

DOI:10.13203/j.whugis20170035



文章编号:1671-8860(2018)10-1595-08

# 出租车轨迹数据的地域间移动模式分析

龚 希<sup>1,2</sup> 陈占龙<sup>1,2</sup> 谢 忠<sup>1,2</sup>

1 中国地质大学(武汉)信息工程学院,湖北 武汉,430074

2 国家地理信息系统工程技术研究中心,湖北 武汉,430074

**摘 要:**基于地域的移动模式(zone-based movement pattern, ZMP)的发掘通过对出租车轨迹的聚类分析,同步发掘地域与移动轨迹。该方法通过 ZMP 的合并达到新地域发掘的目的,并加以距离和专题属性组成的相邻约束以保留移动的方向性、地域的功能属性以及地域间的距离关系。通过连接矩阵迭代计算得到最优合并的 ZMP 进行合并,从而发掘 ZMP,同时通过覆盖度、精准度以及基于这两者的平衡评估因子等对合并得到的 ZMP 进行评定。通过现实世界的出租车数据进行实验,结果表明该方法高效可行,能合理地实现合并现有区以发掘新地域。

**关键词:**地域分析;地域发掘;移动模式分析;移动模式发掘;出租车轨迹

**中图分类号:**P208

**文献标志码:**A

城市化发展使城市产生不同功能的区域,发掘并理解功能区域和人类移动规律可制定更好的城市规划便利人类生活<sup>[1]</sup>。Fusco 等<sup>[2]</sup>将地域定义为地理上有相邻关系的站点,本文在此基础上定义基于地域的移动模式(zone-based movement pattern, ZMP)为一种移动轨迹及该轨迹关联的一对地域——起始地域  $O$  和目的地域  $D$ ,任意 ZMP  $p$  表示为  $p = O \rightarrow D$ 。对 ZMP 的发掘即可达到对地域和移动模式的双重发掘。随着 GPS 技术的快速发展,车辆的移动记录更容易获取,因此被广泛应用于交通<sup>[3-4]</sup>、人类活动<sup>[5]</sup>分析等。本文利用出租车 GPS 数据对城市居民的出行移动路径进行分析,根据乘客移动轨迹的分布与特征得到地域之间的关系,从而实现对 ZMP 的发掘。

目前对移动轨迹的研究有如下几点:①针对城市区域中特定地域间的运动方式的挖掘,如通过基于动态规划和基因算法的方法分析特定区域间的活动模式<sup>[6]</sup>;②通过移动数据进行行为分析,如通过结合空间特征和社会经济学特征的模型分析家庭的旅行模式<sup>[7]</sup>,通过乘客的乘车记录定性分析乘客的搭乘行为<sup>[8]</sup>等。此外,移动模式的分析还可通过不同领域的数据进行分析<sup>[9]</sup>,如通过

智能卡数据评估未来的公共交通情况<sup>[10-11]</sup>。而在地域的发掘方面,普通的空间聚类方法如  $K$ -means<sup>[12]</sup>、BIRCH<sup>[13]</sup>等已发展完善,由此产生基于普通空间算法的改进算法,如基因算法<sup>[14]</sup>、基于层次结合 Voronoi 的空间聚类算法<sup>[15]</sup>等。此外,部分方法将聚类方法与约束方法相结合,如通过变化的人口普查数据进行区域的划分<sup>[16]</sup>,通过离散的 Voronoi 图结合位置关系和种子权重进行区域的划分<sup>[17]</sup>等。

以上方法在同时发掘地域和移动轨迹时仍具有一定的局限性。Kim 等<sup>[18]</sup>开创性地提出了一种通过地铁搭乘记录同时发掘地域及乘客移动模式的方法。但该方法只考虑了空间关系而忽略了属性关系,这将产生有失实际的结果,如将距离很近但功能性质相差较远的地域当成相似类地域进行合并。且地铁只能在固定路线行驶,这导致该方法使用范围受限。相较地铁,出租车可从任意出发点行驶到任意目的点,为研究人类移动动态提供更独特的视窗<sup>[3]</sup>。因此本文在此研究基础上提出一种通过出租车轨迹发掘基于地域的移动模式的方法,它通过空间和属性特征的双重约束,保证了地域间的无重复性,使发掘的 ZMP 关联的

收稿日期:2017-09-13

项目资助:国家重点研发计划(2017YFC0602204);国家自然科学基金(41401443,41671400);中央高校基本科研业务费专项资金优秀青年基金(CUG160226)。

第一作者:龚希,博士生,主要从事交通 GIS 方面的研究。gongxi\_cug@126.com

通讯作者:谢忠,博士,教授。xiezhong@cug.edu.cn

地域的功能更集中,为使用者提供更有力的决策选择。

### 1 地域及移动方式发掘分析

本文方法基于聚类分析迭代的合并同一等级的相似实体,通过 ZMP 的合并实现保留移动方向性前提下的 ZMP 发掘,达到发掘地域和移动模式的目的。距离和专题属性组成的相邻约束保证了地域专题属性和地域间距离关系的相近性。如图 1 所示,ZMP 的发掘是通过连接矩阵进行迭代的过程,连接矩阵的元素表征所有 ZMP 间合并的可能程度,选出最大值后检查其是否满足迭代停止条件,若不满足,则将最大值对应的两个 ZMP 进行合并并更新连接矩阵,否则返回当前 ZMP 结果。在了解具体方法前,先介绍数据预处理及与方法相关的概念定义。

