DEVELOPMENT OF DESTINATION CHOICE SET GENERATION METHOD USING A BEHAVIOR-SIMILARITY HUMAN NETWORK

Yu Fujiwara[1], Junji Urata[2], Makoto Chikaraishi[3], Akimasa Fujiwara[4]

[1]Member of JSCE, Dept. of Advanced Science and Eng., Hiroshima University
(1-5-1, Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8529, Japan)
E-mail: m213720@hiroshima-u.ac.jp (Corresponding Author)
[2]Member of JSCE, Dept. of Civil Eng., The University of Tokyo
(7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan)
E-mail: urata@bin.t.u-tokyo.ac.jp
[3]Member of JSCE, Dept. of Advanced Science and Eng., Hiroshima University
(1-5-1, Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8529, Japan)
E-mail: chikaraishim@hiroshima-u.ac.jp
[4]Member of JSCE, Professor, Dept. of Advanced Science and Eng., Hiroshima University
(1-5-1, Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8529, Japan)
E-mail: afujiw@hiroshima-u.ac.jp

This study proposes a method of generating choice sets, focusing on travel to non-regular destinations. More specifically, We first develop an association network based on their personal travel histories. We then generate an individual-specific choice set of destinations, using travel histories of people who have the high behavioral similarity index obtained from the developed association network. The prediction accuracy of the proposed model is empirically confirmed using public transit smart card data in Hiroshima. Specifically, the results show that the proposed method can significantly reduce the number of alternatives in the choice set, while maintaining its prediction performance.

Key Words : *Destination choice, Smart card data, Choice set, Non-regular trip, Behavior similarity*

1. INTRODUCTION

It is challenging to predict the non-daily travel behavior, for example, transportation for leisure purposes, etc. Traditional travel demand forecasting approaches used for the planning of public transport and road networks have been constructed not to predict such non-daily travel but to predict daily travel with high regularity.

Emerging transportation network companies (TNCs) such as Micro-Transit and e-hailing have a great potential to efficiently provide services to non-daily travel. In general, TNCs provide their services only in areas with high demand, while the service provision to areas with low demand is desiable from the perspective of improving social welfare (Hensher et al., 2020). The potential to provide services to low demand areas would depend on the accuracy of short-term prediction. For example, if the accurate prediction is possible, TNCs can efficiently provide services to non-daily travel in addition to travel with high regularity.

Accurate prediction of non-daily travel is difficult essentially because it is not regularly conducted (hereinafter, we call it non-regular travel), implying that individual's travel history is less useful to improve the prediction accuracy. While time series and machine learning methods have been utilized to improve the prediction accuracy, they rely on individual's travel history.

In this study, in order to improve the prediction accuracy of destination choice for non-regular travel, we propose a novel approach of utilizing travel history of others. The proposed approach is inspired by (1) e-commerce recommendation systems (Schafer et al., 1999; Sarwar et al., 2000; Schafer et al., 2000), and (2) dicusssions on choice set generation in the development of discrete choice models with the large choice set (For route choice, Frejinger et al., 2009; Yao et al., 2020. For destination choice, Crompton and Ankomah, 1993, Decrop, 2010). The approach consists of two steps. The first step is to develop an association network based on their personal travel histories, which has been widely used in e-commerce

recommendation systems. The second step is to generate an individual-specific choice set of destinations, using travel histories of people who have the high behavioral similarity index obtained from the developed association network. We empirically compare the prediction accuracy of the proposed method with those of other conventional methods including a deep neural network model, using public transit smart card data.

2. METHODOLOGY

(1) Association network and generation

In this sub-section, we show the main idea of our approach. First, an association network is formed based on the degree of similarities identified by means of the spatiotemporal patterns of non-regular trips using long-term behavior history data. A similarity is defined as the ratio of the number of stations commonly used in the past to the total number of stations used by two users. The similarity of the past destinations of users i and j, S_{ij} is defined as follows:

$$S_{ij} = \frac{\left|\{D_i\} \cap \{D_j\}\right|}{\left|\{D_i\} \cup \{D_j\}\right|} \tag{1}$$

where $\{D_i\}$ is the set of destinations visited by user i, and $|-|$ denotes the number of elements in the set $\{D\}$. S_{ij} is the ratio of the number of stations used by the two users in common in the past to the number of stations used by them. This similarity is well known as the Jaccard index (Jaccard, 1912). We then form an association network based on this similarity.

