Imperial College London
Department of Earth Science and Engineering
MSc Environmental Data Science and Machine Learning

Independent Research Project
Project Plan

Predicting spatiotemporal trends in intraurban CO_2 using land use regression and machine learning algorithms

by
Anna Carina Smith

Email: anna.smith23@imperial.ac.uk
GitHub username: edsml-acs223

under the supervision of
Fangxin Fang
Linfeng Li
Jiansheng Xiang

June 14, 2024

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 urban CO_2 emissions, and there is a need to study spatiotemporal trends in CO_2 to better inform sustainable spatial planning of cities. This study investigates the viability of using land use regression (LUR) to predict intraurban CO_2 in the San Francisco Bay Area using data from the BEACO$_2$N monitoring network. LUR is commonly used to predict urban air pollution but there has been no evidence of using LUR to predict intraurban CO_2. The integration of machine learning (ML) algorithms, such as XGBoost, has improved LUR predictive accuracy for pollutants such as carbon monoxide. Therefore, the integration of LUR with ML methods will also be explored in this study. This project represents a novel contribution to LUR research and intraurban CO_2 modelling aiming to clarify the relationship between land use and intraurban CO_2.

INTRODUCTION

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 of air pollution (Li et al., 2021). The most common use of LUR is to produce exposure assessments for epidemiological studies to 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) analysed 12 studies and found that independent variables used in LUR can be broadly categorized into four categories: (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 of predictions, and Ryan and LeMasters (2007) found that the analysed LUR models accounted for between 54% and 81% of the variability in air pollutant concentrations. Furthermore, the integration of machine learning (ML) algorithms, particularly 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 and 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 due to spatially skewed 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).

Intraurban Carbon Dioxide

Urbanization is a key trend of 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). 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. 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. Better understanding the trends and variability in intraurban CO_2 is important to inform stakeholders and policymakers in the development of sustainable spatial planning strategies (Mitchell et al., 2018; Wang, 2018).

Like the air pollutants typically modelled using LUR, CO_2 demonstrates intraurban spatiotemporal variability. The heterogeneous nature of urbanization and land use activities results in a heterogeneous landscape of CO_2 levels within cities (Wang, 2018). Wang (2018) studied the relationship between zoning plans and emissions resulting from major economic sectors in the Taipei metropolitan area; the study found that total sector emissions increased alongside growth in total zoned area, and that individual sector activities, and associated sector emissions, had unique spatial distributions. Another study identified increased emissions resulting from suburban development and population growth in rural areas (Mitchell et al., 2018). These studies confirm that urbanization and land use change contribute to the spatiotemporal variability of CO_2 within cities.

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 activity or 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 efforts 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 CO_2 data (Duren and Miller, 2012). Gaps persist in the understanding of urban carbon dynamics, and there is a need for more long-term, spatially distributed urban CO_2 monitoring networks (Mitchell et al., 2018).

Research Gap

A review of existing literature has produced no evidence of using LUR to model intraurban CO_2, 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 conditions. 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 project aims to address this research gap by developing an LUR model to predict intraurban CO_2.

Objectives

The objective of this study is to understand how land use affects intraurban CO_2. Specifically, the viability and model performance of using LUR to predict intraurban CO_2 will be explored. The study will also investigate the potential of integrating ML algorithms like XGBoost with LUR. The central research question is: How does land use contribute to the spatiotemporal trends of intraurban CO_2 in the San Francisco Bay Area? The sub-questions are: (1) Can LUR effectively predict intraurban CO_2? (2) Can ML algorithms improve LUR model performance? (3) What are the key predictors of intraurban CO_2 emissions in the San Francisco Bay Area?

Methodology

Predictor variable selection and LUR modelling will be informed by previous LUR studies, including Larkin et al. (2023), Lee et al. (2017) and Li et al. (2021). The integration of ML algorithms will be informed by Wong et al. (2021), which explored deep neural networks (DNN), random forest (RF), and extreme gradient boosting (XGBoost). Since XGBoost performed best, this will be the algorithm of primary interest. Depending on model performance, additional algorithms may be considered.

