Filling in the Map: Spatial Interpolation of Small Particle Air Pollution in Houston from Multi-Source Air Quality Data

Isabelle Adeyinka, Bartu Citci, Connie Huang, Andrew Kim, Reid Westervelt, and Matthew Ye

Faculty Mentor: Dr. Lorenzo Luzi
PhD Mentor: Ali Azizpour
Houston Chronicle Sponsors: Rebekah F. Ward, Alexandra Kanik, Matt Zdun

March 24, 2025

Abstract

Air pollution, specifically fine particulate matter $PM_{2.5}$, is an issue that disproportionately affects urban communities, yet the methods used to measure and regulate it often fail to provide an accurate representation of pollution levels across a city. The government regulatory agency monitoring Houston, namely the Texas Commission on Environmental Quality (TCEQ), relies on a small number of strategically placed air quality monitors, leaving significant gaps in pollution data for communities outside of these locations. This study explores spatial interpolation techniques as a means to fill these gaps, using a combination of regulatory and non-regulatory air quality data, meteorological conditions, traffic volume, and industrial emissions to interpolate $PM_{2.5}$ concentrations in Houston. The analysis compares Ordinary Kriging and Regression Kriging models to determine the most effective approach for estimating pollution in unmonitored areas. In addition to assessing the accuracy of these models, this study examines the spatial distribution of pollution sources and their correlation with demographic and socioeconomic factors. By developing a more comprehensive method for estimating air pollution, this research contributes to the broader discussion on environmental justice and public health, providing a tool that empowers Houston residents with the ability to better understand and advocate for improved air quality monitoring in their neighborhoods.

Contents

1 Introduction

1.1 Background

The issue of air pollution and its harmful effects on both the environment and human health have been studied for decades (World Health Organization, 2025). One type of pollution in particular, $PM_{2.5}$, has uniquely adverse impacts on human health outcomes and environmental patterns. (Environmental Protection Agency, 2025; New York State Department of Health, 2025). Airborne particulate matter (PM) is not a singular pollutant but a mixture of various chemicals, ranging from aerosols to metallic compounds. $PM_{2.5}$ refers to particulate matter 2.5 micrometers in diameter or less, about 1/30th the width of a human hair (Nazarenko et al., 2020).

Since it is small enough to be ingested by humans and enter their bloodstream, $PM_{2.5}$ can cause various health problems for those who are exposed to it. Decreased lung function, irregular heartbeats, heart attacks, premature death in people with heart or lung disease, and aggravated asthma are just some of the adverse health impacts caused by exposure to $PM_{2.5}$ (California Air Resources Board, 2025). A significant presence of $PM_{2.5}$ can also contribute to climate change. Some components of $PM_{2.5}$, such as black carbon, have climate-cooling properties, while others, such as nitrate and sulfate, have climate-warming effects (Strasert et al., 2019). Studies show that soil quality, plant growth, and crop yield are also negatively affected by the presence of $PM_{2.5}$ (Chao et al., 2025).

For these reasons, the causes of increased $PM_{2.5}$ in the atmosphere have been the subject of many recent studies. The most prevalent anthropogenic sources of $PM_{2.5}$ include vehicle emissions, industrial processes, heating and burning activities, and industrial emissions (Thangavel et al., 2022). Natural sources include wildfires, volcanic eruptions, and organic compounds released by plants and trees (VanCuren and Gustin, 2015). $PM_{2.5}$ pollution is an especially pressing issue for Houston, which has the worst air quality in Texas, a state that ranks 45th in air quality nationwide (U.S. News & World Report, n.d.).

Houston residents are uniquely affected by such pollution, which can cause a slew of respiratory and cardiopulmonary health effects (Sexton et al., 2007). Some demographic groups are more affected by particulate matter pollution than others, which is another motivator for this project. Biased housing policies, discriminatory zoning laws, and other factors have allowed racial and ethnic disparities in $PM_{2.5}$ exposure to persist over time (Tessum et al., 2021). These disparities highlight the need for more accurate and accessible methods of measuring $PM_{2.5}$ concentrations, particularly in neglected communities. Recently, there has been growing concern among Houston residents regarding air pollution and its monitoring (Ferrell, n.d.). Namely, the Texas Commission on Environmental Quality (TCEQ) bases their pollution regulations and permit issuances based on only twelve monitors that measure $PM_{2.5}$ across the entire Houston Metropolitan Area. As a result, communities that are not near these monitors have little information about the level of $PM_{2.5}$ pollution they face, leading to a lack of awareness regarding the health impacts the pollutant can cause. This project aims to address these concerns regarding the TCEQ's monitoring of $PM_{2.5}$ in the Houston-The Woodlands-Pasadena Metropolitan Area.

1.2 Objectives

In order to support the Houston Chronicle's mission of educating the public about the sources of air pollution and the communities it impacts, we aimed to achieve the following project objectives:

- **(Objective 1) Create a Spatial Interpolation Model for Air Pollution**: Our primary goal was to develop a machine learning model that uses data from existing air quality sensors in Houston, along with additional factors such as weather, traffic volume, and known nearby polluters, to interpolate $PM_{2.5}$ levels in areas without sensors. This model helps identify regions of Houston that are likely to have high levels of air pollution, allowing us to advocate for the strategic placement of new sensors in those areas. The models developed for this objective can be found in Sections 5.1 - 5.5. From these models, we also performed a hotspot analysis of neighborhoods likely to have high air pollution in Section 5.6, supplementing

descriptive analyses of neighborhood demographics associated with worse air pollution in Sections 4.2.1 and 4.2.5.

- **(Objective 2) Compare Regulatory vs. Non-Regulatory Air Quality Data**: To test the hypothesis that government air quality monitors are not geographically distributed in a way that accurately represents air quality across Houston, we compared the descriptive statistics of air quality and sensor coverage between TCEQ (regulatory) sensors and PurpleAir (non-regulatory) sensors in Section 4.2.1. We also our interpolation model using only data from PurpleAir sensors or TCEQ sensors, to assess how the removal affects the model's interpolation quality. This ablation study, found in Section 5.7, helps understand the impact of sensor distribution on the accuracy of pollution estimates, as well as what hotspots are captured when only non-regulatory or regulatory sensors are considered.
- **(Objective 3) Identify Major Houston-Area Polluters**: As part of our descriptive analysis, we identified the primary sources of emissions impacting the air quality of Houston communities. Obtaining a clearer understanding of these polluters can help government agencies implement stronger regulations in affected areas. In addition to conducting descriptive analyses of high-polluting industrial facilities as in Section 4.2.2, we used our spatial interpolation model to obtain coefficient values representing the magnitude of different sources' (e.g., industrial facilities, vehicular traffic) contributions to air pollution.
- **(Objective 4) Map Air Pollution for At-Risk Communities**: For our final deliverable, we created an interactive map that allows Houston Chronicle readers to assess the level of $PM_{2.5}$ at their own address. By providing a more comprehensive map that incorporates data beyond just regulatory sensors, this tool allows Houston residents to better understand local risks for pollution and how these risks impact their communities. We implement the results of our strongest interpolation model for this interface in Section 5.8.

2 Literature Review

2.1 $PM_{2.5}$ Pollution in Houston

Various studies have detailed the causes and effects of $PM_{2.5}$ pollution, specifically in Houston. Factors such as temperature, wind speed, wind direction (Tai et al., 2012), traffic emissions (Zhang et al., 2017), and single-point factory pollutants (Sadeghi et al., 2020) have all been shown to increase $PM_{2.5}$ levels in Houston. We will incorporate this information to enhance our model, allowing it to interpolate $PM_{2.5}$ measurements in areas that are not comprehensively covered by TCEQ regulatory data monitors. We also use previous research as a motivating factor to provide such data for $PM_{2.5}$ concentrations.

On February 7, 2024, the United States Environmental Protection Agency (EPA) strengthened regulation regarding annual $PM_{2.5}$ standards, lowering the limit from 12 $\mu g/m^3$ to 9 $\mu g/m^3$ (Texas Commission on Environmental Quality, n.d.). Currently, the short-term standard for the daily average of $PM_{2.5}$ concentrations is 35 $\mu g/m^3$ (California Air Resources Board, n.d.). However, some studies assert that there is no true "safe" level of $PM_{2.5}$ exposure. For example, a time-series analysis of mortality in Boston found that for every 10 $\mu g/m^3$ increase in short-term $PM_{2.5}$ exposure, there was a 2.8 percent increase in mortality (Kloog et al., 2013). Critics argue the EPA's threshold for safe $PM_{2.5}$ concentration is arbitrary and lenient, especially when compared to that of the World Health Organization (WHO), which states that the annual average concentration of $PM_{2.5}$ should not exceed 5 $\mu g/m^3$. This is especially relevant to Texas, which follows the EPA's guidelines, rather than the WHO's. This growing concern over $PM_{2.5}$ exposure reflects a need for accurate and reliable methods to monitor and predict concentrations, particularly in areas lacking direct sensor measurements.

2.2 Spatially Distributed Data

A critical component of our study is the use of spatially distributed data, which involves the measurement of a physical phenomenon across the Earth's surface (Doreian, 1981). Often, such

Longitude	Latitude	PM2.5 (μg/m³)
-95.28	29.83	188
-95.24	29.73	21
…	…	…

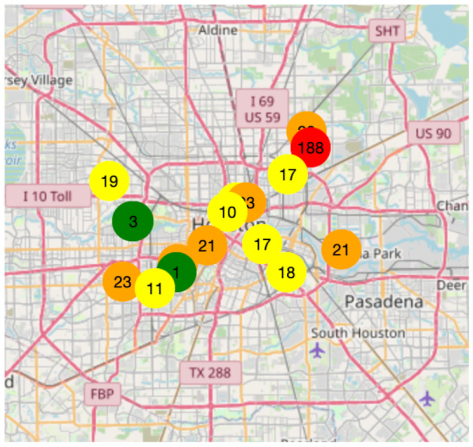

Figure 1: Example of spatially distributed data. Depicted above are the $PM_{2.5}$ concentrations of selected air quality sensors in the Houston area from both PurpleAir and TCEQ in May 2024.

data is tabular in nature, with geographic coordinates (e.g., a pair of longitude, latitude) being assigned to the phenomenon. In this case, the physical phenomenon is the concentration of $PM_{2.5}$ $\mu g/m^3$, and the measurements are obtained by either regulatory sensors (i.e., those of the TCEQ, whose recordings are used for policy decisions) or non-regulatory sensors (e.g., those of PurpleAir, whose recordings are not used for policy decisions).

2.3 Spatial Autocorrelation

Crucial to spatially distributed data is the concept of spatial autocorrelation, which refers to the relationship between measurements based on spatial proximity (that is, the "correlation" based on "self," dependent on spatial distribution) (Getis, 2009). Positive spatial autocorrelation occurs when similar values cluster together (e.g., nearby sensors both having low $PM_{2.5}$ readings). Meanwhile, negative spatial autocorrelation occurs when dissimilar values are adjacent in a "checkerboard" pattern (e.g., a sensor with low $PM_{2.5}$ next to one with high $PM_{2.5}$). A lack of spatial autocorrelation indicates no discernible pattern between points in space.

The Moran's I test (Moran, 1950) is used to assess spatial autocorrelation. The test statistic is defined as:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{1}$$

where N is the number of sensors, x_i and x_j are the values of the measured phenomenon at sensors i and j respectively, $\overline{x}$ is the mean value, and w_{ij} is a spatial weight derived from a distance matrix. When constructing this matrix, sensors within a specified distance (e.g., 1 km) are neighbors with a weight w_{ij} of 1, while those beyond this distance have a weight w_{ij} of 0. This is computed for every sensor/sensor pair in the dataset, with W being the sum of all these weights. Positive Moran's I values indicate positive spatial autocorrelation, while negative values suggest negative spatial autocorrelation. To determine statistical significance of the observed Moran's I statistic, the recorded values are randomly shuffled between the sensors to create a distribution of theoretical Moran's I statistics. The observed Moran's I is compared to this distribution, and a p-value is calculated. If the p-value is less than 0.05, the data is considered to exhibit significant spatial autocorrelation.

Related to the concept of spatial autocorrelation is hotspot analysis. The global Getis-Ord G_i statistic, which can itself serve as a measure of spatial autocorrelation, provides information on whether a given spatially distributed dataset displays clustering (Getis and Ord, 1992). Meanwhile, the local G_i^* statistic identifies specific locations where recorded values are significantly higher or lower than the average across the entire region. These are referred to as hotspots and

coldspots, respectively (Getis and Ord, 1992). The local G_i^* statistic essentially quantifies the intensity of the clustered hotspot or coldspot, with a p-value determining the statistical significance. This method allows us to identify which areas within the entire region are specifically impacted. Running this analysis is especially useful for finding the source of polluters that are hard to trace such as $PM_{2.5}$. A study done in China found hotspots directly along the Danjiang River, which were areas with high land use for farming (Xie et al., 2024). Not only were hotspots detected, but a link between high pollution and highly populated areas with intensive crop farming and livestock was also identified. Although common uses of this statistic are for finding pollution hotspots (Ruidas and Pal, 2022), it can also be used for social sciences such as a study done in India to find states with upper primary educational development (Jana and Sar, 2016).

2.4 Spatial Interpolation Techniques

Predicting $PM_{2.5}$ concentrations in a location without sensors may be framed as a spatial interpolation problem. Spatial interpolation is defined as the prediction of a physical phenomenon at some point where no measurement is available (Kyriakidis and Goodchild, 2006). By learning on known locations, the model generates predictions of measurements across a specified grid, in our case $PM_{2.5}$ concentrations across the entire Houston metro area. Since each interpolation model has its own assumptions and constraints, it is important to determine which model is most applicable to this project. There was a comprehensive review of 61 spatial interpolation methods across 53 comparative studies to assess their effectiveness (Li and Heap, 2011). Broadly speaking, there are two general types of models: geostatistical models, which account for spatial autocorrelation, and deterministic models, which do not. Ordinary Kriging and Ordinary Co-Kriging are among the most frequently used approaches in the former, while Inverse Distance Weighting is commonly used in the latter. In this section, we discuss each model in terms of our objective of estimating $PM_{2.5}$ in Houston.

2.4.1 Inverse Distance Weighting

Inverse distance weighting is a prototypical example of deterministic modeling. The core assumption is that points close to each other will behave similarly (Choi and Chong, 2022). The interpolated value of $PM_{2.5}$ at a point s then follows the formula:

$$Z(s) = \frac{\sum_i \frac{Z_i}{d(s,i)^p}}{\sum_i \frac{1}{d(s,i)^p}} \tag{2}$$

where $Z(s)$ is the predicted $PM_{2.5}$ value at a location s; Z_i is the known $PM_{2.5}$ value at point i; $d(s,i)$ is the distance between the unknown location s and the known point i; and p is an exponent used to determine how strong distance affects weighting. Note that this is essentially an average of the neighboring $PM_{2.5}$ observations, weighted by distance raised to some exponent. Inverse distance weighting is most effective when one expects nearby points to have similar values and there are many known values. Inverse distance weighting has historically been used to interpolate pollution in large cities such as Valencia, with moderate success (Contreras and Ferri, 2016). However, the choice of the exponent value is relatively controversial and requires careful tuning (De Mesnard, 2013). Furthermore, small distances between the point of interpolation and the observation create a small denominator, resulting in a very large interpolated value. Lastly, the simplicity of the equation does not take into account spatial autocorrelation.

2.4.2 Kriging Models

Geostatistical interpolation models, meanwhile, frame interpolation as a sum of three components:

$$PM_{2.5_{\text{ground truth}}}(s) = PM_{2.5_{\text{trend}}}(s) + PM_{2.5_{\text{SA}}}(s) + \varepsilon(s) \tag{3}$$

Specifically, the ground truth at a location in space s is a sum of:

1. A deterministic, "trend"-driven component that represents large-scale changes in the data across the grid. This can be modeled using a regression model (e.g., linear regression), with features such as rainfall, traffic volume, or external polluters.

2. An "SA" component that captures the spatial autocorrelation between nearby points in the grid.

3. "White noise" error that is not captured by either of the other two components.

In Ordinary Kriging, it is assumed that the "trend" is constant over the grid—that is, $PM_{2.5_{\text{trend}}}(s) = \mu$. Therefore, any variation in $PM_{2.5}$ is solely due to distance in space (i.e. $PM_{2.5_{\text{SA}}}(s)$). This assumption, known as first-order stationarity, means that the only inputs needed for the model are the geographical coordinates of each sensor and their corresponding $PM_{2.5}$ values (Finne and Sauzet, 2025). The estimation $\hat{PM_{2.5}}$ is then essentially a weighted sum of the following form:

$$\widehat{PM_{2.5}}(s) = \sum_{i=1}^{n} \lambda_i PM_{2.5_{\text{ground truth}}}(s_i) \tag{4}$$

where $PM_{2.5}(s_i)$ represents known values, and λ_i are weights (Cressie, 1989).

To compute the weights, a function known as the variogram is derived from the observed data. The variogram is expressed as:

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{n} \left[PM_{2.5_{\text{ground truth}}}(s_i) - PM_{2.5_{\text{ground truth}}}(s_i + h) \right]^2 \tag{5}$$

where h represents the distance between two given sensors and n is the total number of sensor/sensor pairs. In this way, the variogram essentially measures how the variation in $PM_{2.5}$ changes over a given space–and thus captures spatial autocorrelation (Oliver and Webster, 2014). With the variogram fit, the model then imposes a constraint through first-order stationarity, which states that the expectation value of $PM_{2.5_{\text{ground truth}}}$ is also equal to μ. Therefore:

$$E(PM_{2.5_{\text{ground truth}}}) = E(\widehat{PM_{2.5}}) = \mu \tag{6}$$

$$E(\hat{PM_{2.5}}) = E[\sum_{i=1}^{n} (\lambda_i PM_{2.5_{\text{ground truth}}})] = \sum_{i=1}^{n(\lambda_i E[PM_{2.5_{\text{ground truth}}}])} \tag{7}$$

$$E(\hat{PM_{2.5}}) = \sum_{i=1}^{n} \mu\lambda_i \tag{8}$$

In order for $E(\widehat{PM_{2.5}}) = \mu$, $\sum(\lambda_i)$ must equal 1. Therefore, while μ is not explicitly calculated during ordinary kriging, it is implicitly taken into account by normalizing the sum of the weights to 1. This then defines an optimization problem in which the objective is to minimize the variance of the estimation error by selecting the optimal set of weights:

$$Var[\hat{PM_{2.5}} - PM_{2.5_{\text{ground truth}}}] = E[(\hat{PM_{2.5}} - PM_{2.5_{\text{ground truth}}} - E[\hat{PM_{2.5}} - PM_{2.5_{\text{ground truth}}}])^2] \tag{9}$$

Through algebraic manipulation, this loss function can be rewritten in terms of the variogram, which is fit from the data. By taking the partial derivative of the function with respect to each weight λ_i in the data, a series of simultaneous equations is established. The solution of these equations then provides the weights, from which the interpolation can be made.