When configuring the association network, the optimal network size n is searched for each individual with two indicators. ①Cover ratio (*CR*): the percentage of user i's visited destinations that is covered by the generated alternatives using the proposed method. The equation is defined as follows:

$$CR_i(n) = \frac{\left|D_i^{rh-in}(n)\right|}{\left|D_i'\right|} \tag{2}$$

$$\left\{D_i^{rh-in}(n)\right\} = \left\{D_i^{rh}\right\} \cap \left\{\{D_{1i}'\} \cup \cdots \cup \{D_{ni}'\}\right\}, \quad (\{1i, \ldots, ni\}\ in\ SL_i(n)) \tag{3}$$

where, $SL_i(n)$ is a set of users who are within n th similarity and $\{D_i^{rh}(n)\}$ is the set of non-regular destinations for user i. It should also be noted that the set $\{D_i\}$ in equation (1) and the set $\{D_i'\}$ is neighboring users' destination choice sets and that in equation (3) do not necessarily coincide. ② Generated destinations (*GDs*): the number of generated alternatives. The equation is follows:

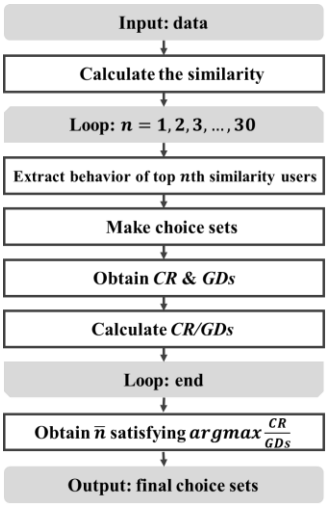

Fig. 1 Framewoek of our generation algorithm

$$GD_i(n) = |\{D_{1i}'\} \cup \ldots \cup \{D_{ni}'\}| \tag{4}$$

This indicator is introduced to evaluate whether the choice set size is kept as small as possible. We prefer to keep the choice set as small as possible in principle, largely because the low predictability of the destination choice model essentially comes from the large number of alternatives that cannot really be distinguished by their alternative-specific attributes. Note that a trade-off relationship exists between the *CR* and the *GD*s: if the number of *GD*s is too small, the choice set contains only a few or no stations that an individual has visited before, resulting in the lower value of the *CR*. Therefore, the optimal network size n for each individual should be determined by taking the trade-off into account. Since S_{ij} ranks the similarity between users, we calculated the values of $CR_i(n)$ and $GD_i(n)$ according to this ranking, and found the number of people n for which the maximum value of $CR_i(n)/ED_i(n)$ is obtained.

$$\bar{n} = \text{argmax}\frac{CR_i(n)}{GD_i(n)} \tag{5}$$

The ratio measure is the probability of predicting the correct destination when a destination is randomly selected from the reduced set of alternatives, weighted by the coverage ratio.

Given the above discussions, we propose an generation algorithm shown in Fig. 1. In the algorithm, a behavior-similarity network is constructed using top $\bar{n}$ similarity users, where $\bar{n}$ is optimized to maximize the ratio of the *CR* and the *GDs* for each individual.

(2) Evaluation

To evaluate the prediction accuracy of our proposed algorithm, we conducted an out-of-sample validation by comparing the average of predictability. The average predictability is defined as:

$$\text{Predictability} = \frac{\sum_{i \in I} \delta_i P_i}{|I|}, \quad \delta_i = \begin{cases} 1 \; if \; i = s \\ 0 \; if \; i \neq s \end{cases} \tag{6}$$

where, I is the number of samples, s is the actual station chosen, and P_i is the probability of selecting the observed destination i, and is calculated for each model. The predictability means the average choice probabilities for actual chosen stations. If the choice set does not contain the alternative, P_i is set as zero.

