Ann Oper Res (2006) 144: 235–247
DOI 10.1007/s10479-006-0009-8

Evaluation of choice set generation algorithms for route choice models

Shlomo Bekhor · Moshe E. Ben-Akiva · M. Scott Ramming

Published online: 24 May 2006

Abstract This paper discusses choice set generation and route choice model estimation for large-scale urban networks. Evaluating the effectiveness of Advanced Traveler Information Systems (ATIS) requires accurate models of how drivers choose routes based on their awareness of the roadway network and their perceptions of travel time. Many of the route choice models presented in the literature pay little attention to empirical estimation and validation procedures. In this paper, a route choice data set collected in Boston is described and the ability of several different route generation algorithms to produce paths similar to those observed in the survey is analyzed. The paper also presents estimation results of some route choice models recently developed using the data set collected.

Keywords Route choice · Choice set · Model estimation · Logit models

Introduction

This paper discusses choice set generation and route choice model estimation for large-scale urban networks. Recent Intelligent Transportation Systems (ITS) applications have highlighted the need for better models of the behavioral processes involved in route choice. In particular, the desire to provide route guidance based on real-time traffic information to drivers highlights the fact that drivers have imperfect knowledge of traffic conditions and

S. Bekhor (✉)
Technion—Israel Institute of Technology, Faculty of Civil and Environmental Engineering, Technion Campus, Haifa 32000, Israel
e-mail: sbekhor@tx.technion.ac.il

M. E. Ben-Akiva
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: mba@mit.edu

M. S. Ramming
J.F. Sato and Associates, Littleton, CO, USA
e-mail: sramming@jfsato.com

limited information processing ability. Given these limitations, it is not surprising to observe drivers making sub-optimal (from the individual point of view) route choices. Further, drivers also exhibit a wide range of knowledge of network topology and route selection criteria, such as minimizing time or stress, or maximizing the aesthetic experience of a trip.

Route choice modeling is typically divided into a two-stage process. First, possible alternative routes are generated to form the choice set. Then the probability a given route is chosen from a specified choice set is calculated. These two procedures may correspond to non-compensatory and compensatory decision rules. The two-step methodology presented has the advantage that by explicitly specifying the set of available routes, we can examine possible selection criteria, and reduce computational time by not generating unrealistic routes. With a finite, known choice set, theoretically-based corrections for route overlapping can be applied.

This paper is organized as follows. First, several methods for generating unique alternative routes are described, and how these methods may be compared. Then a particular network database is described and a set of route generation algorithms using route choice data from Boston is examined. Properties of the final choice set are summarized. Next, the well-known Logit and Probit route choice models are discussed, and several recent models that have been proposed to overcome the overlapping problem are outlined. The paper selects two route choice models to illustrate the different model structures. Finally, some estimates of route choice model coefficients based on the Boston data are presented.

1. Choice set generation algorithms

In a roadway network, there may be numerous alternative routes. However, many of these possible routes may be overly circuitous, or otherwise unsuitable for a particular origin-destination pair. Since our modeling task is to predict route choice from among the routes that a particular traveler considers, we would like to identify all the routes that any traveler might consider. Specifically, we want to be able to identify algorithmic rules for generating the observed routes to avoid introducing biases in the estimation procedure, and to have useful algorithms for navigation systems. Such algorithms should be able to reflect drivers' knowledge of the transportation network and their perceptions of travel times and other network variables. Further, there is no benefit to enumerating routes that no traveler would consider. That is, computational effort is one criterion by which to evaluate potential path generation algorithms.

The effectiveness of different path generation techniques is defined in terms of the generated routes' *coverage* of the observed routes. Ideally, a generated route would match the observed route link-for-link; in this case, the algorithm has *replicated* the survey route. Other routes may not be replicated, so the distance that the generated route shares in common with the survey route is considered. Thus, the *overlap* typically is expressed as a percentage of the survey route distance. Finally, coverage is defined as the percentage of observations for which an algorithm or set of algorithms has generated a route that meets a particular threshold for overlap.

There are many dimensions in which a path generation algorithm may be designed. A well-known method, known as the *K*-shortest Path algorithm, generates the first "*k*" shortest paths for a given origin-destination pair. Two popular heuristics may be classified as *link penalty* and *link elimination* methods. Both techniques proceed iteratively after identifying a shortest path. In a link penalty heuristic (see for example De La Barra, Perez, and Anez, 1993), the impedance of all links on the shortest path is gradually increased. In a link elimination

technique, links on the shortest paths are removed from the network in sequence to generate new routes.

