Fundamental Study on the Issues of using Probe Data for OD Estimation and Route Identification

Masao Kuwahara
Professor, Institute of Industrial Science, University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
TEL: +81-3-5452-6418, FAX: +81-3-5452-6420
E-mail: kuwahara@iis.u-tokyo.ac.jp

Edward Chung
Visiting Professor, Center for Collaborative Research, University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan
TEL: +81-3-5452-6529, FAX: +81-3-5452-6420, E-mail: edward @iis.u-tokyo.ac.jp

Tomotaka Ishida
Gyoda Filtration Plant, Busuness Enterprize Bureau, Saitama Prefecture

ABSTRACT

This study discusses methodologies and issues on OD estimation and route identification from probe information. Although a number of studies on probe information have been reported, most studies have mainly focused on the characteristics of probe information itself such as the frequency and accuracy of the information and on algorithms of map-matching. This research however intends to explore how probe information can be utilized for transportation planning and traffic operation. More specifically, how OD volumes and trip routes can be estimated from probe information and how the accuracy depends on probe characteristics, zone configuration and network density are discussed. Firstly, the theoretical framework is constructed and then the proposed methodology for OD estimation using simulation and for route identification using the Tokyo metropolitan network is demonstrated.

INTRODUCTION

This study discusses methodologies and issues on OD estimation and route identification from probe information. Probe in this case is position detection tool such as GPS, mobile phone or PHS (Personal Handy phone System) which measure precise and detailed data of traveler's movements. This means that complete information of traveler's time-space trajectory from his/her origin to destination can be obtained.

Although quite a few studies on probe information have been reported (Wang and Nakamura 2004, Chung et al. 2003), these studies so far mainly focused on the characteristics of probe information such as the frequency and accuracy of the information and on algorithms of map-matching. This research however intends to explore how probe information can contribute to transportation planning and operation. More specifically, how OD volume and route identification are affected by the accuracy and frequency of probe information as well as the network and zone configurations, are discussed.

PROBE DATA CHARACTERISTICS

Positional data collected with GPS measures the distances to GPS satellites. The accuracy of positional data collected depends on the type of GPS such as kinematic GPS has an accuracy of 5 to 15 cm, and differential GPS has an accuracy of 2 to 10 meters. On the other hand, PHS and mobile phone use network based positioning technology, rely on various means of triangulation of the signal from cell sites serving a mobile phone or radio bases serving a PHS. The radio bases of PHS are distributed densely because the strength of radio wave used by this system is weaker than normal cellular phone, and therefore the level of accuracy of PHS is higher than mobile phone. PHS and mobile phone has an accuracy of 50-100m and 150-500m, respectively. Some mobile phones equipped with GPS can use Hybrid Wireless Assisted GPS that make use of the mobile network based positioning when the GPS system (the most accurate) is unsuccessful in acquiring enough satellites. Assisted GPS provides high yield of location fixes in all mobile coverage areas including urban canyons and inside buildings. Thus increases the capture positional data using mobile phone and also the accuracy level to 5-50 meters.

STUDY FRAMEWORK

In transport applications, probe data are often measured using GPS, mobile phones and PHS. As discuss above, all methods of measurement have a degree of error in the positional data and it is dependent of the equipment, network, landscape and environment. The number of probe depends on the number of positional equipments in the target population. This in part affects the frequency of data available for analysis. However, the frequency of data transmission is also constraint by the cost of transmission and bandwidth.

Positional data are aggregated into zones to form OD matrix. The size of the zones with respect to measurement error affects the accuracy of the OD matrix. Route identification is mapped straight onto transport network and therefore the accuracy of the map matching may depend on the density of the transport network.

The following sections analyses these factors, measurement error, data frequency, zone size and transport network configuration on OD estimation and route identification.

OD ESTIMATION

The conventional method of collection OD data is via questionnaire survey. This is not only an expensive exercise but the frequency of OD data collection is also restricted to every 5 years if the coverage area is large. Probe data offers the opportunity to estimate OD continuously by time of day and day of week, and a much lower cost. This section deals with a fundamental analysis to explore how a uniformly distributed location error, zone size and sample distribution has affected on OD estimation.

THEORETICAL ANALYSIS

Suppose that probe information of n_{ij} trips from zone i to zone j is received. Since the probe information has some errors (we call 'probe information' for information provided from a probe car), it cannot be taken for granted that actually n_{ij} trips are made from i to j. It is expected that the accuracy of the estimated OD volumes generally depend upon the following three variables:

1) Accuracy of probe information: Probe information always contains some amount of error and the magnitude of the error would affect the estimated OD volumes.

2) Zone characteristics: Zone characteristics such as the size, shape, and distribution of population and employment over the zone, affect OD estimates. If the size of a zone is relatively large enough compared to the range of error of probe information, it may be possible to aggregate the probe locations to estimate OD volumes by simply accepting all locations provided from probes. If not however, a location provided from a probe may not be accurate enough to identify a zone in which the probe actually exists. Also, the zone shape as well as distribution of population and employment over the zone may affect the OD estimates because the expected number of samples (ie. probe information) would be different.

