中图分类号：TP391 论文编号：1028716 22-SZ085
学科分类号：085212

硕士学位论文

移动对象数据清洗和质量评估方法研究

研究生姓名 方成龙
专 业 类 别 工程硕士
专 业 领 域 软件工程
指 导 教 师 许建秋 教授

南京航空航天大学

研究生院 计算机科学与技术学院

二〇二二年三月

Nanjing University of Aeronautics and Astronautics
The Graduate School
College of Computer Science and Technology

Research on Data Cleaning and Quality Assessment Methods of Moving Objects

A Thesis in
Software Engineering

by
Fang Chenglong

Advised by

Prof. Xu Jianqiu

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Engineering

March, 2022

承诺书

本人声明所呈交的硕士学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京航空航天大学或其他教育机构的学位或证书而使用过的材料。

本人授权南京航空航天大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本承诺书）

作者签名：_____

日　　期：_____

摘　要

近年来，GPS 定位技术和存储技术的发展，使得移动对象数据的应用越来越普遍。由于设备故障、传感器错误以及恶劣环境等因素的影响，采集的原始数据中通常包含错误和异常信息。这些异常信息严重影响原始数据质量，造成低质量数据无法满足科学研究的需求。因此，要彻底清洗原始数据的异常信息，从而为数据查询和分析提高数据质量，提高结果的可靠性。同时，对数据进行快速的质量评估可以从海量的原始数据中获取高质量数据。高质量的移动对象数据对于智慧城市、无人驾驶技术的发展具有巨大的推动作用。（1-2 句话介绍这个数据清洗的技术难度）

本文针对移动对象数据信息，实现了不同类型的异常检测、修复算法以及对数据质量进行快速评估算法。主要的研究工作如下：

（1）详细深入分析了现有移动对象数据并以此总结数据中存在的各种异常信息。对于（最为常见、最为典型、最复杂的）重复数据和无序数据的处理，提出了基于堆排序的最大无序偏移算法，极大的提高了排序去重的效率。针对缺失数据和无效数据的填充，提出了基于近邻数据的填充算法，利用周围正确的数据变化提高填充的准确率。对于漂移数据和模糊数据的修复，通过范围约束和数据变化最大概率组合的算法遵循数据修改的最小原则，极大的维护了原始数据的特征。（实验结果）

（2）为了快速获取原始数据集中高质量的数据，保证原始数据的特征，本文提出了基于 AHP 的多属性层次结构模型对移动对象数据质量进行评估。通过分析移动对象数据的特征和属性构建层次模型，计算各属性的影响权重对数据进行质量评估。为了提高算法评估的效率，在此基础上提出了基于多分段的快速连续抽样，在稳定质量评分的前提下减少数据的采样来提高评估效率。（在什么数据规模下）对算法的有效性进行了实验验证，数据质量评估的效率提高了约 50%。

（3）为了清洗原始数据集中的异常数据，提高数据质量，设计了针对移动对象数据异常检测和修复工具 GPSClean。该工具是嵌入在开源的可扩展移动对象数据库 SECONDO 中。将原始数据集上传到工具中，经过数据预处理、数据填充、数据修复、数据转换等操作，则可以获取清洗之后高质量的移动对象数据。工具将清洗前后的数据分布通过界面展示，可以直观的看出异常数据清洗的效果。通过类型转换模拟车辆历史运动轨迹，对于城市交通预测具有非常大的实用价值。

关键词：移动对象，轨迹数据，数据清洗，质量评估，清洗工具

ABSTRACT

In recent years, the development of GPS positioning technology and storage technology has led to a surge in the number of applications of moving object data. Raw data usually contain abnormal information caused by equipment failures, sensor errors, and environment influences. Low-quality data fails to support application requirements and therefore raw data will be comprehensively cleaned before usage. Therefore, there is an urgent need to comprehensively clean the abnormal information in the raw data to improve the data quality. At the same time, a quick quality assessment of the data can obtain high-quality data from a large amount of raw data. High-quality moving object data will greatly promote the development of smart cities and driverless technology.

According to the moving object data, this paper implements the corresponding anomaly detection and repairing algorithm and the fast assessment algorithm for the data quality. The main research work are as follows:

(1) Through analyzing the existing moving object data and summing up various abnormal information in the data. For the processing of duplicate data and disordered data, a maximum disorder offset algorithm based on heap sort is proposed, which improves the efficiency of sorting and deduplication. For missing data and invalid data, this paper proposes a filling algorithm based on nearest neighbor data, which improves the accuracy of filling through the change information of correct data around abnormal data. For the repairing of drift data and fuzzy data, to follow the minimum principle of data modification, the characteristics of the raw data are greatly maintained through the algorithm of combining range constraints and the maximum probability of data changes.

(2) To quickly obtain high-quality data in the raw dataset and ensure the characteristics of the raw data, this paper proposes a multi-attribute hierarchical structure model based on AHP to evaluate the data quality of moving objects. The paper constructs a hierarchical model by analyzing the characteristics and attributes of the moving object data, and calculates the influence weight of each attribute, so as to evaluate the quality of the data. To improve the efficiency of algorithm evaluation, a fast continuous sampling based on multi-segment is proposed on this basis, and the sampling of data is reduced under the premise of stable quality score to improve the evaluation efficiency. The paper verifies the effectiveness of multi-segment continuous sampling evaluation through experiments, and the evaluation efficiency is increased by about 50%.

(3) To clean the abnormal data in the raw dataset and improve the data quality, GPSClean is designed to detect and repair the abnormal data of moving objects. The tool is embedded in the open source moving object database SECONDO. By uploading the raw dataset to the tool, after data

preprocessing, data filling, data repairing, data conversion and other operations, we can obtain high-quality moving objects after cleaning data. The tool displays the cleaned data and the raw data distribution through the interface, which can intuitively see the accuracy of abnormal data cleaning. The historical trajectory of vehicles is simulated through type conversion, which has great practical value for urban traffic prediction.

Keywords: Moving Objects, Trajectory Data, Data Cleaning, Quality assessment, Cleaning tool

目 录

图表清单

注释表

P	GPS 轨迹数据集	P'	数据集 P 修复之后的数据
p_i	数据集中第 i 个记录	c	最大无序偏移
v_i	p_i 记录的速度	lon, lat	经纬度
$\gamma_{lon}, \gamma_{lat}$	经纬度范围	T_f	数据采集频率
L_e	数据误差范围	N	数据集记录数
$Dist(p_i, p_j)$	p_i 与 p_j 记录之间的距离	W	属性权重
T	层次结构模型目标层	S	层次结构模型准则层
A	层次结构模型方案层	M	不同属性数据模型
λ	特征值	CI	一致性指标
CR	一致性比率	Q	数据评分

缩略词

缩略词	英文全称
AHP	Analytic Hierarchy Process
AR	Autoregressive
ARMA	Autoregressive Moving Average
CFD	Conditional Functional Dependencies
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DTW	Dynamic Time Warping
EWMA	Exponential Weighted Moving Average
GK	Gauss-Kruger
GPS	Global Positioning System
MA	Moving Average
NNF	Nearest Neighbor Filling
RCSWS	Range Constraints and Sliding Window Statistics
RFID	Radio Frequency Identification
RMS	Root Mean Square
SNM	Sorted Neighborhood Method

第一章 绪论

近几年定位技术和存储技术的发展，使得越来越多的移动对象数据得以保存和研究。由于定位设备在各个领域的普遍使用以及地理环境和气候因素的影响，采集的原始 GPS 数据质量参差不齐并包含各种类型的异常数据信息。低质量的移动对象数据中存在大量的异常和误导信息，这给数据分析，城市规划，道路建设等领域带来了诸多的困扰。高质量的移动对象数据具有准确的数据信息不仅给实验分析提供了基础保障，对于科学技术的发展也起到了推进的作用。为了更好的利用现有采集的移动对象数据，则需清洗原始数据中的异常信息来提高数据质量。同时，为了从海量数据中快速获取高质量的原始数据信息，对大规模数据进行快速的质量评估也显得十分重要。

1.1 研究背景

随着大数据处理和存储技术的发展，各个研究领域的数据量都与日俱增。当今，低质量的数据已经不能够满足科学研究和工程建设的需求，机器学习和人工智能技术的发展对高质量数据的要求越来越强烈。高质量的数据是科学研究的基础，对于一些存在异常的数据在工程中使用则会带来巨大的经济损失。同样，高质量的 GPS 数据会给很多科研领域，如路况预测[1]，轨迹分类[2]，旅行地点推荐[3]等带来便利。

数据在现代社会就像石油一样珍贵，对于各个研究领域，高质量的数据是科学研究发展的基石。移动对象数据在日常生活中每时每刻都在产生，不管是交通设备还是生活用品产生的移动对象数据都存在着或多或少的异常数据信息。由于地理环境以及气候因素的影响使得定位设备再精确也会产生一定的异常数据，这些异常数据严重的影响原始数据集的质量，阻碍科学研究的推进。不同定位设备的误差以及环境因素的影响产生的异常数据的类型也有所不同，这些异常数据主要包括重复数据、无序数据、缺失数据、无效数据、漂移数据、模糊数据等。为了能够更好的利用现有的移动对象数据推动科学研究的发展，清洗数据集中的异常信息提高数据质量则是非常重要且具有挑战的事情。

在海量的移动对象数据中，并不是所有的数据都具有科学研究的价值。低质量的数据可以通过异常数据清洗来提高数据质量，但是相比于同等情况下原始数据质量本就很高的数据集，对于低质量的数据则可以无需对其进行修复。以原始高质量的移动对象数据集进行科学实验研究则更加的具有说服力。为了能够快速有效的利用现有的移动对象数据集，获取数据集中质量较高的数据，则需要对数据质量进行评估。移动对象数据质量的评估不仅需要考虑数据的表面属性信息，还需要分析数据潜在的一些特性。对不同类型的数据属性分析其在整个数据集中的占比，并提供合理的影响权重是非常具有挑战性。在大规模的原始移动对象数据集中通过快速的质量评估来获取较高质量的数据不仅节省了对低质量数据异常修复的时

间，对于原始数据质量要求较高的科学研究也提供了非常大的帮助。

目前，移动对象数据异常清洗已经有了一些研究工作。针对不同的异常数据类型具有相应的清洗修复方法。对于单点异常的检测和修复方法目前有很多，主要是通过基于平滑处理和基于约束处理的方法。基于平滑处理的技术主要包括 MA、AR[4]、ARMA[5]等，通过计算最近观测到的 n 条数据记录的加权平均值作为数据的预测值来对异常数据进行检测和修复。在基于约束处理的方法上，Song 等人[6]提出了基于速度约束的方法来判断单点数据的变化是否在前一个点的基础了超过了变化的最大值来判断异常数据。对于连续异常点的判断由于数据点较多判断的复杂程度也相应的提升。Keogh 等人[7]提出了一种序列异常检测方法，通过计算子序列和其最接近的序列之间的匹配距离来判断子序列的异常数据。在其他类型的异常清洗方法上，DBSCAN[8][9]算法通过数据的密度将数据划分为一个聚类来对异常数据清洗也起到了不错的效果。

1.2 选题依据和意义

近年来，信息技术的发展，推动着社会经济实现快速增长。全球信息化进程的推进，使得互联网技术发展迅速，各行业的数据量也迅猛增长。数据犹如信息时代的矿产和石油，推动着现今社会技术的发展。高质量的数据是人工智能技术发展的前提，作为解决科学问题的基础。通过利用对数据的解释和提炼，来挖掘数据中潜在的一些价值提高算法的可行性。但是，原始数据中或多或少会存在一些异常信息降低数据质量，低质量的数据严重干扰科学实验的准确性。为了获取高质量的原始数据，则需要对数据中的异常进行检测和修复，并对数据质量进行分析和管理，在获取大规模数据的同时保证数据质量的稳定性。

在大多数的情况下，企业会获得大量的数据，但是对数据却知之甚少，包括数据存在的一些特性以及数据优劣的分析。这就使得许多企业具有一种尽管有数据但是却不能让数据产生其应有价值造成数据浪费的通病。完整的高质量的数据对于模型的训练和学习起到的作用是非常有利的，能够从数据中发现数据潜在的规律从而达到较好的训练效果。高质量的数据分析对金融领域数据分析人士而言是必不可少的，一点微小的异常数据都会引起股市的动荡对社会产生巨大的影响。若是获取的数据中存在大量类似这种异常的数据，那么低质量的数据对于股市的分析几乎没有帮助，甚至影响后续股市波动的判断。若是能够准确获取每一次的数据，那么通过专业的分析则能对股市的每一次波动进行较好的预测。判断数据质量的优劣，并通过判断的情况来对质量较低的数据进行异常的修复从而获取高质量的数据，从中获取潜在的价值已经成为了目前技术发展的重要问题。

移动对象数据的规模之大，以一个城市的出租车运行轨迹来说每日产生的数据量都在 TB 甚至 PB 级以上，设备采集频率的大小也同样影响着数据量的大小。虽然每天都会产生非常多轨迹数据，但是由于多种原因如设备定位误差，人为误操作，高楼隧道及恶劣天气的影响使得数据中存在较多的异常数据信息，这些异常信息严重影响了数据质量，使得大规模的移动对象数据不能进行很好利用。在数据挖掘领域，对于移动对象数据的分析大约 50%

以上的时间都是花费在数据的预处理上，为了保证实验的可行性和准确性提高数据质量是势在必行的一项任务。由于各种原因的影响使得异常数据的种类也多种多样，主要包括数据的重复、无序、缺失、无效、漂移、模糊。通过研究和探索数据异常检测和修复算法，对数据中存在的异常信息进行相应的清洗从而能够得到高质量的数据。高质量的移动对象数据对基于位置服务的应用如导航定位，路径优化，地点推荐中发挥了巨大的作用，不仅推动着科学研究的进步，同时也改善了人们的生活提高了生活的质量。

大数据以及人工智能等技术的兴起，其中大部分的算法和实验模型都是在高质量数据的基础上实现的，如果把关注点只放在基于高质量数据的模型与算法，而不关注数据质量，实验获取的实际结果将会有巨大的偏差。为了能从现有的数据中挖掘出潜在的价值，提高数据质量成为了当下的迫切需求。移动对象数据的质量评估和异常修复技术获取的高质量数据能够为城市计算，交通预测[1]，移动对象数据挖掘与分析[2]，异常交通模式[10]，空间索引等领域提供可靠的研究基础和实用价值。

1.3 主要研究工作

日常生活中每时每刻都在产生移动对象数据，但是这些数据中往往存在着或多或少的异常数据信息，这些异常数据严重影响原始数据的质量，对于基于数据研究的领域带来了巨大的挑战。异常数据清洗则是提高数据质量的一种非常有效的方法，清洗之后干净的数据不仅对于人工智能，大数据分析能够提供良好的实验数据基石，而且对于科学技术的发展起到了推进作用。移动对象数据规模之大，但是并不是所有的移动对象数据都是存在研究价值的，这就需要从海量的数据集中获取快速的获取高质量的数据，对移动对象数据进行质量评估则是非常重要的一种措施。

针对移动对象数据清洗和数据质量评估，本文具体的研究工作如下：

（1）对于移动对象数据进行异常检测并对不同的异常信息提出相应的异常修复算法。本文所分析的异常类型主要为数据的重复、无序、缺失、无效、漂移、模糊。不同的异常数据存在不同的数据特点，针对异常数据的特点本文提出了相应的算法对其进行检测和分析。为了保证原始数据的特点遵循原始数据修改的最小原则，对于异常数据的修复，则保证在正确数据的最大可接受范围内。

（2）针对移动对象数据质量评估，对移动对象的数据特征进行了分析，并从数据多属性角度出发分析数据表面和潜在存在的一些属性并建立对应的属性数据模型。为了使得数据中各属性对应数据质量的影响权重更有说服力，提出了基于 AHP（Analytic Hierarchy Process）的多属性数据质量评估方法。以层次分析来比较属性之间的重要性关系建立判断矩阵，并将其最大特征值对应的归一化特征向量作为属性的影响权重。为了提高对规模较大数据集的质量评估效率，提出了基于多分段连续抽样评估算法，在稳定数据质量评分的前提下，多分段连续抽样评估相比于对整个数据集的评估在效率上有了明显的提高。

（3）设计移动对象数据清洗和修复的方案，实现异常数据清洗的可视化，并在数据库 SECONDO 中实现了数据清洗的工具。分析不同类型的移动对象数据集，根据数据属性导入原始数据集并通过 GUI 展示原始数据的分布。通过不同的异常检测和修复算法对原始数据集进行清洗，通过比较清洗前后数据的分布变化可以直观的看出清洗之后的数据更加的符合实际情况。将清洗之后干净的数据进行数据转换为移动对象数据类型 MPoint，则可以观察物体的历史运动轨迹，对于移动对象数据的历史轨迹分析提供了帮助。

1.4 本文的组织结构

本文分为六个章节说明移动对象数据清洗及质量评估，组织结构如图 1.1 所示。

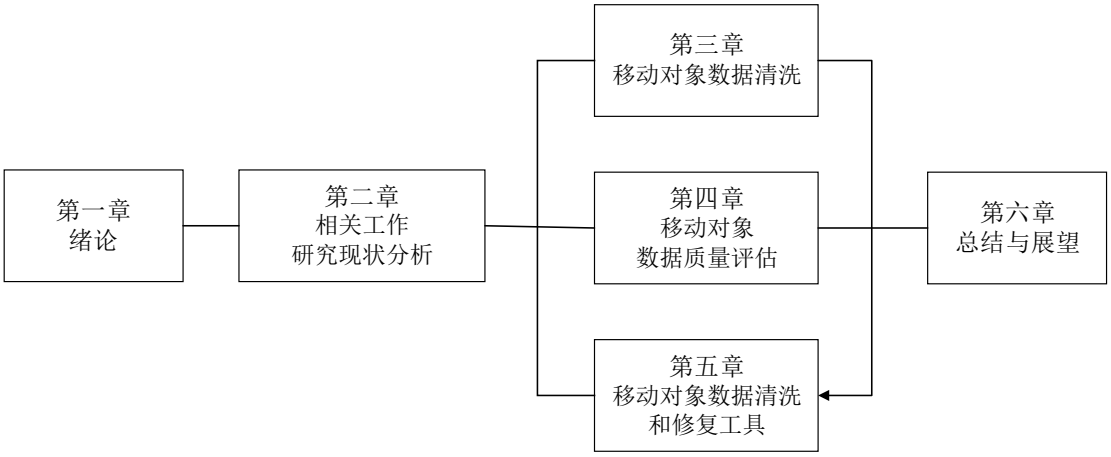


图 1.1 文本组织结构图

各个章节的内容介绍如下：

第一章对本文的研究背景、选题依据和意义、主要研究工作进行了介绍，然后对文章的组织结构进行了说明。

第二章对移动对象数据清洗和质量评估的研究现状做了详细的介绍。首先介绍了数据清洗领域的一些工作，包括数据异常检测及数据异常修复。然后介绍了在数据质量评估领域现有的一些工作，对不同数据进行质量评估的差异，最后对现有一些异常检测及修复的工具的应用场景进行阐述。

第三章针对移动对象数据中的异常数据进行检测和修复。分别对不同类型的异常如无序数据、重复数据、漂移数据、无效数据等提出了对应的异常检测和修复方法。为了保证原始数据的特点，基于数据的最小修改原则提出了基于约束和概率统计模型对漂移数据进行修复的方法，以最小的数据修复代价提高了原始数据的质量。

第四章对移动对象数据质量进行了评估。由于移动对象数据存在的属性特征复杂，通过分析数据属性建立数据模型来分析原始数据中各种类型数据的占比。在对各类型的数据属性对整体数据质量影响的权重上面设计了基于 AHP 的多属性质量评估方法，应对海量的移动对象数据为了提高质量评估的效率，提出了基于多分段的连续抽样方法保证了数据质量评分稳定的同时提高了质量评估的效率。

第五章介绍了移动对象数据清洗和修复工具 GPSClean。该工具嵌入在移动对象数据库 SECONDO 中经过数据清洗和修复来提高数据的质量，并将数据转换成移动对象类型用以观测清洗之后数据的历史运动轨迹，通过 GUI 可以直观的看出清洗之后的数据质量明显提高，对移动对象历史轨迹的研究提供了巨大的帮助。

第六章总结了文章中针对现有移动对象数据问题提出方法的创新和不足。通过方法中现有的一些不足和待改进的方面展望后期需要进行的研究工作。

第二章 相关工作研究现状分析

近年来，定位技术和大数据的快速发展使得移动对象数据的规模越来越大，在各种研究邻域的应用也越来越多。移动对象数据中潜藏着无限的数据价值，对于城市规划、智慧交通、兴趣点推荐等方面起到了十足的帮助。由于人为因素以及各种环境因素的影响使得定位设备采集的数据中存在着大量的异常数据信息，这些异常信息严重的干扰数据的质量使得一些低质量的数据难以进行使用和研究。虽然移动对象的数据规模大，但是其中能够较好的利用到科学研究和生活帮助上的数据则较为稀少。因此，对移动对象数据集中的异常数据进行清洗则成为了科学研究的必要条件。通过数据质量评估不仅可以快速的获取原始数据集中的高质量数据提高科学实验的可行性，而且保证了原始数据的特征使其能够发挥更大的研究价值。下面介绍关于移动对象数据清洗和评估的相关工作研究现状。

2.1 移动对象数据清洗现状

现有移动对象数据清洗主要包括了对异常数据的检测和修复两个部分。由于异常的数据信息有多种类型因此对于不同类型的异常数据具有相应的检测和修复算法。异常数据中主要包括了重复数据、无序数据、缺失数据、无效数据、漂移数据、模糊数据等。对于重复数据和无序数据的检测和修复相比于其他类型的异常数据而言相对简单，异常数据清洗的重点则更加的偏向于对其他类型异常信息。文中异常数据检测和修复的相关工作介绍也是重点针对这类异常进行说明。

2.1.1 移动对象数据异常检测

异常数据检测是异常修复的基础，只有异常检测功能相对全面才能提高数据的修复精度，提高数据的质量。目前，各研究领域在数据异常检测方面的工作做的相对较多。在异常数据中，重复数据和无序数据的检测相对简单。重复数据主要分为两种类型一种为连续性的数据重复，另一种为间断性的数据重复。对于连续性的重复数据而言只需要判断当前检测数据前后数据是否相同，而对于间断性的数据重复检测会相对复杂。首先可以对数据进行排序将间断性的重复转变为连续性的重复，但是在数据去重之前排序的时间消耗较大。现有的排序-检测-合并的 SNM[11]算法，通过滑动窗口的方式对窗口内的数据进行排序比较来去重，但是重复数据的数量是变化的，对于窗口大小的设定成为了一项难题，窗口过小会使得重复数据的检测不够全面，窗口过大同样也会导致窗口中较多无意义数据之间的比较，浪费了大量的计算资源。以 Hash 方法来进行判断间断性的重复数据相比于 SNM 算法来讲，在算法的去重效率上有了显著的提升，但是需要有足够大的内存空间存放移动对象数据，对去重之后的无序原始数据，则还需要进行数据的排序操作。

在对单点数据的异常检测方面，最主要的异常分析方法是采用构建数据预测模型，通过将模型中获取的预测值与原始数据中的观测值对比，来确定数据的异常情况。文章[12]通过以一段时间戳上数据的中位数为预测模型，来预估时间线上当前检测数据的数值，进而计算预测值与观测值之间的差，将差值和设置的误差阈值进行比较，来检测当前数据的异常情况。对于使用中位数与阈值比较的方法来说，可以很好的检测偏离原始数据较大的异常数据信息，但是在误差阈值的设置上面则需要根据具体的数据情况来分别处理。移动平均 MA 算法[17]也被广泛的应用于单点异常的检测。通过计算最近观测到的 n 个时间序列的平均值作为当前检测数据的预测值来判断异常点，以平滑的处理方法来检测异常数据。在对移动对象数据进行检测中可以根据两点之间的距离变化情况来判断数据的变化是否平滑，对于数值变化情况较为陡峭的数据标记为异常数据。对于加权移动平均模型 WMA 来说则是根据对当前检测数据的影响情况对数据中不同位置的观测数据赋予不同的权重计算当前的预测值，相比于 MA 算法，WMA 使用不同的影响权重则可以起到更加精确的异常检测效果。Song 等人[6]提出的基于速度约束的 SCREEN 方法，通过对前一段时间序列数据的变化情况来约束当前数据的变化范围。判断当前检测的数据是否在前几个数据正常变化约束的范围内，将已经超出约束范围的数据作为已经识别的异常数据信息。SCREEN 算法可以很好的检测偏离原始数据较大的异常数据，但是由于方法中设置的约束范围是数据速度变化最大的情况，因此当数据变化小于速度最大阈值的偏离较小的异常则难以对其进行检测。自动回归模型 AR[4]和外因输入自回归模型 ARX[13]在众多领域都被广泛的应用于数据的异常检测，对于移动对象数据的异常检测通过平滑的方式可以很好的检测出偏离原始数据较小的异常信息，但是过度平滑处理会使得原本正确的数据被修改造成过度修复的现象。

在对连续数据的异常检测方面，连续的异常数据有多种情况主要包括整体数据漂移异常和连续单点数据漂移异常。连续数据整体漂移的异常表现为在移动对象数据集中存在一段时间轨迹数据正常但是对于原始数据的运动估计具有明显的偏移，连续单点数据漂移异常表现为连续的多个数据点由于地理环境或者气候因素的影响导致定位的数据点没有规律可循，以各个数据点作为参考相互检测互为异常。对于连续数据的异常检测，Keogh 等人[7]提出了一种序列异常检测的方法，通过计算数据集中子序列和其最接近的序列之间的匹配距离是否大于设定的阈值来判断子序列是否为异常序列。通过子序列的方式可以很好的检测移动对象数据中的异常信息，对于整体数据漂移或者是连续单点漂移的异常具有很好的异常检测效果。在整体数据漂移中，通过子序列的比较可以较好的发现连续的数据集中在整体漂移的数据时间段呈现出断层的效果。在连续单点的数据漂移中，连续的单点异常会造成原始数据集在对应时间段显示出波动幅度较大的数据现象，通过子序列的比较数据的波动幅度较大则可以对其进行很好的检测。但是序列异常检测算法通过数据集中各个子序列之间的匹配距离计算会造成较大的比较浪费和时间消耗。当然，单点数据异常检测方法也可以检测多个数据异常，对于连续数据异常也具有检测效果。在通过基于密度的异常检测方法上，DBSCAN 算法[8][9]

