

# 基于上下车人数的公交客流起迄点估计方法

暨育雄<sup>1</sup>, 妥思佳<sup>1</sup>, 米苏拉尼莱比<sup>2</sup>, 麦寇德马克<sup>2</sup>

(1. 同济大学 道路与交通工程教育部重点实验室, 上海 201804; 2. 俄亥俄州立大学 工程学院, 美国 哥伦布 43210)

**摘要:** 自动计数器 (APC) 系统可自动采集大量乘客上下车数据. APC 数据以聚集的方式表示公交乘客起迄点 (OD) 交通流, 因此, 可用于估计公交 OD 交通流. 据此, 提出基于 APC 数据统计分布的公交乘客起迄点交通流估计模型, 并开发高计算效率的 HEM (heuristic expectation maximization) 算法. 在模型评价比较中, 考虑了 APC 样本量、APC 测量误差和 OD 调查样本量对 OD 估计精度的影响. 结果表明, HEM 法比传统方法更有效地利用 APC 数据, OD 估计精度更高. 另外, HEM 法在计算效率和估计精度两方面都优于近年来新提出的基于仿真的 OD 估计方法.

**关键词:** 公共交通; 起迄点估计; 期望最大化算法; 自动计数器系统

**中图分类号:** U491

**文献标志码:** A

## Estimation of Transit Passenger Origin-destination Flows Based on Boarding and Alighting Counts

Ji Yuxiong<sup>1</sup>, Tuojijia<sup>1</sup>, Mishalani<sup>2</sup> G Rabi, McCord R Mark<sup>2</sup>

(1. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China; 2. The Ohio State University, College of Engineering, Columbus OH 43210, USA)

**Abstract:** A large amount of passenger boarding and alighting counts at stops can be collected by automatic passenger counter (APC) systems. APC data describe passenger origin-destination (OD) flows at the aggregate level and, therefore, can be used to estimate passenger OD flows. In view of this, an OD flow estimation formulation based on the distribution of the APC data is proposed and a computationally efficient heuristic expectation maximization (HEM) algorithm to solve the proposed formulation is developed. The effects of the quantity of the APC data, APC measurement errors and OD survey sample size are considered in the methodology evaluation and comparison. It is revealed that the HEM

algorithm takes better advantage of the APC data than traditional methods and that it outperforms a simulation-based approach that was recently proposed in the literature in terms of both estimation accuracy and computational efficiency.

**Key words:** transit; origin-destination (OD) estimation; expectation maximization (EM) algorithm; automatic passenger counter (APC) system

公交乘客起迄点 (origin destination, OD) 交通流矩阵提供公交网络上两个车站间的乘客交通量. 该矩阵可分为线路级和网络级两个层面. 在公交线路 OD 矩阵中, 起迄点车站属于同一公交线路, 无需考虑乘客换乘; 在公交网络级 OD 矩阵中, 起迄点车站可属于不同公交线路, 需考虑乘客换乘. 本文关注公交线路乘客 OD 矩阵的估计. 线路 OD 矩阵可作为网络 OD 矩阵估计的输入, 提高网络 OD 矩阵估计的精度<sup>[1]</sup>. 公交线路 OD 矩阵在公交规划 (如线路合并、分离和新线路设计) 和公交控制 (如跳站、快车和短线车) 等方面已得到广泛的应用<sup>[2-3]</sup>.

公交 OD 矩阵一般通过交通调查获取. 由于交通调查耗时长, 费用高, 因此调查的样本量一般较低, 调查 OD 矩阵变异性大. 另外, 调查 OD 矩阵的精度还受到调查误差的影响. 相对而言, 获取精确的乘客上下车人数较容易. 乘客上下车数据以聚集的方式表示乘客 OD 交通流. 在一给定车站上车或下车的人数等于以此车站为起点或终点的 OD 交通流的和. 基于乘客上下车数据, 许多研究者提出不同的方法估计公交线路 OD 矩阵. 由于一组乘客上下车数据对应无数个 OD 矩阵, 一些方法提出融合小样本的 OD 调查数据和乘客上下车数据估计 OD 矩阵<sup>[4-5]</sup>, 或使用小样本 OD 调查数据估计乘客出行模式, 然后结合乘客上下车数据估计 OD 矩阵<sup>[6-7]</sup>. 当无其他 OD 信息时, 一些方法则在假设乘客出行模

