XINJUN LAI, Ph.D.[1,2]
E-mail: xinjun.lai@gdut.edu.cn [1], laixinj@mail2.sysu.edu.cn [2]
JUN LI, Ph.D.[2]
E-mail: stslijun@mail.sysu.edu.cn
ZHI LI, Ph.D.[1]
E-mail: lizhi_piers@gdut.edu.cn
[1] School of Electro-Mechanical Engineering,
Guangdong Provincial Key Lab of Computer Integrated Manufacturing,
Guangdong University of Technology
Guangzhou 510006, China
[2] School of Engineering, Sun Yat-sen University
Guangzhou 510006, China

Traffic in the Cities
Preliminary Communication
Submitted: Mar. 30, 2015
Accepted: Feb. 10, 2016

A SUBPATH-BASED LOGIT MODEL TO CAPTURE THE CORRELATION OF ROUTES

ABSTRACT

A subpath-based methodology is proposed to capture the travellers' route choice behaviours and their perceptual correlation of routes, because the original link-based style may not be suitable in application: (1) travellers do not process road network information and construct the chosen route by a link-by-link style; (2) observations from questionnaires and GPS data, however, are not always link-specific. Subpaths are defined as important portions of the route, such as major roads and landmarks. The cross-nested Logit (CNL) structure is used for its tractable closed-form and its capability to explicitly capture the routes correlation. Nests represent subpaths other than links so that the number of nests is significantly reduced. Moreover, the proposed method simplifies the original link-based CNL model; therefore, it alleviates the estimation and computation difficulties. The estimation and forecast validation with real data are presented, and the results suggest that the new method is practical.

KEY WORDS

cross-nested Logit; stochastic route choice; subpath; correlation;

1. INTRODUCTION

Route choice models capture the travellers' choosing behaviours in order to reproduce their choosing preferences and probabilities, and hence to compute traffic flows on each road for the demand forecast and management. Discrete choice models, especially the Logit-based models are the most frequently used for its closed-form structure and its simplicity in estimation and prediction [1]. The utilities of alternative routes are assumed to be stochastic, and the error terms are included into the utility function to represent the uncertainty or the unobserved stochastic attributes. In particular, the assumption that the errors are independently identically extreme value distributed (IID), lead to the multinomial Logit (MNL) model with a closed form. Generally, Logit requires less computational time than the multinomial Probit model, which assumes that error terms are normally distributed but lead to a non-closed-form expression; therefore, its estimation and computation require simulation-based methods.

1.1 Literature review

The IID assumption of Logit, however, leads to a major drawback which is that it cannot represent the correlation of alternatives, and therefore leads to inaccurate results when routes overlap. Advanced methods are proposed to address this issue, such as Path Size Logit (PSL) [1], CNL [2], mixed Logit (also called error component) [3],etc. Prato [4] provided a literature review on the most frequently used methodologies in route choice modelling. Recently, Fosgerau et al. [5] proposed a recursive Logit model that does not require the enumeration of paths. Papola and Marzano [6] proposed a joint network generalized extreme value (GEV) approach to model the route choice that does not require choice set generation as well. Given that the choice set is sampled beforehand, Frejinger et al. [7] used the sampling correction term for the MNL model, and Lai and Bierlaire [8] studied the sampling issues in GEV route choice model.

Data used for the mentioned route choice models are frequently collected either from questionnaires or from GPS devices. With the former approach, the respondents are required to report the itineraries from the origin to destination (OD) [9]. However the respondents do not always report with a link-by-link style; instead, they just recall the major roads or landmarks

they passed, but not necessarily the minor streets or smaller intersections, such as "home - TV tower – Sunset Av. - workplace", instead of "home – 1st alley – 34th street - TV tower – 52nd street – Sunset Av. – 5th alley - workplace". Most importantly, travellers do not plan their trips by a link-specific style. Recently, the GPS technology has become the mainstream to collect the route choice data [10]; however, errors in route reconstructions are inevitable when processing GPS points, especially in the urban area with skyscrapers where the signals are disturbed. Besides, the same problem occurs to the studies with floating car data [11-13] and smartphone data [10] where they are even sparser. Regarding this issue, Frejinger and Bierlaire [14] proposed a network-free approach to link this gap, and the domains of data relevance (DDR) is used to determine the relevance of the collected data points to the physical network. A probabilistic map-matching algorithm is proposed based on the DDR method and several candidate re-constructed paths are given likelihoods of being the true one [10]. However, it still raises a concern whether the link-based route choice models are suitable with these data collecting and processing approaches. In particular, the information of minor details is not essentially useful. From the perspective of travellers, they might not consider the correlation of alternatives in links; on the other hand, from the view of analysts, considering and processing link-based details might be time-consuming.

