Bulgarian Journal of Meteorology and Hydrology

Национален институт по метеорология и хидрология

National Institute of Meteorology and Hydrology

Bul. J. Meteo & Hydro 30/2 (2026)

Geospatial analysis of air pollution and health risk in Sofia using GIS and GeoAI tools of ESRI

Kaloyan Ivanov*

ESRI Bulgaria Ltd., 35 Nikola Vaptsarov Blvd., floor 4&5, Sofia 1407, Bulgaria

Abstract: Fine particulate matter (PM10) poses a well-documented health risk in urban environments, and its spatial distribution across cities is far from uniform. This study applies a geospatial analytical framework to PM10 concentrations recorded at 15 monitoring stations in Sofia, Bulgaria, during the winter season of 2018/2019 (synthetic dataset generated for educational purposes) — a period during which multiple Sofia stations registered exceedances of the annual average standard of 40 µg/m³. Three spatial methods were applied and compared: Inverse Distance Weighting (IDW), Empirical Bayesian Kriging (EBK), and Hot Spot Analysis (Getis-Ord Gi*). All analyses were performed in ArcGIS Pro, incorporating Corine Land Cover 2018 data as a thematic contextual layer. Cross-validation revealed that EBK substantially outperformed IDW, yielding a Root Mean Square Error (RMSE) of 5.8 µg/m³ versus 20.8 µg/m³ for IDW. Hot Spot Analysis identified statistically significant clusters of elevated PM10 at the 99% confidence level in peripheral areas of the city, while cold spots were located in central zones. In parallel, a conceptual framework for a GeoAI model based on a Multi-Layer Perceptron (MLP) neural network is presented, integrating spatial coordinates and Corine Land Cover categories as input variables for PM10 regression. Although the current sample size (n=15) limits full model training, the framework demonstrates the scalability of GeoAI methods for urban air quality prediction when broader datasets become available. An international reference case from the Marche Region, Italy, illustrates how similar GIS workflows combined with Augmented Reality (AR) can support real-time public air quality communication (Sanità et al., 2024). The findings confirm that EBK is the preferred method for small-network spatial interpolation of PM10, and that GeoAI approaches hold significant potential for future integration into operational air quality monitoring systems.

Keywords: PM10, spatial interpolation, GeoAI, ArcGIS Pro

* k.ivanov@esribulgaria.com

1. INTRODUCTION

Urban air quality is one of the most pressing environmental and public health challenges of the 21st century. Fine particulate matter with an aerodynamic diameter of 10 micrometers or less (PM10) has been extensively linked to respiratory and cardiovascular disease, premature mortality, and reduced quality of life (Goyal et al., 2021). The European Union has set a daily limit value of 50 µg/m³ (not to be exceeded more than 35 days per year) and an annual mean standard of 40 µg/m³ for PM10, and Bulgarian cities — particularly Sofia — have a documented history of frequent exceedances during the winter heating season.

Understanding the spatial distribution of PM10 across urban areas is essential for targeted policy interventions. Monitoring networks provide point-based measurements at discrete locations, but urban decision-making requires spatially continuous information: which districts are most affected, where hot spots persist, and how pollution gradients relate to land use and urban morphology. Geographic Information Systems (GIS) provide the methodological infrastructure to bridge this gap, enabling the integration of station measurements, land cover classifications, demographic data, and other spatial layers within a unified analytical environment (Janssen et al., 2008).

Classical geostatistical interpolation methods have been widely applied to air quality data. Inverse Distance Weighting is computationally simple and intuitive, assigning prediction values as distance-weighted averages of neighboring observations (Masroor et al., 2020). Kriging-based methods, particularly Empirical Bayesian Kriging, account for the spatial autocorrelation structure of the data and additionally provide estimates of prediction uncertainty — a feature of particular value when monitoring networks are sparse (Gribov & Krivoruchko, 2020). Spatial cluster detection via the Getis-Ord Gi* statistic complements interpolation by identifying zones of statistically persistent elevated or depressed concentrations, supporting evidence-based prioritization of intervention areas (Goyal et al., 2021; Esri, 2025b).

More recently, the convergence of geospatial analysis and Artificial Intelligence — collectively termed GeoAI — has opened new avenues for spatial prediction of air quality. Machine learning methods such as Multi-Layer Perceptron neural networks, Random Forest, and Gradient Boosting can capture nonlinear relationships between pollutant concentrations and a diverse set of predictor variables including land use, topography, meteorological conditions, and traffic density. While these approaches currently require larger training datasets than are typically available from conventional monitoring networks, their integration into operational GIS platforms such as ArcGIS Pro renders them increasingly accessible to environmental practitioners (Esri, 2025a).

