Integrating Household Travel Survey and Social Media Data to Improve the Quality of OD Matrix: A Comparative Case Study

Zesheng Cheng, Sisi Jian, Taha Hossein Rashidi, Mojtaba Maghrebi, and Steven Travis Waller

***Abstract*— Collecting effective data is a fundamental step in developing transport networks and related research. Social media have become an emerging source of data for traffic analyses. In this paper, we demonstrate that the function of a city influences the utility of social media data in travel demand models by generating models for eight US cities with different functions. Data from Twitter and Foursquare, as well as other socio-demographic information, are considered as independent variables in Origin-Destination trip regression models generated via a Random Forest regression technique. Model performance with and without use of social media data are compared via 10-fold cross-validation. The results indicate that the accuracy of the models for all eight cities improved when independent variables based on social media data were included. The performance was most improved in metropolitan areas, followed by rural and tourist areas. Inspired by this finding, we conclude that the city function influences the utility of social media data in travel demand models. Meanwhile, we create models based on trip purpose and transport mode to explore other factors that may impact the efficiency of applying social media data in transport research.**

***Index Terms*— Twitter, Foursquare, random forest regression, travel demand estimation, multi-city model.**

I. Introduction

COLLECTING effective transport data is the first step in developing traffic networks and related research. Due to improvements in techniques, there is an increasing number of available data sources [1]. However, it usually takes traditional data sources, such as a household travel survey (HTS), and a large budget, labour force and time period to collect transport data. In order to address this problem, novel data sources such as social media [2], smart card tracking systems [3] and taxi trajectories [4], [5] have been considered by researchers. These may provide additional information that helps researchers to develop new ideas in transport research, including prediction of people's locations [6]–[8], their individual behaviours [9] and mobility patterns [10], [11].

Manuscript received December 5, 2018; revised June 16, 2019 and November 11, 2019; accepted November 18, 2019. Date of current version May 29, 2020. The Associate Editor for this article was S. C. Wong. *(Corresponding author: Taha Hossein Rashidi.)*

Z. Cheng, T. H. Rashidi, M. Maghrebi, and S. T. Waller are with the School of Civil and Environmental Engineering, University of New South Wales (UNSW), Sydney, NSW 2052, Australia (e-mail: rashidi@unsw.edu.au).

S. Jian is with the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong.

Digital Object Identifier 10.1109/TITS.2019.2958673

Compared to traditional data acquisition methods, use of novel data sources can save both time and money [12]. However, some of them, such as taxi trajectories, are not easily accessible to most researchers. Among these new data sources, social media are considered to have relatively easy access via online applications and have large user coverage. Especially in the past decade, user groups have expanded dramatically. According to statistics, the number of active Twitter user accounts was greater than 336 million at the end of quarter 1, 2018 [13]. Meanwhile, the information posted is increasingly rich and may include users' real-time positions, timelines and publicly-registered information. Accordingly, social media data mining has been applied to various areas of transport research. In the field of mobility pattern analysis, for instance, Hasan et al. presented an appropriate method for extracting large-scale data from social media to predict users' missing activity patterns in the timeline of social media data. This method tries to solve a major problem of social media-based data: the activity patterns provided are often disaggregated [14]. In addition, Huang et al. declared that Twitter is a useful data source for activity pattern analysis. By combining Twitter data with American Community Survey data, study [15] introduced an approach to predicting users' home and work locations. Liu et al. studied the mobility patterns of local and visiting Twitter users and reported that short-distance movement comprised the majority of the activity of both of them [16]. Other studies that have combined social media and transport data include studies on users' activity spaces [17], tourist destination and accommodation choices [18]–[20] and other individual behaviours [21]–[23].

As well as facilitating mobility pattern analysis, social media data can provide additional information for the creation of travel demand models. Hence, social media are potential data sources for transport network planning [24]. In 2014, Gao et al. provided a method for extracting trips from collected *Tweets* (messages posted on the Twitter online platform) based on their posting times and locations. The study reported that if a user posts *Tweets* at different locations within four hours, an origin-destination (OD) trip can be inferred. The model has been tested in the Greater Los Angeles Area [25]. This research provided a valid method for using OD trips extracted from social media data for transport research. A more recent study proposed an approach to applying Twitter data to the validation of a travel demand model. The author applied

latent class analysis and a Tobit regression model, which is a linear parametric-based model [26], to estimate travel demand among different sub-regions in Los Angeles based on Twitter data and other socio-demographic variables. The study presented an appropriate model for converting an OD matrix extracted from Twitter data into an official statistical matrix [6]. Using a similar approach, Lee et al. compared OD matrixes generated by social media data and the California State-wide Travel Demand Model (CSRDM). The study found four classes of relationship between CSTDM ODs and Twitter-based ODs, which represent four different types of trips made in California [27].

Based on a review of relevant studies, there are two main research gaps remaining. 1) In travel demand models based on social media data, most variables extracted from socio-demographic or social media information are not linearly related to OD travel demand. Therefore, to model demand more accurately, a non-linear or non-parametric regression technique may be required. In fact, several non-parametric regression models based on machine learning techniques have recently been applied to travel demand models. For example, Djukic discussed the use of dimensionality reduction and principal component analysis on real OD demand estimation. They defined a new transformed variable called *demand principal components* and demonstrated an improvement in OD estimation accuracy [28]. In addition, Cheng proposed a new model for OD matrix estimation based on a random forest (RF) algorithm and validated it with HTS statistics [29]. Due to the fact that OD trip data extracted from social media are strongly correlated with HTS OD travel demand [25], it has been suggested that social media data and machine learning regression techniques be combined to improve model accuracy. 2) It is worth noting that most of the research mentioned above has focused on a single city, such as Greater Los Angeles [25] or New York City [30]. However, the utility of social media data may also be influenced by the functions of the target city. Comparing the performance of models in regions with different city functions could help increase the efficiency of social media data utilization in travel demand models.

