# 目 录

[**摘要 I**](#_Toc448270580)

[**ABSTRACT III**](#_Toc448270581)

[**第1章 绪论 1**](#_Toc448270582)

[1.1 研究背景 1](#_Toc448270583)

[1.2 本文主要工作 4](#_Toc448270584)

[1.2.1 基于平面滑动平均模型的流量预测方法 4](#_Toc448270585)

[1.2.2 基于道路捆绑模型的路线推荐方法 5](#_Toc448270586)

[1.3 组织结构 5](#_Toc448270587)

[**第2章 相关工作 6**](#_Toc448270588)

[2.1 交通流预测 6](#_Toc448270589)

[2.2 路线推荐 7](#_Toc448270590)

[2.3 本章小结 8](#_Toc448270591)

[**第3章 流量预测 9**](#_Toc448270592)

[3.1 PMA模型定义 9](#_Toc448270593)

[3.2 PMA建模 10](#_Toc448270594)

[3.2.1 个体模型 10](#_Toc448270595)

[3.2.2 相似模型 12](#_Toc448270596)

[3.2.3 模型聚合 13](#_Toc448270597)

[3.3 实验及分析 13](#_Toc448270598)

[3.3.1 数据集与实验环境 13](#_Toc448270599)

[3.3.2 评价指标 14](#_Toc448270600)

[3.3.3 结果与对比 14](#_Toc448270601)

[3.4 本章小结 18](#_Toc448270603)

[**第4章 路线推荐 19**](#_Toc448270604)

[4.1 Spica模型定义 19](#_Toc448270605)

[4.2 Spica建模 20](#_Toc448270606)

[4.2.1 衔接技术分析 20](#_Toc448270607)

[4.2.2 TDJ图的构建 21](#_Toc448270608)

[4.2.3 旅行时间估计 22](#_Toc448270609)

[4.3 时间依赖启发式算法TDHA 22](#_Toc448270610)

[4.3.1 启发估价函数 23](#_Toc448270611)

[4.3.2 时间依赖的宽度优先搜索算法 24](#_Toc448270612)

[4.3.3 路线推荐 25](#_Toc448270613)

[4.4 实验及分析 26](#_Toc448270614)

[4.4.1 数据集与实验环境 26](#_Toc448270615)

[4.4.2 评价指标 26](#_Toc448270616)

[4.4.3 结果与对比 27](#_Toc448270617)

[4.5 本章小结 28](#_Toc448270619)

[**第5章 总结和展望 29**](#_Toc448270620)

[5.1 总结 29](#_Toc448270621)

[5.2 未来工作 29](#_Toc448270622)

[**参考文献 31**](#_Toc448270623)

[**致谢 34**](#_Toc448270624)

[**攻读学位期间发表的学术论文 35**](#_Toc448270625)

[**攻读学位期间参与科研项目情况 36**](#_Toc448270626)

# CONTENTS

[**Chinese Abstract I**](#_Toc445222513)

[**English Abstract III**](#_Toc445222514)

[**Chapter 1 Introduction 1**](#_Toc445222515)

[1.1 Background 1](#_Toc445222516)

[1.2 Main work 4](#_Toc445222520)

[1.2.1 Plane Moving Average Model 4](#_Toc445222521)

[1.2.2 Path Bundling Model 5](#_Toc445222522)

[1.4 Organization 5](#_Toc445222523)

[**Chapter 2 Related works 6**](#_Toc445222524)

[2.1 Traffic flow prediction 6](#_Toc445222525)

[2.2 Route recommendation 7](#_Toc445222526)

[2.3 Summary 8](#_Toc445222525)

[**Chapter 3 Traffic flow prediction 9**](#_Toc445222528)

[3.1 PMA definition 9](#_Toc445222529)

[3.2 PMA modeling 10](#_Toc445222529)

[3.2.1 Individual model 10](#_Toc445222530)

[3.2.2 Similarity model 12](#_Toc445222531)

[3.2.3 Integration 13](#_Toc445222532)

[3.3 Experiment and analysis 13](#_Toc445222529)

[3.3.1 Dataset and environment 13](#_Toc445222532)

[3.3.2 Goodness-of-fit statistics 14](#_Toc445222532)

[3.3.3 Evaluation 14](#_Toc445222532)

[3.6 Summary 18](#_Toc445222541)

[**Chapter 4 Route recommendation 19**](#_Toc445222542)

[4.1 Spica definition 19](#_Toc445222543)

[4.2 Spica modeling 20](#_Toc445222533)

[4.2.1 Motivation of joint graph 20](#_Toc445222534)

[4.2.2 Building of TDJ 21](#_Toc445222535)

[4.2.3 Travel time estimation 22](#_Toc445222536)

[4.3 Time-dependent heuristic algorithm 22](#_Toc445222537)

[4.3.1 Heuristic evaluation function 23](#_Toc445222538)

[4.3.2 Time-dependent breadth-first search algorithm 24](#_Toc445222539)

[4.3.3 Route generation 25](#_Toc445222540)

[4.4Experiment and analysis 26](#_Toc445222529)

[4.4.1 Dataset and environment 26](#_Toc445222532)

[4.4.2 Goodness-of-fit statistics 26](#_Toc445222532)

[4.4.3 Evaluation 26](#_Toc445222532)

[4.5 Summary 28](#_Toc445222541)

[**Chapter 5 Conclusion and prospect 29**](#_Toc445222554)

[5.1 Conclusion 29](#_Toc445222555)

[5.2 Prospect 29](#_Toc445222556)

[**References 31**](#_Toc445222557)

[**Acknowledgement 34**](#_Toc445222558)

[**Publication 35**](#_Toc445222559)

[**Research projects 36**](#_Toc445222560)

# 摘 要

自20世纪末以来，由于交通路网的日益复杂、汽车等交通工具数量的快速增长，交通流量的变化越来越影响人们的生活，交通流量的不合理疏导甚至会导致交通堵塞和交通事故等问题。短期交通流量的预测对于缓解城市交通问题，提高城市交通运输效率和建设智慧城市都有着积极的作用。另外，个人驾驶轨迹信息，特别是运营类的司机行驶轨迹，其中隐含了影响效率和司机真实行为的特征，在提高效率和反作弊的需求下，司机效率影响因素的挖掘和轨迹行为的发现有着不可忽略的作用。

短时交通流量预测方面，主要有模拟仿真、回归分析、神经网络等类型的方法。模拟仿真的方法需要对复杂的交通网络建立模型，需要很高的先验知识支持，并且计算代价较大，因此在实际中应用比较困难；回归分析方法主要有自回归模型（AR）、滑动平均模型（MA）和自回归滑动平均（ARIMA）模型等；包括深度学习网络模型中的栈式自动编码机（SAE）和深度置信网络（DBN）在内的神经网络方法也都运用在了交通流量预测任务中，深度学习方法的主要优点是可以利用多层的网络结构优化特征空间，从而使得新的特征更具表现力。如何获取更有意义的特征值成为了目前研究的重点。这些方法一般是依据历史交通流量数据来预测接下来一个时间段内的交通流量，并且大多只针对有少量数据采集点的交通流量进行预测，没有在整个城市区域的角度进行预测。目前广泛应用的智能交通系统（ITS）中，大量的二维城市时空交通数据已经可以获取，并且数据量正在以惊人的速度增长，这就使得整个城市区域的交通流量预测有了更大规模的数据支持，而更大规模数据的产生也使我们面临着更多的挑战。随着计算机硬件计算能力的提高，深度学习的方法已经开始得到了越来越多的探索和应用，深度学习方法的高维特征抽取能力对于预测任务有极其重要的意义。我们提出了一种基于数据重组和卷积神经网络的交通流量预测方法，主要特点是：1. 通过车辆轨迹数据挖掘路口车流量间的影响因子；2. 将路口间车流量的影响因子应用到构建针对短时交通流量的特征矩阵中，即构建拥有时间-空间二维的特征矩阵；3. 将卷积神经网络用于预测交通流量，利用其局部感知原理发挥空间因素对短时流量的影响。

司机行为研究方面，我们主要针对的是司机轨迹中的停留点意义和轨迹切割问题。

我们将提出的模型应用到实际的交通数据集中，进行了充分的实验，结果证明了模型的有效性和先进性，在提到的评价指标中已经由于原有模型，具有在实际中应用的价值，并可以在相关领域内迁移应用。

**关键字：**深度学习；时空数据；短时流量预测；轨迹行为分析

# ABSTRACT

Since the end of the 20th century, due to the increasingly complex road network, and the rapid growth in the number of vehicles, traffic flow changes more and more people's lives. traffic flow unreasonable grooming and even lead to traffic congestion and traffic accidents and other issues The forecast of short-term traffic flow has a positive effect on alleviating urban traffic problems, improving urban transport efficiency and building intelligent cities. In addition, the personal driving trajectory information, especially the operating class of the driver's trajectory, which implied the impact of efficiency and the characteristics of the driver's real behavior, in improving the efficiency and anti-cheating demand, the driver efficiency factors of excavation and trajectory behavior discovery Has a role that can not be ignored.

Short-term traffic flow forecast, there are simulation, regression analysis, neural network and other types of methods. The simulation method requires the establishment of a complex traffic network model, requires a high priori knowledge support, and the calculation cost is large, so it is difficult to apply in practice. The regression analysis method mainly includes autoregressive model (AR), sliding average Model (MA) and autoregressive moving average (ARIMA) model. Neural network methods including stack type automatic coding machine (SAE) and deep confidence network (DBN) in depth learning network model are also used in traffic flow In the prediction task, the main advantage of the depth learning method is that it can use the multi-layer network structure to optimize the feature space, which makes the new features more expressive. How to obtain more meaningful eigenvalues ​​has become the focus of the current research. These methods are generally based on historical traffic flow data to predict traffic flow for the next time period, and most of them only for a small amount of data collection point of traffic flow forecast, not in the entire urban area perspective. At present, the widely used intelligent transportation system (ITS), a large number of two-dimensional urban space-time traffic data has been available, and the amount of data is growing at an alarming rate, which makes the entire urban area traffic flow forecast with a larger scale Data support, and the emergence of larger-scale data also makes us face more challenges. With the improvement of computer hardware computing ability, the method of deep learning has begun to get more and more exploration and application. The high-dimensional feature extraction ability of depth learning method is of great significance to the forecasting task. We propose a traffic flow forecasting method based on data reorganization and convolution neural network. The main features are as follows: 1. Exploring the influencing factors of traffic flow between vehicles by vehicle trajectory data; 2. Applying the influencing factors of traffic flow between roads In the feature matrix of short-term traffic flow, we construct a feature matrix with time-space two-dimensional. 3. Convolution neural network is used to predict traffic flow, and its local perception principle is used to exert the influence of spatial factors on short-term flow The

We have applied the model to the actual traffic data set, carried out the full experiment, the result proves the validity and the advancement of the model, in the mentioned evaluation index has been due to the original model, has the practical application value , And can migrate applications in related areas.

