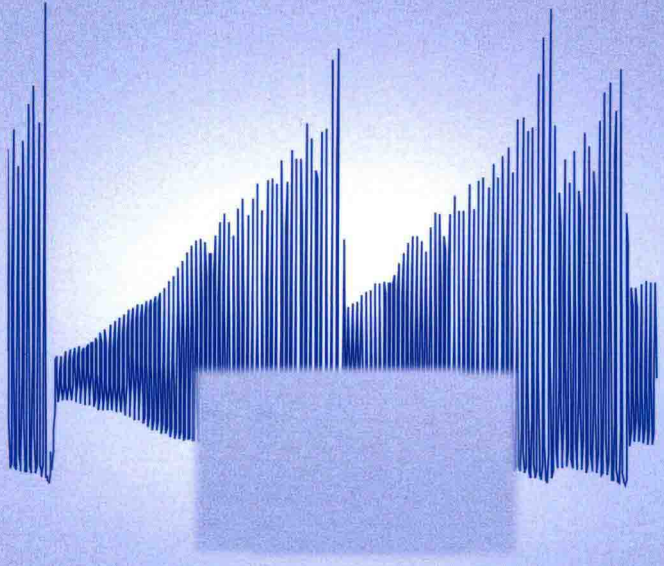
Theories and Approaches on Intelligent Short-term Prediction on Traffic Information

短时交通信息智能预测理论及方法

王 扬 陈艳艳 著



人民交通出版社股份有限公司
China Communications Press Co.,Ltd.

短时交通信息智能预测理论及方法

Theories and Approaches on Intelligent Short-term Prediction on Traffic Information

王 扬 陈艳艳 著

人民交通出版社股份有限公司
China Communications Press Co.,Ltd.

内 容 提 要

本书从交通信息预测的研究意义、基本理论及分类等基本背景知识入手，介绍了作者在交通数据的异常值祛除、噪声抑制及缺失数据填补等预处理方法上的研究成果，对基于不确定性理论的一些预测模型进行了比较研究，并在此基础之上提出了两种基于非监督学习方法的单步智能预测模型，在本书最后着重介绍了作者在多步预测方面的研究成果，并结合实验研究论证了五种多步预测新方法的有效性。

本书可供同行学者参考阅读，也可作为参考书供相关专业研究生学习。

图书在版编目(CIP)数据

短时交通信息智能预测理论及方法 / 王扬，陈艳艳著. —北京：人民交通出版社股份有限公司，2016.9

ISBN 978-7-114-13148-6

Ⅰ.①短… Ⅱ.①王… ②陈 Ⅲ.①信息技术—应用—交通运输管理—智能控制—预测控制—研究 Ⅳ.①U495

中国版本图书馆 CIP 数据核字(2016)第 144516 号

书　　名：短时交通信息智能预测理论及方法
著 作 者：王　扬　陈艳艳
责任编辑：戴慧莉
出版发行：人民交通出版社股份有限公司
地　　址：(100011)北京市朝阳区安定门外外馆斜街 3 号
网　　址：http://www.ccpress.com.cn
销售电话：(010)59757973
总 经 销：人民交通出版社股份有限公司发行部
经　　销：各地新华书店
印　　刷：北京市密东印刷有限公司
开　　本：787×980　1/16
印　　张：9.75
字　　数：220 千
版　　次：2016 年 9 月　第 1 版
印　　次：2016 年 9 月　第 1 次印刷
书　　号：ISBN 978-7-114-13148-6
定　　价：36.00 元

前　言

随着我国城镇化的步伐不断加快，经济和社会活动日趋频繁，城市交通系统正经历着前所未有的演变历程，不确定无序的交通状态已成为城市交通的新常态，短期突发的交通拥塞也趋于常态化。为了更好地应对这种短时突发的交通演变，并及时做出科学的控制和诱导策略，需要实时准确的交通预测信息。交通预测研究是进一步提升城市交通系统智能化水平的一项重要内容。

