DOI:10.13203/j.whugis20140221　　文章编号:1671-8860(2016)06-0797-06

利用经验约束规则和证据理论进行出租车异常轨迹检测

周　洋[1]　方志祥[2]　李清泉[2,3]　郭善昕[1]

1　武汉大学遥感信息工程学院,湖北 武汉,430079
2　武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079
3　深圳大学空间信息智能感知与服务深圳市重点实验室,广东 深圳,518060

摘　要:提出了一种根据择路经验特征,利用证据理论检测出租车异常轨迹的方法。该方法考虑最短路径与轨迹长度的比值、规避路径代价、出行发生时间 3 个因素,利用证据理论综合这 3 个证据来判别轨迹的异常程度,检测出行距离和路径选择明显不同于正常情况下的异常轨迹。实验结果表明此方法能有效识别不符合正常认知的异常轨迹,不依赖于单一起始点和终点对(origin-destination,OD)中的轨迹数目,能快速处理海量 GPS 数据,可用于大规模浮动车数据择路行为分析前期的数据过滤。

关键词:异常轨迹检测;D-S 证据理论;经验约束规则;出租车轨迹

中图法分类号:P208　　**文献标志码**:A

出租车轨迹广泛应用于城市动态监测[1]、商圈吸引力分析[2]、城市内人群出行模式分析[3-4]等方面。基于数据传输效率、存储空间等方面的考虑,大规模浮动车系统普遍采样率较低(40～120 s),使得发生在采样间隔中的乘客换乘无法记录,于是两段旅程有可能会被记录为一段轨迹。另外,有些出租车司机为赚取更多车费故意带乘客绕行较远路径等欺诈行为的发生,也会导致产生行驶距离和路径选择不合常理的路径。这些行为产生的异常轨迹淹没于真实数据集中,对后续的正常情况下司机择路行为分析、起始点和终点(origin-destination, OD)对之间的旅程开销分析等造成一定的干扰,而检测出租车异常轨迹有助于数据分析前期的数据过滤。另一方面,大多数出租车宰客纠纷还停留在投诉后人工判别的基础上,自动检测开销明显超出正常范畴的异常轨迹能为提高出租车服务的规范性提供一定的技术支持。

常见异常轨迹的探测方法如支持向量机(support vector machine, SVM)、马尔柯夫时序分析等机器学习方法[5]需要较大样本数目用于学习训练,在出行量较少的 OD 区域的效果并不显著,且城市中出租车轨迹的 OD 对分布广泛且多样,用 SVM 等机器学习方法检测异常轨迹,需要对每一 OD 对都建立样本数据,花费开销较大。另一类基于密度和子轨迹特征等的轨迹分类法[6]难以应付采样稀疏和以路网为基础的 GPS 轨迹,且出租车轨迹分布并不均匀,难以确定密度阈值。文献[7]提出了一种基于 iBAT 的方法,构建伪轨迹序列探索同一或相似 OD 对的所有轨迹中"少"且"异于大多数"的异常轨迹。文献[8]提出了一种基于速度的出租车欺诈行为探测方法。文献[9]通过距离和密度特征利用证据理论检验异常轨迹。另外,也有方法将空间划分为格网[6,9],GPS 轨迹为伪格网序列,方法中所计算的行驶距离并不是出租车在实际路网中的行驶距离,并没有考虑到出租车行驶的城市背景,如路网、时间等因素。本文认为出租车的轨迹符合一定的时空约束和行为习惯,异常轨迹的行驶长度、择路行为会偏离正常认知,因此提出一种不依赖于样本数目,侧重于出行时空约束和经验规律的异常轨迹探测方法。

收稿日期:2014-10-15

项目资助:国家自然科学基金(41371377, 41231171);国家 863 计划(2012AA12A403-4);深圳市科技研发资金(ZDSY20121019111146499,JSGG20121026111056204);深圳市战略性新兴产业发展专项资金(JCYJ20121019111128765)。

