Bootstrap standard error estimations of nonlinear transport models based on linearly projected data

Wai Wong[a]; S. C. Wong[b]; and Henry X. Liu[a,c]

[a]*Department of Civil and Environmental Engineering, University of Michigan, United States*
[b]*Department of Civil Engineering, The University of Hong Kong, Pokfulam, Hong Kong*
[c]*University of Michigan Transportation Research Institute, United States*

Abstract

Linear data projection is a commonly leveraged data scaling method for unbiased traffic data estimation. However, recent studies have shown that model estimations based on linearly projected data would certainly result in biased standard errors. Although methods have been developed to remove such biases for linear regression models, many transport models are nonlinear regression models. This study outlines the practical difficulties of the traditional approach to standard error estimation for generic nonlinear transport models, and proposes a bootstrapping mean value restoration method to accurately estimate the parameter standard errors of all nonlinear transport models based on linearly projected data. Comprehensive simulations with different settings using the most commonly adopted nonlinear functions in modelling traffic flow demonstrate that the proposed method outperforms the conventional method and accurately recovers the true standard errors. A case study of estimating a macroscopic fundamental diagram that illustrates situations necessitating the proposed method is presented.

Keywords: Big data era; Linear data projection; Heteroscedasticity; Bootstrap standard error; Macroscopic fundamental diagram

1. Introduction

The advancement of information technology has ushered in an era of big data. The vast quantity of traffic data generated every day give transportation and traffic researchers the opportunity to gain an unprecedented understanding of the world's transportation systems. Transport models are fundamental tools helping us to understand the past, predict the future and control the systems. Therefore, accurate and reliable transport model estimations are crucial to many transportation studies. Accurate and unbiased traffic data must be input to ensure these models' accuracy and reliability.

Although the accuracy and efficiency of traffic data collection have been substantially improved, it remains difficult in practice to retrieve traffic data from an entire network due to the myriad limitations of high-technology devices. On-road fixed detectors (e.g., loop detectors) can generally be used to accurately acquire traffic data, but their installation and maintenance usually incur huge costs, impeding their universal deployment (Herrera and Bayen, 2010; Herrera et al., 2010). Therefore, their coverage is generally confined to a subset of links (Caceres et al., 2012). The travel times of vehicles traveling through a roadway can be obtained using a vehicle re-identification system; a vehicle's signature is matched when it

passes sensors installed at the two ends of the roadway (Kwong et al., 2009). Any technological utilities or devices that can recognize vehicle identities, such as wireless magnetic sensors (Kwong et al., 2009), license plate recognition systems (Herrera et al., 2010) and radio frequency identification transponders (Wright and Dahlgren, 2001; Ban et al., 2010) can be readily adopted for these schemes. However, in addition to the huge costs of installation and implementation, the potential risk to privacy is a major obstacle to the deployment of these utilities or devices over an entire network. The cellular systems introduced a decade ago (Zhao, 2000; Bolla and Davoli, 2000; Ygnace and Drane, 2001) provide a potential solution to the problems of cost and coverage (Herrera et al., 2010). However, as mobile phones are generally considered a distraction for drivers (Liang et al., 2007), their use to collect traffic data is discouraged or even prohibited in many countries. Global Positioning System (GPS) devices offer another favorable means of collecting traffic data at a relatively low cost from probe vehicles circulating across a network (Miwa et al., 2013). However, data generated from vehicle fleets, such as UPS, FedEx or taxis (Bertini and Tantiyanugulchai, 2004; Moore et al., 2001; Wong et al., 2014; Schwarzenegger et al., 2009), may create bias due to their distinct patterns of operation and travel. Moreover, the global application of such systems is significantly impeded by the additional capital and installation costs and potential privacy issues related to the use of GPS trackers.

To overcome the practical obstacles to direct and accurate traffic data measurement, various mathematical techniques, such as data filtering, scaling and sampling, have been leveraged for accurate traffic data estimation. Linear data projection is a common and highly transferable data scaling method that can be used to infer unobservable population traffic data by extrapolating observable traffic data based on the sampled mean of a set of scaling factors. As more and more connected vehicle (or probe vehicle) data become available in the big data era, such a method is often necessary to scale up observable traffic data. In addition, many transport models in the big data era rely heavily on various sources of data for calibration, estimation and validation. Therefore, a data scaling method that can fuse data collected from various sources for unbiased traffic data estimation is urgently needed. Unobservable traffic data can often be written as a linear combination of observable traffic data and scaling factors. Scaling factors are the ratios bridging unobservable and observable traffic data. They can be defined differently to fit certain specified physical contexts. Due to the complexity and stochastic nature of transportation systems, scaling factors are typically random variables rather than constants, and are therefore usually assumed to follow certain distributions. The variance of a scaling factor can quantify several types of heterogeneity, such as spatial, temporal or spatiotemporal heterogeneity, depending on the sampling approach. In practice, as the values of the scaling factors are not known, they are replaced by the estimated scaling factor mean in a linear data projection for the unbiased estimation of unobservable traffic data.

Many transportation studies have used linear data projection to estimate unobservable traffic data. For instance, this method can be used to unbiasedly estimate the hourly total traffic flow passing through a link without a detector. If the probe vehicle flow is observable on every road and the total traffic flow is only observable on a subset of roads with detectors installed, then the total-traffic-to-probe-vehicle ratio at a link with a detector can be defined

as the scaling factor in this context. As both the total traffic flow and the probe vehicle flow are available, scaling factors can be sampled from these links. Sampled scaling factors usually differ across space, due to the heterogeneity of land-use patterns and stochasticity, but they can generally be assumed to follow a distribution over a network due to geographical proximity. In such cases, the spatial heterogeneity across the network can be quantified using the variance of the scaling factor. As the expected traffic composition ratio is the mean of the scaling factor, if its value is 80 when the hourly probe vehicle flow observed on the road of interest is 20 veh/h, the unbiased estimator for total hourly traffic across this road can be estimated by their product, i.e., 1600 veh/h. This method has been used for traffic flow estimation in many studies (Wong and Wong, 2015; 2016a; 2016b; 2016c; 2018). Similarly, the transit travel distribution can be expanded from a set of sampled vehicles fitted with passenger counting devices, and a zonal travel diary can be extrapolated from a sample of household surveys based on linear data projection. In a recent study of multiple-vehicle crash frequency, Meng et al. (2017a) also applied linear data projection in estimating time exposure, which accounts for variation in crash count at a specific site.

Macroscopic transport model estimations are a representative example of the use of linear data projection for unobservable population traffic data estimation. Due to their enormous potential benefits in various applications, such as initial land-use planning (Yin et al., 2013; Ho and Wong, 2007), urban network traffic control (Daganzo, 2007; Geroliminis et al., 2013; Aboudolas and Geroliminis, 2013) and road pricing schemes (Zheng et al., 2012; Geroliminis and Levinson, 2009), these models have rapidly gained considerable attention in recent years. Generally, they can be classified as macroscopic cost flow (MCF) functions or macroscopic fundamental diagrams (MFDs). MCF functions are normally used for static analysis, and MFDs for dynamic analysis. As accurate and known macroscopic models are the necessary input for these beneficial applications, it is essential to estimate these models based on accurate traffic data collected from a whole network. However, due to the various limitations mentioned, direct measurements of essential traffic data are normally infeasible. Linear data projection offers a useful and practical framework for the unbiased estimation of these unobservable data, based on data assembled from various sources. In a study of MFDs, Geroliminis and Daganzo (2008) used linear data projection to derive the accumulation in a network, using the scaling factor of total-traffic-to-occupied-taxi ratio. In terms of MCFs, Wong and Wong (2015, 2016b, 2016c) estimated macroscopic Bureau of Public Road functions for networks in Hong Kong using real-world data assembled from GPS-equipped taxis and counting stations. With the total-traffic-to-occupied-taxi ratio as the chosen scaling factor, data scaling was used to estimate the total hourly traffic flows entering the sampled networks. The heterogeneity in the scaling factor stemmed from several factors, such as the various land-use purposes of lots (Meng et al., 2017b).

As linear data projection provides unbiased estimators of unobservable traffic data, it is intuitive to estimate models using linearly projected data. However, the implications of this assumption—most importantly, its effects on estimated parameters and their standard errors—have been largely ignored in the field. Directly estimating a model using linearly projected data is equivalent to neglecting the effects of scaling factor heterogeneity (or

variability). Ignoring spatial heterogeneity inherently assumes a constant scaling factor across space. Similarly, if information on temporal heterogeneity is lost, a uniform scaling factor is implied. Recently, Wong and Wong (2015) generically proved that systematic bias is embedded in the model parameters estimated from linearly projected data, disregarding the distribution of the scaling factor and the form of the model to be calibrated, as long as the scaling factor is subject to heterogeneity and the model is a nonlinear function of the scaling factor. In their study, however, only generalized multivariate polynomial (GMP) functions with fixed exponents, which are essentially linear regression models, were chosen for in-depth examination. Both analytical expressions quantifying the extent of the bias and global adjustment factors significantly reducing such bias were derived for GMP functions. However, many transport models are categorized not as GMP functions but as generic nonlinear models that necessitate nonlinear regressions. Therefore, Wong and Wong (2018) further explored the domain of nonlinear transport models. The authors found that the nonexistence or complexity of derivation of simple closed-form adjustment factors and the model specification error induced by linear data projection were the major pragmatic concerns raised by the adjustment factor approach. Inspired by the mechanism of systematic data point distortion induced by linear data projection, the authors proposed a mean value restoration (MVR) method that required only the first two moments of the scaling factor, and an extended MVR (EMVR) method that further captured higher-order moments with an assumed scaling factor distribution to remove embedded systematic biases in nonlinear transport models. In their study of MFD estimation, Du et al. (2016) proposed an algorithm based on k-means clustering analysis to address the unrealistic assumption of a constant scaling factor across the network. However, since the algorithm was designed for and focused on MFD data point estimation (i.e., network flow and density), it is usually not directly transferrable to other situations.

Even if unbiased parameters can be easily estimated using the methods proposed above, their statistical significance, and hence the validity of the models estimated, must be determined and tested by their standard errors. Thus, the effects of linear data projection on estimated standard errors are another critical research direction. Wong and Wong (2016c) generically proved that when linear data projection is used, heteroscedasticity is intrinsically introduced regardless of the form of model estimated, as long as the scaling factor is subject to heterogeneity. The resulting complicated error structures, which contain both random errors independent and identically distributed (*i.i.d.*) as normal and heteroscedasticity, undesirably violate the homoscedasticity assumption and result in biased standard errors. As these biased standard errors may be greater or smaller than the true values, they may lead either to the mistaken rejection of the true null hypotheses in statistical tests of significance, causing a type I error; or to the erroneous failure to reject the false null hypotheses, resulting in a type II error. To collectively re-capture lost heterogeneity and accurately estimate model parameters and their standard errors, the authors then focused on the family of GMP functions with fixed exponents and proposed analytical distribution free (ADF) and equivalent scaling factor (ESF) methods. However, these proposed methods are applicable only to GMP functions with fixed exponents, which are essentially linear regression models, although their shapes can be nonlinear. Methods of reducing the biases embedded in standard

error estimates have not been extended to generic nonlinear transport models requiring nonlinear regressions. Although unbiased nonlinear transport models can be estimated from linearly projected data using the proposed MVR and EMVR methods, the statistical significance of the parameter estimates remains unknown, and hence the validity of application of the estimated models is still unclear. This poses a new and important research question concerning the unbiased standard error estimation of generic nonlinear transport models based on linearly projected data.

