

基于出租车 GPS 大数据的城市区域间可达性评估模型

王英博¹ 单晓晨² 孟 煜³

(辽宁工程技术大学创新实践学院 辽宁 阜新 123000)¹

(辽宁工程技术大学软件学院 辽宁 葫芦岛 125105)² (东北大学计算机科学与工程学院 沈阳 110819)³

摘 要 区域间可达性的评估对城市地面交通出行效率的提高有着重要作用。传统区域间可达性评估方法使用区域间直线距离计算区域间的平均旅行时间,其平均值与实际值的偏差较高,而且基于出租车乘降热点统计的区域间可达性量化方法对于旅行目的地分布不均的区域量化结果过低。针对以上两点不足导致的区域间可达性评估不准确的问题,文中构建了基于 GPS 的区域间可达性评估模型,从出租车 GPS 数据中提炼出完整的旅行来计算实际的旅行时间,以提高平均旅行时间的准确性。在此基础上还提出了一种基于四维 OD 矩阵的可达率计算模型,并以此可达率作为可达性量化标准,从而解决部分区域因发生旅行的目的地分布不均而导致的区域可达性评估不准确的问题。实验表明,提出的可达性评估模型较传统方法而言评估的准确性提高了 9.4%~28.7%,特别是在旅行目的地分布不均的结果区域中,可达性评估准确性的提高更为显著。

关键词 可达性, GPS, OD 矩阵, 大数据, 交通
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.042

Inter-regional Accessibility Evaluation Model of Urban Based on Taxi GPS Big Data

WANG Ying-bo¹ SHAN Xiao-chen² MENG Yu³

(College of Innovation and Practice, Liaoning Technical University, Fuxin, Liaoning 123000, China)¹

(School of Software, Liaoning Technical University, Huludao, Liaoning 125105, China)²

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)³

Abstract The evaluation of inter-regional accessibility plays an important role in improving the efficiency of ground traffic in cities. Traditional inter-regional accessibility evaluation methods make use of the inter-regional linear distance to calculate the regional average travel time, leading to big error between average value and actual value, and the result of inter-regional accessibility measurement method based on hotspot statistics of taxi boarding area quantifying the areas with uneven travel destination distribution is unsatisfactory. In order to solve the problem of inaccurate inter-regional accessibility evaluation caused by the above two points, this paper constructed an inter-area accessibility evaluation model based on GPS, and extracted a complete trip from the taxi GPS data to calculate the actual travel time, so as to improve the accuracy of average travel time. On this basis, this paper proposed a quantitative calculation model of accessibility rate based on four-dimensional OD matrix, and used the accessibility rate as the quantification standard of accessibility to solve the problem of inaccurate evaluation of inter-regional accessibility caused by uneven travel destination distribution of some areas. Experiments show that the accuracy of the proposed accessibility evaluation model is 9.4%~28.7% higher than the traditional method, especially in the area with uneven distributed travel destination, the improvement of accessibility evaluation is significant.

Keywords Accessibility, GPS, OD matrix, Big data, Transportation

1 引言

城市的道路交通网四通八达,为人们的出行提供方便。城市内区域间的可达性能够反映出区域间交通活动的便利程度。随着城市的发展,城市内的部分地区在交通条件不变的前提下可达性很差。造成可达性差的原因有很多,如道路规划不科学和车流激增等。近些年,区域可达性的评估逐渐成为交通领域的研究热点^[1-2]。此外,道路的可达性并不是在所有时间段内都一样,评估区域间各时段的可达性既可以帮助市民合理地规划旅行,也可以为交通管理部门提供解决潮汐

到稿日期:2017-12-19 返修日期:2018-03-22 本文受国家科技支撑计划基金资助项目(2013BAH12F00, MK2013008),辽宁省教育厅科学技术研究一般项目(L2015216)资助。

王英博(1964—),男,博士,教授,CCF 会员,主要研究领域为软件工程、数字矿山、大数据, E-mail: wybustb@126.com; 单晓晨(1994—),女,硕士生,CCF 会员,主要研究领域为大数据、数字图像处理, E-mail: 867352646@qq.com(通信作者); 孟煜(1990—),男,博士生,CCF 会员,主要研究领域为数据挖掘、云计算, E-mail: mengyu@stumail.neu.edu.cn。

交通问题的决策依据。

对于区域间道路可达性的研究,有学者使用 GIS 对可达性进行预测^[7],使用 GIS 地理信息系统是以搜索最短路径为前提的,数据收集工作相对较少;还有学者使用个人信息数据^[8]对不同时间段的可达性进行计算,这可以更加真实地反映不同时间段内旅行路线与旅行时间的关系,但需要大量的数据才能计算出人群普遍的旅行分布。基于位置的方法^[9]的核心是对旅行时间的挖掘与学习,其已被广泛应用于城市规划和地理学研究。Curl 等^[10]认为基于位置的可达性评估是最常用的,因为这种方法评估的准确率较高也更易于解释区域间的沟通。研究人员通常认为基于位置的方法更加适合于通报和预测运输计划的可实现程度。

因此,在之后的很长一段时间内对于区域间可达性的评估一般都是以可达的乘降热点数作为评估标准。可达热点数的计算方法多样,多数方法首选通过聚类生成热点,然后根据旅行平均时间筛选出可达热点。在筛选可达热点时,若以往的方法认为区域间平均旅行时间较短,则所有目的区域涉及的热点都可达,否则均不可达。这种方法在筛选热点的过程中很容易出现以偏概全的情况,特别是对于旅行目的地分布不均的区域来说,由于其到达的目的区域的热点较少,因此其可达性会被评估得较低,但这种情况下其真实的可达性应该是良好的,由此容易造成对这些区域间可达性的误判。一些学者认为区域间旅行耗时近似服从正态分布,因此其利用正态分布规律推测可达热点数,对之前的筛选方法做出改进^[2]。改进后的可达热点数考虑到了正态分布下区域间所有旅行小于平均旅行时间的概率。但是 Cui 等^[2]的研究满足正态分布是要以大量数据为前提的,并且在产生正态分布的过程中是将近似正态分布的数据平滑成正态分布,最终产生区域间旅行时间分布的概率。这一平滑的过程会忽略一部分有效的旅行信息,这样的概率分布与真实的情况依然存在偏差,评估区域间可达性最关键的环节就是对旅行时间分布进行研究,无法准确地计算区域间旅行时间的分布会造成评估结果的不准确。

