IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 25, 2019, accepted March 4, 2019, date of publication March 12, 2019, date of current version April 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904645

A Multiple Regression Approach for Traffic Flow Estimation

LILIAN PUN[1], PENGXIANG ZHAO[2], AND XINTAO LIU[1]
[1]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong
[2]Institute of Cartography and Geoinformation, ETH Zürich, Zürich, Switzerland

Corresponding author: Pengxiang Zhao (pezhao@ethz.ch)

This work was supported in part by the Hong Kong SAR Government through the General Research Fund under Grant B-Q43R, and in part by the PolyU Fund under Grant G-YN99.

ABSTRACT Traffic flow information is of great importance for transport planning and related research. The conventional methods of automated data collection, such as annual average daily traffic (AADT) data, are often restricted by limited installation, while the state-of-the-art sensing technologies (e.g., GPS) only reflect some types of traffic flow (e.g., taxi and bus). Complete coverage of traffic flow is still lacking, thus demanding a rigorous estimation model. Most studies dedicated to estimating the traffic flow of the entire road network rely on single to only a few properties of the road network and the results may not be promising. This paper presents an idea of integrating five topological measures and road length to estimate traffic flow based on a multiple regression approach. An empirical study in Hong Kong has been conducted with three types of traffic datasets, namely floating car, public transport route, and AADT. Six measures, namely degree, betweenness, closeness, page rank, clustering coefficient, and road length, are used for traffic flow estimation. It is found that each measure correlates differently for the three types of traffic data. Multiple regression approach is then conducted, including multiple linear regression and random forest. The results show that a combination of various topological and geometrical measures has proved to have a better performance in estimating traffic flow than that of a single measure. This paper is especially helpful for transport planners to estimate traffic flow based on correlation available but limited flow data with road network characteristics.

INDEX TERMS Traffic flow estimation, topological and geometrical Properties, correlation analysis, multiple linear regression, random forest.

I. INTRODUCTION

Urban traffic flow information has always been essential for individual travelers, transportation planning, vehicle management and urban development. In particular, a complete real traffic flow of the entire road network is of great interest for both practitioners and researchers across many fields. It is significant to assist travelers in selecting reasonable travel time and path. Specifically, with the rapid development of Intelligent Transportation Systems (ITS), the demand for traffic flow information in real-time is increasing. To supplement incomplete datasets in space and time, it is necessary to research on reliable methods of traffic flow estimation.

There are currently various methods to collect traffic flow information such as manual street surveys, probe vehicles or floating car data (FCD), road-side detectors and closed-circuit television (CCTV) camera video images, among which loop-detected data and FCD are typical stationary and mobile collection methods respectively. Annual traffic census data collected by road-side detectors have higher precision due to less influence of external factors and the number of all types of vehicles passing the sensor-installed location can be counted. Various researches have dedicated to the estimation and prediction of traffic flow based on traffic census data [7], [14], [18], [25], [28], [31]. Compared with loop-detected data collected at fixed locations and with low coverage, GPS-enabled floating cars have a wider coverage, which is a necessary supplement to the stationary data collection technology. However, floating car data normally only reflect traffic flow information about one particular vehicle fleet, such as a taxi or truck. In addition, low sampling frequency and limited spatial coverage in several periods are drawbacks for estimating traffic flow. Therefore, in reality missing traffic flow information is still common for the whole

The associate editor coordinating the review of this manuscript and approving it for publication was Safdar Hussain Bouk.

road network. How to estimate traffic flow based on loop-detected data, floating car data or other traffic data source becomes a critical issue in urban traffic flow estimation.

The scope of this paper is to investigate a multiple regression approach for traffic flow estimation, which integrates both topological and geometrical characteristics of the road network. In this work, Hong Kong Island is selected as a case study. Specifically, three types of data sources of traffic flow namely Annual Average Daily Traffic (AADT) from roadside detectors, Public Transport from its schedule of service and GPS-enabled fleets of GoGoVan are compared to evaluate their relative association of the model. The results indicate that superior traffic flow estimation can be achieved with a multiple regression approach by combining the topology and geometry features of the road network.

The next section reviews the past researches on traffic flow estimation based on road network characteristics. Section III introduces the multiple variables regression approaches including multiple linear regression and random forest regression. The selected geographic and topological measures of road network are displayed in Section IV. Section V presents the study area and the three types of traffic data. Section VI analyzes and discusses the traffic flow estimation results. The contributions and future directions of this study are summarized in the last section.

II. REVIEW OF RESEARCHES ON TRAFFIC FLOW

In recent years, under the growing requirement of traffic flow information, various traffic data collection methods have been evolving considerably such as pneumatic road tubes, microwave radar, probe vehicles or floating car data (FCD), road-side sensors [34] and closed-circuit television (CCTV) camera video images [19]. Annual traffic census data are collected by installing inductive loops and pneumatic tubes on the road to record the passing vehicle flows. However, due to the limited coverage and expensive costs of implementation and maintenance, this data source is not sufficient to cover the entire road network. In addition, installations of these sensing devices are often restricted to major roads or highways only. Yet, serious congestion problems or bottleneck conditions at the small arterial roads and their junctions are not uncommon. Similarly, CCTV camera can capture 7×24 hours' data but is restricted to particular road segments only. Automatic derivation of traffic flow information from the video is still not very mature. The emergence of FCD provides an effective method for collecting traffic data with its advantages of lower cost, real-time collection and wider coverage [2], [3], [22]. However, the consent of the vehicle drivers or companies such as taxi, bus, truck needs to be sought in which privacy is still a concern. Furthermore, FCD is normally collected based on one type of company fleet, thus the traffic flow information collected is only partial. Although the above-mentioned data sources have been gradually applied to estimate traffic flow [10], [18], [30], [38], it seems so far no single method enables a complete set of data both spatially and temporally. For places and times without traffic flow captures, it is necessary to develop a method for accurately estimating urban traffic flow based on available data sources.

An enormous amount of studies has indicated that the topological and geometrical properties of road network have a significant influence on urban traffic flow [5], [8], [12], [13]. For topology, Jiang *et al.* [13] studied how road centrality measures correlated to traffic flow from the perspective of natural roads and examined the impact of join principles of segments on metric-flow correlation. The results indicated that weighted PageRank, PageRank and connectivity were the best metrics in terms of metric-flow correlation with the correlation coefficient greater than 0.7. Jiang and Liu [14] predicted traffic flow of Hong Kong using topological measures of road connectivity, path length and clustering coefficient to correlate with AADT datasets. The correlation coefficients were just about 0.3 due to the diverse topography of the study area. Kazerani and Winter [17] argued that it was not appropriate to analyze traffic flow with traditional betweenness or centrality measures alone as the dynamics of travel behavior was neglected. Leung *et al.* [21] proposed a framework which combined traditional betweenness measures with travel speed to predict traffic flow using taxi trajectory data. It was found that mere topological properties were usually insufficient for the prediction of traffic flow. Gao *et al.* [6] also carried out urban traffic flow estimation using taxi trajectory data to correlate with conventional betweenness or centrality measures of the road network, but the results were not satisfactory. Jayasinghe *et al.* [11] examined the capability of centrality measures of connectivity, global integration, local integration and choice to predict traffic flows of different types of vehicles. Ye *et al.* [35] proposed a modified betweenness measure to predict traffic flow using GPS taxi trajectory data. The results showed that the modified measure had better correlation with observed taxi traffic flow. Zhao *et al.* [40] proposed an improved network centrality measure framework to analyze urban traffic flow using GPS taxi trajectory data, which considered both the geometrical and topological properties of the road network. However, the estimated traffic flow from GPS taxi trajectory data was incapable of representing actual traffic flow with only a rectangular region in the downtown selected as the study area.

