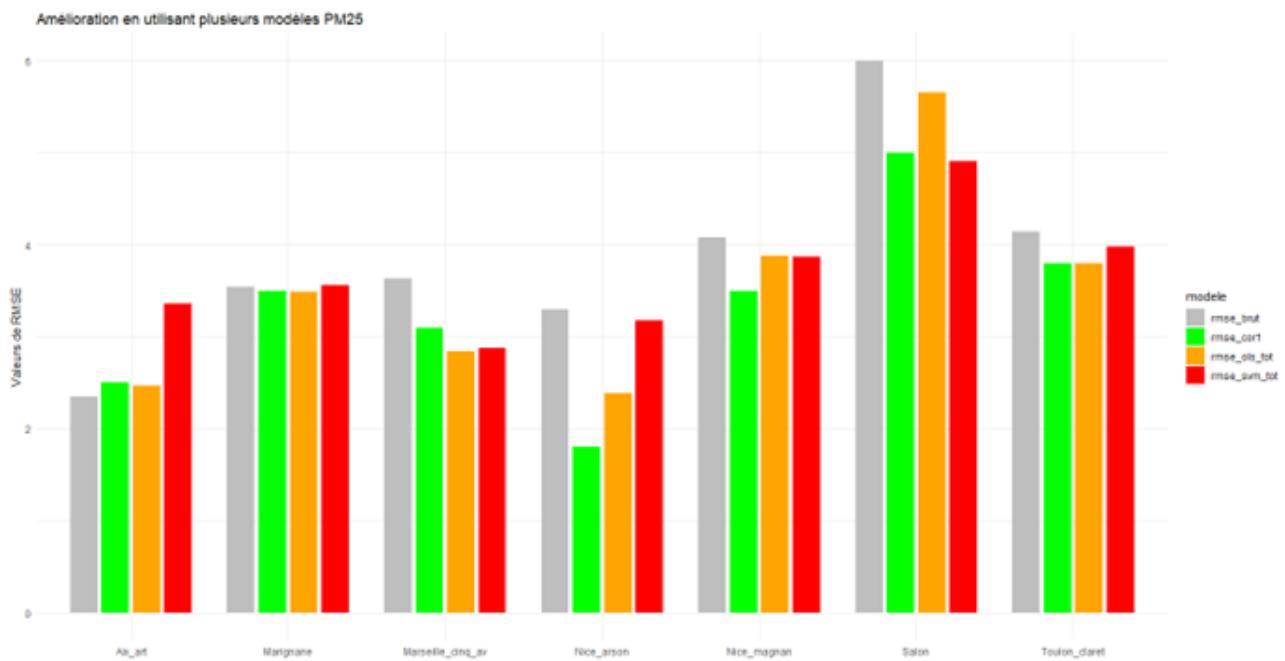


Fig. 2.7 – Visualisations des RMSE des trois modèles utilisées

Choix final de la méthode à adopter

Bilan de la comparaison des modèles utilisés

Les résultats indiquent que la régression simple montre généralement la meilleure performance, avec une réduction moyenne de 16% du RMSE par rapport aux mesures brutes. La régression multiple suit de près avec une amélioration moyenne de 12%, tandis que la correction SVM présente une amélioration moyenne de 9%.

Ces observations montrent que la régression simple est souvent la méthode la plus efficace pour améliorer la précision des corrections dans les villes étudiées.

Incertitude sur les mesures microcapteurs

Méthode de calcul de l'incertitude élargie : L'incertitude élargie est calculée à partir de l'écart type d'erreur dans une gamme de valeurs donnée, rapportée à la valeur seuil et multipliée par 1.96 pour tenir compte de l'intervalle de confiance à 95%. Pour le seuil de 10 µg/m³, l'écart type est mesuré entre 5 et 15 µg/m³, tandis que pour le seuil de 25 µg/m³, il est mesuré entre 20 et 30 µg/m³.

Formule :

$$\text{Incertitude élargie} = \frac{\sigma_{\text{erreur}}}{\text{valeur seuil}} \times 1.96 \quad (2.3)$$

- Incertitudes horaires

Le **Tableau 5** montre les résultats des incertitudes horaires sur les données brutes et les données corrigées avec la régression simple pour les seuils de 10 µg/m³ et 25 µg/m³.

Tab. 2.5 – Incertitudes élargies et biais pour les seuils de 10 µg/m³ et 25 µg/m³

Méthode	Seuil 10 µg/m ³		Seuil 25 µg/m ³	
	Brut	Régression simple	Brut	Régression simple
Incertitude élargie en %	33	36	32	40
Biais en µg/m³	-0.7	-1.10	-0.1	-2.9

Comme indiqué dans le tableau, les incertitudes élargies en pourcentage sont légèrement plus élevées pour le seuil de 25 µg/m³ par rapport à celui de 10 µg/m³. Les biais observés sont également plus importants pour le seuil de 25 µg/m³, ce qui suggère une plus grande variabilité dans les mesures pour des concentrations plus élevées.

- Incertitudes journalières

Le **Tableau 6** présente les incertitudes journalières détaillées pour les seuils de 10 µg/m³ et 25 µg/m³, tant pour les valeurs brutes que corrigées.

Tab. 2.6 – Incertitudes élargies et biais pour les seuils de 10 µg/m³ et 25 µg/m³

Méthode	Seuil 10 µg/m ³		Seuil 25 µg/m ³	
	Brut	Régression simple	Brut	Régression simple
Incertitude élargie en %	34	31	28	23
Biais en µg/m³	-1.74	-0.66	-0.7	-0.94

En observant les résultats dans le tableau, on note que les incertitudes élargies journalières varient entre les données brutes et corrigées, avec une tendance générale à des incertitudes plus faibles pour les valeurs corrigées. Les biais varient également selon le seuil avant et après la correction, reflétant des différences dans la précision des mesures entre les sites.

- Incertitudes annuelles (Sur toute la période)

Le **Tableau 8** résume les incertitudes annuelles observées sur toute la période de mesure. Il présente les moyennes et écarts pour les sites Aix Art, Marseille, et Marignane.

Tab. 2.7 – Incertitudes annuelles par site

	Aix Art	Marseille	Marignane
Moyenne en µg/m ³	10	11	13
Écart en µg/m ³	0.7	1.4	1

Le tableau montre que les valeurs moyennes des concentrations varient légèrement entre les sites, avec des écarts plus importants pour Marseille par rapport aux autres sites. Ces écarts peuvent refléter des différences dans les conditions locales ou les méthodes de mesure.

2.2.2 Evaluation d'une autre période d'entraînement des modèles

Pour une seconde évaluation de la période (QA/QC) concernant le PM_{2,5}, on également examiné une autre période, du 25/11/2023 au 07/12/2023. Voici les résultats obtenus :

Dans le cadre de notre étude, nous avons exploré l'influence de la période d'entraînement sur la performance des différents modèles. Afin de comparer les résultats, nous avons utilisé deux périodes distinctes pour les données de QA/QC : janvier et novembre. Les résultats sont résumés dans le tableau suivant, où les améliorations moyennes en termes de RMSE sont présentées pour chaque méthode :

Tab. 2.8 – Améliorations moyennes pour les méthodes de régression et SVM pour les périodes QA/QC de janvier et novembre

Période QA/QC	Régression simple	Régression multiple	SVM
Janvier	16%	12%	9%
Novembre	6%	10%	-5%

Analyse des résultats

On constate que pour la période QA/QC de janvier, la méthode de régression simple avec exposant montre la meilleure amélioration, suivie par la régression multiple avec humidité relative (HR). En revanche, pour la période QA/QC de novembre, la régression multiple avec HR dépasse la régression simple, offrant une performance deux fois supérieure. Cependant, le SVM dégrade considérablement les résultats comparativement à la première période (Janvier).

La figure suivante 2.8 représente les améliorations des scores calculés avec une autre période d'entraînement :

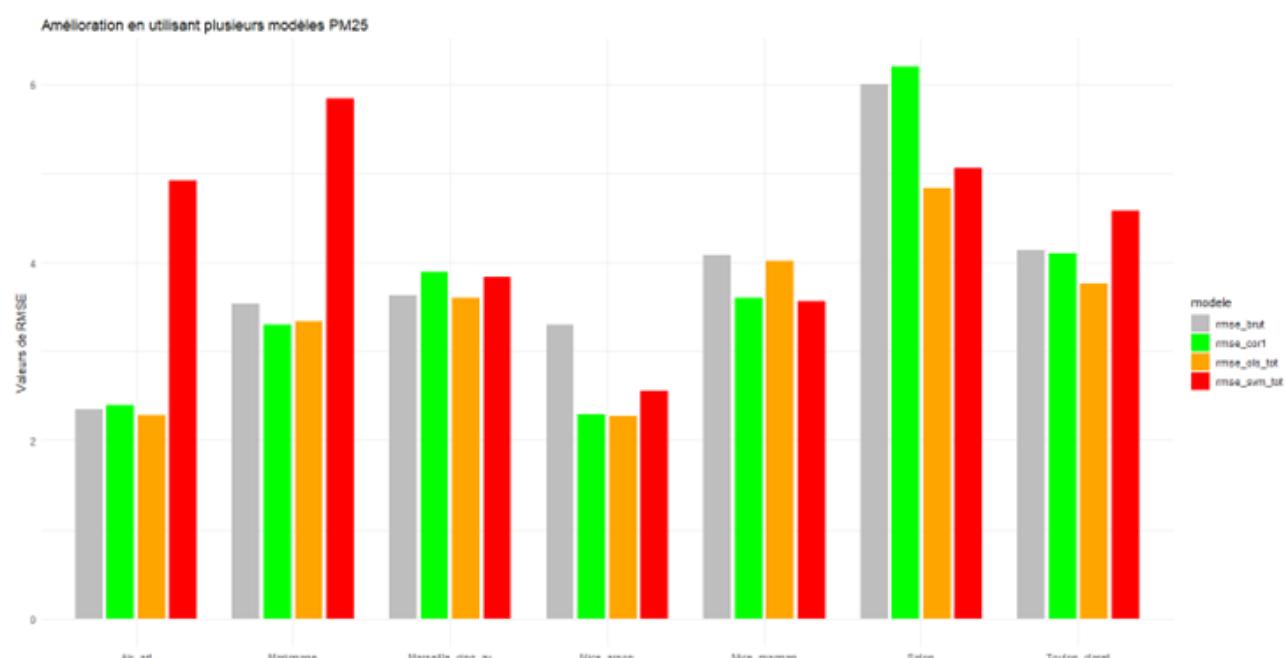


Fig. 2.8 – Les scores des différentes méthodes pour la deuxième période QA/QC - Novembre

Bilan de la comparaison des modèles avec deux périodes d'entraînement différentes

Synthèse

Les résultats suggèrent que l'utilisation de l'humidité relative (HR) comme variable explicative permet de mieux gérer les périodes contenant des valeurs élevées de PM_{2.5}. Cela pourrait indiquer que la régression multiple avec HR est plus robuste face aux variations saisonnières des niveaux de pollution.

2.2.3 Synthèse des résultats de la deuxième correction - en temps réel

Prenons à chaque fois une station de référence différente et regardons les résultats pour les PM_{2.5} :

1. Analyse des résultats de la première correction sur les couples microcapteurs/stations choisies comme référence pour le calcul des ratios

La première correction appliquée aux données des PM_{2.5} montre une nette amélioration par rapport aux données brutes, comme illustré par les deux figures ci-dessus. La figure 2.9 présente les données brutes (en gris) comparées aux données corrigées (en bleu) sur une base horaire. On observe que la correction réduit considérablement les écarts et rapproche les mesures des microcapteurs des valeurs de référence.