城市交通系统日趋庞大复杂，表现出极强的非线性和不确定性，传统的预测模型与方法已无法继续满足智能交通控制与诱导等要求。为此，一批学者近年来开展了许多有益的探索工作。为了更好地满足交通系统新常态下的短时交通信息预测需求，本书作者近年来持续开展了一些相关研究工作。本书是作者结合已有相关研究成果，经过系统梳理后，将近期零散的研究工作及成果汇集成书。本书从交通信息预测的研究意义、基本理论及分类等基本背景知识入手，介绍了作者在交通数据的异常值祛除、噪声抑制及缺失数据填补等预处理方法上的研究成果，在论述了预测效果检验及性能评价方法之后，对基于不确定性理论的一些预测模型进行了比较研究，并在此基础之上提出了两种基于非监督学习方法的单步智能预测模型，由于单步预测在实践中存在着较大的局限性，而多步预测理论和方法还尚不成熟，故在本书最后着重介绍了作者在多步预测方面的研究成果，并结合实验研究论证了五种多步预测新方法的有效性。

本书不仅汇集了作者近期在相关领域方面的主要研究成果，亦有些经验教训，而且在撰写方式上力求系统地展现短时交通信息预测的整个流程。因此，本书不仅适合同行学者参阅，也可作为参考书供研究生学习。由于作者学识浅薄、水平有限，加之经验不足，文中不乏纰漏与拙见，恳请读者不吝批评指正。

本书所介绍的研究工作曾经得到了北京市教育委员会科技发展计划面上项目（KM201010005021）“基于复融合的嵌套式综合交通预测系统”、北京市自然科学基金重点项目（8131001）“多方式出行链协同机理及公交一体化关键技术研究”、教育部留学回国人员科研启动基金资助项目（32004011201201）“轨道交通换乘通道的视频预警信息研究”、交通运输部建设科技项目（2015318J37130）“基于个体出行链的公交客流动态感知与特征提取技术”等的资助，在此一并表示感谢。

著　者

2016 年 6 月

目　录

第1章 绪 论

在远古时代，茹毛饮血的先人出于求生的本能，便有着强烈地想要预知未来的冲动。然而，先人们对于复杂演变事物认知的肤浅，使得早期的预测往往笼罩在一层玄虚迷信的色彩下。在现代社会，如能早于他人获知未来走势，便能在激烈的竞争中占据优势。

在现代社会中，道路交通无论是在人们的经济生活还是休闲出游中都具有非常重要的作用。然而，在大力推动汽车行业发展的同时，人们并没有充分地意识到交通资源的有限性和稀缺性，出行需求的快速增长必定给道路交通系统带来巨大压力。而且，效率低下的道路交通系统也无疑促使了交通紧张的进一步恶化。人们为了避免交通拥堵带来的出行成本增长，渴望提前了解并预知未来的交通状态。

在另一方面，作为道路交通系统的管理者更是千方百计地致力于提高交通系统的运行效率，以期实现以最小的交通资源来尽可能地满足所有民众的出行需求。而有效实现这些目标的前提之一，就是能够准确地掌握未来交通状态的变化和走势。因此，无论是交通系统的使用者还是管理者，交通系统的预测信息都是实现高效出行和交通运行的基础。

随着智能交通系统的快速发展，先进的出行信息服务系统扮演着越来越重要的角色。只有预先获取前方道路的状态，才能主动地做好决策。而且，在交通诱导与控制中，交通信息的获取是智能交通系统的最前端，是控制系统中流动信息的来源。然而，交通控制及诱导的作用效果，通常需要经历一段时间的积累，才能逐渐显现出来。这种控制及诱导的时滞性也需要交通信息预测的支持，才能提前预判，从而提高控制及诱导策略的作用效果。因此，交通预测在智能交通系统领域中占据着重要地位。

1.1 交通信息预测研究的意义

1.1.1 预测的概念

预测是一门广泛运用于社会、经济、科学技术等各个领域的新科学。在研究事物发生、发展所呈现的规律性以及分析现状条件、环境因素制约和影响的基础上，运用科学的理论、方法和各种经验，对未来所发生的事情进行合理的估计，并合理地推测事物未来演变的状态和发展的趋势。

