sustainability

Article

Quantifying Individual $PM_{2.5}$ Exposure with Human Mobility Inferred from Mobile Phone Data

Zhaoping Hu [1,2,†], Le Huang [1,2,†], Xi Zhai [3], Tao Yang [3], Yaohui Jin [1,2,*] and Yanyan Xu [1,2,*]

[1] MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China; zhaopinghu@sjtu.edu.cn (Z.H.); huanglelele@sjtu.edu.cn (L.H.)
[2] Data-Driven Management Decision Making Laboratory, Shanghai Jiao Tong University, Shanghai 200240, China
[3] Shanghai Urban and Rural Construction and Traffic Development Research Institute, Shanghai 200240, China; jessie_zx28@163.com (X.Z.); yangtaocoolboy@163.com (T.Y.)
* Correspondence: jinyh@sjtu.edu.cn (Y.J.); yanyanxu@sjtu.edu.cn (Y.X.)
† These authors contributed equally to this work.

Abstract: Treatment of air pollution and health impacts are both crucial components of long-term sustainability. Measuring individual exposure to air pollution is significant to evaluating public health risks. In this paper, we introduce a big data analytics framework to quantify individual $PM_{2.5}$ exposure by combining residents' mobility traces and $PM_{2.5}$ concentration at a 1-km grid level. Diverging from traditional approaches reliant on population data, our methodology can accurately estimate the hourly $PM_{2.5}$ exposure at the individual level. Taking Shanghai as an example, we model one million anonymous users' mobility behavior based on 60 billion Call Detail Records (CDRs) data. By integrating users' stay locations and high-resolution $PM_{2.5}$ concentration, we quantify individual $PM_{2.5}$ exposure and find that the average $PM_{2.5}$ exposure of residences in Shanghai is 60.37 $\mathrm{ug \cdot h \cdot m^{-3}}$ during the studied period. Further analysis reveals the unbalance of the spatiotemporal distribution of $PM_{2.5}$ exposure in Shanghai. Our $PM_{2.5}$ exposure estimation method provides a reliable evaluation of environmental hazards and public health predicaments confronted by residents, facilitating the formulation of scientific policies for environmental control, and thus advancing the realization of sustainable development.

Keywords: human mobility; mobile phone data; $PM_{2.5}$ exposure

Citation: Hu, Z.; Huang, L.; Zhai, X.; Yang, T.; Jin, Y.; Xu, Y. Quantifying Individual $PM_{2.5}$ Exposure with Human Mobility Inferred from Mobile Phone Data. *Sustainability* **2024**, *16*, 184. https://doi.org/10.3390/su16010184

Academic Editor: Elena Cristina Rada

Received: 1 November 2023
Revised: 9 December 2023
Accepted: 19 December 2023
Published: 25 December 2023

1. Introduction

Air pollution, which includes emissions from vehicles, industrial processes, and other sources, contributes to environmental degradation. Pollutants can harm ecosystems, damage vegetation, and affect water quality, thus undermining the sustainability of natural resources. On the other side, poor air quality caused by pollution can have severe health consequences, leading to respiratory diseases, cardiovascular problems, and even premature death. Therefore, quantifying the health impact of air pollution plays an important role in urban sustainable development.

Exposure refers to the dynamic interaction between air pollutants and the surface of human body, delineating the interplay between the environment and the human body. Assessing the level of pollutant exposure involves evaluating both the duration of contact and the concentration of associated pollutants [1]. As one of the environmental problems derived from industrialization, the air pollutant $PM_{2.5}$ has a serious impact on the health of residents. Both long-term [2] and short-term [3] exposure to $PM_{2.5}$ will have harmful effects on human health, especially increasing the risk of cardiovascular and respiratory diseases, as well as lung cancer, thus directly affects the health of residents. Since there is no established research indicating that $PM_{2.5}$ concentration below a certain threshold is entirely harmless to humans, it is important to minimize exposure levels as much as possible. As $PM_{2.5}$ has significant detrimental influences on human health, it is crucial to quantify exposure to $PM_{2.5}$ throughout the day. The estimation of exposure can stimulate more discussions about public health concerns [4], provide more precise health guidance for individuals, and also offer a scientific basis for comprehensive health management of residents.

In the past few years, the quantification of human exposure to pollutants has been constrained by the availability of extensive data and computational resources. Early studies predominantly focused on aggregate level exposure assessments, such as community level or neighborhood level [5–8]. This approach, however, is susceptible to the Modified Area Unit Problem (MAUP) [9], where the outcomes are influenced by the geographical units or spatial scales employed in the studies. Such aggregated analyses may only offer a partial representation of actual exposure scenarios, potentially leading to imprecise conclusions. Another limitation of aggregate level studies is the omission of the mobility behavior of individuals. These studies, by basing exposure assessments predominantly on pollutant concentrations at individuals' residences and overlooking their mobility behaviors, give rise to the Neighborhood Effect Averaging Problem (NEAP) [10]. Park et al. [11] validated this problem, emphasizing the necessity of incorporating spatiotemporal variations in both human mobility and pollutant concentrations to enhance the accuracy of exposure assessments.

In urban built environment, particularly where traffic congestion and long-distance commutes are prevalent [12], individuals may spend substantial time away from their residences, making the consideration of mobility patterns in exposure assessments indispensable. Therefore, it is necessary to assess exposure from the individual level rather than aggregate level. Traditional methodologies for individual level exposure quantification have relied on questionnaire surveys to gather trajectory information [13,14]. However, these methods are labor-intensive, costly, and impractical for large-scale population studies.

The advent of big trajectory data, such as Call Detail Records (CDR), has provided new opportunities for modeling human mobility [15–17]. Notable studies have leveraged mobile phone data to elucidate disparities in $PM_{2.5}$ exposure. For instance, Xu et al. [18] utilized CDR data to investigate environmental justice aspects of $PM_{2.5}$ exposure in Beijing, revealing economic disparities in exposure levels. Similarly, Guo et al. [19] examined exposure disparities across multiple temporal scales, although their study was limited by the short duration of mobile phone data available, restricting the ability to assess long-term stable exposure patterns of residents. Besides, it is worth noting that research on the disparity of environmental air pollution exposure mainly focuses on developed countries [20–25], while developing countries, despite suffering from more severe air pollution, have relatively limited research on such exposure inequalities [26–28]. To the best of our knowledge, there is also no paper studying the residents' individual exposure to $PM_{2.5}$ in Shanghai, which is one of the most iconic cities in China.

In this paper, we propose a big data analytics framework to accurately quantify individual $PM_{2.5}$ exposure in Shanghai by coupling mobile phone data with $PM_{2.5}$ concentration data at a fine scale. The mobile phone data is generated by the interaction between mobile phones and communication base stations in daily life from January to April 2014. By performing stay point detection on mobile phone data, we can identify the user's place of residence and work, as well as the complete daily trajectory. Moreover, to infer fine-grained $PM_{2.5}$ concentration, we combine two types of data: station monitoring data with high temporal resolution and China High Air Pollutants (CHAP) data with high spatial resolution. This paper proceeds to compute the individual's exposure to $PM_{2.5}$ by utilizing their stay behavior and environmental corresponding $PM_{2.5}$ concentration. By comparing the results with residence-based exposure, we demonstrate the importance of mobility in measuring exposure. In addition, we analyze the spatial and temporal variations in individual exposure, providing new perspectives and data support for policy formulation.

This paper is organized as follows: Section 2 describes the data utilized in this study and the methodology we propose, Section 3 presents the obtained results. In Section 4, we further discuss the obtained results, and Section 5 presents our conclusions and future work.

