High-resolution mapping of regional traffic emissions by using land-use machine learning models

Xiaomeng Wu[#,1], Daoyuan Yang[#, 2], Ruoxi Wu[1], Jiajun Gu[3], Yifan Wen[1], Shaojun Zhang[1, 2, 4, 5], Rui Wu[2], Renjie Wang[2], Honglei Xu[2], K. Max Zhang[3], Ye Wu*[,1, 2, 4, 5] and Jiming Hao[1, 4, 5]

[1] School of Environment and State Key Joint Laboratory of Environment Simulation and Pollution Control, Tsinghua University, Beijing 100084, China, Tsinghua University, Beijing 100084, P. R. China

[2] Laboratory of Transport Pollution Control and Monitoring Technology, Transport Planning and Research Institute, Ministry of Transport, Beijing 100028, China

[3] Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14853, U.S.A.

[4] State Environmental Protection Key Lab of Sources and Control of Air Pollution Complex, Beijing 100084, P. R. China

[5] Beijing Laboratory of Environmental Frontier Technologies, Beijing 100084, P. R. China

[#] These authors contributed equally to this work.

Correspondence to: Ye Wu (ywu@tsinghua.edu.cn)

Abstract. On-road vehicle emissions are a major contributor to significant atmospheric pollution in populous metropolitan areas. We developed an hourly-based, link-level emissions inventory of vehicular pollutants using two land-use machine learning methods based on the datasets of road traffic monitoring in the Beijing-Tianjin-Hebei (BTH) region. The results indicate that a land-use random forest (LURF) model is more capable of predicting traffic profiles than other machine learning models on most occasions for this study. The inventories under three different traffic scenarios depict a significant temporal and spatial variability in vehicle emissions. NO_X, fine particulate matter ($PM_{2.5}$) and black carbon (BC) emissions from heavy-duty trucks (HDTs) in general have higher emission intensity on the highways connecting to regional ports. The model finds a general reduction in Light-Duty Passenger Vehicles when traffic restrictions were implemented, but a much more spatially heterogeneous impact on HDTs, with some road links experiencing up to 40% increases in HDT traffic volume. This study demonstrates the power of machine learning approaches to generate data-driven and high-resolution emission inventories, which provides a platform to realize the near real-time process of establishing high-resolution vehicle emission inventories for policy makers to engage in sophisticated traffic management.

1. Introduction

The rapid social and economic growth in China has driven the development of road transportation systems and mobility services over the past few decades. This macrotrend also aligns with the faster pace of urban expansion and agglomeration, creating higher travel activities that are not only caused by urban commuting but also are caused by intercity connections. Consequently, on-road transportation systems have resulted in substantial challenges regarding traffic congestion, carbon emissions, air pollution, and land-use issues (Uherek et al. 2010; Waddell 2002; Chapman 2007). To address traffic-related air pollution issues, previous studies have developed link-level emission inventories for metropolitan areas or their urban cores. Notably, more studies have recognized the considerable environmental impact from nonlocally registered vehicles, especially the nonlocal heavy-duty diesel trucks (HDDTs) serving for regional freight purposes. For example, nonlocal HDDTs are estimated to contribute nearly 30% to 40% of the total on-road emissions of nitrogen oxides (NO_X) and fine particulate matter ($PM_{2.5}$), which are even greater than the contributions of the 5 million local passenger cars (Wang et al. 2011; Yang et al. 2019). Undoubtedly, for transportation hubs, such as Beijing, we see a clear need to support policymaking that road emission inventories should be enlarged to the multiprovince level to improve the management of transportation emissions.

The technological evolution in intelligent transportation systems has facilitated emission inventories for megacities. For example, Gately et al. applied global positioning system (GPS)-informed speed data from mobile phones and vehicles (Gately et al. 2017) to map the emission fluxes from vehicles in the Greater Boston region. In addition to such trajectory data (Sun, Zhang, and Shen 2018), open-accessed traffic congestion indexes could also be derived from navigation companies or municipal government agencies to dynamically estimate road speeds. For traffic volume and fleet mix, radio-frequency identification (RFID) and traffic cameras are capable of reporting detailed vehicle counts by using license plate numbers (Zhang et al. 2018). These real-world traffic datasets are useful for elucidating temporal and spatial variations in traffic emissions. However, we are still confronted with a few challenges to construct multiprovince, link-level emission inventories by utilizing these developed methods that are applicable to smaller research domains. First, the annual averaged daily traffic (AADT) data, for example, which could be assessed from the U.S. Highway Administration, typically uses the traffic profiles of a select portion of a roadway system (i.e., “sample panel”) to represent the “full extent” of the system. Second, simple assumptions and empirical adjustments of vehicle kilometers traveled (VKT) are often used to downscale state-level or national-level AADT profiles to traffic patterns of specific counties (Gately, Hutyra, and Sue Wing 2015). Both of these factors could result in estimates of the spatial variations in traffic volumes that may not represent real-world patterns. Furthermore, the AADT datasets are updated per year according to the annual submission from all states. Therefore, the AADT datasets could support the average analysis of seasonal or day-of-week variations(McDonald et al. 2014) but are limited to reflecting emissions in a quasi-dynamic fashion (e.g., hourly).

Traffic demand modelling is a useful complement to measurements and simple empirical downscaling, which has been utilized to assist the development of emission inventories(Gately et al. 2017; Zhang et al. 2018). However, transportation simulation methods are often time-consuming, machine learning method then represents yet another complementary way of estimating traffic flows in a particular context, a way that is faster than full traffic demand modelling and more able to adapt to local conditions than simpler empirical approaches.

The research domain of this study, the Beijing-Tianjin-Hebei (BTH) region (see Fig. S1), geographically covers three provincial-level administrative regions with a total land area of 217,000 km^2. As the national political center, the BTH region has also developed the busiest freight system in northern China but has suffered from the worst air quality since the 2000s. We utilized hourly traffic profiles including volume, speed and fleet mix obtained from the governmental intercity highway monitoring network to pioneer land-use machine learning methods for developing link-level emission inventories. The methodology can potentially be used either to map traffic characteristics on a larger scale (at the national level) or to deal with real-time urban traffic data streams that need to overcome the challenges of computational accuracy and efficiency.

2. Methodology and Data

2.1 Research domain and emission calculation

The government traffic monitoring datasets mainly cover main intercity highways, such as expressways, national highways and provincial highways. Notably, we did not include urban sections or minor roads in the research domain because the traffic profiles of these roads were administered by local governmental agencies. The network of intercity main roads in the BTH region has a total length of 50,660 km, including 18,824 km of expressways, 8,989 km of national highways, and 22,847 km of provincial highways (See the definition of road types in Table S1). The entire region has held 22.38 million registered vehicles (motorcycles excluded) by 2017, with an average annual growth rate of 7% since 2013. In addition, freight trucks for mass transportation of coal and steel from other provinces pass through the BTH region in very large numbers because this region accounts for approximately one quarter of the total steel production in China.