Regression Kriging (or Kriging with External Drift), meanwhile, relaxes the assumption of first-order stationarity (Figure 2). Critically, $PM_{2.5_{\text{trend}}}$ is a regression model that is trained on the selected auxiliary features. The trend at each measured location is then quantified. A process of "de-trending" follows, in which $PM_{2.5_{\text{trend}}}$ is subtracted from the ground truth. This leaves the residuals, which represent the variance not explained by trend (Zimmerman et al., 1999). Ordinary Kriging is then applied to the resulting residuals to obtain $PM_{2.5_{\text{SA}}}$ for all points on the grid by the same methods described above (the subscript SA stands for "spatially autocorrelated"). Finally, a "re-trending" step is applied, where $PM_{2.5_{\text{trend}}}$ is calculated for all points on the grid and added to the previously calculated spatially autocorrelated component

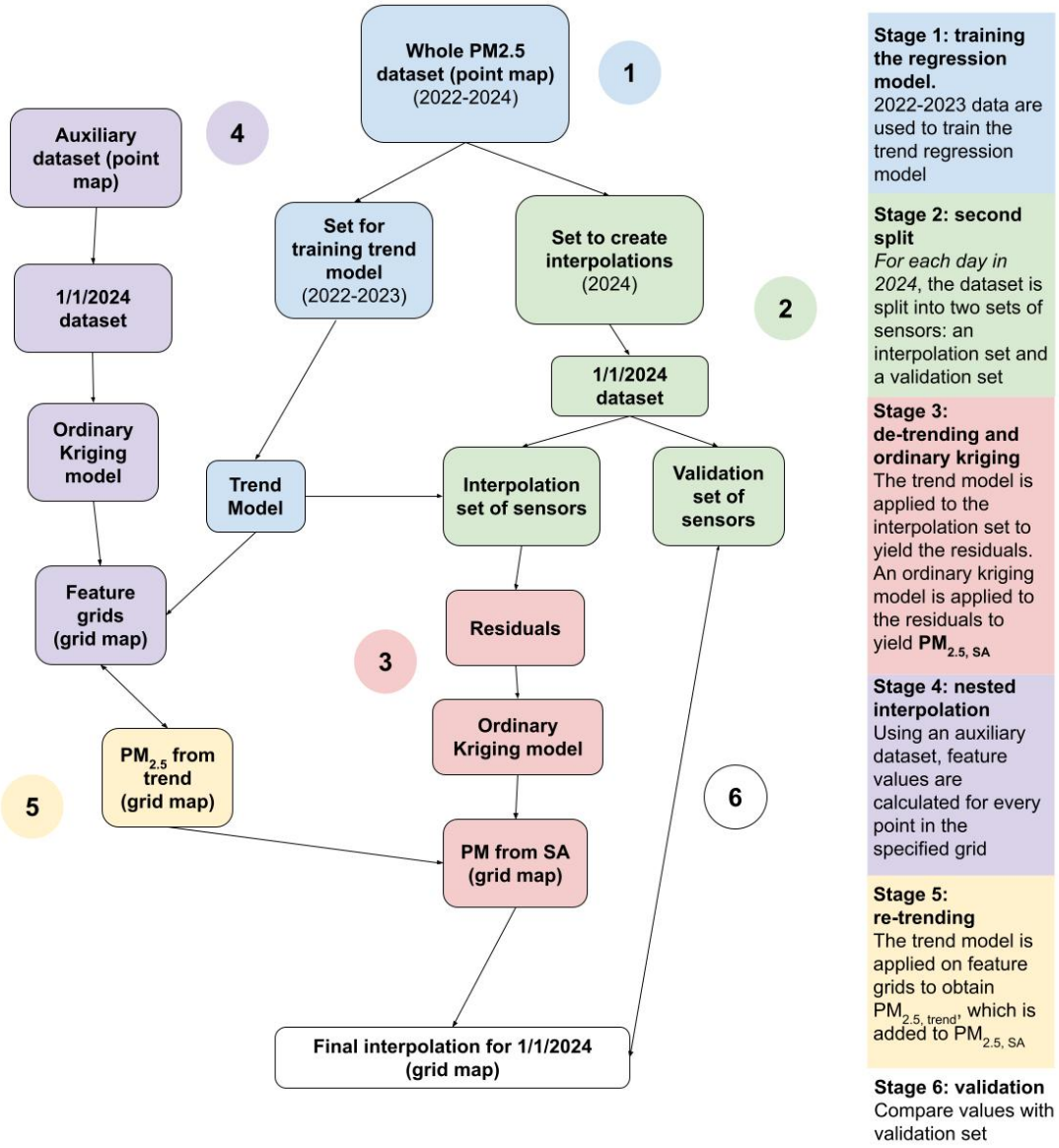

Figure 2: Schematic of Regression Kriging. The ground truth is first used to train a model. The model then predicts $PM_{2.5}$ for each sensor to yield the trend component of the model. De-trending is then applied to obtain the residuals. Ordinary Kriging is then applied over the residuals to obtain the spatially autocorrelated component of the model in the form of a grid. Through re-trending, the model is applied over a series of feature grids to add the trending component and yield the final prediction.

to yield the final prediction (Mesić Kiš, 2016). The findings of Li and Heap suggested that geostatistical interpolation methods generally outperformed deterministic models. Among these, Regression Kriging was concluded to be the most effective model due to the inclusion of auxiliary variables in the predictions (Li and Heap, 2011). Past work has utilized Regression Kriging to interpolate the concentration of pollutants across a city, such as Mercer et al.'s work on gaseous oxides of nitrogen in Los Angeles (Mercer et al., 2011). To the best of our knowledge, however, no study has utilized a combination of meteorological, traffic volume, and single-point factory emissions as auxiliary features to interpolate $PM_{2.5}$ concentrations.

2.5 Regression Techniques

Regression is the prediction or learning of a numerical feature for a new case, given example cases. In contrast to classification, which predicts categorical features, regression is used to predict numerical features. In this section, we expound upon three regression techniques implemented across our models to estimate $PM_{2.5}$ levels across Houston, the three models being 1) linear regression, 2) random forest, and 3) k-nearest neighbors.

2.5.1 Linear Regression

Linear regression is among the simplest forms of regression, involving a linear combination of predictors. Its results are easily interpretable and mathematically straightforward to explain. For a given number of p variables, the model can be simply explained as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{10}$$

where y is the feature being predicted, $x_1, x_2, \cdots, x_p$ are the independent variables or predictors, β_0 is the intercept, $\beta_1, \beta_2, \cdots, \beta_p$ are the regression coefficients associated with each predictor x_j, and ε is the random error term that captures variation not explained by the linear combination of the predictors, typically represented as a Gaussian distribution with mean zero and constant variance.

There are four assumptions of the features made by the model, being:

1. **Linearity:** There exists a linear relationship, positive or negative, between each of the independent variables $x_1, x_2, \cdots, x_p$ and the dependent variable y.

2. **Independence:** The error terms ε are assumed to be statistically independent across observations.

3. **Homoscedastity:** The error terms have equal variance σ^2. In other words, this assumption implies that the variability of the residuals does not depend on the values of the predictors.

4. **Normality:** The error terms are normally distributed. This has no effect on the consistency on the predictions themselves, but does impact hypothesis testing.

While these assumptions are what the model is founded upon, many properties of linear models remain valid even without all four assumptions strictly met.

To estimate the regression coefficients, one must minimize the Ordinary Least Squares (OLS) loss function, as shown below:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2$$

where $\mathcal{L}(\boldsymbol{\beta})$ is the total loss of the function, y_i is the observed value, $\mathbf{x}_i^\top$ is the vector of predictor variables, and β is the regression coefficient. As previously mentioned, the model will find regression coefficients that minimize the total squared differences between observed (truth) and calculated (prediction) values of y_i.

Now that we have established the classic linear regression model, we now introduce regularization. Regularization is the implementation of additional penalties to the loss function to keep the model from getting too complex, preventing overfitting and improving generalization. There are many variants to how a model can be regularized, but in particular, two penalties are commonly used in linear regression.

The first is L1 regularization, or Lasso regularization, in which the sum of the absolute values of the model's coefficients is added to the loss function. By rewarding the reduction of less important coefficients to zero, it encourages sparsity and effectively performs variable selection. Mathematically, the loss function with the L1 penalty can be expressed as:

$$\mathcal{L}_{\text{lasso}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ controls the strength of the regularization overall.

The second mentionable regularization technique is L2 regularization, or Ridge regularization. This penalty adds the sum of the squared values of the model's coefficients to the loss function. Instead of forcing coefficients all the way to zero, they are instead rewarded for being small, such that features with larger influences will be spread across multiple, smaller features. This prevents overfitting to any singular feature and is particularly advantageous when features are correlated with each other. The loss function with the L2 regularization penalty is as follows:

$$\mathcal{L}_{\text{ridge}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

In practice, the L1 and L2 regularizations are often combined and referred to as Elastic Net regularization. It is written as follows:

$$\mathcal{L}_{\text{elastic}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2\right]$$

where $\alpha \in [0, 1]$ controls the balance between L1 and L2 regularization in the Elastic net. Thus, both the L1 and L2 regularization terms are incorporated into the loss function during training, retaining both of their advantages. The L1 term confers sparsity and feature selection, while the L2 term prevents any one feature from dominating.

2.5.2 Random Forest Regression

A random forest typically uses a collection of decision-making trees to make predictions accurately and consistently, as first created by Breiman, 2001. Because multiple trees are used, it is called ensemble learning. Each tree is trained on a subset of the data, which is created by a process known as Bootstrap sampling. Rows are sampled with replacement, as are features, and thereby introducing the randomness aspect of random forests. This also ensures a low correlation between the trees, allowing them to produce different results and prevents overfitting to the training data. Once each tree is trained, their outputs will be averaged together to yield the final prediction in a process known as aggregation.

In lieu of a global loss function, each decision tree is self-contained. How each decision tree decides to split the data at each node is a rigorous process in which some metric is minimized, such as Gini importance and mean decrease in impurity (MDI), or mean square error (MSE). The Gini impurity score can be defined as:

$$G = \sum_{i=1}^{C} p_i(1 - p_i) = 1 - \sum_{i=1}^{C} p_i^2 \tag{11}$$

where p_i is the proportion of samples belonging to class i at a given node, and C is the number of classes. For more information on any specific algorithm, please refer to Breiman, 2001.

The random forest model has many benefits, like a general high accuracy with predictions and a robustness against overfitting. Most importantly, it handles non-linearity well, unlike linear regression. However, its drawbacks lie in its long computation times and low interpretability.

2.5.3 K-Nearest Neighbors Regression

K-Nearest Neighbors (KNN) regression makes predictions based on the similarity between input data points. Specifically, it estimates the target value of a new data point by averaging the values of its k closest neighbors in the training dataset, where closeness is typically determined using Euclidean distance, although Manhattan distance is possible as well. The formula for Euclidean distance is

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

where $\hat{y}$ is the predicted value for the test data point, k is the number of nearest neighbors, and y_i represents the target values of the k nearest neighbors to the test point.

This straightforward approach makes KNN particularly well-suited for problems where the data distribution is irregular or nonlinear. Moreover, KNN regression is highly intuitive and interpretable: predictions are based on actual observed data points, and the influence of nearby observations is straightforward to understand. Compared to other models, it is also relatively computationally light.

3 Data Description

3.1 Data Overview: Seven Distinct Datasets

Dataset	Rows	Columns	Size	Format	Purpose
TCEQ	12,188	3	331 KB	CSV or TSV	Sensor data
PurpleAir	74,636	3	1.3 MB	CSV or JSON	Sensor data
Point Source Emissions	364,889	34	54.5 MB	XLSX	Auxiliary variables
Traffic	40,579	35	7.5 MB	CSV or XLSX	Auxiliary variables
Weather	26,304	41	7.7 MB	CSV or JSON	Auxiliary variables
U.S. Census	6,896	12	681 KB	CSV or JSON	Auxiliary variable, TCEQ vs. PurpleAir analysis
Land Use	N/A	N/A	17.04 MB	GDB	Auxiliary variable

Table 1: Description of datasets utilized in this project. Used formats are underlined, and alternative formats are shown.

The table provides an overview of the datasets used for our analysis and model, detailing their size, format, and purpose. Our model seeks to interpolate $PM_{2.5}$ using sensors from regulatory (TCEQ) and non-regulatory (PurpleAir) data. Together, these two datasets constitute the "core dataset" containing measurements of the outcome variable, $PM_{2.5}$, across Houston. Meanwhile, Regression Kriging will utilize auxiliary variables such as point source emissions, traffic, and weather to train a trend regression model, capturing non-spatial factors in $PM_{2.5}$ concentration. These datasets together comprise an "auxiliary" dataset. In addition to providing a median neighborhood income variable for our auxiliary dataset, the U.S. Census dataset is included for comparative analysis between TCEQ and PurpleAir data. All datasets are tabular in format, with the exception of land use, which we sourced as a Geodatabase (GDB) spatial file from the Houston-Galveston Area Council. This comprehensive data integration supports a more detailed examination of air pollution patterns and contributing factors, as well as demographic analysis of monitor placement and pollution trends.

3.2 TCEQ Data

TCEQ is the state agency responsible for regulating environmental quality in Texas. TCEQ oversees a wide range of environmental issues, including air quality, water quality, waste management, and more, ensuring that Texas complies with both state and federal environmental laws and regulations. TCEQ owns several air quality sensors located across the state to monitor air quality and ensure regulatory compliance. Hourly readings from TCEQ's sensors, along with data from other smaller contributors, can be accessed through the Texas Air Monitoring Information System (TAMIS) database via the TAMISWeb portal at: https://www17.tceq.texas.gov/tamis/index.cfm?fuseaction=home.welcome. The full list of sensors accessible via TAMISWeb can also be found at: https://www17.tceq.texas.gov/tamis/index.cfm?fuseaction=report.site_list

We acquired TCEQ data through a Raw Data Report (AQS) and fetched hourly reading data points, from January 1, 2022, to December 31, 2024, with the target list of "$PM_{2.5}$ Parameters." All sensors in the Greater Houston Metropolitan Area with available $PM_{2.5}$ readings, specifically from Brazoria, Galveston, Harris, and Montgomery counties, were included.

Within this initial dataset, there are two distinct parameters present for reporting $PM_{2.5}$: "$PM_{2.5}$ (Local Conditions)" and "$PM_{2.5}$ (Local Conditions Acceptable)". Both parameters are reported in units of ($\mu g/m^3$), and present near real-time measurements of particulate matter.

The difference between the two parameters is that "$PM_{2.5}$ (Local Conditions) Acceptable" readings aren't taken into account for regulatory purposes, due to differences in the ways they were calculated. Nevertheless, both parameters represent valid measurements.

We altered the initial TCEQ dataset, averaging out the hourly readings for every day, taking both "$PM_{2.5}$ (Local Conditions)" and "$PM_{2.5}$ (Local Conditions) Acceptable" into account. The sole columns retained were the site identifier, the mean of $PM_{2.5}$ readings ($\mu g/m^3$), and the date of the readings. As a result, we were left with a final TCEQ dataset of 12,188 rows and 3 columns in the format of a comma-separated value (CSV) file.

3.3 PurpleAir Data

PurpleAir is a company that manufactures and sells air quality sensors. Purchasers may opt to contribute their sensor recordings, registering their sensors as "public." Otherwise, the sensor is registered as "private" and accessible only to the owner. Each sensor is assigned a unique "sensor index," which serves as an identifier for querying. PurpleAir provides an application programming interface (API) that allows both purchasers and non-purchasers to query historical data from each public sensor.

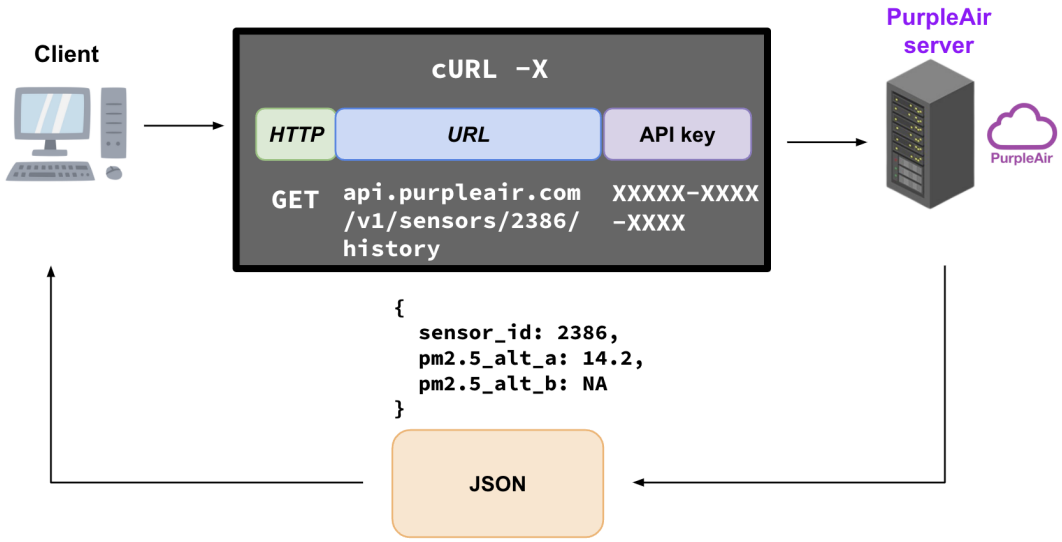

Figure 3: Schematic of PurpleAir's REST API architecture. A client requests a sensor's history, which is a resource (noun), by utilizing an HTTP command (verb). The server then provides this in the form of a JSON in exchange for a point deduction.

