一种基于乘客出行轨迹的公交断面客流估算方法*

胡继华[1,2], 邓 俊[1,2], 黄 泽[1,2]
(1. 中山大学 工学院, 广州 510006; 2. 广东省智能交通系统(ITS)重点实验室, 广州 510006)

摘 要: 为充分利用公交车 GPS 数据和 IC 卡数据来估算公交断面客流,提出一种对刷卡乘客分类推断上、下车站点并扩样叠加轨迹的方法。首先通过融合公交车 GPS 数据和 IC 卡数据来推断不同类别刷卡乘客的上、下车站点;然后,结合投币乘客的出行特征,用低出行频次刷卡乘客的出行轨迹拟合投币乘客的出行轨迹,并对全部乘客的出行轨迹进行叠加;最后以广州公交为例进行了实验。实验结果表明提出的方法可以免交通调查近似得到单条公交线路的断面客流。

关键词: 断面客流; 出行链; 下车站点; 数据挖掘; IC 卡; GPS 数据
中图分类号: U491.1 **文献标志码**: A **文章编号**: 1001-3695(2014)05-1399-04
doi:10.3969/j.issn.1001-3695.2014.05.027

Travel path based method for estimating public transit section flow

HU Ji-hua[1,2], DENG Jun[1,2], HUANG Ze[1,2]
(1. *School of Engineering, Sun Yat-sen University, Guangzhou* 510006, *China*; 2. *Guangdong Provincial Key Laboratory of Intelligent Transportation System, Guangzhou* 510006, *China*)

Abstract: To make full use of smart card data and GPS data recorded in the process of bus running and estimate public transit section flow, this paper put forward a classification method to identify boarding stations and alighting stations of passengers and superpose the passengers' travel paths. At first, it identified boarding stations and alighting stations of smart card passengers by fusing smart card data and GPS data, making the smart card passengers' travel paths acquired. Then it fitted travel paths of passengers who paid by inserting coins according to travel paths of smart card passengers, and superposed all the passengers' travel paths. At last, it took transit system of Guangzhou city for example and conducted some experiments. The experimental results show that the proposed method can approximately estimate public transit section flow of a single bus line without traffic surveys.

Key words: section flow; trip chain; alighting station; data mining; smart card; GPS data

0 引言

公交断面客流是指一定单位时段内某一道路断面的一条公交线路的断面公交客运量或者该断面所有公交线路的断面客运量[1]。单条公交线路断面客流及其时变特征是规划公交车发车频次、考量是否设置区间车的重要依据。

传统的公交断面客流调查方法主要是人工调查法。该方法耗费较多的人力,成本较高。在人工交通调查的基础上,有学者提出了基于历史数据的客流预测方法。Zhao 等人[2]首先对吉林公交历史客流曲线进行小波分解,然后采用神经网络模型对分解的曲线进行客流预测。Chen 等人[3]利用 Hilbert-Huang transform(HHT)方法对台北公交换乘客流数据进行预测分析。杨军等人[4]也以各类大型活动的历史 OD 客流数据为基础,利用灰色预测算法对客流数据建立灰色模型,然后建立马尔可夫修正模型,最后利用预测误差对灰色预测结果进行修正得到最终的预测大客流值。这些方法的数据来源依然是人工调查。近年来,随着公交车 IC 卡收费系统和 GPS 系统在全国各地陆续投入使用,为研究乘客的出行轨迹提供了大量的数据,对个体出行轨迹进行集计研究为获取公交断面客流提供了一种新的思路。然而,由于我国大部分城市的公交线路采取一票制收费,乘客只在上车时刷卡,下车时不需要刷卡,无法直接通过数据融合来获取乘客下车站点,只有部分刷卡乘客的下车站点能根据出行链分析法间接得到。根据吴祥国[1]、章玉[5]对重庆和北京部分一票制公交线路的研究,能用出行链分析法得到完整 OD(即上、下车站点)的乘客占全部刷卡乘客的比例较低。吴祥国通过乘以一个较大的扩样系数来得到全部刷卡乘客的 OD 矩阵,但适用出行链分析法的部分乘客通常乘坐公交的频率较高,不是对刷卡乘客的一个随机抽样。章玉则舍弃出行链分析法采用概率方法来判断下车站点,避免了样本代表性不足的缺点,但没有充分利用已有数据,也难以保证下车站点判断的精度。此外,投币乘客也是公交乘客的重要组成部分,其出行特征与刷卡乘客存在一定差异,难以直接用高频次出行乘客的出行轨迹来拟合得到投币乘客的出行轨迹。因而章玉在计算断面客流时并没有考虑投币乘客。

