基于 OD 数据的群体行为可视分析

黄文达, 陶煜波*, 屈 珂, 林 海

(浙江大学 CAD&CG 国家重点实验室 杭州 310058)
(taoyubo@cad.zju.edu.cn)

摘 要: 公共自行车服务系统记录了人们日常出行的移动数据, 研究这些数据有助于了解群体行为的轨迹和模式. 已有的基于公共自行车数据的群体行为研究只分析了站点之间的群体流动, 没有提供更高层次的群体行为分析. 本文通过群体定义从自行车数据中提取群体行为, 并设计了一个多视图合作的可视分析系统, 支持从区域深入到站点的群体行为交互分析. 日历图展示群体行为随时间的变化. 群体行为分布地图支持在区域或站点层面对群体行为进行分析, 并配以时间和流量筛选实现更细粒度的探索分析. 堆叠时钟图对比群体行为在工作日和周末的模式. 最后, 通过四个案例来展示群体行为的时空, 轨迹模式, 证明本系统的有用性和有效性.

关键词: OD 数据; 群体行为; 区域划分; 可视分析; 时空数据
中图法分类号: TP391.41

Visual Analysis of Group Behavior Based on Origin-Destination Data

Huang Wenda, Tao Yubo*, Qu Ke, and Lin Hai
(State Key Laboratory of CAD&CG, ZheJiang University, Hangzhou 310058*)*

Abstract: The public bike service system records the routine movement data produced by people, and research on these data is beneficial to perceive the trajectory and pattern of group behavior Existing research based on public bike data which study group behavior only analyze group flows between stations, without providing analysis of group behavior at a higher level. In this paper, group behavior is extracted from bike data according to the definition of group. A visual analysis system with multiple coordinated views is designed as well, which supports interactive analysis of group behavior in a regions-to-stations way. Calendar view shows the temporal variation of group behavior. Group behavior distribution map is used to analyze group behavior at station level or at region level, with time filter and flow filter to achieve finer granularity in exploration and analysis. Stacked clock view compares patterns of group behavior in weekend and those on weekday. Finally, four case studies are used to show the spatial, temporal and trajectory pattern of group behavior, which prove the usefulness and effectiveness of our system.

Key words: OD data; Group behavior; Region Segmentation; Visual analysis; Spatial-temporal data

轨迹数据在社会科学、生物、城市交通等诸多领域都已经存在大量的工作和研究. 在城市交通领域, 研究轨迹数据具有十分重要的意义和价值, 比如可以研究城市道路的交通状况来识别拥堵区

收稿日期: 2017-6-30; 修回日期: 20**-**-**. **基金项目**: **国家自然科学基金**(61472354, 61672452),NSFC-**广东省人民政府大数据科学研究中心项目**(U1611263). 黄文达(1995—), 男, 硕士, 主要研究方向为信息可视化; 陶煜波(1980—), 男, 博士, 副教授, CCF 会员, 论文通讯作者, 主要研究方向为数据可视化、可视分析; 屈 珂(1994—), 女, 硕士, 主要研究方向为信息可视化; 林 海(1965—), 男, 博士, 教授, 博士生导师, 主要研究方向为数据可视化、可视分析、电磁计算.

域和地段，根据用户推荐个性化路线，挖掘人们坐地铁的主要换乘模式，识别套牌车等等. 轨迹数据存在关联关系，并非孤立产生，比如由结伴而行所产生的一组轨迹，这种行为被称之为群体行为. 蕴藏在公共自行车数据中的群体行为通常都表现出其独有的时空特征. 比如结伴骑车出行的人们都会更倾向于在周末，午后和晚上出行，且选择的地点都是类似公园，广场等人群密集的地方. 研究群体行为有助于发现群体的移动规律以及分析群体形成的原因. 比如，研究发现女性相比男性更倾向于群体骑行是出于一种自我保护的意识[1]，因为女性单独骑行存在更大的风险. 也有研究发现群体骑行是引导那些童年以后就不再骑自行车的人重新骑车的重要原因之一[2].

公共自行车出发点-目的地(origin-destination)数据记录了借车站点，还车站点等属性，提供了借车还车的站点位置以及时间信息. 然而，由于自行车站点数目较多，倘若需要从宏观上一一研究两个站点之间的流量关系，那么 n 个站点就会产生 n^2 数量级的关联关系，这将会给可视化带来了难度不小的挑战. 同时，考虑到人们借车还车的习惯：倘若在站点 A 借不到车，那么很有可能会从 A 相邻的站点借车(还车亦然)，以及考虑到单个站点的流入流出量受天气，活动事件等不确定因素的影响具有很大的不稳定性，即人们并不会固定的选择某一站点进行有规律的骑行.

本文选择对城市进行区域划分，对一片区域进行观察和分析. 具体地，本文选择采用一个迭代双聚类算法对站点进行聚类，将群体功能相似且位置相邻的站点聚为一类，以便更有效地进行数据分析以及更好的可视化展示.

本文先从自行车数据中提取出需要研究的群体行为，并以此为依据，利用群体流量对自行车站点进行聚类从而实现区域划分，进而探索群体行为的时空分布，区域与区域之间的群体流量关系，再深入到站点与站点之间的群体流量关系. 最后设计并实现了一个包含日历图，群体行为分布地图，流量散点图，堆叠时钟图，和年龄直方图的可视分析交互系统.

本文的主要贡献如下：

- 基于群体流量对自行车站点进行聚类，旨在将群体功能相似的站点聚成一个类，从而得到划分后群体功能相似的区域.
- 设计了一个支持区域模式，和站点模式的群体行为分布地图，支持从区域深入到站点的自顶向下的可视分析流程.
- 实现了一个多视图合作，协同高效且交互功能丰富的可视分析系统，实现群体行为挖掘和语义分析.

1 相关工作

与本文工作相关的主要工作有：轨迹数据的可视分析，OD 数据可视化，和空间区域的划分.

1.1 轨迹数据的可视分析

轨迹数据[3][4]可以分为个体移动轨迹数据和群体移动轨迹数据. 目前国内外对群体移动行为的可视分析研究工作相对较少，一般都关注于个体移动行为的分析，并针对此设计了诸多应用性的可视分析系统. Shamal 等[5]设计的 SematicTraj 系统用一种直观高效，语义丰富的手段来管理和检索出租车的移动轨迹；Wang 等[6]利用出租车 GPS 轨迹对城市交通拥堵以及拥堵传播模式进行可视分析；Zeng 等[7]针对公共交通系统研究了人们在不同时刻从某一站点出发到达另一站点所需的时间问题；

对于群体移动轨迹具体的可视分析，Andrienko 等[8]结合动画以及用户自定义时间窗来动态展示物体移动的轨迹，然而动画在展示多个物体的移动时会给用户带来认知负担和记忆负担，且对轨迹的直接可视化只能胜任一些定义明确，操作简单，意义单薄的可视分析任务. Hoang 等[9]基于区域流量来预测人群流动从而预防大规模的踩踏事件. 与本文工作联系较为密切的是 Beecham 等的工作. 他们针对伦敦公共自行车数据，从年龄，性别等方面研究了对群体租车这一行为的影响因素和决定因素[10]. 特别地，他们基于站点流量将群体行为的租车轨迹进一步分类，得到若干种类型的群体租车行为模式，进而对这几种模式进行讨论和对比[11]. 虽然已有工作确实得到了不少有价值的结论，但可视分析方法要么是基于站点，要么是基于区域，没有将二者有机地结合起来，且缺乏一个针对群体行为进行高效分析的交互式可视化系统. 因此本文结合出发点-目的地聚集的可视化和面向群体行为的特征可视化来支持更加系统全面的可视分析任务，设计并实现一个界面友好，高效凝练，多视图合作的可视化系统，将群体移动的

规律从区域到站点的层面直观地展现出来.