**Keyword:**  deep learning; temporal-spatio data; short term traffic forecasting;

# 绪论

本章首先介绍短时交通流量预测和司机行为分析的研究背景及意义；其次对该领域近些年的相关研究工作进行说明；然后介绍本文在以上两个课题中的贡献；最后描述了整篇文章的结构。

## 1.1 研究背景

在快速发展的社会背景下，交通问题是一个极为重要的现实问题。而交通数据的大量增长为解决此类问题提供了新的解决途径。

**交通流量预测** 城市交通流量预测是智能交通系统中非常重要的一环，影响着城市路网的结构和规划。为了建设智慧城市，实时地准确预测交通流量是最现实的需求。在现代城市中，交通流的变化对人们的日常生活有着极为重要的影响，例如出行路线的选择等。长时交通流量的预测通常是指每月甚至每年的交通流量预测。由于车辆的动态变化，长时交通流量的预测在实际中无法很好得应用。因此短时的交通流量预测对我们会更有用，例如，我们在驾车出行时想知道未来十分钟前方各个路口的流量情况来决定行驶路线。

大量的模型技术已经应用到了短期交通流量预测上。包括滑动平均模型，k近邻模型，自回归模型，周期性的ARIMA模型和神经网络模型等。预测短期交通流量往往很大程度上依赖于历史和实时的交通数据，这些数据可以从各种传感器，例如线圈，雷达，GPS以及多媒体数据中获取。

**路线推荐** 没有良好的路线推荐将带来城市交通拥挤问题。城市交通拥挤不仅将导致社会经济多项功能的衰退，延缓城市发展的步伐，还会引发城市居民生活环境的持续恶化，成为阻碍城市现代化建设的“城市顽疾”。交通拥挤对社会生活最直接的影响是增加了居民的出行时间和成本，而出行成本的增加不仅直接影响工作效率，使城市活力大打折扣，还会抑制人们的日常活动，降低居民的生活质量和幸福指数。另外，交通拥挤还将增大交通事故发生的概率，而交通事故的增多又反过来加剧了交通的拥挤，因此形成了一种恶性循环。据相关统计，21世纪以来，欧洲每年因交通事故造成的经济损失就高达500亿美元之多。此外，交通拥挤还破坏了城市自然生态环境。在机动车迅速增长的今天，交通拥挤对自然环境的污染也在不断增加，并且在逐步成为城市空气质量恶化的主要根源。因为在交通拥挤的情况下，车辆只能在低速状态行驶，而这种频繁地在停车和启动之间切换，不仅增加了汽车的能源消耗还增加了汽车尾气排放量。根据英国伦敦20世纪90年代的检测报告，大气中74%的氮氧化物来自汽车尾气排放，而且这个百分比还在日益增加。可见路线推荐研究对城市的健康发展至关重要。路线推荐指的是给定始发点和目的地，利用已有的交通轨迹数据推荐合理的出行路线，合理性可以从时间与空间两个维度来衡量，最佳的推荐路线应该是时间越短且路程最近。路线推荐和我们每一个人的出行息息相关，诸如高德地图、百度地图等各大地图类软件都在致力于提供合理的路线推荐，以节省人们的出行成本，提高生活质量。总的来说，车辆行驶过一条道路的时间长短主要由三个因素决定。其一，道路的本身情况，包括路面状况、长度、分道数量以及红路灯的情况；其二，不同的时间段意味着不同的车辆密度，也决定了通过的时间；其三，司机的驾驶经验与技术，通常驾驶经验丰富的司机比初次上路的新手表现更好，这样的经验包含了对当前道路的熟悉程度。因此，对于高效的路线推荐模型来说，以上三种因素缺一不可。目前较常用行的路线推荐方法包括A\*算法、Dijkstra算法、贝尔曼-福德（Bellman-Ford）算法等。

本文工作是针对以上两个核心热点研究课题的现存方法所存在的不足展开的。针对交通流量预测问题我们发现，单源流量预测的解决方法虽然已经比较成熟，但是针对一个城市的交通道路网络来说，该类方法并未考虑多源数据之间的关联；不仅如此，像这样的预测往往只是在类似高速公路这样流量变化趋于平缓的数据源上才能取得不错的结果。而针对路线推荐问题，我们发现现存的模型是将道路通行时间独立出来进行建模，并未考虑相连接的道路之间的通行关系；而且往往这样的模型都是基于出租车GPS采集的数据进行的路线推荐，这样的单一车辆数据集容易产生数据倾斜，并不能代表所有车辆的行驶情况，所以其无法反映城市道路的真实路况。因此，对于这两方面的问题，我们将在后文分别给出详细的解决方法和建模策略，来弥补上述两类方法存在的缺陷。

## 1.2 本文主要工作

本文的主要工作分为两个方面，一是提出一种基于平面滑动平均模型的流量预测方法，二是提出一种基于道路捆绑模型的路线推荐方法。这两种方法从全新的角度剖析流量预测以及路线推荐两个热点问题，取得了一定的进展。本小节接下来的内容将对这两种全新的方法分别阐述其做出的贡献。

### 1.2.1 基于平面滑动平均模型的流量预测方法

我通过分析过去的流量预测方法我们发现了以下几个问题：其一，这些方法用的数据集通常采自高速公路的车辆监控系统，这样的数据集通常趋于平稳，因此预测的结果理所应当不会很差；其二，这些方法的数据集通常只采集一个数据点（如道路监控视像头）的信息，而单源数据集的结果很难运用于更加复杂的交通道路网；其三，现存的方法并未考虑多个数据采集点之间可能存在的联系。比如，毫不相关的交通路口之间的流量可能呈现相同的变化规律。如果能够将这些潜在的关联挖掘出来，将对交通流量预测结果的准确性带来较大的提升。

因此，我们提出一种基于平面滑动平均模型PMA（plane moving average）的流量预测方法。该方法的核心思想是：利用最相似的历史交通流量模式来进行预测。PMA是一种简单高效的模型，它建立在两个子模型之上：个体模型和相似模型。个体模型融合预测对象本身的历史数据特征，而相似模型则融合与预测对象呈现高相似度的其他对象的数据特征，这种相似度我们用一种叫做WWL（where we like）的算法来评估，两种模型分别得到一个流量预测结果。然后我们将两种结果用神经网络融合在一起就得到了PMA模型的最终交通流量预测结果。

### 1.2.2 基于道路捆绑模型的路线推荐方法

在针对路线推荐方面现存方法的不足，我们提出了一种基于道路捆绑模型的路线推荐方法Spica。Spica从一种全新的视角来构建路线推荐模型，其核心思想是利用一种衔接技术JT（joint technique）从旅行时间维度描述相连道路关系，然后建立路线推荐模型进行预测。具体步骤为：首先，使用JT建立一种经过改进的TDN，我们称之为TDJ（time-dependent joint）图；然后对TDJ图中的每一条边进行时间依赖的旅行时间预估；最后基于TDJ图以及每条边的旅行时间分布情况，使用一种时间依赖启发式算法TDHA（time-dependent heuristic algorithm）计算图中的最短路径，得出合理的路线推荐方案。

## 1.3 组织结构

本文共包含五个章节，每个章节的具体内容组织如下：

第一章主要介绍了流量预测和路线推荐两个交通领域问题的研究背景，并对本文的主要工作贡献进行了简单介绍。

第二章对这两个交通热门领域问题的相关工作进行回顾，并说明了本论文的实验环境以及模型数据集情况。阐述了我们提出的模型的，分别对平面滑动平均模型与道路捆绑模型将要用到的，为下一章的模型实现做好准备。

第三、四章分别对平面滑动平均模型与道路捆绑模型的实现细节进行了全面详细的阐述。包括各种概念及参数的定义描述、公式描述、图形描述、代码描述、数据预处理、数据训练方法以及模型具体的工作机制。然后分别介绍了实验的评价标准以及作为对比的现存方法。最后，通过在真实数据集上进行大量的实验，验证了我们提出的这两种模型得到的结果优于其他的对比模型。

第五章对本文提出的两种模型进行总结以及对该方向未来发展的展望。

# 相关工作

本章节将分别阐述交通流预测和线路推荐的两个交通热门领域的研究现状以及实验环境和所用数据集情况。交通流预测方面，由于短时流量预测对解决交通问题更为重要，因此本章节将简单介绍长时交通流预测，详细介绍短时交通流预测；路线推荐方面，将从最短路径推荐、最快路径推荐以及基于出租车GPS数据集进行研究的相关路线推荐方法进行简单介绍。

## 2.1 交通流预测

在二十世纪前期，交通流量预测的作用主要是为交通控制系统服务。第一代城市交通控制系统（Urban Traffic Control Systems，UTCS）使用历史的交通流数据进行离线的流量预测；第二代的UTCS通过实际测量的数据对历史平均的数据进行了补充修正，预测精度获得了一定程度的提高；而第三代的UTCS则全部使用实测数据来进行预测。UTCS这套系统所使用的算法都存在时滞问题[1]。此后，不少研究者以及科研人员投入了相当之多的精力积极地探索着交通流量预测的其他方法。1974年谱分析法的运用在交通流预测上取得了比较满意的结果，但是该方法无法应对突发流量对预测结果所带来的影响。1979年研究人员利用Box-Jerkins技术来预测高速公路上的流量，但结果表明该方法的精度有限；之后的一年他们又将Box-Jerkins技术与一路段时长4年的交通流量数据相结合对该路段进行了流量预测，精度较高但是其要求的历史数据较多，因此也并非是一个很好的方法。1984年研究人员采用卡尔曼滤波理论构建了一个流量预测模型，预测精度优于第二代的UTCS方法。1991年研究人员应用非参数回归模型进行流量预测，在某些特定的情况下比时间序列模型的预测精度要高。1993年神经网络的运用，使得交通流预测的结果大幅提升同时还解决了时滞问题。1996年随着混合模型的加入，神经网络与传统的ARIMA[27][28]时间序列模型相结合使得ARIMA模型具有更加广泛的适用性和可移植性。

