Analyzing spatiotemporal traffic line source emissions based on massive didi online car-hailing service data

Daniel (Jian) Sun\textsuperscript{a,b,⁎}, Kaisheng Zhang\textsuperscript{c}, Suwan Shen\textsuperscript{d}

\textsuperscript{a} State Key Laboratory of Ocean Engineering, School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai 200240, China
\textsuperscript{b} China Institute of Urban Governance, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai 200240, China
\textsuperscript{c} Shanghai Urban Construction Design & Research Institute (Group) Co., Ltd, No. 3447 Dongfang Road, Pu-Dong New District, Shanghai 200215, China
\textsuperscript{d} Department of Urban and Regional Planning, University of Hawaii, Manoa, No. 2424 Maile Way, Honolulu, HI 96822, United States

ARTICLE INFO

Keywords:
- Online car-hailing service
- NO\textsubscript{x} emission
- Fuzzy C-means clustering
- Spatial analysis
- Built environment

ABSTRACT

Nowadays, the massive car-hailing data has become a popular source for analyzing traffic operation and road congestion status, which unfortunately has seldom been extended to capture detailed on-road traffic emissions. This study aims to investigate the relationship between road traffic emissions and the related built environment factors, as well as land uses. The Computer Program to Calculate Emissions from Road Transport (COPERT) model from European Environment Agency (EEA) was introduced to estimate the 24-h NO\textsubscript{x} emission pattern of road segments with the parameters extracted from Didi massive trajectory data. Then, the temporal Fuzzy C-Means (FCM) Clustering was used to classify road segments based on the 24-h emission rates, while \textit{Geographical Detector} and MORAN’s I were introduced to verify the impact of built environment on line source emissions and the similarity of emissions generated from the nearby road segments. As a result, the spatial autoregressive moving average (SARMA) regression model was incorporated to assess the impact of selected built environment factors on the road segment emission rate based on the probabilistic results from FCM. It was found that short road length, being close to city center, high density of bus stations, more ramps nearby and high proportion of residential or commercial land would substantially increase the emission rate. Finally, the 24-h atmospheric NO\textsubscript{2} concentrations were obtained from the environmental monitor stations, to calculate the time variational trend by comparing with the line source traffic emissions, which to some extent explains the contribution of on-road traffic to the overall atmospheric pollution. Result of this study could guide urban planning, so as to avoid transportation related built environment attributes which may contribute to serious atmospheric environment pollutions.

1. Introduction

With the increasing development of motorization and urbanization, transportation has gradually become one of the main sources of air pollution in many cities throughout the world. For example, according to statistics, on-road transportation accounts for about 47% of CO, 42% of NO\textsubscript{x}, and 18.4% of PM emissions, respectively in Europe (Kousoulidou and Sarmaras, 2008); and 27.5% of total greenhouse gas in U.S., in which 41.9% results from passenger cars (Mausami and Weitz, 2017). It also contributes to about 19% of...
nitrogen oxide emissions in Canada, rising up to 85% in Montreal (IJC, 2012). In China, it is estimated that transportation related activities account for approximately 46% of NOx, 78% of CO, and 83% of HC in metropolises (Wang et al., 2008). Consequently, it’s important to identify the hotspot and the mechanism of on-road emissions to provide guidance for built environment and land use during the planning stage.

Traditional emission measurements are based on laboratory-similar bench test (Silva et al., 2015), remote sensing (Ning and Chan, 2007), and tunnel test (Zhang et al., 2015). However, the emitting characteristics of motor engines extracted from these studies were only applicable to experimental environment. They are expensive and impossible to be used for dynamic monitoring or multi-vehicle test. In recent years, the portable emission measurement system (PEMS) has been widely used in mobile monitoring, which is easy to assemble. Defries et al. (2014) found that PEMS was more compatible to chassis testing than other mobile methods in detecting common pollutants. Gallus et al. (2016) also disclosed that the correlation tests between particle number PEMS and chassis dynamometer tests are in good agreement. However, a set of PEMS could just be used for one-vehicle at one time, taking months or even years to complete the regional data collection.

In such case, simulation models and detailed estimation formulas are developed to simplify the emission estimation. For example, in U.S., Environmental Protection Agency (EPA)’s M0tor Vehicle Emission Simulator (MOVES) is a state-of-the-science emission modeling system that estimates emissions for mobile sources at the national, county, and project level for criteria air pollutants, greenhouse gases, and air toxics. The emission calculation is based on predefined speed bins and vehicle specific power (VSP) distribution, which is suitable from microscopic to macroscopic. Abou-Senna et al. (2013) integrated MOVES and VISSIM microscopic simulation to predict emissions from vehicles on limited-access highways. Liu et al. (2013) input vehicle operational parameters extracted from GPS trajectories into MOVES for estimating vehicle emission factors in Shanghai. International Vehicle Emission (IVE) model is also a VSP-based algorithm, which generated 60 VSP bins to calculate emissions. The model was evaluated by utilizing an available dataset of remote sensing measurements on a large number of vehicles in Hangzhou, China, and found that it performed well on estimating HC and CO emissions (Guo et al., 2007). Nagpure et al. (2011) applied the fundamental equations of IVE model to estimate the emission rates from the light-duty commercial vehicles, and revealed the increases of CO and VOCs and a decrease of NOx emission rate along with altitude. Computer Program to Calculate Emissions from Road Transport (COPERT) is another widely used emission model developed by European Environment Agency (EEA) using speed as the main independent variable (Broderick and O’Donoghue, 2007; Lozhkina and Vladimir, 2015). Since COPERT complies with the European emission standard which is similar to the one adopted in China, the package is capable for emission estimation in Chinese cities (Lang et al., 2016). Bellasio et al. (2007) implemented COPERT to calculate the emission inventory of Sardinia, Italy, and proved its reliability by comparison with field experiments. However, all above referred models require sufficient data when applied to a citywide level, which is unfortunately difficult to be obtained in practical.