3) The number of probe trips (samples): The number of probe information (ie. the number of samples) clearly influence the OD estimates in the expansion process.

Probability that probe k is actually in zone i

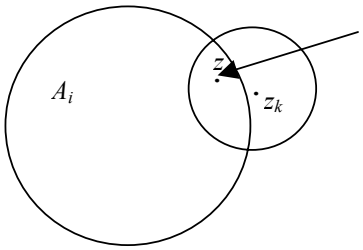
Suppose a location provided from probe k is $z_k = (x_k, y_k)$, where z_k is a two-dimensional vector of x_k and y_k coordinates. The location z_k could be any place either inside or outside of target

zone i. Then, the probability that the probe is actually within zone i can be evaluated using the following error distribution of the probe information. The error distribution could be evaluated based upon characteristics of the positioning system such as GPS, mobile phone, and PHS.

$e(z|\ z_k)$ = probability that a probe is actually at z when it provides information of z_k. (1)

$p_{ik}(z_k)$ = Prob{probe k is actually within zone i , when probe k provides information z_k}

$= \int_{A_i} e(z\ |z_k)dz$, where A_i = an area of zone i. (2)



The actual location of a probe which provides its information at z_k is distributed with probability density $e(z|\ z_k)$.

Evaluation Trip Generation from Zone *i*, G_i

First of all, the following variable g_{ik} is introduced:

$$g_{ik} = \begin{cases} 1 & \text{if probe } k \text{ actually generated from zone } i, \text{ with probability } p_{ik} \\ 0 & \text{otherwise, with probability } 1- p_{ik} \end{cases} \quad (3)$$

Then, the number of trips actually generated from zone i, G_i, is written as the sum of g_{ik}'s:

$$G_i = \sum_{k=1}^{n} g_{ik}\ , \qquad n = \text{the total number of all probes} \quad (4)$$

Given location z_k of probe information, since the actual location is distributed by $e(z|\ z_k)$, the g_{ik} clearly follows the Binomial distribution with parameter $p_{ik}(z_k)$. The expectation and variance of g_{ik} are written as follows provided that the probe k informs its location of z_k:

$$E\{g_{ik}|\ z_k\} = p_{ik}(z_k) \qquad Var\{g_{ik}|\ z_k\} = p_{ik}(z_k)\{1 - p_{ik}(z_k)\} \quad (5)$$

Considering the probability density of location z_k, which is denoted by $h(z_k)$, we evaluate the probability that g_{ik} is equal to 1. The $p_{ik}(z_k)$ is a probability of g_{ik} = 1 given z_k, and a probability that a probe information is sampled at location z_k is equal to $h(z_k)dz_k$. Therefore, probability p_{ik} that g_{ik} becomes equal to 1 is written as:

p_{ik} = probability that g_{ik} is equal to 1 = $\int_A p_{ik}(z_k)\ h(z_k)dz_k$ (6)

where A = a whole study area

$h(z_k)$ = probability density function of location z_k, $z_k \in A$

The expectation and variance of g_{ik} are hence written as

$$\begin{aligned} &E\{g_{ik}\} = p_{ik} \\ &Var\{g_{ik}\} = p_{ik}(1-p_{ik}) \end{aligned} \tag{7}$$

and those for G_i are

$$\begin{aligned} &E\{G_i\} = \sum_k E\{g_{ik}\} = \sum_k p_{ik} \\ &Var\{G_i\} = \sum_k Var\{g_{ik}\} = \sum_k p_{ik}\{1-p_{ik}\} \end{aligned} \tag{8}$$

Evaluation of Trips Attraction in Zone *j*, S_j

Similarly, for the number of trips actually attracted in zone *j*, S_j, the expectation and variance are written as:

$$S_j = \sum_{k=1}^{n} s_{jk} \qquad s_{jk} = \begin{cases} 1, \text{ if probe } k \text{ actually attracted in zone } j \text{ with probability } p_{jk} \\ 0, \text{ otherwise with probability } 1\text{- } p_{jk} \end{cases}$$

p_{jk} = probability that s_{jk} is equal to $1 = \int_A p_{jk}(z_k)\, h(z_k) dz_k$

$$\begin{aligned} &E\{S_j\} = \sum_k E\{s_{jk}\} = \sum_k \int_A p_{jk}(z_k)\, h(z_k) dz_k = \sum_k p_{jk} \\ &Var\{S_j\} = \sum_k Var\{s_{jk}\} = \sum_k p_{jk}(1-p_{jk}) \end{aligned} \tag{9}$$