Apart from road topology, the geometrical characteristics of the road network have also been used to estimate traffic flow. Cheng [4] proposed a regression model based on road functional classification – road width, surface type and population in geographical areas to estimate traffic flow. Xia *et al.* [32] studied a model to estimate AADT for non-state roads in urbanized areas in Florida, which involved both road geometry (e.g. the number of lanes, road functional classification) and socioeconomic variables (e.g. population, dwelling units). Zhao and Chung [37] developed several regression models for estimating traffic flow based on several variables, including road functional classification, the number of lanes, access to expressways etc. Anderson *et al.* [1] estimated annual average traffic daily flow

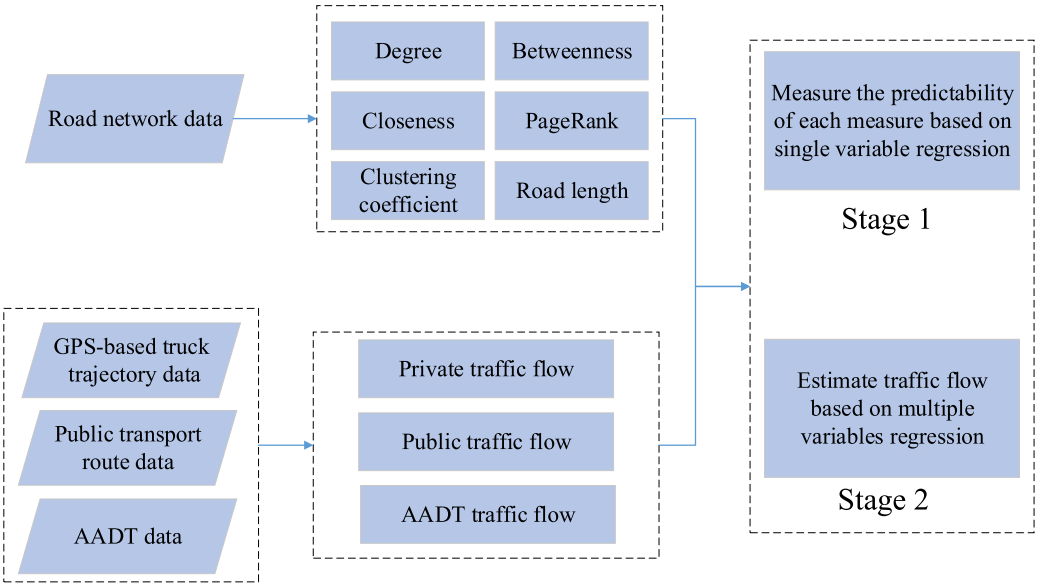


FIGURE 1. The workflow of the proposed approach.

through developing a model with five explanatory variables, including road functional classification, number of lanes, population and employment within a half mile and whether the road was a through the street. Lowry and Dixon [24] estimated AADT taking into account the number of lanes and speed limit based on linear regression. A regression model was developed to estimate AADT using a combination of socio-economic factors and geometrical properties of the road network including the number of lanes and road functional classification [5]. All these studies indicated that it was more appropriate to estimate traffic flow by combining multiple related factors. Yet, the topological characteristics of the road network in the traffic flow estimation were ignored.

All these previous studies demonstrate that urban traffic flow is closely associated with either the topological or geometrical characteristics of the road network. Integrating both multiple topological and geometrical measures to estimate traffic flow has been paid little attention. Generally speaking, a single topological or geometrical measure normally reflects one type of road characteristic. For instance, for two roads with the same number of lanes, the road with higher connectivity is normally characterized by a higher traffic flow. Therefore, it is worthwhile to explore traffic flow estimation by combining multiple topological and geometrical characteristics of the road network.

Additionally, there are also a bunch of studies that focus on short-term traffic prediction based on historical traffic data. For instance, Zhang and Liu [36] applied LS-SVMs to forecast traffic flow of one week based on the traffic flow in the former 23 weeks. Karlaftis and Vlahogianni [16] discussed differences and similarities between statistical methods and neural networks in the field of transportation such as traffic flow analysis and forecasting. Tang *et al.* [27] proposed a method to forecast travel speed by constructing a fuzzy neural network based on 2-minute travel speed data. Results were found to be better when compared with six traditional models. Traffic flow estimation in this work was based on both topological and geometrical characteristics of the road network.

III. MULTIPLE VARIABLES REGRESSION APPROACH

For many previous studies, available traffic flow data are not comprehensive enough to cover all roads in time and in space. An accurate and reliable method to estimate traffic is needed based on the known data points and time, preferably with diverse data sources for wider coverage. In this study, a multiple regression approach is proposed to improve the traffic flow estimation. It consists of two stages:

(a) To measure the predictability of each geometrical and topological measure with respect to the individual type of traffic flow data.
(b) To develop multiple regression models through integrating the topological and geometrical measures.

The workflow of the proposed approach is shown in Figure 1. In this study, two models of multiple regression are introduced – multiple linear regression and random forest representing typical linear and non-linear multiple regression models respectively. There are certainly many other regression models such as finite mixture regression model, Bayes method [41]. In this paper, the focus is on examining whether multiple regression analysis is superior to univariate analysis in traffic flow estimation, instead of comparing the performance of various regression approaches.

Multiple regression analysis refers to constructing a prediction model by analyzing the correlations between two or more independent variables and a dependent variable. If there is a linear relationship between the dependent variable and independent variables, it is called multiple linear regression.