The idea of modelling travel behaviours with simplified network has been firstly used in traffic assignment to reduce the computational burden [15-17]. Frejinger and Bierlaire [18] used the subnetwork idea in route choice analysis to capture the perceptual correlation of routes in travellers, and the idea is used by an error component (EC) model with a specification of Path Size term in the utility. The subnetworks can be motorways and main roads in the network hierarchy or by the most frequently used names when people describe itineraries. The "subnetwork" they used is actually a portion of a path, but not a network; therefore, there is no grid or loop in the "subnetwork". as the notation "subpath" has been preferred in this paper. Their research explores and provides more insights in interpreting travellers' behaviours in an abstract level of the road network. However, the employed error component (EC) model is based on the Probit mixture and thus based on simulation-based estimation and computation, which are highly time-consuming. Kazagli and Bierlaire [19] used the mental representation item to capture the travellers route choice behaviours, thus simplifying the modelling. Li et al. [20] provide the equivalent impedance idea to simplify road networks into one virtual link, similar to resisters simplification in an electric circuit. Papola and Marzano [6] used link aggregation method in a network GEV structure and thus simplified the network analysis. These studies all provide methodologies and prospects in simplifying the route choice and network analysis.

1.2 Motivation and paper structure

The aim of the paper is to explore the usage of simplified network in route choice modelling. The merits of the subpath idea can thus reduce the complexity in models and provide more behavioural information on travellers' perception in the route choice context. The cross-nested Logit (CNL) [21] model was used because of its closed form, which is valuable and supposed to be more time-saving than the EC model that was used by Frejinger and Bierlaire [18]. Moreover, Fosgerau et al. [16] have shown that any random utility model can be approached as close as needed by a CNL model. The original CNL in route choice with a link-based style [2] is not always flexible in application, because estimating a large set of nesting parameters is time-consuming, and it is difficult to successfully and significantly estimate all of them. Therefore, empirical approximation is often required. However, with a loss in behavioural interpretation [4]. The proposed subpath-based CNL model aims at capturing route choice behaviours by considering the travellers' conceptual correlation of routes. The new model simplifies the structure of the link-based CNL model and thus reduces its difficulties in estimation and computation. In particular, we are able to utilize the non-link-specific data that can be easily obtained, such as questionnaires and sparse GPS points, etc.

The rest of the paper is organized as follows: Section 2 provides the methodology of the subpath-based CNL model, with an illustrated example. In Section 3 the proposed method is then applied with the sparse GPS data in the urban area, estimation and forecast validation results are provided. Finally, Section 4 provides a conclusion.

2. METHODOLOGY

2.1 A subpath-based cross-nested Logit model

A subpath is defined as a portion of a route, and it can be a major road, a landmark or the most frequently used location name from the surveys, e.g. "TV tower" or "Sunset Avenue". In particular, when the raw and sparse GPS data in the urban area are processed, it is not necessary to accurately match the data to a specific link; instead, matching it to a subpath is sufficient. Routes that share a same subpath item may not physically overlap, but they might seem to be partially identical for the unobserved attribute, the correlated subpath. In order to capture such features we proposed the subpath-based method and defined that routes that share a same subpath are correlated. The cross-nested Logit (CNL) model is adopted to

explicitly capture this correlation, and each subpath is represented by a nest. Each alternative route uses one subpath item that belongs to that nest, as shown in *Figure 1*, and the correlation of alternatives are captured by this structure. It can be interpreted as a special case of the link-based CNL model where each link in the network is a subpath. This allows for simplicity and flexibility of modelling because it greatly reduces the computation time as the number of nests decreases; besides, the estimation would be easier since the number of estimated scale parameters is much smaller. In particular, it enables to capture the perceptual correlation between the paths, even though they might not physically overlap, e.g. two paths that partly use the Sunset Av. might share unobserved attributes even though they do not overlap at all.

The CNL model is used for its flexibility in explicitly modelling the correlation and its tractability in model solving because of its closed-form structure. The proposed method is a two-level CNL model; the upper-level represents the subpaths, and the lower-level represents the alternative routes. The subpaths M are mutually exclusive subsets, which are called subsets of the study network, and they are denoted by $\beta_1\{E_1,K_1\}$, ..., $\beta_M\{E_M,K_M\}$, where E_m is the set of links in subpath S_m, and K_m is the set of nodes. Besides, two subpaths should not have correlated links, which is

$$S_m \cap S_g = \{\emptyset, K_{mg}\}, \forall m \neq g \tag{1}$$

where K_{mg} is the intersection set of K_m and K_g.

Let C be the full choice set of paths from origin O to destination D, and the utility of alternative path i is

$$U_i = \beta_i x_i + \varepsilon_i, \forall i \in C \tag{2}$$

where x_i is a vector of attributes, β_i is the corresponding parameter vector to be estimated, the term $\beta_i x_i$ can be interpreted as the deterministic part of the utility, named V_i; ε_i is the error part of utility. The distribution of the error and the correlation of alternatives in a CNL model are discussed by Abbe et al. [21].