The emissions of primary vehicular pollutants (carbon monoxide, CO; total hydrocarbon (THC); nitrogen oxide, NO_X, $PM_{2.5}$, and black carbon, BC) were calculated with a high-resolution method in a temporal and spatial framework. A link-level emission inventory modeling framework, called EMBEV-Link, was used to complete the emission calculation (Yang et al. 2019). For each road link, hourly emissions are the product of the traffic volume, link length and speed-dependent emission factors (see Eq. 1).

$$E_{h,j.l} = \sum_t \mathrm{EF}_{c,j}(v) \times \mathrm{TV}_{c,h,l} \times L_l \tag{1}$$

where $E_{h,j,l}$ is the total emission of pollutant j on-road link l at hour h, units in g h^{-1}; $\mathrm{EF}_{c,j}(v)$ is the average emission factor of pollutant j for vehicle category c at speed v, units in g km^{-1}; $\mathrm{TV}_{c,h,j}$ is the traffic volume of vehicle category c on-road link l at hour h, units in veh h^{-1}; and L_l is the length of road link l, units in km. According to the resolution of traffic mix data, six vehicle categories were defined, namely, light-duty passenger vehicles (LDPVs), medium-duty passenger vehicles (MDPVs), heavy-duty passenger vehicles (HDPVs), light-duty trucks (LDTs), and heavy-duty trucks (HDTs) (see Table S2). Different from the city-scale emission inventories, we did not separate the traffic volumes of transit buses and taxis from the HDVPs and LDPVs, respectively. Additionally, motorcycles were not included because we could barely observe the presence of motorcycles on these intercity highways. For each vehicle category, speed-dependent emission factors were developed based on the EMBEV model. The EMBEV model embodied the detailed fleet configurations based on vehicle registration data, which was developed based on thousands of in-lab dynamometer tests and hundreds of on-road tests (Zhang et al. 2014). Since 2015, this model has become the architype of the official National Emission Inventory Guidebook in China (Wu et al. 2016; Wu et al. 2017). Based on the EMBEV model, this study updated the BTH emission database, taking full account of the differentiated vehicle emission characteristic of Beijing, Tianjin and Hebei. The main influencing factors include: implementation timetable of vehicle emission standards, fuel quality, intensity of in-use vehicle supervision, proportion of high-emission vehicles, etc. Fig. S2 shows the fleet-average emission factors of CO, NO_X and BC for LDPVs and HDTs estimated by the updated EMBEV model in Beijing, Tianjin and Hebei. Since the traffic monitoring stations cannot obtain the emission standard information of the vehicle, the proportions of emission standard as well as vehicle age/mileage (used to estimate mileage deterioration of emissions) were assumed to be consistent with the default fleet composition data in the EMBEV model. We also made adjustment based on the restriction policy, such as the HDTs older before China III are not allowed to drive within the fifth rings in Beijing. Fig. S3 presents speed-dependent emission factors modified by different regions representing the traffic configurations (e.g., fuel type, emission standard and vehicle size) and operating conditions (e.g., fuel quality) estimated for the calendar year of 2017. Evaporative THC emissions were not included in the current EMBEV-Link model because we were limited to spatially specifying the evaporative off-network emissions.

2.2 Traffic scenarios under various transportation management schemes

Three traffic scenarios were generated as inputs to the EMBEV-Link model to observe the impact of major transportation management schemes. Scenario Weekday (*S1*) represents the average traffic patterns during weekdays (Monday to Friday) with regular driving patterns. Scenario Weekend (*S2*) represents the average traffic patterns during weekends (Saturday and Sunday), possibly with more leisure travel. Furthermore, Scenario Restriction (*S3*) represents the traffic patterns under special restrictions. The Chinese governments have launched comprehensive actions to alleviate air pollution during seriously polluted episodes. For the on-road sector, transportation restrictions are implemented during polluted days when the $PM_{2.5}$ concentrations or air quality index (AQI) were forecasted to be higher than certain thresholds. As one of the most polluted regions in China, the BTH region has implemented a package of traffic control measures, especially during the winter, with the worst meteorological conditions for pollution dispersion. In this study, Scenario Restriction (*S3*) estimated the real-world traffic patterns during a special week of November 4th to 8th, 2017, facing serious haze pollution. More extensive bans than usual were adopted during that week, so that coal-related freight trucks and high-emitting vehicles (e.g., gasoline cars in compliance with pre-China 2 standards) were restricted from many roads in the BTH region, the traffic composition (especially for the configurations of emission standard) would be adjusted accordingly in this scenario. Different from the averaged diurnal patterns in *S1* and *S2*, hourly emissions were continuously estimated throughout the period with traffic restrictions.

2.3 Generating dynamic traffic profiles based on land-use machine learning models

2.3.1 Data collection

The input data for modeling development of this study is mainly including traffic data and the land-use data. An overview of the data used to train the land-use machine learning models is summarized in Table S3 and detailed below.

Traffic data. The Ministry of Transport has established nationwide networking to monitor intercity traffic conditions(Yang et al. 2019; Zhang et al. 2018). Twenty-four-hour diurnal traffic profiles, including the volume, speed and fleet mix, were obtained from 848 intercity highway sites in the BTH region (see Fig. S1). The hourly traffic profiles of a whole week (20th to 27th) were collected in January, April, July and November in 2017 to represent the average scenarios (*S1* and *S2*). For *S3*, we further collected the hourly profiles from a special week including traffic restrictions (November 5th and 6th). Fig. S4 shows the distribution of the annual average daily traffic profiles used to train the models to see the capability for predicting the spatial distribution of these traffic profiles. Notably, the monitoring data of the traffic volumes could not separate the LDPVs and MDPVs from the light-medium-duty passenger vehicles (LMDPVs) on the whole. Therefore, we assembled the two categories when predicting the traffic volumes and separated the predicted traffic volumes according to the estimated total vehicle activity (i.e., registered population × annual VKT).

Land-use data. As land-use machine learning models have rarely been used to develop traffic emission inventories, we selected candidate spatial predictors based on previous research on the air pollution concentration predictors (Hoek et al. 2008; Lee et al. 2017). Many of these predictors, such as the population density, road density and distance from transportation facilities (e.g., airports and transit stations), are expected to affect traffic activities as well. The potential predictor variables were divided into two groups: a) variables representing a point value and b) variables representing the cumulative values of an area (buffer variables). The buffered variables were represented as a density value (standardized by the buffer area). In total, 139 spatial predictor variables (see Table S3) regarding the land-use data were calculated considering the data availability. Gong et al. transferred the global training sample set developed in 2015 at a 30-m resolution to map a 10-m resolution global land cover in 2017 (Gong et al., 2019), which was assigned into each buffer to calculate the land-use variables. Next, we utilized the points of interest (POIs) mined from the Amap API to calculate the POI density in every buffer and the distance variables (Gaode Map, 2017). The population density in the buffer was extracted from WorldPop, which estimates the numbers of people per pixel (ppp) and people per hectare (pph) and adjusts the national totals to match the UN population division with a spatial resolution of 100 m (WorldPop, 2015). The China Digital Road-network Map (CDRM) data developed by NavInfo were used in this study. The road information included the location (latitude and longitude), administration, number of lanes and designed speed limit of the monitoring sites. We categorized the intercity traffic monitoring sites by road type (e.g., expressways, national-level highways and provincial-level highways) and thus calculated the road density in each buffer.

2.3.2 Development and validation of the Machine Learning models

In order to find the most capable model for this study, five machine learning models commonly used in the environment or transport field were selected based on a comprehensive literature review and developed to estimate the spatial distributions of link-level traffic volumes, speeds and mixes for the research domain. They are land-use random forests (LURF), grading boosting decision trees (GBDT), support vector regression (SVR), Gaussian process regression (GPR) and linear regression (LR). LURF are often implemented in prediction analyses (e.g., spatial distribution of pollutant concentrations) because of their increased accuracy and resistance to multicollinearity and complex interaction problems compared to linear regression (Hastie 2009; Brokamp et al. 2018). GPR is a flexible nonparametric Bayesian model that has been successfully applied to predict traffic characteristics (e.g., traffic congestion(Liu, Yue, and Krishnan 2013) and traffic volumes (Xie et al. 2010)) with state-of-the-art results. GBDT models are also well adopted for traffic prediction in various study, due to the ability to continuously reduce error during the execution of each iteration (Xia and Chen 2017; Yang et al. 2017; Yang, Zheng, and Sun 2019; Li and Bai 2016). SVR method showed great competence over other traditional algorithms while dealing with non-linear, high-dimensional and small sample problems (Li and Xu 2021). LR is a fundamental and popular algorithm in the field of traffic flow prediction, which can establish a statistical relationship between dependent variables and independent variables. (Alam, Farid, and Rossetti 2019; Boukerche and Wang 2020). The detailed advantages and disadvantaged of each model are summarized in Table S4.

