STUDY ON DATA COLLECTION PROCESS FOR FEATHERS ON HOUSEHOLD TRAVEL SURVEY IN KOREA

S. CHO [a], B. KOCHAN [a], T. BELLEMANS [a], W.D. Lee [b], J.H. HWANG [b] and C.H. JOH [b]
[a] Transportation Research Institute (IMOB), Hasselt University, Belgium
Email: sungjin.cho@uhasselt.be; bruno.kochan@uhasselt.be; tom.bellemans@uhasselt.be
[b] Department of Geography, Kyung Hee University, Republic of Korea
Email: whiteowl@khu.ac.kr; hwangjhn@naver.com; bwchjoh@khu.ac.kr

ABSTRACT

The growing interest in activity-based approaches has led to more necessity of activity-travel diary data. Since there has been no concrete framework of activity-travel survey, most of required data are obtained from a typical travel survey. Hence, a prerequisite of the activity-based approaches is essential to figure out erroneous data that might occur in deriving activity information from such a trip-based survey and to eliminate an unexpected error in implementation. The goal of this study is therefore to propose an appropriate data processing method for FEATHERS on household trip survey in Korea. In line with the goal, this paper will be followed by a brief explanation of using a household travel survey in Korea. A type of errors in the survey will be closely dealt with in accordance with the data requirement for FEATHERS. A data processing method will be then proposed by producing some cases of those types in the survey. In the end, it will conclude with summary of this study and a research agenda for the future.

Keywords: activity-based model, household travel survey, FEATHERS, data processing method

1. INTRODUCTION

The growing interest in ABM (activity-based transport model) has led to more necessity of activity-trip diary data. In previous researches, data from HTS (household travel survey) have been typically used for ABM. Since HTS collects household/ personal socio-economic data and also individual travel diary by asking respondents to record their daily trips over given time period, it does not always meet the data requirement for the ABM. Because implementing ABM needs not only trip data that can be achieved from HTS, but also activity information executed by individual in the household. Hence, the activity information is necessarily derived from the trip data reported by respondents in HTS. As a result, some biases or errors might be included in some of deriving activity information. Moreover, if the survey data are incomplete or inconsistent no matter by respondents' mistake or else, then it makes more difficult to achieve elaborate data from the survey. Thus, a prerequisite of ABM is necessary to figure out those erroneous data that might occur in deriving activity information from HTS and to process it to avoid an unexpected error in implementing ABM.

Since Clarke et al (1981), abundant research have dealt with issues on data collection (i.e. HTS) for implementing ABM. Two of the research are significantly related to this study. One is that Arentze et al. (1999) set a number of rules for verifying inconsistent and incomplete data in the survey and adopt some solutions to the rules, and at the end, they developed an interactive computer system for the logical verification and inference of activity, which called SYLVIA, for their activity-based model system, called ALBATROSS. The other research is that Přibyl et al. (2003) specifies particular rules of data cleaning process applied for the CentreSIM survey data, which is to improve a transportation modeling in CentreCounty, Pennsylvania. Those two research closely diagnose all the possible cases arose in data collection process and present some strategy of treating the inconsistent and incomplete records in the survey. Although the concept and direction of those research is quite similar with this study, this study intends to concentrates on the structure and pattern of erroneous data occurred in applying trip-based survey to activity-based research, in contrast with two other studies that concern only activity-based survey. It can differentiate this study form the other studies because there might be more challenges in data processing for trip-based survey. We will discuss it later on.

In line with that, this study aims to propose an appropriate data processing method for ABM by treating such issues on FEATHERS, which is an activity-based simulation system, using HTS in

South Korea. This paper will be followed by an overview of the household travel survey in South Korea, 2006. The system overview and the data requirement of FEATHERS will be then briefly explained in the next section. Followed by, a type of errors will be dealt in accordance with the FEATHERS requirement and a data processing method as a solution to those error types will be discussed by introducing some cases of those types of errors in HTS. In the end, this paper will conclude with summary of this study and a research agenda for the future.

2. DATA AND MOETHOD

2.1.1 Household travel survey in Korea

A Korean household trip survey has been conducted every 5 year by MTA (Metropolitan Transportation Authority) since 2002. The survey aims to predict travel demands and to assist policy makers and practitioners by providing a valuable information about policy impacts on individual travel behaviour in Seoul metropolitan area. The survey asks respondents to report socio-demographic characteristics of individual and household and trips executed in a given day. In 2006, the survey used for this study collected trip diary from 213,610 households (2.9% of population in the study area), which is chosen by using cluster and random sampling methods (MTA, 2007).

