

Surveillance intelligente et prediction de la qualite de l'air au Grand Marche de Kinshasa basee sur l'Internet des objets et l'apprentissage automatique

[Intelligent Monitoring and Air Quality Prediction at the Kinshasa Central Market Based on the Internet of Things and Machine Learning]

¹Donatien KADIMA MUAMBA, ²Bernice MAMBULA ZUIYA

¹Faculté des Sciences Informatiques de l'Université Reverend KIM

²Faculté des Sciences Informatiques de l'Université Protestante au Congo

Abstract-Air pollution is a major environmental and public health concern in rapidly growing urban areas. In many developing cities, the lack of real-time monitoring systems limits the availability of reliable information for environmental management and decision-making. This study presents the design and implementation of an intelligent air quality monitoring and prediction system deployed at the Grand Market of Kinshasa. The proposed system combines Internet of Things technologies, cloud-based data management, machine learning techniques, and web-based visualization tools. Environmental data were collected through air quality sensors connected to an embedded microcontroller and transmitted to a cloud platform for storage and analysis. A web dashboard was developed to provide real-time monitoring, historical data visualization, and predictive insights. Furthermore, association rule mining using the Apriori algorithm was applied to discover relationships among pollutants, while a Random Forest model was employed to forecast future pollutant concentrations based on historical observations. The experimental results demonstrate the capability of the proposed system to monitor air quality and extract meaningful knowledge from environmental data. The predictive model achieved satisfactory performance for particulate matter prediction, particularly for PM_{2.5} ($R^2 = 0.84$) and PM₁₀ ($R^2 = 0.94$), while lower predictive accuracy was observed for gaseous pollutants, highlighting the complexity of their temporal behavior. Overall, the model obtained a Mean Absolute Error of 0.30, a Root Mean Square Error of 0.80, and a coefficient of determination of 0.38. The study confirms that the integration of Internet of Things technologies, data mining techniques, and machine learning methods provides a practical framework for continuous air quality monitoring and supports data-driven environmental management in urban environments.

Resume

La pollution atmosphérique constitue un enjeu majeur de santé publique et de protection de l'environnement dans les zones urbaines à forte croissance démographique. Dans de nombreuses villes en développement, l'absence de systèmes de surveillance en temps réel limite la disponibilité d'informations fiables pour la gestion environnementale et l'aide à la décision. Cette étude présente la conception et la mise en œuvre d'un système intelligent de surveillance et de prédiction de la qualité de l'air déployé au Grand Marché de Kinshasa. La solution proposée combine les technologies de l'Internet des objets (IoT), le stockage des données dans le cloud, les techniques de fouille de données et les méthodes d'apprentissage automatique.

Les données environnementales ont été collectées à l'aide de capteurs de qualité de l'air connectés à un microcontrôleur embarqué, puis transmises vers une plateforme cloud pour leur stockage et leur analyse. Un tableau de bord web a été développé afin d'assurer la visualisation en temps réel des mesures, la consultation de l'historique et le suivi des indicateurs environnementaux. En complément, l'algorithme Apriori a été utilisé pour extraire les relations significatives entre les polluants, tandis qu'un modèle Random Forest a été mis en œuvre pour prédire l'évolution future des concentrations à partir des observations historiques.

Les résultats obtenus démontrent la capacité du système à surveiller efficacement la qualité de l'air et à extraire des connaissances pertinentes à partir des données collectées. Le modèle de prédiction présente de bonnes performances pour les particules fines, notamment les PM_{2.5} ($R^2 = 0,84$) et les PM₁₀ ($R^2 = 0,94$), alors que les performances demeurent plus limitées pour les polluants gazeux, traduisant la complexité de leurs dynamiques temporelles. Globalement, le modèle

a obtenu une erreur absolue moyenne (MAE) de 0,30, une racine de l'erreur quadratique moyenne (RMSE) de 0,80 et un coefficient de détermination (R^2) de 0,38.

