International Journal of Geo-Information

MDPI

Supplementary Information

Black Carbon Concentration Estimation with Mobile-Based Measurements in a Complex Urban Environment

Minmeng Tang [1,2,*], Tri Dev Acharya [3] and Deb A. Niemeier [4]

[1] Department of Land, Air, and Water Resources, University of California Davis, Davis, CA 95616, USA
[2] Department of Civil and Environmental Engineering, Cornell University, New York, NY 14850, USA
[3] Institute of Transportation Studies, University of California Davis, Davis, CA 95616, USA; tdacharya@ucdavis.edu
[4] Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn Martin Hall, College Park, MD 20742, USA; niemeier@umd.edu
* Correspondence: mmtang@ucdavis.edu

This supplementary PDF file includes the following 5 sections including references:

1. Land use variables
2. RF model tuning parameters
3. SVR model feature selection, dimension reduction, and model tuning
4. LASSO model regression coefficients
5. Sensitivity analysis with the five most sensitive features
6. Reference

1. Land use variables

The land use variables are calculated using Messier et al [1] as a guide and the detailed instructions are available in the Messier et al. (2018) supplementary material. In general, we construct 108 land use variables from the following datasets with six buffer sizes including 50 m, 100 m, 250 m, 500 m, 1000 m, and 2500 m:

we construct the binary road classification variable based on the OpenStreetMap dataset, which is open-source data that provides roads, trails, cafes, railway stations, and so on all over the world[2]. In the OpenStreetMap dataset, the road systems are classified into multiple categories based on their importance within the local road system as a whole. To simplify the road classification variable, we construct three categories from tens of categories in the OpenStreetMap dataset, which are highways, arterials, and residential roads. The highways category contains a motorway, motorway link, and trunk link; the arterials category contains a primary, primary link, secondary, secondary link, service, tertiary, tertiary link, and unclassified; the residential category contains living street and residential.

For each category (highways, arterials, and residential roads), we calculate the total road length within each buffer size based on the OpenStreetMap data[2].

Two binary variables indicating whether a road segment is on a designated heavy-duty truck route or on a road where heavy-duty trucks are prohibited are created based on the Oakland Truck Routes (2017) report created by the city of Oakland which is available online at https://www.arcgis.com/home/item.html?id=0fe7f165a9274b1182002ff9c0f4851d[3].

Three binary variables indicating commercial, industrial, and residential zonings are created based on the City of Oakland zoning classifications, which is available online at https://oakgis.maps.arcgis.com/apps/webappviewer/index.html?id=3676148ea4924fc7b75e7350903c7224[4]. These three land use zonings are

Citation: Tang, M.; Acharya, T.D.; Niemeier, D.A. Black Carbon Concentration Estimation with Mobile-Based Measurements in a Complex Urban Environment. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 290. https://doi.org/10.3390/ijgi12070290

Academic Editors: Wolfgang Kainz, Xiao Li, Xiao Huang and Zhenlong Li

Received: 27 April 2023
Revised: 2 July 2023
Accepted: 18 July 2023
Published: 20 July 2023

generalized from the complex city zoning codes based on the explanations in Table S2 in the supplementary material in Messier et al. (2018)[1].

We also calculate the average Normalized Difference Vegetation Index (NDVI) within each buffer to represent the coverage of vegetation around each road segment. The NDVI is calculated based on the measurement from LANDSAT 8 in Google Earth Engine[5].

To include land cover information in the LUR model, we created 6 land cover types based on the USGS National Land Cover Database (NLCD) 2016[6,7] at 30 m resolution. These 6 land use types are Developed Open, Developed low, Developed medium, Developed high, evergreen forest, and mixed forest. For each land cover type, six buffer sizes are used to calculate the corresponding variables, except evergreen forest and mixed forest, which only have 2500 m buffer since the rest buffer sizes lead to variables with all zeros. For the land cover variables, the percentage of the land cover type within the buffer area is calculated by using the number of pixels representing the corresponding land cover type divided by the total number of pixels within the buffer area.

We use the 2010 census tract population data[8] to calculate the population density within each buffer. By assuming the population is evenly distributed within each census tract, we can compute the population density of each buffer based on buffer area and population density in each intersected census tract.