在出行链分析法中,IC 卡数据和 GPS 数据也没有得到充分利用,通过提高数据挖掘深度,可分析得到出行轨迹的乘客比例还可以进一步提高[6~8],从而获得更具代表性的乘客样本。另一方面,投币乘客通常为低出行频次乘客,在刷卡乘客

收稿日期: 2013-06-26; **修回日期**: 2013-08-19 **基金项目**: 国家自然科学基金资助项目(41271181);国家"863"计划资助项目(2011AA110306)

作者简介:胡继华(1971-),男,河南人,讲师,博士,主要研究方向为地图学与地理信息系统、时态 GIS(hujihua@mail.sysu.edu.cn);邓俊(1990-),男,湖南人,硕士研究生,主要研究方向为公共交通规划;黄泽(1989-),男,广东人,硕士研究生,主要研究方向为城市时空可达性.

中存在与投币乘客出行特征较为相似的群体,确定并分离出这个群体,就可以刷卡乘客的出行轨迹来拟合投币乘客的出行轨迹。

本文以广州市公交系统为例,在准确判断出乘客上车站点的基础上,研究通过将刷卡乘客分类并对不同出行频次乘客采取不同下车站点推断方法来兼顾所有刷卡乘客,提高乘客出行轨迹的取样率,并对不同类别的刷卡乘客进行研究,选取适当出行频次的刷卡乘客来代表投币乘客,进而叠加所有乘客轨迹,得到完整的公交断面客流。

1 单个乘客出行轨迹提取

1.1 乘客上车站点判断

在公交车辆上的 GPS 设备能够实时采集车辆的运行状态信息如时间、速度、经纬度等,进而与该条线路的公交站点位置进行匹配,以获取每条 GPS 数据记录的站点位置信息,确定该条线路每个车次到达各个站点的时刻[9];之后,通过对相应车次的 IC 卡数据记录依据时间进行匹配,就可以确定每条 IC 卡数据记录的上车站点。

1.1.1 公交车到站时间判断

进行公交车到站时间判断所需的基础数据表包括公交 IC 卡刷卡信息表、公交车 GPS 数据表和公交站点基本信息表。各表的主要字段与它们之间的相互关系如图 1 所示。

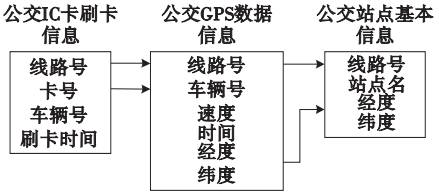

图1 公交车上站时间判断各数据表之间的关联关系

公交车到站时间判断的具体步骤如下:

a)删除 GPS 数据表中速度大于 10 km/h 的数据。

b)计算剩余数据点到各站点的距离,若该数据点与任何一个站点的距离在 100 m 之内,保留该数据点,否则删除。

c)剩余数据点依据其对应的站点分组并按时间依次排列,取这些数据点中最早的时间为公交车到达对应站点的时间。

d)根据每辆车当天刷卡上车的第一位乘客的上车时间,向前推移 5 min 作为公交车到达始发站的时间。

在步骤 a)中,由于广州市公交车的 GPS 数据采集间隔时间为 30 s,采样率较低,所以为了保证数据点的数量,并没有完全剔除非零速点,而是保留了所有速度不超过 10 km/h 的数据。同时由于公交车并不是全天候进行 GPS 数据采样,在早晨首次发车时通常在未进行 GPS 数据采样时已有乘客刷卡上车,因此在步骤 c)之后还应通过步骤 d)来得到公交车到达始发站的时间。在 GPS 采样率较低的情况下,郭婕等人[10~12]提出可根据同一站上车乘客的刷卡时间通常比较集中来对公交 IC 卡乘客的刷卡数据进行聚类分析,进而对公交车到站时间来进行补充和修正。但这种方法需要人工识别,不太适合处理大量数据。本文方法得到的公交车到站时间通常会略早于实际到站时间,但是由于乘客刷卡时间集中在公交车到站后的一小段时间内,所以并不会对上站判断结果造成影响。

