Imperial College London

Department of Earth Science and Engineering

MSc in Environmental Data Science and Machine Learning

Independent Research Project

Final Report

Exploring Land Use Effects on Intraurban CO_2 using Machine Learning Algorithms for Urban Decarbonization

by

Anna Carina Smith

Email: anna.smith23@imperial.ac.uk

GitHub username: edsml-acs223

Repository: https://github.com/ese-msc-2023/irp-acs223

Supervisors:

Fangxin Fang

Linfeng Li

Jiansheng Xiang

September 2024

Table of Contents

ABSTRACT

Carbon dioxide (CO_2) is a key driver of anthropogenic climate change, and cities have been identified as major sources of emissions. Both urbanization and land use change are positively correlated with rising CO_2 emissions, highlighting the need to study spatiotemporal trends in CO_2 to better inform sustainable city planning and urban decarbonization strategies. This study is the first to investigate the viability of using land use regression (LUR) to predict intraurban CO_2 concentrations in the San Francisco Bay Area using data from the BEACO_2N monitoring network. Furthermore, LUR is compared to machine learning (ML) and deep learning (DL) algorithms that explore non-linear relationships, representing a two-fold novel contribution to the literature. Model performance is evaluated using reserved data from training CO_2 sensors, as well as data from unseen sensor locations. The highest predictive accuracy is achieved using extreme gradient boosting (XGBoost) and a convolutional neural network (CNN), both with R^2 values of 0.58, outperforming traditional LUR, which achieved an R^2 of 0.34. XGBoost and CNN also outperformed traditional LUR for unseen sensor locations, accounting for up to 42% of the variability in observed CO_2 concentrations. The performance of the models is constrained by the inconsistent validity of the CO_2 data, as well as the limited availability of environmental feature data with fine temporal resolution.

INTRODUCTION

Intraurban Carbon Dioxide

Urbanization is a key trend in the twenty-first century, and 70% of energy-related CO_2 emissions globally are associated with urban areas (Intergovernmental Panel On Climate Change (IPCC), 2023). Currently, over half of the world's population lives in cities, and by 2100 this rate is projected to increase up to 80-90% (Riahi et al., 2017). Carbon dioxide (CO_2) is a critical greenhouse gas (GHG) that is produced during the combustion of fossil fuels and is a key driver of anthropogenic climate change. Research has identified positive correlations between urbanization and CO_2 emissions (Poumanyvong and Kaneko, 2010; Wang, 2018). Therefore, cities around the world must be treated as critical contributors to climate change, and increased efforts should be dedicated towards understanding and mitigating their climate impact. Better understanding trends and variability in intraurban CO_2 is important to inform decision-makers in the development of urban decarbonization strategies (Mitchell et al., 2018; Wang, 2018).

Intraurban CO_2 demonstrates considerable spatiotemporal variability. The heterogeneous nature of urbanization and land use activities results in heterogeneous landscapes of ambient CO_2 concentrations within cities (Wang, 2018). Urban areas have been found to be net sources of CO_2, with distinct seasonal and diurnal fluxes: urban CO_2 is highest in the mornings and lowest in the evenings, with summer fluxes being less pronounced than in the winter (Coutts et al., 2007; Velasco et al., 2005). An increase in wintertime concentrations can be linked to more heating fuel combustion and less vegetation cover (Bergeron and Strachan, 2011; Coutts et al., 2007). Patterns in ambient CO_2 are also closely related to traffic volumes, with an increase in emissions observed during rush hours periods (Coutts et al., 2007). Finally, suburban areas tend to have lower ambient concentrations than urban cores (Bergeron and Strachan, 2011; Velasco and Roth, 2010). Mitchell et al. (2018) identified an increase in emissions resulting from suburban development and population growth in rural areas. These studies illustrate the spatiotemporal variability of intraurban CO_2 and its association with urbanization and land use change.

Unlike air pollutants, ambient CO_2 is not commonly measured using sensors at the intraurban level. Most often, CO_2 emissions are calculated to attribute responsibility to governmental or corporate entities using aggregated emissions or energy consumption data (Duren and Miller, 2012; Mitchell et al., 2018). Efforts to establish monitoring networks are rising, and research in the past decade indicates the deployment of urban CO_2 sensors primarily in the U.S. (Bréon et al., 2015; Briber et al., 2013; Duren and Miller, 2012; Lauvaux et al., 2016; Mitchell et al., 2018; Rice and Bostrom, 2011). Most notably, the Megacities Carbon Project represents significant effort to establish long-term multisite CO_2 monitoring networks in megacities around the world (Duren and Miller, 2012); however, many existing networks are limited in the number of nodes as some feature only between one and five sensors (Bréon et al., 2015; Briber et al., 2013; Helfter et al., 2016; Mitchell et al., 2018; Rice and Bostrom, 2011). Most networks have been established recently, providing a limited historical record of ambient CO_2 (Duren and Miller, 2012). Gaps persist in the understanding of urban carbon dynamics, and there is a need for more long-term, spatially distributed CO_2 monitoring networks (Mitchell et al., 2018).

Land Use Regression

Land use regression (LUR) is a specialized application of multiple linear regression used to estimate ambient air pollution. It operates under the principle that environmental features, such as land use, population density, road networks, topography, and meteorological conditions are relevant predictors for air pollutant concentrations (Li et al., 2021). The most common use of LUR is to produce exposure assessments for epidemiological studies that predict what levels of air pollution survey participants may be exposed to at unmonitored locations, such as their places of residence (Larkin et al., 2023; Li et al., 2021; Ryan and LeMasters, 2007). Researchers have used LUR to predict concentrations of nitrogen dioxide (NO_2), ambient respirable suspended particulates (PM_{10}), fine suspended particulates ($PM_{2.5}$), ozone (O_3) and carbon monoxide (CO) (Larkin et al., 2023; Li et al., 2021; Wong et al., 2021).

In a literature review synthesizing the development of LUR models, Ryan and LeMasters (2007) analyzed 12 studies and found that independent variables used in LUR can be broadly categorized into four groups: (1) road type, (2) traffic count, (3) elevation, and (4) land cover; traffic count was generally the most important predictor variable. LUR can achieve considerable accuracy, and Ryan and LeMasters (2007) found that the LUR models reviewed in the study accounted for between 54% and 81% of the variability in air pollutant concentrations. Furthermore, the integration of machine learning (ML) and deep learning (DL) algorithms, particularly Extreme Gradient Boosting (XGBoost), has proven capable of improving LUR accuracy (Wong et al., 2021).

One major advantage of LUR is its ability to capture fine-scale spatial and temporal patterns (Larkin et al., 2023; Ryan and LeMasters, 2007). Intraurban air pollution is characterized by high spatial and temporal variability due to seasonal and daily variations in traffic and meteorological conditions, as well as the decay of pollutants over space and time (Larkin et al., 2023; Ryan and LeMasters, 2007). Therefore, granular data is key when attempting to accurately capture spatiotemporal variability in intraurban air pollution. Limitations of LUR include poor transferability between cities and limited global generalizability, partly due to uneven spatial distributions of sensors (Larkin et al., 2023; Li et al., 2021). The performance of the models is sensitive to the quality and quantity of training data, the location of sensors, and the choice of predictor variables (Ryan and LeMasters, 2007).

Research Gap

A review of existing literature has produced no evidence of using LUR to model intraurban CO_2 concentrations, nor any discussion about the feasibility of such an approach. Possible explanations include that CO_2 is a GHG rather than an air pollutant associated with human health risks. Therefore, it has little relevance to LUR's most common application of creating exposure assessments for epidemiological studies. Furthermore, the scarcity of long-term, spatially distributed urban CO_2 monitoring networks may have stalled the development of LUR models that rely on temporally and spatially distributed CO_2 data. This study aims to address this research gap by developing an LUR model to predict intraurban CO_2.