On the country, vehicle-based GPS data as a source of driving trajectory record linking to the estimation and influencing mechanism of vehicular emissions, have gained increased attentions (Liu et al., 2009), which covers city wide road network continuously and consequently is believed more accurate. However, studies based on data from the new emerging car-hailing mode are scarce, while most similar studies were focused on taxi vehicle trajectories. Wang et al. (2008) utilized GPS to collect the driving pattern data including 140 taxis, and combined them with IVE to estimate the total emissions of CO, VOC, NOx, and PM from vehicles in Shanghai. However, this study was just of coarsely resolved data collection with limited vehicle number and land scope. Nyhan et al. (2016) obtained GPS trajectory data from over 15,000 taxis in Singapore, which enabled to quantify the instantaneous driving cycle parameters in high spatio-temporal resolution with the input of real-time emission model in calculating CO2, NOx, VOCs, and PM. Luo et al. (2016) extracted speed information from taxi GPS data of Shanghai to quantify spatial and temporal hotspot emissions using COPERT model, disclosing a distribution of dual-core cyclic structure. However, their study failed to distinguish vehicles of different type and emission standard, and the final the output contains solely plotting-based figures.

The existing literatures about on-road emissions mainly were associated with three gaps as follows. First, the majority of studies requires huge amount of survey data as the input for the simulation stage. Since some research are in a rather large scale, the temporal difference was neglected, by just analyzing day, month or year as a whole. Second, most studies just estimate emissions from the original GPS sources without further reckoning on various vehicle types and emission standards. Finally, the results are mainly qualitative plotting or simple statistics without explanation with additional information such as land use and built environment. In fact, from the perspective of sustainable human settlements, it’s necessary to explore the correlation between the road emissions and the air pollution concentrations, which should be the ultimate objective in analyzing on-road emissions.

The purpose of this study is to investigate the 24-h line source emission pattern of the road network so as to analyze the problematic segments with continuous or unconventional high emissions. Formulas from COPERT were selected to estimate NOx emission of different vehicle types and emission standards. The geographical methods including geographical detector, MORAN’S Index and spatial regression were applied to analyze the spatial-temporal characteristics, as well as impacts from the built environment factors in the vicinity. In comparison, the previous study of the same authors (Zhang et al., 2017) has focused on discussing the spatiotemporal distribution and the cause of road congestion based on taxi GPS trajectories, revealing how land use and built environment affect road congestions. The main objective is to propose a modeling frame to combine Fuzzy C-Means (FCM) clustering and spatial regression, while this paper mainly focuses on road emissions, and utilizes COPERT model to calculate emission factor of road traffic, which is relatively innovative for the transportation big data analytics. One of the key innovation is to approximate the traffic flow composition for all roads by incorporating the surveyed results from 40 selected segments (both primary and secondary arterials), which extends the applicability of the sampled Didi on-line car-hailing data. Then, the 24-h trend pattern of spatial correlation between road emissions and air pollution concentration was further analyzed to emphasize the spatial and temporal relationship of NOx traffic emission to the overall atmospheric air quality. More specifically, the two sub-objectives of this study are
To use Didi massive car-hailing data to extract line source emission pattern of road segments, and then classify segment-based 24-h emission pattern into various clusters through FCM clustering method.

To assess the impact of selected built environment and land use factors on segment-based emission rate using the spatial auto-regressive moving average (SARMA) regression model, and determine the factors that may significantly contribute to the traffic emissions.

2. Materials and data

2.1. Didi service GPS data

The massive Didi hailing-car GPS data from August 8 to 12, 2016 (Monday to Friday), sunny, were provided by Didi Inc., including both trajectories and orders from the registered taxis and the regular Didi hailing cars. The raw trajectory datasets include Order ID, Signal Received Time, Longitude, Latitude, Speed, etc. (Sun et al., 2014), with the time interval as 3 s. The overall number of records from August 8 to 12, 2016, are 9,973,471 (2,855,013 taxi records), 10,755,260 (3,184,219 taxi records), 9,544,299 (2,800,285 taxi records), 9,211,886 (2,911,388 taxi records), and 10,296,053 (2,931,417 taxi records) respectively, all in a similar volume level. The trajectories from both taxis and regular cars were first aggregated, so that the hourly average speeds for each day were obtained, as presented in Fig. 1. It was found that the 24-h average speed demonstrates a high similarity (with the largest difference at 4.97%), indicating that the 5-day data can represent multi-day patterns from the perspective of time-of-day variation.

Then, the space projection of GPS records for both AM and PM peak hours were obtained and are shown in Fig. 2, which also demonstrate rather similarities for both morning and evening peaks. Thus, the temporal and spatial similarity of multi-day data have been validated, and the 5-day datasets were chosen as the research subjects for the entire trajectories.