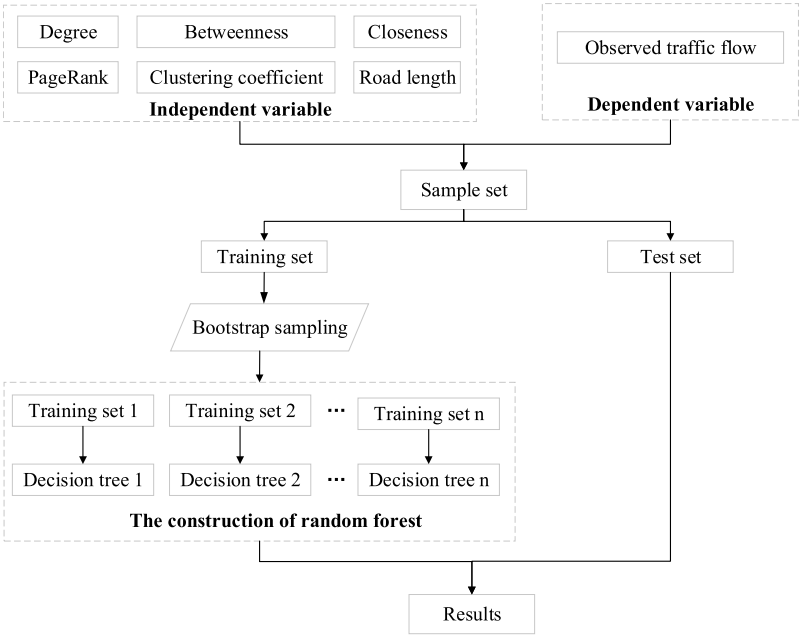


FIGURE 2. The flowchart of the random forest.

Compared with a univariate linear regression model, the dependent variable such as urban traffic flow is normally associated with multiple factors.

As one of the machine learning techniques of classification and regression, random forest is developed by combining a large set of regression trees based on ensemble learning (Breiman, 2001). In random forest regression, each tree stands for a set of conditions or restrictions. Based on a deterministic algorithm, it is built by selecting a random sample and a random set of variables from the training dataset. Specifically, for the problem of regression prediction, random forest algorithm selects a weighted average method to improve the prediction accuracy by summing up a great number of trees. The superiority of random forest is its high prediction accuracy under the same operation rate and better fitting of nonlinear data compared with traditional statistical methods. It has then been commonly applied to urban traffic flow prediction [9], [20], [33]. However, little attention has been paid to traffic flow estimation from the perspective of road network characteristics.

In this study, a random forest algorithm is implemented in the "RandomForest" package [23] within R environment software [26] to estimate traffic flow. The selected topological and geometrical measures are regarded as independent variables, whereas the observed traffic flow is the dependent variable. The process consisting of three main steps is displayed in Figure 2. First, the topological and geometrical characteristics of the road network as well as observed traffic flow of the sample dataset is randomly divided into two parts, namely training set and test set. Second, simple random sampling with replacement is executed multiple times for the training set from which n bootstrap training sets are obtained. Each bootstrap training set can be used to construct a decision tree. The random forest model is constructed by n decision trees. Third, the test set is further used to verify the model and obtain the results.

IV. GEOMETRICAL AND TOPOLOGICAL MEASURES

To implement multiple variable regression model, urban traffic flow is correlated with both the geometrical and topological characteristics of the road network. Road length is the geometrical property adopted in this study. There are two frequently-used approaches for deriving length – named road and natural road. The former relies on the road name to merge the road segments. The latter merges the adjacent road segments according to their continuation. Here, the length of each road by name is used. This is because drivers are used to be attracted more to familiar and major roads implied by its name. Also psychologically, they tend to avoid turning to another road as possible unless really necessary.

Five topological measures are selected to quantify the topological characteristics of road network – degree, betweenness, closeness, PageRank, and clustering coefficient.

(1) *Degree* measures the local connectivity of the road segment. It is defined as the number of other segments directly connected to it and is also called connectivity in space syntax [15].

(2) *Betweenness* refers to the number of times a road segment acts as a bridge along the shortest path between all pairs of nodes in the entire planar graph network. It is calculated by the following formula:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{d_v(s,t)}{d(s,t)} \quad (1)$$

where $d_v(s,t)$ denotes the number of shortest paths between segment s and t through segment v and $d(s,t)$ represents the number of shortest paths between segments s and t.

(3) *Closeness* is the inverse of the total graph-theoretic distance of a given road segment from all other road segments in the entire planar graph network. It is defined as:

$$C_C(v) = \frac{1}{\sum_{v \neq t} d_{vt}} \tag{2}$$

where d_{vt} stands for the graph-theoretic distance of segment v and segment t.

(4) *PageRank* is an algorithm used by Google Search to rank websites in their search results [39]. When applying to the road network, a road segment is important if linked to many other segments. The PageRank value will be high and so is the value of the linked segment. PageRank satisfies the following equation:

$$p_i = q \sum_j a_{ji} \frac{p_j}{L(j)} + \frac{1-q}{N} \tag{3}$$

where a_{ji} represents the topological relation between the segment j and i. If segment j and i are intersected, a_{ji} is 1, or 0 otherwise. $L(j) = \sum_i a_{ji}$ is the number of segments intersected with the segment j. N is the number of segments in the road network. q is attenuation factor which is normally set as 0.85.

(5) *Clustering coefficient*is the measure of the clustering degree of a road segment. It is defined as the probability that two adjacent segments of a given segment are intersected. Given a segment v, $E = \{e_1, e_2, \cdots, e_n\}$ stands for the segments intersected with v. The clustering coefficient CC of the segment v can be denoted as follows:

$$CC(v) = \frac{\text{Num (actual segments)}}{\text{Num (possible segments)}} \tag{4}$$

where numerator and denominator represent the number of actual segments intersected with v and the number of possible segments intersected with v respectively.

Figure 3 illustrates the derivation of these centrality measures. Figure 3(a) is a simple sketch map of a road network of 10 road segments. The corresponding segment-based network model is displayed in Figure 3(b) in which nodes represent the road segments and edges the corresponding intersections. Refer to Figure 3(a), the indices of degree, betweenness, and closeness for road segment d are 6, 10.5 and 0.75 respectively. For road segment d, its six adjacent segments (a, b, c, e, f, g) would form $(6 \times 5) \div 2 = 15$ pairs of intersected segments, while there are only six pairs of intersected segments (i.e. a and c, a and f, c and f, b and e, b and g, e and g), so the clustering coefficient of segment d is $6 \div 15 = 0.4$.

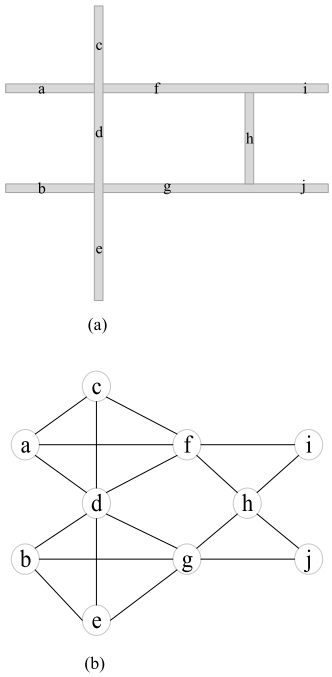


FIGURE 3. An illustrative example of the planar graph road network and its corresponding network model.