针对这种由于区域间可达性量化方法不足而导致的可达性评估不准确的问题,本文提出一种基于四维 OD 矩阵的区域间可达率计算模型,并利用区域间可达率作为区域间可达性的量化标准。区域间可达率是区域间发生的所有旅行中旅行时间小于城市内平均旅行时间的旅行数占区域间总旅行数的比率。通过建立四维 OD 矩阵,可以利用所有可用的旅行信息计算出区域间真实的可达旅行的概率。提出的可达性评估模型可有效地解决区域间在旅行分布不均情况下的可达性评估不准确问题,在此基础上还可以避免因旅行涉及到的热点过少而产生可达性不佳的假象。

此外,传统方法在计算区域间旅行时间时需要知道区域间的距离,区域间的距离大多是基于地理信息系统(GIS)数据计算得出的^[3-4],但是通过 GIS 数据只能得到区域间起止点间的直线距离,然而一般来说旅行路线都不是直线,因此其无法反映出真实的旅行线路。这样就会导致区域间平均旅行时间的计算不准确,最终对可达性评估造成较大误差。城市内出租车因为工作特性会选取最合理、用时最短的路线,能够体现起止点区域间的交通特性。本文则在可达性评估模型中利用海量出租车 GPS 数据来弥补 GIS 的不足,从出租车 GPS 大数据中抽出完整的有效旅行,进而计算区域间的平均旅行

时间,有效地解决因利用 GIS 数据无法准确计算区域间平均旅行时间而导致的区域间可达性评估不准确的问题。

本文第 2 节介绍其他学者对道路可达性的研究;第 3 节详细阐述本文提出的区域间可达性评估模型,以及实现的方法和步骤;第 4 节以沈阳市为例对本文提出的模型进行实例分析;第 5 节将实验结果与传统方法的实验结果进行对比,分析两者的不同与优劣;最后总结本文所提模型并得出结论。

2 相关工作及研究

道路可达性这一概念由来已久,KOENIG 等将道路可达性定义为借助某种交通系统从某一区域到达另一区域的完成程度和便捷程度^[5-6],它是度量交通建设水平、路网结构合理性的重要指标。在解决评估区域间可达性这一问题时,可以将问题分为两部分:1)如何对区域间可达性的好坏进行量化;2)如何确定量化标准值。在探索如何量化区域可达性好坏这一问题时,Páez 等^[11]提供了可达性评估的另一种观点,即关于如何进行可达性的评估才能使结果既反映出旅途可以完成多远,又能对不同距离的旅行在基于经验的旅行条件下进行可达性评估。对于可达性的量化方式,华松逸等^[12]用区域间的平均旅行时间作为衡量标准。李龙等^[13]以速度恒定、距离加权的方法计算旅行时间及可达的热点数来量化可达性。但上述方法主要用到统计学的方法,因为无法考虑各区域的特点,所以不可避免地会产生一些错误的估计。本文将区域间不同的旅行分布特点与出租车 GPS 数据相结合,运用大数据的思想对已有的旅行信息进行挖掘,以可达的旅行数量占总旅行数的比率来量化可达性。

结合此量化策略并考虑区域与区域之间的可达性,本文着重研究基于位置信息的方法。该方法首先将被调查的地理区域分为若干个小型分析区域,再从中确定研究区域。从地理位置角度,融合旅行时间、道路规划等客观因素来研究各区域间的可达性,同时结合旅行时间等相关因素产生 OD 矩阵,进而利用 OD 矩阵得出量化值并对区域间的可达性进行评估。

迄今为止,出租车 GPS 大数据已被广泛应用于科学研究,尤其是旅游需求的建模^[14]和旅行时间的估计^[15-16]。张红等^[17]的研究表明,对出租车 GPS 大数据进行计算及知识挖掘,可得到路段行程时间、路段平均速度和道路拥塞程度等信息,还可以获知司机选择路线的倾向以及乘客乘降的密集地点等信息,进而反映城市交通流的信息,既帮助乘客了解出行信息,又帮助司机优化导航路线。然而,在区域间可达性评估领域,前人对于出租车 GPS 大数据的利用较少。本文恰是利用出租车 GPS 大数据可以提供准确的地理位置以及最优路线这一特点,对出租车数据进行挖掘学习,使得到的区域间可达性评估更加客观与准确。

3 区域间可达性评估模型

判定区域间可达性好坏的主要因素是区域间旅行耗费的时长。本节将详细介绍如何进行区域间可达性的评估。首先对已有的 GPS 数据进行筛选;其次从剩下的 GPS 点中抽出完整的旅行,并将旅行分为 8 类;然后以旅行的起点和终点进行聚类,并用聚类出的热点确认学习区域;最后建立 OD 矩阵,从矩阵中计算出区域间可达率,从而评估区域间的可达性。