Each path that uses subpath S_m belongs to the nest m, and each route at least belongs to one subpath. The probability that a path i is chosen from choice set C can be interpreted as

$$\Pr(i|C) = \sum_{m=1}^{M} \Pr(S_m|C_s)\Pr(i|S_m) \tag{3}$$

where

$$\Pr(i|S_m) = \frac{\alpha_{im}^{\mu_m/\mu} e^{\mu_m V_i}}{\sum_{j \in C} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_j}} \tag{4}$$

is the conditional choice probability of path i given that subpath m is chosen, and

$$\Pr(S_m|C_S) = \frac{\left(\sum_{j \in C} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_j}\right)^{\mu/\mu_m}}{\sum_{p=1}^{M} \left(\sum_{j \in C} \alpha_{jp}^{\mu_p/\mu} e^{\mu_p V_j}\right)^{\mu/\mu_p}} \tag{5}$$

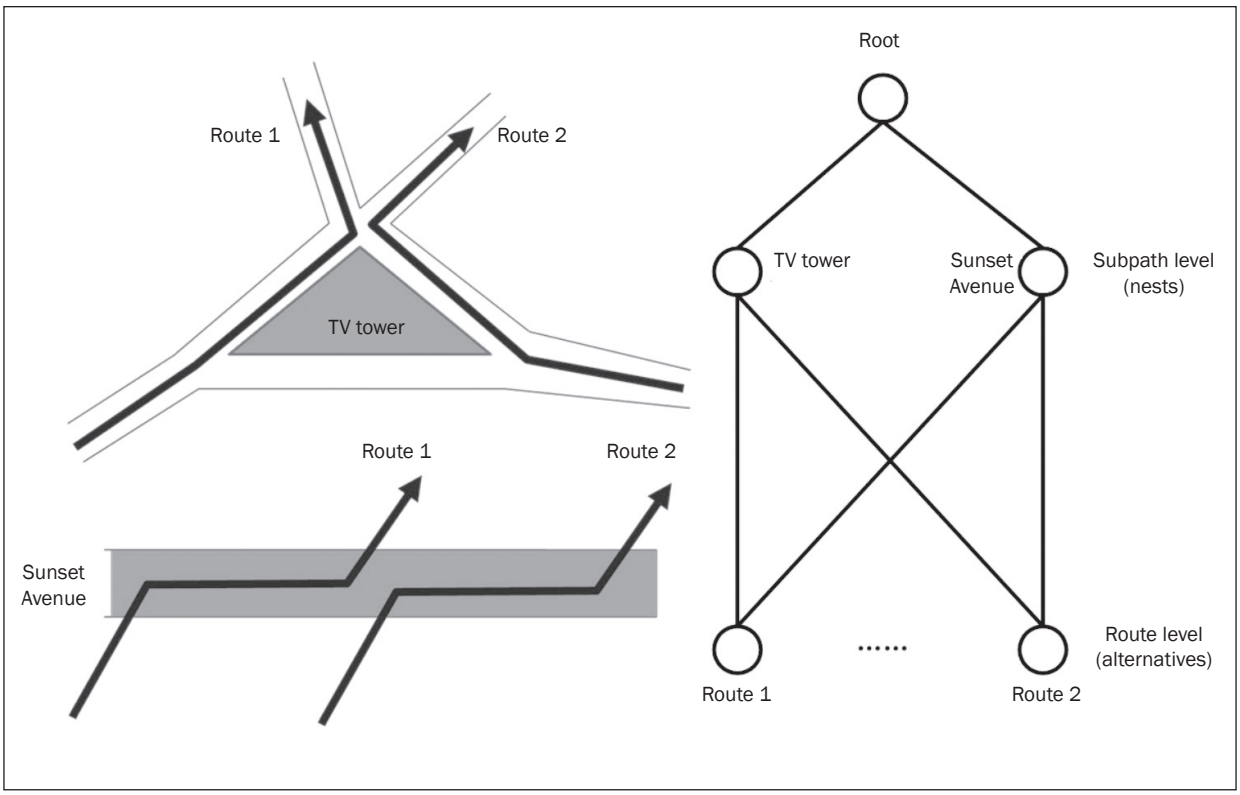

Figure 1 – Routes 1 and 2 are not physically correlated but share the same subpaths, and the correlation is captured by a two-level CNL structure

is the marginal probability that subpath m is chosen; C_S is the set of subpaths; μ is the root scale parameter and usually normalized to 1; α_{im} is the inclusive parameter to capture the level of membership of alternative i in subpath m, defined as $\alpha_{im} \propto l_m/L_i$ and

$$\sum_{m=1}^{M} \alpha_{im} = 1, \forall i \in C \tag{6}$$

where l_m is the length of path i in subspath m, and L_i is the length of path i; μ_m is the nesting scale parameter of subpath m. Besides, this condition $0 \le \mu \le \mu_m$ holds for all m.

The proposed model can be rewritten as an MNL-like structure with the theories of the multivariate extreme value (MEV) models [22], as shown in *Equation 7*

$$\Pr(i|C) = \frac{\exp[\mu V_i + \ln G_i]}{\sum_{j \in C} \exp[\mu V_j + \ln G_j]} \tag{7}$$

with the choice probability generating function, the G function, as

$$G(e^{Vi}) = \sum_{m=1}^{M} \left[\sum_{i=1}^{|C|} \alpha_{im}^{\mu m/\mu} e^{\mu Vi} \right]^{\mu/\mu_m} \tag{8}$$

Therefore, the logarithm of the derivative of (8) is

$$\ln G_i = \ln \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left[\sum_{j \in C} \alpha_{jm} e^{\mu_m V_j} \right]^{\frac{\mu - \mu_m}{\mu_m}} \right] \tag{9}$$

It should be noted that if μ_m=v, it means there is no correlation between alternatives. The G function (8) takes the form

$$G(e^{V_i}) = \sum_{i=1}^{|C|} e^{\mu V_i} \tag{10}$$

and the model collapses to an MNL model. Besides, the proposed method collapses into a link-based CNL model when each subpath is one link, then $\ln G_i$ is

$$\ln Gi = \ln \sum_{q \in Q} \left[\mu \alpha_{iq}^{l} e^{V_i(\mu_q - 1)} \left[\sum_{j \in C} \alpha_{jq}^{l} e^{\mu_q V_j} \right]^{\frac{\mu - \mu_q}{\mu_q}} \right] \tag{11}$$

where Q is the set of links, the subscript q represents link $q, \forall q \in Q$, α_{iq}^{l} is the inclusive parameter that $\alpha_{iq}^{l} = l_q/L_i$.

