

**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：待定

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**

In order to study……

………….......

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

**目 录**

(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

**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**

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。

**第1章 一级标题**

1.1 二级题目

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

……………………………………………………………………………………………………………………………………………………………………………………

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。

1.1.1 三级题目

**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 聚类分析

聚类分析指的是按照一定的分类规则（距离、相似性等）将一组数据划分为多个类簇的过程，是一种无监督的探索性数据分析（数据挖掘）方法，在大数据分析、计算机图形学、模式识别等专业领域应用广泛。聚类分析过程往往涉及许多算法与子任务迭代处理，且并非完全自动化，一般需要根据结果手动调整参数优化。

聚类分析的核心问题是聚类模型，典型的聚类模型可以分为以下几类：1）连接性模型：层次聚类是连接性模型的典型代表。它通过不断合并距离相近的数据对象来构建层次化的簇结构。该方法的优点是易于理解和实现，但对初始度量方法的选择比较敏感。2）质心模型：k均值算法是质心模型的代表，它通过迭代地移动簇的质心来使得簇内数据的距离平方和最小化，最终将数据划分到k个簇中。3）分布模型：使用统计分布对簇进行建模，例如期望最大化算法使用的多元正态分布。4）密度模型：DBSCAN和OPTICS算法是密度模型的代表，它们将数据空间中密度较高的区域定义为簇，并通过密度阈值或可达距离来识别簇中的数据点。5）基于图的模型：HCS聚类算法是基于图的模型的代表，它将数据点之间的相似性关系转换为图结构，并通过图的社区发现算法来识别簇中的数据点。6）神经模型：自组织映射（SOM）是神经模型的代表，它是一种无监督神经网络，将数据点映射到一个二维网格上，并通过相邻神经元的激活模式来识别簇中的数据点。

在上述聚类模型的基础上，可以进一步设计用于区分不同类别的规则，这些规则一般依赖一种相似性度量来划分，例如欧几里得距离、曼哈顿距离、闵可夫斯基距离、皮尔逊相似度、豪斯多夫距离等。根据分类规则对于分类算法的约束程度（是否允许重复分类、是否允许存在未分类目标等），又可以分为硬性聚类与模糊聚类（软聚类）。

2.2.2 时空轨迹聚类分析

时空轨迹建模为某一对象按时间顺序记录的空间位置序列，对其的聚类分析就是根据轨迹间的时空相似性度量划分为不同类簇的过程。时空轨迹数据具有一定的特殊性，首先，它是以时间顺序记录的移动对象位置序列，内含时间因素代表其具有动态性，同一个对象在不同时刻具有不同位置属性，这与一般聚类分析的静态数据（如带有诸多静态属性数据的人或物）具有较大差别。其次，它有明确的线状空间结构，相较于一般的（多维空间中）点状数据模型，轨迹数据本质上为连续的同质顺序点列，轨迹间的相似性度量无法直接套用一般聚类的概念为团块的聚类模型已有的度量方法，而是需设计与之适应的特殊算法。

考虑到时空轨迹的特殊性，无法直接套用一般的聚类方法，而是需要根据轨迹的时空间结构设计特定的相似性度量算法，部分典型的时空轨迹相似性度量算法：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。

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

本章以行为地理学、城市职住平衡理论和过剩通勤理论为理论基础，并探讨聚类分析、用户行为分析和时空轨迹聚类分析等相关研究技术。行为地理学强调以人为本，探讨人与地理环境的互动关系，并在可持续发展和社会公平等领域发挥重要作用。城市职住平衡理论指出有效缓解城市交通拥堵的关键是合理规划就业与住房地点，而过剩通勤理论提供了衡量城市通勤状况的量化指标。在相关研究技术方面，聚类分析和用户行为分析为数据挖掘和模式分析提供了重要手段，特别是时空轨迹聚类分析针对轨迹数据的特殊性提出了一系列相似性度量算法。这些理论和技术的研究与应用有助于更好地理解和解决城市发展中的诸多问题，促进城市可持续发展和社会治理的进步。