In light of these research gaps, the objective of this study was to determine whether the city's function influences the utility of social media data in travel demand models. In this paper, eight US cities or regions with three different functions— metropolis, rural or tourist—were selected as target regions. Travel demand models were built in those cities based on a random forest (RF) technique. The model contains variables for OD trips extracted from Twitter, *check-ins* to locations collected from Foursquare, and other socio-demographic information. Comparison of the performance of regression models that do or do not contain social media data for different cities demonstrates that social media data generally improve travel demand estimates. The models work best for metropolitan cities, followed by rural and tourism areas. This paper also tries to determine why the worst model improvement was for tourism areas. The influences of trip purpose and transport mode on the models are also discussed.

The paper contains six sections. Section II introduces the data used in the regression model, while Section III discusses the methodology. The results are presented in Section IV and discussed in Section V. Finally, Section VI concludes by presenting the key findings and highlighting further research directions.

II. Data Description

In this study, we use social media as a data source and apply travel demand regression models to eight cities with different functions. The subsections below provide a detailed description of the data applied in the model, which contain 1) descriptions of target regions; 2) data from HTS and relevant travel demands; 3) information extracted from social media data, including Twitter and Foursquare; and 4) socio-demographic data and other land-use data.

A. Description of Target Regions

Eight US target cities or regions with different city functions were used: Atlanta, Georgia; Baltimore City, Maryland; Chicago, Illinois; Seattle, Washington; the Champaign-Urbana-Savoy (CUS) urban agglomeration in Illinois, Idaho; Daytona Beach, Florida; and the Northeast (NE) Florida urban agglomeration. These cities have different functions and industry mixes of different proportions. Compared with other regions in the world, there are two advantages to selecting target cities from the USA. 1) High data quantity and quality: The majority of US states have detailed and accurate statistics available. The origins and destinations of trips collected from HTSs, as well as socio-demographic data, could be detailed collected in block or TAZs level. Meanwhile, there are high numbers of active social media users and posted tweets. 2) Most relevant research has been done in the US, which provides a useful benchmark. The eight target regions can be divided into three groups according to their functions. 1) Metropolis group: Compared with rural and tourist regions, metropolitan cities always have larger areas and populations. Meanwhile, they may have different central functions and better foundations for residents' survival and development. This group includes Atlanta (US Metro ranking 9 [31]), an important industrial center and key transportation hub; Baltimore (US Metro ranking 22 [31]), the largest independent city and one of the US's major harbors; Chicago (US Metro ranking 3 [31]), a centre of economics, trade, light and heavy industry and culture; and Seattle (US Metro ranking 15 [31]), a computer science development center. 2) Rural group: including the CUS region, one of the major rural agglomeration areas in Illinois; and Idaho, the largest rural state in America. 3) Tourist group: including Daytona Beach and Northeast Florida. In rural and tourist areas, a major part of the economy is involved in the tourist industry [32], [33]. Rural and tourist areas have relatively small populations and areas, and rarely provide the functionality of a city. Accordingly, they usually form urban agglomerations with other nearby cities that have similar functions.

B. Dependent Regression Variable: HTS-Based Travel Demand

The Household Travel Survey (HTS) is published by the Metropolitan Travel Survey Archive [34]. It is a public website

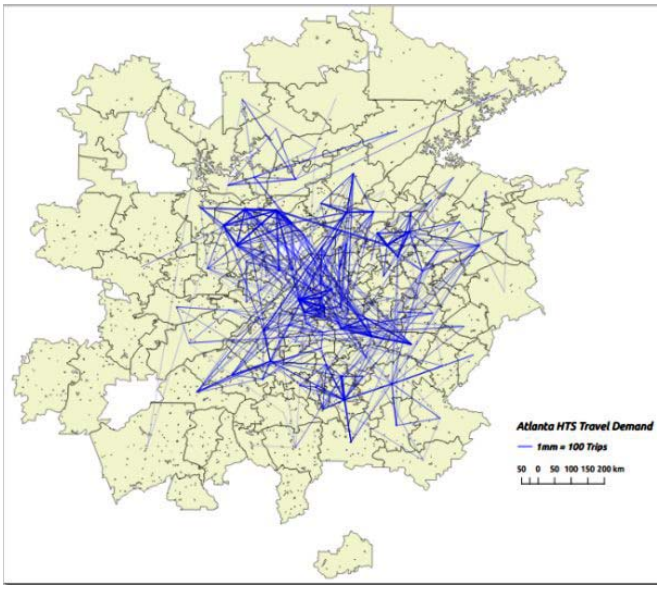

Fig. 1. Atlanta travel demand based on HTS data.

providing statistics on local residents' trips together with other information such as personal information, transport methods and travel purposes. For each target region, the data were last updated in around 2010. In order to generate an OD travel demand matrix, each target city was divided into subareas according to the ZIP Code Tabulation Areas (ZCTAs) system. Then, the origins and destinations of HTS trips were aggregated into different regions to create an OD matrix. Only links with more than ten trips were considered in the model. This helps reduce error under the premise of considering as many trips as possible. Then, the OD matrix was reshaped to an n-by-1 vector. This vector is dependent variable set of the regression models, which is $\mathbf{X^{(1)}} \ldots \mathbf{X^{(n)}}$ in Equation (1), Section III. Taking Atlanta as an example, firstly, based on the ZCTAs, the city was divided into 142 regions. Then, the origin and destination of trips extracted from HTS data were aggregated to each region to create an OD travel demand matrix with 20,164 (142×142) links. After selecting links with more than ten trips and matrix reshaping, an 847-by-1 vector was created and used as the dependent variable of the Atlanta regression model. Figure 1 illustrates the ZCTA regional division of Atlanta and the 847 links considered in the model. In the figure, thicker lines link OD pairs with higher travel demand based on HTS statistics.