1.2 OD 数据可视化

OD 数据属于轨迹数据中的一种, 是由起点终点, 起止时间以及一些其他附加属性所构成的轨迹数据. 例如卡口数据, 公共自行车数据, 人口迁移数据等都属于此类数据. OD 数据的特点是只提供了起点和终点的位置, 但不记录具体的轨迹路径. 因此比较适合回答诸如"从站点 A 到站点 B 平均每天有多少车辆经过"等问题. Jiang 等[12]对出租车的 OD 数据进行了可视分析. 对于 OD 数据的可视化, 可视化研究者们已经提出了不少方法. 其中主要分为 3 大类: 流图, OD 矩阵和 OD 图.

流图是一种将起点和终点用直线或曲线连接起来, 并用线的宽度来编码流量大小的可视化方法. 毫无疑问, 这种方法简单明了且通俗易懂. Wang 等[13]在对稀疏轨迹数据进行分析的时候,用流图来比较不同链路之间流量的大小, 以及链路流量大小与该链路相邻站点的交通状态的联系.但是流图只适合于单个起点的展示, 对于多个起点, 大量交叉和错综复杂的线条将引起严重的相互遮挡. 针对这一问题, 有不少研究者们提出了各种不同的边捆绑以及边过滤技术.

OD 矩阵是一个 m 行 n 列, 并用其中的小方格的颜色来编码流量大小的可视化方法. 这种方法相对于流图可扩展性更高. 但因为其丢失了地理位置信息, 往往需要和另外的地图结合起来才能发现出和地理位置相关的规律.

OD 图[14]是指将整个地图分割成大小一样的方格, 然后再在每个小方格里面嵌套一个小的经过分割的整体地图的一种可视化方法. 这样一来,方格 A 的小地图里面对应的方格 B 的颜色编码的就是从 A 到 B 的流量. Yang 等[15]设计了 OD 图的另一个变种 MapTrix 并在文中用定量的方法比较 Maptrix 和 OD 图的优劣. 该方法主要是用连线将一个存放起点的地图, 一个存放终点的地图和一个 OD 矩阵图用直线一一对应连接起来, 从而充分发挥了了 OD 矩阵的简洁性, 又保留了地理空间位置信息.

OD 数据除了包含空间属性以外, 往往还带有时间属性. 如何将这两种属性编码在一起一直给可视化研究者们带来很大的挑战. Boyandin 等对流图中用动画或者 small multiples 编码时间属性的差异做了定量对比[16]. 他同时也提出了一个较为新颖的可视化图表 Flowstrates[17]来编码时间属性.在此基础上, Zeng 等[18]研究了经过某一路径点的所有 OD 轨迹的流量在一段时间内的变化.

本文综合考虑群体行为分析的主要需求, 从直观和易用的角度出发, 采用流图作为主要方法来编码 OD 数据之间的流量关系.

1.3 区域划分

区域划分的实质是按照一定的规则将地图划分成一块块区域, 然后将属于该区域的所有个体的统计信息聚集在一起作为一个代表整体的信息, 典型的例子就是将自行车站点聚类, 将同一个区域内所有站点的流量聚集在一起, 作为该区域的流量.

区域划分的规则依据可以是均匀网格划分, 行政区域划分, 人口密度划分, 功能, 活动划分等. 经过均匀网格划分或行政区域划分后的地图往往可以直接在方格内用颜色来编码统计信息的大小, 形成一个被称之 choroplethmap[19]的地图. 然而上述两种划分都没有考虑到人口密度的分布以及区域的功能性质. 因此不少研究学者希望通过利用机器学习等方法将地图上的个体聚类, 每一个类代表拥有一定功能或性质的区域. Yuan 等[20]利用文本分析中的主题模型, 并用 LDA 算法检测人们乘车上下班中存在的轨迹模式, 最后用这些模式对区域进行划分, 从而得到一个富含语义和价值的划分结果. 在此基础上, Wu 等[21]同样利用主题模型对区域建模, 从区域中提取出事件, 利用 NMF 将事件矩阵分解成一个模式概率矩阵和模式解释矩阵, 最后通过模式概率矩阵聚类, 将模式相同的区域聚成一个更大的区域.

与大多数区域划分算法的功能不同, 本文区域划分的目的是将群体功能相似的站点划分为一个区域, 进而对群体行为的时空分布进行分析.

2 概述

首先介绍本文所用的数据结构, 在此基础上介绍如何从这些数据中提取出群体行为, 然后详细地给出利用提取出的群体进行区域划分的迭代双聚类算法, 接着罗列出可视分析任务, 最后给出整个系统工作流程的概览.

2.1 数据结构

本文采用的是典型的 OD 数据. 每条记录 R 由

起始站点 P_o, 起始时间 T_o, 终止站点 P_d, 终止时间 T_d, 统计属性 A 等主要字段构成. 如下所示:

$$R =< P_o, T_o, P_d, T_d, A >$$

除此以外, 还包括各个站点的数据, 即每个站点的经纬度和名称.

2.2 群体行为的定义

群体行为的构成需要满足三个条件: 1. 群体中的成员从同一站点出发. 2. 群体中的成员到达了同一个站点. 3. 群体中所有成员的出发时间, 到达时间均小于某一阈值. 对于数据库中存储的任意两条记录 R_i、R_j, 即它们要满足如下关系:

1. $R_i.P_o = R_j.P_o$
2. $R_i.P_d = R_j.P_d$
3. $|R_i.T_o - R_j.T_o| < \Delta t, |R_i.T_d - R_j.T_d| < \Delta t$

对于获得的所有群体 G, 规定群体 G 的一些属性:

1. G 的大小为 G 包含的记录的数目
2. G 的起始时间为 G 包含的所有记录中起始时间的最小值
3. G 的终止时间为 G 包含的所有记录中终止时间的最大值
4. G 的起始站点为 G 中任意一条记录的起始站点
5. G 的终止站点为 G 中任意一条记录的终止站点

值得注意的是, 通过这种方法提取出来的群体之间的成员可能并不互相认识, 也不一定是一起约好租车. 他们只是恰好满足群体出行这个特征.