1.1.2 乘客上车站点判断

得到公交车上车站点数据之后,就可以结合乘客刷卡数据进行上车站点判断。

进行公交车到站时间判断所需的基础数据表包括公交 IC 卡刷卡信息表、公交车到站数据表和羊城通刷卡数据中车辆号和 GPS 数据车辆号的匹配表。羊城通刷卡数据和 GPS 数据中采用不同的车辆号来标志同一辆车,但两种车辆号的后四位是相同的,在相同线路号的限定下,可以通过简单的 SQL 语句来得到两种数据匹配表。以上各表的主要字段与它们之间的相互关系如图 2 所示。

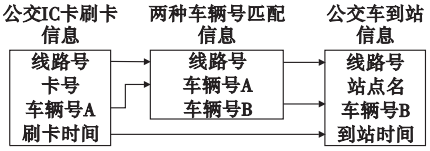

图2 上车站点判断各数据表之间的关联关系

乘客上车站点判断的具体步骤如下:

a)读取一条待推断上车站点的乘客刷卡信息;

b)根据车辆号匹配到对应的公交车到站信息,保留在刷卡时间前后 10 min 之内的公交车到站数据;

c)依次对比刷卡时间和步骤 b)保留的车辆到站时间,若刷卡时间位于某两相邻站点的到站时间之间,则相邻站点中前一个站点即为乘客的上车站点。

1.2 乘客下车站点判断

由于乘客在下车时不需要刷卡,其下车站点无法采用与判断上车站点相同的方法。部分乘客的下车站点可以根据他们多次乘车的上车站点基于出行链分析方法间接推断得出。出行链分析法是在公交乘客的单日全部出行可以形成完整环状出行链结构的情形下,将乘客后一次乘车的上车站点推断为前一次乘车的上车站点[13]。由于乘客可能综合采用公交、出租车、自行车或步行等多种出行方式,并不是所有乘客的公交出行链都可以形成一个完整的环状结构[5]。本文通过重新组织已有数据,提出了适用性更加广泛的下车站点判断方法。

1.2.1 数据预处理

进行乘客下车站点判断所需数据表为乘客刷卡数据表、公交车 GPS 数据表和公交线路表。以这些数据为基础可以得出以下数据表:a)待定下车站点的刷卡乘客列表;b)表 1 中乘客的近期出行信息表;c)上、下行站点序列表;d)上、下行方向站点上站人数表。

其中:数据表 a)c)为已有数据,得到数据表 b)和 d)是数据预处理的关键。用上文的方法依次得到数据表 a)中所有乘客多日出行的上车站点,然后将这些多日出行记录按卡号、刷卡时间排序即可得到数据表 b)。得到数据表 a)中所有乘客当日出行的上车站点后,根据他们的上车站点进行统计即可得到各站点全天的刷卡乘客上车人数,分上、下行方向区分即得到数据表 d)。

以上各数据表的主要字段与它们之间的联系如图 3 所示。

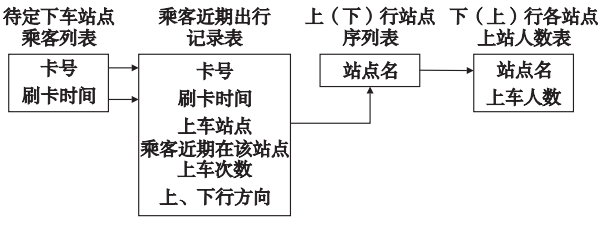

图3 下车站点判断各数据表之间的关联关系

1.2.2 不同类型乘客下车站点判断方法

本文根据乘客多日出行记录将待判断下车站点的乘客分为 A、B、C 三种类型。其中:A 类型乘客为当天公交出行链闭

合的乘客;B 类型乘客为当次公交出行链断裂但多日出行记录中有相似的完整公交出行链的乘客;C 类型乘客为当次公交出行链断裂且多日出行记录中不存在相似的完整出行链的乘客。对于 A 类型乘客,可直接用上文提到的出行链分析法进行判断,乘客的下次乘车上车站点即视为他们的当次乘车下车站点;对于 B 和 C 类型乘客,他们的下车站点是不确定的,本文使用带权重的随机算法来获取他们的下车站点。