3.1 数据清洗

为使评估结果更准确,应尽量确保 GPS 序列点有较高的

有效性。但由于数据上传失误、GPS 度量误差及 GPS 信号异常等原因,出租车 GPS 大数据中存在一些无效的数据,如果不对其进行清洗,则会对计算行车时间、行车速度等造成严重的影响。本文主要清洗由以下 3 点原因造成的 GPS 无效数据。1)因信号中断等问题而出现 GPS 断点,为使整个 GPS 序列连贯,使用位置坐标为 0 的 GPS 点替代 GPS 断点。对于这些位置坐标为 0 的 GPS 点应采取过滤措施。2)GPS 的度量量和上传问题导致的一些 GPS 点的位置坐标记录错误。因为 GPS 序列是基于时间序列产生的,所以相邻 GPS 点的时间间隔是很小的,只要坐标有所异常就会导致速度明显异常。计算相邻时间 GPS 数据点间的平均速度 V_i :

$$V_i = \frac{\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2}}{T_i - T_{i+1}} \quad (1)$$

其中, x_i 和 y_i 是第 i 个 GPS 点的位置坐标; x_{i+1} 和 y_{i+1} 是第 $i+1$ 个 GPS 点的位置坐标;根据两点的位置坐标可以求出两点间的距离; T_i 和 T_{i+1} 是第 i 个点和第 $i+1$ 个点的当前时间。设定当 V_i 大于 200 km/h 时表示速度异常,需要进行过滤。选定 200 km/h 是因为正常配置的出租车的行驶速度无法超出 200 km/h。3)当 GPS 点的位置坐标记录有误时,两点间的平均速度 V_i 可能在阈值范围内但加速度会存在明显异常。计算相邻平均速度的加速度 a_i :

$$a_i = \frac{v_i - v_{i+1}}{T_i - T_{i+1}} \quad (2)$$

a_i 超过 30 m/s^2 对于出租车的正常配置来说也是不现实的。对于此类致使加速度超出 30 m/s^2 的 GPS 点,本文采用直接删除的方法对数据进行清洗。

还有一类情况是因为司机对道路的选择不够合理或者乘客乘坐同一辆出租车先后到达不同目的地造成旅行的起点与终点之间距离比实际旅程短太多的现象,这会降低起始区域与终点区域间的可达性。本文将过滤掉这一类旅行,对这类旅行的判断将于 3.2 节中介绍。

3.2 对旅行进行分类

对旅行分类的主要目的是方便判断造成区域间可达性差的原因。将旅行分类这一过程分为两步,首先确定出一条完整的旅行,然后对已经确认的旅行按照旅行条件进行分类。GPS 数据本身只是 GPS 序列,以 $G_i(P_i, T_i, S)$ 表示,其中 P_i 为位置坐标, T_i 为当前时间, S 为是否载客的状态标记。区域间可达性的评估是基于可达率而言的,而可达率的获取是从旅行时间这一角度来讲的,因此应先从 GPS 大数据中抽取完整的旅行。一次完整的旅行可以记为 $trip(on, off, time, period, workday, L_{real})$, on 为旅行的起点坐标, off 为旅行的终点坐标, $time$ 为旅行所耗时长, $period$ 是旅行所发生的时段, $workday$ 是旅行是否于工作日发生的标识。

GPS 点中的载客标记在空载时将一直为 0,载客的过程中该状态标记将一直为 1 直至乘客下车该状态才恢复为 0。即如果 G_i 的状态标记 S 为 0,但 G_{i+1} 的状态标记 S 为 1,则表示 G_{i+1} 是开始载客的点,也就是 $trip.on$,再搜索 G_i 的 S 为 1 但 G_{i+1} 的 S 为 0 的点,即卸客的点 $trip.off$ 。由此确定了起始点与终点,从而可以确定一条旅行。在搜索到载客的点 $trip.on$ 之后,将经历若干个 $G_i, S=1$ 且 $G_{i+1}, S=1$ 的 GPS 点 G_i 才能搜索到卸客点 $trip.off$ 。这些 GPS 点记录了整个旅行的实际路程,由这些 GPS 点可计算出旅行的实际距离 L_{real} :

$$L_{real} = \sum_{s=1}^s \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \quad (3)$$

累加条件 $s=1 \rightarrow s=0$ 将可以累加的点控制在由刚载客的 GPS 点开始到刚卸客的 GPS 点为止,以累加相邻 GPS 点的直线距离为旅途的实际距离。 $trip.time$ 表示旅行时间由卸客点的 GPS 所记录的時刻减载客点的 GPS 所记录的時刻。

因不同的旅行条件会严重影响区域间的可达性,接下来将对旅行进行分类。旅行的时间段 $trip.period$ 被分为早晨: 7:30-9:00、中午: 9:00-15:00、晚上: 16:00-19:00、深夜: 19:00-次日 7:30。以旅行的起始点时间为准,将根据起始时间确认每一个旅行的 $trip.period$ 属于上述哪一时间段。又因周末与工作日的旅行分布差异较大,所以将上述 4 类旅行分别分为工作日和非工作日两类: $trip.workday$ 和 $trip.weekend$ 。这样旅行可以被分为 8 类,分别是工作日的早晨、中午、晚上、深夜,非工作日的早晨、中午、晚上、深夜。旅行分类的流程图如图 1 所示。虚线框内的内容为判断是否存在绕道现象,是否应该过滤掉此条旅行。其他内容是从 GPS 点中抽取出完整旅行的步骤,其中 k 是旅行的序号,如果该旅行不需过滤则序号加 1 并搜索下一条旅行;若旅行应被过滤则序号不变,直接搜索下一条旅行并对原旅行信息重新赋值。

