Estimation of scaling factors for traffic counts based on stationary and mobile sources of data

Fanyu Meng[a]*, S.C. Wong[a], W. Wong[a], Y.C. Li[a]

[a] *The University of Hong Kong, Pofulam Road, Hong Kong Island, Hong Kong*

Abstract

To combine mobile sources and stationary sources, a modeling approach to quantify the variability of the linear projection function using a non-linear regression method is established in this study. Weights that vary spatial-temporally are assigned to neighboring scaling factors. Together with a normalized weighted average function, the subject scaling factor is determined. The framework is applied to a case study in Hong Kong combining Global Positioning System data and the annual traffic counts from 85 fixed stations in Annual Traffic Census database. The performance of the models is assessed based on relative root mean square error and Akaike information criterion.

Keywords: traffic counts; scaling factors; stationary data sources; mobile data sources; GPS; taxi

1. Introduction

Traffic information such as volume and speed is crucial to studies of traffic flow modeling and traffic exposure measurement. Such information is mainly collected from two major sources: stationary sources and mobile sources, each of which has its own strengths and drawbacks. Stationary sources are well known fixed detectors such as loop detectors and radars. Video camera recording, a widely applied approach for data collection in traffic control and management, is also considered a typical stationary source. Because it is expensive to install and maintain such sources [1], their distribution usually cannot cover the whole arterial network. Therefore, to enlarge the coverage area of traffic information, mobile sources, such as the global positioning system (GPS), are gradually taking the place of traditional fixed detectors. The prevalence of mobile sources is due not only to their relatively low cost but also to the flexibility and accessibility of the probe vehicles. With enhancements in technology, the accuracy and availability of mobile source data is persuading researchers and traffic managers to adopt such sources instead of conventional fixed detectors. However, probe vehicles, no matter what their flexibility is, are always merely samples of the total population, which potentially reduces the quality of the data acquired.

To make full use of the strengths of these two data sources, researchers have identified various methods to merge them in multiple ways, one of which is by data projection. Scaling factors are collected in the whole network and considered to be random variables with a certain distribution representing certain types of variance. Linear data projection is a way of projecting the data from samples to the whole population by the mean of a group of random scaling factors [2]. Any systematic bias, which increases with the degree of variability of the random scaling factors, is usually ignored using this approach. If the variability can be reduced by identifying explanatory factors, such as spatial and temporal characteristics, by which the change in the scaling factor can be tracked, the embedded

* Fanyu Meng. Tel.: +852-65701444;
E-mail address: 101366@tongji.edu.cn

systematic bias may be reduced. Therefore, an approach is needed to measure the relationship between scaling factors and potential explanatory variables so that the reliability of the projected data can be improved.

This study used the continuity of data recorded from stationary sources together with the wide distribution coverage of mobile sources by combining the data collected from both sources. The combination of the data was achieved by modeling data projection functions using the observed scaling factors and variables that can possibly influence the scaling factors. A case study regarding taxi flow projection in Hong Kong was conducted by means of linear and non-linear regression approaches as an example and method of validating the proposed data projection approach. Various independent variables and forms of projection function were tested and analyzed.

2. Literature Review

Data extracted from stationary sources once prevailed among studies focusing on traffic state estimation and modeling. The loop detector, one of the stationary sources with the widest application, is usually embedded in the pavement of arterial roads and generates an inductive response when a vehicle passes over it. Various studies have used such detectors to acquire traffic information data because most of the major world cities are equipped with inductive loop detector systems, and the information of all vehicles that pass above has been recorded. Thus, high-quality temporal samples were guaranteed [3]. Faouzi, Leung and Kurian [4] analyzed the space mean speed and time mean speed using double-loop detector data by assuming that the speed followed a normal distribution and applied the approach to a highway in Spain. Cao et al. [5] proposed a discrete model to estimate queue characteristics for both free-flow and congested conditions using fixed-loop detector data, including a case study on a major highway in Canada. Microsimulation was applied to validate the utility of the calibrated model. Simulations were also adopted in the research on applying a formulation of boundary conditions for scalar conservation laws to traffic networks after proving that this solution, although weak, certainly exists [6]. Besides loop detectors, other stationary sources were also considered to be traditional and stable methods used to acquire traffic information data. Wireless magnetic sensors were applied to estimate the travel time of vehicles on arterial networks by matching the noise generated using a statistical model [7]. Some researchers have proposed or updated traffic stream models by means of collecting real-time traffic information through fixed detectors without identifying the detector types [8] [9]. However, although stationary sources allow wide temporal ranges and large sample sizes, their high installation and maintenance fees greatly decrease their spacing distribution density. Hence, the lack of spatial range and the relatively large systematic errors [3] of the detectors weakens the availability of stationary data to be applied to some complicated real-time models.

With the skyrocketing development of communication techniques, multiple mobile sources are gradually becoming attractive and convenient for data mining, among which GPS is the most common. The application of GPS data analyses appear prominently in estimations and predictions of travel time because the GPS dataset is straightforward and meaningful in traffic operation. Hofleitner et al. [10] proposed a method to predict how travel time distributes throughout a network, which included a dynamic Bayesian network to release the dependence on the network, and applied the modeling framework to the network in San Francisco using GPS data from 500 probe vehicles. Jenelius and Koutsopoulos [11] extracted data from scarce GPS probe vehicles and depicted the vehicle trajectories to estimate link travel times and delays caused by intersections. Origin-destination information was collected from GPS data and analyzed to estimate travel time and network operation status in New York [12]. Based on an analytical model, a neural network was applied to travel time estimation using GPS probe data and validated by microsimulation [13]. Apart from GPS information data, other mobile sources such as Bluetooth data [14] and mobile phone cellular data [15] were also popular mobile sources for traffic data acquisition. Jie et al. [16] compared travel time estimation availability and accuracy among videos, GPS and Bluetooth devices and proposed that Bluetooth scanners were not adequate to categorize traffic modes, whereas GPS and video recorded this information in their datasets.

