

ZARROUK Moez

Promotion 2024

Année universitaire 2023/2024

Diplôme d'ingénieur Télécom Physique Strasbourg
Mémoire de stage de 3ème année

**Evaluation des performances des capteurs de pollution
atmosphérique à faible coût**



Inria, laboratoire CITI
56 Boulevard Niels Bohr,
69100 Villeurbanne
<https://www.inria.fr/fr>

 **citi** Lab

BOUBRIMA Ahmed
+33 6 13 89 10 00
Du 18 Mars 2024
au 31 Août 2024

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué à la réalisation de mon stage.

Tout d'abord, je remercie chaleureusement mes directeurs de stage, RIVANO Hervé et BOUBRIMA Ahmed , pour leur encadrement enrichissant et leur soutien constant tout au long de cette expérience.

Je tiens également à remercier toute l'équipe Agora à INRIA pour leur accueil chaleureux, leur disponibilité, et les nombreuses discussions enrichissantes.

Je remercie également mon tuteur école Madame MEILLIER Céline pour son soutien et ses conseils pendant le stage.

Enfin, je souhaite exprimer ma reconnaissance à ma famille et à mes amis pour leur soutien moral et leur encouragement constant.

Table des matières

Liste des figures	vi
Liste des tableaux	viii
1 Introduction	1
2 Environnement et contexte de la mission	3
2.1 Présentation de l’Inria	3
2.2 Centre Inria de Lyon	4
2.3 Inria, INSA-Lyon et laboratoire CITI	5
2.4 Equipe Agora	6
2.5 Horaires	8
3 État de l’Art	9
3.1 Les particules polluantes	9
3.1.1 Définition	9
3.1.2 Unités de mesures	10
3.2 Méthodes de mesure des particules polluantes	11
3.2.1 Les méthodes de mesure de référence	11
3.2.2 Beta Attenuation Method (BAM)	13
3.2.3 Le fonctionnement de la BAM	13
3.2.4 Les capteurs <i>PM</i> à coût faible	15
3.3 Évaluation des performances des micro-capteurs	17
3.4 Métriques de performance et mesures statistiques	18
3.4.1 Le coefficient de détermination r^2	18
3.4.2 RMSE et MAE	19
3.4.3 Erreur moyenne de biais : MBE	20
3.4.4 Coefficient de variation : CV	20
3.5 Analyse des données	21
3.5.1 Préparation des données	21

3.5.2	Relation entre les capteurs à faible coût et le capteur de référence	21
3.5.3	Variabilité intra-modèle	23
3.5.4	Reproductibilité des capteurs	23
3.5.5	Impact de la température et de l'humidité relative	23
3.6	Techniques de détection des anomalies des micro-capteurs	25
3.6.1	Approches basées sur les mesures des capteurs	25
3.6.2	Approches basées sur la consommation des capteurs	26
3.7	Méthodologie retenue	27
4	Analyse préliminaire des mesures de pollution	28
4.1	Jeu de donnée	28
4.1.1	API	28
4.1.2	Site utilisé	30
4.2	Prétraitement des données	31
4.3	Analyse préliminaire des données	33
5	Étude comparative des méthodes de calibration	37
5.1	Paramètres de calibration	37
5.2	Processus de calibration	38
5.3	Méthodes de calibration implémentées	39
5.3.1	Régression linéaire simple (RLS)	39
5.3.2	Régression linéaire multiple (RLM)	39
5.3.3	Régression polynomiale multivariable (RP)	40
5.3.4	Random Forest (RF)	41
5.3.5	Gradboost	42
5.4	Résultats de comparaison	42
6	Analyse de l'impact du vieillissement	46
6.1	Modélisation mathématique	46
6.2	Évolution des performances des fonctions de calibration	47
6.3	Comparaison du vieillissement de différents capteurs	50
7	Conclusion	53

Table des figures

Figure 2.1	Répartition des équipes Inria Lyon	4
Figure 2.2	Organigramme du centre Inria Lyon	5
Figure 3.1	Les types de particules polluantes	10
Figure 3.2	BAM 1020	14
Figure 3.3	Exemple d'un micro-capteur de pollution	16
Figure 3.4	Principe de fonctionnement des capteurs à faible coût des <i>PM</i> .	17
Figure 3.5	Les caractéristiques des capteurs à faible coût issus de fabricants différents et utilisant la même technique de mesure.	22
Figure 3.6	L'impact de l'humidité relative sur 12 capteurs à faibles coût. .	24
Figure 4.1	Extraction des données	30
Figure 4.2	Répartition des mesures en 2021	32
Figure 4.3	Distribution des valeurs	34
Figure 4.4	Comparaison des distributions en Février et Juin 2021	35
Figure 5.1	Résultats des différents modèles de calibration sur le capteur 1 .	45
Figure 6.1	Evolution de R^2 et RMSE en fonction des mois	48
Figure 6.2	Analyse des performances des capteurs en fonction de la corrélation avec l'humidité.	49
Figure 6.3	Comparaison du vieillissement de différents capteurs	51

Figure A.1	Concentrations mesurées par les capteurs 2 et 3 pendant la dernière semaine de février 2021	56
Figure A.2	Courbe d'apprentissage du modèle Random Forest	56

Liste des tableaux

Table 3.1	Caractéristiques des méthodes de mesure de la pollution de l'air.	12
Table 4.1	Id des polluants et des données météorologiques	29
Table 4.2	Id des appareils de mesure	29
Table 4.3	Statistiques descriptives des mesures de capteurs pour une période de 3 ans	34
Table 4.4	Performance des Modèles par Mois et Année	36
Table 5.1	Performance des modèles pour les trois capteurs	43

1 Introduction

Ce stage de fin d'étude s'inscrit dans le cadre de l'amélioration de la qualité d'air dans la région Auvergne Rhône Alpes et la région Île de France. La pollution de l'air a un impact important sur l'environnement et surtout sur la santé [1]. C'est pour cette raison que la surveillance de la qualité de l'air est exigée par les agences environnementales [2], les régions et les états afin d'évaluer l'exposition environnementale de la population à de multiples polluants de l'air. L'instrument de référence utilisé pour répondre aux normes internationales en matière de mesure de la qualité d'air se caractérise par un coût élevé et un niveau de maintenance important [1]. De plus, sa résolution spatiale et temporelle est insuffisante pour caractériser avec granularité fine les variations des concentrations des polluants atmosphériques. C'est pour cela que grâce aux avancements technologiques, les agences environnementales et les régions comme Auvergne Rhône Alpes et Île de France commencent à s'intéresser à des micro-capteurs à forte résolution temporelle et spatiale (grâce à leur faible coût). Les capteurs à faible coût ont le potentiel de compléter et d'étendre les capacités des réseaux existants de surveillance de l'air ambiant. Ils permettent aussi de fournir des mesures à l'échelle du quartier (grâce à leur coût très faible), ce qui contribue à l'amélioration de la prise de décision concernant la réduction des effets de la pollution atmosphérique [3]. Cependant, les capteurs à faible coût sont susceptibles la plupart du temps à fournir des mesures imprécises et loin de décrire la réalité. L'évaluation des performances des capteurs de particules polluantes est le principal sujet d'intérêt de ce stage. En effet, les particules polluantes et principalement les particules fines

sont très néfastes à la fois pour la santé et pour l'environnement. L'objectif de ce stage est de déterminer la capacité des capteurs à faible coût à décrire le niveau de pollution de l'air en étant le plus proche de la réalité. Cela revient à analyser leurs performances en fonction de tous les facteurs pouvant les impacter (météo, encrassement, vieillissement, etc.). Dans ce rapport, nous allons dans un premier lieu aborder les généralités relatives à la mesure des particules polluantes (définitions, unités de mesure, etc.). Puis, nous récapitulons les méthodes de fonctionnement des moniteurs de référence et des capteurs à faible coût tout en abordant les limites de chaque technologie de mesure. Ensuite, nous aborderons les différentes techniques de calibration, qui seront comparés entre eux. Enfin, nous présenterons les différents tests effectués afin d'analyser le vieillissement des micro-capteurs.

2 Environnement et contexte de la mission

2.1 Présentation de l’Inria

L’Inria est l’Institut National de Recherche en Informatique et en Automatique. Il a été fondé en 1967 dans le cadre du « plan calcul » du gouvernement sous le nom d’INRIA. Il est sous la tutelle du ministère de la Recherche et du ministère de l’Industrie. L’Inria est un établissement public à caractère scientifique et technologique et possède 10 centres de recherche. Ils sont situés pour la plupart dans des campus en France mais l’un d’entre eux est aussi situé au Chili. Le siège de l’Inria est situé à Rocquencourt à proximité de Versailles.

L’Inria fonctionne avec un modèle d’équipes-projet constituées d’environ une vingtaine à une trentaine de membres rassemblés autour d’un responsable scientifique. Aujourd’hui il y a 225 équipes-projet répartis sur les 10 centres de l’Inria. Chaque équipe-projet dispose d’une autonomie scientifique et financière, avec un budget composé de ressources attribuées par le centre et de "ressources propres" provenant de subventions régionales, nationales, européennes et de contrats avec des entreprises. Les équipes travaillent fréquemment en collaboration avec des établissements partenaires à l’Inria.

2.2 Centre Inria de Lyon

Le centre Inria de Lyon a été créé le 1er janvier 2022. Il est membre associé de l’Université de Lyon. Ses équipes de recherche sont communes avec l’ENS de Lyon, l’Université Claude Bernard Lyon1, l’INSA Lyon, le CNRS et l’INRAE. Le centre de Lyon héberge 300 personnes dont 130 employées par l’Inria réparties sur 17 équipes de recherche. Le centre a aussi la particularité d’accompagner la création de startups et 9 ont été créées depuis 2005. Certaines équipes sont bilocalisées entre les centres de Lyon et Grenoble. Le centre de Lyon est spécialisé dans les recherches sur l’intelligence artificielle, la modélisation en biologie et santé, les réseaux et la communication ainsi que les systèmes embarqués et leur architecture.

Domaines	Bilocalisées	Lyon
Mathématiques appliquées, Calcul et Simulation		
Algorithmique, Programmation, Logiciels et Architectures	PRIVATICS QINFO	ARIC, CASH
Réseaux, Systèmes et Services, Calcul distribué		AGORA, AVALON, OCKHAM, MARACAS, ROMA, EMERAUDE
Perception, Cognition, Interaction	CHROMA	
Santé, Biologie, Planète numériques		BEAGLE, DRACULA, ERABLE, MOSAIC, AISTROSIGHT

Figure 2.1 : Répartition des équipes Inria Lyon

ORGANIGRAMME DU CENTRE INRIA DE LYON (au 1^{er} mai 2024)

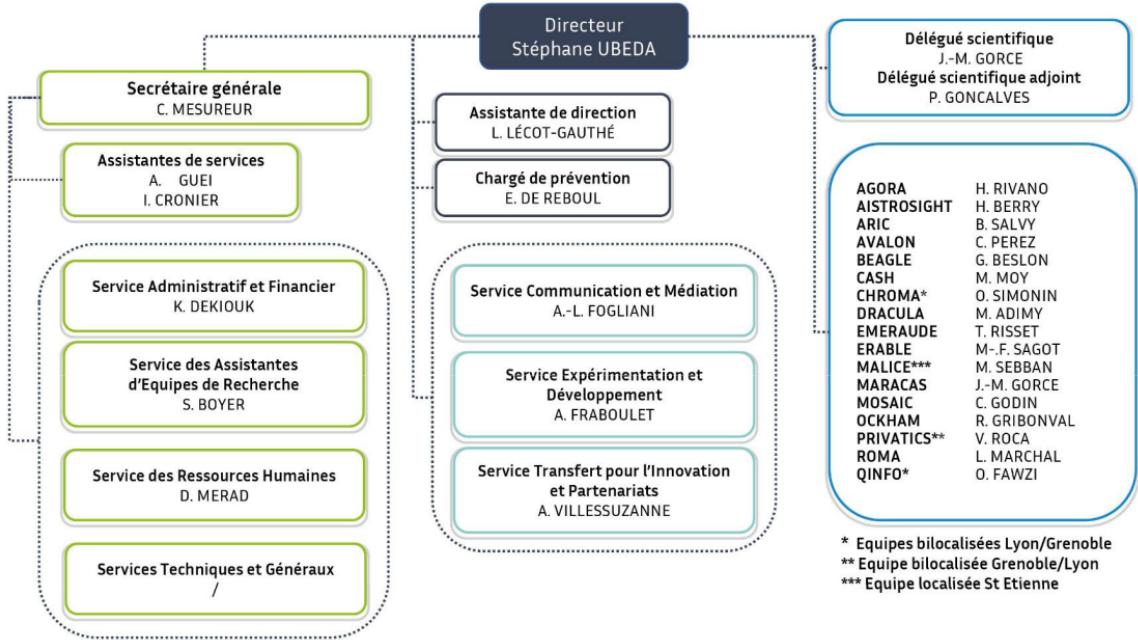


Figure 2.2 : Organigramme du centre Inria Lyon

2.3 Inria, INSA-Lyon et laboratoire CITI

J'ai été recruté pour mon stage au sein de l'équipe Agora (Algorithmes et Optimisation de Réseaux Autonomes). L'équipe Agora fait partie des 8 équipes structurant le laboratoire CITI : Chroma, Dynamid, Emeraude, Macaras, Phenix, Privatics et Wired. Le laboratoire CITI a été fondé en 2001 au sein de l'Insa Lyon. Le laboratoire est par la suite passé sous double tutelle de l'INSA et l'Inria. Grâce à cette tutelle, je suis recruté et financé sur des fonds de l'Inria. Le laboratoire CITI est réparti sur 3 sites sur Lyon : le bâtiment de l'Inria à la Doua où j'ai effectué mon stage, le bâtiment Chappe de l'INSA et les locaux de l'institut Rhônalpin des Systèmes Complexes (IXXI) de l'ENS.