Objectives

The objective of this study is to develop machine learning models to help predict intraurban CO_2 concentrations in the San Francisco Bay Area. Specifically, the viability and model performance of using LUR to predict intraurban CO_2 will be explored. The study will also investigate how well ML or DL algorithms with non-linear features can predict intraurban CO_2 concentrations. The central research question is: How well do the models explored in this study simulate the distribution and variability of intraurban CO_2 concentrations in the San Francisco Bay Area? The sub-questions are: (1) Can LUR effectively predict intraurban CO_2 concentrations? (2) Can ML and DL algorithms improve upon LUR model performance? (3) What are the key predictors of intraurban CO_2 concentrations in the San Francisco Bay Area? My working hypotheses are that more advanced ML and DL algorithms will outperform traditional LUR for predicting intraurban CO_2. Furthermore, I hypothesize that feature variables related to traffic count or road activity will be amongst the most significant predictors of CO_2.

To address these objectives and research questions, I collected and processed CO_2 and predictor variable data for modelling. Next, I trained and evaluated three models to predict intraurban CO_2 at the chosen study site. Finally, I analyzed and compared the final models and their performance to draw conclusions about their viability.

METHODS

Study Site: BEACO$_2$N Network

The Berkeley Environmental Air-quality & CO_2 Network (BEACO$_2$N) was identified as a promising source of spatiotemporally distributed intraurban CO_2 data. The sensor network was established in the San Francisco Bay Area by a team of researchers at the University of California, Berkeley (Shusterman et al., 2016). A total of 74 unique sensors, or nodes, collected real-time CO_2 data primarily across the San Francisco Peninsula, the East Bay, and North Bay between 2012 and 2024. The sensors in the network record data for CO_2, NO, NO_2, O_3, CO, and aerosol as well as for meteorological conditions including temperature, pressure, and relative humidity. The data has a temporal resolution of one minute and a spatial resolution of approximately one mile, which was the finest spatial and temporal resolution identified among any intraurban CO_2 monitoring network. The downloaded BEACO$_2$N dataset included over 2.4 million raw observations. Similar to previous studies, CO_2 and meteorological data was aggregated temporally for modelling, by calculating daily averages when a given sensor recorded at least 18 out of 24 hours' worth of valid data in a single day (Larkin et al., 2023; Lee et al., 2017).

Feature Data Collection and Processing

The entire modelling framework, starting with data collection and processing, is summarized in Figure 1. The choice of predictor variables was informed by previous LUR studies conducted by Larkin et al. (2023), Lee et al. (2017), and Li et al. (2021). Variables related to land use, road traffic, vegetation, urbanization, and meteorology were chosen and downloaded based on relevance and availability. Table 1 summarizes the features and data sources used. Since most BEACO$_2$N data was observed between 2022 and 2024, feature data was chosen to align as closely as possible with this timeframe.

Table 1. Feature variables included in feature selection and modelling

Feature Category	Variable	Year	Source
buffer radii: 50 m, 100 m, 200 m, 300 m, 500 m, 1000 m, 1500 m, 2000 m, 3000 m, 4000 m, 5000 m			
Land Use *total area [m^2] in buffer*	Built Area Rangeland Trees Water Bare Ground Crops Flooded Vegetation	2021	ESRI Sentinel-2 10m Land Use/Land Cover
Industrial Areas *total area [m^2] in buffer*	Industrial	2021-2023	California General Plan Land Use
Annual Average Daily Traffic (AADT) *total count [AADT] in buffer*	AADT	2022	Caltrans 2022 Traffic Volumes (AADT)
Roads *total length [m] in buffer*	Road Length	2022	U.S. Census Bureau 2022 California Roads TIGER/Line
Normalized Difference Vegetation Index (NDVI) *mean NDVI in buffer*	NDVI	2022	Landsat 8-9 OLI/TIRS C2 L2 Surface Reflectance-derived NDVI
Population Density *people per km^2 in buffer*	Population Density	2022	California Hard-to-Count Index U.S. Census Bureau 2022 Census Tract TIGER/Line
Temperature *[°C]*	Temperature	2012-2024	BEACO$_2$N
Pressure *[kPA]*	Pressure	2012-2024	BEACO$_2$N
Relative Humidity *[%]*	Relative Humidity	2012-2024	BEACO$_2$N

The `geopandas`, `shapely`, and `rasterio` Python libraries were key to processing spatial feature data. Land use data for the study site was downloaded from the global ESRI Sentinel-2 10m Land Use/Land Cover database as a TIF image, where land use information is stored in pixels using unique integer values. The respective land use labels were assigned to each integer value, and the data was stored as a GeoDataFrame, where pixels were converted to polygon geometries. The transformed data was dissolved and saved as a shapefile. The California General Land Use Plan is another source of land use data available for download as a shapefile. Given limitations around spatial coverage, this data was only used to supplement information about industrial sites and activities in the study region represented as polygon geometries. Road and AADT data were downloaded as shapefiles, where roads are represented as lines and AADT information is stored in point geometries. NDVI data was calculated using Landsat 8-9 OLI Collection 2 Surface Reflectance TIF images, where pixels contain NDVI information as float values between -1.0 and 1.0. Finally, population density data was obtained by combining a shapefile featuring census tract polygons with a .csv file containing population density data by census tract, where geographic identifier GEOIDs were used as a common feature to produce a GeoDataFrame.

Geographic coordinates for BEACO$_2$N sensor locations were converted to point geometries and saved as a GeoDataFrame to be plotted and overlayed with spatial feature data. In accordance with previous studies, buffers were created around each node to extract relevant feature data. The buffer radii used were 50, 100, 200, 300, 500, 1000, 2000, 3000, 4000, and 5000 meters. Feature data was extracted within each of the ten buffers zones, for each of the following twelve spatial features: built area, rangeland, trees, water, bare ground, crops, flooded vegetation, industrial areas, annual average daily traffic (AADT), road length, normalized difference vegetation index (NDVI), and population density. Feature extraction methods were informed mainly by Lee et al. (2017) and are summarized in Table 2. A total of 120 spatial features (12 variables times 10 radii) were produced during feature extraction.

Temperature, pressure, and relative humidity data was collected alongside CO$_2$ by BEACO$_2$N sensors; these meteorological features have the same temporal resolution as the CO$_2$ data and were aggregated as daily averages alongside CO$_2$. By contrast, daily data for spatial features described above was not available, and the temporal resolution of these features is instead constant. While land use, NDVI, and population density features are less variable over time, the constant temporal resolution for traffic data is a greater compromise. Since the meteorological features lacked a spatial dimension, they were not processed using buffers. BEACO$_2$N data for CO$_2$ and meteorological conditions was merged with buffered feature data using `node_id` as a common feature, yielding the final dataset of 123 numerical feature variables and CO$_2$ data as the target variable. All features were standardized.

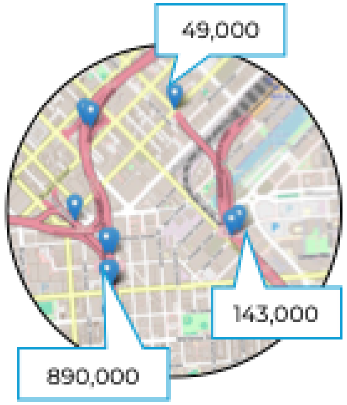
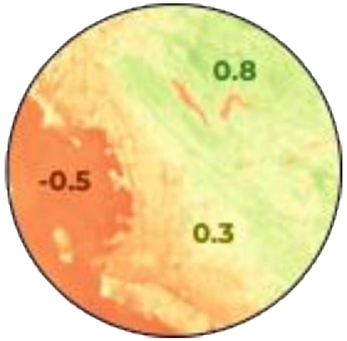
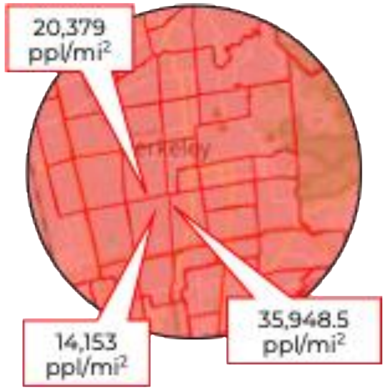
Feature Selection

The feature selection methods from the literature were tested, including approaches by Larkin et al. (2023), Lee et al. (2017), Li et al. (2021), and Wong et al. (2021). Ultimately, a new protocol was defined that first considered features with an absolute Spearman's correlation coefficient of at least 0.03, before iteratively filtering out features with a variable inflation factor (VIF) over 3 (Figure 1). Spearman's correlation coefficient assesses the strength and direction of the relationship between predictor and target variables, with values ranging from -1.0 to 1.0. Since individual correlations were relatively small, the cut-off was chosen to be low. Spearman's correlation coefficient was preferred over the Pearson correlation coefficient and other feature selection metrics seen in the literature as it does not assume a linear relationship between the feature and target variables, which seems more appropriate for non-linear models employed in this study. A VIF cutoff of 3 was observed in previous studies and helps minimize multicollinearity between features (Lee et al., 2017; Li et al., 2021; Wong et al., 2021). Features with all zero observations were dropped.

