作品编号：TJJM20250311044071

2025年（第十一届）全国大学生统计建模大赛

参 赛 作 品

参赛学校： 同济大学

论文题目：上海市出租车动态区域选择模型-深度学习与用户反馈融合研究

参赛队员： 刘浩田 王晓涵 黄家祺

上海市出租车动态区域选择模型-深度学习与用户反馈融合研究

# 摘要

“十四五”现代综合交通运输体系发展规划指出，需依托大数据、人工智能等技术推动交通治理模式创新，构建高效、绿色、安全的城市交通体系。在此背景下，设计一套由大数据人工智能推动的交通模型尤有重要的意义。本研究以上海市出租车为研究对象，融合GPS数据和用户反馈数据，使用聚类、网络爬虫、深度学习等算法，建立模型帮助出租车司机进行路线决策。

过往的研究主要集中在利用历史交通数据进行流量预测，尚未融合用户反馈信息。相比之下，本研究通过整合真实用户投诉数据，构建风险矩阵，并与交通流量密度和上车点热度等特征融合，提供更全面的交通决策支持，且本研究实现GPS轨迹（热度、密度）、用户投诉（风险矩阵）及距离成本的特征权重自适应分配，克服传统固定权重方法的局限性。

本文研究主要分为三个部分。第一部分，本研究采用香港科技大学的290万条上海市出租车GPS数据集，经数据清洗与空间划分处理，基于K-means聚类算法，实现空间划分，对轨迹点进行聚类建模；第二部分，本研究采用python爬虫逆向获取黑猫投诉信息，通过文本分块、地理映射等处理得到风险矩阵；第三部分，本研究采用LSTM对交通流量密度与上车点热度进行时序预测，接着，引入可学习的加权融合机制，基于Softmax函数实现多源特征的自适应权重分配，最后，程序提出递归优化决策算法，结合剪枝阈值机制，规划未来多步行动路径，可为每辆车提供未来5小时的最优路径，并针对研究成果，为企业与部门提出建议。

关键词： 深度学习；交通决策；用户反馈；可学习加权融合；LSTM

目 录

[摘要 I](#_Toc196133429)

[一、 引言 1](#_Toc196133430)

[(一) 研究背景与意义 1](#_Toc196133431)

[(二) 研究方法 2](#_Toc196133432)

[(三) 创新点 3](#_Toc196133433)

[二、 文献综述 4](#_Toc196133434)

[三、 数据选择与处理 5](#_Toc196133435)

[(一) 数据来源 5](#_Toc196133436)

[(二) 数据处理 6](#_Toc196133437)

[四、 模型介绍 9](#_Toc196133438)

[(一) K-means聚类模型 9](#_Toc196133439)

[(二) LSTM模型 9](#_Toc196133440)

[(三) 可学习加权融合技术 11](#_Toc196133441)

[五、 模型研究实践 13](#_Toc196133442)

[(一) 基于行政区划的K-means聚类模型 13](#_Toc196133443)

[(二) 爬虫获取黑猫投诉内容 16](#_Toc196133444)

[(三) 构建LSTM模型并预测密度与热度 17](#_Toc196133445)

[(四) 可学习加权融合模型的构建训练 23](#_Toc196133446)

[(五) 动态区域推荐 26](#_Toc196133447)

[六、 结论与建议 28](#_Toc196133448)

[（一）结论 28](#_Toc196133449)

[（二）建议 29](#_Toc196133450)

**表格与插图清单**

[图 1 本项目技术路线 2](#_Toc196133403)

[图 2 爬虫数据处理工作流程 6](#_Toc196133404)

[图 3 风险集群特征中心 8](#_Toc196133405)

[图 4 数据处理结果 8](#_Toc196133406)

[图 5 聚类模型流程 13](#_Toc196133407)

[图 6 聚类区域地理位置图 14](#_Toc196133408)

[图 7 各区聚类区域可视化 15](#_Toc196133409)

[图 8 距离矩阵热度图 16](#_Toc196133410)

[图 9 黑猫投诉内容爬取流程图 16](#_Toc196133411)

[图 10 数据邻接矩阵 18](#_Toc196133412)

[图 11 流量变化数据图 18](#_Toc196133413)

[图 12 模型训练和验证指标变化曲线 20](#_Toc196133414)

[图 13 LSTM流量数据预测 21](#_Toc196133415)

[图 14 上车点热度预测数据预处理流程图 21](#_Toc196133416)

[图 15 LSTM热点数据预测流程图 23](#_Toc196133417)

[图 16 区域-特征热力图 24](#_Toc196133418)

[图 17 部分区域特征平行坐标图 24](#_Toc196133419)

[图 18 特征相关性散点图矩阵 24](#_Toc196133420)

[图 19 模型权重训练演变 26](#_Toc196133421)

[图 20 部分模拟路线规划 28](#_Toc196133422)

[表 1 出租车GPS数据集字段介绍 5](#_Toc196133423)

[表 2 黑猫投诉网站爬取内容示例 6](#_Toc196133424)

[表 3 K-means聚类模型变量介绍 9](#_Toc196133425)

[表 4聚类模型参数介绍 14](#_Toc196133426)

上海市出租车动态区域选择模型-深度学习与用户反馈融合研究

# 引言

## 研究背景与意义

### 研究背景

随着中国城市化进程的加速和智慧城市建设的深入推进，交通系统作为城市运行的“血管网络”，其智能化升级已成为国家战略发展的重要方向。《“十四五”现代综合交通运输体系发展规划》明确提出，需依托大数据、人工智能等技术推动交通治理模式创新，构建高效、绿色、安全的城市交通体系。然而，传统出租车调度系统多依赖静态经验规则或单一数据源，难以动态响应实时供需变化，亟需技术赋能实现精细化运营。

近年来，深度学习技术在时空预测领域展现了显著优势。长短期记忆网络（LSTM）等模型能够有效捕捉交通数据的时序依赖性和空间异质性，为动态区域推荐提供了新思路。同时，用户反馈数据蕴含丰富的需求痛点和区域风险特征，却长期未被充分挖掘。现有研究多聚焦于单一数据源建模，缺乏多源异构数据的深度融合，导致模型决策维度单一、适应性不足。

上海市出租车行业正面临重大挑战，与网约车竞争加剧，传统出租车需通过智能化提升服务竞争力。结合多种数据源与深度学习方案的模型，对出租车行业有重大意义。

### 研究目的与意义

本文构建了能够基于实时和预测分析为出租车司机推荐最佳区域的动态模型。采用长短期记忆LSTM网络，通过捕捉交通数据的时序特征，准确预测交通模式和热点区域。同时，将用户反馈数据融入模型的决策过程，使其更贴近用户实际需求。最后，通过可学习加权机制融合多种异构数据源，包括GPS轨迹、用户投诉等等，以提升模型的适应性和精确性。

本文提出的动态、数据驱动的出租车区域选择方案，能够有效提升上海出租车行业的运营效率，减少无效行驶里程，有助于缓解城市交通拥堵，降低碳排放;其次，本文将用户反馈作为核心要素纳入模型设计，开辟了出租车调度研究的新方向。通过分析用户投诉数据，模型能够识别服务中的痛点并加以改进，从而提升乘客满意度和行业竞争力,具有重要的实践价值。

## 研究方法

本文主要采用以下技术路线：

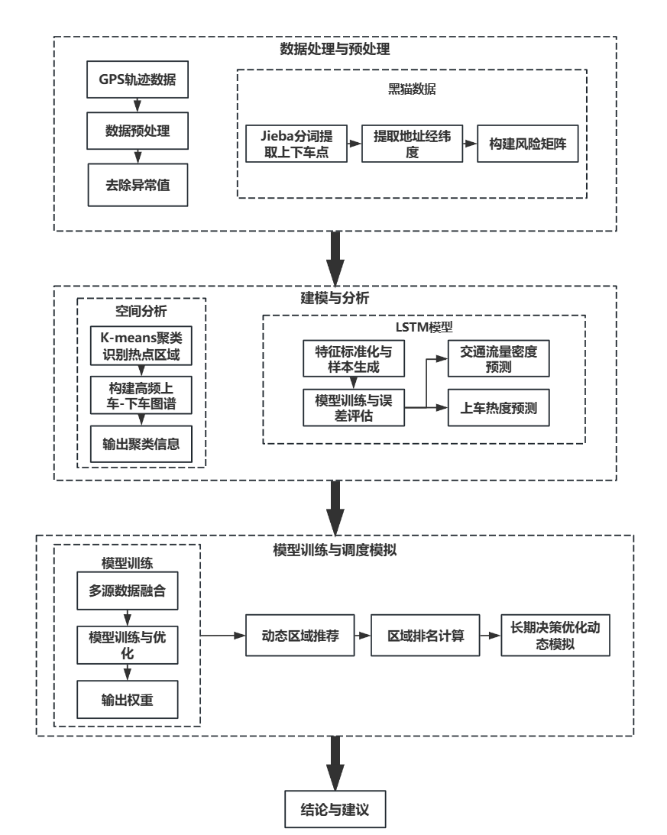


图 1 本项目技术路线

### 研究数据采集与预处理

首先，本文从香港科技大学公开发布的数据集中获取上海市出租车的GPS轨迹数据，该数据集包含约290万条记录，涵盖车辆编号、时间戳、经纬度、车速、航向角以及载客状态等关键信息。为了确保数据质量，对原始数据进行清理。

### 投诉数据结构化建模

本文研究采集的用户反馈数据通过构建文本爬虫从黑猫投诉平台抓取，为使非结构化数据转化为可用于分析的结构化数据，本文采用分词工具对投诉文本进行预处理，识别文本中的上车点和下车点信息，构建上车点与下车点的风险矩阵。

### 聚类分析与热点区域识别

在对GPS轨迹数据进行预处理后，本文运用K-means聚类算法对出租车的上下车点进行空间聚类分析，在上海市每个行政区内独立进行聚类分析，以确保聚类结果具有地理意义，计算每个聚类簇的中心点、影响半径等参数，为后续研究基础数据支持。

### 时序预测建模：LSTM网络

为了实现对出租车流量和上车点热度的动态预测，本文构建了基于长短期记忆网络LSTM的时序预测模型。模型可以成功预测每个区域在若干时间步内的流量和上车点热度，为出租车的动态调度提供决策依据。

### 多源数据融合：可学习加权**机制**

为实现多源异构数据的融合，本文引入了可学习加权融合技术。该技术通过神经网络自动学习不同数据源特征的重要性权重，实现多个视角对决策系统的联合优化，生成综合评分。基于该评分，本文构建动态区域推荐系统，为每辆出租车计算最优的目标区域，并通过递归优化决策算法实现长期决策规划。

## 创新点

### 融合用户满意度反馈与轨迹数据的决策系统设计

整合用户投诉数据与GPS轨迹数据，构建以用户满意度为核心的决策优化模型，突破传统单一数据源的局限。

### 引入LSTM深度学习网络解决预测问题

采用LSTM网络捕捉交通数据的时序特征，提升热点区域与流量预测精度，为动态推荐提供可靠支持。

### 可学习加权融合机制

利用可学习加权融合技术机制，实现多源特征的端到端加权融合，使模型具备自适应决策策略调优能力，通过深度学习构建全面科学的决策依据。

1. 动态区域推荐与长期决策优化

设计实时动态区域推荐模型，结合递归优化算法实现长期规划，超越传统短期决策框架。

# 文献综述

近年来，伴随城市智能交通系统的发展，基于出租车GPS轨迹的聚类分析与寻客路径优化研究逐步成为智慧出行领域的热点方向。相关研究大致经历了基于传统聚类方法的探索阶段、融合智能算法的方法扩展阶段，以及结合深度学习的模型革新阶段。