2.4 Equipe Agora

L'équipe Agora est spécialisée dans la recherche sur les télécommunications et les réseaux sansfil. Les principales thématiques de recherche d'Agora sont :

- Le déploiement de réseaux sans fil via le développement de modèles d'optimisation avec une attention particulière aux réseaux de capteurs et les communications par satellite.
- La collecte des données des réseaux sans fil.
- L'exploitation des données des réseaux sans fil.

L'équipe Agora a plus d'une trentaine de membres dont 2 chercheurs permanents, 5 enseignants-chercheurs permanents, 12 doctorants, 2 post-doctorants, 4 ingénieurs et 3 stagiaires dont moi. Le chef actuel de l'équipe-projet Agora est Prof. Hervé Rivano. Les différents membres de l'équipe travaillent sur plusieurs projets qui sont en lien avec une ou plusieurs thématiques parmi celles susmentionnées. Les projets actuels principaux de l'équipe sont notamment :

- DRON-MAP : c'est un projet de réseaux de drones coopératifs conçus pour suivre des panaches de pollution lors de situations d'urgence telles que des accidents industriels ou des catastrophes naturelles. Le projet est coordonné par Dr. Walid Bechkit depuis 2021.
- DOLL : c'est un projet coordonné par Dr. Oana Iova ayant pour but d'améliorer la technologie de réseau LoRaWAN1 en proposant un nouveau protocole pour

le débit descendant. Un réseau LoRaWan est constitué d'équipements à basse consommation communiquant leurs données à une passerelle. Ce réseau de type étoile est cependant exploité en majorité avec des communications partant des nœuds vers la passerelle (liaison ascendante) car les capacités de faire de la communication descendante sont limitées. Le projet DOLL vise donc à mettre au point un protocole novateur pour renforcer l'efficacité des réseaux LoRaWAN.

- DEMON : coordonné par Dr. Razvan Stanica, le projet DEMON vise à transformer les architectures de réseaux cellulaires via l'implémentation de stations de base mobiles. Les objectifs du projet sont de démontrer les avantages d'un réseau auto-déployable face à des solutions dites traditionnelles. DEMON pourra ainsi contribuer dans des sujets innovants tels que les véhicules autonomes ou la virtualisation de réseaux.

Ces projets ont été soumis à l'Agence Nationale de Recherche (ANR) et ont reçu des financements de leur part afin qu'ils puissent être menés à bien. La soumission d'un projet à l'ANR est constituée de 2 phases :

1. La phase de pré-sélection : on soumet un premier document présentant le projet de manière simple en 4 à 6 pages.
2. La phase de sélection : on soumet cette fois un autre document expliquant le projet de manière détaillée en une vingtaine de pages.

En règle générale, un projet sur dix réussit à franchir la phase de sélection pour obtenir des financements. En cas de refus par le jury, il est toujours possible de soumettre à nouveau le projet ultérieurement.

2.5 Horaires

Les horaires de travail étaient fixes : de 8h30 à 12h30 et de 13h30 à 18h.

3 État de l'Art

Dans ce chapitre, nous allons examiner les méthodes de référence de mesure reconnues par les agences environnementales, ainsi que les méthodes utilisées par les capteurs à faible coût pour mesurer les particules polluantes.

3.1 Les particules polluantes

3.1.1 Définition

Les polluants de type particules (Particulate Matter ou PM) se composent d'un mélange de substances organiques et minérales sous forme liquide ou solide en suspension dans l'air. Elles sont regroupées par leur taille (longueur du diamètre de la particule). En effet, la taille détermine le taux de pénétration des particules dans les voies respiratoires. La figure 3.1 montre les différents types de particules polluantes.

Les deux principales classes de taille des particules fines sont PM_{10} , particules ayant un diamètre inférieur à 10 micromètres, et $PM_{2.5}$, particules ayant un diamètre inférieur à 2.5 micromètres. Ces deux classes, selon l'EPA (Environmental Protection Agency) et WHO (World Health Organisation), ont un impact néfaste sur la qualité de l'air ainsi que sur la santé humaine causant des maladies respiratoires et cardio-vasculaires, et des cancers pulmonaires. Cela est dû à leur composition chimique qui détermine la capacité des particules à réagir avec les organes cibles telles que les poumons. Cette composition chimique est extrêmement hétérogène avec des matières organiques telles

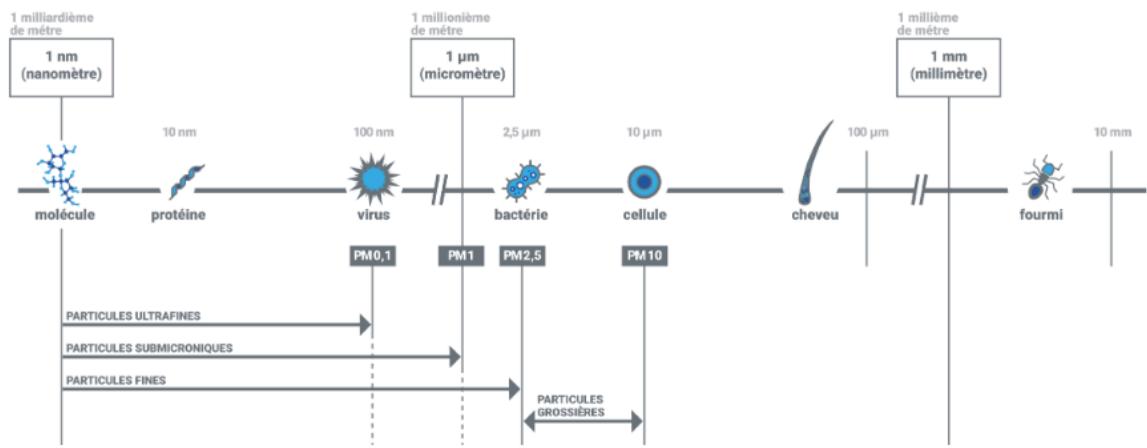


Figure 3.1 : Les types de particules polluantes

que le carbone noir, et inorganique telles que le nitrate et l'ammonium. Ces particules sont issues des émissions liées aux trafics routiers, chauffage au bois et la combustion de gaz et de fioul.

3.1.2 Unités de mesures

Les concentrations chimiques dans l'air sont généralement mesurées en unités de masse de la substance (micro-gramme ou gramme) par unité par volume d'air (mètre cube). Ces concentrations peuvent aussi être exprimées en ppb (parts of billion) grâce à la formule de conversion ci-dessous en se basant sur la masse molaire du polluant, la température et la pression atmosphérique.

$$\text{Concentration (ppb)} = 24.45 \times \frac{\text{Concentration } (\mu\text{g}/\text{m}^3)}{\text{Masse Molaire}}$$

Tel que $24.45 \text{ m}^3 \text{ mol}^{-1}$ est le volume molaire à 25°C et 1 atm.

Cependant, cette unité de mesure ne fonctionne pas avec les particules fine vu leur composition chimique, elles sont donc exprimées en $\mu\text{g m}^{-3}$ ou g m^{-3} .

3.2 Méthodes de mesure des particules polluantes

3.2.1 Les méthodes de mesure de référence

Les capteurs de pollution de référence sont reconnus par l'EPA grâce à leur conformité aux normes exigées par le NAAQ (National Ambient Air Quality). Ils reposent sur des méthodes de référence fédérales (FRM) pour garantir la cohérence et la fiabilité des données sur la qualité de l'air collectées par les agences gouvernementales et les organismes de réglementation. Ces méthodes s'appuient principalement sur des filtres pour collecter les particules fines en fonction de leur taille sur des filtres distincts. Chaque filtre est ensuite pesé (après l'équilibrage de l'humidité) avant et après le passage des particules pour déterminer la différence de masse due aux particules fines collectées. Le volume d'air est calculé à partir du débit d'air entrant par unité de temps. La concentration des particules est donc la masse totale divisée par le volume d'air, en micro-grammes par centimètre cube.

Malgré leur précision, ces méthodes ne sont pas optimales car elles permettent des mesures uniquement une fois toutes les 24 heures. De plus, leur coût élevé et la limitation dans le déploiement des capteurs sont des obstacles à la réalisation d'un suivi continu et en temps réel, essentiel pour prendre des décisions éclairées en vue

d'améliorer la qualité de l'air.

Il y a aussi d'autres méthodes de mesure de référence qui sont les méthodes fédérales équivalentes (FEM) et qui sont conçues pour être équivalentes aux FRM en termes de fiabilité et qui utilisent des technologies pour effectuer les mesures de concentration des particules. Ces méthodes permettent d'avoir des mesures en temps réel. Le tableau ci-dessous récapitule les méthodes de référence ainsi qu'un exemple de modèle de capteurs l'utilisant et leurs caractéristiques telles que la précision, le débit d'entrée d'air mesurée en litre par minute ($L \cdot min^{-1}$), la plage de température et de pression dans laquelle se font les mesures.

Méthode	Précision	Débit	Temp.	Pression	Capteur
Instack Particulate filtration		$14\text{-}25 L \cdot min^{-1}$	30-45°C	600-800mmHg	RAAS10-200
TEOM (Tapered element oscillating monitor)	$\pm 0.75\%$	$0.5\text{-}4.0 L \cdot min^{-1}$	8-25°C		TEOM1405-DF
Dichotomous Air Sampler		$10\text{-}19 L \cdot min^{-1}$	40-50°C		Thermo Scientific Partisol 2000-D
Beta Attenuation Method	$\pm 2\%$	$16.7 L \cdot min^{-1}$	10-40°C		BAM1020

Table 3.1 : Caractéristiques des méthodes de mesure de la pollution de l'air.

Dans la sous-section suivante nous détaillons la méthode de référence la plus utilisée, à savoir la Beta Attenuation Method (BAM).

3.2.2 Beta Attenuation Method (BAM)

La méthode d'atténuation bêta est l'une des techniques de mesure en temps réel les plus utilisées pour la surveillance des particules dans l'air ambiant. C'est une méthode fédérale équivalente en regard de l'utilisation des particules bêta pour mesurer la concentration des particules.

3.2.3 Le fonctionnement de la BAM

Une fois le flux d'air est rentré et filtré à travers une entrée à $16L \cdot min^{-1}$. Les particules sont donc piégées sur une bande filtrante. La source bêta qui est le ^{14}C , au-dessus du filtre, émet des rayons bêta ; ces particules sont atténuées et atteignent le détecteur à scintillation sensible aux rayons bêta après avoir traversé le ruban filtrant déposé sur les particules.

La figure 3.2 montre le schéma de fonctionnement du capteur "BAM 1020".