The *labeling* approach of Ben-Akiva et al. (1984) exploits the availability of multiple link attributes, such as travel time, distance and functional class to formulate different "generalized cost" functions that produce alternative routes. These routes may be labeled according to the criteria such as "minimize time," "minimize distance," "maximize use of expressways," etc., that yielded it.

Simulation methods produce alternative feasible paths by drawing impedances from different probability distributions. The distribution type (for example, Gaussian, Gumbel, Poisson), distribution parameters, number of draws and the seed of the pseudo-random number generator are design variables. In this paper, we used a Gaussian distribution with a mean and standard deviation calculated from travel times. (The choice of the Gaussian distribution was primarily for computational convenience, rather than for any theoretical reason.) Up to 48 draws were simulated for each observation, as this was estimated to take roughly the same computational time as the link elimination and link penalty algorithms.

2. The Boston data set

We have performed route generation experiments using a highway network database developed by Central Transportation Planning Staff (CTPS), the Metropolitan Planning Organization (MPO) for the Boston region. The highway network covers an area of approximately 2,800 square miles where about 4 million inhabitants reside. The network consists of over 800 zones, about 13,000 nodes, and about 34,000 one-way links.

Link attributes in the database include distance, free-flow time, estimated time (that is, the output of the CTPS traffic assignment model), capacity, number of lanes, tolls, assigned volume, functional class, presence of government-numbered signage (e.g., Interstate 93, U.S. Route 1, Massachusetts Route 16), and indicators of security such as neighborhood income and employment. With these attributes, it is possible to construct many different labels. Of course, many attributes will be correlated—such as distance, free-flow time and estimated time.

Route choice data come from a 1997 Transportation Survey of Faculty and Staff conducted by the MIT Planning Office. Drivers were asked to provide a written description of their habitual route. When route descriptions contained gaps, the least-distance path was used to connect known portions of the survey respondent's route. We omitted observations where the respondent made stops along the way, or did not provide enough information from which to construct a coherent route. A total of 188 respondents met the screening criteria and thus formed the origin-destination pairs on which the various route generation algorithms described above were performed.

3. Evaluation of choice set generation algorithms

Several variations of the four broad types of route generation algorithms described above were examined: labeling, link elimination, link penalty and simulation. Table 1 shows the coverage results of individual labels. That is, each algorithm generates exactly one route by minimizing a particular label. In the instances where a label has parameters, such as the trade-off between time and distance, the set of parameters producing the greatest coverage was used.

Table 1 Coverage of individual single-route generation algorithms for Boston

Algorithm description and parameters	Overlap required for coverage					
	100%		90%		80%	
1. Least time	64	34%	69	37%	84	45%
2. Least free-flow time	63	34%	70	37%	87	46%
3. Minimize *Generalized Cost* Minimize 0.4 ∗ Time + 0.4 ∗ Distance + 0.2 ∗ Toll	62	33%	67	36%	77	41%
4. Minimize V/C-weighted time Minimize CC Time + 0.8 Time 1($V/C = 0$) + Time 1($0 < V/C < 0.9$) + 0.9 Time 1($V/C \geq 0.9$)	61	32%	67	36%	81	43%
5. Minimize left turns path Double or triple left turn penalty	58	31%	66	35%	81	43%
6. Maximize capacity-weighted time path	55	29%	64	34%	74	39%
7. Maximize time in secure neighborhoods Weighted by median income	55	29%	60	32%	76	40%
8. Maximize high capacity roads path Min (High Cap + 2 Low Cap + CC) Time	45	24%	50	27%	65	35%
9. Turn-penalty hierarchy path (1.5 min for one level higher or lower)	42	22%	49	26%	63	34%
10. Maximize freeways path Minimize (Freeway + 2 Expressway + 4 Arterial + 4 Local + CC) Time	38	20%	46	24%	56	30%
11. Least distance	38	20%	42	22%	53	28%
12. Minimize number of links	33	18%	55	29%	57	30%
13. Maximize expressways path Minimize (2 Freeway + Expressway + 2 Arterial + 2 Local + CC) Time	33	18%	34	18%	43	23%
14. Maximize arterials path Minimize (4 Freeway + 2 Expressway + Arterial + Local + CC) Time	27	14%	27	14%	30	16%
15. Minimize tolls and turn penalties	18	10%	19	10%	28	15%
16. Minimize stop lights	15	8%	17	9%	26	14%
Combination of all above algorithms	136	72%	143	76%	160	85%

Notes: 188 observations total. Algorithms are sorted in descending order of coverage at the 100 percent overlap threshold.