Due to the spatial coverage advantage of mobile sources and the population coverage strength of stationary sources, a combination of the two would tend to emphasize the advantages of and compensate for the disadvantages of each source [4]. Several classical data fusion techniques have been developed and constantly upgraded over the past decades. Nanthawichit, Nakatsuji and Suzuki [17] successfully applied the Kalman filter to fuse the data from fixed detectors and probe vehicles and estimated traffic states with a macroscopic traffic flow model. Herrera and

Bayen [18] reconstructed traffic density by comparing a Lagrangian approach that included a correction term and a Kalman measurement for data fusion. The differences between the two proposed approaches were calculated by a standard algorithm for which data from California were chosen to test the accuracy of the results. Xie, Cheu and Lee [19] used multi-layer perception and multi-layer regression neural networks to fuse data collected from loop detectors and probe vehicles. Apart from the broadly adopted methods, some operational approaches such as ordered weighted averaging, which were mainly applied for decision making [20], were also considered useful in traffic state modeling.

With the development of data fusion approaches to combine stationary and mobile sources, researchers have begun to examine the pros and cons of each approach. Numerous studies have paid attention to the comparison of different data fusion methods of gathering traffic information because each classic fusion approach has its intrinsic flaws. Bachmann et al. [3] applied common fusion approaches such as the simple convex method, Single-Constraint-At-A-Time (SCAAT) Kalman filter, ordered weighted averaging and fuzzy integrals to freeway traffic speed estimation and compared the resulting root-mean-square error (RMSE) values to search for the best approach under various circumstances. Similar work was completed by Canaud et al. [21] using a microscopic simulator to model arterial networks in Paris, and they chose the weighted fusion, Kalman filter and SCATT Kalman filter methods to integrate data from loops, GPS and Bluetooth devices. Detailed analyses of the applicability of various fusion techniques were conducted to provide suitable methods for diverse traffic situations. By simply comparing various measures, [22] made it possible to merge voting technique, fuzzy regression and Bayesian technique for travel time estimation using GPS and fixed detector data, which enabled a more robust data fusion approach to estimate traffic states. Bhaskar, Chung and Dumont [23] proposed a slicing measure to predict the cumulative plots of the quartile of travel time while integrating the two sources to reduce the relative deviation that most data fusion methods can achieve. The revised algorithms for analyzing hybrid datasets have improved the possibility of source combination methodology for use in future applications.

Despite the advancement of data fusion techniques, which greatly reduce systematic errors when combining heterogeneous sources, data projection is always regarded as a straightforward approach by defining scaling factors to project certain quantities from a small population to a larger one. Linear data projection, the most popular scaling method, applies the mean of a set of previously known scaling factors to the data projection process. The scaling factors, such as passenger car units and traffic composition ratios [2], are random variables with certain types of distribution and spatial-temporal variances. Geroliminis and Daganzo [24] used linear projection to combine data acquired from fixed detectors and floating vehicle probes and illustrated the mechanism by which the spatial variance of the sample data could explain the scattered points with experimental errors, in which traffic composition ratio was considered a random scaling factor. Wong and Wong [2] showed the existence of a systematic bias in linear projection functions due to ignorance of the variability of scaling factors and proposed a method to reduce it.

Although various approaches to integrating mobile and stationary sources have been proposed, gradually revised and considerably upgraded, the systematic error of each data fusion method still exists, and assuring complete accuracy has proved to be extremely arduous work. Rather than applying these approaches with high complication and irremovable bias, linear data projection is a better approach to combine the traffic information from various sources, especially when future application of the data requires less integrated information (such as traffic volume estimation). However, although the bias of linear projection can be managed by an adjustment term, how the bias varies with the change of temporal and spatial characteristics has not been clearly addressed in the previous literature. Therefore, based on the use of data projection, this paper first proposes a modeling framework to quantify the relationship between scaling factors and spatial-temporal variables such as time and distance by nonlinear regression. Second, GPS data and fixed detector data obtained in 2011 in Hong Kong is combined by calibrating the proposed projection model. Finally, the results of the performance of the projection model in the case study together with the feasibility and validity of the modeling approach are discussed in Section 4.

3. Methodology

The purpose of the following modeling framework is to project traffic volume by integrating the data acquired from stationary sources and mobile sources. First, the scaling factors of all known stations with fixed detectors are

calculated as observations in the following modeling section. Then, the projection function model is proposed using nonlinear regression approaches.

3.1. Scaling factor

Define K as the total population of stations with fixed detectors and J as the total population of probe vehicles with a fixed time interval, Δt, between two records. For each station, S_k, $\forall k \in K$, N_{sk} and N_{mk} are defined as the traffic counts for station S_k in observation period T for stationary sources and mobile sources, respectively, in which the former is able to be directly measured from the stationary data.

To measure the number of probe vehicles that pass station S_k located at (x_k, y_k) in period T, a gate G_k for station k with two ends E_{1k} and E_{2k} with the locations (x_{1k}, y_{1k}) and (x_{2k}, y_{2k}) is set. S_k, E_{1k} and E_{2k} locate on the same line, which is perpendicular to the tangent line of the link at the point of S_k. In principle, the two ends of the gate for station S_k should be set on the edge of each side of the road. However, to compensate for the system error in the location recorded by mobile devices, a certain range of adjustment in the width of the gate can be considered accordingly. Let s_k^g represent the straight line segment connecting S_k, E_{1k} and E_{2k}, whose ends are E_{1k} and E_{2k}.

