# Comparing Spatial Interpolation Methods for $PM_{2.5}$ as Inputs to Urban Decision-Making in Greater Boston

**Soomi Lee, Stratis Ioanndis, Amy Mueller**

Northeastern University

## Abstract

Exposure to ambient $PM_{2.5}$ is estimated to be associated with ˜3 million deaths globally in 2017. Reducing this number requires targeted health-protective interventions, especially in urban areas with higher pollution burden, however obtaining high-resolution $PM_{2.5}$ data in urban environments is challenging due to sparse sensor and regulatory monitor distribution as well as the complex spatial heterogeneity of urban air quality. In this study, we evaluate various spatial interpolation methods to estimate $PM_{2.5}$ concentrations in Brookline, a municipality in Greater Boston, explicitly examining the trade-offs between sensor network size and interpolation performance. Random Forest achieves RMSE of $0.68 \mu g/m^3$ (MAPE of 7.5%), significantly outperforming other methods. The RMSE for all methods decreased by less than 0.02 $\mu g/m^3$ when 15 (40%) fewer sensors were used to train the models. These findings highlight the potential of data-driven spatial interpolation techniques in mitigating tradeoffs between cost and sensor network comprehensiveness in complex urban environments.

## 1 Introduction

$PM_{2.5}$ refers to suspended particles in the air for which the aerodynamic diameter is 2.5$\mu$m or less (U.S. EPA, 2009). Ambient (outdoor) $PM_{2.5}$ poses a significant health risk, estimated to be responsible for ˜3 million deaths and ˜83 million disability-adjusted life years (DALY) worldwide in 2017 (Bu et al., 2021). Access to local data on $PM_{2.5}$ level, particularly in urban environments, could help individuals to avoid exposure to poor air quality and municipalities to implement measures aimed at reducing air pollution. Regulatory monitors are, however, sparsely located due to cost constraints, and interpolating between existing stations is challenging due to the significant spatial variability of urban air quality arising from the heterogeneous structures and diverse emission sources. In the Greater Boston area, regulatory monitors are 2-10 km apart, but urban air quality can vary between city blocks (100m) as shown for our study area in Figure 1, where measurement differences reach up to $20 \mu g/m^3$ (around 50%) across distances of less than 1.5 km.

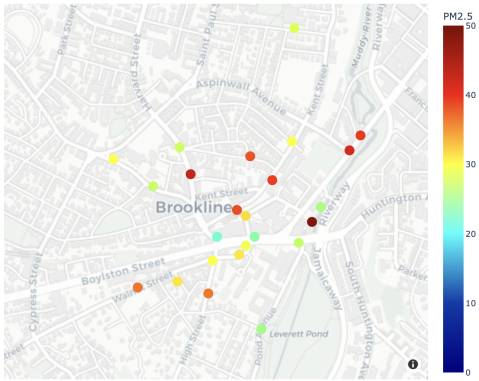


Figure 1: $PM_{2.5}$ concentrations ($\mu g/m^3$) in Brookline, MA on July 28,2024 at 15:00 as reported by QuantAQ air quality monitoring sensors. Significant spatial variation is observed across this area (approximately 2.25 $km^2$ shown), with points located one block apart having concentration difference up to $20 \mu g/m^3$.

In this study, we apply various spatial interpolation methods to $PM_{2.5}$ data in one municipality of the Greater Boston Area while varying the number of sensors used to train the models. Our study yields insights into how interpolation performance changes with sensor density, which can provide guidance for decision-making in designing next-generation systems for urban air quality monitoring.

## 2 Related Work

Spatial interpolation methods, broadly categorized into statistical and machine learning approaches, are widely used to analyze and predict spatial data in air quality monitoring (Lin et al., 2020). For instance, Zhang et al. (2018) employed statistical approaches (inverse distance weighting, Kriging, and spline interpolation) using $PM_{2.5}$ data from 54 air monitoring stations in Shanghai, China, and achieved a minimum root mean square error(RMSE) of $4.3\mu g/m^3$ with inverse distance weighting. In recent years, machine learning techniques have gained popularity, offering flexibility in handling non-linear relationships and complex data patterns. For example, Guo et al. (2022) utilize an XGBoost model that estimates the $PM_{2.5}$ levels at a 500 m $\times$ 500 m spatial resolution using data from 448 micro stations in Lanzhou City, China with an RMSE of $11.0\mu g/m^3$. Cheng et al. (2018) developed ADAIN, a neural attention model that integrates air quality data from 36 monitoring stations with POIs, road networks, and meteorological data achieving an RMSE of 42.6 for $PM_{2.5}$.