From Table 1, we can note that no single label performs very well. Minimizing free-flow time produces the best results, and even then, less than one-half the respondents appear to choose a minimum-free-flow-time path. Even fewer appear to follow a minimum-distance path.

The analyst may be tempted to combine paths generated by multiple labels in hopes of achieving coverage equal to the sum of each label's coverage. Unfortunately, this result requires a very special case. Since link attributes are likely correlated, for some OD pairs, the least time path might be identical to the least generalized cost path; for another OD pair, the least free-flow time path might be identical to the Maximize Freeways path. Therefore, the coverage of multiple labels will be at least the maximum of the individual labels' coverage, and at most the sum of each label's coverage.

It can further be noted that combining the 16 algorithms presented in Table 1 does not produce a satisfying result, as 15 to 25 percent of observations do not have sufficient overlap

Table 2 Coverage of multiple-route generation algorithms for Boston

Algorithm description and parameters	Overlap required for coverage					
	100%		90%		80%	
Combination of all labeling algorithms (From Table 1)	136	72%	143	76%	160	85%
Combination of min. distance, free-flow time and time	74	39%	82	44%	97	52%
K-Shortest paths–link penalty 40 unique routes	102	57%	120	67%	143	80%
K-Shortest paths–link penalty 15 unique routes	101	56%	118	66%	139	78%
K-Shortest paths–link elimination	113	60%	119	63%	134	71%
Combination of all above algorithms	156	83%	164	87%	175	93%
Minimize simulated time 48 draws	94	50%	120	64%	148	79%
Minimize simulated time 32 draws	92	49%	115	61%	143	76%
Minimize simulated time 16 draws	82	44%	106	56%	133	71%
Minimize simulated time 8 draws	71	38%	95	51%	121	64%
Combination of all above algorithms	157	84%	165	88%	177	94%

Notes: 188 observations total. Algorithms are sorted by type, and then in descending order of coverage at the 100 percent overlap threshold.

with any of the generated routes, depending on the threshold chosen. Therefore, we examine algorithms that specifically generate multiple paths, such as the link elimination and link penalty *K*-Shortest Path heuristics, and simulation. Results of these algorithms are compared with labeling in Table 2.

The distributional parameters used for simulating travel times were calculated similarly to the parameters of generalized cost labels. We found good coverage results when we drew link travel times from a distribution having a standard deviation twice that of the mean.

Table 2 shows that the *K*-Shortest Path heuristics do increase coverage over labeling alone. As expected, the simulation approach shows diminishing returns with respect to the number of draws. At 48 draws, simulation provides better coverage than the three labels that require no parameters: distance, free-flow time and estimated time. However, simulation does not do better than any individual *K*-Shortest Path heuristics, or the labeling approach with all 16 labels.

In evaluating route choice generation algorithms, we also need to consider computational performance. An algorithm that yields a five percent increase in the number of observations covered may not be cost-effective if it takes months to run, for example. The results of computational time experiments are shown in Table 3 below. Minimizing one label is the fastest, as this simply requires a call to the built-in shortest-path routine. Minimizing a

Table 3 Computational time of alternative algorithms

Algorithm description	Time for 1 OD pair	Time for 188 OD pairs
Minimize one label	32 s*	1 h 40 min
Minimize a random draw	35 s*	1 h 50 min
Minimize 48 random draws	3 min 20 s*	10 h 30 min
Link elimination (DynaMIT)	7 min	22 h*
Link penalty (De la Barra) for 15 unique routes	25 min	3 d 6 h*
Link penalty (De la Barra) for 40 unique routes	1 h 40 min	13 d*

Notes: * Indicates a calculated quantity. Computational experiments were conducted using TransCAD 3.1 on a 400 MHz Pentium II workstation with 256 MB RAM running Windows NT 4.0.

random draw is almost as fast; time must be allowed to make the draws of random travel time before constructing the shortest paths. The link elimination and link penalty heuristics, which involve multiple shortest-path calls, take successively longer.

All algorithms were run with conventional transportation software (TransCAD), which is a GIS platform. Therefore, the results shown may be affected by the GIS file structure. The link penalty approach seems to perform particularly poorly because updating the costs on a few links requires re-writing the whole network database. In comparison, the link elimination heuristic can be fairly efficiently implemented—a "link in use" bit can be turned on or off.