Evaluation of OD Trips, N_{ij}

As for the OD trips, N_{ij}, the expectation and variance are written as:

$$N_{ij} = \sum_{k=1}^{n} g_{ik} s_{jk}$$

$$\begin{aligned} &E\{N_{ij}\} = \sum_k E\{g_{ik} s_{jk}\} = \sum_k \int_A p_{ik}(z_k) p_{jk}(z_k)\, h(z_k) dz_k = \sum_k p_{ik} p_{jk} \\ &Var\{N_{ij}\} = \sum_k Var\{g_{ik} s_{jk}\} = \sum_k p_{ik} p_{jk}(1 - p_{ik} p_{jk}) \end{aligned} \tag{10}$$

Linear City Example

Using a linear city with an area A of length $2w$, zone size of A_i, measurement error e and with the following functions as an example (see Figure 1):

$h(x) = 1/w \qquad 0 < x < w$

$e(z|z_k) = 1/2e \quad 0 < x < w, \quad$ for all k

$w > 1$

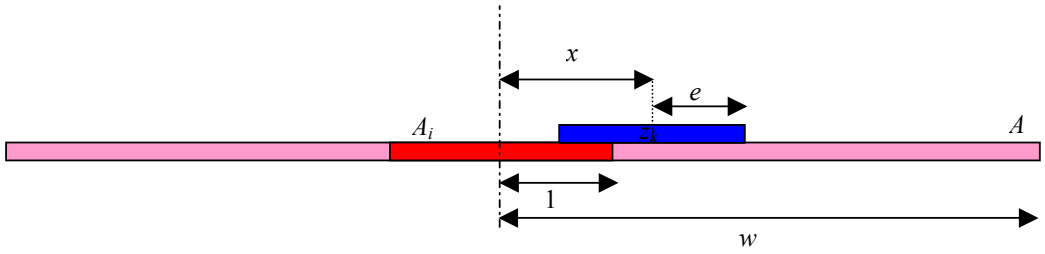
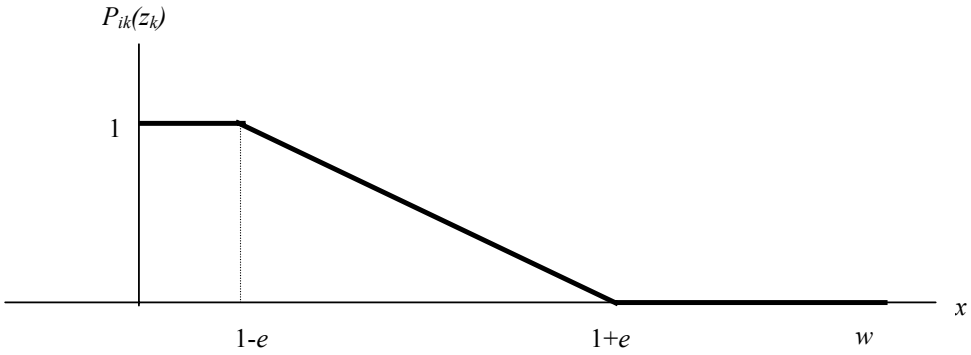


Figure 1 – Linear city

As shown above, we employ uniform distributions for both probe samples *h(x)* and error distribution $e(z|z_k)$. Also, the size of the zone A_i has the length of 2. Since the whole study area is symmetric, we consider only the half of the area as shown above.

Here, we summarize the major result and for the detailed derivation, please look at Appendix for more details. For $e < w\text{-}1$, the expectation and variance of g_{ik} and G_i can be evaluated as follows:

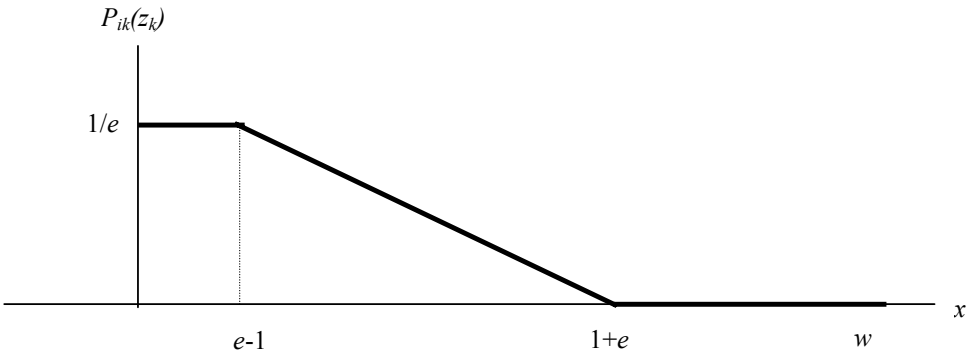
$$
\begin{aligned}
&E\{g_{ik}\} = p_{ik} = \frac{1}{w} \\
&Var\{g_{ik}\} = p_{ik}(1-p_{ik}) = \frac{1}{w}(1-\frac{1}{w}) \\
&E\{G_i\} = \frac{n}{w} \\
&Var\{G_i\} = \frac{n}{w}(1-\frac{1}{w})
\end{aligned}
\qquad (11)
$$



For a large error such as $e > w\text{-}1$, the detailed calculation is omitted. However, if the error increases very large, probability $p_{ik}(z_k)$ becomes $1/e$ for almost all location z_k and variance $Var\{g_{ik}\}$ converges to

$$Var\{g_{ik}\} = 1/e\{1 - 1/e\},$$

and hence $Var\{g_{ik}\}$ goes to zero as error e increases to infinity.



