Testing the proportionality condition with taxi trajectory data

Article *in* Transportation Research Part B Methodological · May 2017

DOI: 10.1016/j.trb.2017.05.003

CITATION
1

READS
463

3 authors:

Jun Xie
Southwest Jiaotong University
11 PUBLICATIONS **57** CITATIONS
SEE PROFILE

Yu Nie
Northwestern University
94 PUBLICATIONS **1,920** CITATIONS
SEE PROFILE

Xiaobo Liu
Southwest Jiaotong University
60 PUBLICATIONS **375** CITATIONS
SEE PROFILE

Transportation Research Part B 104 (2017) 583–601

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Testing the proportionality condition with taxi trajectory data

Jun Xie[a], Yu (Marco) Nie[b,*], Xiaobo Liu[a]

[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan 610031, China
[b] Department of Civil and Environmental Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA

ARTICLE INFO

Article history:
Received 16 December 2016
Revised 2 May 2017
Accepted 3 May 2017
Available online 13 May 2017

Keywords:
User equilibrium
The proportionality condition
Paired alternative segments
Taxi trajectory data

ABSTRACT

The proportionality condition has been widely used to produce a unique path flow solution in the user equilibrium traffic assignment problem. However, it remains an open question whether and to what extent this condition accords to real travel behavior. This paper attempts to validate the behavioural realism of the proportionality condition using more than 27 million route choice observations obtained by mining a large taxi trajectory data set. A method is first developed to uncover more than three hundred valid paired alternative segments (PAS), on which the proportionality condition is tested by performing linear regression analysis and chi-square tests. The results show that the majority of the PASs tested (up to 85%) satisfy the proportionality condition at a reasonable level of statistical significance.

1. Introduction

Traffic assignment is one of the most widely used tools in transportation planning. While the standard user-equilibrium (UE) traffic assignment model (Beckmann et al., 1956) uniquely determines the total flow on each network link, it is generally unable to produce unique route flows or multiple-class link flows (see e.g. Chapter 3 Sheffi, 1985). However, many useful planning applications, such as *select link analysis*, do use route flows. The lack of uniqueness thus can result in inconsistency and instability in the analyses that depend on these solutions (Lu and Nie, 2010; Bar-Gera et al., 2012).

Additional assumptions are needed to assure that route flows and multi-class link flows are uniquely determined. A generally accepted assumption is to find the *most likely route flows*, which is the solution to an entropy maximization problem (Rossi et al., 1989). Bar-Gera and Boyce (1999) derive the proportionality condition from entropy maximization, which states that *flows should be distributed to two paired alternative segments (PAS) with equal cost according to the same proportion regardless of user class, origin or destination*. By enforcing this condition, Bar-Gera (2010) develops a new assignment algorithm - known as Traffic Assignment by Paired Alternative Segments (TAPAS) - that is capable of determining *nearly unique* route flows. By “nearly unique” we highlight the fact that the proportionality condition is not sufficient to ensure entropy maximization. As pointed out by Bar-Gera (2006) and Borchers et al. (2015), *higher-order* proportionality conditions may be needed to secure a solution to the entropy maximization problem. Nevertheless, empirical evidence suggests that higher-order proportionality conditions rarely arise in real-life networks, and that a solution satisfying the proportionality condition offers approximation of high quality to the maximum entropy solution for most practical purposes (Bar-Gera, 2006).

* Corresponding author.
E-mail address: y-nie@northwestern.edu (Y. (Marco) Nie).

http://dx.doi.org/10.1016/j.trb.2017.05.003
0191-2615/

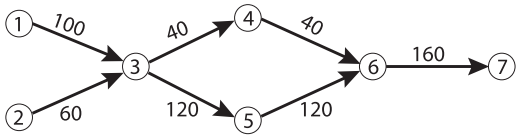


Fig. 1. Illustration of proportionality with one set of paired alternative segments (link labels are link flows at UE).

Table 1
Multiple UE path flow solutions for the UE link flow solution reported in Fig. 1 (h^* is the unique path flow solution that also satisfies proportionality).

ID	Path	h^*	h_1	h_2
1	1-3-4-6-7	25	40	0
2	1-3-5-6-7	75	60	100
3	2-3-4-6-7	15	0	40
4	2-3-5-6-7	45	60	20

Not surprisingly, the concept of proportionality has been gradually introduced into mainstream transportation planning software[1] since 2010 (Boyce et al., 2010; Bar-Gera et al., 2012).

One may argue that the proportionality condition is merely an artificial instrument for producing a unique solution that can be consistently used as a benchmark for scenario comparison. In theory, Lu and Nie (2010) showed that a unique UE route flow may be determined by maximizing any objective function that possesses certain properties. Among these competing "unique" route flow solutions, there seems no sound reason (other than for the sake of convenience) to prefer one to the other. It remains an open question whether and to what extent the proportionality condition, which is the choice of the current practice, accords to real travel behavior.

In light of the above, we document in this paper the first attempt to validate the behavioral realism of the proportionality assumption. The validation study lags behind in the literature mainly because it requires a large amount of route choice data, which is not readily available from traditional data sources. This critical gap is filled in this study thanks to the availability of GPS log data collected by the GPS receiver installed in taxis, referred to as *taxi trajectory data* hereafter. Taxi trajectory data have several appealing characteristics that make it a desirable choice for studying human mobility patterns in general (see e.g. Li et al., 2011; Yuan et al., 2010; Zhang et al., 2016; Wang et al., 2014; Yue et al., 2009; Liu and Qu, 2016; Zheng et al., 2013; Liu et al., 2010), and route choice behavior in particular. First, it does not have the privacy issue, because no personal information can be recovered from such data (unlike mobile phone or social media data). Second, taxi drivers can be seen as "experts" in path planning (Lin et al., 2015) because they usually know the road network and traffic conditions well. Thus, the path taken by a taxi driver is usually the "optimal" choice, which is a prerequisite for applying the proportionality condition. Third, thanks to passenger boarding information, it is relatively easy to break the continuous trajectory data into taxi trips with a specific OD pair. Last but not least, the real travel time of a taxi through any path segment can be easily estimated based on the position and time information, which is essential in guaranteeing equal travel time on selected PASs.

The reminder of this paper is organized as follows. Section 2 briefly reviews the proportionality condition and Section 3 introduces the proposed procedure for its validation. Section 4 presents the details of the taxi trajectory data, as well as the necessary data processing procedures. Section 5 reports and discusses the validation results, and Section 6 concludes the paper with a summary of main findings.

2. Concept of proportionality

If we assume that paths with identical costs are equally attractive to all travelers, then their probability of choosing either segment of any two paired alternative segments (PAS) with equal cost would be identical, regardless of their origin, destination and user class (provided the class identification does not affect a traveler's evaluation of segment costs). Accordingly, the *most likely* UE path solution, or the UE path flow solution that has the highest probability to realize (or be observed), must satisfy the proportionality condition.

Fig. 1 illustrates the concept of proportionality. The link flows shown in Fig. 1 satisfy the user equilibrium (UE) condition, that is, the paired alternative segments (3-4-6 and 3-5-6) both have positive flows and exactly the same cost. Given the UE link flows, multiple path flow solutions exist that all produce the same link flow solution. Table 1 lists three feasible UE path flow solutions, but only the path flow solution h^* is determined according to the proportionality condition. The reader can verify that the ratio between the flow on path 1 and 2 (as well as path 3 and 4) is 1:3, equal to the ratio of the flow

[1] see e.g. https://www.inrosoftware.com/en/news-and-events/posts/traffic-assignment-choices-at-emme-4-1/; http://www.trafikanalysforum.se/sites/default/files/bibliotek/latest_developments_in_visum_-_klaus_nokel.pdf.

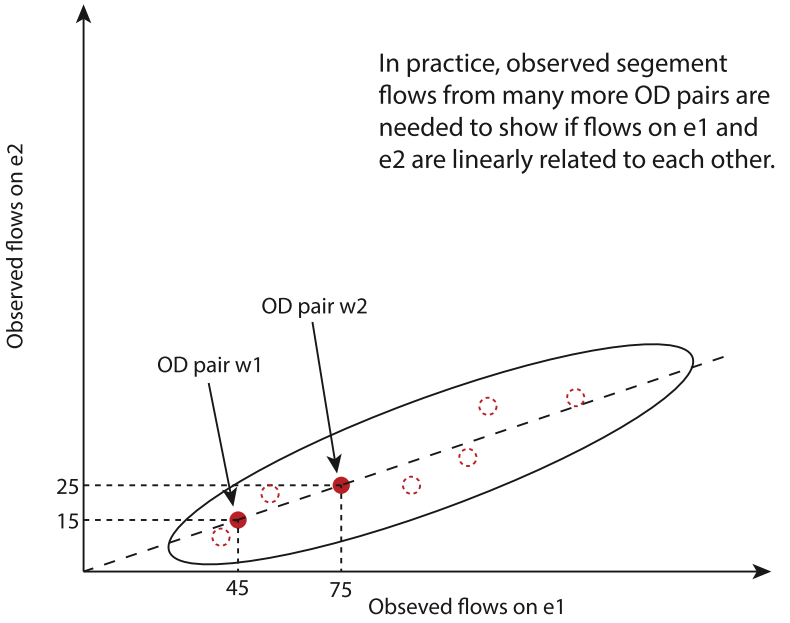


Fig. 2. Proportionality test based on observed segment flows.

on the two PASs (3-4-6 and 3-5-6)[2]. In other words, the distributions of flows between 3-4-6 and 3-5-6 is always 1:3 for both O-D pairs 1-7 and 2-7.

3. Testing methods

To explain our idea for testing the proportionality condition, let us use $p(e_1, e_2)$ to denote a PAS in a valid PAS set $\mathbf{P}$, where e_k represents segment $k = 1, 2$. Also, let $x_{e_k}^{p,w}$ denote the number of recorded trips associated with O-D pair w that use the segment e_k of PAS p; and $\mathbf{x}_{e_k}^{p}$ the vector of the recorded trip numbers for all O-D pairs using e_k of PAS p. Once $\mathbf{x}_{e_k}^{p}$ is obtained for both $k = 1, 2$ of a given $p \in \mathbf{P}$, a statistical analysis can be performed to test the strength of the linear relation between the two vectors of $\mathbf{x}_{e_1}^{p}$ and $\mathbf{x}_{e_2}^{p}$.