## 3 Methodology

### 3.1 Problem Statement

The study aim is to develop a model that can infer the $PM_{2.5}$ value of a geographic location based on $PM_{2.5}$ values from available data sources (e.g., sensors). Formally, suppose we have sensors $\mathcal{S} = \{1, 2, 3 \ldots, N\}$, monitored over a time window $\mathcal{T} = \{1, 2, 3 \ldots, T\}$ with $T$ time steps. Let $y_{i,t} \in \mathbb{R}^+$ be the $PM_{2.5}$ value observed at sensor $i \in S$ at time $t \in T$. We divide sensors in to two non-overlapping sets: context set $\mathcal{C}$ and target set $\mathcal{G}$. We wish to solve the following regression problem: at each time $t \in \{1, 2, 3, ..., T\}$ given (a) the sensor measurements from sensors in $\mathcal{C}$ and (b) the distance between a target sensor $j \in \mathcal{G}$ and each context sensor $i \in \mathcal{C}$ denoted as $d(i, j)$, predict the $PM_{2.5}$ value $\hat{y}_{j,t}$ at sensor $j$. (Generally, such an algorithm could estimate $PM_{2.5}$ at any location, however limiting targets to locations with sensors provides a pathway for model evaluation.) Stated another way, for each sensor $j$ in target set $\mathcal{G}$ we wish to regress a parametric model $f(\mathbf{x}_{j,t}; \beta)$ with hyperparameters $\beta$ that predicts $y_{j,t} \in \mathbb{R}$ from vector $\mathbf{x}_{j,t} \in \mathbb{R}^{2|C|}$. The latter contains (a) measurements from sensors in the context set $\mathcal{C}$ ($\in \mathbb{R}^{|\mathcal{C}|}$), and (b) distances from each of the context set ($\mathcal{C}$) sensors ($\in \mathbb{R}^{|\mathcal{C}|}$).

We explore the following two questions: (a) What is the spatial interpolation performance of various algorithms? (b) How does the performance differ depending on the dimensionality of $\mathcal{C}$? We will investigate this by applying various spatial interpolation methods on $PM_{2.5}$ concentration data from a network of sensors in the town of Brookline, a municipality located adjacent to the city of Boston, MA.

### 3.2 Dataset

We use hourly $PM_{2.5}$ data retrieved from 35 QuantAQ (QuantAQ, 2025) air quality sensors located in Brookline, MA, a town of ˜17.6 $km^2$ (see Figure 2). The analysis period spans from May 24 to September 19, 2024, and all sensors used had over 90% complete data during the study period. Table 1 presents summary statistics across all sensors for the study period; Figure 3(b) presents the distribution of $PM_{2.5}$ values, which follows a lognormal pattern. Figure 3(a) illustrates hourly mean dynamics in $PM_{2.5}$ concentrations in Brookline over 10 days in August, highlighting the significant temporal fluctuations. The overlay in Figure 3(b) shows that the standard deviation of $PM_{2.5}$ among sensors is generally low ($\leq 2\mu g/m^3$) but, as can be observed in Figure 3(a), tends to increase when $PM_{2.5}$ concentrations are high. This is consistent with the presence of spatially and temporally diverse sources of air

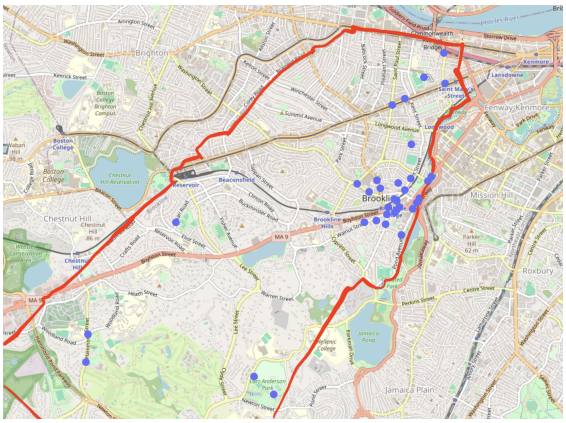

Figure 2: Map of sensors used in this study: violet dots show individual sensors, and the red line indicates municipal limits. Figure created using OpenStreetMap (OpenStreetMap contributors, 2017).

