Analysis, Prediction and Mitigation of Exposure to Vehicular Air Pollution Based on Multi-Source Urban Data

Gurban Aliyev[1,2,*]

[1]The University of Pisa, Pisa, Italy
[2]KDD Lab, ISTI-CNR, Italy

Abstract

An increasing amount of vehicular emissions in urban air pollution create a health risk for urban residents. Meanwhile, calculation and analysis of vehicular pollution using GPS trajectories and microscopic models is getting more popular as this method proves to be more useful and reliable in comparison to other methods. However, GPS-trajectory-based estimations suffer from the lack of GPS data and absence of validation/calibration of estimated emission amounts. Another problem is in the assessment of pollution levels using GPS trajectories as previous studies only consider changes in total vehicular emissions and ignore air quality guideline levels. In this paper, the methodology and preliminary results of experiments conducted for imputation of missing emission data are reported. An existing graph convolutional network model which is designed to predict traffic flows is adopted to estimate vehicular emissions in Pisa. This approach is based on the assumption that the same model can predict traffic emissions as a traffic flow and resulting emission are correlated. In the end of the paper, there is a discussion of future research directions planned to be taken during my PhD period to address issues in the estimation, analysis and mitigation of exposure to vehicular emissions in cities.

Keywords

road networks, vehicular emissions, missing data imputation, graph convolutional network, graph embedding

1. Introduction

Ambient air pollution is one of the main barriers to the sustainable development of urban areas, which are expanding fast recently [1]. Air pollution is a more serious concern in cities than in rural areas as cities are densely populated. Also, primary sources of anthropogenic emissions, such as energy production and transportation, are concentrated around urban clusters. As a result, the concentration of air pollutants causes low air quality in cities, with more and more people exposed to air pollution every year. This problem can have severe effects on public health and the economy, which is why the United Nations call to reduce the adverse per capita environmental impact of cities [2].

We focus on estimation and mitigation of vehicular emissions as it is harder to estimate and control emissions of tens of thousands cars traveling. In contrast, it is relatively easy to control

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy
*Corresponding author.
g.aliyev@studenti.unipi.it (G. Aliyev)

industrial air pollution caused by fewer stationary factories. While the number of cars in cities is increasing, the total emission amount of some vehicular pollutants continues to rise despite emission standards set [3]. In order to quantify vehicles' emissions, studies using GPS traces offer the best trade-off between the highly-detailed human mobility [4, 5] and representative vehicle fleet [6, 7]. A number of recent studies [6, 8, 9] assess spatial and temporal distributions of vehicular emissions using estimations from vehicular trajectories. Some of these studies help to quantify the impact of existing green mobility policies [8], next-generation routing principles for vehicles [9] or electrification of gross polluters [6] in terms of the decrease in emissions.

However, existing literature has some gaps to be addressed: firstly, most of the currently available trajectory datasets cover only a portion of all vehicles and roads in the road network. So, to have a more complete view of spatial and temporal emission patterns, network management requires an inventory of vehicular emissions that represent the whole population. Another issue is that previous studies have not validated or calibrated emission estimations using ground truth data. Emissions are estimated using microscopic emission functions which take some input parameters inferred from GPS points (such as the speed and acceleration), and some vehicle-related information (engine size, age, etc.). There is also a drawback in the assessment of emissions' mitigation policies as previous studies only consider changes in total emissions and ignore air quality guideline levels [10]. For example, the air quality in some parts of a city can get worse while total vehicular emissions decrease.

In this paper, I list existing models in missing road network data imputation to be adopted, with preliminary results of experiments (Section 2). In the end (Section 3), I present planned activities on future directions of my research.

2. Mising Data Imputation

Because of the limited amount of data collected, the problem of sparse emission data on spatial distribution of urban roads is common. Although there is no study which tries to impute missing emission values on the spatial distribution, there are several new researches [11, 12, 13, 14] which use state-of-the-art approaches to estimate various road features. Some of these models are claimed to be task-agnostic [12, 13], while others are task-specific [11, 14]. However, task specific models are used to estimate missing traffic flows. We decided to consider [11] and [14] as well, taking an assumption that traffic flows and vehicular emissions should have a correlation.

Some of the studies mentioned above [11, 12] recommend applying graph convolutional network (GCN), which is a special type of neural networks adopted from computer vision, to impute road network data. GCN is useful in missing data imputation (MDI) on road networks as it can work with graph-structured data. In other words, using geographic location and other features of graph entities (roads or road intersections), target feature values can be estimated based on similarities between embedded vector representations of those entities. Among existing GCN models applied on missing road network data imputation, SI-GCN [11] appears to be task-specific (estimating traffic flows only), while RFN is task-agnostic and has been used to in driving speed estimation and speed limit classification [12]. Also, SI-GCN estimates traffic flows not on roads, but between geographic units of a city, while RFN has a more microscopic

approach and estimates target features on road segments.