2.2 Illustrated example

A demonstrated network shown in *Figure 2* is presented to illustrate the proposed model. It is a grid network from origin O to destination H. The lengths of links are marked above arrows. Three major roads in the network are shown in bold arrows, a+b, c+d and e+f, and they are chosen as the subpaths, named as S_1, S_2 and S_3. Other links which are drawn in dashed arrows are minor streets. Five shortest paths are selected as the considerable choice set for all the travellers and shown in *Figure 2*. Each of them uses at least one subpath.

The relations of five paths and three subpaths can be represented as the structure in *Figure 3*. Therefore, with the proposed method the probability of choosing path i is as *Equation 7*, where

$$\ln Gi = \ln \sum_{m=1}^{3} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left[\sum_{j \in C} \alpha_{jm} e^{\mu_m V_j} \right]^{\frac{\mu - \mu_m}{\mu_m}} \right] \tag{12}$$

C={1,2,3,4,5}; $\alpha_{1\text{-}1}$=1, $\alpha_{2\text{-}1}$=1/2.2, $\alpha_{2\text{-}2}$=1.2/2.2, $\alpha_{3\text{-}2}$=1, $\alpha_{4\text{-}3}$=1, $\alpha_{5\text{-}2}$=1/2 and $\alpha_{5\text{-}3}$=1/2; μ_1, μ_2 and μ_3 are parameters that should be estimated.

If each link in the network is one subpath, then the model is a link-based CNL model, and its structure is shown in *Figure 4*. The choice model is (7) and (11), where $|Q|$=12, and the number of estimated nesting scale parameters μ_q is 12. We can see that the proposed method increases the simplicity of the model, and it also reduces the estimation difficulties because the number of nesting scale parameters is much lower. Moreover, it should be noted that not all parameters in a link-based CNL model, shown in *Figure 4*, can be successfully estimated because the structure is too complex. As a consequence, it is not unusual that analysts do not estimate the nesting scale parameters in *Figure 4*'s model but to postulate them (Vosha, Ramming), but with a loss of precision. The simplified structure of the proposed model enables us to estimate all the nesting scale parameters.

Besides, it should be noted that if one path does not belong to any subpath, the choice probability of this path would be zero according to *Equation 3*. Therefore, when the analysts design the model and specify the subpaths, they should be certain that each path at least belongs to one subpath. In particular, the relevance and the specifications of subpaths should be tested after estimation to confirm or reject various hypotheses.