Cette étude confirme que l'intégration de l'Internet des objets, de la fouille de données et de l'apprentissage automatique constitue une approche pertinente pour la surveillance continue de la qualité de l'air et l'aide à la gestion environnementale en milieu urbain.

Keywords: Air quality monitoring; Internet of Things; ESP32; SDS011; MQ123; Random Forest Regressor; Apriori algorithm; Environmental monitoring; Air pollution prediction; Supabase.

Remerciements

Les auteurs remercient l'Université Protestante au Congo et Université Révérend KIM pour l'encadrement scientifique ayant permis la réalisation de cette étude. Ils expriment également leur gratitude à toutes les personnes ayant contribué, directement ou indirectement, à la collecte des données, au développement du système de surveillance de la qualité de l'air et à l'aboutissement de ce travail.

1. INTRODUCTION

La dégradation de la qualité de l'air est devenue l'un des principaux défis en matière d'environnement et de santé publique et suscite le plus d'inquiétude, notamment dans les zones surpeuplées [3] avec une circulation humaine accrue et une densité de population élevée, comme la vie urbaine. Le Grand Marché de Kinshasa est pollué en raison des émissions provenant des transports routiers, des activités commerciales et de l'utilisation diversifiée de combustibles. Néanmoins, avec la disponibilité limitée d'outils de surveillance, le suivi continu des agents aéroportés reste difficile, compliquant la surveillance en temps réel de l'exposition aux polluants [5].

Dans l'ensemble, l'objectif ultime de cette recherche est de créer un système intelligent par nature qui contrôlera, analysera et prévoira la qualité de l'air avec des gadgets intelligents intégrés comme l'Internet des objets (IoT), l'algorithme Apriori [2] et un modèle basé sur les forêts aléatoires [1]. Le système visera à recueillir les données environnementales collectées en temps réel, à corrélérer les éléments de différents polluants et à prédire l'évolution des futurs polluants environnementaux pour une meilleure prise de décision.

2. MATERIELS ET METHODES

2.1. SITE D'ETUDE

L'étude a été réalisée au Grand Marché de Kinshasa, caractérisé par une forte densité de population, une circulation intense et diverses activités commerciales susceptibles d'influencer la qualité de l'air.



Fig.1 : Grand Marché de Kinshasa

2.2. Matériels utilisés

Le système développé comprend :

- Un microcontrôleur ESP32 ;

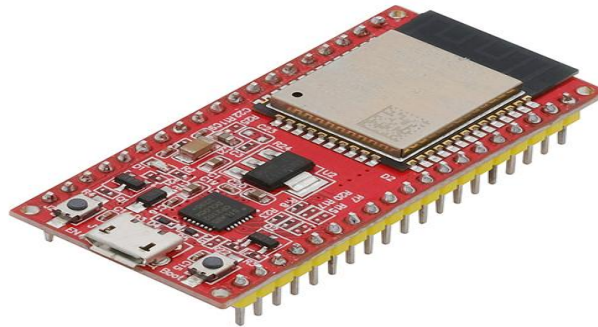


Fig.2 : Microcontrôleur ESP32-S2

- Un capteur SDS011 pour la mesure des particules PM2.5 et PM10 ;



Fig.3 : Capteur SDS011

- Un capteur MQ123 pour la détection des gaz atmosphériques ;



Fig.3 : Capteur MQ123

- Une LED de signalisation ;



Fig.3 : LED Rouge

- Une plateforme Supabase pour le stockage des données ;