过去二十多年时间里，交通拥堵问题日趋严重的前提下，各大国际期刊和会议上涌现了大量的流量预测相关研究文献，提出了许多行之有效的预测方法，这些方法可以大致归为两大类：长时流量预测和短时流量预测。文献[13][14]通过分析宏观上的流量时间序列的数学特征来研究长时流量预测问题，探索了交通网络的自身特性，提出了针对月和年时间维度的流量预测方法论，并在实际运用当中取得了不错的效果。短时流量预测在文献[21-28]中被广泛研究，这些文献的作者只关心相邻的最近时段的流量值，提出的线性模型方法均是针对下一时段的流量进行预测。其中的一些方法只用了单独的预测模型，比如文献[27][28]；而另外一些方法则运用了混合模型[22][23][26]，取长补短。文献[26]首次将滑动平均法，指数平滑法以及ARIMA模型三者利用神经网络进行预测结果的融合，建立了一个聚合模型DA（aggregation model），其预测结果较之以往的单一模型取得了较大的提升。

在提高交通流量预测的准确度方面，除了流量预测模型本身地不断发展演变，数据的预处理技术也至关重要。通过去除数据噪点，降低脏数据的比例，良好的数据处理技术能很大程度上提高流量预测的准确性[20][29][30]。文献[29][30]提出了一种简便但是高效的交通流数据预处理方法对BP神经网络的训练数据进行降噪处理，使得最终的预测精度有了较大的提升。将数据集分解成突发（burst data）数据集和非突发（non-burst）数据集来分别进行预测，文献[20]在预测结果的精度上取得了进一步的突破。这些方法充分说明了经过预处理的数据集能带来更加准确的预测结果。

## 2.2 路线推荐

在城市交通道路的推荐研究中求解最优路线问题，与求静态路网中两点之间的最短路径不同，该路径通常具有时间依赖性。由于科学技术的迅猛发展，我们已经可以较为容易的获取城市道路车辆的位置以及其行驶速度，进而建立起一种以时间为自变量函数的时间依赖网络TDN（time-dependent networks）。谭国真等学者对TDN给出了较为详细的介绍[2]，将其划分为FIFO（first in first out）和非FIFO网络两大类，并介绍了FIFO弧的概念。FIFO弧的代价变化满足条件∀s > 0，∀t，Cost(i，j，t)≤ Cost(i，j，t + s) + s，其中s与t表示时间，i和j表示FIFO弧连接的两个节点，Cost(i，j，t)表示弧ij在t时刻的时间消耗代价。非FIFO弧则不满足上述性质，网络中只要存在一条非FIFO弧就是非FIFO网络。文献[3]直接用Dijkstra算法求出了FIFO网络的最小时间路径，但是对非FIFO网络却无能为力。针对这个问题，文献[4]针对非FIFO网络的特征加了约束条件，使约束后的网络成为了一种FIFO网络。之后关于FIFO网络的研究中，文献[5]通过提出几种非FIFO网络模型，解决了可等待模型中的最短路问题。而文献[2]则提出一种SPT-DN算法解决了非FIFO-TDN在不可等待情况下的最短路径问题。

针对弧代价的预测与计算是路径规划问题中至关重要的一部分。在允许等待的情形下，文献[6]将非FIFO网络中的非FIFO弧通过数学变换转化成了FIFO弧。文献[7]对[6]中未考虑的特殊情况作了补充。文献[8]将交叉口转弯系数与延误系数引入改进的A\*算法中进行计算。在信息熵理论的基础之上文献[9]通过计算路径的综合属性值在多属性随机TDN中选取最优路径。文献[10]基于短时路况不变的前提下提出了一种代价计算方法。文献[12]提出了一种运用于城市交通网络联合运输的最优路径算法。在TDN中文献[13]使用了一种基于实时数据的启发式算法，这种方法能够根据路况的实时变化做出相应的路由选择。总之，以上这些路线推荐的方法都能在某些特定情况下一定程度地解决路线规划的问题，但是其前提条件是建立在弧代价的计算总能通过某种方法而准确得到的基础之上的。然而实际情况中对代价预测的准确性却不尽如人意，其往往会随着时间的推移而大打折扣，以至于很难获得较为理想的推荐路线结果。对此，本文提出的道路捆绑模型不受代价预测的影响，有效的避免了该问题的影响。

## 2.3 本章小结

本章节详细描述了交通流预测和线路推荐问题的相关工作情况以及实验环境的相关情况，为后文对模型进行展开描述做了铺垫

# 流量预测

本章对我们提出的滑动平均模型进行描述。首先对模型用到的概念进行定义；然后详细描述模型建模、模型训练以及模型的预测方法；接下来介绍模型的实验情况，包括数据集、实验环境和实验结果；最后与若干流行模型在相同数据集上进行预测结果的比较，说明该模型的优越性。

## 3.1 PMA模型定义

在具体阐述平面滑动平均模型PMA的实现机制之前，本小节先对该模型将要用到的一些概念进行定义与说明，在这之中将用到一些数学表达式。我们将给出什么是交通流和流量模式的定义，为后文描述模型的实现细节打下基础。针对交通流量的预测问题，结合PMA模型要求，我们需要定义两个方面的问题：其一，交通流的定义以及具体形式；其二，时间依赖的流量模式。二者的具体定义如下：

**定义1(交通流)** 交通流量数据是随时间收集的一种时间序列数据，每一个具体的流量数据值对应的是在一段特定时间内通过的所有车辆的总数。我们用Δt分钟来表示这段特定时间的长度，时刻t的流量数据值q(t)表示在时间间隔(t-Δt,t]内通过的全部车辆数量，其中Δt，t，q(t)均是非负整数。

**定义2(流量模式)**不难发现q(t)的具有周期性的变化规律，比如呈现出天之间相似，周之间相似等。因此，我们定义两种流量模式如下，其中代表预测值：

***Ⅰ.***sa(t)代表t时刻的前ka个时段的流量模式。

sa(t)={q(t-kaΔt),q(t-(ka-1)Δt),……,q(t-Δt),}

***Ⅱ.***Sb(t)代表同一t时刻的前kb天的流量模式。

Sb(t) = {q(t-1440kb), q(t-1440(kb-1)), ……, q(t-1440), }

本小节通过数据层面和建模层面，从数学角度定义了平面滑动平均模型PMA中的交通流、流量模式，为下文对模型的具体描述打下了基础。

## 3.2 PMA建模

平面滑动平均模型PMA是滑动平均模型MA的一种改进模型，它使用一种称之为平面流量模式PFP（plane flow pattern）的二维数据结构来描述其算法实现的过程。

为了将交通数据流个体本身与相似的交通数据流相结合进行预测，我们使用个体历史流量数据建立一种个体模型和相似流量数据建立一种相似模型，二者通过神经网络进行聚合得到最终的流量预测值。这里需要注意的地方是，这两个子模型都运用了平面滑动平均算法PMA。接下来我们将分别阐述个体模型和相似模型的定义，PMA的训练方法以及最后模型的聚合。

### 3.2.1 个体模型

PFP中，我们定义sa(t)和sb (t)分别为水平流量模式与垂直流量模式，提出一种建立PFP的方法。如图3-1所示，图中展示了sa(t)、sb(t)、a个Δt时间间隔之前的sb (t-kaΔt)、kb天之前的sa (t-1440ka)等构成的平面流量模式PFP，可知图中相邻的两个流量点要么相隔Δt分钟，要么相隔1天的时间。不难发现kaΔt有可能等于1440，这是由于一个天级流量模式将在PFP中被使用多次。

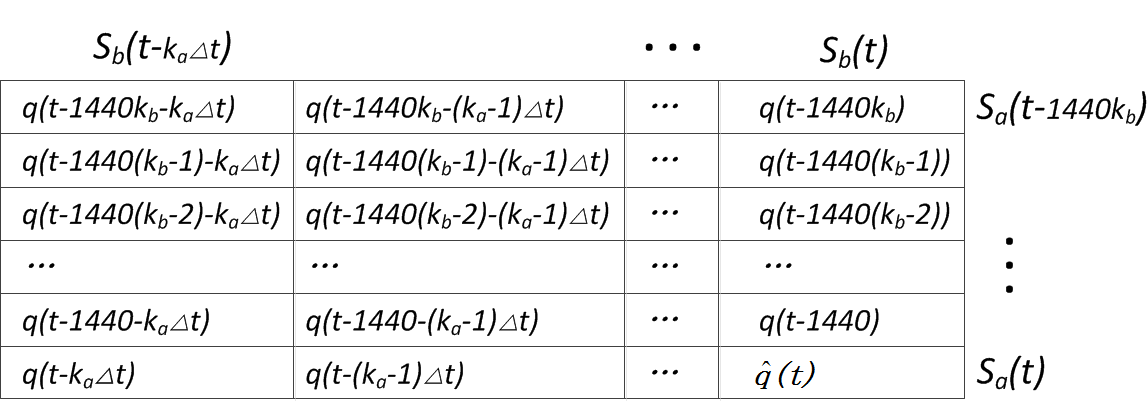


图3-1 个体模型PFP，其中代表预测值。

PMA用来产出一个PFP的预测结果。为了方便起见，我们用y(i,j)表示PFP中第i行和第j列的真实流量值，表示PMA得出的流量预测值。为了描述PMA的过程，我们先给出两个计算公式如下：

**模式权重** 每一个流量模式都能对预测准确度提供帮助，我们定义一种模式权重ρ来描述每一个流量模式对目标预测值的贡献程度。

(1)

其中L是PFP的水平长度，相邻间隔时长为Δt分钟（见图3-2）。δ是列数，m代表目标天数。

**PMA预测值**PMA的计算方法与MA滑动平均的计算方法相似，只是从一维的情况扩展到了二维的情况。

(2)