2. Materials and Methods

2.1. Data Description

In this section, we introduce the datasets used in this study. including Call Detail Records (CDR) data and $PM_{2.5}$ concentration recordings.

When a mobile phone user performs operations such as turning on and off the phone, making a call, sending a text message, or using the mobile data network, his mobile phone exchanges information with the base station on a regular or occasional basis to ensure the quality of communication and to perform billing operations. In this process, communication operators will record these interaction timestamps and interactive base station code or location and other information in real-time. In addition to the data generated by the active operation of the mobile phone user, the base station also periodically detects the signal of the mobile phone and interacts with the mobile phone according to a specific time period.

Considering that the mobile phone always interacts with the nearest base station, the CDR data can reflect the user's location. In this paper, we use the Call Detailed Record data of the Shanghai area provided by the communication operator, which contains the records of one million anonymous users exchanging information with the base station from 1 January 2014, to 31 April 2014. Each record contains the user's anonymous ID, the timestamp of the interaction with the base station, and the latitude and longitude of the base station, as shown in Table 1. These one million users generated nearly 60 billion records.

To protect users' privacy, the CDR data we use is not the most recent. This CDR dataset is only utilized to support this research and has not been made public. Considering that the land use types in Shanghai are relatively stable, correspondingly, the spatial distribution of the population and daily mobility behavior in Shanghai are also stable. Therefore, even though the data we use is collected in 2014, we can still obtain reliable information about users' daily mobility behavior for our research.

Table 1. Feature description in CDR data.

Feature	Type	Description
uid	String	Anonymous user ID
time	String	Timestamp of record, e.g., 20140101002656
lon	Float	The longitude of the base station, e.g., 121.469116
lat	Float	The latitude of the base station, e.g., 31.225176

The calculation of individual $PM_{2.5}$ exposure requires $PM_{2.5}$ concentration data with high spatial and temporal resolution. In this paper, we fuse two $PM_{2.5}$ datasets to generate hourly $PM_{2.5}$ concentration data for each 1-km grid.

Firstly, we collected the air quality data from the national fixed monitoring stations provided by China National Environmental Monitoring Station. Specifically, we selected the real-time concentration data of $PM_{2.5}$ monitored hourly by fourteen fixed monitoring stations in Shanghai for research. The spatial distribution and the number of monitoring stations are shown in Figure 1.

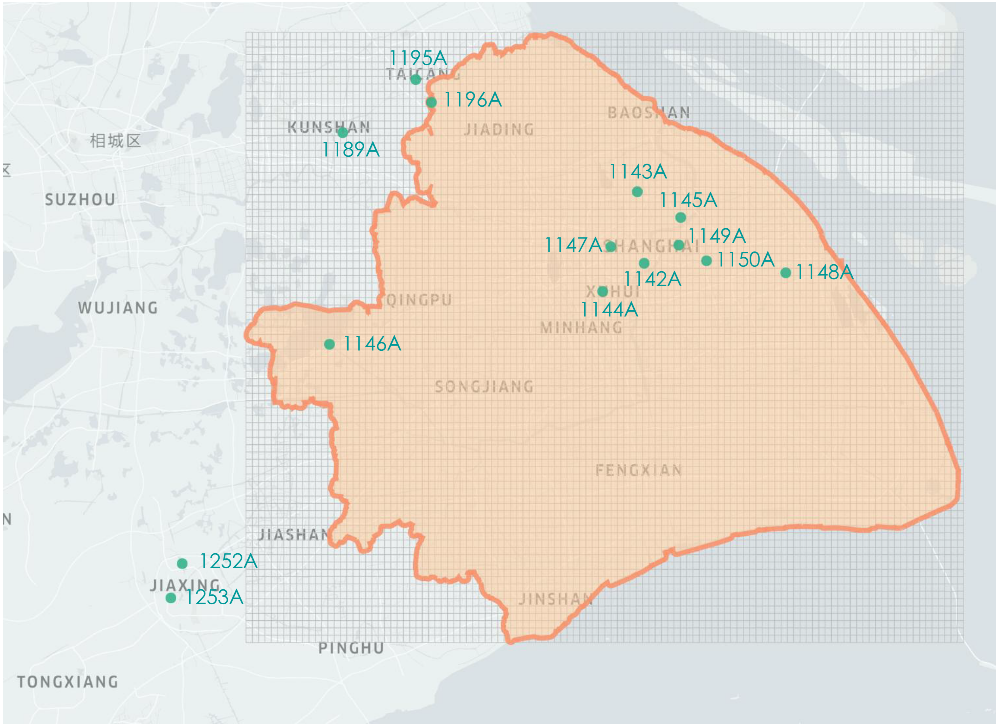

Figure 1. Study area in Shanghai. The orange part represents the study area, the green dots represent the monitoring station locations, and the grey grids represent the specific study spaces.

Shanghai's environmental monitoring stations were only able to provide hourly $PM_{2.5}$ monitoring data after May 2014. We compared the daily average $PM_{2.5}$ concentration of nine monitoring stations in Shanghai within four months in 2014 and 2015. The results in Table 2 show that there is slight difference in the average $PM_{2.5}$ concentration data between 2014 and 2015 in Shanghai. Therefore, in this paper, we used the $PM_{2.5}$ data from January to April 2015 provided by the environmental monitoring stations in Shanghai to extract the daily pattern of $PM_{2.5}$.

Table 2. Comparison of daily average $PM_{2.5}$ concentration in 2014 and 2015.

Average $PM_{2.5}$ Concentration	2014 (ug/m^3)	2015 (ug/m^3)	Diff (ug/m^3)
January	76.61	83.03	−6.42
February	52.32	64.39	−12.07
March	56.87	54.10	2.77
April	52.93	55.80	−2.87

The other dataset is China High Air Pollutants (CHAP), a high spatial resolution and high-quality near-surface $PM_{2.5}$ pollutant dataset in China reconstructed by Jing et al. [29,30]. They constructed a Space-Time Extra-Trees (STET) model by fusing aerosol optical depth (AOD) data, meteorological data, land surface conditions, and population distribution to estimate the concentration of $PM_{2.5}$. This dataset provides the daily average $PM_{2.5}$ surface concentration on a 1-km grid over China from 2000 to 2021. The cross-validation determination coefficient and root mean square error of the model used to estimate $PM_{2.5}$ in this dataset were 0.92 and 10.76 respectively. In this paper, we select the daily $PM_{2.5}$ concentration data in Shanghai from January 2014 to April 2014 within the scope of 1-km grids to calculate the individual $PM_{2.5}$ exposure.

For the daily $PM_{2.5}$ data in the Shanghai area provided by the CHAP dataset, we first performed the visual analysis, as shown in Figure 2. From the temporal perspective, the pollution problem is more prominent in January of winter 2014, with the monthly average concentration of 62 ug/m^3, while the average concentration of February, March, and April are 42 ug/m^3, 44 ug/m^3 and 38 ug/m^3 respectively. From the spatial perspective, the spatial distribution of $PM_{2.5}$ concentration in Shanghai showed a trend of high in the west and low in the east.

