Combined estimation of activity generation models incorporating unobserved small trips using probe person data

Sohta ITOH [a], Eiji HATO [b]
[a] *Department of Urban Engineering, School of Engineering, The University of Tokyo*
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[b] *Department of Civil Engineering, School of Engineering, The University of Tokyo*
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[a] *E-mail: itoh@bin.t.u-tokyo.ac.jp*
[b] *E-mail: hato@bin.t.u-tokyo.ac.jp*

Abstract: Probe Person (PP) survey with Global Positioning Systems (GPS) and web diary enable us to collect much higher resolution data in real time. The new method of travel behavior expected to be fused with Questionnaire type travel survey which has been the basis of travel demand data in transportation planning. In the conventional questionnaire based survey, the problem of missing trips and activities is pointed out that caused non-response bias. This study considers the mechanism of missing activities in Person Trip (PT) surveys and proposes a combined estimation method using a sample selection model. The method contributes to correcting the bias of missing data by data fusion of PP data and PT data. The empirical results show that several personal attributes and contexts of the activities are found to be significant factors of missing of activities in paper-based PT surveys.

Keywords: Probe Person Survey, Person Trip Survey, Household Survey, Combined Estimation, Missing Trip, Non-Response Bias

1. INTRODUCTION

1.1 Research Background

Recently, tracking method with mobile instruments is used as a travel survey to replace conventional survey methods. Probe Person (PP) survey is the new survey approach based on the mobile assisted tracking method. PP survey method using GPS (Global Positioning Systems) phone and web diary is proposed in Hato et al. (2006). Originally, the method is based on tracking systems using PHS (Asakura and Hato, 2004). As both mobile instrument technologies and GPS have rapidly advanced in recent years and are wide spread, accurate travel information can be observed continuously in real time. Probed tracking data are available for transportation planning and on-line monitoring of travelers in urban area.

Traditional questionnaire type travel surveys have been widely conducted for establishing urban transportation planning. Person Trip (PT) Survey or Household Travel Survey belongs to this type of surveys. In PT Surveys, respondents are sampled at random from population in the target urban area and asked to fill in paper-based travel diaries. Though PT Surveys are useful to estimate inter-zone OD travel demands, it is pointed out that respondents underreport trips and activities when asked their travel behavior. In questionnaire surveys, the respondents have to rely deeply on their memories because the questions are directed the trips undertaken in the previous days. Engineers engaged in transportation planning are eager to collect detailed attributes and accurate travel information to analyze travel demands but it is not easy for respondents to remember complete information of all trips they did. Especially, it is difficult to answer exact arrival times, departure times and places of activities. Research shows that short trips and activities are often underreported because the survey participants cannot recollect such trivial activities or they do not think of these short trips and activities as meaningful travel information to report. Brog (1982) identified three types of non-reported trips: (a) trips that were not reported by the respondents due to increasing lack of care in case of survey periods of several days' length, (b) trips that were not reported by the respondents because they forgot or considered them redundant, and (c) trips that the respondents did not want to report on the basis of their own deliberate decisions. The lack of activities causes biases that may result in erroneous activity-travel analysis. Too many inquiries to answer may produce non-response biases.

1.2 Panel Data Analysis

Prior to the advent of the Probe Person surveys using cell phones, panel analysis has been applied in travel behavior planning using longitudinal data (Kitamura, 1990). Panel analysis is the type of statistical analysis that consists of data observed repeatedly for the same samples. Panel study enables evaluating day-to-day behavioral dynamics. The advantages of panel data is that variation in travel behavior and its long-term change can be observed.

Though panel surveys give us important findings, surveyors have to pay attention to some problems. One of the major difficulties is the problem of attrition bias due to drop-out early from the survey. Attrition bias is a selection bias caused by dropping out of respondents between waves of surveys. Kitamura and Bovy (1987) considered attrition by constructing a probabilistic model of attrition and introducing the weigh that took advantage of mobility and other available information for the data.

1.3 Development of GPS-based Surveys

However, GPS-based surveys have made rapid progress in this decade. In GPS-based surveys, survey burden for respondents is much less than that for PT surveys. Easy recording should prevent recording omissions. In this respect, GPS-based surveys are considered to be superior to conventional questionnaire type travel survey methods. PP Surveys using web diary and GPS phones, as mentioned later, are also GPS-based surveys. The GPS-based methods have been developed since late in the twentieth century. For example, Zitto and D'este (1995) and Sermons and Koppleman (1996) used GPS in surveys to measure vehicle trips. Murakami and Wagner (1999) applied a tracking method using GPS and PDA (Personal Digital Assistant) for drivers. Their results show that self-reported distances by using written diaries are much longer than distances recorded by GPS. Bricka and Bhat (2006) compared vehicle driver trips of a GPS-based survey and a conventional travel survey and provided the factors affecting trip underreporting. Stopher and Greaves (2010) also compared GPS data and diary records and indicated discrepancies between the duration of trip and distance of travel between two types of data. According to the study of Wolf (2001) and Wolf (2004), underreporting rates of traditional travel surveys have ranged from 11 to 81 percent, depending on a range of socio-demographic attributes and trip situations. Itsubo and Hato (2006) concluded the difference of 5 percent was shown between the average number of trips per day of a GPS-based survey and a paper-based survey. In the study of Bricka and Bhat (2006), 40 percent of drivers missed trips.