其中W是PFP的垂直高度（见图3-2）。每一个PMA的预测值都对应一个Li×Wi大小的PFP。

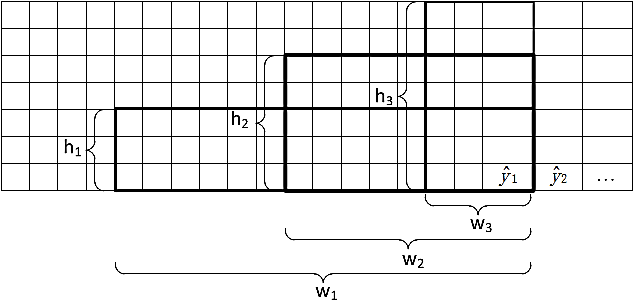


图3-2 在PMA训练过程中不同大小的PFP，其中代表预测值。

例如，PFP包含了3种sa流量模式和4种sb流量模式，4是这个PFP的目标预测值。如果我们只是单独用流量模式{1,2,3}和{10, 8}来做历史平均预测，我们将得到预测值2和9，这两个结果都与目标预测值4相去甚远。使用模式权重我们可以得到目标流量模式的前两个sa流量模式的权重分别为0.478和0.639，然后可以如此计算最后的预测值：(10 × 0.478 + 8 × 0.639) / 2 = 4.946。不难看出4.946与目标预测值更为接近。

在PMA训练方法中，决定PFP的最优大小是核心任务。算法1中详细的描述了PMA算法如何选择最适合的具有较强泛化能力的PFP大小。我们将所用的数据集按时间维度划分成了训练集和测试集。

表3-1 PMA算法实现

|  |
| --- |
| **算法1：**PMA实现代码 |
| 1: Firstly, for the PFP, work out the pattern weight ρ by (1);  2: <X, Y> is the size of the training set;  3: for L = 1…X do  4: for W = 1…Y do  5: To each ⟨L,W⟩, compute PMA forecasting value of all training samples by (2);  6: Calculate the average RMSE between the each pair of forecasting value and the actualvalue;  7: Update the optimal PFP size according to the minimum RMSE;  8: end for  9: end for  10: Use the optimal PFP size to make predictions. |

### 3.2.2 相似模型

通常情况下时间序列在相同场景类别下展现出相似的特征，比如同一个区域。一个每小时通过1000辆车的路口是不可能出现在偏僻的山村，而只能出现在城市的主干道之上。因此，针对这样的相似性，我们提出一种由欧拉距离和皮尔森相关系数共同决定的相似度计算的WWL算法（where we like）。通过该算法选出与目标流量模式最相似的n个流量模式来进行预测。由于不能加入未来因素进行预测，这里我们规定所选取的相似流量模式与目标流量模式之间的时间距离至少为1天。

WWL算法首先，利用皮尔森相关系数公式计算目标流量模式与其他流量模式之间的相关性，根据皮尔森相关系数的定义，我们筛选pcc的值大于0.8的高相似度流量模式数据。然后，用欧拉公式计算筛选出来的流量模式与目标模式的欧拉距离以保证最后得到的相似流量模式不仅要在形状上与目标流量模式相似，在数值大小上也要与之相近。我们将最后筛选出来的相似流量模式排序后与目标流量模式一起得到如下式子：

(3)

其中s(t)表示目标流量模式，si(t)代表与目标流量模式第i相似的流量模式，d是WWL算法的参数。

直观上看，流量模式越相似，对目标流量模式的预测帮助就越大。因此我们将筛选出来的相似流量模式按相似度的高低从下至上排列成像PFPindividual(t)一样，同样用PMA算法来计算对应PFP预测值，作为相似模型的预测结果。

### 3.2.3 模型聚合

选择合适的模型来合并个体模型和相似模型的预测结果也是一项很重要的工作。目前有很多流行的方法能做模型的聚合，比如线性回归、逻辑回归和神经网络等。为了简单高效起见，我们选取神经网络来做以上两种模型的融合。

通过上述两种模型的计算，我们能分别得到一个预测值的结果序列，，其中ω= 1表示个体模型的预测结果序列，ω= 2表示相似模型的预测结果序列，m是预测的时间下标值，代表第i个预测值。同时为了验证，我们也有一组对应的真实流量值作为预测目标值，从而得到一个神经网络的数据集(P1, P2, R)。

我们分别选取双曲正切函数tansig和线性函数purelin作为隐层和输出层的激活函数。学习方法我们选用了BP反向传播算法，因为BP算法是最流行的算法且具有较快的收敛速度。

(4)

(5)

其中n代表隐层神经元的个数，hi代表隐层的输出，net则是神经网络的输出。在训练集大小的选取上虽然目前没有精确的规则进行判断，我们建议训练集尽量不要太小。

## 3.3 实验及分析

### 3.3.1 数据集与实验环境

实验用的数据集取自某大城市交通治安卡口系统所采集的车辆行驶数据[15]，数据采集时间是2013年4月1日到2013年5月5日。本数据集一共记录了140,440,933条真实的车辆行驶信息，采集的数据一共覆盖了全市308个监控摄像头。我们将这些数据按时间维度划分成了一个训练集和一个测试集：将4月1日到5月2日的数据划为训练集，5月3日到5月5日的数据划为测试集。实验环境用的是戴尔OPTIPLEX990型号的台式机，因特尔i5-2400处理器，CPU主频3.1GHz，内存大小为4GB。

然后我们对数据集进行预处理操作。通过计算每个Δt时间间隔内的车流量值，训练集和测试集加起来我们一共得到了10780种sa流量模式和7392种sb流量模式（详见**3.2.1小节**）。

### 3.3.2 评价指标

我们使用两个流行的评价指标来对比评估PMA模型的预测准确性。这两个指标的描述如下：

1. 均方根误差**RMSE（Root Mean Squared Error）**是衡量预测值平均误差的一种流行方法。

6)

1. 绝对误差百分比**MPAE（Mean Percentage Absolute Error）**是衡量比例误差最流行的方法。

(7)

这里的和分别表示真实流量值与模型预测值，N表示对比的真实流量值的总数量。

### 3.3.3 结果与对比

由于取到的原始数据集中的每一条数据是一个四元组：（时间戳，卡口编号，车牌号，行驶方向），因此需要对数据做预处理才能够适合PMA模型的要求。在PMA模型中，我们用Δt表示最小时间单位，将数据集中的时间戳落在[t, t+Δt)区间的车辆进行加和得到交通流量模式，根据文献[18]中的建议，我们将PMA中的时间单位Δt设为60分钟，我们用计算前后流量值平均值得方法解决缺失流量值的问题。

我们通过评估个体模型、相似模型以及它们聚合后的模型来评价PMA模型的预测效果。对于每个评估实验，我们对数据集中所有卡口点的流量数据进行预测，并取他们的平均值作为实验对比的依据。

首先，我们循环个体模型与相似模型中参数对〈L, W〉（在公式(1)和(2)中提出）的值。由于均方根误差RMSE能够反应交通流量值之间的差异程度，因此我们用该误差值来帮助模型训练出误差最小的参数对〈L, W〉。图3-3(a)和3-3(b)展示了个体模型和相似模型在不同参数对〈L, W〉数值下的RMSE大小，每一个RMSE的计算涵盖了一天的24个数据点（Δt为60分钟），然后取24个RMSE值的平均值作为这个训练数据点当前的RMSE值。从图3-3中我们可以看出RMSE最小值的位置，从而得出个体模型取L = 6和W = 2、相似模型取L = 7和W = 2作为这两个子模型PFP的大小为最优选择。

