Estimation of Origin-destination matrices using coarse-grained mobile phone data

Qian GE[1], Daisuke FUKUDA[2]

[1], Doctorate Student, Dept. of Civil Eng., Tokyo Institute of Technology
(12-1 , Ookayama 2, Meguro-ku, Tokyo 152-8552, Japan)
E-mail: ge.qian@plan.cv.titech.ac.jp
[2]Member of JSCE, Associate Professor, Dept. of Civil Eng., Tokyo Institute of Technology
(12-1 , Ookayama 2, Meguro-ku, Tokyo 152-8552, Japan)
E-mail:fukuda@plan.cv.titech.ac.jp

In this study, we propose a methodology to estimate and update OD matrices using coarse-grained (i.e. aggregated by mesh) mobile phone-based population data. The dataset we use has no individual positioning information but alternatively can describe the variation of travel demand while causing no legal issue. Based on past studies that try to estimate and update OD matrices using traffic counts, we develop a variation of the general model framework. An estimator in which relative entropy is adopted to describe divergence of matrices has been put forward to update OD matrices. Empirical study is being conducted in Tokyo using mobile phone data of one day in 2012, in conjunction with travel survey of the year 2008.

Key Words: *OD estimation, entropy, mobile phone data, coarse-grained,*

1. Introduction

The problem of origin-destination (OD) matrices estimation had been studied for several decades. Traditional approaches usually estimate OD matrices through large-scale travel survey. The person-trip (PT) survey in Japan is an instance. PT survey is conducted every 10 years and collects detailed person attributes and record all trip purposes, duration and destination of about 1% of total household members in each city of Japan on some particular day. It is obvious that the survey is too costly to conduct and the data becomes outdated soon because tremendous data collection and processing work should be done.

Since the 1970s, many approaches have been developed for estimating or updating the prior matrix using traffic counts collected on some links of the network, or partial network. As depicted in Cascetta (2013) [1], studies in this line have been extended from estimation or updating OD flows in static system to with-in day dynamics, and then further to incorporate on-line data with off-line data to get filtered data. Several thorough reviews of approaches in this area may be found in Bera and Rao (2011) [2] and Cascetta *et al.* (2009) [3].

On the other hand, pervasive computing devices such as smart phone, tablet and other GPS applications have become indispensable to most citizens. Those devices can provided location information of individuals in unprecedented detail. It is not a new idea to try utilizing detailed or simple mobile phone data to for traffic management and control. For the specific goal of measuring OD flows, different types and fineness of dataset help to bring about different approaches to estimate traffic flow. In other words, estimating OD matrices from mobile phone location data is a data-oriented and data-intensive study.

Billing data in association with cell phone tower information when a phone received a message or make a phone call was firstly used by White *et al.* (2002) [4]. With-in dynamics had been studied by location data of mobile phones every two hours in Caceres *et al.* (2007) [5]. Location update (Pan *et al.*, 2006) [6] and cell phone tower handover information (Sohn *et al.*, 2008) [7] were employed to infer individual travel movement and then aggregated to OD matrices.

However, these studies focused on individual location data, which always has good performance in art but intangible to practice. Moreover, public anxiety have become intensified over privacy violation in the past several years. Estimating traffic flow using dataset that does not contain any individual location information is an alternative to the past studies. On the basis of these facts, we attempt to develop OD estimation approaches using aggregated mobile phone data to estimate and update prior ma-

trix in this article.

This paper is organized as follows: Section 2 describes the mobile phone dataset to use. Section 3 introduces the theoretical foundation and mathematical properties of the general framework. An OD estimation approach that incorporates population with activity-based approach will be proposed in Section 4. In the end of this article, we will draw some conclusions.

2. Mobile Phone Dataset

(1) Coarse-grained Mobile phone data

Coarse-grained mobile phone data is several types of location dataset in relatively coarse granularity. This concept is brought about in González *et al.* (2008) [8] when studying the movement of people with call, dial record (CDR) data. This dataset records individual's location when he sends a message, makes a phone call or connects the internet. Calabrese *et al.* (2011) [9] tries to utilize the same dataset to estimate OD matrices by reconstructing the travel trajectory of individuals, and aggregating the number of trajectory by OD pairs.