V. STUDY AREA AND AVAILABLE TRAFFIC DATA

Hong Kong is a mega city with highly developed urbanization. The total land area is about 1104 km^2 with a population of more than 7 million. It is comprised of Hong Kong Island, Kowloon, and the New Territories. Hong Kong Island with its diverse commercial and residential land uses is taken as a study case. The Island is also characterized by the uneven and non-symmetrical distribution of roads. A dense network of elongated-shaped major roads is found in the northern part stretching from west to east, separated from a few very winding major roads on the south by mountains and a tunnel in between. Topology obviously varies significantly from the coast (a denser network) to the central part (only a few links). Yet, traffic flow does not vary in the same pattern as commuting is always high between the commercial and residential areas around the Island. Such geographical setting does form a representative and challenging example to prove the validity of any estimation model compared with a more symmetrical grid pattern of the road network.

Three types of traffic data in Hong Kong are used to conduct linear regression analysis with each of the above-described topological and geometrical measure. They are the annual average daily traffic (AADT) data, GPS truck trajectory data, and public transport route data. These three types of data correspond to covering all types of vehicles, a portion of private vehicles (without fixed route and schedule) and all public transport fleets (with fixed routes and pre-planned schedule) respectively, and geographically correspond to data from a few major roads to almost all

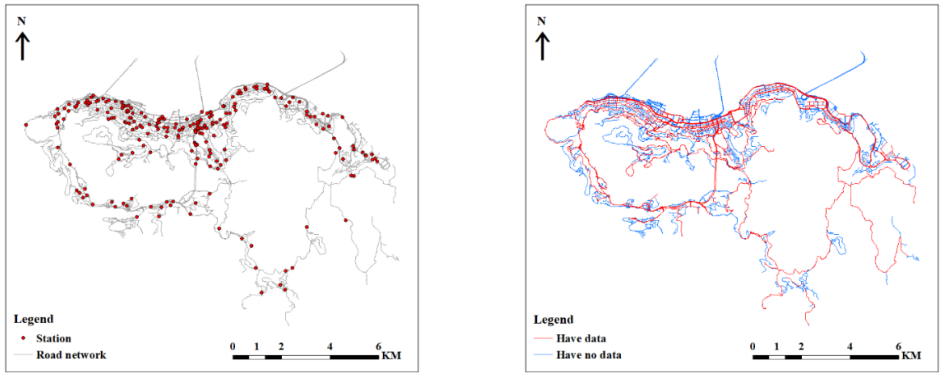


FIGURE 4. The spatial distribution of stations and AADT data in Hong Kong Island. (a) Counting stations. (b) AADT data.

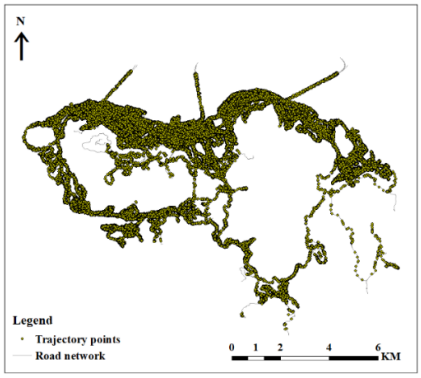


FIGURE 5. The spatial distribution of GPS-based truck trajectory data in Hong Kong Island.

road segments respectively. Each is described in detail as follows:

(1) *AADT* – In Hong Kong, there are a total of 853 counting stations, taking a census of about 87% of the 2,089 km of trafficable roads [29]. An Annual Average Daily Traffic (AADT) is derived from hourly, daily and monthly variations of traffic counts. Such information is useful to understand the relative traffic volume at different parts of the road network. In this study, data from a total of 216 counting stations (accounting for about 34% of traffic flow data on Hong Kong Island) has been used with its distribution shown in Figure 4.

(2) *GPS-based truck trajectory data* – this dataset is provided by GoGoVan, the largest online logistics company in Hong Kong. Data come from more than 1000 trucks for five workdays from July 3 to 7, 2017. In Figure 5, each dot represents a trajectory point. The trajectories cover almost all roads. This may reflect the relative flow of private car for each named road, based on a total of over 552,000 private cars in Hong Kong.

(3) *Public transport route data* – there are 506 public transport routes on the Island including bus, mini-bus, and tram. According to the departure frequency per day of 24-hour of each route, the traffic flow of each road segment traversed by public transport can be estimated.

VI. RESULTS AND ANALYSIS

A. TRAFFIC FLOW ESTIMATION BASED ON SINGLE MEASURE

Traffic flow estimation is first conducted based on a single measure respectively before applying the multiple linear regression model. Figure 6(a)-(c) show the flow per day derived from each data set. Since these are essentially different in coverage, volume and format, classification with the same values cannot be performed. Yet for easy comparison, they are all classified qualitatively with equal interval into three categories of high (red), medium (yellow) and low (green) flow according to their own data range. All three show a higher flow in the northern part of the Island, especially around the Central Business District. For truck trajectory data, only a few backbone roads exhibit higher traffic flow, whereas both public transport route data and AADT data have higher traffic flow extended furthermore on the two sides in the north as well as to the trunk roads and tunnel linking to the west and the south respectively. It is also noted that some main roads have a particularly high flow of public transport only as these are designated as bus-only-lane.

On the other hand, with the topological measures of each road segment defined using formula (1)-(4), the relationship with each data set using single variable regression analysis is performed with the results shown in Table 1. R values are the correlation coefficients between traffic flow and various measures. R^2 values are the goodness of fit for the single variable models. For both truck trajectory data and public transport data, the coefficients of closeness are low and inappropriate for estimating both types of traffic flow. It is conjectured that such a weak correlation is probably due to a neglect of the actual road length. This is reinforced by the high correlation coefficient of 0.61 and 0.48 of road length. According to the results of the t-test, the significance values are smaller than 0.05, indicating that the linear relationships between the independent variables and dependent variable are significant in the 5% significance level. A different picture occurs for AADT data. It is found that all the R^2 values are significant except for the clustering coefficient. This is probably

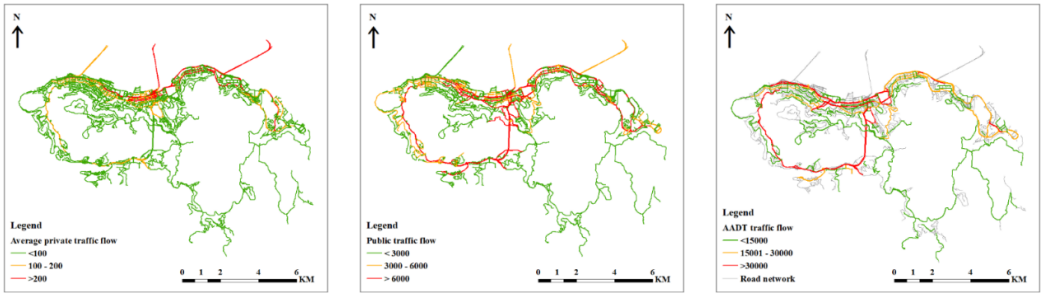


FIGURE 6. The spatial distribution of (a) private traffic flow, (b) public traffic flow, and (c) AADT traffic flow.

TABLE 1. Single regression results based on three types of traffic flow.