To calculate the mean elevation within each buffer area, we use the National Elevation Data (NED)[9] in Google Earth Engine[5], which has approximately a 10 m resolution in the US.

Some other point sources may also contribute to air pollution concentrations. To include the point source contributions in the LUR model, we select four point source categories, which are ports, airports, National Priority Listing (NPL) sites, and Toxic Release Inventory (TRI) sites, and calculate the exponentially decayed contributions from these sources by using Equation 1 in supplementary from Messier et al. (2018)[1]. We use 8 decay distances (λ_l) including 50 m, 100 m, 500 m, 1,000 m, 2,500 m, 5,000 m, 10,000 m, and 50,000 m for all the 4-point sources listed above.

Ports data is downloaded from the Bureau of Transportation Statistics (https://data-usdot.opendata.arcgis.com/datasets/major-ports/data). Airports data is also available from the Bureau of Transportation Statistics (http://osav-usdot.opendata.arcgis.com/datasets/831853ab8b714a81b6a3e21d0b164a4e_0). All ports and airports within the US are used in the calculation of point source contributions. NPL data is available from US Environmental Protection Agency (USEPA) (https://epa.maps.arcgis.com/home/item.html?id=c2b7cdff579c41bbba4898400aa38815#overview). We only use the NPL sites within Alameda County to prepare the point source contribution variables. TRI data is also available from USEPA (https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-2018). We use the TRI sites within the surrounding 10 counties to prepare the corresponding variables.

Meanwhile, we also calculate the inverse distance of the nearest point sources for each road segment with respect to all the 4 abovementioned point sources.

2. RF model tuning parameters

We are using the scikit-learn package in Python to train the Random Forest (RF) regression model[10]. Within this package, the RF regression model has multiple hyper-parameters, and we are tuning five free parameters with the rest parameters using the default values, which are max_depth, max_features, n_estimators, min_samples_split, and min_samples_leaf. To achieve better performance, we introduce the Hyperopt optimization algorithm, which uses a form of Bayesian optimization methods to consider information from previous iterations. Figure S1 shows the general steps to use Bayesian optimization in model tuning with the Hyperopt algorithm. To apply the Hyperopt optimization algorithm, we need to first pre-define the search space, which limits the hyper-

parameters range. The pre-defined search space for the RF regression model is shown in Table S1.

