Geospatial Embeddings for Pollution Prediction: NO_2 in the Bay Area

Peter F. Grinde-Hollevik[a]

[a] *Rausser College of Natural Resources, University of California, Berkeley,*

Abstract

The ever-increasing challenge of air pollution necessitates innovative approaches for accurately monitoring and predicting pollutant concentrations, particularly nitrogen dioxide (NO_2). This study introduces a predictive modeling framework that utilizes a novel data structure, geospatial embeddings derived from satellite imagery, to predict census block median NO_2 concentrations across Bay Area cities, including Berkeley, Palo Alto, San Francisco, and Oakland. These predictions are based on in-situ pollution measurements from Google Street View car data.

Aiming to bridge the gap in high-resolution, accurate NO_2 concentration predictions across extensive areas, including those without sensor coverage, our research evaluates the performance of various machine learning models. These include Neural Networks, Support Vector Regression (SVR), as well as multiple forms of Regression.

Key findings reveal that SVR and Ridge Regression models outperform others in capturing the intricate spatial patterns of NO_2 pollution across the Bay Area. Furthermore, the integration of Earth Index and MOSAIKS geospatial embeddings features through a Random Forest Regressor significantly enhances results, highlighting the importance of leveraging diverse data sources for environmental monitoring. Our model achieves high precision within the sample, with prediction errors bounded within $\pm 0.1\%$. However, when applied to out-of-sample areas, such as Palo Alto, Redwood City, and Millbrae, the prediction accuracy broadens to $\pm 10\%$, suggesting a variability that warrants cautious interpretation for direct application in unmonitored regions. Our spatial analysis of prediction errors emphasizes the model's capacity to identify pollution hotspots, providing valuable insights that could guide targeted interventions and policy formulation for effective pollution control in a resource-scarce environment.

Keywords:
Remote Sensing, Machine Learning, Neural Networks, Pollution Prediction, Geospatial Embeddings

Contents

1. Acknowledgements

This honors thesis would have not been possible without the tremendous support I've received from Professor Duncan Callaway and Professor Sofia Villas-Boas.

Research Advisor: Prof. Duncan Callaway, Energy and Resources Group, Rausser College of Natural Resources University of California, Berkeley

Faculty Sponsor: Prof. Sofia Villas-Boas EEP/ARE, Rausser College of Natural Resources University of California, Berkeley

I also kindly thank Caleb Kruse, formerly at the non-profit Earth Genome, for sharing the Earth Index embeddings data and sound advice.

2. Introduction

Air quality is a critical environmental and public health concern that affects ecosystems, climate, and human health worldwide. Nitrogen dioxide (NO_2), a prevalent air pollutant primarily produced by combustion processes, is associated with various adverse health outcomes, including respiratory issues, and environmental impacts such as acid rain and smog formation. The accurate monitoring and prediction of NO_2 concentrations are essential for effective pollution management and mitigation strategies. (1)

Traditional approaches to monitoring NO_2 have relied heavily on in-situ measurements using ground sensors. While these methods provide accurate local readings, they suffer from limited spatial coverage and cannot offer insights into larger geographic trends or areas without sensor deployment. (2) Recent advancements in machine learning and remote sensing technologies, particularly the use of geospatial embeddings, have opened new avenues for enhancing pollution prediction by leveraging the rich spatial context of environmental data. However, there remains a significant gap in applying these advancements to produce high-resolution, accurate NO_2 concentration predictions across wide areas, including those lacking ground sensor coverage. (3; 4)

This research aims to address the identified gaps by employing a novel approach that applies these geospatial embeddings as features in a series of machine learning models. By setting our study within the Bay Area, a region characterized by significant urbanization and traffic—and consequently, higher levels of NO_2 pollution—we leverage both Google Street View car data and Earth Index embeddings, followed by an exploration of MOSAIKS embeddings. Our methodology not only compares these distinct sources of embeddings but also investigates the potential of their combined use for improving predictive accuracy. (5)

Our contributions are twofold: first, we demonstrate the efficacy of geospatial embeddings in predicting NO_2 concentrations, offering a significant improvement over traditional in-situ monitoring methods in terms of spatial coverage and prediction resolution. Second, through a comparative analysis of Earth Index and MOSAIKS embeddings and the combination of these datasets, we present a simple model that enhances predictive accuracy across the Bay Area. This model utilizes over 4000 features derived from the Multi-task Observation using SAtellite Imagery & Kitchen Sinks (MOSAIKS) project and temporal-spatial reductions of Sentinel-2 satellite bands, with the MOSAIKS being easily accessible to the public. (6; 7; 4)

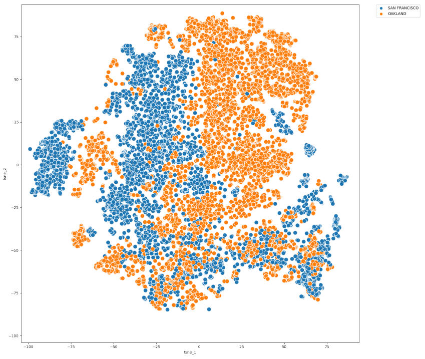

Figure 1: Earth Index Embeddings reduced to 2-d using t-distributed stochastic neighbor embedding (t-SNE). Observations from San Francisco in blue and Oakland in orange.

Geospatial embeddings, particularly those generated by the MOSAIKS method, extract comprehensive information from satellite imagery. The process involves applying small, randomly selected patches from these images as filters across the entire image. By convolving these patches over the image, the method captures detailed spatial patterns and characteristics within the imagery. These patterns are then represented as high-dimensional vectors, which are referred to as geospatial embeddings in this paper. This approach effectively compresses and encodes the rich, complex information present in satellite images into a manageable, analytical format. (4)

By analyzing these diverse data sources and embedding techniques, this study contributes to the evolving field of air quality monitoring and prediction, offering insights into the capabilities and limitations of current methodologies and the potential of machine learning algorithms to bridge these gaps.

3. Background

The evolving landscape of prediction modeling using remote sensing data has witnessed the emergence of diverse and innovative frameworks tailored to address the complex challenges of air quality monitoring. Among these, Muthukumar et al.'s integration of in-situ meteorological data and pollutant measurements with Convolutional Long-Short Term Memory (ConvLSTM) networks represents a significant leap forward, enabling 10-day ahead forecasts of PM2.5 concentration in Los Angeles County, US (8). Similarly, Lim et al. have developed an algorithm that leverages remote-sensed reflectance measurements for direct conversion into PM10 concentrations on Penang Island, Malaysia (9). The work of Kamińska further enriches the field by utilizing traffic flow and meteorological factors—specifically wind direction and speed—alongside two sets of Random Forest models to predict atmospheric NO_2 concentrations in Warsaw, Poland (10).

Expanding the methodological diversity, Yan et al. employ LTSM to learn patterns between spatially distributed monitoring sites for 1-hour ahead forecasts of Beijing's AQI, showcasing the potential of deep learning in local air pollutant predictions (11). Complementing these efforts, Pak et al. propose a 2-stage process for predicting 8-hour average ozone concentrations, further illustrating the adaptability of prediction models to various pollutants (12). The potential of spatial-temporal predictions is further demonstrated by Kruse et al., who explore the global terrestrial aggregation of plastic waste using geospatial embeddings, building upon Karra et al.'s training of a land-use/land-cover (LULC) classifier on human-labeled Sentinel-2 pixels (13; 14).

In this dynamic context, the work by Novotny et al. on the national satellite-based land-use regression for NO2 across the United States provides an essential reference point (15). Their study harnesses satellite and ground-based NO2 measurements alongside geographic characteristics to offer high-resolution estimates of ambient NO2 pollution. Their methodology demonstrates the significant potential for enhancing air quality modeling through the integration of satellite data.

Apte et al.'s method of equipping Google Street View cars with air quality sensors introduces a novel approach to hyper-local pollution measurement, demonstrating the critical role of mobile monitoring in capturing spatial variability in urban air quality (16). This innovative sampling technique allows our target variable to be at finer spatial granularity than the aforementioned studies.

In parallel, the comprehensive review by Zhang et al. (2022) on deep learning methods for air pollutant concentration prediction and the critical evaluation of land-use regression (LUR) models by Hoek et al. (2008) provide a foundational understanding of the current state and challenges in the field. These works highlight the significance of dynamic correlations in pollutant concentrations and the integration of auxiliary features such as Points of Interest (POIs), traffic, and meteorological data for model accuracy enhancement (17; 18).

Building on these advancements, this paper seeks to carve a unique niche by combining the strengths of geospatial embeddings with the predictive power of deep learning models. By eschewing the recurrent temporal component in favor of predicting median pollutant concentrations and incorporating a geospatially embedded remote sensing component, this research aims to explore new frontiers in air quality prediction. The critical proposition tested in this investigation is whether such a novel combination can yield highly accurate predictions of census block-level local air pollution, thereby offering a promising new direction for environmental monitoring and public health protection.

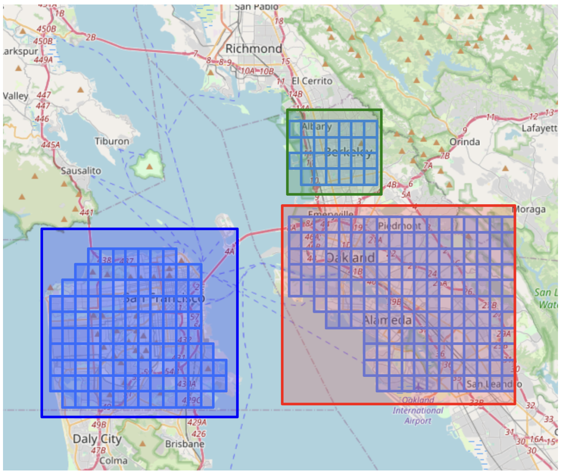

Figure 2: MOSAIKS & Earth Index Embeddings coverage area of San Francisco (Blue), Oakland (Red), and Berkeley (Green).

4. Methodology

4.1. Analysis goals

Our goal is to most accurately predict census block-level median concentrations NO_2 using variations of our geospatial embedding features as outlined below.

4.2. Predictive model framework

Given a set of features $\mathbf{X}$ and a target variable Y representing NO_2 concentrations (units: ppb) at the census block level, our predictive model can be formalized as:

$$\log(Y) = f(\mathbf{X}) + \epsilon \qquad (1)$$

where:

- Y is the NO_2 concentration (units: ppb) we aim to predict, measured at the census block level,
- $\mathbf{X}$ is the feature vector, which can be:
  - $\mathbf{X}_{\text{EI}} \in \mathbb{R}^{390}$ for the Earth Index features, or
  - $\mathbf{X}_{\text{MOSAIKS}} \in \mathbb{R}^{4000}$ for the MOSAIKS features,
  - $\mathbf{X}_{\text{EI + MOSAIKS}} \in \mathbb{R}^{4390}$ for the combined EI and MOSAIKS features,
- f represents the predictive model mapping the feature space to NO_2 concentrations,
- ϵ is the error term capturing the difference between predictions and actual NO_2 concentrations.

