# Driving Big Data : A First Look at Driving Behavior via a Large-Scale Private Car Dataset

Li, Tong

IEEE
2019-04

Li , T , Alhilal , A , Zhang , A , Hoque , M A , Chatzopoulos , D , Xiao , Z , Li , Y & Hui , P 2019 , Driving Big Data : A First Look at Driving Behavior via a Large-Scale Private Car Dataset . in 2019 IEEE 35th International Conference on Data Engineering Workshops : ICDEW 2019, Proceedings . IEEE ... International Conference on Data Engineering Workshop , IEEE , pp. 61-68 , IEEE International Conference on Data Engineering Workshops , Macau , China , 08/04/2019 . https://doi.org/10.1109/ICDEW.2019.00-34

http://hdl.handle.net/10138/328919
https://doi.org/10.1109/ICDEW.2019.00-34

unspecified
acceptedVersion

# Driving Big Data: A First Look at Driving Behavior Via a Large-scale Private Car Dataset

Tong Li[†], Ahmad Alhilal[†], Anlan Zhang[¶], Mohammad A. Hoque[‡],
Dimitris Chatzopoulos[†], Zhu Xiao[£], Yong Li[§], Pan Hui[†‡]
[†] Department of Computer Science & Engineering, The Hong Kong University of Science and Technology
[¶] School of Computer Science & Engineering, Beihang University
[‡] Department of Computer Science, University of Helsinki
[£] College of Computer Science & Engineering, Hunan University
[§] Department of Electronic Engineering, Tsinghua University
E-mail: {t.li@connect.ust.hk, aalhilal@ust.hk, zal1506@buaa.edu.cn, mohammad.a.hoque@helsinki.fi,
dcab@cse.ust.hk, zhxiao@hnu.edu.cn, liyong07@tsinghua.edu.cn, panhui@cse.ust.hk }

***Abstract*—The increasing number of privately owned vehicles in large metropolitan cities have contributed to congestion, increased energy waste due to congestion, raised $CO_2$ emissions, and impacted our living conditions negatively. Analysis of data representing human mobility and citizens' driving behavior can provide insights to reverse these conditions. This article presents a large-scale driving status and trajectory dataset consisting of 426,992,602 records collected from 68,069 vehicles over a month. From the dataset, we analyze the driving behavior and produce random distributions of trip duration and millage to characterize the car trips. We have found that a private car has more than 17% probability to make four trips per day, and a trip has more than 25% probability to last 20-30 minutes and 33% probability to travel 10 Kilometers during the trip. The collective distributions of trip mileage and duration follow Weibull distribution, whereas the hourly trips follow the well known diurnal pattern and so the hourly fuel efficiency. Based on these findings, we have developed an application which recommends the drivers to find the nearby gas stations and possible popular places from the historical trips. We further highlight that our dataset can be applied for developing dynamic Green maps for fuel efficient routing, modeling efficient Vehicle-2-Vehicle (V2V) communications protocols, verifying existing V2V protocols, and understanding user behavior in driving their private cars.**

***Index Terms*—Green Map, Trajectories, Fuel Efficiency, Vehicular Communications, $CO_2$ emission.**

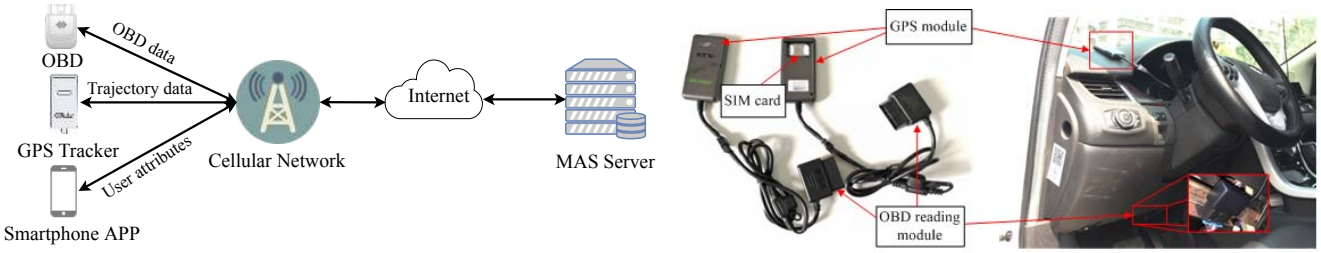
## I. Introduction

The continuously increasing number of automotive vehicles in the cities have been changing the travel experience of citizens. On one hand, they have become a daily necessity that facilitates people's modern lives. On the other hand, their fastest growth has caused a series of problems in urban areas, such as increased traffic accidents, traffic jams, and environmental pollution. These severe problems have triggered multiple actions of urban planning and governmental regulations, and both the research community and the industry have been looking for opportunities to solve these problems [1–3].

Although fast-growing vehicles bring lots of problems for modern big cities, their produced vast volumes of trajectory data provide valuable information to investigate and understand driving behavior and bring a new opportunity and perspective to solve these tricky problems. For example, Wang et al., [4] collected 4.8 TB taxi GPS data from 2013 to 2018 and comprehensively investigated the evolving patterns of electric taxi networks and paved the way for future shared autonomous vehicles to mitigate traffic jams and air pollution in urban areas.