Traditional LUR (OLS)

Traditional LUR observed in the literature consists of a multiple linear regression model used to predict air pollutant concentrations. In this project, ordinary least squares (OLS) was used to fit a multiple linear regression model to the selected features and the observed CO$_2$ data.

Table 2. Feature data extraction methodologies for spatial features, including all land use categories , industrial sites, annual average daily traffic, road length, normalized difference vegetation index and population density.

Land Use & Industrial	AADT	Road Length	NDVI	Population Density
"Rangeland" "Water" "Built Area"		ancisco South of Market District		 20,379 ppl/mi² erkeley 14,153 ppl/mi² 35,948.5 ppl/mi²
Total area [m^2] in buffer, per Land Use type	Total AADT sum in buffer	Total length of roads [m] in buffer	Average NDVI in buffer	Total population density in buffer area
$\sum_{i=1}^{n}(\text{area})_i,$	$\sum_{i=1}^{n}(\text{AADT})_i,$	$\sum_{i=1}^{n}(\text{road length})_i,$	$\frac{1}{n}\sum_{i=1}^{n}(\text{NDVI})_i,$	$\frac{1}{\text{buffer area }[km^2]}\sum_{i=1}^{n}\left(\frac{\text{ppl}}{\text{mi}^2}\right)_i \times (mi^2)_i,$
$n =$ total # polygons per LU type in buffer	$n =$ total # AADT observations in buffer	$n =$ total # of roads in buffer	n = total # of pixels in buffer	$n =$ total # of polygons in buffer

Extreme Gradient Boosting (XGBoost)

Previously, Wong et al. (2021) have explored XGBoost in the context of LUR by "incorporating" the two methods. In essence, LUR was used for feature selection, by fitting a multiple linear regression model between the feature and target variables, and exploring the strength of correlations, as well as processing the significance and direction of coefficients for individual features. The chosen features were then used to develop an XGBoost model to predict the target variable. Since using LUR for feature selection imposes a linear relationship on the features and the target variable, while XGBoost can instead capture non-linear relationships in the data, this approach was not chosen for this study. Instead, features were selected using the protocol described in the Feature Selection section, and selected, scaled features were used to fit an XGBoost model. Gradient descent is used to minimize the loss function, in this case the mean squared error (MSE).

Convolutional Neural Network (CNN)

Finally, a one-dimensional convolutional neural network (1D CNN) was trained to predict CO_2 concentrations (Figure 1). The selected feature data was reshaped to explore the relationship between predictor variables. The model architecture includes two convolutional layers, the first applying 64 filters and the second applying 128 filters, each with a kernel size of 3 and ReLU activation function. The convolutional layers are each followed by a batch normalization layer to stabilize training, a max pooling layer with a pool size of 2 to reduce spatial dimensions, and a dropout layer with dropout rate of 20% to minimize overfitting. The output is then flattened and fed to a dense layer of 128 nodes activated using ReLU, followed by a batch normalization and a 20% dropout layer. The final output layer is a single node representing predicted CO_2 concentrations. The CNN was compiled with the Adam optimizer, an initial learning rate of 0.001 and an MSE loss function. Early stopping, learning rate reduction on plateau, and model checkpointing based on validation loss were included as callbacks. The CNN was trained for up to 150 epochs with a batch size of 64 (Appendix E). Hyperparameters were tuned using a 3-fold cross-validated GridSearch.

Model Validation and Testing

Model performance was evaluated using the coefficient of determination (R^2), the root mean squared error (RMSE), mean squared error (MSE) and the mean absolute error (MAE) (Figure 1). After preprocessing the data by removing invalid observations and calculating daily averages for CO_2 and meteorological features, the dataset was balanced to ensure a more consistent spatial and temporal coverage. Select nodes that were eliminated during the balancing process were used to evaluate model performance at unseen sensor locations. These test nodes were excluded from the training process and split into two groups: central test nodes (sensors that are surrounded by training nodes), and fringe test nodes (sensors that are geographically far removed or located near the fringes of the training network). Meanwhile, the data from balanced nodes was randomly split into training and test sets, where 80% of the dataset was used for training and validation, while 20% of the data was reserved for testing. The traditional LUR and XGBoost models were scored using 10-fold cross-validation, in addition to the reserved test set; given the complexity of the CNN and its training time, the CNN scores were not cross-validated. Model performance was evaluated and plotted for individual training and test nodes (see Results Explorer, Appendix A).

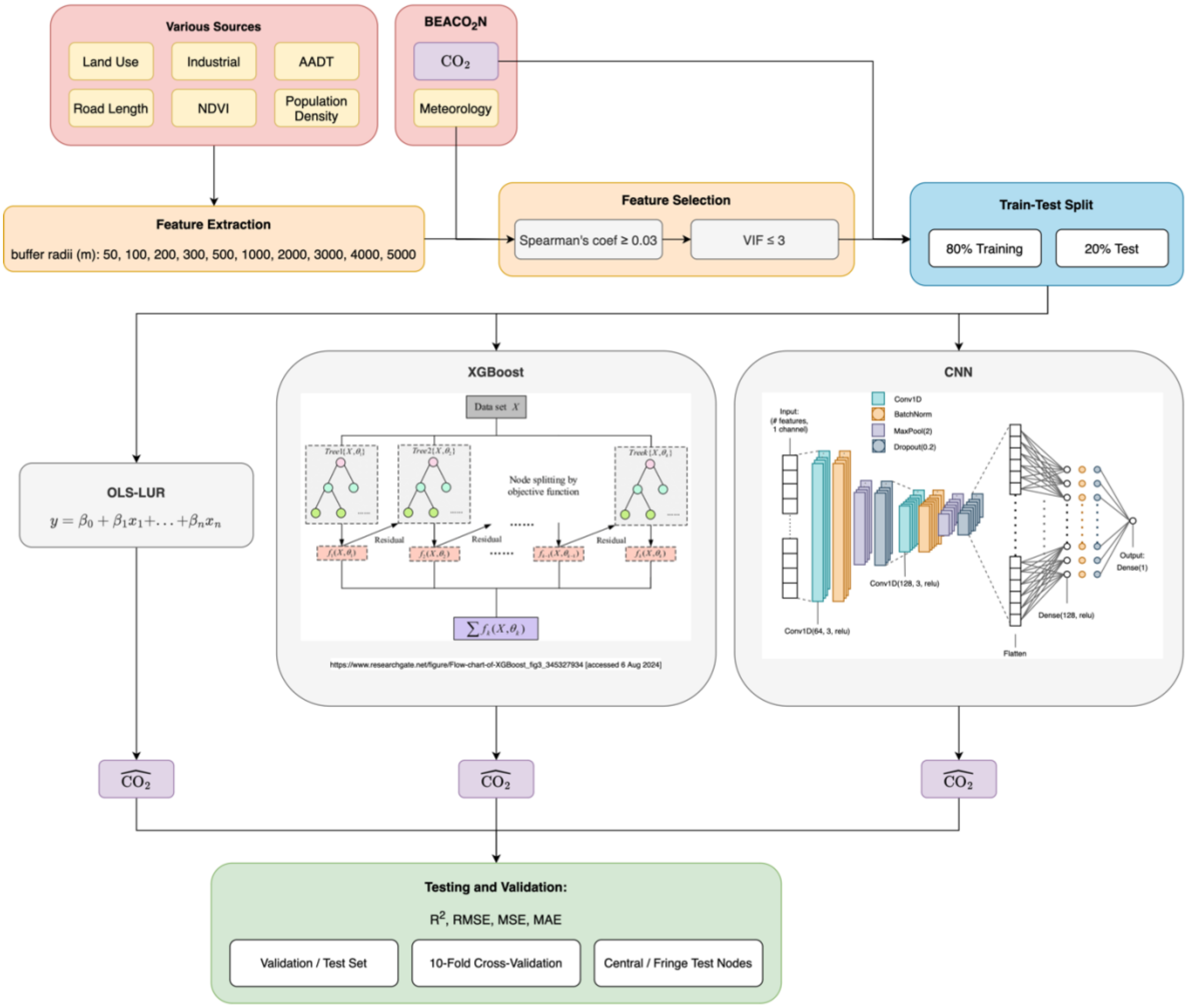


Figure 1. The modelling workflow. Data collection (red), data processing by feature extraction and feature selection (orange), train-test split (blue), modelling using traditional LUR, XGBoost, and CNN (grey) and model evaluation (green).