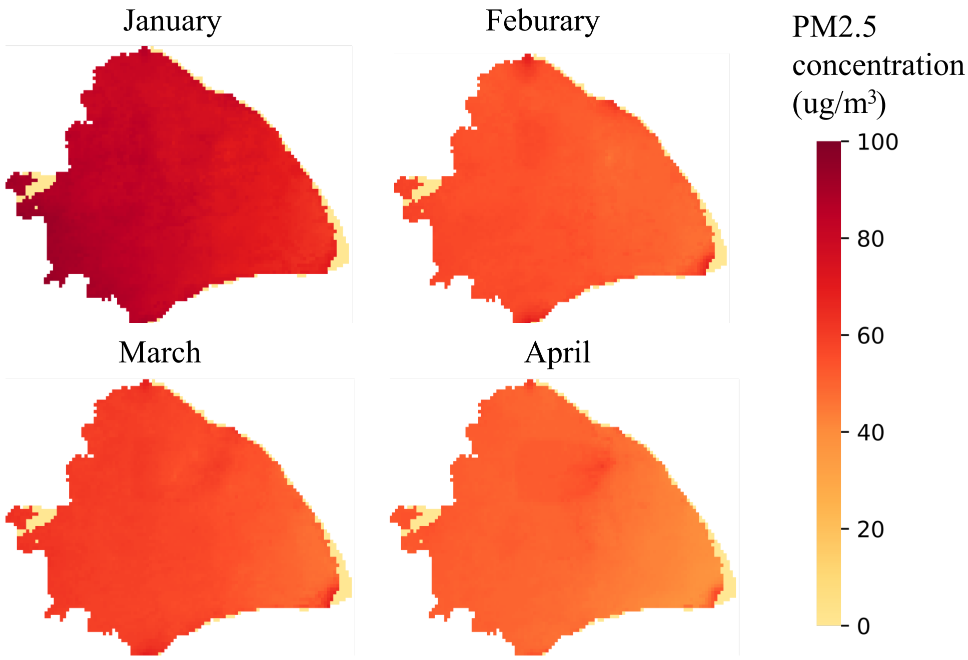

Figure 2. The daily average $PM_{2.5}$ concentration in Shanghai from CHAP dataset.

2.2. High-Resolution $PM_{2.5}$ Data

As introduced above, we used two $PM_{2.5}$ datasets in our research. The $PM_{2.5}$ data provided by the fixed monitoring station has high temporal resolution and low spatial resolution, while the CHAP dataset has low temporal resolution and high spatial resolution. Therefore, we combine these two spatiotemporal $PM_{2.5}$ concentration datasets and derive a new $PM_{2.5}$ dataset that can provide hourly $PM_{2.5}$ concentration data in every 1-km grid in Shanghai. We used the hourly $PM_{2.5}$ concentration provided by the fixed monitoring station to correct the daily average $PM_{2.5}$ concentration data of the 1-km grid provided by the CHAP dataset to infer the hourly $PM_{2.5}$ concentration of each 1-km grid in Shanghai. In order to obtain the hourly average $PM_{2.5}$ concentration data based on the corresponding daily average $PM_{2.5}$ concentration data, we define a parameter named *correction factor*. For each fixed monitoring station, we estimate its hourly average $PM_{2.5}$ concentration in the m month based on the data provided by the fixed pollutant monitoring station and then calculate the average $PM_{2.5}$ concentration for the whole month. The ratio of these two concentration records is the hourly correction factor for the monitoring station this month. Specifically, for monitoring station f, we use $C_{f,m,d,h}$ to represent its $PM_{2.5}$ concentration on day d hour h in month m and the correction factor $CF_{f,m,h}$ of the station at the h hour of the m month is calculated as follows:

$$C_{f,m,h} = \frac{\sum_{d \in M} C_{f,m,d,h}}{|M|} \tag{1}$$

$$C_{f,m} = \frac{\sum_{d \in M} \sum_{h=0}^{23} C_{f,m,d,h}}{|M| \times 24} \tag{2}$$

$$CF_{f,m,h} = \frac{C_{f,m,h}}{C_{f,m}} \tag{3}$$

where M represents the set of days in the m-th month. Since we have 14 fixed monitoring stations, we can calculate a total of $14 \times 4 \times 24 = 1344$ correction factors. Section 3.1 provides an example of calculated correction factors for monitoring stations.

Subsequently, we map the daily average $PM_{2.5}$ concentration data of the 1-km grids in the CHAP dataset to the grids in Shanghai and delete the grids without mapping values. Following this, we calculate the average of the grids with multiple mapping values. We regard $\hat{C}_{(x,y),m,d}$ as the daily average $PM_{2.5}$ concentration of the $Grid(x,y)$ in the d-th day, the m-th month. Then for every grid, we use the correction factor of the fixed monitoring station that has the nearest distance to it to correct its $PM_{2.5}$ concentration data and obtain the hourly average $PM_{2.5}$ concentration in this grid. This operation is shown in the following equation:

$$\hat{C}_{(x,y),m,d,h} = CF_{f_{(x,y)},m,h} \times \hat{C}_{(x,y),m,d} \tag{4}$$

Now, we have obtained the $PM_{2.5}$ concentration of 24 h per day from January to April 2014 in Shanghai and the high-resolution (per hour for each 1-km spatial grid) $PM_{2.5}$ data is of great importance to calculate the individual $PM_{2.5}$ exposure.

2.3. Recognizing Individual's Stay Locations

As we introduced above, the CDR data reveals users' mobility behavior. In order to infer the mobility trace of residents, we must know the specific location where the user stayed. There are two types of stay behavior [31]. One is when the user's coordinate is completely kept at the same location for a period of time, which is unusual because even at the same location, the user's mobile phone usually produces slightly different records. The second type of stay behavior is more common and shows that the individual moves or stays within a certain range of the same location, but the presence of different base stations in the vicinity leads to subtle differences in their location record data. Therefore, it is not credible to determine the user's historical stay location only based on their coordinate changes.

In this paper, we utilize the clustering method proposed by Jiang et al. [16] to recognize users' stay behavior based on their CDR data. As Figure 3 shows, we cluster the CDRs from the temporal dimension and spatial dimension respectively. By doing this, we can filter out the disturbance of the user's coordinates among base stations and delete the outliers so that we can cluster different records near the same location into a single point.

Firstly, we apply clustering in the temporal dimension to filter out the disturbance of locations. We cluster the points, which are temporally and spatially close in the record sequence into a single location, and take the difference value between the first record and the last record in the clustering as the dwelling time of this point. For example, assuming that user i has a CDR sequence $D_i = (d_i(1), d_i(2), \ldots, d_i(n_i))$, where $d_i(k) = (t(k), lon(k), lat(k))$ is a 3-tuple recording the timestamp and coordinates of the k-th record. By setting a distance threshold Δd_1 (500 m), we cluster CDRs within the threshold to their center point (the point with the smallest sum of distances to other points), and we calculate the time difference between the earliest record and the latest record as the user's dwelling time at this cluster point. After this process, D_i is transferred to a new sequence $D_i' = (d_i'(1), d_i'(2), \ldots, d_i'(n_i'))$, where $n_i \neq n_i'$, $d_i'(k) = (t(k), dur(k), lon(k), lat(k))$ is a four-tuple to record the arrive time, dwelling time and coordinates after clustering of the k-th record.