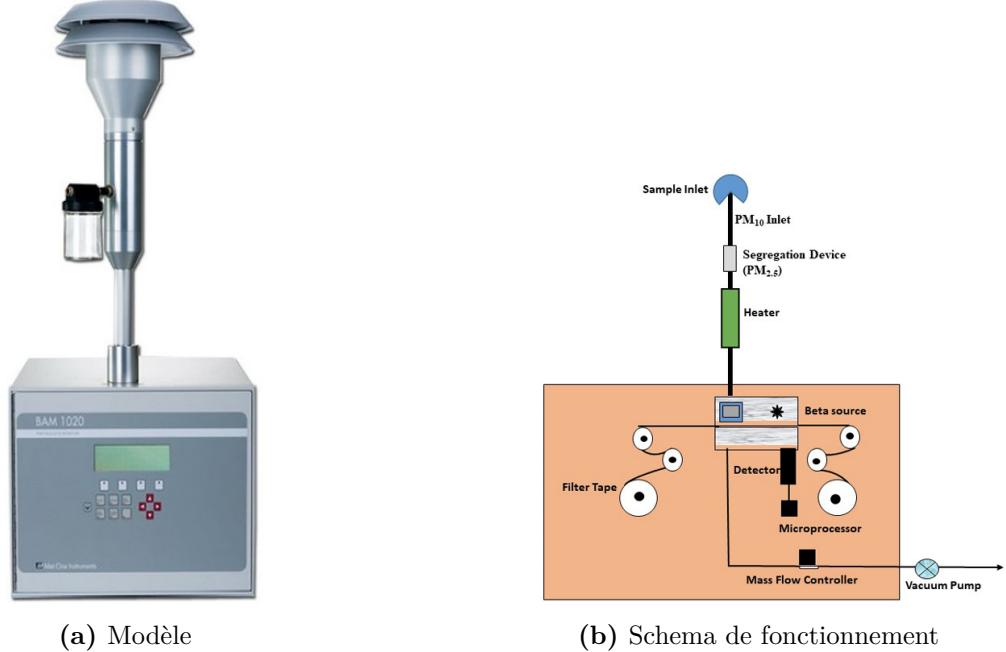
Le détecteur transforme la lumière en un signal électrique et calcule la concentration en micro-gramme par centimètre cube suivant cette formule : $C = \frac{A \ln\left(\frac{I_0}{I}\right)}{\mu Qt}$

C = Concentration en micro-gramme par cm^{-3}

A = Surface du filtre par laquelle le flux d'air est passé en cm^2

I = Intensité du rayon bêta atténuée

I_0 = Intensité initiale



(a) Modèle

(b) Schema de fonctionnement

Figure 3.2 : BAM 1020

μ = Coefficient d'atténuation en $\text{cm}^2 \mu\text{g}^{-1}$

Q = Débit d'air en L min^{-1}

t = Temps en minutes

Cette méthode de mesure a plusieurs limites. En effet, elle dépend des conditions météorologiques telles que la température et l'humidité relative ce qui implique une sous-estimation ou surestimation des concentrations des particules ainsi que sa résolution spatiale est faible.

3.2.4 Les capteurs *PM* à coût faible

Les capteurs *PM* à bas coût suscitent de plus en plus d'intérêt grâce à leurs compositions et leur faible coût par rapport aux capteurs de référence. En effet, ils sont caractérisés par un prix bien inférieur à 1000 euros (pour la partie sonde). Cependant, ces capteurs n'ont pas encore atteint la performance voulue pour être déployés sans l'accompagnement des capteurs de référence pour vérifier les valeurs mesurées. A cela s'ajoute qu'aucune agence ne se repose sur ses mesures à cause de leur dégradation rapide avec le temps.

Il n'existe pas de définition commune d'un capteur, il convient de faire la distinction entre le module de détection du polluant et l'ensemble de système de surveillance comprenant un ou plusieurs modules de détection ainsi qu'un boîtier de protection, un système d'alimentation (batterie), un microprocesseur pour convertir le signal de la lumière en unités de concentration, du matériel électronique et des composants pour le stockage et la transmission des données. Un exemple d'un boîtier capteur fait par l'équipe AGORA est dans la figure 4.3.

Dans tout le rapport, le terme capteur désigne le module de détection.

3.2.4.1 Fonctionnement des capteurs à coût faible

Les capteurs à faible coût sont des capteurs optiques basés sur la technique de dispersion de la lumière qui est le principe fondamental de leur fonctionnement. Ces capteurs sont équipés d'un ventilateur pour faire circuler l'air et d'une LED émettant



Figure 3.3 : Exemple d'un micro-capteur de pollution

des faisceaux lumineux avec une longueur d'onde entre le domaine du visible et le domaine infrarouge pour garantir la détection de particules ayant un diamètre entre 0.1-100 μm . Quand le faisceau lumineux traverse les particules, la lumière dispersée est collectée par une photo-diode. La taille de la particule est ensuite déterminée et comptée en comparant l'intensité lumineuse à une courbe standard calibrée à l'aide des particules dont la taille est déjà connue. Cela permet donc de calculer la concentration. Chaque modèle de capteur utilise une méthode de calcul de la concentration propre à son fabricant, ce qui explique les petites différences entre les modèles de capteurs à faible coût. La figure 3.5 présente quelques modèles de capteurs à faible coût et leurs caractéristiques.

3.2.4.2 Caractéristiques et limites

Les capteurs à faible coût font des mesures en temps réel (une mesure toutes les secondes), ainsi qu'ils sont sensibles à des concentrations de l'ordre de 1 $\mu\text{g}/\text{m}^3$. Cependant, ce sont des capteurs qui ont une durée beaucoup plus inférieure à celle

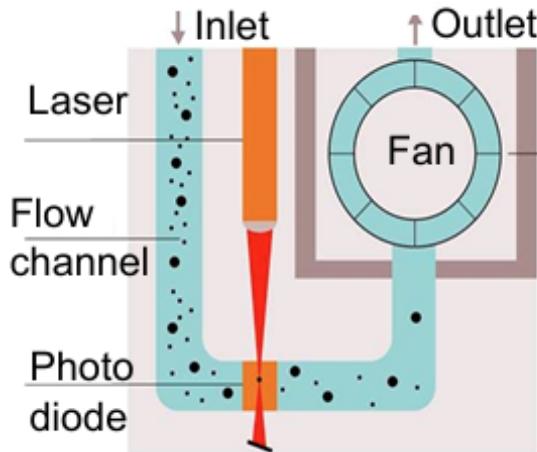


Figure 3.4 : Principe de fonctionnement des capteurs à faible coût des PM

annoncée par les fabricants. Leurs composants tels que le ventilateur ou la LED peuvent tomber en panne peu de temps après le déploiement des capteurs. Ils sont très sensibles aussi aux conditions météorologiques surtout l'humidité relative dès qu'elle devient supérieure à 85%, ce qui peut perturber les mesures et user les composants. De plus, les mesures sont basées sur des calibrations effectuées sur des particules de taille connue et sous des conditions bien précises différentes des conditions réelles dans lesquelles elles sont faites en pratique.

3.3 Évaluation des performances des micro-capteurs

Pour assurer une bonne qualité de mesure, les capteurs à faible coût sont évalués par différentes approches pour déterminer leur capacité à décrire le niveau de pollution d'air avec la façon la plus proche de la réalité en les comparant à des capteurs de références validés par les agences de protection de l'environnement [4].

Les données issues des capteurs sont analysées afin de déterminer le niveau de leur

performance, l'impact des conditions météorologiques ainsi que les anomalies causant leur détérioration.

3.4 Métriques de performance et mesures statistiques

Les concentrations des particules fines mesurées par les capteurs, regroupées par taille, sont stockées dans une base de données accompagnée par les facteurs météorologiques susceptibles d'influencer l'évaluation des capteurs tels que les conditions météorologiques. L'évaluation des capteurs repose donc sur des métriques et des mesures statistiques.

Les articles scientifiques parcourus mettent en évidence les mesures statistiques de trois indicateurs clés de performance des capteurs : La corrélation, l'erreur de mesure et la précision. La corrélation est utilisée pour évaluer les degrés de relation entre les mesures estimées du capteur et des valeurs de référence. L'erreur de mesure compare la différence entre les estimations du capteur et les mesures de référence correspondantes. Quant à la précision, elle mesure l'accord entre les mesures répétées d'une même propriété dans des conditions identiques.

3.4.1 Le coefficient de détermination r^2

Le coefficient de détermination r^2 fournit une mesure de l'importance de la relation entre les variations des concentrations de particules fines générées le capteur à faible coût et le capteur de référence. Ce coefficient est une mesure de l'ajustement du modèle de la régression linéaire aux valeurs observées qui sont les valeurs du cap-

teur de référence. Dans l'équation ci-dessous, y_i représente les valeurs observées qui sont les valeurs du capteur de référence, \hat{y} représente la valeur prédictive par le modèle de régression linéaire et \bar{y} représente la moyenne des mesures du capteur de référence :

$$r^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

D'après [5], dans une situation idéale où le capteur correspond parfaitement au capteur de référence, la pente serait égale à 1, le terme constant serait égal à 0 et le coefficient r^2 serait proche de 1. D'après [5] et [3], il existe une relation entre le capteur et le capteur de référence à partir d'une valeur seuil de r^2 qui est 0.7.

3.4.2 RMSE et MAE

Le MAE (erreur moyenne absolue) et le RMSE (erreur quadratique moyenne) mesurent l'ampleur de l'erreur de mesure. Le RMSE donne plus de poids aux erreurs les plus significatives, tandis que le MAE traite chaque erreur d'une manière égale [5].

Les équations pour RMSE et MAE sont données ci-dessous :

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$\text{Mean absolute error (MAE)} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

3.4.3 Erreur moyenne de biais : MBE

Le MBE entre le capteur de particules fines à bas coût et le capteur de particules fines de référence décrit la direction de l'erreur et indique si le capteur a tendance à sous-estimer ou surestimer le capteur de référence [5]. L'équation pour MBE est ci-dessous :

$$\text{Mean Bias Error (MBE)} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$

3.4.4 Coefficient de variation : CV

Le coefficient de variation est une métrique recommandée par l'EPA [6] pour mesurer la précision des capteurs à bas coût. En effet, Il est calculé comme suit :

$$CV = \frac{\sigma}{\mu} \times 100\%$$

Tel que σ est l'écart-type μ est la moyenne des données sur une minute provenant d'unité du même modèle du capteur. La valeur finale de CV pour chaque modèle de capteur a été déterminée comme la moyenne de toutes les valeurs temporaires de CV. D'après [6], les capteurs à faible CV (inférieur à 10%), ont une reproductibilité élevée ce qui veut dire qu'ils fournissent des mesures fiables et cohérentes.

3.5 Analyse des données

3.5.1 Préparation des données

En général, les données des capteurs à faible coût sont comparées à celles du capteur de référence, puis téléchargées et stockées sur un serveur distant, pour être affichées sur une page web privée actualisée toutes les minutes. La page permet donc un téléchargement direct des données [3]. Une fois disponible, un nettoyage des données est effectué en éliminant les valeurs négatives, les valeurs nulles, les valeurs invalides (NAN) et les valeurs qui se trouvent en dehors de la plage de mesure ou en dehors des limites de concentrations typiques de particules fines [3]. Les données sont ensuite moyennées en général sur des intervalles de temps d'une heure et mises en correspondance par date et heure avec les données horaires du capteur de référence. Selon [3], les données récupérées devraient représenter au minimum 75% des données brutes.

3.5.2 Relation entre les capteurs à faible coût et le capteur de référence

Dans [3], les auteurs mesurent la corrélation de quatre capteurs $PM_{2.5}$ à faible coût utilisant la technique de dispersion de la lumière et issus de 4 fabricants différents (voir la figure 3.5) avec un capteur de référence TEOM, mentionné dans la table ???. Trois capteurs de chaque modèle(fabricant) ont été placés sur le même site en Pologne pour comparer la performance des capteurs sous les mêmes conditions de mesures.

Les auteurs trouvent sur la période de mesure un coefficient de détermination r^2 élevé, ce qui montre la relation linéaire entre les capteurs et la référence pour les moyennes

Sensor model	SDS011	ZH03A	PMS7003	OPC-N2
Manufacturer	Nova Fitness	Winsen	Plantower	Alphasense
Approximate price (\$)	20	20	20	500
Dimensions (mm)	71 × 70 × 23	50 × 32.4 × 21	48 × 37 × 12	75 × 63.5 × 60
Approximate weight (g)	50	30	30	105
Power supply voltage (V)	5	4.5–5.5	4.5–5.5	4.8–5.2
Working current (mA)	220	70–140	≤100	175
Detectable size range (μm)	0.3–10	0.3–10	0.3–10	0.38–17
Size bins	Not available	Not available	6 size bins	16 size bins
Estimated PM_{x} concentration	$\text{PM}_{2.5}, \text{PM}_{10}$	$\text{PM}_1, \text{PM}_{2.5}, \text{and } \text{PM}_{10}$	$\text{PM}_1, \text{PM}_{2.5}, \text{and } \text{PM}_{10}$ Effective range: 0–500 Maximum range: above 1000 (for $\text{PM}_{2.5}$)	$\text{PM}_1, \text{PM}_{2.5}, \text{and } \text{PM}_{10}$ 0.01–1500·10 ³ (for PM_{10})
Concentration range ($\mu\text{g}/\text{m}^3$)	0–999.9	0–1000 (for $\text{PM}_{2.5}$)		

Figure 3.5 : Les caractéristiques des capteurs à faible coût issus de fabricants différents et utilisant la même technique de mesure.

sur une heure sur avec PMS7003 avec une valeur de 0.89. Cependant, le capteur OPC-N2 présente un coefficient modéré (0.61) à cause de la dispersion des données.