	Private traffic flow		Public traffic flow		AADT traffic flow	
Measures	R	R^2	R	R^2	R	R^2
Degree	0.69	0.48**	0.60	0.37**	0.48	0.23**
Betweenness	0.59	0.34**	0.46	0.21**	0.43	0.19**
Closeness	0.07	0.005**	0.08	0.006*	0.39	0.15**
PageRank	0.58	0.34**	0.53	0.28**	0.35	0.12**
Clustering coefficient	-0.32	0.10**	-0.22	0.05**	-0.08	0.007
Length	0.61	0.37**	0.48	0.23**	0.28	0.08**

* Coefficient significant at $p < 0.05$

** Coefficient significant at $p < 0.01$

due to the low coverage of counting stations. In addition, the coefficients of degree are highest for all types of data.

Previous studies have indicated that topological measures (e.g. centrality) can be used to predict traffic flow at the aggregate level [14], [40] but mainly focus on correlation analysis based on one type of traffic data, such as floating car data, or AADT data. In this study, traffic flow estimation based on three types of traffic flow information has been conducted separately and has proved that all six measures have different correlations with different datasets. For instance, closeness has low correlations with private and public traffic flow, while it has a high correlation with AADT data. On the contrary, the clustering coefficient can be used to estimate private and public traffic flow but is inappropriate for estimating AADT data. Therefore, there is a potential to improve estimation accuracy through selecting appropriate measures.

B. TRAFFIC FLOW ESTIMATION BASED ON MULTIPLE MEASURES

Based on the finding that topological and geometrical measures have different correlation coefficients with traffic flow, traffic flow estimation using these measures in multiple variable regression models is carried out. Both the multiple linear regression model and a random forest model are implemented, which are compared with support vector regression and artificial neural networks methods.

1) MULTIPLE LINEAR REGRESSION MODEL

Multiple linear regression analysis is conducted to explore the relationships between multiple measures of the road network and three types of traffic flow. Table 2 summarizes the results of the multiple linear regression analysis based on the five measures. The correlation coefficient R and R^2 of the model reach 0.81 and 0.65 respectively for the private traffic flow, which is higher than the measure of degree, bearing the maximum R and R^2 values (i.e. 0.69 and 0.48) among all measures in Table 1. It is concluded that multiple measures are more appropriate for estimating private traffic flow than that of a single measure. Coefficients with one asterisk and two asterisks mean that the corresponding independent variables are significant at the 5% and 1% significant levels in the regression model. For the public traffic flow, the correlation coefficient R and R^2 of the model reach 0.66 and 0.43 respectively, which are slightly greater than the maximum R and R^2 values (i.e. 0.60 and 0.37). It is found that only the degree and betweenness are significant. The correlation coefficient R and R^2 of the model for the AADT traffic flow reach 0.64 and 0.41 respectively, which are also higher than the maximum R and R^2 (i.e. 0.48 and 0.23). Therefore, it can be concluded that performance of estimating traffic flow is better by combining multiple

TABLE 2. Results of multiple linear regression based on three types of traffic flow.

Multiple linear regression models			
	Private traffic flow ($R = 0.81$ $R^2 = 0.65$)	Public traffic flow ($R = 0.66$ $R^2 = 0.43$)	AADT traffic flow ($R = 0.64$ $R^2 = 0.41$)
	Coefficients	Coefficients	Coefficients
Degree	1.893**	1.536**	1.915**
Betweenness	0.103**	0.049	0.072
Closeness	-0.045**	-0.005	0.017
PageRank	-1.574**	-1.090**	-1.521*
Clustering coefficient	-0.120**	-0.030	0.056
Length	0.274**	0.111*	-0.015

* Coefficient significant at $p < 0.05$

** Coefficient significant at $p < 0.01$

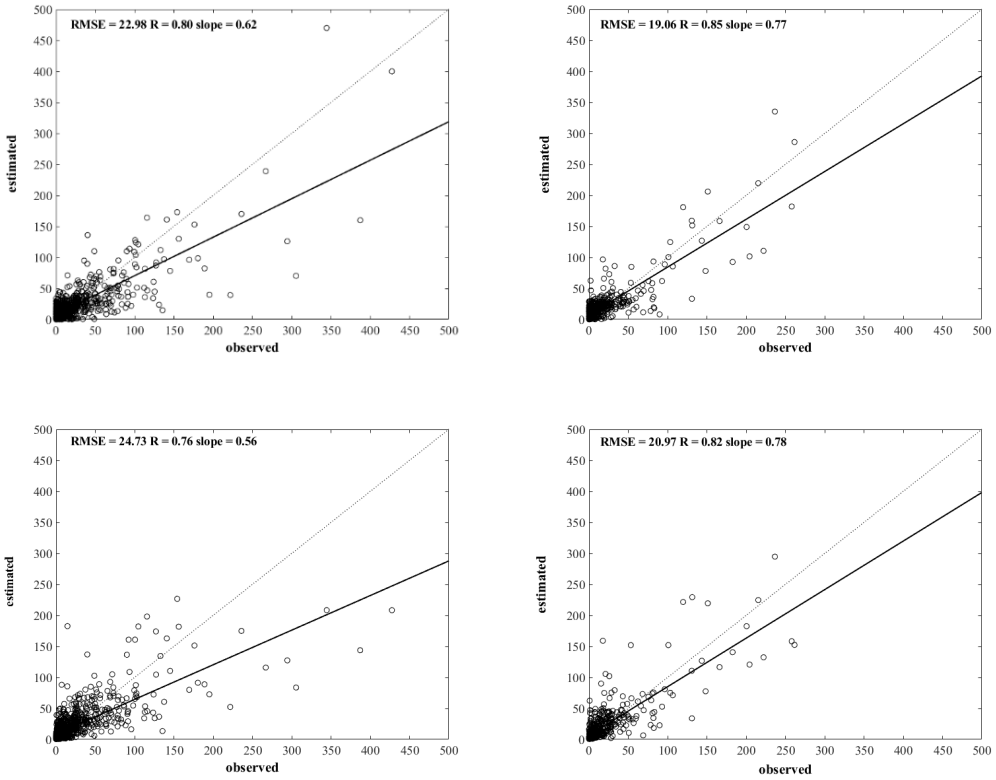


FIGURE 7. The relationships between observed traffic flow and estimated traffic flow based on trajectory data. (a) Training dataset using multiple linear regression, (b) test dataset using multiple linear regression, (c) training dataset using the random forest, (d) test dataset using the random forest.

road network properties than that of a single road network property.