In this study, in the process of model development, we used the annual daily average traffic profiles (i.e., traffic volume by category and speed) as the observed responses of the models, and the land-use datasets (detailed in Section 2.3.1) were used for the input of the models. We conducted an overall 10-fold cross-validation to evaluate the model performance. The entire dataset, including the hourly traffic datasets and the land-use datasets, was randomly split into 10 groups, with each group containing ~10% of the data. In each cross-validation, nine groups of data were employed as training sets to fit the models and make predictions on the remaining group. This process was repeated 10 times until every group was predicted. The Pearson R (r), root mean squared prediction error (RMSE) and mean absolute prediction error (MAPE) between the model predictions and observations were calculated to evaluate the model performance.

The validation results indicated that the LURF model is the most capable of estimating the traffic characteristics (see Section 3.1), which was used to further develop the hourly specific LURF models. The relative importance of the predictors from the trained LURF model shows their prediction ability. For each predictor, we permuted the values of this predictor across every observation in the dataset and measured the increase in the mean standard error (MSE) per permutation. Repeating this process for each predictor, a metric stores the increase in the MSE due to the permutation of out-of-bag observations across each input predictor averaged over all trees in the forest and divided by the standard deviation taken over the trees. The larger this value is, the more important the predictor should be. The metric stores the strength of the relationships between the predictors and projections to indicate the relative importance between various predictors.

3. Results and Discussion

3.1 The results of the traffic prediction

Table 1 illustrates the Pearson R used to evaluate the model performance for predicting the daily averaged traffic parameters based on the constructed machine learning models, the overall performance of prediction (Pearson R, RMSE and MAPE) were summarized in Table S5. The performances were assessed using a 10-fold cross-validation. First, the LURF models consistently derive higher correlation coefficients between the predicted and observed traffic profiles than almost all the other models. The Pearson's r values range from 0.62 (LDTs) to 0.79 (LMDPVs) for the LURF models, which are higher than the corresponding correlation coefficients for the other four models. The R values of GBDT for LMDPV are slightly better than LURF (0.81 vs. 0.79). Furthermore, the RMSE values of the category-resolved traffic volumes by using the LURF models are significantly lower than those for the other models. The RMSE values of GBDT and SVR on HDT traffic simulations are also relatively low. For the value of MAPE, the performance of LURF, GBDT and SVR are comparable and better than GPR and LR. In general, LURF has the best performance in the simulation of all traffic profiles, while the performance of GBDT and SVR on individual indicators (such as traffic flow of LMDPV and HDT) is also acceptable. Therefore, the LURF model is more capable of predicting traffic profiles than other models on most occasions for this study and will be used in following simulations.

We further developed an hourly specific LURF model to predict the hourly averaged traffic profiles under different scenarios. We ranked the variable importance in the training process utilizing the method above to construct a variable importance measure and then averaged the hourly ranks for each variable to the final importance index listed in Table S6. Table S6 illustrates the top 10 important variables for all land-use predictors in our LURF models under *S1*. In general, in addition to the road location (province, city and county), the important input variables for predicting traffic volumes are primarily related to road information, such as the road type, road density, number of lanes and designed speed. Especially for the HDTs and speed simulations, the top 10 important variables are almost all related to road information. The effects of land cover and POI information for estimating traffic characteristics are relatively less important than the role of road information. For passenger fleets (LMDPVs and HDPVs) and light-duty trucks (LDTs and MDTs), some important inputs related to the land-use datasets (e.g., population, POI information) are indicated by the analysis. Notably, these important land-use variables often represent large buffers (e.g., 2000 m to 5000 m). This effect is because intercity monitoring sites are typically located far away from urban areas, where POI information is more densely available.

3.2 Traffic activity characteristics of road networks under various scenarios

Different from urban weekday-weekend patterns (Yang et al. 2019), a higher traffic activity is estimated during weekends (*S2*; 931 million veh km) than on weekdays (*S1;* 841 million km). This data implies that more leisure trips during weekends could be captured by intercity highway monitoring data (see Fig. S5). The lower traffic activity during weekdays is mainly observed in Hebei Province (63%); among all vehicle categories, the LDPVs account for ~70% of the total reduced VKT, followed by HDTs (16%) and LDTs (10%).

We annualized the allocation of the total traffic activity by vehicle category in the BTH region by aggregating the daily patterns under *S1* and *S2* (see Fig. S6). Among all the fleets, the LDPVs account for the largest proportion of the total annual traffic activity (64%), followed by the HDTs (18%) and LDTs (9%). For the HDTs, the fractions of the total traffic activity decrease from Hebei (20%) to Tianjin (18%) and thus to Beijing (11%), probably because of two major reasons. First, the passenger traffic activity by LDPVs is denser in Beijing than in Hebei. Second, many HDTs are strictly limited within the Sixth Ring Road in Beijing, which could also shrink the freight activity outside the restriction area (Yang et al. 2019). Table S7 illustrates the allocations of traffic activity by the vehicle category and road type. For freight transportation, more than 50% of the HDT traffic activity is estimated to occur on expressways. In Tianjin and Hebei, HDTs travel more frequently on expressways than LDTs to improve efficiency.

In terms of temporal variations, the rush hour phenomenon, shown as an increase in the traffic activity, occurred from 8:00 to 10:00 GMT+8 and from 15:00 to 17:00 GMT+8 on weekdays (*S1*). Compared to the rush hour effect within urban areas in Beijing (e.g., peaked at the hour of 7:00 GMT+8)(Yang et al. 2019), the morning rush hours occur later, while evening rush hours are earlier for these intercity highways. Similar trends are observed during weekends (*S2*), with a higher traffic activity reflecting more casual intercity trips (see Fig. S5). In contrast, the diurnal fluctuations of the average speeds depict quite close characteristics between weekdays and weekends (see Fig. S7) because the level of congestion for intercity highways (even during rush hours) is not comparable to urban areas. It should be noted that the speed in Beijing are obviously lower than that in Tianjin and Hebei, mainly due to the congestion occurred in Beijing resulting from its larger travel demands.

The traffic activity of Scenario *S3* with special policy interventions was clearly reduced in the BTH region. The daily traffic activity of *S3* was decreased by 23% compared with that of normal weekdays. However, traffic reductions are heterogeneous among the various vehicle categories and different regions. As Fig. 1 shows, the LDPVs show uniform reductions of approximately 25% on all the intercity highways in the BTH region; only 4% of roads are identified with increased LDPV volumes. For HDTs, reduced traffic could also be observed in Beijing and the expressways and national highways in Tianjin, but there is a certain number of provincial highways in Tianjin (~20%) and expressways (~30%), national highways (~15%) and provincial highways (~20%) in Hebei with increased flow. Four subregions are indicated in Fig. 1 with significantly increased HDT volumes (more than 50%). Subregions 1-3 represent the roads heading to several large ports (the ports of Qingdao, Tianjin, and Qinhuangdao for subregions 1-3, respectively), and subregion 4 represents the areas near the boundary of the BTH region. These differences in traffic volume between the LDPVs and HDTs indicate that for passenger travel, the restrictions could uniformly reduce the travel demand across the region. In contrast, decreases in HDT traffic volumes are expected in areas with the stricter enforcement of traffic restrictions (i.e., Beijing and highway and national roads in Tianjin). The traffic restrictions were more effective in Beijing, resulting in a 29% decrease among all vehicle fleets, especially for the HDTs (a 52% decrease) and MDTs (a 42% decrease). In Hebei, such traffic restrictions could result in a detour of the HDTs, as the operators and drivers of the HDTs conduct their business even on restrictive days. The decrease in the total traffic activity in Hebei was primarily due to the LMDPVs, and the traffic activities of MDTs and HDTs were only 5% and 14% lower than those on normal weekdays (*S1*).