The survey consists of three categories: household, individual and trip. Each of these categories includes related information. The household category takes into account basic information about household residence and composition and trip-related data. The person category covers socio-demographic characteristics of each household member. At last, the trip category describes all components of trips conducted by the individual member of the household, which contains a trip order, purpose, and origin and destination time and location.

2.2 FEATHERS

FEATHERS, (Forecasting Evolutionary Activity-Travel of Households and their Environmental RepercussionS) is an activity-based transportation model system for supporting transportation politician and planner in Flanders, Belgium. The system was designed to extend the applicability and transferability with its modular architecture. An ALBATROSS is embedded as a main scheduling module inside the system (Bellemans *et al.*, 2010). FEATHERS can predict individual daily activity-trip schedule in order to analyze activity-trip patterns in the study area of interest. Thus, using FEAHTHERS allows us to assess an alternative transport policy and suggest the future plan by means of measuring its impact on individual daily life.

FEATHERS requires several data sets as an input for implementing the system. The FEATHERS input contains diary, population and environment data. Among those data sets, the diary data is going to be further discussed in this research. Household data describes a residence location, household composition, an age class of the household members (which are elder and child), income level and a number of vehicles in the household. Person data depicts a basic socio-demographic attribute belong to the each members of the household. Note that FEATHERS only considers a householder and his/her partner as a study target so that the person data are only correspondent to those two adults in the household. In contrast to the composition of the HTS, the FETHERS uses activity data as input that accounts activity facets, such as activity type, location and time (departure and duration, and day of week). At last, the components of the trip data are almost same as those of HTS.

3. EXPERIMENT

3.1 Classification of Errors

As mentioned in the introduction, it is indispensable to dispose biases and errors introduced in deriving required information from HTS in order to prevent from a failure in implementing ABM, like FEATHERS. In this study, as understanding the cause and the symptom of a problem enables us to find an efficient solution to the problem, we classified errors into multiple types according to the cause and the pattern (or structure) of the errors: missing data, practical error and logical error.

3.1.1 Type I: Missing data

An error in Type I is resulted from missing data in an observation. There are several reasons why the data is missing in HTS. First, respondents tend to avoid answering to somewhat delicate questionnaire due to a personnel reason. For example, how much do you earn money? Second, they easily forget to report complicate or obscure questionnaire. In HTS, reporting the time and location information of all the trips executed seems much burdensome for respondents.

In statistics, missing data are typically classified into three categories according to the mechanism of the missing data (Little and Rubin, 1987): MCAR (missing completely at random), MAR (missing at random) and MNAR (missing not at random). In detail, MCAR implicates that there is no relationship between missingness of the data and any values of either the observed of missing. On the other hand, MAR may depend on the other observed data, but not the missing data. MNAR is related to the missing value itself. Since each of these categories has a different strategy to deal with the missing data, it is necessary to identify the pattern of the missing data. In this study, we used a function of *Missing Value Analysis* in SPSS to figure out the pattern of the missing data in each.

Table 1. Univariate Statistics (only household category)

				missing	
	presents	mean	std. deviation	n	%
htruck	44,890	.10	.323	2,219	4.7
htaxi	44,890	.01	.117	2,219	4.7
hcycle	44,890	.04	.200	2,219	4.7
hother	39,911			7,198	15.3
hincome	46,805			304	.6

* variables with less than 0% missing are not displayed.

As can be seen in Table 1, *hother* (whether or not possessing another house) shows the majority of the missing data in the household category, and *htruck* (number of a truck in household), *htaxi* (number of a taxi) and *hcycle* (the number of a cycle) are followed by. *hincome* (household income) also has a small number of missing cases. In the person category, *pemployee* (type of employee; 19.9%), *pstyle* (working type; 18.3%) and *pwith* (whether or not living together with householder; 12.7%) account the biggest proportion of the missing data. This is because these attributes are somewhat sensitive to respondents' privacy. In the trip category, most of the missing cases are involved in *card* (whether or not using card for a trip fare; 14.1%), which is followed by *fare* (whether or not pay for a trip fare; 4.9%), *park* (whether or not parking; 4.8%) and *ori* (origin location type; 4.4%). The missing data in this category might come from respondents' burden on reporting such a complicate questionnaire in HTS. Separate-variance t-tests was then examined to identify variables whose patterns of missing data may be influencing the other variables (see Table 2). Note that we only experimented for the household category here because a quantitative attribute with missing data is included in this category.