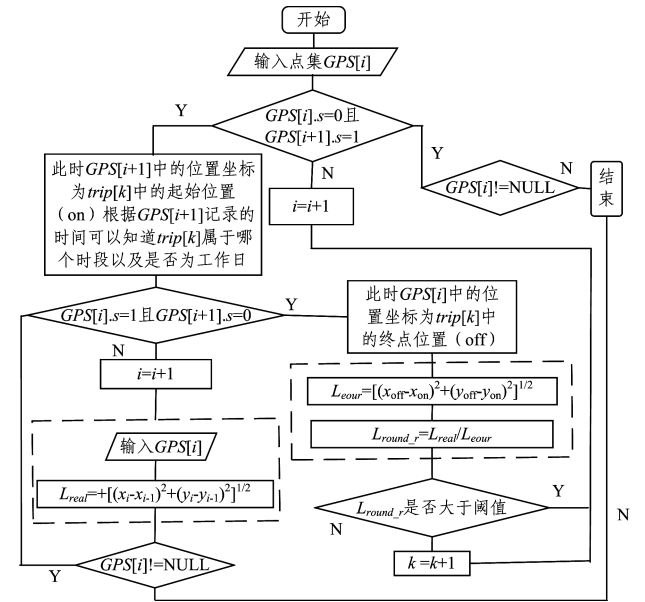


图 1 旅行分类流程图

Fig. 1 Flowchart of travel classification

3.1 节中提到的应被过滤掉的旅行可在每次确定出旅行的起始点后判断。判断方法为:首先计算出旅行的直线距离 L_{our} :

$$L_{our} = \sqrt{(X_{on} - X_{off})^2 + (Y_{on} - Y_{off})^2} \quad (4)$$

其中, $X_{on}, Y_{on}, X_{off}, Y_{off}$ 是一次旅行的起始点与终点的坐标。

旅行的实际距离 (L_{real}) 和直线距离 L_{our} 的比值为 L_{round_r} :

$$L_{round_r} = \frac{L_{real}}{L_{our}} \quad (5)$$

因此, L_{real} 理应比 L_{our} 大,但是相差太多会被认为是存在绕道现象。 L_{round_r} 的值过大则认为该条旅行应被删除。

3.3 确认学习区域

确认学习区域的核心概念就是找到包含热点的区域。一

个城市可以被分成 $m \times n$ 个区域,但是只有少数区域包含热点,如国道、海域、河流、步行街、封闭式小区等是不会包含热点的。

本文使用聚类的方法生成热点,用 3.2 节中挑出的旅行的起点和终点作为点集进行聚类。因热点的聚类与旅行的密度相关,所以本文选用 DBSCAN 方法进行聚类。生成的热点以坐标的形式表现, $m \times n$ 个区域有坐标及序号。学习区域确认的示意图如图 2 所示。 $m \times n$ 的表格代表将区域分为 $m \times n$ 个,并有唯一标号圆点为热点,其中只有 14 号区域包含热点,则 14 号区域为学习区域,应赋予新的标号。每个区域只要包含一个热点就可以被认为是学习区域,并使学习区域有新的序号,新的序号应与旧的标号保持映射,以方便 OD 矩阵的建立。本文将区域右上顶点的坐标定为区域的坐标,这样只需将热点坐标向上取整就可以判断热点属于哪一区域。

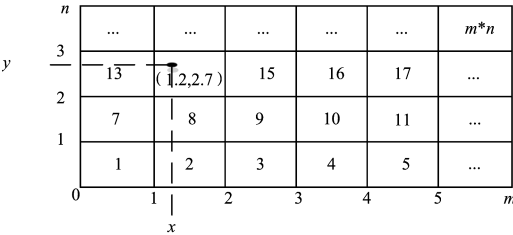


图 2 确认学习区域示意图

Fig. 2 Schematic diagram of confirming learning area

3.4 建立 OD 矩阵评估区域间的可达性

大多数 OD 矩阵(origin-destination)是二维的,用于记录区域间旅行的某一个属性信息。本文采用四维 OD 矩阵记录旅行的多维信息。矩阵的第一维和第二维与正常的 OD 矩阵一样,表示目的区域和起始区域;第三维表示旅行条件,由 *trip.period* 和 *trip.workday* 组成,共计 8 种;第四维是旅行信息,包括旅行时间 *trip.time*、旅行的起点 *trip.on*、旅行的终点 *trip.off* 以及旅行的实际距离 *trip.L_{real}*。矩阵的结构示意图如图 3 所示。起始区域 R_q 到 R_j 的所有旅行对应出现在矩形框内。方框为旅行条件,方框内所承载的信息为具体的旅行信息。

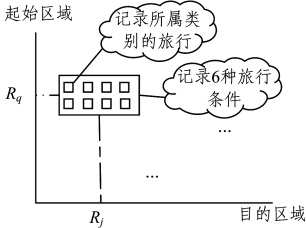


图 3 OD 矩阵维度示意图

Fig. 3 Diagram of OD matrix dimension

产生 OD 矩阵的目的是将抽取出的同一类旅行存储到同一个结构中,以方便计算区域间的可达率。整个 OD 矩阵的生成可以概括为将已知的旅行信息 *trip[k]* 作为输入,最终产生形如 $OD[i][j][][]$ 的矩阵。其中 i 和 j 为学习区域的序号,可由旅行信息 *trip[k].on*, *trip[k].off* 的坐标向上取整得到起止坐标所对应的原区域的标号,再通过原区域的标号与学习区域的序号间的映射关系可确定 *trip[k].on* 和 *trip[k].off* 所属的学习区域,进而确定 i 和 j 的值。第三维

