# Clearer skies in Beijing – revealing the impacts of traffic on the modeling of air quality

Yanyan Xu*
Department of Civil & Environmental Engineering, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: 857-285-0982; Email: yanyanxu@mit.edu

Ruiqi Li
School of Systems Science, Beijing Normal University
No. 19 Xinjiekouwai St., Haidian District, Beijing 100875, China
Department of Civil & Environmental Engineering, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: 857-207-5631; Email: liruiqi@mit.edu

Shan Jiang
Department of Civil & Environmental Engineering, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: 857-654-5066; Email: shanjang@mit.edu

Jiang Zhang
School of Systems Science, Beijing Normal University
No. 19 Xinjiekouwai St., Haidian District, Beijing 100875, China
Tel: 011-86-10-58802732; Email: zhangjiang@bnu.edu.cn

Marta C. González
Department of Civil & Environmental Engineering, MIT
Center for Advanced Urbanism, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: 617-715-4140; Email: martag@mit.edu

* Corresponding author

5339 words + 8 figures × 250 words + 0 tables = 7339 words
November 16, 2016

**ABSTRACT**

Urban air pollution imposes major environmental and health risks worldwide, and is expected to become worse in the coming decades as cities expand. Detailed monitoring of urban air quality at high spatial and temporal resolution can help to assess the negative impacts as a first step towards mitigation. Improvement of air quality needs a variety of measures working together, including controlling industrial pollution and mitigating automobile emissions. In contrast to the measurable industrial pollution, in many of the developing countries, the impact and control of automobile emissions on air quality is neither well understood nor well established. Moreover, the automobile emission data sets are difficult to collect. In this paper, we present a data analysis framework to uncover the impact of urban traffic on estimating air quality in different locations within a metropolitan area. To that end, we estimate the traffic surrounding 24 air quality (AQ) monitoring stations in Beijing, combining mobile phone data and road networks with a traffic assignment model. We investigate how the amount of traffic surrounding each station can impact the modeling of air quality index (AQI) observed by the stations. We separately estimate the contribution of traffic information to the modeling of AQI with regression models in the summer and winter. Further, we group the AQ monitoring stations into four classes, and show that in the summer, air pollution in the inner city is generally more severe than that in the suburbs due to urban traffic; while in the winter, air pollution in the south of Beijing surpasses that in the inner city, most likely due to heating using coal.

## INTRODUCTION

With the rapid urbanization and the acceleration of industrialization, today's air pollution has become a global threat of human health, especially for the large scale and densely populated cities in developing countries (*1, 2*). As pointed by the World Healthy Organization (WHO), in 2012 around 3.6 million people died – 16% of total global deaths – as a result of ambient air pollution exposure, which makes it the largest environmental risk to the health of human beings. Moreover, exposure to air pollutants is largely beyond the control of individuals and requires action by public authorities at the national, regional and even international levels.

It is important to detect pollutants in the air, to explore their sources, and to model their temporal and spatial patterns, in order to make policy recommendations to mitigate their negative impacts. To better predict air quality (AQ), the relationship between the sources and AQ needs to be examined and clarified. The sources of air pollution are usually divided in 4 categories: stationary, such as industries; mobile, such as transportation sources; area, such as agricultural areas, cities, and wood burning fireplaces; and natural, such as dust and wildfires. The first two of them are human related factors and represent research priorities in the literature. Mobile sources include motor vehicles, marine vessels, and air-crafts. Among them, the exhaust emission of motor vehicles is one of the primary factors that influence AQ in urban areas (*3, 4*). Consequently, clear impacts between traffic and AQ may inform environmental policies. To examine the impact of traffic on air pollution, McHugh *et al.* updated an atmospheric dispersion modeling system with a traffic emissions database (*5*). Several studies measured the impacts of traffic and meteorology on air pollution measuring data near roads (*6, 7*). While these studies are detailed on the chemical processes, they do not cover the entire city. Using a data analysis perspective, Zheng *et al.* studied the variations of air quality in space and time in the entire Beijing region via machine learning techniques, combining multiple data sources including taxi data, number of facilities, and the road network data (*8*). In a related work (*9*), they predicted air quality in each station, informed by historical AQ and meteorological data, and weather forecasts without considering traffic conditions.

We focus our study in Beijing, which is one of the most congested and polluted cities in China. Improving AQ in Beijing, is a top-priority locally, that has attracted the world's attention in the past few years. These efforts are compromised by the rapid growth of motorization and urbanization (*4*). Fig. 1 shows the noise-removed values of air quality index (AQI), wind speed, and humidity from April 1, 2014 to May 1, 2015 in Beijing. The figure represents the variation of AQI at 24 AQ monitoring stations within the Sixth Ring Road of Beijing. Higher AQI values indicate worse air quality. Specifically, AQI values in the range of 0-50 are established as good air, 51-100 moderate, 101-150 unhealthy for sensitive groups, 151-200 unhealthy, 201-300 very unhealthy, and 301-500 hazardous. We see that in general the AQIs in Beijing in the observation period are from moderate to unhealthy. However, they are more stable and lower in the summer (May to October) than in the winter (November to March). Also the wind speed and humidity show different patterns in the two seasons. In the present work, we seek to uncover the contributions of traffic to the air pollution modeling in the summer and winter separately.

Our work contributes to the types of studies presented in Refs. (*8, 9*) in three major aspects. First, we establish separate models—one for the winter and one for the summer—to gain better understanding of seasonal effects of AQI. Second, to investigate the different spatial impacts of urban traffic on AQI, we separately model their relationship by station, taking into account a set of online publicly available daily traffic congestion index (TCI) reported by the local transportation committee to reflect realistic daily traffic conditions. Finally, we enrich the on-line TCI with a

