Hyperpath or Shortest Path: An Evaluation Method and a Case Study with GPS Probe Data

Jiangshan Ma [a], Daisuke Fukuda [b]
[a,b] *Department of Civil Engineering, Tokyo Institute of Technology, Tokyo, 152-8552, Japan*
[a] *E-mail: ma.js@plan.cv.titech.ac.jp*
[b] *E-mail: fukuda@plan.cv.titech.ac.jp*

Abstract: Because of travel time uncertainty in traffic networks, the shortest path determined with a priori information may consist of links for which there are long delays and the path may not be optimal. Adaptive routing would suggest more reliable guidance by providing alternative links for en-route switching in the case of traffic congestion and a risk-averse hyperpath particularly provides an alternative method of adaptive route guidance. To evaluate hyperpath routing in terms of behavioral reality, we propose a method based on the cosine similarity index that takes advantages of huge quantities of route choice data. As an empirical study, we use taxi GPS probe data collected for about four years in Tokyo and compare hyperpath routing with the popular shortest-path routing in terms of reality of driver behavior. The empirical results indicate that hyperpath routing is closer to the real route choice than shortest-path routing.

Keywords: Routing, Shortest path, Hyperpath, GPS probe data, Cosine similarity index

1. INTRODUCTION

The development of shortest-path (SP) algorithms, digital maps and positioning technologies has resulted in the widespread use of vehicle navigation systems. Furthermore, the vehicle navigation market is still expanding. Since the development of the well-known Dijkstra algorithm (Dijkstra, 1959), the SP problem has been well studied. Among algorithms, the A-star (A*) algorithm (Hart et al., 1968) and its variants have played the most pivotal role in car navigation and have been widely implemented in most in-vehicle Global Positioning System (GPS) navigation systems.

The SP problem, however, tends to become non-trivial in networks characterized by travel time uncertainty or variability. There are some sophisticated routing algorithms dealing with travel time variability (Fu and Rilett, 1998; Miller-Hooks and Mahmassani, 2000; Nie and Wu, 2009) but these algorithms work only when travel time distributions are available. Kaparias and Bell (2009) took stochastic characteristic based reliability index into consideration, which results in a more efficient reliable routing algorithm. However, the routes recommended by these algorithms pre-determined and are still risky because of travel time uncertainty. Chen et al. (2010) proposed a risk-averse route guidance algorithm employing a constrained A* search. Other literature has suggested that multipath navigation could be more favorable for drivers (Chi-kang, 1994; Chen et al., 2007). Alternatively, instead of focusing on routes, a route choice can be considered as a result of sequential link choices (Fosgerau et al. 2009) and can consequently be similar to a multipath recommendation. These studies focused on recommending prior routes but neglected the adaptability of routing to revealed travel time.

In traffic assignment, the well-known Dial algorithm (Dial, 1971) excludes links of zero likelihood and finds reasonable SPs in the remaining link set. However, link delays are not considered in Dial's algorithm when generating the sub-network comprising the remaining links. In frequency-based transit assignment, additional waiting times are involved and a hyperpath sub-network can be found by minimizing the expected total travel time (Nguyen and Pallotiino, 1988; Spiess and Florian, 1989). The resulting hyperpath is based on the concept of optimal strategy and leads to a link-to-link strategic route choice. Since the basic implementation of the original Spiess and Florian algorithm (Spiess and Florian, 1989), Cominetti and Correa (2001) proposed a hyperpath–Dijkstra algorithm, which resembles the structure of Dijkstra's algorithm but theoretically has less time complexity. Noting that the waiting time for a transit line is comparable to the delay for a specific road link, Bell (2009) and Bell et al. (2012) initiated research into risk-averse hyperpath-based vehicle navigation referring to the idea of the optimal strategy, and proposed the hyperstar algorithm as an acceleration inspired by the A* algorithm. Ma et al. (in press) further accelerated the hyperstar algorithm to the Dijkstra–hyperstar algorithm using optimistic heuristics pre-calculated with Dijkstra's algorithm and a node-directed search that is similar to the best-first search proposed by Dechter and Pearl (1985). Empirical performance tests of hyperstar, hyperpath–Dijkstra and Dijkstra–hyperstar algorithms were also carried out by Ma et al. (in press).

The behavioral reality is undoubtedly important in vehicle routing but it is seldom considered in most routing algorithms mainly because of the rigorous performance requirement of algorithms. The most popular SP algorithm implicitly assumes that all travelers definitely take the SP and also has been adopted in all-or-nothing traffic assignment (1984). Morikawa et al. (2005) evaluated route choice decisions based on GPS data taken from 1500 taxis in Nagoya, Japan over 2 months and found a high percentage of non-shortest paths. Zhu and Levinson (2012) evaluated the SP assumption using GPS data in the Twin Cities Metropolitan Area. In the field of route choice modeling, it is popular to model route choices with behavioral models based on random utility maximization and how to construct drivers' route choice set in the model is crucially important for demand forecasting. Bekhor et al. (2006) emphasized that many route choice models presented in the literature pay little attention to empirical estimation and validation procedures, and evaluated several route choice set generation algorithms with data collected in Boston by asking travelers for route descriptions. According to Zhu and Levinson. (2012), current route choice set generation cannot reveal the majority of observed paths.

In recent years, with the development of technologies such as the GPS, it is possible to collect a large amount of route choice data with probe vehicles equipped with GPS devices. This leads to many possible uses of the huge quantity of GPS data. As a result, data-driven route choice modeling has become popular in recent years. In route choice modeling, the first step is to extract several routes from an entire network as a choice set. Conventional methods are based on the biases of modelers instead of actual data. For example, K-SP algorithms are usually used to generate such choice sets. It is unavoidable that such biased route choice set generation is an obstacle in route choice modeling. Fafieanie (2009) examined the possibility of calibrating the route choice set generation by map-matching GPS data. In contrast, Frejinger et al. (2009) proposed a method for sampling the route choice set directly from observations. Both studies aimed to build a route set that is more indicative of reality.

To sum up the above-mentioned literature review about vehicle routing algorithms, multipath traffic/transit assignment, and choice set generation for route choice modeling, there are two aspects which ensure the behavioral reality of modeling methodologies: (1) are the modeled routes the same as those used in reality and (2) do the estimates of link/route importance (probability) reflect travelers link/route preferences? In the context of vehicle navigation, the algorithm performance is quite rigorous so that the probabilistic methods widely used in route choice modeling are generally not used. Alternatively, we anticipate that hyperpath routing balances the performance and behavioral reality, which may be promising in behavior-consistent route guidance. This motivates us to study the real route choices and consider whether the hyperpath routing reflects the reality better than popular SP routing. The availability of huge quantities of GPS probe vehicle data makes it possible to compare the behavioral reality of the hyperpath and SP by posterior processing. The objective of the paper, therefore, is to evaluate hyperpath routing in terms of behavioral reality by employing the proposed method based on the cosine similarity index that takes advantages of huge quantities of route choice data.

