



<b>Title</b>	<b>Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection</b>
<b>Author(s)</b>	<b>Wong, W; Wong, SC</b>
<b>Citation</b>	<b>Transportation Research Part B: Methodological, 2016, v. 88, p. 72-92</b>
<b>Issued Date</b>	<b>2016</b>
<b>URL</b>	<b><a href="http://hdl.handle.net/10722/224885">http://hdl.handle.net/10722/224885</a></b>
<b>Rights</b>	<b>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</b>

# Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection

Wai Wong and S.C. Wong

*Department of Civil Engineering, The University of Hong Kong, Pokfulam, Hong Kong*

---

## Abstract

Reliable transport models calibrated from accurate traffic data are crucial for predicating transportation system performance and ensuring better traffic planning. However, due to the impracticability of collecting data from an entire population, methods of data inference such as the linear data projection are commonly adopted. A recent study has shown that systematic bias may be embedded in the parameters calibrated due to linearly projected data that do not account for scaling factor variability. Adjustment factors for reducing such biases in the calibrated parameters have been proposed for a generalized multivariate polynomial model. However, the effects of linear data projection on the dispersion of and confidence in the adjusted parameters have not been explored. Without appropriate statistics examining the statistical significance of the adjusted model, their validity in applications remains unknown and dubious. This study reveals that heteroscedasticity is inherently introduced by data projection with a varying scaling factor. Parameter standard errors that are estimated by linearly projected data without any appropriate treatments for non-homoscedasticity are definitely biased, and possibly above or below their true values. To ensure valid statistical tests of significance and prevent exposure to uninformed and unnecessary risk in applications, a generic analytical distribution-free (ADF) method and an equivalent scaling factor (ESF) method are proposed to adjust the parameter standard errors for a generalized multivariate polynomial model, based on the reported residual sum of squares. The ESF method transforms a transport model into a linear function of the scaling factor before calibration, which provides an alternative solution path for achieving unbiased parameter estimations. Simulation results demonstrate the robustness of the ESF method compared with the ADF method at high model nonlinearity. Case studies are conducted to illustrate the applicability of the ESF method for the parameter standard error estimations of six Macroscopic Bureau of Public Road functions, which are calibrated using real-world global positioning system data obtained from Hong Kong.

*Keywords:* Heteroscedasticity; Standard error estimation; Linear data projection; Macroscopic Bureau of Public Roads; GPS

---

## 1. Introduction

Accurate and reliable model calibration is vitally important for transportation studies, because it helps to establish a better understanding of the interaction between transportation

infrastructure, vehicles and road users. Hence, such calibration allows appropriate urban traffic planning, traffic management and control measures to be implemented. Moreover, the irreversible development patterns that are caused by infrastructures and the critical roles they play in promoting economic growth (Carlsson et al., 2013) necessitate that reliable transport models be estimated at the planning stage to prevent the misuse of public budgets and resources.

The observational nature of most transportation studies makes them different from typical science and engineering subjects, in which systems can be investigated using the desired values of independent variables under well-controlled experimental setups. In contrast, accurate traffic data must be collected from huge transportation systems for reliable transport model calibrations. The advent of various high-tech devices has significantly improved the accuracy and efficiency of traffic data collection over the past several decades. However, various factors limiting the applications of these detectors and sensors still make it impractical to collect traffic data from the entire population. On-road fixed detectors, such as inductive loop detectors, can collect data at an acceptable level of accuracy with minimal effort, but their high installation and maintenance costs hinder the ubiquitous deployment of detectors all over the network (Herrera and Bayen, 2010; Herrera et al., 2010). Thus, the coverage of such detectors is normally limited to a subset of links within a network (Caceres et al., 2012). A vehicle re-identification system can measure the travel time of a vehicle across a link by matching the vehicle signature as the vehicle passes through the two ends of a link outfitted with sensors (Kwong et al., 2009). The radio frequency identification transponders (Wright and Dahlgren, 2001; Ban et al., 2010), license plate recognition systems (Herrera et al., 2010), wireless magnetic sensors (Kwong et al., 2009) and other unique tags are readily available utilities for such systems. However, the risk of privacy issues and the high installation and implementation costs are major obstacles to deploying most such schemes over the entire arterial network. The cellular systems introduced a decade ago (Bolla and Davoli, 2000; Ygnace and Drane, 2001; Zhao, 2000) offer a resolution to the cost and coverage problems (Herrera et al., 2010). Nevertheless, their application is prohibited or discouraged in many countries, because the use of cell phones while driving disrupts the drivers' attention (Liang et al., 2007). Global positioning systems (GPSs) are another promising means of collecting traffic data from almost the entire network at a relatively low cost (Miwa et al., 2013). However, GPS data collected from vehicle fleets (e.g., FedEx, UPS, or taxis) (Moore et al., 2001; Bertini and Tantiyanugulchai, 2004; Schwarzenegger et al., 2009; Wong et al., 2014) involve biases due to the fleets' specific operational or travel patterns. In addition, the extra capital and installation costs of GPS trackers, along with the potential for privacy issues, impede the application of GPS tracking systems on a global scale.

Despite technological advancements, traffic data collection from huge transportation systems using specific devices is still limited by various factors. Thus, different mathematical techniques such as data scaling, filtering and sampling are commonly used to estimate traffic data and overcome these difficulties. Linear data projection is a prevalent data scaling scheme that infers population traffic quantities by projecting the observable traffic quantities from a subset of the population, using the mean of a set of sampled scaling factors. The scaling factor used in a linear data projection varies according to each situation. Because transportation

systems are dynamic and non-steady, scaling factors are usually random variables that are subject to variability and thus are assumed to follow distributions. Depending on the sampling approach adopted, the scaling factor variance may require measuring different types of variability, such as spatial or temporal variability.

Linear data projection has been used for traffic data estimation in numerous transportation studies. For example, an hourly total traffic flow across a link that is not outfitted with an on-road fixed detector can be estimated using linear data projection. Assuming that the total traffic flow is observable on a subset of links outfitted with detectors in a network, and that the occupied taxi flow is observable on every link in the network, the total-traffic-to-occupied-taxi ratios that are sampled at the links outfitted with detectors can be chosen as the scaling factors. Given the heterogeneities of the road hierarchy and the land use pattern, the sampled scaling factors can be different from each other. They are assumed to follow a distribution over the network due to geographical proximity. The scaling factors are sampled across the network, and thus their variance measures the spatial variability. As the scaling factor mean is the most probable observed traffic composition ratio across the network, if the sampled scaling factor mean is 100 and the hourly occupied taxi flow on the link of interest is 10 veh/h, then the total hourly traffic on this link can be estimated by the product of these factors; that is, 1000 veh/h.

In accident analysis, exposure expressed in vehicle kilometrage (i.e., the product of annual traffic volume and road length) is usually a typical explanatory variable accounting for the variations in a road's annual crash levels. The corresponding parameter associated with the variable is known as the accident rate. Due to limited resources, detailed traffic data throughout a year are usually collected for only a subset of links, whereas short-term (e.g., a weekday) traffic data are surveyed for other links. In such cases, the exposure of a link with only short-term traffic volume can be estimated using linear data projection with the annual-to-short-term-traffic-volume ratios of nearby links with full-year traffic data as the chosen scaling factors. The scaling factor variance measures spatial variability. The product of a road's short-term traffic volume, the scaling factor mean and the road length provide a good estimate of that road's exposure.

Another example uses linear data projection as the equivalent traffic flow estimation expressed in passenger car units (PCUs). Unlike the usual assumption, a PCU is not necessarily static (Chandra et al., 1995) due to the varying traffic composition across time. For a road installed with an on-road fixed detector recording vehicle counts 24 hours a day, the hourly PCU value is not always known because surveyors can only be sent on-site to identify the vehicle types for several hours a day, certain days a year, according to a strategic sampling plan influenced by budget constraints. The product of an hourly traffic count and the sampled PCU means can estimate the hourly equivalent traffic flow. The PCU variance measures the temporal variability. Moreover, other traffic quantities such as trip completion rates, vehicular accumulations and space-mean speed can also be inferred using linear data projections (Geroliminis and Daganzo, 2008).

Macroscopic transport models have gained much attention in recent decades due to their potential applications in area-wide traffic management and control (Aboudolas and Geroliminis, 2013; Geroliminis et al., 2013), and initial land use planning (Ho and Wong, 2007; Yin et al., 2013). However, calibrations of these models require traffic flow data across networks via different links that may be unobservable. Linear data projection's simplicity has made it a popular choice in many studies and real-world traffic data estimation situations. Geroliminis and Daganzo (2008) leveraged linear data projection to infer the total traffic flows of sites without detectors from occupied GPS taxi flows, using the traffic composition ratio as the scaling factor in their study of a macroscopic fundamental diagram. Similarly, Wong and Wong (2015, 2016) estimated the total hourly traffic flow entering 1 km × 1 km sampled networks in Hong Kong using linear data projections and calibrated macroscopic cost flow functions.

Models that are directly calibrated on linearly projected data that do not account for scaling factor variability may result in systematically biased model parameters. The study conducted by Wong and Wong (2015) focused on parameter estimations, and proved that such a bias is introduced if the model is a non-linear function of the scaling factor and the scaling factor is subject to variability, regardless of the model form or the distribution of the scaling factor. In that study, the generalized multivariate polynomial model was examined, and adjustment factors that could efficiently and effectively reduce the systematic bias of the calibrated parameters were derived. Nevertheless, such parameter adjustments only guarantee unbiased (or slightly biased) parameters in the long run, whereas the effects of linear data projection on the dispersion and variability of the adjusted parameters, measured by their standard errors, have not been explored. Without appropriate statistics examining the parameter statistical significance of the adjusted models, their validity for applications remains unknown and doubtful. This study focuses on standard error estimations, and reveals that heteroscedasticity is inherently introduced when a data projection scheme is adopted, regardless of the model form. Parameter standard errors that are estimated by linearly projected data without any appropriate treatments for heteroscedasticity are definitely biased, and possibly above or below their true values. In other words, a true null hypothesis may be wrongly rejected, leading to a type I error, or a false null hypothesis may fail to be rejected, leading to a type II error in the statistical tests of significance for the adjusted parameters. Thus, unbiased standard error estimations are vitally important to ensure valid statistical tests of significance and prevent exposure to uninformed and unnecessary risk when the applying the adjusted models.

To demonstrate the existence of such bias in standard error estimations when a data projection scheme is adopted, a simple numerical example of the calibration of the following model is presented.

$$y = \beta_0 + \beta_1 X + \varepsilon = \beta_0 + \beta_1(fx) + \varepsilon$$

where  $x$  is the observable independent variable;  $f$  is the scaling factor of  $x$  (which is assumed to follow a distribution);  $X = fx$  is the projected value;  $\varepsilon$  is the random error (which is

assumed to be normally distributed);  $y$  is the observable dependent variable, and  $\beta_0$  and  $\beta_1$  are the model parameters.

Ten thousand data points of  $x$ , serving as the observable independent variables, are sampled from a negative exponential distribution with a mean of 0.2. As scaling factors are usually positive, a lognormal distribution with  $\bar{f} = 1$  and  $\sigma_f = 0.2$  is chosen to sample the corresponding scaling factors for the 10,000 samples of  $x$ .  $\bar{f}$  and  $\sigma_f$  are the mean and the standard deviation of the scaling factor  $f$ , respectively. Ten thousand  $\varepsilon$ , serving as the random errors of the 10,000 observations, are sampled from a normal distribution with zero mean and 0.1 standard deviation (i.e.,  $\sigma_\varepsilon = 0.1$ ). Assuming that  $\beta_0 = 3$  and  $\beta_1 = 1$ , the corresponding 10,000 points of  $y$  (which serve as the observed data for the dependent variable) can be calculated based on the assumed values of the parameters and the sampled  $x$ ,  $f$  and  $\varepsilon$ . Suppose that the values of all individual  $f$  and  $\varepsilon$  are no longer available, and  $X$  can only be estimated via a linear data projection based on the mean value of  $f$  (i.e.,  $\bar{f}x$ ), which is a common real-world occurrence. Regression analysis conducted between  $y$  and the linearly projected  $X$  (using any standard statistical package) results in a set of calibrated parameters,  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , and a set of reported standard errors of the calibrated parameters,  $RSE(\hat{\beta}_0)$  and  $RSE(\hat{\beta}_1)$ . By using the same set of  $x$ , and repeating the above steps 10,000 times, 10,000 sets of calibrated parameters and reported standard errors are obtained. The standard deviations of the calibrated parameters,  $SD(\hat{\beta}_0)$  and  $SD(\hat{\beta}_1)$ , are the unbiased estimators of the standard errors of the calibrated parameters. Table 1 shows the mean and standard deviations of the calibrated parameters, the mean of the reported standard errors and their corresponding percentage errors.