Table I in Section II.C details the dependent variable information used in each city's regression model. It can be seen that although only links with more than ten trips were considered, at least 70% of trips from the HTS data were considered in each city's regression model.

C. Independent Regression Variables: Data Extracted From Social Media

Twitter data were collected via the Twitter REST application programming interface (REST API). APIs are public platforms that allow developers to access features or data for Twitter and its related applications [35]. REST APIs can collect all *tweets* posted within a specific area and period. The memory limit is 10 days and the download limit is 600 *tweets* per minute [36].

TABLE I
ESSENTIAL INFORMATION OF TARGET CITIES

	Atlanta	Baltimore	Chicago	CUS
ZCTA No.	142	33	183	25
Links	20,164	1,089	33,489	625
Links 10+ Trips	847	119	607	30
Total Trips	39,407	6,695	12,416	3,207
Trips Considered	26,652	5,656	9,463	3,073
Tweets	4,432,841	1,633,867	2,355,971	149,076
Users	520,315	62,233	156,240	48,037
Trips	90,006	151,256	209,764	25,909
Trips Considered	40,601	119,568	49,084	21,295
	Daytona Beach	**Idaho**	**NE Florida**	**Seattle**
ZCTA No.	73	67	35	146
Links	5,329	4,489	1,225	21,316
Links 10+ Trips	109	231	261	998
Total Trips	7,011	23,859	20,197	61,602
Trips Considered	5,984	22,827	14,663	51,259
Tweets	813,869	622,934	166,688	2,048,574
Users	93,698	40,825	55,491	238,733
Trips	62,598	20,016	43,015	207,336
Trips Considered	44,663	15,226	19,196	149,310

REST APIs were used to collect tweets posted in each ZCTA suburb of the target cities. Then, tweets with geotagged information were selected and OD trips were extracted based on the algorithm mentioned above [25]. When two tweets were posted by a single user from different suburbs within a certain time period (4 hours), it was regarded as a single trip. Overly-frequent posts by a user with the same location or trajectory were considered to originate from social robots and were excluded. Table I details the dependent variable and travel demand information extracted from Twitter for each target city.

Foursquare data were also used. Check-ins to seven types of venue—Entertainment, School/College, Food, Nightlife, Outdoor/Sports, Professional and Shopping—in different ZCTA regions were collected by a third-party API. The API provided the top-50 registered landmarks with the most check-ins in the search area of a selected venue type. The sum of check-in numbers, to some extent, represents the function and land use of the target region. For instance, if the number of check-ins to Shopping venues is dramatically higher than that of other venues in a specific suburb, that suburb was considered to be a shopping centre. Landmarks registered on Foursquare have geolocations attached. By providing the centroid and range of a given suburb, the API provides the number of check-ins at a selected venue within that range. There were seven venue types mentioned above. Therefore, for each link, Foursquare provided 14 variables with the total number of check-ins for different venue types. For these 14 variables, seven described the origin of the link while seven described the destination.

D. Other Independent Variables

Besides the independent variables extracted from social media, ten more variables were applied in the regression: area, population and its density, housing number and the density of origin and destination suburbs respectively. These variables were collected from a census and socio-demographic database [37] and related calculations. On the one hand, these variables are the most common choices for use in travel

demand models. On the other hand, the eight target regions belonged to different US states, which have different statistical systems. These variables were provided by Proximity One and the Census 2010 ZIP code Demographic Profile Dataset. Another two variables used were the straight-line distance (km) between the centroid of origin and destination suburbs, and a binary variable named 'same OD mark'. For this binary mark, 1 indicates the link has the same origin and destination, while 0 indicates they are different.

III. Methodology

A. Random Forest Regression

The random forest (RF) algorithm is a highly flexible machine learning technique that can be applied to both regression and classification tasks [38]. It is claimed to be "unexcelled in accuracy among current algorithms" [39]. It can estimate the significance or correlation of each independent variable automatically. Meanwhile, due to the randomly-selected nature of the training samples and features, the probability of over-fitting is relatively low. Cheng's research reported that, compared with linear regression, regression tree algorithms and simple neural networks, random forest has better performance when applied to travel demand models [29]. The RF approach can give a model greater explanatory power as there is a direct indicator of which independent variables are more important in the regression process. Therefore, RF regression was used as the basic regression model in this paper.

The learning unit of RF is called a *classification and regression tree* (CART). The basic idea of a CART algorithm is to divide a given space into a set of rectangular areas and then fit the points in each area to a constant or to a simpler model. The most common CART algorithm is called a *binary tree*, which divides each area into two subareas recursively and decides the output for each one. Mathematically, for a given training data set D, we have [40]:

$$D = \left\{ \left(\boldsymbol{x^{(1)}}, y^{(1)} \right), \left(\boldsymbol{x^{(2)}}, y^{(2)} \right), \ldots, \left(\boldsymbol{x^{(m)}}, y^{(m)} \right) \right\} \quad (1)$$

where $\boldsymbol{x^{(1)}} \ldots \boldsymbol{x^{(m)}}$ is a vector containing dependent variables for sample 1 to sample m, and $y^{(1)} \ldots y^{(m)}$ is an independent variable for sample 1 to sample m.

After training, the space is divided into J different subareas. For a given testing sample n, the output of a regression tree could be expressed as [40]:

$$m\left(\boldsymbol{x^{(n)}}\right) = \sum_{j=1}^{J} v_j * I(x \in R_j) \quad (2)$$

where:

$\boldsymbol{x^{(n)}}$ = a vector containing the dependent variables of a given testing sample n;

J = the total amount of subareas;

j = an index of each subarea;

v_j = regression output of subarea j;

$I(.)$ = an indicator function returning 1 if its argument is true or 0 otherwise; and

R_j = subarea j, where $\bigcup_{j=1}^{J} R_j = 1$, $\bigcap_{j=1}^{J} R_j = \emptyset$.