This paper first outlines risk-averse hyperpath routing. A general method for evaluating routing algorithms is then proposed and an empirical study for Tokyo is presented. Finally, the paper presents conclusions and a discussion on data limitation and future work.

2. HYPERPATH ROUTING: MODEL AND ALGORITHM

The following notation is used throughout the paper:

r / s:	origin/destination node
I / A:	set of nodes/links
i / j:	tail/head node of a link
A_i^+ / A_i^-:	outgoing/incoming links of node i
H:	hyperpath link set
H_i^- / H_i^+:	$H_i^- = H \cap A_i^-$, $H_i^+ = H \cap A_i^+$
I^O / I^C:	set of open/closed nodes
A^O / A^C:	set of open/closed links
p_i :	probability that node i is used in H
p_a / P_a:	probability that link a is used in H / H_i^+
c_a:	uncongested travel time on link a
d_a:	potential maximum delay on link a
u_i:	shortest travel time from node i to the destination node expected by pessimists minimizing the maximum exposure to delay
h_i:	node potential of i with respect to r
N:	a large number that depends on the precision of the computation
K_i:	the number of attractive links when facing the link choice decision at node i

The shortest path can hardly been found when travel times become uncertain. In this case, travelers usually follow a strategy which turns out to be a set of potentially optimal path (hyperpath). According to Bell (2009) and Ma et al. (in press), the risk-averse hyperpath problem, which minimizes expected total travel time by pessimistic travelers, is expressed as

$$\min \sum_{a \in A} (c_a + d_a) \cdot p_a, \tag{1}$$

subject to

$$\sum_{a \in A_i^-} P_a = 1, \ \ i \neq r$$

$$\sum_{a \in A_i^+} P_a = 1, \ \ i \neq s$$

$$\frac{P_a}{P_b} = \frac{d_b}{d_a}, \forall a, b \in H_i^+$$

$$p_a = P_a \cdot p_i, \quad \forall a \in H_i^+$$

$$p_i = 1, \ \ i \in \{r, s\}$$

$$p_a \geq 0, P_a \geq 0, p_i \geq 0$$

We assume that travelers are risk averse and make probabilistic sequential link choices that are sensitive to link delays. The link choice probabilities are assumed to be inversely proportional to the potential delay:

$$\frac{P_a}{P_k} = \frac{d_k}{d_a}, a \in H_i^+ \text{ and } k \in H_i^+. \tag{2}$$

Since

$$\sum_{k \in H_i^+} P_k = \sum_{k \in H_i^+} \left(\frac{d_a}{d_k} \cdot P_a \right) = 1, \tag{3}$$

we have

$$P_a = \frac{1}{\sum_{k \in H_i^+} \left(\frac{d_a}{d_k} \right)} = \frac{\frac{1}{d_a}}{\sum_{k \in H_i^+} \left(\frac{1}{d_k} \right)}. \tag{4}$$

Furthermore, we assume more attractive links help mitigate the exposure to the risk of delay and introduce the "delay exposure to attractive links" e_i at a node i as

$$e_i = \frac{\sum_{k \in H_i^+} (P_k d_k)}{K_i} = \frac{\sum_{k \in H_i^+} \left(\frac{1}{d_k} \cdot d_k \right)}{K_i \cdot \sum_{k \in H_i^+} \left(\frac{1}{d_k} \right)} = \frac{1}{\sum_{k \in H_i^+} \left(\frac{1}{d_k} \right)}. \tag{5}$$

The travel time expected by pessimists at node i is then denoted by

$$T_i = \sum_{k \in H_i^+} (P_k \cdot s_k) + e_i = \frac{1 + \sum_{k \in H_i^+} \left(\frac{s_k}{d_k} \right)}{\sum_{k \in H_i^+} \left(\frac{1}{d_k} \right)}, \tag{6}$$

where s_k is the expected travel time (including potential delay exposures) of the remainder of the trip. Eq. [1] can be transformed to Eq. [7] by considering that $e_i = P_a \cdot d_a$ and further introducing $\omega_i = \left(\sum_{a \in H_i^+} p_a \right) \cdot e_i = p_i \cdot e_i$. If we set $f_a = 1/d_a$, P_a and e_i respectively become $f_a \,/\, \sum_{k \in H_i^+} f_k$ and $[1 + \sum_{k \in H_i^+} (f_k s_k)] / \sum_{k \in H_i^+} f_k$, which are of the same mathematical form as the relations in transit assignment. Eq. [7] is exactly the same as the optimal-strategy-based hyperpath problem in Spiess and Florian (1989), which was also used by Bell (2009):

$$\min \sum_{a \in A} c_a p_a + \sum_{i \in I} \omega_i, \tag{7}$$

subjected to

$$\sum_{a \in A_i^+} P_a - \sum_{a \in A_i^-} P_a = g_i,$$

$$p_a \cdot d_a \leq \omega_i, \quad a \in H_i^+,$$

$$p_a \geq 0,$$

$$g_i = \begin{cases} 1 & \text{if } i = r \\ -1 & \text{if } i = s \\ 0 & \text{otherwise} \end{cases}.$$

An accelerating algorithm named SF[di] for fast hyperpath generation was given by Ma et al. (in press) with optimistic node potentials and best-first search. For completeness, the algorithm is illustrated as follows.

Step 0: Initialization

$u_i \leftarrow \infty, \forall i \in I - \{s\}, u_s \leftarrow 0$; $f_a \leftarrow 1/d_a$ **if** $d_a > 0$, **else** $f_a \leftarrow M$; $f_i \leftarrow 0, \forall i \in I$;

$p_i \leftarrow 0, i \in \mathrm{I} - \{r\}$; $p_r \leftarrow 1, H \leftarrow \emptyset$; $\omega \leftarrow s; A^O \leftarrow \emptyset$, $A^C \leftarrow \emptyset$.