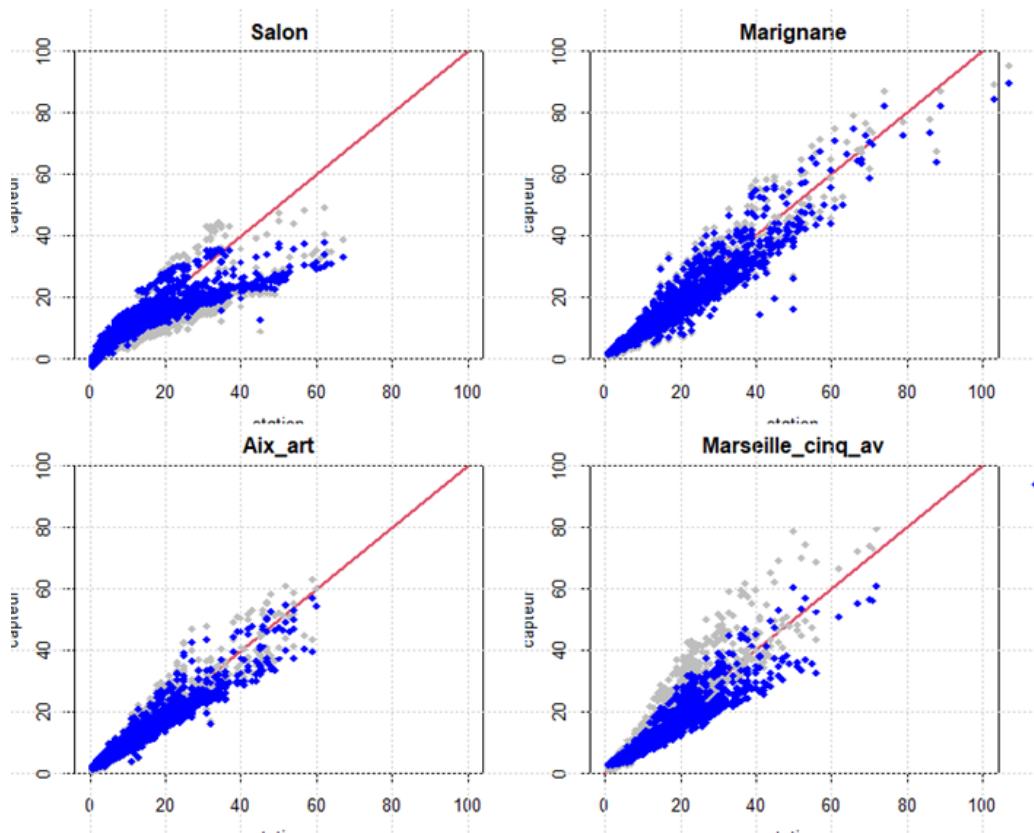


Fig. 2.9 – Les données brutes (gris) et corrigées 1 (bleu) des PM_{2.5} en horaire (en $\mu\text{g}/\text{m}^3$) pour les stations qui serviront comme station de référence

De plus, la figure 2.10, qui montre les données journalières, confirme cette tendance, soulignant que la correction permet de mieux capturer les tendances journalières des PM_{2.5}.

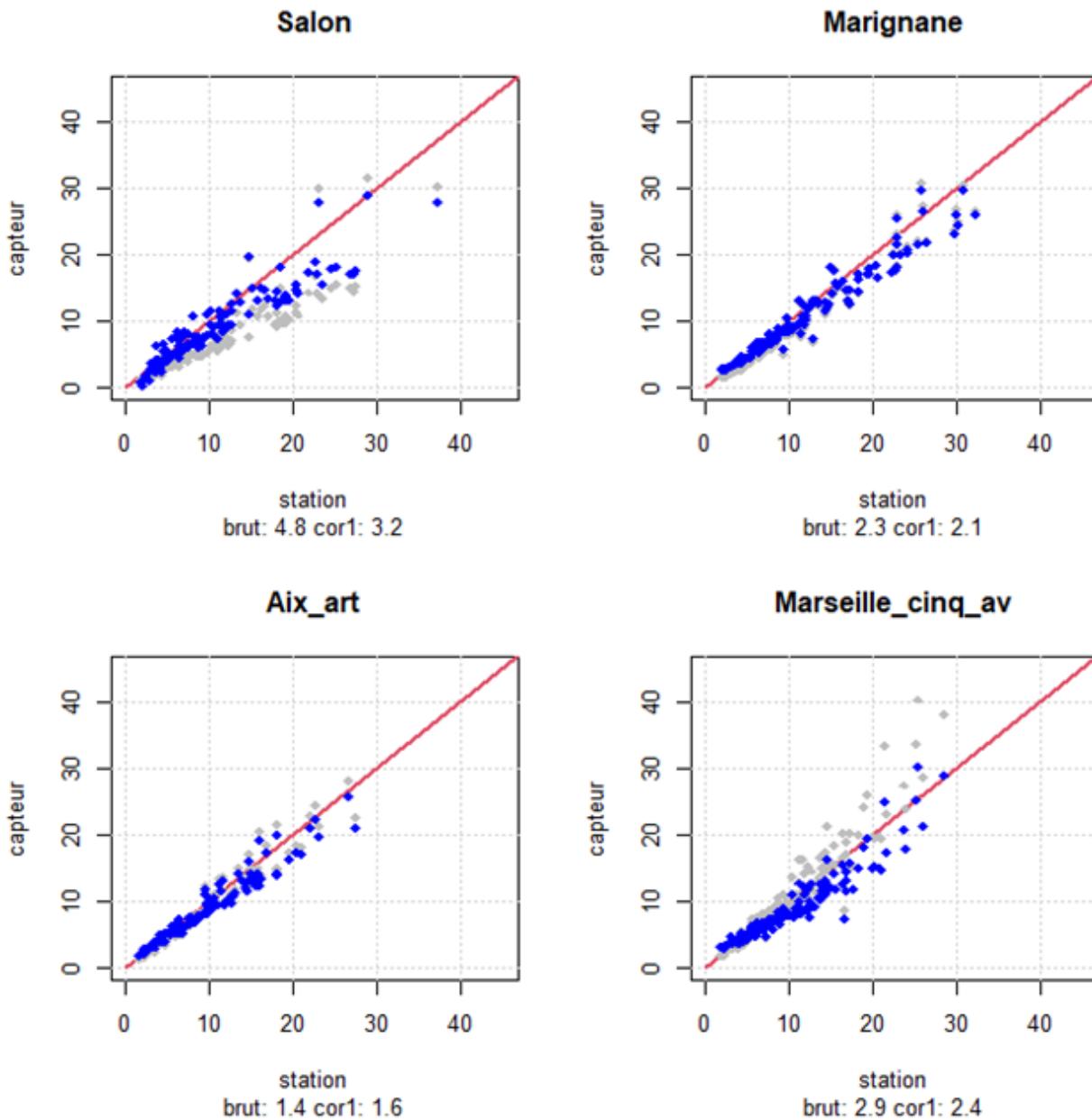


Fig. 2.10 – Les données brutes (gris) et corrigées 1 (bleu) des PM_{2.5} en journalier (en $\mu\text{g}/\text{m}^3$) pour les stations qui serviront comme station de référence

2. Calcul et application des ratios de correction entre stations de fond

Nous allons présenté les résultats de comparaison entre différents sites de fond pour la période QA/QC du mois de **janvier**. Chaque station de référence est analysée séparément pour mieux comprendre l'impact des corrections appliquées aux données brutes.

2.1 Pour la période QAQC du mois de janvier

i. Aix Art comme station de référence

Le tableau 2.9 présente les scores RMSE pour les données brutes et corrigées en appliquant différents ratios (1h, 3h, 6h, 12h, 24h). Aix Art est utilisée comme station de référence pour le calcul des ratios.

Tab. 2.9 – Les scores RMSE brut, correction1 et correction 2 en appliquant différents ratios (1h, 3h, 6h, 12h, 24h)

ville_ref	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Aix_art	Marignane	3.49	3.41	2.77	2.6	2.72	2.85	2.86
Aix_art	Salon	5.94	4.92	4.1	4.03	4.04	4.08	4.18
Aix_art	Marseille_5av	3.46	3.2	2.25	2.07	2.16	2.43	2.53
Moyenne		3.84	3.55	2.57	2.68	2.81	2.97	3.03

Analyse

La correction 2 avec ratio_3h et ratio_6h à partir de la station d'Aix Art vient améliorer en moyenne de 1 µg/m³ par rapport aux données corrigées 1 et avec 1.2 µg/m³ par rapport aux données brutes.

Tab. 2.10 – Les améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1

ville_ref	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Aix_art	Marignane	2	26	19	24	20	16	16
Aix_art	Salon	17	32	17	18	18	17	15
Aix_art	Marseille_5av	8	40	30	35	32	24	21
Moyenne		9	33	22	26	23	19	17

Analyse

On remarque qu'en partant de Aix Art comme station de référence, on a une amélioration en ratio 3h en moyenne de 26 % (ratio_6h 23% et 22% en ratio 1h), ce qui fait qu'en global on a une amélioration de 33 % en moyenne après deux corrections. On peut corriger Marseille_5Avenue (40%), Marignane (26%) et Salon (32%) à partir de la station d'Aix Art en ratio 3h ou 6h, ou bien même avec le ratio 1h.

ii. Marseille comme station de référence

Tab. 2.11 – Les scores RMSE brut, correction1 et correction 2 en appliquant différents ratios (1h, 3h, 6h, 12h, 24h)

ville_ref	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Marseille_5av	Marignane	3.51	3.43	3.49	3.39	3.36	3.35	3.3
Marseille_5av	Salon	6.03	5	4.03	3.98	3.98	3.98	3.99
Marseille_5av	Aix_art	2.31	2.53	2.26	2.17	2.21	2.35	2.49
Moyenne		3.95	3.65	3.26	3.18	3.18	3.23	3.26

Analyse

La correction 2 avec ratio_3h et ratio_6h à partir de la station de Marseille vient améliorer en moyenne de $0.5 \mu\text{g}/\text{m}^3$ par rapport aux données corrigées 1 et avec $0.8 \mu\text{g}/\text{m}^3$ par rapport aux données brutes.

Tab. 2.12 – Les améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1

ville_ref	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Marseille	Marignane	2	3	-2	1	2	2	4
Marseille	Salon	17	34	19	20	20	20	20
Marseille	Aix Art	-10	6	11	14	13	7	2
Moyenne		3	14	9	12	12	10	9

Analyse

On remarque qu'en partant de Marseille comme station de référence, on a une amélioration en ratio 3h en moyenne de 12 % identique au ratio 6h, ce qui fait qu'en global on a une amélioration de 14 % en moyenne après deux corrections. On peut corriger Aix art (14%), Marignane (1%) et Salon (20%) à partir de la station de Marseille en ratio 3h ou 6h.

iii. Marignane comme station de référence

Tab. 2.13 – Les scores RMSE brut, correction1 et correction 2 en appliquant différents ratios (1h, 3h, 6h, 12h, 24h)

ville_ref	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Marignane	Salon	5.88	4.86	4.16	4.03	3.87	3.72	4.05
Marignane	Aix art	2.26	2.48	1.95	1.77	1.78	1.9	2.1
Marignane	Marseille_5av	3.46	3.2	2.35	2.24	2.3	2.35	2.5
Moyenne		3.87	3.51	2.82	2.68	2.65	2.66	2.88

Analyse

La correction 2 avec ratio_3h et ratio_6h à partir de la station de Marignane vient améliorer en moyenne de $0.8 \mu\text{g}/\text{m}^3$ par rapport aux données corrigées 1 et avec $1.2 \mu\text{g}/\text{m}^3$ par rapport aux données brutes.

Tab. 2.14 – Les améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1

ville_ref	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Marseille	Salon	17	31	14	17	20	2	4
Marseille	Aix art	-10	22	21	29	28	20	20
Marseille	Marseille_5Av	8	35	27	30	28	7	2
Moyenne		5	29	21	25	25	24	18

Analyse

On remarque qu'en partant de Marignane comme station de référence, on a une amélioration en ratio 3h en moyenne de 25 % identique au ratio 6h, ce qui fait qu'en global on a une amélioration de 29 % en moyenne après deux corrections. On peut corriger Aix art (29%), Marseille (30%) et Salon (17%) à partir de la station de Marignane en ratio 3h ou 6h.

La Figure 2.11 illustre la comparaison entre les données brutes (en gris) et les données corrigées 2 (en bleu) des PM_{2.5}, enregistrées sur une base horaire à partir des différents couples de référence.