通过数据的密度来将数据划分为一个聚类。通过聚类计算出数据中的离群点，将偏离较大的数据视为异常数据。对于比较集中的数据类型异常检测效果比较好，但是将检测到的异常点进行抛弃则破坏了原有数据的完整性。在通过机器学习方式来检测数据集中的异常点方面，通过生成对抗网络[14][15][16]的方式通过同时训练了两种模型，其中一个生成数据的分布模型，一个生成判别数据异常信息的判别模型。然后以反向传播和梯度下降来计算两个模型的误差并更新权重，使得生成的两个模型的损失最小化。

2.1.2 移动对象数据异常修复

数据数量和数据质量就像硬币的两面，对于数据管理同样重要[18]。相比于将检测到的异常数据点作为无用的噪声丢弃[19]，数据的异常修复避免了大量连续的数据丢失，保存了原有数据的完整性。

为了充分的利用现有的移动对象数据集从中获取高质量的数据，则需要对数据中的异常进行清洗，各个邻域的数据修复方法也是层出不穷。对于上文提及到的数据异常检测的方法大部分也可以进行数据的异常修复。在基于平滑的数据清洗技术方面，MA 序列清洗算法[17]通过在计算一段时间内数据的平均值作为数据的修复值来对异常的数据进行清洗。AR 模型[4]的数据清洗算法是将当前数据作为回归变量，通过前 k 条数据记录的加权线性组合来描述数据的回归过程。文章[20]-[22]通过卡尔曼滤波模型对数据进行一些清洗。卡尔曼滤波模型不仅能够处理平稳信号还能处理非平稳信号和多维信号。平滑的异常修复算法对于一些异常值比较小的数据修复的准确率是很高的，但是对于一些异常值较大的数据来说，基于平滑的修复算法会由于对峰值数据平滑过渡的过程使得原始数据集中正确的低值数据被修改，造成过度修复异常数据量不降反增的现象。

在基于约束的数据清洗技术方面，顺序依赖 ODs 的方法通过数据之间的依赖关系来判断数据是否违反了相应的约束，比如一辆汽车的油耗只能为正数，如果当数据为负数的时候则违反了约束将其视为异常的数据。对于移动对象数据中的经纬度信息则可以对其进行约束经纬度必须是在[0, 90]或[0, 180]的范围内，根据轨迹的数据信息可以通过行驶的城市，道路等来缩小约束的范围更加精确的对数据进行检测和修复。Lopatenko 等人[23]提出了基于拒绝约束 DCs 的方法来对数值型时序数据进行异常的检测和修复。其原理也是和 ODs 大致相同，通过数据之间的依赖关系判断并修复数据。论文[24]提出了顺序依赖 SDs，通过比较时序数据中连续数据之间的差值来判断数据存在的异常。该方法通过当前数据点和前一个数据点之间的差值是否满足设定的阈值进行初步的判断，进而判断数据是否在对应的区间上来对数据进行修复。论文[6]提出了速度约束的方法 SCREEN，设定窗口大小利用当前检测数据窗口内的正确数据依据最大的数据变化情况设定对当前检测数据的变化范围，对于窗口内多个正确数据形成多个对当前检测数据的约束范围，当检测数据不在多个约束范围内根据数据修改的最小原则[25]将异常的数据修复到所有的约束范围中距离当前点最近的数据点。SCREEN 算法很好的遵循了数据修改的最小原则，以最小的数据改动使得异常数据的修复满足了窗口内所

有正确数据的约束情况。该算法能够很好的检测和修复违反速度约束较大的漂移数据点的异常情况，但是对于在速度约束内的较小的漂移异常则难以进行检测和修复。其他的基于约束的数据修复方法还有方差约束[26]，相似规则约束[27]等。

在基于统计的数据清洗技术方面，Zhang 等人[28]提出了基于最大似然估计的方法来对时序数据中的异常数据进行了清洗。该方法通过统计时序数据中数据变化快慢的分布情况做一个概率聚合函数，并根据整体数据中数据变化的最大概率情况来对异常数据进行修复。对于移动对象数据的检测和修复，基于最大似然估计的方法可以很好的检测原始数据集中物体运动速度和加速度的概率分布，分析数据集中检测到的概率分布，将概率小于设定的最小概率阈值的数据标记为异常数据，并使用最大概率数值情况对异常数据进行修复。例如在概率分布中，概率低于阈值的数据对应的速度为 200km/h，最大的概率数据对应的速度为 100km/h，由于异常的 200km/h 的速度在整个数据集中出现较少因此以较低的概率可以对其较好的检测，并以最大的概率数据 100km/h 的速度对其运动的数据进行修复。最大似然估计方法对于漂移较大，分布概率较小的异常数据修复的差值也较大，在对于漂移较小，分布概率也较小的异常数据的修复则相比于约束方法具有良好的修复效果。论文[29]-[31]中都使用了极大似然估计的方法对不同种类的数据进行了异常的检测和修复。

2.1.3 其他数据类型清洗

高质量的数据对于各个领域的研究工作都起着至关重要的作用，只有在数据准确的前提下才能降低风险发生的概率。下文介绍了其他领域关于数据清洗的相关工作。

异常数据的种类很多，不同研究领域中主要针对的异常信息也有所不同。在对医疗数据的异常检测修复中，武森等人[32]提出了在不完备的数据集上进行数据填充的方法，较好的解决了医疗数据中的缺失数据问题。医疗数据信息的准确性是关系人体健康的重要因素，通过人体健康医疗大数据[33]的分析，从数据中预测潜在的一些健康风险并对其进行有效的预防成为了医疗发展的趋势。杨等人[34]提出了一种网络流量异常检测模型，通过对数据的全局感知提高异常检测的性能。邵等人[35]基于生成对抗神经网络提出了时间序列异常检测模型，对复杂的时间序列建模并以数据填补方法合理的对缺失值进行处理。对于传感器中异常数据的处理，Zhuang 等人[21]使用卡尔曼滤波器模型来预测传感器的数据并使用 WMA 对异常的数据信息进行平滑的处理来修复异常数据。Marczak 等人[20]将增强的卡尔曼滤波用于状态空间模型鲁棒估计的数据清理。在天文时空数据的异常检测上，Fang 等人[36]利用中位数均值的方法快速的获取星等数据的峰值，通过将峰值和整个数据集的中位数均值作比较将差值大于设定阈值的作为疑似的异常数据点，再对疑似点的左右两边数据进行判断进而分析出微引力透镜和恒星耀发两种异常的数据现象。通过中位数和均值的快速异常定位方法相比于现有的使用动态时间规整 DTW 算法的异常数据匹配节省了大量的时间。DTW 算法也常用于轨迹数据相似性的比较，通过两个相似的轨迹数据进行匹配可以将一个轨迹数据作为参考

用于检测和修复另一个轨迹数据中的异常信息。

对于文本中的异常数据检测，函数依赖 FDs 可以很好的检测文本中的异常数据信息，例如江苏省南京市对应的邮编为 210000，则存在一对（江苏省南京市->210000）的函数依赖，在对其他地址数据的检测中如有江苏省南京市但不是对应邮编的则被检测为异常数据。在函数依赖的基础上，Fan 等人[37]提出了条件函数依赖 CFD 对关系数据库中的异常信息进行处理。CFD 通过执行语义相关数据值来捕获数据一致性，给关系数据库的异常清洗提供约束条件，提高数据质量。Rammelaere[38]等人将 CFD 视为关联规则的扩展并提出了三种近似 CFD 的数据清洗方法，使用不同的组合模式挖掘分析函数依赖中的条件，提高了数据清洗的性能。Arocena 等人[39]在文中给出了一个新的对称性的数据质量约束方法，用于清洗数据库中引入的错误信息问题。在对重复数据的检测中，Naumann 等人[40]通过语义规则匹配的方法来解决重复数据的问题。时珉等人[41]在文中通过分析了光伏阵列异常数据信息的来源和特性，给出了利用滑动标准差的异常数据信息处理算法。文章[42]在分布式的环境中使用基于聚类的方法对电源数据进行异常检测，然后使用指数加权平均方法进行数据修复，通过将数据的预测值和观测值进行分析比较来判断数据是否异常。对于 RFID 的数据清洗，Jeffery 等人[43]提出了一种自适应的数据清洗方法，利用基于采样和平滑处理的技术提高了 RFID 数据的质量。Baba 等人[44]提出了室内 RFID 多变量隐马尔可夫模型 IR-MHMM 来捕捉室内 RFID 数据的不确定性并对其进行清洗。Gupta 等人[45]使用隐马尔可夫 HMM 去预测股票的价格走势，文章[46]也使用 HMM 对数据中存在一些异常数据进行检测和清洗。论文[47]使用了时空概率模型 STPM 去学习历史数据的一些数据特征，进而用来清洗当前的数据。

2.2 数据质量评估现状

数据清洗的目的是为了修复原始数据集中的异常数据，获取高质量的数据集。通过修复算法修复的异常数据相比于原始正确的数据具有一定的数据误差。通过对数据进行质量评估，从原始的数据集中获取高质量的数据，同等质量的未经修改数据集相比于算法修复之后的数据集而言具有较好的数据特征，更好的体现出了原始物体运动的轨迹。但是现有的数据质量评估标准还不完善，下文则从多个角度对现有的质量评估工作进行介绍。

大数据技术的发展使得各个领域对数据的重视程度越来越高，虽然数据规模较大，但是各行业对数据质量评估的研究工作却相对较少。各行业之间数据的不一致性，使得在数据质量的评估方面设定判断的标准是非常具有挑战的。Wang 等人[48]通过将出租车的 GPS 轨迹数据可视化直观的分析数据的质量，通过结合了半监督学习和可交互式视觉探索分析数据质量问题，并自动提取已知的问题。对于存在多个相同轨迹的数据质量比较方面也是非常的方便。通过半监督学习的方法不仅可以直接的判断数据中肉眼可以观察到的异常信息，对于数据密集部分难以肉眼观察和分析异常也给予了一定的帮助。Lee 等人[49]提出了一个新的基于窗口的方法计算一个质量控制参数来判断 GPS 数据的质量，并通过他们实现全局定位传感质量控制来提高定位的精度，解决 GPS 数据丢失和误差而费时的问题。文章[50]通过 GPS 日志信

息的数量和质量来判断旅行数据中 GPS 定位的数据质量问题，并将数据与调查产生的位置数据进行比较来对数据质量进行分析。论文[51]从多径效应、信噪比、定位精度等多方面对数据质量进行了分析，较好的利用现有的 GPS 数据定位的原理来分析数据的质量，通过定位设备分析接收信号的强弱来提高定位的精度。论文[52]从具有 GPS 定位功能的手机中提取混合位置的 GPS 数据进行了研究，并开发了一种利用本地环境信息的插补策略，通过将原始数据与估算数据与参与者的在线调查响应进行比较来评估所提出的估算策略进而分析估算数据的质量。

对数据分析的质量判断，Sadiq 等人[53]提出判断特定数据质量的主要指标，包括整体性、准确性、一致性等。目前为止，由我国信息标准化技术委员会提出的数据质量评价指标，主要为规范性、完整性、准确性、一致性、时效性、可访问性。不同的数据质量评价指标对应的质量也会有所不同，对于不同指标的侧重也会有所偏向。对于股票行业的数据在众多的指标中则更加的偏重于数据的时效性，因为及时的数据更新才能更好的显示市场的波动情况，虽然历史的数据信息非常的准确具有一定的参考意义，但是对于实时性的数据需求而言则显得没有数据时效性的影响权重大。现有 GPS 观测数据质量评价分级没有统一的指标，郑等人[54]以 2000 个点位的 GPS 观测数据根据正态分布的数学模型，确定了分级评价指标。古等人[55]通过 TEQC 软件对原始观测的数据进行转换，并对数据质量评估进行了分析。对于时间流的数据质量，Dasu 等人[56]针对流数据提出了四种约束类型的方法对数据质量进行分类识别可能潜在的数据异常问题，并通过使用统计失真的方法计算理想数据和真实数据之间的差值来判断数据的质量。

2.3 移动对象数据清洗工具及应用

由于对高质量数据的需求，使得各个领域数据清洗的工具有很多。高质量的移动对象数据不仅对于智能交通规划与优化[57]-[59]、道路规划[60]、城市管理[61]-[63]具有重要的意义，而且对大数据和无人驾驶技术的发展起到决定性的作用。

对于数据异常的检测和修复的工具上，Ding 等人[64]提出了一个时序数据异常检测工具 Cleanits，该工具提供了良好的界面交互，通过考虑数据中的多个特征属性并结合专业的领域知识对时序数据进行异常分析。Huang 等人[65]提出了一个新的框架 TsOutlier 用于检测和解释物联网数据中的异常数据信息。TsOutlier 使用多种算法检测时序数据中的异常数据，并支持数据的批量处理和流处理。Rong 等人[66]提出了一个基于平滑方法检测和修复异常的工具 ASAP，该工具主要用于解决传统数据库的清洗问题，保持了自适应优化降噪和趋势之间的权衡关系，但是 ASAP 违反了数据修改的最小原则导致数据失真，对于移动对象数据的清洗容易造成数据过度修复的现象。Yu 等人[67]提出了一个基于统计的概率交互式数据清洗系统 Piclean，使用低秩近似产生概率错误和概率修复，它隐式地发现并使用数据集列之间的关系进行清理。文章[68]提出了一个由概率推理驱动的整体数据修复框架 HoloClean，框架

利用输入数据的统计特性，依赖于完整性约束或外部数据源使用定量数据的修复方法修复异常信息。基于数据修复的概率统计分布，HoloClean 能够很好的对不同类型错误的数据集进行数据修复。Krishnan 等人[69]通过迭代清洗过程来训练统计模型并提出了清洗模型 ActiveClean，该模型允许在统计建模问题中进行渐进式和迭代式的数据清洗同时能够保持收敛性。ActiveClean 支持一类重要的模型（例如线性回归和 SVM），并优先清洗这些记录来影响数据的结果对结果进行分析，然后根据清洗的结果再对数据中的异常进行清洗。Wang 等人[70]针对社交网络的实体数据信息提出了数据清洗模型 EDCleaner，该模型将实体数据信息转化为带有属性标签的结构化数据。EDCleaner 中提出了对半结构化数据进行属性识别和数据归一化的方法，有效识别数据的属性标签关系，得到规范统一的结构化数据，并利用机器学习分类器进一步提高属性识别的准确率，最终形成高效准确的数据清洗方法能够有效的处理多数据场景。Huang 等人[71]提出了一个隐私感知数据清理即服务框架 PACAS，通过数据定价方案促进客户和服务提供商之间的沟通，在该方案中，客户发出查询，服务提供商以一定的价格返回干净的答案，同时保护隐私数据。PACAS 为一种新的数据清洗即服务模型，该模型允许客户端与托管敏感数据的数据清洗提供商进行交互。PACAS 有效地保护了语义相关的敏感值，并保证了隐私感知清理技术的准确性。

移动对象数据管理与分析对人们生活中的方方面面都起到非常重要的作用。通过路网信息以及 GPS 设备定位的功能，历史的轨迹信息不仅能够帮助导航寻路，还能通过周围的地理信息进行兴趣点推荐。对于交通规划和优化，通过分析道路历史移动对象数据信息对于道路的拥挤情况以及车辆类型能够及时的进行控制不仅能够很好的疏通车流量，还能避免因拥挤导致的交通事故。日常生活中不同种类应用的步数统计和将康助手通过分析人们的运行轨迹和运动情况来分析人体的健康状态，并对身体可能出现的一些状况进行预测起到良好的预防效果。如今大数据和人工智能技术的快速发展，推动着无人驾驶技术成为了现今社会研究的热点。高质量的移动对象数据给无人驾驶技术模型的训练提供了有力的帮助，通过对汽车历史轨迹的训练进而训练出足够能应对多种路况的数据模型，并通过实时的 GPS 数据定位给无人驾驶技术的发展提供了可能。移动对象数据也可以用于地图映射[72]，通过地图映射可以很好的检测采集的数据中是否存在异常数据并给予修复，同样对于没有及时更新的地图也可以使用较新的移动对象历史轨迹信息进行地图道路的映射从而达到相辅相成的作用。Cheng 等人[73]提出了深度卷积神经网络通过大数据训练进行数据特征提取实现自动分类和检测异常错误场景，自动的清洗来维护数据库的正确性。Qu 等人[74]基于移动对象轨迹大数据的分析，利用 GIS 的密度来获取旅游的热点区域，通过热点区域识别从而能够将设立的服务点能够为游客提供最大化的服务。

第三章 移动对象数据清洗

具有 GPS 定位功能设备的普及导致产生了大量的轨迹数据。由于环境因素以及设备采集精度的影响，使得收集的原始数据通常包含由设备故障、传感器错误和环境影响引起的错误和异常信息。海量的 GPS 数据可能会存在着大量的异常，这些异常信息会严重影响数据使用者的利益。相反，高质量的 GPS 数据不仅对数据分析、城市建设、道路设计具有重要意义，而且对大数据和无人驾驶技术的发展起到决定性的作用。

本章主要分析了在移动对象数据管理中出现的各类异常数据，通过分别对不同类型的异常数据进行检测修复来提高数据的质量。

3.1 问题描述

移动电话等支持 GPS 的设备的广泛使用使得位置数据的记录变得非常容易，因此收集了大量此类数据，使得移动对象数据非常普遍。由于设备故障、传感器错误和环境影响等多种因素，原始数据通常包含脏、异常和不正确的信息[43][75]，这将严重影响数据的使用，例如进行数据分析和查询处理。大量的异常不仅导致原始数据不能满足科学研究的需求，而且影响了交通预测和无人驾驶技术的发展。现有的一些异常清洗方法难以对 GPS 数据做出全面的异常修复。对 GPS 数据进行清洗主要有两个挑战：（i）异常数据检测和（ii）异常数据修复。事实上，异常数据的检测是异常修复的基础。相比于将检测到的异常数据点作为无用的噪声丢弃[19]，数据的异常修复算法避免了大量连续的数据丢失，保存了原有数据的完整性。提高异常修复算法准确率的前提是具有比较全面的异常检测算法，但是绝大部分的 GPS 数据的异常检测算法很难做出全面的异常检测。

本文主要检验了以下六种类别的异常数据信息，包括重复数据、无序数据、缺失数据、无效数据、漂移数据、模糊数据。这些异常数据信息严重影响原始数据的质量，不同类型的异常数据也给大数据的分析和挖掘带来了巨大的挑战。重复数据是由于设备或其他因素问题导致的数据重复存储造成的，无序数据是网络传输延迟造成的不正确的数据存储顺序。其他异常数据大多是设备定位问题引起的，其中设备定位问题引起的噪声数据是最主要的部分。检测无效数据和缺失数据很简单，无效数据由数据的纬度或经度是否超出范围来确定。缺失数据是通过比较两个数据之间的时间差和数据收集的频率来确定的。对于漂移数据，本文提出了基于范围约束和滑动窗口统计方法进行检测和修复。在原始数据的基础上，本文提出了一种利用行驶方向的算法，并通过分析数据的变化进一步提高了准确性。范围约束的方法不仅可以检测漂移较大的异常数据点，还可以通过统计方法实现了漂移较小异常数据的检测。重复数据检测的效率取决于原始数据的乱序程度，通过基于原始数据中存在的约束条件提出了有效的排序方法，避免了对整个数据集进行排序所产生的时间消耗。

3.2 相关定义

本节主要对移动对象数据异常处理中遇到的一些概念和定义进行介绍，并结合异常处理的流程对各个部分进行说明。

假设 $P=\{p_1, p_2, \mathrm{K}, p_n\}$ 为 GPS 记录的数据集，集合 P 中每一个 p 都由一组属性组成，其中主要的一些属性为 t（time），lon（longitude），lat（latitude），其余的一些属性信息为速度、方向、加速度等。异常清洗中的主要概念在下文进行介绍。

定义 3.1（最大无序偏移）设 P'为 GPS 数据集 P 对应时间 t 的一个有序集合，如果对于 $\forall p_i \in P$ 有 $\exists p_j^{'} \in P', p_i = p_j^{'} \wedge |j-i| \leq c$，则认为$c$ 为 GPS 数据集 P 的最大无序偏移。

最大无序表示原始数据集通过时间进行排序后获取一个新的数据集，排序前后的数据可以在最大固定的范围内找到对应的数据位置。假设 P 为原始 GPS 数据集的一个无序集合，数据集 P 的最大无序偏移为c，其中 $p_i \in P$，并且 P'是原始数据集 P 的一个有集合，其中 $p_i^{'} \in P'$，因此可以得到 $p_i \in \{p_{i-c}^{'}, \mathrm{K}, p_{i+c}^{'}\}$。例如：如图 3.1 所示，原始数据集合 P 的最大无序偏移为 2，原始 GPS 数据集经过时间排序后得到有序集合 P'，P 中 p_3 则被修复到了对应集合 $\{p_1^{'}, p_2^{'}, p_3^{'}, p_4^{'}, p_5^{'}\}$ 中的 $p_1^{'}$ 位置。

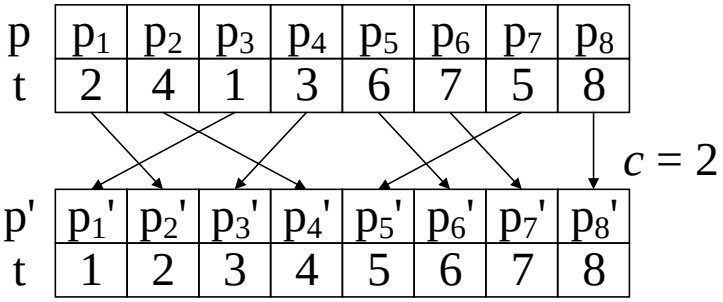


图 3.1 最大无序偏移（$c=2$）

定义 3.2(无序异常)设一个原始的数据集合 $P=\{p_1, p_2\mathrm{K}, p_n\}$，集合 P 中如果 $\exists p_i, p_j \in P$ 有 $i<j \wedge p_i.t > p_j.t$，那么集合 P 存在无序数据异常。

GPS 数据集合中的无序异常数据信息通常是由信号延迟和存储问题导致的，使得定位的数据信息延后了一段时间才得以保存。如图 3.2 所示，原始数据集合 P 中的 p_3，p_4，p_7 即为无序数据异常。

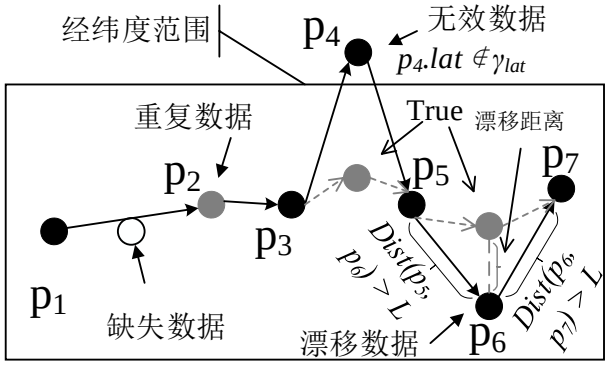


图 3.2 不同类型异常数据

表 3.1 符号表

符号	描述
P	GPS 轨迹数据集，包含多个 p
P'	数据集 P 修复之后的数据
p_i	数据集中第 i 个记录
v_i	p_i 近似的速度
lon, lat	经纬度
T_f	数据采集频率
L_e	设备采集数据的误差范围
$\gamma_{lon}, \gamma_{lat}$	设备采集数据的经纬度范围
$Dist(p_i, p_j)$	p_i 和 p_j 两个数据点之间的距离

定义 3.3（重复异常）设一个原始的数据集合 $P=\{p_1, p_2 \mathrm{K}, p_n\}$，集合 P 中如果 $\exists p_i, p_j \in P$ 有 $i \neq j \wedge p_i = p_j$，那么集合 P 存在重复数据异常。

定义 3.4（无效异常）给定一个原始 GPS 数据集合 P，其中 γ_{lon} 和 γ_{lat} 分别为集合 P 的经度和纬度数据范围，如果 $\exists p_i \in P$ 符合 $p_i.lon \notin \gamma_{lon} \vee p_i.lat \notin \gamma_{lat}$，则原始数据集合 P 中存在无效数据异常。

定义 3.5（缺失异常）设 T_f 为原始 GPS 数据集合 P 的采样频率，如果 $\exists p_i, p_j \in P$ 符合 $i < j \wedge p_j.t - p_i.t \geq T_f * (j - i + 1)$，则集合 P 中存在缺失数据异常。

定义 3.6（漂移异常）给定一个函数 $Dist(p_i, p_j)$ 用于计算 GPS 数据集合 P 中第 i 个元素 p_i 与第 j 个元素 p_j 之间的距离，并且 L 表示数据采样频率 T_f 时间内物体运行通过最大距离的阈值，在集合 P 中，如果对于 $\forall p_i \in P, p_i.lon \in \gamma_{lon} \wedge p_i.lat \in \gamma_{lat}$ 满足 $\exists p_k \in P$ 且 $Dist(p_{k-1}, p_k) > L \wedge Dist(p_k, p_{k+1}) > L$，则集合 P 中存在漂移数据异常。

如图 3.2 所示，图中对重复异常、无效异常、缺失异常、漂移异常进行了举例说明。缺失异常数据用空心点表示，重复的数据异常具有多个相同 GPS 数据记录，例如 p_2。无效异常的数据是指大于原始数据集的经纬度的最大范围，例如 p_4。漂移数据异常是指在正常经纬度范围内，但是和周围正常数据点的距离超过了设定的最大阈值，例如 p_6。表 3.1 给出了本章中常用的一些符号信息。