**Table 1**

The mean and standard deviation of the calibrated parameters and the mean of the reported standard errors for the simple linear function

Mean		Standard deviation		Mean	
$\hat{\beta}_0$ (%error)	$\hat{\beta}_1$ (%error)	$\hat{\beta}_0$	$\hat{\beta}_1$	$RSE(\hat{\beta}_0)$ (%error)	$RSE(\hat{\beta}_1)$ (%error)
3.0000 (0.00%)	1.0000 (0.00%)	0.00174	0.00854	0.00163 (-6.17%)	0.00579 (-32.19%)

Simulation results show that the calibrated parameters are unbiased, which is consistent with Wong and Wong (2015), because the model is a linear function of the scaling factor. However, the deviations between the means of the reported standard errors and the standard deviations of  $\hat{\beta}_0$  (-6.17%) and of  $\hat{\beta}_1$  (-32.19%) reveal that the estimated standard errors are biased. Statistical tests based on these biased standard errors may lead to biased inferences. One important implication demonstrated by this simple numerical example is that even if parameter adjustments are not required for a linear model, the standard errors are biased and an appropriate standard estimation method is still necessary.

To remove such bias, it is necessary to gain a better understanding about the origin of the bias and of the potential for more efficient methods of standard error estimation. Standard error estimations are more difficult to make than parameter estimations, because a higher-order moment for an estimated parameter measurement of its dispersion is considered. This study

aims to fill these knowledge gaps. The study shows that the heteroscedasticity inherently introduced by data scaling schemes is the source of the bias. Although adjustment factors for reducing systematic biases in the calibrated parameters of generalized multivariate polynomial models have already been derived by Wong and Wong (2015), the standard error estimations for such model forms have remained unexplored. A generic analytical distribution-free (ADF) method is proposed to estimate the standard errors of the calibrated/adjusted parameters of the generalized multivariate polynomial model. However, the estimation accuracy of ADF method deteriorates at high model nonlinearity, due to the flexibility of unspecified scaling factor distributions. Therefore, an equivalent scaling factor (ESF) method, which generally provides more robust standard error estimations at high nonlinearity, is also proposed. As the ESF method transforms the model to a linear function of an ESF before calibration, the calibrated model parameters are unbiased. In other words, adjustments for the calibrated parameter are not required, and this method provides an alternative solution path for unbiased parameter estimations when a linear data projection is adopted. The Bureau of Public Roads (BPR) function (as adopted in the *Highway Capacity Manual* (Transportation Research Board, 2000)) is a member of the generalized multivariate polynomial family. The BPR function, which depicts the travel time against traffic flow relationship in a link, is commonly leveraged in many European countries and the United States (Dowling et al., 1998; Lum et al., 1998). This function plays a crucial role in static user equilibrium analysis (García-Ródenas and Verastegui-Rayo, 2013). Macroscopic Bureau of Public Roads (MBPR) functions, which model the relationship between travel time and traffic flow over a network, are calibrated in this study based on real-world GPS data, using the proposed ESF method to estimate the model parameters and their corresponding standard errors.

The remainder of the study is structured as follows. Section 2 proves that heteroscedasticity is inherently introduced when a data projection scheme is adopted, and that direct calibrations based on the linearly projected data lead to biased standard error estimations. A generic formula estimating the parameter standard errors of the generalized multivariate polynomial model is derived in Section 3. Simulation results based on the ADF method are presented. The subsequent section proposes the ESF method. Simulation results reveal that the ESF method generally performs better at high model nonlinearity than the ADF method. In Section 5, the proposed ESF method is applied in case studies of MBPR function calibrations and standard error estimations using real-life GPS data acquired in Hong Kong. The final section provides a conclusion on the findings of the study and discusses the potential directions for future research.

## **2. Biased standard error estimations due to heteroscedasticity**

This section proves that heteroscedasticity is inherently introduced in a data projection scheme when the scaling factor adopted is subject to variability. This finding violates the usual assumption of homoscedasticity in model calibrations, which has led to biased standard error estimations.

## 2.1 Heteroscedasticity arising from data projection

This subsection unveils the inherent heteroscedasticity arising from data projections. Consider  $y = G(z) + \varepsilon$ , where  $G(z)$  is a function of any form;  $\varepsilon$  is an independent random error, assumed to be normally distributed with zero mean and constant variance  $\sigma_\varepsilon^2$ ;  $z$  is constituted by the sum of the products of a set of scaling factors and a set of observable independent variables (i.e.,  $z = \sum_{i=1}^m f_i x_i$ ;  $x_i$  is the observable independent variable  $\forall i \in [1, m]$ );  $f_i$  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 used to construct the quantity  $z$ .

In most cases, the data collection for  $z$  is impossible or impractically expensive and labor-intensive compared to data collection for  $x_i$ . In practice, however, the observable variable  $x_i$  can generally be collected in relatively cheaper ways. The scaling factor  $f_i$  of each individual  $x_i$  is assumed to follow a distribution. Theoretically, this factor can be assumed to follow any distribution, but the properties of the chosen distribution should be in line with the conditions of the given situation. For instance, if the scaling factor is always non-negative, with a lower relative frequency at high values, then a lognormal distribution can be chosen as the assumed candidate distribution. The first and second moments of the distribution can be estimated by using another set of scaling factors sampled from independent sources under similar conditions. In many situations, the impracticability of direct data collection for  $z$  necessitates the use of a data scaling scheme. However, the data projection method fundamentally leads to non-homoscedasticity, due to the variability of the scaling factors as stated in Proposition 1.

**Proposition 1.** In addition to the random error  $\varepsilon$ , with a normally assumed constant variance  $\sigma_\varepsilon^2$ , heteroscedasticity is inherently introduced, as  $z$  is constituted via a data projection scheme in which  $z$  is expressed as a linear combination of the scaling factors  $f_i$  and the observable independent variables  $x_i$  (i.e.,  $z = \sum_{i=1}^m f_i x_i$ ) regardless of the distribution of the scaling factor or the form of the model  $G(z)$ , as long as the scaling factors are subject to variability.

**Proof.** Approximate  $G(z)$  by a Taylor series expansion with the center at  $f_i = \bar{f}$ ,  $\forall i \in \mathbb{N}^+$ .

$$y = G(\bar{\mathbf{f}}) + \sum_{i=1}^m \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} (f_i - \bar{f}) + \dots + \varepsilon \quad (1)$$

Grouping all of the terms with random variables,  $\varepsilon'$  is defined as a composite random term, which is expressed in the following form:

$$\varepsilon' = \sum_{i=1}^m \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} (f_i - \bar{f}) + \dots + \varepsilon \quad (2)$$

Substituting Eq. (2) into Eq. (1), Eq. (3) gives the following results:



$$y = G(\bar{\mathbf{f}}) + \varepsilon' \quad (3)$$

Taking the variance on both sides of Eq. (3) for any given  $\mathbf{x}$ ,

$$\text{Var}(y|\mathbf{x}) = \text{Var}[G(\bar{\mathbf{f}}) + \varepsilon']$$

As  $G(\bar{\mathbf{f}})$  is a constant,  $\text{Var}[G(\bar{\mathbf{f}})] = 0$  and  $\text{Cov}[G(\bar{\mathbf{f}}), \varepsilon'] = 0$ . Therefore,

$$\text{Var}(y|\mathbf{x}) = \text{Var}(\varepsilon') \quad (4)$$

Substituting Eq. (2) into Eq. (4) and ignoring higher order terms,

$$\text{Var}(y|\mathbf{x}) \cong \text{Var} \left[ \sum_{i=1}^m \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} (f_i - \bar{f}) + \varepsilon \right]$$

The independence between  $\varepsilon$  and  $f_i, \forall i \in [1, m]$  leads to  $\text{Cov} \left[ \sum_{i=1}^m \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} (f_i - \bar{f}), \varepsilon \right] = 0$ . It follows that

$$\text{Var}(y|\mathbf{x}) \cong \text{Var} \left[ \sum_{i=1}^m \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} (f_i - \bar{f}) \right] + \text{Var}(\varepsilon)$$

$$\begin{aligned} \text{Var}(y|\mathbf{x}) \cong & \sum_{i=1}^m \left[ \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} \right]^2 \text{Var}(f_i - \bar{f}) \\ & + 2 \sum_{1 \leq i < j \leq m} \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} \frac{\partial G(\bar{\mathbf{f}})}{\partial f_j} \text{Cov}[(f_i - \bar{f}), (f_j - \bar{f})] + \sigma_\varepsilon^2 \end{aligned}$$

Assuming that  $f_i, \forall i \in \mathbb{N}^+$  are independent of each other,  $\text{Cov}[(f_i - \bar{f}), (f_j - \bar{f})] = 0, \forall i, j \in \mathbb{N}^+ \setminus [i = j]$ . Thus,

$$\text{Var}(y|\mathbf{x}) \cong \sum_{i=1}^m \left[ \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} \right]^2 \{E[(f_i - \bar{f})^2] - [E(f_i - \bar{f})]^2\} + \sigma_\varepsilon^2$$

where  $E(f_i - \bar{f}) = 0$  and  $E[(f_i - \bar{f})^2] = \sigma_f^2$ . Thus,

$$\text{Var}(y|\mathbf{x}) \cong \sigma_f^2 \sum_{i=1}^m \left[ \frac{\partial G(\bar{\mathbf{f}})}{\partial f_i} \right]^2 + \sigma_\varepsilon^2 \quad \square \quad (5)$$

Eq. (5) is an approximated scedastic function of  $y$ , which expresses the variance of  $y$  in terms of both  $\sigma_\varepsilon^2$  and  $\sigma_f^2$ . In addition to the random error  $\varepsilon$  with a constant variance  $\sigma_\varepsilon^2$ , it is apparent that the variability of  $y$  for a given  $\mathbf{x}$  is also dependent the scaling factor variance  $\sigma_f^2$ . As  $\sum_{i=1}^m \left[ \frac{\partial G(\bar{f})}{\partial f_i} \right]^2$  is a function of  $\mathbf{x}$ , the effect of  $\sigma_f^2$  on the variance of  $y$  changes with  $\mathbf{x}$ , and hence heteroscedasticity exists. The variability composition of  $y$  is complex, because it is a mixture of both the constant variance  $\sigma_\varepsilon^2$  and the varying  $\sigma_f^2$ . Thus, as long as the scaling factor is subject to variability, heteroscedasticity is introduced when a data projection scheme is leveraged. It should be stressed that Proposition 1 is generic, as  $G(z)$  can be a function of any model form.

## 2.2 Violation of the homoscedasticity assumption

Model calibrations normally require an assumption of homoscedasticity that is a constant variance of the dependent variable along with the independent variable being modeled. However, heteroscedasticity introduced by a data projection scheme undesirably violates this assumption, and this is largely ignored in the field during model calibrations. Hence, biased standard error estimates may be above or below the true standard errors. These biased statistics may wrongly reject the true null hypotheses, leading to a type I error, or wrongly fail to reject the false null hypotheses, leading to a type II error in the statistical tests of significance for the adjusted parameters. Direct applications of the adjusted models without valid statistical tests of significance result in exposure to an uninformed and unnecessary risk of making wrong decisions.

Although the scedastic function of the composite error term can be easily obtained via Eq. (5), the conventional approaches to the problem of heteroscedasticity (generalized least squares (GLS) and weighted least squares (WLS) methods) are rather indirect to address heteroscedasticity arising from linear data projection in this case. Given the additive relationship between the term with  $\sigma_f^2$  and  $\sigma_\varepsilon^2$ , the formulated weighting factors cannot be independent of the unknown  $\sigma_\varepsilon^2$ . Estimation of the unknown  $\sigma_\varepsilon^2$  requires most of the formulae derived in the ADF method, which is proposed in Section 3. However, once the ADF method is applied, the problem can be addressed directly. Most importantly, when a linear data projection is leveraged, the expectation of the composite error term,  $E(\varepsilon')$ , is nonzero when the model to be calibrated is a nonlinear function (Wong and Wong, 2015). In such cases, the usual assumption of zero expected value of the error term is violated and biased parameters and standard errors are the result. Thus, these conventional approaches are not only indirect to this problem, but may also be incapable of solving it. Thus, to ensure reliable and valid statistical tests, efficient and appropriate standard error estimation procedures must be derived that can estimate the unbiased statistics as functions of these projected data, involving both the constant  $\sigma_\varepsilon^2$  and the varying  $\sigma_f^2$ .

### 3. ADF method

A generalized multivariate polynomial model is examined in this study. It is useful for approximations of many functional functions and can be easily reduced to a polynomial of the required number of terms with desired orders in practice. In this section, the generic ADF method, which estimates standard errors based on the reported residual sum of squares, is derived based on the generalized multivariate polynomial model. The ADF method is flexible, because no assumption concerning the model's scaling factor distribution is necessary. Comprehensive simulations using the proposed ADF method are conducted to demonstrate its effectiveness in making estimations. Although the generalized multivariate polynomial model is the only model chosen for the demonstrations in this study, it should be noted that the concepts and methodologies proposed in this and the subsequent sections can be extrapolated to other cases.