的取值设定为 1~8,分别代表 3.2 节中提出的 8 种旅行条件,最后将 *trip[k]* 中剩余的旅行信息写入 OD 矩阵的最后一维中。构建 OD 矩阵的伪码如下。

```
输入:旅行信息 trip[m]
输出:OD 矩阵 OD[i][j][ ][ ]
Trip OD[i][j][8][ ]; //定义 OD 矩阵
for(i=0;i< Learn_A,i++)// Learn_A 为确定的学习区域的个数
{
    for(m=0;m<length(on);m++)
    {
        if(floor(on[m].x)*regionx+floor(on[m].y)=learn[i])
            //查找由学习区域 i 始发的旅行
        {
            for(n=0;n<Learn_A;n++)
            if(learn[n]=floor(off[m].x)*regionx+floor(off[m].y))
                //查找当前选中旅行的目的区域所对应的学习区域
            j=n;//n 为目的学习区域
        }
        If(type.workday[m]=d and period[m]=iod)
            //d,iod 代表 3 种旅行条件的组合,每一种组合对应一个 TY 值
        OD[i][j][TY][KTY]=trip[m]; //TY 取值为 0~7,将对应的 trip
            记录到对应旅行条件下的 KTY 中
        KTY++;
    }
}
```

下面给出一条 OD 矩阵所记录的数据的实例:

$$\begin{aligned} \overrightarrow{OD}_{103,104} &= (\vec{C}_{103,104,1} \cdots \vec{C}_{103,104,8}) \\ \vec{C}_{103,104,1} &= (\overrightarrow{trip}_1 \cdots \overrightarrow{trip}_n) \\ \overrightarrow{trip}_2 &= (7:53:21, 8:03:12, 4911, 1, 1, 5, 13) \end{aligned}$$

该记录为学习区域 103 和 104 之间旅行条件为工作日早晨时发生的第二条旅行。由时刻 7:53:21 开始至时刻 8:03:12 为止,历时 9min51s,行驶 5.13km。*trip* 的含义可以参照 3.2 节。

在评估区域间可达性时,传统的方法如轮廓测量法以热点数 AC_{qj} 来评估,即:

$$AC_{qj} = \sum_{t=1, \overline{TIME}_{qj} \leq TM_{max}}^8 \quad (6)$$

其中, a_{ij} 为 t 条件下所涉及目的区域 R_j 的热点数,区域间平均旅行时间 TM_{max} 为可容忍的时间范围,一般取整个城市所有旅行的平均时间。轮廓测量法认为在旅行条件为 t 的情况下,若区域 R_q 到 R_j 的平均旅行时间超出容忍范围 TM_{max} ,则认为 t 条件下 R_q 无法达到 R_j 的任何热点,显然 t 条件下 R_q 到 R_j 的可达性不好;对于区域间平均旅行时间小于城市平均旅行时间的情况而言,认为 t 条件下 R_q 可达 R_j 的任何热点,但无法表示可达性好与不好的程度,甚至会对区域间的可达性产生错误的评估。此外,对于某些旅行分布较单一的区域而言,平均旅行时间小于城市平均旅行时间,但是旅行涉及到的热点较少,因此向上累加的 a_{ij} 个数也很少,最终结果就是该区域可达热点少,可达性不好。但事实上,这些区域的可达性是好的,因此对于部分区域,传统方法会产生错误的可达性评估。

本文以大数据的思想为主淡化统计学的方法,利用大量

出租车 GPS 数据进行分析挖掘,提出以区域间可达率来量化区域间可达性。本文所说的可达率是基于旅行信息计算的而非旅行所涉及的热点。因此,由区域间旅行的完成情况来表达区域间的可达性,可以在评估出可达性好坏的基础上,进一步评估出区域间可达性的准确结果,以表示区域间的可达程度。对于部分旅行分布单一的区域,区域间可达率只考虑这些区域所发生的旅行的完成情况,避免了传统的轮廓测量法对这些区域造成的错误估计。本文根据生成的 OD 矩阵逐个排查每两个区域间每一条旅行的耗时(包含在第四维旅行信息中)是否超出城市平均旅行时间 TM_{max} 。以区域间未超出平均旅行时间的旅行数占区域间发生的全部旅行数的比例 PAC_{qj} 作为综合情况下的区域间可达率。此外还要求出各旅行条件下区域间未超出城市平均旅行时间 TM_{max} 的旅行概率,通过生成的 OD 矩阵可以将两个区域间发生的旅行分为 8 类,对区域间的每一类旅行按上述方法求解特定旅行条件下的区域间可达率 PAC_{qjt} 。这样每两个区域间会产生一个综合可达率 PAC_{qj} ,以评估区域间的可达性,还会产生 8 个特定条件下的区域间可达率 PAC_{qjt} ,以辅助探索造成可达性差的原因。该方法充分考虑了各地区的旅行分布特点并以概率表达区域间的可达性和旅行的完成程度。这样可以得到一个较为准确的 PAC_{qjt} 和 PAC_{qj} 来对区域间可达性进行量化。人们可以依靠这两个值来规划自己的出行安排甚至出行路线。

$$PAC_{qjt} = \frac{trip_{jtava}}{count_{ODjt}} \tag{7}$$

$$PAC_{qj} = \frac{trip_{jva}}{count_{ODqj}} \tag{8}$$

式(7)中, $trip_{jtava}$ 为 t 条件下区域 q 和区域 j 间可以完成的旅行个数, $count_{ODjt}$ 表示 t 条件下区域 q 和区域 j 间发生的所有旅行的个数。式(8)中, $trip_{jva}$ 表示区域 q 和区域 j 间可以完成的旅行个数, $count_{ODqj}$ 表示区域 q 和区域 j 间发生的所有旅行的个数。

对道路设计和交通网的决策者而言,利用得到的 PAC_{qj} 和 PAC_{qjt} 可以分析造成区域间可达性差的原因。首先分析 PAC_{qj} 较差的区域,查找对应区域各条件下的 PAC_{qjt} ,如果所有条件下的 PAC_{qjt} 都不理想,应先考虑是否是因为旅途遥远导致的旅行时间过长。可用平均旅行速度 V_{ij} 进行验证,即:

$$V_{ij} = \frac{L_{real_{ij}}}{time_{ij}} \tag{9}$$

若平均速度 V_{ij} 正常,则证明 PAC_{qj} 较小是旅途遥远造成的;如果平均速度较慢则考虑是路况的问题,可从邻域着手研究如何对道路做出合理规划。对旅行条件进行分类并求各条件下区域间可达率 PAC_{qjt} 的目的是,在区域间的综合可达率 PAC_{qj} 不理想且各条件下可达率相差较大时,可分析各条件下的平均旅行速度差异,进而针对不同的旅行条件对道路做出因地制宜的规划。