Conduct Dijkstra's algorithm: $h_i^O \leftarrow$ least optimistic travel time from r to $i, \forall i \in I$

Step 1: Selecting link *a*

1.1 for each link $k \in A_\omega^-$

if $k \notin A^O$ and $k \notin A^C$ **then** $A^O \leftarrow A^O + \{k\}$

1.2 select link a in A^O with the minimum $u_j + c_a + h_i^O$, $A^O \leftarrow A^O - \{a\}, A^C \leftarrow A^C + \{a\}$

Step 2: Updating node label

if $u_i \geq u_j + c_a$ **then**

if $u_i = \infty$ *and* $f_i = 0$ **then** $u_i \leftarrow \frac{1+f_a(u_j+c_a)}{f_i+f_a}$

else $u_i \leftarrow \frac{f_i u_i + f_a(u_j+c_a)}{f_i+f_a}, f_i \leftarrow f_i + f_a, H \leftarrow H + \{a\}, \omega \leftarrow i$

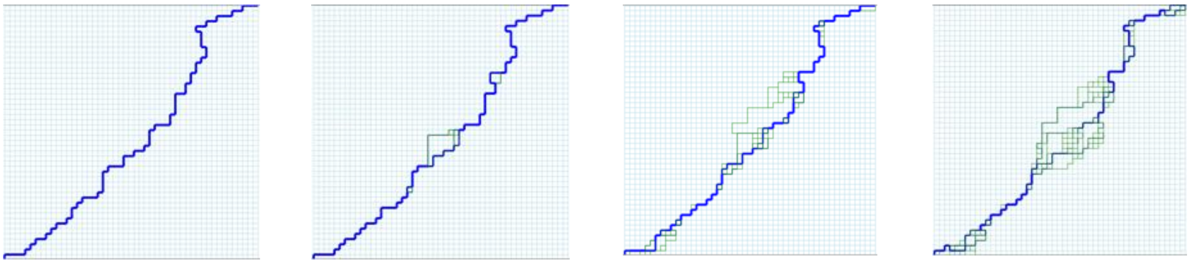
if $u_j + c_a + h_i^O > u_r$ *or* A^O is empty, **then** go to step 3 **else** loop steps 1 and 2

Step 3: Loading step

for each link $a \in A$ in decreasing order of $u_j + c_a + h_i^O$,

if $a \in H$ **then** $p_a \leftarrow (f_a/f_i) \cdot p_i$ and $p_j \leftarrow p_j + p_a$ **else** $p_a = 0$.

Heuristic search is involved in the SF^{di} so the actual complexity is dependent on the searching scenario. Empirical comparisons among existing hyperpath algorithms in computational speed are provided in Ma et al. (in press). Although heuristic node potentials are used, SF^{di} still gives an exact hyperpath since the node potentials are admissible. Hyperpath can be represented by a set of links and each link is associated with a link choice possibility according to behavioral assumptions (e.g. Eq. (2)). Figure 1 illustrates hyperpaths increasing with potential congestion on the network.



(a) No delay (b) Low delay (c) Medium delay (d) High delay

Figure 1 Hyperpath increase with network delay levels

(Link colors change from green to blue as choice possibilities increase)

One may concern the computational speed problem which is the main reason that many sophisticated algorithms cannot be used in practice. The original hyperpath is at a complexity of $O(m \cdot \log m)$ (m denotes the number of links) which is comparable to that of the shortest path algorithm ($O(n \cdot \log n)$, n denotes the number of nodes) if both algorithms utilizes priority queue based data structures. Consequently, the A-star-like SF^{di} algorithm is applicable in practical route guidance systems. Running time experiments are included in Ma et al. (in press).

3. A ROUTING EVALUATION METHOD BY USING COSINE SIMILARITY INDEX

3.1 Cosine similarity index

In route choice modeling, the problem of path overlap has not been considered in conventional logit model since it is based on the interdependent and identically distributed (iid) assumption of error terms. To deal with path overlap, models such as the C-logit model (Casscetta, 1996), link-nested logit model (Vovsha and Bekhor, 1998) and path size logit model (Ben-Akiva and Bierlaire, 1999) have been widely used. Additionally, the overlap problem has been considered in finding dissimilar routes (Akgün et al., 2000; Lim and Kim, 2005; Zijpp and Catalano, 2005).

The above-mentioned studies are based solely on the overlapping length. Although these studies have various specific representations of path overlap, they employ the common idea that two routes are more similar if they share a greater length. This is a natural and clear approach but difficult to apply when comparing link sets for two reasons: (1) the similarity of "link exclusion" should also be considered in addition to that of link inclusion (i.e. overlapped links) and (2) link choice frequencies or probabilities should also be considered.

To measure the extent that the hyperpath reflects reality, we propose to use the cosine similarity index (CSI) in vector space model (Tan et al., 2006). Let $\mathbf{X} = [x_1, x_2, \cdots, x_n]$ and $\mathbf{Y} = [y_1, y_2, \cdots, y_n]$ be link set vectors for a theoretical link set and real link set respectively, where x_n and y_n represent link counts and link choice possibilities respectively. With the nature of the cosine function, the CSI for the two link sets takes a value of 1 if the two link sets are identical and a value of zero if they are completely disjoint:

$$\mathrm{CSI} = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| \cdot \|\mathbf{Y}\|} \tag{8}$$

3.2 A simple example of the evaluating process

Figure 2 (a) shows a simple road network containing six nodes, eight links and one origin–destination (OD) pair. Figure 2 (b), (c) and (d) shows the situation for each hypothetical result (the hyperpath, SP and real route choice respectively). In this example, the vectors for hyperpath and SP in Figure 2 (b) and (c) are given by link choice probabilities in accordance with the link ID in Figure 2 (a). The link choice probabilities for hyperpath are calculated from the aforementioned algorithm. In contrast, travelers who follow SP definitely choose the links on SP, which means the probabilities of SP links are ones. The vector for real route choice set in Figure 2 (d) is obtained from route observations as shown in Table 1. Assuming the map-matched route choices obtained by analyzing the GPS probe data are those given in Table 1, the vector of real route choices and the CSI of the SP and hyperpath can be calculated. Route 1-3-6 is the SP without considering delays while it may suffer large delay in revealed traffic conditions. Route 1-2-6 could be longer than route 1-3-6 for undelayed travel but may be less exposed to possible large delays. The resulting hyperpath emphasizes travelers' behaviors in making risk-averse decisions (links on route 1-2-6 are more preferred). Neither link 3-5 nor link 5-6 is chosen in the hyperpath or SP routing because they may be unattractive in terms of either the undelayed travel time or potential delay. Nevertheless, because of the data accuracy and complexity of decision making, such links are still observed.