For example, consider Fig. 1 again, and let O-D pair $w1$ and $w2$ be $2-7$ and $1-7$ respectively, and $e_1 = 3-5-6, e_2 = 3-4-6$. Then $\mathbf{x}_{e_1}^{p} = [45, 75], \mathbf{x}_{e_2}^{p} = [15, 25]$. The solid circles in Fig. 2 are plotted using $\mathbf{x}_{e_1}^{p}$ and $\mathbf{x}_{e_2}^{p}$ as x and y coordinates, respectively. Clearly, the straight line that connects the two solid circles has a slop of 1/3 and an intercept of 0, which means the proportionality condition is satisfied perfectly among these two O-D pairs.

Intuitively, for the two observed flow vectors $\mathbf{x}_{e_k}^{p}, k = 1, 2$, testing the proportionality may be viewed as testing the following linear relationship

$$\mathbf{x}_{e_2}^{p} = \beta \mathbf{x}_{e_1}^{p} + \gamma \tag{1}$$

Note that the segment with higher flows on average are always treated as segment 1 (the independent variable). Accordingly, the extent to which the proportionality condition is satisfied may then be measured by the goodness-of-fit to linear relationship (the R^2 value). However, this straightforward approach has a couple of shortcomings. First, the standard linear regression are based on the assumption that the independent variable is normally distributed. In the current context both variables ($\mathbf{x}_{e_1}^{p}$ and $\mathbf{x}_{e_2}^{p}$) are count variables, therefore more likely to follow Poisson or negative binomial distribution. When the magnitude of each component in $\mathbf{x}_{e_1}^{p}$ is small, this drawback is especially acute, making the linear regression approach much less reliable. Second, for PASs with extreme ratios (i.e. the vectors fit to a line with near zero slope), achieving a satisfactory goodness-of-fit is much more difficult than the cases when the ratio falls into a moderate range.

To mitigate the above shortcoming, we propose to supplement the linear regression analysis with Pearson's chi-square test (Pearson, 1900), which is widely used to determine whether observed frequencies are significantly different from expected frequencies. In our context, for a given PAS p with n O-D pairs and segment flow vectors $\mathbf{x}_{e_k}^{p} = \{x_{e_k}^{p,w_1}, \ldots, x_{e_k}^{p,w_n}\}, k = 1, 2$, testing proportionality amounts to the test of the following null hypothesis with a degree of freedom $n-1$:

$$H_0: \quad \frac{x_{e_1}^{p,w_1}}{x_{e_2}^{p,w_1}} = \frac{x_{e_1}^{p,w_2}}{x_{e_2}^{p,w_2}} = \ldots = \frac{x_{e_1}^{p,w_n}}{x_{e_2}^{p,w_n}} \tag{2}$$

[2] The segment flow is formally calculated by formulation (12) in (Bar-Gera, 2010).

Table 2
Description of COST database.

Set	Period	#Taxis	Total GPS Points (million)	Total GPS Points with passengers (million)
1	1/1/2011 - 1/31/2011	14171	658.71	242.98
2	1/15/2015-1/21/2015	15726	420.19	194.13
3	3/25/2015-3/31/2015	15665	389.52	183.37
4	5/25/2015-5/31/2015	15414	369.50	174.33
5	7/25/2015-7/31/2015	15608	334.42	145.19
6	9/24/2015-9/30/2015	15564	358.16	151.78
7	11/24/2015-11/30/2015	15425	352.61	148.84

The chi-square statistic χ^2 is computed as

$$\chi^2 = \sum_{i=1}^{n}\left(\sum_{k=1}^{2}\frac{\left(x_{e_k}^{p,w_i} - \bar{x}_{e_k}^{p,w_i}\right)^2}{\bar{x}_{e_k}^{p,w_i}}\right), \tag{3}$$

$$\bar{x}_{e_k}^{p,w_i} = \left(x_{e_1}^{p,w_i} + x_{e_2}^{p,w_i}\right)\frac{\sum_{j=1}^{n} x_{e_k}^{p,w_j}}{\sum_{j=1}^{n} x_{e_2}^{p,w_j} + \sum_{j=1}^{n} x_{e_1}^{p,w_j}}, k = 1, 2. \tag{4}$$

If $\chi^2 < \chi^2_{(n-1)\alpha}$ (where $\chi^2_{(n-1)\alpha}$ is the critical value at a significance level of α), the null hypothesis is not rejected, implying that no significant difference can be found among the proportions of different O-D pairs at α. We note that the chi-square test usually requires $\bar{x}_{e_k}^{p,w_i}$ be greater than certain threshold (usually five or ten) for all *i, k*. Five is adopted in this study.

The chi-square test has its own drawbacks. For one thing, it does not give much information about the goodness-of-fit, or the strength of linear relationship. In addition, the chi-square statistic is affected by the size of O-D pairs (n). For cases with a large number of O-D pairs, the chi-square test tends to reject the null hypothesis more easily. This could happen, as our experiments confirmed, even when the linear regression suggests satisfactory fitness to the linear relationship.

In light of the above discussions, we propose to test the proportionality assumption by jointly considering both the linear regression and the chi-square test. Specifically, a PAS that satisfies either of the following two conditions is considered *proportionality conforming*.

$$R^2 > \bar{R}^2;\ \chi^2 < \chi^2_{(n-1)0.05} \tag{5}$$

where $\bar{R}^2$ is a preselected lower bound for the goodness-of-fit and $\chi^2_{(n-1)0.05}$ is the chi-square critical value at a significance level of 0.05 and a degree of freedom of $n-1$. A PAS that satisfies both conditions is considered *proportionality super-conforming*.

Testing (1) or (2) using real-world data poses a number of practical challenges. For one thing, meaningful PASs are not easy to come by. PASs may be obtained by solving an UE traffic assignment problem using an algorithms such as TAPAS or a post-process procedure such as described in Xie and Xie (2016). However, because travel forecasting models are based on strong route choice assumptions that may not always be fulfilled in reality, PASs found in this way may be incompatible with observed route choice patterns. Another issue has to do with observing O-D specific segments flows, which can only be aggregated from path flows. While directly observing path flows is no easy task, having a large enough sample that would cover any given PAS with meaningful amount of segment flows is even more challenging. In what follows, we will explain how to overcome the above challenges using a large sample of taxi GPS trajectory data.

4. Data and processing methods

4.1. Overview

The taxi data set used in this study is a subset of the so-called "City Of Shenzhen Taxi" (COST) database, which includes ten separate raw GPS trajectory data sets, each collected in a continuous period of time between 2011 and 2015. Each GPS trajectory includes a time series data of location (latitude/longitude), time stamp, instantaneous speed, heading, and passenger state (on or off). The reader is referred to Nie (2017) for a more detailed description of the COST data. In this study, seven data sets are used, as detailed in Table 2. Fig. 3 shows the Shenzhen Transportation Planning Network, which consists of 21,114 road segments and 3,199 traffic analysis zones (TAZ). The TAZs will be used to identify the origin and the destination of each individual taxi trip, and the GPS trajectory of each trip will then be matched to a sequence of road segments in the network.

The data shown in Table 2 are separated into two sample groups: Set 1 (the one-month data collected in 2011) as Sample 1, and Sets 2 - 7 (the six-week data collected in six weeks in 2015) as Sample 2. The test for the proportionality condition will be carried out for each data sample independently, because transportation infrastructure and travel patterns

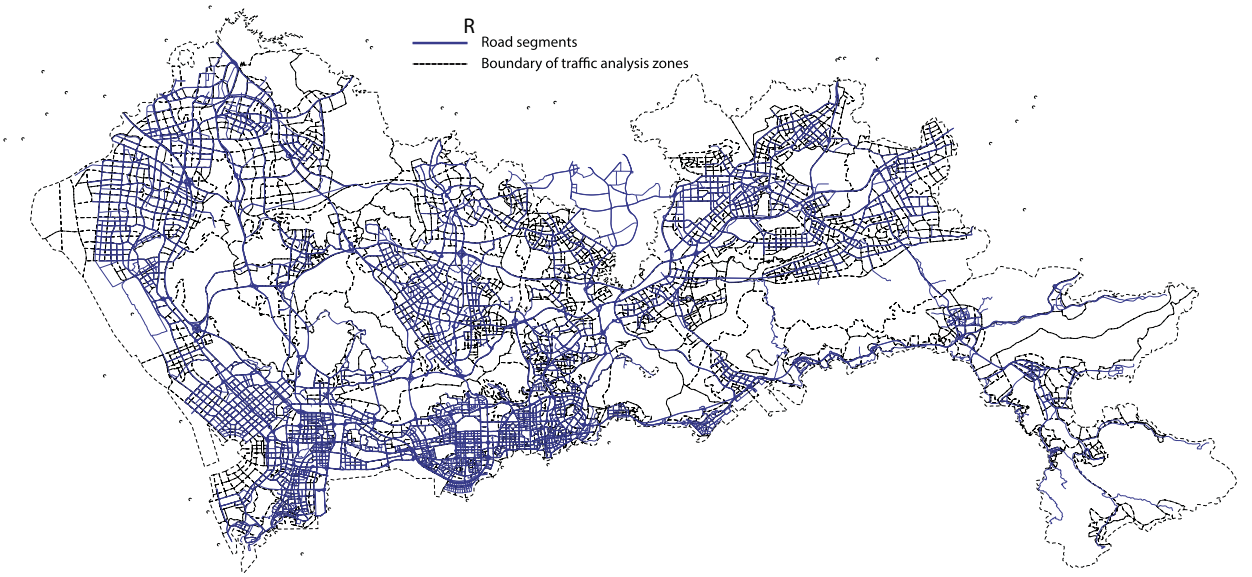


Fig. 3. GIS data of the city of Shenzhen.

might have been significantly changed in a period of four years[3]. For each sample, the following procedures are proposed to obtain sufficient segment flow data for the test.

1. **Trip identification:** Break the GPS trajectory into separate "trips", each corresponding to a unique passenger state (occupied or vacant, but only occupied trips are used in this study), an origin TAZ, a destination TAZ, a starting and an ending time.
2. **Map matching:** Match the trajectory of each occupied (i.e. the trip with a passenger) trip onto segments in the road network, and convert each trip into a path that consists of a sequence of consecutive road segments.
3. **PAS generation:** Generate candidate paired alternative segments (PASs) from all generated paths.
4. **PAS selection:** Create a valid PAS set by filtering out PASs not suitable for the proportionality test.
5. **Proportionality test:** For each valid PAS p, count the number of paths that connect all O-D pairs and use segment e_k as the corresponding segment flow, denoted as $\mathbf{x}^p_{e_k}, k = 1, 2,$ where the length of each vector equals the number of O-D pairs. Then a linear regression and a chi-square test are performed for the two vectors $\mathbf{x}^p_{e_k}, k = 1, 2$.