式的基础上,结合乘客上下车数据估计 OD 矩阵。如, Tsygalnitsky 假设在某一给定车站,车上所有乘客的下车概率相同<sup>[8]</sup>。朱从坤等引入权系数表征用地性质对乘客出行模式的影响<sup>[9]</sup>。Li 和 Cassidy 将车站分为主要车站(如商业区、地铁站区域等)和次要车站,并认为乘客下车的概率为车站类型的函数<sup>[10]</sup>。Li 则假设乘客出行模式服从马尔科夫链分布<sup>[11]</sup>。周雪梅等假设乘客出行模式为乘车距离和用地性质的函数<sup>[12]</sup>。

在诸多公交线路 OD 矩阵估计方法中,实践中最为常用的方法为 IPF (iterative proportional fitting) 法<sup>[5, 13]</sup>。IPF 法计算效率高,人工介入少,易于实施。该方法从一给定的基本矩阵出发,通过反复调整基本矩阵行向量和列向量的和修正基本矩阵,直至基本矩阵的行向量和列向量的和分别收敛于给定的对应车站的上下车人数。基本矩阵可通过小样本的 OD 调查数据或先验信息构建。如无其他 OD 信息,可采用无信息基本矩阵。无信息基本矩阵赋予可行 OD 对相同的正值(如 1)、不可行 OD 对 0 值。在特定条件下,IPF 法和许多其他方法<sup>[8, 10-11]</sup>估计的 OD 矩阵相同。在一般条件下,IPF 法和一些基于频率论和贝叶斯的方法<sup>[4-5]</sup>估计的 OD 矩阵类似。实践应用表明,IPF 法估计的公交 OD 矩阵具有较高的精度<sup>[14-15]</sup>。

从数据利用的角度来看,多数已有的公交线路 OD 矩阵估计方法(包括 IPF 法)主要利用乘客上下车数据的一阶矩(first moment)信息。即,如有多辆公交车的乘客上下车数据,最终 OD 矩阵估计主要满足乘客在每站上车和下车人数的平均值(或总和)。然而,乘客上下车数据的统计分布包含比平均值更多的信息。Hazelton 建立了基于乘客上下车数据分布的统计模型<sup>[16]</sup>。由于模型的复杂性,作者采用马尔科夫链蒙特卡罗(markov chain monte carlo, MCMC)仿真算法估计 OD 矩阵。但是该算法计算效率低,不易于工程应用,并且作者没有对算法的性能进行详细验证。

如今,许多公交网络引入乘客自动计数器

(automatic passenger counter, APC)系统自动采集每个车站上下车人数。APC 系统可持续采集大量的乘客上下车数据,为拟合 APC 数据统计分布提供了数据基础。据此,本文建立了基于 APC 数据统计分布的公交线路乘客 OD 矩阵估计模型,并开发了基于(expectation maximization, EM)法的启发式算法。该算法编程简单,估计精度高于传统方法,在计算效率上远优于 MCMC 法,可用于常规的交通应用。