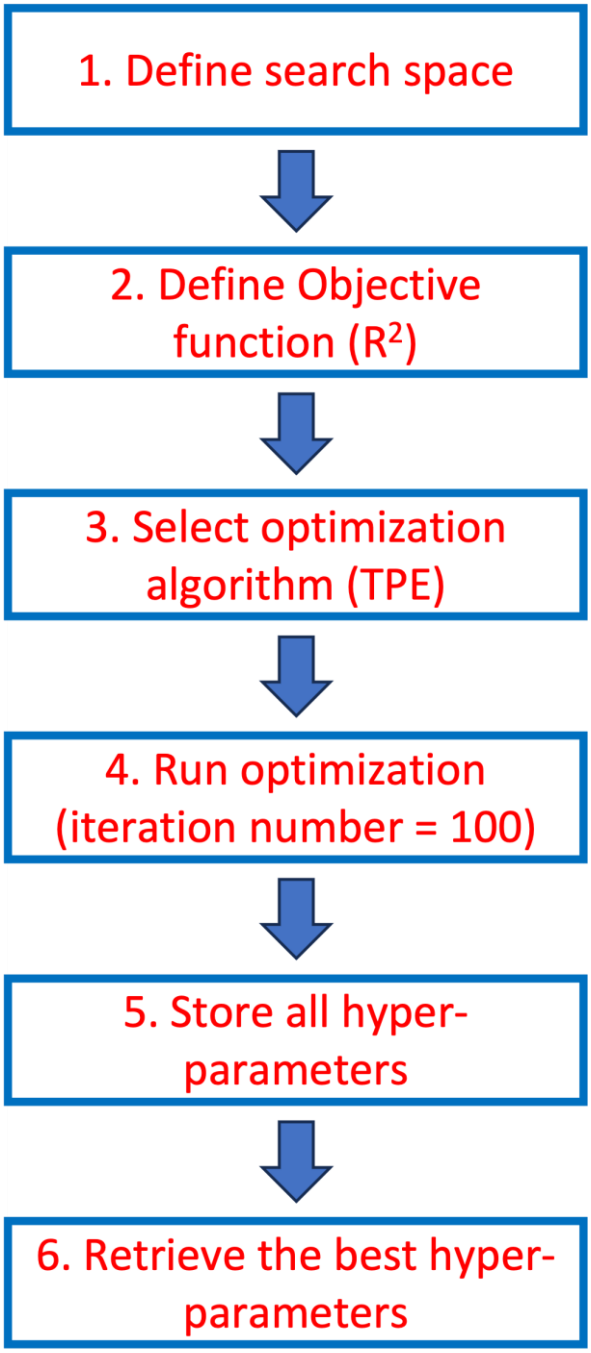

Figure S1. General procedure for using Hyperopt optimization algorithm to tune machine learning models (TPE is Tree of Parzen Estimators, which is a widely used optimization algorithm in Hyperopt algorithm).

Table S1. Search space for RF model hyper-parameters.

Hyper-parameters	Values
max_features	'auto', 'sqrt', 'log2', 1, 0.5, 0.1, 0.05, 0.01

	Minimum value	Maximum value	Increment

max_depth	10	50	1
n_estimators	450	1000	1
min_samples_split	2	10	1
min_samples_leaf	2	10	1

With the Hyperopt optimization algorithm, the optimized hyper-parameters for the RF regression model are listed in Table S2 below. Based on these hyper-parameters' value, we train the RF model over the train set and predict BC concentrations over the validation set, which is then used to calculate the RF model's performance.

Table S2. Tuned values of RF model hyper-parameters.

Hyper-parameters	Optimized value
max_features	0.5
max_depth	24
n_estimators	781
min_samples_split	5
min_samples_leaf	2

3. SVR model feature selection, dimension reduction, and model tuning

3.1 FOCI feature selection

The FOCI method selects 13 features out of the 108 input features and the 13 selected features are listed in Table S3. Based on the FOCI selected features, we use the Hyperopt optimization algorithm to train the SVR model over the train set and make predictions over the validation set, which is used to evaluate SVR model performance.

Table S3. The FOCI method selected 13 features for the SVR model.

NDVI_1000	port_5000	residential_2500	residential_land_type
Truck_prob	port_500	population_1000	Commercial_land_type
Truck_route	airport_100	Mixed_forest_2500	Industrial_land_type
Developed_open_2500			

3.2 Genetic Algorithm (GA) parameter setup

In the GA model, the "chromosome" is set to have two sections. The first section contains three hyper-parameters of the SVR model, which are regularization parameter C, kernel coefficient gamma, and the loss function penalty parameter epsilon; the second section contains the selected features, which are the ID numbers of the corresponding features. Given the different nature of these two sections, the mating and mutation processes are calculated separately. The detailed mating and mutation processes are available in Zhang et al. (2015)[11]. In our GA model, the initial population size is set to be 100, a mating cross rate of 0.8, a mutation rate of 0.1, elite size 2, and offspring number 8. The GA-optimized hyper-parameters of the SVR model are listed in Table S4 and the corresponding features are given in Table S5.

Table S4. GA optimized hyper-parameters of the SVR model.

Hyper-parameter	GA optimized value
C	2.35478
gamma	0.0831661
epsilon	0.129904

Table S5. GA selected 45 features for the SVR model.

Developed_open_250	NDVI_100	npl_50	Population_250
Developed_open_500	NDVI_250	npl_100	Population_500
Developed_open_1000	NDVI_500	npl_500	Population_1000
Developed_open_2500	NDVI_1000	npl_1000	Population_2500
Developed_low_250	tri_50	npl_2500	port_500
Developed_low_1000	tri_2500	npl_5000	port_1000
Developed_low_2500	tri_10000	highway_100	port_5000
Developed_medium_250	residential_50	highway_1000	port_50000
Developed_medium_2500	residential_500	Elevation_100	port_inverse_dist
Developed_high_100	airport_1000	Index_Hwy	Latitude
Developed_high_250	npl_10000	truck_route	Speed_Med
Developed_high_1000			

4. LASSO model regression coefficients

The LASSO model gives 77 non-zero features. These features and the coefficients are listed in Table S6.