RESULTS

BEACO$_2$N Data Exploratory Data Analysis and Processing

The number of reporting nodes was not consistent over the entire 12-year period. Nodes were installed and removed over time, with the total number of nodes growing from 12 in 2012 to a maximum of 54 in 2023. Individual nodes also experienced periods of defect or inactivity introducing additional temporal inconsistencies to the data. For example, all CO_2 measurements reported in 2012 were invalid (recorded as -999) and therefore eliminated. Six sensors reported only invalid data, while all nodes reported at least one invalid observation over the entire reporting period. Over 90% of days included in the reporting period had at least one invalid observation across the network. Invalid observations for CO_2 and meteorological variables were dropped, as well as data from COVID-19 years 2020 and 2021.

Calculating daily averages after filtering out invalid observations and daily averages with less than 18 hours' worth of valid data produced 67,044 observations from a total of 65 nodes. This data was imbalanced, seeing as some nodes reported over 1,700 valid daily averages, while others reported less than ten; complete daily CO_2 averages on some days were reported by only one node, while other daily averages were represented by over 40 nodes. In attempt to balance the daily average CO_2 data, the dataset was filtered so that each day was represented by at least 27 nodes while every node reported at least 200 daily averages. The final dataset included 17,389 daily CO_2 averages reported by 42 sensors. 11,128 observations were used for training, 2,783 for validation, and 3,478 observations were reserved for testing. Five of the sensors that were eliminated during the balancing procedure were used as central test nodes, while seven were used as fringe test nodes.

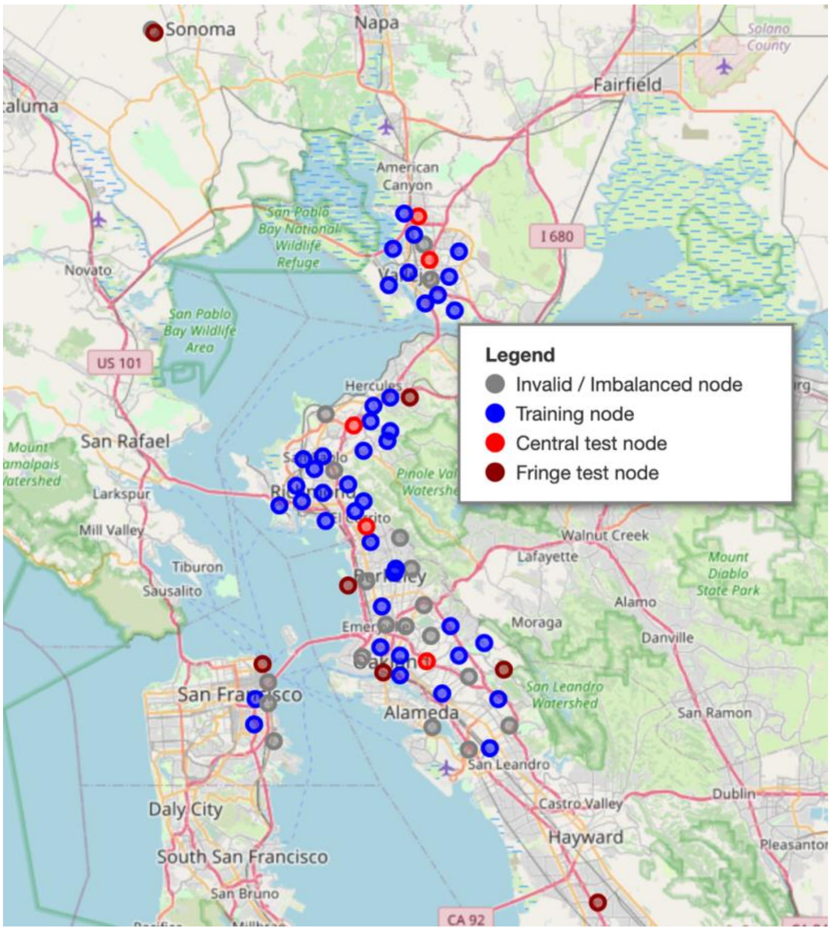


Figure 2. BEACO2N sensor locations and groupings.

Feature Selection

After dropping 18 features with all zero data, 105 features entered the feature selection process. Applying the Spearman's correlation coefficient condition selected 32 features, of which 11 passed the VIF condition. The eleven final features are: temperature, pressure, relative humidity, trees area (50m), total road length (1000m), total road length (200m), built area (2000m), total AADT (3000m), flooded vegetation area (1000m), industrial area (5000m), and average NDVI (50m). The feature selection scores presented in Table 3 highlight that the meteorological features and the trees feature had the strongest absolute relationships with CO_2, followed approximately by population density, built area, NDVI, road length, and AADT. No features related to rangeland, water, bare ground, or crops were selected by either criterion. Overall, absolute Spearman's correlation coefficients were relatively small, with the largest coefficients achieved by temperature and pressure, with coefficients of -0.51 and 0.40 respectively. The reduction from 32 features to 11 by the VIF condition demonstrates that many features, especially features of the same category but of different buffer radii, were strongly intercorrelated.

Table 3. Feature selection scores and outcomes. 32 features selected by Spearman's correlation coefficient, of which 11 were selected using VIF. The three meteorological features had the largest absolute Spearman's coefficients. The sign of Spearman's correlation coefficient indicates direction of effect on CO_2 concentrations.

Feature	Spearman	VIF
temp	-0.51	1.26
pressure	0.40	2.47
rh	-0.11	1.20
Trees_area_100m	-0.09	-
Trees_area_50m	-0.08	2.14
Trees_area_200m	-0.06	-
Trees_area_300m	-0.06	-
Trees_area_500m	-0.05	-
avg_pop_dens_2000m	0.05	-
avg_ndvi_100m	-0.05	-
Built_Area_area_1000m	0.05	-
avg_pop_dens_3000m	0.05	-
avg_pop_dens_4000m	0.04	-
Built_Area_area_3000m	0.04	-
Built_Area_area_4000m	0.04	-
avg_ndvi_200m	-0.04	-
avg_pop_dens_1000m	0.04	-
total_road_length_1000m	0.04	1.65
Trees_area_1000m	-0.04	-
avg_ndvi_300m	-0.04	-
total_road_length_200m	0.04	1.46
Built_Area_area_2000m	0.04	1.93
avg_ndvi_500m	-0.04	-
avg_pop_dens_5000m	0.04	
total_AADT_3000m	0.04	1.40
Flooded_Vegetation_area_1000m	-0.03	1.22
Industrial_area_5000m	0.03	1.53
Built_Area_area_500m	0.03	-
total_AADT_1000m	0.03	-
avg_pop_dens_500m	0.03	-
avg_ndvi_50m	-0.03	1.37
avg_ndvi_1000m	-0.03	-

Training Node Model Performance

Table 4 summarizes the model validation and testing scores. When the models were scored using the reserved test set representing 20% of the data from the training nodes, the traditional LUR model explained 34% of the variability in the observed CO_2 data, while XGBoost and CNN both accounted for 58%. For the traditional LUR model, the RMSE, MSE, and MAE values achieved on the test data were 15.81, 250.02 and 12.04; XGBoost achieved RMSE, MSE and MAE values of 12.56, 157.66 and 9.14, respectively; and CNN achieved RMSE, MSE and MAE values of 12.63, 159.45, and 9.08, respectively. During model validation, XGBoost achieved the best scores across all metrics, whereas during model testing, XGBoost was closely rivaled by CNN. The 10-fold cross-validated scores for traditional LUR and XGBoost matched the validation and test set scores closely, indicating no overfitting within the training set. Geographically, R^2 values were generally highest for training nodes around Richmond and Vallejo. Here, the network of sensors was especially dense and regularly spaced compared to areas in San Francisco, Berkeley and Oakland (see Results Explorer, Appendix A).

