CALIBRAGE IN SITU DE CAPTEURS POUR LA SURVEILLANCE DES $PM_{2,5}$ DANS L'AIR AMBIANT : PERFORMANCE ET TRANSFERABILITE DES MODELES DE MACHINE LEARNING

S. Masmoudi*[1], M. Tossa[2], S. Crunaire[1], N. Redon[1], L. Spinelle[3], A. Gressent[3] et T. Macé[2]

[1]Centre de Recherche Énergie Environnement (CERI EE), IMT Nord Europe, Université de Lille, 59000 Lille, France
[2]Laboratoire national de métrologie et d'essais (LNE), 75724 Paris, France
[3]Institut National de l'Environnement Industriel et des Risques (INERIS),60550 Verneuil-en-Halatte, France
*sahar.masmoudi@imt-nord-europe.fr

TITLE

In situ calibration of low-cost sensors used for $PM_{2.5}$ measurement in ambient air: Performance and transferability of Machine Learning models

RESUME

La surveillance des particules fines dans l'air ambiant est cruciale pour la santé publique et l'environnement. Les capteurs à bas coût, offrent une alternative pour collecter des données localisées et en temps réel qui permettent de compléter avantageusement le réseau de stations de mesures fixes. Néanmoins, une fois déployés sur le terrain, la qualité des données de ces dispositifs est susceptible de se dégrader. Il est donc nécessaire de proposer des techniques de calibrage facile à implémenter sans déplacer les capteurs pour garantir des mesures plus fiables dans le temps. Cette étude vise à évaluer les performances de différents modèles de Machine Learning (régression linéaire multiple, forêts aléatoires, Long Short Term Memory LSTM, etc.) pour le calibrage de capteurs de particules $PM_{2,5}$ en colocalisation avec un instrument de référence, puis à tester la transférabilité du modèle pour corriger les mesures des capteurs déployés dans des zones plus éloignées. Les données de la base SensEURcity, collectées à Anvers, révèlent que le modèle LSTM est le plus précis, avec un RMSE moyen de 8,79 µg/m³, et capture bien les dépendances temporelles des mesures. L'évaluation de la transférabilité montre une fiabilité accrue des capteurs déployés, avec un Coefficient de Corrélation Intra-classe (CCI) supérieur à 0,75 et une nette réduction du RMSE normalisé après calibrage.

ABSTRACT

Monitoring fine particulate matter in ambient air is critical for public health and environment. Low-cost sensors provide an alternative for collecting real-time data, complementing regulatory measurement station networks. However, once deployed in the field, the data quality from these devices can degrade over time. It is therefore necessary to propose calibration techniques that are easy to implement without moving sensors to guarantee reliable measurements over time. This study evaluates the performance of various Machine Learning models (e.g., multiple linear regression, random forests, Long Short-Term Memory networks LSTM) for calibrating $PM_{2.5}$ sensors collocated with a reference instrument, and then to test the transferability of the model to correct measurements from sensors deployed in distant areas. Using the SensEURcity dataset collected in Antwerp, the LSTM model demonstrated the highest accuracy, achieving an average RMSE of 8.79 µg/m³ while effectively capturing temporal dependencies. Transferability assessments revealed improved reliability for deployed sensors, with an Intraclass Correlation Coefficient (ICC) above 0.75 and a significant reduction in normolized nRMSE after calibration.

MOTS-CLÉS: capteur, particules fines, calibrage, Machine Learning, LSTM, transférabilité
KEYWORDS: sensor, fine particulate, calibration, Machine Learning, LSTM, transferability

1. INTRODUCTION

Compte tenu des effets néfastes des particules fines (PM_{10}, $PM_{2,5}$), la surveillance de la pollution particulaire est considérée comme de toute nécessité en matière d'environnement et de santé. Toutefois, les réseaux classiques de stations de mesure de la qualité de l'air, bien qu'efficaces, sont coûteux à installer et à maintenir, et leur nombre réduit limite la couverture géographique. Pour pallier ces insuffisances, les réseaux de capteurs à faible coût ont émergé comme une solution complémentaire. Ces capteurs permettent une densification du réseau de mesure et offrent la possibilité d'obtenir des données localisées en temps réel (Karagulian et al., 2019; Kumar et al., 2015). Cependant malgré leurs avantages, ces systèmes capteurs présentent une qualité de données nécessitant des ajustements pour garantir leur fiabilité et leur exploitation. Le calibrage sur le terrain de ces capteurs devient alors une étape indispensable du processus d'assurance qualité/ contrôle qualité (QA/QC) des données.