The objective of our modeling effort is to minimize the Mean Squared Error (MSE) between the predicted and actual NO_2 concentrations, defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\log((Y_i)) - \log(\hat{Y}_i))^2 \qquad (2)$$

where:

- n is the number of observations,

- Y_i is the actual NO_2 concentration (units: ppb) for the ith observation, measured at the census block level,

- $\hat{Y}_i$ is the predicted NO_2 concentration (units: ppb) for the ith observation.

Minimizing the MSE guides the selection and optimization of our model both in the training and testing phase, ensuring the most accurate predictions of NO_2 concentrations and comparability across our models.

4.3. Data Acquisition and Preprocessing

Our study utilized three main datasets:

1. **Google Street View Car Data:** This dataset includes NO2 median concentration levels for census blocks across the Bay Area, collected via sensors on Google Street View cars over a 32-month period between May 2015 to December 2017. We have 4137 unique values of our target variable. See Figure 2 for its coverage area.
2. **EarthRise Geospatial Embeddings:** Comprising detailed land characteristic measurements from Sentinel-1 and Sentinel-2 satellite imagery, this dataset covers 12 bands including UV intensity and vegetation index for 100m x 100m blocks. We have 390 geospatial embedding features.
3. **MOSAIKS**: Following Rolf et. al's framework and sourced from their project website. 4000 features covering 2990 0.01x0.01 degree tiles.

Data Preprocessing was carried out to clean and normalize the embeddings data. The Earth Index embeddings were originally larger JSON files, but was converted into .shp files, then manipulated using the geoPandas package. The MOSAIKS embeddings were sourced as .csv files, then converted to geoDataFrames, just as the Earth Index embeddings for easier manipulation. Both datasets were geospatially inner joined onto our NO_2 target dataset ahead of the modelling.

4.4. Model Development and Evaluation

Train-Test Split:. The datasets were divided into training and testing sets to evaluate model performance comprehensively. 80% of data being left for training and 20% for testing.

To address the prediction task, we evaluate a range of predictive models that include both linear and non-linear algorithms accessed through the `sklearn` library.

- **Linear Models:** OLS Regression as a baseline model. Ridge and Lasso Regression for their ability to handle multicollinearity through regularization.

- **Ensemble Methods:** Random Forest Regressor, utilizing multiple decision trees to enhance prediction accuracy and control overfitting.

- **Support Vector Machines:** SVR, chosen for its effectiveness in high-dimensional spaces, suitable for the complex nature of geospatial data.

- **Neural Networks:** MLP Regressor, exploring the capability of deep learning techniques to model intricate nonlinear relationships.

4.5. Model Selection and Hyperparameter Tuning

In our training phase, we tune the following models on our training set over set of hyperparameters for optimization. The models and their respective hyperparameters are as follows:

1. `Ridge()`: The model employs Ridge regularization with the regularization strength, `alpha`, varied log-uniformly within the range 10^{-4} to 10^0. The specific hyperparameter tuning strategy for this model involves:
   - `alpha`: `loguniform(1e-4, 1e0)`
2. `Lasso()`: Similar to Ridge, Lasso introduces sparsity into the coefficients, with `alpha` also varied log-uniformly from 10^{-4} to 10^0. Its hyperparameters include:
   - `alpha`: `loguniform(1e-4, 1e0)`
3. `RandomForestRegressor()`: This ensemble method uses multiple decision trees to improve prediction accuracy and control overfitting. The hyperparameters are:
   - `n_estimators`: `randint(100, 500)`
   - `max_depth`: `randint(3, 10)`
   - `min_samples_split`: `randint(2, 11)`
   - `min_samples_leaf`: `randint(1, 5)`
   - `max_features`: Chosen from ['auto', 'sqrt', 'log2']
4. `SVR()`: Employed for its effectiveness in both linear and non-linear predictions, the SVR model's hyperparameters subject to optimization include:
   - `C`: `loguniform(1e-2, 1e2)`
   - `kernel`: Selected from ['linear', 'rbf']
   - `gamma`: `loguniform(1e-4, 1e-1)`
5. `MLPRegressor()`: As a neural network approach, its complexity and learning rate are finely tuned, with hyperparameters including:
   - `hidden_layer_sizes`: Selected from [(50,), (100,), (50, 50), (100, 50)]
   - `activation`: Chosen from ['tanh', 'relu']
   - `solver`: Selected from ['sgd', 'adam']
   - `alpha`: `loguniform(1e-4, 1e-2)`
   - `learning_rate_init`: `loguniform(1e-4, 1e-2)`

4.6. Cross-Validation:

This hyperparameter tuning is executed through `RandomizedSearchCV()`, incorporating cross-validation to ensure that the selection of hyperparameters is not only optimal but also robust across different subsets of the training data. We employ 5-Fold Cross-Validation to evaluate model performance reliably. This technique involves partitioning the data into five subsets, iteratively using one subset for validation and the remaining for training, ensuring each data point is used for both training and validation across folds. This approach allows us to systematically explore a wide hyperparameter space for each model, identifying configurations that yield the best predictive performance.

4.7. Training and Model Evaluation

Post hyperparameter tuning, models are trained on the entire train dataset using the best hyperparameters identified. The Mean Squared Error (MSE) metric is then calculated to quantify prediction accuracy, allowing us to compare models. Each best model as measured by training MSE is then applied to our test set. See figures 1 and 2 for test and training MSEs.

Due to the structured approach of utilizing cross-validation for both hyperparameter tuning and model evaluation, a separate validation set is not explicitly carved out from the dataset. Instead, the cross-validation process ensures each segment of the data is used for validation, offering a comprehensive and robust evaluation of model performance.

5. Discussion: Data & Models

This section delves into the interpretation of the results obtained from our study, focusing on the comparison of model performances, the potential for overfitting, the interpretation of the generated prediction error maps, and the implications of our findings in the context of pollution prediction in a resource constrained environment and the promise of the use of embeddings in future work.

5.1. Analysis of Target Variable Distribution

Original NO_2 Concentrations:. Figure 3 presents the distribution of the original target variable, showing a pronounced skewness towards higher NO_2 concentrations. This skewness can influence the performance of the predictive models, especially those that assume a normal distribution of the target variable, like our 3 linear models: OLS, Ridge, and Lasso.

Log-Transformed NO_2 Concentrations:. To address the skewness observed in Figure 4, we applied a logarithmic transformation to the target variable. Figure 4 demonstrates how this transformation results in a distribution that approximates normality. This motivates our use of a log transformation as seen in equation (1) in Chapter 3.2.

5.2. Interpretation of Model Performance

In this section, we briefly summarize our main model performances across train and test sets, as well as the 3 different combinations of datasets we have available. In addition, we take a short, but important look at the potential for overfitting on these complex datasets.

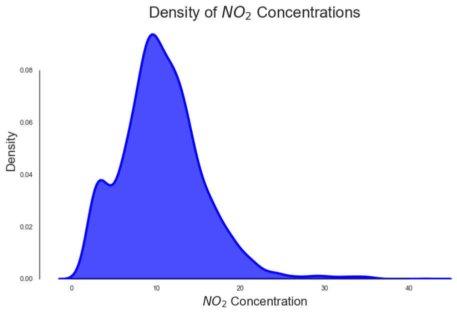

Figure 3: Distribution of our original target variable, NO_2 (ppb)

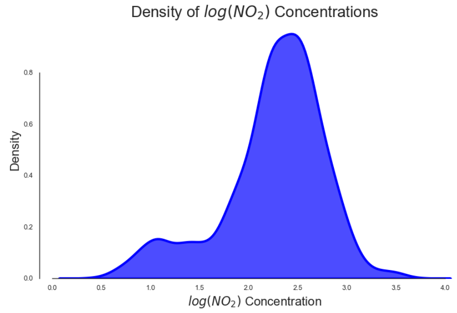

Figure 4: Density Distribution of our target variable, $\log NO_2(ppb)$

5.2.1. Earth Index Dataset Performance

The Support Vector Regression (SVR) model demonstrated the best performance on the Earth Index dataset, with the lowest MSE on the test set among all models evaluated (MSE Test Set: 0.0396). This suggests that SVR, with its specific configuration of $C = 13.1451$ and $\gamma = 0.0062$, is particularly adept at capturing the spatial patterns and complexities inherent in the Earth Index data for predicting NO_2 concentrations. The superior performance of SVR over linear models and even more complex models like MLP and Random Forest indicates its robustness in handling non-linear relationships between features and the target variable.

Table 1: Model Performance using Earth Index: MSE on Training and Test Sets

Model	**MSE Training Set**	**MSE Test Set**
Linear	0.0425	0.0519
Ridge	0.0584	0.0505
Lasso	0.0626	0.0542
Random Forest	0.0539	0.0506
SVR	0.0438	0.0396
MLP	0.0497	0.0509

5.2.2. MOSAIKS Dataset Performance

For the MOSAIKS dataset, Ridge Regression emerged as the most effective model, showcasing its strength in dealing with high-dimensional data and potentially correlated features through regularization (MSE Test Set: 0.0678). The use of Ridge Regression with $\alpha = 0.0254$ implies that adding a

Table 2: Model Performance on MOSAIKS: MSE on Training and Test Sets

Model	MSE Training Set	MSE Test Set
Linear	0.0688	1.5679×10^{14}
Ridge	0.0802	0.0678
Lasso	0.0916	0.0733
Random Forest	0.0797	0.0683
SVR	0.0863	0.0704
MLP	0.1123	0.0916

Table 3: Model Performance on Combined EI + MOSAIKS Features: MSE on Training and Test Sets

Model	MSE Training Set	MSE Test Set
Linear	0.0071	0.0100
Ridge	0.0119	0.0102
Lasso	0.0138	0.0117
Random Forest	4.73×10^{-5}	8.23×10^{-5}
SVR	0.0041	0.0035
MLP	0.0058	0.0025

penalty on the size of coefficients helps to prevent overfitting, making it a reliable choice for datasets with many features like MOSAIKS.

The extraordinarily high MSE observed for the Linear model on the test set (1.5679×10^{14}) highlights the challenges of linear models in handling complex, high-dimensional datasets without regularization.

5.2.3. Combined EI + MOSAIKS Dataset Performance

The combination of Earth Index and MOSAIKS features yielded remarkably improved results, with the Random Forest Regressor showing the best performance (MSE Test Set: 8.23×10^{-5}). This indicates that the ensemble method, capable of capturing non-linear relationships and interactions between a vast number of features, is exceptionally effective when leveraging a rich, combined feature set. The specific configuration of the Random Forest model underscores the importance of tuning hyperparameters to the characteristics of the dataset for optimal performance.

5.2.4. Overfitting Analysis

Indicators of Overfitting:. A model's performance gap between the training and testing datasets can serve as an indicator of overfitting. For instance, the Linear model's results on the MOSAIKS dataset, with its MSE on the test set skyrocketing compared to the training set, suggest overfitting. It indicates the model's tendency to memorize rather than generalize from the training data, capturing noise instead of discerning the underlying relationship (see Table 2). This is typically a sign that the model's complexity is not appropriately counterbalanced by regularization or other methods to prevent it from fitting to the noise.

