DOI:10. 3969/j. issn. 1672-0032. 2024. 02. 013

基于网格聚类优化的区域热点路径识别

翁旭艳,郑淑妮
浙江省交通运输科学研究院,浙江 杭州　310023

摘要:针对轨迹聚类方法难以准确识别高相似热点路径的问题,提出能区分起讫点或局部路段的热点路径识别方法,将出行轨迹映射并压缩为移动网格序列,分别从边界与内部区分序列间的空间相似性度量,整合转化为距离并进行基于方格序列密度的空间聚类(grid sequencedensity-based spatial clustering of applications with noise, GS-DBSCAN)。以青岛市市南区部分出租车的轨迹数据为例,与只考虑内部相似性且分别以对比序列中较短序列和较长序列为基准的聚类方法进行对比验证。结果表明:同时考虑边界与内部相似性且以较长序列为基准的 GS-DBSCAN 算法能正确区分多出行起讫点分布下长度差异较大的分离、汇合与交叉耦合热点路径,受路径长度、网格尺寸等变量差异的影响小于 2%,且聚类运算效率较高。

关键词:热点路径;边界;内部;轨迹聚类
中图分类号:U491　　**文献标志码**:A　　**文章编号**:1672-0032(2024)02-0089-08
引用格式:翁旭艳,郑淑妮. 基于网格聚类优化的区域热点路径识别[J]. 山东交通学院学报,2024,32(2):89-96.
WENG Xuyan, ZHENG Shuni. Regional hotspot path identification based on grid clustering optimization [J]. Journal of Shandong Jiaotong University,2024,32(2):89-96.

0　引言

在交通领域广泛应用全球定位系统(global positioning system,GPS),可得到大量车辆轨迹数据,为车辆路径规划、交通运行等研究提供数据支持。Wang 等[1]、Zhang 等[2]通过出租车轨迹 GPS 数据检测车辆异常轨迹;Liu 等[3]、陈振华[4]通过出租车 GPS 轨迹数据分析城市交通拥堵模式;Lin 等[5]通过 GPS 轨迹数据预测城市交通路网流量。通过 GPS 轨迹数据分析热点路径的研究较多,热点路径指在一段时间内车辆频繁经过的路段,通过识别热点路径,可为绿波带设置、主路优先等交通运行管理提供决策依据。采用聚类方法识别热点路径的关键是相似性度量[6]。李颖等[7]分别采用离散弗雷歇算法、动态时间规整算法、最长公共序列(longest common subsequence,LCS)算法和实序列编辑距离算法对货车轨迹数据进行聚类,认为 LCS 算法的有效性最优。Kim 等[8]以 LCS 算法下车辆行驶距离与较短轨迹距离之比作为轨迹空间相似度进行基于密度的聚类(density-based spatial clustering of application with noise,DBSCAN)算法,分析交通网络中空间和时间出行模式。冯琦森[9]、赵欣[10]改进 DBSCAN 算法,前者结合重庆地形特点考虑海拔高度因素,认为比对轨迹中的最大海拔高度差不应超过一定范围,否则相似度为 0;后者在轨迹点空间相似度度量上增加车辆转角差异性,定义时间维度上的轨迹相似度。Kang 等[11]通过直接度量轨迹网格序列间重叠区域表征轨迹相似性。吴俊伟等[12]将网格抽象为图模型并通过图搜索进行网格聚类。王超[13]考虑时空事件的密度和类型对出租车轨迹点进行聚类。Lee 等[14]提出基于轨迹分段的相似性度量,对速度或方向上快速变化的轨迹点分段。在轨迹映射至网格空间的聚类研究中,可通过直接度量轨迹网格序列间重叠区域表征轨迹相似性[15],也可将网格抽象为图,以邻接网格的共有轨迹定义网格间可

收稿日期:2023-02-02
基金项目:浙江省交通运输科技计划项目(2023001)
第一作者简介:翁旭艳(1993—),女,上海人,工程师,工学硕士,主要研究方向为城市交通,E-mail:2445951226@qq.com。

达性,通过图搜索进行网格聚类及热点路径识别[16]。通过优化相似度算法、增加相关因素变量等全面、准确地分析轨迹数据,但以较短轨迹为基准的 LCS 相似度可能造成过于重视不同热点路径的重叠部分,导致将重叠度较高的不同路径归为同一类热点路径。起讫点(origin-destination,OD)矩阵和路径流量的研究较多,但考虑轨迹数据的边界起讫点和内部中间点属性的研究较少。双层规划模型[17]在 OD 矩阵估计中应用较多,Ou 等[18]基于机器学习算法,提出估计动态 OD 流量的双层框架;Tang 等[19]将改进的三维卷积神经网络模型用于动态 OD 矩阵估计;Cao 等[20-21]提出的 OD 矩阵,采用极大似然法拟合各路段的速度-密度函数,建立基于动态交通分配的双层广义最小二乘法估计器,分析浮动车采样频率较低时的 OD 矩阵估计方法,保证正常估计路段流量;Huang 等[22]采用长短期记忆(long short-term memory,LSTM)模型估计 OD 矩阵。以较短轨迹为基准的 LCS 相似度识别热点路径轨迹,易将不同起讫点的路径归为同一类热点路径。

本文优化基于轨迹数据的路径识别方法,以区域网格化为基础,将车辆轨迹压缩为仅关注移动特征的网格序列,序列间的相似性度量包含边界与内部 2 部分,考虑 GPS 定位漂移及出行产生与吸引的集聚性,边界网格的相似性通过基于热点网格密度的空间聚类(hot grid density based spatial clustering of application with noise, HG-DBSCAN)得到热点小区度量,内部网格以基于网格的改进 LCS 相似度表征,二者融合后转化为距离度量,最后进行基于方格序列密度的空间聚类(grid sequence density-based spatial clustering of applications with noise,GS-DBSCAN),优化识别热点路径。