This paper presents a geospatial analysis of PM10 concentrations in Sofia, Bulgaria, applying and comparing IDW, EBK, and Hot Spot Analysis within ArcGIS Pro, and proposes a conceptual GeoAI framework using a Multi-Layer Perceptron. The study also references an international case study from the Marche Region, Italy, as an example of how GIS-based air quality analysis can be extended to immersive Augmented Reality visualization for public engagement (Sanità et al., 2024; Renault et al., 2024).

2. DATA AND STUDY AREA

2.1. Study area

The study focuses on the municipality of Sofia, the capital and largest city of Bulgaria, situated in a topographic basin at approximately 550 m above sea level and surrounded by mountain ranges. This orographic setting promotes temperature inversions and limits the dispersion of pollutants during winter months, making the city particularly susceptible to elevated PM10 concentrations from domestic heating, road traffic, and industrial emissions.

2.2. PM10 monitoring data

Concentration data for PM10 were obtained from 15 air quality monitoring stations operating within the Sofia urban area during the winter season of 2018/2019. Daily average concentrations were used for the analysis. The 2018/2019 winter period was characterized by elevated PM10 levels, consistent with the broader pattern of recurring exceedances in Sofia: multiple automatic monitoring stations recorded exceedances of the annual average standard of 40 $\mu g/m^3$ during that reference period. The winter heating season consistently produces the highest PM10 concentrations in Sofia, driven primarily by solid-fuel combustion in areas outside the district heating network.

NOTE: The dataset used in this study was generated for educational and demonstrational purposes and does not represent actual measurements from the operational Bulgarian air quality monitoring network. The spatial distribution and concentration values of the 15 stations are based on realistic assumptions derived from published literature but should not be used for regulatory or policy purposes.

2.3. Ancillary spatial data

Corine Land Cover 2018 (CLC 2018) was used to characterize land use across the study area. CLC categories relevant to the urban context include continuous and discontinuous urban fabric, industrial and commercial units, road and rail networks, green urban areas, and agricultural land. Land cover data serve both as a contextual reference for interpreting interpolation results and as a categorical input variable for the conceptual GeoAI model (Janssen et al., 2008). Reference layers in ArcGIS Pro — including imagery basemaps, administrative boundaries, road network layers, and demographic data — were used to support spatial referencing and contextual interpretation of results.

3. METHODS

3.1. Inverse Distance Weighting (IDW)

IDW is a deterministic spatial interpolation method in which the predicted value at any unsampled location is computed as a weighted average of surrounding observations, with weights proportional to the inverse of the distance to each observation point (Masroor et al., 2020). The method is computationally efficient and produces intuitively interpretable results, but it does not account for the spatial autocorrelation structure of the data, nor does it provide prediction uncertainty estimates. IDW was applied in ArcGIS Pro using the default power parameter of 2.

3.2. Empirical Bayesian Kriging (EBK)

EBK is a geostatistical interpolation method that automates the estimation of the semivariogram by generating a large number of semivariogram realizations through subsetting and simulation, and combining them within a Bayesian framework (Gribov & Krivoruchko, 2020; Esri, 2025a). This approach addresses a fundamental shortcoming of classical kriging: rather than fitting a single semivariogram from the observed data and treating it as the true model, EBK explicitly accounts for the uncertainty in semivariogram estimation, thereby producing more accurate prediction standard errors (Gribov & Krivoruchko, 2020). EBK has been shown to be particularly advantageous for small datasets, where classical semivariogram estimation is inherently unstable. In the present study, EBK was applied in ArcGIS Pro using the Geostatistical Wizard, with automatic parameter selection.

3.3. Hot Spot analysis (Getis-Ord Gi*)

Hot Spot Analysis was performed using the Getis-Ord Gi* statistic as implemented in ArcGIS Pro (Esri, 2025b). This method evaluates, for each observation point, whether the local concentration — together with the values of its spatial neighbours — is significantly higher or lower than the global mean. The output is a set of z-scores and associated p-values that indicate the statistical significance of spatial clustering. Zones classified as hot spots (high positive z-scores) represent areas of statistically persistent elevated PM10, while cold spots (high negative z-scores) indicate zones of consistently low concentrations (Goyal et al., 2021). Results are reported at the 99%, 95%, and 90% confidence levels.