The details of trip identification is omitted here because they are discussed in Nie (2017). Note that we exclude vacant trips because these trips likely reflect more of drivers' search behavior than of passengers' route choice behavior. In what follows we will discuss Steps 2 - 4 in details.

4.2. Map matching

A trip generated in the trip identification step is typically a sequence of raw GPS points, which must be projected to the road network to create a path that consists of road segments. This process is often known as map matching in the literature, for which numerous methods have been proposed in the last two decades. The reader is referred to Zheng (2015) for a recent review of these methods. For completeness, a simple but well-performing map matching method implemented in this study is briefly described in Algorithm 1.

The method described in Algorithm 1 consists of two steps. The first step, known as *correct link identification*, matches each GPS point first to a link in the network and then associate it with the closer end node of the link (cf. Line 5-19 in Algorithm 1). Given a GPS point b, we define for each candidate link ij adjacent to b a composite index I_{ij}, which combines the distance (i.e., the distance between b and ij) and the directional information (i.e., the difference between the vehicle's moving direction and the link direction). The smaller is the index I_{ij}, the more likely is it that b falls on link ij. Once b is matched to the link with the smallest I_{ij}, it is projected to a closer end of the link (i or j). Accordingly, the time when vehicle arrived at this node can be estimated[4] (cf. Line 17 in Algorithm 1). To summarize, the first step of the matching algorithm produces, for a given trip, a list of nodes with corresponding arrival times (denoted by $\mathbf{N}_g$).

[3] The GDP of Shenzhen is still growing at the pace of about 8.4% in 2015, see https://en.wikipedia.org/wiki/Shenzhen.

[4] The node arrival time will be set by the closest GPS point if there are several observations for the same node.

Algorithm 1 Map Matching.

Input: The set of trips **G** obtained from trip generation; the network.
Initialize: Initialize path set $\mathbf{H} = \varnothing$.
for each trip $g \in \mathbf{G}$ such that $|g| \geq 10$ **do**
  correct link identification: Line 5 - 19 (cf. Fig 4(a)).
  for each point $b \in g$ **do**
    Let a and c denote the preceding and subsequent points of b, respectively.
    Collect all links within a certain distance, i.e., $\bar{d}$, to point b and save them in $\mathbf{C}_b$.
    Calculate the angles of ab and bc with the horizontal line, i.e., α and β, respectively.
    for each link $ij \in \mathbf{C}_b$ **do**
      Calculate the shortest distance d_{ij}^b from point b to the closest point l in link ij.
      Calculate the angle of line il with the horizontal line, i.e., α_{ij}.
      Calculate the angle of line lj with the horizontal line, i.e., β_{ij}.
      Calculate the indicator $I_{ij} = \gamma \cdot d_{ij}^b + (1-\gamma) \cdot \left(|\alpha - \alpha_{ij}| + |\beta - \beta_{ij}|\right)$, $\gamma \in (0,1)$.
    end for
    Choose the link with minimum I_{ij} in $\mathbf{C}_b$ as the correct link for b.
    Project point b to the closer end of the link ij, say i.
    Estimate the time when the vehicle arrived at node i by $t_i = t_b \pm \frac{d_{il}}{(d_{ab}+d_{bc})/(t_c - t_a)}$.
    Add node i with a time label into the matched node set $\mathbf{N}_g$ from the back.
  end for
  Determination of most likely segments: Line 21 - 28
  for each node i in $\mathbf{N}_g$ **do**
    Let j be the subsequent node of i in $\mathbf{N}_g$.
    Compute the shortest path segment e_{ij} from i to j with an A-Star algorithm.
    Compute the average travel speed through e_{ij} by $v_{e_{ij}} = d_{e_{ij}}/(t_j - t_i)$.
    Add links in segment e_{ij} into path h_g with speed labels.
  end for
  Add matched path h_g into path set **H**.
end for
Output: A matched path set **H** where each path $h_g \in \mathbf{H}$ consists of a list of links with travel speed labels.

(a) Identification of the correct link (b) Example of map matching for a trip

Fig. 4. Illustration of map matching.

The second step determines the segment that connects each pair of adjacent nodes to form a path (cf. Line 21-28 in Algorithm 1). Following Wenk et al. (2006), the shortest segment connecting the two nodes is used. While more sophisticated methods exist (Quddus et al., 2006; Lou et al., 2009; Newson and Krumm, 2009), our experiments indicate the shortest segment matches the trajectory satisfactorily. This may be due to two reasons. First, the average sampling interval in our data is relatively short, ranging from about 22 seconds in Sample 1 to 12 - 14 seconds in Sample 2 (Nie, 2017). Short sampling intervals mean that consecutive points are close to each other, often appearing on the same or adjacent links. Second, even if these points are separated by a few links, they are likely to be connected by direct (hence short) paths rather than those with detours (Lou et al., 2009). An A-Star shortest path algorithm is employed to determine the segment that connects two adjacent nodes. Since each node has an arrival time label, the average speed on the connecting segment can also be easily estimated (cf. Line 24 in Algorithm 1). Fig. 4(b) shows how a trip is matched to links in the network through the above procedures.

Except the key steps outlined in Algorithm 1, a few remarks are in order here on detecting and excluding outliers. First, a taxi may stay at the same location for a relative long period - due to congestion, traffic signal or other delays - which may cause quite a few points to concentrate in a small area. These cases often render the directional information useless or even misleading. To avoid misidentification, all succeeding points that are "too close" (defined as <15 meter) to the present point is discarded in the first step. Second, due to close proximity of roads in some areas (e.g. freeway interchanges), mismatching single points to wrong road segments is often unavoidable. In light of this, we propose to compare the average segment speed to the posted speed limit. If one of the points used to construct the segment is mismatched, the shortest distance between the two points in the network is likely to be significantly longer than the true distance, leading to unusually high average speed. Once such abnormality is detected, a trial-and-error procedure is used to remove either point and connect the remaining one to the point further up- or down-stream, until all segments in the path have an average travel speed within the posted speed limit.

4.3. PAS generation

All paths obtained by map matching are saved in the path set **H**. PASs are then generated by scanning each pair of paths with the same OD pair in **H**. Algorithm 2 describes the PAS generation procedure.

Algorithm 2 PAS Generation.

1: **Input:** Path set **H** obtained by map matching.
2: **Initialize:** Initialize the PAS index $\kappa = 1$; initialize the PAS list $\mathbf{P} = \varnothing$, initialize the PAS set $\mathbf{H}_w = \varnothing$ for each OD pair w.
3: **for** each path $h_1^w \in \mathbf{H}$ **do**
4: **for** each path $h_2^w \in \mathbf{H}_w$ **do**
5: Initialize the status of each node i (denoted by m_i) in path h_1^w and h_2^w as $m_i = 0$.
6: Mark each node in h_1^w as $m_i = 1$.
7: Set the PAS set $\mathbf{M} = \varnothing$.
8: Scan each link ij in h_2^w from the destination back to the origin:
9: **if** $m_i = 0$ and $m_j = 1$ **then**
10: j is recorded as the ending node of a new PAS p_κ.
11: **else if** $m_i = 1$ and $m_j = 0$ **then**
12: i is recorded as the beginning node of PAS p_κ and set $\kappa = \kappa + 1$.
13: Add p_κ into the set **M**.
14: **else if then**
15: Continue.
16: **end if**
17: **for** each p_κ in **M do**
18: Build the segments e_1 and e_2 of p_κ from h_1^w and h_2^w, where $e_1 \subseteq h_1^w$ and $e_2 \subseteq h_2^w$.
19: **if** $p_\kappa \notin \mathbf{P}$ **then**
20: Add p_κ into **P**.
21: Add w into $\mathbf{W}_{p_\kappa}$; initialize $x_{e_1}^{p_\kappa,w} = 1$ and $x_{e_2}^{p_\kappa,w} = 1$.
22: **else if** $p_\kappa \in \mathbf{P}$ **then**
23: **if** $w \notin \mathbf{W}_{p_\kappa}$ **then**
24: Add w into $\mathbf{W}_{p_\kappa}$; initialize $x_{e_1}^{p_\kappa,w} = 1$ and $x_{e_2}^{p_\kappa,w} = 1$.
25: **else if** $w \in \mathbf{W}_{p_\kappa}$ **then**
26: $x_{e_1}^{p_\kappa,w} = x_{e_1}^{p_\kappa,w} + 1$.
27: **end if**
28: **end if**
29: **end for**
30: **end for**
31: Add h_1 into $\mathbf{H}_w$.
32: **end for**
33: **Output:** A list of PAS **P**. Each PAS $p_\kappa \in \mathbf{P}$ is associated with a list of OD pairs $\mathbf{W}_{p_\kappa}$, and the recorded path flow counts variables $x_{e_1}^{p_\kappa,w}$ and $x_{e_2}^{p_\kappa,w}$.

A few details of the algorithm warrant further explanation. First, note that we only need to scan a path pair when both are associated with the same O-D pair, because only such a path pair can form a meaningful PAS for the purpose of our test[5] (cf. Line 3-4 and 31 in Algorithm 2). Second, more than one PAS may be generated by scanning one pair of paths. Thus, a set **M** is introduced to temporarily store all PASs generated from the same scan (cf. Lines 7, 13 and 17 in Algorithm 2).

[5] The proportionality condition is tested based on flow aggregated for the same O-D pair, see Fig. 2.

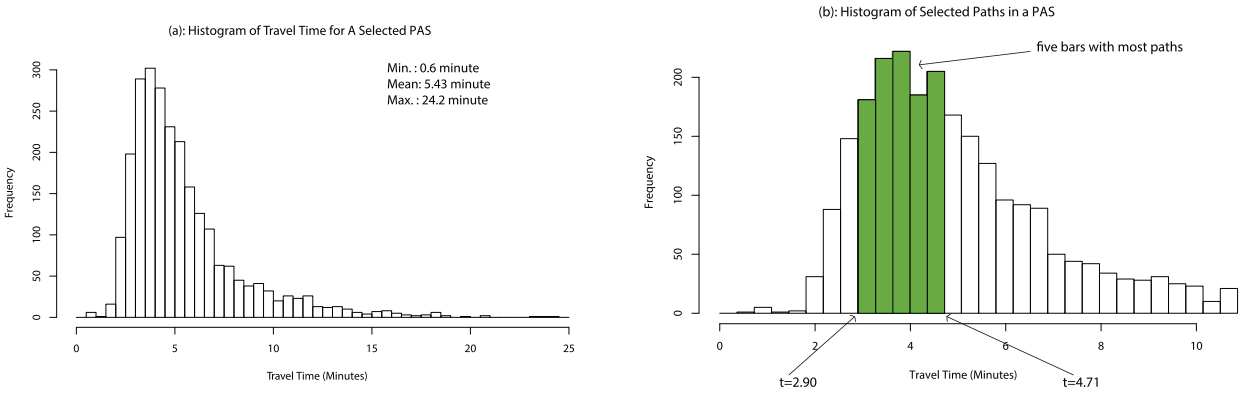


Fig. 5. Example of selecting paths according to travel time on a PAS.

