Transportation Research Part B 88 (2016) 72–92

Contents lists available at ScienceDirect

Transportation Research Part B

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

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

Wai Wong*, S.C. Wong

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

ARTICLE INFO

Article history:
Received 11 September 2015
Revised 2 March 2016
Accepted 2 March 2016
Available online 11 April 2016

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

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.

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

* Corresponding author.
E-mail address: waiwong1012@connect.hku.hk (W. Wong).

http://dx.doi.org/10.1016/j.trb.2016.03.001

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 h 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).