Table 4. Model performance and comparison. XGBoost performed best on the validation and test sets. CNN performed marginally better for central and fringe test nodes.

Evaluation Step	Metric	LUR	XGBoost	CNN
Validation (20% of training set)	R^2	0.35	0.60	0.57
	RMSE	15.42	12.04	12.50
	MSE	237.78	145.02	156.15
	MAE	11.82	8.82	9.14
Test Set (20% of full data)	R^2	0.34	0.58	0.58
	RMSE	15.81	12.56	12.63
	MSE	250.02	157.66	159.45
	MAE	12.04	9.14	9.08
10-Fold CV	R^2	0.34	0.58	-
	RMSE	15.80	12.54	-
	MSE	249.57	157.26	-
	MAE	11.99	9.05	-
Central Test Nodes	R^2	0.31	0.42	0.42
	RMSE	19.13	17.48	17.46
	MSE	366.05	305.67	304.71
	MAE	15.47	12.90	13.01
Fringe Test Nodes	R^2	-0.69	-0.88	-0.47
	RMSE	20.21	21.24	18.77
	MSE	404.76	451.14	352.46
	MAE	17.24	18.11	16.10

Testing Node Model Performance

Model evaluation on unseen nodes demonstrates overall weaker performance than on test data reserved from training nodes. The models performed better for central test nodes, with XGBoost and CNN both accounting for 42% of the variability in the data. While traditional LUR only captured 31% of the variability for central test nodes, this score is very close to performance on training nodes, whereas XGBoost or CNN demonstrated a considerable drop in performance compared to training nodes (Table 4). All three models performed exceptionally poorly on the fringe test nodes, scoring negative R^2 values. Overall, the CNN model achieved the best scores for unseen node locations.

Feature Importances

Beyond the Spearman's correlation coefficient, feature importances for LUR and XGBoost were further analyzed. In the LUR model, the features with the highest partial R^2 were temperature, pressure, and trees area (50m), with partial R^2 values of 0.16, 0.14, and 0.05 respectively (Appendix B). For XGBoost, feature importance measured through gain in accuracy highlighted pressure, temperature, trees area (50m), relative humidity, and total AADT (3000m) as the five most important features (Appendix C). Looking at SHAP (SHapley Additive exPlanations) values for XGBoost highlights pressure, temperature, relative humidity, built area (2000m), and industrial area (5000m) as the five most important features (Appendix D.) By SHAP value, trees area (50m) has the lowest relative importance. This indicates that nearby trees sometimes resulted in significant accuracy improvements, evidenced by a high gain, but its overall contribution to predicted concentration is less consistent across space.

This is likely because many nodes were not located near significant tree areas. While relative humidity and total AADT (3000m) contribute moderately both in terms of improved accuracy and shaping predictions, likely due to their variability and abundance across space, built area (2000m) and industrial area (5000m) seem to play a significant role in shaping predictions given the effect of their presence or absence on ambient CO_2. For both LUR and XGBoost, temperature and pressure are the two strongest predictor variables.

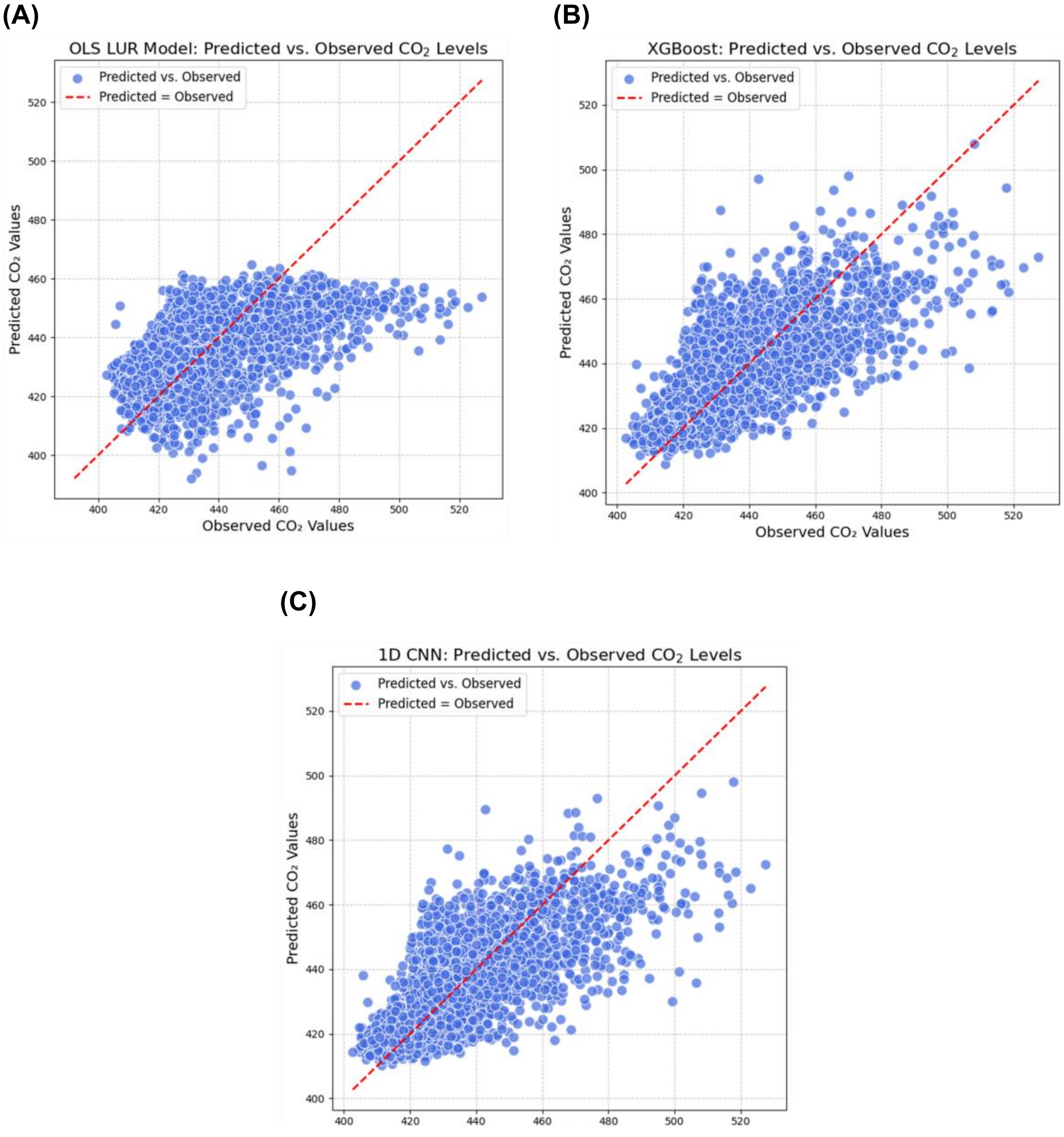


Figure 1. Model CO_2 predictions versus true concentrations for 20% reserved test data from training nodes. (A) Traditional LUR (OLS): R^2=0.34, (B) XGBoost: R^2=0.58, (C) CNN: R^2=0.58.