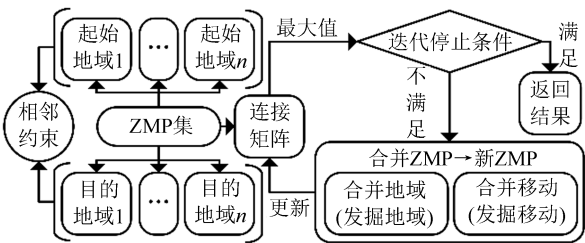


图1 ZMP 发掘的基本流程

Fig.1 The Process of Discovering ZMP

#### 1.1 数据预处理

GPS 数据源包含大量冗余的信息,需筛选出正确有用的信息,并进行格式编辑,以在发掘算法中被高效利用。预处理主要包括如图 2 所示的 4 个步骤。

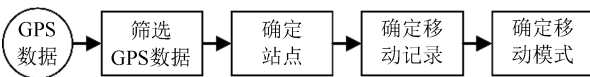


图2 数据预处理流程

Fig.2 Flow Chart of Data Preprocessing

第 1 步:GPS 数据的筛选包括两方面:①剔除错误和无用的 GPS 记录,如点范围不合理、非载客状态、非上下客状态时以及重复和不全的记录,仅保留正确的上下客点记录;②每条 GPS 记录仅保留有用的经纬度值、载客状态等项目。第 2 步:需确定站点,出租车没有固定的停靠点,GPS 点数据是随机分散的,可借助站点进行移动模式确认工作。类似地域的定义,本文将站点定义为地理上有相邻关系的点的聚类结果点,可通过所有上下客点聚类得到站点数据。K-means 是一种广泛使用的简单快速的基本聚类方法,经

过上步筛选后的 GPS 点数据基本无噪声,因此本文通过 K-means 对 GPS 点聚类得到站点,每个上下客点都有一个隶属站点。第 3 步:确定移动轨迹记录,乘客的移动情况只需通过起始点和目的点反映,一条上客点记录和一条下客点记录即可组成一条移动记录。更新移动记录的起始点和目的点为其隶属的站点,则得到最终的移动记录。此外需剔除起始站点和目的站点相同的移动记录,在某种意义上它表示乘客没有移动。第 4 步:对移动记录数据分类以获取不同移动模式,将起始站点相同且目的站点相同的移动记录归为一类移动模式,并统计该类移动模式中移动记录的数目。

#### 1.2 相邻约束

本文相邻约束实现几何和属性的双重约束。几何约束主要通过 Delaunay 三角网与 Voronoi 图<sup>[19]</sup>实现,即一个实体只与其直接 Delaunay 相邻实体间有较为明显的作用。本文定义一个实体与它的 Delaunay 相邻实体的邻接值  $g$  与它们之间的空间距离的平方成反比;而对于 Delaunay 相邻实体之外的对象, $g$  迅速衰减到可以忽略的程度。

$$g(p, q) = k \frac{1}{d_E(p, q)^2}, k = \begin{cases} 1, q \in N_p \\ 0, q \notin N_p \end{cases} \quad (1)$$

式中, $d_E(p, q)$  表示实体  $p$  与其他实体  $q$  间的空间距离; $k$  为邻近标志; $N_p$  表示与实体  $p$  的 Delaunay 相邻的实体集合。

但实体分布较分散时距离相隔很远的实体也可能是 Delaunay 相邻的,这可能导致此类实体的合并。因此需加入空间距离进一步约束,仅当两个实体 Delaunay 相邻且两者间的空间距离  $d_E$  不超过阈值  $\delta$  时才是几何相邻的。从而  $g$  值表示为:

$$g(p, q) = \frac{k}{d'_E(p, q)^2}, k = \begin{cases} 1, q \in N_p \text{ 且 } d_E(p, q) \leq \delta \\ 0, q \notin N_p \text{ 或 } d_E(p, q) > \delta \end{cases} \quad (2)$$

式中, $N_p$  表示实体  $p$  的几何相邻实体的集合; $d'_E(p, q)$  是  $d_E(p, q)$  归一化结果。

属性约束通过专题属性实现。即将空间属性与专题属性归一化后分别计算空间距离与专题属性距离,再进行加权融合即得到相邻距离  $D_E$ ,表达式为:

$$D_E(p, q) = \omega_1 \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} + \omega_2 \sqrt{\sum_{k=1}^n (A_{pk} - A_{qk})^2} \quad (3)$$

式中, $A_{pk}$  表示实体  $p$  的第  $k$  维专题属性值; $\omega_1$ 、 $\omega_2$  表示几何约束与属性约束的权值,默认情况为  $\omega_1 = \omega_2 = 0.5$ 。

通过  $D_E$  替代式(2)中的  $d'_E$  得到两个实体的最终  $g$  值为:

$$g(p, q) = k \frac{1}{D_E(p, q)^2},$$

$$k = \begin{cases} 1, & q \in N_p \text{ 且 } d_E(p, q) \leq \delta \\ 0, & q \notin N_p \text{ 或 } d_E(p, q) > \delta \end{cases} \quad (4)$$

