基于公交车 GPS 数据的
城市道路偶发性拥堵检测与系统实现

重庆大学硕士学位论文

（专业学位）

学生姓名：崔德冠

指导教师：廖孝勇　　讲师

学位类别：工程硕士（控制工程领域）

重庆大学自动化学院

二 o 一五年四月

The Detection and Sytem Implementation of Urban Road Contingency Jam Based on Bus GPS Data

A Thesis Submitted to Chongqing University

in Partial Fulfillment of the Requirement for

Professional Degree

By

Cui Deguan

Supervised by Lecturer Liao Xiaoyong

Specialty： ME (Control Engineering Field)

College of Automation of Chongqing University

Chongqing, China

April 2015

摘　　要

城市道路偶发性拥堵检测可为交通诱导、应急处理等提供支持和参考，有助于提高现代交通管理和服务水平。城市道路偶发性拥堵具有偶然性，发生的时间、地点不确定，现有的算法主要用于高速路，在城市道路偶发性拥堵检测方面难以取得满意的效果。近年来智能交通领域积累了海量的公交车 GPS 数据，这些数据覆盖面积广、实时性高、维护成本低和可靠性高，包含丰富的道路交通拥堵信息，反映了道路交通拥堵的变化过程，可为其偶发性拥堵检测提供支持。但如何利用这些数据检测城市道路偶发性拥堵，目前还缺乏有效的研究。

因此，论文以海量公交车 GPS 数据为基础，研究如何检测发生在城市道路上的偶发性交通拥堵。论文首先分析道路交通状态的历史情况，然后分析道路交通状态的实时变化趋势，结合历史情况和实时变化趋势，以检测城市道路偶发性拥堵，并进行相应的软件系统开发实现工作。因此，论文主要研究了以下 3 点内容：

① 研究了基于公交车GPS数据的城市道路偶发性拥堵检测系统体系结构，将主要工作划分为若干模块，明确各模块的主要内容和各模块之间的关系；并在最后开发和实现基于公交车GPS数据的城市道路偶发性拥堵检测系统。

② 研究了如何利用海量公交车GPS历史数据分析道路交通状态规律的技术：定义“路段延误时间指数”用以表征道路交通状态；针对交通偶发性拥堵的相对性和情景性的特点，提出基于T-检验的K-均值自适应聚类算法进行道路交通状态模式识别；引入“四分位差”的方法明确区分交通平常状态与偶发性拥堵；

③ 研究了道路交通状态实时变化趋势分析与偶发性拥堵检测的技术。针对单一的瞬时速度不完全可靠的特点，提取了 4 个速度参数用以表征道路交通实时状态；引入 CVA（Canonical Variate Analysis，规范变量分析）算法用于检测偶发性拥堵：分析道路交通状态的实时变化趋势，通过比较交通状态实时变化趋势与历史阈值，从而实现偶发性拥堵的检测。

在软件系统实现的基础上，利用实际公交车 GPS 数据进行测试。测试结果表明，该系统不仅能实现城市道路偶发性拥堵的检测，在误报率低于 35%的情况下，检测率达到 90%，平均检测时间为 3.2 分钟，能取得较为满意的效果，可为城市道路交通管理和服务提供支持，具有实用意义。

关键词：城市道路偶发性拥堵，公交 GPS 数据，K-均值自适应聚类，规范变量分析

ABSTRACT

Road traffic contingency jam detection can provide support and advice for traffic guidance, emergency treatment, and so on, so as to improve the traffic management level and service level. However, road traffic contingency jam are those random events, when and where they will happen are highly uncertainty. Therefore the existing algorithms can't obtain acceptable results. In recent years, a huge number of bus GPS data began to accumulate in the field of ITS. These GPS data have many advantages, such as wide area covering, high real time, low maintaining cost and high reliability. Bus GPS data can reflect changes in the process of road traffic states, because of the date Contains abundant information of road traffic states. And without any doubt, it can surely provide something help for road traffic contingency jam detection. Here comes the question, how can we use the GPS data to detect traffic contingency jam, especially on urban road traffic? There is little effective study for the question.

Therefore, based on the huge number of bus GPS data, this paper is aimed to study urban road traffic contingency jam detection. The paper first analyses the historical rule of road traffic states. Next, it focused on road traffic states' real time variation tendency. By means of road traffic states' historical rule and real time variation tendency, it can realize road traffic contingency jam detection. Based on the method, this paper designs and implements the detection systedem.Hence, this paper mainly focused on 3 research word.

Firstly, this paper studies the system structure for road traffic contingency jam detection base on bus GPS data, which divides the main work into several parts and make the main work of each part and the relationship between then more clearly. And in the last, the paper designs and implements the urban road contingency jam detection system.

Secondly, the paper studies on developing a method which can help to analyze the rule of traffic states based on huge number of GPS historical data. As researches showed that instantaneous velocity in bus GPS data can't confidently reveal traffic states, the paper defines a variable which named 'Road Delay Time Index', shorted for RDTI. AS RDTI is more exact to describe traffic states, is will be the key variable in the follow-up work. Aimed at the relativity and conditionity of traffic states, the paper puts forward K-Means self-adaptive clustering algorithm base on T test, which is used for traffic

conditions classification. Using the algorithm, we can divide traffic condition into 8 classes, and therefore determine the characteristic of traffic condition. And in the meantime, introduce ‘interquartile range’, a common statistics concept, to divide normal traffic state and abnormal traffic state clearly, which will play a significant role in training threshold and detection result assessment.

The last main research of the paper is road traffic states’ real time variation tendency analyze and road traffic contingency jam detection. This paper firstly draws 4 different variables from bus GPS data, which can describe the real time traffic states more fully and exactly. Besides, this paper introduces Canonical Variate Analysis (CVA) into road traffic contingency jam detection process. The algorithm using Squared Prediction Error (SPE) as parameter to analyzes the road traffic states’ real time variation tendency, and uses historical traffic contingency jam to train the threshold. By comparing the real time variation tendency and threshold in a number of periods, we can determine whether an abnormal traffic state has happened.

Based on the detection system, the paper uses real GPS to test the contingency jam detection result. Final result shows that on the condition of False Alarm Rate (FAR) under 35%, Detection Rate(DR0 reach to 90%, and the Mean Time To Detection(MTTD) is less than 3.2 minutes. This result can be helpful for urban traffic management and service, and has practical value.

Keywords: Urban road contingency jam detection, Bus GPS data, K-Means self-adaptive clustering algorithm, Canonical Variate Analysis (CVA)

目　录

1 绪　　论

1.1 课题背景

随着社会经济水平以及现代科学技术的发展，我国城市化水平发展不断深入，城市人口数量不断增加，同时各种交通机动车辆数量快速增加，造成了一系列的社会问题。目前，交通拥堵、资源浪费，环境污染等问题日益突出，虽然各级交通管理部门采取了多方面的措施，但是交通运行效率与管理水平仍亟待提高。其中主要的解决方法之一是不断修建交通道路，提供更多的交通基础设施。但是由于基础设施投资巨大，建设周期长，同时受土地资源以及资金等因素等影响，仅仅依靠加大交通道路基础设施建设无法缓解以及彻底解决当前交通压力。目前城市交通基础道路实施的增加速度仍低于机动车辆的增长速度，我国公安部交通管理局 2015 年 1 月 27 日公布数据表明，直到 2014 年末，在我国，汽车保有量共有 1.54 亿辆，机动车保有量总共达到 2.64 亿辆，国内包括重庆市在内的 10 个城市汽车保有量超过 200 万辆，我国交通供需矛盾尖锐。在此背景下，当道路出现偶发性拥堵时，进一步加剧了城市交通发展的矛盾。因此，如何在现有交通基础设施的条件下提高道路交通运行效率及管理水平是当代交通领域面临的主要问题。

除了加大交通基础设施建设之外，还需要利用先进的科学技术和管理手段来提高交通道路运行效率和服务水平，而智能交通是其中一种切实可行的方法。这是目前国内外综合全面治理、解决交通运输问题的主要手段[1][2][3]。所谓智能交通（Intelligent Transportation System，简称 ITS），是将先进的信息技术、数据通讯传输技术、电子控制技术以及计算机技术等有效地集成运用于整个交通运输管理体系，而建立起的一种在大范围内，全方位发挥作用的，实时、准确、高效的综合的运输和管理系统[4][5]。ITS 主要研究的范围包括：先进的交通管理系统，驾驶员信息系统，车辆控制系统，运营车辆调度系统，公共交通系统，乡村系统以及自动公路系统等，其关键技术包括：传感器、电子视频图像处理、位置测量、数据处理、电子数字及移动通讯技术等[4]。ITS 可为应用交通诱导、交通道路维护、车辆运营管理、交通拥堵判别与监控等日常管理、维护和服务提供支持。

随着智能交通技术的研究与应用，目前我国城市交通水平不断提高，从理论研究、观念、技术应用、管理水平、产业发展等方面都可以看出有明显的提高与进步。相关研究表明，先进的智能交通技术可以有效提高道路交通管理和服务水平，可以大幅度降低因交通事故造成的死亡人数比例和提高交通工具的使用效率[6]。《2013 年中国城市智能交通市场研究报告》指出，2012 我国智能交通的市场规模达到了 159.9 亿元，同比增长 21.7%。智能交通技术为提高现代经济发展水平和

社会生活水平做出了巨大的贡献，大大方便了社会大众的出行。但必须承认，目前道路交通管理还存在不少问题亟待进一步的研究解决，而其中包含了道路交通偶发性拥堵检测的问题。

道路交通偶发性拥堵是由发生在道路上的意外的交通事件引起，会造成交通大范围拥挤，增加交通参与者出行成本，严重情况下还会造成二次交通事故的发生，造成生命和财产损失。及时、有效地检测道路交通偶发性拥堵，可为道路交通诱导、车辆调度、应急处理等提供支持与参考，有利于进一步提高道路交通的管理水平和服务水平。但如何实现对道路交通偶发性拥堵（尤其是城市道路交通偶发性拥堵）的检测，是目前亟待解决的问题。

1.2 道路交通拥堵概述

本质上，交通拥堵是一种交通状态，而偶发性拥堵为一种异常状态。交通状态包括正常交通状态和偶发性交通拥堵状态。研究城市道路偶发性拥堵首先需要明确道路交通状态及道路交通偶发性拥堵（Road Traffic Contingency Jam）的含义。

1.2.1 道路交通状态的描述

道路交通状态是一个带有主观色彩的概念，不同的交通参与者对交通状态的感受、理解和定义可能有较大的差异。即使是同一位交通参与者，也可能会在不同的时间、地点以及心态下对交通状态的产生不一样的感知度。因此，虽然国内外学者纷纷提出了不同的定义和解释，但目前还没得到统一的标准。

除了“交通状态”一词的表述之外，与之相关的表述有交通拥挤程度、交通服务水平等。交通状态的含义包括以下 2 点：其一，道路交通的客观情况，一般以车流量、占有率、车速等参数描述，各参数不同的取值反应了不同的道路交通状态；其二，交通参与者的主观感受，乘客、驾驶员、交通管理人员等交通参与者对交通状态的判别标准也不尽相同。

1.2.2 道路交通拥堵程度的划分

如何划分道路交通拥堵是国内外研究的重点。目前国内外对交通拥堵的划分有多种，不同国家之间没有统一的标准，甚至是在同一国家中也存在多个不同的标准，如我国的交通拥堵划分标准是《城市交通管理评价指标体系》中的规定，而北京等地也纷纷针对自身具体的情况对道路交通拥堵的划分作了规定。

首先，在国家标准方面，其中比较著名的有美国的《道路通行能力手册》(Highway Capacity Manual，HCM)中规定的道路交通服务水平(Level of Service)，它在考虑车辆之间的运行条件及其驾驶员和乘客的主观感觉的基础上将高速公路服务水平从低到高依次分成了 A、B、C、D、E、F 等 6 个等级[7]。日本以 HCM 作为参考，根据该国车流量大而道路资源紧张的具体情况提出了 3 个等级的道路交

通服务水平量化标准，依次类似于 HCM 的 C、D、E 三个等级[8]。德国以交通密度为依据，把快速干道的服务水平划分为 5 个区，其中 I 区的交通密度为 0~10辆/KM，II 区的交通密度为 10~20辆/KM，III 区的交通密度为 20~30辆/KM，IV 区的交通密度为 30~40辆/KM，V 区的交通密度>40辆/KM，在 I~III 区的范围内，车流量呈现出稳定的拥堵，从 IV 区开始，道路上车辆量不稳定，各车行驶速度受到相互制约[9]。

其次，在中国，最著名的是公安部 2002 年发布的《城市交通管理评价指标体系》[7]，该体系以机动车在城市主干路上的平均行程速度为参考，将道路交通拥堵程度划分成了 4 个等级，具体如表 1.1 所示：

表 1.1 我国道路交通拥堵等级划分

Table 1.1 Road traffic states classification of China

交通拥挤程度	平均行程速度（单位：km/h）
畅通	>=30
轻度拥挤	[20,30)
拥挤	[10,20)
严重拥挤	<10

其次，在地方标准方面，北京市质量技术监督局的标准首先将城市道路划分为快速路、主干路、次干路和支路 4 种，然后根据路段的平均行程速度将每种道路的交通拥堵划分为 5 个等级，具体如表 1.2 所示：

表 1.2 不同路段拥挤等级划分（单位：km/h）

Table 1.2 Traffic congestion classification on different roads（km/h）

	等级				
快速路	>65	(50,65]	(35,50]	(20,35]	≤20
主干路	>45	(35,45]	(25,35]	(15,25]	≤15
次干路	>35	(25,35]	(15,25]	(10,15]	≤10
支 路	>35	(25,35]	(15,25]	(10,15]	≤10

可以看出，北京的标准是《城市交通管理评价指标体系》的进一步细化。

还有，在行业标准方面，中国公路学会 1998 年出版的《交通工程手册》中，提出了基于 Q－v－H 关系模型的混合交通的交通拥堵评价体系，具体如下表所示，其中 Q、v、H 分别为交通量、平均速度和交通熵 $Q_{基}$、$v_{基}$、$H_{基}$ 分别为对应的基准

值，如表 1.3 所示。

表 1.3 混合交通流的交通拥堵评价体系

Table1.3 Evaluation system of Mixed traffic states

交通拥堵	范围
I 型	$Q \le Q_{基}, H \le H_{基}, v \ge v_{基}$
II 型	$Q > Q_{基}, H \le H_{基}, v \ge v_{基}$
III 型	$Q < Q_{基}, H > H_{基}, v \ge v_{基}$
IV 型	$Q \le Q_{基}, H > H_{基}, v < v_{基}$
V 型	$Q > Q_{基}, H < H_{基}, v < v_{基}$
VI 型	$Q > Q_{基}, H > H_{基}, v < v_{基}$

由以上分析可以看出，不管是国家标准还是地方标准，或者行业标准，主要都是以速度或者交通密度作为交通拥堵的划分依据，但是由于不同国家（地区）的具体情况不同，对交通拥堵的划分也不尽相同，或分类数目不同，或每类范围不同，或分类数目和每类范围都不相同。同时，在划分过程中均考虑了交通参与者的主观因素。因此，对交通拥堵的划分不能一概而论，关键在于是否符合当地的实际情况，是否具有实用意义。

1.2.3 道路交通偶发性拥堵分析

道路交通偶发性拥堵和正常交通拥堵程度相对。道路交通偶发性拥堵随机性大，发生的时间、地点及影响程度都具有高度不确定性，而且有的交通偶发性拥堵发生之前没有明显可观测的征兆，或者其征兆信息没有得到及时的关注和处理，就会造成偶发性拥堵的发现和处理不及时，会引发一系列后果。对社会大众而言，轻则引起堵车甚至二次拥堵，增加出行者的时间成本，重则引发交通事故，造成生命财产损失（如损害车辆及交通基础设施，人员伤亡等）。交通拥堵包括了常发性拥堵和偶发性拥堵。对常发性拥堵，可以利用大量的历史数据进行检测及预测，找出其发生及其变化的规律；而偶发性拥堵的发生则难以预测，能做到的是及时检测出来以及分析其影响。一般情况下，道路交通偶发性拥堵（Traffic Contingency Jam）由发生在道路上的异常事件（Abnormal Accident）引起[10]。道路交通异常事件包括常发性异常事件和偶发性异常事件。其中，常发性异常事件包括：天气原因（雨、雪、雾等）、节假日突发高峰流、道路常年失修或者长期施工、司机驾驶行为习惯等，而偶发性事件包括：交通基础设施原因（树木、广告牌或建筑物倒

塌等)、驾驶人员原因、特殊大型活动、临时交通管制、交通事故（撞车、货物散落、车辆故障）等。

尽管引起交通偶发性拥堵的原因各异，而且拥堵的严重程度也不一样，但是如果没有得到及时、有效的处理，则会造成更加严重的后果。因此，为了减轻道路交通偶发性拥堵所带来的负面效应，首先需要实现道路交通偶发性拥堵及时、准确的检测，并在此基础上采取有效的应急措施。然而，如何及时检测交通偶发性拥堵的发生，这是一个首先需要解决的问题，也是本文研究的重点。

通过对国内外相关研究的分析总结，本文对道路交通偶发性拥堵作如下定义：道路交通偶发性拥堵是指在道路上由于发生了异常事件而造成或者即将造成交通运行缓慢，导致道路通行能力下降，引发交通延误甚至是生命财产损失的情况。

对应于常发性异常事件和偶发性异常事件，道路交通拥堵也可分为常发性拥堵和偶发性拥堵。其中，常发性拥堵主要是受气候、出行规律、道路特性、驾驶行为习惯等固定的或者具有规律性的因素的影响，通过对大量数据的深入分析可以得到其中的规律，进而可以进一步提出改善或者应对措施。

而偶发性拥堵则主要受意外事件的影响而发生，这是本文研究的重点，在后文中（除特别说明外）所说的道路交通拥堵均指偶发性偶发性拥堵。道路交通偶发性拥堵发生的随机性大，其发生的时间地点难以预测，而且内部机理不明确，难以建立精确的数学模型进行描述分析。尤其是在城市道路上，由于城市道路交通拥堵受早晚高峰期的影响，其交通拥堵会出现上下浮动的情况。在路况变化不定的情况下研究道路交通偶发性拥堵更加困难，而也更有必要。因此，在后文中如果没有特别说明，本文所研究的道路交通偶发性拥堵是指发生在城市道路上的偶发性偶发性拥堵。此外，道路交通偶发性拥堵具有异常点的特征，主要包括以下三个特点[11]：

① 相对性，即道路交通的偶发性拥堵是相对与正常拥堵而言的。正常情况下道路交通拥堵受早晚高峰期和平峰期的影响而上下浮动，相同的拥挤程度在高峰期间属于正常现象，而在夜间则很可能是异常现象。如公交车某次经过某路段的时间为 5 分钟，但在其他很多路段上可能属于正常的现象，不是异常的交通拥堵，而在平时几乎不会发生拥堵的路段上很可能就是发生了异常。

② 情景性（或条件性），即道路交通偶发性拥堵与其所处的情景（时间和地点）有关，在不同的时间，不同的地点，偶发性拥堵的拥挤程度不同。如在交通拥挤常发路段，一次的交通拥挤属于正常的拥堵现象，而在一直保持畅通的路段上发生了拥挤，则属于一种异常。

③ 集群性，即道路交通偶发性拥堵是一种连锁反应，偶发性拥堵下不仅仅是某一辆车的拥堵发生异常，而且是多辆车在连续一条路段、连续一段时间的拥堵

也会发生异常。

充分考虑这些特点，有助于实现及时、准确地检测道路交通偶发性拥堵，从而降低偶发性拥堵所造成的负面影响。

1.3 国内外研究现状

及时、准确识别道路交通的偶发性拥堵，对道路交通系统实现及时、有效的引导和控制，可以提高城市交通的运行效率与管理和服务水平。先进的交通拥堵检测技术是有效防止和缓解交通拥堵、实现道路交通流在路网上的合理分布的方法，目前在各国得到了较大的发展。然而先进的交通拥堵检测需要以大量的交通信息为基础，因此交通信息采集技术是实现及时、准确、有效的交通拥堵检测的基础。目前，关于交通信息采集技术可以概括为两类：固定的交通信息采集技术和移动的交通信息采集技术。

1.3.1 交通信息采集技术现状

① 固定交通信息采集技术

固定交通信息采集技术是一种将交通检测器设置在道路的固定地点，对通过特定地点的车辆交通参数信息进行采集和检测的技术。其中按照工作方式和电磁波长范围分类，具有代表性的有：磁频采集技术、波频采集技术以及视频采集技术三种类型[12]。固定检测器收集到的交通信息主要包括：时间，检测器编号，检测地点，检测对象，方向，车道占有率，速度，平均车头距等。目前交通管理部门仍在很大程度上依赖于固定检测器收集的道路拥堵信息，但固定检测器并非在适应所有环境条件的要求，其不足之处主要在于：1）天气条件的限制。固定交通信息检测主要技术的环形感应线圈和磁强器在白天和黑夜以及大部分天气条件下都能达到很好的检测效果，但是在下雪天气中难以达到检测精度要求；而被动声传感器则相反，它能在雪天发挥很好的检测效果，但是在晴天的检测效果不明显[13]；2）大范围安装检测器成本很高，而且大量检测器安装和维护工作十分复杂，需要耗费大量的人力物力[14]；3）检测器位置固定，相邻两个检测器之间的距离较远，只能收集到有限空间范围的交通数据，难以满足智能交通实时、全面控制的要求[15]。

② 移动交通信息采集技术

移动交通信息采集技术可以弥补固定采集技术的不足。移动交通信息采集技术是指在移动车辆（如公交车）上安装的终端设备通过卫星定位或检测道路特征来获取道路动态交通参数的方法。目前应用较多的主要有：基于 GPS 浮动车、电子标签或者图像识别的动态交通信息采集技术。

目前我国大部分城市广泛在出租车和公交车上安装 GPS 车载终端系统来采集

车辆信息，以便于对运营车辆和交通拥堵实现有效的监控。以公交车 GPS 数据为例，其主要信息包括：时间、车辆 IP、位置信息（如经纬度和海拔高度）、方向、所在站点、速度、里程距离等。以重庆市为例，该市 500 多条线路的近 9000 辆公交 GPS 车载终端每隔 10 秒即传回一条 GPS 数据，每天收集到的数据达 1700 万条，详细地记录了公交运行的轨迹，有效反映了道路交通拥堵的实际情况，可为交通管理和服务提供了坚实的基础。

公交车 GPS 数据的优点在于：1）数据量大，能全程采集公交数据，采集时间间隔短，信息量充足，覆盖时空范围广；2）无需大量车辆检测器，投资维护成本较低；3）受环境影响较小；4）数据接受采集终端安装在移动车辆上，不需在道路上新增设施，对环境无负面影响；5）采集信息能实时反映车辆拥堵及道路运行拥堵信息等[16]。其不足之处主要是：1）受城市高大建筑物、地下通道或隧道的遮挡，存在一些 GPS 检测“盲区”，导致部分地方数据缺失；2）受 GPS 车载终端故障影响，传回数据可能出现错误；3）GPS 数据精度要求较高，但受网络传输的影响，接收到的 GPS 会出现延迟。但美国 ADVANCE 系统研究表明，GPS 数据比环形线圈数据能提供更加精确的参数，可得到更加精确的行程时间，在 50000 个检测报告中，99.4%是可靠的[17]。

通过以上的分析，不难看出：1）公交车 GPS 数据数量巨大，蕴含丰富的道路交通规律信息，可以从中挖掘出有效的信息为道路交通管理提供支持和服务；2）公交车 GPS 数据可靠性高，从中提取出来的信息更加可靠、有效；3）公交车 GPS 数据覆盖范围广，维护成本低，充分发挥公交车 GPS 数据的作用可以有效降低减少对固定检测设备的投入，从而可以达到节约资源降低交通管理成本效果。

1.3.2 国内外文献综述

目前国内外在道路交通偶发性拥堵研究方面主要是针对异常事件的研究，包括异常事件的检测、判别和预测等。由于道路交通偶发性拥堵一般由异常事件引起，因此，对偶发性拥堵的检测实际上就是对异常事件的检测。道路交通偶发性拥堵检测需要在研究交通模型的基础上，选择有效的参数，建立交通偶发性拥堵的检测算法，在交通偶发性拥堵发生前，可以对潜在的、将要发生的偶发性拥堵进行预测；在交通异常发生后，迅速确定其发生的时间、位置及其影响程度，从而可以提高城市道路的交通管理与服务水平[18]。经国内外不少专家学者长期的研究实践，发展出了许多有效检测道路交通偶发性拥堵的方法，其中应用较为广泛的有：ARIMA 时间序列算法、神经网络算法、加州算法、聚类算法、支持向量机、主成分分析、D-S 证据理论等。这些算法各有优劣，都能实现对交通偶发性拥堵的有效检测，但其侧重点有所不同。

① 交通拥堵检测算法分类

可以根据算法的数据对象、用途、范围等方面属性对这些算法进行分类，其中较为常见的交通异常事件检测方法分类有：

1）根据所应用数据来源的不同可分为基于固定检测器检测算法技术、基于浮动车 GPS 数据的检测技术和基于多源数据的检测技术等三类[19]；

2）根据交通异常事件检测技术可以分为自动检测技术(Automatic Incident、Detection，AID)和非自动检测技术(NAID)两类[20]，或者分为不借助模型的人工检测方法和通过模型的自动检测方法两类[21]，而自动检测技术又可以进一步分为直接检测算法和间接检测算法；

3）根据研究的范围可以分为宏观、中观、微观三类[22]；

4）从参数的类型来看可以分为离散型和连续型 2 类。

以下将针对各种已有算法进行逐一介绍。

② 交通拥堵检测算法综述

1）加利福尼亚算法，即著名的“加州算法”。它是最早用于道路交通异常事件检测的算法之一[23][24]。它是由美国加利福尼亚州的运输部在上世纪 60 年代末开发的一种模式识别算法。它以道路占有率参数，通过比较某一路段及其上下游之间的占有率，当该路段和上下游的占有率均超过一定的阈值（如 20%），则认为发生了异常事件。该算法原理简单，易于理解和实现，但由于只用到占有率一个参数，没有考虑车流量和车辆速度的问题，对交通偶发性拥堵的判别不够全面。