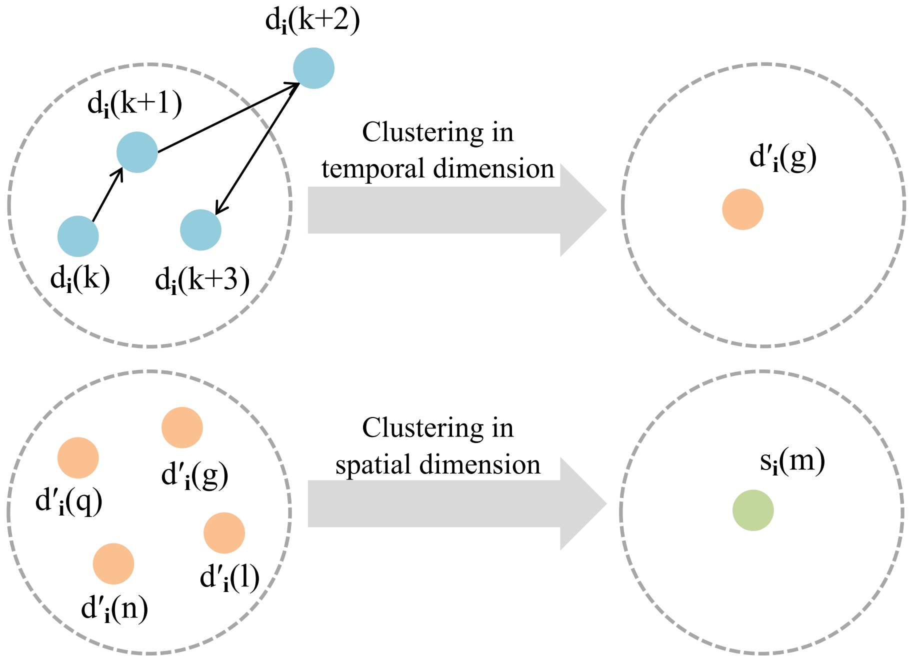

Figure 3. Clustering users' CDRs to recognize their stay behavior.

Then we apply the clustering operation in spatial dimension to filter out outlier locations. Specifically, we set Δd_2 (500 m) as the distance bar to further cluster locations in the CDR sequence. Here we only merge spatial-closed locations and delete records whose dwelling time is less than Δd_t (10 min in this paper) after spatial clustering. Then we get the final stay behaviors $S_i = (s_i(1), s_i(2), \ldots, s_i(m_i))$, where $s_i(k) = (t(k), dur(k), lon(k), lat(k))$. In this way, we finally filter out the locations that users pass by and retain the long-time stay behaviors that are conducive to downstream modeling tasks.

After generating users' daily stay locations from their CDR data, we further identify the location of the users' residences. On the one hand, the location of the user's home is convenient for us to calculate the pollutant exposure based on residence. On the other hand, it is important to understand the location of the home in the mobility trajectory, because the environment and landuse around residences of users affect their daily travel and activities, which is related to their mobility pattern. Assuming that most users go out during the day on weekdays and return home from their workplaces at night, we define the location with the highest frequency of visits on weekday nights and all day on weekends as the user's home location. If a user's total number of visiting 'home', which is calculated by the above rule, is less than 10, then we claim that this user is a short-term visitor of the city and delete his records. Detailed results of this subsection can be found in Section 3.2.

2.4. Calculating Individual Exposure

As we mentioned above, the essence of quantitatively describing exposure is to focus on the concentration of pollutants and the duration of contact [1]. Duan [32] has once provided a method to calculate exposure by linearly combining concentration and dwelling time. In this paper, we implement two methods to quantitatively calculate individual exposure: E^R represents the exposure calculated solely based on the home location of an individual, and E^S represents the exposure calculated based on the mobility behavior of an individual. As the traditional method of estimating the exposure based on residences, E^R only uses the location of users' homes and does not take count of their mobility behavior. The exposure calculation method based on mobility behavior, E^S, takes into account people's stay at various locations during one day and evaluates the impact of staying at a specific location on the exposure. The calculation method of these two exposure metrics of user u in the h hour of day d in month m can be expressed as:

$$E^R_{u,m,d,h} = \hat{C}_{(x^u_R, y^u_R),m,d,h} \times 1 \tag{5}$$

$$E^S_{u,m,d,h} = \sum_{i=1} \hat{C}_{(x^u_{S_i}, y^u_{S_i}),m,d,h} \times \frac{dur_{u,m,d,h}(S_i)}{3600} \tag{6}$$

Here, (x^u_R, y^u_R) represents the grid where user u lives, $(x^u_{S_i}, y^u_{S_i})$ represents the grid where user u stays in his trajectory, $dur_{u,m,d,h}(S_i)$ represents the stay time, which is recorded in second, of user u in the grid within the h hour of day d of month m, and the unit of the final $PM_{2.5}$ exposure is ug·h·m^{-3}.

3. Results

3.1. Example of Correction Factor

After calculating the correction factors in line with the method introduced in Section 2.2, in this subsection we select the calculated correction factors from two monitoring stations and provide further explanation on these correction factor values. We take the JingAn region monitoring station (No. 1147A) located in the city center and the Dianshanhu monitoring station (No. 1146A) located in the suburb as examples to explore their correction factors in these four months. Figure 4 shows the correction factor calculated for the two monitoring stations in different months. Where the y-axis represents $CF_{f,m,h} - 1$, if it is greater than 0, it means that the concentration at that hour needs to be adjusted upward relative to the daily mean, otherwise it means that it needs to be adjusted downward, and 0 indicates no change. Figure 4 shows that there are differences in the adjustment rules of monitoring stations in different months and locations. In general, the correction rate per hour is hardly more than 20%. The monitoring station in the city center will have a peak at around 10:00 am, and the peak time of January and February is later than that of March and April, because the sunrise time is later in winter, and the peak time of $PM_{2.5}$ concentration will be delayed. In addition, the concentration of $PM_{2.5}$ in the suburban monitoring stations fluctuates more than that in the city center during the day, and in the early morning hours (around 0:00 to 5:00), the concentration of $PM_{2.5}$ in each monitoring station is lower than the average except in January.

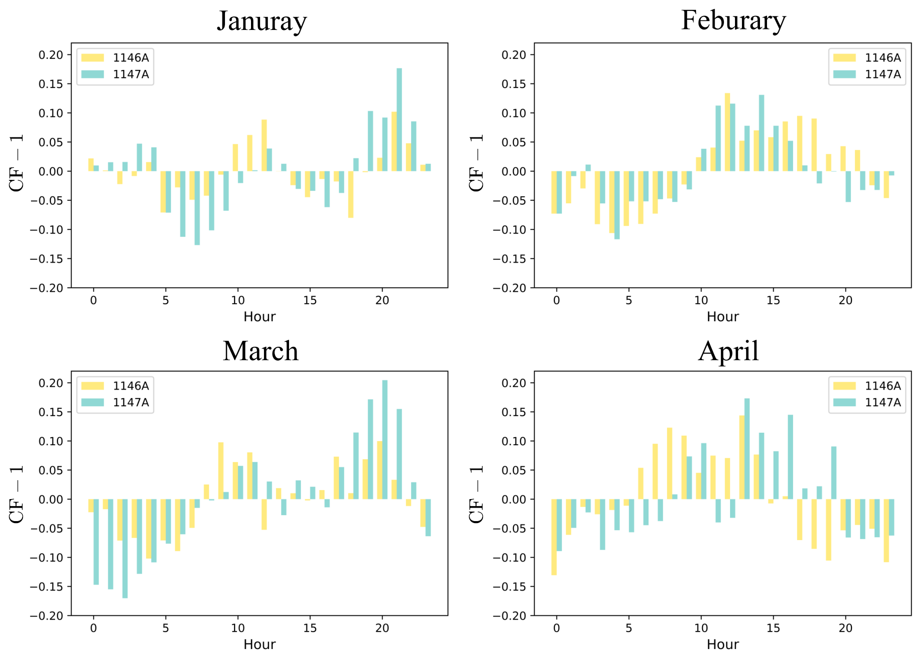

Figure 4. Correction factors for monitoring stations 1146A and 1147A per hour from January 2014 to April 2014

3.2. Stay Behavior and Home Locations

We process the user's raw CDR data according to the clustering method introduced in Section 2.3, filter out outlier records, and recognize the users' stay behavior in their daily travel based on the original CDR data. Figure 5 shows the distribution of the number of users' daily stay locations. We find out that more than 80% residents visit one to three different locations per day and only a small number of users visit more than five locations per day.