早期研究主要聚焦于对出租车轨迹的统计建模与聚类分析。Besse[5]等人（2016）率先提出了一种基于二维高斯混合模型的轨迹分布方法，为后续空间预测模型奠定了基础。随后，Bian[6] 等人（2018）对轨迹聚类技术进行了全面梳理，将相关算法归类为无监督、监督与半监督三类，构建了较为系统的理论框架,刘耀宗[1] 等人（2018）同年提出TR-OPTICS算法，提升了热点路径与动态交通流识别的准确性。

随着方法不断深入，研究者开始融合各种技术，以适应更复杂的场景。Gong[7]等人（2020）构建了一个集成时空聚类、贝叶斯概率计算与蒙特卡洛模拟的双层建模框架，表现出较高的行为模式识别能力。Rathore[8]等人（2020）基于新加坡 Grab 平台的大规模预订关系数据，提出了改进的共聚类算法 sco-iVAT，有效刻画了司机与乘客之间的空间关系。Liu [9]等人（2021）则进一步提出改进型模糊DBSCAN（TC-FDBSCAN）算法，并应用于深圳城市数据中验证其有效性。

在深度学习快速发展的新阶段，轨迹聚类建模逐渐引入神经网络方法。Yue[10]等人（2020）提出的 DETECT 模型利用深度嵌入机制实现了无监督轨迹聚类，具备较强的空间行为尺度适应能力。Liu[11]等人（2022）研究了不同空间粒度对网约车需求预测模型的影响，提出了结合卷积神经网络与LSTM网络的H-ConvLSTM模型。

党鹏飞[2]（2021）基于聚类分析与Voronoi区域划分，构建了两阶段寻客推荐模型，并结合ARIMA模型动态预测交通速度，实现对寻客路径的实时优化。熊敏[3]（2022）则提出了基于自适应网格密度的热点区域聚类算法，构建了改进的PS-MNL模型，显著优于传统Logit与路径重叠修正模型。

# 数据选择与处理

## 数据来源

### 上海市出租车GPS数据集

为实现对上海市出租车运行轨迹进行高精度的空间聚类分析与热点区域识别，本文选取了由香港科技大学智慧城市研究小组公开发布的上海市出租车GPS轨迹数据集作为研究的基础数据源。该数据集收录了290万条上海市范围内出租车车辆的轨迹信息，具有较强的代表性和研究价值。原始数据包含出租车交通行为信息，每条记录对应一次GPS上报，包含以下字段：

表 1 出租车GPS数据集字段介绍

|  |  |
| --- | --- |
| **字段** | **注释** |
| 车辆编号（car\_id） | 用于唯一标识不同的出租车辆 |
| 日期与时间（datetime） | 精确到秒的时间戳，记录该条轨迹点的采集时间 |
| 地理坐标  （longitude, latitude） | 以WGS-84坐标系表示的经纬度，  用于标定出租车的位置信息 |
| 车速（speed, km/h） | 表示车辆在该时刻的瞬时速度 |
| 航向角（azimuth） | 表示车辆行驶方向的角度信息 |
| 载客状态（occupied） | 二值字段，表征出租车在该时刻是否处于载客状态  （0表示空载，1表示载客） |

### 黑猫投诉数据

为了获取真实的用户体验，本文从黑猫投诉平台中抓取了关于“上海出租车”“上海出租车”等关键词相关的用户投诉数据。该类数据具有鲜明的语义特征和用户视角，其内容真实，为研究出租车行驶区间风险度提供了重要信息。

本文总共抓取到数千条数据，数据特点为来源真实可靠，信息结构完整，具有完备的投诉标题，公司，摘要，类型和诉求，该数据以文本内容为主，数据涵盖了各个平台，与不同时间段和地段的用户反馈，涉及多种出行问题，内容具有较强的地理指向性。

表 2 黑猫投诉网站爬取内容示例

|  |  |  |
| --- | --- | --- |
| 投诉编号 | 公司名称 | 投诉摘要 |
| 17381681652 | 滴滴出行客服 | 我在25年4月六号晚上10送客到上海南站附近后由于场站爆单平台派场站订单，群里说有运管有图 由于我刚来上海...被迫导致取消了两笔上海南站订单 接连申诉平台不给通过导致扣服务分… |
| 17381695544 | T3出行 | 本人为一名合规双证出租车司机….从顺港楼公交站到佛山宏悦酒店（顺德大良店），该订单为一个起步价订单，全程约3公里，当天为清明节假期，人流量和车流非常大，在接近拐弯去终点的掉头位置因为有大量的车在待转，路口堵塞，无法变线掉… |
| 17376852679 | 聚的出租 | 我是出租车司机，从上海市奉贤区海湾镇龙腾阁接单到上海市浦东新区龙泽大厦7号楼的订单，总里程67公里。并按照平台导航行驶，当时正赶上早高峰，路上严重堵车，当到达乘客目的地后，实际车费是324元，高速费是24元。乘客说怎么不和平台预估的不一样... |

由网站爬取示例看出，许多投诉内容都与地点有关，用户/司机可以有效的反映出此地区风险的特征，且此类户投诉对司机造成的打击较大，探究地点相关的风险指标对出租车决策的影响具有重大意义。

## 数据处理

### 投诉爬虫数据处理

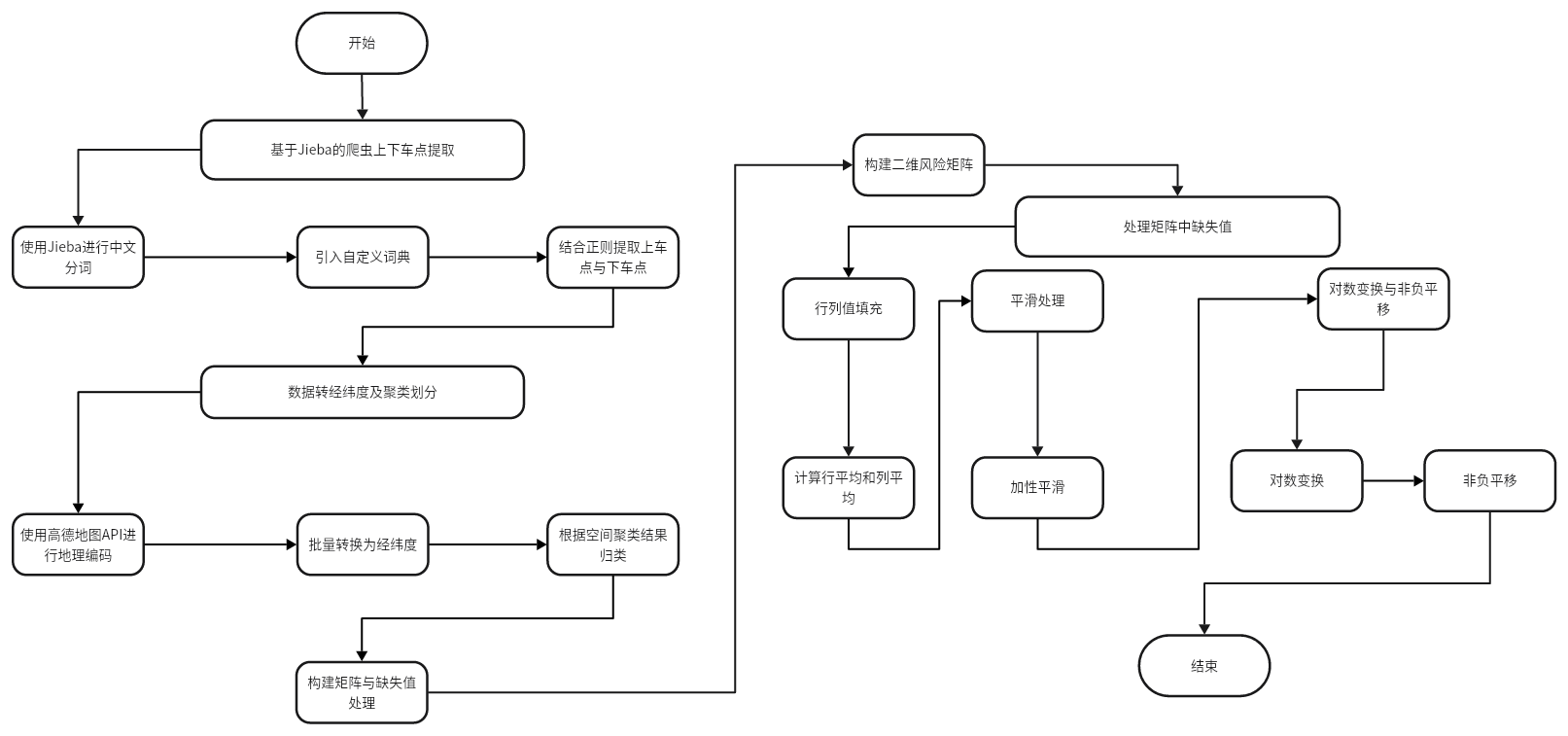


图 2 爬虫数据处理工作流程

#### 基于Jieba的爬虫上下车点提取

为了实现对文本中上车点与下车点的识别，本文引入中文分词工具 Jieba 进行预处理。Jieba是专门用于分词的算法，能够有效地对中文句子进行分词处理并提取词性信息用于文本分析。

由于上车点和下车点应具有明显的地理词性，所以在Jieba的基础词典上，进一步引入自定义词典，结合正则提取，提取具有空间指向性的结构。利用语句中“从…到”,“由…至”等词来分开上车点与下车点。

#### 数据转经纬度及聚类划分

所收集到的数据均为文本类型，无法直接参与空间分析。本文使用高德地图地理编码API服务。服务具有高精度、可扩展的地理编码处理，同时可以批量处理得到定位。

本文根据已经定义的空间聚类结果，将每个地址根据其地理位置映射到聚类类别中，以便区域统计分析。

#### 构建矩阵与缺失值处理操作

本文将上车点和下车点结合映射为二维风险矩阵图，其中表示上车点i到下车点j的风险值。

在矩阵中存在数据缺失，需要进行处理。首先，本文选用行列值填充是考虑了一般性的数值缺失处理方法。

对于每一个单元格通过行平均和列平均计算得出。其中是第i行的非零平均，代表第j列的非零平均,为平衡行列贡献的加权因子。本文使用平滑处理，加性平滑，去除极端小值带来的结果。为了避免估值过小或存在极端值导致后续转换失真，操作有效保持了矩阵整体的结构特征。同时，采用对数变换与非负平移，拉大差距，考虑到风险值应该全部为单向，故使用线性平移使得最小值为0。