根据调查数据资料对事物进行科学地分析，并挖掘其发展变化的规律对未来进行预测，这一系列工作就称为预测分析。预测工作的基础是充分的客观数据资料，预测分析中所采用的方法和手段称为预测技术。预测分析和预测技术两者总称为预测的理论和方法。将预测的理论和方法作为研究的科学则称为预测科学，简称为预测[1]。

随着社会的不断进步，人们对客观事物的发展有了更为深刻的认知，预测的科学性也因此不断提高。那么，预测的科学性是什么？预测的科学性就是指预测有科学基础，是在总结事物发展规律的基础上，对事物未来发展做出合理的推测，这里的科学基础主要涉及预测的理论、方法、资料和计算等其他因素。

预测可分为广义的预测和狭义的预测。广义预测包括在同一时期根据已知事物推测未知事物的静态预测，也包括根据某一事物的历史和现状推测其未来的动态走势，而狭义的预测，仅关注动态预测。预测理论既可以应用于研究自然现象，又可以应用于研究社会现象。

人类经过千百年来的生产实践和社会实践，总结出了“凡事预则立，不预则废”的经验。也就是说，预测是给决策系统为制定决策提供必须的未来参考信息。即，预测是为决策服务的，是为了提高科学决策及管理的水平，减少决策及管理中的盲目性，降低决策和管理中可能遇到的各种风险，使决策和管理目标能够得以圆满顺利地实现。正因为预测具有这样的决策支持功能，长期以来才受到了人们普遍的关注。

回顾历史，可以发现人类的预测活动源远流长。但美好的愿望往往因认知水平的局限，而常常显现出浓重的唯心主义和一定的迷信色彩。现代科学技术的发展，使得预测技术及相关理论逐渐成熟，并成功地应用到诸多领域，产生了许多分支，常见的如人口预测、经济预测、交通预测、科技预测、气象预测等。

1.1.2 交通信息预测的必要性

现代社会的发展使得大城市的交通状况令人堪忧。有关统计数据显示，每年因交通堵塞，美国的经济损失高达约1000亿美元，英国约200亿英镑，欧洲数千亿欧元。我国走

过了改革开放30年的发展,已经逐步成为汽车大国,近年来我国每年因交通堵塞造成的GDP损失达5%~8%。据世界银行的统计资料显示,北京二环至三环之间的干线道路上,高峰时车辆的平均速度,已经由1994年的45km/h,下降到2005年的10km/h以下,已经低于自行车的12km/h。中国社会科学院数量经济与技术经济研究所测算,北京市每天因为堵车造成的社会成本达到4000万元,相当于每年损失146亿元。由于过快增长的机动车出行需求与有限道路资源形成的尖锐矛盾,所带来的诸如道路堵塞严重、交通事故频繁、环境污染加重、能源消耗加大等问题,严重制约着首都经济持续稳定的增长和社会和谐健康的发展。

为了解决交通堵塞这个长期困扰人们出行的痼疾,各国政府以及许多研究院所先后采取了多种措施来缓解交通拥挤,其中智能交通系统(Intelligent Transportation Systems, ITS)成为最有希望解决交通堵塞的途径之一。智能交通系统就是利用通信、控制、计算机及传感等方面的先进技术,建立起高度信息化、智能化的交通管理系统。欧美等国近年来大力开展智能交通系统的研发和应用,并初见成效。我国目前也在积极研究智能交通系统,力求缓解交通拥挤,降低事故率,实现节能减排的科学发展。

作为智能交通系统的核心和关键部分之一,交通信息是实现科学交通管理和有效诱导的重要保障。道路交通是一个时变非线性的高度复杂开放系统,在系统演化过程中存在着许多不确定因素。尽管随着检测技术的不断提升,交通路况实时信息的应用也越来越广泛了,但是实时路况信息在支持交通管理与决策时缺乏先见性,对交通事件不能起到主动预防作用。比如,在利用可变信息板(Variable Message Signs, VMS)进行诱导时,由于信息作用具有一定的时滞性,也就是信息的诱导作用在一段时间后才能显现出来,所以只有预测信息才能真正地反映出诱导的效果。因此,只有掌握超前路况信息,才能及时准确地作出应对决策,从而尽可能地避免可能的不利交通事件。交通预测不仅是有效缓解交通拥挤、保证正确交通诱导的前提和关键,而且也对交通管理、城市规划、城市信息化建设等起到积极作用。