The aim of this paper is to fill a gap in previous research—the absence of unbiased standard error estimation methods for use with nonlinear transport models estimated from linearly projected data. It first identifies the major practical difficulties with the traditional approach to standard error estimation in the situation under consideration. A bootstrapping MVR method combining the bootstrap resampling method and the MVR or EMVR method is proposed to accurately estimate standard errors. Comprehensive simulations with various settings are performed for GMP functions with relaxed exponents and multivariate exponential decay functions, which are the most commonly used nonlinear functions modeling traffic flow, to evaluate the capability and robustness of the proposed method in recovering the true standard errors. The results reveal that the method can be used to accurately estimate parameter standard errors. To illustrate real-world scenarios necessitating the use of linear data projection and the proposed bootstrapping MVR method, an MFD is estimated for a sampled network in Hong Kong, based on real-world data retrieved from GPS-equipped taxis and counting stations. The estimated parameter standard errors are used to assess the statistical significance of the estimated model. Standard errors are crucial when testing the statistical significance of an estimated model. Although the MVR and EMVR methods proposed by Wong and Wong (2018) provide unbiased estimated parameters for nonlinear transport models based on linearly projected data, no existing method offers unbiased standard errors that can be used to determine statistical significance. Therefore, this work is important and necessary. Moreover, the size of the model set that can be handled by the bootstrapping MVR method is much greater than that in Wong and Wong (2016c). Linear data projection is a simple, highly transferable and thus powerful data scaling method of providing unbiased estimators for unobservable traffic data, using traffic data collected with state of the art technology. Thus, developing methods of unbiased standard error estimation for nonlinear transport models estimated from linearly projected data contributes to the field not only by filling the abovementioned research gap, but also by facilitating the full utilization of this powerful data scaling method in the big data era.

The remainder of the paper is structured as follows. Section 2 explains the practical difficulties associated with the traditional approach to standard error estimations in cases of nonlinear transport models estimated from linearly projected data. The bootstrapping MVR method, which is expected to accurately estimate parameter standard errors for all nonlinear models, is proposed in Section 3. Comprehensive simulation studies with various settings, conducted to evaluate the performance of the proposed method, are reported in Section 4. Next, a case study of MFD estimation, illustrating real-world scenarios requiring linear data

projection, and the proposed method are presented. The last section concludes the study with the results, major findings and potential directions for future research.

2. Practical difficulties of the traditional approach to standard error estimation

Using the traditional approach, parameter standard errors (or their approximations) are typically estimated using an empirical analog of an explicit theoretical formula derived from an assumed model (Shao and Tu, 1995). More specifically, if the parameters $\boldsymbol{\beta}$ of a specific model $G(\boldsymbol{\beta})$ can be estimated using a closed-form expression, the explicit expression of $Var(\boldsymbol{\beta})$ can be readily derived by taking the variance of the closed-form expression as a function of certain unknown quantities. The estimated covariance matrix of $\boldsymbol{\beta}$ or the estimator of $Var(\boldsymbol{\beta})$, $\widehat{Var(\boldsymbol{\beta})}$ can be derived by substituting these unknown quantities with their unbiased estimators. The standard error estimates of the model parameters are given by the square roots of the diagonal elements of $\widehat{Var(\boldsymbol{\beta})}$. A typical example of the traditional approach is the ADF method proposed by Wong and Wong (2016c). As only GMP models with fixed exponents, which are essentially linear regression models, were considered, a closed-form expression for the estimator of $\boldsymbol{\beta}$ was obtained. As a result, standard errors could be easily estimated using the traditional approach.

However, the normal equations are nonlinear in their parameters in the context of nonlinear transport models that necessitate nonlinear regression. Hence, they cannot be solved by a finite sequence of standard operations; instead, iterative methods such as the Gauss–Newton method or the method of steepest descent must be used. In other words, closed-form expressions estimating the model parameters $\boldsymbol{\beta}$ may not exist. In such cases, the traditional approach, in which $Var(\boldsymbol{\beta})$ is obtained by taking the variance of the closed-form expression, may not be practicable. Furthermore, according to Wong and Wong (2016c), if the scaling factors used are subject to a certain type of heterogeneity, heteroscedasticity is inherently introduced by linear data projection. The error term, in this case, is neither simply $i.i.d.$ as normal nor heteroscedastic, but a combination of both. Additionally, the nonzero value of the expectation of the composite error term (Wong and Wong, 2015; 2018) further complicates the problem. Thus, the complex structure of the composite random error term tends to make the derivation of $\widehat{Var(\boldsymbol{\beta})}$ immensely cumbersome or even impossible, even if closed-form expressions do exist. These practical difficulties imply that the traditional approach to standard error estimations may not be a viable solution to the research question under consideration. Instead, a generic and flexible method that can accurately estimate the parameter standard errors of any nonlinear transport models based on linearly projected data is needed. A new bootstrap resampling method incorporating either the MVR or the EMVR method, known as the bootstrapping MVR method, is proposed in this study.

3. Bootstrapping MVR method

Bootstrapping is a powerful and appealing statistical method that can be used to estimate the sampling distribution of almost any statistic through random sampling with

replacement. The properties of the statistic, such as its variance, can then be derived from the approximated sampling distribution. However, it should be stressed that bootstrapping alone cannot solve the problem considered here, as the expected value of the distorted composite error term is not equal to zero (Wong and Wong, 2015; 2018). Thus, either the MVR or EMVR method must be incorporated. This section first briefly sets out the central principle of the MVR and EMVR methods. It then introduces the notion of bootstrapping and proposes the bootstrapping MVR method for unbiased standard error estimations in cases in which generic nonlinear transport models are estimated using linearly projected data.

3.1 Central principle of MVR and EMVR methods

To remove the embedded systematic biases in nonlinear transport models, Wong and Wong (2018) investigated the mechanism of systematic data point distortion induced by linear data projection and proposed the MVR and EMVR methods to account for the nonzero mean value of the complex error term arising from linear data projection. Define $G(\boldsymbol{\beta}; z)$: $\mathbb{R}^m \rightarrow \mathbb{R}$ as a highly differentiable function of any form, where $\boldsymbol{\beta}$ is the model parameter vector. The unobservable independent variable z is represented by linear combinations of a set of observable independent variables and scaling factors; that is, $z = \sum_{i=1}^{m} f_i x_i \,, \forall i \in \{1, 2, \dots, m\}$, where x_i represents the observable independent variables, f_i is the scaling factor of x_i, which is assumed to follow any distribution with mean $\bar{f}$ and variance ${\sigma_f}^2$ and m is the number of terms constituting z. The authors found that when linear data projection was used, the unobservable data points originally lying on the true model, $G(\boldsymbol{\beta}; z)$ or $G(\boldsymbol{\beta}; \bar{z})$, were projected systematically onto the projection plane such that the mean value of the set of resulting linearly projected data, $\bar{y}^*$, deviated from the true model by distance Δy and lay on the expectation function of the linearly projected data, $E[G(\boldsymbol{\beta}; z)]$. The diagram on the left-hand side of Fig. 1 illustrates the mechanism described. For more details, interested readers may refer to Wong and Wong (2018).

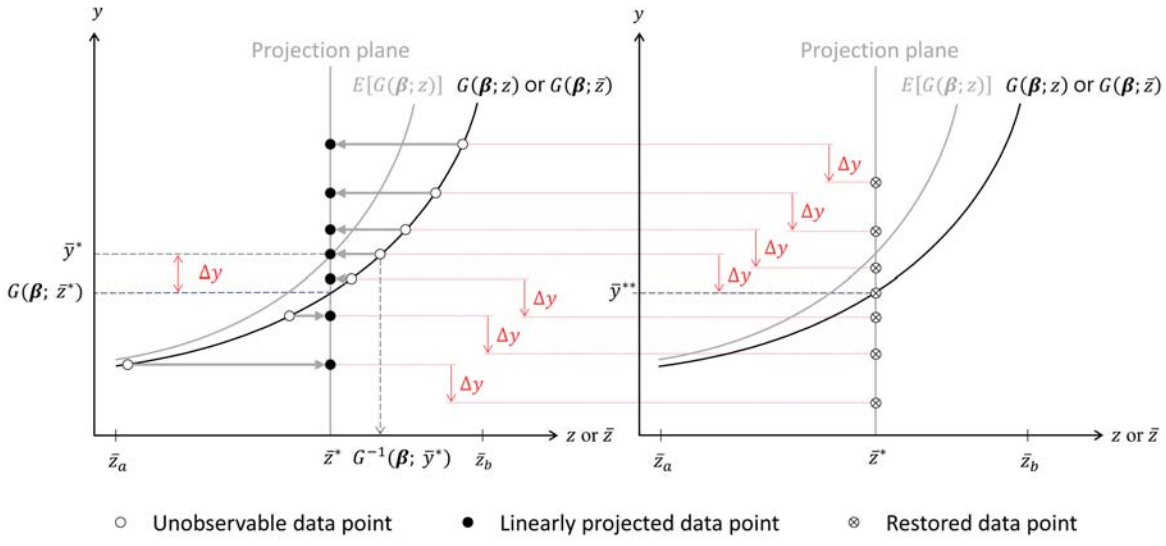


Fig. 1. Illustration of the central principle of the MVR and EMVR methods

The general form of the expectation function of linearly projected data (Wong and Wong, 2018) can be represented by Eq. (1):

$$E[G(\boldsymbol{\beta};z)] = G(\boldsymbol{\beta};\bar{z}) + \frac{1}{2!}{\sigma_f}^2 \sum_{i=1}^{m} \frac{\partial^2 G(\boldsymbol{\beta};\bar{z})}{\partial {f_i}^2} + \frac{1}{3!} S_f \sum_{i=1}^{m} \frac{\partial^3 G(\boldsymbol{\beta};\bar{z})}{\partial {f_i}^3} + \frac{1}{4!} K_f \sum_{i=1}^{m} \frac{\partial^4 G(\boldsymbol{\beta};\bar{z})}{\partial {f_i}^4} + \cdots, \quad (1)$$

where S_f and K_f are the skewness and kurtosis, respectively, of the scaling factor. The rth order approximation of the expectation function, $E_r[G(\boldsymbol{\beta};z)]$, can be obtained by truncating all of the terms behind the rth term in Eq. (1), $\forall r \in \mathbb{N}^+$. It should be noted that the linear approximation $G(\boldsymbol{\beta};\bar{z})$ is identical to the true model $G(\boldsymbol{\beta};z)$; thus the deviation Δy can generally be approximated by Eq. (2):

$$\Delta y \cong E_r[G(\boldsymbol{\beta};z)] - G(\boldsymbol{\beta};\bar{z}), \quad (2)$$

where $r \geq 2$. As illustrated in Fig. 1, the central goal of both the MVR method and the EMVR method is to restore the mean value of linearly projected data $\bar{y}^*$ by accounting for the deviation Δy in model estimations such that the mean value of the restored data, $\bar{y}^{**}$, is very close to or lies on the true model $G(\boldsymbol{\beta};z)$. For a dataset of size N, the corresponding least square function, S, is given by Eq. (3):

$$S = \sum_{j=1}^{N} \left[(y_j - \Delta y_j) - G\left(\widehat{\boldsymbol{\beta}};\bar{z}_j\right)\right]^2 = \sum_{j=1}^{N} \left\{y_j - E_r\left[G\left(\widehat{\boldsymbol{\beta}};\bar{z}_j\right)\right]\right\}^2. \quad (3)$$

The MVR method, which is equivalent to the direct model estimation of linearly projected data based on the quadratic approximation of their expectation function (i.e., $r = 2$), is flexible, because usually only the first two moments of the scaling factor are available. In contrast, the EMVR method can further enhance the accuracy of estimation by capturing higher-order moments of the scaling factor (i.e., $r > 2$) based on an assumed distribution of the scaling factor. Generally, the quartic approximation of the expectation function (i.e., $r = 4$) can achieve a satisfactory level of accuracy (Wong and Wong, 2018).