第一作者:周洋,博士生,主要研究方向为移动对象轨迹分析和时空模式挖掘。cleverzhouyang@whu.edu.cn

通讯作者:方志祥,博士,教授。zxfang@whu.edu.cn

Dempster-Shafer (D-S)证据理论[10]是一种不确定性推理和处理方法,不需要先验概率和条件概率就可以进行证据推理,同时可以依靠证据的积累不断缩小假设集,可以将“不确定”区分开来。利用D-S合成法则可以综合考虑多证据进行决策推理,处理不确定性问题和异常检测。本文从出租车出行的时空经验约束入手,以旅程距离、规避路径代价、出行发生时间等约束条件为判别证据,利用证据理论检验出行距离和路径选择明显不同于正常情况下的异常轨迹。该方法并不依赖于单一OD对中的轨迹数目,能快速处理海量GPS数据。

1 基于证据理论的异常检测

1.1 D-S证据理论

D-S证据理论中的识别框架Θ包含了所有可能的判别假设,各假设相互独立且互斥,有2^{Θ}个子集。本文中用于异常轨迹检测有两种判别,分别为$A=\{$异常$\}$、$B=\{$正常$\}$,则识别框架$\Theta=\{A,B\}$,有$\{A\}$、$\{B\}$、$\{A,B\}$、$\varnothing$共4种子集。

对$\forall X\subseteq\Theta$,存在基本概率分配(basic probability assignment, BPA)函数$m(X)$,表示证据对命题X的信任程度,且$m(X)$满足$m(\varnothing)=0$且$\sum_{X\subseteq 2^{\Theta}} m(X)=1$。BPA可通过检测数据或者人的经验给出。根据BPA,可计算证据对某假设X的信任函数Bel(X)和似然函数Pl(X)[10]:

$$\mathrm{Bel}(X)=\sum_{Y\subseteq X} m(Y) \tag{1}$$

$$\mathrm{Pl}(X)=\sum_{Y\cap X=\varnothing} m(Y) \tag{2}$$

式中,Bel(X)为包含在X中的所有子集的BPA之和,表示证据对X为真的信任程度;Pl(X)表示证据不怀疑X的程度;Bel(X)和Pl(X)分别表示证据对假设X为真信任程度的下限和上限,每个命题都有形式为[Bel(X), Pl(X)]的判别结果,[Bel(X), Pl(X)]称为信度区间。

本文选取距离约束、经验规避路段约束和时间约束3个指标作为证据判别轨迹的异常程度。

1.2 距离约束

乘客对距离、费用等因素的考虑,使得正常出租车的行驶路径满足一定的距离约束。引入参数R_d衡量实际行驶距离(GPS trajectory, GT)L_{GT}与最短路径(shortest distance path, SDP)L_{SDP}间的比率:

$$R_d=\frac{L_{\mathrm{SDP}}}{L_{\mathrm{GT}}} \tag{3}$$

R_d的范围为$(0,1]$,R_d越大,表明实际行驶路径与最短路径的长度差别越小;R_d越趋近于0,表明相比于最短距离,实际路径绕行的越远。一般而言,由于宰客和信息缺失所造成的异常轨迹都会有较大的绕路,即R_d的值较小。

1.3 经验规避路段约束

由于出租车司机对城市路网十分熟悉,且对城市交通状态具有充分全面的了解,所以其对路径规划有自己的经验和习惯。许多出租车由于拥堵、转向、驾驶习惯等原因,会牺牲一定的行驶距离来保证驾驶的顺畅和舒适度,有意识地规避最短路径中的某些路段而选择另外的替代路径。近年来,许多研究基于出租车寻路经验知识构建经验层级路网,进而进行导航路径的规划[11-12],均以出租车频繁通过的路段作为出租车司机的经验知识,结合速度、时间、联通度等指标进行路径优化导航。与前述研究有所区别的是,本文将最短路径中出租车司机经常拒绝的路段称为经验规避路段(experiential avoid road, EAR),EAR是属于最短路径但出租车并没有实际采纳的路段。EAR的寻找方式与经验路段的构建方式类似,本文以浮动车规避最短路径的频繁程度将所有EAR划分为N个等级。等级越高,说明规避该路段的轨迹越多。等级最低的路段说明规避属于偶发性现象。将出租车为了规避EAR而产生的绕行代价记为I_c,计算公式为:

$$I_c=\begin{cases}0, L_{\mathrm{GT}}^{d}=0\\ 1-\dfrac{\sum_i w_i\times L_{\mathrm{SDP}}^{d}(\mathrm{EAR}_i)}{L_{\mathrm{GT}}^{d}}, L_{\mathrm{GT}}^{d}\neq 0\end{cases} \tag{4}$$

式中,L_{GT}^{d}为实际轨迹中区别于最短路径部分的路段总长度;$L_{\mathrm{SDP}}^{d}(\mathrm{EAR}_i)$表示被规避最短路径中EAR路段的长度;$w_i$为权重,由对应EAR所属的等级$N$决定。当$L_{\mathrm{GT}}^{d}=0$时,说明实际路径与最短路径一致,没有花费代价,$I_c=0$;当$L_{\mathrm{GT}}^{d}\neq 0$时,$I_c$越大,表明轨迹为了规避EAR绕行的代价越大。与正常绕行相比,异常轨迹的规避路段并不含有EAR或者存在EAR但绕行的代价十分离谱。

1.4 时间约束

一般而言,出租车在白天交通条件复杂的情况下产生绕行的概率较大,而在夜晚由于道路交通通达,不会产生拥堵,所以夜间发生的较大绕行通常为异常轨迹。本文为简便处理,将时间划分为白天模式(07:00~21:00)和夜间模式(22:00~06:00)。在白天模式中交通状况复杂,发生一定程度的绕路比较正常。在夜间模式中,交通畅通,

绕行行为属于异常轨迹的可能性较大。时间约束判别能力较弱,更适合用作辅助性判别。

1.5　异常指数

本文以距离约束 R_d 和经验规避路段约束 I_c 作为主要证据逐一检验每一轨迹为异常的确定程度,时间约束为辅助判别,检验流程如图 1 所示。多证据可通过 Dempster 合成规则进行正交运算得到综合的 BPA 函数,反映证据的联合作用。为避免证据冲突时产生的 Zadeh 悖论,本文采用 Yager 提出的合成规则[13]。Bel(A)表示认为轨迹属于异常的程度,Bel(A)越大 ,表示轨迹异常指数越高;Pl(A)表示不怀疑轨迹为异常的程度,Pl(A)越小,表明轨迹正常的指数越大,异常指数越低。一般来说,Bel(A)＞50％的轨迹可判别为异常,Pl(A)＜50％的路径可判别为正常,对于其他不确定情况,再结合时间进行辅助判别。例如,若 Bel(A)处于 40％ ～ 50％之间,且发生时间为夜间,仍可判定为异常。

图 1　算法流程

Fig. 1　Flowchart of the Algorithm

2　实验与分析

2.1　数　据

选取武汉市二环内区域作为实验区域(见图 2),采用武汉市超过 10 000 辆出租车的浮动车数据,数据采集时间为 2012-09-07～2012-09-16(其中 09-12～09-13 数据丢失)共 8 d 的数据,采样间隔为 60 s 左右,经过地图匹配和轨迹恢复选择载客状态且起始点均处于实验区域的轨迹共 587 820 条。部分轨迹分布见图 2。

2.2　证据指标

证据的选择直接影响最终的判别结果,所以有必要考察每个证据指标的特征。计算每条轨迹对应的最短路径和 R_d,并统计基本信息如表 1。

图 2　研究区域和出租车轨迹分布示例

Fig. 2　Study Area and Samples of Taxi GPS Trajectories

表 1　实验数据统计

Tab. 1　Statistical Description of Traces in Different R_d

指标	$R_d \leqslant 1$	$R_d < 1$	$R_d \leqslant 1/1.25$	$R_d \leqslant 1/1.5$
轨迹数目	587 820	362 008	23 547	7 101
占所有轨迹百分比	100％	61.6％	4％	1.2％
平均 GT 长度/m	3 795	5 052	7 013	8 208
平均 SDP 长度/m	3 606	4 557	4 708	4 408
平均 R_d	0.97	0.92	0.69	0.56

