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

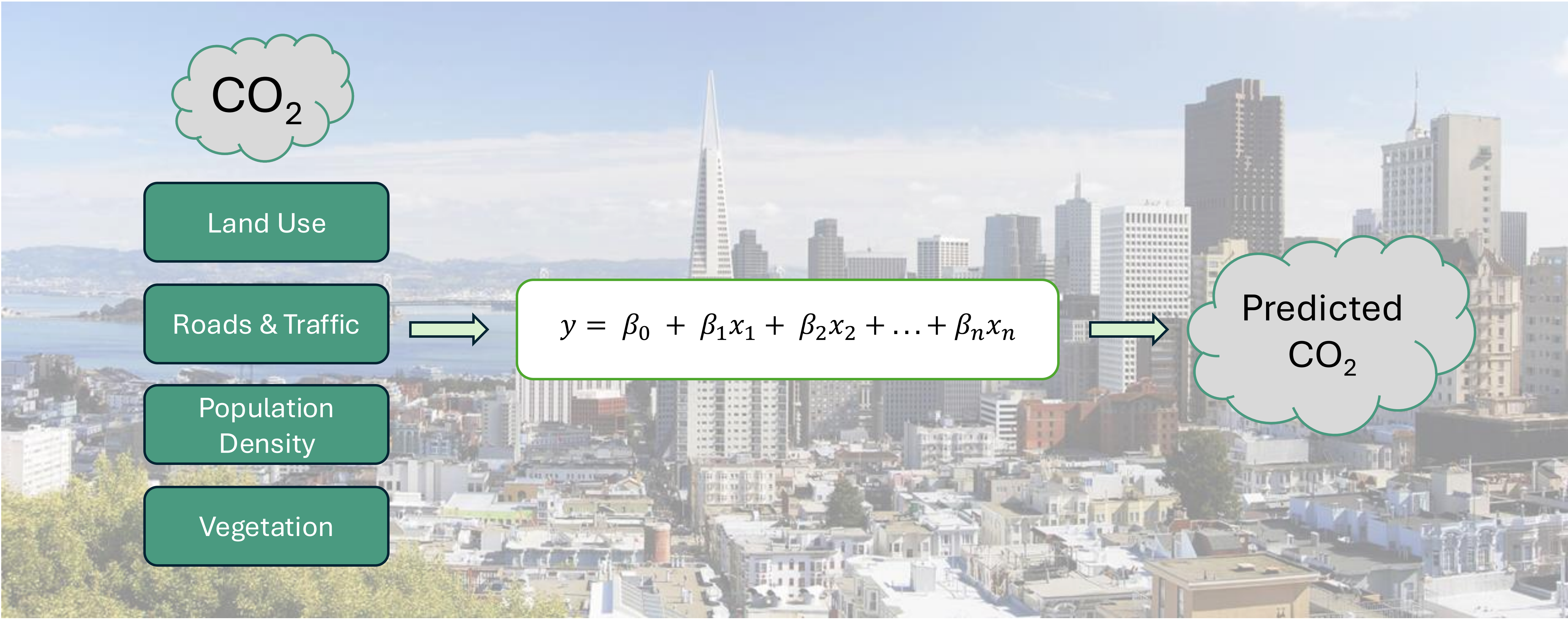
M.Sc. Environmental Data Science and Machine Learning

IRP Presentation

September 9th, 2024

Anna C. Smith

Intraurban CO_2 & Land Use Regression



Research Questions

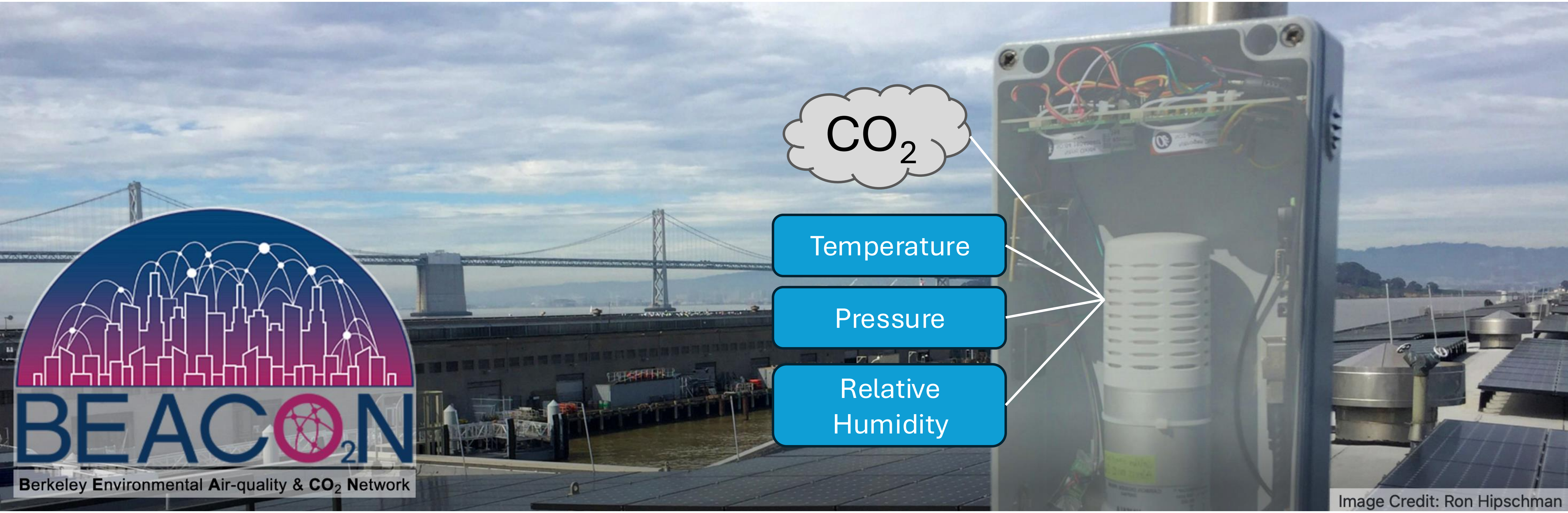
CRQ: How well do the models explored in this study simulate the distribution and variability of intraurban CO_2 concentrations in the San Francisco Bay Area?

SQ1: Can LUR effectively predict intraurban CO_2 concentrations?