1.2 交通信息预测的基本理论

人们在大量的实践中,逐渐总结了许多预测方法。这些预测方法都是根据事物的发展规律,或在事物发展过程中出现了随着事物发展而显现出来的现象。因此,可将预测的基本原理归纳为以下几种。

(1)整体性原理。

整体性原理是基于系统的思想,认为事物是由若干元素构成的有机整体,因此事物发展变化过程便具有了整体性的特征。

(2)惯性原理。

事物的发展变化与其过去历史(尤其是近期过去)的行为总有着千丝万缕、或大或小的联系,即过去的行为影响现在,也势必影响未来,这种在时间上影响作用的现象称之为“惯性现象”。所谓惯性原理,就是通过研究对象的现在对过去的依赖,根据其所表现出来的惯性,预测其未来演化的状态。惯性原理是趋势外推法的主要理论依据之一。

(3)相似性原理。

根据已知的某事物的发展变化特征,推断具有近似特征的预测对象在未来的状态。相似性原理就是从已知领域过渡到未知领域的探索,是一种重要的创造性方法。类比物之间的相似特征越多,类推预测的则越准确,类比越可靠。

(4)相关原理。

相关原理,就是研究预测对象与其相关事物间的相关关系(尤其是其中可能存在的因果关系)性,利用相关事物的演变特性来推断预测对象的未来走势状态,主要表现形式为因果关系。

(5)概率推断原理。

概率推断原理就是指,当某个预测结果相对于其他可能的结果以较大概率出现时,则认为该预测结果可以成立。

(6)反馈原理。

通过反馈不断进行修正是基于人们在实践中总结出的经验教训来指导未来工作的基本思路。利用反馈原理能够更好地处理事物的动态演变过程,在具有较好的适应性的同时,能够提高预测的鲁棒性。

1.2.1 基于影响因素的预测理论

事物的发展总是有前因后果。即便是在没有探明其事物发展的内在因果关系时,或是在无法清晰描述事物发展的起因时,也总会在事物发展的过程中,存在着一些对其发展具有深刻影响作用的因素。然而,不论这些因素是否是造成事物发展的本质原因,人们发现可以借助这些影响因素来预测事物的发展。这是基于影响因素进行预测的基本思想。

在进行预测时,首先要对收集的大量数据进行分析,确定具有较强影响作用的因素,再建立它们相互影响的关系模型,并进行参数标定,最后还需对建立的预测模型进行验证。在交通信息预测中,由于影响着交通运行的因素很多,在这些影响因素中,有些容易测得,而另外一些因素则具有较强的不确定性;另一方面,有些因素的影响作用较强,有些则影响作用较小。因此,在建立预测模型时,通常需要筛选影响因素,以便在确保预测精度能够达到满意效果的同时,能够以较小的代价获取可靠的影响因素数据。然而,对于复杂的交通演变,其影响因素可能会随着时间的推移而发生变化,还有一些影响因素可能因测量精度或误差造成数据质量不高,从而导致预测的精度无法保证或显现出时好

时坏的动态变化。

在智能交通系统(ITS)中,不仅实时交通流对提供动态的交通控制和引导是非常重要的,而且生动和准确的交通流预测可以帮助减少意外故障,提高交通系统的效率。然而,道路交通系统是一个复杂的、开放的、并随时间变化的系统,该系统通常表现出高度随机性和不确定性。这个具有挑战性的问题激发了人们相当大的研究热情,不断地致力于这一领域进行探索。因此,已经开发了诸多的预测算法,诸如卡尔曼滤波器及其扩展、支持向量机、贝叶斯网络和混合方法等。从不同的角度可以对这些预测算法进行多种不同的分类,如单道路连接或相关联道路网(子网络)、市区街道或高速公路、参数或非参数模型、分析或数据驱动方法、单变量或多变量方法等。然而,这些预测方法背后的基本思想或多或少像“昨日重现”,也就是说,从历史的经验中提取信息和知识用于推断和预测未来的状态。因此,必须对过去有很好的了解,以试图发现事物进化的规律。