3.2 Formulation of the bootstrapping MVR method

Statistics are random quantities whose probability distributions constitute their sampling distributions. As these sampling distributions and their properties typically depend

on underlying unknown populations, the distributions and properties are naturally unknown (Shao and Tu, 1995). The bootstrapping method, first introduced by Efron (1979), provides a simple and powerful means of approximating sampling distributions. These approximations make it possible to estimate statistical properties, such as standard errors. As bootstrapping involves random sampling with N replacements from an empirical distribution constituted by an observed dataset of size N, this method is categorized as nonparametric resampling in the statistics field.

Bootstrapping provides a simple and powerful means of making statistically precise estimations, primarily because it offers two distinct advantages. First, the resampling method replaces the theoretical derivations required by traditional methods. Bootstrapping uses the Monte Carlo approximation to estimate the precision of statistics. This approximation bypasses the complicated and tedious theoretical derivations of analytical formulae. As discussed in Section 2, these derivations are usually very complicated or even impossible to perform when generic nonlinear transport models that necessitate nonlinear regression and complex error structures caused by linear data projection are considered. Second, bootstrapping frees the estimation method of certain restrictive parametric assumptions usually imposed by the traditional approach in the specific contexts under consideration. Instead, it accommodates the complex composite error term and indirectly assesses the properties of the sampling distributions of the statistics via resampling, making it feasible for use in a wide range of cases. However, the two main advantages of the bootstrap resampling method are achieved at the expense of greater computational cost. Fortunately, due to the availability of powerful and inexpensive modern computers, this cost is acceptable. Thus, this relatively computer-intensive method is applicable to a broad range of problems, and has gained considerable attention in the field of applied statistics, although it has rarely been applied in the transportation field.

In the context of generic nonlinear transport models estimated from linearly projected data, however, the expected value of the composite error is nonzero (Wong and Wong, 2015; 2018). Thus, to accurately estimate the parameter standard errors, a new bootstrap resampling method that incorporates either the MVR or the EMVR method is proposed. Given N observed two-dimensional vectors $(y_i, \bar{z}_i), \forall i \in [1, N]$, where y_i represents the observed dependent variable and $\bar{z}_i$ represents the linearly projected data (i.e., $\bar{z}_i = \sum_{j=1}^{m} \bar{f} x_j$) generated from an unknown two-dimensional population distribution Ψ, the N observed vectors constitute an empirical distribution Ψ_N. Assuming that $G(\boldsymbol{\beta})$ is a candidate model estimated on the basis of linearly projected data using either the MVR or the EMVR method, the following formulated bootstrapping MVR method can be used to estimate the bootstrap standard errors of the estimated parameters, $\widehat{\boldsymbol{\beta}}$.

1) To obtain a bootstrap sample, randomly sample N vectors with replacements from the empirical distribution, Ψ_N. Ψ_N^* denotes the distribution of the bootstrap sample.

2) Based on the bootstrap sample, use the MVR or the EMVR method to estimate the candidate model $G(\boldsymbol{\beta})$ and generate a bootstrapping MVR estimate $\boldsymbol{\beta}^*$.

3) To generate M bootstrapping MVR estimates, repeat steps 1 and 2 M times, where M is large (typically $1{,}000$ or $10{,}000$). The M bootstrapping MVR estimates constitute a histogram that is a Monte Carlo approximation of the bootstrap distributions for these bootstrap estimates.

4) Compute the standard deviations of the M bootstrapping MVR estimates, which are the bootstrap standard error estimates of $\widehat{\boldsymbol{\beta}}$.

The fundamental concept of bootstrapping MVR method, illustrated in Fig. 2, is to mimic inferences about an unknown population distribution Ψ from sample data Ψ_N using inferences about an empirical distribution Ψ_N from resampled data Ψ_N^*. Ideally, if the population distribution Ψ is known, a set of MVR parameter estimates $\widehat{\boldsymbol{\beta}}$ can be obtained using a set of sample data Ψ_N drawn from Ψ. By repeating this step R times, histograms can be formed gradually for the MVR parameter estimates, and their standard deviations can serve as the unbiased standard error estimators. As the number of repetitions, R, tends to infinity, these histograms tend to the sampling distributions of the MVR parameter estimates.

However, the population distribution Ψ is not known in reality. Therefore, using the bootstrap resampling method, the observed empirical distribution Ψ_N is analogous to the unknown population distribution Ψ, and the resampled data Ψ_N^* play the role of data sample Ψ_N. According to the Glivenko–Cantelli theorem, the empirical distribution Ψ_N uniformly converges to the population distribution Ψ as the sample size, N, increases. Thus, when N is sufficiently large, the observed empirical distribution Ψ_N can be regarded as a reasonable and approximate distribution of the unknown population distribution Ψ. The resulting bootstrap standard errors are estimates of the parameter standard errors. As the bootstrap resampling method comprises repetitions of random sampling with replacements N times from an observed dataset of size N, the total number of possible ordered bootstrap samples is N^N. Theoretically, all of the bootstrap estimates can be obtained by enumerating all of the possible ordered bootstrap samples (Chernick and LaBudde, 2011). The true bootstrap distributions of the bootstrapping MVR estimates, which are approximations of the unknown sampling distributions of the MVR parameter estimates, can be constituted by the N^N set of bootstrapping MVR estimates.

However, this approach is only feasible if the observed dataset is sufficiently small (Fisher and Hall, 1991); if N is too large, the computational cost may be huge. In practice, a commonly used alternative is Monte Carlo approximation with a large number of repetitions of M. As M increases, the histograms constituted by the bootstrapping MVR estimates approach the true bootstrap distributions for the bootstrapping MVR estimates. However, it must be stressed that increasing the number of bootstrapping MVR estimates, M, does not increase the amount of information in the original data sample; it only reduces the effects of the random sampling error arising from the bootstrap procedure itself. An M with a size of $1{,}000$ or $10{,}000$ can generally achieve a satisfactory level of accuracy.

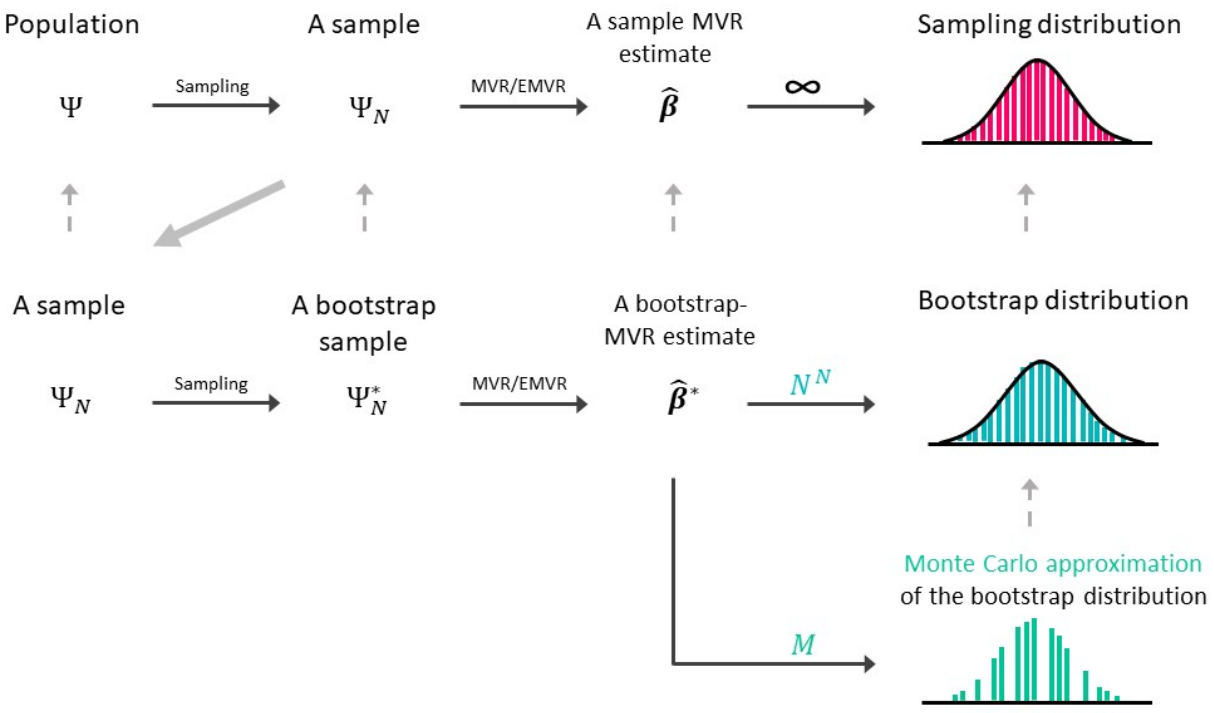


Fig. 2. Illustration of the fundamental principle of the bootstrapping MVR method

4. Numerical examples

Section 3 proposes the bootstrapping MVR method capable of accounting for the complex error structure of a mixture of *i.i.d.* random errors and heteroscedasticity resulting from linear data projection. The method also accurately estimates the standard errors of nonlinear transport models based on linearly projected data. This section evaluates the capability and robustness of the proposed method in recovering true standard errors via 12 simulation studies based on the two nonlinear functions most commonly used to model traffic flow relationships. To illustrate a real-world scenario necessitating the use of linear data projection and the proposed method, an MFD is estimated for a sample network in Hong Kong.

4.1 Simulation studies

Comprehensive simulation studies based on GMP functions with relaxed exponents and multivariate exponential decay functions are conducted in this subsection to evaluate the performance of the proposed method in recovering true standard errors. The two selected functions are the functions most commonly used to depict traffic flow relationships. In addition to variation in the chosen functional forms, the robustness of the proposed method is examined with regard to different scaling factor distributions and the numbers of linear combinations of the scaling factor and the observable independent variable.