#### 3.1 Formulation

Let  $P(z)$  be a function in polynomial form with  $n + 1$  terms and define  $p = n + 1$ , which is the total number of sensitivity parameters of the polynomial. Consider the following model,  $y = P(z) + \varepsilon$ :

$$y = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_n z^n + \varepsilon$$

where  $z = \sum_{i=1}^m f_i x_i$ ;  $\beta_0, \beta_1, \beta_2, \dots, \beta_{n-1}$  and  $\beta_n$  are the model parameters. Thus,

$$y = \beta_0 + \beta_1 \left( \sum_{i=1}^m f_i x_i \right) + \beta_2 \left( \sum_{i=1}^m f_i x_i \right)^2 + \cdots + \beta_n \left( \sum_{i=1}^m f_i x_i \right)^n + \varepsilon \quad (6a)$$

Eq. (6a) can also be written in matrix form, as shown by Eq. (6b), in terms of observations:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6b)$$

where  $\mathbf{y}^T = [y_1, y_2, \dots, y_N]$ , which is a  $(1 \times N)$  transposed column vector of observations of the dependent variable;  $N$  is the total number of observations;  $\mathbf{X}$  is an  $(N \times p)$  matrix of observations of the independent variables;  $X_{rk} = (\sum_{i=1}^m f_{ir} x_{ir})^k$ , which is the  $r$ th row and the  $k$ th column entry of  $\mathbf{X}$ ,  $\forall r \in [1, N]$  and  $\forall k \in [0, n]$ ;  $\boldsymbol{\beta}^T = [\beta_0, \beta_1, \dots, \beta_n]$ , which is a  $(1 \times p)$  transposed column vector of model parameters;  $\boldsymbol{\varepsilon}^T = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N]$ , which is a  $(1 \times N)$  transposed column vector of random errors, and  $\varepsilon_r$  is assumed to be independent identically distributed (i.i.d.) as normal with zero mean and constant variance,  $\sigma_\varepsilon^2$ ,  $\forall r \in [1, N]$ .

However, the individual value of each scaling factor is usually not available, which is a common real-world occurrence. We assume that the first and second moments of the scaling factor distribution can be estimated by using another set of scaling factors sampled from independent sources under similar conditions. Thus, a linear data projection is usually adopted,

in which each individual scaling factor is replaced with its mean value. Using Eq. (3) and expanding  $P(\bar{\mathbf{f}})$ , Eq. (7a) gives results as follows.

$$y = \beta_0 + \beta_1 \left( \sum_{i=1}^m \bar{f} x_i \right) + \beta_2 \left( \sum_{i=1}^m \bar{f} x_i \right)^2 + \dots + \beta_n \left( \sum_{i=1}^m \bar{f} x_i \right)^n + \varepsilon' \quad (7a)$$

Eq. (7a) can also be written in matrix form, as shown Eq. (7b), in terms of observations:

$$\mathbf{y} = \bar{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}' \quad (7b)$$

where  $\bar{\mathbf{X}}$  is an  $(N \times p)$  matrix of linearly projected independent variables;  $\bar{X}_{rk} = \left( \sum_{i=1}^m \bar{f} x_{ir} \right)^k$ , which is the  $r$ th row and the  $k$ th column entry of  $\bar{\mathbf{X}}$ ,  $\forall r \in [1, N]$  and  $\forall k \in [0, n]$ , and  $\boldsymbol{\varepsilon}'^T = [\varepsilon'_1, \varepsilon'_2, \dots, \varepsilon'_N]$ , which is a  $(1 \times N)$  transposed column vector of composite random components.

The model parameters are usually calibrated using Eq. (8), based on the linearly projected data with an (invalid) assumption that the composite random components  $\boldsymbol{\varepsilon}'$  are the error terms i.i.d. as normal, with zero mean and a constant variance of  $\sigma_{\boldsymbol{\varepsilon}'^2}$ . Making parameter estimations using Eq. (8) is equivalent to minimizing the residual sum of squares. The minimized residual sum of squares reported by a standard statistical package is obtained based on the composite random components  $\boldsymbol{\varepsilon}'$ . We define this result as the reported residual sum of squares, which is denoted by  $RSSE$ .

$$\hat{\boldsymbol{\beta}} = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^T \mathbf{y} \quad (8)$$

where  $\hat{\boldsymbol{\beta}}^T = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n]$ , which is a  $(1 \times p)$  transposed column vector of calibrated model parameters. The fitted regression model is given by

$$\hat{\mathbf{y}} = \bar{\mathbf{X}} \hat{\boldsymbol{\beta}} \quad (9)$$

where  $\hat{\mathbf{y}}^T = [\hat{y}_0, \hat{y}_1, \dots, \hat{y}_N]$ , which is a  $(1 \times N)$  transposed column vector of predicted response variables, based on the linearly projected independent variables and the calibrated parameters. The corresponding hat matrix,  $\mathbf{H}$ , which is known as the influence matrix or projection matrix, is defined as follows.

$$\mathbf{H} = \bar{\mathbf{X}}(\bar{\mathbf{X}}^T \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^T \quad (10)$$

Thus, Eq. (9) can be alternatively expressed as shown in Eq. (11).

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad (11)$$

However, Wong and Wong (2015) have recently proved that the calibrated parameters based on linearly projected data are systematically biased, i.e.,  $E(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$ , regardless of the distribution of the scaling factor or the form of the model to be calibrated, as long as the data involve a non-linear function of the scaling factor and the scaling factor is subject to variability. In other words, the calibrated parameters are unbiased only if the model is a linear function of the scaling factor. In particular, for the case of a generalized multivariate polynomial model, these authors have shown that  $E(\mathbf{X}) \cong \bar{\mathbf{X}}\bar{\mathbf{F}}$ , and they have proposed to reduce the systematic bias by incorporating the scaling factor variance  $\sigma_f^2$  into the calibrated parameters  $\hat{\boldsymbol{\beta}}$  based on Eq. (12).

$$\bar{\boldsymbol{\beta}} = \bar{\mathbf{F}}^{-1}\hat{\boldsymbol{\beta}} \quad (12)$$

where  $\bar{\boldsymbol{\beta}}^T = [\bar{\beta}_0, \bar{\beta}_1, \dots, \bar{\beta}_n]$ , which is a  $(1 \times p)$  transposed column vector of the adjusted model parameters;  $\bar{\mathbf{F}}$  is an  $(p \times p)$  diagonal matrix, and  $\bar{F}_k = 1 + \left[ \frac{k(k-1)}{2} \right] \left[ \left( \frac{\sigma_f}{\bar{f}} \right)^2 \right] \left[ \frac{\sum_{i=1}^m x_i^2}{(\sum_{i=1}^m x_i)^2} \right]$ , which is the  $k$ th row and the  $k$ th column diagonal element of  $\bar{\mathbf{F}}$ ,  $\forall k \in [0, n]$ . Each calibrated parameter can be corrected independently. The simulation conducted by Wong and Wong (2015) demonstrated the effectiveness and efficiency of the proposed methodology, and it is indeed true that  $E(\bar{\boldsymbol{\beta}}) \cong \boldsymbol{\beta}$ . It should be noted that  $E(\bar{\boldsymbol{\beta}})$  is an approximation to  $\boldsymbol{\beta}$  only when the model to be calibrated involves at least one term with an order greater than two and the scaling factor moment of the corresponding order is nonzero (e.g., a model involves a cubic term and the skewness of the scaling factor is nonzero.). Otherwise,  $E(\bar{\boldsymbol{\beta}})$  is exactly equal to  $\boldsymbol{\beta}$ , because the adjustment factor proposed by Wong and Wong (2015) was derived based on Taylor series expanded up to the second-order term.

Nevertheless, according to Proposition 1, it is apparent that the response variable  $y$  and the composite random components  $\varepsilon'$  are not randomly distributed with either a constant variance or with pure heteroscedasticity, but instead are distributed with a combination of both patterns. The composite random components  $\varepsilon'$  should thus be regarded as distorted error components. The invalid assumption of i.i.d. normally distributed  $\varepsilon'$ , with zero mean and a constant variance  $\sigma_{\varepsilon'}^2$ , definitely leads to incorrect standard error estimations. Accurate and reliable standard error estimations are important, because accurate estimates are necessary to perform statistical tests and to provide evidence on whether the corrected model parameters are statistically significant. For the case of a generalized multivariate polynomial model involving a linear data projection, the standard errors of the calibrated/adjusted sensitivity parameters can be estimated analytically, based on the reported residual sum of squares,  $RSSE$ , which is obtained via the usual model calibration procedures as stated in Proposition 2.

**Proposition 2.** The standard error of the  $k$ th calibrated/adjusted sensitivity parameter of a generalized multivariate polynomial model,  $\bar{\beta}_k$ , can be estimated by the square root of the  $k$ th row and the  $k$ th column diagonal element of the estimated covariance matrix of  $\bar{\boldsymbol{\beta}}$ ,  $\forall k \in [0, n]$ . The estimated covariance matrix of  $\bar{\boldsymbol{\beta}}$  is a  $(p \times p)$  symmetric matrix given by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) \cong (\overline{\mathbf{F}}^{-1})(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}^T \left[ \widehat{Var}(\overline{\mathbf{X}}\widehat{\boldsymbol{\beta}}) + \frac{RSSE - \text{tr}[(\mathbf{I}_N - \mathbf{H})\widehat{Var}(\overline{\mathbf{X}}\widehat{\boldsymbol{\beta}})]}{N - p} \mathbf{I}_N \right] \overline{\mathbf{X}}(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} (\overline{\mathbf{F}}^{-1})^T$$

where  $\mathbf{I}_N$  is an  $(N \times N)$  identity matrix.

**Proof.** Taking variance on both sides of Eq. (12),

$$Var(\overline{\boldsymbol{\beta}}) = Var(\overline{\mathbf{F}}^{-1}\widehat{\boldsymbol{\beta}}) \quad (13)$$

Substituting Eq. (8) into Eq. (13),

$$Var(\overline{\boldsymbol{\beta}}) = Var[\overline{\mathbf{F}}^{-1}(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}^T \mathbf{y}]$$

As both  $\overline{\mathbf{F}}$  and  $\overline{\mathbf{X}}$  are constants, and  $\mathbf{y}$  is the only quantity possessing randomness,

$$Var(\overline{\boldsymbol{\beta}}) = (\overline{\mathbf{F}}^{-1})(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}^T Var(\mathbf{y}) \overline{\mathbf{X}}(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} (\overline{\mathbf{F}}^{-1})^T$$

The independence between  $\varepsilon$  and  $f_i, \forall i \in [1, m]$  leads to

$$Var(\overline{\boldsymbol{\beta}}) = (\overline{\mathbf{F}}^{-1})(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}^T [Var(\mathbf{X}\boldsymbol{\beta}) + \sigma_\varepsilon^2 \mathbf{I}_N] \overline{\mathbf{X}}(\overline{\mathbf{X}}^T \overline{\mathbf{X}})^{-1} (\overline{\mathbf{F}}^{-1})^T \quad (14)$$

Replacing  $Var(\mathbf{X}\boldsymbol{\beta})$  and  $\sigma_\varepsilon^2$  by their unbiased estimators, which are denoted by  $\widehat{Var}(\overline{\mathbf{X}}\widehat{\boldsymbol{\beta}})$  and  $\widehat{\sigma}_\varepsilon^2$ , respectively, gives the estimated covariance matrix of  $\overline{\boldsymbol{\beta}}$ ,  $\widehat{Var}(\overline{\boldsymbol{\beta}})$ .