2）基于贝叶斯定理的单因素检测算法。该算法由 Levin M 和 Krause G M 等人在 1978 年提出[25]，它同样以占有率作为交通流特征参数，需要计算上下游占有率的差分比，通过上游占有率和历史事件信息的比较判定事件发生的概率，从而对道路交通事件进行检测。这种算法利用概率论的思想对交通异常事件进行检测，其不足之处在于：其一，需要知道先验概率；其二，对发生频率较低的交通异常事件的预测效果不尽人意。

3）标准偏差算法(SND)。这是 Dudek C L 等人在 1974 年提出并经其验证的一种利用交通参数（如占有率）的标准偏差进行交通事件自动检测的算法[26]，该算法首先计算出交通参数的标准差，然后将得到的标准差与预定的阈值进行比较以实现对交通异常事件的检测。若连续多个（2 个及以上）检测周期的标准偏差均超出指定范围，即判定有异常的交通事件发生。

4）自回归移动平均算法 ARIMA(Auto Regressive Integrated Moving Average)。ARIMA 是由 Box 和 Jenkins 在上世纪 70 年代提出的一种著名的时间序列算法，主要包含三个过程：移动平均过程、自回归过程以及自回归移动平均过程，该算法广泛应用于各领域研究和实践中。Ahmed 和 Cook 在 70 年代末首先利用该方法实现了交通事件自动检测[27]。在使用该算法之前需要对原始检测数据进行平滑去噪

声处理，同时需要实现对进行模型的识别与估计。

5）人工神经网络算法（ANN）。人工神经网络通过模拟人类大脑神经突触联接的结构进行信息处理。在 20 世纪 90 年代左右美国加州大学的 Chew 等人开始将神经网络算法引入交通异常事件检测的过程中。1995 年，Chew 和 Ritchie 等人将神经网络模型分成输入层、中间层和输出层，提出一种基于多层前馈网络的神经网络算法，用于检测交通事件及非交通事件的信号，能取得较好的效果[28]。

在我国，马黎和赵丽红等人在 2010 年提出了基于 BP 神经网络的交通异常事件检测模型，同时利用广深高速实际的交通异常数据进行验证。实验结果表明，该算法不仅检测速度快，还具有高检测率及低误报率等特点[18]。

6）MacMaster 算法。该算法是一种基于突变理论的模型，不仅可以检测出道路交通拥堵的拥堵，还可以检测出道路交通拥堵的类型（偶发性或常发性），在计算效率和实时运行效率等方面具有明显的优越性[29][30]。该算法在对交通拥堵进行判别的过程中综合考虑交通流、车速和道路占有率等方面因素的影响。它优点是同时考虑了占有率和流量两个指标，检测时间比加州算法短，而其不足之处在于：在对非阻塞和阻塞的定义过程中，需要考虑道路几何线形等客观因素的差异，需要对不同路段的进行重新定义[31]。

7）基于模糊数学的算法。应用模糊数学的方法对道路交通异常事件进行检测是在上世纪 90 年代初期由 Hsiao 首先提出，利用隶属度函数拓展了以往的集合理论，较好地解决了高维模型下的交通模式分类的问题[32]。随后，Chang E.C.等人在对突发交通时间的判别过程中引入了不确定推理的思想，对基于模糊逻辑的交通事件检测算法作了改进，在数据不够精确或者数据不完整的情况下利用模糊边界来计算出了突发交通事件发生的概率[33]。接着，Srinivasan 等人将模糊逻辑与遗传算法结合起来，提出了基于混合模糊逻辑的算法对道路交通事件进行检测。该算法对不精确和不完整的数据具有较高的容错性，可以达到较好的效果[34]。在我国，周伟等人在 2001 年在分析速度、流量及占有率等交通参数的基础上，通过对这些参数各模糊集隶属函数的计算，提出了一种基于模糊综合识别的交通事件检测算法，不仅可以检测交通拥挤，还能检测其发生的原因。

8）支持向量机算法（SVM）。支持向量机算法在交通领域应用比较普遍，该算法通过构造一个超平面或者高维空间，将低维向量映射到更加复杂的高维空间，通过对少数信息的训练即可实现对小样本数据和非线性数据的分析，在交通领域常用于对交通拥堵的分类与预测。2003 年，Fang Yuan 等人将该算法应用到交通事件自动检测过程中，通过对模拟数据和实际数据的测试，得到了较好的事件检测效果[35]。

9）基于单个检测设施的 AID 算法。姜桂艳等人于 2001 年在人工神经网络的

基础上，利用单个检测设施的信息设计了一种 AID 算法用于高速路事件检测，并设计了一种可实现三级报警制度的高速公路交通事件自动检测系统[36]。

10）非参数回归算法。宫晓燕等人为实现短时交通流的预测和交通事件检测，从基于密集度的变 K 搜索算法与基于动态聚类和散列函数的历史数据组织方式对传统的非参数回归算法进行改进，提出一种基于非参数回归的综合算法。改进的算法具有“无参数”、可移植、预测精度高等特点[37]。

1.3.3 现有算法的不足

通过对以上各算法的分析，可以发现，目前在交通偶发性拥堵研究方面还存在一些不足，具体如下：

① 以上算法大部分针对快速路或高速公路，缺乏对城市道路交通拥堵的分析，而城市道路交通拥堵比高速路的交通拥堵更为复杂，具有波动性，在不同时间、不同路段的交通拥堵不同，目前的算法没有考虑交通拥堵的时空差异性；

② 如何区分城市道路交通正常拥堵与偶发性拥堵，目前还没有明确、统一的指导性指标；

③ 海量公交车 GPS 数据的积累，为城市道路交通偶发性拥堵检测提供了契机，但如何利用这些数据检测城市道路交通偶发性拥堵，目前还缺乏有效的研究。

针对以上问题，本文以海量的公交车 GPS 数据为基础，研究发生在城市道路上的偶发性交通偶发性拥堵检测的方法。

1.4 研究内容、意义与目的

1.4.1 研究内容

在分析国内外在道路交通偶发性拥堵检测领域的研究现状、总结已有的模型和算法的优点和不足的基础上，针对以上的不足本文主要进行以下几方面的研究：

① 研究道路交通偶发性拥堵检测的系统体系结构和模型，从整体上把握道路交通偶发性拥堵研究的主要工作和范围，将主要工作划分为若干个相互独立又彼此关联的模块，明确每一模块的主要工作和各模块之间的关系，为后续工作奠定基础；

② 以海量的公交车 GPS 数据为基础，针对道路交通偶发性拥堵的相对性和情景性的特点以及城市道路交通状态波动较大的情况，研究道路交通状态的历史规律，包括道路交通状态的整体情况分析、道路交通状态模式识别以及偶发性拥堵的量化定义；

③ 针对道路交通偶发性拥堵随机性大、机理模型不清楚等特点，研究道路交通偶发性拥堵的实时变化趋势分析方法，利用 CVA 算法建立基于公交车 GPS 数据的道路交通偶发性拥堵检测模型，最终实现道路交通偶发性拥堵的检测。

1.4.2 研究意义

研究基于公交车 GPS 数据的道路交通偶发性拥堵检测具有以下两方面的意义：

① 学术意义

目前在道路交通偶发性拥堵检测方面，主要集中于高速路（或快速路），而在城市道路交通偶发性拥堵检测方面的研究还比较少；且目前检测方法主要针对高速路，然而城市道路交通状态具有明显的波动性，现有的方法还没有考虑到这点；而且，目前在道路交通偶发性拥堵的研究很大部分是基于仿真的分析，而利用实际数据（特别是海量数据）进行研究的很少。因此，本文以海量公交车 GPS 数据为基础，研究道路交通异常检测的主要方法和模型。

② 实用意义

当今时代已经向大数据时代迈进，在智能交通领域积累了海量的公交车 GPS 数据，这些数据记录了公交车的运行轨迹，具有非常大的价值。如何从海量的公交车 GPS 数据中挖掘出有用的信息以为交通管理和服务提供支持和参考，是目前智能交通领域面临的重要课题和研究热点。本文以海量公交车 GPS 数据为基础，对海量的公交车 GPS 数据进行分析和挖掘，充分发挥公交车 GPS 数据的价值，以实现对道路交通偶发性拥堵的检测。

此外，道路交通检测研究成果可用于交通诱导、交通应急处理等方面，有助于提高道路交通管理水平和服务水平。

1.5 论文章节安排

以上对道路交通偶发性拥堵的特征及国内外研究现状进行了分析，并对本文的研究内容和研究意义进行了说明；接下来的主要工作是研究如何实现对道路交通偶发性拥堵的及时、准确的检测。针对这一目标，对本文的各章节作以下安排：

第一章：绪论，说明研究背景，对道路交通偶发性拥堵的特点进行分析，总结国内外研究现状，说明研究内容及意义；

第二章：基于公交车 GPS 数据的道路交通偶发性拥堵检测系统体系结构的设计和模型的建立，提出整体方案；

第三章：以海量公交车 GPS 数据为基础，分析道路交通状态的历史情况，在此基础上对道路交通状态进行模式识别，划分交通情景，确定类情景下的道路交通状态特征，在此基础上明确道路交通正常拥堵和偶发性拥堵的量化区分；

第四章：利用 CVA 算法实现对交通状态的实时趋势分析与偶发性拥堵检测；

第五章：道路交通偶发性拥堵检测系统的实现与应用，包括系统的实现与检测结果的分析评价；

第六章：总结与展望。

最后，致谢与参考文献。

2 总体方案设计

2.1 本章引言

从一章的分析中可以知道，目前国内外在道路交通偶发性拥堵方面主要集中于对高速路(快速路)交通偶发性拥堵的检测，而在城市道路交通偶发性拥堵检测的研究比较少。此外，由于近年来海量公交车 GPS 数据的积累，如何充分发挥这些数据的作用，为智能交通各具体课题的研究提供支持和参考，目前也亟待进一步的研究。而道路(尤其是城市道路)交通偶发性拥堵的检测作为智能交通领域的一个重要课题，目前利用公交 GPS 数据检测道路交通偶发性拥堵的研究比较缺乏。

因此，本文以海量的公交车 GPS 为基础，研究城市道路交通偶发性拥堵的检测，为城市道路交通诱导或应急处理等提供支持和参考，以帮助提高城市道路交通的管理水平和服务水平。

为了明确利用公交车 GPS 数据进行道路交通偶发性拥堵检测，首先设计整体的技术方案，以帮助明确整体的工作内容和范围，明确整体的框架和技术路线。因此，本章为总体方案设计，具体内容包括系统体系结构的设计和检测方法两部分。

2.2 基于公交 GPS 数据的道路交通偶发性拥堵检测系统体系结构

城市道路交通偶发性拥堵检测与高速路的交通偶发性拥堵检测有较大的区别。一般而言，在一般的高速路上，交通拥堵相对比较稳定，对偶发性拥堵的确定和判别都相对比较简单。城市道路交通具有明显的早高峰和晚高峰，拥堵是常见的交通拥堵，对城市道路交通偶发性拥堵的判别不能仅凭简单的畅通和拥堵进行判别，还应该根据交通拥挤发生的时间、地点以及拥挤程度等多因素进行综合考虑之后才能判定是否有偶发性拥堵发生。

公交车 GPS 数据覆盖面广、可靠性高、实时性高、维护成本低，更重要的是它包含了丰富的城市道路交通拥堵的变化信息，可为道路交通拥堵的研究提供可靠的支持和参考。基于公交 GPS 数据的道路交通偶发性拥堵检测需要通过对公交车 GPS 数据的采集、集成和预处理（如清洗、转换和装载等），监控交通系统的运行拥堵，检测交通系统的异常信息，分析产生偶发性拥堵的原因，并实现对交通偶发性拥堵动态变化趋势的预测，以便于实现对交通系统的实时监测和控制，达到降低交通出行者的出行成本，提高道路交通管理水平和服务质量的目的。

而利用公交车 GPS 数据检测道路交通偶发性拥堵，首先需要明确所需完成的工作和范围，然后对各项工作进行合理地划分，明确主要工作所包含的模块，同

时明确各模块之间的关系，确定后续工作的主要目标。在这一基础上，明确实现所有工作的思路，设计出合理的解决方案和技术路线。

利用公交车 GPS 数据进行道路交通偶发性拥堵检测的整体思路是：第一，建立基于公交车 GPS 数据的道路交通偶发性拥堵检测系统的体系结构，将主要工作划分为若干个具体的模块，明确每部分的内容和作用，从整体上把握本文的内容和方向；第二，对海量历史数据进行预处理，并选择合理的参数用以表征和描述道路交通状态，为后续工作奠定基础。第三，以海量公交车 GPS 数据为基础，分析道路交通状态的历史情况，统计出道路交通状态历史的整体情况，在此基础上进行道路交通模式识别，根据交通状态的拥挤程度划分交通情景，并统计分析每类交通情景下交通状态的特征。在此基础上，明确每类情景下交通正常拥堵和偶发性拥堵的差异，确定历史的交通偶发性拥堵。第四，分析道路交通状态的实时变化趋势，利用历史的偶发性拥堵训练每类情景下的阈值，将道路交通状态实时变化趋势与阈值进行对比，从而确定是否有偶发性拥堵的发生。第五，实现基于公交车 GPS 数据的道路交通偶发性拥堵检测系统，并且应用实际数据对道路交通偶发性拥堵检测的结果进行验证。具体如图 2.1 所示。

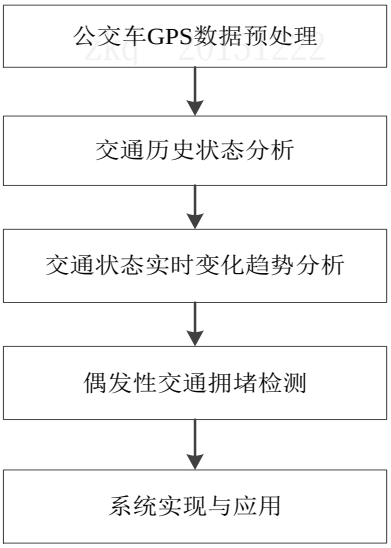


图 2.1 基于公交车 GPS 数据的道路交通偶发性拥堵检测整体思路

Figure 2.1 The whole idea for road traffic contingency jam detection base on bus GPS data

基于公交车 GPS 数据的道路交通偶发性拥堵检测应该包括道路交通状态历史规律分析、道路交通状态实时变化趋势分析以及道路交通偶发性拥堵检测等部分，而进一步，各部分又包括多个具体需要解决问题和难点。根据这一思路，首先需要建立基于公交车 GPS 数据的道路交通偶发性拥堵检测系统的体系结构。为利用公交车 GPS 数据进行道路交通偶发性拥堵检测建立系统体系结构是为了从整体上

把握后续工作的方向和范围，将主要工作进行划分为若干相互关联的模块，突出重点工作，同时明确主要模块的主要内容及其各模块之间的关联关系。

因此，本文设计基于公交车 GPS 数据的道路交通偶发性拥堵检测系统体系结构具体如 2.2 图所示。

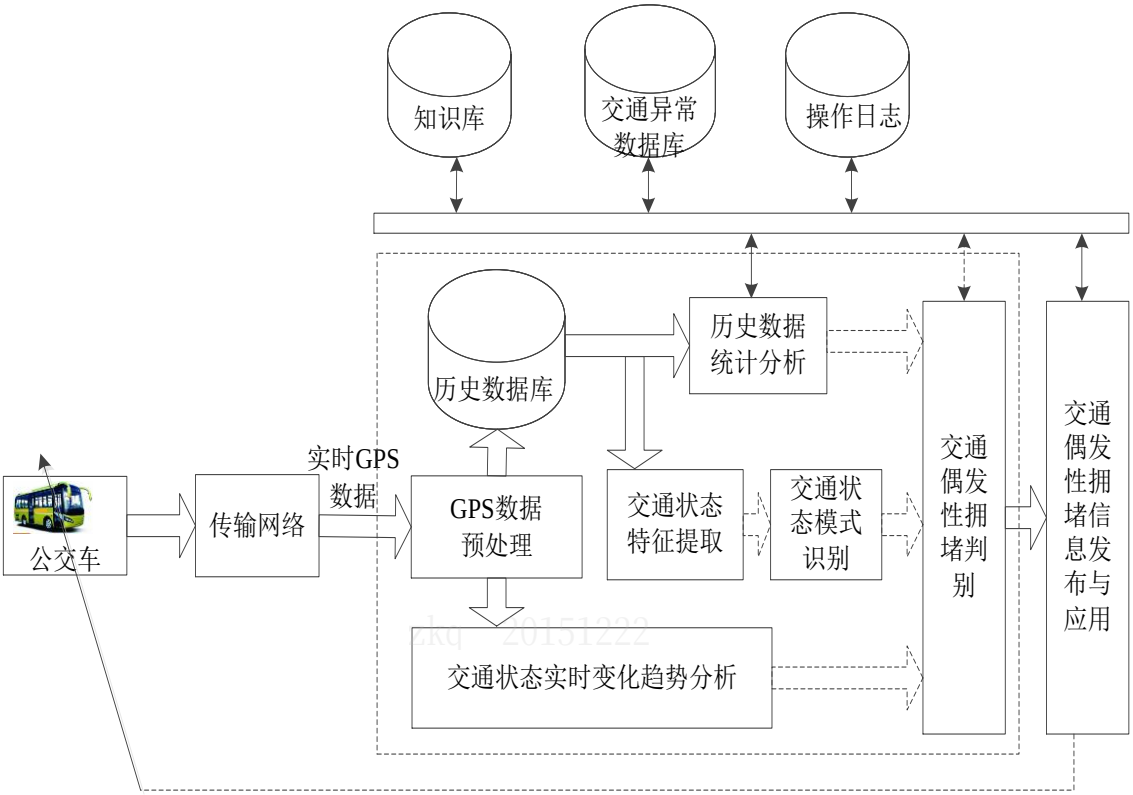


图 2.2 基于公交车 GPS 数据的道路交通偶发性拥堵检测系统体系结构

Figure2.2 System structure of road traffic contingency jam detection base on bus GPS data

其中各模块的主要内容及其作用具体如下：

① GPS 数据预处理模块，这是后续工作的基础：包括数据的接收、解析、清洗，转换和装载，实现数据质量保障，以及后续参数的计算和转化。其中数据接收是指利用数据接收程序从远程 GPS 终端接收 GPS 字符串序列，解析是对 GPS 字符串序列进行解析，将字符串序列转化为可视的具有特点含义的数据项。清洗是指删除明显错误的数据，同时对缺失的数据进行补偿处理，以保证数据的正确性和完整性。数据装载是指将公交车 GPS 数据保存到数据库中，并定时进行备份。数据集成是指建立具有特定主题的数据仓库，将 GPS 数据、车辆信息、线路信息等多类信息进行融合。由于这些具体的操作内容繁多，目前已有成熟的解决方案，且不是本文的重点，具体操作细节在此不作深入介绍；此外，值得一提的是，该

部分除了为后续工作提供完整可靠的公交车 GPS 数据之外，还需要对这些数据作进一步的计算出来，计算出后续工作所必须的一些相关参数，如 3.2 节中的路段行程时间、路段延误时间指数等，以及 4.2 节中的各种实时速度参数，具体在后文中再作详细分析说明；

② 历史数据库模块：主要包括为对历史数据的存储、备份，为增删查改等日常操作提供支持。此外，该部分将公交车 GPS 数据按一定的要求进行存储、转换和更新，为后续的历史数据的统计分析和道路交通模式识别等工作提供支持（由于该部分目前已有较为成熟的技术，且不是本文的重点，在本文中不作进一步的介绍说明）。

③ 历史数据统计模块：利用海量的历史数据对交通历史状态情况的整体分析，分析道路交通历史分布情况，为把握交通状态的整体情况提供参考（3.2 节）；

④ 交通状态特征提取：首先分析道路交通状态在不同路段不同时间段的交通拥堵程度及其波动程度，分析道路交通状态的历史变化趋势，为交通状态模式识别提供基础（3.3.1 节）；

⑤ 交通状态模式识别：根据海量的公交车 GPS 数据，统计出每个交通情景下道路交通状态的拥堵及波动程度，根据拥堵程度的大小对交通情景进行划分，使得同一类情景的交通状态差异最小化而不同类情景下的交通状态差异最大化，确定每一情境下的道路交通状态的情况（3.3 节）。在此基础上，确定每一种情境下交通正常拥堵和偶发性拥堵的量化区分，确定历史的偶发性拥堵，为后续利用历史数据训练各情景下的阈值和评价道路交通偶发性拥堵检测效果提供支持和参考（3.4 节）；

⑥ 交通状态实时趋势分析：根据实时的公交车 GPS 数据，首先选择合适的参数来表征道路交通的实时状态，然后选择合适的方法分析道路交通状态的实时变化趋势（4.3~4.4 节）；

⑦ 交通偶发性拥堵判别模块：利用历史的偶发性拥堵训练每种情景下的阈值，确定每类情境下交通状态变化的范围，将实时变化趋势与阈值进行对比，从而实现对偶发性拥堵的判定（4.5 节）；

⑧ 交通偶发性拥堵信息发布及应用模块：是指将偶发性拥堵研究的结果应用到实际中，以期提高城市智能交通系统的服务水平。

⑨ 知识库主要其他各模块；交通异常数据库则包括与交通偶发性拥堵有关的信息和知识，包括除 GPS 数据外其他一切在偶发性拥堵的判别、影响程度分析等模块中的输入信息和输出结果。

在图 2.2 的各模块中，公交 GPS 数据预处理是后续工作的基础，而历史数据统计分析是为从整体上把握道路交通历史的状态提供参考，道路交通特征提取是

在海量历史数据的基础上分析各路段在各时间段的交通特征，为道路交通状态模式识别提供支持。道路交通状态模式识别是为了确定各情景下的交通状态特征，为明确区分交通正常状态和偶发性拥堵从而最终实现交通偶发性拥堵的量化定义提供支持。而道路交通状态实时变化趋势分析则是为了确定道路交通拥堵的实时变化趋势，将道路交通拥堵变化趋势与预定值相比较，最终实现道路交通偶发性拥堵的判别。

由于道路交通系统是一个受多方面因素相互作用、相互影响的复杂系统，具有高度动态性和非线性，对道路交通偶发性拥堵的研究是一个持续深入和改进的过程。在本文中，将针对一些关键的问题，如道路交通状态的表征、利用海量公交 GPS 数据分析道路交通历史状态、交通状态特征提取、交通状态模式识别、交通状态实时趋势分析以及最终的交通偶发性拥堵的判别等问题（即图 2.1 中虚线框内包含的部分）进行深入研究分析。而对于其他部分目前已有成熟的方案，在此只作简单说明。

2.3 基于公交车 GPS 数据的道路交通偶发性拥堵检测方法

正常情况下，道路交通状态会随着时间和地点的不同而可能出现较大的波动。但道路交通状态具有历史趋势性，即在同一路段上的连续一段时间内，道路交通拥堵程度保持稳定或者变化很小。而当出现偶发性拥堵时，道路交通拥堵程度在短时内会产生较大的波动。同时，受实时路况的影响，公交在运行过程中其速度变化差异也可能比较大，不能因公交车辆在一个周期内的速度发生急剧的变化而说明发生了偶发性的交通拥堵。

由前面的分析可以，道路交通偶发性拥堵具有集群性，在偶发性拥堵的情况下，往往不止一辆车会出现异常，也不止在一个周期内会出现异常，而是同一路段上的多辆车在多个周期内都会出现异常的反应。因此，对道路交通偶发性拥堵的判定需要同时考虑同一路段上的多辆车在多个周期内拥堵的情况。因此，本文提出一种基于公交车 GPS 数据的道路交通偶发性拥堵检测方法，如下所示：

$$AS(r,k,t)=\begin{cases}1, & \forall j\in(t,t-1,\ldots,t-p+1), SPE_R(r,k,j)>SPE_H(r,k)\\ 0, & \exists j\in(t,t-1,\ldots,t-p+1), SPE_R(r,k,j)\le SPE_H(r,k)\end{cases} \quad (2.1)$$

式（2.1）中 r 和 k 分别指的是路段和时段，t 为检测周期，p 为检测周期数，$AS(r,k,t)$ 为路段 r 在 k 时段 t 时刻的偶发性拥堵检测标志，$AS(r,k,t)=1$ 说明有偶发性拥堵发生，$AS(r,k,t)=0$ 说明没有偶发性拥堵发生。$SPE_R(r,k,j)$ 为路段 r 在 k 时段 j 时刻的实时检测值（需要注意的是该值为多车参数融合的结果，具体计算方法及其物理意义在第 4 章中再作进一步说明），$SPE_H(r,k)$ 为路段 r 在 k 时段的历史值。当 t 时刻前的连续 p 个周期内都有 $SPE_R(r,k,j)>SPE_H(r,k)$，才能说明有偶发性拥

堵发生，此时 $AS(r,k,t)=1$；否则没有偶发性拥堵发生，此时 $AS(r,k,t)=0$ 。

要实现该方法需要解决三个具体的问题：第一，$SPE_R(r,k,j)$ 和 $SPE_H(r,k)$ 的具体物理意义是什么？第二，如何计算 $SPE_R(r,k,j)$ 和 $SPE_H(r,k)$ 的值？第三，p 值的大小该如何确定？实际上，图 2.2 中的历史数据统计分析和道路交通模式识别等都是为求得 $SPE_H(r,k)$ 值提供支持，而进行道路交通状态实时趋势的分析首先就要确定选择何种参数来表征 $SPE_R(r,k,j)$ 值，在此基础上计算出 $SPE_R(r,k,j)$ 的取值大小。再根据历史的数据训练得到 $SPE_H(r,k)$ 值。通过将 $SPE_R(r,k,j)$ 和 $SPE_H(r,k)$ 进行对比，即可判定是否有偶发性拥堵发生。

因此，基于公交车 GPS 数据的道路交通偶发性拥堵检测可分成四部分的内容：

① 以海量公交车 GPS 历史数据为基础的道路交通状态规律分析，主要包括数据统计分析、道路交通状态特征提取和交通模式识别，确定每个路段每个时间段对应的交通状态的特征，为得到 $SPE_H(r,k)$ 做准备；

