Using taxi GPS data for macroscopic traffic monitoring in large scale urban networks: calibration and MFD derivation

Shoufeng Lu\textsuperscript{a,\ast}, Victor L.Knoop\textsuperscript{b}, Mehdi Keyvan-Ekbatani\textsuperscript{c}

\textsuperscript{a}Traffic and Transportation College, Changsha University of Science and Technology, Wanjiali road No.960, Changsha 410114, China
\textsuperscript{b}Department of Transport Planning, Delft University of Technology, 2628 CN Delft, Netherlands
\textsuperscript{c}Civil and Natural Resources Engineering, University of Canterbury, Private Bag 4800 Christchurch 8140, New Zealand

Abstract

A two-Fluid Model (TFM) of urban traffic provides the macroscopic description of traffic state. The TFMs parameters are hard to calibrate, particularly for the dynamic traffic conditions. This leads to the TFM often being used to compare the quality of service through the plot of stopping time versus trip time of the vehicles in the network. Recently, the taxi GPS data has been applied to predict the traffic condition at the network level. Despite the network-wide coverage of the taxi GPS probe data, the penetration rate of taxis in the network traffic is still a vital and challenging issue for traffic estimation purpose. It is necessary to estimate penetration rate of taxis by combining with other data sources. Here, we propose a novel approach to fill two gaps: TFM parameter calibration and the taxis penetration rate. This method stretches the description of TFM to a zone size. The method is applied to real Changsha city GPS data, calibrating the parameters. The macroscopic fundamental diagram of the large-scale city is derived. For the Changsha case, running speed is the super-linear power law of the fraction of running cars; the fraction of stopping time is nearly linear power law of density, which can be an alternative of the density. The proposed method enables the calibration of TFM parameters and macroscopic traffic monitoring at urban scale using only GPS data.

© 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
"Peer-review under responsibility of the scientific committee of the International Symposium of Transport Simulation (ISTS’18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS’18)."

Keywords: macroscopic fundamental diagrams; taxi GPS data; two-fluid model; parameter calibration

1. Introduction

With the development of intelligent transportation systems, a wide range of data sources provide an opportunity for macroscopic monitoring and modelling in large scale. Geroliminis and Daganzo (2008) showed empirically a well-defined relationship between average flow and average density in the urban scale defined as MFDs (Macro-
scopic Fundamental Diagrams). This work inspires many people to focus on MFDs research. MFDs are regarded as a promising approach of modelling urban traffic. Because the simulation software brings convenience for setting arbitrary traffic scenarios, most studies used the simulation method to generate traffic states from free flow to congestion, and complete MFDs can be generated. Many studies have examined various aspects of MFDs, such as the impact of route choice, turn maneuvers and density distribution on MFDs (Gayah and Daganzo (2011); Knoop et al. (2013); Leclercq and Geroliminis (2013); Mahmassani et al.(2013)), MFD shape under adaptive traffic control strategy (Keyvan-Ekbatani et al., 2016) and MFD estimation using a variety of data sources (Leclercq et al. (2014); Du et al. (2016); Gayah and Dixit (2013); Gayah, et al. (2014)). Keyvan-Ekbatani et al. (2013) proposed a reduced MFDs concept, which uses a reduced amount of measurements, but exhibit a critical range of traffic states equivalent to the complete MFDs. Several studies have developed real-time control strategies using the MFD to limit vehicle entries within a protected network using traffic signals to ensure that it does not become oversaturated (Keyvan-Ekbatani et al.(2012); Haddad et al.(2013); Haddad and Shraiber(2014); Keyvan-Ekbatani et al.(2015); Ramezani et al.(2015)).

