#### DISAGGREGATE MODELLING BY MULTI-SOURCE DATA

T.  $Yai<sup>1</sup>$ , S. Morichi<sup>2</sup>), K. Fan<sup>3)</sup>

1) T. Yai 2) S. Morichi 3) K. Fan 3) K. Fan Tokyo Institute of Technology Tokyo Institute of Technology Chuo University 2-12-1 Ookayama, Tokyo, Japan 2-12-1 Ookayama, Tokyo, Japan 3-12-1 Assuga, Tokyo, Japan

## 1. INTRODUCTION

Nearly a couple of decades have passed since disaggregate modelling were developed. They have been widely applied in many regions also in Japan (1). The models have a distinct advantage that they can be calibrated by few observations for predictions into the future. For this reason, various transport behaviours have been modeled recently. It is obvious that the model parameters will be less stable as samples decrease. Although many other important results have been derived in the earlier discussions on this topic, the sufficient size of observations required to estimate the model has not been proposed yet

On the other hand, the statistical aggregate data, which is relatively more reliable as an amount than disaggregate small data, like traffic road-side counts or aggregated results of large scale interview surveys, are often collected on each transportation mode. These data are generally available for transportation planning studies, though it is difficult to estimate the model solely with these data. Utilizing disaggregate and aggregate data to calibrate the model in such cases, we are able to expand applications of the models. This is because combining disaggregatc data, which is small but fine, with aggregate data, which is large but accurate, improves the predictability.

The purpose of this paper is to provide a new method to estimate disaggregate models with the foregoing concept and to examine the applicabilities for inter and infra city passenger demand. In Section 2, we briefly review the methods proposed in previous studies. Section 3 presents our new method and we apply them to intercity travel demand analysis in Section 4. Section 5 presents other applications for passengers in a recreational area and in a city.

#### 2. METHODS TO COMBINE DIFFERENT DATA

At the end of 1970s, estimation techniques for choice based samples were developed by Lerman & Manski (2) and others. This is a kind of data combination methods because aggregated shares from different sources are also utilized in estimating disaggregate models.

A lot of studies to estimate OD matrices with link flow data have been carried out by Fisk & Boyce (3), Maher (4), Ben-Akiva (5) and others. Maher developed a new technique to estimate them using Bayesian updating method. Supposing OD trips as parameters, he utilized link flow data to

compose likelihood functions and to estimate OD matrices. Ben-Akiva proposed a general framework for methods to combine different data sources. Applications for several combinations were demonstrated in the study.

Then, Quevedo (6) provided combined likelihood functions to calibrate a disaggregate continuous choice model. A simultaneous maximization of disaggregate and aggregate likelihood functions was conducted in the paper. Morichi & Yai (7) updated constant terms of disaggregate models additionally with trip attraction volumes in estimating OD tables.

Besides, few studies were conducted for updating disaggregate models with aggregate data, though there are many researches on this topic. Since a theory of disaggregate model is based on the random utility and the individual behaviour, utilizing aggregated information in the estimation of the model may be contradicting the theory. We attempt to address this issue using Bayesian approach after Maher, Atherton & Ben-Akiva (8), and others.

## 3. ESTIMATION OF DISAGGREGATE MODEL USING ADDITIONAL AGGREGATE DATA

Aggregated data is the summation of individual choices. Thus, the estimated data from individual models should be close to the observed. Hence, we updated the model by modifying parameters considering the precision of aggregated data. We employed Bayesian estimation of parameters. A prior distribution of parameters and additional posterior information associated with choice results are necessary for the completion. Obtaining both information, we can provide a new vector of parameters. Any data as long as it corresponds to aggregated volumes from disaggregate models, in practice, can be utilized in the combine method.

In this study, a prior distribution was assumed to be that of discrete model parameters which were calibrated singly with disaggregate data. The updated parameters can be calculated with the distribution and a likelihood function composed of additional aggregate data. The prior distribution was normal, and the later was supposed to be normal. A maximum point of the posterior distribution will be an optimal solution.

Morichi et.al. (9) have already proposed this technique and investigated principal properties of the estimator in intracity passenger demand. They developed another type of estimation, which is an approximate technique equivalent to the generalized least square method and is also utilized in setting the initial values for the exact estimation, though it is not addressed here.