2) RANDOM FOREST

In this section, random forest is conducted to explore the relationship between multiple measures of the road network and three types of traffic flow obtained from the aforesaid data. Both the topological and geometrical measures are selected as independent variables including degree, betweenness, closeness, PageRank, clustering coefficient and road length. The observed traffic flows from three types of traffic data are regarded as dependent variables respectively. The dataset is randomly divided into two parts, namely training dataset and testing dataset, representing two thirds and one-third of the

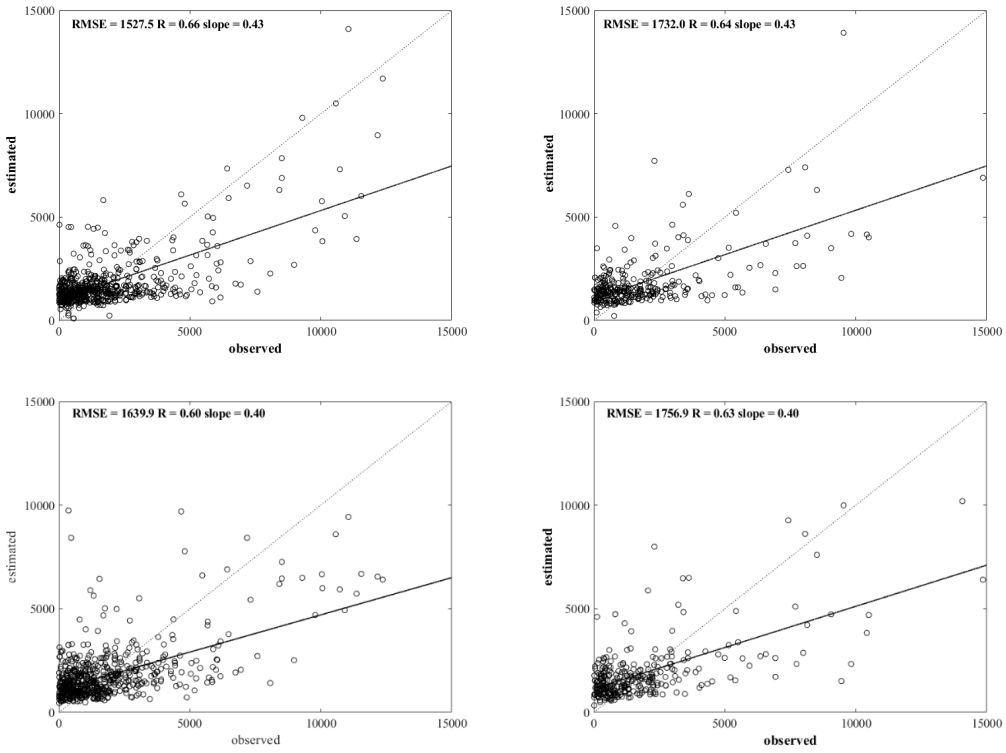


FIGURE 8. The relationships between observed traffic flow and estimated traffic flow based on route data. (a) Training dataset using multiple linear regression, (b) test dataset using multiple linear regression, (c) training dataset using the random forest, (d) test dataset using the random forest.

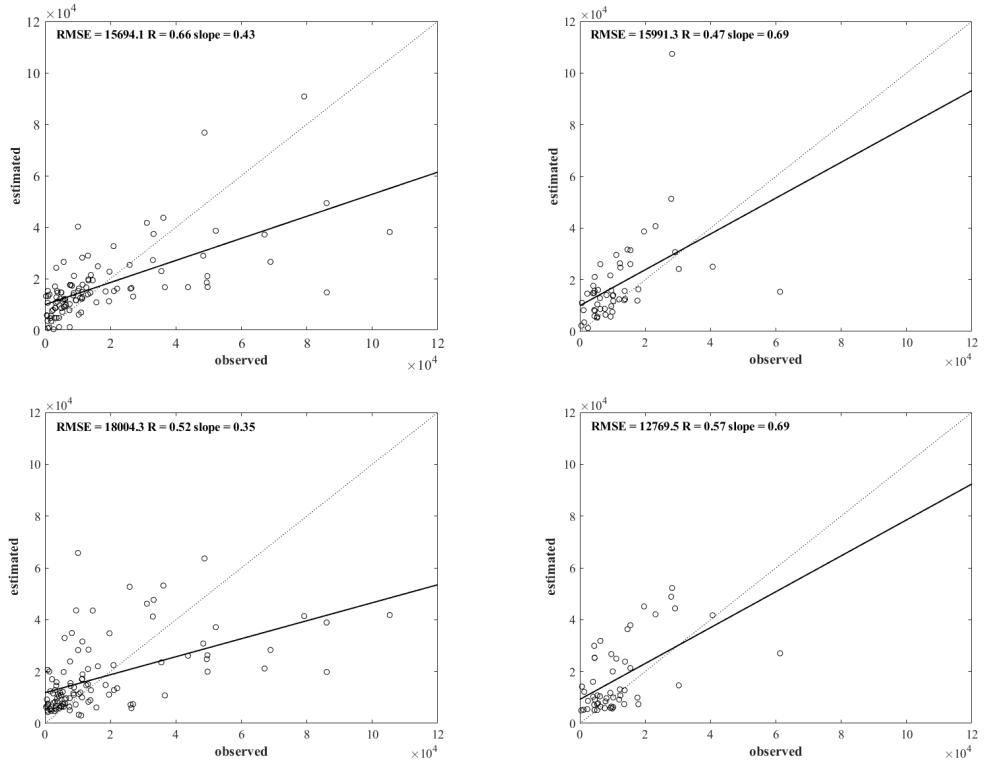


FIGURE 9. The relationships between observed traffic flow and estimated traffic flow based on AADT data. (a) Training dataset using multiple linear regression, (b) test dataset using multiple linear regression, (c) training dataset using the random forest, (d) test dataset using the random forest.

TABLE 3. Comparison results of different regression methods.

	Multiple Linear Regression			
	Training Dataset		Test Dataset	
	R	RMSE	R	RMSE
Private traffic flow (Figure 7)	0.80	22.98	0.85	19.06
Public traffic flow (Figure 8)	0.66	1527.5	0.64	1732.0
AADT traffic flow (Figure 9)	0.66	15694.1	0.47	15991.3
	Random Forest			
	Training Dataset		Test Dataset	
	R	RMSE	R	RMSE
Private traffic flow (Figure 7)	0.76	24.73	0.82	20.97
Public traffic flow (Figure 8)	0.60	1639.9	0.63	1756.9
AADT traffic flow (Figure 9)	0.52	18004.3	0.57	12769.5
	Support vector regression			
	Training Dataset		Test Dataset	
	R	RMSE	R	RMSE
Private traffic flow	0.54	37.4	0.68	33.5
Public traffic flow	0.56	2132.9	0.55	2354.8
AADT traffic flow	0.49	28944.9	0.65	20818.3
	Artificial neural networks			
	Training Dataset		Test Dataset	
	R	RMSE	R	RMSE
Private traffic flow	0.54	829.2	0.61	857.8
Public traffic flow	0.45	2746.3	0.48	2753.4
AADT traffic flow	0.38	28237.9	0.64	17052.4

total dataset respectively. Specifically, the training dataset is used to train the model whereas the testing dataset is used to verify the model. The number of trees determined is 100.