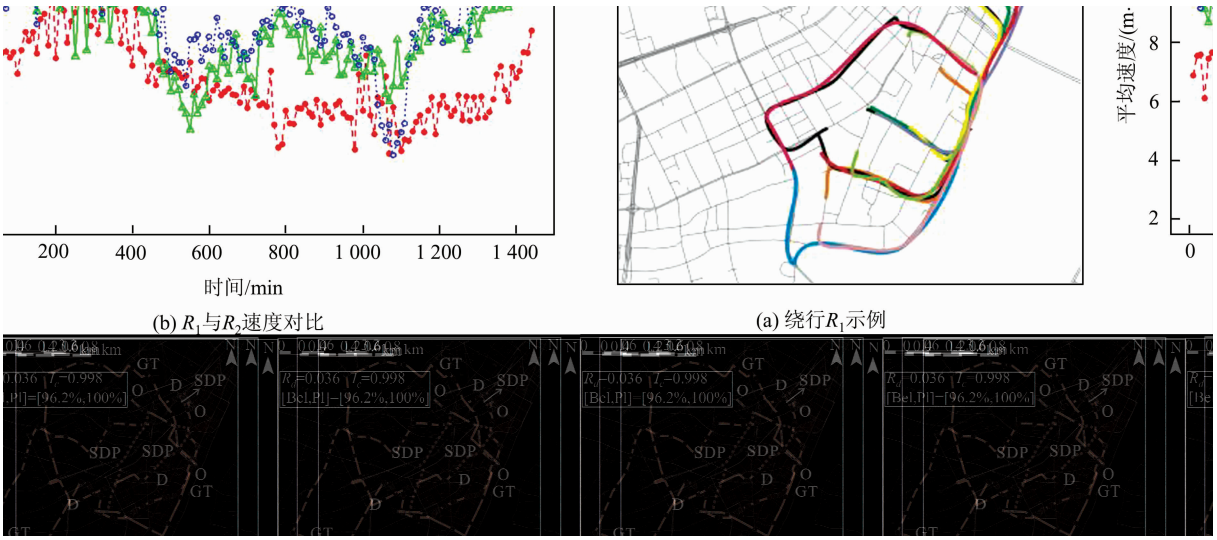
从表 1 可以看出,实际路径长度超过最短路径长度 1.5 倍的轨迹(即 $R_d \leqslant 1/1.5$)在数据中所占的比例并不算高(1.2％),但其平均最短路径长度和实际轨迹长度都高于整体数据,说明绕路较易发生在距离较远的出行中。

图 3 为 EAR 层级图,以道路 R_1 为例具体说明司机绕行 EAR 的原因。图 3(b)中 R_1 位于主干道中山大道,由于该处地处商业最繁华地段,交通流量大,所以许多司机绕行 R_2(沿江大道)和其他低等级道路,如图 4(a)。图 4(b)为 R_1 与 R_2 在 24 h 内每 10 min 的平均速度对比,可以看见 R_2 南北走向的两条路段的平均速度明显高于 R_1。另外,长江隧道也是绕行发生较大的地方,主要原因在于长江隧道一旦发生拥堵,司机将没有其他应变措施以选择其他替代路径。

2.3　检测结果与分析

本文以指标 R_d、I_c 为主要判断证据,在计算绕行代价 I_c 时,采用 w 表示 EAR 等级从高到低的权重[11]。初始 BPA 设定如表 2,设置 $m(\Theta=\{A,B\})=0.1$ 有助于避免证据冲突所带来的判定错误。