Best Models and Overfitting:. Conversely, the best performing models in our study, such as the Support Vector Regression (SVR) for the Earth Index and the Random Forest Regressor for the combined dataset, present minimal differences between their training and testing MSEs. This closeness indicates a good generalization capability, implying that these models have learned the underlying patterns without overfitting to the training data (see Tables 1 and 3). The presence of regularization in Ridge Regression for the MOSAIKS dataset also suggests that it has effectively addressed overfitting, hence providing more reliable test set performance.

Table 4: Summary of Best Models for EI, MOSAIKS, and Combined Datasets

Dataset	Best Model Configuration
EI	SVR with $C = 13.1451$ and $\gamma = 0.0062$
MOSAIKS	Ridge Regression with $\alpha = 0.0254$

Combined	Random Forest Regressor with:
	`max_depth = 9,`
	`max_features = 'auto',`
	`min_samples_leaf = 3,`
	`min_samples_split = 6,`
	`n_estimators = 406`

6. Discussion: Spatial Analysis of Prediction Errors

This section provides a detailed analysis of the spatial distribution of prediction errors for our best predictior on the combined datasets, the Random Forest for the NO_2 concentrations in San Francisco, CA. We delve deeper into two maps: one representing the prediction errors on the training dataset and the other on the test dataset. Equivalent maps have been generated for Berkeley and Oakland, CA and can be found in the appendix.

Each bar on the map represents a census block, with the color indicating the magnitude and direction of the error. Bright yellow indicate overprediction, whereas dark blue indicate underprediction following this formula:

$$\text{Prediction Error (PE)} = \log(Y_{\text{obs}}) - \log(Y_{\text{pred}}) \quad (3)$$

To convert the error to percentages over/under the observed values, we use the following formula:

$$\text{PE (\%)} = (e^{\text{PE(Log Units)}} - 1) \times 100 \quad (4)$$

In the maps for San Francico, Berkeley, and Oakland, our axis of log errors run from -0.001 to 0.001, which is approximates to:

For a log scale error of -0.001, the percentage error is:

$$\text{Percentage Error for } -0.001 = (e^{-0.001} - 1) \times 100 \approx -0.1\% \quad (5)$$

Similarly, for a log scale error of 0.001:

$$\text{Percentage Error for } 0.001 = (e^{0.001} - 1) \times 100 \approx +0.1\% \quad (6)$$

In other words, we have relatively tight error bounds (±0.1% for predictions on our original 3 locations. Later in the analysis, we'll compare this to our out-of-sample performance in Palo Alto, Redwood City, and Millbrae, CA.

6.1. Analysis of Training Data Prediction Errors

Figure B.5 displays the prediction errors on the training dataset, using 80% of our data available for San Francisco.

6.1.1. General Trends

Central Urban Accuracy:. The predictive model demonstrates a higher accuracy in the central urban areas of San Francisco, particularly in the downtown sector characterized by a denser grid structure in the upper right quadrant of the maps. This is evidenced by the prevalence of green hues indicating minor prediction errors, suggesting a robust model performance where the density of activities and pollution sources might be higher yet more consistent or well-represented in the training data.

Suburban Area Variability:. Contrasting with the central urban consistency, the suburban regions to the North and South of Golden Gate Park exhibit a more varied error pattern. The lack of a dominant trend in prediction errors, as reflected by a mix of blue and yellow markers, suggests the presence of local factors or conditions not fully captured by the model. Notably, there are distinct clusters where the model consistently underpredicts (indicated by dark blue) or overpredicts (indicated by yellow), which could be indicative of micro-environments or localized emission sources that the model is not able to pick up on.

6.1.2. Connection to underlying NO_2 concentration

Observing the spatial distribution of NO_2 pollution in figure B.15 in the form of a heatmap, we easily deduce that this issue of central urban accuracy and more suburban variability in predictions might be tied to the underlying concentrations we are trying to predict. Downtown San Francisco appears to have higher concentrations of NO_2 than the two other areas, and at the same time more accurate predictions. However, if we zoom into the map of downtown San Francisco, as we do in figure B.16, we see something interesting: Places that exhibit variability are census blocks closer to higher polluted areas like along the beginning of Columbus Avenue, Chinatown, and Transbay / SOMA.

6.1.3. Considering test errors

As seen in figure B.6, it's harder to extract patterns in prediction errors in a smaller spatial sample. However, the variability in errors in the areas North and South of Golden Gate Park seem consistent with the training set. In contrast, the errors seem more variable in downtown San Francisco than in the training set: Census blocks close to those with low prediction errors in the training set, like the the western end of SOMA, now exhibit higher prediction errors.

6.1.4. Implications for Model Improvement

This, together with the overall variability in errors, leaves room for discussion as to what information one loses by not taking into the consideration the predictions of neighboring census blocks. Future research could incorporate this by potentially building stacked ensembles, or similar models.

6.2. Applying our model to unseen data: Palo Alto, Redwood City, and Millbrae, CA

In this section, we turn our attention to what seems to be our most promising model taking data availability into concern. The accuracy of our best Random Forest Regressor applied to our combined Earth Index and MOSAIKS embeddings is potentially outweighed by the lack of readily available Earth Index embeddings online. The MOSAIKS, following the Rolf et al. framework are easily accessed on the project website. (4) Here, we apply our best model on the MOSAIKS dataset, Ridge Regression with $\alpha = 0.0254$, 3 cities in the South-Western part of the Bay Area that was also included in the Google Street View pollution project by Apte et al. (16). Across our 3 sites, our model achieves a test MSE of 0.1117.

Model	Mean Squared Error (MSE)
Ridge Regression	0.1117

6.2.1. Widening error bars on fully unseen data

The first observation when mapping our prediction errors in our three new cities, as seen in Figures B.12, B.11, and B.13, is the widening of our error bars from ±0.001 on the logarithmic scale to ±0.1. This adjustment significantly impacts the interpretation of prediction accuracy, expanding the potential error margin. A quick 'on-the-napkin' calculation helps contextualize this change in terms of percentage error. Converting log-scale errors to percentages using the exponential function yields the following insights:

For an error of -0.1 in log scale, the percentage error is:

$$\text{Percentage Error for } -0.1 = (e^{-0.1} - 1) \times 100 \approx -9.52\% \quad (7)$$

Similarly, for an error of 0.1 in log scale:

$$\text{Percentage Error for } 0.1 = (e^{0.1} - 1) \times 100 \approx +10.52\% \quad (8)$$

Thus, the widening of error bars from ±0.001 to ±0.1 on the logarithmic scale effectively translates to a shift from errors marginally affecting predictions to potentially influencing them by up to ±10%, marking a substantial increase in the potential deviation from actual values. This is important to consider as we delve deeper into the predictions in one of our 3 cities, Palo Alto.

6.2.2. Prediction errors in Palo Alto

Reaping the benefits from a richer test dataset (all data from Palo Alto is used for testing), we make an attempt in interpreting the patterns of prediction errors in Palo Alto. Looking at figure B.11, it is evident that the Eastern side (right side) of the map mostly suffers from overpredictions all hovering around 10%, whereas the Western side (left side) of the city has higher variability in the direction of the errors. On that side, following Middlefield road (around the middle of the map), we see a stripe of underprediction around 10%. This appears to be alongside a busy thoroughfare, but as shown in figure B.14, it is not the most polluted area, which would be the Bayshore Highway entrances in the upper right corner. A potential, yet untested explanation for this could be that the algorithm has picked up on medium-sized roads as less polluted than what they in fact are, and therefore systematically underpredict an area like this. A quick search with Google Earth Engine shows that this area is also populated with a lot of trees, which could lend itself to another explanation: Could the algorithm have picked up on features representing trees and deem them as usually less polluted areas? Both potential explanations speak to the complexeties in interpretation when dealing with uninterpretable features in larger, more complex algorithms. At the very best, we're left with guesses.

7. Discussion: Strategic Resource Allocation

7.1. Setting up the problem

We imagine a social planner that aims to maximize knowledge about local NO_2 air pollution and use this knowledge to minimize local health impacts. This social planner is constrained by funds to monitor air pollution through widespread installment of sensors, as well as costs for mitigation. We posit that our model may help increase knowledge of the potential general distribution of pollution in unmonitored areas, as well informing targeted interventions in high-pollution hotspots. Informed decisions should be made, and the widening error bars as evident in our models performance in Palo Alto, Redwood City, and Millbrae should be taken into consideration.

7.2. Maximizing knowledge

The social planner could apply our best performing model to their Bay Area district, and we estimate that the prediction will be within the ±10% error range. Especially areas with lower rates of pollution monitoring could benefit from this, as well as places were existing monitoring solutions are limited to high-polluted areas, like the EPA's sensors in San Francisco. A more spatially granular picture could be developed relative to existing solutions that do not provide census-block NO_2 median concentrations. To further enhance this, we suggest collecting more spatially granular in-situ data similar to our Google Street View dataset, as well as adding a temporal component to the data, which we discuss further below.

7.3. Minimizing harm

We assume that within a limited resource setting, the social planner will have funds left over from reduced monitoring costs using our model. Particularly as our model depends on fully public data, be it either MOSAIKS or Google Street View data published by researchers. Given this, such funds could be directed towards harm-reducing solutions, especially in areas that the model has identified as high-pollution zones, like highway intersections or census blocks close to busy thoroughfares. Without going into too much details, solutions could include regulations against idling, subsidies towards indoor air purifiers in private and public spaces, with more.

7.4. Ensuring impact

To ensure the impact of solutions like these, accessibility and interpretability of our models is of paramount importance. As these models improve in scale and accuracy, online dashboards updated frequently could be of benefit to a social planner, be it of local or national importance. Another way one can ensure impact is to allow prediction like these to improve on existing frameworks that policymakers are used to, like the Cal EnviroScreen map, which incorporates socio-economic factors as well as multiple forms of pollution to map out environmental justice in a given census tract. Furthermore, this modeling framework could be expanded to other sources of pollutants (like Black Carbon, NO, SO_2, or PMs). Widening the scope, while ensuring the quality of predictions could increase the potential for impact on a social planner's decisions when resources are scarce.

8. Discussion: Limitations of Model

While our model advances the prediction of air pollution concentrations, it does not incorporate socio-economic factors or account for temporal variations in pollution levels. This has the potential to improve prediction models, increase generalizability, and inform better decision making. Adding a temporal aspect to the predictions may also increase the usefulness of our solutions, giving way to increased impact among social planners. A last obvious limitation is the uncertainty associated with predicting on totally unseen data. We suggest ways to reduce this risk, but acknowledge the presence of irreducible errors.

8.1. Socioeconomic Factors

As earlier mentioned in 7.4, our model solely depends on remote sensing data for its features, and does not take into the account of socioeconomic factors like median age, income, education, nor the demographics of the census blocks. Our current work has limited itself to focus on the impact geospatial embeddings can have on future prediction pollution, yet this does not mean that potential correlates between our local air pollution and such factors does not exist nor does it negate its potential importance in improving the prediction models. Such information could be easily integrated through governmental data sources such as the aforementioned Cal EnvironScreen tool.