Nevertheless, the existing datasets are merely based on data collected from public transport or floating cars, e.g., buses and taxis [4–6]. These datasets are constrained by predetermined trajectories (buses) or points of interest (taxis). Based on these existing datasets, we cannot grasp the typical driving behavior of citizens, i.e., the owners of private cars. The private cars refer to a class of small motor vehicles that are registered by individuals and for personal use. In some cities, private cars account for 80% of vehicles in China [7]. In terms of a report from the University of Michigan Transportation Research Institute, the average number of private cars per household is up to 1.97 in US [8]. Although private cars take the most significant part of automotive vehicles in modern cities, rare studies focus on this 'elephant in the room.'

A high-quality private car trajectory dataset has important application values, since it offers an effective way to understand not only the traffic dynamics in cities to improve the transport services but also the driving behavior of individuals. In this work, we first build a platform to collect the trajectory data for privately owned vehicles (POVs). Specifically, we use GPS trackers and on-board diagnostics (OBD) monitors to record trajectories and driving status of POVs. Up to now, our platform supports over 68,069 POVs in the mainland of China. We present and analyze a subset of our whole dataset, which covers one month from 1st of July 2016 to 31st of July 2016. In detail, we compare citizen's driving behavior during weekdays and weekends. We find that the number of trips on weekdays is less than that on weekends, while there is no big difference regarding trip durations and trip mileage between weekdays and weekends. We further extract the probability distributions regarding the characteristics of the recorded trips. Both the duration and mileage of all trips follow Weibull distributions, while they follow Gaussian distribution for each user. These distributions also suggest that a private car has more than 17% probability to make four trips per day, and a trip has more than 25% probability to last 20-30 minutes and 33% probability to travel 10 Kilometers during the trip. The average fuel efficiency




(a) Framework

(b) Hardware instruments

Fig. 1: The MAS framework and hardware instruments installed in the private cars.

TABLE I: The important attributes of private car trips collected by the MAS platform.

| Item | Description | Data type | | | Responsible device | | |
|---|---|---|---|---|---|---|---|
| | | Trips | Trajectories | Driving status | OBD | GPS | SIM Card |
| `ObjectID` | ID of the vehicle | | | | | | ✓ |
| `StartTime` | Local time when start engine | ✓ | | | ✓ | ✓ | |
| `StopTime` | Local time when stop engine | ✓ | | | ✓ | ✓ | |
| `StartLon` | Longitude of the vehicle when engine starts | ✓ | | | ✓ | ✓ | |
| `StartLat` | Latitude of the vehicle when engine starts | ✓ | | | ✓ | ✓ | |
| `StopLon` | Longitude of the vehicle when engine stops | ✓ | | | ✓ | ✓ | |
| `StopLat` | Latitude of the vehicle when engine stops | ✓ | | | ✓ | ✓ | |
| `TripMileage` | Mileage of the trip | ✓ | | | ✓ | ✓ | |
| `TripOil` | Fuel consumption of the trip | ✓ | | | ✓ | ✓ | |
| `TripPeriod` | Duration of the trip | ✓ | | | ✓ | ✓ | |
| `Lon` | Longitude of the vehicle | | ✓ | | | ✓ | |
| `Lat` | Latitude of the vehicle | | ✓ | | | ✓ | |
| `GPSStatus` | Status of GPS signal | | ✓ | | | ✓ | |
| `GPSTime` | Local time | | ✓ | | | ✓ | |
| `Speed` | Instantaneous speed of the vehicle | | | ✓ | ✓ | | |
| `Direct` | Current direction of the vehicle | | | ✓ | ✓ | | |
| `Mileage` | Current mileage of the vehicle | | | ✓ | ✓ | | |
| `OilNum` | Current volume of vehicle fuel | | | ✓ | ✓ | | |
| `AlarmDesc` | Description of alarms | | | ✓ | ✓ | | |
| `RPM` | Revolutions per minute of the vehicle | | | ✓ | ✓ | | |
| `AccPos` | Position of the accelerator pedal | | | ✓ | ✓ | | |

of the trips peaks at around 5 AM while is of the minimum at near 10 AM.

We also developed an application which recommends the nearby gas stations and possible popular places. The recommendation is based on the recent fuel status of the car and the historical trips of an individual. We further discuss that our datasets can be applied for developing dynamic Green maps for fuel efficient routing, modeling efficient Vehicle-2-Vehicle (V2V) communication protocols, verifying V2V protocols through simulation, and understanding user behavior in driving their private cars. It is worth mentioning that the data collection process is ongoing and we present only a fraction of the total dataset in this work.

The rest of this paper is organized as follows. We present the data collection process in the Section (Section II). Next, in Section III we list the basic characteristics of the dataset presented in this work, and we discuss the driving behavior of the users and characterize the fuel efficiency of the trips. After that, in Section IV we describe the applicability of the introduced dataset and present three use cases. Section V concludes this work.

## II. Data Collection, Attrobutes, and Privacy

To collect the trajectories of POVs, we cooperate with Mapgoo[1] and build a cloud platform, called Mapgoo Automotive Services (MAS), for vehicle networks.

### A. Car Instrumentation

Fig. 1(a) depicts the MAS platform. MAS collects data from both hardware, i.e., OBD and GPS trackers, and software, i.e., a smartphone app. They upload data to the MAS server via cellular networks.

