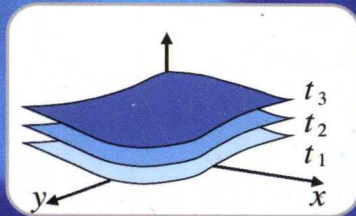
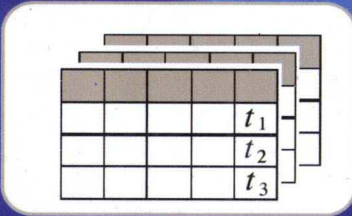
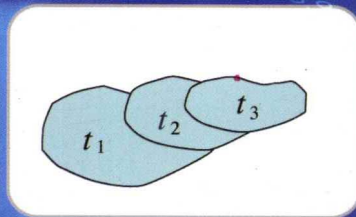
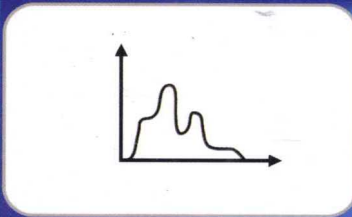




地球观测与导航技术丛书

时空序列数据分析和建模

王佳璆 邓敏 程涛 黄健柏 著



科学出版社

地球观测与导航技术丛书

时空序列数据分析和建模

王佳璆 邓敏 程涛 黄健柏 著

科学出版社

北京

内 容 简 介

时空数据分析是从海量的地理时空数据中提取信息、知识的有效手段。本书全面介绍了时空序列数据分析和建模的方法。在系统介绍时空数据分析的概念、内容、结构体系和研究进展的基础上,对时空数据的性质进行了探讨;然后以时空数据类型为线索,分别研究了空间点、线和面时空数据的时空一体化分析和建模方法,以及基于智能学习算法的时空数据分析和建模方法,并分别采用社会经济统计数据、环境温度数据、交通路网数据对各种时空模型进行了验证;分析比较了不同模型之间的优缺点及各模型的适用范围;最后对时空序列数据分析和建模方法未来的发展进行了展望。

本书结构严谨,理论、方法和应用结合紧密,是面向 GIS 专业时空数据分析课程的教材,既可供高等院校测绘、地理、经济、环境等专业的研究生使用,同时也可作为相关领域研究和技术人员参考书。

图书在版编目(CIP)数据

时空序列数据分析和建模/王佳璆等著. —北京:科学出版社,2012
(地球观测与导航技术丛书)
ISBN 978-7-03-033341-4

I. ①时… II. ①王… III. ①地理信息系统-研究 IV. ①P208

中国版本图书馆 CIP 数据核字(2012)第 005403 号

责任编辑:孙 芳 / 责任校对:宋玲玲
责任印制:赵 博 / 封面设计:陈 敬

科学出版社出版

北京东黄城根北街 16 号
邮政编码:100717

<http://www.sciencep.com>

新科印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2012 年 1 月第 一 版 开本:720×1000 1/16

2012 年 1 月第一次印刷 印张:9 3/4

字数:215 000

定价:48.00 元

(如有印装质量问题,我社负责调换)

《地球观测与导航技术丛书》编委会

顾问专家

徐冠华 龚惠兴 童庆禧 刘经南

王家耀 李小文 叶嘉安

主 编

李德仁

编 委 (按姓氏汉语拼音排序)

鲍虎军 陈 戈 程鹏飞 房建成 龚建华 龚健雅

顾行发 江碧涛 江 凯 景贵飞 李加洪 李 京

李 明 李增元 李志林 林 琿 林 鹏 卢乃锰

孟 波 秦其明 施 闯 史文中 吴一戎 许健民

尤 政 郁文贤 张继贤 张良培 周成虎 周启鸣

《地球观测与导航技术丛书》出版说明

地球空间信息科学与生物科学和纳米技术三者被认为是当今世界上最重要、发展最快的三大领域。地球观测与导航技术是获得地球空间信息的重要手段,而与之相关的理论与技术是地球空间信息科学的基础。

