www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Urban link travel speed dataset from a megacity road network

Feng Guo, Dongqing Zhang, Yucheng Dong & Zhaoxia Guo

Received: 4 October 2018

Accepted: 3 April 2019

Published online: 16 May 2019

Link travel speeds in road networks are fundamental data in many research areas of traffic, transportation, and logistics. To support the research in these areas, we develop a dataset, containing the travel speeds on each road link and in different time periods together with the real road network map. The dataset is collected from a representative megacity in Western China, Chengdu. The road network of this city involves different urban road network structures. The dataset shows the realistic variations and randomness of urban link travel speeds. This enables the research of real data-driven decision-making problems in traffic, transportation and logistics areas.

Background & Summary

With the rapid advancement of Global Positioning System (GPS) and information technology, the traffic and transportation sector is experiencing a massive increase in the amount of traffic data (e.g., vehicles' travel trajectory data) collected. More and more real data-based studies were conducted and reported in recent years. It is reported that real data-based research could result in 30% time reduction in congestions[1], 5% carbon emission reduction[2], or 30% reduction in fleet size[3] in the road network.

In the research areas of traffic, transportation and logistics, the travel speeds or travel times on road links are fundamental data for various decision-making problems, such as (a) traffic assignment[4,5], (b) vehicle routing[2,6,7], (c) ridesharing[8,9], and (d) the fleet minimization problem[3] in urban road network. These problems are generally defined in a road network, the weights of whose links are traditionally travel speeds or corresponding travel times or costs. However, to the best of the authors' knowledge, no publicly available travel speeds dataset suitable for these decision-making problems has been reported.

Various randomnesses exist in travel speeds on real-world urban links. It has been reported that travel speeds could follow different probability distributions[10–12], and there exist spatial and temporal correlations between travel speed on different links and in different time periods[13–15]. It is thus critical to share and publish the link travel speed dataset with real-world distributions and correlations.

On the other hand, there exist various road network structures in the real world, such as modified linear, branch, grid, 3-directional grid, 1-ring web, and 2-ring web[16]. To make the dataset more representative, it is crucial to collect the data from a city with different road network structures.

This research takes Chengdu, a megacity in Western China, as the case city, and shares the link travel speed dataset from its road network. The dataset contains the link travel speed data from June 1, 2015 to July 15, 2015 on each link and in different time periods together with the Chengdu road network map.

To obtain the link travel speed dataset, we first collect the real-time GPS trajectory data of floating vehicles in Chengdu. Then, we perform map matching to output the projected paths of the trajectories on the map and estimate the travel speeds on each link in different time periods based on the map matching results. Finally, we check the data for errors, and validate the variations and randomness of link travel speeds.

The main purpose of publishing this dataset is to facilitate real data-driven research on decision-making problems in traffic, transportation and logistics areas. Moreover, the dataset can be used in various other scenarios as well. For instance, it can be used as input data to forecast the vehicle travel speeds or travel times in urban road network[17]. The data can reflect the real traffic conditions and enable to identify the congestions[18].

Methods

Figure 1 shows the flowchart of methodology to obtain the link travel speeds in Chengdu's road network. The steps involved are described in detail below.

Business School, Sichuan University, Chengdu, 610065, China. Correspondence and requests for materials should be addressed to Z.G. (email: zx.guo@alumni.polyu.edu.hk)

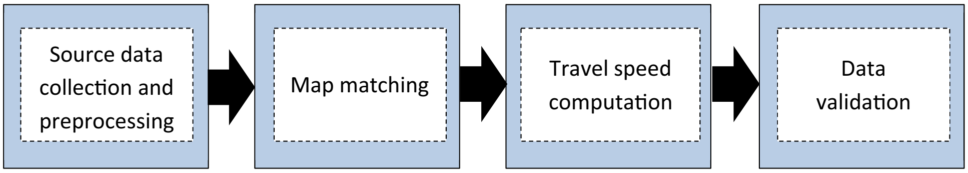

Fig. 1 Flowchart of methodology. The figure shows the flowchart of methodology to obtain the link travel speeds in Chengdu's road network.

