

利用 GPS 数据估计路段的平均行程时间

张和生^{1,2}, 张毅³, 温慧敏⁴, 胡东成³

(1. 北京交通大学 电气工程学院, 北京 100044; 2. 轨道交通控制与安全国家重点实验室, 北京 100044; 3. 清华大学 自动化系, 北京 100084; 4. 北京交通发展研究中心, 北京 100055)

摘要:对利用出租车 GPS 数据估计路段平均行程时间的方法进行了研究。在 GPS 数据误差修正的基础上, 根据 GPS 数据量的不同, 对大样本数据量采用样本均值估计路段平均行程时间, 对小样本数据量采用顺序统计量中位数估计路段平均行程时间, 并计算了估值的置信区间和置信度。采用实际出租车 GPS 数据进行估计并与线圈数据估计值进行了比较, 结果相差较小。说明该估计方法能够用于实际路段平均行程时间的估计。

关键词:交通运输系统工程; 路段平均行程时间; 出租车 GPS 数据; 顺序统计量

中图分类号:U491 **文献标识码:**A **文章编号:**1671-5497(2007)03-0533-05

Estimation approaches of average link travel time using GPS data

Zhang He-sheng^{1,2}, Zhang Yi³, Wen Hui-min⁴, Hu Dong-cheng³

(1. School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China; 2. State Key Laboratory of Rail Traffic Control and Safety, Beijing 100044, China; 3. Department of Automation, Tsinghua University, Beijing 100084, China; 4. Beijing Transportation Research Center, Beijing 100055, China)

Abstract: An approach to estimate the average link travel time (ALTT) using taxi GPS data was developed. In this approach the GPS data is calibrated first. Then according to the amount of GPS data samples different estimation methods are employed. For large samples, the average value of the samples is used to estimate ALTT; while for small samples, the order statistics is used. The confidence interval and the degree of confidence of the estimated ALTT are also calculated. The estimation of ALTT using real taxi GPS data was compared with that using loop data, and result shows that the estimation error using GPS data is lower than that using loop data, indicating that the proposed approach can be applied in practice.

Key words: engineering of communications and transportation system; average link travel time; GPS data from taxi; order statistics

路段平均行程时间估计是智能交通系统的重要理论问题之一^[1], 目前通常根据环形线圈检测数据估计路段平均行程时间。由于线圈检测数据是截面数据, 而路段平均行程时间是反映路段区间的参数, 所以对于动态时变交通系统, 利用线圈

数据估计平均行程时间很难满足实时性和准确性约束条件^[2]。国外已经利用 GPS (Global Positioning System 全球定位系统) 数据估计交通状态、路段行程时间^[3-5]。在国内, 随着 ITS 的发展, 一些出租车安装了 GPS 接收器。GPS 数据记

收稿日期: 2006-09-18.

基金项目: “863” 国家高技术研究发展计划项目 (2006AA11Z231); “973” 国家基础研究发展计划项目 (2006CB705506); 轨道交通控制与安全国家重点实验室开放课题资助项目 (SKL2007K011).

作者简介: 张和生 (1970-), 男, 副教授, 博士. 研究方向: 控制工程, 信息处理. E-mail: hszhang@bjtu.edu.cn

录了出租车的位置(经纬度)、运行速度等参数,能够反映车辆的路段运行状态。加之出租车流动性大、分布广,弥补了线圈检测数据作为截面数据难以估计路段平均行程时间和环形线圈埋设仅限于主干道,无法得到次要道路状态的缺陷。

受 GPS 接收器本身精度、GPS 数据通信等影响,用于路段平均行程时间估计的有效 GPS 数据量较小,利用出租车 GPS 数据进行路段行程时间估计时,必须能对 GPS 数据自身误差进行修正,并且能根据不同的 GPS 数据样本估计平均行程时间及计算置信区间。

本文利用北京出租车 GPS 数据研究了 GPS 数据误差修正方法和大、小样本 GPS 数据估计路段行程时间的方法。对于小样本 GPS 数据,提出了利用顺序统计量方法估计路段平均行程时间。