## 1 模型

由于公交线路总的乘客出行量可较方便地从 APC 数据中估计出,本文采用概率 OD 交通流矩阵表征乘客在某一时间段内的出行模式。概率 OD 交通流矩阵给出乘客在给定时间段内在任意两个车站间出行的概率。假设公交车  $l$  上的乘客 OD 交通流服从 Multinomial 多项分布

$$\mathbf{T}_l \sim M(\mathbf{V}_l, \boldsymbol{\alpha}) \quad (1)$$

式中: $\mathbf{T}_l$  表示公交车  $l$  上的乘客 OD 矩阵; $\mathbf{V}_l$  表示公交车  $l$  上的总乘客出行量; $\boldsymbol{\alpha}$  表示概率 OD 交通流矩阵。乘客在公交车  $l$  上的上下车人数可由  $\mathbf{T}_l$  矩阵中得到

$$b_l(i) = \sum_{j=i+1}^{N_s} \mathbf{T}_l(i, j) \quad (2a)$$

$$a_l(j) = \sum_{i=1}^{j-1} \mathbf{T}_l(i, j) \quad (2b)$$

式中: $b_l(i)$  和  $a_l(j)$  分别表示公交车  $l$  在  $i$  站的上车人数和在  $j$  站的下车人数; $\mathbf{T}_l(i, j)$  表示公交车  $l$  上从站  $i$  到站  $j$  的乘客数; $N_s$  表示总的车站数。另外,使用  $x_l$  表示公交车  $l$  上下车数据的集合;使用  $x$  表示所有公交车上下车数据的集合。

如有小样本的 OD 调查数据,则也假设调查 OD 矩阵服从 Multinomial 多项分布

$$\mathbf{z} \sim M(\mathbf{S}, \boldsymbol{\alpha}) \quad (3)$$

式中: $\mathbf{z}$  表示调查 OD 矩阵; $\mathbf{S}$  表示调查样本量。

基于以上假设,可推出概率 OD 交通流矩阵  $\boldsymbol{\alpha}$  的似然函数  $\mathcal{L}(\boldsymbol{\alpha} | x, \mathbf{z})$ ,如下式:

$$\mathcal{L}(\boldsymbol{\alpha} | x, \mathbf{z}) \propto \prod_{l=1}^L \left[ \sum_{\mathbf{T}_{l,k} \in \Gamma} \prod_{i=1}^{N_s-1} \prod_{j=i+1}^{N_s} \frac{\boldsymbol{\alpha}(i, j)^{\mathbf{T}_{l,k}(i, j)}}{\mathbf{T}_{l,k}(i, j)!} \right]^{A_l} \times \prod_{i=1}^{N_s-1} \prod_{j=i+1}^{N_s} \boldsymbol{\alpha}(i, j)^{\mathbf{z}(i, j)} \quad (4)$$

式中: $L$  表示有 APC 数据的公交车总数; $\boldsymbol{\alpha}(i, j)$  表示乘客在车站  $i$  和  $j$  之间出行的概率; $\mathbf{z}(i, j)$  表示调查得到的乘客在车站  $i$  和  $j$  之间出行的总数; $\mathbf{T}_{l,k} \in \Gamma$  表示所有满足式(2a)和(2b)的 OD 矩阵。式(4)中的

项  $A_l$  表示观测到 APC 数据  $x_l$  的概率。项  $A_l$  的乘积表示观测到所有公交车上的 APC 数据的概率,即 APC 数据的统计分布。另外,式(4)中的项  $B$  表示观测到调查 OD 数据  $\mathbf{z}$  的概率。概率 OD 交通流矩阵的

估计可通过最大化似然函数获得.

如前所述,式(4)中考虑了 APC 数据的统计分布.但是,计算该分布需要对每辆公交车枚举所有满足 APC 上下车人数  $x_l$  的可行 OD 矩阵.即使可能对 OD 矩阵进行枚举,其庞大的计算量也使得模型的应用变得困难.

## 2 算法

上文描述的 OD 估计模型可视为数据遗失问题:每辆公交车的 OD 矩阵为遗失数据;APC 和 OD 调查数据为观测数据;概率 OD 交通流矩阵  $\alpha$  为模型参数.较为常见的处理遗失数据的方法为 EM 法<sup>[17]</sup>.该方法主要由 E(expectation)-步骤和 M(maximization)-步骤组成. E-步骤计算遗失数据的期望值;M-步骤假设 E-步骤计算的期望值为真实观测值,计算参数的极大似然估计.通过重复以上两个步骤,估计的参数将逐步逼近给定似然函数的模式.但是,在本研究中使用 EM 法依然无法避免 OD 矩

阵枚举的问题.因此,本文开发了启发式的 EM 法,称为 HEM (heuristic expectation maximization)法,用于最大化式(4),估计概率 OD 交通流矩阵.其主要步骤为:

(1) 给定一个概率 OD 交通流矩阵的最初估计  $\alpha_0$ .