To obtain  $\widehat{Var}(\overline{\boldsymbol{\beta}})$ , consider  $Var(\mathbf{X}\boldsymbol{\beta})$ , which is an  $(N \times N)$  diagonal matrix. Where  $\forall r \in [1, N], \forall k, l \in [0, n]$  and  $k < l$ , the  $r$ th row and  $r$ th column diagonal element of  $Var(\mathbf{X}\boldsymbol{\beta})$  is given by

$$\begin{aligned} Var \left[ \sum_{k=0}^n \beta_k \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \right] &= \sum_{k=0}^n \beta_k^2 Var \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \right] \\ &+ 2 \sum_{0 \leq k < l \leq n} \beta_k \beta_l Cov \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k, \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^l \right] \end{aligned} \quad (15)$$

Consider the first expression on the right hand side of Eq. (15),  $\forall k \in [0, n]$ .

$$\beta_k^2 \text{Var} \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \right] = \beta_k^2 \left\{ E \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^{2k} \right] - E \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \right]^2 \right\}$$

Using the result from Appendix A,

$$\begin{aligned} & \beta_k^2 \text{Var} \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \right] \\ & \cong \beta_k^2 \left\{ \begin{aligned} & \left[ 1 + \frac{2k \cdot (2k-1)}{2} \left( \frac{\sigma_f}{\bar{f}} \right)^2 \frac{\sum_{i=1}^m x_{ir}^2}{(\sum_{i=1}^m x_{ir})^2} \right] \\ & - \left[ 1 + \frac{k \cdot (k-1)}{2} \left( \frac{\sigma_f}{\bar{f}} \right)^2 \frac{\sum_{i=1}^m x_{ir}^2}{(\sum_{i=1}^m x_{ir})^2} \right]^2 \end{aligned} \right\} \left( \sum_{i=1}^m \bar{f} x_{ir} \right)^{2k} \end{aligned} \quad (16)$$

Consider the second expression on the right hand side of Eq. (15),  $\forall k, l \in [0, n]$  and  $k < l$ .

$$\begin{aligned} & \beta_k \beta_l \text{Cov} \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k, \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^l \right] \\ & = \beta_k \beta_l \left\{ E \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^l \right] - E \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k \right] E \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^l \right] \right\} \end{aligned}$$

Again, using the result from Appendix A,

$$\begin{aligned} & \beta_k \beta_l \text{Cov} \left[ \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^k, \left( \sum_{i=1}^m f_{ir} x_{ir} \right)^l \right] \\ & \cong \beta_k \beta_l \left\{ \begin{aligned} & \left[ 1 + \frac{(k+l) \cdot (k+l-1)}{2} \left( \frac{\sigma_f}{\bar{f}} \right)^2 \frac{\sum_{i=1}^m x_{ir}^2}{(\sum_{i=1}^m x_{ir})^2} \right] - \\ & \left[ 1 + \frac{k \cdot (k-1)}{2} \left( \frac{\sigma_f}{\bar{f}} \right)^2 \frac{\sum_{i=1}^m x_{ir}^2}{(\sum_{i=1}^m x_{ir})^2} \right] \left[ 1 + \frac{l \cdot (l-1)}{2} \left( \frac{\sigma_f}{\bar{f}} \right)^2 \frac{\sum_{i=1}^m x_{ir}^2}{(\sum_{i=1}^m x_{ir})^2} \right] \end{aligned} \right\} \left( \sum_{i=1}^m \bar{f} x_{ir} \right)^{k+l} \end{aligned} \quad (17)$$

$\widehat{\text{Var}}(\mathbf{X}\boldsymbol{\beta})$  can thus be obtained via substitutions of the observed data and the corrected model parameters into each diagonal element of  $\text{Var}(\mathbf{X}\boldsymbol{\beta})$ .

For the unbiased estimator of  $\sigma_\varepsilon^2$ , consider the reported residual sum of squares,  $RSSE$ , as obtained from the standard regression procedures.

$$RSEE = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) \quad (18)$$

Substituting Eq. (11) into Eq. (18),

$$RSEE = \mathbf{y}^T (\mathbf{I}_N - \mathbf{H}) \mathbf{y} \quad (19)$$

Taking expectation on both sides of Eq. (19) and using the trace-variance formula,

$$E[RSEE] = \text{tr}[(\mathbf{I}_N - \mathbf{H})\text{Var}(\mathbf{y})] + E[\mathbf{y}]^T (\mathbf{I}_N - \mathbf{H}) E[\mathbf{y}]$$

As  $\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta}) + \sigma_\varepsilon^2 \mathbf{I}_N$  and  $E[\mathbf{y}] \cong \bar{\mathbf{X}}\bar{\mathbf{F}}\bar{\boldsymbol{\beta}}$ ,

$$E[RSEE] \cong \text{tr}[(\mathbf{I}_N - \mathbf{H})\text{Var}(\mathbf{X}\boldsymbol{\beta})] + \sigma_\varepsilon^2 \text{tr}[\mathbf{I}_N - \mathbf{H}] \\ + (\bar{\mathbf{X}}\bar{\mathbf{F}}\bar{\boldsymbol{\beta}})^T (\mathbf{I}_N - \mathbf{H}) (\bar{\mathbf{X}}\bar{\mathbf{F}}\bar{\boldsymbol{\beta}})$$

Also, as  $\text{tr}[\mathbf{I}_N - \mathbf{H}] = N - p$  and  $(\bar{\mathbf{X}}\bar{\mathbf{F}}\bar{\boldsymbol{\beta}})^T (\mathbf{I}_N - \mathbf{H}) (\bar{\mathbf{X}}\bar{\mathbf{F}}\bar{\boldsymbol{\beta}}) = 0$ ,

$$E[RSEE] \cong \text{tr}[(\mathbf{I}_N - \mathbf{H})\text{Var}(\mathbf{X}\boldsymbol{\beta})] + \sigma_\varepsilon^2 (N - p) \quad (20)$$

By substituting  $\text{Var}(\widehat{\mathbf{X}\boldsymbol{\beta}})$  into Eq. (20),  $\hat{\sigma}_\varepsilon^2$  can be obtained after rearrangement.

$$\hat{\sigma}_\varepsilon^2 \cong \frac{RSEE - \text{tr}[(\mathbf{I}_N - \mathbf{H})\text{Var}(\widehat{\mathbf{X}\boldsymbol{\beta}})]}{N - p} \quad (21)$$

After substituting  $\text{Var}(\widehat{\mathbf{X}\boldsymbol{\beta}})$  and  $\hat{\sigma}_\varepsilon^2$  into Eq. (14), the estimated covariance matrix of  $\bar{\boldsymbol{\beta}}$ ,  $\text{Var}(\bar{\boldsymbol{\beta}})$  results, as shown in Eq. (22).

$$\text{Var}(\bar{\boldsymbol{\beta}}) \cong (\bar{\mathbf{F}}^{-1})(\bar{\mathbf{X}}^T \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^T \left[ \text{Var}(\widehat{\mathbf{X}\boldsymbol{\beta}}) \right. \\ \left. + \frac{RSEE - \text{tr}[(\mathbf{I}_N - \mathbf{H})\text{Var}(\widehat{\mathbf{X}\boldsymbol{\beta}})]}{N - p} \mathbf{I}_N \right] \bar{\mathbf{X}} (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^{-1} (\bar{\mathbf{F}}^{-1})^T \quad \square \quad (22)$$

The square roots of the diagonal elements of the estimated covariance matrix indicate the standard errors of the calibrated/adjusted sensitivity parameters. This proposed methodology is flexible, because no assumption is made concerning the form of the scaling factor distribution, and only the first and second moments of the scaling factor are required. Thus, the method is analytical and distribution-free. Moreover, the covariance matrix is estimated based on the residual sum of squares (as indicated by any standard statistical package) upon a regression on the linear projection data, as if the scaling factor is subject to zero variability. Thus, the proposed ADF method can be easily integrated with the existing



standard model calibration procedures and with the parameter adjustment procedure proposed by Wong and Wong (2015). The combination of the parameter adjustment procedure and the proposed ADF method is an extension of the classical model calibration, as it further considers the effects of scaling factor variability on both the parameter and the standard error estimations.

### 3.2 Simulations

This subsection presents comprehensive simulations of regressions on the generalized multivariate polynomial function, as shown in Eq. (23), and on standard error estimations using the proposed ADF method.

$$y = \beta_0 + \beta_n z^n + \varepsilon = \beta_0 + \beta_n \left( \sum_{i=1}^m f_i x_i \right)^n + \varepsilon \quad (23)$$

Twelve simulation cases with different combinations in terms of the numbers of linear combinations of the scaling factors and observable independent variables (i.e.,  $m = 1$  and 5), scaling factor distributions (i.e., normal and lognormal) and model nonlinearity (i.e.,  $n = 1, 2$  and 3) were considered for investigating the performance of the ADF method under different situations. Polynomial functions with two terms of orders ranging from zero to three were selected for simulations because they resemble practical functions of transportation analyses, such as the BPR function adopted in the *Highway Capacity Manual* (Transportation Research Board, 2000).

As  $m$  was chosen to be either one or five, two sets of 10,000 observations of  $x$  serving as the observable independent variable had to be generated. The observable independent variable can be any traffic variable following any distribution. We chose a negative exponential distribution with a mean of 0.2 for the purpose of data generation for all of the simulation cases. For  $m = 1$ , each observation was comprised of one  $x$ . In contrast, for  $m = 5$ , five  $x$  were independently sampled from the distribution for each observation. To facilitate comparisons between the performances of the ADF method under different situations, any effects on the results due to the sampling error arising from data generations of the observable independent variables should be avoided. Thus, these two sets of sampled observable independent variables were used for all of our simulation cases.

The scaling factor  $f$  of each  $x$  was sampled from either a normal or lognormal distribution, with  $\bar{f} = 1$  and  $\sigma_f = 0.2$ , according to the simulation case of interest. In addition, 10,000  $\varepsilon$ , which served as the set of random errors for the 10,000 observations, were sampled from a normal distribution with zero mean and 0.1 standard deviation. We set  $\beta_0 = 3$  and  $\beta_n = 1$ . The corresponding 10,000  $y$ , serving as the observed data for the dependent variable, were calculated based on the assumed model parameter values, the sampled  $x$ ,  $f$  and  $\varepsilon$ , and the  $n$  value of that particular simulation case. Assuming that the values of all individual  $f$  and  $\varepsilon$  were no longer available, the observed independent variable could only be estimated by leveraging a linear data projection based on the scaling factor mean (i.e.,  $\sum_{i=1}^m \bar{f} x_i$ ). Regression analysis based on the linearly projected data resulted in the calibrated model

parameters (i.e.,  $\hat{\beta}_0$  and  $\hat{\beta}_n$ ) and the reported standard errors (i.e.,  $RSE(\hat{\beta}_0)$  and  $RSE(\hat{\beta}_n)$ ). The calibrated model parameters of the nonlinear models (i.e.,  $n \neq 1$ ) were corrected by the adjustment factors proposed by Wong and Wong (2015). Using the proposed ADF method, the standard errors of the calibrated/adjusted model parameters (i.e.,  $SE(\hat{\beta}_0)$  and  $SE(\hat{\beta}_n)$ ) were subsequently estimated.

As with the simple simulation presented in Section 1, each simulation case was repeated 10,000 times to obtain the mean and standard deviation of the calibrated/adjusted parameters, the mean of the reported standard errors and the mean of the estimated standard errors based on the ADF method. For each simulation case, all of the 10,000 repeated simulations used the same set of observable independent variables  $x$ , but a different and resampled scaling factor  $f$  and a random error  $\varepsilon$ . Thus, the variability of the calibrated/adjusted model parameter originated purely from the randomness of the scaling factor and the random error. The standard deviation of a calibrated/adjusted parameter for measuring such variability was an unbiased standard error estimator. Table 2 summarizes the results for the 12 simulation cases.

**Table 2**

The means and standard deviations of the calibrated/adjusted parameters and the means of the reported standard errors and estimated standard errors, based on the ADF method

Assumed scaling factor distribution	$m$	$n$	Mean		Standard deviation		Mean		Mean		
			$\hat{\beta}_0$ (%error)	$\bar{\beta}_n$ (%error)	$\hat{\beta}_0$	$\bar{\beta}_n$	$RSE(\hat{\beta}_0)$ (%error)	$RSE(\bar{\beta}_n)$ (%error)	$SE(\hat{\beta}_0)$ (%error)	$SE(\bar{\beta}_n)$ (%error)	
Normal Distribution		1	3.0000 (0.00%)	1.0000 (0.00%)	0.00175	0.00865	0.00163 (-7.00%)	0.00579 (-33.08%)	0.00175 (-0.10%)	0.00862 (-0.38%)	
		1	2	3.0000 (0.00%)	1.0004 (+0.04%)	0.00196	0.02677	0.00138 (-29.67%)	0.00748 (-72.05%)	0.00196 (-0.16%)	0.02650 (-1.03%)
		3	3.0000 (0.00%)	0.9994 (-0.06%)	0.00295	0.06995	0.00155 (-47.41%)	0.00893 (-87.23%)	0.00280 (-5.28%)	0.06456 (-7.71%)	
		1	3.0001 (0.00%)	1.0000 (0.00%)	0.00433	0.00475	0.00394 (-9.12%)	0.00360 (-24.19%)	0.00432 (-0.22%)	0.00475 (-0.06%)	
		5	2	3.0000 (0.00%)	1.0000 (0.00%)	0.01042	0.01050	0.00574 (-44.95%)	0.00352 (-66.51%)	0.01033 (-0.91%)	0.01034 (-1.53%)
		3	2.9997 (-0.01%)	1.0001 (+0.01%)	0.04267	0.02850	0.01309 (-69.33%)	0.00425 (-85.10%)	0.04139 (-3.00%)	0.02742 (-3.76%)	
Lognormal Distribution		1	3.0000 (0.00%)	1.0000 (0.00%)	0.00174	0.00854	0.00163 (-6.17%)	0.00579 (-32.19%)	0.00175 (+0.79%)	0.00862 (+0.94%)	
		1	2	3.0000 (0.00%)	1.0001 (+0.01%)	0.00206	0.02839	0.00141 (-31.60%)	0.00764 (-73.07%)	0.00198 (-3.94%)	0.02653 (-6.53%)
		3	3.0000 (0.00%)	1.0045 (+0.45%)	0.00336	0.08076	0.00166 (-50.45%)	0.00957 (-88.15%)	0.00287 (-14.54%)	0.06496 (-19.56%)	
		1	3.0000 (0.00%)	1.0000 (0.00%)	0.00425	0.00467	0.00394 (-7.39%)	0.00360 (-22.91%)	0.00432 (+1.67%)	0.00475 (+1.64%)	
		5	2	3.0000 (0.00%)	1.0000 (0.00%)	0.01068	0.01067	0.00588 (-44.98%)	0.00360 (-66.24%)	0.01041 (-2.59%)	0.01037 (-2.79%)
		3	2.9995 (-0.02%)	1.0009 (+0.09%)	0.04435	0.02959	0.01378 (-68.94%)	0.00447 (-84.90%)	0.04163 (-6.13%)	0.02748 (-7.15%)	

\*  $RSE(\hat{\beta}_0)$  and  $RSE(\bar{\beta}_n)$  were the reported standard errors of the calibrated parameters, but not of the adjusted parameters.