To create a binary regression tree, one algorithm involves choosing an optimized split variable and its split value, then dividing one space into two subareas recursively. After repeating the steps for each subarea to meet a stopping criterion, for instance, an error threshold, a regression tree will be generated [41].

RF regression is kind of ensemble learning technique which uses a bagging algorithm to integrate several regression trees [42]. These regression trees are independent of each other and the estimates of the forest are determined by their voting and mode. The training algorithm can be described as:

1) For a provided training set with N samples and M features, each regression tree selects N samples randomly. The same sample could be selected repeatedly, which is called a *bootstrap sample method* [43].
2) Train each regression tree with m randomly selected features where $m < M$. Repeat the step of creating CART until each regression tree meets the requirements.
3) For a given test input, estimate its output with each regression tree and vote to determine the final results, which is called a *bagging process*.

The importance of the variables in the regression can be tested following the steps below. This method is called *out-of-bag estimation of feature importance* [42].

1) Compute the regression root mean squared error (RMSE) for the given regression forest.
2) Permute the values for the selected variables, train and test the model again to calculate its new RMSE.
3) Repeat steps 1) and 2) several times to reduce bias. The average difference between the old and new RMSEs can reflect the importance of the variables. The higher the value, the more important the variable is.

B. K-Fold Cross-Validation

K-fold cross-validation is a model testing technique for examining the performance of the model using data collected iteratively. Theoretically, during the process, the primary database A is randomly divided into k packages of equal size. Each package contains M/K samples (round down). One of the packages will be selected for testing and the rest of the samples are used as training data. The cross-validation process contains k iterations until each package has been used as testing data exactly once [44]. K-fold cross-validation is an appropriate method for model testing, especially in cases of insufficient data [45]. For our regression model, to make full use of the collected Twitter data, 10-fold cross-validation was applied. For one testing fold, the regression residuals as well as the RMSE will be reported, which are important standards for evaluating our regression model.

IV. Results

A. Model

By applying the variables mentioned in Section III above, two regression models based on an RF algorithm were generated for each target city. The first one contained distance, 'same OD mark', area, population and its density, housing number and housing density. The second model considered

TABLE II

REGRESSION RESULTS FOR EIGHT TARGET CITIES

City	Function	Trip mean	Trip standard deviation	RMSE without social media data	RMSE with social media data	RMSE improvement
Atlanta	Metropolis	31.46	40.61	31.1	25.8	17.2%
Baltimore		47.53	92.68	29.1	26.7	8.2%
Chicago		15.59	15.52	7.1	6.2	12.7%
Seattle		51.36	92.0	50.2	44.1	12.1%
CUS	Rural	102.4	141.2	84.7	81	4.3%
Idaho		98.81	202.9	85.0	81.5	4.1%
Daytona Beach	Tourist	54.89	71.23	51.2	50.6	1.1%
NE Florida		56.18	248.2	68.8	68.2	0.8%

social media data, which contains all the mentioned variables. All the RF regression models contained 1000 binary regression trees and each of them was trained by a maximum of six features. Then, 10-fold cross-validation was applied to test the performance of the models.

B. Regression Result

Table II indicates the regression results of the models with and without data from social media. In this paper, the decrease in the regression RMSE, which determines the differences between predicted and ground truth values, was selected as a standard to measure performance improvements.

As shown in the table, all of the regression models were improved by applying social media data but to different degrees. The travel demand models were most improved in metropolitan areas ($>8\%$), followed by rural areas (around 4%) and tourist area (around 1%).

Table III illustrates the importance of each variable in the model using the out-of-bag estimates mentioned in Section III.A. In the table, Model 1 is the logogram of the model without social media data while Model 2 is the model with social media data. The variables with higher out-of-bag importance values are more significant in the regression. The table shows that for most models, link variables, inner marker and distance were more important than other variables. Moreover, social media data played a more important role in models of metropolitan areas, while distance was more important for the other target regions.

C. Suburb-Based Analysis

This section presents the reason why the improvement was better in metropolitan areas and worst target tourist areas. Figure 2 is a series of graphs that illustrate the problem by taking Atlanta, Daytona Beach and NE Florida as an example.

Basically, the main cause of the problem is that social media and HTS data do not match well in the target tourist areas. A similar conclusion can be drawn from Table III. As shown in Figs. 2(a), 2(c), and 2(e), the distribution of trips according to HTS data was distributed relatively evenly around the city and more likely to aggregate towards the city centre. However, unlike Atlanta, for which trips extracted from Twitter had similar distribution as HTS data (as shown in Fig. 2(b)), travel demand extracted from Twitter was mostly aggregated to several specific links in the target tourist areas in Figs. 2(d) and 2(f). Although the origins and destinations of those links are popular zones in the tourist area, such as the beach, famous attractions, and airport or nearby transport stations, social media data are unable to provide enough help to other links in the transport network. Therefore, the improvements in travel demand models in tourist areas were the worst among the target areas Moreover, most HTS data were collected from local residents. However, in tourist areas, a large percentage of *tweets* were posted by tourists. Considering this condition of mismatch, predictably, the mentioned problem may influence the utility of social media data in other tourist areas and in related research. Therefore, suitable data

TABLE III

VARIABLE OUT-OF-BAG IMPORTANCE ESTIMATION (% INCREASE IN RMSE)

