Towards Building an Optimal LUR Model for Air Quality Prediction Using Machine Learning Approach

Dinushani Senarathna

Department of Mathematics

July 30 2020

Mentors: Dr. Sumona Mondal, Dr. Suresh Dhaniyala, Dr. Shantanu Sur, Dr. Supraja Gurajala
Collaborators: Vijay Kumar

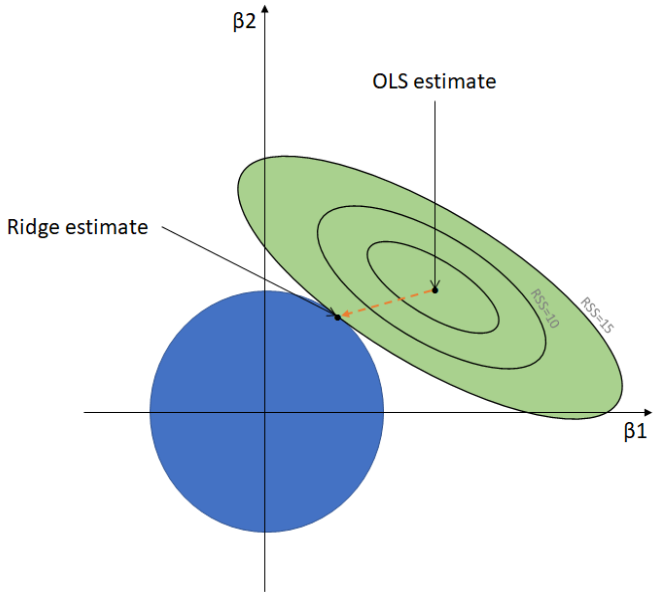
Introduction

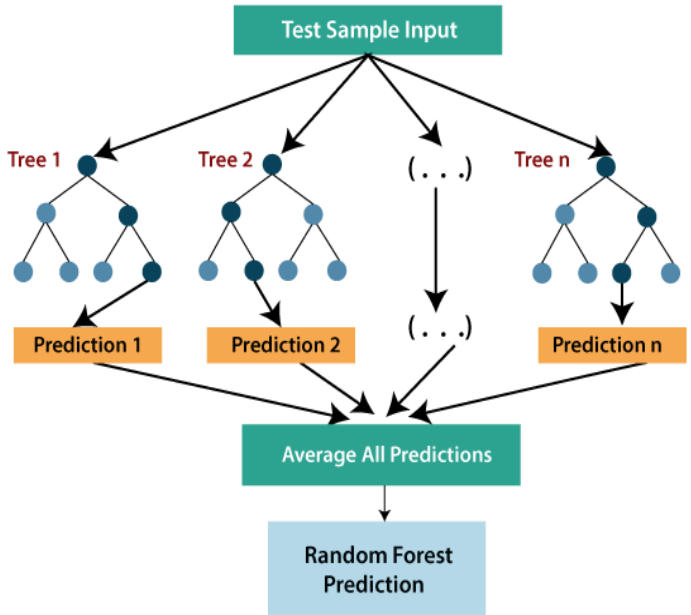
- **Motivation:** Environmental epidemiology requires knowledge of variations in the ambient air quality at high spatiotemporal precision.
- **Objective:** Apply machine learning-based Random Forest Regression (LURF) technique against LUR and Ridge Regression techniques using air quality and land use data.
- **Research question:** Compare three different (Ordinary Least Square (OLS), Ridge, and Random Forest) regression techniques, and find out the most optimal technique to predict air quality index more precisely.

Ridge and Random Forest Regression

- **Ridge Regression:** This is use for analyzing multiple regression data when suffering from multicollinearity. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