\* Adjustments on the calibrated parameters of the linear terms were not required.

After adjustments on the calibrated parameters of the nonlinear terms, the calibrated/adjusted model parameters were extremely close to their true values (i.e.,  $\beta_0 = 3$  and  $\beta_n = 1$ ). In terms of magnitude, all of their percentage errors were less than or equal to 0.45%, which demonstrated the significant correction power of the adjustment factors proposed by Wong and Wong (2015). Except for the two cases involving cubic terms and log-normally distributed scaling factors, the expected values of the adjusted sensitivity parameters of the nonlinear terms,  $\bar{\beta}_n$ , should be exactly equal to their corresponding true values,  $\beta_n$ . Any small discrepancy between them purely emanated from the remaining sampling errors of the scaling factor and random error among the 10,000 repeated simulations. For the two cases with cubic terms and log-normally distributed scaling factors, the mean values of  $\bar{\beta}_n$  were close approximations of  $\beta_n$  due to the unrecovered minimal effects of the third moments of the scaling factors on their corresponding parameters. In addition to the sampling errors, the unrecovered minimal effects led to the slightly larger percentage errors of the mean values of  $\bar{\beta}_n$  (i.e., +0.45% for  $m = 1$  and +0.09% for  $m = 5$ ). Nevertheless, such minimal percentage errors of values less than 0.5% are acceptable in practical terms. The standard deviations of the calibrated/adjusted parameters were the unbiased estimators of their standard errors. However, the reported standard errors of the calibrated parameters seriously deviated from those of the unbiased estimators. The percentage errors of the reported standard errors, ranging from around 6% to 88%, generally increased rapidly with  $n$ .

The last two columns shown in Table 2 present the standard errors of the calibrated/adjusted model parameters, estimated according to the proposed ADF method. The simulation results revealed that the magnitudes of the percentage errors for the estimated standard errors were much smaller than those for the reported standard errors. These results suggested that the ADF method performed much better than the classical model calibration procedures (which do not account for the heteroscedasticity arising from a data projection scheme). Moreover, a similar pattern persisted in which the percentage errors generally increased with the exponent  $n$ . For example, in the case of  $m = 1$  and the normally distributed scaling factors, the magnitudes of the percentage errors of  $SE(\hat{\beta}_0)$  and  $SE(\hat{\beta}_n)$  increased from 0.10% and 0.38% (when  $n = 1$ ), to 0.16% and 1.03% (when  $n = 2$ ) and then to 5.28% and 7.71% (when  $n = 3$ ). In spite of these results, when  $n = 1$ , the errors of the estimated standard errors could be well controlled by the ADF method, to around 1%. It should also be noted that the simulation case with  $m = 1$ ,  $n = 1$  and a log-normally distributed scaling factor was exactly the same simulation case as was presented in Section 1. In that case, the estimated standard errors of  $\hat{\beta}_0$  and  $\hat{\beta}_n$  (as indicated by the proposed ADF method) were 0.00175 (+0.79%) and 0.00862 (+0.94%), respectively. A lack of information concerning the higher order moments of the scaling factor in the derivation of the ADF method could be the major cause for the increasing trend of the percentage error with the exponent  $n$ , because these higher order moments (that may be effective in governing the variability of the calibrated/adjusted parameters at higher model nonlinearity) were not captured in the proposed method.

Comparing the simulation cases of normally distributed scaling factors with the cases of log-normally distributed scaling factors, the magnitudes of the percentage errors of the estimated standard error in the former cases were generally lower than those of the latter cases, with the same values of  $m$  and  $n$ . For instance, when  $m = 1$  and  $n = 3$ , the magnitudes of the percentage errors of  $\widehat{SE}(\widehat{\beta}_0)$  and  $\widehat{SE}(\widehat{\beta}_n)$  in the cases of normally distributed scaling factors were 5.28% and 7.71%, respectively, which were lower than those in the cases of log-normally distributed scaling factors. The magnitudes of the percentage errors of  $\widehat{SE}(\widehat{\beta}_0)$  and  $\widehat{SE}(\widehat{\beta}_n)$  in the log-normally distributed scaling factor cases were 14.54% and 19.56%, respectively. These results may provide evidence for the postulate that the higher order moments of the scaling factors may be effective in governing the variability of the calibrated/adjusted parameters at higher nonlinearity, because a lognormal distribution is asymmetric, and its third moment and higher order odd number moments are non-zero.

Moreover, as the numbers of the linear combinations of the scaling factors and the observable independent variables (i.e.,  $m$ ) increased, the errors of the estimated standard errors were reduced for cases with scaling factors sampled from the same distribution and the same values of  $n$ . Taking the case of a log-normally distributed scaling factor and  $n = 3$  as an example, the magnitudes of the percentage errors of  $\widehat{SE}(\widehat{\beta}_0)$  and  $\widehat{SE}(\widehat{\beta}_n)$  were 14.54% and 19.56%, respectively, when  $m = 1$ . However, as  $m$  increased to 5, these magnitudes dropped to 6.13% and 7.15%, respectively. The cancellation of random effects among different scaling factors as  $m$  increased could be a reason for the reductions of the estimated standard errors.

These simulation results demonstrated that the proposed ADF method, which accounts for the heteroscedasticity arising from the variability of a linear data projection, can generally estimate the standard errors of the model parameters at an acceptable level of accuracy. However, although the proposed ADF method was analytical and distribution-free, the estimation accuracy deteriorated when the model nonlinearity (i.e.,  $n$ ) was increased and the higher order moments of the scaling factor governing the calibrated parameter variability were ignored.

### *3.3 Mild violation of i.i.d. random error*

The ADF method formulated in the previous subsection is based on an assumption of i.i.d. random error. However, this might not always be the case in reality. The random error can possibly be heteroscedastic and its variance may vary with  $z$ . Thus, a set of sensitivity analyses for all of the simulation cases were conducted to review if the ADF method can still perform at a satisfactory level under situations with mild violation of i.i.d. random error. Keeping all the other settings the same, the sensitivity analyses were performed at a 10% elasticity of the standard deviation of  $\varepsilon$  with respect to  $z$ . For instance, when  $m = 5$ , random error were sampled from a normal distribution with mean of zero and standard deviation of  $0.1 + 0.01(z - 1)$ , which varies with  $z$ . This represents a 10% elasticity at  $(\sigma_\varepsilon = 0.1, z = 1)$ . Under such settings, heteroscedasticity originating both from linear data projection and random error are introduced.

To test the performance of the ADF method under such a mild assumption violation, each of the 12 simulation cases were repeated for 10,000 times using the conventional regression analysis and the ADF methods as if the random error was i.i.d. as normal. The mean of the reported standard errors (i.e.,  $RSE(\hat{\beta}_0)$  and  $RSE(\hat{\beta}_n)$ ) and mean of the estimated standard errors (i.e.,  $SE(\hat{\beta}_0)$  and  $SE(\hat{\beta}_n)$ ) could then be subsequently computed. All of the simulation cases revealed that the percentage errors of the reported and estimated standard errors obtained from situations with and without the mild assumption violation were comparable. Taking the case with  $m = 5$ ,  $n = 3$  and a normally distributed scaling factor as an example, the magnitudes of the percentage errors of both the reported and estimated standard errors (i.e.,  $RSE(\hat{\beta}_0) = -69.17\%$ ,  $RSE(\hat{\beta}_n) = -84.93\%$ ,  $SE(\hat{\beta}_0) = -2.48\%$  and  $SE(\hat{\beta}_n) = -2.66\%$ ) only decreased slightly compared to the same case shown in Table 2 (i.e.,  $RSE(\hat{\beta}_0) = -69.33\%$ ,  $RSE(\hat{\beta}_n) = -85.10\%$ ,  $SE(\hat{\beta}_0) = -3.00\%$  and  $SE(\hat{\beta}_n) = -3.76\%$ ). For the case with  $m = 5$ ,  $n = 3$  and a log-normally distributed scaling factor, comparing to the same case shown in Table 2 (i.e.,  $RSE(\hat{\beta}_0) = -68.94\%$ ,  $RSE(\hat{\beta}_n) = -84.90\%$ ,  $SE(\hat{\beta}_0) = -6.13\%$  and  $SE(\hat{\beta}_n) = -7.15\%$ ), the magnitudes of the percentage errors of both the reported and estimated standard errors (i.e.,  $RSE(\hat{\beta}_0) = -69.63\%$ ,  $RSE(\hat{\beta}_n) = -85.21\%$ ,  $SE(\hat{\beta}_0) = -8.32\%$  and  $SE(\hat{\beta}_n) = -9.15\%$ ) only increased slightly. Most importantly, results of all these simulations indicated that under situations where the assumption of i.i.d. random error was mildly violated, the use of the proposed ADF method could yield much better and more accurate standard error estimates with significantly smaller percentage errors in magnitudes compared to those reported standard errors by the conventional calibration procedures, because the proposed method at least could address the problem of heteroscedasticity arising from linear data projection.

#### 4. ESF method

To address the accuracy deterioration of the proposed ADF method at high model nonlinearity, the ESF method is introduced in this section. The ESF method is a combination of a simple transformation simulation (which transforms the generalized multivariate polynomial function into a linear function of ESFs) and the ADF method. A series of simulations demonstrated that the proposed ESF method can significantly improve the standard error estimation accuracy at high model nonlinearity.

##### 4.1. Formulation

According to Wong and Wong (2015), if the model to be calibrated is a linear function of the scaling factor, no systematic bias will be embedded in the calibrated parameters, even if the scaling factor is subject to variability. Taking advantage of this property, the ESF method uses a simple transformation simulation to alter the generalized multivariate polynomial function into a linear function of a set of ESFs, based on the additional assumption about the

distribution of the scaling factor. The assumed distribution should follow the properties of the scaling factor. For instance, a lognormal distribution can be the assumed candidate if the scaling factor is always non-negative, with a lower relative frequency at high values. Such a transformation has two major advantages. It bypasses parameter adjustments after a model calibration, because the calibrated parameters are unbiased. Also, it increases the accuracy of standard error estimation, because information about the higher order moments of the scaling factor are captured by the first and second moments of the ESF.

Consider a generalized multivariate polynomial function,  $G(\mathbf{f})$ ,  $\exists G(\mathbf{g})$ , as shown in Eq. (24), such that  $G(\mathbf{g})$  is equivalent to  $G(\mathbf{f})$ .

$$y = \beta_0 + \beta_1 g_1 w + \beta_2 g_2 w^2 + \dots + \beta_n g_n w^n + \varepsilon \quad (24)$$

where  $g_j, \forall j \in [1, n]$  is the ESF of the  $j$ th term that is assumed to follow an unknown distribution, with mean  $\bar{g}_j$  and standard deviation  $\sigma_{g_j}$ , and  $w = \sum_{i=1}^m x_i$ , which is the sum of all observable independent variables,  $x_i, \forall i \in [1, m]$ .

Consider and compare the  $k$ th terms of Eq. (6a) and Eq. (24),

$$\begin{aligned} \beta_k g_k w^k &= \beta_k \left( \sum_{i=1}^m f_i x_i \right)^k \\ g_k &= \left[ \sum_{i=1}^m \left( \frac{x_i}{\sum_j^m x_j} \right) f_i \right]^k \end{aligned} \quad (25)$$

Therefore, the ESF of the  $k$ th term  $g_k$  is the  $k$ th power of the weighted sum of the scaling factor. The weighting factors are constituted by one observation set of the observable independent variable  $x_i, \forall i \in [1, m]$ .

Suppose that there are  $N$  observation sets of  $x_i, \forall i \in [1, m]$ . For each observation, one simulated  $g_k$  can be evaluated using Eq. (25) by inputting a set of  $f_i$  as sampled from the assumed scaling factor distribution, with mean  $\bar{f}$  and variance  $\sigma_f^2$ . Then, the  $N$  simulated  $g_k$  form a distribution that approximates the true, but unknown distribution of  $g_k$ . The mean  $\bar{g}_k$  and the standard deviation  $\sigma_{g_k}$  can be estimated from these  $N$  simulated  $g_k$ . It should also be noted that the first and second moments of the ESF can be estimated analytically in some cases. For instance, this factor can be estimated when the assumed scaling factor distribution is lognormal and  $m = 1$ , and  $\bar{g}_k$  and  $\sigma_{g_k}$  can be evaluated based on  $\bar{f}$  and  $\sigma_f^2$  using the well-known standard transformation equations.