The interesting feature of the example is that the expectation and variance of G_i stay at constant values for $e < w$-1 regardless of the value of e. This means that we could obtain the same estimate of trip generation G_i with the same variance under the uniform error and population distributions. However, in the real world, the population (trip generating points) is normally distributed in various ways different from the error distribution. In such realistic conditions, larger e cannot capture the actual distribution of population but it may skew the distribution by the error distribution itself. This is because, within a certain range of error, we assume the actual probe location is distributed according to the error distribution rather than the population distribution.

SIMULATION ANALYSIS

The theoretical analysis above discussed the OD expected value and variance for N trips given the population is uniformly distributed. Although all the equations to estimate OD based on Eq. (2), (6) and (10) are derived, the practical calculation using this theoretical method can be a bit tedious. The other method to estimate OD is by simply summing up all the probe information. Thus, in this section, simulation analysis focuses on the later method and examines the probability that OD information provided by the probe is in fact correct.

A simulation program was developed to simulate the generation of probe information. The simulated study area consists of a 5 by 5 square grid and with uniform density distribution was simulated. The measurement error is assumed to be circular in shape and to be uniformly distributed. The density distribution of the study area is uniform. A sample size of 100,000 probe information was randomly generated. The steps for generating the OD pairs are as follows:

1. Origin zone is randomly selected and followed by the destination zone.
2. A point within each origin and destination zones are generated.
3. Given the measurement error range (ie. radius of circle), the proportion of the circles inside the origin zone, P_O and destination zone P_D are calculated (see Figure 2).
4. The probability of the OD pair is P_{OD} is simply $P_O P_D$.

5. Steps 1-4 are repeated for N samples.
6. The expected value and variance for each OD pair from a probe are calculated.

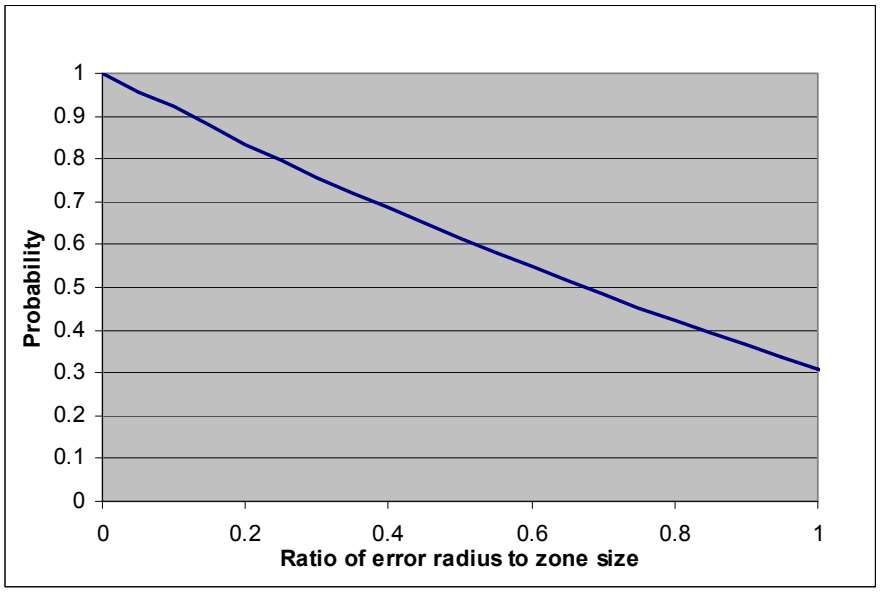


Figure 2 - Probability that a probe inside the subject zone (for uniform distribution)

The result from the simulation is expressed in terms of the ratio between the zone size *l* and measurement error *r* (ie. radius of circle), as shown in Figure 3. As expected, when the probe information is perfect ie. $r/l = 0$, the expected value is 1. However, for $r/l = 0.2$ and 1.0, the expected values are 0.7 and 0.1, respectively. The variance peaked at about $r/l = 0.2$ with a value of 0.492.