随着遥感、地理信息、导航定位等空间技术的快速发展和航天、通信和信息科学的有力支撑,地球观测与导航技术相关领域的研究在国家科研中的地位不断提高。我国科技发展中长期规划将高分辨率对地观测系统与新一代卫星导航定位系统列入国家重大专项;国家有关部门高度重视这一领域的发展,国家发展和改革委员会设立产业化专项支持卫星导航产业的发展;工业与信息化部和科学技术部也启动了多个项目支持技术标准化和产业示范;国家高技术研究发展计划(863 计划)将早期的信息获取与处理技术(308、103)主题,首次设立为“地球观测与导航技术”领域。

目前,“十一五”计划正在积极向前推进,“地球观测与导航技术领域”作为 863 计划领域的第一个五年计划也将进入科研成果的收获期。在这种情况下,把地球观测与导航技术领域相关的创新成果编著成书,集中发布,以整体面貌推出,当具有重要意义。它既能展示 973 和 863 主题的丰硕成果,又能促进领域内相关成果传播和交流,并指导未来学科的发展,同时也对地球观测与导航技术领域在我国科学界中地位的提升具有重要的促进作用。

为了适应中国地球观测与导航技术领域的发展,科学出版社依托有关的知名专家支持,凭借科学出版社在学术出版界的品牌启动了《地球观测与导航技术丛书》。

丛书中每一本书的选择标准要求作者具有深厚的科学研究功底、实践经验,主持或参加 863 计划地球观测与导航技术领域的项目、973 相关项目以及其他国家重大相关项目,或者所著图书为其在已有科研或教学成果的基础上高水平的原创性总结,或者是相关领域国外经典专著的翻译。

我们相信,通过丛书编委会和全国地球观测与导航技术领域专家、科学出版社的通力合作,将会有一大批反映我国地球观测与导航技术领域最新研究成果和实践水平的著作面世,成为我国地球空间信息科学中的一个亮点,以推动我国地球空间信息科学的健康和快速发展!

李德仁

2009 年 10 月

前 言

20 世纪 60 年代,计量革命推动了地理学研究从定性走向定量,与此同时诞生的地理信息系统(GIS)为空间信息采集、管理和分析提供了工具或技术,地理学理论、知识、分析方法等随着 GIS 技术的广泛应用而得到重视和发展。70~80 年代,空间分析方法的发展成为主流,空间分析方法的框架和体系逐渐建立起来,空间分析成为 GIS 的核心。90 年代以来,空间分析的发展进入了成熟期。随着对地观测技术、计算机技术、网络通信技术的飞速发展,空间数据获取不再成为分析地理问题的瓶颈,空间数据特征的研究受到了前所未有的关注。空间数据分析领域的研究十分活跃,新的方法和模型不断提出。由于受到不断增长的大量空间数据的驱动,从数据出发的探索性空间分析技术、可视化技术、空间数据挖掘技术等面向海量空间数据的分析方法受到重视,并且得到了深入的发展。不难发现,从 20 世纪 60 年代计量方法引入地理学研究到 90 年代空间数据挖掘技术的出现,主要还是以 GIS 为计算环境的空间分析方法的发展时期,在这期间,空间分析理论与方法体系基本形成。当前,随着海量地理时空数据的出现,迫切需要发展地理时空数据分析和建模方法来处理海量的地理时空数据,特别是发展时空一体化的分析和建模方法,这对 21 世纪地理学研究的发展提出了新的挑战。因此,构建时空一体化的时空分析模型,从海量的地理时空数据中挖掘知识,对于研究地理时空现象具有十分重要的科学意义。

2006 年 10 月 18~22 日,美国电气和电子工程师协会(IEEE)在中国香港举办了第一届空间和时空数据挖掘(Spatial and Spatio-Temporal Data Mining, SSTDM)国际研讨会。此次会议就空间和时空数据挖掘算法,时空数据分析、时空对象关系建模、时空数据挖掘的多尺度表达,时空数据模糊性及不确定性处理等专题进行了广泛的研讨。2008 年 9 月 10~12 日,在兰州召开的中国地理信息系统协会/GIS 理论与方法专业委员会学术研讨会上,将时空信息机理、时空数据分析、挖掘及地理系统建模、时空数据融合、模型集成及空间决策支持系统作为会议主题,会议主要围绕当前时空分析的理论与方法等前沿问题进行深入交流与系统研讨,促进了国内 GIS 时空数据分析和地理系统建模理论与方法的研究与发展。2011 年 4 月 12~16 日,在西雅图举行的美国地理学家(AAG)年会将时空一体化的地理学与地理信息科学作为会议的主题,体现了国际上对时空数据分析和建模方法研究的重视。