For each probe vehicle j, $\forall j \in J$, define $M_{j,t}$ as the location of this vehicle at time t , $\forall t \in T$, with the coordinates $(x_{j,t}, y_{j,t})$, where a record should exist in the location log of mobile data at time t for application purposes. Let $s_{j,t}^m$ be the curve segment following the shape of road alignment that connects locations of mobile probe vehicles $M_{j,t}$ and $M_{j,t+\Delta t}$. If there is an intersection of line segments $s_{j,t}^m$ and s_k^g, the vehicle number count, N_{mk}, increases by one, which means vehicle j has passed station S_k during time interval $[t, t + \Delta t)$.

Upon acquiring the sample vehicle count, N_{mk}, for all of the stations S_k, $\forall k \in K$, in period T, the observed scaling factor, $\widehat{f_k}$, to be applied to project traffic volume from the sample probe vehicles to the whole population of the observation area is defined as follows:

$\forall k \in K$,

$$\widehat{f_k} = \frac{V_{sk}}{V_{mk}} = \frac{N_{sk}/T}{N_{mk}/T} = \frac{N_{sk}}{N_{mk}} \tag{1}$$

where $\widehat{f_k}$ is the observed scaling factor for station S_k; V_{sk} and V_{mk} are the traffic flows for stationary sources and mobile sources, respectively; T is the observation period; and N_{sk} and N_{mk} are traffic counts for stationary sources and mobile sources, respectively, in period T. Because the traffic flows of both sources are estimated using the mean flow in a certain period, the scaling factor, $\widehat{f_k}$, for each station S_k is calculated using the notion of linear data projection, which applies the temporal mean of the time-dependent scaling factors in this case.

3.2. Projection function

As the scaling factor is not constant, we assume that a set of independent variables, $X_1, X_2, \cdots, X_n$, are possible to influence the scaling factors. The independent variables can represent all kinds of characteristics of a certain location (or a certain area) at a certain time (or in a certain period). Define I as a set of locations within the network such that $K \subseteq I$. Because the scaling factor observation, $\widehat{f_k}$, for each station S_k is readily prepared, scaling factor f_i at location i in the network, $\forall i \in I$, can be estimated by all of the other observed scaling factors, $\widehat{f_k}$, for station S_k, $\forall k \in K$.. The projection function for f_i is in the form of a weighted average of all of the other stations S_k by applying a weight, w_k, to the observed scaling factor, $\widehat{f_k}$, of each station S_k. The weight term, w_k, is a function of the independent variables:

$\forall k \in K$,

$$w_k = g_k(\boldsymbol{X}, \boldsymbol{\beta}) = g_k(X_{1k}, X_{2k}, \cdots, X_{nk}, \beta_1, \beta_2, \cdots, \beta_r) \tag{2}$$

where w_k is the weight of observed scaling factor, $\widehat{f_k}$, given by $g_k(\boldsymbol{X}, \boldsymbol{\beta})$; $\boldsymbol{X^T} = [X_{1k}, X_{2k}, \cdots, X_{nk}]$, which is a transposed vector $(1 \times n)$ of a set of selected independent variables representing spatial or temporal attributes of station k, and $\boldsymbol{\beta^T} = [\beta_1, \beta_2, \dots, \beta_r]$, which is a $(1 \times r)$ transposed vector of the corresponding unknown parameters

for the independent variables that have to be calibrated. Note that the values of $X_{1k}, X_{2k}, \cdots, X_{nk}$ for all stations should be known; thus, the unknown model is only related to $\beta_1, \beta_2, \cdots, \beta_r$, which are unrelated to k. The weight function, $g_k(\boldsymbol{X}, \boldsymbol{\beta})$, can be in either linear or nonlinear form, which should be decided according to the independent variable combination and the potential tendency of the change in w_k while a change in $\boldsymbol{X}$ takes place.

After the weight for each observed scaling factor, $\widehat{f_k}$, $\forall k \in K$, is calculated according to the weight function, the estimated scaling factor, f_i, $\forall i \in I$, can be defined using the form of weighted average together with the correction terms as follows:

$\forall i \in I$ and $k \in K$,

$$f_i = \frac{\sum_{\forall k}\left(w_k \widehat{f_k}\right)}{\sum_{\forall k} w_k} + \gamma_1 Y_{1i} + \gamma_2 Y_{2i} + \cdots \gamma_q Y_{qi} \tag{3}$$

where f_i is the estimation of scaling factor for station S_i; $Y_{1i}, Y_{2i}, \cdots, Y_{qi}$ are the correction terms for each estimation, f_i; and $\gamma_1, \gamma_2, \cdots, \gamma_q$ are the corresponding parameters for these correction terms, and q is the number of correction terms. The correction terms are established to measure the influence of some independent variables that cannot change the weight for the estimated points but that can still possibly influence the overall scaling factors.

According to the nonlinear regression approach, the unknown parameters $\beta_1, \beta_2, \cdots, \beta_r$ will be estimated by minimizing the least square function, $\varepsilon(\boldsymbol{\beta}, \boldsymbol{\gamma})$:

$\forall k \in K$,

$$\varepsilon(\beta, \gamma) = \sum_{\forall k}\left(f_k(\beta, \gamma) - \widehat{f_k}\right)^2 \tag{4}$$

Therefore, all of the parameters, $\beta_1, \beta_2, \cdots, \beta_r$ and $\gamma_1, \gamma_2, \cdots, \gamma_q$, can be obtained and the projection function can be calibrated. Given the values of $X_{1i}, X_{2i}, \cdots, X_{ni}$ and $Y_{1i}, Y_{2i}, \cdots, Y_{qi}$, the scaling factor at location i at each moment can be calculated. Hence, the two sources of data are combined using linear data projection.

4. Case Study

4.1. Introduction

A case study that shows the applicability of the proposed projection modeling framework and illustrates the process of applying the framework to a real-world situation was performed in Hong Kong. GPS probe taxis and fixed detectors were chosen to be the mobile and stationary sources, respectively, in this case. To project the volume from the GPS sample taxis to all vehicles running on the road network of Hong Kong in 2011, the linear projection method was adopted and the spatial distribution of the scaling factors were modeled using the proposed model.

