Comparison of Non-Compensatory Models of Driver's Choice on Dynamic Park and Ride

by

Toshiyuki Yamamoto,
Department of Geotechnical and Environmental Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
Tel: +81-52-789-4636, Fax: +81-52-789-3738
e-mail: yamamoto@civil.nagoya-u.ac.jp

Shinya Kurauchi
Department of Geotechnical and Environmental Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
Tel: +81-52-789-3565, Fax: +81-52-789-3738
e-mail: kurauchi@civil.nagoya-u.ac.jp

and

Takayuki Morikawa
Graduate School of Environmental Studies, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
Tel: +81-52-789-3564, Fax: +81-52-789-3738
e-mail: morikawa@civil.nagoya-u.ac.jp

January 2002
Revised June 2002

Prepared for presentation and publication at the
9th World Congress on Intelligent Transportation Systems,
Chicago, Illinois, October 2002

Comparison of Non-Compensatory Models of Driver's Choice on Dynamic Park and Ride

Toshiyuki Yamamoto[1], Shinya Kurauchi[1] and Takayuki Morikawa[2]

[1] Department of Geotechnical and Environmental Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

[2] Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

SUMMARY

Non-compensatory models of driver's choice on dynamic park and ride are developed and examined on the predictability in this study. One of the data mining tools, C4.5, is used to develop decision tree and production rules of driver's choice. The generated decision tree and production rules are compared with the semi-ordered lexicographic model developed in the preceding study. Compared are the similarity of the estimated decision making structures and the distribution of the segments not correctly represented by the models as well as goodness-of-fit and hit ratio. The comparison shows that the semi-ordered lexicographic model has a higher goodness-of-fit and hit ratio than models with data mining tools. The results also suggest that the models developed in this study represent different decision rules from, but have similar distributions of the segments not represented by the models with the semi-ordered lexicographic model. Empirically, the consistent results by models with data mining tools and the semi-ordered lexicographic model suggest that the parking congestion level in CBD has a significant effect on the choice behavior.

KEYWORDS: decision tree, data mining, dynamic park and ride, non-compensatory model, production rules, semi-ordered lexicographic model, stated preference

INTRODUCTION

The park and ride (P&R) system is one of the travel demand management measures to reduce vehicle use. P&R system offers parking spaces connected to transit, and encourages using the transit to enter the congested area instead of driving their vehicles into the congestion. Recently, intelligent transportation system (ITS) enables to attract drivers more actively to P&R by showing drivers real time information such as estimated travel time to the destination by car and transit, congestion at the parking in the area of the destination, etc. The dynamic P&R (DP&R) system uses variable message signs on the road and/or in-vehicle information devices to show drivers such information, so the drivers are able to choose the DP&R en-route considering real time travel information. The usage of the DP&R system relies on en-route decisions by the drivers who originally intend to drive their vehicles toward their destinations. Thus, the driver's en-route choice behavior should be thoroughly understood to know what kinds of conditions are required, and to determine which information to be offered for a successful implementation of DP&R.

Most of the former studies on mode or route choice behavior have applied discrete choice models with the linear-in-attributes utility function. The linear-in-attributes utility function assumes the compensation of the utility in the sense that a low score with one attribute can be compensated by a high score with another attribute. As many researches inside and outside the transportation field suggested, the compensatory model is one of the many candidates representing a choice behavior, but many other types of models such as the satisficing concept of Simon (1950), or non-compensatory rules (Foerster, 1979) outperform the compensatory model in some cases. These researches suggest that individuals employ more simplified decision rules rather than the normative rule which evaluates all the attributes of all available alternatives simultaneously, especially under the choice situations with a large number of alternatives or attributes, or under the time pressure.

The en-route choice on DP&R use investigated in this study is processed within a small amount of time while the driver keeps driving, and more likely tends to form non-compensatory nature because of the time pressure. A semi-ordered lexicographic model (Coombs, 1964), which is one of the non-compensatory models, was developed and applied for the driver's choice on DP&R use in the preceding study (Kurauchi and Morikawa, 2001), and the results showed a higher goodness-of-fit and AIC (Akaike, 1973) than a conventional compensatory model. The results of the preceding study suggest that drivers form non-compensatory decision making process with a reasonable probability. However, the maximum likelihood estimation used in the preceding study has a difficulty in converging because of a high non-linearity of the parameters, so the estimated model in the study includes some insignificant coefficients.