The latest papers mentioned that GPS-based surveys are studied for further development. Stopher et al. (2011) conducted GPS-only survey for over 3,500 households.

Sneade (2011) reported effectiveness of introducing measuring three-axis acceleration in travel surveys.

The objective of this research is to make good use of PP data and PT data for accurate travel behavior. This paper also focuses on the missed activities of PT data. It has long been suspected that people often omit their travel in paper questionnaire-based surveys and researchers have analyzed the gap between self-reported trip data and recorded data by GPS, as we mentioned before. However, most of these studies deal with only vehicle trips and they do not sufficiently investigate factors that cause missing data. In this study, we assume that PP data does not miss any activities because respondents record their activities on the spot with their portable instruments. We compose a combined estimation model to investigate the socioeconomic and spatial variables correlated with missed activities. In addition, we propose the method to correct the bias of missing activities of PT data by weighting each activity with the multiplicative reciprocal of the probabilities of its observation.

1.4 Outline of Probe Person Survey

PP survey method is the method using an automatic position and time recording systems based on GPS and Internet communications. The aims of PP survey system is 1) ensuring accurate travel records through space-time position determination functions of high accuracy, 2) reducing recording omissions through timely travel behavior recording functions, and 3) ensuring improved efficiency of data coding by the investigator and improving the sense of participation in surveys of subjects through a survey system that emphasizes real-timeness and interactiveness. (Timmermans and Hato, 2009) In PP surveys, respondents carry portable travel-activity measuring instruments. In addition to basic trip data, PP surveys also enable researchers to observe various data by using sensors of mobile instruments. Hato (2010) proposed a method for estimating behavioral contexts using a portable instrument called BCALs (Behavioral Context Addressable Loggers in the Shell). BCALs has several environmental sensors that can measure its location, sound, air pressure, and acceleration without any handling by a respondent. The most remarkable characteristic of PP survey is measuring various kinds of data such as text, photo, sound and acceleration. With PP Survey, it is expected that we can know by far more data about travel behavior. These data can be obtained during trips and activities with sensors and explosive spread of smartphones enable us to get these data easily. By applying SVM (Support Vector Machine) and HMM (Hidden Markov Model), activity patterns are estimated automatically from travelers' location and acceleration records. Measured data can also be applied to evaluate the level of service in modeling.

The methods of PP survey and PT survey are compared in Figure 1. Activities and trips are recorded by cell phone with GPS function when leaving the trip origin, transferring the mode, and arriving at the destination. Respondents operate cell phones at each time. Paper-based surveys depend on respondents' vague memories, while accurate observation is expected in PP surveys, as data are recorded in real time. An application of the cell phone records the data of trip OD, travel mode, trip purpose, time of departure, time of mode change, time of arrival, and trajectory data during trip. Figure 2 shows the typical example of activity tour in PP survey. As shown in the figure, travel trajectory data are recorded by GPS every several seconds. By analyzing this data, it can be easily confirmed what path they went along and how fast they traveled. In PP surveys, respondents can correct their data on the Internet. For example, if a respondent forget to enter the data in the cell phone when leaving a previous activity place, he/she can simply connect to web and correct the departure time. To ensure