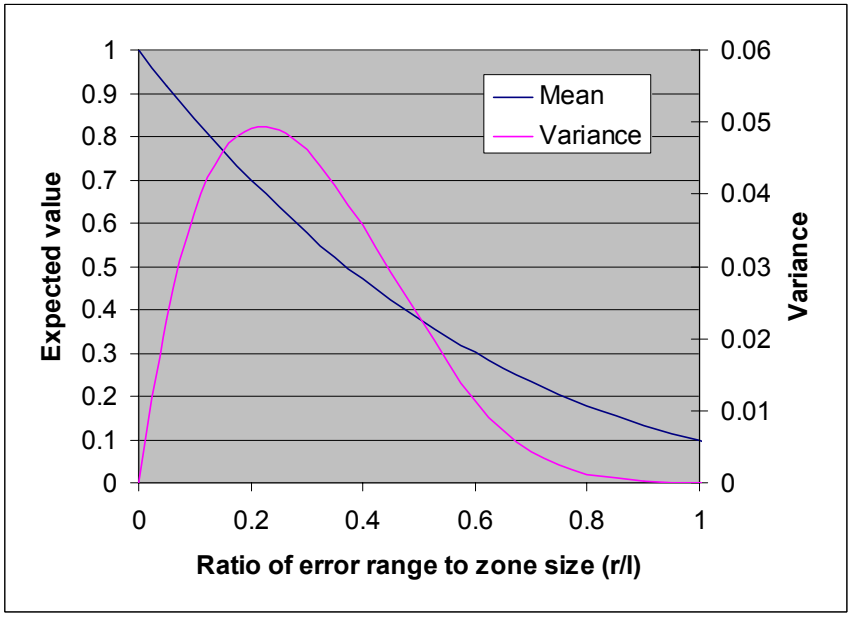


Figure 3 - Origin destination analysis for uniform density distribution

The outcome of the OD estimation is also dependent of the density distribution of the study area. A comparative analysis was carried out for 3 different density distributions (see Figure 4):

1. uniform (activities evenly distribution over the zone),
2. center (activities only occur at the center of the zone 36% of zone area), and
3. fringe (activities only occur at the fringe of the zone 64% of zone area).

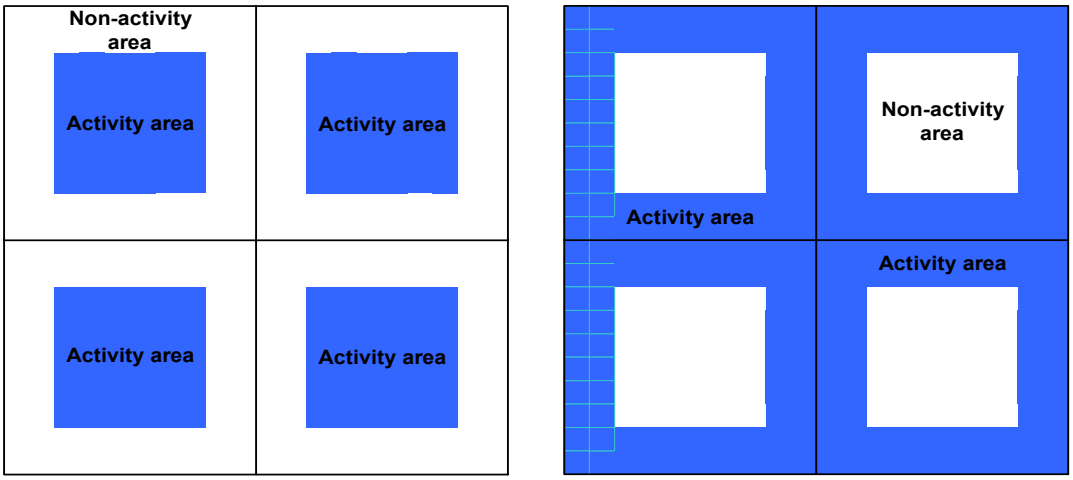


Figure 4 - Center and fringe activity distribution

Figure 5 shows for activities occurring at the centre of the zone, even when r/l=0.2, the expected value is still 1. However, for the same r/l value if all activities occur at the fringe, the expected value is only 0.55. Note that all the expected values converge to 0.1 at r/l=1 (ie. measurement error radius equals the zone size).

In summary, depending on the level of confidence required for OD estimation, the zone size, measurement accuracy and distribution of activities in the study area needs to be selected judiciously.