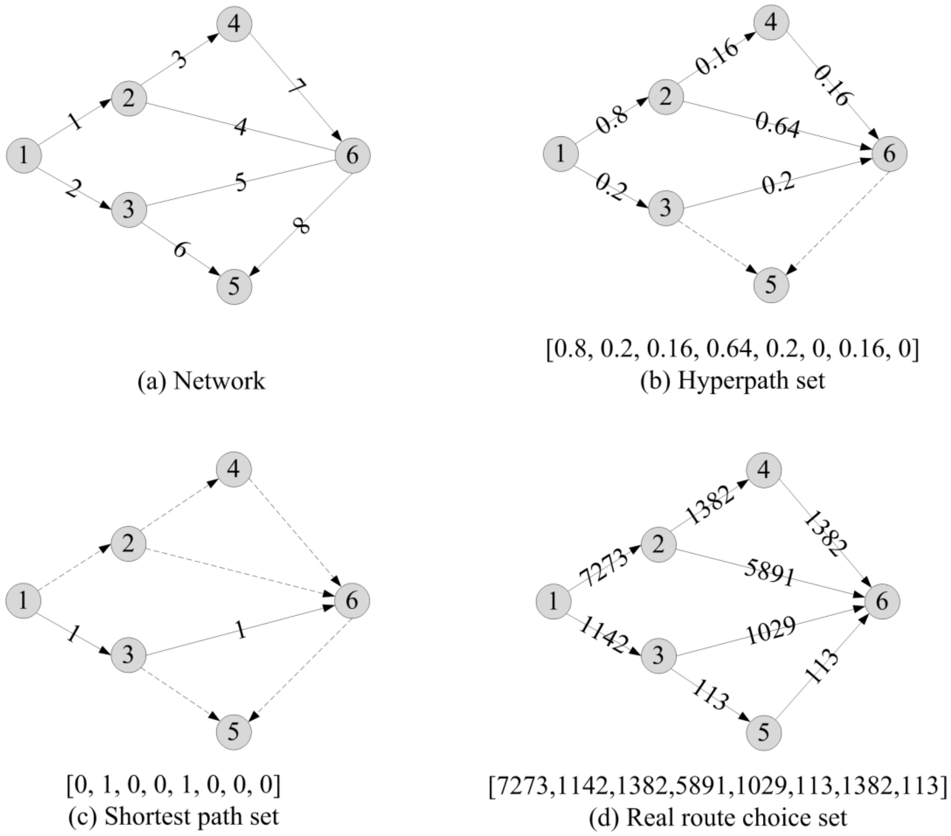


Figure 2. A simple network example (numbers on links are link IDs in (a), link choice probabilities in (b) and (c), and link counts in (d) respectively, and dashed links in (b) and (c) are unattractive links determined by the assumptions of each model)

Table 1. Example path set vector calculation

Observed Path	Count	Link	1	2	3	4	5	6	7	8
1→3→7	1382	Count	1382	0	1382	0	0	0	1382	0
1→4	5891		5891	0	0	5891	0	0	0	0
2→5	1029		0	1029	0	0	1029	0	0	0
2→6→8	113		0	113	0	0	0	113	0	113
Real route choice vector		Sum	7273	1142	1382	5891	1029	113	1382	113

In this example, the CSI for the hyperpath and real data is calculated as 0.994 while the CSI for the SP and real route choice is calculated as only 0.159. We thus say that hyperpath routing outperforms SP routing because it better reflects reality in this example. Note that the CSI is a relative index and a very large vector dimension will result in the CSI approaching one so that the differences among link sets are very small. Nevertheless, the vector dimension will not affect the relative relations among link sets. For example, in a network with a million links, a hyperpath for a OD pair consists of 50 links and the observed set consists of 150 links (30 links are captured by the hyperpath), and if all links are indexed, the CSI calculated for the two vectors (in dimensions of millions) closely approaches 1. However, this does not mean the two link sets are almost the same. In this case, the CSI might be close to one because most items in the vectors are zero. To avoid this, only the links involved in link sets for comparison are indexed in practice.

4. CASE STUDY WITH TAXI GPS PROBE DATA IN TOKYO

To evaluate hyperpath against the shortest path (in terms of travel time) in the aspect of behavioral reality, taxi GPS probe data collected in Tokyo are utilized. Sophisticated algorithms such as ELB which calculates least expected time (LET) paths (Miller-Hooks and Mahamassmi, 2000) are not evaluated for several reasons: Firstly, LET path requires assumptions or estimations about travel time distributions which is impractical for large scale networks because of data availability; Secondly, the computation of LET is too high to be conducted for the Tokyo network we studied. For example, ELB algorithm gives LET path at a complexity of $O(I^2n^3P)$ where I and P are the number of discretized time intervals and the maximum number of possible values of discrete arc travel time random variable of a time interval respectively; Thirdly, although LET path do gives a set of path, behavioral assumptions (about choice possibilities) have not been made. Instead of comparing with LET path, comparisons with network loading algorithms (e.g. Dial's algorithm or route choice models such as Path Size Logit model in Ben-akiva and Bierlaire, 1999) are possible but the motivation of comparison is still not clear since such algorithms are not designed for route guidance. At the current stage, we only focus on the comparison with the shortest path, which is the most widely used algorithm in practical applications.

4.1 Data description

The taxi GPS probe data used in this study was collected from October 2004 to July 2008 by the Ministry of Land, Infrastructure and Transport of Japan. GPS devices were installed in about 80 taxis in the Tokyo metropolitan area. The recorded area covers 15 cells. Each cell has dimensions of 11.35 km × 9.35 km, and one with high road density is selected as the study area (Figure 3). The “all roads” layer of the study area consists of 38,111 nodes and 108,363 directed links, of which 27,035 directed links are major roads.

A taxi position was recorded in real time at a sampling interval of 1 second and the tracking files were map-matched to the Tokyo base map with the SIS-based (CadCorp: http://www.cadcorp.com) software PROLIMAS by the Institute of Behavioral Sciences. The threshold for stopping was predetermined as the taxi being at rest for 120 seconds; i.e., when a taxi has a velocity of 0 km/h for more than 120 seconds, the “stopkey” attribute in a route table is recorded as true. Since no passenger boarding information has been recorded, it is difficult to choose a reasonable time interval to split a one-day route into trip-based routes. On the one hand, if the threshold is too long, trips may be combined into a single trip if a passenger alights from the taxi quickly, and on the other hand, a single trip may be split because of stops at intersections if the threshold is too short. With the “stopkey” records, stops longer than 120 seconds at non-intersection locations are considered as trip start/end points when splitting daily recorded routes into trip routes.