图 5 所示为本文方法检测的异常轨迹。由于 Bel(A)表示怀疑轨迹属于异常的程度,Pl(A)表示不怀疑轨迹为异常的程度,从[Bel,Pl]值看,



(b) R_1与R_2速度对比　　(a) 绕行R_1示例

图 3　出租车经验规避路段等级(EAR)

Fig. 3　Experience Levels of Taxi EAR

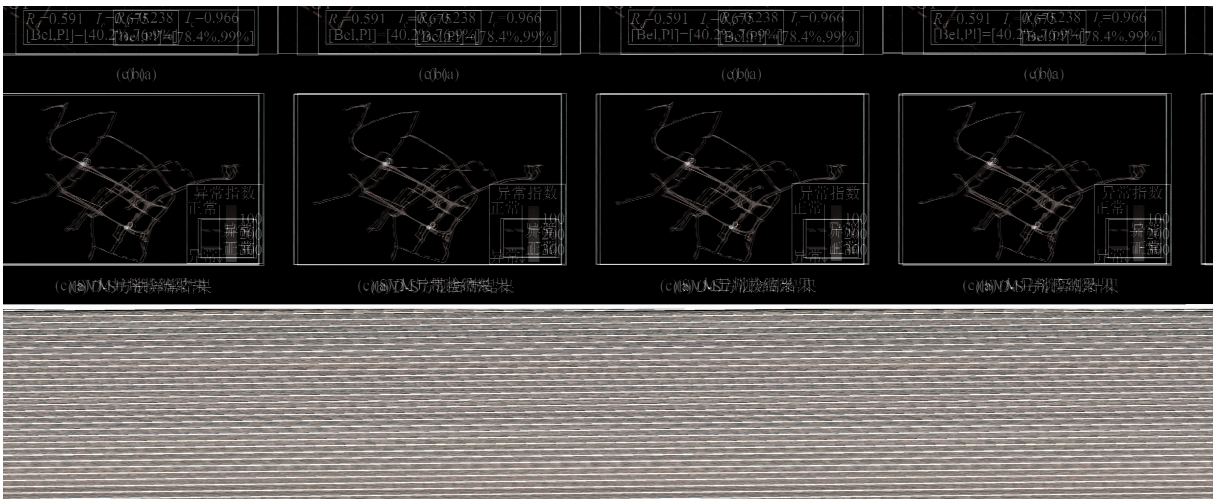


图 4　出租车经验规避路段说明

Fig. 4　Illustrations of Taxi EAR

表 2　初始概率分配函数 BPA 值

Tab. 2　Default BPA Valuse for Each Set

	R_d		I_c	
	$0.1<R_d<0.9$	$R_d\leqslant 0.1$ 或 $R_d\geqslant 0.9$	$0.1<I_c<0.9$	$I_c\leqslant 0.1$ 或 $I_c\geqslant 0.9$
$m(\{A\})$	$1-R_d-0.05$	$1-R_d$	$I_c-0.05$	I_c
$m(\{B\})$	$R_d-0.05$	R_d	$1-I_c-0.05$	$I_c-0.05$
$m(\Theta=\{A,B\})$	0.1	0	0.1	0

Bel(A)＞50%的轨迹可判别为异常，Pl(A)＜50%的路径可判别为正常。图 5(a)轨迹虽然异常度只有 40.2%，但其发生时间为 23:00，所以该绕行不合理程度较大，归类为异常。图 5(b)和图 5(c)中轨迹 Bel 值较大，为 78.4%和 96.2%，属于较明显异常，一般情况下，司机不会主动选择这些路径，其产生原因可能为中间的乘客换乘没有记录使两段旅程合并为一段旅程或者恶意宰客。