**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

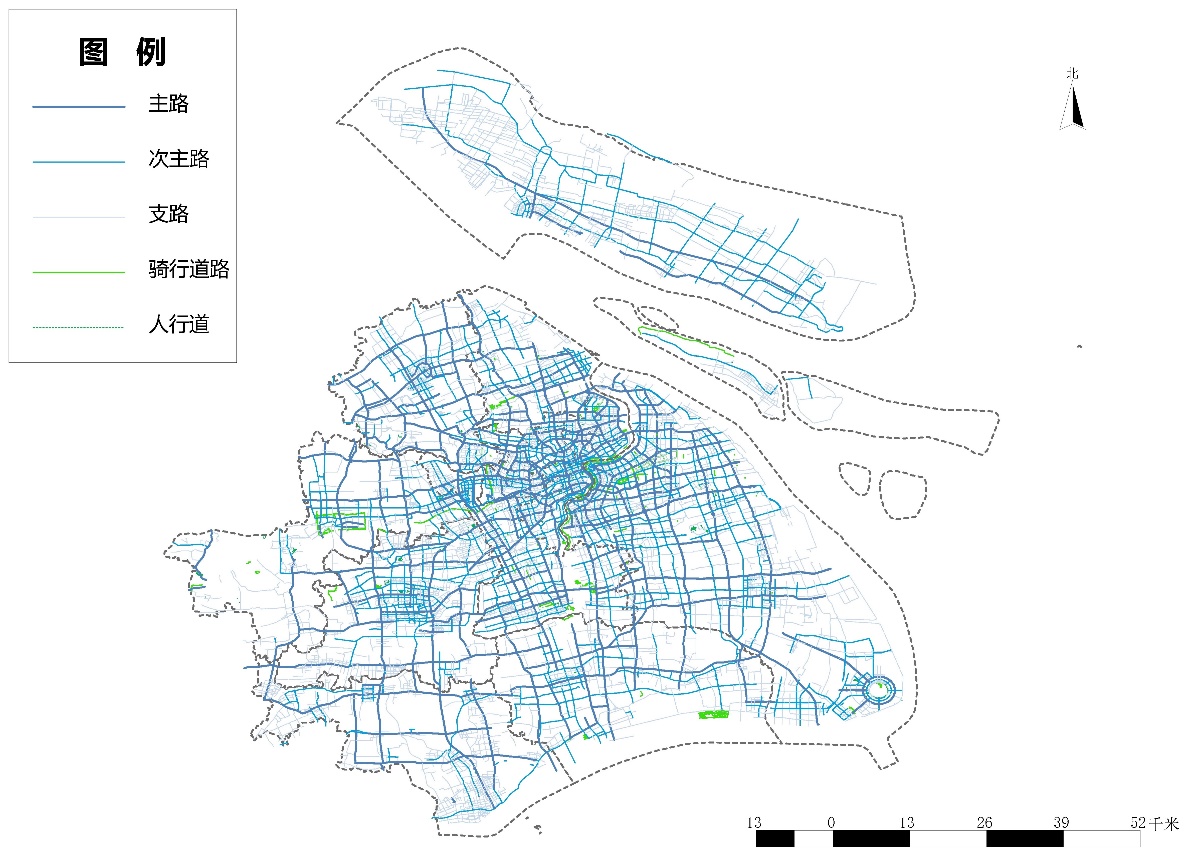
本文的研究数据包括骑行轨迹数据与上海市基本地理信息数据，其中骑行轨迹包括由互联网采集整理得到的 2016 年 8 月整月 102361 条上海市某品牌共享单车骑行轨迹数据及 2020 年 8 月整月 12793 条骑行记录，上海市基本地理信息数据包括由 Open Street Map 开源地理信息数据网站获取上海市路网矢与兴趣点分布矢量数据及由高德地图开放平台获取的上海市矢量边界数据。

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“#”分隔的骑行轨迹点集字段，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（具体方法见 3.2.2），并且骑行实际距离也需要根据正确的轨迹数据计算（计算方法见 3.2.4）。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 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#... |

上海市基本地理信息数据的预处理使用 QGIS 软件处理，使用 QGIS 软件解析 Open Street Map 网站内部格式（ \*.OSM ）的矢量数据将其转换为常用的 shapefile 文件格式( \*.shp )并使用上海市矢量边界数据裁剪掉区域外的部分。从上海市路网数据集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。