By focusing on certain amount of traffic, only segments, defined as the link between two intersections, from the primary and secondary roads within the Central City (Outer Ring) of Shanghai were considered. A criterion was set as the segment length has to be longer than 300 m to avoid the excessive influence of intersections. To this end, a total of 853 road segments were screened out by this stage.

2.2. Data processing

As the raw data contain Order ID, Signal Received Time, Longitude, Latitude, Speed, etc., the flawed records are mainly caused by location deviation and data missing. Consequently, the trajectory points that do not match to the nearest roads within 15 m, stay within same location for over five minutes, or are with a speed over 120 km/h or with a sudden distance deviation over 100 m were eliminated. Moreover, ten vehicles per hour were set as the threshold for each segment to avoid error propagation, and finally an overall 536 segments were kept with these criteria. Then, the hourly average speed of each segment was calculated by averaging the speeds of all Didi-hailed cars passed through, and the sampled traffic flow for each segment was also extracted. More details for data processing procedure can be found from the previous study (Zhang et al., 2017).

To calculate the line source emissions, the Didi hailed cars alone are not enough. Supplementary field surveys were carried out to approximate the traffic composition, e.g. passenger cars, autobuses and trucks. As the urban expressway system in Shanghai has three
rings: inner ring, middle ring and outer ring, which typically distinguish the city center from the suburb, locations for the proposed traffic composition survey were scheduled to be scattered at each zone along the expressways. Finally, 40 road segments were selected to investigate the vehicle composition, as shown in Fig. 3, among which 10 roads were covered for each area zone, with five primary and five secondary, respectively. The results of surveying traffic composition for primary and secondary arterial segments within each zoning area are provided in Table 1. It can be figured out that passenger cars generally consist of more than 85% traffic volume, except for the primary roads in Outside Outer Ring, Outer-middle Rings, and Middle-inner Rings, in which the percentage of passenger cars are between 80% and 85%. Also, the percentages of trucks within the outer ring, particularly within the middle ring, are smaller (within 5%), as trucks are forbidden to enter into many central urban areas in Shanghai.

2.3. Traffic volume estimation

With the surveys, the total number of vehicles as well as the traffic compositions in each category, namely passenger car, autobus and truck on the 536 filtered roads were obtained. By combining with the Didi-hailed taxis data, which are all passenger

![Geospatial distribution of trajectory points for AM and PM peak hours.](image)

Fig. 2. Geospatial distribution of trajectory points for AM and PM peak hours.
cars, the percentage of the Didi hailing cars to the total passenger cars on each road segment were calculated, respectively. The ratio value was estimated as 12.9%, with no large differences across the zoning area and road type. Then, the number of on-road passenger cars, autobuses and trucks for the rest 496 segments can be estimated as follows:

1. For each segment \( k \), calculate the number of overall passenger cars based on the estimated ratio, 12.9%:

\[
Q_{1k} = q_{1k}/0.129
\]

where \( Q_{1k} \) is the number of passenger cars (including both Didi and the regular passenger cars) on road segment \( k \); and \( q_{1k} \) is the number of Didi hailing vehicles (including taxis and ordinary on-hailing cars) on road segment \( k \).

2. For each segment \( k \), calculate the number of autobuses and trucks according to the surveyed traffic composition provided in Table 1:

![Fig. 3. Traffic volume and composition investigations on 40 selected road segments.](image-url)

Table 1
Surveying traffic composition for each zoning area per vehicle type (%).

<table>
<thead>
<tr>
<th>Zoning area</th>
<th>Passenger car</th>
<th>Autobus</th>
<th>Truck</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Taxi</td>
<td>Ordinary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outside outer ring</td>
<td>Primary 14.3</td>
<td>68.1</td>
<td>8.2</td>
<td>9.4</td>
</tr>
<tr>
<td>Secondary</td>
<td>13.6</td>
<td>74.5</td>
<td>6.3</td>
<td>5.6</td>
</tr>
<tr>
<td>Outer-middle rings</td>
<td>Primary 16.3</td>
<td>67.2</td>
<td>10.4</td>
<td>6.1</td>
</tr>
<tr>
<td>Secondary</td>
<td>14.1</td>
<td>71.7</td>
<td>9.6</td>
<td>4.6</td>
</tr>
<tr>
<td>Middle-inner rings</td>
<td>Primary 6.9</td>
<td>73.9</td>
<td>16.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Secondary</td>
<td>8.4</td>
<td>76.9</td>
<td>12.3</td>
<td>2.4</td>
</tr>
<tr>
<td>Within inner ring</td>
<td>Primary 7.7</td>
<td>81</td>
<td>9.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Secondary</td>
<td>8.1</td>
<td>80.6</td>
<td>8.6</td>
<td>2.7</td>
</tr>
</tbody>
</table>
where \(Q_2_k\) and \(Q_3_k\) are the number of autobuses and trucks on road segment \(k\), and \(R_{1k}\), \(R_{2k}\), and \(R_{3k}\) are the proportions of passenger cars (including both Didi and the regular passenger cars), autobuses and trucks on the road segment \(k\), which should be located on the corresponding zoning area and the road type.