2.3 站点聚类

本文希望选取一个能同时对两个或以上的加权因素进行聚类的算法思路, 从而实现将地理位置相近且群体功能相似的站点聚成一个类. 相比于传统的基于权重的相似度聚合算法, 即定义两个站点的相似度为:

$$sim(i, j) = \alpha * sim_1(i, j) + (1-\alpha) * sim_2(i, j)$$

其中 α 的取值为 0~1, 用来控制两个因素的占比权重,然后将相似度大的站点聚合成一个类, Li 等[22]提出的迭代双聚类在本文中有更好的应用场景, 其主要有优势体现在如下方面:

1. 无需先验知识,即权重 α 的设置
2. 不需要考虑相似度定义的选取, 即应该选取余弦距离相似度还是欧式距离相似度等.
3. 该算法在两个因素互为相关联时, 比如本文中功能因素是基于地理因素的聚类结果计算, 而功能因素的聚类结果又需要地理因素去完善, 这种通过因素之间不断迭代完善结果的过程相比于一次性将两个因素加权求值的聚合算法能获得更好的聚类效果.
4. 能够灵活的控制迭代次数来选取最优的聚类结果.

然而, 由于迭代双聚类算法涉及的参数较多, 会给聚类结果的调试带来一定的困难.

综上, 本文最终采用 Li 等的聚类算法并进行一定的调整和改善. 具体地, 在数据上, 鉴于群体行为是主要的研究对象, 本文使用了群体流量应用在该算法中. 在方法上, 除了考虑流出量外, 同时增加流入量使得功能性约束更强. 在结果上,群体功能相似且位置相近的站点被聚成了一个类, 比如公园外围的一周被提取成了单独的一个类, 这是 Li 等的聚类结果中所没有展现出的有用信息.

具体算法步骤如下:

输入: 位置聚类的类个数 k_1, 流量矩阵聚类的类个数 k_2, 迭代次数 m, 群体记录 G, 站点经纬度 (lng,lat), 站点数目 n

输出: 每个站点所属的类的 ID

Step1. 先利用站点位置, 即站点的经纬度将站点聚成 k_1 个类, 这是依据地理位置相近.

Step2. 为了达到群体功能相似, 本文选择将站点的群体流出量作为依据, 利用刚刚得到的 k_1 个类, 并规定 4 个时间段: 早高峰[7,11), 白天[11,16), 晚高峰[16,21), 夜晚[21,7), 那么可以得到由每个站点在这 4 个时间段分别到 k_1 个类的流出量以及 k_1 个类到每个站点

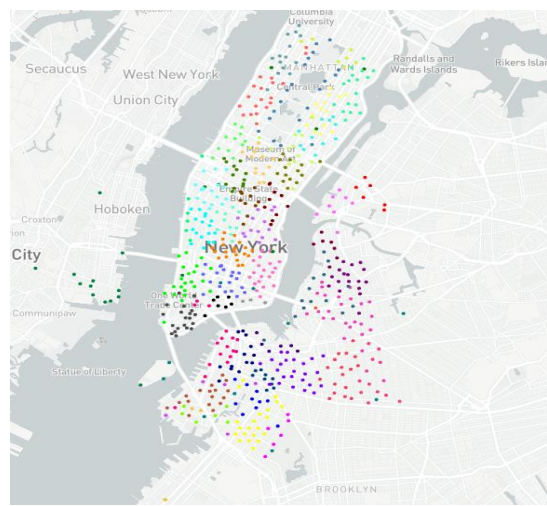
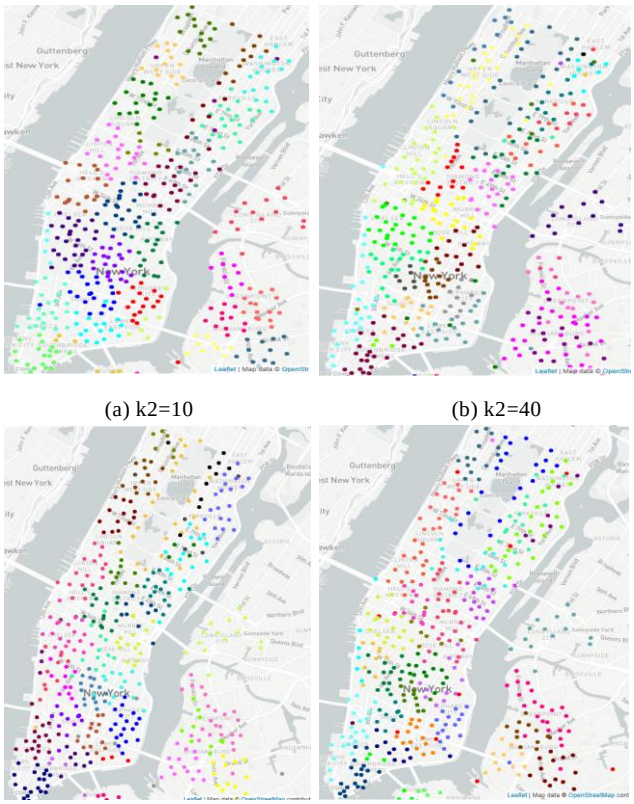


图 1 纽约市自行车站点聚类结果

的流入量,将这些数据排列成一个个大小为 8 x k_1 的矩阵，根据这些矩阵将站点聚成 k_2 个类.

Step3. 对于 k_2 个类里面每一个类 C_i，假定该类共包含 N_i 个站点，对这些站点按照地理位置聚成 $[\![N_i \times k_1 / n]\!]$ 个类，其中 $[\![]\!]$ 可以是向下取整也可以是向上取整，使得 $\sum_{i=1}^{k_2} N_i \times k_1 / n = k_1$. 再回到 Step1 迭代，直到 Step3 生成的 k_1 个类和 Step1 生成的 k_1 个类不再发生改变而收敛，或迭代至设定的迭代次数 m.



(a) k2=10　　(b) k2=40

(c) k2=23　　(d) k2=27

图 2 当 k_1=50, m=50, k_2 取值为 10, 40, 23, 27 时纽约市自行车站点聚类结果

本文使用 *Kmeans* 算法进行上述两次聚类，在本文中，令 k_1=50, k_2=25, m=50，聚类结果如图 1 所示.

k_1 是最终生成的类的个数，用户可结合站点分布的面积大小以及先验知识来决定最终希望看到多少个类. 鉴于群体行为是主要的研究对象，并且聚类的两个依据中，功能相似这一聚类依据所占的比重不能太小，k_2 建议设置为 k_1 的一半左右，图 2 展示了 k_1=50, m=50, k_2=10, 40, 23, 27 的聚类结果. 若 k_2 设置得太小会因为功能因素的权重不够导致类的功能性不明显，如 k_2=10 的聚类结果，可以看到公园并没有聚成一圈，而是被分成了若干个类. 倘若设置得太大，则会因地理因素权重不够导致类和类之间产生过度的重叠和交叉，如 k_2=40 的聚类结果. 观察 k_2=23, 25, 27 的聚类结果，没有出现较为剧烈的震荡.