将本文实验结果与文献[5]中的 SVM 方法以及人工判读结果相比较(见图 6)。本文方法选取通过某一 OD 对的所有轨迹(共有 391 条轨迹，其中 9 条与最短路径相同)，将所提基于 D-S 的方法检测的异常指数区间以 Bel(A)值升序排列、Pl(A)值降序排列，序列号越大异常度越大，如图 6(b)。SVM 方法仅考虑了将轨迹的线型形态作为特征向量，其结果较大程度地依赖于样本训练的好坏，所以与样本中的异常轨迹形态差别较大的异常轨迹无法得到识别。本文所提基于 D-S 的方法对较大程度的异常检测良好，但对某些轨迹的局部异常效果不佳。由图 6 可以看出，本文结果基本与人工判读结果较一致，明显好于 SVM 的检测效果。

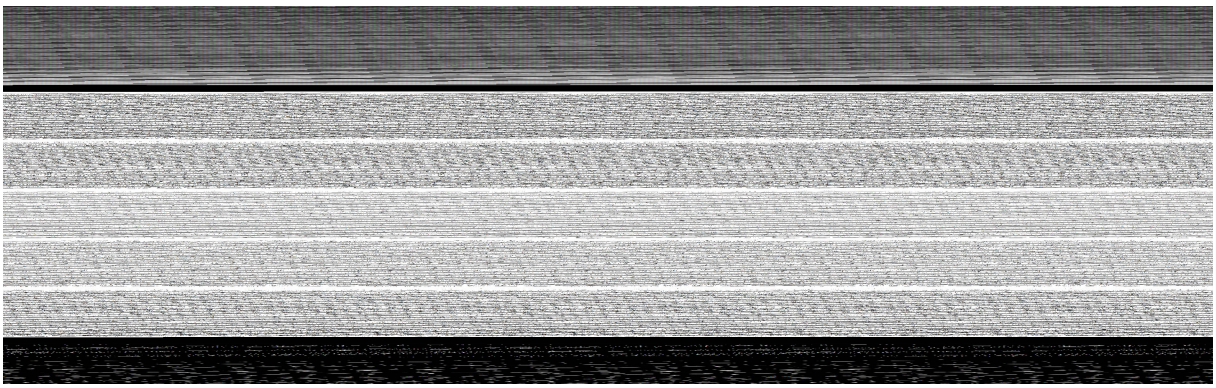


图 5　异常检测结果及 D-S 异常指数示例

Fig. 5　Examples of Anomalous Detection Results and Their D-S Indices

图 6　3 种方法结果对比

Fig. 6　Result Comparison of Three Methods

3　结　语

对于大量的浮动车数据而言,OD 分布广泛多样,以往以同一 OD 对中大多数相似轨迹为参照寻找异常轨迹的方法计算开销非常大,且其效果在样本量少的情况并不理想。本文提出了一种基于证据理论和出租车出行时空约束和择路经验特征的出租车异常轨迹检测方法,通过证据理论对每条轨迹都给定一个异常度区间[Bel, Pl],能有效检测行驶轨迹超出正常范畴较为明显的轨迹。本文提出的方法不基于 OD 对之间的样本数目,通过约束条件判定轨迹的异常度。实验结果表明该方法对异常轨迹效果很好,可用于司机择路行为分析前期的数据过滤。

下一步工作可结合道路等级、速度限制、突发事件等更丰富的交通信息,以及司机行驶过程中的转向次数等更丰富的行为证据检测异常轨迹。

参　考　文　献

[1] Zheng Yu, Liu Yanchi, Yuan Jing, et al. Urban Computing with Taxicabs[C]. ACM International Conference on Ubiquitous Computing, Beijing, 2011

[2] Wang Handong, Yue Yang, Li Yuguang, et al. Spatial Correlation Analysis of Attractiveness of Commercial Facilities[J]. *Geomatics and Information Science of Wuhan University*, 2011, 36(9): 1 102-1 106 (王汉东,乐阳,李宇光,等. 城市商业服务设施吸引力的空间相关性分析[J]. 武汉大学学报·信息科学版,2011,36(9):1 102-1 106)

[3] Liu Yu. Understanding Intra-urban Trip Patterns from Taxi Trajectory Data[J]. *Journal of Geographical Systems*, 2012, 14: 463-483