### 2.4. Calculation of NO\(_x\) line source emission

As the emission standard in China is similar to that in Europe, for example, Euro IV is equal to China Stage IV, this study mainly focused on the standards from China Stage III to V, issued in 2008, 2009 and 2014, respectively within Shanghai. Each newly-purchased vehicle should comply with the emission standard of the year, from which the standard distribution could be estimated. According to *Shanghai Municipal Comprehensive Transportation Operation Annual Report* (SRIUTD, 2015, 2016), the annual retired rate of vehicles in Shanghai is around 6%, including passenger cars, autobuses, and trucks. This study first obtained the number of newly-purchased vehicles of each model, and then incorporated the annual retiring rate to estimate the number of vehicles of China Stage III to V. Since China Stage I standard had been obsoleted in 2015 (Wang, 2015), the number of vehicles in Stage II equals to the total vehicle amount minus the number of vehicles from Stages III to V, with the results summarized in Table 2 below.

In this study, the emission rates of vehicles were calculated by type and standard, as follows.

\[
TE_k = \sum_{i=1}^{3} \sum_{j=2}^{5} EF_{i,j,k} \times q_{i,j,k} \times L_k
\]

\[
EF_k = TE_k / q_k
\]

Table 2

<table>
<thead>
<tr>
<th>China stage</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passenger car</td>
<td>17.34</td>
<td>12.22</td>
<td>55.63</td>
<td>14.81</td>
<td>100</td>
</tr>
<tr>
<td>Autobus</td>
<td>11.79</td>
<td>16.74</td>
<td>49.82</td>
<td>21.65</td>
<td>100</td>
</tr>
<tr>
<td>Truck</td>
<td>45.71</td>
<td>12.02</td>
<td>37.85</td>
<td>4.42</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Vehicle type</th>
<th>China stage</th>
<th>Eq.</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passenger car</td>
<td>II (6)</td>
<td>2.84E – 1</td>
<td>–2.34E – 2</td>
<td>–8.69E – 3</td>
<td>4.43E – 4</td>
<td>–1.68E – 4</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>III (6)</td>
<td>9.29E – 2</td>
<td>–1.22E – 2</td>
<td>–1.49E – 3</td>
<td>3.97E – 5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IV (6)</td>
<td>1.06E – 1</td>
<td>0</td>
<td>–1.58E – 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>V (6)</td>
<td>1.89E – 1</td>
<td>1.57</td>
<td>8.15E – 2</td>
<td>2.73E – 2</td>
<td>–2.49E – 4</td>
<td>–2.68E – 1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Autobus</td>
<td>II (7)</td>
<td>1.09E01</td>
<td>–2.18E – 1</td>
<td>1.11E02</td>
<td>–1.13</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>III (7)</td>
<td>3.26E02</td>
<td>–1.77</td>
<td>4.01E01</td>
<td>–5.77E – 1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IV (7)</td>
<td>2.47E01</td>
<td>5.30E – 1</td>
<td>2.19E03</td>
<td>3.54</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>V (7)</td>
<td>–4.81E01</td>
<td>1.44E01</td>
<td>2.26E – 1</td>
<td>6.62E01</td>
<td>1</td>
<td>–6.33E – 1</td>
<td>2.18E – 1</td>
<td></td>
</tr>
<tr>
<td>Truck</td>
<td>II (7)</td>
<td>2.03</td>
<td>0</td>
<td>–3.18E – 1</td>
<td>0</td>
<td>2.41E – 4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>Decrease by 16% compared with China Stage II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>Decrease by 32% compared with China Stage III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V (7)</td>
<td>5.14E – 1</td>
<td>–1.41E – 2</td>
<td>–9E – 2</td>
<td>–</td>
<td>–</td>
<td>3.43</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>
where, $EF$ is the average emission factor, $V$ is the hourly average speed of a certain road segment, and $a, b, c, d, e, f, g$ are coefficient parameters.

The hourly NOX emission factor $EF_k$ for the 536 road segments were calculated, and then aggregated to the average emission factor as a vector with 24 dimensions, $\{EF_1, EF_2, \ldots, EF_{24}\}$. Fig. 4 presents the vehicle average emission factor by each hour, in which the emission pattern demonstrates obvious higher values during peak hours, indicating the severest pollution, although the emission pattern is related to multiple factors, including speed, traffic composition, and traffic control (Sun and Zhang, 2017; Pan et al., 2017).

3. Result analysis and discussion

3.1. Clustering analysis based on road segment emission factors

Clustering analysis, an unsupervised machine learning approach, was used to classify road segments into different categories based on their emission patterns. First, the 24-h emission pattern of each segment $i$ was expressed by a 24-dimension vector: $X_i = \{EF_1, EF_2, \ldots, EF_{24}\}$. FCM Clustering method (Bezdek et al., 1984) was introduced to classify each segment-based 24-dimension vector into different groups by probabilities, which may be further integrated with mathematical regression, so as to assess the connection between the 24-h emission and the built environment and land use attributes.

As the cluster number has to be predetermined and requires to be evaluated by validity indexes to reach a meaningful and explicable result, the value was tested from 2 to 7. Four cluster validity indexes were introduced and calculated as the selection criteria to assess the cluster results, taking both similarity within a cluster and difference across clusters into account: Fuzzified PBM (PBMF) (Pakhira et al., 2005), Minimum centroids’ distance (MCD) (Zhu and Nandi, 2014), Partition coefficient (PC) (Bezdek, 1973) and Fukuyama-Sugeno index (FSI) (Pal and Bezdek, 1995). Finally, four clusters was found that can balance cluster dispersion and in-cluster compactness, and consequently the best clustering number was chosen as four. Based on the maximum probability, the number of objects fallen within each cluster are 130, 202, 103 and 101, respectively.