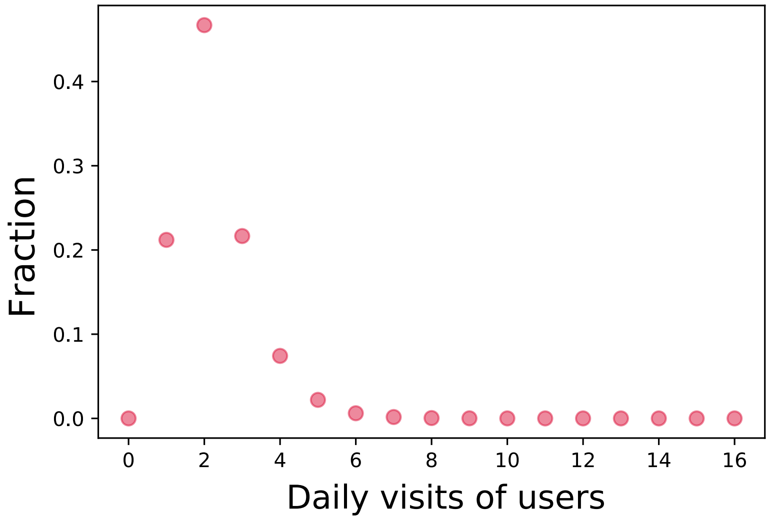

Figure 5. Distribution of the number of locations visited by mobile phone users per day. Each point located at (x, y) represents that the proportion of users who have an average daily record count of x is y relative to the total number of users.

In line with the home location identification method introduced in Section 2.3, we identify the location of users' residences based on their stay behaviors and filter out short-term visitors to Shanghai. After this process, our data finally includes 647,010 users. Subsequently, we calculate the region of the users' homes based on the inferred coordinates of their homes and normalize the number of residences within the administrative district. The distribution of home locations obtained from the CDRs data in Shanghai is shown in Figure 6.

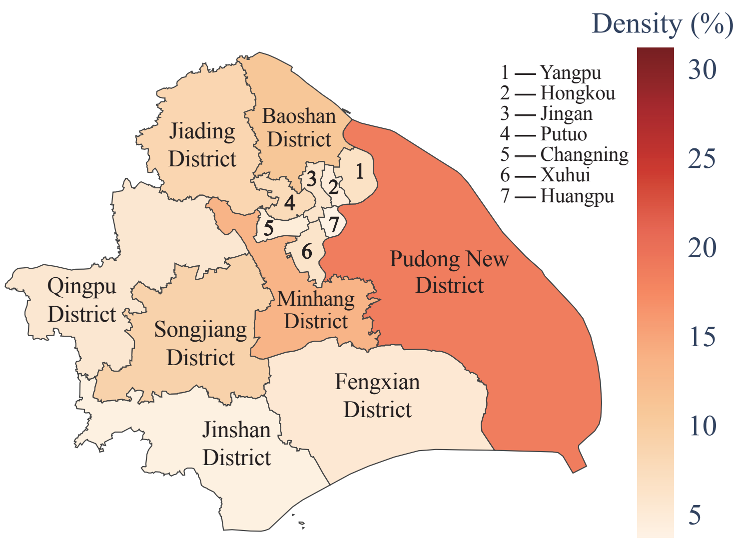

Figure 6. Distribution map of users' home locations. We have normalized user home location counts across districts, showing each district's proportion to Shanghai's total. Darker colors indicate more home locations.

3.3. Calculated $PM_{2.5}$ Exposure and Analysis

Utilizing the methods introduced in the above sections to process the CDR data and $PM_{2.5}$ concentration data, we calculate different exposures according to Equations (5) and (6). The results are shown in Table 3. Where we find that the average individual exposure calculated by users' home location is less than that based on stay behaviors.

Table 3. The average $PM_{2.5}$ exposure of different calculation methods.

$PM_{2.5}$ Exposure ($ug{\cdot}h/m^3$)	January		February		March		April		Overall	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Home location based	75.325	43.561	52.234	32.620	57.107	27.449	52.115	18.016	60.328	33.763
Stay behavior based	75.350	43.764	52.196	32.718	57.136	27.682	52.263	18.167	60.374	33.925

Subsequently, we further analyzed the differences of individual exposure in the spatial and temporal dimensions. First, we analyzed the differences in hourly average $PM_{2.5}$ exposure over 24 h every day. We calculated the distribution of the average exposure of all users at different times based on the hourly stay behavior of 647,010 users and the spatiotemporal $PM_{2.5}$ data with high resolution. Results are shown in Figure 7.

Results in Figure 7 reveal that the average $PM_{2.5}$ exposure of users is on the high side during the morning peak and evening peak (9 am to 11 am and 6 pm to 8 pm) periods of each day. We suggest that one conceivable reason is that high emissions from automobile exhaust during commuting hours result in high airborne $PM_{2.5}$ concentrations, and thus the calculated exposure level is higher than the daily average exposure level.

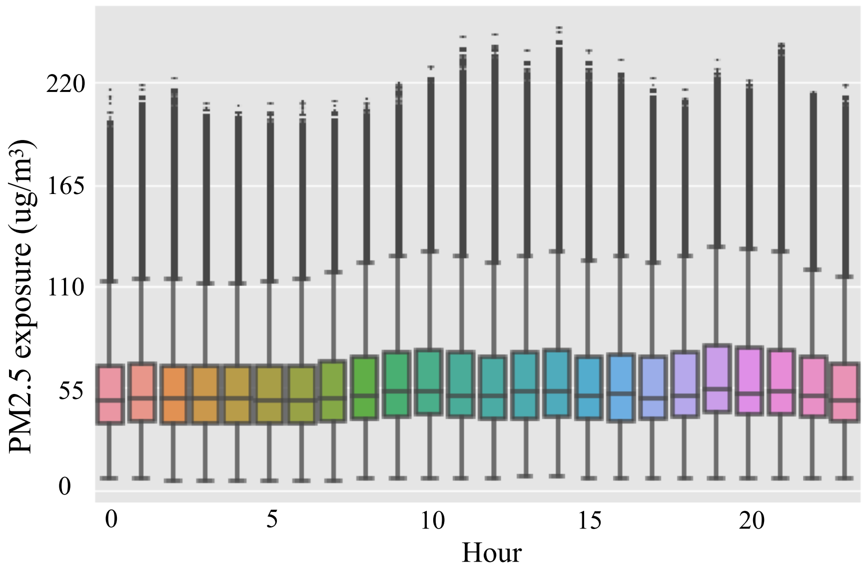

Figure 7. The box plot of $PM_{2.5}$ exposure of all mobile phone users per hour per day. (The colors of the different boxes are automatically designated by the plotting library and do not convey any particular meaning in this figure.)

Moreover, we analyzed the differences of individual $PM_{2.5}$ exposure in different months. We calculated the average hourly exposure of all users in different months and visualized this part of the data in Figure 8. We found that the average exposure level in January was significantly higher than the other months because the overall $PM_{2.5}$ concentration in January was higher. The average exposure levels in February and March are similar, and the interquartile range in April is the shortest, indicating that there is no significant difference in the average hourly exposure of all residences within April.