1 GPS 数据误差修正方法

利用出租车 GPS 数据分析交通状态,需要解决出租车 GPS 接收器自身误差的修正问题^[6]。车辆运行时由于 GPS 接收器收不到信号或有较强的干扰信号,GPS 经纬度数据与真实值相差较大。车辆静止时由于 GPS 接收器计算精度和外界多径效应引入的干扰,位置数据不是同一值而是小范围值,造成车辆行驶的假象。因此消除 GPS 数据误差主要是消除位置漂移和速度漂移。

1.1 消除车辆运行时误差的方法

消除车辆运行时误差可采用 GPS 经纬度数据与 GIS 数据进行匹配的方法,计算步骤如下:

步骤 1:确定选择路段,在 GIS 中定位,找出其经纬度范围。

步骤 2:检索 GPS 数据中的相关数据。将任意时刻 GPS 数据的经纬度记为 $P_i(x_i, y_i)$,按

$$\begin{cases} x_i \in [\min(x_0, x_D), \max(x_0, x_D)] \\ y_i \in [\min(y_0, y_D), \max(y_0, y_D)] \end{cases} \quad (1)$$

进行判断,将在路段经纬度范围的数据初步认定为所选路段的 GPS 数据。

步骤 3:根据时间顺序区分不同运行方向的 GPS 数据。

步骤 4:选择路段的某一方向,在初选定的 GPS 数据 $P_i(x_i, y_i)$ 中,按照下式匹配 GIS 数据

$$\begin{cases} x_i \in [x - \Delta, x + \Delta] \\ y_i \in [y - \Delta, y + \Delta] \end{cases} \quad (2)$$

式中: Δ 是匹配裕度。

若 $P_i(x_i, y_i)$ 在路段经纬度范围内,则此

GPS 数据与 GIS 数据匹配良好。对于误差较大的数据,可以采用投影方式把其投影到相关路段中或者在进行速度估计时摒弃。

1.2 消除车辆静止时误差的方法

图 1 所示为对出租车在北京西四大街电子配套市场(经纬度为(11622.0241,3955.5861))停车 58 s 采集的 GPS 数据的经纬度变化。图中 * 为经纬度的变化,共 21 次。

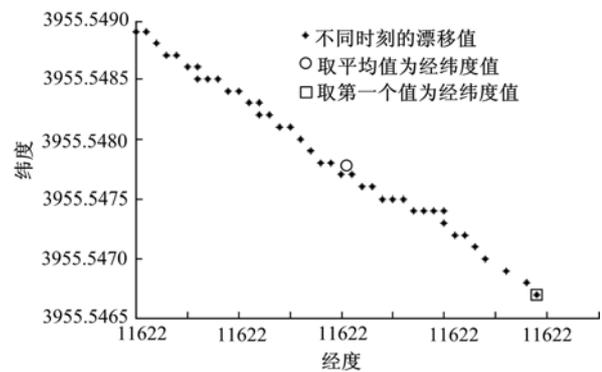


图 1 在(11622.0241,3955.5861)的速度漂移

Fig. 1 Speed drift at (11622.0241,3955.5861)

最大漂移距离是经度 9.24 cm 和纬度 6.78 cm。由于车辆静止时误差的绝对数值不大,消除这类误差可以采用两种方法:一是对所有速度为零点的经纬度进行平均。设某时段共有 N 个连续速度为零点,经纬度为 $P_{0i}(x_{0i}, y_{0i})$,则速度为零对应的经纬度 $P_0(x_0, y_0)$ 为

$$\begin{cases} x_0 = \sum_{i=1}^N \frac{x_{0i}}{N} \\ y_0 = \sum_{i=1}^N \frac{y_{0i}}{N} \end{cases} \quad (3)$$

图 1 中“○”标记为取平均值速度为零的经纬度。

二是取第一个速度为零点的经纬度值,如图 1 中“□”标记。经过误差修正后的 GPS 数据可用于估计路段平均行程时间。

2 估计路段平均行程时间的方法

在实际交通管理中,通常把实时检测的交通数据整合为 5 min 一个状态分析时段。用 5 min 交通参数平均值表示路段平均交通状态,因此一天共有 288 个时段。为表示方便,用变量 k 表示一天中的 $[t_k, t_{k+1}]$ ($k=1, 2, \dots, 288$)时段。由于 GPS 数据包含经纬度信息和时间信息,因此可直接根据 GPS 数据计算每辆车的路段行程时间。

