A Data-Integration Analysis on Road Emissions and Traffic Patterns

Ao Qu, Yu Wang, Yue Hu, Yanbing Wang, and Hiba Baroud(✉)

Vanderbilt University, Nashville, TN, USA
{ao.qu,hiba.baroud}@vanderbilt.edu

Abstract. Understanding human activities and urban mobility patterns is key to solving many urban issues such as congestion and emissions. With the abundant data sets available at different levels of fidelity, one of the main challenges is the sparsity and heterogeneity of data sources. The integration of such data sources is essential to better inform system design and community-level strategies. In this paper, we incorporate a variety of data sources including land use, vehicle emissions and building footprint to comprehensively visualize and analyze traffic patterns in the Chicago Loop area. We first implement and compare three different nearest-neighbor-search algorithms to determine building occupancy assignment, and then perform a spatial-temporal correlation analysis of vehicle emissions focusing on factors such as land use, public transit and demographic. Lastly, we discuss the traffic characteristics from data analysis, such as traffic congestion formation and rush hours etc.

Keywords: Vehicle emissions · Traffic patterns · Nearest neighbor search

1 Introduction

Motivation and Contribution. According to the inventory of U.S. Greenhouse Gas (GHG) Emissions and Sinks 1990–2018, transportation accounted for the largest portion (28%) of total U.S. GHG emissions in 2018 [1]. Amongst all the sources, passenger-cars contribute to nearly 60%. The majority of the use cases are for daily commute. Therefore, it is central to understand the commute patterns of city dwellers, and the integral relationship amongst other factors such as land use, building occupancy, road network and emissions, to consequently inform energy-efficient and sustainable community strategies specifically to each city.

The fundamental problem we address in this paper is the lack of data integration procedures for city-scale traffic impact analysis. Because of the lack of direct data sources for traffic impact analysis such as daily commute schedule

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-63393-6_34) contains supplementary material, which is available to authorized users.

J. Nichols et al. (Eds.): SMC 2020, CCIS 1315, pp. 503–517, 2020.
https://doi.org/10.1007/978-3-030-63393-6_34

or block-level emissions, and the inconsistency of the fidelity and scale of various data sources, we design a workflow (Fig. 1) by (1) developing algorithms to realistically assign vehicles' last-seen locations to nearby buildings, (2) preparing grid-based data that incorporates multiple data sources for regression analysis, and (3) conducting feature selection and impact analysis for vehicle emissions. The question of finding the determining factors in the urban areas that contribute to traffic congestion and vehicle emissions is city-dependent, and doing so could potentially help city planners and policy makers target specific areas to estimate and reduce traffic-related emissions. In this paper, we focus on the Chicago Loop area, a major business district in Chicago, IL.

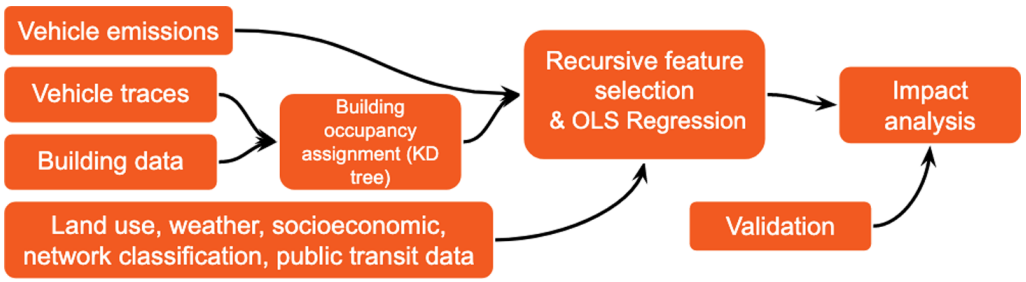


Fig. 1. Workflow for traffic impact analysis.

The main contributions are the following: (1) we estimate a realistic building occupancy schedule by efficiently assigning vehicle occupants to nearby buildings using three nearest neighbor algorithms, with our customized metric, *nearest end point distance*. We demonstrate numerically the superiority of running time and accuracy of our approach by comparing it with others; (2) we propose a method to analyze the impact of city land use, populations, and public transit on vehicle emissions. We integrate various data sources that contribute to vehicle emissions, and perform an area-wide correlation analysis on the selected features using a linear regression model and XGBoost for validation. Specifically, we investigate the impact of land use, population, building occupancy schedule and weather on local vehicle emissions; (3) we lastly characterize traffic patterns by locating the traffic hot spots, popular roads, and rush hours, among others.

Data Sources. Most of the data sources used in this project are provided by Oak Ridge National Laboratory. The data sources are listed following.

1. Commute data: (1) simulated morning commute vehicle traces data at 30 s intervals for one day. The data include road segment (link) ID, driver ID and vehicle speed at each time step. The simulation software is TRansportation ANalysis SIMulation System (TRANSIMS) [2,3]. (2) schedule data for morning commute from National Household Travel Survey (NHTS) [4] and (3) vehicle type distribution data.

