Daily O-D Matrix Estimation using Cellular Probe Data

Yi Zhang*
Department of Civil and Environmental Engineering,
University of Wisconsin-Madison,
Madison, WI 53706
Phone: 1-608-262-2524
E-mail: zhang34@wisc.edu

Xiao Qin
Department of Civil and Environmental Engineering,
South Dakota State University
Brookings, SD 57007
Phone: 1-605-688-6355
E-mail: xqin@cae.wisc.edu

Shen Dong
Department of Civil and Environmental Engineering,
University of Wisconsin-Madison,
Madison, WI 53706
Phone: 1-608-262-2524
E-mail: sdong2@wisc.edu

Bin Ran
Department of Civil and Environmental Engineering,
University of Wisconsin-Madison,
Madison, WI 53706
Phone: 1-608-262-0052
E-mail: bran@wisc.edu

* *Corresponding Author*

ABSTRACT

With the fast-growing wireless-communication market, the cellular positioning technologies are becoming one of the important means to monitoring real-time traffic status, providing traveler information, measuring system operations performance, and estimating travel demand. An innovative methodology is presented in this paper to estimate the daily O-D demand using cellular probe trajectory information. Taking advantages of the emerging cell-phone tracking technologies, the cellular probe trajectories are obtained by recording all the signal-transition events and period location update events of cellular probes to determine the trip origins and destinations. To apply the O-D estimation to a broader spectrum, the probability of cell-phone ownership was treated as a conditional probability depending on users' socio-economic factors available in the census data such as age, rage, household income, etc.. A mathematic model was designed to convert the cellular counts into equivalent vehicle counts, using the posterior information obtained from the characteristics of cellular probe trajectories. Next, the traveling population daily O-D demand was estimated via a robust Horvitz-Thompson estimator. Finally, the methodology was tested via a VISSIM simulation and results were compared with a conventional simple random sampling (SRS) method. The comparison outcome shows great potential of using cellular probe trajectory information as a means to estimating daily O-D travel demand.

Key words: Daily O-D demand estimation, Cellular probe data, Cell-phone tracking technology, Horvitz-Thompson estimator

1. INTRODUCTION

With respect to the increasing needs on the traffic demand forecasting, the estimation and prediction of O-D matrix has become an important issue in the current transportation planning and operation scope. The O-D estimation is the essential source for traffic demand information.

Generally, there are two types of the O-D estimation methods. One is the survey-based O-D estimation method, which utilize the trip survey data to generate the O-D matrix (*1, 2*). The other one is the traffic-counts-based O-D estimation method, which uses the observed link traffic counts to reversely derive the O-D matrix (*3-5*).

Traditional survey-based trip diary approach to estimating trip generation and distribution is time-consuming and cost-prohibitive (*1, 6*). The estimation may vary from one study as a result of the limitation of the survey sample size and sampling randomness. The counts-based methods used the existing traffic devices such as loop detectors and video cameras to obtain the link traffic counts. The O-D matrix is derived by an opposite way of traffic assignment (*3*). But the naturally most of the models are underdetermined (*3, 7*).

In recent years, some positioning technologies such as GPS and cell phone emerged to be used to monitor real-time traffic status(*8, 9*), provide traveler information (*10, 11*), and estimate travel demand (*12-16*). With the popularity of cell phone and emerging cell-phone tracking technologies, using cellular probe data have the great potential to provide a larger sample size in a timely manner.

Pan et al. *(12)* proposed a method to record the cell-phone positions every 2 hours and aggregated them to obtain the trip distribution between each O-D pairs. Caceres *(13)* proposed another method to record the Location Updates (LU) events to count the O-D trip flows between each Location Area (LA) and convert the cell-phone counts into vehicle counts. Sohn and Kim *(14)* developed an idle Handoff (HO) technology for cell-phone positioning to get the "virtual" traffic counts on observation links, and use the synthetic method to derive the time-dependent O-D matrix from the link traffic counts.

There are three major limitations existing in current literatures. The first one is the signal-transition events are not fully used. Since the LA includes tens of cells, and its coverage is much larger than cell, sometimes, the data fusion of the LU and HO events will increase the complexity of the O-D estimation problem. Most of literatures use either the HO events or the Location Update events (Includes periodic location update (PLU)) to determine the trajectories of cell phones or cell-phone counts. The limitation of using only one type of transition events is that it only records a part of information of cellular probe trajectories, in which case it may lead to inaccurate positioning results.

The second one is that the socio-economic difference of cell-phone owners is omitted. Considering the cell-phone owner group as a sample selected from the population, naturally the sample can be treated as a simple random sample. This is the so-called "Simple Random Sampling" (SRS) strategy. Most of literatures adopted this method *(12, 13, 15, 16)*. However, whether a person owns a cell phone depends on several important factors, such as age, household income, race etc.. Disregarding the difference among those factors may leads to socio-economic "bias".

The third one is that cell-phone counts are not properly converted into vehicle counts. As we know, aggregation of the cellular probe trajectories will return the cell-phone counts. However, in transportation area, the interests mainly focus on the vehicle counts. Typically, the cell-phone counts are not equal to vehicle counts, since different vehicles may carry different number of cell-phone owners. It needs to be converted before it can be used. In current literatures, this problem is either omitted (*10-12, 14*), or treated by predefining an equivalent factor to do the conversions *(13)*.

This paper proposed a method trying to cover the above three limitations. The method uses the full information on the signal-transition events to produce cellular probe trajectories. Also the socio-economic factors are taken into consideration to generate the probability of cell-phone ownerships. Then, the vehicle counts are aggregated by using the characteristics of the cellular probe trajectories.

This paper is organized as follows: The section 2 introduce the cell-phone tracking technology; The section 3 introduce the proposed method to estimate the daily O-D demand; The section 4 gives a simulation based experiment to demonstrate and verify the proposed method; The last section gives out the major conclusion of this paper.

2. CELL-PHONE TRACKING TECHNOLOGY

The cell-phone tracking technology uses the signal transition between two conterminous cells to determine the location of the object (*10*). Signal transition refers to a phenomenon that some parameters change their values at some “virtual” boundaries of its defined location region. In practice, a cell size and boundary changes with time due to the fluctuation of signal coverage.

Generally, in the GSM network, the parameters which can be used to track the signal transitions are Location Area Code (LAC), serving cell ID (Cell ID) and Timing Advance (TA). The corresponding signal-transition events in GSM network are Location Update (LU) for LAC, HO for Cell ID and the transition of TA values, respectively (*17*).

When a cell-phone with an on-going phone call crosses the boundary of different cells, a HO operation, in which the cell id and the time stamp are recorded automatically by the system, will be executed. If the cell phone is turned on but not on call, a LU event will be automatically recorded when it crosses the boundary of different Location Areas (LA). The timing advance is used to compensate for the time that takes a wireless signal to travel at the speed of light between a Base Transceiver Station (BTS) and the cell phone (*17*). Multiplying TA and 550 meters can give the minimum distance to a BTS. The maximum distance will be (TA+1) multiplying with 550 m. Similar to a HO, the timing advance transition can only be collected when a cell phone is in on-call mode.

In addition to the signal-transition events, the cellular system also provides a periodic location update for the cell ID information *(12)*. Generally, the cellular system will update each cell phone’s cell IDs periodically and add it into a Database. Here the location information provided by this event is the cell ID and timestamp. This event is called Periodic Location Update (PLU), and the length of the period can be adjusted by the mobile carrier. Usually, the update period is set to 2 hours by default.