定义 3.7（数据范围）给定一个 GPS 数据集合 P，其中 p_i 为集合 P 中的数据点，p_i 点对应的速度为 v_i，在固定的数据采集频率 T_f 内，p_i 点以最大加速度运动通过的距离为 L，以 p_i 数据点为圆心，距离 L 为半径形成的区域则为 p_i 的数据范围。

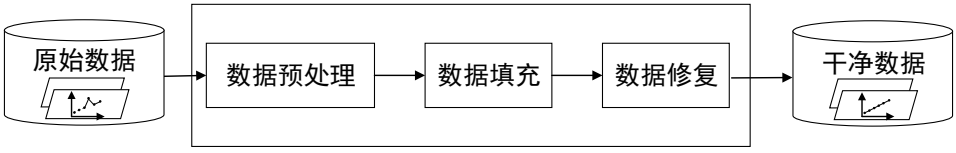


图 3.3 异常数据清洗流程

针对移动对象数据清洗，本文针对数据中可能存在的异常数据信息设计了异常检测和修复的处理流程如图 3.3 所示。流程中主要由三个部分组成：(i)数据预处理；(ii)数据填充；(iii)

数据修复。数据预处理执行数据排序并对重复数据进行去重工作。数据填充是对缺失的 GPS 数据根据有效的算法进行填充，数据修复是针对无效数据和漂移数据的异常检测和修复。漂移数据是指不符合行驶规则，明显偏离行驶轨迹的数据记录。为了遵循数据修改的最小原则，将偏离原始数据较大的无效数据作为漂移数据进行修复处理，这样不仅提高了处理效率，相比于将无效数据删除更加的保证了原始数据的特征。

3.3 数据预处理

3.3.1 数据依赖关系

移动对象数据集中 t_i 时间点的数据记录一定程度上依赖 t_{i-1} 时间点的数据信息情况。考虑到移动对象数据集中连续的数据之间存在依赖关系，则检测点周围数据情况的分析对于数据的异常检测和修复具有非常重要的参考依据。距离检测点时间较远的数据的参考价值较小，因此根据检测点的位置，设定适当大小的滑动窗口则可以很好的解决数据之间的依赖关系。为了实现异常数据清洗的高效性和准确性，文中设计了一种基于滑动窗口的统计方法。为了支持对不同类型移动对象数据集的自动异常清洗，在对数据集进行分析之前通过获取不同数据集中共有的属性列保存进 SECONDO[76]数据库中，从而进行后续的数据清洗工作。

3.3.2 数据排序和去重

由于定位设备信号延迟以及存储技术的影响，原始移动对象数据集中数据存储的时间可能会晚于实际的定位时间，这将导致移动对象数据记录出现乱序的结果。文中利用原始数据的最大无序偏移设计了一种最大偏移堆排序 MOHSort（Maximum Offset HeapSort）算法以提高对无序数据处理的效率。MOHSort 算法是对窗口中的数据进行排序，窗口的大小即为排序算法中堆的大小。该算法的思想是通过堆结构的特性将设备采集到的数据预先存储到堆结构中，等待延迟数据的到达并对其进行排序。在对数据排序之前，首先通过外部设备确定数据延迟存储的最大时间，并根据设备采集数据的频率计算延迟的时间内所能采集的数据点的个数。然后，将堆的大小初始化为延迟时间内存储数据的个数大小，并预先在堆中填入对应的数据个数。在对数据进行排序的过程中，将堆顶点的数据时间与排序后的数据时间进行比较。如果比较的数据不在连续的采集频率内，则将原始的堆大小扩大到当前大小的两倍。添加扩展后的新数据后，再次将堆顶的数据与排序后的数据进行比较。如果两个比较的数据时间连续，则将堆顶的数据取出，如果时间不连续，则将堆的大小继续扩大到当前大小的两倍。最大乱序排序算法如算法 3.1 所示。

假设定位设备存储原始数据集的最大延迟时间为 60 秒，数据的采集频率为 2 秒。因此，在数据最大延迟的期间内的数据记录个数为 $60/2=30$，所以数据集的最大无序偏移为 30。为了较好的实现堆结构的比较，将堆大小默认设置为 31，以数据时间作为比较参数构建小顶堆。首先使用前 31 个数据点构建一个堆如图 3.4（a）所示。堆中数字的大小代表数据记

算法 3.1 最大偏移堆排序 MOHSort

输入：原始轨迹轨迹集 *P*

数据集最大的无序偏移 *c*

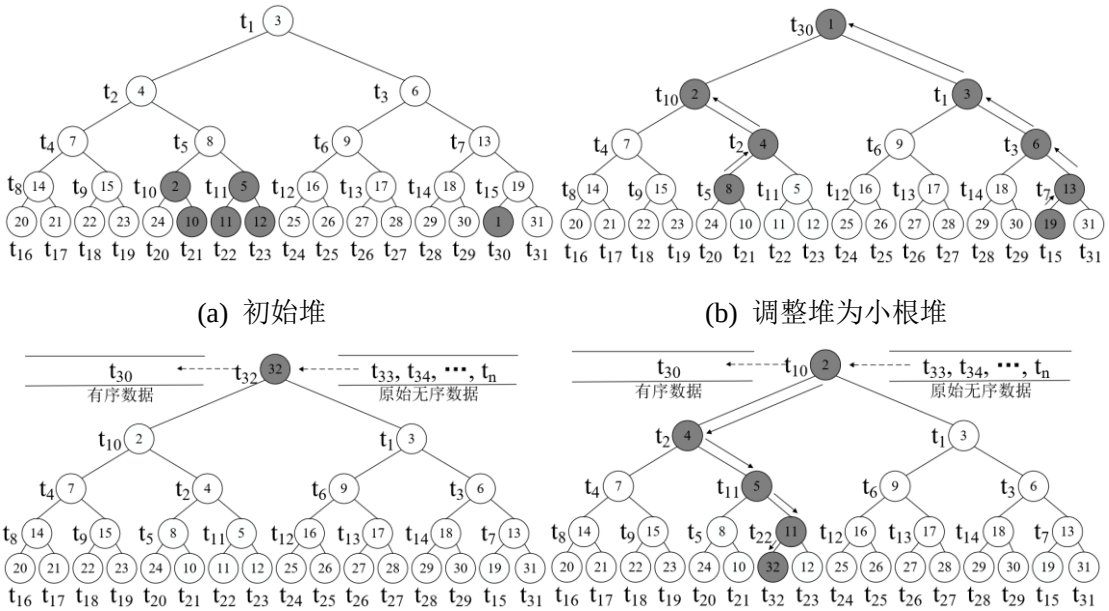
输出：干净的数据集 *P′*

1. *H* ← 通过数据集 *P* 的前 *c* 个记录构建小根堆
2. *N* ← |*P*|
3. **For** *i* ← *c* + 1 **to** *N* **do**
4. **If** *H.peek()* 不是下一个连续的数据 **then**
5. *c* ← *c* * 2
6. **While** *i* < *N* ∧ *H.size* < *c* **do**
7. *H.push*(p_i)
8. ++*i*
9. **Else**
10. *P′* ← *H.pop()*
11. *H.push*(p_i)
12. **While** *H* ≠ *null* **do**
13. *P′* ← *H.pop()*
14. **Return** *P′*

录的时间大小，灰色的节点表示数据集中延迟存储造成的无序数据。小顶堆以时间的大小进行排序，如图 3.4（b）所示灰色的节点为堆结构按照时间比较进行调整的节点。由于原始数据集的最大无序偏移为 30，因此即使一条数据记录到达无序偏移的最大偏移量如节点 t_{30}，在对堆进行排序后仍然可以获得正确的时间顺序。所以，小顶堆中的顶部元素即为排序的第一个数据记录。将堆的顶部数据放入有序队列中，并将未排序原始数据集中的下一个数据记录放入堆中，如图 3.4（c）所示。然后，将新插入的数据进行堆的调整并将堆顶的数据记录取出按照时间顺序存储在有序队列中，如图 3.4（d）所示。重复上述步骤，从小顶堆中获取顶部元素，并将未排序的数据集中的下一个数据放入堆中，直到处理完所有的数据。

定理 3.1 给定一个数据集合，其中包含 *N* 个数据记录，*c* 为数据集合的最大无序偏移常数，那么 MOHSort 算法的时间复杂度为 $\mathrm{O}(N \cdot \lg c)$。

证明：由于数据集合的最大无序偏移是常数 *c*，因此对于无序的数据记录都可以在最大偏移量 *c* 的范围内找到正确的数据时间顺序。对于每次堆中新插入的节点，维护堆所需的时间为 $\mathrm{O}(\log_2 c)$，其中 *c* 为堆中的节点个数。堆调整操作的次数是所有数据点的数量 *N*，所以 MOHSort 算法的时间复杂度为 $\mathrm{O}(N \cdot \lg c)$。



(a) 初始堆　　(b) 调整堆为小根堆

(c) 取出堆顶数据，插入下一个数据　　(d) 调整堆结构

图 3.4 基于堆排序的最大无序偏移

对于有序的移动对象数据集，删除其中的重复数据记录则变的相对简单。算法可以通过检测当前数据记录是否与下一条记录相同来判断重复的数据记录。如果记录数据相同，则删除检测点的数据。如果两个数据记录不相同，则继续检测数据集中存在的其他类型的数据异常，所以去重算法的时间复杂度为$O(N)$。

3.4 缺失数据填充

GPS 定位设备可能由于信号影响或意外关闭，使得设备采集的移动对象数据在此期间存在数据缺失的异常现象。例如，当车辆行驶在信号较弱的偏远地区或者驾驶员关闭了汽车的引擎时，设备此时则会有可能采集不到任何数据。如果车辆停放了很长时间，通过算法识别很可能将停车时间未采集的数据识别为大量的缺失数据。为了避免算法错误缺失数据的检测，文中对两个连续数据记录之间的时间差大于 10 分钟的数据进行分段处理。通过分段处理方法不仅避免了将大量数据加载到内存，而且可以正确处理错误识别的缺失数据。

由于连续的 100 条原始数据记录可以充分反映数据集采集的频率，因此在对数据集进行数据缺失检测之前，首先需要统计数据集中前 100 条数据记录的时间差。通常 GPS 定位设备采集数据的时候会以固定的频率，所以文中取数据集的前 100 个数据中连续数据时间差占比概率最大的时间差作为当前数据集的采集频率。对于以时间排好序的数据集，计算连续数据记录之间的时间差，并将时间差与设备的采集的频率进行比较以确定数据集中的缺失数据。对于两个相邻的连续数据之间的时间差大于设备频率的两倍或更多倍时，将这两个数据记为缺失数据的最左端和最右端。

$$num = \frac{t_r - t_l}{T_f} \tag{3.1}$$

$$L = v_l \cdot T_f + \frac{\frac{v_r - v_l}{t_r - t_l} \cdot T_f^{\,2}}{2} \tag{3.2}$$

$$\begin{cases} x = x_l + (y - y_l)(x_r - x_l)/(y_r - y_l) \\ (x - x_l)^2 + (y - y_l)^2 = L^2 \end{cases} \tag{3.3}$$

$$\begin{cases} (x_l + u(x_r - x_l)/(y_r - y_l), y_l + u), y_r > y_l \\ (x_l - u(x_r - x_l)/(y_r - y_l), y_l - u), y_r < y_l \\ (x_l + L, y_l), y_r = y_l \wedge x_l \le x_r \\ (x_l - L, y_l), y_r = y_l \wedge x_l > x_r \end{cases}, u = \left(\frac{L}{\sqrt{1 + \left[(x_r - x_l)/(y_r - y_l)\right]^2}}\right) \tag{3.4}$$

将数据集中相邻两个连续数据记录的时间差大于 10 分钟的进行分段，这样可以有效避免了由于发动机停机导致数据被错误识别为大量缺失数据的情况，而且 GPS 定位设备的采集频率通常为秒级，对于缺失 10 分钟的数据即使对其进行数据填充也不会起到很好的修复效果，反而会填充更多的异常数据造成数据质量低下。对于适当的数据缺失，为了达到良好的数据填充效果，则需要通过计算缺失数据周围正常数据的变化来对数据进行填充。文中的方法利用了正常数据的变化和缺失数据的持续时间来解决填充问题。这是通过基于近邻数据的填充方法完成的。为了预测缺失数据这段时间内的数据点情况，本文将这段数据大致设置为匀加速进行分析。根据上文统计的方法得到的原始数据的收集频率，首先通过公式 $num = \frac{t_r - t_l}{T_f}$（3.1 计算缺失数据的个数，其中 t_l 为缺失数据最左边的时间，t_r 是缺失数据最右边的时间，T_f 是数据的采样频率。然后，通过对缺失数据前一段时间正确数据的加权平均法，得到缺失数据前速度的近似值。以加权平均的方法同样可以得到缺失数据后一段时间正确数据的速度近似值。根据缺失数据相邻数据的速度变化情况，则可以形象地模拟车辆的运行状态。以车辆进行匀加速的情况下，可以根据速度变化值和数据缺失持续时间计算出车辆在数据缺失期间内的加速度。车辆以固定频率时间内的可以通过的距离可由公式 $L = v_l \cdot T_f + \frac{\frac{v_r - v_l}{t_r - t_l} \cdot T_f^{\,2}}{2}$（3.2 计算得出，其中 v_l 为缺失数据前一段时间的近似速度，v_r 为缺失数据后一段时间的近似速度。文中通过高斯-克吕格 GK 投影将经纬度数据信息映射到二维平面上，(x_i, y_i) 是经纬度映射对应的坐标。公式 $\begin{cases} x = x_l + (y - y_l)(x_r - x_l)/(y_r - y_l) \\ (x - x_l)^2 + (y - y_l)^2 = L^2 \end{cases}$（3.3 中的联立方程是由缺失数据的两个相邻边上的正确数据点的连线和以左边相邻的正确数据点为圆心，以在一个频率时间内通过的距离为半径的圆组合得到的。通过公式 $\begin{cases} x = x_l + (y - y_l)(x_r - x_l)/(y_r - y_l) \\ (x - x_l)^2 + (y - y_l)^2 = L^2 \end{cases}$（3.3）的联立方程，可以得到公式 $\begin{cases} (x_l + u(x_r - x_l)/(y_r - y_l), y_l + u), y_r > y_l \\ (x_l - u(x_r - x_l)/(y_r - y_l), y_l - u), y_r < y_l \\ (x_l + L, y_l), y_r = y_l \wedge x_l \le x_r \\ (x_l - L, y_l), y_r = y_l \wedge x_l > x_r \end{cases}, u = \left(\frac{L}{\sqrt{1 + \left[(x_r - x_l)/(y_r - y_l)\right]^2}}\right)$（3.4 中对应

缺失数据的坐标位置。近邻数据填充 NNF（Nearest Neighbor Fill）算法在算法 3.2 中给出。NNF 算法首先计算数据集中缺失数据点的个数，然后通过遍历缺失数据来计算当前位置与前一个正确数据点的距离，最后计算出当前检测位置的缺失数据坐标并对其进行填充数据。

算法 3.2 近邻数据填充 NNF

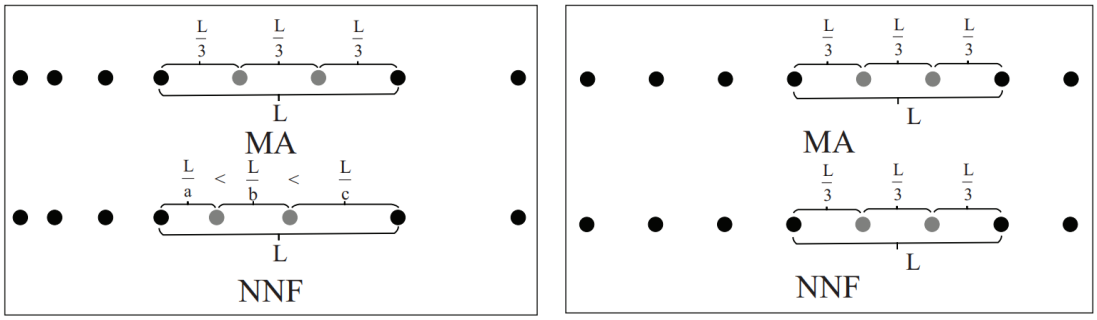
输入：轨迹数据集 P

左边界 l

右边界 r

输出：干净的数据 P'

1. T_f ← 获取数据集 P 的频率
2. $num \leftarrow (p_r.t - p_l.t) / T_f$
3. $v_l, v_r \leftarrow$ 计算缺失数据左边两边的近似速度
4. $L_f \leftarrow$ 根据公式 3.2 计算在频率 T_f 内通过的距离
5. $L \leftarrow L_f$
6. **While** num != 0 **do**
7. $p'.t \leftarrow$ 计算缺失数据时间
8. $(p'.x, p'.y) \leftarrow$ 根据公式 3.4 计算填充数据的坐标
9. ——num
10. $L \leftarrow$ 计算下一个频率内通过的距离
11. $P' \leftarrow P.add(p')$
12. **Return** P'



(a) 匀变速运动　　(b) 匀速运动

图 3.5 不同方法修复缺失数据比较

如图 3.5（a）显示了在加速条件下填充缺失数据的不同方法的结果。基于平滑处理的移动平均（MA）模型[17]，根据缺失点的数量平均划分正常数据之间的距离。MA 方法没有利用相邻正确数据的速度变化情况，以及不同行驶条件下不同距离的变化因素。因此，在加速行驶的情况下，使用 MA 分配的数据填充距离远大于缺失数据开始之前 NNF 分配的距离。同时，分配相同的行驶距离也违反了车辆行驶的规律，即在直线行驶时车辆行驶速度高通过的距离应比车速低通过的距离要长。与加速情况相反的是减速情况，均匀分配距离的方法同

样也违反了驾驶规律。图 3.5（b）显示了以恒定速度填充缺失数据的不同方法的结果。如果缺失数据的相邻速度没有变化，则认为当前缺失数据是当汽车以近似恒定的速度行驶时造成的。MA 平滑填充的方法可以很好地反映车辆行驶的数据记录。由于汽车匀速行驶，加速度为$0m/s^2$，这时 NNF 算法填充的数据也趋于平均，填充数据的结果与 MA 方法类似。NNF 算法不仅具有 MA 算法均等填充数据的优势，而且对于逐渐演化的数据也有很好的填充优势，在变化的数据中，NNF 比 MA 有更好的填充效果。

3.5 漂移数据清洗

为了检测和修复漂移数据，本文设计了一种结合范围约束和最大似然估计的算法。范围约束修复漂移较大的数据异常，避免了使用基于统计方法修复较大漂移异常而违反数据修改的最小原则。通过最大似然估计方法修复漂移较小的异常，也解决了基于约束方法检测不到的这类异常的问题。由于车辆不能一直加速行驶，所以使用当前速度来设置加速度的变化情况。欧洲共同体和联合国欧洲经济委员会的制动标准见表 3.2。

表 3.2 制动标准

车辆类型	M_1	M_2 M_3	N_1 N_2 N_3
制动减速（m/s²）	5.8	5	434
制动时间（s）	0.36	0.54	0.54
制动距离（m）	≤0.1V+V²/150	≤0.15V+V²/135	≤0.15V + V²/115

M_1 指不超过八座的客车；M_2 指八座以上，总质量不大于五吨的客车；M_3 指八座以上，总质量在五吨以上的客车。N_1 指总质量不大于三点五吨的货车或拖拉机；N_2 指的是总质量的范围在从三点五吨至十二吨之间的货车；N_3 指的是总质量大于十二吨的卡车。本文的移动对象数据集主要由 M_1 和 N_2 两种车辆收集。出租车的速度变化率一般不超过$3.5m/s^2$，运输车辆的速度变化率一般不超过 $2.5m/s^2$。对于两种不同数据集中数据速度变化率的参考阈值分别设置为$[-5.8m/s^2, 3.5m/s^2]$和$[-4.4m/s^2, 2.5m/s^2]$。

表 3.3 不同设备采集的原始数据样例

	编号	时间	经度	纬度	…	方向	速度
No.1	1	2008-02-05 22:05:42	116.69155	39.85164			
	1	2008-02-05 22:15:41	116.69155	39.8518			
	1	…	…	…			
	1	2008-02-05 23:55:41	116.69159	39.85182			
No.2	AA00001	2018/8/4 1:23:54	115.944571	28.65114	…	114	11
	AA00001	2018/8/4 1:23:55	115.944603	28.651126	…	110	12
	AA00001	…	…	…	…	…	…

	AA00001	2018/8/4 1:25:02	115.950913	28.648358	…	117	42

为了检测和修复数据集中的漂移异常，数据的方向属性是算法分析的重要因素。不同的数据采集源使得在原始数据的存储上对应的属性信息存在一些细微的差异。如表 3.3 所示，分别展示了 1 号 GPS 定位设备和 2 号 GPS 定位设备采集的原始数据信息。1 号设备采集的数据仅包含编号、时间、经度、纬度。除了上述四个属性外，2 号装备采集到的数据集中还有方向和速度等附加属性信息。为了更充分利用原始数据中包含的数据属性信息，如果数据集的属性包含方向，则对方向属性进行分析以进一步改进算法。

3.5.1 范围约束和统计

对于数据集中没有方向属性的异常修复，文中通过基于滑动窗口的限制数据变化范围和最大似然估计相结合的方法来对漂移数据进行异常修复。基于约束的修复方法 SCREEN[6] 通过设置数据的最大变化约束范围（s_{min}，s_{max}）来对异常数据进行检测，但是车辆的行驶速度在正常情况下不能总是一直增加的，其中 s_{min} 是数据最大减少变化的范围阈值，s_{max} 是数据最大增加的变化范围阈值。基于约束的异常检测方法不能很好地检测数据集中漂移较小的异常更不用说对这类异常进行修复。如果使用最大似然估计方法计算整个数据集的加速度会使算法运行缓慢，因此对部分数据进行计算使用局部滑动窗口的方法可以很好的提高算法的效率。为了减少对原始数据的修改提高异常修复的准确率，文中设计了一种范围约束的方法来修复漂移较大的数据异常。对于漂移较小的数据异常，设计了一种基于滑动窗口统计的方法来检测和修复数据。范围约束和滑动窗口统计 RCSWS（Range Constraints and Sliding Window Statistics）算法在算法 3.3 中给出。

基于速度约束，范围约束可以很好地实现对检测到的异常进行修复。数据范围是指数据采集的一个频率内当前数据点通过的最大距离所形成的圆形区域。检测一个数据点时，不仅需要前一个正确数据点的范围，还需要下一个正确数据点的范围。为了反映数据之间的速度变化，假设后一个数据点的速度大于前一个数据点的速度。因此，检测点的前一个数据点的范围半径小于后一个数据点的范围半径。同时，当前的检测点应在上一个正确数据点的范围和下一个正确数据点的范围交界内。对于不在前后正确数据范围交界内的数据，将检测的数据标记为漂移的异常数据并按照数据修改的最小原则对异常数据进行修复。通过范围约束对检测到的漂移异常数据进行修复时情况分为以下三种：

情况 1：检测的数据点不在周围正确数据的范围内。在这种异常数据情况下，数据点与原始数据点相差较大。为了满足数据修改的最小原则，文中通过将异常数据点设置为最接近正常数据范围上的点来对异常数据进行修复。以两个数据点为圆心，则范围的半径要取决于两个数据自身速度在采样频率内能够经过的最大距离。如果两个数据范围之间没有交集，如图 3.6（a）所示，则对数据集中下一个数据点进行检测判断是否违反了范围约束。如果仍然违反范围约束，则继续对下一个数据点进行检测直到数据满足约束条件为止。将检测点 p_i

与前一个正确数据点 p_{i-1} 进行连接，连接与 p_{i-1} 的数据点的范围相交于 p_1^r。p_2^r 和 p_3^r 是满足范围约束条件的两个数据点范围的交点。对于异常点 p_i 则可以使用公式（3.5）进行计算修复。如果 p_1^r 在两个圆相交的范围内，则用 p_1^r 作为 p_i 的异常修复点。在其他情况下，使用 p_2^r 和 p_3^r 中最接近 p_i 的数据点进行修复。剩余的异常点根据新修复的点进行检测和修复。如果以

$$p_i' = \begin{cases} p_1^r, p_1^r \in (c\ p_{i-1} \cap c\ p_i) \\ p_2^r, Dist(p_i, p_2^r) \le Dist(p_i, p_3^r) \\ p_3^r, Dist(p_i, p_2^r) > Dist(p_i, p_3^r) \end{cases} \quad (3.5)$$

算法 3.3 范围约束和滑动窗口统计 RCSWS

输入：原始轨迹数据集 P

检测位置 i

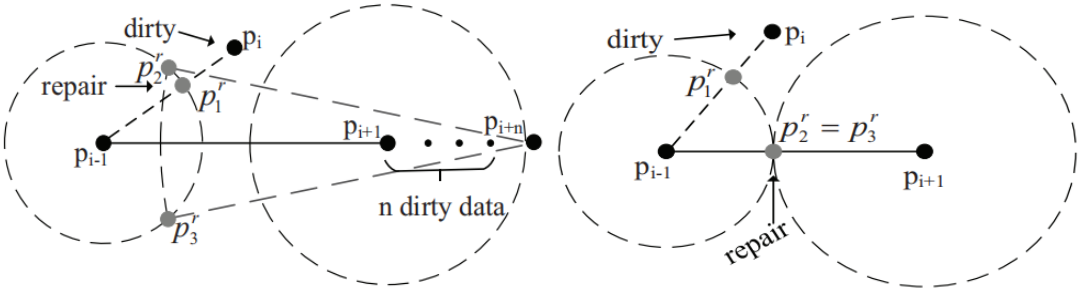
数据集频率 T_f

概率阈值 σ

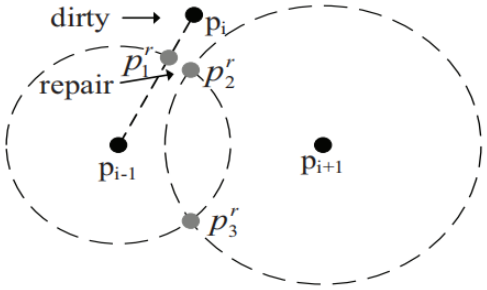
输出：干净数据集 P'