Fig. 5 presents the 24-h average emission factor of the four clusters, in which the membership is expressed by the color and size of the bubble. Fig. 5(a) has the lowest emission rate per vehicle with a mean of 0.238 g/km. Cluster 2 has the mean value of 0.286 g/km with a stable dispersion. Lower rate and stable trend mean perfect balance between traffic volume and operational speed. Cluster 3 with a mean of 0.358 g/km, which represents a notable increase of average emission. Cluster 4 has the highest value, 0.439 g/km, and the largest deviation, which has obvious extreme peak-hour phenomenon.

The spatial distribution of clustering result for all segments are illustrated in Fig. 6, in which each segment was allocated to the affiliated cluster based on highest membership value. It is shown that segments with high emission rates (Cluster 4) mainly concentrate between inner and middle rings, carrying majority of commuting vehicles. Additionally, this region also has the highest proportion of heavy-duty vehicles according to the field observation. Cluster 3 with the second highest emission rate is mainly located on outskirts and city center, which may be due to the low travel efficiency within the downtown CBDs. Clusters 1 and 2 indicate comparatively low emission rate, for which additional spatial regression is needed for further explanation.

3.2. Spatiotemporal characteristics of emission pattern

3.2.1. Selection of influencing factors

To explain the formation of extreme high emission rate, various influencing factors were chosen for further investigations with spatial analysis methods including Geographical Detector, Moran’s I and SARMA. The possible built environment factors were selected as follows:

F1: Road type (primary or secondary road). During the numeric analysis, 1 indicates the primary road and 2 for secondary road.
Road type may largely impact the travel speed of vehicles passed by, which is the key factor for emissions.

**F2:** Road segment length. Here, the segment was defined as the road component between two intersections. The starting and ending intersections were intentionally left out, as those are generally considered as the severest polluted nodes (He et al., 2015), and should be analyzed separately.

**F3:** Distance to the nearest ramp (defined by Euclidean distance). Traffic volume close to ramp area is always with strong oscillation, which may cause serious congestion and emission problems (He et al., 2016). Particularly, low speed and braking nearby would induce extremely high emissions and deserve further investigations.

**F4:** Number of bus stations along the road segment per 100 m. Roads with large number of bus stations are always related to the commuting corridors or even transportation hubs in the vicinity, indicating more vehicles and possible aggravated traffic congestion and emissions (Long et al., 2010).

**F5:** Number of schools within 500 m per 100 m in segment length. In general, elementary schools or even middle schools may produce traffic pressure on the surrounding roads during daily attending or leaving periods by the picking-up behavior of parents or school buses (Yu and Liu, 2011), which may impact on-road emissions, and is particularly significant for large cities.

**F6:** Distance to the nearest metro station. Metro stations always carry large amount of passenger flow and transship of buses or passenger cars, which may obstruct the roads around (Huang et al., 2017).

**F7:** Distance to the nearest hospital. Hospital as one of the most important built environment factors and public service facility

Fig. 5. Average emission by each cluster (a) Cluster 1 (# of points = 130) (b) Cluster 2 (# of points = 202) (c) Cluster 3 (# of points = 103) (d) Cluster 4 (# of points = 101).
Fig. 5. (continued)

Cluster 3

Average Emission Factor (g/km)

Time
1:00-2:00 3:00-4:00 5:00-6:00 7:00-8:00 9:00-10:00 11:00-12:00 13:00-14:00 15:00-16:00 17:00-18:00 19:00-20:00 21:00-22:00 23:00-24:00

Cluster 4

Average Emission Factor (g/km)

Time
1:00-2:00 3:00-4:00 5:00-6:00 7:00-8:00 9:00-10:00 11:00-12:00 13:00-14:00 15:00-16:00 17:00-18:00 19:00-20:00 21:00-22:00 23:00-24:00

Fig. 6. Clustering result on segment emission factor by primary membership.
would attract large amount of traffic, which has been assessed as a key factor for on road emissions (Zhong and Bushell, 2017).

F8: The relative location to the urban expressway rings. The urban expressway system in Shanghai has three rings, inner ring, middle ring and outer ring, which typically distinguish the center with the suburb, and are considered as the spatial layout of the city. The regions divided by the three expressway rings are denoted by 0 to 3 for the zones from within the inner ring to outside of the outer ring.

F9: Number of public parking lots within 500 m. The impact of public parking lots on congestion and emissions is similar to that of hospitals, as both are important traffic generation and attraction nodes.

F10-12: Land use factors. Emissions are sensitive to trip speed and vehicular characteristics, which may be largely affected by land use status (Cao et al., 2006; Hong and Goodchild, 2014). Here, land use mainly considers commercial area (F10), residential area (F11) and transportation area (F12), as other attributes, such as education and hospital have been considered within the built environment factors before (F5 & F7). Transportation land cover mainly includes railway station, airport, transportation hubs, etc. The three factors are quantitatively expressed by the area proportion within a radius of 500 m.

The 12 variables for analysis are summarized in Table 4.