3. DATA

This study is empirically verified using one-year smart card data (SCD) in the Hiroshima metropolitan area in Japan. The smart card that was introduced in the Hiroshima area in 2008 has been used by 31 companies as of January 2021, including railway, bus and cab companies. Our study area was Hiroshima city. In Hiroshima, most daily commuters in the central business district use public transportation. Thus, people who cannot drive a car use public transportation. However, the composition of the representative modes of transportation in Hiroshima City is heavily dependent on automobiles. The SCD used in this study was collected in a 365-day period, from July 1, 2017 to June 30, 2018. During this period, 977,518 card holders completed a total of 86,170,009 trips. The study analyzed 236,179 holders who had travelled more than 100 times during this period. The number of observed stations of public transportation was 3654. The number is a maximum candidate of destination choice if we do not apply any choice set generation process. Note that SCD observed a destination station, and this study regarded this station as the traveler's destination. This study focuses on destinations that are not visited daily or those which are not attractive. We defined a non-regular destination (NRD) as that which has not been visited during the past 30 days, and/or destinations associated with 10 recorded trips in the passenger's travel history.

The total number of trips with NRDs as destinations was 6,245,856, and the total number of trips for the cardholders was 66,272,706, indicating that about one-tenth of the trips were for NRDs. In addition, the median number of NRDs for each user is 9 stations, suggesting that people who use public transportation on a daily basis have a certain number of NRDs. The top 20 stations with the highest number of passengers per hour were selected as the main arrival stations, and 71 stations with the highest number of visitors in the area were selected. As a result, we found that

- The ratio of stations other than the main arrival stations to the observed NRD was 0.52 on average.
- 36% of all NRDs are concentrated in 71 major arrival stations.

From the personal behavior histories, approximately 6% of all trips were non-regular trips and did not involve major destinations. Although it is difficult to deploy mass transit services for these trips, the volume is enough for a Micro-Transit transportation system for sustainable human life.

4. RESULTS

(1) Association network and generation

To validate our generation algorithm, we divided the data set into two: data for the first 10 months are used as a training set, and those the last two months are used as a validation set. For this analysis, we randomly selected 9,872 users whose travel histories contained NRDs. First, we show how the *CR* and *GDs* change with the number of neighboring users defined based on similarity using the training data. Fig. 2-1 shows the box plot of the *CR* by the number of neighboring users. Fig. 2-2 is a box plot of the *GDs* by the number of adjacent users. From these two figures, it can be seen that both the *CR* and the *GDs* tend to increase as the number of neighboring users n increases, suggesting that it is appropriate to determine the size of the network by calculating the optimal n for each individual, rather than determining an arbitrary n for every choice generation.

Fig. 3 shows the distributions of *CR* and *GDs* for the selected 9,872 users who had NRD in their travel histories. However, the figure for the validation data (Fig.3-2) excludes users who had zero NRDs in the last two months of the study period; the total number of users was 7,948. In the validation, we also used the set of alternatives learned by the training data. The results of the training data show higher *CR* compared to the validation data. The average *CR* of the training data is 0.435 and that of the validation data is 0.225. One of the reasons for the small *CR* of the validation data was the small number of NRDs in the last two months. The average number of the generated choice sets is 15.4 and the average number of users in the network is 2.3. This number is less than 1/200 of the full choice set, which means that our algorithm was

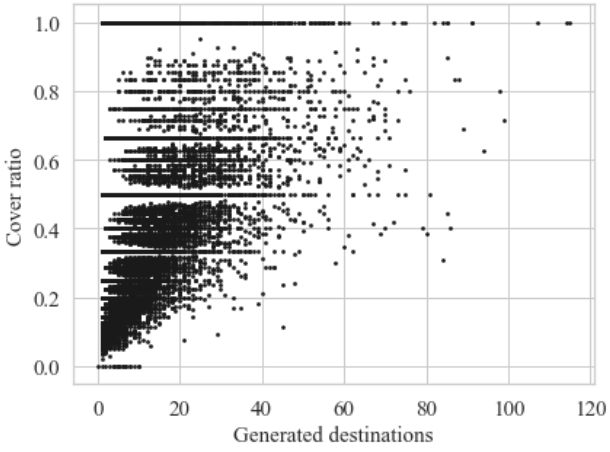

Fig. 2-1 Cover ratio variations as a function of the number of generated choice sets for training sets