2. Emission data: (1) road-level traffic volumes (aggregated from TRANSIMS outputs). (2) Road-level emissions generated using MOVES [5], an emissions simulator.
3. Road network: this data includes link IDs and road type, GeoJSON file of the road network used for the TRANSIMS and MOVES runs, and definition of different link types.
4. Building data: (1) building footprints from Microsoft [6]. (2) Land use data from Chicago Metropolitan Agency for Planning (CMAP) [7], including GeoJSON file containing polygon data with land use attributes.
5. Socioeconomic data: (1) Population from CMAP/Census (2010) [8], (2) community snapshots (2017) [9] and (3) Chicago commute time (2017) [9].

Additional data was collected. The list with corresponding references are provided below. (1) OpenStreetMap: natural cover data [10], (2) DATA.GOV: Chicago bus routes and Chicago rail system ("L") shapefiles [11], (3) Weather Underground: historical weather data [12], (4) Chicago Data Portal: building height data, Chicago population by census block, and census block boundaries [13].

Related Work. Daily commute has a high impact on city traffic and vehicle emissions. In order to analyze the factors that affect vehicle emissions, we need to understand the commute behaviors in terms of when and where people travel to work, based on survey data such as National Household Travel Survey (NHTS), vehicle traces data and building location information. The highly spatio-temporally varying commute patterns have posed many challenges to modeling building occupancy on a high-resolution level. Studies such as [14,15] develop high-resolution building occupancy models using surveyed time-based data, which underpin further analysis such as building energy demand modeling. In [16], a realistic building occupancy assignment is accomplished using a quadtree based approach to allocate agents' first and last seen locations to nearby buildings. Our paper uses a similar approach but compares and analyzes different agent assignment algorithms along with the quadtree.

Integration of other data sources and modeling techniques are also important to understand the relationship amongst human activities, land use and traffic emissions. For example, meteorological data [17] and social media data [18] are adopted to explore the potential influence of human activities on urban traffic congestion and emissions. Integrated models of land use and transportation are also applied to study city dynamics [19–21]. All of the related works present a comprehensive model for evaluating the effect of human activities on cities' microclimates.

The challenge of assigning vehicle occupants to nearby buildings is essentially a nearest neighbor search (NNS) problem [22]. There are numerous algorithms to solve the NNS problem and they are classified into two types: exact methods and approximation methods. In exact methods, the simplest algorithm is the purely brute force one, which is the most accurate but most computationally demanding of all. This running time can be further improved by employing space partitioning

methods, such as KD-tree and Hilbert R-tree [23], which skip computations on some branches and increase efficiency. In approximation methods, the quadtree is widely used due to its superior performance and simple implementation. The details of the quadtree can be seen in [16].

2 Methodology

2.1 Challenge 1: Algorithms to Assign Vehicle Occupants to Buildings

We formulate the building occupancy assignment as a nearest-neighbor-search (NNS) problem. Specifically, we want to assign the last seen locations of vehicles (given by the simulated vehicle trace data) to their nearest buildings (given by the building footprint data) [16]. We apply three search algorithms, brute-force, quadtree, and KD-tree. We compare their performance with respect to efficiency and accuracy. Regarding the building type, we assume that people work in non-residential buildings, and filter out residential buildings based on the land use codes.

We first apply brute-force search algorithm to obtain the baseline running time and accuracy of the building occupancy assignment. Since this algorithm finds the exact solution using a double loop: for each agent find the nearest building, we use the results to benchmark the accuracy of other algorithms. Secondly, we apply the quadtree algorithm used in [16] to assign occupants to buildings. We further introduce the KD-tree algorithm to solve the same problem.

The KD-tree algorithm iteratively bisects the search space and constructs a tree where the leaf nodes correspond to the building locations and the branch nodes correspond to the higher subspaces. If the distance between a vehicle and a subspace is larger than the minimum distance, we can skip this branch of the tree such that the search efficiency can be improved.

For each of the above three algorithms, we assign vehicles based on three distance metrics: *Euclidean distance* (ED), *weighted Euclidean distance* (WD) and our heuristic version *nearest end point distance* (ND). The WD is measured by multiplying the ED with a weight factor proportional to the inverse of building area. The detailed definition has been mentioned in [16]. The ND is the distance from a vehicle to the nearest end point of a building polygon. For vehicle i and building j , ND (d_{ij}) is defined as:

$$d_{ij} = \min_{k \in P_{\text{poly}_j}} \|r_i - r_k\| , \forall i \in V, j \in B \tag{1}$$

where V is the set of vehicle last seen locations and B is the set of building polygons; P_{poly_j} denotes the set of points on the boundary of building polygon j. r_i's are the coordinates of the vehicle location i.

Furthermore, we calculate building capacity by multiplying building size with per capita area and count the number of overload buildings to compare the performance of the three distance metrics.