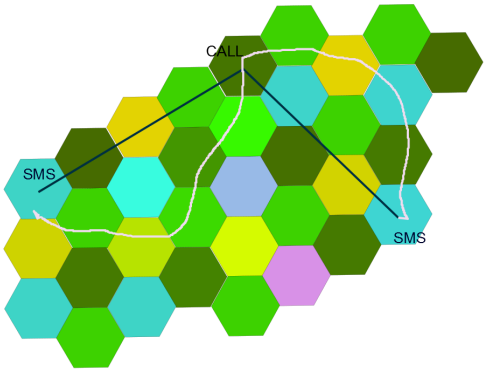

Fig.1 Reconstruction of Individual's Trajectory Using CDR **Data (black line shows the estimated trajectory a trip by mobile phone data, grey line shows the actual trajectory)**

However, this dataset is still too "fine" when we considered people's privacy concern. Location data without individual information is the dataset we attempt to use in this study.

(2) Aggregated Population Data

Our methodology uses aggregated mobile phone data, or population data to estimate dynamic OD matrices.

The raw dataset provided by a major mapping corporation of Japan records individual's location every one hour. This dataset is a fine-grained description of individual's movement in the metropolitan area. To avoid privacy issue, the dataset has been processed to become aggregated population of several specific groups of people in mesh level. The size of each mesh is 250*250 m^2. Each mesh covers several blocks. This aggregate dataset covers information as follows:

1. Total number of people who are staying at home at the recording time (AT-HOME);
2. Total number of people who are staying in their workplace at the recording time (AT-WORK);
3. Total number of people who are moving in the mesh at the recording time (MOVING);

The interval of recording is one hour. The dataset records location information start from 0:00 and end at 23:00. In total, we had 24 records for each mesh.

Each data item in the dataset is characterized by the maximum and minimum of longitudes and latitudes of the area, and a time stamp to show the time slice when it is recorded.

Individual location is collected by GPS devices. The precision of the raw dataset is generally considered better than location data by cell tower. However, sampling bias may introduce error to the estimators.

Fig.2 shows the distribution of people who stay at their workplace in Tokyo on Dec 30, 2012. This figure is composed by mapping mobile phone data to the corresponding mesh. Different colors show the number of people in each mesh. The AT-WORK population of meshes become denser with it color turns from blue to red.