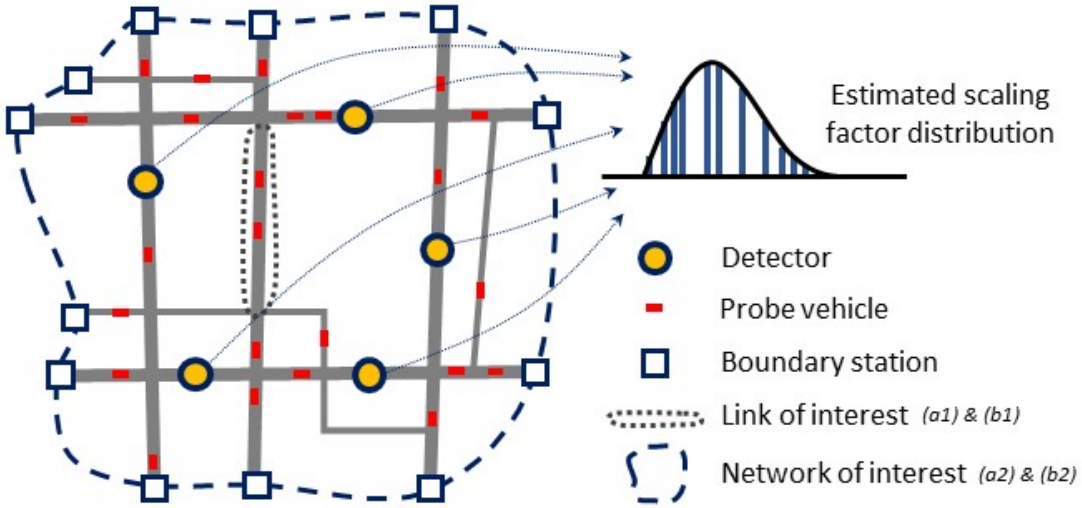
4.1.1 Two chosen classical traffic flow models

The two nonlinear models chosen for the simulation studies are the GMP function with relaxed exponents (Eq. 4) and the multivariate exponential decay function (Eq. 5). These are the functions most commonly used to depict traffic flow relationships. The GMP function with relaxed exponents is defined as follows:

$$y = \beta_0 + \beta_n z^n + \varepsilon = \beta_0 + \beta_n (\sum_{i=1}^{m} f_i x_i)^n + \varepsilon, \quad (4)$$

where β_0, β_n and n are the model parameters and ε is the random error normally assumed to follow $N(0, \sigma^2)$. The GMP function is commonly used to model an increasing relationship between traffic quantities. Cost flow functions or volume delay functions, which can be categorized as link-based ($m = 1$) or area-based ($m > 1$), are among the typical classes of traffic flow models showing such an increasing relationship. Both the Bureau of Public Roads (BPR) function adopted in the *Highway Capacity Manual* (Transportation Research Board, 2000) and the macroscopic BPR (MBPR) function studied by Wong and Wong (2015, 2016b, 2016c) have the same form as the GMP function. In this physical context, β_0 represents the free-flow travel time per unit of distance; β_n is the product of the congestion sensitivity parameter and free-flow travel time; n is the model nonlinearity parameter; y represents the travel time per unit of distance across the link or network under consideration; and z is the total traffic flow associated with the corresponding link or network.

Given that total traffic flow is only observable at a subset of links in a network with detectors installed, whereas probe vehicle flow is observable at every link within the network, model estimations of both the link- and the area-based BPR functions require the use of both linear data projection and the proposed method. When estimating the BPR function of a link without a detector (see scenario (a1) in Fig. 3), y is the travel time per unit of distance associated with that link. Assuming that these probe vehicles travel at speeds similar to those of the surrounding traffic, the reciprocal of the hourly space-mean speed of a probe vehicle is an estimator for y. The total-traffic-to-probe-vehicle ratio at the link can be defined as the scaling factor f, and x is the hourly probe vehicle flow across the link. Their product is the hourly total traffic flow. However, the link's f is unknown due to the absence of a detector. Only the scaling factors of nearby links with detectors are observable. Due to geographical proximity, the scaling factors of all of the nearby links can be assumed to follow a distribution. The sampled observable scaling factors can be used to estimate the mean and variance of the distribution. As the scaling factor mean is the most commonly observed total-traffic-to-probe-vehicle ratio, it is used to estimate the unbiased hourly total traffic flow across the link of interest in a linear data projection (i.e., $\bar{f}x$). If the exponent of the BPR function is relaxed and allowed to vary, the function is a nonlinear regression model; thus, the proposed bootstrapping MVR method is required for unbiased standard error estimation.



GMP function with relaxed exponent (Eq. (4))

Scenario (a1): Link-based cost flow function (BPR function)

- m: Number of links of interest ($m = 1$)
- β_0: Free-flow travel time per unit distance (T_f)
- β_n: Product of free-flow travel time and congestion sensitivity parameter ($T_f\alpha$)
- n: Model nonlinearity
- y: Travel time per unit distance associated with the link of interest (T)
- f: Total-traffic-to-probe-vehicle ratio associated with the link of interest
- x: Hourly probe vehicle flow associated with the link of interest (q_p)

Scenario (a2): Area-based cost flow function (MBPR function)

- m: Number of boundary stations ($m > 1$)
- β_0: Free-flow travel time per unit distance (T_f)
- β_n: Product of free-flow travel time and congestion sensitivity parameter ($T_f\alpha$)
- n: Model nonlinearity
- y: Travel time per unit distance associated with the network of interest (T)
- f_i: Total-traffic-to-probe-vehicle ratio of boundary station i
- x_i: Hourly probe vehicle flow entering the network via boundary station i (q_{pi})

Multivariate exponential decay function (Eq. (5))

Scenario (b1): Link-based speed-density relationship (Underwood's model)

- m: Number of links of interest ($m = 1$)
- a: Free-flow speed (u_f)
- b: Optimal density (k_o)
- y: Space-mean speed associated with the link of interest (u)
- f: Total-traffic-to-probe-vehicle ratio associated with the link of interest
- x: Average probe vehicle density associated with the link of interest (k_p)

Scenario (b2): Area-based speed-density relationship (MFD)

- m: Number of links within the network of interest ($m > 1$)
- a: Network free-flow speed (u_f)
- b: Network optimal density (k_o)
- y: Space-mean speed associated with the network of interest (u)
- f_i: Total-traffic-to-probe-vehicle ratio of boundary station i
- x_i: Average probe vehicle density at link i (k_{pi})

Fig. 3 Real-world cases of estimating link- and area-based cost flow functions and the speed-density relationship that necessitate the use of linear data projection and the proposed bootstrapping MVR method for unbiased standard error estimation

However, if an entire network is of interest (see scenario (a2) in Fig. 3), an MBPR function should be considered. In this case, m is the number of links intercepting the boundaries of the sampled network, which are defined as boundary stations (Wong and Wong, 2015, 2016a, 2016b). The reciprocal of the hourly space-mean speed of the probe vehicles in the sampled network can be used to approximate the travel time per unit of distance across the network. f_i is the total-traffic-to-probe-vehicle ratio at boundary station i, and x_i represents the hourly probe vehicle flow entering the network through boundary station i. Again, as detectors may be absent, the values of the scaling factors may be unknown. By inferring the scaling factor distribution using sampled total-traffic-to-probe-vehicle ratios at links with detectors in the network, both the mean and variance of the scaling factor can be obtained. The linear data projected can then be used to obtain an unbiased estimate of the hourly traffic flow entering the network (i.e., $\sum_{i=1}^{m} \bar{f} x_i$). The proposed method is necessary to estimate unbiased standard errors in this case.

The second model chosen for the simulation studies is the multivariate exponential decay function:

$$y = a\, exp\left(\frac{-z}{b}\right) + \varepsilon = a\, exp\left(\frac{-\sum_{i=1}^{m} f_i x_i}{b}\right) + \varepsilon, \tag{5}$$

where a and b are the model parameters and ε is the random error normally assumed to follow $N(0, \sigma^2)$. The multivariate exponential decay function is normally used to depict strictly decreasing relationships between traffic quantities. Link-based ($m = 1$) and area-based ($m > 1$) speed-density relationships are among the typical classes of traffic flow models with such a decreasing relationship. Thus, this functional form can be a candidate model for speed-density relationships. If a single link is considered, Eq. (5) reduces to the classical Underwood model. If the MFD of a network is of interest, the multivariate exponential decay function can be a candidate model for statistical fitting. In these physical contexts, parameter a represents the free-flow speed, parameter b represents the optimal traffic density at which the throughput is maximum and z is the traffic density of the link or network considered.

Similarly, estimating both the link- and the area-based speed-density relationship necessitates both linear data projection and the proposed method if probe vehicle flow is observable for each link in the sampled network and the observation of total traffic flow is limited to a subset of links outfitted with detectors. Using Underwood's model of a link (see scenario (b1) in Fig. 3), the space-mean speed of the probe vehicles traveling on the link over a short period (e.g., 10 min) is an unbiased estimator for y, assuming that all of the vehicles on the same link travel at similar speeds. The total-traffic-to-probe-vehicle ratio at that link can be taken as the scaling factor f, and x represents the average probe vehicle density at the same link. As the scaling factor f of the link of interest is unobservable, linear data projection must be used for unbiased estimations of traffic density, using the scaling factor mean inferred from the sampled total-traffic-to-probe-vehicle ratios of nearby links (i.e., $\bar{f} x$). The

proposed bootstrapping MVR method must then be applied for unbiased standard error estimation, as Underwood's model is a nonlinear regression model.

In the case of area-based speed-density relationships (scenario (b2) in Fig. 3), m is the total number of links in the selected network. The space-mean speed of all of the probe vehicles in the sampled network in a short period (e.g., 10 min) can be taken as an approximation of y. f_i is the total-traffic-to-probe-vehicle ratio at link i, and x_i is the average probe vehicle density at link i. Again, due to the possible absence of detectors, the values of the scaling factors may be unobservable. Linear data projection has to be leveraged to estimate the network traffic density using the scaling factor mean estimated from the sampled total-traffic-to-probe-vehicle ratios (i.e., $\sum_{i=1}^{m} \bar{f} x_i$). As the multivariate exponential decay function is a nonlinear regression model, the proposed method is necessary for unbiased standard error estimations.

4.1.2 Data generation

To evaluate the performance and robustness of the proposed method in recovering true standard errors, 12 simulation cases with various combinations of the chosen models (GMP functions with relaxed exponents and multivariate exponential decay functions), number of linear combinations of the scaling factors and observable independent variables (i.e., $m = 1, 2$ and 3) and distributions of the scaling factors (i.e., normal and lognormal) were considered. In addition, model estimations of both the link- and the area-based cost-flow relationships and the speed-density relationships in the situations described above could be mimicked by these simulation studies.

We set $\beta_0 = 3$, $\beta_n = 1$, $n = 3$, $a = 30$ and $b = 2000$. As m was set at 1, 2 or 3, three sets of observable independent variables with $10{,}000$ observations were sampled for each selected model. x_i can be used to represent the hourly probe vehicle flow across link i (or boundary station i) or the average probe vehicle density at link i, according to the above-described physical contexts. In practice, the observable independent variable can represent any traffic variable with any probability distribution. Each observation comprised one x when $m = 1$. Similarly, two or three xs were independently generated from the selected distribution for each observation for cases in which $m = 2$ or 3, respectively. Without a loss of generality, uniform distributions were chosen to generate the observable independent variable. For the GMP function, x was sampled from $Unif(0, 1)$. For the multivariate exponential decay function, x was sampled from $Unif(0, 100)$. These six sets of observable independent variables were used in all of the simulations to avoid sampling errors during data generation. The scaling factor f for each x was generated from a normal or a lognormal distribution, with $\bar{f} = 1$ and $\sigma_f = 0.2$ for the GMP function and $\bar{f} = 100$ and $\sigma_f = 20$ for the multivariate exponential decay function. In addition, $10{,}000$ random errors, ε, for the $10{,}000$ observations were generated from $N(0, 0.1)$ for the GMP function and $N(0, 1)$ for the multivariate exponential decay function. The corresponding $10{,}000$ dependent variables, y, were evaluated using the true values of the model parameters and the sampled x, f and ε for each simulation case. Assuming that the value of each f was not available, z could only

be estimated based on a linear data projection in which the unknown f was replaced by the scaling factor mean (i.e., $\bar{z} = \sum_{i=1}^{m} \bar{f} x_i$).