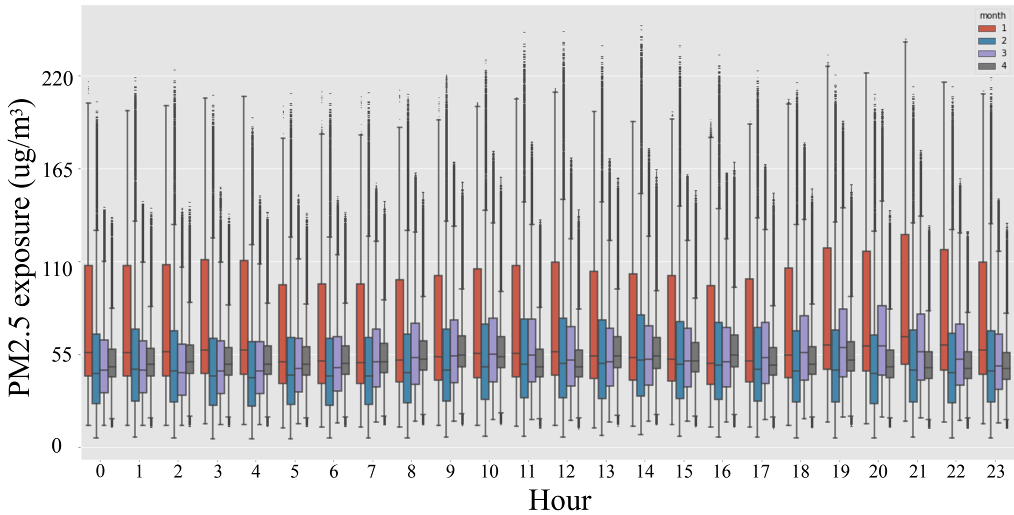

Figure 8. The box plot of $PM_{2.5}$ exposure of all mobile phone users per hour per month.

For the spatial difference analysis, we compared the average $PM_{2.5}$ exposure levels of residents living in different regions. We mapped each user's home location to the corresponding grid and then calculated the average exposure of all users living in that grid under different computational methods. These results shown in Figure 9 reveal that the individual $PM_{2.5}$ exposure in Shanghai has a decreasing trend from west to east, which is relatively close to the distribution of $PM_{2.5}$ concentration.

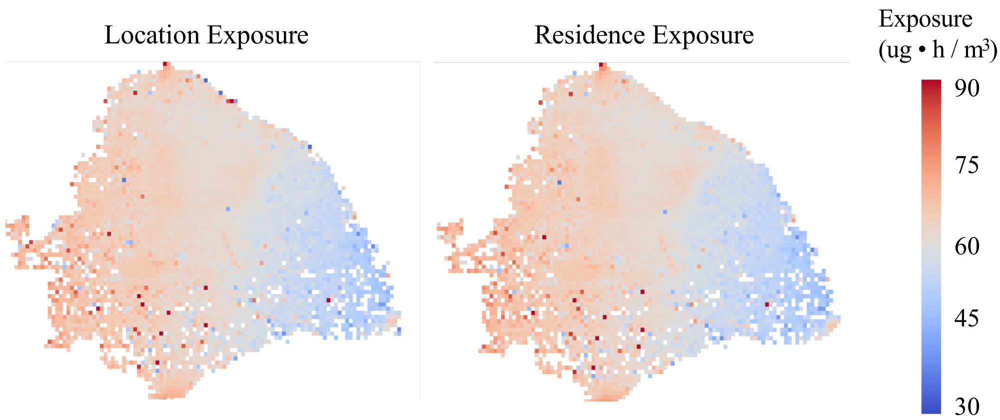

Figure 9. Spatial distribution of $PM_{2.5}$ exposure under different calculation methods

4. Discussion

In our experimental segment, we calculated the exposure suffered by residents at the individual level. In addition to calculating the average exposure of all residents, we also conducted further analysis of the individual exposure in the temporal and spatial dimensions. We also analyzed the results obtained from the two different methods of calculating exposure. In Figure 10, we not only show the discrepancy between two different exposure estimating methods but also illustrate the geographical environment around Shanghai. Additionally, We collect data on population, industry, and urban construction for each administrative district in Shanghai from the official website of the Shanghai Bureau of Statistics (https://tjj.sh.gov.cn/tjnj/20170629/0014-1000201.html accessed on 9 December 2023) for the year 2014. These data are exhibited in Figure 11.

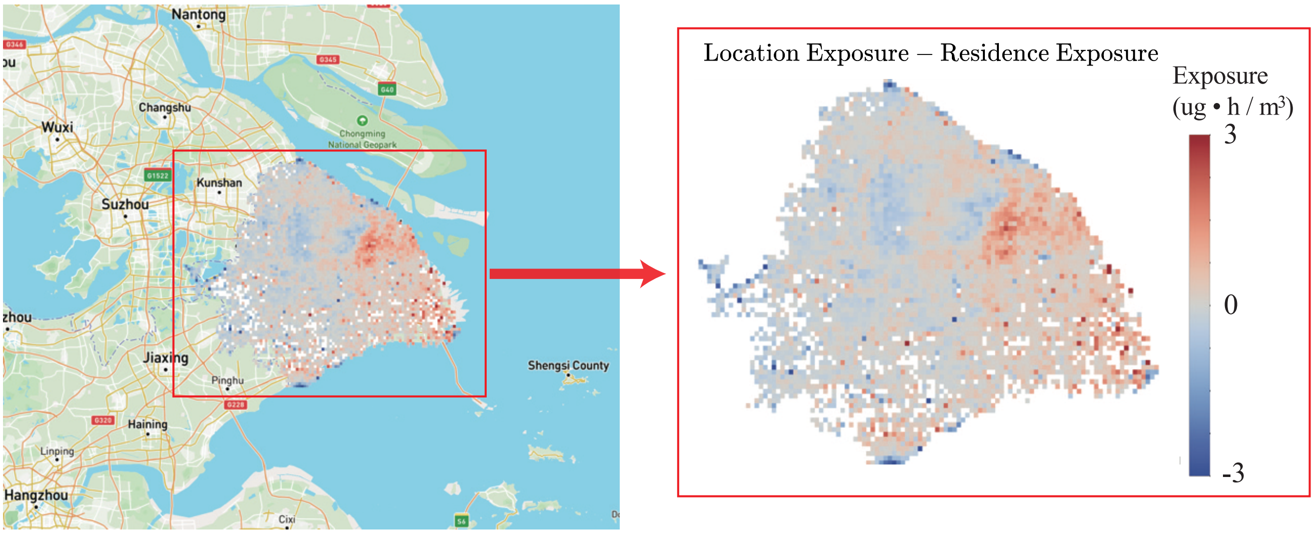

Figure 10. The spatial distribution of the difference in $PM_{2.5}$ exposure under different calculation methods, coupled with the geographical environment map of the vicinity surrounding Shanghai.