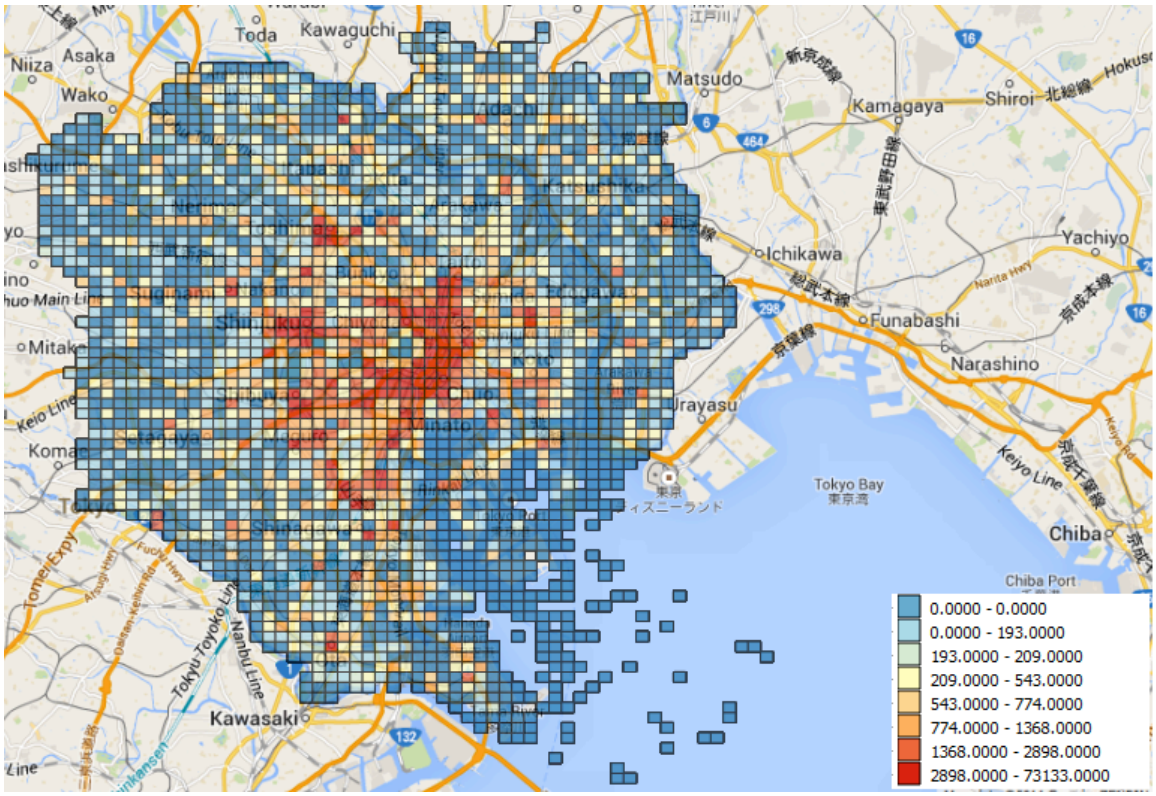

Fig.2 AT-WORK Population of Tokyo

3. Problem Formulation

Variables and denotations for the proposed model are listed as follows:

$\mathbf{P}^*$	the optimal OD matrix
$\hat{\mathbf{P}}^t$	the target OD matrix in time slice t
$\mathbf{P}^t$	feasible set OD matrix in time slice t
$\hat{\mathbf{C}}^t$	estimation moving population of in time slices t from mobile phone
$\mathbf{C}^t$	vector of moving population of in time slices t
$\hat{c}_r^t$	estimated moving population of mesh r in time slices t from mobile phone
c_r^t	moving population of mesh r in time slice \$t\$
$\hat{\mathbf{v}}^t$	estimated traffic flow of links in time slice t from mobile phone
$\mathbf{v}^t$	traffic flow of links in time slice t
$\hat{v}_l^t$	estimated traffic flow of link l in time slice t from mobile phone
v_l^t	traffic flow of link l in time slice t

Cascetta (1988) [10] had shown that most OD demand static estimators can be generalized in a framework of constrained optimization problems and extended it to dynamic cases in Cascetta (1993) [11]. This framework is in this form:

$$\mathbf{P}^* = arg\ \min[f_1(\mathbf{P}^t, \hat{\mathbf{P}}^t) + f_2(\mathbf{v}^t, \hat{\mathbf{v}}^t)] \quad (1)$$

The most studies focus on specify functions $f_1()$ and $f_2()$. The prior matrices $\hat{\mathbf{P}}^t$ are obtained by comprehensive historical survey or simple survey.

Accordingly, the problem in this article shares a similar form:

$$\mathbf{P}^* = arg\ \min[f_1(\mathbf{P}^t, \hat{\mathbf{P}}^t) + f_2(\mathbf{c}^t, \hat{\mathbf{c}}^t) \quad (2)$$

After finding out the relationship between population data and traffic flows and selecting indices to describe the divergence of the prior and posterior traffic flow matrices, we may get an estimation or update method for calibrating OD matrices. For some cases, for example, the commuting trips, the relationship may be not difficult to find out (see Section 3).

4. A Entropy-based Estimator of Travel Demand

In this section, we take the commuting trips as an instance to describe an estimator in the general framework of Equation 2.

(1) Data processing

As the first step, mesh-level population dataset should be converted to zone-level dataset. In most cities, the size of zone is usually much larger than

mesh. Assuming that people are evenly distributed in each mesh, we may calculate the zone-level AT-WORK population by aggregating the population of each mesh it covers. This conversion can be completed by spatial analysis tools. The results of this step is the AT-WORK population of zone k in time slice t, denoted by P_{kw}^{t} .

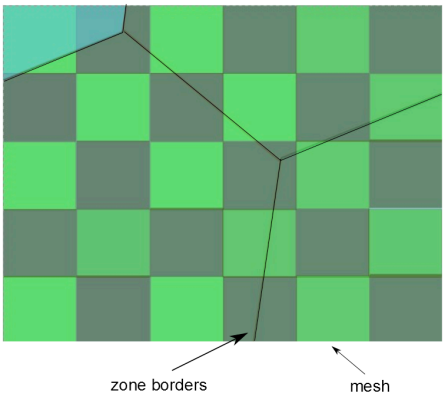

Fig.3 The Relationship Between Zones and Meshes

(2) Model Formulation

It is obvious that population difference between two consecutive time slices equals to number of trips happened in this specific zone. That means $\Delta P_{kw}^{t} = P_{kw}^{t} - P_{kw}^{t-1}$ is total number of population change in zone k during time slice t. If $\Delta P_{kw}^{t} > 0$, there are more people who have gone to their workplace in zone k than people who leave their workplace during the past time slice t and *vice versa*.