The knowledge discovery and data mining methods have been developed in conjunction with the information technology. Data mining methods have a higher computability with a large sample size, and some data mining methods also have flexible framework in representing the effects of the attribute and capable of non-compensatory decision rules. One of the data mining methods, C4.5 (Quinlan, 1993), is used to represent en-route driver's choice on the use of DP&R in this study. C4.5 generates a decision tree and production rules, which are the non-compensatory decision rules. Wets et al. (2000) and Yamamoto et al. (2002) have applied C4.5 in the transportation field recently. C4.5 is applied to generate a decision tree and production rules of driver's choice on DP&R in this study. The generated decision tree and production rules are compared with the semi-ordered lexicographic model to examine the efficiency of the method. The comparison includes the similarity of the estimated decision making structures and the distribution of the segments not correctly represented by the models as well as goodness-of-fit and hit ratio.

MODELS

In this section, the semi-ordered lexicographic model developed in the preceding study is described briefly, and then, C4.5 is described to clarify the similarities and differences between them.

SEMI-ORDERED LEXICOGRAPHIC MODEL

The semi-ordered lexicographic model is an extension of the lexicographic model. The semi-ordered lexicographic model assumes that decision maker has his own importance rank of attributes, and compares the alternatives in the most important attribute. The alternative with a better attribute value is chosen in binary choice if the difference in the attribute values of the alternatives is larger than a specific threshold, which is assumed to follow lognormal distribution in the study. If the difference is smaller than the threshold, the second most important attribute is used to compare in basic semi-ordered lexicographic rules. The successive process assumed here, however, results computational difficulty in the parameter estimation, so a conventional compensatory utility function and binary logit model is used instead of the comparison of the second most important attribute in the preceding study. Thus, the semi-ordered lexicographic model developed in the preceding study can be considered as a general model including conventional compensatory choice model as a special case when all the thresholds are set to be large numbers. For details, see Kurauchi and Morikawa (2001).

In the preceding study, a set of the attributes are predetermined, and the membership functions representing latent segments who consider one of the attributes as their most important attribute are estimated by the model estimation. The maximum likelihood estimator is used to obtain the parameters, thus the statistical inferences are obtained in the same way as the traditional utility-based binary choice models. The approach, however, has a difficulty in such a case that the preliminary knowledge on the choice behavior is limited or the number of the attributes is large, because the likelihood function includes a high non-linearity of the parameters and requires a computational burden.

DECISION TREE AND PRODUCTION RULES

The C4.5 algorithm is one of the supervised learning algorithms. It generates a decision tree and production rules in order. A decision tree represents choice behavior as sequential examinations of attributes, as in the theory of elimination by aspects (Tversky, 1972). On the other hand, production rules represent choice behavior by a set of IF-THEN rules that determine the choice according to the conditions as indicated by sub-sets of the attributes, as in a production system (Newell and Simon, 1972). Both the decision tree and production rules represent non-compensate choice processes like the semi-ordered lexicographic model, elimination by aspect and the production system. The difference between C4.5 and these theories are, however, that the former generates the choice structure to best represent observed choices inductively without any presumptions, while the latter predetermine the choice structure before the estimation.

In the C4.5 algorithm, the sample cases are subdivided recursively into segments based on explanatory variables, until each segment finally contains only those sample cases that made the same choice. Next, the resulting decision tree is simplified by pruning some divisions to merge the segments to generate the final decision tree. To make a production rule set, the decision tree before pruning is transformed into a set of production rules that uniquely segment the sample, and the resulting rule set is simplified by deleting some conditions in a rule, or eliminating some rules to generate the final production rules. For details, see Quinlan (1993).

The generations of the decision tree and production rules have multiple steps as shown above. On each step, the information theory rather than more general statistical theory is applied: the recursive subdivision is based on the concept of entropy, the pruning of the tree is based on the confidence limit of the expected errors, and the generalization of the production rules uses the confidence limit of the expected errors and the Minimum Description Length (MDL) principle (Rissanen, 1983). This unable to obtain straightforward overall statistical inferences on the resulting decision tree and production rules. Also, the data mining tools such as decision tree and production rules are thought to have a higher risk of not representing a population, but over-fitting the sample cases.

DATA SET

The data set used in this study is obtained from the survey conducted in Nagoya, Japan, in 1997 (Nakamura et al., 1999). The survey was conducted to investigate the drivers' adoption to the DP&R system in making commuting and shopping trips whose origins are in suburbs and destinations are in CBD. Seven hypothetical scenarios for both commuting and shopping trips were presented to each respondent, and the respondents were asked to choose DP&R or driving directly to the destination. The attributes describing the scenarios consist of travel times and travel costs for both alternatives, the distance from parking lot to the transit station for DP&R, and the parking congestion level in CBD for the drive directly to the destination.