聚类后，每个类由若干个站点组成，对应一个具有特定群体功能性质的区域. 通过在站点中聚合记录数或在区域中聚合站点记录数，本文规定区域/站点中的一些统计属性:

1. 流入量: 从其他区域/站点流入到当前区域/站点中的流量

2. 流出量: 从当前区域/站点流出到其他区域/站点的流量

3. 自流量: 从当前区域/站点流入自身的流量

4. 总流量: 当前区域/站点的流入量，流出量，自流量的总和

2.4 可视任务分析

在已有的对群体行为的可视分析研究中，Beecham 等[11]通过对群体行为的时空分布进行全局观察(T1)，从宏观上回答了“where are group-cycling journeys, when are they made”等问题，并基于站点之间群体流动的时空属性来挖掘不同的群体行为模式. 通过分析相关文献[1][2][23]的任务和对群体行为分析的调研，本文提出了从区域深入到站点的群体行为可视分析流程，支持单点观察(T2)，两点观察(T3)，模式对比(T4).

T1. 全局观察

全局观察提供了群体行为的时间，空间，统计信息等属性的一个整体概览. 能初步回答诸如“哪片区域群体行为较多”，“群体行为在哪几天比较活跃”等宏观，抽象的问题.

T2. 单点观察

单点观察能够基于区域或站点进行更加细致入微的观察. 它能帮助用户分析感兴趣的区域或站点在哪个时间段群体行为比较活跃，与之流量关联较强的又分别是哪些区域或站点.

T3. 两点观察

两点观察可以基于两个特定的区域或站点进行分析，对比. 特别是它能对两个区域或站点之间流量流动的因果关系，逻辑关系进行挖掘.

T4. 模式对比

主要通过研究和分析区域或站点在工作日和非工作日两种模式下的特征，对这些特征进行类比和对比，从而发现群体行为的规律.

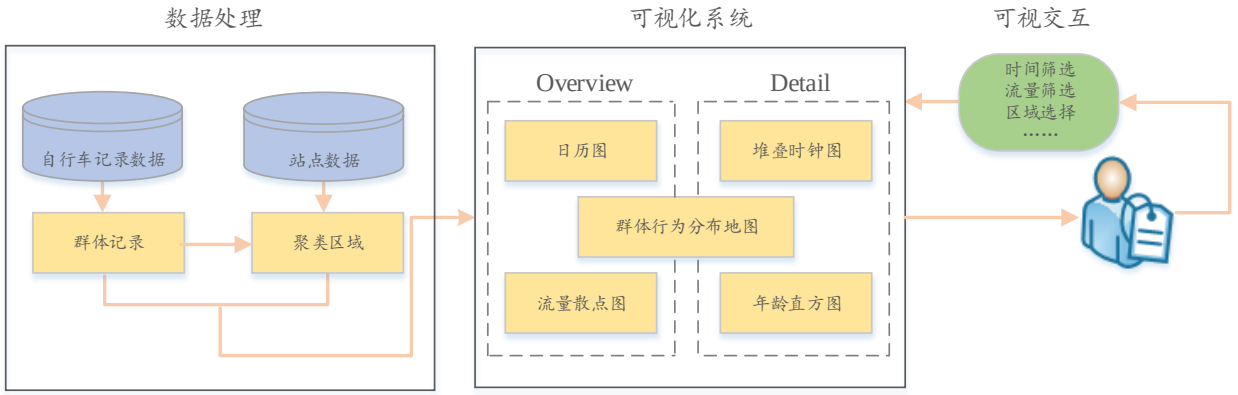


图 3 系统工作流程概览. 在数据处理阶段, 先从自行车记录数据中提取出群体记录, 利用群体记录和站点数据对站点聚类得到群体功能相似的区域, 接着将这些聚类结果和群体记录通过可视化系统展现出来. 可视化系统主要包含用于静态全局探索的 overview 和动态局部探索的 detail, 用户可通过丰富的交互手段参与人机交互的过程

2.5 系统工作流程

图 3 展示了整个系统的工作流程. 本系统主要包含数据处理和可视化系统两部分. 其中在数据处理阶段, 主要完成群体记录的获取以及站点的聚类. 在可视化系统中, 视图根据其作用被归类为概览(overview)图和细节(detail)图. 概览图中的视图用于全局观察且数据不会发生改变, 而细节图中的视图用于局部观察且数据会随用户的交互做出相应的改变. 最后, 用户可以通过一系列操作实现交互式地探索自行车记录中的群体行为. 整个系统充分用了机器处理大数据的计算能力和强大的绘图能力, 又利用了人在决策和分析过程的判断能力和推理能力.

3 可视化设计与交互

根据数据和任务驱动, 本文设计了一个多视图合作的可视化交互系统, 如图 4 所示. 其中群体行为分布地图, 日历图, 流量散点图提供了对群体行为属性在全局上的观察. 群体行为分布地图同时也提供了对区域和站点的探索. 堆叠时钟图能较好地支持工作日和周末两种模式的分析. 年龄直方图展示群体的年龄分布. 本系统的设计严格遵守了“ Overview First, Zoom and Filter, Details on Demand”[25] 的设计准则, 使用户能够从宏观到微观对群体行为有一个从面到点的探索过程.

3.1 可视化设计

3.1.1 日历图

日历图对群体行为在时间维度上的分布给出了一个总体直观的概览, 如图 4(a)所示. 日历图以日历为隐喻来表示时间, 通俗易懂, 每一个小方格代表一天, 每块方格的颜色编码一个类别或顺序属性.具体地来说, 本文采用由绿色到红色, 黄色为过渡的一个渐变色带来编码群体行为的记录数目. 颜色越红表示当天群体记录越多, 越绿表示当天群体记录越少. 从图中, 可以清楚地看到群体流量在时间维度上呈现出的一个整体规律: 周末的记录数明显多于工作日的记录数. 这与本文一开始对群体行为特征的猜测一致.

3.1.2 群体行为分布地图

地图具有很强的地理位置信息表达能力, 几乎是所有研究轨迹数据的可视化系统的标配. 本文设计了一个群体行为分布地图.

在已有的研究中, 自行车站点常常被编码为地图上的一个点, 用区域的边界来编码站点所属的区域. 然而由于本文侧重研究群体行为的功能, 在迭代双聚类算法中将群体流出量这一标准设置的权重相对较大, 导致最后的聚类结果不可避免地出现多处区域与区域之间的重叠与交叉的情况, 因此最后决定用区域的中心点来编码标识一个区域, 如图 4(c)群体行为分布地图所示. 具体的来说, 本文用圆来编码区域的中心点, 圆的颜色来编码区分不同的的区域, 圆的半径大小来编码区域的总流量. 同时在地图中嵌入流图来编码流的信息, 流的大小用线的宽度编码, 流的方向由线的方向和颜色编码. 具体的来说, 流的大小和线宽成正比. 两点相连, 规定对该点来说在左侧一方的流为出流, 在右侧一方的则为入流.