Table.1 Typical signal-transition events in Figure.1

Events	Area	Timestamp

PLU	Cell 1	08:00
TA	Cell 1	08:33
HO	Cell 1 –> Cell 2	08:38
LU	Cell 8 –> Cell 11	08:59
PLU	Cell 12	10:00
PLU	Cell 12	12:00
PLU	Cell 12	14:00
PLU	Cell 12	16:00
HO	Cell 14 –> Cell 16	17:09
TA	Cell 16	17:16
LU	Cell 13 –> Cell 10	17:45
PLU	Cell 4	18:00
PLU	Cell 4	20:00

Combining the above cellular location technologies, the cellular probe trajectories can be obtained by recording the signal-transition (HOs, LUs and TAs) and the PLU events. Figure.1 gives an example to illustrate the process of cellular tracking method. Assume a cell phone starts traveling at Cell 1. Its trajectory can be tracked by the signal transitions and the PLUs. Table.1 describes the different events recorded by the system.

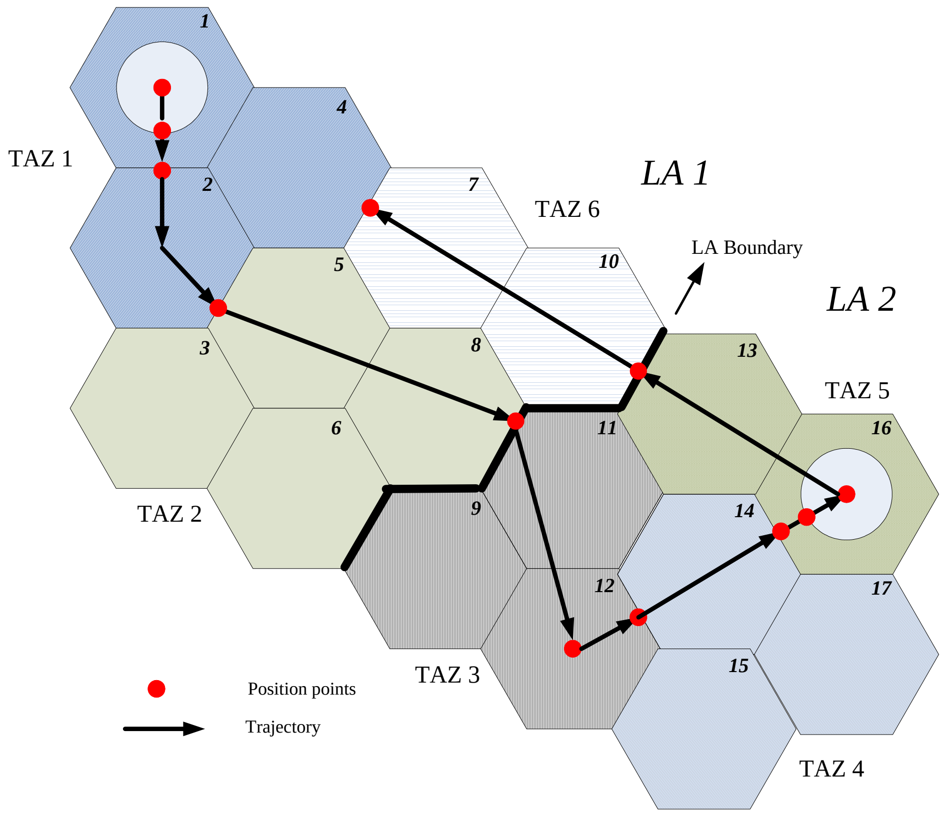


Figure.1 Illustration of cellular tracking technology

The cellular tracking method provides a possibility to record the Origin-Destination information by analyzing the trajectory of cell-phone users. Gur Y.J. et. al. *(18)* considered the location of the first signal transition event (mostly is the event that the first time turn on the cell phone in the morning and register to GSM network) as the trip origin. In this paper, we adopted this method to identify the trip origins. The problem is how to decide the trip end. One possible solution is to consider the TAZs (Transportation Analysis Zone) with the longest distance from the trip origin and most PLUs recorded as the destinations. Taking the Figure.1 as an example, the trip starts at TAZ 1 since the first events happens at TAZ 1. According to Table.1, TAZ 3 has the longest distance and it has the longest duration in a day. It is easy to conclude that TAZ 3 is the trip end.

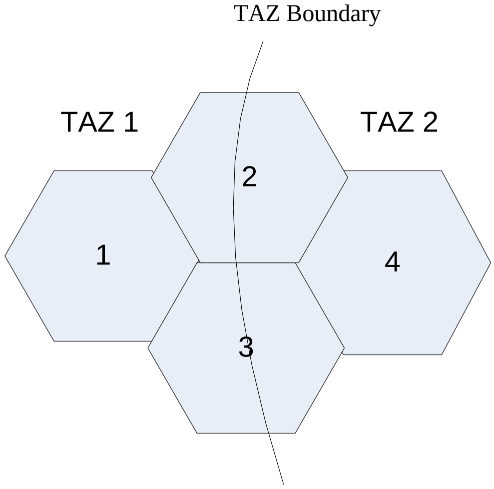


Figure.2 Illustration of imperfectly overlapping between TAZs and cells

Typically, TAZs will not match the boundaries of cells perfectly. A cell may be covered by multiple TAZs, meanwhile a TAZ may cover one or more cells entirely or just a part of a cell. As shown in Figure. 2, TAZ 1 and TAZ 2 cover the entire cell 1 and cell 4 respectively, and share the cell 2 and cell 3. If there is no signal transition event, the system can only tell the cell ID information by the PLU events, which may cause spatial errors because it is hard to determine which TAZ the cell belongs to. Pan et al. *(12)* used a probability of one cell belonging to a TAZ according to the proportion of area covered in each TAZ to determine it is covered by which TAZ, if there is no additional demographical information provided. We adopted this method in this paper.

3. METHODOLOGY

3.1 Study Design

The signal-transition events and PLUs associated with the corresponding cell-phone ids and timestamps can be collected and stored in a database at the operating center of the cellular carrier. It is easy to get a specific cell-phone owner’s trajectory by just doing a query in the database. The proposed method will first generate the individual cellular probe trajectory, in which the cell-phone signal transition and PLU events are recorded to form the trajectory.

After the collection of cellular probe trajectories, the identification of trip origins and ends will be executed. Then a vehicle-per-cellphone equivalent factor will be calculated to covert the cell-phone trips into vehicle trips, since a vehicle may carry different number of cell-phone owners. Till now, what we got is the cellular trips and the corresponding equivalent vehicle trips. However, those trip makers who don’t own cell phones should also be considered. A data aggregation process will be carried out to project the sampled vehicle O-D matrix to the population vehicle O-D matrix. As a result, the actual O-D matrix can be obtained following these above procedures. Figure.3 shows the procedures of the O-D estimation method.

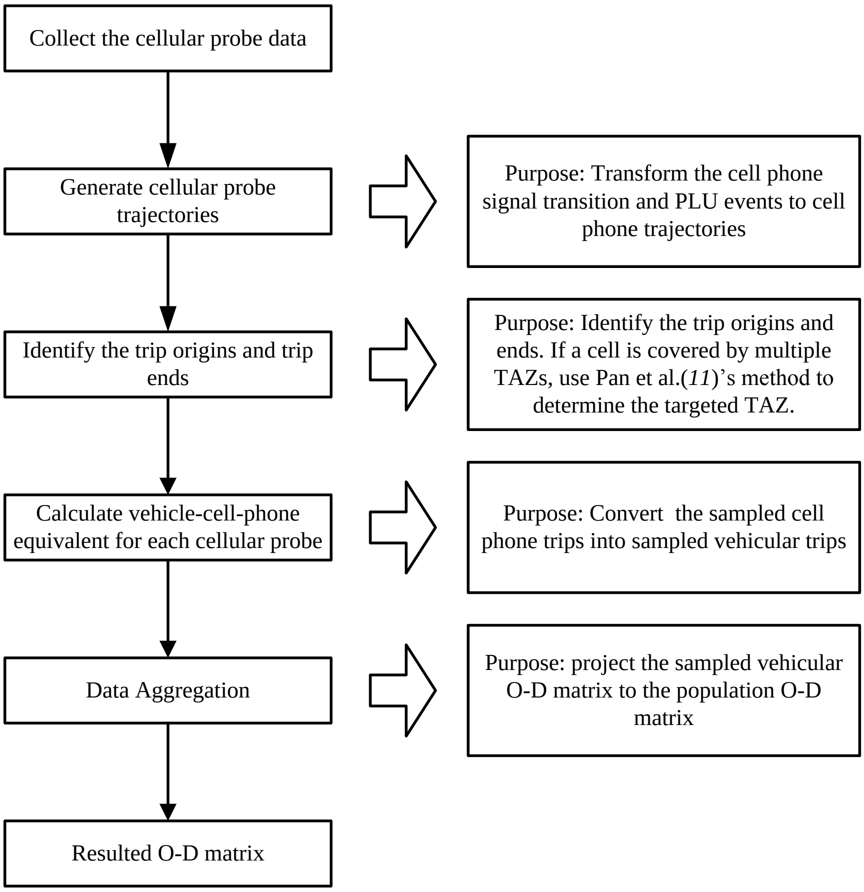


Figure.3 Flow chart of the estimation process

Notations:

$p(cp \mid f_1, f_2, \dots f_n)$ - the conditional probability of cell-phone ownership depending on factors $(f_1, f_2, \dots f_n)$

$p(cp \mid f_n)$ - the conditional probability on factor f_n

p_{c_m} - the market share of a specific carrier m

$p(cp)$ - the market penetration of cell phones

f_{vc} - the vehicle-per-cellphone equivalent factor

$\overline{f}_{pvo}$ - the average passenger-vehicle occupancy

N_i - the total population in TAZ i

T_{ij} - the O-D flows from TAZ i to j

$\hat{T}_{ij}$ - the estimated value of T_{ij} of SRS method

$\tilde{T}_{ij}$ - the estimated value of T_{ij} from cell-phone owner group of our method

S^i - the cell-phone owner group in TAZ i

p_k^i - the posterior probability of cell-phone ownership for one or several carriers of the k-th people in TAZ i. It may vary in terms of ages, income and sex.

Y_k^{ij} - the indicating variable, 1 if k -th people in population has a trip between TAZ i and j, otherwise 0.

y_k^{ij} - the indicating variable. 1 if k -th people in cell-phone owner group has a trip between TAZ i and j, otherwise 0.

P_{ij} - the proportion of people have trips between TAZ i and j

$\hat{P}_{ij}$ - the estimated value of P_{ij} of SRS method

$\tilde{P}_{ij}$ - the estimated value of P_{ij} from cell-phone owner group

3.2 Important Assumptions

Before introducing the daily O-D demand estimation method, two important assumptions should be made in order to make the cell-phone tracking method can be used to estimate the O-D demand between TAZ pairs.

1. *There might be multiple cellular carriers existing in the research areas. Each of the carriers is operated independently.* It means that the owner groups and the signal coverage of each cellular carrier are independent. In this case, each cell-phone owner group which belongs to a specific cellular carrier can only be treated as individual sample set.
2. *The cell-phone ownership pattern is identically distributed among different cellular carriers*. That means different cellular carriers have the same distribution of their owner's age, income, race etc.. This assumption holds in the general conditions, although there are some cellular carriers having different distribution in terms of the subscriber's demographics, such as MetroPCS has a heavy emphasis on prepaid phone plans, Nextel had a strong business focus.

The first assumption guarantees the generation of cellular probe trajectories and the identification of trip origins and destinations can be carried out independently among various cellular carriers. After the trip origins and destinations are determined, the difference on the signal coverage of different carriers will no longer influence the accuracy estimation results. The second assumption guarantees the generation of the cell-phone ownership probabilities can be applied to multiple carriers.

3.3 Determine the Probability of Cell-phone Ownership

In traditional trip survey methods, the SRS strategy are generally adopted to design the trip surveys, in which at most 5% sampling rate are used to get the unbiased estimation of trips in a TAZ. But sometimes a lower sampling rate makes the sampling error intolerable.

The cellular probe data gives another way to aggregate the trip data because of its unique advantages:

1. Cellular probe data is easy to be collected.
2. The size of cell-phone owner group is much larger than the sample size in traditional surveys.

For example, in United States, the number of cell-phone users reaches approximately 87% of the total population in 2008 *(19)*, in which the major 3 cellular carriers add up to nearly 80% market share (Verizon: 32%, AT&T: 29%, Sprint: 18%) *(20)*. Considering the large size of the existing cell-phone owner group, it would be clear that the sampling error should be substantially less than the traditional O-D surveys. It should be noticed that although the cell phone market penetration rate reaches 80% of the market share, in practice, since the cellular probe data are collected independently among different cellular carriers, it is more possible to collect the data from one or two cellular carriers. Therefore, we need to consider both the cell phone market penetration rate and the market share of individual cellular carriers.

Many researches treated the cellular probe trajectory data as the SRS survey data. Each individual in the sample is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process. In other words, each individual has the same probability to own a cell phone. Here the probability of owing a cell phone is a prior probability and equals to the cell-phone market penetration rate. However, the market penetration rate is a kind of prior probability which is obtained from some market research reports or papers. It may lead to inaccuracy if disregarding the possible social-economical bias. Typically, the probability of whether a person owns cell phones should be related to many factors, such as age, income and race, etc. *(21)*. For example, young people consist of the largest group of the cell-phone owners in terms of the cell-phone owners' age distribution. Therefore, the young people will have larger probability to own cell phones than the old ones. And in some cases, the higher income people will have larger probability to own cell phones.

The conditional probability of cell-phone ownership can be assumed to have the following linear relationship:

$$p\left(cp \mid f_1, f_2, \dots f_n\right) = \alpha_1 p(cp \mid f_1) + \alpha_1 p(cp \mid f_2) + \dots + \alpha_n p(cp \mid f_n) \qquad (1)$$

where $\left(\alpha_1, \alpha_2 \dots, \alpha_n\right)$ is the coefficients where $\sum_{i=1}^{n} \alpha_i = 1$. Generally, equation (1) needs to be calibrated to determine the coefficients. In many situations, the following equation is used to calculate the probability $p(cp \mid f_n)$:

$$p(cp \mid f_n) = \frac{p\left(cp, f_n\right)}{p\left(f_n\right)} = \frac{p\left(f_n \mid cp\right) p\left(cp\right)}{P\left(f_n\right)}$$

Considering the situation for a specific cellular carrier m, the probability of a person owns a cell phone in TAZ i turns to be:

$$p_k^i = p_{c_m} p\left(cp \mid f_1, f_2, \dots f_n\right) = p_{c_m} \left[\alpha_1 p(cp \mid f_1) + \alpha_1 p(cp \mid f_2) + \dots + \alpha_n p(cp \mid f_n)\right] \qquad (2)$$

If multiple carrier data are available, the equation (2) turns to be:

$$p_k^i = \left[\alpha_1 p(cp \mid f_1) + \alpha_1 p(cp \mid f_2) + \dots + \alpha_n p(cp \mid f_n)\right] \sum_{m=1}^{M} p_{c_m} \qquad (3)$$

where M is number of cellular carriers from which the data are obtained. Note that

In practice, due to the privacy concerns, most of the personal information required for the equations (1 – 3) cannot be obtained directly from cellular carriers or operators. However, the U.S. census provides a large amount of demographic survey data for us to produce the distributions of the personal information (age, income, race, etc.). We can utilize the information to calculate the probability of cell-phone ownership. The case study part will give a detailed procedure to determine the cell-phone ownership probabilities.

3.4 Vehicle-per-cellphone equivalent factor

Typically, the cell-phone tracking technology will return the cellular probe counts. However, in transportation planning field, the main interest is on vehicle flows rather than the cellular probe flows. Consequently, a vehicle-per-cellphone equivalent factor f_{vc} will be used in our method to convert cellular probe flows into equivalent vehicle flows *(13)*. We designed a method to estimate the f_{vc} based on the posterior information obtained from the characteristics of the cellular probe trajectories.