A two-Fluid Model (TFM) of urban traffic proposed by Herman and Prigogine (1979) stems from kinetic theory of multilane highway traffic (Prigogine and Herman,1971). This assumes that the urban traffic flow consists of two fluids (flows). The first one is that of the moving vehicles in the city network, and the second is that of the vehicles stopped as a result of traffic signals, stop signs, and other traffic control devices. Parked vehicles, however, are not considered as a part of the second flow. TFM is used to evaluate network performance (Herman and Ardekani, 1984), analyze the relationship between TFM parameters and network topology, traffic control and traffic incidence (Mahmassani et al. 1984), and to utilize the two-fluid trend in the plot of stopping time versus trip time to compare quality of service (Vo et al. 2007). Recently, behavioural foundations of TFM (Dixit, 2013), the combination of TFM and safety (Dixit et al., 2011), the combination of TFM and MFD (Dixit and Geroliminis,2015) have been investigated. Specifically, Dixit and Geroliminis (2015) testified that TFM should be defined by aggregated level data instead of individual level data, and derived a formula similar to the flow rate-density relationship by replacing the flow rate with production rate of stopped vehicles, and the density with the fraction of stopped vehicles. The difference between our work and their paper is that our proposed approach can calibrate two TFM parameters, and their paper can only calibrate one parameter. TFM describes two relations. First, the average running speed in an urban network is proportional to the nth power of the fraction of the vehicles that are moving.

\[ v_r = v_m(f_r)^n \]  

where \( v_r \) is running speed, \( f_r \) is the fraction of running vehicles, \( v_m \) is maximum speed, \( n \) represents one of the two-fluid model parameters. Secondly, the fraction of the stopped vehicle is proportional to the pth power of the normalized density.

\[ f_s = \left( \frac{k}{k-m} \right)^p \]  

The chase car technique was adopted to collect individual vehicle trajectories to calibrate two parameters \( n \) and \( p \) mentioned above (Herman and Prigogine 1979; Ardekani, 1984; Mahmassani et al., 1984; Mahmassani et al., 1987; Herman, Malakhoff, Ardekani, 1988 ; Mahmassani et al., 1990; Williams et al. 1985; Williams et al. 1995 ; Ayadh 1986; Ardekani et al., 1992; Vo et al., 2007; Dixit et al., 2011; Dixit et al., 2012). Using individual vehicle data heavily relies on the ergodic and steady state assumptions. An ergodic traffic system is one which time mean performance of any single vehicle over a sufficiently long period of time would be identical to the mean performance of all the vehicles in the ensemble over the same period. However, traffic is a dynamic process, an individual vehicles could result in biased sampling of the different parts of the network depending on their distribution in a given time interval. Dixit and Geroliminis (2015) utilized regression analysis to evaluate the differences in TFM parameter estimation using individual vehicle data and network level data. They found that it would not be appropriate to utilize individual data to determine TFM due to the violation of the steady state conditions. Also, Aerial photographic technique was used to obtain the network vehicle density, but it is time consuming and costly. Since network vehicle density is difficult to observe, after the work of Ardekani (1984), the calibration of parameter \( p \) is rarely studied. This paper aims to stretch the description of the TFM to a zonal approach. Using real-world data from taxis, we also calibrate parameters \( n \) and \( p \) for the equations. This way, an MFD is created for a city with real data, for which the penetration rate of taxis is unknown. This is indeed one of the novelties and contributions presented in this paper. The paper is organized to
first provide the method of calibrating parameters, followed by calibration procedure, then presents the macroscopic fundamental diagram of two-fluid model, and finally provide the results and key finding.

2. Methodology

The goal is to derive MFD using only taxi GPS data. The proposed method combines the existing TFM and the recently studied notion of MFD. Both TFM and Edies definitions (1965) rely on vehicle trajectories, and therefore connecting TFM and Edies definitions would provide opportunities to estimate urban traffic state using taxi GPS data of only an unknown fraction of the traffic. The main idea of relating them is to utilize mean travel time and mean travel distance.

Nomenclature

\begin{align*}
  v_{\text{ave}} & \quad \text{Average speed of all cars} \\
  v_{\text{taxi}} & \quad \text{Average speed of taxis} \\
  v_r & \quad \text{Running speed} \\
  v_m & \quad \text{Maximum speed} \\
  T_m & \quad \text{Minimum travel time} \\
  f_r & \quad \text{The fraction of running cars} \\
  f_s & \quad \text{The fraction of stopped cars} \\
  T_s & \quad \text{Stop time} \\
  T & \quad \text{Travel time} \\
  T_{s,i} & \quad \text{Stop time of ith car} \\
  T_{r,i} & \quad \text{Running time of ith car} \\
  T_i & \quad \text{Travel time of ith car} \\
  d_i & \quad \text{Travel distance of ith car} \\
  N & \quad \text{The number of all cars} \\
  N_{\text{taxi}} & \quad \text{the number of taxis} \\
  \tau & \quad \text{The length of time period}
\end{align*}