# 3.1 Parameter estimation of disaggregate model

Maximum likelihood method is usually employed for estimating disaggregate models. That is expressed as follows,

$$
\max L(\theta) = \sum_{n} \sum_{i \in C_n} \delta_{i_n} \ln(P_{i_n}). \tag{1}
$$

In the above,  $P_{in}$  represents a choice probability of individual n for alternative *i*,  $C_n$  is choice set of *n.*  $\delta_{in}$ takes 1 only if individual *n* chose alternative i and takes *zero* elsewhere. Logit formulation was adopted as disaggregate model, that is,

$$
P_{in} = \frac{e^{V_{in}}}{\sum_{r \subset c_n} e^{V_{in}}} \tag{2}
$$

where  $V_{i_n}$  is a strict utility of n for i and is usually described as,

$$
V_{i\sigma} = \sum_{k} \Theta_k X_{i\sigma k} \tag{3}
$$

An linear function on parameter  $\theta_k$  and attribute  $X_{ink}$  are usually assumed. Maximization of  $L(\theta)$  leads the estimator  $\theta$  which is consistent and distributed normal in statistical large samples,

$$
\Theta \sim MVN[ \theta, \Sigma ] \qquad \Sigma = (-\nabla^2 L(\theta))^{-1} . \qquad (4)
$$

The above equation is rewritten as follows,

$$
P(\theta) = \frac{1}{(2\pi)^{K/2} |\Sigma|^{1/2}} \exp[-\frac{1}{2} (\theta - \hat{\theta})^t \Sigma^{-1} (\theta - \hat{\theta})].
$$
 (5)

### 3.2 Updating of Disaggregate Model Parameters using Bayesian Approach

The parameter solely from disaggregate data is described as  $\theta^d$  here, and the covariance matrix is also written by  $\Sigma_d$ . Additional data Q is an amount of individuals and should be aggregated with disaggregate models. For instance, Q are passenger counts on board survey or tourists passed entrances of destination facilities. The observed  $Q$  is supposed to have random variances  $\Sigma$  caused by observation and aggregation errors. Then, aggregate volume  $Q$  is expressed as,

$$
Q = Q(\theta) + \epsilon \tag{6}
$$

In eq.(6), Q is assumed to be normally distributed because observation and sampling errors generally randomly distributed.

A probability density function of the error term is described,

$$
P(Q|\theta) = \frac{1}{(2\pi)^{K/2} |\Sigma|^{1/2}} \exp\{-\frac{1}{2}(Q-Q(\theta))'\Sigma^{-1}(Q-Q(\theta))\} \ . \tag{7}
$$

Q is  $J \times I$  and  $\Sigma$  is  $J \times J$  matrix. Eq.(7) represents a likelihood function concerned with  $\theta$ .

Now, we describe  $Q^{\circ}$  and  $\Sigma_{\circ}$  as observed values. Bayesian procedure gives a posterior distribution of parameters as follows,

$$
P(\theta|Q^o) \propto P(\theta) P(Q^o|\theta).
$$
 (8)

For obtained  $P(Q<sup>o</sup>|\theta)$ , maximizing the above function yields an optimal solution of parameters. That is an

updated Bayesian estimator. Using (5) and (7),

$$
P(\theta|Q^o) \propto exp[-\frac{1}{2}(\theta - \theta^d)^t \Sigma_d^{-1}(\theta - \theta^d) - \frac{1}{2}(Q^o - Q(\theta))^t \Sigma_o^{-1}(Q^o - Q(\theta))]
$$
\n(9)

can be derived directly. Maximization of (9) is equivalent to minimization of the following equation,

$$
\min_{\theta} \left[ (\theta - \theta^d)' \Sigma_d^{-1} (\theta - \theta^d) + (Q^{\circ} - Q(\theta))' \Sigma_o^{-1} (Q^{\circ} - Q(\theta)) \right]. \tag{10}
$$

Statistical modes of the distribution is equal to parameters at the minimum point of the above function. An iteration technique such as Newton-Raphson method is adopted to solve it. The above distribution is not symmetric associated with  $\theta$  because  $Q(\theta)$  is a non-linear function on  $\theta$ .