City	Atlanta		Baltimore		Chicago		Seattle	
Model #	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Link-related variables								
Twitter trips		94		47		63		101
Inner marker	80	74	50	44	19	29	84	58
Distance	94	78	66	49	33	24	95	59
Origin suburb-related variables								
Area	17	9	12	0	2	11	29	5
Population	18	17	0	2	11	7	36	2
Housing	29	19	4	0	32	9	46	12
Entertainment		11		18		12		8
School		25		1		9		12
Food		17		10		14		12
Nightlife		16		11		18		28
Sports		17		11		18		17
Professional		14		15		19		14
Shopping		12		9		12		10
Destination suburb-related variables								
Area	25	10	11	0	33	2	31	1
Population	23	19	0	5	0	0	34	8
Housing	30	22	0	0	3	0	45	16
Entertainment		16		15		27		7
School		23		0		3		10
Food		16		13		12		16
Nightlife		17		11		13		28
Sports		15		10		12		18
Professional		13		15		11		16
Shopping		12		15		8		13

City	CUS		Idaho		Daytona		NE Florida	
Model #	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Link variables								
Twitter trips		33		36		14		2
Inner marker	15	13	41	16	65	54	22	21
Distance	60	59	98	97	96	78	94	98
Origin suburb variables								
Area	0	0	5	1	7	3	16	49
Population	25	19	45	39	19	8	30	15
Housing	24	17	48	43	10	4	20	18
Entertainment		0		2		4		7
School		21		6		4		4
Food		1		11		2		13
Nightlife		0		4		2		19
Sports		0		0		0		3
Professional		0		0		0		6
Shopping		0		8		0		5
Destination suburb variables								
Area	0	0	4	2	6	2	13	46
Population	25	21	44	40	18	8	30	23
Housing	26	19	49	48	13	0	20	19
Entertainment		0		4		5		6
School		21		6		3		2
Food		0		4		2		1
Nightlife		1		3		1		9
Sports		0		3		4		6
Professional		0		6		1		7
Shopping		0		11		5		4

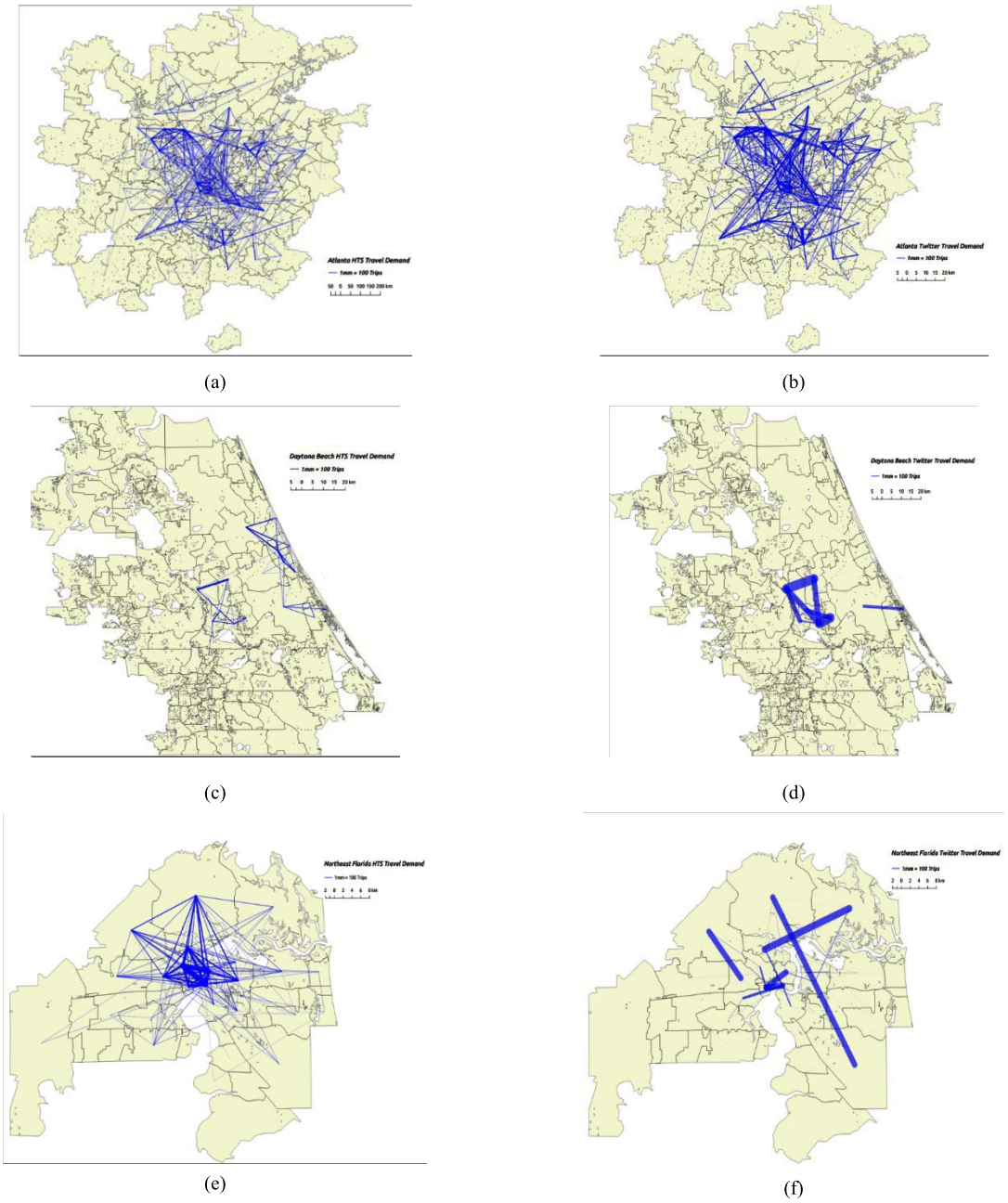

Fig. 2. (a). Atlanta HTS travel demand. (b). Atlanta Twitter travel demand. (c). Daytona Beach HTS travel demand. (d). Daytona Beach Twitter travel demand. (e). NE Florida HTS travel demand. (f). NE Florida Twitter travel demand.

cleaning should be performed when using social media data from tourist areas.