1 热点路径识别优化

1.1 区域网格化

假设研究区域的经度范围为$[\psi_{\min},\psi_{\max}]$,纬度范围为$[\gamma_{\min},\gamma_{\max}]$,正方形网格边长为 k,将研究区域划分为若干正方形网格。综合考虑路网密度、车辆 GPS 轨迹点密度及运算效率等因素确定网格尺寸,网格过大,易覆盖同方向相互平行的若干道路;网格过小,会导致大部分网格没有 GPS 轨迹点落入且运算效率低。

用 $G_{x,y}$ 表示网格,x、y 分别为 $G_{x,y}$ 在网格坐标系下的横、纵坐标。网约车订单 i 的轨迹点列表为 $\boldsymbol{O}_i=\{\boldsymbol{P}_1^i,\boldsymbol{P}_2^i,\cdots,\boldsymbol{P}_n^i,\cdots,\boldsymbol{P}_N^i\}$,其中,$N$ 为轨迹点数;$\boldsymbol{P}_n^i(n\in[1,N])$ 为第 n 个轨迹点,是包含 3 项基本信息的一维向量,$\boldsymbol{P}_n^i=(t_n^i \quad \psi_n^i \quad \gamma_n^i)$,$\psi_n^i$、$\gamma_n^i$ 分别为 t_n^i 时刻车辆的经、纬度坐标。以 $\boldsymbol{P}_n^i$ 为例说明车辆 GPS 轨迹点与网格的映射关系,公式为:

$$\begin{cases} x=\lfloor(\psi_n^i-\psi_{\min})/k\rfloor \\ y=\lfloor(\gamma_n^i-\gamma_{\min})/k\rfloor \end{cases}, \tag{1}$$

式中$\lfloor\ \rfloor$为取整符号。

1.2 序列边界

GPS 定位有随机漂移特征,挖掘热点出行区域时不直接基于轨迹点进行聚类,而是先将轨迹抽象为热点网格,再通过 HG-DBSCAN 算法聚类,提升搜索效率,且可处理任意形状和大小的簇。采用 HG-DBSCAN 对热点网格进行空间聚类时,作以下 3 个基本定义。

定义 1 热点网格。$G_{x,y}$ 包含的 GPS 轨迹起讫点数 w 应大于预先设定的判断阈值 λ。

定义 2 网格经、纬度坐标。考虑到 $G_{x,y}$ 中 GPS 轨迹点分布不均匀,采用网格内轨迹点的平均经、纬度表示网格经、纬度坐标,公式为:

$$\begin{cases} \overline{\psi}_{x,y}=\sum_{q=1}^{w}\psi_q/w \\ \overline{\gamma}_{x,y}=\sum_{q=1}^{w}\gamma_q/w \end{cases}。$$

定义 3　网格球面距离。任意 2 个网格 $G_{x,y}$、$G_{x',y'}$在经、纬度坐标下的球面距离[23]

$$d(G_{x,y},G_{x',y'})=\Delta\eta R\ ,$$

式中:$\Delta\eta$ 为 $G_{x,y}$ 的平均经、纬度坐标 $(\bar{\psi}_{x,y},\bar{\gamma}_{x,y})$、$G_{x',y'}$的平均经、纬度坐标 $(\bar{\psi}'_{x',y'},\bar{\gamma}'_{x',y'})$ 连线与地球中心线的夹角,$\Delta\eta=2\arcsin(\sqrt{\sin^2(\Delta\psi/2)+\cos\bar{\psi}\cos\bar{\psi}'\sin^2(\Delta\gamma/2)})$,其中 $\Delta\psi$ 为 2 个网格经度之差,$\Delta\gamma$ 为 2 个网格纬度之差;R 为地球半径。

HG-DBSCAN 的网格密度是以热点网格的经纬度坐标为圆心,在指定半径 E_{ps} 内的热点网格数。若网格密度大于给定的最小核心网格判别数 N_{um},则该网格为核心网格。边界网格指落在某核心网格的邻域内,但本身不是核心网格;噪声网格是既非核心也非边界的网格。HG-DBSCAN 算法流程包括输入、初始化、运行、输出。

1)输入。输入 λ、E_{ps}、N_{um}。

2)初始化。将分析时间内所有订单轨迹的起讫点映射至对应网格;依据定义 1 确定热点网格集合 $\boldsymbol{H}=\{\boldsymbol{G}_1^{\mathrm{h}},\boldsymbol{G}_2^{\mathrm{h}},\cdots,\boldsymbol{G}_b^{\mathrm{h}},\cdots,\boldsymbol{G}_B^{\mathrm{h}}\}$,其中,$B$ 为热点网格数,$\boldsymbol{G}_b^{\mathrm{h}}$ 为第 b 个热点网格,$\boldsymbol{G}_b^{\mathrm{h}}=(x\quad y\quad \bar{\psi}\quad \bar{\gamma}\quad l)_b^{\mathrm{h}}$,其中 l 为网格的类标签,$l=-2$,簇编号 $C=-1$。