4.2. Data

The GPS taxi database contains 460 probe taxis that can travel throughout the transport network of Hong Kong. Vehicle information including time, coordinates (in WGS84 format), speed and occupancy is recorded every 30 seconds. The GPS probe taxi data from 2011 were chosen for application in the data projection model as the mobile sources.

The Transport Department of Hong Kong conducts an Annual Traffic Census (ATC) to measure traffic volume conditions in the Hong Kong road network. The ATC has reported traffic flow information, public transport statistics, cross-harbor traffic and vehicle registration status for each year since 1983. The proposed methodology was first initiated in 1961 and was revised in 1965 and 1985. Data extracted from ATC 2011 were applied in the case study, which recorded the traffic volume and characteristics of 1813 km of the total 2083 km of trafficable roads in Hong Kong [25]. Among the 844 counting stations equipped with stationary traffic counters that are distributed within the road network, 217 are on Hong Kong Island, 307 are in Kowloon and 320 are in the New

Territories. The stations are divided into two categories, core stations and coverage stations, according to observation times and number of traffic characteristics covered.

In this special case, 85 ATC stations that recorded the vehicle classification and annual average daily traffic (AADT) were selected as stationary sources to match the data accuracy and coverage of the probe vehicle database and fulfil the requirements of linear projection. Of these 85 stations, which are distributed nearly equally across the three broad regions of Hong Kong, 29 are located on Hong Kong Island, 32 in Kowloon and 27 in the New Territories. The average taxi percentage for 2011 is available for all of the chosen stations, which allows taxi percentage to be applied as one of the known scaling factors.

Apart from the mobile source and stationary source, the independent variables mentioned in section 3.2 were defined before modeling. In this case study, the Traffic Characteristics Survey (TCS) database of 2011 was used to collect traffic characteristics used as independent variables. The TCS 2011 contains the Household Interview Survey (HIS), Stated Preference Survey and Hotel/Guesthouse Tourists Survey [26]. The aggregated land use variable in this case study was extracted from the trip destination database, which is subordinate to the HIS. In total, 35,401 households were surveyed in HIS 2011, of which 71% provided usable results. All of the trips, and the destination attributes, made in a typical day by all of the household members were recorded. The reported trip destinations were categorized and assembled into a trip destination database, which was adopted in land use analyses.

The zoning system adopted in the case study was established by the Planning Department of the Hong Kong Special Administrative Region (HKSAR) in a detailed land survey. All of the territories in Hong Kong were divided into 406 zones of irregularly shaped polygons, including 18 cross-boundary zones and 388 normal zones. The zoning system was used to aggregate independent variables in the proposed projection model.

4.3. Results and Discussion

The GPS taxi data for 2011 were extracted from the database, and the annual traffic count, N_{mk}, for GPS probe taxis for the kth ATC station, S_k where $\forall k \in [1, 85]$, was measured according to the coordinates of each GPS record and the location of the ATC stations. Dividing the annual probe vehicle count, N_{mk}, by 365 days gives the annual average daily flow of the probe taxis, Z_{mk}. The scaling factor for S_k can be calculated as follows:

$$\widehat{f_k} = \frac{Z_{sk}}{Z_{mk}} \tag{5}$$

where Z_{sk} is the AADT for station S_k, and Z_{mk} is the annual average daily flow of the probe taxis that have passed station S_k. 85 scaling factors were subsequently obtained.

The model calibration was based on the assumption that temporal variations in the total traffic and taxis within the same period of time are basically similar. Therefore, the independent variables include only spatial characteristics of the stations. Because the scaling factors are believed to be similar for stations that are more adjacent to each other, the distance, D_{ik}, between two stations i and k is thought to be one of the independent variables. Apart from this, the consistency of land use, L_{ik}, and the broad region of Hong Kong, R_{ik}, are also considered to influence the spatial distribution of the scaling factors.

To quantify the land information for each of the 406 zones of Hong Kong, trip destination attributes for each zone were applied and categorized because the categorized destination of a certain trip can represent the possible land use type of the destination. In the 2011 TCS database, 57 destination categories can be chosen (see Appendix A). For each zone, the percentage of number of trips for each destination category among the total number of trips arriving in the zone was calculated; hence, a distribution of trips for 57 destination categories was developed for each zone. The distribution referred to the relation between a certain zone with different trip purposes.

An agglomerative hierarchical cluster analysis was then conducted to divide the zones into different clusters according to the distributions. Squared Euclidean distance was used to measure the distances between clusters based on Ward's method, which classifies the quantities by minimizing the distance function step by step. Finally, a clustering scheme with 6 clusters was adopted: (i) mainly residential area; (ii) mainly work place; (iii) residential and miscellaneous area; (iv) work place and miscellaneous area; (v) retail area; and (vi) cross-boundary area. The number of cases and percentage of each cluster is shown in Table 1. The cluster map, which displays the

distribution of all 6 clusters in the 406 zones of the Hong Kong territories, is shown in Fig. 1. Cluster no. 0 in the cluster map represents zones with no trip distributed in the TCS trip survey in 2011, and these were deleted prior to clustering.

The land use variable, L_{ik}, was then created with 6 branches according to the 6 clusters: if S_i and S_k are in the zone with the same land use cluster, $L_{ik} = 1$; otherwise, $L_{ik} = 0$. Apart from the land use variable, region R_{ik} was defined similarly to L_{ik}: if S_i and S_k are in the same broad region (i.e., Hong Kong Island, Kowloon and the New Territories), $R_{ik} = 1$; otherwise, $R_{ik} = 0$.