We assume that during a time period of 5 min, the mean travel distance, mean travel time and mean speed of taxis is equal to those of other cars.

\begin{align}
  \frac{T_{TD_{\text{taxi}}}}{N_{\text{taxi}}} & = \frac{T_{TD}}{N} 
\end{align}

where \( T_{TD_{\text{taxi}}} \), \( T_{TT_{\text{taxi}}} \) are total travel distance and total travel time of taxis. \( N_{\text{taxi}} \) is the number of taxis. \( v_{\text{taxi}} \) is average speed of taxis. \( v_{\text{ave}} \) is average speed of all cars. TTD and TTT are total travel distance and total travel time of all vehicles. \( N \) is the number of all vehicles. This assumption relies on the amount and spatial distribution of taxis. Only taxis with passengers are analyzed, since they provide the most reasonable sample of the traffic flow, and do not have any non-traffic related stops. Edies definitions of the flow rate and the density are based on the vehicle trajectories. The average flow \( q \) and density \( k \) in a network for network length \( L \) and the time interval are given by:

\begin{align}
  q & = \frac{\sum_k d_k}{L \times \tau} = \frac{T_{TD}}{L}; \quad k = \frac{\sum_k \tau_k}{L \times \tau} = \frac{T_{TT}}{L}
\end{align}

where \( d_k \) and \( \tau_k \) are respectively distance travelled and the time spent by a trip \( k \) within the space-time window defined by network length \( L \) and time interval \( \tau \). TTD is the total travel distance of all vehicles. TTT is the total travel time of all vehicles. In most real-world applications, the total time and distance travelled will only be known if the detailed trajectories of all vehicles are provided. If these data are only provided by a subset of vehicles in the network (e.g., those serving as mobile probes), Eq. (4) cannot be applied directly. This difficulty can be solved by computing mean...
vehicle density and mean flow of a single car, which are derived by dividing N, the number of vehicles served by road network, on both side of Eq. (4), then yields the Eq. (5).

\[
\bar{q} = \frac{1}{N} \frac{TTD}{L \times \tau}, \bar{k} = \frac{1}{N} \frac{TTT}{L \times \tau}
\]  

(5)

Note that, TTD and TTT is the aggregated network level data, which are not subject to the assumptions regarding ergodicity and steady state. The derivation of the proposed methodology starts from the relation between network average flow and average speed in TFM is given by

\[
q_{\text{ave}} = k_m v_{\text{ave}} (1 - \left(\frac{v_{\text{ave}}}{v_m}\right)^{1/n})^{\frac{1}{p}}
\]  

(6)

where \(k_m\) is jam density, \(q_{\text{ave}}\) is average flow, \(v_{\text{ave}}\) is average speed, \(v_m\) is average maximum speed, \(n\) and \(p\) are parameters. According to Eq. (5), \(q_{\text{ave}}\) can be rewritten as

\[
q_{\text{ave}} = \frac{TTD}{L \times \tau}
\]  

(7)

Substitute Eq. (7) into Eq.(6) then yields

\[
\frac{TTD}{L \times \tau} = k_m v_{\text{ave}} (1 - \left(\frac{v_{\text{ave}}}{v_m}\right)^{1/n})^{\frac{1}{p}}
\]  

(8)

According to another definition of density, density is defined as the number of vehicles per unit length of the roadway, \(k_m = \frac{N_m}{L}\), where \(N_m\) is the maximum number of vehicles in a network with a network length \(L\). Now, the two sides of the equation (8) are divided by the number of vehicles served by road network \(N\), then yielding

\[
\frac{TTD}{N \times L \times \tau} = \frac{N_m}{N \times L} v_{\text{ave}} (1 - \left(\frac{v_{\text{ave}}}{v_m}\right)^{1/n})^{\frac{1}{p}}
\]  

(9)