Le calibrage sur le terrain consiste à colocaliser les capteurs à faible coût avec une station de référence certifiée sur une période déterminée, voire en continu, afin de développer un modèle de correction alignant les données des capteurs avec celles des instruments de référence. Parmi les méthodes de calibrage couramment utilisées figurent les régressions linéaires simples et multiples, mais ces approches ne permettent

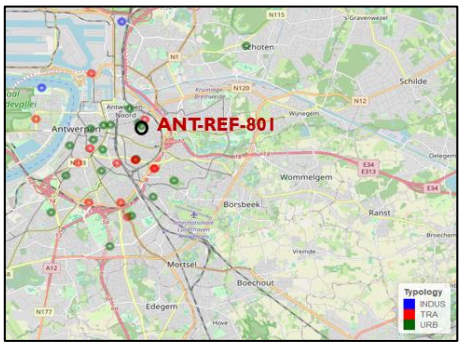
DOI : 10.25576/ASFERA-CFA2025-43920 Cet article est publié sous la seule responsabilité de ses auteurs

pas de capturer les non-linéarités ni les processus atmosphériques complexes. Des erreurs peuvent ainsi survenir, notamment en raison du vieillissement des capteurs et des variations des conditions environnementales. Les approches de machine Learning (ML) apparaissent prometteuses pour gérer ces non-linéarités et exploiter plus efficacement les variables (Villanueva et al., 2023). Dans ce contexte, cette étude vise deux objectifs principaux : i) évaluer les performances de différents modèles de Machine Learning pour le calibrage in situ de capteurs de particules colocalisés avec des stations de référence, ii) examiner la capacité de ces modèles à se généraliser pour calibrer des capteurs déployés dans des zones éloignées des stations de référence.

2. DONNEES ET METHODES

2.1. Collecte de données

La base de données SensEURcity (Van Poppel et al., 2023), récemment rendue public (mai 2023), a été utilisée pour cette étude. Elle regroupe des mesures de plusieurs polluants et paramètres météorologiques issues d'instruments de référence et de capteurs déployés dans trois grandes villes européennes : Anvers, Oslo et Zagreb. Pour cette étude, seules les données collectées à Anvers ont été exploitées, plus précisément celles provenant de la station urbaine de référence **ANT-REF-801**. Ce choix s'explique par une importante densité de capteurs sur la zone étudiée et la durée de collecte de mesures la plus longue. La figure 1 présente un résumé des métadonnées utilisées.



Données utilisées	Description
Station référence à Anvers	ANT-REF-801 station urbaine
Nombre de capteurs colocalisé	2 capteurs colocalisés avec la REF
Période de collecte de données	Avril 2020 - Avril 2021
Nombre de capteurs déployés	11 capteurs déployés en zone urbaines
Variables collectées	Particules fines $PM_{2,5}$, Température, Humidité relative
Type de capteurs	PMS 5003
Instrument de référence	Fidas
Résolution temporelle	Capteur : minute, Réf : horaire

Figure 1. Localisation des sites d'échantillonnage à Anvers et description des données collectées dans la base SeneEURcity (Van Poppel et al., 2023)

2.2. Méthodologie

- **Développement du modèle de calibration**

La méthodologie repose sur un processus en trois étapes visant à valider et corriger les données collectées par les capteurs (*c.f.* figure 2).

La première étape consiste à prétraiter et fusionner les données. En raison des imperfections présentes dans les données brutes des capteurs, des étapes préliminaires de prétraitement ont été choisies et effectuées avant le développement et l'implémentation du modèle de correction. Ces étapes incluent : *i)* l'élimination des valeurs dépassant les maxima mesurés par l'instrument de référence, *ii)* la suppression des valeurs figées pendant 10 minutes et *iii)* l'exclusion des valeurs manquantes sans remplacement. Les données de référence, collectées à une fréquence horaire, n'étant pas synchrones avec les mesures des capteurs enregistrées chaque minute, les données des capteurs ont été moyennées à une résolution horaire afin d'harmoniser les fréquences de mesure. Après ce prétraitement, la corrélation entre les capteurs colocalisés a été évaluée, révélant un coefficient de détermination élevé (R^2 = 0,92). Sur cette base, les ensembles de données de chaque capteur ont été fusionnés en une base de données unifiée, noté $D_{fusion} = concaténation\ (D_1, D_2, \ldots\ldots, D_k)$, afin de constituer un ensemble de données consolidé.