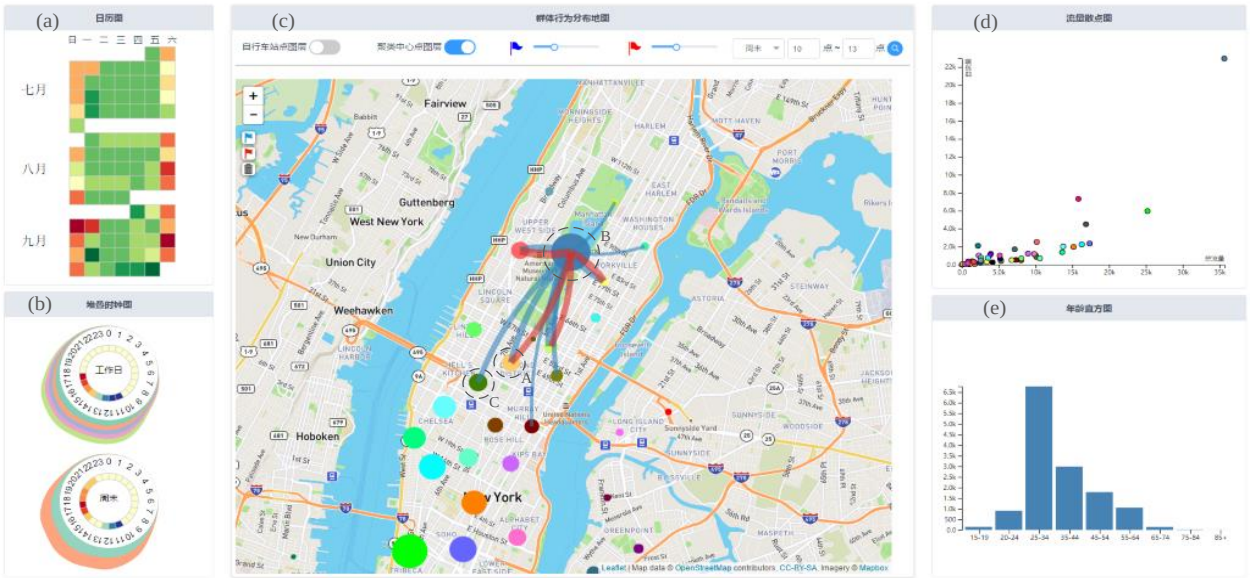


图 4　可视化系统界面总览，图为在群体行为分布地图中点击区域 B，并把时间段设为周末的 10 点-13 点后的状态：(a) 日历展示了 7-9 三个月总体的群体流量随时间的变化. (b) 堆叠时钟图展示了单个区域分别在工作日和周末的群体行为模式. (c) 群体行为分布地图里嵌入了流图，并在上方的菜单栏里配有控制浮层显隐的开关按钮，流量筛选的滑动条以及时间选择的数字输入框等交互控件. (d) 流量散点图用以研究群体总流量和自流量的关系. (e) 年龄直方图可以观察单个区域的群体年龄分布

群体行为分布地图支持两种模式：区域模式和站点模式. 一开始地图处于区域模式，地图上只会显示所有的区域，单击某个区域会显示和此区域关联的流，且蓝色代表入流,红色代表出流. 当通过交互，将蓝色的旗子放在地图中的某个区域将其标识为起点，将红色的旗子放在地图上的某个区域将其标识为终点时，地图会进入站点模式，如图 7 所示，此时地图只会显示选中的两个区域的所有站点，以及这些站点间的流，其中蓝色代表起点的出流，红色代表终点的出流. 值得一提的是，当起点终点都标识为同一个区域时，地图会显示该区域内部站点的所有流，流的方向不再由颜色编码，只能由线的方向决定. 流的颜色均为蓝色.

3.1.3　堆叠时钟图

日历时间图提供了对群体行为在时间上分布的总览，粒度较大，属于 overview 的范畴. 对于时间，仍需要向下钻取到一个更细的层面，同时也需要一个更加直观高效的图来研究单个区域在工作日与周末的群体行为模式.

根据这个原则，本文先是设计了一个圆形的堆叠图，为了描述方便且将其称为外圆,如图 4(b) 所示. 具体的来说,每一圈代表 24 个小时，每一层代表一天，每一层的宽度编码流量的大小. 工作日由内到外分别对应周一至周五，周末则对应周六到周日. 此外，本文还设计了一个圆形的热力图嵌入其中，称为内圆. 具体地来说，每一格对应一个小时，格子的颜色编码流入量与流出量的差. 考虑到颜色的统一，即为了对应群体行为分布地图中蓝色代表入流，红色代表出流，圆形热力图中每一格越蓝代表流入得越多，越红代表流出得越多. 这样一来，外圆的整体形状以及内圆的颜色排列便提供了一种直观的群体行为模式的反应和对比.

3.1.4　流量散点图

为了提供一种辅助手段帮助用户对特殊站点的筛选，以支持 overview+detail 的设计准则，本文认为自流量与总流量的关系是值得关注的一个点. 因此利用如图 4(d) 所示的流量散点图来发现和定位一些离群的异常点，再对这些异常点进行特殊分析.

3.1.5　年龄直方图

除了时空以外的信息，诸如年龄，性别，客户类型，收入状况，居住地等统计信息对于群体行为的研究同样具有价值. 由于数据源的限制，在本文的系统里面只展示了年龄信息. 直方图由于其简洁明了的特性，常常被用来展示数据的分布，因此

本文采用直方图来展示群体年龄分布的规律. 从图 4(e) 可以看到: 年龄分布呈现出“中间多, 两头少”的正态分布的特征, 并且 25-34 岁年龄阶段的人占据大多数.

