

**本科毕业设计（论文）**

基于共享单车骑行数据的上海市通勤模式分析

**Commuting Patterns Unveiled through Shared Bicycle Data**

学 院：测绘与空间信息学院

专业班级：地理信息科学 2 班

姓 名：潘志清

学 号：202001020717

指导教师：牟乃夏

完成日期：2024年5月17日

教务处制

**学位论文原创性声明**

本人呈交给山东科技大学的学位论文，除所列参考文献和世所公认的文献外，全部是本人攻读学位期间在导师指导下的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

若有不实之处，本人愿意承担相关法律责任。

本人签名： 日 期： 年 月 日

**学位论文使用授权声明**

本人完全了解山东科技大学有关保留、使用学位论文的规定，同意本人所撰写的学位论文的使用授权按照学校的管理规定处理。

作为申请学位的条件之一，学校有权保留学位论文并向国家有关部门或其指定机构送交论文的电子版和纸质版；有权将学位论文的全部或部分内容编入有关数据库发表，并可以以电子、网络及其他数字媒体形式公开出版；允许学校档案馆和图书馆保留学位论文的纸质版和电子版，可以使用影印、缩印或扫描等复制手段保存和汇编学位论文；为教学和科研目的，学校档案馆和图书馆可以将公开的学位论文作为资料在档案馆、图书馆等场所或在校园网上供校内师生阅读、浏览。

（保密的学位论文在解密后适用本授权）

作者签名： 导师签名：

日 期： 年 月 日 日 期： 年 月 日

**摘 要**

中文摘要应具有高度的概括性，语言精炼、明确，扼要叙述论文（设计）的主要内容，包括研究目的与意义、研究内容与方法以及研究结论等，同时需要突出论文（设计）的新论点、新见解或创造性成果。要求用中、英文分别书写，字数不少于400字。英文摘要内容应与中文摘要一致，语句通顺，语法正确，准确反映论文（设计）内容。

关键词是供检索用的主题词条，应采用能覆盖毕业论文（设计）主要内容的通用技术词条（参照相应的技术术语标准），可从标题或正文中选择3-5个最能表达主要内容的词语作为关键词，按词条的外延层次排列（外延大的排在前面）。关键词有中、英文对照，分别附于中、英文摘要后。

关键词：山东科技大学；本科生；毕业设计（论文）

**ABSTRACT**

二级标题：黑体4号加黑，行距固定值20磅，段前0.5行，段后0.5行

(一级标题宋体4号加黑，二级及以下标题宋体小4，1.5倍行距)

从该部分开始，外语专业使用相应语种，格式要求与中文模板一致

一级标题：中文黑体3号加黑，英文Times New Roman 小2加粗，单倍行距，段前0.8行，段后0.5行。

三级标题：黑体小4号加黑，行距固定值20磅，段前0.5行，段后0.5行。

In order to study……

………….......

**KEY WORDS:** undergraduate; graduation project (thesis); SDUST

**目 录**

**第1章 绪论 1**

1.1 二级题目 1

1.1.1 三级题目 3

**第2章 一级题目 4**

2.1 二级题目 5

2.1.1 三级题目 5

**参考文献 6**

**附录 7**

**致谢 8**

**目 录**

[**第1章 绪论 5**](#_Toc166618787)

[1.1 研究背景与意义 5](#_Toc166618788)

[1.1.1 研究背景 5](#_Toc166618789)

[1.1.2 研究意义 5](#_Toc166618790)

[1.2 国内外研究现状 6](#_Toc166618791)

[1.3 研究内容与步骤 8](#_Toc166618792)

[1.4 论文组织结构 8](#_Toc166618793)

[1.5 本章小结 9](#_Toc166618794)

[第2章 相关理论基础 10](#_Toc166618795)

[2.1 相关理论基础 10](#_Toc166618796)

[2.1.1 行为地理学 10](#_Toc166618797)

[2.1.2 城市职住平衡与过剩通勤理论 10](#_Toc166618798)

[2.2 相关研究技术 12](#_Toc166618799)

[2.2.1 相似性分析 12](#_Toc166618800)

[2.2.2 密度分析 13](#_Toc166618801)

[2.2.3 用户行为分析 14](#_Toc166618802)

[2.3 本章小结 14](#_Toc166618803)

[**第3章 研究数据与方法 15**](#_Toc166618804)

[3.1 研究范围 15](#_Toc166618805)

[3.2 数据来源与预处理 15](#_Toc166618806)

[3.2.1 数据来源与数据清洗 15](#_Toc166618807)

[3.2.2 轨迹重排序及长度量测 17](#_Toc166618808)

[3.2.3 轨迹简化 21](#_Toc166618809)

[3.3 研究方法 23](#_Toc166618810)

[3.3.1 轨迹终点最近邻匹配算法 23](#_Toc166618811)

[3.3.2 数学形态学 25](#_Toc166618812)

[3.3.3 Jaccard相似系数 25](#_Toc166618813)

[3.4 本章小结 26](#_Toc166618814)

[**第4章 结果与讨论 27**](#_Toc166618815)

[4.1 时间通勤模式分析 27](#_Toc166618816)

[4.1.1 通勤时距模式 27](#_Toc166618817)

[4.1.2 通勤间隔分析 29](#_Toc166618818)

[4.1.3 单车周转率分析 31](#_Toc166618819)

[4.2空间通勤模式分析 32](#_Toc166618820)

[4.2.1 通勤目的地分析 32](#_Toc166618821)

[4.2.2 通勤轨迹分析 34](#_Toc166618822)

[4.3本章小结 38](#_Toc166618823)

[**第5章 总结与展望 39**](#_Toc166618824)

[5.1研究结论 39](#_Toc166618825)

[5.1规划建议 39](#_Toc166618826)

[5.1研究展望 39](#_Toc166618827)

[**参考文献 40**](#_Toc166618828)

[**附录 41**](#_Toc166618829)

[**致谢 42**](#_Toc166618830)

**第1章 绪论**

1.1 研究背景与意义

1.1.1 研究背景

摩拜单车于2016年4月首次在上海投入运营，并掀起了一轮以共享单车市场迅速膨胀为代表的共享经济热潮。33个大中小城市积极引进共享单车服务，当年全国累计投放共享单车约200万辆，日活跃车辆最高时达到了惊人的55万辆。此时小时单价稳定在一元，共享单车也逐步以“廉价、便捷、绿色”的美好形象打入城市通勤市场，快速扩张。在初步站稳脚跟，共享单车企业们于2019年掀起一轮涨价风波，小蓝单车、青桔单车、摩拜单车、哈啰单车等众多企业上调起步价与每小时单价，有些企业涨幅超过三倍。当年年底，共享单车小时单价正式步入“三元时代”。

交通运输部部长李小鹏认为共享单车是城市慢行系统的一种模式创新，是使用互联网思维解决交通运输问题的一种可行方式，尤其对解决人民群众出行的“最后一公里”问题十分有效。在一些城市，共享单车确实有效缓解了中短途、慢速交通系统的压力，打通了固定的交通基础设施与居民个性化目的地之间的“最后一公里”。共享单车的出现对城市发展，居民通勤行为塑造都带来深远的影响。

城市通勤数据反映了城市空间内中小尺度、周期性的人员流动情况。人们骑行共享单车属于一种缓慢的匀速移动，且具有较高的灵活性，其轨迹往往能够渗透到城市的每一条毛细血管般狭窄的道路中，相较于固定线路的通勤方式更能反映人的行为特性。结合以上两点，借助共享单车通勤轨迹数据，我们既可以从宏观上把握城市主要交通干线、主要产业热点区域一周七个工作日的人口规律性流动情况，分析总体城市职住分布与产业布局情况；我们也可以从城市的微观结构着手，探究渗入城市细小角落的骑行轨迹究竟反映了怎样的行为与细碎环境的相互作用。

1.1.2 研究意义

无桩电子围栏式共享单车是一种具有浓厚互联网气息的产品服务，也是共享经济快速发展阶段的标志性产物。共享单车一般依赖 GPS 定位设备与互联网接入，借助电子围栏算法及主动式 GPS 定位圈定借还车区域及使用范围，并结合移动客户端与后台调度算法规范用户骑行行为（如骑行红包奖励及出圈扣除押金惩罚等）以达到维持正常经营的目的。相较于传统的有桩式公共自行车，共享单车具有极大地便利性优势与极低的准入门槛，因此一经投放市场，共享单车的用户规模便快速扩大，其本身也成为解决城市通勤“最后一公里”、缓解城市污染与交通拥堵等问题的重要解决方案。然而，基于共享单车骑行轨迹的人员与物质流动模型其实是一种耗散模型，与任何自然发生的能量耗散一样，共享单车网络对于城市通勤的贡献也始终随着一天中时间的推移自然下降到低水平，也就意味着原本相对集中的易于使用的共享单车分布会由于通勤等原因变得分散和难以使用。想要维持共享单车的合理分布及单车流动“借-还”循环畅通，就需要依赖基于通勤模式经验指导下的强制调度，这种调度一般发生在非通勤时间，共享单车服务公司使用卡车或平板三轮车搜集散乱在城市各处的共享单车并整齐放置到合理的地点以支撑第二天的骑行通勤需求。因此，由共享单车骑行轨迹数据中挖掘出的通勤模式具有一定的现实指导意义。

综上所述，本研究旨在探究城市中共享单车通勤的时空模式，为城市规划和交通管理提供科学参考，同时为共享单车服务的优化提供建议。

1.2 国内外研究现状

李文翔等人[1]从环保的角度，使用问卷调查法、统计分析法及核密度分析法等方法结合减排因子计算公式评估了2016年上海市全年共享单车对于六种主要污染物的减排效益及空间分布模式，指出上海市人口密集区域具有更大的减排潜力。王若萱等人[2]使用随机森林模型，结合一套基于兴趣点数据的目的地区域建模指标体系重点分析了上海市区域性共享单车通勤模式，为后续该领域内的模拟仿真系统提供了研究基础。全雨霏[3]从宏观大数据角度，综合考虑骑行轨迹的起点、终点及起-终连结的时空特征，采用统计分析、问卷调查、实地调研等方法研究了南京市共享单车通勤模式，分析出共享单车使用时段的峰值规律、空间热点分布规律及其影响因素（绿地分布、公共交通换乘站点分布等）并结合此为城市规划提出自己的建议。常新[4]对深圳市龙岗区时间跨度为半个月的共享单车轨迹数据进行回归分析，总结出骑行行为模式并反向推求城市环境对骑行行为的支撑作用，最后使用空间自相关理论分析原因，指出道路的繁忙程度对于骑行频次具有影响，并根据研究结果提出了合理的城市规划建议。谢国微[10]从天气与城市建成对骑行行为的影响出发，综合采用统计方法、地理加权回归模型(GWR)、核密度分析法、多元线性回归法等多种研究手段探究共享单车使用情况的时空模式及其影响因素。

随着数据驱动的智能算法逐步发展，也有一部分研究者注意到共享单车轨迹中蕴含的智能，他们使用新兴的数据驱动智能算法直接将数据本身转化为解决问题的能力，为交通预测、道路规划、大数据感知等应用场景提供助力。Jie Bao等人[5]基于上海市共享单车大数据记录开发出一套城市自行车道规划算法，该算法具有一个用于描述轨迹长度与收益关系的启发式函数，能够根据项目预算、施工周期等因素灵活调整自行车道设计结果，具有较高的工程实践价值。谢光明[7]为了解决共享单车时空供需不平衡问题，基于大量骑行轨迹数据研究出一套单车流量预测方法。他首先使用两级聚类及 K-means 算法基于密度找到虚拟站点，然后根据用户访问频次结合鲁汶(Louvain)算法得到虚拟站点间的社区网络最后使用图注意力网络(GAT)及LSTM网络并融合气象参数完成单车流量预测。该方法对于共享单车的动态调度，缓解单车分布时空不平衡具有重要的指导意义。刘冰等人[8]采用相似的研究思路，使用聚类及社区发现算法（鲁汶算法）研究基于共享单车的生活圈，他们发现骑行生活圈的空间范围相较于步行生活圈更大也更复杂，是居民日常通勤的基础性组成部分，合理地布置骑行网络对于便利居民生活具有重要意义。王俊等人[9]使用带有卷积层的长短期记忆网络（ConvLSTM），结合南京地区共享单车需求量，天气，季节等参数训练深度学习预测模型，对于需求量的预测精度相较于单一网络（卷积网络或长短期记忆网络）取得一定提升。

虽然共享单车是一种新兴互联网经济下发展起来的环境友好型便利交通设施具有诸多优势，共享单车市场快速膨胀也带来了许多城市治理难题，最为显著的问题之一就是大量损坏、报废共享单车的该如何处置。刘泉宏[6]等人重点考虑了共享单车生命周期末环的回收处理问题，基于武汉市共享单车报废点位采用先求取重心后进行 K-means 聚类的处理方法，计算出合理的共享单车回收站部署位置。该研究对于共享单车生命周期中后管理维护站点的选址提供了一定的技术支持。

上文提到的大多研究都是将共享单车轨迹视作一个整体，轨迹点视作点云，往往采用基于密度的空间聚类与统计方法，忽略了共享单车骑行轨迹作为矢量线段集合所内含的空间结构信息。任丹[11]使用 TRACLUS 算法船舶轨迹进行聚类，提取出公共航行轨迹，与一般的基于点密度聚类方式不同，该方法考虑到了轨迹数据本身作为向量的几何特征，对于共享单车骑行轨迹研究领域具有启发意义。

过分析城市共享单车领域的国内外研究成果，本文作如下总结与思考：

（1）研究细分方向：大致可分为污染物与环境、城市道路规划、模拟仿真算法、社交网络分析、热点路段、时空模式分析及城市道路及空间规划质量评估等。可以发现，共享单车骑行轨迹中蕴含极其丰富的信息，是研究城市、自然、人及其关系的绝佳窗口。

（2）研究数据及方法：除用户骑行轨迹数据外，研究者也会根据研究的细分方向补充部分数据，包括居民出行问卷调查数据、城市基础地理数据（公交站台、绿地分布等）、天气数据、共享单车服务设施数据（报废处理站、虚拟停车桩等）以及其他类型的轨迹数据（包括步行、出租车等）。研究方法大致可分为统计法、聚类法（基于点、基于轨迹线）、地理统计法、及数据驱动的智能算法等。可以发现，共享单车轨迹数据具有丰富的时空结构信息，需要酌情选择合理的分析方法。

（3）思考：骑行轨迹数据往往与城市路网分布、主要基础设施分布以及生活设施分布具有一定的关联性，尤其需要考虑其与主要道路间的关联。骑行轨迹目的地暗含了用户的真实意愿，也可以根据目的地类别对轨迹进行分类，实现更进一步的分析。用户的行为特征及单车流动特征也可以成为重要的分析抓手，例如调查一段时间内某一用户的骑行频率，某段时间内单车的周转率等。