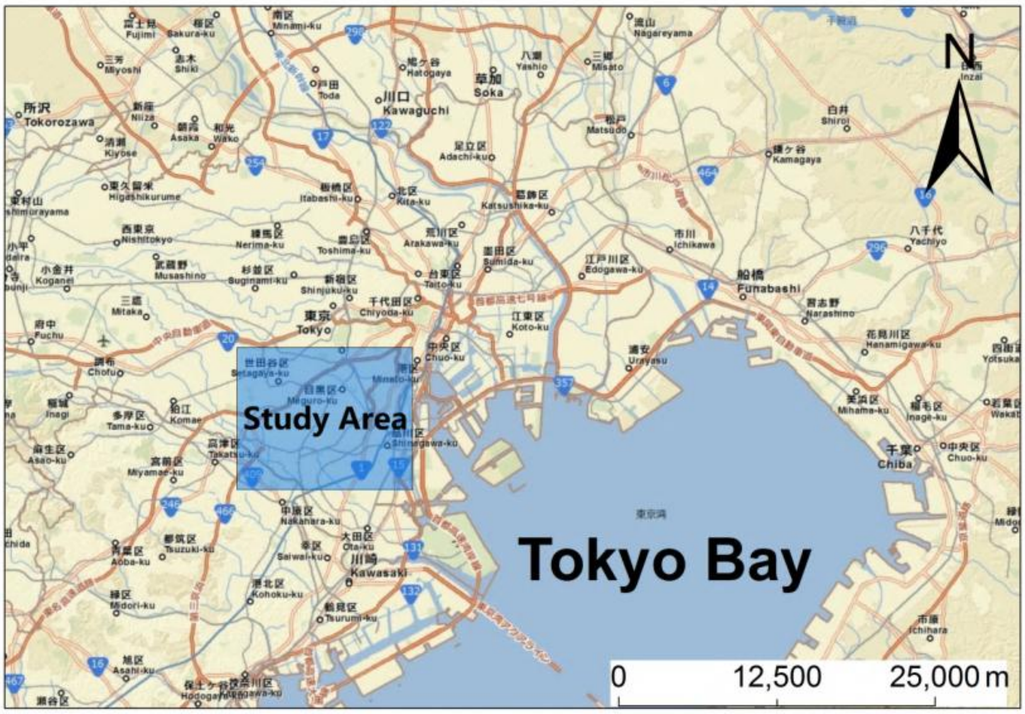


Figure 3. Selected study area (southwest Tokyo)

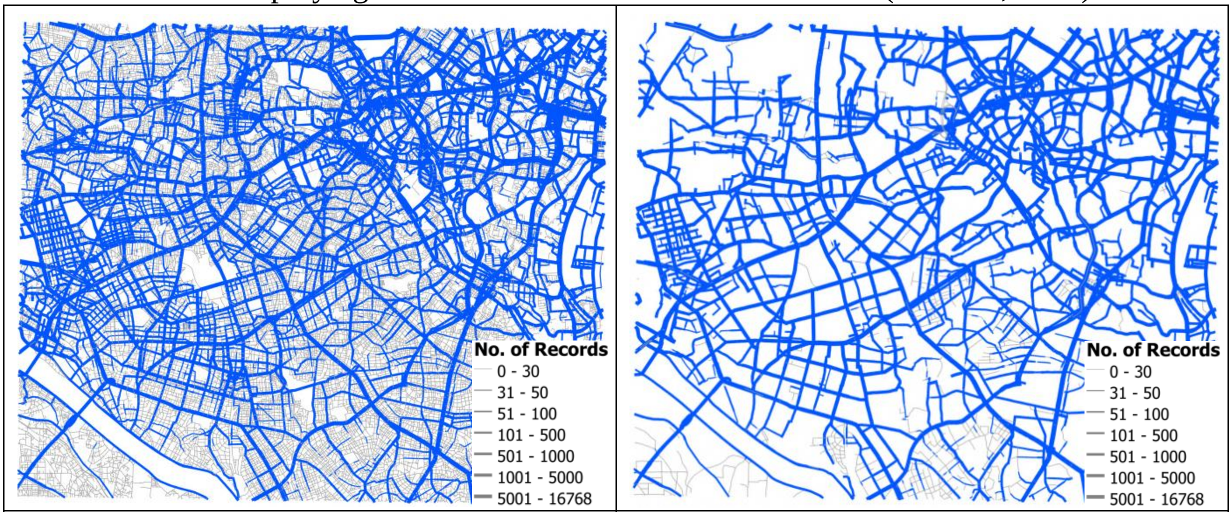
4.2 Calculation of link travel time

Bell (2009) assumed that travelers minimize their exposure to "maximum" delay. However, maximum delays may only occur in extreme cases related to unusual weather events or traffic incidents. Extremely risk-averse travelers will never use a link if there have been extremely long time delays on that link previously. Therefore, we implement the expected delay, which can be calculated from the historical probe vehicle data, as the maximum delay used in the hyperpath assumption. This setting would be reasonable to some extent because the original assumptions of hyperpath-based routing focus on uncertainty in travel times and care little about the possibility of delay.

To calculate the hyperpath, the undelayed link travel times and expected delays should be prepared for all links. The undelayed travel time is obtained from the length and legal speed limit of each link, and the delayed travel time (i.e., average travel time) can be directly calculated from the probe vehicle data. For each link for which we know the undelayed travel time and (with-delay) average travel time, the expected delay is calculated as the difference between the two.

In terms of variation in the link travel time, we only focus on the morning peak-traffic time from 7:00 a.m. to 9:00 a.m.. To avoid losing results for other time periods, the trip routes are related to a timetable. The prefix of the time ID indicates the weekday and the suffix indicates the time segment when the data were recorded, which auto-increases every 5 minutes. For example, the ID 1_2 represents the case of 00:05:00–00:10:00 on Monday. As a result, the time ID table starts from 1_1 and ends at 7_288. In this study, we only consider the peak-traffic data for weekdays in estimating the average link travel time.

There are many outliers in GPS data that need to be removed. The outliers are very long time intervals and may arise because taxis sometimes stop on a link but quickly restart within 120 seconds for various reasons. In practice, however, the sample size of 30 tends to be used as a cut-off point. We analyzed the number of travel time records on two different levels of GIS map layers (the main-roads layer and all-roads layer) and found that only 57.3% of links have enough GPS records in the all-roads layer (Figure 4 (a)). In the main-roads layer, 82.9% of links have a sample size meeting the cut-off point (Figure 4 (b)). Although we tend to simulate the travel times of unrecorded links, low data coverage would introduce problems in simulation. Therefore, we used the main-roads layer with a higher data coverage ratio and excluded outliers employing a statistical method of outlier detection (Ben-Cal, 2005).



