International Journal of Engineering Research and Development
eISSN : 2278-067X, pISSN : 2278-800X
Volume 2, Issue 2 (July 2012), PP. 65-71

Spatial Transferability of Mode Choice Model

Md Shahid Mamun[1], Md Ahsan Sabbir[2]

[1]*Assistant Professor, Department of Civil Engineering, Ahsanullah University of Science and Technology, 141-142 Love Road, Tejgoan Industrial Area, Tejgoan, Dhaka-1208, Bangladesh.*

[2]*Senior Lecturer, Department of Civil Engineering, Stamford University Bangladesh, 51, Siddeswari Road, Dhaka-1217, Bangladesh.*

Abstract*—Many studies have been performed in the field of transferability of mode choice model. Both temporal and spatial transferability were covered in the previous studies although temporal transferability got main focus in this respect. Most of the research on spatial transferability was performed on identical data sets. This research examines the effect of omission of one of the most important variables (travel cost) in spatial transferability. Multinomial logit models are used for building mode choice models. Two sets of data from two different regions are used: one from Dallas-Fort Worth (DFW) and another from the San Francisco Bay Area (BATS). Each of the datasets is randomly divided into two samples: estimated dataset (90% of the whole data) and application dataset (10% of the sample). Model estimated from BATS's estimated dataset is applied to BATS's application dataset as well as DFW's estimated dataset. In the same way model estimated from DFW's estimated dataset is applied to DFW's application dataset as well as BATS's estimated dataset. No adjustment is done for transferring the models. The study finds that the parameters of explanatory variable of the models estimated from two regional datasets are similar in terms of sign and magnitude. Modal shares predicted by models are almost same as sample shares within the region and outside the region.

***Keywords*— Mode Choice, Discrete Model, Disaggregate Model, Multinomial Logit, Spatial Transferability,**

I. INTRODUCTION

Transferability is an important issue in travel demand modeling and forecasting. It is invoked when models are used to predict behavior in contexts that have different characteristics from those that existed in the data collection environment. Temporal transfer occurs when a model estimated in one time period in a specific geographic context is used in future forecasting in the same area. Spatial transfer involves applying a model estimated on data from one particular spatial entity to another geographic context. Without transferability in time, the model has no use in forecasting and without transferability between regions, models developed in one region cannot be applied elsewhere. The task of this research is to examine the spatial transferability of mode choice models when one or more important variables (travel cost, travel time) are missing in the data set. For this purpose, two sets of data from two different regions are used: one from Dallas-Fort Worth and another from the San Francisco Bay Area. Two multinomial logit mode choice models are built with these data sets and then examined for cross transferability.

Rest of the paper is organized as follows. Section II provides literature review on model transferability. Data descriptions and data preparation are provided in Section III. Methodology is described in Section IV. Section V presents the empirical results. Finally, summary and conclusions are provided in section VI.

II. LITERATURE REVIEW

Transferability is an issue in two dimensions, space and time. Many studies have been performed on both temporal and spatial transferability. McCarthy [1] analyzed the temporal characteristics of work trip behavior in the San Francisco Bay Area. From a pre Bay Area Rapid Transit (BART) set of observations, a multinomial logit model of work trip modal choice was estimated and an updated model was developed by transferring the BART model's coefficient vector and freely estimating the parameters associated with BART alternative specific variables. The paper supported the hypothesis that the coefficient estimates, in a short run framework, remain stable. The paper also mentioned that the model would be transferable to a different population facing similar transportation alternatives and exhibiting a similar socioeconomic character. Badoe and Miller [2] also mentioned that a well context-specified model should be able to represent the decision-making process in other contexts, as long as the basic nature of the decision-making process remains the same. Andrade et al. [3] have studied the temporal transferability of a Multinomial Logit Model (MNL) and a hybrid Neuro-Fuzzy Multinomial Logit Model (NFMNL). The models were estimated by 2000 data and applied on 2004 data of same residential location. Their study suggested that directly transferred models may not be able to capture the changing aspects of the transportation system between the estimation and the application context.