3.3 Temporal and spatial characteristics of traffic emissions

The total daily emissions for the intercity highways in the BTH region are estimated to be 1443 tons for CO, 152.3 tons for THC, 1158 tons for NO_X, 37.30 tons for $PM_{2.5}$ and 18.73 tons for BC, during weekdays (*S1*, See Fig. 2; see provincial-level total emissions and emission intensity in Fig. S8. During weekends (*S2*), the total daily emissions are estimated to increase by approximately 10% for all pollutants due to increased traffic activity. Comprehensive traffic restrictions under *S3* triggered decreased vehicle emissions, e.g., 33% to 38% for CO and THC, 23% for NO_X, and 15% to 17% for $PM_{2.5}$ and BC in the entire domain relative to *S1*. For LDPVs traveling on the highways, the greater reductions in CO and THC result from the lower traffic volumes due to traffic restrictions, especially in Tianjin and Hebei. In these areas, the controls for vehicles are behind those in Beijing, and therefore, the restrictions on pre-China 2 gasoline cars could result in a larger reduction. Diesel trucks contribute significantly to the emission reductions of NO_X, $PM_{2.5}$ and BC, and their lower reduction percentages compared to CO and THC are related to smaller decreases in the HDT traffic volume in Tianjin and Hebei.

The major temporal difference in the diurnal patterns between *S1* and *S2* lies in the higher emissions during the rush hours of weekends. We thus refer to the weekday scenario (*S1*) to elucidate the temporal and spatial emission patterns (see Figs. 2 and 3). The emission peaks of CO and THC during morning (9:00 to 10:00 GMT+8) and evening (16:00 to 17:00 GMT+8) rush hours, which are apparently associated with diurnal fluctuations in the passenger travel demand, are shown in Fig. S5. As Figs. 4a and 4b visualize, the CO emission intensity close to urban areas is significantly higher than that in outlying areas during both peak and off-peak periods. The highest hourly emissions of CO and THC, which are more associated with passenger traffic activity, were estimated during the morning rush hour (10:00 GMT+8); these emissions were higher than their 24-h averages by approximately 40%-50%. The emission allocations show high resemblance between CO and THC; specifically, the LDPVs dominate the contributions, and the proportions in Beijing and Tianjin are higher than those in Hebei. This increase is because these two metropolitan areas have a higher density of residential population (indicated by *pop_5000m*), business units (indicated by *POI_office_5000m*), and urban lands (indicated by *urbanland_5000m*) than in Hebei; these variables have been identified as the most important variables for traffic activity predictions of LDPVs by using LURF modeling (see Table S6).

Diesel fleets are responsible for much greater shares of on-road NO_X, $PM_{2.5}$ and BC emissions than CO and THC emissions. Consequently, distinctive traffic behaviors of these diesel fleets will result in disparate temporal and spatial emission patterns from those for CO and THC. First, we have observed that the total emissions of NO_X, $PM_{2.5}$ and BC during the night (0:00 to 4:00 GMT+8, Fig. 2) are closer to the emissions during the daytime, but the nighttime-daytime differences in the emission patterns are less than those of CO and THC. This finding is because a considerable part of long-haul freight trucks in China are operated by two drivers, who could shift duty and travel during nighttime. Second, the NO_X, $PM_{2.5}$ and BC emission contributions by HDTs in Tianjin and Hebei are higher than those in Beijing by approximately 10%. Additionally, the higher percentages of the total NO_X, $PM_{2.5}$ and BC emissions by HDPVs (tourist and intercity coaches) in Beijing are higher than those in Tianjin and Hebei. The comparison results indicated higher passenger travel demand in Beijing due to its attraction of tourists and lower freight transportation activity due to truck restrictions.

The emission maps discern Tianjin Port, the largest port in northern China with an annual freight handling amount of 466 million tons in 2018, as a significant hotspot. A large amount of HDTs flood into Tianjin Port (i.e., Subregion 2 in Fig. 1), leading to significantly higher emissions on adjacent highways throughout the day (Figs. 3C and 3D). The daily variation in the NO_X emission intensity in the port area is more obvious than that in Tianjin and the BTH region, with a peak period from 7:00 to 18:00 GMT+8. The hourly average NO_X emission intensity of Tianjin Port is 1.87±0.42 kg km^{-1} h^{-1}, which is 47% and 123% higher than the average levels of Tianjin and the BTH region, respectively (see Fig. S9).

3.4 Discussion

An efficient protocol of dynamically modeling hourly based, link-level emissions. This study provides a universal analytical method for a high-resolution vehicle inventory at a regional scale, especially for regions including many cities suffering from the difficulty of addressing traffic data at a high spatial resolution. Fig. 4 shows the hourly variations in the traffic activities of LDPVs and HDTs and total vehicle emissions by region during a special week when traffic restrictions were implemented (November 5th to 6th, 2017). During November 2nd to 5th, the traffic activity resembles that during a normal week (e.g., April 20th to 27th, 2017; see Fig. S10). When the traffic restrictions were being implemented during November 5th and 6th, we observed an ~20% reduction in the VKT of LDPVs and HDTs, an ~30% reduction in CO emissions and an ~20% reduction in NO_X emissions, which resembled *S1* and *S3* because of traffic restrictions. Therefore, the high efficiency of the calculation based on the LURF model provides a platform to realize the near real-time process of establishing high-resolution vehicle emission inventories, and can dynamically support the further evaluation of environmental benefits from traffic policies and management measures. However, the spatiotemporal dependencies are not as clearly modeled in machine learning as they are in general linear model frameworks, and future work could derive methods for optimizing the predictors used in machine learning models to improve the accuracy of the prediction.

Comparison of the best machine learning method against the standard empirical bottom-up approach. Addressing the spatial characteristics of traffic profiles based on limited datasets is a significant challenge for establishing high-resolution emission inventories of on-road vehicles. To overcome this barrier, the allocation of the VKT based on-road information was the most typical way to establish simplified bottom-up inventories in previous studies. Zheng et al. allocated the VKT of each county based on the same weighing factors considering the vehicle category and road type. Gately et al. used the same allocation of the total VKT considering the differences in weighing factors according to the road type but regardless of distinguishing the vehicle category. This section compared the differences between the above empirical methods and machine learning method obtained in this study, M1 denotes the best machine learning method (i.e., LURF) in this study and M2 denotes the allocation method based on the standard empirical bottom-up approach. The VKT of M2 was allocated based on the combination of the methods from Zheng and Gately (according to Eq. 2).

$$\mathrm{VKT}_{l,vc}^{M2} = \mathrm{VKT}_{vc}^{M1} \times \frac{\mathrm{TV}_{l,vc} \cdot RL_{l,vc}}{\sum_l \mathrm{TV}_{l,vc} \cdot RL_{l,vc}} \quad (2)$$

where VKT_{vc}^{M1} is the total VKT of vehicle category *vc* of the BTH region calculated in M1; $\mathrm{VKT}_{l,\,vc}^{M2}$ is the allocation VKT on-road link *l* of vehicle category *vc* in M2; $\mathrm{TV}_{l,vc}$ is the preallocated traffic volume based on its road type and location on-road (Yang et al. 2019) link *l* of vehicle category *vc*; and $\mathrm{RL}_{l,vc}$ is the length of road link *l*. CO and $\mathrm{NO_X}$ are discussed, as they represent the gasoline and diesel featured emissions, respectively.