According to the cellular probe trajectory characteristics, the set of trajectories can be divided into three subsets:

1. The set of trajectories crossing at least two LA boundaries, σ_1.
2. The set of trajectories crossing just one LA boundaries, σ_2.
3. The set of trajectories without crossing any LA boundaries, σ_3.

For the first two subsets of trajectories, here are three assumptions:

1. *Phones in close proximity (i.e. the same car) generate signal transition events at exactly the same time.* In practice, this assumption needs to be relaxed since phone variation is quite high and signal events may have quite large differences in timing even for phones in the same car. The following two assumptions hold based on this assumption.
2. *There cannot be two vehicles crossing two continuous LA boundaries at same timestamps.* Typically, the dimension of a LA is 3-5 miles by 3-5 miles. There is a very small possibility that some parallel travelling cars crossing at least two LA boundaries at two same timestamps. If two cellular probe trajectories crossing two continuous LA boundaries at two timestamps t and $t+\tau$, they should be in the same vehicle.
3. *Within the saturation headway, there is only one vehicle crossing LA boundaries in each lane.* The default saturation headway is 2.0 seconds. Within two timestamps t and $t+2$, there is only one vehicle crossing LA boundaries in each lane.

The estimation of f_{cpv} for the trips crossing at least two boundaries will be based on the first assumption. Assuming a set σ_s of cellular probe trajectories $(1, 2 \ldots i, \ldots m)$ crossing at two LA boundaries at timestamps t and $t+\tau$, so the expected value of the number of passengers in σ_s will be:

$$\psi_{\sigma_s} = \sum_{k \in \sigma_s} \frac{1}{p_k^i} \tag{4}$$

The average passenger-vehicle occupancy $\bar{f}_{pvo}$ (passengers per vehicle) *(22)* is applied to determine the number of vehicles crossing the two LA boundaries at timestamp t and $t+\tau$:

$$Vehs = 1 + \left(\psi_{\sigma_s} - \left\| \sigma_s \right\| \Big/ \bar{f}_{pvo} \right)$$

So the vehicle-per-cellphone equivalent factor for cell-phone owner i in set σ_s is:

$$f_{vc}^{i} = \frac{1 + \left(\psi_{\sigma_s} - \left\| \sigma_s \right\| \Big/ \bar{f}_{pvo} \right)}{\psi_{\sigma_s}}, \quad i \in \sigma_s$$

For the second subset of the trips, the third assumption is used. Suppose a cell-phone owner i in set σ_2 crosses a LA boundary at timestamp t. There are Ω links located at the boundary. Each of the link j has several lanes. The number of average occupied lanes (the average number of lanes which are occupied by vehicles) at link during peak hour is π_j. A set σ_t of cell phones crossing the boundary between time t and $t+2$. Note that σ_t consists of both the cell phones crossing only one LA boundary and those crossing at least two LA boundaries at time t and $t+2$. So the vehicle-per-cellphone equivalent factor for set σ_2 will be:

$$f_{vc}^{i} = \frac{\sum_{j=1}^{\Omega} \pi_j - \left(\psi_{\sigma_t \cap \sigma_1} \Big/ \bar{f}_{pvo} \right)}{\psi_{\sigma_t} - \psi_{\sigma_t \cap \sigma_1}}, \quad i \in \sigma_2 \qquad (5)$$

For the third subset of cellular probe trajectories, the average value of f_{vc} of the first two subsets is assigned to them:

$$f_{vc}^{i} = \frac{\sum_{j \in \sigma_1} f_{vc}^{j} + \sum_{j \in \sigma_2} f_{vc}^{j}}{\left\| \sigma_1 \right\| + \left\| \sigma_2 \right\|}, \quad i \in \sigma_3 \qquad (6)$$

where the operator $\left\| \bullet \right\|$ means that the size of the set.

3.5 Trip Generation and Distribution

The trip generation and distribution are the first two steps in the traditional four-step transportation planning process. The trip generation is to decide the number of trips which are produced or attracted in a specific TAZ. The trip distribution process is to distribute the productions and attractions predicted by trip generation model to the O-D flows from each production zone i to each attraction zone j.

Due to the limitations on the sample sizes of surveys, the traditional trip generation and distribution model cannot secure an accurate result. The cell-phone tracking technology provides a larger sample. Here we introduce a new method to obtain the population O-D demand combining trip generation and distribution together.

For the total population, the proportion of people who have trips between TAZ i and j should be:

$$P_{ij} = \overline{Y}^{ij} = \frac{1}{N}\sum_{k=1}^{N} Y_k^{ij}$$

To get the value of T_{ij}, it only needs to multiply P_{ij} with N :

$$T_{ij} = \sum_{k=1}^{N} Y_k^{ij} = NP_{ij}$$

If treating the cell-phone owner group as a "simple random sample". The sampling results can directly be estimated by the following equation:

$$\hat{P}_{ij} = \frac{1}{\left\|S^i\right\|}\sum_{k=1}^{n} y_k^{ij}$$

Note that in SRS method, the size of cell-phone owner group can be estimated by:

$$\left\|S^i\right\| = Np(cp)\,p_{c_m}$$

So the estimated value of T_{ij} in SRS survey method is:

$$\hat{T}_{ij} = N\hat{P}_{ij} = \frac{\sum_{k=1}^{n} y_k^{ij}}{p(cp)\,p_{c_m}} \tag{7}$$

Since the distribution of cell-phone owners cannot be considered as the "simple random sample", the $\hat{T}_{ij}$ cannot be inferred directly using equation (7). A Horvitz – Thompson (HT) estimator *(23)* of the P_{ij} is proposed:

$$\tilde{P}_{ij} = \sum_{k \in S^i} \frac{y_k^{ij}}{Np_k^i} \tag{8}$$

From equation (8), it can be seen that the higher probability of owning cell phones, the less weight the corresponding y_k^{ij} is given, in this way the HT estimator uses probability to weight the responses in the estimating the total. The HT estimator of T_{ij} can be defined as follows:

$$\tilde{T}_{ij} = N\tilde{P}_{ij} = \sum_{k \in S^i} \frac{y_k^{ij}}{p_k^i} \tag{9}$$

Note that T_{ij} is the O-D trips between TAZ i and j, but what we need is the vehicle O-D flows. So the vehicle-per-cellphone factor should be added in the estimator:

$$\tilde{T}_{ij}^{\mathrm{Veh}} = \sum_{k \in S^i} \frac{y_k^{ij} f_{vc}^k}{p_k^i} \tag{10}$$

Now to prove the HT estimator of P_{ij} is an unbiased estimator. Let

$$\delta_k^i = \begin{cases} 1 & \text{if } k \in S\text{, that is to say the } k\text{th ppl has cell phone in TAZ } i \\ 0 & \text{Otherwise} \end{cases}$$

Then the estimator $\tilde{P}_{ij}$ can be expressed in following form:

$$\tilde{P}_{ij} = \sum_{k=1}^{N} \frac{Y_k^{ij} \delta_k}{N p_k^i}$$

The expectation of $\tilde{P}_{ij}$ is:

$$E\left(\tilde{P}_{ij}\right) = \sum_{k=1}^{N} E\left(\frac{Y_k^{ij}\delta_k^i}{N p_k^i}\right) = \sum_{k=1}^{N}\left(\frac{Y_k^{ij} E\left(\delta_k^i\right)}{N p_k^i}\right) = \sum_{k=1}^{N}\left(\frac{Y_k^{ij} p_k^i}{N p_k^i}\right) = \overline{Y}^{ij} = P_{ij} \tag{10}$$

The estimator $\tilde{P}_{ij}$ is the unbiased estimator of P_{ij}.