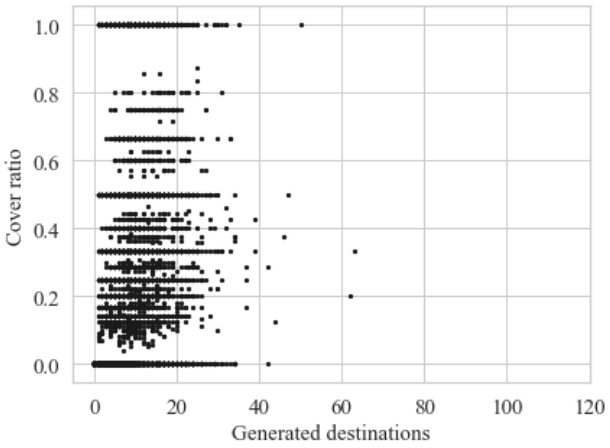

Fig. 2-2 Cover ratio variations as a function of the number of generated choice sets for validation sets

able to largely reduce the choice set. From this verification, we can see that the generation algorithm is not significantly wrong, but there is room for improvement in the association network construction.

(2) Model predictability

Here, we show the performance of the proposed algorithm for predicting the NRD of the generated choice sets compared to the following destination choice models:

(i) Multinomial logit (MNL) model with the proposed choice set

(ii) MNL model with choice set from user's own history

(iii) MNL model with choice set from major destinations

(iv) Deep learning model with full choice set

(v) MNL model with full choice set

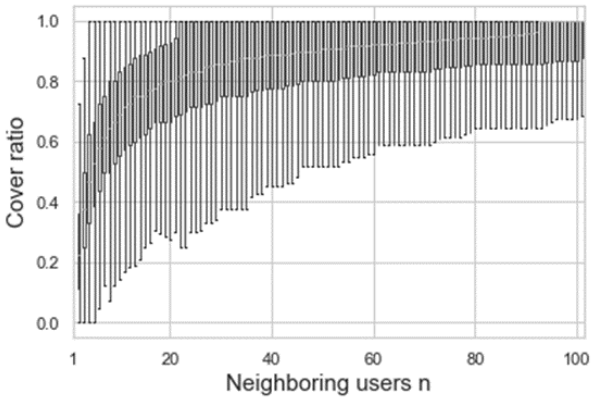

Fig. 3-1 Cover ratio as a function of the number of neighboring users

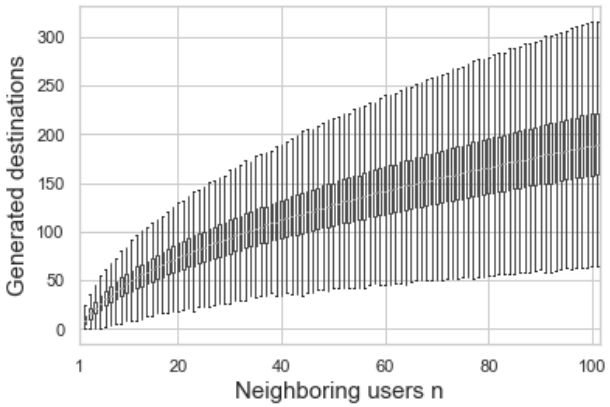

Fig. 3-2 Generated destinations as a function of neighboring users

In (i), the choice set is generated using the approach described above, applying the conventional MNL model. In (ii), the choice set is generated using the past non-regular trips of the users. For the model (iii), 71 major destinations are used as a choice set. The models (iv) and (v) used all the 3654 destinations. Models from (i) to (iii) used the same parameters obtained by estimating the MNL model (v). The parameters were estimated by maximum likelihood estimation method using about 1,000 randomized trips for NRD. The observed utility of destination s for discrete choice models is,

$$V_s = \alpha(TC_s + \beta TT_s + \gamma \ln(WT_s + 1) + \delta CF_s) \qquad (6)$$

where, TC_s is the travel cost to destination s [100 yen], TT_s is the travel time [h], WT_s is the transfer waiting time [h], CF_s is the ratio of commercial facilities to all facilities around the destination, and $\alpha, \beta, \gamma,$ and δ are parameters. The estimation results are shown in Table 1. Additionally, we compared a model with alternative-specific constants only, for the comparison. The ASC-only model resulted in a probability equal to the share of the selected destinat-