1.3 研究内容与步骤

以往的研究大都以轨迹本身为研究重心，并以原因分析（考察天气等的影响）及副作用分析（考察减排效益等）为目的，并未充分考察骑行者的行为模式及轨迹与基础设施建的相互关系。因此，本文以上海市共享单车通勤用户为研究对象，以经预处理及清洗后的上海市共享单车骑行轨迹数据作为数据源，使用最邻近匹配算法、轨迹预处理算法（重排序、简化）、用户行为分析法、线密度算法、数学形态学操作、Jaccard相似系数、时空格网分析等研究方法，探究共享单车轨迹时空模式及其与通勤基础设施之间的关系。进一步的，本文为城市规划、共享单车管理等相关领域提出合理化的建议，以更好地促进上海市慢行交通系统的发展。本文的主要研究内容及研究步骤如下：

（1）共享单车通勤数据建库：首先获取2016年8月及2020年8月上海市某品牌共享单车通勤数据（每条轨迹带有时间戳及起止点，其中2016年数据带有详细的骑行轨迹）；其次，进行数据预处理与清洗。针对2016年骑行轨迹数据，设计并实现轨迹重排序算法（3.2.2）、轨迹长度量测算法（3.2.2）及轨迹简化（抽稀）算法（3.2.3）。而后，针对轨迹长度、骑行时长、骑行速度、字段缺损状况等因素设计合理的数据清洗规则，剔除异常值；最后，计算每一段轨迹的起止点的GeoHash值，并将2016年骑行轨迹数据保存为便于存取的二进制格式（\*.pickle），以便后续处理。

（2）上海市基础地理信息数据建库：首先从 Open Street Map 网站上下载上海市地理数据集（包含兴趣的、各级别路网等数据），并对数据进行裁剪及清洗；对于兴趣点数据，仅保留八类意义明确类型的数据；对于路网数据，仅提取主路、次主路及支路三个级别路网。然后，将兴趣的数据转化为 GeoHash 格式，并构建树状空间索引，以便快速查找最邻近的兴趣点；最后将空间索引及数据集保存为易于存取的二进制格式。

（3）共享单车通勤时间模式分析：借助预先建立的通勤数据库，从时间角度探究骑行通勤模式，分别从轨迹时距分布模式、单车日周转趋势，用户骑行时间间隔模式角度对用户骑行通勤模式进行分析，并探究背后的可能原因，总结相关规律。

（4）共享单车通勤空间模式分析：综合通勤数据库及上海市兴趣点数据库，从空间角度考察轨迹的空间分布模式。首先使用轨迹终点最近邻匹配算法（3.3.1）为每一段轨迹终点匹配最邻近兴趣点，并据此分析轨迹终点的类型占比及空间分布模式。然后使用线密度栅格来具体分析通勤轨迹空间分布模式及其与路网间的相似关系。最后根据潮汐指数来考察共享单车潮汐区域空间分布，并总结相关规律。

1.4 论文组织结构

本文共分为五章：

第一章为绪论。主要从宏观角度论述本文的研究背景及研究意义，并对国内外共享单车轨迹分析、城市通勤分析等领域的研究成果进行总结概括为国内外研究现状，参考国内外研究现状确定本文研究内容及技术路线，为后续研究明确方向，打下基础。

第二章介绍理论基础与相关技术。该部分首先从学术理论入手，介绍了行为地理学、城市职住平衡与过剩通勤理论，进一步明确所研究问题的理论本质。然后从研究技术着手，介绍了相似性分析、密度分析、用户行为分析等领域内常用的技术手段。

第三章介绍本研究用到的数据及开展研究时所用到的具体方法。首先介绍数据来源及数据细节，并介绍数据清洗规则,而后结合数据特征，介绍轨迹点重排序、轨迹长度量测、轨迹抽稀、数学形态学操作及Jaccard相似系数等数据处理分析算法。

第四章着重介绍本文的研究结论。从时、空两个方向，结合轨迹时距分布特征、目的地分布特征、综合使用线密度计算、数学形态学操作等方法，得出骑行热点时段分布、骑行热点路段分布、各级别道路对骑行通勤支撑情况。

第五章为总结与展望，结合本文研究内容提出概括性结论，针对研究过程提出不足与未来进一步完善的方向。同时，结合得到的上海市骑行通勤时空分布特征，为城市规划建言献策。

1.5 本章小结

本章首先介绍本文的研究背景及意义，指出研究共享单车通勤模式对于理解及优化以共享单车为代表的城市慢行交通系统具有重要意义。而后，本文考察了国内外研究现状，提出相关总结与思考，并进一步介绍本文的主要研究内容与步骤框架。最后，该章节对本文组织结果做出简要介绍。

**第2章 相关理论基础**

2.1 相关理论基础

2.1.1 行为地理学

行为地理学是人文地理学的一个重要分支[12]，该学科在 20 世纪初期诞生，是地理学、心理学、行为科学的交叉学科，在后来的发展中逐步融合行为主义地理学与时间地理学两大理论基础，形成了一套相对完善的研究范式与理论体系。该学科主要探讨人与地理环境之间的互动，强调“以人为本”，从微观人地关系的研究出发探索宏观时空规律。受到行为经济学中将非理性的经济主体作为微观研究的对象启发，行为地理学所研究的“人”是一种有限理性的个体，这些个体的行为虽然受到地理环境等的客观制约而呈现出一定规律性，但也会出于主观做出一些随机行为。个体可以依据属性汇总为群体、可以依据社交网络与空间活动范围汇总为社群，这些不同类别不同层次的群体同样被归为“人”这一研究对象。行为地理学所研究的“地”指的是一种抽象的、主客观相结合的行为空间，比如城市意向、认知空间等。行为地理学强调人地关系的统一性，人与地相互依赖、相互影响、相互塑造，这二者是不可分割的。

与传统的将研究区域视作静态的“块”不同，行为地理学的研究强调动态性，这种动态性体现在：1）其研究图景是一个具有复杂时空流动模式的网络：个体与个体之间、个体与环境之间均存在复杂的互动关系，甚至这些关系本身也相互制约相互影响，这种有机的网络模型所具有的动态性是传统的静态区域块模型所无法比拟的；2）研究视野综合了宏观与微观并可动态调整：通过汇总个体行为，总结区域规律来不断扩大研究视野。通过不断缩小研究视野，深挖“人”这一复杂个体全生命周期内在认知的流变规律，更好地关怀个体生命质量；3）数据及对应处理方法层面的动态性：移动互联网、GPS 和移动位置服务(location-based services，LBS)等技术提供实动态数据支撑，人工智能、大数据技术等先进的数据处理手段提供快速动态的分析能力。

随着时代的发展，“人本主义”日益得到社会各界的认同，着重研究人的日常行为及其与环境的交互关系的行为地理学也越发受到重视。行为地理学以其独特的研究视角，在可持续发展、社会公平、国土空间规划与社会治理等方面发挥出越来越大的作用。

2.1.2 城市职住平衡与过剩通勤理论

随着中国城镇化水平不断提高，城市交通拥堵问题日益凸显，其背后城市交通巨大供需失衡问题也逐渐被社会各界所认识。对于城市居民而言，住所与工作地点相距甚远往往会带来巨大的通勤压力，这也是城市交通拥堵的一个重要原因。研究并达成职住平衡可以有效缓解城市交通拥堵，职住平衡理论也可以为城市规划提供有用的指导。职住平衡理论（Jobs-Housing Balance）[13]最早由 Howard 在1902年出版的书籍《明日田园城市》中提出，他认为应当建设一种具有“自力性”（Self-contained）的社区，这种社区能够确保居民的就业需求与住房需求一并得到满足，最终由众多具有“自力性”的社区组成“田园城市”。这样的“田园城市”可以有效缓解交通拥堵及环境污染，减轻居民通勤压力，缩减中低收入家庭开支提升幸福感。

将多种城市功能理想化地集成在许多相互离散的小块土地上，这种理想化的城市模式在现实中往往难以实现。“职住分离”的现状是市场长期自主选择的结果，一方面企业追求产业聚集带来的规模效益，企业以零散的方式分布于小片地块会降低生产效率提高管理成本，另一方面，求职者更愿意为了高薪前往较远的工作地就职，随着城市公共交通系统的发展，通勤成本不断下降，这种职住距离还会被进一步拉远。另外，对于一个具体的家庭而言，求职并不是选择居住地唯一需求，医疗保健、子女求学及休闲放松等综合需求同样重要，尤其是优质的公共服务资源（重点学校、三甲医院等）往往成为一个家庭选择居住地的主要因素，单纯强调职住一体并不能很好地满足家庭的综合性需求。因此，需要客观看待城市职住分离的现状，着重研究城市通勤现状，过剩通勤可以作为衡量一个城市职住平衡的量化指标，较为客观地反映城市通勤现状。

过剩通勤（Excess Commuting）指的是实际通勤距离与理论通勤距离之间的差值，该理论最早在1982年由 Hamilton 首先提出，经过三十多年的发展逐步形成了一套较为完善的理论体系。在城市建模方面，该领域最早将城市建模为单中心模型，认为城市中心提供了所有的就业机会。后来随着计算机技术及地理信息系统的发展，有学者使用泰森多边形将城市划分为多个职住格网，更精细地建模城市职住空间分布情况。在理论通勤距离计算方面，主要计算方法包括：1）交通问题线性规划最优化函数（Transportation Problem in Linear Programming, TPLP）：采用线性规划的方法，以交通成本最小为优化目标，职住人数分布守恒为约束条件求解理论通勤距离；2）最大熵值法与蒙特卡罗仿真模型：最大熵值法是一种基于信息论的随机模拟方法，计算简单，不需要大量数据。蒙特卡罗模拟是一种基于随机抽样的统计方法，灵活性高，计算量大，结果受随机抽样影响；3）等比例匹配通勤：认为就业岗位分配与企业所占区域在区域中的比例相同。其中，TPLP 的其基本原理是将城市中的就业岗位和居住地视为节点，将通勤视为弧线，并以交通成本最小化为目标，求解所有居民的通勤路径和通勤距离。

其中， 为最小平均通勤距离， 为城市通勤总人口， 为城市居民总数， 为城市就业岗位总数， 为居民 与就业岗位 之间的通勤成本（一般为距离）， 为居民 与就业岗位 之间的通勤量（一般为人数）。

上述公式需要满足如下约束条件：

其中， 为居民 的出发地， 为就业岗位 的目的地。

居民实际通勤情况可以使用问卷调查、GPS 轨迹数据、公交刷卡数据等方式获取，通过计算实际通勤距离 结合理论通勤距离 可以计算出过剩通勤率 ：

2.2 相关研究技术

2.2.1 相似性分析

相似性分析指的是依据某种相似性度量指标，计算并判别两个或多个对象间的相似程度，常用于医学、市场营销、犯罪学、计算机视觉等领域。地理学领域内的相似性分析一般强调对象间空间分布的相似性，有些研究也会同时考虑时间维度的相似性，即对象间的时空分布相似性。相似性分析的核心是相似性度量指标，由于这些指标一般针对高维数据间的度量设计，空间数据若不考虑属性至多只有3个纬度，所以也可以直接套用多数指标。常用的相似性度量指标有：欧氏距离、曼哈顿距离（3.2.2）、余弦距离、Jaccard 系数（3.3.3）等。

但是，时空数据（尤其是时空轨迹数据）有其自身的特殊性，它是以时间顺序记录的移动对象位置序列，具有动态性，即同一个对象在不同时刻具有不同位置属性。这与一般相似性分析的静态数据（如带有诸多静态属性数据的人或物）具有较大差别。因此，也存在一些细致考察了轨迹的时空结构特征的相似性度量算法：1）轨迹间欧氏距离：将待分类轨迹视作随时间变动的点集，在每一个时间断面计算轨迹点间的欧氏距离并在全局时间轴方向上累积。对噪声敏感，需要采样频率一至，否则会有较大误差。2）最小外包络距离：首先使用轨迹特征点（方向快速变化的点）切分轨迹，然后提取这些子段的外包络矩形并组成顺序序列作为轨迹的概化表征，最后累计矩形间的距离以度量轨迹间的相似性。该方法对原始轨迹平滑去噪，但也会丢失一部分信息。3）子轨迹相似性度量[11]（TRACLUS）：该算法首先根据最小描述长度原则将轨迹划分为多个子轨迹，然后计算子轨迹间的三种距离（垂直距离、平行距离、角度距离）并加权计算的综合相似性量值。4）历史最近距离：在一段给定的时间范围内，检查各个时刻两个轨迹之间的距离（欧氏距离或其他），选取最小距离作为历史最近距离。

相似性分析可以揭示大量数据间的多种共性特征，因此也是其他更深一步的数据分析方法（聚类分析）的重要前置步骤和关键组成部分。时空数据由于有时需要考虑细致的时空结构，所以需要设计相应的特殊相似性度量算法。

2.2.2 密度分析

密度分析是一种空间分析方法，用于计算和可视化点要素或线要素的集中程度以及分布模式，其中，密度是指单位面积内的要素数量（点的个数或线的长度）。密度分析的本质是将离散的点要素或线要素数据转换为连续的密度表面。密度表面可以直观地反映出要素的分布模式，并为进一步的空间分析提供基础。根据计算方法的不同，密度分析可大致分为以下几类：1）简单点密度分析: 该方法将每个点的测量值除以其邻域的面积来计算密度。邻域可以是固定大小的圆形或方形区域，也可以是某种自定义形状。2）距离加权密度分析：在计算密度时，考虑要素到参考点的距离，根据距离赋予权值。3）核密度分析: 该方法使用核函数来计算输出栅格像元周围的点要素的密度。

线密度分析与点密度分析具有类似的思想，即统计一定邻域内要素分布情况。线密度一般定义为单位面积内累计线要素的长度，使用圆形作为邻域来统计线要素密度。线密度计算有时需要考虑权重，即不同线要素对密度的贡献不同。线密度可用于野生动物迁徙路径、城市管网、城市道路等领域的分析，基于其生成的密度栅格还可进行相似性分析、路网提取等进一步分析。假设有两类线要素，分别用L1和L2表示，其权重分别为V1和V2，搜索半径为R，则线密度D可以表示为：

2.2.3 用户行为分析

用户行为分析（User behavior analytics）是一种通过分析用户使用数据来了解用户行为模式的方法。企业可以收集用户点击次数、登录时间、浏览时长等数据，并通过数据挖掘、模式分析等手段进行分析。用户行为分析可以帮助企业在不打扰用户的情况下（相较于发放用户问卷）更好地了解用户群体，从而提升服务质量。此外，高效及时的用户行为分析还可以防范恶意攻击（例如DDoS攻击），过滤异常账号（刷单诈骗）。一些常用的用户数据挖掘方法包括统计用户使用频率并划分用户等级、计算并绘制用户贡献累计分布曲线（CDF曲线）、计算用户留存率等。其中，用户留存率（User Retention Rate）是一个用于衡量产品吸引力及用户粘性的重要指标，它指的是在某一时间段内开始使用产品或服务的用户，经过一段时间后仍然继续使用该产品或服务的用户的比例。

夏琼燕等人[14]下载并研究了2005年到2007年 Open Street Map （OSM）开源地理信息网站中志愿者用户数据（记录了用户的贡献行为），通过计算用户贡献累计分布曲线，他们发现该网站的贡献者符合“二八定律”，即该网站两成的贡献者贡献了约80%的数据，并根据贡献率将用户分为多个活跃等级。通过合理选取时间窗口，计算用户留存率，结合用户活跃度分类，他们发现OSM社区中低活跃度的用户留存率偏低，而高活跃度用户总体保持稳定，社区稳定存续主要依靠中高活跃度用户。