地理学中,由于连续的时空数据都是经过离散化抽样提取并存储的,因此能够将时空数据看作是在空间上有相关关系的时间序列的集合,即时空序列(space-time series)。时空序列分析和建模属于时空建模的范畴。对时空序列数据的分析、建模及预测称为时空序列分析。目前,在地理学研究中,有关时空序列分析和建模方面的研究成果较少。近年来,神经网络、支持向量相关等智能计算方法的发展为时空序列建模提供了新的手段,面对日益庞大的海量时空数据,迫切需要探索新的时空数据分析工具才能发现其中的地理知识,掌握其变化规律。本书在集成时空自相关移动平均模型的基础上,借鉴神经网络、支持向量相关等智能计算方法应用于时间序列分析和空间数据分析中的思想,从寻找易

用、准确、可靠、实用性强的时空序列建模方法的角度出发,在对时空自相关移动平均模型进行深入系统研究的基础上,构建了几种新的时空序列模型,并通过这些模型在实例中的表现,对它们的性能进行了全面的分析和评估。

时空数据分析包含的内容十分广泛,作为一本教科书,既要考虑内容的全面性,又要有所取舍。因此,在参照国外教材体系及相关文献内容的前提下,考虑国内教学和课程体系的需要,本书的内容主要包括时空数据的性质、探索性时空数据分析、时空自相关移动平均模型、非平稳时空序列建模、空间点和面时空数据的时空神经网络模型、时空支持向量机模型等。第1章对时空序列分析建模的概念、内容、框架体系和研究进展作了系统介绍;第2章探讨了时空数据的性质及有关问题;第3章研究了近年来日益受到重视的时空自相关移动平均模型,并以空间点、线和面时空数据为线索,分别研究了在社会经济统计、交通路网、环境监测方面的应用;第4章研究了时空序列混合建模框架和模型;第5章对时空神经网络进行了研究;第6章研究了时空序列支持向量机模型;第7章对各种时空序列模型的优缺点进行了比较和总结。全书较系统地反映了时空数据分析方法和技术的最新进展与主要成果。

本书的研究工作得到了国家863计划项目(2009AA12Z206)、教育部新世纪人才支持计划(NECT-10-0831)、中南大学前沿研究计划(2010QYZD002)和中国博士后科学基金(20090461019)的资助。另外,国家973计划项目(2006CB701305)和中山大学“985工程”GIS与遥感的地学应用科技创新平台也给予了资助。

在本书撰写过程中得到了很多人的帮助。感谢中山大学黎夏教授对本书研究工作的指导!感谢刘启亮等在文献资料整理和文稿校核方面付出的辛勤劳动!本书的成书过程亦得到了中南大学各级领导的关心与支持。感谢中南大学人事处唐忠阳副处长、科技部吴厚平副部长、李启厚副处长、研究生院刘少军副院长、陈立章主任给予的鼓励与帮助!感谢中南大学博士后流动站周元敏老师的大力支持!感谢地球科学与信息物理学院朱建军教授、柳建新教授、刘兴权教授、邹峥嵘教授等学院领导在本书撰写过程中给予的指导、关心和支持!