The long computational times of the link penalty approach disqualified it from further consideration. We also had reservations about the realism of paths generated by the link elimination approach. Since we eliminated only one link at a time, it was feared the other generated paths would closely resemble the original shortest path, with the exception of a brief deviation. We were pleased with the computational time of the simulation algorithm, and its ease of implementation. By considering both coverage and computational time, we decided to use simulation with 48 draws and labeling with the three parameter-free objective functions for our "final" choice set generation. Other labels produced paths similar to those from minimizing distance, free-flow time or estimated time.

Practitioners may be concerned that the simulation algorithm we selected may still be to slow for realistic applications. Fortunately, path generation needs only be run once for each estimation data set; paths can be stored, making it much quicker to calculate path attributes from "new" link variables. Similarly, for forecasting, it is only necessary to generate paths once per network configuration. When considering highway construction alternatives, path sets for build alternatives can be created efficiently by generating paths that are encouraged (through strategic choice of label or link attribute values) or constrained to use the new links, and combining this new set of paths with the base case path set.

The choice set generation procedure selected for route choice estimation generated up to 51 alternative routes: from three deterministic labels and 48 random draws. Some origin-destination pairs would have fewer alternatives available, as some labels or draws might yield duplicate paths. The distribution of the number of unique paths in the choice sets of the 188 auto users in our sample is shown in Figure 1. It can be seen that the median size of the generated choice set is about 30 routes, and that about one-quarter of the observations have a choice set with 40 or more feasible routes.

Also notice that some observations have a very small choice set. These observations generally correspond to employees who live close to the MIT campus. The density of streets in the network is such that these people have few reasonable alternative routes to MIT. That is, the nearest parallel facility may be quite far from the best route, when considered in relation to the total distance between origin and destination. After eliminating observations with only a single route in the choice set, 159 observations were selected to compose the estimation data set.

Figure 2 shows the distribution of the number of links in the choice sets of the 159 respondents used to estimate route choice models. The number of links is an important statistic because it determines the complexity of some of the route choice models, as explained further in this paper. The observation using the greatest number of links has 856 links in its choice set. The smallest choice set among the 159 respondents has 19 links.

The large variation in the number of links is heavily dependent on the distance between the origin and destination. All observations have a common destination (MIT), but a variety