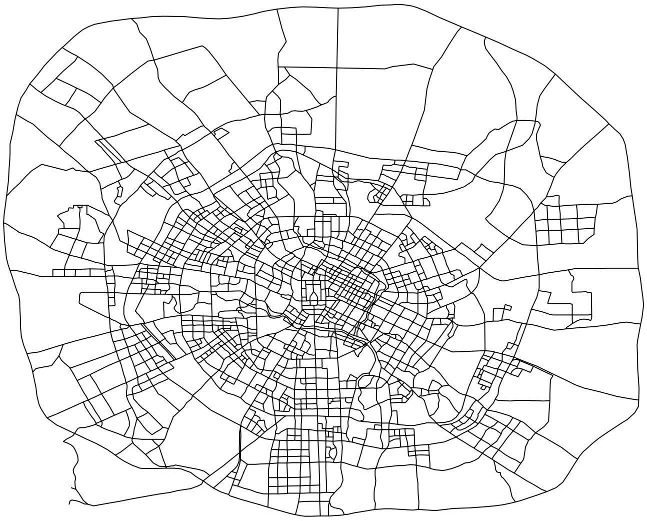

Fig. 2 Road Network of Chengdu. The figure shows the road network of Chengdu within the ring expressway, which consists of 1,902 nodes and 5,943 directed links.

Step 1. Source data collection and preprocessing. *Source data collection.* The source data for the link travel speed dataset consist of road network data and GPS trajectory data of floating vehicles. Based on OpenStreetMaps data, we use the method proposed by Karduni *et al.*[19] to obtain the road network data, which contain the road network topology and the length of each link. The road network of Chengdu within the ring expressway is shown in Fig. 2, which consists of 1,902 nodes and 5,943 directed links. We don't consider those links with few or no GPS trajectories in the road network. The trajectory data of floating vehicles are usually collected by the GPS-enabled devices installed in each floating vehicle during specified time intervals. This research collects the GPS trajectory data of taxis in Chengdu, China. Each trajectory sample (record) consists of the geographic location in latitude and longitude, taxi status, real-time travel speed and sampling time. All taxis use the same type of GPS-enabled devices, which ensures that the trajectory samples collected from different taxis have the same precision. The sampling rate of trajectories keeps unchanged, which is once per 10 seconds. The status of a taxi in operation switches between vacant and occupied when the taxi picks up or drops off passengers. Forty-five-day data, from 0:00 on June 1, 2015, to 23:59 on July 15, 2015, are collected. These data contain a total of 3.01 billion raw GPS trajectory samples produced by a total of more than 12,000 taxis during the data collection period.

Preprocessing of trajectory data. GPS signals in taxi operations could be affected or even blocked sometimes due to various electromagnetic signal shielding and interference in the city. As a result, some abnormal trajectory samples could be collected and included in the source data. It is thus crucial to preprocess the raw GPS trajectory data before map matching. On the basis of the analysis on our raw GPS trajectory data, we consider the following three types of trajectory samples as abnormal samples, and remove them from the source data.

1. Trajectory samples without key information, including taxi ID and speed values.
2. Trajectory samples with a latitude or longitude of 0°.
3. Trajectory samples having the same location information but different speed values with their proceeding trajectory points.

Step 2. Map matching. Reliable map matching is critical to identify the accurate location of the vehicles, which outputs the projected paths of the trajectories on the map. This research performs the map matching based on the method proposed by Li *et al.*[20]. This method is selected because (1) this method is one of the most cited ones in recent years that can handle both low-frequency and high-frequency probe data of large data size well, and (2) the effectiveness of this method has been validated by comparing with the hidden Markov model-based and other map matching algorithms. The main steps are described as follows.

1. Generating the set S of all order-k links from the given Chengdu map M. Each vertex v in the road map has one or more order-k links. k is a given distance constant. A path is called as vertex $v's$ order-k link if it satisfies the following conditions: (a) it starts from the vertex v, (b) it is composed of n non-repeating links $l_1, l_2, \ldots, l_n$ such that $\sum_{i=1}^{n-1} d^{l_i} < k$ and $\sum_{i=1}^{n} d^{l_i} \geq k$, where d^{l_i} denotes the length of link l_i, and (c) all vertices on this path are not revisited. The larger the value of k is, the higher the map matching accuracy is[20]. But the number of segments increases drastically as k increases, which leads to the much higher computation cost. We have compared the results at different k and set $k = 0.8$ after considering the trade-off between map matching accuracy and computational efficiency. The map matching accuracy is 95.7% at $k = 0.8$ in terms of the accuracy criterion used in Li *et al*.'s paper[20].
2. Joint link selection. This step assigns an order-k link from S to each trajectory sample point. Let t denote a travel trajectory, which is a set of τ trajectory sample points. We have $t = \{a_{t,1}, a_{t,2}, \ldots, a_{t,\tau}\}$ where $a_{t,i}$ is the i^{th} trajectory sample point in trajectory t. The trajectory t is considered as a sparse and noisy sampling of an underlying path on map M. Note that the underlying path may start and end in the middle of links of M. We define the matching distance $d^m(a_{t,i}, o)$ between a trajectory sample point $a_{t,i}$ and an order-k link o as