**图3.1 上海市分级路网数据**

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，此外，还从 OSM 数据中提取了相关兴趣点数据。

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轨迹简化

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。

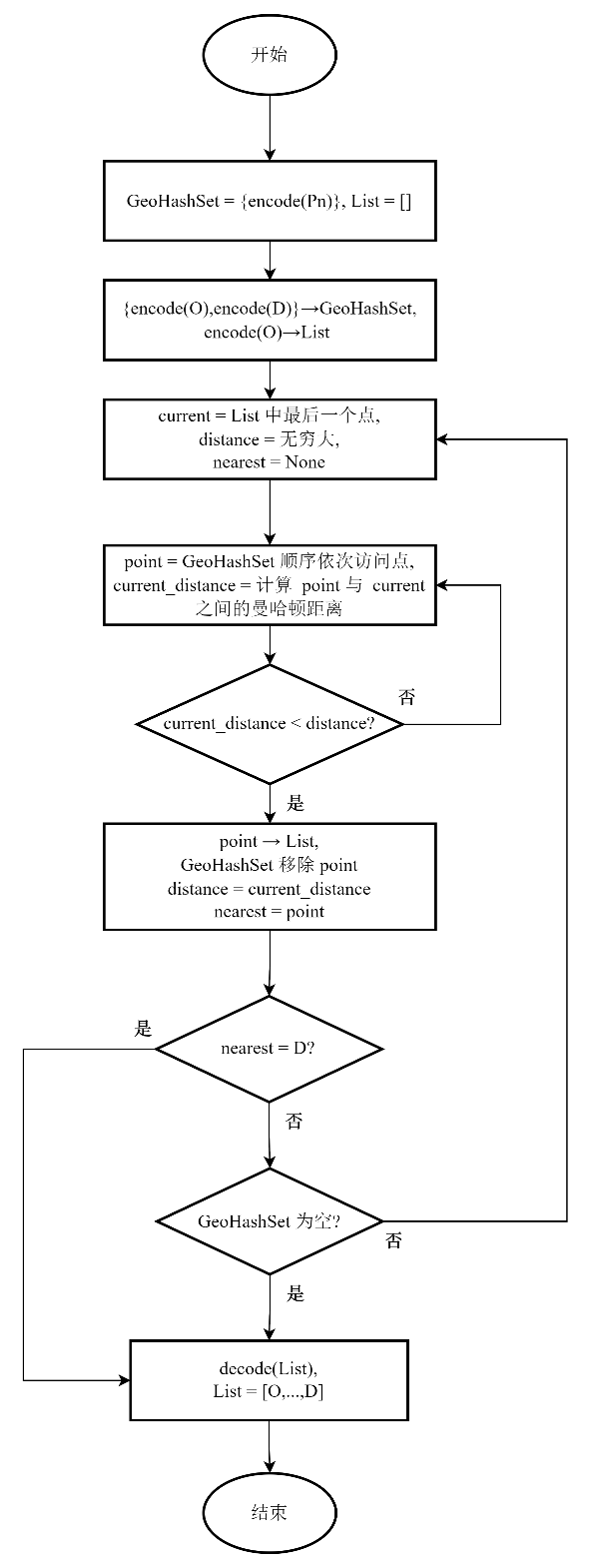


**图3.2 GeoHash算法流程**

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：

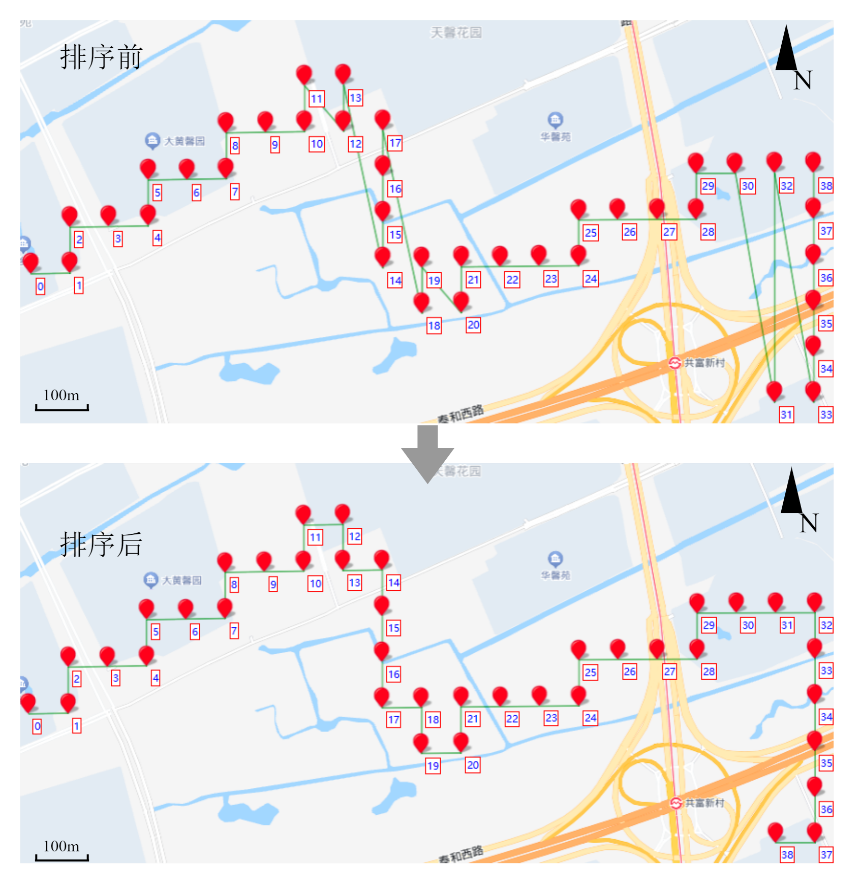
考虑到轨迹点密度较高，相邻点之间的距离很短，且经纬度变化范围较小（绝大部分相邻点之间的经纬度跨度不超过0.001°），因此曼哈顿距离能够有效捕捉轨迹点的局部运动特征。且曼哈顿距离的计算公式简单，仅涉及绝对值运算，计算效率高。对于具有大量轨迹点的GPS数据，使用曼哈顿距离可以显著降低计算成本。因此，本算法通过计算两个坐标点之间的曼哈顿距离来判断两个点之间实际的空间距离。

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。以下是对某一真实路径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。