4 实例学习

4.1 确定学习区域

为验证提出的方法可以有效评估区域间的可达性,本文对 2016 年 3 月—2016 年 5 月沈阳市所有出租车 GPS 行车大数据进行实例研究。

本文将整个沈阳市主城区按 46×46 个区域来划分,可以得到 2116 个区域。由于这些区域中大部分不包括行车路线

或者出租车乘降热点,因此需要从其中选择出包含乘降热点的区域作为可达性评估的学习区域以缩小学习的范围。根据 OD 矩阵中旅行信息的起始坐标与终点坐标,以 DBSCAN^[18-20]方法进行热点聚类,当某一乘降点周围所包含的其他乘降点的个数是否足够多可以使之之前选中的乘降点成为热点,并将该点周围的点也列入热点的候选集。本文以 100m 为搜索半径,包含的最小乘降点的样本数为 15。热点的确定结果如图 4 所示,其中卫星地图上所有的点是聚类出的热点区域。按上述 DBSCAN 方法进行聚类,所有可用的旅行信息数据可聚类出 4369 个热点,而这些热点分布在所有 2116 个区域中的 521 个区域中,因此本文以这 521 个区域作为可达性评估的学习区域。

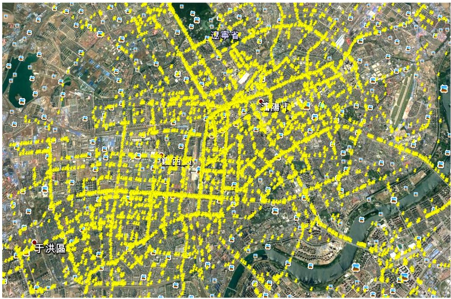


图 4 热点聚类结果
Fig. 4 Hotspot clustering results

4.2 各区域间可达率的计算

本文将对这 521 个学习区域进行排列组合,使其两两一组并计算每一组的两个区域间的可达率,在计算 8 种旅行条件下的区域间可达率时需要将 OD 矩阵遍历一遍,在求区域间综合可达率时又要遍历一遍 OD 矩阵,因此本文首先按照 3.4 节中提到的计算区域间可达率的方法并行计算 8 种旅行条件下的区域可达率和综合条件下的区域可达率,以提高运算效率,其中 TM_{max} 约为 28 min。然后对综合可达率较差的区域进行分析,此时 8 种旅行条件下的可达率将成为分析阶段的重要依据。以一对区域为例,其在 8 种旅行条件下的区域间可达率如图 5 所示,每组柱状图右侧部分代表非工作日下午各时间段的区域间可达率,每组柱状图左侧部分代表工作日下午各时间段的区域间可达率,纵轴为相应旅行条件下的区域间可达率的情况。

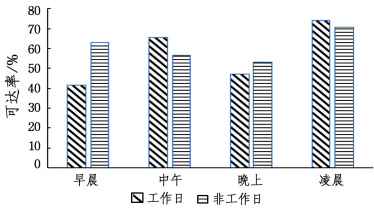


图 5 区域间的可达率
Fig. 5 Inter-regional availability

5 实验与结果分析

本文提出了一种不依赖区域间热点数和平均旅行时间的方法来评估区域间可达性。从旅行的角度出发,计算各区域间发生可达程度较好的旅行的概率,将此概率作为区域间的可达率来评估区域间的可达性。为了验证本文提出的以道路可达率评估道路可达性的结果的准确性,本文将沈阳市所有出

出租车在 2016 年 3 月—2016 年 5 月这 3 个月期间的 GPS 行车数据作为区间可达性评估的训练数据,将 2016 年 6 月 1 日—2016 年 6 月 15 日这 15 日内的 GPS 行车数据作为测试数据。训练数据用于评估各种算法下的区域可达性,测试数据则用于生成可达性评估的真实值,以此对比本文方法、基线法与轮廓测量法的区间可达性评估准确性。基线法与轮廓测量法都是从可达热点个数的角度来评估区域间的可达性,仅以区域间的平均旅行时间为拦截条件决定区域间可达热点的个数。两者的差别在于:基线法在计算区域间的平均旅行时间时,先由 GPS 数据求出区域间发生的每一条旅行的平均速度,再对这些旅行的平均速度取均值作为区域间的平均速度,最后通过区域间的直线距离和区域间的平均速度求得区域间的平均旅行时间。本文算法模型主要由 JAVA 语言实现,并在由 8 台 Sugon I620-G20 服务器(CPU: E5-2630V4 * 2, 内存: 32 GB) 搭建的 Hadoop 分布式环境中进行计算,操作系统为 CentOS7.1。

实验 1 3 种方法判定的可达性好的区域对数与真实可达性好的区域对数的对比如表 1 所列。由 4.1 节可知,共产生了 521 个学习区域,又因为区域间旅行是双向的,所以会产生 521×521 对区域组合。其中可达性被判定为好的区域对数是指由 3 种方法计算出的可达性好的区域组合数;与真实值判定一致的区域组合数是指由 3 种方法计算出的可达性好的区域对中与真实结果相吻合的区域组合的个数。

表 1 不同方法的区域间可达性评估结果

Table 1 Inter-regional accessibility evaluation results for different methods		
方法	可达性被判定为好的区域组合数	与真实值判定一致的区域组合数
真实值	119872	119872
基线法	153581	93765
轮廓测量法	120390	96688
本文方法	122508	109896