The PurpleAir API follows a Representational State Transfer (REST) architecture (Figure 3). Under this paradigm, resources ("nouns")–such as sensors and their historical data–can be requested through HTTP commands ("verbs") (Fielding, n.d.). The results of all queries are provided in a JavaScript Object Notation (JSON) format, though a CSV format can also be specified. After creating an account, users can generate an API read key (to retrieve data) and API write key (to contribute data or edit sensors). By using the cURL command line tool and the HTTP GET command to query historical data, users can obtain $PM_{2.5}$ data for a given sensor (identified by the sensor index) over a specified timeframe and resolution. The full API documentation is available at https://api.purpleair.com/. A limitation, however, lies in the number of queries an individual can make. Each query deducts "points" from the owner of the user-specific API read key, with the cost varying based on the number of records and fields requested. Each user starts with 1,000,000 "points" and may purchase additional points as needed. For this reason, the Houston Chronicle provided us with $PM_{2.5}$ time-series data at a resolution of daily averages, stretching from January 1, 2022, to December 31, 2024. This data was obtained for all public sensors within the latitude range of (28.579°, 30.819°) and longitude range of (-96.755°, -93.898°), totaling 75 sensors.

The only fields we selected from the PurpleAir API, in addition to the timestamp, were "pm2.5_alt_a" and "pm2.5_alt_b." The two recordings result from the ability of PurpleAir sensors to make two simultaneous readings of $PM_{2.5}$ concentration. However, as recommended by our sponsors, we excluded records in which the two readings differed by a factor of 70% or greater from the dataset, indicating low-quality readings. This resulted in the exclusion of 5,939 records. For records that were retained, we took the average between "pm2.5_alt_a" and "pm2.5_alt_b" as the value for $PM_{2.5}$. This resulted in a dataset of a size of 74,636 records and 3 columns (the sensor index, the date, and the $PM_{2.5}$ reading). Non-selected variables included meteorological data (e.g., pressure, humidity, temperature) and data on other pollutants (e.g., $PM_{1.0}$, $PM_{10.0}$)).

It is worthwhile to note that the $PM_{2.5}$ concentrations that PurpleAir sensors report have undergone an internal calibration process. In this process, PurpleAir sensors multiply the reading recorded by their Plantower laser detectors by a factor of 3.0 to calibrate it to the standards of regulatory agencies such as the EPA and TCEQ. However, a study conducted by industry expert Lance Wallace found that this calibration factor is still insufficient (Wallace, 2022). Instead, he recommends multiplying the raw recording by a calibration factor of 3.4 rather than 3.0. Because the raw recording is not directly retrievable by the PurpleAir API, we multiplied the reported $PM_{2.5}$ concentrations (which had been improperly calibrated by a factor of 3) by a factor of 3.4/3 in order to make the PurpleAir and TCEQ recordings comparable.

3.4 TCEQ Point Source Emissions Inventory

Each year, TCEQ conducts a survey of the state's chemical plants, refineries, electric utility plants, and other industrial sites that meet reporting criteria. This survey includes polluter identifiers such as company name, site, and account number; location features such as city, ZIP code, TCEQ region, and county; and concentrations of pollutants such as nitric oxides, $PM_{2.5}$, and $PM_{10.0}$. We will use this dataset to identify known air polluters in the Houston area under the assumption that air quality will worsen around these sites. Unlike other datasets that track only facilities that have been granted federal permits to emit pollution, this dataset provides records of actual polluters with site-specific pollution levels.

The point source emissions inventory dataset we used is the most recently available (2022) and can be downloaded as an XLSX file at https://www.tceq.texas.gov/airquality/point-source-ei. Although we lack precise emissions data for 2023 and 2024, the opening of new large industrial facilities and significant changes in polluter activity are rare. Therefore, the 2022 report is assumed to be a sufficient representation of polluters in Texas for dates after 2022.

3.5 TDOT Traffic Data

The Texas Department of Transportation (TDOT) collects traffic data from locations throughout Texas. The dataset is publicly available and can be downloaded as a CSV or XLSX file at https://gis-txdot.opendata.arcgis.com. Average Annual Daily Traffic (AADT) data, measured in vehicles per day (VPD), are collected at each traffic station from roadway traffic sensors that track every car passing over said sensor. Most sensors collect traffic count data from one 24 hour period and average the whole year based on that finding which is calculated using this formula:

$$AADT = V_{24} \times F_A \times F_S \tag{12}$$

where V_{24} refers to the "24-hour volume," which is the raw count of vehicles recorded over 24 hours. F_A is an axle correction factor applied to sensors that track the number of axles instead of the number of vehicles. Lastly, F_S is a monthly adjustment factor to account for seasonality. Instead of using this, we decided to filter for only stations with sensors that track each day year round, or permanent stations. The permanent traffic counting machines and data are maintained by TDOT TPP Traffic Section. The usage of permanent sensors allowed for daily traffic values which was important to our model as $PM_{2.5}$ has large fluctuations each day. The dataset includes traffic counts from 464 permanent stations throughout Texas with 24 active stations in the Greater Houston Area. Notable features of the dataset include traffic station ID, county/district information, and VPD values.

3.6 Visual Crossing Weather Data

Visual Crossing is a weather data provider that offers historical weather data reports. The platform provides daily or hourly weather data for any location in the United States. Notable features of the dataset include temperature (e.g., maximum, minimum, mean, "feels like" temperature), precipitation (e.g., "in inches," chance, type), humidity, wind (e.g., speed, gust, direction), cloud cover, and UV index (Figure 4). Additional features include visibility, sunrise and sunset times, and a text description of weather conditions. Users can query data from https://www.visualcrossing.com/weather-query-builder/ as a CSV file, or utilize a REST API in ways similar to those described in Section 3.1.

We queried and downloaded the daily weather data recorded by each of Visual Crossing's 24 weather stations located in the Houston area (including the KHOU and KIAH Houston news stations) from January 1, 2022 to December 31, 2024 (1,096 total days), totaling 26,304 observations.

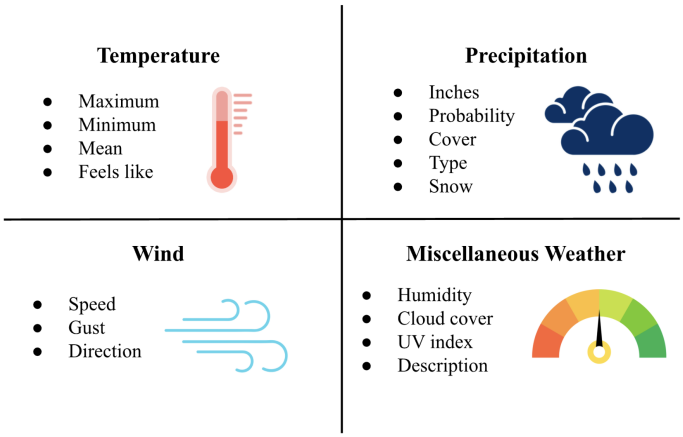

Figure 4: Important features from Visual Crossing weather dataset. This includes categories like temperature, precipitation, wind, and miscellaneous descriptors.

3.7 American Community Service US Census Data

The American Community Service (ACS) provides detailed demographic Census data for all geographies in the United States, including states, counties, congressional districts, tracts, and block groups. There were over 27,000 variables in the ACS survey database for 2020. For this project, several demographic variables were selected with guidance from the Houston Chronicle to explore whether any of these variables exhibit patterns in relation to TCEQ monitor placement and $PM_{2.5}$ pollution. Currently, the data extracted from the ACS for each Census tract using an REST API includes median age, median household income, racial demographic breakdowns, and household size. Specifically, we utilize median income in our regression kriging model. Documentation to access the API can be found at this link: https://www.census.gov/data/developers/data-sets.html.

3.8 H-GAC Land Use Data

The Houston-Galveston Area Council (H-GAC) is a voluntary association of local governments in the Gulf Coast Planning Region of Texas, largely concerned with community planning and economic development. Since 2002, H-GAC has intermittently developed and released in-house land use datasets categorizing how humans have utilized land across the Houston-Galveston area. This includes construction intensity levels for developed areas, as well as classifications of undeveloped natural land such as wetlands, pastures, and forest. H-GAC utilizes this data inform analysis on community and environmental planning, but we use the most recent land use data from 2022

to understand how human development and different natural environments influence the $PM_{2.5}$ concentrations in the surrounding atmosphere.

The 2022 data can be downloaded as a Geodatabase raster file at , where the technical document and metadata can also be found. The area covered includes the 13-county region of H-GAC, as well as San Jacinto and Grimes Counties—roughly 12,500 square miles. Contrary to the other auxiliary datasets we use, the land use data is not tabular, but geospatial. We document our process for extracting land use values across the Houston Metropolitan Area in Section 4.1.5.

4 Data Exploration and Modeling

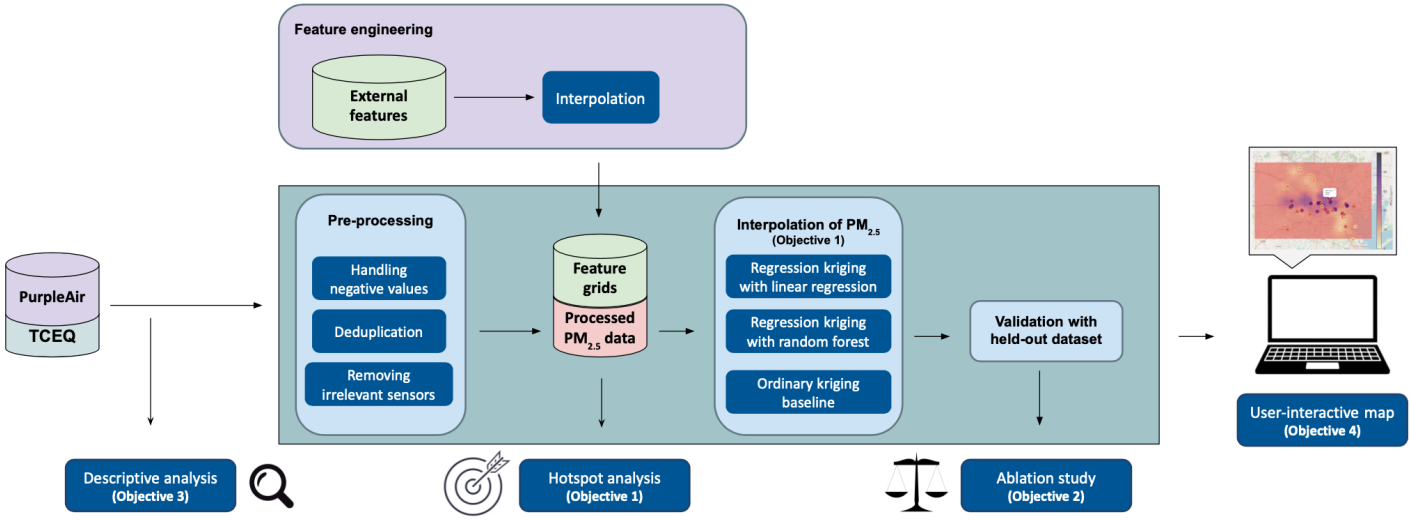

Figure 5: Overview of our data science pipeline.

A summary of our data science pipeline is displayed in Figure 5. We first preprocess the core datasets (TCEQ and PurpleAir $PM_{2.5}$ data) (Section 4.1.1) and each of the auxiliary datasets (Sections 4.1.2 - 4.1.5). Next, we perform exploratory data analysis on each of the datasets (Sections 4.2.1 - 4.2.5) to further understand their structures, as well as achieve Objective 3 (to identify major Houston polluters). Meanwhile, we apply feature engineering on the auxiliary datasets by the methods described in Section 4.1.6 to obtain structures known as feature grids. We then interrogate the processed core dataset through hotspot analysis (Section 5.6) to identify areas that experience disproportionately high or low values of $PM_{2.5}$ (Objective 1). After feature selection (Section 5.6), the processed core dataset, along with the auxiliary feature grids, undergoes modeling by the interpolation methods described in Section 5.2. We then conduct an ablation study using the trained model (Section 5.7) to examine the importance of individual regulatory and non-regulatory sensors as part of Objective 2. Lastly, we use the interpolation models to build a user-interactive map for Houston residents (Objective 4).

4.1 Data Wrangling

4.1.1 Combined $PM_{2.5}$ Dataset

Both the TCEQ and PurpleAir datasets were structured with three features (i.e., the sensor index, timestamp, and $PM_{2.5}$ value). Therefore, we employed a vertical join to combine these datasets. The resulting shape of the dataset was (87,322 rows, 3 columns), obtained from 87 sensors.

Of the 86,824 records remaining, 28,447 were identified as duplicates. As the dataset was provided by the Houston Chronicle, no explanation was available for the existence of these duplicates.

However, queries performed by us did not result in duplicates. After we dropped these records, the shape of the dataframe was (58,377 rows, 3 columns).

A limitation of both datasets was the absence of geographical coordinates for each sensor. While geographical coordinates for PurpleAir sensors could, in theory, be obtained through the PurpleAir API, this method proved unreliable in practice, with NA values often being returned. PurpleAir hosts a web page that provides a graphical user interface available as a map, allowing users to track real-time recordings of $PM_{2.5}$. From this map, latitude and longitude can be manually identified for each sensor by clicking on the sensor and checking the url (e.g., map.purpleair.com/.../**29.7032/-95.246629**). Similarly, the TCEQ website provides coordinates for each TCEQ sensor. This information, along with the status of each sensor as "indoor" or "outdoor," was provided to us by the Houston Chronicle through a sensor information CSV file. We performed a left join between the combined dataset and the sensor information dataset. Despite this, we were unable to assign coordinates to five PurpleAir sensors, which we therefore dropped from downstream analysis. The resulting shape of the dataset was (60,048 rows, 7 columns), with 82 sensors remaining.

a)

site_id	timestamp	pm2.5_alt_a
24	2022-01-01	...
24	2022-01-02	...
...	...	...
46	2022-01-01	...
...	...	...

b)

site_id	indoor_outdoor	longitude	latitude
2386	outdoor	-95.07647	29.532282
27009	indoor	-95.585686	29.733050
...	...	...	...
24	outdoor	-95.326139	29.901027
46	outdoor	-95.283981	29.828516

c)

site_id	timestamp	pm2.5_alt_a	source	indoor_outdoor	longitude	latitude
2386	2022-01-01	...	PurpleAir	outdoor	-95.07647	29.532282
2386	2022-01-02	...	PurpleAir	outdoor	-95.07647	29.532282
...	...	...	...	...	...	...
24	2022-01-01	...	TCEQ	outdoor	-95.326139	29.901027
24	2022-01-02	...	TCEQ	outdoor	-95.326139	29.901027
...	...	...	TCEQ	...	...	...

Figure 6: Samples of combined $PM_{2.5}$ datasets, before and after merging with sensor-level information. a) Structure of combined PurpleAir and TCEQ $PM_{2.5}$ datasets. b) Structure of sensor information dataset provided by Houston Chronicle c) Structure of dataset after a left join between the $PM_{2.5}$ dataset and sensor information datasets.

Because the PurpleAir dataset included indoor sensors while the TCEQ dataset utilized exclusively outdoor sensors, we dropped indoor sensors from downstream analysis. Ten sensors were identified as indoors. This resulted in a data shape of (52,228 records, 7 rows), with 72 unique sensors.

We detected four records with negative $PM_{2.5}$ values, all sourced from the TCEQ dataset. Since negative measurements are typically recorded when the $PM_{2.5}$ value is extremely low, we imputed the $PM_{2.5}$ concentration as zero for these rows. The inclusion of outliers (such as those caused by fireworks on holidays like the Fourth of July) was advised by the Houston Chronicle.

Lastly, in order to ensure each row accounted for the temporality of our time-series data, we added two new columns to the combined $PM_{2.5}$ dataset. The first column represents the $PM_{2.5}$ reading taken at the same location one day before ("lag" of one day), and the second represents the reading taken seven days before ("lag" of one week). The use of a lag of one day has been

employed in the past in studies that sought to explore the effects of $PM_{2.5}$ concentrations on mortality (Staniswalis et al., 2009). We used a one-week lag, assuming that industrial sites exhibit similar emissions on different days of the week (i.e., a Monday behaves similarly to the previous Monday due to work schedules). The final dataset had a shape of (52,228 rows, 9 columns).

Data Cleaning Step	**Rows**	**Columns**
Base TCEQ $PM_{2.5}$ dataset	12,188	3
Base PurpleAir $PM_{2.5}$ dataset	74,636	3
Vertical join of $PM_{2.5}$ datasets	86,824	3
Left join between combined $PM_{2.5}$ dataset and sensor information dataset to add coordinates	87,322	7
Dropping duplicates	64,316	7
Dropping sensors that have no location or are outside of Houston	60,568	7
Dropping indoor sensors	52,716	7
Detecting records with negative $PM_{2.5}$, 4 records imputed as zero	52,716	7
Adding $PM_{2.5}$ reading columns with one-day and one-week lag	52,716	9

Table 2: Summary of data cleaning procedures applied to the $PM_{2.5}$ datasets.

4.1.2 Point Source Emissions Inventory Dataset

Datasets described in this section and onward (Sections 4.1.2 - 4.1.5) will be incorporated as auxiliary variables, external to the core $PM_{2.5}$ dataset. As a result, they were cleaned independently prior to incorporation.

The dataset of polluter characteristics was limited to features mentioning company identity, location, amount of pollution, and particulate class of the pollution. To that end, we removed unnecessary features used for record keeping such as the name, codes, and descriptions of the Source Classification Code (SCC), Facility Identification Number (FIN), Contaminant name, etc. "Annual Routine TPY" is a numerical feature that we retained, representing annual emissions from routine operations in tons per year.

Under the assumption that concentrations of $PM_{2.5}$ would worsen around the sites of pollution, we kept only records related to $PM_{2.5}$ emissions within or near Houston. Sites located outside of counties in the greater Houston area were removed, as the model was intended to be trained on and used to interpolate only within Houston. Finally, we removed records with zero Annual Routine TPY to exclude polluters who were permitted to pollute $PM_{2.5}$ but did not. After consolidating the different instances of pollution from the same registration number, the final dataset included 5,776 records of polluters producing $PM_{2.5}$ within Houston.

4.1.3 TDOT Traffic Dataset

The TDOT traffic dataset included AADT values dating back to 2004, along with several identification and location-based variables. To get permanent sensor information, we had to manually pull the dataset from TDOT's interactive interface called Statewide Traffic Analysis and Reporting System (STARS II) in which we filtered for active, permanent stations in Houston using the TCDS filtering system and exported the XLSX file using the report center. This dataset included the station ID, date, and VPD value for each hour of a given day. The dataset did not have missing values, although some sensors had days that were missing which led to fewer inputs for our interpolation models on that given day. To get the latitude and longitude information for the stations, we had to merge this dataset with the TDOT Permanent Count Stations dataset provided directly on the TDOT website. This dataset gave location based data for each permanent station in Texas. We merged on the Station name, and obtained the latitude and longitude data from the TDOT Permanent Count Stations dataset.