Furthermore, the variance of the estimator $\tilde{P}_{ij}$ is.

$$Var\left(\tilde{P}_{ij}\right) = Var\left[\sum_{k=1}^{N} \frac{Y_k^{ij}\delta_k^i}{N p_k^i}\right] = \frac{1}{N^2}\left\{\sum_{k=1}^{N} \frac{\left(Y_k^{ij}\right)^2 Var\left(\delta_k^i\right)}{\left(p_k^i\right)^2} + \sum_{k \neq m}^{N} \frac{Y_k^{ij} Y_m^{ij} Cov(\delta_k^i, \delta_m^i)}{p_k^i p_m^i}\right\} \tag{11}$$

Note that

$$Var\left(\delta_k^i\right) = E\left[\left(\delta_k^i\right)^2\right] - \left[E\left(\delta_k^i\right)\right]^2 = p_k^i(1 - p_k^i)$$

and

$$Cov(\delta_k^i, \delta_m^i) = E\left(\delta_k^i \delta_m^i\right) - E\left(\delta_k^i\right) E\left(\delta_m^i\right) = p_{km}^i - p_k^i p_m^i$$

where p_{km}^i is the joint probability of both the kth people and the mth people own cell phones.

Considering two people in sample have an independent probability to own cell phones, the equation (11) turns to be:

$$Var\left(\tilde{P}_{ij}\right)=Var\left[\sum_{k=1}^{N}\frac{Y_k^{ij}\delta_k^i}{Np_k^i}\right]=\frac{1}{N^2}\left\{\sum_{k=1}^{N}\frac{\left(Y_k^{ij}\right)^2 p_k^i(1-p_k^i)}{\left(p_k^i\right)^2}\right\}=\frac{1}{N^2}\left\{\sum_{k=1}^{N}\frac{\left(Y_k^{ij}\right)^2(1-p_k^i)}{p_k^i}\right\} \quad (12)$$

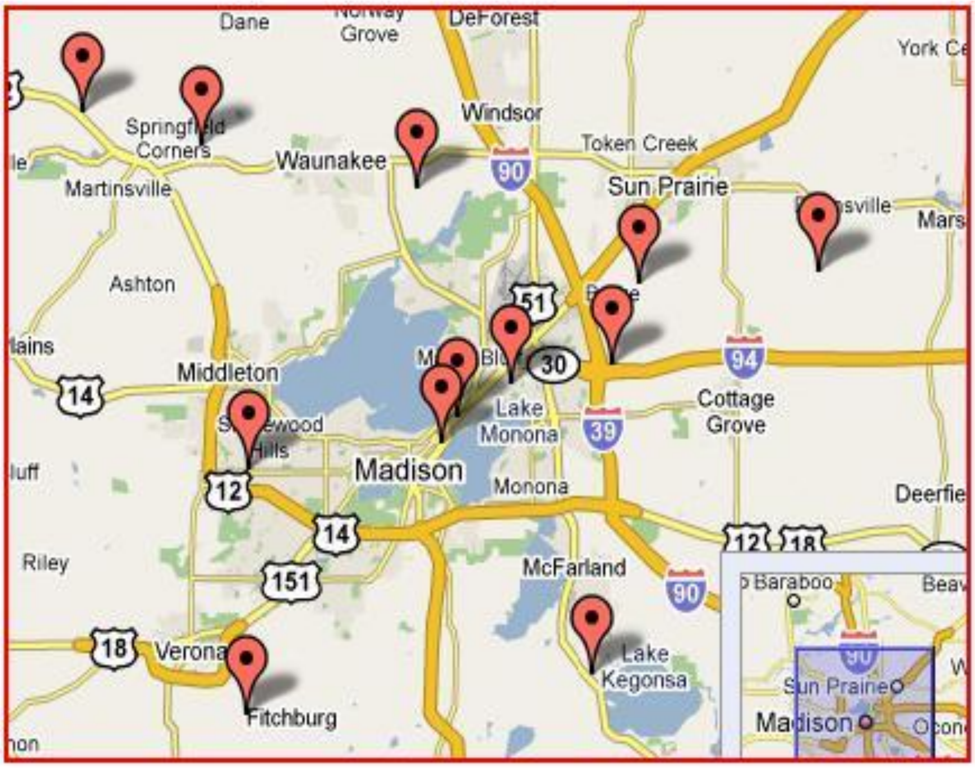
If assume the people in analyzed TAZ have the same probability p of cell-phone ownership, the expected value of variance turns to be:

$$E\left[Var\left(\tilde{P}_{ij}\right)\right]=E\left[\frac{(1-p)}{N^2 p}\sum_{k=1}^{N}\left(Y_k^{ij}\right)^2\right]=\frac{(1-p)}{Np}E\left[P_{ij}\right]=\frac{(1-p)\tilde{P}_{ij}}{Np} \quad (13)$$

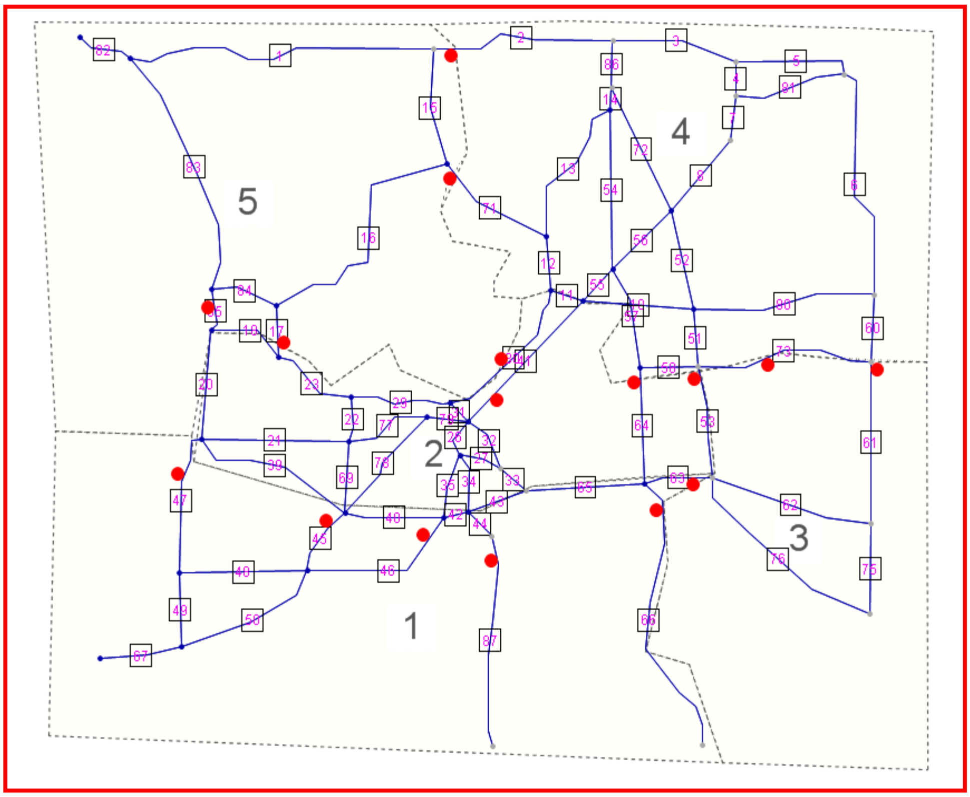
Then the expected value of standard deviation should be:

$$E\left[SD\left(\tilde{P}_{ij}\right)\right]=E\left[\sqrt{\frac{(1-p)P_{ij}}{Np}}\right]=\sqrt{\frac{(1-p)\tilde{P}_{ij}}{Np}} \quad (14)$$

4. CASE STUDY – SIMULATION EXPERIMENTS



(a)

