Geostatistical and Machine Learning Approaches for Air Pollution Interpolation

Aminata SALL

African Institute for Mathematical Sciences, AIMS-Senegal

Supervised by Monica PIRANI
from Imperial College,UK

June 11, 2025

Definition of Air Pollution

Air pollution refers to the presence of harmful or excessive quantities of substances in the air we breathe. These substances, known as pollutants, can be gases, particulates, or biological molecules that pose risks to human health and the environment.
Air pollution can be both visible and invisible, and its sources are diverse, including natural events and human activities such as industrial emissions, vehicle exhaust, and burning of fossil fuels.

Source: World Health Organization, 2021

fig1:Air pollution

Main Sources of Air Pollution

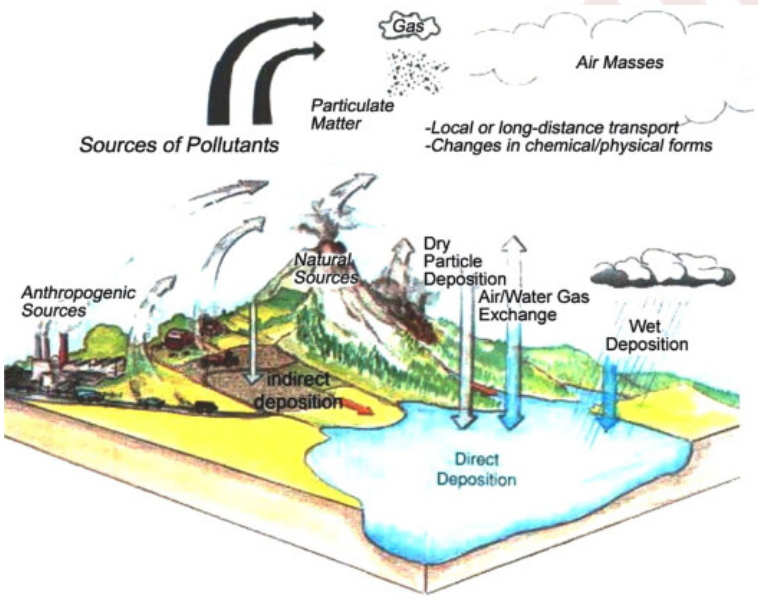


fig2:Anthropogenic and natural sources of air pollution.

Air pollution originates from both **anthropogenic** and **natural** sources:

Anthropogenic Sources:

- **Transport:** CO, NO_x, PM, VOCs from vehicles.
- **Industry:** SO_2, NO_x from fossil fuels.
- **Agriculture:** CH_4 (livestock), N_2O (fertilizers).

Natural Sources:

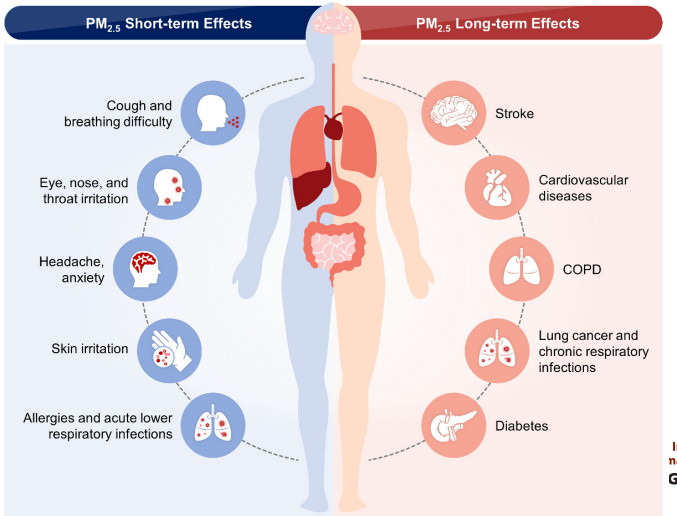
- **Volcanoes:** SO_2, ash, aerosols.
- **Forest fires:** CO, CO_2, $PM_{2.5}$, PM_{10}.
- **Dust and pollen:** Natural PM_{10}.

AIMS African Institute for Mathematical Sciences SENEGAL

Motivation and Rationale

- **Air pollution is a major urban challenge.**
- $PM_{2.5}$ is a key pollutant with serious health impacts (WHO, 2021).
- Urban areas concentrate sources such as traffic and industry.
- Accurate exposure mapping is crucial for public health strategies.
- Monitoring networks are often sparse or unevenly distributed.
- Spatial interpolation allows estimation at unmonitored locations.

fig3:traffic and industry pollution



Institute for natical Sciences
GAL

Research Question and Objective

Research Question:
How to best interpolate high-resolution $PM_{2.5}$ concentrations in urban areas with limited monitoring?

AIMS African Institute for Mathematical Sciences SENEGAL

Specific Objectives

- Acquire and preprocess $PM_{2.5}$ data from PurpleAir sensors.
- Implement and analyze IDW and ordinary kriging.
- Develop and apply Random Forest Spatial Interpolation.
- Evaluate model accuracy via cross-validation metrics.
- Visualize spatial predictions and uncertainties.
- Critically discuss strengths and limitations of each method.

AIMS African Institute for Mathematical Sciences SENEGAL

Definition of Geostatistical Data

Geostatistical data consist of measurements of a spatially continuous phenomenon collected at fixed spatial locations.