One complication was that the 2024 traffic data was not yet available. Therefore, to get traffic volumes that would support our final model, we trained a Random Forest Regression model using the interpolated traffic volumes from 2022 and 2023 along with other features such as day of the week, month, and weekend. The Random Forest Regression then predicted daily 2024 traffic volume for each $PM_{2.5}$ sensor in 2024 which allowed for a clean modeling process despite the lack of 2024 traffic data provided by TDOT.

Lastly, we were informed by our sponsors that we should consider secondary $PM_{2.5}$ production, in which $PM_{2.5}$ particles form indirectly through chemical reactions between the atmosphere and substances from vehicle emissions. To account for this, we lagged our traffic values by 2 days (which was the time estimated by our sponsors in which secondary $PM_{2.5}$ would form from car emissions).

4.1.4 Visual Crossing Weather Dataset

We selected the features available in the Visual Crossing weather dataset that were related to air quality, such as inches of precipitation, wind speed, and wind direction. Features that were constant across all dates were removed: name (Houston, TX for all rows) and a severe weather risk indicator (none for all rows). We also removed features with textual information that were largely redundant (i.e., precipitation type, conditions, weather description, and weather icon). Finally, features that we deemed to be unrelated to air quality were removed (e.g., stations).

None of the remaining features contained any missing values, eliminating the need for missing value management or imputation. The available features were relatively rich, so the only features engineered were specific year, month, and day values extracted from the datetime column to facilitate analyses of varying temporal granularity. Finally, we adjusted the wind direction feature, which was originally represented in degrees from which the wind was coming from, relative to North, by adding 180 degrees so that it represented the direction the wind was blowing towards, relative to North. The final dataset contained 1,096 rows (one per day over three years) and 28 columns.

4.1.5 H-GAC Land Use Dataset

Given the land use data was not in a tabular format, we first converted the original GDB file into a Tagged Image File Format (TIFF), increasing the size of the file to 89.2 MB. This TIFF file was compatible with querying a land use value for each provided geographical point in Houston. As the data only includes one variable and has no missing data for our area of interest, the next step for wrangling was to reassign land use values from their original format as 2-digit numerical codes (e.g., 22, 41, 90) to textual definitions (e.g., "Developed, Low Intensity," "Deciduous Forest," "Woody Wetlands") as defined in H-GAC's technical document. Two land use codes, 0 and -128, were not defined in the technical document; therefore, they were retained as text representations of the numeric code ("0" and "-128"). Lastly, we one-hot encoded the land use variable, which refers to the process of transforming each distinct land use category into a column of its own with binary values: 0 indicating an observation does not belong to the category, or 1 indicating that it does. This encoding allows our regression models interpret categorical data as numerical inputs. We designated "Developed, Medium Intensity" as the reference category, meaning our regression models interpret the coefficients of other land use categories in comparison to this baseline.

4.1.6 Auxiliary Feature Grids

Recall the re-trending stage of Regression Kriging (Stage 5 in Figure 2). In this process, the trending model is applied to every point in the grid of interpolation to obtain $PM_{2.5_{\text{trend}}}$ for all points of the grid. Therefore, every point in the grid needs to have a value for each auxiliary feature used to train the trend model. To address this issue, we use an approach of "nested interpolation." In this approach, the auxiliary feature at each point in the grid is first interpolated by either Inverse Distance Weighting (Section **??**) or Ordinary Kriging (Section 2). This yields a grid of values for each auxiliary feature (herein referred to as a "feature grids"). Utilizing the Visual Crossing weather dataset, we employed Ordinary Kriging to obtain feature grids for all available weather features, such as temperature, dew point, precipitation level, wind speed, and

atmospheric pressure. We also used Ordinary Kriging to obtain feature grids for the two lagged features. Meanwhile, Inverse Distance Weighting was employed for point source emissions and traffic volume (including lagged values), as recommended by previous work (Shi et al., 2020). Interpolation was not necessary to obtain a land use feature grid, for which H-GAC already publishes values for every geographical point in our grid of interest. The same applied for income. Lastly, a trivial feature grid, consisting of only the latitude of each point in the grid, was included to account for spatial dependence in the trend model. Using the feature grids, we assigned values for features that were not initially present to the sensors of the core $PM_{2.5}$ dataset (i.e., weather, traffic, emissions, land use) by taking the corresponding value of each feature at each sensor location.

4.2 Data Exploration

4.2.1 Distinct Air Quality Snapshots From Regulatory vs. Non-Regulatory Sensors

We examined the volume of $PM_{2.5}$ data in the combined $PM_{2.5}$ dataset by year (Figure 7). The number of data points appeared to increase over time, with 12,474 records in 2022 compared to 23,120 records in 2024. Only 7 sensors–all sourced from PurpleAir–had data for the entire time frame from January 1, 2022, to December 31, 2024. This may reflect the emergence of new sensors in the dataset that did not exist at earlier dates. Across all years, there were more PurpleAir recordings than TCEQ recordings, which was expected. The observation that sensors have the most "complete" set of recordings in 2024 further justifies our selection of 2024 as the year over which we perform interpolations of $PM_{2.5}$.

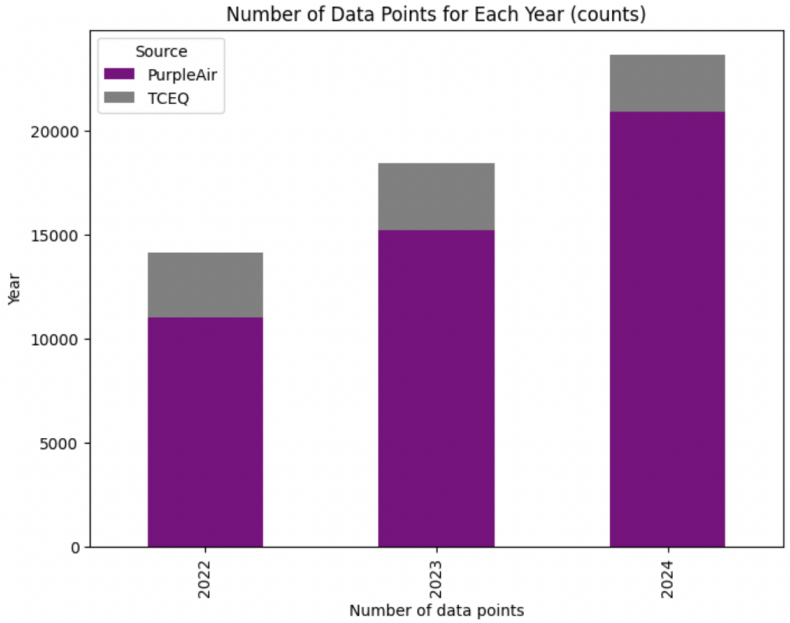

Figure 7: The number of $PM_{2.5}$ recordings increases with time. There were 12,474 records in 2022; 17,122 records in 2023; and 23,120 records in 2024. The bars are colored by the source of the data (either PurpleAir or TCEQ).

We also compared the $PM_{2.5}$ concentrations of TCEQ and PurpleAir sensors through a Student's t-test, which found that the mean $PM_{2.5}$ concentration of TCEQ (9.877 $\mu g/m^3$) was significantly greater than that of PurpleAir (7.623 $\mu g/m^3$) ($p < 0.001$), even after the calibration process described in 3.1. This is a relatively surprising finding: if the TCEQ is strategically placing its sensors in areas of low pollution, then we would expect the opposite result. One likely explanation of this finding is that many PurpleAir monitors are purchased by consumers for placement outside their homes, therefore these monitors more strongly tend to be located in affluent neighborhoods,

where pollution is not so severe. This explanation is supported by the significant cost of each PurpleAir monitor, ranging in the hundreds.

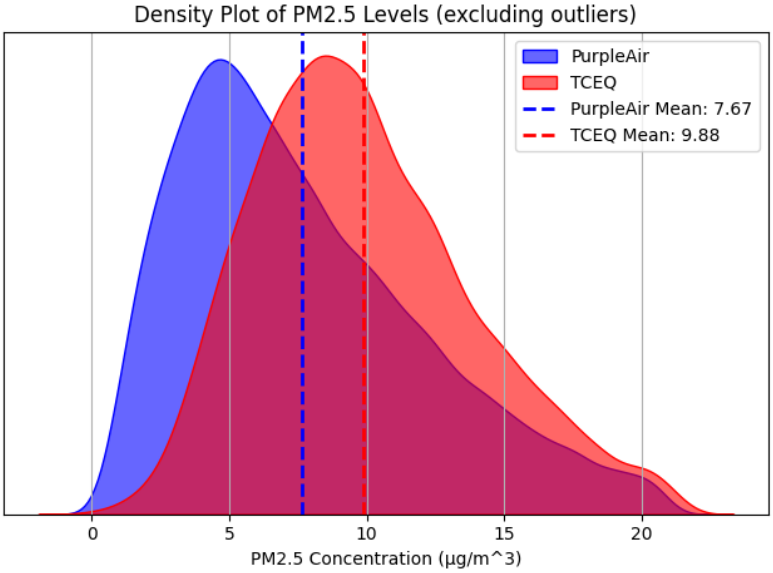

Figure 8: PurpleAir sensors, on average, read out lower $PM_{2.5}$ measurements than those of the TCEQ.

Histograms with kernel density estimate are shown for each subset of data. The vertical lines of each color correspond to the means of their respective distribution (blue for PurpleAir, red for TCEQ) (Figure 8).

The most- and least-polluted sensors were also surveyed. The degree of pollution of each sensor was then taken as the median daily $PM_{2.5}$ concentration over each sensor's lifetime. The most-polluted sensors (95th quantile) were spread among East Houston, Interstate 610, and Galena Park. Galena Park in particular, located along the Houston Ship Channel, has become an area of interest for many air quality advocates (Figure 9). They had a median daily $PM_{2.5}$ concentration of 9-10 $\mu g/m^3$. All but one was sourced from PurpleAir. Meanwhile, the least-polluted sensors (5th quantile) did not have a noticeable spatial pattern. All of these were PurpleAir sensors, with a median daily concentration of 2-3 $\mu g/m^3$. It should also be noted that the ZIP codes of the most-polluted sensors had a substantially lower median household income (an average of $44,743.66) compared to those of the least-polluted sensors (an average of $144,958.50), a finding that is supported by the exploration of Census data described in Section 4.2.5. We conduct a more in-depth hotspot analysis of neighborhoods consistently reporting clean or polluted air in Section 5.6.

4.2.2 Polluters Are Concentrated in Harris County along Buffalo Bayou

Upon preliminary analysis of the TCEQ-published Point Source Emissions Inventory, we found that a large amount of polluters were located in Harris County compared to any other Houston area (3,494 polluters in Harris County versus 2,287 polluters outside combined). A visualization of the distribution of polluters revealed that polluters in Harris County clustered around Buffalo Bayou and the Houston Ship Channel (Figure 10). It should be acknowledged that each entry in the dataset represents an instance of pollution at a given industrial site, meaning that a company may have many instances of pollution listed. These instances were averaged to yield one value for a given industrial site.

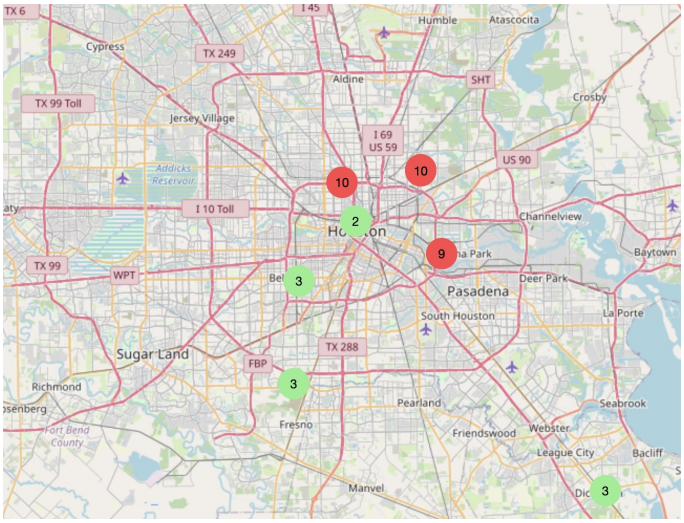

Figure 9: The most-polluted sensors are located near Interstate 610, Galena Park, and East Houston. Red markers correspond to the sensors with a median daily $PM_{2.5}$ concentration recording within the 95th percentile of sensors. Of the four sensors (two are placed in the same place, causing overlapping), only one was from the TCEQ. Green markers correspond to those with a median concentration within the 5th percentile of sensors. All sensors were from PurpleAir. The number within each marker is the median daily $PM_{2.5}$ concentration.

4.2.3 Traffic Volume Follows a Consistent Weekly Pattern

When comparing the weekend average of the 2022 and 2023 traffic volume datasets to the weekday average, we found that the weekday traffic volume had an average traffic volume of 3,303 VPD, higher than the weekend's 2,955. We also examined whether the hourly traffic volume differed from weekends and weekdays, as this difference would justify using a binary weekday value to help predict the 2024 traffic volume. A noticeable difference in hourly traffic trends was observed when comparing weekend and weekday traffic volumes. (Figure 11).

After running the Random Forest Regression to impute 2024 traffic, we sought to visually compare traffic volumes across different years to assess whether the general trend over time was similar. A line chart comparing the traffic volumes by year shows that the predicted 2024 traffic volumes closely resemble the trends of the actual 2022 and 2023 volumes. This indicates that our predicted values align well with the historical data (Figure 12).

4.2.4 High Variation in Weather Over Time

While analyzing daily weather summary data representing the entire Houston area over the three-year period of our analysis, we found that many features displayed significant changes over time, with warmer months (approximately May to October) exhibiting distinct weather patterns compared to the cooler months of November through April (e.g., much greater precipitation). The significant variation in feature values supports the inclusion of weather as a feature in interpolating air pollution. For instance, we examined patterns of wind speed and direction over the timespan of the dataset. The mean wind speed was 17.91 mph, with a standard deviation of 5.06 mph. Wind direction had a large standard deviation of 86.85 degrees (out of a 360 degree circle), revealing high variability in wind direction from day to day. A radial quiver plot illustrates that Houston winds tend to consistently follow particular directions—toward the northeast from November to January, and otherwise toward the southeast for the majority of the year (Figure 13). While wind speed and direction may vary greatly between observations, the presence of prevailing wind directions, rather than purely random wind patterns, suggests that air pollution may drift predominantly towards certain sections of the Houston area, in accordance with past studies indicating that wind can carry, dilute, and diffuse $PM_{2.5}$ (Weiwei et al., 2011; Gu et al., 2015). At the same time, these works argue that this effect on air quality can vary greatly by wind direction in a given locale.

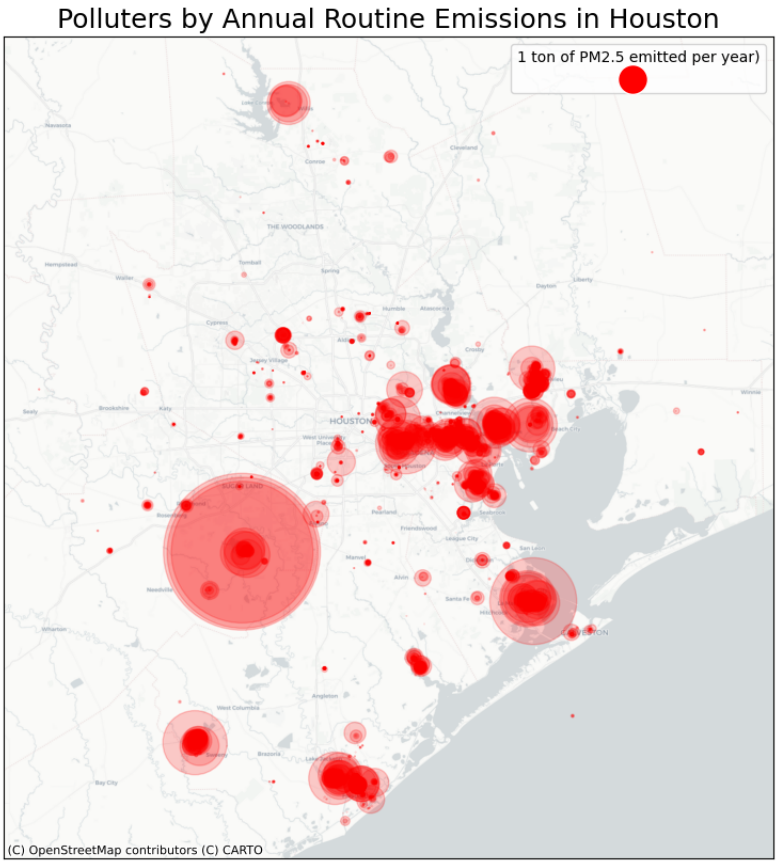

Figure 10: Harris County polluters cluster around Buffalo Bayou. A map view of active $PM_{2.5}$ polluters within the Greater Houston Area in the TCEQ Point Source Emissions dataset is shown above. Each point represents a registered instance of pollution while its size reflects its reported annual routine emissions.

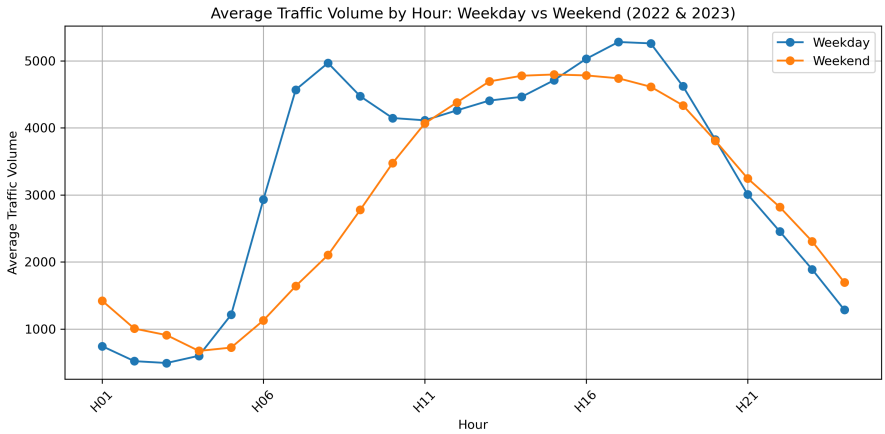

Figure 11: Average traffic volume by hour comparing weekdays (blue) and weekends (orange) for 2022 and 2023. We can see a sharp contrast in hourly traffic volume as weekday traffic peaked at hour 16–18, which are rush hours. Weekend traffic, meanwhile, had a more gradual increase throughout the day.