Up to now, MAS platform covers 68,069 cars. As shown in Fig. 1(b), each car is installed with a lightweight and low-cost On-board Diagnostics (OBD) monitor which is compatible with ISO 14230 (KWP2000) and Society of Automobile Engineers (SAE) protocols [9]. OBD monitors are in the charge of recording the status of vehicle subsystems, such as the engine, braking system, cooling system, and the electronic control module. In addition, we have installed a Global Positioning System (GPS) tracker which is a cheap commercial GPS receiver (ublox LEA-6T) [10] and a communication unit with

[1] http://www.mapgoo.net/html/MAS.aspx

TABLE II: A sample trip record in the dataset.

| ObjectID | StartTime | StopTime | StartLon | StartLat | StopLon | StopLat | TripMileage | TripOil | TripPeriod |
|---|---|---|---|---|---|---|---|---|---|
| 556605 | 01/07/2016 09:27:33 | 01/07/2016 11:10:00 | 109.822249 | 40.641596 | 110.522649 | 40.597578 | 85,611 | 6,979 | 6,147 |

TABLE III: A sample trajectory and the driving condition record in the dataset.

| ObjectID | Lon | Lat | GPSStatus | GPSTime | Speed | Direct | Mileage | AlarmDesc | RPM | AccPos |
|---|---|---|---|---|---|---|---|---|---|---|
| 556605 | 109.822249 | 40.641596 | Strong (9) | 01/07/2016 09:27:33 | 0 | 10 | 6,383 | None | 0 | 0 |

a SIM card. The GPS receiver is responsible for collecting vehicle trajectories, and the communication unit is used to upload the data collected by the OBD monitor and GPS receiver to the MAS servers. Thanks to the development of cellular networks. These data can be uploaded real time and with low latency. The International Mobile Equipment Identity (IMEI) number is used as the unique ID for each vehicle and is one-to-one mapped to a bit string as an anonymized ID (`ObjectID`) for privacy protection. We also developed a smartphone application to help the users to manage their trips and fuel consumption efficiently (Section III-B).

### B. Dataset Attributes

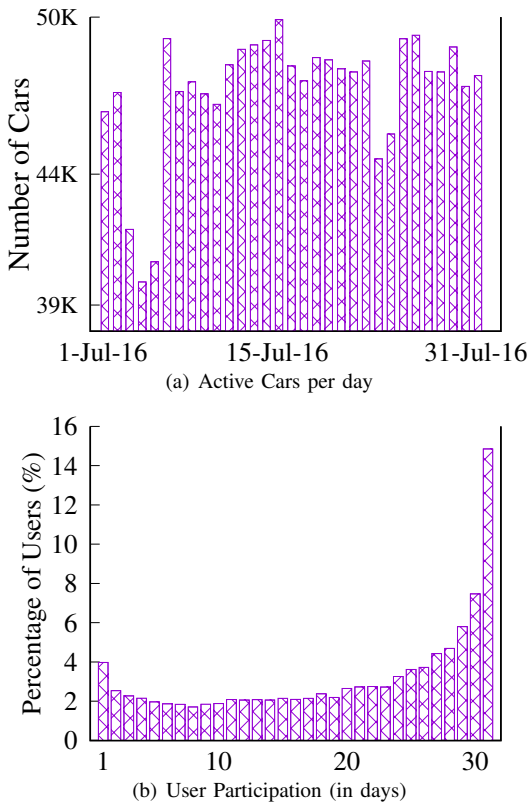
In detail, we capture three types of data, *trips*, *trajectories*, and *driving status*, which are summarized in Table I.

1) *Trips*. Both OBDs and GPS trackers collect the trips data which include information of the start and stop time (`StartTime`, `StopTime`) the start and stop locations (`StartLon`, `StartLat`, `StopLon`, and `StopLat`), the millage (`TripMileage`) in meters, the duration (`TripPeriod`) in seconds, and fuel consumption (`TripOil`) in milliliters.
2) *Trajectories*. The trajectories are collected by the GPS trackers and uploaded to the MAS cloud server after every 30 seconds. Apart from vehicle locations (`Lon`, `Lat`), GPS trackers also upload the time (`GPSTime`) and the status of the GPS signals (`GPSStatus`) that can help us to detect the outliers and improve the accuracy of the collected trajectories. The trajectory dataset provides the most detailed and comprehensive records of POV movements, even user behavior.
3) *Driving status*. Driving status data are collected by the OBD module every 30 seconds, and include the vehicle speed (`Speed`), the driving direction (`Direct`), the current mileage (`Mileage`), the revolution per minute (`RPM`), the accelerator pedal position (`AccPos`) and the description of activated alarms (`AlarmDesc`).

Table II contains an example of a trip entry and Table III shows a sample of a trajectory data point and the driving status.

### C. Privacy Protection Measures

We have received the approval from every POV owner to collect data from their cars. Additionally, they gave their consent to educational institutions to study their data for research purposes. We are very aware of their privacy and have taken active steps to protect the MAS platform users. First, the vehicle ID has been anonymized (as a bit string) by our collaborating company, and we do not have access to the true vehicle ID or even vehicle model. Second, all the researchers are regulated by a strict non-disclosure agreement. The dataset is stored in a server protected by authentication mechanisms and firewalls in our collaborating company's network. Our collaborator overlooks the data processing on their servers.



(a) Active Cars per day

(b) User Participation (in days)

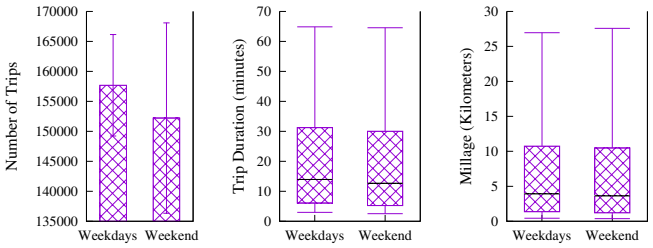
Fig. 2: Driver participation in MAS framework during July 2016.