Table S6. LASSO selected features and the coefficients.

Feature	Coefficient	Feature	Coefficient
intercept	-0.5849	Developed_low_50	0.0108
Road_Type	-0.0548	Developed_low_100	0.0130
Index_Hwy	0.1299	Developed_low_500	-0.0429
truck_route	0.1655	Developed_low_1000	-0.0067
truck_prob	-0.0281	Developed_low_2500	0.0897
commercial	0.0177	Developed_medium_50	0.0012
industrial	0.0659	Developed_medium_100	0.1765
Speed_Med	0.0993	Developed_medium_250	0.0694
highway_50	-0.0145	Developed_medium_500	0.0265
highway_100	0.1065	Developed_medium_1000	-0.0041
highway_250	-0.0182	Developed_high_50	0.0089
highway_500	-0.0114	Developed_high_100	0.2376
highway_1000	-0.0796	Developed_high_250	0.0941
highway_2500	-0.0593	Developed_high_1000	0.0695
arterial_50	0.0289	Developed_high_2500	-0.2894
arterial_100	-0.0316	Mixed_forest_2500	0.0041
arterial_250	-0.0199	Population_100	0.0269
arterial_500	-0.0247	Population_250	0.0356
arterial_1000	0.0225	Population_500	-0.0537
arterial_2500	0.0678	Population_1000	-0.0623
residential_50	-0.0507	Population_2500	-0.1372
residential_100	-0.0026	port_100	-0.0036
residential_250	-0.0168	port_500	-0.0484
residential_500	-0.0584	port_1000	0.1377
residential_2500	0.5946	airport_50	0.0018
NDVI_50	0.0034	airport_100	-0.0034
NDVI_100	0.0272	airport_500	-0.1449
NDVI_2500	0.0697	airport_1000	0.3026

Elevation_50	-0.0978	airport_10000	-0.2506
Elevation_100	-0.0919	npl_500	0.0004
Elevation_250	-0.0523	npl_1000	-0.1053
Developed_open_50	-0.0051	npl_5000	0.0220
Developed_open_250	0.0090	npl_10000	0.2574
Developed_open_500	0.0568	npl_50000	0.1345
Developed_open_1000	0.0069	tri_50	0.0571
Developed_open_2500	-0.1301	tri_100	-0.0266
tri_inverse_dist	-0.0370	tri_1000	0.0734
npl_inverse_dist	0.0095	tri_5000	-0.0984
tri_10000	-0.0013	tri_50000	-0.0530

5. Sensitivity analysis with the five most sensitive features

The top 5 most sensitive features for all four models based on the OAT sensitivity analysis are shown in Figure S2. For each column from top to bottom, it shows the five most sensitive features (most sensitive to least sensitive) for a specific model.