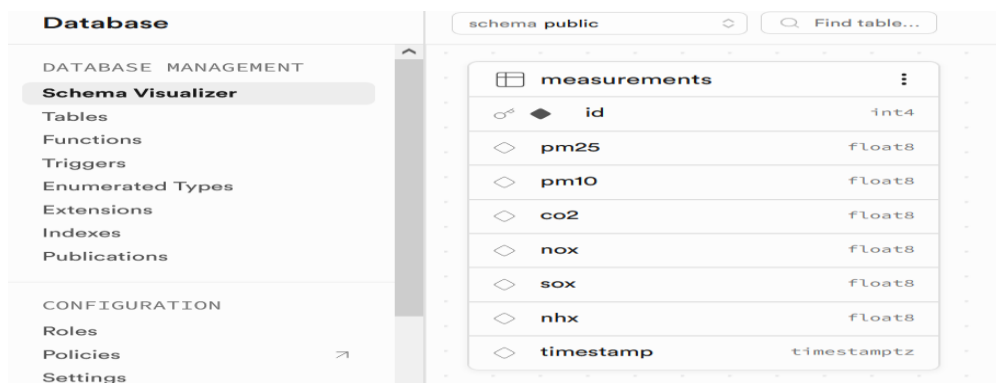


Fig.4 : Supabase

- Une API FastAPI pour le traitement des données ;

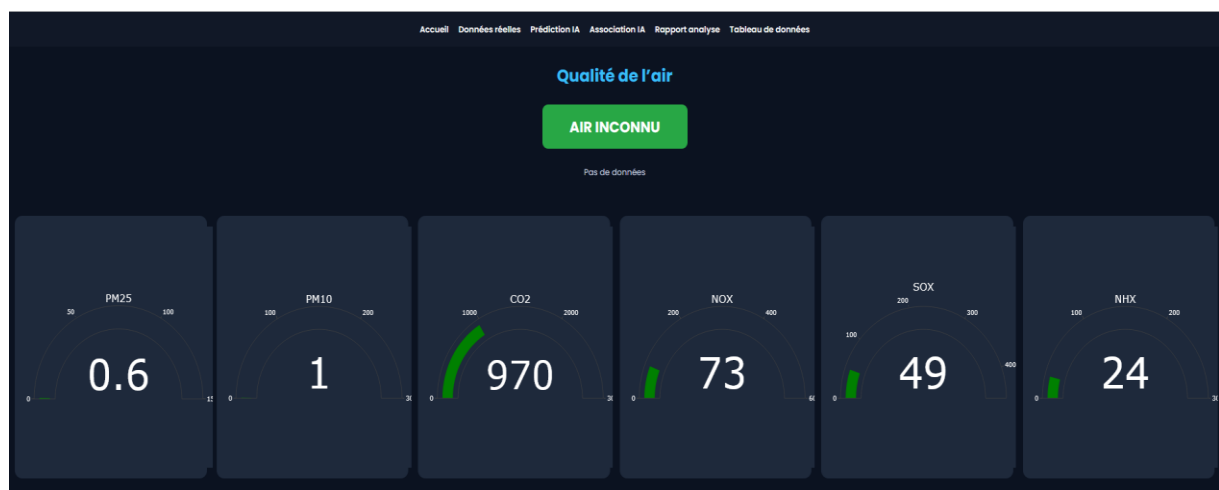


Fig.4 : API FastAPI

- Un tableau de bord web pour la visualisation des résultats.

Données brutes (Dataset) via Capteurs

Nombre total d'enregistrements : 500

Télécharger CSV

PM2.5	PM10	CO2	NOx	SOx	NHx
Min : 0.25	Min : 0.26	Min : 888.40	Min : 66.63	Min : 44.42	Min : 22.21
Moy : 3.96	Moy : 10.21	Moy : 974.58	Moy : 73.09	Moy : 48.73	Moy : 24.36
Max : 4.78	Max : 14.68	Max : 1048.60	Max : 78.64	Max : 52.43	Max : 26.21

Date	PM2.5	PM10	CO2	NOx	SOx	NHx
09/06/2026 18:27:24	1.32	1.39	905.98	67.95	45.3	22.65
09/06/2026 18:27:18	1.31	1.38	906.47	67.99	45.32	22.66
09/06/2026 18:27:12	1.29	1.36	908.91	68.17	45.45	22.72
09/06/2026 18:27:05	1.28	1.35	906.47	67.99	45.32	22.66

Fig.5 : Interface web de visualisation

2.3. Collecte et stockage des données

Les informations environnementales sont périodiquement recueillies par des capteurs, puis envoyées via Wi-Fi à la plateforme Supabase. Les mesures collectées concernent les concentrations de PM2.5, PM10, CO₂, NO_x, SO_x et NH_x. Les données sont ensuite conservées dans une base de données centralisée pour des usages futurs.