3.2.2. Geographical detector

To assess the impact of the built environmental and land use factors to certain clusters, Geographical detector (Wang et al., 2010) was introduced to preliminarily explain the extent that each factor may be responsible for clustering, as shown in Eq. (8):

$$PD_R = 1 - \frac{1}{n^2_R} \sum_{i=1}^{4} n_{Di} \sigma^2_{Di,R}$$

(8)

where $PD_R$ is the power of determinant on clustering result for factor $R$, $\sigma^2_R$ is the global variance of factor $R$ within the overall spatial region, $\sigma^2_{Di,R}$ is the divisional variation in cluster $Di$. Eq. (8) interprets the ratio of $n_{Di}$ weighted divisional variation over the global variance. Larger $PD_R$ indicates factor $R$ differs largely across clusters, and the determinant power of $R$ is stronger. If $PD_R$ equals to 1, factor $R$ alone could perfectly differentiate objects.

Fig. 7 presents the explanatory power of determination for each factor ascendingly. F1: Road type (0.074) has the highest $PD$, which may greatly impact vehicles’ emission rate. F4: Number of bus stations alongside got the second (0.057), as oversize autobuses not only reduce the operation speed of road but also bring higher emission rate compared with other vehicle types. The third factor is the proportion of transportation land (F12: 0.049), which mainly reflects the occurrence of transportation hubs, as huge amount of mixed traffic flow pick-up passengers or cargos may occur along with hubs. F7: Distance to the nearest hospital (0.037), F5: Number of schools around (0.032), F9: Number of public parking lots (0.029), F10: Proportion of commercial land (0.029), and F6: Distance to the nearest metro station (0.027) have similar explanatory power. These are all hotspots of traffic attraction, which may cause traffic congestion or brake along with high emissions. As shown in Fig. 7, F8: Relative location to the expressway rings (0.024) also has a considerable impact on emissions, while the rest factors have little power of determinant. Moreover, factors might also cause individual impact on single cluster, which is unable to be explained by $PD$, and consequently further regression is required for detailed investigation.

3.2.3. Moran’s $I$ test

GLOBAL MORAN’s $I$ (Moran, 1950) tests whether adjacent objects have spatial similarities with each other, ranging $[-1, 1]$. The definition of MORAN’s $I$ is given in Eq. (9) as follows:

$$I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (X_i - \bar{X})^2} (i \neq j)$$

(9)

where $X_i$ and $X_j$ are observed values, indicting the membership of clusters here; $w_{ij}$ is the spatial weight matrix describing the spatial relationship between objects. $I = 0$ means total independence, $I > 0$ reflects positive correlation, and vice versa. As road network

Table 4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: Road type</td>
<td>Rd_type</td>
<td>1</td>
<td>–</td>
<td>2</td>
</tr>
<tr>
<td>F2: Road segment length (m)</td>
<td>Rd_len</td>
<td>300</td>
<td>1314</td>
<td>4510</td>
</tr>
<tr>
<td>F3: Distance to the nearest ramp (m)</td>
<td>Dist_ramp</td>
<td>8.33</td>
<td>990.90</td>
<td>4304</td>
</tr>
<tr>
<td>F4: Number of bus stations along the road segment per 100 m (stations/100 m)</td>
<td>Num_bus</td>
<td>0</td>
<td>0.24</td>
<td>2.47</td>
</tr>
<tr>
<td>F5: Number of schools within 500 m per 100 m (schools/100 m, r = 500 m)</td>
<td>Num_scho</td>
<td>0</td>
<td>0.46</td>
<td>2.92</td>
</tr>
<tr>
<td>F6: Distance to the nearest metro station (m)</td>
<td>Dist_metro</td>
<td>8.05</td>
<td>814.9</td>
<td>3213</td>
</tr>
<tr>
<td>F7: Distance to the nearest hospital (m)</td>
<td>Dist_hosp</td>
<td>13.47</td>
<td>979.8</td>
<td>6207</td>
</tr>
<tr>
<td>F8: Relative location to the expressway rings</td>
<td>Loca_ring</td>
<td>0</td>
<td>–</td>
<td>3</td>
</tr>
<tr>
<td>F9: Number of public parking lots within 500 m</td>
<td>Parking_lot</td>
<td>0</td>
<td>3.15</td>
<td>39</td>
</tr>
<tr>
<td>F10: Commercial area proportion (%)</td>
<td>Com_pro</td>
<td>0</td>
<td>5.25</td>
<td>59.12</td>
</tr>
<tr>
<td>F11: Residential area proportion (%)</td>
<td>Res_pro</td>
<td>0</td>
<td>29.29</td>
<td>99.41</td>
</tr>
<tr>
<td>F12: Transportation area proportion (%)</td>
<td>Trans_pro</td>
<td>0</td>
<td>15.63</td>
<td>100</td>
</tr>
</tbody>
</table>
has spreading characteristics, one segment may influence another object far away. Consequently, $w_{ij}$ should be based on distance decay as below (Gibbons and Overman, 2012):

$$
\begin{cases}
\exp \left(-0.5 \left( \frac{d_{ij}}{b} \right)^2 \right), & d_{ij} < b \\
0, & \text{otherwise}
\end{cases}
$$

(10)

where $d_{ij}$ is the distance between the two adjacent objects, and $b$ is chosen as 1000 m in this study which was set as the upper limit of buffer size when analyzing impact of road emissions, as suggested by Liu et al. (2016). Moran’s $I$ mainly tests whether the membership of objects to a cluster has aggregation effect spatially.

Table 5 provides the MORAN’s $I$ value as well as the corresponding statistical p-value calculated for each cluster. It was found that all clusters have a positive $I$ value and passed the 5% level of significance test, which means the segments belonging to the same cluster would spatially gather together. This further indicates the effectiveness of FCM clustering. The values of all indexes are relatively low, as the global Moran’s $I$ was used here for each cluster, which was calculated based on the membership of all 536 road segments to certain cluster.