3)运行。(1)随机选取 1 个类标签 $l=-2$ 的热点网格 $\boldsymbol{G}_b^{\mathrm{h}}$,若 $\boldsymbol{G}_b^{\mathrm{h}}$ 的 E_{ps} 邻域内至少有 N_{um} 个热点网格 $\boldsymbol{G}^{\mathrm{h}}$,则 $C+1$,并将第 b 个热点网格的类标签 l_b^{h} 赋值为 C,令 $\boldsymbol{H}'$ 为 $\boldsymbol{G}_b^{\mathrm{h}}$ 邻域中的热点网格集合,循环遍历 $\boldsymbol{H}'$ 中的每个热点网格直至结束,若类标签 $l_{b'}^{\mathrm{h}}=-2$,则将 $l_{b'}^{\mathrm{h}}$赋值为 C,同时若 $\boldsymbol{G}_{\mathrm{b}'}^{\mathrm{h}}$邻域内至少有 N_{um} 个 $\boldsymbol{G}^{\mathrm{h}}$,则将 $\boldsymbol{G}_{b'}^{\mathrm{h}}$邻域中的热点网格追加至 $\boldsymbol{H}'$;(2)若 $\boldsymbol{G}_b^{\mathrm{h}}$ 的 E_{ps} 邻域内没有大于或等于 N_{um} 个热点网格 $\boldsymbol{G}^{\mathrm{h}}$,$l_{\mathrm{b}}^{\mathrm{h}}=-1$;(3)重复步骤(1)(2),直到没有 $l=-2$ 的 $\boldsymbol{G}_b^{\mathrm{h}}$。

4)输出。依据簇内轨迹点数确定热点小区集合 $\boldsymbol{Z}=\{\boldsymbol{Z}_1^{\mathrm{h}},\boldsymbol{Z}_2^{\mathrm{h}},\cdots,\boldsymbol{Z}_a^{\mathrm{h}},\cdots,\boldsymbol{Z}_A^{\mathrm{h}}\}$,其中,$A$ 为热点小区数,$\boldsymbol{Z}_a^{\mathrm{h}}$ 为第 a 个热点小区,含若干类标签相同且大于-1 的 $\boldsymbol{G}_b^{\mathrm{h}}$。

1.3　序列内部

将轨迹点的 LCS 扩展至网格序列可提升搜索效率,且 LCS 允许序列部分形变(车辆运动的随机性)。订单轨迹 O_i 通过式(1)映射为网格序列 S_i,删除 S_i 中的重复网格,得到压缩后的 S_i'。考虑网格化下,同一路径的不同轨迹-网格序列受驾驶行为、GPS 定位漂移及网格大小等因素影响,网格的重叠率可能较低,建立基于网格的 LCS,需定义网格间的空间相似度。

给定 2 个网格序列 $\boldsymbol{S}_i'=\{\boldsymbol{G}_1^{i'},\boldsymbol{G}_2^{i'},\cdots,\boldsymbol{G}_{N'}^{i'}\}$ 与 $\boldsymbol{S}_j'=\{\boldsymbol{G}_1^{j'},\boldsymbol{G}_2^{j'},\cdots,\boldsymbol{G}_{M'}^{j'}\}$,计算 $\boldsymbol{S}_i'$中的第 n 个网格 $\boldsymbol{G}_n^{i'}=(t\ x\ y)_n^{i'}$ 和 S_j'中的第 m 个网格 $\boldsymbol{G}_m^{j'}=(t\ x\ y)_m^{j'}$ 间的欧式距离

$$d_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})=\sqrt{(x_n^{i'}-x_m^{j'})^2+(y_n^{i'}-y_m^{j'})^2}\ ,$$

式中:$x_n^{i'}$、$y^{i'}{}_n$ 分别为 $\boldsymbol{G}_n^{i'}$ 的横、纵坐标,$x_m^{j'}$、$y_n^{j'}$ 分别为 $\boldsymbol{G}_m^{j'}$ 的横、纵坐标。

对网格相似度 $s_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})$ 进行归一化处理,公式为:

$$s_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})=\begin{cases}0, d_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})>\delta\\ 1-d_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})/\delta, d_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})\leqslant\delta\end{cases},$$

式中 δ 为网格间允许的最大临近距离。

基于网格相似度,通过动态规划方法确定 $\boldsymbol{S}_i'$与 $\boldsymbol{S}_j'$间的最长公共序列,需将待求解的问题划分为若干阶段,定义各阶段的状态及变量,确定阶段间的状态转移方程。$\boldsymbol{S}_{i(n)}'$为 $\boldsymbol{S}_i'$前 n 个网格,$\boldsymbol{S}_{j(m)}'$为 $\boldsymbol{S}_j'$前 m 个网格,将 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')$状态定义为 $\boldsymbol{S}_{i(n)}'$与 $\boldsymbol{S}_{j(m)}'$间匹配网格相似度总和的最大值。当 $n=0$ 或 $m=0$ 时,$s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')=0$;若 n、m 均不为 0,满足无后效性前提下,$s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')$与 $\boldsymbol{S}_i'$前 $n-1$ 个网格与 $\boldsymbol{S}_j'$前 $m-1$ 个网格的相似度 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n-1)}',\boldsymbol{S}_{j(m-1)}')$、$\boldsymbol{S}_i'$前 $n-1$ 个网格与 $\boldsymbol{S}_j'$前 m 个网格的相似度 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n-1)}',\boldsymbol{S}_{j(m)}')$、$\boldsymbol{S}_i'$前 n 个网格与 $\boldsymbol{S}_j'$ 前 $m-1$ 个网格的相似度 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m-1)}')$有关,分别对应 M_1、M_2 及 M_3 3 种匹配模式,$M_1=s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n-1)}',\boldsymbol{S}_{j(m-1)}')+s_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})$,$M_2=s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n-1)}',\boldsymbol{S}_{j(m)}')$,$M_3=s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m-1)}')$,即:

$$s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')=\begin{cases}0, m=0 \text{ 或 } n=0\\ \max(M_1,M_2,M_3), m、n\neq 0\end{cases}^{\circ}$$