Some works have been performed to see whether the models need to be updated or not, to get better transferability to accommodate the future changed conditions. Sanko and Morikawa [4] proposed to update alternative-specific constant so that estimated disaggregate discrete choice models can be used in applied context. They also investigated the factors affecting the transferability of the updated constants. Their analysis showed that the factors could depend on regional

characteristics and past travel behaviors (inertia) and were anti symmetric and path-dependent on changes in the level of service. Agyemang-Duah and Hall [5] analyzed the spatial transferability of an ordered response model of shopping trip generation in Metropolitan Toronto. The paper investigated both the performance of direct model transfer without updating the transferred coefficients and the performance of a scaling updating procedure that adjusted the model parameters by using small-sample data from the region to which the model was to be applied. The results of this spatial transferability analysis showed that a directly transferred ordered response model performed reasonably well in predicting the aggregate shares in the application (new) context. Revising the constant terms and the scalars in the model substantially improved the predictive ability of the transferred model.

Karasmaa [6] compared the alternative methods of spatial transfer as a function of sample size. The Helsinki Metropolitan Area database was used to estimate the model and the Turku database was used for application context. Bayesian updating, combined transfer estimation, transfer scaling and joint context estimation transfer procedures were examined. Model transferability was tested for six different sample sizes. The model transferability was examined by comparing the transferred models to the models estimated using the entire set of the data which regarded as the best estimate representing ''the real situation''. The results indicated that joint context estimation gives the best prediction performance in almost all cases. In particular, the method is useful if the difference in the true parameters between the two contexts is large or only some of the model coefficients are precise. The applicability of joint context estimation can be improved by viewing the coefficients as variable-oriented, as well as by emphasizing precise and imprecise coefficients differently.

In practice, the estimated data and application data might not be identical. One or more important variables might not exist in both of the data sets. The effect of omission of relevant explanatory variables on model transferability was investigated by Koppelman and Wilmot [7]. They analytically formulated the relationship between variable omission and its impact on estimation and prediction. An empirical analysis of transferability of mode choice model among three geographic sectors of the Washington, D.C. was undertaken to verify and clarify the analytical results. Specification effects on model transferability were examined in the context of partial model transfer in which alternative specific constants were adjusted to give the best local fit conditional on the transfer of the remaining parameter values. The research gave following tentative conclusions: (1) improvements in model specification lead to improvement in absolute transfer effectiveness measured against some fixed reference point; (2) improvements in model specification may or may not lead to improvement in relative transfer effectiveness measured against the corresponding predictive ability of a similarly improved local model; (3) there exists some minimum adequate specification quality level which must be achieved in order to obtain reasonable levels of model transferability.

Many studies have been performed in the field of transferability. Both temporal and spatial transferability were covered in the previous studies, although temporal transferability got main focus in this respect. Most of the research on spatial transferability was performed on identical data sets. Only Koppelman and Wilmot [7] investigated the effect of omission of relevant explanatory variables on the level of transfer effectiveness. In this research, we will examine the effect of omission of one of the most important variables (travel cost) in spatial transferability.

III. DATA

A. Data Description

For this research, two data sets are used: one from the 2000 San Francisco Bay Area Travel Survey (BATS) and the other from the 1996 Dallas-Fort Worth (DFW) Household Activity Survey. The BATS data set consists of four data files (trip data, household data, personal data and level of service data) and the DFW data set consists of six data files (trip data, household data, personal data, employment data, school data and level of service data).

Both the BATS and DFW datasets do not contain all the important explanatory variables for building the mode choice model. As the objective of this study is to examine the transferability of mode choice model between these two regions, only the variables exist in both datasets are selected as explanatory variables. As for example, DFW dataset does not have travel cost variable, therefore travel cost is not included for BATS dataset. After some preliminary statistical analysis travel time, travel distance, age, gender, license driver (whether a person is a license driver or not), number of vehicles in a household and household income are selected as explanatory variables. Some descriptive statistics of age, household size, number of vehicles in household and household income are provided in Table I. From the table it can be observed that age and number of person per household are same in two areas, but people in San Francisco Bay area have higher income and won higher number of vehicles than those of DFW area. Frequency distribution of gender and choice alternatives are provided in Tables II and IV.

Table I: **Descriptive Statistics of Age, Household Size, Number of Household Vehicle and Household Income**

Variable	BATS			DFW		
	No. Observation	Mean	Std. Deviation	No. Observation	Mean	Std. Deviation
Age	33402	38.42	20.83	6166	37.76	21.43
Number of persons in household	14529	2.30	1.25	2750	2.24	1.2
Total number of HH vehicles	14529	1.85	0.96	2750	1.68	0.90
Household income in $1000s	14529	80.69	48.40	2750	49.71	33.1

Table II: **Frequency Distribution of Gender**

Gender	BATS		DFW	
	Frequency	Percent	Frequency	Percent
Female	17230	51.58	3251	52.72
Male	16172	48.42	2915	47.28
Total	33402	100	6166	100