对数变换与非负平移可以放大不同路径之间的风险差异，并且使估值更加敏感[18]。程序随后进行非负处理。

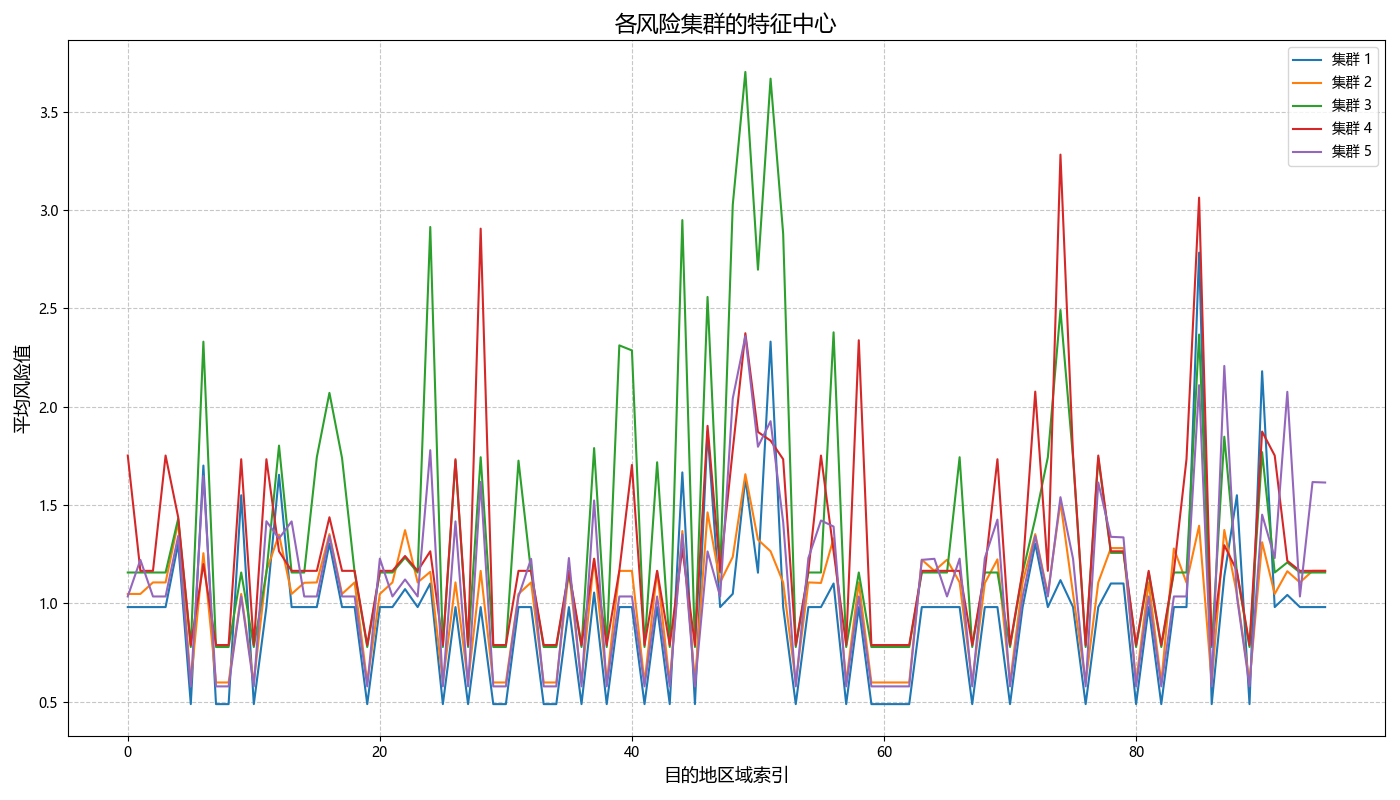


图 3 风险集群特征中心

### LSTM模型GPS预清理

首先，LSTM程序对原始数据的完整性进行校验。通过遍历数据集的每一行记录，检测关键字段的空值，直接删除包含缺失值的记录。

接着，程序设定速度有效范围为0至150 km/h，对计算结果进行严格筛选：速度计算结果不正确时，直接标记为无效数据；超过150 km/h的记录被视为物理不可行，执行删除操作。若有连续位置不变但速度非零的轨迹段，判定为停车未熄火状态，执行逻辑删除。

同时，程序验证经纬度值的物理有效性，删除超出合理范围的记录：经度范围：-180°至180°；纬度范围：-90°至90°。

最后，程序以车辆ID与时间戳为复合主键，识别并删除完全重复的记录

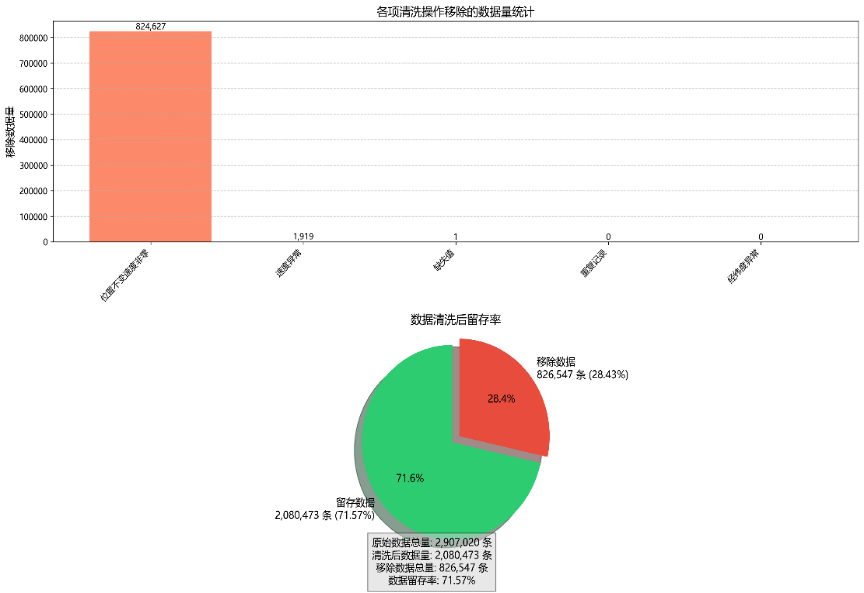


图 4 数据处理结果

程序从原始数据总量2,907,020条中移除了626,547条，占总数据的28.43%，表明程序具有较强的异常检测能力，能够有效识别并过滤大量低质量或无效数据。

# 模型介绍

## K-means聚类模型

K-means 是一种经典的非监督学习算法，广泛应用于空间聚类、图像分割及用户行为分析等领域。其基本思想是将数据点划分为 K 个簇，使得同一簇内的数据点尽可能相似，而不同簇之间的差异尽可能大。在地理空间场景中，这种算法特别适用于发现数据的空间分布结构与聚集特征[12]。

1. means 的核心目标是最小化簇内样本点到各自簇中心的平方距离[14]，具体的目标函数如下：

表 3 K-means聚类模型变量介绍

|  |  |
| --- | --- |
| **变量名称** | **含义解释** |
| K | 聚类的簇数 |
| Ci | 第i个簇的点集合 |
| ui | 第i个簇的中心（质心） |
|  | 样本点x与簇中心之间的欧几里得距离平方 |

## LSTM模型

本文使用了LSTM模型作为热点与流量的预测模型。传统循环神经网络(RNN)在处理序列数据时存在梯度消失和梯度爆炸问题。

数学上，考虑RNN的梯度计算：

其中连乘项会导致梯度不稳定。

而LSTM通过细胞状态这一创新设计建立了贯穿整个时间序列的信息高速公路，配合三重门控机制实现了对信息流的精确控制。

### 细胞状态

LSTM单元的核心是细胞状态，该状态通过线性操作在不同时间步之间传递信息，避免了非线性变换导致的梯度衰减。

细胞状态的更新机制：

遗忘门控制历史信息的保留比例，输入门调控新信息的加入程度。这种线性组合方式在数学上保证了梯度可以长期保持稳定，当且时，理论上可实现无限长时间的记忆保持。

### 三重门控机制

三重门控机制构成了LSTM的控制中枢。遗忘门

产生0到1之间的遗忘系数，决定从细胞状态中丢弃哪些历史信息。

输入门包含两个协同工作的部分：

门控单元：

和候选记忆：

前者决定更新强度，后者提供新的记忆内容。

输出门：

则通过：

的计算方式控制当前时间步的隐藏状态输出。

在前向传播过程中，LSTM单元依次执行门控计算和状态更新。首先接收前一隐藏状态和当前输入的拼接向量，并行计算三个门控信号和候选记忆。随后更新细胞状态并生成当前时间步的输出。这种结构使得网络可以自主决定何时记住新信息、何时遗忘旧信息以及何时输出当前状态。

反向传播阶段，LSTM展现出优异的梯度保持能力。由于细胞状态路径采用线性更新，其梯度，确保梯度可以长距离传播。门控单元使用的sigmoid函数导数范围在0-0.25之间，既防止了梯度爆炸，又避免了完全消失。。

作为循环神经网络的重要变体，LSTM凭借其卓越的时序建模能力，已在众多领域展现出强大的应用价值。近年来，LSTM因其优越的时间序列建模能力被广泛应用于智慧城市场景的交通预测中[15]。

## 可学习加权融合技术

可学习加权融合(Learnable Weighted Fusion）技术是一种通过神经网络自动学习特征融合权重的先进方法，其核心在于动态整合多源异构特征。与传统固定权重或启发式加权方法相比，该技术实现了三个关键突破：非线性权重生成机制通过可微分函数动态产生权重分布，端到端优化框架使融合权重与下游任务联合训练，自适应调整能力可根据输入特征自动平衡各特征的贡献度。

### 数学模型

设输入特征为，则权重生成过程为：

LearnableParams

它通过softmax函数进行归一化处理，得到概率分布。这个设计确保了三个重要性质：所有权重之和严格等于1，每个权重保持非负，且整个系统保持完全可微。它建立了可微分的权重调节机制，使得模型能够通过梯度反向传播自动优化各特征的贡献比例。

特征融合部分，它采用双路径架构，显式加权路径实现线性组合：

它保留了传统加权平均的可解释性优势。而隐式融合路径通过多层感知机实现深度非线性交互，其数学表达为：

### 模型架构

可学习加权融合技术采用了层次化的组件结构，每个模块都具有明确的功能定位。这种设计实现了两个重要功能：一方面通过概率约束保证权重分配的合理性，另一方面通过可微变换支持端到端训练。特征编码器采用标准的神经网络结构：

其核心作用是通过多层非线性变换提取高阶特征表示，其中批归一化层加速训练收敛，Dropout层增强泛化能力，ReLU激活函数引入非线性，这些设计共同保证了特征编码的效率和效果。同时，整个架构通过梯度传播：

实现端到端优化，使得权重生成与特征学习能够协同进化，这是模型能够自适应不同数据分布的关键所在。

### 动态特性

该技术的动态特性主要体现在权重演化机制和特征交互能力两个方面。权重演化过程遵循动力学规律：

其核心价值在于实现了融合策略的自主进化：训练初期系统快速识别主导特征并分配较高权重，中期通过非线性交互发掘次要特征的价值，后期达到各特征权重的稳定平衡。其设计具有三大功能优势：能够建模特征间的互补效应，提升信息利用率；可以抑制冲突特征的负面影响，增强融合鲁棒性；还能捕捉条件依赖关系，实现情境感知的融合策略。

# 模型研究实践

## 基于行政区划的K-means聚类模型

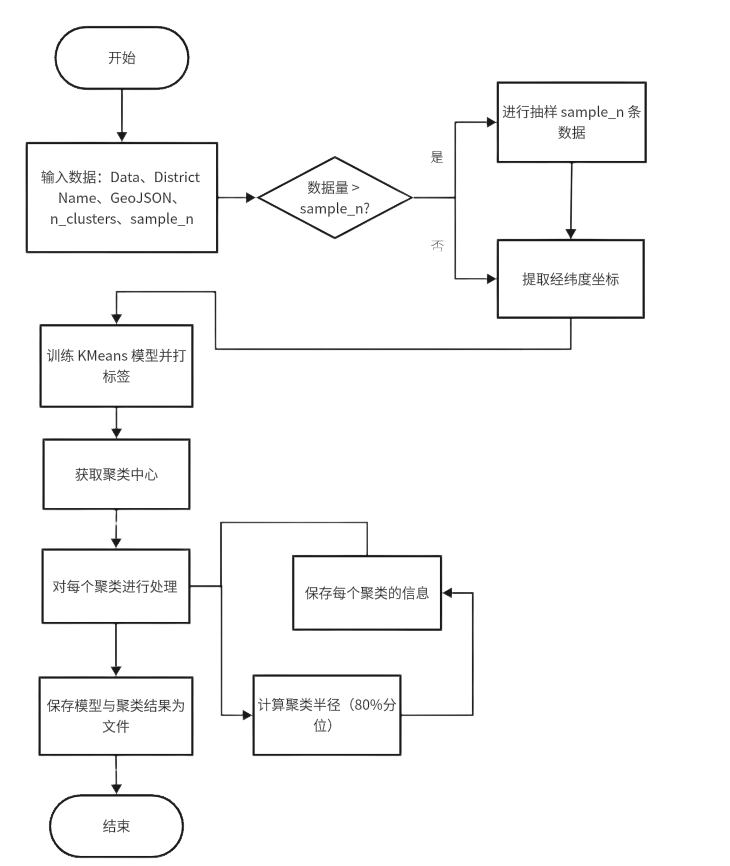


图 5 聚类模型流程

### 聚类基本模型

在完成空间划分后，本文以行政区为单位，对每个区域内的数据分别进行聚类建模处理。每个行政区的数据被视为独立子集，输入至 K-means 聚类算法中，提取该区域内的热点中心及相关统计参数。