[4] Fang Zhixiang, Shaw Shilung, Tu Wei, et al. Spatio Temporal Analysis of Critical Transportation Links Based on Time Geographic Concepts: A Case Study of Critical Bridges in Wuhan, China[J]. *Journal of Transport Geography*, 2012, 23: 44-59

[5] Oliva J B. Anomaly Detection and Modeling of Trajectories[D]. Pittsburgh: Carnegie Mellon University, 2012

[6] Lee J G, Han Jiawei, Li Xiaolei. Trajectory Outlier Detection: A Partition-and-Detect Framework[C]. IEEE International Conference on Data Engineering, Cancun, Mexico, 2008

[7] Zhang Daqing, Li Nan, Zhou Zhihua, et al. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces[C]. ACM International Conference on Ubiquitous Computing, Beijing, 2011

[8] Liu Siyuan, Krishnan R. Fraud Detection from Taxi's Driving Behaviors[J]. *IEEE Transactions on Vehicular Technology*, 2014, 63(1):646-672

[9] Ge Yong, Xiong Hui, Liu Chuanren, et al. A Taxi Driving Fraud Detection System[C]. IEEE International Conference on Data Mining, British Columbia, Canada, 2011

[10] Shafer G. Mathematical Theory of Evidence[M]. Princeton: Princeton University Press, 1976

[11] Tang Luliang, Chang Xiaomeng, Li Qingquan. The Knowledge Modeling and Route Planning Based on Taxi Experience[J]. *Acta Geodaetica et Cartographic Sinica*, 2010, 39(4): 404-409 (唐炉亮, 常晓猛, 李清泉. 出租车经验知识建模与路径规划算法[J]. 测绘学报, 2010, 39(4): 404-409)

[12] Li Qingquan, Zeng Zhe, Zhang Tong, et al. Path-Finding Through Flexible Hierarchical Road Networks: An Experiential Approach Using Taxi Trajectory Data[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2011, 13(1): 110-119

[13] Yager R R. On the Dempster-Shafer Framework and New Combination Rules[J]. *Information System*, 1989, 41(2): 93-137

Anomalous Taxi Trajectory Detection Based on Experiential Constraint Rules and Evidence Theory

ZHOU Yang[1] *FANG Zhixiang*[2] *LI Qingquan*[2,3] *GUO Shanxin*[1]

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

2 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

3 Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China

Abstract: In the context of urban transportation, large-scale collections of floating car trajectories are constrained by low sampling rates due to concernsabout data processing and storage. This creates uncertainty when identifying movement trajectories that reflect true route choice behaviors. To reduce uncertainty, this paper presents an approach using experiential constraints based on evidence theory to detect anomalous trajectories in taxi GPS trajectories. The approach employs three factors including the ratio of travel length between GPS traces and shortest distance path, the cost index of experiential avoid roads, and travel start times. The evidences based on the three measurements are combined in an evidence theory framework in order to get the anomalous degree of each trajectory so that the anomalous trajectories whose travel distance and travel route are significantly different from normal ones can be detected. A case study is presented using real world GPS trajectories of over 11,000 taxis. The experimental results in Wuhan show that our method, which is not influenced by the number of trajectories between a single OD pair, has the ability to detect anomalous trajectories and can be applied to clean biased data before route choice analysis using a large fleet of floating cars.

Key words: anomalous trajectory detection; D-S evidence theory; experiential constraint rule; taxi trajectories

First author: ZHOU Yang, PhD candidate, specializes in trajectory analysis and space-time pattern mining of moving objects. E-mail: cleverzhouyang@whu.edu.cn

Corresponding author: FANG Zhixiang, PhD, professor. E-mail: zxfang@whu.edu.cn

Foundation support: The National Natural Science Foundation of China, Nos. 41371377, 41231171; the National 863 Program of China, No. 2012AA12A403-4; Shenzhen Scientific Research and Development Funding Program, Nos. ZDSY20121019111146499, JSGG20121026111056204; Shenzhen Dedicated Fund of Strategic Emerging Industry Development Program, No. JCYJ20121019111128765.