3.2 交互

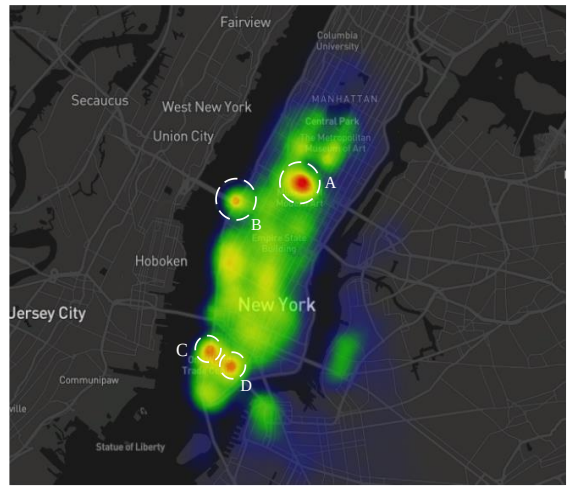
由于目前尚未存在一种可靠有效的可视编码方案能将时空信息, 统计信息, 关系信息紧凑地编码在同一个视图中. 本文采用了多视图联动, 协同合作的方法, 并运用了大量的交互手段来支持多种复杂的可视分析与探索. 以下是本系统支持的一些交互操作:

- 高亮: 在流量散点图中, 当鼠标悬停在代表某个区域的点时, 地图中对应的区域中心点会高亮,其他则变暗, 以形成“Pop out”的效果让用户在群体行为分布地图中的众多区域中心点中迅速、准确的定位.
- 标识: 用户可以将地图左侧中的蓝色或红色旗子放到地图中的区域中心点上来选取起点和终点, 使地图进入站点模式.
- 联动: 当群体行为分布地图中的某个区域中心点被鼠标点击以后, 地图中只会显示该区域关联的入流、出流. 同时, 堆叠时钟图, 年龄直方图会显示对应的区域的信息.
- 时间筛选: 当用户通过地图上方工具栏中的时间控件选择时间以后, 地图中流量的大小会相应地反应在该时间段的状态.
- 图层筛选: 自行车站点图层与区域中心点图层均可通过地图上方的开关按钮控制, 用户可根据需要隐藏或显示图层.
- 流量筛选: 地图上方配有两个滑动条用以辅助用户筛选地图中流量高于某个阈值的流. 从而避免了流太多, 相互遮挡等问题.

4 案例分析

本文采用的是纽约 2016 年 7-9 三个月份的自行车数据, 约为 420 万条记录. 先过滤掉起始站点, 终止站点, 起始时间, 终止时间中存在空缺的记录, 将时间差阈值设为 3 分钟, 从 420 万条自行车记录中提取了约为 32 万条群体记录.

针对 2.4 中提出的四个可视分析任务, 分别展示了四个案例, 对应任务 T1-T4, 来说明本系统的有用性和有效性..



(a)

(b)

图 5 热力图(a)和群体行为分布地图(b)展示群体行为的空间分布

4.1 纽约市群体行为时空分布概览

本案例通过观察纽约市群体行为的总体时空分布,对所研究的数据形成一个初步,宏观的概览.

首先通过设置流量阈值在所有流中筛选出流量较大的流,结果如图 5(b)所示.

在时间维度上,通过观察日历图 4(a)可以清楚地看到群体流量呈现出的一个大致的周期性规律, 即周末的记录数明显多于工作日的记录数. 这与已有研究的结论吻合: 群体行为更倾向于出现在周末.

在空间维度上,通过热力图 5(a)可以看到:四个红色的主要热点 A, B, C, D , 和最上方的“U”型的区域, 以及中部偏黄的区域是群体行为较为密集的区域.结合群体行为分布地图可以发现,面积最大的红色热点 A 和“U”型区域对应的是公园内部大量活跃的群体行为模式,即图 5(b)中的模式 1; 左

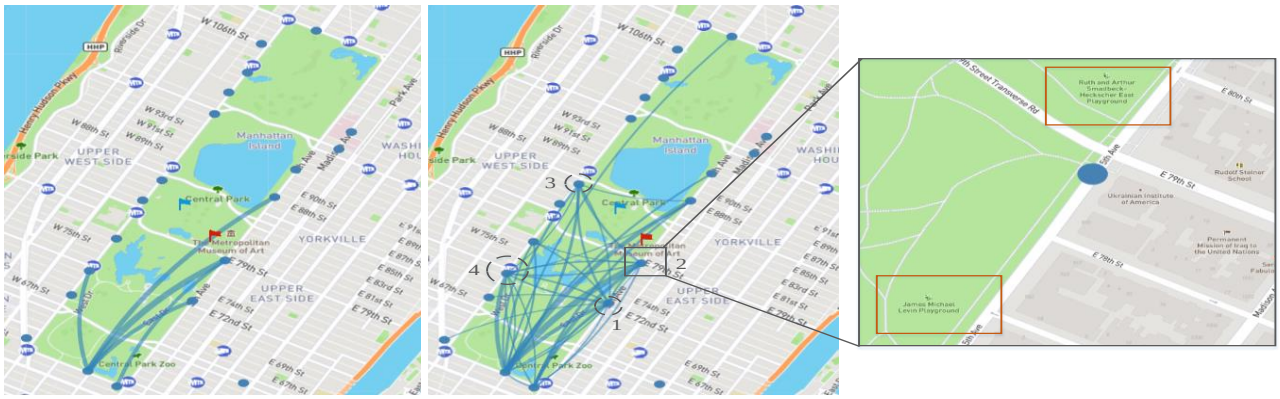


图 6 将起点终点都设置为代表公园的区域，并选择时间段为周末 13 点-15 点，设定较大流量阈值(a) ,以及较小流量阈值(b)后，公园内部站点之间流量的流动情况．放大地图可以发现站点 2 周围存在个游乐场(c)

边的红色热点 B 对应的自行站点位于河隧道附近，从图 5(b)中的模式 2 可以发现与这些起始站点相关联的终点都位于河岸附近,且距离相对较远,因此推测终点位置附近设置了几个轮渡点,人们骑车到该站点后转别的交通工具去对岸.而最下方的两个红色热点 C 和 D 对应的行为模式 3 描述了人们骑自行车通过“Brooklyn”大桥到达对面的“Brroklyn”镇．这些是纽约市群体租车行为的主要三类空间模式．

4.2 公园内部群体移动模式探索

本案例通过研究单个区域的内部流量流动来探索公园内部群体移动模式的特征．具体的探索流程如下:

首先，通过图 5(a)的热力图可以看到地图上方有一个成“U”型的区域对应公园和广场的外围．而另一方面，在图 4(d)中，一个离群的异常点分外显眼．该点具有两个特征：总流量最大，自流量占了总流量的一半以上．通过将鼠标悬浮在该点，在群体行为分布地图中一个分布在地图右上角且半径最大的圆被高亮．通过观察地图，发现该圆对应的正是围绕曼哈顿广场和中心公园外围的“U”型区域．

接着，进一步分析这些占了总流量一半以上的自流量在公园内部的流动模式．当用鼠标在地图中点击该圆时，通过对应的堆叠时钟图,如图 4(b)所示，可以看到周末的内圈分成了蓝-黄-红三个区间段．结合常识和猜测，该颜色排列很有可能对应人们进入公园-在公园里游玩-离开公园的模式．因此将时间锁定在周末 13-15 点这片连续的黄色中，并将起点终点设置为该区域使地图进入站点模式．通过拖动在地图上方的工具栏里的滑动条设定流量阈值，筛选出大于某一阈值的流．图 6(a), (b) 分别对应设置了一个较大阈值和较小阈值后的结果．