为保证各行政区的聚类效果具备可比性与一致性，本文对每个行政区划分为6个聚类簇。考虑到部分行政区数据量极大，为提高聚类效率并避免模型受极端密度点影响，可以数据量超过100,000条的区域进行了固定抽样数为100,000的随机抽样处理，以保留数据的代表性同时提升聚类速度[13]。

聚类模型建成后，为每个聚类计算如下关键参数：

表 4聚类模型参数介绍

|  |  |  |
| --- | --- | --- |
| **参数名称** | **变量名称** | **参数解释** |
| 聚类中心坐标 | center\_lat, center\_lon | 每一簇的空间中心点， 即热点地理中心位置 |
| 聚类半径 | radius\_km | 基于地理空间距离，取各聚类点到中心点 距离的第80百分位作为该聚类的影响半径 |
| 聚类热度指标 | heat | 即每个聚类中包含的数据点数量， 用于衡量该区域的长时间行车热度强弱 |
| 区域名称/编号 | area\_ID | 结合行政区名与聚类编号共同构成聚类标识符“行政区名-序号” |

聚类完成后，模型输出每个簇的中心点，并基于地理空间距离计算各聚类簇内点至中心点的 80% 分位距离，作为该簇的影响半径。此外，本文还统计了每簇所包含的轨迹点数量，作为热点区域的长时间行车热度值指标（与后文中热度值，即上车点数量不同）。