4.1.3 Simulation results

Regression analysis was conducted on the basis of linearly projected data using the MVR and EMVR methods proposed by Wong and Wong (2018). For the GMP function, the chosen value of n was 3, so its expectation function could only be approximated up to its cubic form. In particular, for the simulation case with a scaling factor assumed to follow a normal distribution, the cubic approximation was automatically reduced to a quadratic approximation due to the zero skewness of the normal distribution. Therefore, the quadratic approximation and the cubic approximation were used in the simulation cases associated with normally distributed and lognormally distributed scaling factors, respectively. In contrast, the multivariate exponential decay function was infinitely differentiable. In general, because the fourth-order approximation yielded a satisfactory level of accuracy, the quartic approximation of the expectation function was used in the regression analysis. Calibrated model parameters (i.e., $\hat{\beta}_0$, $\hat{\beta}_n$, $\hat{n}$, $\hat{a}$ and $\hat{b}$) and reported standard errors (i.e., $\widehat{RSE(\hat{\beta}_0)}$, $\widehat{RSE(\hat{\beta}_n)}$, $\widehat{RSE(\hat{n})}$, $\widehat{RSE(\hat{a})}$ and $\widehat{RSE(\hat{b})}$) were obtained from the model estimation. The proposed bootstrapping MVR method was then leveraged to obtain the bootstrap standard errors of the estimated parameters (i.e., $\widehat{SE_B(\hat{\beta}_0)}$, $\widehat{SE_B(\hat{\beta}_n)}$, $\widehat{SE_B(\hat{n})}$, $\widehat{SE_B(\hat{a})}$ and $\widehat{SE_B(\hat{b})}$). The number of bootstrap samples, M, was set at $10{,}000$.

The number of repetitions for each simulation case R was set at $10{,}000$. Therefore, to obtain the means and standard deviations of the parameter estimates, the means of the reported standard errors and the bootstrap standard errors, $10{,}000$ repetitions were conducted for each simulation case with resampled scaling factors and random errors. Tables 1 and 2 present the simulation results for the six cases using the GMP model. Similarly, Tables 3 and 4 summarize the simulation results for the six cases using the multivariate exponential decay function.

Table 1

Means and standard deviations of the parameter estimates of the GMP functions based on the MVR or the EMVR method

Assumed scaling factor distribution	m	Mean			Standard deviation		
		$\hat{\beta}_0$ (% $error$)	$\hat{\beta}_n$ (% $error$)	$\hat{n}$ (% $error$)	$\hat{\beta}_0$	$\hat{\beta}_n$	$\hat{n}$
Normal distribution	1	3.000 (0.00%)	1.000 (−0.01%)	3.001 (+0.03%)	0.0028	0.0127	0.0752
	2	3.000 (0.00%)	1.000 (−0.01%)	3.001 (+0.03%)	0.0186	0.0328	0.0669
	3	3.000 (0.00%)	1.001 (+0.05%)	3.001 (+0.04%)	0.0625	0.0566	0.0679
Lognormal distribution	1	3.000 (0.00%)	1.000 (+0.01%)	3.003 (+0.11%)	0.0030	0.0137	0.0840
	2	3.000 (0.00%)	1.000 (+0.02%)	3.001 (+0.02%)	0.0199	0.0353	0.0720
	3	3.001 (+0.03%)	1.000 (−0.01%)	3.002 (+0.08%)	0.0672	0.0590	0.0704

Table 2

Means of the reported standard errors and bootstrap standard errors for the GMP functions

Assumed scaling factor distribution	m	Mean			Mean		
		$\widehat{RSE(\hat{\beta}_0)}$ (% $error$)	$\widehat{RSE(\hat{\beta}_n)}$ (% $error$)	$\widehat{RSE(\hat{n})}$ (% $error$)	$\widehat{SE_B(\hat{\beta}_0)}$ (% $error$)	$\widehat{SE_B(\hat{\beta}_n)}$ (% $error$)	$\widehat{SE_B(\hat{n})}$ (% $error$)
Normal distribution	1	0.0048 (+70.73%)	0.0085 (−32.51%)	0.0586 (−22.04%)	0.0028 (+0.53%)	0.0125 (−1.17%)	0.0750 (−0.26%)
	2	0.0217 (+16.53%)	0.0243 (−25.81%)	0.0382 (−42.83%)	0.0186 (−0.14%)	0.0328 (−0.07%)	0.0670 (+0.05%)
	3	0.0597 (−7.28%)	0.0339 (−40.15%)	0.0350 (−48.40%)	0.0639 (−0.76%)	0.0565 (−0.99%)	0.0673 (−0.90%)
Lognormal distribution	1	0.0053 (+74.28%)	0.0096 (−29.98%)	0.0650 (−22.62%)	0.0030 (−0.53%)	0.0136 (−0.52%)	0.0837 (−0.36%)
	2	0.0232 (+16.71%)	0.0261 (−26.03%)	0.0410 (−43.09%)	0.0198 (−0.26%)	0.0352 (−0.43%)	0.0715 (−0.61%)
	3	0.0626 (−6.83%)	0.0356 (−39.63%)	0.0368 (−47.75%)	0.0668 (−0.70%)	0.0586 (−0.62%)	0.0704 (−0.04%)

The model estimation results for the GMP functions with relaxed exponents, shown in Table 1, revealed that the parameter estimates and their true values were extremely close (i.e., $\beta_0 = 3$, $\beta_n = 1$ and $n = 3$). In terms of magnitude, the percentage errors of all of the parameter estimates were less than or equal to 0.11%. This demonstrated the effectiveness of the MVR and EMVR methods proposed by Wong and Wong (2018). Quadratic approximation and cubic approximation were the exact expectation functions of the linearly projected data for the simulation cases associated with the normally distributed and the lognormally distributed scaling factors, respectively. Thus, any remaining percentage errors in the parameter estimates should have emanated purely from sampling errors in the random errors and from the scaling factors in the $10{,}000$ repetitions. Unbiased estimators of the parameter standard errors were given by the standard deviations of the parameter estimates. However, as shown in Table 2, the reported standard errors seriously deviated from these unbiased estimators. The magnitudes of the percentage errors of the reported standard errors ranged from around 7% to 74%.

The last three columns of Table 2 reveal the bootstrap standard errors of the parameter estimates for the GMP functions estimated using the proposed bootstrapping MVR method. As demonstrated by the simulation results, the magnitudes of the percentage errors of the bootstrap standard errors were all well within 1.17%. The results indicated that the proposed bootstrapping MVR method outperformed conventional regression procedures, which were unable to account for the complex error structure comprising both the heteroscedasticity induced by linear data projection and random errors $i.i.d$ as normal. No specific patterns in the percentage errors of the bootstrap standard errors were identified along the dimensions of m or in the choice of scaling factor distribution. For instance, the magnitudes of the percentage errors of $\hat{\beta}_0$ dropped from 0.53% to 0.26% and then increased to 0.70% as m increased from 1 to 3, in the case of the lognormally distributed scaling factor. Similarly, when $m = 2$, the magnitude of the percentage error of $\hat{\beta}_n$ in the case with the lognormally distributed scaling factor (0.43%) was greater than that in the case with the normally distributed scaling factor (0.07%). However, when $m = 3$, the magnitude of the percentage error of $\hat{\beta}_n$ in the situation with a lognormally distributed scaling factor (0.62%) was smaller than that in the case with normally distributed scaling factors (0.99%). The remaining minimal errors among the bootstrap standard errors stemmed from two types of error that are thoroughly explained later in this subsection.

Table 3

Means and standard deviations of the calibrated parameters of the multivariate exponential decay functions based on the EMVR method

Assumed scaling factor distribution	m	Mean		Standard deviation	
		$\hat{a}$ (% $error$)	$\hat{b}$ (% $error$)	$\hat{a}$	$\hat{b}$
Normal distribution	1	29.999 (0.00%)	2000.293 (+0.01%)	0.0667	8.1346
	2	29.992 (−0.03%)	2001.076 (+0.05%)	0.2914	14.3951
	3	29.978 (−0.07%)	2002.523 (+0.13%)	0.9915	32.6607
Lognormal distribution	1	30.005 (+0.02%)	1999.135 (−0.04%)	0.0645	7.3904
	2	29.998 (−0.01%)	2000.094 (0.00%)	0.2797	13.6727
	3	30.002 (+0.01%)	2000.944 (+0.05%)	0.9769	31.7319

Table 4

Means of the reported standard errors and bootstrap standard errors for the multivariate exponential decay function

Assumed scaling factor distribution	m	Mean		Mean	
		$\widehat{RSE(\hat{a})}$ (% $error$)	$\widehat{RSE(\hat{b})}$ (% $error$)	$\widehat{SE_B(\hat{a})}$ (% $error$)	$\widehat{SE_B(\hat{b})}$ (% $error$)
Normal distribution	1	0.0064 (−90.36%)	58.9611 (+624.81%)	0.0666 (−0.16%)	8.0991 (−0.44%)
	2	0.0470 (−83.89%)	135.2453 (+839.52%)	0.2914 (0.00%)	14.4448 (+0.35%)
	3	0.4450 (−55.12%)	628.0568 (+1822.97%)	1.0332 (+4.20%)	33.6018 (+2.88%)
Lognormal distribution	1	0.0056 (−91.31%)	51.1547 (+592.18%)	0.0648 (+0.48%)	7.5551 (+2.23%)
	2	0.0442 (−84.21%)	126.7450 (+826.99%)	0.2848 (+1.83%)	13.8660 (+1.41%)
	3	0.4402 (−54.94%)	618.1656 (+1848.09%)	1.0085 (+3.24%)	32.7512 (+3.21%)

As shown in Table 3, the model parameters of the multivariate exponential decay functions estimated using the EMVR method were extremely close to their true values (i.e., $u_f = 30$ and $k_o = 2000$). In terms of magnitude, the percentage errors of all of the cases were less than or equal to 0.13%. Although the multivariate exponential decay function is infinitely differentiable, a quartic approximation of the expectation function was expected to attain a satisfactory level of accuracy in the model estimation. Therefore, the remaining percentage errors in the parameter estimates were assumed to have stemmed primarily from unrecovered minimal effects due to the scaling factors' higher-order moments, and from the sampling errors of the random errors and scaling factors. Again, unbiased estimates of the parameter standard errors were given by the standard deviations of the parameter estimates. However, as shown in Table 4, the reported standard errors were seriously biased. The magnitudes of the percentage errors of the standard errors ranged from around 55% to 1848%.