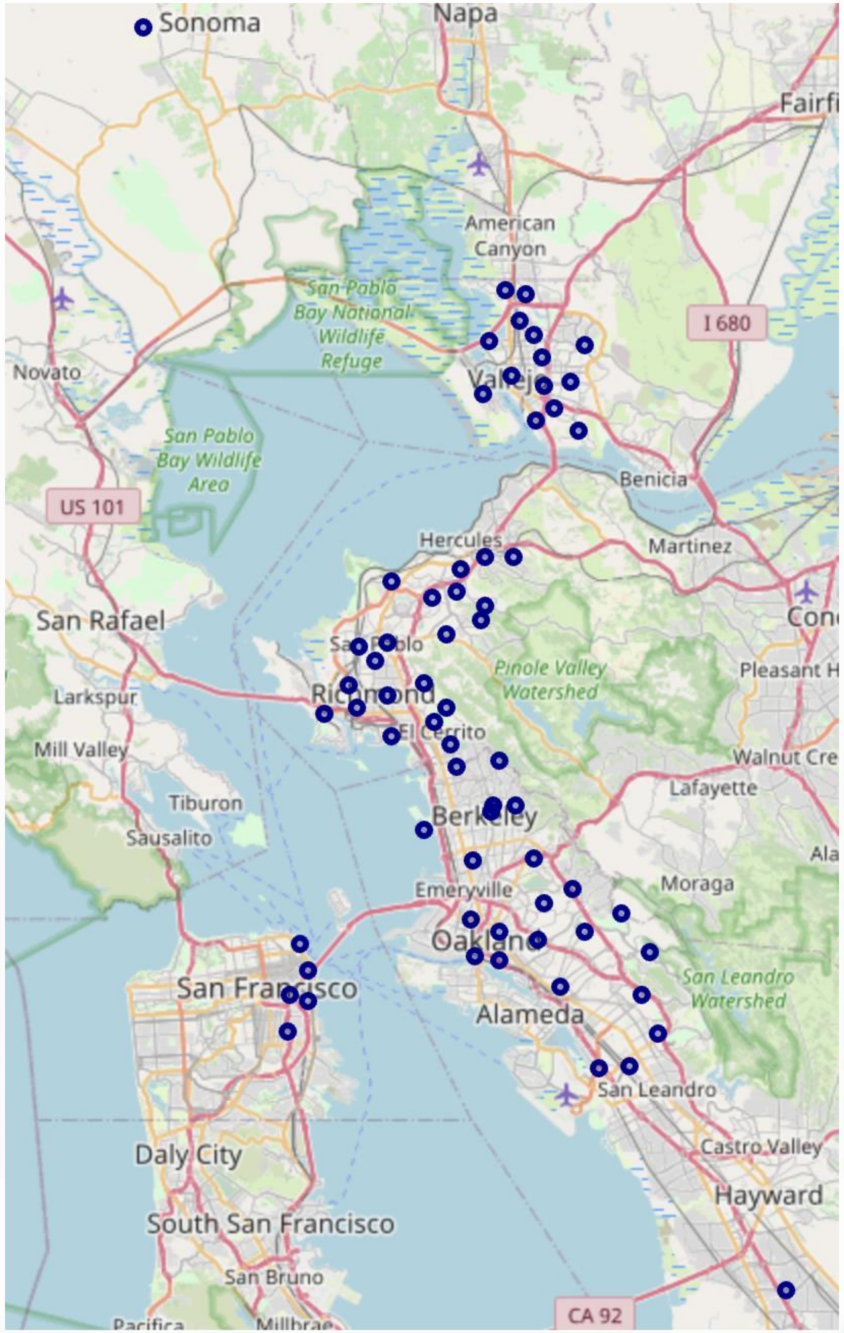
SQ2: Can ML / DL algorithms improve upon LUR model performance?

SQ3: What are the key predictors of intraurban CO_2 concentrations in the San Francisco Bay Area?