② 利用实时的公交车 GPS 数据分析道路交通状态的实时变化趋势，以确定 $SPE_R(r,k,j)$；

③ 比较道路交通状态的实时变化趋势的大小，实现对道路交通偶发性拥堵的判别，包括：阈值的训练，即确定 $SPE_H(r,k)$ 的值；道路交通偶发性拥堵的判定，即通过 $SPE_R(r,k,j)$ 和 $SPE_H(r,k)$ 的比较从而最终确定 $AS(r,k,t)$；

④ 通过实际数据进行检测，比较不同 p 值下的检测结果，可以实现检测效果的优化，从而进一步提高检测效果。

2.4 本章小结

本章为利用公交车 GPS 数据进行道路交通偶发性拥堵检测的总体方案设计，以明确主要的工作内容和范围，确定具体的技术路线，主要包括两部分：首先建立了基于公交车 GPS 数据的道路交通偶发性拥堵检测系统的体系结构，以便从整体上把握利用公交车 GPS 数据检测道路交通偶发性拥堵的范围和方向，划分主要工作模块，明确各模块的主要内容及各模块之间的关系。其次，为对道路交通偶发性拥堵实现准确的描述，建立了基于公交车 GPS 数据的道路偶发性拥堵检测模型。

在后续的章节中，第 3 章利用海量的公交车 GPS 数据分析道路交通状态的历史规律，而第 4 章则利用实时的公交车 GPS 数据分析道路交通状态的实时变化趋势。结合第 3 章中交通实时的历史规律和第 4 章的道路交通实时的实时变化趋势，即可实现对道路交通偶发性拥堵的检测。在第 5 章实现偶发性拥堵检测系统，并以实际的 GPS 数据对本文所提出的道路交通偶发性拥堵检测方案进行验证。

3 基于公交车 GPS 历史数据的道路交通状态规律分析

3.1 本章引言

由第 2 章的分析可知，基于公交车 GPS 数据的道路偶发性拥堵检测主要包括基于公交车 GPS 历史数据的道路交通拥堵规律分析、基于 GPS 实时数据的实时变化趋势分析和道路交通偶发性拥堵检测 3 部分的内容，而道路交通实时规律分析又包括历史数据统计分析、道路交通状态模式识别，其中历史数据统计分析是以海量的公交车 GPS 历史数据为基础，从中宏观的角度统计分析公交车 GPS 数据的总体情况；道路交通状态模式识别是分析道路交通拥堵在不同情景下的差异，对具有相同交通拥堵变化模式的情景进行分类，确定每一类情景的交通拥堵特征；而实时变化趋势分析则根据公交车 GPS 实时数据是从微观的角度分析道路交通拥堵，以实现对交通拥堵变化趋势的准确把握；最后道路交通偶发性拥堵检测则是融合前三者的结果，以实现对道路交通偶发性拥堵的实时检测。

本章分析道路交通状态的规律，包括公交 GPS 数据统计分析和道路交通状态模式识别两部分，具体又可以分为以下三部分：① 利用统计学方法对公交 GPS 历史数据进行统计分析，包括参数分析、路段和时段划分、整体分析和差异分析等部分，以实现对道路交通历史状态特征的整体把握，为后续工作提供支持和参考；② 提出基于 T 检验的 K-均值聚类算法进行道路交通状态模式识别，以实现对交通情景的划分以及确定每种情景下交通状态的拥堵程度和波动范围；③ 引入四分位差的思想明确划分道路交通正常状态和偶发性拥堵。

3.2 道路交通历史拥堵程度统计分析

道路交通历史状态的统计分析为了从整体上把握道路交通状态的情况，同时分析从时间和空间方面分析道路交通拥堵程度在不同的路段不同时段的特征及差异，从而为道路交通状态模式识别提供基础。包括以下四部分：参数分析、路段和时段划分、道路交通拥堵程度整体统计分析以及其差异分析。

其中参数分析是选择有效的参数用以表征道路交通拥堵程度；而路段和时段的划分，是将时间和空间划分别划分为若干个的单位，确定道路交通状态分析的最小单位；历史数据统计分析是为了从整体上把握道路交通拥堵程度的历史情况，而道路交通状态差异分析则是比较交通拥堵在不同星期、不同时间段、不同路段的差异，以便为后续的道路交通状态模式识别提供支持和参考。

3.2.1 历史参数分析

道路交通状态特征描述即选择有效的参数对道路交通状态进行准确的描述，

这是检测道路交通偶发性拥堵的基础。目前对道路交通状态的研究主要基于固定检测数据、浮动车 GPS 数据以及仿真数据。对于不同的数据，可选择的参数基础也不尽相同。其中固定检测数据对应车流量、速度、占有率等参数，而浮动车 GPS 数据则对应地有速度、经纬度、GPS 时间、里程等参数。此外，还可从这些不同的基础参数中计算或转换成其他特征参数，如文献[38]定义了“路段拥挤度”指标来描述交通拥堵特征，文献[39]提出以路段行程时间、路段交通拥堵系数和路段拥挤程度等参数来描述交通网络的实际信息和交通拥堵信息。

由于公交车 GPS 数据中包含的参数包括路段、里程、GPS 时间、瞬时速度等，而能够直接拥有表征交通状态的只有“瞬时速度”一个参数。然而，由于公交在运行过程中停靠站上下客、信号灯区域排队等候、重庆市区不同路段上下坡度差异较大、临时路况障碍等因素的影响，单一的瞬时速度作为指标参数不能完全有效表征路况的实时交通状态。因此，需要结合其他参数，以便有效表示路段的交通状态。

对于过去的历史数据，可以从中计算出每个车次公交车在每个路段的路段行程时间，用路段行程时间来表征该车次所经过的路段的交通状态。由于路段行程时间反应了公交车辆经过相应路段的情况，比速度信息更加稳定可靠。因此，本文主要利用路段行程时间来表征车辆所经过的路段的历史状态。

由于不同路段的长度及路段等级不同，不同路段的路段行程时间差异较大，为了消除不同路段之间的差异，在此定义“路段延误时间指数”来表征一个路段的交通拥堵程度。具体计算公式如下：

$$\lambda(r,i,k)=\frac{t(r,i,k)-\min\{t(r,k)\}}{\min\{t(r,k)\}} \tag{3.1}$$

其中 $\lambda(r,i,k)$ 表示车辆 i 在一天中第 k 个时段经过路段 r 的路段延误时间指数，$t(r,i,k)$ 为车辆 i 在每天第 k 个时段经过路段 r 的路段行程时间，$\min\{t(r,k)\}$ 为所有公交车辆在 k 时段经过路段 r 的最小的路段行程时间。

3.2.2 时段和路段划分

① 时段划分

公交车在固定路线上行走，具有线路和站点固定的特点，其运行过程包含了公交运行规律以及道路交通状态变化的特点。正常情况下，从时间的角度看，道路交通状态会因时间段的不同而不同，在同一地点的道路交通状态会随着时间的变化而呈现周期性特征。从长期的角度来看，道路交通状况在一定的时间范围（即一个时段）内的变化趋势整体保持稳定拥堵，具有历史趋势性。研究道路交通状态在不同时段的差异，关键在于如何确定每个时间段的范围。若时段范围划分太大，则不能体现交通状态的差异性，若时段范围划分太小，则会将具有相同规律

的时段划分为多个更小的时段，降低效分析效率。经统计分析，每个时间段应以半小时为一个时段为宜。因此，一天 24 小时可以被平均划分为 48 个时间段，而一周 7 天可以划分为 336 个时间段。

② 路段划分

同样，路段是城市道路路网的基本组成单位，需要对路段进行科学合理地划分。由于公交车定线运行、定点停靠的特点，对公交线路上路段的划分首先应以公交站作为节点，同一线路上的相邻两个站点之间的道路为一个路段。然而，在同一路段上可能有多条线路的公交车辆经过，为了充分利用多线路的公交车 GPS 数据，提高道路交通偶发性拥堵的检测效果，应该考虑不同线路路段重合的问题。还有，由于公交站点的设置是以客流量为主要依据，在同一公交路段上的交通状况（车流量和客流量等）和地理条件（坡度和宽度等）可能存在较大的差异，此外，受到信号灯控制的影响，同一路段上的公交行驶速度也可能有较大的差异。因此，需要根据具体情况对公交路段作进一步的划分。其中划分的依据包括：信号灯交叉口、地理环境差异、车道数量、交通拥堵相同等，具体划分过程在此不作详细介绍。

3.2.3 交通状态整体分析

为了从整体上把握道路交通状态的情况，为后面利用公交车 GPS 数据进行道路交通偶发性拥堵判别提供参考，需要对海量公交车 GPS 数据对道路交通状态的历史情况进行统计分析，主要包括两方面的内容：整体描述分析和道路交通历史状态在不同星期、路段和时间段的分布分析。

在此，选择重庆市日间车公交线路 886 线路下行方向 2-19 号路段从 2013 年 6 月 11 日到 2014 年 1 月 10 日连续 7 个月的 GPS 数据作为研究对象，首先从中统计出每个车次在每条路段的路段行程时间，然后根据式（3.1）计算出对应的路段延误时间指数，最终得到各车次在各路段的路段延误时间指数数据共 54522 条。以路段延误时间指数为指标，以描述统计方法为手段分析道路交通历史状态整体情况，具体如表 3.1 和 3.2 所示。

表 3.1 道路交通状态历史情况统计

Table3.1 Statistics result of road traffic history states

有效个数	均值	中值	众数	标准差	方差	偏度	峰度	全距	极小值	极大值
54522	0.715	0.407	0.111	0.897	0.804	3.268	16.917	12.105	0.000	12.105

表 3.2 道路交通状态百分位数分布情况

Table3.2 Percentiles of Road traffic states

百分位	1	10	50	70	80	90	95	98	99
路段延误时间指数	0.012	0.100	0.407	0.745	1.010	1.786	2.549	3.447	4.210

从表3.1和表3.2可以看出，整体上路段延误时间指数的均值约为0.71，标准差接近0.90，说明886线路下行方向的运行时间超出理想时间的71%，也说明了道路交通平均处于轻度拥挤状态，而且波动程度比较大。最大值为12.1052所表示的意义是：在最拥挤的拥堵下，正常情况下5钟能走完的路程，结果花了一个多小时（65.5分钟）。偏度大于0，说明路段延误时间指数为右偏，其分布在右边有拖长的尾巴；峰度大于0（达到16.917），说明其有更加陡峭的顶峰，说明道路交通在某个拥堵下的集中程度非常高。百分位数分析中，1对应的值为0.011765，说明有1%的路段延误时间指数的值小于0.011765。同理，可以看出，10%的路段延误时间指数在0到0.1以内，有80%的路段延误时间指数在0到1.01以内，有98%的路段延误时间指数在3.5以内，说明在98%的拥堵下，公交的行驶时间是在正常行驶时间的4.5倍以内。路段延误时间指数大于4.21的次数不到1%，说明只有不到1%的拥堵下，公交行程时间是正常情况下行程时间的5.2倍以上。道路交通拥堵的历史分布情况如图3.1所示。

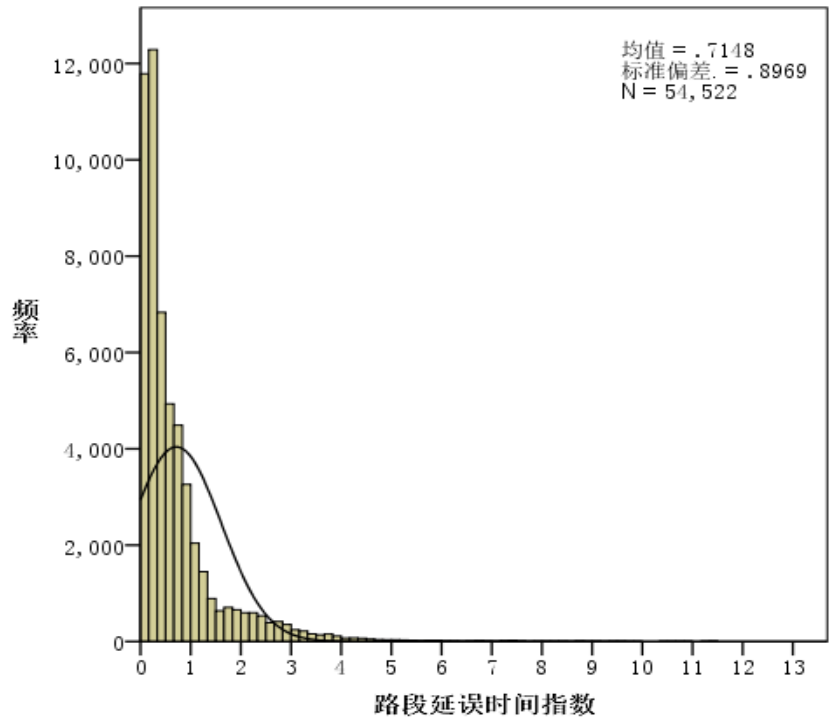


图 3.1 路段延误时间指数分布图

Figuer3.1 The total distribution of road delay time index

图3.1与表3.1、表3.2对应，从上图可以看出，道路交通的历史状态主要为畅通或者轻度拥挤的状态。随着路段延误时间指数（拥挤程度）的增大，出现的次数呈逐渐减少的趋势。说明了历史上的道路交通状态主要以畅通或者轻度拥挤为主，而处于严重拥挤的状态所占的比例相对很小，也说明历史上偶发性拥堵所占的比例的幅度不大。

3.2.4 交通状态差异分析

交通状态差异分析的目的是为了检验道路交通状态堵在不同路段、不同星期和不同时间段之间的差异。在此，仍选择重庆市 886 线路下行方向 2-19 号路段从 2013 年 6 月 11 日到 2014 年 1 月 10 日期间的 GPS 数据作为分析对象。

① 道路交通状态在不同星期天之间的差异分析

利用方差分析方法分析一周内 7 天的交通状态的差异，结果如表 3.3 所示。

表 3.3 道路交通状态在一周各天之间的差异分析

Table 3.3 ANOVA of road traffic states between weekdays

星期	个数	均值	标准差	极小值	极大值	F	显著性 方差齐性检验	
1	7229	0.743	0.965	0.000	12.105			
2	8292	0.713	0.877	0.000	11.375			
3	8518	0.720	0.925	0.000	11.500			
4	8418	0.727	0.887	0.000	11.500	24.471	0.000	0.000
5	7445	0.796	1.012	0.000	10.750			
6	7028	0.662	0.810	0.000	10.525			
7	7592	0.640	0.769	0.000	9.675			
总数	54522	0.715	0.897	0.000	12.105			

从上表可以看出，首先进行方差齐性检验，检验结果中对应的P值为0.000，说明方差不齐，需要进一步进行Welch修正，修正后的F值为24.471，对应的显著性水平为0.000，说明不同天的交通状态存在明显的差异。

同理，可从时间段和路段两个方面来对交通状态的差异进行检验，其显著性水平均为 0.000，说明交通状态在时间段和路段等因素方面都具有显著性差异。

从以上分析可以看出，不同路段、不同星期、不同时间段的交通状态可能有明显的差异。因此，检测道路交通偶发性拥堵，考虑道路交通状态在不同路段、不同星期天和不同时间段之间的差异是非常有必要的，同时这也说明了进行道路交通状态模式识别的必要性。

3.3 基于 K-均值聚类自适应算法的交通状态模式识别

道路交通偶发性拥堵具有相对性和情景性。从 3.1 节中可知，可以从空间的角度将一条公交线路（或者一个区域）划分为若干个路段，而从时间的角度则包括可以将一天划分为多个时间段，而且一个路段和一个时间段构成一个基本的时空单元，即一个交通情景。根据道路交通状态的历史趋势的特点可知，从长期看来，同一情景下的交通状态具有历史趋势性，但不同情景的交通状态差异可能较大，但从整体上看，又有多个情景下的交通状态相同。

因此，所谓“交通状态模式识别”，就是要以海量的历史数据为基础，根据道路交通的拥堵程度实现对不同交通情景的划分，使具有相同（或相似）交通状态的情景划分为一类，而交通状态差异较大的情景划分为不同的类，并确定每一类情景对应的交通状态特征。

3.3.1 道路交通状态特征提取

道路交通状态特征提取主要是首先统计各初始情景下的道路交通的历史状态特征，进一步分析道路交通历史状态在不同星期、时段和路段的分布情况，确定每个路段的交通状态在不同时间段的交通拥堵的拥挤程度和波动程度。道路交通拥堵在不同星期、时间段和路段上的整体分布具体如图3.2-3.4所示。

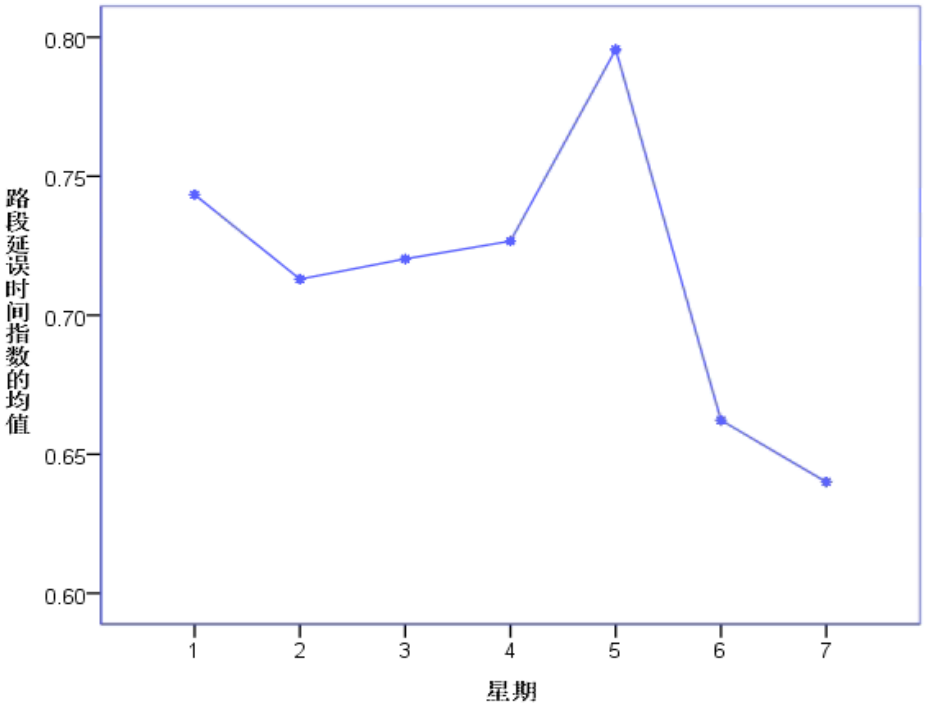


图 3.2 道路交通状态在不同星期天的整体分布

Figure 3.2 Road traffic states’ distribution in different weekdays

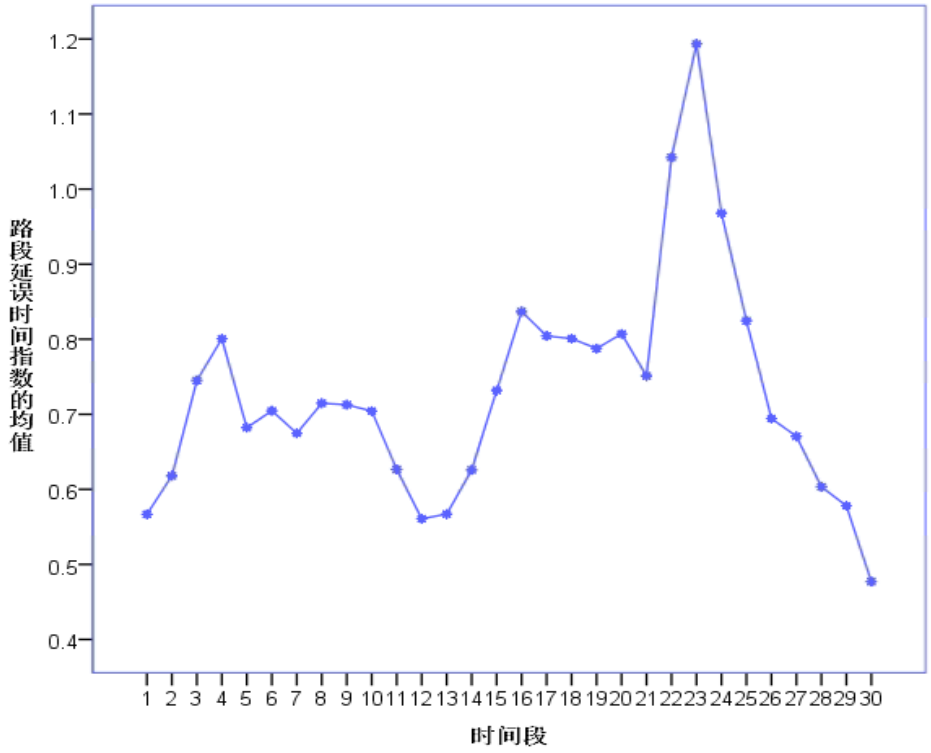


图 3.3 道路交通状态在不同时段之间的整体分布

Figure 3.3 Road traffic states' distribution in different time sections

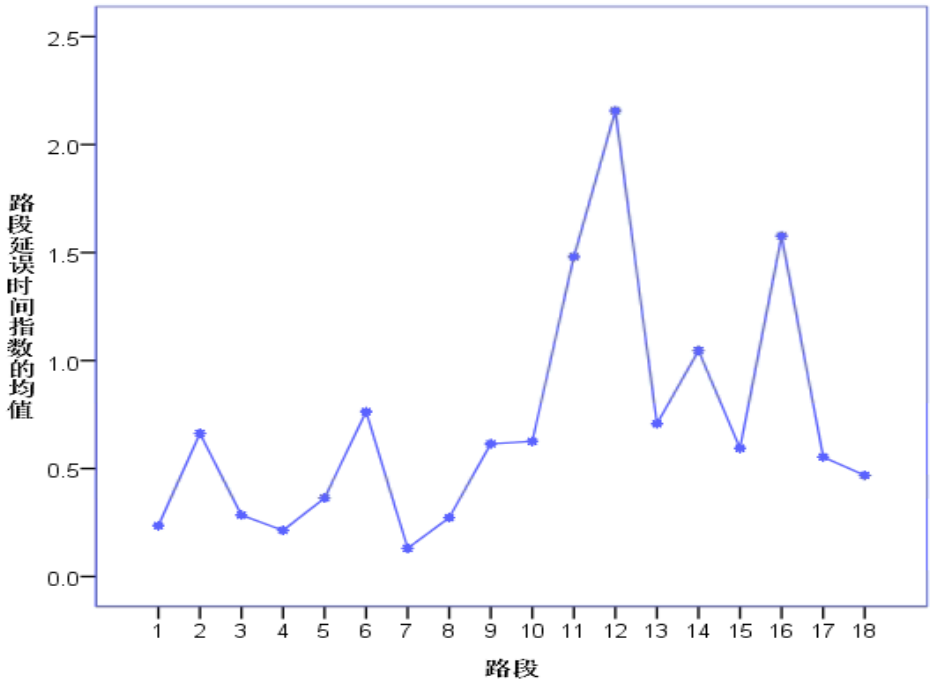


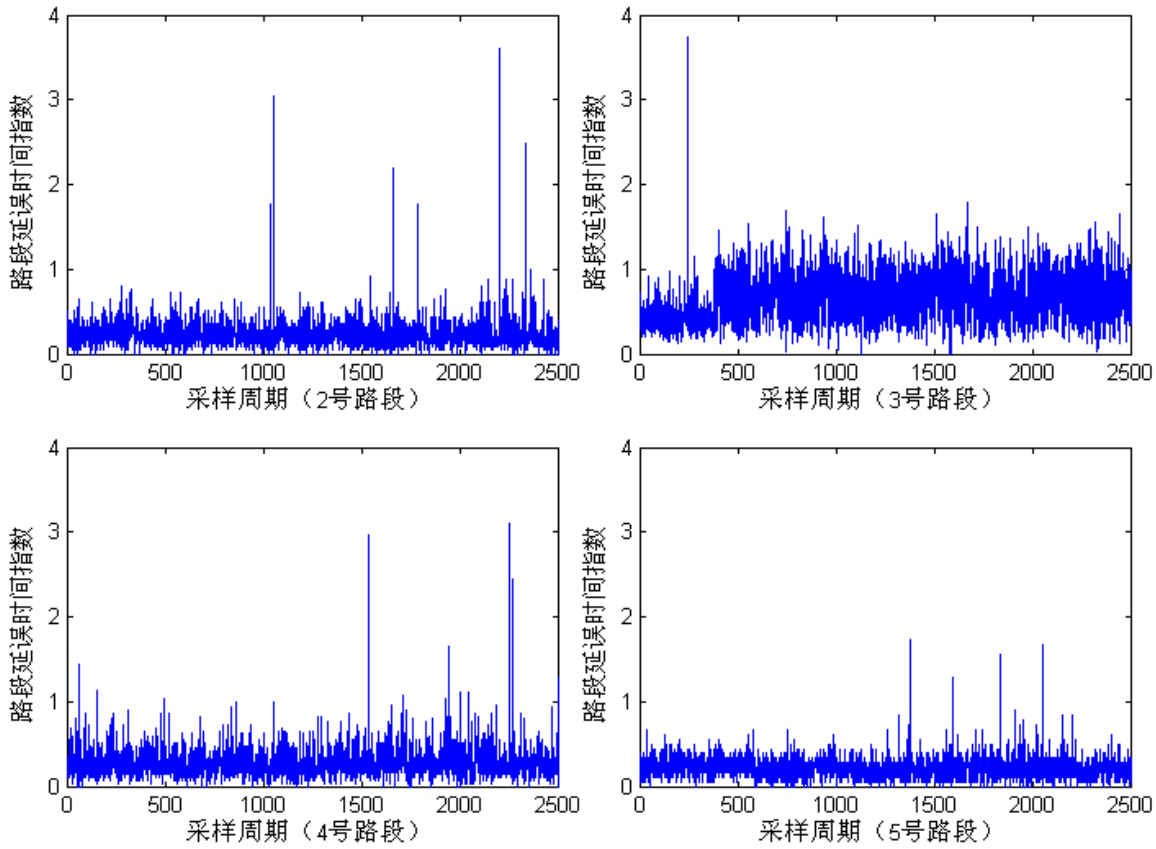
图 3.4 道路交通状态在不同路段之间的整体分布

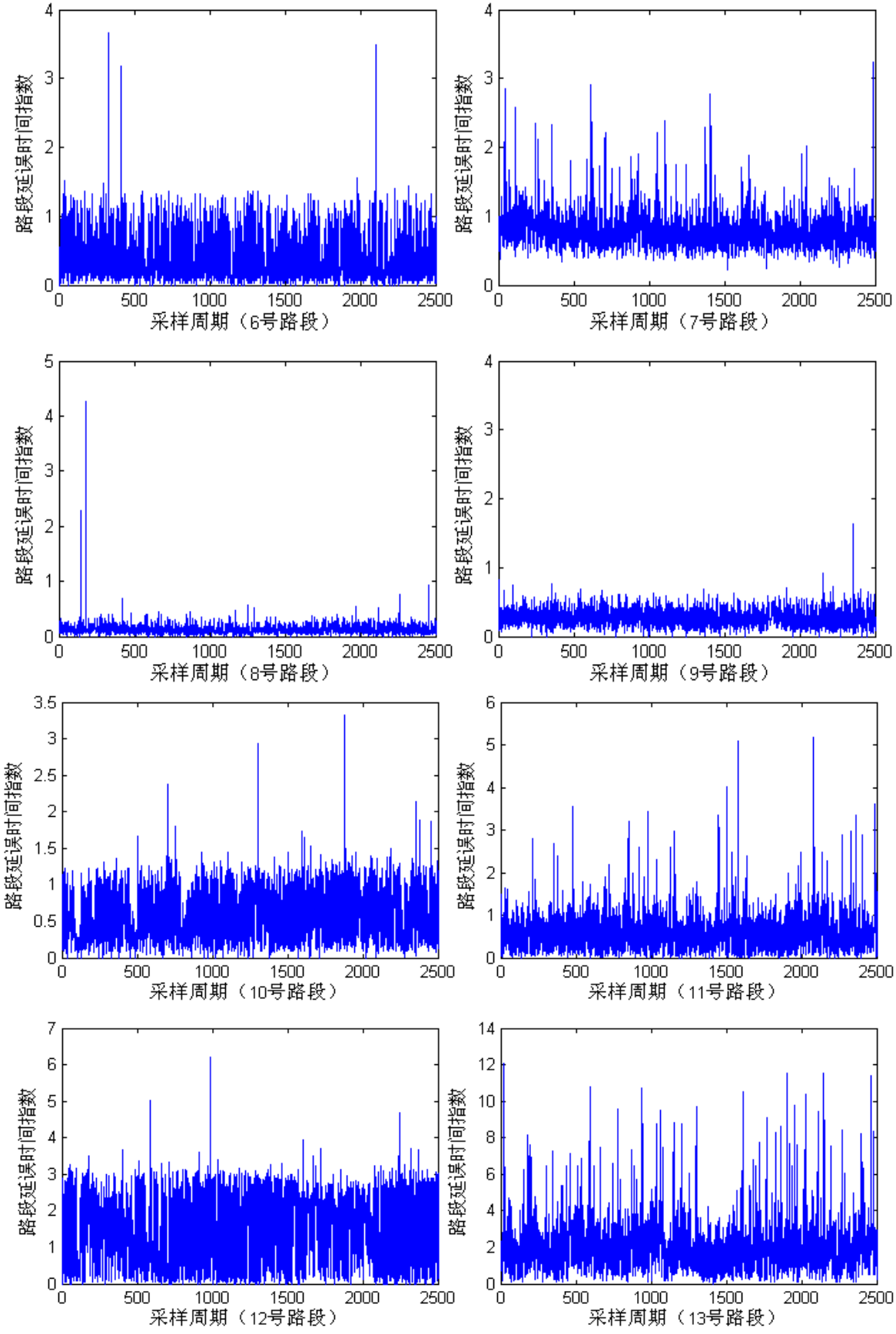
Figure 3.4 Road traffic states' distribution in different road sections