The third issue has to do with properly counting the number of paths that pass each segment of a PAS. For a given PAS p_κ that is first identified with an O-D pair w, the OD-specific segment flows are initialized as $x_{e_1}^{p_\kappa,w} = 1$ and $x_{e_2}^{p_\kappa,w} = 1$ (cf. Line 21 and 24 in Algorithm 2); otherwise, only the flow on the segment that the path h_1^w actually passes will be updated (cf. Line 26 in Algorithm 2).

4.4. PAS selection

Not all PASs generated from the previous section can be used for testing the proportionality condition. Some of these PASs are rarely used (created by occasional route choices associated with non-recurrent conditions such as accidents) or simply results of misidentification. These PASs must be first excluded because the segments are not really considered competitive alternatives by taxi drivers. In addition, PASs with very short segments are not desirable because (1) they typically represent trivial alternatives, and (2) correctly identifying such PASs through trajectories is difficult.

Those candidate PASs, which are long enough and frequently used, will then be subject to further scrutiny. By definition the two segments of each PAS should offer the identical travel cost at equilibrium. As straightforward as it sounds, this requirement is difficult to fulfill in real data for a couple of reasons. First, measuring "travel cost" requires the knowledge of each individual's utility function, which is not directly observable. Instead, the segment travel time is adopted in this study as a surrogate because it is the most important element in the utility of a trip and is often correlated with the trip distance. Yet, the segment travel time may vary significantly within a day, and aggregating data over fine time intervals (such as 15 to 30 minutes) to avoid temporal variations may greatly increase the demand for data. Last but not least, since imperfection is a hallmark of real data, one has to accept segment travel times within a range as approximately identical.

With the above discussions in mind, we select the PASs used for testing the proportionality condition by the following rules.

Rule 1 All PASs with an average segment length shorter than 0.5 mile are filtered out.

Rule 2 For a given candidate PAS $p(e_1, e_2)$, only a subset of the paths with similar travel times will be kept for analysis. These paths are called *valid* paths for $p(e_1, e_2)$. To determine the valid paths, the following method is used.

1. Let the travel time of any path that use either segment of PAS $p(e_1, e_2)$ be in $\Omega = [t_0^p, t_1^p]$. Divide its subset $[t_0^p, 2E_p]$ into B intervals, where E_p is the average of all observed travel times for the PAS.
2. Let r_m be the number of paths whose travel time falls into the mth interval, and $R_m^l = \sum_{j=m}^{m+l} r_j$, $m \in [1, B]$, $m + l \leq B$, where l is a given parameter that measures the tolerance on travel time variations. Find $m^* = \arg\max_m R_m^l$.
3. All paths that fall into $m^*, ..m^* + l$ are then considered as valid paths.

Fig. 5 demonstrates the idea. In the example, the segment travel times ranges from a minimum of 0.6 minutes to a maximum of 24.2 minutes. With $B = 30$ and $l = 4$, all paths with a travel time between 2.9 to 4.71 are accepted as valid paths. Clearly, more restrictive parameters (large B and small l) would yield a set of valid paths with travel times closer to each other. Yet, such sets would also be smaller, endangering the statistical significance of the test. Our experiments indicate that $B = 30$ and $l = 4$ strike a reasonable balance between the two concerns.

Rule 3 The valid paths will be further aggregated based on their O-D pairs. An O-D pair is considered as a *valid* O-D pair for the PAS if it has enough valid paths covering each of the two segments. Specifically, according to the requirement of chi-square test, for each valid O-D pair w the condition $E(x_{e_1}^{p,w}) \geq 5$ and $E(x_{e_2}^{p,w}) \geq 5$ must be satisfied[6].

[6] Rule 3 effectively eliminates PASs with one and only one segment that is deemed to have too few flow observations. The rule may introduce biases because these PASs may be more likely to deviate from proportionality. Yet, including these PAS is also problematic because the chi-square test will be rendered practically invalid (the rule of thumb dictates that each count must be no less than 5). More data are needed to alleviate the potential impact of this problem on the test.

Table 3
Preliminary results of trip identification, map matching and PAS generation.

Dataset	Trips	Matched Path	Generated PAS
Sample 1 (2011)	8,521,331	4,647,656	666,177
Sample 2 (2015)	18,486,794	9,733,779	1,499,880

Table 4
Changes on the numbers of selected PAS.

Dataset	No Rule	Rule 1	Rule 1 & 4	Rule 1,2,3 & 4
Sample 1 (2011)	666,177	522,447	26,133	80
Sample 2 (2015)	1,499,880	1,210,775	58,257	291

Table 5
Statistics on PASs selected for the analysis.

Dataset	PAS Number	Ave. OD pair number per PAS	Ave. path count per OD pair
Sample 1 (2011)	80	20	42
Sample 2 (2015)	291	21	43

Rule 4 The PAS $p(e_1, e_2)$ is *valid* for testing the proportionality condition only if the number of its valid O-D pairs exceeds certain threshold so that a meaningful statistical analysis can be performed. In our implementation, a PAS is selected for test only when it has at least eight valid O-D pairs.

5. Results

5.1. Preliminary results

Table 3 summarizes the preliminary results after processing data in samples 1 and 2. The second column shows the number of occupied trips obtained from trip identification; the third column reports the number of paths identified from map matching, and the fourth column gives the number of PASs generated from Algorithm 2. Roughly, we are able to convert half of all occupied trips into paths, and generate a unique PAS for every seven paths.

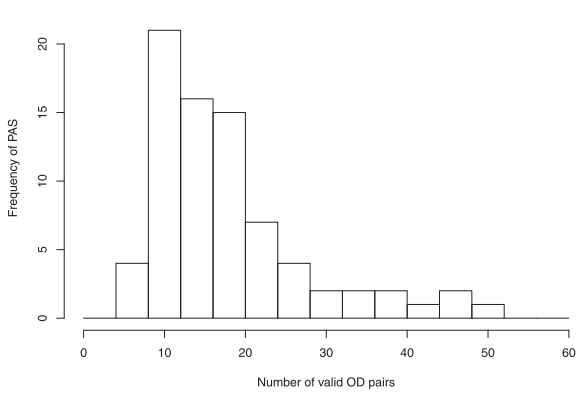
The PASs generated in each sample are then selected following the proposed four rules in the previous section. Table 4 shows how the number of PASs change when different rules are applied. For Sample 1, Rule 1 excludes about 21.58% of the total PASs. Adding Rule 4 excludes additional 74.5% of all PASs, leaving a little more than 3.92% intact. Finally, when Rules 2 and 3 are applied, the number of valid PASs drops to about 0.01% of the original number. The effects of the rules on Sample 2 data are very similar. The fact that Rules 1 and 4 can disqualify more than 96% of all PASs implies that most of the PAS generated are either trivial alternatives (Rule 1) or only serve traffic of very limited spatial diversity (Rule 4). Of those PASs that survive Rules 1 and 4, more than 99% are ultimately eliminated because there are simply not enough trips with similar travel times. Suffice it to say that the extremely low qualification rate is not a result of cherry-picking, but rather highlights the intensive data requirements to perform such tests (recall that we start with about three billion GPS points).

Fig. 6 shows the histogram plot for the number of valid O-D pairs across all PASs, and the histogram plot for the number of valid paths across all O-D pairs in both samples. Interestingly, while the two samples have very different size and collection time, their histograms are very similar in terms of both shape and range. As shown in Table 5, for both data samples, on average each PAS has about 20-21 valid OD pairs, and each O-D pair has about 42-43 valid paths. This unexpected side finding seems to indicate that these statistics may be affected more by the distribution of taxi travel demand and the network topology than by the amount of data.

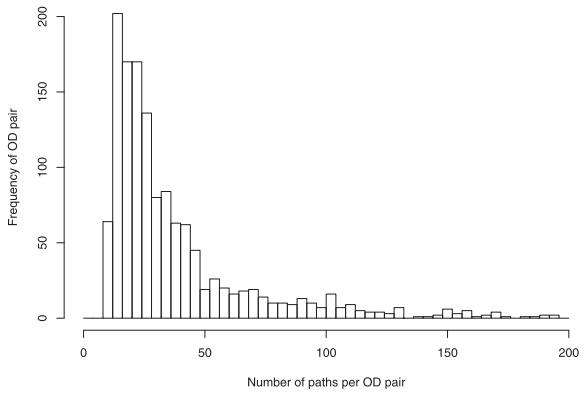
5.2. Goodness-of-fit to the proportionality condition

In this section we first offer an overview of how the proportionality condition is satisfied among all the valid PASs. As mentioned before, the two statistics from the linear regression analysis and the chi-square test are considered here: the R^2 value and the (p-value of) chi-square statistic.