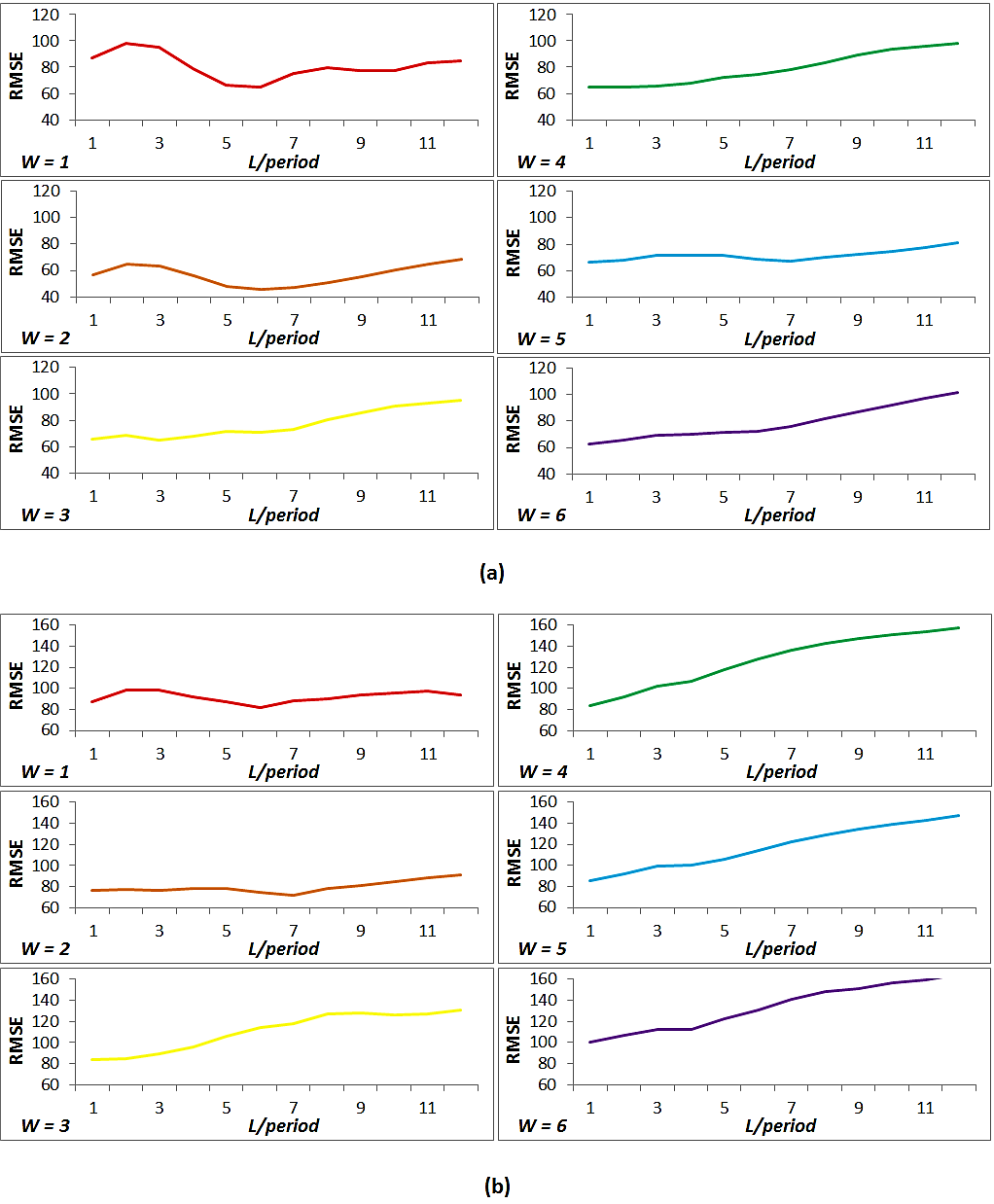


图3-3 个体模型与相似模型在不同〈L, W〉参数对下的RMSE值

然后，我们将这两个训练好的子模型的预测结果用神经网络聚合起来。由于本实验的聚合较为简单，我们选取一个三层模型的（输入层、隐层和输出）神经网络模型来做这样的聚合。该三层模型的中的输入层为两个神经元，输出层一个为神经元，而隐层的神经元我们将通过训练来决定其个数。设定隐层神经元的个数从4到12变化，循环的每一次我们都计算出对应神经网络模型的RMSE值。图3-4(a)展示了RMSE的值随着隐层神经元个数的变化而变化的趋势。从图3-4(a)中我们发现最优的隐层神经元个数为8，因此，我们选取2-8-1神经网络模型来做最终的模型聚合。图3-4(b)展示了个体模型、相似模型以及神经网络聚合模型之间的性能评估结果。从图中我们得出，聚合后的模型的RMSE明显降低，而MPAE则无明显变化。

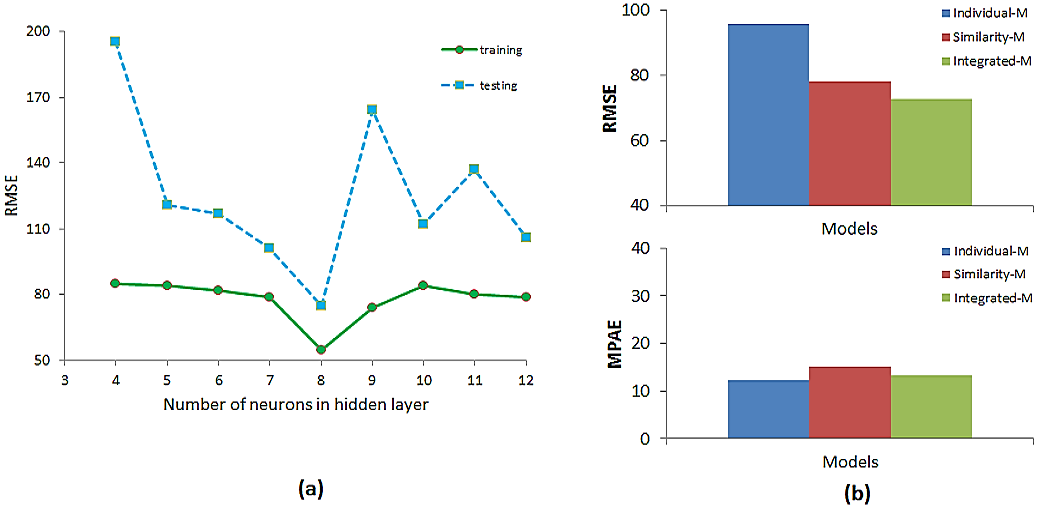


图3-4 不同神经元个数下的训练集与测试集的RMSE值，

以及个体模型、相似模型与聚合模型之间的性能比较

在经过数据提取、归总和预处理的交通数据集上，我们用几个流行的单源模型naïve、MA、ARIMA和NN模型，以及聚合模型DA[26]来与PMA模型做对比评测。这几个对比模型简述如下：

1. **Naïve(or no-change)**模型的形式比较简单，其中表示的是t时刻下的流量预测值：

(8)

1. **MA**模型的计算方式如下，其中k是MA模型中的参数[19]。MA的核心思想是处理最近的k个采样数据；k的值在每一次的平均计算中随着时间保持不变。

(9)

1. **ARIMA**模型的参数为(r, d, s)，其中d是差分参数，r和s是模型AR与MA的操作运算符。
2. **Artificial神经网络**模型是一种单源输出模型，通常用来拟合非线性数据集。模型的输入是时刻t的前个连续的流量值，而模型的输出则是在时刻t处的流量预测值。该模型的的长度以及隐层的神经元个数同样由实验评估决定，限于篇幅，此处不再赘述。

(10)

1. **DA**模型通过神经网络将三个子模型（MA，ES和ARIMA）的输出结果进行融合作为模型的预测结果。

Naïve，MA，ARIMA，NN以及DA这些模型每一个需要的历史数据长度都不一样。因此针对每一个模型的不同数据要求我们给出不同的数据样本空间。例如，naïve模型训练数据的长度不能超过预测值之前一个小时；而对于ARIMA模型来说，其训练数据集至少要跨越48个小时才能使模型进行有效的预测。对于以上每一个模型，我们定参后与PMA模型在相同的测试集上进行对比评估。对于naïve模型，我们只需要简单的取历史最近一次的流量值作为其预测值；对于MA模型，通过实验我们我们得出当k等于3的时候流量预测值的平均误差最小；而对于ARIMA模型，经过试验我们知道参数选取为(1, 1, 0)的时候该模型的流量预测误差最小；通过实验，我们设定ES模型中的参数α= 0.1和γ= 0.1以及NN为3-12-1来构建DA模型，这样能达到流量预测的最小误差。然后同样的，我们预测未来一个时间单位内的交通流量值。

由于针对工作日和周末两份数据集实验结果较为相似，因此我们只给出针对工作日数据集的实验结果。图3-5展示了各个流量预测模型之间以及与PMA模型之间RMSE与MPAE指标的对比结果。根据图(a)和图(b)的实验结果可知，PMA模型的流量预测结果要优于以上任何一种流量预测模型的预测结果。

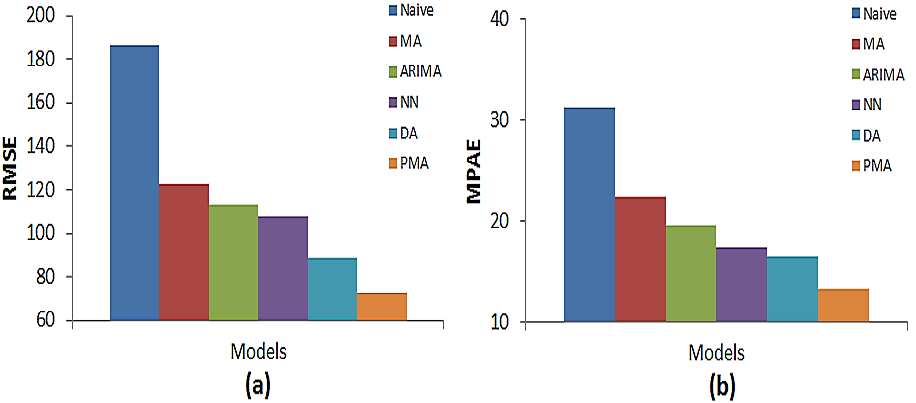


图3-5 不同模型与PMA模型关于RMSE与MPAE指标的对比结果

## 3.4 本章小结

本章节分别从模型定义、建模过程、模型训练、模型预测到实验结果与对比这五个方面逐步详细阐述了PMA模型的设计与实现方式。通过在从真实的交通治安卡口系统处获取的原始数据提取并处理得到实验数据集上进行大量的实验，具有一定的实际意义。并与若干现存模型来做性能的对比，验证了我们提出的PMA模型所具有的优越性。

# 路线推荐

本章对我们提出的道路捆绑模型进行描述。首先对模型用到的概念进行定义；然后详细描述模型的建模过程以及模型的推荐策略；接下来介绍模型的实验情况，包括数据集、实验环境和实验结果；最后与T-Drive模型在相同数据集上进行预测结果的比较，说明该模型的可行性。

## 4.1 Spica模型定义

在具体阐述道路捆绑模型Spica的实现机制之前，本小节先对该模型将要用到的一些概念进行定义与说明，我们将依次给出路段、路网、路线、车辆轨迹、衔接边以及衔接图的定义，为后文描述模型的实现细节打下基础。我们定义说明如下六个概念。

**定义3(路段)**一条路段r指的是一种单向或者双向的边，它包含两个终端点（r.s，r.e，本文中具体指的是道路交通监控摄像头的位置）。如果一条路段是单向边，则r只能允许从r.s行驶到r.e。

**定义4(路网)**一个路网Gr其实就是一个有向图，Gr=(Vr, Er),其中Vr是一个路段顶点的集合，Er是路段的集合。路网中每个路段的通行时间是动态变化的，在一天中随着时间的不同而不同。

**定义5(路线)**一条路线R指的是一个相连的路段集合，R：r1→r2→…→rn，其中ri.e = ri+1.s,(1 ≤i< n)。一条路线的起始点和终止点可以写成R.s = r1.s和R.e = rn.e。

**定义6(车辆轨迹)**一条车辆轨迹Tr指的是某个车辆被道路交通监控摄像头所记录的某一次行驶轨迹。车辆轨迹中的每一个节点p∈Pr包含了一个时间戳p.t(分钟计)。Tr : p1→p2→…→pn,其中0< pi+1.t – pi.t <ΔT,(1 ≤i< n)。Pr是所有监控摄像头的位置集合，ΔT代表一个路段的最大旅行时间。