$$d^m(a_{t,i}, o) = \max\left\{d^e(a_{t,i}, o),\ \max_{\rho\in o} d^e(\rho, t)\right\} \tag{1}$$

where $d^e(a_{t,i}, o)$ and $d^e(\rho, t)$ denote the Euclidean distances between the sample points ($a_{t,i}$ and ρ) and its closest point in the piecewise linear curve (o and t). While it is straightforward to use the distance between a sample point and an order-k link, we add a regularizing term $\max_{\rho\in o} d^e(\rho, t)$ to reflect the consistency between order-k link o and trajectory t. If $d^e(a_{t,i}, o)$ is small but $\max_{\rho\in o} d^e(\rho, t)$ is very large, order-k link o is likely incompatible globally to trajectory t because the former only indicates the local compatibility between point $a_{t,i}$ and order-k link o. We allocate each trajectory sample point to the order-k link with the minimal matching distance.
3. Map matching using the selected order-k links. For each trajectory, a path is constructed from the selected order-k links, and take the links with trajectory sample points as this trajectory's projected path. If two adjacent links l_i and l_{i+1} in the projected path are not directly connected on map M, we use the shortest path between l_i and l_{i+1} to connect them.

Step 3. Travel speed computation. This step is to compute the travel speed on each link in each time period based on the map matching results, which involves 3 procedures: postprocessing of trajectory data, link travel speed estimation and speed data imputation.

Postprocessing of trajectory data. This procedure is to remove the trajectory samples that could decline the accuracy of travel speed estimation. We consider the following three cases.

1. Sometimes several continuous trajectory samples from a taxi have the same location and the speed value of 0 over a time period due to temporary parking. In this case, these trajectory samples are removed; otherwise, the link travel speed obtained will be smaller than the actual link travel speed.
2. The travel speeds of taxis cannot reflect the real link travel speeds during the periods of picking up or dropping off passengers. We thus find trajectory samples collected from taxis that have status switches between vacant and occupied in each period. Among the samples found, those with a speed less than the current reference speed on the link in each period are removed. That is, we only remove low-speed samples from taxis with status changes although there may be many taxis in the traffic stream that have speeds less than the reference speed. The reference speed is set as the average speed of other taxis without status changes on the same link in the same or adjacent periods.
3. The travel speeds on a link could have a very high variance in a time period due to irrational driving behaviors (e.g., speeding and stopping) or data capture errors. Speed v is considered as an outlier if $|v - \text{median}(\mathbf{v})| \geq n \cdot \text{median}(\mathbf{v})$, where $|x|$ represents the absolute value of x, $\text{median}(\mathbf{v})$ represents the median of speeds of all samples on this link in the trajectory. By analysing the samples on randomly chosen 1,189 links (20% of 5,943 links), we find that setting $n = 0.59$ is able to remove outliers most effectively. This research thus takes $n = 0.59$. The corresponding trajectory sample is removed if its speed is an outlier.

Link travel speed estimation. The travel speeds on a link are estimated based on the location, sampling time and real-time travel speed information of trajectory sample points on the link. The sampling time indicates the time when the record is sampled, while the real-time travel speed indicates the instant travel speed of the taxi recorded by the GPS-enabled devices installed in the taxi. This research sets two minutes as one period, during which about 50,000–60,000 trajectory sample points are usually collected according to the real raw data. Thus, on average, there are about 8–10 points collected on each link in one period. In addition, the average link length is about 436.6 m in the network and the lengths of 76.7% links are less than 500 m. Considering the short-link feature, an average of 8–10 or less sample points are usually sufficient to estimate the travel speed on each link in one period. We estimate the travel speeds on a link from the following two perspectives.