They are modeled as realizations of a spatial stochastic process:

$$\{Z(s) : s \in D \subset \mathbb{R}^2\}$$

where $Z(s)$ denotes the value of the process at location s in domain D.

These data assume smooth spatial variation, with observations taken at irregularly spaced sites.

Objectives of Geostatistical Analysis

The main goal is to model the spatial structure and predict values at unobserved locations.

In this work, three interpolation methods are considered:

- Inverse Distance Weighting (IDW)
- Ordinary Kriging
- Random Forest Spatial Interpolation (RFSI)

Inverse Distance Weighting (IDW)

Prediction at location s_0:

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{n} w_i Z(s_i)}{\sum_{i=1}^{n} w_i} \quad \text{where} \quad w_i = \frac{1}{d(s_0, s_i)^p} \tag{6}$$

Where:

- $d(s_0, s_i)$: distance between prediction and observed locations.
- p: power parameter controlling decay (commonly $p = 2$).
- Nearby points exert more influence.
- Easy to implement, deterministic.
- Does not quantify prediction uncertainty.

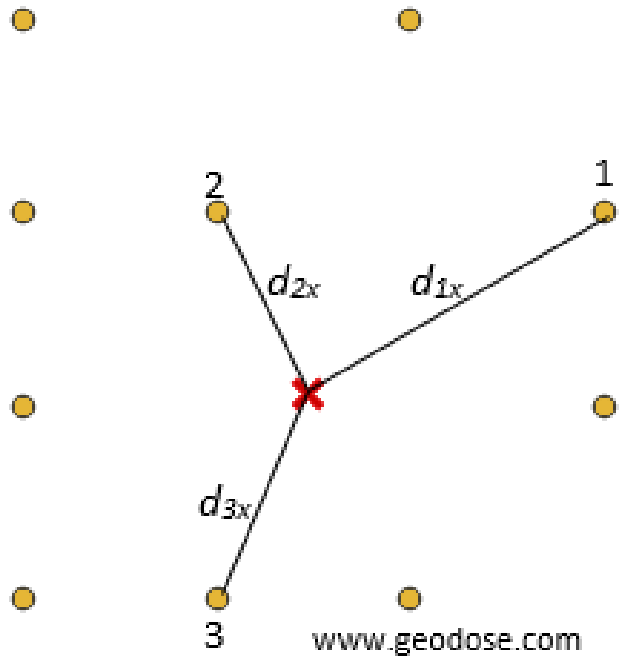


fig5:Illustration of IDW interpolation

Why the Variogram?

- Kriging relies on spatial autocorrelation.
- The variogram measures how similarity decreases with distance.
- It helps model spatial structure before interpolation.
- Key step: fit a theoretical model to the data.

AIMS African Institute for Mathematical Sciences SENEGAL

Empirical Variogram

$$\gamma(h) = \frac{1}{2}\mathrm{Var}\left[Z(s) - Z(s+h)\right] \quad (7)$$

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{(i,j)\in N(h)}\left(Z(s_i) - Z(s_j)\right)^2 \quad (8)$$

Where:

- h: spatial lag distance.
- $N(h)$: set of all pairs separated by distance h.
- Measures spatial structure and dependence between observations.

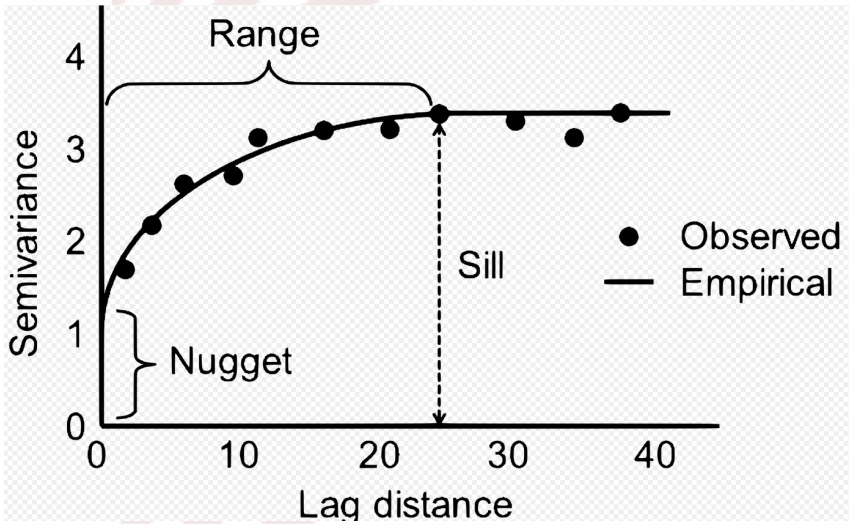


fig6:Example of emperical variogram

Common Variogram Models

Model forms:

- *Spherical:* rises linearly then plateaus at range a.
- *Exponential:* smooth increase approaching sill asymptotically.
- *Gaussian:* smooth, with parabolic initial behavior.
- *Matérn:* flexible, includes smoothness parameter ν.

Ordinary Kriging: Prediction

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \quad \text{with} \quad \sum_{i=1}^{n} \lambda_i = 1 \tag{9}$$

- Weights λ_i derived to minimize prediction variance.
- Captures spatial autocorrelation modeled via variogram.
- Provides best linear unbiased prediction (BLUP).