La deuxième étape consiste à évaluer les performances de différentes approches de Machine Learning pour calibrer les capteurs de particules $PM_{2,5}$, en utilisant les données de colocalisation en continu de l'instrument de référence. Pour développer un modèle généralisé de correction, plusieurs approches de Machine Learning ont été testées : régression linéaire multiple, arbres de décision, forêt aléatoire, etc... Les variables utilisées pour le développement des modèles sont les mesures de $PM_{2,5}$, températures et humidité relatives. De ces données, 80% ont été utilisés pour l'entrainement et 20% pour le test, et pour préserver la continuité temporelle des séries, le test est placé juste après l'entraînement. Afin de garantir une évaluation robuste du modèle une validation croisée des séries temporelles est répétée 10 fois.

La troisième étape consiste au transfert du modèle. Après avoir comparé et évalué différents modèles selon les métriques présentées dans les paragraphes suivants, le meilleur est transféré aux capteurs déployés dans les zones urbaines d'Anvers, garantissant ainsi des données corrigées et harmonisées pour l'ensemble du réseau de capteurs.

DOI : 10.25576/ASFERA-CFA2025-43920 Cet article est publié sous la seule responsabilité de ses auteurs

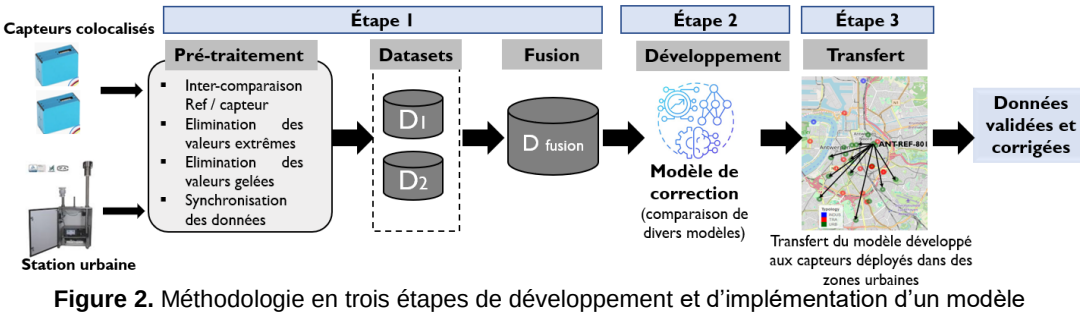


Figure 2. Méthodologie en trois étapes de développement et d'implémentation d'un modèle

- **Evaluation des performances des modèles**

La comparaison des performances des modèles testés a été réalisée en se basant sur les métriques suivantes : la racine de l'erreur quadratique moyenne (RMSE) et le coefficient de corrélation de Pearson (r). Soit $x(t)$ la concentration prédite par le modèle et $y(t)$ celle mesurée par la référence à l'instant *t*.

Racine de l'écart quadratique moyen (RMSE) évalue l'écart entre les concentrations $PM_{2,5}$ prédites par un modèle et celles mesurées par l'instrument de référence.

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{t=1}^{N_{test}} (y_i(t) - y_i(t))^2} \qquad \text{Équation 1}$$

Coefficient de corrélation pearson (r) mesure l'intensité de la relation linéaire entre les concentrations $PM_{2,5}$ prédites par un modèle et celles mesurées par l'instrument de référence.

$$r = \frac{\sum_{t=1}^{N_{test}} (x_i(t) - \bar{x})\,(y_i(t) - \bar{y})}{\sqrt{\sum_{t=1}^{N_{test}} (x_i(t) - x)^2 \sum_{t=1}^{N_{test}} (y_i(t) - \bar{y})^2}} \qquad \text{Équation 2}$$

- **Évaluation de la transférabilité du modèle sélectionné**

Pour évaluer l'effet de la transférabilité du modèle dans la calibration des données issues des capteurs, les méthodes statistiques proposées par Qiao et al. (Qiao et al., 2021) ont été utilisées pour vérifier la cohérence entre les mesures des capteurs X et celles de l'instrument de référence Y, tant en termes de dynamique temporelle qu'en valeur vraie. Deux métriques sont définies :

Le coefficient de corrélation intra-classe (CCI) vérifie si $\boldsymbol{X}$ et $\boldsymbol{Y}$ suivent la même tendance et possèdent la même dynamique. Un CCI supérieur à 0,75 indique une bonne fiabilité du capteur.

$$CCI = \frac{D(X+Y) - D(X-Y)}{D(X+Y) + D(X-Y)} \qquad \text{Équation 3}$$

Où $\boldsymbol{D}$ représente l'opérateur arithmétique de la variance.