The last two columns of Table 4 reveal the bootstrap standard errors in the parameter estimates for the multivariate exponential decay functions estimated using the proposed bootstrapping MVR method. The simulation results showed that the magnitudes of the percentage errors for the bootstrap standard errors were less than or equal to 4.20% for all of the simulation cases. The magnitudes of the percentage errors for the reported standard errors were much greater than those for the bootstrap standard errors. Compared with conventional regression, therefore, the proposed bootstrapping MVR method dramatically improved the accuracy of the standard error estimates. As in the cases of GMP functions, no clear patterns in the percentage errors of the bootstrap standard errors were identified with certainty along the dimensions of m or in the choice of scaling factor distribution. For instance, in cases with lognormally distributed scaling factors, the magnitude of the percentage errors of $\hat{b}$ dropped from 2.23% to 1.41% and then increased to 3.21% as m increased from 1 to 3. Similarly, when $m = 2$, the magnitude of the percentage error of $\hat{a}$ in the case of the lognormally distributed scaling factor (1.83%) was greater than that in the case of the normally distributed scaling factor (0.00%). However, when $m = 3$, the magnitude of percentage error of $\hat{a}$ for the case of the lognormally distributed scaling factor (3.24%) was smaller than that in the case with a normally distributed scaling factor (4.20%). The minimal errors remaining among the bootstrap standard errors in these cases also originated from two types of error.

As the proposed bootstrapping MVR method accurately estimated the parameter standard errors under different simulation settings for each of the selected models, it was considered robust and flexible. Nevertheless, a minimal amount of error, although insignificant, was observed in the bootstrap standard error estimates. These remaining minimal errors stemmed from two sources, namely the two leveraged approximation mechanisms in the bootstrapping MVR method (see Fig. 2). First, the histograms formed by the M sets of bootstrap estimates were only Monte Carlo approximations of the bootstrap distributions for the bootstrap estimates. This type of sampling error could be reduced by increasing the size of M. Second, Ψ_N was only an approximation of the distributions for the unknown population distribution Ψ. By collecting more data and hence increasing the size, N, of the dataset, these errors could be further minimized. In practice, however, when N and M

were sufficiently large, these minimal percentage errors in the bootstrap standard error estimates were insignificant and acceptable.

Most importantly, the results of the comprehensive simulation study demonstrate that compared with conventional models, the proposed bootstrapping MVR method can better accommodate the complex error structure containing both the heteroscedasticity inherently induced by linear data projection and random errors $i.i.d.$ as normal. Therefore, to estimate accurate statistics indicating the significance of the parameter estimates, the proposed bootstrapping MVR method should always be leveraged with generic nonlinear transport models based on linearly projected data.

4.2 Application to MFD

In this subsection, to illustrate the real-world situations necessitating the use of linear data projection and the proposed bootstrapping MVR method, an MFD was estimated for a $1\,\text{km} \times 1\,\text{km}$ network in Choi Hung, Hong Kong. The data were extracted from GPS-equipped taxis and on-road fixed detectors. The total-traffic-to-occupied-taxi ratio was selected as the scaling factor used to estimate network traffic density in a linear data projection.

4.2.1 Chosen model depicting the decreasing area-wide speed-density relationship

To model the decreasing area-wide speed-density relationship, five candidate models were chosen for statistical fitting. The five selected models took the functional forms of the classical Greenshield, Pipes-Munjal, Greenberg, Underwood and Drake models[1], except that density was expressed as the sum of the number of vehicles on all the links within the sampled network (i.e., $k = \sum_{i=1}^{m} k_i$). Of the five candidate models, the function in the form of the Underwood model performed best as measured by the Akaike information criterion (AIC). The AIC, encapsulating the parsimony principle, can be used to select the best fitted model with consideration of both goodness of fit and model complexity (Sakamoto et al., 1987). Thus, the following model, which shared the form of the multivariate exponential decay function, was selected for presentation:

$$u = u_f \exp\left(\frac{-k}{k_o}\right) = u_f \exp\left(\frac{-\sum_{i=1}^{m} k_i}{k_o}\right), \tag{6}$$

where u_f is the free-flow space-mean speed (measured in km/h); k_o is the optimal traffic density per unit of area (measured in veh/km^2); k_i is the number of vehicles on link i in time T; m is the total number of links in the sampled network; k is the traffic density per unit of

[1] Greenshield model corresponds to linear function. Pipes-Munjal model corresponds to polynomial function. Greenberg model corresponds to logarithmic function. Underwood model corresponds to exponential decay function. Drake model corresponds to exponential model of a quadratic function.

area (measured in $\mathrm{veh/km^2}$) in time T, which can be expressed as the sum of the number of vehicles on all the links within the sampled network; and u is the space-mean speed (measured in $\mathrm{km/h}$) over time T .

4.2.2 Databases

To model the macroscopic space-mean speed and traffic density relationship, traffic data for the sampled 1 km × 1 km network in Hong Kong were estimated and extracted from the *Annual Traffic Census 2010* (ATC) (Transport Department, 2010) and a 2010 taxi GPS database, which store one-year traffic data from stationary sources and mobile sources, respectively.

The ATC report comprised comprehensive traffic data collected from more than 1,500 counting stations covering approximately 90% of Hong Kong's trafficable area (Tong et al., 2003; Lam et al., 2003). The average annual daily traffic (AADT) across each of the counting stations associated with the selected network was extracted to constitute the sampled scaling factors, which were used to estimate traffic density.

The taxi GPS database provided detailed travel information on approximately 480 GPS-equipped taxis in 2010. These taxis sent their real-time locations, expressed in terms of World Geodetic System 1984 (International Terrestrial Reference Frame 96) data in decimal degrees, timestamps, instantaneous speed, travel direction and occupancy, to the traffic center every 30 seconds. These taxi data covered Hong Kong's entire transportation network.

4.2.3 Data constitution and the necessity of linear data projection

Space-mean speed u and traffic density k over time T were essential to the MFD estimation. These data were obtained from the one-year ATC report and the taxi GPS database described in the previous subsection. As the behavior of the occupied taxis resembled that of normal traffic, their speed and density data were retrieved from the database to constitute space-mean speed and traffic density. Fig. 4 illustrates the patterns of normalized occupied taxi flow and normalized traffic flow at a few stations (defined as core stations in the ATC report) at which temporal traffic counts were available in the network under study. Although both normalized flows varied through the day, their patterns were remarkably similar, which suggested that the assumption was reasonably valid.

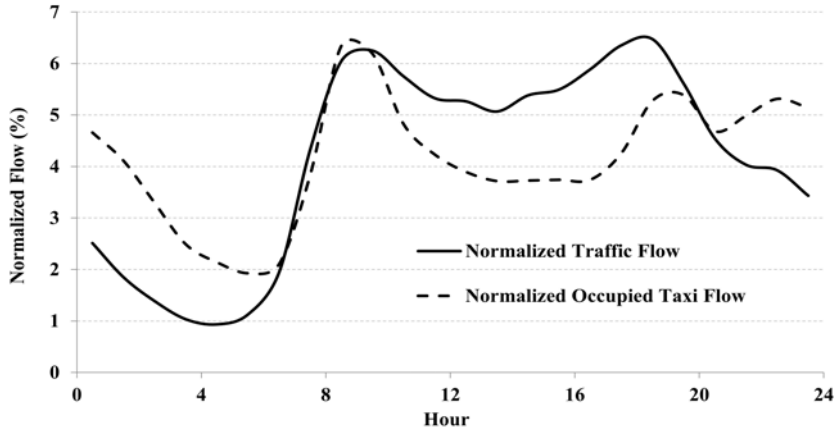


Figure. 4. Normalized occupied taxi flow and normalized total traffic flow at several locations in the network under study

The choice of the sampling (or aggregation) period T reflected a trade-off between data quality and traffic state resolution. A greater T may have increased the number of vehicles in the samples and improved their reliability, but it would also have undesirably averaged out the inhomogeneous traffic states across T. We chose a ten-minute sampling period for T because this appeared to strike a suitable balance between the two factors. Each pair of space-mean speed u and traffic density k were thus average values representing the traffic state of the network in a ten-minute sampling period. As each GPS-equipped taxi reported its travel information to the traffic center twice per minute, we defined a unit time slice, Δt, as 30 seconds. Therefore, a unit sampling period T consisted of 20 time slices. To alleviate the problem of significant errors arising from insufficient occupied taxi coverage, only samples with observed occupied taxis in each of the 20 time slices were included in the case study to filter out and discard unreliable samples and ensure the credibility of the sampled data.

The network traffic density, k, was the total number of vehicles on all of the links in the sampled network. However, as on-road fixed detectors had only been installed on a subset of links in the network, which is usually the case worldwide, it was impossible to directly measure network traffic density. However, using the taxi GPS database, the number of occupied taxis on any link was readily obtainable. Alternatively, the network traffic density k can be expressed as the sum of the linear combinations of f_i and x_i as follows:

$$k = \sum_{i=1}^{m} k_i = \sum_{i=1}^{m} f_i x_i, \tag{7}$$

where f_i is the scaling factor of link i, defined as the total-traffic-to-occupied-taxi ratio on that link and assumed to follow a certain distribution with mean $\bar{f}$ and standard deviation σ_f; and x_i is the average number of occupied taxis on link i in a unit sampling period T, which

was adjusted in accordance with the ratio between the normalized traffic and occupied taxi flows to account for temporal effects (Wong and Wong, 2015, 2016c). The data on x_i could be readily retrieved from taxi GPS data; however, the individual value of each f_i was unknown.

Due to the geographical proximity of the links in the sampled network, the scaling factors of all of the links were assumed to follow a distribution subject to a certain spatial heterogeneity. Such spatial heterogeneity may arise from a number of factors. Meng et al. (2017b) quantified spatial heterogeneity using various measures, such as the land-use purposes of different lots. Fig. 5 is a schematic sketch of the Choi Hung network, illustrating the approximately uniformly distributed ATC stations across the network. The network consisted of 51 links and 17 ATC stations. As the ATC stations were equipped with on-road fixed detectors, the AADT across these stations could be extracted from the ATC report, and the links associated with the ATC stations were selected as the sampling sites for the scaling factors. By dividing the AADT of an ATC station by the average annual daily occupied taxi flow across the same station, each sampled scaling factor was obtained. The mean of the scaling factor distribution, the most probable observed total-traffic-to-occupied-taxi ratio in the network, was estimated by the average value of the sampled scaling factors. The spatial heterogeneity of the traffic composition ratio was quantified by the scaling factor variance. The scaling factor mean was 198.6, indicating that each observed occupied taxi in Choi Hung represented approximately 199 vehicles in the network. The scaling factor variance was 139.9. As not all of the scaling factors were known, a direct evaluation of network traffic density based on Eq. (7) was not possible. Instead, linear data projection (i.e., $\bar{k} = \sum_{i=1}^{m} \bar{f} x_i$) was necessary for unbiased traffic density estimation.