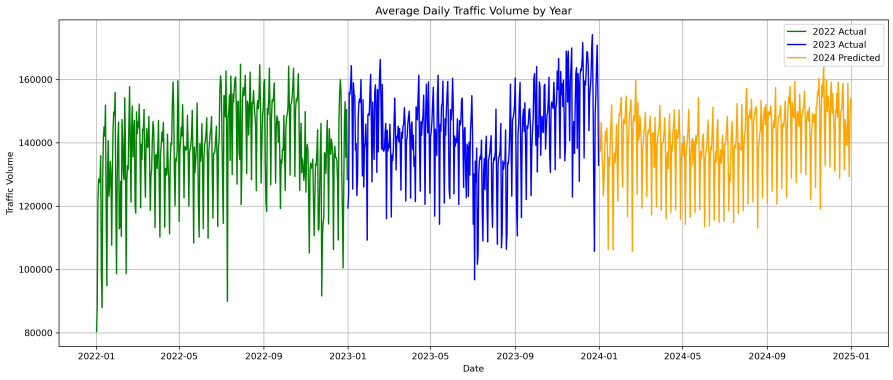

Figure 12: Average daily traffic volume across the years 2022 (green), 2023 (blue), and 2024 (orange). Traffic data for 2022 and 2023 reflect actual measured volumes, while 2024 values are predicted using a random forest model.

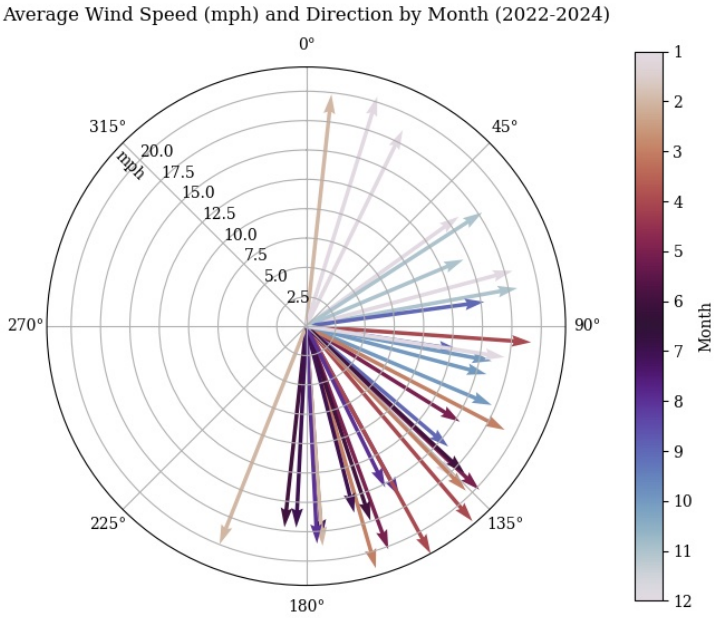

Figure 13: Houston winds tend to blow northeast from November to January (light blue to orange), and southeast for all other months (red to purple). A radial quiver plot of monthly average wind speed and direction from 2022 to 2024 is shown above. The arrows are colored by month, and the length of each arrow corresponds to the wind speed (as defined by the key at angle 315º). The direction of each arrow corresponds to wind direction.

4.2.5 Low-Income Households Are More Vulnerable to Air Pollution

Using the US Census API, we extracted several demographic features believed to be correlated with $PM_{2.5}$ and the distance from the closest TCEQ monitor. Currently, this includes demographic data such as race, average household size, median income, and median age. The Houston Chronicle has also advised us to include variables from the Social Vulnerability Index (e.g., housing type, English language proficiency) (CDC, 2024).

In Figure 14, each $PM_{2.5}$ emitter within the Houston Metropolitan area from the point source emissions dataset is plotted. The x-axis represents the amount of $PM_{2.5}$ emissions released yearly, and the y-axis represents their distance from the closest TCEQ monitor. A statistically significant difference was found in the average distance to the closest TCEQ monitor between the top 20 $PM_{2.5}$ polluters in Houston and the rest of Houston polluters. Sites with the 20 highest an-

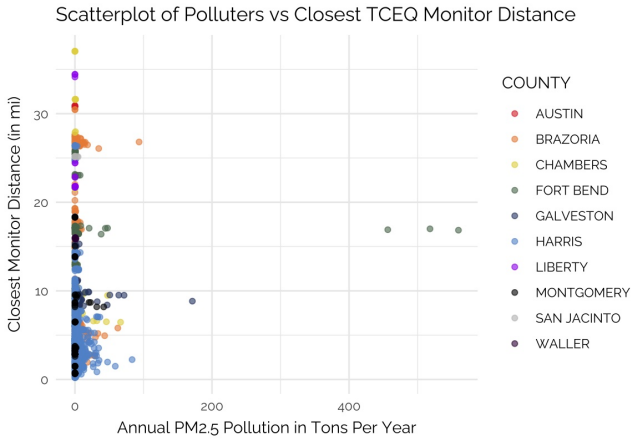

Figure 14: The largest $PM_{2.5}$ plant, WA Parish, depicted as the 3 rightmost green dots is 16 miles away from any TCEQ monitor. The y-axis represents the distance (miles) from the closest TCEQ monitor, while the x-axis represents the amount of $PM_{2.5}$ emissions per year (tons).

nual $PM_{2.5}$ emissions were, on average, located approximately 41.99 miles away from the closest TCEQ $PM_{2.5}$ monitors, while the remaining sites were located only 20.77 miles away ($p = 0.004$). This preliminary analysis may support concerns that TCEQ monitors do not adequately capture areas where the most particulate matter is being released into the atmosphere and are instead strategically placed to yield more favorable results. Alternatively, it could be that the polluters are strategically selecting locations away from the sensors.

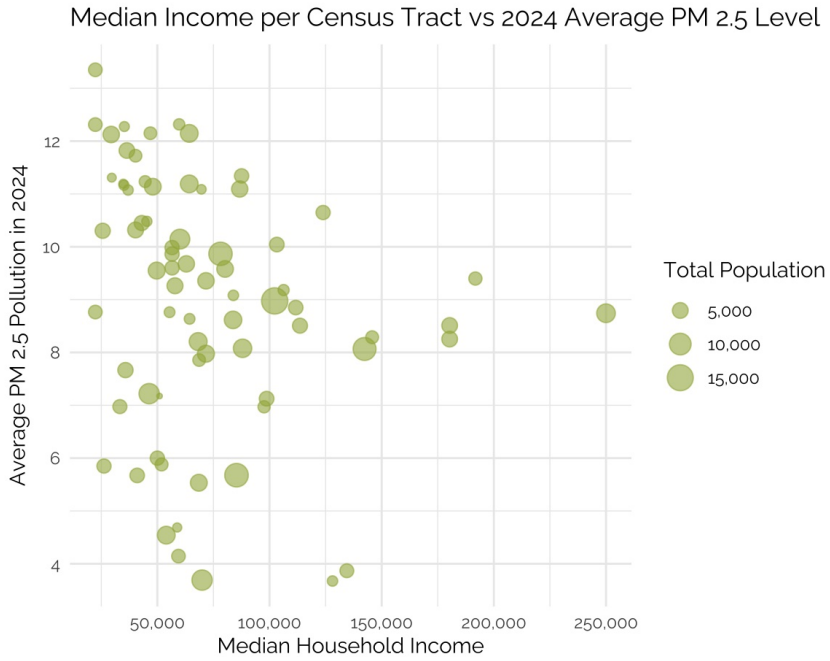

Figure 15: Wealthier households live in areas with less particular matter. A scatterplot showing average annual $PM_{2.5}$ pollution from 2024 in tons (y-axis) for every Census tract versus median household income with a PM 2.5 monitor (x-axis) in the Houston Metropolitan Area is shown above.

Additionally, a roughly negative correlation was observed between median household income per Census tract and average $PM_{2.5}$ pollution for those Census tracts that were in the same ZIP

code as TCEQ monitors. This may suggest that wealthier households live in areas less exposed to particulate matter—and, by extension, that poorer households are more vulnerable.

4.2.6 Houston is a Center of Intense Land Development

We uploaded the TIFF land use file into QGIS geospatial visualization software to perform a preliminary visual inspection of Houston's land use classifications on a map. Although much of the land surveyed by H-GAC constitutes pastures, forests, and cultivated crops, almost no land in the urban core of the Houston Metropolitan Area is undeveloped. Figure 16 highlights the highest intensities of development in the Houston downtown area, around major roads and highways, and along the Houston Ship Channel, with only scattered pockets of forests and woody wetlands.

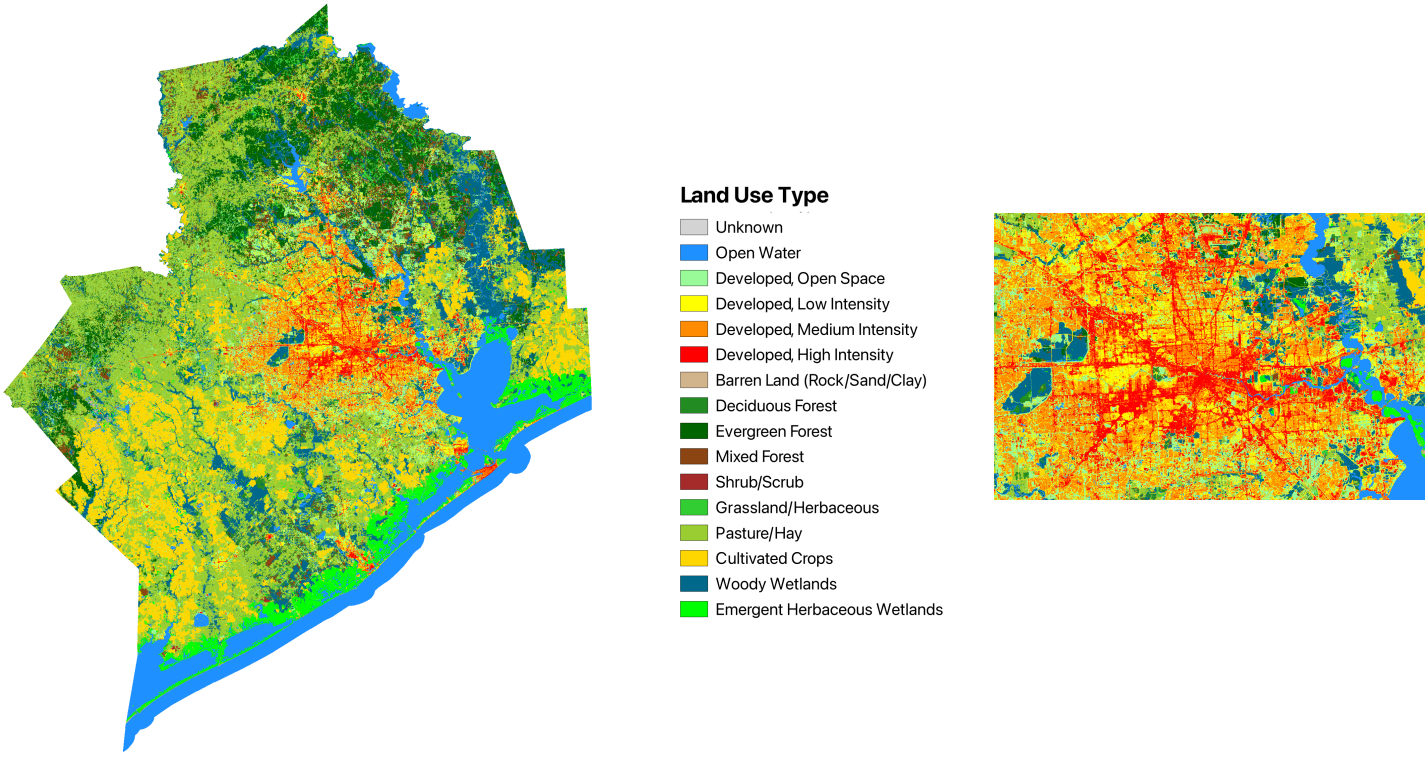

Figure 16: Land use is diverse across H-GAC's 13-county map, but the Houston Metropolitan Area is almost entirely developed. In particular, land highlighted in red represented high-intensity developed areas, largely concentrated in downtown Houston, along major roads, and surrounding the Houston Ship Channel.

In order to hypothesize whether more industrialized areas are at greater risk for air pollution, we determined the average $PM_{2.5}$ recording for each land use type. Of the 15 classifications provided by H-GAC, $PM_{2.5}$ sensors were only placed in 9: wetlands, pasture/hay, developed areas, open water, and cultivated crops (see Figure 17). Contrary to our expectation for developed areas to report the highest average concentrations of $PM_{2.5}$, woody wetlands (10.25 $\mu g/m^3$) and pasture/hay areas (9.89 $\mu g/m^3$) had the greatest $PM_{2.5}$ recordings, potentially as a product of proneness to fires and high-pollution agricultural practices like livestock farming, respectively. Even so, more intensely developed land appeared to exhibit greater air pollution than less developed areas—sensors in areas labeled "Developed, High Intensity" recorded an average $PM_{2.5}$ concentration of 9.72 $\mu g/m^3$, while those in "Developed, Low Intensity" areas averaged 8.31 $\mu g/m^3$. In conjunction with the last finding that Houston areas surrounding high vehicular traffic (including the Houston Ship Channel) are likely to fall in the "Developed, High Intensity" classi-

fication, descriptive analysis suggests that these neighborhoods are at higher risk for heightened air pollution.

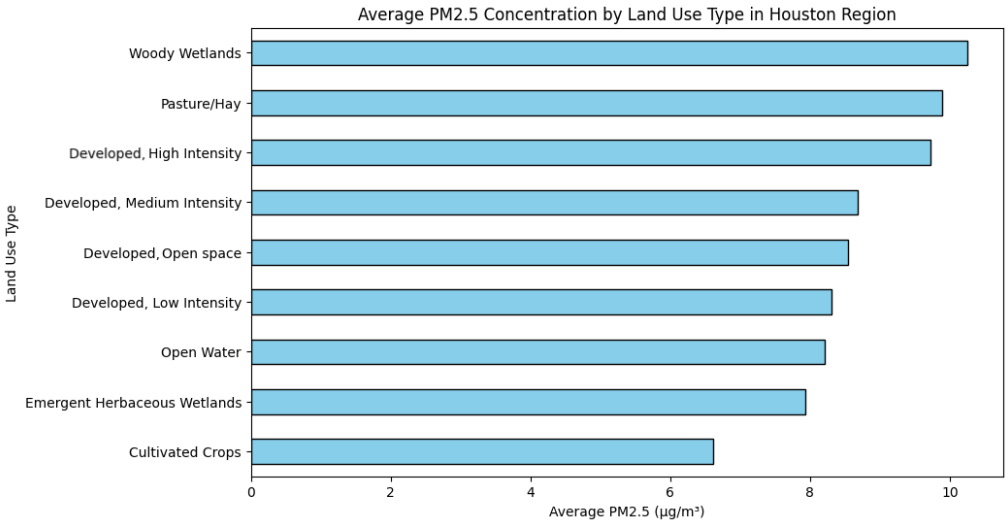

Figure 17: $PM_{2.5}$ sensors in woody wetlands and pasture/hay areas reported the highest average $PM_{2.5}$ concentrations. Still, sensors on land with high-intensity development record greater air pollution than those on land with lower-intensity development.

4.3 Modeling

The ideal interpolation model for our project would meet the following properties:

1. **Lightweight and scalable:** Since an interpolation model will be established for each day, it is crucial that the training and testing of the model are efficient.
2. **Interpretability:** The Houston Chronicle's mission is to educate the public not only on the levels of $PM_{2.5}$ in a local area, but also on what factors influence these levels on a day-to-day basis.
3. **Accuracy:** The model should be able to predict each day with similar accuracy (minimal variance). Additionally, the model should produce results within an acceptable error (minimal bias).

Table 3 summarizes the scalability, interpretability, and accuracy of each model that will be discussed in this section.

While past studies have utilized deterministic interpolation models such as Inverse Distance Weighting to interpolate pollutant concentration (Contreras and Ferri, 2016), some scholars express concerns about their use. One such criticism is the fine-tuning of parameters, such as the Inverse Distance Weighting exponent, a process that is often ad hoc and provides little room for interpretability (De Mesnard, 2013). This arbitrary nature results in a severe lack of interpretability. Furthermore, Inverse Distance Weighting has been observed to be generally less accurate than geostatistical interpolation models in the context of airborne pollution (Shukla et al., 2020). Nevertheless, their use in interpolating other auxiliary features, particularly traffic and polluter emissions (Shi et al., 2020), is still recommended.

Meanwhile, past studies that utilized geostatistical interpolation models to predict pollutants like $PM_{2.5}$ found success. Many of these studies have utilized Ordinary Kriging as a baseline model to compare against more sophisticated models (Janssen et al., 2008). Ordinary Kriging is particularly attractive due to its inherent ability to capture spatial autocorrelation and its lightweight nature, which is an important consideration for the back-end development of a user-interactive web page. For this reason, we have selected Ordinary Kriging as a baseline model.

However, Ordinary Kriging faces a critical limitation in its lack of complexity. If we were to analogize the interpolation of $PM_{2.5_{\text{trend}}}$ to a regression problem, then Ordinary Kriging, which utilizes only longitude and latitude as "features," would encounter the issue of underfitting. It is unable to incorporate auxiliary features, such as weather, traffic volume, and point emissions. This limitation arises from its assumption of first-order stationarity, namely that there is no deterministic process or large-scale trend over the Houston area (i.e., $PM_{2.5_{\text{trend}}}$ is a constant μ). Given the heterogeneity in the distribution of polluters and traffic as described in Sections 4.2.2 and 4.2.3, this assumption is unrealistic.

Regression Kriging, in contrast, accounts for large-scale trends by modeling $PM_{2.5_{\text{trend}}}$ as a regression model trained on non-spatial features. Furthermore, the use of a regression model like linear regression offers the added benefit of interpretability. A major challenge of Regression Kriging, however, is the process of re-trending, in which the trend model is applied on every part of the specified grid using a panel of feature grids. The use of entire feature grids for each auxiliary feature for each day in 2024, in theory, imposes a burden on system memory.