1. $count \leftarrow 0$
2. $a \leftarrow$ 计算 p_{i-1} 记录的加速度
3. $t \leftarrow T_f$
4. $dis_{max} \leftarrow$ 计算 p_{i-1} 在时间 t 内以加速度 a 通过的最大距离
5. $dis \leftarrow Dist(p_{i-1}, p_i)$
6. **While** $dis \ge dis_{max}$ **do**
7. $count$++
8. $t \leftarrow$ 计算 p_{i-1} 和 $p_{i+count}$ 之间的时间差
9. $dis_{max} \leftarrow$计算 p_{i-1} 在时间 t 内以加速度 a 通过的最大距离
10. $dis \leftarrow Dist(p_{i-1}, p_{i+count})$
11. **If** $count > 1$ **then**
12. $P' \leftarrow$ 将 p_{i-1} 到 $p_{i+count}$ 的数据变化均匀的划分以修复连续异常
13. **Return** $\leftarrow P'$
14. **If** $count = 1$ **then**
15. $P' \leftarrow$通过范围约束修复 p_i
16. **Return** P'
17. **Else**
18. $O \leftarrow$统计滑动窗口中数据加速度的变化
19. $O(i) \leftarrow$ 在 p_i 记录处速度变化的概率
20. **If** $O(i) < \sigma$ **then**
21. $P' \leftarrow$使用速度变化情况概率大于 σ 的数值来修复 p_i
22. **Return** P'
23. **Return** P

两个数据点为圆心的范围相交且存在一个交点，如图 3.6（b）所示，相交点就是异常数据点 p_i 的修复点，其中 $p_2^r = p_3^r$。如果以两个数据点为圆心的两个范围有两个相交点，如图 3.6（c）所示，根据检测异常数据点的位置用 p_2^r 和 p_3^r 中离 p_i 最近的一个数据点进行异常修复。



(a) 无交点　　(b) 一个交点



(c) 两个交点

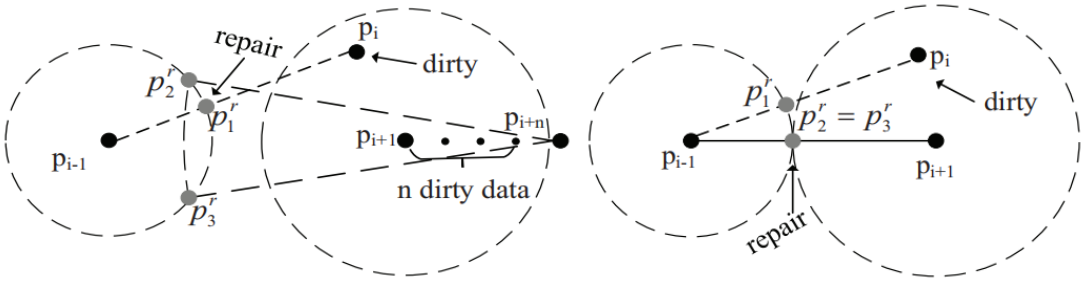
图 3.6 违反周围数据范围约束异常情况

$$p_i' = \begin{cases} p_1^r & , p_1^r \in (\mathrm{e}\ p_{i-1} \wedge \mathrm{e}\ p_{i+1}) \\ p_2^r & , Dist(p_i, p_2^r) \le Dist(p_i, p_3^r) \\ p_3^r & , Dist(p_i, p_2^r) > Dist(p_i, p_3^r) \end{cases} \quad (3.5)$$

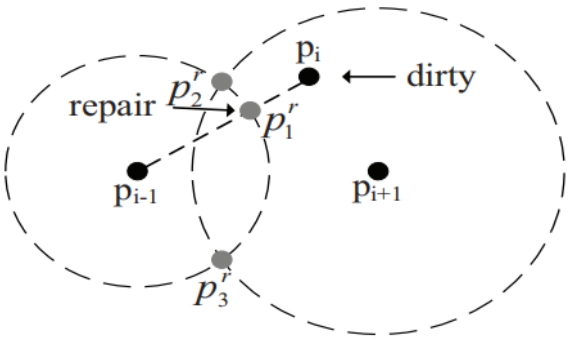
情况 2：对于检测到的数据点仅在前后一个数据的正常范围内的情况，因为前一个数据点已经被检测或者修复为正常的数据，所以检测点是违反了前一个数据的范围约束。这种情况即为当前检测点 p_i 超过了上一个数据点 p_{i-1} 的约束范围，但是在下一个数据点 p_{i+1} 的约束范围内，如图 3.7 所示。如果 p_{i-1} 和 p_{i+1} 也违反约束，则继续判断后续数据，直到两个数据的范围相交。对于情况 1 中对于异常数据获取三个异常的候选修复点，并根据公式

$$p_i' = \begin{cases} p_1^r & , p_1^r \in (\mathrm{e}\ p_{i-1} \wedge \mathrm{e}\ p_{i+1}) \\ p_2^r & , Dist(p_i, p_2^r) \le Dist(p_i, p_3^r) \\ p_3^r & , Dist(p_i, p_2^r) > Dist(p_i, p_3^r) \end{cases}$$

（3.5 修复 p_i，如图 3.7 所示。



(a) 无交点　　(b) 一个交点



(c) 两个交点

图 3.7 违反前一个正常数据范围约束异常情况

情况 3：对于检测的数据点 p_i 在前后两个数据的范围内的情况。由于满足范围约束条件如果当前检测点具有较小的漂移异常，则使用基于约束的方法将无法对 p_i 点进行检测和修复异常数据。为了准确检测和修复满足范围约束且具有较小漂移的数据异常，采用基于滑动窗口的最大似然估计可以很好的统计正确数据的变化的概率修复，修复统计中变化概率较小的异常数据。与整体数据统计相比，基于滑动窗口的统计不仅减少了统计的数据量，而且避免了数据之间间隔长以及较多错误数据的干扰问题。由于窗口中的前半部分数据已经修复，统计结果则可以正确的反映窗口中数据的变化情况。如果当前数据变化在统计结果中所占的概率较小，将当前数据视为与原始运动的轨迹具有较小的偏差并标记为较小的漂移异常点。由于对窗口中的前一半数据原始正确或已经被检测和修复，基于数据最小修改原则将使用统计结果中占比 50%的数据变化情况的数值来修复异常。

对于检测数据集中漂移数据较小的异常，现有的统计方法会统计整个数据集中数据属性变化的情况使得整体算法运行缓慢。如果原始数据集中原本异常数据就比较多，则统计出来的结果同样不能反映真实的数据现象，较多的异常数据占比较大反而会将概率小的正确数据进行错误的修复造成较差的异常修复效果。

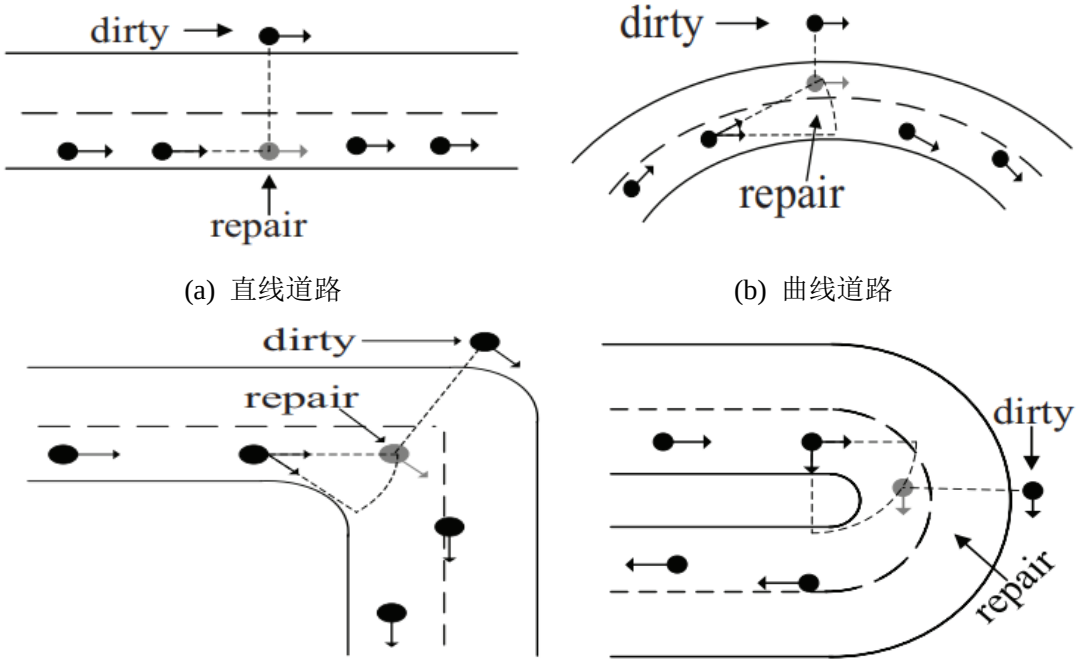
3.5.2 方向分析

对于原始数据集中带有方向属性的数据，可以通过利用前后数据速度方向的变化情况更精确的对异常数据进行异常检测和修复。行驶方向属性在数据的异常检测修复中的分析主要分为以下三个情况进行介绍：

情况 1：在滑动窗口中，检测点前后的连续一段时间内数据对应的车辆行驶方向几乎相似。这种行车轨迹数据通常是在车辆沿直线道路行驶时采集的，如图 3.8（a）所示。在车辆沿着直线道路行驶的情况下，行驶方向几乎相同。当前检测点的行驶方向与前一个正确的数据点合并时，两个数据的行驶方向一致。根据检测点前一个正确数据点的近似速度和加速度，可以得到前一个数据点在两点的时间差内通过的距离。前一个正确数据点行驶距离的终点即可以作为直线行驶时漂移数据的修复点。

情况 2：在滑动窗口中，检测点前后的连续一段时间内数据对应的车辆行驶方向存在一定偏差。这种类型的行车轨迹数据通常是由相对较小的曲线道路引起的，如图 3.8（b）和图 3.8（c）所示。将当前检测点的行驶方向与前一个数据点进行合并，两个行驶方向之间会有一定的偏差。通过计算前一段数据的近似速度和加速度来计算两点时间差内两个不同方向车辆可以通过的距离。由于两点的方向为最终的速度方向且车辆在行驶的过程中行驶方向是由前一段数据方向过度到后一段数据的方向，因此车辆在这段时间内的平均行驶方向在这两个数据点的方向之间，则这两点不同方向所形成的扇形区域则为当前检测点所在的候选修复区域。为了遵循数据最小变化的原则[25]，文中将扇形区域中离异常点最近的点作为异常数据的修复点。如果车辆在弯道上的行驶速度较为缓慢且短时间内速度变化的方向较小，则这种情况将转变成情况 1 对其进行分析。

情况 3：在滑动窗口中，检测点前后的连续一段时间内数据对应的车辆行驶方向有接近 180 度的偏转。对于这种类型的行车轨迹数据，通常是车辆在 U 型道路上行驶收集的数据，如图 3.8（d）所示。在这种道路情况下，将当前检测点的方向与前一个数据点的方向进行合并。根据两点方向的行驶距离形成的扇形区域，选择距离当前检测异常点最近的数据点作为异常的修复点。在车辆行驶较慢的情况下即使行驶在弯曲幅度较大的道路上，由于速度较小方向变化的也较为缓慢根据具体的观测数据则可以转换为情况 1 或 2 对数据进行分析。



(a) 直线道路　(b) 曲线道路

(c) 直角道路　(d) U 型道路

图 3.8 不同行驶方向对应的异常修复

在对数据集中存有方向属性的异常数据修复中，文中分析了不同道路场景下修复点可能存在的数据范围。在不同的道路场景下，车辆行驶距离和两个方向的夹角所形成的扇形区域总能包含正确的数据点。与整个圆形区域范围相比，扇形区域可选择的数据修复点的范围更小。在异常数据修复中，通过方向分析利用形成的修复扇形候选区域既可以满足数据修改的

最小原则，又可以缩小异常数据修复的范围，从而可以提高异常数据修复的准确性。带方向的范围约束和滑动窗口统计 RCSWSD（Range Constraints and Sliding Window Statistics with Direction）算法在算法 3.4 中介绍。

算法 3.4 方向范围约束和滑动窗口统计 RCSWSD

输入：轨迹数据集 P

检测位置 i

数据集频率 T_f

概率阈值 σ

输出：干净数据集 P'

1. $P' \leftarrow RCSWS(P, i, T_f, \sigma)$
2. $a_{i-1} \leftarrow$ 计算 p_{i-1} 的加速度
3. $v_{i-1} \leftarrow$ 计算 p_{i-1} 的速度
4. $L \leftarrow$ 计算 p_{i-1} 在频率内通过的距离
5. $\alpha \leftarrow$ 计算 p_{i-1} 到 p_i 两个记录行驶方向的角度
6. $S \leftarrow$ 形成角度 α 的候选扇形区域并构造等式
7. **If** $p_i \in S$ **then**
8. $\quad p_i' \leftarrow$ 计算扇形候选区域最接近 p_i 的点
9. $\quad P' \leftarrow p_i'$
10. **Return** P'

3.6 实验与性能评估

本节的实验在操作系统 Ubuntu 14.04（64 位，内核版本 4.8.2-19）的台式电脑（i7-4790CPU，3.6GHz，8GB 内存，1TB 硬盘）上运行评估。使用 C/C++语言进行编写，在可扩展的移动对象数据库系统 SECONDO[76]中进行开发。

3.6.1 数据集及实验配置

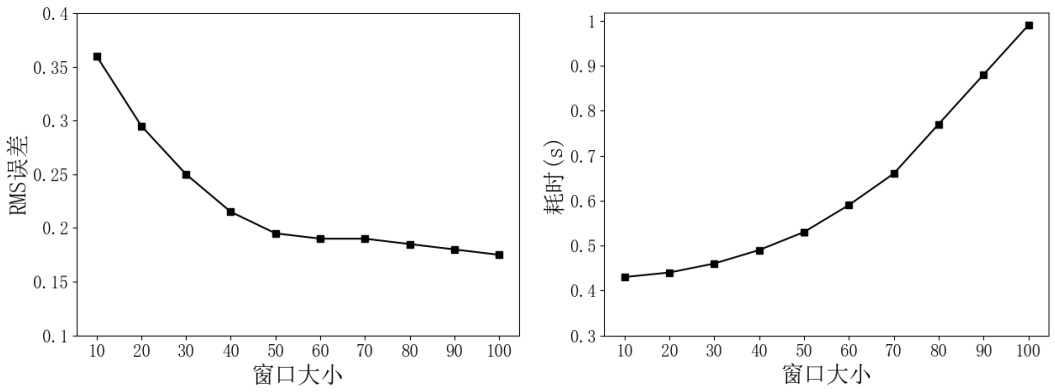
数据集。文中使用的数据集分别为从两种不同类型的车辆中收集的 GPS 数据。关于这两种数据集的一些信息在表 3.4 中介绍。一个是微软采集的出租车数据集[77][78]，包括 id、time、longitude、latitude。该数据集包含了 2008 年 2 月 2 日至 2 月 8 日的 10357 辆出租车的运动轨迹。数据记录数约为 1.5 千万，数据采样频率为 10 秒。另一个数据集是由中国交通运输部提供的运输车辆（卡车）的轨迹数据，其中包含了 13 个数据属性信息，例如 device_num, direction_angle, lng, lat, acc_state, location_time, gps_speed 和其他的一些属性信息。该数据集主要包含了 2018 年 7 月 30 日至 2018 年 10 月 26 日共 450 辆运输车辆的轨迹数据，数据记录约为 4.5 千万，数据采样频率为 1 秒。

表 3.4 数据集介绍

类型	属性数量	时间范围	经度范围	纬度范围	车辆数量	记录数量	城市	频率
出租车	4	[2008-02-02, 2008-02-08]	[115.117, 117.067]	[39.067, 41.1]	10357	1.5 千万	中国北京	10 s
卡车	13	[2018-07-30, 2018-10-26]	[109.418705, 121.99981]	[22.100031, 31.9953]	450	4.5 千万	中国南昌	1 s

评估标准。对于移动对象数据评估的标准文中采用了均方根误差 RMS（Root-Mean-Square error）[43]来评估数据修复的准确性。RMS 表示修复数据与真实原始数据之间的距离。RMS 的值越小，异常修复的数据越接近真实的轨迹数据，数据修复的结果也越准确。假设 x^{truth} 为干净的原始真实的轨迹 GPS 序列数据，x^{repair} 为对原始数据中的异常数据进行检测和修复后的 GPS 序列数据。RMS 的结果可以通过公式 $\Delta(x^{truth}, x^{repair}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{truth} - x_i^{repair})^2}$（3.6 计算得出。由于没有真实的干净的移动对象数据集，实验中使用的干净原始真实的轨迹数据集通过选择收集的原始轨迹数据中近似真实的数据记录作为参考。对于没有异常的原始数据集，通过手动去除正常数据并相应地添加一些噪声数据来对修复的结果进行比较计算。

$$\Delta(x^{truth}, x^{repair}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{truth} - x_i^{repair})^2} \quad (3.6)$$



(a) RMS　　(b)时间消耗

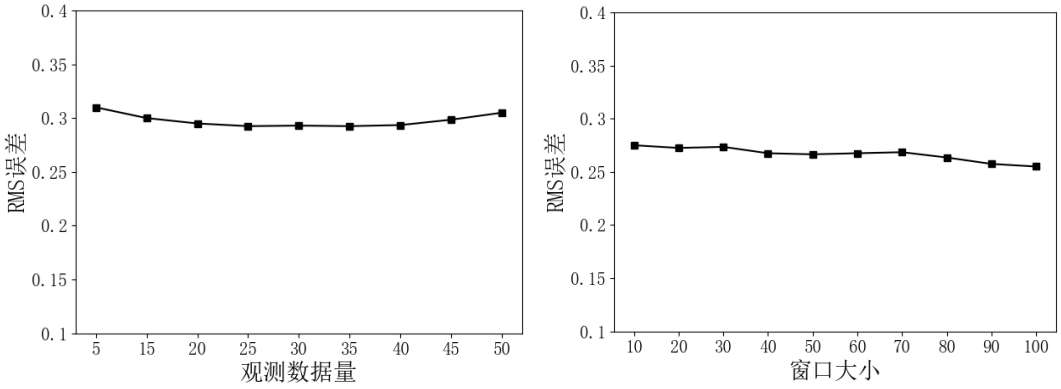
图 3.9 数据记录为 100k 时不同窗口大小的 RMS 和时间消耗

窗口大小。滑动窗口的大小对于原始数据中异常值较小的数据的检测和修复起到了相当重要的作用。为了获得合适的窗口大小，文中使用了近 10 万个真实原始移动对象数据集来测试 RCSWS 算法。通过实验控制不同的窗口大小，比较异常数据修复的 RMS 误差结果与相应的时间消耗以平衡数据清洗的精度和效率。随着窗口大小的增加，RMS 误差的值也逐渐减小并呈现出越来越缓慢的趋势，如图 3.9（a）所示。在算法的时间消耗方面，随着窗口大小的增加，相应的算法时间也越来越大，如图 3.9（b）所示。当窗口大小大于 50 时，RMS

误差值没有显示出明显的降低，反而偶尔呈现出增加的趋势。这种现象是由于窗口设置过大，可能导致统计结果不能反映检测点附近的数据状态，从而会对异常的修复产生误导性的影响。大量数据的统计结果也容易使得异常数据的修复趋于一致性，从而造成对数据修复不准确的现象。为了获取异常数据修复的准确性和效率之间的平衡，并在短时间内获得较高的准确率，文中将滑动窗口的大小设置为 50，并在后续的对比实验中使用该值。

3.6.2 实验结果分析

文中通过以无序的数据集且数据集中无序程度具有最大偏移量来对 MOHSort 算法和普通的堆排序算法进行比较并分析对应的时间消耗。对于异常的修复，文中将提出的算法 RCSWS 和 RCSWSD 与两种目前先进的方法进行比较包括：（1）EWMA[22]和（2）SCREEN[6]。EWMA 算法将每个观测值的加权系数设置为随时间呈指数下降的值。观测数据的时间越接近当前检测数据的时间，数据对应的数值加权系数越大。EWMA 根据最终的加权移动平均计算预测值并对当前检测的异常数据进行修复。SCREEN 使用前一段时间内的数据点的最大和最小变化约束来计算当前检测数据的大致范围并确定异常修复的候选点。候选点是通过检测点周围的数据对当前数据进行约束范围所形成的异常可能的修复点。然后，SCREEN 将候选点集中距离当前异常点最近的候选点作为异常的修复点并判断修复点是否在当前数据的约束范围内，进而对异常数据进行修复。



(a) EWMA 观测数据的范围　　(b) SCREEN 窗口大小

图 3.10 对比算法调整

为了在数据集上获得 EWMA 和 SCREEN 算法的最佳实验结果，文中通过使用真实数据进行实验并分别调整了两种方法对应的参数。在 EWMA 算法参数的调整中，对观测数据的范围选择进行了实验，如图 3.10（a）所示。图中可以看出，在对移动对象数据集的实验中 EWMA 算法对应的观测数据范围再 40 秒左右显示异常数据修复效果最佳。随着算法中观测数据量的增加，数据的异常修复将趋于平滑从而使得正确的数据可能也被当做异常数据被修改。对于 SCREEN 算法的设置，文中使用了其中的局部最优算法 *Local*。文中设置汽车的最大速度为$120km/h \approx 33.33m/s$，即 1 秒内行驶的最大距离为 33.33 米，对应的经度变化为 0.0003902，对应的纬度变化为 0.0002998，因此局部最优算法 *Local* 在经度上的约束为

$s_{lon}=(-0.0003902,0.0003902)$ 在纬度上的约束为 $s_{lat}=(-0.0002998,0.0002998)$。SCREEN 算法对应的窗口大小调整的实验结果如图 3.10（b）所示。随着算法窗口大小的增加，SCREEN 的局部最优修复结果也略有提高。随着窗口的增大也会导致 SCREEN 的局部最优算法演变成全局最优算法造成异常修复的时间消耗。为了获取一个较好的局部最优算法进行实验比较，文中将实验的窗口大小设置为 60 作为 SCREEN 之后实验对比的参数配置。

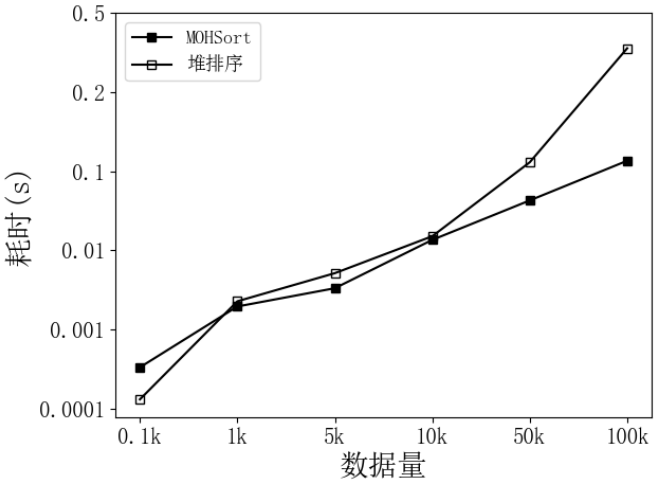
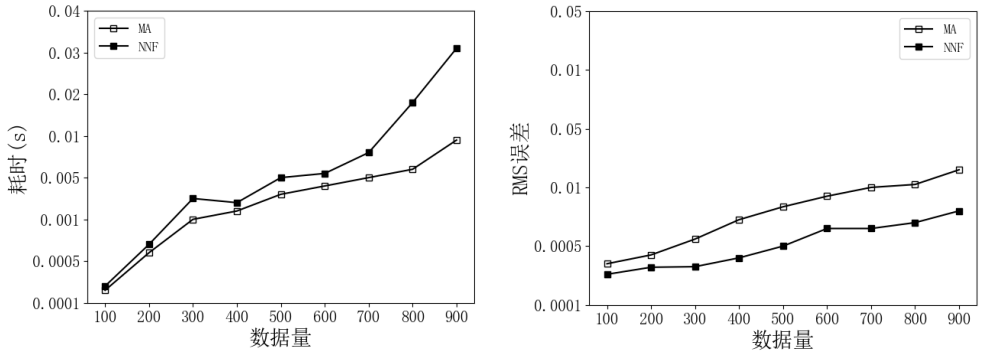


图 3.11　MOHSort 与普通堆排序时间比较

文中使用真实的移动对象数据集对具有最大无序偏移特性的 MOHSort 和普通堆排序算法进行实验对比。对原始数据集进行分析，由于数据采集的最大延迟为 1min，数据采集的频率为 1s，因此在 MOHSort 算法中的最大偏移量则设置为 60。MOHSort 和普通堆排序算法对比的排序时间实验结果如图 3.11 所示。当数据集大小为 100 时，由于数据量较小，原始数据基本保持真实轨迹中正确的时间顺序，而 MOHSort 算法的冗余操作计算造成了一定时间的浪费，使得普通堆排序的时间消耗略小于 MOHSort。随着数据集规模的增大，MOHSort 相对于普通堆排序的优势越来越明显。当数据集的大小达到 10 万条数据记录时，MOHSort 算法的效率要比普通堆排序算法快 3 倍。随着数据集的增加，原始数据中对应的无序数据的数据也随之变多。由于 MOHSort 在大规模数据中的数据调整堆的大小要比普通堆排序算法的开销小很多，所以在大规模的数据集中效率提升非常可观。

