Proceedings of the 29th Chinese Control Conference
July 29-31, 2010, Beijing, China

基于模糊C均值聚类的路段平均行程时间估计方法*

陈　宁

浙江科技学院机械与汽车工程学院，浙江 杭州 310023
E-mail: ningchenzust@yahoo.com.cn

摘　要： 路段平均行程时间是城市交通管理部门、交通行业部门制定调度方案及管控预案的重要信息依据。为了能够有效利用低频采样GPS数据，实现路段平均行程时间的精确估计，本文提出一种基于模糊C均值聚类的路段平均行程时间估计方法：首先利用改进地图匹配算法，对低频采样浮动车数据进行预处理，将路段的上下游交叉口详细分析，准确估计车辆经过路段交叉口的时刻，通过两交叉口的时差计算单车路段行程时间；然后基于模糊C均值聚类(FCM)方法，将某时段的单车路段行程时间分成快、中、慢速三类，利用各类的聚类中心和分类后数据，按一定的权值，估计平均路段行程时间。计算结果表明：平均行程时间估计值与实测值比较接近，能够满足城市城市交通管理部门、交通行业部门的需求。

关键词： 路段平均行程时间，低频采样浮动车数据，模糊c均值聚类，地图匹配算法

Estimation Approach for Average Link Travel Time Based on Fuzzy C-Mean Method*

CHEN Ning

Zhejiang University of Science and Technology, Hang Zhou 310023, P. R. China
E-mail: ningchenzust@yahoo.com.cn

Abstract: In order to availably utilize taxi GPS data with low frequency and precisely estimate the average link travel time (ALTT), an estimation approach of ALTT and Improved Map-matching algorithm was proposed. After analyzing upstream and downstream intersections of link particularly, more accurate estimates the moment of a vehicle through the intersection .Using the time difference of upstream and downstream intersections as single vehicle link travel time (SLTT) .We generate patterns for the high, middle and low speed level by calculating the SLTT using FCM. ALTT calculated by means of a certain weight using the center value of each cluster after confirming of cluster membership of SLTT data. Computation result shows that the relative tolerance is less between the measure value and computation value of ALTT.

Key Words: Average Link Travel Time, GPS Data of Vehicles, Fuzzy C-Mean Method, Map-matching Algorithm

1　引言 (Introduction)

路段平均行驶时间的估计是当今智能交通系统的研究热点[1]。它反映了路段交通状态的重要指标，为交通控制及交通诱导提供可靠的指标，是先进出行者信息系统（ATIS）的基础之一[2]。

传统的路段平均行驶时间估计是“间接”的估测方式，比较有代表性的是线圈检测技术[2,3]。但是，当采集点特别拥挤，使该点流量达到饱和或者过饱和时，大部分方法的误差将大幅度增大，甚至失效。随着GPS技术研究与应用的发展，路段平均行驶时间可以通过采集数据“直接”进行估计[4]。

由于浮动车的技术优势，近几年国内外涌现出大量的相关文献，然而目前的研究方式都是针对理想的高频采样数据进行分析[5,6]。高频采样虽然能一定程度上提高估计的精度，但是数据样本量越大，意味着数据通讯、计算存储负荷以及投资成本更大，故不适合目前我国国情，导致我国城市交通中低频采样应用更加普遍。但是，在我国目前混合式交通现况下，低频采样的瞬时速度无法有效代表路段平均速度[7]。为了解决这一矛盾，本文改进了地图匹配算法，提出了基于模糊C均值聚类的低频采样浮动车数据（FCD）对路段平均行驶时间估计模型，并以杭州的城市交通为例，利用低频采样（30s）出租车GPS数据进行实验验证

2　低频采样浮动车原始数据预处理(Pre-processing for GPS Data with Low Frequency)

GPS浮动车作为一种新型的动态实时交通采集技术，它使用GPS车载装置采集车辆的行驶参数(如时间、速度、坐标、方向等),并将这些数据通过GPRS网络传送到浮动车信息中心，经过汇总、预处理后生成

*此项工作得到浙江省自然科学基金资助, 项目批准号: Y107034

实时的路况交通信息。这些经过预处理的信息输出给交通部门以提供实时交通状态信息。

2.1 改进的地图匹配算法(Improved Map-matching Algorithm)