Pour déterminer en détail la relation entre les mesures des capteurs et la référence, les auteurs dans [3], [5] et [6] ont essayé de voir la relation entre l'erreur par heure et les concentrations mesurées par le capteur de référence dans le but d'identifier l'impact des concentrations mesurées par les capteurs à faible coût . En effet, [5] a comparé 12 capteurs à faible coût avec un instrument de référence qui est le BAM, dont le fonctionnement est expliqué dans la partie 2.2.3. Afin d'examiner l'impact potentiel des concentrations de $\text{PM}_{2.5}$ sur la réponse des capteurs, les erreurs de biais horaires ont été tracées en fonction du BAM pour les 12 capteurs. Cela permet de voir en effet la sensibilité des capteurs à faible coût aux niveaux de pollution, et ainsi déterminer si le capteur à faible coût sous-estime ou surestime l'instrument de référence. Les auteurs ont ensuite tracé l'erreur relative de $\text{PM}_{2.5}$ (concentration en %) en fonction du BAM ($\text{PM}_{2.5}$, concentrations horaires) pour voir l'impact de la plage de concentration sur la performance du capteur.

3.5.3 Variabilité intra-modèle

La variabilité intra-modèle au sein d'un certain nombre de capteurs du même modèle (3 pour le cas de [3]) se définit comme les degrés d'accord entre les capteurs. Cela est déterminé en calculant les concentrations moyennes de particules fines mesurées par les capteurs individuels, puis en les comparant avec la moyenne des moyennes et l'écart-type pour cette moyenne des moyennes. L'écart-type pour la moyenne des moyennes fournit une mesure de la variabilité intra-modèle. Un écart-type élevé pour la moyenne des moyennes indique une forte variabilité intra-modèle, tandis qu'un écart-type faible indique une faible variabilité intra-modèle.

3.5.4 Reproductibilité des capteurs

Le calcul du coefficient de variabilité permet de déterminer la reproductibilité du modèle du capteur, en effet, il permet de déterminer le degré de fiabilité du modèle et degrés de répétabilité entre les capteurs de ce modèle. Cette caractéristique est évaluée dans [3] et [6] vu qu'elle est essentielle pour déterminer la performance des capteurs d'après EPA.

3.5.5 Impact de la température et de l'humidité relative

Un facteur important causant la dégradation du capteur à faible coût est l'humidité relative. En effet, d'après [7], à partir d'un certain seuil, la performance du capteur se dégrade. Cela est aussi examiné par les papiers, en effet,dans [5], les erreurs de

biais horaires ont été tracées en fonction de l'humidité relative horaire pour les 12 capteurs. Idéalement, la pente de la ligne de meilleur ajustement devrait être nulle et se situer sur l'axe $y = 0$ (voir figure 3.6). Cela détermine à quelle valeur de l'humidité, les mesures du capteur commencent à s'éloigner de celles de référence.

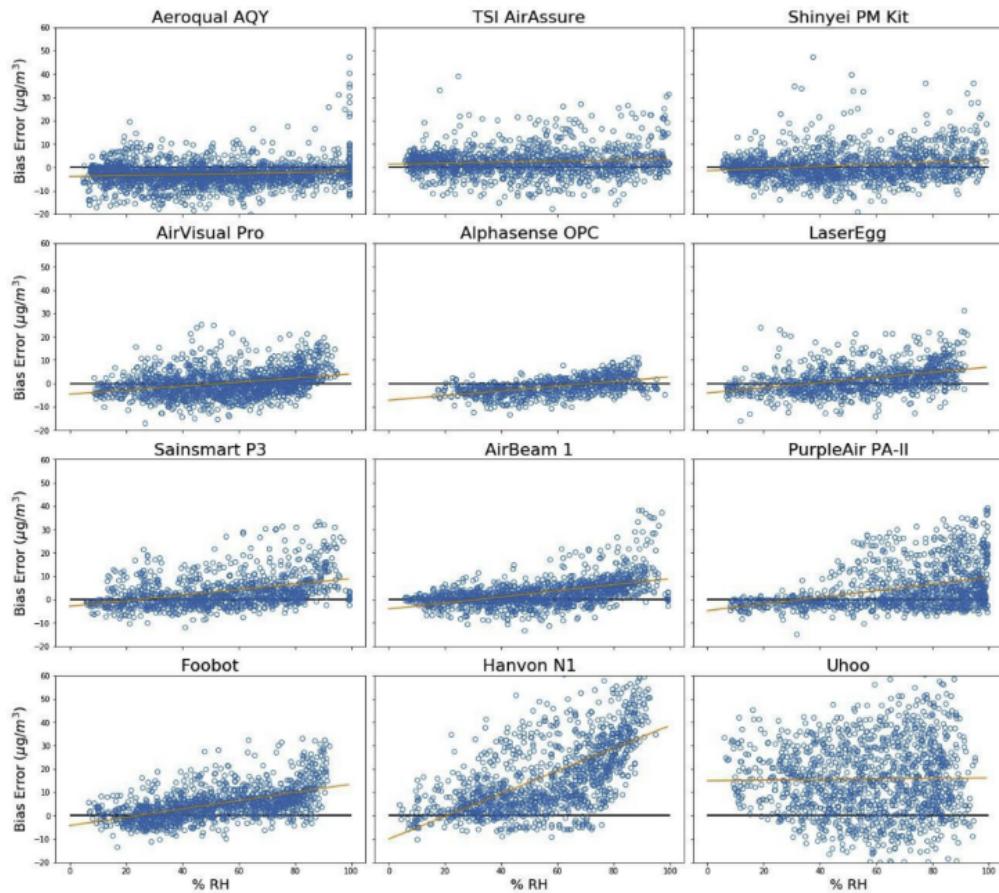


Fig. 1. Impact of Relative Humidity (RH) on the bias error between Sensor and Met One BAM.

Figure 3.6 : L'impact de l'humidité relative sur 12 capteurs à faibles coût.

Dans [3], les auteurs ont tracé les concentrations mesurées par le capteur de référence en fonction des concentrations mesurées par le capteur à bas coût pour trois intervalles de valeurs d'humidité relative. Ils ont ensuite réalisé une régression linéaire pour étudier l'effet de l'humidité sur le capteur à bas coût (figure ci-après). Leurs analyses

ont montré que pour une humidité élevée, les capteurs sont moins performants. Ainsi, les auteurs de [8] ont montré que la température exerce une influence sur les capteurs à faible coût en diminuant le coefficient de détermination r^2 .

Les articles [8], [9], et [10] ont étudié l'influence des conditions météorologiques sur la performance des capteurs à bas coût, et ils sont arrivés à constater que l'humidité a un impact plus fort que celui de la température. En effet, lorsque l'humidité relative est élevée, les particules polluantes absorbent l'eau, ce qui entraîne une augmentation du signal de lumière diffusée par les particules. Par conséquent, la concentration de ces dernières est surestimée et les composants deviennent défaillants.

3.6 Techniques de détection des anomalies des micro-capteurs

3.6.1 Approches basées sur les mesures des capteurs

Les approches centrées sur les données décrites se reposent sur des techniques statistiques basées sur des données collectées des capteurs voisins [11],[12]. Ces approches divisent les données continues des capteurs en tranche de temps et identifient trois types d'anomalies :

1. **Anomalie spatiale** est détectée si les données de mesure sont beaucoup plus élevées (ou plus basses) que la moyenne de son capteur de référence haut de gamme voisin dans la même tranche de temps. En effet si l'écart entre les deux capteurs est supérieur à un seuil prédéfini (fixé à 2 écarts types dans le papier), cela est donc considéré comme une anomalie.
2. **Anomalie temporelle** est détectée lorsqu'il y a un écart significatif (supérieur

à 2 écarts types) dans le comportement attendu du capteur par rapport à ses données historiques.

3. **Anomalie spatio-temporelle** est détectée si les données de mesure sont identifiées comme des anomalies spatiales et temporelles dans la même tranche de temps.

3.6.2 Approches basées sur la consommation des capteurs

La mauvaise performance des capteurs à bas coût est attribuée à l'impact des facteurs environnementaux associés aux limites des composants bon marché. Une part importante des défaillances des capteurs est causée par la défaillance de ces composants, entraînant ainsi des inexactitudes dans les mesures. Les auteurs de [11] proposent une approche appelée CurrentSense, qui permet de diagnostiquer l'état du capteur $PM_{2.5}$ sans aucune information sur les conditions météorologiques ainsi que sur les données historiques de mesure. Cette méthode échantillonne le courant consommé par le capteur pour produire une unique empreinte électrique indiquant l'état du capteur (défaillant ou non) et pour isoler le composant responsable de la défaillance. D'après les auteurs, les composants qui sont à l'origine des défaillances sont la LED émettant les faisceaux de lumière, et le ventilateur qui fait passer le flux d'air à l'intérieur du capteur. Ainsi, CurrentSense échantillonne le courant à deux fréquences, à savoir 30 Hz et 5 kHz, afin de capturer toutes les variations et de surveiller les composantes à basse fréquence, où l'on observe la consommation de courant par la LED, et à haute fréquence, où l'on observe la consommation de courant par le ventilateur. Si un changement dans la consommation de courant par un composant est détecté, alors

ce composant est considéré comme défaillant. Cette méthode s'est montrée efficace à classifier les capteurs défaillants avec une précision de 99%.

3.7 Méthodologie retenue

Dans le cadre de ce stage, nous allons nous concentrer sur l'approche centrée sur les données tout en utilisant les métriques évoquées dans ce chapitre afin de modéliser la performance des capteurs à faible coût déployés par les agences de suivi de qualité de l'air en France. Cela nous permettra de caractériser entre autres l'impact de l'humidité et de la température (mais aussi l'encrassement et le vieillissement des sondes) sur les concentrations mesurées.

4 Analyse préliminaire des mesures de pollution

Afin d'évaluer les performances des micro-capteurs de particules fines, je me suis basé sur un jeu de données obtenues avec des capteurs déployées à Grenoble par l'agence de surveillance de qualité de l'air Atmo Auvergne-Rhône-Alpes (ATMO-AuRA). ATMO-AuRA est l'observatoire de référence pour la surveillance de la qualité de l'air de la région AuRA. Les stations de surveillance réalisent des mesures toutes les heures. Ces mesures horaires sont envoyées en WIFI à un serveur dans les locaux de l'agence. Ce serveur va les stocker dans une base de données. Les clients accèdent donc à la base de données via une API (Application Programming Interface).

4.1 Jeu de donnée

4.1.1 API

Dans la région d'Auvergne Rhône Alpes, nous comptons près de 256 sites de mesures, dans lesquelles des appareils de mesure (analyseur d'air et micro-capteurs) sont installés pour mesurer la concentration des polluants (gaz ou particules) et les données météorologiques (humidité et température). Chaque site, chaque type de polluant et chaque type d'appareil de mesure est caractérisé par un identifiant. L'assemblage de ces identifiant forme l'identifiant du capteur de pollution. En effet, les tableaux [4.1](#) et [4.2](#) présentent les identifiants utilisés dans le cadre de mon stage. Un identifiant

d'une capteur s'écrit donc : `id_site + _ + id_Polluant + _ + id_Appareil`.

Polluant	Id_Polluant
$PM_{2.5}$	39
PM_{10}	24
Humidité	38
Température	54

Appareil de mesure	Id_Appareil
Analyseur d'air	11
Micrôcapteur	57

Table 4.2 : Id des appareils de mesure

Table 4.1 : Id des polluants et des données météorologiques

Pour obtenir les données des mesures d'un polluant d'un site, il faut donc entrer le numéro de département dans lequel le site est situé, l'identifiant de ce site, l'identifiant du polluant et la période de mesure qu'on souhaite avoir. Dans le cadre de stage, on s'est intéressé sur le bassin d'air Grenoblois, alors nous avons choisi de travailler sur le site Grenoble les Frênes, comportant un analyseur d'air (capteur de référence), 5 micro-capteurs dont trois mesurant le $PM_{2.5}$, un mesurant l'humidité et un mesurant la température. Les trois micro-capteurs sont déployés depuis janvier 2021. Il existe aussi un autre micro-capteur déployé une année après ces derniers.

Pour récupérer ces données, j'ai automatisé le processus de récupération des données en réalisant un code en langage python permettant de prendre en paramètres les informations mentionnées au paragraphe de dessus, et transforme les données horaires obtenues sous format JSON en tableau csv.