| **Statistic** | **Value** ($\mu g/m^3$) |
|---|---|
| Mean | 8.8 |
| Median | 7.2 |
| Standard deviation | 5.7 |
| Minimum | 0.4 |
| Maximum | 65.0 |
| Lowest mean for 1 sensor | 6.7 |
| Highest mean for 1 sensor | 11.5 |

Table 1: Summary Statistics for $PM_{2.5}$ data measured using 35 QuantAQ sensors across the Town of Brookline from May 24 to September 19, 2024.

pollution in cities, highlighting a challenge for achieving accurate spatial interpolation in a highly imbalanced data scenario.

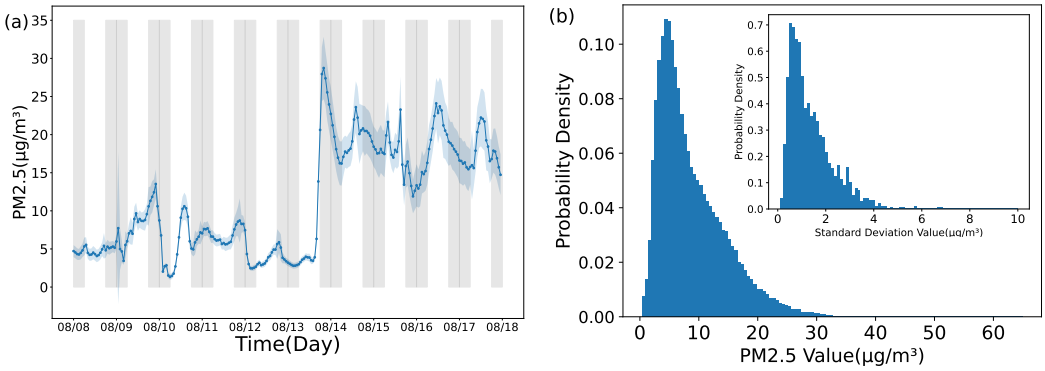


Figure 3: (a) Hourly mean ($\pm$ 1 $\sigma$) of $PM_{2.5}$ across all 35 sensors over 10 days (gray bars indicate nighttime). The values show significant temporal variation and higher standard deviation during hours with elevated $PM_{2.5}$. (b) Probability distribution of hourly $PM_{2.5}$ values for all 35 sensors during the entire study period with overlay showing the probability density of hourly standard deviation of $PM_{2.5}$.

### 3.3 Interpolation Methods and Metrics

We implement Inverse Distance Weighting (IDW) (Li and Heap, 2008), Ordinary Kriging (OK) (Oliver and Webster, 2014; Lin et al., 2020; Webster and Oliver, 2007), and Random Forest (RF) (Breiman, 2001; Yang et al., 2016) for spatial interpolation (see details in Appendix A). We also use the mean (MEAN) as a baseline method against which to evaluate performance of the more sophisticated methods. Performance is assessed using Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), $R^2$ and the mean/median of $R^2$ (see Appendix B). We split the dataset into 80% training, 10% validation, and 10% test set (see Appendix C). Hyperparameter ranges and their optimal values (with respect to the validation set RMSE) are listed in Table 2 in Appendix A.

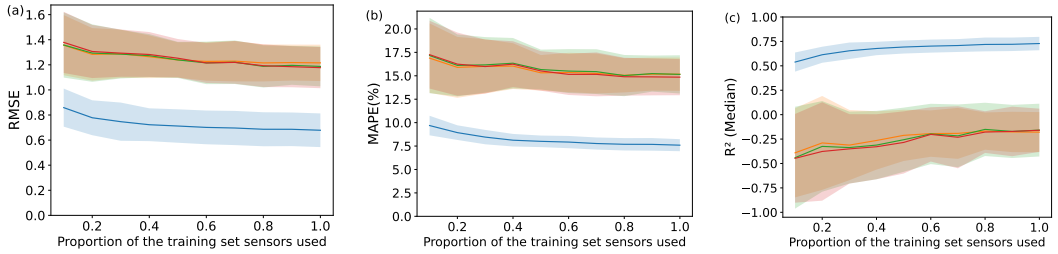


Figure 4: X-axis indicates the fraction of $\mathcal{C}$ used in training the models (i.e., 0.1 is 10% of $\mathcal{C}$ used). Shaded areas in these plots indicate the standard deviation of the performance. RF performs the best while the performance of IDW and OK are comparable to that of MEAN.