目前适用于浮动车系统的预处理方法主要为地图匹配算法，分三种：点到点匹配、点到线匹配和线到线匹配三种，模式识别是其支撑技术[8]。浮动车对实时性要求较高，且数据量较大，故算法复杂度高的线线匹配及数据库布置复杂的点点匹配都无法符合要求。我们改进了点到线匹配算法，用来对低频浮动车采样数据进行定位。

传统点线匹配算法，忽略车辆定位的连续性，将车辆与道路的关系视为简单散点到线的距离关系，选择投影距离最短的投影点作为所求的匹配点。鉴于此方法仅考虑垂直距离最短的路段作为匹配路段，相对误差比较大，故引入车辆定位数据中的方向，以及道路的车道宽度作为约束条件，如图1所示，具体步骤如下：

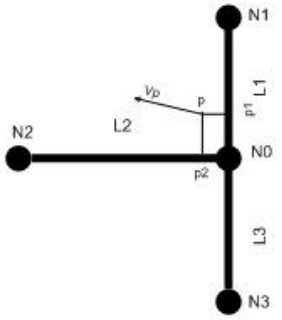

图1 改进的点线匹配示意图

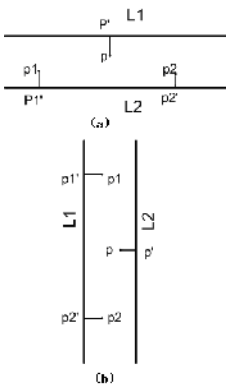

图2 错误匹配

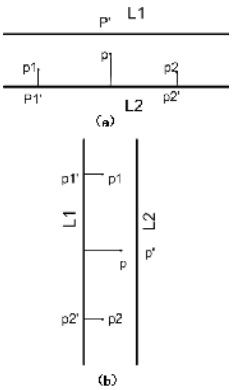

图3 错误匹配修正

1）获取车辆定位点坐标及方向信息。

2）引入道路车道信息，将单纯的线形道路数据变成具体宽度的面形道路。

3）根据车辆定位坐标搜索误差范围内的面形道路作为匹配候选道路。

4）引入车辆方向约束，剔除不符合的候选道路。

5）将最小距离的剩余候选道路作为匹配道路，最终匹配点为p2。

定点匹配算法所得到的初步定位点数据没有考虑车辆行驶的连续性。此算法虽已引入方向约束，仍会出现不稳定情况，如图2（a）和（b）所示。显然这个结果是不符合车辆连续性的要求。故引入历史数据，对将上述错误情况进行修正，结果如图3（a）和（b）所示。

2.2 车辆轨迹跟踪 (Trajecto Tracking)

车辆路径跟踪是确定车辆行驶路线，也是计算路段平均行驶时间的基础。该算法根据车辆ID及道路的拓扑结构，基本上能够确定车辆行驶路线，不足之处在于车辆数据采集频率过低，会出现跨越多个路口的复杂情况，如图4所示，p1、p2为连续采样点，p1、p2的路径可能是ABEF、ADEF、ABCF等。为了确定车辆轨迹，这里引入了图论知识、空间数据库及路网拓扑结构，且认为驾驶员都是正常驾驶状态。

最短路径搜索分为静态最短路径搜索和动态最短路径搜索。目前静态路径主要有Dijkstra算法和A*搜索算法，其中A*算法是比较有效的，但A*算法需要确定评价函数，评价函数的设计将决定搜索时耗以及能否寻找到最优解，过程比较复杂。

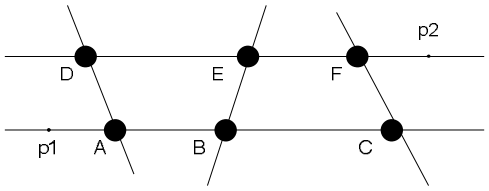

图4 车辆轨迹歧义情况

传统的Dijkstra算法复杂因素在于需要遍历所有路网结构的节点，需要大量的搜索时耗。可是Dijkstra算法对于小区域内（交叉口少）的静态最短路径搜索算法非常有效， 20个路口可以在1秒左右搜索完毕，故能解决对图4之类的歧义情况。GPS低频采样的时间间隔大，即使中间还可能存在被剔除的错误点，两个车辆定位点间的距离也不会很长，经过的路口也相对较少。结合以上因素，基于空间数据库（如Oracle Spatial）的小区域Dijkstra最短路径搜索算法，参考图4，车辆轨迹确定主要步骤如下：