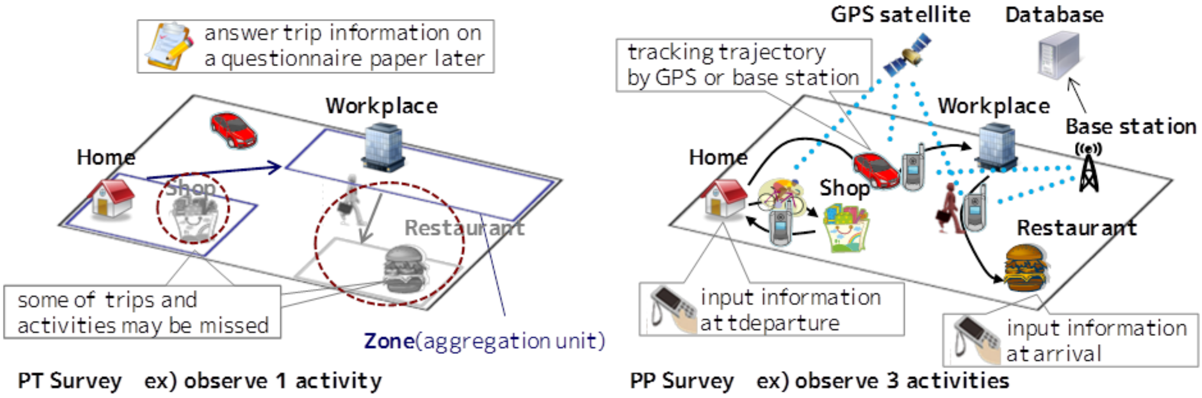

Figure 1. The methods of PP survey and PT survey

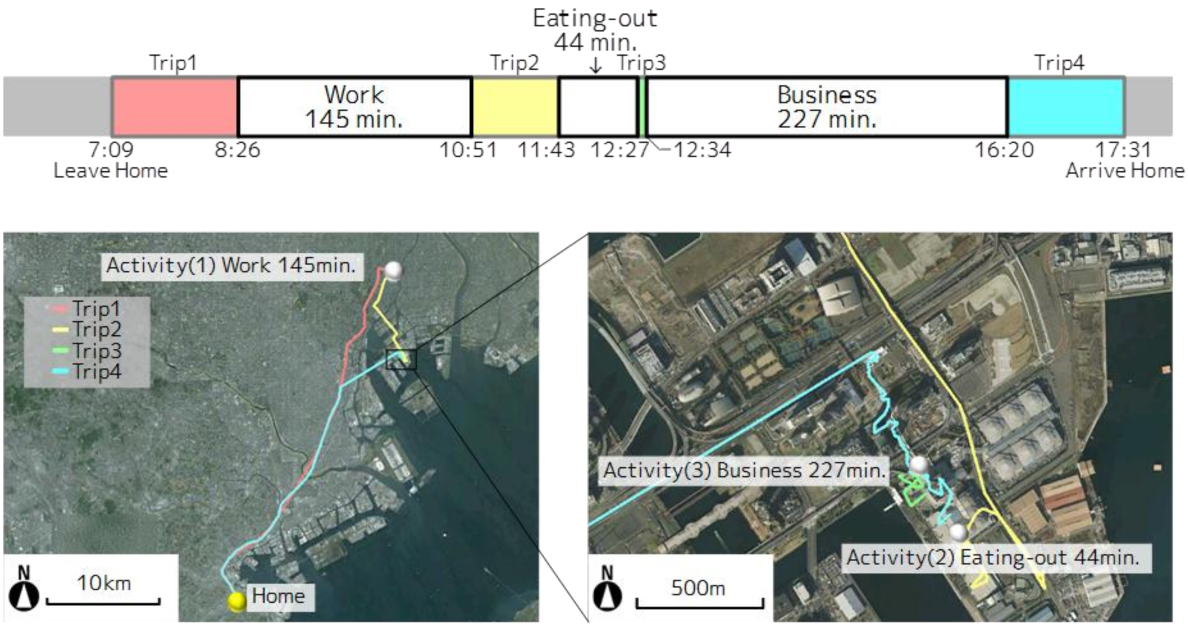

Figure 2. Example of activity tour and trajectory observed in Probe Person survey

respondents' convenience, the design of the survey is important. Timmermans and Hato (2009) discussed instrument design and interfaces of the surveys so as to improve data quality. They suggested that visualization motivated respondents to participate in the survey.

Another advantage of PP surveys is that long term data can be observed for same respondents as panel data. With PP data, we can distinguish variability within individuals and variability among individuals. Hato (2006) investigated variability of trip-activity patterns by monitoring PP data of the same individuals over 5 weeks.

As the resolution is higher, respondents' privacy should be taken care of. The trajectory data enable researchers to detect respondents' residence and workplace directly. Some people refuse to join the survey because they are reluctant to offer their detailed behavior, so protecting private information is a problem still to be solved.

Table 1. Outline of Probe Person survey in Yokohama

Surveillance period	35 days (2010/07/05 - 2010/08/08)
Survey methods	Probe Person survey with GPS cell phone + Web diary
Area	Yokohama Metropolitan Area, Japan
The number of samples	40 people
The number of trips	3,617 trips
The number of location data	789,074 points