Table 2. Separate Variance *t* Tests (Household class)

		hnum	*hkidnum*	*hcar*
htruck (*htaxi* & *hcycle*)	t	3.3	1.6	-1.3
	df	2,465.8	2,433.3	2,448.4
	# present	44,890	4,4890	44,890
	# missing	2,219	2,219	2,219
	mean(present)	3.43	.25	.77
	mean(missing)	3.35	.23	.79
hother	t	12.6	-5.5	-4.2
	df	9,829.9	9,498.7	9,973.1
	# present	39,911	39,911	39,911
	# missing	7,198	7,198	7,198
	mean(present)	3.45	.25	.77
	mean(missing)	3.26	.29	.80
hincome	t	3.4	5.2	4.1

	df	306.7	310.1	307.3
	# present	46,805	46,805	46,805
	# missing	304	304	304
	mean(present)	3.42	.25	.77
	mean(missing)	3.19	.13	.63

* variables with less than 0.1% missing are not displayed.

Table 2 shows that smaller households are likely to report the attributes, *hother* and *hincome*. When *hother* is missing, the mean *hnum* (household size) 3.26 compared to 3.45 when present (3.35 versus 3.43 for *htruck*, *htaxi* and *hcycle*; 3.19 versus 3.42 for *hincome*). The result implies that the data may not be categorized as MCAR because the missingness of above attributes seems to influence the mean of several other variables.

In addition, tabulated patterns was then tested to determine whether the data are jointly missing or not. In consequence, *hcycle*, *htruck* and *htaxi* (in the household category), *pemployee*, *pstyle* and *pwith* (in the person category), and *card*, *fare* and *park* (in the trip category) are missing together more than other pairs of the missing data. This is not surprising because those pairs of the missing data indicate similar information, such as *hcycle*, *htruck* and *htaxi,* about all describe a vehicle type, and *pemployee and pstyle* on job status (except for *pwith*). This might be the reason of such a delicate attribute in HTS. It implies that respondents are not likely to report a pair of similar attributes at one time whatsoever. At last, Little’s MCAR test (Little, 1988) was conducted to examine whether the missing data are MCAR or not.

Table 3. Results of MCAR test (Household category)

hnum	*hkidnum*	*hcar*	*hvan*	*Htruck*	*htaxi*	*hcycle*	*hetcveh*
3.45	.25	.77	.08	.09	.01	.04	.02

* Little's MCAR test: Chi-Square = 1531.433, DF = 5, Sig. = .000

The null hypothesis for Little’s MCAR test is that the data are MCAR. As you can see in Table 3, since the significance value (Sig.=0.000) is less than 0.05, the null hypothesis is rejected in this case, which means that the missing data in the household category are not MCAR. Two other categories also reject the null hypothesis. Therefore, we confirmed that the data are not missing completely at random. which means that it is unfeasible to apply *listwise* delete or *single imputation* for the missing values in HTS. The alternative methods for the missing data will be discussed in the following subsection 4.2.

3.1.2 Type II: Practical error

A practical error, Type II, is resulted from that respondents answer to a questionnaire, but the answer has an incorrect form or other than the given categories of an attribute in HTS. We found an amount of problematic cases in the trip category, especially trip location, which mostly recorded zero or other than the given categories in the attributes. For example, ones report a value of 3 in the attribute of *ori* (origin type) that has only two categories; 1 for home and 2 for other place. Moreover, the others might misunderstand the range of the given time of day in HTS, such as *ohour* (hour of trip departure time) and *dhour* (hour of trip arrival time). While they were asked to report activities and trips executed within 24 hours in a given day, they reported activities or trips executed even in the following day. Such a type of errors can result in system error when implementing ABM, because the system cannot easily operate incorrect data. Hence, the practical error in Type II should be revised prior to implementing the system. The solution to this problem will be proposed in the next subsection.