**定义7(衔接边)**每条衔接边ej都是由相连的两条路段组成（ej.r1:pa→pb,ej.r2:pb→pc），因此包含了两个旅行时间ej.t1和ej.t2。旅行时间的计算将在后文进行描述。衔接边将从每一条真实的车辆轨迹中提取。

**定义8(衔接图)**衔接图Gj=(Vj,Ej)是一个由衔接边顶点集合Vj和衔接边集合Ej组成的一个有向图。

本小节通过数据层面和建模层面，从数学角度定义了道路捆绑模型Spica中的路段、路网、路线、车辆轨迹、衔接边和衔接图等概念，为下文对模型的具体描述打下了基础。

## 4.2 Spica建模

Spica模型利用了车辆轨迹特征，将相连的路段用一种衔接技术“捆绑”在一起，因此我们称之为道路捆绑模型。

本小节首先我们将阐述什么是衔接技术以及其实现方式；接下来阐述时间依赖衔接图TDJ的构建方式；然后针对TDJ中的每一条边我们描述其相应的旅行时间估计方法；最后阐述在TDJ图上如何用时间依赖启发算法TDHA进行最优路径的计算，进而得出Spica的路线推荐结果。

### 4.2.1 衔接技术分析

现代城市尤其是大中型城市的道路错综复杂，基本都是多道并行，而这些车道又可以分为快车道、慢车道、专用车道、左转车道、右转车道以及直行车道等。在不同车道上行驶的车辆通常具有不同的速度，不同的速度就意味着不同的旅行时间。然而，现存的时间依赖旅行时间估计模型只是将每一个路段的旅行时间视为条件独立的个体进行分析（如文献[16][17]），而这样的方式将造成一些原本不可能情况发生。比如，某个时刻右行车道上的行车速度远大于直行车辆的行车速度的时候，通常意味着直行路段堵车或者在等直行的红绿灯。在这种情况下，已经在直行车道实线内等待的车辆是无法右转至右行车道上的；同样的，已经驶入右行车道实线内的车辆也无法降速插入直行车道。这些车辆只能沿着自己原有的路线行驶下去，这也就说明了他们的行驶速度无法完成“高低跨越”，而在现有的旅行时间估计模型则有可能出现这样的非正常现象。本文中，我们将这样的非正常现象称之为路线衔接问题。

为了避免非正常的的路线推荐，保证驾驶线路能够在正常的行驶时间内完成，因此我们提出了一种将相连路段通过旅行时间连接在一起衔接边，而不是针对每一个路段给出单独的一个旅行时间。图4-1(a)展示了衔接边的三种形态，图4-1(b)描述了其中一种右转衔接边的实际例子。

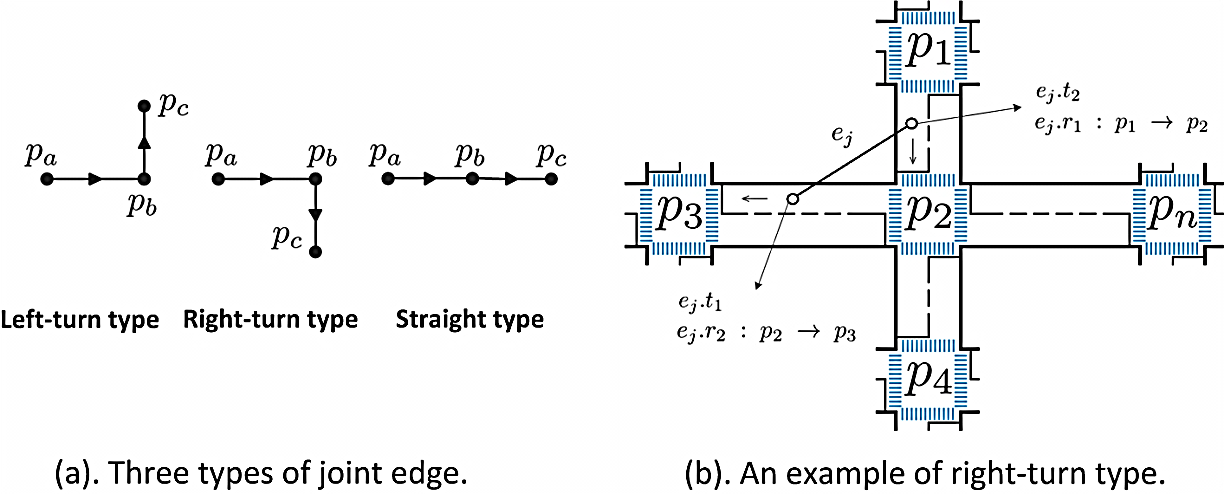


图4-1 衔接边的描述

### 4.2.2 TDJ图的构建

为了构建TDJ图，首先我们通过交通卡口监控系统获取一段时间的车辆行驶记录；然后从该数据集中提炼抽取出车辆的轨迹数据集；接下来将车辆轨迹分解成若干衔接边，组成衔接边数据集；最后我们用这个衔接边数据集来构建TDJ图。

从车辆轨迹数据集中我们发现，每周相同天的数据轨迹呈现出相似的变化规律，且工作日的轨迹数据之间的相似度高，周末的轨迹数据相似度较高。因此，我们又将衔接边的数据集按时间划分成了工作日数据集与周末数据集。图4-2阐明了一个包含24条衔接边的TDJ图的转化构建过程，其中r代表路段（见**定义3**）。

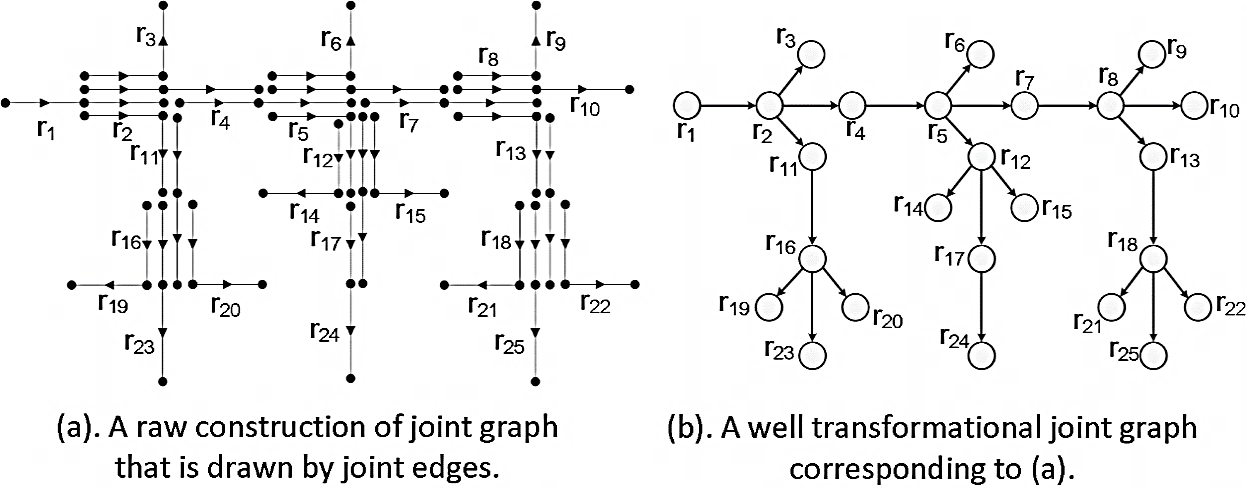


图4-2 一个构建TDJ图的示例

### 4.2.3 旅行时间估计

这里我们将对一天中每一种衔接边的旅行时间进行计算。**定义4**中我们知道路网的旅行时间是动态变化的，因此我们不能简单的对每一条衔接边给出一个固定的旅行时间。为了简单起见，我们用Δt表示最小的时间单元，其值的选取会在第四章的实验评估中提到。对衔接边的旅行时间估计见如下公式(11)。

(11)

其中n表示车辆轨迹数据集的容量；ej的表示衔接边（见**定义7**）；t1表示车辆轨迹从采集点pa到pb的旅行时间，t2表示从采集点pb到pc的旅行时间；{pa, pb, pc}表示衔接边按车辆轨迹的行驶方向分别经过的三个数据采集点的标号；pa.t，pb.t，pc.t表示从该点出发的时刻。

相同路段的车辆轨迹数据的旅行时间往往存在差别，这是由于不同的司机具有不同的驾驶经验。我们用参数σ来描述司机穿过衔接边三个点{pa, pb, pc}的驾驶经验，用参数表示司机穿过车辆轨迹{p1, p2, …, pn }的平均驾驶经验。一般情况下，司机的驾驶经验越好，其旅行时间就越短。例如，意味着该司机的驾驶技术胜过其他80%的司机。σ和计算公式如下：

(12)

(13)

其中τ代表一天中的时间单元序号，单位为Δt。

## 4.3 时间依赖启发式算法TDHA

由于时间依赖启发式算法TDHA略为复杂，特此单独列一小节出来对其进行描述。不同于静态图的最短路计算，TDJ图中各个顶点的时间距离随着时间的变化而变化。在这种情况下，我们不能简单的使用一些静态的最短路算法如迪杰斯特拉算法、贝尔曼佛德算法和佛洛伊德算法来计算最短时间路径。本小节提出的TDHA启发式搜索算法类似于A\*算法，算法分成两个阶段：第一阶段，计算启发式估价函数，用来降低下一阶段的运行时间成本；第二阶段，进行时间依赖的宽度有限搜索寻找时间最优路线。

本小节接下来将从启发估价函数、时间依赖的宽度优先搜索TBFS（time-dependent breadth-first search）以及最终的路线推荐这三个部分描述TDHA的实现过程。

### 4.3.1 启发估价函数

在执行TBFS之前，我们先计算启发估价函数来降低搜索的成本。首先，我们定义启发估价函数，作为每一对起点和终点(pi, pj)在第τ时间单位下的时间距离预估值并将其初始化为0。然后，检索车辆轨迹数据集中的每一条轨迹数据来更新的值，让其越小越好（0除外）。算法2的代码详细描述了更新的过程。

表4-1 启发式估价函数算法实现

|  |
| --- |
| **算法2：**启发式估价函数 |
| 1: Firstly, initialize each to zero;  2: 0 <i< j ≤ n;  3: for each vehicle trajectory Tr : p1→...→pi→...→pj→...→pn do  4: for i = 1...n - 1 do  5: for j = i + 1...n do  6: ;  7: if > 0 then  8:  9: else  10:  11: end if  12: end for  13: end for  14: end for |

### 4.3.2 时间依赖的宽度优先搜索算法

对于一个给定的ts时刻的路线查询（一个起始点pstart和一个目的点pend），我们用TBFS算法来进行最短路的计算。TBFS是BFS的一个变种，TBFS的实现分成两个部分，第一部分是TBFS算法的初始化（见算法3），第二部分是TBFS算法的实现主体，具体计算细节见算法4。