(a) all-roads layer (b) main-roads layer

Figure 4. GPS probe data coverage in the map layers

Assuming the travel time of link a follows a normal distribution $N(\mu_a, \sigma_a^2)$, where μ_a and σ_a^2 are the mean and variance of the travel time on link a respectively, the average travel time for each link is calculated after excluding the outliers outside the confidence interval $[\mu_a - z_{1-\alpha/2} \cdot \sigma_a,\ \mu_a + z_{1-\alpha/2} \cdot \sigma_a]$, where $z_{1-\alpha/2}$ is the z-score at a confidence level of α. We take $\alpha = 0.05$; thus, the corresponding confidence interval is $[\mu_a - 1.96\sigma_a,$ $\mu_a + 1.96\sigma_a]$. Because the deletion of outliers also affects μ_a and σ_a, we loop the outlier exclusion iterations until newly calculated μ_a^* and σ_a^* become stable ($| \mu_a^* - \mu_a | < \epsilon_1$ and $| \sigma_a^* - \sigma_a | < \epsilon_1$, $\epsilon_1 = \epsilon_2 = 0.01$ in our case). For the links with sample size smaller than 30, we take the interval of $[0.5T_a, 5T_a]$ (where T_a is the travel time calculated from the design speed and length of link a) for outlier detection. Because of the huge data quantity but limited physical memory, an online variance algorithm introduced by Terriberry (2007) is used to calculate the mean and variance.

4.3 Data Generation for Links with no Travel Time Data

Since not all links were covered by probe taxis, there are links with no or only a few records of travel times. These links are referred to as failed links. The travel times on these links cannot be calculated directly, or the directly calculated travel times are not reliable, because of the sample size. Alternatively, we implemented a simple geometry-based simulation to generate the average travel time for the failed links. We simply assumed that the speed of a link should be consistent with that of its adjacent links and the speed of the failed link can be calculated by averaging the speeds of connected links oriented at small angles to the failed link (we adopted $15°$ as the cut-off point). An example is illustrated in Figure 5, where links 1, 2 and 3 have been covered by GPS probe data while link 4 has not ($\alpha \leq 15°$, $\beta \leq 15°$ and $\gamma > 15°$). The simulated speed of link 4 is thus obtained as

$$V_4 = \frac{V_1 + V_2}{2} = \frac{l_1/t_1^d + l_2/t_2^d}{2}, \tag{9}$$

where l and t^d represent the link distance and the with-delay travel time respectively. After the travel time simulation with speed correlations, the travel times of 97.2% links are available. The travel times of the remaining 2.8% links are calculated by assuming that their speeds are the same as the average speeds calculated for all links in the same road hierarchy.

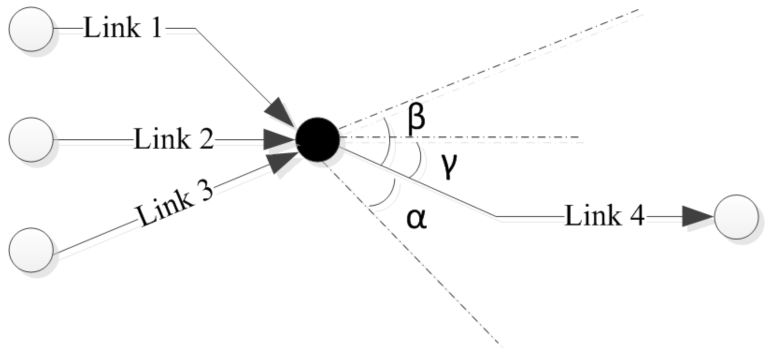


Figure 5. Travel time data calculation for links having low GPS data coverage (Links 1 and 4 are line strings on the GIS map)

4.4 Empirical Results Obtained from Tokyo Taxi GPS Probe Data

Although there is a large quantity of GPS probe data, an exact dot-based OD pair may lack sufficient support data. To obtain more route observations, we created buffers with the radius of 100 meters to aggregate the routes under the assumption that 100 meters is the expected acceptable walking distance to take a taxi in general.

To find the OD pairs with buffers for which there are many route observations, we put the mesh in a 12 km × 12 km box and then split the box into small cells (200 m × 200 m). These square cells are approximations of the circular buffers for exact origins or destinations. Ten exact OD pairs are then randomly selected (
Table 2 and Figure 6) from OD pairs having more than 300 route observations and distances between 3000 and 5000 m, which is considered as the typical interval for taxi trips. It is found that the OD pairs beyond this distance interval have few route observations and they therefore cannot be used in the evaluation. The average distance of these OD pairs is 4114 m and the average number of observed routes is 799. The hyperpath of these selected OD pairs is calculated with the main-roads layer because of the higher data coverage.

Table 2. Ten selected OD pairs and CSI results

ID	Origin ID	Destination ID	Distance (m)	Records	CSI_HP	CSI_SP
1	579	831	3165	376	0.2682	0.2371
2	21786	3577	4523	431	0.5581	0.3521
3	13520	2132	4611	577	0.5129	0.2577
4	256	1348	3362	326	0.607	0.4125
5	1335	21130	4102	521	0.5255	0.1654
6	17791	2588	3876	1217	0.5133	0.2786
7	2190	173	4816	763	0.5122	0.4317
8	1579	11369	4458	1213	0.6117	0.3512
9	1765	5337	4614	1749	0.4656	0.2217
10	328	12511	3617	819	0.5349	0.216