Another novel approach in missing road network data imputation is the application of graph embedding separately to obtain vector representation of the road network entities and using that data to estimate the target variable [13, 14]. Graph embedding is adopted from word embedding [15] and aims to estimate target feature values based on similarities between vector representations of road segments or geographic units. Spatial Structure-Aware Road Network Embedding Model (SARN) is the only task-agnostic and microscopic model predicting feature values on the road-level. The model has been used to predict several features, like road property, trajectory similarity and shortest-path distance [13]. The other model, Geocontextual Multitask Embedding Learner (GMEL), is task-specific and estimates commuting flows between geographic units [14], which similar to what SI-GCN does.

Apart from comparing different models of MDI, we succeeded to run experiments with SI-GCN during the first semester of the PhD research. In Subsection 2.1, we describe SI-GCN model and report preliminary results obtained from experiments with SI-GCN.

2.1. Experiments with SI-GCN and Preliminary Results

SI-GCN model is based on the given idea: Given two flows f_{ij} and f_{mn}, if the origins v_i and v_m, as well as destinations v_j and v_n are similar, then these two flows should have approximately equivalent intensities. The model considers two similarities: closeness of attributes between geographical units (the first-order similarity) and neighborhood structural proximity between geographical units in a spatial interaction network (the second-order similarity) [16].

The architecture of the model consists of three main modules, which are the spatial representation layer, the encoder and the decoder. The dataset accessed is T-Drive [17], which has taxi trajectories in Beijing during 5 days between 13-17 May in 2013. In the first part, the model captures the spatial representation of taxi flows. Specifically, this layer constructs the local graph of geographic units with taxi flows connecting them, conducts negative sampling and organizes features of each geographic unit - grid. Grids are created by a squared spatial tessellation of 30x30, while taxi flows represent the number of taxi trips between two grids during five days. Next, the encoder generates vector representation for each geographical unit with graph convolution. Finally, the decoder generates missing flows from the vector representation.

The model has been experimented by its designers on the dataset of T-Drive and evaluated using root mean square error (RMSE), mean absolute percentage error (MAPE) and common part of commuters (CPC). SI-GCN outperformed three baseline mobility models with MAPE value of 24.3%. We adapted the model to the emission dataset of Pisa (available in our laboratory at ISTI-CNR), where aggregated emission values are mapped to road segments of Pisa. An emission amount mapped to each road segment represents aggregated CO_2 amount emitted by cars on a given road in 2017. These values are calculated using a microscopic model [1] from vehicular trajectories dataset provided in our laboratory.

The adaptation of the model is in the following way: instead of dividing the city into geographic units and representing them as nodes, we consider road intersections as nodes in a graph representation of the network. Consequently, links between nodes represent road segments, where the task of the model is to estimate CO_2 emissions (instead of traffic flow) on road

segments with missing data. Also, we constructed 26 various features for road intersections in contrast to 3 features (centroid coordinates, the number of pick-ups and the number of drop-offs of grids) used in the original model. A larger number of features is supposed to improve the model performance as suggested by the authors of SI-GCN [16].

Preliminary results after repeating the experiment for 3 times show 80-85% of MAPE. This result is still lower than 100%, but significantly higher than MAPE of 24.3%. The main problem is that the model underestimates extremely high emission values, sometimes by 6 times. In the next section, we try to discuss possible reasons of SI-GCN's lower performance rate and reconsider our methodology for MDI.

3. Discussion and Future Directions

It is worth to discuss preliminary results to define the shape our research direction in MDI (Subsection 3.1). Also, we share some ideas to be applied during the next steps of the research.

3.1. Data Imputation & Enrichment

Lower performance of SI-GCN during our experiments can be attributed to several reasons. First, many of neighboring road intersections are close to each other and have stronger similarity in comparison to spatial grids used in the original model (which are easier to distinguish). Having many similar nodes (road intersections), SI-GCN might not be able to estimate extremely low or high emission values. In this case, more features of road intersections might needed to let the model find such a feature which can capture differences between different pairs of nodes. Another implication can be that node-related information on a primal graph (where roads are edges and intersections are nodes) simply cannot help to estimate emissions on edges. In this case, we can use dual graph representation, where nodes represent road segments, while links between road segments would represent intersections of roads. RFN can be a good alternative to be experimented as the author considers sparsity of a road network, road-related features, and shows that road networks have volatile homophily [12]. Volatile homophily means that there are neighborhoods where roads have similar values in the traffic flow or driving speed. However, the value can significantly change at intersections of road segments having different types (e.g., between a residential street and a motorway).