Table 1. Cluster analysis results

Cluster number	Cluster name	No. of cases	Percentage
1	Mainly residential area	156	38.4%
2	Mainly work place	42	10.3%
3	Residential and miscellaneous area	88	21.7%
4	Work place and miscellaneous area	65	16.0%
5	Retail area	49	12.1%
6	Cross-boundary area	6	1.5%
Total		406	100%

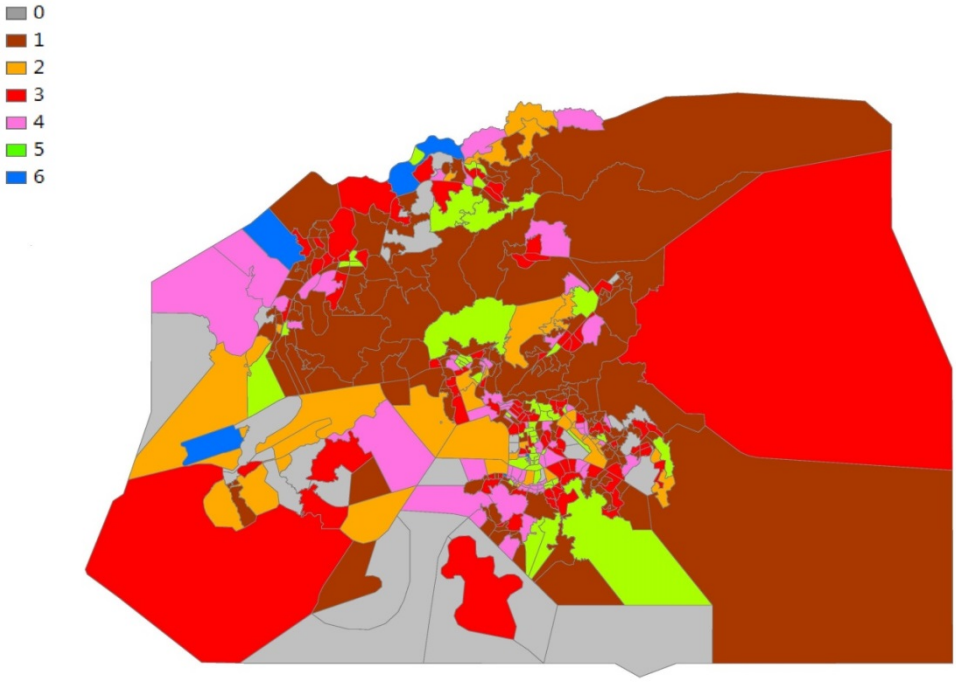

Fig. 1. Distribution of land use clusters

To exclude less useful information from the projection function and simplify the model, a linear regression projection model using the “stepwise” method in SPSS 20.0 was first performed, in which significant independent variables were identified to be applied in the projection function. The independent variables in this model included L_{ik}, R_{ik} and all of the interaction terms of L_{ik} and R_{ik}, which are denoted as $L_{ik} \cdot R_{ik}$. No cross-boundary zones were included in the selected 85 observed scaling factors. Therefore, land use cluster no. 6 was removed from the

linear regression model. The explanatory variables adopted in the linear model together with their means and standard deviations are listed in Table 2.

The results of the linear regression model are shown in Table 3. In the stepwise linear regression, the SPSS program chose the "backward" or "forward" method automatically. In this case, the "backward" method was chosen, and all of the variables were first tested in the model together. After removing the insignificant variables, three variables with significant coefficients at the 0.05 level remained: constant, "KC5" (p-value = 0.14) and "NTC1" (p-value = 0.48). The results show that the potential independent variables, L_{ik} and R_{ik}, are not significant in the linear projection function at the 0.05 level. However, two interaction terms can significantly influence the scaling factors, which indicates that the scaling factors of the stations in Kowloon with retail land usage and in the New Territories with residential areas are distinctive from the others. The coefficient of the constant term ($\beta_1 = 28.6$) shows that the average scaling factor of the stations except those belong to "KC5" and "NTC1" is 28.6, with a standard error of 2.99. The coefficients of "KC5" ($\beta_2 = 32.6$) and "NTC1" ($\beta_3 = 17.1$) indicate that the average scaling factors of the stations belonging to "KC5" and "NTC1" are considerably larger than the others. These two variables were further applied to the nonlinear projection model.

Table 2. Explanatory variables in the linear regression model

Name	Description	Mean	Std. deviation
f	Observed scaling factor	32.1	26.3
HK	HK Island	0.34	0.477
Kowloon	Kowloon	0.38	0.487
NT	New Territories (base)	-	-
C1	Land use cluster 1	0.39	0.490
C2	Land use cluster 2	0.08	0.277
C3	Land use cluster 3	0.25	0.434
C4	Land use cluster 4	0.20	0.402
C5	Land use cluster 5 (base)	-	-
HKC1	Interaction of HK Island & land use cluster 1	0.20	0.402
HKC2	Interaction of HK Island & land use cluster 2	0.06	0.237
HKC3	Interaction of HK Island & land use cluster 3	0.01	0.108
HKC4	Interaction of HK Island & land use cluster 4	0.06	0.237
HKC5	Interaction of HK Island & land use cluster 5	0.01	0.108
KC1	Interaction of Kowloon & land use cluster 1	0.07	0.258
KC2	Interaction of Kowloon & land use cluster 2	0.00	0.000
KC3	Interaction of Kowloon & land use cluster 3	0.14	0.350
KC4	Interaction of Kowloon & land use cluster 4	0.12	0.324
KC5	Interaction of Kowloon & land use cluster 5	0.05	0.213
NTC1	Interaction of New Territories & land use cluster 1	0.12	0.324
NTC2	Interaction of New Territories & land use cluster 2	0.02	0.152
NTC3	Interaction of New Territories & land use cluster 3	0.09	0.294
NTC4	Interaction of New Territories & land use cluster 4	0.02	0.152
NTC5	Interaction of New Territories & land use cluster 5 (base)	-	-

Table 3. Linear regression result

Variable	Coefficient	Standard error	t-statistics	p-value
Constant	28.6	2.99	9.549	.000
KC5	32.6	13.00	2.519	.014
NTC1	17.1	8.52	2.011	.048

Upon determination of the two most influential interaction terms, several nonlinear projection models were tested using different forms and different independent variables. The weighting function selected in these models was the exponential function:

$$w_k = exp(\beta_1 X_{1k} + \beta_2 X_{2k} + \cdots + \beta_n X_{nk}) \quad (6)$$

The notation has the same meaning as those in section 3.2. The parameters $\beta_1, \beta_2, \cdots, \beta_n$ to be calibrated in Equation (6) should be controlled to be positive or negative according to the potential relationship between X_{ik} and w_k, which should also be a method of model validation. The projection function with correction terms is settled according to Equation (3). The explanatory variables for the nonlinear regression models are shown in

Table 4.