本文为提升模型在后续系统中的可用性与扩展性，将所有聚类模型及其结果参数序列化保存。

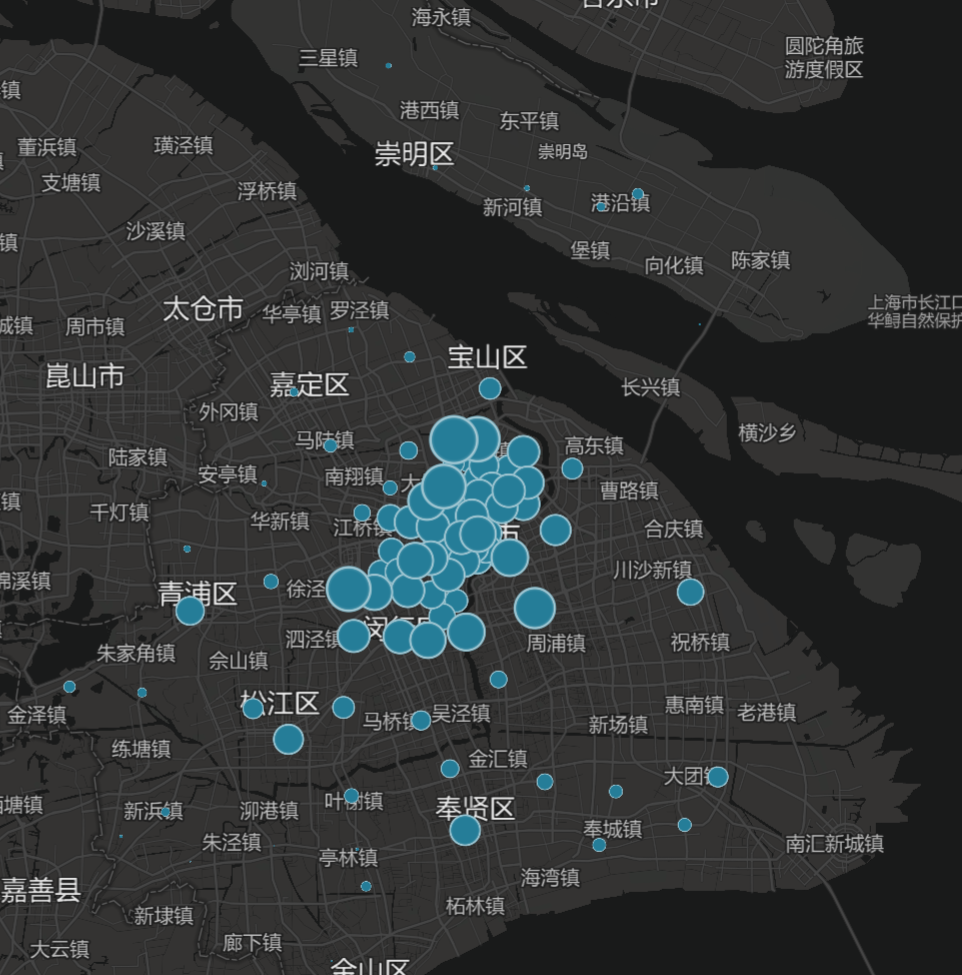


图 6 聚类区域地理位置图

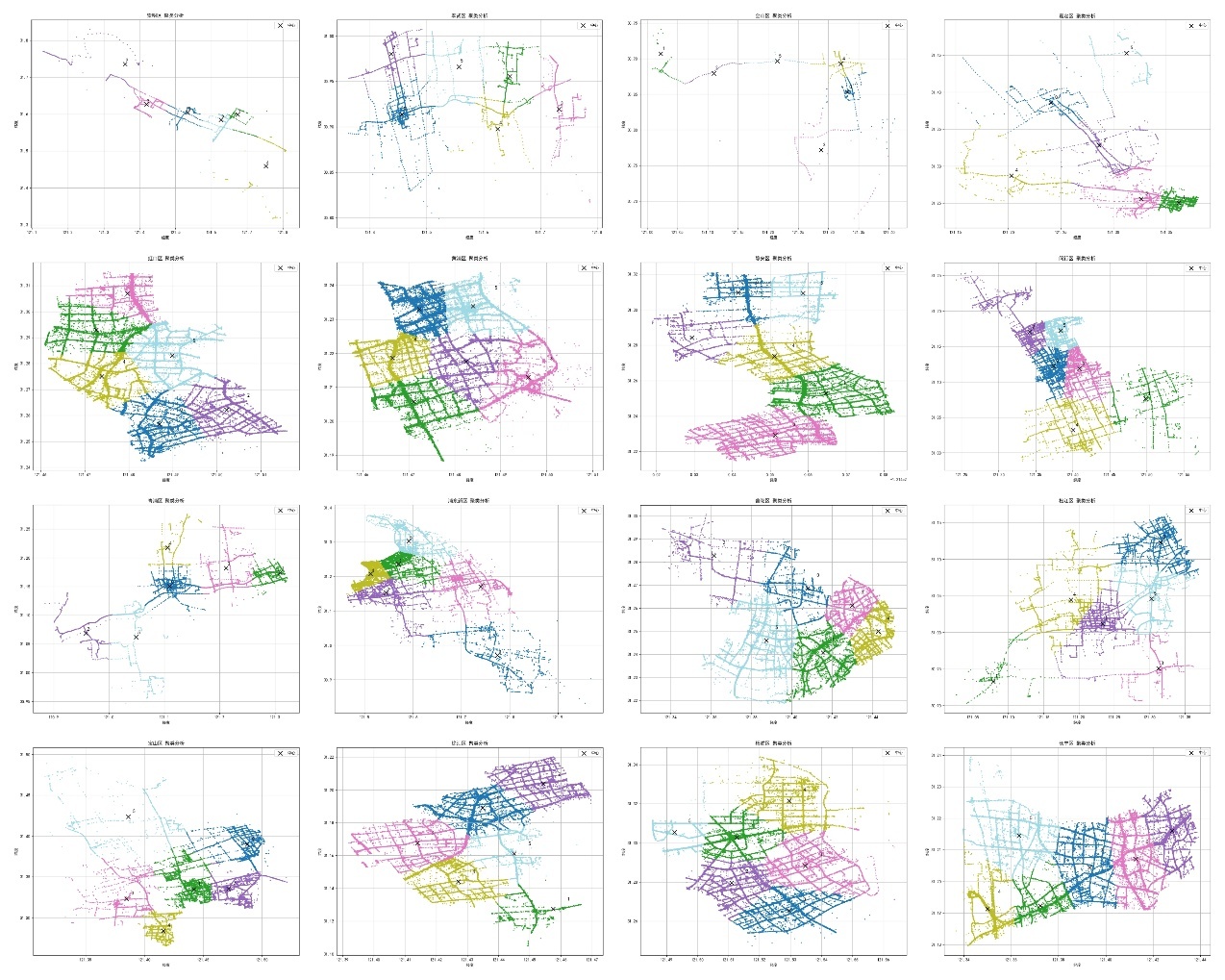


图 7 各区聚类区域可视化

### 基于聚类结果的距离矩阵分析

为进一步分析不同聚类区域之间的空间分布关系与邻近程度，本文基于聚类中心点的地理坐标，构建区域间的距离矩阵，并进行可视化展示与统计分析。

为准确计算各聚类中心之间的地理距离，本文采用了 Haversine公式，以计算球面坐标系统中两点之间的最短路径计算。：

在实际实现过程中，首先从聚类结果中提取每个聚类区域的中心点坐标，随后构建距离矩阵，其中每个元素表示第*i*个聚类中心到第*j*个聚类中心的直线距离。

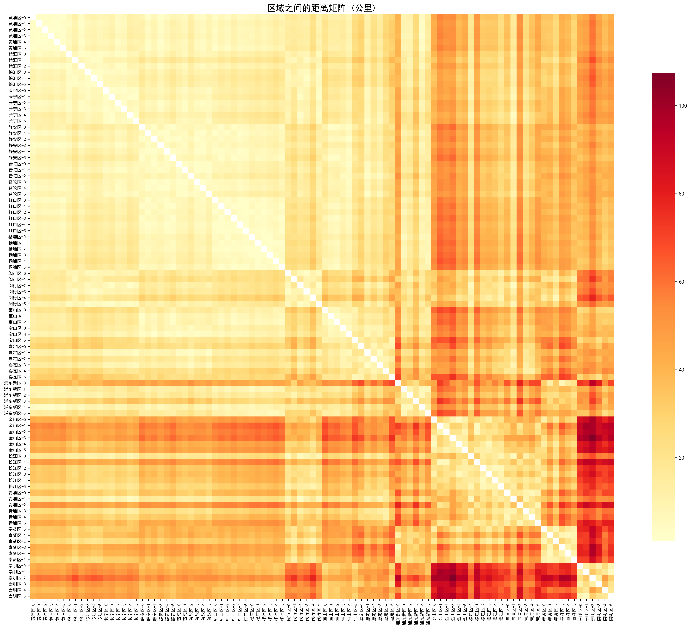


图 8 距离矩阵热度图

## 爬虫获取黑猫投诉内容

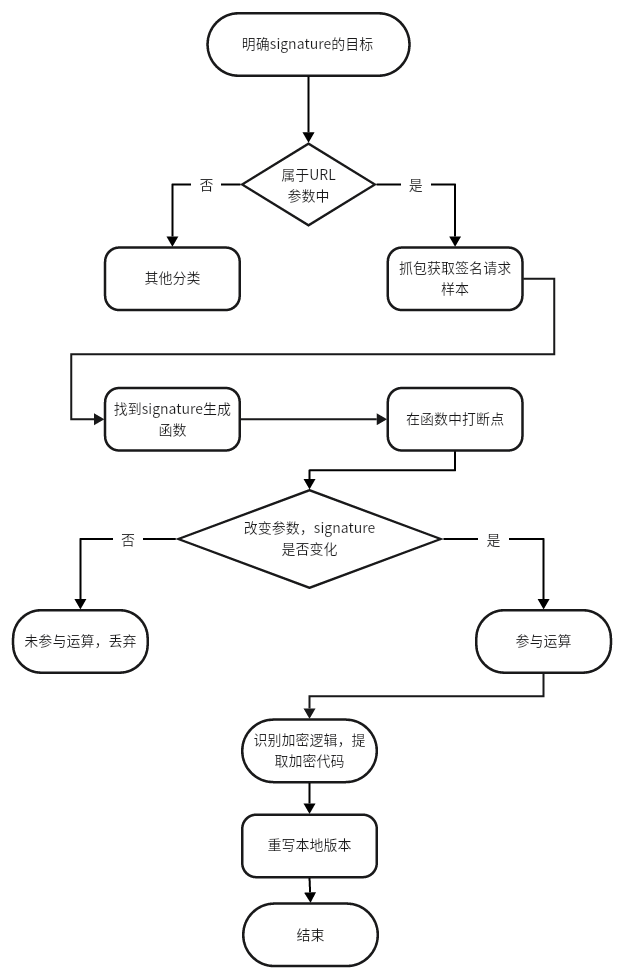


图 9 黑猫投诉内容爬取流程图

黑猫投诉是新浪旗下的网站，采用基于 *signature* 的加密机制，在爬取的时候需要逆向获取加密方法。

*signature*是签名内容，需要查看网站的签名生成方法，证明URL中的签名，其生成逻辑需通过逆向分析获取。本文通过*DEVTools*进一步查看签名所在函数，在关键位置添加断点查看参数，发现总共有6个参数涉及签名，*l,p,b,keyword,c,d*。为了实现本地签名加密，需要具体查看加密方法和参数来源[4]。该签名采用标准的SHA256加密算法。参数中*l*是时间戳，*p*和*b*是固定值，*keyword*是搜索关键词，*c*是一页显示数量，也是固定值，*d*与页面滚动有关。对上述参数进行拼接后产生哈希值，在本地环境中复现加密逻辑。

同时，请求URL页面需要避免被拒绝。由于网站一般为了放置恶意爬取部署多重反爬机制，所以采用多种方式模拟用户请求数据。本研究通过完成Headers 模拟，包含user-agent，等真实请求头，绕过UA检测，来源检测；登录后注入真实Cookies值，用于识别真实身份；添加时间戳形成动态签名，签名随时间，页面，关键词变动，绕过JS加密限制。

## 构建LSTM模型并预测密度与热度

### 交通流量密度预测

#### 交通数据处理与矩阵构建

在获取干净的GPS数据后，程序利用预定义的聚类信息将每个GPS点分配到对应的区域。采用*KDTree*算法实现高效的最近邻搜索，计算每个GPS点到聚类中心的欧几里得距离，并判断其是否在聚类半径内。距离计算公式为：

根据以上数据，程序计算区域间距离矩阵，结合动态阈值生成邻接矩阵，其元素定义为：

这种自适应策略考虑了区域大小差异的影响，避免了大区域过度连接或小区域连接不足，大区域覆盖小区域的问题。

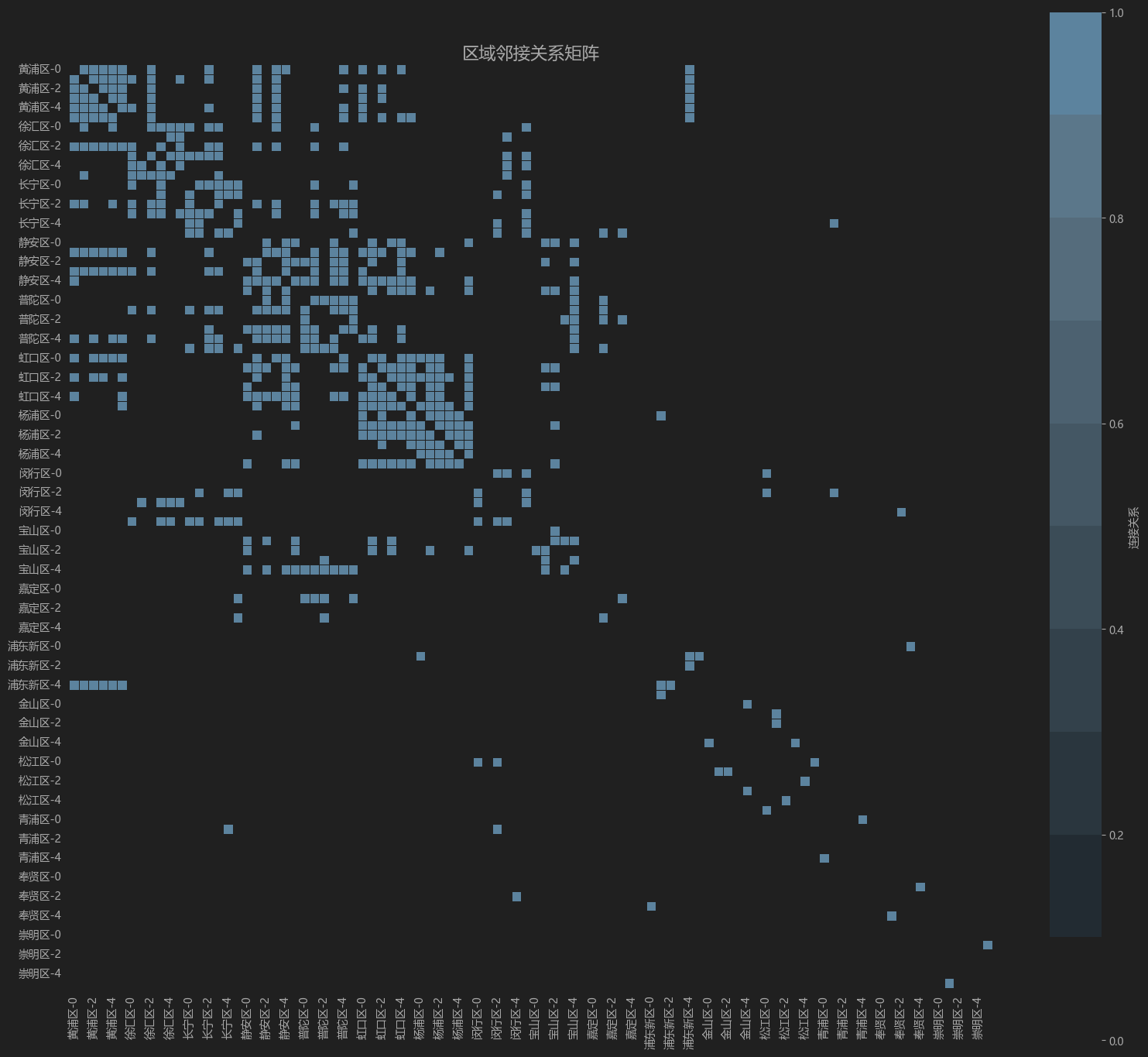


图 10 数据邻接矩阵

随后，程序以5分钟为时间间隔对数据进行分组，计算每个聚类区域内的平均速度和流量，生成三维特征矩阵，其形状为***时间步数,区域数,特征数***。矩阵元素表示在时间步、区域的特征（为二维，表示流量与速度）,通过*Min-Max*标准化消除量纲差异。标准化公式为：

此步骤可提升模型收敛速度并缓解梯度爆炸风险。这种方法保留了时间序列的连续性，捕捉了空间分布的异质性。

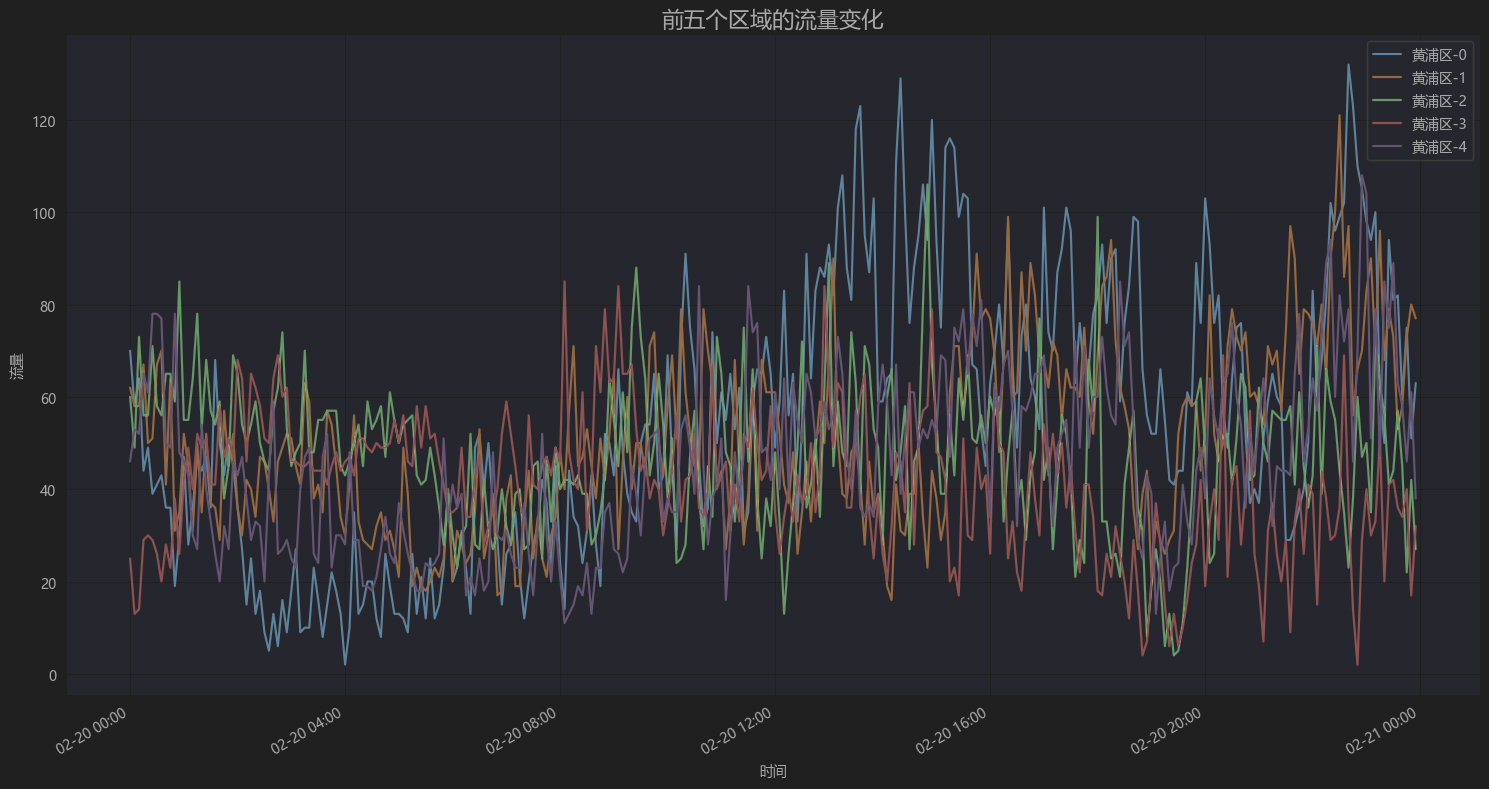


图 11 流量变化数据图

#### LSTM模型架构与训练机制

首先，模型对输入数据进行了时序切分处理，将长时间序列按照滑动窗口方式切分为多个样本，每个样本包含输入序列和对应的输出序列。形式化表示为：

程序按照 7:2:1 的比例将数据集划分为训练集、验证集和测试集，

同样的，研究对特征进行了标准化处理，将各特征转换到统一的数值范围，避免了量纲不同带来的训练困难。标准化采用了如下公式：

其中*μ*和*σ*分别为训练数据的均值和标准差。标准化的必要性在于，流量和速度的量级差异会导致模型收敛困难，通过将数据缩放到相同尺度，程序提高了训练的稳定性。

本文设计的LSTM 模型架构包含输入层、LSTM 隐藏层和全连接输出层。输入层接收维度为2，对应流量和速度的特征向量；隐藏层包含 16 个 LSTM 单元；输出层将 LSTM 的隐藏状态映射到2个维度的预测结果。模型参数初始化采用Xavier方法。