## III. Driving Behaviour Analysis

The collected dataset is composed of 68,069 unique vehicles that conducted 4,844,563 trips in 12 cities during the July of 2016. The total records from the OBD and GPS devices (i.e., the trajectories and the driving status records) are 426,992,602. Fig. 2(a) shows the exact number of vehicles per day. The number of active vehicles participated in the data collection varies per day, and the average is 47,211. Fig. 2(b) shows

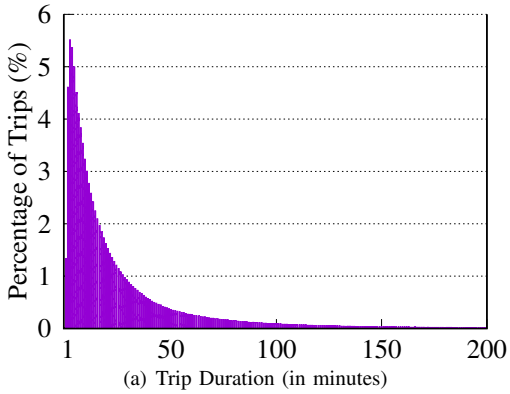
TABLE IV: Characteristics of the collected dataset.

| Vehicles | Trips | Cities | Collection Period | Trajectory and driving status records | | | |
|---|---|---|---|---|---|---|---|
| 68,069 | 4,844,563 | 12 | 1/07/2016 - 31/07/2016 | 426,992,602 | | | |
| City | Vehicles | Records | Size ($km^2$) | City | Vehicles | Records | Size ($km^2$) |
| Shenzhen | 11,403 | 36,808,679 | 2,050 | Guangzhou | 9,617 | 43,089,864 | 7,433 |
| Shanghai | 6,062 | 41,525,996 | 6,340 | Changsha | 4,647 | 22,091,783 | 11,819 |
| Zhengzhou | 4,273 | 17,661,095 | 7,507 | Wuhan | 4,055 | 16,102,250 | 8,494 |
| Xiamen | 3,113 | 10,676,474 | 1,699 | Xian | 2,900 | 17,096,606 | 10,097 |
| Kunming | 2,666 | 11,094,456 | 21,501 | Nanning | 2,360 | 9,819,783 | 22,189 |
| Chongqing | 1,692 | 7,544,379 | 82,300 | Chengdu | 1,640 | 5,756,221 | 14,378 |

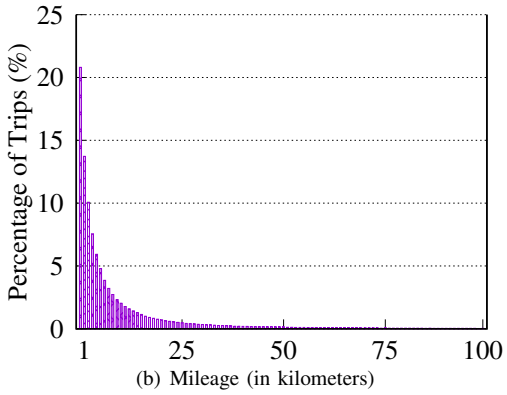



(a) Trips per day (b) Trip Duration (c) Trip Millage

Fig. 3: Characteristics of trips.




(a) Trip Duration (in minutes)




(b) Mileage (in kilometers)

Fig. 4: Distributions of duration and millage characteristics of the trips.

the distribution of drivers based on their participation. The percentage of vehicles (i.e., participating drivers) that provided data every day of that month is 15% (i.e., 10,210).

We first look into the geographical distribution of the collected data. Since the exact location of the cars had been changing during the data collection period, we first define the typical location of each collected private car. The typical longitude and latitude are the private car's `StartLon` and `StartLat` whose average dissimilarity to all the other `StartLon` and `StartLat` of this car is minimal. Hence, the typical location of each collected car can be mathematically expressed as,

$$\begin{aligned} \mathcal{L}_{lon} &\leftarrow \arg \min_{\texttt{StartLon}} \sum (\texttt{StartLon} - \texttt{StartLon}_i), \\ \mathcal{L}_{lat} &\leftarrow \arg \min_{\texttt{StartLat}} \sum (\texttt{StartLat} - \texttt{StartLat}_i), \end{aligned} \tag{1}$$

where $\mathcal{L}_{lon}$ and $\mathcal{L}_{lat}$ are the typical longitude and latitude respectively. Using this information we counted the number of vehicles and records in each of the twelve identified Chinese cities. Table IV presents the characteristics of the collected dataset in detail. Over 20,000 collected private cars are located in the Pearl River Delta region including Guangzhou and Shenzhen, which accounts for the significant part of our dataset. In addition, there are still some cars distributed in other big cities in the mainland of China, such as Shanghai, Xian, Zhengzhou Changsha, and Wuhan.

It is worth to note that our MAS platform is live and the number of cars using our platform is increasing every day.