DISCUSSION

The results demonstrate that of the three models used to predict intraurban daily average CO_2 concentrations within the BEACO_2N monitoring network, XGBoost achieved some of the highest scores on validation and test data reserved from training nodes. CNN achieved comparable performance to XGBoost for test data from training nodes but most notably demonstrated better performance on unseen node locations. Overall, XGBoost and CNN outperformed traditional LUR, indicating that non-linear relationships are more representative of the true relationship between the environmental predictor variables and CO_2 concentrations (Figure 3). Like carbon monoxide models produced by Wong et al. (2021), this study evidences the ability of ML and DL algorithms to outperform traditional LUR, indicating that the non-linear trend is not specific to only CO_2. This helps make the case that modelling of intraurban ambient gases like air pollutants or GHGs should move away from linear regression methods and instead focus on the potential of ML and DL algorithms to produce better predictions.

Model performance at unseen node locations was lower than for training node locations. The models therefore have a low external validity, which echoes conclusions reached about the transferability and generalizability of traditional LUR models (Larkin et al., 2023; Li et al., 2021). Comparing the performance between central and fringe test nodes indicates that predictive accuracy at unseen locations that are more centrally situated within the network of training nodes is higher than at node locations far removed from the main cluster of training nodes. Models to predict intraurban CO_2 concentrations in a desired region should therefore be specific to that area and cover as much ground as possible at a fine spatial resolution. Poor performance at fringe test nodes is likely also the result of proximity to land use types that were underrepresented near training nodes, such as water bodies or forested areas. This resembles conclusions drawn by Ryan and LeMasters (2007) about the importance of the variability of land use characteristics captured by the network in determining model performance. Nodes with unusual distributions of observed CO_2 were also more difficult to predict (Figure 4). Any network of nodes used to train prediction models for intraurban CO_2 concentrations should capture as much variability in predictor and target variables as possible.

The analysis of feature importances highlights the significant contribution of meteorological features, especially temperature and pressure, in predicting ambient CO_2 concentrations. The importance of these features is likely linked to their granular temporal resolution consistent with the target variable. Compared to land use features, which were individually not present throughout the entire study area and remained constant for all observations from a given node and buffer radius, meteorological features demonstrated an inherently higher degree of variability across observations. While at least one feature was selected from both the road length and AADT categories, their relative feature importance was lower than expected considering results from previous LUR studies (Larkin et al., 2023; Li et al., 2021; Ryan and LeMasters, 2007; Wong et al., 2021). The tree feature was shown to improve accuracy for both LUR and XGBoost, although the feature was not abundant across the entire study area and therefore did not shape most predicted values as evidenced by its low SHAP score. Instead, built area as the most abundant land use type and industrial activities in the greater vicinity were shown to shape predicted CO_2 concentrations more consistently across the space. Overall, feature importances were more closely linked to spatiotemporal variability and abundance of a given feature near the network of training sensors than to their absolute contribution to true ambient levels of CO_2.

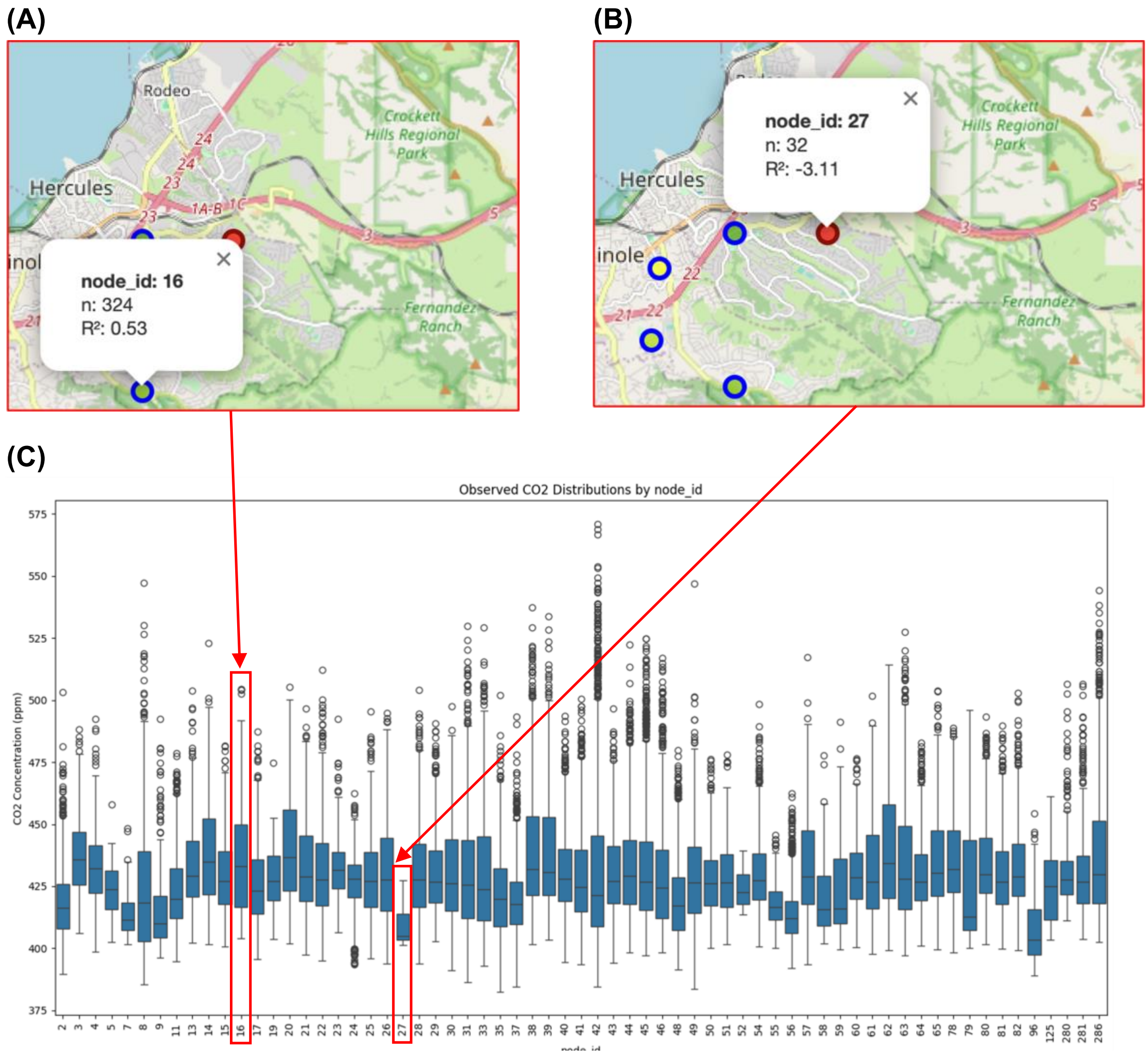


Figure 4. Comparing XGBoost model performance for (A) node 16 and (B) node 27. Node 16 has better R^2 than node 27 despite similar locations. Trend consistent across all models. Notable differences are the number of observations (n) and (C) distributions of CO_2 observations.

The quality and availability of data for intraurban CO_2 and predictor variables was a major limitation for this study. While the BEACO_2N network has an impressive spatiotemporal resolution compared to other intraurban CO_2 monitoring networks, the inconsistent validity of observations across nodes created a strong spatiotemporal imbalance in the dataset that had to be overcome. Furthermore, the generalizability of the network was compromised, especially in areas with greater proportions of land use types that were underrepresented near the network of training sensors. Irregularity and coarse spatial resolution in the distribution of sensors also seemed to contribute to poor performance in certain areas, compared to Richmond and Vallejo, where the network was denser and more regular. The constant temporal resolution of environmental features, especially traffic data, was another notable limitation. Ideally, data for predictor variables should be collected with the same spatial and temporal resolution as for CO_2. The novelty of this study lies in modelling intraurban CO_2 using LUR and ML algorithms. Future directions include modelling intraurban CO_2 using different models, predictor variables, feature selection methodologies or CO_2 monitoring networks.