表4-2 TBFS算法初始化

|  |
| --- |
| **算法3：**TBFS初始化 |
| 1: TBFS (Gj = (Pr, Vj, Ej), pstart);  2: // program initialization;  3: for each p∈Pr except pstart do  4: distance[p] = infinite;  5: end for  6: i = 0, k = 0;  7: prefix[0] = 0, parent[0] = pstart;  8: for each joint edge e∈Ejdo  9: if pstart = then  10: distance[] = e. t1;  11: ENQUEUE(Qr, e. r1 ());  12: k = k + 1;  13: ;  14: end if  15: end for |

Gj是一个衔接图（见**定义8**），Pr在**定义6**中已说明，在**定义3**中描述了Vj中的点而Ej中的衔接边指的是**定义7**中描述的路段{pa, pb, pc}。算法3中distance、prefix与parent是三个数组，其中：数组distance记录以pstart为起点的单源最短距离，数组prefix存储TBFS过程的前驱节点号，数组parent存储当前节点的父节点号。算法4用计算好的dPre来优化其时间复杂度。

表4-3 TBFS算法实现主体描述

|  |
| --- |
| **算法4：**TBFS算法 |
| 1: TBFS (Gj = (Pr, Vj, Ej), pstart);  2: // calculate the time-dependent shortest path;  3: while not(empty(Qr)) do  4: i = i + 1;  5: r = DEQUEUE(Qr);  6: for each joint edge e(pa, pb, pc) do  7: if and equals to then  8: τ1= , τ1= ;  9: if < then  10: continue;  11: end if  12: if then  13:  14: ENQUEUE(Qr, e. r2 ());  15: k = k + 1;  16: ;  17: end if  18: end if  19: end for  20:end while |

### 4.3.3 路线推荐

经过TBFS过程之后，我们得到distance、prefix和parent三个结果数组，通过一个回溯的方法我们就能得到模型最终的路线推荐结果。回溯的方法一共分成三步：首先，在parent数组中反向找出pend第一次出现的位置并将之压入栈中；然后我们从该点出发根据prefix数组记录的关系进行反向查找直到当前数组下标为0并在过程中依次将回溯到的节点压入栈中；最后，我们按栈的输出顺序输出栈中所存的位置序列，就得到了Spica模型从起始点pstart到目的点pend的路线推荐结果。

## 4.4 实验及分析

### 4.4.1 数据集与实验环境

实验用的数据集取自某大城市交通治安卡口系统所采集的车辆行驶数据[15]，数据采集时间是2015年3月9日到2015年3月15日。本数据集一共记录了44,593,706条真实的车辆行驶信息，采集的数据一共覆盖了全市900个监控摄像头。我们将这些数据按时间维度划分成了一个训练集和一个测试集：将工作日3月9日到3月12日以及周末3月14日的数据划为训练集，3月13日和3月15日的数据划为测试集。实验环境用的是戴尔OPTIPLEX990型号的台式机，因特尔i5-2400处理器，CPU主频3.1GHz，内存大小为4GB。

然后我们对数据集进行预处理操作。根据同一车辆行驶信息的时间先后顺序，对其进行Δt时间间隔的路径归总，转化成车辆轨迹数据。经过预处理后，训练集和测试集加起来我们一共得到了4,415,329条车辆轨迹数据。

### 4.4.2 评价指标

我们使用匹配率MR（Match Rate）来评价Spica模型中衔接技术的合理性，用Rα（α率）和Rβ（β率）两个统计指标来评价不同路线推荐模型的路线推荐的合理性。三项评价指标的描述如下：

1. 指标MR用来衡量路线衔接问题的准确性与合理性。

(14)

1. 指标Rα表示在A方法中有多少条路线推荐的方案优于B方法。

(15)

1. 指标Rβ用来反应A方法给出的推荐路线优于B方法的程度。

(16)

### 4.4.3 结果与对比

在Spica模型中，数据集也需要经过预处理才能得到实验可用的数据。我们把Δt作为时间最小单位，来归总衔接边的顶点，这里我们根据实验结果（详见**4.5小节**）设定时间单位Δt为5分钟。经过对原始数据的预处理后，我们一共得到18172个流量模式和4,415,329条车辆轨迹。

时间单位Δt的大小决定之衔接边的数量以及模型的时间复杂度。图4-3展示了衔接边的数量以及模型从执行到结束的时间长度随着Δt大小的变化曲线。根据图4-3(a)和(b)中的结果曲线，我们将时间单位Δt设为5分钟。

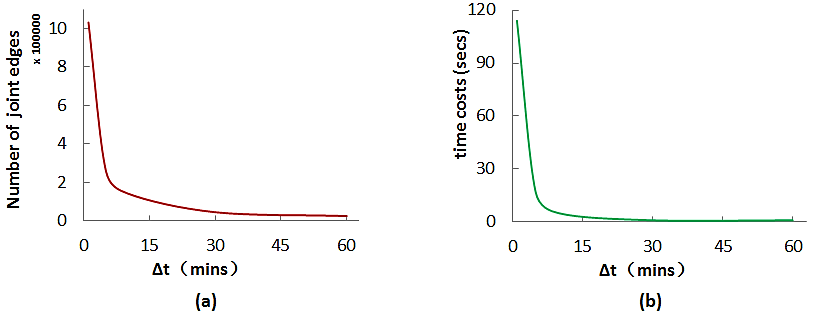


图4-3 衔接边数量与模型的时间复杂度随着Δt的变化曲线

在经过数据提取、归总和预处理的交通数据集上，我们对所有的车辆轨迹进行对应的路线推荐并将T-Drive方法[17]作为对比的方法进行路线推荐结果的比较。

首先，基于**4.2.1**小节的理论，相连路段之间的旅行时间往往不相同。从文献[16][17]中我们了解到同样类型的一类司机所产生的车辆轨迹的旅行时间往往处在同一个旅行时间簇之中。我们取1至5不同的旅行时间簇[16][17]数目以及当时间簇数目为2时不同的车辆轨迹长度对MR指标分别进行评估。图4-4(a)和4-4(b)展示了分别展示了以上两种实验情况。从图中我们能看出随着时间簇数目的增多和车辆轨迹的变长，T-Drive模型的MR指标变得越来越低，与之相反，衔接技术则能相对抑制路线衔接问题的出现。

然后，我们利用指标Rα和Rβ来评估Spica模型的性能并选取真实的路线作为模型的基准对比。对于不同司机的驾驶经验，通过**4.2.3**小节我们计算出了每一条车辆轨迹的驾驶经验值来对模型预测出的旅行时间进行先衰减相乘，进而得到模型最终的路线推荐结果。

同样的，由于针对工作日和周末两份数据集实验结果较为相似，这里我们也只给出针对工作日数据集的实验结果。图4-4(c)展示了Spica模型与T-Drive模型对于指标Rα在一天中不同起始时间τ下的实验结果，图4-4(d)则展示了Spica模型与T-Drive模型分别在不同的Rβ值下推荐路线的占比。从图(c)和图(d)中看出，虽然T-Drive的Rβ值评估效果略优于Spica模型，但Spica模型的实验结果也很高效而且这两种模型的Rα几乎持平。因此，实验结果证明了Spica模型的有效性，对给出合理的路线推荐能达到令人满意的结果。

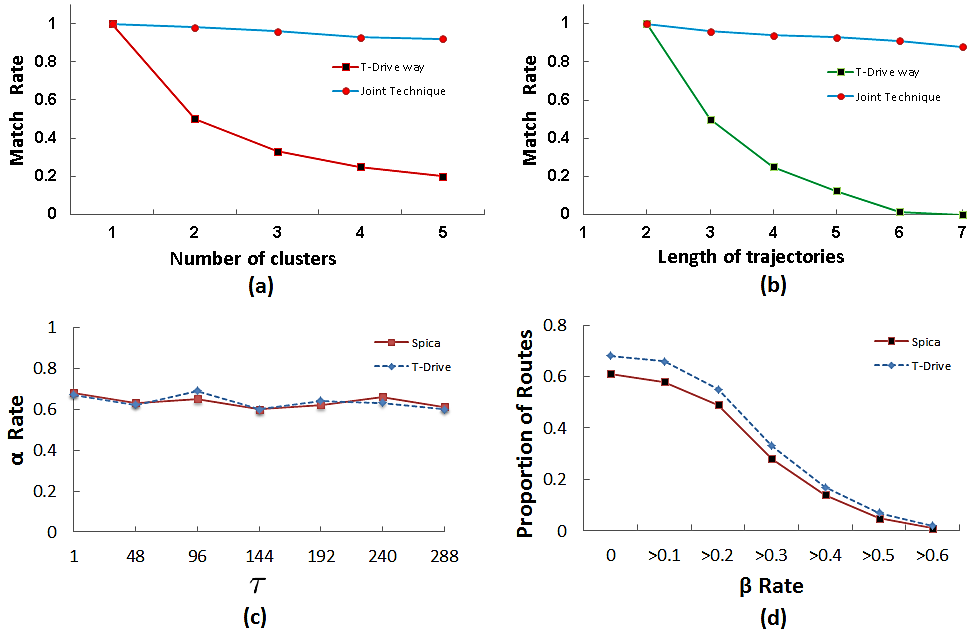


图4-4 MR指标下路线衔接问题的影响及Spica模型与T-Drive模型效果对比

## 4.5 本章小结

本章节分别从模型定义、建模过程、模型训练、模型预测到实验结果这五个方面逐步详细阐述了Spica模型的设计与实现方式。通过在从真实的交通治安卡口系统处获取的原始数据提取并处理得到实验数据集上进行大量的实验，具有一定的实际意义。并与T-Drive模型做性能的对比，结果具有较强的说服力，充分验证了我们提出的Spica模型所具有的合理性和可行性。

# 总结和展望

## 5.1 总结

本文针对交通数据挖掘领域的流量预测与路线推荐两个热点问题进行了研究和探索。针对交通流量预测问题存在的不足，本文提出了一种采用平面滑动平均算法的聚合模型，该模型从新的视角出发将交通流量数据之间的关系提取出来，与传统的统计手段相结合提出一种新的流量预测方式，并通过在真实的交通流量数据集上进行了大量的实验，验证了PMA模型的可行性。针对交通路线推荐问题存在的不足，本文提出了一种道路捆绑模型，该模型同样从新的视角出发研究车辆轨迹的特征，提出衔接边的新概念将相连接的不同路段的旅行时间与车辆轨迹关联起来使得推荐的结果变得更加可靠，通过在真实的数据集上进行的实验证明了Spica模型的可行性。