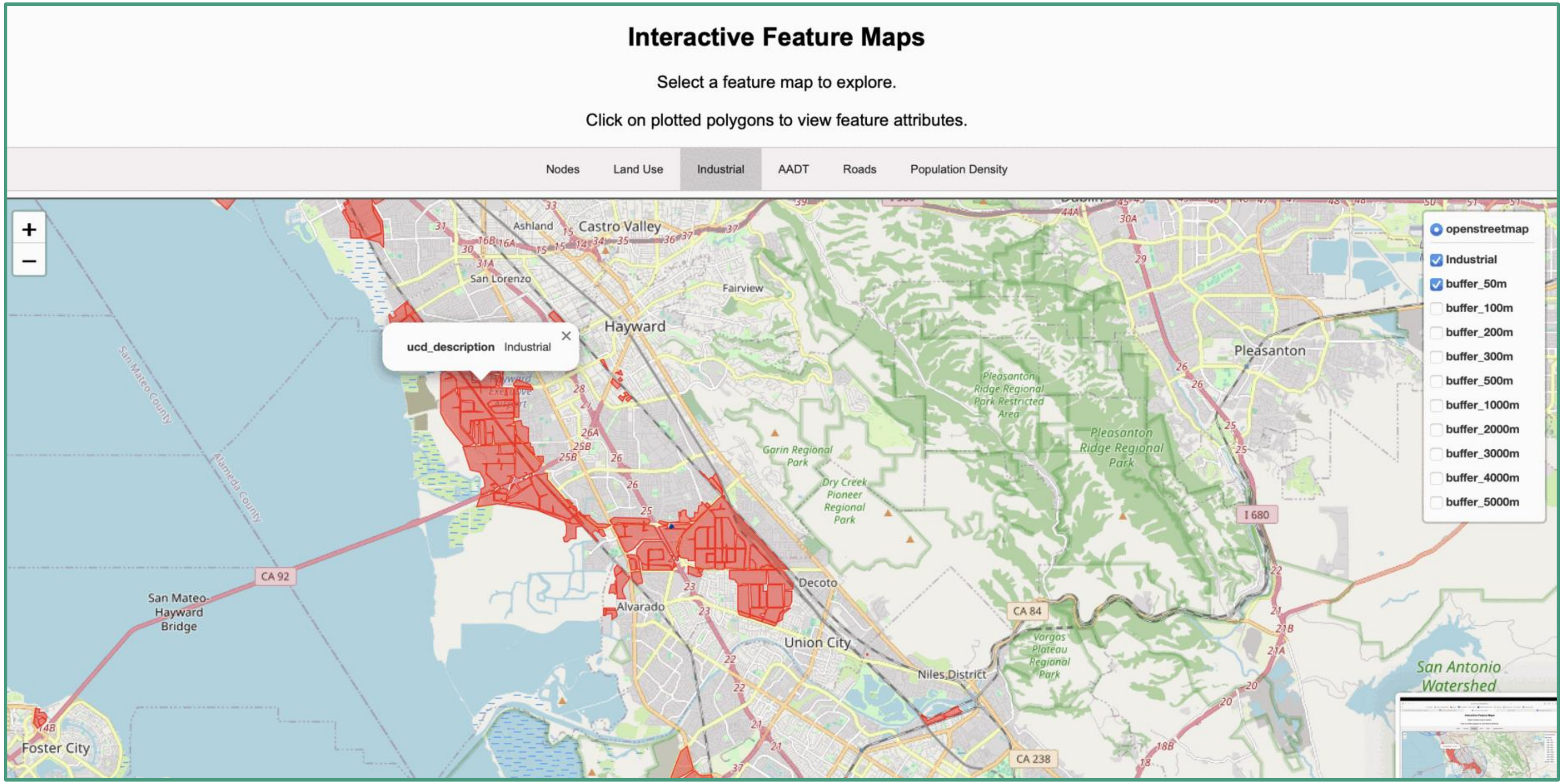
BEACO$_2$N Data



BEACO$_2$N Data

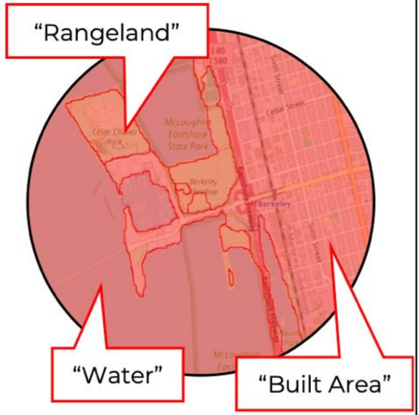


Feature Data Collection & Extraction



Feature Data Collection & Extraction

Land Use & Industrial

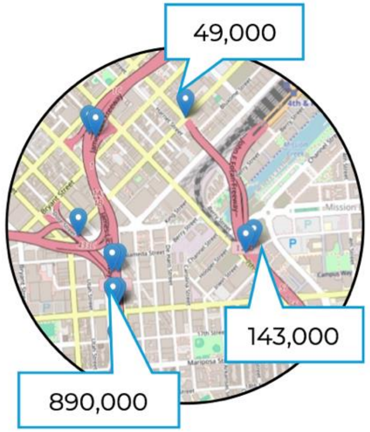


Total area [m^2] in buffer, per Land Use type

$$\sum_{i=1}^{n} (\text{area})_i ,$$

$n =$ total # polygons per LU type in buffer

Annual Average Daily Traffic



Total AADT sum in buffer

$$\sum_{i=1}^{n} (\text{AADT})_i ,$$

$n =$ total # AADT observations in buffer

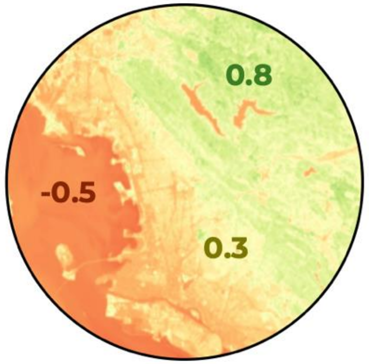
Road Length

Total length of roads [m] in buffer

$$\sum_{i=1}^{n} (\text{road length})_i ,$$

$n =$ total # of roads in buffer

NDVI

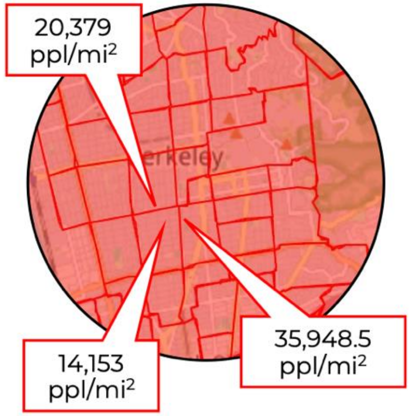


Average NDVI in buffer