Three nonlinear models were finally established using the proposed framework, and the combinations of independent variables of the three nonlinear projection functions are listed in Table 5. The calibrated parameters and results of the nonlinear regression models are shown in Table 6.

Table 4. Explanatory variables for nonlinear regression models

Variable	Description
f_i	Scaling factor of station S_i
D_{ik}	Distance between station S_i and S_k; km
R_{ik}	If station S_i and S_k are in the same broad region: 1; otherwise: 0
L_{ik}	If station S_i and S_k are in the zones with same land use cluster: 1; otherwise: 0
δ_i^{kc5}	If station S_i is in Kowloon with land use cluster 5: 1; otherwise: 0
δ_i^{ntc1}	If station S_i is in the New Territories with land use cluster 1: 1; otherwise: 0

Table 5. Combination of independent variables

Model no.	1	2	3
X	D_{ik}	D_{ik}, R_{ik}, L_{ik}	D_{ik}
Y	-	-	δ_i^{kc5}, δ_i^{ntc1}

Table 6. Results of nonlinear regression modeling

Model no.	Variable	Parameter	Std. error	t-statistics	p-value	rRMSE	AIC
1	D_{ik}	-0.630	0.140	4.495	0.000	61.9%	931.3
2	D_{ik}	-1.255	0.558	-2.251	0.027	58.1%	924.4
	R_{ik}	-4.315	1.612	-2.677	0.009		
	L_{ik}	0.139	2.535	0.055	0.956		
3	D_{ik}	-0.549	0.230	-2.385	0.019	59.3%	927.8
	δ_i^{kc5}	26.081	15.633	1.668	0.099		
	δ_i^{ntc1}	14.613	9.682	1.509	0.014		

In Model 1, only the distance between stations S_i and S_k was adopted as an independent variable. The estimated parameter of D_{ik}has a p-value of 0.000, which means that this parameter is significant at the 0.01 level in Model 1. This estimated parameter is negative, which indicates that the greater the distance between two stations, the smaller the weight to be given to S_k while estimating f_i, which is consistent with the basic assumption. This model is applicable to the future volume projection from the GPS taxi samples to the total traffic in the Hong Kong road network.

Models 2 and 3 assumed that a broad region and land use type were two of the three major independent factors (with distribution of probe taxi traffic being the third factor) that could change the distribution of the scaling factors. The difference between nonlinear Models 2 and 3 is that Model 2 used variables representing region and land use consistency in the weighting function, whereas Model 3 applied "KC5" and "NTC1", the two variables proved to be significant in the linear regression model, as the two correction terms in the overall scaling factor calculation. The results of Model 2 show that distance D_{ik} (estimated parameter is -1.255) and broad region R_{ik} (estimated parameter is -4.315) can significantly influence the distribution of scaling factors (p-values equal to 0.027 and 0.009, respectively). However, the parameter D_{ik} in Model 2 is nearly twice that in Model 1, whereas the parameter R_{ik} is negative, which decreases the weight if stations S_i and S_k are in the same broad region. This result indicates that after adopting the region characteristics, the larger influence of distance should be considered: if the two stations, S_i and S_k, are not in the same region ($R_{ik} = 0$), the influence of distance should be emphasized; if the two stations, S_i and S_k, are in the same region ($R_{ik} = 1$), the influence of distance should be weakened compared to the influence of R_{ik}. Basically, the negative parameters of D_{ik} and R_{ik} in the exponential weighting function validate that distance and region are not independent, which quantifies the intricate relationship between the two independent variables and the scaling factor. In Model 3, the parameters of the two correction terms are consistent with the linear model at a certain scale. The parameter of δ_i^{ntc1} is significant at the 0.05 level, whereas the parameter of δ_i^{kc5} is significant at the 0.1 level. Compared with Model 1, the extra correction terms in the projection functions of Model 3 diminish the absolute parameter of distance D_{ik}. This result indicates that if station S_i is in a zone with other land use types (other than "KC5" and "NTC1"), a comparatively larger weight should be assigned to each station S_k where $k \neq i$. Therefore, to some extent, δ_i^{ntc1} and δ_i^{kc5} strengthen the dependence of estimated factor, f_i, to the distance D_{ik}.

To compare the proposed models statistically, the Akaike information criterion (AIC) was adopted as the main statistic, which indicates the information lost in each model. The AIC of a model is defined as [27]:

$$AIC = (-2)\, log(maximum\ likelihood) + 2a \tag{7}$$

where a is the number of parameters. For least square cases, AIC is defined by:

$$AIC = N \cdot ln\left(\frac{RSS}{N}\right) + 2a \tag{8}$$

where RSS is the residual sum of squares and N is the sample size used for regression. According to the results in Table 6, the lowest AIC comes from Model 2, which cuts the residual sum while adopting three independent variables, indicating that Model 2 loses the least information among the three models. The relative RMSE (rRMSE) for the three proposed nonlinear models in this case provides the same results as the AIC. Therefore, Model 2 is the first choice among these three nonlinear models for further application with regard to taxi spatial projection in Hong Kong in 2011.