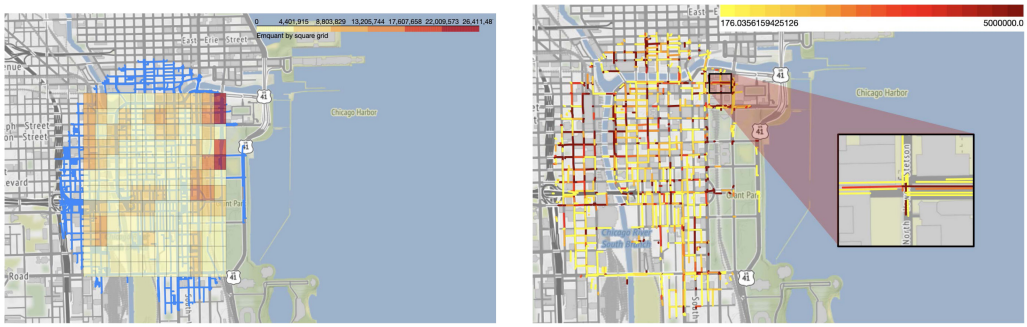
2.2 Challenge 2: Vehicle Emissions and Correlation Analysis

Data Preparation: Generate Grid-Based Data. We enrich the provided data with data from additional sources (listed in Data sources).

One of the main challenges in conducting traffic emissions analysis and exploring the impact of other factors is data reconciliation. Geographical features are often based on different scopes, such as points, lines, and polygons. To address this problem, we first select a target area that fully covers our study region. Then we introduce a grid-based data integration technique to normalize and aggregate various data sets into $N \times N$ grids as shown in Fig. 2a. Specifically, the feature variables are aggregated as follows. The *population* is the total number of residents in the grid. The *inflow population* is the total number of people commute to the grid area each day. The *Public transit (bus & rail)* and *road types* measure the total length of corresponding bus, rail or road line within the grid. *Land use types* and *natural cover types* are the total area of the corresponding type, and the *foot print area* is multiplied by the number of stories if the building type information is available.

If one line or polygon intersects with more than one square grid, then we assume that the corresponding feature is evenly distributed on this line/polygon. For example, the emission data for a certain square grid is calculated as following:

$$\text{total grid emission} = \sum_{\text{all roads}} \text{road emission} \times \frac{\text{length of road within grid}}{\text{total length of road}}$$



(a) A 12×12 grid example and its relationship with roads.

(b) Road emission visualization.

Fig. 2. Spatial variation of aggregated emission. Darker red color indicates more emissions. The partially enlarged view in (b) shows that even within a small space, there is a large variation among road links. Thus, the grid-based method as illustrated in (a) is used to reduce the noise. Note: The dark area on the left of (b) doesn't look dark in (a) because roads in that area are actually more sparse and our grids don't cover some of the roads with heavy emission due to the difficulty of incorporating spreading road network into regularly shaped grid. This issue should be minimized when a broader range of data becomes available.

Another way to combine all the data is a road centered method, which defines fixed areas centered around the centroid of each road, and then measures each quantity within every defined area as the independent variables. However, this method is not suitable for our case since the emission data is generated by simulation and contains inherent noise. Emission per unit length is calculated as $\frac{\text{road emission}}{\text{road length}}$ and we observe a large variation in this measure even among roads that are within the same intersection (Fig. 2b). Therefore by averaging all roads in a specific area, the grid-based method effectively reduces the noise.

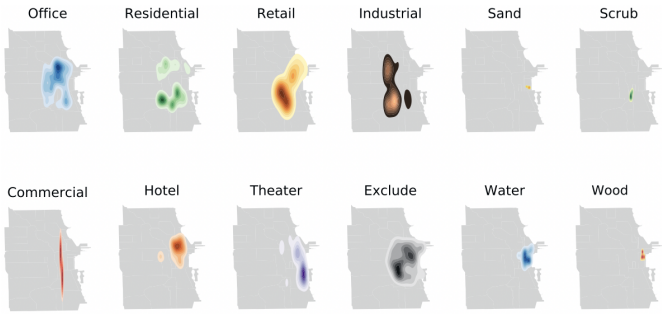


Fig. 3. The spatial distribution of each land use and natural cover type. We can see that different land use type is concentrated in different areas. For example, residential buildings are concentrated in the south while office is more likely to be seen in the north.

Regression Model and Feature Selection. Our primary analysis examines the relationship between vehicle emissions and other factors. We first perform a multivariate regression analysis by partitioning the study area into 12×12 grids, and assessing the contribution of each factor to the road emissions nearby. The first model intends to examine the spatial correlation only, so the time-varying variables are averaged. For example, emissions for a certain grid are calculated as total emissions in a day divided by 24 h.

Figure 3 and 4 show land use and natural cover types distribution using kernel density estimation (KDE) and the spatial correlations among all features within the study area, respectively. From the correlation matrix (Fig. 4), we notice that the correlation coefficients between some features (e.g., population and residential areas) indicate the presence of a strong multicollinearity (Pearson correlation coefficient $\rho \geq 0.7$), which increases the standard errors of the coefficients when doing regression analysis, and in turn may cause some independent variables to be not significant. To address this issue, we employ recursive feature elimination(RFE) to repeatedly remove the least important variables. For spatial correlation analysis of vehicle emissions, we regress the averaged emissions on other variables selected by RFE using an Ordinary Least Squares(OLS) model.