对长度分别为 N'与 M'的序列 $\boldsymbol{S}_i'$与 $\boldsymbol{S}_j'$,建立行列数分别为 $N'+1$ 与 $M'+1$ 的矩阵 $\boldsymbol{D}_{(N'+1,M'+1)}$,通过 $\boldsymbol{D}_{(N'+1,M'+1)}$说明动态规划求解最长公共序列长度及内容的过程,如图 1 所示。矩阵 $\boldsymbol{D}_{(n,m)}$保存当前 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')$及匹配的网格序列长度 $l_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')$,$M_1$、$M_2$ 及 M_3 对应 $\boldsymbol{D}_{(n,m)}$由 $\boldsymbol{D}_{(n-1,m-1)}$、$\boldsymbol{D}_{(n-1,m)}$及 $\boldsymbol{D}_{(n,m-1)}$分别沿对角线、向下及向右方向传递得到。以 $\boldsymbol{S}_{i(3)}'$(长度为 3 的网格序列 $\boldsymbol{S}_i'$)与 $\boldsymbol{S}_{j(3)}'$(长度为 3 的网格序列 $\boldsymbol{S}_j'$)为例,设置 $\delta=2$,分析 $\boldsymbol{S}_{i(3)}'$与 $\boldsymbol{S}_{j(2)}'$匹配至 $\boldsymbol{S}_{i(3)}'$与 $\boldsymbol{S}_{j(3)}'$匹配的转移过程。对 $\boldsymbol{S}_{i(3)}'$与 $\boldsymbol{S}_{j(2)}'$,只有在 M_1 下 $\boldsymbol{G}_3^{i'}$ 与 $\boldsymbol{G}_2^{j'}$ 相匹配,$s_{\mathrm{LCS}}(\boldsymbol{S}_{i(2)}',\boldsymbol{S}_{j(1)}')+S_G'(G_3^{i'},G_2^{j'})$取得最大值,并沿对角线由 $\boldsymbol{D}_{(2,1)}$传递至 $\boldsymbol{D}_{(3,2)}$。对 $\boldsymbol{S}_{i(3)}'$与 $\boldsymbol{S}_{j(3)}'$,$s_{\mathrm{LCS}}(\boldsymbol{S}_{i(3)}',\boldsymbol{S}_{j(2)}')$并未继续传递,$\boldsymbol{D}_{(3,3)}$中的 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(3)}',\boldsymbol{S}_{j(3)}')$由 $\boldsymbol{D}_{(2,2)}$传递得到,说明该方法从全局确定网格间的最佳匹配。$\boldsymbol{D}_{(m,n)}$中的 $l_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')$遵循 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(n)}',\boldsymbol{S}_{j(m)}')$确定的传递方向,但当 $\boldsymbol{D}_{(n-1,m-1)}$沿对角线向 $\boldsymbol{D}_{(m,n)}$传递时,$l_{\mathrm{LCS}}(\boldsymbol{S}_{i(n-1)}',\boldsymbol{S}_{j(m-1)}')$是否加 1 还需检验 $s_G(\boldsymbol{G}_n^{i'},\boldsymbol{G}_m^{j'})$是否大于 0,若大于 0,说明 $\boldsymbol{G}_n^{i'}$ 与 $\boldsymbol{G}_m^{j'}$ 真正匹配且最长公共序列长度加 1。当 n 或 m 从 0 变化至 N'或 M'时,$\boldsymbol{D}_{(M',N')}$的 $s_{\mathrm{LCS}}(\boldsymbol{S}_{i(N')}',\boldsymbol{S}_{j(M')}')$及 $l_{\mathrm{LCS}}(\boldsymbol{S}_{i(N')}',\boldsymbol{S}_{j(M')}')$分别表示序列间网格相似度总和最长公共序列长度的最大值。若获取最长公共序列中的匹配网格对,需借助矩阵 $\boldsymbol{D}_{(M'+1,N'+1)}$反向搜索,但耗时长。