(2) 对于  $h = 1, 2, \dots$  重复执行以下两个步骤——① E-步骤,近似计算每辆公交车每个 OD 对的期望 OD 值;② M-步骤,假设 E-步骤计算的期望 OD 值为真实观测值,计算概率 OD 交通流矩阵的极大似然估计  $\alpha_h$ .

当收敛标准满足时,步骤(2)的迭代结束.本文使用的收敛标准为:相邻两个循环计算的概率 OD 交通流的最大差异小于  $0.001/N$ ,其中  $N$  为可行 OD 对的总数,或循环次数超过 500. HEM 与 EM 算法的不同在于:EM 需要在 E-步骤进行 OD 枚举;HEM 使用近似的方法计算期望 OD 值,从而避免 OD 枚举.在第  $h$  个循环中,公交车  $l$  上 OD 对  $(i, j)$  的期望 OD 值通过下式近似获得:

$$E_h(T_l(i, j) | x_l, \alpha_{h-1}) \approx a_i(j) \times \frac{E_h(D_l(j | i)) \cdot \alpha_{h-1}(i, j) \left/ \sum_{s=j}^{N_s} \alpha_{h-1}(i, s) \right.}{\sum_{r=1}^{j-1} \left[ E_h(D_l(j | r)) \cdot \alpha_{h-1}(r, j) \left/ \sum_{s=j}^{N_s} \alpha_{h-1}(r, s) \right. \right]} \quad (5)$$

式中:  $E_h(D_l(j | r))$  表示当公交车  $l$  到达车站  $j$  时,公交车上从车站  $r$  上车的期望乘客数.  $E_h(D_l(j | r))$  等于从车站  $r$  上车的总人数减去从车站  $r$  上车、在车站  $j$  上游车站下车的期望乘客数,即

$$E_h(D_l(j | r)) = b_l(r) - \sum_{s=r+1}^{j-1} E_h(T_l(r, s) | x_l, \alpha_{h-1}), \text{ 当 } j > r+1 \quad (6a)$$

$$E_h(D_l(j | r)) = b_l(r), \text{ 当 } j = r+1 \quad (6b)$$

式(5)的推导如下:在一特定车站  $j$ ,假设终点在车站  $j$  的 OD 交通流服从 Multinomial 多项分布,其中总采样数为在  $j$  站下车的人数  $a_j$ ,在  $j$  站下车的乘客来自站  $i$  的概率由式(5)右侧第 2 项近似表示.则式(5)可通过多项分布期望值的公式获得.

HEM 法中的 E-步骤从第一个下客车站开始,对每个车站使用式(5)计算从上游车站上车,在该站下车的期望 OD 值,直到最后一个车站.当对某一车站  $j$  使用式(5)时,在站  $j$  上游车站下车的期望 OD 交通流已知,可先被用于式(6a)和(6b),再代入式(5)计算终点站  $j$  的期望 OD 值.