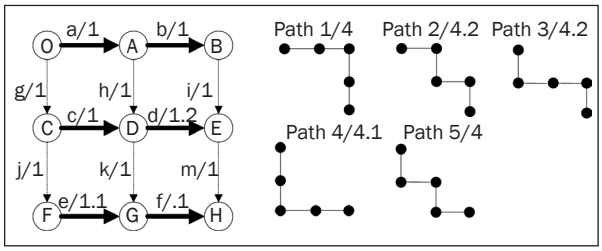

Figure 2 – Demonstrated grid network

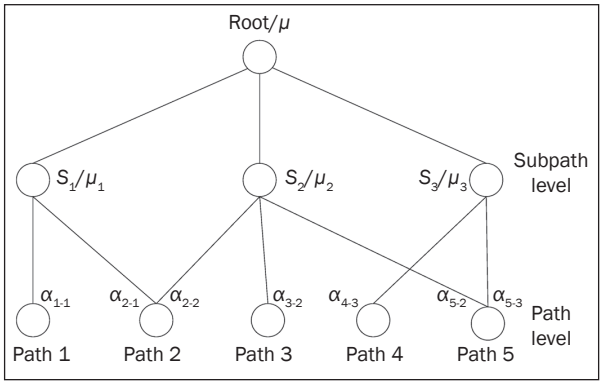

Figure 3 – Structure of a subpath CNL model

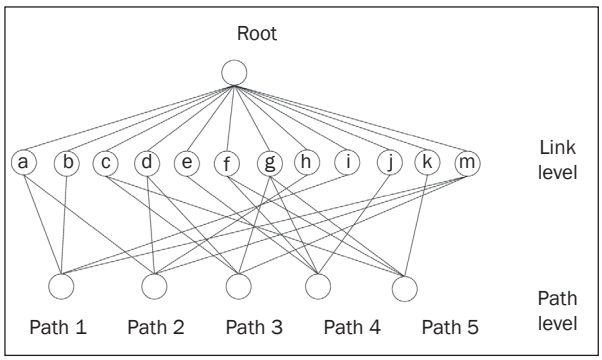

Figure 4 – Structure of a link-based CNL model

3. EMPIRICAL RESULTS

In order to evaluate the performances of the proposed method with other models, this section applies the new model with real data. A case study of taxi drivers choosing routes in the city centre is presented. The studied city, Guangzhou, is situated in the southern China and has approximately ten million inhabitants. Only the central business district, the Tianhe region as shown in *Figure 5*, is studied. The information on the studied network is shown in *Table 1*. The data set for the estimation is from GPS-equipped taxis when they were carrying passengers. The data was collected by a management company for monitoring purposes but not for navigation, so the route choice behaviour is based on the drivers' own judgement. Besides, the GPS points are collected every 30 seconds; therefore, the data are very sparse and sometimes difficult to be accurately matched to a specific link; however, it is easier to identify which major roads or major intersections they passed, so the proposed method is suitable. The vehicles were monitored within a radius of 2 km in the CBD, and 7,810 trips from 1,066 OD pairs are collected for case study, as shown in *Table 2*. The statistics of the observations show that the maximum, minimum and average number of the chosen routes between an OD pair are 12, 2 and 4, respectively. This indicates that more than one route is considered and chosen by the drivers, suggesting the need to investigate their route-choosing behaviours and, particularly, to analyze how they perceive and learn the network and how they change their behaviours when attributes change.

Figure 5 – The studied region

We chose three subpaths which are the major roads that traverse the studied region. The bold dashed lines illustrated in the figure are the most frequently used major roads when travellers describe itineraries. Two of them are east-west directional, and one is south-north directional, with their IDs marked in the figure. We also used the GPS data to count how many trips used these subpaths, and the results are that out of 7,810 trips, 4,884 trips used subpath 1 (denoted as S1), 1,189 trips used S2, and 1,150 used S3. Besides, they are the most frequently used roads among others in this region. Therefore, they are selected as the sub-

Table 1 – Information on the studied network

Nodes	Unidirectional links	Major roads	Arterial streets	Minor streets	Signal-controlled intersections
208	662	24	34	32	57

Table 2 – Information on the collected data

Taxis	Observations	*OD* pairs
2569	7810	1066

Table 3 – Information on the studied subpaths

ID	Length (km)	Nb. of lanes (bi-direction)	Signal-controlled intersections	Used times*	Remarks
1	3.2	8	9	4,884	Traffic lights are set as green bandwidth for south-north directions
2	3.8	12	0	1,189	
3	4.1	12	0	1,150	There is one bottleneck and the capacity decreases in half

* Out of 7,810 trips

paths in the model, and their attributes are provided in *Table 3*.

3.1 Choice set generation

We used the GPS data and the link elimination method (Azevedo et al., 1993) to generate the considerable choice set for each observation. According to one day's GPS data in the studied region, 7,810 trips are made between 1,066 OD pairs, and the number of actual chosen paths between any OD are not larger than 12; moreover, the average is just four paths. It suggests that in the studied region, the number of the actual chosen paths by the taxi drivers is not large. In order to build a considerable choice set, for each OD pair we firstly include the chosen paths into the choice set, and then we use the link elimination method (Azevedo et al., 1993) to generate more paths until there are 20 paths in the choice set. This algorithm finds the shortest path and then adds it to the choice set; afterwards it eliminates one link from the original shortest path, and another shortest path is generated from the modified network and added to the choice set, if it was not generated before. In this data set we cannot have access to the information on each driver, but since they are taxi drivers they are assumed to be equally familiar with the transportation network and traffic conditions. Without loss of generality we assume that the choice sets between one OD pair are the same for all the taxi drivers.

3.2 Model specification

We compare the new model with the MNL, PSL, CNL and EC models. The MNL model is chosen for comparison because it is the traditional model; however, it cannot account for the correlation of alternatives. The PSL model is chosen because it is the most used model in route choice and because it considers the correlation; however, it only captures the physical overlap. Additionally, the PS term is derived from an approximation. The CNL model systematically captures the correlation, but has a more complex form. Finally, the EC model, which also considers the subpath effect, similar to the proposed model, is compared as well. From the comparison among these models, we can assess the performance of the proposed model in terms of the calculation time, fitting of the data, and forecasting ability. The choosing probability of path i from choice set C by the MNL model is

$$P(i) = \frac{\exp(V_i)}{\sum_{j \in C} \exp(V_j)} \tag{13}$$

where V_i is the deterministic utility of path i. As for the PSL model, a path-size term lnPSi [1] is added to V_i as

$$\ln(PS_i) = \ln\left[\sum_{\alpha \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C} \delta_{aj}}\right] \tag{14}$$

where l_a is the length of link a, Γ_i is the set of links belonging to route i, δ_{aj} is a route-link incidence dummy which equals one if route j uses link a, and zero otherwise. The individual heterogeneity is ignored here for illustrative simplicity thus n is not in the model.

The link-based CNL model is shown in *Equations 7* and *11*, where the scale parameter of each nest is computed by the specification of Vovsha and Bekhor [2] as

$$\mu_m = 1 - \frac{\sum_{j \in C} \alpha_{mi}}{\sum_{j \in C} \delta_{mi}} \tag{15}$$

and it is denoted as L-CNL$_1$. We also test another link-based CNL model, denoted as L-CNL$_2$, where the nests scale parameters are all assumed to be the same, and they are estimated from the observations.