2.3 本章小结

本章主要内容为相关理论概念及研究技术。相关理论部分首先介绍行为地理学，探讨人与地理环境灵活多变的互动关系，指出“以人为本”的思想观念越来越受到社会各界重视。而后，该部分开始探讨职住平衡理论，该理论认为职住空间应当尽可能均匀地分布于城市，然而绝大多数城市的实际情况却是“职住分离”与“产业聚集”，为了客观地考察城市职住动态分布情况，人们从城市通勤入手提出了过剩通勤系列指标。相关研究技术部分主要从空间分析及行为分析两个方向入手，空间分析方向主要介绍了常见的聚类分析及密度分析方法，行为分析方向主要介绍了用户粘性分析等几种常见的用户数据挖掘方法。

**第3章 研究数据与方法**

3.1 研究范围

本文选取上海市作为研究区域。上海市北界长江，南枕杭州湾，西接江浙两省，东临东海，全市面积 6340.5 平方公里，现辖16个市辖区：黄浦区、徐汇区、长宁区、静安区、普陀区、虹口区、杨浦区、闵行区、宝山区、嘉定区、浦东新、金山区、松江区、青浦区、奉贤区、崇明区。参考第七次全国人口普查主要数据，上海市常住人口为24870895人。近几十年来经济飞速发展的同时，上海市逐步形成了多中心、高人口密度及高混杂度的城市格局，这也进一步引起人口拥挤、交通拥堵和环境污染等大城市病。这些城市特征也同时存在于中国大多数大型、超大型城市中，选取作为中国经济中心的上海为研究区域具有一定的典型意义及泛化意义。

3.2 数据来源与预处理

3.2.1 数据来源与数据清洗

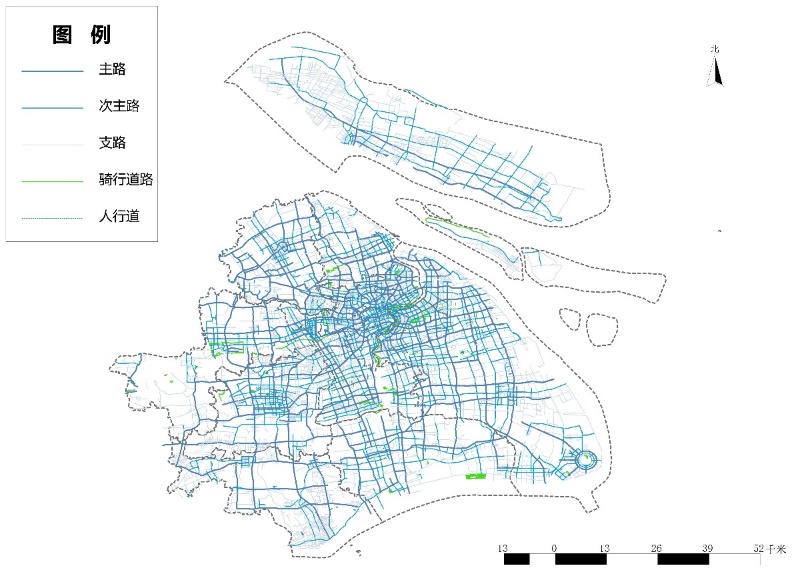
本文的研究数据包括骑行轨迹数据与上海市基本地理信息数据，其中骑行轨迹包括由互联网采集整理得到的 2016 年 8 月整月 102361 条上海市某品牌共享单车骑行轨迹数据及 2020 年 8 月整月 12793 条骑行记录，上海市基本地理信息数据包括由 Open Street Map 开源地理信息数据网站获取上海市路网矢与兴趣点分布矢量数据及由高德地图开放平台获取的上海市矢量边界数据。

2016 年骑行记录数据较为详细，除了常见的骑行起止时间及起止点位置字段外，还包含了一个以“#”分隔的骑行轨迹点字段，但是由于网络延迟、设备故障等原因该字段内的轨迹点为乱序排列，需要设计算法以恢复正确的骑行轨迹，并且骑行实际距离也需要根据正确的轨迹数据计算（见 3.2.2）。2020 年骑行记录并未包含详细的骑行轨迹点集，但提供骑行距离。对于骑行记录数据，按照如下规则清洗：1）去除骑行时长小于1分钟大于 8 小时的骑行记录；2）对于有轨迹点集的数据，去除记录到的轨迹点数量小于 3 个的记录；3）去除存在字段缺失及异常符号的记录；4）去除骑行终点位于上海市之外的记录，该记录超出研究区域。

表3.1 轨迹数据示例

|  |  |  |
| --- | --- | --- |
| 字段 | 说明 | 字段值 |
| ID | 编号 | 1 |
| BID | 单车编号 | 79699 |
| UID | 用户编号 | 759 |
| ST | 开始时间 | 2016/8/1 0:23 |
| SX | 起点经度 | 121.52 |
| SY | 起点纬度 | 31.309 |
| ET | 结束时长 | 2016/8/1 0:32 |
| DU | 骑行时间 | 9 |
| EX | 终点经度 | 121.525 |
| EY | 终点纬度 | 31.316 |
| WD | 星期 | 星期一 |
| track | 轨迹点 | 121.520,31.309#... |

从 Open Street Map (OSM) 网站下载上海市区域的 OSM 数据，该数据为基于 XML 格式的矢量数据，包含道路、建筑物、兴趣点等地理要素信息，并以节点、关系和区域等形式组织。使用 QGIS 软件解析 OSM 数据，将其转换为常用的 Shapefile 文件格式 (\*.shp)。使用上海市矢量边界数据裁剪 OSM 数据，去除区域外的部分。从预处理后的 OSM 数据中抽取出五个级别道路（主路、次主路、支路、骑行道路、人行道）共计 46445 条数据，考虑到骑行道路及人行道较为稀疏，故最终研究仅选取主路、次主路、支路三个级别的道路数据进行研究。主路一般连接城市各个分区，红线宽度为30至45米；次主路起辅助主路的作用，一般连接主路并作为分区内的主路，红线宽度为30至40米；支路一般作为生活道路，连接各个街坊片区，红线宽度为12至15米。此外，还从 OSM 数据中提取了相关兴趣点数据集。



**图3.1 上海市分级路网数据**

从上海市兴趣点数据集中抽取出八个类别的兴趣点集（交通设施服务、体育休闲服务、公司企业、商务住宅、科教文化服务、购物服务、风景名胜、餐饮服务）共计 122608 条数据。

表3.2 兴趣点类别及编码

|  |  |  |
| --- | --- | --- |
| 类名 | 英文名 | 编码 |
| 交通设施服务 | Transportation | 0 |
| 体育休闲服务 | Sports and Leisure | 1 |
| 公司企业 | Company | 2 |
| 商务住宅 | Business Residence | 3 |
| 科教文化服务 | Science and Education Culture | 4 |
| 购物服务 | Shopping | 5 |
| 风景名胜 | Scenic Spots | 6 |
| 餐饮服务 | Catering | 7 |

3.2.2 轨迹重排序及长度量测

2008年 Gustavo Niemeyer 提出了 GeoHash 地理编码系统[16]，它能将地理位置编码成由字母和数字组成的短字符串。GeoHash 是一种多层级的空间数据结构，通过网格状的划分及 Z 形空间填充曲线 (Z-order curve)方式将空间进行编码。GeoHash 编码精度可以通过增加编码字符长度无限拓展，同时可以通过去除编码末尾的字符来降低精度 (空间精度随之降低)。

由于其编码结构具有层级性，GeoHash 保证了两个共享前缀越长的编码，其代表的地理位置越接近；由于其采用具有空间突变特性的 Z 形空间填充曲线，GeoHash 无法保证距离很近的两个地理位置一定具有相同的编码前缀，且具有多位共享前缀的编码实际距离也有可能很远（多发生与边界处）。这种前缀匹配程度并不完全等同于地理距离的接近程度的现象被称为距离失真。

计算GeoHash首先需要将经纬度坐标分别转换为二进制表示。然后按照“经度值占偶数位，纬度值占奇数位”的规则将经纬度交织为一个二进制串，最后使用 Base32 编码转换为字符串。GeoHash算法的计算过程可以概括为以下公式：

其中，Interleave 函数表示经纬度二进制值的交织操作，Base32 函数表示二进制值转换为 Base32 编码的操作。

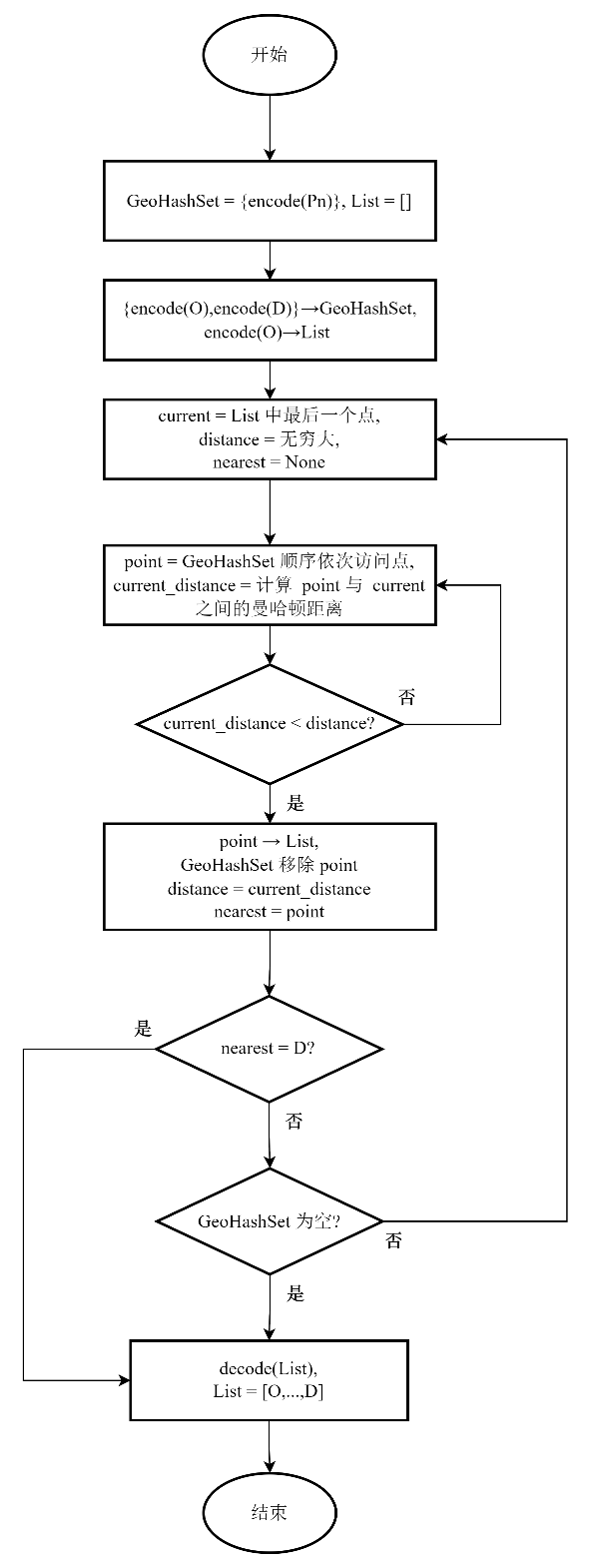


**图3.2 GeoHash算法流程**

曼哈顿距离由纽约市曼哈顿的棋盘式街道布局而得名，定义为两个点的坐标轴投影长度之。在二维平面上，两点 (x1, y1) 和 (x2, y2) 之间的曼哈顿距离为：

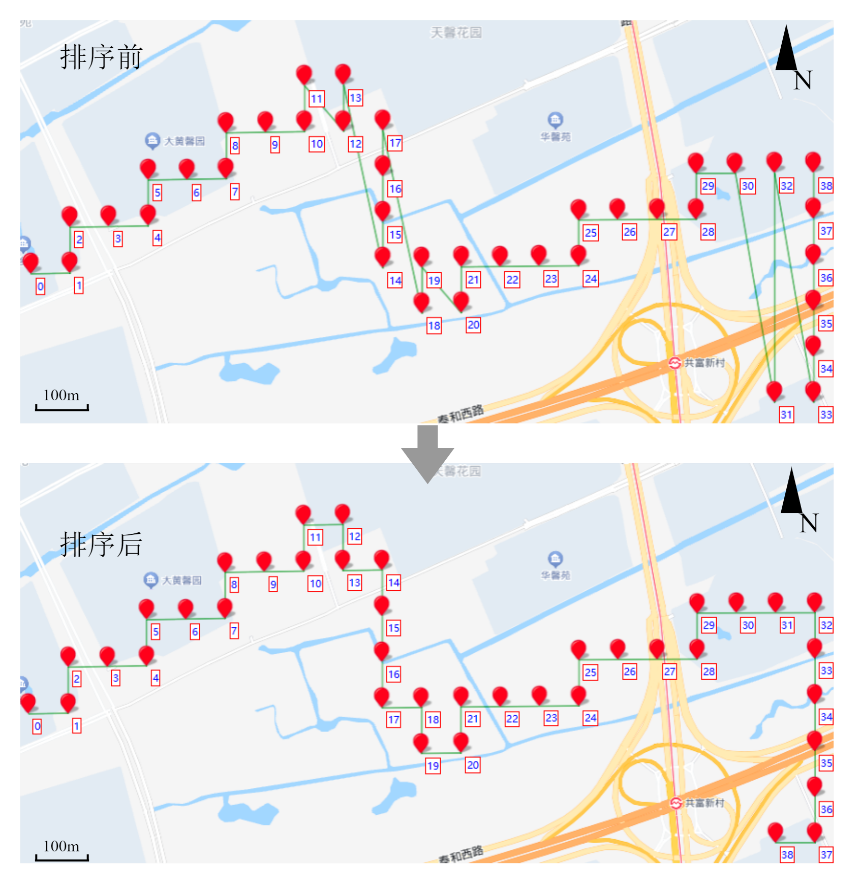
考虑到轨迹点密度较高，相邻点之间的距离很短，且经纬度变化范围较小（绝大部分相邻点之间的经纬度跨度不超过0.001°），因此曼哈顿距离能够有效捕捉轨迹点的局部运动特征。且曼哈顿距离的计算公式简单，仅涉及绝对值运算，计算效率高。对于具有大量轨迹点的GPS数据，使用曼哈顿距离可以显著降低计算成本。因此，本算法通过计算两个坐标点之间的曼哈顿距离来判断两个点之间实际的空间距离。

对于乱序轨迹点集，在知道始末点的前提下，本文提出了一种基于 GeoHash 空间索引及曼哈顿距离及贪婪策略轨迹重排序算法:认为所有记录中的轨迹点集都是乱序状态，对于任意一轨迹点集，首先创建空列表 List 并将包含了起点和终点的所有点组成的列表 Pn 逐点计算 GeoHash 编码并放入集合 GeoHash 中（该步骤同时借助集合的特性实现了去重复）。然后，循环迭代，直到轨迹点列表为空。在每次迭代中，算法会计算当前点与轨迹末尾点的曼哈顿距离，并找到距离最近的点。如果找到的最近点就是终点，则算法终止，否则将最近点加入轨迹并从 GeoHash 集合中删除。最后，将轨迹列表 List 中的 GeoHash 编码转换为坐标并返回简化后的轨迹。