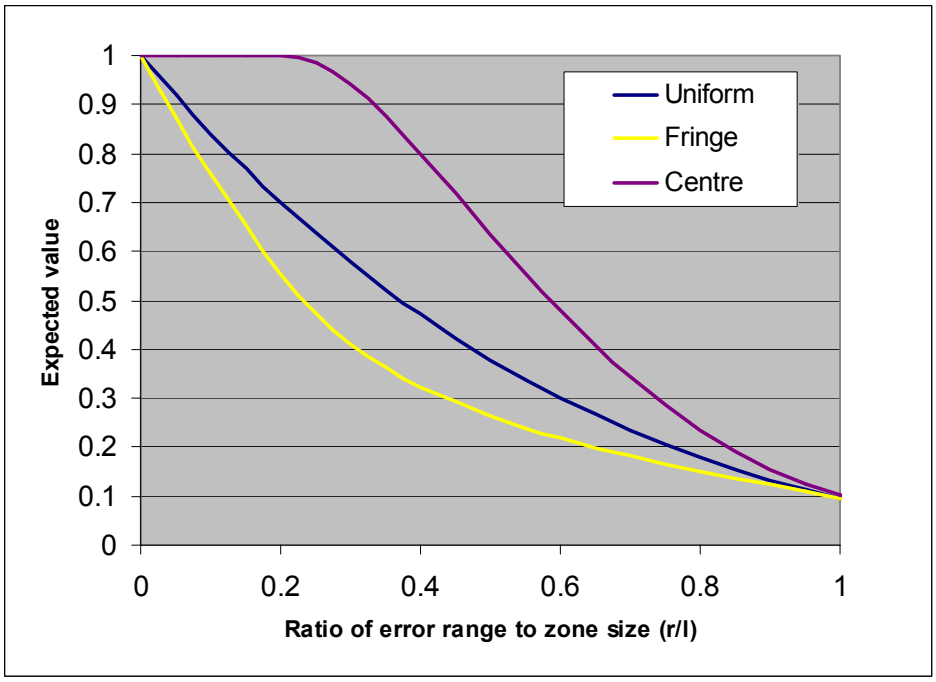


Figure 5 - Origin destination analysis for different density distribution

ROUTE IDENTIFICATION

In-vehicle map matching use in car navigation integrates GPS, navigable map database, driving direction and odometer. Correlating the raw data from GPS, odometer and other sensors to a navigable map database enables accurate car location to be determined. Although off-line map matching can consider all probe locations at once to find the route in contrast with in-vehicle map matching, it is more difficult in accurately matching the right route because no other vehicle sensors data are available for off-line matching.

This section describes a basic method for off-line map-matching and presents a case study in the center of Tokyo.

Off-line Map-matching Algorithm

Extracting Candidate Links

When information from probe k, $z_k = (x_k, y_k)$, is provided such as shown by a red circle in the figure below, a candidate link set is first extracted. A distance from link a to probe information z_k, $d(a,k)$ is defined as the distance of a perpendicular line from z_k to link a. Then,

link a is included in the candidate link set, if the minimum of $d(a,k)$ over k is less than the threshold value ϕ. In this analysis, the threshold value $\phi = 500$ [m] is employed.

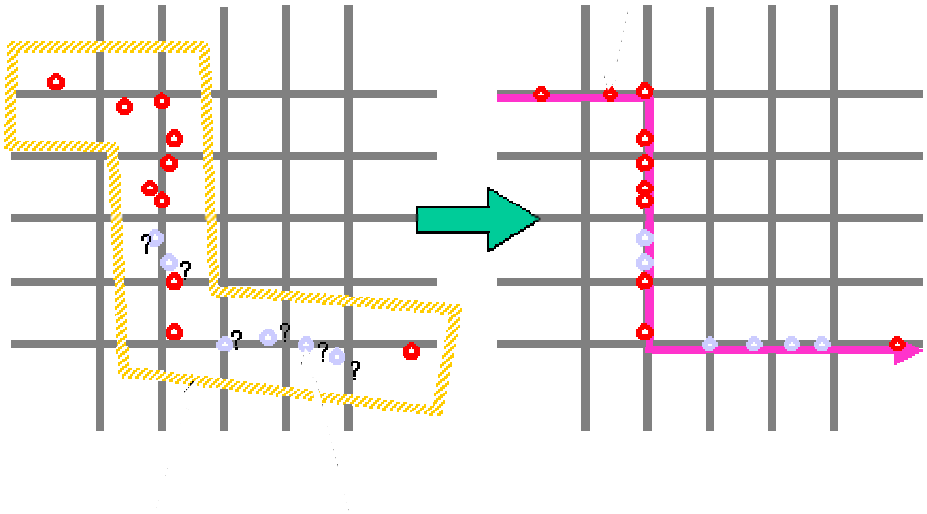


Figure 6 – Off line map matching algorithm

Defining Link Costs

For link a with its link length of L_a in the candidate set, link cost C_a is calculated using $d(a,k)$ as follows:

$$C_a = \alpha \cdot L_a \cdot \frac{\sum_k d(a,k)}{n_a}, \tag{12}$$

where n_a = the number of probe information which satisfy $d(a,k) < \phi$. In this definition, C_a becomes smaller if probe information z_k is closer to link a so that the link is more likely on the route.

Searching the Shortest Route

Using the calculated link costs C_a's, the route of probe k is identified by searching the shortest route from the initial link to the final link of probe k.