设在第 k 时段进入某路段的所有 N 辆车中,有 $n_k (n_k > 0)$ 辆装有 GPS 接收器的车行驶完整个路段(包括下游路口)。第 j 辆车 ($j=1, 2, \dots, n_k$) 驶入路段时刻为 $t_i(j)$, 驶出路段下游路口时刻为 $t_o(j)$, 则第 k 时段第 j 辆车的行程时间 $t_k^G(j)$ 为

$$t_k^G(j) = t_o(j) - t_i(j) \quad (4)$$

由于路段平均行程时间是在第 k 时段内所有车辆在该路段上行程时间的平均值^[7], 所以从统计意义上看, 用 GPS 数据估计路段平均行程时间就是用抽样数据推断总体(交通流)参数。

在北京 SCOOT 系统的检测范围内, 每天 7:00~20:00 的每 5 min 时段中, 路段上每个车道平均交通流在 [60, 150] 范围。根据中心极限定理, 可以假定车辆的交通状态参数服从正态分布。

受 GPS 数据接收的准确性和通信的影响, 每个时段能够利用的有效 GPS 数据量不同。为准确估计路段平均行程时间, 对不同抽样数据量应采取不同的估计方法。根据文献[7]的仿真结果, 当某时段 GPS 数据量(样本)达到该时段路段车流量(总体)的 20% 时, 估计行程时间值与行程时间真值误差在 3% 以内; 当样本量为总体的 15% 时, 估计误差在 5% 以内; 而当样本量为总体的 10% 时, 则估计误差在 10% 左右。

2.1 大样本 GPS 数据的估计方法

在第 k 时段进入路段的车辆数为 N_k , 接收的有效 GPS 数据量为 n_k , 当 $n_k \geq 15\% N_k$ 时, 按大样本计算方法估计路段平均参数。此时用 GPS 数据构造的统计量可认为符合正态分布。

(1) 大样本 GPS 数据估计路段平均参数

第 k 时段 GPS 数据为 i. i. d. 样本数据(独立同分布 Independent Identically Distribution)。计算出的每辆车的行程时间 $t_k^G(1), t_k^G(2), \dots, t_k^G(n_k)$ 也是 i. i. d., 且服从正态分布。

第 k 时段所研究路段的平均行程时间 T_k^G 为均值 μ_k 、方差 σ_k^2 的正态分布, 记为 $T_k^G \sim N(\mu_k, \sigma_k^2)$ 。其中均值和方差是该时段进入路段所有车辆参数的统计量。由于实际接收到的 GPS 数据量的限制, 仅能通过 GPS 数据样本估计总体均值 μ_k 和总体方差 σ_k^2 。

在大样本 GPS 数据条件下, 行程时间样本均值、样本方差是路段平均行程时间总体均值和方差的无偏估计。总体车流的路段平均行程时间可采用行程时间样本均值和方差以一定的置信区间表示。

行程时间样本均值 T'_k 为

$$T'_k = \sum_{j=1}^{n_k} t_k^G(j) / n_k \quad (5)$$

行程时间样本方差 S_k^2 为

$$S_k^2 = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} (t_j - T'_k)^2 \quad (6)$$

行程时间样本标准差 S_k 为

$$S_k = \sqrt{\frac{1}{n_k - 1} \sum_{j=1}^{n_k} (t_j - T'_k)^2} \quad (7)$$

(2) 大样本 GPS 数据估计参数的置信区间

已知 $t_k^G(1), t_k^G(2), \dots, t_k^G(n_k)$ i. i. d. 为行程时间总体 $T_k^G \sim N(\mu_k, \sigma_k^2)$ 的样本, 用样本均值 T'_k 和方差 S_k^2 构造 t 分布统计量估计路段平均行程时间(总体均值 μ_k)的置信范围

$$\frac{T'_k - \mu_k}{S_k / \sqrt{n_k}} \sim t(n_k - 1) \quad (8)$$

给定置信度 $1 - \alpha$, 可计算出

$$P\left\{T'_k - \frac{S_k}{\sqrt{n_k}} t_{\alpha/2}(n_k - 1) < \mu_k < T'_k + \frac{S_k}{\sqrt{n_k}} t_{\alpha/2}(n_k - 1)\right\} \quad (9)$$