对于 B 类型乘客,由于他们有可供借鉴的相似出行链,下车站点在多日出行记录中位于当次上车站点下游的站点中选取,各站点被选中的概率为在该站点上车次数占总数的比重。对于 C 类型乘客,下车站点在当次上车站点的所有下游站点中选取,各站点被选中的概率为当天该站点上车人数占所有下游站点上车人数总数的比重。下车站点判断的具体算法步骤如下:

a)读取一条待定下车站点乘客信息记录。

b)匹配该卡号在乘客近期出行记录表中的位置。

c)得到该乘客该次出行上车站点,根据当次行驶方向站点序列表得到下游站点列表。

d)查找乘客近期出行记录表的下一条记录,与上一条记录卡号和上车站点作对比。

e)若卡号与上条记录相同,且上车站点在下游站点列表中,则乘客属于 A 类型乘客,该上车站点即为上次乘车的下车站点,转步骤 h);否则转步骤 f)。

f)求该乘客的近期出行站点集合和下游站点集合的交集,若该交集不为空,则乘客为 B 类型乘客,根据乘客近期在该交集中各站点上车次数带权重随机选择下车站点,转步骤 h);若该交集为空则转步骤 g)。

g)进入到此步骤的乘客为 C 类型乘客,根据站点全天单方向上车人数在下游站点列表中带权重随机选择下车站点。

h)判断本条记录是否为最后一条记录,若是则结束;否则转步骤 a)。

步骤 f)中得到 B 类型乘客下车站点的带权重随机算法的流程如图 4 所示。

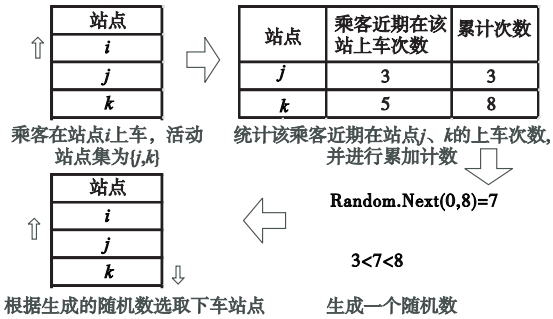

图4　带权重随机算法流程

得到 C 类型乘客下车站点的随机算法与之类似,只需把决定权值的因素换成各站点上车客流量。

2　乘客扩样与出行轨迹叠加

从刷卡乘客扩样到所有乘客,通常使用的方法是以所有刷卡乘客作为样本,除以刷卡乘客占全部乘客的比重直接扩样得到总体。而实际上刷卡乘客和投币乘客通常具有不同的出行特征,投币乘客的出行通常为偶然性的出行,与刷卡乘客中的 C 类型乘客类似,他们的出行目的地通常不是工作地点。因而本文用刷卡乘客中的 C 类型乘客的出行轨迹来拟合投币乘客的出行轨迹,选择对 C 类型乘客进行扩样来得到 C 类型乘客和投币乘客的整体出行情况。

用 a_{ij} 表示 C 类型刷卡乘客 j 出行轨迹引发的 i 和 $i+1$ 站点间断面的客流量,轨迹覆盖该断面则取值 1,否则取值 0。用 $a_{ij}^{'}$ 表示其他刷卡乘客 j 出行轨迹引发的 i 和 $i+1$ 站点间断面的客流量,取值方法同上,n、m 分别为 C 类型刷卡乘客和其他刷卡乘客人数,η 为投币乘客占所有乘客的比重,则站点 i 和 $i+1$ 站点间断面客流为

$$A_i = \sum_{j=0}^{n} a_{ij} + \sum_{j=0}^{m} a_{ij}^{'} \times \left(1 + \frac{(n+m)\eta}{(1-\eta)n}\right) \tag{1}$$

3　实例

本文以广州市 448 路公交 2012 年 8 月 28 日全天刷卡乘客的 IC 卡数据为算例。早高峰时段为 7:00 ~ 9:00,晚高峰为 17:00 ~ 19:00。东圃客运总站开往科林路总站(共 26 站)记为上行方向,科林路总站开往东圃客运总站(共 24 站)记为下行方向。当天 448 路共有 2 203 条刷卡记录,通过匹配判断出上车站点的记录为 2 024 条,其中 A 类乘客 823 人,B 类乘客 974 人,C 类乘客 227 人。投币乘客占总乘客比重根据实地调查结果取 90% 。

本算例中程序采用 C#语言编写,数据预处理耗时约 20 h(由于乘客上车站点为下车站点判断基础数据,上车站点判断耗时全部计入数据预处理耗时中),程序计算耗时约 3 min。