1）获取车辆定位匹配点p1、p2。

2）搜索距离两定位点空间最短路口，如A，F。

3）按照正常驾驶行为，利用步骤2中所确定的路口通过空间搜索算法（如矩形区域搜索、圆形区域搜索）搜索两路口间可能通过的路口。

4）根据步骤2确定的路口及步骤3所搜索的路口行程小区域的路网，采用Dijstra算法搜索的最短路径

作为该车辆行驶轨迹。

3 路段平均行程时间估计(Estimation of Average Link Travel Time)

3.1 改进端点时差法 (Improved Map-matching Algorithm)

对匹配后的浮动车数据分析后，发现在交通高峰时段及平峰时段，单车瞬时速度为零（包括极低车速，下同）的样本占绝大部分，这是因为我国城市普遍存在的机动车与非机动车并行的混合交通。斑马线不仅分布在道路交叉口，还分布在很多路段的路段中间，这将导致浮动车所采集的单车速度样本不符合实际交通状况。

按照采样车辆瞬时速度估计路段平均速度（或路段平均行程时间），会造成很大误差。根据实验中匹配后地图的显示，以及文献9交叉口停车次数及停车率分析，发现采样数据在有信号灯交叉口的入口比较密集，在交叉口的出口相对松散，路段中间相对很少。

因此，我们将对交叉口车辆分析，根据端点时差计算路段行程时间。传统的端点时差法依据交叉口相邻采样点的距离之比等于相邻点的时间之比，即两对相邻点之间的平均速度相等，确定t_a：

$$t_a = \frac{L_1 t_p + L_2 t_{p-1}}{L_1 + L_2} \tag{1}$$

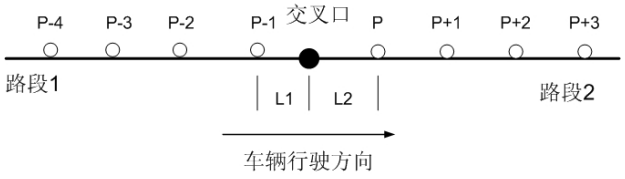

图5 端点时差法

显然采样频率越高，行程时间的精度越高，反之，则精度越低，甚至会错误。

鉴于此，交叉口某车辆的采样数据主要分为三种情况：

1）交叉口入口及出口范围都有单车数据;

2）仅在交叉口出口范围有单车采样数据;

3）仅在交叉口入口范围有单车采用数据

上面提及的交叉口范围根据文献10定义，由于采样频率比较低，可适当延长交叉口出口范围，以便获取更多采样数据。将该三种方法视为三种计算策略，输入采样点的信息，选择自己所需策略来计算车辆经过此交叉口的时刻，使端点时差法适应低频采样。

1）对于第一种情况，若v_p不为零，则有

$$t_a = t_p - (L_2/v_p + \varepsilon) \tag{2}$$

其中:ε表示交叉口对车辆所造成的延误(下同)。

若v_p为零，v_{p-1}不为零，则当p点为空车数据，即为交叉口乘客上车或者下车数据。

则有

$$t_a = t_{p-1} + L_1/v_{p-1} + \varepsilon \tag{3}$$

若v_{p-1}为零，则车辆处于交叉口等待状况中，则有

$$t_a = t_{p-1} + L/TL_1 \tag{4}$$

其中:L为平均排队长度;T为该相位平均红灯时间。

2）对于第二种情况，若v_p不为零，则采用公式1来计算;若v_p为零，视该采样点为空车数据，将其剔除。

3）对于第三种情况，若v_{p-1}不为零，则采用公式2来计算;若v_{p-1}为零，则采用公式3来计算。

3.2 FCM路段平均行程时间估计(Estimation of Average Link Travel Time Based on FCM)

在智能交通应用中，通常把实时检测的交通数据整合为一个状态分析时段（本文中以5分钟为例）。用5min中所有车辆路段行程的平均值表示路段平均行驶时间。单车数据通过GPS出租车采集，受很多因素影响，如乘客上下车，空车寻找乘客和交叉口延时，若简单的采用算术平均法来估计路段平均行驶时间，则误差很大。因此，我们引入了模糊C均值聚类（FCM）方法，使计算的路段平均行程时间尽量与实际行驶时间接近。