Ordinary Kriging: Prediction Variance

$$\text{Var}(\hat{Z}(s_0)) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) \tag{10}$$

- $\gamma(h)$ is the variogram function expressing spatial dependence.
- Allows quantification of uncertainty in predictions.

AIMS African Institute for Mathematical Sciences SENEGAL

Random Forest Regression

$$\hat{\theta}_B(x) = \frac{1}{B}\sum_{b=1}^{B} t_b^*(x) \qquad (11)$$

Where:

- x is the input feature vector (location or covariates).
- $t_b^*(x)$ is the prediction from the b^{th} tree trained on a bootstrap sample.
- B is the total number of trees.
- $\hat{\theta}_B(x)$ is the aggregated prediction.

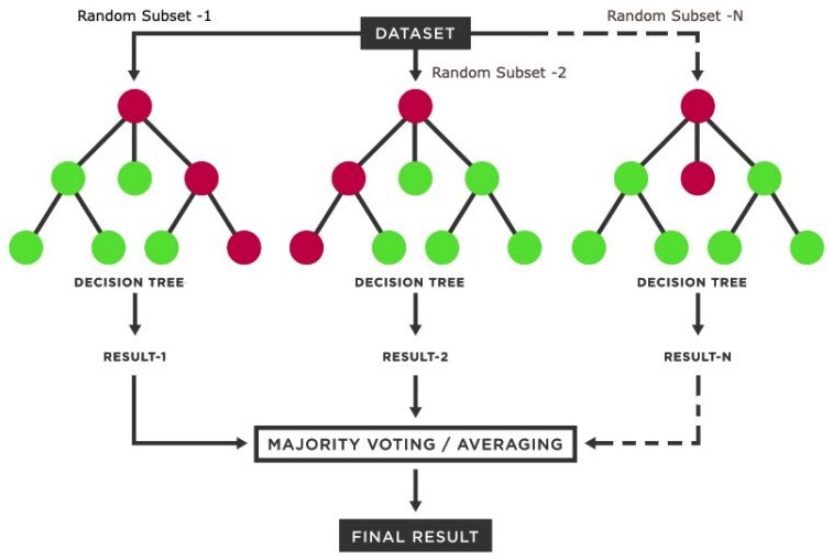


fig7:Illustration of Random Forest ensemble learning

Random Forest Spatial Interpolation (RFSI)

$$Y(s) = f\left(v_{p_1}, d_{p_1}, \ldots, v_{p_K}, d_{p_K}\right) \tag{12}$$

- v_{p_k}: PM2.5 values at k nearest neighbors.
- d_{p_k}: distances to these neighbors.
- f: function learned by random forest capturing spatial structure.

K-Fold Cross-Validation

- Split dataset into K folds.
- Iteratively train on $K-1$ folds, test on the remaining.
- Aggregate prediction errors over all folds.
- Reduces overfitting and estimates generalization accuracy.

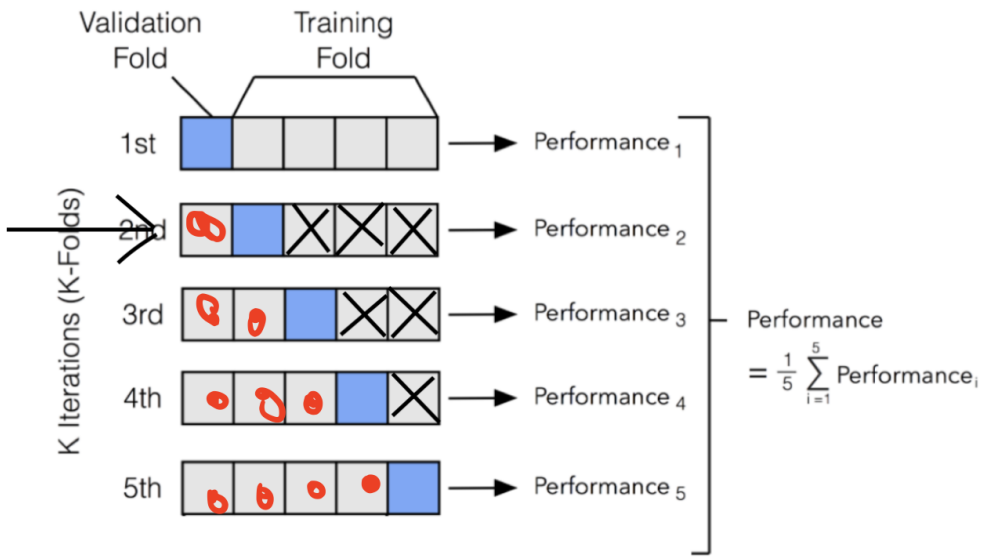


fig8:Cross-Validation Process: K-folds

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

- Measures average magnitude of errors.
- Sensitive to large deviations.

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{16}$$

- Measures average absolute deviations.
- More robust to outliers than RMSE.

R^2 - Coefficient of Determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

- Proportion of variance explained by the model.
- Ranges from 0 (no explanatory power) to 1 (perfect fit).