$$\frac{1}{n} \sum_{i=1}^{n} (\text{NDVI})_i ,$$

$n =$ total # of pixels in buffer

Population Density

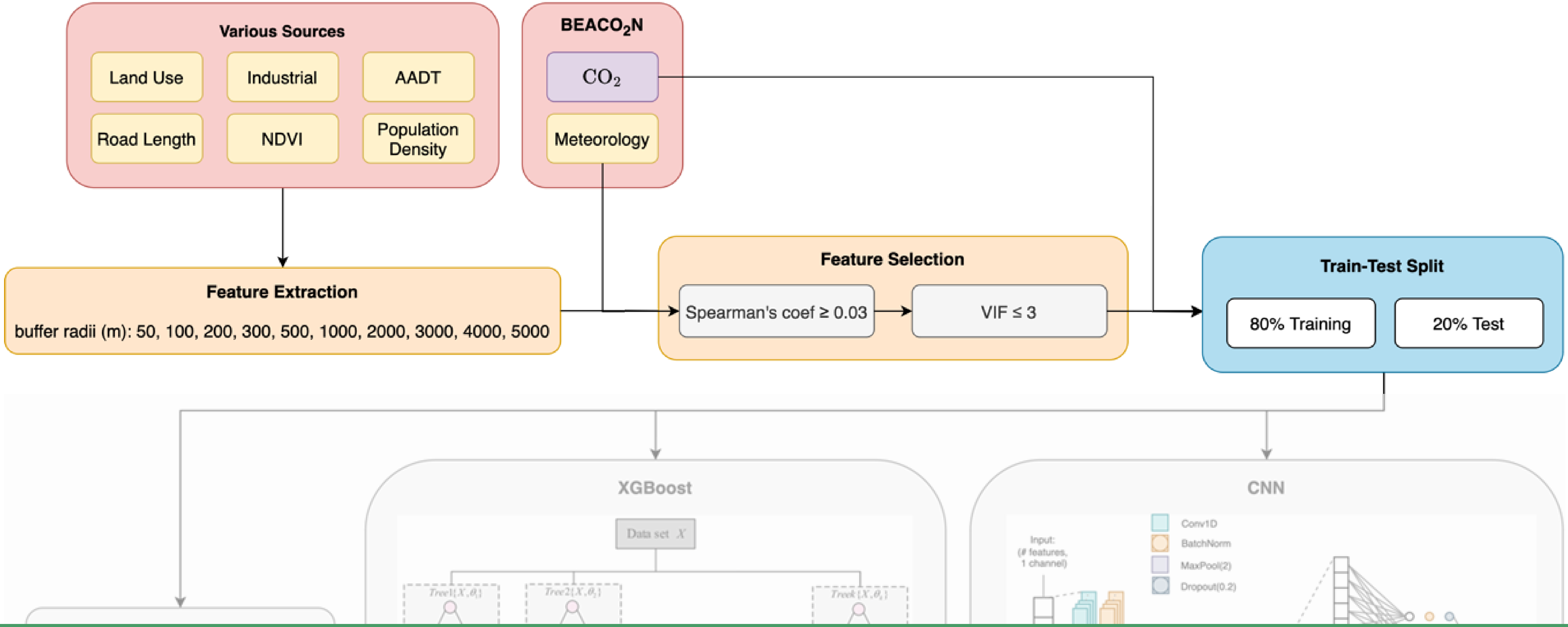


Total population density in buffer area

$$\frac{1}{\text{buffer area } [km^2]} \sum_{i=1}^{n} \left(\frac{\text{ppl}}{\text{mi}^2}\right)_i \times (mi^2)_i ,$$

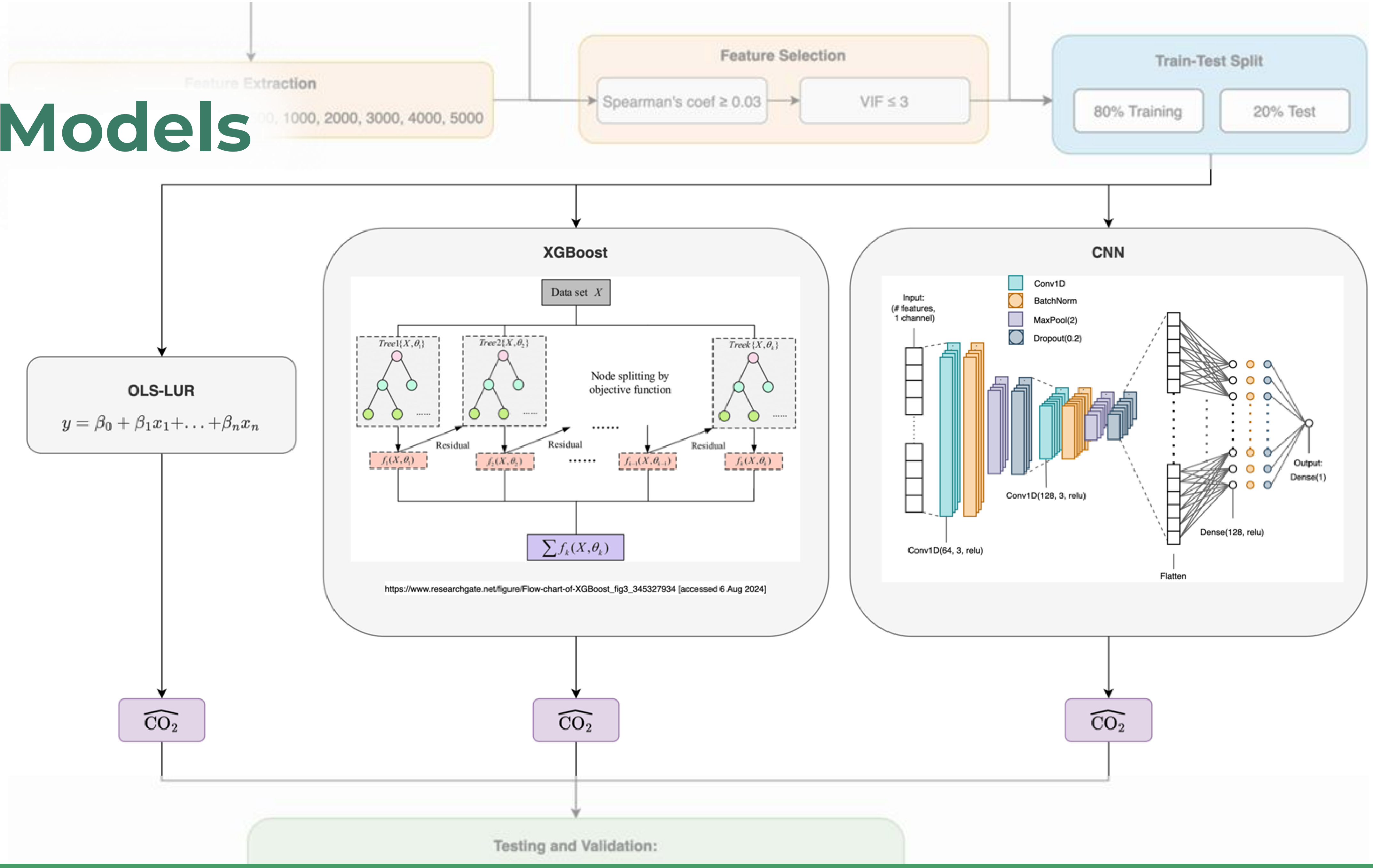
$n =$ total # of polygons in buffer

Methods



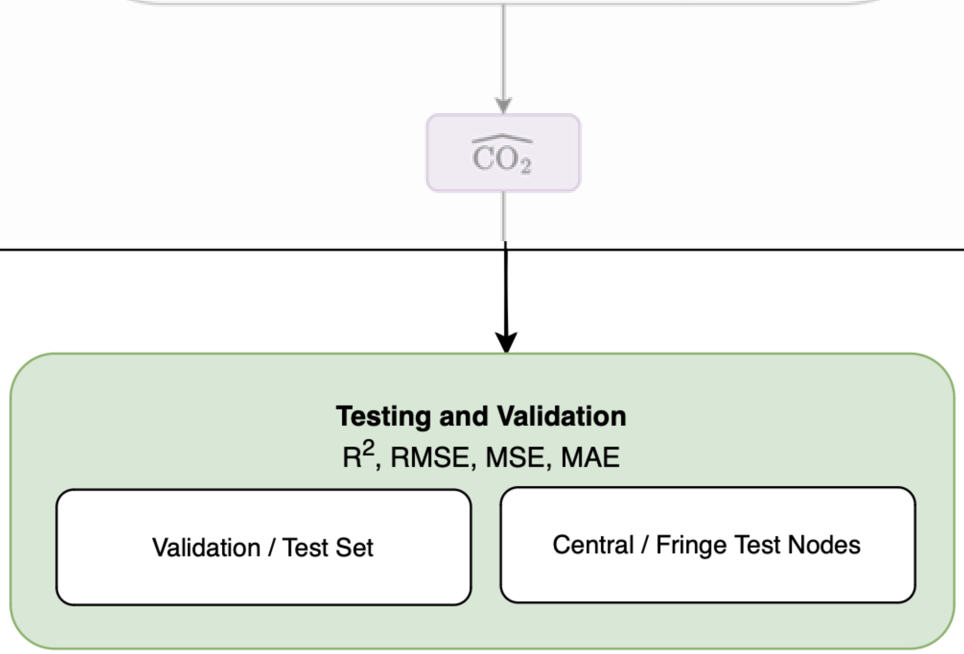
Various Sources
Land Use
Industrial
AADT
Road Length
NDVI
Population Density
BEACO$_2$N
CO_2
Meteorology
Feature Extraction
buffer radii (m): 50, 100, 200, 300, 500, 1000, 2000, 3000, 4000, 5000
Feature Selection
Spearman's coef ≥ 0.03
VIF ≤ 3
Train-Test Split
80% Training
20% Test
XGBoost
CNN
Data set X
Conv1D
BatchNorm
MaxPool(2)
Dropout(0.2)