Table III: **Frequency Distribution of Trip Mode (BATS Data)**

Trip Mode	Frequency	Percent
Drive Alone	99462	50.49
Shared Ride	65576	33.29
Transit by walk access	4779	2.43
Transit by drive access	1922	0.98
Walk	16301	8.27
Bike	2618	1.33
Air	26	0.01
Taxi	135	0.07
Others	987	0.50
Is a pure recreation trip beginning and ending at home	5192	2.64
Total	196998	100

Table IV: **Frequency Distribution of Trip Mode (DFW Data)**

Travel Mode	Frequency	Percent
Car	21295	91.92
Bus	401	1.73
School bus	395	1.71
Walk	974	4.20
Bike	68	0.29
Other and unknown	34	0.15
Total	23167	100

B. Data Preparation

In order to estimate models, data were prepared for the nlogit software. The personal data and household data were added to the trip record data file by matching the household id and person id. For the DFW data, the employment and school data were also added to the trip data file. Then the level of service data were added to the trip files by matching the origin and destination of each trip.

Data cleaning was performed on both of the BATS data and DFW data. If there was any missing value in origin, destination, mode, distance, license driver, gender, household vehicle, household income, travel time that trip record was eliminated from the datasets. Final BATS dataset contains 191,806 trip records and DFW dataset contains 8,078 records. Both of the datasets were randomly split into two datasets: 90% data (estimated data) were selected for model estimation and

10% data (application data) were selected for model prediction. The estimated dataset of BATS contains 172,623 trips and DFW dataset contains 7,273 trips; whereas predicted dataset of BATS contains 19,183 trips and DFW contains 805 trips.

From Tables III and IV, it can be seen that the available mode alternatives are not the same in both of the data sets. In the DFW data set, the number of alternatives is six, but the modal shares of bus, school bus, bike and others are very small. Therefore, bus and school bus were merged together under a "transit" category, while walk, bike and others were merged together under an "other" mode. In the BATS data, there are nine alternative modes. Here, drive alone and shared ride were merged together under a "car" mode, transit by walk access and transit by drive access were merged together under a "transit" mode and walk, bike, air, taxi and other were merged together under an "other" mode. So for both of the data sets, the choice alternatives are car, transit and other. Modal shares of car, transit and other in BATS are 86%, 3.5% and 10.5% respectively, while in DFW data the shares are 88.6%, 4.6% and 6.8%.

IV. METHODOLOGY

Multinomial logit model with tree choice alternatives (car, transit and other) is used for developing discrete mode choice model. Each choice alternative has its own utility function. Mode specific constants are added to the transit and other modes. Personal attributes (age, gender license driver, etc.), household attributes (number of vehicles in household, household income, etc.) and level of service attributes (in vehicle travel time, out of vehicle travel time) are tried in the utility functions. The model building started with constant-only model and was developed gradually by adding one variable at a time. If the added variable was found statistically significant, the variable was kept in the utility function otherwise the variable was deleted from the utility function. Finally, the eliminated variables were tried once again to see if those variables were still insignificant or not. In final model, the sign of the parameter are according to the expectations and all explanatory variables are statistically significant at 95% confidence level. After trying different specifications, the best model was selected based on performance measures (log likelihood function, adjusted rho square). The utility functions of the final models of these two regions are provided below.

BATS Model:
$U(Car) = b3*OVTT + b4*NVehC + b5*LicDvrC + b6*HHIncC$
$U(Transit) = b0*unotr + b3*OVTT$
$U(Other) = b1*unoor + b3*OVTT + b7*DistO + b8*GenO$

DFW Model:
$U(Car) = b4*NVehC + b5*LicDvrC + b6*HHIncC$
$U(Transit) = b0*unotr$
$U(Other) = b1*unoor + b7*DistO + b8*GenO$

Where,
OVTT = Out of vehicle travel time
NVehC = Total Number of vehicles in a household
LicDvrC = Binary variable, whether the person is a licensed driver (1) or not (0)
HHIncC = Household Income
DistO = Travel distance
GenO = Binary variable, Male=1, Female=0

It can be seen from the above utility functions that some important variables (in vehicle travel time, travel cost, age, etc.) are missing in the models. We tried in vehicle travel time, but as we did not get the correct sign, we deleted that variable from the model. Person's age was also tried in the utility function as mode specific variable but turned out to be statistically insignificant. Travel cost is one of the most important variables for mode choice model. This variable is not in the model because: (1) DFW data set does not have cost variable; (2) The objective of this research is to examine the transferability of the mode choice model when one/more important variables are missing in data set.