Study Area: San Francisco

- Coastal city in Northern California, USA.
- Located on a narrow peninsula between the Pacific Ocean and San Francisco Bay.
- Neighbored by urban and industrial areas such as Oakland (east) and Daly City (south).
- Mediterranean climate: cool, wet winters and dry summers.
- Frequent marine fog due to cold Pacific currents.
- Pollution influenced by road traffic, port activity, and nearby industrial zones.
- Elevated $PM_{2.5}$ concentrations due to combined natural and anthropogenic factors.

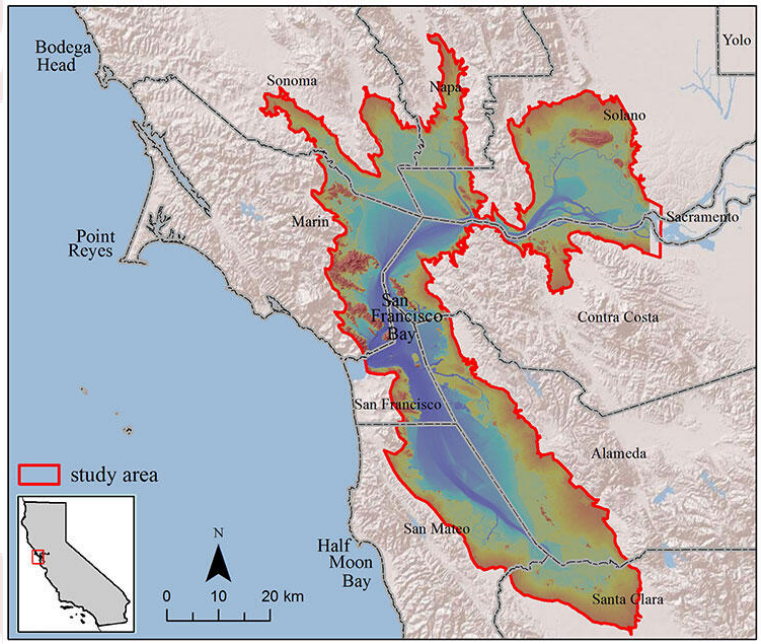


fig9: Study area : San Francisco

AIMS African Institute for Mathematical Sciences SENEGAL

Data Description

- 44 monitoring sites distributed unevenly across the San Francisco.
- The dataset covers the period from **January 1, 2020 to May 31, 2020**.
- These sensors provide **high-resolution** spatial and temporal measurements of PM concentrations.
- Original 4-hour resolution data were **aggregated to daily means** to reduce short-term variability.
- Most sensors are located in the **southern and eastern** parts of the city.

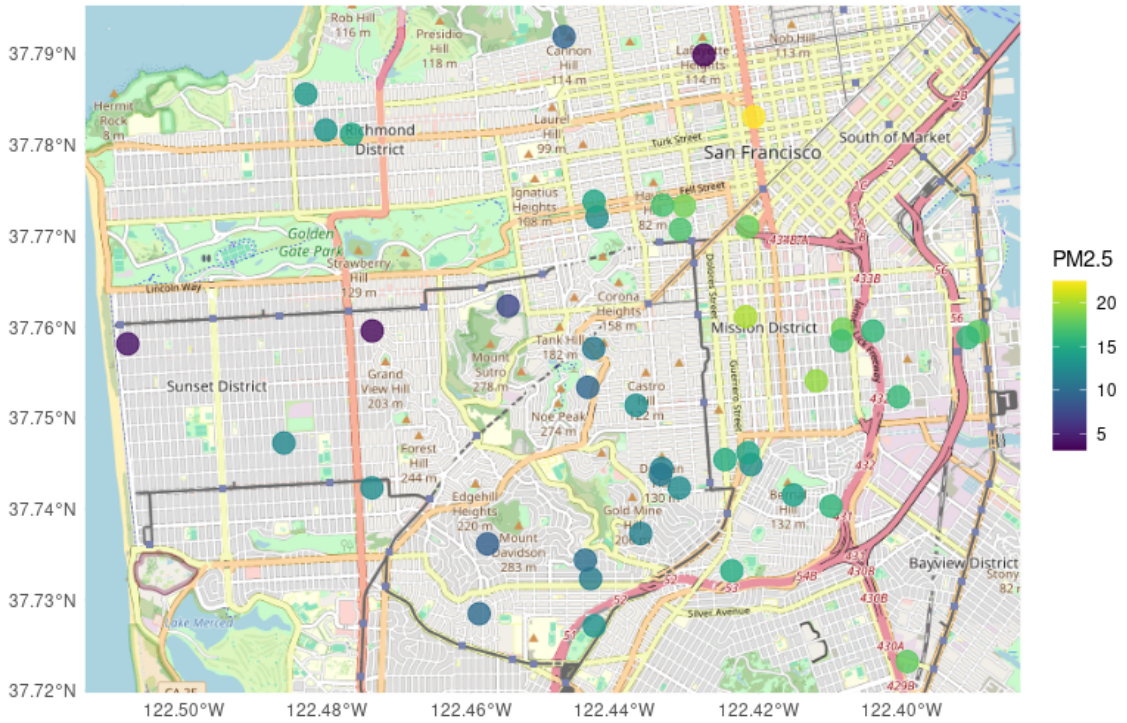


fig10:Location of PurpleAir sensors in San Francisco .