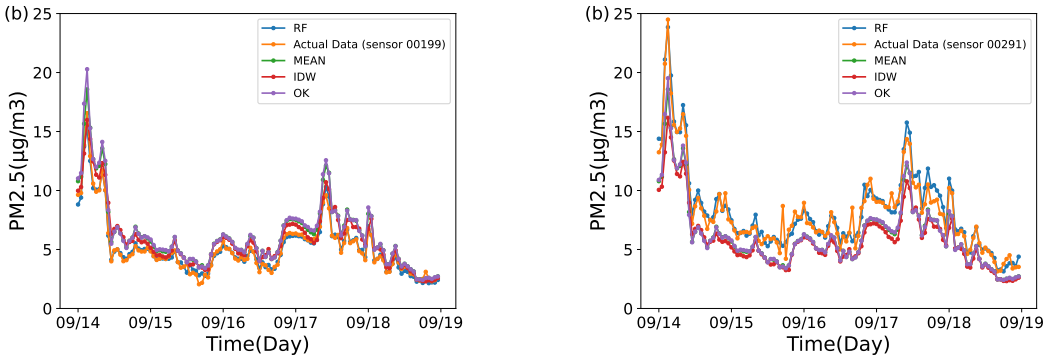


Figure 5: Sample of interpolation results for (a) $PM_{2.5}$ during September 14-19, 2024 at sensor 00199 where all methods use 100% of $\mathcal{C}$ and all methods predict $PM_{2.5}$ relatively accurately. (b) $PM_{2.5}$ during September 14-19, 2024 at sensor 00291 where all methods use 100% of $\mathcal{C}$ but all methods fail to effectively capture the peaks.

## 4 Results

Figure 4 illustrates the spatial interpolation performance with respect to the size of $\mathcal{C}$ used in training. RF significantly outperforms other methods. In contrast, the performance of IDW and OK is comparable to that of MEAN. As both IDW and OK share the assumption of stationarity in the data, this suggests that the variability in $PM_{2.5}$ values cannot be fully explained by spatial distance alone. Furthermore, all methods fail to capture peaks in $PM_{2.5}$ concentrations, which is critical for accurate air quality modeling. Figure 5 illustrates this limitation: a significant discrepancy in predictions is observed during peaks, with Figure 5(b) showing more prominent errors due to its sharper peaks. Additional results including the mean of $R^2$ values are available in Appendix D.

Another notable observation from Figure 4 is that, across all algorithms, the regression performance saturates at around 60%. Thus, retaining only 60% of sensors in $\mathcal{C}$ causes minimal performance degradation. This suggests that simply adding more sensors may not improve performance, and more sophisticated site selection and/or algorithm development is needed. Furthermore, as shown in Figure 4(c), the maximum $R^2_{median}$ value achieved is 0.75, indicating room for improvement.

## 5 Conclusion

We evaluate several algorithms for spatial interpolation of air quality data in an urban environment. Among the methods tested Random Forest performed best even when 40% of available training data were omitted but still struggled to adequately capture $PM_{2.5}$

variability. These findings illustrate the potential of spatial interpolation techniques in balancing the trade-off between data comprehensiveness and cost while emphasizing the need for improved methods that can better capture the heterogeneity of urban air quality.

## References


Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Xiang Bu, Zhonglei Xie, Jing Liu, Linyan Wei, Xiqiang Wang, Mingwei Chen, and Hui Ren. Global pm2. 5-attributable health burden from 1990 to 2017: Estimates from the global burden of disease study 2017. *Environmental Research*, 197:111123, 2021.

Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rong Guo, Ying Qi, Bu Zhao, Ziyu Pei, Fei Wen, Shun Wu, and Qiang Zhang. High-resolution urban air quality mapping for multiple pollutants based on dense monitoring data and machine learning. *International journal of environmental research and public health*, 19(13):8005, 2022.