	Scalability	Interpretability	Accuracy
Inverse distance weighting	**Very high**	**Very low**	**Low - Moderate**
Ordinary Kriging	**High**	**Very low**	**Moderate**
Regression Kriging	**Moderate**	**High**	**High**

Table 3: Summary of advantages and disadvantages of different models considered.

5 Experiment: Regression Kriging and Ordinary Kriging

We developed three kriging implementations to interpolate $PM_{2.5}$ for each day in 2024: Ordinary Kriging (OK) as a baseline, Regression Kriging with linear regression (RK-LR), and Regression Kriging with random forest regression (RK-RF). While linear regression is commonly utilized with Regression Kriging when modeling airborne pollutants (Shi et al., 2020), there is little previous work that utilizes a random forest trend model to capture non-linear relationships in this context.

5.1 Feature Selection

Collinear features may negatively impact our workflow, particularly for the linear regression trend model, by making coefficients less stable and interpretable. Therefore, we removed potentially collinear features by assessing the variance inflation factor (VIF) of each feature. The variance inflation factor for a given feature X_j can be described mathematically as:

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2} \tag{13}$$

$$R_j^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

Here, R_j^2 is the coefficient of determination obtained by regressing the feature X_j on all the other features, $\hat{y}_i$ is the predicted value of the observed feature value y_i, and $\bar{y}$ is the mean of the observed values. If R_j^2 is high (i.e. X_j is dependent on other features), then the resulting VIF will also be high. We removed all features with a VIF greater than 5. This resulted from potentially training on 39 features to 24 features. Collinear features included those related to temperature (minimum and maximum temperature, feels-like temperature, minimum and maximum feels-like temperature), those related to wind gust (mean and max wind speed), those related to cloud cover (solar radiation and energy, UV index), and those related to humidity (dew point).

In order to improve model performance, we also utilized feature selection to identify the non-colinear features that helped best improve the performance of each trend model. The set of 24 features examined included numerical meteorological features (i.e. cloud cover, temperature, humidity, pressure, wind gust, precipitation, moon phase), traffic AADT (with lagged values up to 2 days), $PM_{2.5}$ emissions from pollutants from the auxiliary dataset, land use, income, the previous day's and previous week's $PM_{2.5}$ concentrations, and latitude.

Backward elimination is part of a larger family of feature selection algorithms known as stepwise regression models and is the preferred method for feature selection for Regression Kriging (Hengl et al., 2004). The method begins by using all inputted features to train the trend model. The performance is then assessed using cross-validation, which is used as a baseline. The algorithm then trains another model, but this time with one of the features omitted. The performance is reassessed and compared to the baseline performance. This process is systematically repeated for each subset of features, until one subset is determined as the most important. A disadvantage of this process is the computational expense of sampling each subset of features. We utilized backward elimination to decide the best features for the linear regression trend model in RK-LR and the random forest regression trend model in RK-RF.

5.2 Training and Interpolating Models

For the baseline models, we only retained relevant columns (i.e., the sensor's index, longitude, latitude, $PM_{2.5}$ reading, and timestamp). For each day in 2024, an interpolation set (80 percent of sensors) was used to train the OK model. After training, we used a separate held-out validation set of sensors (20 percent of sensors) as a ground truth to compare the interpolated $PM_{2.5}$ at their locations.

Meanwhile, we developed the Regression Kriging models by the workflow outlined in Section 2 and summarized in Figure 2. We used a two-year subset spanning 01/01/2022 to 12/31/2023 to train the trend regression model. This subset of data consisted of $PM2.5$ readings accompanied by values for all selected features, including several numerical weather features and $PM_{2.5}$ readings lagged by one day and one week. Latitude was also selected as important. The 2024 subset was handled in the same manner as the OK model, with each day's data being split into an interpolation set and a validation set. The trend model was then applied onto the interpolation set to obtain $PM_{2.5_{\text{trend}}}$ for each sensor. The residuals were obtained by subtracting $PM_{2.5_{\text{trend}}}$ from the observed $PM_{2.5}$ values of each sensor. An Ordinary Kriging model was then applied to these residuals to obtain $PM_{2.5_{\text{SA}}}$ over the grid of interpolation. Meanwhile, Ordinary Kriging was applied for each selected feature in the auxiliary dataset to produce feature grids as part of nested interpolation. The trend model was applied onto each of these feature grids, yielding $PM_{2.5_{\text{trend}}}$ over the grid of interpolation. Lastly, the $PM_{2.5_{\text{trend}}}$ and $PM_{2.5_{\text{SA}}}$ were added together to yield the final interpolated values for the given day. This process was repeated for each day in 2024. The grid over which interpolation occurred was fixed for each day, spanning from -96.090° to -94.618° of longitude and from 29.363° to 30.277° of latitude.

All numerical features in the training 2022-2023 dataset and interpolation 2024 dataset were normalized using Z-normalization, where, for a given feature x_i with mean μ and standard deviation σ in the 2022-2023 training set, the normalized value z_i is:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{15}$$

This is crucial to our trend models, especially given the inclusion of outliers and the wide range of feature scales. Without normalization, features with larger numerical ranges could disproportionately influence the trend models' learning.

The variogram model used for all kriging models was exponential, in line with similar interpolation implementations in the literature (Shi et al., 2020).

5.3 Baseline Models

We must confirm that regression kriging would outperform lighter, simpler models, or else our efforts in this report would be wasted. To this end, we constructed four baseline models to

compare the regression kriging models to. The three baseline models are, 1) inverse distance weighting as outlined in section 2.4.1, 2) K-Nearest Neighbors Regression as outlined in section 2.5.3, 3) a dummy model that predicts the daily average $PM_{2.5}$ level for all sensors, which we will call "average", and 4) an ordinary kriging model as described in section 2.4.2.

5.4 Selecting Hyperparameter Values For Regression Models

We selected values for the hyperparameters of the linear regression and random forest regression models using 5-fold Grid Search Cross-Validation. This is a technique that systematically searches through a predefined grid of hyperparameter values. The technique then evaluates each combination by training the model on different subsets ("folds") of the training set (i.e. 2022-2023 data). Four folds are used to train the model using the given combination, and the remaining held-out fold is used to validate and evaluate the model. This is repeated five times, each time using a different fold for validation. The hyperparameter combination with the best average performance is selected.

The hyperparameters searched for linear regression were regularization strength and L1:L2 ratio. The values searched for regularization strength ranged from 0.0001 (indicating very weak regularization) to 1000 (indicating very strong regularization, shrinking coefficients aggressively). The values searched for L1:L2 ratio ranged from 0.0001 (which is near purely Ridge (L2) Regression) to 1 (which is purely Lasso (L1) Regression). The best combination identified was a regularization strength of 0.001 and a L1:L2 ratio of 0.0001. Therefore, the linear regression model performed better on the training data with very light L2 regularization. This could indicate that the features used were all contributing usefully to estimating PM2.5, with little redundancy.

The hyperparameters searched for random forest, meanwhile, were the maximum tree depth and number of trees. The values searched for the number of trees were 25, 50, 100, 200, 500, 750, and 1000. Meanwhile, the values searched for the maximum depth of trees were 5, 10, 20, 30, 50, 100, and None (which allows the decision trees to keep splitting until all leaves are pure, with no limit to the depth). The best combination identified was a maximum depth of None and a number of 500 trees.

5.5 Experimental Results: Regression Kriging with Linear Regression Outperforms Other Models

Three evaluation metrics were obtained for each of the three models implemented, for each day in 2024. First, the validation root mean squared error (RMSE), which is commonly used to measure the differences between interpolated values and actual values, was calculated. It is defined as:

$$RMSE_i = \sqrt{\frac{\sum(PM_{2.5\text{truth}} - PM_{2.5\text{pred}})^2}{n}} \tag{16}$$

where i is a day in the 2024 subset and n is the number of sensors in the validation set. Meanwhile, the validation median percent error (MPE) of each day was also taken to provide a relative measure of the error, which we define, for each day i, as:

$$\text{MPE}_i = \text{median}\left(\left(\frac{|PM_{2.5,\text{actual}} - PM_{2.5,\text{pred}}|}{PM_{2.5,\text{actual}}} \times 100\right)_i\right) \tag{17}$$

Lastly, for the regression models, the "trend RMSE," which is the RMSE computed from comparing the predictions made by the model against the interpolation set, was also computed for each day. A table showing the median of each of these three daily metrics is shown in Table 4. The training RMSE, which is the RMSE computed comparing the predictions made by the model against the data it was trained on (i.e., the 2022-2023 data), is also reported.

Of the models explored, RK-LR and RK-RF appear to generally outperform the others, having the lowest median validation RMSE and median validation MPE. Indeed, the Regression Kriging models outperformed every baseline model for both validation RMSE and MPE metrics. We can therefore confirm that the Regression Kriging models have captured greater complexity—either due to the fundamental structure of the model or the complexity of the external features.

Model	Training RMSE ($\mu g/m^3$)	Median Validation RMSE ($\mu g/m^3$)	Median Validation MPE (%)	Median Trend RMSE ($\mu g/m^3$)
RK-LR	4.283	1.789 ± 0.139	12.9 ± 0.6	2.862 ± 0.288
RK-RF	0.790	1.744 ± 0.136	13.18 ± 0.546	3.156 ± 0.22
OK	-	7.799 ± 0.199	46.24 ± 1.779	-
IDW	-	3.973 ± 0.235	38.3 ± 2.408	-
KNN	-	2.793 ± 0.118	21.5 ± 0.690	-
Average	-	2.782 ± 0.116	23.0 ± 0.70	-

Table 4: Summary of model performance metrics. The later three columns are medians taken over the span of 2024, and standard errors are reported.

However, there was no significant difference in performance between the two Regression Kriging models themselves. While RK-LR had a high training RMSE, suggesting a lower model complexity (or high bias), it had a lower median trend RMSE, indicating a better generalizability (or low variance). Meanwhile, RK-RF had a lower training RMSE (or low bias), but had a higher median trend RMSE (or high variance). This may indicate some degree of overfitting on the 2022-2023 training data in RK-RF's random forest trend model, even with the use of 5-fold grid-search cross validation. Nevertheless, both models significantly outperform the baselines, namely IDW, OK, the average baseline model, and the kNN baseline model.

The weights of the linear regression trend model are also noteworthy. Table 5 shows the coefficients and significances assigned to each normalized feature retained after removing collinear features. Meteorological features found to be significantly associated with $PM_{2.5}$ include humidity and cloud cover in the positive direction, which some of the literature attributes to $PM_{2.5}$'s effect as an aerosol (Zhou et al., 2023). Meanwhile, stronger wind gust and precipitation are associated with lower $PM_{2.5}$, in line with the literature's definition of both wind and rain as airborne pollutant-diffusing agents (Liu et al., 2016). One unexplained feature deemed significant by the model is "moon phase." It is likely that the moon phase, which essentially acts as a proxy of time of year, was significant due to the temporality of the trend. This is further supported by the fact that both lagged features ($PM_{2.5}$ one day ago and $PM_{2.5}$ seven days ago) were also found to be significant, positive predictors of $PM_{2.5}$.

Contrary to our expectations, point source $PM_{2.5}$ emissions was not a significant predictor of $PM_{2.5}$. This is likely due to the fact that the feature was measured at an annual level. We attempted to introduce temporality at the daily level using a technique previously developed in the literature (Contreras and Ferri, 2016). In this method, only polluters that are within a 30 °sector upwind of the $PM_{2.5}$ sensor with a 1.5-mile radius are considered. Critically, wind direction changes daily, which theoretically improve the resolution of this feature. Nevertheless, the feature still was not a significant predictor of $PM_{2.5}$. Meanwhile, daily traffic volume reported a negative coefficient, counterintuitively supporting that traffic decreases air pollution. However, we also found that two-day lagged traffic has an opposite, positive relationship with $PM_{2.5}$, which supports our hypothesis that traffic majorly emits gases that contribute to secondary $PM_{2.5}$ formation over several days rather than immediate $PM_{2.5}$ emissions. It is also possible that the negative relationship between the immediate daily traffic volume and $PM_{2.5}$ is similar to the negative relationship between wind gust and $PM_{2.5}$that high activity disperses $PM_{2.5}$ particles, and thus decreases the concentration in the short term. Median neighborhood income was a significant negative predictor of $PM_{2.5}$; lower-income neighborhoods appear to be at greater risk of higher $PM_{2.5}$ concentrations given their proximity to major polluters. Similarly, keeping in mind that we omitted medium-intensity land development as the reference category for our land use features, high-intensity development had the greatest, positive coefficient of all land use types. This finding affirms that highly developed neighborhoods in downtown Houston, around major roads, and along the Houston Ship Channel are likely to have greater $PM_{2.5}$ levels. Finally, it appears that the $PM_{2.5}$ trend is spatially dependent, with sensor latitude determined to be a significant feature. In sum, our RK-LR coefficients reaffirm significant effects of weather on $PM_{2.5}$ concentration, as well as points to high-traffic, low-income, and highly developed Houston neighborhoods as having

significantly greater risk of higher air pollution.

Among these significant predictors of $PM_{2.5}$ backward elimination (as described in Section 5.1) only certain features were retained by backward elimination. A summary is shown for each model in 18. Both models used meteorological features like cloud cover, wind direction, precipitation, and wind gust, as well as traffic (immediate and lag of 2 days) and temporal features (moon phase, yesterday's $PM_{2.5}$ and last week's $PM_{2.5}$). Only RK-LR included humidity and the land use categories of pasture and high-intensity several. Furthermore, unlike RK-RF, RK-LR also incorporated latitude, indicating that the trend has a spatial component as well. Meanwhile, RK-RF was the only one to include the land use category of wetland and cultivated crops, as well as yesterday's traffic volume and temperature. Interestingly, all of these had a non-significant coefficient for linear regression (as indicated in 5). This may indicate a non-linear relationship in these features with $PM_{2.5}$ that is not properly captured by the linear regression trend model.

Variable	**Coef**	**P** $> \|t\|$
Humidity (%)	0.3190	<0.0001
Temperature (°)	0.0391	0.388
Atmospheric pressure (mb)	0.0266	0.554
Average wind gust speed (km/hr)	-0.8446	<0.0001
Precipitation level (mm)	-0.5750	<0.0001
Wind direction (°)	-0.3548	<0.0001
Moon phase	-0.2423	<0.0001
Cloud cover (%)	0.2681	<0.0001
Yesterday's $PM_{2.5}$ value ($\mu g/m^3$)	3.0136	<0.0001
Last week's $PM_{2.5}$ value ($\mu g/m^3$)	0.4615	<0.0001
Wind-sensitive emissions (TPY)	-0.0279	0.357
Traffic volume (AADT)	-0.1481	<0.0001
Yesterday's traffic volume (AADT)	-0.0477	0.175
Two days ago's traffic volume (AADT)	0.1779	<0.0001
Median income ($)	-0.1634	<0.0001
Land use - Cultivated Crops	-0.4457	0.106
Land use - Developed, High Intensity	0.5773	<0.0001
Land use - Developed, Low Intensity	0.2021	0.004
Land use - Developed, Open Space	0.2933	<0.0001
Land use - Emergent Herbaceous Wetlands	-0.3010	0.081
Land use - Open Water	-1.3142	0.267
Land use - Pasture/Hay	0.4300	<0.0001
Land use - Woody Wetlands	0.4760	<0.0001
Latitude (°)	0.2828	<0.0001

Table 5: Coefficients and significances of variables in RK-LR trend model. Significant p-values are denoted in red. High lagged traffic volume, low neighborhood median income, and high-intensity land development are all significantly associated with greater $PM_{2.5}$ levels.

Meanwhile, OK performed markedly worse compared to all other models, with a much higher median validation RMSE and validation MPE value. One potential cause of OK's poor performance lies in its assumption that $PM_{2.5}$ values are spatially autocorrelated. Therefore, we applied Moran's I statistical test to each day's distribution of $PM_{2.5}$ values in 2024 to assess for spatial autocorrelation. Of the days surveyed, only 79 exhibited significant evidence of spatial autocorrelation ($p < 0.05$, all were negatively autocorrelated). The violation of spatial autocorrelation likely hindered the performance of OK. As can be seen in the plot of RMSE over time (Figure 19), dates with the highest RMSE had no spatial autocorrelation. Furthermore, IDW, which does not assume spatial autocorrelation, generally outperformed OK.

Somewhat paradoxically, however, application of Moran's I statistical test to the residuals after de-trending also showed a lack of spatial autocorrelation in the residuals on most dates, with only 49 exhibiting significant evidence of spatial autocorrelation ($p < 0.05$, all were negatively autocorrelated). That is, despite having more dates in which spatial autocorrelation was violated,

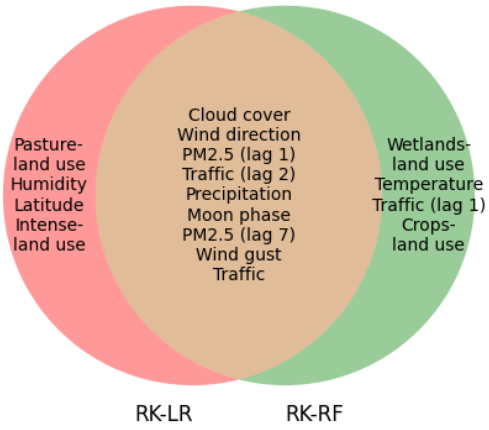

Figure 18: Venn diagram of features selected by backward elimination for each model. The left red region represents features used by only RK-LR. The right green region represents features used by only RK-RF, which are likely non-linear with $PM_{2.5}$. The middle yellow intersection represents features used by both models.

the Regression Kriging models still outperformed pure OK. Shi et al. reported a similar finding (Shi et al., 2020). One possible explanation would be that the poor performance of the Ordinary Kriging portion of Regression Kriging was compensated by the presence of the trend component, which accounts for non-spatial variables. In order to further explore this finding, we developed two new models similar to RK-LR and RK-RF, except we applied IDW to the residuals after de-trending instead of OK (i.e. a sort of "regression IDW"). This resulted in a much worse performance compared to the RK models, with a median validation RMSE of 8.481 ($\mu g/m^3$) for the linear regression model and a median validation RMSE of 8.764 ($\mu g/m^3$) for the random forest regression model.