在训练优化部分，本文采用了Adam优化器训练模型，训练过程中，模型在每个 epoch 后在验证集上评估性能。损失函数则采用了掩码平均绝对误差作为主要损失函数：

其中，是掩码矩阵，用于排除异常或缺失值的影响。同时，引入了流量特征和速度特征的加权机制，通过参数*θ*平衡两者的贡献：

同时加入评估指标，全面衡量模型性能：

训练结束后，模型在测试集上进行最终评估，并将训练好的模型参数保存，以供后续模拟和应用使用。

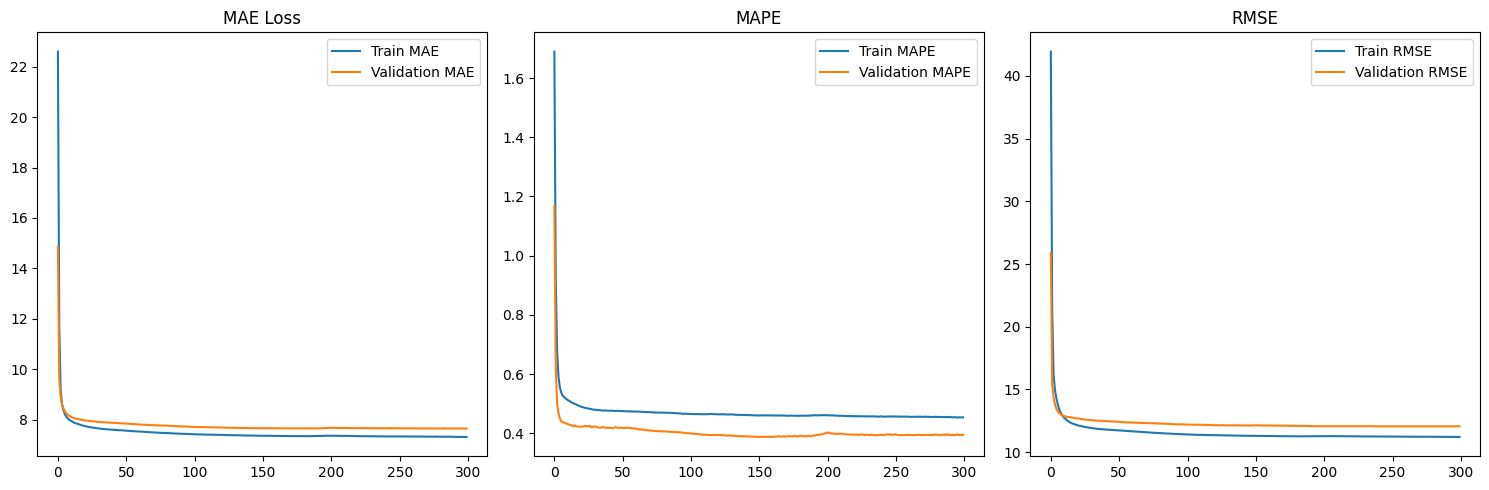


图 12 模型训练和验证指标变化曲线

从图中可以看出，无论是MAE、MAPE还是RMSE，训练集和验证集的误差都在迅速下降，并在很短的时间内趋于稳定。这表明模型在训练数据上表现良好。误差的快速收敛显示出模型学习能力强，能够有效捕捉数据中的趋势。同时，训练误差和验证误差之间的差距很小，几乎没有过拟合的迹象。

#### 交通流量模拟与可视化分析

预测模块采用滑动窗口策略，将最新预测结果反馈至输入序列以保持时间连续性. 程序设计了历史窗口机制，连续输入前𝑇个时间步的历史数据，通过模型递推预测下一个时间步的输出。为了适配LSTM的输入需求，特征矩阵需转化为的形式，即将所有区域在一个批次中并行处理，提升预测效率。具体地，对于每个时间步*t*：

1.将当前窗口 输入模型，获取预测值

2.逆标准化还原真实流量，计算区域流量密度 ,其中 为流量，为区域半径；

3.更新历史窗口为 ,实现自回归预测。

预测输出后，程序通过逆变换复原为实际物理值，其变换公式为：

预测结果包括每个区域的未来流量与速度。程序关注于区域的流量密度，即单位面积上的车辆流量，其计算公式为：

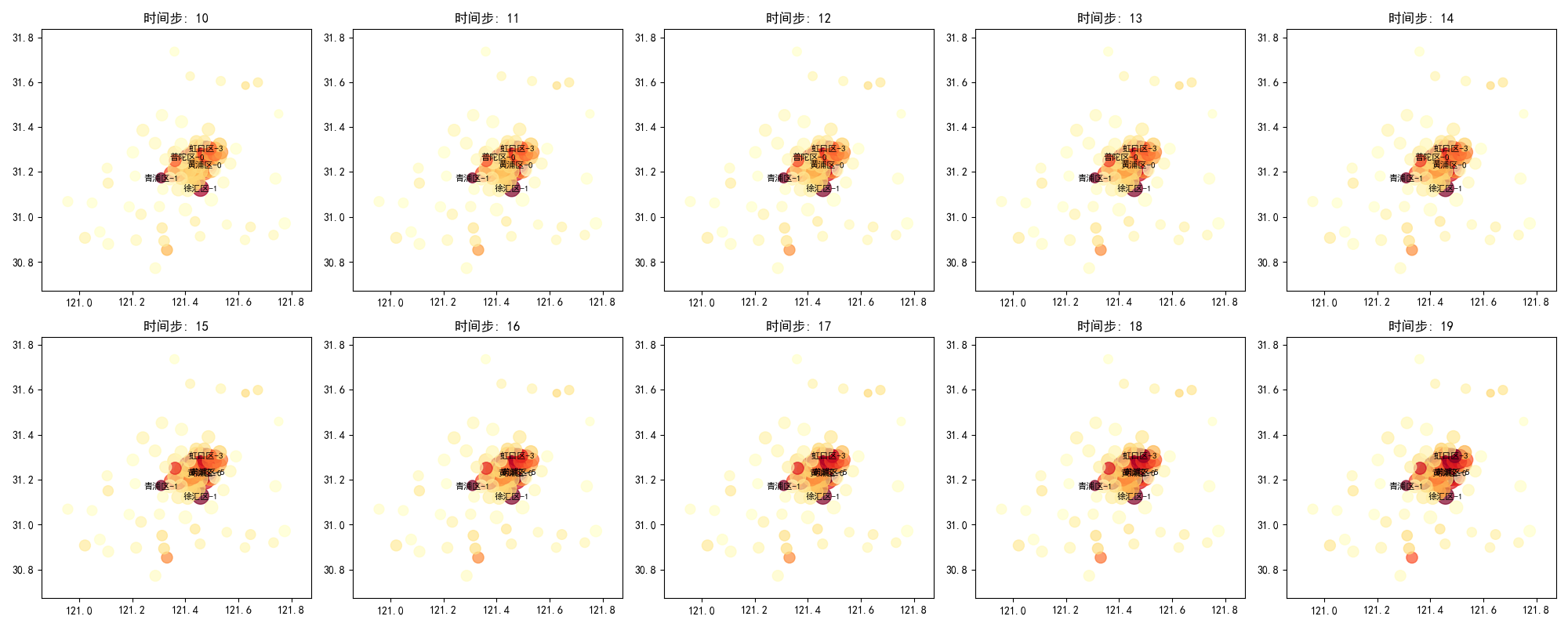


图 13 LSTM流量数据预测

由输出结果可见，本文的LSTM模型较好的预测并可视化了各个聚类地区的交通流量数据，输出流量密度处于前列的几位聚类区域，为后续加权融合模型的建立提供了基础。结合多源数据的LSTM模型被成功用于城市交通流预测，证明其可应对非结构化数据挑战[16]。

### 上车点热度预测

#### 数据预处理

程序通过检测占用状态（occupied）从0（空车）变为1（载客）的时刻，提取上车点。具体而言，程序对每个数据文件按车辆标识和时间戳排序，确保时序连续性。然后，应用条件判断当前记录占用状态为1且前一记录为0，同时排除跨车辆的错误检测，即确保前后记录属于同一车辆。

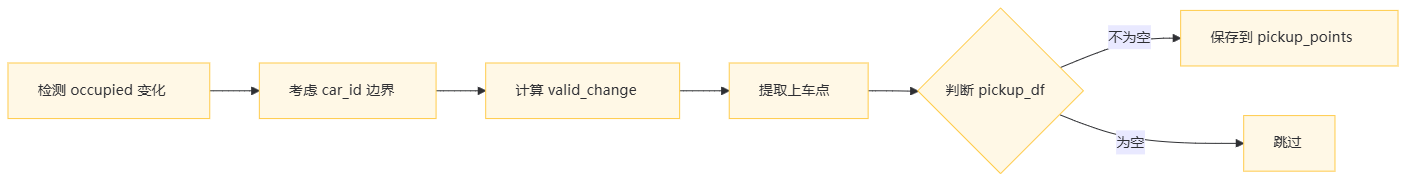


图 14 上车点热度预测数据预处理流程图

在上车点提取后，程序同样利用预定义的聚类数据，将上车点映射到最近的聚类区域。随后，程序以1小时为时间间隔，统计每个聚类区域的上车点数量，生成特征矩阵  
。其中其中*T*为时间步数，*N*为聚类数。数据经Min-Max标准化后输入模型，公式为，

若某些聚类在某时间窗口内无上车点，程序为其填充零值，确保矩阵包含所有预定义聚类。

#### LSTM模型架构与训练机制

在热度预测部分，为了和流量预测部分LSTM模型可以配合使用，模型使用参数与后者相同。模型采用单层LSTM结构，输入为历史12小时的热度序列 ，输出未来1小时的热度预测 。