# 3.3 Aggregation in the Model Estimation

The above method is available for the popular aggregation methods of disaggregate models such as Enumeration, Classification and Naive. The only condition to apply aggregate data here, is that it must correspond to the aggregated data from the models. For example, in the mode choice model, we can apply the total passenger trips by each mode to the aggregated results of individual choices. For Enumeration,

$$
Q_i(\theta) = \frac{T}{N} \sum_{n=1}^{N} P_{i,n}
$$
 (11)

where  $N$  is total samples and  $T$  is expressed as,

$$
T = \sum_{i} Q_i^o \tag{12}
$$

for Classification,

$$
Q_i(\theta) = T \Sigma w_{\mathbf{r}} P_{i\mathbf{r}} \tag{13}
$$

Using these equations, we can estimate the model parameters which reflect the introduced aggregation method.

### 4. APPLICATION FOR INTERCITY TRAVEL DEMAND

In this Section, we examined the applicability of Bayesian method in the prediction of intercity travel demand. Although there are a few intercity data, there exist some difficulties in analysing the passenger demand solely with those data. So, we adopted another individual data from Intercity Person Trip (IPT) Survey, which was conducted in some cities by JTERC (10), to calibrate disaggregate transportation models. Available statistics and surveys, Interregional Passenger Transportation (IPT) Statistics, Air passenger (AP) Survey and Road traffic (RT) Survey, were utilized as aggregated data in reestimating disaggregate model parameters. The difference of regional model structures will be discussed here.

# 4.1 Individual Data and Calibration of the Models

Pre-studies of UT-Survey were conducted in Tokyo and some other local cities in 1982. We employed the data of Tokyo, where more than 2000 households were sampled and almost a thousand recreational trips are available for modelling, to build disaggregate models. Every data contains several trip attributes of travels beyond one-day-trip areas.

We divided individual data into two segments by the trip length as competition between transport modes differs in regions. Using segmented data, we estimated the following two models; Model-1 is for a segment whose trips are more than 300km long and Model-2 is for less than 600 km long trips. The reason that trips between 300km and 600km are overwrapped in both is chiefly because size of sample is inadequate to estimate model-parameters. Model-2 may be so unstable in the prediction. However, the purpose of this section is to examine the feasibility of our data combination method, so the results that models with small samples are improved by additional aggregate data are also necessary to be analyzed. These models have three transport modes ( Rail, Air, and Car ) and contain four variables and two constant terms.

Table-1 shows the parameters, which are used as the original in the later updating process. Time values of them can be obtained by dividing parameters of line-haul time and terminal time by those of total travel cost. The results appeared quite different; the time value of line-haul in Model-1 is 1820 yen/h and that in Model-2 is 4080 yen/h and the values of terminal in them are 1890 yen/h and 1020 yen/h. These differences will be discussed after the revision of models in 4.4.

Variable	model 1 (T-statistics)	model 2 (T-statistics)
Line-haul travel time(min.)	$-0.01439$	$-0.01647$
	(6.07)	(6.67)
Terminal travel time(min.)	$-0.01501$	$-0.004106$
	(7.09)	(2.06)
Total travel cost(yen)	$-0.0004757$	$-0.0002424$
	(6.20)	(2.51)
Car ownership(1:own,0:otherwise)	1.175	0.8418
	(4.88)	(1.75)
Constant(Car)	1.643	3.558
	(2.26)	(4.91)
Constant(Rail)	0.8853	2.045
	(1.20)	(2.74)
Likelihood ratio	0.38	0.50
Num. of observations	576	168

Table-1 Estimated parameters of mode choice model (Car, Rail, Air) for recreational trips

#### 4.2 Aggregate data in model updating

# 4.2.1 Statistics for intercity passengers

Although various passenger-statistics are collected periodically in Japan (11), some of them have lack of accuracy, especially in OD flow data in intercity travel. The reason is that the statistics are reprocessed from those of each transportation industry and have some aggregation errors. In addition, sampling surveys such as Air Passenger Survey (every 2 years) or Road Traffic Census (every 3 years) are available for transportation demand studies though differences in their time periods and types of questionaries provide the difficulty in mode choice modelling by choice-based estimation. If possible to combine the above statistics in demand forecasting, planning of intercity transports will be more rational.