HEM 法中的 M-步骤相对简单.在第  $h$  个循环中,概率 OD 交通流矩阵中 OD 对  $(i, j)$  的极大似然

估计可通过下式获得:

$$\alpha_h(i, j) = \frac{\sum_{l=1}^L E_h(T_l(i, j) | x_l, \alpha_{h-1}) + z(i, j)}{\sum_{l=1}^L V_l + S} \quad (7)$$

## 3 数值评价

本研究对某一 8.34 km 长、包含 18 个站点、153 个可行 OD 对的公交线路进行大规模的 OD 调查,获取了早高峰时段(8:00~10:00)36 辆公交车上 2 244 位乘客的 OD 信息.由于 OD 调查样本远大于可行 OD 对的总数,认为获取的调查 OD 矩阵可很好地反映真实乘客出行模式,因此假设标准化后的调查 OD 矩阵为真实的概率 OD 矩阵,采用本文介绍的模型框架仿真每辆公交车上的 OD 交通流以及 APC 上下车人数,并使用仿真的 APC 数据估计概率 OD 交通流矩阵.通过将估计的概率 OD 交通流和假设的真实概率 OD 交通流进行对比,评价 OD 估计方法的性能.

基于仿真的500辆公交车APC数据,图1采用散点图对估计的概率OD值和真实的概率OD值进行对比。图中的点越接近对角线,则模型估计精度越高。除了HEM法外,图1还显示了IPF和MCMC法估计的概率OD值。从图1可以看出,对于多数可行OD对,HEM法估计的概率OD值最接近真实的概率OD值;而IPF法高估低概率OD值、低估高概率OD值。总体来说,HEM法估计的概率OD值精度最高,IPF法估计的概率OD值精度最低,而MCMC法估计的概率OD值精度介于IPF和HEM法之间。

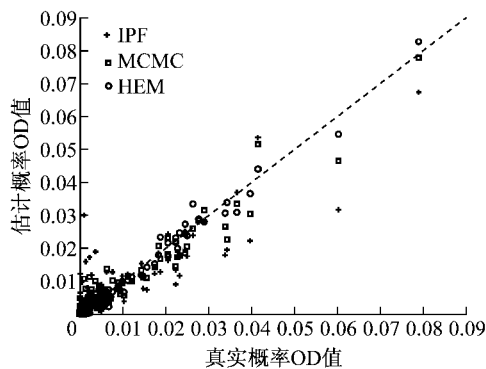


图1 估计概率OD流量与真实OD流量的比较

Fig.1 Comparison of the estimated and true probability OD flows

另外,本文采用 $d_h$ (hellinger distance)指标表示概率OD矩阵估计的整体误差, $d_h$ 定义为

$$d_h = \left[ \sum_{i=1}^{N_s-1} \sum_{j=i+1}^{N_s} (\sqrt{p(i,j)} - \sqrt{\alpha(i,j)})^2 \right]^{1/2} \quad (8)$$

其中, $p(i,j)$ 表示估计的可行OD对 $(i,j)$ 的概率OD交通流。低 $d_h$ 值表示高的估计精度。对应图1的结果,IPF,MCMC和HEM法的 $d_h$ 值分别为0.40,0.26和0.16。同样可以看出,HEM法估计精度最高。另外,从计算效率的角度看,使用苹果笔记本(2.5 GHz处理器),IPF,HEM和MCMC法的计算时间分别为10 s,29 s和16 h。前两种方法更适用于工程实践。

以上的OD估计基于给定公交车数量的无测量误差APC数据。除估计方法外,OD估计精度还受到APC样本量、APC测量误差以及OD调查样本量的影响。图2表示4种情况下以 $d_h$ 指标代表的OD估计精度与APC样本量(即有APC数据的公交车数量)之间的关系。这4种情况考虑了有无APC测量误差和有无OD调查数据的组合。为了量化APC测量误差对OD估计精度的影响,假设观测APC值 $y$ 为真实值 $Y$ 的线性函数: $y=\beta Y$ 。其中 $\beta$ 服从 $(1-w, 1+w)$ 区间内的三角对称分布。 $w$ 值越大,则APC测量误差越大。