### A. Trip Patterns

We next focus on understanding users' driving behavior from the characteristics of the recorded trips. We initially categorize the trips according to the ones conducted on the weekdays and the ones conducted on the weekends. As shown in Fig. 3(a), the average number of trips in a weekday is higher than the ones in a weekend, but the standard deviation is higher on the weekends. This is explained by the fact that during weekdays drivers follow their routines while they are more unpredictable during the weekends. Further, in this direction, we calculate the distributions of duration and millage of the trips on weekdays and weekends. As shown in Figs. 3(b) and 3(c), both the distributions for weekdays and weekends are very close with each other, while the trip duration and the mileage on weekdays are slightly longer. The candlesticks in the box-plots of Figs. 3(b) and 3(c) depict the bottom 10% and the upper 90% of the collected values while the sketched area contains the values between 25% and 75%. The horizontal line inside the box depicts the median of the distribution.

To further investigate the characteristics of trips, we produce

the plots of Fig. 4. In Fig. 4(a) we depict the long tail distribution of trip duration which can be approximated by the following Weibull distribution[2],

$$P(T_D) = 0.0754 \cdot T_D^{-0.0724} \exp(-0.0813 \cdot T_D^{0.9276}), \quad (2)$$

with Root Mean Square Error (RMSE) 0.002485. $T_D$ denotes the trip duration and is in minutes. $P(\cdot)$ represents the probability. As expected, most of the trips are shorter than an hour, but also there are trips with a duration of more than two hours.

Next, Fig. 4(b) depicts the distribution of the trip mileage. Most trips are shorter than 25 kilometers, but there exist trips of more than 75 kilometers. This distribution can also be approximated by a Weibull distribution. The formula is,

$$P(T_M) = 0.2782 \cdot T_M^{-0.1715} \exp(-0.3358 \cdot T_M^{0.8285}), \quad (3)$$

where $T_M$ is the trip mileage and in kilometers. The RMSE is 0.002485.

We then calculate the distribution of the number of trips per day for all of the registered users, as shown in Fig. 5(a). For one user, he/she is of the highest probability around 0.17 to take four trips per day. This distribution also can be approximated with a Weibull distribution with a negligible RMSE of 0.005564,

$$P(N_T) = 0.0376 \cdot N_T^{1.3540} \exp(-0.0160 \cdot N_T^{2.3540}), \quad (4)$$

where $N_T$ stands for the average number of trips per day.

The distribution of the average trip duration and the average trip mileage for each user are depicted in Fig. 5(b) and 5(c), respectively. We notice that a trip has more than 25% probability to last 20-30 minutes and around 33% probability to travel 10 KM during the trip. Nevertheless, both of these distributions can be approximated with a Gaussian distribution[3].

The distribution of average trip duration for each user is approximated by,

$$P(\overline{T_D}) = 0.3062 \cdot \exp\left(-\left(\frac{\overline{T_D} - 11.42}{8.786}\right)^2\right), \quad (5)$$

with RMSE of 0.01778. Here, $\overline{T_D}$ is the average trip duration for each user.

The distribution of average mileage for each user is approximated by,

$$P(\overline{T_M}) = 0.2761 \cdot \exp\left(-\left(\frac{\overline{T_M} - 28.61}{18.03}\right)^2\right), \quad (6)$$

with RMSE of 0.01917. $\overline{T_M}$ is the average trip mileage for each user.

We further explore the variance of car usage between weekdays and weekends, as shown in Fig. 6. We consider trips to represent the car usage and employ `StopTime` to describe the temporal feature of each trip. From the results, we observe that the use of private cars decreases during the night and increases during the daytime, due to drivers' daily routines. At about 5 AM, the usage of cars is the lowest since

[2]The general form of Weibull distribution is: $y = a \cdot b \cdot x^{b-1} e^{-a \cdot x^b}$

[3]The general form of Gaussian distribution is: $y = a \cdot e^{-(\frac{x-b}{c})^2}$

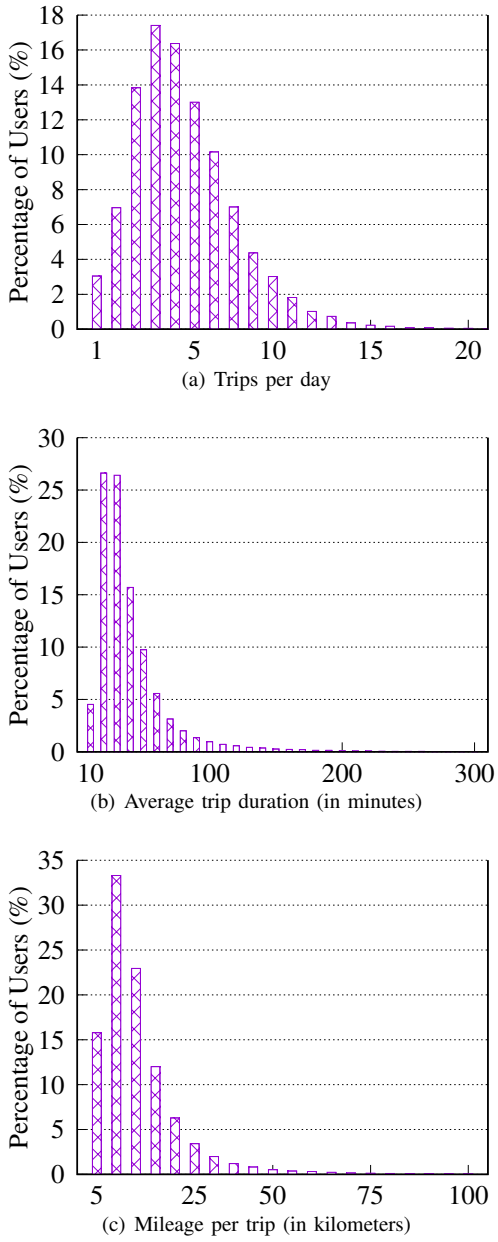



Fig. 5: Distributions of trips' characteristics for users.