**图3.3 轨迹重排序算法**

由于 Python 编程语言的集合（Set）原生的散列值计算函数作用于元组（tuple）类型的数据时会发生二进制精度损失，所以需要首先使用 GeoHash 算法对位置元组进行编码（encode）再将得到的字符串存入集合中，在需要用到实际坐标时再使用对应的解码方法（decode），这样可以规避使用存入及去除集合时由于散列值计算带来的精度损失。考虑到GeoHash编码存在距离失真的问题，本算法使用两经纬度坐标间的曼哈顿距离来估算实际距离。以下是对某真实轨迹进行排序的实际效果：



**图3.4 轨迹重排序效果示例**

假设地球是一个球体，球面半径为 r（取6371000米），用经度（lon）和纬度（lat）表示球面上的点，则在已知有序骑行轨迹 n 个点组成的列表()的情况下，可以使用 Haversine 算法计算得每字轨迹的长度再累加得总轨迹长度。

使用上述方法对一段轨迹进行处理，计算得轨迹起点与终点间的直接半正弦距离为 2118.25 米，未排序轨迹点集的累计半正弦距离（认为地球半径r为 6371000 米）为 8685.89 米，经过算法重排序后的累计半正弦距离为 3605.04 米。可以发现，由于未经处理的轨迹点集中存在部分不合理的乱序点，导致轨迹总长度明显偏大，经算法重排序后的轨迹点集较为合理。

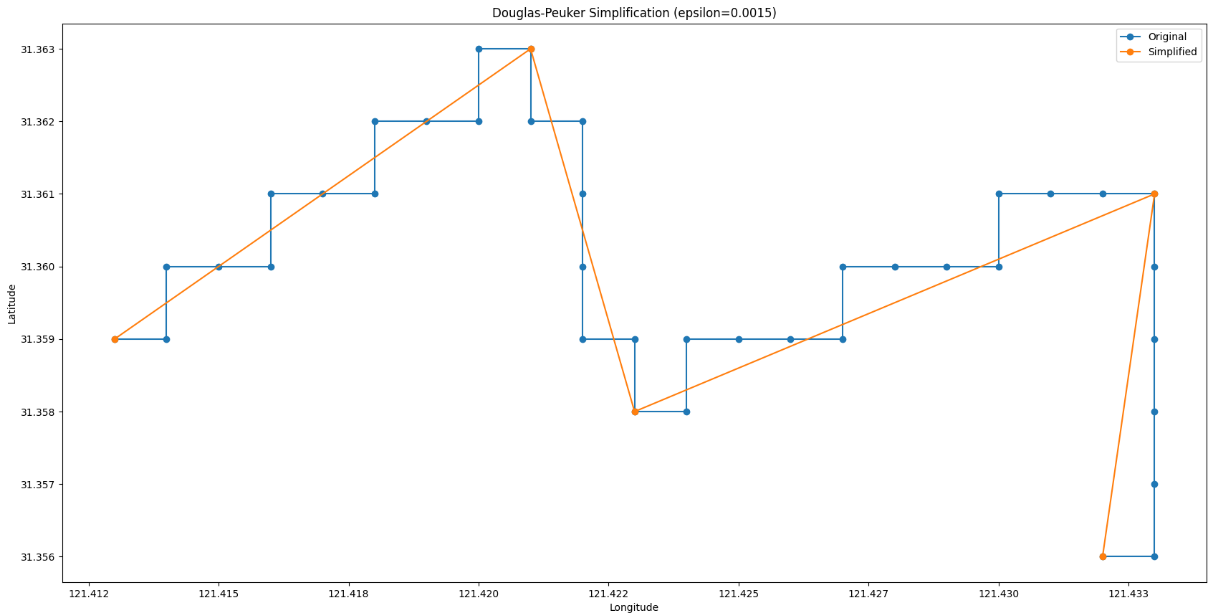
在中央处理器为 i7-10870H，机带RAM 16GB 的 Windows11 笔记本上，使用该算法处理十万条轨迹数据总耗时约1小时31分。

3.2.3 轨迹简化

共享单车位置记录设备通常以固定频率记录并上传数据，导致即使经过轨迹重排序，仍存在大量冗余点，例如，平直路段可能出现过于密集的轨迹点，增加后续分析负担。此外，由于定位精度限制，用户沿着某些路径运动时，轨迹会呈现锯齿状噪声，这也会影响后续分析，尤其是对于轨迹聚类，这种曲折的轨迹会严重影响聚类质量，导致最终拟合轨迹出现剧烈的波动。

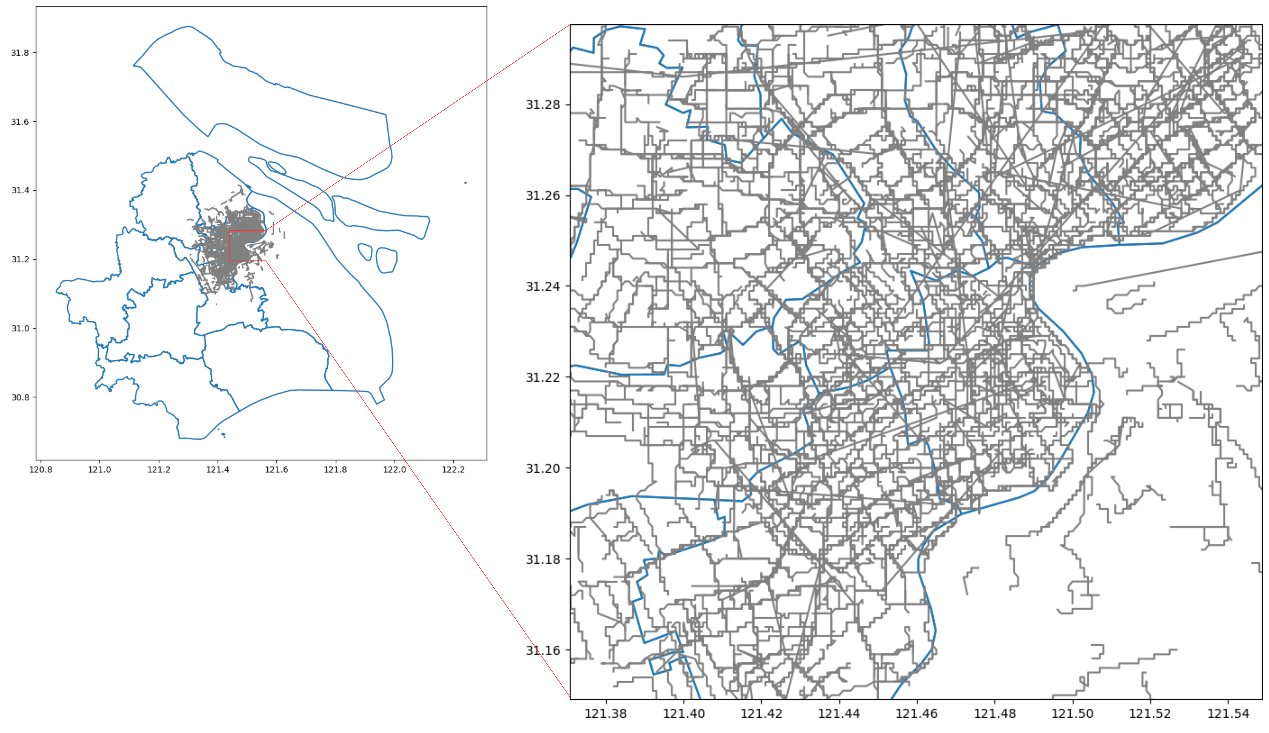
为了解决上述问题，本文采用道格拉斯-扑克算法简化用户骑行轨迹。道格拉斯-扑克法是一种迭代逼近的线抽稀算法，主要算法思想是基于给定阈值保留系列关键点以近似地表示原始曲线，该算法在地理信息系统（GIS）中被广泛应用于简化地图数据、轨迹压缩等方面，以实现在保持曲线的几何特征的前提下，减少数据存储和处理的开销。

具体算法步骤如下：1）首先，从给定的曲线中找到与起始点和结束点之间距离最大的点。2）如果这个最大距离大于给定的阈值n，则将曲线在这个点处切分成两段，并对每一段分别递归地应用该算法。3）如果最大距离小于等于阈值n，则直接将起始点和结束点作为结果返回。4）递归过程中，将所有满足条件的点连接起来形成近似曲线，最终得到简化后的曲线。



**图3.5 轨迹简化序效果示例**

本文对重排序后的轨迹数据进行阈值为0.0015的抽稀操作，挑选数据集中8月1日全天的轨迹数据绘制成图，对比抽稀前后可以发现通过将阈值设置为0.0015，我们可以有效地去除大部分冗余点以及由于低定位精度引入的锯齿状噪声，同时保留轨迹的整体形状。另外，在经过线简化步骤后，可以发现轨迹分布更为集中，某些空间分布规律已经有所显现。



**图3.6 未化简轨迹**



**图3.7 化简后轨迹**

这一优化措施显著减少了线段的数量，从而降低了后续分析的计算复杂度。此外，消除了轨迹中的锯齿状噪声也提高了轨迹数据的质量，使得后续分析结果更加可靠和准确。

3.3 研究方法

3.3.1 轨迹终点最近邻匹配算法

骑行轨迹数据缺乏对骑行目的的直接标注，限制了对其深入分析。本文提出一种基于兴趣点数据的骑行终点分析方法，通过将骑行终点匹配到最近的兴趣点类别，推断骑行者的行为目的。该方法可有效识别购物、通勤、休闲等行为模式，为共享单车服务优化提供依据。

对于某一段轨迹的终点，可以提取距离其最近兴趣点的类别，这种查找最邻近点并赋值的思想是本算法的指导思想。但是，本文使用到的上海市兴趣点数据集有十二万多条数据，轨迹数据在清洗后也有近十万条。如果直接对所有点进行暴力搜素，算法时间复杂度约为 ，粗略估计将执行120亿次半正弦距离计算（haversine）。据本机测算，执行一次 haversine 算法大约消耗 15 μs，粗略估计，在不考虑排序操作的情况下，算法将执行超过三小时，因此需要建立用于支持高效查找的空间数据结构。考虑到 GeoHash 编码的层次结构，本文构建了 GeoHash 树索引，支持高效的兴趣点查找。GeoHash 树是一种基于前缀树的树形索引结构，能够快速筛选出与目标编码字符串相似度较高的兴趣点，本文针对 GeoHash 编码字符串实现了这种树形索引，该索引支持插入（insert）及查询（query）特定的编码字符串。

该算法首先对兴趣点数据集构建树形索引（GeoHashTree），而后对某一个轨迹终点执行逐步扩大搜索空间范围的贪心算法（greedyQuery）。贪心搜索算法对输入编码字符串首先执行一次搜索操作，若未得到返回值，就会逐个去除最后一位（扩大空间范围）并再次搜索，直到找到最邻近的兴趣点（或者待匹配编码为空）。该算法的流程如下：



**图3.8 最邻近匹配算法流程**

高效的空间索引大大加快了最邻近查找的速度，经试验，在十二万条兴趣点数据集中使用该算法对近十万条轨迹数据的终点进行最邻近匹配共耗费约两分钟。

3.3.2 数学形态学

数学形态学（Mathematical Morphology）是一门建立在集合论和拓扑学基础上的图像分析学科，最早于1964年由法国科学家马瑟荣（G. Matheron）和赛拉（J. Serra）提出。它采用数学集合来描述图像对象，并利用集合间操作（如交、并、补等）来处理和分析图像，主要用于图像处理领域。数学形态学的基本思想是用具有一定形状的结构元素去探测和分析图像中的目标，通过对图像与结构元素进行一系列的运算来提取图像中的几何特征。数学形态学具有非线性、结构化等特点，在图像去噪、边缘检测、特征提取等方面具有独特的优势。数学形态学的基本操作包括：

腐蚀（Erosion）：用结构元素对图像进行腐蚀，使图像中前景像素的数量减少。数学上，图像A的腐蚀B，记为，定义为：

(3.4)

膨胀（Dilation）：用结构元素对图像进行膨胀，使图像中前景像素的数量增加。数学上，图像A的膨胀B，记为，定义为：

(3.5)

开运算（Opening）：先对图像进行腐蚀，然后再进行膨胀。数学上，图像A的开运算B，记为，定义为：

(3.6)

闭运算（Closing）：先对图像进行膨胀，然后再进行腐蚀。数学上，图像A的闭运算B，记为，定义为：

(3.7)

通过选择合适的结构元素及数学形态学操作，可以降低栅格数据中的噪声，在保持原有线状目标整体形状不变的情况下剔除冗余栅格连通细小缺口，从而有效提高栅格数据质量。

3.3.3 Jaccard相似系数

Jaccard相似系数（Jaccard similarity coefficient）定义为两个样本交集大小与并集大小之比，适用于有限样本数据间的相似性度量。Jaccard相似系数的数学定义为：

Jaccard相似系数的取值范围为 [0, 1]，值越接近1，表示两个栅格数据越相似。特定大小的栅格数据是一种典型的有限样本数据，所以可以使用Jaccard相似系数来度量两个栅格数据之间的相似性。对于栅格数据，可以在对其进行二值化操作后使用布尔运算来计算二者间交集和并集，从而快速计算Jaccard相似系数。