Regarding EC model [18], an error component is added to the utility *Equation 2* to represent the correlation between subpaths, and it is defined as FTζ, where $F_{(JxQ)}$ is the loading matrix (J is the number of paths and Q is the number of subpaths), and an element f_{ij} of $F_{(Nxm)}$ equals $\sqrt{l_{ij}}$ where l_{ij} is the length by which path i overlaps with path j $(i,j \in C)$; $T_{(QxQ)}$=diag(σ_1, σ_2 ,..., σ_Q) (σ_q is the covariance parameter associated with subpath q, to be estimated), $\zeta_{(Qx1)}$ is a vector of IID N(0,1) variates. It is denoted as the SP-EC model.

Four attributes are chosen for the utility function, as shown in the following

$$V_i = \beta_L \cdot Length_i + \beta_{ARR} \cdot ArteryRoadRatio_i + \\ + \beta_S \cdot Signal_i + \beta_{PS} \cdot \ln PS_i \tag{16}$$

The length and time are two highly similar and correlated attributes, so only one of them is considered in the utility function. However, it is difficult to obtain a precise actual travelling time before departure. Although the actual consumed time can be captured by the GPS device in this case, it is not the time perceived by the drivers. Note that in route choice we are actually modelling the drivers perception and perception error. Therefore, we choose length rather than time as the attribute in the utility. The unit of *Length* is kilometers; therefore, its magnitude is similar to other attributes for the convenience of the estimation. The *ArteryRoadRatio* is the length of the artery road (major roads and arterial streets) divided by the total length of the trip and is used to test the assumption that travellers prefer to drive on the artery links, so a higher ratio is expected to have a larger utility. The information on traffic light settings, such as waiting time and coordinated green wave, is also useful; however, we cannot obtain such information. Instead, we use the number of signal-controlled intersections in the utility; it is expected to have a significant effect on the urban route choice. The more intersections with traffic lights, the lower is the utility of the path. The $\ln PS_i$ is the path-size term of path i, which represents the impact of the overlapping degree of alternative paths. For comparison, the model without $\ln PS_i$ is also estimated. The parameters of the coefficients in the utility are assumed to be the same for all of the drivers.

3.3 Estimation

Estimation results are shown in *Table 4*, where the estimation of SP-EC model is based on 100 draws of simulation. SP-CNL$_1$ is the proposed model whose utility is without the term $\ln PS_i$, and the utilities of SP-CNL$_2$ and SP-EC are with this term. We perform a scaled estimation to facilitate comparison among models, because we cannot solely estimate μ and β, and the coefficient estimates are really estimates of $\mu\beta$ jointly [9]. Therefore, we fixed the estimates of length, which is actually $\mu\beta_L$, of the MNL model for the rest of the models, and the scale for all the models is consequently the same.

The estimates all have the expected signs and they have similar values for the same attribute among models, except the one of *AteryRoadRatio* of the SP-EC model, which is approximately 10 times smaller than other models. The SP-EC model also has the smallest value of the adjusted likelihood ratio index $\overline{\rho}^2$, possibly due to the small value of draws in the simulation. The L-CNL$_2$ model with an estimated nesting scale outperforms the L-CNL$_1$ where the nesting scale is approximated by empirical method, in the sense that it has larger value of $\overline{\rho}^2$. This tells us that if we use the CNL model in route choice analysis, explicitly estimating the scale parameter will improve the fitting of the data and will not increase much the estimating time. The new model with a Path Size term in the utility is better than the one without, according to the $\overline{\rho}^2$ value, and both of them have estimates of μ_2 approximate to 1 which shows a low correlation. The estimation results of $\overline{\rho}^2$ suggest that the SP-CNL$_2$ model has worse data fit than the original CNL model. Regarding the estimation times, the new model excels in its computational time as compared to the original CNL model and particularly the original subpath model, SP-EC, whose time is approximately 100 times larger with only 100 draws. It suggests that the new model is advanced in computational efficiency, especially when applied to big network analysis and traffic assigning computation. The new model requires approximately 50 times less time than the CNL models, due to the simplified structure. Note that the studied network only includes 662 links and the estimation only includes 7,810 observations. If we enlarge the studied region and use more observations to train the models, the cost in time may be larger. The L-CNL$_2$ model produces the highest value of $\overline{\rho}^2$, but it may cost too much time in its application. Moreover, if we want to consider more effects in the model, e.g., the sampling of alternatives [8], the estimation time will be too large if we use the original L-CNL$_2$ model. Therefore, introducing the proposed model is appropriate to simplify the original one, to reduce the consumed time, and to obtain a usable trade-off between data fitting and cost.

3.4 Forecasting

The route choice model is mostly used in predicting travel behaviours. Thus, the model prediction ability is more important than its data-fit in estimation; therefore we also perform a validation experiment to further examine the forecast power of different models.

An out-of-sample forecasting method is presented, and the term “out-of-sample” means the data sets for validating forecasting have not been used for estimating the models. There are 200 randomly selected OD pairs including 3,345 observations that are used for the forecasting validation. The models with the estimated parameters are used back to calculate the choosing probabilities of these trips, and then they are compared with the actual choosing behaviours. The out-of-sample error is used for comparing the forecasting power of models, and the smaller value indicates a better performance.