得到总体均值 μ_k 的置信度为 $1 - \alpha$ 的置信范围

$$\left[T'_k - \frac{S_k}{\sqrt{n_k}} t_{\alpha/2}(n_k - 1), T'_k + \frac{S_k}{\sqrt{n_k}} t_{\alpha/2}(n_k - 1)\right] \quad (10)$$

总体方差置信范围需构造 χ^2 分布的统计量

$$\frac{(n_k - 1) S_k^2}{\sigma_k^2} \sim \chi^2(n_k - 1) \quad (11)$$

可计算出

$$P\left\{\frac{(n_k - 1) S_k^2}{\chi_{\alpha/2}^2(n_k - 1)} \leq \sigma_k^2 \leq \frac{(n_k - 1) S_k^2}{\chi_{1-\alpha/2}^2(n_k - 1)}\right\} = 1 - \alpha \quad (12)$$

得到总体方差 σ_k^2 置信度为 $1 - \alpha$ 的置信范围

$$\left[\frac{(n_k - 1) S_k^2}{\chi_{\alpha/2}^2(n_k - 1)}, \frac{(n_k - 1) S_k^2}{\chi_{1-\alpha/2}^2(n_k - 1)}\right] \quad (13)$$

在大样本 GPS 数据条件下, 从 GPS 数据估计路段平均行程时间是以置信度 $1 - \alpha$ 的置信区间给出的无偏估计。

2.2 小样本 GPS 数据的估计方法

若第 k 时段进入路段的车辆数为 N_k , 当 $n_k < 15\% N_k$ 时, 按小样本计算方法估计路段平均参数。由于实际 GPS 数据样本数量小, 样本分布出现非对称的概率较大, 直接用 GPS 数据样本估计总体路段平均行程时间均值误差也较大。此时采

用非参数方法,用 GPS 数据的顺序统计量及中位数估计总体路段平均行程时间。顺序统计量是充分统计量,即 GPS 样本包含车流总体分布中未知参数信息,总体的路段平均行程时间可以从 GPS 数据的行程时间样本顺序统计量推断。

(1)小样本 GPS 数据的样本中位数

第 k 时段有 n_k 个有效 GPS 数据样本 $t_k^G(1), t_k^G(2), \dots, t_k^G(n_k)$ 。将行程时间样本从小到大排序后,得到行程时间样本顺序统计量 $t_{(1)}, t_{(2)}, \dots, t_{(n_k)}$ 。从行程时间样本顺序统计量可计算行程时间样本中位数 M 为

$$M = \begin{cases} (t_{n_k/2} + t_{1+n_k/2})/2, n \text{ 为偶数} \\ t_{n+1/2}, n \text{ 为奇数} \end{cases} \quad (14)$$

用 GPS 数据计算的行程时间中位数可以表示路段总体车流的平均行程时间,但还需计算置信区间。

(2)小样本 GPS 数据估计参数的置信区间

行程时间 $t_k^G(1), t_k^G(2), \dots, t_k^G(n_k)$ i. i. d., 其顺序统计量为 $t_{(1)}, t_{(2)}, \dots, t_{(n)}$, 则顺序统计量 $t_{(r)}$ 分布函数表示至少有 r 个行程时间样本 t_i 小于或等于 t 的概率,即

$$F_r(t) = P(t_{(r)} \leq t)$$

$$F_r(t) = P(t_{(r)} \leq t) = \sum_{i=r}^n C_n^i F^i(t) [1 - F(t)]^{n-i} \quad (15)$$

式中: C_n^r 表示从 n 个数中取 r 个数的组合。

$P(t_{(j)} < m_p < t_{(k)})$ 表示在 m_p 前有 j 个点,并且在 m_p 前不能多于 k 个点,则

$$P(t_{(j)} < m_p < t_{(k)}) = \sum_{i=j}^{k-1} C_n^i F^i(t) [1 - F(t)]^{n-i} = 1 - \alpha \quad (16)$$