The questionnaires were handed out to randomly sampled residents and collected at the households later. The sample size is 1102 respondents, and the response rate is 90.4%. From the full sample, 4778 cases on the shopping trips for which pertinent explanatory variables are available were used in the model estimation. The possible biases caused by the multiple observations from the same respondent are not considered in this analysis. Table 1 shows the explanatory variables used in the analysis, which are consistent with Kurauchi and Morikawa (2001) for comparison.

RESULTS

GENERATED DECISION TREE AND PRODUCTION RULES

Presented in Figure 1 is the generated decision tree. In the figure, the left-hand side of the colon represents the condition with which a segment is divided into sub-segments, and the right-hand side is the predicted choice for each sub-segment. If the right-hand side is blank, there are additional divisions for the sub-segment, which are placed in the next rows. In Figure 1, for example, *Park-full* is chosen for the first division, and if *Park-full* is 1, then the segment is divided again by *Cost*. If *Cost* is not greater than 0.4, then the choice is DP&R, and if *Cost* is greater than 0.4, then the segment is divided again by *Time.*

Table 1. Explanatory variables used in the analysis

Variable	Definition
Cost	Total travel cost (DP&R – Drive in 1,000 yen).
Time	Total travel time (DP&R – Drive in hour).
Dist-near	1, if the station is close to the present point; 0, otherwise.
Park-f-v	1, if parking lot in CBD is few vacant; 0, otherwise.
Park-full	1, if parking lot in CBD is fully occupied; 0, otherwise.
Child	1, if the individual has a child (or children) in the household; 0, otherwise.
Male	1, if the individual is a male; 0, otherwise.
Age30	1, if the individual's age is less than 30; 0, otherwise.
Age60	1, if the individual's age is over 60; 0, otherwise.
Student	1, if the individual is a student; 0, otherwise.
High-inc	1, if the individual's income is over 10 million yen per year; 0, otherwise.
Carown2	1; if the individual owns more than 2 cars in the household; 0, otherwise.

```
Park-full = 1 :
|   Cost ≤ 0.4 : DP&R
|   Cost > 0.4 :
|   |   Time > -0.333 : Drive
|   |   Time ≤ -0.333 :
|   |   Child = 1 : Drive
|   |  Child = 0 :
|   |   |   Male = 1 : Drive
|   |   |   Male = 0 : DP&R
Park-full = 0 :
|   Time > -0.5 : Drive
|   Time ≤ -0.5 :
|   |   Cost > 0 : Drive
|   |   Cost ≤ 0 :
|   |   |   Male = 1: Drive
|   |   |   Male = 0 : DP&R
```

Figure 1. Generated decision tree of driver's choice on DP&R use

The result suggests that the driver first changes the decision structure according to the parking congestion level in CBD. If the parking lot in CBD is fully occupied, the difference in travel costs is considered as the next attribute to determine the choice. On the other hand, if the parking lot in CBD is not fully occupied, the difference in travel time is considered. The result also suggests that the thresholds for the differences in travel cost and time are severer for choosing DP&R if the parking lot in CBD is not fully occupied, and that male has a simpler decision structure than female, because if male, the difference in travel time is not considered when the parking lot in CBD is fully occupied, nor the difference in travel cost is considered when the parking lot in CBD is not fully occupied.

The production rules generated by the algorithm are presented in Figure 2. After **if**, the conditions are presented where all the conditions must be satisfied to be included in the segment, and the predicted choice for the segment is presented after **then**. The rules are aligned in the order of priority. If the conditions of the upper rule are satisfied, the lower rules are not considered.

The result shows the same tendency as the decision tree in this empirical analysis. The parking congestion level in CBD, the differences in travel times and costs are considered for determining the choice, and male has a simpler decision structure than female, because the first two rules are applied only to females. Male has only one set of conditions to choose DP&R that the parking lot in CBD is fully occupied and the travel cost for DP&R is not higher in 400 yen than the drive to the destination.