3.1.3 Type III: Logical error

A logical error in Type III arises when people respond to a questionnaire in a correct form but it is typically incomplete or inconsistent data in spatial and/or temporal dimensions. For instance, two consequent times of either activities or trips reported by respondents are overlapped. Another example with respect to a spatial dimension is that one reports going home as the purpose of the last trip, but the destination of that trip is not matched to his home location. An error in Type III only occurs in the trip category because the trip category only deals with spatial and temporal facets in the attributes. Type III was then classified into Type III-1~4 according to a specific case as followed:

- $T^E_i > T^B_{i+1}$ $(i < i+1)$: travel time overlapped (Type III-1)
- $T^E_i = T^B_{i+1}$: no gap between two consequent trips (Type III-2)
- $T^D_i \neq T^O_{i+1}$: a trip location inconsistent (Type III-3)
- $T^D_i \neq HL$ (only if a trip i is back home): a home location mismatched (Type III-4)
- $T^O_1 \neq HL$ and/or $T^D_n \neq HL$: non-residential schedule (Type III-5)

Where T^B_i, T^E_i are the departure time and the arrival time of trip i $(1 \leq i \leq n)$. T^O, T^D are a trip origin and destination location, and HL is a home location. For Type III-1, an error is generally emerged when a respondent forget updating an activity (or trip) time in the diary. Type III-2 normally has a similar reason with Type III-1, but it may not be illogical in the sense that the following activity needs no time to be executed. For example, a tour (or bring/get activity) is executed through a trip itself. Another possible reason for this error is that people tend to overlook reporting a rather short-term activity or trip. Errors in Type III-3 might be resulted from that a respondent forgets adding an activity that is suddenly emerged in the middle of two existing consecutive activities. Otherwise, the trip purpose is to transfer from one trip mode to another. In that case, there is normally no time gap between two consecutive trips. A Type III-4 error occurs by respondents' mistake or again forget inserting emerged activity, which happened just before going home, in the schedule. A Type III-5 causes a system error in FEATHERS due to the fact that it assumes a home-based schedule that begins with a home-originated trip and finish end with a back-home trip. Both the origin and destination location of the first and last trips should be matched with a home location. This type of error does not make any problem with trip-based transportation models, but it can be crucial in FEATHERS because it is difficult to deriving activity information from inconsistent or incomplete trip data. The error can also effect on the system performance. Hence, the error in Type III should be processed to prevent from an unexpected error or a poor performance resulted in the system.

3.2 Data processing results

3.2.1 Type I

Related studies have proposed several methods for treating missing data (here, Type I). Conventional methods are listwise and pairwise deletion. Listwise omits the case with missing values in any of the analysis variables, but on the other hand, pairwise uses the case with non-missing values for pairs of variables. Since both methods assume that the pattern of the missing data is MCAR, it works fine in the case of MCAR. Otherwise, it can introduce biased estimation into the missing data (Allison, 2001). For MAR (in other words, non-MCAR), EM (expectation maximization) can be used to deal with missing data. EM is a numerical algorithm that repeatedly cycles through two steps: expectation and maximization. These two steps are iterated until the estimation do not change from one iteration to the next (Dempster *et al.*, 1977). Suppose a set of the observed data X consisting of observed O and missing M. An unknown parameters θ can be estimated by maximizing the observed data log-likelihood as follows:

$$\hat{\theta} = \arg\max p(\theta|O) \qquad (1)$$

$$Q(\theta|\theta^t) = E(\log p(O, M|\theta)|O, \theta^t) \qquad (2)$$

$$\theta^{t+1} = \text{argmax}\, Q(\theta|\theta^t) \qquad (3)$$

where θ^t is the current estimate of the parameter. For the expectation step, the conditional expectation is calculated by (2). The maximization step then updates the estimate with the maximized parameter computed by (3). These steps are repeated till no improve in θ.

As mentioned in the previous sub-subsection (see 4.1.2), since the pattern of the missing data in HTS is MAR, we applied the EM estimation to process the missing data. For that, we used *Multiple Imputation* using EM in SPSS to generate a possible value for the missing data. Note that *Multiple Imputation* generally shows better performance than *single imputation* for solving the problem of missing values (Buuren, 2007).