由于时间、资料、知识等多方面的局限性,书中疏漏在所难免。望对这一领域有兴趣的读者、专家不吝赐教,以使作者完善本书。

作者

2011年10月

目 录

《地球观测与导航技术丛书》出版说明

前言

第 1 章 绪论	1
1.1 时空序列分析建模的发展背景	1
1.2 时空序列分析建模的研究概况	2
1.3 时空序列分析建模的应用	5
1.4 本书的主要研究内容及结构安排	8
1.5 本章小结	9
参考文献	9
第 2 章 时空数据的表达及基本性质	14
2.1 地理时空的理解	14
2.2 时空数据的表达	15
2.3 时空数据的基本性质	16
2.4 本章小结	26
参考文献	26
第 3 章 时空自相关移动平均模型	28
3.1 自相关模型	28
3.2 移动平均模型	30
3.3 时间自相关移动平均模型	32
3.4 空间/时空自相关移动平均模型	34
3.5 实例 1——空间面状数据的时空预测	44
3.6 实例 2——交通路网数据的时空预测	55
3.7 本章小结	60
参考文献	61
第 4 章 时空序列混合框架和模型	62
4.1 非平稳时空过程模型方法	62
4.2 非平稳时空序列混合建模框架	63
4.3 实例——空间点数据的时空预测	69
4.4 本章小结	82
参考文献	82
第 5 章 时空序列神经网络模型	86
5.1 神经网络模型	86

5.2	时空神经网络模型	88
5.3	网络的结构及工作方式	91
5.4	网络的学习方法和算法	93
5.5	时空非平稳建模	95
5.6	实例 1——空间面状数据的时空预测	98
5.7	实例 2——空间点数据的时空预测	104
5.8	本章小结	110
	参考文献	110
第 6 章	时空序列支持向量相关模型	112
6.1	机器学习概论	112
6.2	统计学习理论	114
6.3	支持向量机的发展及应用	116
6.4	多输出支持向量相关算法	117
6.5	构造时空核函数	122
6.6	实例 1——空间面状数据的时空预测	126
6.7	实例 2——空间点数据的时空预测	131
6.8	本章小结	135
	参考文献	135
第 7 章	总结与展望	137
7.1	模型比较及讨论	137
7.2	主要研究结论	138
7.3	研究展望	140
附录	194 个国际气象交换站描述性统计表	141

第 1 章 绪 论

1.1 时空序列分析建模的发展背景

20 世纪,科学技术的快速发展促进了地理学研究的飞跃。对地观测技术、计算机技术、网络通信技术的重大突破是推动地理学大发展的动力。地理信息技术的出现促进了地理分析方法的发展和应用,开创了地理学研究的崭新时代(王远飞等,2007)。地理信息技术的出现不仅为地理学研究提供了新的技术支持,而且形成了以地理信息科学为核心的新的研究领域、理论和方法。空间分析是地理信息科学的核心领域,它是地理学分析方法中区别于其他定量分析的推演分析框架和体系。空间数据分析是围绕地理学问题的“空间”本质建立模型的方法。随着地理信息系统(GIS)、遥感(RS)和全球定位系统(GPS)等技术的发展和广泛应用,积累了海量的地理时空数据,如经济统计数据、卫星影像数据、环境监测数据、交通流量数据等。这些数据呈现复杂的时间关系和空间关系,具有多源、多变量、异构、海量、多尺度、多时相等特点(王劲峰等,2006)。长期以来,由于缺乏有效的时空数据建模方法,分析和处理这些时空数据非常困难,导致不能充分地认识和利用隐含在时空数据中的有用信息。因此,有效地分析和建模地理时空数据,构建时空一体化的时空预测模型,对于研究地理时空现象具有十分重要的科学意义(王劲峰等,2006; Yao,2003;徐冠华,1994)。