从以上分布图可以看出，周五的路段延误时间指数的均值最大，说明周五这

一点的交通拥堵情况最严重，其次是周一和周四，而周六和周日的交通状况最好。在时间段方面，道路交通历史状态具有明显的早高峰和晚高峰，其中早高峰期在时段 3 和时段 4(即早上 8:00-9:00，其中时段 1 为早上 7:00-7:30，时段 2 为 7:30-8:00，依次类推，直到晚上 21:30-22:00 为时段 30)，晚高峰期在时段 17:30-19:00。而在路段方面，10 号路段之前的路段延误时间指数的均值较小，说明这些路段的交通状况相对比较稳定，而 11 号路段及其下游路段的路段延误时间指数均值较大，说明这些路段交通状况相对拥堵，其交通拥堵波动比较大。

进一步，分析交通历史拥堵在不同路段上的具体分布，如图 3.5 所示，该图从左到右，从上到下依次为 886 线路下行方向 2 到 19 号路段共 18 条路段的交通状态分布情况。




路段延误时间指数
采样周期（6号路段）
采样周期（7号路段）
采样周期（8号路段）
采样周期（9号路段）
采样周期（10号路段）
采样周期（11号路段）
采样周期（12号路段）
采样周期（13号路段）

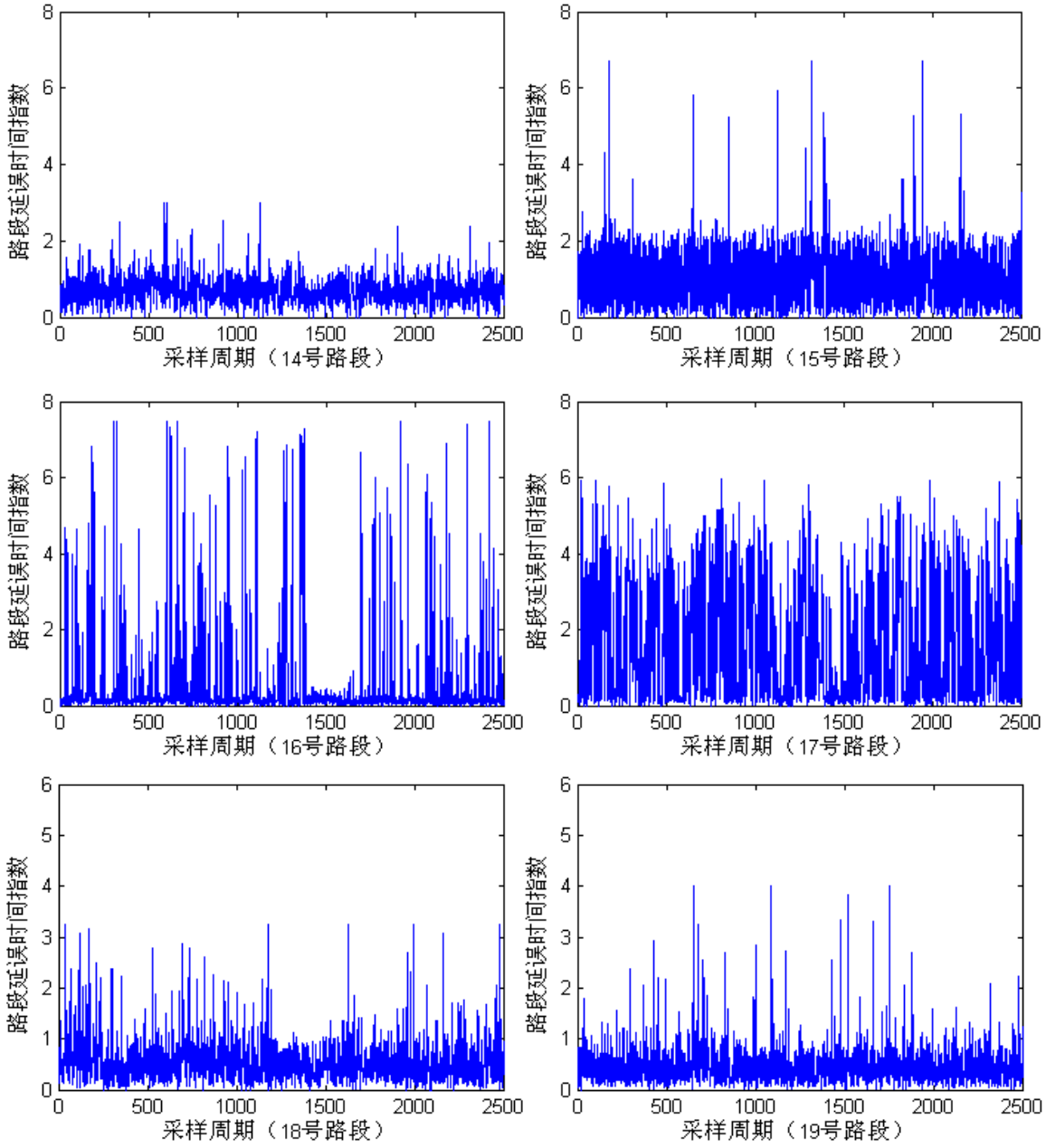


图 3.5 886 线路 2-19 号路段的路段延误时间指数历史分布

Figure3.5 Road delay time index history distribution of 2 to 19 road sections of bus line 886

从图 3.5 可以看出，不同路段、不同时间段的道路交通状态有明显的差异，其中交通状态较为稳定的有 2 号、4 号、8 号等路段，而从 10 号路段到 19 号路段区间内，各路段的交通状况有较大的波动，尤其是 13 号、16 号和 17 号等路段。说明了在不同的路段上，对是否发生了偶发性拥堵的判定范围是有明显差异的。

从上述的分析可以看出，不同交通情景（或路段不同、或时间段不同、或两者都不同）下的交通状态很可能不同，因此需要统计出每种情景下的交通状态的拥挤程度和变化程度（即每种情景下的路段延误时间指数的均值和标准差）。每一

种情景下的路段延误时间指数的均值和标准差即表征了该情景的交通特征。

3.3.2 交通情景划分的必要性

对交通情景进行划分，实质上是对交通状态的划分，每一种情景对应一种交通状态。在绪论中已经提到目前已经存在多个国家、地区和行业的交通拥堵划分标准，这些标准在制定的过程中，考虑了交通参与者的主观感受因素，因此在对交通状态进行科学合理地划分的前提下，还要考虑到标准的普遍适用性和社会大众的接受程度，因此，这些标准中的交通状态分类数目要尽可能小，而使每一类交通状态的范围均较大。

但经实验分析发现，若交通状态划分标准划分范围太大，则不适用于道路交通偶发性拥堵的检测。如果交通拥堵划分不当，则会造成较大的漏报率和误报率的问题。例如，每天早上 7:00-7:30 期间，2 号路段的路段延误时间指数长期保持在 0.25 以下，但在某天连续一段时间增大到 0.6 左右，这是一种异常，但若将畅通状态下路段延误时间指数的正常值控制在 0.667 以下，则会造成交通偶发性拥堵的误报。再如，某路段在晚高峰期路段拥堵比较严重，这是一种常发性拥堵，属于正态拥堵的范畴。拥堵状态下如果阈值太小，就会造成误报，如果阈值太大，则很可能造成偶发性拥堵的漏报。而且，拥堵状态下偶发性拥堵漏报和误报的矛盾不能仅靠调节阈值的大小消除，可行的方法是根据交通状态对交通情景进行更加精细、准确地划分。

对交通情景的划分有两个关键的问题：其一，划分为多少类？其二，每一类的范围多大？对不同时空下交通情景进行划分，可以利用 K-均值聚类算法，该算法原理比较简单，实用性强，在多个不同领域得到广泛的应用，但该算法也存在一些不足之处。因此，结合道路交通情景划分的需求和 K-均值聚类算法的不足之处，本文提出一种基于 T 检验的 K-均值自适应聚类算法以实现对交通情景的划分。

3.3.3 K-均值自适应聚类算法

① K-均值聚类算法简介

K-均值聚类算法（K-Means Clustering）是从大量数据中挖掘知识的一种常用的聚类方法，其主要思想是使同类的差异最小化而不同类的差异最大化。其基本原理是：首先指定聚类数目（即 K 值），然后为每一类取一个初始聚类中心，计算每个样本到 K 个聚类中心的距离；然后根据距离最小原则将每个样本分配给距离聚类中心最近的那一类，将所有样本数据划分为 K 类；计算平均每一类的均值，将这 K 类的均值作为聚类中心再次聚类；同理依次循环直到迭代次数超过预定的阈值或者满足聚类判据的要求为止。其中计算每个样本到每一类聚类中心距离的方式有欧几里得距离、曼哈顿距离等。本文中使用的是欧几里得距离，判别截止条件为平方误差准则，具体可参考文献[11]，在此不作详细介绍。

② 基于 T-检验的 K 均值自适应聚类算法

K-均值聚类算法的优点是简单实用，在各个领域得到广泛的应用，不足之处在于它是一种静态的、离线划分方法，其聚类数目（K 值）需预先设定，不适用于海量数据的聚类，且其聚类结果是静态的，不能随着环境条件的变化而改变。针对这些不足，本文提出一种基于 T 检验的 K 均值自适应聚类算法。其中 T-检验用于检验同一类聚类成员之间的差异性，为 K 值的确定提供参考和依据，而引入自适应的思想是为了使该算法具有自学习的功能，能够对系统参数的变化做出及时的调整，以增强算法的适应性。

利用 K 均值聚类算法对交通情景进行聚类，需要确定聚类数目。以往的聚类算法中聚类数目一般人为确定，具有一定的主观性，且没有对同一类聚类成员之间的差异性（或者说聚类数目的合理性）进行检验。而判别聚类数目是否的合理性，而需要看同一类聚类成员之间的相似性，比较各聚类成员两两之间的相似度，可以利用 T 检验的方法进行检验。

T 检验常用于比较两个样本之间的均值是否有显著性的差异，从而判定两样本之间的相似性，其基本原理是：设两个总体的均值分别为 $\mu(i,j)$ 和 $\mu(m,n)$，且相互独立。设 $\bar{\lambda}(i,j)$ 和 $\bar{\lambda}(m,n)$ 为两个样本集的均值，样本容量分别为 n_1 和 n_2，$s(i,j)$ 和 $s(m,n)$ 分别为对应的标准差，T 检验的基本步骤具体包括：

1）建立原假设。H0：两个样本集的均值相等，即 $\mu(i,j)=\mu(m,n)$；H1：两个样本集的均值不相等，即 $\mu(i,j)\neq\mu(m,n)$；

2）构造统计量，由于 $s(i,j)$ 和 $s(m,n)$ 的方差不一定相等，所以，选择统计量为

$$t=\frac{(\bar{\lambda}(i,j)-\bar{\lambda}(m,n))-(\mu(i,j)-\mu(m,n))}{\sqrt{\frac{s^2(i,j)}{n_1}+\frac{s^2(m,n)}{n_2}}} \qquad (3.2)$$

3）计算 t 值及其对应的 P 值；

4）比较 P 值和显著水平 α 之间的关系，若 $P>=\alpha$（一般取 α=0.05），则拒绝假设，说明两个样本集的均值不存在显著的差异，否则说明存在显著的差异。

由于聚类结果中的聚类数目，聚类成员、聚类中心等均为未知，对于不同的输入，得到的聚类结果不同，在此引入自动控制领域的自适应的思想（所谓自适应即“系统改变自身的特性以适应环境的变化”），建立基于 T 检验的 K-均值自适应聚类算法模型。具体如图 3.6 所示：

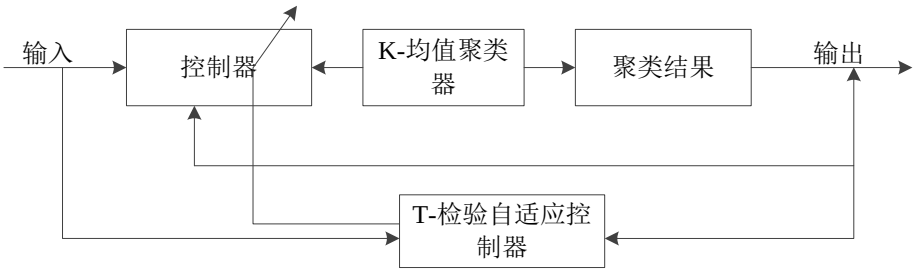


图 3.6 基于 T 检验的 K-均值自适应聚类模型

Figure 3.6 K-means self-adaptive clustering algorithm base on T test

其中各部分的内容及其作用包括：

1）输入：各车次在情景下的路段延误时间指数的均值；

2）输出：聚类数目、聚类成员、聚类中心；

3）控制器：T 检验计算，K 值的控制；

4）K-均值聚类器：根据给定的 K 值，对情景的路段延误时间指数的均值聚成 K 类；

5）聚类结果：将聚类结果模块输出，同时将输出结果中的 K 值传给控制器，将每类的聚类成员输出到 T 检验自适应控制器；

6）T 检验自适应控制器：根据聚类结果得到每一类的聚类成员，从输入中得到每个聚类成员的各车次在情景下的路段延误时间指数，对每一类聚类成员两两之间的均值进行检验，得到检验结果返回给控制器。

需要注意的是，设某一类的聚类成员数目为 n，若将各聚类成员两两之间进行 T 检验，则需要检验 n(n-1)/2 次，当 n 较大时，则检验次数将非常可观。因此，需要对该算法作进一步的改进。

注意到，比较两个样本数据集的差异性除了与两者的均值有关外，还与两者的方差（或标准差）有关。进一步，对各路段在各时间段的路段延误时间指数的均值与方差进行 Pearson 相关性分析和曲线拟合，得到各均值及对应标准差的关系如表 3.4 和表 3.5 所示：

由表 3.4 和表 3.5 可以看出，不同路段不同时间段的路段延误时间指数的均值及其标准差之间的 Pearson 相关系数达到 0.808，说明两者之间有很强的相关关系。进一步的曲线拟合结果表明，两者之间可以用线性、二次型、三次型、S 型等模型表示，其 R 方均超过 0.81，说明了路段延误时间均值中包含了相应标准差的至少 90%以上的信息。因此，可以考虑只比较同一类中均值和标准差最大的聚类成员与最小的聚类成员之间的关系（即将比较均值最小成员和均值最大成员、均值最小成员和标准差最大成员、标准差最小成员和均值最大成员、标准差最小成员和标准差最大成员 4 者的关系）即可，这样可以大大降低比较次数，提高聚类效率。

表 3.4 路段延误时间指数均值及其标准差相关性分析

Table 3.4 Correlation between road delay time index's means and standard deviation

		路段延误时间指数均值	路段延误时间指数标准差
路段延误时间指数均值	Pearson 相关性	1	**0.808**
	显著性（双侧）		0.000
	平方与叉积的和	1734.599	944.651
	协方差	0.459	0.250
	N	03780	3778
路段延误时间指数标准差	Pearson 相关性	**0.808**	1
	显著性（双侧）	0.000	
	平方与叉积的和	944.651	788.167
	协方差	0.250	0.209
	N	3778	3778

表 3.5 路段延误时间指数均值及其标准差曲线拟合结果

Table 3.5 Road delay time index's means and standard deviation curve fitting

方程	模型汇总					参数估计值		
	R 方	F	df1	df2	Sig.	b1	b2	b3
线性	**0.819**	17077.836	1	3777	0.000	0.582		
对数	0.000	0.731	1	3777	0.393	-0.008		
倒数	0.077	316.750	1	3777	0.000	0.051		
二次	**0.832**	9381.872	2	3776	0.000	0.691	-0.048	
三次	**0.833**	6264.836	3	3775	0.000	0.715	-0.070	0.003
复合	0.074	303.545	1	3777	0.000	0.653		
幂	**0.833**	18893.071	1	3777	0.000	1.333		
S	**0.845**	20665.324	1	3777	0.000	-0.406		
增长	0.074	303.545	1	3777	0.000	-0.427		
指数	0.074	303.545	1	3777	0.000	-0.427		

因变量：路段延误时间指数标准差；自变量：路段延误时间指数均值。

因此，基于 T 检验的 K-均值自适应聚类算法的具体流程如图 3.7（其中 3.7（a）为基于 T 检验的 K-均值自适应算法的整体流程图，3.7(b)为对每一类聚类成员进行 T 检验的流程图）。

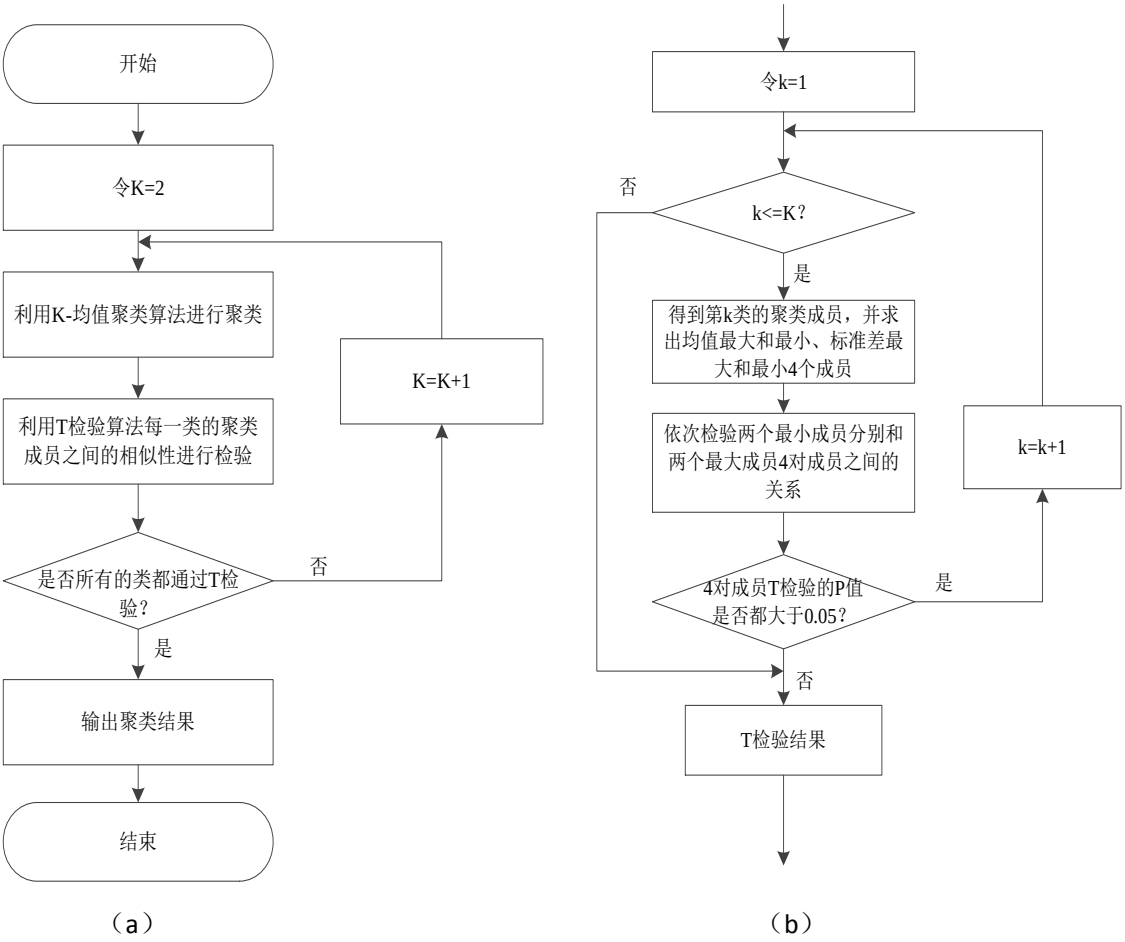


图 3.7 基于 T 检验的 K-均值自适应聚类算法流程图

Figure 3.7 The flow chart of K-means self-adaption clustering algorithm base on T test

利用基于 T 检验的 K 均值自适应聚类算法的进行交通情景的划分，具体如下：

首先统计出各初始情景下的路段延误时间指数均值 $\overline{\lambda}(r,i)$，其中 r 为路段编号，i 为时段编号。设路网中有 n 条路段（$r=1,2,\ldots,n;i=1,2,\ldots,7m$），其中 m 为每天的时段数。

其次，以 $\overline{\lambda}(r,i)$ 作为聚类成员，令 K=2 为初始值进行聚类，利用 K 均值自适应算法将 $7mn$ 个初始聚类成员（初始情景）划分为 k 类 $C_1,C_2,\ldots,C_k$；

最后，对于聚类成员数目大于 1 的类，利用 T 检验方法检验其内部各聚类成员之间的差异性。若对于所有的类都有 T 检验显著性水平（P 值）大于 0.05，则说明聚类有效，否则说明聚类无效，需要增大 K 值进一步聚类，直到满足所有类的内部成员之间的关系满足 T 检验显著性要求或 K 值达到最大预定值为止。

其中，自适应的应用主要体现在两方面：其一，在聚类过程中根据 T 检验的结果对 K 值进行调整；其二，根据过去时间的长短调整历史数据的时效性，对不同时期的数据的权重进行调整，使得过去时间越短，权重越大，而过去时间越长，

权重越小。

3.3.4 道路交通情景的划分实验结果与交通状态特征分析

道路交通情景划分的意义在于：针对不同时空下的交通拥挤程度，确定正常情况下情景划分的数目以及每一种情景所对应的交通拥堵。以 $\bar{\lambda}(r,i)$ 表征情景（r，i）交通拥堵，其中 r 表示路段而 i 表示时段，以 $\bar{\lambda}(r,i)$ 作为初始聚类成员，利用本文提出的基于 T 检验的 K-均值自适应聚类算法对道路交通情景进行划分，结果如表 3.6 所示。

由 3.6 表可以看出，利用基于 T 检验的 K-均值自适应聚类算法实现了对道路交通情景的划分，在满足 T 检验显著性水平检验（P>0.05）的基础上，根据交通拥堵历史分布的差异，将道路交通情景划分为了 8 类，每一种交通情景下交通拥挤程度及其变化程度都不同（表中 $\bar{x}$ 为均值，s 为标准差）。其中交通拥挤程度最小的是第一类情景，其路段延误时间指数的均值只有 0.231，标准差为 0.154 说明其变化较小，该情境下交通状态相对稳定。而拥挤程度最大的情景的路段延误时间指数为 5.880，其方差为 2.548，说明其交通状态波动程度较大。而且，随着不同情景下交通拥堵程度的增大，其波动程度也越大。