$[t_{(j)}, t_{(k)}]$ 为样本置信度为 $1 - \alpha$ 的置信区间。大于 M 和小于 M 的样本点数都服从两项分布 $b(n, 0.5)$, 则置信度 $1 - \alpha$ 满足

$$\begin{aligned} P(t_{(1)}, t_{(2)}, \dots, t_{(j-1)} < M) + \\ P(t_{(j+1)}, \dots, t_{(n)} < M) = \alpha \end{aligned} \quad (17)$$

利用样本中位数 M 可以计算置信度、置信区间和样本数 n 的关系

$$\begin{aligned} P(t_{(1)} \leq M \leq t_{(n)}) = \\ 1 - P(M < t_{(1)}) - P(M > t_{(n)}) = \\ 1 - P(M < t_{(1)}, \dots, M < t_{(n)}) - \\ P(M > t_{(1)}, \dots, M > t_{(n)}) = \\ 1 - P(M < t_1, \dots, M < t_n) - \\ P(M > t_1, \dots, M > t_n) \end{aligned} \quad (18)$$

$$P(t_{(1)} \leq M \leq t_{(n)}) =$$

$$\begin{aligned} 1 - \prod_{i=1}^n P(M < t_i) - \prod_{i=1}^n P(M > t_i) = \\ 1 - \left(\frac{1}{2}\right)^n - \left(\frac{1}{2}\right)^n = 1 - \left(\frac{1}{2}\right)^{n-1} \end{aligned} \quad (19)$$

GPS 数据的样本数目 n_k 确定后,置信度为 $1 - 0.5^{n_k-1}$ 的置信区间是 $[t_{(1)}, t_{(n)}]$ 。当 GPS 数据量 $n_k > 7$, 根据 GPS 数据样本中位数估计的车流平均行程时间置信度 $1 - 0.5^{n_k-1} > 0.99$ 。

(3)小样本 GPS 数据估算方法

用两步搜索最优置信区间确定路段行程时间。

步骤 1: 确定满足置信度 $1 - \alpha$ 要求的备选区间 $[t_{(i)}, t_{(j)}] (i < j)$ 。

步骤 2: 从备选区间中选择长度最短的区间。

3 实际算例

采用北京市实际出租车 GPS 数据,以 5 min 为一个时段估计路段平均行程时间,选择府右街南至南长街路段作为计算实例。在该路段接收的 GPS 数据相对较多,而且该路段位于长安街沿线,进入车辆大多为小汽车和客车,其车辆长度的分布比较均匀,有利于利用线圈数据估计路段平均行程时间。

采集出租车 GPS 数据的时间为 2005 年 12 月 26 日上午 6:30~9:30。共有 39 个可用 GPS 数据样本分布在不同的时段。采用小样本数据估计方法估计路段平均行程时间,并计算置信度,如表 1 所示。

表 1 采用 GPS 数据估计路段平均行程时间

Table 1 Average link travel time estimation using GPS data

时 段	置信度	路段平均行程时间/s	
		估值	估计范围
6:45-6:50	0.75	109	[103, 121]
7:10-7:15	0.5	145.5	[141, 150]
7:20-7:25	0.875	290	[276, 301]
7:35-7:40	0.875	230	[213, 250]
7:40-7:45	0.75	261	[244, 272]
8:00-8:05	0.5	202.5	[190, 215]
8:10-8:15	0.5	127.5	[124, 131]
8:20-8:25	0.5	158	[150, 166]
8:30-8:35	0.5	124	[118, 130]
8:40-8:45	0.5	115.1	[110, 125]
9:00-9:05	0.5	94	[90, 98]
9:05-9:10	0.968	102	[87, 116]

把采用 GPS 数据估计的路段平均行程时间与采用线圈数据估计的路段平均行程时间进行比较。由于环形线圈能够检测进入路段的所有车辆,从数据量上看,利用线圈数据估计的路段平均行程时间优于采用 GPS 数据。

把采用 GPS 数据估算的路段平均行程时间与采用线圈数据估算的路段平均行程时间进行比较,详细计算过程可参阅文献[8]。

图 2 所示为对线圈数据采用 Markov 排队模型估计的路段平均行程时间和对 GPS 数据采用顺序统计量中位数估计的路段平均行程时间的对比。两者估计的路段平均行程时间有一些差别,但差异值不大。