L'écart quadratique moyen normalisé (nRMSE) utilisée pour évaluer si X et Y sont en concordance.

$$nRMSE = 1 - \frac{e^{\frac{-\sqrt{E[(X-Y)^2]}}{E(Y)}}}{1 + \frac{\sqrt{E[(X-Y)^2]}}{E(Y)}} \qquad \text{Équation 4}$$

3. RÉSULTATS ET DISCUSSIONS

- **Comparaison des modèles de calibrage**

La figure 3 présente les résultats de comparaison des performances de divers algorithmes de calibrage en termes de RMSE moyen. Le modèle LSTM présente le plus faible RMSE, égal à 8,79 µg/m³. À l'inverse, l'algorithme Adaboost présente le RMSE le plus élevé, à 13,78 µg/m³. Ces résultats mettent en évidence l'importance de la capacité des réseaux de neurones LSTM à capturer les dépendances temporelles entre les mesures. En effet, contrairement aux approches classiques, les LSTM sont spécialement conçus pour traiter les données séquentielles, ce qui leur permet de reconnaître et d'exploiter les relations temporelles dans les séries de données. Cette caractéristique est particulièrement utile pour le calibrage in situ des données de capteurs, car leurs mesures présentent souvent des fluctuations et des tendances qui évoluent avec le temps. Dans ce contexte, le LSTM se révèle performant pour corriger ces dérives, en intégrant des informations issues des variations précédentes et en prédisant les valeurs futures de manière plus précise.

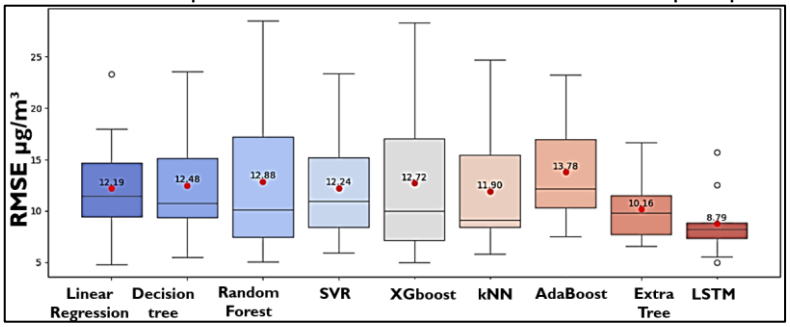


Figure 3. Comparaison des RMSE des modèles de calibrage appliqués à des séries de données capteurs $PM_{2,5}$

DOI : 10.25576/ASFERA-CFA2025-43920
Cet article est publié sous la seule responsabilité de ses auteurs

- **Transférabilité du modèle**

La figure 4 illustre les concentrations de $PM_{2,5}$ après correction par le modèle LSTM pour l'un des capteurs déployé dans la zone urbaine. La courbe en bleu représente les valeurs prises en tant que référence mesurées à l'aide d'un FIDAS, la courbe en vert correspond aux données brutes du capteur et la courbe en rouge montre les mesures corrigées par le modèle LSTM. La correction permet de réduire considérablement les écarts entre les mesures du capteur et l'instrument de référence, ce qui confirme bien l'efficacité du modèle LSTM pour ajuster les données sur le terrain en fonction des variations et des erreurs de mesure précédentes.