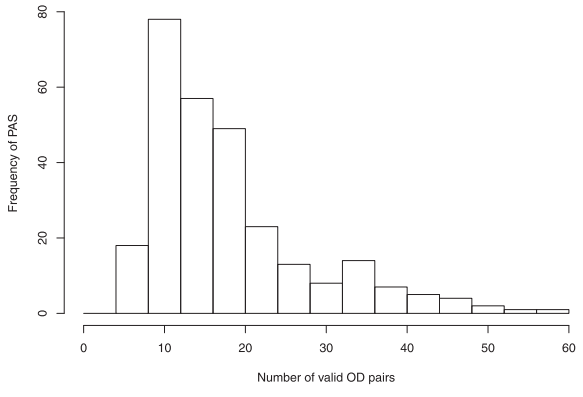
Fig. 7 draws the empirical cumulative distribution function (CDF) of (1) the R^2 value in the linear regression, (2) the p-value of chi-square statistic at $\alpha = 0.05$, and (3) the intercept-count ratio in the linear regression (ideally the intercept γ in formulation (1) should be close to zero), for both data samples. One can easily read from such plots the percentage of the PASs that have R^2 value, p-value or intercept-count ratio below/above any given threshold. Let us focus on the Sample 1 data first. Fig. 7(b) shows that about 70% of PASs has a p-value larger than 0.05. This suggests that, for an overwhelming majority of PASs, their segment flows ratios are deemed to have no significant difference by the chi-square test (note that



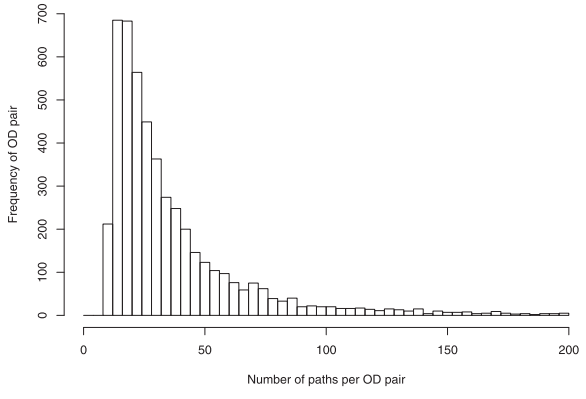
(a) Histogram of valid OD pairs per PAS for 2011



(b) Histogram of valid paths per OD pair for 2011



(c) Histogram of valid OD pairs per PAS for 2015



(d) Histogram of valid paths per OD pair for 2015

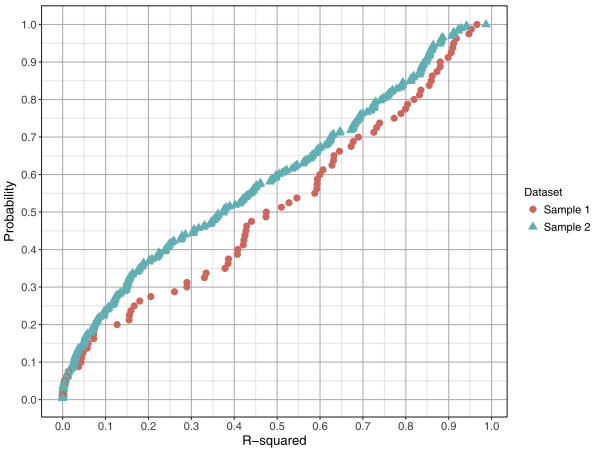
Fig. 6. Comparison of the distributions of valid OD pairs per PAS and valid paths per OD pair for each sample.

a small p-value implies that rejecting the null hypothesis is likely). Moreover, segment flows of many PASs tend to fit well to a straight line based on the R^2 values from the regression. Fig. 7(a) shows that about 40% of PASs has a R^2 value higher than 0.6, and about 60% has a R^2 greater than 0.4. Fig. 7(c), whose horizontal axis represents the ratio of intercept to the maximum O-D flow count on segment two (y-axis), shows that about 90% of the PASs with $R^2 > 0.4$ have ratios less than 0.2, suggesting that the majority of well-fitted PASs also have an intercept close to zero. The overall goodness-of-fit pattern is similar in the two data samples, with the second sample has a slightly lower portion of qualified PASs corresponding to the same criteria.

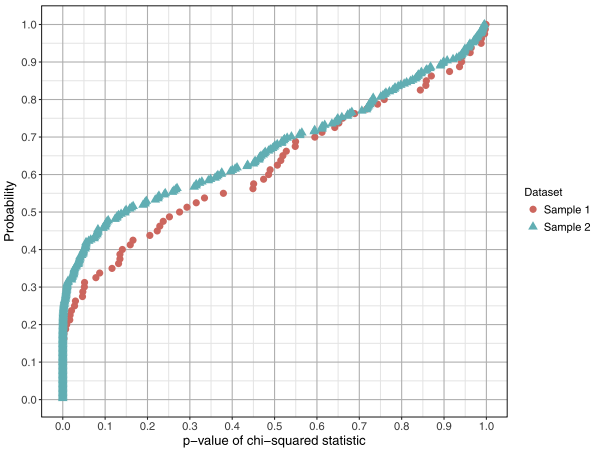
Thus, the results generally assert that for a large portion of the PASs tested, the proportionality condition is satisfied according to the definition proposed in Section 3. Specifically, we classify all valid PASs to four categories based on the two tests: *proportionality super-conforming* ($R^2 > 0.4$, p-value> 0.05), *proportionality* R^2*-conforming* ($R^2 > 0.4$ and p-value ≤ 0.05), *proportionality* χ^2*-conforming* ($R^2 \leq 0.4$ and p-value> 0.05) and *proportionality non-conforming* ($R^2 \leq 0.4$, p-value$\leq$ 0.05). Fig. 8 reports the share of PASs in each category. For Sample 1, about 49% of all valid PASs are super-conforming, about 36% are χ^2-conforming, and 14% are R^2-conforming. Overall, about 85% of all tested PASs in Sample 1 are proportionality conforming one way or another. The portion of conforming PASs in Sample 2 is 76%, slightly worse than that of Sample 1. A possible reason for the discrepancy might be that the data in Sample 2 comes from six different months in year 2015. As a result, drivers' route choice behavior may be subject to greater temporal fluctuations.

5.3. Detailed analysis

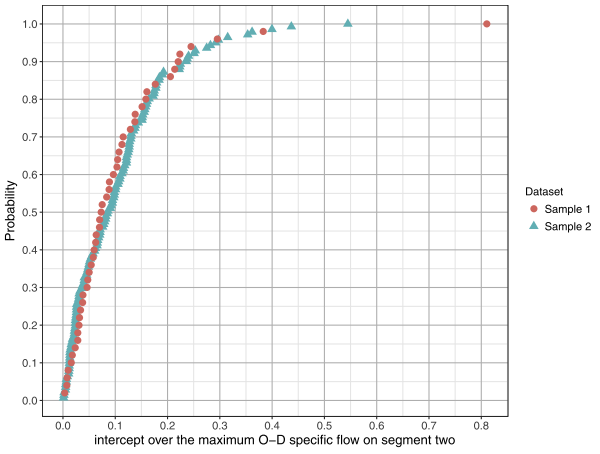
In this section, we proceed to perform detailed proportionality tests for PASs in different categories. Two PASs are analyzed in each category, and for each PAS three plots are created to visualize (1) the proportionality test; (2) the detailed topology as seen in the map; and (3) the corresponding origins and destinations.



(a) CDF of R-square for both data samples

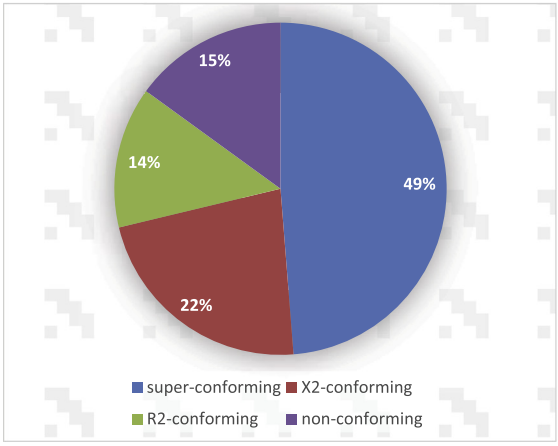


(b) CDF of p-value of chi-square statistic for both data samples

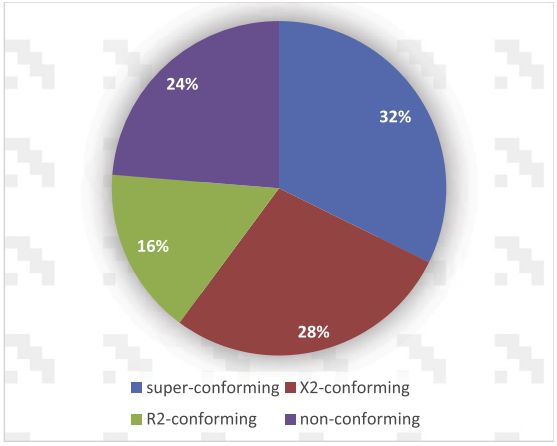


(c) CDF of intercept for the subsets of PASs with $R^2 \geq 0.4$ in both data samples

Fig. 7. Results of cumulative distribution function (CDF).



(a) Sample 1 (2011)



(b) Sample 2 (2015)

Fig. 8. Share of the four categories of PASs in both data samples.

5.3.1. Super-conforming cases

Two super-conforming PASs are selected from the first category. The two PASs have very similar p-values from chi-square test (0.24 vs. 0.23) but somewhat different R^2 values (0.95 vs. 0.83). This is designed to highlight the different focus of the two statistical tests.

Fig. 9(a) and (b) show the test results, where the horizontal and vertical axes represents respectively the valid path counts of segment one and two. Each point in a plot represents a valid O-D pair associated with the PAS and the line is the fitted linear function. The statistics of both the linear regression and the chi-square test are reported in the plots. If a PAS perfectly satisfies the proportionality condition, all the points should fall on a straight line. For PAS 1, at $R^2 = 0.95$, most points are closely aligned with the fitted line. In contrast, noticeable deviations from the fitted line can be found for PAS 2. In addition, note that the intercept should be zero if the proportionality is perfectly satisfied. The intercept for PAS 1 and PAS 2 are -1.1 and 2.75, reasonably close to zero compared to the range of the segment flows.

The p values of chi-square statistic are very close in both cases, even though PAS 1 evidently shows a better goodness-of-fit than PAS 2. A closer look reveals that the degree of freedom for PAS 1 (df=47) is significantly larger than that of PAS 2 (df=37). These observations verify that a larger degree of freedom tends to make it easier to reject the null hypothesis in the chi-square test.

Fig. 9(c) and (d) show the topology of the two PASs with key attributes. In both cases, the alternative segments are similar to each other in most attributes: road type, length and average segment travel time. Take PAS 2 as an example. Segment 2 consists of links of type 3 and 4 (urban arterial with 3 or 2 lanes), and Segment 1 consists of links of type 4 and 5 (urban arterial with 2 or 1 lane). The difference in the length and the average travel time is only 4.4% and 0.6% respectively. Such a high degree of similarity may explain why they are both used by the taxi drivers. However, it is important to note that, despite the similarities, the flows do not distribute *evenly* between the two segments. Rather, they follow certain proportion independent of their origin and destination. Why do taxi drivers seem to prefer Segment 1 to 2 for PAS 2 by a ratio of 2.1:1? A possible explanation is that Segment 2 can carry less traffic than Segment 1 under prevailing (or "equilibrium") conditions. For PAS 1, although segment 1 is about 14.3% longer than segment 2, their average travel time are very similar (with a 0.61% difference), mostly because part of segment 1 (from A to B) is expressway.