COMPARISON WITH SEMI-ORDERED LEXICOGRAPHIC MODEL

The generated decision tree and production rules are compared with the semi-ordered lexicographic model developed and estimated in the preceding study (Kurauchi and Morikawa, 2001), though the estimation results of which are not shown here because of the limited space. The estimated semi-ordered lexicographic model has four latent classes that each class considers one of the four explanatory variables: the differences in travel time, travel cost, the parking congestion level in CBD, and the distance from parking lot to the transit station. The aggregate shares for each latent class are 16% for travel time, 17% for travel cost, 5% for the parking congestion in CBD, and 62% for the distance from parking lot to the transit station. The results of the decision tree and the production rules generated by data mining tools and the semi-ordered lexicographic model are not consistent that the generated decision tree and production rules do not include the distance from parking lot to the transit station as the explanatory variable in the final models, while the variable is considered as the most important variable by the largest latent class in the semi-ordered lexicographic model. The results imply that the decision tree and production rules represent different choice structures from the semi-ordered lexicographic model, although the all models represent non-compensatory choice structures.

```
  if  Child = 0
      Male = 0
      Park-full = 1
      Cost ≤ 0.4
      Time ≤ -0.333
then  DP&R

  if  Male = 0
      Cost ≤ 0
      Time ≤ -0.5
then  DP&R

  if  Cost ≤ 0.4
      Park-full =1
then  DP&R

  if  Cost > 0.4
      Time > -0.333
then  Drive

  if  Child = 1
      Cost > 0.4
then  Drive

  if  Time > -0.5
      Park-full = 0
then  Drive

  if  none of the above
then  Drive
```

Figure 2. Generated production rules of driver's choice on DP&R use

However, *Male* and *Child* have highly significant coefficients in the membership functions for the latent class with travel time as the most important variable in the semi-ordered lexicographic model, which is consistent with the decision tree and production rules, where the two variables alone are included in the final models. Also, the average probability of the drive alternative being rejected by the parking congestion level in CBD is 100%, which is much higher than that of other attributes in the semi-ordered lexicographic model. This result means that the parking congestion level in CBD has a significant effect on the choice behavior, which is consistent with the decision tree, where the same variable is used as the primal variable considered among all the variables.

In order to examine the predictive performance, the hit ratios are calculated as the fraction of the cases whose choices are predicted correctly. The prediction is regarded to be correct if the alternative actually chosen by the respondent has a higher probability predicted by the model. The hit ratio is a primitive index if compared to the statistical inferences. However, the statistical inferences are not obtained from the decision tree and production rules directly as stated above, while log-likelihood and t-statistics are used as the goodness-of-fit of the model, and statistical significances of the parameters, respectively, in the maximum likelihood estimations. In order to compare the goodness-of-fit of the decision tree and the production rules with the semi-ordered lexicographic model, and examine the statistical significances of the segments divided by the decision tree and the each rule in the production rules, binary logit models only with the dummy variables representing segments divided by the decision tree, and that representing the conditions used in the production rules are estimated, respectively.

Table 2 shows hit ratios and Log-likelihood values at convergence by models, which include the results of conventional binary logit model as a reference. The table suggests that the hit ratios of decision tree and the production rules are the same and higher than that of binary logit model, but lower than the semi-ordered lexicographic model. The log-likelihood values at convergence of the decision tree and production rules are smaller than both semi-lexicographic model and binary logit model, although all the rules in the production rules and almost all the segments divided by the decision tree have statistically highly significant coefficients. The results suggest that the semi-ordered lexicographic model outperforms the decision tree and the production rules in this empirical analysis.

In order to investigate the similarity among the models from the viewpoint of the distribution of the predictability in the sample population, Pearson's correlations of the probabilities of choosing the alternative actually chosen, and the cross tabulations and Spearman's correlations of the distributions of the correct/incorrect predictions are calculated.

Table 2. Hit ratio and log-likelihood at convergence

	Hit ratio	Log-likelihood
Binary logit model	0.690	-2748
Semi-lexicographic model	0.697	-2728
Decision tree	0.693	-2798
Production rules	0.693	-2810

Pearson's correlations are shown in Table 3. The table shows that all the four models have quite high correlations of the predictions. The table also shows that the semi-ordered lexicographic model has a higher correlation with the binary logit model than the decision tree and the production rules, and that the decision tree and the production rules have higher correlations with each other than the semi-ordered lexicographic model.

Table 3. Pearson's correlation of the predictions by models

	(1)	(2)	(3)	(4)
(1) Binary logit model	1.000	0.979	0.916	0.913
(2) Semi-lexicographic model		1.000	0.911	0.898
(3) Decision tree			1.000	0.958
(4) Production rules				1.000

The cross tabulations and Spearman's correlations of the distributions of the correct/incorrect predictions are shown in Table 4. The decision tree is excluded because the predictions by the decision tree and the production rules are found to be identical. The tables show that the distributions of the incorrect predictions are highly correlated among the models, and that the binary logit model and production rules have the highest correlation. The results suggest that particular segments in the population are not correctly represented by the models regardless of the choice structures assumed by the models. On the other hand, the choices of other segments are correctly represented by both compensatory and non-compensatory models.

