Probabilistic Estimation of Travel Behaviors Using Zone Characteristics

Masatoshi Takamiya, Kosuke Yamamoto, and Toyohide Watanabe

Department of Systems and Social Informatics
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{takamiya,yamamoto,watanabe}@watanabe.ss.is.nagoya-u.ac.jp

Abstract. There are many prior works of modeling travel behaviors. Most of them are investigated under the assumption that many kinds of data such as that of Person Trip (PT), which surveys travel behaviors, are available. Therefore, they do not consider an application to cities where the survey is not examined. In this paper, we propose a method for estimating travel behaviors using zone characteristics which is obtained from structural data of city. Focusing on dependent relationships between travel behaviors and city structure, we estimate the travel behaviors by means of the relationships. We first define trip and zone characteristics, and then introduce our method. With our method, we make use of Bayesian network constructed with PT data and the structural data. In addition, we show the effectiveness of our method through evaluation experiments.

Keywords: travel behaviors, Bayesian network, K2 algorithm.

1 Introduction

Generally, traffic simulations are utilized to measure the effect of new intelligent transportation systems in real world[1,2]. Although these studies exhibit several notable results, there are doubts about the validity of the results. Therefore, it is necessary to represent travel behaviors, which contain population flow, with computational simulations.

There are some studies to represent travel behaviors. Most of them estimate population flow using data of surveys which are examined on travel behaviors in real world, such as Person Trip (PT). Kitamura constructed a framework which simulates the living activities with activity-based approach which models individual travel behaviors[3]. Some studies adopt Neural network to the problem: for instance, Mozolin et al. compared[4] the performance of multilayer perceptron neural networks with that of maximum-likelihood doubly-constrained model, which is conventional model, for commuter travel behaviors, and Zhou et al. explored the application of back-propagation network to travel demand analysis [5]. However, those studies are under the assumption that many kinds of data such as residential information and survey data about travel behaviors

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 615–622, 2009.

of the city are available. It is expensive to examine the travel behaviors survey. Moreover, the survey is not examined in every metropolitan area. Therefore, in order to apply an estimation of general cities, we need a new method.

Our objective is to estimate travel behaviors using only structural data of city. The data contains position information about important facilities for travelers such as stations and schools. In addition, the data is able to be obtained without the survey about travel behaviors. We consider that there are general patterns of travel behaviors regardless of city and the patterns are involved by various factors such as time and location of the facilities. We estimate the behaviors using Bayesian network constructed by using the patterns.

The remainder of this paper is organized as follows. In Section 2, we refer to our approach. Section 3 mentions construction of Bayesian network. In Section 4, we explain how to estimate travel behaviors with our method. In Section 5, we report our experiments and results. Section 6 concludes this paper and offers our future work.

2 Approach

In order to achieve our objective, we propose a method for estimating travel behaviors using zone characteristics. Focusing on dependent relationships between travel behaviors and city structure, we extract and utilize trip patterns to estimate travel behaviors using zone characteristics. A city is divided into the several zones and the trip is defined as a personal movement from an origin zone to a destination zone for one purpose. Their characteristics are defined in Section 3.2. Using zone characteristics, we apply another city with only its structural data of city. With our method, probabilities which zones are selected as a destination zone with are calculated about an origin zone. Calculating them about overall origin zones, we can estimate travel behaviors in the entire city.

In order to represent the dependent relationship, we use Bayesian network. It is one of probabilistic models which represent conditional probability and indicate causal relationships by graph structure. Interpreting the dependent relationships as causal relationships, we are able to describe the travel behaviors using Bayesian network. Moreover, we discover the graph structure using K2 algorithm.

A flowchart of our method is shown in Fig.1. *Calculation* denotes a process which calculates probabilities for each zone characteristic, and *assignment* denotes a process which assigns the probabilities to all zones according to the characteristics. Bayesian network is constructed with PT data and structural data of city, and represents general travel behaviors. Therefore, we consider our method to be able to predict travel behaviors of other city.