However, commuting trips are usually share common pattern. Most commuters leave their home to workplace on the morning and go back home in afternoon. That means we can approximate positive ΔP_{kw} s to the sum of commuting trips from other zones to k, i.e., $\Delta P_{kw} = \sum_{i}^{n} p_{ikw}^{t}$ and negative ΔP_{kw} s to the sum of commuting trips from zone k to other zones.

We may find a estimator of commuting travel demand by solving:

$$\mathbf{P_w}^{*} = arg\ \min f(\mathbf{P}_{kw}^{t}, \hat{\mathbf{P}}_{kw}^{t}) \tag{3}$$

s.t.,

$$\sum_{i}^{n} P_{ikw}^{t} = \Delta P_{kw}^{t} \tag{4}$$

,where n denotes the total of zones,

P_{ikw}^{t} denotes the number of commuting trips from zone i to k.

$\mathbf{P_w}^{*}$ denotes the posterior OD matrices,

$\hat{\mathbf{P}}_{kw}^{t}$ denotes the prior matrix of commuting trips.

In this model, mobile phone data is mainly used for updating the existing matrices since the population dataset is not enough to make reliable estimation of traffic flow. However, most OD matrices collected by travel survey don't have time stamp and trip purposes. They are not sufficient to be the prior matrices since commuting trips and their time stamps cannot be differentiated. To deal with this problem, we are trying to employ an activity-based framework to get the prior.

(3) Activity-based prior generation

Activity based travel demand models predict travel behavior as a derivative of activities. Therefore, by predicting which activities are performed at particular destinations and times, trips and their timing and locations are implicitly forecast in activity based models (Jovicic, 2001) [13].

The activity-based models are characterized by three features by Davidson *et al.* (2007) [14]:

1. An activity-based platform that implies the modeled travel be derived within a general framework of the daily activities undertaken by households and persons;
2. Tour based structure of travel where the tour is used as the base unit of modeling travel instead of elemental trip;
3. micro-simulation modeling techniques that are applied at the fully-disaggregated level of persons and households, which convert activity and travel related choices from fractional-probability model outcomes into a series of discrete choices.

State-of-practice activity-based approaches for travel demand can be categorized into utility-maximization models and rule-based models. Utility-maximization models have strong behavior theoretical basis. These models lies in two basic ideas. Firstly, people's need for travel is derived from their demand for activity. One will take a trip when he can gain more utility from the activity than disutility from it. Second, people face spatial and temporal constrains when move between locations (Bowman, 2001) [12]. Rule-based models usually don't consider the behavior basis but view people's choice of activities as a result of a set of heuristic rules. A thorough review of activity-based model may be found in Pinjari *et al.* (2011) [15].

In our study, we tries to apply the model described in Bowman *et al.* (2001) [12] to study the travel demand of Tokyo, since the model have been applied in cities such as Portland, Sacramento and other cities in the US.

We plan to use PT survey data of Tokyo on the year 2008 as the dataset to estimate travel pattern of each individuals in the metropolitan area. The result

of this component is individual trips with timestamp and purposes. We will extract all commuting trips from individual's activity pattern to form a prior of commuting trip distribution.

(4) Minimizing the divergence

From Equation.(3), we can utilize estimators that had been adopted in the previous studies, such as Generalized Least Square (GLS), Maximum Likelihood Estimation and Bayesian approach to estimate or updating the traffic flows. In this section, we display the estimation of cross entropy method.

The divergence of matrix is commonly describe by the relative entropy, or Kullback–Leibler (K-L) divergence of these two matrices.

For discrete probability distributions P and Q, the K–L divergence of Q from P is defined to be

$$D_{\mathrm{KL}}(P\|\,Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right)P(i).$$

The K–L divergence is only defined when P and Q both sum to one.