most drivers are sleeping and their vehicles are inactive. The vehicle usage first peaks during the morning rush hour, i.e., 10 AM, on weekdays. Further, the vehicle usage reaches its maximum at evening rush hour, i.e., 7 PM, both on weekdays and weekends since most people start their leisure time at that time and drive their private cars to go outside. We also notice that the vehicle usage on weekdays is higher in the morning than on weekends. Also, on the weekends, vehicles are more active until later in the night than during the weekdays. This may be explained as people going to bed later and sleeping longer during the weekend. We can see there are two small peaks at about 0.30 AM and 1.30 AM on weekends. We infer that after entertainment people return home at that time. Similar patterns also exist in our other aspects of life, such as

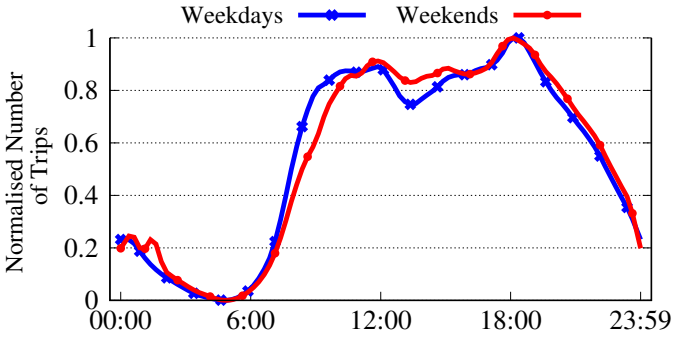



Fig. 6: Min-max normalized number of trips.

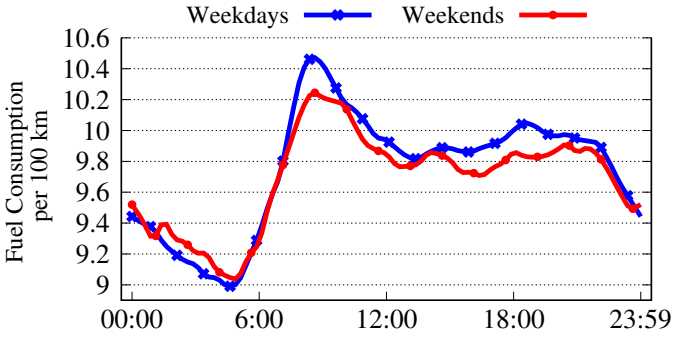



Fig. 7: Fuel consumption (L) per 100 km (Fuel efficiency).

the mobile data usage of smartphone users over a day [11].

### B. *Fuel Efficiency of the Trips*

Fuel consumption of the cars is an important metric in describing the driving behavior. Therefore, we try to understand the fuel efficiency of the recorded trips. We compute the amount of fuel used per unit distance. For example, liters per 100 kilometers, i.e., Fuel Efficiency = (L/100 km). Therefore, the lower the value, the less fuel is required to travel a certain distance. These numbers can be mapped to compute the fuel efficiency of a trip as well, which also encompass the contribution of other practical factors such as traffic congestion, the speed limit of the roads, users' driving behavior. All the cars in our dataset use *Gasoline* as fuel. Fig. 7 demonstrates the average fuel efficiency of daily trips regarding L/100km for all trips in the dataset. We notice that fuel efficiency begins to degrade as the amount of traffic starts to increase after 6 AM, i.e., when the rush morning hour begins. As soon as the rush hour ends, namely after 10 AM, the fuel efficiency of the trips increases implying less traffic. Fuel efficiency is correlated with the time and the volume of trips. This is clearly depicted in Fig. 7, and we notice that weekends are more fuel efficient than weekdays. Fuel efficiency also depends on the car model, however, it is not possible to find the contribution of a car itself from our dataset, as we do not collect the car model for privacy reasons.

We then further investigate the application of this dataset by utilizing the historical data about the trips from a vehicle's perspective. This application recommends places and fuel stations from individuals past trips. We conduct the approach shown in

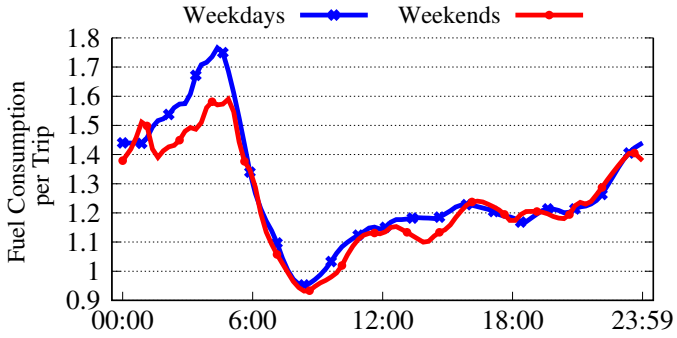



Fig. 8: Fuel consumption (L) per trip.