乘客上、下车站点判断的部分计算结果如表 1 所示。

表 1　乘客上、下车站点判断部分计算结果

卡号	上车站点	下车站点	乘客类型
5100000196752570	东圃客运站总站	科林路总站	A
5100000303473060	东圃客运站总站	谭村路口	A
5100000431304180	黄村东路口	玉树新村	A
5100000400207510	黄村东路口	玉树新村	B
5100000328882550	黄村东路口	玉树新村	B
5100000319439080	黄村训练基地	光谱西路中	A
5100000324048120	黄村训练基地	科学城路口	A
5100000301033720	国土资源工程学校	科林路总站	A
5100000309941320	国土资源工程学校	科林路总站	C
5100000333120030	科学城管委会	科林路总站	A

对刷卡乘客进行扩样,并把所有乘客的出行轨迹叠加起来,得到相邻站点间断面客流量,如图 5、6 所示。

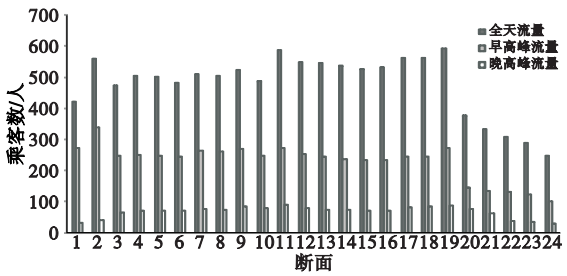

图5　上行方向断面客流量

图 5 中断面 1 表示东圃客运总站至黄村东路口之间的断面,其余依次为上行方向后续相邻站点间的断面。站点序列为东圃客运站—黄村东路口—黄村立交南—边防指挥学校—环场路—地铁黄村站—黄村地铁站—黄村训练基地—广东体育职业学院—奥体路—农工商学院—国土资源工程学校—科学城路口—光宝路口—光谱西路—光谱西路中—谭村路口—科学城管委会—科学大道中—广东软件园—玉树新村—南翔二

路—南云三路口—南云三路中—南云三路南—科林路总站。

图 6 中断面 1 表示科林路总站至南云三路南之间的断面，其余依次为下行方向后续相邻站点间的断面。站点序列为科林路总站—南云三路南—南云三路中—南云三路口—南翔二路—玉树新村—彩频路—谭村路口—光谱西路中—光谱西路—光宝路口—科学城路口—国土资源工程学校—农工商学院—奥体路—广东体育职业学院—黄村训练基地—黄村地铁站—地铁黄村站—环场路—边防指挥学校—黄村立交南—珠村路—东圃客运站。

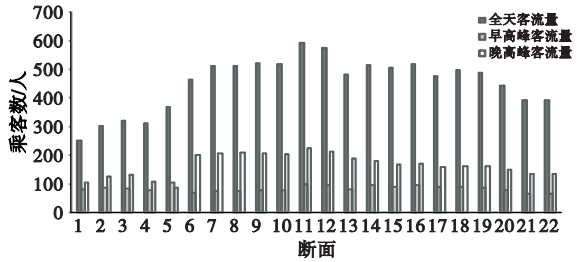

图6 下行方向断面客流量

从图 5、6 可以看出，国土资源工程学校到光宝路口段的断面客流量在整条线路中是最大的，而从玉树新村到科林路总站的客流量则明显小于同时期的其余断面客流量。这与实际调查结果以及乘客的感受是一致的。

4 结束语

本文提出了一种根据乘客多日出行记录对刷卡乘客进行分类的乘客下车站点判断思路，并以广州市公交系统为例详细介绍了判断刷卡乘客上、下车站点的方法，在此基础上提出了通过对刷卡乘客的扩样和叠加乘客的出行轨迹来得到单条公交线路各站间的断面客流的方法，最后结合实例对本文方法进行了验证。算例结果表明，应用这种方法可以近似得到单条公交线路时变的断面客流。本文在下车站点判断中只考虑到了一些来源于常识并可以直观理解的判断方法，尚未考虑用机器学习等方法来研究数据本身可能存在的其他规律，数据挖掘深度还可进一步提高。

参考文献：

[1] 吴祥国. 基于公交 IC 卡和 GPS 数据的居民公交出行 OD 矩阵推导与应用[D]. 济南：山东大学，2011.

[2] ZHAO Shu-zhi, NI Tong-he, WANG Yang, *et al*. A new approach to the prediction of passenger flow in a transit system[J]. Computers & Mathematics with Applications, 2011, 61(8): 1968-1974.

[3] CHEN Mu-chen, WEI Yu. Exploring time variants for short-term passenger flow[J]. Journal of Transport Geography, 2011, 19(4): 488-498.

