SUPPLEMENTARY INFORMATION

Unraveling Environmental Justice in Ambient $PM_{2.5}$ Exposure in Beijing: A Big Data Approach

Yanyan Xu[1,2,3], Shan Jiang[4,5], Ruiqi Li[6,7], Jiang Zhang[6], Jinhua Zhao[8], Sofiane Abbar[9], Marta C. González[1,2,3*]

[1]Department of Civil & Environmental Engineering, MIT, Cambridge, MA, 02139, USA

[2]Department of City and Regional Planning, UC Berkeley, 406 Wurster Hall, Berkeley, CA 94720

[3]Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, CA 94720

[4]Department of Urban and Environmental Policy and Planning, Tufts University, Medford, MA 02155, USA

[5]Department of Civil & Environmental Engineering, Tufts University, Medford, MA 02155, USA

[6]School of Systems Science, Beijing Normal University, Beijing 100875, China

[7]College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

[8]Department of Urban Studies & Planning, MIT, Cambridge, MA 02139, USA

[9]Qatar Computing Research Institute, HBKU, Doha 5825, Qatar

*To whom correspondence should be addressed; E-mail: martag@berkeley.edu

List of Figures and Notes:

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 S1: Pseudo code of R-tree algorithm. 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.

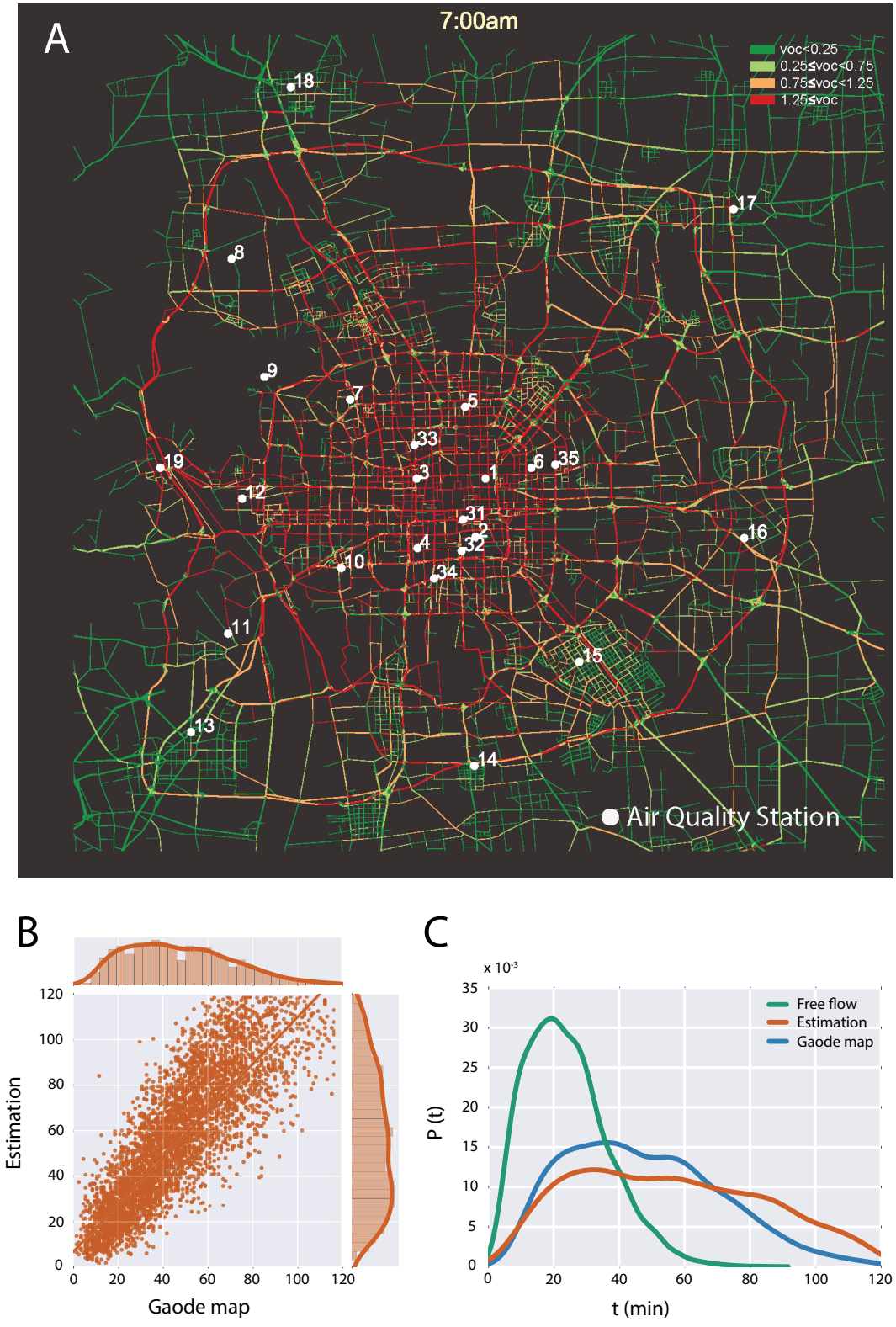

Figure S2: Estimated traffic states during the morning peak hour and the validation. **A** The four level of colors implies the traffic stats on each road segment. The colors are defined with the value of volume-over-capacity. **B** The estimated travel time of major OD pairs versus that of the travel time from Gaode map. **C** The distribution of travel time of free flow, estimation, and Gaode map during the morning peak hour. The estimated travel time has a good agreement with Gaode map.

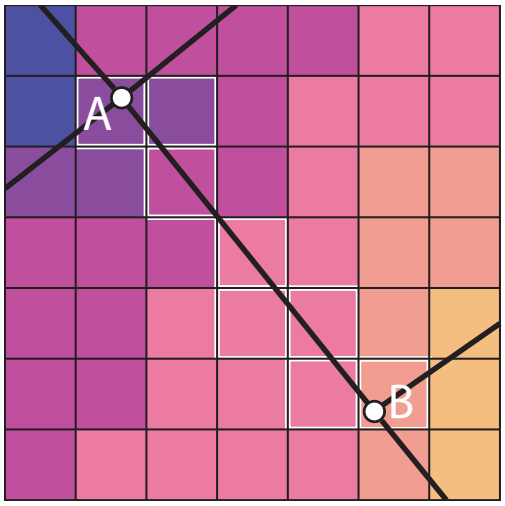

Figure S3: $PM_{2.5}$ concentration estimation for a given road segment. The size of each grid is 200 × 200 m^2. The road segment A→B covers 8 grids labeled with white edges. The $PM_{2.5}$ concentration of A→B equals to the average concentration of the 8 grids.

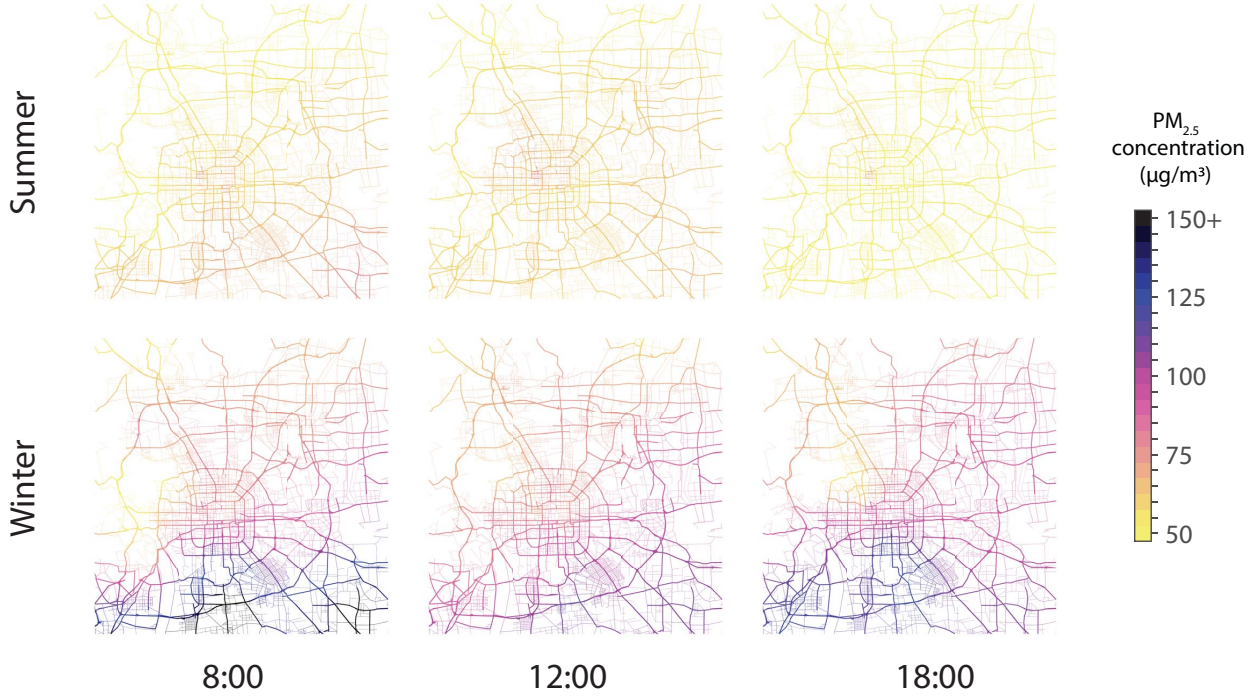

Figure S4: Estimated $PM_{2.5}$ concentration on road segments at 8:00, 12:00, and 18:00 during the summer and winter of 2015 in Beijing.