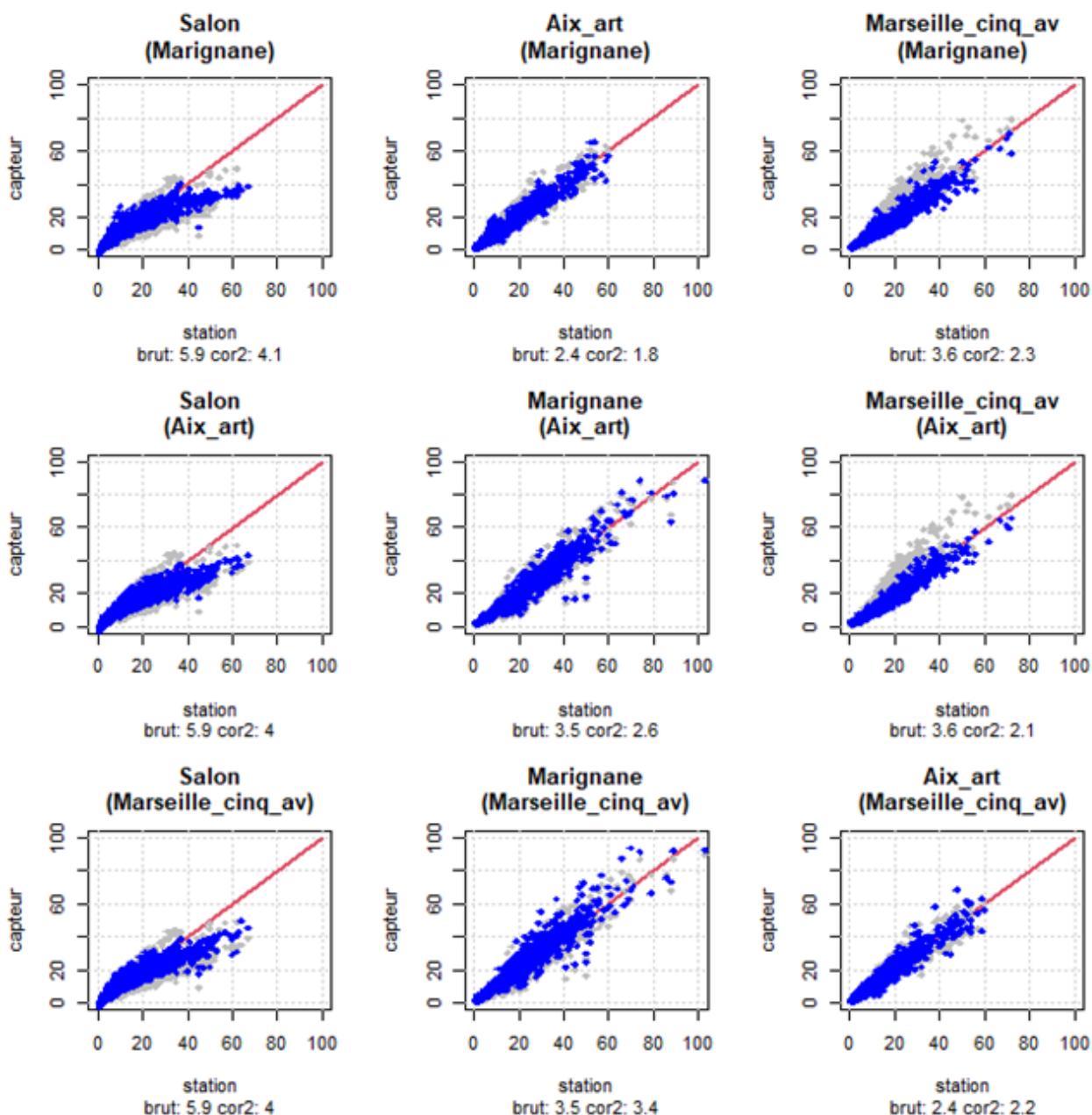


Fig. 2.11 – Les données brutes (gris) et corrigées 2 (bleu) des PM_{2.5} en horaire avec différentes stations de référence(en $\mu\text{g}/\text{m}^3$)

En conclusion, cette figure 2.11 démontre que la méthode de correction utilisée a considérablement amélioré la précision des mesures des PM_{2.5} en horaire. Les données corrigées 2 offrent une meilleure approximation des valeurs attendues.

La figure 2.12 présente une comparaison entre les données brutes (en gris) et les données corrigées 2 (en bleu) des PM_{2.5} sur une base journalière.

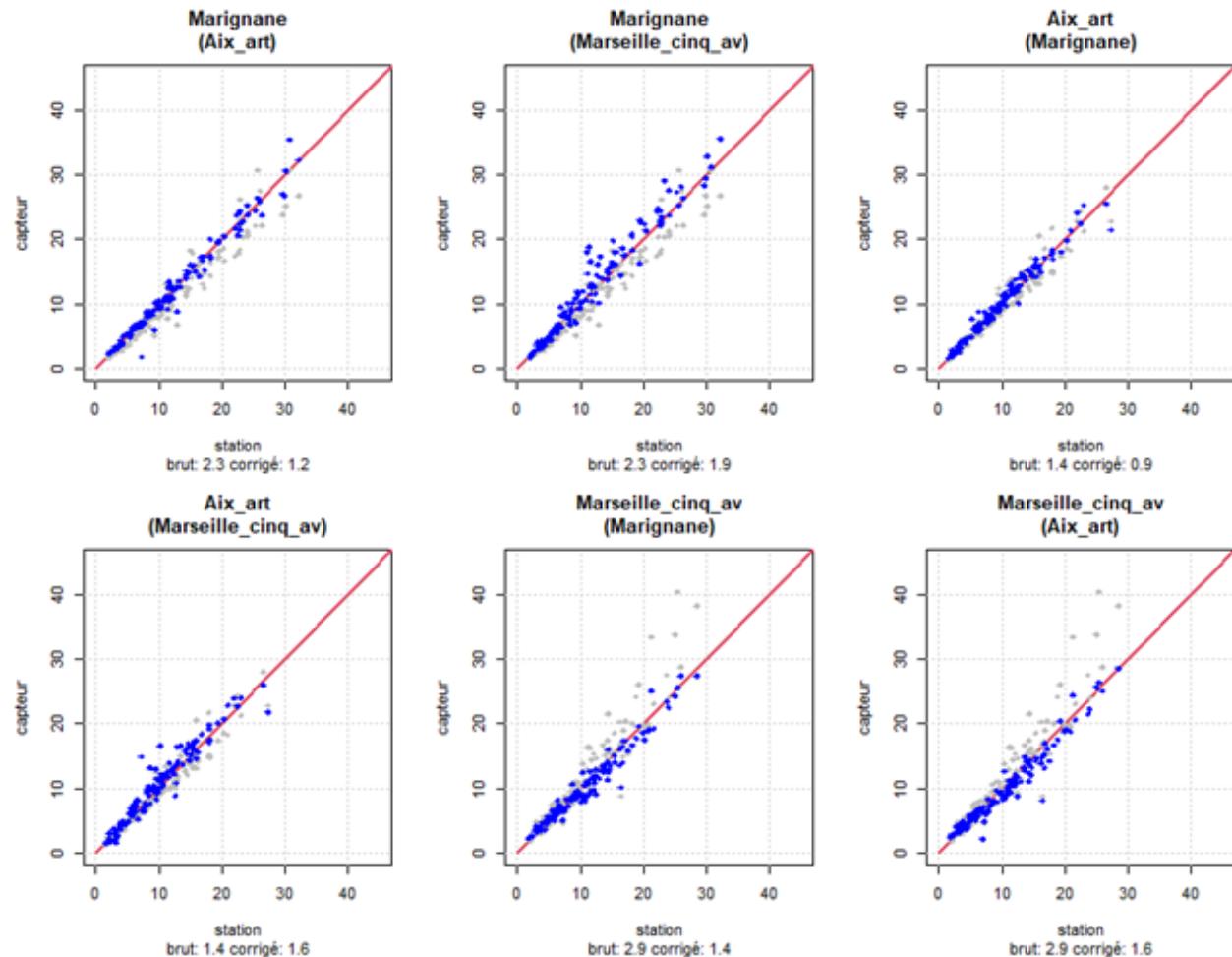


Fig. 2.12 – Les données brutes (gris) et corrigées 2 (bleu) des PM_{2.5} en journalier avec différentes stations de référence(en $\mu\text{g}/\text{m}^3$)

2.2 Pour la période QA/QC du mois de novembre

i. Aix Art comme station de référence

ville_ref	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Aix_art	Marignane	3.49	3.27	2.94	2.65	2.63	2.72	2.86
Aix_art	Salon	5.94	6.06	4.95	4.83	4.88	5.07	5.03
Aix_art	Marseille_5Av	3.46	3.87	3.24	3.19	3.24	3.62	3.73
Moyenne		3.84	3.89	3.12	3.23	3.30	3.53	3.60

Tab. 2.15 – Les scores RMSE brut, correction1 et correction 2 en appliquant différents ratios (1h, 3h, 6h, 12h, 24h)

Analyse

La correction 2 avec ratio_3h et ratio_6h à partir de la station d'Aix Art vient améliorer en moyenne de $0.7 \mu\text{g}/\text{m}^3$ par rapport aux données corrigées 1 et avec $0.61 \mu\text{g}/\text{m}^3$ par rapport aux données brutes.

ville_ref	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Aix_art	Marignane	6	24	10	19	20	17	13
Aix_art	Salon	-2	19	18	20	19	16	17
Aix_art	Marseille_5Av	-12	8	16	18	16	6	4
Moyenne		-2	19	26	20	17	10	8

Tab. 2.16 – Les améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1

Analyse

On remarque qu'en partant de Aix Art comme station de référence, on a une amélioration en ratio_3h en moyenne de 20 % identique au ratio_6h (17%) et 26% en ratio 1h, ce qui fait qu'en global on a une amélioration de 32 % en moyenne après deux corrections. On peut corriger Marseille_5Avenue, Toulon Claret, Marignane et Salon à partir de la station d'Aix Art en ratio 3h ou 6h. Même si la correction 1 dégrade les données, la correction 2 vient rattraper cette dégradation.

ii. Marseille cinq avenue comme station de référence

ville_ref	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Marseille_5Av	Marignane	3.51	3.3	3.99	3.9	3.83	3.77	3.92
Marseille_5Av	Salon	6.03	6.17	4.11	4.06	4.01	3.93	3.9
Marseille_5Av	Aix_art	2.31	2.38	2.42	2.38	2.4	2.41	2.55
Moyenne		3.95	3.95	3.51	3.45	3.41	3.37	3.46

Tab. 2.17 – Les scores RMSE brut, correction1 et correction 2 en appliquant différents ratios (1h, 3h, 6h, 12h, 24h)

Analyse

La correction 2 avec ratio_3h à partir de la station de Marseille 5Avenue vient améliorer les scores en moyenne de $0.5 \mu\text{g}/\text{m}^3$ par rapport aux données corrigées 1 et aux données brutes.

On remarque aussi que si on enlève les deux micro capteurs de Nice pour être un peu plus cohérents :

ville_ref	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Marseille_5Av	Marignane	6	-11	-21	-18	-16	-14	-19
Marseille_5Av	Salon	-2	33	33	34	35	36	37
Marseille_5Av	Aix_art	-3	-3	-2	0	-1	-1	-7
Moyenne		0	6	3	5	6	7	4

Tab. 2.18 – Les améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1 (Sans les sites de Nice)

Analyse

On remarque qu'en partant de Marseille Cinq Avenue comme station de référence, on a une amélioration en ratio_3h en moyenne de 14 %, ce qui fait qu'en global on a une amélioration de 28 % en moyenne après deux corrections. On peut corriger Aix Art, Marignane et Salon à partir de la station de Marseille 5Avenues en ratio 3h ou 6h. En revanche, il faut un ratio_24h pour corriger Toulon Claret avec un pourcentage de 7%.

Prenons à chaque fois une station de référence différente et regardons les résultats maintenant pour les PM₁₀ :

1. Calcul et application des ratios de correction entre stations de fond et trafic d'une même ville

Nous allons présenté les résultats de la transférabilité des corrections entre différents sites de fond et trafic dans une même ville pour la période QA/QC du mois de **janvier**.

1.1 Pour la période QA/QC du mois de janvier

i. La ville de Nice

Le tableau 2.19 présente les scores RMSE pour les données brutes et corrigées en utilisant différents ratios (1h, 3h, 6h, 12h, 24h) pour les stations de fond et de trafic à Nice.

Tab. 2.19 – Scores RMSE brut, correction 1 et correction 2 pour différentes stations à Nice

ville_réf	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Nice Magnan	Nice Arson	11,2	7,46	7,79	7,22	7,39	7,97	7,39
Nice Arson	Nice Magnan	11,48	6,89	8,98	8,32	7,61	7,56	7,06

Le tableau 2.20 montre les améliorations en pourcentage par rapport aux mesures brutes et corrigées 1 pour les ratios différents.