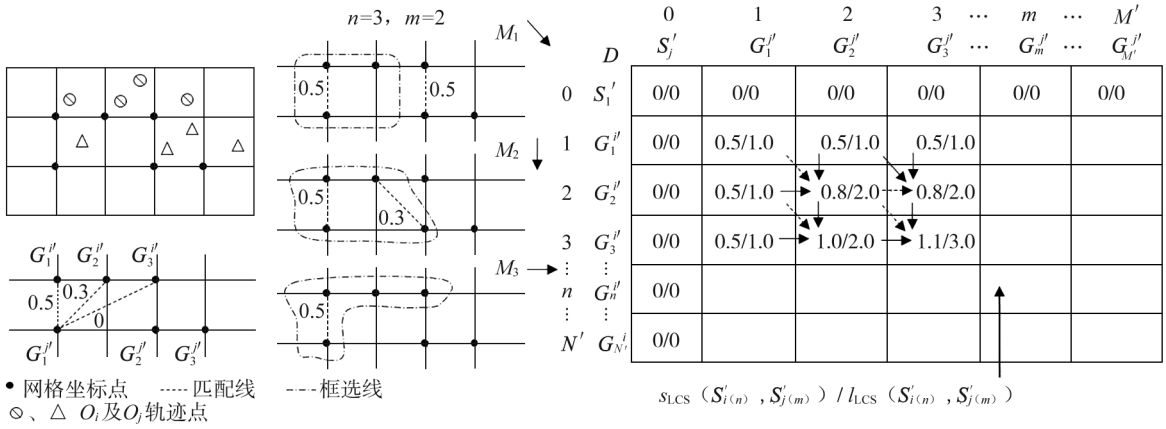

图 1 动态规划求解最长公共序列长度及内容

1.4 序列整体相似度及 GS-DBSCAN

车辆的出行不仅受路网约束,还与起讫点的用地相关,故轨迹-网格序列的相似性度量应充分考虑边界与内部的特征。对热点小区间的任意 2 个网格序列 $\boldsymbol{S}_i'$与 $\boldsymbol{S}_j'$,首先结合 $G_{x,y}$ 与热点小区 z_a^{h} 的隶属关系,将边界网格分别替换成对应的热点小区 $\boldsymbol{S}_i'^{(\mathrm{od})}=\{z_a^{\mathrm{h}(i)},\boldsymbol{G}_2^{i'},\cdots,\boldsymbol{G}_{N'-1}^{i'},z_{a'}^{\mathrm{h}(i)}\}$、$\boldsymbol{S}_j'^{(\mathrm{od})}=\{z_a^{\mathrm{h}(j)},\boldsymbol{G}_2^{j'},\cdots,\boldsymbol{G}_{M'-1}^{j'},z_{a'}^{\mathrm{h}(j)}\}$,根据移动序列 $\boldsymbol{S}_i''=\{\boldsymbol{G}_2^{i'},\cdots,\boldsymbol{G}_{N'-1}^{i'}\}$与 $\boldsymbol{S}_j''=\{\boldsymbol{G}_2^{j'},\cdots,\boldsymbol{G}_{M'-1}^{j'}\}$计算序列相似度 $s_{\mathrm{Seq}}(\boldsymbol{S}'^{(\mathrm{od})}_i,\boldsymbol{S}'^{(\mathrm{od})}_j)$,其包含 2 部分:1)起讫热点小区间的比较,若存在不同,则惩罚因子 $p_{\mathrm{f}}=-1$,保证相似序列小于等于 0;2)$\boldsymbol{S}_i''$与 $\boldsymbol{S}_j''$比较,以序列中较长的 1 个为基准,计算最大公共子序列与其长度之比,计算公式为:

$$s_{\mathrm{Seq}}(\boldsymbol{S}'^{(\mathrm{od})}_i,\boldsymbol{S}'^{(\mathrm{od})}_j)=\frac{l_{\mathrm{LCS}}(\boldsymbol{S}_i'',\boldsymbol{S}_j'')}{\max(l_{\mathrm{Seq}}(\boldsymbol{S}_i''),l_{\mathrm{Seq}}(\boldsymbol{S}_j''))}+p_{\mathrm{f}},$$

$$p_{\mathrm{f}}=\begin{cases}0, z_a^{\mathrm{h}(i)}=z_a^{\mathrm{h}(j)} \text{ 且 } z_{a'}^{\mathrm{h}(i)}=z_{a'}^{\mathrm{h}(j)}\\ -1, z_a^{\mathrm{h}(i)}\neq z_a^{\mathrm{h}(i)} \text{ 或 } z_{a'}^{\mathrm{h}(i)}\neq z_{a'}^{\mathrm{h}(i)}\end{cases},$$

式中:$l_{\mathrm{LCS}}(\boldsymbol{S}_i'',\boldsymbol{S}_j'')$为 $\boldsymbol{S}_i''$与 $\boldsymbol{S}_j''$间匹配的最长网格序列长度,$l_{\mathrm{Seq}}(\boldsymbol{S}_i'')$、$l_{\mathrm{Seq}}(\boldsymbol{S}_j'')$分别为 $\boldsymbol{S}_i''$与 $\boldsymbol{S}_j''$的网格序列长度。

耦合路径轨迹因起讫点或局部路段不同分为分离、汇入及交叉 3 种情况,如图 2 所示。对同时与 $\boldsymbol{S}_3$ 相耦合且长度差异较大的 $\boldsymbol{S}_4$ 及 $\boldsymbol{S}_5$,以较长序列为基准的度量可正确表征 $\boldsymbol{S}_4$ 与 $\boldsymbol{S}_5$ 分别对 $\boldsymbol{S}_3$ 的相似性,即受长度差异影响较小。对长度相近的路径,如处于同一对小区的 $\boldsymbol{S}_1$ 与 $\boldsymbol{S}_2$,以较短或较长序列为基准均

可,前者比后者更接近 1;但对不同小区的 $\boldsymbol{S}_3$ 与 $\boldsymbol{S}_6$,2 种度量方法均存在局限性,此时需增加起讫边界小区不同的限制条件进行区分。

计算 $\boldsymbol{S}_i'^{(\mathrm{od})}$ 与 $\boldsymbol{S}_j'^{(\mathrm{od})}$ 间的相似度后,需转化为距离度量才可进行 GS-DBSCAN 聚类,公式为:

$$d_{\mathrm{Seq}}(\boldsymbol{S}_i'^{(\mathrm{od})},\boldsymbol{S}_j'^{(\mathrm{od})})=1-s_{\mathrm{Seq}}(\boldsymbol{S}_i'^{(\mathrm{od})},\boldsymbol{S}_j'^{(\mathrm{od})})。$$

GS-DBSCAN 的伪代码与 HG-DBSCAN 类似,研究对象变成网格序列,不再赘述,得到热点路径集合 $\boldsymbol{R}=\{\boldsymbol{r}_1^{\mathrm{h}},\boldsymbol{r}_2^{\mathrm{h}},\cdots,\boldsymbol{r}_U^{\mathrm{h}}\}$,其中 U 为热点路径数。