Table 1 Parameter estimation of MNL

Param.	Estimate	t-value
α	-0.403	-2259***
β	4.19	879***
γ	20.3	473***
δ	-6.17	-2702***
Initial log-liklihood	-3767585.0	
Final log-liklihood	-551116.2	
Sample	98339	

ion. For (ii) and (iii), Fig.4 shows the distribution of *GDs* for model (ii) and the proposed model, and Fig.5 shows the distribution of *CR* for model (iii) and the proposed model.

The results of model predictability are shown in Table 2. . The predictability of our proposed model is 0.1020 and outperforms all other models. From these results, we have obtained three important insights.

1. The prediction of NRD by discrete choice models with a very large number of candidate choices is very difficult, and that candidate generation plays an important role.
2. Focusing on the method of generating the choice sets, rather than increasing the complexity of the model structure as in the case of methods such as a deep learning model, contributes to improving the model accuracy.
3. Focusing on the method of generating choice sets, it is clear that generating choice sets based on the behavioral history of those with similar behavior is more predictive than an approach based on individuals' own past behavioral history.

For (ii) and (iii), Fig.4 shows the distribution of *GDs* for model (ii) and the proposed model, and Fig.5 shows the distribution of *CR* for model (iii) and the proposed model. From Fig.4, the proposed model has more *GDs* than model (ii), which covers stations that a user has never used and has the potential to predict NRDs. Figure 5 also shows that model (iii) with a fixed set of alternatives has a low *CR* overall and may not be able to represent individual behavior, making it difficult to predict non-regular trips.

5. CONCLUSION

In this study, we proposed a new choice set generation approach for NRD choice models with large choice sets. Specifically, we introduced an

