Is social media an appropriate data source to improve travel demand estimation models?

Zesheng Cheng
Research Centre for Integrated Transport Innovation (rCITI)
School of Civil and Environmental Engineering, UNSW, Sydney, Australia
UNSW Australia
E-mail: z5121075@ad.unsw.edu.au

Sisi Jian
Research Centre for Integrated Transport Innovation (rCITI)
School of Civil and Environmental Engineering, UNSW, Sydney, Australia
UNSW Australia
E-mail: s.jian@unsw.edu.au

Mojtaba Maghrebi
Research Centre for Integrated Transport Innovation (rCITI)
School of Civil and Environmental Engineering, UNSW, Sydney, Australia & Department of Civil Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
E-mail: maghrebi@unsw.edu.au

Taha Hossein Rashidi
Research Centre for Integrated Transport Innovation (rCITI)
School of Civil and Environmental Engineering, UNSW, Sydney, Australia
E-mail: taha.hosseinrashidi@unsw.edu.au

Steven Travis Waller
Research Centre for Integrated Transport Innovation (rCITI)
School of Civil and Environmental Engineering, UNSW, Sydney, Australia
E-mail: s.waller@unsw.edu.au

World count: 5250+ (2 Tables + 5 Figures)*250 = 6500 Words

Submission Date: 1st August, 2017
ABSTRACT

Social media data has emerged as an innovative data source for traffic analysis. In this paper, we evaluate the effectiveness of including Twitter data into the Origin-Destination (OD) trip estimation. 1.3 million of geo-tagged tweets in the Greater Sydney Area for more than two months are collected, and information such as Twitter OD trips, the number of friends and followers of Twitter users are extracted as the independent variables in the OD trip regression model. The Random Forest regression technique is applied to develop the OD trip regression. The performance of the models considering Twitter data and not including Twitter data are compared via 10-fold cross-validation method. The results indicate that the accuracy and stability of the RF regression model can be improved if we consider Twitter data in the independent variables. Inspired from this finding, we conclude that social media data can be an effective data source to improve the prediction of traditional travel demand models. The regression results at the suburb level also suggest that the heterogeneity of socio-demographic features across suburbs will affect the model performance. To further improve the prediction, it is necessary to categorize suburbs into groups based on socio-demographic characteristics such as population density and distance to city center, and develop a separate OD trip regression model for each group.

Key words: Social media data; Geo-tagged tweets; Machine learning; Random Forest regression.
1. INTRODUCTION
Nowadays, with the growth in population and the development in economy, traffic demand has dramatically increased especially in large cities all over the world, resulting in problems such as congestions, imbalanced transport infrastructure utilizations and decline in urban travel efficiency (1). Facing these issues, it is significant for metropolitan planners to create a more efficient urban transport network (2). Accurate prediction of travel demand is the fundamental step to ensure an efficient transport network planning (3). Official traffic diaries, such as Household Travel Survey (HTS), have been utilized as the primary data source for travel demand estimations (4). However, HTS will take a large amount of budget and labor force and an extended time period to collect data. In order to address this problem, new data sources, such as social media (5), smart phone (6) and taxi trajectory systems (7), have been applied to estimating travel demand due to their cost-efficiency and convenience to obtain.

Among the new data source, social media has the advantage of low cost and high impact user coverage (8). According to the statistics, the number of active user accounts is larger than 328 million at the end of quarter 2, 2017 (9). Some of the users post tweets with their coordinates (latitude and longitude). The information could help to determine their locations or even tract their mobility patterns. It becomes the basis of taking social media data into transport studies.

Social media data analysis applying on urban transport research is a novel research topic emerged recently. Majid et.al (10) estimated the destination and accommodation of tourists in unfamiliar city using the geo-tagged information from social media data. Since then, harvested social media data has been applied to different areas of transport research. Ruths and Pfeffer (11) discussed the prospects and advantages of estimating individual behavior using social media as data source. The team proposed a filter model to exclude the nonhuman accounts in database collected from social media. They compared the results of several analysis methods applying on the same database and concluded that social media data could reduce the bias of individual behavior estimation. In addition, Lee et.al (12) collected geo-tagged tweets around southern Santa Barbara, American for 17 weeks. They studied users’ activity spaces based on them. Since the results showed the growth of activity space was not influenced by user’s tweet habits, it was concluded that Twitter was a valuable data source for long-term activity space studied.