Tab. 2.20 – Améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1 pour les stations à Nice

ville_réf	ville	brut-vers-cor1	brut-vers-cor2_3h	cor1-vers_ratio1h	cor1-vers_ratio3h	cor1-vers_ratio6h	cor1-vers_ratio12h	cor1-vers_ratio24h
Magnan	Arson	33	36	-4	3	1	-7	1
Arson	Magnan	40	28	-30	-21	-10	-10	-2

Analyse

On remarque que les deux sites de Nice en PM_{2.5} ne peuvent pas se corriger mutuellement de manière efficace. Les variations importantes des améliorations observées dans les ratios indiquent des différences significatives entre les sites, ce qui pourrait être lié à la spécificité du site de Magnan en PM_{2.5}.

ii. La ville de Toulon

Le tableau 2.21 présente les scores RMSE pour les données brutes et corrigées en utilisant différents ratios (1h, 3h, 6h, 12h, 24h) pour les stations de fond et de trafic à Toulon.

Tab. 2.21 – Scores RMSE brut, correction 1 et correction 2 pour différentes stations à Toulon

ville_réf	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Toulon Claret	Foch	26,26	12,47	10,51	10,55	11,49	12,15	12,26
Toulon Foch	Claret	12,88	8,6	6,99	6,79	7,1	7,75	7,75

Le tableau 2.22 montre les améliorations en pourcentage par rapport aux mesures brutes et corrigées 1 pour les ratios différents.

Tab. 2.22 – Améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1 pour les stations à Toulon

ville_réf	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Claret	Foch	53	60	16	15	8	3	2
Foch	Claret	33	47	19	21	17	10	10

Analyse

On remarque que les deux sites de Toulon en PM₁₀ se corrigeent bien mutuellement. Les améliorations sont significatives avec des réductions de RMSE notables, en particulier pour les ratios 3h et 6h, indiquant une bonne compatibilité des sites pour la correction des données.

iii. La ville d'Aix

Le tableau 2.23 présente les scores RMSE pour les données brutes et corrigées en utilisant différents ratios (1h, 3h, 6h, 12h, 24h) pour les stations de fond et de trafic à Aix.

Tab. 2.23 – Scores RMSE brut, correction 1 et correction 2 pour différentes stations à Aix

ville_réf	ville	rmse_brut	rmse_cor1	rmse_ratio_1h	rmse_ratio_3h	rmse_ratio_6h	rmse_ratio_12h	rmse_ratio_24h
Aix Art	Aix Roy René	14,03	12,39	11,4	11,46	11,66	11,82	11,86
Aix Roy René	Aix Art	7,97	7,17	8,96	6,48	6,95	6,94	6,48

Le tableau 2.24 montre les améliorations en pourcentage par rapport aux mesures brutes et corrigées 1 pour les ratios différents.

Tab. 2.24 – Améliorations des scores en %, par rapport aux mesures brutes et/ou corrigées 1 pour les stations à Aix

ville_réf	ville	brut_vers_cor1	brut_vers_cor2_3h	cor1_vers_ratio1h	cor1_vers_ratio3h	cor1_vers_ratio6h	cor1_vers_ratio12h	cor1_vers_ratio24h
Aix Art	Roy René	12	18	8	8	6	5	4
Roy René	Aix Art	10	19	-25	10	3	3	10

Analyse

En partant d'un site de trafic vers un site de fond ou inversement à Aix, on observe une amélioration de 8% à 10% en ratio 3h, ce qui représente une amélioration globale de 19% après deux corrections. Le ratio 24h montre également une amélioration notable dans le cas spécifique de Roy René à Aix Art.

Conclusion

- La correction entre un site de trafic et un site de fond (ou inversement) est efficace pour les villes d'Aix et Toulon, avec des améliorations significatives observées dans les ratios de correction.
- En revanche, cette correction ne semble pas fonctionner efficacement pour la ville de Nice, ce qui pourrait être lié à la spécificité du site de Magnan en PM₁₀.

2.3 Discussion : Exploration des données

Pour améliorer la précision des microcapteurs Nexlec, nous disposons de deux ensembles de données distincts :

- Ensemble d'entraînement (période QA/QC) – Train set : Période du 1er janvier 2024 au 13 janvier 2024.
- Ensemble de test (période hors QA/QC) – Test set : Couvre la période totale du 10 octobre 2023 au 10 février 2024, à l'exception de la phase d'entraînement.
→ **Objectif** : Évaluer l'efficacité de la correction sur des situations non incluses dans la phase d'entraînement, mettant ainsi à l'épreuve la robustesse de la correction sur des données nouvelles.

Représentativité des données de l'échantillon d'entraînement par rapport à l'ensemble des données test Cette section présente des histogrammes superposés décrivant la distribution des données provenant des stations fixes pour les microcapteurs Nexlec. Chaque graphique compare les valeurs mesurées pendant la période d'entraînement (Train set) et celle de test (Test set), permettant une évaluation visuelle des variations entre les deux ensembles. Dans la [figure 2.13](#), on a affiché deux exemples des mesures des station en PM_{2.5}.

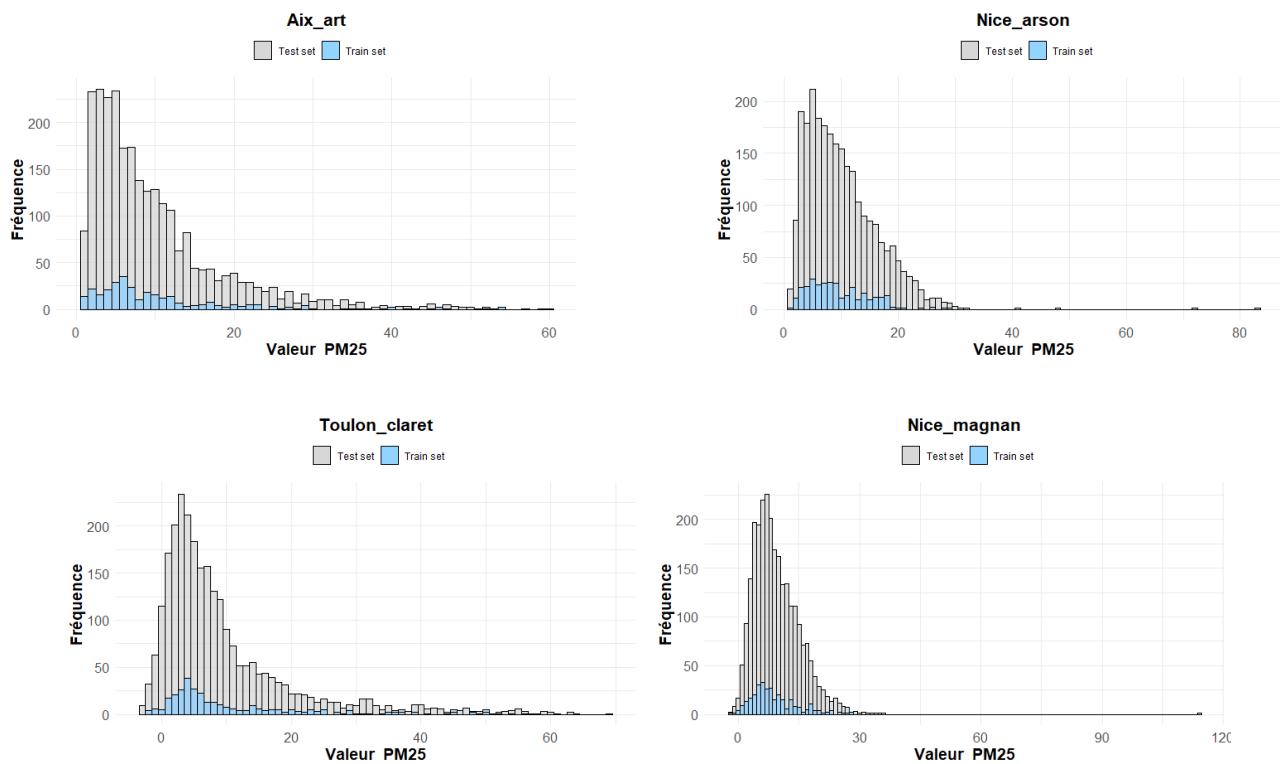


Fig. 2.13 – Graphiques de la représentativité du train set par rapport au test set.

Analyse des graphiques : Les résultats finaux suggèrent que, pour la plupart des stations, l'échantillon d'entraînement demeure représentatif des gammes de valeurs observées. Ces constatations reposent sur l'analyse des distributions des données des stations fixes, comparant les périodes d'entraînement et de test.

Chapitre 3

Méthodes d'Interpolation pour la Cartographie en Temps Réel

Après avoir effectué des corrections sur les données brutes provenant d'une dizaine de microcapteurs, en les ajustant en fonction des mesures des stations fixes, la suite logique de cette mission a été de cartographier ces résultats pour obtenir une vue d'ensemble précise de la qualité de l'air. La création des cartes horaires précises est essentielle pour une gestion efficace des polluants atmosphériques tels que les particules fines **PM_{2.5}**.

Dans le cadre du projet de déploiement futur de plus de 1000 microcapteurs dans la région PACA, il est crucial de développer des méthodes d'interpolation capables de générer des cartes en temps réel à partir des données disponibles. Ce chapitre explore et compare différentes méthodes d'interpolation, telles que l'inverse distance weighting (IDW), le krigeage, le k-nearest neighbors (KNN) et la fonction `interp`.

Étant donné que le réseau de microcapteurs est encore limité, j'ai procédé à des simulations de données pour des ensembles de 10, 50, 100, 300 et 1000 points de mesure. Ces simulations permettent d'explorer la performance des différentes méthodes d'interpolation dans des conditions variées, et d'anticiper les défis que posera un déploiement massif des capteurs.

L'objectif est d'identifier les méthodes les plus adaptées pour fournir des cartes horaires précises, tout en garantissant l'absence d'artefacts, la rapidité de calcul, et l'exactitude aux points de mesure des stations fixes. Les résultats obtenus fourniront des recommandations pour la future gestion des données lorsque le réseau de microcapteurs sera étendu à l'ensemble de la région.

3.1 Interpolation avec Inverse Distance Weighting (IDW)

Principe de Fonctionnement

L'Interpolation par Distance Inverse (IDW) est une méthode géostatistique utilisée pour estimer les valeurs de données en un point donné à partir des valeurs observées à des points voisins. Le principe fondamental de cette méthode repose sur l'idée que les points plus proches

ont une influence plus importante sur la valeur estimée que les points plus éloignés. Mathématiquement, la valeur estimée $z(x)$ en un point x est calculée comme une moyenne pondérée des valeurs observées, où les poids sont inversément proportionnels à une fonction de la distance entre les points :

$$z(x) = \frac{\sum_{i=1}^n \frac{z_i}{d_i^\lambda}}{\sum_{i=1}^n \frac{1}{d_i^\lambda}}$$

où z_i est la valeur observée au point i , d_i est la distance entre le point d'estimation x et le point i , et λ est un paramètre de puissance qui contrôle l'influence des points voisins (7).

Résultats des Simulations

Localisation des 100 points de mesure simulés sur la carte brute

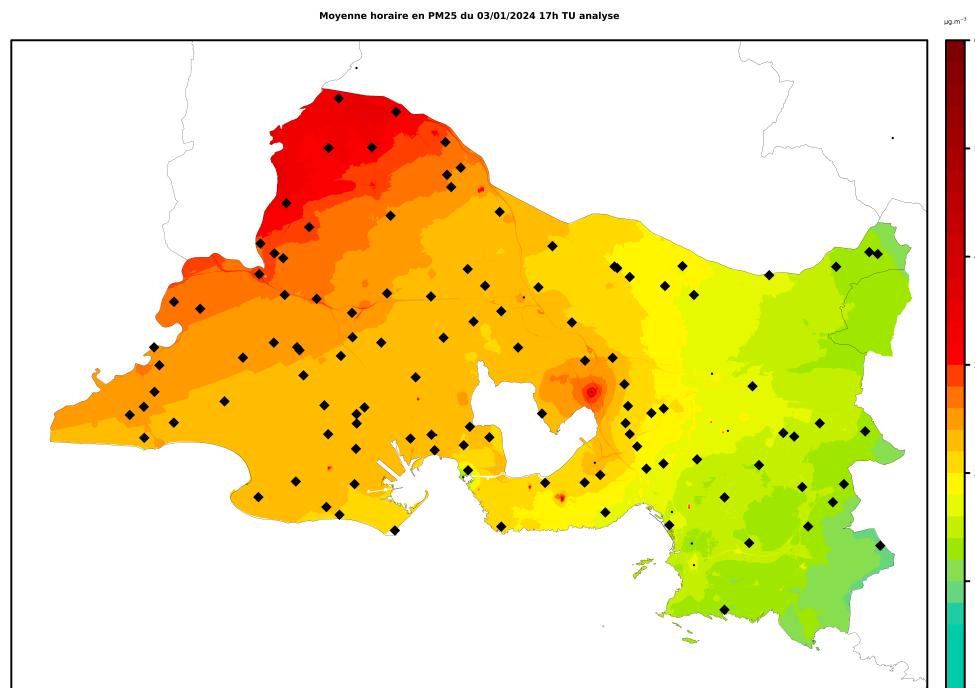


Fig. 3.1 – Carte brute avec 100 points de mesure simulés.