The out-of-sample error is defined as

$$\varepsilon_f^z = \sqrt{\sum_{n=1}^{N} \left(P_n^R(C) - P_n^Z(c, \beta^Z)\right)^2} \tag{17}$$

where N is the number of observations; Z is the type of models, which are the seven compared models in this section; $P_n^Z(c, \beta^Z)$ is the predicted probability that traveller n chooses the chosen path c, computed by

Table 4 – Estimation of compared models

Parameters	MNL	PSL	L-CNL$_1$	L-CNL$_2$	SP-CNL$_1$	SP-CNL$_2$	SP-EC
Length (fixed)	-1.26	-1.26	-1.26	-1.26	-1.26	-1.267	-1.26
Std. err.	0.0444	0.0742	0.0774	0.0713	0.345	0.0585	0.148
t-test	28.3	16.9	16.2	17.7	3.65	21.7	8.54
Artery road ratio	5.52	5.24	5.73	2.94	5.32	5.01	0.621
Std. err.	0.0729	0.0689	0.0798	0.0332	0.0465	0.0469	0.0968
t-test	75.7	65.6	71.8	88.5	114	106	6.41
Nb. Signals	-0.646	-0.526	-0.561	-0.330	-0.611	-0.495	-0.709
Std. err.	0.0161	0.0184	0.0204	0.0105	0.0192	0.0226	0.0934
t-test	40.1	28.5	27.5	31.4	31.8	22.0	7.60
Path Size	-	0.371	0.435	0.250	-	0.361	0.204
Std. err.		0.0391	0.0400	0.0489		0.024	0.102
t-test		9.48	10.8	5.11		14.6	2.00
μ_{link}				5.17			
Std. err.				0.173			
t-test				29.8			
μ_1					1.07	1.07	
Std. err.					0.0273	0.0230	
t-test					39.2	46.5	
μ_2					1.01	1.05	
Std. err.					0.0543	0.0392	
t-test					18.6	26.8	
μ_3					1.06	1.13	
Std. err.					0.0380	0.0454	
t-test					28.0	24.8	
σ_1							1.60
Std. err.							0.265
t-test							6.04
σ_2							1.36
Std. err.							0.0857
t-test							15.8
σ_3							1.60
Std. err							0.331
t-test							4.85
Final L-L	-610×10	-600×10	-606×10	-584×10	-609×10	-599×10	-744×10
L(0)	104×10^2	104×10^2	-104×10^2	104×10^2	104×10^2	104×10^2	104×10^2
$\overline{\rho^2}$	0.413	0.423	0.418	0.439	0.415	0.424	0.285
Elapsed time (s)	3.10	3.32	469	532	10.4	12.6	151×10

model Z with the estimated parameters β^Z; $P_n^R(c)$ is the real probability that traveller n chooses the chosen path c in reality, which is one.

The out-of-sample forecasting errors of the models are shown in *Table 5*. The error of the SP-EC model is the largest, probably due to the small draws in estimation. The new model with the Path Size correction term is superior to the compared models, but the one without the term has a slightly larger error than the L-CNL$_2$ and PSL model. This suggests that adding the Path Size term into the utility does increase the model's ability in modelling travellers' route choice behaviours. The results from this experiment suggest that the new model indeed outperforms others, which is different from what the estimation results suggest, where the new model shows an inferior data-fit than the original CNL model. We thus conclude that the SP-CNL$_2$ model

Table 5 – Out-of-sample forecast validation

Model	MNL	PSL	L-CNL$_1$	L-CNL$_2$	SP-CNL$_1$	SP-CNL$_2$	SP-EC
ε_f^M	40.5	39.5	40.5	39.4	39.9	36.2	45.2

performs the best in reproducing the expected results in the sense that it has the smallest out-of-sample error in forecasting. It should be noted that this conclusion is based on a one-time test, and more validation exercises and more networks should be performed and tested to comprehensively evaluate the proposed model. Moreover, because this is a route choice model of taxi drivers, estimates should not be directly used for analyzing all types of drivers. Because this "pilot" exercise on taxi drivers shows that the new model is practical, it will be useful and suitable for analyzing other drivers' route choice behaviours.

4. CONCLUSION

The paper presents a simplified structure of the cross-nested Logit model to capture travellers' perception in higher-level of the road network, where the correlation of alternatives is not specified from a link-by-link style, but from the subpath perspective. The subpath idea is applied to capture drivers conceptual correlation of routes, and a CNL model is employed to (1) decrease the computational time of the original EC-based model, then non-simulated method is required; (2) to explore the usage of the CNL model because the original model is not always flexible in route choice for its large set of parameters to be estimated. Results with real data suggest that the new model is practical and capable of reproducing the expected results. The proposed approach shows superiority in computational time saving and its ability to capture travellers' choosing behaviours on a perceptual level.

Future directions of the subpath idea would be the expansions to: (1) the choice set generation / sample step; (2) data map-matching step, where in this paper we still use the link-specific style. Besides, it would be useful to examine and discuss the appropriate size of the subpath in the CNL model. Moreover, the proposed idea can be applied with the DDR [14] or the recursive Logit [5] methods; therefore, the "data processing"-"choice set generation"-"route choice model" steps can be obtained at the same time, instead of doing them step-by-step.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 51178475, 51405089), the Science and Technology Planning Project of Guangdong Province (2015B010131008), the National Development and Reform Commission General Office's High-tech project ([2011]2448), and National enterprise Internet service support software engineering technology research center (2012FU125Q09).