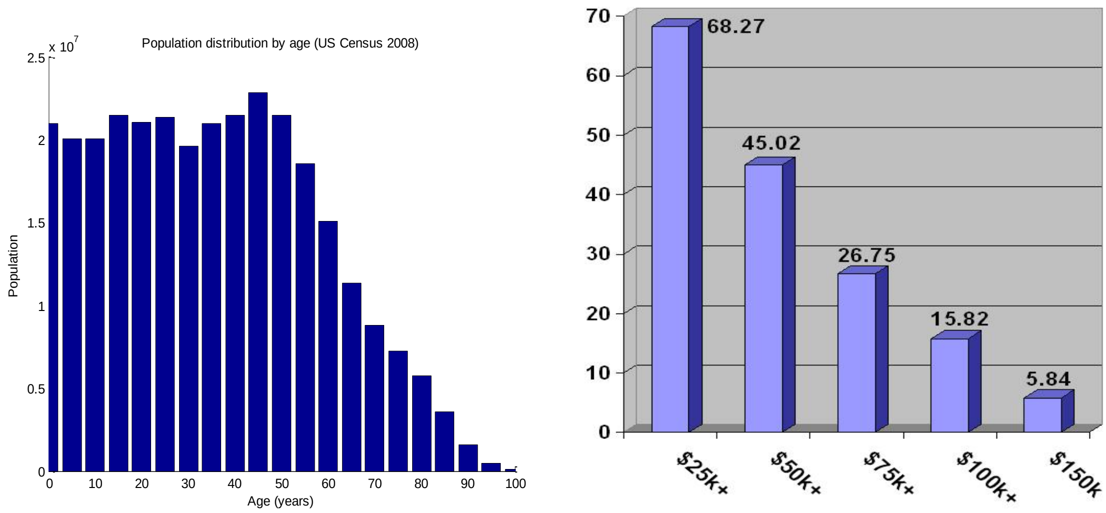


(b)

Figure.4 (a) Cell tower location map of the research area

(b) Corresponding simulation network

This simulation aims to provide a verification of the proposed O-D estimation method. The proposed research area is Dane county in the southwest of the state of Wisconsin. To be simplified, it is divided into 5 TAZs. Figure.4 (a) shows the cell-phone tower locations and Figure.4 (b) shows the corresponding VISSIM simulation layout of the research area. The red circles in Figure.4 (b) are the intersections of links and LA boundaries.



(a) (b)

Independent Variables	Cell-phone sample
Age	

18 - 30	41%
31 - 62	53%
63 – up	6%
Income	
Less than $50,000	51%
$50,000 – up	37%
Don't know/Refused	12%

(c)

Figure.5 (a) U.S. census data of age distribution, 2008 *(24)*

(b) U.S. census data of income distribution, 2006 *(25)*

(c) Demographic information on cell-phone ownership pattern *(21)*

The input data involve with 3 input modules: the trip survey data module, the cell phone ownership distribution module and the vehicle occupancy distribution module. The simulation period is set to be 24 hours to estimate the daily O-D demand data. The 3 input modules prepared the input parameters as well as the input O-D matrix to start the VISSIM simulator. A cell phone signal transition events module will be paired with the VISSIM simulator module to provide the random events such as call-in and call-out events, which can be used to generate the HO events.

The trip survey data module uses the Wisconsin State-wide Trip survey data *(26)* as the aggregated input daily O-D matrix. Also the trip survey data contains the age and household income information which can be used in the cell phone ownership distribution module.

The cell phone ownership distribution module is designed to assign the cell phone to each trip maker with cell phone ownership probabilities. To simplify the demonstration, only the income and age are taken into consideration for the cell-phone ownership probability. Figure.5 (a) and (b) illustrate the U.S. census demographic data for the population by age and income.

Figure.5 (c) shows the population age and income distribution, from which it is easy to get the conditional probability $p(cp \mid age)$ and $p(cp \mid income)$. To get the value of $p(cp \mid age)$ and $p(cp \mid income)$, it can be calculated as following:

$$p(cp \mid age) = \frac{p(cp, age)}{p(age)} = \frac{p(age \mid cp)\, p(cp)}{P(age)}$$

$$p(cp \mid income) = \frac{p(cp, income)}{p(income)} = \frac{p(income \mid cp) * p(cp)}{p(income)}$$

where $p(cp)$ is the market penetration rate of cell phones. In this simulation, it is set to be 0.8, and the market share of a specific carrier is set to be 0.25.

Use the equation (1) to determine the probability of cell-phone ownership:

$$p(cp \mid age, income) = \alpha\, p(cp \mid age) + (1 - \alpha)\, p(cp \mid income)$$

where α is the coefficient between 0 and 1. In this simulation, it is set to be 0.5.

In 2008 U.S. census data *(24)*, Dane county has a 482,705 residents. The cell-phone ownership probability for each people is generated based on the above demographic information. Then the cell-phone owners are assigned to each individual vehicle.

The vehicle occupancy distribution module is to generate the random numbers of passengers assigned to each individual vehicle. It is used to convert the trip counts into vehicle counts. The average Passenger-Vehicle Occupancy in United States is 1.89 *(27)*. A [1,3] discrete uniform distribution is used to generate the random numbers.

The VISSIM simulation tool is employed to simulate the vehicle movements between each O-D pair. The input O-D table is assigned by VISSIM's built-in Dynamic Traffic Assignment (DTA) algorithm to generate the vehicle flows on links. The centers and boundaries of cells and LAs are predefined without any time-dependent fluctuations in the simulation network. The radius of cell coverage is set to 1000 ft. The boundaries in VISSIM are set as data collection points on links where the cell or LA boundaries intersect with. The data collection points can record each vehicle's ID and timestamp when the vehicle crosses them. In this simulation, the cell phones are assumed to be set in turn-on mode automatically.

During each simulation time step, the system will check whether there is any signal-transition event happened. The cell phones are assigned with a small probability to determine the occurrence of call-in and call-out events. The durations are determined by assigned a random number. The HO events will be recorded when the cell phones are in on-call mode and cross the data collection point at cell boundaries. The TA events will be record with its corresponding cell centers when cell phones are in on-call mode. The LU events will be record when cell phones cross the data collection points at LA boundaries. The PLU events will be recorded every two hours with its corresponding cell centers as well. After collecting the cellular probe trajectories, the O-D estimation introduced in Figure.3 is employed to get the estimated O-D matrix. The simulation process is shown in Figure.6.