API Atmo Auvergne-Rhône-Alpes

GET

[https://api.atmo-aura.fr/api/v1/valeurs/horaire?
api_token=fb0d9e1ded2c253ba5be472679933df&date_debut=yesterday&sites=FR20062&label_court_polluant=PM10](https://api.atmo-aura.fr/api/v1/valeurs/horaire?api_token=fb0d9e1ded2c253ba5be472679933df&date_debut=yesterday&sites=FR20062&label_court_polluant=PM10)

Paramètre(s)	Paramètre	Valeur	Description
doc	get	Valeurs possibles : get <i>affiche la documentation en fonction de la méthode http demandée</i>	
format	json	Valeurs possibles : json csv <i>format de sortie</i>	
separateur_decimal_point_csv		Valeurs possibles : true <i>Uniquement pour le format CSV. Permet de définir le séparateur décimal par un point (ex: 23.8).</i>	
sites	FR15043	<i>liste de sites séparés par une virgule</i>	
mesures		<i>liste de mesures séparés par une virgule</i>	
valeur_brute	1	Valeurs possibles : 1 true <i>Inclus les valeurs brutes</i>	
type_appareil_id		<i>liste d'id de type d'appareils séparés par une virgule</i>	
date_debut	2021-01-01	<i>date de début des valeurs au format YYYY-MM-DD ou "12 hours" ou "7 days"</i>	
date_fin	2024-03-31	<i>date de fin des valeurs au format YYYY-MM-DD</i>	
code_polluant		<i>code_polluant ou liste de code_polluant séparés par une virgule</i>	
id_poll_ue	39	<i>id_poll_ue ou liste de id_poll_ue séparés par une virgule</i>	
label_court_polluant	PM2.5	<i>label_court_polluant ou liste de label_court_polluant séparés par une virgule</i>	
order_by_date		Valeurs possibles : asc desc <i>(Défaut: asc). Trie les données par date de façon ascendante ou descendante</i>	
combler		Valeurs possibles : true 1 <i>Permet d'avoir une série de date sans trou. Nécessite de renseigner une date de début et de fin.</i>	
groupe_id		<i>Retourne les mesures définies dans un groupe de mesures (voir api/v1/groupes).</i>	
select		<i>Permet de filtrer et d'ordonner les champs retournés. Saisissez le nom des champs séparés par une virgule (ex: mesure_id,date, valeur)</i>	
timezone_csv	Europe/Paris	<i>Uniquement pour le format CSV. Permet de définir le fuseau horaire pour les dates. (Par défaut UTC, pour l'heure local, mettez Europe/Paris)</i>	
dateformat_csv		<i>Uniquement pour le format CSV. Permet de définir le format de la date. (Par défaut jour/mois/année heure minutes seconde) Voir les formats possibles sur la page Format datetime. Le format classique AAAA-MM-JJ HH:MM:SS se fait avec Y-m-d H:i:s</i>	

Figure 4.1 : Extraction des données

4.1.2 Site utilisé

Pour voir la performance des capteurs à bas coût, j'ai utilisé trois micro-capteurs déployés depuis 2021 mesurant $PM_{2.5}$ et un analyseur d'air, le capteur de référence se trouvant sur le bassin d'air Grenoblois notamment au site Grenoble les frênes portant l'identifiant FR15043. Ce site porte notre intérêt vu qu'il compte plus un micro-capteur issus des différents fabricants, localisés avec la référence contrairement

aux autres sites. Cela permettrait de se rendre compte si des tests ou des corrections sont valables pour tous ou uniquement pour certains capteurs. De plus, on retrouve dans ce site deux micro-capteurs pour mesurer la température et l'humidité.

4.2 Prétraitement des données

Une fois les données sont récupérées, un pré-traitement a été effectué afin d'avoir une claire visualisation et une bonne compréhension des données. Les valeurs horaires de concentrations de $PM_{2.5}$ pour le site les Frênes ont été extraites pour la période allant du premier janvier 2021 jusqu'à mars 2024.

Ces valeurs sont presque équilibrées à l'image de la figure 4.2 montrant la répartition des mesures pour chaque capteur en 2021. En effet, il existe des heures dans lesquelles des mesures de concentration de quelques capteurs manquent. Alors, pour assurer une visualisation claire sur les mêmes heures et par la suite une bonne évaluation de performance, j'ai créé un tableau de données pour chaque micro-capteur dans lequel on retrouve ses valeurs horaires de concentration accompagnées des valeurs de référence et des valeurs de température et humidité afin d'avoir à chaque heure ces quatre mesures. Cela a été réalisé à l'aide de la librairie Pandas notamment avec la fonction "merge" qui permet de faire l'intersection entre le tableau des données de la pollution et le tableau des données météorologiques. Il est vrai qu'il peut y avoir certaines heures où les données seront manquantes, mais cette méthode reste efficace et reflète fidèlement la réalité. Elle permet de conserver 97% des valeurs mesurées par chaque capteur, comme le montre le tableau ci-dessous. Pour la suite de ce rapport, les trois micro-capteurs seront désignés comme suit : capteur de référence

Répartition des mesures en 2021

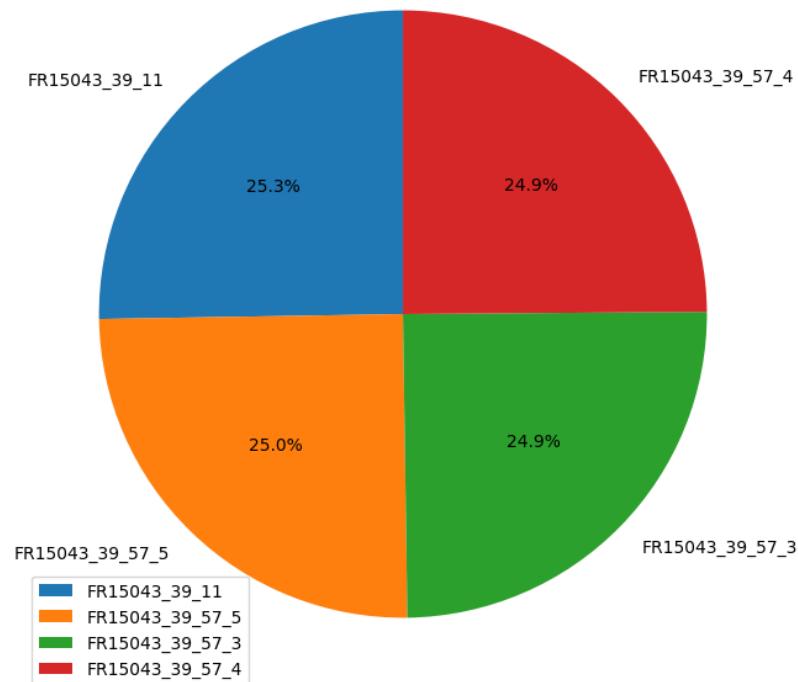


Figure 4.2 : Répartition des mesures en 2021

(FR15043_39_11), capteur 1 (FR15043_39_57_3), capteur 2 (FR15043_39_57_4) et capteur 3 (FR15043_39_57_5).

4.3 Analyse préliminaire des données

Les données récupérées sont des données brutes issues des micro-capteurs, alors une compréhension de ces données est nécessaire pour leur évaluation.

Sur une vue d'ensemble sur les micro-capteurs, ils mesurent souvent des valeurs similaires sur la plage de valeurs basse. En effet, il y a une bonne superposition des distributions pour les valeurs basses (0 à 20), ce qui montre que dans cette plage de valeurs, les deux capteurs sont relativement cohérents. Cependant, la distribution des capteurs à bas coût a une plus grande dispersion, ce qui signifie qu'il mesure parfois des valeurs beaucoup plus élevées que celles du capteur de référence, indiquant une possible divergence de performance ou de précision.

Pour des valeurs plus élevées (au-delà de 30), la densité pour le capteur à faible coût reste non négligeable tandis qu'elle devient très faible pour le capteur de référence, ce qui suggère que les mesures élevées du capteur à faible coût pourraient être des valeurs aberrantes ou des erreurs de mesure. Au-delà de cette plage, les différences deviennent plus marquées, indiquant une possible divergence de performance ou de précision. D'après le tableau 4.3, les capteurs ou les mesures identifiés par les capteurs présentent une variabilité relativement élevée, avec des coefficients de variation proches ou supérieurs à 100%. Cela indique que pour chaque mesure, l'écart type est presque aussi grand, voire plus grand que la moyenne, ce qui peut signaler des pro-

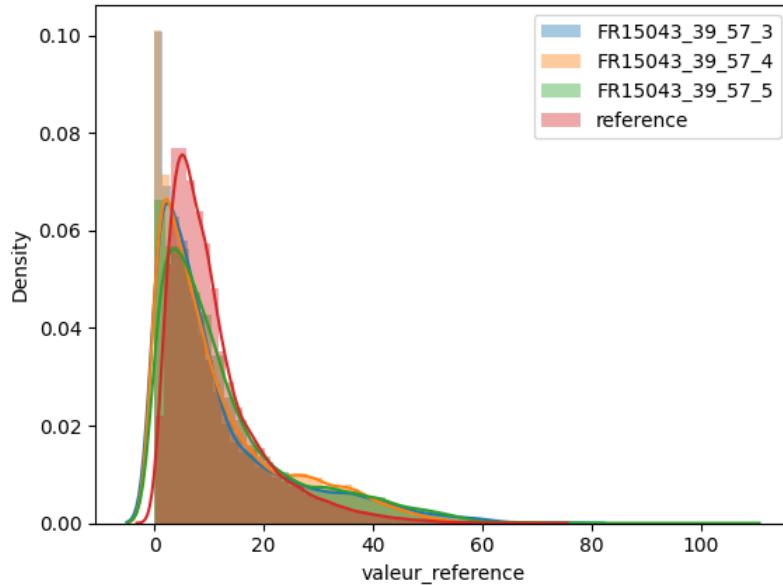


Figure 4.3 : Distribution des valeurs

blèmes de précision ou de stabilité des mesures. Dans des contextes où la précision est cruciale, cette variabilité pourrait être problématique et nécessiter des ajustements, des calibrations supplémentaires.

mesure_id	moyenne	écart_type	CV %
FR15043_39_11	10.89	8.60	79.00
FR15043_39_57_3	11.79	12.67	107.48
FR15043_39_57_4	11.71	11.71	99.92
FR15043_39_57_5	12.96	12.43	95.89

Table 4.3 : Statistiques descriptives des mesures de capteurs pour une période de 3 ans

Cette variabilité élevée des micro-capteurs est identifiée aussi sur une période mensuelle. En effet, le diagramme à moustache dans la figure 4.4 montre une distribution de valeurs plus large des micro-capteurs par rapport à la référence. Ainsi, de nombreuses valeurs aberrantes sont présentes significatives d'un pic de pollution qui s'est

produit près des capteurs.

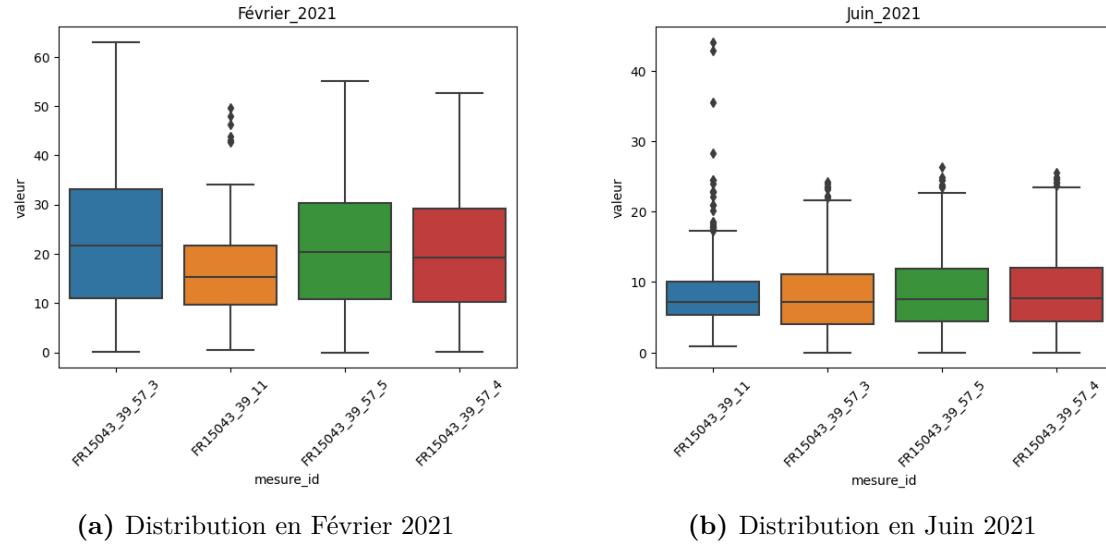


Figure 4.4 : Comparaison des distributions en Février et Juin 2021

Nous remarquons également que quelques mois tels que le mois de Février 2021 et Juin 2021 dans lesquels les micro-capteurs ne parviennent pas à détecter les pics de pollutions. En effet, on obtient un indice de corrélation de 0.66 pour février 2021 et 0.70 pour juin 2021 comme le montre le tableau 4.4.