5.6 Hotspot Analysis: Vulnerable Houston Areas Lie Along the Houston Ship Channel

As part of Objective 1 (identifying vulnerable regions of Houston that are likely to have high levels of air pollution), we conducted a hotspot analysis to investigate spatial patterns in $PM_{2.5}$ concentrations. Getis-Ord spatial statistics (2), like Moran's I, measure how spatially correlated a sensor is with its neighbors (defined by us as the three nearest sensors). However, it does so by identifying clusters of unusually high or low values relative to the rest of the sample. The analysis began by calculating the global Getis-Ord G_i statistic for each day from 2022 through 2024. This produced a global G_i score representing the degree of spatial correlation among $PM_{2.5}$ values on a given day, along with a p-value to test whether the clustering was statistically significant.

For each date with significant global G_i score ($p \leq 0.05$), indicating some amount of clustering, we applied the local Getis-Ord G_i^* statistic. This analysis provided a local G_i^* score and p-value for each sensor on each selected date, allowing us to detect sensors that were part of statistically significant clusters of high or low $PM_{2.5}$ values. Sensors with local G_i^* scores greater than 1 were labeled as hotspots, indicating $PM_{2.5}$ values that were significantly higher than those of other sensors. Meanwhile, those with local G_i^* scores less than -1 were labeled as cold spots, indicating significantly lower $PM_{2.5}$ levels relative to other sensors. The resulting hotspot and cold spot sensors were recorded for each date with clustering.

The sensor that most frequently was identified as a hotspot was "Glen Manor" (detected on

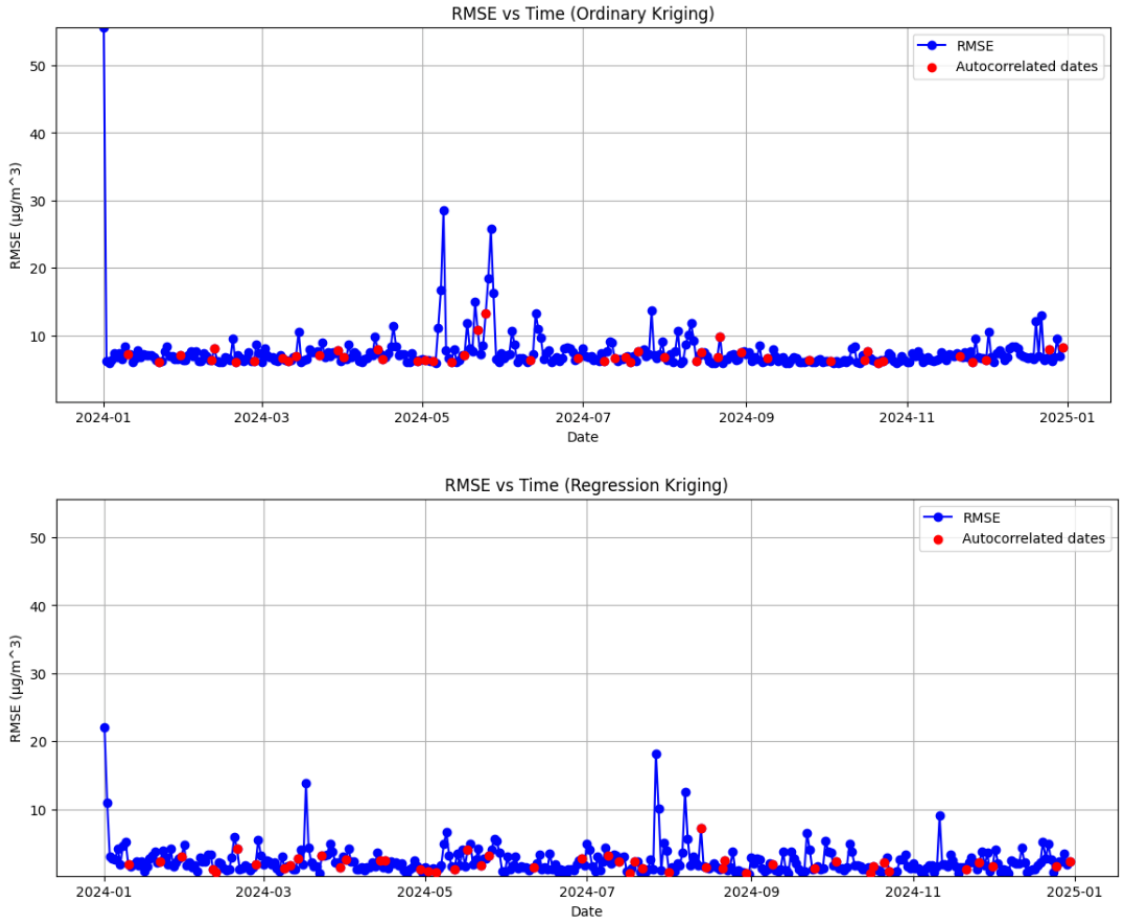

Figure 19: OK and RK-LR perform poorly in times of no spatial autocorrelation. The RMSE of the OK model (above) and the RK-LR model with Linear Regression (below) was plotted over time for 2024. Points with no spatial autocorrelation are marked in blue and tend to show a higher error, while those with spatial autocorrelation are marked in red and tend to show a lower error.

199 separate dates), which is located in Galena Park, near the Houston Ship Channel. Many other hotspots were similarly located along the channel or in neighborhoods near the channel (i.e. Deer Park, Baytown) (Figure 20). It is likely that traffic from cargo ships and industrial sources of pollution (as revealed in Section 4.2.6) contributed to the increased $PM_{2.5}$ emissions. Meanwhile, areas in Northeast Houston were also filled with hotspots, such as Settegast, which has the lowest life expectancy in Houston (Schuetz, n.d.), and Kashmere Gardens, a low-income neighborhood that has been designated as a cancer cluster–an area with disproportionately high rates of cancer (Ryan, n.d.). It is then crucial that these neighborhoods, which are especially vulnerable to high levels of $PM_{2.5}$, are reliably monitored.

Meanwhile, the sensor most frequently identified as a coldspot (detected on 113 separate dates), was "Royal Oaks Houston Tx." This area, in the Memorial area, is generally more affluent, with a median household income of $115,937 according to ACS data. These results suggest that certain areas of the city had reliably cleaner air over time.

5.7 Ablation Study: Excluding PurpleAir Sensors Causes Vulnerable Neighborhoods to be Overlooked

As part of Objective 2 (to compare the influence of regulatory and non-regulatory sensors on model performance), we conducted an ablation study by removing either TCEQ or PurpleAir sensors.

First, we aimed to train and evaluate RK-LR models with the features specified in Section 5.1:

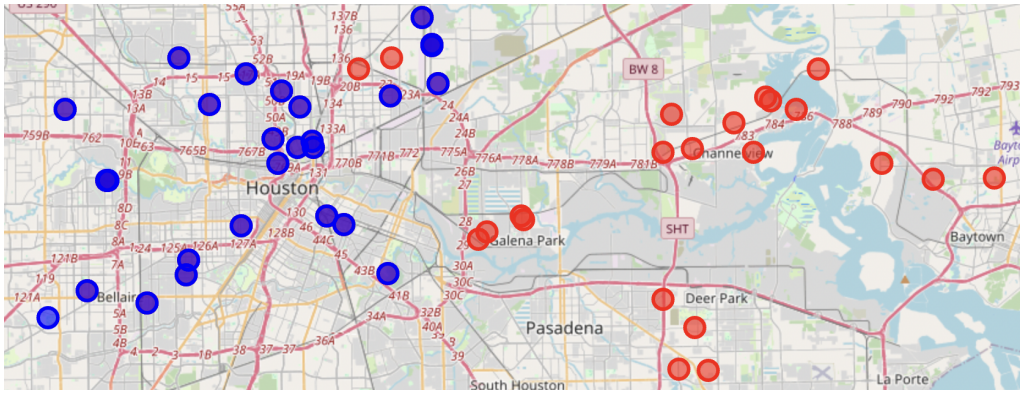

Figure 20: Significant hotspots (red) over time are located along the Houston Ship Channel. Meanwhile, coldspots (blue) are located more inland, such as in Memorial.

one that used only PurpleAir sensors, and one that only used TCEQ sensors. Daily validation RMSE and validation MPE were calculated across the 2024 interpolation set, and the median value of each metric was used to summarize the impact of removing each subset of sensors. For the RK-LR model that used only PurpleAir sensors, there were modest increases in both median validation RMSE (1.789 vs. 2.485 $\mu g/m^3$) and median validation MPE (12.9% vs. 20.086%), indicating a slightly worse performance after excluding TCEQ sensors. This is likely due to reduced coverage in areas like Baytown, where there are no PurpleAir sensors. It should be noted, however, that this performance is still better than those of the baseline models. Meanwhile, the RK-LR model that used only TCEQ sensors could not be trained, as there were days in which the volume of the data was not sufficient to generate a variogram. This reflects the sheer sparsity of the TCEQ network.

Secondly, we conducted hotspot analysis on the same ablated datasets. Figure 21 compares the hotspots captured by each dataset. The PurpleAir dataset revealed most of the same hotspots as the unablated dataset, especially Settegast, Kashmere Gardens, the Houston Ship Channel, Deer Park, and Galena Park. The TCEQ dataset, meanwhile, only highlighted Settegast and Galena Park. By failing to detect Kashmere Gardens, the Houston Ship Channel, and Deer Park—low-income areas with a high risk of chronic $PM_{2.5}$ exposure—the TCEQ dataset risks omitting these vulnerable communities from regulatory interventions. This emphasizes the importance of geographic breadth in monitoring $PM_{2.5}$

5.8 Interactive Air Pollution Map

With the modeling and supporting tools completed, we developed an interactive, user-friendly dashboard to fulfill our final objective (Objective 4). This dashboard enables users to quickly train a Regression Kriging model to predict $PM_{2.5}$ levels across the Houston Metro Area using customizable parameters. It is designed to provide accurate predictions for a selected date and location, while also offering advanced options such as feature selection, model selection, and prediction export. Users can choose between RK-LR and RK-RF and toggle the inclusion of any feature that was not filtered out by VIF in Section 5.1.

After the user has made their choices, selects an interpolation date, and enters their address, the dashboard trains their chosen model using the features specified to generate an interactive map displaying $PM_{2.5}$ levels across Houston. In addition to exploring the absolute values of $PM_{2.5}$ concentrations for that given day, users can also toggle an option that displays the relative $PM_{2.5}$ values, enhancing contrast to make areas with higher $PM_{2.5}$ concentrations darker (Figure 22). At the top of the screen, the predicted $PM_{2.5}$ concentration at their specified location is displayed, accompanied by a brief explanation of the health implications associated with the predicted levels (as discussed in Section 2). For instance, a level of 16.6 $\mu g/m^3$ would yield the following output: "This concentration (>15.0 $\mu g/m^3$) is considered to be concerning at a 24-hour concentration according to the World Health Organization. However, it is not considered so by the Texas

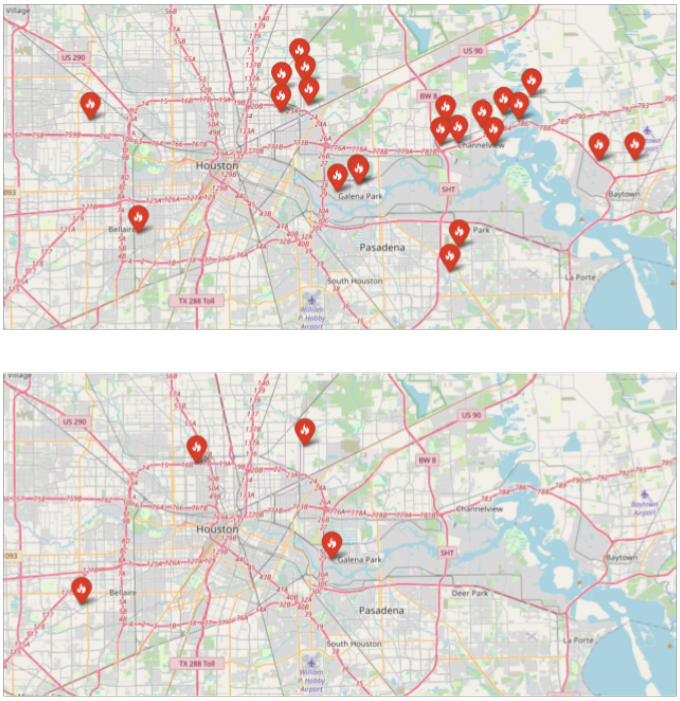

Figure 21: Loss of PurpleAir sensors results in vulnerable areas being overlooked. Hotspots when using only PurpleAir data (above) or TCEQ data (below) are shown. The ablated dataset using only PurpleAir sensors captures hotspots in Galena Park, Deer Park, the Houston Ship Channel, Settegast, and Kashmere Gardens. Meanwhile, the ablated dataset using only TCEQ sensors captures only Galena Park and Baytown.

Commission on Environmental Quality, who recently changed their standards to >35 $\mu g/m^3$." At the request of our sponsors, users can export the interpolated values as a CSV file containing coordinates for each point in the grid of interpolation, accompanied by their corresponding $PM_{2.5}$ levels. This enables custom visualizations or further analysis for those interested in exploring the results in more depth.

The dashboard serves as both a communication tool and a public resource. Our dashboard features a checkbox that lets users enable hotspot analysis—highlighting areas that experience disproportionately high concentrations of $PM_{2.5}$ on that day. If the nearest hotspot is in one of the neighborhoods discussed in Section 5.6, the user will be notified of their distance to the nearest hotspot. Furthermore, a short article will be displayed, presenting information about nearby industrial facilities, incidents, and observed health risks, accompanied by news articles for further information (Figure 23. This helps residents make more informed health decisions by providing accessible air quality information for Houston. At the same time, it brings attention to areas with concerning $PM_{2.5}$ concentrations, supporting broader public discourse around air quality and environmental justice in the region.

6 Conclusions

6.1 Impact

By creating spatial interpolation models and visualizing them through a user-interactive interface, our project educates Houston residents by providing accurate estimates of particulate pollution in their area, thereby increasing overall environmental knowledge and awareness. This information can influence decision-making, such as choosing a place to live or selecting a career location.

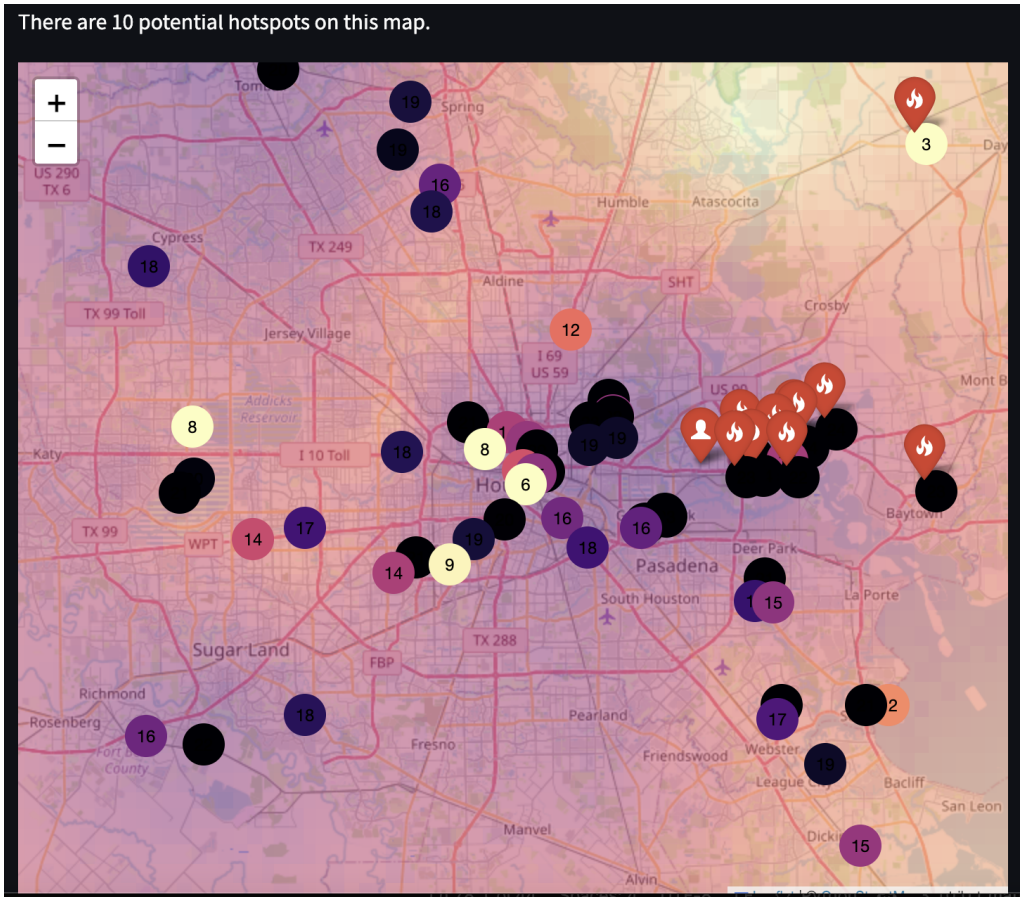

Figure 22: A screengrab of our dashboard's display of an interactive map. This has toggled on the "Visualize relative $PM_{2.5}$levels" and the "Want to know more about hotspots?" setting, displaying red markers (indicating hotspots) and visualizing relative $PM_{2.5}$ levels, with darker spots corresponding to areas of higher $PM_{2.5}$ concentration.

Furthermore, our hotspot analysis revealed how certain areas had reliably cleaner air compared to other areas. Through our ablation study, we were able to demonstrate the importance of the placement of sensors in creating our interpolation model, as well as the detection of high-risk neighborhoods. Lastly, we believe our project can lead to change at the legislative level, potentially advocating for a more robust, fair, and equitable distribution of TCEQ monitors across the Greater Houston area.

6.2 Future Work

Our results with Regression Kriging, particularly in combination with Linear Regression are promising, and they pave the way for further exploration. We found that income is a significant predictor of $PM_{2.5}$ concentration, and it would be worthwhile to see if other demographic factors, such as percentage Black, percentage Hispanic, or social vulnerability index, may also be significant predictors. Secondly, certain areas, such as Fifth Ward and Sunnyside, are low-income neighborhoods that have neither PurpleAir nor TCEQ sensors. Installing sensors in these areas through the help of advocate groups like Air Alliance Houston (or perhaps obtaining readings from another source of sensors) could enhance our model further, and shed light on the concentrations of $PM_{2.5}$ on those areas. Thirdly, obtaining polluter data at a finer resolution than yearly averages might allow our model to utilize point source emissions data more appreciably. By implementing these refinements, we anticipate further improvements in our interpolation models, supporting our goal of comprehensively "filling the map" of air pollution for Houston residents.