As Fig. 5 illustrates, compared with M1, M2 tends to underestimate the CO emissions on the provincial highways close to urban areas (80% of the provincial highway links are underestimated) and overestimates the expressways and national highways in remote rural areas, especially in Beijing and Hebei (77% and 62% of the expressways are overestimated in Beijing and Hebei, respectively). This miscalculation is because, for remote rural areas without monitoring traffic data, we tend to preallocate volumes according to the road rank, which means expressways and national highways will be allocated more traffic volumes than provincial roads. For $\mathrm{NO_X}$, we observe a similar misestimation in the emission distribution of CO; that is, approximately 70% of provincial highway links are underestimated, and ~70% and 60% of expressways and national highways are overestimated. The long-tailed distributions of the relative difference of CO and $\mathrm{NO_X}$ are illustrated in Fig. S11. Overall, the differences between the two methods are not extreme because 79% and 82% of road links' relative differences for CO and $\mathrm{NO_X}$ are within ±50%. The analysis indicates that we may use the simplified M2 with the absence of land-use data while needing to pay attention to the uncertainty in the project-level emission calculations (e.g., port-related areas).

Temporal and spatial patterns of air pollutant emissions from on-road vehicles are of substantial interest because of the associated potential public health impact. The population exposure to vehicular pollutants are greatest in areas with high vehicle usage and population density (Marshall et al. 2005). Air quality simulation models with fine-grained input from high-resolution vehicle emission inventories will be valuable for evaluating the potential health benefits from vehicle emission control measures (Ke et al. 2017). However, we are limited to estimates of detailed link-level emissions for urban areas in Tianjin and Hebei due to the data availability (e.g., traffic volume of urban local roads). Currently, as noted in the introduction, there are already link-level emission inventories in large cities with traffic datasets from the ITS, but a strengthening of the portability of the analytical method is still needed. This study points a promising way to smart management of traffic emissions in various regions and cities by combining advanced data-driven techniques and multi-source ITS datasets, which will provide policymakers with a better understanding of how air quality impacts regional and local transportation activities.

4. Conclusion

This paper developed an hourly-based, link-level emissions inventory of vehicular pollutants using land-use machine learning methods based on the datasets of road traffic monitoring in the BTH region of China under three traffic scenarios. The methodology can potentially be used either to map traffic characteristics on a larger scale or to deal with real-time urban traffic data streams that need to overcome the challenges of computational accuracy and efficiency. The major findings can be summarized as follows:

(1) The land-use random forest (LURF) model is more capable of predicting traffic profiles than other models on most occasions in this research.

(2) The important input variables for predicting traffic volumes are primarily related to road information, such as the road type, road density, number of lanes and designed speed.

(3) A higher traffic activity is estimated during weekends than on weekdays, and the traffic activity of *Scenario S3* with special policy interventions was deceased by 23% compared with that of normal weekdays.

(4) The model finds a general reduction in Light-Duty Passenger Vehicles when traffic restrictions were implemented, but a much more spatially heterogeneous impact on Heavy-Duty Trucks (HDTs).

Acknowledgments

This work is supported by Ministry of Science and Technology of China's International Science and Technology Cooperation Program (2018YFE0106800, MOST of China and European Union Horizon 2020 Collaborative Project), the National Natural Science Foundation of China (52170111 and 41977180), and sponsored by Tsinghua-Toyota Joint Research Institute Cross-discipline Program. K. Max Zhang would like to acknowledge the support from the Cornell China Center and the National Science Foundation (NSF) through grant No. 1605407.

Code/Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author contribution

Y. W and S. Z. conceived the research idea; X. W., D. Y., R. Wu, J. G, and Y. Wen contributed to new analytic tools; D. Y., R. Wu, R. Wang, and H. X. prepared the data; X. W., S.Z., K.M.Z. and Y. Wu provided valuable discussions on modeling development and paper organization; X. W., D. Y., Y. W, K. M. Z. and S.Z. wrote the paper with contributions from all authors.

Competing interests

The authors declare that they have no conflict of interest.

References

Alam, Ishteaque, Dewan Md. Farid, and Rosaldo J. F. Rossetti. 2019. "The Prediction of Traffic Flow with Regression Analysis." In, 661-71. Singapore: Springer Singapore.

Boukerche, Azzedine, and Jiahao Wang. 2020. 'Machine Learning-based traffic prediction models for Intelligent Transportation Systems', *Computer Networks*, 181: 107530.

Brokamp, Cole, Roman Jandarov, Monir Hossain, and Patrick Ryan. 2018. 'Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model', *Environmental Science & Technology*, 52: 4173-79.

Chapman, Lee. 2007. 'Transport and climate change: a review', *Journal of Transport Geography*, 15: 354-67.

Gaode Map. 2017. "Search for Point of Interests." In. lbs.amap.com: Alibaba Group.

Gately, C. K., L. R. Hutyra, S. Peterson, and I. S. Wing. 2017. 'Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone GPS data', *Environmental Pollution*, 229: 496-504.

Gately, C. K., L. R. Hutyra, and I. Sue Wing. 2015. 'Cities, traffic, and CO2: A multidecadal assessment of trends, drivers, and scaling relationships', *Proc Natl Acad Sci U S A*, 112: 4999-5004.

Gong, Peng, Han Liu, Meinan Zhang, Congcong Li, Jie Wang, Huabing Huang, Nicholas Clinton, Luyan Ji, Wenyu Li, Yuqi Bai, Bin Chen, Bing Xu, Zhiliang Zhu, Cui Yuan, Hoi Ping Suen, Jing Guo, Nan Xu, Weijia Li, Yuanyuan Zhao, Jun Yang, Chaoqing Yu, Xi Wang, Haohuan Fu, Le Yu, Iryna Dronova, Fengming Hui, Xiao Cheng, Xueli Shi, Fengjin Xiao, Qiufeng Liu, and Lianchun Song. 2019. 'Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017', *Science Bulletin*, 64: 370-73.

Hastie, Trevor. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.

Hoek, Gerard, Rob Beelen, Kees de Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. 2008. 'A review of land-use regression models to assess spatial variation of outdoor air pollution', *Atmospheric Environment*, 42: 7561-78.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*.

Ke, W., S. Zhang, Y. Wu, B. Zhao, S. Wang, and J. Hao. 2017. 'Assessing the Future Vehicle Fleet Electrification: The Impacts on Regional and Urban Air Quality', *Environmental Science & Technology*, 51: 1007-16.

Lee, Martha, Michael Brauer, Paulina Wong, Robert Tang, Tsz Him Tsui, Crystal Choi, Wei Cheng, Poh-Chin Lai, Linwei Tian, Thuan-Quoc Thach, Ryan Allen, and Benjamin Barratt. 2017. 'Land use regression modelling of air pollution in high density high rise cities: A case study in Hong Kong', *Science of The Total Environment*, 592: 306-15.

Li, Cong, and Pei Xu. 2021. 'Application on traffic flow prediction of machine learning in intelligent transportation', *Neural Computing and Applications*, 33: 613-24.

Li, X., and R. Bai. 2016. "Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree." In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1010-15.

Liaw, Andy, and Matthew Wiener. 2001. 'Classification and Regression by RandomForest', *Forest*, 23.

Liu, Siyuan, Yisong Yue, and Ramayya Krishnan. 2013. *Adaptive Collective Routing Using Gaussian Process Dynamic Congestion Models*.

Marshall, Julian D., Thomas E. McKone, Elizabeth Deakin, and William W. Nazaroff. 2005. 'Inhalation of motor vehicle emissions: effects of urban population and land area', *Atmospheric Environment*, 39: 283-95.