Table 4.4 : Performance des Modèles par Mois et Année

Mois_Année	r ²	MAE	RMSE
Janvier (2021)	0.97	14.01	17.09
Février (2021)	0.16	11.50	14.2
Juin (2021)	0.44	2.37	4.14
Juillet (2021)	0.75	2.25	3.18
Juin (2022)	-0.5	3.64	4.93
Juillet (2022)	0.68	2.33	2.67
Août (2022)	0.54	2.12	2.47
Septembre (2022)	-0.49	2.48	2.88
Octobre (2022)	-0.86	4.07	5.24
Septembre (2023)	-1.46	3.41	5.56
Octobre (2023)	-2.18	4.76	5.64

5 Étude comparative des méthodes de calibration

Les données des micro-capteurs de pollution de l'air ne peuvent pas être utilisées directement et doivent être d'abord corrigées (ou calibrées). Dans le but de corriger les mesures des capteurs, j'ai implémenté et comparer des méthodes de calibration ayant pour but d'ajuster les valeurs des capteurs en prenant en compte les données météorologiques telles que la température et l'humidité relative. L'analyse des résultats de ces méthodes de calibration nous permettra aussi de déterminer la performance des capteurs et identifier les facteurs influant le plus la dégradation de leur bon fonctionnement.

5.1 Paramètres de calibration

Afin de calibrer les capteurs à bas coût, j'ai pris en considération les trois paramètres suivants dans le processus d'inférence des niveaux de pollution de référence :

1. Le niveau de concentration de particules fines mesuré par le micro-capteur. Il s'agit de la mesure brute calculée à partir de l'estimation faite par le fabricant.
2. La température, un facteur important pouvant avoir une influence sur les composants des capteurs
3. L'humidité relative, étant un facteur important pour la calibration. En effet, l'eau piégée dans l'air pourrait modifier la taille des particules mais aussi pourrait être considérée comme une particule. Par la suite, les mesures des micro-capteurs sont faussées.

5.2 Processus de calibration

Le processus de calibration a été effectué sur chaque mois et sur chaque micro-capteur, alors j'ai divisé les données de chaque micro-capteur par mois accompagnés par les données de référence et les données météorologiques. Ainsi, pour chaque mois je prends 80% pour entraîner le modèle et 20% pour le test.

De plus, pour éviter le sur-apprentissage à l'entraînement, alors j'ai opté pour la validation croisée qui peut être effectuée par deux méthodes principales. La première est la validation croisée exhaustive, qui consiste à trouver et à tester toutes les combinaisons possibles pour diviser l'échantillon original en un ensemble de d'entraînement et un ensemble de test. L'autre méthode, plus courante et celle que j'ai utilisée, est la validation croisée non exhaustive, connue sous le nom de validation k-fold. La technique de validation k-fold consiste à diviser les données d'entraînement en k groupes et à réserver l'un de ces groupes pour tester le modèle d'apprentissage à chaque tour.

Pour effectuer la validation k-fold, les données sont d'abord affectées de manière aléatoire à un nombre k de godets de taille égale. Un godet est ensuite réservé comme godet de test et est utilisé pour mesurer et évaluer la performance des (k-1) autres godets.

La validation croisée est répétée k fois, le processus est répété jusqu'à ce que tous les godets aient été utilisés à la fois comme un godet d'entraînement et de test. Les résultats sont ensuite agrégés et combinés pour formuler un modèle unique. Pour l'entraînement des différents modèles, j'ai fixé k à 5.

5.3 Méthodes de calibration implémentées

5.3.1 Régression linéaire simple (RLS)

J'ai commencé avec une méthode simple qui est la régression linéaire simple prenant en compte la concentration de $PM_{2.5}$ mesurée par le micro-capteur. Cette méthode a pour but de voir si le modèle linéaire simple arrive à bien corriger les capteurs.

$$y = a \cdot x_{\text{capteur}} + b$$

Les données brutes peuvent contenir des valeurs aberrantes qui perturbent l'apprentissage du modèle. Pour atténuer cet effet, nous avons utilisé 'RobustScaler' de scikit-learn, qui normalise les données en utilisant la médiane et l'écart inter-quartile. Cette méthode est moins sensible aux valeurs extrêmes, comparée à d'autres techniques comme la standardisation ou la mise à l'échelle min-max.

5.3.2 Régression linéaire multiple (RLM)

Pour prendre en compte les autres paramètres de calibration (température, humidité, etc.), j'ai ensuite implémenté une fonction de calibration basée sur la régression multi-linéaire.

La régression multi-linéaire prend en compte plusieurs variables explicatives afin

d'améliorer la calibration des micro-capteurs comme suit :

$$y = \begin{bmatrix} x_{\text{hum}} & x_{\text{temp}} & x_{\text{cap}} & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

Les données brutes peuvent contenir des valeurs aberrantes qui perturbent l'apprentissage du modèle. Pour atténuer cet effet, j'ai normalisé les données en utilisant la médiane et l'écart inter-quartile comme suit :

$$x_{\text{norm}} = \frac{x_{\text{brute}} - \text{médiane}}{Q_3 - Q_1}$$

Le choix de cette technique de normalisation est motivé par le fait qu'elle est moins sensible aux valeurs extrêmes comparées à d'autres techniques comme la standardisation ou la mise à l'échelle min-max.

5.3.3 Régression polynomiale multivariable (RP)

La régression polynomiale multi-variable cherche à approximer la relation entre la variable dépendante y et les variables indépendantes x_{hum} , x_{temp} et x_{cap} à l'aide d'une fonction polynomiale de degré d . Ce modèle est efficace car en rajoutant des variables polynomiales, le modèle gagne en complexité. La relation est donnée par l'équation

suivante :

$$y = X\beta$$

où X est la matrice des termes polynomiaux des variables indépendantes et β est le vecteur des coefficients des termes polynomiaux.

Après la normalisation des données, pour capturer les relations non linéaires entre les caractéristiques et la variable cible (le capteur de référence), nous avons utilisé PolynomialFeatures pour transformer les caractéristiques d'origine en caractéristiques polynomiales. Ensuite, nous avons appliqué une régression linéaire sur ces nouvelles caractéristiques.

Pour identifier le degré de polynôme optimal, j'ai utilisé GridSearchCV, qui effectue une recherche exhaustive sur une grille de paramètres spécifiée. GridSearchCV utilise la validation croisée pour évaluer les combinaisons de paramètres et sélectionner celle qui minimise l'erreur quadratique moyenne.

5.3.4 Random Forest (RF)

La méthode Random Forest est un algorithme d'apprentissage utilisé pour les tâches de classification et de régression. Elle combine plusieurs arbres de décision pour créer un modèle plus robuste et précis. Chaque arbre est construit à partir d'un sous-échantillon aléatoire du jeu de données, et les prédictions finales sont obtenues en agrégeant les prédictions de tous les arbres. Cela permet de réduire le risque de sur-apprentissage et d'améliorer les performances globales du modèle.

Pour calibrer un capteur à bas coût, Random Forest est particulièrement utile car il peut gérer des données bruitées et des relations non linéaires complexes entre les variables. En appliquant ce modèle, on peut obtenir des prédictions précises malgré les variations et les incertitudes inhérentes aux capteurs bon marché. La robustesse de Random Forest aide à extraire des informations pertinentes des données brutes du capteur, améliorant ainsi la fiabilité et la précision des mesures obtenues.

5.3.5 Gradboost

Le Gradient Boosting est une technique d'apprentissage supervisé qui combine plusieurs modèles faibles, généralement des arbres de décision, pour créer un modèle global plus robuste et performant. Contrairement à Random Forest, qui construit des arbres en parallèle et agrège leurs prédictions, le Gradient Boosting construit des arbres de façon séquentielle, chaque nouvel arbre corrigeant les erreurs des arbres précédents. Cela permet de capturer des relations complexes et non linéaires dans les données.

5.4 Résultats de comparaison

J'ai appliqué les méthodes implémentées aux données des trois capteurs présentées dans le chapitre 4. Pour obtenir des résultats plus précis, j'ai d'abord segmenté les données par mois, ce qui permet de réduire l'impact de la saisonnalité sur les observations et d'améliorer la robustesse des résultats.

En utilisant le coefficient de détermination R^2 et le $RMSE$ comme métriques d'éva-

luation, je présente les premiers résultats de performances mensuelles dans le tableau 5.1 où je montre pour chaque méthode les valeurs moyennes, minimales et maximales de chaque métrique.

Modèle	Capteur 1						Capteur 2						Capteur 3					
	R ²			RMSE			R ²			RMSE			R ²			RMSE		
	Min	Moy	Max	Min	Moy	Max	Min	Moy	Max	Min	Moy	Max	Min	Moy	Max	Min	Moy	Max
RLS	-0.61	0.71	0.96	0.77	2.12	6.14	-1.85	0.6	0.98	0.83	2.34	6.15	-0.63	0.71	0.96	0.81	2.19	6.29
RLM	-0.18	0.80	0.97	0.63	1.85	5.24	-1.13	0.73	0.98	0.73	2.06	5.4	-0.09	0.79	0.97	0.59	1.94	6.02
RP	0.19	0.84	0.99	0.50	1.59	5.02	-0.5	0.8	0.99	0.47	1.73	5	0.15	0.84	0.99	0.45	1.72	6.12
RF	0.34	0.85	0.99	0.47	1.60	5.03	0.09	0.83	0.99	0.48	1.72	5.34	0.34	0.86	0.99	0.46	1.61	4.68
GradBoost	0.31	0.85	0.99	0.49	1.58	4.98	0.3	0.82	0.99	0.47	1.72	5.16	0.35	0.84	0.99	0.45	1.62	5.33

Table 5.1 : Performance des modèles pour les trois capteurs

D'après le tableau, nous remarquons que quelque soit le modèle de calibration, plus le coefficient de détermination est proche de 1 plus l'erreur quadratique moyenne est minime. Par exemple, avec la régression linéaire simple, la valeur moyenne minimale de r^2 est de 0.6 et correspond au plus grand RMSE moyen qui est de 2.34. Cela veut dire donc que plus le modèle est ajusté sur les données de référence moins sera l'erreur quadratique.

Nous remarquons aussi que les performances de calibration diffèrent d'un modèle à un autre. En effet, la valeur moyenne de r^2 et la valeur moyenne de RMSE varient respectivement entre 0.6 et 0.86 et entre 1.58 et 2.34. La figure 5.1 montre également ces variations avec la régression linéaire étant la méthode la moins performante comparée aux autres méthodes. De plus La régression linéaire montre des performances plus variables et parfois moins fiables entre les trois capteur. Cependant, elle atteint parfois un bon niveau de performance avec un r^2 proche de 1 (0.98) et un RMSE très faible (0.77). Cela est du au fait qu'il existe quelques mois pendant lesquels la corrélation entre le capteur et la référence est élevé ou la variabilité des données météorologiques

sont stables et présentent une légère variabilité.

De même, bien que la régression multiple offre une amélioration par rapport à la régression simple, ses performances demeurent un peu trop variables, contrairement ce que montrent les méthodes de régressions non linéaires, qui sont moins variables, ne présentent pas une grande différence entre elles (une différence de $\pm 1\%$ en r^2 et de $\pm 0.1\%$).

Enfin, les méthodes Random Forest et Gradboost se distinguent de la régression polynomiale en terme de performance, par exemple, la calibration du capteur 3 donne un RMSE maximal de 6.12 ce qui est inférieur ce que donnent Random Forest et Gradboost. Cela est expliqué par le fait que la régression polynomiale est sensible aux grandes concentration de $PM_{2.5}$. Nous remarquons que le capteur 2 est légèrement moins performant que les deux autres capteurs, en effet, il présente à chaque modèle le plus faible coefficient de détermination et la plus grande erreur quadratique. Nous voyons ça bien surtout avec la régression linéaire simple et la régression linéaire multiple.