### 3.2.4. SARMA regression of road segments

The advantage of using FCM is to conduct regression based on continuous membership (dependent variable). In addition to the referred 12 factors, neighboring segments may also influence traffic demand on a road as well as emissions, and the lagging influence and spatial dependency may play important roles. To this end, a spatial lag model called mixed spatial autoregressive moving average model (SARMA) (Anselin et al., 1996) is introduced to consider both dependence and errors with closer objects having a greater impact. The structure of SARMA is shown in Eq. (11):

$$
\begin{align*}
\gamma_i &= (1-\rho W)^{-1}X\beta + (1-\rho W)^{-1}u \\
u &= (1 + \lambda W)\varepsilon, \varepsilon \sim N(0,\sigma^2I)
\end{align*}
$$

(11)

where $\gamma_i$ is the segment’s membership belonging to Cluster $i$, and $X$ is the vector of 12 environment characteristics, both were standardized to $[-1, 1]$ before regression. $\rho$ is the spatial autoregressive parameter measuring neighborhood effects, $\rho > 0$ means positive correlation and vice versa. $\lambda$ is the spatial error coefficient, disclosing and quantifying the inherent similarity or dissimilarity. $\varepsilon$ is the random error term, $W$ is the spatial weight matrix as presented in Eq. (10), $\beta$ is the coefficient vector, and $\sigma^2I$ is the variance of the random error. The result of SARMA regression is shown in Table 6.

### Table 5

<table>
<thead>
<tr>
<th>Index</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MORAN’s $I$</td>
<td>0.232</td>
<td>0.324</td>
<td>0.199</td>
<td>0.536</td>
</tr>
<tr>
<td>P value</td>
<td>7.008e–09</td>
<td>1.449e–15</td>
<td>5.486e–07</td>
<td>2.2e–16</td>
</tr>
</tbody>
</table>
### Table 6
Results of SARMA regression for cluster membership prediction.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Cluster 1 lowest</th>
<th>Cluster 2 moderate</th>
<th>Cluster 3 abnormal</th>
<th>Cluster 4 highest</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.7</td>
<td>0.0178</td>
<td>0.5758</td>
<td>0.0183</td>
</tr>
<tr>
<td>F1: Road_type</td>
<td>-0.0455</td>
<td>0.0119</td>
<td>0.0015</td>
<td>0.0091</td>
</tr>
<tr>
<td>F2: Rd_len</td>
<td>0.0165</td>
<td>0.0128</td>
<td>0.0202</td>
<td>0.0104</td>
</tr>
<tr>
<td>F3: Dist_ramp</td>
<td>0.011</td>
<td>0.0122</td>
<td>0.0109</td>
<td>0.0089</td>
</tr>
<tr>
<td>F4: Num_bus</td>
<td>0.0025</td>
<td>0.0115</td>
<td>0.0082</td>
<td>0.0095</td>
</tr>
<tr>
<td>F5: Num_scho</td>
<td>0.012</td>
<td>0.0134</td>
<td>-0.004</td>
<td>0.0112</td>
</tr>
<tr>
<td>F6: Dist_metro</td>
<td>0.0269</td>
<td>0.0132</td>
<td>-0.0038</td>
<td>0.0113</td>
</tr>
<tr>
<td>F7: Dist_hosp</td>
<td>-0.0113</td>
<td>0.0142</td>
<td>-0.0028</td>
<td>0.0114</td>
</tr>
<tr>
<td>F8: Loca_ring</td>
<td>0.0226</td>
<td>0.0159</td>
<td>-0.0125</td>
<td>0.0127</td>
</tr>
<tr>
<td>F9: Parking_lot</td>
<td>-0.0186</td>
<td>0.0144</td>
<td>-0.018</td>
<td>0.0114</td>
</tr>
<tr>
<td>F10: Com_pro</td>
<td>0.004</td>
<td>0.0129</td>
<td>0.0038</td>
<td>0.0097</td>
</tr>
<tr>
<td>F11: Res_pro</td>
<td>0.0063</td>
<td>0.0139</td>
<td>0.0039</td>
<td>0.0103</td>
</tr>
<tr>
<td>F12: Trans_pro</td>
<td>0.0221</td>
<td>0.0125</td>
<td>0.0021</td>
<td>0.0099</td>
</tr>
<tr>
<td>Rho</td>
<td>0.0164</td>
<td>0.0369</td>
<td>0.1424</td>
<td>0.0351</td>
</tr>
<tr>
<td>Lambda</td>
<td>0.1173</td>
<td>0.1032</td>
<td>-0.9419</td>
<td>0.0972</td>
</tr>
<tr>
<td>LR test</td>
<td>1.833</td>
<td>15.171</td>
<td>0.808</td>
<td>11.58</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>87.99</td>
<td>132.82</td>
<td>77.45</td>
<td>69.79</td>
</tr>
<tr>
<td>AIC</td>
<td>-143.98</td>
<td>-233.65</td>
<td>-122.9</td>
<td>-107.57</td>
</tr>
</tbody>
</table>