Fig.6 : Interface web de collecte de données

2.4. Analyse des données par l'algorithme Apriori

L'algorithme Apriori a été appliqué aux données collectées afin d'identifier les associations fréquentes entre les différents polluants atmosphériques [2]. Cette approche permet de mettre en évidence les relations de cooccurrence entre les niveaux de pollution observés.

Dataset transformé (1 = dépassement, 0 = sous seuil)

Show 10 entries

CO2_ELEVE	NHX_ELEVE	SOX_ELEVE
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0

Showing 1 to 10 of 50 entries

Previous 1 2 3 4 5 Next

Fig.7 : Binarisation des données collectées

Légende :

- ✓ ELEVE → dépassement du seuil normal [3] (niveau d'alerte)
- ✓ TRES_ELEVE → dépassement sévère (niveau critique)

Règles Apriori et metrics

Show 10 entries

ANTECEDENTS	CONFIDENCE	CONSEQUENTS	CONVICTION	LEVERAGE	LIFT	SUPPORT
co2_ELEVE,sox_ELEVE	1	nhx_ELEVE	Infinity	0.12	6.99	0.14
nhx_ELEVE	1	sox_ELEVE	Infinity	0.12	6.99	0.14
nhx_ELEVE	1	co2_ELEVE,sox_ELEVE	Infinity	0.12	6.99	0.14
nhx_ELEVE	1	co2_ELEVE	Infinity	0	1	0.14
nhx_ELEVE,co2_ELEVE	1	sox_ELEVE	Infinity	0.12	6.99	0.14
nhx_ELEVE,sox_ELEVE	1	co2_ELEVE	Infinity	0	1	0.14
sox_ELEVE	1	nhx_ELEVE	Infinity	0.12	6.99	0.14
sox_ELEVE	1	nhx_ELEVE,co2_ELEVE	Infinity	0.12	6.99	0.14
sox_ELEVE	1	co2_ELEVE	Infinity	0	1	0.14

Showing 1 to 9 of 9 entries

Previous 1 Next

Fig.8 : Règles Apriori et metrics

REMARQUES : Les **règles d'association fiables** sont générées à partir de ces motifs en filtrant celles ayant une forte confiance ($\geq 80\%$) et en les classant selon leur pertinence (lift). Enfin, les règles les plus significatives sont sélectionnées pour analyser les **relations entre polluants dominants et situations de dégradation de la qualité de l'air**.

2.5. Modélisation prédictive par Random Forest

Un modèle Random Forest Regressor a été entraîné à partir des données historiques afin de prédire les concentrations futures des polluants [1]. Une fenêtre temporelle de dix observations successives a été utilisée pour construire les variables d'entrée du modèle. Les prédictions portent sur les concentrations futures de PM2.5, PM10, CO₂, NO_x, SO_x et NH_x.