### 1.3 地域及 ZMP 相关的定义

给定  $N$  个站点的集合  $S = \{s_1 \cdots s_N\}$ , 则有如下定义。

定义 1 地域  $z_i$  是由一个或若干站点组成的, 即  $z_i = \{s_i \cdots s_k\} (1 \leq i \leq k \leq N)$ 。不同地域不包含相同的站点, 即对任意地域  $z_i, z_j (i \neq j)$  都有  $s_p \neq s_q (s_p \in z_i, s_q \in z_j)$ 。任意站点可看作单地域, 即由单个站点组成的地域。

定义 2 两个地域间的邻接值为两个地域  $z_i, z_j$  中站点间邻接值的平均值:

$$g(z_i, z_j) = \frac{\sum g(s_p, s_q)}{n}, s_p \in z_i, s_q \in z_j \quad (5)$$

式中,  $g(s_p, s_q)$  为站点  $s_p$  和站点  $s_q$  间的邻接值;  $n$  为  $z_i, z_j$  中站点对的数目。

定义 3 地域  $z_i$  的属性平均值  $\overline{A}(z_i)$  为该地域中所有站点的属性的平均值:

$$\overline{A}(z_i) = \frac{\sum_{i=1}^n A(s_p)}{n}, s_p \in z_i \quad (6)$$

式中,  $A(s_p)$  为站点  $s_p$  归一化后的专题属性值;  $n$  为地域  $z_i$  中的站点数目。

定义 4 地域  $z_i, z_j$  间的相邻系数为邻接值与专题属性平均值差异的比值, 记为  $G(z_i, z_j)$ :

$$G(z_i, z_j) = \frac{g(z_i, z_j)}{|\overline{A}(z_i) - \overline{A}(z_j)|} \quad (7)$$

$G(z_i, z_j)$  值越大, 表明地域  $z_i, z_j$  空间距离越小, 属性越接近, 反之亦然。

定义 5 地域  $z_i, z_j$  是相邻的, 当且仅当  $G(z_i, z_j) \geq \gamma$  成立, 否则不相邻。  $\gamma$  为地域合并最小值, 只有相邻的地域才可能合并。

在以上地域的定义的基础上, 对 ZMP 可定义为: 已知初始状态下包含有  $N_1$  个地域的集合  $Z = \{z_1 \cdots z_{N_1}\}$  关联了  $N_2$  种 ZMP, 可表示为 ZMP 集  $M = \{m_1 \cdots m_{N_2}\}$ , 其中  $m_k = z_0 \rightarrow z_d (1 \leq k \leq N_2, z_0 \in Z, z_d \in Z \text{ 且 } z_0 \neq z_d)$ 。通过该 ZMP 集  $M$  迭代发掘到的第  $i$  种 ZMP 为:  $p_i = O_i \rightarrow D_i (O_i \subset Z, D_i \subset Z \text{ 且 } O_i \neq D_i)$ 。

定义 6 对 ZMP  $p_i = O_i \rightarrow D_i$  和  $p_j = O_j \rightarrow D_j$ , 若  $O_i$  和  $O_j$  相邻且  $D_i$  和  $D_j$  相邻, 则这两个

ZMP 相邻。

定义 7 两个 ZMP  $p_i = O_i \rightarrow D_i$  和  $p_j = O_j \rightarrow D_j$  间的连接值为两者的起始地域的合并地域中站点到目的地域的合并地域站点的平均数目<sup>[18]</sup>:

$$\rho_{i,j} = \rho(p_i, p_j) = k \frac{n(O_i \cup O_j \rightarrow D_i \cup D_j)}{|O_i \cup O_j| \cdot |D_i \cup D_j|},$$

$$k = \begin{cases} 1, & p_j \in N_{p_i} \\ 0, & p_j \notin N_{p_i} \end{cases} \quad (8)$$

式中,  $n(O_i \cup O_j \rightarrow D_i \cup D_j)$  表示从合并起始地域  $O_i \cup O_j$  到合并目的地域  $D_i \cup D_j$  的移动数目;  $N_{p_i}$  为  $p_i$  相邻的 ZMP 集合。两种相邻的 ZMP 覆盖的轨迹数目越多, 关联的站点越少,  $\rho_{i,j}$  就越大, 两者越可能合并。且根据定义有  $\rho_{i,j} = \rho_{j,i}$ 。

定义 8 已知两个 ZMP  $p_i = O_i \rightarrow D_i$  和  $p_j = O_j \rightarrow D_j$ , 定义两者合并得到的新 ZMP 为:

$$p_t = O_t \rightarrow D_t (O_t = O_i \cup O_j, D_t = D_i \cup D_j) \quad (9)$$

定义 9 已知 ZMP  $p_i = O_i \rightarrow D_i$  和  $p_j = O_j \rightarrow D_j$ , 若  $O_j \subset O_i$  且  $D_j \subset D_i$ , 则  $p_j$  是  $p_i$  的子集。

定义 10 已知 ZMP  $p_i = O_i \rightarrow D_i$  和  $p_j = O_j \rightarrow D_j$ , 若  $O_i = O_j \cup D_j$  或  $D_i = O_j \cup D_j$ , 则  $p_j$  是  $p_i$  的子集。子集解决不同地域出现相同站点的问题, 使得新发掘的地域无交叠, 更易理解。