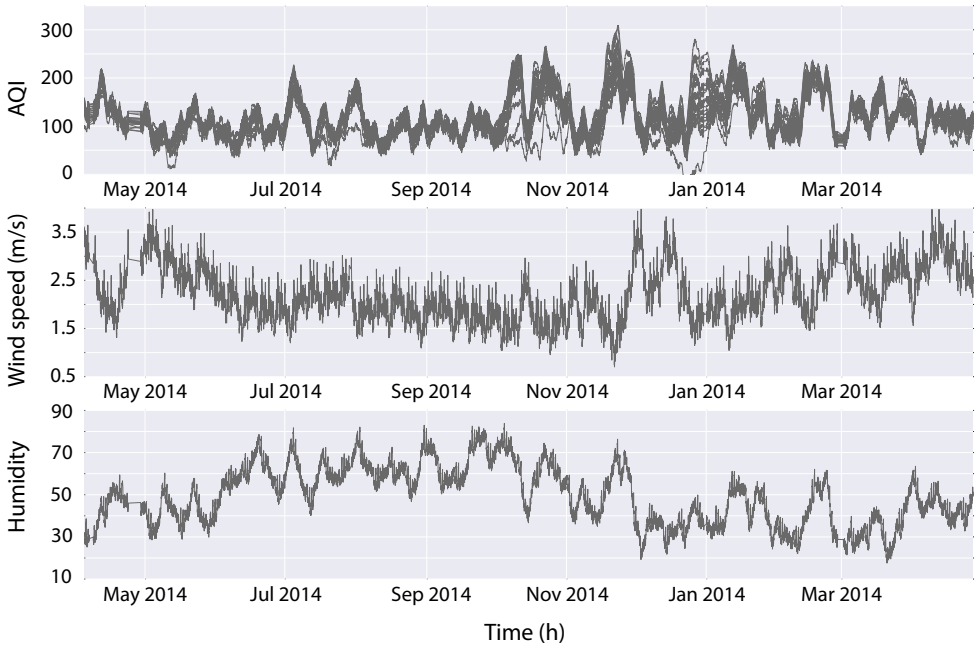


**FIGURE 1 Variations of AQI, wind speed, and humidity from April 1, 2014 to May 1, 2015 in each of the 24 air quality monitoring stations within the Sixth Ring Road of Beijing.**

travel demand model. We calculate the collective travel time (CTT) of all vehicles surrounding the AQ monitoring station, which is estimated from a mobile phone data based travel demand model and traffic assignment model integrated with the TCI. Relating traffic with the actual number of drivers and their origins and destinations is crucial to mitigate congestion in the urban road network, which can take into account AQ impact.

In the next sections, first, we discuss the mobile phone meta data and results from the call detail records (CDR) to inform a travel demand model. Second, we analyze the importance of traffic information to the prediction of AQI and the diversity of AQ in space per season. Concluding remarks and directions for future work are given in the last section.

## DATA AND METHODOLOGY

### Travel demand estimation from mobile phone data

We estimate the travel demand for the 19.4 million residents living in the urban area of Beijing. This is commonly referred to the region within the Sixth Ring Road, shown in Fig. 2a, which has 5.6 million privately-owned vehicles registered in 2013 (*10*). To our knowledge, our work constitutes the first traffic estimates of the region based on mobile phone data for Beijing.

Alexander *et al.* and Colak *et al.* outlined a general framework to obtain Origin-Destination (OD) matrices from massive mobile phone data (*11, 12*). We apply the same methods to extract trips of users, and estimate the person and vehicle travel demand by combining them with census data within the Sixth Ring Road of Beijing. Fig. 2a shows the map of Beijing with the AQ stations marked by blue circles. We focus our study in the inner area marked in darker green.

First, we extract stay locations of massive anonymous users from raw mobile phone data, and labeling activities with *home*, *work* and *other*. Second, we infer number of trips among the

stay locations of users by different time of the day and by purpose. Combining with census data, we expand mobile phone users to total population, and estimate an OD matrix for an average day. Next (an innovative step proposed by this study), we generate a series of day-specific OD matrices by using local reported daily traffic congestion index for the city, which allows us to fluctuate the average daily OD to reflect the realistic daily traffic conditions. We then assign the daily vehicle demand to the road network.

*Mobile phone data*

The mobile phone dataset contains 100,000 users with their call detailed records (CDR) and data detailed records (DDR) for December 2013. Each record of the CDR and DDR data has a hashed ID, time-stamp, longitude, and latitude of the cell tower when the phone communicated with it. According to Voronoi tessellation, the average distance between towers is 332 meters (with a median of 254 meters), representing the spatial resolution in the study. Fig. 2b shows the flow between tracts for the morning peak (6am-10am), obtained using the mobile phone data as proxy for surveys, with the methods detailed below. Fig. 3a shows the average number of phone usage records per day that a user has during the whole month. As we see the majority of users are active with an average of 15 records per day.

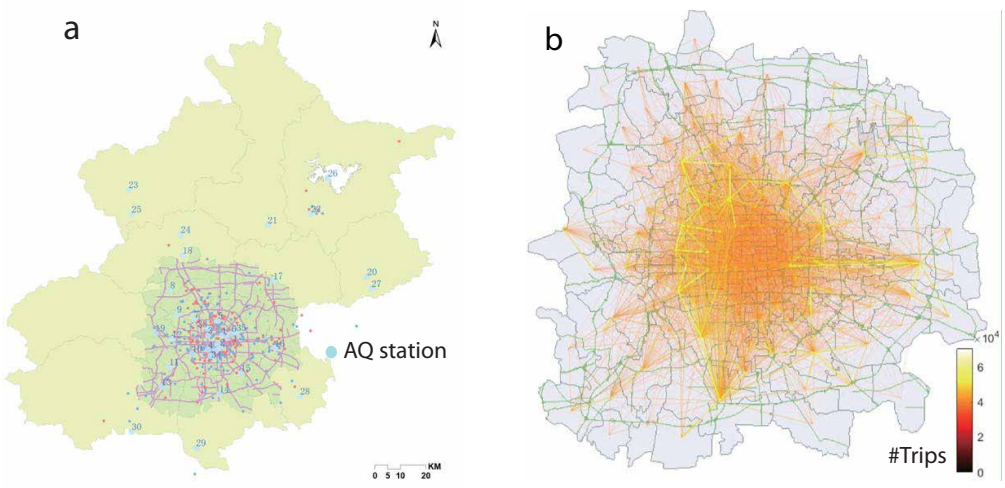


**FIGURE 2 Study Area. (a) The boundary of the whole Beijing, it is about 16,410$km^2$. The city area is the greener area, within the Sixth Ring Road, marked by the outer purple line. The blue circles are 35 air quality (AQ) station. (b) Trips between origin destination (OD) pairs in the Morning peak (7am-10am) in urban Beijing.**

Mobile phone carriers use methods to execute tower-to-tower call balancing to improve their service. This generates signal jumps that introduce noises, appeared as fast and long movements beyond a travel speed limit. To eliminate this artifact, various methods have been reviewed in (*13*). One of the simplest yet effective methods is to remove the next record if the the inferred speed between two records is beyond reasonable speed limit. However, it heavily relies on the correctness of the first record. To improve its accuracy, we check if the first record is a noise —if the speed between the first and the second record is beyond a predefined speed limit, we then remove the first record. We repeat this process until there is no artificial jumps between two records. Next, we distinguish stay-point and pass-by from the remaining records.