表 3.6 基于 T 检验的 K-均值自适应聚类算法的道路交通情景划分结果

Table 3.6 Road traffic condition classification base on K-means self-adaption clustering algorithm

聚类	聚类成员个数	$\bar{x} \pm s$	t 值*	P 值*
1	1608	0.231±0.104	1.356	0.187
2	537	0.436±0.310	1.482	0.153
3	565	0.658±0.498	0.894	0.378
4	539	1.063±0.838	0.596	0.556
5	325	1.638±1.123	1.961	0.063
6	149	2.328±1.365	1.153	0.260
7	48	3.373±1.677	1.687	0.105
8	9	5.880±2.548	1.011	0.324

注：*此处为均值最小成员和均值最大成员之间的 t 值和 P 值。

3.4 基于四分位差的道路交通偶发性拥堵量化界定

3.4.1 基于四分位差的道路交通偶发性拥堵量化界定方法

要检测道路交通偶发性拥堵，首先需要确定是否出现异常的状态。而要确定是否出现异常，首先需要明确对道路交通的正常拥堵和偶发性拥堵进行量化区分。对于高速公路（或快速路），一般均假设其交通状态具有稳定性，当出现较为严重

的拥挤时，即可认为是发生了异常。而且高速公路上偶发性拥堵（或异常事件）的确定一般都有交通管理部门的记录和实际数据作为依据。而在城市道路具有明显的早高峰和晚高峰，在高峰期间都会出现不同程度的拥堵。目前在城市道路交通方面，对于正常拥堵和偶发性拥堵的区分还没有统一、明确的标准可供参考。因此，研究道路交通偶发性拥堵检测的一个关键的问题是：如何界定道路交通正常拥堵和偶发性拥堵？

在对交通情景聚类划分的基础上，需要进一步区分每种情景下正常拥堵和偶发性拥堵的差异。针对目前对城市交通偶发性拥堵还没有统一、明确的定义和划分的情况，在此引入统计学中常用的四分位差的方法，提出了一种基于四分位差的道路交通偶发性拥堵和正常拥堵的量化区分方法。

在统计学中，四分位差法是一种广泛用于区分正常值和异常值的方法。该方法在医学、心理学、教育学等众多领域都具有广泛的应用[40][41]，其基本原理是：

首先，将样本数据作标准化处理，计算出样本数据的四个百分位数（分别用 $Q1$，$Q2$，$Q3$，$Q4$ 表示）。其中 $Q1$ 表示第一四分位数，其意义是：将所有样本数据从小到大排序，排在第 25%的位置所对应的数值就是第一四分位数；同理，$Q2$ 为排在中间（第 50%的位置）所对应的值，$Q3$ 为排在第 75%的位置上所对应的数值；$Q4$ 为样本数据中排在最后位置所对应的值（即最大值，或称极大值）。

其次，计算样本数据集的四分位差（$Interquartile\ Range, IQR$，又称四分位距），其中 $IQR = Q3 - Q1$。根据统计学中对离群点（即异常值）的定义，位于区间 $[Q1 - 1.5 * IQR，\ Q3 + 1.5 * IQR]$ 范围内的都是正常值，而超出这个范围的就是离群点。进一步，还可以将异常值分为温和异常值和极端异常值。所谓极端异常值就是超出 $[Q1 - 3 * IQR，\ Q3 + 3 * IQR]$ 范围的异常值。

由于利用四分位差的方法来计算异常值在统计学中应用非常普遍，认可度高，而偶发性拥堵就是一种异常的交通状态。因此，在此参考四分为差的方法以实现对道路交通正常拥堵和偶发性拥堵的划分。具体操作如下：

① 将每种情景下的路段延误时间指数进行标准化处理；

② 计算每种情景下的 $Q1$ 和 $Q3$，进一步计算 IQR；

③ 计算每种情景下的正常取值范围 $[Q1 - 1.5 * IQR，\ Q3 + 1.5 * IQR]$；

④ 道路交通偶发性拥堵的判别。当同一情境下连续两辆公交车的路段延误时间指数的标准化值均超出 $[Q1 - 1.5 * IQR，\ Q3 + 1.5 * IQR]$，则认为道路上发生了异常的交通拥堵。

3.4.2 道路偶发性拥堵量化界定实验结果及分析

根据四分位差的方法计算出 12 种交通情景正常交通拥堵的上限和下限，具体如表 3.7 和图 3.8 所示：

表 3.7 各交通情景的道路交通正常拥堵和偶发性拥堵的划分

Table 3.7 Difference between road traffic normal jam and contingency jam in each condition

交通情景	样本数	$\bar{x} \pm s$	极小值	极大值	百分位数		下限	上限
					25	75		
1	23041	-0.517+0.239	-0.800	3.972	-0.724	-0.629	0.800	-0.087
2	6761	-0.211+0.425	-0.800	4.512	-0.675	-0.551	0.800	0.765
3	6568	-0.063+0.467	-0.800	6.808	-0.651	-0.486	0.800	0.893
4	10620	0.257+0.817	-0.800	7.567	-0.609	-0.330	0.800	1.715
5	4807	1.034+1.257	-0.800	7.567	-0.510	0.000	0.800	4.721
6	2057	1.806+1.528	-0.800	10.981	-0.345	0.151	0.800	5.845
7	539	2.976+1.877	-0.800	12.073	-0.220	0.353	0.800	6.917
8	129	5.782+2.852	-0.492	12.105	-0.354	1.138	0.492	12.144
汇总	54522	0.000+1.000	-0.800	12.105	-0.595	0.163	-1.731	1.299

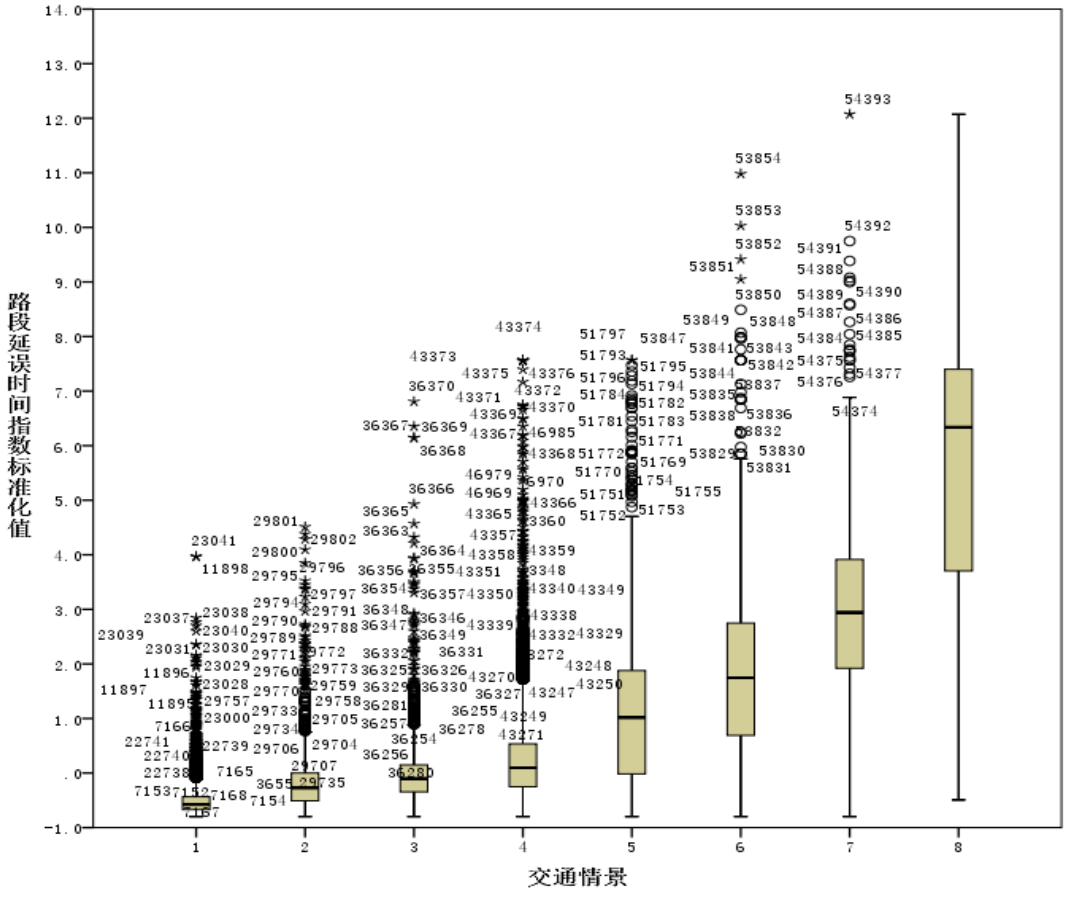


图 3.8 各情景的道路交通状态箱型图

Figure 3.8 Box-plots of road traffic states in each condition

利用箱型图将每种情景下的偶发性拥堵表示出来，如图3.8所示（图中显示的数据标签为数据行号，“o”为温和异常值，“*”为极端异常值。此外，当Q1-1.5*IQR<极小值时，阈值的下限为极小值）。图中箱型图最底下的直线为下限，最上端的直线为上限。从箱型图可以看出，不同交通情景的下限差异不大，均处于-0.9到-0.5之间，而且处于下限直线的偶发性拥堵次数极少，绝大部分的偶发性拥堵都处于控制区间的上限以上，主要对应的是平时不拥堵而现在拥堵，或者平时拥堵但现在堵的更加严重的情况。一方面，由于交通拥堵由平时拥堵而现在不拥堵的情况下控制线下限难以确定，另一方面，人们研究交通偶发性拥堵更加关注的是发生的拥堵是偶发性拥堵。因此，本文所针对的道路交通偶发性拥堵，主要是平时不拥堵而现在发生拥堵或者是平时拥堵而现在更加拥堵的情况。

经验证，利用基于“四分位差”的方法可以确定：在886线路下行方向的2-19号路段上，从2013年6月11日到2014年1月10日期间共出现偶发性拥堵次数共824次。从中随机抽取50次偶发性拥堵利用原始的公交车GPS历史数据进行逐一验证，可以发现，其中有48次为正确区分的偶发性拥堵，对交通正常拥堵和偶发性拥堵的区分正确率达到96%，说明利用“四分位差”的方法可以准确区分交通正常拥堵和偶发性拥堵。其中的两次正常拥堵误判为偶发性拥堵，其原因均是连续两辆车数据均出现错误，导致路段行程时间和路段延误时间指数计算有误。总的来说，可以看出，利用基于“四分位差”的方法可以准确实现各情境下交通正常拥堵和偶发性拥堵的量化区分。

此处所定义的道路交通偶发性拥堵将成为第5章中训练各情景阈值的基础。同时，基于四分位差的道路交通偶发性拥堵定义的方法将成为本文中划分交通偶发性拥堵的依据，并在第5章“系统实现与应用”中为检验道路交通偶发性拥堵实时检测的效果提供参考标准。

3.5 本章小结

本章首先选择合适的参数以表征道路交通状态，同时简单介绍路段和时段的划分工作，以为后续工作奠定基础；其次，以海量公交车 GPS 数据为基础分析了道路交通状态的历史分布情况；接着，针对道路交通偶发性拥堵具有相对性和情景性的特点，提出了一种基于 T 检验的 K-均值聚类算法实现道路交通状态模式识别，将交通拥堵差异较大的交通情景划分为 8 类，并确定每一类情景下交通拥堵拥挤程度及波动程度的大小；此外，利用四分位差的方法确定每类交通情景下的道路交通拥堵程度，确定道路交通正常拥堵和偶发性拥堵的界限，为后续阈值的训练和偶发性拥堵检测效果的评价提供基础。

4 基于 CVA 的道路交通偶发性拥堵检测

4.1 本章引言

在第 3 章中，基于海量公交车 GPS 历史数据的统计结果表明，不同情境下的交通状态具有显著的差异，而道路交通状态模式识别实现了对交通情景的划分，并且确定了每一类情景的交通拥堵特征。在此基础上明确了道路交通正常拥堵和偶发性拥堵的差异，作为后续工作中各情景下阈值的训练以及检测效果的评价提供依据。

道路交通系统是一个动态、非线性的复杂系统，由车辆、驾驶人员、行人、道路、天气、交通管理规章制度等众多因素组成，这些因素之间相互影响，内部机理复杂多变，若仅从历史数据统计分析的角度难以解释清楚道路交通系统的内在机理，不适合建立精确的数学模型。数据驱动是一种以数据为基础的过程工业系统故障诊断技术，它不需要建立精确数学模型即可实现对系统故障的诊断且能取得很好的效果[42][43]，而海量 GPS 数据的积累正好为利用数据驱动检测道路交通偶发性拥堵提供契机。从本质上讲，道路交通系统与工业过程系统有着比较多的相似之处，两者都是实时动态变化的系统，都希望能够维持稳定，但有不可避免会出现偶发性拥堵。相对于正常的拥堵而言，在道路交通系统的偶发性拥堵与工业系统的故障在本质上是相同的。

因此，本文将交通偶发性拥堵看成道路交通系统的故障，引入工业控制系统中故障诊断的思想和方法——数据驱动技术用于道路交通偶发性拥堵的检测过程，用于分析道路交通拥堵的实时变化趋势。

4.2 实时参数分析

对于实时的公交 GPS 数据，由于车辆处于运行状态，还没有走完相应的路段，不能得到准确的路段行程时间。实时 GPS 数据直接可用的参数为瞬时速度，但如前面所述，用实时 GPS 速度不能完全表征实时路况交通拥堵的信息。进一步深入分析发现，实时数据中除了瞬时速度之外，有可以计算其他速度参数用以表征道路交通状态的特征及其趋势，具体如下：

① 瞬时速度 $v_1(i,t)$：直接参数，表征公交车 i 在 t 时刻的瞬时状态；

② 周期平均速度 $v_2(i,t)$：车辆 i 在每周期内平均速度，表征车辆在 t 时刻前一周期内的趋势，计算公式为：

$$v_2(i,t)=\frac{l(i,t)}{\tau} \tag{4.1}$$

其中$l(i,t)$为车辆i在周期 t 内所走的距离，τ 为采样周期长度；

③ 加权滑动平均速度$v_3(i,t)$：表征车辆i在 t 时刻前n个周期内所有瞬时速度的均值，计算公式为：

$$v_3(i,t)=\sum_{j=1}^{n}\frac{v_j(i)}{n} \tag{4.2}$$

其中$v_j(i)$为车辆i在前第j个周期内的瞬时速度，n为周期数，此处各瞬时速度加权系数相等，均为 1/*n*；

④ 多车平均速度$v_4(t)$：在同一周期内经过同一路段的所有线路车辆的平均速度，其计算公式为：

$$v_4(t)=\begin{cases}\sum_{j=1}^{c}\frac{v_j}{c} & c\geq 1\\ y_i(t-1) & c=0\end{cases} \tag{4.3}$$

其中 c 为采样周期 t 内在相同路段上的公交车数量，v_j为车辆j的瞬时速度。

其中瞬时速度表征的是某一车辆的瞬时状态，周期平均速度表征的是某一车辆在过去一个周期内的状态，加权滑动平均速度表征某辆车过去连续多周期内的序列值，多车平均速度表征多辆车在同一周期内整体的状态。理想情况下，当车辆在路段上匀速行走时，有$v_1(t)=v_2(t)=v_3(t)=v_4(t)$，但在实际中受不同路段路况和司机驾驶行为差异等因素的影响，各个速度参数变化不稳定，以上等式一般不成立。结合一辆车的瞬时状态和过去变化趋势以及同一路段上多车的变化趋势，更能加有利于实现对交通偶发性拥堵及时、准确的检测。

另外，需要注意的是：当车辆处于停靠站上下客的拥堵时，属于正常的公交停车行为，此时的公交 GPS 瞬时速度为 0，不足以表征道路的交通状态，因此，需要对此种状态下的车辆行驶速度进行修正。具体的方法是：首先确定车辆是否站间停车的状态，若是，则利用车辆进站前两个周期的瞬时速度的均值修正站内的车辆速度。同理，可以对信号灯影响区域内处于停车拥堵的公交车辆速度进行修正。

当车辆处于运行拥堵时，可以利用速度参数来计算“路段延误时间指数”，如式 4.4 所示：

$$\lambda(r,i)=\frac{l_r/v_{(r,i)}-l_r/v_{r\max}}{l_r/v_{r\max}}=\frac{v_{r\max}}{v_{(r,i)}}-1 \tag{4.4}$$

上式中l_r为路段r的长度，$v_{r\max}$为在路段r上的最大限速,$v_{(r,i)}$为车辆在路段r上第i时刻的瞬时速度且有$v_{(r,i)}>0$。当$v_{(r,i)}=0$时，则需要进一步判断车辆是否处于站内停车拥堵或信号灯区域等待状态，若是，则利用上述方法对车辆行驶速度

进行修正；若否，则令 $\lambda(r,i)=\lambda_{\max}$（其中 $\lambda_{\max}$ 为利用式 3-1 从海量历史数据中统计出来的最大值）。

参考我国公安部在 2002 年发布的《城市交通管理评价指标体系》里的交通拥挤等级划分方法，得到 4 个道路交通状态等级。根据重庆市公交车 GPS 数据，公交车在城市主干路的最大限速为 $45\mathrm{km/h}$。当车辆保持最大限速经过某路段时，则其路段行程时间达到最小，因此可以计算出每个等级阈值对应的路段延误时间指数阈值，如表 4.1 所示：

表 4.1 路段延误时间指数与交通拥挤等级关系

Table 4.1 The relationship between Road delay time index and traffic congestion level

交通拥堵	速度 v 范围（km/h）	路段延误时间指数 λ 范围
畅通	>=30	<=0.667
轻度拥挤	[20,30)	(0.667-1.667]
拥挤	[10,20)	(1.667-4.667]
严重拥挤	<10	>4.667

由表 4.1 可以看出，“路段延误时间指数”不仅可以消除不同路段长度和等级的差异，它同样能描述路段的交通拥挤程度，且该参数还可以由微观上的实时速度参数和中观上的路段行程时间参数计算得到，可以同时用于计算历史的交通状态和实时道路交通状态，更具有灵活性。此外，该参数与速度参数具有倒数关系，如果用速度表征交通状态，则在畅通状态下的速度变化范围比较大，而随着拥挤程度的增大，速度的变化范围越来越小。而路段延误时间指数则减小了畅通状态下参数的变化范围，放大了拥挤拥堵下的参数变化范围，更加有利于对拥堵状态下的偶发性拥堵的检测。

实时参数相关性分析主要检验以上 4 个实时参数两两之间的相关性以及各参数内部的自相关性，对参数的有效性进行验证，同时也为后续道路交通偶发性拥堵检测方法的选择提供参考。其中参数间的相关系数采用 Pearson 相关系数表示，参数内部的自相关关系采用 Box-Ljung 统计量作为评价指标[44]，以重庆市的日间公交线路 886 线路的 16 号路段在 2014 年 1 月 10 号一天的数据为例，对 4 个速度参数的之间的相关性和各变量内部的自相关性进行检验。

① 不同变量之间相关性分析

分析 4 个速度变量两两之间的相关关系，结果如表 4.2 所示：

表 4.2 不同速度变量之间的相关关系分析

Table 4.2 Correlation coefficient between different velocity variables

		瞬时速度	周期平均速度	加权滑动平均速度	多车平均速度
瞬时速度	Pearson 相关性	1	0.008	0.599	0.544
	显著性（双侧）		0.859	0.000	0.000
	N	15468	15468	15468	15468
周期平均速度	Pearson 相关性	0.008	1	-0.008	-0.016
	显著性（双侧）	**0.859**		0.852	0.712
	N	15468	15468	15468	15468
加权滑动平均速度	Pearson 相关性	0.599	-0.008	1	0.383
	显著性（双侧）	0.000	**0.852**		0.000
	N	15468	15468	15468	15468
多车平均速度	Pearson 相关性	0.544	-0.016	0.383	1
	显著性（双侧）	0.000	**0.712**	0.000	
	N	15468	15468	15468	15468

表 4.2 的 Pearson 相关系数及其显著性水平表明，4 个变量两两之间的 Pearson 相关系均小于 0.6，其中瞬时速度与加权滑动平均速度的相关系数为 0.599，恰好小于 0.6，说明这些变量两两之间的相关性不强。而周期平均速度与其他速度的显著水平以及多车平均速度和加权滑动平均速度的显著水平均大于 0.05，说明这些速度变量之间不存在明显的相关关系，周期速度包含的信息绝大部分都是其他速度变量中没有包含的信息。

此外需要注意的是，为了使加权滑动平均速度更多包含其他变量未包含的信息，其周期数的确定使得该变量与其他变量的相关系数尽可能地小，通过试凑的方法得到 n=8 得到的效果最好。

② 各变量的自相关性分析

利用统计学中常见的Box-Ljung统计量作为评价指标分析4个实时参数的在不同滞后周期的自相关性（其中最大滞后周期取 16），其结果如表 4.3、表 4.4 和图 4.1 所示。

从自相关分析结果可以看出，除了周期平均速度外，其他三个指标在不同的滞后周期下的Box-Ljung统计量均超过100，对应的Sig.值均为0.000，说明这三个速度变量均有较强的自相关性，而周期平均速度在不同的滞后周期下的Box-Ljung统计量均小于15，对应的Sig.值均大于0.05，说明该变量没有明显的自相关性。考虑

到大多数变量的自相关性，为下文采用CVA提供理论基础。

表 4.3 瞬时速度和周期平均速度自相关关系分析

Table 4.3 Autocorrelation coefficient of instantaneous velocity and average velocity in a cycle

滞后周期	瞬时速度			滞后周期	周期平均速度		
	自相关系数	Box-Ljung 统计量	Sig.		自相关系数	Box-Ljung 统计量	Sig.
1	0.696	259.35	0.000	1	0.008	0.04	0.846
2	0.433	359.98	0.000	2	0.082	3.64	0.162
3	0.252	394.23	0.000	3	-0.019	3.84	0.280
4	0.126	402.85	0.000	4	-0.016	3.97	0.410
5	0.119	410.47	0.000	5	-0.005	3.99	0.552
6	0.186	429.15	0.000	6	0.078	7.29	0.295
7	0.199	450.62	0.000	7	-0.046	8.44	0.295
8	0.209	474.43	0.000	8	0.016	8.58	0.379
9	0.221	501.06	0.000	9	-0.015	8.69	0.466
10	0.182	519.12	0.000	10	0.035	9.36	0.499
11	0.153	531.96	0.000	11	0.015	9.47	0.578
12	0.137	542.21	0.000	12	-0.023	9.76	0.637
13	0.099	547.60	0.000	13	0.026	10.13	0.683
14	0.086	551.68	0.000	14	-0.042	11.09	0.679
15	0.076	554.87	0.000	15	0.066	13.46	0.567
16	0.059	556.78	0.000	16	0.041	14.39	0.569

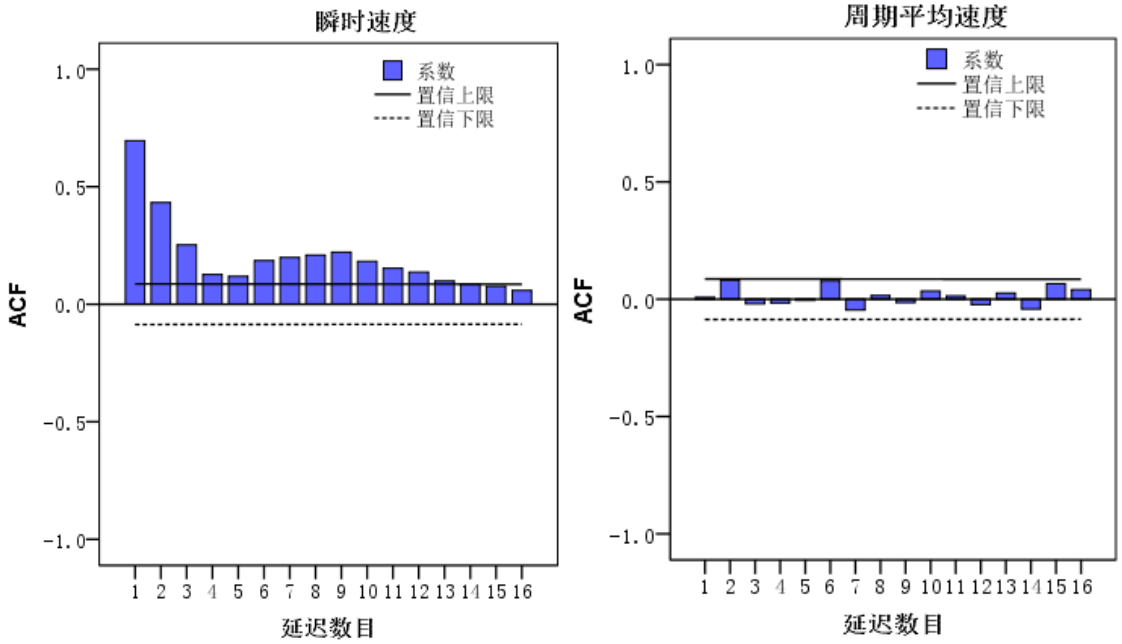
表 4.4 加权滑动平均速度和多车平均速度自相关关系分析

Table 4.4 Autocorrelation coefficient of weighted moving average velocity and average velocity of multi-bus

滞后周期	瞬时速度			滞后周期	周期平均速度		
	自相关系数	Box-Ljung 统计量	Sig.		自相关系数	Box-Ljung 统计量	Sig.
1	0.799	342.11	0.000	1	0.663	235.77	0.000
2	0.599	534.68	0.000	2	0.383	314.66	0.000
3	0.432	635.04	0.000	3	0.185	333.08	0.000
4	0.327	692.55	0.000	4	0.087	337.17	0.000

续表 4.4：

滞后周期	瞬时速度			滞后周期	周期平均速度		
	自相关系数	Box-Ljung 统计量	Sig.		自相关系数	Box-Ljung 统计量	Sig.
5	0.277	733.97	0.000	5	0.055	338.78	0.000
6	0.272	774.10	0.000	6	0.090	343.19	0.000
7	0.262	811.36	0.000	7	0.110	349.79	0.000
8	0.239	842.44	0.000	8	0.159	363.46	0.000
9	0.213	867.23	0.000	9	0.210	387.44	0.000
10	0.166	882.23	0.000	10	0.175	404.12	0.000
11	0.156	895.46	0.000	11	0.105	410.14	0.000
12	0.176	912.48	0.000	12	0.089	414.48	0.000
13	0.175	929.28	0.000	13	0.062	416.57	0.000
14	0.169	944.90	0.000	14	0.051	418.03	0.000
15	0.128	953.98	0.000	15	0.020	418.25	0.000
16	0.084	957.89	0.000	16	-0.012	418.34	0.000