1.2.2 基于时间序列分析的预测理论

时间序列是事物发展过程中根据时间先后顺序而得到的一系列观测值。许多事物的发展都能以时间序列的形式呈现出来,如交通事故发生的周度序列、出行需求变化的小时序列、交通流量变化的分钟观测序列等。

时间序列里蕴藏着事物发展的丰富信息,具有其他形式所不能代替的知识,而其最典型的特征之一,便是相邻观测时刻之间存在着难以分割的依存关系。而这种在时间维度上的依赖关系具有极大的实用价值,是进行预测的根本依据。

时间序列通常是在固定时间间隔下在每一个观测时刻记录下来的观测值,这样的时间序列从数学的角度可定义为:一个时间序列就是指根据时间顺序所记录的一系列观测值$\{r_i\}_{i=1}^{N}$,N为观测值的个数。其中,每个记录可以是一维或者是多维数据,如$r_i=\{a_1, a_2, \cdots, a_m\}$是$m$维数据。而且,这些记录数据可以是连续实数,也可以是离散数据。如果记录的数据随着时间的推移而发生变化,则称其具有动态特性,否则具有静态特性。

在进行预测时,可以根据数据在时间上的相互依赖特性,在历史数据及当前数据的分析基础上,对未来时刻的数据进行推测。此外,时间序列的预测也可以建立在序列的相似性基础之上,借鉴“昨日重现”的思想,通过对历史数据分析,找出相似的历史演变片段,以此作为预测未来演变的基础依据。

交通系统的发展会受到许多因素的影响,比如天气、道路施工和大型群体事件(如奥运会)等。将这些影响因素作为交通系统的潜在输入变量,可对交通状态未来发展趋势进行预测。然而预测精度在很大程度上取决于这些影响因素数据收集、分析和处理的质量,而且其中一些影响因素较难获取或是获取的数据精度不高或是获取的数据格式之间存在着较大的差异。历史观测的时间数据序列(即时间序列)嵌入了丰富的信息,充分挖掘时间序列的内涵,可以用于推断未来的状态。综上所述,相对于基于影响因素的预测

方法来说，基于时间序列的预测方法在数据收集方面易于实现，而且统一的数据格式和采集手段便于后续处理。因此，基于历史数据的交通流预测是本书主要阐述的对象。

1.3 交通信息预测的分类

1.3.1 依据时间跨度划分

由于预测数据的最终用途存在着差异，如交通规划可能需要对未来几年后的交通演变进行预测，并以此预测数据作为制定未来建设规划的主要依据。如果是制定管理措施，可能需要预测几个月后的交通信息即可；如果是为了制定临时的交通管制或诱导策略，则需要对近期未来的交通状态进行预测。

因此，根据不同的目的，交通预测的时间也有所差别。目前，多数研究对预测时间大致划分为长时、中时及短时预测。即，长时范围有设定为数小时、数天、数月，甚至是数年等[2]；中时范围则以天计或是月计；而短时范围有从5min到15、30、60min等[3-5]。也有一些研究把预测时间划分为长时和短时两种。

由于时间序列是一组在不同时刻获取的观测值，那么根据是对下一个时刻进行预测还是对未来几个时刻进行预测，对未来的预测可分为单步预测和多步预测。其中，多步预测又可分为两种。一是直接对未来若干个时刻后的状态进行预测，这种多步预测方法比较简单，只需将历史数据与预测时刻的状态建立起关系即可，因此也被称为直接多步预测；二是建立在单步预测的基础上，在每次单步预测之后，再将预测的结果作为已知输入，对下一时刻进行预测，如此循环滚动，直至到所需的预测时刻[6]。