V. Further Discussion

A. Purpose-Based Discussion

HTS always reports trips generated by local residents together with their travel purposes. In this paper, the influence of travel purpose on the utility of social media data is discussed as well. To do this, relevant trips were considered in three groups: Professional (School and Work trips), Shopping and Entertainment. Although other categories, for example, health (medical care) can generate a number of trips as well, it lacks data from either socio-demographic or social media. So we do not take them into consideration at this stage. The numbers of trips for each purpose are shown in Fig. 3.

Applying similar methods, the results of the RF regression model and the regression results with and without social media data were compared. For each purpose group, the dependent variable was HTS-data trips of different purposes. The independent variables included all Twitter-data trips, distance, other socio-demographic data and different venues from Foursquare data. The model for Professional purposes contained School/College and Professional check-ins. The Shopping group model contained check-ins for Shopping venues, and the Entertainment model contained check-ins for

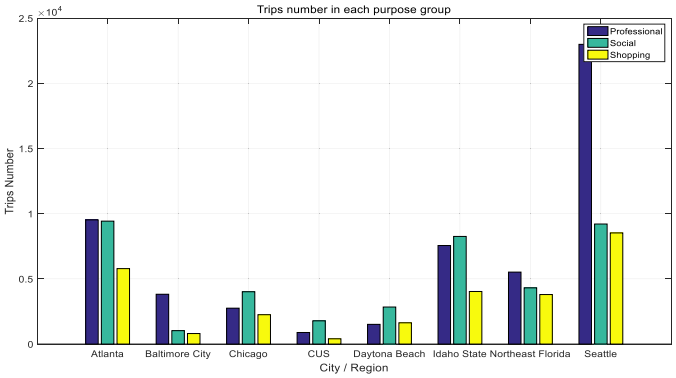

Fig. 3. Trips taken for each type of purpose.

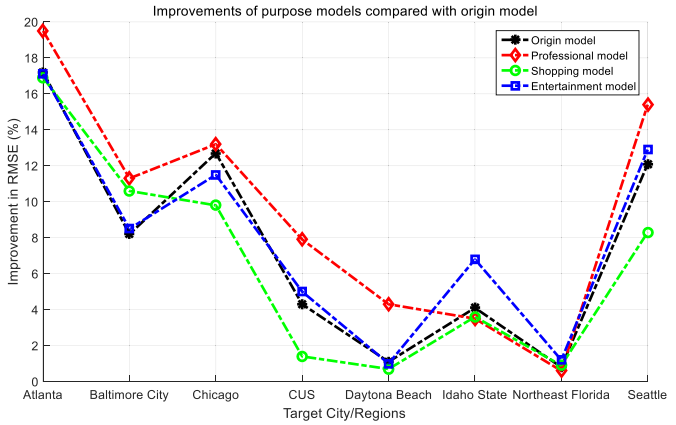

Fig. 4. Improvement in purpose models compared with origin model.

the other four venues. Using the improvement in RMSE as a standard, the results are shown in Fig. 4 below:

The figure suggests that the trends of each group were not significantly different to those of the origin model (black line). However, the professional model (red line) usually performed better. This may have two reasons. Firstly, it might be easier to label schools or colleges by check-ins from Foursquare because they usually have larger areas and fewer distractions. That offers the models more information for determining trip origins and destinations. In addition, compared with other groups, School or Professional trips are repeated daily and, therefore, are more predictable. Once one School or Professional trips is extracted from Twitter data, it represents a large number of trips with similar origins and destinations, resulting in a more efficient utilization of social media data.

B. Mode of Transport-Based Discussion

Applying a similar model, we now discuss the influence of transit mode on the travel demand regression with social media data. Modes of transport can be broadly divided into three groups: walk & bike, private vehicle and public transport. Figure 5 shows the trip numbers in each mode for different target regions according to HTS data. It can be seen that in most target regions, trips reported by HTS were made by private car. Because there may be small numbers of trips in some groups, the regression results for those groups are meaningless. Therefore, only groups with enough samples

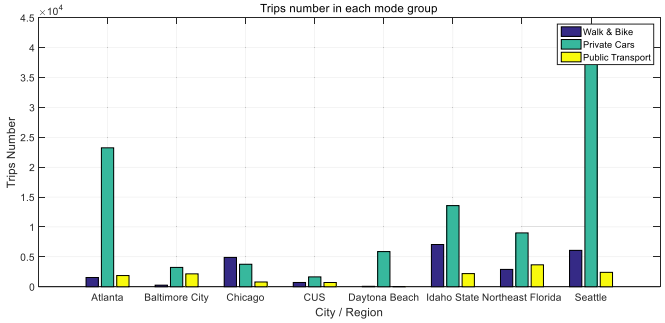

Fig. 5. Trip numbers in each transport mode for the eight cities.

TABLE IV
MODE-BASED ANALYSIS RESULTS

Target city	Function	Origin model	Walk & bike	Private car	Public transport
Atlanta	Metropolis	17.2%	-	25.0%	-
Baltimore City		8.2%	-	9.1%	6%
Chicago		12.7%	9%	11.5%	-
Seattle		12.1%	4.1%	15%	-
CUS	Rural	4.3%	3.5%	4.3%	1.2%
Idaho State		4.1%	2.2%	5.0%	1.9%
Daytona Beach	Tourist	1.1%	-	1.5%	-
Northeast Florida		0.8%	0.4%	0.8%	

(HTS trips) will be considered and others are represented by '-' in Table IV.