Linearly projected data result from replacing each individual ESF by its estimated mean value. As only the first and second moments of the scaling factor contribute to the standard error estimations in the case of a linear model, and as the first two moments of the ESF capture the information of the higher order moments of the original scaling factor, the standard error

estimations based on the transformed case using the previously proposed generic ADF can significantly improve the estimation accuracy.

#### 4.2. Simulations

To demonstrate the estimation accuracy and effectiveness of the proposed ESF method, this subsection presents simulations on the regressions of Eq. (23) with exactly the same settings as those used in Section 3.2.

Twelve simulation cases with the same combinations in the numbers of linear combinations of the scaling factors and observable independent variables (i.e.,  $m = 1$  and  $5$ ), scaling factor distributions (i.e., normal and lognormal) and in the model nonlinearity (i.e.,  $n = 1, 2$  and  $3$ ) were conducted, based on the proposed ESF method. To enable comparison, the exact same two sets of 10,000 observations of  $x$  as sampled previously in Section 3.2 were adopted for all of the simulation cases.

For each simulation, the scaling factors  $f$  were sampled from either a normal or a lognormal distribution, with  $\bar{f} = 1$  and  $\sigma_f = 0.2$ . The random errors  $\varepsilon$  were sampled from a normal distribution with zero mean and 0.1 standard deviation. The corresponding dependent variables  $y$  were evaluated based on these sampled data, the assumed parameter values (i.e.,  $\beta_0 = 3$  and  $\beta_n = 1$ ) and the  $n$  value of each particular case. Supposing that the individual value of  $f$  and  $\varepsilon$  were unavailable, the simple transformation simulation proposed in Section 4.1 was applied to transform the model into a linear function of the ESF and to simulate the mean and standard deviation of the ESF. Instead of a direct application of the linear data projection based on  $\bar{f}$ , the sampled observable independent variables were linearly projected using the simulated ESF mean,  $\bar{g}_n$ . Regression analysis on the linearly projected data resulted in the calibrated model parameters (i.e.,  $\hat{\beta}_0$  and  $\hat{\beta}_n$ ) and the reported standard errors (i.e.,  $RSE(\hat{\beta}_0)$  and  $RSE(\hat{\beta}_n)$ ). No adjustments of the calibrated model parameters were required. The standard errors of the calibrated model parameters (i.e.,  $SE(\hat{\beta}_0)$  and  $SE(\hat{\beta}_n)$ ) were subsequently estimated based on the generic ADF method.

In a manner similar to the previous analysis, each simulation case was repeated 10,000 times, using the same set of sampled observable independent variables  $x$ , but with a different and resampled scaling factor  $f$  and random error  $\varepsilon$ . Table 3 presents the means and standard deviations of the calibrated parameters, the means of the reported standard errors and the means of the estimated standard errors for each simulation case. The standard deviations of the calibrated parameters were the unbiased estimators for their standard errors.



**Table 3**

The means and standard deviations of the calibrated parameters and the means of the reported standard errors and estimated standard errors, based on the ESF method

Assumed scaling factor distribution	$m$	$n$	Mean		Standard deviation		Mean		Mean		
			$\hat{\beta}_0$ (%error)	$\hat{\beta}_n$ (%error)	$\hat{\beta}_0$	$\hat{\beta}_n$	$RSE(\hat{\beta}_0)$ (%error)	$RSE(\hat{\beta}_n)$ (%error)	$SE(\hat{\beta}_0)$ (%error)	$SE(\hat{\beta}_n)$ (%error)	
Normal Distribution		1	3.0000 (0.00%)	1.0020 (+0.20%)	0.00177	0.00861	0.00163 (-7.75%)	0.00580 (-32.64%)	0.00175 (-0.90%)	0.00864 (+0.30%)	
		1	2	3.0000 (0.00%)	1.0091 (+0.91%)	0.00197	0.02714	0.00138 (-29.99%)	0.00726 (-73.26%)	0.00199 (+1.13%)	0.02759 (+1.66%)
		3	3.0000 (0.00%)	0.9935 (-0.65%)	0.00300	0.07049	0.00155 (-48.20%)	0.00792 (-88.76%)	0.00299 (-0.27%)	0.07035 (-0.19%)	
		1	3.0000 (0.00%)	1.0013 (+0.13%)	0.00429	0.00471	0.00394 (-8.10%)	0.00361 (-23.44%)	0.00431 (+0.63%)	0.00474 (+0.65%)	
		5	2	2.9999 (0.00%)	1.0003 (+0.03%)	0.01046	0.01051	0.00574 (-45.18%)	0.00347 (-66.99%)	0.01046 (-0.01%)	0.01053 (+0.15%)
		3	2.9993 (-0.02%)	1.0031 (+0.31%)	0.04243	0.02821	0.01309 (-69.15%)	0.00409 (-85.49%)	0.04244 (+0.04%)	0.02822 (+0.01%)	
Lognormal Distribution		1	3.0000 (0.00%)	1.0027 (+0.27%)	0.00176	0.00876	0.00163 (-7.58%)	0.00581 (-33.69%)	0.00175 (-0.62%)	0.00868 (-0.84%)	
		1	2	3.0000 (0.00%)	1.0027 (+0.27%)	0.00205	0.02849	0.00141 (-31.30%)	0.00737 (-74.12%)	0.00206 (+0.25%)	0.02854 (+0.19%)
		3	3.0000 (0.00%)	0.9969 (-0.31%)	0.00338	0.08108	0.00166 (-50.79%)	0.00848 (-89.54%)	0.00334 (-1.37%)	0.07944 (-2.02%)	
		1	3.0001 (0.00%)	1.0004 (+0.04%)	0.00435	0.00477	0.00394 (-9.39%)	0.00360 (-24.36%)	0.00431 (-0.96%)	0.00472 (-0.97%)	
		5	2	3.0000 (0.00%)	0.9974 (-0.26%)	0.01068	0.01067	0.00588 (-44.97%)	0.00355 (-66.77%)	0.01064 (-0.39%)	0.01066 (-0.16%)
		3	2.9999 (0.00%)	1.0019 (+0.19%)	0.04519	0.03004	0.01380 (-69.45%)	0.00431 (-85.65%)	0.04494 (-0.54%)	0.02983 (-0.71%)	

\* No parameter adjustment was required for any of the simulation cases.

The calibrated model parameters of the 12 simulation cases were extremely close to their assumed values (i.e.,  $\beta_0 = 3$  and  $\beta_n = 1$ ). In terms of magnitude, the percentage errors of all of the calibrated parameters were less than or equal to 0.91%. In particular, most of the percentage errors for  $\hat{\beta}_n$  in Table 3 were slightly greater than that of  $\bar{\beta}_n$  in Table 2. The exact values of the scaling factor mean and standard deviation (i.e.,  $\bar{f} = 1$  and  $\sigma_f = 0.2$ ) were used for linear data projection and parameter adjustments in Section 3.2. In contrast, the simulated ESF means,  $\bar{g}_n$ , which were subject to sampling errors arising from the scaling factor,  $f$ , were adopted for linear data projections in this section. By increasing the sample size of the observable independent variables, and hence the number of the scaling factor,  $f$ , from 10,000 sets to 100,000 sets, the significant reductions in the percentage errors of  $\hat{\beta}_n$  in these repeated simulation cases provided evidence that the sampling errors embedded in the simulated ESF means,  $\bar{g}_n$ , caused the slightly greater percentage errors, as shown in Table 3. In practical terms, minimal percentage errors of less than 1% are acceptable. These results demonstrated that the proposed ESF method indeed provided an alternative path for unbiased parameter estimations when a linear data projection was applied. However, the reported standard errors of the calibrated parameters were severely biased. The percentage errors of the reported standard errors ranged from about 7% to 90%, and they escalated rapidly as  $n$  increased.

The information about the higher order moments of the scaling factor distribution (which may contribute to the calibrated parameter variability) were captured by the first and second moments of the ESF. Thus, the proposed ESF method could better account for the effects of different scaling factor distributions and of model nonlinearity on the calibrated parameter variability. The proposed method should also significantly improve the standard error estimation accuracy compared to that of the ADF method. The means of estimated standard errors for the 12 simulation cases based on the ESF method are presented in the last two columns of Table 3. The magnitudes of their percentage errors for all of the cases were well controlled, to about 2%. These results provided evidence for a great improvement in estimation accuracy by using such a method. Moreover, due to the small magnitudes of these percentage errors, errors arising purely from random sampling might account for a considerable portion of the entire error structure. Thus, the magnitudes of the percentage errors were anticipated to be less sensitive to the model nonlinearity and the distribution of the scaling factor. Hence, the general patterns of their magnitudes with the model nonlinearity and the scaling factor distribution might not be obvious. For instance, in the case of  $m = 1$  and with normally distributed scaling factors, the magnitudes of the percentage errors of  $\widehat{SE}(\hat{\beta}_0)$  and  $\widehat{SE}(\hat{\beta}_n)$  were 0.90% and 0.30%, respectively, when  $n = 1$ . These magnitudes then increased to 1.13% and 1.66% when  $n = 2$ , and dropped to 0.27% and 0.19%, respectively, when  $n = 3$ . The clear increasing trend in the magnitudes of percentage errors with the model nonlinearity as shown in Section 3.2 was not observed in this case. Moreover, the magnitudes of the percentage errors for the estimated standard errors in cases with a log-normally distributed scaling factor could be lower than in cases with a normally distributed scaling factor. For example, when  $m = 1$  and  $n = 2$ , the magnitude of the percentage errors of  $\widehat{SE}(\hat{\beta}_0)$  and  $\widehat{SE}(\hat{\beta}_n)$  in the case of a normally distributed scaling factor were 1.13% and 1.66%,

respectively. However, in the case of a log-normally distributed scaling factor, these magnitudes were reduced to 0.25% and 0.19%, respectively. The minimal errors in the standard error estimations in situations of different model nonlinearity and scaling factor distribution demonstrated the robustness of the proposed ESF method.

In addition to the improved estimation method, a greater number of linear combinations between the scaling factors and the observable independent variables could generally cancel the random effects among different scaling factors, which could result in more accurate estimated standard errors. Taking the case of a log-normally distributed scaling factor and  $n = 3$  as an example, the magnitudes of the percentage errors of  $SE(\widehat{\beta}_0)$  and  $SE(\widehat{\beta}_n)$  were 1.37% and 2.02%, respectively, when  $m = 1$ . However, their magnitudes dropped to 0.54% and 0.71%, respectively, when  $m = 5$ .

#### 4.3. Mild violation of i.i.d. random error

Similar to Section 3.3, the same set of sensitivity analyses were conducted to examine the performance of the ESF method under situations with mild violation of i.i.d. random error. All of the simulation results showed that the magnitudes of the percentage errors of the reported and estimated standard errors obtained from situations with and without the mild assumption violation were comparable. For instance, for the case with  $m = 5$ ,  $n = 3$  and a normally distributed scaling factor, magnitudes of the percentage errors of both the reported and estimated standard errors (i.e.,  $RSE(\widehat{\beta}_0) = -69.34\%$ ,  $RSE(\widehat{\beta}_n) = -85.58\%$ ,  $SE(\widehat{\beta}_0) = 0.17\%$  and  $SE(\widehat{\beta}_n) = 0.19\%$ ) only increased slightly compared to the same case shown in Table 3 (i.e.,  $RSE(\widehat{\beta}_0) = -69.15\%$ ,  $RSE(\widehat{\beta}_n) = -85.49\%$ ,  $SE(\widehat{\beta}_0) = 0.04\%$  and  $SE(\widehat{\beta}_n) = 0.01\%$ ). When  $m = 5$ ,  $n = 3$  and the scaling factor was sampled from a log-normally distribution, the percentage errors of the reported and estimated standard errors (i.e.,  $RSE(\widehat{\beta}_0) = -69.21\%$ ,  $RSE(\widehat{\beta}_n) = -85.52\%$ ,  $SE(\widehat{\beta}_0) = 0.62\%$  and  $SE(\widehat{\beta}_n) = 0.56\%$ ) were also comparable to that of the same case shown in Table 3 (i.e.,  $RSE(\widehat{\beta}_0) = -69.45\%$ ,  $RSE(\widehat{\beta}_n) = -85.65\%$ ,  $SE(\widehat{\beta}_0) = -0.54\%$  and  $SE(\widehat{\beta}_n) = -0.71\%$ ). Moreover, these simulations also demonstrated that even if heteroscedastic random error did exist, the ESF method could substantially reduce the magnitudes of the percentage errors of the estimated standard errors compared to the reported standard errors obtained via the conventional calibration procedures.

#### 4.4. Comparisons of the ADF and ESF methods

Although both the proposed ADF and ESF methods can be easily integrated with the existing standard model calibration procedures, they have different properties and different pros and cons. Therefore, depending on the situation, the most appropriate method should be chosen and adopted in each application.