1.3.2 依据空间跨度划分

除了可以按照预测时刻离当前时刻的远近关系对交通预测进行分类外，还可以从预测所涵盖的范围来进行划分。

(1)城市道路交通预测。

随着我国城镇化发展的步伐逐步加快，城市交通成为城市可持续发展的重要支撑和保障环节。由于城市经济的快速发展，城市交通问题日趋凸显，交通预测在城市规划、交通规划、交通管理等方面具有重要的参考价值。城市交通不同于公路交通，在城市交通系统中，交叉路口的信号灯控制系统在影响和调节城市路网的交通流量上起着重要的作用。

(2)公路交通预测。

公路承载着城市外部的交通，成为省际、城际及城乡交通的重要载体之一。城市间经贸活动的日趋频繁，相应的物流需求迅速增长，公路在货物运输方面承担着重要的角

色。随着我国居民汽车拥有量的快速增长,跨省市的自驾出行需求增长迅速,尤其是在重大节假日期间高速公路的拥堵问题十分突出。因此,交通预测在缓解公路拥堵和保障运输顺畅方面起到了积极的作用。

对于城市交通系统来说,交通预测还可分为:路段和路网两种类型。其中,对路段进行交通预测时,主要针对的是基本路段上的交通参数(如交通流量、行车速度、占有率等),而一般不涉及交叉路口的交通情况。另一方面,在对路网上的交通参数(如行程时间等)进行预测时,就需要考虑交叉路口的影响,例如在预测行程时间时需要考虑交叉路口的延误。此外,由于路网中相邻路段(如被预测路段的上下游等)之间存在着一定的相关性,可以利用这种空间相关特性来进行预测。路网的交通参数可以先从各路段预测的交通参数通过集计的方式获得,但也可从宏观的角度通过分析相关的影响因素来预测未来时刻的交通参数。

1.3.3 依据预测方法划分

按照预测方法可将现有的方法分为以下6种。

(1)基于统计理论。

通过统计历史交通数据得出其固有规律,通常假设未来走向具有与历史数据相同的特性。代表性方法有历史平均模型(History Average Model)、线性回归预测模型、时间序列预测模型、卡尔曼滤波模型(Kalman Filtering Model)[7]等。大多数方法是建立在线性基础上,在时变性强的情况下,预测效果不佳。

(2)基于交通仿真。

建立仿真模型进行预测[8]。仿真模型可分为3种:基于连续性描述的流体力学模型、基于概率性描述的气体动力模型、基于离散性描述的跟驰模型和元胞自动机模型。从物理角度也可认为是宏观、介观和微观模型。宏观模型便于把握交通流的整体特性,简化后容易得出解析解,但不适于非线性强的情况。微观模型虽然捕捉到交通系统的离散性和非线性,但模型参量很难客观准确确定。介观模型在考虑交通整体特性的同时有所涉及系统中的离散个体,但是待定参数增多且确定烦琐。

(3)基于模式匹配。

从历史交通数据中总结出一组模式,把当前采集的数据与历史模式进行比对,找到最为接近的一个或几个模式作为预测的依据[9]。具有代表性的方法是非线性回归法。该方法不需要建立模型和先验知识,直接从历史数据中挖掘信息,因此较适合有特殊事件发生时的预测,但不易提炼总结潜藏的规律,且需大量数据以便建立完整的数据库。

(4)基于神经网络。

通过对历史数据学习确立网络模型,再利用建成的神经网络进行预测[10]。因其具有自学习的特点,近年来得到广泛关注。不足之处在于:建模需要大量的历史数据;经学习

确定的输入与输出关系不易被理解；由于利用经验最小化原理进行学习，因此不能实现期望风险最小化，这是其理论上的缺陷。

(5)基于非线性理论。

此类方法是在混沌理论、耗散理论、协同论、分形理论等非线性理论的基础上形成的[2][7]，其中研究的热点较集中在基于混沌理论的预测方法和小波分析预测法。此类方法对于短时内非线性较强的预测结果不错，但不适于中长时预测。

(6)综合预测方法。

各种预测方法均有优缺点，综合预测方法试图借助融合技术综合多种不同预测方法，达到“取长补短”的效果[11,12]。研究重点是如何有效地把多种方法结合起来，如何分配及调整各种方法的权重。