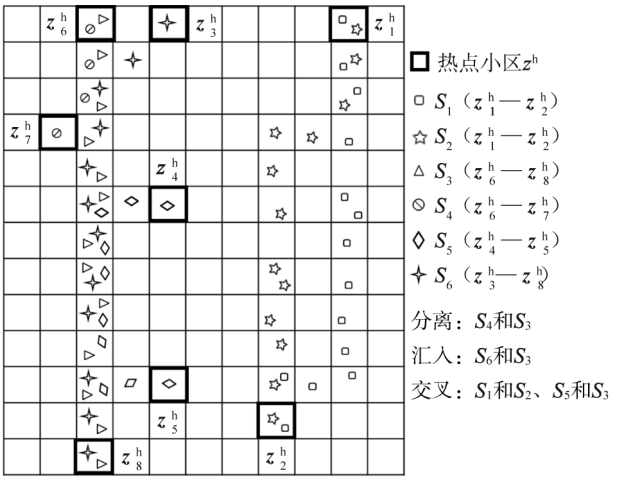

图 2　轨迹-网格序列

2　案例分析

以 2017-01-13 青岛市市南区的网约车订单数据为案例验证方法的有效性。区域覆盖的经度区间为(120.375 6°,120.400 0°),纬度区间为(36.063 1°,36.075 9°),面积约为 4.79 km^2。案例共采集 33 538 条订单轨迹数据,每条订单数据包含脱敏处理后的订单编号及轨迹点列表 2 个字段,如订单编号 3e7a3f2d231 的轨迹点列表为{(1 484 236 791 s　120.398 56°　36.007 014°),(1 484 236 794 s　120.398 57°　36.070 17°),(1 484 236 977 s　120.398 55°　36.070 18°)},轨迹点列表中包含 Unix 时间戳及 WGS84 坐标系下的经、纬度。

2.1　热点路径识别结果

先聚类识别热点小区,再将热点小区作为网格边界,增加约束条件,聚类识别热点路径。取网格尺寸为 30 m×30 m,小于次干路及以上等级道路间的最小平行距离。假设车辆在网格内的位置均匀分布且速度为 36~55 km/h,轨迹点平均采样时间为 3 s,则车辆每次移动的网格数为 1.0~1.5。首先建立参数不同间隔取值下的交叉组合,通过试算,设置 $\lambda=30$,$E_{\mathrm{ps}}=30$ m,$N_{\mathrm{um}}=2$。根据所选组合进行 HG-DBSCAN 聚类,聚类后的簇总数为 82,根据簇中包含的出行起讫点数降序排列选取前 13 个簇作为热点小区,如图 3 所示。由图 3 可知:该区域内的热点小区多分布在研究范围内道路的端点处,表示联系研究范围外热点区域的主要途经点。z_{12}^{h} 和 z_{13}^{h} 位于研究范围内道路中间,结合该点的用地可知,周边商业综合体华润万象城成为农历新年前最后 1 个工作日的出行热点。从热点小区间出行期望线可知 z_{12}^{h} 与 z_1^{h} 间的联系最强,说明华润万象城是研究范围内有强吸引力的出行热点,同时对研究范围外从 z_1^{h} 方向的居民有较强吸引力。

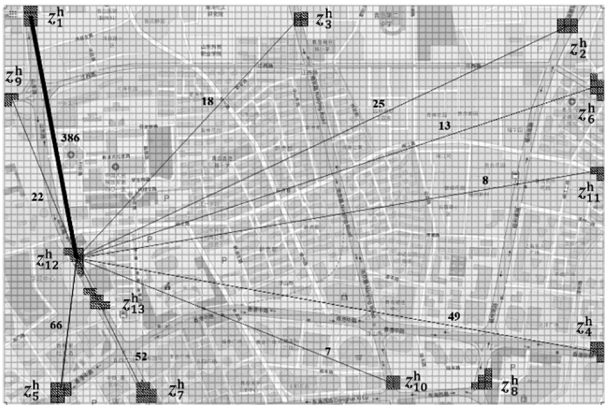

图 3　热点小区分布与轨迹期望线

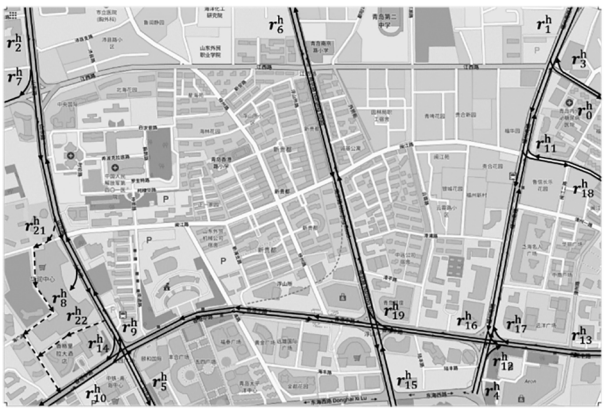

图 4　热点小区间的热点路径集合

选取早、晚高峰时段(07:00—10:00,18:00—21:00)数据识别热点路径。通过试算,设置 $\delta=2$,$E_{\mathrm{ps}}=0.2$ m,$N_{\mathrm{um}}=3$。对 13 个热点小区间的轨迹-网格序列进行 GS-DBSCAN 聚类,选取其中序列数较多的前 20 个簇作为热点路径集合,如图 4 所示。由图 4 结合热点小区分布可知:热点路径主要分布在主干道,受相交道路或道路一侧邻近热点小区的

吸引,又存在若干条转入、转出的高相似路径。

2.2 基于 LCS 的序列相似性度量算法对比与网格敏感性分析

保持 $\delta=2$,$E_{ps}=0.2$ m,$N_{um}=3$ 不变,对比分析不同计算方法下的聚类效果。方法 1:只考虑内部相似性且以对比序列中较短序列为基准。方法 2:只考虑内部相似性且以对比序列中较长序列为基准。方法 3:本文建立同时考虑边界与内部相似性的 GS-DBSCAN 聚类且以对比序列中较长序列为基准。