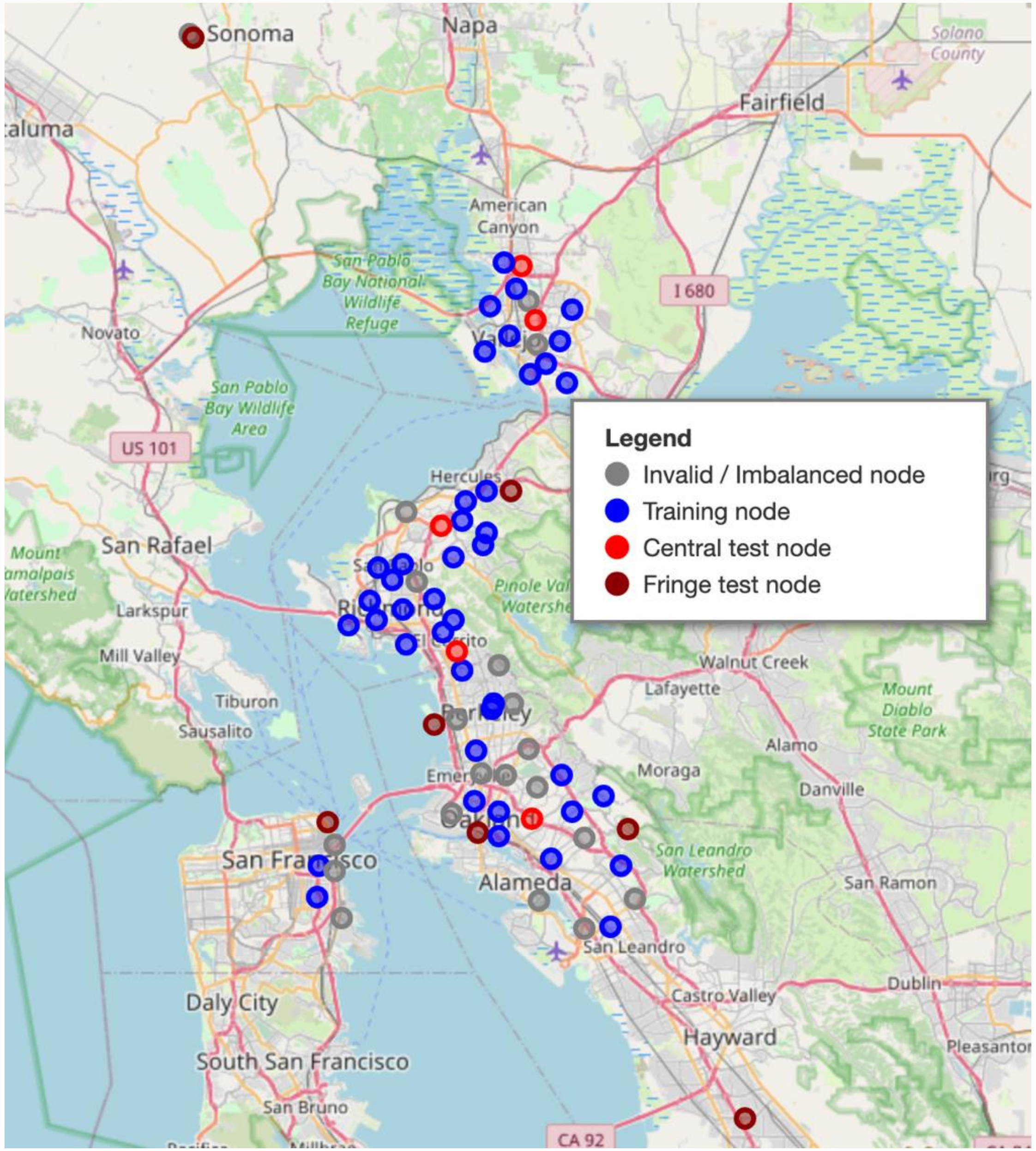
Models



Feature Extraction
00, 1000, 2000, 3000, 4000, 5000
Feature Selection
Spearman's coef ≥ 0.03
VIF ≤ 3
Train-Test Split
80% Training
20% Test
OLS-LUR
$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$
XGBoost
Data set X
Node splitting by objective function
Residual
$\sum f_k(X, \theta_k)$
https://www.researchgate.net/figure/Flow-chart-of-XGBoost_fig3_345327934 [accessed 6 Aug 2024]
CNN
Conv1D
BatchNorm
MaxPool(2)
Dropout(0.2)
Input: (# features, 1 channel)
Conv1D(64, 3, relu)
Conv1D(128, 3, relu)
Flatten
Dense(128, relu)
Output: Dense(1)
$\widehat{CO_2}$
$\widehat{CO_2}$
$\widehat{CO_2}$
Testing and Validation:

Model Evaluation





Feature Selection

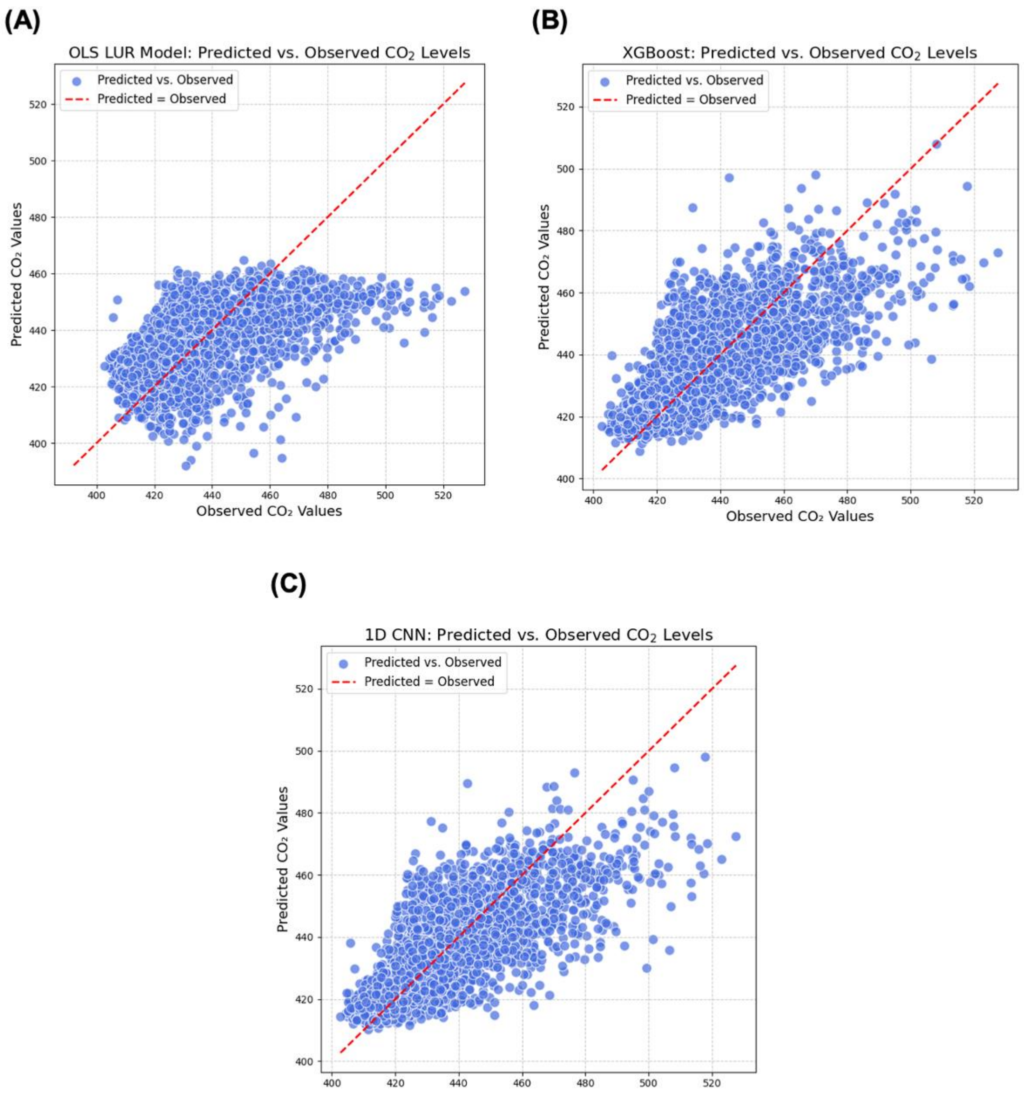
11 selected features:

- Temperature
- Pressure
- Relative humidity
- Trees (50 m)
- Total road length (1000 m)
- Total road length (200 m)
- Built area (2000 m)
- Total AADT (3000 m)
- Flooded Vegetation (1000 m)
- Industrial area (5000 m)
- Average NDVI (1000 m)

Feature	Spearman's ≥ 0.03	VIF < 3
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	-

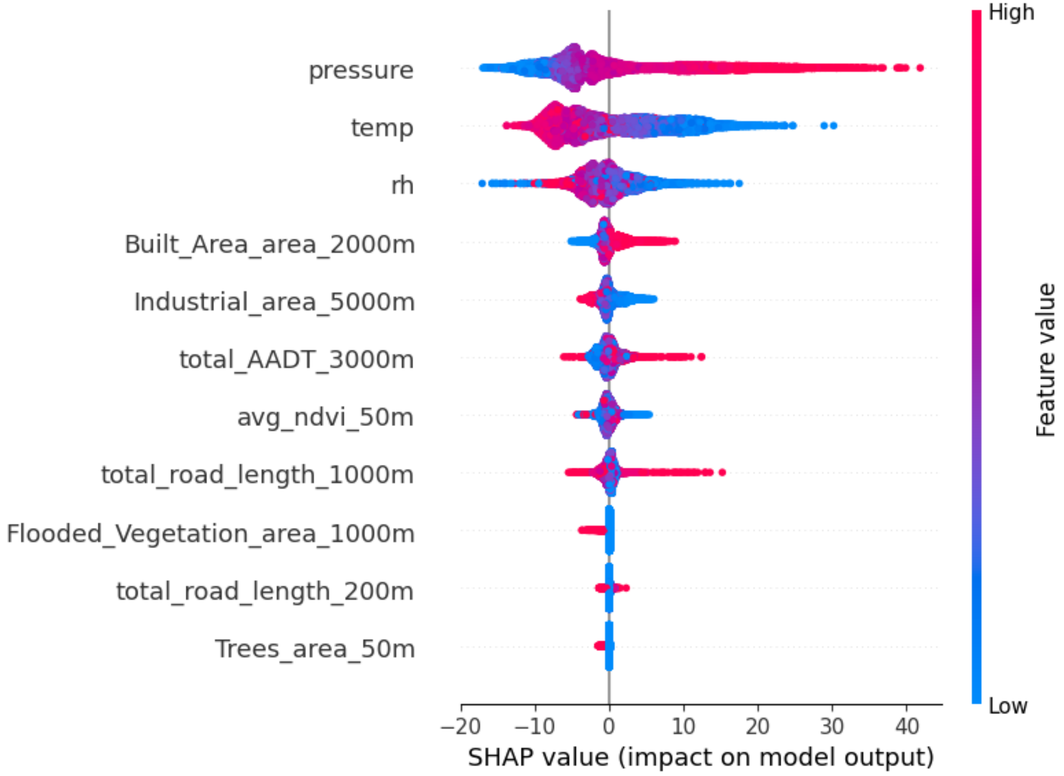
Model Performance

Evaluation Step	Metric	LUR	XGBoost	CNN
Test Set (20% of training node 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
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