CASE STUDY

The study area is in the center of Tokyo covering an area of about 30 km by 30 km (see Figure 7) and consists of approximately 60,000 links. Two study routes are considered from A to B and A to C as shown by the red lines in the figure. Then, probe locations z_k's are generated along the route with uniform errors within several ranges at several different

frequencies. In this study, we use error ranges of 0, 10 and 100 meter and frequencies of 5, 60 and 300 seconds. For instance, for the error range of 10m at the frequency of 60 seconds, probe location z_k is generated at every 60 seconds with uniform error between 0 and 10 meters.

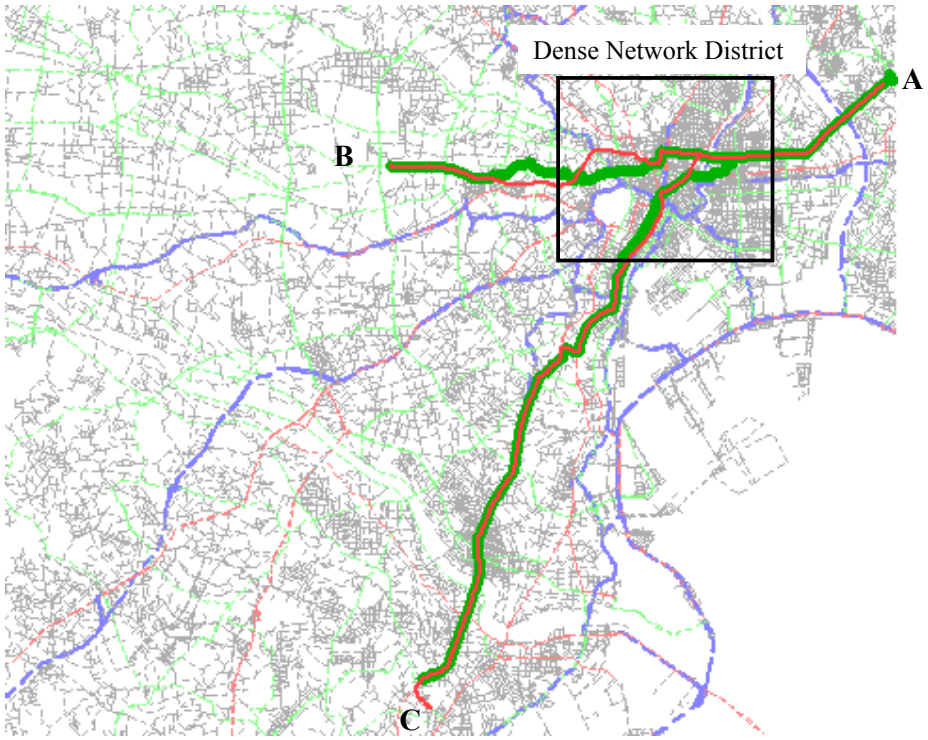


Figure 7 – Study Area in the Center of Tokyo

Off-line Map-matching Results

For probe locations generated with a combination of the error range and the frequency, the off-line map-matching is applied to estimate the route. As an example, routes shown by green lines in Figure 7 are those estimated when the error range is 100 m at frequency of 300 sec. For each of the estimated routes, the percentage of the route distance correctly identified is calculated as shown in Table 1.

Percentage of correctly identified Route Distance (%)		Frequency (seconds)					
		5		60		300	
Error Range (m)	0	A to B	100	A to B	100	A to B	68
		A to C	100	A to C	95	A to C	80
	10	A to B	100	A to B	100	A to B	68
		A to C	100	A to C	95	A to C	80
	100	A to B	97	A to B	100	A to B	68
		A to C	92	A to C	93	A to C	80

Table 1 - Percentage of Route Distance correctly Identified

In this experiment, the accuracy of route identification depends only on frequency of the probe information but almost independent of the error range up to 100 m. The mismatching mostly happens in the dense network district especially when the probe vehicle makes turns as shown in Figure 7.

CONCLUSION

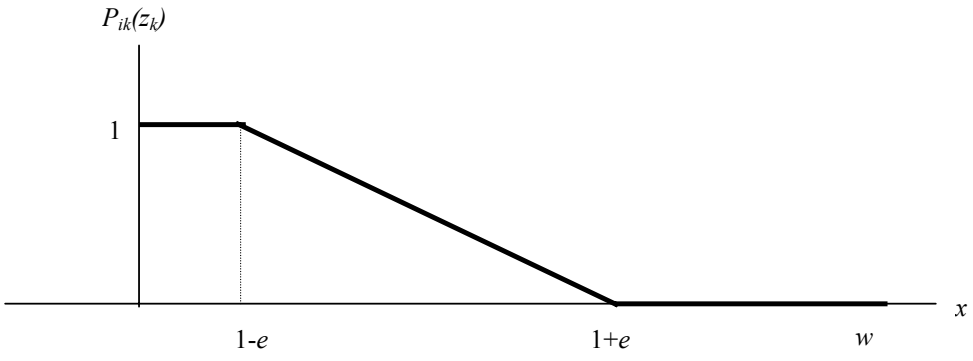
A theoretical analysis of OD estimates using probe information with measurement error for different zone size was analysed. The study found that for uniform distribution of activities, the expected value for N trips has the same value regardless of the magnitude of error range. Using simulation, the accuracy of correctly estimating OD trips as a function of the ratio between zone size and measurement error for uniform distribution and non uniform distribution was presented. The results demonstrated that depending on the level of confidence required for OD estimation, the zone size, measurement accuracy and distribution of activities in the study area needs to be selected judiciously.