**Note:**
* Indicates p < 0.1.
** Indicates p < 0.05.
*** Indicates p < 0.01.
For Cluster 1 (Lowest emission rate), almost all significant factors (i.e. F1: Road_type, F2: Rd_len, F3: Dist_ramp, F6: Dist_metro, F8: Loca_ring, F9: Parking_lot, F12: Trans_pro) cause less impact on operational speed. Although primary roads (F1) carry more traffic, they may associate with higher speed, and consequently cause less pollution. Longer segments (F2) are less affected by signal control, where vehicles suffer little braking and lining up, thus emitting less NOx. Ramps (F3) are the merging points of traffic flow, and vehicles on segments far away from ramps have low emission rate. The role of distance to metro stations (F6) is similar to that of ramps. Relative location to expressway rings (F8) is positive, indicating low emission rate segments interspersed on outskirts. More parking lots around (F9) would impede the traffic, and reduce the membership to Cluster 1. The proportion of transportation land (F12) is a special factor, since hubs such as airports or rail stations, may have better traffic organization, resulting in better operation and less single-vehicle emission.

For Cluster 2 (Moderate emission rate), emission rate per vehicle rises slightly. Road length and number of parking lots around remain key factors as Cluster 2. A new significant built environment is the number of schools within 500 m, which represents the amount of picking ups during attending and leaving periods, especially in large cities. This would largely disrupt the vehicle speed passing by, as well as aggravate emission rate. The proportion of residential land is significantly positive. High density of residential land may cause tidal phenomena, however the roads around residential land may have extreme lower traffic during non-peak hours, driving down the average emission rate.

Emission rate in Cluster 3 (Abnormal emission rate) have higher deviation within 24 h, and obvious rush hour effect appears. The significant factors in Cluster 3 all somewhat demonstrate characteristics of commuting pressure. Longer roads composite commuting corridors, even though some may not be arterials. The traffic problems near ramps and metro stations are intensified during peak hours, as the closer to these areas means higher emission rate. Residential land here largely emphasizes the time dependency.

Cluster 4 (Highest emission rate) is the key to explain high NOx emission. Short segments generally experience more signal control and are related to high emission rates. Relative location to ring is negative, indicating vehicles within the middle and inner rings have higher emission rate, which is consistent with rational recognition of downtown congestion. However, the parking lot effect changes to positive, indicating that more parking lots may lead vehicles hunting for possible parking lots and parking positions, which in turn induces increased congestion and vehicle emissions. A new factor is the number of bus stations alongside. According to COPERT, buses generally have higher emission rate and frequently entering and exiting station also influence the traffic flow, and thus more stations mean higher emissions. A higher proportion of commercial land always attracts more traffic and tends to become the center of congestion accompanied by high NOx emission.

3.3. Correlations between atmospheric concentration (NO2) and line source emission

To investigate the potential correlation between on-road emissions and atmospheric pollutant concentration, the hourly NO2 concentrations from the atmospheric environmental monitoring stations within the studying area were extracted from the website of Shanghai Metropolitan Environment Monitoring Center (www.semc.gov.cn/aqi/home/Index.aspx). Since NO within atmosphere could be easily oxidized into NO2, it is reasonable to use NO2 as a surrogate of NOX emissions, with the geo-spatial concentration trends that traffic and the built environment impacts. It was found that the short road segments with more signal controls would largely impact NOx emission, and green wave control may mitigate the situation. Being close to city center, high density of bus stations, and high proportion of residential or commercial land would also increase the emission rate. The high-density bus stations on the commuting corridor should be reduced and relocated to the neighboring roads. To mitigate traffic emissions, bus stations is not recommended to be located within 500 m to the parking lot.
area. Since ramps are crucial to high emission rate, the land use in the vicinity should avoid having high trip generations. Moreover, as high proportion of transportation land use tend to reduce the traffic emissions, which may emphasize the importance of the planning and construction of integrated transport hub within city instead of suburban or urban fringe areas. However, as the emission factors of autobuses and trucks are generally much higher than the regular passenger cars (Sun et al., 2017), it is essential to have them electrified for reducing daily traffic emissions. The proposed framework could be used for any spatial problems investigation, particularly to estimate the traffic emission rate from Didi service GPS data and COPERT model by incorporating different vehicle types, which is an important innovation. The result could guide urban planning and assist to identify road segments with serious traffic emission problems.

This paper also compared the hourly emission trend with the Kriging interpolated atmosphere NO$_2$ concentration, which found positive correlation and revealed the temporal variation of transportation emission contribution. While the results are promising, further studies need to be conducted to improve the performance of the model. First, this study just chose Shanghai as a case study, the conclusions should be generalized after analyzing more cities and regions. Second, the correlation of urban traffic line source emissions and atmospheric pollution concentration was just a preliminary investigation. The five days trajectories chosen as the research samples are mainly representatives for typical day situations instead of comprehensive circumstances. Additional data sources, including both vehicle related data and atmospheric factors, as well as dispersion models may be incorporated for further research and investigations. Finally, as many driving behaviors may be extracted from raw trajectory and order data, it’s also necessary to analyze the spatiotemporal relationship between congestion and emissions, which could better explain formation of abnormal emissions from the perspective of traffic operation.
Acknowledgments

The research was supported in part by the Major Project of National Social Science Foundation of China [16ZDA048], the Shanghai Municipal Natural Science Foundation, China [17ZR1445500], and the Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, China (Online Car-Hailing Service and Urban Transport) [TJDDZHCX006]. Any opinions, findings or conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.trd.2018.04.024.

References