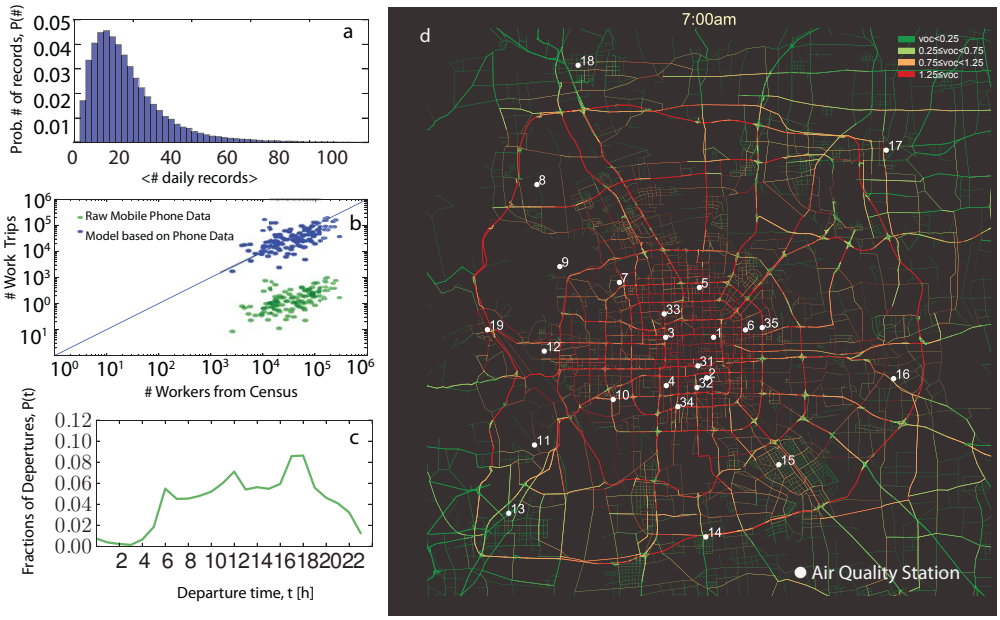


**FIGURE 3 Traffic model of Beijing. (a) The distribution average daily records among the 100,000 mobile phone users (b) Validation of the estimated number of work trips vs. employment information from Census data (c) Estimated fraction of trip departures per hour of the day. (d) Estimated volumes of cars in the streets per time of the day.**

We improve upon the stay-point algorithm presented in (*13, 14*) as follows. (i) we apply a temporal agglomeration algorithm. The temporally consecutive records within a certain radius (e.g., 500 meter) are bundled together with a updated stay duration from the start time of first record to the end time of last one. (ii) We then label the records as pass-by points and stays, according to the stay duration threshold (e.g.,10 minutes) based on the local context in Beijing. In analysis hereafter, we only focus on the stays. We then combine all the spatially adjacent stay points for a user (within a threshold) as his or her stay regions, which will be later labeled as *home*, *work*, and *other*. For this spatial agglomeration, we use R-tree to accelerate the computation (*15*). R-tree is a type of spatial B-tree, a spatial search balancing tree that checks the boundaries of elements to make the search faster (see details in Fig. 4). We then get a mapping relation between stay points and stay regions.

*Stay detection and activity labeling*

We then estimate the type of each stay location for every user, classified as *home*, *work* or *other*. The most visited location during weekday nights and weekends are labeled as *home*, and the most visited one during weekday working hours (at least 500 meters away from home) is labeled as *work*, and the rest are labeled as *other*. We assume that within 500 meters, it is not necessary to travel by car.

*Vehicle demand estimation*

After labeling the activity type, we estimate residential and working population within each zone (i.e, a Voroni polygon generated from towers), and calculate an expansion factor by dividing the number of phone users by total population for each zone. We aggregate the population data at

**Algorithm 1:** Spatial Agglomeration by R-tree (Python)

```
import index from rtree;
tempStay2Stay = dict();
idx = index.Index();
for node in tempStays do
    idx.inserts(tempStay);
    #degenerate the rectangular to a point when inserting the tempStays;
for node in tempStays do
    VectorState[node] = idx.intersection(node's square buffer);
    #search the buffer region of each node to see how many nodes are in its neighbor in the
    constructed rtree (i.e., idx above);
while the sum of StateVector is not 0 do
    choose the node_i with maximum value in StateVector;
    intersection = idx.intersectionnode i's square buffer if the len(intersection)==StateVector[i]
    then
        #cluster the nodes within node i's buffer, get the mapping relationship to the most central
        one for nodes j within the square buffer;
        for node_j in intersection do
            tempStay2Stay[node_j] = node_i;
            StateVector[node_j]=0;
            idx.delete(node_j);
    else
        StateVector[i] = len(intersection);
        continue;
return tempStay2Stay;
```

**FIGURE 4 Algorithm used for detection of stay regions from mobile phone traces**

the 100 by 100 meter grid level obtained from WorldPop [1] to the *Jiedao* level (census zones comparable to towns in U.S.). We compared the total population obtained from WorldPop with the Beijing Census data (2010) at the *Jiedao* resolution, and they are in good agreement. We compare the home-work trips generated by our model with the census employment statistics at the *Jiedao* level, only taking into account the phone users with labeled work location. We find that our employment estimation is in reasonable agreement with the Beijing $2^{nd}$ Economic Census (see Fig. 3b).

Trips are then assigned a trip purpose: home-based-work (commuting), home-based-other, and non-home-based, according to the inferred locations of two consecutive stays. We then get an overall average departure time distribution from all the trips normalized by the number of active days, and an expansion factor for each user. Although a travel survey from Beijing is not available to us at the moment, this method has been approved in other cities with their travel surveys (*11, 12, 16*). In Fig. 3c, we show the estimated fraction of trips per hour in an average day.