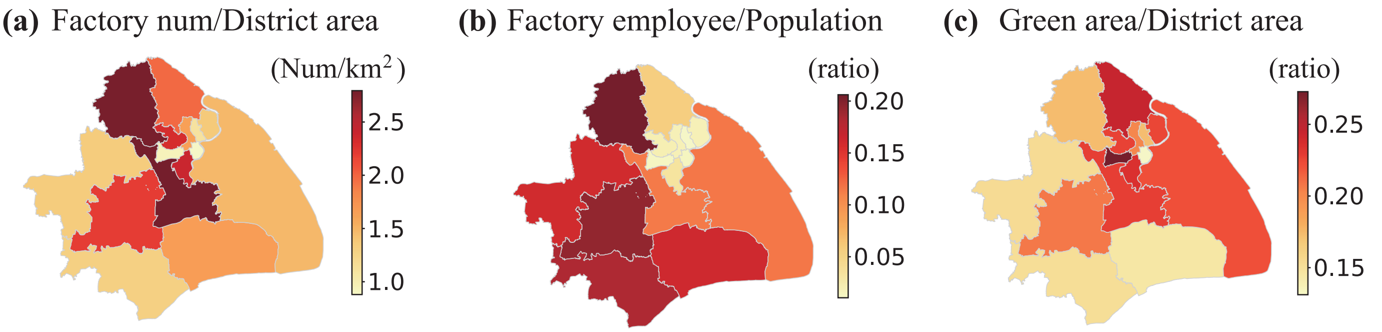

Figure 11. (**a**) The density, which is denoted in units per square kilometer, of industrial enterprises. This value is calculated by dividing the number of industrial enterprises in the district by the area of the district. (**b**) The proportion of the population employed by industrial enterprises to the total population in different districts. (**c**) The proportion of green space area to the total area in different districts.

The result in Figure 10 indicates that the exposure calculated based on stay behavior is slightly higher than that calculated based on the residence in eastern Shanghai. However, the outcome is the contrary in western Shanghai. The spatial distribution of pollutant concentration in Shanghai is the main reason for this result. Figure 11a shows that the density of industrial enterprises in the eastern region of Shanghai is relatively low, and Figure 11c reveals a high proportion of green space in the same area. Moreover, according to the geographical location of Shanghai shown in Figure 10, the eastern part of Shanghai is near the sea and can benefit from sea breezes. These factors provide a good explanation for the west-high and east-low trend in the spatial distribution of $PM_{2.5}$ concentration in Shanghai as shown in Figure 2.

Due to the influence of this pollutant distribution trend, when we estimate the $PM_{2.5}$ exposure, we always find that residents living in the eastern part of Shanghai suffer lower levels of $PM_{2.5}$ exposure than those living in the western part of Shanghai, irrespective of whether their spatial movement was considered or not. Residents in the Pudong new area have the lowest average $PM_{2.5}$ exposure per hour, and residents in the Jinshan district have the highest $PM_{2.5}$ exposure. Although mobility behaviors can somewhat reduce the effect of the residential environment on individual $PM_{2.5}$ exposure to be closer to the overall average, our experimental results show that the individual $PM_{2.5}$ exposure in Shanghai is still highly correlated with the residential environment. If the concentration of $PM_{2.5}$ around the user's residence is high, then his overall exposure to $PM_{2.5}$ is high. And vice versa, if the concentration of $PM_{2.5}$ around the user's residence is low, his exposure is low.

Specifically, from the temporal perspective, our experimental results indicate that individuals have a relatively high exposure to $PM_{2.5}$ during morning and evening rush hours. Therefore, the government can introduce policies to encourage the public to travel green during commuting by implementing measures such as moderate vehicle restrictions and constructing more bicycle lanes. Besides, the pollution exposure levels in January are significantly higher than in February, March, and April, highlighting the severity of pollution problems in winter. From the perspective of environmental sustainability, therefore, it's necessary to promote winter pollution prevention and control initiatives further. For example, government departments can actively promote to residents the use of renewable energy sources such as biomass, solar, and geothermal energy for heating, so as to reduce coal burning. From the spatial perspective, as we have introduced above, the $PM_{2.5}$ exposure levels of residents in Shanghai display a distinct pattern of higher in the west and lower in the east. One reason is that the pollutant concentration is relatively higher in the western part of Shanghai. Additionally, as shown in Figure 11b, the proportion of residents engaged in industrial production is higher in the western part of Shanghai. These residents are exposed to higher levels of pollutants during their daily work, which is also a reason for the higher average exposure level in the western part of Shanghai. Therefore, policy-making should prioritize addressing pollution control efforts in the western regions of Shanghai, reducing the generation of pollutants such as $PM_{2.5}$ from the source. Residents living in the western region can also consider equipping their homes with air purifiers to mitigate the health impacts of pollution exposure.

Furthermore, we noticed a limitation during the collection of concentration data from stationary monitoring stations in Shanghai. These pollutant concentration monitoring stations are mainly concentrated in the city center, while the suburban areas lack monitoring stations. This imbalance may affect the accuracy of the overall study results. Therefore, we believe that cities should pay attention to the balance of monitoring station selection when setting up pollutant monitoring stations. Pollutant concentrations in urban centers are highly variable and higher on average, which is why monitoring stations are more concentrated in urban centers. However, monitoring pollutants in suburban areas can help researchers better study the overall distribution of pollutants and the exposure of residents. Therefore, additional pollutant monitoring stations in suburban areas can facilitate air pollution research and provide more reliable health guidance to residents living in the suburbs. At the same time, low-cost air quality sensors [33,34] might present a significant solution to the uneven spatial distribution of monitoring stations.

5. Conclusions

In this paper, we initially apply a clustering method to recognize users' stay behavior based on their CDR data and propose a reasonable approach to estimate the high-resolution $PM_{2.5}$ concentration in every 1-km grid every 1-h slot. Subsequently, we propose a big data analysis framework for individual exposure estimation, which is the main work of this paper. This framework can quantify large-scale estimation of individual exposure based on users' stay behaviors and high-resolution $PM_{2.5}$ concentration data.

When it comes to future work, we believe it is possible to further differentiate user dwell behavior, calculate different exposure levels for indoor and outdoor spaces, and even more accurately assess the impact of whether the user wears a mask on exposure estimation. Additionally, we believe that we can further improve the precision of individual exposure estimation by calculating the exposure of an individual's transition based on the detailed travel behavior of the user. However, these efforts require finer-grained user behavioral data and detailed trajectory data. In a word, this paper proposed a novel individual exposure estimation framework, offering fresh viewpoints and substantiating data to guide the development of environmental policies for mitigating individual-level pollutant exposure.