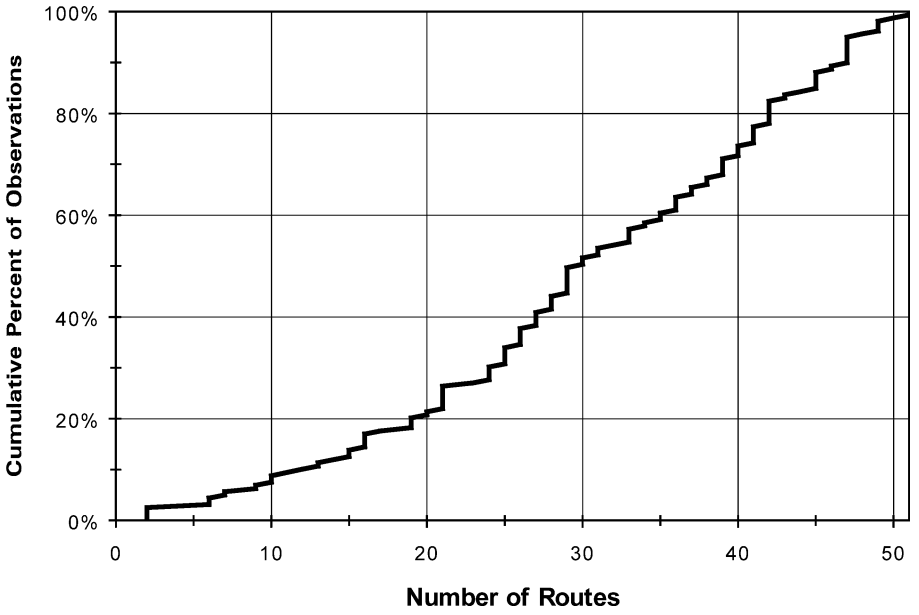

Fig. 1 Cumulative distribution of choice set size

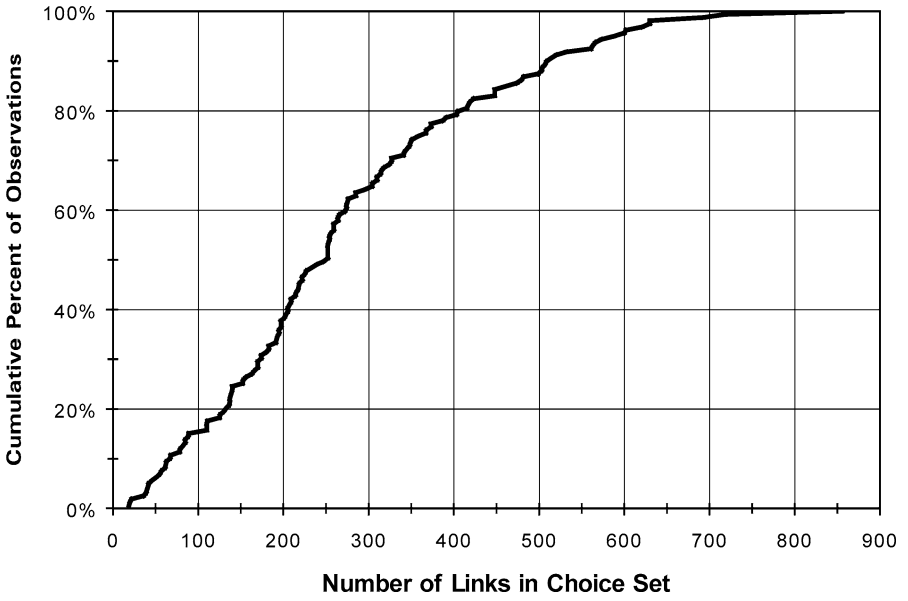

Fig. 2 Distribution of links in choice sets

of origins. Thus, a more distant origin will have a larger number of links in its choice set for two reasons: (1) because it is more distant, a driver will need to traverse more links to reach his or her destination; and (2) because of the greater distance between origin and destination, the driver will likely have more alternative paths available.

4. Route choice models

The deterministic shortest path is the simplest route choice model, which is used in deterministic traffic assignment models. The multinomial Logit (MNL) and Probit models were proposed long ago as generalizations to the deterministic model. Probit is based on the normal (or Gaussian) distribution, and thus requires simulation. In comparison, MNL is based on the Gumbel distribution and has a well-known analytical form. However, the MNL model is not suitable to model route choice, because it cannot account for similarities among routes.

Several types of models have been recently proposed to overcome the MNL drawbacks. These models represent modifications or generalizations of the Logit structure. *C*-Logit, proposed by Cascetta et al. (1996) and Path-Size Logit (PSL) presented in Ben-Akiva and Bierlaire (1999) may be considered modifications to the MNL model, as they add a correction term to path utilities but maintain the MNL model structure. Thus, they can be estimated using existing Logit software.

The Cross-Nested Logit (CNL) model of Vovsha (1997) and the Paired Combinatorial Logit (PCL) model of Chu (1989) were adapted for route choice situation in Prashker and Bekhor (1998). Gliebe, Koppelman, and Ziliaskopoulos (1999) also adapted the PCL model for route choice. These models have a more general (and therefore more complex) error structure. These models are members of the Generalized Extreme Value (GEV) family of models developed by McFadden (1978), which also includes MNL and Nested Logit.

Some researchers have examined the suitability of the Probit model for route choice. Because the Probit model is based on error terms having a multivariate normal distribution—as opposed to a Type I Extreme Value distribution as assumed in MNL and other GEV models—an arbitrary covariance structure may be specified. Daganzo and Sheffi (1977) were the firsts to use the Multinomial Probit (MNP) to model route choice.

Yai, Iwakura, and Morichi (1997) provide a recent example of an application of the MNP route choice model in Japan. The authors assume the covariance of route utilities is proportional to overlap length. Routes are also assumed to have heteroskedastic error terms where variance is proportional to route length or impedance.

Choice models with combinations of Gaussian and Type I Extreme Value error terms have been proposed by researchers such as McFadden and Train (1998), who call the resulting model Mixed Logit, and by Ben-Akiva and Bolduc (1996), who refer to the resulting system as Multinomial Probit with Logit Kernel, or simply Logit Kernel. These models are specified so that if cross-alternative correlations are estimated to be zero, the model reduces to MNL. Bekhor, Ben-Akiva, and Ramming (2002) presented an adaptation of the Logit Kernel model to route choice.

The following sections present in brief two route choice models: the Path-Size Logit Model and the Cross-Nested Logit model. These two models were purposely selected to exemplify the differences in the model structures. Furthermore, estimation results of these models will be compared against the simple MNL model. For a full review on the different route choice models, see Ramming (2001).