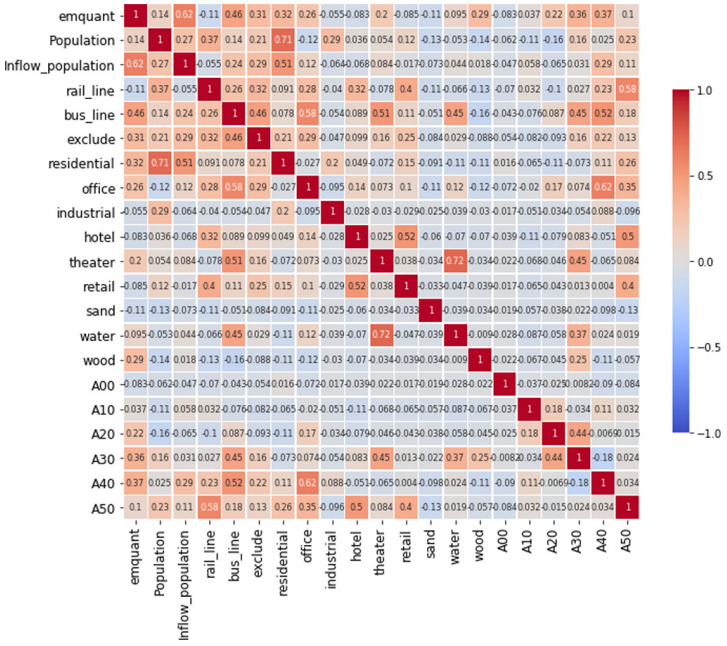


Fig. 4. A matrix showing correlation coefficients between variables

Robustness Testing. Since the number of grid cells may affect the correlation results, we test the robustness of the area division by employing three approaches to validate our result. First, we repeat the same procedure on 8×8, 10×10, and 15×15 grid dividing the same area to check the consistency of the significance of independent variables. Second, we implement the road centered method although some variations in feature importance caused by the inherent data noise are expected. Third, we use tree-based XGBoost regression to calculate the feature importance ranking. The feature importance reported by XGBoost is the average information gain across all decision trees when the feature is used as a splitting node. In each robustness test, we also rank the feature importance of each variable so that we can check whether the variables used in our primary model remain stable.

Temporal Variation. Since the vehicle emissions and some other variables are also time-varying quantities, we intend to investigate the temporal correlation of vehicle emissions as well. To this end, we include features such as weather. Overall, 94.8% of all roads demonstrate an increase in emission from January to July which is clearly an evidence of the presence of seasonal effect. However, we are unable to extract more detailed insights regarding temporal variation. The reason is twofold: first, the current emission data covers only two very short time periods (Jan 9th and July 4th to 10th) and the simulated emission data in July is the same each day. Second, the spatial coverage is too small to include the diversity of weather conditions. We intend to address the temporal correlation in future work when a broader range of data becomes available.

2.3 Challenge 3: Traffic Patterns Characterization

Traffic Hot Spots, Congestion, and Popular Roads. According to INRIX [24], a leading traffic analytics company, traffic hot spots are defined as traffic jams that occur at the same locations along a stretch of road. The measure we use is based on the idea that traffic state can be reflected by the average speed. To identify traffic jams, we apply *speed performance index* (SPI) formerly developed by *Beijing Traffic Management Bureau* (BTMB) to evaluate the traffic condition of each road during each hour [25]. The index, defined as the ratio between the current speed and the maximum possible speed, can be applied here. SPI ranges from 0 to 1 with 1 indicating a very smooth traffic and 0 extremely congested traffic. However, we do not count zero in this study because zero average speed for an hour is more likely a sign of no vehicle passing through. According to BTMB, heavy congestion occurs when $SPI < 0.25$. In our study of hot spots, we first use k-means algorithm to cluster roads into 20 small groups by their spatial locations and calculate the average number of occurrences of heavy congestion for each cluster. We also calculate the ratio between the weekly average speed in a week and maximum possible speed for each road so that we can identify specific hot spots. Popular roads are measured by their traffic volume instead of average speed. We aggregate the traffic volume provided in the simulation data and select the top ranked roads to highlight in the map.

Travel Time. We pre-process the data to eliminate outliers in two steps. First, we select all the commute trips from home to work that are less than 2.5 h (as the rest are obviously outliers, e.g., 10+ hours for a single trip), which cover 99.5% of all the trips. Second, we only keep the trips that start from home between 5:00 and 13:00 since people typically go to work in the mornings. Then to analyse the travel time, we divide the time window between 5:00 and 13:00 into 10-second-intervals. We treat the travel time for each time interval as a random variable, and calculate the mean and the 95% confidence interval based on the travel time of trips occurring in this interval.

Busy Times and Comparison with NHTS. To compare the simulation results and the survey conducted by NHTS, we first average the simulation output from Monday to Friday to compute the average total traffic volume of each hour in a day. The provided NHTS trip distribution is the same for each day so it suffices to make comparison on a one-day distribution. Then, we proceed to obtain the fraction of traffic volume per hour in the busy-time distribution plot (Fig. 9). Note that the busy-time distribution sums up to 1 overtime, and thus can be treated as a probability distribution. Therefore, a commonly used measure, Jensen–Shannon divergence [26], can be used to quantify the resemblance between two probability distributions. Jensen-Shannon divergence is calculated as the entropy of the mixture of two distributions minus the sum of the entropy of each distribution such that a disparity in the two inputs would lead to higher score.