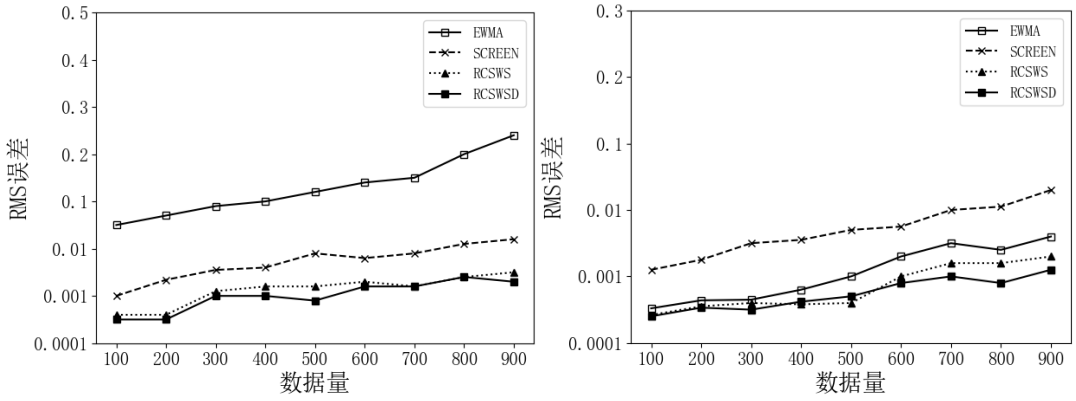


(a) 时间消耗　　(b) RMS

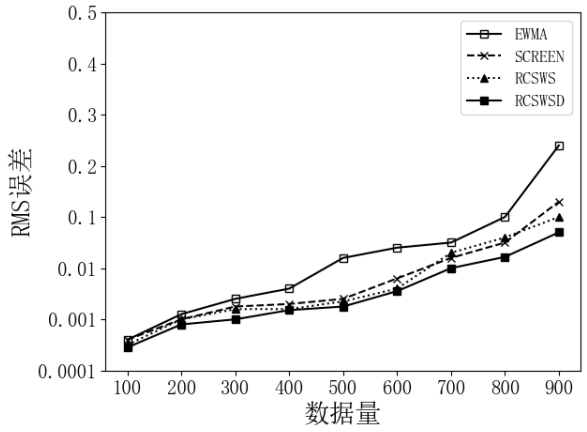
图 3.12 填充算法比较

文中从包含 1000 个数据记录的真实数据集中删除了一些正确的数据记录以进行数据填

充的比较。由于 MA 方法通过计算周围数据的平均值来填充缺失数据，因此在算法的效率上面要高于 NNF 算法，如图 3.12（a）所示。但是，在数据填充的准确性方面，NNF 利用相邻数据的变化情况，根据缺失数据周围速度的变化来计算缺失点的大致位置。与 MA 的填充方法相比，NNF 算法更符合车辆行驶的运行规律，不仅可以在匀速的时候具有 MA 的填充效果，对于变速的运动 NNF 通过计算周围数据的近似速度的变化来预测缺失数据点的阈值相比于 MA 的均分效果在填充上更能体现出车辆速度变化的状态，从而使得 NNF 在数据填充的精度上要优于 MA，如图 3.12（b）所示。



(a) 偏离较大异常数据集　　(b) 偏离较小异常数据集

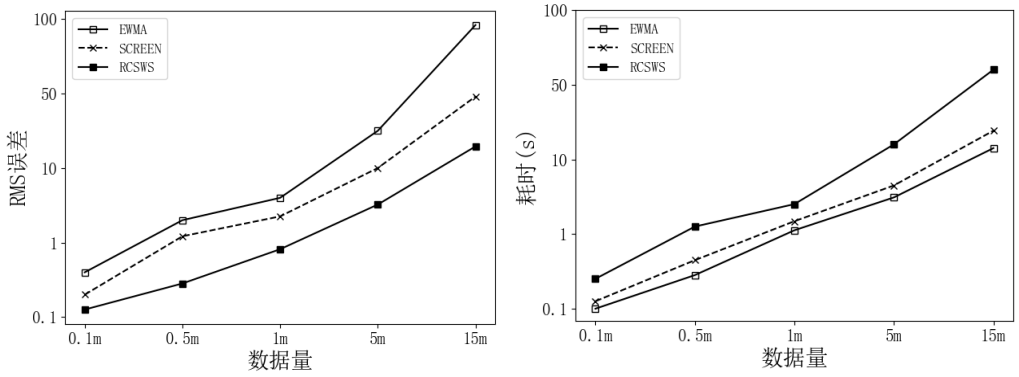


(c) 多种异常数据集

图 3.13 不同类型异常的影响

对不同类型异常的修复实验比较，文中通过手动添加三种不同类型的异常数据集来验证算法的修复效果。这三个数据集中分别包含大量漂移较大的异常数据、大量漂移较小的异常数据和各种类型的异常混合的数据。图 3.13（a）显示了具有大量漂移较大异常数据的实验结果。 根据 RMS 的数据值可以直观的判断出 RCSWS、RCSWSD、SCREEN 算法在修复漂移较大异常数据方面要优于 EWMA 算法。由于基于约束的方法难以对漂移较小的异常进行检测，因此文中的在 RCSWS 算法中添加的滑动窗口的统计检测方法可以很好的避免小异常不被检测的缺点。检测点前一段的数据趋于稳定，使得不同权重的乘积之和得到的预测值仍然可以用于修复漂移较小异常。所以 EWMA 在修复漂移较小异常的方面要优于 SCREEN 算

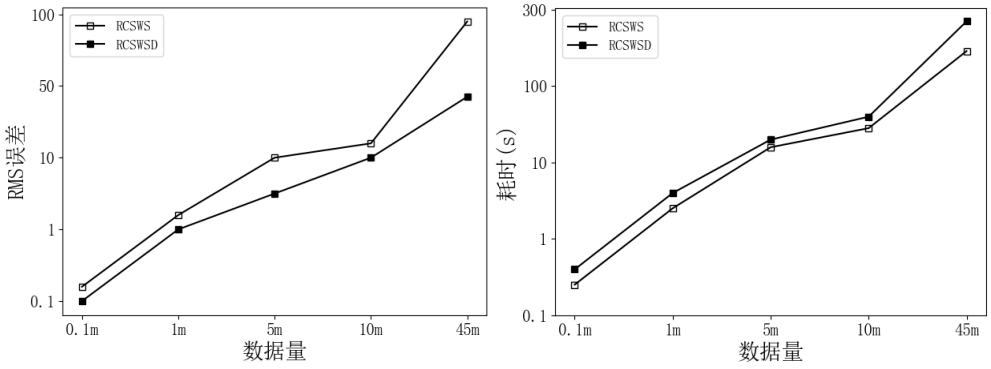
法，并且 RCSWS 和 RCSWSD 修复的结果差别不大，如图 3.13（b）所示。对于多种不同类型的异常修复对比，由于一些无效点的干扰，比如数值为 0 的一些数据会使得平滑的 EWMA 算法过度修复导致正确的数据也被修改从而显示出最坏的修复结果，如图 3.13（c）所示。RCSWSD 算法通过使用方向分析可以更加精确的去修复各种不同类型的异常数据情况，使得算法的修复效果要稍微优于 RCSWS。



(a) RMS　　(b) 时间消耗

图 3.14 无方向属性的不同大小数据集（出租车）

关于各种大小数据集的实验比较，首先使用不同数据量的出租车数据进行比较实验。当数据量较小的情况下，三种算法得到的均方根误差的差别不大，如图 3.14（a）所示。当数据量达到 10 万条数据记录时，EWMA 和 SCREEN 的 RMS 误差值略有增加。RCSWS 对应的 RMS 误差波动小，仍然具有良好的数据修复效果。当数据量达到 1500 万条数据记录时，SCREEN 算法无法处理满足约束内异常的缺点被无限放大。EWMA 对于处理一些较大的数据异常过于平滑，使得数据中原本正确的数据也被修改，导致 RMS 值比 SCREEN 的要大。RCSWS 算法不仅使用范围约束方法解决漂移较大的数据异常，也通过使用滑动窗口的统计方法修复满足约束内的漂移较小的数据异常，使得数据整体修复的精度要优于其他两种方法。由于 RCSWS 需要对原始数据进行约束检测，并对满足约束的数据进行数据变化统计，因此 RCSWS 在大规模数据修复上花费的时间最多，如图 3.14（b）所示。SCREEN 算法需要统计当前点对应数据的候选点，而 EWMA 只需要计算不同权重对应的前一个数据的乘积之和，因此在时间的消耗上面 EWMA 的算法耗时最少。



(a) RMS　　(b) 时间消耗

图 3.15 带有方向属性的不同大小数据集（卡车）

对于数据集中包含不同属性的对比，原始数据中不同的数据属性也会影响算法的清洗效果。文中通过使用运输车辆的数据集对 RCSWS 和 RCSWSD 算法进行实验，实验的 RMS 结果如图 3.15（a）所示。由于运输车辆数据集包含了车辆的运动行驶方向，RCSWSD 可以很好地利用原始数据中的属性信息。当实验数据量较小时，两种算法的修复效果没有明显的差别。当数据量达到 4500 万条时，RCSWSD 算法的 RMS 值明显低于 RCSWS。由于 RCSWSD 增加了对数据的方向分析，算法时间的消耗要高于 RCSWS，如图 3.15（b）所示。对于 4500 万条的数据记录，RCSWSD 的时间消耗几乎是 RCSWS 的两倍。

RCSWS 算法不仅具有 SCREEN 算法对较大异常数据的检测和修复的优势，还可以利用统计方法弥补 SCREEN 在约束范围内不能对较小数据异常检测的缺陷。滑动窗口方法不仅提高了算法的运行效率，而且避免了对整体数据集中数据变化的统计从而使得异常数据的修复方向变得单一。同时，通过近邻的数据还能较好的获取当前检测数据的特征，以及避免原始数据中存在较多的异常数据对数据统计所形成的干扰。

3.7 本章小结

本章主要介绍了对移动对象数据集中存在异常数据的检测和修复问题。通过相关定义移动对象数据集中的不同种类的异常数据信息进行分析，并给出不同异常数据存在的数据特征，并针对不同的异常数据提出了相应的解决方法。对于无序数据而言，文中通过利用原始数据中存在的最大乱序特征提出了 MOHSort 算法，相比于对原始整个数据集的排序 MOHSort 算法对于无序数据排序的效率具有明显的提升，数据集越大，排序效果约明显。对于缺失数据的填充，文中利用缺失数据周围正确数据速度的变化情况，提出了近邻数据的填充方法 NNF，通过过渡数据的这种数据变化来实现车辆原始的运行轨迹实现数据的填充。对于漂移异常数据的修复，文中以范围约束和滑动窗口统计方法相结合的方式不仅修复了违反约束的较大的漂移异常对于约束内较小的漂移异常同样具有修复效果。在此基础上通过对运行轨迹的方向分析进一步提高了异常数据修复的准确率。

第四章 移动对象数据质量评估

移动对象数据在日常生活中每时每刻都在产生，但是规模如此之大的数据能够被很好利用的却很少。首先，定位设备存在一定的误差，容易受到地理位置、气候环境等因素的影响。再者就是采集数据设备的不健全，导致了一些移动对象数据存在缺失、错误等异常问题。这些低质量的移动对象数据不仅对数据研究少有帮助，而且浪费了大量的人力，物力和财力。移动对象数据质量评估可以通过判断原始数据的质量以此来提高设备采集数据的精度。在科学研究上，移动对象数据质量评估也可以快速的判断出大量移动对象数据的质量问题，从而可以选取较高质量的原始移动对象数据来提高实验的可行性。

本章分别从移动对象数据质量评估的准确率和效率两方面进行了研究。首先提出了基于AHP（Analytic Hierarchy Process）的多属性移动对象数据质量评估模型提高评估的准确率，在保证准确率的基础上提出了基于多分段的连续抽样来提高质量评估的效率。

4.1 问题描述

由于 GPS 定位技术和存储技术的进一步发展，大量的移动对象数据已得到了存储和研究。由于地理位置和气候环境的影响，使得设备在同一地点采集到的数据质量也会有所不同。因此，同种数据以及不同种数据之间的质量评估则急需解决。目前，在 GPS 定位研究领域定位的精度还有待提高。通过移动对象数据质量评估方法评估 GPS 定位设备采集到的原始数据并做出相应的质量评分，以此数据质量评分来对 GPS 定位设备进行质量反馈作为后期设备参数维护和分析具有非常重要的参考意义。在大数据时代，每时每刻都会产生大量的移动对象数据，将所有的移动对象数据都进行一遍研究是非常不切实际的。虽然存在大量的移动对象数据，但是并不是所有的数据都是可以为研究使用的。以往研究为了提高移动对象数据的质量，通常会对原始数据进行异常检测和清洗，但是众多的检测和清洗方法都存在或多或少的误差，在清洗方法参数设置中，参数设置的一些偏差则有可能造成对原始数据中异常数据的过度清洗，从而表现出清洗之后的数据反而比原始数据质量低的现象。为了遵循清洗数据中对原始数据修改的最小原则，本文使用移动对象数据质量评估方法对海量的原始数据进行质量评估并从中获取质量较高的数据作为之后研究分析的主要数据。原始数据中挑选出来的数据相比于清洗之后获取的同等质量的数据而言，原始数据的“零”修改更能遵循清洗数据中最小修改的原则，也能直接的反馈出原始移动对象数据信息的价值。

针对移动对象数据质量评估的问题，主要对于质量评估的准确率和效率两个方面进行了研究。移动对象数据质量评估通常需要对移动对象数据中的各种属性进行分析，充分利用现有的属性以及数据中存在的潜在属性，确定各个属性之间权重的关系对于质量评估准确率是至关重要的。判断各个属性之间的权重关系使用人为设定的权重没有较大的说服力。因此，

对于权重的设定设计了基于 AHP 的属性之间相互比较方法，对于不同属性的重要因素可以较为准确的计算出来。由于对整个大规模的数据集进行评估分析非常的耗时，也同样需要提高数据质量评估的效率。使用普通的抽样方法减小数据量可能会造成评估准确的大幅降低，不连续的抽样也不能反映数据的原始信息。为了平衡移动对象数据质量评估准确率和效率之间的关系，提出了基于多分段的快速连续抽样来保证准确率在可接受的阈值范围内来提高数据评估的效率。

4.2 基于 AHP 的多属性数据质量判断

除了分析移动对象数据的表面属性和潜在属性外，还需要对数据属性进行权重的设定。不同属性对于数据集质量评估的影响也是有所不同的，因此使用相同的权重对属性进行设定是不合理的。在现有的权重设定方法中，调查统计法在权重的设定上面消耗的人力物力都比较大。对于公式法和数理统计法来说，使用公式对移动对象数据中上文提及到的属性进行统一的设定，对于同一种特征下的数据权重是比较方便的，但是对于不同特征情况下的数据设定则又要设计另外的权重比例模型。例如数据完整性和数据准确性中的数据缺失和数据精度属性的权重设定在公式法和数理统计法中是难以估计的，因为不同特征不同属性之间的判断规则也会有所不同。目前，在移动对象大数据质量评价研究领域中相应的评价办法还不健全，还需要更深入的研究和探讨。不同属性对于移动对象数据质量的影响也是难以进行人为评定，通过人为设定不同属性的权重缺乏一定的说服力。为了提高质量评估的准确性，文章提出了一种基于层次分析法的移动对象数据多属性质量评估模型，以在同一特征下进行属性的两两比较并通过判断矩阵和一致性检验来更加精确的计算不同特征下对应属性对移动对象数据质量评估的影响。

4.2.1 移动对象数据 AHP 层次结构模型

AHP 是一个将定性与定量研究相结合的、成体系的、高层次的政策分析工具，为多目标、多准则以及无明显结构特征的复杂问题提出了十分重要的决策依据。在多属性的移动对象数据质量评估中，基于 AHP 的方法可以将不同的属性进行两两比较，通过相同性质属性对上层特征的影响来设置影响权重的大小。在对相同特征下属性比较的同时减少了性质不同的诸多属性之间相互比较的困难，提高了移动对象数据质量评估的准确性。以移动对象数据质量为目标构建移动对象数据的 AHP 层次结构模型如图 4.1 所示。其中以数据质量为层次结构的目标层，数据的各种特征作为层次结构的准则层，数据的各种属性作为层次结构的方案层。

移动对象数据质量特征及属性框架如图 4.2 所示。移动对象数据质量特征主要包含了完整性、准确性、一致性、及时性，这些特征对应着层次结构模型中的准则层，分别用 S_1, S_2, S_3, S_4 表示。每个数据特征模块下又包含了不同的数据属性，这些属性对应着层次结构模型

中的方案层，各特征关联着各自的属性，分别用 $A_1, A_2, \ldots, A_{12}$ 表示。

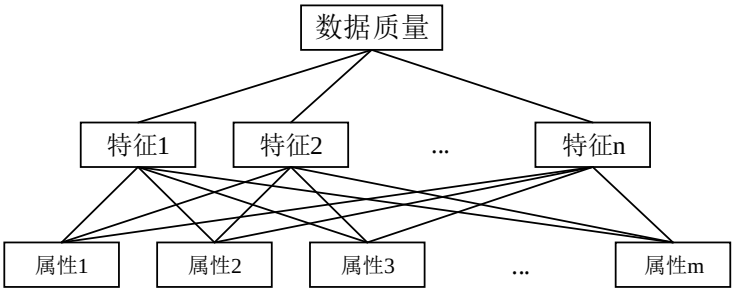


图 4.1 数据质量层次结构模型

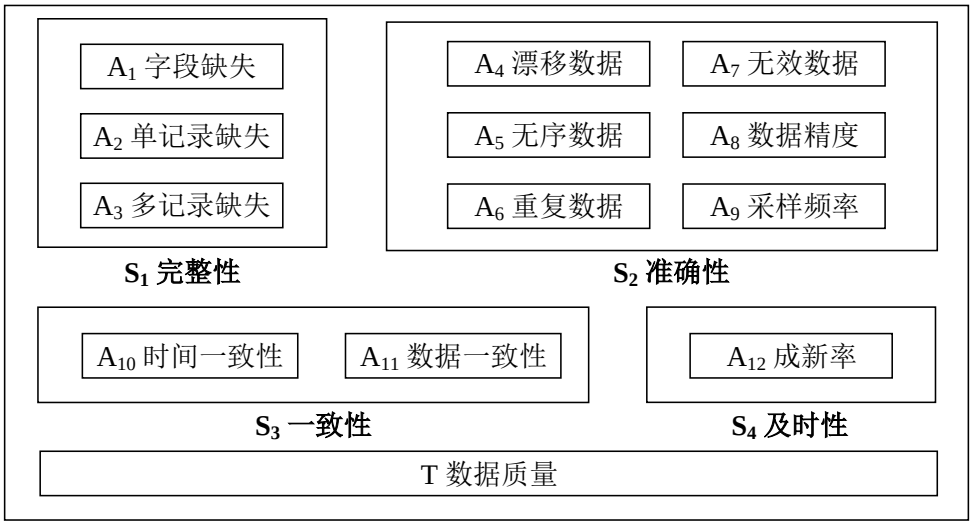


图 4.2 移动对象数据质量特征属性框架

设移动对象数据中各特征对应的模型分别表示为 M_i，各特征对数据质量的影响权重表示为 W_i，则数据质量评分 Q 可以通过公式 $Q=\sum_{i=1}^{4} M_i \cdot W_i = \sum_{j=1}^{12} m_j \cdot w_j$（4.1）得出，其中 m_j 为第 j 个属性模型，w_j 为对应属性模型对数据质量的影响权重。

$$Q=\sum_{i=1}^{4} M_i \cdot W_i = \sum_{j=1}^{12} m_j \cdot w_j \qquad (4.1)$$

4.2.2 移动对象数据多属性约束分析

移动对象数据集中通常包含许多的数据属性，但是不同的定位设备采集存储数据的格式也有所不同。为了使得评估方法适用多数的移动对象数据，文章以基本的移动对象 GPS 数据进行说明。移动对象数据中通常都会包含编号，时间、经度、纬度这四个基本信息。通过原始数据直接显示的属性进行判断数据质量是远远不够的，还需要挖掘出数据中潜在的一些属性信息来提高质量评估的准确率。

数据质量评估包含对移动对象数据中存在的多个特征属性的判断比较。针对移动对象数据管理的主要特点，文章的重点主要从数据的完整性、准确性、一致性、及时性四个数据特征方面来分别分析其对移动对象数据质量判断的影响，并构建相应的特征属性模型。表 4.1 给出了本章中常用的一些符号信息。

表 4.1 符号表

符号	描述
P	移动对象数据集，包含多个数据记录 p_i
N	移动对象数据集的记录数$\|P\|$
W	数据集中对应的权重，包含多个 w_i
T_f	数据集的采样频率
T	层次结构模型中的目标层
S	层次结构模型中的准则层
A	层次结构模型中的方案层
M	不同属性对应的数据模型
λ	判断矩阵特征值
CI	一致性指标
CR	一致性比率

（1）数据完整性。数据完整性是指原始数据集中是否存在数据缺失的情况。因为移动对象数据都是不断采集的，因各种因素的影响，在数据处理中难免产生遗漏现象。数据缺失主要存在于三种情况，分别为属性缺失，单记录缺失、多记录缺失。属性缺失，例如一条记录中经纬度属性任意一个缺失都会使得该条记录的定位数据变得模糊。但是对数据进行修复时相比于单条记录缺失的填充，属性缺失记录可以通过现有属性对缺失属性进行约束来提高数据填充的准确性，对于数据完整性的判断影响较小。多记录缺失是指数据中连续记录缺失的情况，这种缺失情况使用填充修复方法相比于单记录缺失的填充修复误差较大，对于数据完整性的判断影响也同样较大。由于数据缺失的越多对数据完整性的影响越大，数据缺失模型的值越趋近于 1，反之则对数据完整性的影响越小，数据缺失模型的值越趋近于 0。设存在移动对象数据集 $P=\{p_1,p_2,\mathrm{K},p_n\}$，其中 N 为移动对象数据集 P 中记录数的总量，$n=\{n_1,n_2,n_3\}$，其中 n_1 为属性缺失的记录数，n_2 为单记录缺失的记录数，n_3 为多记录缺失的记录数，m_i 为对应的属性模型，则数据完整性模型 M_1 的可由公式 $M_1=(m_1,m_2,m_3),m_i=\frac{n_i}{N},i\in[1,3]$（4.2）计算得出。

$$M_1=(m_1,m_2,m_3),m_i=\frac{n_i}{N},i\in[1,3] \quad (4.2)$$

（2）数据准确性。数据准确性是指数据处理中是否出现了非正常和出错的信息。由于 GPS 定位设备以及恶劣地理环境的影响会造成移动对象数据定位模糊与实际车辆行驶轨迹的真实地理位置具有一些偏差从而形成一些不准确的异常数据信息。其中，不正确的异常数据信息主要分为漂移数据、无序数据、重复数据、无效数据等。原始数据集中，异常数量越多对数据分析准确性的负面影响越严重，使得数据质量也相对较低。$n=\{n_4,n_5,n_6,n_7\}$，其中 n_4 为漂移数据异常的总记录数，n_5 为无序数据异常的总记录数，n_6 为重复数据异常的

总记录数，n_7 为无效数据异常的总记录数，m_i 为对应的属性模型，N 为数据集的总数据记录数量，那么异常数据对应的数据模型则可以用公式 $m_i = \frac{n_i}{N}, i \in [4,7]$（4.3）表示。

$$m_i = \frac{n_i}{N}, i \in [4,7] \quad (4.3)$$

由于不同的定位设备采集数据的精度和频率有所不同，当然对应的原始数据质量也会有所差别，数据的采样频率和定位精度同样会影响数据的准确性。为了较好的评估移动对象数据质量，提高数据质量评估的准确率，文章将移动对象数据定位精度的质量判断标准设置为 1 米，数据采样频率以 1 秒作为质量评估的标准。由于移动对象数据格式不同，以 1 米为单位设置 σ 为定位精度的标准，∂ 为原始数据集的定位精度。$\partial - \sigma$ 的数值越大则表示原始数据集的定位精度与标准定位的差值越大，数据质量越小，因此对于数据精度计算的模型则可以用公式 $m_8 = \frac{1}{\partial - \sigma}$（4.4）表示。

$$m_8 = \frac{1}{\partial - \sigma} \quad (4.4)$$

设 T_f 为原始数据的对应设备的采样频率，采样频率越大，同一时间内的采集的数据记录则越少。在数据集其他情况条件相同时，数据记录数越多，对应的潜在数据价值越大。由于频率小的数据集中已经包含了频率大的数据，因此数据质量越高，则采样频率和数据质量成反比，对应的数据频率模型可以用公式 $m_9 = \frac{1}{T_f}$（4.5）所表示。

$$m_9 = \frac{1}{T_f} \quad (4.5)$$

数据准确性的模型 M_2 可由公式 $M_2 = (m_4, m_5, m_6, m_7, m_8, m_9)$（4.6）表示。

$$M_2 = (m_4, m_5, m_6, m_7, m_8, m_9) \quad (4.6)$$

（3）数据一致性。数据一致性是指数据中的格式或数值是否遵循数据规范。移动对象数据中不同设备采集数据的格式也不一样，例如在时间上的不同格式有“年-月-日 时-分-秒”和“年/月/日 时/分/秒”之分，经纬度的取值上面有“度”、“度-分”、“度-分-秒”以及数值等。不一致的数据格式使得数据处理的复杂度增加，数据质量也将随之变低。除了不一致的格式外，例如数据中经纬度为 0 的数值不一致也违反了数据一致性的准则，影响了数据质量。设 $d = \{d_1, d_2\}$，其中 d_1 表示数据时间不一致的记录数，d_2 表示数据数值不一致的记录数，则对应的数据一致性模型可以用公式 $M_3 = (m_{10}, m_{11}) = (\frac{d_1}{N}, \frac{d_2}{N})$（4.7）表示。

$$M_3 = (m_{10}, m_{11}) = (\frac{d_1}{N}, \frac{d_2}{N}) \quad (4.7)$$

（4）数据及时性。数据及时性是指数据采集的时间到查看或使用数据的时间间隔。文

章使用数据及时性来评估数据从采集到之后每次使用评估之间的时间差来表示数据的成新率。考虑到现今社会基础建设的快速发展，由于道路的损坏和城镇化进程的推进，大量的道路被翻新和重建，这也会影响数据质量评估的准确性。历史移动对象数据虽然具有一定的参考性，但是时代的发展造就了道路调整和技术迭代使得与现今时间相差较小的数据集相比于时间相差较大的数据集在数据的更新上的优势就凸显出来。因为数据采集是持续在一段时间内的活动，将整个数据集的时间都进行一遍计算或从数据采集开始或采集结束记为历史数据集的时间都是不明智的选择，不仅造成在时间间隔上计算浪费大量的时间，而且用数据集的开始或者结束作为数据集的时间都是不准确的。为了计算历史数据集与当前时间之间的时间差，提高数据质量判断的准确性，本文取数据集中开始数据记录和结束数据记录两个时间的平均值用于表示整体数据集的时间。这样不仅节省数据集中所有记录和当前时间之间的时间比较，对于计算所有数据记录时间差的均值具有同样的效果。设 T_{cur} 为数据质量评估的当前时间，t_{begin} 为数据集中记录开始的时间，t_{end} 为数据集中记录结束的时间。由于和当前时间相差越大，数据更新越多，因此数据质量也就越低，则数据一致性模型可以用公式 $M_4=(m_{12})=(1/(1+T_{cur}-\frac{t_{end}-t_{begin}}{2}))$ （4.8）表示。

$$M_4=(m_{12})=(1/(1+T_{cur}-\frac{t_{end}-t_{begin}}{2})) \tag{4.8}$$

4.3 多属性数据质量评估

在移动对象数据层次结构模型中可以看出不同特征下具有不同的数据属性，这样则不需要将所有的属性都放在一起两两比较，只需要比较同一个节点下对应的数据信息即可，避免了性质不同的多属性之间相互比较的困难，提高了数据质量评估的准确性。