通过式(8)可构建表示 ZMP 集中任意 ZMP 间连接值的连接矩阵。由于  $\rho_{i,j} = \rho_{j,i}$ , 可采用上三角连接矩阵表示  $M$  中  $N_2$  个 ZMP 间的连接值:

$$C = \begin{pmatrix} 0 & \rho_{1,2} & \cdots & \rho_{1,N_2-1} & \rho_{1,N_2} \\ 0 & 0 & \cdots & \rho_{2,N_2-1} & \rho_{2,N_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \rho_{N_2-1,N_2} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \quad (10)$$

### 1.4 新 ZMP 的发掘方法

新 ZMP 的发掘算法通过连接矩阵  $C$  获取当前迭代中最优合并的两个 ZMP。整体流程如下:

```

Initialize /*初始化阶段*/
Zone: Set each station as a single zone
ZMP: set each movement pattern as an initial ZMP
C: calculate initial connect matrix C(1) //计算初始连接矩阵
Iteration /*迭代计算阶段*/
Find  $\rho_{u,v} = \rho^{*(k)} = \max(\rho_{i,j}^{(k)})$ 
Merge two ZMP  $p_u$  and  $p_v$ , into  $p_t$ 
if ( $\rho_s$  is a subset of  $\rho_t$ ) Merge  $\rho_s$  to  $\rho_t$  //检查子集
if ( $\rho_s$  is a self-set of  $\rho_t$ ) Remove  $\rho_s$  from  $M$  //检查子集
Update  $Z, M, C^{(k)}$  respectively
if ( $\rho^{*(k)} < \theta$ ) Return  $Z, M$  /*返回结果*/
    
```

在初始化阶段, 将每个站点初始化为单地域

得到地域集  $Z$ , 移动模式初始化为 ZMP 得到 ZMP 集  $M$ , 则可得初始连接矩阵  $C^{(1)}$ ; 第  $k$  次迭代时对当前  $M$  包含的  $R_k$  条 ZMP 计算连接矩阵  $C^{(k)}$ , 将矩阵最大元素  $\rho^{*(k)} = \max(\rho_{i,j}^{(k)}) (i, j = 1 \dots R_k \text{ 且 } i \neq j)$  关联的两个 ZMP 合并成新 ZMP  $p_i$ , 在剩余 ZMP 中找到  $p_i$  的子集合并到  $p_i$ ,  $M$  将移除所有合并入  $p_i$  的 ZMP 并加入新 ZMP  $p_i$ , 同时移除的 ZMP 关联的地域会从  $Z$  中移除, 而  $p_i$  关联的两个新地域将被加入。此外,  $p_i$  的自集关联的两个地域在更新关联地域后同属一个, 需从  $M$  中移除。一次迭代后,  $M$  中会减少至少 2 个 ZMP, 新增 1 个 A 类 ZMP, 有若干 C 类 ZMP 可能更新为 B 类 ZMP。更新完  $Z, M$  及连接矩阵后继续下一次迭代, 直到剩余的 ZMP 不能合并或连接矩阵中的最大值比阈值  $\theta$  小, 停止迭代, 将  $M$  返回得到最终结果。

本文将类似  $p_i$  这样由其他 ZMP 合并成的新 ZMP 记为 A 类 ZMP。若被移除的地域在  $M$  中仍有其他关联的 ZMP, 则需用其所属的新地域更新这些 ZMP 的关联地域, 本文记这类与新地域关联但不是新合成的 ZMP 为 B 类 ZMP, 剩下与新地域无关联的 ZMP 记为 C 类 ZMP, 三类 ZMP 的区别如表 1 所示。

表 1 三类 ZMP 的区别

Tab.1 Difference of Three Types of ZMP

条件	A 类	B 类	C 类
由两个 ZMP 合并成	✓	×	×
关联地域中有新地域	✓	✓	×

## 2 实验与评价

### 2.1 数据预处理

本文采用北京市连续 3 天的出租车 GPS 数据。筛选后得到 195 622 条移动轨迹记录, 覆盖了 391 244 个上下客站点(见图 3(a))。通过 K-means 对其初聚类得到站点数据,  $K$  值决定了初始地域的粒度, 继而影响最终结果地域的粒度,  $K$  值越大, 产生的新地域的覆盖范围越小、越精细, 根据不同的数据及需求可选择不同的  $K$  值。本文以中粒度 100 个站点为例进行实验(见图 3(b)), 本文专题属性采用不同类型的兴趣点在北京的分布来表示(见图 3(c)、图 3(d)), 选取的 5 种类型的兴趣点分别为购物点、教育点、景点、企业点和居民区。这与人类的日常生活关联较为紧密, 因此足够表达出城市的不同的功能, 本文兴趣

点数据来源于百度地图。

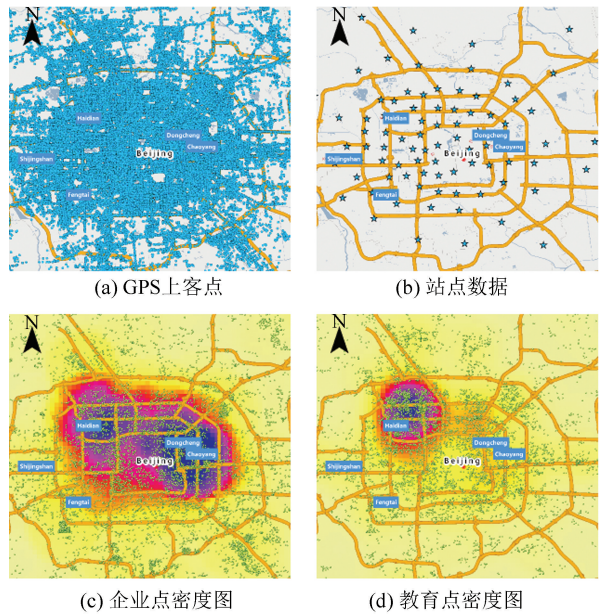
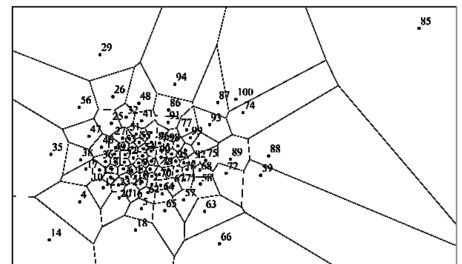