Carte des écarts spatialisés calculés avec la méthode IDW

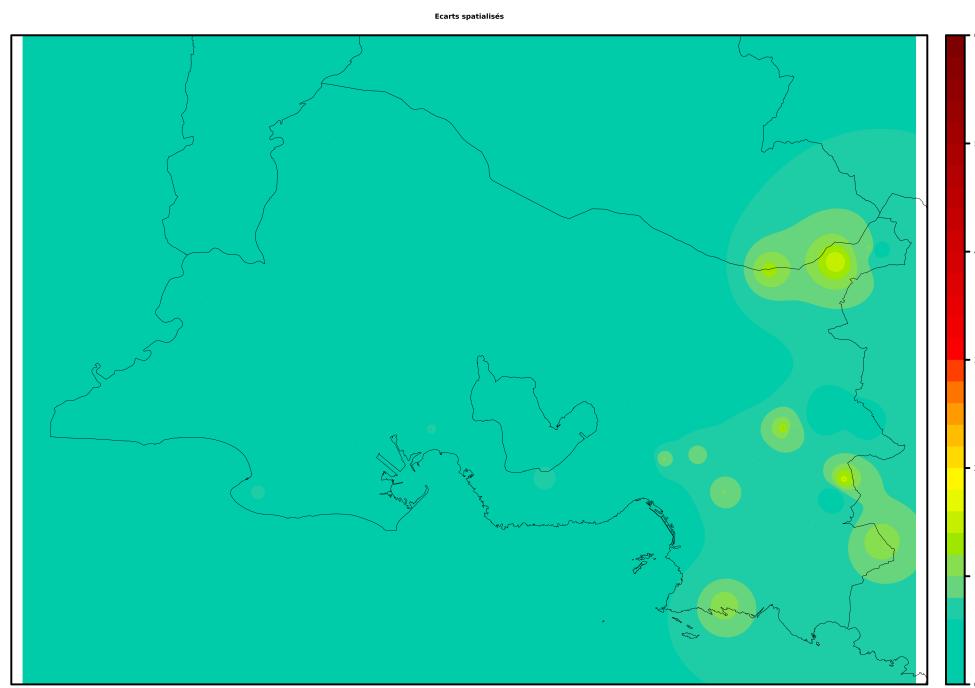


Fig. 3.2 – Carte des écarts calculés avec IDW avec 100 points de mesure.

Carte d'interpolation corrigée avec la méthode IDW

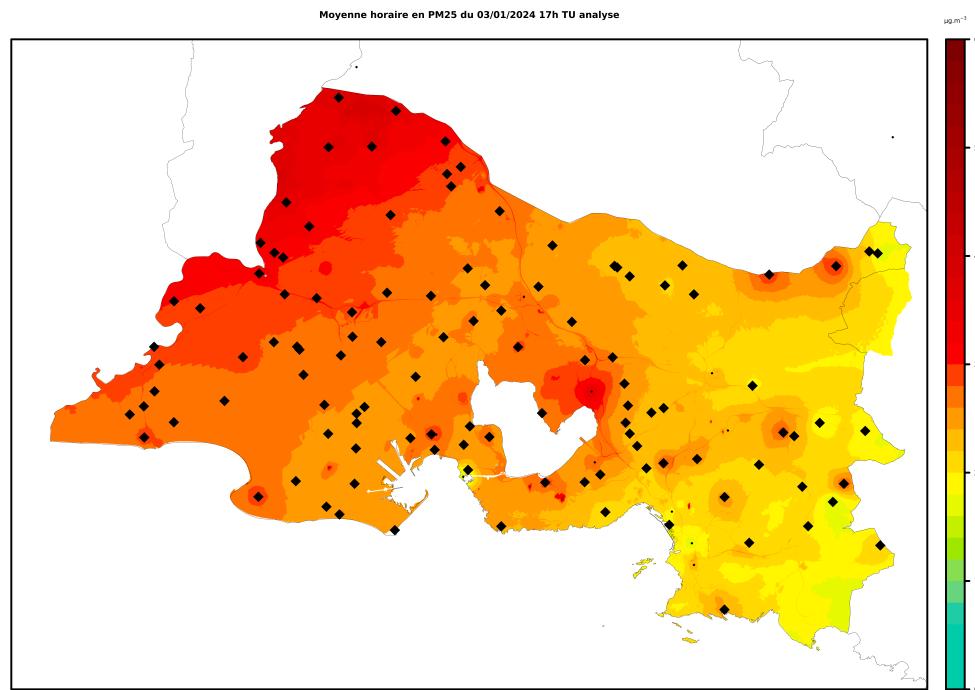


Fig. 3.3 – Carte d'interpolation corrigée IDW avec 100 points de mesure.

Analyse des résultats

Avec l'algorithme IDW (Inverse Distance Weighting), les valeurs interpolées aux points des stations sont exactes comme on peut le voir dans le graphe de la figure 3.4, ce qui garantit une bonne précision aux emplacements des mesures. Cependant, cette méthode présente un inconvénient majeur : elle nécessite environ 20 minutes pour traiter seulement 100 points de mesure, ce qui la rend peu efficace pour des volumes de données plus importants.

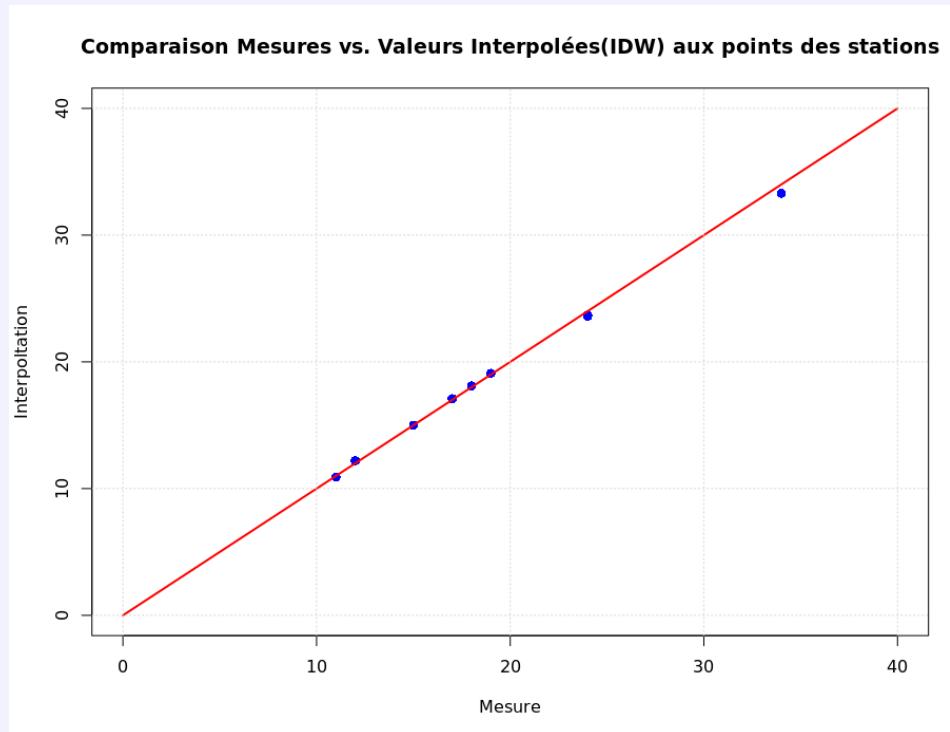


Fig. 3.4 –

3.2 Méthode d'Interpolation 4 : K-Nearest Neighbors (kNN)

3.2.1 Principe de Fonctionnement

La méthode d'interpolation utilisant l'algorithme des K plus proches voisins, ou **K-Nearest Neighbors** (kNN), est une approche non paramétrique largement utilisée en analyse de données spatiales. Contrairement à d'autres techniques d'interpolation, KNN ne repose pas sur des hypothèses précises sur la distribution des données, mais sur la proximité spatiale entre les points.

Le principe de base du KNN consiste à estimer la valeur d'un point inconnu en prenant la moyenne (ou parfois une pondération) des valeurs des K points les plus proches de ce point. La distance utilisée pour définir les plus proches voisins est souvent la distance euclidienne, mais d'autres métriques peuvent être appliquées en fonction du contexte.

L'estimation de la valeur d'un point peut se formuler mathématiquement de la manière suivante :

$$z(x) = \frac{1}{K} \sum_{i=1}^K z_i$$

où z_i représente les valeurs des K points les plus proches du point x . Cette méthode est simple à implémenter et peut s'adapter à différents types de données spatiales.

3.2.2 Résultats des Simulations

Localisation des 100 points de mesure sur la carte brute

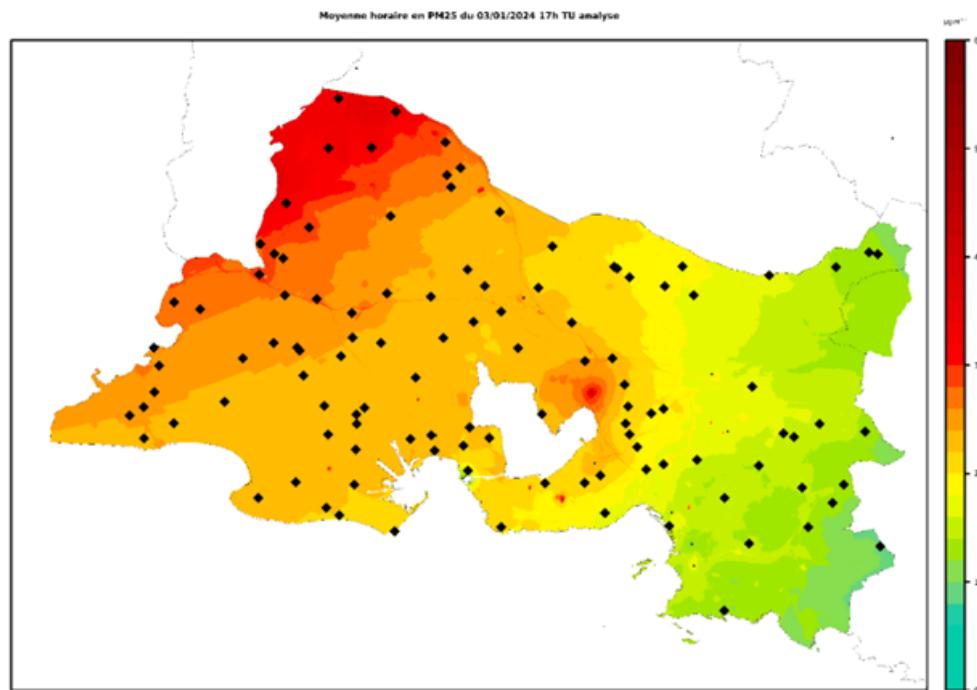


Fig. 3.5 – Carte brute avec 100 points de mesure.