3.4 本章小结

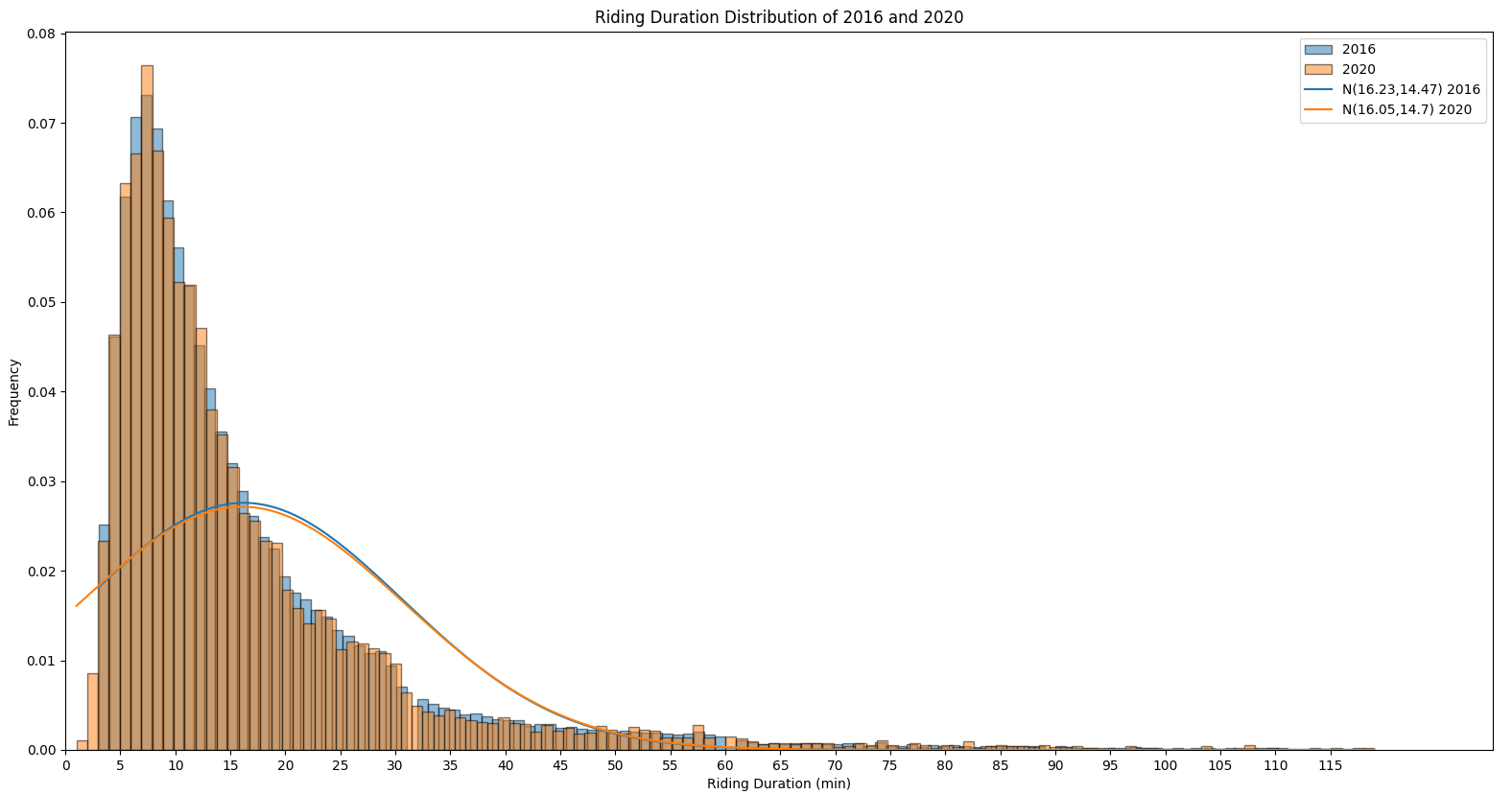
正文

**第4章 结果与讨论**

4.1 时间通勤模式分析

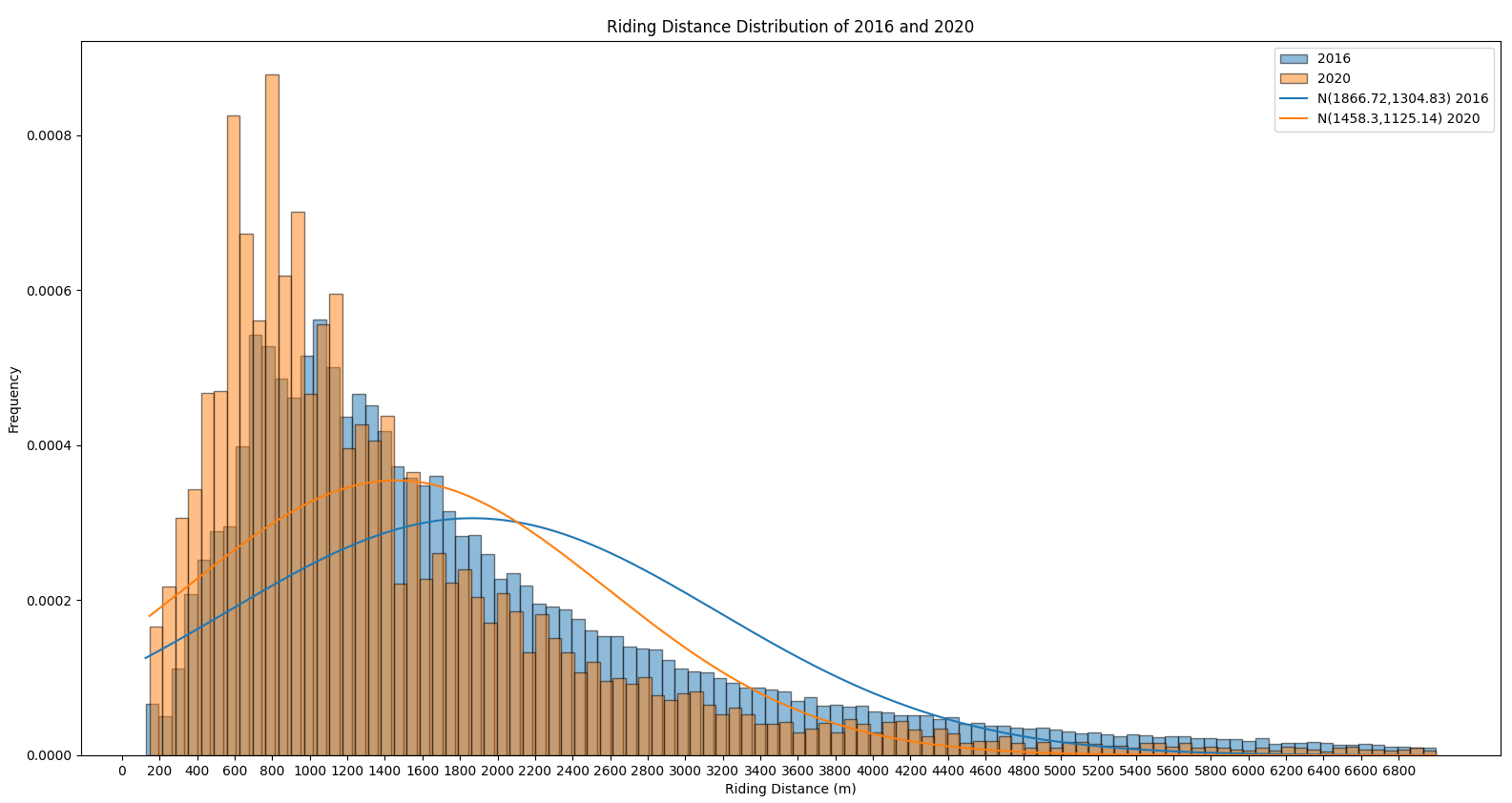
4.1.1 通勤时距模式

对共享单车通勤时长进行汇总分析，计算得每一个骑行时长区间内的频率，并拟合出正态分布曲线。频率直方图显示数据的直观频率分布情况，拟合得出的正态分布曲线显示数据趋势及统计学分布。可以发现，共享单车平均通勤时长为16分钟，骑行时长8分钟频率最高，极少有人选择骑行一个小时以上。对比2016年与2020年数据，总体分布及趋势保持稳定，骑行平均时长略有下降。共享单车在多数情况下还是一种短途、短时间的通勤工具。



**图4.1 骑行时长分布**

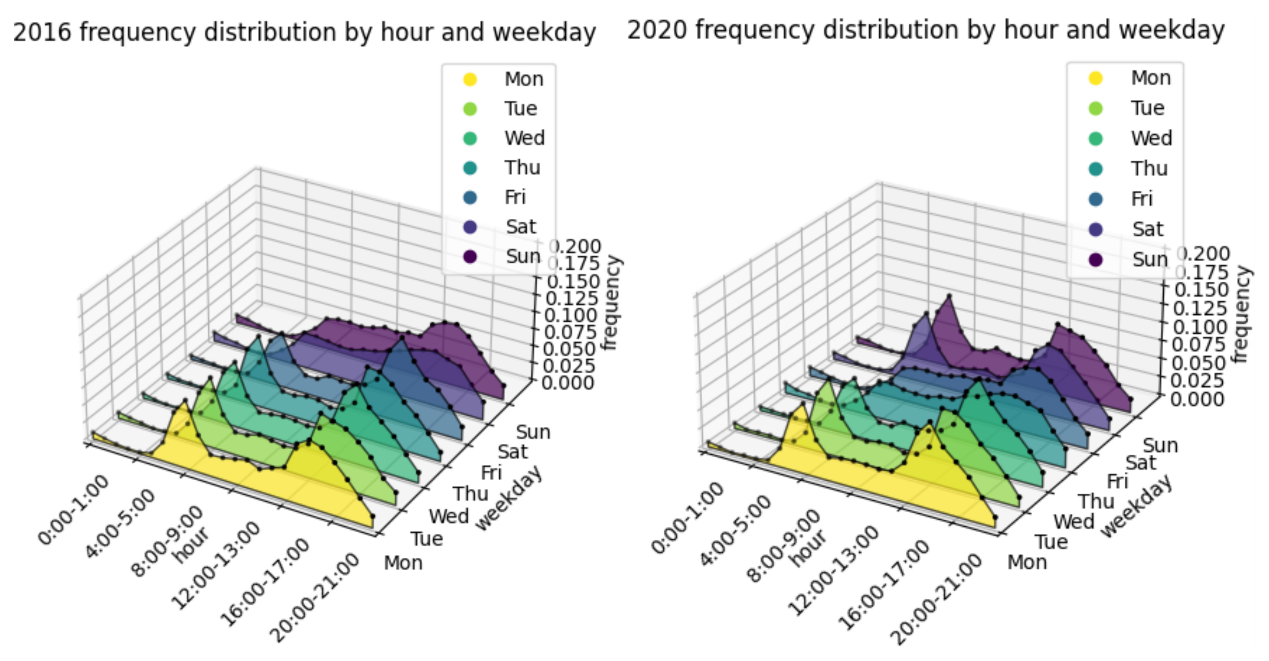
对共享单车通勤距离进行汇总分析，绘制出频率分布直方图及对应的拟合正态分布曲线，可以发现2016年平均骑行距离为1.8公里，而2020年平均骑行距离仅为1.4公里。2020年数据高度集中在小于1公里的范围内，数据标准差更小，分布也更为紧凑，0.8公里左右的短途骑行成为最多人的选择。说明人们越来与认可单车是解决“最后一公里”通勤问题的有效交通工具，共享单车市场也逐步积累了一批有稳定短途通勤需求的用户群体。



**图4.2 骑行距离分布**

共享单车的通勤时距模式分析表明，其使用场景十分典型，符合推广期宣传的短途、快速、方便等特点。

对共享单车通勤时段进行汇总分析，研究不同工作日一天二十四小时不同时段的通勤规律。首先将骑行数据的开始时间按照工作日分组汇总，再对每一个工作日按二十四小时统计每一个小时内的订单数量，最后将每个小时的订单数量除以当前工作日的总订单数量，得到每个小时的订单占比。分别对2016年及2020年的骑行数据进行汇总，可以得到如下统计图，结合该统计图可以探究上海市一周七天每天二十四小时内的订单数量分布。



**图4.3 骑行时段分布**

2016年工作日骑行具有明显的早晚高峰特征，早高峰集中在8:00至9:00，晚高峰则以16:00至17:00较为突出，与用户工作日上班、上学等通勤需求贴合。周末骑行曲线总体平缓，且8:00至15:00总体高于工作日，其中16:00至19:00呈现出一个平滑单峰，这样的分布与用户周末休闲放松需求相符合。

2020年工作日骑行时段分布较为不同，其中周一、周二、周三及周末呈现出明显的早晚高峰特征，而周四及周五呈现出类似于2016年周末的平缓单峰特征。并且，周末的早高峰强度甚至大于一般工作日。这种现象与当时新冠疫情防控政策有关，当时某些企业采取居家办公或弹性工作策略，学校也有一定的网络授课安排。另外，大规模核酸检测也会造成非一般的使用时段分布，一些人可能为了不影响工作日正常上班，而选择在周末上午做核酸检测。

4.1.2 通勤间隔分析

间隔，指的是用户的相同行为之间的时间间隔长度，例如用户两次登陆账号的时间间隔等。用户骑行共享单车通勤通常是一种重复性动作，同一用户两次骑行之间的时间间隔可以有效反映使用者的行为习惯。间隔分析就是在研究大量个体用户骑行时间间隔特征的基础上，汇总出总体用户间隔的概率分布并把握宏观用户行为模式的一种研究方法。

用户骑行数据中有用于区分单个用户的编号（UID）及用户开始骑行的时间戳（ST），通过筛选用户编号再顺序求取两次骑行时间戳之间的差值即可获得用户骑行间隔，可以将所有用户的骑行间隔汇总至一个代表了整体用户行为模式的超级用户中。对于单个用户，可以采用如下公式计算骑行间隔：

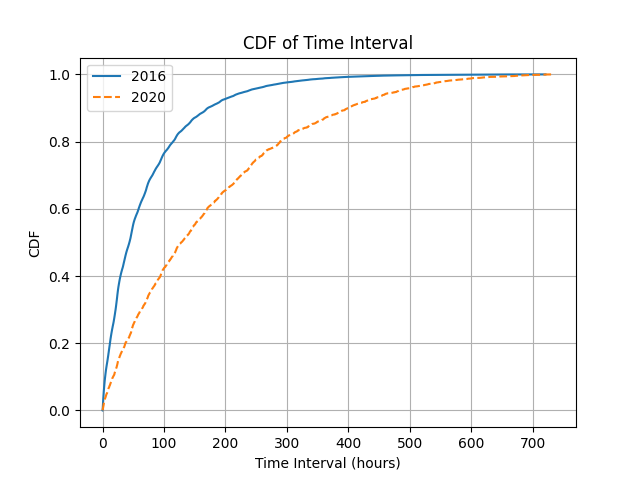
可以将超级用户的骑行间隔数据记录为：

其中，为用户骑行间隔时间，$n$为用户骑行次数。

对汇总了所有用户骑行间隔的超级用户，我们可以使用累积分布函数（Cumulative Distribution Function, CDF）来描述用户骑行间隔的分布情况。累计分布函数是一个在[0,1]区间上的单调非减函数，数学定义如下：

在计算机中可以使用如下公式估算CDF：

分别汇总计算2016年和2020年上海共享单车用户两次骑行之间间隔的累计分布曲线：



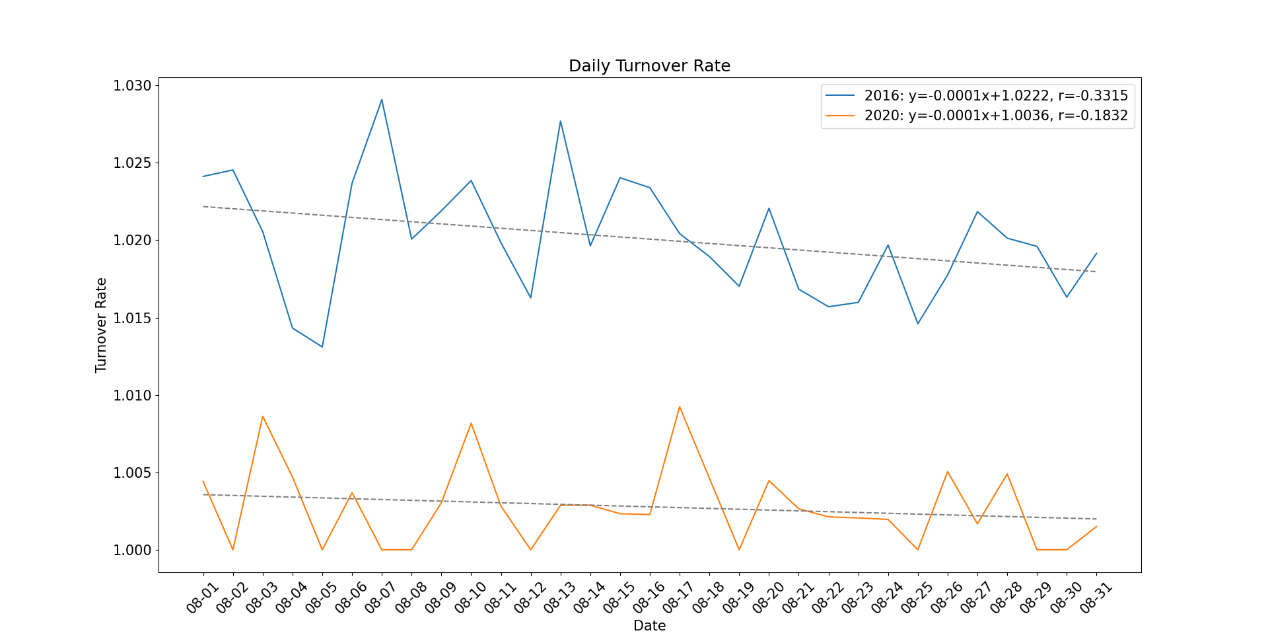
**图4.4 用户骑行间隔累计分布曲线**

2016年和2020年的两次骑行间隔分布总体相似，数据相对集中于短间隔方向（间隔300小时以下），这也意味着大部分用户频繁使用共享单车，两次骑行间隔较短，多用于日常通勤。