Spatial-Temporal Analysis of Speed. We analyze the spatial temporal variation and summarize our finding in a dynamic visualization. Again, we assume that zero-speed roads imply zero-traffic so those roads are colored green.

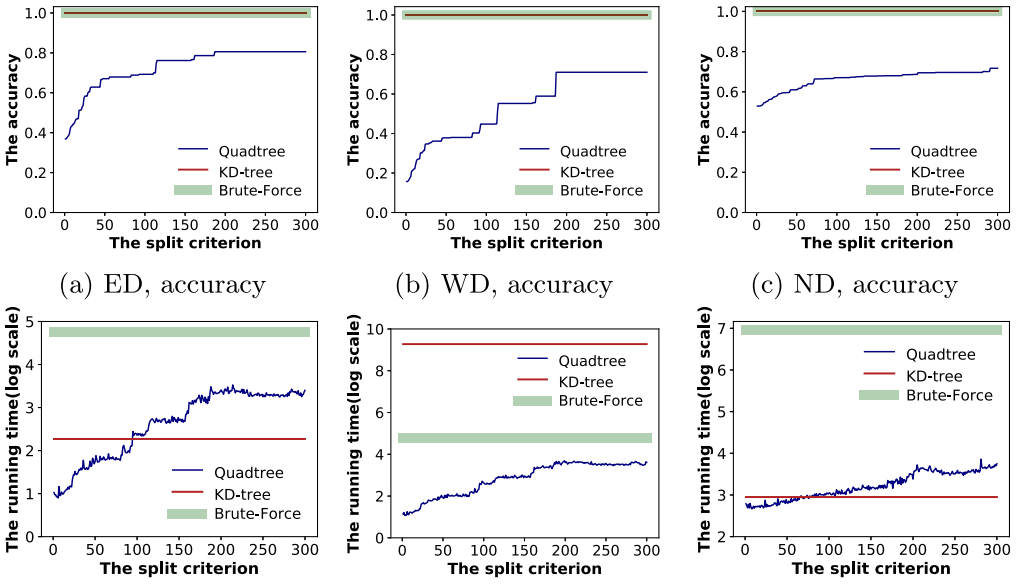
3 Results

3.1 Challenge 1: Performance Comparison of NNS Algorithms

The accuracy of quadtree, KD-tree and brute-force algorithms for building occupancy assignment are shown in the first row of Fig. 5. Both KD-tree and brute-force algorithms achieve 100% accuracy because they compute the exact solution to the NNS problem, no matter what distance metric we choose. However, the accuracy of quadtree only improves when the partition is coarser (i.e., the number of leaf nodes becomes smaller), and no-split quadtree becomes equivalent to the brute-force method). In terms of the distance criteria, ND metric can achieve a higher accuracy than WD or ED, this is because ND metric as in Eq. (1) considers the geometric shape of the buildings and not just the centroids, leading to a better approximation of the actual distance. The running time of each of three algorithms is shown in the second row of Fig. (5). Brute-force algorithm has the longest running time when using ED or ND metric. This is due to the double loop structure in the brute-force algorithm which requires going through all the vehicles' last seen locations and all the building locations to find the nearest building for each occupant. KD-tree has a consistently low running time for ED or SD metric, but fails to outperform brute-force when using WD metric. This is because WD requires reconstructing the search tree when each new vehicle location is added, which significantly slows down the computational time for KD-tree. As for quadtree, higher accuracy can be achieved when using lower-fidelity split, but this also increases the running time. The vehicle assignment and the overload buildings are shown in the supplementary materials. The total number of office buildings is 665 and the number of overload buildings is 27, 10, 10 for ED, WD and SD metrics. Considering both the accuracy and the running time, KD-tree with ND metric consistently outperforms brute-force and quadtree.

3.2 Challenge 2: Area-Wide Correlation Analysis of Vehicle Emissions

Regression Analysis. To reduce standard errors caused by feature multicollinearity, we first perform a recursive feature selection (RFE) on the grid-based data. RFE is a method that keeps removing the weakest feature, which also allows us to evaluate the rankings of features. We find that for 12×12 grid, the adjusted R^2 is the highest(0.71) when top 13 features are used in regression model and most of them are statistically significant (Fig. 6). This high adjusted R^2 indicates that a large portion of variance in emission can be explained by the features chosen by the model.



(a) ED, accuracy (b) WD, accuracy (c) ND, accuracy

(d) ED, running time (e) WD, running time (f) ND, running time

Fig. 5. The performance of algorithms based on different distance metrics

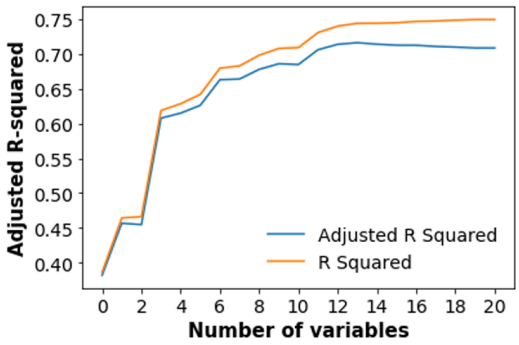


Fig. 6. Adjusted R^2 increases as more features are added to the model

The regression result (Table 1) shows that the main contribution to vehicle emissions comes from inflow population, and some certain types of road including A50 (Vehicular trail, road passable only by four-wheel drive vehicle) and A40 (Local, neighborhood, and rural road, city street), which are positively correlated with vehicle emissions with significance (p-value) $p < 0.001$. Rail line length and vehicle emissions are negatively correlated with p-value $p < 0.001$, which implies the important role of Chicago rail system in alleviating road transportation. One interesting finding is that wood coverage has a strong positive correlation with emissions. One possible interpretation is that wood coverage represents urban parks which are often built near city busy corridors. We also want to emphasize that correlation does not imply causation. Our analysis only explores