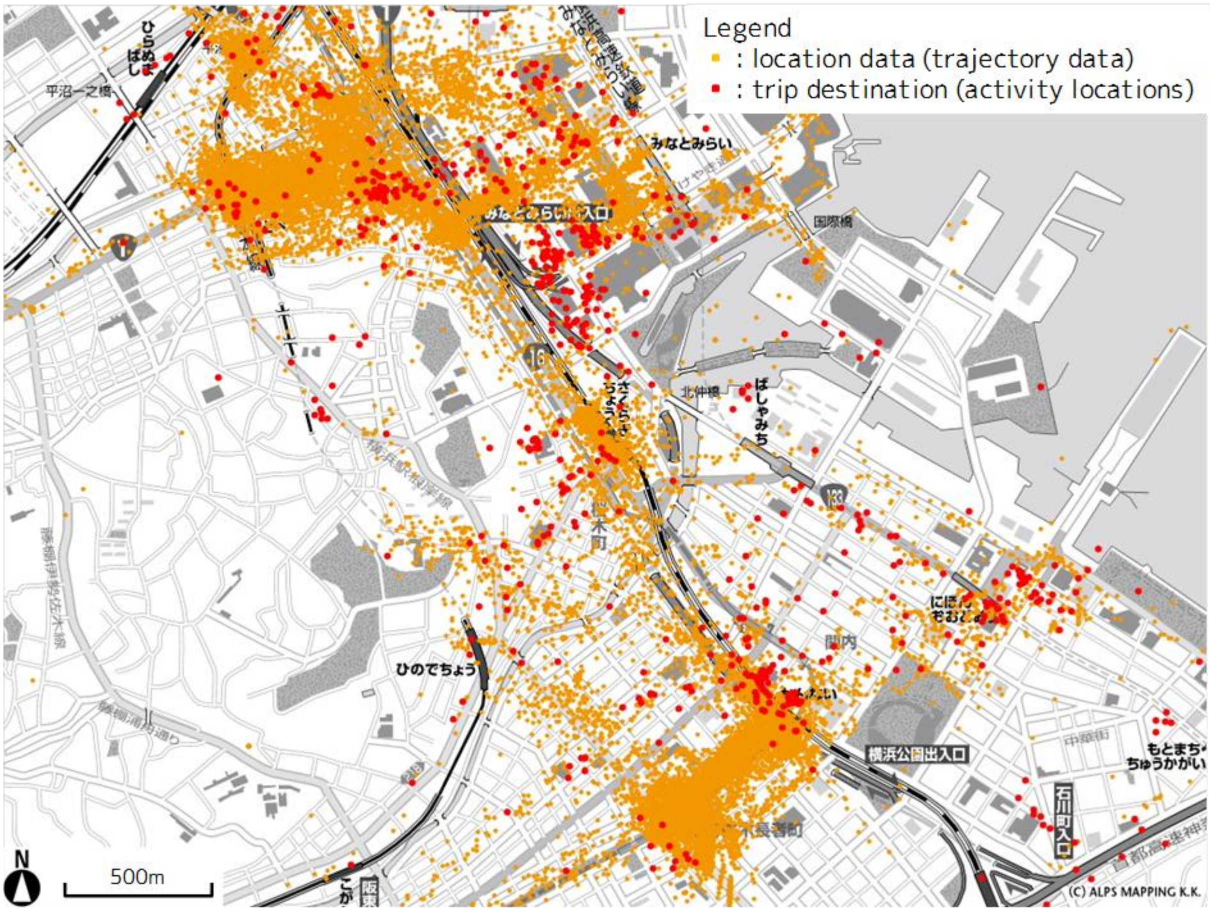

Figure 3. Mapping of trajectory data and trip destinations

2. ELEMENTAL ANALYSIS USING PROBE PERSON SURVEY

2.1 Probe Person Survey

The PP data we used was obtained in Yokohama, Japan. Table 1 shows the overview of the data. Surveillance period was from July 5th to August 8th in 2010. 40 respondents participated in the survey. All the respondents were resident of Yokohama and used to commute daily to the city center. They agree to the privacy policy of the survey. During this survey, the respondents recorded their trips and activities by cell phones with GPS function. Their location data were obtained automatically during their trips. The trajectory data and activity locations observed by GPS in the PP survey are shown in Figure 3. Mapping these data indicates that the density of trajectories and activities differ from area to area. In the 35-day surveillance period, 789,074 GPS location data and 3,617 trip data were recorded.

GPS Instrument records its location every 5 seconds during trip. Although GPS has

Table 2. Outline of Person Trip survey

Surveillance period	2008/10 - 2008/11 (each respondent answers his/her traffic behavior of 1 day in surveillance period)
Method	Paper questionnaire
Area	Tokyo Metropolitan Area, Japan (including Yokohama)
The number of trips	1,906,032 trips
The number of trips in Yokohama	253,737 trips

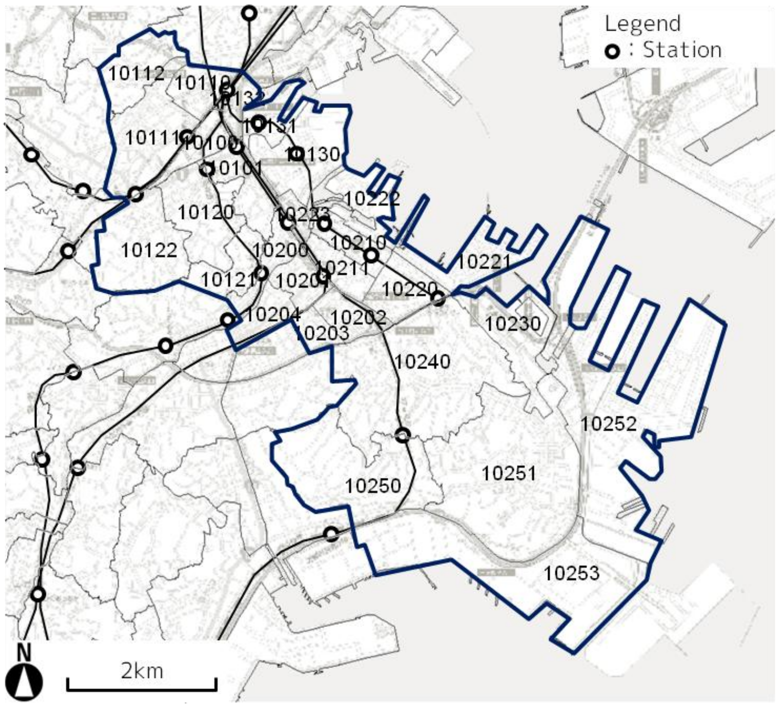

Figure 4. Analysis area and zone code in Yokohama

some measurement error especially among high buildings, horizontal accuracy of an instrument GPS is about 100 meters, so accuracy is very minor issue to detect origin and destination zone. In analyzing activities in metropolitan region, this measurement error has little effect to estimation.

2.2 Person Trip Survey

The PT data used in this study is extracted from 5th Person Trip Survey in Tokyo Metropolitan Area, Japan. The survey was conducted in 2008. Table 2 shows the overview of the data. The survey was executed by having respondents complete paper questionnaire forms and collecting the forms by mail. Respondents registered all of the trips of one day. The sampling rate of the survey was 2.7 percent and 1,906,032 data were obtained.

In the PT survey, zone is used as an aggregation unit. Zones around city centre area of Yokohama are shown in Figure 4. In the figure, five-digit numbers stand for zone codes. In this study, because the trip data are aggregated to these zones, we consider the centroid of zones as activity locations.

We translate these massive data to activity data including purpose, duration, starting and ending time, and place of activities.