8.2. Temporal Aspect

As described above, our current model predicts median NO_2 concentration in census blocks as measured over a 6-months period. However, geospatial embeddings has the potential to be generated ad-hoc or updated regularly following existing frameworks such as Rolf et al. (4). Attempts were made in earlier phases of this research project to source temporally corresponding NO_2 measurements, that is, a panel dataset of geospatial embeddings and our pollutant. The TROPOMI NO_2 instrument as presented in Goldberg et. al is particularly promising and worthwhile of further exploration. (19) With its higher spatial granularity and hourly updates, the upcoming TEMPO instrument may also be a way to accomplish both high spatial- and temporal granularity in our NO_2 predictions. (20) Currently, though, there seems to be a trade-off between spatial and temporal granularity: Both the TEMPO (4.7×2.1 km) and TROPOMI (7×3.5 km) instruments are updated frequently, but at much lower spatial granularity than the Google Street View car used in this study. This, on the other hand, may have trouble covering the same area on a hourly basis. However, solutions such as installing NO_2 and sensors for other pollutants on fleets of autonomous vehicle fleets (e.g the Google-backed company Waymo) might be a solution to remove the trade-off between the spatial and temporal granularity of our target variable.

8.3. Uncertainties in Predictions for Unseen Areas

The extrapolation of model predictions beyond the geographical scope of the training data introduces a degree of variability that must be carefully managed. Not placing trust in the predictions may reduce the possibility for positive impact as aforementioned, while doing the opposite may misinform policies impacting millions of people. To navigate these uncertainties, social planners can adopt a multi-tiered approach, combining model predictions with periodic ground-truthing exercises. One way could be to randomly sample predictions in unseen areas, and then monitoring the actual median concentrations over a time period. This has the potential to reduce monitoring costs, while ensuring the truthfulness of our model. New data could then be incorporated to our model to further enhance new predictions in yet an unseen area. Another aspect not yet mentioned, as seen in figure 2, is the possibility of applying dimensionality reduction algorithms to separate out observations across cities, using the inherent similarities that exist across city boundaries. Further improvements to the models could take this into account, for example by training models on clusters of similar areas, then applying that very model on a unseen, yet similar cluster (as identified by the embeddings). This is merely suggestive, but could be a worthwhile endeavour in future research.

9. Conclusion

In this study, we developed a predictive model using geospatial embeddings to estimate NO_2 pollution levels across the Bay Area. By integrating data from Google Street View cars and satellite imagery through Earth Index and MOSAIKS embeddings, we demonstrated the potential of machine learning in improving pollution prediction accuracy. Our results highlight the effectiveness of Support Vector Regression (SVR) and Ridge Regression models, with a combined approach using a Random Forest Regressor showing the best performance.

However, the application of the model to out-of-sample areas revealed significant prediction accuracy variability, emphasizing the importance of cautious model extrapolation to unmonitored regions. This research underlines the utility of predictive modeling in identifying pollution hotspots and supporting targeted environmental interventions. Future work should focus on incorporating temporal data and socio-economic factors to enhance model accuracy and applicability, moving towards more dynamic and comprehensive pollution monitoring solutions.

References

[1] Y. O. Khaniabadi, G. Goudarzi, S. M. Daryanoosh, A. Borgini, A. Tittarelli, and A. De Marco, “Exposure to pm 10, no 2, and o 3 and impacts on human health,” *Environmental science and pollution research*, vol. 24, pp. 2781–2789, 2017.

[2] V. A. Southerland, S. C. Anenberg, M. Harris, J. Apte, P. Hystad, A. van Donkelaar, R. V. Martin, M. Beyers, and A. Roy, “Assessing the distribution of air pollution health risks within cities: a neighborhood-scale analysis leveraging high-resolution data sets in the bay area, california,” *Environmental health perspectives*, vol. 129, no. 3, p. 037006, 2021.

[3] S. Costa, J. Ferreira, C. Silveira, C. Costa, D. Lopes, H. Relvas, C. Borrego, P. Roebeling, A. I. Miranda, and J. P. Teixeira, “Integrating health on air quality assessment—review report on health risks of two major european outdoor air pollutants: Pm and no2,” *Journal of Toxicology and Environmental Health, Part B*, vol. 17, no. 6, pp. 307–340, 2014.

[4] E. Rolf, J. Proctor, T. Carleton *et al.*, “A generalizable and accessible approach to machine learning with global satellite imagery,” *Nat Commun*, vol. 12, p. 4392, 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-24638-z

[5] B. C. Singer, A. T. Hodgson, T. Hotchi, and J. J. Kim, “Passive measurement of nitrogen oxides to assess traffic-related pollutant exposure for the east bay children’s respiratory health study,” *Atmospheric environment*, vol. 38, no. 3, pp. 393–403, 2004.

[6] S. Chambliss, C. P. Pinon, K. Messier, B. LaFranchi, C. Upperman, and M. Lunden, “Bay area mobile monitoring multi-pollutant block medians,” https://doi.org/10.6084/m9.figshare.15070314.v1, 2021.

[7] C. Kruse, E. Boyda, S. Chen, K. Karra, T. Bou-Nahra, D. Hammer, J. Mathis, T. Maddalene, J. Jambeck, and F. Laurier, “Satellite monitoring of terrestrial plastic waste,” *PloS one*, vol. 18, no. 1, p. e0278997, 2023.

[8] P. Muthukumar *et al.*, “Predicting pm2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data,” *Air quality, atmosphere, & health*, vol. 15, no. 7, p. 1221–1234, 2022.

[9] H. S. Lim, M. Z. MatJafri, K. Abdullah, and C. J. Wong, “Air pollution determination using remote sensing technique,” https://doi.org/10.5772/8319, 2009.

[10] J. A. Kamińska, “A random forest partition model for predicting no2 concentrations from traffic flow and meteorological conditions,” *Science of The Total Environment*, vol. 651, no. Part 1, pp. 475–483, 2019.

[11] R. Yan *et al.*, “Multi-hour and multi-site air quality index forecasting in beijing using cnn, lstm, cnn-lstm, and spatiotemporal clustering,” *Expert Systems with Applications*, vol. 169, p. 114513, 2021.

[12] U. Pak *et al.*, “A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction,” *Air Qual Atmos Health*, vol. 11, p. 883–895, 2018.

[13] C. Kruse *et al.*, “Satellite monitoring of terrestrial plastic waste,” *arXiv preprint arXiv:2204.01485*, 2022.

[14] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, “Global land use / land cover with sentinel 2 and deep learning,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 4704–4707.

[15] E. V. Novotny, M. J. Bechle, D. B. Millet, and J. D. Marshall, “National satellite-based land-use regression: No2 in the united states,” *Environmental Science & Technology*, vol. 45, no. 10, pp. 4407–4414, 2011.

[16] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg, “High-resolution air pollution mapping with google street view cars: Exploiting big data,” *Environmental Science Technology*, vol. 51, no. 12, pp. 6999–7008, 2017.

[17] B. Zhang, Y. Rong, R. Yong, D. Qin, M. Li, G. Zou, and J. Pan, “Deep learning for air pollutant concentration prediction: A review,” *Atmospheric Environment*, vol. 290, p. 119347, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231022004125

[18] G. Hoek, B. Brunekreef, S. Goldbohm, P. Fischer, and P. A. van den Brandt, “Association between mortality and indicators of traffic-related air pollution in the netherlands: a cohort study,” *The Lancet*, vol. 360, no. 9341, pp. 1203–1209, 2002.

[19] D. L. Goldberg, S. C. Anenberg, G. H. Kerr, A. Mohegh, Z. Lu, and D. G. Streets, “Tropomi no2 in the united states: A detailed look at the annual averages, weekly cycles, effects of temperature, and correlation with surface no2 concentrations,” *Earth’s Future*, vol. 9, no. 4, p. e2020EF001665, 2021. [Online]. Available: https://doi.org/10.1029/2020EF001665

[20] The Center for Astrophysics — Harvard & Smithsonian, “Tempo instrument captures its first images of air pollution over greater north america,” https://www.cfa.harvard.edu/news/2023-08-$24_TEMPO_Instrument_Captures_Its_First_Images, Aug2023, accessed$: $2024 - 04 - 02$.

Appendix A. Supplementary Code

Appendix A.1. train_and_evaluate

The ‘train_and_evaluate‘ function iterates through a list of models and their associated hyperparameters (if any), trains each model on the training data, and evaluates its performance on both the training and testing datasets.

```python
def train_and_evaluate(X_train, y_train, X_test,
    y_test, models, n_iter=10):
    results = {}
    for name, model_info in models:
        print(f"Training and evaluating {name}...")
        if "hyperparams" in model_info:
            random_search = RandomizedSearchCV(
    model_info["model"],
                model_info["hyperparams"],
                n_iter=n_iter, cv=5,
                scoring='neg_mean_squared_error',
                n_jobs=-1, random_state=42, verbose
    =1)
            random_search.fit(X_train, y_train)
            best_model = random_search.
    best_estimator_
        else:
            model_info["model"].fit(X_train, y_train)
            best_model = model_info["model"]

        predictions_train = best_model.predict(
    X_train)
        mse_train = mean_squared_error(y_train,
    predictions_train)
        predictions_test = best_model.predict(X_test)
        mse_test = mean_squared_error(y_test,
    predictions_test)

        results[name] = {"mse_train": mse_train, "
    mse_test": mse_test}
    return results
```