#### 4.2.2 Aggregate data for model updating

Annual OD tables of total passengers, named "Interregional Passenger Transportation Statistics", mentioned above were adopted as aggregate data for model updating because they are the unique statistics in Japan, whose unit of origin and destination is a prefecture. The poor accuracy of the data comes to be clear by comparing their time series variations. Figure-1 shows an example of changes of trips from Tokyo by private car from 1978 to 1986. It seems that tendencies of volume changes have errors of some kinds though various changes in social and economic conditions arise every year. Another problem on this statistics is that some OD pairs have zero volumes as destinations of air passengers are fixed on the prefectures in which airports are located and no passenger drive to far destinations.



The passenger data in 1982 have to be estimated approximately, as disaggregate models were built with individual data in the year. So, we tried to lighten these problems using other data sources. For car nips, since time series data are gathered in every OD pair, it is possible to calculate average shares using the pooling data. So, we used the shares to assign OD volumes again under the constraint that total in the year is fixed. In addition, shares of destinations by RTC in 1985, which were limited in highway

users, were also utilized for re-assignments of total volumes for zero-OD in some regions. On the other hands, for air passengers, actual destinations from the above survey were adopted to re-estimate OD volumes.

Further, we reduced the data to recreational trips using the ratios which are calculated as 47% for rail, 69% for private car, and 36% for air, in the IPT-Survey because of the necessity to focus on the behaviours. One-way trips from Tokyo after these processes were presented at Table-2. Some OD still remain at zero because the above additional data could not change them. Prefectures are classified into 8 groups, which represent directions from Tokyo and partially differ from ordinary regions in Japan.

Destination	(One way uips pu year)			
	Car	Rail	Air	
<b>Tohoku1</b>				
Aomori	64365	314403	80765	
Iwate	103410	445235	16146	
Akita	90198	247899	92579	
Yamagata	193238	393962	43640	
Tohoku2				
Miyagi	211363	896619	71063	
Hukushima	500834	1100103	0	
Tochigi	3610877	3920744	0	
Hokuriku				
Nagano	3266474	1480774	$\bf{0}$	
Nigata	286022	1244747	33043	
Ishikawa	28872	116169	161249	
Toyama	4892	173402	26564	
Tokai				
Shizuoka	6613291	2850156	0	
Aichi	562374	2256257	$\bf{0}$	
Gifu	213592	176571	0	
Kinki				
Osaka	388788	2793822	366216	
Hyogo	359156	448830	96021	
Kyoto	137519	1090785	8026	
Nara	70380	8021	7452	
Wakayama	49699	37840	7697	

Table-2 Examples of recreational passenger demand from Tokyo in 1982 (One way trips per year)

#### 4.3 Model updating for intercity travel

# 4.3.1 The accuracy of aggregate data

It is necessary to evaluate the accuracy of aggregate data when we utilize our re-estimation method. They were defined by the following co-variance matrices  $V_{ij}$  to simplify the analyses.

$$
V_{ij} = Q_i Q_j \alpha^2 \quad (i=j)
$$
\n
$$
=0 \qquad (i \neq j)
$$
\n(14)

where  $\alpha$ , which corresponds to the coefficient of variation of aggregate data, is a scale value to determine the accuracy.

We fixed them as follows; the value of air is 0.3, that of rail is 0.5 and that of car is 0.7. It is impossible to estimate the precise values from applicable data because of strict difficulties to divide errors into every effect. Although it may be less accurate than actual, under-estimation of the reliability does not increase undesirable effects on revised parameters. Obviously the accuracy of air is best and that of car is worst, so we decided to adopt above values as the accuracies. The distributions of the errors of aggregate data is assumed as normal and as mutually independent each other. Thus, our method is directly available because these assumptions provide no necessity to consider correlations of the errors between transport modes.

# 4.3.2 Parameter updating of disaggregate model

As mentioned above, destination areas from Tokyo were divided into 8 segments ( we call them "regions"): Tohokul, Tohoku2 (nearer than Tohokul ), Hokuriku, Tokai, Kinki, Chugoku, Shikoku, and Kyushu. Every region contains a few prefectures, for example, Tohokul is composed of 4 prefectures: Aomori, Iwate, Akita and Yamagata, shown in Table-2.