Compared with other new data source, the user groups of social media got a dramatically expand in the past decade (13). Meanwhile, the posted contents were increasingly rich, which provided a large amount of real-time data for analysis (14). Under this circumstance, social media became a potential data source for travel demand estimation as well. The fundamental theory of taking social media data into travel demand estimation was emerged from Dr.Gao and his team’s (15) research. The team collected geo-tagged tweets around the Greater Los Angeles Area to create an Original-Destination (OD) trip estimation algorithm. The algorithm declared that if one user posted a tweet in different locations within 4 hours, it could be considered as one OD trip. Those extracted trips were place-based aggregated and validated with the
data from American Community Survey (ACS). The results suggested a strong
 correlation between OD trips extracted from Twitter data and from ACS data (Person
correlation coefficient = 0.91, p value = 0.0017).

Based on Dr. Gao and his team’s (16) algorithm, a more recent study proposed an
approach to apply Twitter data on validating travel demand models. The authors
applied the latent class analysis and a Tobit regression model to estimate travel
demand among different sub-regions in Los Angeles using Twitter data and
socio-demographic data. They concluded that it is an appropriate approach to covert
Twitter OD matrix into the official travel demand model. The Tobit model developed
in this research considered only non-negative terms of dependent variables with a
normal distributed error term (17). However, for a given origin or destination, most
demographic variables considered in the model are not linearly related to the
dependent variable. Therefore, to predict the OD travel demand more accurately, a
non-linear or non-parametric regression technique needs to be proposed to improve
the modeling performance.

Recently, several non-parametric regression model based on machine learning
techniques have been applied on travel demand estimation. Djkic, Van Lint and
Hoogendoorn (18) discussed the application of dimensionality reduction and principle
component analysis on real OD demand estimation. They defined a new transformed
variable called ‘demand principal components’ and demonstrated a dramatically
improvement of OD estimation accuracy. Zhan et al. (19) applied a hierarchical
regression tree model to estimate travel demand, frequency and mode choice of
university student in China. The paper declared that the model revealed the features of
students’ travel behaviors. Moreover, Saadi et al. (20) proposed a new model for OD
matrix estimation based on random forest algorithm. They adopted data from a travel
survey and validated their model by Belgium National HTS. However, none of those
studies used social media data as data source. Due to the fact that OD trips extract
from social media data has strong correlation with practical OD trips (15), it will
improve the performance of the model by taking social media data into consideration.

This paper develops a regression model to improve the prediction of the OD travel
demand estimated by the HTS in the Greater Sydney Area using Twitter data and
census and geographic data. The regression model is built based on a machine
learning technique, Random Forest (RF). Compared with other regression methods,
the main advantages of random forest are its flexibility and higher accuracy (21). The
independent variables include OD trips extracted from tweets, users’ Twitter social
network information, and socio-demographic data. The research demonstrates that
Twitter data is a possible data source to improve the accuracy of RF regression model
on OD matrix estimation. But the prediction accuracy varies across areas with
different socio-demographic characteristics. Though the regression residual analysis
across different suburbs, it is found that, first, the distance between a suburb and CBD
could be another possible feature which influences the estimation accuracy. Second,
in the suburbs with lower population density, the collected variables in the regression
model might not be able to reflect their actual travel demands. Taking Social media
data into machine learning OD estimation model is a topic which has not been
touched in travel demand studies. Since RF is a flexible, high accurate regression methods (21) and Twitter trips is highly correlated with actual statistics trips (15). This paper suggests that it might a possible selection to improve the accuracy of OD trips estimation by the combination of them.

The paper contains five sections. Section 2 introduces the data used in the regression model. Section 3 discusses the methodology, followed by the discussion of the regression results. Finally, Section 5 concludes the key findings and highlights further research directions.

2. DATA DESCRIPTION

In this study, we apply three datasets for the regression analysis. The three datasets are: 1) New South Wales (NSW) HTS data (22); 2) Extracted information from Twitter data including geo-tagged location data and personal Twitter network information, including number of followers, friends and favorites; and 3) Census data obtained from Australian Bureau of Statistics (23) and other land use data.