Table 4. Descriptive Statistics for *hcycle* (Quantitative)

data	n	mean	std. deviation	minimum	maximum
original data	44,890	.04	.200	.00	5.00
imputed values	2,219	.17	.125	.00	.78
complete data after imputation	47,109	.04	.200	.00	5.00

Table 5. Descriptive Statistics for *pemployee* (Categorical)

data	category	n	%
original data	1	3,510	5.5
	2	36,620	57.4
	3	5,505	8.6
	4	18,172	28.5
	5	1	.0
	6	1	.0
	8	1	.0
imputed values	1	4,853	5.3
	2	52,216	57.4
	3	7,570	8.3
	4	26,282	28.9
	5	4	.0
	8	1	.0
complete data after imputation	1	8,363	5.4
	2	88,836	57.4
	3	13,075	8.4
	4	44,454	28.7
	5	5	.0
	6	1	.0
	8	2	.0

Table 4 and 5 describe statistics for the original data, imputed values, and complete dataset (combining the original data and imputed values) with imputed values. In Table 4, *hcycle* (number of cycle) shows that statistics for complete dataset are completely equal to that for the original data, meaning that the imputed values for the missing data are estimated by EM. In Table 5, *pemployee* (type of employee) also shows an enough identical pattern of the values between the original data and the complete data. Due to a limited space, above two variables are only dealt with in this paper. The rest of the variables in the missing data also shows satisfactory results.

3.2.2 Type II

To process Type II, the same method for Type I was used for modifying wrong-formed data with right formation in the attributes. In detail, incorrect values in a trip location was revised based on the other observation. For instance, if *ori* (origin type; 1 for home, 2 for other place) has a wrong value (i.e., 3), then it can be modified by comparing the origin location (T^D) of the trip with the home location (*HL*) of the person conducted the trip. If *ocode*=*hloc*, then *ori* is set to 1, otherwise 2. The rest of the errors in Type II are resulted from that respondents report other than the given categories in an attribute. For example, while two categories (1=yes, 2=no) are given in the attribute of *park* (whether or not park), 0 and 3 are recorded in the attribute. In this case, a value of 0 can be considered as missing data due to enough cases to apply for missing data method, whereas a value of 3 cannot do that because of the mere sample in this case.

Table 6. Result of Type II

class	case	before		after		method
		n	%	n	%	
household	*hincome*	123	0.3	0	0.0	missing data method for 0
person	*plicense*	2,281	1.4	0	0.0	missing data method for 0
	pemployee	64,039	40.1	3	0.0	missing data method for {0, 9}
	pstyle	64,039	40.1	48	0.0	missing data method for {0, 9}
	sex	111	0.0	0	0.0	missing data method for 0
trip	*ori*	11,245	3.7	0	0.0	*ori*={1 if $HL=T^D$, 2 otherwise}
	ohour	1,186	0.4	1,186	0.4	
	dhour	3,142	1.0	3,142	1.0	
	card	2,463	0.8	17	0.0	missing data method for 0
	park	18,514	6.1	4	0.0	missing data method for 0
	fare	18,588	6.1	11	0.0	missing data method for 0

Table 6 illustrates the result of data processing for Type II in HTS. Most of the problems in Type II was resolved by using the missing data method, which is the EM estimation, except *ohour* (hour of trip departure time) and *dhour* (hour of trip arrival time) in the trip category. This is because we have no choice but to omit those cases from the data sets. In the end, 97.6% (181,320 of 185,731) of the errors in Type II were successfully corrected in total.

3.2.3 Type III

An error in Type III has to be dealt with according to its symptom, which is already classified into Type III-1~4. The solutions to each symptom are as followed:

- Type III-1 (travel time overlapped): $T^B_{i+1} \Rightarrow T^E_i$
- Type III-2 (no time gap between two trips): $(T^B_{i+1} - \alpha) \Rightarrow T^E_i$ $(\alpha \geq 0)$
- Type III-3 (a trip location inconsistent): $T^O_{i+1} \Rightarrow T^D_i$
- Type III-4 (a home location mismatched): $HL \Rightarrow T^D_i$
- Type III-5 (non-residential schedule): add $T^O_{(1-1)}$ (= HL) and/or $T^D_{(n+1)}$ (= HL)