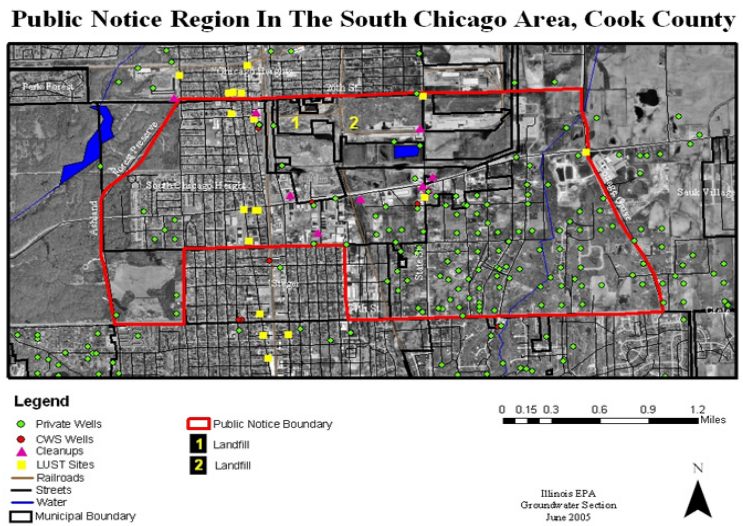
- **Random Forest Regression:** This is use for nonlinear multiple regression. This technique can produce different regression trees which have distribution for the output variable in each leaf.

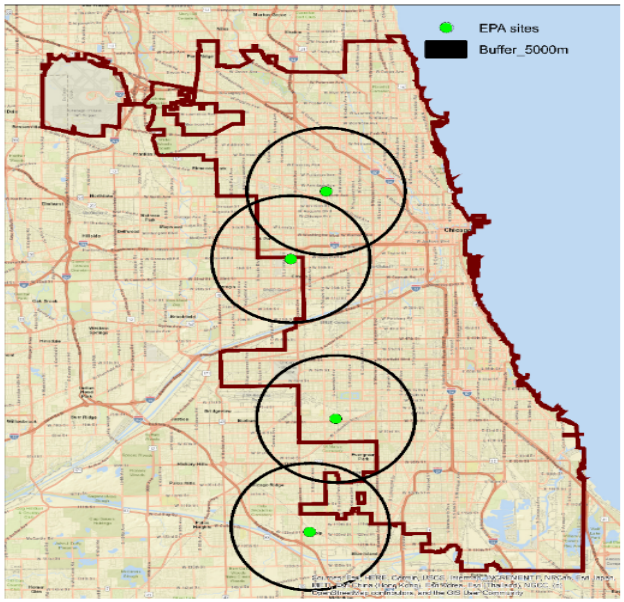




Method-Data Collection

- Choose Chicago Cook County, IL as target area
- Collect Data from EPA site and National Land Cover Data (NLCD) from 2017-2019
- Make buffers (Land use buffer analysis is used to quantify features of the urban landscape within a defined distance of a point of interest) from 500m to 5000m for Land used and NLCD variables





Method-Model Building

- Build univariate regression models for $PM_{2.5}$ with all predictors
- Check mulicolinerity between predictors and $PM_{2.5}$
- Select optimal buffers using highest R^2 from each cluster
- Apply three different (OLS, Ridge, and LURF) regression techniques
- Model validation and assumption checking
- Compare and select best model

Results

Variable Selection-Using univariate model R^2 value

Meta Var.	R^2	Adj. R^2
Wind Speed	0.070	
Relative Humidity	0.030	0.080
Outdoor Temperature	0.013	0.101
Wind Direction	0.003	0.101
Barometric Pressure	0.002	0.101

Univariate models for meteorological parameters

Buffer Size	Bus Stop	Annual Average Daily Traffic	Railway Road Length	Street Length
0.5K	0.030	0.0200	0.040	0.021
1.0K	0.017	0.0001	0.170	0.028
2.0K	0.040	0.0001	0.030	0.135
3.0K	0.010	0.0010	0.080	0.140
4.0K	0.010	0.0001	0.135	0.050
5.0K	0.005	0.0003	0.138	0.012
Adj. R^2	0.140	0.274	0.366	0.366

R^2 values of Univariate models for Land Used parameters

- In meteorological parameters, wind speed has the highest model R^2 with $PM_{2.5}$

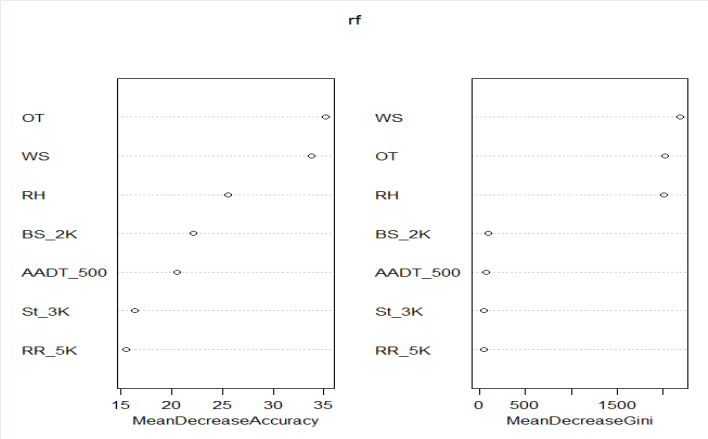
Multiple linear regression vs Ridge regression coefficients