Table IV presents the performance of the created models grouped by transport mode compared with the improvements in origin models. Although there are some missing results due to a lack of samples ('-'terms), Table IV illustrates that applying social media data in the regression of private car trips provided better performance than those for other modes. It was also better than the origin models, which are aggregations of all modes. That is to say, information extracted from social media to some extent reflects users' travelling mode choices. Based on this result, it is believed that data from social media may provide more information on trips made by private vehicles than by other modes in travel demand estimation.

VI. CONCLUSION

The main contribution of the study is to demonstrate that the city function influences the utility of social media data in travel demand models. To prove this, OD trip models were generated for eight US cities with different functions (metropolis, rural and tourist). Independent variables obtained from Twitter and Foursquare were introduced into the model. Then, the results with and without use of social media-derived variables were examined by 10-fold cross-validation. The results show that travel demand models are mostly improved in metropolitan areas (>8%), followed by rural areas (>4%) and tourist areas (around 1%). Moreover, by analyzing the results based on suburb division, it is suggested that data from social media do not match HTS data well in the tourist areas, resulting in lower improvement to travel demand models. Meanwhile, the purpose-based analysis illustrates that although improvements in each purpose group model were significantly different

to those in origin models, estimates of School or Professional trips usually had better results.

According to the findings of this paper, it is believed that social media are appropriate data sources for use in traffic demand estimation. However, they play different roles in regions with different functions, suggesting that existing trip models using social media data must select the target city carefully. In metropolitan and rural areas, it could help to improve the performance of daily trip models. This would provide a time- and budget-saving data source for city planning that has more acceptable accuracy than HTS-based models. In tourist areas, although model performance rarely improved, with relevant data cleaning, social media data could be valuable for tourist-purpose trips and behavioural research.

For further studies, similar models could be applied to other areas of transport-related research to determine the utility of social media data. In addition, it is believed that a combination of social media data and machine learning techniques may be a helpful supplement for travel demand modelling in metropolitan areas. Therefore, other valuable variables could be introduced to generate travel demand models with higher performance and efficiency.

REFERENCES

[1] J. Dodson, P. Mees, J. Stone, and M. Burke, *The Principles of Public Transport Network Planning: A Review of the Emerging Literature With Select Examples*, no. 15, Griffith Univ., Mar. 2011.

[2] S. Bregman, "Uses of social media in public transportation," 99th Transp. Res. Board, Washington, DC, USA, Tech. Rep., 2012.

[3] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *J. Intell. Trans. Syst.*, vol. 11, no. 1, pp. 1–14, 2007.

[4] Y. Yue, Y. Zhuang, Q. Li, and Q. Mao, "Mining time-dependent attractive areas and movement patterns from taxi trajectory data," in *Proc. 17th Int. Conf. Geoinf.*, Aug. 2009, pp. 1–6.

[5] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[6] J. H. Lee, S. Gao, and K. G. Goulias, "Can Twitter data be used to validate travel demand models," in *Proc. 14th Int. Conf. Travel Behaviour Res.*, 2015.

[7] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," in *Proc. 6th Int. AAAI Conf. Weblogs Social Media*, 2012.

[8] Z. Wang, S. Y. He, and Y. Leung, "Applying mobile phone data to travel behaviour research: A literature review," *Travel Behav. Soc.*, vol. 11, pp. 141–155, Apr. 2018.

[9] Y. Asakura and E. Hato, "Tracking individual travel behaviour using mobile phones: Recent technological development," in *The Expanding Sphere of Travel Behaviour Research*. Bingley, U.K.: Emerald Group Publishing Ltd., 2009, pp. 207–236.

[10] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011.

[11] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput.*, 2013.

[12] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2009.

[13] (Jun. 1, 2018). *Statista*. [Online]. Available: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users.

[14] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transp. Res. C, Emerg. Technol.*, vol. 44, pp. 363–381, Jul. 2014.

[15] Q. Huang and D. W. Wong, "Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us?" *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 9, pp. 1873–1898, 2016.

[16] Q. Liu, Z. Wang, and X. Ye, "Comparing mobility patterns between residents and visitors using geo-tagged social media data," *Trans. GIS*, vol. 22, no. 6, pp. 1372–1389, 2018.

[17] J. H. Lee, A. W. Davis, S. Y. Yoon, and K. G. Goulias, "Activity space estimation with longitudinal observations of social media data," *Transportation*, vol. 43, no. 6, pp. 955–977, 2016.

[18] A. Majid, L. Chen, G. Chen, H. T. Mirza, I. Hussain, and J. Woodward, "A context-aware personalized travel recommendation system based on geotagged social media data mining," *Int. J. Geograph. Inf. Sci.*, vol. 27, no. 4, pp. 662–684, 2013.

[19] M. M. Hasnat and S. Hasan, "Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data," *Transp. Res. C, Emerg. Technol.*, vol. 96, pp. 38–54, Nov. 2018.

[20] M. M. Hasnat and S. Hasan, "Understanding tourist destination choices from geo-tagged tweets," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3391–3396.

[21] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science*, vol. 346, no. 6213, pp. 1063–1064, 2014.

[22] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller "Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges," *Transp. Res. C, Emerg. Technol.*, vol. 75, pp. 197–211, Feb. 2018.

[23] X. Zhang and S. T. Waller, "Implications of link-based equity objectives on transportation network design problem," *Transportation*, no. 46, pp. 1–31, 2018.

[24] M. Batty, "Big data, smart cities and city planning," *Dialogues Human Geogr.*, vol. 3, no. 3, pp. 274–279, 2013.

[25] S. Gao, J.-A. Yang, B. Yan, Y. Huand, K. Janowicz, and G. McKenzie, "Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area," in *Proc. 8th Int. Conf. Geogr. Inf. Sci. (GIScience)*, 2014.