Fig. 5. Schematic sketch of the network in Choi Hung, showing the approximately uniform distribution of ATC stations

As the occupied taxis interacted with the surrounding vehicles with which they traveled in the sampled network, all of the vehicles were assumed to travel at similar speeds. The space-mean speed of all traffic during sampling period T was obtained by dividing the total distance traveled by the total time spent. Assuming that the traffic composition ratio of each link remained approximately uniform during a short (ten-minute) sampling period, T, the space-mean speed was given by Eq. (8):

$$u = \frac{\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij} \Delta t f_i}{\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} \Delta t f_i}, \tag{8}$$

where N_{ti} is the total number of occupied taxis on link i, $\forall i \in [1, m]$, u_{ij} is the speed of the jth occupied taxi on link i, $\forall j \in [1, N_{ti}]$ and Δt is 30 seconds. Similarly, as individual values could not be identified for the scaling factors, direct evaluation was impossible. However, the second-order approximation of the expectation of the space-mean speed, which was dependent on the coefficient of variation of the scaling factor, is given by Eq. (9). A detailed derivation is provided in Appendix A.

$$E(u) \cong \frac{\sum_{i=1}^{m}\sum_{j=1}^{N_{ti}} u_{ij}}{\sum_{i=1}^{m} N_{ti}} + \left(\frac{\sigma_f}{\bar{f}}\right)^2 \frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{N_{ti}} u_{ij}\right)\left(\sum_{i=1}^{m} {N_{ti}}^2\right) - \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{N_{ti}} u_{ij}\right) N_{ti}\right]\left(\sum_{i=1}^{m} N_{ti}\right)}{\left(\sum_{i=1}^{m} N_{ti}\right)^3} \quad (9)$$

The first term on the right-hand side of Eq. (9) (i.e., the arithmetic mean of the speeds of the occupied taxis) is the first-order approximation of the space-mean speed of all traffic and an unbiased estimator of the space-mean speed of the occupied taxis. Although this approximation was used in Geroliminis and Daganzo's (2008) study of MFDs, its adoption carries the assumption of a homogeneous traffic composition ratio across a network, which is rarely true in reality. The second-order approximation used in this study improves the accuracy of the space-mean speed estimate because the second term shown in Eq. (9), which can be viewed as a correction to/adjustment of the first term, further incorporates information about the spatial heterogeneity of the traffic composition ratio and the spatial distribution of the occupied taxis.

4.2.4 Standard error estimations based on the proposed bootstrapping MVR method

The multivariate exponential decay function was a nonlinear function of the scaling factor, and the scaling factor was subject to spatial heterogeneity. Therefore, direct model estimation using linearly projected data would have resulted in biased parameters (Wong and Wong, 2015). As the histogram of the natural logarithm of the sampled scaling factors was approximately symmetrically bell-shaped, a lognormal distribution was a candidate distribution for the scaling factor. A Kolmogorov-Smirnov goodness of fit test was performed to test the null hypothesis that the sampled scaling factors were consistent with a specified lognormal distribution. As the resulting statistic was smaller than the critical value of the test at the 0.05 level of significance, insufficient evidence was obtained to reject the null hypothesis. Therefore, the scaling factor was assumed to be lognormally distributed and the EMVR method with a quartic approximation of the expectation function was leveraged to ensure that the parameter estimates were sufficiently accurate (Wong and Wong, 2018). Fig. 6 depicts the scatter plot and the best-fitted multivariate exponential decay function based on the EMVR method.

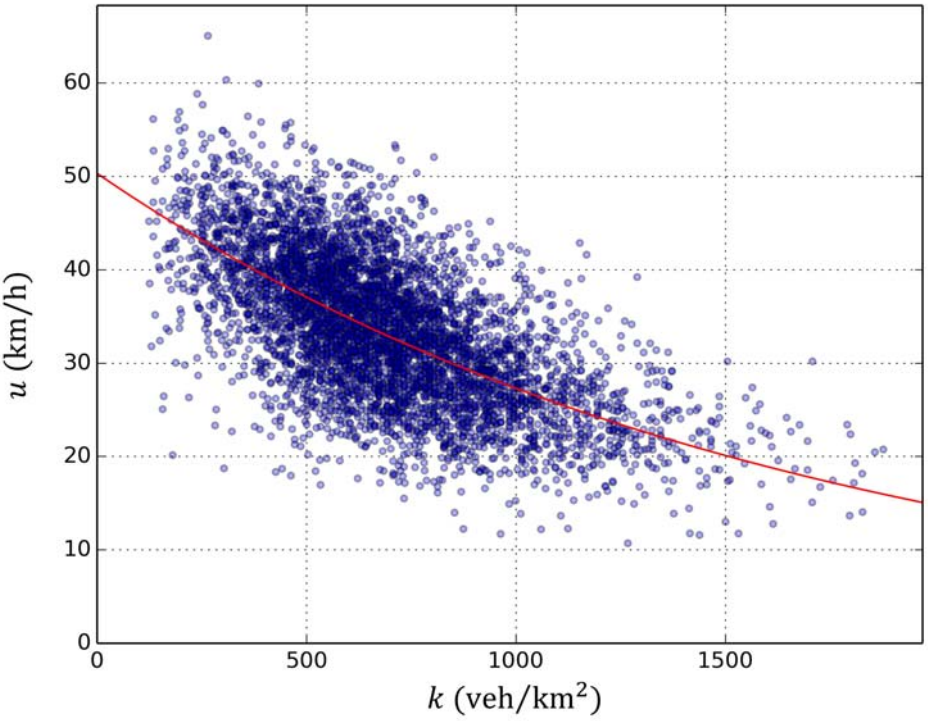


Fig. 6. Scatter plot of speed against density and the best fitted multivariate exponential decay function using the EMVR method for a 1 km × 1 km network in Choi Hung, Hong Kong

However, the heteroscedasticity inherently introduced by linear data projection results in biased standard error estimates (Wong and Wong, 2016c). Thus, without accurate and suitable statistics indicating the significance of the estimates, the validity of application of the model remained unclear. In addition, as the multivariate exponential decay function is a nonlinear transport model, necessitating nonlinear regression, the proposed bootstrapping MVR method was adopted to account for the complex error structure and accurately estimate the standard errors. The number of repetitions M was set at $10{,}000$. In addition, $5{,}254$ two-dimensional vectors (i.e., $N = 5{,}254$) were observed in the dataset. Thus, in each repetition, $5{,}254$ vectors were randomly sampled with replacements from the observed dataset to form a bootstrap sample. Based on the bootstrap sample, the EMVR method was used to obtain a set of bootstrap estimates (i.e., $\hat{u}_f^*$ and $\hat{k}_o^*$). The bootstrap standard errors of the estimated parameters were given by the standard deviations of the M sets of bootstrap estimates. The parameter estimates and bootstrap standard errors, as estimated by the EMVR method and the proposed bootstrapping MVR method, respectively, are presented in Table 5.

Table 5
Parameter estimates and bootstrap standard errors of the multivariate exponential decay function using the EMVR method and the proposed bootstrapping MVR method, respectively

	Model parameter	
	$\hat{u}_f$ (km/h)	$\hat{k}_o$ (veh/km^2)
Calibrated parameter	50.3	1634.9
$\widehat{RSE}$ (% $error$)	0.3538 (-2.31%)	29.6215 ($+2.33\%$)
Bootstrap standard error	0.3622	28.9478
t-statistic	138.95	56.48
p-value	0.00	0.00

Using conventional nonlinear regression in standard statistical package, the estimation of standard errors was based on the assumption that the distorted composite error terms are $i.i.d.$ as normal. The row for $\widehat{RSE}$ in Table 5 above shows the reported standard errors estimated using conventional regression procedures. The biased standard errors were above or below the true values. The comprehensive simulation studies presented in Section 4.1 demonstrated that the proposed bootstrapping MVR method accounted for both heteroscedasticity and $i.i.d.$ random errors in standard error estimations when nonlinear transport models were estimated using linearly projected data. Therefore, the bootstrap standard errors were the best estimates. Compared with these best estimates, the reported standard errors of $\hat{u}_f$ and $\hat{k}_o$ were underestimated by 2.31% and overestimated by 2.33%, respectively, in this case study. The t-statistics and p-values of the parameter estimates were evaluated according to the parameter and bootstrap standard errors. All of the parameter estimates were statistically significant at the 0.001 level, as their p-values were much smaller than 0.001. Although the magnitude of scaling factor variability differs between cases, the proposed method guarantees more accurate standard error estimates; therefore, it should always be adopted as a precaution to minimize uninformed and unnecessary risks when performing statistical tests.

5. Conclusions

In today's era of big data, linear data projection is an important and powerful data scaling method that can be used to infer population traffic and transportation quantities from a set of samples. The method provides an unbiased estimator of unobservable traffic data, using data that can be collected via state of the art technology without incurring extra capital or operational costs, due to sensor deployment. As more and more connected vehicle (or probe vehicle) data become available, linear data projection is increasingly necessary to scale up observable traffic data for this estimation process. However, recent studies have shown that direct model estimations using linearly projected data may cause biased parameters, and definitely result in biased standard errors. Therefore, it is vitally important to develop methods that can remove such biases to realize the full strength of this powerful data scaling method. Methods of removing biases embedded in both the parameters and the standard

errors of linear regression models have been proposed (Wong and Wong, 2015; 2016c). For nonlinear transport models necessitating nonlinear regression, methods of unbiased standard error estimation remain unexplored, although MVR and EMVR methods have recently been established for unbiased parameter estimation (Wong and Wong, 2018). In the absence of accurate standard errors indicating the statistical significance of parameter estimates, the validity of application of an estimated model remains unclear. It is vital to explore methods of filling this research gap.

This study was conducted to develop a method of unbiased standard error estimation for generic nonlinear transport models using linearly projected data. This work is new and important. Typically, standard error estimation is much more challenging than parameter estimation, as it demands attention to the variability and dispersion of parameter estimates. The study first presented the practical difficulties associated with the use of the traditional standard error estimation approach to deal with generic nonlinear transport models estimated using linearly projected data. Simple close-form formulae for unbiased standard error estimation may not exist for this type of model, or they may be too complicated to derive. As an alternative, a bootstrapping MVR method incorporating either the MVR or the EMVR method was proposed. The proposed method accounts for the complex error structure comprising both the heteroscedasticity induced by linear data projection and random errors $i.i.d$ as normal.

To evaluate the capability and robustness of the proposed method in recovering true parameter standard errors, a series of simulation cases with different settings was conducted based on the GMP function with a relaxed exponent and the multivariate exponential decay function, which are the most commonly used nonlinear transport models depicting traffic flow relationships. The simulation results showed that the proposed method substantially outperformed the conventional estimation method, and that it accurately estimated the standard errors in all of the cases considered. Thus, the proposed method was considered robust and flexible. The remaining minimal percentage errors in the standard error estimates arose from two of the approximation mechanisms used in the bootstrapping MVR method. In practice, these minimal errors could be further minimized by increasing the size N of the dataset and the number of repetitions M in the resampling procedures. Most importantly, the simulation results suggested that the proposed bootstrapping MVR method should always be used for accurate standard error estimations with generic nonlinear transport models using linearly projected data.

To illustrate real-world situations necessitating the use of linear data projection and the proposed bootstrapping MVR method in standard error estimations, an MFD was estimated for a sampled network in Choi Hung, Hong Kong using real-world GPS and counting station data. All of the parameter estimates of the MFD were statistically significant at the 0.001 level, because the p-values of the parameter estimates were much smaller than 0.001.