Spatial Distribution of Sensors

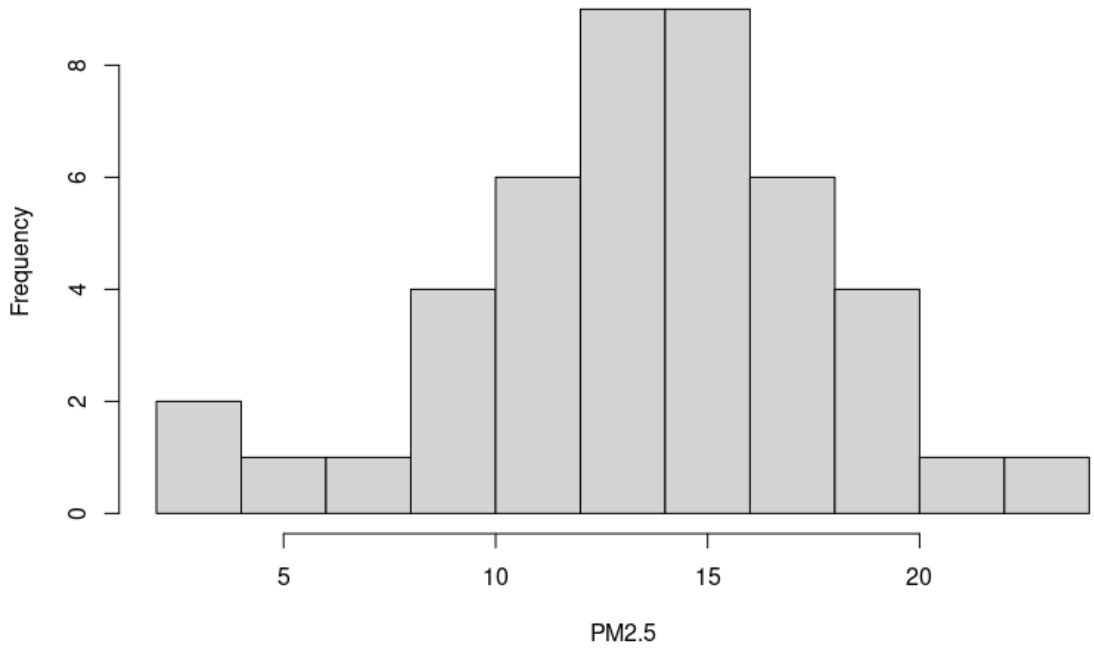


fig11:Histogram showing the number of sensors per neighborhood.

Temporal Distribution of $PM_{2.5}$

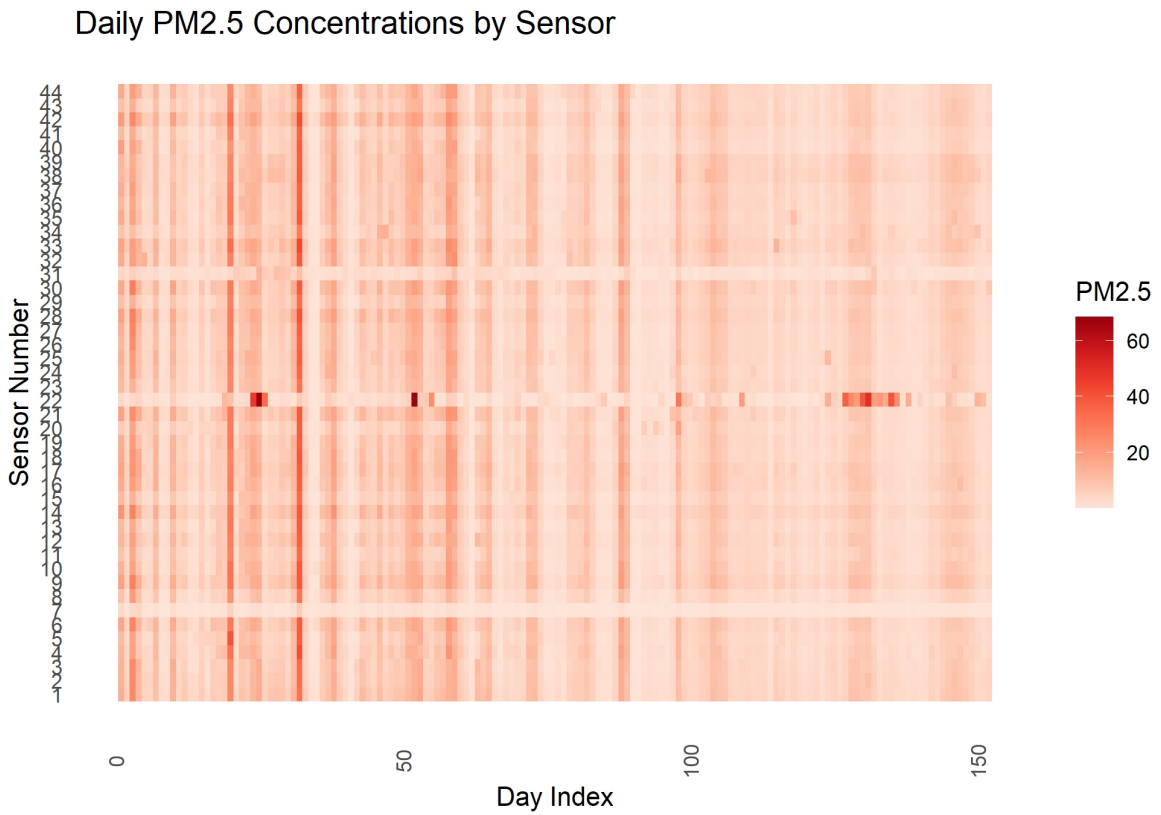


fig12:Heatmap illustrating daily $PM_{2.5}$ levels over time across all sensors.