同时，本部分研究采用加权混合损失函数：

其中MAE衡量预测偏差的绝对值，MAPE关注相对误差，RMSE惩罚大误差项。通过Adam优化器动态调整参数，并引入早停机制防止过拟合

#### 长期预测与动态可视化

本部分程序加载训练好的LSTM模型和标准化器，并以测试集的首个12小时序列作为历史窗口，提供连续预测的起点，使用采用自回归滚动预测策略。

模拟过程生成未来50个时间步的热度预测，采用滚动预测机制。每一步，程序标准化历史窗口，输入模型预测下一时间步热度，剔除最旧数据，添加新预测值，形成：

同时逆标准化，将预测值还原为实际上车次数，移除最早时间步，生成长期预测序列。

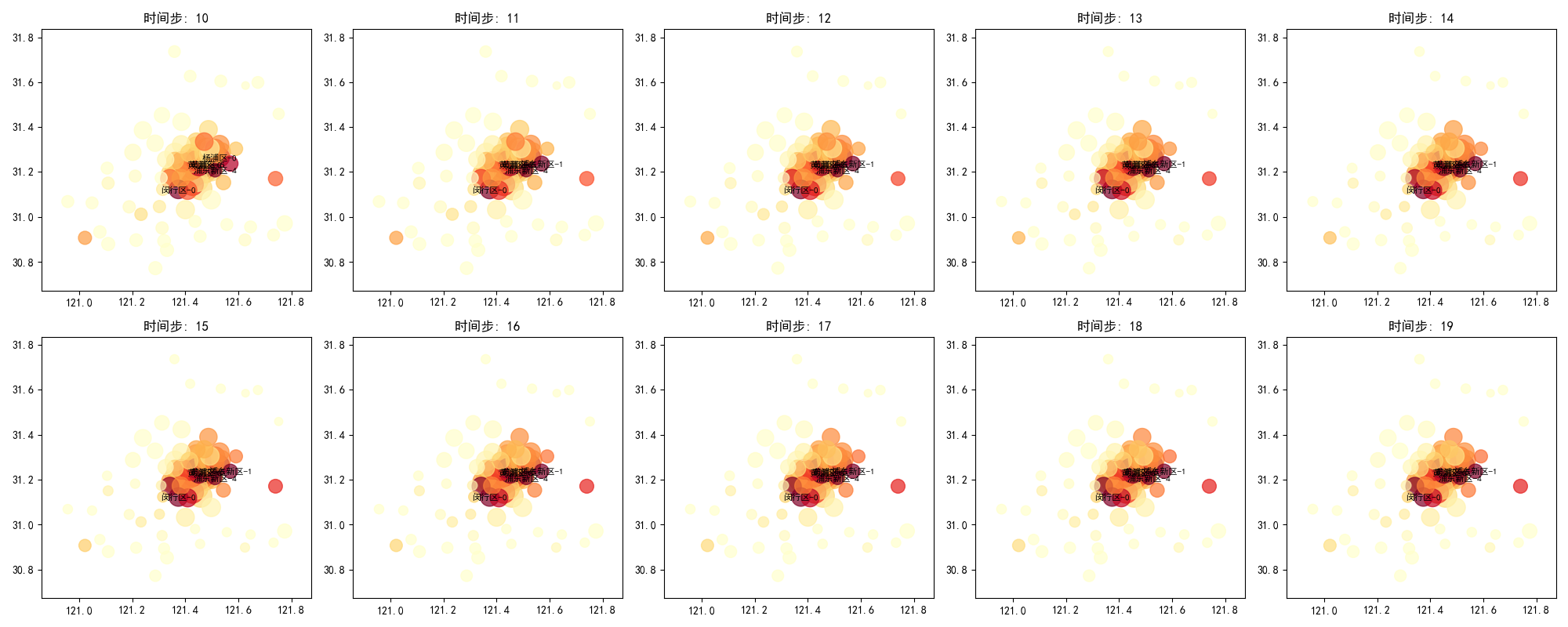


图 15 LSTM热点数据预测流程图

可见，LSTM模型预测并可视化了各个聚类地区的热点数据，并输出处于前列的几位聚类区域，为后续加权融合模型的建立提供了坚实基础。

## 可学习加权融合模型的构建训练

### 数据加载获取及训练数据准备

首先，本文依据先前代码结果，构建综合性的数据环境，包含三类关键数据：区域聚类结果、风险矩阵及距离矩阵。

本文使用先前LSTM进行时序预测的历史数据，输出预测未来若干个时间段的时段的区域热度与流量密度分布预测。多变量堆叠式LSTM模型被证实在短期交通预测任务中具有更高的预测精度[17]。

为了让不同量纲的特征具有可比性，程序采用了最小-最大缩放（Min-Max Scaling），将特征值映射到区间[0,1]。对于一个特征向量，标准化后的值 计算公式为：

这种标准化处理确保了不同量纲的特数值上具有可比性，避免因量纲差异导致模型训练偏向某一特征。标准化后的特征被进一步整合为一个四维向量，包含热度 、密度 、风险 和距离（距离在此阶段设置为使用到其他所有区域的平均距离，后续在车辆决策时实际进行区域间距填写）。

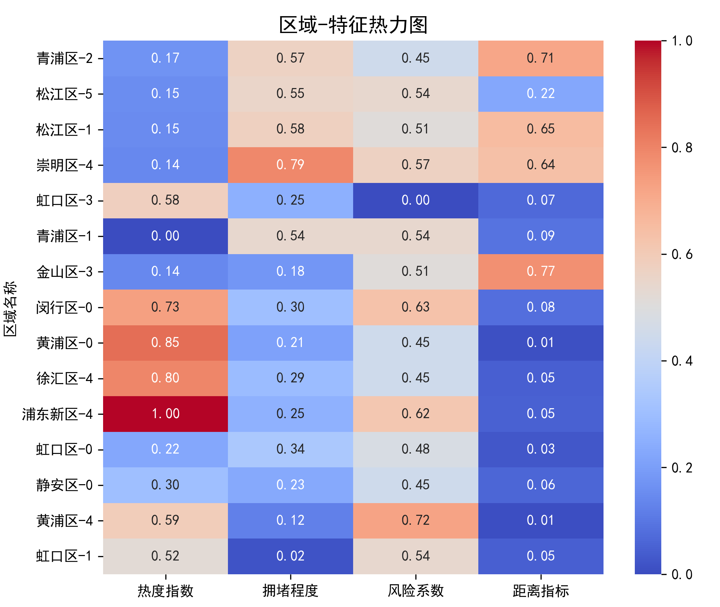


图 16 区域-特征热力图

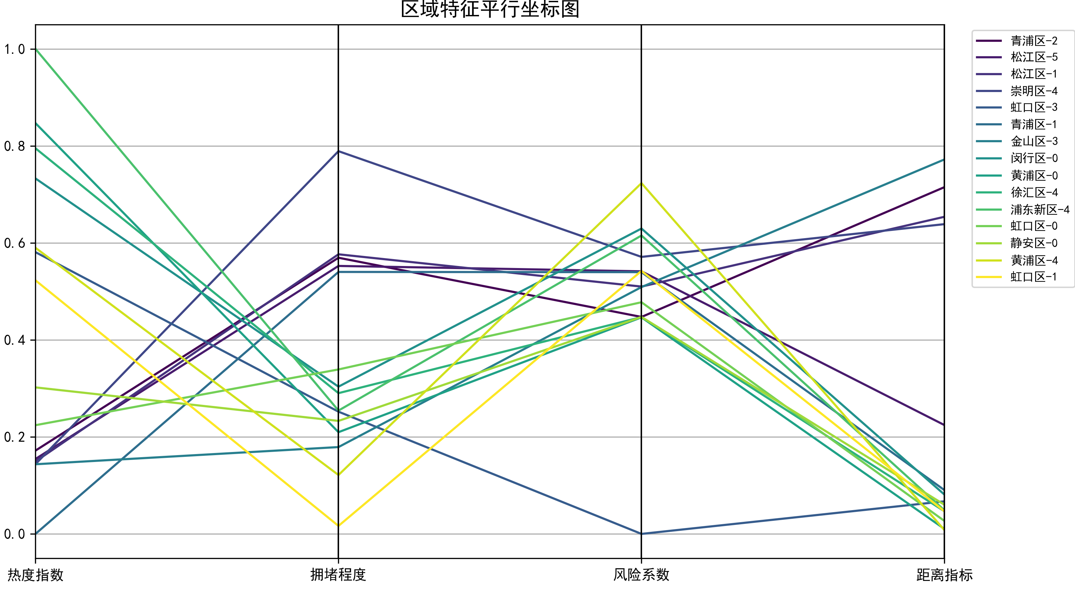


图 17 部分区域特征平行坐标图

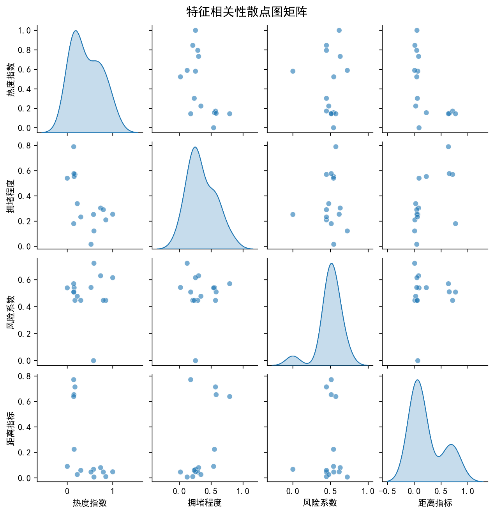


图 18 特征相关性散点图矩阵

### 融合模型建立

在数据准备就绪后，研究进入模型设计阶段，构建了可学习加权融合模型以实现多因素的融合预测。该模型由两个核心组件构成：可学习权重层和多层感知机MLP。可学习权重层负责动态调整热度、密度、风险和距离四个因素的重要性，并通过训练过程优化。

程序设计的要点为可学习权重层，使用*Softmax*函数将原始权重归一化处理，确保所有权重为正且总和为1。将原始权重 转换为归一化的权重 ，计算公式为：

特征融合网络采用多层感知机（MLP）架构，用于学习各因素间的复杂交互关系。该网络的输入维度为 4（对应四个因素），通过 [64,32,16] 三个隐藏层，最终输出单一评分值。每个隐藏层后跟随 ReLU 激活函数、批量归一化和 Dropout（率为 0.2），有效增强了模型的泛化能力和稳定性。网络形式化表示为：

此融合模型的必要性在于，出租车区域密度预测涉及多个相互作用的因素，单一特征无法全面反映区域的优劣。可学习权重层提供了灵活性，使模型能够根据训练数据调整各因素的贡献，而MLP的深层结构则能够挖掘特征间的复杂关系，从而生成更准确的评分。

### 融合模型训练

模型训练是研究的优化阶段，旨在通过迭代更新参数，使可学习加权融合模型更好地预测区域密度。训练过程利用准备好的数据加载器，采用Adam优化器调整模型参数，学习率为0.001，训练轮次设为150。每一轮训练中，模型接收批次特征输入，计算区域评分，并通过自定义的损失函数评估预测效果。

此部分的关键在于设计的复合损失函数，综合考虑了各因素的奖励和惩罚。损失函数定义为：

其中 为基础预测损失，为热度奖励项（负值，因为热度高的区域更受欢迎）， 为密度惩罚项，为风险惩罚项，为距离惩罚项，、、、 为各项的权重系数，这些系数通过可学习权重层动态调整。

据此，函数鼓励模型选择热度高的区域，惩罚模型选择高密度（拥堵）、高风险、长距离区域。

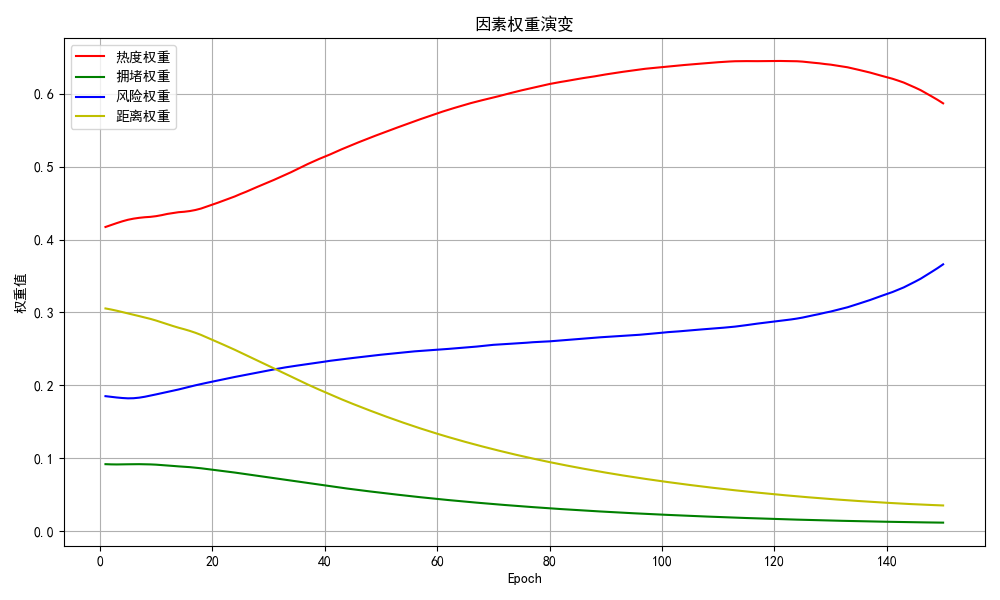


图 19 模型权重训练演变

根据动态权重演变结果，本文发现权重变化：

#### 热度权重变化

初期快速上升表明模型优先学习需求热度的强相关性，符合出租车中"热区优先"的直觉；中期达到峰值后缓慢下降，但保持主导地位。

#### 风险权重持续上升

单调递增曲线显示模型逐步重视安全因素；训练后期增速变高，最终成为第二大权重，证明了本文的创新点：风险值对车辆区域的决策有确实存在的意义与作用。

## 动态区域推荐

### 计算区域排名

基于训练好的融合模型，本文设计了区域排名算法，为每辆出租车计算最优的目标区域。

对于每辆车，本文计算其当前位置到各区域中心的距离。与传统方法不同，本文采用了接近度转换策略，将原始距离 *d* 转换为接近度 *p* ：

这种双曲转换使得距离特征与其他特征具有相同的方向性（值越大越好），便于模型一致处理。接近度同样通过标准化器进行归一化处理。

接下来，本文为每辆车构建特征矩阵，形状为 [1, num\_areas, 4]，包含各区域的热度、密度、风险和距离（接近度）特征。将特征矩阵输入融合模型，得到各区域的评分。此处，融合模型的作用是将多元特征综合评估为单一评分值，评分计算过程为：