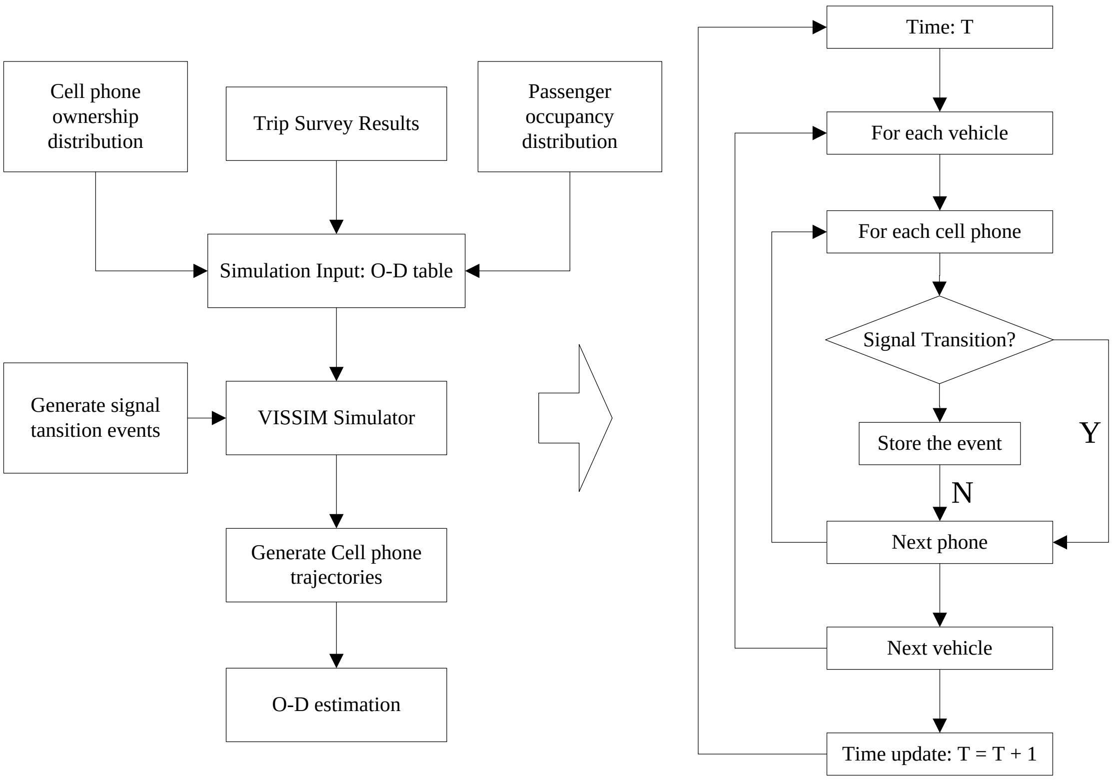


Figure.6 Illustration of the simulation process

The SRS method is also implemented in the simulation, and the average passenger-vehicle occupancy is used to convert the population trips into vehicle trips.

The results are shown in Table.2. It can be found that most of the estimated O-D flows have less percentage error than the SRS method. The average percentage error of the proposed method is 7.56%, while the SRS method returns 15.45%. Since the cell phone user group is not naturally a random sample, the HT estimator can give a more accurate estimation results comparing to SRS method. Moreover, our method uses the vehicle-per-cellphone factor to convert the cell phone counts into vehicle counts, while the SRS method employs the prior Passenger-Vehicle Occupancy information. The simulation results fully show the advantage of our method over SRS method.

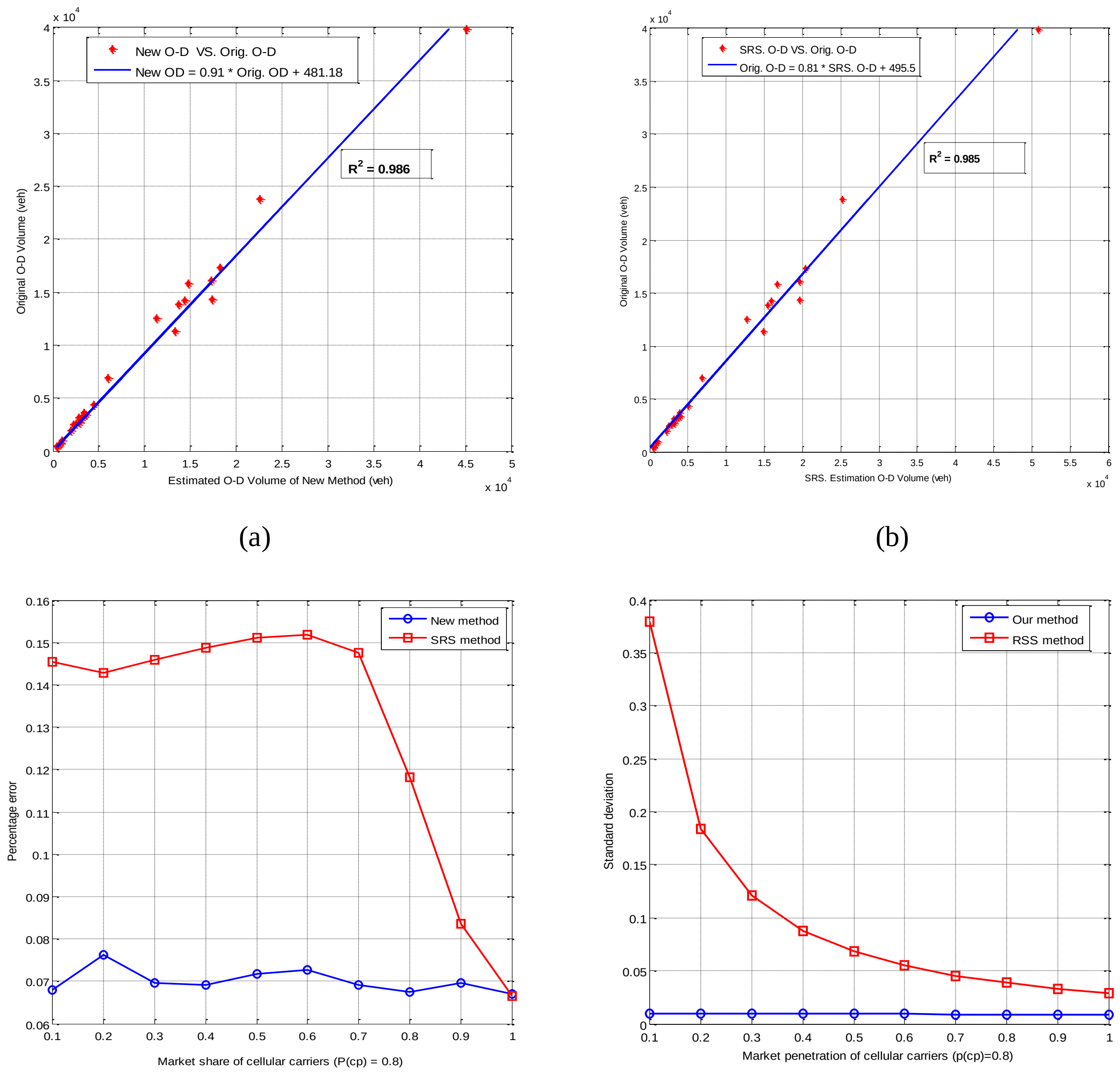
Table.2 Simulation results of proposed method and SRS method

O-D Pair	Orig. O-D	Esti. O-D	Per. Error	SRS O-D	Per. Error
1->1	16074	17325	7.78%	21080	22.23%
1->2	13872	13749	0.89%	15785	11.22%
1->3	750	796	6.13%	1043	21.33%
1->4	2562	2543	0.74%	3238	11.24%
1->5	2442	2274	6.88%	2543	2.70%
2->1	11334	13379	18.04%	16600	31.20%
2->2	23810	22581	5.16%	23413	6.08%
2->3	4332	4441	2.52%	5443	16.41%
2->4	14256	14405	1.05%	16230	12.06%
2->5	15828	14764	6.72%	15048	5.53%
3->1	960	972	1.25%	1000	15.94%
3->2	3384	3645	7.71%	4008	19.77%
3->3	6948	6026	13.27%	6035	1.87%
3->4	3300	3366	2.00%	3900	14.24%
3->5	456	481	5.48%	498	20.61%
4->1	2712	2899	6.90%	3613	21.05%
4->2	17322	18290	5.59%	21863	17.83%
4->3	3138	2823	10.04%	2863	0.06%
4->4	39840	45142	13.31%	55298	27.60%
4->5	3642	3418	6.15%	3730	7.08%
5->1	1947	1943	0.21%	2268	12.12%
5->2	14316	17358	21.25%	22353	37.25%
5->3	390	506	29.74%	585	37.95%
5->4	3174	3202	0.88%	3538	10.84%
5->5	12522	11347	9.38%	11123	2.02%

Furthermore, the estimated O-D volumes versus original O-D volumes are plotted in Figure.7 (a) and (b). It can be seen that both the results from the proposed method and SRS method have strong relationship with the original O-D flow. The regression shows both the lines fits the data very well, in which the coefficients of determination R^2 of our method are 0.986 and 0.985, respectively.

To better illustrate the advantages of the proposed method, a sensitivity analysis is carried out to see the influence of the cell-phone owner group size (market share of cellular carriers). The cell-phone market penetration rate is fixed at 0.8. The market share of cellular carriers is increased from 0.1 to 1. Note that here the market share of cellular carriers is the total market share of the carriers which are available to provide cellular data. Figure.7 (c) and (d) show the comparison between our method and SRS method in term of the percentage error and standard deviation of P_{ij} with the increasing of market share of cellular carriers.