定义 4.1（判断矩阵）设 Z 为移动对象数据层次结构模型中的一个非叶子节点，X_1,X_2,L,X_n 为 Z 节点的子节点，设 $1\le i\le j\le n$，子节点 X_i,X_j 对父节点 Z 的影响权重分别为 x_i,x_j，X_i 与 X_j 的影响权重之比为 $x_{i,j}$，满足 $x_{i,j}>0,x_{i,j}=1/x_{j,i},x_{i,i}=1$，以 X_1,X_2,L,X_n 的影响权重之比所构成的矩阵称为节点 Z 的判断矩阵记为 $Z(X)$，矩阵公式如公式（4.9）所示。

$$Z(X)=\begin{array}{c} \\ X_1 \\ X_2 \\ \mathrm{M} \\ X_n \end{array}\begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \mathrm{K} & X_n \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & \mathrm{K} & x_{1,n} \\ x_{2,1} & x_{2,2} & \mathrm{K} & x_{2,n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ x_{n,1} & x_{n,2} & \mathrm{K} & x_{n,n} \end{pmatrix} \end{array}=\begin{pmatrix} 1 & x_{1,2} & \mathrm{K} & x_{1,n} \\ 1/x_{1,2} & 1 & \mathrm{K} & x_{2,n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 1/x_{1,n} & 1/x_{2,n} & \mathrm{K} & 1 \end{pmatrix} \tag{4.9}$$

对于移动对象数据层次结构模型，根据判断矩阵的定义则可以对各层的非叶子节点定义出对应的判断矩阵。分别为 $T(S),S_1(A),S_2(A),S_3(A),S_4(A)$ 五个判断矩阵，矩阵公式如

$$S_2(A)=\begin{pmatrix} a_{4,4} & a_{4,5} & \text{K} & a_{4,9} \\ a_{5,4} & a_{5,5} & \text{K} & a_{5,9} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ a_{9,4} & a_{9,5} & \text{K} & a_{9,9} \end{pmatrix},\quad S_3(A)=\begin{pmatrix} a_{10,10} & a_{10,11} \\ a_{11,10} & a_{11,11} \end{pmatrix},$$

（4.10）所示。

$$T(S)=\begin{pmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} \\ s_{2,1} & s_{2,2} & s_{2,3} & s_{2,4} \\ s_{3,1} & s_{3,2} & s_{3,3} & s_{3,4} \\ s_{4,1} & s_{4,2} & s_{4,3} & s_{4,4} \end{pmatrix},\quad S_1(A)=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

$$S_2(A)=\begin{pmatrix} a_{4,4} & a_{4,5} & \text{K} & a_{4,9} \\ a_{5,4} & a_{5,5} & \text{K} & a_{5,9} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ a_{9,4} & a_{9,5} & \text{K} & a_{9,9} \end{pmatrix},\quad S_3(A)=\begin{pmatrix} a_{10,10} & a_{10,11} \\ a_{11,10} & a_{11,11} \end{pmatrix},\qquad (4.10)$$

$$S_4(A)=\left(a_{12,12}\right)=1$$

定义 4.2 （绝对一致性）设 *Z(X)*为节点 *Z* 对子节点集合 *X* 的一个判断矩阵，其中 $1\le i\le j\le k\le n$，设 $x_{i,j}=p, x_{j,k}=q$，如果 $x_{i,k}=p\cdot q$，那么判断矩阵 *Z(X)*具有绝对一致性，且为一致阵。

在判断矩阵 *Z(X)*为一致阵时，矩阵的转置 $Z(X)^T$ 也为一致阵，矩阵 *Z(X)*的一个特征根 λ 的值为 n，其他的值都是零。根据一致阵的性质，可得矩阵中任一列均是 λ 的特征向量，以 $W=\{w_1,w_2,\text{K},w_n\}$ 表示对应的特征向量，那么 *W* 归一化的向量 W' 则可以由公式

$$W'=\begin{pmatrix} w_1' \\ w_2' \\ \text{M} \\ w_n' \end{pmatrix}=\frac{1}{\sum_{i=1}^{n} w_i}\begin{pmatrix} w_1 \\ w_2 \\ \text{M} \\ w_n \end{pmatrix}$$

（4.11）所得出，且 $\sum_{i=1}^{n} w_i'=1$，则 w_i 为 *X* 层中第 i 个节点对上层 *Z* 的影响权重。

$$W'=\begin{pmatrix} w_1' \\ w_2' \\ \text{M} \\ w_n' \end{pmatrix}=\frac{1}{\sum_{i=1}^{n} w_i}\begin{pmatrix} w_1 \\ w_2 \\ \text{M} \\ w_n \end{pmatrix}\qquad (4.11)$$

根据判断矩阵的定义可知其也是互反阵，此时矩阵的最大特征根 λ 大于等于 n。当判断矩阵不具有一致性时，$\lambda-n$ 的值越大，不一致性越强烈，当且仅当 $\lambda=n$ 时，矩阵为一致阵。为了提高移动对象数据属性对数据质量影响权重的计算精度，文章通过一致性指标 CI 来对矩阵的不一致性进行约束， CI 可由公式

$$CI=\frac{\lambda-n}{n-1}$$

（4.12）计算得出。CI 的值越接近于零，代表矩阵越接近于一致性，而 CI 的数值越大，代表矩阵的一致性就越糟糕。为了判断 CI 的大小范围，引入了随机一致性指标 RI，表 4.2 中给出了 1 到 10 维的判断矩阵一千次的随机一致性指标值，CR 可由公式

$$CR=\frac{CI}{RI}$$

（4.13）计算得出。当 CR < 0.1 时，则可视为判断矩阵的不一致性在可接受的范围内。对于不在可接受范围内的判断矩阵，则需要重新建立判断矩阵并进行一致性检测。对于符合不一致性约束的判断矩阵，则计算出最大特征值对应的归一化特征向量，向量中的数值即是判断矩阵所对应的子节点的属性对于父节点的影响权重。

$$CI = \frac{\lambda - n}{n - 1} \tag{4.12}$$

表 4.2 随机一致性指标值

维数	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49

$$CR = \frac{CI}{RI} \tag{4.13}$$

算法 4.1 基于 AHP 的多属性数据质量评估算法

输入：各属性计算模型 $m_1, m_2, \ldots, m_n$

输出：质量评分 Q

1. α ← false;
2. /**层次单排序**/
3. **While** !α **do**
4. 构建 $T(S)$的判断矩阵;
5. CR ← 计算 $T(S)$矩阵的一致性
6. **If** $CR < 0.1$ **then**
7. α ← true
8. α ← false;
9. /**层次总排序**/
10. **While** !α **do**
11. **For** i ← 1 **to** 4 **do**
12. ε ← false;
13. **While** !ε **do**
14. 构建 $S_i(A)$的判断矩阵
15. CR_i ← 计算 $S_i(A)$的一致性比率
16. **If** $CR_i < 0.1$ **then**
17. ε ← true
18. CR ← 计算 A 层的总排序一致性比率
19. **If** $CR < 0.1$
20. α ← true
21. $Q \leftarrow 0$
22. **For** i ← 1 **to** 12 **do**
23. Q += $m_i \cdot \sum_{j=1}^{4} s_j a_{i,j}$

24. **Return** *Q*

层次结构模型中各个层次对应的判断矩阵具有一致性或者在非一致性可接受的范围内时，则可以对各层进行层次总排序来获取各层对总目标的影响权重。以移动对象数据层次结构模型为例：S 层的层次单排序即为对 T 层的层次总排序记为 s_1, s_2, s_3, s_4，满足 $CR < 0.1$，记 A 层的 A_1, A_2, K, A_n 对 S 层中各个节点的层次单排序一致性指标分别为 CI_1, CI_2, CI_3, CI_4，随机一致性指标分别为 RI_1, RI_2, RI_3, RI_4，则 A 层的层次总排序一致性比率可由公式（4.14）计算得出。若 A 层的一致性比率 $CR < 0.1$，则层次总排序满足一致性检验。则 A 层第 i 个节点对目标 T 层的影响权重可以由公式 $\sum_{j=1}^{4} s_j a_{i-j}, 1 \le i \le 12$（4.15）计算得出，其中 s_j 表示 S 层第 j 个节点对 T 的影响权重，a_{i-j} 表示 A 层第 i 个属性对 S 层第 j 个属性的影响权重。

$$CR = \frac{s_1 CI_1 + s_2 CI_2 + s_3 CI_3 + s_4 CI_4}{s_1 RI_1 + s_2 RI_2 + s_3 RI_3 + s_4 RI_4} = \frac{\sum_{i=1}^{4} s_i CI_i}{\sum_{i=1}^{4} s_i RI_i} \tag{4.14}$$

$$\sum_{j=1}^{4} s_j a_{i-j}, 1 \le i \le 12 \tag{4.15}$$

基于 AHP 的多属性移动对象数据质量评估算法如算法 4.1 所示。首先，根据移动对象数据的层次结构模型构建 S 层对目标 T 层的判断矩阵并计算一致性比率，满足一致性的范围后构建下一层 A 层对 S 层中各节点的判断矩阵并计算一致性比率。对于满足一致性比率的 A 层进行层次总排序，计算满足层次总排序中 A 层的各个属性计算模型和总目标的影响权重的乘积之和即为移动对象数据质量分数。

4.4 基于多分段的快速连续抽样

对于海量的移动对象数据集，如果将数据集中所有的数据都进行异常检测分析其质量，在评估前期对数据的异常检测工作便成了计算量和耗时量非常大的一项任务。为了节省整体算法处理的时间，提高算法的效率，则不能对移动对象数据中所有的数据都进行异常检测，应通过获取移动对象数据中的部分数据样例来表示整个数据集的特征，以小样本的数据用例进行异常分析，从而减小异常检测所消耗的时间提高算法的效率。根据移动对象数据的特性，文章提出了基于多分段的快速连续抽样方法采样原始数据来获取具有原始数据集特征的代表性数据，以此减小异常检测分析的数据量，提高移动对象数据质量评估的效率。

4.4.1 数据特性

移动对象数据的连续性。在定位设备的运行中，移动对象数据的采集是连续的。在没有数据缺失的情况下，定位设备采集的下一个数据点和上一个数据点的时间差必为设备采集

数据的固定频率。根据移动对象数据的连续性特点，在检测原始数据中是否存在缺失数据的情况时，必须要保证检测的数据是在一段时间内持续采集的，这样才能准确的对数据进行评估，若是选取离散的数据点进行检测那么缺失的数据量会明显偏高，从而评估出不准确的数据质量。根据数据连续性的特点，因此采样的数据需要具有连续性。

移动对象数据的依赖性。移动对象数据中下一个定位数据的数值情况会依赖前一个数据点的数值并在前一个数据点规定的约束范围内。移动对象数据的依赖不仅体现在时间上的依赖在空间位置上也具有依赖关系。在连续的数据集中，以前一个数据点的数值为例，在时间上，后一个数据的时间应和前一个数据时间相差一个频率。在空间位置上，后一个数据点的经纬度信息依赖前一个正确数据点的经纬度信息，两点之间的距离不能超过运动的物体以其最大速度在频率内经过的距离。根据数据的依赖性，在对数据进行取样的时候为了检测数据中的漂移数据对其进行质量评估也需要获取连续的数据集。若选取离散的数据进行异常检测，会由于数据点之间的时间相距较远而使得异常数据检测不准确影响质量评估，因此采样的数据需要具有连续性。

移动对象数据的相似性。在数据采集期间，运动的物体通常会以速度变化较小以及速度方向变化较缓的方式运动，这样的运动方式会使得在不同的时间段内存在相似运动的移动对象数据。相似的移动对象数据不仅在多数据集中出现，在单一的数据集中也会出现相同的移动对象数据。根据数据相似性特点，在对数据集进行抽样来表现数据特征时则不能把数据集归为一段进行提取，提取一段连续数据相当于原始数据集中其他的移动对象数据都相似于提取出来的数据，从而影响数据质量评估。为了体现原始数据集中多种轨迹特点，对数据集进行多分段处理来表示原始数据集中不同情况下的数据情况，因此采样的数据需要进行多分段采样保证评估的准确性。

4.4.2 多分段的快速连续抽样算法

算法 4.2 多分段的快速连续抽样

输入：移动对象数据集 P

分段数 n

分段大小 m

输出：代表性数据集 P'

1. $N \leftarrow P.size/n$
2. $P' \leftarrow$ null
3. **If** $N < m$ **then**
4. **Return** P'
5. **For** $i \leftarrow 1$ **to** n **do**
6. /**获取每一个分段中间一部分连续数据**/

7. $p_i' \leftarrow$ 获取数据集 P 中 $(i-1)\cdot N+(N-m)/2$ 到 $(i-1)\cdot N+(N+m)/2$ 的数据记录
8. $P'.push(p_i')$
9. **Return** P'

通过移动对象数据的特性分析得出在对数据进行样本采样时以多分段的连续抽样方法获取的数据样本可以很好的满足原始数据的特征。对移动对象数据集进行随机抽样的方法处理简单执行效率高，对于随机抽样的数据同样可以检测出数据集中数据一致性的问题，但是对于抽样数据中的缺失数据及异常数据则难以进行判断。以多分段连续抽样的方法来提取移动对象数据集中的数据不仅具有随机抽样方法可以检测数据一致性问题，还可以通过连续的数据集检测数据中缺失数据和异常数据。将整个数据集进行多个分段并从各个分段中抽取连续的数据作为对应分段的代表数据相比于将数据集作为一个整体抽取一段连续的数据而言，多分段的抽样数据可以很好的满足原始数据在不同运动时间，不同运动状态所具有的数据特点。在多分段的数据连续抽样中，为了使得连续的数据更能符合对应分段的特性，在连续数据的抽样上面抽取了分段数据集中间连续的数据作为采样样本。多分段的连续抽样算法如算法 4.2 所示。

为了获取原始数据集的质量评分需要把分段采样的数据进行评估，对于多分段的数据采样将每一分段数据获取的质量评分以加权平均的方式获取质量评分作为整体数据集的总评分。设将移动对象数据分为 n 段，每个分段提取的数据对应的数据质量评分分别为 q_1,q_2,K,q_n，不同的分段对应的权重分别为 w_1,w_2,K,w_n，那么整体数据集的质量评分可以由公式 $Q=\dfrac{\sum_{i=1}^{n}q_i\cdot w_i}{\sum_{i=1}^{n}w_i}$（4.16）计算得出。基于多分段的快速连续抽样的质量评估算法如算法 4.3 所示。

$$Q=\frac{\sum_{i=1}^{n}q_i\cdot w_i}{\sum_{i=1}^{n}w_i} \qquad (4.16)$$

算法 4.3 基于多分段的快速连续抽样的质量评估算法

输入：各属性计算模型 $m_1, m_2, \ldots, m_n$

移动对象数据集 P

分段数 n

分段大小 m

各分段权重 W

输出：质量评分 Q

1. /**调用算法 4.2**/
2. $P' \leftarrow$ 多分段连续抽样(P, n, m);
3. While $p_i' \in P'$ **do**

4. /**调用算法 4.1**/
5. $q_i \leftarrow$ 多属性质量评估($m_1, m_2, \ldots, m_n$)
6. $++i$
7. $Q \leftarrow \sum_{i=1}^{n} q_i \cdot w_i \Big/ \sum_{i=1}^{n} w_i$
8. **Return** Q

4.5 实验与性能评估

本小节在 Ubuntu14.04 环境下，基于可扩展的时空数据库 SECONDO 以 C/C++语言进行对基于 AHP 的多属性数据质量评估算法及多分段连续抽样评估算法进行了实验。实验数据集采用卡车数据集以及微软采集的出租车数据集[77][78]，以共有的 id，时间，经度和纬度四个属性进行实验评估。

4.5.1 AHP 多属性的质量评估实验

针对数据质量的评估，实验分别对相同的数据集及不同的数据集进行了比较，对于不同的数据集实验采用了相同的属性进行评估。移动对象数据层次结构模型中各层因素对上一层的重要性比较参数采用了 Santy 的 1-9 标度方法进行设置，1 表示两个因素相比具有同样重要的作用，2-9 中数字越大表示两个因素相比一个因素比另一个因素越重要，倒数则表示两个因素相反比较的结果。虽然数据集类型不同，但是为了方便比较相同属性的数据质量，对于不同的数据集在层次因素中的相互比较采用了相同的判断矩阵进行实验。

数据的完整性是对数据质量影响最大的特征因素，没有数据谈何数据质量，其次数据准确性会严重干扰数据相比于其他两种数据特征对数据质量的影响要大，数据及时性对于数据质量的影响固然重要但是对于其他的特征及时性只会在特定的实时应用场景影响才会较大，例如股票市场。因此文章中认为及时性相比于其余数据特征对于数据质量的影响较弱，当然也可以对不同的应用场景侧重分析特有的数据特征。在移动对象数据层次结构模型中 S 层中的数据特征对 T 层的移动对象数据质量影响的相互比较产生的判断矩阵信息如表 4.3 所示。为得到更加精确的判断矩阵，文章将新构建的判断矩阵一致性比率以最大程度降低并保持在非一致性可接受的范围内，其判断矩阵的最大特征值 λ 为 4.0080，对应归一化的特征向量为 $w = (0.5375, 0.3027, 0.1055, 0.0543)$，归一化的特征向量即为 S 层中各数据特征对数据质量影响的权重。

表 4.3 数据质量特征判断矩阵信息

	S_1	S_2	S_3	S_4	CR	特征值	特征向量
S_1	1	2	5	9			0.5375
S_2	1/2	1	3	6	0.0030	4.0080	0.3027
S_3	1/5	1/3	1	2			0.1055

S_4	1/9	1/6	1/2	1		0.0543

A 层的数据属性对 S 层中各个数据特征影响产生的判断矩阵设置如下。其中 S_1 数据完整性属性判断矩阵信息如表 4.4 所示，为一致阵一致性比率 CR = 0，最大特征值 λ 为 3，对应的归一化特征向量为(0.1429，0.2857，0.5714)，S_2 数据准确性属性判断矩阵信息如

表 4.5 所示，为非一致阵，一致性比率 CR = 0.0059，满足 CR < 0.1 在非一致性可以接受的范围内，最大特征值 λ 为 6.0370，对应的归一化特征向量为(0.1948, 0.0672, 0.0432, 0.3896, 0.1948, 0.1104)，S_3 数据一致性属性判断矩阵信息如表 4.6 所示，为一致阵，最大特征值为 2，对应归一化特征向量为(0.2000, 0.8000)。S_4 数据及时性属性的判断矩阵信息如表 4.7 所示，由于只有一个属性，因此对应特征的影响权重即为属性的影响权重。

表 4.4 S_1 数据完整性属性判断矩阵信息

	A_1	A_2	A_3	CR	特征值	特征向量
A_1	1	1/2	1/6			0.1111
A_2	2	1	1/3	0	3.000	0.2222
A_3	6	3	1			0.6667

表 4.5 S_2 数据准确性属性判断矩阵信息

	A_4	A_5	A_6	A_7	A_8	A_9	CR	特征值	特征向量
A_4	1	3	4	1/2	1	2			0.1948
A_5	1/3	1	2	1/6	1/3	1/2			0.0672
A_6	1/4	1/2	1	1/8	1/4	1/3	0.0059	6.0370	0.0432
A_7	2	6	8	1	2	4			0.3896
A_8	1	3	4	1/2	1	2			0.1948
A_9	1/2	2	3	1/4	1/2	1			0.1104

表 4.6 S_3 数据一致性属性判断矩阵信息

	A_{10}	A_{11}	CR	特征值	特征向量
A_{10}	1	1/4	0	2	0.2000
A_{11}	4	1			0.8000

表 4.7 S_4 数据及时性属性判断矩阵信息

	A_{12}	CR	特征值	特征向量
A_{12}	1	0	1	1

由于以上判断矩阵满足一致性或在非一致性的可接受范围内，则 A 层上各个属性对总目标层 T 层的影响权重以公式 $\sum_{j=1}^{4} s_j a_{i-j}, 1 \le i \le 12$ （4.15）计算出的结果如表 4.8 所示。分析原始数据集中各个属性模型占比情况，模型占比与对应权重的乘积之和即为原始数据集的质量评估分数。

表 4.8 属性权重信息

属性	影响权重
A_1 属性缺失	0.0597
A_2 单记录缺失	0.1194
A_3 多记录缺失	0.3583
A_4 漂移数据	0.0590
A_5 无序数据	0.0203
A_6 重复数据	0.0131
A_7 无效数据	0.1180
A_8 数据精度	0.0590
A_9 采样频率	0.0334
A_{10} 时间一致	0.0211
A_{11} 数据一致	0.0844
A_{12} 成新率	0.0543

针对多属性移动对象数据质量评估，以出租车和卡车两种数据集进行实验，分别以不同数据量的原始数据集进行分析多属性的模型占比，得出的实验结果如图 4.3 所示。在质量评分的时间消耗上面，随着数据量的增加消耗的时间也随之增加。在数据的评分上面，随着数据量的增加，数据质量评分存在上升后又逐渐下降的现象。由于较小的数据量是截取整个数据集的一部分，使得数据量较小时存在的缺失或异常数据占比情况也随之变少。在到达一定数据量时异常数据的占比情况上升降低了数据质量使得数据评分随之下降。对于数据量越来越大的数据集异常数据的占比情况稍微减小且趋于稳定使得数据评分也随之有一些微小的上升且呈现出逐渐稳定的趋势。

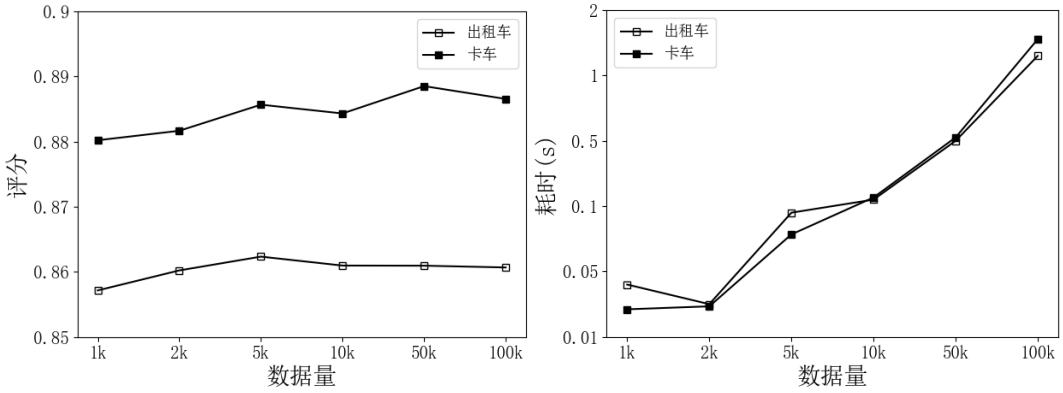


图 4.3 多属性的轨迹数据质量评估

4.5.2 多分段连续抽样质量评估算法

在对移动对象数据集进行分段抽样评估实验中，以固定大小 100k 的出租车数据集作为原始数据集对多分段连续抽样质量评估算法进行测试，实验结果如图 4.4 所示。在质量评估的评分上，随着抽样数据的占比减小，分段次数较小的数据由于异常数据或正常数据比较集中使得数据质量评分波动幅度也较大。抽样数据的减小，不同分段次数之间的质量评分也呈

现出下降的趋势。当抽样数据大于 60%时的质量评分和整体数据集的质量评分相比误差较小。抽样数据占比小于 40%之后的质量评分开始有了明显的下跌，且质量评分之间的误差逐渐变大。在质量评估的时间消耗上，抽样数据的占比越少所需要的评估时间也随之减少。因此需要在保证质量评分稳定的前提下来提高质量评估的效率。由实验结果分析得出当采样数据占比整体数据集 60%时，多分段连续抽样评估算法以较小的评分误差在效率上相比于对整体数据集评估算法而言有了明显的提高。

