Transportation Research Part A 111 (2018) 41–52

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Reproducible generation of experimental data sample for calibrating traffic flow fundamental diagram

Jin Zhang[a,b], Xiaobo Qu[b,*], Shuaian Wang[c,d]

[a] Griffith School of Engineering, Griffith University, Gold Coast, QLD 4222, Australia
[b] Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg, Sweden
[c] The Hong Kong Polytechnic University Shenzhen Research Institute, Nanshan District, Shenzhen, China
[d] Department of Logistics & Maritime Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Traffic flow fundamental diagram
Sample selection bias
Reproducible sample generation
Experimental data

ABSTRACT

Speed – density relationship, which is usually referred to as the traffic flow fundamental diagram, has been considered as the foundation of the traffic flow theory and transportation engineering. Speed - density relationship is the foundation of the traffic flow theory and transportation engineering, as it represents the mathematical relationship among the three fundamental parameters of traffic flow. It was long believed that single regime models could not well represent all traffic states ranging from free flow conditions to jam conditions until Qu et al. (2015) pointed out that the inaccuracy was not caused solely by their functional forms, but also by sample selection bias. They then applied a new calibration method (named as Qu-Wang-Zhang model hereafter) to address the sample selection bias. With this Qu-Wang-Zhang model, the result calibrated from observational data sample can consistently well represent all traffic states ranging from free flow conditions to traffic jam conditions. In the current paper, we use a fundamentally different approach that is able to yield very similar and consistent results with the Qu-Wang-Zhang model. The proposed approach firstly applies reproducible sample generation to convert the observational data to experimental data. The traditional least square method (LSM) can subsequently be applied to calibrate accurate traffic flow fundamental diagrams. Two reproducible sample generation approaches are proposed in this research. Based on our analyses, the first approach is somewhat affected by outliers and the second approach is more robust in dealing with potential outliers.

1. Introduction

It has been 80 years since Greenshields et al. (1935) proposed a linear function for the speed-density relationship. After this seminal paper, numerous studies have been carried out to improve such an over-simplified relationship (e.g. Greenberg, 1959; Underwood, 1961; Newell, 1961; Drake et al., 1967; Del Castillo and Benítez, 1995a,b; MacNicholas, 2008; Wang et al., 2011, 2013; Qu et al., 2017). These models have been derived from different car following theories, based on the follow-the-leader concept, i.e., rules on how drivers follow their leading vehicles (Zhang and Kim, 2005; Jin et al., 2015). Many researchers believe that single-regime models cannot represent the empirical data consistently well ranging from the free-flow regime to the congested regime (Edie, 1961; Drake et al., 1967). As such, multiple regime models were proposed by using various different curves to represent distinct traffic states. Although the performance of multiple-regime models perform well in terms of accuracy, it has been criticized due to the

* Corresponding author.
E-mail addresses: drxiaoboqu@gmail.com (X. Qu), wangshuaian@gmail.com (S. Wang).

https://doi.org/10.1016/j.tra.2018.03.006
Received 8 March 2017; Received in revised form 1 March 2018; Accepted 7 March 2018

Table 1
Six single-regime speed-density models (Qu et al., 2015).

Single-regime models	Function	Parameters
Greenshields (Greenshields et al., 1935)	$V = v_f\left(1-\frac{k}{k_j}\right)$	v_f, k_j
Greenberg (1959)	$V = v_m\ln\left(\frac{k_j}{k}\right)$	v_f, k_j
Underwood (1961)	$V = v_f\exp\left(-\frac{k}{k_j}\right)$	v_f, k_j
Northwestern (Drake et al., 1967)	$V = v_f\exp\left(-\frac{1}{2}(\frac{k}{k_o})^2\right)$	v_f, k_o
Newell (1961)	$V = v_f\left[1-\exp\left\{-\frac{\lambda}{v_f}\left\{\frac{1}{k}-\frac{1}{k_j}\right\}\right\}\right]$	v_f, k_j, λ
3PL model (Wang et al., 2011)	$V(k,\theta) = \frac{v_f}{1+\exp\left(\frac{k-k_c}{\theta}\right)}$	v_f, k_c, θ

lack of mathematical elegance. Table 1 lists six well-known single regime models with two or three parameters which will be discussed in the following sections.

Qu et al. (2015) recently pointed out that the inaccuracy of the calibrated models to represent mid to high traffic flow conditions (in particular traffic jam conditions) was not solely caused by the functional forms, but by the sample selection bias. They suggest that there are two types of data samples: the first type refers to the uniformly distributed experimental data sample; and the latter type is observed based on the occurrence of samples in the given survey period which is called observational data. The observational data might be highly skewed and experimental data is usually uniformly distributed. Let us take the speed – density data collected from Georgia State Route 400 (referred to as GA400 dataset hereafter) as an example. The data includes 47,815 observations of speeds and densities, 86.3% (41,259) of which refer to free flow conditions ($<$ 20 veh/km). Obviously, this data is a typical observational data and the calibrated models are likely to be dominated by free flow conditions, which result in poor performances of these models in representing congested traffic states. Qu et al. (2015) proposed a new calibration approach (named as Qu-Wang-Zhang model hereafter) that is able to largely unbiasedly represent all traffic flow conditions from free flow to traffic jam conditions. As such, the single-regime models calibrated by the new approach are able to represent distinct traffic flow conditions consistently well. Fig. 1 shows the GA400 dataset against the calibrated Greenberg model by the traditional least square method and the results by Qu-Wang-Zhang model. It is apparent that the new model outperforms the traditional method, especially for high volume traffic conditions. Note that the mid to high traffic conditions are the focus for traffic control and other transport engineering research and applications (Aboudina et al., 2016; He, 2016). In this regard, these traffic states should be unbiasedly modelled and not be dominated by free flow conditions, although they have relatively low occurrence.

In this research, we propose a fundamentally different approach to the Qu-Wang-Zhang model for addressing the issue of sample selection bias. We propose three sample generation approaches to convert the observational GA400 data into uniformly distributed experimental data so that the traditional least square method can be used for the calibration. The first approach is on the basis of Monte Carlo simulation which is basically to randomly select quasi-uniform samples (Guevara, 2015). Although the performance is acceptable, the result is not reproducible, namely, the calibration results will be different from one run to another, although the exactly same procedure is followed. In this regard, we first design a reproducible sample generation approach that is able to remedy the deficiency mentioned above. However, although the result is reproducible and the performance is good, this sample generation approach is largely affected by the outliers. As a result, a larger number of generated samples do not necessarily lead to better modelling performance. In order to address this issue, we propose another sample generation approach that takes into account all

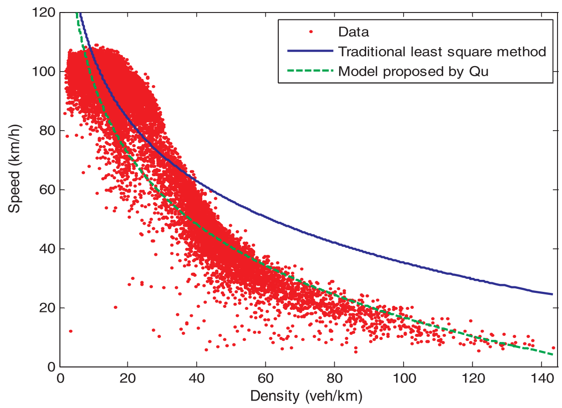

Fig. 1. Calibrated Greenberg models by two methods.