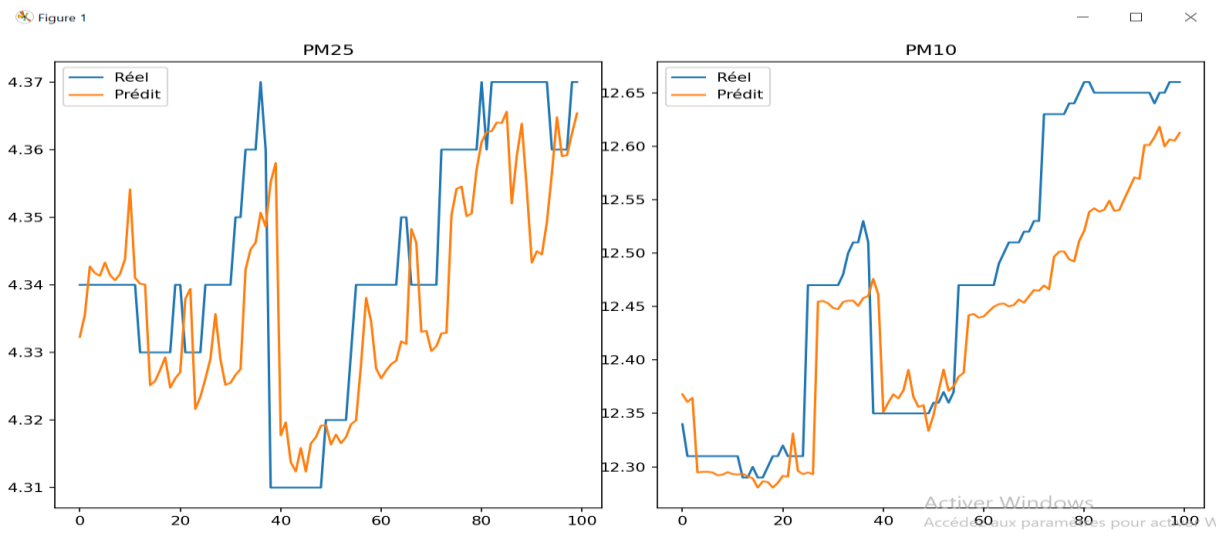


Fig.9 : Entra  nement du mod  le

3. Results

3.1. Surveillance en temps r  el

Le syst  me d  velopp   a permis la collecte, la transmission et le stockage en temps r  el des donn  es environnementales issues des capteurs install  s sur le site d'  tude. Les mesures sont consultables    travers un tableau de bord web interactif.



Fig.10 : Entra  nement du mod  le

REMARQUES :

Le syst  me met en place une API de surveillance et de pr  diction de la qualit   de l'air en temps r  el. Il utilise un mod  le pr  -entra  n   bas   sur une fen  tre de 10 observations pour pr  dire simultan  ment six polluants, tout en comparant les valeurs actuelles et futures. Une seconde interface fournit les donn  es r  centes pour la visualisation des tendances. L'ensemble int  gre des m  canismes de nettoyage et de validation des donn  es afin d'assurer la fiabilit   des r  sultats en production.

3.2. Analyse des associations entre polluants

L'application de l'algorithme Apriori a permis d'identifier plusieurs règles d'association entre les polluants atmosphériques. Ces résultats mettent en évidence certaines relations fréquentes entre les concentrations observées des différents contaminants.