Table 3. The number of activities and the sum of activity duration of activity tours

	The number of activities					The sum of activity duration				
	mean		s.d.		t-statistics	mean (min.)		s.d. (min.)		t-statistics
	PT	PP	PT	PP		PT	PP	PT	PP	
age 20s	1.26	1.39	0.69	0.75	2.62*	457.0	544.0	213.1	221.0	5.95*
age 30s	1.40	1.60	0.91	0.85	3.12*	426.9	389.0	251.3	274.3	1.84
age 40s	1.53	1.74	1.06	1.22	2.63*	445.0	288.5	267.2	277.9	8.60*
age 50s	1.55	1.80	1.02	1.20	1.98*	412.2	325.9	249.3	219.4	3.73*
age 60s+	1.56	1.58	0.96	0.85	0.19	233.3	298.2	205.7	324.2	1.63
male	1.49	1.78	1.02	1.19	4.86*	459.7	497.9	249.8	262.6	2.90*
female	1.43	1.43	0.85	0.73	0.00	309.1	281.7	232.5	254.9	2.14*
total	1.46	1.60	0.94	0.73	5.39*	383.0	389.5	252.6	280.4	0.65

* : reject the null hypothesis of no difference between the mean of PT data and that of PP data at 5% significant level

2.3 Comparison between PT Data and PP Data

In order to predict activity generation, it is desirable to enhance advantages of both PT and PP data. PP survey enables collecting accurate and high-resolution data than conventional methods but application is limited to small survey so far. On the other hand, conventional PT survey is a type of large scale survey and representativeness of population is ensured statistically.

Actually, some trial large-scale PP surveys using smart phones are reported. (Transportation Research Board's Travel Survey Methods Committee, 2012) For instance, “The Quantifable Traveler app” was developed by UC Berkeley and acted as passive data collector with 80 people in a survey. Future Mobility Survey in Singapore also designed smartphone-based survey and approximately 150 people joined the survey in 2012. However, these respondents are much fewer than respondents in large-scale person trip surveys. The estimation error is due to sample size so small PP surveys are inferior to large-scale PT surveys in reliability. Furthermore, according to the survey by Google, the rate of smartphone is 20% in Japan and 44% in U.S., which are not high enough to conduct random sampling surveys so far. Because of limitations, PP surveys cannot totally replace PT surveys so it is worth developing a method to use both PT and PP data.

In this subsection, we will describe elemental analysis of PT data and PP data about the significant difference between two data. Table 3 provides the number of activities and the sum of activity duration of activity tours. In the table, the mean and the standard deviation of the duration and the number of activities are calculated. In addition, to examine the difference statistically, we apply Student’s t-tests for the two sets of PT and PP data when comparing. As can be seen from Table 3, in almost all of categories, the number of activities of the tour of PT data is smaller than that of PP data. As of ages, young people tend to show larger difference than older people. In age 20s, the reported trips of PT survey average 1.26 while the recorded trips of PP average 1.39. However, in age 60s, the difference is not seen at significant level. The means of trips of PT in age 60s is 1.56 while that of PP is 1.58. Between male and female, male respondents seem to be more likely to omit their activities than female respondents. In male, there is statistically significant difference between PP and PT but the PP trips and PT trips of female averages same.

The average sum of activity duration of a tour for all respondents is 383.0 minutes in PT and 389.5 minutes in PP data. As a whole, there is not much significant difference of activity duration.

3. PT/PP COMBINED ESTIMATION

3.1 Estimation Method

This section deals with a combined estimation about activity generation using both PT and PP data. In previous studies, combined estimation method using individual term for different type of data and handling error term was applied for revealed preference (RP) and stated preference (SP) data (Morikawa, 1993). The gap of the mean and the standard deviation of the error terms between PT data and PP data is able to be handled same as the gap between RP and SP data. We describe activity generation by a probit selection model. The merit of a probit selection model using normal distribution as an error term is the operability when considering data's interdependence and comparing the measurement of error terms. On each day, binary variable y_{in} which represent if individual i does an activity in zone n or not is described as

$$y_{in}^{*} = \beta x_{in} + \varepsilon_{in}$$
$$\begin{cases} y_{in} = 1 & if \;\; y_{in}^{*} > 0 \\ y_{in} = 0 & if \;\; y_{in}^{*} \le 0 \end{cases} \tag{1}$$

where y_{in}^{*} is the latent variable about possibility of activity generation; β is a vector of coefficients; x_{in} is a vector of explanatory variables of individual i and zone n; and ε_{in} is a random term which is independent of each activity and it is normally distributed ($\varepsilon_{in} \sim N(0,1)$).

We describe that activities of PT data are likely to be estimated lower than those of PP data in Section 3. To consider non-reported activities, we assume that PP data would be closer to the real value about activity generation. Based on this assumption, we introduce a correction term for considering the systematic bias caused by the difference of survey styles. Activity generation model for PT data can be rewritten as

$$y_{in}^{PT*} = \beta x_{in} + \gamma + \varepsilon_{in}^{PT} \tag{2}$$

γ includes specific bias only in the PT survey, which is a correction term for distributions of activity generation.

In combined estimation, variance of the error term can also be written as

$$Var(\varepsilon_{in}^{PT}) = \mu^2 Var(\varepsilon_{in}) \tag{3}$$

The error term of the PT data has $N(0, \gamma)$ normal distribution.