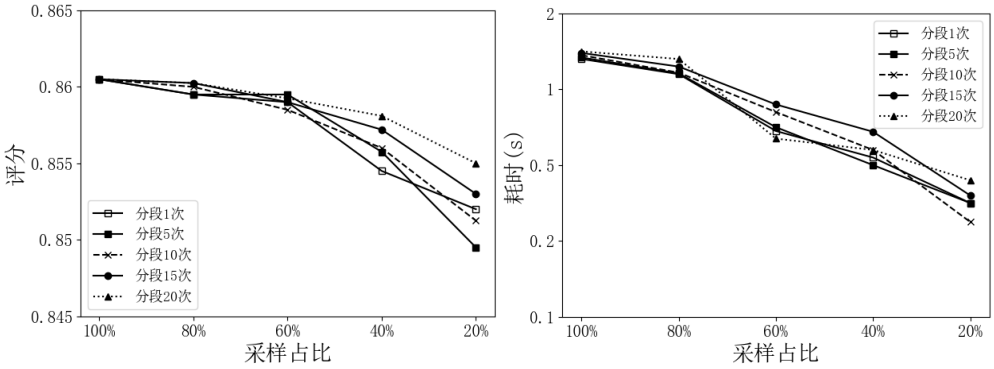


图 4.4 多分段连续抽样质量评估

4.6 本章小结

本章节介绍了移动对象数据质量所带来的一些问题并针对数据质量提出了基于 AHP 的多属性数据质量评估算法和多分段连续抽样评估算法。通过分析不同的数据属性建立数据模型，以多属性之间重要性的相互比较分析对应数据属性的权重，相比于现有的均值、指数权重的设定方法更具有说服力。为了保证数据评分误差在可接受的前提下提高移动对象数据质量评估的效率，提出了以多分段连续采样的方法抽取数据中的数据进行分析。相比于其他的抽样方法，多分段连续抽样也最大化的保存了原始数据的特征。实验结果表明，多分段连续抽样的评估算法在保证数据评分的前提下在评估的效率方面相比于对整个数据集的评估效率有了明显的提高。

第五章 移动对象数据异常清洗工具

移动定位设备的发展使得对大规模移动对象数据研究成为了热点。海量的原始轨迹数据通常包括许多的错误和异常数据信息。为了充分利用现有海量的 GPS 数据，主要的任务是检测修复原始轨迹数据中不同种类的异常信息，以支持应用程序需求。本章设计开发了一个名为 GPSClean 的工具去检测和修复轨迹数据中的异常点。这并非一个单独的软件，而是内嵌到开源的移动对象数据库系统 SECONDO 中的框架。该框架可以很好的支持轨迹数据中六种异常的修复并通过 GUI 界面展示修复之后数据记录的分布情况，通过修复之后的干净数据和原始数据对比来展示框架清洗的效果。对于干净的移动对象数据，还可以通过 GUI 界面观察车辆的历史运动轨迹，这对城市交通建设具有较好的引导作用。

5.1 清洗工具实现

定位设备的普及使得对海量移动对象数据的研究成为了可能，但是由于不确定因素的干扰，原始采集的数据记录中总会包含一些各种类型的异常数据信息。这些异常的错误数据信息严重的干扰了科学数据实验的效果，影响了原始数据的特征。为了提高移动对象数据的质量，较好的修复数据中的各类异常数据成为了迫在眉睫的事情。高质量的移动对象数据不仅对给数据分析提供了基石，而且对于数据挖掘和无人驾驶技术的发展起到了重要的推动作用。为了能够自动清洗 GPS 数据中存在的异常数据获取干净的数据记录，文中研发了一种面向移动对象数据异常清洗工具 GPSClean。该工具是在 Ubuntu 环境下以 C/C++语言在可扩展的移动对象数据库 SECONDO 中实现的。通过自定义的操作符来实现对数据的输入，清洗和转换的功能。通过 SECONDO 的 GUI 界面对比原始数据和清洗之后干净的数据可以直接的看出异常数据清洗的效果。

5.1.1 GPSClean 处理流程和框架

移动对象数据集中的异常数据主要包括重复数据、无序数据、缺失数据、无效数据、漂移数据、模糊数据等。由于移动对象数据为时间序列具有随着时间排序的特性，后一个采集的数据情况会依赖于前一个数据的数值，以前一个数据记录作为基础进行相应的变化，因此在对各种类型异常数据处理之前，无序数据异常的处理必须首先进行检测和修复。若是没有检测数据的有序情况，而直接依赖原始顺序对数据进行填充或者漂移异常的修复，那么填充或者修复的数据结果也可能因为数据乱序的情况使得前一个较大的数据值来影响当前检测修复的数据值，使得数据修复没有可靠的准确性。对于有序的数据集，将数据集中的重复数据进行去重则变成了一项简单的工作，只需要对比下一个数据记录是否相同则可以修复重复数据的异常问题。在前项工作完成之后，则需要对数据中的缺失数据进行填充。因为对于漂

移数据的检测依赖于前一段时间内数据的变化情况，如果前一段时间原始数据缺失过多则会降低漂移数据修复的准确性。对于无效数据例如经纬度为 0 的数值，由于此时的数值没有参考意义，如果简单的将无效数据进行删除则会转变成了异常数据类型中的缺失数据。为了遵循数据修改的最小原则，文中将无效数据归并为缺失数据类型异常进行填充。由于模糊数据是由于设备信号影响，从而显示数据的定位精度不准确造成数据模糊。模糊数据同样也是偏离原始运行轨迹的异常数据信息，在对数据的修复中将其归纳为漂移数据进行处理，使用漂移异常的修复方法对其进行修复尽可能的保证了数据的原始特征。

移动对象数据清洗工具 GPSClean 主要包含了四大部分，依次是数据预处理，数据填充，数据修复和数据转换。GPSClean 的系统框架图如图 5.1 所示。在数据的预处理部分，工具的主要功能是对原始数据集中的无序数据和重复数据进行清洗。对于数据填充的部分，工具主要利用近邻数据的填充方式对原始数据集中的缺失数据或者无效数据进行补充。在数据的修复部分，工具针对数据集中的漂移数据和模糊数据进行了检测修复。为了能够观测清洗之后移动对象数据的历史的运行轨迹，系统还将清洗之后干净的数据转换为移动对象数据类型 MPoint，通过查看车辆的运行轨迹更能直观的看出数据清洗的效果。

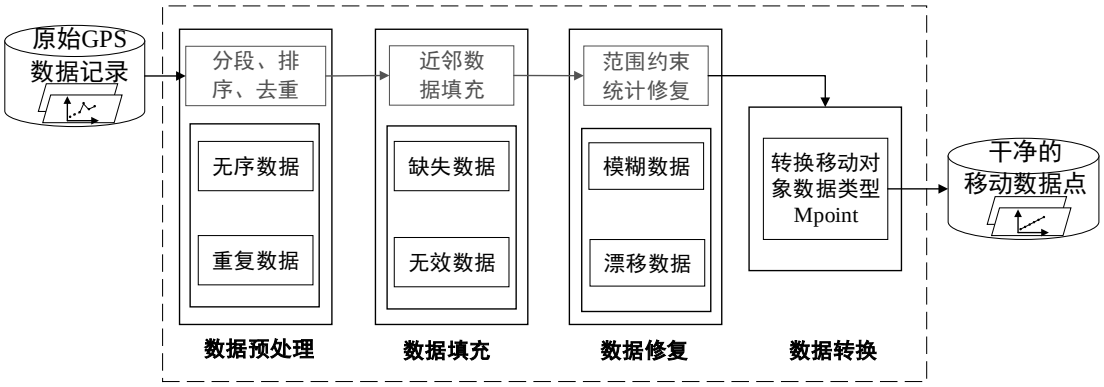


图 5.1 GPSClean 系统框架

5.1.2 数据预处理

移动对象数据的预处理是准确检测和修复异常数据信息的基石，主要的功能包括对数据的分段、排序、去重。有序的数据对于后续的数据处理包括数据填充、数据修复、数据转换都有至关重要的作用。

对于无序数据的检测和排序，通过对原始数据顺序的遍历来查看是否有后一个数据的时间小于前一个数据的时间以此来判断数据的有序性。对于无序的数据，为了提高对大规模数据排序的效率，提出了基于堆排序的原始数据最大无序偏移的排序算法，通过少量数据的比较提高排序效率。对于有序数据集中的重复数据，则只需要判断下一个数据是否和当前检测数据相同就可以对重复数据起到去重的效果。

5.1.3 数据填充

由于带有 GPS 定位的设备在关闭时不收集数据，因此在数据填充之前需要根据两个连续点之间的时差将数据进行分段。无效数据是具有不规则偏移轨迹的数据，对原始数据本身没有实际意义，例如经纬度的数据 0。为了修复无效数据，维护数据的完整性，工具中将无效数据视为缺失数据进行处理。根据对缺失数据记录附近正常数据的变化，考虑到车辆行驶的规律对数据的这种变化进行匀加速的过渡处理。

近邻数据填充算法可以很好地填充轨迹数据中的缺失记录。根据正确的数据计算出缺失数据周围的近似速度，因此可以很容易地计算缺失数据周围正确数据速度的变化情况。以初始速度和加速度，可以计算出在固定频率内经过的距离。在缺失数据前一个的正确数据行驶方向上取距离相同的位置点作为数据的填充点。将新填充的数据点作为正确的数据点，更新下一个缺失数据点。近邻数据填充法不仅具有平均填充法的优点，而且在填充变速运动轨迹时更符合车辆行驶规律。

5.1.4 数据修复

漂移数据的修复是移动对象数据清洗的重要组成部分。为了减少对原始数据的修改，工具通过使用范围约束的方法对违反约束、偏离运动轨迹较大的数据进行修复。在此基础上设计了基于滑动窗口的最大似然估计方法，解决了约束范围内偏离轨迹较小的异常数据不能检测和修复的问题。通过统计窗口中数据的变化概率，使用数据变化情况中概率占比最大的数值来对占比概率小于阈值的数据进行异常修复。

范围约束可以很好地检测出明显偏离轨迹的异常数据信息。根据当前检测点前一段正确数据的初始速度和加速度来计算在频率内可以通过的最大距离。将不在前一个正确数据的有效范围内的检测数据记为偏差较大的异常。基于滑动窗口的概率统计算法主要用于范围约束内异常数据的检测和修复。对于范围约束修复的数据，则不会用滑动窗口统计方法对其进行二次修复。通过计算窗口中每个时间点对应的加速度，计算窗口中每个加速度的概率，以窗口中最大概率对应的加速度对概率低于阈值的数据点进行修复。

5.1.5 数据转换

对于清洗之后干净的 GPS 数据记录，工具可以对其进行数据类型的转换并将其存储在数据库中，类型转换格式如图 5.2 所示。通过 convertP2UU、convertUU2UP、convertUP2MP 三个操作符则可以将系统中的 Point 类型转化为移动对象的数据类型 MPoint。类型转换不是必须的，用户可以根据自己的需求进行选择。对原始 GPS 数据进行清洗后，可以通过 DBMS 的数据导出功能直接获得干净的数据。当然，清洗后的干净数据也可以转换为移动对象数据类型用于观察车辆行驶的历史轨迹。通过历史轨迹可以方便用户对移动对象执行时空查询和任务分析，例如检测交通拥挤区域。

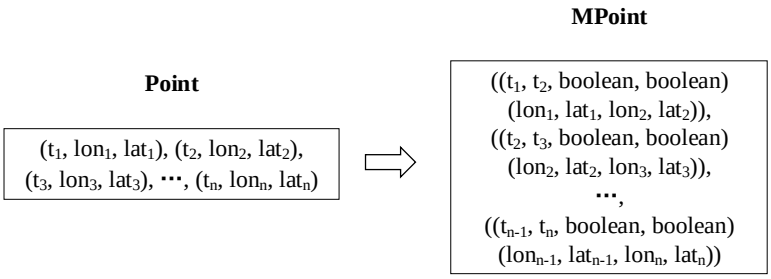


图 5.2 数据类型转换

5.2 系统演示

本章介绍的工具是嵌入在开源的可扩展移动对象数据库 SECONDO 中，并以数据库现有功能为基础进行扩展实现的。以微软采集的出租车数据集[77][78]的一部分作为实验数据进行系统演示，其中包含了 99494 条数据记录。将清洗前后的数据分布通过 SECONDO 的 GUI 界面进行展示，可以直观的看出工具清洗异常数据的效果。

将出租车的原始轨迹数据上传到工具中，车辆的 GPS 数据记录分布如图 5.3 所示。从图中可以看出，由于无效的 GPS 数据记录，例如 0，导致了数据的分布变大，使得原本经纬度信息变化就很小的原始数据聚集在了图中的右上角，无法很好的显示正常数据的特征。这种异常数据给数据分析和显示带来了严重的干扰。

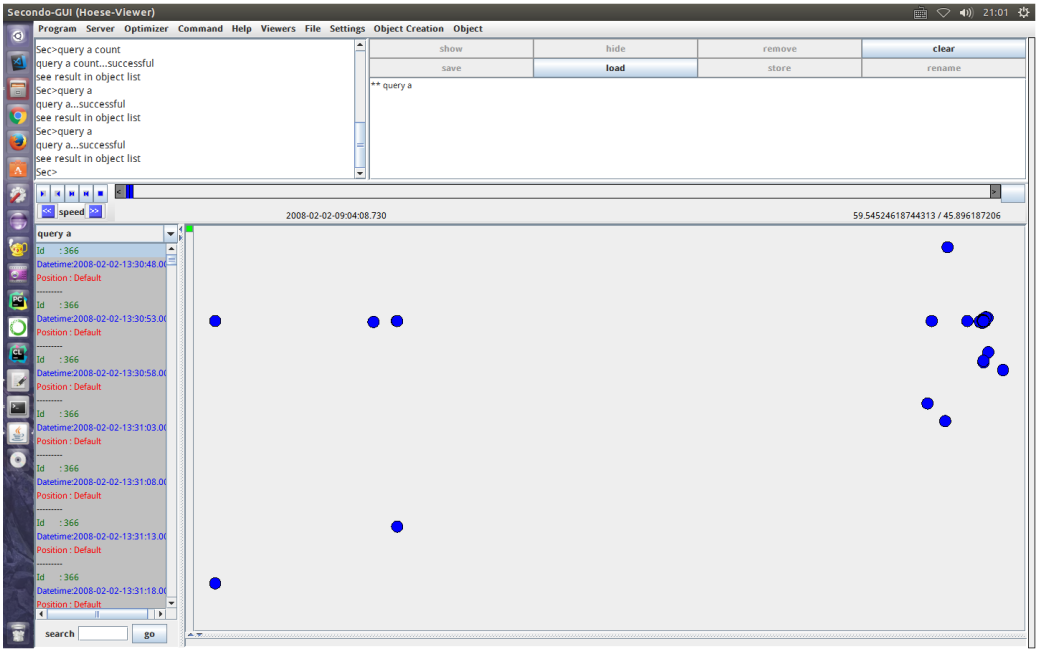


图 5.3 原始数据轨迹分布

为了方便观察原始轨迹数据中的信息，则将图中右上角数据集中的部分进行放大如图 5.4 所示。对比原始没有放大的数据分布来说，可以明显的看出无效数据记录对于数据分布的影响，使得分布范围扩大，原本正常数据轨迹范围则变得相对较小。

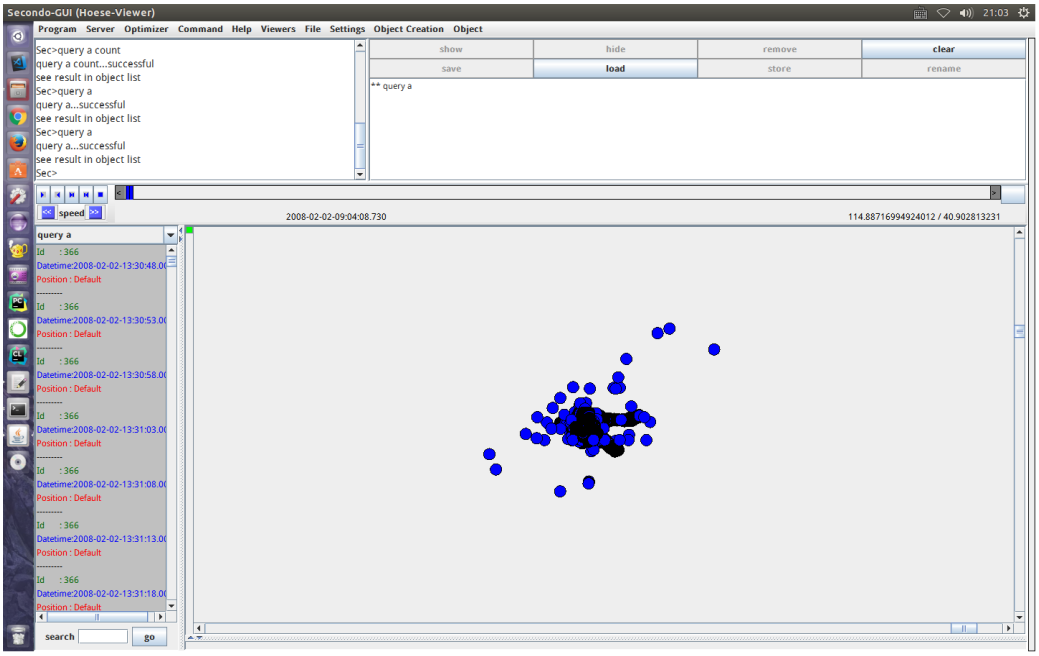


图 5.4 原始数据局部放大

图 5.5 是图 5.4 中数据集中部分的进一步放大，可以直观的看出车辆运动的历史轨迹，且轨迹数据中存在大量的异常数据信息。这些异常数据信息严重的影响原始数据的质量。对于图中颜色较深区域的数据情况是由于数据记录过于密集引起的，可能因为车辆拥堵造成的重复定位经纬度相同的问题，也可能存在设备重复存储数据造成重复数据异常的问题。对于连续轨迹上缺失的一段数据则为原始数据集中的缺失数据异常。对于轨迹周围的一些离散的数据点，有的偏离轨迹较远有的偏离轨迹较近，则为数据集中漂移或模糊数据。

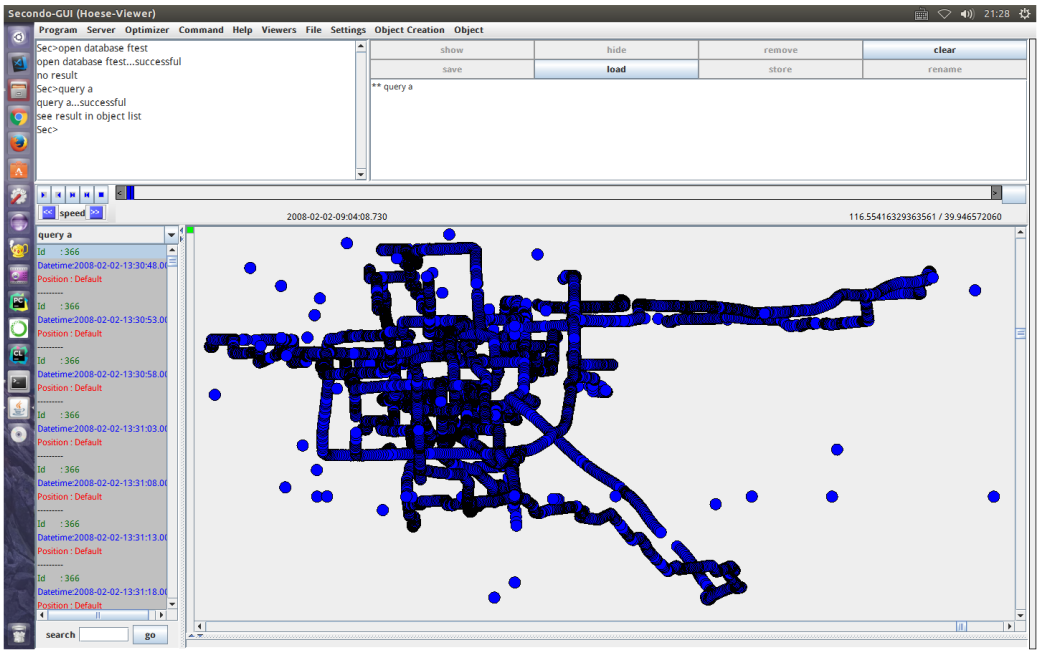


图 5.5 真实轨迹分布

将原始轨迹数据通过工具清洗之后的数据分布如图 5.6 所示，可以清晰的显示出租车的

历史运动轨迹。相比于原始轨迹的数据分布，清洗之后干净的轨迹数据没有了无效数据的干扰，则可以在界面上清晰的展示所有数据记录，将运动轨迹铺满整个界面而不用放大局部过于聚集的区域。对于缺失的轨迹数据段，从图中也可以看出工具具有很好的数据填充效果。相比于放大的原始轨迹数据中的漂移数据，对于偏离轨迹数据较远和偏离轨迹数据较近的异常数据都得到了非常好的修复。

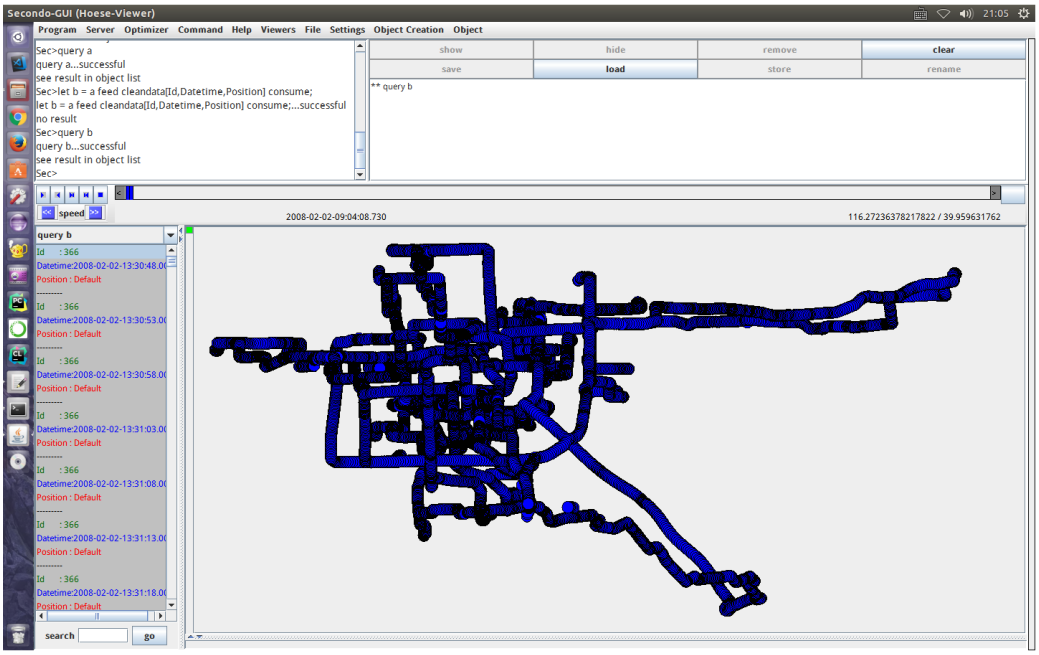


图 5.6 清洗之后的数据图

图 5.7 中展示了将清洗之后干净的数据转换成移动对象数据类型以及移动数据点的运行状态。可以通过界面上的功能按钮来对数据点的运动进行快进、暂停、倒退等操作，以更加方便分析车辆的历史运行轨迹。

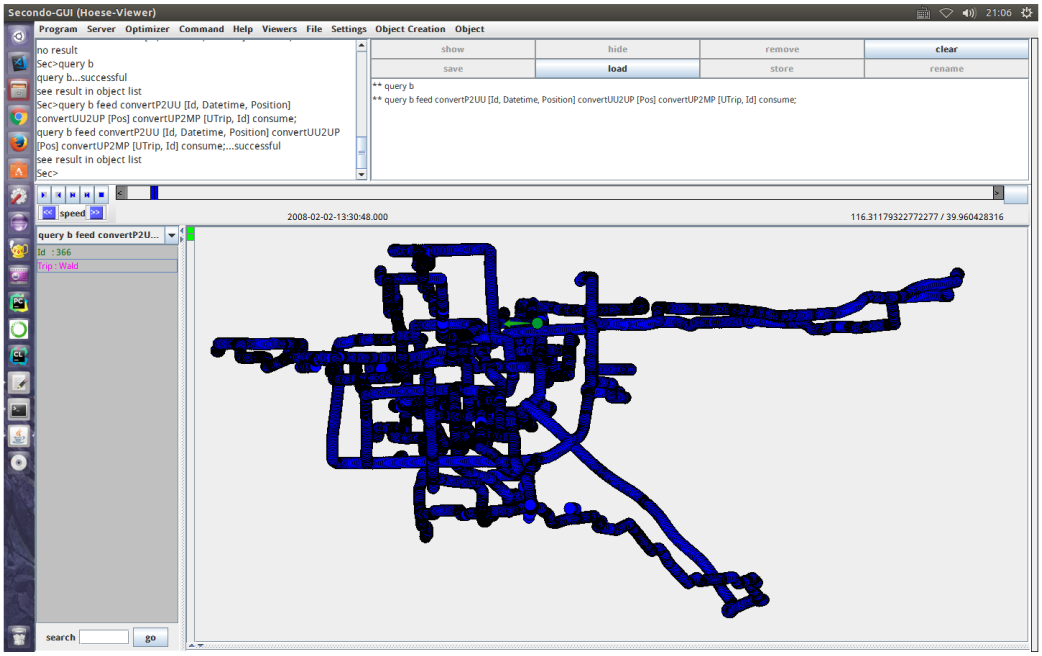


图 5.7 运动轨迹图

5.3 实验与性能测试

本章的工具是在可扩展的移动对象数据库 SECONDO 中使用 C/C++在 Ubuntu 环境下开发实现。为了验证工具清洗原始数据集中异常信息的准确性和有效性，工具通过使用真实轨迹数据集其中包含出租车轨迹数据[77][78]，卡车轨迹数据和飞机飞行数据来对工具的清洗效果进行评估。数据集的统计如表 5.1 所示。