When the ADF method is used, the observable independent variables are linearly projected by the scaling factor mean  $\bar{f}$ . After model calibration based on these projected data, the calibrated parameters are corrected by the adjustment factors proposed by Wong and Wong (2015), and the standard errors are estimated based on the reported residual sum of squares using the ADF method. This method is generic and flexible, because no assumption about the distribution of the scaling factor is required. However, the estimation accuracy declines with the increase of model nonlinearity. When the ESF method is leveraged, however, an assumption about the scaling factor distribution is necessary. The assumed distribution should follow the properties of the scaling factor. A transformation simulation converts the generalized multivariate polynomial function into a linear function of the ESFs. The observable independent variables are then linearly projected, using the ESF mean,  $\bar{g}$ . As the calibrated parameters are unbiased, the standard errors can be estimated subsequently, after the model calibration. The estimated standard errors are more robust at high model nonlinearity, although an additional assumption on the scaling factor is required. Thus, making a choice between the ADF and ESF methods involves a trade-off between flexibility and accuracy. In addition, although both the ADF and ESF methods can generally perform at a satisfactory level when the assumption of i.i.d. random error is mildly violated, the ESF method can usually yield standard error estimates with smaller percentage errors in magnitudes than that obtained using the ADF method. Thus, the ESF method is in general superior to the ADF method, and it should be adopted to ensure high estimation accuracy as long as sufficient information about the scaling factor properties is available.

## 5. Case studies

Many real-world situations necessitate the applications of linear data projection. One typical example is the model calibration of the MBPR function, using data acquired from both stationary and mobile sources. The MBPR function, which is a member of the generalized multivariate polynomial family, is an essential ingredient for the continuum modeling of urban city systems (Wong, 1998; Yang and Wong, 2000; Ho and Wong, 2006; Ho et al., 2013; Yin et al., 2013). Direct model calibration based on linearly projected data may result in biased calibrated parameters, and this can definitely lead to biased estimated standard errors. This section illustrates the application of the proposed ESF method for the parameter and standard error estimations in the MBPR function calibrations for six 1 km  $\times$  1 km networks in Hong Kong, using data obtained from the on-road fixed detectors and GPS-equipped taxis.

### 5.1. The MBPR function

The MBPR function, which depicts the monotonically increasing relationship between the travel cost and traffic flow over an area-wide network, is defined as follows (Wong and Wong, 2015):

$$T = T_f + T_f \alpha Q^2 \quad (26)$$

where  $T$  is the travel time per unit of distance (measured in h/km);  $T_f$  is the free-flow travel time (measured in h/km);  $Q$  is the hourly total traffic flow entering the sampled network via different links (measured in veh/h);  $\alpha$  is the congestion sensitivity parameter (measured in h<sup>2</sup>/veh<sup>2</sup>).

## 5.2. Databases

The databases that store data obtained from stationary sources (*The Annual Traffic Census* (or ATC) 2010 (Transport Department, 2010)) and those from mobile sources (the taxi GPS database 2010) were both used for the case studies of MBPR function calibrations.

The ATC report presents detailed traffic data from more than 1500 stations covering approximately 90% of trafficable area in Hong Kong (Lam et al., 2003; Tong et al., 2003). The data for the average annual daily traffic (AADT) across each of these stations were highly useful for the constitution in this study, and were therefore extracted from the ATC report.

The taxi GPS database recorded detailed trip information from 480 GPS-equipped taxis over the course of 2010. Each probe vehicle reported its real-time location expressed in WGS84 (the ITRF96 reference frame) in decimal degrees, and the data concerning time, instantaneous speed, traveling direction, and occupancy were sent to the traffic center at a frequency of twice per minute. The full coverage of the taxi data over the entire transportation network ensured the possibility of obtaining the occupied taxi flow on any link. As the behavior of occupied taxis resembled that of normal traffic, only the speed and flow data of these taxis were retrieved from the database for data constitution.

## 5.3. Data constitution and the necessity of linear data projection

The travel time per unit of distance  $T$  and the hourly total traffic flow entering a 1 km  $\times$  1 km sampled network  $Q$  were the essential ingredients for calibrating the MBPR function of each network. As occupied taxis interacted with other vehicles around them as they traveled within the sampled networks, all of these vehicles should have traveled at similar speeds. In other words, the arithmetic mean of the speeds of the occupied taxis within a sampled network during an hour was the unbiased estimator of the hourly space-mean speed of all the traffic within that network in that hour. The reciprocal of the estimated hourly space-mean speed could be taken as the estimate of the hourly travel time per unit of distance.

The hourly total traffic flow entering a sampled network was the sum of the hourly traffic flows entering the network through all of the links intercepting the 1 km  $\times$  1 km boundary. The intercepting location of each link and the network boundary was defined as a boundary station of that sampled network. However, as only a subset of links had been outfitted with on-road fixed detectors, direct measurements of the hourly traffic flow across these boundary stations were not possible. Nevertheless, the occupied taxi flow across any link was

readily attainable through reference to the taxi GPS database. The hourly total traffic flow could be alternatively expressed as the sum of linear combinations of  $f_i$  and  $v_i$  as follows:

$$Q = \sum_{i=1}^m f_i v_i \quad (27)$$

where  $m$  is the total number of boundary stations of a sampled network;  $f_i$  is the total-traffic-to-occupied-taxi ratio of boundary station  $i$  (defined as the scaling factor of that boundary station);  $v_i$  is the observed hourly occupied taxi flow entering the sampled network through boundary station  $i$ , which was measured in veh/h and adjusted according to the ratio between the normalized traffic and occupied taxi flows hour by hour to account for the temporal effects (Wong and Wong, 2015, 2016). Although the data of  $v_i$  could be extracted from the taxi GPS database, the value of each individual  $f_i$  was still unknown.

Given the geographical proximity of each link within a sampled network, the scaling factors of all the links (including the boundary stations and the links within the sampled network) could be assumed to follow a distribution subject to certain spatial variability patterns arising from the heterogeneities of the road hierarchy and land uses of different lots. As ATC stations were outfitted with on-road fixed detectors and the AADT across them could be obtained from the ATC report, those ATC stations within the sampled network were chosen as the scaling factor sampling sites. Each sampled scaling factor was evaluated by dividing the AADT of an ATC station by the average annual daily occupied taxi flow across that station. The average value of the sampled scaling factors served to estimate the scaling factor distribution mean, which was probably the most observed total-traffic-to-occupied-taxi ratio of that network, and the variance of the sampled scaling factors measured the network's spatial variability.

As the value of each individual scaling factor was unknown, the mean and the variance of the scaling factor distribution could be estimated by using the sampled scaling factors. Therefore, a linear data projection was adopted to estimate the hourly total traffic flow entering a sampled network. If a linear data projection (in which each individual scaling factor was replaced with the estimated mean of scaling factor) was applied directly, the calibrated parameter had to be corrected by the adjustment factor proposed by Wong and Wong (2015), and the standard errors had to be estimated by using the ADF method. In contrast, if the ESF method was adopted, the ESF mean and variance had to be simulated via a transformation simulation before the model calibration, based on the linearly projected data and using the ESF mean. The generic ADF method was subsequently adopted to estimate the parameter standard errors.

#### 5.4. Model calibrations based on the ESF method

Due to its relatively robust and accurate standard error estimations, the ESF method should be adopted if sufficient information concerning the scaling factor is available. The scaling factor used in this case was the total-traffic-to-occupied-taxi ratio. An AADT of an

ATC station consists of flows of any vehicle type, including the occupied taxi flow. Thus, by definition, the scaling factor was always greater than zero. The approximately symmetrical bell-shaped histograms based on the natural logarithm of the sampled scaling factors of most of these six cases suggested that a log-normal distribution can be a potential candidate scaling factor distribution. A Kolmogorov-Smirnov (KS) goodness-of-fit test was then conducted for each sampled network with a null hypothesis that the sampled scaling factors were consistent with the specified log-normal distribution. Because the KS statistics of all six of these cases were smaller than their corresponding critical values for a level of significance of 0.05, there was insufficient evidence to reject the null hypotheses. Thus, the scaling factors were assumed to be log-normally distributed for the demonstration of the proposed ESF method. Using the transformation simulation proposed in Section 4.1, the mean and standard deviation of the ESF were obtained, and Eq. (26) was transformed into Eq. (28) as follows:

$$T = T_f + T_f \alpha g_2 \left( \sum_{i=1}^m v_i \right)^2 \quad (28)$$

Table 4 presents the sampled means and standard deviations of the sampled scaling factors and the simulated means and standard deviations of the ESFs of the six sampled networks. The column for the number of ATC stations presents the scaling factor sample sizes of the sampled networks. Column  $m$  indicates the total number of boundary stations in each sampled network. The scaling factor mean was the average total-traffic-to-occupied-taxi ratio of a sampled network. Taking Tin Hau as an example,  $\bar{f}$  was about 193, meaning that each observed occupied taxi in Tin Hau represented roughly 193 vehicles. The scaling factor standard deviation measured the spatial variability, because the scaling factors were sampled at different locations within the networks. Conversely, the ESF mean did not possess a clear physical meaning, but instead captured the information of higher order moments of the original scaling factor that contributed to the mean of the response variable. Similarly, the standard deviation of the ESF did not have a clear physical meaning, but instead captured the information of higher order moments of the original scaling factor that governed the calibrated parameter variability.

Table 5 summarizes the calibrated parameters and the estimated standard errors of the MBPR function for each sampled network, based on the ESF method. As the MBPR function was transformed into a linear function of the ESF, the calibrated model parameters were unbiased and no adjustment was required. The column for the R-squared value reveals the goodness-of-fit of calibrated MBPR functions for the six selected networks. The resultant free-flow speed  $\hat{v}_f$  was given by the inverse of the calibrated free-flow travel time  $\hat{T}_f$ , and the congestion sensitivity parameter  $\hat{\alpha}$  was evaluated by dividing  $\widehat{T}_f \hat{\alpha}$  by  $\hat{T}_f$ . A recent study revealed that the values of  $\hat{T}_f$  and  $\hat{\alpha}$  are highly associated with the junction density and road density of a network, respectively (see Wong and Wong, 2016). The column for  $\widehat{SE}$  shows the standard errors of the calibrated parameters, based on the ESF method. In addition to the random error, the effects of heteroscedasticity, which was inherently introduced by a linear

data projection with a varying scaling factor, were taken into account for better standard error estimations. In contrast, the column for  $\widehat{RSE}$  presents the reported standard errors based on the general regression procedures in which parameters and standard errors were estimated as if the distorted composite error terms were i.i.d. as normal. The biased reported standard errors could be greater or smaller than their true values. With reference to the best standard error estimates obtained by the ESF method, the reported standard errors estimated based on the invalid assumption of homoscedasticity could be overestimated to as large as 49.15% and underestimated to an extent of 39.93% in these case studies. The use of these biased reported standard errors in hypothesis testing could lead to an unnecessary risk of making the wrong decision. The  $t$ -statistics and  $p$ -values of the calibrated parameters were calculated based on the calibrated parameters and the standard errors estimated using the ESF method. It was apparent that all of the calibrated parameters were statistically significant, because their  $p$ -values were much smaller than 0.001, which was the chosen level of significance.