[26] J. F. McDonald and R. A. Moffitt, "The uses of Tobit analysis," *Rev. Econ. Statist.*, no. 2, pp. 318–321, May 1980.

[27] J. H. Lee, A. Davis, E. McBride, and K. G. Goulias, "Statewide comparison of origin-destination matrices between California travel model and Twitter," in *Mobility Patterns, Big Data and Transport Analytics*. Amsterdam, Netherlands: Elsevier, pp. 201–228, 2018.

[28] T. Djukic, J. W. C. Van Lint, and S. P. Hoogendoorn,, "Methodology for efficient real time OD demand estimation on large scale networks," in *Proc. 93rd Annu. Meeting Transp. Res. Board*. Washington, DC, USA: Transportation Reseach Board, Jan. 2014, pp. 12–16.

[29] Z. Cheng, "Is social media an appropriate data source to improve travel demand estimation models?" Transp. Res. Board, Washington, DC, USA, Tech. Rep., 2018.

[30] S. Hasan and S. V. Ukkusuri, "Reconstructing activity location sequences from incomplete check-in data: A semi-Markov continuous-time Bayesian network model," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 687–698, Mar. 2018.

[31] *Annual Estimates of the Resident Population*, Bureau USC, Combined Stat. Area, Suitland, MD, USA, Apr. 2019.

[32] *Annual Report*, Council NFR, Daytona Beach, FL, USA, Apr. 2019.

[33] DBOC Website. (2018). *Daytona Beach CRA Annual Report 2018*. Accessed: Apr. 16, 2019. [Online]. Available: http://www.codb.us/documentcenter/view/14251

[34] Nexus. (2018). *Metropolitan Travel Survey Archive*. Accessed: Jul. 1, 2018. [Online]. Available: http://www.surveyarchive.org/

[35] S. Lee and J. Kim, "WarningBird: Detecting suspicious URLs in Twitter stream," in *Proc. NDSS*, 2012.

[36] Twitterdev. (2018) *Twitter API Reference Index*. Accessed: Jul. 1, 2018. [Online]. Available: https://developer.twitter.com/en/docs/api-reference-index.html.

[37] ProximityOne. (2018) *Cencus 2010 ZIP code Demographic Profile Dataset*. Jul. 16, 2018 [Online]. Available: http://proximityone.com/cen2010/zcta/dp.html.

[38] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Montreal, QC, Canada, Aug. 1995, vol. 1, pp. 278–282.

[39] L. Breiman and A. Cutler, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2005.

[40] L. Breiman, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[41] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Res.*, vol. 35, pp. W339–W344, 2007.

[42] T. Hastie and R. J. Tibshirani Friedman, "Overview of supervised learning," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009, pp. 9–41.
[43] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.
[44] Y. Zhao and J. Zobel, "Effective and scalable authorship attribution using function words," in *Proc. Asia Inf. Retr. Symp.* New York, NY, USA: Springer, 2005.
[45] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of Database Systems*. New York, NY, USA: Springer, pp. 532–538, 2009.

Zesheng Cheng is currently pursuing the Ph.D. degree with the Research Centre for Integrated Transport Innovation (rCITI), School of Civil and Environmental Engineering, University of New South Wales, Australia. His research interest is about applying large-scale data analysis technologies in urban transport networks.

Sisi Jian received the Ph.D. degree in transport engineering from the University of New South Wales, UNSW. She is currently a Research Associate with the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology. Her major research interests are travel demand modelling, shared mobility, travel time reliability, transport network planning, and optimization.

Taha Hossein Rashidi is currently an Associate Professor in transport engineering with the School of Civil and Environmental Engineering, UNSW, and a member of the Research Centre for Integrated Transport Innovation (rCITI). His research and teachings demonstrate how effectively transport engineering can draw upon the strengths of a broad range of disciplines to inform smart-city solutions. He is currently leading research into the interconnectivity between travel behaviour and time use and the potential of new mobility technologies to influence this paradigm as well as working on an industry partnership project with GoGet to undertake research on autonomous driving. He is also examining the capacity of social media data to complement existing data resources as a part of the development of an integrated multilevel modeling framework to demonstrate the relationships between land use and transport systems and the consequences this has for city planning and travel behaviour more broadly.

Mojtaba Maghrebi received the B.Sc. degree in civil engineering and the M.Sc. degree in construction engineering management from the University of New South Wales (UNSW), and the Ph.D. degree from UNSW, under the supervision of Prof. Waller (Research Center for Integrated Transport Innovation-rCITI) and Prof. Sammut (School of Computer Science). In his Ph.D. thesis, he focused on solving large scale dispatching problem with the machine learning approach. Around two years of his Ph.D. was spent at the School of Computer Science, UNSW, when he was trying to bridge between operation research (OR) and machine learning in vehicle routing problem (VRP) context. He is currently an Adjunct Lecturer (Assistant Professor) with rCITI, UNSW, and a meantime holds an Assistant Professor Position with the Ferdowsi University of Mashhad (Link). His research interests include innovative branching techniques in Mixed Integer Programming, supervised and ensemble learnings, and intelligent decision support systems. He was awarded the British Council Prize for a presented paper at the 10th Civil Engineering Student Conference. In 2013, he was bestowed first place in the IEEE Technologies of the Future and also highly commended IT Innovation Certificate which was received at the Australia ICT Celebration CeBit.

Steven Travis Waller is currently the Evans & Peck Professor of transport innovation with the School of Civil and Environmental Engineering, and the Director of the UNSW Research Centre for Integrated Transport Innovation (rCITI). He is a member of the Transportation Research Board of the US National Research Council, and the winner of the 2011 Hojjat Adeli Award for Innovation in Computing. He focusses on dynamic traffic assignment, stochastic routing, network design, planning for ITS, adaptive equilibrium, network behaviour under information, and bi-level optimization of transport networks.