Dividing both side of Equation.(3) by the total number of trips, we can get a statistical form of the OD estimation problem.

$$\mathbf{p}^* = \arg\ min\ \mathbf{D}(\mathbf{p}^t, \hat{\mathbf{p}}^t) = \sum_r \hat{p}_r^t \log \frac{\hat{p}_r^t}{p_r^t} \tag{5}$$

s.t.,

$\sum_r p_r^t = 1$

$\sum_i^n p_{ikw}^t = \Delta p_{kw}^t$

$p_r^t \geq 0 \quad \forall r \in N$, N is the set of OD pairs

, where $\mathbf{p}^*$ denotes the estimated posterior of OD distribution

$\mathbf{p}^t$ denotes the vector of the distribution of traffic flows in time slice *t*

$\hat{\mathbf{p}}^t$ denotes the prior vector of the distribution of traffic flows in time slice *t*

p_r^t denotes the fraction of traffic flows of OD pair *r* in time slice *t*

$\hat{p}_r^t$ denotes the estimated fraction of traffic flows of OD pair *r* in time slice *t*

Δp_{kw}^t denotes the variation of AT-WORK population change in zone *k* in time slice *t*

This model is based Wilson's spatial interaction model of trip distribution[16].

(5) Calibration of Traffics Flow

This optimization problem describe in Equation.(5) can be solved by *Lagrangian Multiplication*. With some manipulation, we may get the solution:

$$p_r^t = \frac{\hat{p}_r^t e^{\lambda_i}}{\sum_r \hat{p}_r^t e^{\lambda_i}} \tag{6}$$

, where λ_i is the Lagrangian multiplier corresponding to OD pair *r*.

By substituting λ_i back to the constraints, we can obtain each p_r^t .

(6) A Numerical Example

Since the prior generation method and updating method are relatively independent, we may illustrate the updating model by a simple numerical example.

Considering a study area with 10 zones, and we have calculated the prior of the trip distribution in time slice t:

As shown in Fig.4, each element in the matrix is the fraction of flow

We also know the population change in each zone as in vector $\Delta\mathbf{p}$ = 0.10, 0.13, 0.07, 0.12, 0.11, 0.08, 0.09, 0.06, 0.14, 0.10}.

In other words,

$\sum_j p_{1,j} = 0.10$

$\sum_j p_{2,j} = 0.13$

… …

$\sum_j p_{10,j} = 0.10$

$\sum_{ij} p_{ij} = 1$

$p_{ij} \geq 0, \forall\, i, j$

, where p_{kj} denotes proportion of commuting trip flows from zone *k* to *j*.

Using *Newton's Method*, we may get the posterior matrix as shown in Fig. 5

5. Discussion and Conclusions

In this article, we proposed an approach to estimate OD matrix that incorporates mobile phone population data and existing model framework. The proposed methodology in this article can be viewed as a variation of OD estimation models from traffic count.

By understanding the characteristics of commuting trips, we further put forward a practical estimator of commuting travel flow that updates prior trip distribution generated from an activity-based model. Relative entropy or K-L divergence of distributions

0	0.0011	0.0081	0.0169	0.0188	0.0051	0.0012	0.0166	0.0025	0.0028
0.0062	0	0.0029	0.0073	0.0070	0.0032	0.0054	0.0114	0.0105	0.0017
0.0017	0.0165	0	0.0000	0.0227	0.0118	0.0108	0.0025	0.0226	0.0063
0.0110	0.0099	0.0044	0	0.0165	0.0179	0.0189	0.0018	0.0210	0.0105
0.0083	0.0217	0.0096	0.0075	0	0.0120	0.0205	0.0106	0.0122	0.0003
0.0012	0.0164	0.0130	0.0048	0.0090	0	0.0194	0.0185	0.0164	0.0205
0.0157	0.0139	0.0208	0.0054	0.0198	0.0108	0	0.0091	0.0192	0.0008
0.0025	0.0114	0.0012	.0040	0.0166	0.0173	0.0219	0	0.0226	0.0149
0.0013	0.0207	0.0013	0.0029	0.0051	0.0071	0.0154	0.0117	0	0.0050
0.0203	0.0223	0.0199	0.0192	0.0037	0.0189	0.0195	0.0067	0.0011	0