**Table 4**

Simulated mean and standard deviations of the ESF for each sampled network

Sampled network	Number of ATC stations	$\bar{f}$	$\sigma_f$	$m$	$\bar{g}_2$	$\sigma_{g_2}$
Tin Hau	31	193.4	74.0	28	38991.6	15420.0
Ma Tau Wai	18	193.9	81.9	26	39754.9	20486.0
Fortress Hill	25	232.2	93.8	18	55783.0	25598.3
Admiralty	15	211.2	71.4	22	45973.8	14383.1
Jordan	31	157.4	69.2	35	25131.9	7519.0
Kowloon Tong	13	193.4	45.4	19	37626.1	8625.6

**Table 5**

The calibrated parameters and estimated standard errors of each sampled network based on the ESF method

Sampled network	Model parameter	Calibrated parameter	$\widehat{RSE}$ (%error)	$\widehat{SE}$	$t$ -statistic	$p$ -value	$R^2$	$\hat{v}_f$ (km/h)	$\hat{\alpha}$ (h <sup>2</sup> /veh <sup>2</sup> )
Tin Hau	$\hat{T}_f$ (h/km)	0.0331	$3.002 \times 10^{-4}$ (+19.24%)	$2.518 \times 10^{-4}$	131.61	0.000	0.483	30.2	$2.558 \times 10^{-9}$
	$\widehat{T}_f\alpha$ (h <sup>3</sup> /km/veh <sup>2</sup> )	$8.478 \times 10^{-11}$	$1.954 \times 10^{-12}$ (-7.43%)	$2.111 \times 10^{-12}$	40.16	$1.378 \times 10^{-259}$			
Ma Tau Wai	$\hat{T}_f$ (h/km)	0.0229	$2.514 \times 10^{-4}$ (+17.11%)	$2.146 \times 10^{-4}$	106.90	0.000	0.715	43.6	$4.285 \times 10^{-9}$
	$\widehat{T}_f\alpha$ (h <sup>3</sup> /km/veh <sup>2</sup> )	$9.831 \times 10^{-11}$	$1.382 \times 10^{-12}$ (-39.93%)	$2.300 \times 10^{-12}$	42.74	$1.259 \times 10^{-284}$			
Fortress Hill	$\hat{T}_f$ (h/km)	0.0226	$1.947 \times 10^{-4}$ (+49.15%)	$1.305 \times 10^{-4}$	172.90	0.000	0.600	44.3	$2.828 \times 10^{-9}$
	$\widehat{T}_f\alpha$ (h <sup>3</sup> /km/veh <sup>2</sup> )	$6.383 \times 10^{-11}$	$1.162 \times 10^{-12}$ (-20.51%)	$1.462 \times 10^{-12}$	43.66	$1.366 \times 10^{-293}$			
Admiralty	$\hat{T}_f$ (h/km)	0.0225	$2.137 \times 10^{-4}$ (+20.97%)	$1.767 \times 10^{-4}$	127.28	0.000	0.690	44.5	$1.939 \times 10^{-9}$
	$\widehat{T}_f\alpha$ (h <sup>3</sup> /km/veh <sup>2</sup> )	$3.133 \times 10^{-11}$	$4.684 \times 10^{-13}$ (-16.68%)	$5.621 \times 10^{-13}$	55.73	0.000			
Jordan	$\hat{T}_f$ (h/km)	0.0356	$3.341 \times 10^{-4}$ (+26.45%)	$2.642 \times 10^{-4}$	134.92	0.000	0.630	28.0	$4.735 \times 10^{-9}$
	$\widehat{T}_f\alpha$ (h <sup>3</sup> /km/veh <sup>2</sup> )	$1.688 \times 10^{-10}$	$2.881 \times 10^{-12}$ (-6.67%)	$3.087 \times 10^{-12}$	54.68	0.000			
Kowloon Tong	$\hat{T}_f$ (h/km)	0.0264	$2.191 \times 10^{-4}$ (+2.43%)	$2.139 \times 10^{-4}$	123.35	0.000	0.729	37.9	$9.502 \times 10^{-9}$
	$\widehat{T}_f\alpha$ (h <sup>3</sup> /km/veh <sup>2</sup> )	$2.507 \times 10^{-10}$	$3.406 \times 10^{-12}$ (-20.83%)	$4.302 \times 10^{-12}$	58.28	0.000			

## 6. Conclusions

Linear data projection is currently prevalent, and is expected to become a widespread data inference method in many observational transportation studies that necessitate traffic data estimations. Recently, Wong and Wong (2015) unveiled the possibility that systematic bias is introduced into the calibrated parameters of model calibrations that are based on linearly projected data. These researchers proposed adjustment factors for the generalized multivariate polynomial function to reduce such bias. Nevertheless, the study only guaranteed unbiased (or slightly biased) adjusted parameters in the long run, and did not shed light on the variability of and confidence in them. Thus, the validity of the applications of the adjusted models remains unknown and suspicious. To avoid any exposure to unnecessary risk, the effects of linear data projection on the standard errors estimations must be investigated, and accurate standard error estimation procedures must be derived for credible statistical tests of the significance of the adjusted parameters. The results of this study fill this research gap and address the problem.

This study makes several important contributions to the existing body of knowledge. First, the study unveils the fact that heteroscedasticity is inherently introduced when a data projection scheme is adopted, regardless the form of the model to be calibrated. This implies that even if the adjustment factors proposed by Wong and Wong (2015) are not required for a linear model, the estimated standard errors are always biased and the methodology proposed in this paper is also always necessary. The factor of heteroscedasticity arising from linear data projection complicates the composite error structure, as the heteroscedastic data is mixed with random errors, which violates the usual assumption of homoscedasticity in model calibration. A direct model calibration based on linearly projected data without any appropriate treatment of the distorted error structure definitely leads to biased standard errors, which could be possibly greater or smaller than their true values. Moreover, due to the additive relationship between the varying  $\sigma_f^2$  and constant  $\sigma_\varepsilon^2$  and the possible nonzero expected value of the composite errors, the conventional GLS and WLS methods are not only indirect to, but also may not be able to address the problem of heteroscedasticity originating from linear data projection.

Estimating standard errors is a more difficult task than estimating parameters, because the variability of a calibrated parameter in measuring its dispersion is under consideration. This study proposes the ADF and ESF methods, which take the problem of heteroscedasticity into account when estimating the standard errors of calibrated/adjusted parameters in a generalized multivariate polynomial model. The ADF method is not only purely analytical, but is also flexible, because no assumption on the scaling factor distribution is required. However, the simulation results revealed that estimation accuracy based on the ADF method declined when the model nonlinearity was increased, and when the higher order moments of the scaling factor that governs the calibrated parameter variability were ignored. The ESF method (which is a combination of a transformation simulation and the ADF method) is less flexible, as an additional assumption about the scaling factor distribution is necessary. The transformation simulation alters the generalized multivariate polynomial model into a linear function of the ESFs. Thus, no adjustment is required for the calibrated parameters, and this method can also

be regarded as an alternative path for unbiased parameter estimation, without the use of the adjustment factors proposed by Wong and Wong (2015). The results of this study's simulation exercises demonstrated significantly improved estimation accuracy by using the ESF method. This improvement resulted from the additional information captured from the higher order moments of the original scaling factor through the ESF mean and standard deviation. Making a choice between the ADF and ESF methods in fact involves a trade-off between flexibility and accuracy. Moreover, the ESF method can usually provide more accurate standard error estimates with smaller percentage errors in magnitudes when the assumption of i.i.d. random error is mildly violated. Thus, the ESF method is generally superior to the ADF method and should always be applied if sufficient information concerning the scaling factor is available. Most importantly, these two proposed standard estimation methods can be easily incorporated with the existing classical model calibration procedures, and thus they can be considered as extensions of the standard model calibration method. These proposed methods serve to further enable the consideration of complex error structures that involve both heteroscedasticity and random errors i.i.d. as normal.

To illustrate the application of the proposed ESF method in real-life situations, model calibrations of the MBPR functions (which are members of the generalized multivariate polynomial family) were conducted for six sampled networks in Hong Kong. These models used data retrieved from both on-road fixed detectors and GPS-equipped vehicles. The total-traffic-to-occupied-taxi ratio was chosen as the scaling factor, and this factor was assumed to be log-normally distributed. The  $p$ -values of the calibrated parameters, which were evaluated based on the estimated standard errors, were much smaller than the chosen level of significance (i.e., 0.001), which indicated that all of the calibrated parameters were statistically significant.

In addition, it should be stressed that heteroscedasticity is inherently introduced for models of any form when a data projection scheme is adopted. However, the proposed ADF and ESF methods for solving this problem are only applicable to generalized multivariate polynomial models. Although the concepts involved in the proposed methodologies can generally be extrapolated, it can be far more difficult to take the complex error structure into account in cases that involve highly complicated model forms or calibration methods. The challenges of finding new extensions for parameter and standard error estimations that involve linear data projections for other complicated model forms or calibration methods may present interesting directions for future work. Furthermore, random errors are assumed to be independent identical normally distributed in the formulations of the ADF and ESF methods. However, in most occasions, random errors can be heteroscedastic. Although simulation results showed that the ADF and ESF methods generally can perform at a satisfactory level when such an assumption is mildly violated, detecting heteroscedastic random error and accurate standard error estimations under situations with the presence of heteroscedasticity arising from both linear data projection and heteroscedastic random error can be difficult, because the effects of heteroscedasticity arising from these two sources are mixed and the variance structure of the heteroscedastic random error is usually unknown. An exploration of methodologies that address heteroscedasticity originating from both sources would be a possible future research direction.

## Acknowledgements

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

## References

- Aboudolas, K., Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. *Transportation Research Part B: Methodological* 55, 265-281.
- 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, pp. 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.
- Carlsson, R., Otto, A., Hall, J.W., 2013. The role of infrastructure in macroeconomic growth theories. *Civil Engineering and Environmental Systems* 30 (3-4), 263-273.
- Chandra, S., Kumar, V., Sikdar, P.K., 1995. Dynamic PCU and estimation of capacity of urban roads. *Indian Highways* 23 (4), 17-28.
- Dowling, R.G., Singh, R., Cheng, W.W.-K., 1998. Accuracy and performance of improved speed-flow curves. *Transportation Research Record: Journal of the Transportation Research Board* 1646, 9-17.
- García-Ródenas, R., Verastegui-Rayó, D., 2013. Adjustment of the link travel-time functions in traffic equilibrium assignment models. *Transportmetrica A: Transport Science* 9 (9), 798-824.
- 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.
- 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., 2006. Two-dimensional continuum modeling approach to transportation problems. *Journal of Transportation Systems Engineering and Information Technology* 6 (6), 53-68.
- Ho, H.W., Wong, S.C., 2007. Housing allocation problem in a continuum transportation system. *Transportmetrica* 3 (1), 21-39.
- Ho, H.W., Wong, S.C., Sumalee, A., 2013. A congestion-pricing problem with a polycentric region and multi-class users: a continuum modelling approach. *Transportmetrica A: Transport Science* 9 (6), 514-545.

- Kwong, K., Kavalier, 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.
- Lum, K.M., Fan, H.S.L., Lam, S.H., Olszewski, P., 1998. Speed-flow modeling of arterial roads in Singapore. *Journal of Transportation Engineering* 124 (3), 213-222.
- 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, CA.
- Schwarzenegger, A., Bonner, D.E., Iwasaki, R.H., Copp, R., 2009. *2008 State Highway Congestion Monitoring Program (HICOMP)*, Annual Data Compilation, Caltrans, Sacramento, CA.
- 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, S.C., 1998. Multi-commodity traffic assignment by continuum approximation of network flow with variable demand. *Transportation Research Part B: Methodological* 32 (8), 567-581.
- 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., 2016. Network topological effects on the macroscopic Bureau of Public Roads function. *Transportmetrica A: Transport Science* 12 (3), 272-296.
- 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, CA.
- Yang, H., Wong, S.C., 2000. A continuous equilibrium model for estimating market areas of competitive facilities with elastic demand and market externality. *Transportation Science* 34 (2), 216-227.
- 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, CA, pp. 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.

## Appendix A. Approximate solution to the expectation of the $k$ th term of a generalized multivariate polynomial model

Here, we provide an approximate solution to the expectation of the  $k$ th term of a generalized multivariate polynomial model.

Define the  $k$ th term of a generalized multivariate polynomial model to be  $T_k = \beta_k (\sum_{i=1}^m f_i x_i)^k$ ,  $\forall k \in [0, n]$ . Approximate  $T_k$  by a Taylor series expansion with the center at  $f_i = \bar{f}$ ,  $\forall i \in [1, m]$ .

$$T_k = T_k(\bar{\mathbf{f}}) + \sum_{i=1}^m \frac{\partial T_k(\bar{\mathbf{f}})}{\partial f_i} (f_i - \bar{f}) + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 T_k(\bar{\mathbf{f}})}{\partial f_i \partial f_j} (f_i - \bar{f})(f_j - \bar{f}) + \dots \quad (\text{A1})$$

Ignoring higher order terms and taking the expectation on both sides,

$$E(T_k) \cong T_k(\bar{\mathbf{f}}) + \sum_{i=1}^m \frac{\partial T_k(\bar{\mathbf{f}})}{\partial f_i} E(f_i - \bar{f}) + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 T_k(\bar{\mathbf{f}})}{\partial f_i \partial f_j} E[(f_i - \bar{f})(f_j - \bar{f})] \quad (\text{A2})$$

As  $E(f_i) = \bar{f}$ ,  $E(f_i - \bar{f}) = 0$ . Assuming that  $f_i, \forall i \in \mathbb{N}^+$  are independent of each other,  $E[(f_i - \bar{f})(f_j - \bar{f})] = 0$ ,  $\forall i, j \in [1, m] \setminus [i = j]$  and  $E[(f_i - \bar{f})^2] = \sigma_f^2$ . It follows that

$$E(T_k) \cong T_k(\bar{\mathbf{f}}) + \frac{1}{2} \sigma_f^2 \sum_{i=1}^m \frac{\partial^2 T_k(\bar{\mathbf{f}})}{\partial f_i^2} \quad (\text{A3})$$

As  $\frac{\partial^2 T_k}{\partial f_j^2} = \beta_k k(k-1) (\sum_{i=1}^m f_i x_i)^{k-2} x_j^2$ ,  $\forall i, j \in [1, m]$ ,

$$E(T_k) \cong \beta_k \left[ 1 + \frac{k \cdot (k-1)}{2} \left( \frac{\sigma_f}{\bar{f}} \right)^2 \frac{\sum_{i=1}^m x_i^2}{(\sum_{i=1}^m \bar{f} x_i)^2} \right] \left( \sum_{i=1}^m \bar{f} x_i \right)^k \quad (\text{A4})$$