The standard deviation of the observed scaling factors is 26.3 (Table 2), and the RMSE of Model 2 is 24.1. In linear projection, the average value of scaling factors is adopted to project data from a sample to the target value, and thus the standard deviation of the observed scaling factors is equal to RMSE of linear projection. In other words, the variability of the scaling factors is reduced by quantifying the relationship among the scaling factors and exploratory variables. If the model is restricted by other unknown factors in other situations, or if more variability such as temporal variance will be incorporated, the similar procedures are capable of being introduced into other projection models to measure certain relationships.

5. Conclusion

Mobile sources and stationary sources are two indispensable data sources in transportation research and management. Stationary sources provide traffic information of the whole population, but it is usually impossible to cover the whole network of a city. Mobile sources are capable of moving anywhere within a network, but they are limited by sample size. Therefore, combining these two sources to use their advantages and compensate for their disadvantages is becoming a trend in traffic information data acquisition, for which the linear projection function is a straightforward and efficient approach with regard to data fusion. However, because it is difficult to measure and completely remove the systematic bias of linear projection, the results of such an approach need to be further purified.

In this paper, a modeling framework that introduces the procedures to establish a non-linear projection model for integrating mobile sources and stationary sources is proposed. A weighting function is proposed to offer various weights to the observed scaling factors, and a weighted summation function that combines the observed scaling factors is defined to estimate the scaling factor. A nonlinear data regression method is then adopted to calibrate the unknown parameters using the known information of the stations and mobile data by minimizing the sum of square residuals. The projection model is capable of measuring the influence of several spatial-temporal independent variables on the scaling factor of a certain location. The relationship between a scaling factor and the corresponding spatial-temporal information can be quantified, and the variability can be reduced by the proposed modeling framework, in which spatial and temporal variances in the scaling factors are measured and applied to further projections.

A case study that exemplifies the procedures to apply the proposed modeling framework to a real-world situation is then performed on data from the Hong Kong road network. ATC stations and GPS probe vehicles are chosen as the respective stationary and mobile sources in this case. In considering the ratio of the AADT to annual average daily probe taxi flow, land use type, distance and broad region are thought to be influential to the spatial distribution of scaling factors. To acquire the land-use characteristics of the 406 zones, an agglomerative cluster analysis approach is conducted, which classifies the trip destination into 6 clusters: mainly residential area, mainly work place, residential and miscellaneous area, work place and miscellaneous area, retail area and cross-boundary area. By applying the results of the cluster analysis, a stepwise linear regression model with all of the interaction terms of land use characteristics and broad region is established. As a result, Kowloon with land use cluster 5 (KC5) and the New Territories with land use cluster 1 (NTC1) are found to have a significant influence on the scaling factor. Thus, 3 nonlinear regression models with different combinations of independent variables are proposed, and the forms of the weighting function are all exponential functions. The results illustrate that distance is undeniably a significant variable in this case, with a negative parameter in the weighting function. Besides, the broad region attribute, if treated as an independent variable in the power of the weighting function, can also significantly influence the projection result, whereas land use type is not significant in the same model. Of the interaction terms found in the linear model, “NTC1” is significant at the 0.05 level and “KC5” is significant at the 0.1 level, where the two variables are considered as correction terms in the weighted average function (final projection function). Finally, AIC is applied to accomplish the model comparison process, which proves that the second model contains more information in this case, and the variability of the scaling factors is proved to be reduced. The whole process of the case study in Hong Kong is capable of certifying the availability of the proposed modeling framework by establishing a volume projection function for probe taxis and fixed detectors. The error of the projection is shown to be reduced compared with traditional linear projection, when the influences of various explanatory factors on the scaling factors are accounted for. The results can be applied to traffic analyses such as measurement of taxi exposure and other variables.

The present case study tests the availability of the modeling framework in integrating the probe taxi source and fixed detector source and proposes three possible variable combinations of independent variables. However, no time-related independent variable is tested, and temporal variation is not quantified in this case. For future research, the fitness of other mobile and stationary sources that provide different types of traffic information data need to be examined, and the applicability of the framework to modeling temporal variation needs to be validated. Moreover, the complexity of the weighting function and projection function is likely to be raised when introducing more realistic factors into the model. For weight calculation, functions such as logarithmic, proportional and power

functions can be tested in various situations, whereas for the scaling factor calculation, the normalized weighted average serves merely as an example of combining the weights and observed scaling factors. Other function and modeling approaches can be proposed and incorporated based on different weighting functions and the various objectives of the data projection.

Acknowledgements

This work was supported by a Research Postgraduate Studentship from the University of Hong Kong, and grants from the University Research Committee of the University of Hong Kong (201511159015), and Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 717512, 17208614). We gratefully acknowledge the Transport Department of HKSAR for providing the ATC and TCS traffic information data and Motion Power Media Limited and Concord Pacific Satellite Technologies Limited for offering the GPS data.

Appendix A. Trip Destination Categories in TCS 2011

No	Trip Destination Category
1	Home
2	Usual place of work
3	Other places of work
4	School/educational institution that oneself studies in
5	School/educational institution that other people studies in
6	Bank
7	Market/shopping arcade/shops/supermarket (excluding restaurant, cinema, karaoke)
8	Food premise: restaurant, café, pub etc.
9	Cinema/concert, karaoke and other entertainment facilities (including nightclubs/bathhouses/massage/establishments/club mahjong rooms/maghong-tin kau premises)
10	Park/playground/sports ground/court/gym centre/swimming pool/beach
11	Hospital/ medical center/ clinic
12	Cultural/ community center, library and arts/ museum
13	Church/ temple
14	Elderly homes
15	Home of family member/ relative/ friend
16	Boundary control point (departure)
17	Boundary control point (arrival)
18	Anywhere in the mainland
19	Anywhere in the overseas
20	Another home/ second home (occasional)
21	Police station/ immigration department/ other government offices
22	Post office
23	Petrol station
24	Parking lot
25	Bus/ public light bus terminal/stop/MTR/LRT station/ pier (non-transport purpose)