On one hand, according to the method proposed by Quiroga and Bullock[21], we estimate the travel distance of single vehicle on the link by using the real-time travel speed and sampling time information of trajectory samples, thereby get the corresponding travel speed of each vehicle on the link. We then obtain the travel speed on the link in each time period by averaging the speeds of all vehicles on the link in the same period.

Consider a link l. The vehicles travelled on the link are indexed by j ($j = 1, \ldots, J$), and the trajectory sample points on the link are indexed by k ($k = 1, \ldots, K$). Let $p^l_{j,k}$, $t^l_{j,k}$, $v^l_{j,k}$ denote the position, time, and speed of vehicle j's k^{th} trajectory sample on link l, respectively.

We can calculate the travel distance d^t_j of vehicle j between $t^l_{j,1}$ and $t^l_{j,K}$ as follows.

$$d^t_j = \int_{t^l_{j,1}}^{t^l_{j,K}} v dt \approx v^l_{j,1}\left(\frac{t^l_{j,2} - t^l_{j,1}}{2}\right) + \sum_{k=2}^{K-1} v^l_{j,k}\left(\frac{t^l_{j,k+1} - t^l_{j,k-1}}{2}\right) + v^l_{j,K}\left(\frac{t^l_{j,K} - t^l_{j,K-1}}{2}\right) \tag{2}$$

If the first and the last (K^{th}) trajectory samples on the link are close to the two ends of the link, d^t_j approximates the length of link l. Then, we can use Eqs (3) and (4) to estimate the travel speed v^l_j of vehicle j and the link travel speed v^l_{avg1}.

$$v^l_j = \frac{d^t_j}{t^l_{j,K} - t^l_{j,1}} \tag{3}$$

$$v^l_{\mathrm{avg1}} = \frac{1}{J}\sum_{j=1}^{J} v^l_j \tag{4}$$

The travel speed v^l_{avg1} obtained above is usually smaller than the actual link travel speed. The reason is simple. The time interval for sampling two continuous trajectory points is 10 s in our source data, so the lasting time of travel speed $v^l_{j,k}$ ($2 \le k \le K - 1$) is 10 s as well according to Eq. (2). However, in the real world, the taxi tends to travel at the maximum speed allowed on links and low-speed travels usually last for a short time period. Therefore, the lasting time of 10 s tends to lengthen the low-speed travel time and lower down the travel speeds obtained.

On the other hand, we can first obtain the time of entering and exiting the link of each vehicle based on the location and sampling time information of trajectory samples. Then we estimate the travel speed of each vehicle v^l_j on the link l, and obtain the travel speed v^l_{avg2} on the link by averaging the speeds of all vehicles on the link in the same time period[21].

Let $v^l_{j,1}$ and $v^l_{j,K}$ denote the travel speeds of vehicle j passing through the entry and exit ends of link l respectively. We use Eqs (5) and (6) to obtain the time of vehicle j arriving the entrance and exit ends of link l, $t^l_{j,\mathrm{ent}}$ and $t^l_{j,\mathrm{exit}}$ respectively,

$$t^l_{j,\mathrm{ent}} = t^l_{j,1} - \frac{d^l_{j,\mathrm{ent}}}{v^l_{j,1}} \tag{5}$$

$$t^l_{j,\mathrm{exit}} = t^l_{j,K} + \frac{d^l_{j,\mathrm{exit}}}{v^l_{j,K}} \tag{6}$$

where $d^l_{j,\mathrm{ent}}$ denotes the distance between the entrance end and $p^l_{j,1}$, and $d^l_{j,\mathrm{exit}}$ denotes the distance between $p^l_{j,K}$ and the exit end of link l. Then, we estimate v^l_j and v^l_{avg2} by Eqs (7) and (8).

$$v^l_j = \frac{d^l}{t^l_{j,\mathrm{exit}} - t^l_{j,\mathrm{ent}}} \tag{7}$$

$$v^l_{\mathrm{avg2}} = \frac{1}{J}\sum_{j=1}^{J} v^l_j \tag{8}$$

where d^l denotes the length of link l. The travel speed v^l_{avg2} obtained above is usually larger than the actual link travel speed. The reason is simple. According to Eqs (5) and (6), $t^l_{j,\mathrm{ent}}$ tends to be larger and $t^l_{j,\mathrm{exit}}$ tends to be smaller because the vehicle's actual speeds at the ends of link are usually less than $v^l_{j,1}$ and $v^l_{j,K}$ due to the effects of traffic signals and turning vehicles. Thus, v^l_j and v^l_{avg2} tend to be larger.