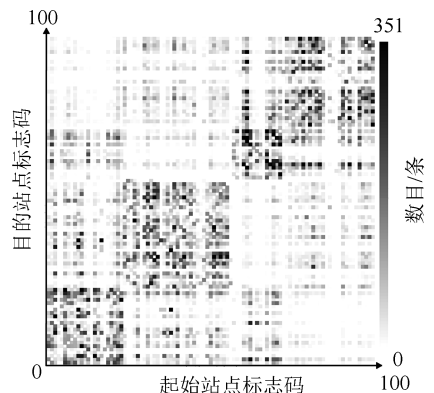
图 3 实验数据成图

Fig.3 Visualization of the Data

图 4(a)是 100 个站点间的 Voronoi 图, 图中的数字是相应站点的标识码, 由此可推断站点间 Delaunay 相邻关系。对 195 622 条移动记录分类, 并去掉起始站点与目的站点相同的记录, 得到 674 种移动模式共覆盖 182 519 条移动轨迹。对移动模式覆盖的移动记录数目的统计结果如图



(a) Voronoi图



(b) 移动轨迹数目统计

图 4 100 个站点间的关系

Fig.4 Relation Between the 100 Stations

4(b)所示,横、纵坐标分别为起始和目的站点标识码,站点标志码从 0 增到 100,图中格子颜色越深,表示移动数目越多,其能反映更多人的活动轨迹,会优先被用到算法中。但该图中浅色占大部分,这类移动轨迹数目较少的移动模式在发掘算法中大部分是不被用到的,去掉这部分移动模式可减少算法的计算和时间耗费,并消除噪声。本文以去掉移动数目小于 100 的移动模式为例进行实验,经筛选后剩余 336 种移动模式共覆盖 48 644 条移动轨迹。将每个站点初始化单地域,则 336 种移动模式可全部初始化 C 类 ZMP 进行后续发掘。

为定量评估新产生的 ZMP 和地域,本文采用 3 种评估值<sup>[30]</sup>:评估覆盖度的  $v$  值、评估精准度的  $a$  值和对前两者折中评估的  $c$  值。如对 ZMP  $p_i=O_i \rightarrow D_i$  的评估方法如式(11)所示。

$$\left\{ \begin{aligned} v(p_i) &= r(O_i \rightarrow D_i) = \frac{n(O_i \rightarrow D_i)}{m} \\ a(p_i) &= \frac{r(O_i \rightarrow D_i)}{r(O_i \rightarrow D_i) \cdot r(O_i \rightarrow D_i)} = \frac{mn(O_i \rightarrow D_i)}{n(O_i \rightarrow D_i) \cdot n(O_i \rightarrow D_i)} \\ c(p_i) &= \sqrt{v(p_i)a(p_i)} = \frac{n(O_i \rightarrow D_i)}{\sqrt{n(O_i \rightarrow D_i) \cdot n(O_i \rightarrow D_i)}} \end{aligned} \right. \quad (11)$$

其中,  $r(O_i \rightarrow D_i)$ 、 $r(O_i \rightarrow D_i)$ 、 $r(O_i \rightarrow D_i)$  依次为从  $O_i$  到  $D_i$  的移动轨迹数目比例、以  $D_i$  为目的地域的移动轨迹数目比例、以  $O_i$  为起始地域的移动轨迹数目的比例;  $n(O_i \rightarrow D_i)$  是从  $O_i$  到  $D_i$  移动轨迹数目;  $m$  是移动轨迹的总数目。  $v$  值通过 ZMP 覆盖的移动轨迹数目计算,  $v$  值越高,则相应 ZMP 的关联地域间的联系更紧密;  $a$  值通过 ZMP 关联的起始地域和目的地域的独立度计算,其暗示关联的两个地域间有内在联系;  $c$  值是覆盖度和准确度的平衡,它可挖掘到性能较好但不易发现的隐藏 ZMP。

### 2.2 实验结果及分析

地域合并最小值  $\gamma$  的影响将在后文详细描述。本文以  $\gamma=10$  为例,使算法迭代到没有可合并的 ZMP 为止,最终共迭代了 19 次后返回最终结果。图 5(a)-5(d)依次是第 5、10、15、19 次迭代后产生的 A 类 ZMP 关联的新地域。

最终 ZMP 集中有 238 个 ZMP,包含 14 个 A 类 ZMP(见图 6),124 个 B 类 ZMP 以及 100 个 C 类 ZMP。最终结果产生了 13 个新地域(见表 2)。