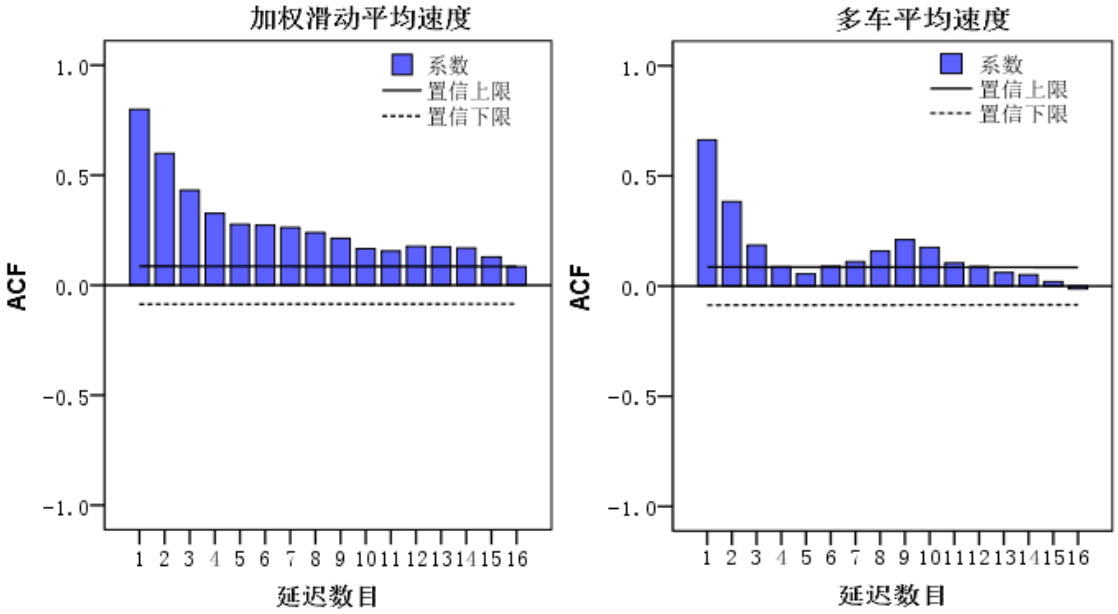


图 4.1 各速度变量在不同滞后周期的自相关系数分布

Figure4.1 Autocorrelation coefficient of different variables in different cycle

4.3 CVA 原理简介

故障诊断技术包含多种不同的算法，其中较为传统的有主元分析、小波分析、偏最小二乘法等[45]。这些算法的共同点是假定样本数据均有独立同分布的特点。从 4.1 节的分析中可以看出，可以利用瞬时速度、周期平均速度、加权滑动平均速度和多车平均速度这 4 个实时参数来表征道路交通的实时状态，经过对这 4 个速度变量的相关分析结果表明：除了周期平均速度之外，瞬时速度、加权滑动平均速度和多车平均速度这三个参数都具有很强的自相关性，传统算法往往忽略了该特点，而规范变量分析（Canonical Variate Analysis，CVA，也叫 Canonical Correlation analysis，CCA）算法则考虑到了这一点[46][47]。CVA 算法发展于最优的统计推理原理，并且其最优的统计精度已得到证明[48]。

CVA 一般用于工业系统过程的故障诊断，并且能取得很好的效果，但却没有用于道路交通偶发性拥堵的检测。从本质上讲，道路交通出现偶发性拥堵实际上就是道路交通系统发生了故障。因此，在此引入 CVA 算法用于分析道路交通状态的实时变化趋势，包括以下 3 个部分的内容：CVA 原理简介，模型构建和道路交通偶发性拥堵检测的实现。

CVA 的思想是最大化两个数据集的相关关系[48]。对于过程变量数据集（尤其是具有时间序列特性的数据集），可以分为历史数据集与未来数据集两部分，而 CVA 就是通过最大化过去数据集与未来数据集的关系，从而提取系统特征信息，建立分析模型。

设有 $m+n$ 维序列值（ $\boldsymbol{x}_{k-m}, \boldsymbol{x}_{k-m+1}, \ldots, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k}, \boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_{k+n-1}$ ），以 k 时刻为界分

为 $\boldsymbol{X} \in R^{m\times q}$ 和 $\boldsymbol{Y} \in R^{n\times q}$ 两部分，其中 q 为变量个数，$\boldsymbol{X}$ 为时刻 k 前 m 个周期的序列值，$\boldsymbol{Y}$ 为时刻 k 及其以后 n 个周期的序列值，则有 $\boldsymbol{X} = [\boldsymbol{x}_{k-m}, \boldsymbol{x}_{k-m+1}, \cdots, \boldsymbol{x}_{k-1}]^T$，$\boldsymbol{Y} = [\boldsymbol{y}_k, \boldsymbol{y}_{k+1}, \cdots, \boldsymbol{y}_{k+n-1}]^T$。对 $\boldsymbol{X}$，$\boldsymbol{Y}$ 的每个样本值进行标准化变换，变换公式为：

$$\boldsymbol{x}^* = \frac{\boldsymbol{x} - \boldsymbol{\mu}_x}{\boldsymbol{\sigma}_x} \tag{4.5}$$

其中 $\boldsymbol{\mu}_x$ 为 $\boldsymbol{x}$ 的均值，$\boldsymbol{\sigma}_x$ 为 $\boldsymbol{x}$ 的标准差，则 $\boldsymbol{X},\boldsymbol{Y}$ 变换后得到 $\boldsymbol{X}^*,\boldsymbol{Y}^*$，其中 $\boldsymbol{X}^* = [\boldsymbol{x}^*_{k-m}, \boldsymbol{x}^*_{k-m+1}, \cdots, \boldsymbol{x}^*_{k-1}]^T$，$\boldsymbol{Y}^* = [\boldsymbol{y}^*_k, \boldsymbol{y}^*_{k+1}, \cdots, \boldsymbol{y}^*_{k+n-1}]^T$。此时，分析随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的相关关系，可转化为分析 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的相关关系。

记 $\mathrm{cov}(\boldsymbol{X}^*,\boldsymbol{Y}^*)$ 为 $\boldsymbol{X}^*,\boldsymbol{Y}^*$ 的协方差矩阵，则有

$$\mathrm{cov}(\boldsymbol{X}^*,\boldsymbol{Y}^*) = E(\boldsymbol{X}^* - E\boldsymbol{X}^*)(\boldsymbol{Y}^* - E\boldsymbol{Y}^*)^T \tag{4.6}$$

记 $\mathrm{cov}(\boldsymbol{X}^*,\boldsymbol{Y}^*)$ 为 $\sum{}^* = (\sigma_{ij})_{m\times n}$，其中

$$\sigma_{ij} = \mathrm{cov}(\boldsymbol{x}^*_i, \boldsymbol{y}^*_j)(i = k-m, k-m+1, \cdots, k-1;\ j = k, k+1, ?\quad k+n-1) \tag{4.7}$$

将 $\sum{}^*$ 进行分解，有

$$\sum{}^* = \begin{bmatrix} \sum_{x^*x^*} & \sum_{x^*y^*} \\ \sum_{y^*x^*} & \sum_{y^*y^*} \end{bmatrix} \tag{4.8}$$

其中 $\sum{}^*_{xx}$ 为变量 $\boldsymbol{X}^*$ 的协方差矩阵，$\sum{}^*_{yy}$ 为变量 $\boldsymbol{Y}^*$ 的协方差矩阵，$\sum{}^*_{xy} = (\sum_{yx})^T$ 为变量 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的互协方差矩阵。

记 $\boldsymbol{\rho}_{ij} = \boldsymbol{\rho}(\boldsymbol{x}^*_i, \boldsymbol{y}^*_j)(i = k-m, k-m+1, \cdots, k-1;\ j = k, k+1, \cdots, k+n-1)$ 为 $(\boldsymbol{x}^*_i, \boldsymbol{y}^*_j)$ 的相关系数，其中

$$\boldsymbol{\rho}(\boldsymbol{x}^*_i, \boldsymbol{y}^*_j) = \frac{\mathrm{cov}(\boldsymbol{x}^*_i, \boldsymbol{y}^*_j)}{\sqrt{D\boldsymbol{X}^*}\sqrt{D\boldsymbol{Y}^*}} \tag{4.9}$$

记 $\boldsymbol{R}^* = \boldsymbol{\rho}(\boldsymbol{X}^*,\boldsymbol{Y}^*)$ 为 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的相关系数矩阵同样，将 $\boldsymbol{R}^*$ 进行分解，则有

$$\boldsymbol{R}^* = \boldsymbol{\rho}(\boldsymbol{X}^*,\boldsymbol{Y}^*) = (\rho(\boldsymbol{x}^*_i, \boldsymbol{y}^*_j))_{m\times n} = \begin{bmatrix} \boldsymbol{R}_{x^*x^*} & \boldsymbol{R}_{x^*y^*} \\ \boldsymbol{R}_{y^*x^*} & \boldsymbol{R}_{y^*y^*} \end{bmatrix} \tag{4.10}$$

将 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 进行线性变换，得到两个新变量 $\boldsymbol{U}$ 和 $\boldsymbol{V}$，其中

$$\begin{cases} \boldsymbol{U} = \boldsymbol{a}^T\boldsymbol{X}^* = a_1\boldsymbol{x}^*_{k-m} + a_2\boldsymbol{x}^*_{k-m+1} + \ldots + a_m\boldsymbol{x}^*_{k-1} \\ \boldsymbol{V} = \boldsymbol{b}^T\boldsymbol{Y}^* = b_1\boldsymbol{y}^*_k + b_2\boldsymbol{y}^*_{k+1} + \ldots + b_n\boldsymbol{y}^*_{k+n-1} \end{cases} \tag{4.11}$$

其中 $\boldsymbol{a} = [a_1, a_2, \ldots, a_m]$，$\boldsymbol{b} = [b_1, b_2, \cdots, b_n]$

此时，分析变量 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的相关关系转化为了分析变量 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的相关关系，则有

$$\mathrm{cov}(\boldsymbol{U},\boldsymbol{V}) = \mathrm{cov}(\boldsymbol{a}^T\boldsymbol{X}^*, \boldsymbol{b}^T\boldsymbol{Y}^*) = \boldsymbol{a}^T\,\mathrm{cov}(\boldsymbol{X}^*,\boldsymbol{Y}^*)\boldsymbol{b} \tag{4.12}$$

$$\begin{cases} D(\boldsymbol{U}) = D(\boldsymbol{a}^T\boldsymbol{X}^*) = \boldsymbol{a}^T D(\boldsymbol{X}^*)\boldsymbol{a} = \boldsymbol{a}^T\sum_{x^*x^*}\boldsymbol{a} \\ D(\boldsymbol{V}) = D(\boldsymbol{b}^T\boldsymbol{Y}^*) = \boldsymbol{b}^T D(\boldsymbol{Y}^*)\boldsymbol{b} = \boldsymbol{b}^T\sum_{y^*y^*}\boldsymbol{b} \end{cases} \tag{4.13}$$

因此，可得$\boldsymbol{U}$和$\boldsymbol{V}$的相关系数为

$$\boldsymbol{\rho}(\boldsymbol{U},\boldsymbol{V})=\frac{\boldsymbol{a}^{T}\sum_{x^{*}y^{*}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^{T}\sum_{x^{*}x^{*}}\boldsymbol{a}}\sqrt{\boldsymbol{b}^{T}\sum_{y^{*}y^{*}}\boldsymbol{b}}} \tag{4.14}$$

CVA 的思想就是要使变量$\boldsymbol{U}$和$\boldsymbol{V}$的相关关系最大化，即$\boldsymbol{\rho}(\boldsymbol{U},\boldsymbol{V})$最大化。

由于$\boldsymbol{X}^{*}$和$\boldsymbol{Y}^{*}$是经过标准化处理的变量，有$D(\boldsymbol{X}^{*})=1$，$D(\boldsymbol{Y}^{*})=1$，又因为$\boldsymbol{U}$和$\boldsymbol{V}$的是$\boldsymbol{X}^{*}$和$\boldsymbol{Y}^{*}$的线性函数，因此有

$$\begin{cases}D(\boldsymbol{U})=\boldsymbol{a}^{T}\sum_{x^{*}x^{*}}\boldsymbol{a}=1\\ D(\boldsymbol{V})=\boldsymbol{b}^{T}\sum_{y^{*}y^{*}}\boldsymbol{b}=1\end{cases} \tag{4.15}$$

因此，

$$\rho(\boldsymbol{U},\boldsymbol{V})=\frac{\boldsymbol{a}^{T}\sum_{x^{*}y^{*}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^{T}\sum_{x^{*}x^{*}}\boldsymbol{a}}\sqrt{\boldsymbol{b}^{T}\sum_{y^{*}y^{*}}\boldsymbol{b}}}=\boldsymbol{a}^{T}\sum_{x^{*}y^{*}}\boldsymbol{b} \tag{4.16}$$

因此，求两组变量$\boldsymbol{X}$、$\boldsymbol{Y}$的相关关系最大化的问题就转化成了在满足（4.15）的约束下，使得式（4.16）的值最大化。为求得在约束条件下的极值，在此引入拉格朗日（Lagrange）系数，则转化为求式（4-13）的极大值。

$$g(\boldsymbol{a},\boldsymbol{b})=\boldsymbol{a}^{T}\sum_{x^{*}y^{*}}\boldsymbol{b}-\frac{\lambda}{2}(\boldsymbol{a}^{T}\sum_{x^{*}x^{*}}\boldsymbol{a}-1)-\frac{\gamma}{2}(\boldsymbol{b}^{T}\sum_{y^{*}y^{*}}\boldsymbol{b}-1) \tag{4.17}$$

分别对上式求$\boldsymbol{a}$和$\boldsymbol{b}$的偏导，并令其等于 0，可得

$$\begin{cases}\frac{\partial g}{\partial \boldsymbol{a}}=\sum_{x^{*}y^{*}}\boldsymbol{b}-\lambda\sum_{x^{*}x^{*}}\boldsymbol{a}=0\\ \frac{\partial g}{\partial \boldsymbol{b}}=\sum_{y^{*}x^{*}}\boldsymbol{a}-\gamma\sum_{y^{*}y^{*}}\boldsymbol{b}=0\end{cases} \tag{4.18}$$

根据上式变换求解可得

$$\lambda=\gamma=\boldsymbol{a}^{T}\sum_{x^{*}y^{*}}\boldsymbol{b} \tag{4.19}$$

这说明了$\lambda=\gamma$即为线性组合$\boldsymbol{X}^{*}$和$\boldsymbol{Y}^{*}$的相关系数。对式（4.19）进行变换处理可得

$$\begin{cases}\sum^{-1}_{x^{*}x^{*}}\sum_{x^{*}y^{*}}\sum^{-1}_{y^{*}y^{*}}\sum_{y^{*}x^{*}}\boldsymbol{a}-\lambda^{2}\boldsymbol{a}=0\\ \sum^{-1}_{y^{*}y^{*}}\sum_{y^{*}x^{*}}\sum^{-1}_{x^{*}x^{*}}\sum_{x^{*}y^{*}}\boldsymbol{b}-\lambda^{2}\boldsymbol{b}=0\end{cases} \tag{4.20}$$

令

$$\begin{cases}\sum^{-1}_{x^{*}x^{*}}\sum_{x^{*}y^{*}}\sum^{-1}_{y^{*}y^{*}}\sum_{y^{*}x^{*}}=\boldsymbol{A}\\ \sum^{-1}_{y^{*}y^{*}}\sum_{y^{*}x^{*}}\sum^{-1}_{x^{*}x^{*}}\sum_{x^{*}y^{*}}=\boldsymbol{B}\end{cases} \tag{4.21}$$

则有

$$\begin{cases}\boldsymbol{A}\boldsymbol{a}=\lambda^{2}\boldsymbol{a}\\ \boldsymbol{B}\boldsymbol{b}=\lambda^{2}\boldsymbol{b}\end{cases} \tag{4.22}$$

由此可见，λ^2 同时为 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的特征根，而 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 即为 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 对应的特征向量。由（4.22）可以看出，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的特征根相同且在0和1之间。令 $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_m^2 > 0$ 表示 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的特征根，则 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ 即为 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的相关系数。$\lambda_1, \lambda_2, \cdots, \lambda_m$ 对应的特征向量 $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \cdots, \boldsymbol{a}^{(m)}$ 和 $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \cdots, \boldsymbol{b}^{(m)}$，则可得到 m 对 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的线性组合：

$$\begin{cases} u_1 = \boldsymbol{a}^{(1)T}\boldsymbol{X}^*, v_1 = \boldsymbol{b}^{(1)T}\boldsymbol{Y}^* \\ u_2 = \boldsymbol{a}^{(2)T}\boldsymbol{X}^*, v_1 = \boldsymbol{b}^{(2)T}\boldsymbol{Y}^* \\ \cdots\cdots \\ u_m = \boldsymbol{a}^{(m)T}\boldsymbol{X}^*, v_m = \boldsymbol{b}^{(m)T}\boldsymbol{Y}^* \end{cases} \tag{4.23}$$

称 $(u_1, v_1), (u_2, v_2), \ldots, (u_m, v_m)$ 为 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的 m 对规范变量。因此，求 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的相关系数和规范变量归根到底就是要求矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的特征根及对应的特征向量。

此外，为了保证 $\boldsymbol{X}^*$ 和 $\boldsymbol{Y}^*$ 的相关度达到最大（即防止前面变量已经提取出来的信息在后面重复表达），m 个特征向量组 $(\boldsymbol{a}^{(1)}, \boldsymbol{b}^{(1)}), (\boldsymbol{a}^{(2)}, \boldsymbol{b}^{(2)}), \cdots, (\boldsymbol{a}^{(m)}, \boldsymbol{b}^{(m)})$ 两两之间互不相关，即

$$\begin{cases} \mathrm{cov}(u_i, u_j) = \mathrm{cov}(\boldsymbol{a}^{(i)T}\boldsymbol{X}^*, \boldsymbol{a}^{(j)T}\boldsymbol{X}^*) = \boldsymbol{a}^{(i)T}\boldsymbol{X}^*\boldsymbol{a}^{(j)} = 0 \\ \mathrm{cov}(v_i, v_j) = \mathrm{cov}(\boldsymbol{b}^{(i)T}\boldsymbol{Y}^*, \boldsymbol{b}^{(j)T}\boldsymbol{Y}^*) = \boldsymbol{b}^{(i)T}\boldsymbol{Y}^*\boldsymbol{b}^{(j)} = 0 \end{cases} \tag{4.24}$$

式（4.24）中 $1 \leq i \leq m, 1 \leq j \leq m$ 且 $i \neq j$。

由于总体样本的均值向量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 一般为未知，一般可用 $\boldsymbol{\mu}$ 的极大似然值（即样本的均值向量 $\overline{\boldsymbol{x}}^T$ 和 $\overline{\boldsymbol{y}}^T$）来估计 $\boldsymbol{\mu}$，用样本的协方差矩阵 $\boldsymbol{S}$ 或相关矩阵 $\boldsymbol{R}$ 估计 $\boldsymbol{\Sigma}$。其中各变量样本值经过标准化处理，且

$$\begin{cases} \overline{\boldsymbol{x}}^T = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\right)^T \\ \overline{\boldsymbol{y}}^T = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i\right)^T \end{cases} \tag{4.25}$$

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{xx} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{S}_{yy} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{xx} & \boldsymbol{R}_{xy} \\ \boldsymbol{R}_{yx} & \boldsymbol{R}_{yy} \end{bmatrix} = \boldsymbol{R} \tag{4.26}$$

$$\begin{cases} \boldsymbol{S}_{xx} = \dfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \\ \boldsymbol{S}_{yy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T \\ \boldsymbol{S}_{xy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T \\ \boldsymbol{S}_{yx} = \dfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \end{cases} \tag{4.27}$$

可用 $\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{yx}$ 和 $\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}$ 来估计 $\sum_{xx}^{-1}\sum_{xy}\sum_{yy}^{-1}\sum_{yx}$ 和 $\sum_{yy}^{-1}\sum_{yx}\sum_{xx}^{-1}\sum_{xy}$，以非 0 特征值 $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_m^2$ 来估计相关系数 $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_m^2$，用其对应的特征向量 $\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{a}}_2, \cdots, \hat{\boldsymbol{a}}_m$ $\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2, \cdots, \hat{\boldsymbol{b}}_m$ 来估计 $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m$ 和 $\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m$，进一步计算出样本 i 的第 j 个规范变量的值，如下式所示。

$$\begin{cases} u_{ij} = \hat{\boldsymbol{a}}_j^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) \\ v_{ij} = \hat{\boldsymbol{b}}_j^T(\boldsymbol{y}_i - \overline{\boldsymbol{y}}) \end{cases} \tag{4.28}$$

其中 i 为第 i 个样本，而 j 代表第 j 个规范变量。对于规范变量 j，画出 (u_{ij}, v_{ij}) 的散点图，根据该散点图的分布可以表征原始数据的变化趋势。

4.4 CVA 模型构建

设给定时间序列的过程变量输入数据集为 $\boldsymbol{u}(t) \in R^{m\times q}$，其输出为 $\boldsymbol{y}(t) \in R^{n\times q}$，则可建立一个带输入的回归滑动平均拥堵空间模型如下：

$$\begin{cases} \boldsymbol{x}(t+1) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{w}(t) \\ \boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) + \boldsymbol{H}\boldsymbol{w}(t) + \boldsymbol{v}(t) \end{cases} \tag{4.29}$$

其中 $\boldsymbol{x}(t) \in R^{k\times q}$ 为 k 阶拥堵向量，$\boldsymbol{A} \in R^{k\times k}$、$\boldsymbol{B} \in R^{k\times m}$、$\boldsymbol{C} \in R^{n\times k}$、$\boldsymbol{D} \in R^{n\times m}$、$\boldsymbol{H} \in R^{n\times k}$，$\boldsymbol{w}(t)$ 和 $\boldsymbol{v}(t)$ 为白噪声序列，t 为任意时刻，q 为变量个数。

CVA 是在过去状态和未来状态相关程度最大化的基础上，根据过去状态信息得到未来状态的预测信息，将预测信息与实际信息对比从而检测偶发性拥堵。对于一个时间序列样本数据 $t = 1, 2, \ldots, n$，以 t 为任意当前时刻，则包含过去信息的输出向量为：

$$\boldsymbol{y}_p(t) = [\boldsymbol{y}^T(t-1), \boldsymbol{y}^T(t-2), \cdots, \boldsymbol{y}^T(t-p), \boldsymbol{u}^T(t-1), \boldsymbol{u}^T(t-2), \cdots, \boldsymbol{u}^T(t-p)]^T \tag{4.30}$$

包含过去信息的输出向量为：

$$\boldsymbol{y}_f(t) = [\boldsymbol{y}^T(t), \boldsymbol{y}^T(t+1), \cdots, \boldsymbol{y}^T(t+f-1)]^T \tag{4.31}$$

Hankel 矩阵为：

$$\begin{aligned} \boldsymbol{Y}_p &= [\boldsymbol{y}_p(t) \quad \boldsymbol{y}_p(t+1)) \quad \cdots \quad \boldsymbol{y}_p(t+\mathrm{N}-1)] \in R^{(n+m)p\times N} \\ \boldsymbol{Y}_f &= [\boldsymbol{y}_f(t) \quad \boldsymbol{y}_f(t+1)) \quad \cdots \quad \boldsymbol{y}_f(t+\mathrm{N}-1)] \in R^{mp\times N} \end{aligned} \tag{4.32}$$

其中 N 为模型数据长度，p 和 f 分别为过去和未来信息的长度，且一般取 $f > p$。令 Σ_{pp} 和 Σ_{ff} 分别为过去数据集 $\boldsymbol{Y}_p(t)$ 和未来数据集 $\boldsymbol{Y}_f(t)$ 的协方差矩阵，Σ_{pf} 两个数据集的交叉协方差矩阵。

实现 CVA 首先要实现对未来和过去两个数据集的奇异值分解（SVD，Singular Value Decomposition），则可得到下式

$$\boldsymbol{\Sigma}_{ff}^{-1/2}\boldsymbol{\Sigma}_{fp}\Sigma_{pp}^{-1/2} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{4.33}$$

其中 Σ 为从大到小依次排序的规范系数矩阵（$\Sigma = diag(\lambda_1, \lambda_2, \ldots, \lambda_r, \ldots)$）。令 $\boldsymbol{J} = \boldsymbol{U}^T\Sigma_{pp}^{-1/2}$，当过程噪声小于一定的阈值时，一般取前 k 个特征值及其相应的特征向量即可包含绝大部分的信息。则时刻 t 对应的拥堵空间向量的求取如下式所示：

$$\boldsymbol{x}(t) = \boldsymbol{U}_k^T\boldsymbol{\Sigma}_{pp}^{-1/2}\boldsymbol{y}_p(t) = J_k\boldsymbol{y}_p(t) \tag{4.34}$$

其中 $\boldsymbol{U}_k$ 为奇异值分解矩阵中 $\boldsymbol{U}$ 的前 k 列（k 值根据特征根的累积贡献率选取），此时 $\boldsymbol{x}(t)$ 为 k 阶最优拥堵。

进一步，求出状态空间矩阵的各项系数。由状态空间方程得

$$\begin{bmatrix} \boldsymbol{m}(t+1) \\ \boldsymbol{y}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}\begin{bmatrix} \boldsymbol{m}(t) \\ \boldsymbol{u}(t) \end{bmatrix} + \begin{bmatrix} I & \boldsymbol{0} \\ \boldsymbol{E} & I \end{bmatrix}\begin{bmatrix} \boldsymbol{w}(t) \\ \boldsymbol{v}(t) \end{bmatrix} \tag{4.35}$$

