Transit Route Choice Decision-making Using an Integrated Bayesian Statistical Inference

Project proposal for University of Virginia SYS 6014 Decision Analysis Spring 2020

Zechen Hu – Updated April 20, 2020

INTRODUCTION

Smart card automated fare collection systems are being used more and more by public transit agencies. While their main purpose is to collect revenue, they also produce large quantities of very detailed data on onboard transactions. These data can be very useful to transit planners, from the day-to-day operation of the transit system to the strategic long-term planning of the network. Since travel information provided by smart card is on individual level, it brings an opportunity of understanding individual travel behavior and improve the accuracy of existing transit route assignment model.

According to this data, we have a chance to make a decision-making model on the individual level. Since we have all the public transit data in Incheon, we can use those data to build a predictive model and generate the probability of each transit route in a route choice set. Here the route choice set is a set including all alternative routes for a certain OD (origin to destination) pair which also implies the set of options for decision-maker. Our decision-maker is the passenger. And then apply the decision-making analysis which in this study is based on transit utility function. In this study, we choose the origin of Bupyeongu Office and the destination of Onsu for the OD pair as our research object.

MODEL OF THE DECCISION PROBLEM

Problem Initialization

To specify the stated problem, assume a pair of origin and destination as showed in figure 1. The information provided by smart card data only includes tap-in and tap-out time, but cannot match to specific routes. Since travel time can be inferred from the smart card data, the main objective of this study is to find out if through travel time, the individual route choices or the probability of it could be derived.

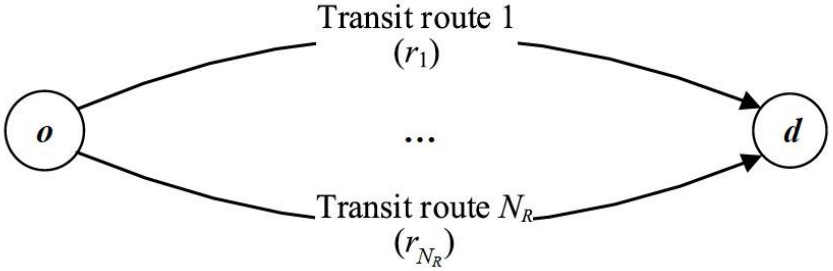

Figure 1 Illustration of Origin-Destination and Route Alternatives

The smart data comes from a Korea transit operation company named T-money. A sample of data is showed below, which include information of origin, destination, travel time, route number, and cost for each single card use.

First_stop	Last_stop	Unlinked_Trip	Length	Time	Ttime	Chain	mode1	route1	route1_obs	vehicle1
2755	2801376	2	8198	23	29	Y	203	20000000		.
2801394	2755	2	7336	23	25	Y	480	28206004	1	128717257
2755	2801417	2	7788	26	30	Y	203	20000000		.
2801394	2755	2	7336	26	29	Y	480	28206004	1	128717254
2755	2756	1	1000	16	16	Y	203	20000000		.
2801636	2801384	1	4083	16	16	Y	480	28217002	1	128717066
2801394	2755	2	7336	27	33	Y	480	28206004	1	128717253
2756	2801631	2	6724	26	34	Y	203	20000000		.
2801394	2755	2	7336	21	27	Y	480	28206004	1	128717252
2755	2801417	2	7788	33	41	Y	203	20000000		.
2755	2801376	2	8198	27	32	Y	203	20000000		.
2801394	2755	2	7336	21	24	Y	480	28206004	1	128717251
2755	2756	1	1000	4	4	Y	203	20000000		.
2756	2801384	2	7119	31	38	Y	203	20000000		.
2801394	2755	2	7336	20	23	Y	480	28206004	1	128717252
2801379	2756	2	7176	21	25	Y	476	28030014	1	128721340
2756	2801376	2	7198	21	26	Y	203	20000000		.
1816	1806	1	4400	10	10	Y	202	20000000		.

Figure 2 A Sample of Smart Card Data

To achieve better benefit of this study, the proper origin and destination selected for this study should have route alternatives that compete with each other instead of having one option is obviously better than the rest. With this consideration, the origin of Bupyeongu Office and the destination of Onsu in Incheon, Korea were selected, as showed in Figure 3. There are two route options as showed, 1) the orange one has longer distance but without transfer, while 2) the yellow one has shorter distance but will need to transfer. A week of data is available and will be used in the model analysis in the following sections.

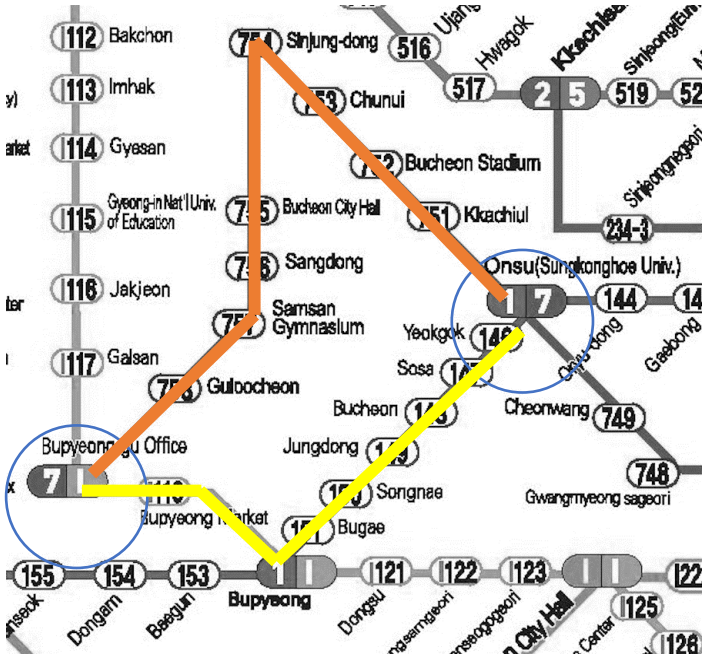

Figure 3 Origin and Destination Locations and Route Alternatives

Decision Analysis

Back to this transportation scenario, we have the information about the average time consumption for each route, the transit fare for each route and the posterior probability of an individual passenger choosing a specific route, conditional on the observation of the passenger's journey time. The cost set here is including time cost and money cost. Our utility function is generally defined as follows:

$$EU(r_i) = -\Pr(choice_{qr}|t_q) \cdot (t_q + fare_q)$$

Our utility value is estimated in terms of the time consumption and the ticket cost. For the decision-maker, which is the passenger here, is definitely going to make his/her decision and choose to go through a cheaper and faster path for his/her trip.

PREDICTIVE MODEL

Bayesian inference method applied to the question mentioned above was first introduced by Richardson and Green (1). Rephrasing our objective in Bayesian language is that what is the probability of an individual passenger choosing a specific route, conditional on the observation of the passenger's journey time. In conformity with Bayes' rule, the translated equation including such relationship is showed below:

$$\Pr(choice_{qr}|t_q) = \frac{Pr(choice_{qr}) * \Pr(t_q|choice_{qr})}{Pr(t_q)}$$

Where,

$\Pr(\mathrm{t_q}|\mathrm{choice_{qr}})$ represents a likelihood that the observed travel time would be $\mathrm{t_q}$ given an evidence of route r being actually chosen by the passenger;

$\Pr(\mathrm{choice_{qr}})$ represents the prior probability, which reflects intrinsically the possibility that, among the population Q, route r itself could potentially be used relative to other alternatives.

Additionally, based on the law of total probability, $Pr(t_q)$ is included for normalization, thus the above equation could be simplifying to the following:

$$\Pr(choice_{qr}|t_q) \propto Pr(choice_{qr}) * \Pr(t_q|choice_{qr})$$

In such way, the solution of our problem focuses on two terms: one, the prior probability

$\Pr(\mathrm{choice_{qr}})$; and two, the likelihood function $\Pr(\mathrm{t_q}|\mathrm{choice_{qr}})$. The following sections will discuss the method of solving these two terms in details.

Mathematical linkage between the problem and the method(s)

Assuming the travel time for a given route between a pair of origin and destination follows a Gaussian distribution, the proposed model is a weighted sum of all possible alternatives by mixing

probabilities. Let m(t) donate the proposed mixture Gaussian distribution, and t is the travel time between a pair of origin and destination. The number of route alternatives is N, and the travel time distribution of each route is $c_i(t;\theta_i)$. The mixture Gaussian is formulated as followed:

$$m(t) = \sum_{i \in R} \omega_i * c_i(t;\theta_i);$$

$$\sum_{i \in R} \omega_i = 1$$

With sample smart card data of observations of individual travel time, an empirical mixture distribution can be gained. However, little knowledge is known about the component distributions. If given more information to each transit service, their average travel time is supposed to be distinguishable. With properly identified component distributions, for any given observation, the posterior probability could be given by:

$$\Pr(choice_{qr}|t_q) = \frac{\omega_i * c_i(t;\theta_i)}{\sum_{j \in R} \omega_j * c_j(t;\theta_j)}$$

As $c_i(t;\theta_i)$ is not known, in this study, it is assumed two types of distribution, Gaussian and lognormal. Then the question shifted to find out the parameters of each distribution. On the basis of existing data of travel time observations, those parameters could be figured out if the route-specific travel time data can be completely identified from among pooled observations. In practice, this may not be the case. In this context, the Expectation-Maximization (EM) algorithm is widely acknowledged to be an effective approach to estimate the parametric model by fitting observation data. It is an iterative algorithm implemented by the following 4 steps:

i. Initialize parameters;

ii. For 'Expectation' (E-step), calculate $\Pr(choice_{qr}|t_q)$ using intial parameters;

iii. For 'Maximization' (M-step), evaluate and update it to be new values;

iv. Repeat (ii) and (iii) to converge to an optimum solution.

For initializing the parameters, for instance, a set of well-qualified starting values could be derived from K-means clustering. It partitions all the actual observations into a specified number, 'K', of 'clusters' each having a centroid, i.e. a mean, where $K \geq 2$. In the context of passengers' route choices that being discussed in this paper, a cluster should be equivalent to a group of passengers who chose the same transit route of which the mean travel time is considered the centroid. That is, each observation of travel time would be assigned to one cluster, and the member observations are supposed to be tightly close to its mean. This could be achieved by minimizing the sum of all the mean squared errors between the members and their centroid, over all clusters.

REFERENCE

Lijun Sun, Yang Lu, Jian Gang Jin, Der-Horng Lee, Kay W. Axhausen. An integrated Bayesian approach for passenger flow assignment in metro networks, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2015.01.001.

Richardson, S., and P. J. Green. On Bayesian analysis of mixtures with an unknown number of 33 components (with discussion). Journal of the Royal Statistical Society: Series B (Statistical 34 Methodology), Vol. 59, No. 4, 1997, pp. 731-792.

Yongsheng Zhang, Enjian Yao, Junyi Zhang, Kangning Zheng. Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2018.04.019.