Although the proposed method offers a remarkable improvement in the accuracy of standard error estimations, the method, in some cases, is relatively computationally

demanding and time-consuming. Typically, analytical approaches are much less computer-intensive. Thus, although traditional standard error estimation has some practical difficulties, this approach is still worth exploring. Some interesting challenges for future research include deriving analytical formulae for estimating standard errors in situations of interest, and making comparisons between standard error estimates based on different approaches in terms of their level of bias, efficiency, consistency, robustness or computational cost.

Acknowledgments

The work presented in this study was supported by a Research Postgraduate Studentship and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 17208614). The second author was also supported by the Francis S Y Bong Professorship in Engineering. We would like to express our sincere gratitude to Concord Pacific Satellite Technologies Limited and Motion Power Media Limited for providing taxi GPS data, and to the Transport Department of the HKSAR government for providing traffic flow data from the ATC.

Appendix A. Second-order approximation of the expected space-mean speed of all traffic, capturing the spatial heterogeneity of the traffic composition ratio

Here, we provide the second-order approximation of the space-mean speed of all traffic during the unit sampling period T. The space-mean speed of all traffic is the total distance traveled divided by the total time spent. It can be expressed in terms of the individual speeds of occupied taxis, the number of occupied taxis and the traffic composition ratio on each link $i \in [1, m]$, as shown in Eq. (A1):

$$u = \frac{\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij} \Delta t f_i}{\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} \Delta t f_i}. \quad \text{(A1)}$$

As $\sum_{j=1}^{N_{ti}} 1 = N_{ti}$ and Δt is independent of i and j, Eq. (A1) can be simplified to Eq. (A2):

$$u = \frac{\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij} f_i}{\sum_{i=1}^{m} N_{ti} f_i}. \quad \text{(A2)}$$

As the value of each f_i was unknown, it was impossible to directly evaluate the space-mean speed using Eq. (A1) or Eq. (A2). However, the first and second moments of space-

mean speed could be estimated using the sampled scaling factors. Therefore, the second-order approximation of the expected space-mean speed, which further captured the spatial variability of the traffic composition ratio across links, was used instead (Wong and Wong, 2015; 2018).

The first and second partial derivatives of $u\ w.r.t.\ f_k, \forall k \in [1, m]$ are given by Eq. (A3) and Eq. (A4), respectively.

$$\frac{\partial u}{\partial f_k} = \frac{(\sum_{i=1}^{m} N_{ti} f_i) \sum_{j=1}^{N_{tk}} u_{kj} - \left(\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij} f_i\right) N_{tk}}{(\sum_{i=1}^{m} N_{ti} f_i)^2} \tag{A3}$$

$$\frac{\partial^2 u}{\partial {f_k}^2} = \frac{2N_{tk}\left[\left(\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij} f_i\right) N_{tk} - (\sum_{i=1}^{m} N_{ti} f_i)\left(\sum_{j=1}^{N_{tk}} u_{kj}\right)\right]}{(\sum_{i=1}^{m} N_{ti} f_i)^3} \tag{A4}$$

Thus, the second-order approximation of the expected space-mean speed of all traffic is given by Eq. (A5), as follows:

$$\begin{aligned} & E(u) \\ & \cong \frac{\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij}}{\sum_{i=1}^{m} N_{ti}} \\ & + \left(\frac{\sigma_f}{\bar{f}}\right)^2 \frac{\left(\sum_{i=1}^{m} \sum_{j=1}^{N_{ti}} u_{ij}\right)\left(\sum_{i=1}^{m} {N_{ti}}^2\right) - \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{N_{ti}} u_{ij}\right) N_{ti}\right](\sum_{i=1}^{m} N_{ti}).}{(\sum_{i=1}^{m} N_{ti})^3} \end{aligned} \tag{A5}$$

In particular, if $N_{t1} = N_{t2} = \cdots = N_{ti} = \cdots = N_{tm}$, implying that occupied taxis are uniformly distributed across space, or $\sigma_f = 0$, implying a spatially homogeneous traffic composition ratio, then the second term on the right-hand side of Eq. (A5) equals zero. In such cases, the expected value of the space-mean speed of all traffic in the long term is equivalent to the expected space-mean speed of the occupied taxis, which is simply the arithmetic mean of all of the speeds of the occupied taxis.

References

Aboudolas, K., Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. *Transportation Research Part B: Methodological* 55, 265-281.

Ambühl, L., Menendez, M., 2016. Data fusion algorithm for macroscopic fundamental diagram estimation. *Transportation Research Part C: Emerging Technologies* 71, 184-197.

Ban, X.J., Li, Y., Skabardonis, A., Margulici, J.D., 2010. Performance evaluation of travel-time estimation methods for real-time traffic applications. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 14 (2), 54-67.

Bertini, R.L., Tantiyanugulchai, S., 2004. Transit buses as traffic probes: Use of geolocation data for empirical evaluation. *Transportation Research Record: Journal of the Transportation Research Board* 1870, 35-45.

Bolla, R., Davoli, F., 2000. Road traffic estimation from location tracking data in the mobile cellular network. In: *Proceedings of the Wireless Communications and Networking Conference, IEEE,* Chicago, USA, volume 3, 1107-1112.

Caceres, N., Romero, L.M., Benitez, F.G., del Castillo, J.M., 2012. Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems* 13 (3), 1430-1441.

Chernick, M.R., LaBudde, R.A., 2011. *An introduction to bootstrap methods with applications to R.* Hoboken, New Jersey, John Wiley & Sons.

Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological* 41 (1), 49-62.

Du, J., Rakha, H., Gayah, V.V., 2016. Deriving macroscopic fundamental diagrams from probe data: Issues and proposed solutions. *Transportation Research Part C: Emerging Technologie* 66, 136-149.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7, 1-26.

Fisher, N.I., Hall, P., 1991. Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference* 27 (2), 157-169.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological* 42 (9), 759-770.

Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. *IEEE Transactions on Intelligent Transportation System* 14 (1), 348-359.

Geroliminis, N., Levinson, D.M., 2009. Cordon pricing consistent with the physics of overcrowding. In Lam, W.H.K., Wong, S.C., Lo, H.K. (eds). *Transportation and Traffic Theory 2009: Golden Jubilee*. New York, Springer US, 219-240.

Herrera, J.C., Bayen, A.M., 2010. Incorporation of Lagrangian measurements in freeway traffic state estimation. *Transportation Research Part B: Methodological* 44 (4), 460-481.

Herrera, J.C., Work, D.B., Herring, R., Ban, X., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies* 18 (4), 568-583.

Ho, H.W., Wong, S.C., 2007. Housing allocation problem in a continuum transportation system. *Transportmetrica* 3 (1)**,** 21-39.

Kwong, K., Kavaler, R., Rajagopal, R., Varaiya, P., 2009. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies* 17 (6), 586-606.

Lam, W.H.K., Hung, W.T., Lo, H.K., Lo, H.P., Tong, C.O., Wong, S.C., Yang, H., 2003. Advancement of the annual traffic census in Hong Kong. *Proceedings of the Institution of Civil Engineers – Transport* 156 (2), 103-115.

Liang, Y., Reyes, M.L., Lee, J.D., 2007. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems* 8 (2), 340-350.

Meng, F., Wong, W., Wong, S.C., Pei, X., Li, Y.C., Huang, H., 2017a. Gas dynamic analogous exposure approach to interaction intensity in multiple-vehicle crash: Case study of crashes involving taxis. *Analytic Methods in Accident Research* 16, 90-103.

Meng, F., Wong, S.C., Wong, W., Li, Y.C., 2017b. Estimation of scaling factors for traffic counts based on stationary and mobile sources of data. *International Journal of Intelligent Transportation Systems Research* 15 (3), 180-191.

Miwa, T., Ishiguro, Y., Yamamoto, T., Morikawa, T., 2013. Allocation planning for probe taxi devices based on information reliability. *Transportation Research Part C: Emerging Technologies* 34, 55-69.

Moore II, J.E., Cho, S., Basu, A., Mezger, D.B., 2001. *Use of Los Angeles Freeway Service Patrol Vehicles as Probe Vehicles*. California PATH Research Report UCB-ITS-PRR-2001-5, California PATH Program, Institute of Transportation Studies, University of California, Berkeley, California.

Nagle, A.S., Gayah, V.V., 2014. Accuracy of networkwide traffic states estimated from mobile probe data. *Transportation Research Record: Journal of the Transportation Research Board* 2421, 1-11.

Schwarzenegger, A., Bonner, D.E., Iwasaki, R.H., Copp, R., 2009. *2008 State Highway Congestion Monitoring Program (HICOMP)*, Annual Data Compilation, Caltrans, Sacramento, California.

Shao, J., Tu, D., 1995. *The jackknife and bootstrap*. New York, Springer-Verlag.

Tong, C.O., Hung, W.T., Lam, W.H.K., Lo, H.K., Lo, H.P., Wong, S.C., Yang, H., 2003. A new survey methodology for the annual traffic census in Hong Kong. *Traffic Engineering and Control* 44 (6), 214-218.

Transport Department, 2010. *The Annual Traffic Census 2010*. Traffic and Transport Survey Division, Transport Department, Government of the Hong Kong SAR, Hong Kong.

Transportation Research Board, 2000. *Highway Capacity Manual*. National Research Council, Washington, D.C.

Wong, R.C.P., Szeto, W.Y., Wong, S.C., Yang H., 2014. Modelling multi-period customer-searching behaviour of taxi drivers. *Transportmetrica B: Transport Dynamics* 2 (1), 40-59.

Wong, W., Wong, S.C., 2015. Systematic bias in transport model calibration arising from the variability of linear data projection. *Transportation Research Part B: Methodological* 75, 1-18.

Wong, W., Wong S.C., 2016a. Evaluation of the impact of traffic incidents using GPS data. *Proceedings of the Institution of Civil Engineers – Transport* 169 (3), 148-162.

Wong, W., Wong S.C., 2016b. Network topological effects on the macroscopic Bureau of Public Roads function. *Transportmetrica A: Transport Science* 12 (3), 272-296.

Wong, W., Wong, S.C., 2016c. Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection. *Transportation Research Part B: Methodological* 88, 72-92.

Wong, W., Wong, S.C., 2018. Unbiased estimation methods of nonlinear transport models based on linearly projected data. *Transportation Science*. In press.

Wright, J., Dahlgren, J., 2001. *Using Vehicles Equipped with Toll Tags as Probes for Providing Travel Times*. California PATH Working Paper UCB-ITS-PWP-2001-13, California PATH Program, Institute of Transportation Studies, University of California, Berkeley, California.

Ygnace, J.L., Drane, C., 2001. Cellular telecommunication and transportation convergence: a case study of a research conducted in California and in France on cellular positioning techniques and transportation issues. In: *2001 IEEE Intelligent Transportation Systems Conference Proceedings*, Oakland, California, 16-22.

Yin, J., Wong, S.C., Sze, N.N., Ho, H.W., 2013. A continuum model for housing allocation and transportation emission problems in a polycentric city. *International Journal of Sustainable Transportation* 7 (4), 275-298.

Zhao, Y., 2000. Mobile phone location determination and its impact on intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems* 1 (1), 55-64.

Zheng, N., Waraich, R.A., Axhausen, K.W., Geroliminis, N., (2012) A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. *Transportation Research Part A: Policy and Practice* 46 (8), 1291-1303.