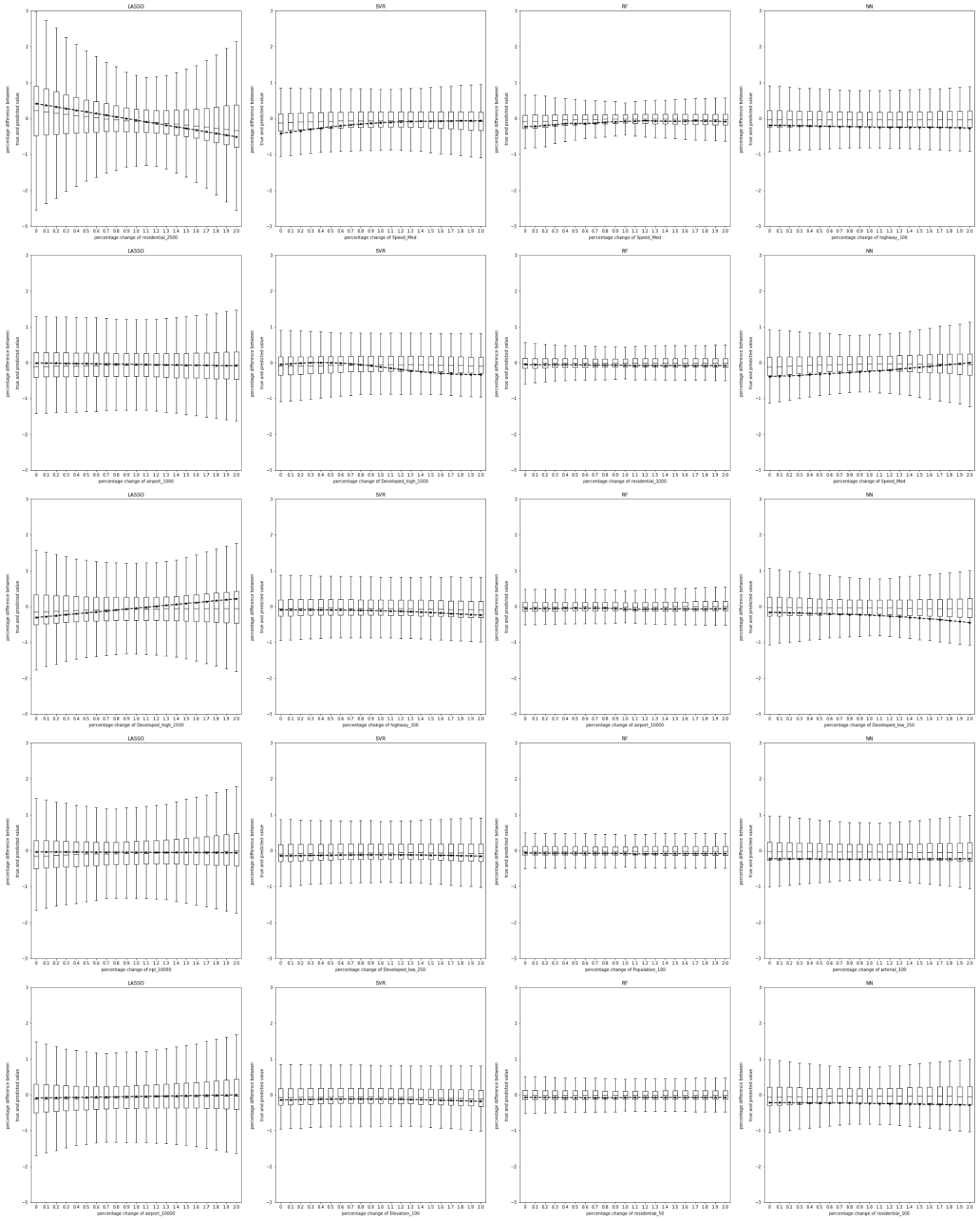

Figure S2. Top 5 most sensitive features for each model and how their variations influence model performance in BC prediction (box shows 25th, 50th, and 75th percentiles; dot means mean value).

References

1. Messier, K.P.; Chambliss, S.E.; Gani, S.; Alvarez, R.; Brauer, M.; Choi, J.J.; Hamburg, S.P.; Kerckhoffs, J.; LaFranchi, B.; Lunden, M.M.; et al. Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression. *Environ. Sci. Technol.* **2018**, *52*, 12563–12572, doi:10.1021/acs.est.8b03395.
2. Contributors, O.S.M. OpenStreetMap.
3. Davidlok Oakland Truck Routes (2017).
4. City of Oakland Planning and Zoning Map.
5. Gorelick, Noel and Hancher, Matt and Dixon, Mike and Ilyushchenko, Simon and Thau, David and Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, doi:10.1016/j.rse.2017.06.031.
6. USGS NLCD 2016 Land Cover (CONUS).
7. Yang, L.; Jin, S.; Danielson, P.; Homer, C.; Gass, L.; Bender, S.M.; Case, A.; Costello, C.; Dewitz, J.; Fry, J.; et al. A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 108–123, doi:https://doi.org/10.1016/j.isprsjprs.2018.09.006.
8. U.S. Census Bureau Population.
9. Gesch, D.B.; Evans, G.A.; Oimoen, M.J.; Arundel, S. The National Elevation Dataset. In; American Society for Photogrammetry and Remote Sensing, 2018; pp. 83–110.
10. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
11. Zhang, D.; Xiao, J.; Zhou, N.; Zheng, M.; Luo, X.; Jiang, H.; Chen, K. A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction. *Biomed Res. Int.* **2015**, *2015*, doi:10.1155/2015/292683.