Author Contributions: Data curation, L.H.; methodology, L.H. and Y.X.; validation, L.H. and Z.H.; formal analysis, L.H. and Z.H.; investigation, L.H. and Z.H.; resources, X.Z., T.Y. and Y.J.; writing—original draft preparation, Z.H. and L.H.; writing—review and editing, Z.H., L.H. and Y.X.; visualization, L.H. and Z.H.; supervision, Y.J. and Y.X.; funding acquisition, X.Z., T.Y., Y.J. and Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was jointly supported by the National Key Research and Development Program (2022YFC3303102), the National Natural Science Foundation of China (62102258), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Watson, A.Y.; Bates, R.R.; Kennedy, D. Assessment of human exposure to air pollution: Methods, measurements, and models. In *Air Pollution, the Automobile, and Public Health*; National Academies Press: Washington, DC, USA, 1988.
2. Lepeule, J.; Laden, F.; Dockery, D.; Schwartz, J. Chronic exposure to fine particles and mortality: An extended follow-up of the Harvard Six Cities study from 1974 to 2009. *Environ. Health Perspect.* **2012**, *120*, 965–970. [CrossRef] [PubMed]
3. Deryugina, T.; Heutel, G.; Miller, N.H.; Molitor, D.; Reif, J. The mortality and medical costs of air pollution: Evidence from changes in wind direction. *Am. Econ. Rev.* **2019**, *109*, 4178–4219. [CrossRef] [PubMed]
4. Torretta, V.; Tolkou, A.; Katsoyiannis, I.; Schiavon, M. Second-hand smoke exposure effects on human health: Evaluation of PM10 concentrations in the external areas of a university campus. *Sustainability* **2020**, *12*, 2948. [CrossRef]
5. Morello-Frosch, R.; Pastor, M.; Sadd, J. Environmental justice and Southern California's "riskscape" the distribution of air toxics exposures and health risks among diverse communities. *Urban Aff. Rev.* **2001**, *36*, 551–578. [CrossRef]
6. Buzzelli, M.; Jerrett, M.; Burnett, R.; Finklestein, N. Spatiotemporal perspectives on air pollution and environmental justice in Hamilton, Canada, 1985–1996. *Ann. Assoc. Am. Geogr.* **2003**, *93*, 557–573. [CrossRef]
7. Brulle, R.J.; Pellow, D.N. Environmental justice: Human health and environmental inequalities. *Annu. Rev. Public Health* **2006**, *27*, 103–124. [CrossRef]
8. Bravo, M.A.; Anthopolos, R.; Bell, M.L.; Miranda, M.L. Racial isolation and exposure to airborne particulate matter and ozone in understudied US populations: Environmental justice applications of downscaled numerical model output. *Environ. Int.* **2016**, *92*, 247–255. [CrossRef]
9. Fotheringham, A.S.; Wong, D.W. The modifiable areal unit problem in multivariate statistical analysis. *Environ. Plan. A* **1991**, *23*, 1025–1044. [CrossRef]
10. Kwan, M.P. The neighborhood effect averaging problem (NEAP): An elusive confounder of the neighborhood effect. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1841. [CrossRef]
11. Park, Y.M.; Kwan, M.P. Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. *Health Place* **2017**, *43*, 85–94. [CrossRef]
12. Chang, Y.S.; Lee, Y.J.; Choi, S.S.B. Is there more traffic congestion in larger cities? Scaling analysis of the 101 largest US urban centers. *Transp. Policy* **2017**, *59*, 54–63. [CrossRef]
13. Marshall, J.D.; Granvold, P.W.; Hoats, A.S.; McKone, T.E.; Deakin, E.; Nazaroff, W.W. Inhalation intake of ambient air pollution in California's South Coast Air Basin. *Atmos. Environ.* **2006**, *40*, 4381–4392. [CrossRef]
14. Hajat, A.; Diez-Roux, A.V.; Adar, S.D.; Auchincloss, A.H.; Lovasi, G.S.; O'Neill, M.S.; Sheppard, L.; Kaufman, J.D. Air pollution and individual and neighborhood socioeconomic status: Evidence from the Multi-Ethnic Study of Atherosclerosis (MESA). *Environ. Health Perspect.* **2013**, *121*, 1325–1333. [CrossRef] [PubMed]
15. Alexander, L.; Jiang, S.; Murga, M.; González, M.C. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 240–250. [CrossRef]
16. Jiang, S.; Ferreira, J.; Gonzalez, M.C. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Trans. Big Data* **2017**, *3*, 208–219. [CrossRef]
17. Huang, L.; Xia, F.; Chen, H.; Hu, B.; Zhou, X.; Li, C.; Jin, Y.; Xu, Y. Reconstructing human activities via coupling mobile phone data with location-based social networks. *Travel Behav. Soc.* **2023**, *33*, 100606. [CrossRef]

18. Xu, Y.; Jiang, S.; Li, R.; Zhang, J.; Zhao, J.; Abbar, S.; González, M.C. Unraveling environmental justice in ambient $PM_{2.5}$ exposure in Beijing: A big data approach. *Comput. Environ. Urban Syst.* **2019**, *75*, 12–21. [CrossRef]
19. Guo, H.; Li, W.; Yao, F.; Wu, J.; Zhou, X.; Yue, Y.; Yeh, A.G. Who are more exposed to $PM_{2.5}$ pollution: A mobile phone data approach. *Environ. Int.* **2020**, *143*, 105821. [CrossRef]
20. Chaix, B.; Gustafsson, S.; Jerrett, M.; Kristersson, H.; Lithman, T.; Boalt, Å.; Merlo, J. Children's exposure to nitrogen dioxide in Sweden: Investigating environmental injustice in an egalitarian country. *J. Epidemiol. Community Health* **2006**, *60*, 234–241. [CrossRef]
21. Gray, S.C.; Edwards, S.E.; Miranda, M.L. Race, socioeconomic status, and air pollution exposure in North Carolina. *Environ. Res.* **2013**, *126*, 152–158. [CrossRef]
22. Collins, T.W.; Grineski, S.E. Environmental injustice and religion: Outdoor air pollution disparities in metropolitan Salt Lake City, Utah. *Ann. Am. Assoc. Geogr.* **2019**, *109*, 1597–1617. [CrossRef]
23. Samoli, E.; Stergiopoulou, A.; Santana, P.; Rodopoulou, S.; Mitsakou, C.; Dimitroulopoulou, C.; Bauwelinck, M.; de Hoogh, K.; Costa, C.; Marí-Dell'Olmo, M.; et al. Spatial variability in air pollution exposure in relation to socioeconomic indicators in nine European metropolitan areas: A study on environmental inequality. *Environ. Pollut.* **2019**, *249*, 345–353. [CrossRef] [PubMed]
24. Richardson, E.A.; Pearce, J.; Tunstall, H.; Mitchell, R.; Shortt, N.K. Particulate air pollution and health inequalities: A Europe-wide ecological analysis. *Int. J. Health Geogr.* **2013**, *12*, 1–10. [CrossRef] [PubMed]
25. Smith, J.D.; Mitsakou, C.; Kitwiroon, N.; Barratt, B.M.; Walton, H.A.; Taylor, J.G.; Anderson, H.R.; Kelly, F.J.; Beevers, S.D. London hybrid exposure model: Improving human exposure estimates to NO_2 and $PM_{2.5}$ in an urban setting. *Environ. Sci. Technol.* **2016**, *50*, 11760–11768. [CrossRef] [PubMed]
26. Fan, X.; Lam, K.c.; Yu, Q. Differential exposure of the urban population to vehicular air pollution in Hong Kong. *Sci. Total Environ.* **2012**, *426*, 211–219. [CrossRef] [PubMed]
27. Huang, G.; Zhou, W.; Qian, Y.; Fisher, B. Breathing the same air? Socioeconomic disparities in $PM_{2.5}$ exposure and the potential benefits from air filtration. *Sci. Total Environ.* **2019**, *657*, 619–626. [CrossRef] [PubMed]
28. Zhao, X.; Cheng, H.; He, S.; Cui, X.; Pu, X.; Lu, L. Spatial associations between social groups and ozone air pollution exposure in the Beijing urban area. *Environ. Res.* **2018**, *164*, 173–183. [CrossRef]
29. Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications. *Remote Sens. Environ.* **2021**, *252*, 112136. [CrossRef]
30. Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; et al. Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space–time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273–3289. [CrossRef]
31. Zheng, Y. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol. (TIST)* **2015**, *6*, 1–41. [CrossRef]
32. Duan, N. Models for human exposure to air pollution. *Environ. Int.* **1982**, *8*, 305–309. [CrossRef]
33. Day, R.F.; Yin, P.Y.; Huang, Y.C.T.; Wang, C.Y.; Tsai, C.C.; Yu, C.H. Concentration-Temporal Multilevel Calibration of Low-Cost $PM_{2.5}$ Sensors. *Sustainability* **2022**, *14*, 10015. [CrossRef]
34. deSouza, P.N. Key concerns and drivers of low-cost air quality sensor use. *Sustainability* **2022**, *14*, 584. [CrossRef]