CONCLUSION

This study was the first of its kind to evaluate the viability of using LUR to predict intraurban ambient CO_2 concentrations. The study has demonstrated that ML and DL algorithms can outperform traditional LUR in terms of predictive accuracy, illustrating that the relationship between environmental features and CO_2 should not be presumed to be linear. This conclusion corroborates the findings of a previous study, underscoring the potential and value of expanding upon traditional LUR using novel modelling approaches. Evaluating models using data from unseen node locations further illustrated the models' limitations in terms of generalizability and transferability. The availability and spatiotemporal variability of data from intraurban CO_2 monitoring networks is scarce and remains a major limitation for research in this field. In light of rising urbanization and the increasingly drastic effects of climate change, efforts to establish intraurban CO_2 monitoring networks with high spatiotemporal resolution should be expanded. More research into the modelling of ambient intraurban CO_2 is necessary to better understand and predict the effects of anthropogenic urban activities and land use change on climate change. This research is vital to inform decision-makers on sustainable development and urban decarbonization in the 21st century.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Fangxin Fang and Linfeng Li for their guidance and valuable feedback throughout this thesis project. I also thank Marijan Beg for his logistical support during the MSc program and IRP. I am very thankful for the friends in the EDMSL course who have been a tremendous source of joy, motivation, and inspiration. Lastly, I am profoundly grateful to my friends and family for their constant encouragement and support.

REFERENCES

Bergeron, O., Strachan, I.B., 2011. CO2 sources and sinks in urban and suburban areas of a northern mid-latitude city. Atmospheric Environment 45, 1564–1573. https://doi.org/10.1016/j.atmosenv.2010.12.043

Bréon, F.M., Broquet, G., Puygrenier, V., Chevallier, F., Xueref-Remy, I., Ramonet, M., Dieudonné, E., Lopez, M., Schmidt, M., Perrussel, O., Ciais, P., 2015. An attempt at estimating Paris area CO&lt;sub&gt;2&lt;/sub&gt; emissions from atmospheric concentration measurements. Atmos. Chem. Phys. 15, 1707–1724. https://doi.org/10.5194/acp-15-1707-2015

Briber, B., Hutyra, L., Dunn, A., Raciti, S., Munger, J., 2013. Variations in Atmospheric CO2 Mixing Ratios across a Boston, MA Urban to Rural Gradient. Land 2, 304–327. https://doi.org/10.3390/land2030304

Coutts, A.M., Beringer, J., Tapper, N.J., 2007. Characteristics influencing the variability of urban CO2 fluxes in Melbourne, Australia. Atmospheric Environment 41, 51–62. https://doi.org/10.1016/j.atmosenv.2006.08.030

Duren, R.M., Miller, C.E., 2012. Measuring the carbon emissions of megacities. Nature Clim Change 2, 560–562. https://doi.org/10.1038/nclimate1629

Helfter, C., Tremper, A.H., Halios, C.H., Kotthaus, S., Bjorkegren, A., Grimmond, C.S.B., Barlow, J.F., Nemitz, E., 2016. Spatial and temporal variability of urban fluxes of methane, carbon monoxideand carbon dioxide above London, UK. Atmos. Chem. Phys. 16, 10543–10557. https://doi.org/10.5194/acp-16-10543-2016

Intergovernmental Panel On Climate Change (Ipcc) (Ed.), 2023. Climate Change 2022 - Mitigation of Climate Change: Working Group III Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009157926

Larkin, A., Anenberg, S., Goldberg, D.L., Mohegh, A., Brauer, M., Hystad, P., 2023. A global spatial-temporal land use regression model for nitrogen dioxide air pollution. Front. Environ. Sci. 11, 1125979. https://doi.org/10.3389/fenvs.2023.1125979

Lauvaux, T., Miles, N.L., Deng, A., Richardson, S.J., Cambaliza, M.O., Davis, K.J., Gaudet, B., Gurney, K.R., Huang, J., O'Keefe, D., Song, Y., Karion, A., Oda, T., Patarasuk, R., Razlivanov, I., Sarmiento, D., Shepson, P., Sweeney, C., Turnbull, J., Wu, K., 2016. High-resolution atmospheric inversion of urban CO $_2$ emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX). JGR Atmospheres 121, 5213–5236. https://doi.org/10.1002/2015JD024473

Lee, M., Brauer, M., Wong, P., Tang, R., Tsui, T.H., Choi, C., Cheng, W., Lai, P.-C., Tian, L., Thach, T.-Q., Allen, R., Barratt, B., 2017. Land use regression modelling of air pollution in high density high rise cities: A case study in Hong Kong. Science of The Total Environment 592, 306–315. https://doi.org/10.1016/j.scitotenv.2017.03.094

Li, Z., Ho, K.-F., Chuang, H.-C., Yim, S.H.L., 2021. Development and intercity transferability of land-use regression models for predicting ambient PM10, PM2.5, NO2 and O3 concentrations in northern Taiwan. Atmos. Chem. Phys. 21, 5063–5078. https://doi.org/10.5194/acp-21-5063-2021

Mitchell, L.E., Lin, J.C., Bowling, D.R., Pataki, D.E., Strong, C., Schauer, A.J., Bares, R., Bush, S.E., Stephens, B.B., Mendoza, D., Mallia, D., Holland, L., Gurney, K.R., Ehleringer, J.R., 2018. Long-term urban carbon dioxide observations reveal spatial and temporal dynamics related to urban characteristics and growth. Proc. Natl. Acad. Sci. U.S.A. 115, 2912–2917. https://doi.org/10.1073/pnas.1702393115

Poumanyvong, P., Kaneko, S., 2010. Does urbanization lead to less energy use and lower CO2 emissions? A cross-country analysis. Ecological Economics 70, 434–444. https://doi.org/10.1016/j.ecolecon.2010.09.029

Riahi, K., Van Vuuren, D.P., Kriegler, E., Edmonds, J., O'Neill, B.C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J.C., Kc, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Da Silva, L.A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J.C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., Tavoni, M., 2017. The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. Global Environmental Change 42, 153–168. https://doi.org/10.1016/j.gloenvcha.2016.05.009

Rice, A., Bostrom, G., 2011. Measurements of carbon dioxide in an Oregon metropolitan region. Atmospheric Environment 45, 1138–1144. https://doi.org/10.1016/j.atmosenv.2010.11.026

Ryan, P.H., LeMasters, G.K., 2007. A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. Inhalation Toxicology 19, 127–133. https://doi.org/10.1080/08958370701495998

Shusterman, A.A., Teige, V.E., Turner, A.J., Newman, C., Kim, J., Cohen, R.C., 2016. The BErkeley Atmospheric CO&lt;sub&gt;2&lt;/sub&gt; Observation Network: initial evaluation. Atmos. Chem. Phys. 16, 13449–13463. https://doi.org/10.5194/acp-16-13449-2016

Velasco, E., Pressley, S., Allwine, E., Westberg, H., Lamb, B., 2005. Measurements of CO fluxes from the Mexico City urban landscape. Atmospheric Environment 39, 7433–7446. https://doi.org/10.1016/j.atmosenv.2005.08.038

Velasco, E., Roth, M., 2010. Cities as Net Sources of CO $_2$: Review of Atmospheric CO $_2$ Exchange in Urban Environments Measured by Eddy Covariance Technique. Geography Compass 4, 1238–1259. https://doi.org/10.1111/j.1749-8198.2010.00384.x

Wang, S.-H., 2018. Can spatial planning really mitigate carbon dioxide emissions in urban areas? A case study in Taipei, Taiwan. Landscape and Urban Planning.