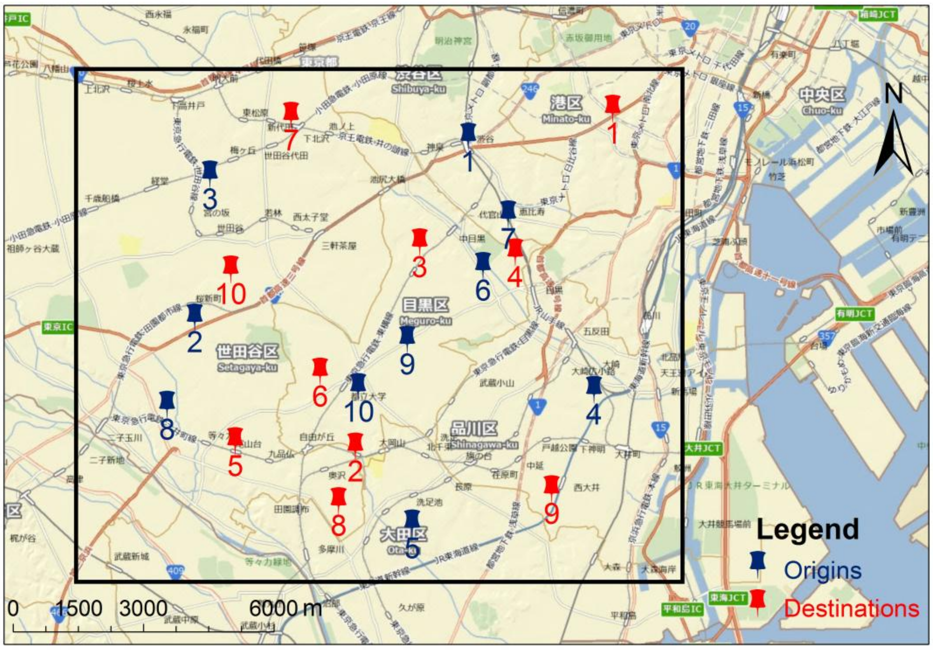


Figure 6. Locations of 10 selected OD pairs

The CSI results are presented in

Table 2. The average CSI for the hyperpath and real data (CSI_HP) for 10 OD pairs is 0.5110 while that for the SP and real data (CSI_SP) is 0.2924. This result is reasonable because the SP assumes very simple driver behavior (i.e., all drivers use the same route) and the assumption is far from what is observed in real cases. The results for CSI_HP, however, are unsatisfactory to argue that the observed routes chosen by taxi drivers are “quite” similar. This may be due to the limitation of the data we used, and more empirical studies with different data sources need to be carried out. In fact, it may be difficult to find a model which can fully explain the real behavior and we recommend using CSI as a comparative index. As an instance, Figure 7 (a) and (b) show the results of one of the OD pairs (from Shibuya to Roppongi, ID = 1) with low CSI_HP. It is seen that the real route choices are much more complex than the theoretically generated hyperpath. Cycles can sometimes be found for some revealed routes. One possible reason is that taxi drivers sometimes stopped for less than 120 seconds, which was the threshold we used to determine trip end points, and the route is actually a combination of multiple trips.

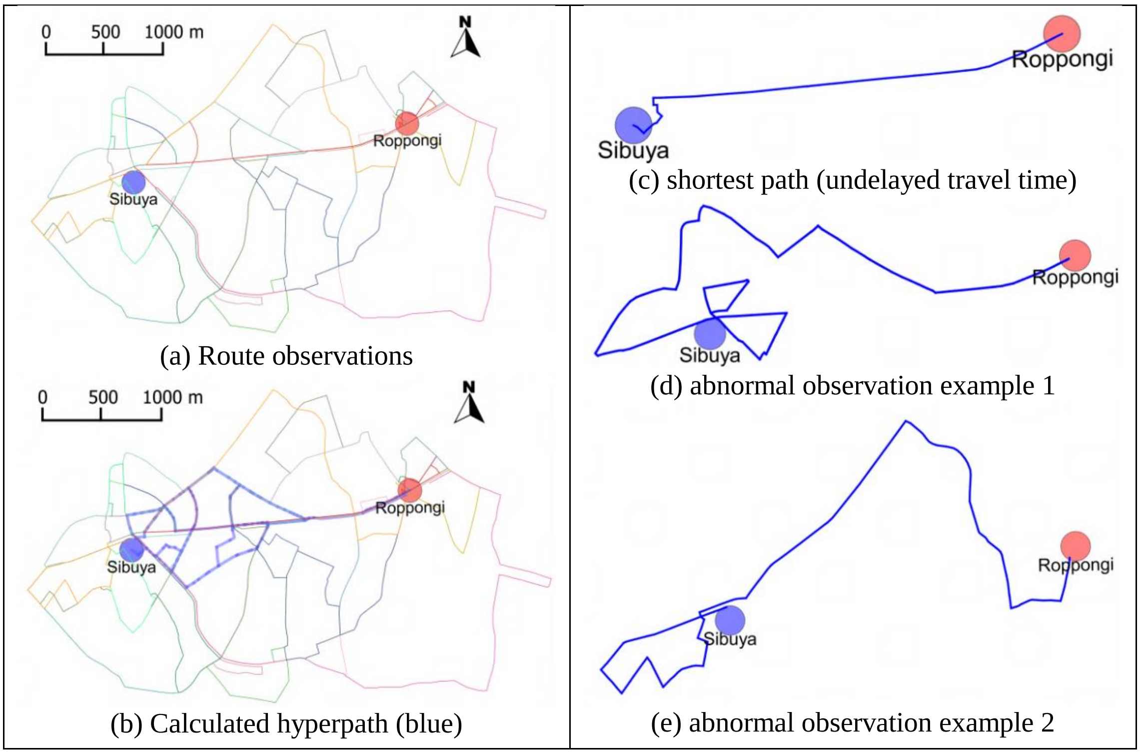


Figure 7. Observed routes and calculated hyperpath from Sibuya to Roppongi

The boarding or alighting information is unavailable in our data source; hence, some abnormal routes (Figure 7 (d), (e)) have been recorded for unknown reasons (the shortest path is shown in Figure 7 (c) for comparison). Some trip routes may actually consist of multiple trips. Additionally, the driving behavior of taxi drivers is supposed to be quite different when carrying passengers compared with when seeking passengers. Some route preferences may result from driver behavior when seeking passengers in areas with possibly high taxi demand. The customer-seeking behavior may partially explain some of the abnormal routes.

Because of the limitations of the data source, we found that many trip route data were missing and some abnormal routes had been taken into calculation. It would be more preferable to use collective routes for each individual driver instead of aggregated routes across different drivers.

5 CONCLUSION

This paper introduced a method of evaluating the behavioral reality of hyperpath routing and carried out an empirical study using taxi probe data for Tokyo. The proposed method, which uses a CSI, is an easy way to compare a theoretical link set based on routing algorithms with a revealed link set, especially when a large quantity of GPS data is available. The empirical results indicate that hyperpath routing is closer to the real route choice than shortest-path routing. Although no benchmark has been proposed for the similarity index so far, generally a routing algorithm with the larger CSI, as a comparative index of behavioral reality, would be recommended.

The proposed evaluation method is general since it is not limited to hyperpath routing. In our empirical study, only the SP routing is compared with hyperpath because SP is the most popular one. Other multi-path routing algorithms may also be compared if we are able to appropriately give weights for each link based on route choice models. Outside the context of vehicle routing, it is also applicable to evaluating the behavioral reality of route choice models.