RFN consists of K relational fusion layers ($K \geq 1$) and takes the node, edge, and between-edge attributes as inputs. These inputs are propagated through each layer, where each layer has a node-relational and edge-relational fusion. It helps to learn representations from the node-relational and edge-relational views. Node relational fusion is performed on a primal graph where road intersections represent nodes, while edge-relational fusion is based on a dual graph where roads represent nodes. Dual graph attributes are constructed by aggregating node and between-edge attributes. The aggregation function also has an attention function to exclude noise-contributing neighbors. While performing aggregation during the fusion, a relational fusion operator allows RFN to rely on homophily only conditionally.

An alternative method to capture graph structure of road networks and use edge-relative information is edge embedding learning (SARN [13]). The advantage of embedded learning is that vector representation of the network data can be used to predict the target feature

with a simple prediction model (the model depends on the feature to be predicted). SARN is task-agnostic as its embedding is based on self-supervised graph contrastive learning (GCL). The basic idea of GCL is based on generation of a pair of graph variants (i.e., graph views) by augmenting the graph G. This includes masking random edges or vertex attributes. Next, vertices (road segments) are mapped to embeddings using a graph encoder F. In this case, graph views should have similar embeddings for the same vertex $s_i \in G$ and dissimilar embeddings for different vertices.

The other model, GMEL, is specialized in prediction of commuting flows and learns embeddings of geographic units, like SI-GCN does. However, geographic units in graph representation are census tracts instead of square grids. GMEL embeds the information using Graph Attention Network (GAT), which is based on the idea that nearby units are more related than distant units [18]. In the end, embeddings are used to train a gradient boosting machine (GBM) and predict commuting flows. We would like to apply GMEL to use results as one of the baselines to RFN and SARN.

3.2. Data Validation/Calibration

The next objective during the research is to find out how to validate emission data estimated from trajectory data with the ground truth data. Several studies [16, 19, 20] try to assess estimated emissions by comparing on-road estimations to on-road measurements only. Our aim is to go further and validate emissions estimations using measurements from air pollution monitoring stations or satellites. The reason for such a choice is that station and satellite measurements are up-to-date and available for most cities. In a case of significant differences during the validation, some calibration can be applied to improve the dataset.

However, there are a series of obstacles while comparing two different types of data. For example, satellites offer emission maps, while vehicular emissions are represented on roads. Also, emission maps represent pollution amounts from different sources (transportation, manufacturing, etc.). Moreover, the share of vehicular emissions received by roads can increase or decrease depending on different factors, such as weather conditions (temperature, precipitation, etc.). Yet, vehicles are the only source of some pollutants (such as NO_x) in non-industrial areas of cities. Based on such pollutants, we can link road-level emissions to measured emissions in non-industrial zones using machine learning or deep learning methods, which can consider weather and other background parameters, too. In addition, the validation process can be facilitated by comparing relative changes, instead of focusing on absolute values [20].

3.3. Application of the Emission Data on Urban Ecology

Another objective during the research is to see whether it is possible to decrease health costs of vehicular emissions by applying sustainable urban planning and mobility policies. It is planned to test the policy impact of changing urban configurations, like pedestrianizing streets, adding bike lanes and relocation of traffic-attracting sites. During this process, a tool developed by Yeghikyan et al. [21] can be used to assess the impact of green urban policies on the traffic flow. This tool is based on spatial interaction models and neural networks and predicts how traffic toward some point of interest may change after some urban project development in a

given location. Resulting traffic changes should be transformed into changes in air quality using our "emission health cost model". In order to measure the exposure to air pollution, it is needed to integrate population density data [22] which is publicly available. Involvement of epidemiologists would also be helpful to define threshold exposure levels, which vary depending on the pollutant and environment. For example, threshold levels can be defined in a maximum or average amount for a period from 8 hours to a year [10].

The expected impact of this project is the contribution to the development of sustainable urban transportation and improvement of air quality in cities. This will be achieved through accurate emission estimations of the vehicle fleet (using massive amounts of mobility data and novel machine learning approaches) and generation of sustainable mobility policies.

References

[1] M. Nyhan, S. Sobolevsky, C. Kang, P. Robinson, A. Corti, M. Szell, D. Streets, Z. Lu, R. Britter, S. R. Barrett, C. Ratti, Predicting Vehicular Emissions in High Spatial Resolution Using Pervasively Measured Transportation Data and Microscopic Emissions Model, Atmospheric Environment 140 (2016) 352–363. URL: https://www.sciencedirect.com/science/article/pii/S1352231016304502. doi:https://doi.org/10.1016/j.atmosenv.2016.06.018.

[2] Transforming Our World: The 2030 Agenda for Sustainable Development, 2018. URL: http://connect.springerpub.com/lookup/doi/10.1891/9780826190123.ap02. doi:10.1891/9780826190123.ap02.