We obtained OD matrices by different time periods of an average weekday according to the departure time at both the Voronoi polygon and census tract level, where the number of trips are expanded by the expansion factors. To consider trips made by motorized vehicles, we weigh obtained person trips by vehicle ownership rates at the district level which is larger than *Jiedao*

[1] `http://www.worldpop.org.uk/data/methods/`

(e.g, with 18 districts in Beijing). According to the 2013 Beijing Year Book (*10*), due to local traffic regulation policy, around 20% of cars are restricted not to travel on the road according to their car license numbers. We multiply 0.8 by all trips, as each day two license ending-numbers are restricted by the city. The other factor is the vehicle usage rate— many people who own cars tend to use subways rather than driving to avoid traffic congestion in peak hours. Consequently, we assume a factor of 80% for all tracts, and this step is yet to be improved with more accurate car usage rate data, which is not available at high resolution. Finally, with a traffic assignment model (*17*), we assign the vehicle ODs to the road network resulting estimates of travel time and car volumes for each segment of the road network.

*Day-specific travel demand estimation*

We extend the average 24-hour demand calculated from mobile phone data to day-specific ODs using data reported on traffic congestion index (TCI). TCI is published by Beijing Transportation Research Center (BJTRC) (*18*) and ranges from 0 to 10. As explained by BJTRC, 0 indicates all vehicles in the road network traveling in free flow speed; 10 indicates the travelers on average take double free-flow travel time on the road segments. TCI reflects the degree of congestion, other than the faction of travel demand. We use them, however, as a source of of information to generate variations in demand, with the following equation:

$$f_d = \frac{TCI_{max} + TCI_d}{TCI_{max} + TCI_{mean}} \tag{1}$$

where $f_d$ is the demand factor on the $d$th weekday in our data set; $TCI_d$ is the value of TCI on the $d$th weekday; $TCI_{max}$ and $TCI_{mean}$ are the maximum and mean TCI of all weekdays, respectively. As a result, the zone-to-zone OD matrix is scaled with the demand factor $f_d$ for each weekday. In our experiments, $f_d$ ranges from 0.65 to 1.31. This means that during weekdays from April 2014 to May 2015, we allow for fluctuations in traffic congestion, introducing a degree of uncertainty in the proposed travel demand estimates, enriched by the variations reported by in the TCI on the same days over which we will model the AQIs.

*Traffic Assignment*

To estimate the traffic state and travel time of drivers, we assign the vehicle demand to the road network using a user equilibrium (UE) model. A UE model assumes that all of the travelers in the road network try to find their routes with respect to the shortest travel time (*19, 20*). The road network of Beijing within the Sixth Ring Road is extracted from OpenStreetMap (*21*). We extracted or estimated requisite attributes of road segments, including free flow speed, capacity, length, and number of lanes, from OpenStreetMap.

The road network is represented as a directed acyclic graph (DAG), $\mathcal{G}(\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of all nodes, $\mathcal{E}$ is the set of all edges. In our implementation of the UE model, the anticipated travel time on each edge $e$ is calculated by the Bureau of Public Roads (BPR) function:

$$t_e = \left(1 + \alpha \left(\frac{v_e}{C_e}\right)^{\beta}\right) \times t_e^f \tag{2}$$

where $v_e$ is the number of vehicles attempting to use edge $e$ per hour; $C_e$ is the capacity of the edge; $t_e^f$ is the free flow travel time on edge $e$ and is estimated using the limit speed of the edge; $\alpha$ and $\beta$ are two coefficients and we are using $\alpha = 0.18$ and $\beta = 4$ in our experiments.

To solve the UE model, we minimize the distance between the optimal solution and the current solution in an iterative process (*22, 23*). In our work, the distance is measured using the following equation:

$$r_g = 1 - \frac{\sum_{o\in\mathcal{O},d\in\mathcal{D}} t'_{od} f_{od}}{\sum_{e\in\mathcal{E}} t_e v_e} \tag{3}$$

where $\mathcal{O}$ and $\mathcal{D}$ are the set of origin and destination nodes in the road network; $f_{od}$ is the demand of flow from $o$ to $d$; $t'_{od}$ is the shortest travel time of trip $(o,d)$ in the current iteration. Further details of the implementation of assignment can be found in (*17*).

Fig. 3d shows the assignment results during morning peak hour. The color of each road segment reflects the volume-to-capacity (VoC). A larger VoC indicates that the road is used by a larger number of vehicles compared with its capacity. As seen from the figure, a large proportion of the urban roads are in congestion during the morning peak.

To verify that the assignment results are reliable and robust, we compare the travel time of 5,000 OD pairs with top number of commuters during the morning peak hour with the travel time provided by Gaode (*24*), which is a leading traffic navigation company in China. Fig. 5a shows the comparison of travel times, suggesting that our estimated travel times and Gaode's are quite close for most of the trips. Fig. 5b presents the distribution of commuting time of the top 5000 OD pairs. The distribution indicates that our assignment model provides reliable estimates of travel time delay in the peak hour.

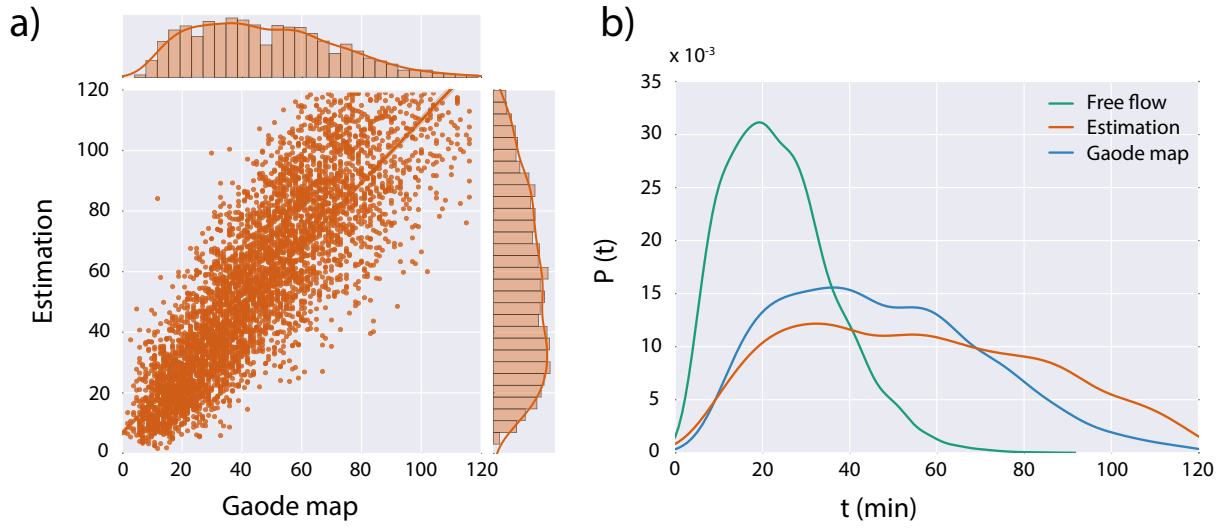


**FIGURE 5 Commuting time validation with Gaode travel time. (a) The scatter plot of** 5000 **trips with top commuters. (b) The distribution of travel time with three modes: free flow, our estimation, and Gaode map.**