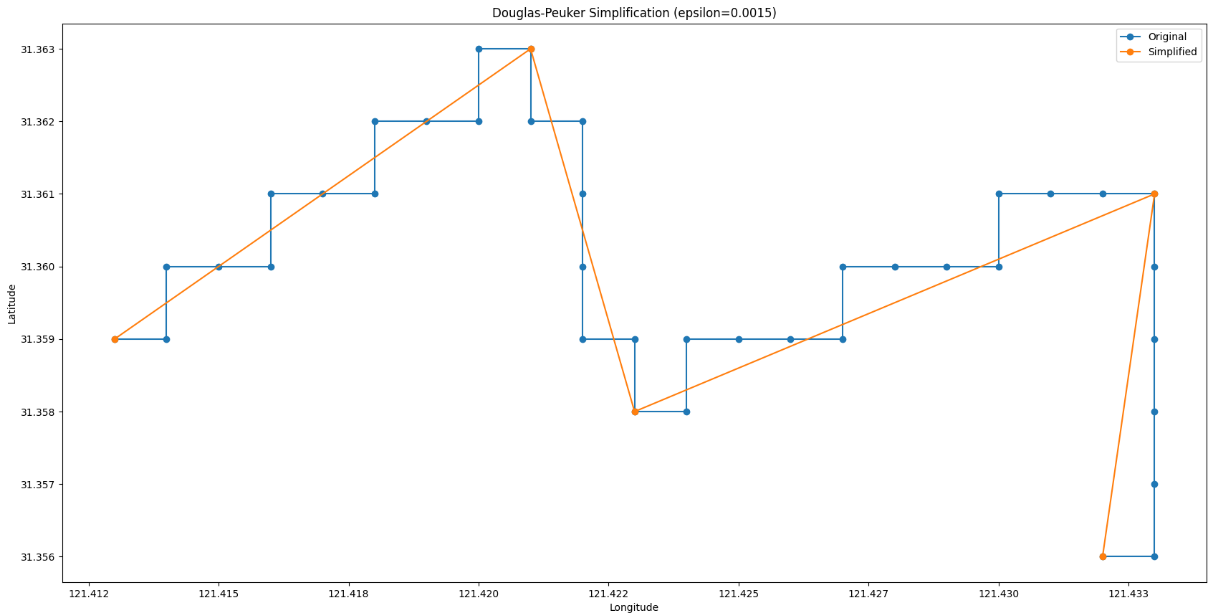
使用上述方法对一段轨迹进行处理，计算得轨迹起点与终点间的直接半正弦距离为 2118.25 米，未排序轨迹点集的累计半正弦距离（认为地球半径为 6371000 米）为 8685.89 米，经过算法重排序后的累计半正弦距离为 3605.04 米。可以发现，由于未经处理的轨迹点集中存在部分不合理的乱序点，导致轨迹总长度明显偏大，经算法重排序后的轨迹点集较为合理。

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

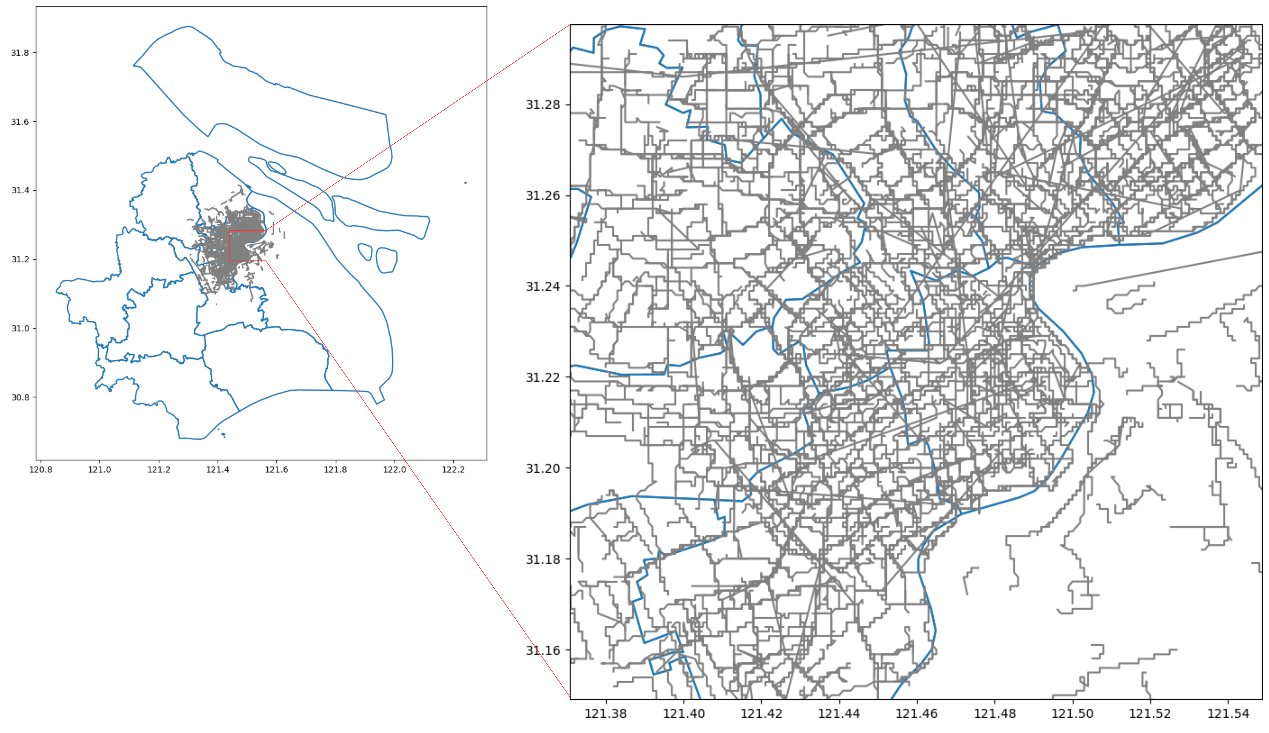
共享单车位置记录设备通常以固定频率记录并上传数据，导致即使经过轨迹重排序，仍存在大量冗余点，例如，平直路段可能出现过于密集的轨迹点，增加后续分析负担。此外，由于定位精度限制，用户沿着某些路径运动时，轨迹会呈现锯齿状噪声，这也会影响后续分析，尤其是对于轨迹聚类，这种曲折的轨迹会严重影响聚类质量，导致最终拟合轨迹出现剧烈的波动。