以 $\boldsymbol{z}_1^h$ 至 $\boldsymbol{z}_5^h$、$\boldsymbol{z}_7^h$、$\boldsymbol{z}_9^h$ 的网格序列集为例分别进行 3 种方法下的轨迹聚类,得到的簇如图 5 所示。由图 5 可知:方法 3 中的 5 个轨迹簇 $\boldsymbol{C}_0$、$\boldsymbol{C}_1$、$\boldsymbol{C}_2$、$\boldsymbol{C}_3$、$\boldsymbol{C}_4$ 对应区域中的路径 $\boldsymbol{r}_2^h$、$\boldsymbol{r}_7^h$、$\boldsymbol{r}_{14}^h$、$\boldsymbol{r}_{21}^h$、r_{22}^h;方法 1 中以较短序列为基准,$\boldsymbol{C}_1$ 与其他簇轨迹的相似度均较高,在密度可达的传递作用下将其归于 1 类;方法 2 中虽以较长序列为基准,但 $\boldsymbol{C}_0$ 与 $\boldsymbol{C}_2$ 序列长度相近且重叠度较高,导致 $\boldsymbol{r}_2^h$ 与 $\boldsymbol{r}_{14}^h$ 无法区分;方法 3 中 $\boldsymbol{C}_0$、$\boldsymbol{C}_1$、$\boldsymbol{C}_2$、$\boldsymbol{C}_3$、$\boldsymbol{C}_4$ 可较好地区分 $\boldsymbol{r}_2^h$、$\boldsymbol{r}_7^h$、$\boldsymbol{r}_{14}^h$、$\boldsymbol{r}_{21}^h$ 与 $\boldsymbol{r}_{22}^h$ 的轨迹,说明本文考虑边界与内部相似性,且以对比序列中较长序列为基准的方法合理。

a)方法 1 b)方法 2 c)方法 3

图 5 不同方法下的轨迹聚类簇

分析 GS-DBSCAND 方法对网格尺寸的敏感性,在 $k=30$ m 的基础上分别缩小、扩大 1 倍,相应地调整 δ 进行聚类,得到不同路径下的正确轨迹数,如表 1 所示。

表 1 不同 k、δ 下各路径识别的正确轨迹数

k/m	δ	正确轨迹数					聚类时间/s
		$\boldsymbol{r}_2^h$	$\boldsymbol{r}_7^h$	$\boldsymbol{r}_{14}^h$	$\boldsymbol{r}_{21}^h$	$\boldsymbol{r}_{22}^h$	
15	4	149	100	51	3	4	440
30	2	150	100	51	5	4	307
60	1	149	100	49	4	4	118

由表 1 可知:聚类时间随 k 的减小而增大,当 $\delta=4$,$k=15$ m 时,聚类时间约为 440 s,仅比 $\delta=2$,$k=30$ m 时的聚类时间增大 43.32%,GS-DBSCAN 方法的轨迹聚类运算效率较高;不同 k 和 δ 下,$\boldsymbol{r}_2^h$、$\boldsymbol{r}_7^h$、$\boldsymbol{r}_{14}^h$、$\boldsymbol{r}_{21}^h$ 及 $\boldsymbol{r}_{22}^h$ 的聚类结果整体偏差小于 2%,受 k 和 δ 的影响较小。

3 结束语

为准确识别热点路径,本文提出细分边界与内部两部分度量序列间相似性的轨迹数据热点路径识别优化方法,以区域网格化为基础,将车辆轨迹映射为移动网格序列,通过 HG-DBSCAN 算法识别热点区域,再将热点区域作为网格边界,增加约束条件,通过 GS-DBSCAN 算法识别热点路径。以青岛市市南区的网约车订单数据为案例进行热点路径识别,验证方法的有效性。结果表明:本文提出的移动网格序列研究方法能准确识别热点小区,正确区分多出行起讫点分布下长度差异较大的分离、汇合与交叉耦合热点路径,且聚类运算效率较高,可为城市范围内交通组织设计、交通拥堵治理、智慧交通出行路径规划等城市交通规划设计与管理提供参考依据。后续将结合实际工作,采集私家车、共享单车等更丰富的轨迹数据,扩大案例范围,开展更全面的研究。

参考文献:

[1] WANG Y L, QIN K, CHEN Y X, et al. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data[J]. ISPRS International Journal of Geo-Information, 2018,7(1):25-45.

[2] ZHANG D Q, LI N, ZHOU Z H, et al. Ibat: detectiong anomalous taxi trajectories from GPS traces[C]//Proceedings of 13th International Conference on Ubiquitous Computing. Beijing, China: Ubicomp,2011:17-21.

[3] LIU C K, QIN K, KANG C G. Exploring time-dependent traffic congestion patterns from taxi trajectory data[C]//2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM). Fuzhou, China: IEEE,2015:39-45.

[4] 陈振华. 基于轨迹数据的城市交通拥塞传播模式挖掘[D]. 吉林:吉林大学,2019.

[5] LIN S, SCHUTTER B D, XI Y G, et al. Efficient network-wide model-based predictive contral for urban traffic networks [J]. Transportation Research Part C: Emerging Technologies, 2012,24:122-140.

[6] MAZIMPAKA J D, TIMPF S. Trajectory data mining: a review of methods and applications[J]. Journal of Spatial Information Science,2016,13:61-99.

[7] 李颖,赵莉,赵祥模,等. 基于大货车 GPS 数据的轨迹相似性度量有效性研究[J]. 中国公路学报,2020,33(2):146-157.

[8] KIM J, MAHMASSANI H S. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories[J]. Transportation Research Part C: Emerging Technologies, 2015,7(10):164-184.

[9] 冯琦森. 基于出租车轨迹的居民出行热点路径和区域挖掘[D]. 重庆:重庆大学,2016.