由表 1 可以看出,基线法因为低估旅行时间而高估区域间的可达性,所以其评估出的可达性好的区域组合的个数是最大的,但是与实际情况相吻合的组合个数是最少的,因此基线法计算的准确率是最低的。造成轮廓测量法计算结果较小的原因是在各条件下以轮廓测量法计算出的可达热点数的取值不是 0 就是 a_j ,这就意味着 8 个条件中必须有半数以上达到要求才能认为是该区域间的可达性好。显然这是很难的,因此轮廓测量法得到的可达性好的区域组合的个数较少,此外轮廓测量法的评估结果与实际情况相吻合的组合个数相对较少,因此该方法的准确率也较低。无论是基线模型法还是传统的轮廓测量法都十分依赖旅行的平均时间,基线法平均旅行时间 $timeb_{qj}$ 的计算方式为:

$$timeb_{qj} = \frac{LD_{qour}}{V_{qj}} \tag{10}$$

其中, LD_{qour} 为区域间的直线距离:

$$LD_{qour} = \sqrt{(X_q - X_j)^2 + (Y_q - Y_j)^2} \tag{11}$$

V_{qj} 为区域间的平均旅行速度:

$$V_{qj} = \frac{\sum_{n=1}^{OD_{qj}} V_k}{OD_{qj}} \tag{12}$$

$$V_k = \frac{L_{real}}{trip[k].off - trip[k].on}$$

其中, OD_{qj} 代表区域 q 到区域 j 的旅行总次数。然而用直线距离会比用 L_{real} 更容易得出一个较小的平均旅行时间,这会高估区域间的可达性,因此用直线距离求解平均旅行时间是不准确的。并且区域间平均旅行速度 V_{qj} 的计算是对区域间所有旅行的平均速度 V_k 再取平均值得到的,前后需经过 3 次除运算,会将之前的误差不断向后传递,产生不好的结果。因此,基线模型希望依靠平均旅行时间来判断区域可达性,但是却无法产生准确的平均旅行时间。对于轮廓测量法而言,其在平均旅行时间的计算方面是比较准确的,但是在量化方面认为只要平均旅行时间不符合要求就否定整个区域,反之则肯定整个区域。这会夸大区域间真实的可达性,即低估可达性较差的区域,使计算结果变得很低,高估可达性较好的区域,使计算结果变得很理想,这会大大影响可达性评估结果的准确性。

而本文提出的方法评估可达性好的区域组合的个数最接近真实结果,并且与真实情况相吻合的区域组合的个数是最多的,准确率比基线法高出 28.7%,比轮廓测量法高出 9.4%。因此,本文提出的方法的计算准确率较高。产生这一结果的主要原因是:1)3 种方法都比较依赖区域间的平均旅行时间,但是轮廓测量法和基线法都无法准确地计算出旅行时间,因此准确率的下降是难免的。2)区域间可达性的好坏直观来讲就是区域间的通信是否顺畅,而旅行时间是最直观地反映这点的。而轮廓测量法是以可达热点数来量化区域间可达性的,但轮廓测量法在确定可达热点个数时不够准确,因此会使准确率下降。

实验 2 本文模型对区域间可达性评估准确性较高的主要原因是模型采用的量化方式较为科学。在实验中发现,在 521 个学习区域中,在每对学习区域的可达率的计算过程中,本文算法有 15.6%~27.5% 的准确率提升。轮廓测量法和本文方法计算出的某一区域到其他区域间道路可达率与真实可达率的对比如图 6 所示。由于量化方式的不同,为了对比实验结果,将轮廓测量法得出的可达热点数转化为其与所有涉及到的热点数的比率来进行比较。

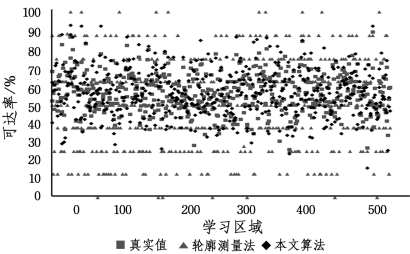


图 6 可达率计算结果的对比

Fig. 6 Comparison of reachability calculation results

图 6 中,横轴为 521 个学习区域,纵轴为区域间可达率,其中矩形点为真实的值,菱形点为本文方法所评估出的值,三角形点为经转化后的轮廓测量法的值。很容易看出,矩形点与菱形点的吻合度较高。三角形点也就是轮廓测量法只产生 9 种比率即 0,0.125,0.25,0.375,0.5,0.625,0.75,0.875,1。这是由轮廓测量法不严谨的量化导致的,而且明显可以看出,当真实的可达率越高或越低时,轮廓测量法的结果越趋于两端的极值 0 或 1。

结束语 本文针对传统方法评估区域间可达性准确率不

高这一问题做出两点改进。首先,传统方法在量化方式上均以可达热点数作为评估标准,这忽略了区域间旅行的方向性即旅行目的地分布不均的问题,这会低估某些区域间的可达性。本文提出可达率这一概念充分考虑了区域间旅行的方向性,纠正被低估的区域间的可达性。此外,传统方法在计算区域间可达热点的个数时,认为没有一个热点可达或全部热点都可达,这会使评估结果向两极发散。本文利用大数据的思想,基于出租车 GPS 行车数据构建了区域间可达率计算模型,并以区域间可达率来评估可达性,对区域间车辆旅行信息进行挖掘,减少了传递误差,评估结果也相对准确。为了验证本文提出模型的准确性,分别设计了两组对比实验来对比区域间可达率与可达性评估的准确性。实验结果表明,本文方法计算的区域间可达率较其他传统方法有 15.6%~27.5% 的提高,同时区域间可达性判定的准确性也提高了 9.4%~28.7%。

由于实际影响区域间可达性的因素有很多,在今后的研究中可以加入其他因素的相关数据使评估结果更接近真实情况。此外,本文提出的可达率这一量化方式会对区域间可达性的评估产生积极影响,但也可能不够准确。接下来将着手研究如何减少产生可达率这一过程中的误差传递及细化可达率的计算方式。

参 考 文 献

[1] CHENG Y, LIU L, REN J L, et al. Study on Traffic Satisfaction and Economic Development Level Measurement and Spatial Pattern of Jinan Metropolitan Area[J]. Chinese Economic geography, 2013, 33(3): 59-64. (in Chinese)
程钰, 刘雷, 任建兰, 等. 济南都市圈交通可达性与经济发展水平测度及空间格局研究[J]. 经济地理, 2013, 33(3): 59-64.