3.2 Sample Selection Model

In Section 3.1, we assume that the bias affects PT data uniformly but the effect of missing activities is estimated to depend on its activity duration, activity environment, context, and personal attribute. If an activity which is actually done is not recorded in survey and if these missing activities have some characteristics in common, sampling bias affects the estimation result. It is quite probable that some factors influence the propensities to record activities.

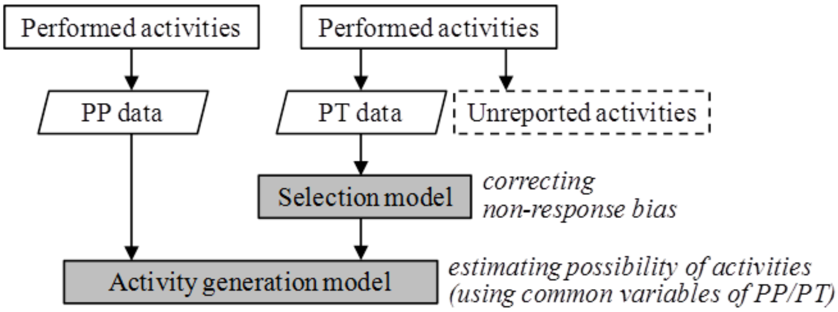

Figure 5. Overview of combined estimation considering non-response bias

This type of bias is often called non-response bias. In order to consider this issue, we developed sample selection model to estimate activity generation. The overview of this method is shown in Figure 5.

The latent variable about activity generation is written in simple linear model same as previous section.

$$
\begin{aligned}
& y_{in1}^{*} = \beta_1 x_{in1} + \varepsilon_{in1} \\
& \begin{cases} y_{in} = 1 & if\ \ y_{in1}^{*} > 0 \\ y_{in} = 0 & if\ \ y_{in1}^{*} \le 0 \end{cases}
\end{aligned} \tag{4}
$$

Besides this outcome model, we insert selection model in activity generation model to describe missing effect for PT data as

$$y_{in2} = \beta_2 x_{in2} + \varepsilon_{in2} \tag{5}$$

$$\begin{pmatrix} \varepsilon_{in1} \\ \varepsilon_{in2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \eta\sigma_1 \\ \eta\sigma_1 & 1 \end{pmatrix} \right) \tag{6}$$

where y_{in2} is an unobserved variable for selection; x_{in2} is a vector of explanatory variables which have influences on missing mechanism. For equation (5) and (6), only when unobserved y_{in2} variable exceeds a particular threshold, y_{in1} is observed. Let the threshold be zero. If $y_{in2} > 0$, latent variable y_{in1} is valid to calculate the possibility of activity in each zone. Then, the expected value of latent variable y_{in1} after considering bias of selection is written as

$$
\begin{aligned}
E(y_{in1} \mid y_{in2} > 0) \quad &= \beta_1 x_{in1} + E(\varepsilon_{in1} \mid \varepsilon_{in2} > -\beta_2 x_{in2}) \\
&= \beta_1 x_{in1} + (\eta\sigma_1)\frac{\varphi(\beta_2 x_{in2})}{\Phi(\beta_2 x_{in2})}
\end{aligned} \tag{7}
$$

where Φ is the cumulative distribution function of the standard normal distribution and ϕ is the probability density function. We apply the selection model to expression of unrecorded activities in PT data.

3.3 Estimation Results

In this section, we present our estimation results and discuss the model of activity generation

Table 4. Estimation results

Independent variables	The normal activity generation model			The sample selection model		
	Parameter	t score		Parameter	t score	
For activity generation model						
Constant	-1.902	-76.64	*	-1.808	-79.24	*
Male	0.091	12.59	*	0.069	7.51	*
Age ≧ 60	-0.116	-15.37	*	-0.106	-10.89	*
Single-member household	0.090	8.79	*	0.100	7.73	*
Car ownership	-0.003	-0.42		-0.002	-0.17	
Distance from home (km)	-0.108	-98.83	*	-0.117	-58.97	*
Distance from workplace (km)	-0.025	-43.52	*	-0.028	-35.70	*
Store space (ha) [1)]	0.043	71.31	*	0.035	39.55	*
γ	0.125	5.09	*	-	-	
η	-	-		0.435	16.94	*
For selection model						
Male	-	-		0.466	14.18	*
Age 20-39 years	-	-		-0.545	-7.07	*
Age ≧ 60	-	-		0.355	4.20	*
Distance from home (km)	-	-		0.071	0.66	
Distance from workplace (km)	-	-		0.020	0.23	
Stay Duration (min.)	-	-		0.044	4.99	*
μ	-	-		3.557	17.67	*
Observations (PT) n	1,780,164			1,780,164		
Observations (PP) n	23,000			23,000		
Initial log-likelihood	–1,249,858			–1,249,858		
Final log-likelihood	–65,013			–64,272		
Rho-squared $\bar{\rho}^2$	0.948			0.949		

- Not relevant; * Significant at 1% level.

1) : The sum of space about retail stores in the zone

and the effect of sample selection of missing activities in PT data. Estimation results are shown in Table 4. The normal activity generation model is the model described in Section 3.1. The sample selection model is the model in Section 3.2 considering non-response bias in PT surveys.