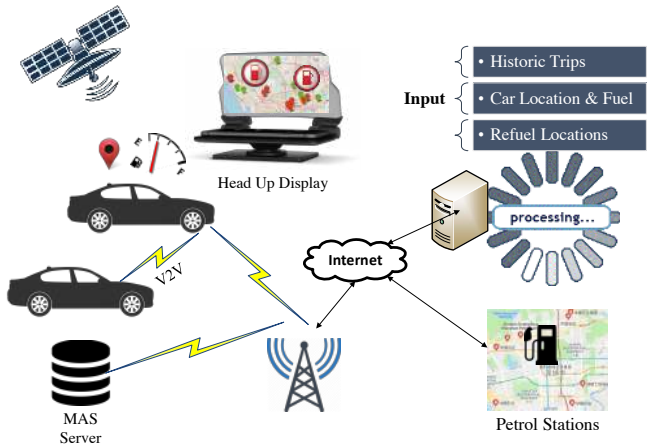



Fig. 9: Popularity and Remaining fuel-based PoI recommendation framework, and other use cases of private car dataset.

Fig. 9. In this approach, we feed the information about the car, i.e., remaining fuel and current location, and the related car's past trips along with the list of positions of the petrol stations. Afterward, we calculate the fuel consumption of this car from the historical traces. For efficient calculation, we classify these trips according to the day of the week (weekday, weekend) and time of the day (rush hour or usual traffic hour), and select the one that best fits the current day and time. More specifically, construct fuel efficiency profile of a car for every visited route in the history. Fig. 8 shows an example profile for all the trips in the dataset, which infers citizens drive long-distance trips at night while short-distance trips during the daytime. We then calculate the maximum reachable distance based on the remaining fuel and recommend the locations of reachable stops and petrol stations starting from the current location. A car can simply request the backend server with the remaining fuel and present location. This application can be packaged as a smartphone application as well. For faster response, the backend can be commissioned at the base stations.

## IV. Use Cases for the Dataset

Many applications can employ the up-to-date version of this dataset. Fig. 9 illustrates a framework to utilize it efficiently. To better assist drivers, some applications can retrieve the past trips and trajectories to recommend an optimal route during

an ongoing trip, as discussed in Section IV-A, or find the economic path regarding fuel, or even analyze the driver's behavior for safety concerns real time and provide feedback. In this section, we list three promising use cases which also are our potential future researches of the dataset.

### A. Collaborative Route Recommendations

The collected dataset can be utilized to provide a personalized travel route recommendation and plan the optimal travel route between two geographical locations [12, 13]. Since we have the whole historical trajectories of each car for each trip, we can get the driving profile of each user, specifically the trip duration, distance, and fuel consumption. Similar to [14], we can design the personal driving assistant for each driver and utilize this extracted knowledge to recommend the best route based on community preferences like the one of less fuel consumption, less time or more popularity [15]. Moreover, we can feed remaining fuel and the location of the gas stations to improve the recommendation by considering the situation that whether remaining fuel can support the planned trip or not.

### B. V2V Communications

Vehicle-to-Vehicle (V2V) communications as a networking paradigm enable vehicles with network interfaces that are able, apart from connecting to Internet services, to interconnect. Such connected vehicles can securely assist each other on V2V applications [16]. Example V2V applications are collision avoidance, content and route sharing, and remote vehicle software updates and diagnosis [17]. All these enable road safety, efficient traffic management for reducing congestion and $CO_2$ emission, pedestrian safety, and other urban developments.

In practice, however, it is difficult to evaluate the performance of such applications or new models in a large scale deployment. The collected dataset can be used to identify the obstacles towards efficient V2V networks by modeling automobile and communication networks in twelve Chinese cities. Additionally, the dataset also can be used for trace-driven simulation to evaluate the performance of various V2V protocols and applications.

### C. Profiling Driver's Behavior

Aggressiveness driving behavior have been identified as a critical risk factor in road traffic injuries [18], which accounts for 14% of all crashes resulting in death and influences both the risk of an accident and the severity of the injuries [19]. As introduced in Section II, we have comprehensive information about the car motion, namely the items related to trajectories such as `Lat`, `Lon`, `GPSTime`, and `speed`. We can match that into the Google map to get the information about the road, for instance, the speed limit, the location and the environment of this road. Since the majority of private cars are connected with their drivers, we can compare both the data obtained in the dataset with the information getting from Google map, and assess the behavior of the driver. For instance, if the driver changes lanes abruptly, veering left or right in the lane, or takes a different route instead of talking the pre-specified route to follow, where this can be detected by checking the trajectory of the car. This use case aims to assess the quality of driving and check the likelihood of being high-risk driving. Automobile insurers consider many factors when calculating your car insurance premiums and such use case helps them to assess the risk taken by a specific car, where good driving record decreases the premium and risky driving increases it.

## V. Discussion and Conclusions

The number of datasets related to vehicular studies are limited. The most popular datasets are related to taxi trips or shared riding in New York city[4]. The datasets contain pickup and drop off events, trip duration, fair, the number of passengers, payment methods, and trajectory information. These datasets are constrained by predetermined trajectories (buses) or points of interest (taxis). Therefore, previous studies mostly developed tools for exploring the datasets [20] and investigated the congestion pattern, predicted the travel time in New York city, and presented novel methods to improve the traffic congestion [21, 22]. Nevertheless, the amount of private cars are significant in different countries and increasing. Every US household has almost two cars and 80% of the transport vehicles are private cars in China. Based on the existing taxi datasets, we cannot grasp the typical driving behavior of private car owners and their contributions in our urban lives.