McDonald Brian C., McBride Zoe C., Martin Elliot W., and Harley Robert A. 2014. 'High ‐ resolution mapping of motor vehicle carbon dioxide emissions', *Journal of Geophysical Research: Atmospheres*, 119: 5283-98.

Sun, Daniel, Kaisheng Zhang, and Suwan Shen. 2018. 'Analyzing spatiotemporal traffic line source emissions based on massive didi online car-hailing service data', *Transportation Research Part D: Transport and Environment*, 62: 699-714.

Uherek, Elmar, Tomas Halenka, Jens Borken-Kleefeld, Yves Balkanski, Terje Berntsen, Carlos Borrego, Michael Gauss, Peter Hoor, Katarzyna Juda-Rezler, Jos Lelieveld, Dimitrios Melas, Kristin Rypdal, and Stephan Schmid. 2010. 'Transport impacts on atmosphere and climate: Land transport', *Atmospheric Environment*, 44: 4772-816.

Waddell, Paul. 2002. 'UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning', *Journal of the American Planning Association*, 68: 297-314.

Wang, Xing, Dane Westerdahl, Ye Wu, Xiaochuan Pan, and K. Max Zhang. 2011. 'On-road emission factor distributions of individual diesel vehicles in and around Beijing, China', *Atmospheric Environment*, 45: 503-13.

WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton). 2015. "China 100m Population." In.

Wu, X., Y. Wu, S. Zhang, H. Liu, L. Fu, and J. Hao. 2016. 'Assessment of vehicle emission programs in China during 1998-2013: Achievement, challenges and implications', *Environ Pollut*, 214: 556-67.

Wu, Y., S. Zhang, J. Hao, H. Liu, X. Wu, J. Hu, M. P. Walsh, T. J. Wallington, K. M. Zhang, and S. Stevanovic. 2017. 'On-road vehicle emissions and their control in China: A review and outlook', *Science of the Total Environment*, 574: 332-49.

Xia, Ying, and Jungang Chen. 2017. "Traffic Flow Forecasting Method based on Gradient Boosting Decision Tree." In *Proceedings of the 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*, 413-16. Atlantis Press.

Xie, Yuanchang, Kaiguang Zhao, Ying Sun, and Dawei Chen. 2010. 'Gaussian Processes for Short-Term Traffic Volume Forecasting', *Transportation Research Record*, 2165: 69-78.

Yang, D., S. Zhang, T. Niu, Y. Wang, H. Xu, K. M. Zhang, and Y. Wu. 2019. 'High-resolution mapping of vehicle emissions of atmospheric pollutants based on large-scale, real-world traffic datasets', *Atmos. Chem. Phys.*, 19: 8831-43.

Yang, Jie, Linjiang Zheng, and Dihua Sun. 2019. "Urban Traffic Flow Prediction Using a Gradient-Boosted Method Considering Dynamic Spatio-Temporal Correlations." In *Knowledge Science, Engineering and Management*, edited by Christos Douligeris, Dimitris Karagiannis and Dimitris Apostolou, 271-83. Cham: Springer International Publishing.

Yang, Senyan, Jianping Wu, Yiman Du, Yingqi He, and Xu Chen. 2017. 'Ensemble Learning for Short-Term Traffic Prediction Based on Gradient Boosting Machine', *Journal of Sensors*, 2017: 7074143.

Zhang, S. J., T. L. Niu, Y. Wu, K. M. Zhang, T. J. Wallington, Q. Y. Xie, X. M. Wu, and H. L. Xu. 2018. 'Fine-grained vehicle emission management using intelligent transportation system data', *Environmental Pollution*, 241: 1027-37.

Zhang, Shaojun, Ye Wu, Xiaomeng Wu, Mengliang Li, Yunshan Ge, Bin Liang, Yueyun Xu, Yu Zhou, Huan Liu, Lixin Fu, and Jiming Hao. 2014. 'Historic and future trends of vehicle emissions in Beijing, 1998–2020: A policy assessment for the most stringent vehicle emission control program in China', *Atmospheric Environment*, 89: 216-29.

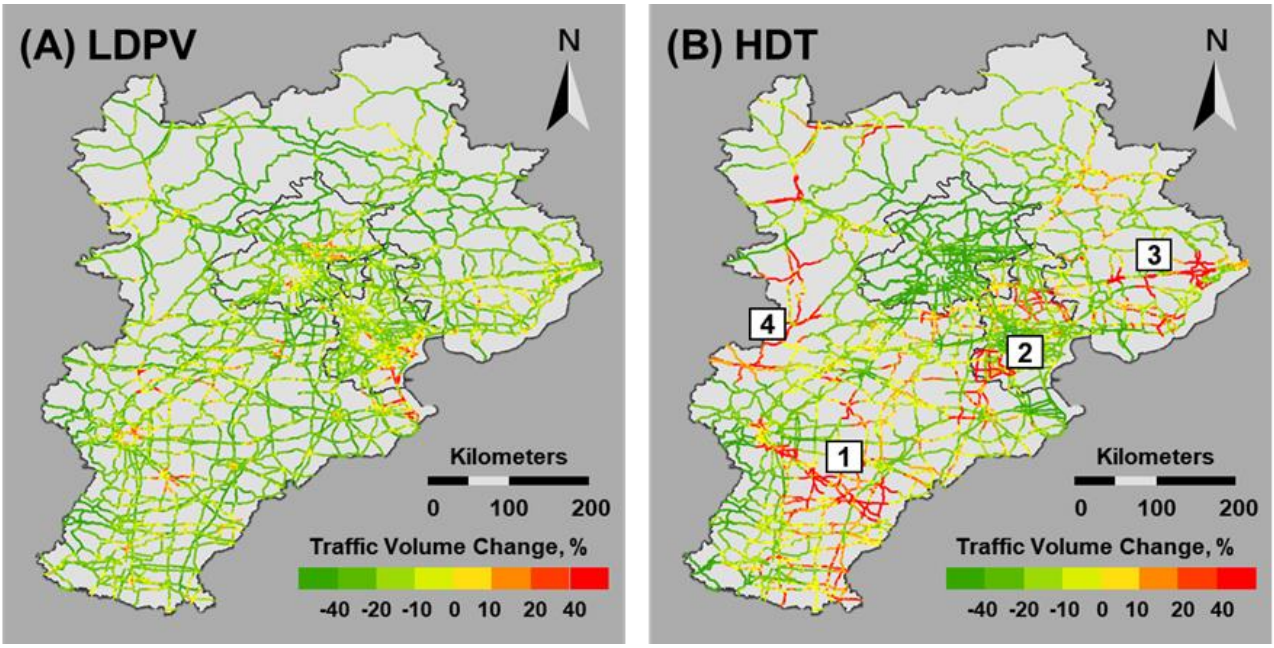

Figure 1. Estimated traffic volume reductions of the LDPVs and HDTs under traffic restrictions (*S3)* compared to normal weekdays (*S1)*.