Jin Li and Andrew D Heap. A review of spatial interpolation methods for environmental scientists. *Geoscience Australia Canberra*, Record 2008/23, 2008.

Yuan-Chien Lin, Wan-Ju Chi, and Yong-Qing Lin. The improvement of spatial-temporal resolution of pm2. 5 estimation based on micro-air quality sensors by using data fusion technique. *Environment international*, 134:105305, 2020.

MA Oliver and R Webster. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena*, 113:56–69, 2014.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org` , 2017.

Inc QuantAQ. Quantaq, 2025. URL `https://quant-aq.com/products/modulair`. Accessed: 2025-02-D2.

U.S. EPA. Integrated science assessment (isa) for particulate matter (final report, dec 2009). Technical Report EPA/600/R-08/139F, U.S. Environmental Protection Agency, Washington, DC, December 2009.

Richard Webster and Margaret A Oliver. *Geostatistics for environmental scientists*. John Wiley & Sons, 2007.

Ren-Min Yang, Gan-Lin Zhang, Feng Liu, Yuan-Yuan Lu, Fan Yang, Fei Yang, Min Yang, Yu-Guo Zhao, and De-Cheng Li. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological indicators*, 60:870–878, 2016.

Xuanyi Zhang, Runhe Shi, and Maosi Chen. Comparative study of the spatial interpolation methods for the shanghai regional air quality evaluation. In *Remote Sensing and Modeling of Ecosystems for Sustainability XV*, volume 10767, pages 188–196. SPIE, 2018.

## A Interpolation Methods

We review the three methods we implemented below. Hyperparameters explored and their optimal values (selected using the validation set) are listed in Table 2.

**Inverse Distance Weighting** Inverse distance weighting (IDW) estimates the values of an attribute at unsampled points by using a linear combination of values from sampled points, weighted by an inverse function of the distance between the point of interest and the sampled points where $p \in \mathbb{N}$ is an exponent that determines the decay in distance. Formally, it can be written as Equation 1.

$$\hat{y}_{j,t} = \frac{\sum_{i \in \mathcal{C}} \frac{1}{d(i,j)^p} y_{i,t}}{\sum_{i \in \mathcal{C}} \frac{1}{d(i,j)^p}}. \tag{1}$$

The underlying assumption is that sampled points closer to the unsampled point are more similar in value than those farther away (Li and Heap, 2008).

**Ordinary Kriging** Kriging is a term used for a family of least-squares regression methods used for spatial interpolation (Oliver and Webster, 2014). Ordinary Kriging assumes that the spatial random variable satisfies the second-order stationarity assumption, meaning its mean is constant and its covariance depends only on the distance between points (Lin et al., 2020). OK predicts the random variable at an unobserved location as a weighted sum of observed values, with weights determined by solving a system of linear equations that incorporate semivariances derived from a fitted variogram model (Webster and Oliver, 2007). We use the Pykrige library to implement Ordinary Kriging.

**Random Forest** Random Forest is a machine learning technique that combines the predictions of multiple decision trees to reduce overfitting and improve accuracy (Breiman, 2001). During training, each tree is built using a random subset of the original data, sampled with replacement, and a randomly selected subset of predictors is chosen for splitting at each node (Breiman, 2001; Yang et al., 2016). We use the scikit-learn library for our Random Forest experiments.

Table 2: Hyperparameters explored with the optimal values identified via the validation set (with respect to RMSE) in bold. For IDW, $p$ represents the power of the inverse distance weight. In RF, 'number of trees' specifies the number of trees in the forest, while 'max_features' denotes the number of features to consider when determining the best split. Among 'max_features' used, 'sqrt' indicates that the maximum number of features considered for a split is the square root of the total number of features, and 'log2' specifies that it is the base-2 logarithm of the number of features. 'oob_score' determines whether out-of-bag samples are used to estimate the generalization score.

| **Algorithms** | **Hyperparameters** |
|---|---|
| IDW | $p$ = [**1**,2,3] |
| OK | variogram_model = [**linear**, power, gaussian, spherical] |
| RF | number of trees : [40,60,80,**100**, 120, 140], max_features = [1.0, **sqrt**, log2], oob_score=[True, **False**] |

## B Metrics

The following equations define the Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), $R^2$, $R^2_{mean}$, $R^2_{median}$. Let $\hat{y}_{i,t} \in \mathbb{R}^+$ denote the predicted $PM_{2.5}$ value for sensor $i$ at time $t$. The target set is denoted by $\mathcal{G}$, and the number of time steps is represented by $T$.