5. The path-size logit model

The Path-Size Logit model is similar to C-Logit in that a correction term is added to a path's utility. However, PS-Logit has a different theoretical basis. The notion of "size" comes from the theory of aggregate alternatives, which was first employed for destination and residence

choice. However, unlike destination choice, where zones may have a size representing thousands of *elemental* destinations (e.g., workplaces), the largest size a path may have is one. The log of the path size is added to the path utility to form the Path-Size Logit model from MNL:

$$P(i|C_n) = \frac{e^{V_{in}+\ln PS_{in}}}{\sum_{j\in C_n} e^{V_{jn}+\ln PS_{jn}}} = \frac{PS_{in}\, e^{V_{in}}}{\sum_{j\in C_n} PS_{jn}\, e^{V_{jn}}} \tag{1}$$

Where: PS_{in} is the size of path i for person n; V_{in} is the utility of path i for person n; C_n is the path-set for person n.

A path with no overlapping links needs no utility adjustment and has a size of one. The extreme case of two paths being created by "duplicating" or "splitting an existing path down the middle" results in each having a size of one-half. Path sizes may be calculated based on the length of links within a path, and the relative lengths of paths that share a link. Therefore, the calculation of path sizes is dependent on the specification of the choice set. We propose the following definition for the Path-Size as follows:

$$PS_{in} = \sum_{a\in\Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j\in C_n} \frac{L_i^\gamma}{L_j^\gamma}\delta_{aj}} \tag{2}$$

Where: a indexes links (arc); Γ_i is the set of links in path i; l_a is the length of link a; L_i is the length of path i, so $L_i = \sum_{a\in\Gamma_i} l_a$; $\delta_{aj} = 1$ if path j uses link a and 0 otherwise; γ is a parameter to be calibrated.

Note that when $\gamma = 1$ is similar to the Ben-Akiva and Bierlaire (1999) definition:

$$PS_{in} = \sum_{a\in\Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j\in C_n} \frac{L^*_{C_n}}{L_j}\delta_{aj}} \tag{3}$$

Where: $L^*_{C_n} = \min_{k\in C_n} L_k$, that is, the length of the shortest path in C_n.

6. The cross-nested logit model

The Cross-Nested Logit model is a member of the broad Generalized Extreme Value (GEV) model. Assumptions and properties of the GEV model are discussed in Ben-Akiva and Lerman (1985), and are beyond the scope of this discussion. The Cross-Nested Logit model differs from the well-known Nested Logit model in that lower-level alternatives may belong to more than one nest. That is, we define a set of parameters α_{mi} for each alternative i and each nest m ($0 < \alpha_{mi} < 1$), which represents the degree of "membership" or the inclusion weight of alternative i in nest m. The sum of α_{mi} over all nests is generally normalized to one for each lower-level alternative, i. The choice probabilities of the Cross-Nested Logit model are as follows:

$$P_i = \frac{e^{V_i} \sum_m \alpha_{im} \left(\sum_i \alpha_{im}\, e^{V_i}\right)^{\mu-1}}{\sum_m \left(\sum_i \alpha_{im}\, e^{V_i}\right)^{\mu}} \tag{4}$$

Where μ is the nesting coefficient. As with the Nested Logit model, when the nesting coefficient is equal to 1, the model collapses to the simple MNL model. It is possible to rewrite the expression for the choice probability as follows:

$$P_i = \sum_m P_{i|m} \cdot P_m \tag{5}$$

Where the conditional probability of a route i being chosen in link (nest) m is:

$$P_{i|m} = \frac{\alpha_{im}\, e^{V_i}}{\sum_i \alpha_{im}\, e^{V_i}} \tag{6}$$

And the marginal probability of a nest m being chosen is:

$$P_m = \frac{\left(\sum_i \alpha_{im}\, e^{V_i}\right)^{\mu}}{\sum_m \left(\sum_i \alpha_{im}\, e^{V_i}\right)^{\mu}} \tag{7}$$

The CNL model is adapted to route choice situation by suitably defining the inclusion coefficients as dependent on network topology as follows:

$$\alpha_{mi} = \left(\frac{l_m}{L_i}\right)^{\gamma} \delta_{mi} \tag{8}$$

Where γ is parameter to be calibrated. In this paper, we assumed $\gamma = 1$ for convenience. Note that the dimension of this parameter may be quite large for real size networks. Recall that in the data set used in this paper, the dimension of this parameter is 856 (maximum number of links for a single observation) times 51 (maximum number of alternatives).