从图中，可以发现公园里存在三个比较明显的内部群体移动规律:

1) 从图 6(b)可以看到流量主要集中在下半部分，从而推断出可能是因为公园太大，大多数人骑到一半就放弃继续骑行，选择回头．

2) 结合图 6(a)可以更清楚地发现，公园下方两个入口的最大流出量均是到达站点 2，而非离入口较近的站点 1．通过放大地图，如图 6(c)所示，发现离该站点较近的地方有两个游乐场，进而推断出从公园入口进入的人们有很大一部分选择直奔游乐场游玩．

3) 在公园左侧一排的站点中，这些流的起止点更倾向于集中在公交站附近的站点，如图中的站点 3 和站点 4，这样方便人们直接坐车回家或者去别的地方．

4.3 增设双座位自行车站点

通过对群体租车数据的统计分析，其中有 289302 个群体都是由两人组成，占总群体数量的 87.8%．因此,在交通法规和安全条例允许下,可以考虑增加双人自行车,方便情侣或家人一起骑行游玩.本案例通过研究两个区域的流量关系来寻找双座位自行车站点投放的候选站点．

从案例 2 可以了解到人门进入公园的时间段主要集中在周末的 10 点-13 点．接下来想进一步知道在这个时间段内，哪个区域流入到公园的流量较多，在这两个区域里面又分别是哪些站点之间的流量较多，为什么？

通过将时间选择在周末 10 点-13 点并在地图中点击公园所在的区域的中心点，将筛选流入量和流出量的滑动条的阈值设为大小相等，如图 4(c)所示．发现此时公园的入流从流数目和流量大小上确实都胜过出流，并且注意到了下方有一个区域 A 流进公园区域 B 的流量较多．通过将 A 设置为起点，公园 B 设为终点，使地图进入站点模式，并通过滑动条控件筛选出流量最多的前几条,得到结果如图 7 所示．

在起点的出流中，即图中显示的蓝色线条中，这些 OD 轨迹的起始点大多数集中在站点 1，终止点集中在站点 3．通过地图可以发现站点 1 周围有两个公交站，推断出人们坐公交到此再骑自行车去公园．而站点 3 是公园的两个入口之一．在终点的出流中，即图中显示的红色线条中，这些 OD 轨迹的起始点相对分散，但流量最大的前两条集中在公园的其中一个入口：站点 3．终止点则相对集中在站点 1 和站点 2，通过地图可以发现站点 2 周围也有公交站．根据这些有用的信息，初步判断这三个站点是群体在区域 A 和区域 B 流动的重要枢纽，可以考虑在这三个站点: Broadway & W 49 St, W 52 St & 5 Ave, Central Park S & 6 Ave 分别设置多座位的自行车站点，方便人们集体出行去公园或从公园返回到区域 B．此可视分析方法可应用于任意选中的两个类进行两点观察．

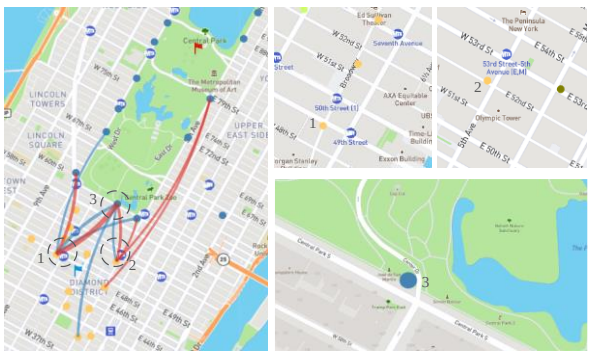
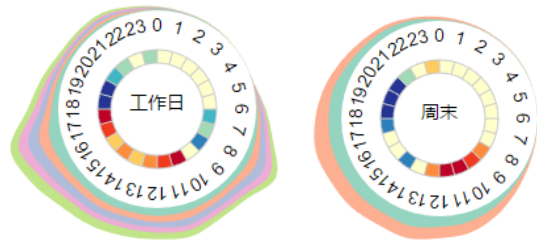


图 7 将起点终点设为区域 A, B 后地图进入站点模式

4.4 工作日和周末群体行为模式对比

本案例研究单个区域的群体行为模式在工作日和周末的差异．

在案例 3 的分析中，当点击了公园正下方的区域 A 时，对应显示的堆叠时钟图，如图 8 所示，从外圈的形状以及内圈的颜色排列上看，工作日的

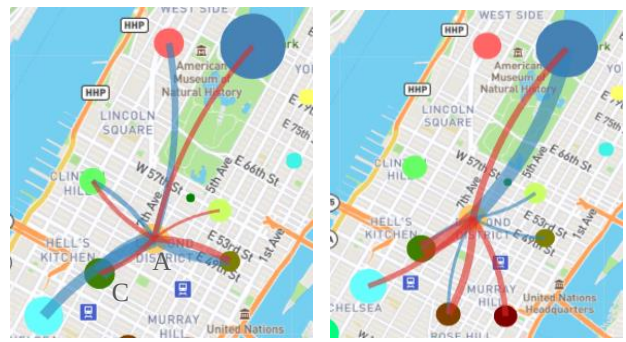


(a) 区域 A 工作日模式 (b) 区域 A 周末模式

图 8 区域 A 工作日和周末的群体行为模式．通过观外圈以及内圈，可以发现存在较明显的差异

模式和非工作日的模式存在明显的差异．具体地来说，工作日的外圈在 6 点-9 点以及 16 点-18 点这两个区间段上有两个凸起，暗示着流量突然增多，周而末却没有．并且，工作日的内圈在 6 点-9 点是蓝色，10 点-13 点是红色，16 点-18 点是红色，18 点-22 点是蓝色，呈现出蓝-红-红-蓝的颜色排列．反观周末，在 8 点-13 点是红色，17 点-22 点是蓝色，呈现出红-蓝的颜色排列．

为了探究这个差异，先在地图中点击区域 A，把时间段锁定在工作日的 6 点-9 点，并设定流量阈值.从图 9(a)中可以发现有一条很大的流量从左下方的一个区域 C 流入，当把时间锁定在工作日的 16 点-18 点时,结果却恰恰相反：图 9(b) 有一条很大的流量流出左下方的区域 C．通过结合地图观察，发现区域 A 里面存在较多的带“building”标识的建筑物，如图 9(c) 所示.这些带有办公性质的大楼都表明该区域有很强的工作区性质，由此便很好地解释了工作日的内圈中的一对蓝红匹配对代表上班下班，即人们从 6 点-9 点集体骑车(也许是在互相不认识的情况下) 去区域 A 上班，然后 16 点-18 点回家．而工作日剩下的那一对红-蓝匹配对和周末的红-蓝匹配对则对应人们在区域 A 骑车去公园和从公园骑回来．值得注意的是,人们在周末 8 点就开始出发去公园了，工作日则会更晚．



(a) (b)

(c)

图 9 设定流量阈值后筛选出在 6 点-9 点(a)和 16 点-18 点(b)与区域 A 关联较强的流. 通过放大地图可以发现区域 A 中存在着大量办公楼(c)

4.5 讨论

基于站点的群体行为分析只关注于局部的群体行为,而基于区域的群体行为分析则只关注于宏观的群体行为.本文结合两者并针对群体行为设计了一个交互式的可视分析系统,支持用户从全局-区域-站点三个层次观察和分析群体行为,通过相关文献的调研抽象出四个任务,最后用四个案例来证明本系统的有用性和有效性.然而,本文针对群体行为的可视分析工作仍有需要改进地方,具体从以下几方面进行考虑:

通过本文对群体的定义所获得的群体尽管在大多数情况下都是正确的,但不排除会有把两个不认识的人标识为同一个群体的可能.可以考虑采用更精确的群体定义或者通过问卷调查等形式获取真实的群体数据,得出更严谨和可靠的分析结论.