Appendix B. Maps

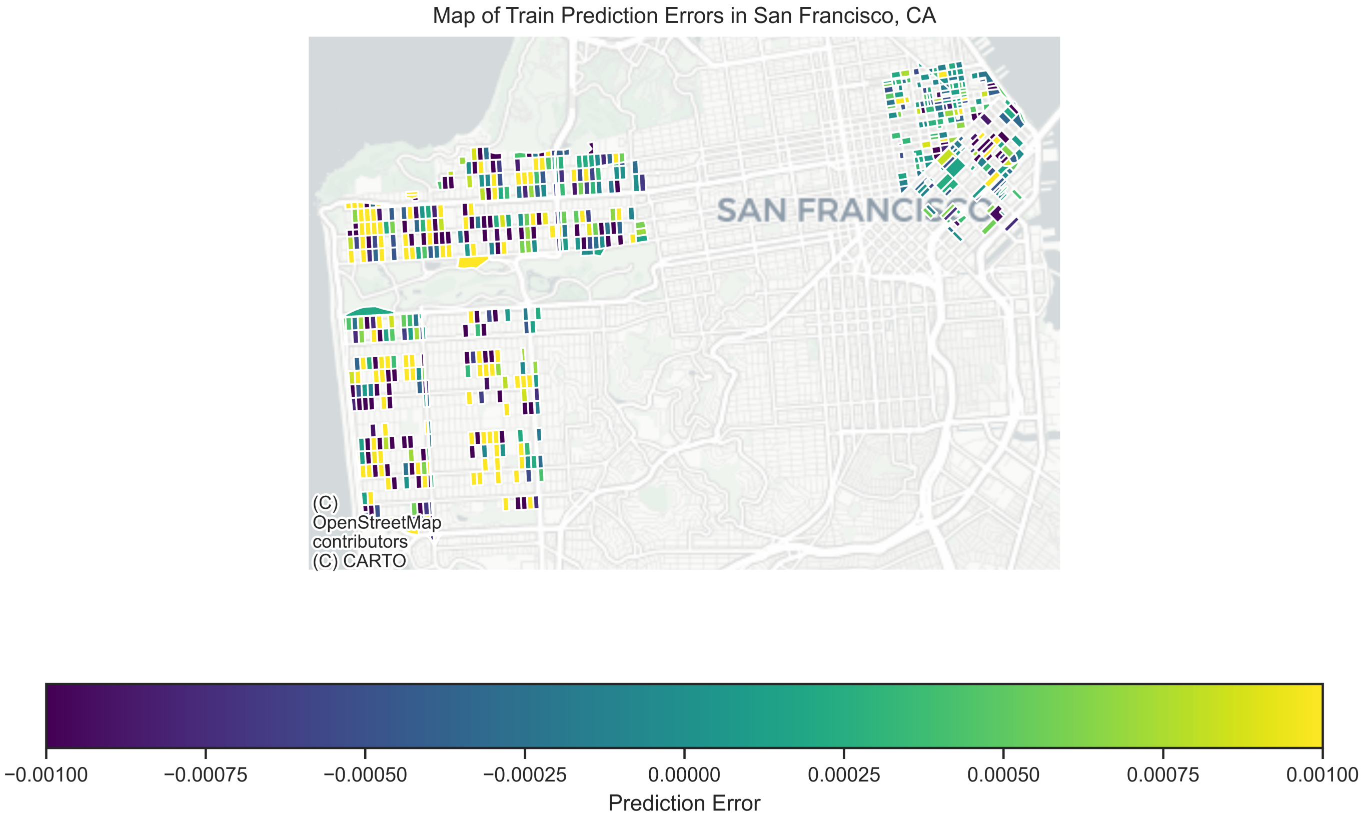

Figure B.5: Train error map for San Francisco, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

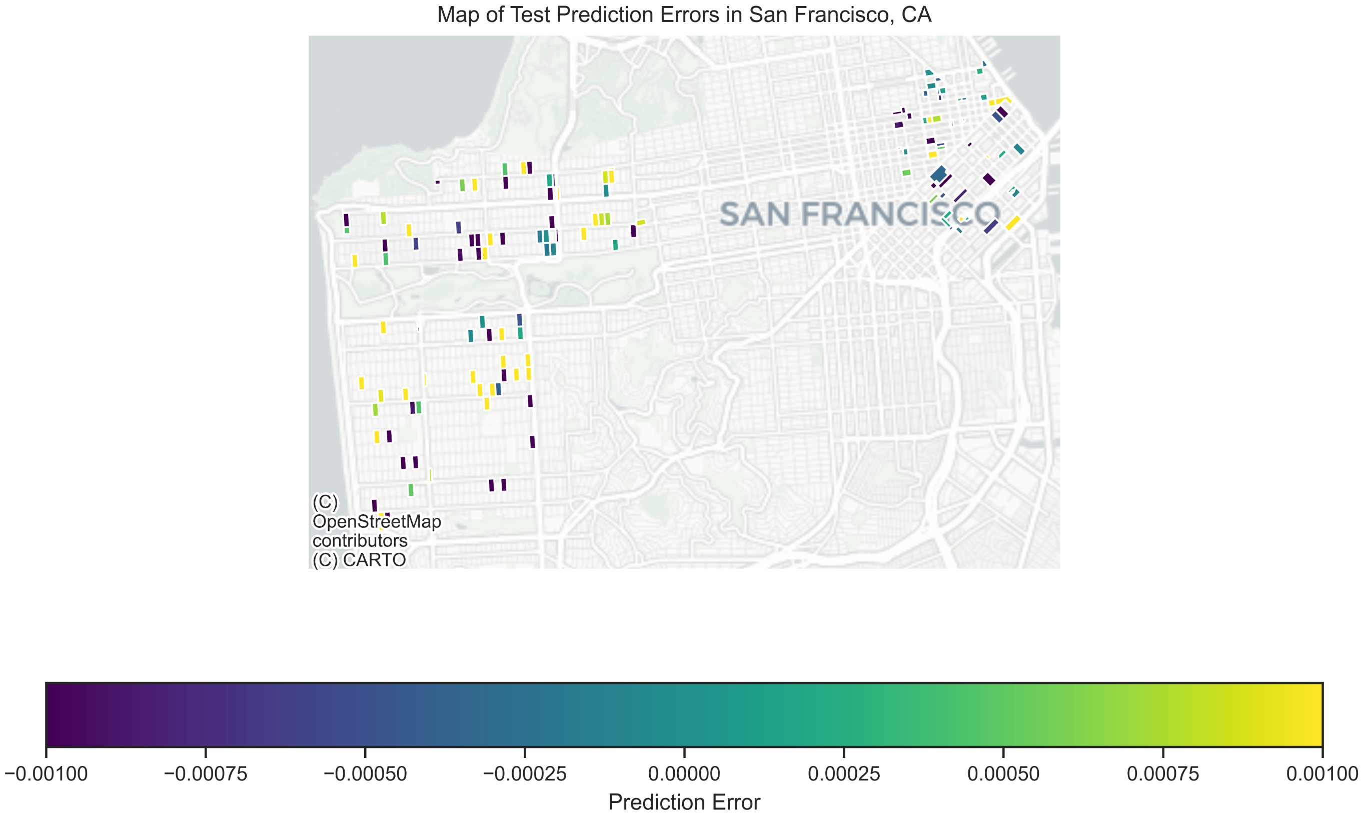

Figure B.6: Test error map for San Francisco, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

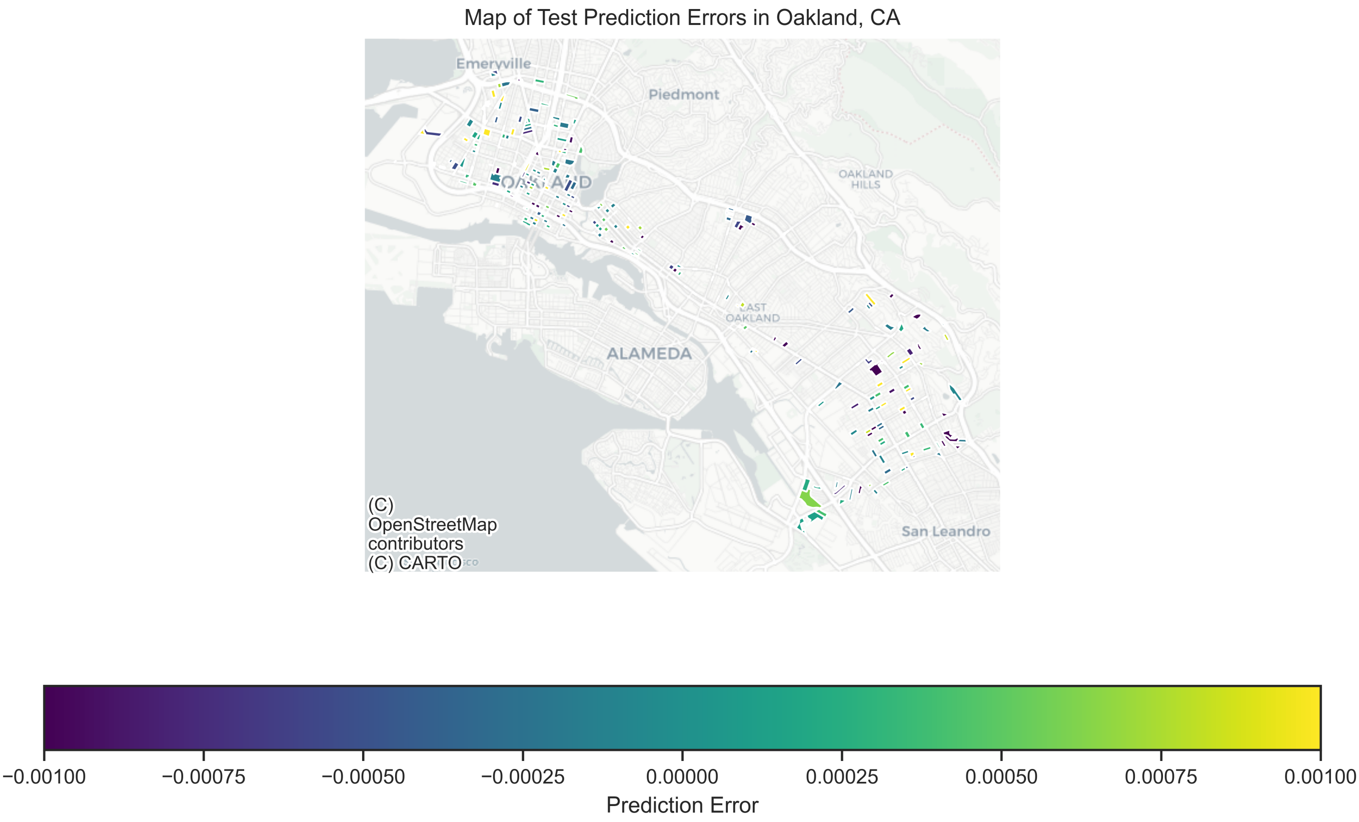

Figure B.7: Test error map for Oakland, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