由于 $\boldsymbol{w}(t)$ 和 $\boldsymbol{v}(t)$ 为白噪声序列，有 $E(\boldsymbol{w}(t)) = E(\boldsymbol{v}(t)) = 0$。已知 $\boldsymbol{u}(t)$、$\boldsymbol{y}(t)$、$\boldsymbol{m}(t)$ 和协方差矩阵 $\boldsymbol{\Sigma}$，则可利用多重线性回归的方法分别求出矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$、$\boldsymbol{D}$、$\boldsymbol{H}$ 的估计值，

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \Sigma\left[\begin{pmatrix} \boldsymbol{m}(t+1) \\ \boldsymbol{y}(t) \end{pmatrix}, \begin{pmatrix} \boldsymbol{m}(t) \\ \boldsymbol{u}(t) \end{pmatrix}\right]\Sigma^{-1}\left[\begin{pmatrix} \boldsymbol{m}(t) \\ \boldsymbol{u}(t) \end{pmatrix}, \begin{pmatrix} \boldsymbol{m}(t) \\ \boldsymbol{u}(t) \end{pmatrix}\right] \tag{4.36}$$

多重回归协方差矩阵为

$$\begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix} = \Sigma\left[\begin{pmatrix} \boldsymbol{m}(t+1) \\ \boldsymbol{y}(t) \end{pmatrix}, \begin{pmatrix} \boldsymbol{m}(t+1) \\ \boldsymbol{u}(t) \end{pmatrix}\right] - \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}\Sigma^T\left[\begin{pmatrix} \boldsymbol{m}(t+1) \\ \boldsymbol{y}(t) \end{pmatrix}, \begin{pmatrix} \boldsymbol{m}(t) \\ \boldsymbol{u}(t) \end{pmatrix}\right] \tag{4.37}$$

则 $\boldsymbol{H} = \boldsymbol{S}_{21}\boldsymbol{S}_{11}^+, \boldsymbol{Q} = \boldsymbol{S}_{11}, \boldsymbol{R} = \boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^+\boldsymbol{S}_{12}$ （“+”为伪逆矩阵）。具体推导过程可参考文献[48]，在此不作详细描述。求出各系数矩阵之后，可进一步求出 $\boldsymbol{x}(t+1)$ 和 $\boldsymbol{y}(t)$ 的估计值 $\hat{\boldsymbol{x}}(t+1)$ 和 $\hat{\boldsymbol{y}}(t)$。状态变量 $\boldsymbol{x}(t+1)$ 和输出 $\boldsymbol{y}(t)$ 的估计值 $\hat{\boldsymbol{x}}(t+1)$ 和 $\hat{\boldsymbol{y}}(t)$ 与其观测值之间的误差分别为

$$\begin{aligned} \boldsymbol{w}(t) &= \boldsymbol{x}(t+1) - \hat{\boldsymbol{x}}(t+1) \\ \boldsymbol{v}(t) &= \boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t) \end{aligned} \tag{4.38}$$

此外，需要注意的是：在 CVA 模型中，过去数据集长度 p 值不能小于变量个数，且未来数据集长度 f 一般需大于过去数据集长度 p（即 $f > p$），这就要求 p 和

f 不能太小。而 CVA 模型中存在时间滞后，f 为模型滞后时间周期数，将导致模型判别结果滞后。在道路交通偶发性拥堵检测的过程中，希望能及时检测出偶发性拥堵，滞后时间周期 f 值越小越好，理想状态下当偶发性拥堵发生时即可马上检测出来（即 $f=0$）。因此，f 值大小的确定在检测精度和效率之间存在矛盾：一方面，在一定范围内 f 值越大，包含的信息越多，则检测精度越高，而检测效率会下降；另一方面，f 值越小，则模型的检测效率越高。因此，合理确定 p 和 f 的值是模型检测精度和效率之间的折中。

4.5 道路交通偶发性拥堵的检测

4.5.1 参数选择

根据 4.1 节的参数分析可以得到实时情况下，可以得到 $v_1(i,t)$、$v_2(i,t)$、$v_3(i,t)$、$v_4(t)$ 这 4 个的实时速度变量，按式（4.4）计算这 4 个变量在 t 时刻对应的路段延误时间指数，分别用 y_1, y_2, y_3, y_4 表示，用以共同描述道路的交通状态。令 $\boldsymbol{y}^T=[y_1, y_2, y_3, y_4]$，令 $\boldsymbol{y}_p(t)=[\boldsymbol{y}^T(t-1), \boldsymbol{y}^T(t-2), \cdots, \boldsymbol{y}^T(t-p)]^T$ 和 $\boldsymbol{y}_f(t)=[\boldsymbol{y}^T(t), \boldsymbol{y}^T(t+1), \cdots, \boldsymbol{y}^T(t+f-1)]^T$ 分别为 4 个变量在过去 p 个周期和未来 f 个周期内的样本矩阵。则有

$$\boldsymbol{y}_p(t)=\begin{bmatrix} y_1(t-1) & y_2(t-1) & y_3(t-1) & y_4(t-1) \\ y_1(t-2) & y_2(t-2) & y_3(t-2) & y_4(t-2) \\ \cdots & \cdots & & \\ y_1(t-p) & y_2(t-p) & y_3(t-p) & y_4(t-p) \end{bmatrix} \tag{4.39}$$

$$\boldsymbol{y}_f(t)=\begin{bmatrix} y_1(t) & y_2(t) & y_3(t) & y_4(t) \\ y_1(t+1) & y_2(t+1) & y_3(t+1) & y_4(t+1) \\ \cdots & \cdots & & \\ y_1(t+f-1) & y_2(t+f-1) & y_3(t+f-1) & y_4(t+f-1) \end{bmatrix} \tag{4.40}$$

利用 CVA 算法对 $\boldsymbol{y}_p(t)$ 和 $\boldsymbol{y}_f(t)$ 进行计算处理，可以得到 t 时刻的预测值，将 t 时刻的变化趋势与实际的状态值进行比较，即可得到道路交通状态的实时变化趋势。

4.5.2 判别指标选择

选择合适的判别指标实际上就是选择合适的参数来表征式（2.1）中的 $SPE_R(r,k)$ 和 $SPE_H(r,k)$ 值。

在过程工业系统中对故障的检测常用的判别指标有 T^2 统计量与 Q 统计量、I^2 统计量等[49]，这些统计量都是通过与预定值进行比较，当其超出预定范围，即可判断有异常事件发生。道路交通系统偶发性拥堵与过程工业的故障都是与正常拥堵进行相比较，因此可以借用这些统计量作为判别指标。

根据以上 CVA 模型的输出 $\hat{y}(t)$，本文采用 SPE 统计量（Squared Prediction Error，即平方预测误差）来表征模型的精度。其中 SPE 统计量的计算方式如下：

$$SPE(i) = \sum_{j=1}^{\omega} (y_{ij} - \hat{y}_{ij})^2 \qquad (4.41)$$

其中 $SPE(i)$ 为 t 时刻的平方预测误差，y_{ij} 为第 j 个规范变量在第 i 个周期的观测值，$\hat{y}_{ij}$ 第 j 个规范变量在第 i 个周期的估计值，ω 为最终输出的变量个数且有 $\omega <= n$，$y_{ij} - \hat{y}_{ij}$ 为第 j 个规范变量在第 i 个周期的预测误差。在过程工业故障诊断过程中，若 $SPE(i)$ 大于预定的阈值则可判定发生了异常。

4.5.3 阈值的确定

阈值的确定实际上就是确定式（2.1）中的 $SPE_H(r,k)$ 值。由 3.4 节已经利用基于 T 检验的 K-均值自适应聚类算法实现了对不同情景的划分，确定每种初始情景（即路段 r 在第 k 个时段）及其对应的类（以 i 表示），则情景（r,k）的阈值就应该 $SPE_H(r,k)$ 值和对应的第 i 类情景的阈值一样，即有

$$SPE_H(r,k) = SPE_H(i) \qquad (4.42)$$

其中 r 和 k 分别为路段和时段编号，而 i 为对应的情景的分类（i=1,2,…,8）.

在过程工业控制系统中以 SPE 统计量作为指标进行故障诊断，其阈值可以采用核函数的方法计算得到。在本文中，为了充分发挥海量的历史数据的优势，利用 3.5 节中四分位数的方法，把历史数据中的正常拥堵和偶发性拥堵进行划分，从海量的历史数据中计算出历史条件下的每一类交通拥挤程度下正常的交通拥堵下 SPE（i）范围，从而确定交通正常拥堵与偶发性拥堵之间的临界值。

利用海量历史数据训练不同交通拥挤程度下正常交通拥堵与异常交通拥堵的临界值，其分布如图 4.2 所示，其中超出阈值的是偶发性拥堵，而小于阈值的是正常拥堵。

由图 4.2 可以看出，对于较为畅通的情况下，不同交通情景下的 SPE（i）控制上限差异不大，对交通偶发性拥堵的判别是“由不堵到堵”的偶发性拥堵的判别，阈值范围较小。而对于拥堵的情况下，对交通偶发性拥堵判别是“由堵到不堵”或者“由堵到更堵”的偶发性拥堵的判别，因为交通拥堵差异比较大，阈值范围也更大。

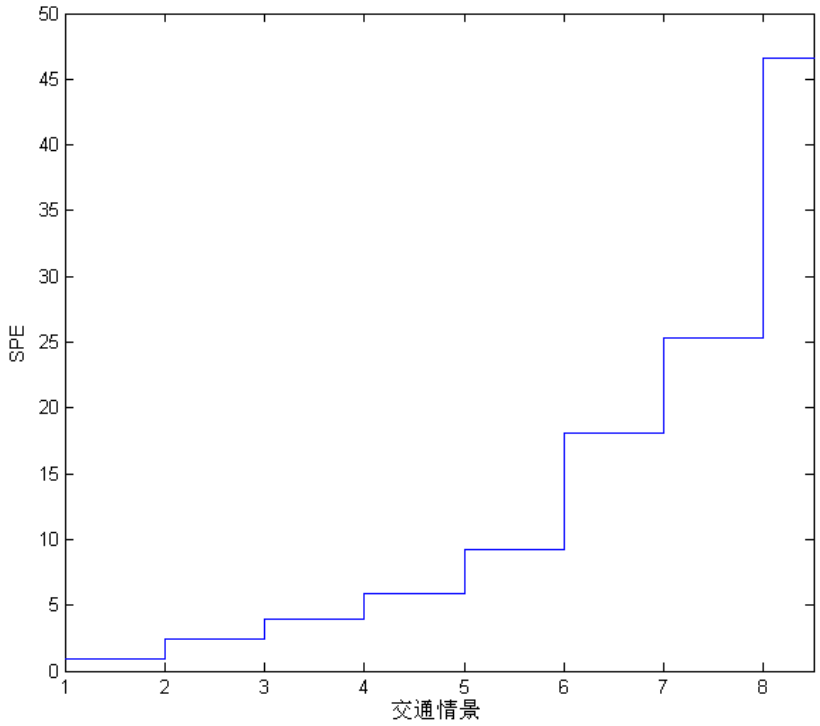


图 4.2 不同情景下 SPE 上限

Figure 4.2 Upper limit line of SPE in each traffic condition

4.5.4 道路交通偶发性拥堵的最终判定

对道路交通偶发性拥堵的判别不能像工业系统故障检测那样，当检测到一个或者连续两个周期的 SPE 值超过阈值即可判定为偶发性拥堵。车辆在路段上行驶，受实际路况、车辆、驾驶员等多种因素的影响可能会出现突然减速或者暂时停车的情况，此时计算出来的 SPE 值可能会突然增大，但并不能说明是出现了偶发性拥堵。此外，对于拥堵的交通模式下偶发性拥堵的判别，本来在拥堵状态下车辆已经处于停止或者速度很小的情况，此时 SPE 值可能也比较大，但也并不能马上判别是正常拥堵或者是异常拥堵。正常状态和异常状态的差异在于：正常情况下状态时间比较短，当拥堵超过一定的时间长度，即可视为异常拥堵。因此，为提高道路交通偶发性拥堵的检测率，同时降低误判率，需要对判别条件作进一步的细化。

在此，对判别条件进行改进，利用当前时刻的前 p 个周期的预测平方误差 $SPE(i-p+1), SPE(i-p+2), \ldots, SPE(i)$ 作为判别指标。只有当连续 p 个周期的 SPE 值均超过阈值，才能判断有偶发性拥堵发生。

4.6 本章小结

本章首先针对单一的瞬时速度参数不完全可靠的情况，抽取了 4 个速度参数

用以共同表征道路交通实时状态。针对公交车实时数据具有显著的自相关的特点，本文选择 CVA 算法建立道路交通偶发性拥堵检测模型。因此，本章首先对实时参数分析进行分析，确定实时交通状态的表征方法；其次，介绍了 CVA 算法的基本原理，然后建立基于 CVA 的道路交通偶发性拥堵检测模型，以实现对道路交通拥堵的实时变化趋势的分析。同时，利用历史的偶发性拥堵信息训练出不同情景下的阈值。将道路交通状态的实时变化趋势与阈值相比较，可以实现道路交通偶发性拥堵的判断。

5 道路交通偶发性拥堵检测系统实现与应用

5.1 本章引言

在第 3 章已完成了利用海量公交车 GPS 历史数据对交通历史状态规律分析，确定了每类情景下道路交通状态的特征；而第 4 章分别利用 CVA 建立了模型以检测道路交通偶发性拥堵。为了验证道路交通偶发性拥堵的检测模型的有效性，需要实现第 2 章中的基于公交车 GPS 数据的道路交通偶发性拥堵检测系统，并利用实际的公交车 GPS 数据实现对该系统的应用效果进行测试。

因此，本章的内容主要包括：第一，基于公交车 GPS 数据的道路交通偶发性拥堵检测系统的实现；第二，道路交通偶发性拥堵检测的评价指标体系介绍；第三，对道路交通偶发性拥堵检测结果的分析和总结。

5.2 系统的实现

本文在得到公交车 GPS 数据的基础上，将公交车 GPS 数据划分为历史数据和实时数据两部分，其中历史数据用于统计分析道路交通状态的历史情况（道路交通历史状态统计分析、交通状态模式识别、偶发性拥堵的量化定义和阈值的训练等），而实时数据用于分析道路交通拥堵的实时变化趋势。其中对于历史数据的处理主要以 SQL Server 2008 数据库和 Visual Studio 2008 等为主要实现平台；而对于实时数据的处理以及偶发性拥堵的实时检测则同样是以 SQL Server 2008 和 Visual Studio 2008 为主要平台。其中，SQL Server 2008 数据库主要用于存储和管理海量的公交车 GPS 数据，同时建立专门的数据仓库实现对公交车 GPS 数据以及其他必要信息（如车辆信息、站点信息等）的集成和规范化处理等；Visual Studio 2008 用于编程实现对公交车 GPS 数据预处理，如 GPS 数据修正，计算历史参数和实时参数等，并且实现基于 T-检验的 K-均值自适应聚类算法以实现道路交通模式识别等；最终在 Visual Studio 2008 平台上实现基于 CVA 算法的道路交通偶发性拥堵的检测。

在设计过程中，使用 UMI（统一建模语言）的方法进行系统软件设计，在实现过程中，以 C#作为编程语言，采用 OOP（面向对象的编程）方法进行编程实现，并且采用 OLE DB 接口对数据库进行操作。由于本文篇幅有限，具体的设计和实现过程内容繁多，在此不一一介绍，而仅简单介绍部分主要系统界面及其作用。

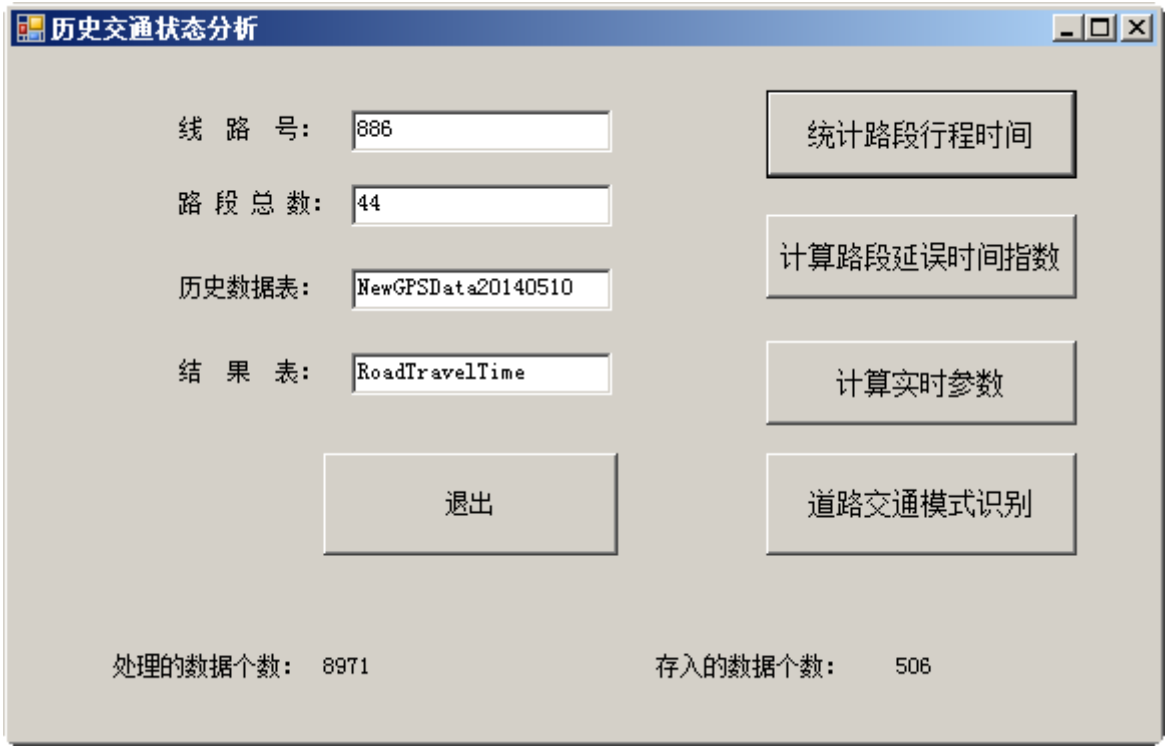


图 5.1 公交 GPS 数据预处理及历史状态统计分析界面

Figure 5.1 Bus GPS data preprocess interface

图 5.2 城市道路偶发性拥堵检测系统主界面

Figure 5.2 The main interface of urban road contingency jam detection system

图 5.3 道路交通偶发性拥堵检测报警

Figure 5.3 Road traffic contingency jam alarm

如以上各图所示，基于海量公交车 GPS 数据的道路交通偶发性拥堵检测系统的实现主要包括利用 Visual Studio 2008 数据的预处理和历史状态的统计分析（路段行程时间的统计，路段延误时间指数的计算，实时参数的计算和实现基于 T 检验的 K-均值自适应算法以实现道路交通模式识别）；对于 CVA 算法的主要计算过程，在 Visual Studio 2008 平台实现，根据 CVA 算法可得到各周期对应的预测值，将预测值与实际值相比较即可得到最终的输出结果，将输出来的结果存储于数据库中。当出现异常时，则跳出报警框，提示有异常发生。最后用 MATLAB 2014a 将输出结果展示出来，具体过程在此不一一介绍。

5.3 评价指标体系的建立

在实现道路交通偶发性拥堵检测系统的基础上，应用该系统实现对道路交通偶发性拥堵检测，并对其应用效果作进一步的统计分析和评价。

目前国内外在道路交通拥堵检测方面所采用的评价指标主要用于评价正常拥堵检测的效果，不适用于作为偶发性拥堵检测的指标。实际上，城市道路偶发性拥堵的发生与异常事件有关，异常的交通拥堵由道路上的异常事件引起，而在道路交通异常事件检测方面一般采用检测率（Detection Rate，DR）、误判率（False Alarm Rate，FAR）以及平均检测时间（Mean Time To Detection，MTTD）等多个

指标对异常事件检测的算法和模型进行评价[50]。因此，在此借用道路交通异常事件的评价方法，以检测率、误判率和平均检测时间为指标，对本文中的道路交通偶发性拥堵判别算法进行评价。三个指标的具体计算方式如下：

$$DR = \frac{TP}{S} * 100\% \tag{5.1}$$

$$FAR = \frac{FN}{DT} * 100\% \tag{5.2}$$

$$MTTD = \frac{1}{TP} \sum_{i=1}^{TP} [TI(i) - AT(i)] \tag{5.3}$$

其中 TP 为准确检测到的偶发性拥堵次数，S 为实际发生的偶发性拥堵次数（包括检测到的和未检测到的），FN 为错误检测（即不是偶发性拥堵却误判为偶发性拥堵）的次数，DT 为检测偶发性拥堵次数，$TI(i)$ 为被算法检测到真正发生的第 i 次偶发性拥堵发生的时间，$AT(i)$ 为真正发生的第 i 次偶发性拥堵实际发生的时间。

理想情况下，有 $DR = 100\%, FAR = 0\%, MTTD = 0$，即所有的偶发性拥堵均能被检测出来，而且没有误检的情况，且偶发性拥堵一发生就能马上被检测出来。实际情况下，由于检测算法的滞后延迟，数据传输、处理和发布也需要一定的时间，目前在道路交通事件检测方面，DR 能达到 $80\% \sim 100\%$，在 DR 超过 80%时，FAR 在 $70\% \sim 30\%$ 之间，而 MTTD 则因检测周期的长短而可能有较大的差异。

5.4 检测结果分析

本文以重庆市 886 线路下行方向的 2-19 号路段共 18 个路段作为研究对象，利用 2014 年 5 月 5 号到 6 月 1 号连续四周的数据对基于 CVA 的道路交通偶发性拥堵检测模型进行测试。

5.4.1 CVA 规范变量系数及其显著性分析

规范变量是经过 CVA 规范化处理后得到的新的测量指标，最终得到的同一组规范向量之间互不相关，而描述过去信息和未来信息的同一对规范变量 (u_i, v_i) 之间的相关系数 λ_i，且不同对规范变量 u_i 和 v_j 之间互不相关（其中 $i, j = 1, 2, \ldots, m$，m 为规范变量个数），λ_i 表明对应的规范变量之间的密切程度，λ_i 越大则其关系越密切，且对模型的解释能力也越强。因此在实际中，只考虑其中相关性较强的规范变量，而忽略相关性较弱的相关变量。经分析，在不同交通模式下的规范变量相关系数差异较大，各种交通情景下的 CVA 规范化分析结果如表 5.1 所示。

表 5.1 各情景下道路交通状态规范变量分析举例

Table 5.1 Example of canonical correlations analysis of each traffic condition

情景	规范变量对数	特征根	累积百分比	卡方值	自由度	显著性水平
情景 1	1	0.687	36.86%	54.241	16	0.000
	2	0.506	64.00%	25.838	9	0.002
	3	0.463	88.84%	12.678	4	0.013
	4	0.208	100.00%	1.961	1	0.161
情景 2	1	0.569	32.46%	92.546	16	0.000
	2	0.529	62.64%	55.588	9	0.000
	3	0.436	87.51%	24.610	4	0.000
	4	0.219	100.00%	4.663	1	0.031
情景 3	1	0.985	46.59%	2194.325	16	0.000
	2	0.731	81.17%	446.188	9	0.000
	3	0.356	98.01%	67.903	4	0.000
	4	0.042	100.00%	0.894	1	0.344
情景 4	1	0.991	48.58%	2315.562	16	0.000
	2	0.673	81.57%	345.312	9	0.000
	3	0.287	95.64%	46.571	4	0.000
	4	0.089	100.00%	3.955	1	0.047
情景 5	1	0.784	50.94%	72.665	16	0.000
	2	0.516	84.47%	19.814	9	0.019
	3	0.214	98.38%	2.626	4	0.622
	4	0.025	100.00%	0.035	1	0.851
情景 6	1	0.888	64.21%	95.044	16	0.000
	2	0.369	90.89%	8.792	9	0.457
	3	0.108	98.70%	0.669	4	0.955
	4	0.018	100.00%	0.018	1	0.893
情景 7	1	0.829	42.62%	95.044	16	0.000
	2	0.513	69.00%	8.792	9	0.457
	3	0.419	90.54%	0.669	4	0.955
	4	0.184	100.00%	0.018	1	0.893
情景 8	1	0.879	48.94%	397.039	16	0.000
	2	0.518	77.78%	91.314	9	0.000
	3	0.349	97.22%	26.857	4	0.000
	4	0.05	100.00%	0.000	1	1.000

在显著性水平取 $\alpha = 0.05$ 的情况下，不同情景所选择的规范变量的对数不一样。具体看来，在选择对模型的解释能力达到 85%以上的规范变量对数，从情景 1 到情景 8 依次可以选择 3、3、3、3、3、2、3、3 对，通过规范化处理之后的变量对模型的解释能力依次为 88.84%、87.51%、98.01%、95.64%、98.38%、90.89%、90.54%和 97.22%。从情境 1 到情景 8 依次 3、4、3、4、2、1、1 和 3 对规范变量之间的相关系数是显著的（显著性水平<0.05）。不同的情景所应选择的规范变量对数不仅相同，而且对模型的解释能力也不同。因此，在利用 CVA 进行道路交通偶发性拥堵检测时，考虑不同情景交通状态之间的差异是必要的，对道路交通偶发性拥堵的检测更有利。

5.4.2 道路交通状态实时变化趋势分析

为了分析道路交通状态的实时变化趋势，在此以 886 线路下行的 16-19 号路段为例，利用 CVA 算法对这四条路段的公交车 GPS 数据进行计算，最终得到 4 条路段的 SPE 分布情况，其中在相同一段时间（连续 500 个周期）内 4 条路段的 SPE 分布情况如图 5.4 所示。