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，在经过线抽稀步骤后，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 改进的 TRACLUS 算法

正文

3.4 本章小结

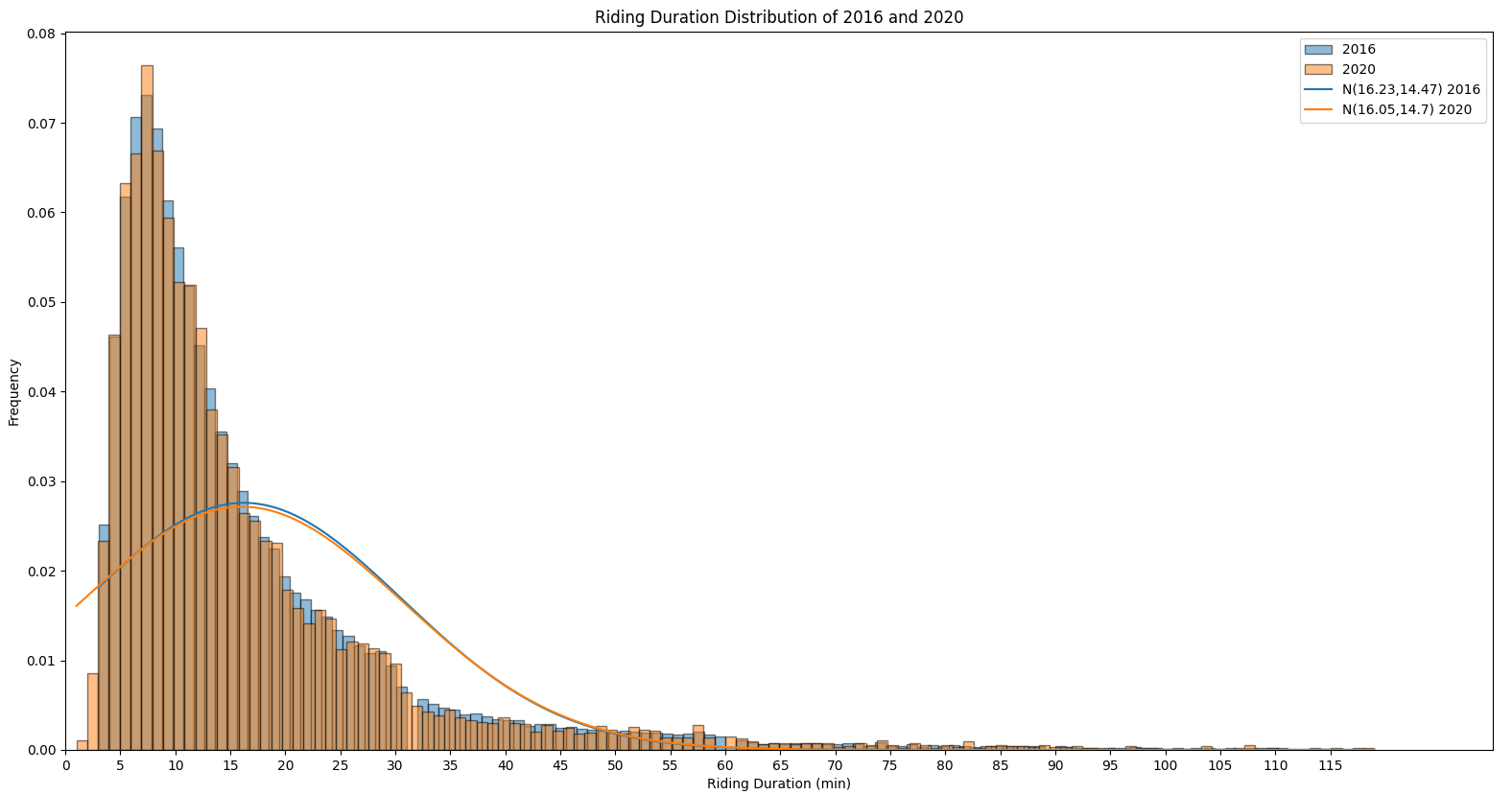
正文

**第4章 结果与讨论**

4.1 时间通勤模式分析

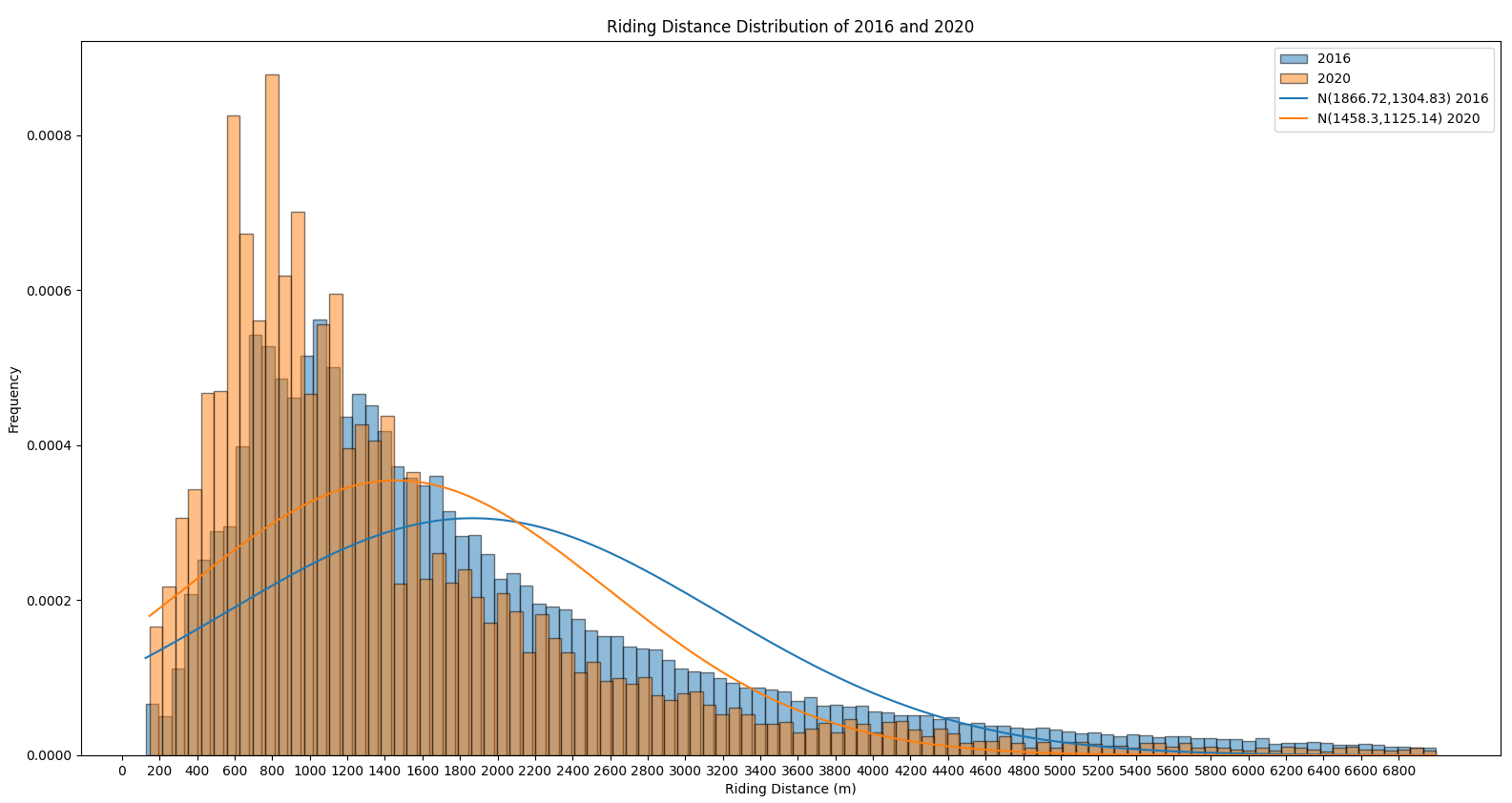
4.1.1 通勤时距模式

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**

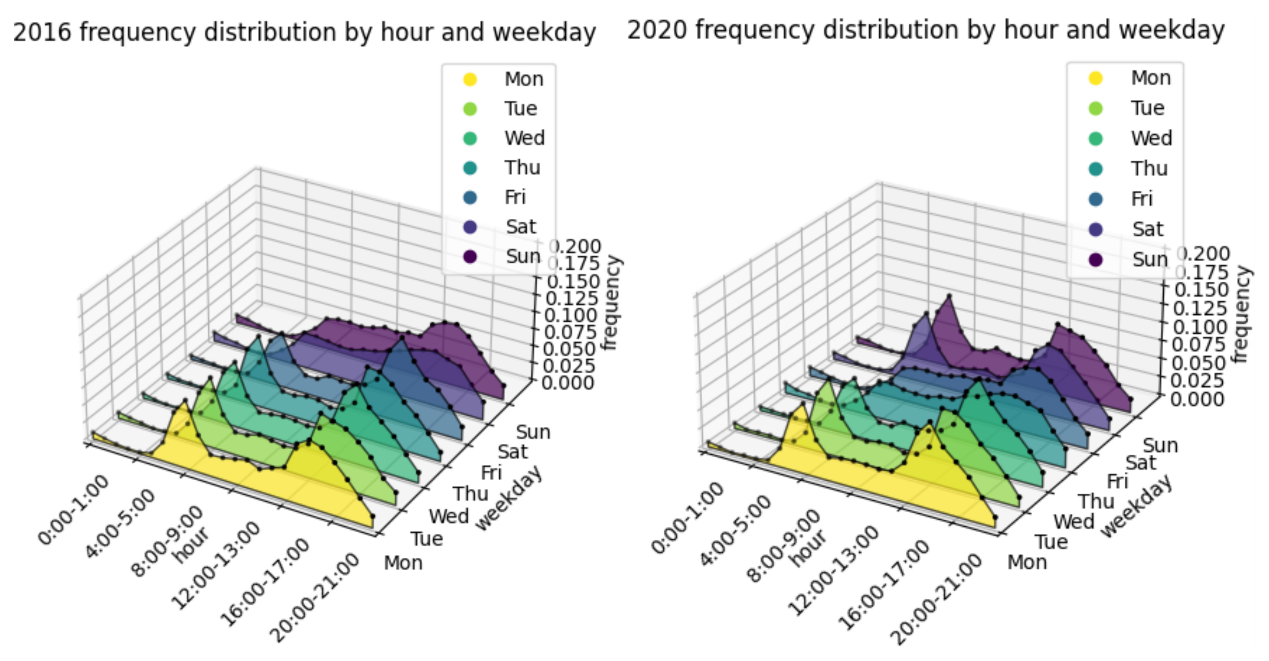