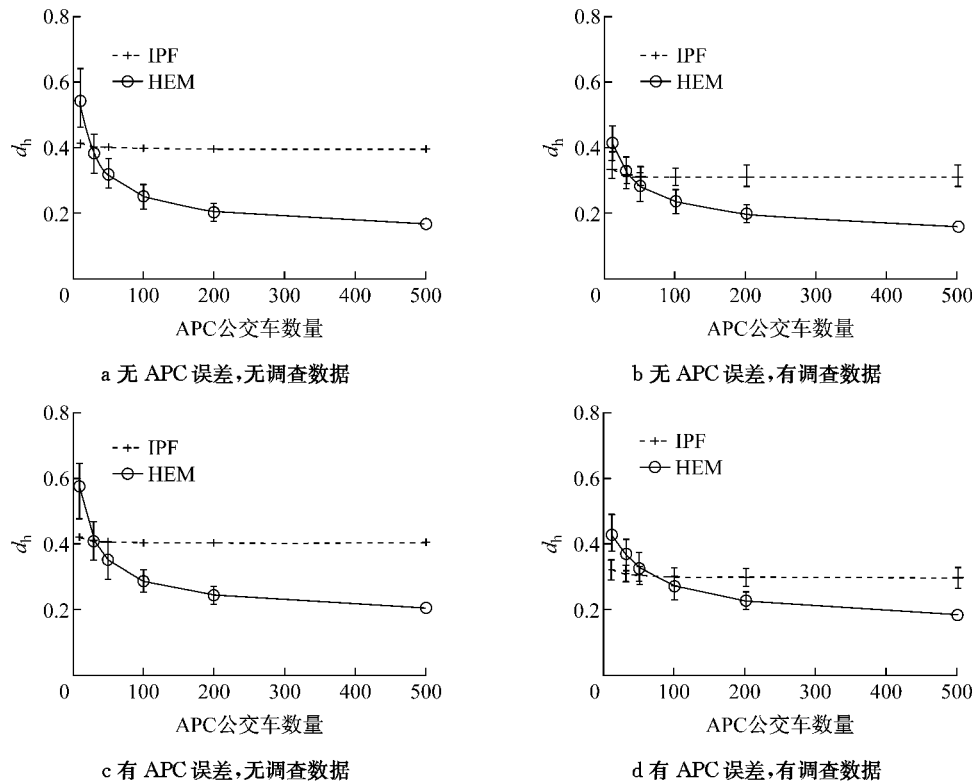


图2 APC样本量、APC测量误差和调查OD样本量对OD估计精度的影响

Fig.2 Effect of the quantity of the APC data, APC measurement errors and OD survey sample size on the accuracy of OD flow estimates

图 2 中有 APC 误差的情况采用  $w=0.15$  仿真观测的 APC 数据. 在这种情况下, 观测的 APC 值  $y$  在 90% 可信度下的置信区间为  $(0.9Y, 1.1Y)$ . 如 APC 数据包含测量误差, 则在 OD 估计前, 对每辆车的 APC 数据进行平衡, 使得总上车人数等于总下车人数, 并且车上人数不为负值. 另外, 图 2 中有调查 OD 数据表示采用随机产生的 100 名乘客的 OD 信息和 APC 数据估计概率 OD 交通流矩阵. 对于给定的 APC 公交车数量、APC 测量误差和 OD 调查样本, 通过重复仿真 100 次获得  $d_h$  性能指标的分布. 图 2 中竖轴的上下限分别表示给定 APC 公交车数量下 97.5% 和 2.5% 分位数的  $d_h$  性能指标. 竖轴和曲线的交点表示  $d_h$  性能指标的平均值. 由于 MCMC 法计算量大, 并且其估计精度低于 HEM 法, 图 2 中没有提供 MCMC 法的结果.

从图 2 中可以看出, 当 APC 公交车数量低时, IPF 法优于 HEM 法. 但是随着 APC 公交车数量的增加, HEM 法优于 IPF 法. 另外, 当 APC 公交车数量较低时, 增加 APC 公交车数量可略微提高 IPF 法的性能. 但是当 APC 公交车数量大于 50 时, APC 公交车数量变化对 IPF 法的性能几乎没有影响. 相反, 当 APC 公交车数量增加时, HEM 法的估计精度有持续明显的提高. 以上结果表明, 与 IPF 法相比, HEM 法可更好地利用大量的 APC 数据以提高 OD 估计精度.