We consider individual attributes, and the distance from home and a workplace to an activity place as explanatory variables. As of the sample selection model, we integrated gender and age on the hypothesis that these factors affect trip reports in PT data, as a result of the comparative analysis in section 2.3. Bricka and Bhat (2006) found that trip distance affects the magnitude of trip under-reporting propensity, so we integrate the distance from home, the distance from workplace, and activity duration into the model. We defined the distance as a length in a straight line from the activity place to home or a workplace.

In general, both models show almost same tendency about activity generation. Male tend to do more activities than female and elder people show fewer activities in common. The distance from home and workplace is also an important factor to activity. As of the goodness of fit, the sampling selection model is slightly superior to the normal model. Inserting the selection model improves accuracy of activity demand prediction.

The sample selection model result shows that demographic variables and

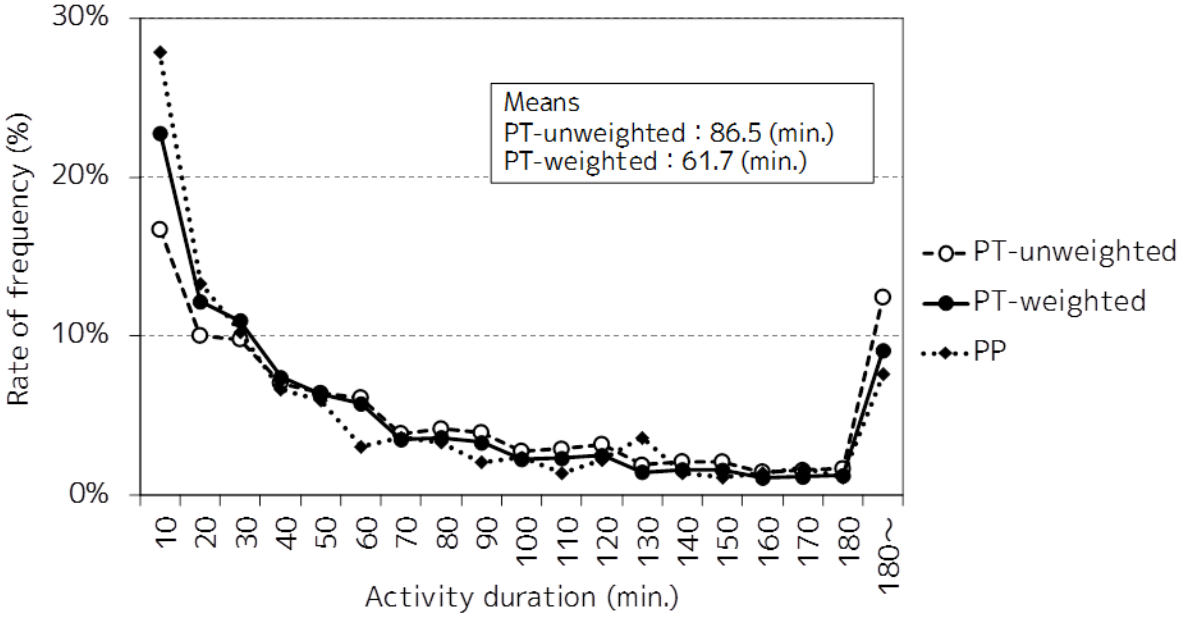

Figure 6. The frequency of activity duration

characteristics of activities associate with activity under-reporting. As of demographic variables, female and less than 40 years of age are likely to miss their activities. Stay duration is also found to have impact on activity missing at significant level. Judging from parameter value, activity duration seems to be one of the strongest factors of missing. The result indicates that activities of short duration have a higher propensity to be unobserved activities.

4. CORRECTING SAMPLING BIAS

In considering traffic planning, analyzing survey data is necessary as a basic study. When aggregating PT data, sampling bias due to missing activities should be incorporated for accurate travel demand. From the proposed model, the probabilities of observation for activities can be obtained. In order to correct the bias and obtain the true distribution of activity generation, the reciprocal of observation probability is considered as the weight. Dividing by the observation probability, it is able to estimate the true activity generation. For example, if the type of activity with some specific attributes is observed with a probability of 25%, true activity generation is estimated to be four times (1/0.25) as much as observed. We define the weight of activity of individual i in zone n as:

$$w_{in} = \frac{1}{p(y_{in2} > 0 \mid x_{in2})} = \frac{1}{\Phi(\beta_2^* x_{in2})} \tag{8}$$

where β^* is the parameter estimated in the previous section.

We show the example of aggregate result with this weighing. As a result of using the method for PT data, the frequency of activity duration compared between unweighted data and weighted data is shown in Figure 6. The figure indicates that the composition of duration of weighted PT data is quite similar to that of PP data over 30 minutes. Compared with weighted data and unweighted data, we can find that short activities are underestimated in questionnaire type PT survey. The component ratio of the activities whose duration is less than 10 minutes is 22.8% in weighted PT data, which is 37% higher 16.7% in unweight PT data. The ratio of long duration activities is also clearly different between unweighted and weighted data because the great part of these activities is estimated to be observed in PT surveys. Weighting leads to increase the proportion of short activities and lessen that of long activities. As a result, the mean of activity duration is 86.5 minutes in unweigted PT data and 61.7 minutes in weighted PT data. The result agrees with previous studies in respect of omitting short trips and activities in self-reported data.