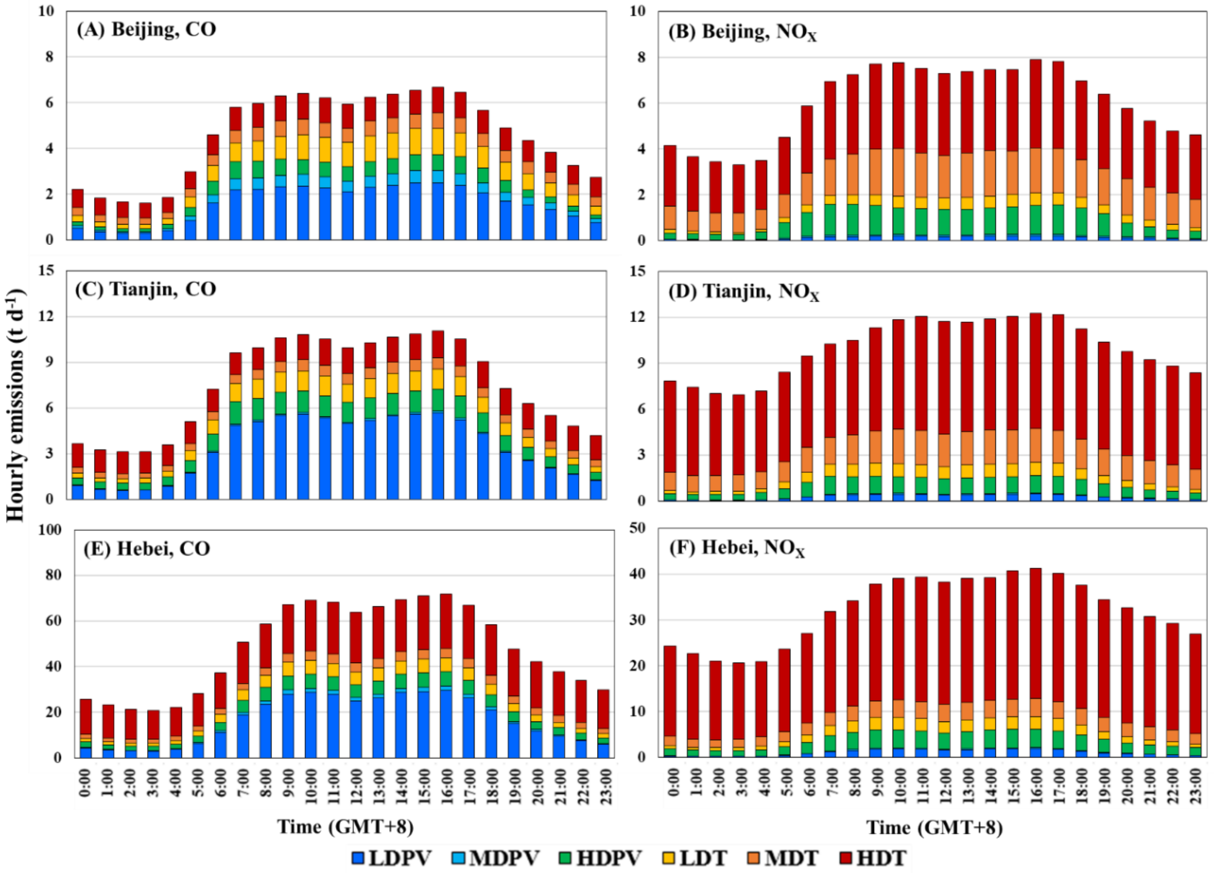

Figure 2. Estimated hourly emissions by vehicle category under *S1*.

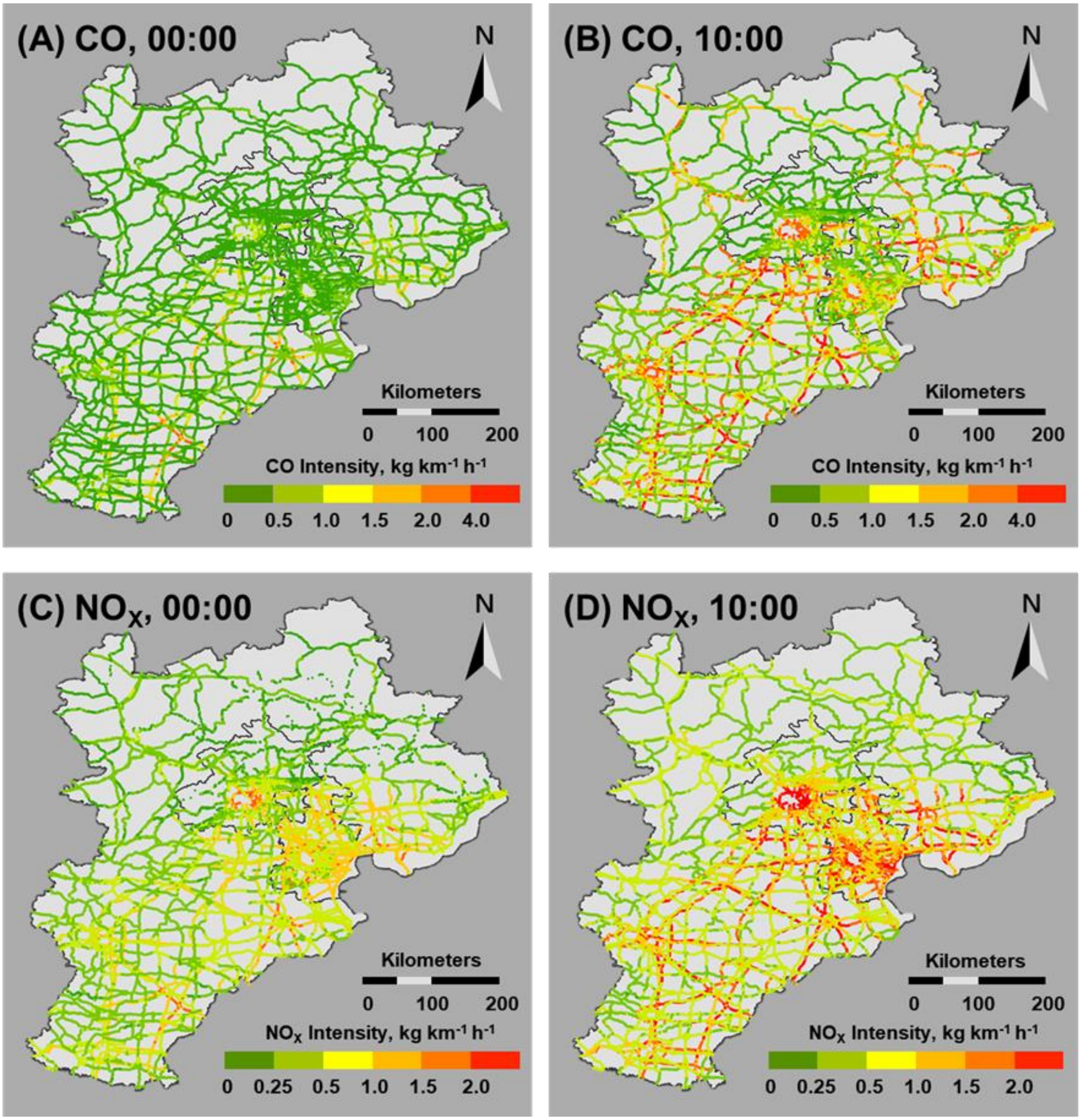

Figure 3. Link-based emission intensity of CO (panels A and B) and NO_X (panels C and D) during a midnight hour (0:00 GMT+8) and a rush hour (10:00 GMT+8).