For the verification of the regression model, the correlation coefficient (R) between the observed traffic flow and estimated traffic flow is used to evaluate the goodness-of-fit of the model. The root mean square error (RMSE) is used to assess accuracy, which indicates the deviation between observed values and estimated values. In addition, the slope of the least square line of best fit is a measurement of how much estimated values deviate from the observed values. There is normally either a positive or negative association between the slope of the least squares line of best fit and correlation coefficient. Generally speaking, the performance of the method is evaluated by comparing the differences of R, RMSE and slope values in the estimated-versus-observed values plots. Higher R and lower RMSE values correspond to higher precision and accuracy of the method.

Figures 7 to 9 respectively display the estimation results from private fleet trajectory data, public transport data and AADT data based on random forest model, from which the relationships between observed traffic flow and estimated traffic flow based on training (left figure) and testing (right figure) datasets can be observed. For both training and testing results, the correlation coefficients R, RMSE and slope are reported.

Table 3 summarizes the comparison results of four regression models using the different types of datasets. It is readily observed that for all datasets no matter of varying road or vehicle coverages, RMSE of training data from Random Forest is higher than that of Multiple Linear Regression, with R from Multiple Linear Regression higher than that from Random Forest. Similar patterns occur also for test data of GPS-based truck trajectory data as well as public transport route and schedule data. It can be concluded that the relationship between traffic flow and the measures adopted is linear and that Multiple Linear Regression models are more appropriate for estimating traffic flow than the Random Forest Model. Compared with GPS-based truck trajectory data, public transport route and schedule data have a lower R and higher RMSE. This is probably due to its nature of fixed routes, resulting that some road segments have no traffic, and

may not be suitable for cities where residents drive more than taking public transport services. Yet in Hong Kong where more than 90% of people take public transport and is a place with dense network and high frequency of varying public transport modes, this dataset with its reliable and derivable traffic flow forms a good supplement to trajectory data from private fleets and taxi.

In addition, compared with the results of multiple linear regression and random forest, the results of support vector regression and artificial neural networks display lower R and higher RMSE for all the datasets except test data of AADT. It is demonstrated that multiple linear regression and random forest are more superior in the performance of traffic flow estimation. Moreover, the higher R values of non-linear multiple regression models for the test data of AADT are probably related to its insufficient sample size.

VII. CONCLUSION

This paper discusses several data collection methods to obtain and derive traffic flow information, using Hong Kong road network and some observed data as empirical cases. Owning to the discrepancy of data format and description of road information from various data sources, integration is sometimes difficult, especially for real-time performance. It examines the potential of combining both topological and geometrical properties of the road network to estimate traffic flow. The objective is to derive a more accurate approach to traffic flow estimation. This is an important input to not only transport planning and problem-solving but also provides a better methodology for navigation software engineering when real-time traffic flow may not be every-where and every time available. The contributions of this paper are mainly in the following two aspects.

First, three types of traffic data are used to estimate traffic flow, namely GPS-based truck trajectory data, public transport route data, and AADT data. Specifically, the relationships between traffic flow and topological and geometrical properties of the road network are analyzed. Through comparing and analyzing the results of these three types of traffic flow, it is found that topological and geometrical measures have different correlations with traffic flow for the three different types of traffic data.

Second, traffic flow estimation is enabled through integrating topological and geometrical properties using multiple regression models. Both topological and geometrical measures are regarded as independent variables, and traffic flow is taken as the dependent variable. By comparing the estimation results, it is found that a combination of topological and geometrical measures results in higher R. The proposed multiple regression approach is therefore more superior to estimating traffic flow based on mere road network properties.

Overall, this study proposes a feasible systematic methodology for estimating urban traffic flow using three types of traffic data. However, there are still several related research issues in the current study, which could be regarded as directions for future work. First, this study models the road network in an undirected graph without consideration of traffic constraints, such as turns and direction. These are more complex and dynamic data but have a significant influence on urban traffic flow. Second, given that this paper only uses Hong Kong as the case study, the method should be verified using the traffic data of other cities.

Acknowledgement

The authors would like to thank GoGoVan and Transport Department, HKSAR Government for supplying data to conduct this research work.

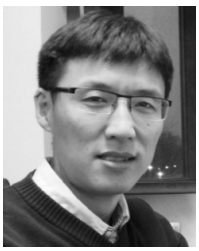
REFERENCES

[1] M. Anderson, K. Sharfi, and S. Gholston, "Direct demand forecasting model for small urban communities using multiple linear regression," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1981, no. 1, pp. 114–117, 2006.
[2] X. Chang, B. Y. Chen, Q. Li, X. Cui, L. Tang, and C. Liu, "Estimating real-time traffic carbon dioxide emissions based on intelligent transportation system technologies," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 469–479, Mar. 2013.
[3] B. Y. Chen, H. Yuan, Q. Li, W. H. K. Lam, S.-L. Shaw, and K. Yan, "Map-matching algorithm for large-scale low-frequency floating car data," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 1, pp. 22–38, 2014.
[4] C. Cheng, "Optimal sampling for traffic volume estimation," Ph.D. dissertation, Univ. Minnesota, Minneapolis, MN, USA, 1992.
[5] M. Doustmohammadi and M. Anderson, "Developing direct demand AADT forecasting models for small and medium sized urban communities," *Int. J. Traffic Transp. Eng.*, vol. 5, no. 2, pp. 27–31, 2016.
[6] S. Gao, Y. Wang, Y. Gao, and Y. Liu, "Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality," *Environ. Planning B, Urban Anal. City Sci.*, vol. 40, no. 1, pp. 135–153, 2013.
[7] M. Gastaldi, G. Gecchele, and R. Rossi, "Estimation of annual average daily traffic from one-week traffic counts. A combined ANN-Fuzzy approach," *Transp. Res. C, Emerg. Technol.*, vol. 47, pp. 86–99, Oct. 2014.
[8] B. Hillier, A. Penn, J. Hanson, T. Grajewski, and J. Xu, "Natural movement: Or, configuration and attraction in urban pedestrian movement," *Environ. Planning B, Planning Des.*, vol. 20, no. 1, pp. 29–66, 1993.
[9] Y. Hou, P. Edara, and C. Sun, "Traffic flow forecasting for urban work zones," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1761–1770, Aug. 2015.
[10] T. Idé, T. Katsuki, T. Morimura, and R. Morris, "City-wide traffic flow estimation from a limited number of low-quality cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 950–959, Apr. 2017.
[11] A. Jayasinghe, K. Sano, and H. Nishiuchi, "Explaining traffic flow patterns using centrality measures," *Int. J. Traffic Transp. Eng.*, vol. 5, no. 2, pp. 134–149, 2015.
[12] B. Jiang and C. Claramunt, "Topological analysis of urban street networks," *Environ. Planning B, Urban Anal. City Sci.*, vol. 31, no. 1, pp. 151–162, 2004.
[13] B. Jiang, S. Zhao, and J. Yin, "Self-organized natural roads for predicting traffic flow: A sensitivity study," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 7, 2008, Art. no. P07008.
[14] B. Jiang and C. Liu, "Street-based topological representations and analyses for predicting traffic flow in GIS," *Int. J. Geograph. Inf. Sci.*, vol. 23, no. 9, pp. 1119–1137, 2009.
[15] B. Jiang, X. Liu, and T. Jia, "Scaling of geographic space as a universal rule for map generalization," *Ann. Assoc. Amer. Geograph.*, vol. 103, no. 4, pp. 844–855, 2013.
[16] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, Jun. 2011.
[17] A. Kazerani and S. Winter, "Can betweenness centrality explain traffic flow," in *Proc. 12th AGILE Int. Conf. Geograph. Inf. Sci.*, 2009, pp. 1–9.
[18] W. H. K. Lam, Y. F. Tang, K. S. Chan, and M. L. Tam, "Short-term hourly traffic forecasts using Hong Kong annual traffic census," *Transportation*, vol. 33, no. 3, pp. 291–310, 2006.
[19] G. Leduc, "Road traffic data: Collection methods and applications," *Work. Papers Energy, Transport Climate Change*, vol. 1, pp. 1–55, Nov. 2008.