Spatial Variogram Characteristics

Among tested models (spherical, exponential, Matérn), the Gaussian model showed the best fit with the following parameters:

- **Nugget:** 0.91 – capturing measurement error or microscale variation.
- **Partial sill:** 17.0 – the structured spatial variance.
- **Range:** 1860 m – spatial correlation fades beyond this distance.

This implies $PM_{2.5}$ values are spatially correlated up to 1.86 km.

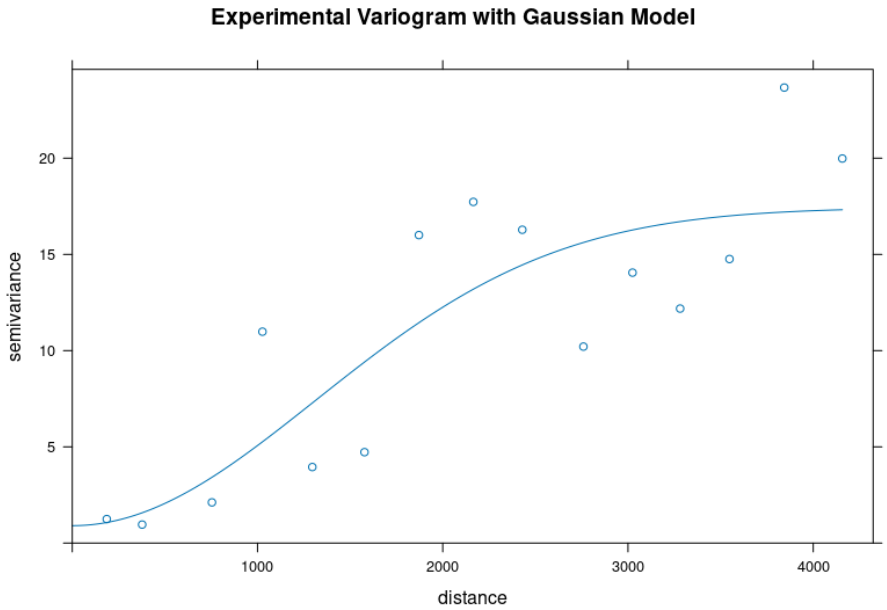
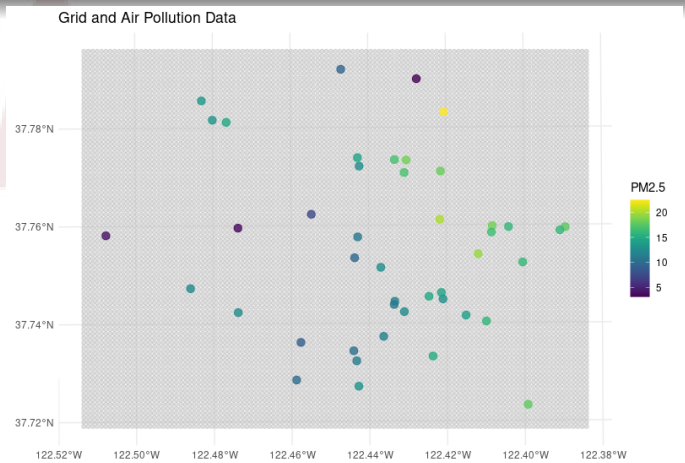
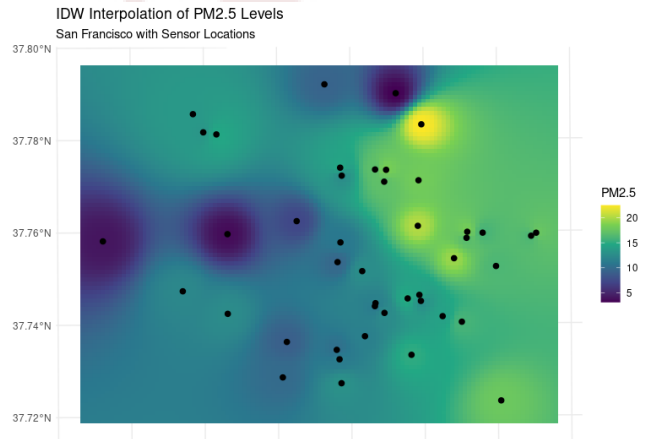


fig13:Empirical variogram with fitted Gaussian model.