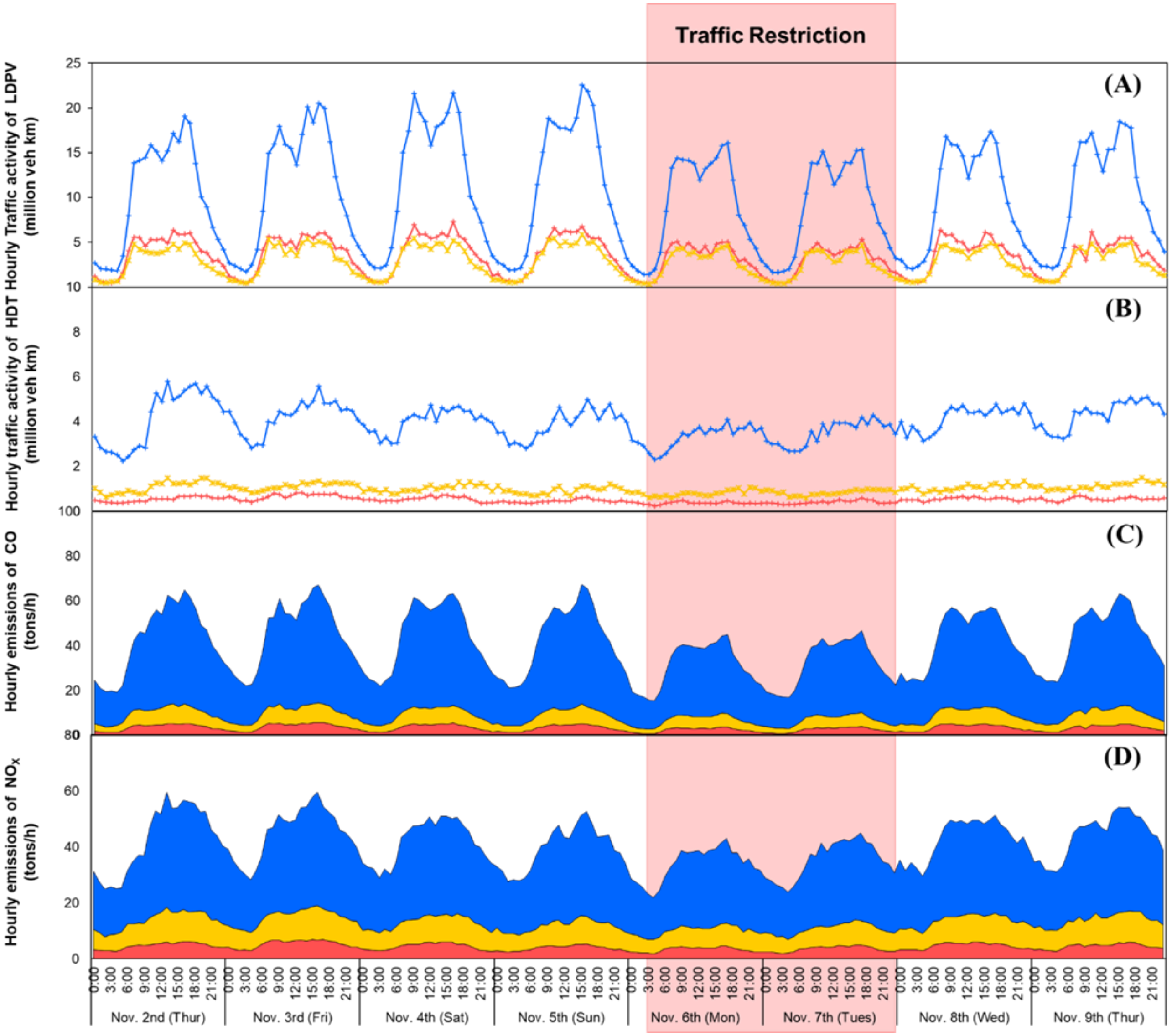

Figure 4. Hourly VKT of LDPV and HDT (A and B) and vehicle emissions of CO and BC (C and D) by region from November 2nd to November 9th, 2017.

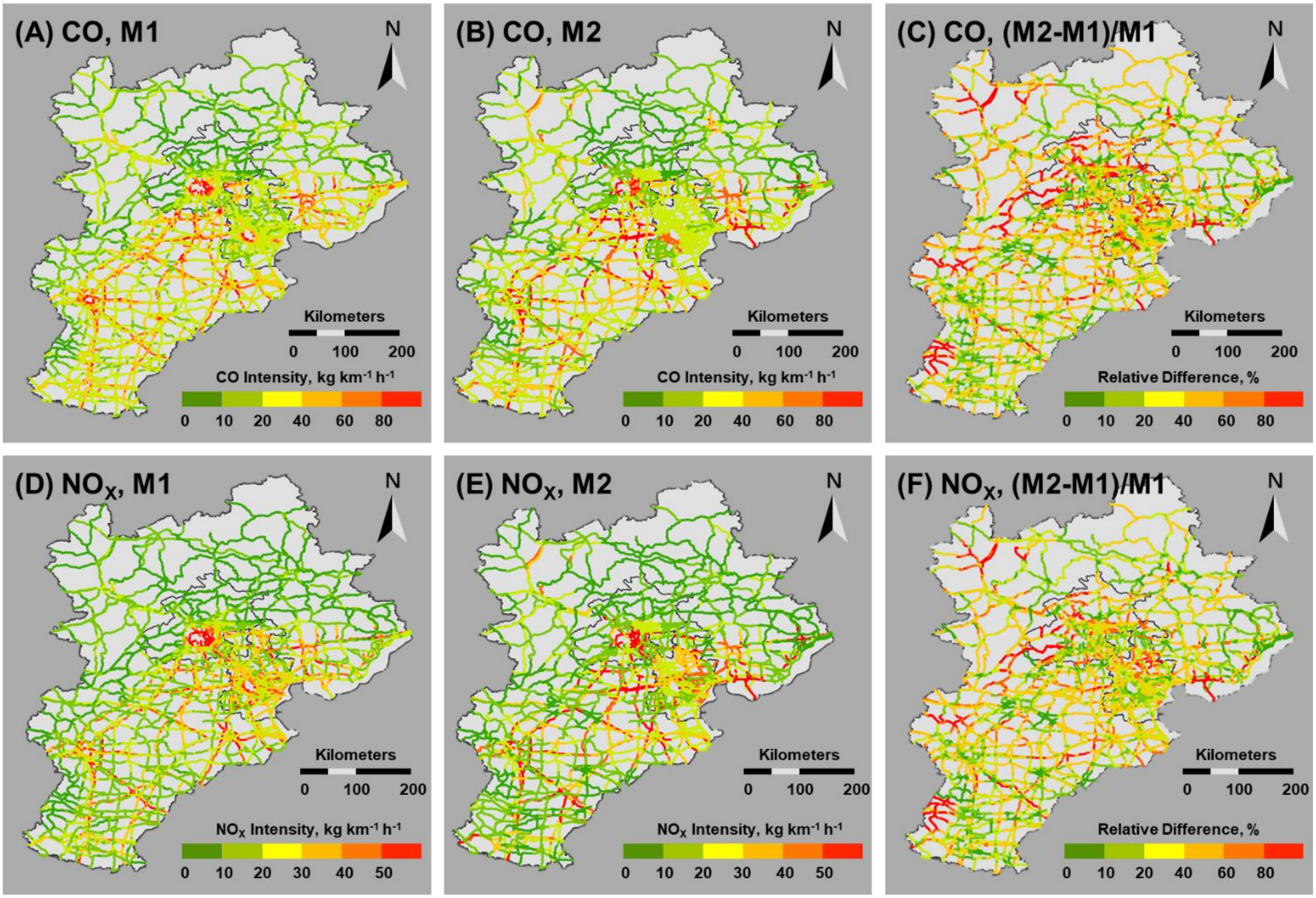

Figure 5. Comparison of the LURF method against the standard empirical bottom-up approach.

Note: M1 denotes the emission inventory based on the best machine learning method (LURF) in this study. M2 denotes the emission inventory based on standard empirical bottom-up approach.

Table 1. Cross-validated mean Pearson R of the traffic prediction by selected machine learning models.

Traffic profiles	**LURF**	**GBDT**	**SVR**	**GPR**	**LR**
LMDPV	0.79	0.81	0.65	0.62	0.48
HDPV	0.61	0.54	0.51	0.46	0.3
LDT	0.62	0.55	0.44	0.49	0.17
MDT	0.64	0.6	0.48	0.47	0.26
HDT	0.65	0.58	0.56	0.58	0.5
Speed	0.75	0.74	0.7	0.71	0.55