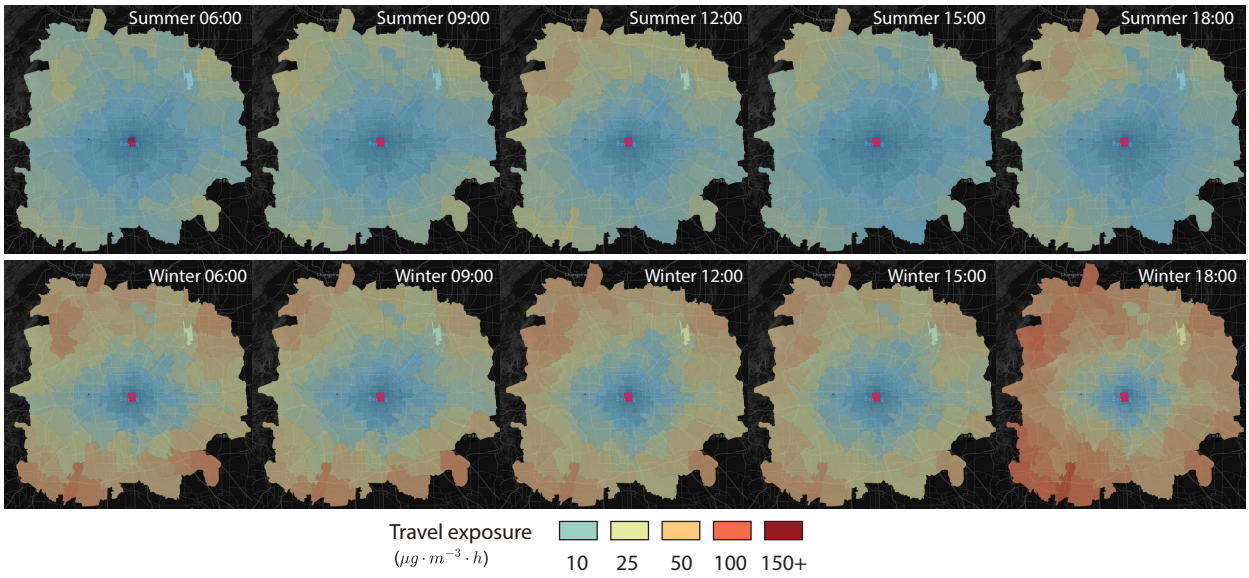

Figure S5: Visualization of travel exposure in Beijing. The travel exposure from the origin zone (purple zone), Forbidden City, to the rest of the city during different hours in summer and winter of 2015. We built an on-line visualization platform for the public: `http://www.mit.edu/~yanyanxu/exposure/`