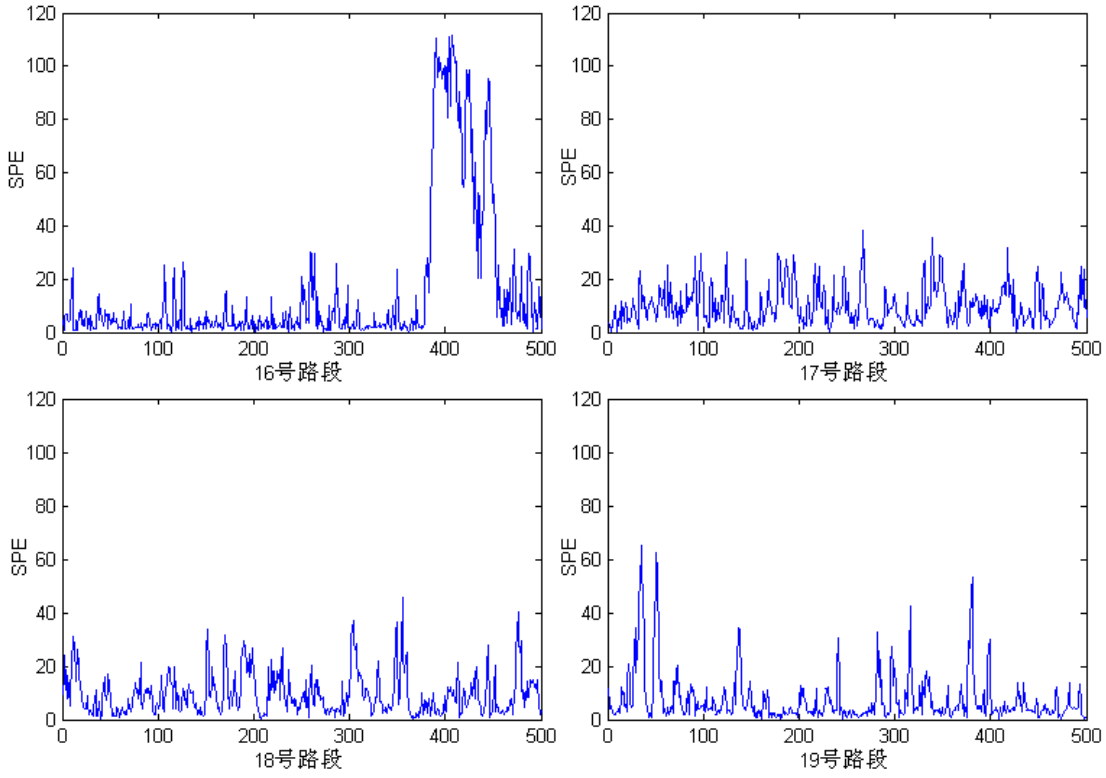


图 5.4 886 线路 16-19 号路段在相同时期内的 SPE 分布曲线

Figure 5.4 SPE distribution of No.16-No.19 road section of 886 bus line in the same cycles

由图 5.4 可以看出，同一时间段内不同路段上交通拥堵的拥挤程度不一样，SPE

越大，说明拥挤越严重；而且，在某一时刻的 SPE 值可能会突然增大，但又会迅速下降，这是车辆在行驶过程中受实时路况的影响突然出现的减速和提速的过程，此时 SPE 虽然大，但不能说明出现了拥堵。因此，对交通偶发性拥堵的判别不能仅凭单周期内 SPE 波动的大小来确定，只有当连续多个周期内 SPE 均超过阈值，才能确定有偶发性拥堵发生。

5.4.3 道路交通偶发性拥堵检测结果分析

2014 年 5 月 5 日到 2014 年 6 月 1 日连续 4 周期间的有效 GPS 数据共达 514339 条，利用 3.4 节的方法确定期间实际发生的道路交通偶发性拥堵共有 143 次。以这些公交车 GPS 数据为基础，对基于 CVA 算法道路交通偶发性拥堵检测模型进行测试，在不同未来数据集长度 f 和过去数据集长度 p 的取值下，得到的检测结果。

表 5.2 基于 CVA 的道路交通偶发性拥堵检测结果

Table 5.2 Road traffic contingency jam detection results base on CVA

f 值	p 值	检测偶发性拥堵次数	检测的实际偶发性拥堵次数	DR	FAR	MTTD（min）
6	5	204	128	89.51%	37.25%	2.44
7	6	196	126	88.11%	35.71%	2.72
8	6	209	129	90.21%	38.28%	2.88
8	7	195	124	86.71%	36.41%	3.04
9	6	196	127	88.81%	35.20%	3.02
9	7	198	130	90.91%	34.34%	3.21
9	8	186	124	86.71%	33.33%	3.33
10	6	197	128	89.51%	35.03%	3.20
10	7	192	126	88.11%	34.38%	3.35
10	8	183	125	87.41%	31.69%	3.51
10	9	181	126	88.11%	30.39%	3.67

由表 5.2 可以看出，利用基于 CVA 算法对道路交通偶发性拥堵，在不同数据集长度下，得到的交通偶发性拥堵检测效果不同，随着数据集长度的增大，误检率（FAR）逐渐降低，但平均检测时间会逐步增大（MTTD），而检测率（DR）变化不大。事实上，当 f 值大于 18（即滞后周期超过三分钟）时，随着 f 值的增大，检测率会下降，而误检率上升，平均检测时间在 5min 以上，检测效果明显变差。因此，f 值和 p 值不能太大。总体上看来，利用 CVA 算法可以有效实现道路交通偶发性拥堵的检测误报率在 35%以下，平均检测时间约为 3 分钟，检测精度达 90%，达到较好的检测结果。

5.4.4 检测结果对比

CVA 方法考虑道路实时参数的自相关的特性，而传统的数据驱动方法却没有考虑到这一点，为了说明 CVA 算法在道路交通偶发性拥堵检测中的效果，以下采用传统的 PCA（Principal Component Analysis，主元分析）算法建立模型分析道路交通的实时变化趋势，以检测道路交通偶发性拥堵。

① SPE分布图对比

将基于CVA算法的道路交通偶发性拥堵检测模型和基于PCA算法道路交通偶发性拥堵检测模型的SPE分布进行对比，如图5.5所示。

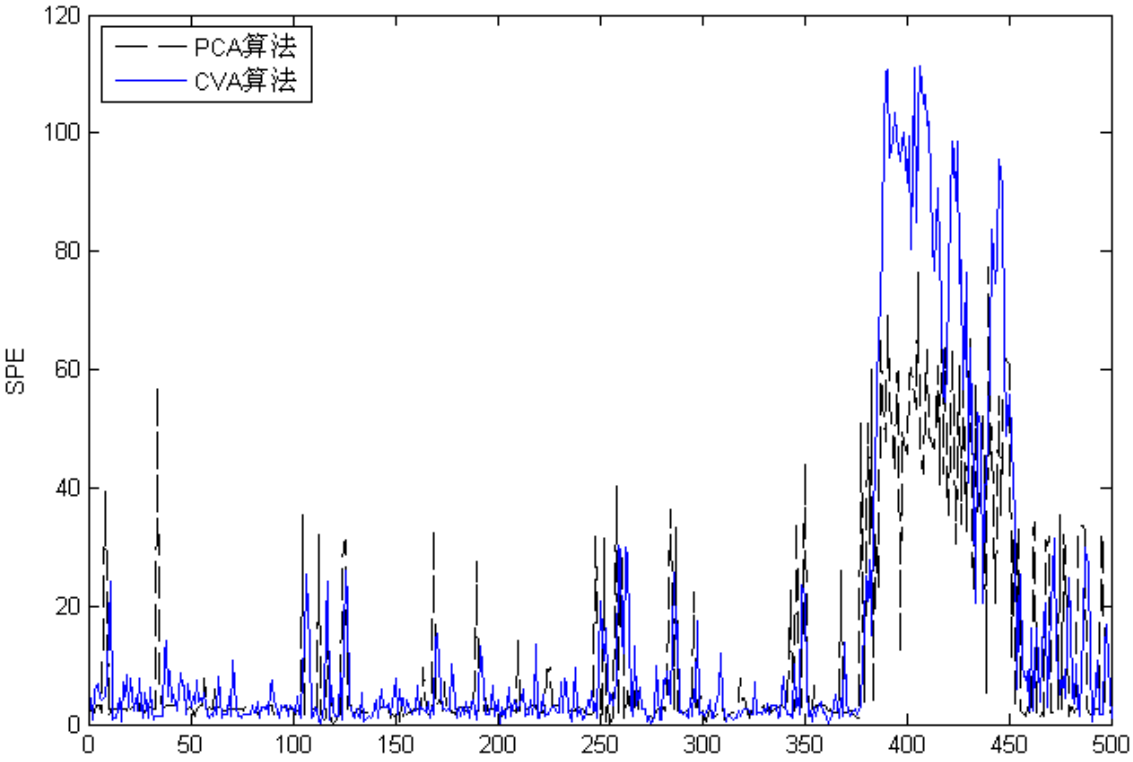


图 5.5 CVA 和 PCA 算法的 SPE 分布曲线分布比较

Figure 5.5 The difference of SPE distribution between CVA and PCA

由图5.5可以看出，基于CVA的道路交通偶发性拥堵检测分布图与基于PCA的道路交通偶发性拥堵检测分布图的分布大体类似，但也有以下三点明显的差异，具体是：其一，整体上CVA的SPE略大与PCA的SPE；其二，正常拥堵下PCA的突变程度较大，而CVA算法的SPE变化相对较为平缓；其三，CVA较PCA有2到3个周期的滞后。

② 检测精度对比

利用PCA检测道路交通偶发性拥堵，同样需要当连续p个周期的SPE值均超出

阈值时，才能确定有偶发性拥堵发生。同样，当*p*取不同值时，得到的检测结果也不同。具体结果如表5.5所示。

表 5.5 基于 PCA 的道路交通偶发性拥堵检测结果

Table 5.2 Road traffic contingency jam detection results base on PCA

p 值	检测偶发性拥堵次数	检测的实际偶发性拥堵次数	DR	FAR	MTTD（min）
1	532	135	94.41%	74.62%	0.88
2	417	133	93.01%	68.11%	1.40
3	321	132	92.31%	58.88%	1.54
4	272	126	88.11%	53.68%	1.70
5	218	124	86.71%	43.12%	1.97
6	192	121	84.62%	36.98%	2.15
7	182	120	83.92%	34.07%	2.30
8	176	120	83.92%	31.82%	2.56
9	169	119	83.22%	29.59%	2.78
10	162	117	81.82%	27.78%	3.01
11	160	117	81.82%	26.88%	3.16
12	156	113	79.02%	27.56%	3.35

同样，利用PCA的方法进行道路交通偶发性拥堵的检测，随着*p*值的增大，检测率下降，当*p*值达到12个周期（即两分钟）时，检测率低于80%。同时，误检率也大大下降，而平均检测时间随之增大。可以看出，在相同*p*值下，CVA的检测率高于PCA的检测率，CVA的误检率低于PCA的误检率，CVA比PCA具有更高的精度，但CVA的平均检测时间大于PCA。相对而言，在一定的范围内，CVA的检测率比较稳定，其检测结果受滞后周期长度的影响较小，而PCA的检测结果则受滞后周期的影响较大。当误报率相近时，CVA的检测率比PCA高于约5个百分点，而平均检测时间则比PCA滞后约半分钟。整体上看，在相同误报率的约束下，CVA算法得到的检测效果高于PCA的检测效果，说明了CVA算法更具优越性。

因此，从整体上来看，本文在海量公交车GPS数据的基础上，分析道路交通状态的规律和实时变化趋势，结合道路交通状态的历史情况和实时变化趋势可以实现对道路交通偶发性拥堵的检测，并且能取得较好的结果。

5.5 本章小结

本章首先设计和实现了基于公交车GPS数据的城市道路偶发性拥堵检测系统；

接着对道路交通偶发性拥堵检测的评价指标体系进行介绍，然后利用实验实现了基于CVA算法的道路交通偶发性拥堵检测模型，实验结果表明：利用CVA算法建立模型用于道路交通偶发性拥堵的检测可以达到较好的结果。

6 总结与展望

6.1 总结

城市道路交通偶发性拥堵随机性大，具有相对性、情境性和集群性等特点，目前的算法难以达到满意的结果。近年来，在智能交通领域积累了海量的公交车GPS数据，这些数据记录了公交运行的轨迹，反映了道路交通状态的历史情况。公交车GPS数据具有数据量大、覆盖面广、实时性高、可靠性高和维护成本低等特点，研究利用公交车GPS数据进行道路交通偶发性拥堵检测有助于提高检测精度。如何从海量的公交车GPS数据中提取出有用的知识为道路交通偶发性拥堵检测乃至为城市道路交通管理和服务提供支持和参考，是智能交通领域面临的重要课题。

因此，本文以海量的公交车GPS数据为基础研究道路交通偶发性拥堵的检测技术，并进行相应的软件系统的开发和实现，主要做了以下三点工作：

① 建立了基于公交车GPS数据的道路交通偶发性拥堵检测模型和系统体系结构，将本文的主要工作划分为若干模块，明确各模块的主要内容和各模块之间相互关系；

② 以海量的公交车GPS数据为基础，研究了道路交通状态规律的分析技术。首先定义“路段延误时间指数”用于表征道路交通状态，其次针对交通偶发性拥堵的相对性和情景性的特定，提出基于T-检验的K-均值自适应聚类算法将道路交通情景划分为8类，并确定每类情景下的交通状态特征，实现了道路交通状态的模式识别；在此基础上，引入统计学中“四分位差”的方法明确了道路交通拥堵正常范围与异常范围的量化区分；

③ 研究了道路交通状态实时变化趋势分析与偶发性拥堵检测的技术。首次引入CVA算法用于检测道路交通偶发性拥堵；以SPE为指标，分析道路交通状态的实时变化趋势，同时利用历史的偶发性拥堵数据训练道路交通正常拥堵与偶发性拥堵之间的阈值。针对道路交通偶发性拥堵的集群性的特点，通过对比连续多个周期内道路交通状态实时变化趋势的SPE与阈值的差异以实现对道路交通偶发性拥堵的检测。

最后，本文设计并实现了基于公交车GPS数据的城市道路偶发性拥堵检测系统，并利用实际的公交车GPS数据进行测试。测试结果表明：该系统能够能实现对道路交通状态实时变化趋势的分析，从而可以实现道路交通偶发性拥堵的检测。整体上，在误报率低于35%、平均延误时间小于3.2分钟时，CVA算法的检测率接近90%，而PCA算法的检测率约为84%，说明CVA算法较PCA算法具有更高的检测精度，且CVA算法检测效果更加稳定。

综上所述，以海量公交车GPS历史数据为基础，结合公交车GPS数据的实时变化趋势，可以实现道路交通偶发性拥堵的检测，而且能取得较为满意的效果。

6.2 展望

道路交通偶发性拥堵检测可为道路交通状态监控、交通诱导、交通应急处理、公交到站时间预测等提供支持和参考，有助于降低交通出行者的出行时间和成本，提高道路交通管理水平和服务水平。本文研究了基于公交车GPS数据的道路偶发性拥堵检测，达到了较高的检测效果，但也不足一些不足之处，主要包括：

① 要发挥道路偶发性拥堵检测效果的作用，需要进一步分析道路交通偶发性拥堵的影响程度，由于道路交通偶发性拥堵的内部机理不清楚，也没有明显的规律可循，如何分析其影响程度将是后续工作进一步研究的内容；

② 城市道路交通偶发性拥堵的检测以大量的GPS数据为基础，而且在实时数据检测过程中CVA算法需要进行大量的矩阵运算，随着系统应用面的扩大，系统在短时内需要处理的数据量也急剧增加，导致系统处理时间较长，内存消耗大，如何提高系统的数据处理效率，也是一个亟待解决的问题；

③ 交通偶发性拥堵的检测效果还有进一步提升的空间，能得到近90%的检测率，但误报率还在30%左右，造成漏报和误报的有算法的因素、数据质量因素等多方面，如果针对这些因素作优化从而进一步提高检测效率，还有待进一步的研究。

致　　谢

三年的研究生生活即将成为过去式，回首过去三年，感触良多。研究学术，对一个问题，钻得越深，发现涉及的范围越广泛，而未知也越多，感觉自己越无知。如陶渊明《桃花源记》里的渔民，经过一个狭窄的洞口之后，豁然开朗走到另一个世界，但这个世界存在着太多的未知。因此，三年研究生的生活中，不敢稍有懈怠。

“研途”走来，遇到不少问题和困难，所幸的是在我最困难的时候总有人能够伸手扶我一把，帮我顺利度过困境。对他们，我满怀感激。

首先，要感谢我的家人，没有他们的支持陪伴，我无法顺利走完三年的研究生生活。《诗经》里说：“孝子不匮，永锡尔类”，这是我最喜欢的一句诗。家人一直都是我奋斗的动力之源，虽然他们没有在具体学术上的问题上给过我帮助，但他们的陪伴和支持就是最重要也是最伟大的帮助。这份恩情，此生难以回报。

其次，感谢我的导师廖孝勇老师，廖老师在平时给我了不少指导，非常感谢他的栽培。还有，要感谢孙棣华教授，孙教授学识渊博，他治学严谨的风格一直令人敬佩叹服。从小学到现在，我是幸运的。因为在多年的求学生涯中往往都能遇到关心我的老师，如我高二时候的班主任张小红老师，他们对我的关心和教诲今生没齿难忘。《礼记·学记》里说：“三王四代唯其师”，老师对学生的教诲往往影响学生的一生。正是他们的关心和帮助才让我走到今天。非常感谢他们。

其次，还有实验室的其他同学和朋友，特别是我研究生三年的好朋友肖军和何伟博士。在这过去的三年里，他们不仅在学业上给了我帮助，同时在生活方面也给我了很大的支持和鼓励。非常感谢他们。

正是认识了这么多的良师益友，正是他们的帮助和指导，我现在才能顺利完成三年的研究生学业。《诗经》里说：“嘤其鸣矣，求其友声”。在以后的道路上，将会结识更多的朋友，但老朋友的恩情将不会忘记。

还有，感谢所有曾经帮助过我、让我成长的那些人。

所谓“来而不往，非礼也”。从小到大，一直都在接受他人的帮助，希望在毕业以后，走在职场上，将读书期间的知识学以致用，能够以自己微小的力量，为社会的发展做出自己的一份贡献。以自己的所学用为国家社稷谋福祉，我想，这应该是对给过我帮助的所有人的最好的回报方式，也是我历来的心愿所在。

崔德冠

二〇一五年四月 于重庆

参考文献

[1] 王笑京, 沈鸿飞, 汪林. 中国智能交通系统发展战略研究[J]. 交通运输系统工程与信息, 2006, 6(4): 9-12.

[2] Ezell S. Explaining international IT application leadership: Intelligent transportation systems[J]. 2010.

[3] Qi L. Research on intelligent transportation system technologies and applications[C].Power Electronics and Intelligent Transportation System, 2008. PEITS'08. Workshop on. IEEE, 2008: 529-531.

[4] 史其信, 陆化普. 智能交通系统的关键技术及研究发展策略[J]. 中国土木工程学会第八届年会论文集, 1998.

[5] 张飞舟. 公交车辆智能调度研究[J]. 交通运输系统工程与信息, 2001, 1(1): 73-80.

[6] 潘峰. 基于 GIS 的道路交通事故预测系统研究[D]. 电子科技大学, 2010.

[7] 戢晓峰. 城市道路交通状态分析方法回顾与展望[J]. 道路交通与安全, 2008, 8(3): 11-15.

[8] 邵敏华. 网络交通评价方法, 指标体系及影响因素研究 [D]. 上海: 同济大学, 2006.

[9] 1995 G A. 道路交通阻塞度及评价方法 [S].

[10] 郭钰愫,邵春福.道路交通异常事件及其交通组织研究[C].2005 年海峡两岸智能交通运输系统学术研讨会暨第二届同舟交通论坛智能交通运输系统研究与实践.2005:402-410.

[11] Jiawei Han M K. Data mining: concepts and techniques[J]. ISBN, 2006, 10: 1-55860.

[12] 庞昊. 基于 FCD 的城市交通动态诱导系统关键技术的研究[D]. 中国科学技术大学, 2009.

[13] Ozbay K, Kachroo P. Incident management in intelligent transportation systems[J]. 1999.

[14] Srinivasan D, Jin X, Cheu R L. Evaluation of adaptive neural network models for freeway incident detection[J]. Intelligent Transportation Systems, IEEE Transactions on, 2004, 5(1): 1-11.

[15] 唐金芝. 数据融合技术在高速公路交通事件检测中的应用 [D]. 长春: 吉林大学, 2007.

[16] 孙棣华, 董均宇, 廖孝勇. 基于 GPS 探测车的道路交通状态估计技术 [J]. 计算机应用研究, 2007, 24(2): 243-245.

[17] Boyce D E, Kirson A, Schofer J L. Design and implementation of ADVANCE: The Illinois dynamic navigation and route guidance demonstration program[C]//Vehicle Navigation and Information Systems Conference, 1991. IEEE, 1991, 2: 415-426.

[18] 马黎,赵丽红,傅惠. 基于 BP 神经网络的交通异常事件自动检测算法[J]. 交通科技与经济, 2010, 12(006): 47-50.

[19] 赵晓娟. 基于多源数据的快速路交通事件自动检测算法研究[D]. 北京工业大学, 2010.

[20] 代磊磊, 姜桂艳, 韩国华. 高速公路事件自动检测算法综述[J]. ITS 通讯, 2005, 6(3): 1-5.

[21] 余勇. 快速路交通事件自动检测算法的研究与实现[D]. 北京: 北京交通大学, 2008.

[22] 戢晓峰. 城市道路交通状态分析方法回顾与展望[J]. 道路交通与安全, 2008, 8(3): 11-15.

[23] Karim A, Adeli H. Incident detection algorithm using wavelet energy representation of traffic patterns[J]. Journal of Transportation Engineering, 2002, 128(3): 232-242.

[24] Karim A, Adeli H. Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm[J]. Journal of Transportation Engineering, 2002, 128(1): 21-30.

[25] Levin M, Krause G M. Incident detection: a Bayesian approach[J]. Transportation Research Record, 1978 (682).

[26] Dudek C L, Messer C J, Nuckles N B. Incident detection on urban freeway[R]. 1974.

[27] Ahmed M S, Cook A R. Analysis of freeway traffic time-series data by using Box-Jenkins techniques[J]. Transportation Research Record, 1979 (722).

[28] Chew R L, Ritchie S G. Automated detection of lane-blocking freeway incidents using artificial neural networks[J]. Transportation Research Part C: Emerging Technologies, 1995, 3(6): 371-388.

[29] Persaud B N, Hall F L, Hall L M. Congestion identification aspects of the McMaster incident detection algorithm[J]. Transportation Research Record, 1990 (1287).

[30] Gall A I, Hall F L. Distinguishing between incident congestion and recurrent congestion: a proposed logic[J]. Transportation Research Record, 1989 (1232).

[31] 徐学才, 刘澜. 自动事件检测算法的比较及评估[J]. 交通科技与经济, 2004, 2: 42-44.

[32] Hsiao C H, Lin C T, Cassidy M. Application of fuzzy logic and neural networks to automatically detect freeway traffic incidents[J]. Journal of Transportation Engineering, 1994, 120(5): 753-772.

[33] Chang E C P, Wang S H. Improved freeway incident detection using fuzzy set theory[M]. 1994.

[34] Srinivasan D, Cheu R L, Poh Y P, et al. Development of an intelligent technique for traffic network incident detection[J]. Engineering Applications of Artificial Intelligence, 2000, 13(3): 311-322.

[35] Yuan F, Cheu R L. Incident detection using support vector machines[J]. Transportation Research Part C: Emerging Technologies, 2003, 11(3): 309-328.

[36] 姜桂艳, 温慧敏, 杨兆升. 高速公路交通事件自动检测系统与算法设计[J]. 交通运输工程学报, 2001, 1(1): 77-81.

[37] 宫晓燕, 汤淑明. 基于非参数回归的短时交通流量预测与事件检测综合算法[J]. 中国公路

学报, 2003, 16(1): 82-86.

[38] 李晓丹, 刘好德, 杨晓光, 等. 城市道路网络交通状态时空演化量化分析[J]. 系统工程, 2009, 26(12): 66-70.

[39] 张和生, 张毅, 胡东成, 等. 区域交通状态分析的时空分层模型[J]. 清华大学学报: 自然科学版, 2007, 47(1): 157-160.

[40] Rousseeuw P J, Croux C. Alternatives to the median absolute deviation[J]. Journal of the American Statistical association, 1993, 88(424): 1273-1283.

[41] Wittstein I S, Thiemann D R, Lima J A C, et al. Neurohumoral features of myocardial stunning due to sudden emotional stress[J]. New England Journal of Medicine, 2005, 352(6): 539-548.

[42] 李晗, 萧德云. 基于数据驱动的故障诊断方法综述[J]. 控制与决策, 2011, 26(1): 1-9.

[43] 曹巍.数据驱动建模方法研究及其工程应用[D].南京工业大学,2012.

[44] 苏金明. 基于数据驱动的流程工业过程性能监控方法研究[D]. 杭州电子科技大学, 2010.

[45] 陈建宏, 永学艳, 杨珊, 等. 基于时间序列模型的矿产品价格分析与预测[J]. 昆明理工大学学报: 理工版, 2009, 34(6): 9-14.

[46] 卢娟, 刘飞. 基于规范变量分析的动态多变量过程故障诊断[J]. 计算机测量与控制, 2007, 15(8): 984-986.

[47] Lee J, Yoo C, Lee I. Statistical process monitoring with independent component analysis[J]. Journal of Process Control, 2004, 14(5):467–485.

[48] Katayama T, Sugimoto S. Statistical Methods in Control & Signal Processing[M]. CRC Press, 1997.

[49] 邓晓刚, 田学民. 基于核规范变量分析的非线性故障诊断方法[J]. 控制与决策, 2006, 21(10):1109-1113.

[50] 张秀媛, 达庆东, 张国伍. 公路自动事件检测技术[J]. 系统工程理论与实践, 2001, 6: 118-124.

附　录

A. 作者在攻读学位期间申请的发明专利目录

[1] 廖孝勇，孙棣华，刘卫宁，古曦，赵敏，郑林江，乔真卿，崔德冠. 一种获取加气站等候车辆队列信息的方法，2013.3，中国，201310083184.8.

[2] 廖孝勇，孙棣华，刘卫宁，古曦，赵敏，郑林江，乔真卿，崔德冠. 一种基于动态修正的公交车到达时间实时预测方法，2013.12，中国，201310414620.

B. 作者在攻读学位期间取得的科研成果目录

[1] 作为主要研究人员参与了重庆市科学技术委员会科技攻关项目“面向电子站牌的公交到达时间预测新技术”（项目编号：cstc2012gg-yyjs00006).