Updating results were described in Table-3, where is a list of updated parameters for each region. We defined the average difference between parameters as follows,

$$
\Delta \theta = \frac{100}{K} \sum_{k=1}^{K} \frac{|\theta_k^2 - \theta_k|}{\theta_k^2} \qquad (%) \tag{15}
$$

where *K* is the number of parameters,  $\theta^d$  is parameters from original disaggregate data and  $\theta^r$  is revised parameters using additional aggregate data. The difference in every region is presented in Figure-2. The parameters in Hokuriku from Model-1 and those in Kinki from Model-2 were much different than those of the original, Shikoku and Kyushu did not change notably, and the change of Tokai were slight. Figure-3 shows differences of aggregated volumes between observed aggregate data modified in 4.2.2 and calculated data from models. The difference was defined as follows,

$$
\Delta Q(\theta) = \frac{100}{J} \sum_{j=1}^{J} \frac{|Q_j^{\alpha} - Q_j(\theta)|}{Q_j^{\alpha}} \qquad (%) \tag{16}
$$



#### Table-3 Revised parameters for each region





Figure-2 Differences of revised parameters from the original

where *J* is the number of aggregated data. In every region, the difference with the revised model come to be smaller than that of pre-revised (original disaggregate) model.





#### 4.4 Time evaluation of passengers from revised models

Figure-4 shows time evaluation values from the revised models. For every region, parameters changed and time values moved. The values may indicate the property of regional transport conditions. The original values of line-haul from disaggregate models were 4080 yen/h ( for longer distance) and 1820 yen/h ( for shorter distance ). Now, the Figure seems to show the following regional differences; the values for high level-of-service areas (Kinki, Chugoku and kyushu) are relatively small, where Tokaido-Sanyo Shinkansen, Tomei-Meshin-Chugoku-Kyushu highway, and airlines are available, and the values in Hokuriku and Shikoku are markedly high, where the service level is totally lower and that of each mode is quite different. In conclusion, it seems that the original value will be changed adequately by regional aggregate data.

Comparison of other parameters are also illustrated at Figure-5. Time evaluation values of terminal travel time which contains access time to departure terminal and egress time from arrival terminal were presented here. The differences between regions partially appear same as the above.



In this section, applying our new method which combines aggregate data to disaggregate models for intercity recreational travel demand, we conclude that it is possible to change original parameters solely with small disaggregate data into regional convenient values using additional data. It will be useful for intercity transport plannings because unadequate data could not provide required information for them until now.

# 5. OTHER APPLICATIONS

### 5.1 Application for traffic volumes in the recreational area

Difficulties to estimate traffic flows in a recreational area still remained because of lack of data and methods, though many studies have dealt with this issue. They are constraints to make a transportation planning into and in the area. Building a system for demand forecasting and explaining exact present

flow patterns are recently requested. Although methods to collect data have to be considered first, the system was tried to developed here and the Bayesian method was also partially applied. The chief purpose of this Section is to apply the method to route choice models as the system is tentatively composed.

## 5.1.1 Data and model system

A system was composed of three disaggregate models, two aggregate models and a few equations to illustrate present traffic flows in the area. As data requested in the system are individual interview data, traffic counts at road side points, and trip attractions in recreational facilities, surveys of two kinds were conducted in the area near Tokyo. We also utilize the data from the survey of reading numberplates conducted with traffic counts to examine the applicability of model updating. The models were revised using the former data after route choice models for car users were proposed with the later samples. Nearly 700 samples were collected from the later survey in 1988, and traffic counts and platereading were conducted at 7 points in the area in the same day.

### 5.1.2 Disaggregate models and the revised results

We calibrated disaggregate models of three kinds. Model-1 represents an entrance route into the area, Model-2 is for a route choice in the area to the first destination, and Model-3 represents an exit route from the area. These were modelled using disaggregate data, explanatory variables of which are limited in the primary attributes.

As it is difficult to estimate adequate volumes of each route traffics solely with disaggregate small samples. Hence, we attempted to improve the representation of flow patterns using additional traffic data. An example of which results are revisions of an entrance route choice model is presented in Figure-6 and 7, where changes of the parameters and those of aggregated volumes are shown. The applicability of our method was briefly tested and was supported through the analysis.

# 5.2 Applications for intracity passengers

### 5.2.1 Applied data

An example for intracity travel demand is addressed here. The combination usage of Person Trip (PT) Survey and National Census in the city of Sapporo were demonstrated for examination of the applicability. PT survey is a popular statistic for urban transport planning in Japan. Cities where more than 30 thousand people live are selected to be surveyed every decade. However, less population cities than the above have few data to utilize in the planning except Road Traffic Census. So, we also considered to apply our method there with additional disaggregate small data, as National Census is available everywhere.