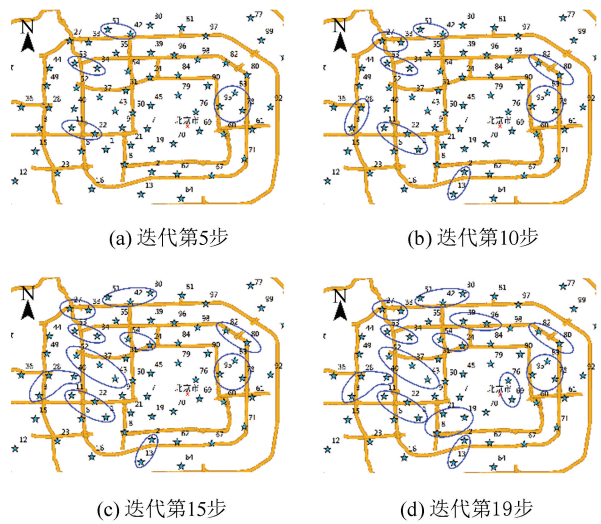


图 5 迭代中产生的新地域

Fig.5 Newly Discovered Zones in Iteration

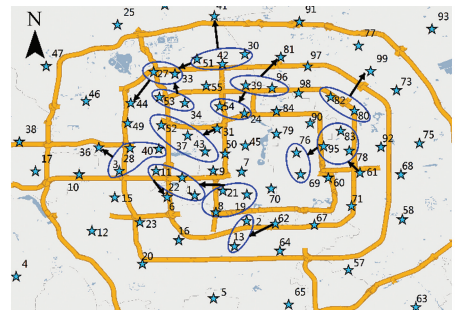


图 6 新发掘的 A 类 ZMP

Fig.6 Newly Discovered A Type ZMP

表 2 新地域

Tab.2 Newly Discovered Zones

地域	组成站点	地域	组成站点
z <sub>1</sub>	95, 83, 78	z <sub>8</sub>	52, 43, 37
z <sub>2</sub>	53, 34	z <sub>9</sub>	40, 28, 3
z <sub>3</sub>	22, 11, 1	z <sub>10</sub>	51, 42, 30
z <sub>4</sub>	13, 2	z <sub>11</sub>	96, 39
z <sub>5</sub>	82, 80	z <sub>12</sub>	76, 69
z <sub>6</sub>	33, 27	z <sub>13</sub>	21, 19, 8
z <sub>7</sub>	54, 24		

为评估 238 个 ZMP,分别计算 3 类 ZMP 的  $v$  值、 $a$  值、 $c$  值的平均值,并与 336 个初始 ZMP 和 238 个结果 ZMP 的总体平均值对比。图 7(a)是平均  $v$  值的对比结果,可发现 A 类 ZMP > B 类 ZMP > C 类 ZMP,即发掘的新地域关联的 ZMP (A、B 类 ZMP 的并集) 相较普通 ZMP 有较好的覆盖度;同时结果 ZMP 相比初始 ZMP,平均  $v$  值有很大提升,从侧面反映出发掘的新 ZMP 有较大的覆盖度。图 7(b)、图 7(c)是平均  $a$  值和平均  $c$  值的对比结果,与平均  $v$  值结果相同,即 A 类 ZMP 不论在覆盖度、精准度还是两者折中的评估

标准上都有最好的评估结果,B类 ZMP 次之,且都优于 C 类 ZMP 以及未经处理的初始 ZMP。由

此反映算法发掘到的新 ZMP 相较其他 ZMP 有更优的性能。

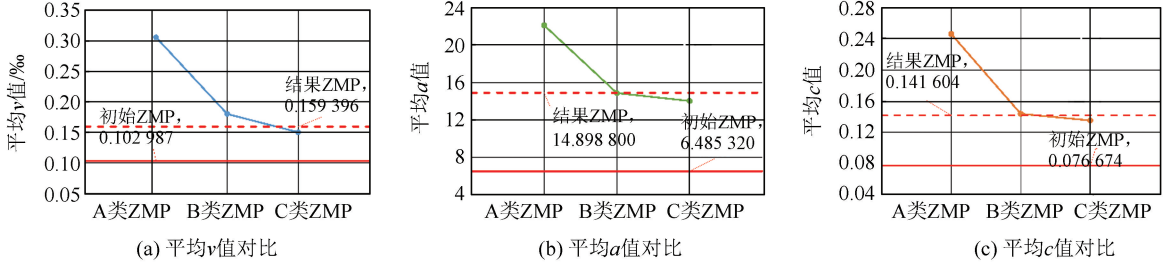


图7 平均v值、a值、c值对比

Fig.7 Comparison of the Average v, a, c Values

为研究  $\gamma$  值影响,固定迭代次数为 15 次,改变  $\gamma$  值,变化如表 3 所示。随着  $\gamma$  的增大,新地域的数目先从 5 增加到 11 后减小至 0,参与合并新地域的初始地域数目也有相同的变化趋势,而新地域中的最大地域(组成地域数目最多的地域)组成的地域数目却一直在减少。当  $\gamma$  较小时如  $\gamma \in [0, 1]$  时,更多地域间的相邻系数满足  $G > \gamma$ ,导致更多地域被合并到同一新地域中,从而产生的新地域不多,且更多次迭代耗费在合并初始地域与另一新地