To reduce the calculating deviation of travel speed values generated by the above two methods, we use Eq. (9) to compromise both values and take the final value as the link travel speed v^l.where w is weight coefficient and we set $w = 0.6$ based on the analysis of a large number of speeds on urban links.

$$v^l = w v^l_{\mathrm{avg1}} + (1 - w) v^l_{avg2} \tag{9}$$

Name	Description
link.csv	Road network file.
speed_[date]_[i].csv	Travel speed files. The value of '[date]' part of the filename is the real date when the travel speed data saved in the file are collected and the value of the '[i]' part is 0 or 1. '[0]' and ['1] correspond to time periods 1–150 and time periods 151–300, respectively. For instance, speed_[601]_[0].csv refers to the travel speed data file for the first 150 periods on June 1, 2015. Forty-five separate speed files are included, which correspond to speed data from June 1 to July 15, 2015.

Table 1. Data files of the dataset.

Column	Data type	Description
Link	integer	Link No. of the directed link between Node_Start and Node_End.
Node_Start	integer	Node No. of the starting node of a link.
Longitude_Start	float	The longitude of the starting node.
Latitude_Start	float	The latitude of the starting node
Node_End	integer	Node No. of the end node of a link.
Longitude _End	float	The longitude of the end node.
Latitude _End	float	The latitude of the end node.
Length	float	The length of the link (unit: m).

Table 2. Summary of fields in link.csv.

Speed data imputation. Some links cannot match with appropriate trajectory points in a specific time period in the map matching step because no trajectory samples are collected on these links in that period. As a result, we cannot obtain their travel speeds in the last procedure. Thus, the following steps are performed in turn until a valid speed value on the link is generated.

1. Obtain speed values on this link in the previous and next two periods of a current period, and take the median of these speed values as the speed on this link in the current period. This approach is called as the temporal imputation approach.
2. Obtain speed values on the immediately adjacent links (with the same direction) of this link in the same time period, and take the median of these speed values as the speed on this link in this period. This approach is called as the spatial imputation approach.
3. Obtain historical speed values on this link in the same period but in neighboring dates, and take the median of these speed values as the speed on this link in this period.

To justify the ordering of using steps 1–2, we have compared the performances of the temporal imputation approach and the spatial imputation approach. We choose out all links (1907 in total), whose travel speeds in 60 different periods are computed without using the imputation process. Then we use both approaches to generate the supplemented average speeds on these links, and compare the relative deviations of both supplemented speeds to the computed speeds. We find that the temporal approach leads to the less relative deviations for 75.6% cases. This research thus uses the temporal approach first for speed data imputation.

Of course, it is possible that there exist some approaches that work better for the data imputation on some links. We do not claim we use the best approach for data imputation, which is not the focus of this paper. Some deviations in speed estimation are inevitable and acceptable. The resulting speed values could be considered as possible real-world realizations due to the randomness and diversity of real world. After all, the dataset is reliable as long as the dataset can pass appropriate technical validation check.

Step 4: Data validation. We perform validation steps for the link travel speed dataset obtained. Please see Section "Technical Validation" for more details.

Data Records

The link travel speed dataset[22] is located in figshare, which is available as 46 separate csv files described in Table 1.

link.csv: This file contains the data of road network topology and the length of each link within the ring expressway of Chengdu. Relevant fields are listed out in Table 2.

speed_[date]_[i].csv: This file contains the data of link travel speeds within the ring expressway of Chengdu in different time periods of a specified date. We only obtained the link speed of 5 representative time horizons, including 3:00–5:00, 8:00–10:00, 12:00–14:00, 17:00–19:00, and 21:00–23:00. These time horizons involve rush hours, normal hours and night hours. Each time period takes 2 minutes. Thus, there are 300 time periods in total in the 5 time horizons. Relevant fields are listed out in Table 3.

Technical Validation

This section is to validate if the link travel speed dataset can reflect the real-world link travel speeds in the road network. We validate the speed dataset by integrating numerical comparison and disciplinary analysis from the following three aspects.

Column	Data type	Description
Period	string	Time period, represented by its start and end time range.
Link	integer	Link No. of link corresponding to the file link.csv.
Speed	float	Travel speed on the link in a specific time period (unit: km/h).