In this paper, we formally introduced MAS platform, the trajectory, driving status, and other trip-related information from more than 68 thousand private cars in 12 cities of mainland China. Along with the detailed trip related information, the dataset contains fuel and car related information. We analyzed the dataset and exported random distribution patterns of various features that characterize drivers' behavior. We investigated the fuel efficiency of the trips and highlighted many interesting and promising use cases, i.e., applications or systems, that can benefit from our dataset. The dataset provides a highly comprehensive view of driving behavior and our findings may have very practical implications for other metropolitan cities. It is very likely that we will find a similar distribution of the trips, and the usage pattern of private vehicles. Therefore, from partial information about those cites, we can infer other useful information.

## Acknowledgment

This research has been supported in part by projects 26211515, 16214817 and G-HKUST604/16 from the Research Grants Council of Hong Kong.

## References

[1] Y. Wang, L. Li, and C. G. Prato, "The relation between working conditions, aberrant driving behaviour and crash propensity among taxi drivers in china," *Accident Analysis & Prevention*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457518301325

[2] E. Gilman, G. V. Georgiev, P. Tikka, S. Pirttikangas, and J. Riekki, "How to support fuel-efficient driving?" *IET Intelligent Transport Systems*, vol. 12, no. 7, pp.

[4] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

631–641, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8436521

[3] H. Nguyen, W. Liu, and F. Chen, "Discovering congestion propagation patterns in spatio-temporal traffic data," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 169–180, June 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7511741

[4] G. Wang, X. Chen, F. Zhang, Y. Wang, and D. Zhang, "Experience: Understanding long-term evolving patterns of shared electric vehicle networks," *arXiv preprint arXiv:1812.07499*, 2018. [Online]. Available: https://arxiv.org/pdf/1812.07499.pdf

[5] C. Liu, K. Deng, C. Li, J. Li, Y. Li, and J. Luo, "The optimal distribution of electric-vehicle chargers across a city," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 261–270. [Online]. Available: https://ieeexplore.ieee.org/document/7837850

[6] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 186–194. [Online]. Available: https://dl.acm.org/citation.cfm?id=2339561

[7] C. N. S. Bureau, *China Statistical Yearbook*. Beijing: China Statistical Publishing House, 2011-2017. [Online]. Available: http://www.stats.gov.cn/tjsj/ndsj/2017/indexeh.htm

[8] M. Sivak, "Has motorization in the us peaked? part 10: Vehicle ownership and distance driven, 1984 to 2016," Tech. Rep., 2018. [Online]. Available: http://www.umich.edu/~umtriswt/PDF/SWT-2018-2.pdf

[9] M. A. K. Niazi, A. Nayyar, A. Raza, A. U. Awan, M. H. Ali, N. Rashid, and J. Iqbal, "Development of an On-Board Diagnostic (OBD) kit for troubleshooting of compliant vehicles," in *Emerging Technologies (ICET), 2013 IEEE 9th International Conference on*. IEEE, 2013, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/6743551

[10] L. Kis and B. Lantos, "Development of state estimation system with INS, magnetometer and carrier phase GPS for vehicle navigation," *Gyroscopy and Navigation*, vol. 5, no. 3, pp. 153–161, 2014. [Online]. Available: https://link.springer.com/article/10.1134/S2075108714030055

[11] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, April 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7762185

[12] G. Cui, J. Luo, and X. Wang, "Personalized travel route recommendation using collaborative filtering based on GPS trajectories," *International Journal of Digital Earth*, vol. 11, no. 3, pp. 284–307, 2018. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/17538947.2017.1326535

[13] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "GreenGPS: A Participatory Sensing Fuel-efficient Maps Application," in *Proceedings of MobiSys*, 2010, pp. 151–164. [Online]. Available: https://dl.acm.org/citation.cfm?id=1814450

[14] V. C. Magaa and M. Muoz-Organero, "Artemisa: A personal driving assistant for fuel saving," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2437–2451, Oct 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7345559

[15] L. Zheng, Q. Feng, W. Liu, and X. Zhao, "Discovering trip hot routes using large scale taxi trajectory data," in *International Conference on Advanced Data Mining and Applications*. Springer, 2016, pp. 534–546. [Online]. Available: https://link.springer.com/chapter/10.1007\%2F978-3-319-49586-6_37

[16] D. Chatzopoulos, C. Bermejo, E. u. Haq, Y. Li, and P. Hui, "D2D task offloading A dataset-based Q&A," *IEEE Communications Magazine*, pp. 1–6, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8594700

[17] K. Zhu, D. Niyato, P. Wang, E. Hossain, and D. In Kim, "Mobility and handoff management in vehicular networks: a survey," *Wireless communications and mobile computing*, vol. 11, no. 4, pp. 459–476, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/wcm.853

[18] A. B. R. González, M. R. Wilby, J. J. V. Díaz, and C. S. Ávila, "Modeling and detecting aggressiveness from driving signals," *IEEE Transactions on intelligent transportation systems*, vol. 15, no. 4, pp. 1419–1428, 2014. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6725676

[19] W. China. (2018) Road safety in china. [Online]. Available: http://www.wpro.who.int/china/mediacentre/factsheets/road_safety/en/

[20] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, Dec 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6634127

[21] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37 – 49, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X13000740

[22] M. M. Vazifeh, P. Santi1, G. Resta, S. H. Strogatz, and C. Ratti1, "Addressing the minimum fleet problem in on-demand urban mobility," *NATURE*, vol. 557, pp. 534–538, 2018. [Online]. Available: https://www.nature.com/articles/s41586-018-0095-1