域而非两个初始地域上,因此参与合并的初始地域总数增长缓慢;当  $\gamma$  在  $[1, 5)$ ,满足  $G > \gamma$  的地域对的数目逐渐减小,更多地产生由初始地域两两合成的新地域,新地域数目增幅明显。同理,更多次迭代会耗费在将合并两个初始地域上,参与合并的初始地域总数随之增多,直到  $\gamma = 5$  达到最大。当  $\gamma$  继续增大,满足  $G > \gamma$  的地域对数目剧烈减少,各项指标也随之降低,直到  $\gamma = 220$  时已无满足  $G > \gamma$  的地域对可合并,几项指标也随之降到 0。

表3  $\gamma$  值对产生的新地域的影响

Tab.3 Influence of  $\gamma$  on the Newly Discovered Zone

地域	$\gamma$ 值										
	0	0.1	1	5	10	20	30	40	50	200	220
新地域	5	5	6	11	10	9	4	3	2	2	0
参与合并初始地域	20	20	21	26	25	18	8	6	4	2	0
最大地域组成地域	12	11	9	4	3	2	2	2	2	1	0

表 4 为不同  $\gamma$  值下的结果 ZMP 的  $c$  值评估。 $\gamma \geq 220$  时不会产生新 ZMP,因此不对该范围讨论。随着  $\gamma$  的增大,C类 ZMP 平均的  $c$  值与结果 ZMP 的平均  $c$  值基本持平,A类、B类 ZMP 的平均  $c$  值大体呈下降的趋势,但一直都高于初始 ZMP 的平均  $c$  值,证明通过该算法可在初始

ZMP 中发掘到性能优良的隐藏 ZMP。且不论  $\gamma$  为何值,都有 A 类 ZMP > B 类 ZMP > C 类 ZMP,即算法发掘的新地域关联的 ZMP 的  $c$  值优于普通 ZMP,反映出该算法在多种条件下都能发掘当前较优的 ZMP 和地域。

表4  $\gamma$  值对  $c$  值的影响

Tab.4 Influence of  $\gamma$  on c Value

ZMP	$\gamma$ 值										
	0	0.1	1	5	10	20	30	40	50	200	220
A 类 ZMP	0.346 8	0.332 8	0.295 8	0.218 1	0.238 9	0.202 4	0.172 3	0.178 6	0.193 6	0.168 9	—
B 类 ZMP	0.142 6	0.139 0	0.137 1	0.138 8	0.136 7	0.142 2	0.130 1	0.145 5	0.151 4	0.135 9	—
C 类 ZMP	0.098 0	0.104 6	0.112 1	0.133 9	0.122 3	0.111 8	0.091 4	0.086 8	0.083 0	0.080 3	—
结果 ZMP	0.103 6	0.109 9	0.118 1	0.138 5	0.127 5	0.114 6	0.092 4	0.087 6	0.083 7	0.080 6	—
初始 ZMP						0.076 7					—

### 3 结 语

本文研究了基于出租车乘车记录的 ZMP 的

发掘即移动模式和地域的双重发掘。通过从出租车轨迹数据筛选出有用的数据来达到同时确认地域和地域之间的移动模式的目的,并以北京出租车数据为例进行了模型验证分析,实验结果显示,本文方法在发掘 ZMP 上能得到满意的结果,新

发掘到的移动模式与其关联地的地域将有助于决策者更好地理解地域的存在以及这些地域之间的关系。下一步的工作将比较预处理中不同聚类算法对 ZMP 发掘的影响,同时会考虑采纳更多属性因素以及对连接矩阵的迭代算法进一步优化,以提高新发掘的 ZMP 的性能与分析效率。