### 5.2.2 An applied model and revised results



Figure-7 Traffic volumes through entrance road to the area

A mode choice model was calibrated with about two thousand observations, which contain 6 transport modes: rail, bus, car, bike, bicycle and walk, and 9 variables: travel time, travel cost, access time to a station, egress time from a station, and five constant terms. Then, the model was revised using 18 OD pairs data from National Census, in which residences and work places of persons were classified in wards. An accuracy of the data is supposed by the ratio to work in this case. An example of revised results were presented at Figure-8, in which the model was updated so as to represent the given OD pattern. As a result, the possibility that the model was varied to reflect the observed data and was improved in predicting future demand was demonstrated. Collecting small samples, we easily apply the method in small local cities, where the above survey is not conducted. The method can also be utilized in some stages of conventional four-step procedures for the prediction of future and/or present passenger demand.

#### 6. CONCLUSION

This paper provided a new method to obtain revised parameters of the disaggregate model using additional aggregate data. We employed Bayesian procedures, in which revised parameters are determined by differences of accuracies between the model parameters and the additional data. Applications to three travel choice situations in Japan are carried out to examine the practicability. Results are



Figure-8 Differences of aggregated volumes between observed and calculated

summarized as follows,

- a) A new method was developed and its properties were investigated through the following applications, and its applicability is proved to be high.
- b) Inadequateness of the data in estimating intercity passenger demand was improved by the present method. Disaggregate models were revised to reflect the differences of regional transport conditions.
- c) Other applications showed the availability of the method to practical planning issues.

Although we did not mention about applications in the estimation of OD matrices, the method can be applied easily to estimate them in other combinations of data. We have already expanded our method to estimate finer OD matrices than the observed, ising disaggregate destination choice model, and roughly divided but accurate distribution or/and attraction data of trips. Moreover, assumptions of the distribution of aggregate data are simplified in this study, though actual errors are not normally distributed and slightly depend on the variations of individual choices. Thus, more complicated expansions on the distribution will be required in further studies.

### REFERENCES

- 1. Yai, T.,(1989), Disaggregate Behavioural Models and Their Applications in Japan, Transportation Research, Vol.23A, No.l, pp.45-52.
- 2. Lerman, S. & Manski, C.,(1979), sample Design for Discrete Choice Analysis of Travel Behavior, Transportation Research, Vol.13A, pp.29-44.
- 3. Fisk, C.S. & Boyce, D. E.,(1983), A Note on Trip Matrix Estimation from Link Traffic Count Data, Transportation Research B, Vol.17, No.3, pp.245-250.
- 4. Maher, M. J.,(1983), Inferences on Trip Matrices from Observations on Link Volumes: a Bayesian Statistical Approach, Transportation Research B, Vol.17, No.6, pp.435-447.
- 5. Ben-Akiva, M.,(1988), Methods to Combine Different Data Sources and Estimate Origin-Destination Matrices. Proc. of the 10th International Symposium on Transportation and Traffic Theory, pp.459-

481, Elsevier.

- 6. Quevedo, S. L. G.,(1985), Combining Survey and Aggregate Data for Model Estimation, Ph.D. dissertation, MIT.
- 7. Morichi, S. & Yai, T.,(1986), Prediction of Trip Distribution by Disaggregate Behavioural Model,<br>Proc. of the World Conference on Transport Research, Vol.2, pp.1582-1599.
- 8. Atherton, T. & Ben-Akiva, M. E.,(1976), Transferability and Updating of Disaggregate Travel Demand Models, TRR610, pp.12-18.
- 9. Morichi, S., Yai, T. & Hirai, S.,(1987), Estimation of Disaggregate Model using Additional Aggre-<br>gate Data, Infrastructure Planning Review, JSCE, No.5, pp.51-58 (in Japanese).
- 10. JTERC,(1983), A Method of Intercity Person Trip Survey, Japan Transport Economics Research Center.
- 11. Nakamura, H. & Morichi, S.,(1989), Transportation Research Activities in Japan, Transportation Research, Vol.23A, No.1, pp.1-6.