Table 1. Vehicle emissions regression analysis and feature rankings. * and *** represent $p < 0.05, p < 0.001$ respectively.

Features	Coefficients	Feature ranking		
		Grid-based OLS	Grid-based XGBoost	Road-centered OLS
Inflow population	62939.816***	1	2	8
Rail line	−52640.761***	4.75	4.25	2
Bus line	40243.822*	4.75	8	7
Office	−35525.439*	10.5	10	9
Water	−31460.252*	12.75	15.75	17
Wood	50688.006***	3.75	7.63	1
A20[1]	27364.166*	10.25	15.13	15
A30[2]	43346.564*	7.5	6	5
A40[3]	70263.346***	6.75	4.75	4
A50[4]	50938.104***	5.5	5	3

[1] Primary road without limited access, U.S. and state highway
[2] Secondary and connecting road, state and county highways
[3] Local, neighborhood, and rural road, city street
[4] Vehicular trail, road passable only by four-wheel drive (4WD) vehicle

the concurrent land use features on vehicle emissions rather than establishing a cause-and-effect relationship.

Validation. To further verify our results of feature selection, we generate three more datasets with different choice of grid size. Then, we calculate the average ranking of each feature based on RFE. We can see from Table 1 that most features presented here, especially those with $p < 0.001$, are consistently top-ranked. Feature importance with XGBoost model and a road-centered model also reports similar ranking, as shown in Table 1. We conclude that our regression model is able to identify the most significant features and the outcome is validated using other methods.

3.3 Challenge 3: Characterize Traffic Patterns

Hot Spot, Congestion and Popular Roads. The visualization in Fig. 7 shows the traffic hot spots and the frequency of heavy congestion for each cluster of roads. The number in the circle indicates on average how many hours the roads around that region are in heavy congestion. The street view images are the six most popular roads ranked by total volume and this result is largely confirmed by the street reviews we find online. An interesting finding is that some very popular roads are not highly congested, which might due to the difference in road design.

Travel Times. Figure (8) visualizes the variation of commute time departing between 5:00 to 13:00 of both NHTS and simulation data. We can clearly see that the mean of the travel time ranges from 0.2 h to roughly 1 h, which is similar

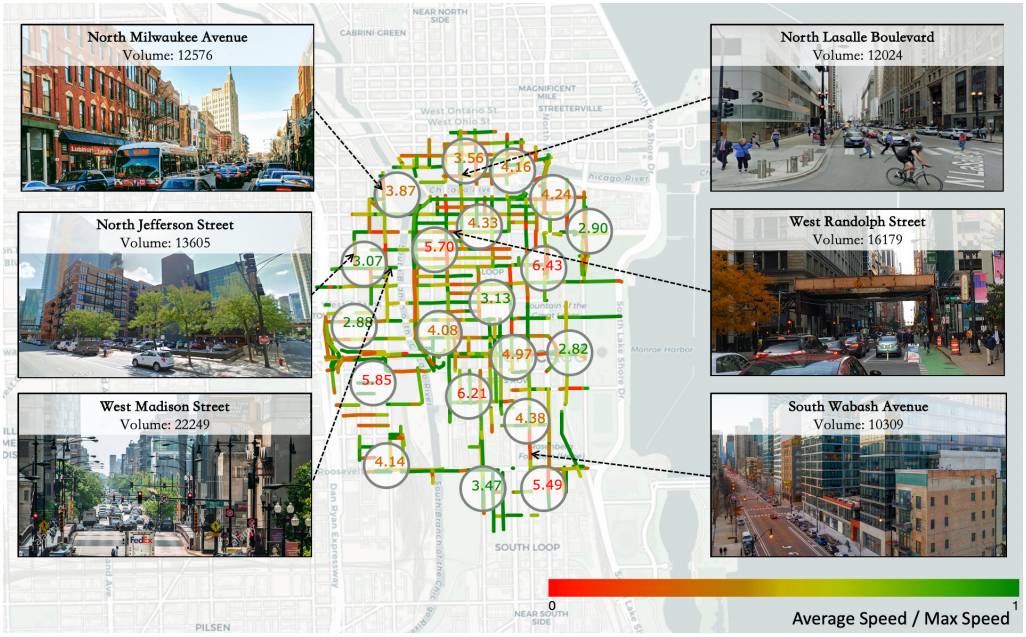


Fig. 7. The variation of travel times throughout the day

to the commute time 58.5 min from a study by Robert Half [27]. In simulation data, the travel time is longest around 7 am and slightly shorter afterwards, possibly because people who have to commute long hours tend to depart early. The travel time between 8 am and 11 am is typically longer than the travel time before 6 am and after 11 am, which might be due to the morning rush hours. In NHTS data, the travel time is highest around 11 am due to the noon peak and decreases afterwards. There is no increasing travel time from 5 am to 7 am in NHTS data as in simulation data, which is caused by the simulation error.