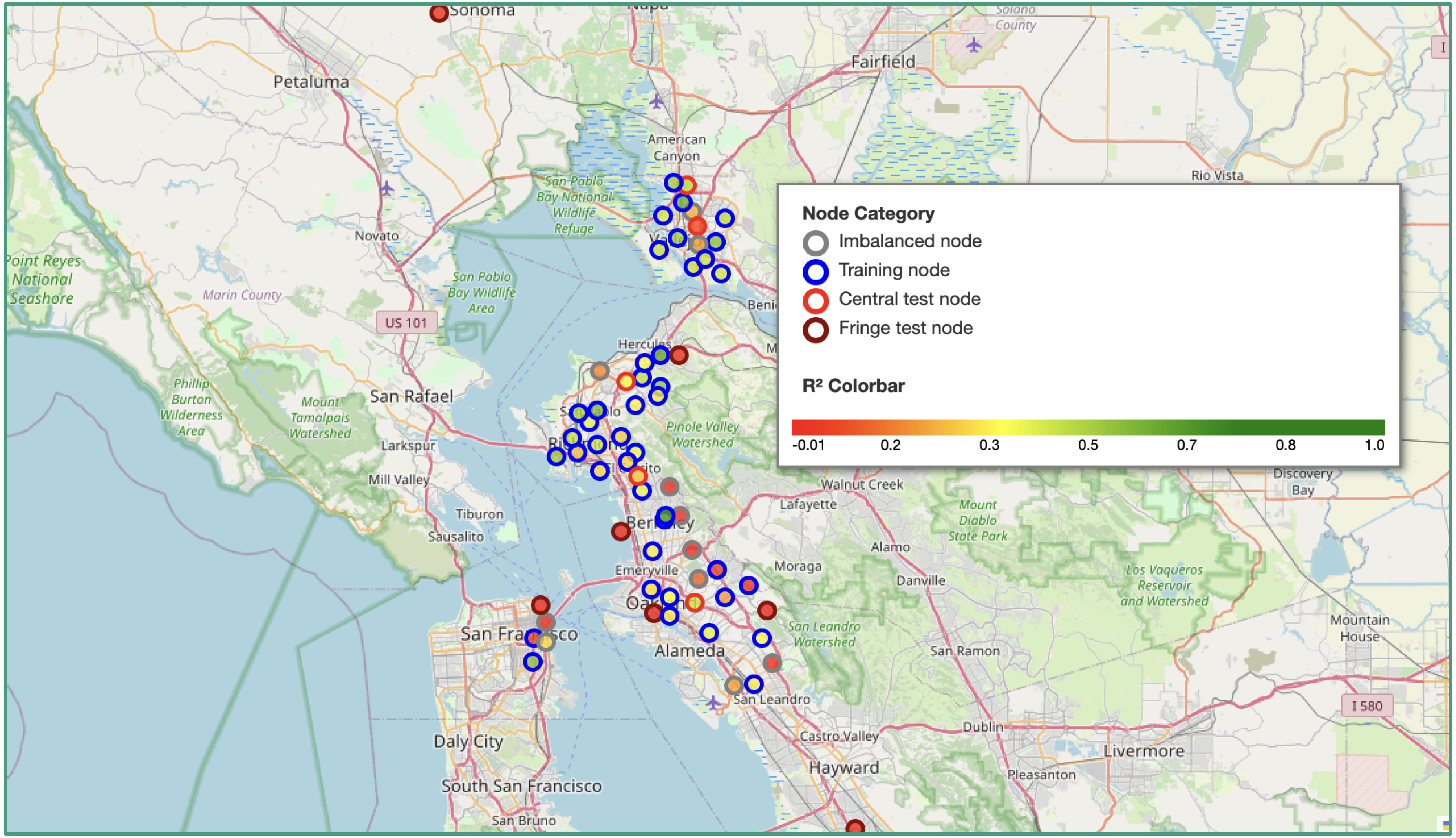


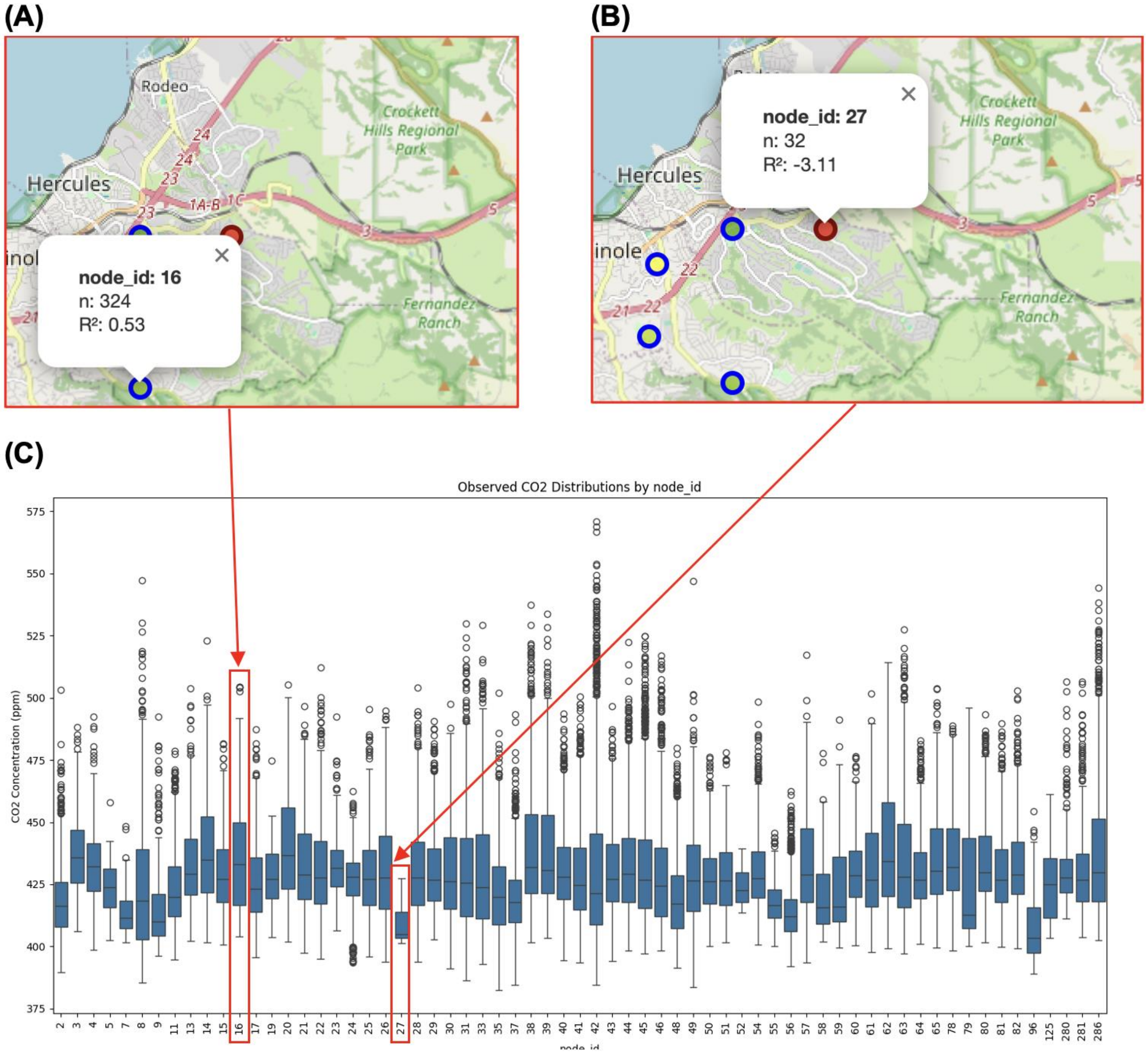
Feature Importance

Feature	Spearman	Partial R^2 (LUR)
temp	-0.51	0.16
pressure	0.40	0.14
rh	-0.11	0.03
Trees_area_50m	-0.08	0.05
total_road_length_1000m	0.04	<0.005
total_road_length_200m	0.04	<0.005
Built_Area_area_2000m	0.04	0.01
total_AADT_3000m	0.04	<0.005
Flooded_Vegetation_area_1000m	-0.03	<0.005
Industrial_area_5000m	0.03	0.01
avg_ndvi_50m	-0.03	<0.005



Spatial Trends



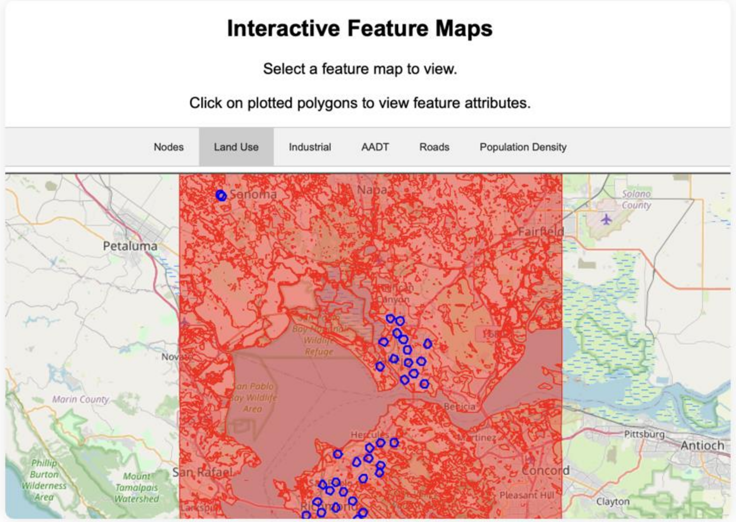


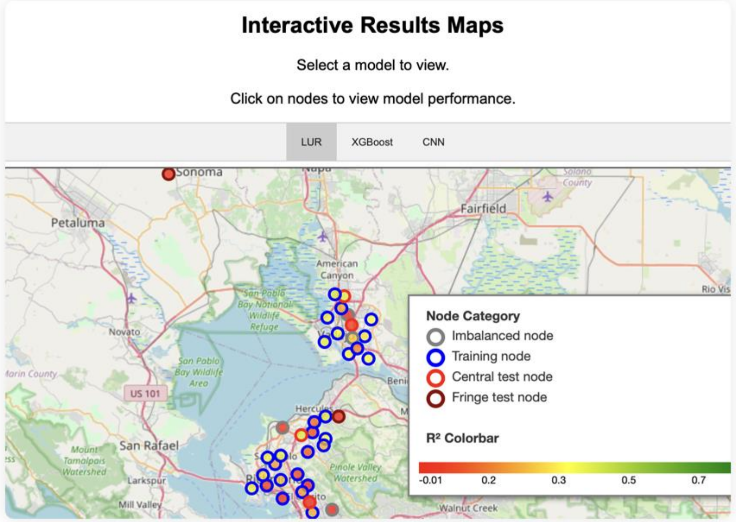
(A)
Rodeo
Crockett Hills Regional Park
Hercules
node_id: 16
n: 324
R²: 0.53
Fernandez Ranch
(B)
node_id: 27
n: 32
R²: -3.11
(C)
Observed CO2 Distributions by node_id
CO2 Concentration (ppm)
575
550
525
500
475
450
425
400
375
2 3 4 5 7 8 9 11 13 14 15 16 17 19 20 21 22 23 24 25 26 27 28 29 30 31 33 35 37 38 39 40 41 42 43 44 45 46 48 49 50 51 52 54 55 56 57 58 59 60 61 62 63 64 65 78 79 80 81 82 96 125 280 281 286
node_id

https://ese-msc-2023.github.io/irp-acs223/

bayareaco2 Explore

Welcome to the bayareaco2 Explore page. This interactive page is designed to help visualize the data and results associated with the bayareaco2 prediction models. The Feature Explorer demonstrates the feature data used to train the models, as well as the BEACO2N sensor locations. The Results Explorer gives a spatial representation of model performance. You can explore performance metrics for different models and for