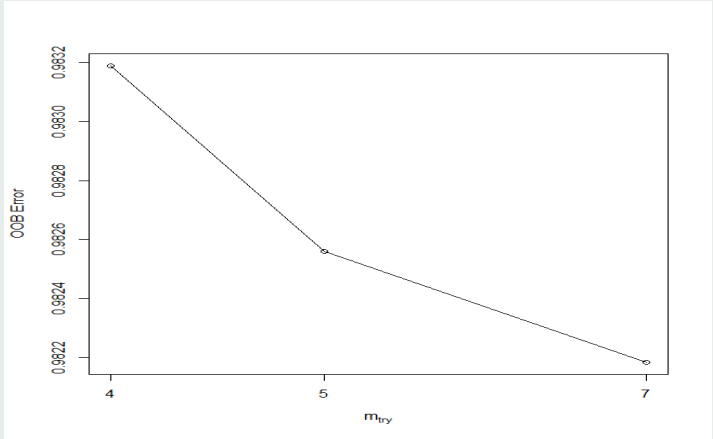
	intercept	WS	RH	OT	BS_2K	AADT_.5K	RR_1K	St_3K	Adj. R^2
OLS 1	31.70	-0.631	0.067	0.068	0.244	-0.0008	NA	-0.0004	.366
OLS 2	31.40	-0.632	0.067	0.068	0.242	-0.0008	-0.0004	NA	.366
Ridge	27.41	-0.634	0.068	0.068	0.137	-0.0005	-0.0005	0.0010	.367

- Both multiple linear and Ridge regression have same model R^2 value as .366.

Results

Variables importance for Random Forest Regression





- Outdoor temperature and wind speed are the most influential predictors for LURF.

Model comparison

Model type	Adj.R^2	RMSE
OLS	0.366	7.256
Ridge	0.367	7.103
Random Forest	0.391	6.904

- Random forest regression has least RMSE and adj. R^2 value.

Conclusion and Future work

- Among all the variables, wind speed played a significant role in model building in all three different regression techniques.
- Under these conditions, Random forest regression has the least RMSE value with the highest model R^2 as.3669 for 4 EPA sites in cook county, Chicago.
- As future work, we plan to investigate, how to improve model performance with more EPA sites and conduct cross- validation technique to validate the model for better air quality prediction.

Reference

1. Philip K. Hopke," Hourly land-use regression models based on low-cost PM monitor data", University of Rochester, Environmental Research 167(2018) 7-14, https://doi.org/10.1016/j.envres.2018.06.052
2. Massimo Stafoggia,"A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in 9Sweden", Lazio Region Health Service/ASL Roma, Atmosphere 2020, 11, 239 doi:10.3390/atmos11030239
3. Marloes Eeftens,"Development of Land Use Regression Models for PM2.5, PM2.5 Absorbance, PM10 and PM coarse in 20 European Study Areas; Results of the ESCAPE Project", Utrecht University, © 2012 American Chemical Society, dx.doi.org/10.1021/es301948k
4. Cole Brokamp, "Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches", University of Cincinnati, Atmospheric Environment 151(2017)1-11, http://dx.doi.org/10.1016/j.atmosenv.2016.11.066
5. Shin Araki,"Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan", Osaka University, Science of the Total Environment, https://doi.org/10.1016/j.scitotenv.2018.03.324

Acknowledgements

- **Mentors:**
  Dr. Sumona Mondal –Department of Mathematics,
  Dr. Suresh Dhaniyala –Department of Mechanical Engineering,
  Dr. Shantanu Sur –Department of Biology
  Dr. Supraja Gurajala–Department of Computer Science, SUNY Potsdam
- **Colleagues:**
  Vijay Kumar

- **Personal Contact:**
  Email: senarasd@clarkson.edu

AIR LAB
AEROSOL INSTRUMANTATION RESEARCH

CLARKSON UNIVERSITY
18 • TECHNOLOGIA • 96
™