Table 3. Summary of fields in speed_[date]_[i].csv.

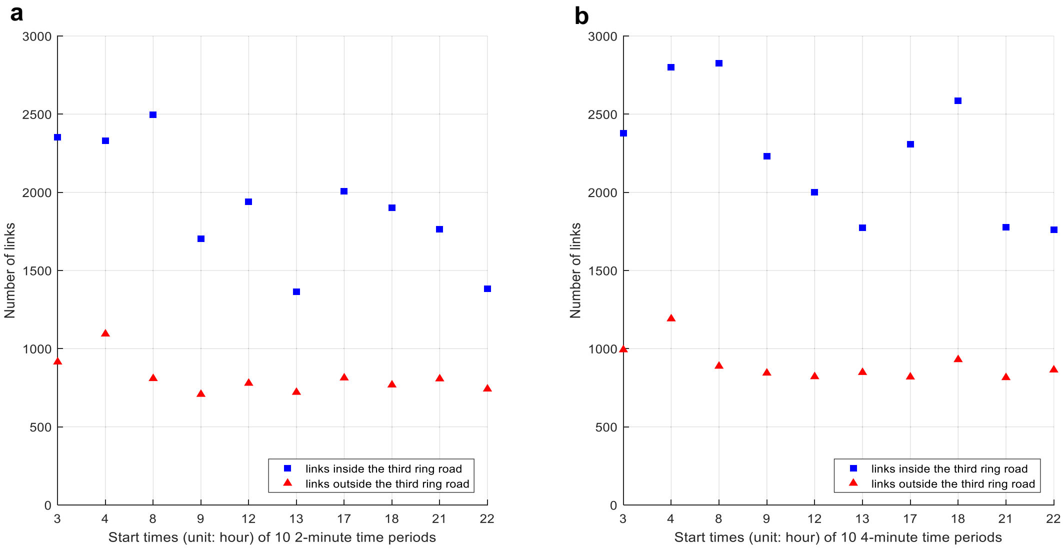

Fig. 3 Number of links on which the temporal correlations of travel speeds are bigger than 0.5 between each of 10 time periods and its next. Subfigures **a** and **b** show the results of 2-minute and 4-minute time periods respectively. Taking subfigure **a** as an example, the two points in the longitudinal axis mean that the speeds on 2,353 links inside the third ring road and 915 links outside the third ring road have temporal correlations bigger than 0.5 between two consecutive 2-minute time periods starting at 3 am, respectively.

Sanity check. The first aspect of technical validation is to detect the actual errors in the link travel speed dataset. We first check that the calculations for travel speeds are inerrant. Then, we check that there are no missing or redundant speeds in the obtained dataset. Next, we check that the speed range in the dataset is valid by examining the largest speed values. We find that only 2 speed values are higher than 140 km/h in the dataset. These large speed values are valid since both of them are collected on airport expressway links during 4:00–5:00.

Validation on variations of travel speeds. The travel speed variations are examined by observing the speeds on 100 representative links and in three different time periods, 8:00–8:02, 12:00–12:02 and 17:00–17:02. These links contain 50 downtown links located within the first ring road and 50 suburban links around the ring expressway of Chengdu. In the three time periods, the speed values (km/h) on the downtown links range within [5.15, 58.75], [6.00, 55.90] and [5.38, 56.25] respectively, while those on the suburban links range within [5.35, 107.65], [5.68, 110.45] and [7.70, 112.20] respectively. It is clear that the travel speeds on the downtown links have smaller upper limits and fluctuate within much smaller ranges than those on the suburban links. It is because there is a relatively fast and smooth traffic flow on suburban links.

Validation on distributions and correlations of travel speeds. It has been reported that the urban link travel speeds obey the normal and lognormal distributions[12]. We fit all random speed variables (5,943 × 300 = 1,782,900 in total) with the normal and lognormal distributions using the maximum likelihood estimation method. On the basis of the one-sample Kolmogorov-Smirnov test[23] with a significance level of 0.05, we find that 97.81% (1,743,920) of them obey the normal distributions and 1.53% (27,213) fit the lognormal distributions, whereas there are only 0.66% (11,767) random speed variables fitting neither distribution. These results are consistent with Wang *et al.*'s findings[12].