赖信君 (博士，广东工业大学，机电工程学院，广东省计算机集成制造重点实验室，广州 510006)
李军 (博士，副教授，中山大学，工学院 A303，广州 510006)
李志 (博士，讲师，广东工业大学，机电工程学院，广东省计算机集成制造重点实验室，广州 510006)

标题: 基于子路径的logit选择模型及重叠路径表征

摘要: 提出一种基于子路径的logit选择模型，用以建模出行者的感知重叠路径和路径选择行为，用以解决传统的基于路段的模型在应用中所存在的问题：（1）出行者对路网信息的记忆和处理，难以细致到每一个具体的路段；（2）利用问卷和GPS收集的出行信息，并不以完整的路段形式表达。本文提出子路径以表征路径的重要部分，如主干道或地标等。交互巢式logit模型用于显式建模路径的重叠部分，模型中的每一个巢为一个子路径，代替原模型中表征路段的巢，用以大幅度减少巢的数目和计算时间。通过实例数据对模型进行参数估计并进行预测检验，显示新模型有较好的实用性。

关键词
交互巢式logit; 随机路径选择；子路径；重叠

REFERENCES

[1] Ben-Akiva ME, Bierlaire M. Discrete Choice Methods and Their Applications to Short Term Travel Decisions. In: Hall R, editor. Handbook of Transportation Science. Dordrecht, Netherlands: Kluwer; 1999. p. 5-33.
[2] Vovsha P, Bekhor S. Link-Nested Logit Model of Route Choice: Overcoming Route Overlapping Problem. Transportation Research Reord. 1998;1645:133-42.
[3] McFadden D, Train K. Mixed MNL Models for Discrete Response. Journal of Applied Econometrics. 2000;15(5):447-70.
[4] Prato CG. Route choice modelling: Past, present and future research directions. Journal of Choice Modelling. 2009;2(1):65-100.
[5] Fosgerau M, Frejinger E, Karlstrom A. A link based network route choice model with unrestricted choice set. Transportation Research Part B. 2013;56:70-80.
[6] Papola A, Marzano V. A Network Generalized Extreme Value Model for Route Choice Allowing Implicit Route Enumeration. Computer-Aided Civil and Infrastructure Engineering. 2013;00:1-21.
[7] Frejinger E, Bierlaire M, Ben-Akiva M. Sampling of alternatives for route choice modelling. Transportation Research Part B. 2009;43:984-94.
[8] Lai X, Bierlaire M. Specification of the cross-nested logit model with sampling of alternatives for route choice

models. Transportation Research Part B: Methodological. 2015;80:220-34.

[9] Ramming MS. Network Knowledge and Route Choice [PhD thesis]. Cambridge, USA: Massachusetts Institute of Technology; 2002.

[10] Axhausen K, Schönfelder S, Wolf J, Oliveira M, Samaga U. 80 weeks of GPS-traces: Approaches to enriching the trip information. IVT, ETH Zurich; 2003.

[11] She X, He Z, Nie P, Zeng W, Cen X, Dai X. Online Map-Matching Framework for Floating-Car Data with Low Sampling Rate in Urban Road Network. Transportation Research Board 91st Annual Meeting; Washington DC; 2011.

[12] Rahmani M, Koutsopoulos HN. Path inference from sparse floating car data for urban networks. Transportation Research Part C: Emerging Technologies. 2013;30:41-54.

[13] Li J, Xie L, Lai X. Route Reconstruction from Floating Car Data with Low Sampling Rate Based on Feature Matching. Research Journal of Applied Sciences, Engineering and Technology. 2013;6(12):2153-8.

[14] Bierlaire M, Frejinger E. Route choice modelling with network-free data. Transportation Research Part C: Emerging Technologies. 2008;16:187-98.

[15] Barton RR, Hearn DW. Network Aggregation in Transportation Planning Models. United States Department of Transportation; 1979.

[16] Fosgerau M, McFadden D, Bierlaire M. Choice probability generating functions. Journal of Choice Modelling. 2013;8:1-18.

[17] Boyles SD. Bushed-based sensitivity analysis for approximating subnetwork diversion. transportation Research B. 2012;46(1):139-55.

[18] Frejinger E, Bierlaire M. Capturing correlation with subnetworks in route choice models. Transportation Research Part B. 2007;41:363-78.

[19] Kazagli E, Bierlaire M. Revisiting Route Choice Modelling: A Multi-Level Modelling Framework for Route Choice Behaviour. 14th Swiss transportation research conference; Ascona, Switzerland; 2014.

[20] Li J, Lai X, Yu Z. A Paired Combinatorial Logit Route Choice Model with Probit-Based Equivalent Impedance. Journal of Transportation Systems Engineering and Information Technology. 2013;13(4):100-4.

[21] Abbe E, Bierlaire M, Toledo T. Normalization and correlation of cross-nested logit models. Transportation Research Part B. 2007;41:795-808.

[22] McFadden D. Modelling the choice of residential location. In: Karlquist, Lundqvist, Snickers, Weibull, editors. Spatial Interaction Theory and Residential Location. Amsterdam: North Holland; 1978. p. 75-96.