3 Construction of Bayesian Network

Bayesian network has nodes corresponding to random variables, and represents static causal relationships among variables by the graph structure. For instance,

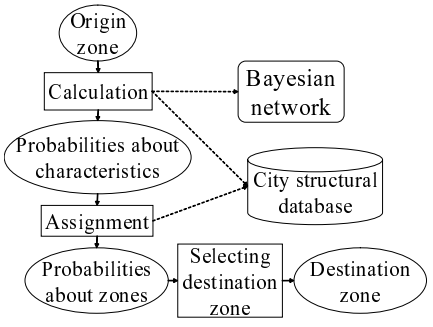

Fig. 1. Flowchart

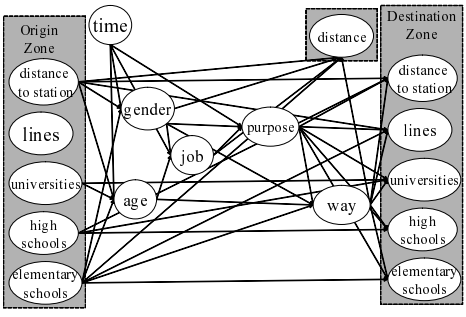

Fig. 2. Bayesian network

we assume that a node "a" corresponding to a variable "A" is conditioned by a node "b" corresponding to a variable "B". In this case, "b" is a parent node of "a" and there is a directed link from "b" to "a". Moreover, the each node has a Conditional Probability Table (CPT). CPT holds conditional probabilities for every combination of values its parent nodes have. Construction of Bayesian network has two phases. Firstly, entries in the CPT are calculated with research data. Secondly, an appropriate set of parent nodes is searched by algorithm for discovering graph structure.

3.1 Research Data

We utilize PT data of Chukyo metropolitan area where is in Japan as trip data, and structural data of Nagoya where is in the area as zone data. The PT examined a sample for 3% of the overall population in the area, which includes 259 zones in Nagoya, and was restricted to the residents of six or more ages. We extracted trips which move between zones in Nagoya from the PT data and the number of the trips is about 140000.

Bayesian network constructed with the data is shown in Fig. 2. *Lines* is a sum of train lines of stations in a zone, *distance* is spatial distance between centroid of an origin zone and that of destination zone, *time* is departure time. *Universities*, *high schools* and *elementary schools* are their sum.

3.2 Zone and Trip Characteristics

Zone. A zone is used as an origin or destination unit of a trip. This means that trips generating from inside a zone are aggregated as trips generating from a centroid of the zone. We divide a district into some zones according to PT. Zone characteristics represent an origin and destination zones in departure time of trips. The trips are dependent on the existences of important facilities for travelers. Therefore, we consider the characteristics to be represented by the information about the number and kinds of facilities in origin and destination zones for a trip. We define zone characteristics as follows:

- the number of elementary and junior high schools
- the number of high schools
- the number of universities
- sum of train lines of stations
- distance to the nearest zone, which has stations, from the zone (This is 0 if zone has station.)

Values of those parameters are obtained from the structural data.

Trip. We obtain trip patterns with PT data. We consider personal attributes and other attributes of trip. Therefore, we define trip characteristics as follows:

- age of travelers
- gender of travelers
- job of travelers
- travel purpose
- way to travel

where age, gender and job are personal attributes of trip. Values of those parameters are obtained from PT data.

3.3 Selecting Destination Order

We consider that there is dependent relationship among departure time, origin zone characteristics, personal attribute, travel purpose and destination zone characteristics. For instance, in the evening, students go home from school, and in the morning, students go to school and workers go to offices or shops. This denotes that personal attributes of trip, such as age and job, are dependent on the origin zone characteristics and its travel purpose is dependent on the personal attributes and origin zone characteristics. Therefore, we define destination selection order as follows:

1. origin zone characteristics
2. personal attribute
3. travel purpose
4. destination zone characteristics.

3.4 Discovery of Graph Structure

We utilize Bayesian network to represent dependent relationship between trip and zone. However, it is not clear to judge whether a node is linked to another node. We use K2 algorithm [6] which is a conventional algorithm for discovering graph structure. K2 algorithm searches an appropriate set of parent nodes for each node. The algorithm requires sets of candidate parent nodes for each node and tries to obtain optimal sets using greedy algorithm.

We use AIC as a scoring function for the search. AIC is a criterion for evaluating a model estimated by the maximum likelihood method, as follows:

$$AIC(M_k) = -2\sum_{n=1}^{N} \log p(x_n, \hat{\theta}_k) + 2p_k \tag{1}$$

where $\hat{\theta}_k$ denotes the parameter of model M_k, and p_k denotes the dimension number of $\hat{\theta}_k$ about a model set $M = \{M_1, M_2, ..., M_k\}$ and a data set $X = \{x_1, x_2, ..., x_N\}$. The value of AIC is small if the model estimates with high accuracy travel behaviors. The large dimension number fits the model the sample data too much. This problem is called overfitting. With the second term, AIC penalizes the dimension number and prevents the overfitting.