V. RESULTS

A. Model Estimation

Models are estimated from the estimated data set (90% of the total sample) of BATS and DFW data. Estimated parameters are provided in Tables V and VI. All the parameters have correct sign and are statistically significant at 95% confidence level.

People generally do not like to spend time for waiting, transferring or other out of vehicle activities. Therefore, the probability of choosing a mode decreases with the increasing value of out of vehicle travel time. This is captured by the negative sign of OVTT. If a household has more vehicles, household members get more available cars for driving and probability of choosing a car increase, which is represented by the positive sign of NVeh. In the same way person having a driving license is more likely to drive, which is captured by the positive sign of LicDvrC. Generally high income people have more cars and like to drive instead of riding public transit. The corresponding HHIncC has positive sign which is according to expectation. The coefficient of distance parameter (DistO) is negative. This indicates that the probability of choosing other mode decreases with the increases of travel distance. This is also according to the expectation. In other mode category the major portion is walk mode and bicycle mode. When the distance is long, people are not willing to walk or

riding bicycle. The coefficient of GenO is positive which indicates that the probability of a person choosing other mode is higher if the person male. This is also according to the expectation. As it mentioned earlier that other mode consists of mainly walk and bike mode. Walking or riding bicycle is laborious task. Male are more likely to take extra physical load than female.

Table V: **Estimated Parameters of BATS Model**

Explanatory variables	Car		Transit		Other	
	Param.	t stat	Param.	t stat	Param.	t stat
Constant	-	-	-0.7260	-18.80	0.4081	13.69
OVTT	-0.0159	-14.16	-0.0159	-14.16	-0.0159	-14.16
NVehC	0.6200	66.97	-	-	-	-
LicDvrC	0.5852	36.25	-	-	-	-
HHIncC	0.0083	21.05	-	-	-	-
DistO	-	-	-	-	-0.1238	-48.75
GenO	-	-	-	-	0.2665	16.32
Number of cases	172623					
Log likelihood at convergence	-75867.10					
Log likelihood for constants-only model	-83395.75					
Adjusted Rho^2	0.09					

Table VI: **Estimated Parameters of DFW Model**

Explanatory variables	Car		Transit		Other	
	Param.	t stat	Param.	t stat	Param.	t stat
Constant	-	-	-0.2975	-2.90	1.0132	8.22
NVehC	0.8378	12.66	-	-	-	-
LicDvrC	1.5717	18.62	-	-	-	-
HHIncC	0.0076	4.24	-	-	-	-
DistO	-	-	-	-	-0.3646	-12.09
GenO	-	-	-	-	0.2694	2.61
Number of cases	7273					
Log likelihood at convergence	-2411.95					
Log likelihood for constants-only model	-3096.45					
Adjusted Rho^2	0.22					

B. Application of the Estimated Model

Estimated model from BATS dataset (90% of the sample) are applied to the rest 10% of the dataset and also applied to the estimated data (90% of the sample) of DFW. The modal shares predicted by the models and the sample shares of the datasets are provided in Table VII. And estimated model from DFW dataset (90% of the sample) are applied to the rest 10% of the dataset and also applied to the estimated data (90% of the sample) of BATS. The modal shares predicted by the models and the sample shares of the datasets are provided in Table VIII. From Tables VII and VIII it can be said that both of the models predict very good. Moreover prediction within the region is better than outside the region.

These results are consistent with Badoe and Miller [2] where they argued that "model should be able to represent the decision-making process in other contexts, as long as the basic nature of the decision-making process remains the same". According to data, these two regions have some similarities in terms of average age, average household size, average number of household vehicles and etc. Although the choice alternatives are different in these two regions, the modal share of the dominant mode (car) is almost same. In this research we also proved that without important variables (cost, in vehicle travel time) model could be transferred to other location.