对共享单车通勤距离进行汇总分析，绘制出频率分布直方图及对应的拟合正态分布曲线，可以发现2016年平均骑行距离为1.8公里，而2020年平均骑行距离仅为1.4公里。2020年数据高度集中在小于1公里的范围内，数据标准差更小，分布也更为紧凑，0.8公里左右的短途骑行成为最多人的选择。说明人们越来与认可单车是解决“最后一公里”通勤问题的有效交通工具，共享单车市场也逐步积累了一批有稳定短途通勤需求的用户群体。



**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。

4.2.1 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：

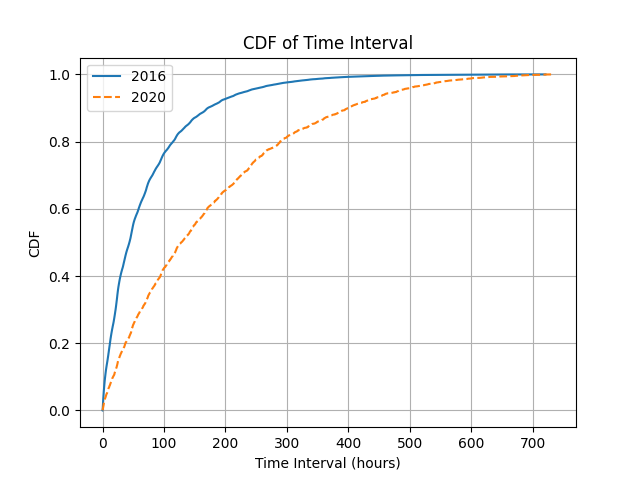