CONCLUSIONS

The decision tree and the production rules of driver's choice on DP&R are developed using stated preference data, and the results are compared with the semi-ordered lexicographic model. The results suggest that the decision tree and the production rules represent a different choice structure from that of the semi-ordered lexicographic model. However, some consistencies are also found among the models, which include that the parking congestion level in CBD has a significant effect on the choice, and that the choice structure varies according to the gender and the presence of the children in the household. These consistencies are considered as the robust inferences of driver's choice structure on DP&R, and should be highly accounted for when implementing the DP&R system. Especially, the provision of the information on the parking congestion level in CBD is found to take a significant effect on the DP&R use in this analysis.

Table 4. Cross tabulation of the predictions by models

(a) Predictions by binary logit model and semi-lexicographic model

	Prediction by semi-lexicographic model	
	Correct	Incorrect
Prediction by binary logit model		
Correct	3086	209
Not correct	245	1238
Spearman's correlation coefficient	0.777 (s.d. = .010)	

(b) Predictions by binary logit model and production rules

	Prediction by production rules	
	Correct	Incorrect
Prediction by binary logit model		
Correct	3129	166
Not correct	183	1300
Spearman's correlation coefficient	0.829 (s.d. = .009)	

(c) Predictions by semi-lexicographic model and production rules

	Prediction by production rules	
	Correct	Incorrect
Prediction by semi-lexicographic model		
Correct	3091	240
Not correct	221	1226
Spearman's correlation coefficient	0.772 (s.d. = .010)	

The comparisons of the predictability suggest that the semi-ordered lexicographic model outperforms the decision tree and the production rules in this empirical analysis.

However, particular segments in the population are not correctly represented by the models regardless of the choice structures assumed by the models. The different choice structures not considered in this study might be combined to the models in this study, and examined its coverage of the population in order to a better understanding of the choice behavior on DP&R system. In the case that we do not have enough preliminary knowledge on the choice behavior, the knowledge discovery and data mining may be one of the tools to explore, because these tools have a possibility to explore the choice behavior endogenously from the data.

ACKNOWLEDGEMENTS

The authors are grateful to Takamasa Akiyama of Gifu University, the discussion with whom the comparison of the models on the segments not correctly represented by the models are based on.

REFERENCES

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle, In Petrov, B. and Csake, F. (eds.) *Second International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267-281.

Coombs, C. H. (1964) *Theory of Data*, John Wiley, New York.

Foerster, J. F. (1979) Mode Choice Decision Process Models: A comparison of compensatory and non-compensatory structures, *Transportation Research*, Vol. 13A, pp. 17-28.

Kurauchi, S. and Morikawa, T. (2001) An Exploratory Analysis for Discrete Choice Model with Latent Classes Considering Heterogeneity of Decision Making Rules. In Hensher, D. (ed.) *Travel Behaviour Research: The Leading Edge*, Pergamon, Oxford, UK, pp. 409-423.

Nakamura, H., Kato, H. and Utsumi, T. (1999) An Applicability Study on the Dynamic Park and Ride System by A Traffic Simulation Incorporating User Behavior Model. Proceedings for the 6th World Congress on Intelligent Transportation Systems, 8 pages in CD-ROM, Toronto.

Newell, A. and Simon, H.A. (1972) *Human Problem Solving*, Prentice-Hall, Englewood Cliffs.

Quinlan, J.R. (1993) *C4.5 Program for Machine Learning*, Morgan Kaufmann

Publishers, San Mateo, CA.

Rissanen, J. (1983) A Universal Prior for Integers and Estimation by Minimum Description Length, *Annals of Statistics*, Vol. 11, pp. 416-431.

Simon, H.A. (1955) Decision Making and the Search for Fundamental Psychological Regularities: What can be learned from a process perspective? Organizational Behavior and Human Decision Processes, Vol. 69, pp. 99-118.

Tversky, A. (1972) Elimination by Aspects: A theory of choice, *Psychological Review*, Vol. 76, pp. 31-48.

Wets, G., Vanhoof, K., Arentze, T. and Timmermans, H. (2000) Identifying Decision Structures Underlying Activity Patterns: An exploration of data mining algorithms, *Transportation Research Record*, No. 1718, pp. 1-9.

Yamamoto, T., Kitamura, R. and Fujii, J. (2002) An Analysis of Driver's Route Choice Behavior by Data Mining Algorithms. *Transportation Research Record*, forthcoming.