地理学中,由于连续的时空数据都是经过离散化抽样提取并存储的,因而能够将时空数据看作是在空间上有相关关系的时间序列集合,即时空序列(Cliff et al.,1975;Curry,1970)。时空序列分析和建模属于时空建模的范畴(Kamarianakis et al.,2005;Cressie et al.,1997;Pfeifer et al.,1980a;Martin et al.,1975;Cliff et al.,1975)。对时空序列数据的分析、建模及预测称为时空序列分析。最早提出这种观点的是加拿大多伦多大学的 Curry。Curry(1970)认为时空序列分析是时间序列分析在空间上的扩展,有必要将空间分析的内容引入到时间序列分析中来建模。英国利兹大学的 Wilson(1972)认为,为了掌握时空动态模式,应该把时空动态过程看成是空间模式随时间演化的观点。1974 年,英国剑桥大学的地理学家 Cliff 和沃里克大学的统计学家 Ord 在 Curry 和 Wilson 等工作的基础上,以时间序列模型为基础,提出了时空序列建模的框架(Cliff et al.,1975),他们在进行时空建模时,基于空间相关性的考虑,提出将空间自相关(spatial autocorrelation, SAR)、空间移动平均(spatial moving average, SMA)和空间相关(spatial regression, SR)模型引入到时间序列模型当中,使之成为时空序列模型,该模型框架实质上是时间序列模型(Box et al.,1970)在空间上的扩展,并首次提出了时空自相关(space-time autocorrelation)的概念。此外,Cliff 和 Ord 在进行时空序列建模时,考虑到空间异质性和时间非平稳性,建议采用时空协方差函数来处理时空交互作用对模型的影响,但他并没有在框架中实现。在 Cliff 和 Ord 提出的框架基础上,剑桥大学 Martin 和 Oeppen 对时空相关性的度

量进行了深入研究,提出了时空自相关函数(space-time autocorrelation function, ST-ACF)、时空偏自相关函数(space-time partial autocorrelation function, ST-PACF),并在英国皇家地理学会的杂志 *Transactions of the Institute of British Geographers* 上发表文章,提出了时空自相关移动平均模型(spatio-temporal autoregressive and moving average, STARMA)(Martin et al., 1975)。STARMA 模型本质上是线性的随机时空动态模型。在 STARMA 模型中, Martin 和 Oeppen 创造性地用时空延迟算子(space-time lag operator)的形式表达时空变量受时间延迟和空间延迟的共同影响,使 STARMA 成为真正的时空一体化模型。在随后的研究中,美国弗吉尼亚大学 Pfeifer 和亚特兰大大学 Deutsch 教授在 *Transactions of the Institute of British Geographers*、*Geographic Analysis*、*Technometrics*、*International Journal of Forecasting* 等国际著名杂志上发表了一系列文章,对 Martin 和 Oeppen 提出的 STARMA 模型进行了完善,并提出了如 Box 和 Jenkins(1976)的三阶段迭代过程(three-stage iterative procedure)的时空建模方法,即模型参数识别、模型参数估计和模型检验,对时空建模过程进行了标准化(Pfeifer et al., 1990, 1981a, 1981b, 1981c, 1980a, 1980b, 1980c, 1980d)。目前, STARMA 模型在经济(Pfeifer et al., 1980b)、交通(韩卫国等, 2007; Kamarianakis et al., 2005)、犯罪学(Pfeifer et al., 1980a)、水文(Cressie et al., 1997; Deutsch et al., 1986)、遥感影像预测(Crespo et al., 2007)、房地产价格预测(Elhorst, 2001; Pace et al., 2000, 1998)等领域都有成功的应用。近年来,时空建模领域又发展出了许多分支,如时空智能数据分析、时空数据挖掘、时空聚类、时空模拟、时空可视化等。然而,时空序列分析与建模的理论与方法并没有得到较大发展。

当前,海量的时空序列数据特别是长时间序列数据的迅速增长,迫切需要探索基于数据驱动的时空序列建模新方法(王劲峰等, 2006)。而传统的基于模型驱动的时空序列建模方法计算较为复杂,建模工作要求较高,不易普及。人工智能等新的智能计算方法发展起来以后,神经网络、支持向量机等基于数据驱动的方法为时空序列建模提供了新的手段。20 世纪 80 年代成立的美国国家地理信息和分析中心(NGCIA)将智能化的地理信息处理可视化的表达列为地理信息系统技术发展的重要内容。1999 年,英国利兹大学计算地理研究中心的 Openshaw 在报告 *Geographical data mining: Key design issues* 中指出地理计算(geocomputation)的智能化是今后地理计算发展的趋势(Openshaw, 1999)。