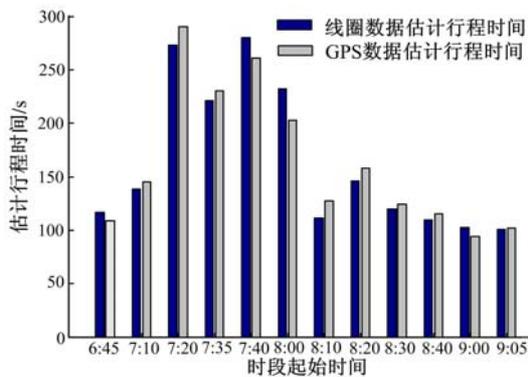


图 2 线圈数据和 GPS 数据估计路段平均行程时间

Fig. 2 Average link travel time estimation with loop data and GPS data

从计算结果可以看出,尽管 GPS 数据量较小,采用顺序统计量的中位数估计路段平均行程时间能够得到比较准确的估计结果。

在估计单个路段的路段行程时间基础上,利用 GPS 数据对北京市闹市口、西单、府右街、南长街、新文化街、西绒线、长椿街、宣武门、和平门、供电局等路口间的路段平均行程时间进行了估计,并与采用线圈数据估计的路段平均行程时间进行对比,最大相对误差小于 12%,说明利用 GPS 数据估计路段平均行程时间是可行的。

4 结束语

通过 GPS 数据与 GIS 数据匹配消除位置漂移,采用算术平均或取第一个速度为零的经纬度数值消除速度漂移。得到反映车辆运行状态的准确 GPS 数据。根据 GPS 数据量大小,采用两种估计方法。对于大样本 GPS 数据,采用 GPS 样本均值估计路段平均行程时间,构成 t 分布和 χ^2 分布计算估计值的置信区间。对于小样本 GPS

数据,采用 GPS 样本构成的顺序统计量中位数估计路段平均行程时间,并计算估值的置信区间。

理论上可证明:大样本 GPS 数据样本均值是对总体车流的路段平均行程时间的无偏估计;小样本 GPS 数据构造的顺序统计量中位数是对总体车流的路段平均行程时间的稳健估计。采用实际 GPS 数据对实际路段进行平均行程时间估计,并与采用线圈数据估计的结果进行了对比,说明利用出租车 GPS 数据估计的路段平均行程时间的方法能够用于交通状态估计、交通控制和交通诱导中。

参考文献:

- [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): 65-67.
Yang Zhao-sheng. Study on the synthetic link travel time prediction model of key theory of ITS[J]. Journal of Traffic and Transportation Engineering, 2001, 1(1): 65-67
- [2] Ran B, Roupail N M, Tarko A, et al. Towards a class of link travel time functions for dynamic assignment models on signalized networks [J]. Transportation Research Part B, 1997, 31(4): 277-290.
- [3] Michael A P Taylor, Jeremy E Woolley, Rocco Zito, et al. Integration of the global positioning system and geographical information systems for traffic congestion studies[J]. Transportation Research, Part C, 2000, 8: 257-285.
- [4] Ruey Long Cheu, Der Homg Lee, et al. An arterial speed estimation model fusing data from stationary and mobile Sensors[C]// Proceedings of IEEE Conference on ITS, Oakland (CA), USA, 2001: 573-578.
- [5] Kerner B, Demir C, Herrtwich R, et al. Traffic state detection with floating car data in road networks[C]// Proceedings of IEEE Conference on ITS, Vienna, Austria, 2005: 700-705.
- [6] 常青, 杨东凯, 寇艳红, 等. 车辆导航定位方法及应用[M]. 北京: 机械工业出版社, 2005.
- [7] 姜桂艳. 道路交通状态判别技术与应用[M]. 北京: 人民交通出版社, 2004.
- [8] 张和生. 基于多源数据的交通状态分析方法研究[D]. 北京: 清华大学自动化系, 2006.
Zhang He-sheng. Approaches for traffic state analyses based on multi-source data[D]. Beijing: Department of Automation, Tsinghua University, 2006.