[2] CUI J X, LIU F, JANSSENS D, et al. Detecting urban road network accessibility problems using taxi GPS data[J]. Journal of Transport Geography, 2016, 51(C): 147-157.

[3] WANG X W, XI Y T, TAO J Q, et al. Study on the Accessibility of Traffic Road Network in the Main Urban Area of Xuzhou City Based on[J]. Chinese Shanxi Architecture, 2016, 42(14): 13-15. (in Chinese)
王晓薇, 奚砚涛, 陶季奇, 等. 基于 GIS 的徐州市主城区道路网络可达性研究[J]. 山西建筑, 2016, 42(14): 13-15.

[4] LI S K. Study on Road Route Selection and Urban Traffic Network Evaluation Based on GIS[D]. Chongqing: Chongqing University, 2005. (in Chinese)
李石科. 基于 GIS 的道路选线及城市交通路网评价研究[D]. 重庆: 重庆大学, 2005.

[5] KONING J G. Indicators of urban accessibility: Theory and application[J]. Transportation, 1980, 9(2): 145-172.

[6] MORRIS J M, DUMBLE P L, WIGAN M R. Accessibility indication for transport planning[J]. Transportation Research A, 1978, 13: 19-109.

[7] LANGFORD M, HIGGS. Accessibility and public service provision: evaluating the impacts of the Post Office Network Change Programme in the UK [J]. Transactions of the Institute of British Geographers, 2010, 35(4): 585-601.

[8] NOVAK D C, SULLIVAN J L. A link-focused methodology for evaluating accessibility to emergency services. Decis[J]. Decision Support Systems, 2014, 57: 309-319.

[9] ANDERSON P, LEVINSON D, PARTHASARATHI P. Accessibility futures[J]. Transactions in GIS, 2013, 17(5): 683-705.

[10] CURL A, NELSON J D, ANABLE. Does accessibility planning address what matters? a review of current practice and practitioner perspectives[J]. Research in Transportation Business & Management, 2011, 2: 3-11.

[11] PÁEZ A, SCOTT D M, MORENCY C. Measuring accessibility: positive and normative implementations of various accessibility indicators[J]. Journal of Transport Geography, 2012, 25: 141-153.

[12] HUA S Y, BAO D W, JIA J H. Research on Measurement Method of Accessibility of Airport Network Based on Impedance Function[J]. Journal of Wuhan University of Technology, 2016, 40(5): 885-890. (in Chinese)
华松逸, 包丹文, 贾俊华. 基于阻抗函数的机场集疏运道路网可达性测度方法研究[J]. 武汉理工大学学报, 2016, 40(5): 885-890.

[13] LI L, WANG Z H, LI B C, et al. A Study on the Spatial Accessibility of Star Hotels in Shanghai and Its Driving Forces[J]. Journal of Hainan Normal University, 2016, 29(4): 425-434. (in Chinese)
李龙, 王朝辉, 李保超, 等. 上海市星级酒店空间可达性及其驱动力研究[J]. 海南师范大学学报, 2016, 29(4): 425-434.

[14] LU Y, LI S. An empirical study of with-in day OD prediction using taxi GPS data in Singapore[C] // Transportation Research Board 93rd Annual Meeting, 2004.

[15] GÜHNEMANN A, SCHÄFER R P, THIESENHUSEN K U. Monitoring traffic and emissions by floating car data[J]. Institute of Transport Studies Working Paper, 2004, 3: 4-7.

[16] MUSTARY N R, CHANDER R P, BAIG. A performance evaluation of VANET for intelligent transportation system[J]. World Journal of Science and Technology, 2012, 2(10): 89-93.

[17] ZHANG H, WANG X M, GUO X C, et al. Taxi GPS track large data in intelligent traffic applications[J]. Journal of Lanzhou University of Technology, 2016, 42(1): 109-114. (in Chinese)
张红, 王晓明, 过秀成, 等. 出租车 GPS 轨迹大数据在智能交通中的应用[J]. 兰州理工大学学报, 2016, 42(1): 109-114.

[18] DING G H, XU Y N, GUO J H. Multi-pattern Matching Based on DBSCAN Clustering Algorithm[J]. Chinese Computer applications and software, 2016(2): 25-29. (in Chinese)
丁国辉, 许莹南, 郭军宏. 基于 DBSCAN 聚类算法的多模式匹配[J]. 计算机应用与软件, 2016(2): 25-29.

[19] RONG Q S, YAN J B, GUO G Q. Research and Implementation of DBSCAN Clustering Algorithm[J]. Chinese Computer Application, 2004, 24(4): 45-46. (in Chinese)
荣秋生, 颜君彪, 郭国强. 基于 DBSCAN 聚类算法的研究与实现[J]. 计算机应用, 2004, 24(4): 45-46.

[20] SCHUBERT E, SANDER J, ESTER M, et al. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN [J]. ACM Transactions on Database Systems (TODS), 2017, 42(3): 19.