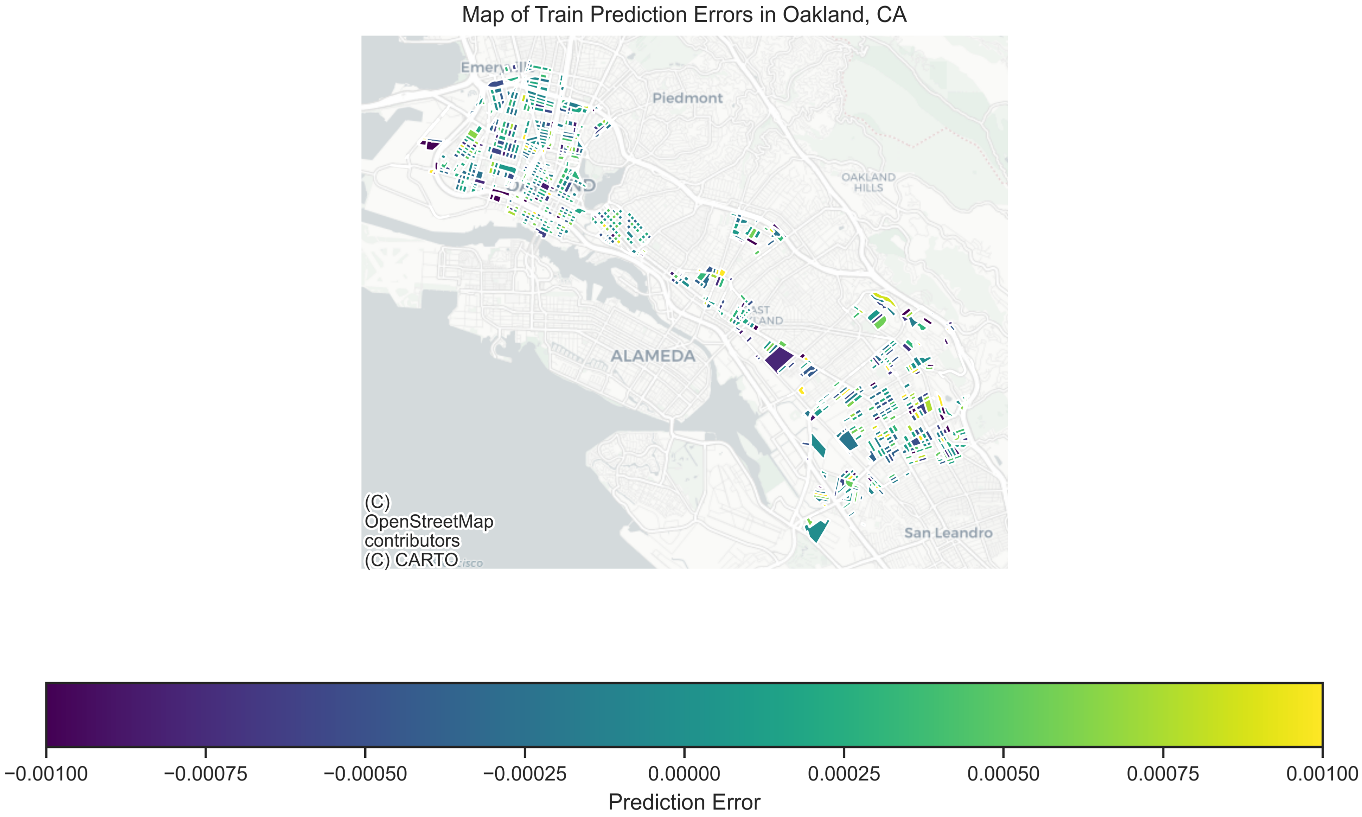

Figure B.8: Train error map for Oakland, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

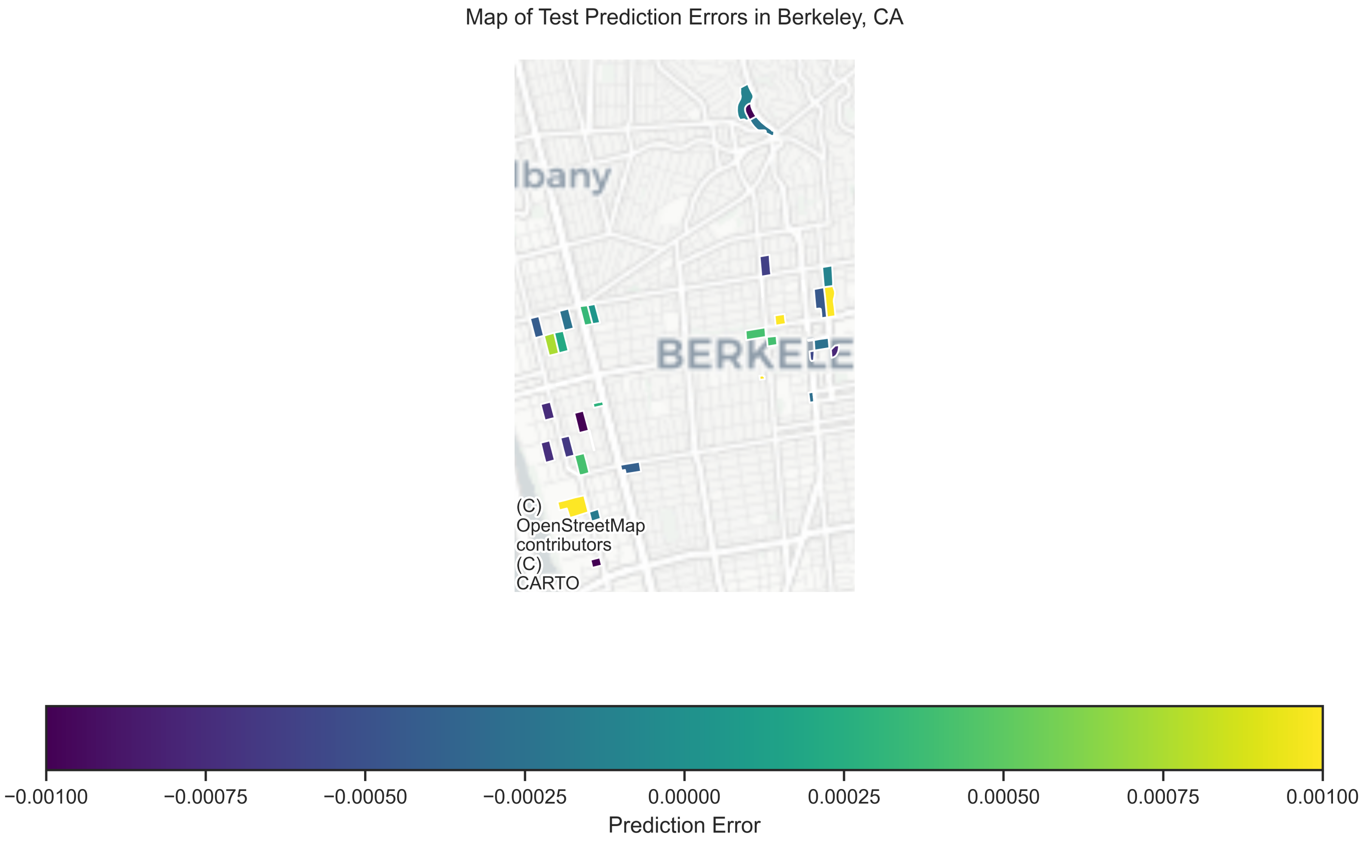

Figure B.9: Test error map for Berkeley, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

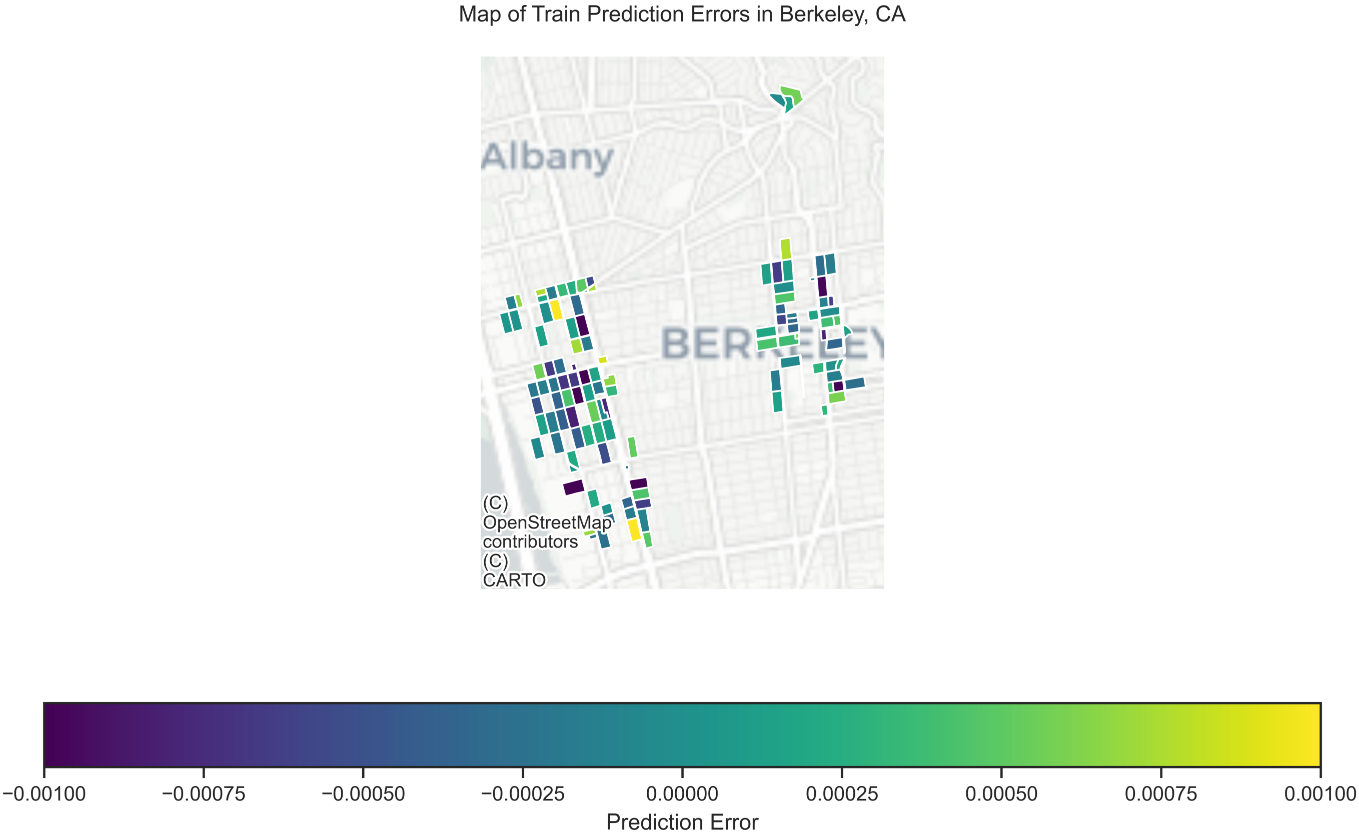

Figure B.10: Train error map for Berkeley, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