上述预测方法还可以按照是否建立模型分为两类：基于模型的预测方法，主要包括历史平均模型、线性回归模型、时间序列预测模型、卡尔曼滤波模型、仿真模型、模糊预测模型及它们的组合预测模型等，此类方法通过建立近似模型进行预测，起到平滑作用的同时有助于发现交通系统中的潜在规律，但需确定参数；无模型的预测方法，主要包括非参数回归、小波预测法、基于混沌理论的预测方法和它们的复合预测方法等，无模型方法直接从数据中寻找信息，适于突发事件的预测，但需大量数据支持且抗干扰能力较差。

1.4 短时交通信息预测的意义

交通运输系统是一个极其复杂的开放系统，运输需求同时受来自系统内和系统外因素的影响，而且这种需求可在任意时间任意地点产生和消除，因此，具有较强的不确定性。一方面，对交通需求进行预测，可以为制定更加合理的政策及规划提供有力的依据。另一方面，对交通流等信息的预测，可以在实际交通运行中起到重要的作用[13]。本书将重点介绍作者在以往科研项目中对交通信息预测所提出的一些数据预处理方法和预测方法。

第2章 数据预处理

2.1 数据预处理概述

交通系统是一个高度复杂的开放系统，具有较强的非线性、时变性及不确定性。在进行交通参数及状态检测时，由于存在着种种干扰，所采集的数据往往含有异常数据、缺失数据、噪声等。这些数据在一定程度上掩盖了交通运行的本质特征，使得其演化趋势不易被洞察。因此，为了确保交通预测的准确性及可靠性，对采集的原始数据进行预处理是一个不可或缺的重要环节。然而，数据预处理尚没有被广泛认可的统一标准，通常需要根据不同数据的数据类型及业务需求，结合专业领域知识，在对数据特性进行充分理解的基础上，再选择相关的数据预处理技术或提出改进的预处理方法。此外，对原始数据的预处理也需视具体情况而定，通常会涉及异常数据的祛除、降噪处理及缺失数据的弥补等，但也可能存在着原始数据只包含其中的一种或两种类型的错误数据。

2.2 异常数据祛除

2.2.1 异常数据

异常数据又称为异常值(Outliers)。有关异常数据的研究可追溯到18世纪，而对异

常数据的最早定义是由 Bernoulli 在 1777 年提出的，但其对异常数据的定义并没有得到广泛认同，而且在此之后又有一些新的定义相继提出。如：Hawkins[14] 认为“异常值就是指那些在数据放到一起后与众不同的数据，使人们质疑这些数据并不是随机偏差所产生，而是由于完全不一样的机制，而这在一定程度上表明了异常值的本质”；Weisberg 结合统计理论，将异常数据的定义是“与集中数据后的其他部分不服从相同的统计模型的数据”；而 Grubbs[15] 则把异常数据定义为“一些明显偏离其余样本的数据，而它们是不符合常见的数据模式”；Beckmen 与 Cook 于 1983 年[16] 提出了关于异常数据的两种看法：一是“将异常值看成是那些与数据集明显不协调的、令人吃惊的数值点，这样异常值就可以解释为是假设分布里的极端值”；二是“将异常值看作杂质点，它与数据集不是同一分布的，是在大多数来自某一同分布的数据集中混入的来自另一分布地少量杂质”。

我国学者张德然[17] 在总结前人定义的基础上，从内涵关系的角度出发，提出了有关异常数据的两种定义，即广义定义和狭义定义。在其广义定义中，将异常数据认定为：“在所获统计数据中相对误差较大的观察数据”，而把这样的数据也称为是奇异值；在其狭义定义中，则将异常数据认定为：“一批数据中有部分数据与其余数据相比明显不一致”，这种数据也被称为是离群点。

数据出现异常的原因很多，其中既有人工采集数据时因操作失误或使用方法或设备不恰当所引起部分数据出现较大偏差，也有因采集手段及设备的局限所造成的。由于异常数据的存在，迷惑或掩盖了事物发展的本质特征，同时也无形当中给处理和分析带来诸多不便。但是，也应注意到，一些异常数据可能蕴含着十分有价值的信息。因此，如何识别并祛除异常数据一直以来都是数据预处理当中的一个必要和重要的环节。