2016年的累计分布曲线在200小时（约间隔8天）处具有明显拐点，数据相对集中于该拐点左侧，呈现出明显的短间隔高频特征，即多数用户一周内要多次骑行共享单车。2020年所有时间间隔的CDF值都小于2016年，这意味着在2020年，用户两次骑行之间的间隔时间总体上比2016年长，考虑到此时处于疫情期间，人们的出行需求受到一定程度的抑制。但是，2020年的曲线没有明显拐点，呈现出一种平缓增加的特征，这说明2020年的两次骑行间隔分布相比2016年更加分散，用户骑行共享单车的通勤模式更为多元，共享单车逐渐为更多人所接受。

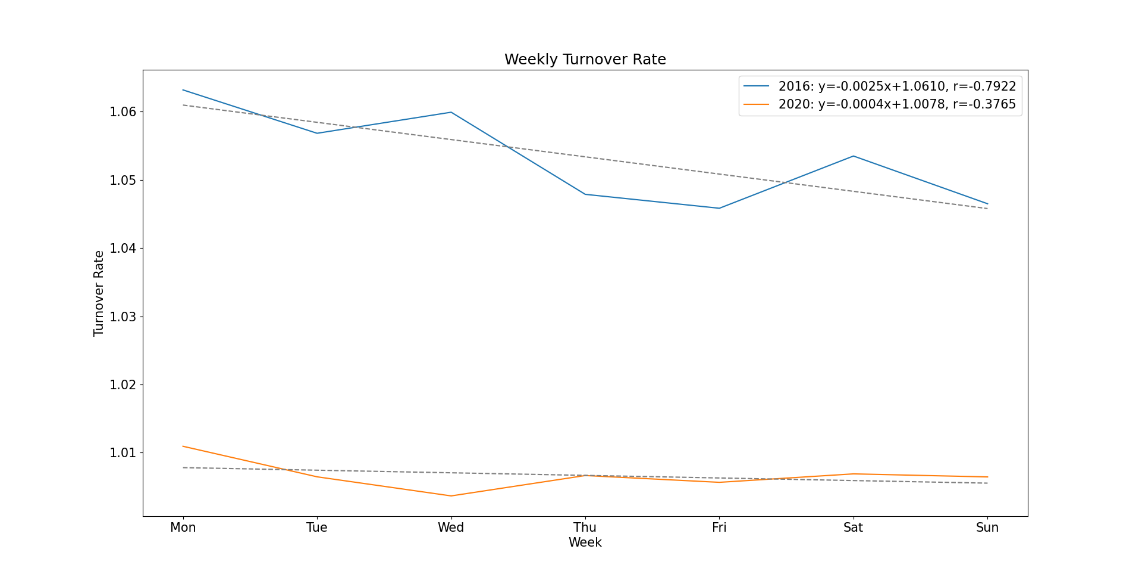
4.1.3 单车周转率分析

周转率原本是一个企业管理中的指标，用来衡量企业资产的使用效率。在共享单车的应用中，我们可以将共享单车的周转率定义为单位时间内单车的使用次数。周转率越高，说明单车的使用频率越高，需求越旺盛。在共享单车的运营中，周转率可以用来衡量共享单车的使用效率，以及共享单车的使用需求。在某一特定时间窗口内，单车周转率具体定义如下为该时间窗口内总订单数与该时间窗口内提供服务的不重复单车总数。现对轨迹数据统计日周转率如下图：



**图4.5 共享单车日周转率**

从总体趋势上来看，由月首至月尾，日周转率均呈现下降趋势，总体波动趋势相近，2020年曲线呈现约三天的迟滞。观察到2016年及2020年均存在三至四个较为明显的周转率尖峰，几乎呈现七天等距分布，并逐步降低。这暗示了共享单车周转状况存在一定的的周内特征，故统计一周七个工作日的周转率状况，绘制为下图：



**图4.6 共享单车工以作日分组的日周转率**

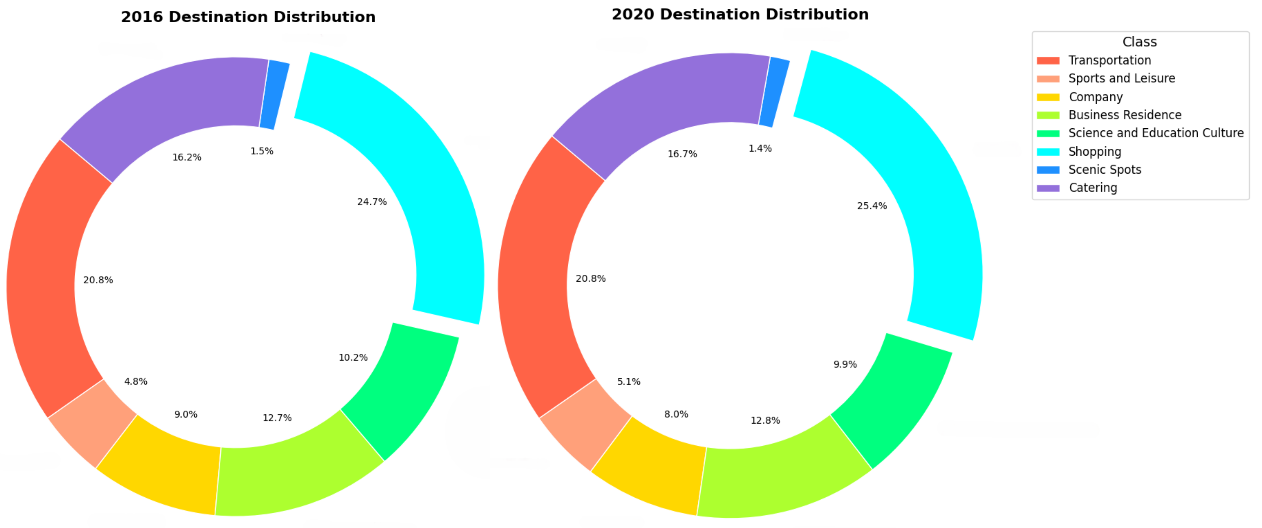
可以发现二者总体趋势相当，工作日日周转率随周一到周日逐步下降，该规律也与日周转率变化相符合。从日周转率来看，高位往往出现在周一、周二、周六及周日，低位一般出现在周三至周五。单车周转率可以从一定程度上反映共享单车企业调度策略及用户通勤需求，通过分析周转率变化，可以帮助企业优化调度策略。由于获取的数据仅为总体骑行数据的部分样本，计算出的周转率会有一定的偏差，但是通过观察其变化趋势仍能发现一定规律。

4.2空间通勤模式分析

4.2.1 通勤目的地分析

骑行作为一种便捷的出行方式，其终点位置往往能够反映骑行者的行为目的。例如，如果骑行终点位于商场附近，则可以推断骑行者可能打算去购物；如果骑行终点位于交通设施附近，例如公交车站台或地铁站，则可以推断共享单车的活动半径可能无法满足骑行者的通勤需求；如果骑行终点位于体育馆或公园附近，则可以推断骑行者可能打算进行体育锻炼或休闲放松。

使用本文提出的骑行终点最近邻匹配算法（3.3.1），可以将骑行终点匹配到最近的兴趣点类别，推断骑行者的行为目的。通过对2016年和2020年的骑行数据进行分析，可以得到如下统计图：



**图4.7 目的地占比分布图**

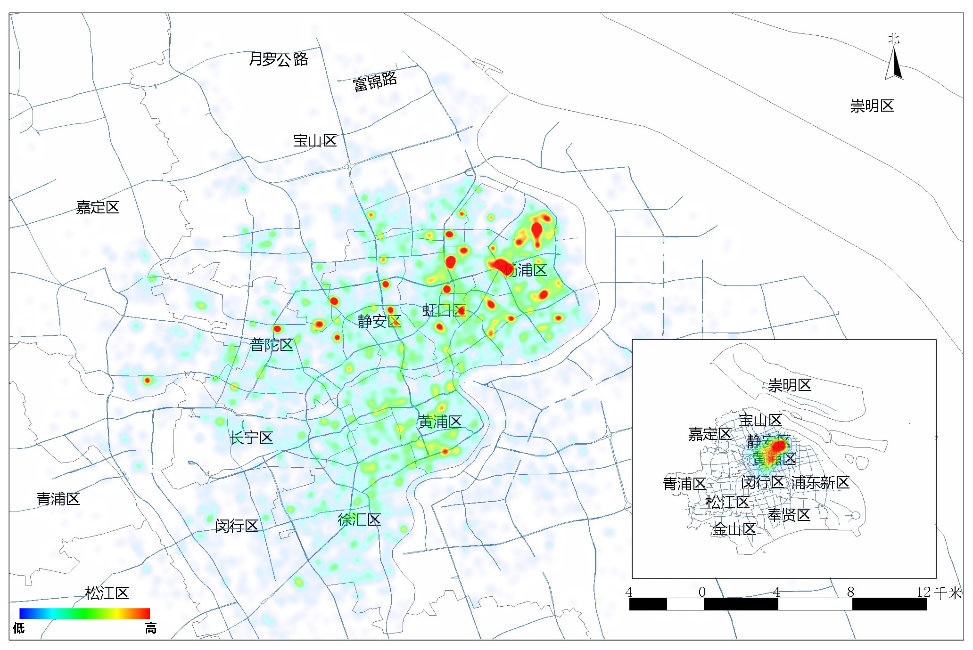
2016年与2020年骑行目的地总体相似，购物服务、交通设施服务和餐饮服务为主要目的地，占比近六成，反映了上海市商业活动活跃、交通发达、生活便利的特点。科教文化服务、商务住宅和公司企业也是较多用户的选择，占比近三成，表明上海市拥有较多较密集的院校及企业。风景名胜和体育休闲服务占比最少，可能与人口密集、生活压力大有关。

2016年至2020年，上海市骑行目的地总体呈现出“三升一降、两稳”的趋势: 1）购物服务、餐饮服务和商务住宅类目的地数量有所增长，分别增长了0.7%、0.5%和0.1%。 这表明该区域的商业活动、餐饮需求和商务氛围有所增强，可能是由于经济发展、居民收入水平提高、消费观念转变等因素导致。2）交通设施服务、公司企业和科教文化服务类目的地数量保持稳定，占比变化不大。 这说明该区域的交通基础设施、商务环境和文化氛围相对稳定，共享单车已经成为人们日常通勤的重要交通工具3）风景名胜类目的地数量略有下降，占比下降了0.1%。 这表明该区域的旅游资源开发需要进一步加强。

通过提取骑行轨迹终点并绘制热力图，我们可以研究上海市共享单车通勤目的地的空间分布模式。结果表明，通勤目的地热点主要沿交通干线分布，呈现连珠状分布，且主要集中在杨浦区。热点区域主要沿逸仙路、共和新路、四平路等路段分布。对比2016年与2020年数据，发现热点区域沿交通干线呈连珠状分布的分布模式保持稳定，但2020年骑行热度总体下降。这表明，共享单车主要还是一种短途中继通勤工具。一般商业设施、交通站点多沿主城区交通干线分布，骑行目的地主要为购物服务及交通设施，也印证了这一观点。



**图4.8 2016年骑行目的地热力图**

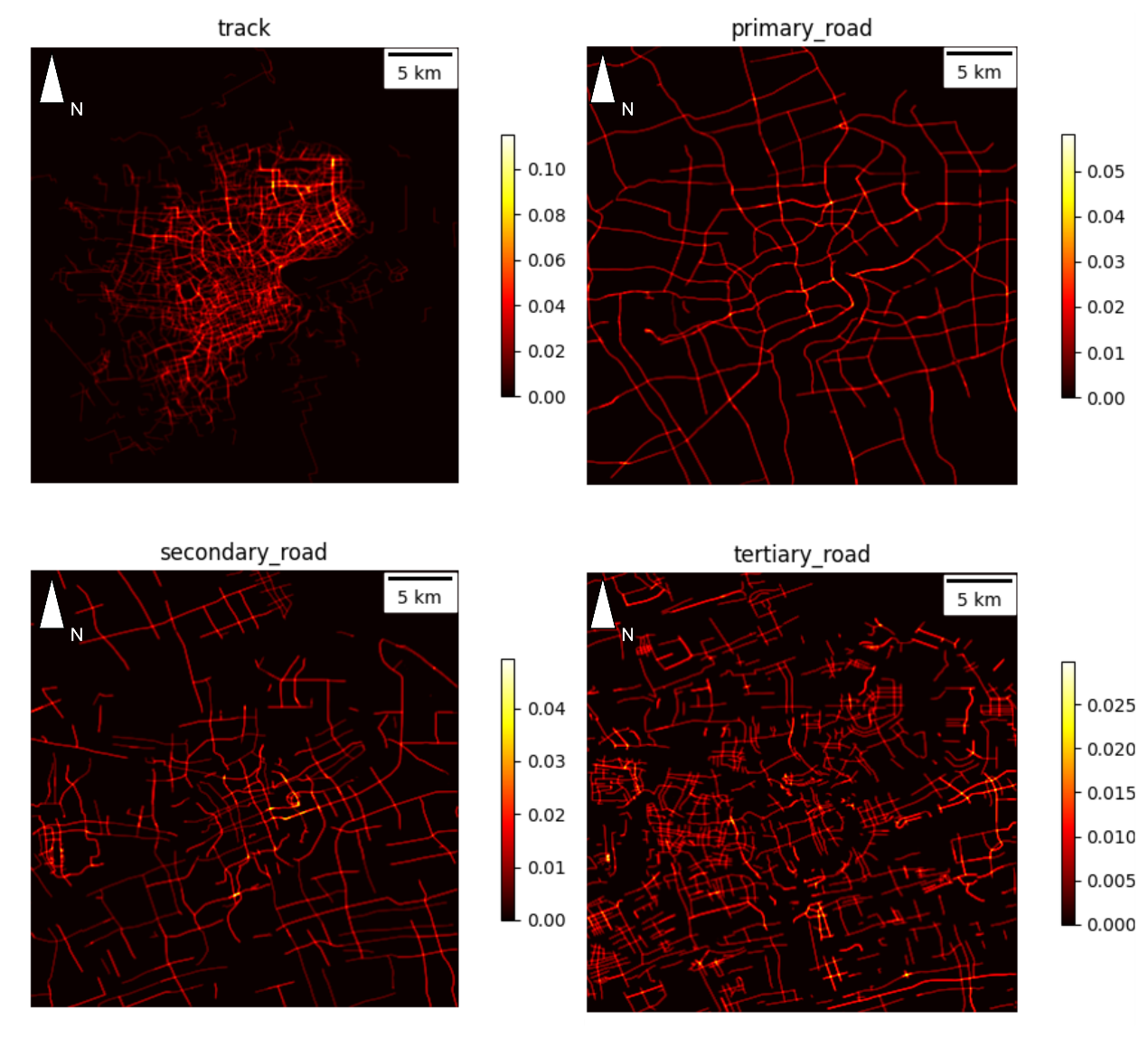


**图4.9 2020年骑行目的地热力图**

总体而言，2016年及2020年上海市骑行目的地组成及分布，可以发现该地区商业资源丰富且分散，居民购物、餐饮需求较为旺盛。交通基础设施发达，生活便利，并且共享单车很好地起到了补充作用。