2.1 Dependent variables: New South Wales (NSW) HTS data

The Household Travel Survey (HTS) published by the transport department of NSW, Transport for NSW, estimates the average number of trips generated between different regions in the Greater Sydney Area. It should be noted that the latest available HTS was published in 2013, which provides a detailed OD trip matrix based on the personal travel data in the year of 2013. According to demographic statistics reports from Australia Bureau of Statistics, the growth rates of generated trips are less than 2% annually in Sydney. Although geo-tagged Twitter trip data was collected in 2017, it is believed that the error of the OD trips estimation is in an acceptable range. Therefore, the OD trip matrix generated via NSW HTS-2013 can still be considered as the dependent variable in the regression model.

In HTS, the Greater Sydney Area is divided into 43 local government areas (LGAs). However, in the created OD matrix (43-by-43), 300 out of the 1849 OD pairs are estimated to have zero demand, which can be too microscopic for the regression model. Therefore, a more aggregated suburb system is required to reduce the impact of this problem. For reporting convenience, we apply the division system adopted by Australian Bureau of Statistics and aggregate the 43 LGAs in HTS into 15 larger census blocks. Based on this regional division, a 15-by-15 OD matrix can be recreated. There are 210 OD links (15*14) excluding the conditions in which origin and destination are in the same suburb. The matrix is reshaped to a 210-by-1 vector and become the dependent variable of the regression model. The range of the number of trips is from 0 to 196,000. The average and standard deviation is 22441 and 31852 respectively. The detailed regional division of the 15 zones and the related OD travel demand aggregated from HTS are shown in figure 1 below. In the figure, a thicker blue line links a pair of suburbs stands for higher travel demand between this OD pair due to HTS statistics.
FIGURE 1: Suburb division of Sydney and HTS statistics Trips

2.2 Extracted information from Twitter data

The Twitter data was collected via the Twitter application programming interfaces (APIs), public platforms for developer to access features or data of Twitter and its relevant applications. Among those APIs, Stream API could collect tweets within a specific area just after it posted instantly (24). We used this API to collect the geo-tagged tweets from 15th Feb to 30th Apr, 2017 for the regression model. However, due to the download rate limitation of Stream API (150 tweets per 10 minutes), the equivalent collection time is around 10 days. During this time period, 1,300,057 geo-tagged tweets have been collected from 171,529 users. On average, 4090 trips between different suburbs have been extracted per day based on Dr. Gao and his team’s algorithm (15). The detailed distribution of the trips and their relationship with HTS OD matrix are shown in figure 2(a) and 2(b) respectively. In figure 2(b), similar to figure 1, thicker lines stand for higher travel demands. In figure 2(b), it suggests that there is roughly a proportional relationship between Twitter trips and HTS trips. However, there are also some outliers existed.
In addition to OD trips, more information could be extracted from collected tweets, including the number of followers, favorites and friends of users travelling along each OD link. Those three features are considered as independent variables of the regression model as well.

3. Other relevant census data and land use data

Besides the 4 independent variables extracted from geo-tagged tweets, there are 29 more variables considered in the regression. These variables are generated based on the census and land use data. Table 1 below lists the data sources and descriptions of the 33 independent variables considered in the regression model.
### Table 1: List of 33 Independent Variables