Generally, it is supposed that the nodes have a linear ordering to operate K2 algorithm. For each node, the set of its candidate parent nodes is built incrementally from the nodes which precede it in the linear ordering. Interpreting the selection of destination order as the linear ordering, we execute K2 algorithm.

4 Method for Estimation

Our method has three steps (Fig.1). Firstly, we calculate probabilities for each characteristic. Bayesian network is utilized to calculate them. Secondly, we assign the probabilities to all zones. Finally, we select a zone as destination from all zones according to the assigned probabilities for each zone.

In order to calculate the probabilities for each characteristic, we make use of Bayesian network shown in Fig.2. We calculate the probabilities using a production rule which is a fundamental rule of probability theory [7,8]. For instance, we assume that node "y" and node "z" are parent nodes of node "x". In this case, a probability of "x" is calculated recursively as follows:

$$p(x) = \sum_{z \in Z} p(z) \sum_{y \in Y} p(y)p(x|y, z) \tag{2}$$

Referring to CPT of node "x", we calculate $p(x|y, z)$.

In the second step, we assign the probabilities for each characteristic to all zones. A probability assigned to one zone is calculated as a joint probability of probabilities for each characteristic the zone has. Moreover, if some zones have the same characteristics, we assign the probability to the zones uniformly. For instance, if there are two zones which have the same number of schools and stations and the same distance from an origin zone, a half of selected probabilities are assigned to each zone.

5 Experiment

We show that travel behaviors are estimated with structural data of city using the proposed method. Therefore, we compare predicted performances in three

Table 1. Information which is available in three environments

	departure time	origin zone characteristics	personal attributes of trip characteristics	others of trip characteristics
Environment1	available	available	unavailable	unavailable
Environment2	available	available	available	unavailable
Environment3	available	available	available	available

environments where their available information is different respectively. We define two indices which are explained in Section 5.1 as the predicted performance.

Available information in the environments is shown in Table 1. Personal attributes denote *age*, *gender* and *job*. In addition, others denote *purpose* and *way*. If the information is available, observed and aggregated values are set in Bayesian network as evidence variables. Env.1, which is short for Environment1, simulates the target environment and Env.2 and Env.3 are ones for comparison with Env.1. The performance in Env.1 is essentially lower than those in others because the available data is restricted. The aim of this experiment is to show how the performance in Env.1 is close to those in others.

5.1 Evaluation Indices

Generally, several indices can be used to validate our model. We measure differences between observed and predicted distributions, and accuracy of predicted result. In order to measure them as the performance of the model, we use two following indices.

ARV. ARV is the average relative variance. This is normalized MSE (Mean Square Error), and often used to validate prediction models. In our study, we define it as follows:

$$\begin{aligned} MSE_t &= \frac{1}{|Z|^2} \sum_{o \in Z} \sum_{d \in Z} x(t,o)^2 (p(d|o,t) - \hat{p}(d|o,t))^2 \\ ARV_t &= \frac{MSE_t}{\sigma^2} \\ &= \frac{\sum_{o \in Z} \sum_{d \in Z} x(t,o)^2 (p(d|o,t) - \hat{p}(d|o,t))^2}{\sum_{o \in Z} \sum_{d \in Z} x(t,o)^2 (p(d|o,t) - \bar{p}(d))^2} \end{aligned} \tag{3}$$

where $x(t,o)$ denotes the number of observed trips from zone o at time t. $p(d|o,t)$ denotes an observed probability of trip toward zone d given all the trips from zone o at time t, and $\hat{p}(d|o,t)$ denotes a predicted probability. $\bar{p}(d)$ denotes an average probability of observed trip toward zone d through a whole day. ARV is 0 if the predicted distribution exactly equals to observed distribution, and is 1 if the model has standard performance. The standard performance means the performance of a model which always predicts average probability regardless of departure time and origin zone. The lower ARV is, the higher the predicted performance is.