3.4. Cross-validation and performance metrics

To assess and compare the accuracy of IDW and EBK, leave-one-out cross-validation was performed: each station was successively withheld, its PM10 value was predicted using the remaining 14 stations, and the prediction was compared to the observed value. Performance was evaluated using the Root Mean Square Error (RMSE), expressed in $\mu g/m^3$. A lower RMSE indicates closer agreement between predicted and observed values. The coefficient of determination R^2 was also computed but, given the small sample size ($n=15$), is interpreted with caution as it is sensitive to the number of observations (Masroor et al., 2020).

3.5. Conceptual GeoAI framework: Multi-Layer Perceptron

As a methodological perspective, a conceptual framework for spatial PM10 prediction using a Multi-Layer Perceptron (MLP) neural network was developed. An MLP is a feedforward artificial neural network architecture consisting of an input layer, one or more hidden layers with nonlinear activation functions (e.g., ReLU or tanh), and an output layer for regression. Unlike linear regression or classical geostatistical methods, an MLP can approximate arbitrarily complex nonlinear functions of the input variables, making it suitable for capturing nonlinear relationships between PM10 concentrations and land use, topographic, and meteorological predictors.

In the proposed framework, the input variables comprise the geographic coordinates (X, Y) of each monitoring location and one-hot encoded Corine Land Cover categories. Future extensions of the model would incorporate meteorological variables (wind speed and direction, temperature, relative humidity), elevation from a Digital Elevation Model, road network density, and traffic intensity as additional predictors. The architecture includes two or three hidden layers, with the output layer producing a single continuous PM10 concentration value.

A critical limitation of the present study is the sample size of n=15 monitoring stations. Machine learning models of the MLP type require substantially larger training datasets — typically a minimum of 50–100 observations — to achieve stable generalization and avoid overfitting. Consequently, the MLP framework is presented here as a scalable conceptual model for future application, rather than as a trained and validated predictive tool. GeoAI tools for spatial regression are available within the ArcGIS Pro environment, including Forest-based Classification and Regression (Random Forest and Gradient Boosting) and deep learning modules with Python integration (scikit-learn, TensorFlow, PyTorch) (Esri, 2025a).

4. RESULTS

4.1. IDW interpolation

The IDW surface reveals a spatially smooth PM10 gradient across Sofia, with higher predicted concentrations in the peripheral areas around the highest-value stations and lower values towards the remaining parts of the municipality. Several zones, particularly in peripheral districts, display predicted daily average concentrations exceeding the EU daily limit value of 50 μg/m³ (Figure 1). The interpolated surface reflects the distance-weighted influence of nearby high-concentration stations, but does not account for directional patterns or land cover heterogeneity. Cross-validation of the IDW model yielded an RMSE of 20.8 μg/m³.

4.2. EBK interpolation and uncertainty

The EBK prediction surface displays a spatially smoother pattern than IDW, with two pronounced concentration maxima: one in the north-western and one in the south-eastern parts of the municipality (Figure 2).

The model additionally generates a prediction standard error surface, which identifies zones of higher interpolation uncertainty — areas where predictions should be interpreted with greater caution, typically located between monitoring stations in areas with sparse spatial coverage (Gribov & Krivoruchko, 2020). Cross-validation of the EBK model yielded an RMSE of 5.8 µg/m³, representing a substantial improvement relative to IDW. The comparison of cross-validation metrics is summarized in Table 1.