**Measuring traffic feature by AQ monitoring station**

The coverage radius for an AQ monitoring stations in the city ranges from 500 meters to 4 kilometers. We define a $2km \times 2km$ square-buffer surrounding each station to examine the relation between traffic around the station and its AQ. This enables us to identify stations that are more sensitive to local traffic. By assigning the day-specific vehicle ODs (extended by the TCIs) to the

road network, we estimate vehicle numbers in the streets by hour for different days. We then estimate the volume of vehicles and travel times for each road segment for each of the days. Since the traffic-related air pollution is not only related to the vehicle volumes but also with the time they spend (to approximate emission) in the road network, we calculate the collective travel time (CTT) within the buffer area of the AQ monitoring station as a traffic feature to model the AQI. The collective travel time is calculated as $t_c = \sum_{e \in \mathcal{B}} v_e t_e$, where $\mathcal{B}$ is the set of roads in the buffer area. Besides, the total VoC is also calculated as the summation of VoC on all roads in the station buffer area.

Fig. 6 shows the CTT and total VoC per hour per station. The CTT and VoC in each station buffer are obtained by assigning the average demand to the road network. As shown in the figure, there are three peak hours on weekdays in Beijing. The CTT at four stations: 1, 3, 5, 31 are significantly higher than others, and three of them also have high VoC. Besides, most stations with heavy traffic are located within the Fourth Ring Road of Beijing. In the next section we discuss the use the CTT as a predictive feature for AQI.

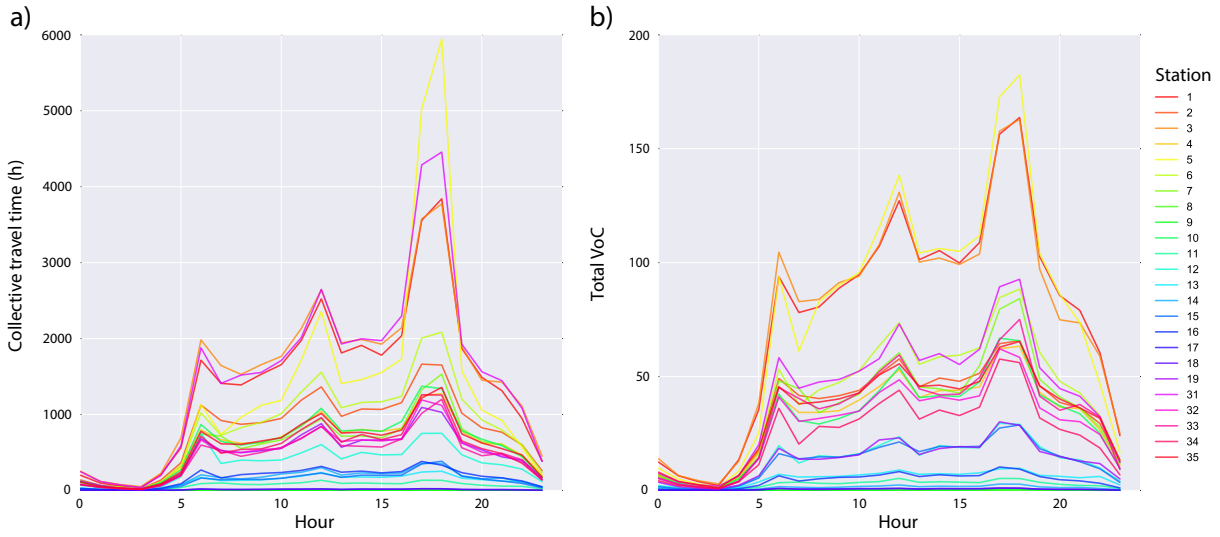


**FIGURE 6 Travel demand information per hour in buffer areas of each station (a) The collective travel time. (b) Total car volume over street capacity (VoC).**

**Impact of traffic to modeling of air quality**

Although traffic is regarded as one of the most critical influences on air pollution in urban areas, the impact of traffic is still not well measured and understood. Zheng *et al.* predicted the AQI in Beijing with features related to meteorology, number of taxi trips, road properties, point of interests (POIs), and traffic related features (e.g., speeds from taxi data) (*8*). They built a single prediction model for the entire city. That is, the model was trained using data from all AQ stations in Beijing, disregarding the spatial variations of AQ. In a later work of the same team, they predicted future AQ in each station, but without considering the traffic factor in the station (*9*). We argue that a city-wide model cannot identify the spatial variations reflecting the importance of local traffic feature for the AQI by station, which is important in relating AQ with transportation policy. In this work, we investigate this aspect, modeling the AQI in each of the 24 monitoring integrating a

travel demand model. The location and ID of the stations are shown in Fig. 3d. We can see that some stations are located in zones with heavier traffic than others.

To evaluate the impact of traffic on air pollution, we model the AQI using the meteorology and traffic information in the same hour. The meteorological features include wind speed, wind direction, humidity, temperature, and pressure. The traffic features include the TCI and proposed CTT. We divide the data set into two parts: summer (from May 1, 2014 to September 30, 2014) and winter (from December 1, 2014 to March 31, 2015). For each part, we train a estimation model for each station under the three aforementioned scenarios, and use the raw AQIs as response. Moreover, as people are more concerned with air quality during daytime, we select the samples from 6:00am to 8:00pm everyday. After eliminating the missing data, the summer data set contains about 430 sample-hours per station; the winter data set contains about 530 sample-hours per station, corresponding to only 31 and 38 days with complete data, respectively. To avoid the overlap between training and testing sets, the first 70% sample-hours are used to train the models, and the last 30% hours are used to test. Subsequently, we estimate the AQI with two distinct models, linear regression and non-linear random forest model (*25, 26*). To obtain stable estimation, we repeat training the model 20 times at each station. At each time, we randomly select 90% data from the training datasets to train the model. The average value of the 20 estimations is regarded as the final estimation of AQI at the station.