Résultats de la méthode kNN, avec $k = 8$

1. Carte des écarts spatialisés calculés avec la méthode kNN

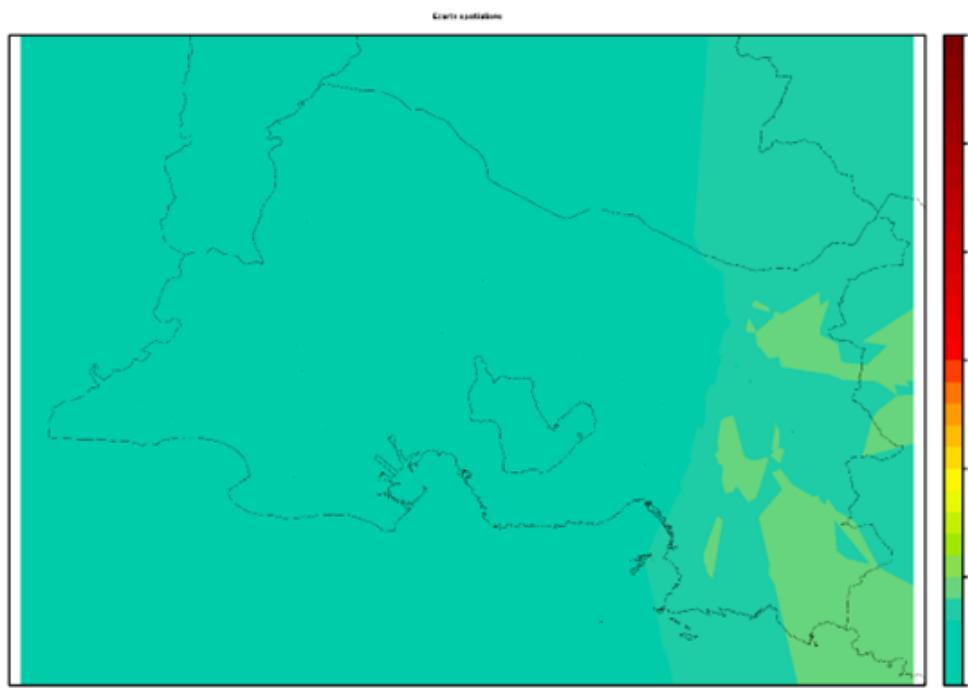


Fig. 3.6 – Carte des écarts calculés avec la méthode kNN, avec 100 points de mesure.

2. Carte d'interpolation avec la méthode kNN

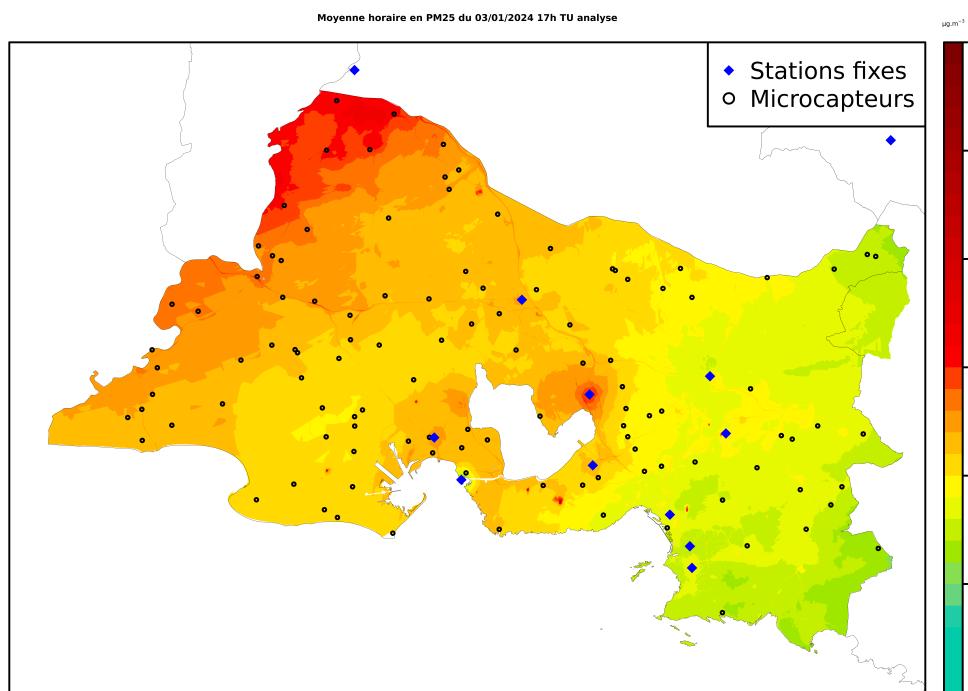


Fig. 3.7 – Carte d'interpolation avec 100 points de mesure.

Résultats de la méthode kNN, avec $k = 1$

1. Carte d'interpolation calculée avec 100 points de mesure

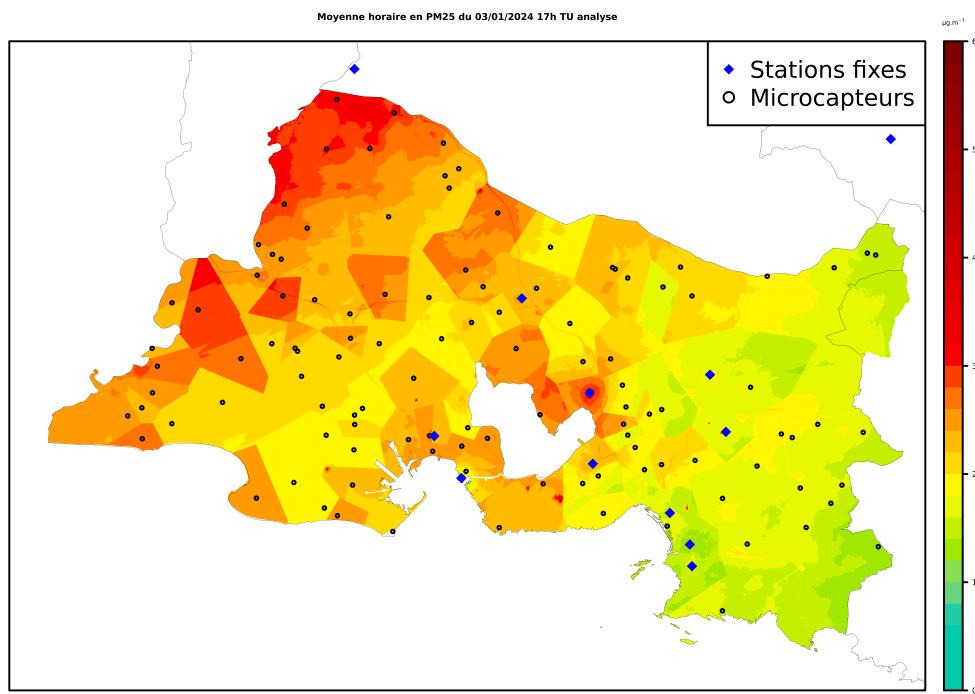


Fig. 3.8 – Carte d'interpolation corrigée avec 100 points de mesure ($k=1$)

2. Carte d'interpolation calculée avec 500 pts de mesure.

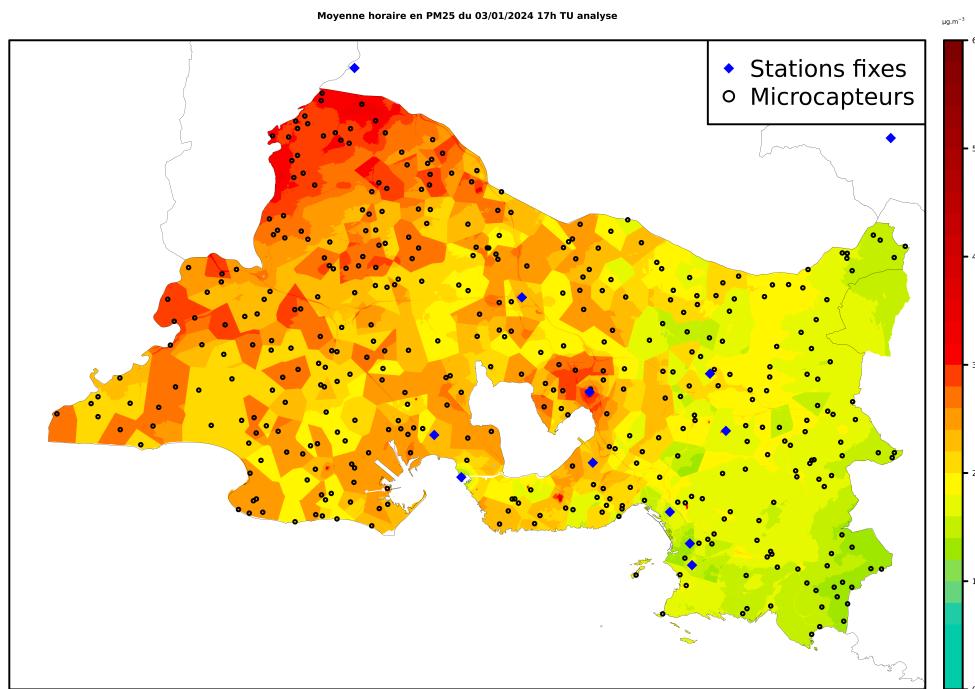


Fig. 3.9 – Carte d'interpolation avec 500 points de mesure ($k=1$)

Analyse des résultats

Lorsqu'on utilise $k=8$ dans l'algorithme kNN (voir Figure 3.10b), les valeurs interpolées aux points des stations fixes ne correspondent pas aux mesures réelles, ce qui montre que ce n'est pas une véritable interpolation. En revanche, avec $k=1$ (voir Figure 3.10a), on obtient les valeurs exactes des mesures, mais au prix d'un lissage insuffisant, rendant les résultats moins homogènes.

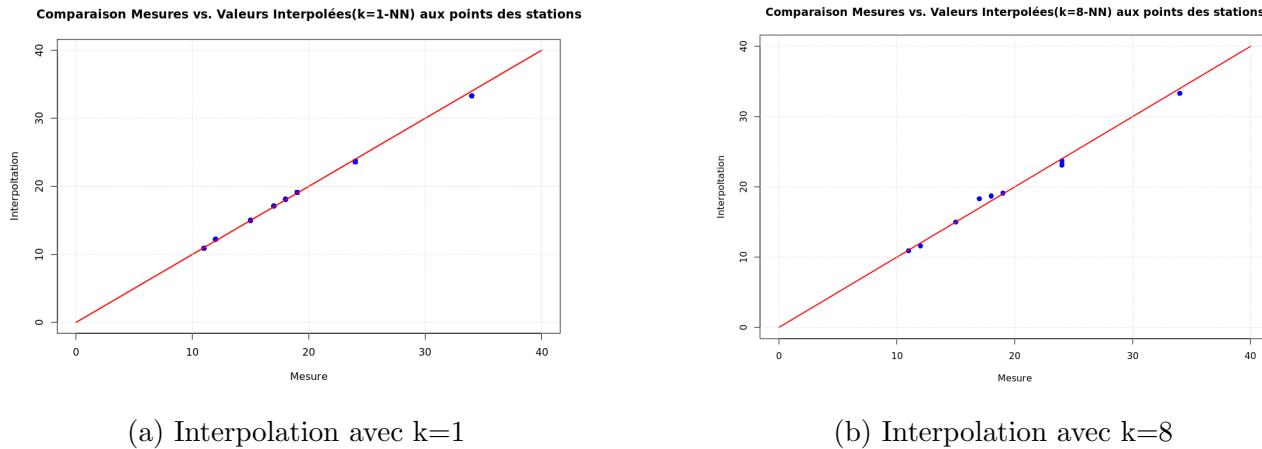


Fig. 3.10 – Comparaison des valeurs interpolées aux points des stations fixes pour $k=1$ et $k=8$.

3.3 Interpolation avec la Fonction interp

Principe de Fonctionnement

La méthode d'interpolation utilisant la fonction `interp` est disponible dans divers logiciels de traitement des données spatiales, tels que le package ‘`interp`’ en R. Cette méthode est souvent utilisée pour interpoler des données à partir de points de mesure dispersés sur une grille régulière.

La fonction `interp` permet d'effectuer des interpolations bidimensionnelles en utilisant diverses méthodes, telles que linéaire, spline, ou inverse distance weighting (IDW). Dans ce contexte, nous utilisons la méthode linéaire, où la valeur estimée en un point est calculée comme une combinaison linéaire des valeurs des points voisins.

La formule générale de l'interpolation linéaire est donnée par :

$$z(x) = \frac{\sum_{i=1}^n w_i \cdot z_i}{\sum_{i=1}^n w_i}$$

où w_i est le poids attribué à la valeur z_i du point i , généralement basé sur la distance au point d'interpolation.

Résultats des Simulations

Les simulations réalisées avec la fonction `interp` en utilisant la méthode linéaire ont montré que cette approche produit des résultats relativement lisses et cohérents, même avec un nombre

réduit de points de mesure.