[4] 杨军，侯忠生. 一种基于灰色马尔科夫的大客流实时预测模型[J]. 北京交通大学学报：自然科学版，2013，37(2)：119-123，128.

[5] 章玉. 基于数据挖掘的动态公交客流 OD 获取方法研究[D]. 北京：北京交通大学，2010.

[6] 师富民. 基于 IC 卡数据的公交 OD 矩阵构造方法研究[D]. 长春：吉林大学，2004.

[7] 闫磊. 基于公交 IC 卡数据的乘客出行时空推算研究[D]. 长春：吉林大学，2012.

[8] 王超. 基于 IC 卡信息的公交客流 OD 推算方法研究[D]. 北京：北京交通大学，2012.

[9] 徐建闽，熊文华，游峰. 基于 GPS 和 IC 卡的单线公交 OD 生成方法[J]. 微计算机信息，2008，24(22)：221-222，218.

[10] 郭婕. 公交 IC 卡通勤乘客 OD 确定方法研究[D]. 南京：东南大学，2006.

[11] 于勇，邓天民，肖裕民. 一种新的公交乘客上车站点确定方法[J]. 重庆交通大学学报：自然科学版，2009，28(1)：121-125.

[12] 陈绍辉，陈艳艳，尹长勇. 基于特征站点的公交 IC 卡数据站点匹配方法研究[J]. 北京工业大学学报，2012，38(6)：885-889.

[13] 陈学武，戴霄，陈茜. 公交 IC 卡信息采集、分析与应用研究[J]. 土木工程学报，2004，37(2)：105-110.

(上接第 1398 页)

参考文献：

[1] NEWTON R M, THOMAS W H. Design of school bus routes by computer[J]. Socio-Economic Planning Sciences, 1969, 3(1): 75-85.

[2] PARK J, KIM B I. The school bus routing problem: a review[J]. European Journal of Operational Research, 2010, 202(2): 311-319.

[3] 党兰学，陈小潘，孔云峰. 校车路径问题模型及算法研究进展[J]. 河南大学学报：自然科学版，2013，43(6)：682-691.

[4] BODIN L D, BERMAN L. Routing and scheduling of school buses by computer[J]. Transportation Science, 1979, 13(2): 113-129.

[5] BODIN L D, GOLDEN B, ASSAD A, *et al*. Routing and scheduling of vehiclesand crews: the state of the art[J]. Computers and Operations Research, 1983, 10(2): 63-211.

[6] 郭强，李育安，郭耀煌. 社区儿童接送服务车辆的线路优化[J]. 西南交通大学学报，2006，41(4)：486-490.

[7] 张苗. 基于双层规划的多目标校车路径优化研究[D]. 成都：西南交通大学，2008.

[8] 张富，朱泰英. 校车站点及线路的优化设计[J]. 数学的实践与认识，2012，42(4)：141-146.

[9] 张玉兵，吴霄翔，任意. 校车安排问题[J]. 高等数学研究，2011，14(1)：122-125.

[10] BRACA J, BRAMEL J, POSNER B, *et al*. A computerized approach to the New York City school bus routing problem[J]. IIE Transactions, 1997, 29(8): 693-702.

[11] PARK J, TAE H, KIM B I. A post-improvement procedure for the mixed load school bus routing problem[J]. European Journal of Operational Research, 2012, 217(1): 204-213.

[12] De SOUZA L V, SIQUEIRA P H. Heuristic methods applied to the optimization school bus transportation routes: a real case[C]//Proc of IEA/AIE. 2010: 247-256.

[13] 党兰学，王震，刘青松，等. 一种求解混载校车路径的启发式算法[J]. 计算机科学，2013，40(7)：248-253.

[14] FÜGENSCHUH A. Solving a school bus scheduling problem with integer programming [J]. European Journal of Operational Research, 2009, 193(3): 867-884.

[15] KIM B I, KIM S, PARK J. A school bus scheduling problem[J]. European Journal of Operational Research, 2012, 218(2): 577-585.

[16] 丁常勇. 合作式校车路径优化问题研究[D]. 大连：大连海事大学，2012.

[17] SAVELSBERGH M W P. The general pickup and delivery problem [J]. Transportation Science, 1995, 29(1): 17-29.

[18] PARK J, TAE H, KIM B I. Corrigendum to "Post-improvement procedure for the mixed load school bus routing problem" [EB/OL]. 2012-11-28. http://dx.doi.org/10.1016/j.ejor.2012. 10.050.