Table VII: **Prediction by Estimated BATS Model**

	BATS (10% Data)		DFW (90%)	
Mode	% Sample Share	% By Model	% Sample Share	% By Model
Car	86.20	87.10	88.80	83.82
Transit	3.50	3.55	4.50	3.35
Other	10.30	9.35	6.70	12.83
Total	100	100	100	100

Table 10: **Prediction by Estimated DFW Model**

	DFW (10% DATA)		BATS (90%)	
MODE	% SAMPLE SHARE	% BY MODEL	% SAMPLE SHARE	% BY MODEL
CAR	87.00	89.28	86.0	92.35
TRANSIT	6.00	4.76	3.50	3.25
OTHER	7.10	5.96	10.5	4.40
TOTAL	100	100	100	100

VI. SUMMARY AND CONCLUSIONS

In this research spatial transferability of a mode choice model was examined. Multinomial logit models were used for building mode choice models. Two sets of data from two different regions were used: one from Dallas-Fort Worth and another from the San Francisco Bay Area. When some important explanatory variables exist in both of the dataset, those variables were considered for developing the models. Both of the datasets were cleaned in order to get rid of missing values of the selected variables. In cleaned datasets, there were 196,998 trip records in BATS dataset and 23,167 trip records in DFW dataset. Originally there were nine travel modes in BATS dataset and six travel modes in DFW dataset. In order to make the datasets similar some modes were combined into one mode in both of the datasets. Finally, both of the datasets contained only three modes (car, transit and other). Each of the datasets was randomly divided into two samples: estimated dataset (90% of the whole data) and application dataset (10% of the sample).

Many explanatory variables were tried to build the mode choice models. If the sign coefficient of a variable was according to expectation and the parameter was statistically significant, only then that variable was kept in the final models. Best models were selected based on log likelihood and adjusted rho square values. Model estimated from BATS's estimated dataset was applied to BATS's application dataset as well as DFW's estimated dataset. In the same way model estimated from DFW's estimated dataset was applied to DFW's application dataset as well as BATS's estimated dataset. The findings of the study are summarized as follows.

A. Findings

- As there was no cost variable in DFW dataset, cost variable was not considered for model building. Even though, the final models turned out to be good in terms of performance measures. This study indicates that without some important variables, credible models could be build.
- The parameters of explanatory variable of the models estimated from two regional datasets are similar in terms of sign and magnitude.
- Modal shares predicted by models are almost same as sample shares within the region and outside the region, but prediction within the region is better than outside the region.
- In this project no adjustment was done in mode specific constants. Even though, both models showed very good transferability.
- In this study, it became clear that without some important variables the models are transferable from one region to another region which was the main objective of this project.

Most of the research on spatial transferability was performed on identical data sets. Only Koppelman and Wilmot [7] investigated the effect of omission of relevant explanatory variables on the level of transfer effectiveness. This study examined the effect of omission of one of the most important variables (travel cost) in spatial transferability. Therefore, the findings of this study are very important. If the model can be transferred between regions, models developed in one region can be applied to other place that will significantly reduce the time, effort and cost of conducting travel survey and estimating model.

B. Recommendations

- Models goodness of fit could be improved by considering the following points:
  - Model could be built according to trip purposes (e.g. work trip, school trip etc.)
  - More explanatory variables could be tried.
  - Number of choice alternatives could be more.
- Weight factor could be used when model is applied to other region.
- In this study both of the dataset have some similarities; the transferability should also be checked with datasets from two different types of region.

REFERENCES

[1]. P. S. McCarthy, "Further evidence on the temporal stability of disaggregate travel demand models," *Transportation Research-B,* vol. 16B, no. 4, pp. 263-278, 1982.

[2]. D. A. Badoe and E. J. Miller, "Analysis of the temporal transferability of disaggregate work trip mode choice models," *Transportation Research Record: Journal of the Transportation Research Board,* no. 1493, pp. 1-11, 1995.

[3]. K. Andrade, K. Uchida, A. Nicholson, S. Kagaya and A. Dantas, "Investigating the temporal transferability of transport modal choice models: An approach based on GIS data base," *in Proc. Eastern Asia Society for Transportation Studies,* vol. 6, 2007.

[4]. N. Sanko and T. Morikawa, "Temporal transferability of updated alternative-specific constants in disaggregate mode choice models," *Transportation,* vol. 37, pp. 203-219, 2010.

[5]. K. Agyemang-Duah and F. L. Hall, "Spatial Transferability of an ordered response model of trip generation," *Transportation Research-A,* vol. 31, no. 5, pp. 389-402, 1997.

[6]. N. Karasmaa, "Evaluation of transfer methods for spatial travel demand models," *Transportation Research-A,* vol. 41, pp. 411-427, 2007.

[7]. F. S. Koppelman and C. G. Wilmot, "The effect of omission of variables on choice model transferability," *Transportation Research-B,* vol. 20B, no. 3, pp. 205-213, 1986.