对于聚类区域的性质,功能和类型,目前仍需要人参与进行推理和判断.比如用户想要判断某一个类的大致功能,是代表一个大学城,公园还是工作区等？此时,用户不得不观察和分析类里面所包含的建筑物的信息来获得结论.对于这个问题,可以考虑融合城市规划，POI，微博等富有语义信息的数据进行研究，使我们的系统更加智能化和自动化.

5 结 语

本文设计了一个多视图合作，协同交互的可视化系统来支持从区域到站点的自顶向下的可视分析流程来研究群体行为. 首先，从自行车记录中提取出群体行为，基于群体流量和站点的地理位置对站点进行聚类. 在此基础上，研究分析了群体行为的时空分布，区域/站点和区域/站点之间流量流动关系，并进一步挖掘群体移动轨迹模式的特征，发现群体移动轨迹模式在工作日和周末的差异.

本系统具有良好的适用性，可以分析诸如道路卡口数据，地铁打卡数据等多种 OD 数据. 本系统同时也具备良好的扩展性，在数据源支持的情况下，可以扩展支持对性别、居住地、国籍、收入等多种统计属性的分析.本工作还可以扩展到共享单车的群体行为分析中.

参考文献(References):

[1] Aldred R. Cycling cultures: summary of key findings and recommendations[J]. 2012.

[2] Bonham J, Wilson A. Bicycling and the life course: The start-stop-start experiences of women cycling[J]. International Journal of Sustainable Transportation, 2012, 6(4): 195-213.

[3] Wang Zuzhao, Yuan Xiaoru. Visual analysis of trajectory data [J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(1): 9-25(in Chinese)
(王祖超, 袁晓如. 轨迹数据可视分析研究[J]. 计算机辅助设计与图形学学报, 2015, 27(1): 9-25)

[4] Pu Jiansu, Qu Huamin, Ni Lionel. Survey on visualization of trajectory data[J]. Journal of Computer-Aided Design & Com puter Graphics, 2012, 24(10): 1273-1282(in Chinese)
(蒲剑苏, 屈华民, 倪明选. 移动轨迹数据的可视化[J]. 计算机辅助设计与图形学学报, 2012, 24(10): 1273-1282)

[5] Al-Dohuki S, Wu Y, Kamw F, et al. SemanticTraj: A New Approach to Interacting with Massive Taxi Trajectories[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 11-20

[6] Wang Z, Lu M, Yuan X, et al. Visual traffic jam analysis based on trajectory data[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2159-2168.

[7] Zeng W, Fu C W, Arisona S M, et al. Visualizing mobility of public transportation system[J]. IEEE transactions on visualization and computer graphics, 2014, 20(12): 1833-1842.

[8] Andrienko, N., Andrienko, G. & Gatalsky, P., 2000. Supporting visual exploration of object movement. Proceedings of the working conference on Advanced visual interfaces AVI 00, p.217 -220.

[9] Hoang M X, Zheng Y, Singh A K. FCCF: forecasting citywide crowd flows based on big data[C]//Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2016: 6.

[10] Beecham R, Wood J. Exploring gendered cycling behaviours within a large-scale behavioural data-set[J]. Transportation Planning and Technology, 2014, 37(1): 83-97.

[11] Beecham R, Wood J. Characterising group-cycling journeys using interactive graphics[J]. Transportation Research Part C: Emerging Technologies, 2014, 47: 194-206.

[12] Jiang Xiarui, Zheng Chunyi, Jiang Li, et al. Visual analysis of large taxi origin-destination data[J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(10): 1907-1917 (in Chinese)
(姜晓睿, 郑春益, 蒋莉, 等. 大规模出租车起止点数据可视

分析[J]. 计算机辅助设计与图形学学报, 2015, 27(10): 1907-1917)

[13] Wang Z, Ye T, Lu M, et al. Visual exploration of sparse traffic trajectory data[J]. IEEE transactions on visualization and computer graphics, 2014, 20(12): 1813-1822.

[14] Wood J, Dykes J, Slingsby A. Visualisation of origins, destinations and flows with OD maps[J]. The Cartographic Journal, 2010, 47(2): 117-129.

[15] Yang Y, Dwyer T, Goodwin S, et al. Many-to-Many Geographically-Embedded Flow Visualisation: An Evaluation[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 411-420.

[16] Boyandin I, Bertini E, Lalanne D. A Qualitative Study on the Exploration of Temporal Changes in Flow Maps with Animation and Small-Multiples[C]//Computer Graphics Forum. Blackwell Publishing Ltd, 2012, 31(3p): 1005-1014.

[17] Boyandin I, Bertini E, Bak P, et al. Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data[C]//Computer Graphics Forum. Blackwell Publishing Ltd, 2011, 30(3): 971-980.

[18] Zeng W, Fu C W, Müller Arisona S, et al. Visualizing Waypoints-Constrained Origin-Destination Patterns for Massive Transportation Data[C]//Computer Graphics Forum. 2015.

[19] Correll M, Heer J. Surprise! Bayesian Weighting for De-Biasing Thematic Maps[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 651-660.

[20] Yuan N J, Zheng Y, Xie X, et al. Discovering urban functional zones using latent activity trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3): 712-725.

[21] Wu W, Zheng Y, Cao N, et al. MobiSeg: Interactive Region Segmentation Using Heterogeneous Mobility Data. IEEE Pacific Visualization Symposium, 2017: 91-100.

[22] Li Y, Zheng Y, Zhang H, et al. Traffic prediction in a bike-sharing system[C]//Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2015: 33.

[23] Garrard, J., Handy, S. & Dill, J. (2012), Women and cycling, in J. Pucher & R. Buehler, eds, 'City Cycling', MIT Press, London, pp. 211–235.

[24] Jacobsen, P. L. (2003), 'Safety in numbers: more walkers and bicyclists, safer walking and bicycling', Injury Prevention 9(3), 205–209.

[25] T.Munzner. Visualization Analysis and Design.CRC Press,2014