Sensitivity of measurement error and frequency of probe information for off-line map matching to identify the route traveled was analysed. The study found that the accuracy of route identification depends only on frequency of the probe information but almost independent of the error range up to 100 m.

APPENDIX : Calculation of Linear City

The expectation and variance of g_{ik} and G_i can be evaluated as follows:
For $e < 1$,

$$p_{ik}(z_k) = \begin{cases} 1 & , \quad 0 < x < 1-e \\ p(x) = \dfrac{1+e-x}{2e}, & 1-e < x < 1+e \\ 0 & , \quad 1+e < x < w \end{cases}$$



$$p_{ik} = \int_A p_{ik}(z_k) h(z_k) dz_k = \frac{1}{w} \int_{1-e}^{1+e} \frac{1+e-x}{2e} dx + 1 \cdot \frac{(1-e)}{w} = -\frac{1}{2ew}(-2e^2) + \frac{(1-e)}{w} = \frac{1}{w}$$

Therefore, the expectation and variance of g_{ik} and G_i are

$$E\{g_{ik}\} = p_{ik} = \frac{1}{w}$$

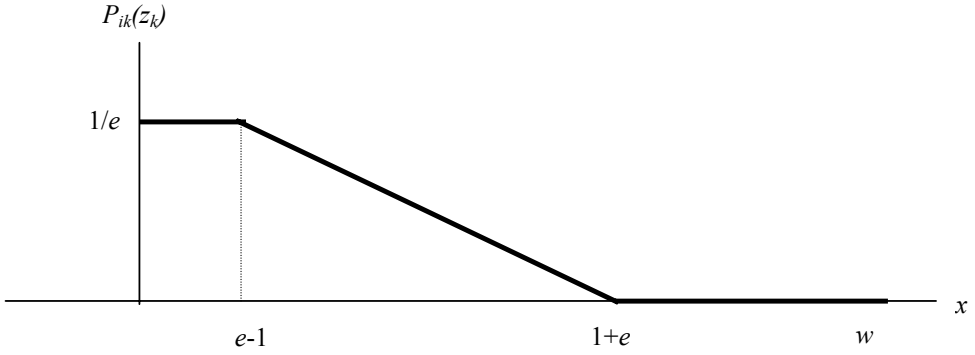
$$Var\{g_{ik}\} = p_{ik}(1-p_{ik}) = \frac{1}{w}(1-\frac{1}{w})$$

$$E\{G_i\} = \frac{n}{w}$$

$$Var\{G_i\} = \frac{n}{w}(1-\frac{1}{w})$$

For $1 < e$ and $e+1 < w$; that is, $1 < e < w-1$:

$$p_{ik}(z_k) = \begin{cases} \frac{1}{e} & , \quad 0 < x < e-1 \\ p(x) = \frac{1+e-x}{2e} & , \quad e-1 < x < e+1 \\ 0 & , \quad e+1 < x < w \end{cases}$$



$$p_{ik} = \int_A p_{ik}(z_k) h(z_k) dz_k = \frac{1}{w} \int_{e-1}^{e+1} \frac{1+e-x}{2e} dx + 1 \cdot \frac{1}{e}\frac{e-1}{w} = -\frac{1}{w} \cdot \frac{-1}{e} + \frac{e-1}{we} = \frac{1}{w}$$

Therefore, the expectation and variance of g_{ik} and G_i are

$$E\{g_{ik}\} = p_{ik} = \frac{1}{w}$$

$$Var\{g_{ik}\} = p_{ik}(1-p_{ik}) = \frac{1}{w}(1-\frac{1}{w})$$

$$E\{G_i\} = \frac{n}{w}$$

$$Var\{G_i\} = \frac{n}{w}(1-\frac{1}{w})$$

REFERENCES

R. Wang and H. Nakamura (2004) "Influence of probe vehicle sample size on expressway travel speed estimation", *Proceedings of 10th WCTR*, Istanbul, July 2004.

Chung, E., Sarvi, M., Murakami, Y., Horiguchi, R. and Kuwahara, M. (2003). "Cleansing of probe car data to determine trip OD". *Proc. 21st ARRB Conf. And 11th REAAA Conf., (CD), Cairns, Australia.* (ARRB Transport Research: Vermont South, Victoria, Australia).