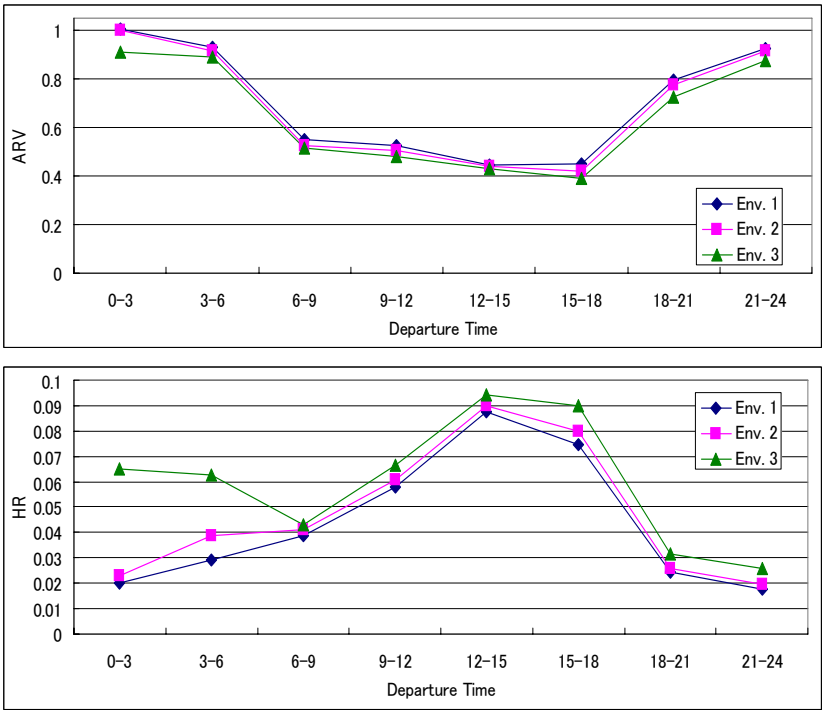

Fig. 3. Experimental results about ARV and HR

HR. HR is hit ratio. It is the ratio of the number of correctly predicted trips to the number of total trips. Unlike ARV, the higher HR is, the higher the predicted performance is.

5.2 Experimental Results

The results are presented in Fig.3. Upper graph shows the value of ARV, and lower graph shows the value of HR in tree environments through a whole day. Horizontal axis of the graph is departure time and vertical axis is value of the index at the departure time.

ARV in Env.1 is larger than that in Env.2 and HR in Env.1 is smaller than that in Env.2 at any time. Both these mean that the performance in Env.1 is less than that in Env.2. However, the difference between Env.1 and Env.2 is a few. In addition, the most difference is 0.04 about ARV and 0.01 about HR at 15-18. Available data in Env.1 is only structural data of city and those in Env.2 and Env.3 also include survey data of travel behaviors. Therefore, the result describes that travel behaviors are estimated efficiently only with the structural data.

However, the performance is remarkably low at 0-3, 3-6 and 21-24 when residents do not travel actively. We consider that this is because of the lack of sample

data. The number of trips is few at the time. Therefore, we consider the sample data insufficiently. In addition, the performance is higher at 0-6 in Env.3 than those in Env.1 and Env.2. This means that it is difficult to predict trip purpose and way to travel especially at the time. This is because that at the time people behaves randomly in comparison with at other time. Moreover, HR is quit small in this experiment. This is because zone characteristics are not sufficient to select one from 259 zones. Therefore, we have to explore more adequate characteristics.

6 Conclusion

In this paper, we proposed a probabilistic method for estimating travel behaviors using zone characteristics. The method is based on a supposition that the travel behaviors are dependent on city structure. With the method, the dependent relationship is represented as Bayesian network. The experimental results show the effectiveness of our method in environments where structural data of city is only available. In our future work, we must explore more adequate characteristics and apply our Bayesian network to other city.

References

1. Uesugi, K., Mukai, N., Watanabe, T.: Optimization of Vehicle Assignment for Car Sharing System. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1105–1111. Springer, Heidelberg (2007)
2. Yamamoto, K., Uesugi, K., Watanabe, T.: Adaptive Routing of Cruising Taxis by Mutual Exchange of Pathways. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 559–566. Springer, Heidelberg (2008)
3. Kitamura, R.: Applications of Models of Activity Behavior for Activity Based Demand Forecasting. In: Activity-Based Travel Forecasting Conference Proceedings (1997)
4. Mozolin, M., Thill, J.C., Usery, E.L.: Trip Distribution Forecasting with Multilayer Perceptron Neural Networks: A Critical Evaluation. Transportation Research Part B 34(1), 53–73 (2000)
5. Zhou, Q., Lu, H., Xu, W.: New Travel Demand Models with Back-Propagation Network. In: Proc. of ICNC 2007, vol. 3, pp. 311–317 (2007)
6. Cooper, G.F., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine learning 9(4), 309–347 (1992)
7. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs (2002)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)