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{G}|T}\sum_{t\in T}\sum_{i\in\mathcal{G}}(y_{i,t}-\hat{y}_{i,t})^2}, \tag{2}$$

$$\text{MAPE} = \frac{100\%}{|\mathcal{G}|T}\sum_{t\in T}\sum_{i\in\mathcal{G}}\left|\frac{y_{i,t}-\hat{y}_{i,t}}{y_{i,t}}\right|, \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i\in\mathcal{G}}(y_{i,t}-\hat{y}_{i,t})^2}{\sum_{i\in\mathcal{G}}(y_{i,t}-\bar{y}_t)^2}, \tag{4}$$

$$R^2_{mean} = \text{Mean}_{t=1}^{T}\left[1 - \frac{\sum_{i\in\mathcal{G}}(y_{i,t}-\hat{y}_{i,t})^2}{\sum_{i\in\mathcal{G}}(y_{i,t}-\bar{y}_t)^2}\right], \tag{5}$$

$$R^2_{median} = \text{Median}_{t=1}^{T}\left[1 - \frac{\sum_{i\in\mathcal{G}}(y_{i,t}-\hat{y}_{i,t})^2}{\sum_{i\in\mathcal{G}}(y_{i,t}-\bar{y}_t)^2}\right], \tag{6}$$

$$\text{where } \bar{y}_t = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} y_{i,t}. \tag{7}$$

For $R^2_{mean}$, we calculate the mean of the $R^2$ values for each time step instead of computing a single $R^2$ value for the entire test set. Likewise for $R^2_{median}$, we compute the median of the $R^2$ values for each time step. This approach is adopted because in our dataset temporal fluctuations are generally much larger than spatial variability. As a result, even when spatial interpolation is suboptimal, the overall $R^2$ for the entire dataset may be disproportionately high. We included $R^2_{median}$ in the results section because we observed that $R^2_{mean}$ was distorted by extreme outliers, which often occur when the variance in the data is very small.

## C Experimental setup

As shown in Figure 6, the dataset is divided in three chunks along both the spatial and temporal axes. Firstly, in the spatial dimension, $\mathcal{C}_{TRN}, \mathcal{C}_{VAL}, \mathcal{C}_{TST} \in \mathcal{S}$ (where $\mathcal{S} = \{1, 2, 3 \ldots, N\}$ is a set of all sensors in the dataset) refer to the context set (observed sensors) used in the training, validation, and testing phase respectively. Likewise, $\mathcal{G}_{TRN}, \mathcal{G}_{VAL}, \mathcal{G}_{TST} \in \mathcal{S}$ denotes the target set for the training, validation, and testing phase respectively. The training and validation phases use an identical set of sensors as context set and target set. The following equations hold for the training/validation data.

$$\mathcal{C}_{TRN} = \mathcal{C}_{VAL},\ \mathcal{G}_{TRN} = \mathcal{G}_{VAL} \tag{8}$$

$$|\mathcal{C}_{TRN}| = |\mathcal{C}_{VAL}| = 0.55 \times |\mathcal{S}| \tag{9}$$

$$|\mathcal{G}_{TRN}| = |\mathcal{G}_{VAL}| = 0.15 \times |\mathcal{S}| \tag{10}$$

Meanwhile, for the testing phase the following equations hold, where the target set $\mathcal{G}_{TST}$ refers to a subset of the data held back both in space and time.