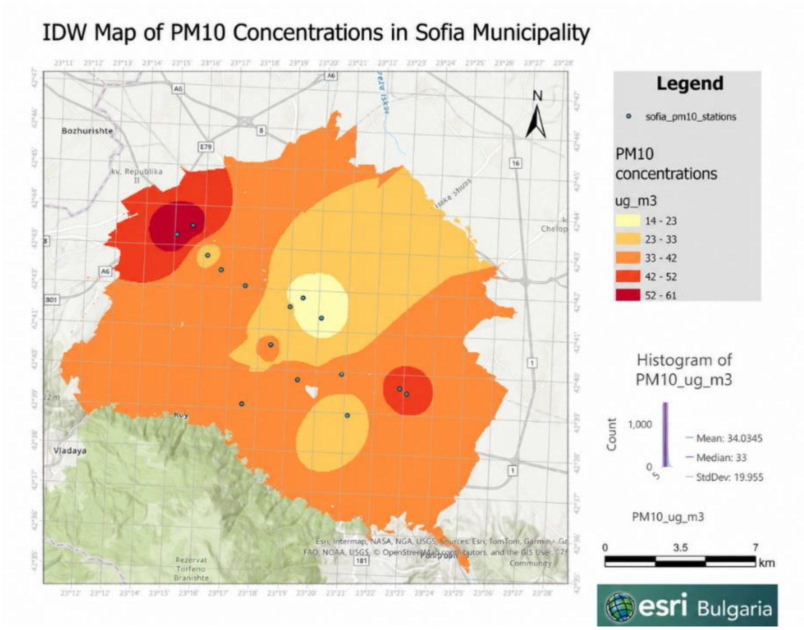

Fig. 1. IDW map of PM10 concentrations in Sofia Municipality

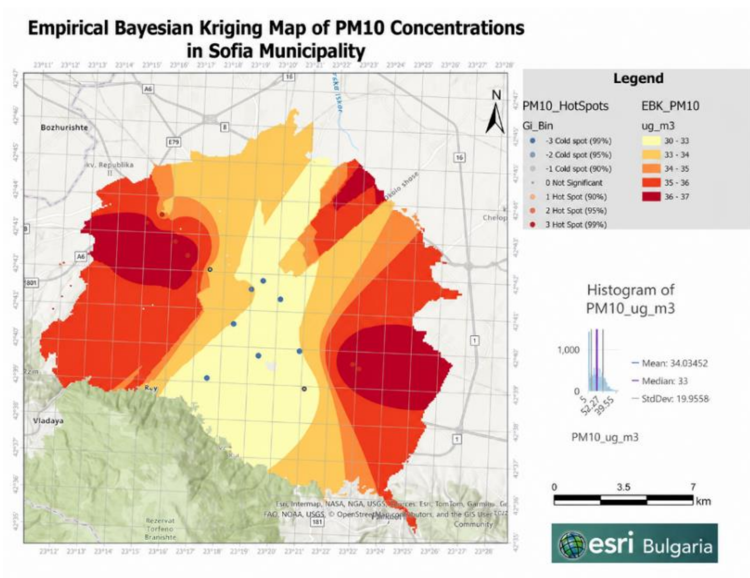

Fig. 2. EBK map of PM10 concentrations in Sofia Municipality

Table 1. Cross-validation performance metrics for IDW and EBK interpolation of PM10 concentrations, Sofia, winter 2018/2019.

Method	RMSE (μg/m³)	R^2	Notes
IDW	20.8	< 0.1	Distance-based weighting; no autocorrelation modelling
EBK	5.8	< 0.1	Bayesian semivariogram simulation; uncertainty surface generated

The R^2 values for both methods were below 0.1, reflecting the limited number of monitoring stations (n=15). At this sample size, leave-one-out cross-validation R^2 is an unreliable model performance indicator; RMSE and the reliability of the prediction confidence intervals are the preferred evaluation criteria (Gribov & Krivoruchko, 2020).

4.3. Hot Spot analysis

The Getis-Ord Gi* analysis identified statistically significant hot spots — zones of persistently elevated PM10 concentrations — in the peripheral areas of Sofia municipality, at the 99% confidence level. Cold spots, indicating areas of statistically consistently low concentrations, were identified in zones located closer to the urban core and green areas (Figure 3).

Areas without statistically significant spatial clustering (neutral z-scores) suggest more spatially variable or weakly autocorrelated PM10 distributions (Goyal et al., 2021; Esri, 2025b). This pattern is consistent with the known spatial structure of winter PM10 pollution in Sofia, driven by domestic solid-fuel heating concentrated in outer residential and semi-rural areas.

5. DISCUSSION

5.1. Comparative assessment of interpolation methods

The substantially lower cross-validation RMSE of EBK (5.8 μg/m³) compared to IDW (20.8 μg/m³) confirms previous findings from comparable studies in which kriging-based methods consistently outperform deterministic distance-weighting approaches for PM10 and PM2.5 spatial prediction (Masroor et al., 2020; Janssen et al., 2008). The advantage of EBK is attributable to its explicit modelling of spatial autocorrelation and its treatment of semivariogram uncertainty, which is particularly beneficial when the number of observations is small (Gribov & Krivoruchko, 2020). At n=15, classical kriging would rely on an unstable semivariogram estimate, whereas EBK's Bayesian simulation framework provides more robust predictions.