FCM技术认为分类样本集合中的每一个样本均以不同的隶属度属于某一类.因此某一类就认为是样本集合上的一个模糊子集，于是每一种这样的分类结果所对应的分类矩阵，就是一个模糊分类矩阵[11]。

设$U=(u_{ik})_{n*c}$为模糊分类矩阵（其中，n表示样本个数，c表示分类数，u_{ik}表示第i个样本属于第k个分类的隶属度），设$X=\{x_1, x_2 ... x_n\}$为被分类样本集合，其中每一个样本x_i均有m个特性指标，即$xi=\{x_{i1}, x_{i2} .. x_{im}\}$。将样本集合$X$分成$c$类（$2<c<n$），设$c$个聚类中心向量为$T$，为了得到最优模糊分类，定义了一个目标函数$J_{FCM}(U,T)$为

$$\mathrm{J_{FCM}(U,T)}=\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}{}^{m} \left\| x_i - t_k \right\|^2 \tag{5}$$

其中:v_k是聚类中心;$m\in[0,+\infty)$为一模糊加权指数，m是为了灵活变动对x_i的隶属度。目标函数表示各类中特征点到聚类中心的距离平方和，聚类问题即要求满足式（5）的U和T，从而使目标函数达到最小值。

模糊C均值算法是通过对公式（5）的目标函数$J_{FCM}(U,T)$的迭代优化来获取对数据集的模糊分类，即迭代

$$u_{ik}=\left[\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}\right]^{-1},d_{ik}=0,1\le i\le n$$
$$u_{ik}=1,\qquad d_{ik}=0,k=1$$
$$u_{ik}=0,\qquad d_{ik}=0,k\ne i \qquad (6)$$

$$t_k=\sum_{i=1}^{n}u_{ik}{}^{m}x_i\Big/\sum_{i=1}^{n}u_{ik}{}^{m},1\leqslant k\leqslant c \qquad (7)$$

使目标函数$J_{FCM}(U,T)$收敛到一个局部极小点或鞍点，得到X的一个最优模糊C划分$U=[u_{ik}{}^{*}]$。

利用FCM，根据聚类中心的t_f、t_m、t_l，将数据分成三类，分别为快速、中等、慢行。通过三类的中心值计算路段平均行驶时间：

$$\bar{t}=\sum_{k=1}^{c}t_k\Big/c \qquad (8)$$

公式（8）计算方式偏向于中间值，忽略了出租车数据频繁停车的特性，与实际交通状况不相符，因此需定义新的计算方式，故引入下面计算公式：

$$\bar{t}=\frac{\sum_{j=1}^{n}T_f+\sum_{j=1}^{m}\left(\frac{m}{m+n}T_{m+l}+\frac{n}{m+n}t_f\right)}{N} \qquad (9)$$

其中:T_f表示以t_f为聚类中心的快速行程时间集合;T_{m+l}表示T_m与T_l两个集合的并集;t_i表示T_f集合中最大值;n表示T_f类中元素的个数;m表示剩下元素个数。

许多城市的浮动车数据来源于GPS出租车，原始数据大多数偏向于慢行数据，因此通过公式（9），对慢行数据处理，使慢行数据按一定的权值进行提速，使计算结果更加精确。

4 计算实例(An Example)

为了验证本文所提出的计算方法，采选取杭州市环城西路某路段北向南方向作为实验路段进行试验。该路段全长约为430m，基本封闭的，上下交叉口均为有信号控制上下游交叉口均为有信号控制的行人过街交叉口。选择这样的路段进行分析，一方面避免车辆丢失，能获取较完整的数据;另一方面路段上下游交叉口都具有信号控制，与大部分的路段类似，具有典型性。

选取出租车GPS数据13:30~14:30时段（平峰时段）及16:30~17:30时段（上下班高峰时段）作为实验数据。采用本文提出的计算方法进行计算，并与杭州市公安局交通警察支队科研所提供的该路段实测值进行对比，结果见表1、2：