# RESULTS

## Analysis of AQI estimation

To assess the impact of traffic features on air quality, we first calculate the relative feature importance of three feature sets, meteorology, TCI, and CTT in two regression models. A linear regression model, and a random forest. For the linear regression, we use the Lindeman, Merenda and Gold (LMG) method to quantify the contribution of individual feature sets to modeling AQI (*27*). For random forest, the importance of a feature set is calculated through the difference of training accuracy with and without the feature set. The estimation accuracy of AQI is calculated by:

$$p = \left(1 - \sum_{i=1}^{N} \frac{|A\hat{Q}I_i - AQI_i|}{AQI_i}\right) \times 100\% \tag{4}$$

where $A\hat{Q}I_i$ is the estimated value of the $i$th sample; $N$ is the number of samples in the testing set.

Fig. 7a and Fig. 7c illustrate the relative feature importance of meteorology, TCI, and CTT in summer, with linear regression and random forest, respectively. As can be seen, meteorology is the leading factor at most stations. However, linear regression suggests TCI is less important than CTT, while random forest suggests TCI and CTT have equal level of contribution to AQI. The importance of features to AQI estimation in winter are divergent for the two regression models, as shown in Fig. 7e and Fig. 7g. Such diversity between two models reflects the AQI in winter is more difficult to model than summer.

Fig. 7b and Fig. 7d present the distribution of the estimated accuracy in all stations in the summer, with the linear regression and random forest, respectively. The red distribution is obtained with model trained with all features, while the blue one is obtained with model trained without traffic features (TCI and CTT). Integrating the traffic features with the meteorological features, the accuracy decreases in some stations. This indicates that the traffic information in a given hour has not direct impact in the AQI in the same hour. The impact of traffic to air quality may be delayed

for more than one hour. Similar results were obtained in winter, as shown in Fig. 7f and Fig. 7h. From these results, we notice that although the traffic information has significant importance in the training phase of regression models, it can not promote the estimation of AQI in the testing phase.

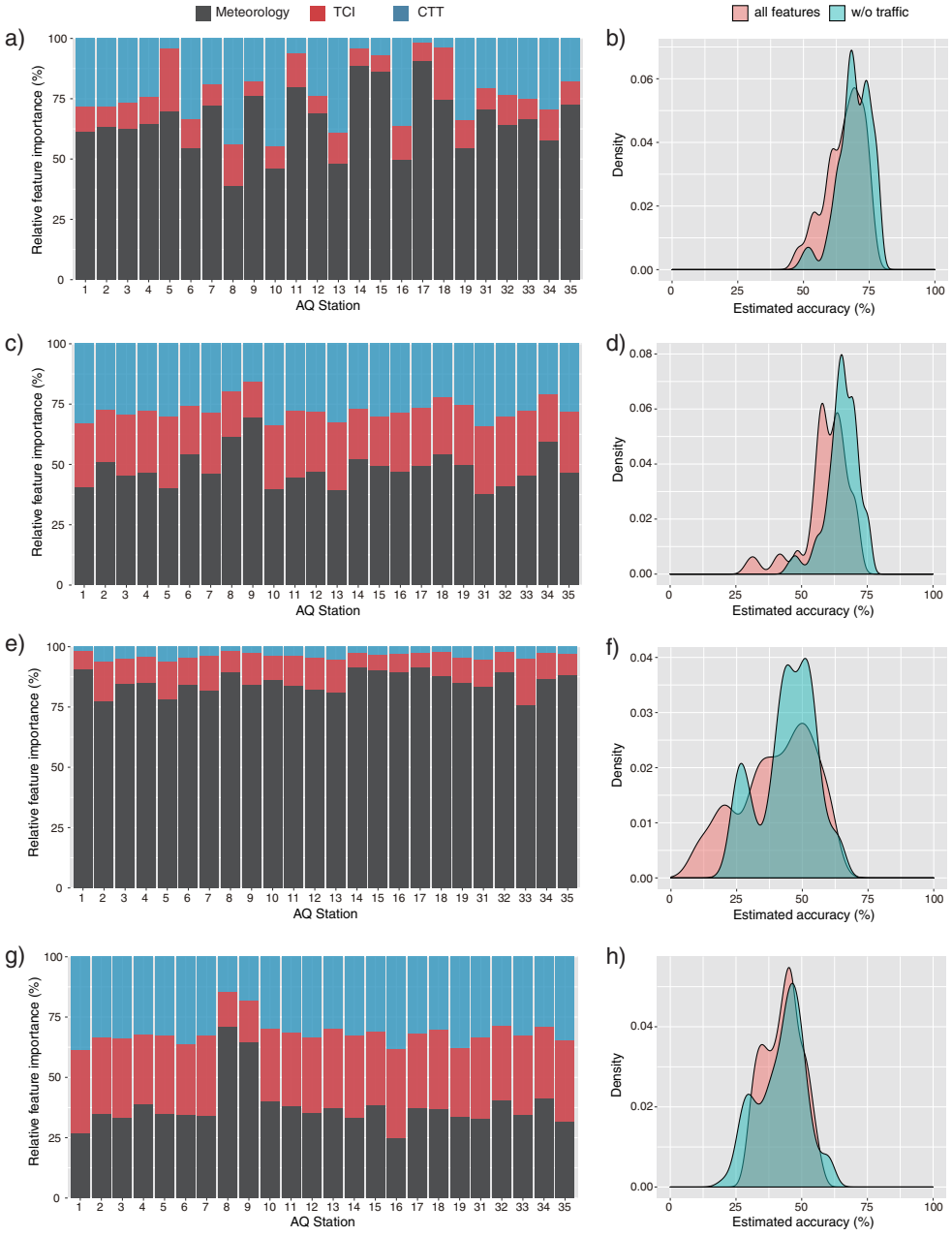


**FIGURE 7 Feature importance and AQI modeling accuracy. (a-b) Relative feature importance per AQ station and the distribution of estimated accuracy of all AQ stations with linear regression in the summer. (c-d) Relative feature importance and the distribution of estimated accuracy with random forest in the summer. (e-h) Results in the winter, e and f are results of linear regression, g and h are results of random forest.**

**Spatial diversity of AQ monitoring stations**

We further analyze the different relationship between AQI and traffic demand information among the 24 stations. In Fig. 8a and 8c, we plot the median value of AQI and CTT at each station in summer and winter, respectively. As shown in the figures, the CTT at the 24 stations are distinctly separated in two groups: heavy and light traffic stations. Heavy traffic stations are located in the inner urban area, while lighter traffic stations are located in the suburbs. To divide the AQIs, we use 100 as a threshold—according to the U.S. Environmental Protection Agency, a AQI higher than 100 is regarded as unhealthy.