总体来说，本文做出的主要贡献可归为以下几点：

1. 提出了一种平面滑动平均模型PMA来预测短时交通流，从一种全新的数据视角将个体交通数据流与相似交通数据流相结合进行预测，并举例说明了该模型合理性与可行性。
2. 提出了一种道路捆绑模型Spica来进行路线推荐，从一种全新的数据角度将相连接的不同路段的旅行时间与车辆轨迹相结合进行路线的规划，并列举了现实生活中的行车例子说明了该模型的合理性与可行性。
3. 在真实的交通治安卡口监控数据集上，对上述提出的两种交通模型进行了大量的实验论证，进一步证明了这两种模型的合理性与可行性。并且通过对核心算法时间复杂度进行优化，让模型执行的时间成本降到了单机可承受程度，为今后实际投入使用做好了准备。

## 5.2 未来工作

交通流量预测方面，本文提出的PMA模型仅仅考虑了交通流量数据中纯数字之间的变化关系。而随着移动互联网、大数据、数据挖掘以及机器学习等技术的迅猛发展，我们可以将交通流量采集点的地理位置因素、车流量的迁移特性以及车辆的移动位置等因素考虑进来，建立包含更多不同的维度交通特征的流量预测模型。这样信息涵盖全面的交通流预测模型将会给出更加准确的流量预测结果。

路线推荐方面，本文提出的Spica模型中仅仅考虑了相连路段之间旅行时间的关系。随着移动互联网、物联网和车载互联网的快速发展，我们可以将更多的移动轨迹信息运用到车辆轨迹的提取中来，然后获得更长的衔接边。比如某个较优路线如果有50%的人都走过，那么Spica模型就能通过更长的衔接边技术把这条路线原封不动的推荐给另外50%的人。

总之，随着科学技术的发展，信息时代的全面普及，能获取的数据种类与方式越多，本文提出的模型的成长空间就越大。

# 参考文献

1. Smith B L, Demetsky M J. Traffic Flow Forecasting: Comparison of Modeling Approaches[J]. Journal of Transportation Engineering, 2015, 123(4):261-266.
2. 谭国真, 高文. 时间依赖的网络中最小时间路径算法[J]. 计算机学报, 2002, 25(2):165-172.
3. Dreyfus S E. An Appraisal of Some Shortest-Path Algorithms[J]. Operations Research, 1969, 17(3):395-412.
4. Smith D E K R L. Fastest paths in time-dependent networks for intelligent vehicle-highway systems applications[J]. Ivhs Journal, 1993, 1(1):1-11.
5. Orda A, Rom R. Shortest-path and minimum-delay algorithms in networks with time-dependent edge-length[J]. Journal of the Acm, 1997, 37(3):607-625.
6. 韩平阳, 罗五明, 王志敏,等. 动态网络中的最短路径改进算法[J]. 军事运筹与系统工程, 2007, 21(1):46-50.
7. 余伟辉, 陈闳中. 时间依赖网络中非FIFO弧的转化研究[J]. 小型微型计算机系统, 2009, 30(1):156-158.
8. 龙科军, 王赛政, 肖向良. 面向驾驶员特性的路径规划算法[J]. 计算机工程, 2011, 37(5):264-266.
9. 陈京荣, 俞建宁, 李引珍. 多属性随机时间依赖网络路径优化[J]. 西南交通大学学报, 2012, 47(2):291-298.
10. 孙奥, 朱桂斌, 江铁. 基于预测信息的时间依赖网络路径规划算法[J]. 计算机工程, 2012, 38(22):133-136.
11. Gao S, Huang H. Real-time traveler information for optimal adaptive routing in stochastic time-dependent networks[J]. Transportation Research Part C Emerging Technologies, 2012, 21(1):196-213.
12. Khani A, Lee S, Hickman M, et al. Intermodal Path Algorithm for Time-Dependent Auto Network and Scheduled Transit Service[J]. Transportation Research Record, 2012, 50(2284):40-46.
13. Papagiannaki K, Taft N, Zhang Z L, et al. Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models[J]. Proceedings - IEEE INFOCOM, 2010, 2(5):1178--1188.
14. Konstantina P, Nina T, Zhi-Li Z, et al. Long-term forecasting of internet backbone traffic.[J]. IEEE Transactions on Neural Networks, 2005, 16(5):1110-1124.
15. Chen M, Liu Y, Yu X. NLPMM: A Next Location Predictor with Markov Modeling[M]// Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2014:186-197.
16. Yuan J, Zheng Y, Xie X, et al. Driving with Knowledge from the Physical World[J]. Kdd ’, 2011:316-324.
17. Yuan J, Zheng Y, Xie X, et al. T-Drive: Enhancing Driving Directions with Taxi Drivers' Intelligence[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(1):220-232.
18. Tan M C, Wong S C, Xu J M, et al. An Aggregation Approach to Short-Term Traffic Flow Prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(1):60-69.
19. Hanke J E, Reitsch A G. Business forecasting[J]. Australian Quarterly, 2009(3):101-109.
20. Q Gao and G Li. A traffic prediction method based on ann and adaptive template matching. 2011.
21. Chan K Y, Dillon T, Chang E, et al. Prediction of Short-Term Traffic Variables Using Intelligent Swarm-Based Neural Networks[J]. IEEE Transactions on Control Systems Technology, 2013, 21(1):263-274.
22. Chan K Y, Dillon T S, Singh J, et al. Neural-Network-Based Models for Short-Term Traffic Flow Forecasting Using a Hybrid Exponential Smoothing and Levenberg–Marquardt Algorithm[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2):644-654.
23. Davarynejad M, Wang Y, Vrancken J, et al. Multi-phase time series models for motorway flow forecasting[C]// International IEEE Conference on Intelligent Transportation Systems. 2011:2033-2038.
24. Guo H, Xiao X, Tang Y. Short-Term Traffic Flow Forecasting Based on Grey Delay Model[M]// Artificial Intelligence and Computational Intelligence. Springer Berlin Heidelberg, 2012:357-364.
25. Shang J, Zheng Y, Tong W, et al. Inferring gas consumption and pollution emission of vehicles throughout a city[C]// AcmSigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2014:1027-1036.
26. Tan M C, Wong S C, Xu J M, et al. An Aggregation Approach to Short-Term Traffic Flow Prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(1):60-69.
27. Williams B. Multivariate Vehicular Traffic Flow Prediction: Evaluation of ARIMAX Modeling[J]. Transportation Research Record Journal of the Transportation Research Board, 2001, 1776(1):194-200.
28. Williams B M, Hoel L A. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results[J]. Journal of Transportation Engineering, 2014, 129(6):664-672.
29. Chan K Y, Dillon T S, Singh J, et al. Traffic flow forecasting neural networks based on exponential smoothing method[C]// Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on. IEEE, 2011:376-381.
30. Chan K Y, Yiu C K F. Development of Neural Network Based Traffic Flow Predictors Using Pre-processed Data[M]// Optimization and Control Methods in Industrial Engineering and Construction. Springer Netherlands, 2014:125-138.

# 致 谢

时间真如白驹过隙，研究生生涯在不知不觉中就要结束了。回首过去的近三年的校园时光，有欢笑、有喜悦，也有失落、有迷茫。无论经历的是喜是忧，无论面对的挫折有多大，身边总有一群温馨的小伙伴，有我室友，有我实验室的师兄弟姐妹，有我的同班同学，也有我的各种球友，他们帮助我进步、帮助我成长，我们一起度过了难忘而意义不凡的研究生生活。想到这里，我的心中便充满了无尽的感激与不舍。

首先，我想对我的导师禹晓辉教授表达敬意与感激之情。从大四保研进入实验室起一共三年多的时光里，禹老师在科研、工作和生活上都给予了我无微不至的关怀与指点。在科研上，他不厌其烦的教授我做研究的方法，教导我要沉下心来打好基础，端正科研态度，不时地向我讲述研究领域内的最新动况，开拓新的思路和视野，引导我研究新的切入点，鼓励我向国际重要会议期刊投稿，并带头将理论转化为实践，将研究成果运用到实际工程中去。在工作上，他以身作则，树立榜样，教会我要有责任心，事情再小也要投入十二分的态度去做好，遇到困难时沉着应对，静下心来思考，往往换个思路就能得到意料之外的效果。在生活上，他与我分享丰富的社会阅历，让我在迈入社会大门之前做好了充分的思想准备，在我遇到挫折时会鼓励我帮我指点迷津，尽自己最大的努力助我走出困境。所有这些禹老师教给我的经验和传递给我的品质，将成为我今后发展当中的一笔宝贵财富。

我想对身边的各位同学表示感谢。研究生一年级在元旦晚会上当主持人的时候，我能想起台下每一位同学天真灿烂的笑脸；研究生二年级在“山大杯”篮球赛上，我能想起我和队员们一起在球场上的挥汗如雨；每逢周末的时候，我能想起和同学们一起去K歌，一起去吃鸡公煲；每逢晚上又和舍友们一起在寝室聊八卦，谈人生，畅想未来。一幕幕温馨的场景长存心中，是你们让我背井离乡的求学之路变得丰富多彩，是你们让我感受到了友情的魅力，同时也让我看到了自身的不足，三人行必有我师，你们的优秀是我进步的动力。

我还要感谢所有关心我的亲朋好友们，感谢你们无私而伟大的付出，感谢你们始终如一的支持与陪伴，感谢你们与我千里共婵娟。

# 攻读学位期间发表的学术论文

[1] Lv L, Chen M, Liu Y, et al. A Plane Moving Average Algorithm for Short-Term Traffic Flow Prediction[M]// Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2015:357-369.

# 攻读学位期间参与科研项目情况

1．国家自然科学基金面上项目（61272092 面向微博的实时流数据处理平台和查询处理关键技术研究，2012.10-2016.12）

2．山东省自然科学基金重点项目（ZR2012FZ004 面向海量时空数据管理的云计算关键技术研究，2012.7-2015.7）