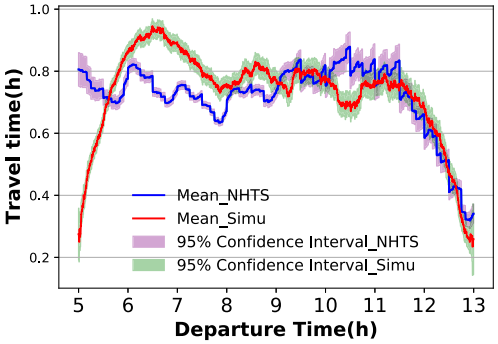


Fig. 8. The variation of travel times throughout the day

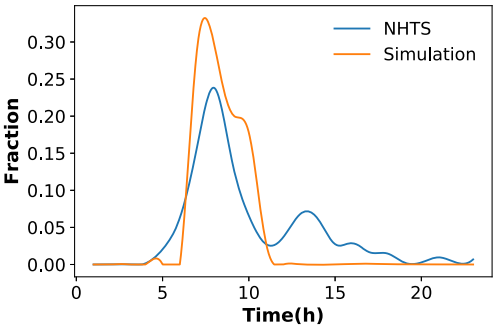


Fig. 9. Busy times according to simulation data and NHTS survey

NHTS Survey Vs Simulation. In general, the simulation has a very similar trend as NHTS with some minor variations. The Jensen-Shannon divergence for these two distributions is 0.38, which indicate a relative similarity between the busy-time distributions. The busy times indicated by both two data sources agree on the morning rush hours (6 am to 10 am). However, there is a second peak in NHTS data after 12 pm which is not found in the simulation. In fact, analyzing on the original simulation data, we find very few trips after 12 pm compared to the NHTS survey. In addition, we observe some unrealistic speeds in the simulation setting. We find that there are about 200 roads with average speeds between 0–1 mph. In fact, excluding the zero speeds, the average speed for all roads is only 8.53 mph. Moreover, we believe that the traffic volume is underestimated. For example, as one of the most popular streets, North Jefferson street is reported to have 8,300 average daily traffic according to Chicago Data Portal. However, in the simulation, the same street has a weekly traffic that is only 13,605.

Spatial-Temporal Variation of Speeds. For this part of the analysis, we plot the speeds on the street map of loop area and generate a GIF to show dynamics (Fig. 10). Basically this visualization aligns with the previous congestion analysis. The speeds are lowest during morning rush hours and roads around the city center tend to have heavy congestion.

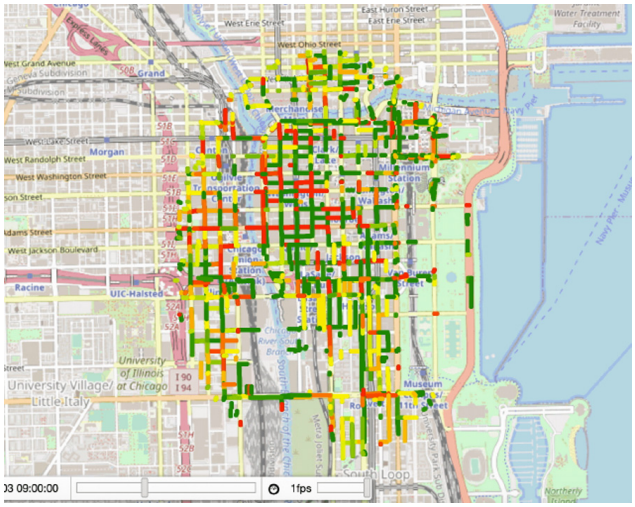


Fig. 10. Spatial-temporal variation of speeds. Green indicates high speed or no traffic while red indicates the opposite. (It may require an Adobe reader to load this GIF. Screenshots of this GIF can also be found here: Appendix) (Colour figure online)

4 Conclusions

In this paper, we provide a framework for data reconciliation and urban traffic patterns characterization. Our solutions to the three challenges contribute to the study of commute patterns and urban transportation systems in the following ways. First, we develop a fast and efficient nearest-neighbor search algorithm, KD-tree with nearest-end-point distance metric, to realistically assign the last seen locations of vehicles to the nearby building. This addresses the lack of direct data sources such as building occupancy schedule, and provides more information on when and where people commute to work. Second, we perform an area-wide analysis of land use, populations and public transit on vehicle emissions. We identify that the inflow population and road types significantly correlates to vehicle emissions. These features are validated using an alternative road-centered data generation approach and a XGBoost model, which produces a similar feature importance ranking. Temporally, a seasonal effect on vehicle emissions is observed but further analysis is hindered due to the lack of high resolution data. Lastly, we explore the traffic simulation data and extract some interesting traffic patterns. We conclude that overall this simulation setup is able to reproduce realistic traffic activities. Most of the travel times are realistic. A good match in busy-time distribution is found between the simulation data and NHTS survey, and major streets are indeed occupied with more vehicles. However, the simulation fails to take into account, for example, the commute back to work after lunchtime that NHTS might indicate.