可以将超级用户的骑行间隔数据记录为：

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。

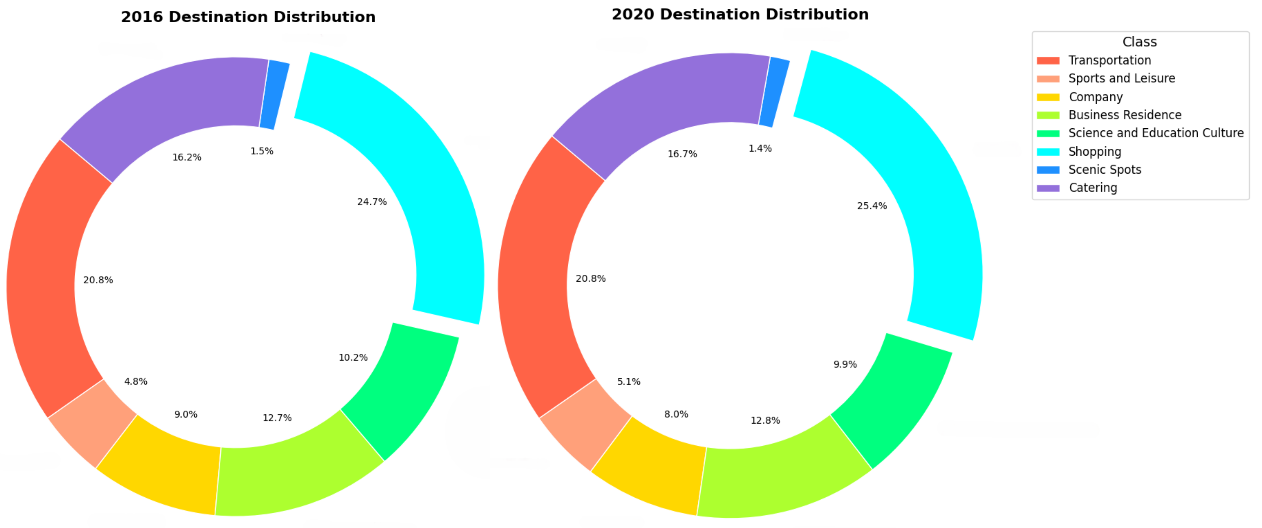
2016年的累计分布曲线在 200 小时（约间隔 8 天）处具有明显拐点，数据相对集中于该拐点左侧，呈现出明显的短间隔高频特征，即多数用户一周内要多次骑行共享单车。2020年所有时间间隔的CDF值都小于2016年，这意味着在2020年，用户两次骑行之间的间隔时间总体上比2016年长，考虑到此时处于疫情期间，人们的出行需求受到一定程度的抑制。但是，2020年的曲线没有明显拐点，呈现出一种平缓增加的特征，这说明2020年的两次骑行间隔分布相比2016年更加分散，用户骑行共享单车的通勤模式更为多元，共享单车逐渐为更多人所接受。