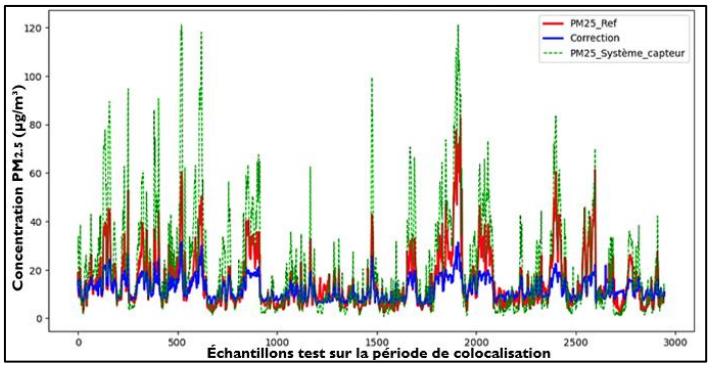


Figure 4. Séries de données de concentrations en $PM_{2,5}$ de référence, brute et corrigée par LSTM

La figure 5 présente l'impact de la correction apportée par le modèle pour les capteurs déployés en s'appuyant sur les indicateurs CCI et nRMSE. Les résultats montrent qu'avant calibrage, plusieurs capteurs ont un CCI inférieur à 0,5, indiquant une faible fiabilité et une mauvaise concordance avec les mesures de référence. De plus, la plupart des capteurs présentent un nRMSE supérieur à 0,8, seuil considéré comme inacceptable (Qiao et al., 2021), ce qui traduit une erreur élevée dans les mesures. Après calibrage par le modèle, les résultats montrent une nette amélioration : la majorité des capteurs affichent un CCI supérieur à 0,75 et un nRMSE réduit en dessous de 0,5 ce qui constitue des seuils de fiabilité acceptable. Cette réduction de l'erreur et l'augmentation de l'ICC démontrent à nouveau l'efficacité du modèle LSTM dans la correction et par la suite l'assurance qualité des données capteurs.

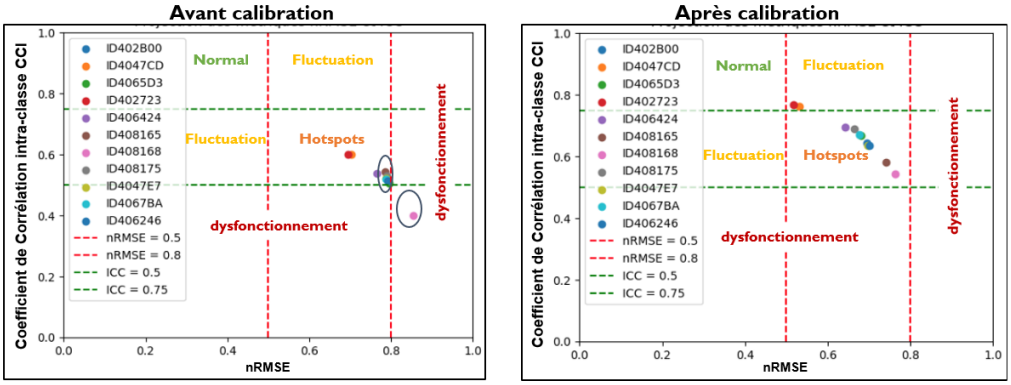


Figure 5. Evaluation de l'effet de calibrage sur les performances des capteurs

REMERCIEMENT

Les auteurs remercient le Laboratoire Central de Surveillance de la Qualité de l'Air (LCSQA) qui a permis le financement de cette étude à travers les travaux menés par le GT micro-capteurs.

RÉFÉRENCES

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., & Borowiak, A. (2019). Review of the Performance of Low-Cost Sensors for Air Quality Monitoring. *Atmosphere*, *10*(9), 506. https://doi.org/10.3390/atmos10090506

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., & Britter, R. (2015). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, *75*, 199–205. https://doi.org/10.1016/j.envint.2014.11.019

Qiao, X., Zhang, Q., Wang, D., Hao, J., & Jiang, J. (2021). Improving data reliability: A quality control practice for low-cost PM2.5 sensor network. *Science of The Total Environment*, *779*, 146381. https://doi.org/10.1016/j.scitotenv.2021.146381

Van Poppel, M., Schneider, P., Peters, J., Yatkin, S., Gerboles, M., Matheeussen, C., Bartonova, A., Davila, S., Signorini, M., Vogt, M., Dauge, F. R., Skaar, J. S., & Haugen, R. (2023). SensEURCity: A multi-city air quality dataset collected for 2020/2021 using open low-cost sensor systems. *Scientific Data*, *10*(1), 322. https://doi.org/10.1038/s41597-023-02135-w

Villanueva, E., Espezua, S., Castelar, G., Diaz, K., & Ingaroca, E. (2023). Smart Multi-Sensor Calibration of Low-Cost Particulate Matter Monitors. *Sensors*, *23*(7), 3776. https://doi.org/10.3390/s23073776