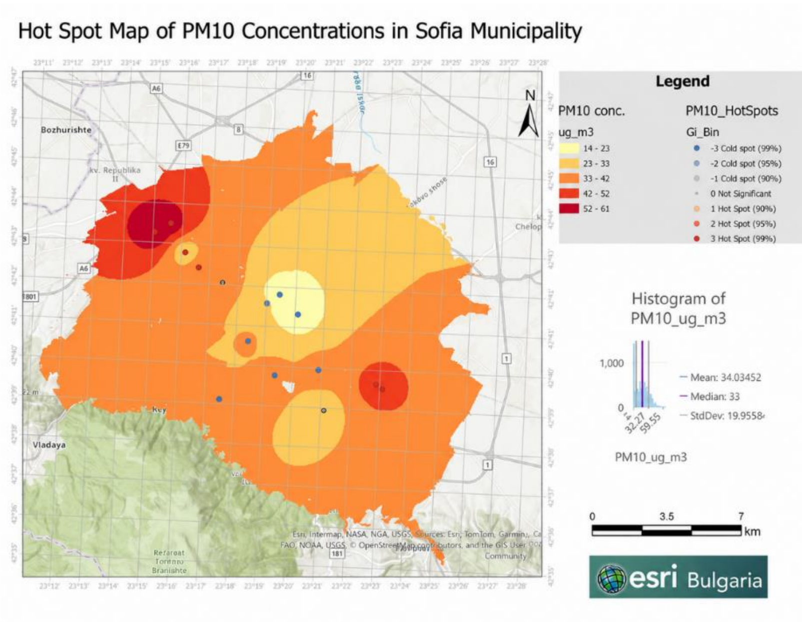

Fig. 3. Hot Spot map of PM10 concentrations in Sofia Municipality

The two distinct concentration maxima identified by EBK — in the north-western and south-eastern parts of the municipality — are spatially consistent with known emission source areas: the north-western area is associated with industrial activity and high-density residential heating zones, while the south-eastern zone corresponds to suburban residential areas with a high prevalence of solid-fuel domestic heating. These patterns are corroborated by the Hot Spot Analysis, which confirms the statistical persistence of elevated PM10 in peripheral areas (Goyal et al., 2021).

The interpretation of the R^2 values below 0.1 for both methods requires methodological clarification. In leave-one-out cross-validation with n=15, the R^2 statistic is highly sensitive to outlier observations and does not reliably reflect model performance (Masroor et al., 2020; Gribov & Krivoruchko, 2020). The RMSE and the proportion of observations falling within the 95% prediction interval of EBK are more informative criteria under these conditions.

5.2. Hot spots and land use context

The peripheral concentration of PM10 hot spots identified by the Getis-Ord Gi* analysis reflects the spatial structure of Sofia's heating infrastructure. Districts further from the district heating network rely predominantly on individual solid-fuel boilers and stoves, which are the dominant source of PM10 during the winter season in Bulgarian cities.

The incorporation of Corine Land Cover data provides a spatial framework for contextualizing these patterns: hot spot zones correspond predominantly to low-density residential and peri-urban land cover categories, consistent with the dispersed heating emission sources (Janssen et al., 2008).

5.3. GeoAI perspective and scalability

The proposed MLP framework represents a methodologically sound approach to extending PM10 spatial prediction beyond the capabilities of classical geostatistical methods. By incorporating land cover categories, topographic variables, and — in future iterations — meteorological inputs and traffic data, an MLP model can capture nonlinear interaction effects that IDW and kriging cannot represent. Studies applying machine learning to urban air quality prediction have demonstrated that neural network and ensemble models substantially outperform classical interpolation when training datasets include sufficient observation points (Esri, 2025a).

The current constraint of $n=15$ stations is a limiting factor that reflects the density of the Bulgarian national air quality monitoring network. Strategies to address this limitation include: integrating satellite-derived PM10 products (e.g., from Sentinel-5P TROPOMI) to create a fused ground-satellite dataset; incorporating data from low-cost sensor networks as supplementary training points; and applying transfer learning from national-scale models trained on denser networks in other European countries.