4.2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：

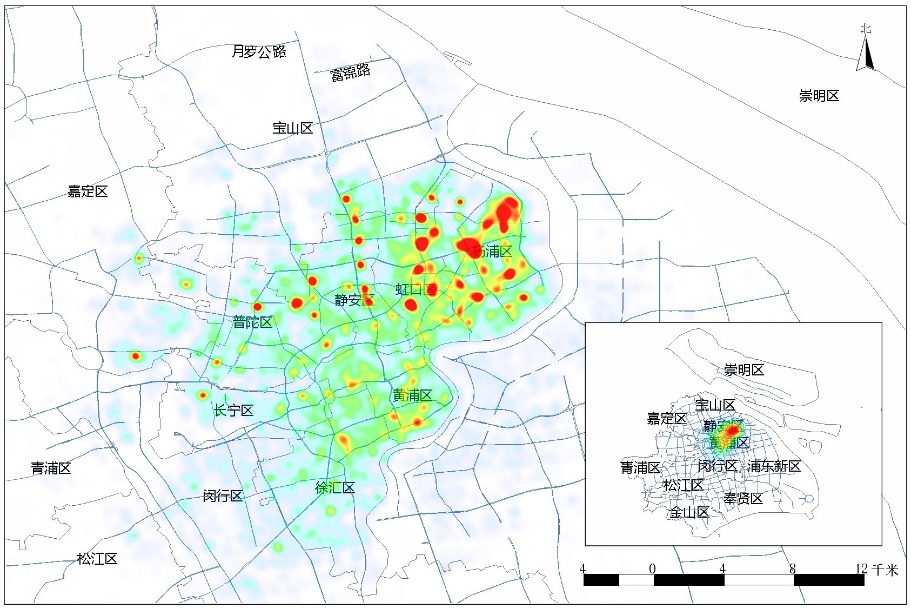


**图4.5 目的地占比分布图**

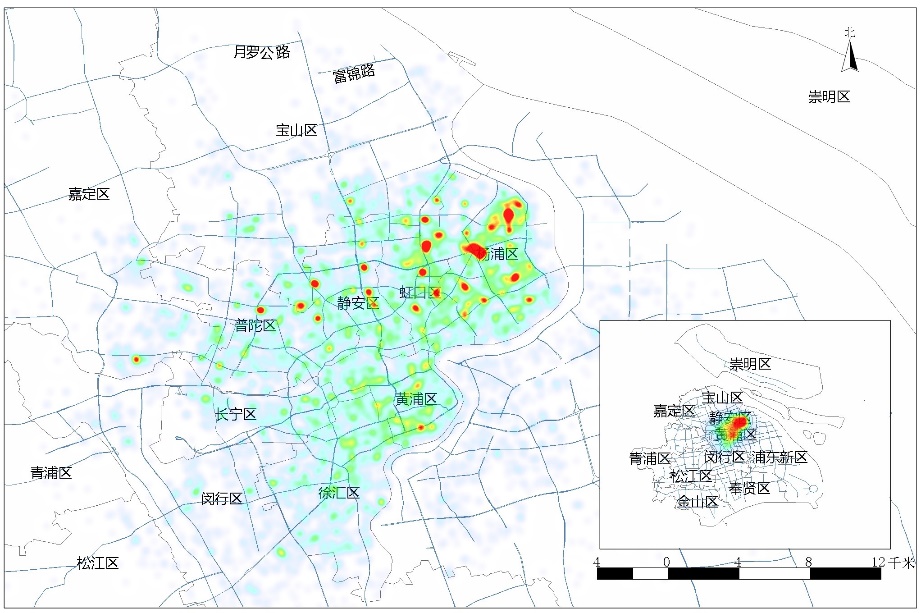
2016年与2020年骑行目的地总体相似，购物服务、交通设施服务和餐饮服务为主要目的地，占比近六成，反映了上海市商业活动活跃、交通发达、生活便利的特点。科教文化服务、商务住宅和公司企业也是较多用户的选择，占比近三成，表明上海市拥有较多较密集的院校及企业。风景名胜和体育休闲服务占比最少，可能与人口密集、生活压力大有关。

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4.6 2016年骑行目的地热力图**



**图4.7 2020年骑行目的地热力图**

总体而言，2016年至2020年上海市骑行目的地变化趋势反映了该区域经济社会发展相对稳定。未来可以考虑加强旅游资源开发、优化商业布局、提升城市宜居性，为市民提供更加丰富多彩的骑行体验。

4.2.1 通勤轨迹分析

该算

4.3本章小结

该算

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1

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