<table>
<thead>
<tr>
<th>Index</th>
<th>Independent Variables</th>
<th>Unit</th>
<th>Mean</th>
<th>Std.</th>
<th>Data Source</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Twitter Trips</td>
<td>Trip</td>
<td>181.8</td>
<td>252.9</td>
<td>Geo-tagged Tweets</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Distance</td>
<td>km</td>
<td>43.7</td>
<td>22.9</td>
<td>ArcGIS</td>
<td></td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>O/D Area</td>
<td>km²</td>
<td>809.0</td>
<td>1025.9</td>
<td>Bureau of Statistics (AU)</td>
<td>Period: 2011 - 2016</td>
</tr>
<tr>
<td>5 &amp; 6</td>
<td>O/D Population</td>
<td>10,000</td>
<td>30.1</td>
<td>11.4</td>
<td>Bureau of Statistics (AU)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>7 &amp; 8</td>
<td>O/D Population Density</td>
<td>1,000/km²</td>
<td>1.8</td>
<td>1.6</td>
<td>Calculated</td>
<td></td>
</tr>
<tr>
<td>9 &amp; 10</td>
<td>O/D Resisted Vehicles</td>
<td>10,000</td>
<td>17.2</td>
<td>6.3</td>
<td>Bureau of Transport Statistics (NSW)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>11 &amp; 12</td>
<td>O/D Average Vehicles per Household</td>
<td>vehicle</td>
<td>1.6</td>
<td>0.4</td>
<td>Bureau of Transport Statistics (NSW)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>13 &amp; 14</td>
<td>O/D Average Travel Distance of Residents</td>
<td>km</td>
<td>9.3</td>
<td>3.7</td>
<td>Bureau of Transport Statistics (NSW)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>15 &amp; 16</td>
<td>O/D Average Travel Time of Residents</td>
<td>min</td>
<td>22.4</td>
<td>1.9</td>
<td>Bureau of Transport Statistics (NSW)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>17 &amp; 18</td>
<td>O/D Housing Number</td>
<td>10,000</td>
<td>10.6</td>
<td>7.3</td>
<td>Bureau of Statistics (AU)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>19 &amp; 20</td>
<td>O/D Housing Density</td>
<td>10,000/km²</td>
<td>0.8</td>
<td>1.4</td>
<td>Calculated</td>
<td></td>
</tr>
<tr>
<td>21 &amp; 22</td>
<td>O/D Number of Employees</td>
<td>10,000</td>
<td>17.3</td>
<td>13.8</td>
<td>Bureau of Transport Statistics (NSW)</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>23 &amp; 24</td>
<td>O/D Employee Density</td>
<td>10,000/km²</td>
<td>1.7</td>
<td>2.7</td>
<td>Calculated</td>
<td></td>
</tr>
<tr>
<td>25 &amp; 26</td>
<td>O/D Average Income</td>
<td>AUD 1,000</td>
<td>1.6</td>
<td>0.2</td>
<td>Australia Realestate Website</td>
<td>Period: 2016 Update</td>
</tr>
<tr>
<td>27</td>
<td>Friends number</td>
<td>10,000</td>
<td>24.0</td>
<td>29.3</td>
<td>Geo-tagged Tweets</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Follower number</td>
<td>10,000</td>
<td>79.4</td>
<td>78.3</td>
<td>Geo-tagged Tweets</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Favorite number</td>
<td>10,000</td>
<td>3.2</td>
<td>6.9</td>
<td>Geo-tagged Tweets</td>
<td></td>
</tr>
<tr>
<td>30 &amp; 31</td>
<td>O/D Property Price</td>
<td>AUD 100,000</td>
<td>12.5</td>
<td>7.5</td>
<td>Australia Realestate Website</td>
<td>Period: 2016 Update, Mean property price of 3-bedroom house</td>
</tr>
<tr>
<td>32 &amp; 33</td>
<td>O/D House Rental Price</td>
<td>AUD 100</td>
<td>6.1</td>
<td>2.2</td>
<td>Australia Realestate Website</td>
<td>Period: 2016 Update, Mean rental price of 2-bedroom house</td>
</tr>
</tbody>
</table>
3. METHODOLOGIES

3.1 Random Forest Regression

Random forest (RF) algorithm is a highly flexible machine learning technique which could be applied on both regression and classification tasks (25). It is the basic regression model used in this paper. Figure 3 below is an overview for creating a binary random forest from given data space and binary trees.

The learning unit of RF is called classification and regression tree (CART). The basic idea of CART algorithm is to divide the given space into a set of rectangular areas and then fit the point in each area to a constant or a simpler model. The most common CART algorithm is called binary tree which divides each area into two subareas recursively and decides the output for each subarea. Mathematically, for a given training data set D, we have (26):

\[
D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}
\]

Where:
- \(x^{(1)} \ldots x^{(m)}\): vectors contain dependent variables for sample 1 to sample m.
- \(y^{(1)} \ldots y^{(m)}\): independent variable for sample 1 to sample m.

After training, the space has been divided into J different subareas. For a given testing sample n, the output of regression tree could be expressed as (26):

\[
m(x^{(n)}) = \sum_{j=1}^{J} v_j \cdot I(x \in R_j)
\]

Where:
- \(x^{(n)}\): A vector contains the dependent variables of given testing sample n.
- J: The total amount of subareas.
- j: Index of each subarea.
- \(v_j\): The regression output of subarea j.
- \(I(\cdot)\): the indicator function returning 1 if its argument is true and 0 for otherwise
- \(R_j\): Subarea j where \(\bigcup_{j=1}^{J} R_j = 1, \cap_{j=1}^{J} R_j = \emptyset\)

To create a binary regression tree, one algorithm is to choose an optimized split variable and its split value and divide 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 (27).