Finally, we partition the 24 stations into four groups: healthy with light traffic, healthy with heavy traffic, unhealthy with light traffic, and unhealthy with heavy traffic, shown in Fig. 8a-d with different colors. As seen from results in Fig. 8a and 8b, the median AQIs of all light traffic AQ stations (green) are under 100, which indicates that around these stations, the air quality on most days in the summer are healthy. For the stations with heavy traffic, only two of them (station 12 and 34) are unhealthy. Station 12 is located at the West Fifth Ring Road; station 34 is located at the South Third Ring Road; and both of them suffer with busy traffic. Fig. 8c and 8d show the results in the winter. In general, the air pollution in winter is much severer than that in the summer. Consequently, AQ at some stations (e.g., station 10, 15, 31, 32 and 35) change from healthy in the summer to unhealthy in the winter, while only stations 12 and 5 improve their AQIs in the winter. Interestingly, these stations are all located in the southern area of Beijing. Meanwhile, from the map of major coal power plants in and around Beijing in Fig. 8e, we observe there are some large-capacity power plants at the south-eastern area of Beijing, e.g. Hebei province and Tianjin. This argument has been demonstrated in literature (*28*): in the winter, the air pollution in the north China is more critical than the south because of the burning of coal for heating. On the other side, the traffic is heavier in the inner core for both winter and summer. Therefore, we argue that the degraded air quality in the southern area of Beijing reaching the unhealthy limits, is likely not related to traffic but due to heating by coal sources.

## CONCLUSION

In this paper, we studied the contribution of traffic related features to the air quality index in the same hour in 24 monitoring stations in Beijing. We integrated mobile phone meta data and publicly available daily traffic congestion index (TCI) to define the traffic features. First, we estimate zone-to-zone vehicle travel using mobile phone data, census data, vehicle usage rate, and road network information. Second, we generate day-specific hourly ODs using TCIs. The day-specific ODs are then assigned to the road network, and the maximum collective travel time (CTT) surrounding each AQ station area is estimated per day in the studied period. Based on the meteorological data, the TCI, and our estimates of CTT, we built two regression models for each station in the summer and the winter. The results show that the traffic information has significant importance in the training phase of the regression model. However, it cannot promote the estimation accuracy in the testing phase. The main reasons may be: (i) the air pollution generated by automobile can not be reflected by AQI immediately; (ii) the regression models do not capture the relations between traffic features and AQI effectively due to the limited period of observation and sample size of to generate the travel demand model.

Moreover, to relate the impact of traffic on air quality in space, we categorize the 24 stations within Sixth Ring Road of Beijing into four groups. We find that the stations with heavy traffic are in the inner core of the city both in winter and summer. The stations with unhealthy levels of

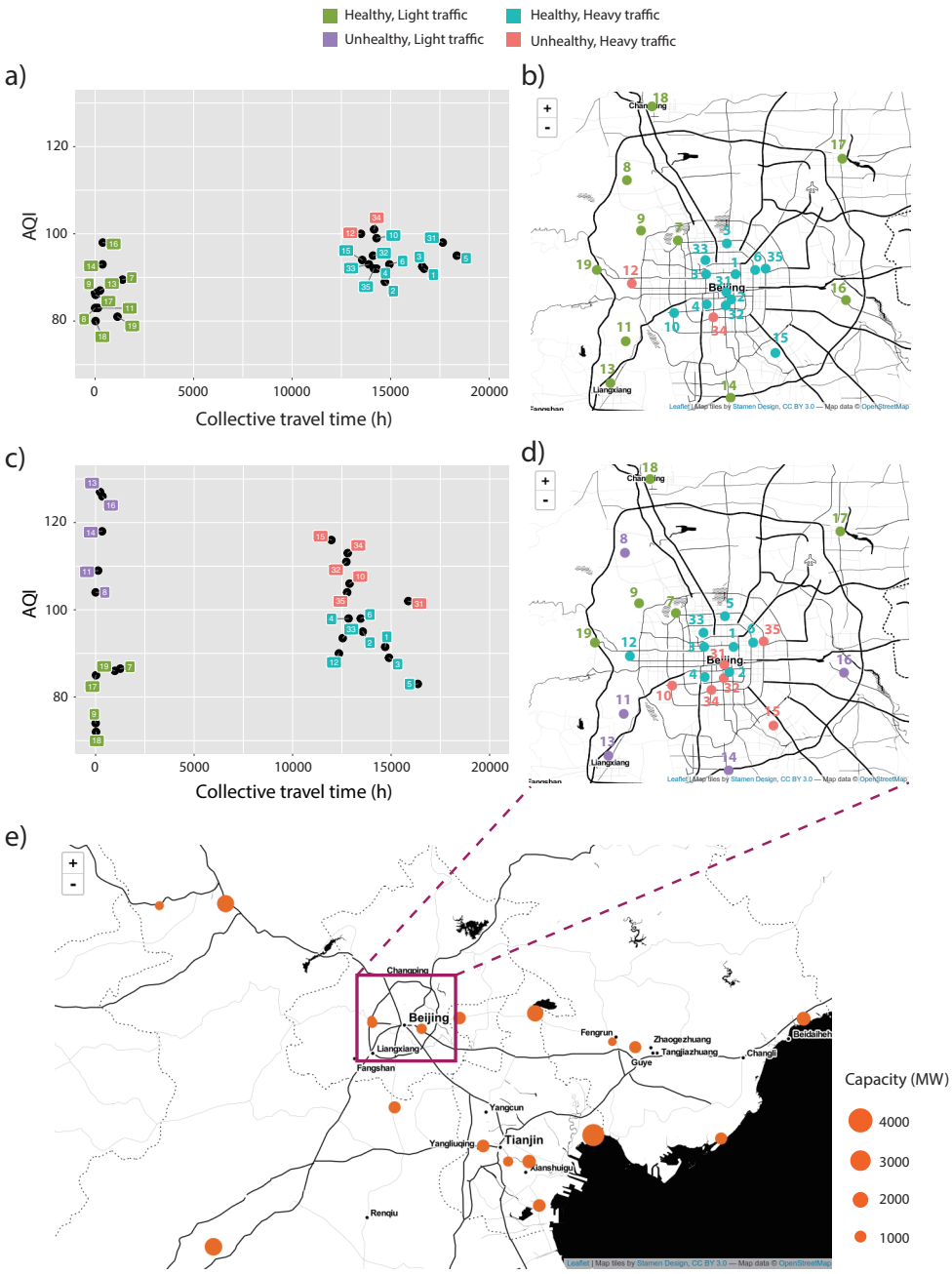


**FIGURE 8 The spatial separation of stations according to AQI and CTT. (a, b) Stations separation results in summer. (c, d) Stations separation results in winter. (e) The location of major coal power plants in and around Beijing.**

air pollution appear in the winter and are concentrated in the southern area of Beijing. Based on these observations, it suggests that the coal heating rather than traffic contributes significantly to the degraded air quality in south Beijing in the winter.