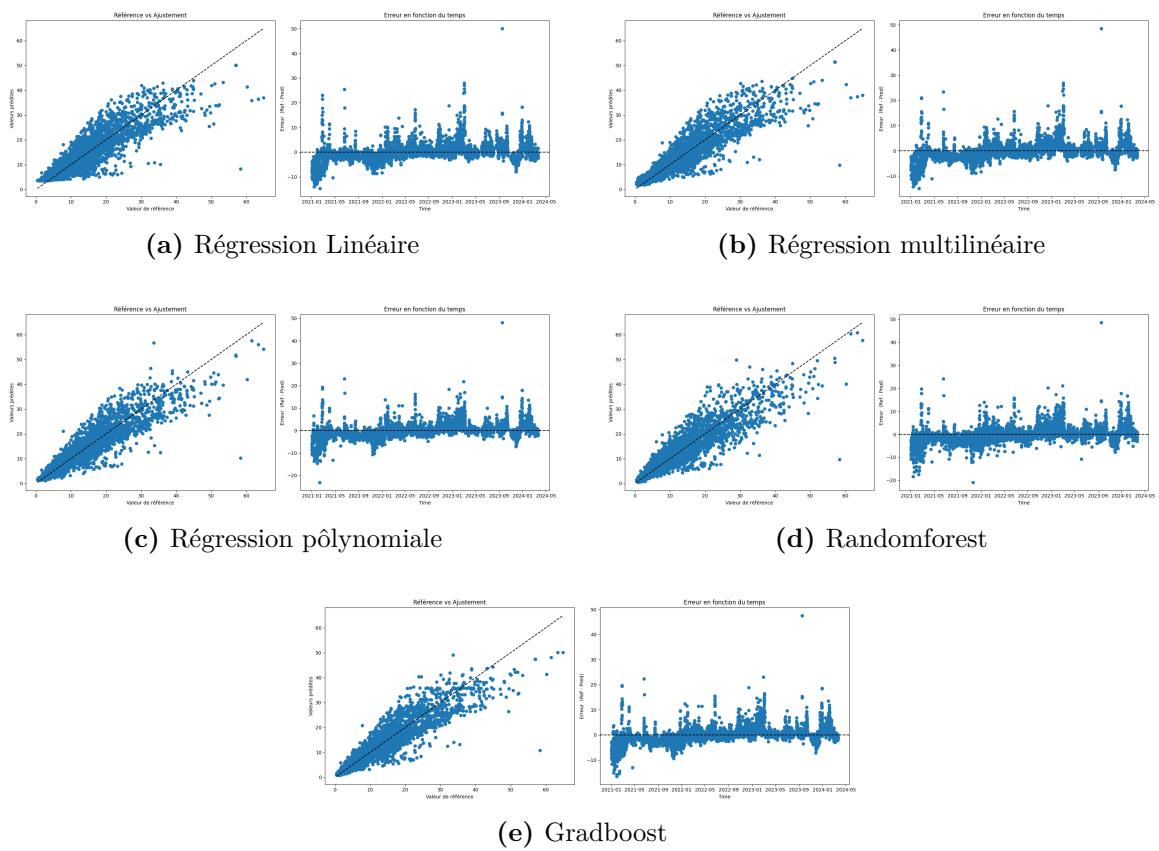


Figure 5.1 : Résultats des différents modèles de calibration sur le capteur 1

6 Analyse de l'impact du vieillissement

Dans ce chapitre, mon objectif est d'analyser l'impact du vieillissement sur les performances des mesures des capteurs de pollution à bas coût. En d'autres termes, je vais analyser l'évolution des performances des capteurs tout au long de leur période de déploiement. Pour ce faire, je me suis basé sur les données corrigées après utilisation des fonctions de calibration présentées dans le chapitre précédent. Plus précisément, je vais utiliser la méthode Random Forest dans le processus de calibration, ce qui me permettra d'isoler l'impact du vieillissement dans mes tests. En effet, les baisses en performance des capteurs calibrées sont dues principalement au vieillissement des composants des capteurs dans le cas où on utilise des fonctions de calibration qui permettent de corriger l'impact de la température et de l'humidité sur les mesures.

6.1 Modélisation mathématique

Soient $valeur_{calibrée}$ et $valeur_{brute}$, respectivement, les valeurs corrigées et les valeurs brutes du capteur de pollution ayant id comme identifiant. La relation entre valeur calibrée et valeur brute peut être écrite comme suit :

$$valeur_{calibrée}^{id} = f_t^{id}(valeur_{brute}^{id}, \text{humidité}, \text{température})$$

où f est la fonction de calibration optimale (théorique) qui permet de corriger au mieux les mesures des capteurs. Notons que f dépend de id car les capteurs ne dérivent pas de la même façon comme nous l'avons montré dans le chapitre précédent. Notons

aussi que f dans ce modèle mathématique dépend du temps car le vieillissement des composants impacte les performances des capteurs.

Mon objectif dans ce chapitre est de répondre aux questions suivantes :

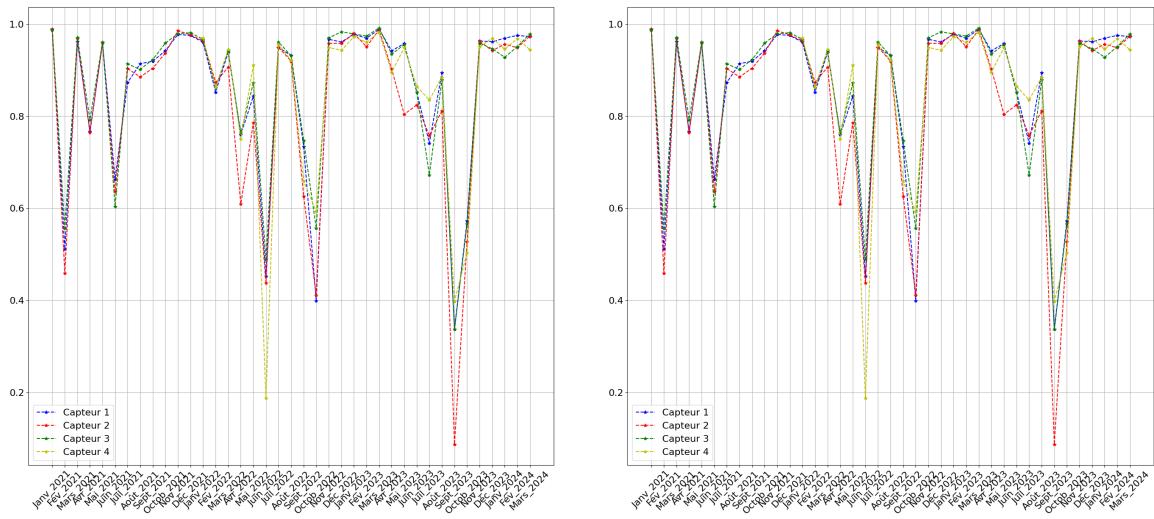
1. Est-ce que les performances des fonctions de calibration évoluent en fonction du temps. En d'autres termes, est-ce que le r^2 de f_t^{id} dépend de t ?
2. Est-ce que les fonctions de calibration sont interchangeables entre les différents capteurs et est-ce que cela évolue en fonction du temps ? En d'autres termes, est-ce que les capteurs vieillissent à la même vitesse ? D'un point de vue de mathématique, on cherche à étudier l'évolution de $f_t^{id2}(m_{brute}^{id1})$ en fonction de t .

6.2 Évolution des performances des fonctions de calibration

J'ai fait le choix de calibrer les micro-capteurs chaque mois pendant toute la période de déploiement des trois ans afin d'éviter l'effet de la saisonnalité (i.e. impact du changement de saison sur la nature des particules mesurées). Ensuite, j'ai divisé les données mensuelles de chaque capteur en deux parties : une première partie de 80% pour entraîner le modèle de calibration et les 20% restants pour les tests de validation.

La figure 6.1, montre le coefficient de détermination de la calibration de chaque capteur pour chaque mois de la période de déploiement. Notons que les trois premiers capteurs sont déployés depuis le premier janvier 2021, alors que le quatrième capteur (moins vieux) n'est déployé sur le site de mesure qu'à partir de janvier 2022.

Nous remarquons d'abord qu'il n'y a pas de différence importante de r^2 entre les premiers mois (par exemple janvier 2021 avec un r^2 de 0.99) et les derniers mois (par



(a) Evolution de R^2 en fonction des mois

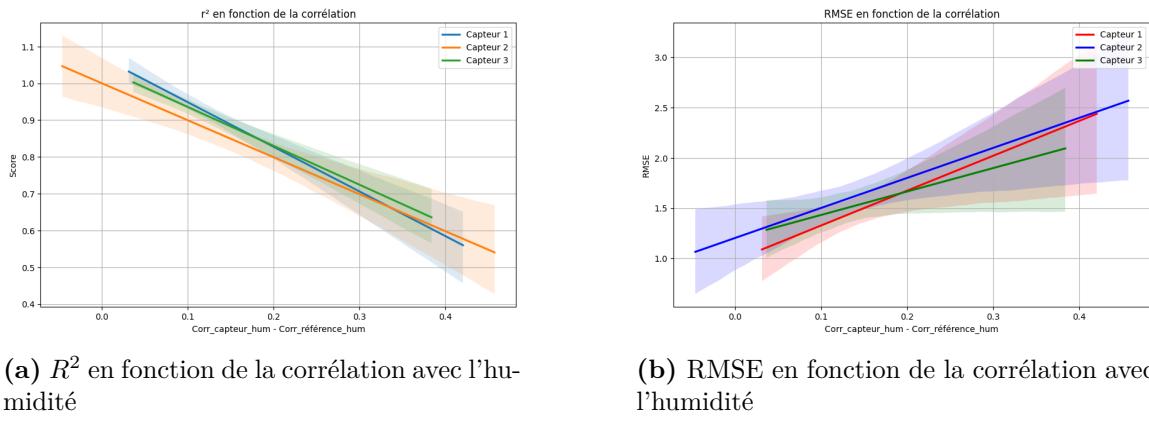
(b) Evolution de RMSE en fonction des mois

Figure 6.1 : Evolution de R^2 et RMSE en fonction des mois

exemple mars 2024 avec un r^2 de 0.97 au minimum). Nous remarquons également qu'il n'y a pas de différence importante entre les valeurs de r^2 du quatrième capteur par rapport aux autres capteurs, même si ce dernier était déployé un an après les autres capteurs.

En utilisant une fonction de calibration permettant de bien corriger régulièrement l'impact de la température et l'humidité, les performances des capteurs à bas coût ne baissent pas de façon importante en fonction du temps.

Cependant, nous remarquons également dans la figure 6.1 qu'il y a une dégradation de performance de calibration pour les quatre capteurs sur quelques mois à l'image du mois de février et juin 2021, et juin, septembre, octobre 2022 ainsi que septembre et octobre 2023. Afin d'expliquer ces variations de performance qui sont très irrégulières, je présente dans la figure 6.2 les résultats de l'analyse des performances des capteurs en fonction de la corrélation avec l'humidité.



(a) R^2 en fonction de la corrélation avec l'humidité

(b) RMSE en fonction de la corrélation avec l'humidité

Figure 6.2 : Analyse des performances des capteurs en fonction de la corrélation avec l'humidité.

La figure 6.2 montre l'évolution du coefficient de détermination r^2 et la RMSE des capteurs 1, 2 et 3 en fonction de la différence entre la corrélation capteur-humidité et la corrélation référence-humidité. Cette différence indique à quel point l'humidité affecte différemment les mesures des capteurs par rapport à la référence.

Nous remarquons que le r^2 diminue de manière linéaire quand la différence de corrélation entre l'humidité et les capteurs par rapport à la référence augmente. Cela montre que plus l'humidité a une influence différente sur les mesures des capteurs par rapport à la référence, moins le modèle de calibration est performant et c'est ce qui explique alors la dégradation irrégulière des performances des micro-capteurs.

Cela nous mène à déduire que l'humidité a un impact significatif sur les capteurs. Une forte différence dans la façon dont l'humidité affecte les capteurs par rapport à la référence rend la calibration moins efficace. Les capteurs sont sensibles aux variations de l'humidité. Cette sensibilité à l'humidité varie d'un capteur à un autre. En effet, l'intervalle de confiance large pour le capteur 2 montre que la performance de la calibration et donc du capteur est plus sensible à l'humidité que les deux autres

capteurs.

6.3 Comparaison du vieillissement de différents capteurs

Dans cette section, nous allons étudier la possibilité d'inter-changer les fonctions de calibration de différents capteurs tout en analysant leur vitesse de vieillissement. En effet, cela nous permettra d'évaluer la possibilité de calibrer les capteurs régulièrement pendant leur période de déploiement en s'appuyant sur les mesures d'un autre capteur déployé près d'une station de référence.

Pour ce faire, j'ai entraîné mon modèle de calibration chaque mois sur 80% des données d'un capteur de chaque mois ensuite je l'ai testé sur 20% des données de l'autre capteur.