Wong, P.-Y., Hsu, C.-Y., Wu, J.-Y., Teo, T.-A., Huang, J.-W., Guo, H.-R., Su, H.-J., Wu, C.-D., Spengler, J.D., 2021. Incorporating land-use regression into machine learning algorithms in estimating the spatial-temporal variation of carbon monoxide in Taiwan. Environmental Modelling & Software 139, 104996. https://doi.org/10.1016/j.envsoft.2021.104996

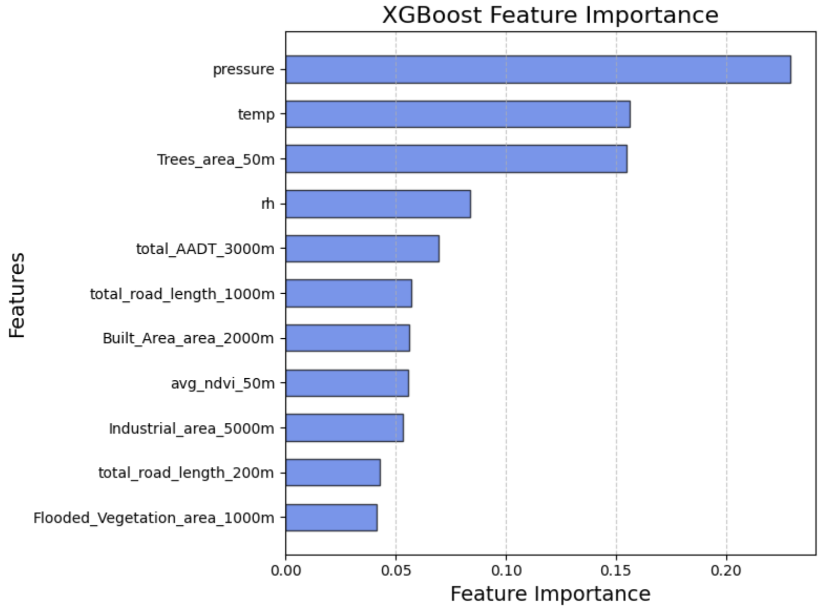
APPENDIX

Appendix A. GitHub repository

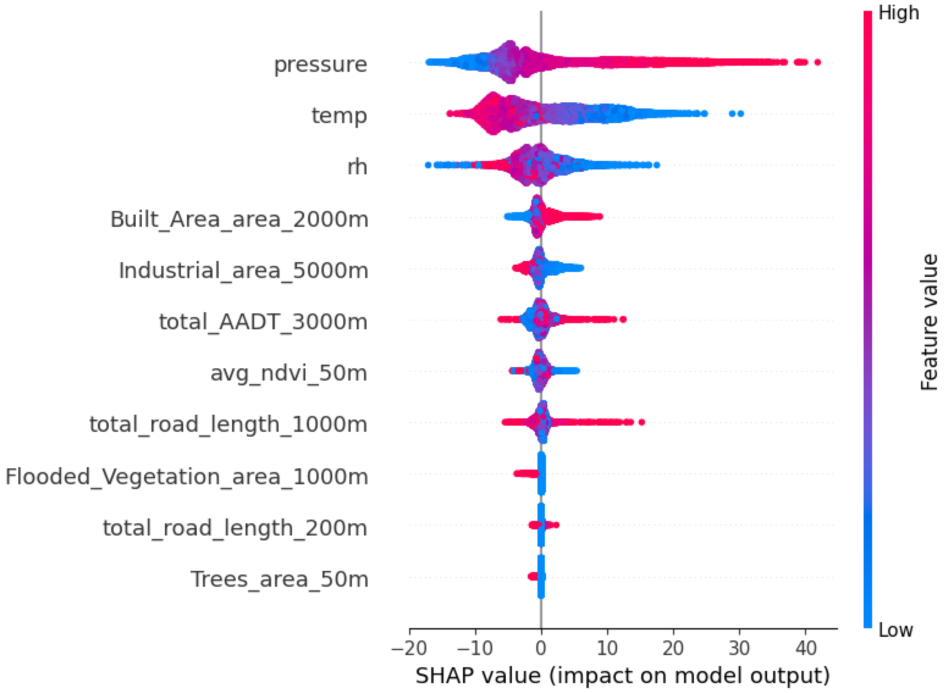
All modelling was done in Python 3.11.9 on a computer with an Apple M2 Max chip and 32 GB of RAM. The code for the project can be found in the GitHub repository `irp-acs223`, containing the `bayareaco2` Python module and as well as Jupyter notebooks that demonstrate model development and implementation. Instructions for installation and links to download data can be found in the `README.md` file. The `explore` folder hosts a GitHub Page, which allow the user to engage more interactively with the data and models used in the `bayareaco2` module. The `Feature Explorer` demonstrates spatial feature variable data and node locations, while the `Results Explorer` showcases model performance for individual nodes.

Appendix B. Traditional LUR model feature statistics to help demonstrate feature importances.

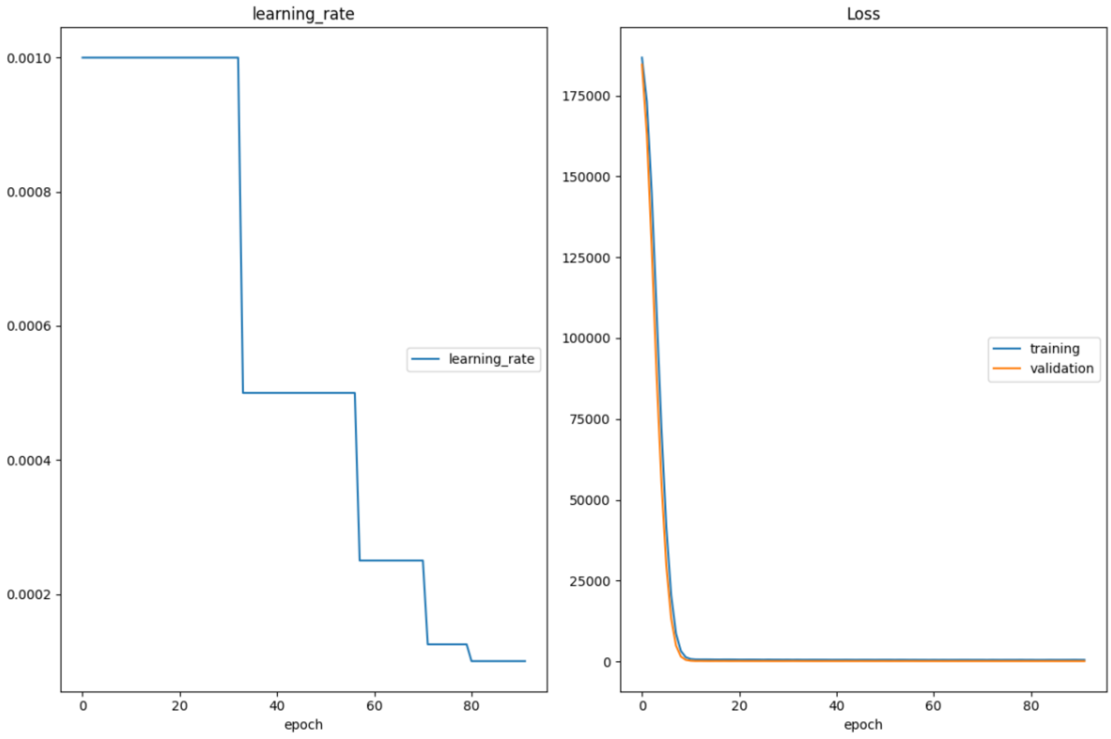
Feature	Coefficient	Standard Error	p-value	VIF	Partial R^2
temp	-7.72	0.15	<0.005	1.25	0.16
pressure	10.00	0.21	<0.005	2.52	0.14
rh	-2.98	0.15	<0.005	1.20	0.03
Trees_area_50m	5.46	0.20	<0.005	2.17	0.05
total_road_length_1000m	-0.78	0.17	<0.005	1.66	<0.005
total_road_length_200m	0.34	0.16	0.04	1.47	<0.005
Built_Area_area_2000m	1.87	0.19	<0.005	1.93	0.01
total_AADT_3000m	0.34	0.16	0.03	1.40	<0.005
Flooded_Vegetation_area_1000m	-0.62	0.15	<0.005	1.22	<0.005
Industrial_area_5000m	-1.91	0.17	<0.005	1.54	0.01
avg_ndvi_50m	0.31	0.16	0.05	1.39	<0.005



Appendix C. XGBoost Gain Feature Importance. The gain statistic describes the improvement in accuracy achieved by adding a given feature to the model.



Appendix D. XGBoost SHAP Feature Importance. The SHAP values describe the contribution of each feature to shaping the model's predictions.



Appendix E. CNN training learning rate and loss curves. Observed learning rate reduction on plateau during training. Training and validation loss stable and decreasing consistently, no indication of overfitting.