5. CONCLUSION

We have discussed the advantages of both new GPS-based PP surveys and conventional PT surveys. PP surveys enable collecting more accurate, high-resolution, and real-time data than paper-based PT surveys. PP surveys have paved the way for detailed traffic planning especially in urban area. However, survey scale is too small to ensure representativeness of population so far because it is difficult to prepare enough devices and applications for large scale surveys.

PT surveys also have disadvantages. In PT surveys, recording respondents' behavior depends on their memories and some activities are not recorded. This effect is not negligible issue when estimating activity generation in traffic planning. Our analysis shows the number of activities per day is statistically different between PP and PT data. The results suggest that age and activity duration is considered to be the factor of non-response bias.

We described two types of combined estimation about activity generation and compare them. To examine the factor that associates to the likelihood of missing activities in PT survey, we introduced a sample selection model into the activity generation model and applied the model in the empirical analysis. As a result, we confirmed not only the improvement of activity generation model but also the presence of missing activities statistically and several demographic attributes and activity characteristics associate if activities are missed or not. Proposed model can estimate how often activities are missed in PT survey quantitatively. This method can be developed to correct the non-response sampling bias due to missing activities in PT data. By multiplying the reciprocal of probabilities of missing obtained from the selection model for each activity, we can assess and correct the bias. The application of the weight shows that we can estimate accurate data by correcting.

This study implicates that GPS-based PP survey help to observe high-resolution travel behavior data and avoid missing trips and activities. With this accurate data, the combined estimation model using PP data and PT data contributes accounting for the bias between different types of surveys, so that data does not lead to inappropriate analysis. PP survey method presents the possibility of collecting much more precise travel behavior data. Using both new method and conventional data enhances travel behavior analysis and contributes to traffic planning in metropolitan region, spatial planning in inner city, urban design, or so.

REFERENCES

Asakura, Y., Hato, E. (2004) Tracking survey for individual travel behavior using mobile communication instruments, *Transportation Research C*, 12(3-4), 273-291.

Bricka, S., Bhat, C. (2006) Comparative analysis of global positioning system-based and travel survey-based data, *Transportation Research Record*, 1972, 9-20.

Brog, W., Erl, E., Meyburg, A. H., Wermuth, M. J. (1982) Problems of nonreported trips in surveys of nonhome activity patterns, *Transportation Research Record*, 891, 1-5.

Google (2012) The our mobile planet. http://www.thinkwithgoogle.com/mobileplanet/

Itsubo, S., Hato, E. (2006) A study of the effectiveness of a household travel survey using GPS-equipped cell phones and a web diary through a comparative study with a paper based travel survey, *Transportation Research Board 85th Annual Meeting*, CDROM.

Kitamura, R., Bovy, P. (1987) Analysis of attrition biases and trip reporting errors for panel data, *Transportation Research A*, 21(4-5), 287-302.

Kitamura, R. (1990) Panel analysis in transportation planning: an overview, *Transportation Research A*, 24(6), 401-415.

Hato, E., Itsubo, S., Mitani, T. (2006) Development of MoALs (Mobile Activity Loggers supported by gps-phones) for travel behavior analysis, *Transportation Research Board 85th Annual Meeting*, CDROM.

Hato, E. (2006) Evaluation of trip-activity pattern variability using probe person data, *Transportation Research Board 85th Annual Meeting*, CDROM.

Hato, E. (2010) Development of behavioral context addressable loggers in the shell for travel-activity analysis, *Transportation Research C*, 18(1), 55-67.

Murakami, E., Wagner, D. P. (1999) Can using global positioning system (GPS) improve trip reporting? *Transportation Research C*, 7(2-3), 149-165.

Rubin, D. B. (1976) Inference and missing data, *Biometrika*, 63(3), 581-590.

Sermons, M. W. and Koppelman, F. S. (1996) Use of vehicle positioning data for arterial incident detection, *Transportation Research C*, 4(2), 87-96.

Sneade, A. (2011) Using accelerometer equipped GPS devices in place of paper travel diaries to reduce respondent burden in a national travel survey, 9th International Conference on Transport Survey Methods.

Stopher, P., Greaves, S. (2010) Missing and inaccurate information from travel surveys: pilot results, *32nd Australasian Transport Research Forum (ATRF)*.

Stopher, P. R., Prasad, C., Wargelin, L., Minser, J. (2011) Conducting a GPS-only household travel survey, *9th International Conference on Transport Survey Methods*.

Morikawa, T. (1994) Correcting state dependence and serial correlation in the RP/SP combined estimation method, *Transportation*, 21, 153-165.

Timmermans, H. J. P., Hato, E. (2009) Electronic instrument design and user interfaces for activity-based modeling, *Transport Survey Methods keeping Up With a Changing World*, Emerald Group Publishing Ltd., 437-462.

Transportation Research Board's Travel Survey Methods Committee (2012) The on-line travel survey manual, Chapter 26, Using smartphones for travel behavior studies. http://www.travelsurveymanual.org/

Wolf, J. (2004) Applications of new technologies in travel surveys, *7th International Conference on Transport Survey Quality and Innovation*.

Wolf, J., Loechl, M., Meyers, J., Arce, C. (2001) Trip rate analysis in GPS-enhanced personal travel surveys, *5th International Conference on Transport Survey Quality and Innovation*.

Zitto, R., D'este, G., Taylor, A. P. (2007) Global positioning system in the time domain: how useful a tool for intelligent vehicle-highway systems? *Transportation Research C*, 3(4), 193-209.