5.4. International reference: AR visualisation in the Marche region

A directly relevant international application demonstrates the potential for integrating the GIS-based analytical workflow described in this study with immersive visualization technologies. Sanità et al. (2024) report a case study from the Marche Region (Ancona, Italy) in which data from 17 ARPAM ground-based monitoring stations were processed in ArcGIS Pro — including spatial interpolation and Air Quality Index (AQI) calculation — and the results were visualized via a WebGIS platform and an Augmented Reality (AR) application. The AR layer overlaid five AQI levels (excellent, good, fair, poor, bad) onto a 3D urban model derived from ArcGIS Online, enabling citizens to perceive spatially differentiated air quality in an immersive urban context. A related study by Renault et al. (2024) further demonstrates the potential of AR for dynamic public exposure visualization of air quality data. These approaches directly parallel the methodological framework applied in the present study and illustrate the operational pathway from the spatial analysis performed here towards real-time public communication tools.

6. CONCLUSIONS

This study demonstrates the applicability of GIS-based spatial analysis for characterizing the distribution of PM10 concentrations in the urban environment of

Sofia, Bulgaria, using a winter 2018/2019 monitoring dataset from 15 stations. The principal findings are as follows:

- EBK substantially outperforms IDW in cross-validation accuracy (RMSE 5.8 vs. 20.8 μg/m³) and is the recommended interpolation method for PM10 spatial prediction under sparse monitoring network conditions, consistent with the broader geostatistical literature (Gribov & Krivoruchko, 2020).
- Hot Spot Analysis (Getis-Ord Gi*) reveals statistically significant clusters of elevated PM10 in the peripheral areas of Sofia municipality, at the 99% confidence level, corresponding to zones dependent on solid-fuel domestic heating (Goyal et al., 2021; Esri, 2025b).
- The concentration patterns observed during the 2018/2019 winter season reflect the broader historical context in which Sofia recorded some of its highest PM10 levels; subsequent years have shown a positive trend in compliance with EU standards.
- A conceptual GeoAI framework based on a Multi-Layer Perceptron demonstrates the feasibility of nonlinear spatial prediction integrating land cover and coordinate variables, but requires a minimum of 50–100 monitoring points for reliable model training.
- The Marche Region (Italy) case study illustrates how the analytical workflow presented here can be extended towards Augmented Reality visualization for citizen engagement and public health communication (Sanità et al., 2024).

Future research directions include: expanding the monitoring network or integrating remote sensing data (e.g., Sentinel-5P) to enable operational GeoAI model training; incorporating meteorological, topographic, and traffic density variables as predictors; developing a temporal prediction component for short-term PM10 forecasting; and connecting the spatial analysis to a real-time dashboard with automated exceedance alerts for stakeholders and the public.

REFERENCES

Esri (2025a). What is empirical Bayesian kriging? *ArcGIS Enterprise Documentation*. Available at: https://enterprise.arcgis.com/en/portal/latest/use/what-is-empirical-bayesian-kriging-.htm (Accessed: May 2026).

Esri (2025b). Hot Spot Analysis (Getis-Ord Gi*). *ArcGIS Pro Tool Reference*. Available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/hot-spot-analysis.htm (Accessed: May 2026).

Goyal, P., Gulia, S., Goyal, S.K. (2021). Identification of air pollution hotspots in urban areas. *Science of the Total Environment*, 798, 149143. https://doi.org/10.1016/j.scitotenv.2021.149143

Gribov, A., Krivoruchko, K. (2020). Empirical Bayesian kriging implementation and usage. *Science of the Total Environment*, 722, 137290. https://doi.org/10.1016/j.scitotenv.2020.137290

Janssen, S., Dumont, G., Fierens, F., Mensink, C. (2008). Spatial interpolation of air pollution measurements using CORINE land cover data. *Atmospheric Environment*, 42(20), 4884–4903. https://doi.org/10.1016/j.atmosenv.2008.02.043

Masroor, K., Fanaei, F., Yousefi, S., Raeesi, M., Abbaslou, H., Shahsavani, A. (2020). Spatial modelling of PM2.5 concentrations in Tehran using Kriging and inverse distance weighting (IDW) methods. *Journal of Air Pollution and Health*, 5(2), 89–96. https://doi.org/10.18502/japh.v5i2.4237

Renault, S., Feldmann, I., Raes, L., Silence, J., Schreer, O. (2024). Dynamic Exposure Visualization of Air Quality Data with Augmented Reality. In *Proceedings of the 10th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM)*, 120–127. https://doi.org/10.5220/0012586700003696

Sanità, M., Fratini, J., Muralikrishna, N., Pierdicca, R., Malinverni, E.S. (2024). Augmented Reality for Air Quality Monitoring: Case Study in the Marche Region (Italy). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4-2024, 389–395. https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-389-2024.