如果所采集的数据量并不大，同时数据的使用目的对实时处理并不作过多的严苛要求，那么此时采用较为原始的人工识别及祛除方法具有一定的可行性，但是可能会受到处理人员主观认知的影响较大。近年来，随着检测技术的普及，许多城市道路上都普遍安装了各种检测器，由此而带来的是海量数据。此外，交通管理和控制等诸多方面对实时数据处理的要求也越来越严格。因此，人工识别及祛除异常数据的方法无法再适应目前发展的需求。以下将首先介绍两种传统的异常数据自动祛除方法，再介绍我们所提出的一种基于密度的异常数据祛除方法。

2.2.2 异常数据祛除的传统方法

目前使用比较普遍的两种异常数据识别及祛除方法都是建立在统计理论基础上的，并且通过一定的前提假设来实现对异常数据的判别。

(1)3δ 准则法。

这种方法建立在假设数据是服从单一高斯分布的基础上，基于所采集的样本，先对样本的期望值进行估算，然后将那些偏离期望值超过一定程度的数据(通常取偏离期望

值超过标准偏差的3倍的数据，即Pauta准则，又称为3δ准则）判定为噪声，进而把这些噪声剔除掉[18,19]。显然，这种异常数据祛除方法属于参数统计法的范畴。由于这种方法是基于数据服从高斯分布的假设基础之上，因此，对于服从其他分布的数据，这种方法无法保证能够有效地识别噪声。此外，由于异常数据的影响，计算出来的期望值及标准偏差等统计量可能存在着不同程度的误差，这些误差也是造成异常数据误判的一大原因。

（2）Tukey测试方法。

首先将采集得到的样本数据从小到大进行排序，再依据四分位数法，将这些数据分为四等份，那些偏离上四分位数和下四分位数超过一定程度（通常设定为偏离上四分位数和下四分位数1.5倍的上下四分位数范围）的数据则被判定为异常数据[20,21]。该方法属于非参数方法。

以上这两种异常数据祛除方法均局限于单个分布的情形，也就是说对于服从多个相同或不同分布的数据，以上这两种方法可能存在着无法正确识别和祛除异常数据的情况，或者存在着漏判等其他情况，甚至可能将那些能够体现实质的数据也当作异常数据剔除掉了。

近几年，有些研究者[22,23]通过聚类来识别并祛除噪声，但是聚类本身就是一个复杂耗时的过程，而且聚类结果的好坏直接影响着噪声的识别。

2.2.3 一种基于密度的异常数据祛除方法

由于数据之间的紧密关系可以通过数据聚集的程度（即密度）来判断，因此，可以通过计算数据的密度来判断哪些数据与其他数据相距较远或相似性较低。正是基于这样的思想，建立了一种基于密度的异常数据祛除方法。该方法的具体计算过程如下。

该方法可大致分为两步，第一步先对数据密度进行估算，下一步则依据所得的密度进行噪声识别并作祛噪处理。可以看出，第一步即计算数据的密度，是识别噪声的基础和关键。由于常用的交通参数时间序列可认为是一个二维数据集，故这里以二维数据的异常值祛除为例进行说明。

首先假设采集得到二维数据集为Z（如图2-1a）中的圆点所示），且其包含了N个数据点$z_k(k\in\{1,2,\cdots,N\})$（在本例中数据点的维度为2）。接着，产生一个数目为M的数据集S（如图2-1a）中的圆圈所示），为了便于描述，这里将该数据集称为种子群。在确定种子群的数目M时，不仅需要保证各个种子点与其相邻种子之间的距离恒等，还需保证种子群的范围能够覆盖待处理的数据集。并且，给每个数据点$z_k(k\in\{1,2,\cdots,N\})$均附有一个初始值为0的种子吸附计数器$c_k$。该种子吸附计数器的作用就是统计该数据点在与其他相邻数据点之间竞争种子时所获得的种子数目。为了判断数据之间在争夺种子中的优胜情况，特设立如下准则：对于每个种子$s_k(k\in\{1,2,\cdots,M\})$来说，先分别根据