RF regression is kind of ensemble learning technique which uses a bagging algorithm to integrate different regression trees together (28). Those regression trees are independent with each other and the estimation results of the forest is 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
selected N sample randomly. Same sample could be selected repeatedly which is called bootstrap sample methods (29).

2) Train each regression tree with m randomly selected features where m<<M. Repeat the step from 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 bagging process.

Compared with other regression techniques, there are two main advantages which make RF a better selection for our regression model. On the one hand, RF regression could estimate the significance or correlation of each independent variable automatically. That is because the worse estimated results from regression trees which trained by unimportant feature will cancel each other by voting. It means feature selection and correlation discussion is not required for the regression model. That is helpful to improve the operational efficiency and keep more detailed information for our regression dataset. On the other hand, due to the randomly selected training samples and features, the probability of over-fitting is relatively low. That made RF claimed to be “unexcelled in accuracy among current algorithm” (30).

The importance of the variables in the regression could be tested followed the step (31): 1) Compute the regression 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 step 1) and 2) several times to reduce its bias. The average difference between the old and new RMSE could reflect the importance of the variables. The higher the value is, the more important the variable is.

3.2 K-fold cross-validation
K-fold cross-validation is a model testing technique to test the performance of the model by using collected data iteratively. Theoretically, during the process, the primary database A is randomly divided into k equal sized packages. Each package contains M/K samples. One of the packages will be selected for testing and the rest of K-1 packages are used as training data for the model. The cross-validation process contains k iterations until each package has been used as testing data exactly once (32).

K-fold cross-validation is an appropriate method for model testing especially under the case of insufficient data. For our regression model, due to the fact that it could make the full use of the collected Twitter data, a 10-fold cross-validation (189 samples for training and 21 samples for testing in each fold), which is the most commonly used k-fold cross-validation process (33), has been applied. For one testing fold, the regression residuals and root mean squared error (RMSE) will be calculated and they will be important standards to evaluate our regression model.
4. RESULTS AND DISCUSSION

4.1 Results of 10-fold cross-validation
To test the effects of twitter data, there will be two databases. The first one including twitter data contains all 33 independent variables (HTS estimation with Twitter Data). The other one excluding twitter data contains the rest 29 variables (HTS estimation without Twitter Data). To apply 10-fold cross validation, the first step is to divide number 1 - 210 randomly into ten groups. Each group corresponds to one testing fold. Same grouping will be used for both of the two databases. Then the testing folds will be created by packaging the samples with same index as number in the group. For each iteration, one fold will be selected as testing fold and the other nine will be selected as training folds. The random forest model used in this paper contains 500 regression trees and each of them will be trained by maximum 6 features.

Table 2 compares the regression results of HTS data with and without Twitter data via four performance metrics, the average regression residuals, RMSE, standard deviation and coefficient of variation (regression std. / regression average) The regression results obtained from the model with Twitter data have lower residual, RMSE, standard deviation, as well as coefficient of variation, which indicates that the RF regression model has been improved after considering Twitter data.

<table>
<thead>
<tr>
<th>Regression Data</th>
<th>Residual</th>
<th>RMSE</th>
<th>STD.</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTS without Twitter Data</td>
<td>11754</td>
<td>16554</td>
<td>28623</td>
<td>1.157</td>
</tr>
<tr>
<td>HTS with Twitter Data</td>
<td>5088</td>
<td>7042</td>
<td>21619</td>
<td>0.934</td>
</tr>
</tbody>
</table>

Also, the importance of the variables from Twitter data has been tested. Table 3 below shows the increase in RMSE after permuting its value for each variable followed the step proposed by Trevor et.al (31). The average number of the 10-fold cross-validation has been reported. It can be concluded that the variable ‘twitter trips’ is the most import new variables. ‘Friend number’ and ‘follower number’ may also play roles in the estimation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>RMSE before permute</th>
<th>RMSE after permute</th>
<th>Increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter Trip</td>
<td>7042</td>
<td>14351</td>
<td>104%</td>
</tr>
<tr>
<td>Friends Number</td>
<td>7042</td>
<td>7533</td>
<td>7%</td>
</tr>
<tr>
<td>Follower Number</td>
<td>7042</td>
<td>8022</td>
<td>13%</td>
</tr>
<tr>
<td>Favorite Number</td>
<td>7042</td>
<td>7105</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Figure 3 below is box plots illustrate the residual, RMSE, standard deviation and coefficient of variation of the 10-fold cross-validation.
Figure 3 presents the box plots comparing the residual, RMSE, standard deviation and coefficient of variation of the 10-fold cross-validation. The box plots suggest that the regression model is more stable when Twitter data is considered in the independent variables than purely using socio-demographic variables. The outlier of figure 3(b) is appearing in 8th fold. By looking inside the test fold, one of the obvious improvements of the model is the ability on estimation of outliers. In testing fold 8, there are 6 out of 21 samples with extremely large or small numbers. Primarily, the regression without Twitter data leaves huge residuals for 5 of those samples, resulting in regression RMSE even larger than the average number of this fold. Twitter data might provide more information for RF regression to distinguish those samples which helps to dramatically improve the estimated results for 3 of them. The estimation results for the rest samples in the fold are improved slightly as well which reduce the RMSE to lower than 50% of average in the fold.