Vovsha and Bekhor (1998) and Papola (2000) estimated CNL models using constant values of μ. Prashker and Bekhor (1998), Papola (2000), and Wen and Koppelman (2001) cite difficulties of making this assumption. Nest-specific μ_m may be estimated if there is sufficient data for their identification. Bekhor and Prashker (2001) proposed the following formulation based on path topology:

$$\mu_m = 1 - \frac{\sum_i \alpha_{mi}}{\sum_i \delta_{mi}} \tag{9}$$

7. Estimation results

Table 4 below presents the best estimates obtained for 4 route choice models: MNL, PSL, and two CNL models. Parameter estimates are shown in bold, followed by t-statistics. T-statistics are for the hypothesis of a zero parameter value. For the ln(Path Size) terms, an additional t-statistic is calculated for the null hypothesis that the coefficient equals one.

Explanatory variables include not only well-known variables such as travel time and distance, but also variables related to network knowledge, such as time spent on government-numbered routes. Further insight on the explanatory variables can be found in Bekhor, Ben-Akiva, and Ramming (2002) and Ramming (2001). In this paper, we skip the discussion on the explanatory variables for brevity. Note that all utility parameter estimates and their standard errors are in the same order of magnitude. Therefore, we focus on measures of model fit, which help us evaluate the usefulness of the Path Size and Cross-Nested specifications.

Table 4 Estimation results

Variable	MNL	PSL	CNL	CNL
1. Distance (miles)	**−0.253**	**−0.212**	**−0.252**	**−0.224**
	−2.4	−2.1	−2.5	−2.3
2. Free-flow time (minutes)	**−0.601**	**−0.513**	**−0.553**	**−0.474**
	−6.6	−6.3	−6.5	−6.0
3. Path uses mass. pike (dummy)	**−0.640**	**−0.490**	**−0.530**	**−0.370**
	−0.9	−0.8	−0.8	−0.6
4. Path uses tobin bridge (dummy)	**2.90**	**2.75**	**2.79**	**2.75**
	3.1	3.1	3.1	3.2
5. Path uses sumner tunnel (dummy)	**2.18**	**1.92**	**2.06**	**1.92**
	1.8	1.7	1.8	1.7
6. Ln(delay) for no income reported	**−5.13**	**−4.45**	**−4.80**	**−4.26**
	−2.6	−2.5	−2.6	−2.6
7. Ln(delay) for income < $100,000 per year	**−0.205**	**−0.583**	**−0.191**	**−0.506**
	−0.5	−1.4	−0.4	−1.2
8. Ln(delay) for income >= $100,000 per year	**−2.562**	**−2.676**	**−2.542**	**−2.624**
	−2.7	−3.0	−2.8	−3.0
9. Time spent on government-numbered route	**0.112**	**0.090**	**0.098**	**0.078**
	3.5	2.9	3.1	2.6
10. Path with least distance label (dummy)	**1.056**	**0.759**	**0.987**	**0.728**
	4.2	3.0	4.0	3.0
11. Path with least estimated time (dummy)	**0.971**	**0.377**	**0.881**	**0.382**
	4.3	1.5	4.0	1.6
12. Ln(path size) based on FF time, $\gamma = \infty$		**0.730**		**0.617**
t-statistic w/r/t 0		6.0		5.2
t-statistic w/r/t 1		−2.2		−3.2
Number of observations	159	159	159	159
Initial log-likelihood	−519.7	−519.7	−519.7	−519.7
Final log-likelihood	−410.8	−393.1	−404.1	−390.6
Number of parameters	11	12	11	12
Rho-Bar squared	0.188	0.221	0.201	0.225

Notice that the model with only the Path Size term outperforms the model with only the Cross-Nested Logit structure. (Using Horowitz's, 1983, non-nested hypothesis test, the probability that CNL is the correct model given the results in Table 4 is about 0.15 thousandths of a percent!) This is interesting for several reasons. First, the PSL model is more easily estimated than the CNL model; with PSL, the Path Size term can be calculated after the path enumeration step and used in standard MNL estimation software. CNL estimation requires specialized code. Commensurate with its complexity, CNL takes longer computational time to estimate. Theoretically, the Path Size term may be thought of as an approximation to CNL as suggested by the model structures presented in the paper.

Why then does the PSL model out perform the CNL model? One reason may be that the PSL term has been calibrated with a value of $\gamma =$ infinity. The CNL specifications estimated in Table 4 above may be more similar to a Path Size term with $\gamma = 1$. The log-likelihood for the PSL model $\gamma = 1$ (not presented in the table) is −409.9. Therefore, CNL is an improvement over PSL with a corresponding specification. Also note that CNL has a better fit than the MNL model, which has a log-likelihood of −410.8.

Also note that the models with both CNL and Path Size specifications out-perform both "pure" CNL and PSL. (The t-statistic of the Path Size coefficient provides a nested hypothesis

test comparing the CNL model without Path Size to the one with Path Size. Horowitz's non-nested hypothesis test reveals a 2.1 percent probability the PSL-only model should be preferred to the CNL with Path Size specification.) This observation may be related to the need to further calibrate the CNL models.