Fig. 9(e) and (f) show the origins and destinations associated with PAS 1 and PAS 2, respectively. Generally, most of these origins and destinations are clustered around the ends of the PAS, because such trips are more likely to use the PAS and be identified. There are interesting exceptions, however. In Fig. 9(e), several "remote" destinations located at the right bottom corner were found to be associated with PAS 1. A close look reveals that these destinations are near an intercity railway station that is potentially a major traffic attractor.

5.3.2. R^2-conforming cases

The detailed test results of two PASs from R^2-conforming category are reported in Figure 10. PAS 3 and PAS 4 both have a p value of chi-square statistic smaller than 0.01, while their R^2 values are 0.91 and 0.47, respectively.

As expected, the linear regression results (cf. Fig. 10(a) and (b)) reveal strong correlations in the two PASs but they fail the chi-square test, mainly due to their relatively large degree of freedom (PAS 3 is with df= 41 and PAS 4 is with df=27). It is somewhat unexpected that a PAS with R^2 as high as 0.91 would fail the chi-square test.

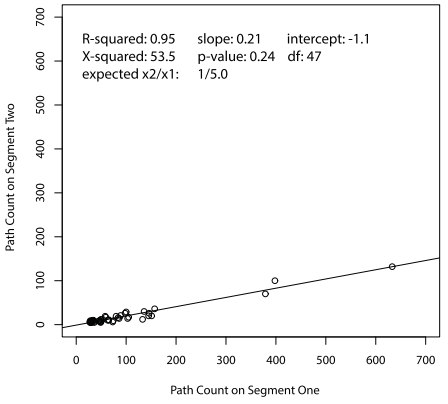
As shown in Fig. 10(d), segment 1 of PAS 4 consists of only urban arterial roads whereas segment 2 consists of both urban expressways (from A to B) and arterial road. Also, segment 2 is about 27.5% longer, but has a slightly lower average travel time thanks to the high speed on the expressway. The results suggest that on average about three quarters of all taxi trips choose the shorter but slower route. Yet, it seems irrational that so many drivers were willing to travel an extra mile for what appears to be a negligible travel time saving (about 2 s). For PAS 3 (Fig. 10(c)), segment 1 has smaller travel cost in terms of both length and average travel time, but the flow distribution on segment 2 is still as large as one quarter regardless of origins or destinations. It is tempting to credit such "irrational behavior" to hidden attributes of the alternatives. That is, taxi drivers' utility may be affected by factors beyond length and travel time. However, a more plausible explanation may be "hidden traffic", i.e. traffic flows originating elsewhere that are invisible to our analysis because they use one but not both segments. The fact that the percentage of "irrational" drivers are relatively steady across all O-D pairs suggests that a route choice preference consistent with the proportionality hypothesis does seems to exist at the aggregated level.

5.3.3. χ^2-conforming cases

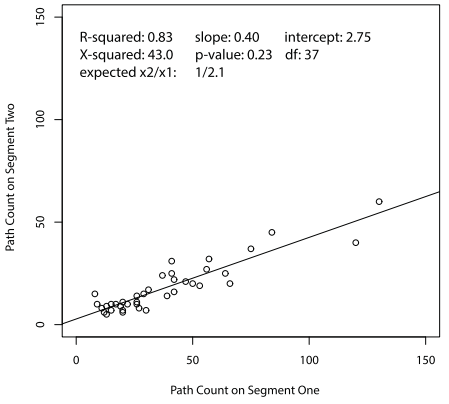
PASs 5 and 6 (as shown in Fig. 11) have a R^2 value less than 0.3, but their p value of chi-square statistic are 0.32 and 0.45 respectively. Since the p values are much larger than the significance level (0.05), we cannot reject the null hypothesis of the chi-square test, which states that the segment flow ratios among different O-D pairs have no difference at the significance level of 0.05.

The expected segment flow ratio for PAS 5 is 1:7.5, i.e. on average only about 11.8% of total flow uses segment 2. As a result, the fitted line appears to be in parallel with the horizontal axis. As mentioned in Section 3, the R^2 value in the linear regression analysis tend to underestimate goodness-of-fit when the slope of the fitted line is closer to zero. In this case, the chi-square test would be more reliable.

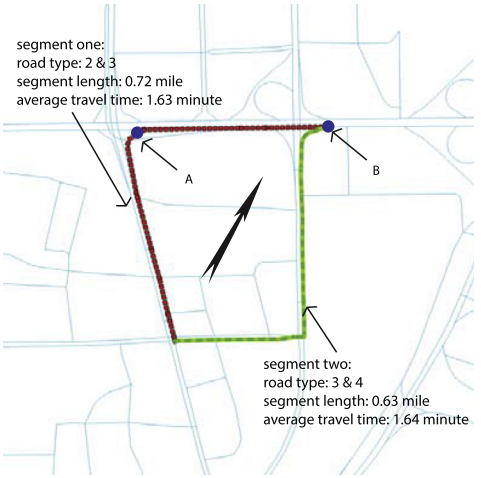
PAS 6 shows another typical case where a strong correlation clearly exist but the linear regression test fails to produce a reasonable R^2 value (see Fig. 11(b)). The failure may be attributed to the fact that the data points are too few and they



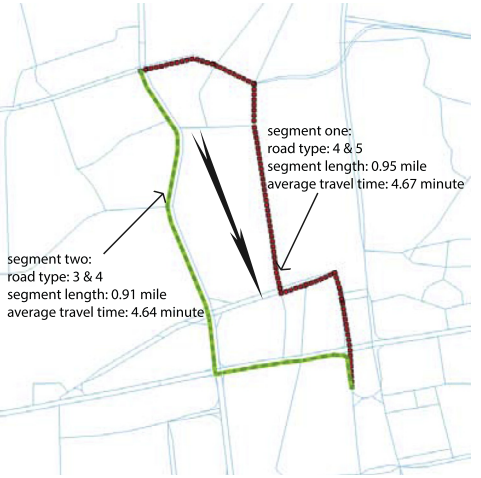
(a) linear relationship of PAS 1



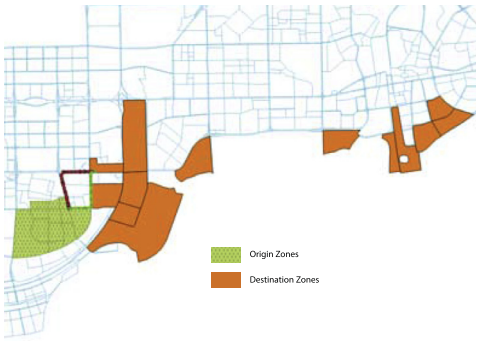
(b) linear relationship of PAS 2



(c) segments of PAS 1



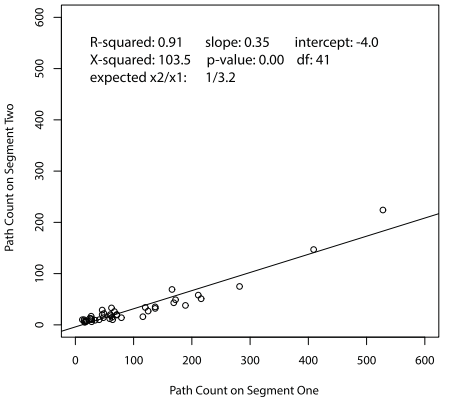
(d) segments of PAS 2



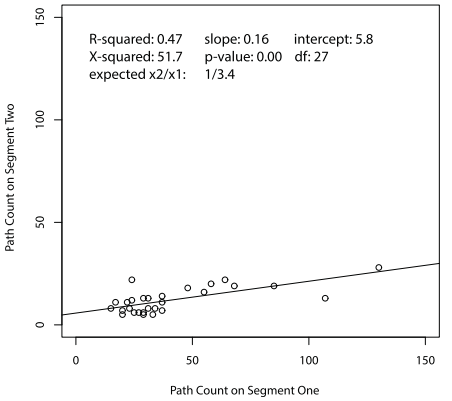
(e) zones associated with PAS 1

(f) zones associated with PAS 2

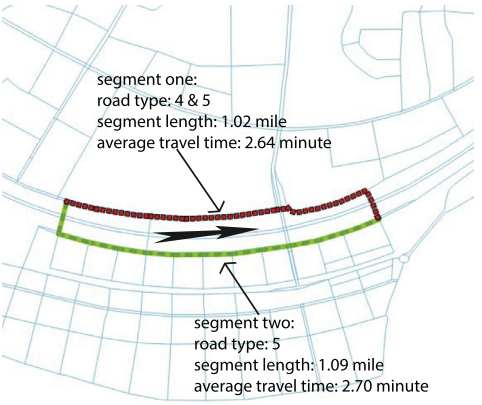
Fig. 9. Selected PASs from the super-conforming category. Road type: 1 denotes freeway with four lanes each direction; 2 denotes urban expressway with four lanes each direction; 3 denotes urban arterial road with three lanes each direction; 4 denotes urban branch road with two lanes each direction; 5 denotes urban branch road with one lane each direction.