26	Places of work of other people (non-work purpose)
27	Industrial and commercial buildings (non-work purpose)
28	Betting braches
29	Racecourse
30	Securities company/ investment center
31	Travel agent
32	Funeral home/ cemetery/ crematorium/ bone ash space
33	Labor organization/ association
34	Learning center
35	Tutorial center
36	Examination venue
37	Beauty salon/ barber shop
38	Hotel (for residence)
39	Hotel (visiting relatives/friends)
40	Barbecue venue
41	Country park/ countryside area
42	Farmland
43	Fishing site
44	Holiday camp
45	Pet shop
46	Volunteer center
47	Tourist attractions
48	Chek lap kok airport (non-departure or arrival)
49	Law firm
50	Garage / car repair center
51	Driving center
52	Waterfront (for recreation)
53	Viewing property
54	Carrying/meeting family
55	Driving / taking a ride
56	Place nearby to home/ downstairs
57	On street/ roaming the street/ no particular destination

References

[1] Herrera, J. C., D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies, 18, No. 4, 2010, p. 568-583.

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

[3] Bachmann, C., B. Abdulhai, M. J. Roorda, and B. Moshiri. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. Transportation Research Part C: Emerging Technologies, 26, 2013, p. 33-48.

[4] Faouzi, N.-E. E., H. Leung, and A. Kurian. Data fusion in intelligent transportation systems: Progress and challenges – A survey. Information Fusion, 12, No. 1, 2011, p. 4-10.

[5] Cao, J., M. Hadiuzzaman, T. Z. Qiu, and D. Hu. Real-time queue estimation model development for uninterrupted freeway flow based on shockwave analysis. Canadian Journal of Civil Engineering, 42, No. 3, 2015, p. 153-163.

[6] Strub, I. S., and A. M. Bayen. Weak formulation of boundary conditions for scalar conservation laws: an application to highway traffic modelling. International Journal of Robust and Nonlinear Control, 16, No. 16, 2006, p. 733-748.
[7] Kwong, K., R. Kavaler, R. Rajagopal, and P. Varaiya. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. Transportation Research Part C: Emerging Technologies, 17, No. 6, 2009, p. 586-606.
[8] Chowdhurya, D., and R. C. Desaib. Steady-states and kinetics of ordering in bus-route models: connection with the Nagel-Schreckenberg model. The European Physical Journal B, 15, 2000, p. 375-384.
[9] Ma, D., D. Wang, Y. Bie, S. Jin, and Z. Mei. Identification of spillovers in urban street networks based on upstream fixed traffic data. KSCE Journal of Civil Engineering, 18, No. 5, 2014, p. 1539-1547.
[10] Hofleitner, A., R. Herring, P. Abbeel, and A. Bayen. Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. Ieee Transactions on Intelligent Transportation Systems, 13, No. 4, 2012, p. 1679-1693.
[11] Jenelius, E., and H. N. Koutsopoulos. Travel time estimation for urban road networks using low frequency probe vehicle data. Transportation Research Part B: Methodological, 53, 2013, p. 64-81.
[12] Zhan, X., S. Hasan, S. V. Ukkusuri, and C. Kamga. Urban link travel time estimation using large-scale taxi data with partial information. Transportation Research Part C: Emerging Technologies, 33, 2013, p. 37-49.
[13] Zheng, F., and H. Van Zuylen. Urban link travel time estimation based on sparse probe vehicle data. Transportation Research Part C: Emerging Technologies, 31, 2013, p. 145-157.
[14] Bhaskar, A., and E. Chung. Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. Transportation Research Part C: Emerging Technologies, 37, 2013, p. 42-72.
[15] Caceres, N., L. M. Romero, F. G. Benitez, and J. M. del. Traffic Flow Estimation Models Using Cellular Phone Data. IEEE Transactions on Intelligent Transportation Systems, 13, No. 3, p. 1430-1441.
[16] Jie, L., H. van Zuylen, L. Chunhua, and L. Shoufeng. Monitoring travel times in an urban network using video, GPS and Bluetooth. Procedia - Social and Behavioral Sciences, 20, 2011, p. 630-637.
[17] Nanthawichit, C., T. Nakatsuji, and H. Suzuki. Application of probe vehicle data for real-time traffic state estimation and short-term travel time prediction on a freeway. TRB 2003 Annual Meeting, 2003.
[18] Herrera, J. C., and A. M. Bayen. Incorporation of Lagrangian measurements in freeway traffic state estimation. Transportation Research Part B: Methodological, 44, No. 4, 2010, p. 460-481.
[19] Xie, C., R. Cheu, and D. Lee. Improving arterial link travel time estimation by data fusion. TRB 2004 Annual Meeting, 2004
[20] Yager, R. R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on Systems, Man and Cybernetics, 18, No. 1, 1988, p. 183-190.
[21] Canaud, M., A. Nabavi, C. Bécarie, D. Villegas, and N. E. E. Faouzi. A realistic case study for comparison of data fusion and assimilation on an urban network – the archipel platform. Transportation Research Procedia, 6, 2015, p. 28-49.
[22] Choi, K., and Y. Chung. A data fusion algorithm for estimating link travel time. ITS journal, 7, No. 3-4, 2002, p. 235-260.
[23] Bhaskar, A., E. Chung, and A.-G. Dumont. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. Computer-Aided Civil and Infrastructure Engineering, 26, No. 6, 2011, p. 433-450.
[24] Geroliminis, N., and C. F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological, 42, No. 9, 2008, p. 759-770.
[25] Transport Department, H. K. The Annual Traffic Census 2011. Transport Department, Hong Kong, Hong Kong Special Adiministrative Region, 2012.
[26] Travel Characteristics Survey 2011 Final Report. Hong Kong Special Administrative Region, 2014.
[27] Akaike, H. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, AC-19, No. 6, 1974, p. 716-723.