[10] 赵欣. 基于时空约束的城市热点区域与热点路径挖掘[D]. 重庆:重庆大学,2017.

[11] KANG H Y, KIM J S, LI K J. Similarity measures for trajectory of moving objects in cellular space[C]//Proceedings of the 2009 ACM Symposium on Applied Computing. Honolulu, Hawaii, USA: Spatial Information Society,2009:9-12.

[12] 吴俊伟,朱云龙,库涛,等. 基于网格聚类的热点路径探测[J]. 吉林大学学报(工学版),2015,45(1):274-282.

[13] 王超. 基于轨迹聚类的频繁模式挖掘方法[D]. 杭州:浙江大学,2021.

[14] LEE J G, HAN J W, WHANG K Y. Trajectory clustering: a partition-and-group framework[C]//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing, China: SIGMOD,2007:593-604.

[15] KANG H Y, KIM J S, LI K J. Similarity measures for trajectory of moving objects in cellular space[C]//Proceedings of the 2009 ACM Symposium on Applied Computing. Honolulu, Hawaii, USA: ACM Symposium on Applied Computing, 2009:1325-1330.

[16] YANG H, SASAKI T, IIDA Y, et al. Estimation of origin-destination matrices from link traffic counts on congested networks[J]. Transportation Research Part B: Methodological, 1992,26(6):417-434.

[17] ZHOU X S , MAHMASSANI H S. Dynamic origin-destination demand estimation using automatic vehicle identification data [J]. IEEE Transactions on Intelligent Transportation Systems, 2006,7(1):105-114.

[18] OU J S, LU J W, XIA J X, et al. Learn, assign, and search: real-time estimation of dynamic origin-destination flows using machine learning algorithms[J]. IEEE Access, 2019, 7:26967-26983.

[19] TANG K S, CAO Y M, CHEN C, et al. Dynamic origin-destination flow estimation using automatic vehicle identification data: a 3D convolutional neural network approach[J]. Computer-Aided Civil and Infrastructure Engineering, 2021,36(1):30-46.

[20] CAO P, MIWA T, YAMAMOTO T, et al. Bilevel generalized least squares estimation of dynamic origin-destination matrix for urban network with probe vehicle data[J]. Transportation Research Record: Journal of the Transportation Research Board, 2013, 2333:66-73.

[21] CAO P, MIWA T, YAMAMOTO T, et al. Estimation of dynamic link flows and origin-destination matrices from lower polling frequency probe vehicle data[J]. Journal of the Eastern Asia Society for Transportation Studies, 2013, 10:762-775.

[22] HUANG T, MA Y T, QIN Z T, et al. Origin-destination flow prediction with vehicle trajectory data and semi-supervised recurrent neural network[C]//2019 IEEE International Conference on Big Data. Los Angeles, CA, USA:IEEE, 2019:1450-1459.

[23] GODAU M, ALT H. Computing the Fréchet distance between two polygonal curves[J]. International Journal of

Computational Geometry & Applications, 1995,5(1):75-91.

Regional hotspot path identification based on grid clustering optimization

WENG Xuyan, ZHENG Shuni

Zhejiang Scientific Research Institute of Transport, Hangzhou 310023, *China*

Abstract: To solve the problem that the trajectory clustering method is difficult to accurately identify high-similarity hotspot paths, a hotspot path identification method that can distinguish between start and end points or local sections is proposed. The travel trajectory is mapped and compressed into a moving mesh sequence, and the spatial similarity measurement between sequences is distinguished from the boundary and the interior, integrated and transformed into distance, and spatial clustering based on grid sequencedensity-based spatial clustering of applications with noise (GS-DBSCAN) is performed. Taking the trajectory data of some taxis in Shinan District, Qingdao as an example, the clustering method that only considers internal similarity and is based on the shorter and longer sequences in the comparison sequence is verified. The results show that the GS-DBSCAN algorithm, which considers both boundary and internal similarity and is based on longer sequences, can correctly distinguish the separation, convergence, and cross-coupling hotspot paths with large length differences under the distribution of multiple travel start and end points. The influence of variable differences such as path length and grid size is less than 2%, and the clustering operation efficiency is high.

Keywords: hotspot path; boundary; interior; trajectory clustering

(责任编辑:赵玉真)

(上接第 88 页)

index of cold chain logistics demand for aquatic pre-made dishes is constructed from four aspects: regional economic development level, market supply and demand level, transportation level, and cold chain technology level. The main factors affecting the change of cold chain logistics demand for aquatic pre-made dishes are studied by gray correlation analysis method. The first-order unitary differential equation GM(1,1) in the gray model (GM) and long short-term memory (LSTM) neural network are used to compare and analyze the cold chain logistics demand for aquatic pre-made dishes in Guangdong Province from 2015 to 2021. The results showed that the main influencing factors affecting the development of cold chain logistics demand for aquatic pre-made dishes in Guangdong Province are cargo turnover and cold chain refrigeration level. The average relative errors of GM(1,1) and LSTM neural network are 2.68% and 0.22%, respectively, and the prediction accuracy of the latter is significantly better than that of the former. Using LSTM neural network to predict the cold chain logistics demand forcargo in Guangdong Province from 2022 to 2024, the demand for cargo in Guangdong Province is on the rise, and it is expected to reach 509.09 million tons in 2024. Guangdong Province should focus on cold chain infrastructure building, ensure stable temperature of aquatic pre-made dishes during storage and transportation, strengthen food supervision of aquatic pre-made dishes, ensure food quality and safety, and continuously promote the development of the aquatic pre-made dishes cold chain industry.

Keywords: aquatic pre-made dish; cold chain logistics demand; GM (1,1); LSTM neural network; grey relational analysis

(责任编辑:赵玉真)