There are limitations to the data that we used. First, Tokyo is a city with highly developed railway systems, which many commuters tend to use especially in rush hours, resulting in less route observations than expected. Second, taxis are owned by companies and shared among a group of taxi drivers, which makes it difficult to study the route choice behavior of individual drivers. Furthermore, the lack of passenger boarding or alighting information made it difficult to obtain trip-based routes. Additionally, the data we used might be biased in terms of the behavior of “taxi” drivers. In an active sense, they are generally more familiar with the roads and network status and are supposed to be aware of recurrent delays, and thus their choices are advisable to common drivers. In a passive sense, a taxi driver may tend to seek customers when the taxi is empty and thus tend to wander around areas with higher taxi demand. More empirical studies with richer data would be highly recommended to examine whether hyperpath routing would be ready for the market.

ACKNOWLEDGEMENTS

This research is supported by the Japanese Committee on Advanced Road Technology and by the Japan Institute of Construction Engineering. The authors would like to thank Haruo Ishida, Naohisa Okamoto, Kazuhiko Makimura and Toshiyuki Nakamura for providing the GPS probe data. The authors are also grateful to Michael Bell, Jan-Dirk Schmöcker and Ioannias Kaparias for their valuable comments.

REFERENCES

Dijkstra, E. (1959) A note on two problems in connection with graphs. Numerical Mathematics 1, 395-412.

Hart, P.E.; Nilsson, N. J. and Raphael, B. (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics SSC4, Vol. 4, No. 2, , pp. 100-107.

Fu, L. and Rilett, L. R. (1998) Expected shortest paths in dynamic and stochastic traffic networks. Transportation Research Part B: Methodological 32(7), 499-516.

Miller-Hooks, E. D. and Mahmassani, H. S. (2000) Least expected time paths in stochastic, time-varying transportation networks. Transportation Science 34 (2), 198-215.

Nie, Y. M. and Wu, X. (2009) Shortest path problem considering on-time arrival probability. Transportation Research Part B: Methodological, 43(6), 597-613.

Kaparias, I. and Bell, M.G.H. (2009) Testing a reliable in-vehicle navigation algorithm in the field. Intelligent Transport Systems 3(3), 314-324.

Chen, Y.Y, Bell, M.G.H. and Bogenberger, K. (2010) Risk-averse autonomous route guidance by a constrained A* search. Journal of Intelligent Transportation Systems 14(3), 188-196.

Chi-kang L. (1994) A Multiple-path routing strategy for vehicle route guidance systems. Transportation Research Part C: Emerging Technologies 2(3), 185-195.

Chen, Y.Y, Bell, M.G.H. and Bogenberger, K. (2007) Reliable pre-trip multipath planning and dynamic adaptation for a centralized road navigation system. IEEE Transactions on Intelligent Transportation Systems 8(1), 14-20.

Fosgerau, M., Frejinger, E. and Karlström, (2009) A. Route choice modeling without route choice. Proceedings of European Transport Conference.

Dial R. B. (1971) A probabilistic multipath traffic assignment model which obviates path

enumeration, Transportation Research 5(2), 83-111.
Nguyen, S. and Pallottino, S. (1988) Equilibrium traffic assignment for large-scale transit networks. European Journal of Operational Research 37(2), 176-186.
Spiess, H. and Florian, M. (1989) Optimal strategies: A new assignment model for transit networks. Transportation Research Part B: Methodological 23(2), 83-102.
Cominetti, R., Correa, J. (2001) Common bus lines and passenger assignment in congested transit networks. Transportation Science, Vol. 35(3), 115-121.
Bell, M.G.H. (2009) Hyperstar: A multipath A-star algorithm for risk averse vehicle navigation. Transportation Research Part B: Methodological 43(1), 97-107.
Bell, M.G.H., Trozzi, V., Hosseinloo, S.H. and Fonzone, A. (2012) Time-dependent Hyperstar algorithm for robust vehicle navigation in time-dependent stochastic road networks. Transportation Research Part A: Policy and Practice 46(5), 790-800.
Ma, J.S., Fukuda, D. and Schmöcker, J.D. (in press) Faster hyperpath generating algorithms for vehicle navigation. Transportmetrica.
Dechter, R. and Pearl, J. (1985) Generalized best-first search strategies and the optimality of A*. Journal of the ACM 32(3), 505-536.
Sheffi, Y. (1984) Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. Prentice-Hall, New Jersey.
Morikawa, T., Miwa, T., Kurauchi, F., Yamamoto, T. and Kobayashi, K. (2005) Drivers route choice behaviour and its implications on network simulation and traffic assignment. Simulation Approaches in Transportation Analysis. Springer Science Business Media, Inc. New York, 135-136.
Zhu S.J. and Levinson, D. (2012) Do people use the shortest path? An empirical test of Wardrop's first principle. Paper presented at the 91st TRB Annual Meeting. Washington D.C.
Bekhor, S., Ben-Akiva, M. and Ramming, (2006) M. Evaluation of choice set generation algorithms for route choice models. Annals of Operations Research 144(1), 235-247.
Fafieanie, M. (2009) Calibrating Route Set Generation by Map Matching GPS Data. Master Thesis, University of Twente.
Frejinger, E.; Bierlaire, M. and Ben-Akiva, M. (2009) Sampling of alternatives for route choice modeling, Transportation Research Part B: Methodological 43(10), 984-994.
Cascetta E., Nuzzolo, A., Russo F. and Vitetta A. (1996) A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. Transportation and Traffic Theory: Proceedings from the 13th ISTTT, 697-711.
Vovsha, P. and Bekhor, S. (1998) The link-nested logit model of route choice: overcoming the route overlapping problem. Transportation Research Record: Journal of Transportation Research Board 1645, 133-142.
Akgün, A., Erkut, E. and Batta, R. (2000) On finding dissimilar paths. European Journal of Operational Research 121(2), 232-246.
Lim, Y. and Kim, H. (2005) A shortest path algorithm for real road network based on path overlap. Journal of the Eastern Asia Society for Transportation Studies 6, 1426-1438.
Zijpp van der, N. J. and Catalano, S. Fiorenzo (2005) Path enumeration by finding the constrained K-shortest paths. Transportation Research Part B: Methodological 39(6), 545-563.
Tan, P.-N., Steinbach, M. and Kumar, V. (2006) Introduction to Data Mining. Pearson Addison Wesley, Boston.
Ben-Akiva, M. and Bierlaire, M. (1999) Discrete choice methods and their applications

to short-term travel decisions. In: Hall, R.W. (ed.) Handbook of Transportation Science, Kluwer, 5-34.

Ben-Gal I. (2005) Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, New York.

Terriberry T. B. (2007) Computing higher-order moments online, Online: http://people.xiph.org/~tterribe/notes/homs.html.