It can be seen that the percentage error keeps unchanged at about 7% with increasing of cellular carriers' market share. On the other hand, the percentage error of SRS method decreases until the market share increased to 0.7.



(a) (b)

(c) (d)

Figure.7 (a) Estimated O-D VS. Original O-D

(b) RSS O-D VS. Original O-D

(c) Market penetration rate VS. Percentage error

(d) Market penetration rate VS. Standard deviation

In Figure.7 (c) and (d), the standard deviation of our method keeps unchanged below 5% with the increasing of market share, while the SRS method will decrease from more than 35% to 5% when the market share increases from 0.1 to 1.

Both the results of percentage error and standard deviation show the proposed method is robust method for the daily O-D matrix estimation. Generally, the smaller market share of the cellular carriers in practice, the less cellular trips can be obtained from the trajectories. Then the accuracy of estimation results will be more difficult to attain. Different with the SRS method, the proposed method can still keep good performance at smaller data set.

5. CONCLUSIONS AND FUTURE RESEARCH EXTENSION

Traditional survey-based trip diary approach to estimating trip generation and distribution is time-consuming and cost-prohibitive. The estimation may vary from one study as a result of the limitation of the survey sample size and sampling randomness. With the popularity of cell phone and emerging cellular tracking technologies, using cellular probe data have the great potential to provide a larger sample size in a timely manner.

In this paper, an exploratory methodology was presented to estimate the daily O-D demand using cellular probe trajectories. They can be obtained by tracking all the signal-transition and periodic location update events of cellular probes to determine the trip origins and destinations. To overcome the potential socio-economic bias, a conditional probability of cell-phone ownership was estimated using traveler's socio-economic factors that are readily available in the census data. Then, a vehicle-per-cellphone equivalent factor was generated based on the posterior information of the characteristics of cellular probe trajectories. In other words, individual cellular trips were converted into equivalent vehicle trips. Next, the trip generation and distribution were obtained simultaneously using a Horvitz-Thompson estimator so that the population O-D demand can be estimated. The Horvitz-Thompson estimator was proved to be an unbiased estimator of the population O-D demand in theory. A VISSIM based simulation was designed to exemplify the proposed method. A "simple random sampling" (SRS) method, the prevailing method in current literature, was also simulated. The comparison between the outcome of cellular probe data and SRS shows that both methods yielded desirable goodness-of-fit in terms of R^2 but the average percentage error of SRS is almost twice of the cellular probe data method, demonstrating the superiority of the proposed methodology. The sensitivity analysis has also shown that the proposed method provides a robust estimation for the daily O-D matrix.

To verify the validity of the assumptions of proposed methodology, a field test is needed in the future study, in which the cellular probe data and cell boundaries will be obtained from cellular carriers. A method should be proposed to eliminate the estimation error caused by variations of cell sizes and boundaries. A more accurate method to determine the trip origins and destinations should be developed. And an additional survey is needed to get accurate demographic information of cell-phone owners. An existing O-D demand matrix will be used as the ground truth to verify the correctness of the estimation results.

REFERENCE

1. Meyer, M.D. and E.J. Miller, *Transportation Planning: A Decision-Oriented Approach*. McGraw-Hill Book Company, INC, New York, NY. 1984
2. Giaimo, G.T., Modifications To Traditional External Trip Models. 2002. pp. p. 163-171.
3. Abrahamsson, T., Estimation of Origin-Destination Matrices Using Traffic Counts - A Literature Survey. INTERIM Report. 1998.
4. Sherali, H.D. and T. Park, Estimation of Dynamic Origin-Destination Trip Tables for A General Network. *Transportation Research Part B: Methodological*, 35(3). 2001. pp. 217-235.
5. Wong, S.C., et al., Estimation of Multiclass Origin-Destination Matrices from Traffic Counts. *Journal of Urban Planning and Development*, 131(1). 2005. pp. 19-29.
6. Stopher, P.R. and S.P. Greaves, Household Travel Surveys: Where Are We Going? *Transportation Research Part A: Policy and Practice*, 41(5). 2007. pp. 367-381.
7. Hazelton, M.L., Some Comments on Origin-Destination Matrix Estimation. *Transportation Research Part A: Policy and Practice*, 37(10). 2003. pp. 811-822.
8. Qiu, Z., et al. State of the Art and Practice: Cellular Probe Technology Applied in Advanced Traveler Information Systems. In *Transportation Research Board 86th Annual Meeting*. Washington D.C.: Transportation Research Board. 2007.
9. Astarita, V., et al., Motorway Traffic Parameter Estimation from Mobile Phone Counts. *European Journal of Operational Research*, 175(3). 2006. pp. 1435-1446.
10. Caceres, N., J.P. Wideberg, and F.G. Benitez, Review of Traffic Data Estimations Extracted from Cellular Networks. *Intelligent Transport Systems, IET*, 2(3). 2008. pp. 179-192.
11. Asakura, Y. and T. Iryo, Analysis of Tourist Behaviour based on the Tracking Data Collected using a Mobile Communication Instrument. *Transportation Research Part A: Policy and Practice*, 41(7). 2007. pp. 684-690.
12. Pan, C., et al., Cellular-Based Data-Extracting Method for Trip Distribution. *Transportation Research Record: Journal of the Transportation Research Board*. 2006. pp. pp 33-39.
13. Caceres, N., J.P. Wideberg, and F.G. Benitez, Deriving Origin Destination Data from A Mobile Phone Network. *Intelligent Transport Systems, IET*, 1(1). 2007. pp. 15-26.
14. Keemin, S. and K. Daehyun, Dynamic Origin-Destination Flow Estimation Using Cellular Communication System. *Vehicular Technology, IEEE Transactions on*, 57(5). 2008. pp. 2703-2713.
15. White, J. and I. Wells. Extracting Origin Destination Information from Mobile Phone Data. In *Road Transport Information and Control, 2002. Eleventh International Conference on (Conf. Publ. No. 486)*. 2002.
16. Lu, J., et al. Applying Cellular-Based Location Data to Urban Transportation Planning. In *Applications of Advanced Technology in Transportation*. Chicago: ASCE. 2006.
17. Mircea, I., S. Emil, and H. Simona. Cell ID Positioning Method for Virtual Tour Guides Travel Services. In *ECAI 2007 - International Conference*. Pitesti, Romania: Electronics, Computers and Artificial Intelligence. 2007.
18. Gur, Y.J., S. Bekhor, and C. Solomon. An Aggregate National Transportation Planning Process in Israel: Formulation and Development. In *Transportation Research Board 88th Annual Meeting*. Washington D.C.: Transportation Research Board. 2009.
19. *Background on CTIA's Semi-Annual Wireless Industry Survey*. 2009; Available from: files.ctia.org/pdf/CTIA_Survey_Year_End_2007_Graphics.pdf.
20. US Wireless Data Market Update - Q1. Chetan Sharma Consulting Co. Ltd. 2009.
21. Cell Phone Nation 2009. Marist Institute for Public Opinion. 2009.
22. Gan, A. and K. Liu, *Vehicle Occupancy Trends in Florida: Evidence from Traffic Accident Records*. Transportation Research Board 87th Annual Meeting. Transportation Research Board. pp. 17p. 2008
23. Konijn, H.S., *Statistical Theory of Sample Survey Design and Analysis*. American Elsevier Publishing Company, INC., New York, NY. 1973
24. Age and Sex Distribution in 2005. U.S. Census Bureau. 2005.
25. Annual Social and Economic Supplement 2006. U.S. Census Bureau. 2006.

26. Statistics, B.o.T. *NHTS/NPTS Database*. 2003; Available from: http://nhts.ornl.gov/download.shtml.
27. Statistics, B.o.T., NHTS 2001 Highlights Report. U.S Department of Transportation, Washington, DC. 2003.