8. Conclusions

This paper focused on the problem of estimating a route choice model for a large network. The approach of the paper was to first generate a choice set and use this choice set to estimate the model parameters.

The choice set generation method proposed falls in the class of deterministic methods. The advantage of such method is that can be applied for any urban network with existent resources, since it is based on successive shortest path calculations using travel time and distance variables.

Several route choice models were proposed to overcome the MNL drawbacks. The Path-Size Logit model was proposed to model route choice since it can capture overlapping among routes, and it can be estimated using conventional software. The model estimation included also network knowledge variables, in addition to standard travel time and distance variables. The initial results presented in this paper suggest that route choice models may be estimated for large urban networks at relatively modest computing resources.

References

Bekhor, S., Ben-Akiva, and M.S. Ramming (2002). "Adaptation of Logit Kernel to Route Choice Situation." *Transportation Research Record*, 1805, 78–85.

Bekhor, S. and J. Prashker (2001). "Stochastic User Equilibrium Formulation for the Generalized Nested Logit Model." *Transportation Research Record*, 1752, 84–90.

Ben-Akiva, M., M.J. Bergman, A.J. Daly, and R. Ramaswamy (1984). "Modelling Inter Urban Route Choice Behaviour." In J. Volmuller and R. Hamerslag (eds.), *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, VNU Press, Utrecht, pp. 299–330.

Ben-Akiva, M. and M. Bierlaire (1999). "Discrete Choice Methods and Their Applications to Short Term Travel Decisions." In R.W. Hall (ed.), *Handbook of Transportation Science*.

Ben-Akiva, M. and D. Bolduc (1996). "Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure." Working Paper, 1996.

Ben-Akiva, M. and S.R. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, MA: MIT Press.

Cascetta, E., A. Nuzzolo, F. Russo, and A. Vitetta (1996). "A Modified Logit Route Choice Model Overcoming Path Overlapping Problems: Specification and Some Calibration Results for Interurban Networks." In J.B. Lesort (ed.), *Transportation and Traffic Theory. Proceedings from the Thirteenth International Symposium on Transportation and Traffic Theory*, Lyon, France, Pergamon pp. 697–711.

Chu, C. (1989). "A Paired Combinatorial Logit Model for Travel Demand Analysis." In *Proceedings of the 5th World Conference on Transportation Research*, 4, Ventura, CA, pp. 295–309.

Daganzo, C.F. and Y. Sheffi (1977). "On Stochastic Models of Traffic Assignment." *Transportation Science*, 11, 253–274.

De la Barra, T., B. Perez, and J. Anez (1993). "Multidimensional Path Search and Assignment." In *Proceedings of the 21st PTRC Summer Meeting*, pp. 307–319.

Gliebe, J.P., F.S. Koppelman, and A. Ziliaskopoulos (1999). Route Choice Using a Paired Combinatorial Logit Model, presented at the 78th TRB Meeting, Washington D.C.

Horowitz, J.L. (1983). "Statistical Comparison of Non-Nested Probabilistic Discrete Choice Models." *Transportation Science*, 17, 319–350.

McFadden, D. (1978). "Modeling the Choice of Residential Location." In A. Karlqvist et al. (eds.), *Spatial Interaction Theory and Residential Location*, North Holland, Amsterdam pp. 75–96.

McFadden, D. and K. Train (2000). "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics*, 15(5), 447–470.
Papola A. (2000). "Some Development of the Cross-Nested Logit Model." In *Proceedings of the 9th IATBR Conference*. Gold Coast, Australia.
Prashker, J.N. and S. Bekhor (1998). "Investigation of Stochastic Network Loading Procedures." *Transportation Research Record*, 1645, 94–102.
Ramming, M.S. (2001). "Network Knowledge and Route Choice." Unpublished Ph.D. Thesis, Massachusetts Institute of Technology.
Vovsha, P. (1997). "The Cross-Nested Logit Model: Application to Mode Choice in the Tel-Aviv Metropolitan Area." *Transportation Research Record*, 1607, 6–15.
Wen, C. and F. Koppelman (2001). "The Generalized Nested Logit Model." *Transportation Research Part B: Methodological*, 35(7), 627–641.
Yai, T., S. Iwakura, and S. Morichi (1997). "Multinomial Probit with Structured Covariance for Route Choice Behavior." *Transportation Research Part B*, 31, 195–207.