Localisation des 40 points de mesure sur la carte brute

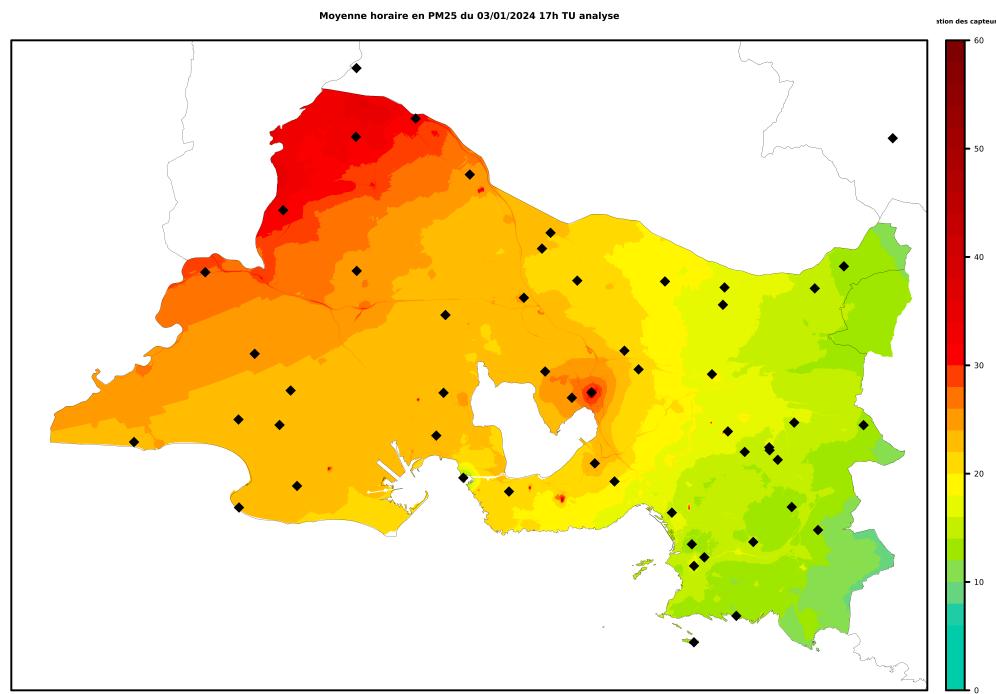


Fig. 3.11 – Carte brute avec 40 points de mesure.

Carte des écarts calculés avec la fonction 'interp'

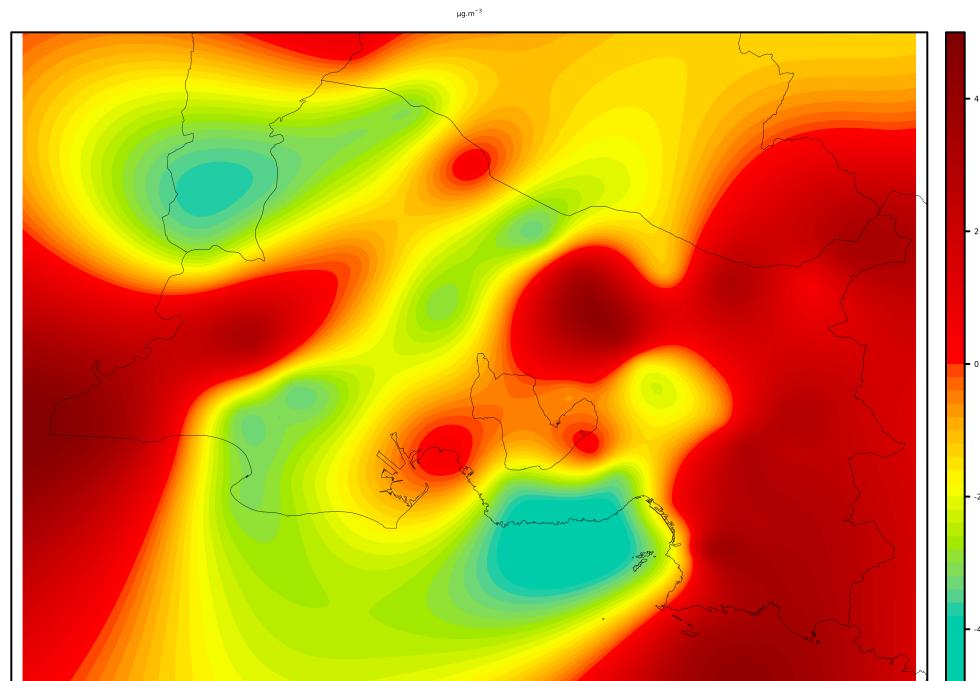


Fig. 3.12 – Carte des écarts calculés avec "interp" avec 40 points de mesure.

Carte d'interpolation avec la fonction "interp" : :interp"

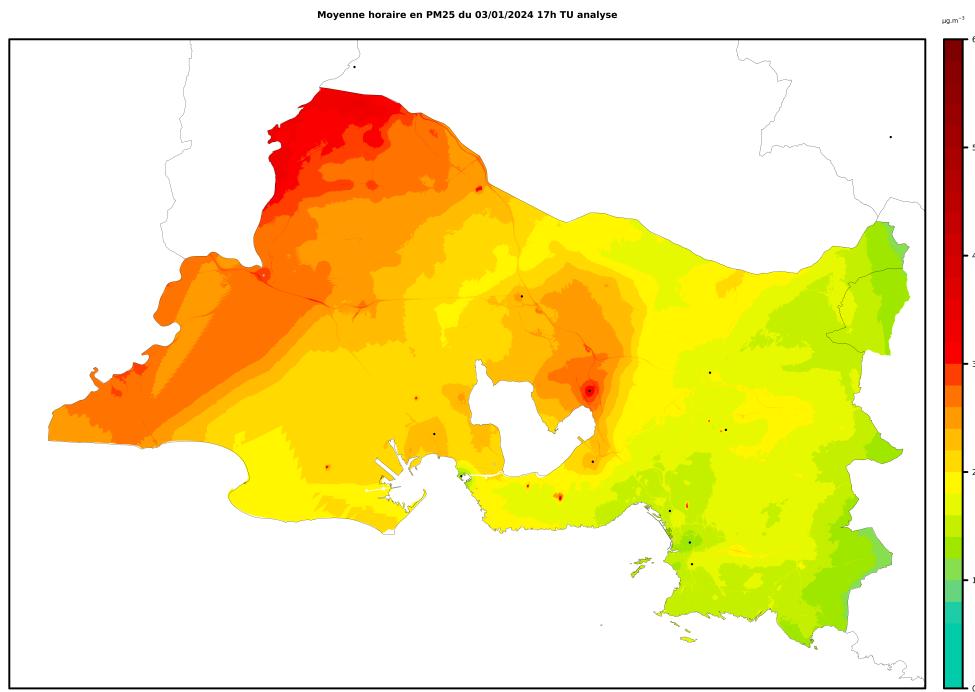


Fig. 3.13 – Carte d'interpolation avec 40 points de mesure.

Analyse des résultats

La méthode `interp`, en plus d'être plus rapide, présente un avantage significatif : aux points des stations fixes, les valeurs interpolées correspondent exactement aux mesures réelles (voir figure 3.14). De plus, elle génère moins d'artefacts que l'algorithme KNN, offrant ainsi des résultats plus cohérents et lissés.

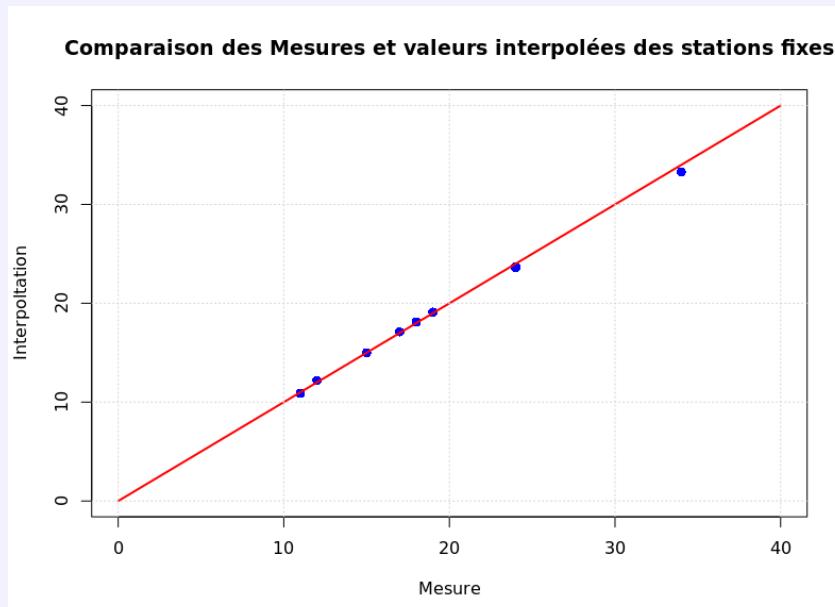


Fig. 3.14 – Graphique illustrant la relation entre les mesures réelles effectuées aux stations fixes et les valeurs obtenues par interpolation à ces mêmes points.

3.4 Comparaison des Méthodes

3.4.1 Analyse Comparative des Résultats

Méthode	Microcapteurs	Temps	Lissage
IDW	10	14.51 min	Performant
Interp	10	20.12 s	Performant
kNN	10	29.08 s	Pas lisse
Kriging	100	15.61 min	Performant
<code>interp</code>	100	40.92 s	Effets de triangles légers
KNN ($k = 1$)	100	1.25 min	Pas lisse (triangles)
IDW	1000	5 min	Bien lissée
<code>interp</code>	1000	6.21 min	Pas lisse
KNN ($k = 1$)	1000	5.21 min	Pas lisse (triangles)

Tab. 3.1 – Comparaison des méthodes d'interpolation selon le nombre de microcapteurs, le temps d'exécution et le lissage visuel.

3.4.2 Évaluation des Avantages et Inconvénients

Analyse Finale des Méthodes

- **Maximiser le lissage** : Pour obtenir un lissage optimal, la méthode IDW est recommandée. Elle assure à la fois l'absence d'artefacts et l'exactitude aux points des stations fixes. Cependant, elle peut être plus lente lorsque le nombre de points de mesure augmente.
- **Interpolation précise avec temps d'exécution optimisé** : Si l'objectif est de garantir l'exactitude des valeurs aux points des stations tout en conservant un temps de calcul raisonnable, la méthode `interp` est la plus adaptée. Elle combine une bonne rapidité d'exécution avec des résultats précis aux points de mesure, surpassant ainsi le krigeage en termes de performance.
- **Minimiser le temps d'exécution** : Pour une exécution la plus rapide, KNN avec $k = 1$ est la méthode à privilégier. Toutefois, ce compromis sur la vitesse engendre un manque de lissage visuel et des effets de triangles visibles dans les interpolations.
- **Exactitude aux points des stations** : Si l'exactitude aux points des stations est primordiale, les méthodes IDW, `interp`, et KNN avec $k = 1$ sont toutes appropriées. Chacune de ces méthodes présente cependant des compromis en termes de lissage et de temps d'exécution.

3.5 Conclusion

Après avoir comparé les méthodes d'interpolation, la méthode `interp` s'avère être la meilleure pour notre usage. Elle garantit à la fois une bonne précision aux points de stations fixes et un

temps de calcul raisonnable. Cependant, il faut noter que le lissage peut diminuer lorsque le nombre de points augmente beaucoup, ce qui peut affecter la qualité des cartes.

D'autres méthodes, comme l'IDW et le krigeage, offrent des résultats lisses et précis, mais peuvent être plus lentes, surtout avec beaucoup de points. La méthode KNN avec $k=1$ est très rapide, mais le lissage est moins efficace.

Il reste à explorer des moyens d'améliorer le compromis entre précision, lissage et temps d'exécution. Des recherches futures pourraient se concentrer sur l'ajustement des paramètres ou sur de nouvelles méthodes d'interpolation adaptées à nos besoins.

Conclusion générale

Ce projet, réalisé pendant mon alternance en tant que data scientist chez AtmoSud, avait pour but d'améliorer l'utilisation des données collectées par des microcapteurs mesurant les particules fines (PM10, PM2.5). J'ai travaillé sur l'optimisation du traitement de ces données en les corrigeant et en les analysant pour assurer leur fiabilité et leur intégration dans des modèles de prévision.