### 参 考 文 献

- [1] Qi G, Li X, Li S, et al. Measuring Social Functions of City Regions from Large-Scale Taxi Behaviors [C]. IEEE International Conference on Pervasive Computing and Communications, Seattle, USA, 2011
- [2] Fusco G, Cagliani M. Hierarchical Clustering Through Spatial Interaction Data. The Case of Commuting Flows in South-Eastern France[C]. International Conference on Computational Science and Its Applications, Santander, Spain, 2011
- [3] Castro P S, Zhang D, Chen C, et al. From Taxi GPS Traces to Social and Community Dynamics: A Survey [J]. *ACM Computing Surveys*, 2014, 46 (2):1 167-1 182
- [4] Ren huiJun, Xu Tao, Li Xiang. Personal Profile Mining Based on Mobile Phone Location Data[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 739-744 (任慧君, 许涛, 李响. 利用车载 GPS 轨迹数据实现公交车驾驶安全性分析[J]. *武汉大学学报·信息科学版*, 2014, 39 (6):739-744)
- [5] Liu Y, Kang C, Gao S, et al. Understanding Intra-Urban Trip Patterns from Taxi Trajectory Data[J]. *Journal of Geographical Systems*, 2012, 14(4): 463-483
- [6] Joh C H, Arentze T A, Timmermans H J P. Multi-dimensional Sequence Alignment Methods for Activity-Travel Pattern Analysis: A Comparison of Dynamic Programming and Genetic Algorithms[J]. *Geographical Analysis*, 2001, 33(3): 247-270
- [7] Srinivasan S, Ferreira J. Travel Behavior at the Household Level: Understanding Linkages with Residential Choice [J]. *Transportation Research Part D: Transport and Environment*, 2002, 7 (3): 225-242
- [8] Hoffman M, Wilson S P, White P. Automated Identification of Linked Trips at Trip Level Using Electronic Fare Collection Data [C]. The Transportation Research Board 88th Annual Meeting, Washington D C, USA, 2009
- [9] Liu Yu, Kang Chaogui, Wang Fahui. Towards Big Data-Driven Human Mobility Patterns and Models [J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 660-666 (刘瑜, 康朝贵, 王法辉. 大数据驱动的人类移动模式和模型研究 [J]. *武汉大学学报·信息科学版*, 2014, 39(6): 660-666)
- [10] Chu K K A, Chapleau R. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling [J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2008, 2 063 (1):63-72
- [11] Munizaga M A, Palma C. Estimation of a Disaggregated Multimodal Public Transport Origin-Destination Matrix from Passive Smartcard Data from Santiago, Chile[J]. *Transportation Research Part C*, 2012, 24(9):9-18
- [12] Macqueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]. The 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, 1967
- [13] Zhang T, Ramakrishnan R, Livny M. BIRCH: An Efficient Data Clustering Method for very Large Databases [C]. The International Conference Management of Data, Montreal, Canada, 1996
- [14] Bação F, Lobo V, Painho M. Applying Genetic Algorithms to Zone Design [J]. *Soft Computing*, 2005, 9(5):341-348
- [15] Deng M, Liu Q, Li G, et al. Field-Theory based Spatial Clustering Method [J]. *Journal of Remote Sensing*, 2010, 14(4):702-717
- [16] Martin D. Extending the Automated Zoning Procedure to Reconcile Incompatible Zoning Systems [J]. *International Journal of Geographical Information Science*, 2003, 17 (2): 181-196
- [17] Moreno-Regidor P, Manso-Callejo M A. Zone Design of Specific Sizes Using Adaptive Additively Weighted Voronoi Diagrams [J]. *International Journal of Geographical Information Science*, 2012, 26 (10): 1 811-1 829
- [18] Kim K, Oh K, Lee Y K, et al. An Analysis on Movement Patterns Between Zones Using Smart Card Data in Subway Networks[J]. *International Journal of Geographical Information Science*, 2014, 28(9): 1 781-1 801
- [19] Gold C M. Problems with Handling Spatial Data—The Voronoi Approach[J]. *CISM Journal ACS-GC*, 1994, 45(1): 65-80

# Analysis on Zone-based Movement Pattern Using Taxi Trajectory Data

GONG Xi<sup>1,2</sup> CHEN Zhanlong<sup>1,2</sup> XIE Zhong<sup>1,2</sup>

1 Faculty of Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China

2 National Engineering Research Center of Geographic Information System, Wuhan 430074, China

**Abstract:** We propose an integrated approach to discover both zones and movement trajectories among zones, which referred to as zone-based movement pattern (ZMP), from taxi trajectory data. This method discovers the zones by merging ZMPs, which keeps the directionality of movement, thematic attributes and distance relationship of zones by the adjacent constraints consists of distant and thematic attributes. By joint average frequencies, we can identify new ZMP by iteratively calculating the best candidate ZMPs to be merged then. In addition, evaluation measures of ZMP are suggested in terms of factors such as coverage, accuracy and a tradeoff of both them. The effectiveness of the proposed approach is demonstrated through a real-world data set obtained, the experiment result shows that the approach can merge the existing zones to discover new ZMP rationally.

**Key words:** zone analysis; zone discovery; movement pattern analysis; movement pattern discovery; taxi trajectory

**First author:** GONG Xi, PhD candidate, specializes in traffic GIS. E-mail: gongxi\_cug@126.com

**Corresponding author:** XIE Zhong, PhD, professor. E-mail: xiezhong@cug.edu.cn

**Foundation support:** The National Key R & D Program of China, No.2017YFC0602204; the National Natural Science Foundation of China, Nos. 41401443, 41671400; the Fundamental Research Funds for the Central Universities, No. CUG160226.

(上接第 1517 页)

# A Method for Automatic Water High Light Detection and Removal in Single UAV Image

YAN Li<sup>1</sup> ZHOU Jiantong<sup>1</sup>

1 School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

**Abstract:** UAV(unmanned aerial vehicle) is convenient and low cost, water reflection can cause high light in the image and has adverse effect on UAV data processing and the DOM(digital orthophoto map) quality. We propose a method for automatic water high light detection and removal in single UAV image. Firstly, we extracted initial candidate highlight regions using multi-scale threshold detection in a high light component proposed in this paper and used the Grabcut algorithm to optimize them. Next, a decision tree was built according to the characteristics of water high light to eliminate the error detection. Then, high light regions were refined by the high light points nearby. Finally, we modified Criminisi algorithm to remove high light regions. In actual UAV images, the experimental results show that our method can remove the water high light well and is superior to other methods proposed by Mallick, Shen and Yoon in terms of PNSR(peak signal to noise ratio) and SSIM(structural similarity index) parameters and visual evaluation, and improves the DOM quality as well.

**Key words:** water high light detection; water high light selection; water high light removal; high light component; improved Criminisi algorithm

**First author:** YAN Li, PhD, professor, specializes in the theories and methods of photogrammetry, remote sensing and LiDAR technology. E-mail: liyan@sgg.whu.edu.cn

**Corresponding author:** ZHOU Jiantong, postgraduate. E-mail: ttong0202@163.com

**Foundation support:** The Special Scientific Research Fund of Land and Resource Public Welfare Profession of China, No.201511009.