According to Eq.(3),

\[
\frac{TTD_{\text{taxi}}}{N_{\text{taxi}} \times L \times \tau} \approx \frac{N_m}{N \times L} v_{\text{taxi}} (1 - \left(\frac{v_{\text{taxi}}}{v_m}\right)^{1/n})^{\frac{1}{p}}
\]  

(10)

Equation (10) relates mean taxi travel distance and mean taxi speed, which contains two parameters \(n\) and \(p\). To proceed, another relation \(f_s = \left(\frac{k}{k_m}\right)^p\) in TFM, the fraction of the stopped vehicle is proportional to the \(p\)th power of the normalized density, is used. Due to the density definition, \(k = \frac{N}{L}\) and \(k_m = \frac{N_m}{L}\), we can derive

\[
\frac{N_m}{N} = (f_s)^p
\]  

(11)

Substitute (11) into Eq. (10), then

\[
\frac{TTD_{\text{taxi}}}{N_{\text{taxi}} \times L \times \tau} \approx \frac{1}{L} (f_s)^p v_{\text{taxi}} (1 - \left(\frac{v_{\text{taxi}}}{v_m}\right)^{1/n})^{\frac{1}{p}}
\]  

(12)

Where \(f_s\) is the fraction of stopped cars, which is measured by the ratio of total stopped time to total travel time of taxis. Note that calculation of \(f_s\) utilizes the aggregated network level data. Therefore, Eq. (12) is the proposed approach for calibrating parameters \(n\) and \(p\) only using taxi GPS data. Note that, the quantity of variables, except \(n\) and \(p\), are measured by aggregated network level taxi GPS data, which ensures Eq. (12) also holds for non-steady state condition. When parameters \(n\) and \(p\) are known, we can estimate the number of vehicles, network average flow and network average density, so that MFD can be created. Network average flow and network average density are computed by

\[
q_{\text{ave}} = k_m v_{\text{ave}} (1 - \left(\frac{v_{\text{ave}}}{v_m}\right)^{1/n})^{\frac{1}{p}}, k_{\text{ave}} = k_m (1 - \left(\frac{v_{\text{ave}}}{v_m}\right)^{1/n})^{\frac{1}{p}}
\]  

(13)
3. Parameter calibration procedure

The procedure for calibrating parameters \( n \) and \( p \) only using taxi GPS data consists of four steps. For each time period \( \tau \), the quantity of all variables are measured from taxi GPS data. For simplicity, the subscript of all variables omits the word taxi.

Step 1 is data filtering. Only taxis with passengers were analyzed, since they provide the most reasonable sample of the traffic flow, and do not have non-traffic related stops. This is achieved by removing taxis which dropped off or picked up passengers within a given time interval. For taxis that did not have passengers throughout the given time interval, their records also are removed.

Step 2 is the estimation of average maximum speed and minimum travel time. Eq.(1) tells us when all vehicles are running (i.e. there are no stopped vehicles in the network, \( f_r = 1 \)), the running speed \( v_r \) is equal to \( v_m \). Therefore \( v_m \) is defined as the average maximum speed, when no vehicles are stopped. \( T_m = \frac{1}{v_m} \), \( T_m \) is the minimum free flow travel time per unit distance. \( T_m \) is calibrated by fitting the scatter of \( T \) and \( T_s \). If the velocity is less than or equal to 5 km/h, the cars is considered standing.

Step 3 is the calibration of parameter \( n \). Parameter \( n \) is calibrated by fitting the scatter of \( v_r \) and \( f_r \), through the formula \( v_r = v_m f_r^n \). The fraction of running vehicles \( f_r \) is given by the ratio of the sum of running time (velocity is larger than 5 km/h) spent by the vehicles and the total time spent by the vehicles in the network. The fraction of stopped vehicles \( f_s \) is given by the ratio of the sum of stop time (velocity is less than or equal to 5 km/h) spent by the vehicles and the total time spent by the vehicles in the network. Theoretically, \( f_r + f_s = 1 \).

Step 4 is the calibration of parameter \( p \). After \( f_r, v_m, n \) are determined in the previous steps, utilize Eq. (12) to fit the scatter of \( \frac{TDD}{NL} \) and \( v \) to estimate parameter \( p \). The nonlinear least-squares regression is used to fit, through nlinfit function built in Matlab software.