其中 *f* 表示经过训练的神经网络模型。通过对评分进行降序排序，获得每辆车的目标区域排名。

这种基于学习的排名方法相比传统固定权重方法具有显著优势：一是权重自适应，能够根据历史数据自动调整各因素的重要性；二是非线性特征融合，能够捕获因素间的复杂交互关系。

### 长期决策优化模拟

为验证决策系统的实际效果，本文设计了包含车辆位置模拟与长期决策规划的实验框架。首先，本文模拟5辆出租车，将它们随机放置在不同的城市区域中心位置。同时，传统的出租车决策通常只考虑当前时刻的最优决策，忽略了决策的长期影响。本文设计了递归优化决策算法，该算法规划未来多个时间步的行动路径，实现了从短期决策到长期规划的跨越。

递归优化决策算法基于以下核心思想：对每辆车，从当前位置开始，规划最多 max\_depth 步的行动路径，每一时间步重新计算区域排名，类似于强化学习中的时序差分学习。算法使用折扣因子来平衡即时收益和未来收益：

其中表示状态 的价值，表示在状态 下采取行动 的即时收益，为折扣因子（0.85）， 表示下一状态的价值。

算法加入了剪枝阈值机制，当某路径的累积评分低于阈值时，停止该路径的进一步探索，有效减少了搜索空间。同时，为防止车辆在短时间内反复访问同一区域，算法对连续访问同一区域的评分进行递减处理：

同时，算法考虑了物理约束，引入了最大移动距离限制，确保规划的路径在物理上可行。

优化后的长期决策方案为每辆车生成了一条由多个时间步组成的路径。每个路径点包含区域 ID、评分和时间步三元组，完整描述了车辆的行动计划，并加入地图可视化。

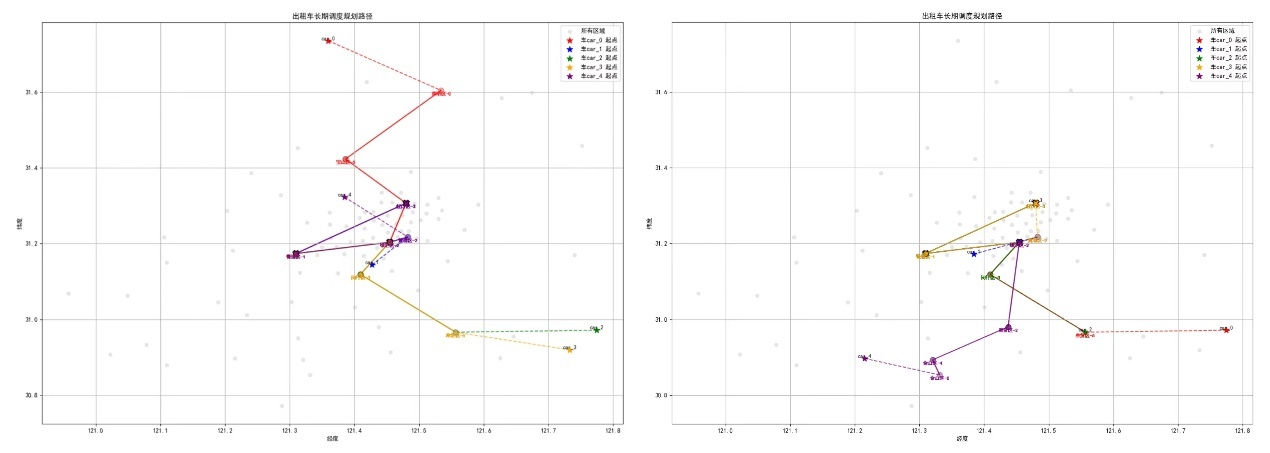


图 20 部分模拟路线规划

由图片显示，本文的模拟路线应用展现了较强的正确性，可以同时模拟多辆车的运作与决策，成功结合了多项因素考量，模拟出租车动态选择区域的实际模型。长期决策规划的优势在于：考虑了决策的时序依赖性，避免了短视决策可能带来的长期不利影响；优化了车辆的整体收益，而不仅仅是单步收益。通过这种方式，系统能够在满足乘客需求的同时，提高车辆运营效率，为司机提供更科学的决策指导。

# 结论与建议

## （一）结论

本文针对上海市出租车运营需求，提出了基于深度学习和用户反馈的动态区域选择模型，设计并验证了该模型的可行性。以下为本文的主要成果：

### 基于LSTM的交通模式预测

本文采用长短期记忆网络（LSTM）对交通流量和上车点热点进行时序预测。LSTM通过捕捉数据的长期依赖关系，准确反映了上海市交通模式的动态变化，为区域推荐提供了可靠依据。模型在训练过程中表现稳定，预测结果与实际交通趋势高度一致。

### 用户反馈的决策优化与多源数据的融合机制

通过整合用户反馈数据（投诉数据），本模型在区域选择中引入了用户体验维度。研究使用了可学习加权融合技术，综合处理GPS轨迹数据、用户反馈等多源输入，确保了推荐结果的全面性和适应性。

### 动态推荐与长期规划

本模型实现了动态区域推荐，并通过递归算法支持长期决策。系统基于多个权重和推荐区域，同时优化司机未来多个时间步的行动路径。模拟中，系统为每辆车生成了实时推荐区域和行动路径，展示了其在提升运营效率方面的潜力。

本文开发的动态区域选择模型通过深度学习技术和多源数据融合，为上海市出租车行业提供了一种智能化的决策工具。模型在预测交通模式、优化区域选择以及提升用户体验方面取得了显著成效，为城市交通管理提供了新思路。

## （二）建议

基于本文的结果和国家政策导向，为进一步推动上海市出租车动态区域选择模型的实际应用和智能交通的发展，提出以下具体、可操作的建议：

### 政策支持与推广应用

建议政府部门加大对智能交通技术的支持力度，推动基于深度学习和用户反馈的动态区域选择模型在出租车行业的广泛应用。可通过政策激励和试点项目，鼓励出租车企业率先采用深度学习相关模型，提升行业竞争力，助力智慧城市建设和“双碳”目标的实现。

### 数据共享与行业合作

建议建立城市交通数据共享平台，促进政府、企业和研究机构之间的协作。通过整合出租车GPS数据、网约车数据、公共交通数据及用户反馈等多源资源，为智能交通研究和应用提供更全面的支持，从而优化城市交通管理。

### 加强用户参与与反馈机制

建议出租车行业借助移动应用或投诉平台，鼓励乘客提供实时反馈，建立完善的用户参与机制，进一步提升服务质量和乘客满意度。

参考文献

1. 刘耀宗, 张宇, 王宇. 出租车乘客轨迹的空间聚类方法[J]. 计算机工程与应用, 2018, 54(14): 249-255.
2. 党鹏飞. 基于GPS数据的出租车寻客行为分析及路径规划研究[D]. 大连理工大学, 2021.
3. 熊敏. 基于出租车轨迹热点区域的寻客路径优化模型和策略研究[D]. 重庆: 重庆交通大学, 2022.
4. 邢羽琪,杨柽.基于逆向技术的深层网络爬虫与数据分析[J].软件工程,2023,26(12):41-45.DOI:10.19644/j.cnki.issn2096-1472.2023.012.009.
5. BESSE P C, GUILLOUET B, LOUBES J M, et al. Destination prediction by trajectory distribution based model[J].
6. BIAN J, TIAN D, TANG Y, et al. A survey on trajectory clustering analysis[J].
7. GONG S, CARTLIDGE J, BAI R, et al. Extracting activity patterns from taxi trajectory data: A two-layer framework using spatio-temporal clustering, Bayesian probability and Monte Carlo simulation[J]. International Journal of Geographical Information Science, 2020, 34(6): 1210-1234.
8. RATHORE M M, PAUL A, AHMED A, et al. Co-clustering-based driver grouping and pickup efficiency prediction in taxi systems[J].
9. LIU Y, ZHANG Y, WANG Y. An improved fuzzy trajectory clustering method for exploring urban travel patterns[J]. Journal of Advanced Transportation, 2021, 2021: 1-14.
10. YUE M, LI Y, YANG H, et al. DETECT: Deep trajectory clustering for mobility-behavior analysis[J].
11. LIU K, CHEN Z, YAMAMOTO T, et al. Exploring the impact of spatiotemporal granularity on the demand prediction of dynamic ride-hailing[J]. Transportation Research Part C: Emerging Technologies, 2021, 128: 103182.
12. An L, Sun X H, Wang Y. K-Means Clustering Algorithm Based on Improved Differential Evolution[J]. Information Dynamics and Applications, 2024, 3(3): 200-210.
13. Bhattacharjee R, Imola J, Moshkovitz M, Dasgupta S. Online k-means Clustering on Arbitrary Data Streams[] Proceedings of The 34th International Conference on Algorithmic Learning Theory. 2023: 204-236.
14. Swier Garst, Marcel Reinders. Federated K-means Clustering[J]. arXiv preprint arXiv:2310.01195, 2023.
15. Papastefanopoulos A, Karyotis C, Sideris A. Multivariate time-series forecasting: A review of deep learning methods in IoT applications to smart cities[J]. Smart Cities, 2023, 6(5): 2242–2265.
16. Essien A, Yu Z, Shu L, et al. A deep-learning model for urban traffic flow prediction with traffic events mined from Twitter[J]. World Wide Web, 2021, 24: 2691–2709.
17. Mondal S, Rehena Z. Stacked LSTM for short-term traffic flow prediction using multivariate time series dataset[J]. Arabian Journal for Science and Engineering, 2022, 47: 9997–10007.
18. Metcalf L, Casey W. Anomaly detection[A]//Cybersecurity and Applied Mathematics. Elsevier, 2016: 57.

致谢

本研究得以顺利开展，离不开香港科技大学智慧城市研究小组提供的上海市出租车 GPS 轨迹数据集。这些数据内容丰富、覆盖广泛，为我们的研究提供了坚实的基础。在此，我们向数据提供方表示衷心的感谢。正是有了这些高质量的数据支持，我们才能深入挖掘出租车动态区域选择的规律，探索出具有创新性和实用性的模型。感谢数据提供方的无私分享，让我们能够站在巨人的肩膀上开展研究工作。

在学习过程中，各位数学学院课程授课老师严谨的治学态度和渊博的专业知识，为我们打开了知识的大门。老师深入浅出的讲解，让我们对复杂的理论知识有了清晰的理解；悉心的指导，帮助我们掌握了一项项研究技能。在遇到困难时，老师总是耐心地为我们答疑解惑，给予我们鼓励和支持，让我们重拾信心，继续前行。感谢授课老师为我们打下的坚实基础，让我们能够在学术道路上稳步前行。

2025年4月17日