[20] G. Leshem and Y. Ritov, "Traffic flow prediction using Adaboost algorithm with random forests as a weak learner," in *Proc. World Acad. Sci., Eng. Technol.*, vol. 19, pp. 193–198, Jan. 2007.
[21] I. X. Leung, S. Y. Chan, P. Hui, and P. Lio. (2011). "Intra-city urban network and traffic flow analysis from GPS mobility traces." [Online]. Available: https://arxiv.org/abs/1105.5839
[22] Q. Li, T. Zhang, and Y. Yu, "Using cloud computing to process intensive floating car data for urban traffic surveillance," *Int. J. Geograph. Inf. Sci.*, vol. 25, no. 8, pp. 1303–1322, 2011.
[23] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[24] M. Lowry and M. Dixon, "GIS tools to estimate average annual daily traffic," National Institute for Advanced Transportation Technology, Bengaluru, Karnataka, Tech. Rep. KLK725, 2012.
[25] D. W. Morley and J. Gulliver, "Methods to improve traffic flow and noise exposure estimation on minor roads," *Environ. Pollut.*, vol. 216, pp. 746–754, Sep. 2016.
[26] R Development Core Team. "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, Tech. Rep., 2009.
[27] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2340–2350, Sep. 2017.
[28] Y. Tang, W. H. Lam, and P. L. Ng, "Comparison of four modeling techniques for short-term AADT forecasting in Hong Kong," *J. Transp. Eng.*, vol. 129, no. 3, pp. 271–277, May 2003.
[29] Traffic and Transport Survey Division, Transport Department of the Government of the Hong Kong SAR. The Annual Traffic Census Tech. Rep., 2013.
[30] D. Xia, H. Li, B. Wang, Y. Li, and Z. Zhang, "A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction," *IEEE Access*, vol. 4, pp. 2920–2934, 2016.
[31] L. Xia and Y. Shao, "Modelling of traffic flow and air pollution emission with application to Hong Kong Island," *Environ. Model. Softw.*, vol. 20, no. 9, pp. 1175–1188, 2005.
[32] Q. Xia, F. Zhao, Z. Chen, L. Shen, and D. Ospina, "Estimation of annual average daily traffic for nonstate roads in a Florida county," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1, no. 1660, pp. 32–40, 1999.
[33] D. Xu and Y. Shi, "A combined model of random forest and multilayer perceptron to forecast expressway traffic flow," in *Proc. IEEE Int. Conf. Electron. Inf. Emergency Commun.*, Jul. 2017, pp. 448–451.
[34] Y. Yan, S. Zhang, J. Tang, and X. Wang, "Understanding characteristics in multivariate traffic flow time series from complex network structure," *Phys. A, Stat. Mech. Appl.*, vol. 477, pp. 149–160, Jul. 2017.
[35] P. Ye, B. Wu, and W. Fan, "Modified betweenness-based measure for prediction of traffic flow on urban roads," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2563, pp. 144–150, Jan. 2016.
[36] Y. Zhang and Y. Liu, "Traffic forecasting using least squares support vector machines," *Transportmetrica*, vol. 5, no. 3, pp. 193–213, 2009.
[37] F. Zhao and S. Chung, "Estimation of annual average daily traffic in a Florida county using GIS and regression," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1769, pp. 113–122, Jan. 2001.
[38] N. Zhao, L. Yu, H. Zhao, J. Guo, and H. Wen, "Analysis of traffic flow characteristics on ring road expressways in Beijing: Using floating car data and remote traffic microwave sensor data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2124, pp. 178–185, Jan. 2009.
[39] P. Zhao, T. Jia, K. Qin, J. Shan, and C. Jiao, "Statistical analysis on the evolution of OpenStreetMap road networks in Beijing," *Phys. A, Stat. Mech. Appl.*, vol. 420, pp. 59–72, Feb. 2015.
[40] S. Zhao, P. Zhao, and Y. Cui, "A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China," *Phys. A, Stat. Mech. Appl.*, vol. 478, pp. 143–157, Jul. 2017.
[41] Y. Zou, J. E. Ash, B.-J. Park, D. Lord, and L. Wu, "Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety," *J. Appl. Statist.*, vol. 45, no. 9, pp. 1652–1669, 2018.

LILIAN PUN is currently an Associate Professor and an Associate Head of the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. Her teachings and research are in the areas of cartography, and GIS data modeling and application, especially in GIS-T and spatial data standards. She is actively engaged in public transport information systems, especially in the areas of data management, web mapping, and optimal route computation. She is also a Council Member of the Intelligent Transport Systems in Hong Kong, a member of the Hong Kong GIS Association, the Hong Kong Taxi Service Committee, and the HKDSE Geography Committee.

PENGXIANG ZHAO received the M.S. and Ph.D. degrees from Wuhan University, in 2012 and 2015, respectively. He is currently a Postdoctoral Researcher in the Institute of Cartography and Geoinformation, ETH Zürich. His research interests include geographic information science, urban mobility, and trajectory data analysis and mining.



XINTAO LIU received the B.Eng. degree in survey from Hohai University, China, in 1998, the M.Sc. degree in cartography and GIS from Nanjing Normal University, China, in 2003, and the Ph.D. degree in geoinformatics from the Royal Institute of Technology, Sweden, in 2012. In 2012, he joined the Department of Civil Engineering, Ryerson University, Canada, where he was a Postdoctoral Fellow in GIS and transportation until 2016. He was a Sessional Lecturer with Ryerson University, since 2015. He is currently an Assistant Professor with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. He is also a PI and a Co-PI of several national projects funded by Sweden, Canada, and Hong Kong. His research interests include GI services and science, urban computing, and GIS in transportation. His research goal is to use the state-of-the-art technologies to advance smart city for a better urban life. He received the Ph.D. Scholarship from Lars Erik Lundbergs. He is a Reviewer of a series of major international journals, such as IJGIS and AAG in his field.