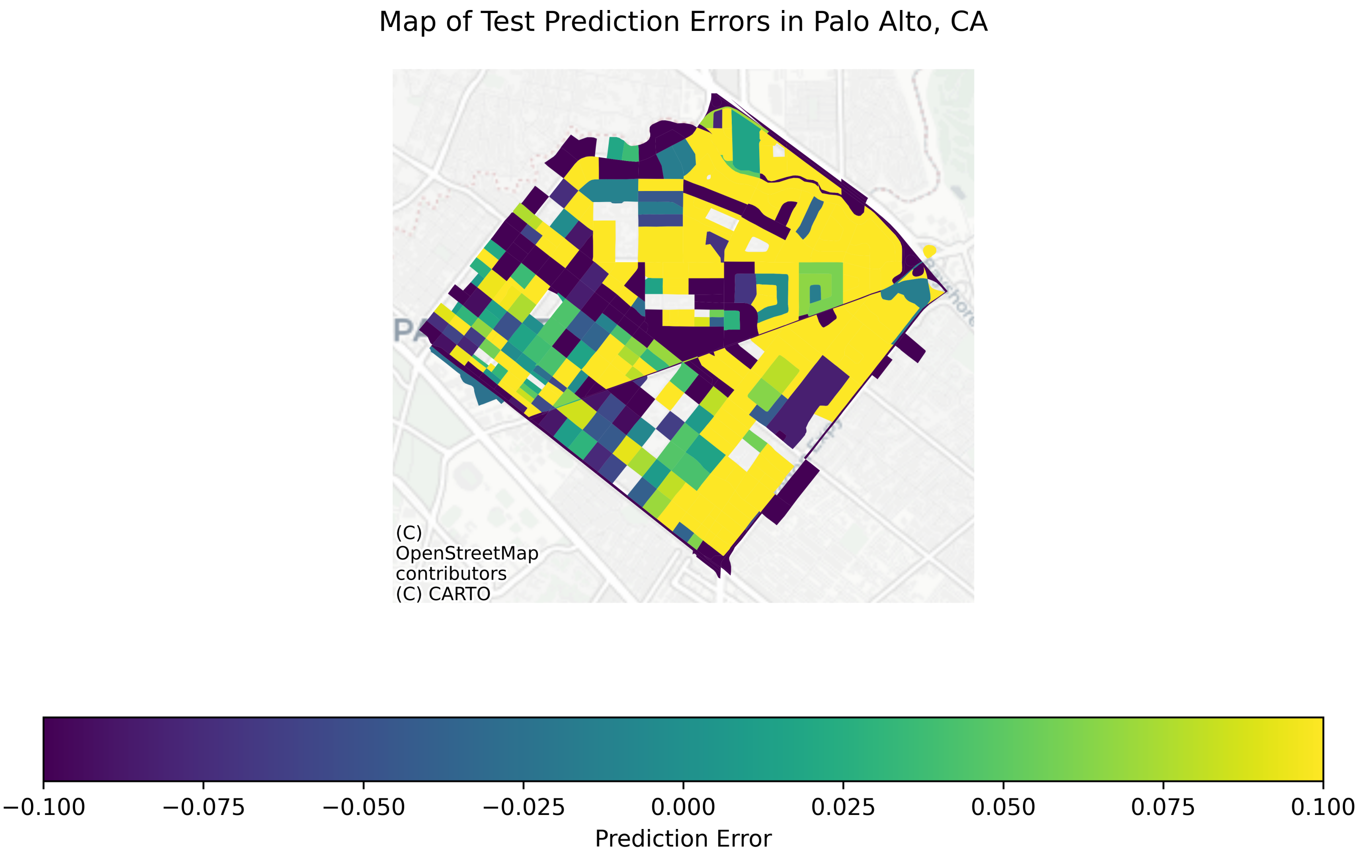

Figure B.11: Test error map for Palo Alto, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

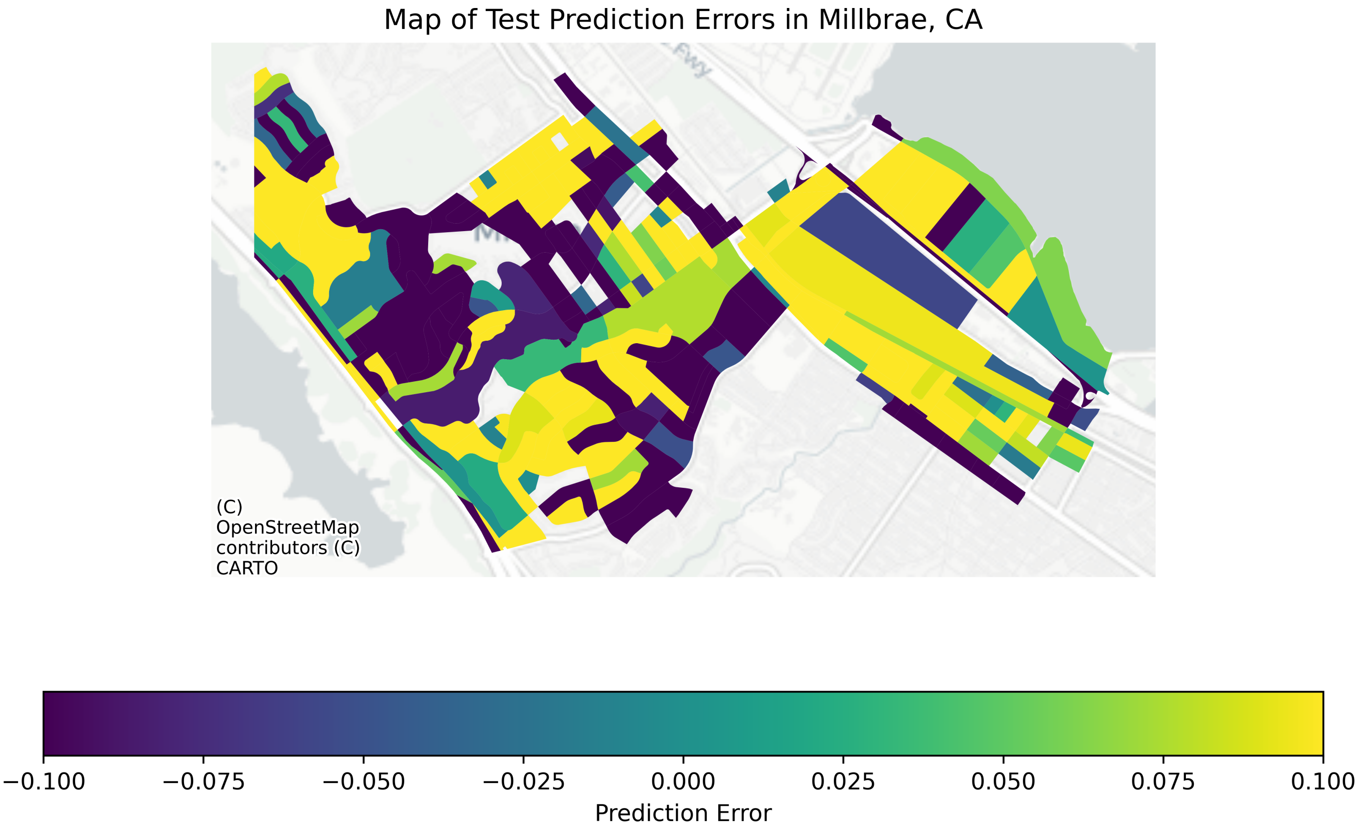

Figure B.12: Test error map for Millbrae, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

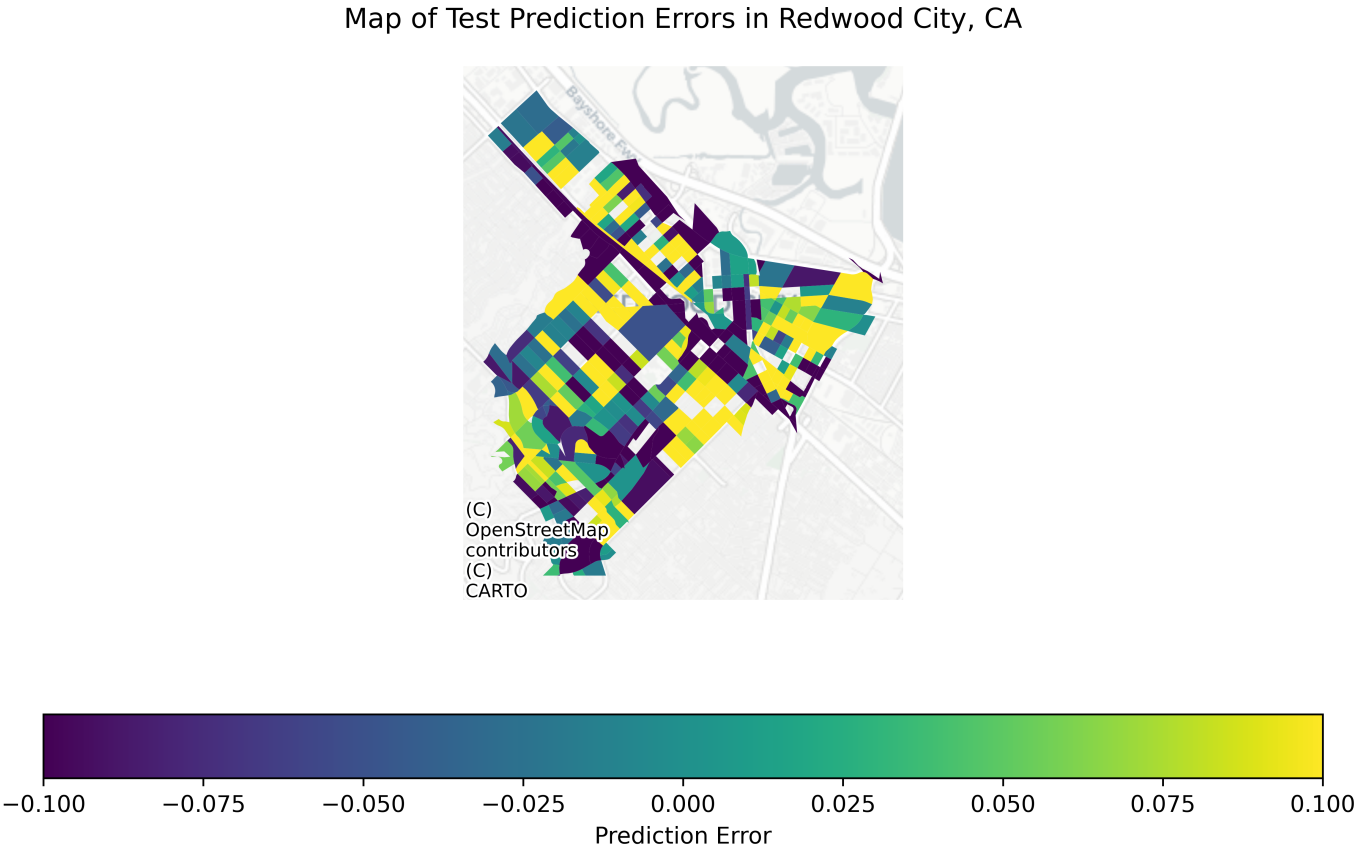
Map of Test Prediction Errors in Redwood City, CA
Bayshore Fwy
(C) OpenStreetMap contributors (C) CARTO
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
Prediction Error

Figure B.13: Test error map for Redwood City, CA showing the spatial distribution of $\log\left(\frac{NO2_{obs}}{NO2_{pred}}\right)$ concentration errors.