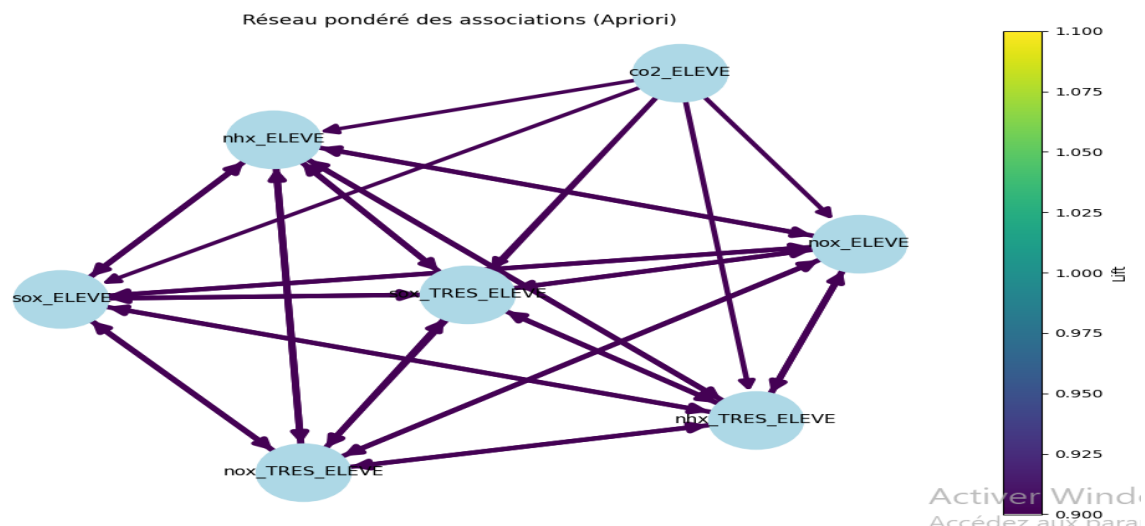


Fig.11: Réseau d'associations de polluants

REMARQUES :

L'image montre un **réseau de relations entre polluants et états de dégradation de l'air**. Les nœuds représentent les variables, et les liens illustrent les associations statistiques fortes. L'épaisseur et la couleur des connexions indiquent l'intensité et la fiabilité des relations, permettant d'identifier rapidement les polluants les plus impliqués dans la mauvaise qualité de l'air.

3.3. Performances du modèle prédictif

Il s'agit d'un modèle de prédiction multi-sorti qui utilise une fenêtre temporelle de 10 observations pour prédire l'évolution simultanée de six polluants. L'évaluation globale montre une performance modérée ($R^2 \approx 0,38$), indiquant que le modèle est partiellement capable d'expliquer la variabilité des données.

Les résultats varient considérablement selon les polluants : les concentrations de PM2.5 et PM10 sont très bien modélisées ($R^2 > 0,84$), indiquant une forte structure temporelle exploitable, tandis que le CO₂, les NO_x, les SO_x et les NH_x montrent une faible performance ($R^2 \approx 0,11-0,12$), suggérant des dynamiques plus complexes ou une moindre corrélation avec les variables historiques utilisées.

4. DISCUSSION

Les résultats obtenus confirment la pertinence des approches de fouille de données basées sur les règles d'association introduites par Agrawal et Srikant [2]. De plus, les performances obtenues avec le modèle Random Forest sont cohérentes avec les travaux de Breiman [1], qui soulignent l'efficacité de cette méthode pour les problèmes de prédiction complexes. L'utilisation de capteurs connectés et de plateformes cloud rejoint également les tendances actuelles des systèmes intelligents de surveillance environnementale [4][5][7].

5. CONCLUSION

Dans cette recherche, nous avons développé et déployé un système intelligent pour surveiller, analyser et anticiper la qualité de l'air au Grand Marché de Kinshasa. Cet appareil intègre des capteurs de l'Internet des objets, l'infrastructure de

stockage et de traitement des données en plus de l'algorithme Apriori et du modèle de forêt aléatoire pour fournir des données environnementales en temps réel.

Les résultats indiquent que le schéma adopté constitue une solution efficace pour la surveillance environnementale urbaine et soutient le processus de prise de décision. Pour les études futures, il serait une amélioration d'ajouter plus de paramètres météorologiques et même d'utiliser des algorithmes de prédiction plus sophistiqués afin d'améliorer l'efficacité du système.

REFERENCES

- [1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- [2] Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th International Conference on Very Large Data Bases*.
- [3] World Health Organization. *Air Quality Guidelines*. WHO, 2021.
- [4] FastAPI Documentation. <https://fastapi.tiangolo.com>
- [5] Espressif Systems. *ESP32 Technical Reference Manual*.
- [6] Nova Fitness. *SDS011 Laser Dust Sensor Specification*.
- [7] Supabase Documentation. <https://supabase.com/docs>
- [8] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. 3rd Edition, Morgan Kaufmann.
- [9] Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. 3rd Edition, Wiley.
- [10] Kumar, P., et al. (2018). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, 75, 199–205.
- [11] Gbegbelegbe, S., et al. (2020). Air quality monitoring challenges in Sub-Saharan Africa. *Atmospheric Environment*, 223, 117289.