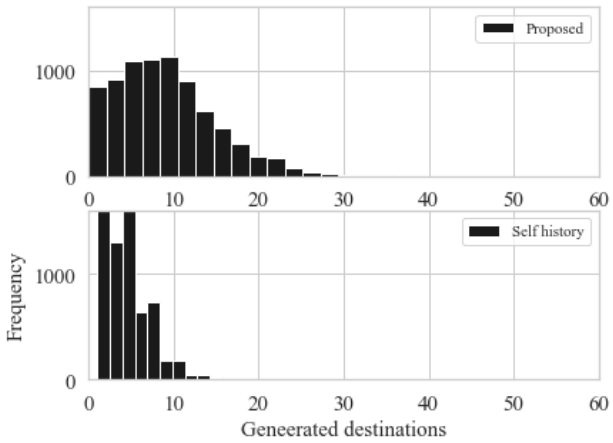

Fig. 4 Histograms of *GDs* of model(i) and model(ii) for validation data

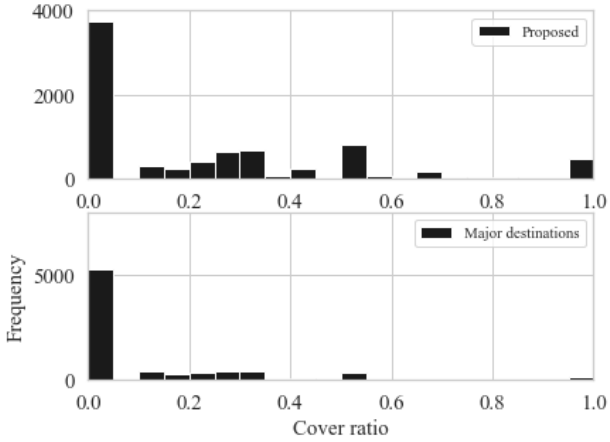

Fig. 5 Histograms of CR of model(i) and model(iii) for validation data

association network based on the similarity of travelers based on the spatio-temporal characteristics of their past travel behavior. To validate the proposed approach, we conducted an empirical analysis using IC card data of public transportation systems, including 110 million trips by approximately one million users in the Hiroshima metropolitan area. The individual behavioral histories showed that 6% of all trips were non-regular, indicating that a non-negligible proportion of trips were non-regular trips. In addition, we developed an association network-based NRD prediction method based on behavioral similarity of public transport usage and showed that the predictability of the proposed model is superior to that of conventional methods that do not generate choice sets or other possible simple choice set generation methods. It is also very interesting to note that for non-patterned behaviors such as non-regular travel behavior, the proposed method performed better than data-driven methods such as deep learning. We have also shown that the choice of gene-

Table 2 Comparison of predicability

Models	*Predictability*
(i) MNL model with the **proposed choice sets**	0.1020
(ii) MNL model with choice sets from **user's own history**	0.0588
(iii) MNL model with choice sets from **major destinations**	0.0230
(iv) **Deep learning model** with full choice sets*	0.0105
(v) MNL model with **full choice sets**	0.0036
(v') MNL-**ASC** with full choice sets	0.0028
(i+iii) MNL model with **proposed choice sets** and **user's own history**	0.0700

ration methods could have a larger impact on predictability than the choice of model structure.

ACKNOWLEDGMENT: The authors would like to acknowledge that a part of this research was conducted under the research project "Short-term travel demand prediction and comprehensive transport demand management", supported by the Committee on Advanced Road Technology under the authority of the Ministry of Land, Infrastructure, Transport, and Tourism of Japan.

REFERENCES

1) Hensher, David., Mulley, Corinne., Ho, Chin., Wong, Yale ., Smith, Goran., Nelson, John. : Understanding mobility as a service (MaaS): Past, present, and future, Elsevier, 2020.
2) Schafer, J., Ben., Konstan, A., Joseph., Riedl, John. : Recommender systems in e-commerce, *Proceedings of 1st ACM Conference on Electronic Commerce*, Denver, Colorado, United States, 1999.
3) Sarwer, Badrul., Karypis, George., Konstan, Joseph., Riedl John. : Analysis of recommendation algorithms for e-commerce, *Proceedings of the 2nd ACM conference on Electronic commerce, Minneapolis, Minnesota*, 2000.
4) Schafer, J., Ben., Konstan, A., Joseph., Ridel, John. : E-commerce recommendation applications, *Data Mining and Knowledgement Discovery*, Vol.5, pp.115-153, 2001.
5) Hensher, David., Mulley, Corinne., Ho, Chin., Wong, Yale ., Smith, Goran., Nelson, John. : Understanding mobility as a service (MaaS): Past, present, and future, Elsevier, 2020.
6) Hensher, David. : Future bus transport contracts under a mobility as a service (MaaS) regime in the digital age: Are they likely to change ?, *Transportation Research Part A*, Vol.98, pp.86-96, 2017.
7) Wright, Steve, John D. Nelson, and Caitlin D. Cottrill, : MaaS for the suburban market: Incorporating carpooling in the mix, *Transportation Research Part A*, Vol.131, pp.206-218, 2020.
8) Frejinger, E., Bierlaire, M., Ben-Akiva, M. : Sampling of alternatives for route choice modeling, *Transportation Research Part B*, Vol.43, pp.984-994, 20, 2009.
9) Yao, R., Bechor, S. : Data-driven choice set generation and estimation of route choice models, *Transportation Research Part C*, Vol.121, pp.102832, 2021.
10) Crompton J.L. and Ankomah, P.K. : Choice set propositions in destination decisions, *Annals of Tourism Research*, Vol.20, pp.461-476, 1993.
11) Decrop, A. : Destination choice sets an inductive longitudinal approach, *Annals of Tourism Research*, Vol.37 pp.93-115, 2010.
12) Jaccard, P. : The distribution of the flora in the alpine zone.1. *New Phytologist*, Vol.11 (2), pp. 37–50, 1912.