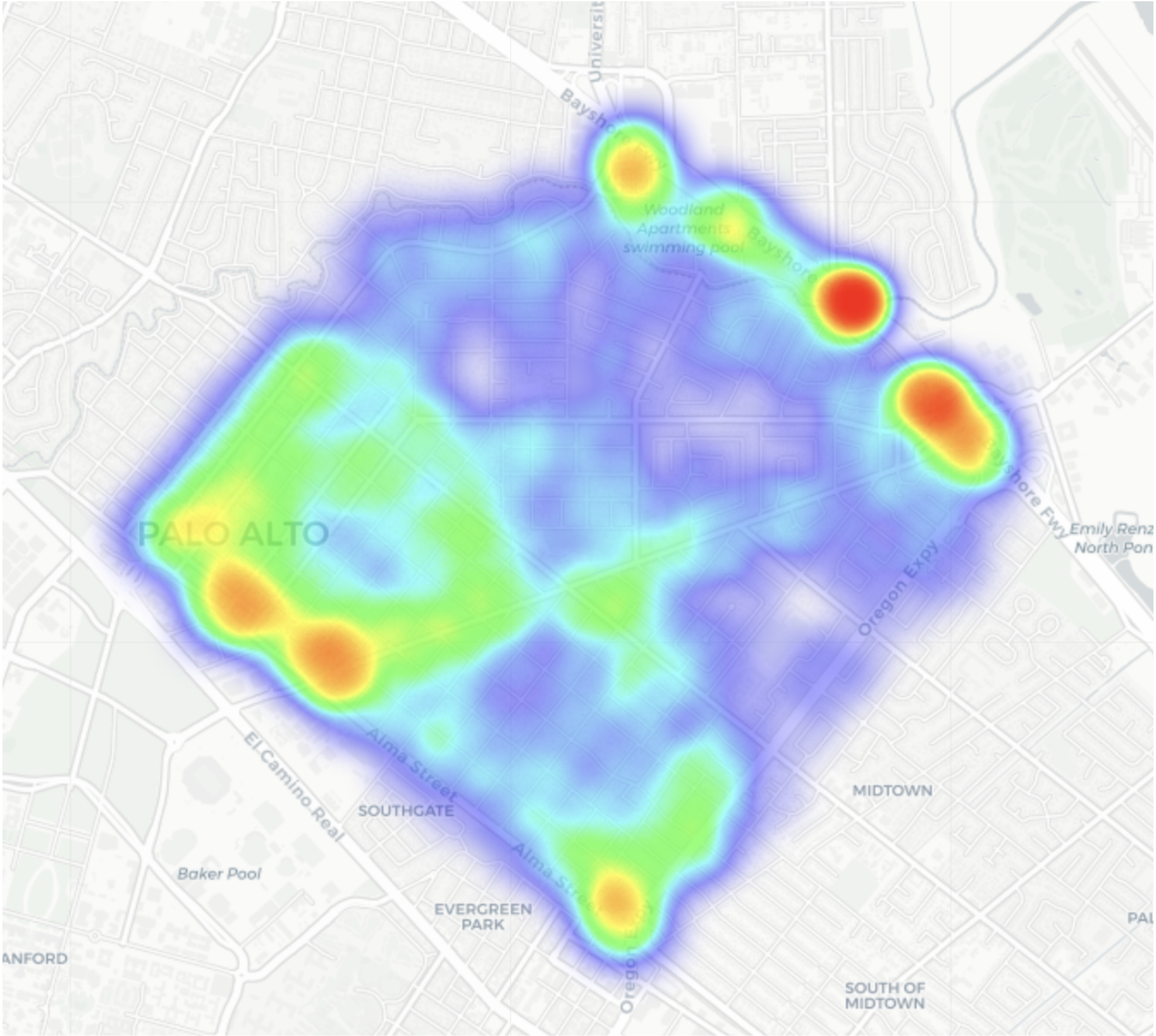

Figure B.14: Heatmap for Palo Alto, CA showing the spatial distribution of NO_2 concentrations.

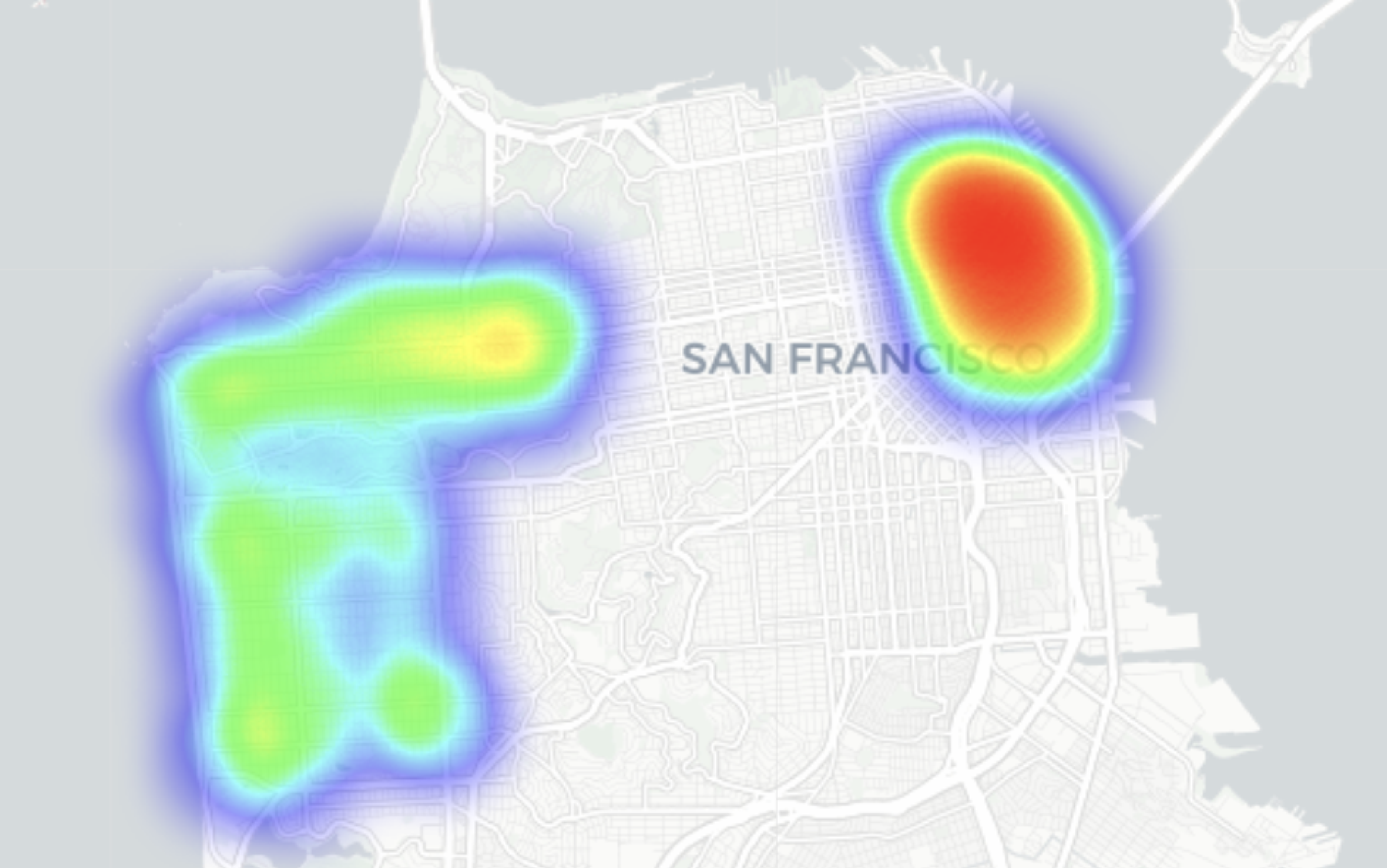

Figure B.15: Heatmap for San Francisco, CA showing the spatial distribution of NO_2 concentrations.

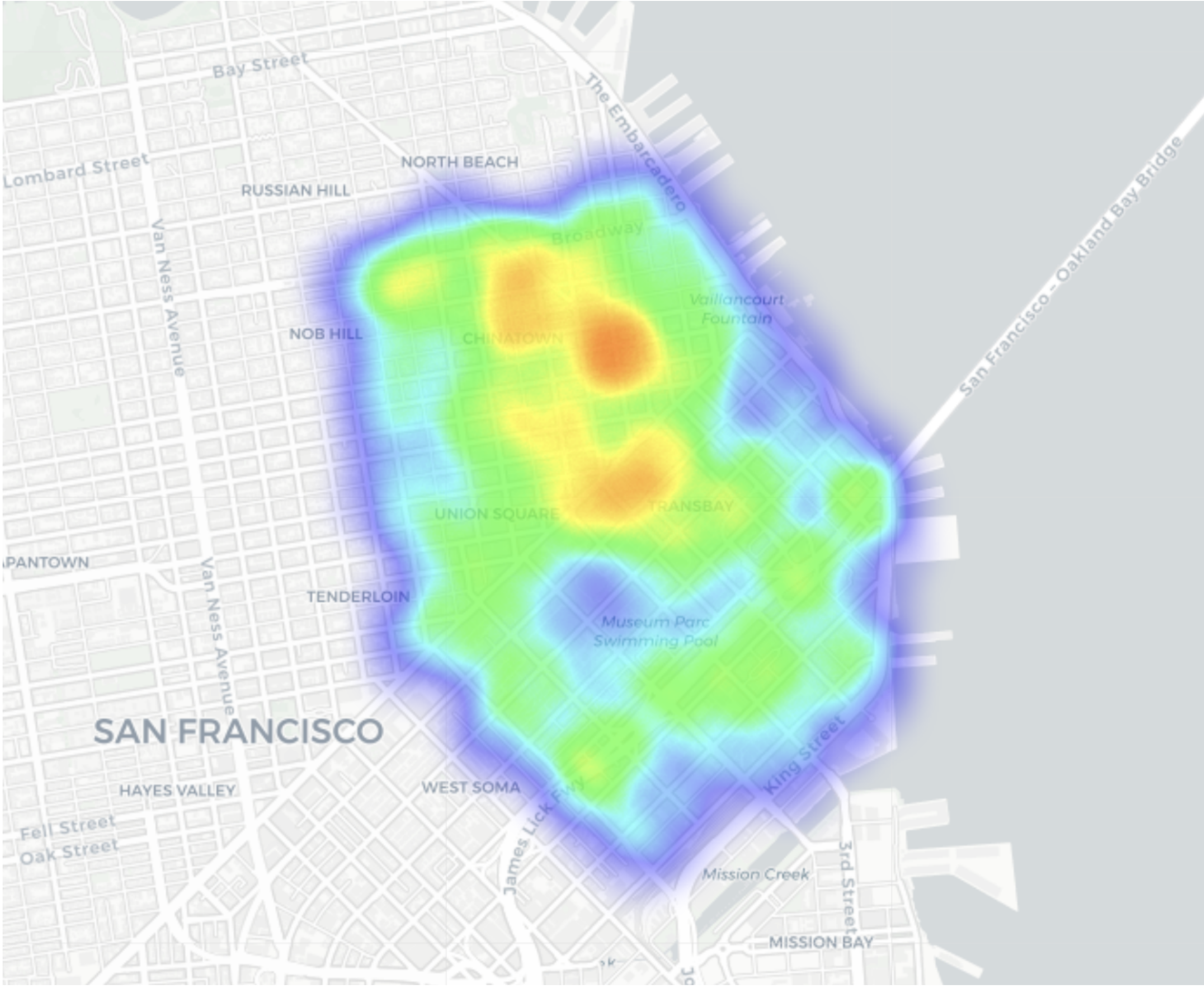

Figure B.16: Zoomed in Heatmap for Downtown San Francisco, CA showing the spatial distribution of NO_2 concentration.