Type III-1 resulted from overlapping travel time was fixed by setting the departure time of the following trip to the arrival time of the trip *i*. Because there is a higher probability of shifting a trip arrival time than a trip departure time. In addition, respondents are more careful to report the departure time of trip than the arrival one because there is more uncertainty in a travel duration. Next, Type III-2 was resolved by shifting the trip arrival time by some amount of time dependent on the trip purpose. When the trip purpose is a trip itself (i.e. tour and bring/get) or trip mode transfer, then the arrival time is shifted by zero because it is unnecessary to insert some time to active after the trip. However, the trip purpose is something else than a trip, then a small amount of time, like 5 minutes, are inserted into the between consecutive trips. In the end, the system has no problem in this type of error and there is no impact on the result because of small change in travel time. Type III-3 is hardly treated by a somewhat big assumption that the departure location of the following trip is set to the origin location of the current trip because this solution might really ignore the probability of inserting new activity emerged between two consecutive activities. Type III-4 was revised by replacing the destination location of a back-home trip with the residence location belonging to the person. Lastly, Type III-5 was processed by manually adding a home-originated trip prior to the first schedule and/or a back-home trip followed by the last one in the non-residential schedule.

Table 7. Result of Type III

case	before		after	
	n	%	n	%
no time gap (Type III-2)	3,187	1.99	1,862	1.16
	- 1,630 (transfer)	1.0	- 1,630	1.0
	- 232 (tour)	.14	- 232	0.14

	- 1,352 (others)	.84	- 0	0.0
non-residential schedule (Type III-5)	991	0.62	0	0.0

Table 7 shows a number of errors in Type III. Before data processing, we found errors of Type III-2 (3,187, 1.99%) and of Type III-5 (991, 0.62%) from the data. As you can see in Table 7, the errors (1,352, 0.84%) in Type III was almost eliminated from the data by the data processing, except for the cases of transferring trip (1,630, 1.0%) and tour (232, 0.14%) typically considered as an activity itself. Hence, it is unnecessarily to fixed the errors from those two kinds of trips. As a result, the result confirms that the data processing works fine for Type III. Note that there is no error of Type III-1, 3 and 4 in this case because those types of errors in the survey was already omitted from the sample, fortunately. Nevertheless, this solution can be still helpful to solve a similar problem in other survey.

4. CONCLUSION

In this paper, a household travel survey in Korea was overviewed at first, and FEATHERS were then explained with its system overview and data requirement. Next, we classified errors into multiple types (Type I, Type II and Type III) according to its symptom and cause in HTS. According to the classification, a number of cases were identified and a reason for those errors was also figured out in each type of errors. Data processing methods were then used to treat the types of errors in HTS. The results from the experiment show that the methods explicitly contribute to improving a quality of data considering the reduced amount of errors in data processing. It is worthwhile to deal with how to treat errors in data and what strategy can be applied for reviving the problematic data by considering the condition of the problem. Based on the findings in this study, we will propose an efficient framework for HTS that enables us to minimize errors occurred by respondents. In addition, we will also develop the data processing module in FEATHERS in the future.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the ion of Korea government (MSIP) (NRF-2010-0029444).

REFERENCES

Allison, P. (2001) Missing Data. Thousand Oaks, Sage Publications, CA.

Arentze, T.A., Hofman, F., Kalfs, P.T.A.M. and Timmermans, H.J.P. (1999). System for logical verification and inference of activity (SYLVIA) diaries. *Transportation Research Record*, (1660), pp. 156-163.

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T. and Timmermans, H.J.P. (2010) Implementation framework and development trajectory of FEATHERS activity-based simulation platform. *Transportation Research Record*, 2175, pp.111-119.

Buuren, S.V. (2007) Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, 16(3), pp. 219–42.

Clarke, M., M. Dix, and Jones, P. (1981) Error and Uncertainty in Travel Surveys. *Transportation*, 10, pp. 105–126.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B,* **39**(1), pp. 1–38.

Little, R.J.A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association,* 83(404), pp.1198-1202.

Little, R.J.A. and Rubin, D.B. (1987) Statistical analysis with missing data. Wiley, New York.

MTA (2012) Passenger OD Synthesis and Future Demand Forecasting, Seoul.

Přibyl, O., Micaelli, J.R., Goulias, K.G. and Patten, M.L. (2004) CHIRAC: A Comprehensive Household Integrated Rectifier for ACtivity diaries. CentreSIM3 Report submitted to McCormick Taylor Associates and the Mid-Atlantic Universities Transportation Center, April 2004, University Park, PA. pp. 38.