另外, 从图 2 中可以看出, 当存在 APC 测量误差或有调查 OD 数据时, IPF 和 HEM 法的性能差别变小, HEM 法需要更多的 APC 数据以提供比 IPF 法更高精度的 OD 估计. 但是, 当 APC 公交车数量超过 100, 在 4 种情况下, HEM 法都优于 IPF 法. 当公交网络采用 APC 系统时, 可以非常容易地累积获得超过 100 辆公交车的 APC 数据. 因此, HEM 法特别适用于已安装 APC 系统的公交线路 OD 矩阵估计.

## 4 结语

如今在公交网络广泛使用的 APC 系统可持续采集大量的乘客上下车数据. 基于此, 本文提出了基于 APC 数据统计分布的公交线路乘客 OD 交通流矩阵估计模型, 并发展了高计算效率的 HEM 算法. 评价表明, HEM 法比传统的 IPF 法更有效地利用大量的 APC 数据, OD 估计精度更高; 另外, HEM 法在计算效率和估计精度两方面都优于最近文献中

提出的 MCMC 法.

本文采用仿真分析对 HEM 法和其他两种方法的性能进行评价比较, 并量化 3 个因素对 OD 矩阵估计的影响. 仿真中使用的真实概率 OD 矩阵基于大量的调查 OD 数据, 可真实地反映乘客出行模式. 仿真分析的优势在于, 可控制不同影响因素的量级以评价方法在不同环境下的适用性.

采用实际数据对模型进行评价验证是本文下一步的研究工作. 实际数据验证需考虑以下几个因素: ① 需采集足够大的乘客样本以代表真实的乘客出行模式, 但是 OD 调查需花费大量的人力和资金; ② 对不同公交车 APC 测量误差进行量化和修正; ③ 在多条公交线路上进行模型验证, 以反映不同乘客出行模式和不同 APC 测量误差对 OD 估计精度的影响. 另外, 在研究中需认识到随着数据采集技术的发展, APC 测量精度可进一步提高.

本文提出的 HEM 算法同样适用于无路径选择的通道车辆 OD 矩阵估计. 如城市快速路或不收费高速公路(收费高速公路的车辆 OD 矩阵可由收费记录直接得到, 无需进行 OD 估计)进口和出口匝道的交通量已知, 通道车辆 OD 矩阵估计类似于本文关注的公交线路乘客 OD 矩阵估计. 虽然研究者已提出许多方法用于通道车辆 OD 矩阵的估计<sup>[18-19]</sup>, 但是将传统方法与 HEM 法进行评价比较值得尝试. 另外, 在将来研究中, 可考虑将模型推广到网络级公交乘客 OD 估计和网络级公路车辆 OD 估计.

## 参考文献:

- [1] Cui J. Bus passenger origin-destination matrix estimation using automated data collection systems [D]. Cambridge, Massachusetts Institute of Technology, 2006.
- [2] Cortes C E, Jara-Diaz S, Alejandro T. Integrating short turning and deadheading in the optimization of transit services [J]. Transportation Research Part A, 2011, 45(5): 419.
- [3] Tirachini A, Cortes C E, Jara-Diaz S. Optimal design and benefits of a short turning strategy for a bus corridor [J]. Transportation, 2011, 38(1): 169.
- [4] Ben-Akiva M E. Methods to combine different data sources and estimate origin-destination matrices [C]//the 10th International Symposium on Transportation and Traffic Theory. Cambridge: Elsevier, 1987: 459-481.
- [5] Ben-Akiva M, Macke P, Hsu P. Alternative methods to estimate route-level trip tables and expand on-board surveys [J]. Transportation Research Record, 1985, 1037: 1.
- [6] Navick D S, Furth P G. Distance based model for estimating a bus route OD matrix [J]. Transportation Research Record, 1994, 1433: 16.

(下转第 1118 页)