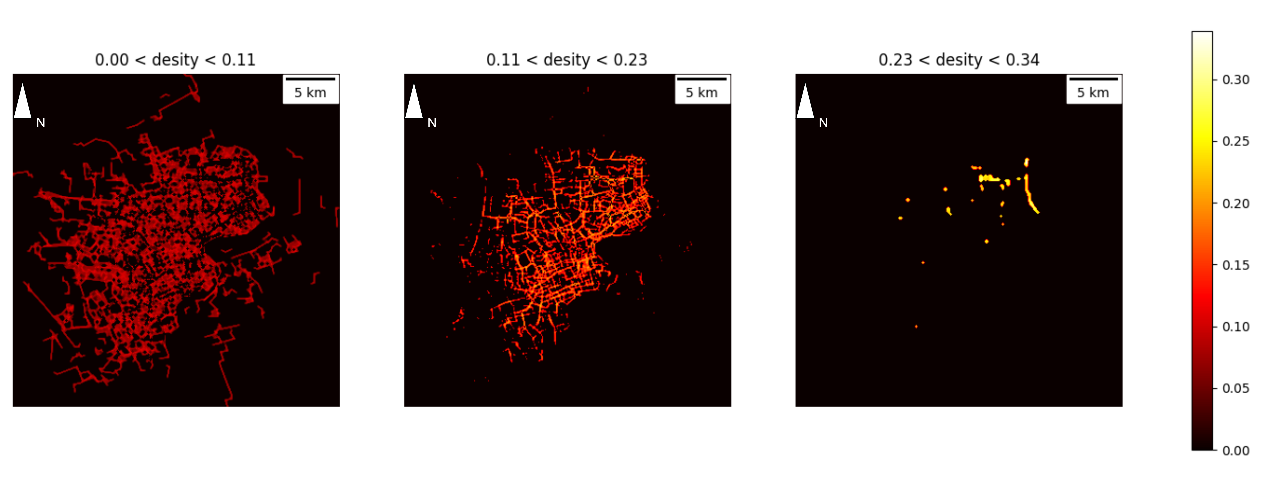
4.2.2 通勤轨迹分析

考虑到工作日通勤轨迹具有相似性，本文选取2016年8月1日（星期一）全天上海市共享单车轨迹数据做详细分析，该数据已预先使轨迹重排序（3.2.2）及道格拉斯-扑克法（3.2.3，阈值为0.0015）处理过。观察到抽稀后轨迹数据粗略呈束状沿道路分布，故采用空间线密度算法轨迹数据进行处理，得到线密度栅格（分辨率约为100m\*100m）。为了探究骑行轨迹数据与各级别路网数据之间的关系，首先使用轨迹格网边界矩形裁剪三级路网数据（主路、次主路、支路），再以与轨迹线密度栅格相同的分辨率计算各级路网的线密度栅格。



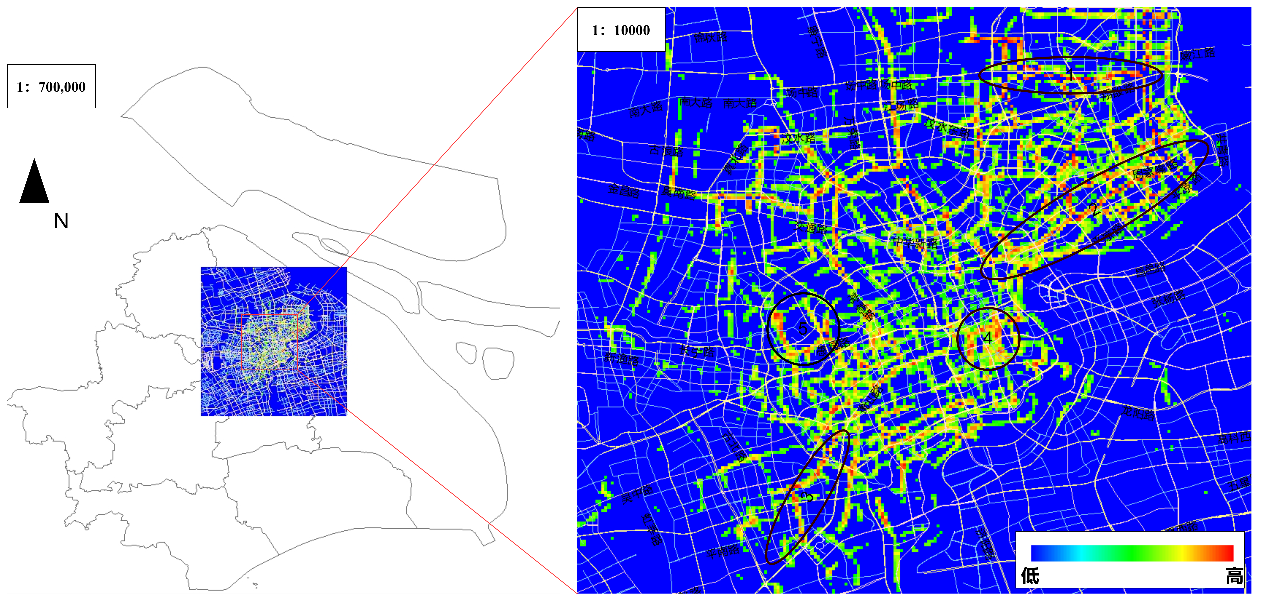
**图4.10 轨迹及各级别道路线密度栅格**

线密度栅格反映了一定空间区域内线数据的密集程度，这对于统计骑行轨迹数据的空间分布具有重要意义。根据计算出的骑行轨迹线密度栅格值分布，划分为三个层级，分别为低密度区域（密度大于0小于0.11），中密度区域（密度大于0.11小于0.23）及高密度区域（密度大于0.23小于0.34）。为了使线形更加清晰，并保持一定的连续性，对数据做了结构元素大小为3的开运算操作。可以发现，骑行轨迹通过最密集的线性区域位于东北方向。



**图4.11 各层级（低、中、高）骑行密度栅格**

选取中密度层级栅格与路网矢量数据叠加制图，图中蓝色代表栅格背景色，绿色代表一般密集区域，黄色代表中等密集区域，红色代表高度密集区域。可以发现骑行轨迹线密度栅格大体上与路网重叠，部分区域因栅格化精度问题存在一到两个格网的偏差。大部分线状骑行轨迹栅格条块沿自身长轴断续存在高度密度路段，尤其易于在路段交汇处出现。少数路段存在连续的高密度栅格分布：政立路（1）、周家嘴路（2）、宜山路及柳州路（3）。还存成团、块、环状分布的高密度栅格区域：黄陵坡路-陆家浜路-河南南路-延安东路所围成的地块（4）、宁夏路-凯旋路-长宁路-曹杨路所组成的环状路（5）。



**图4.12 骑行密度栅格与路网叠加图**

其中，政立路路段位于学校密集区域，沿途分布有上海体育学院宿舍群、上海财经大学教学及宿舍区域以及其他几所小学初中；周家嘴路位于产业密集区域，沿途分布有许多汽修、电子、钢材及汽车用品厂区；宜山路及柳州路位于居住密集区域，周围分布高密度住宅楼；对于团、块、环状高密度栅格区域，黄陵坡路-陆家浜路-河南南路-延安东路所围成的地块为住宅密集区域，其中夹杂分布着幼儿园、小学等学校区域；宁夏路-凯旋路-长宁路-曹杨路所组成的环状路内部分布有上海中山公园及大片的住宅区，苏州河也流经这片区域。

计算骑行轨迹线密度与各级别道路的线密度栅格之间的Jaccard相似系数，得到如下表格：

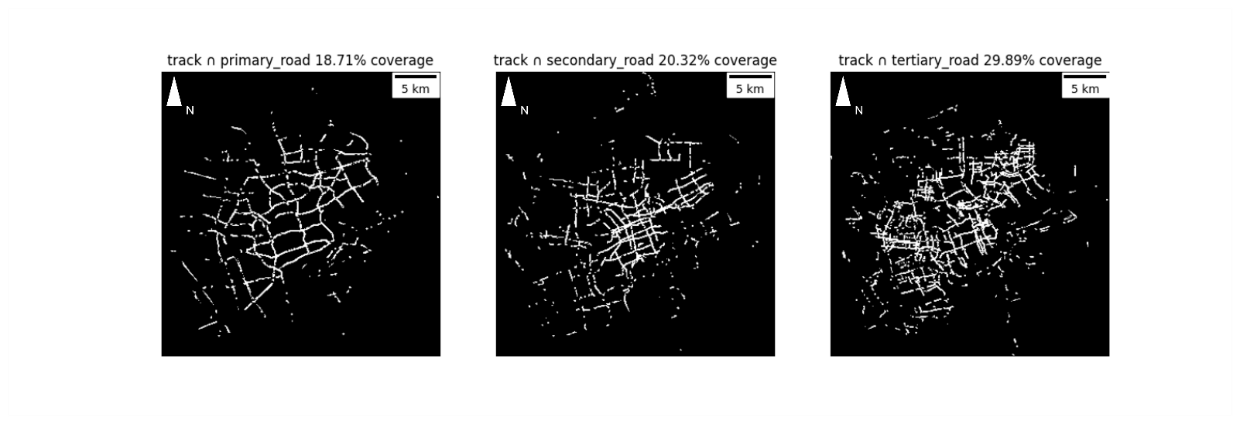
表4.1 轨迹栅格与各级别路网栅格间的Jaccard相似系数

|  |  |
| --- | --- |
| 道路等级 | Jaccard系数 |
| 主路 | 13.21% |
| 次主路 | 14.00% |
| 支路 | 17.55% |

可以发现，骑行轨迹线密度栅格与支路线密度栅格的相似系数最高，次主路次之，主路最低。为了考察各级路网对骑行通勤的支撑状况，现设计轨迹覆盖指标（coverage），对某一级别的路网该指标定义如下：

其中，track为骑行轨迹线密度栅格，road为某一级别道路线密度栅格。该指标参考了Jaccard相似系数的思想，但是将分母替换为轨迹线密度栅格，这样做可以考察某一级别道路栅格对整体骑行轨迹的支撑作用，该指标可以解释为在该道路骑行的轨迹占总骑行轨迹的比率。

为提高栅格数据质量，首先对所有线密度栅格执行系列操作（二值化、结构核元素大小为3 \* 3的开运算操作、结构核元素大小为2 \* 2的闭运算操作），并将轨迹线密度栅格依次与各级道路线密度栅格求交，然后将交集中的非空格网的个数除以轨迹线密度栅格总体的非空格网的个数，即可得到覆盖率。计算得到各级别道路的交集及覆盖率如下下图：



**图4.13 轨迹栅格与各级路网交集（处理后）**

可以发现支路路网对骑行支撑性最好，次主路次之，主路最低，这与Jaccard相似系数相一致。对于三级别道路整体的支撑情况，可以根据道路长度加权求和得到整体覆盖率，即：

(4.6)

其中，为第i级道路的覆盖率，为第i级道路的长度。道路长度采用道路线密度栅格二值化后经过结构核元素大小为2 \* 2腐蚀操作所得非零栅格的累计和估算，相较于轨迹线密度栅格，道路线密度栅格质量较好，故仅做一次腐蚀操作。主路为11640格（未腐蚀为15345格）、次主路为12790格（未腐蚀为16646格）、支路为22293格（未腐蚀为25477格）。根据公式计算得研究区域内总体路网支撑率为24.49%。

4.2.3 通勤潮汐分析

由于通勤需求存在时间与空间上的波动，城市交通往往存在“潮汐现象”，具体来说，早晚高峰虽然都有较大的交通流量，但是影响到的道路及区域往往是不同的，表现在道路的使用上，可能早高峰时一方向拥堵（另一方向通畅），晚高峰情况又完全相反。某些城市为了缓解“潮汐现象”带来的交通流量时空不均衡问题，尝试采用“潮汐车道”来缓解高峰时期的交通压力，例如，某一双向六车道道路，中间两车道被设置为潮汐车道，这两条车道的行驶方向可以根据交通指示灯发生改变，以平衡车流的潮汐现象。

潮汐现象的本质是通勤需求与交通供给在时空上的失衡，即动态波动的通勤需求与相对固定的交通设施供给间的矛盾。由共享单车通勤时段分布可知，共享单车通勤存在明显的时间峰谷特征，即在某一时间窗口内，某一有限区域内，共享单车的数量存在波动，即存在借还失衡的潮汐现象。在空间上找到这些“潮汐点”，并总结出相关分布趋势对于城市共享单车通勤优化具有重要作用。

现考虑研究区域内的某一有限大小区域，在某一时段内的单车借还情况，并定义潮汐指数（Tide Index, TI）为：

其中，Borrow为某一时段内该区域的借车数量，Return为某一时段内该区域的还车数量。潮汐指数为正值时，表示该区域为借车潮汐点，即借车需求大于还车需求；潮汐指数为负值时，表示该区域为还车潮汐点，即还车需求大于借车需求。潮汐指数的绝对值越大，说明潮汐现象越明显，即借还失衡程度越大。