[3] Y. Huang, B. Organ, J. L. Zhou, N. C. Surawski, G. Hong, E. F. Chan, Y. S. Yam, Remote Sensing of On-Road Vehicle Emissions: Mechanism, Applications and a Aase Study from Hong Kong, 2018. URL: https://www.sciencedirect.com/science/article/pii/S1352231018301870?via%3Dihub. doi:10.1016/j.atmosenv.2018.03.035.

[4] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, M. Tomasini, Human Mobility: Models and Applications, 2018. doi:10.1016/j.physrep.2018.01.001.

[5] M. Luca, G. Barlacchi, B. Lepri, L. Pappalardo, A Survey on Deep Learning for Human Mobility, ACM Computing Surveys 55 (2021). doi:10.1145/3485125.

[6] M. Böhm, M. Nanni, L. Pappalardo, Gross Polluters and Vehicle Emissions Reduction, Nature Sustainability (2022). URL: https://www.nature.com/articles/s41893-022-00903-x. doi:10.1038/s41893-022-00903-x.

[7] L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, F. Giannotti, Understanding The Patterns of Car Travel, European Physical Journal: Special Topics 215 (2013). doi:10.1140/epjst/e2013-01715-5.

[8] M. N. Rahman, A. O. Idris, Tribute: Trip-Based Urban Transportation Emissions Model for Municipalities, International Journal of Sustainable Transportation 11 (2017) 540–552. URL: https://www.tandfonline.com/doi/full/10.1080/15568318.2016.1278061. doi:10.1080/15568318.2016.1278061.

[9] G. Cornacchia, M. Böhm, G. Mauro, M. Nanni, D. Pedreschi, L. Pappalardo, How Routing Strategies Impact Urban Emissions, in: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, 2022, pp. 1–4.

[10] WHO, Who Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide, 2021.

[11] X. Yao, Y. Gao, D. Zhu, E. Manley, J. Wang, Y. Liu, Spatial Origin-Destination Flow Imputation Using Graph Convolutional Networks, IEEE Transactions on Intelligent Transportation Systems, vol. 22 (2021) pp. 7474–7484. doi:10.1109/TITS.2020.3003310.

[12] T. S. Jepsen, C. S. Jensen, T. D. Nielsen, Relational Fusion Networks: Graph Convolutional Networks for Road Networks, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 418–429. doi:10.1109/TITS.2020.3011799.

[13] Y. Chang, E. Tanin, X. Cao, J. Qi, Spatial Structure-Aware Road Network Embedding via Graph Contrastive Learning, in: J. Stoyanovich, J. Teubner, N. Mamoulis, E. Pitoura, J. Mühlig, K. Hose, S. S. Bhowmick, M. Lissandrini (Eds.), Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023, OpenProceedings.org, 2023, pp. 144–156. URL: https://doi.org/10.48786/edbt.2023.12. doi:10.48786/edbt.2023.12.

[14] Z. Liu, F. Miranda, W. Xiong, J. Yang, Q. Wang, C. T. Silva, Learning Geo-Contextual Embeddings for Commuting Flow Prediction, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.

[15] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013. URL: http://arxiv.org/abs/1301.3781.

[16] M. Kousoulidou, G. Fontaras, L. Ntziachristos, P. Bonnel, Z. Samaras, P. Dilara, Use of Portable Emissions Measurement System (PEMS) for The Development and Validation of Passenger Car Emission Factors, Atmospheric Environment vol. 64 (2013). doi:10.1016/j.atmosenv.2012.09.062.

[17] Y. Zheng, T-drive Trajectory Data Sample, 2011. URL: https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/, t-Drive sample dataset.

[18] W. R. Tobler, A Computer Movie Simulating Urban Growth in the Detroit Region, Economic Geography 46 (1970) 234–240. URL: https://www.tandfonline.com/doi/abs/10.2307/143141. doi:10.2307/143141.

[19] M. Ekström, A. Sjödin, K. Andreasson, Evaluation of The COPERT III Emission Model with On-Road Optical Remote Sensing Measurements, Atmospheric Environment, vol. 38 (2004) pp. 6631–6641. doi:10.1016/j.atmosenv.2004.07.019.

[20] Y. Wu, G. Song, L. Yu, Sensitive Analysis of Emission Rates in MOVES for Developing Site-Specific Emission Database, Transportation Research Part D: Transport and Environment 32 (2014). doi:10.1016/j.trd.2014.07.009.

[21] G. Yeghikyan, F. L. Opolka, M. Nanni, B. Lepri, P. Lio, Learning Mobility Flows from Urban Features with Spatial Interaction Models and Neural Networks, 2020. doi:10.1109/SMARTCOMP50058.2020.00028.

[22] M. Melchiorri, A. J. Florczyk, S. Freire, M. Schiavina, M. Pesaresi, T. Kemper, Unveiling 25 Years of Planetary Urbanization with Remote Sensing: Perspectives from The Global Human Settlement Layer, Remote Sensing 10 (2018). doi:10.3390/rs10050768.