individual nodes. The data plotted on these maps is meant to be interacted with, so feel free to click and scroll around! Some feature maps have been optimized for the webpage interface. Select an Explore page below to get started:





Feature Explorer

Results Explorer

About This Site

This page is hosted by GitHub Pages from the irp-acs223 GitHub repository. The work presented here was completed by Anna C. Smith, under the supervision of Fangxin Fang and Linfeng Li in completion of her MSc in Environmental Data Science and Machine Learning at Imperial College London. Consult the repository for more code and information about the project. Contact anna.smith23@imperial.ac.uk with any questions.

Contributions

1. Using BEACO_2N data for modeling of intraurban CO_2 concentrations in San Francisco Bay Area
2. Application of LUR to predict CO_2 concentrations
3. Using XGBoost and CNN to predict CO_2 concentrations
4. Using unseen node data to test transferability

Conclusions

- XGBoost and CNN consistently outperformed LUR
- Overall weak transferability to distant unseen locations
- Feature importance related to variability and abundance
- Spatial features limited by constant temporal resolution
- Model performance limited by validity and consistency of CO_2 data

Need for more spatiotemporally distributed CO_2 sensor networks !

Thank you! ☺

M.Sc. Environmental Data Science and Machine Learning

IRP Presentation

September 9th, 2024

Anna C. Smith