 Generally speaking, with the help of Twitter data, random forest OD trips regression model develops the ability to process the data. It can be concluded that Twitter data is probably an appropriate data source for OD matrix regression model to improve the estimation accuracy and stability.

4.2 Suburb-based residual analysis

This section presents the performance of the regression model at the suburb. The Greater Sydney is divided into 15 suburbs based on LGA. Figure 4 illustrates the total residuals obtained from cross-validation for these suburbs. It can be found that for a specific suburb, the closer to the CBD area (City and Inner South), the larger amount of regression residual it has. This can be explained by several reasons. On the one hand, according to HTS statistics, suburbs which are closer to city center generate more trips than suburbs that are further away. It might be more difficult for RF model
to stabilize its estimation on samples with larger number. On the other hand, it also shows that distance between the suburb and CBD could also be a meaningful variable for this regression model.

In addition, the residual ratio (regression residual / average) for each suburb has been analyzed as well. Figure 5(a) shows the distribution of the regression residual ratio when the suburb is the trip origin, and Figure 5(b) presents the distribution of regression residual when considering the suburbs as trip destinations. According to Figure 5(a)(b), each suburb has similar residual ratios no matter its role is origin or destination. It can be inferred that the model might have the ability to distinguish different suburbs after its training by the collected geographical features. Suburbs with higher residual ratio and lower regression accuracy are mainly concentrated in two areas, which are the South and North-East of the Greater Sydney Area. There are several common geographical features of those suburbs. First, these suburbs are farther away from CBD area. They have larger land areas covered by forest or coast, which lead to the lower population density. When modeled with other suburbs, the collected variables might not be able to reveal both of their ability for trip generation in one regression model. Concluding from Figures 4 and 5, the heterogeneity of geographical features across suburbs can affect the performance of the OD trip regression model. Therefore, it can an approach to improve the model performance if we categorize suburbs based on geographical features, and develop separate regression models for different groups of suburbs.
FIGURE 5(a): Residual ratio for suburb as origin

FIGURE 5(b): Residual ratio for suburb as destination
5. CONCLUSION

In this study, we developed and examined an OD trip estimation model in the Greater Sydney Area based on Random Forest regression techniques. Several new independent variables obtaining from Twitter data including Twitter OD trips, numbers of friends and followers have been introduced into the model. We examined the estimated results with and without the variables from Twitter by HTS data using 10-fold cross-validation. The results showed that the accuracy and stability of the regression model could be improved if we considered Twitter data in the model. Inspired from this finding, Twitter data can be an appropriate data source to improve the performance of random forest OD trip regression model. Furthermore, by analyzing the total regression residual and residual ratio at the suburb level, we found that the distance between a given suburb and CBD could influence the accuracy of the prediction. In addition, for suburbs with disparity in demographic features, it is suggested to estimate their OD matrix with separate models.

For further studies, the regression models with the same algorithm could be tested by different datasets from other metropolis around the world. Other valuable variables such as land use characteristics could also be introduced into the regression model. By applying appropriate analytical model, it is believed that the combination of social media data and machine learning techniques will become a helpful supplement for travel demand estimation.


16. Lee, J.H., Gao, S. and Goulias, K.G., 2015, July. Can Twitter data be used to
validate travel demand models. In *14th International Conference on Travel Behaviour Research*.


22. Household Travel Survey, 2013

Bureau of Transport Statistics, Transport of NSW

Election Publication No. D2013-HTS-Table2-linked


Electronic Publication no. E2016-07-LGA-Census:


database systems (pp. 532-538). Springer US.