Les résultats obtenus sont présentés dans la figure 6.3 montant le rapport de performance entre le capteur sur lequel j'ai appliqué la calibration et le capteur testé en fonction des mois. Nous remarquons d'abord qu'en calibrant les capteurs 2, 3, et 4 en utilisant les paramètres de calibration du capteur 1, nous avons une perte de performance remarquable. En effet, les capteurs 2, 3, 4 présentent un rapport de performance qui ne dépasse pas les 85% en général et qui est très variable pendant la période de déploiement. Cela est dû au fait que les différents capteurs ne sont pas parfaitement identiques en termes de composants électroniques même s'ils utilisent tous la même technologie de mesure, à savoir le phénomène de la dispersion de la lumière. Cette différence entre les capteurs résulte en des performances beaucoup plus faibles dans les cas où le problème d'humidité expliqué dans le test précédent

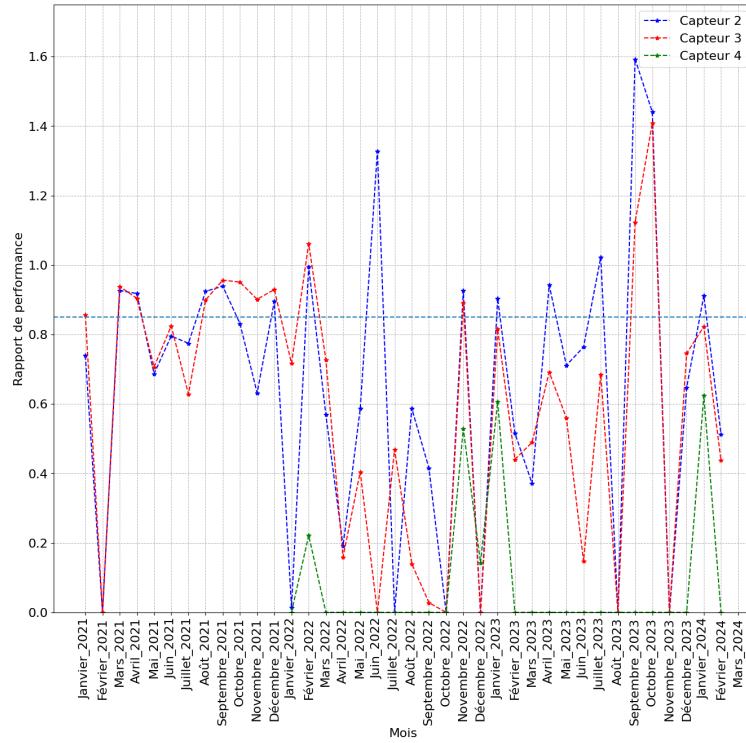


Figure 6.3 : Comparaison du vieillissement de différents capteurs

intervient. Par exemple, on obtient un coefficient de corrélation r^2 négatif au mois de février 2021 pour les capteurs 2 et 3 à cause du fait que le capteur 1 n'était pas performant pendant ce mois-ci.

La calibration d'un micro-capteur de $PM_{2.5}$ basée sur un autre micro-capteur qui est déployé dans un autre site géographique engendre une performance pouvant rendre les mesures inexploitables dans le cas où l'humidité interfère avec le capteur qui sert de base de calibration.

Nous remarquons également dans la figure 6.3 que les performances des capteurs 2 et 3 (qui reportent des données depuis la même date de début du capteur 1, à savoir depuis janvier 2021) baissent en fonction du temps. En effet, en comparant la première

année de déploiement avec la dernière année, nous remarquons que le r^2 obtenu pour les capteurs 2 et 3 est plus souvent variable et moins régulier pendant la deuxième partie de la période de déploiement des capteurs. Cela est expliqué par le fait que les différents capteurs ne vieillissent pas à la même vitesse, ce qui fait que la fonction de calibration obtenue avec l'un des capteurs s'adapte de moins en moins aux autres capteurs en fonction du temps.

Les performances de la calibration d'un micro-capteur de $PM_{2.5}$ basée sur un autre micro-capteur qui est déployé dans un autre site géographique baissent en fonction du temps jusqu'à atteindre un niveau où les mesures obtenues ne sont plus exploitables.

En ce qui concerne le capteur 4 (qui a été déployé bien plus tard que les autres capteurs et qui ne reporte des mesures qu'à partir de janvier 2022), nous remarquons dans la figure 6.3 que les performances obtenues sont tout le temps inférieure aux performances des capteurs 2 et 3. Le r^2 obtenu dans le cas du capteur 4 est en effet quasiment tout le temps négatif. Ce niveau de performance est dû au fait que le capteur 4 étant neuf par rapport au vieux capteur 1 qui sert de base de calibration, la fonction de calibration entraînée sur le capteur 1 n'est pas adapté au capteur 4 (dont les composants électroniques sont encore en bon état comparé au capteur 1).

Il n'est pas donc possible de calibrer un capteur de pollution neuf en se basant les résultats d'un vieux capteur déployé près d'une station de référence. La calibration croisée nécessite en effet d'utiliser des capteurs ayant le même niveau de vieillissement.

7 Conclusion

Dans ce stage, nous avons dans un premier lieu défini les termes clés du sujet du projet de fin d'étude qui se focalise sur les capteurs à faible coût des particules polluantes. Ces polluants suscitent en effet l'intérêt des agences environnementales pour surveiller la qualité de l'air grâce à leur faible coût ainsi que leur forte résolution spatiale et temporelle. Cependant, ces capteurs sont souvent susceptibles de fournir des mesures imprécises (pouvant être impactées par la météo, l'encrassement et le vieillissement), ce qui nécessite donc une évaluation de leur performance afin de confirmer leur fiabilité.

Nous nous sommes focalisés ainsi dans ce stage sur l'impact du vieillissement des capteurs sur leur performance. La calibration nous a servi comme un moyen d'ajustement des valeurs des capteurs afin de pouvoir évaluer les capteurs et déterminer les facteurs causant leur dégradation. En effet, j'ai pu confirmer ce qui a été énoncé par les articles de ma bibliographie, et que l'humidité a un impact sur les capteurs en les rendant de moins en moins fiables même après les avoir calibrés.

J'ai pu monter aussi que les capteurs sont impactés par le vieillissement qui est du à l'usure de ces composants tels que l'intensité du laser qui sert à disperser les particules ou la vitesse à laquelle le ventilateur fait circuler l'air. Ainsi, les micro-capteurs n'ont pas la même vitesse de vieillissement, ce qui montre que les utilisateurs de ces capteurs ne peuvent pas corriger un micro-capteur à l'aide d'un autre s'ils ne sont pas au même niveau de vieillissement.

Bibliographie

1. BORGHI, F., SPINAZZÈ, A., ROVELLI, S., CAMPAGNOLO, D., DEL BUONO, L., CATTANEO, A. & CAVALLO, D. Miniaturized monitors for assessment of exposure to air pollutants : a review. *International journal of environmental research and public health* **14**, 909. doi :[10.3390/ijerph14080909](https://doi.org/10.3390/ijerph14080909) (2017).
2. WAYLAND, R. & MATHIAS, S. Guidance for ozone and fine particulate matter permit modeling.
3. BADURA, M., BATOG, P., DRZENIECKA-OSSIADACZ, A. & MODZEL, P. Evaluation of low-cost sensors for ambient PM_{2.5} monitoring. *Journal of sensors* **2018**, 1-16. doi :[10.1155/2018/5096540](https://doi.org/10.1155/2018/5096540) (2018).
4. BORREGO, C. *et al.* Assessment of air quality microsensors versus reference methods : the EuNetAir joint exercise. *Atmospheric environment* **147**, 246-263. doi :[10.1016/j.atmosenv.2016.09.050](https://doi.org/10.1016/j.atmosenv.2016.09.050) (2016).
5. FEENSTRA, B., PAPAPOSTOLOU, V., HASHEMINASSAB, S., ZHANG, H., BOGHOSSIAN, B. D., COCKER, D. & POLIDORI, A. Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. *Atmospheric environment* **216**, 116946. doi :[10.1016/j.atmosenv.2019.116946](https://doi.org/10.1016/j.atmosenv.2019.116946) (2019).
6. KANG, Y., AYE, L., NGO, T. D. & ZHOU, J. Performance evaluation of low-cost air quality sensors : a review. *Science of the total environment* **818**. doi :[10.1016/j.scitotenv.2021.151769](https://doi.org/10.1016/j.scitotenv.2021.151769) (2022).
7. ALFANO, B. *et al.* A review of low-cost particulate matter sensors from the developers' perspectives. *Sensors* **20**, 6819. doi :[10.3390/s20236819](https://doi.org/10.3390/s20236819) (2020).

-
8. JIAO, W. *et al.* Community air sensor network (CAIRSENSE) project : evaluation of low-costsensor performance in a suburban environment in the southeas-
tern UnitedStates. *Atmospheric measurement techniques* **9**, 5281-5292. doi :[10.5194/amt-9-5281-2016](https://doi.org/10.5194/amt-9-5281-2016) (2016).
 9. RAI, A. C., KUMAR, P., PILLA, F., SKOLOUDIS, A. N., DI SABATINO, S., RATTI, C., YASAR, A. & RICKERBY, D. End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of the total environment* **607-608**, 691-705. doi :[10.1016/j.scitotenv.2017.06.266](https://doi.org/10.1016/j.scitotenv.2017.06.266) (2017).
 10. KARAGULIAN, F., BARBIERE, M., KOTSEV, A., SPINELLE, L., GERBOLES, M., LAGLER, F., REDON, N., CRUNAIRE, S. & BOROWIAK, A. Review of the per-
formance of low-cost sensors for air quality monitoring. *Atmosphere* **10**, 506.
doi :[10.3390/atmos10090506](https://doi.org/10.3390/atmos10090506) (2019).
 11. MARATHE, S., NAMBI, A., SWAMINATHAN, M. & SUTARIA, R. *CurrentSense : a novel approach for fault and drift detection in environmental IoT sensors* in *Proceedings of the international conference on internet-of-things design and im-
plementation* IoTDI '21 : International Conference on Internet-of-Things Design
and Implementation (ACM, Charlottesvle VA USA, 2021), 93-105. doi :[10.1145/3450268.3453535](https://doi.org/10.1145/3450268.3453535).
 12. CHEN, L.-J., HO, Y.-H., HSIEH, H.-H., HUANG, S.-T., LEE, H.-C. & MAHAJAN,
S. ADF : an anomaly detection framework for large-scale PM2.5 sensing systems.
IEEE internet of things journal **5**, 559-570. doi :[10.1109/JIOT.2017.2766085](https://doi.org/10.1109/JIOT.2017.2766085) (2018).

A Annexe

La figure A.1 montre les concentrations mesurées par les capteurs à faibles coût pendant la dernière semaine du mois de février.

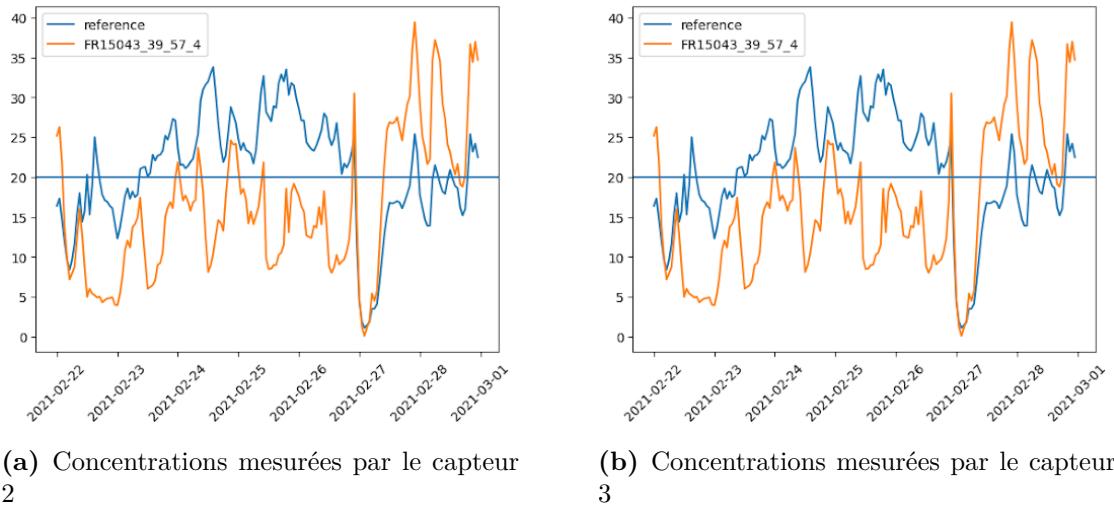


Figure A.1 : Concentrations mesurées par les capteurs 2 et 3 pendant la dernière semaine de février 2021

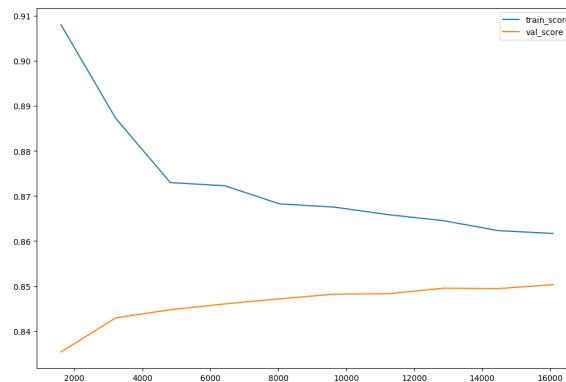


Figure A.2 : Courbe d'apprentissage du modèle Random Forest