Spatial Prediction: IDW

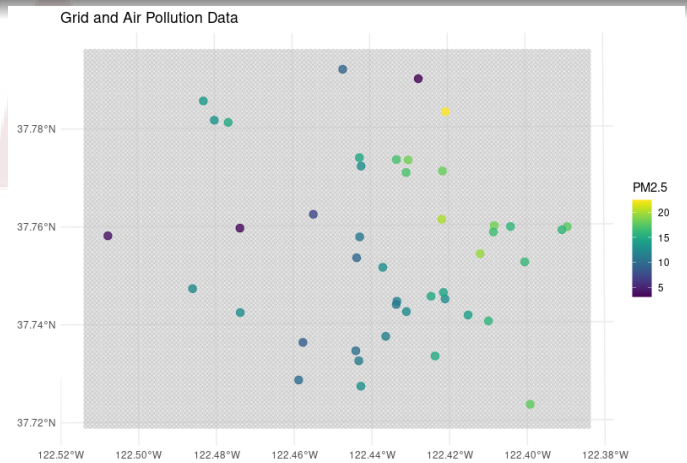
- IDW estimates $PM_{2.5}$ using distance-based weights on a 500 m grid.
- Produces abrupt transitions near sensors, especially downtown.
- Highest values ($\sim$23 g/m^3) in the industrial southeast (Bayview).
- Lower values ($\sim$7 g/m^3) near the western coast (Ocean Beach).
- Limited capacity to model gradual spatial trends.

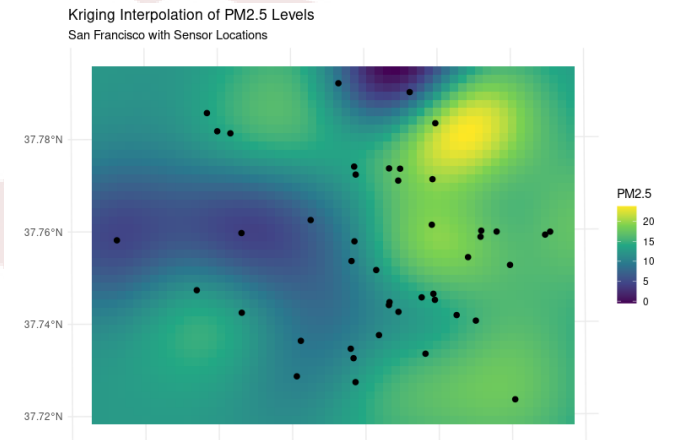




Spatial Prediction: Ordinary Kriging

- Kriging accounts for spatial autocorrelation via a Gaussian variogram (range $\sim$1.8 km).
- Smoother gradients and more realistic spatial patterns.
- Hotspots ($\sim$25 g/m^3) in the southeast, low values in the northwest (6–8 g/m^3).
- Highlights urban air quality disparities relative to WHO and EPA standards.

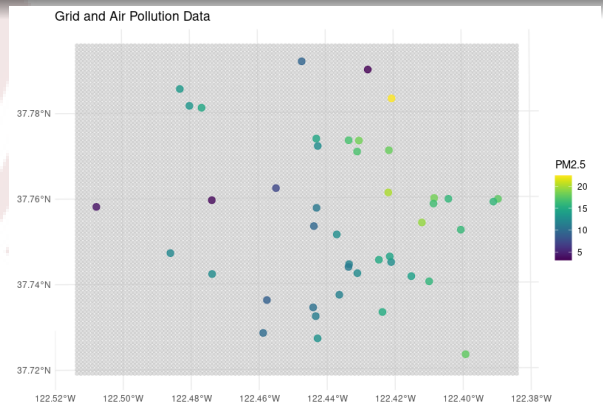


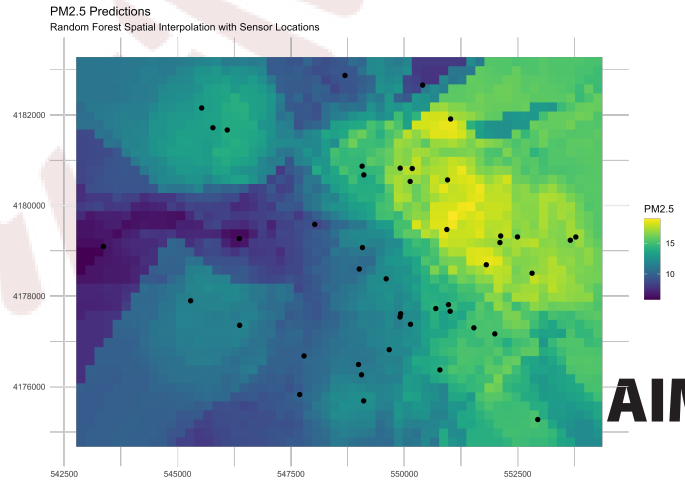


African Institute for Mathematical Sciences SENEGAL

Spatial Prediction: Random Forest (RFSI)

- RFSI uses 10 nearest neighbors as inputs to a 500-tree RF model.
- Captures complex, non-linear spatial patterns and local hotspots.
- Detects elevated values in southeast and Mission District.
- Slight overestimation in dense areas; no covariates used yet.
- Potentially improved by adding external data (e.g., land use, emissions).





AIMS African Institute for Mathematical Sciences SENEGAL

Cross-Validation Results

Method	**RMSE** ($\mu g/m^3$)	**MAE** ($\mu g/m^3$)	R^2
IDW	3.57	2.30	0.28
OK	3.97	2.28	0.31
RFSI	2.43	1.59	0.65

Table:Performance metrics from K-fold cross-validation for each interpolation method.