We further examine the spatial and temporal correlation of travel speeds by calculating the Pearson correlation coefficients. Firstly, we observe the correlations of travel speeds on each link and its neighbouring links in a same time period, which are called as spatial correlations. Taking the speeds in a morning time period (8:00–8:02) as an example, we investigate the correlations of travel speeds on all links and their neighbouring links directly connected in this time period. Based on the Fisher transformation[24] with a significance level of 0.05, we find that the speeds on 49.75% links have significant correlations with the speeds on their neighbouring links in the same direction, and the speeds on 70.68% links have insignificant correlations with the speeds on their neighbouring links in the reverse direction. Speeds in other time periods have similar results. These results are easy-to-understand, and similar correlation findings have been reported in the literature[13–15].

Next, we investigate the correlations between the travel speed on each link in each time period and the speed on the same link but in its adjacent time period, which are so-called temporal correlations. We consider 2 different time period lengths, i.e., 2 and 4 minutes, and the results are shown in Fig. 3. It can be found that, with the increase of time period length, significant temporal correlations of travel speeds on more links can be observed, especially in morning and evening rush hours. It is because the travel speeds collected from a short time period tend to exhibit intense fluctuations and noise[25], and this would weaken the temporal correlation of the travel speeds on some links. Moreover, the number of links with strong temporal correlations outside the third ring road are less than those inside the third ring road in all 10 time periods. It indicates that the travel speeds between two consecutive time periods exhibit the stronger temporal correlation in busy traffic areas, which is in line with Rachtan *et al.*'s findings[14]. The above observations validate the temporal correlations of travel speeds in our dataset.

Usage Notes

Since all data files are provided as csv files, the urban link travel speeds can be analysed and processed using many pieces of software, such as Pyhton, Matlab, and R. As described in Table 1, the road network data and the travel speed data are separated into different files, thus before using the data to study some decision-making problems, these two types of files need to be integrated together according to the link numbers shown in link.csv to get the travel speeds on each link in each time period. In addition, the speed data correspond to 5 representative time horizons, including 3:00–5:00, 8:00–10:00, 12:00–14:00, 17:00–19:00, and 21:00–23:00. The numbers of nodes are set to be dispersed in the road network, which could be changed to form smaller road networks.

Code Availability

We cannot provide access to the raw source data due to their proprietary nature. As mentioned in Step 1 of the Methods section, the source data mainly contain a total of 3.01 billion GPS trajectory samples produced by more than 12,000 taxis during 45 days. As stated by Poulis *et al.*[26], the publication of the trajectories of personal movement could lead to identity disclosure, even if directly identifying information (e.g., names of taxi drivers and passengers) is not published. Moreover, existing trajectory anonymization techniques[26,27] cannot be used in our research because existing techniques do not care about travel speeds in trajectories and do not need the information of taxi status. However, to obtain the travel speed dataset accurately, we have to use the information of taxi status (as described in Step 3) to indicate when each taxi picks up or drops off passengers.

Python (version 2.7.12) is used to produce the link travel dataset in this research. We have not shared the code because the code is dedicatedly designed for our raw source data and researchers cannot benefit from the code without the source data. Meanwhile, the code might reveal the identity of taxi drivers and raw real-time trajectory information of taxis in the road network. However, the code is straightforward, and its steps have been described in detail in the section of 'Methods'. It is easy for a third party to exactly repeat the method.