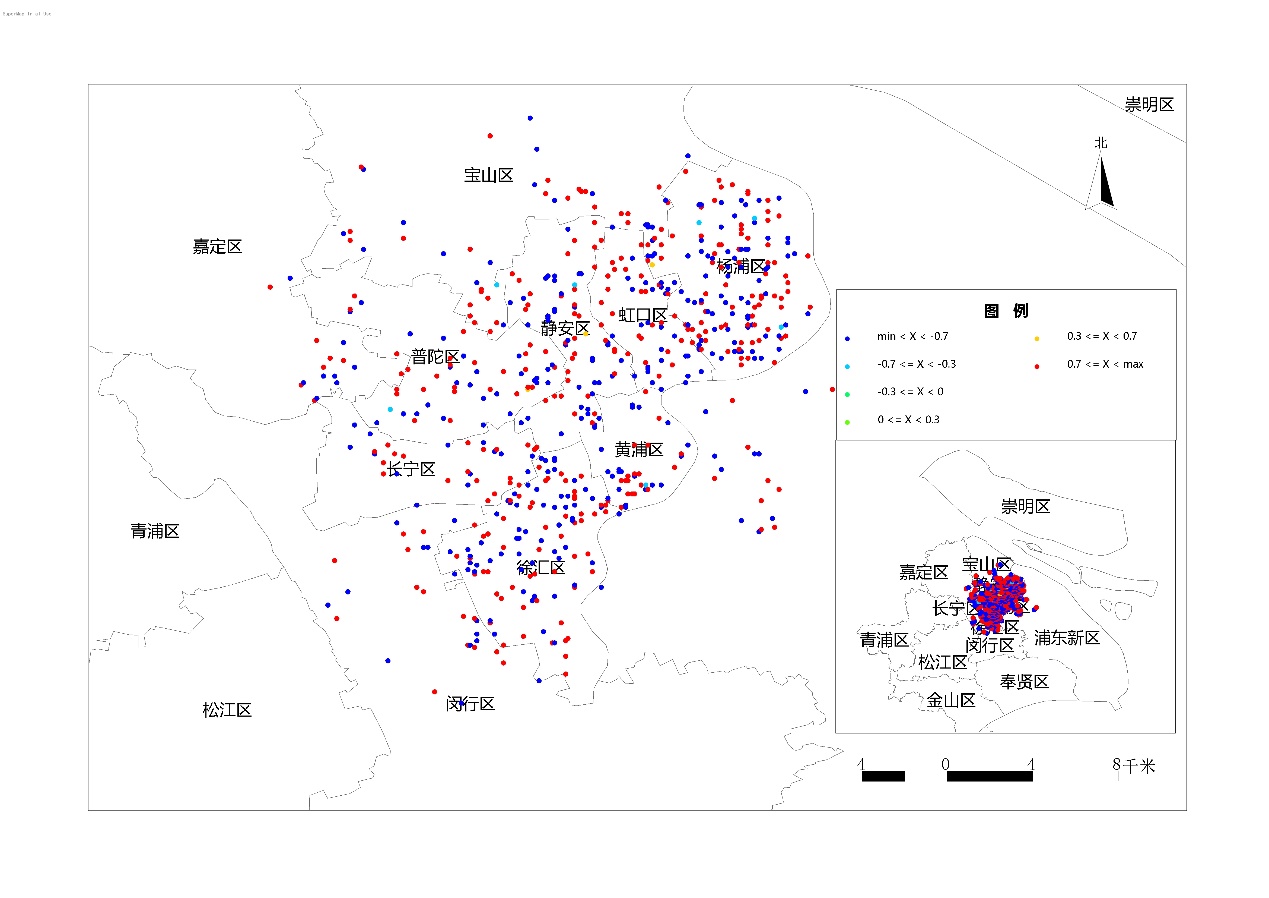
本文采用划分规则时空格网的方式，统计2016年8月1日早晚高峰（7:00-10:00，17:00-20:00）期间上海市1741条共享单车轨迹数据的空间分布，并生成借、还车潮汐点，具体算法步骤如下：

（1）首先求出总轨迹的外包矩形（121.297, 121.629, 31.077, 31.416），并选择合适的格网大小（0.001度），根据时间窗口的数量初始化格网对象。

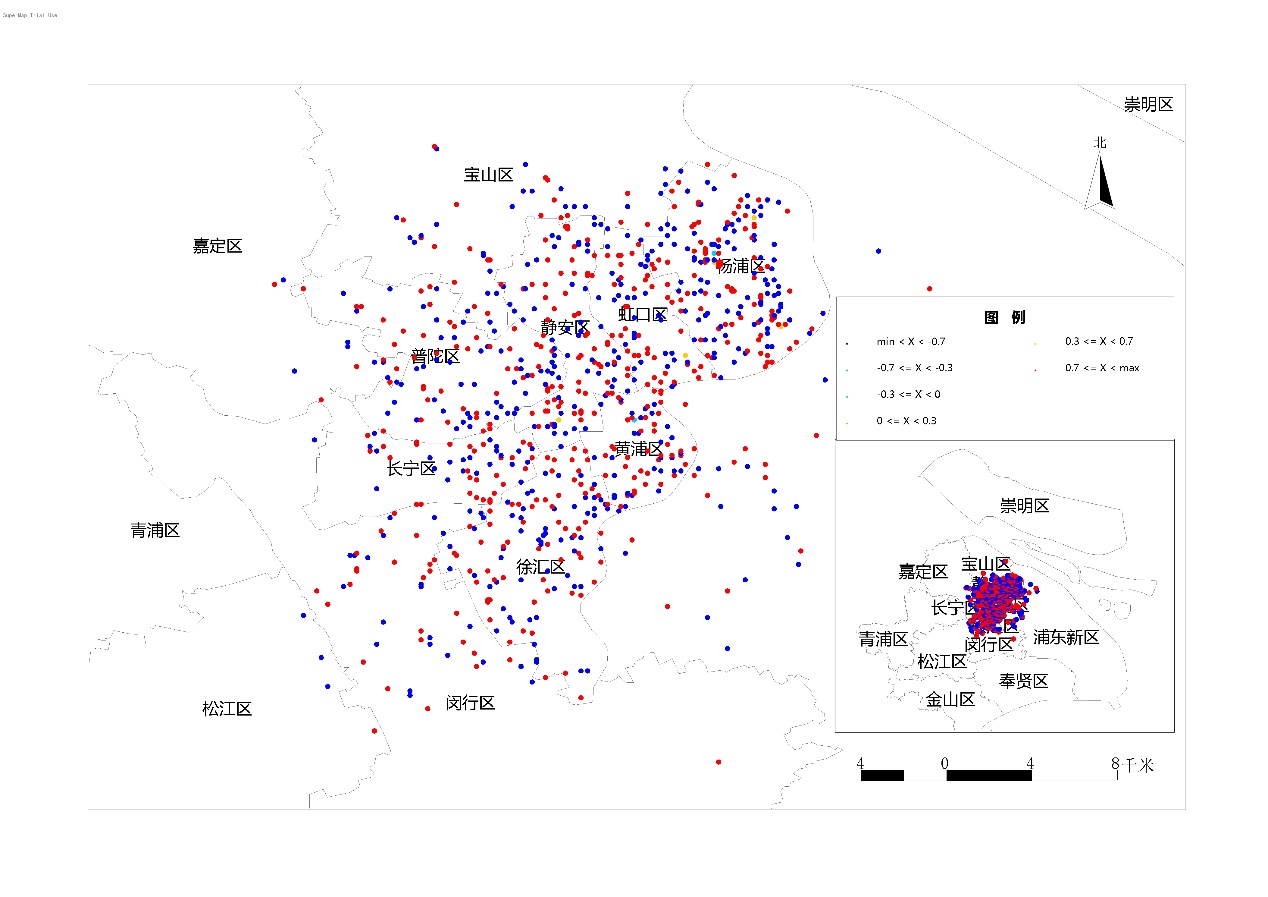
（2）对于某一时间窗口，新建与格网对象相同大小的借、还格网，认为轨迹起点为借车点、轨迹终点为还车点，遍历数据集并累加至借还格网中。结合借还格网，运用潮汐指数计算公式，计算得潮汐指数栅格。重复执行，直至每一个时间窗口都执行过一次该步骤。

（3）提取各时间窗口栅格中非零点为多个点集文件（实际记录栅格中心点），每一个点带有一个用于记录潮汐指数的字段。

得到的潮汐点分布图如下，红色代表借车潮汐点（潮汐指数为正值），蓝色为还车潮汐点。



**图4.14 共享单车潮汐点（早高峰7:00-9:00）**



**图4.15 共享单车潮汐点（晚高峰17:00-20:00）**

通过计算潮汐指数，可以找到潮汐现象最为明显的区域，即潮汐点。仅观察在在空间中离散分布的潮汐点难以有效把握共享单车通勤空间潮汐模式，所以需要对借车潮汐点及还车潮汐点分别计算空间核密度分布。

4.3本章小结

该算

**第5章 总结与展望**

5.1研究结论

研究共享单车通勤时空模式，为致力于打造高效城市慢行系统以缓解城市交通压力的地方政府及相关企业提供了关键信息。此类工作一般依赖传统的信息来源，例如街头问卷调查、静态城市交通设施报告、官方调查报告（年度、季度）及企业对外披露的报告等，这些信息源难以准确反映实时变动的城市通勤状况，基于这些信息所制定的政策及规划也往往具有一定的滞后性。另外，随着投放量增大、大量单车老化，无桩共享单车本身也正在成为一种城市治理难题，如何高效处理大量共享单车产生的数据并从中得出正确的结论，正成为城市治理的下一个关键问题。

相较于传统静态数据，骑行轨迹信息具有足够细的时空粒度，加以合适的可视化手段可以做到实时监控城市共享单车动向，若结合基础地理数据、气象数据等还可以从多个视角深入分析城市现状。但是，骑行轨迹数据也有其棘手之处，数百万辆单车一天产生的轨迹数据是十分巨大的，并且数据中往往存在噪声、缺损及其他异常情况，常规的数据处理及可视化手段难以有效处理它。

本文通过研究上海市共享单车轨迹数据，探索出一套具有一定实用价值的共享单车轨迹数据预处理及分析框架，包括轨迹点重排序、轨迹长度量测、轨迹简化及轨迹线密度栅格分析。同时，本文也使用兴趣点最邻近匹配算法、道路线密度相似性分析等方法研究轨迹数据与基础设施之间的联系。对于用户行为，本文也基于传统的统计学方法对骑行轨迹数据进行统计与可视化，例如骑行时间距离分布直方图、骑行时段折线图、骑行目的地占比饼状图及用户骑行间隔累计概率密度曲线图等。本文的研究结论可概括如下：

5.1规划建议

该算

5.1研究展望

本文选取的研究方法对于分析大量轨迹数据及其模式具有一定的学术合理性。在大量存在噪声的轨迹数据的清洗、处理、化简方面，本文提出的算法具有一定的实际意义。同时，研究结果结合上海市共享单车通勤模式为城市规划等方面建言献策，政府可以根据这些建议优化热点路段的骑行体验，针对薄弱区域加强骑行道路建设。另外，本文所使用的数据处理、分析及可视化方法，经一定的适配操作可以快速投入其他区域的研究中。然而，本文所使用的数据及方法仍然存在一些缺陷和问题，对此本文提出了一些解释及研究展望：

（1）本文的研究数据仅为上海市总体骑行数据的一部分样本，且仅涉及2016年8月及2020年8月这两个时间断面，不可避免地存在以偏概全的情况。尤其需要注意，本文的研究数据在空间上的分布并不均衡，城市边缘地带鲜有轨迹。这可能是通勤需求空间分布不均衡、提供数据的企业业务分布不均衡等多种原因所导致的，未来可以考虑收录更多企业的骑行数据，确保研究的全面性。

（2）数据缺乏详细的骑行者背景资料（如年龄、职业、收入水平等），难以根据骑行者特征对共享单车市场进行诸如逐年龄段、多职业人群、多收入梯度等的细致的分析。未来可以考虑将骑行轨迹与骑行者的身份背景信息相结合，从而从多个角度精细把握共享单车通勤模式及细分市场需求，为企业差异化竞争提供洞察。

（3）本文的研究方法多为传统的统计方法及空间分析方法，并未涉及新型智能算法（神经网络等），并未将数据中蕴含的知识直接转化为解决问题的能力，而是通过结论、建议等方式间接地帮助分析问题。未来可以考虑使用轨迹时空模式学习与预测一体化的先进的算法，从而为城市治理、交通管理等领域提供助力。

**参考文献**

1. 李文翔,唐桂孔,刘博,等.基于摩拜骑行数据的上海市共享单车减排效益时空分析[J].环境科学学报,2021,41(11):4752-4759.DOI:10.13671/j.hjkxxb.2021.0213.
2. 王若萱,吴建平,奇格奇. 基于上海市数据的共享单车用户通勤模式研究（英文）[C]中国仿真学会.第三十三届中国仿真大会论文集.2021:16.DOI:10.26914/c.cnkihy.2021.025005.
3. 全雨霏. 南京市共享单车使用的时空特征及其骑行环境评估[D].东南大学,2024.DOI:10.27014/d.cnki.gdnau.2022.001097.
4. 常新. 基于共享单车轨迹数据的城市街道可骑行性研究[D].哈尔滨工业大学,2021.DOI:10.27061/d.cnki.ghgdu.2020.002483.
5. Jie Bao, Tianfu He, Sijie Ruan , Yanhua Li, and Yu Zheng. 2017. Planning Bike Lanes based on Sharing-Bikes’ Trajectories[c]. In Proceedings of KDD’17, August 13–17, 2017, Halifax, NS, Canada., , 11 pages.
6. 刘泉宏,唐福星.基于K-means聚类算法与重心法的故障共享单车回收中心选址优化[J].运筹与管理,2023,32(07):85-91.
7. 谢光明. 基于改进时空图神经网络的共享单车流量预测[D].华东师范大学,2023.DOI:10.27149/d.cnki.ghdsu.2023.004430.
8. 刘冰,王舸洋,朱俊宇,等.基于共享单车大数据的骑行生活圈识别及其活动网络模式分析[J].城市规划学刊,2023(04):32-40.DOI:10.16361/j.upf.202304005.
9. 王俊,于爱荣.基于ConvLSTM的南京地区共享单车需求预测研究[J].软件工程,2024,27(02):55-59.DOI:10.19644/j.cnki.issn2096-1472.2024.002.011.
10. 谢国微. 天气及建成环境对共享单车出行需求的影响研究[D].南京林业大学,2024.DOI:10.27242/d.cnki.gnjlu.2022.000493.
11. 任丹. 基于TRACLUS算法的船舶轨迹分析系统的设计与实现[D].辽宁师范大学,2021.DOI:10.27212/d.cnki.glnsu.2020.001211.
12. 塔娜,柴彦威.行为地理学的学科定位与前沿方向[J].地理科学进展,2022,41(01):1-15.
13. 杨超,汪超.城市过剩通勤与职住平衡模型[J].同济大学学报(自然科学版),2013,41(11):1712-1716.
14. 夏琼燕,罗冠,张翔,等.OpenStreetMap志愿者贡献与留存分析[J].测绘与空间地理信息,2021,44(02):90-93+97.
15. 郭雨,陈金勇,张新宇,等.OPTICS与离线批处理在轨迹聚类中的应用[J].计算机工程,2020,46(7):72-77,83.
16. 赵雨琪,牟乃夏,祝帅兵,等.基于GeoHash算法的周边查询应用研究[J].软件导刊,2016,15(06):16-18.

**附录**

附录是对于一些不宜放在正文中，但有参考价值的内容，可编入毕业设计的附录中，例如重要数据、表格、公式、图纸、程序等。附录的篇幅一般不要超过正文。

附录编号依次编为附录 1、附录 2。附录标题各占一行，按 一级标题编排。每一个附录一般应另起一页编排，如果有多个较短的附录，也可接排。

标题：黑体小2加黑居中，单倍行距，段前0.5行，段后0行

内容：宋体小4号，行距固定值20磅，英文用Times New Roman，小4号

**致谢**

致谢是对在毕业论文（设计）工作中给予各类资助、指导、协助以及提供各种有利条件的单位、指导教师或其他人员表示感谢，语言应实事求是，切忌浮夸之词。

ХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХХ

标题：黑体小2加粗居中，单倍行距，段前0.5行，段后0行

内容：宋体小4号，行距固定值20磅；

英文用Times New Roman，小4号