The presented framework is portable, as the data sets employed here can be easily obtained for other cities. The traffic estimation model is of low cost in computation and data require-

ments. This work also provides a data pipeline to categorize AQ monitoring stations more affected by traffic congestion, and to estimate AQIs based on meteorology data, traffic congestion index, and travel demand estimates from mobile phone meta data. There are important avenues for future work, these include: (i)to further investigate the variation of specific pollutants such as $NO_2$, PM2.5 and PM10 in space; (ii) to employ disaggregated vehicle models to detect the bottlenecks of congestion in the road network, with sensitivity analyses for the effects of unknown parameters (such as presences of buses and trucks, which are important sources of vehicle emissions); (iii) to validate the potential sources of pollution, integrating aerial images (from providers of remote sensing data such as Planet Labs) with longer and more detailed observations of pollutant sources and presence of vehicles.

## ACKNOWLEDGMENTS

We acknowledge Jinhua Zhao for enlightening discussions. This work would not have been possible without the kind support, information and data provided by Zheng Chang. This work was funded in part by the MIT-Environmental Solutions Initiative, the MIT Samuel Tak Lee (STL) Real Estate Entrepreneurship Lab, the New England UTC 25, and the Center for Complex Engineering Systems (CCES) at KACST.

## REFERENCES

[1] Dockery, D. W., C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris Jr, and F. E. Speizer, An association between air pollution and mortality in six US cities. *New England journal of medicine*, Vol. 329, No. 24, 1993, pp. 1753–1759.

[2] Hoek, G., B. Brunekreef, S. Goldbohm, P. Fischer, and P. A. van den Brandt, Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *The lancet*, Vol. 360, No. 9341, 2002, pp. 1203–1209.

[3] Guo, S., M. Hu, M. L. Zamora, J. Peng, D. Shang, J. Zheng, Z. Du, Z. Wu, M. Shao, L. Zeng, et al., Elucidating severe urban haze formation in China. *Proceedings of the National Academy of Sciences*, Vol. 111, No. 49, 2014, pp. 17373–17378.

[4] Kelly, F. J. and T. Zhu, Transport solutions for cleaner air. *Science*, Vol. 352, No. 6288, 2016, pp. 934–936.

[5] McHugh, C., D. Carruthers, and H. Edmunds, ADMS–Urban: an air quality management system for traffic, domestic and industrial pollution. *International Journal of Environment and Pollution*, Vol. 8, No. 3-6, 1997, pp. 666–674.

[6] Vardoulakis, S., B. E. Fisher, K. Pericleous, and N. Gonzalez-Flesca, Modelling air quality in street canyons: a review. *Atmospheric environment*, Vol. 37, No. 2, 2003, pp. 155–182.

[7] Baldauf, R., E. Thoma, M. Hays, R. Shores, J. Kinsey, B. Gullett, S. Kimbrough, V. Isakov, T. Long, R. Snow, et al., Traffic and meteorological impacts on near-road air quality: Summary of methods and trends from the Raleigh near-road study. *Journal of the Air & Waste Management Association*, Vol. 58, No. 7, 2008, pp. 865–878.

[8] Zheng, Y., F. Liu, and H.-P. Hsieh, U-Air: when urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 1436–1444.

[9] Zheng, Y., X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2015, pp. 2267–2276.

[10] *Beijing Regional Statistic Year Book*. `http://www.bjstats.gov.cn/nj/qxnj/2014/zk/indexch.htm`, 2016, [Online; accessed 12-January-2016].

[11] Alexander, L., S. Jiang, M. Murga, and M. C. González, Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 2015.

[12] Çolak, S., L. P. Alexander, B. G. Alvim, S. R. Mehndiretta, and M. C. González, Analyzing cell phone location data for urban travel: current methods, limitations and opportunities. In *Transportation Research Board 94th Annual Meeting*, 2015, 15-5279.

[13] Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González, A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, ACM, 2013, p. 2.

[14] Zheng, Y. and X. Xie, Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 1, 2011, p. 2.

[15] Guttman, A., *R-trees: a dynamic index structure for spatial searching*, Vol. 14. ACM, 1984.

[16] Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 2015.

[17] Çolak, S., A. Lima, and M. C. González, Understanding congested travel in urban areas. *Nature communications*, Vol. 7, 2016.

[18] *Traffic congestion index in Beijing*. `http://www.bjtrc.org.cn/PageLayout/IndexReleased/Realtime.aspx`, 2016, [Online; accessed 30-March-2016].

[19] Wardrop, J. G., ROAD PAPER. SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESEARCH. *Proceedings of the institution of civil engineers*, Vol. 1, No. 3, 1952, pp. 325–362.

[20] Friesz, T. L. and D. Bernstein, *Foundations of Network Optimization and Games*. Springer, 2016.

[21] *OpenStreetMap*. `https://www.openstreetmap.org`, 2016, [Online; accessed 18-April-2016].

[22] Dial, R. B., A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. *Transportation Research Part B: Methodological*, Vol. 40, No. 10, 2006, pp. 917–936.

[23] Nie, Y. M., A class of bush-based algorithms for the traffic assignment problem. *Transportation Research Part B: Methodological*, Vol. 44, No. 1, 2010, pp. 73–89.

[24] *AMP direction API*. `http://lbs.amap.com/api/webservice/reference/direction/`, 2016, [Online; accessed 16-May-2016].

[25] Breiman, L., Random forests. *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32.

[26] Liaw, A. and M. Wiener, Classification and Regression by randomForest. *R News*, Vol. 2, No. 3, 2002, pp. 18–22.

[27] Lindeman, R. H., P. Merenda, and R. Gold, *Introduction to bivariate and multivariate analysis*, 1980.

[28] Chen, Y., A. Ebenstein, M. Greenstone, and H. Li, Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences*, Vol. 110, No. 32, 2013, pp. 12936–12941.