AIMS African Institute for Mathematical Sciences SENEGAL

Observed vs. Predicted: RFSI

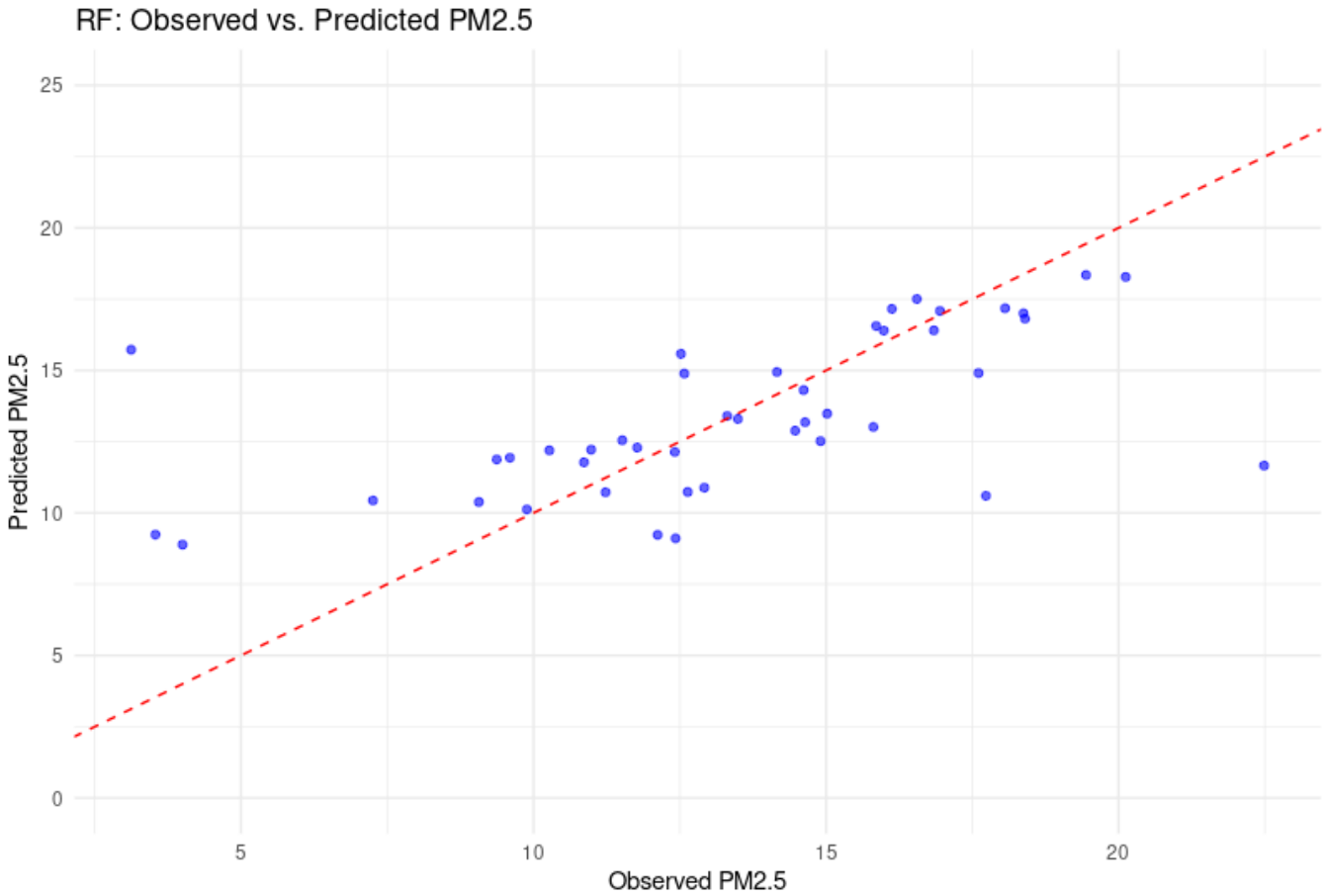


fig19:Scatter plot of observed vs. predicted PM2.5 concentrations using RFSI.

AIMS African Institute for Mathematical Sciences SENEGAL

Conclusion (1/2): Key Findings

- Three interpolation methods were compared: IDW, Ordinary Kriging (OK), and Random Forest Spatial Interpolation (RFSI).
- RFSI achieved the best performance (lowest RMSE, highest R^2), capturing complex spatial patterns.
- IDW was simple but less accurate, as it ignores spatial structure and covariates.
- OK showed smoother results but was limited by stationarity assumptions and variogram fitting.
- These results align with recent studies supporting flexible, data-driven models for urban air pollution.

AIMS African Institute for Mathematical Sciences SENEGAL

Conclusion (2/2): Limitations & Perspectives

- The method is reproducible and adaptable to other urban settings.
- Limitations include:
  - Absence of spatio-temporal modeling;
  - Limited covariate usage in Kriging and RFSI.
- Future directions:
  - Incorporate spatio-temporal models;
  - Integrate traffic, meteorological, and land use data;
  - Extend applications to $PM_{2.5}$ analysis in Senegal.

AIMS African Institute for Mathematical Sciences SENEGAL

Questions?

African Institute for Mathematical Sciences
SENEGAL