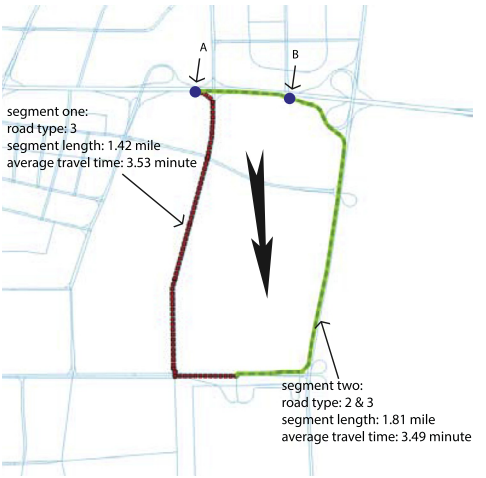
(a) linear relationship of PAS 3



(b) linear relationship of PAS 4

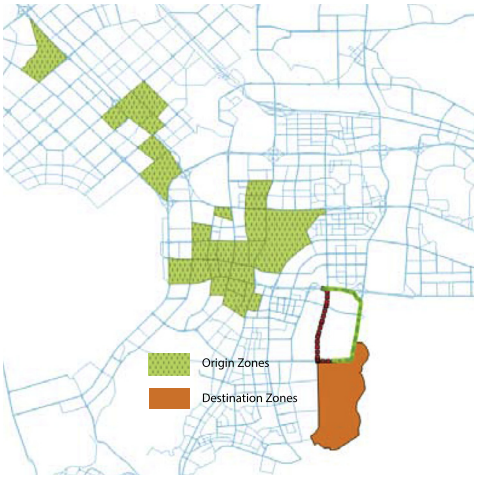


(c) segments of PAS 3



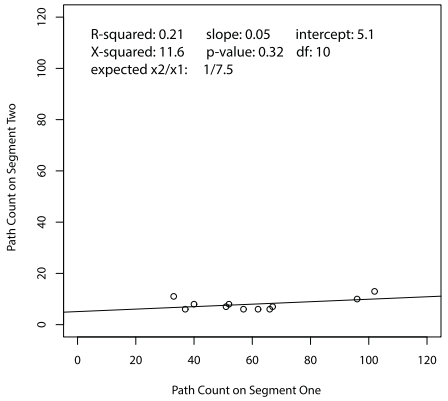
(d) segments of PAS 4

(e) zones associated with PAS 3

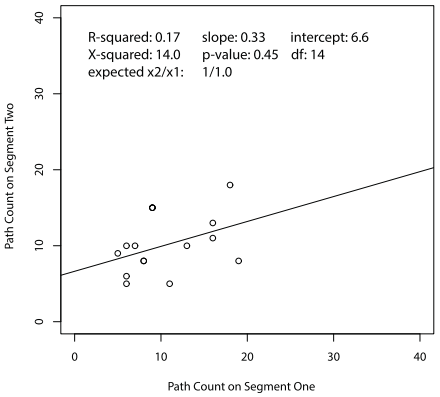


(f) zones associated with PAS 4

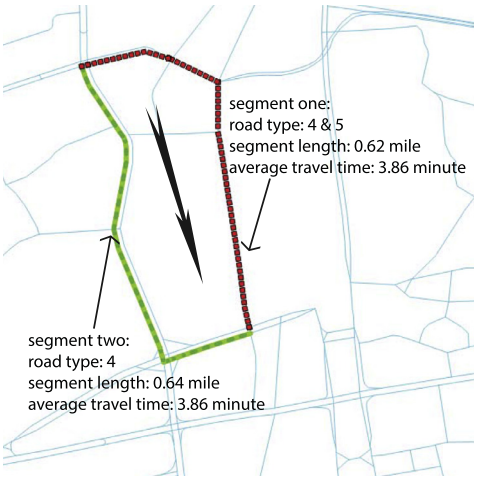
Fig. 10. Test results for PASs in the R^2-conforming category.



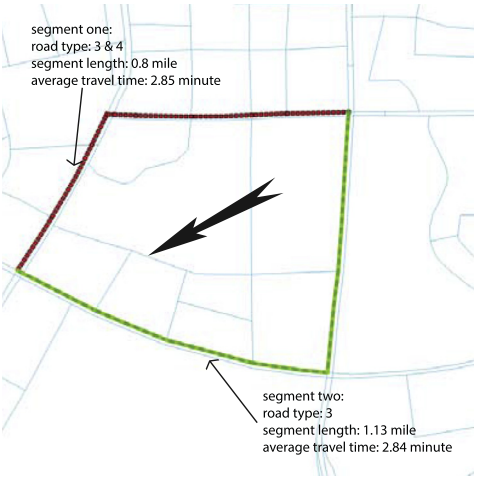
(a) linear relationship of PAS 5



(b) linear relationship of PAS 6



(c) map representation of PAS 5



(d) map representation of PAS 6

(e) zones associated with PAS 5

(f) zones associated with PAS 6

Fig. 11. Test results for PASs from the χ^2-conforming category.

are scattered loosely along the fitted line. Again, the chi-square test clearly works better (in the sense that it confirms the linear correlation) when the sample size is small.

5.3.4. Non-conforming cases

The non-conforming PASs often exhibit two typical patterns, as illustrated in Fig. 12. For PAS 7, the points can be easily clustered into two groups, and each would fit reasonably well with a line of a distinctive slope. A closer look reveals that all the points gathering around the line with smaller slope are associated with the same origin 872 (marked in Fig. 12(e)), whereas the remaining points correspond to other origins. This is clearly a violation of the proportionality condition, because flows from different origins do use both alternative segments, but nonetheless distribute between them with different proportions. A possible explanation is that the size and shape of zone 872, as well as its location relative to the PAS, makes trips originating from it strongly prefer Segment 2. Importantly, unlike in the assignment model, where trips all comes from the "centroid" of a zone, real trips may originate from any intersection in a zone.

Fig. 12(b) illustrates another non-conforming pattern, in which most points do exhibit a linear relationship, but some deviate significantly from the general pattern (e.g. points A and B). Interestingly, if we exclude these outliers, PAS 8 will become R^2-conforming with $R^2 = 0.58$. While not attempted here, it is reasonable to expect that a greater degree of conformity may be achieved if such outliers are removed through a preprocessing operation.

5.3.5. Proportionality by link flows

Following Bar-Gera et al. (2012), we proceed in this section to demonstrate the proportionality condition by aggregating all path flows traversing the PAS over links, as shown in Fig. 13. In the plot, different link color and width represent different flow ratios relative to the flow on the selected segment (see the legend).

Under the proportionality condition, the share of any specific link relative to segment 1 should be identical to the share of the same link relative to segment 2 (Bar-Gera et al., 2012). In other words, the map of segment 1 should look exactly the same as the map of segment 2 except for the PAS itself. Indeed, the paired maps for the PAS 2 (see Fig. 13(a) and (b)) are very similar, which is consistent with the fact that this PAS is classified as super-conforming. Minor differences can be found between Fig. 13(c) and (d), corresponding to PAS 4 (a R^2-conforming case with $R^2 = 0.47$). As expected, Fig. 13(e) and (f) demonstrate greater discrepancies. For example, Fig. 13(e) shows that over 80% of the flows that pass segment one are from origin 872, but the share of flows passing segment two from origin 872 is only about 20% to 40% (see Fig. 13(f)). In a nutshell, the aggregated link flow comparison confirms the findings in the previous sections.

6. Concluding remarks

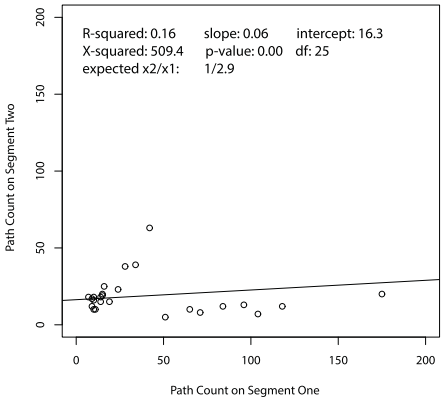
The proportionality condition has been widely used in practice to produce the unique path flow solution for the user equilibrium traffic assignment problem. This study attempted to test, for the first time to the best of our knowledge, whether and to what extent this condition accords to real travel behavior. Using route choice data mined from a large taxi trajectory data set, we obtained 27 million occupied taxi trips, from which a set of valid PASs were uncovered. The proportionality condition is then tested by performing various statistical tests over O-D specific flows on paired segments.

The results suggest that for a commanding majority of the PASs tested, the proportionality condition is satisfied with a reasonable level of statistical significance. It is also worth noting that not all poorly fitted cases (about 15–20%) represent a clear violation of the proportionality condition. Many other factors may be at work. For one thing, since we use the travel time as the surrogate for travel cost, the PAS identified may in fact not be competitive alternatives for some trips. Also, in theory proportionality is defined for "near" equilibrium state that may not always manifest in reality. Finally, as our empirical analysis reveals, the shape, size and relative location of zones may also swing the route choice one way or the other. Taking these noises into consideration, it is remarkable that the proportionality condition is indeed satisfied - be imperfectly it may - for most tested PASs.

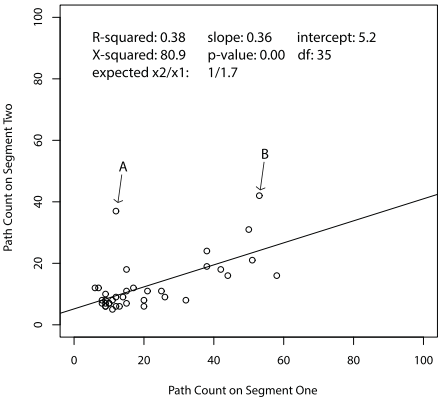
Because it involves route choice, testing proportionality turns out to be extremely data intensive. With three billion GPS points, we were only able to obtain a couple of hundred valid PASs. To put this in perspective, the number of actively used PAS is in the order of several thousands in the case of Chicago (Bar-Gera et al., 2012). Thus, it is almost certain that what we have tested is but a small percentage of all used PASs in a user equilibrium traffic assignment solution. The small size of the taxi fleet (less than 1% of vehicle population in the City of Shenzhen) and the peculiar demand it serves may limit where and how many valid PASs can be uncovered. Thus, simply increasing the amount of taxi data may not be very effective to further strengthen the test for more conclusive results. For future study, large-scale GPS trajectory data of other types of traffic (e.g. regular commuters and truck drivers) may be considered to generalize the applicability of the proportionality condition.

Acknowledgments

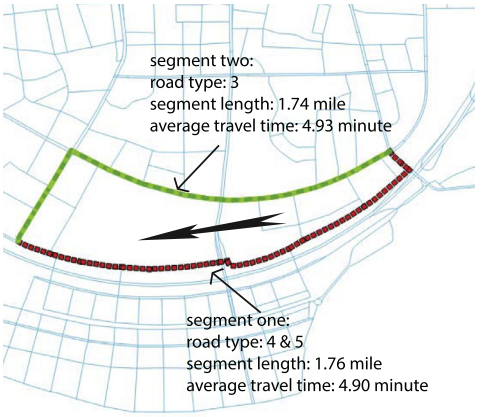
The work was conducted when the first author visited Northwestern University as a visiting postdoctoral researcher. He was funded by Chinese National Nature Science Foundation (Grant NO. 71501129) and Chinese International Postdoctoral Exchange Fellowship Program (NO. 20150045). The work was also partially funded by the United States National Science Foundation under the award number CMMI-1402911. We wish to thank Mr. Jiandong Qiu from Shenzhen Urban Transport Planning Center for providing the COST data used in this study.