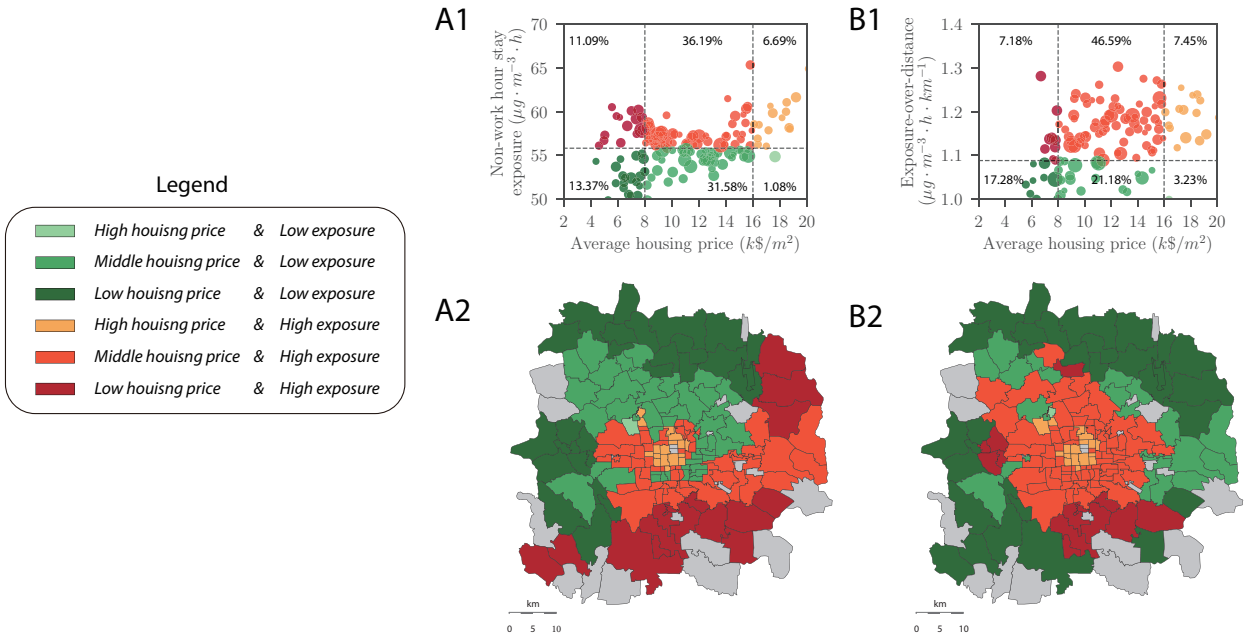

Figure S6: Connection between $PM_{2.5}$ exposure in summer and housing price. **A** PDWE during non-work hours on the selected days in summer versus the average housing price of the same zone. **B** PDWTE from each zone versus the housing prices. **C** Average individual travel exposure versus the housing prices. **D-F** Spatial distribution of six groups in summer.

NOTE 1: URBAN MOBILITY FROM MOBILE PHONE DATA

We estimate the travel demand for the 16.3 million residents living in the urban area of Beijing. This is commonly referred to the region within the Sixth Ring Road, shown in Fig.S2A, which has 5.6 million privately-owned vehicles registered in 2013 [1]. 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 [2, 3]. 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.

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.

1.1 MOBILE PHONE DATA AND STAY DETECTION

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.

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 [4]. 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.

We improve upon the stay-point algorithm presented in [4, 5] 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 [6]. 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. S1). We then get a mapping relation between stay points and stay regions.

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.

1.2 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 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.

Each trip is 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 [2, 3, 7].

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 $Jiedao$ (e.g, with 18 districts in Beijing). According to the 2013 Beijing Year Book [1], 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 [8], we assign the vehicle ODs to the road network extracted from OpenStreetMap [9] resulting estimates of travel time and car volumes for each segment of the road network.

NOTE 2: PERCEIVED AIR QUALITY SURVEY DATA

In the perceived air quality (PAQ) experiment, more than 26,000 individuals received the study invitation, and around 1,000 individuals expressed interest to participate. Among those, 860 individuals downloaded our smartphone-based survey application to track their daily trajectory, and 256 of them finished the survey. The participant fills out a PAQ questionnaire each day during the survey period. They rate the PAQ for home, workplace, and worst spot during their commuting during the two week study period in the winter of 2015 (although some of them didn't complete the whole period).

The perception runs from 1 to 6, respectively representing very bad, bad, just unacceptable, just acceptable, good, and very good. If the rating for a place is equal or under 2, meaning the air quality is bad, then a more specific question pops up. The specific question asks the subject to evaluate the level of air haziness, irritation, bad odor, and headache or dizziness. The evaluation is based on a 1 to 4 rank, respectively representing very bad, bad, just noticeable, and not at all. Overall, the mean of the reported PAQ is 4.68 and the standard deviation is 0.82. Based on this result, the PAQ in Beijing falls to the positive side overall, with a moderate variance across days. More details about the analysis of the PAQ data can be found in [10].

REFERENCES

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

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

[2] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–250, 2015.

[3] Serdar Çolak, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta, and Marta C González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record: Journal of the Transportation Research Board*, (2526):126–135, 2015.

[4] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta 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*, page 2. ACM, 2013.

[5] Yu Zheng and Xing Xie. Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):2, 2011.

[6] Antonin Guttman. *R-trees: a dynamic index structure for spatial searching*, volume 14. ACM, 1984.

[7] Jameson L Toole, Serdar Colak, Bradley Sturt, Lauren P Alexander, Alexandre Evsukoff, and Marta C González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162–177, 2015.

[8] Serdar Çolak, Antonio Lima, and Marta C González. Understanding congested travel in urban areas. *Nature communications*, 7, 2016.

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

[10] Zelin Li. Smartphone-based mobility mapping and perceived air quality evaluation in beijing. Master's thesis, MIT, Cambridge, MA, 7 2016.