Some limitations of this study are also worth noting. First, we are not able to draw any conclusion of the impact of vehicle types on emissions, due to the lack of diversity in vehicle classifications. Information about vehicle types would also help us design a more realistic algorithm since the vehicle type might implicate the building type that the vehicle owner is more likely to work at. Second, the vehicle emissions analysis focuses on a specific region (Chicago Loop area), and may not well generalize to other cities. We acknowledge that these limitations exist due to the scope of this study, and instead we focus on providing a framework of reconciling data of different types, and analyzing emissions using other more accessible data, which can be applied in broader scenarios.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Grant No. CMMI-1727785 (Hu), CMMI-1853913 (Wang), and USDOT Dwight D. Eisenhower Fellowship program under Grant No. 693JJ31945012 (Wang).

References

1. EPA. Fast Facts U.S. Transportation Sector Greenhouse Gas Emissions 1990–2018 (2018). https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions
2. TRANSIMS. https://sourceforge.net/projects/transimsstudio/
3. Williams, M.D., Thayer, G., Smith, L.: Technical Report LA-UR-9782 (1997)

4. National household travel survey. https://nhts.ornl.gov/
5. EPA. Motor vehicle emission simulator (MOVES) (2014). https://www.epa.gov/moves/latest-version-motor-vehicle-emission-simulator-moves
6. Microsoft. U.S. building footprints (2018). https://github.com/Microsoft/USBuildingFootprints
7. CMAP. Land use data. https://www.cmap.illinois.gov/data/land-use
8. CMAP. 2010 census data summarized to chicago community areas (2010). https://www.cmap.illinois.gov/data/land-use
9. CMAP. Community data snapshots raw data, July 2020 release (2020). https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data
10. OpenStreetMap. https://www.openstreetmap.org
11. U.S. Government's open data. https://www.data.gov/
12. Weather underground. https://www.wunderground.com/
13. Chicago data portal. https://data.cityofchicago.org/
14. Richardson, I., Thomson, M., Infield, D.: Energy and Buildings **40**(8), 1560 (2008). https://doi.org/10.1016/j.enbuild.2008.02.006. http://www.sciencedirect.com/science/article/pii/S0378778808000467
15. McKenna, E., Krawczynski, M., Thomson, M.: Energy and Buildings **96**, 30 (2015). https://doi.org/10.1016/j.enbuild.2015.03.013. http://www.sciencedirect.com/science/article/pii/S0378778815002054
16. Berres, A., Im, P., Kurte, K., Allen-Dumas, M., Thakur, G., Sanyal, J.: In: IEEE International Conference on Big Data (Big Data), pp. 3887–3895. IEEE (2019)
17. Shiva Nagendra, S., Khare, M.: Transportation Research Part D: Transport and Environment **8**(4), 285 (2003). https://doi.org/10.1016/S1361-9209(03)00006-3. http://www.sciencedirect.com/science/article/pii/S1361920903000063
18. Huang, W., Xu, S., Yan, Y., Zipf, A.: Cities **84**, p. 8 (2019). https://doi.org/10.1016/j.cities.2018.07.001. http://www.sciencedirect.com/science/article/pii/S0264275118302786
19. Bandeira, J.M., Coelho, M.C., Sá, M.E., Tavares, R., Borrego, C.: Science of the Total Environment **409**(6), 1154 (2011). https://doi.org/10.1016/j.scitotenv.2010.12.008. http://www.sciencedirect.com/science/article/pii/S0048969710013112
20. Namdeo, A., Mitchell, G., Dixon, R.: Environmental Modelling & Software **17**(2), 177 (2002). https://doi.org/10.1016/S1364-8152(01)00063-9. http://www.sciencedirect.com/science/article/pii/S1364815201000639
21. Gualtieri, G., Tartaglia, M.: Transportation Research Part D: Transport and Environment **3**(5), 329 (1998). https://doi.org/10.1016/S1361-9209(98)00011-X. http://www.sciencedirect.com/science/article/pii/S136192099800011X
22. Bhatia, N., et al.: arXiv preprint arXiv:1007.0085 (2010)
23. Kamel, I., Faloutsos, C.: In: Proceedings of the Second International Conference on Information and Knowledge Management, pp. 490–499 (1993)
24. INRIX a smart way to drive (2016). URL https://inrix.com/mobile-apps/. INRIX Inc
25. He, F., Yan, X., Liu, Y., Ma, L.: Green intelligent transportation system and safety. Procedia Eng. **100**(137), 11–12 (2016). https://doi.org/10.1016/j.proeng.2016.01.277. http://www.sciencedirect.com/science/article/pii/S1877705816003040
26. Oesterreicher, F., Vajda, I.: Ann. Insts. Stat. Math. **55**, 639 (2003). https://doi.org/10.1007/BF02517812
27. McGhee, J.: Chicago commute is 2rd longest, but less stressful than in many cities (2017)