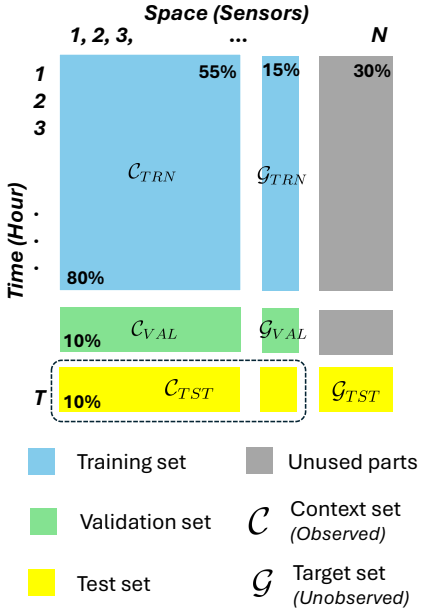


Figure 6: Evaluation strategy and dataset division. The dataset is divided both in the temporal domain and spatial domain.

$$\mathcal{C}_{TST} = \mathcal{C}_{TRN} \cup \mathcal{G}_{TRN} \tag{11}$$

$$|\mathcal{C}_{TST}| = 0.7 \times |\mathcal{S}| \tag{12}$$

$$|\mathcal{G}_{TST}| = 0.3 \times |\mathcal{S}| \tag{13}$$

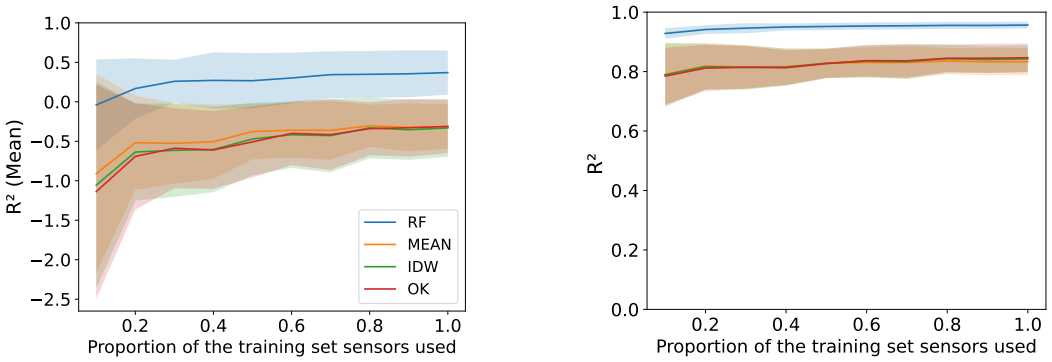


Figure 7: Model performance reported using two $R^2$ metrics where the x-axis indicates the size of the subset of $\mathcal{C}$ used in the models. Shaded areas indicate the standard deviation of the performance. RF outperforms other models which are comparable to using the mean (MEAN).

As can be seen in Equation 11, in the testing stage (using the optimal hyperparameters determined in the validation set) the model is trained using all sensors in the training/validation set ($\mathcal{C}_{TRN} \cup \mathcal{G}_{TRN}$). Then it is tested on the target set $\mathcal{G}_{TST}$. It is important to note that $\mathcal{G}_{TST}$ is not used in the training / validation set, as they are used to evaluate the generalization performance of the models.

In the temporal domain, the data are sequentially divided into three consecutive chunks for the training, validation, and test sets. The first 80% of the data is used for the training set, the next 10% is used for the validation set, and the remaining 10% is used for the test set.

The performance varies significantly depending on the randomly sampled combination of context and target sensors. To account for various combinations of context and target sets, this process is repeated 30 times with different sensor combinations. The mean performance on the validation set (measured using RMSE) is used to determine the best hyperparameters. Additionally, experiments were conducted using varying sizes of the context set $\mathcal{C}$. To achieve this, multiple subsets of the context set, each of different sizes, were randomly sampled and used in the modeling pipeline to estimate the values of the target set. In our experiment, we used subsets with at step sizes of 10 percentage points (10%, 20%, 30%, etc.) each randomly sampled 30 times.

## D Additional Results

Figure 7 shows the performance of the models in terms of the mean of $R^2$ ($R^2_{mean}$) computed at each time step and conventionally used $R^2$ values respectively. Interestingly, while the performance of IDW and OK are reasonable in terms of RMSE, MAPE, and $R^2_{mean}$ values are less than zero. This occurs because low $PM_{2.5}$ levels are much more frequent than high $PM_{2.5}$ levels and the variance is also low when $PM_{2.5}$ levels are low. When the variance is low, small errors can lead to extremely small $R^2$ values. These extreme outliers cause $R^2_{mean}$ to be smaller than $R^2_{median}$.