J'ai utilisé plusieurs méthodes de machine learning, comme la régression simple, la régression multiple en tenant compte de l'humidité relative, et les modèles de Support Vector Machines (SVM). La régression simple s'est révélée la plus efficace pour réduire l'erreur des mesures. Grâce à ces améliorations, les données des microcapteurs sont devenues plus proches de celles des stations de référence, ce qui rend les données plus fiables.

Une partie importante de mon travail a été de corriger les données en temps réel en utilisant des ratios sur 3 à 24 heures. Cela a amélioré la précision des mesures des microcapteurs en les alignant avec celles des stations de référence.

J'ai également testé différentes méthodes d'interpolation, comme IDW et KNN, pour créer des cartes détaillées des concentrations de particules dans la région PACA. Ces cartes horaires, avec une résolution de 25 mètres, offrent des visualisations précises et utiles pour les acteurs locaux.

Cette expérience a été vraiment précieuse pour mon développement professionnel. En travaillant sur tous les projets de Data Science, j'ai acquis une expérience pratique de toutes les étapes du travail, à savoir extraire et nettoyer les données, appliquer les algorithmes de machine learning, visualiser les données dans RStudio, QGIS, ggplot2, et d'autres encore. Mon expérience d'extraction de données à l'aide de SQL et d'API dans d'autres domaines m'a également aidé à garantir la qualité et la cohérence des données. L'encadrement d'un expert en data science m'a aidé à structurer mes analyses et à présenter mes résultats de manière formelle lors des réunions d'équipe. Cette alternance a ainsi consolidé mes compétences techniques tout en me confortant dans mon choix de carrière, en me permettant d'évoluer dans un environnement stimulant, orienté vers les défis de la Data Science.

Annexe A - Extrait du rapport automatisé avec CSS

Microcapteurs

Introduction
Contexte des données
La première correction

- Coefficients des modèles
- Les graphiques
- Performances après la première correction
- Comparaison du RMSE Avant et Après Correction
- La correction au fil de l'eau (en horraire)
- La correction au fil de l'eau en journalier
- La correction 2 sur toute la période
- Annexes

Racha Amina DJAGHLOUL
18 juin, 2024

La première correction

La première correction est effectuée à l'aide d'un modèle linéaire simple, ajustée sur l'échantillon d'entraînement.

Formule de la première correction :

$$Y = a \cdot (x^{\text{exposant}}) + b$$

- Y : Mesure de la station
- x : Mesure brute du capteur
- a : Coefficient de pente
- exposant : Exposant du modèle
- b : Terme constant ajusté sur l'échantillon d'entraînement

Coefficients des modèles

Les coefficients des modèles obtenus après la première correction sont les suivants :

station	capteur	ville	exposant	b	a
2041	BD6BAF	Salon	0.5	-6.41	6.36
2043	BD6AA4	Marignane	1.0	1.18	0.93
3029	BD6BAA	Aix_art	1.0	0.78	0.89
3043	BD6ABC	Marseille_cinq_av	1.1	2.02	0.48
3071	BD6B8B	Toulon_claret	1.1	0.58	0.75
24035	C19D41	Nice_magnan	0.6	-2.51	3.70
24036	C19D40	Nice_arson	0.7	-3.04	3.43

Les graphiques

Les graphiques suivants illustrent les résultats des corrections apportées par les modèles sur les mesures des capteurs par rapport aux références. Chaque graphique représente une station de mesure spécifique, montrant les mesures brutes en gris, les corrections appliquées sauf pour la période QA/QC en rouge, et les corrections spécifiquement pour la période QA/QC en bleu. Les lignes vertes représentent les seuils de qualité de l'air. Les informations sur les performances des modèles, telles que le NRMSE et l'amélioration en pourcentage, sont également fournies en bas de chaque graphique.

Fig. 15 –

Annexe B - Résultats sur l'interface µspot

```
# liste des stations #
station_fixe <- c(24035,3029,3043,3021,2043,2041,3071,3068,24036)
dat1 <- "2023-10-10 00:00:00"
dat2 <- "2024-02-10 23:00:00"

# requête #
conn <- dbConnect(drv, host=host, user=user, pass=pass, dbname=dbname)
dataMes_station <- dbGetQuery(conn, statement=
  paste(
    "
      SELECT dh.val,station
      FROM `",table_station,"`
      WHERE dh>=",dat1,'" AND dh<=",dat2,'"'
      AND station IN
      ",sep="")
  )
  (",paste(station_fixe,collapse="`","`",sep=""),"")
  pol='',code_pol,''
)
dbDisconnect(conn)
names(dataMes_station) <- c("date","mesure", "station")
```

Fig. 16 – Requête d'extraction des mesures des stations sur R

Annexe C - Résultats sur l'interface µspot

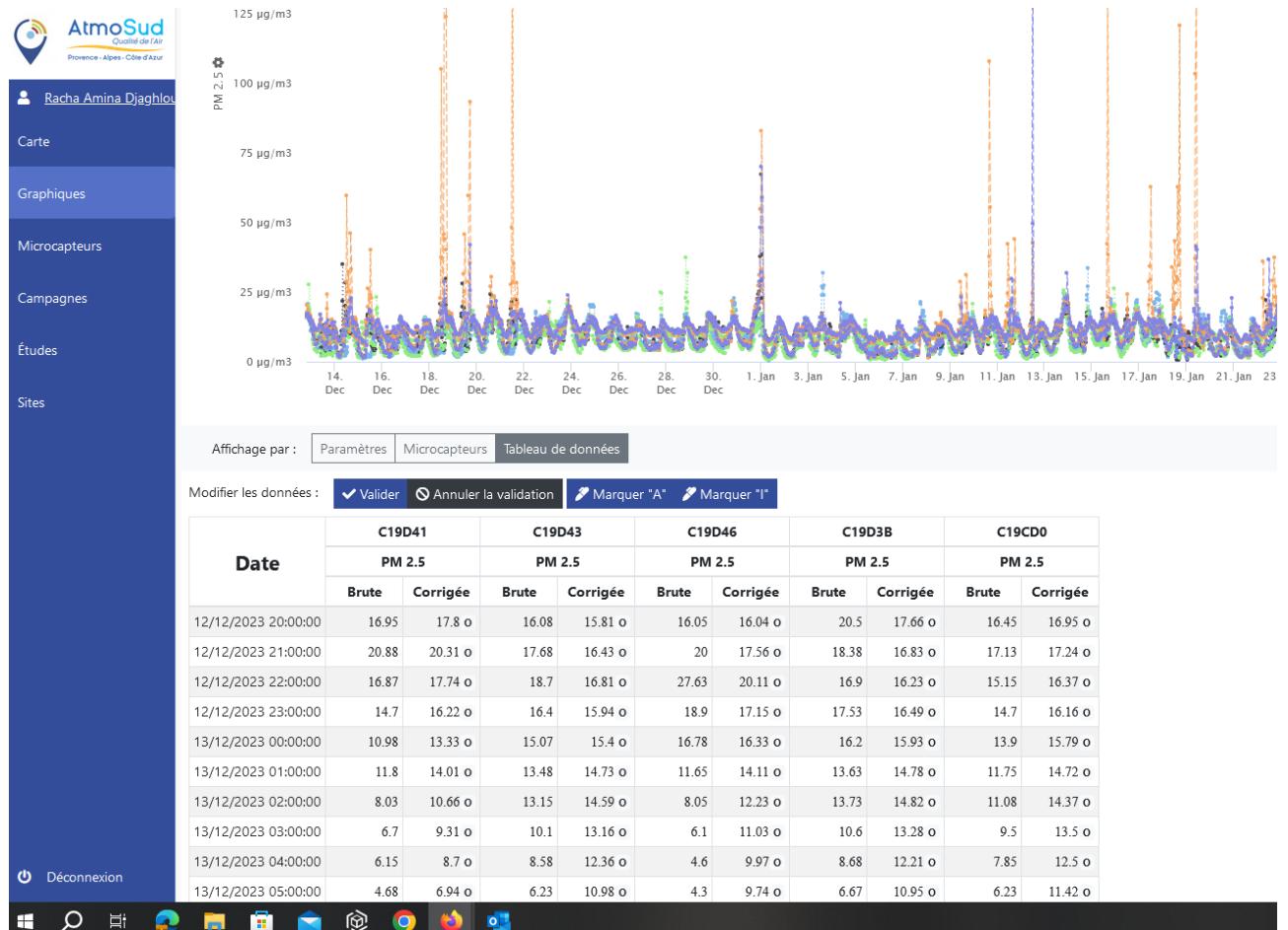


Fig. 17 —

Annexe D - Bulle d'air

- L'expertise AtmoSud des données microcapteurs : une méthodologie en constante amélioration

Pour améliorer la qualité des mesures de PM10 et PM2.5 des micro-capteurs, nous avons développé une méthode de correction en deux étapes pour les corriger.

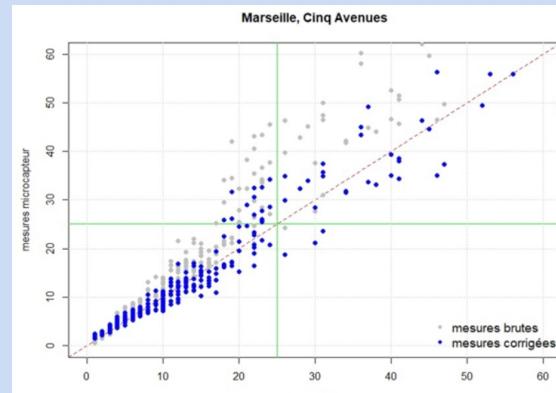
Correction basée sur une inter-comparaison
 Nous comparons les données horaires de nos microcapteurs avec celles des stations de référence. Cette comparaison se fait sur des périodes de temps similaires et est appelée phase de QA/QC (Quality Assurance/Quality Control). Elle permet de corriger les mesures à l'aide d'un modèle polynomial, dont la formulation est la suivante :

$$Y = a \cdot (x^\alpha) + b$$

Y : Mesure à la station de référence
 x : Mesure brute du microcapteur
 a : Coefficient de pente
 α : Exposant du modèle
 b : Terme constant

Correction au fil de l'eau : un couple microcapteur/station de référence est positionné en un même lieu. À partir de ce couple, nous calculons la relation entre la mesure horaire du microcapteur et celle de la station de référence. Nous utilisons cette relation pour corriger les mesures de tous les microcapteurs déployés à proximité jusqu'à plusieurs kilomètres de distance.

Afin de valider cette méthode en deux étapes, nous avons étudié quatre couples microcapteur/station dans les villes suivantes : Marseille, Aix, Marignane et Salon-de-Provence. Nous avons vérifié que la correction calculée dans une ville donnée et appliquée dans les autres villes, améliore la qualité des mesures de microcapteurs.



Comparaison des mesures brutes et corrigées d'un microcapteur à Marseille, montrant les effets de la méthode de correction

Cette étude a montré qu'un capteur de référence d'une ville peut corriger l'ensemble des capteurs déployés à plusieurs kilomètres de distance, qu'il soit en situation de trafic ou de fond.

Fig. 18 –

Bibliographie

- [1] Atmosud. (2023). *Site Web d'Atmosud*. <https://www.atmosud.org>
- [2] Meersens. (2023). *Le dioxyde d'azote (NO₂) : Sources et impacts sur la santé*. <https://meersens.com/le-dioxyde-dazote-no2-sources-et-impacts-sur-la-sante/>. Consulté le 26/06/2023.
- [3] Organisation Mondiale de la Santé (OMS). *Qualité de l'air ambiant et santé*. [https://www.who.int/fr/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/fr/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health). Consulté le 26/06/2023.
- [4] Guyader, A. (2013). *Régression linéaire*. Pages 1 à 11. reglin Kutner2005 StatQuest
- [5] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). Wiley. Extrait de https://users.stat.ufl.edu/~winner/sta4211/ALSM_5Ed_Kutner.pdf.
- [6] StatQuest with Josh Starmer. (2021). *Simple Linear Regression*. <https://statquest.org/statquest-simple-linear-regression/>
- [7] GISGeography. (2024). *Inverse Distance Weighting (IDW) Interpolation*. <https://gisgeography.com/inverse-distance-weighting-idw-interpolation/>
- [8] Interp Package. (2024). *Interpolation and Smoothing of Data*. <https://www.rdocumentation.org/packages/interp/versions/latest>