References

1. Çolak, S., Lima, A. & González, M. C. Understanding congested travel in urban areas. *Nat. Commun.* **7**, 10793 (2016).
2. Ehmke, J. F., Campbell, A. M. & Thomas, B. W. Data-driven approaches for emissions-minimized paths in urban areas. *Comput. Oper. Res.* **67**, 34–47 (2016).
3. Vazifeh, M. M., Santi, P., Resta, G., Strogatz, S. & Ratti, C. Addressing the minimum fleet problem in on-demand urban mobility. *Nature* **557**, 534 (2018).
4. Nie, Y. A note on Bar-Gera's algorithm for the origin-based traffic assignment problem. *Transport. Sci* **46**, 27–38 (2012).
5. Nikolova, E. & Stier-Moses, N. E. A mean-risk model for the traffic assignment problem with stochastic travel times. *Oper. Res.* **62**, 366–382 (2014).
6. Long, J., Huang, H. J., Gao, Z. & Szeto, W. Y. An intersection-movement-based dynamic user optimal route choice problem. *Oper. Res.* **61**, 1134–1147 (2013).
7. Huang, H. & Gao, S. Trajectory-adaptive routing in dynamic networks with dependent random link travel yimes. *Transport. Sci* **52**, 102–117 (2017).
8. Ordóñez, F. & Dessouky, M. M. *Dynamic Ridesharing*, https://doi.org/10.1287/educ.2017.0167/ (2017).
9. Furuhata, M. *et al.* Ridesharing: The state-of-the-art and future directions. *Transport. Res. B-Meth.* **57**, 28–46 (2013).
10. Dey, P. P., Chandra, S. & Gangopadhaya, S. Speed distribution curves under mixed traffic conditions. *J. Transp. Eng* **132**, 475–481 (2006).
11. Park, B. J., Zhang, Y. & Lord, D. Bayesian mixture modeling approach to account for heterogeneity in speed data. *Transport. Res. B-Meth* **44**, 662–673 (2010).
12. Wang, Y. *et al.* Speed modeling and travel time estimation based on truncated normal and lognormal distributions. *Transport. Res. Rec.* **2315**, 66–72 (2012).
13. Cheng, T., Haworth, J. & Wang, J. Spatio-temporal autocorrelation of road network data. *J. Geogr. Syst.* **14**, 389–413 (2012).
14. Rachtan, P., Huang, H. & Gao, S. Spatiotemporal link speed correlations: Empirical study. *Transport. Res. Rec* **2390**, 34–43 (2013).
15. Ermagun, A., Chatterjee, S. & Levinson, D. Using temporal detrending to observe the spatial correlation of traffic. *Plos One* **12**, e0176853 (2017).
16. Amini, B., Peiravian, F., Mojarradi, M. & Derrible, S. Comparative analysis of traffic performance of urban transportation systems. *Transport. Res. Rec* **2594**, 159–168 (2016).
17. Min, W. & Wynter, L. Real-time road traffic prediction with spatio-temporal correlations. *Transport. Res. C-Emer* **19**, 606–616 (2011).
18. Cheng, T., Tanaksaranond, G., Brunsdon, C. & Haworth, J. Exploratory visualisation of congestion evolutions on urban transport networks. *Transport. Res. C-Emer* **36**, 296–306 (2013).
19. Karduni, A., Kermanshah, A. & Derrible, S. A protocol to convert spatial polyline data to network formats and applications to world urban road networks. *Sci. Data* **3**, 160046 (2016).
20. Li, Y., Huang, Q., Kerber, M., Zhang, L. & Guibas, L. Large-scale joint map matching of GPS traces. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 214–223 (ACM, 2013).

21. Quiroga, C. A. & Bullock, D. Travel time studies with global positioning and geographic information systems: An integrated methodology. *Transport. Res. C-Emer* **6**, 101–127 (1998).
22. Guo, F., Zhang, D., Dong, Y. & Guo, Z. Urban link travel speed dataset from a megacity road network. *figshare*, https://doi.org/10.6084/m9.figshare.7140209.v4 (2018).
23. Massey, F. J. Jr. The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* **46**, 68–78 (1951).
24. Fisher, R. A. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3–32 (1921).
25. Vlahogianni, E. & Karlaftis, M. Temporal aggregation in traffic data: Implications for statistical characteristics and model choice. *Transport. Lett.* **3**, 37–49 (2011).
26. Poulis, G., Skiadopoulos, S., Loukides, G. & Gkoulalas-Divanis, A. Apriori-based algorithms for k^m-anonymizing trajectory data. *Transactions on Data Privacy* **7**, 165–194 (2014).
27. Terrovitis, M., Poulis, G., Mamoulis, N. & Skiadopoulos, S. Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Trans. Knowl. Data Eng.* **29**, 1466–1479 (2017).

Acknowledgements

The authors thank the Transport Committee of Chengdu for providing the raw taxi trajectory data. The authors thank the support from the National Natural Science Foundation of China through the general project (No. 71872118), the Ministry of Education in China through the project of Humanities and Social Sciences (No. 18YJC630045), and Sichuan University through projects (Nos 2018hhs-37, sksyl201819, skqx201725).

Author Contributions

F.G. and Y.D. wrote the code used for generating the dataset. D.Z. evaluated the dataset and tested it for errors. Z.G. defined the dataset and provided scientific guidance throughout the dataset generation and validation. F.G. and Y.D. contribute equally and share the first authorship. All authors participated in writing and revising the paper.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.