Fig.4 The Prior Trip Distribution of Study Area

0	0.0015	0.0111	0.0231	0.0257	0.0069	0.0017	0.0227	0.0034	0.0038
0.0142	0	0.0067	0.0168	0.0161	0.0074	0.0123	0.0263	0.0240	0.0062
0.0013	0.0122	0	0.0000	0.0167	0.0087	0.0080	0.0018	0.0167	0.0046
0.0118	0.0106	0.0047	0	0.0177	0.0192	0.0202	0.0020	0.0225	0.0113
0.0089	0.0232	0.0103	0.0080	0	0.0129	0.0220	0.0114	0.0130	0.0004
0.0008	0.0110	0.0087	0.0032	0.0060	0	0.0130	0.0124	0.0110	0.0138
0.0122	0.0108	0.0162	0.0041	0.0154	0.0084	0	0.0071	0.0150	0.0006
0.0014	0.0061	0.0006	0.0021	0.0089	0.0092	0.0117	0	0.0121	0.0080
0.0215	0.0354	0.0022	0.0050	0.0088	0.0121	0.0264	0.0201	0	0.0086
0.0154	0.0169	0.0151	0.0146	0.0028	0.0144	0.0148	0.0051	0.0008	0

Fig.5 The Posterior Trip Distribution of Study Area

is used to describe the difference of the prior and posterior. We then display the estimator can be used through a numerical model.

This model utilizes coarse-grained mobile phone and PT survey, no extra cost will occur on surveying or recording individual's location. Meanwhile, individual's privacy will not be violated since his travel trajectory cannot be inferred from the dataset.

With the aforementioned advantages and more validation works in the future, we are conducting a large-scale empirical study of updating OD matrices of Tokyo based on this proposed model. PT survey data of the year 2008 will be used to calibrate the prior trip distribution and mobile phone population data of the year 2012 will be used for updating. The results of this empirical study is expected to be presented during the conference.

REFERENCES

1) Cascetta, E., Papola, A., Marzano, V., Simonelli, F., & Vitiello, I.: Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data, *Transportation Research Part B: Methodological*, Vol 55, pp 171-187, 2013
2) Bera, S., & Rao, K. V.: Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport*. Vol 49, pp 3-23, 2011.
3) Cascetta, E.: *Transportation systems analysis: models and applications* Vol. 29, Springer, 2009
4) White, J., & Wells, I.: Extracting origin destination information from mobile phone data, *Road Transport Information and Control, Eleventh International Conference on* (Conf. Publ. No. 486) (pp. 30-34), IET, 2002
5) Caceres, N., Wideberg, J. P., & Benitez, F. G.: Deriving origin destination data from a mobile phone network, *Intelligent Transport Systems*, IET, 1(1), pp 15-26, 2007
6) Pan, C., Lu, J., Di, S., & Ran, B.: Cellular-based data-extracting method for trip distribution, *Transportation Research Record: Journal of the Transportation Research Board*, Vol 1945, No.1, pp 33-39, 2006
7) Sohn, Keemin, and Daehyun Kim: Dynamic origin–destination flow estimation using cellular communication system, *Vehicular Technology, IEEE Transactions on*, Vol 57m, No.5, pp 2703-2713, 2008
8) Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L.: Understanding individual human mobility patterns, *Nature*, 453(7196), pp 779-782, 2008
9) Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C.: Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, Vol 10, No.4, pp 36-44. 2011
10) Cascetta, E., Inaudi, D., & Marquis, G.: Dynamic estimators of origin-destination matrices using traffic counts. *Transportation science*, Vol 27, No.4, pp 363-373. 1993
11) Cascetta, E., & Nguyen, S.: A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, Vol, 22, No.6, pp 437-455, 1988
12) Jovicic, G.: Activity based travel demand modelling. Danmarks Transport For Skning, 2001
13) Bowman, J. L., & Ben-Akiva, M. E.: Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, Vol 35, No.1, pp 1-28, 2001
14) Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J & Picado, R.: Synthesis of first practices and operational research approaches in activity-based travel demand modeling, *Transportation Research Part A: Policy and Practice*, Vol 41, No.5, pp 464-488, 2007
15) Pinjari, A. R., & Bhat, C. R.: Activity-based travel demand analysis. *A Handbook of Transport Economics*, No. 10, pp 213-248, 2011.
16) Wilson, A. G.: A family of spatial interaction models, and associated developments, *Environment and Planning*, Vol 3, No.1, 1-32, 1971.