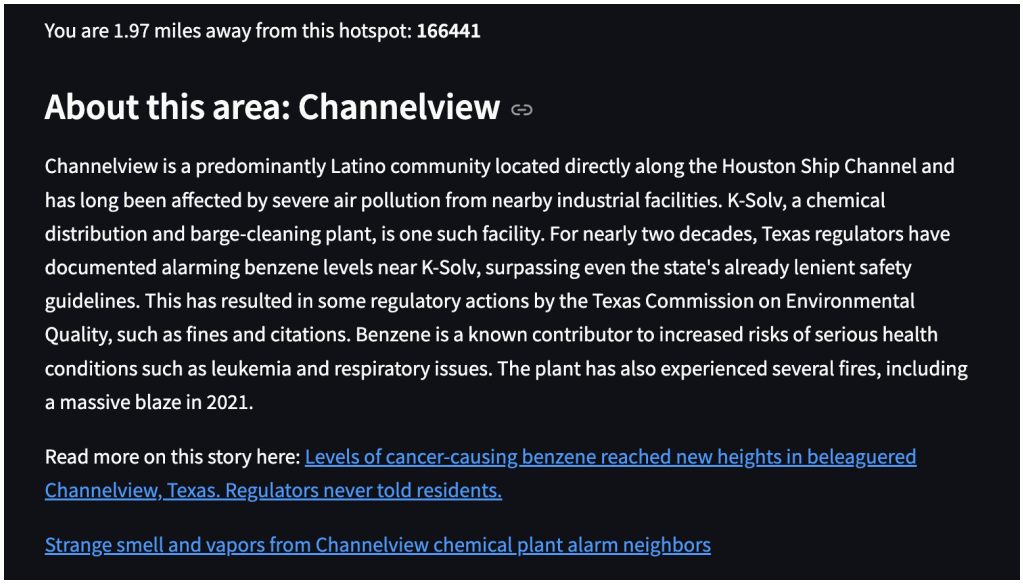

Figure 23: An example of an article about a hotspot, giving the user information about nearby polluted neighborhoods. This resulted from the user inputting the zip code "77015" on March 15, 2024.

7 References

California Air Resources Board. (n.d.). *PM2.5 | California Air Resources Board*. Retrieved February 25, 2025, from https://ww2.arb.ca.gov/our-work/programs/state-and-federal-area-designations/federal-area-designations/pm2-5

CDC. (2024, October 22). *Social Vulnerability Index*. Place and Health - Geospatial Research, Analysis, and Services Program (GRASP). https://www.atsdr.cdc.gov/place-health/php/svi/index.html

Chao, C.-Y., Li, W., Hopke, P. K., Guo, F., Wang, Y., & Griffin, R. J. (2025). Increases in PM2.5 levels in Houston are associated with a highly recirculating sea breeze. *Environmental Pollution, 366*, 125381. https://doi.org/10.1016/j.envpol.2024.125381

Choi, K., & Chong, K. (2022). Modified Inverse Distance Weighting Interpolation for Particulate Matter Estimation and Mapping. *Atmosphere, 13*(5), 846. https://doi.org/10.3390/atmos13050846

Contreras, L., & Ferri, C. (2016). Wind-sensitive Interpolation of Urban Air Pollution Forecasts. *Procedia Computer Science, 80*, 313–323. https://doi.org/10.1016/j.procs.2016.05.343

Cressie, N. (1989). Spatial prediction and ordinary kriging. *Mathematical Geology, 21*(4), 493–494. https://doi.org/10.1007/BF00897332

De Mesnard, L. (2013). Pollution models and inverse distance weighting: Some critical remarks. *Computers & Geosciences, 52*, 459–469. https://doi.org/10.1016/j.cageo.2012.11.002

Doreian, P. (1981). Estimating linear models with spatially distributed data. *Sociological Methodology, 12*, 359–388.

Environmental Protection Agency. (2025). *Health and Environmental Effects of Particulate Matter (PM)*. Retrieved from https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matte

Finne, E., & Sauzet, O. (2025). Feasibility of Using Survey Data and Semi-variogram Kriging to Obtain Bespoke Indices of Neighborhood Characteristics: A Simulation and a Case Study. Geogr Anal, 57, 3-26. https://doi.org/10.1111/gean.12401

Fielding, R. T. (n.d.). Chapter 5: Representational State Transfer (REST). In *Architectural Styles and the Design of Network-based Software Architectures (Ph.D.)*. University of California, Irvine.

Fu, H., Zhang, Y., Liao, C., Mao, L., Wang, Z., & Hong, N. (2020). Investigating PM2.5 responses to other air pollutants and meteorological factors across multiple temporal scales. *Sci-*

entific Reports, 10(1), 15639. https://doi.org/10.1038/s41598-020-72722-z

Getis, Arthur, and J. K. Ord. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis, vol. 24, no. 3*, 1992, pp. 189-206. Ohio State University Press. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Gu, K. H., Shi, H. C., Zhang, S., Fan, S. X., Xu, J. M., & Tan, J. G. (2015). Variation characteristics of PM2.5 levels and the influence of meteorological conditions on chongming island in shanghai. *Resources and Environment in the Yangtze Basin, China*, 24(12), 2108-2116.

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma, 120*(1), 75–93. doi:10.1016/j.geoderma.2003.08.018

Jana, Mrityunjoy, and Nityananda Sar. "Modeling of Hotspot Detection Using Cluster Outlier Analysis and Getis-Ord Gi* Statistic of Educational Development in Upper-Primary Level, India." *Modeling Earth Systems and Environment, vol. 2, no. 60*, 2016, pp. 1-10, Springer International Publishing, doi:10.1007/s40808-016-0122-x.

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

Klemmer, K., Safir, N. S., & Neill, D. B. (2023). *Positional encoder graph neural networks for geographic data*, 1379–1389.

Kloog, I., Ridgway, B., Koutrakis, P., Coull, B. A., & Schwartz, J. D. (2013). Long- and Short-Term Exposure to PM2.5 and Mortality: Using Novel Exposure Models. *Epidemiology, 24*(4), 555–561. https://doi.org/10.1097/EDE.0b013e318294beaa

Kyriakidis, P. C., & Goodchild, M. F. (2006). On the prediction error variance of three common spatial interpolation schemes. *International Journal of Geographical Information Science, 20*(8), 823–855. https://doi.org/10.1080/13658810600711279

Le, V.-D., Bui, T.-C., & Cha, S.-K. (2020). Spatiotemporal Deep Learning Model for Citywide Air Pollution Interpolation and Prediction. 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 55–62. https://doi.org/10.1109/BigComp48618.2020.00-99

Li, J., & Heap, A. D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics, 6*(3–4), 228–241. https://doi.org/10.1016/j.ecoinf.2010.12.003

Liu, S., Ganduglia, C. M., Li, X., Delclos, G. L., Franzini, L., & Zhang, K. (2016). Short-term associations of fine particulate matter components and emergency hospital admissions among a privately insured population in Greater Houston. *Atmospheric Environment, 147*, 369–375. https://doi.org/10.1016/j.atmosenv.2016.10.021

M.A. Oliver, R. Webster, A tutorial guide to geostatistics: Computing and modelling variograms and kriging, CATENA, Volume 113, 2014, Pages 56-69, ISSN 0341-8162, https://doi.org/10.1016/j.catena.2013.09.006.

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., Avol, E. L., Oron, A. P., Larson, T., Liu, L.-J. S., & Kaufman, J. D. (2011). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment, 45*(26), 4412–4420. https://doi.org/10.1016/j.atmosenv.2011.05.043

Mesić Kiš, I. (2016). Comparison of Ordinary and Universal Kriging interpolation techniques on a depth variable (a case of linear spatial trend), case study of the Šandrovac Field. *Rudarsko-Geološko-Naftni Zbornik, 31*(2), 41–58. https://doi.org/10.17794/rgn.2016.2.4

Moran, P. A. P. "Notes on Continuous Stochastic Phenomena." *Biometrika, vol. 37, no. 1/2*, 1950, pp. 17–23. JSTOR.

Murphy, B., Yurchak, R., & Müller, S. (2024). *GeoStat-Framework/PyKrige: V1.7.2* (Version v1.7.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3738604

Nazarenko, Y., Pal, D., & Ariya, P. A. (2020). Air Quality Standards for the Concentration of Particulate Matter 2.5, Global Descriptive Analysis. *Bulletin of the World Health Organization, 99*(2). doi:10.2471/blt.19.245704

New York State Department of Health. (2025). *Particle pollution and health.* Retrieved from https://www.health.ny.gov/environmental/indoors/air/pmq_a

Ruidas, Dipankar, and Subodh Chandra Pal. "Potential Hotspot Modeling and Monitoring of PM2.5 Concentration for Sustainable Environmental Health in Maharashtra, India." *Sustainable Water Resources Management*, vol. 8, no. 98, 2022, Springer Nature Switzerland, doi:10.1007/s40899-022-00682-5

Sadeghi, B., Choi, Y., Yoon, S., Flynn, J., Kotsakis, A., & Lee, S. (2020). The characterization of fine particulate matter downwind of Houston: Using integrated factor analysis to identify anthropogenic and natural sources. *Environmental Pollution, 262*, 114345. https://doi.org/10.1016/j.envpol.2020.114345

Sexton, K., Linder, S. H., Marko, D., Bethel, H., & Lupo, P. J. (2007). Comparative Assessment of Air Pollution–Related Health Risks in Houston. *Environmental Health Perspectives, 115*(10), 1388–1393. https://doi.org/10.1289/ehp.10043

Shamsoddini, A., Aboodi, M. R., & Karami, J. (2017). TEHRAN AIR POLLUTANTS PREDICTION BASED ON RANDOM FOREST FEATURE SELECTION METHOD. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4/W4*, 483–488. doi:10.5194/isprs-archives-XLII-4-W4-483-2017

Shi, T., Dirienzo, N., Requia, W. J., Hatzopoulou, M., & Adams, M. D. (2020). Neighbourhood scale nitrogen dioxide land use regression modelling with regression kriging in an urban transportation corridor. *Atmospheric Environment, 223*, 117218. doi:10.1016/j.atmosenv.2019.117218

Shukla, K., Kumar, P., Mann, G. S., & Khare, M. (2020). Mapping spatial distribution of particulate matter using Kriging and Inverse Distance Weighting at supersites of megacity Delhi. *Sustainable Cities and Society, 54*, 101997. https://doi.org/10.1016/j.scs.2019.101997

Staniswalis, J. G., Yang, H., Li, W.-W., & Kelly, K. E. (2009). Using a Continuous Time Lag to Determine the Associations between Ambient PM_2.5 Hourly Levels and Daily Mortality. *Journal of the Air & Waste Management Association, 59*(10), 1173–1185. doi:10.3155/1047-3289.59.10.1173

Strasert, B., Teh, S. C., & Cohan, D. S. (2019). Air quality and health benefits from potential coal power plant closures in Texas. *Journal of the Air & Waste Management Association, 69*(3), 333–350. https://doi.org/10.1080/10962247.2018.1537984

Tai, A. P. K., Mickley, L. J., Jacob, D. J., Leibensperger, E. M., Zhang, L., Fisher, J. A., & Pye, H. O. T. (2012). Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: Implications for PM2.5 sensitivity to climate change. *Atmospheric Chemistry and Physics, 12*(6), 3131–3145. https://doi.org/10.5194/acp-12-3131-2012

Tessum, C. W., Paolella, D. A., Chambliss, S. E., Apte, J. S., Hill, J. D., & Marshall, J. D. (2021). PM2.5 polluters disproportionately and systemically affect people of color in the United States. *Science Advances, 7*(18), eabf4491. https://doi.org/10.1126/sciadv.abf4491

Texas Commission on Environmental Quality. (n.d.). Air Pollution from Particulate Matter. Texas Commission on Environmental Quality. Retrieved February 25, 2025, from https://www.tceq.texas.gov/airquality/sip/criteria-pollutants/sip-pm#latest

Thangavel, P., Park, D., & Lee, Y.-C. (2022). Recent insights into particulate matter (pm2.5)-mediated toxicity in humans: An overview. *International Journal of Environmental Research and Public Health*, 19(12), 7511. https://doi.org/10.3390/ijerph19127511

U.S. News & World Report. (n.d.). *Urban air quality: Best states rankings.* U.S. News & World Report. Retrieved February 25, 2025, from https://www.usnews.com/news/best-states/rankings/natural-environment/air-water-quality/urban-air-quality

VanCuren, R. (tony), & Gustin, M. S. (2015). Identification of sources contributing to PM2.5 and ozone at elevated sites in the western U.S. by receptor analysis: Lassen Volcanic National Park, California, and Great Basin National Park, Nevada. *Science of The Total Environment, 530–531*, 505–518. doi:10.1016/j.scitotenv.2015.03.091

Wallace, L. (2022). Intercomparison of PurpleAir Sensor Performance over Three Years Indoors and Outdoors at a Home: Bias, Precision, and Limit of Detection Using an Improved Algorithm for Calculating PM2.5. *Sensors, 22*(7), 2755. https://doi.org/10.3390/s22072755

Weiwei, P. U., Xiujuan, Z., & Xiaoling, Z. (2011). Effect of Meteorological Factors on PM 2.5 in Late Summer and Early Autumn of Beijing. *Journal of Applied Meteorological Science, China*, 22(6), 716-723. http://qikan.camscma.cn/en/article/id/20110609

World Health Organization. (2025). *Ambient (outdoor) Air Pollution.* Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Xie, W.-F., Li, J.-K., Peng, K., Zhang, K., & Ullah, Z. (2024). The Application of Local Moran's I and Getis–Ord Gi* to Identify Spatial Patterns and Critical Source Areas of Agricultural Nonpoint Source Pollution. *Journal of Environmental Engineering, 150*(5). doi: 10.1061/JOEEDU.EEENG-7585

Zhang, X., Craft, E., & Zhang, K. (2017). Characterizing spatial variability of air pollution from vehicle traffic around the Houston Ship Channel area. *Atmospheric Environment*, 161, 167–175. https://doi.org/10.1016/j.atmosenv.2017.04.032

Zhang, Y., Tino, P., Leonardis, A., & Tang, K. (2021). A Survey on Neural Network Interpretability. *PIEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5), 726–742. https://doi.org/10.1109/TETCI.2021.3100641

Zhou, L., Sun, L., Luo, Y., Xia, X., Huang, L., Liao, Z., & Yan, X. (2023). Air pollutant concentration trends in China: correlations between solar radiation, PM2.5, and O3. *Air Quality, Atmosphere & Health, 16*(8), 1721–1735. doi:10.1007/s11869-023-01368-3

Zimmerman, D., Pavlik, C., Ruggles, A., & Armstrong, M. P. (1999). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, 31, 375–390.

A Appendix

A.1 Dataset Variable Documentation

The following table names and describes the columns in each of the primary $PM_{2.5}$ and auxiliary feature datasets used in our project.

Combined $PM_{2.5}$ Dataset	
Site_id	Unique identifier for sensor
Sensor_name	Common name of sensor. If there is no name, then this is site_id
Timestamp	Time at which $PM_{2.5}$ recording was collected
PM2.5	PurpleAir: formerly pm2.5_alt, a measurement of $PM_{2.5}$ in which the raw value is multiplied by 3. Based on industry expertise, this value is further scaled by a calibration of factor 3.4/3 TCEQ: an average of "$PM_{2.5}$ Local Conditions" and "$PM_{2.5}$ Speciation Mass" Both are measured in $\mu g/m^3$
Longitude	Longitude of PurpleAir or TCEQ sensor
Latitude	Latitude of PurpleAir or TCEQ sensor
Indoor_outdoor	For PurpleAir sensors, whether or not sensor is indoor or outdoor
Point Source Emissions Dataset	
rn	Registration number
company	Company name
zip.code	ZIP code of polluter site
longitude	Longitude of polluter site
latitude	Latitude of polluter site
county	County of polluter site
annual.routine.tpy	Annual emissions of polluter's routine operations in tons per year
Visual Crossing Weather Dataset	
name	Name of location at which weather was recorded
address	Address of location at which weather was recorded
resolvedAddress	Code of address of location at which weather was recorded
latitude	Latitude of location at which weather was recorded
longitude	Longitude of location at which weather was recorded
datetime	Full date and timestamp on which weather was recorded
day	Day on which weather was recorded
month	Month in which weather was recorded
year	Year in which weather was recorded
tempmax	Maximum temperature in degrees Fahrenheit
tempmin	Minimum temperature in degrees Fahrenheit
temp	Mean temperature in degrees Fahrenheit
feelslikemax	Maximum feels like temperature in degrees Fahrenheit
feelslikemin	Minimum feels like temperature in degrees Fahrenheit
feelslike	Mean feels like temperature in degrees Fahrenheit
dew	Dew point in degrees Fahrenheit
humidity	Relative humidity percentage
precip	Precipitation in inches
precipprob	Precipitation chance percentage
precipcover	Precipitation cover percentage
preciptype	Precipitation type
snow	Snow in inches
snowdepth	Snow depth in inches
windgust	Wind gust in miles per hour
windspeed	Wind speed in miles per hour
windspeedmax	Maximum wind speed in miles per hour

windspeedmean	Mean wind speed in miles per hour
windspeedmin	Minimum wind speed in miles per hour
winddir	Direction wind is blowing from in degrees from North
sealevelpressure	Atmospheric sea level pressure in millibars (mb)
cloudcover	Cloud cover percentage
visibility	Visibility in miles
solarradiation	Solar radiation in W/m^2
solarenergy	Solar energy in MJ/m^2
uvindex	UV index
severerisk	Indicator of severe weather risk
sunrise	Sunrise time
sunset	Sunset time
moonphase	Moonphase
conditions	Short text about weather
description	Description of weather for day
icon	Weather icon to show when displaying data
stations	List of weather station sources
src	Indicator of whether weather was observed by physical station, remote source (i.e., satellite/radar), or both
TDOT Traffic Dataset	
District	District of traffic station
County	County of traffic station
Traffic_Station_ID	Unique traffic station identifier
AADT_2023	Calculated vehicles passing over station per day
Latitude	Latitude of traffic station
Longitude	Longitude of traffic station
Active	Whether station is taking data during given year
County_Cycle	Number of 24 hour cycles recorded
On_Road	Official road name sensor is located on