4. Empirical validation

The daily GPS data of 6184 taxis on Monday, April 22, 2013 is from the city of Changsha in China. Changsha is the capital of Hunan province, situated in the middle part of China. There are 3.62 million population in Changsha metropolitan area at Year 2013, 1658 persons per square kilometer, as expected it has congestion problems. A single taxi records consist of the current time, GPS locations in form of longitude and latitude coordinates, speed, heading angle and passenger information. The length of time interval of recording is 30 seconds. There are approximately 16.8 million records for the day. The area of interest was 34.1 square kilometers, of which 5.5km length from north to south, 6.2km length from east to west. The total length of road is 109.454km. The average number of lanes on the road segment is 4.44 for motor vehicles, and 2 for non-motor vehicles. Due to the error of taxi GPS data, we had to perform a map matching of coordinates of GPS data and road network, which is achieved through ray crossing algorithm (Rourke,1998). We will implement the parameter calibration procedure proposed in this paper for collected, real-world taxi gps data. The data records between 24:00pm and 7:00am are removed due to non-traffic related factors, such as, the taxi drivers may have a rest, car washing or eating foods, and some traffic lights may be managed with yellow flash for all approach directions.

4.1. Calibration of \( T_m \) and \( V_m \)

The trend and scatters of travel time and stop time per kilometer during the period between 7:00am and 24:00pm is shown in Fig. 1. The fitted linear formula is \( T = 1.632 T_s + 1.141 \) with goodness of fit R-square value 0.91. The minimum travel time for one kilometer is 1.141 minute. Correspondingly, the maximum speed is 52.6km/h.

4.2. Calibration of \( n \)

Use the formula \( v_r = v_m f_r^n \) to calibrate the parameter \( n \). \( v_m \) takes the value 52.6 km/h estimated in the previous step. The fitted result with goodness of fit R-square value 0.73 is shown in Fig. 2. and \( n = 1.743 > 1 \). \( v_r \) is the super-linear power law function of the fraction of running cars, which means increasing returns. \( f_r \) changes (increase or decrease) by a certain amount, \( v_r \) will change (increase or decrease) more.
4.3. Calibration of \( p \)

Jam density (or gridlock density) is known to vary in different situations (see Mahmassani et al. (2013)). In this paper, the vaule from Gonzales et al. (2009) has been utilised i.e. \( k_m=90.9\ \text{veh/km/lane} \). Indeed, other values are also found (see e.g. Knoop et al. (2013)). Utilize Eq. (12) to fit the scatters of \( \left( \frac{TTD}{NL}, v \right) \) to estimate parameter \( p \). The estimated \( p \) value is \( 1.0038 \approx 1 \). Hence, \( f_s \) is the nearly linear power law of \( k \), both variables change at the same time the same amount. Because \( f_s \) (the fraction of stop time) is easy to measure through taxis, \( f_s \) can be an alternative of the density.

4.4. MFD creation

Using the above calibrated value of parameters \( v_m, n, \) and \( p \), flow and density can be computed using the formulas (13). The created well-defined MFD is shown in Fig. 3. The critical density is 24.58 veh/km/lane, and critical flow is 547 veh/h/lane.
5. Conclusion

The methodology of TFM parameter calibration and MFD creation using taxi GPS data is proposed, which extends the description of TFM to the zone size. This method does not need to estimate the penetration rate of taxis, enabling the use of only taxi data for traffic state monitoring in large scale urban network. Due to using network level data of taxis, the method relaxes the assumptions of ergodic and steady state and can be applied to dynamic traffic state. For the city of Changsha case, $v_r$ is the power law function of the fraction of running cars with scaling exponents $n=1.743$. The power law is super-linear (the coefficient is larger than 1), which means if $f_s$ changes (increase or decrease) by a certain amount, $v_r$ will change (increase or decrease) more than proportionally. $f_s$ is the nearly linear power law function of density $k$ with the scaling exponents $P=1.0038 \approx 1$. Hence, $f_s$ and $k$ change simultaneously the same amount. Because $f_s$ (the fraction of stop time) is easy to measure, $f_s$ can be an alternative of the density.

6. Acknowledgements

This paper is partially supported by the Project of Hunan Education Department (No.15B011).

References