Expected Outcomes

LUR models in the literature achieved R^2 values roughly between 0.5 and 0.7; this is the expected range of model performance for the LUR model in this study. XGBoost helped increase R^2 from 0.69 to 0.85 in Wong et al. (2021), therefore, a boost in performance is expected upon integrating ML. In line with previous findings, road traffic is expected to be a major predictor of CO_2. This study may be limited by the amount of data and by the aggregation data for different variables from various sources, across slightly different spatial and temporal scales.

PROGRESS TO DATA

A literature review has been conducted to understand the current state of research on LUR, urbanization, land use, spatiotemporal trends in intraurban CO_2, and existing intraurban CO_2 monitoring networks. This review helped identify research gaps and research questions to guide this study. It also helped identify possible case sites and data sources. In particular, the BErkeley Atmospheric CO2 Observation Network (BEACO2N) monitoring network has been identified as a viable source of intraurban CO_2 data featuring over 30 sensors (Shusterman et al., 2016). Possible sources of data on land use and other potential predictor variables have also been identified, mostly using the California Open Data Portal.

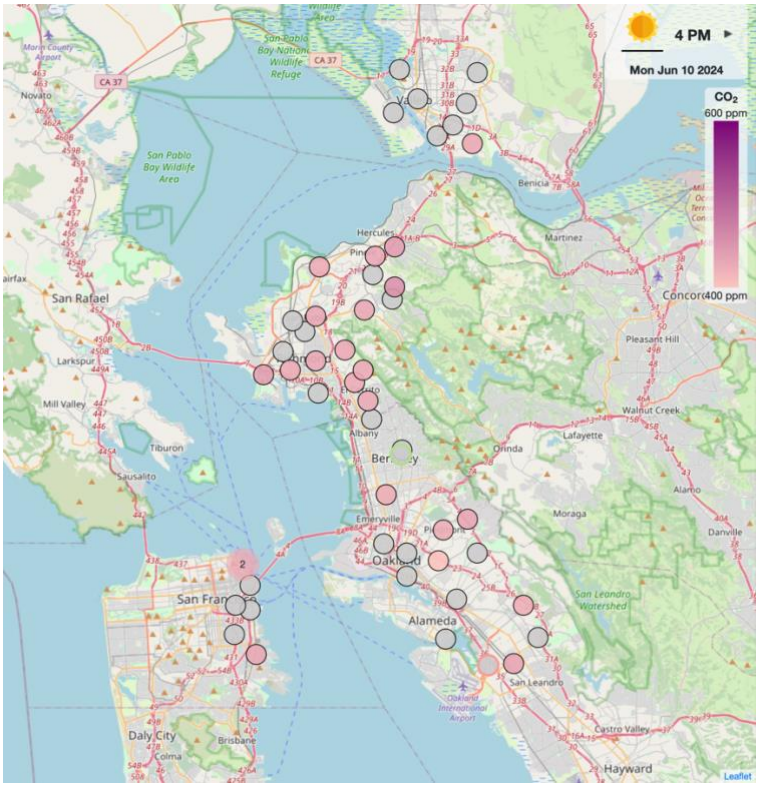

***Figure 1**. Overview of BEACO2N sensors around the San Francisco Bay Area. http://beacon.berkeley.edu/about/*

FUTURE PLAN

May

- Week 1: Literature Review

June

- Week 2: Research Gap Identification and Proposal Writing
- Week 3: Proposal Writing and Data Collection
- ***Submit Project Plan by 12:00pm BST on Friday, June 14th, 2024***
- Week 4: Data Collection and data pre-processing
- Week 5: Data pre-processing and LUR model development

July

- Week 6: LUR model building
- Week 7: LUR model tuning
- Week 8: ML algorithm implementation
- Week 9: ML algorithm implementation
- ***Complete modelling by Friday, July 26th, 2024***
- Week 10: Results, interpretation and analysis

August

- Week 11: Analysis and report writing
- Week 12: Report writing
- ***Complete first draft of Final Report by Friday, August 16th, 2024***
- Week 13: Report writing
- Week 14: Report cleaning and presentation planning
- ***Submit Final Report by 12:00pm BST on Friday, August 30th, 2024***

Bibliography

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 CO₂ 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

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

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 CO₂ Observation Network: initial evaluation. Atmos. Chem. Phys. 16, 13449–13463. https://doi.org/10.5194/acp-16-13449-2016

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