(a) linear relationship of PAS 7

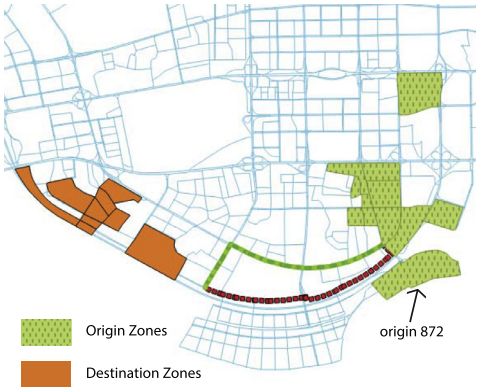


(b) linear relationship of PAS 8

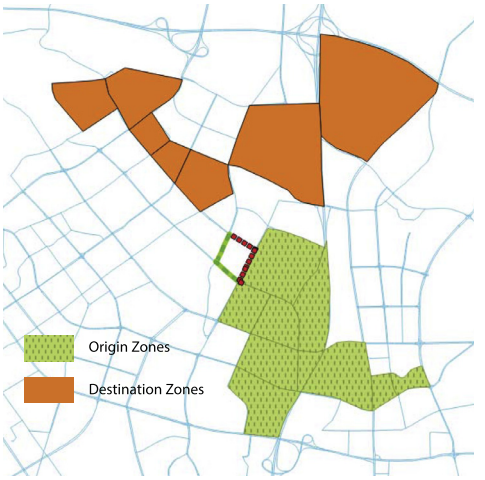


(c) segments of PAS 7

(d) segments of PAS 8

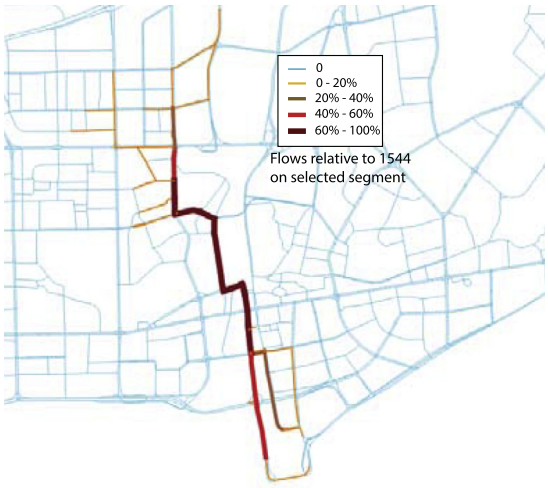


(e) zones associated with PAS 7

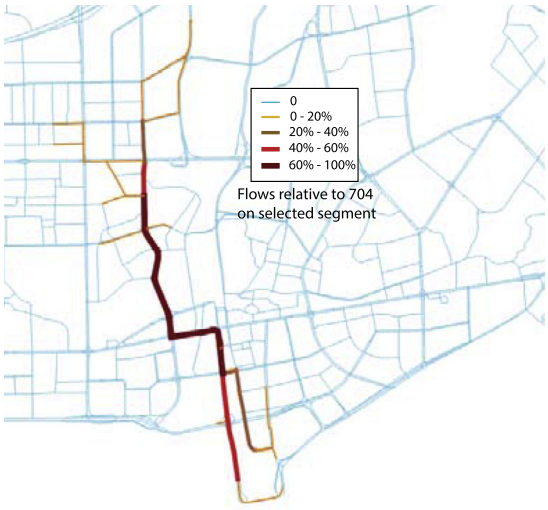


(f) zones associated with PAS 8

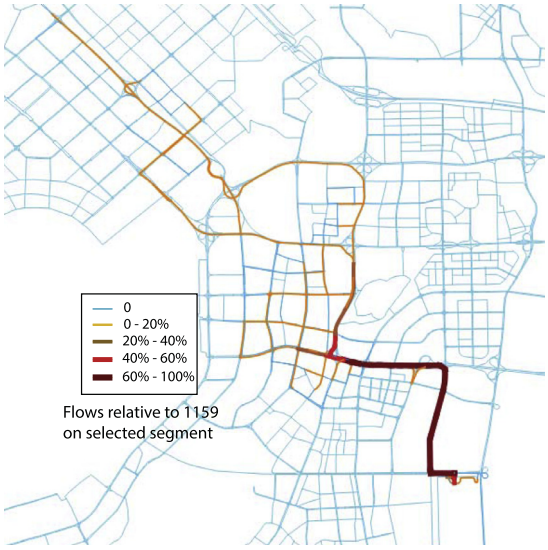
Fig. 12. Test results for PASs from the non-conforming category.



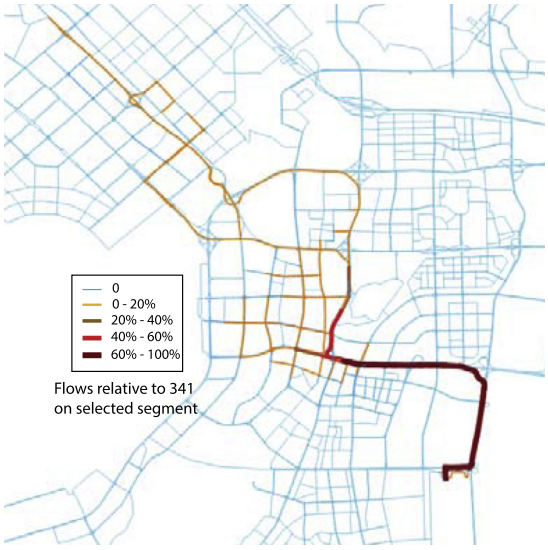
(a) path flow trace for segment 1 of PAS 2



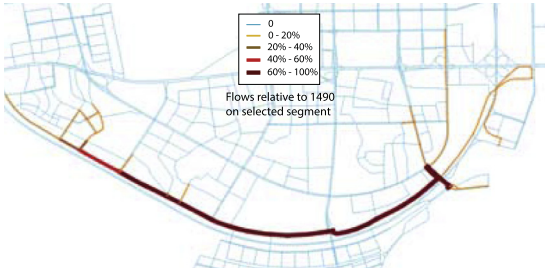
(b) path flow trace for segment 2 of PAS 2



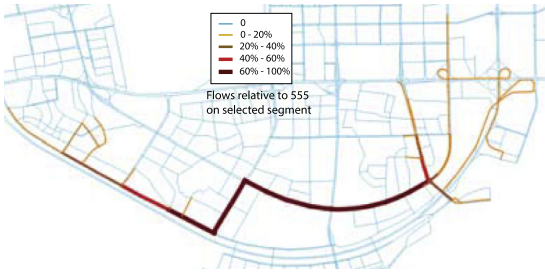
(c) path flow trace for segment 1 of PAS 4



(d) path flow trace for segment 2 of PAS 4



(e) path flow trace for segment 1 of PAS 7



(f) path flow trace for segment 2 of PAS 7

Fig. 13. PAS-based path flow trace for selected PASs.

References

Bar-Gera, H., 2006. Primal method for determining the most likely route flows in large road networks. Transp. Sci. 40 (3), 269–286.
Bar-Gera, H., 2010. Traffic assignment by paired alternative segments. Transp. Res. Part B 44 (8), 1022–1046.
Bar-Gera, H., Boyce, D., 1999. Route flow entropy maximization in origin-based traffic assignment. 14th International Symposium on Transportation and Traffic Theory.
Bar-Gera, H., Boyce, D., Nie, Y.M., 2012. User-equilibrium route flows and the condition of proportionality. Transp. Res. Part B 46 (3), 440–462.
Beckmann, M., McGuire, C., Winsten, C.B., 1956. Studies in the Economics of Transportation. Technical Report.
Borchers, M., Breeuwsma, P., Kern, W., Slootbeek, J., Still, G., Tibben, W., 2015. Traffic user equilibrium and proportionality. Transp. Res. Part B 79, 149–160.
Boyce, D., Nie, Y., Bar-Gera, H., Liu, Y., Hu, Y., et al., 2010. Field test of a method for finding consistent route flows and multiple-class link flows in road traffic assignments. Fed. Highway Administration.
Li, Q., Zeng, Z., Zhang, T., Li, J., Wu, Z., 2011. Path-finding through flexible hierarchical road networks: an experiential approach using taxi trajectory data. Int. J. Appl. Earth Obs. Geoinf. 13 (1), 110–119.
Lin, N., Zheng, Y., Li, J., 2015. Path planning method based on taxi trajectory data. J. Inf. Comput. Sci. 12 (9), 3395–3404.
Liu, S., Liu, Y., Ni, L.M., Fan, J., Li, M., 2010. Towards mobility-based clustering. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 919–928.
Liu, S., Qu, Q., 2016. Dynamic collective routing using crowdsourcing data. Transportation Research Part B: Methodological 93, 450–469.
Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., 2009. Map-matching for low-sampling-rate gps trajectories. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 352–361.
Lu, S., Nie, Y.M., 2010. Stability of user-equilibrium route flow solutions for the traffic assignment problem. Transp. Res. Part B 44 (4), 609–617.
Newson, P., Krumm, J., 2009. Hidden Markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in geographic information systems. ACM, pp. 336–343.
Nie, Y.M., 2017. How can the taxi industry survive the tide of ridesourcing? evidence from Shenzhen, China. Transp. Res. Part C 79, 242–256.
Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London, Edinburgh Dublin Philos. Mag. J. Sci. 50 (302), 157–175.
Quddus, M.A., Noland, R.B., Ochieng, W.Y., 2006. A high accuracy fuzzy logic based map matching algorithm for road transport. J. Intell. Transp. Syst. 10 (3), 103–115.
Rossi, T.F., McNeil, S., Hendrickson, C., 1989. Entropy model for consistent impact-fee assessment. J. Urban Plann. Dev. 115 (2), 51–63.
Sheffi, Y., 1985. Urban transportation networks: equilibrium analysis with mathematical programming methods. Prentice Hall, Englewood Cliffs, NJ.
Wang, Y., Zheng, Y., Xue, Y., 2014. Travel time estimation of a path using sparse trajectories. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and data mining. ACM, pp. 25–34.
Wenk, C., Salas, R., Pfoser, D., 2006. Addressing the need for map-matching speed: localizing global curve-matching algorithms. In: 18th International Conference on Scientific and Statistical Database Management (SSDBM'06). IEEE, pp. 379–388.
Xie, J., Xie, C., 2016. New insights and improvements of using paired alternative segments for traffic assignment. Transp. Res. Part B 93, 406–424.
Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y., 2010. T-drive: driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 99–108.
Yue, Y., Zhuang, Y., Li, Q., Mao, Q., 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In: 2009 17th International Conference on Geoinformatics. IEEE, pp. 1–6.
Zhang, Y., Li, B., Ramayya, K., 2016. Learning individual behavior using sensor data: the case of gps traces and taxi drivers. Available at SSRN 2779328.
Zheng, J., Liu, S., Ni, L.M., 2013. Effective routine behavior pattern discovery from sparse mobile phone data via collaborative filtering. In: Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on. IEEE, pp. 29–37.
Zheng, Y., 2015. Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. (TIST) 6 (3), 29.