表 1　平峰时段路段平均行程时间

时段	估计平均行程时间（s）	实测平均行程时间（s）
13：30-13：35	37.45	48.3
13：35-13：40	51.69	49.6
13：40-13：45	50.3	53.2
13：45-14：50	53.94	4903
13：50-13：55	47.91	45.7
13：55-14：00	42.16	52.3
14：00-14：05	43.73	46.3
14：05-14：10	47.32	50.1
14：10-14：15	51.91	48.2
14：15-14：20	44.13	49.5
14：20-14：25	50.48	52.3
14：25-14：30	55	57.5

表 2　高峰时段路段平均行程时间

时　段	估计平均行程时间（s）	实测平均行程时间（s）
17：00-17：05	55.04	58.2
17：05-17：10	57.43	59.3
17：10-17：15	57.45	62.2
17：15-17：20	64.44	78.9
17：20-17：25	87.18	89.8
17：25-17：30	92.51	100.7
17：30-17：35	84.29	96.2
17：35-17：40	80.74	106.8
17：40-17：45	128.68	115.6
17：45-17：50	123.32	124.5
17：50-17：55	1860.2	118.5
17：55-17：60	112.16	107.2

表1、2的结果表明利用本文所提出的方法整体上的估计值跟实际值比较接近，但是，仍然存在个别时间点上误差很大的现象，经分析主要原因有两个：一方面误差来源于在求单车路段行程时间时，车辆状况没有列入前述分析的三种情况中;另一方面与样本数量过少有关，即使对权值优化，结果还会有所偏差。但从整体上看，估计值与实测值较接近，整体估计结果可信度较高。

5 结论(Conclusions)

利用采用浮动车法估计路段行程时间时，GPS数据仅仅提供瞬时速度这一个交通参数，需要结合GIS来估计单车平均路段行程时间。为了克服传统端点时差无法估计低频采样浮动车的缺点，能够获得较准确的单车行程时间数据，本文对路段上下游采集样本数据进行详细分析，针对各种不同采用情况分别计算方案，很好的覆盖了绝大部分浮动车交叉口出现状况。

利用FCM对某时段单车进行分类以及对分类后的聚类中心及每个类的元素进行加权平滑，既能较好的排除异常数据，也能较好弥补作为浮动车主要数据来源（GPS出租车）本身的缺点。通过实例证明该方法切实可行，能较好的估计路段平均行程时间。

但由于交通状况极为复杂，当浮动车数据样本很少时，计算结果会有所偏离，该方法还需要做进一步的研究和改进。

参考文献(References)

[1] 杨兆升. 关于智能运输系统的关键理论——综合路段行程时间预测的研究[J].交通运输工程学报，2001，1（1）：65267.

[2] 张和生，张 毅，胡东成.路段平均行程时间估计方法[J].交通运输工程学报，2008.2。

[3] 朱 中，杨兆升.基于卡尔曼滤波理论的实时行程时间预测模型[J].系统工程理论与实践，1999，19（9）：74278.

[4] Chen，M. & Chien，S.I.J. Dynamic freeway travel time prediction using probe vehicle data[J].Transportation Research Board， 80th Annual Meeting.2001.

[5] 李筱菁，孟庆春，魏振钢等. GPS技术在城市交通状况实时检测技术中的应用[J].青岛海洋大学学报，2002，32（3）：475-2481.

[6] Y.Li and M. McDonald. Link travel time estimation using single GPS equipped probe vehicle[J].IEEE :Intelligent Transportation System， September 2002，932-937.

[7] 朱 鲤，杨东援.基于低采样频率浮动车的行程车速信息实时采集技术[J].交通运输系统工程与信息，2008.8.

[8] 王 楠，王勇峰，刘积仁.一个基于位置点匹配的地图匹配算法[J].东北大学学报，1999，20（4）：344- 2347.

[9] 李文勇，黄 辉.信号交叉口停车延误调查及分析[J].广西交通科技，2003.5.

[10] 于 泉，孙 玲，荣建基.于浮动车数据调查方法的交叉口延误计算[J].重庆交通大学学报，2009.4.

[11] 李柏年.加权模糊C-均值聚类[J].模糊系统与数学，2007.2.