表 5.1 数据集介绍

类型	记录数量	设备数量	平均距离	时间范围	经度范围	纬度范围	频率
出租车	1.5 千万	10357	826 km	[2008-02-02，2008-02-08]	[115.117, 117.067]	[39.067, 41.1]	10 s
卡车	4.5 千万	450	48 km	[2018-07-30, 2018-10-26]	[109.418705, 121.99981]	[22.100031, 31.9953]	1 s
飞机	0.107 千万	53224	718 km	[2008-04-06, 2008-04-07]	[-10.831388, 11.364724]	[38.253608, 51.809719]	3 min

通过使用多种不同类型的真实轨迹 GPS 数据记录进行测试，工具 GPSClean 可以很好地检测和修复重复数据、无序数据、缺失数据、无效数据、漂移数据、模糊数据。实验各部分的时间消耗如表 5.2 所示。数据预处理部分主要包括数据排序和重复数据删除操作所消耗的时间。由于在数据填充之前已经将其进行分段为多个时间连续的数据集，因此在填充缺失数据上节省了大量的时间。因各种因素影响使得原始数据集中漂移的异常数据较多，使得对于漂移数据的检测和修复所消耗的时间比较于其他异常的检测和修复时间要长。对于异常清洗之后干净的数据集，可以有选择的对数据进行类型转换。三种不同数据集检测出的异常数据信息的统计结果如

表 5.3 所示。表中分别显示各个数据集中检测到的异常点个数以及原始数据中的错误率。因车辆停止运行而造成的 GPS 定位设备获取同样的经纬度信息，可以通过将点类型转换为移动对象数据类型来剔除重复定位的数据。

表 5.2 时间消耗（单位: s）

	出租车	卡车	飞机
数据预处理	19.74	57.78	11.21
数据填充	17.66	38.32	10.18
数据修复	79.5	201.53	42.2
数据转换	177.95	533.88	71.63

表 5.3 实验统计

	出租车	卡车	飞机
重复记录	482280	42600	0
缺失记录	19350	78100	0
异常记录	91857	147556	5253
错误率	3.95%	0.59%	0.52%

表5.4显示了出租车的原始数据与清洗之后干净数据的范围统计以及单位面积数据点的分布情况。单位面积数据点的增加也客观地反映了该工具对移动对象数据具有良好的清洗效果。GPSClean 还可以将 GPS 数据记录转换为车辆历史轨迹上的移动数据点，用于模拟出租车的历史运动路径。通过观察出租车历史轨迹的运动是否平缓来进一步的判断数据中是否存在异常信息。

表 5.4 数据比较

	属性	原始数据集	清洗数据集
出租车	$\#points/km^2$	0.0013	130.8364
	lon_{min}	0	116.275
	lon_{max}	119.633	116.686
	lat_{min}	0	39.748
	lat_{max}	51.003	39.949

以上数据实验结果表明GPSClean工具可以在移动对象数据库SECONDO中很好的对原始数据集进行异常检测和修复。通过清洗前后数据集的分布可以直观的看出工具不仅能够清洗各种异常数据信息，同时在对异常数据的修复中能够遵循数据修改的最小原则，以最大的程度保证原始数据的特征。将清洗之后干净的数据导出不仅提高了原始轨迹的数据质量，对于轨迹数据分析也提供了数据基础。GPS 数据类型的转换使得可以在界面上分析车辆的历史运动轨迹，这对智慧城市的发展提供了重要的依据。

5.4 本章小结

由于各种因素的影响使得现有 GPS 定位设备采集的原始数据中通常都会存在或多或少的异常数据信息，这些异常信息严重的干扰数据质量。为了清洗原始数据集中的异常数据信息提高数据的质量，本章介绍了一个对数据集中现有的各种异常数据信息进行清洗的工具GPSClean。该工具可以很好的清洗原始轨迹数据中的多种异常数据信息，还可以将清洗前后的数据分布通过界面进行展示，直观的看出数据清洗的效果。对于清洗之后干净的数据，将其进行类型转换可以分析车辆历史的运动轨迹。以多种类型的轨迹数据对其进行实验分析可以看出 GPSClean 具有很好的异常检测和修复效果，提高了原始轨迹数据的质量。

第六章 总结与展望

6.1 本文的主要工作和贡献

近年来，随着 GPS 定位设备的普及和发展，每时每刻都会产生大量的移动对象数据。这些数据信息的来源包括手机、车辆、飞机等各种带有 GPS 定位功能的设备，这些数据来源于生活也可以很好的利用于生活。高质量的移动对象数据不仅给日常生活中的导航定位，兴趣点推荐带来了便利，而且智慧城市、无人驾驶、大数据的发展具有非常大的推动作用。但是由于设备误差、人员误操作、恶劣气候环境等的影响，使得海量的原始移动对象数据通常包括许多的错误和异常数据信息。这些异常数据严重的干扰数据的质量，给科研领域中的数据分析带来了巨大的不便。由于多种因素对数据定位的影响，造成原始数据集中存在各种类型异常信息，使得对原始移动对象数据集进行全面的异常检测和修复提高其数据质量成为了数据分析领域的巨大挑战。从海量的数据集中获取高质量的原始数据，对数据质量进行快速的评估则成了必不可少的一项任务。数据质量评估不仅可以获取高质量的原始数据还可以通过每次的质量评估来对设备定位的精度进行反馈从而提高数据采集源的准确性。除此之外，通过移动对象数据清洗工具清洗异常数据获取高质量的数据集，才能给数据分析与挖掘提供实在的数据价值。基于此，本文的主要工作和贡献有以下几点：

（1）首先介绍了移动对象数据的研究现状，根据原始数据集中的各类异常信息介绍现有异常检测和修复方法，并对提出现有方法中数据清洗的不足。然后对各领域的数据质量评估模型以及移动对象数据清洗工具及应用进行了分析和介绍。

（2）分析了现有移动对象数据中能够存在的数据异常，对于不同的异常提出了相应的检测和修复方法。对于重复和无序数据，提出了基于堆排序的最大无序偏移算法，相比于对整个数据集进行排序，该算法对于大规模的数据排序效率具有明显的提升。在数据有序的前提下，重复数据的清洗则相对比较容易。为了避免错误的识别缺失数据首先将数据进行分段，对于分段之后的数据以最近邻的数据填充方法较好的利用了缺失数据周围正确数据的变化来填充数据。对于漂移数据及模糊数据的检测和修复，为了遵循数据修改的最小原则，提出了范围约束的方法来检测和修复偏离轨迹较远的异常数据，以滑动窗口统计的方法来修复窗口内偏离轨迹较小的异常数据信息。通过不同类型异常检测和修复算法不仅清洗了数据中的大多数异常数据，还提高了数据质量。

（3）对于移动对象数据质量评估提出了基于 AHP 的多属性质量评估算法。通过分析移动对象数据特征以及各个特征下的属性，对不同的属性信息进行建模来对数据进行质量评估，以各属性之间的相互比较来提高影响权重设定的说服力。面对大规模的移动对象数据，为了提高数据质量评估的效率，提出了基于多分段的快速连续抽样方法，在保证数据质量评估稳定的前提下来提高数据质量评估的效率。

（4）在开源的可扩展的移动对象数据库 SECONDO 中实现了针对移动对象数据进行高效、准确的异常清洗工具 GPSClean。通过上传原始的数据集，以自定义的操作符对数据集中的各类异常进行检测和修复。清洗前后的数据可以通过界面展示其分布，可以直观的看出工具具有非常好的清洗效果。同时，通过将数据类型进行转换可以模拟车辆的历史运行轨迹，对于轨迹分析具有非常大的帮助。

6.2 未来的研究方向

目前对于移动对象数据异常清洗及质量评估的工作还有待提高，由于研究时间与论文篇幅等因素的限制，虽然目前已经获得了一些成果，但是仍然存在一些不足的地方需要改进。

（1）本文针对的移动对象数据异常清洗中，对于数据集中存在的异常信息的考虑还不全面，或许还会存在着一些其他类型的异常信息。移动对象数据中的其他类型数据比如共享单车也具有一定的数据价值，但是本文中没有对这些数据进行实验分析。对于漂移数据的修复，范围约束阈值的设定以及数据集中其他较多的属性信息也可以利用于数据的修复。从而可以在未来的研究中利用这些信息来进一步提高异常修复的准确率。

（2）本文针对移动对象数据质量评估进行了研究，其中在数据属性权重的设定还有待进一步的提高，由于时间限制，在后期的研究中可以结合专家对属性的评审增强说服力。对于不同的属性模型也可以进一步的改进来提高评估的准确率。

（3）本文针对的移动对象数据清洗的工具中，选用的串行的数据清洗方法，对于大规模的数据集进行清洗会造成大量时间的浪费。在未来的研究中，对于数据的清洗工作可以将其并行化的处理来提高数据清洗的效率。

参考文献

[1] Qi Y, Ishak S. A Hidden Markov Model for short term prediction of traffic conditions on freeways[J]. Transportation Research Part C: Emerging Technologies, 2014, 43: 95-111.

[2] Zheng Y. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(3): 1-41.

[3] Zheng Y, Xie X. Learning travel recommendations from user-generated GPS traces[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1): 1-29.

[4] Hill D J, Minsker B S. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach[J]. Environmental Modelling & Software, 2010, 25(9): 1014-1022.

[5] Dilling S, MacVicar B J. Cleaning high-frequency velocity profile data with autoregressive moving average (ARMA) models[J]. Flow Measurement and Instrumentation, 2017, 54: 68-81.

[6] Song S, Zhang A, Wang J, et al. SCREEN: stream data cleaning under speed constraints[C]. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015: 827-841.

[7] Keogh E, Lin J, Lee S H, et al. Finding the most unusual time series subsequence: algorithms and applications[J]. Knowledge and Information Systems, 2007, 11(1): 1-27.

[8] Corizzo R, Ceci M, Japkowicz N. Anomaly detection and repair for accurate predictions in geo-distributed big data[J]. Big Data Research, 2019, 16: 18-35.

[9] Diao Y, Liu K Y, Meng X, et al. A big data online cleaning algorithm based on dynamic outlier detection[C]. 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE, 2015: 230-234.

[10] Pang L X, Chawla S, Liu W, et al. On detection of emerging anomalous traffic patterns using GPS data[J]. Data & Knowledge Engineering, 2013, 87: 357-373.

[11] Draisbach U, Naumann F, Szott S, et al. Adaptive windows for duplicate detection[C]. 2012 IEEE 28th International Conference on Data Engineering. IEEE, 2012: 1073-1083.

[12] Basu S, Meckesheimer M. Automatic outlier detection for time series: an application to sensor data[J]. Knowledge and Information Systems, 2007, 11(2): 137-154.

[13] Brockwell P J, Brockwell P J, Davis R A, et al. Introduction to time series and forecasting[M]. springer, 2016.

[14] Fang C, Song S, Chen Z, et al. Fine-grained fuel consumption prediction[C]. Proceedings of

the 28th ACM International Conference on Information and Knowledge Management. 2019: 2783-2791.

[15] Li D, Chen D, Jin B, et al. Madgan: Multivariate anomaly detection for time series data with generative adversarial networks: 703–716[J]. 2019.

[16] Sun Y, Peng L, Li H, et al. Exploration on spatiotemporal data repairing of parking lots based on recurrent gans[C]. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 467-472.

[17] Brillinger D R. Time series: data analysis and theory[M]. Society for Industrial and Applied Mathematics, 2001.

[18] Fan W. Data quality: From theory to practice[J]. Acm Sigmod Record, 2015, 44(3): 7-18.

[19] Gupta M, Gao J, Aggarwal C, et al. Outlier detection for temporal data[J]. Synthesis Lectures on Data Mining and Knowledge Discovery, 2014, 5(1): 1-129.

[20] Marczak M, Proietti T, Grassi S. A data-cleaning augmented Kalman filter for robust estimation of state space models[J]. Econometrics and Statistics, 2018, 5: 107-123.

[21] Zhuang Y, Chen L, Wang X S, et al. A weighted moving average-based approach for cleaning sensor data[C]. 27th International Conference on Distributed Computing Systems (ICDCS'07). IEEE, 2007: 38-38.

[22] Gardner Jr E S. Exponential smoothing: The state of the art—Part II[J]. International journal of forecasting, 2006, 22(4): 637-666.

[23] Lopatenko A, Bravo L. Efficient approximation algorithms for repairing inconsistent databases[C]. 2007 IEEE 23rd international conference on data engineering. IEEE, 2007: 216-225.

[24] Golab L, Karloff H, Korn F, et al. Sequential dependencies[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 574-585.

[25] Bohannon P, Fan W, Flaster M, et al. A cost-based model and effective heuristic for repairing constraints by value modification[C]. Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005: 143-154.

[26] Yin W, Yue T, Wang H, et al. Time series cleaning under variance constraints[C]. International Conference on Database Systems for Advanced Applications. Springer, Cham, 2018: 108-113.

[27] Song S, Sun Y, Zhang A, et al. Enriching data imputation under similarity rule constraints[J]. IEEE transactions on knowledge and data engineering, 2018, 32(2): 275-287.

[28] Zhang A, Song S, Wang J. Sequential data cleaning: A statistical approach[C]. Proceedings of the 2016 International Conference on Management of Data. 2016: 909-924.

[29] Gogacz T, Toruńczyk S. Entropy bounds for conjunctive queries with functional dependencies[J]. arXiv preprint arXiv:1512.01808, 2015.

[30] Wang D, Kaplan L, Abdelzaher T F. Maximum likelihood analysis of conflicting observations in social sensing[J]. ACM Transactions on Sensor Networks, 2014, 10(2): 1-27.

[31] Yakout M, Berti-Équille L, Elmagarmid A K. Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes[C]. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013: 553-564.

[32] 武森,冯小东,单志广.基于不完备数据聚类的缺失数据填补方法[J]. 计算机学报,2012, 35(08):1726-1738.

[33] Zhou X, Liang W, Kevin I, et al. Deep-learning-enhanced human activity recognition for Internet of healthcare things[J]. IEEE Internet of Things Journal, 2020, 7(7): 6429-6438.

[34] 杨月麟,毕宗泽.基于深度学习的网络流量异常检测[J].计算机科学,2021,48(S2):540-546.

[35] 邵世宽,张宏钧,肖钦锋,等.基于无监督对抗学习的时间序列异常检测[J].南京大学学报(自然学),2021,57(06):1042-1052.

[36] Fang C, Wang X, Xu J, et al. Efficiently Detecting Light Events in Astronomical Temporal Data[C]. International Conference on Spatial Data and Intelligence. Springer, Cham, 2020: 184-197.

[37] Fan W, Geerts F, Li J, et al. Discovering conditional functional dependencies[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 23(5): 683-698.

[38] Rammelaere J, Geerts F. Revisiting conditional functional dependency discovery: Splitting the “C” from the “FD”[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018: 552-568.

[39] Arocena P C, Glavic B, Mecca G, et al. Messing up with BART: error generation for evaluating data-cleaning algorithms[J]. Proceedings of the VLDB Endowment, 2015, 9(2): 36-47.

[40] Naumann F, Herschel M. An introduction to duplicate detection[J]. Synthesis Lectures on Data Management, 2010, 2(1): 1-87.

[41] 时珉,尹瑞,胡傲宇,吴骥.基于滑动标准差计算的光伏阵列异常数据清洗办法[J].电力系统保护与控制,2020,48(06):108-114.

[42] Wang Z, Yuan X, Ye T, et al. Visual data quality analysis for taxi GPS data[C]. 2015 IEEE Conference on Visual Analytics Science and Technology. IEEE, 2015: 223-224.

[43] Jeffery S R, Garofalakis M, Franklin M J. Adaptive cleaning for RFID data streams[C]. Vldb. 2006, 6: 163-174.

[44] Baba A I, Jaeger M, Lu H, et al. Learning-based cleansing for indoor rfid data[C].

Proceedings of the 2016 International Conference on Management of Data. 2016: 925-936.

[45] Gupta A, Dhingra B. Stock market prediction using hidden markov models[C]. 2012 Students Conference on Engineering and Systems. IEEE, 2012: 1-4.

[46] Baba A I, Jaeger M, Lu H, et al. Learning-based cleansing for indoor rfid data[C]. Proceedings of the 2016 International Conference on Management of Data. 2016: 925-936.

[47] Milani M, Zheng Z, Chiang F. Currentclean: Spatio-temporal cleaning of stale data[C]. 2019 IEEE 35th International Conference on Data Engineering. IEEE, 2019: 172-183.

[48] Wang Z, Yuan X, Ye T, et al. Visual data quality analysis for taxi GPS data[C]. 2015 IEEE Conference on Visual Analytics Science and Technology. IEEE, 2015: 223-224.

[49] Lee D, Cho J, Suh Y, et al. A new window-based program for quality control of GPS sensing data[J]. Remote Sensing, 2012, 4(10): 3168-3183.

[50] Kochan B, Bellemans T, Janssens D, et al. Quality assessment of location data obtained by the GPS-enabled PARROTS survey tool[J]. Journal of Location Based Services, 2010, 4(2): 93-104.

[51] Hong M Q, Zhao W, Chen G, et al. Quality check and analysis of BeiDou and GPS observation data in the experiment of Air-Gun in reservoir[C]. 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing. IEEE, 2017: 104-107.

[52] Yoo E H, Roberts J E, Eum Y, et al. Quality of hybrid location data drawn from GPS‐enabled mobile phones: Does it matter?[J]. Transactions in GIS, 2020, 24(2): 462-482.

[53] Sadiq S, Dasu T, Dong X L, et al. Data quality: The role of empiricism[J]. ACM SIGMOD Record, 2018, 46(4): 35-43.

[54] 郑广伟,徐思达,贾国宪,郑战辉.GPS 观测数据质量评价指标分析[J].海洋测绘,2012,32(03): 37-40.

[55] 古伟洪,田鹏波,王振辉.运用 TEQC 软件对 GPS 数据的预处理与质量评定[J].地理空间信息,2008,6(06):37-39.

[56] Dasu T, Duan R, Srivastava D. Data Quality for Temporal Streams[J]. IEEE Data Eng. Bull., 2016, 39(2): 78-92.

[57] Zhang Q, Chang J, Meng G, et al. Spatio-temporal graph structure learning for traffic forecasting[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 1177-1185.

[58] Chen C, Jiao S, Zhang S, et al. TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(10): 3292-3304.

[59] Chen C, Zhang D, Ma X, et al. Crowddeliver: Planning city-wide package delivery paths

leveraging the crowd of taxis[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 18(6): 1478-1496.

[60] Funke S, Storandt S. Personalized route planning in road networks[C]. Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2015: 1-10.

[61] Shang J, Zheng Y, Tong W, et al. Inferring gas consumption and pollution emission of vehicles throughout a city[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 1027-1036.

[62] Song X, Zhang Q, Sekimoto Y, et al. Prediction and simulation of human mobility following natural disasters[J]. ACM Transactions on Intelligent Systems and Technology, 2016, 8(2): 1-23.

[63] Zheng Y, Yi X, Li M, et al. Forecasting fine-grained air quality based on big data[C]. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 2267-2276.

[64] Ding X, Wang H, Su J, et al. Cleanits: a data cleaning system for industrial time series[J]. Proceedings of the VLDB Endowment, 2019, 12(12): 1786-1789.

[65] Huang R, Chen Z, Liu Z, et al. Tsoutlier: Explaining outliers with uniform profiles over iot data[C]. 2019 IEEE International Conference on Big Data. IEEE, 2019: 2024-2027.

[66] Rong K, Bailis P. ASAP: prioritizing attention via time series smoothing[J]. arXiv preprint arXiv:1703.00983, 2017.

[67] Yu Z, Chu X. Piclean: A probabilistic and interactive data cleaning system[C]. Proceedings of the 2019 International Conference on Management of Data. 2019: 2021-2024.

[68] Rekatsinas T, Chu X, Ilyas I F, et al. Holoclean: Holistic data repairs with probabilistic inference[J]. arXiv preprint arXiv:1702.00820, 2017.

[69] Krishnan S, Wang J, Wu E, et al. Activeclean: Interactive data cleaning for statistical modeling[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 948-959.

[70] Wang J, Zhang H, Fang B, et al. Edcleaner: Data cleaning for entity information in social network[C]. ICC 2019-2019 IEEE International Conference on Communications (ICC). IEEE, 2019: 1-7.

[71] Huang Y, Milani M, Chiang F. Privacy-aware data cleaning-as-a-service[J]. Information Systems, 2020, 94: 101608.

[72] Luo L, Hou X, Cai W, et al. Incremental route inference from low-sampling GPS data: an opportunistic approach to online map matching[J]. Information Sciences, 2020, 512: 1407-1423.

[73] Cheng H Y, Yu C C. Automatic Data Cleaning System for Large-Scale Location Image Databases using a Multilevel Extractor and Multiresolution Dissimilarity Calculation[J]. IEEE Intelligent Systems, 2020.

[74] Qu Z, Wang X, Song X, et al. Location optimization for urban taxi stands based on taxi GPS trajectory big data[J]. Ieee Access, 2019, 7: 62273-62283.

[75] Li L, Chen X, Liu Q, et al. A Data-Driven Approach for GPS Trajectory Data Cleaning[C]. International Conference on Database Systems for Advanced Applications. Springer, Cham, 2020: 3-19.

[76] Güting R H, Behr T, Düntgen C. Secondo: A platform for moving objects database research and for publishing and integrating research implementations[M]. Fernuniv., Fak. für Mathematik u. Informatik, 2010.

[77] Yuan J, Zheng Y, Xie X, et al. Driving with knowledge from the physical world[C]. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 316-324.

[78] Yuan J, Zheng Y, Zhang C, et al. T-drive: driving directions based on taxi trajectories[C]. Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems. 2010: 99-108.

致 谢

时间如雨后的彩虹，缓缓悄悄的消逝于蔚蓝天际。读研的两年半时间，说长不长说短不短，直到快要临别时才会特别珍惜。在学校我结识了诸多良师益友，给我学习和生活带来了莫大的帮助。值此毕业论文完成之际，我向所有关心、爱护和帮助过我的老师、同学和亲友们表示最真挚的感谢。

首先感谢我的导师许建秋教授，感谢您在我硕士研究生期间对我学习上的指导及生活上的关心和帮助。论文的工作大部分是在与您一次又一次的交流和探讨中实践完成的，从研究方向到研究想法，您的谆谆教诲给我开启了科研生活的大门，带我攻克了科研路上一个又一个的难题。当遇到思路上的瓶颈时，您总是不厌其烦的对问题进行反复细致的分析然后给出全面的详细讲解，使得我看待和处理问题的想法更加的仔细和认真。除了学术上的指导外，您也非常关心我们的生活，告诫我们要劳逸结合，学会适当的锻炼和放松。非常荣幸能够成为您的学生，再次感谢许老师对我的科研生活中所付出的关心和帮助。

感谢陈思雨师姐、陈良建师兄、韦建华师姐、郭胜男师姐、宋苏静师姐、王宁师姐，感谢你们对我学习和生活上的指导和帮助。

感谢同门王撷阳、冒艳纯，感谢你们在我科研生活中的一路陪伴和关心，感谢你们付出的帮助和关怀。

感谢吴冰雅、李婷、丁营营、巢成、李璐、刘孟怡师妹，孙滔、刘梦男、邵磊、张旭师弟，感谢你们平日带来的欢声笑语，感谢你们让我的研究生生活多姿多彩。

感谢舍友侯夏晔、胥萌、石正璞，感谢你们创造了一个良好的寝室氛围，感谢你们对我生活上的照顾。

感谢我的女朋友张萌，感谢你对我一直以来的陪伴让我踏入了研究生的生活，感谢你的鼓励才有我现如今的成绩。

特别感谢我的家人，谢谢你们一直以来默默的付出和支持，谢谢你们在我成长路上的陪伴和照顾，谢谢你们对我的教导赋予我一颗进取的心。

最后，向百忙之中抽出时间对我的论文进行评阅和提出宝贵意见的各位专家、教授致以最诚挚的感谢！

方成龙

2021-12-31 于南京航空航天大学

在学期间的研究成果及发表的学术论文

攻读硕士学位期间学术成果及发表(录用)论文情况

1. Chenglong Fang, Feng Wang, Bin Yao, and Jianqiu Xu. GPSClean: A Framework for Cleaning and Repairing GPS Data[J]. ACM Trans. Intell. Syst. Technol. 1, 1, Article 1 (January 2021), 24 pages. (SCI 3 区，录用，第一作者)
2. Chenglong Fang, Jianqiu Xu. GPSClean: An Embedded Tool for Cleaning GPS Data[C]. 2021 22nd IEEE International Conference on Mobile Data Management. IEEE, 2021: 229-232. (CCF-C 类会议，出版，第一作者)
3. Chenglong Fang, Xieyang Wang, Jianqiu Xu, and Feng Wang. Efficiently Detecting Light Events in Astronomical Temporal Data[C]. International Conference on Spatial Data and Intelligence. Springer, Cham, 2020: 184-197. (其他全国学术会议，出版，第一作者)
4. 许建秋, 方成龙, 王撷阳. 基于 GPS 数据异常检测和修复方法. (发明专利, 以受理, 申请号: 202110365091.9)
5. 许建秋, 吴冰雅, 张森, 方成龙. 一种面向 Linux 操作系统时序数据的实时异常检测方法. (发明专利, 以受理, 申请号: 202111236814.1)
6. 方成龙, 王撷阳, 冒艳纯, 许建秋. 华东空管指标分析及报表生成工具. (软件著作, 登记号: 2020SR0265901)
7. 方成龙, 许建秋. WeSignUp. (软件著作, 登记号: 2020SR1144099)

攻读硕士学位期间参加科研项目情况和获奖情况

1. 国家自然科学基金项目（基金编号：NSFC 61972198 2019）:“多模、智能及用户友好的移动对象数据库研究”
2. 江苏省建筑科学研究院绿色建材产品数据库项目
3. 华东空管发展数据及相关指标研究项目
4. 微信小程序 WeSignUp 报名项目
5. 面向 GPS 数据的计亩软件工具项目
6. “两机”专项项目-具有专家知识辅助的进气系统设计
7. 荣获多次三好学生及科研创新先进个人等荣誉称号
8. 天池大数据《科学数据智能发现大赛(SciDI Cup)》三等奖
9. 荣获潍柴动力特别奖学金