鉴于以上原因,本书对地理时空数据的性质进行了较为深入的分析 and 探讨,针对现有的时空序列模型——STARMA 模型进行了深入系统地分析和研究,在掌握时空序列建模原理的基础上,借鉴神经网络、支持向量相关等智能计算方法应用到时间序列和空间数据分析中的思想,从寻找易用、准确、可靠、实用性强的时空序列模型方法的角度出发,构建新的时空序列模型,探索新的时空数据分析和建模方法,这亦是本书的主要目标。

1.2 时空序列分析建模的研究概况

时空序列是时间序列在空间上的扩展,指在空间上有相关关系的多个时间序列的集

合(Kamarianakis et al., 2005; Pfeifer et al., 1980a; Martin et al., 1975; Cliff et al., 1975)。时空建模是指依据给定的时空数据寻找一种分析方法,对未观测时空位置的属性值进行建模和预测的过程(Cliff et al., 1975)。时空序列分析与建模的研究历史可追溯到 Cliff、Ord、Curry 和 Wilson 等科学家们的工作。Cliff 和 Ord 首次提出了时空自相关的概念,并将时间序列模型扩展到空间域,建立了时空序列模型框架(Cliff et al., 1975)。Martin 和 Oeppen(1975)进一步深入研究了时空自相关问题,给出了时空自相关函数和时空偏相关函数的具体定义和计算方法。与此同时,提出了第一个真正意义上的时空相关统计模型——STARMA 模型。STARMA 模型假设时空序列中任一变量值 $z_i(t)$ 都是当前位置 i 及其空间邻接位置过去时期观测值与随机误差项的线性组合(Kamarianakis et al., 2005; Pfeifer et al., 1980a),因此,其本质上是一个线性模型,只能捕捉线性的时空自相关结构。此外,在建模时, Martin 和 Oeppen 假设时空序列是平稳的时空过程,而对于非平稳时空序列,需要进行平稳转化后才能用 STARMA 进行建模。因此, Martin 和 Oeppen(1975)建议用时间序列中的差分法将非平稳时空序列转化为平稳时空序列来处理,这对于空间和时间上都离散的时空序列是适用的,如国民生产总值、土地利用类型、房地产价格等反映社会经济统计的序列数据,因为通常这类时空序列在空间域上描述的是这个空间单元的总特征,与空间位置无关,不需要考虑空间过程的平稳性,而只需要考虑时间过程的平稳性。对于空间上连续时间上离散的时空序列,如降雨量分布、空气污染浓度分布及土壤重金属含量分布等反映自然地理现象的序列数据则要考虑时空平稳特征,用差分法仅能处理时间非平稳特征,对处理空间非平稳(或异质性)无能为力。一般来说,对这类时空序列的分析与处理要比平稳时空序列复杂得多,目前尚无统一、规范的方法(Kamarianakis et al., 2005; Stroud et al., 2001)。此后,美国弗吉尼亚大学的 Pfeifer 和亚特兰大大学的 Deutsch 对 STARMA 的建模过程进行了标准化的定义,提出了与时间序列模型相仿的时空序列建模三阶段迭代过程,即模型识别、参数估计、模型检验。模型识别即通过计算时空自相关函数和偏自相关函数辨别 STARMA 模型的时间自相关阶数、空间自相关阶数、时间移动平均阶数及空间移动平均阶数。参数估计用极大似然法估计模型参数。极大似然估计方法的出发点是对模型中误差项指定联合正态密度函数,从中得到似然函数,然后对该函数进行最大化(Anselin et al., 1998; Cressie, 1993; Anselin, 1988)。模型检验即用时空自相关函数检验模型残差是否为随机误差。

随后,在其他学科领域中也出现了许多时空模型。将这些时空模型按照实现方法的不同进行大致分类,主要包括时空动力学方法、时空相关统计学方法、时空地统计学方法、时空插值方法、状态空间和随机理论方法、卡尔曼滤波方法、层次贝叶斯方法、智能计算方法和时空数据挖掘方法。时空动力学方法首先在掌握时空过程机制和主导因素的基础上建立动力学模型,然后充分考察事件的空间和时间特征,对事件的时间变化及空间分布的影响展开分析和模拟,如传染病时空传播模型 SSIR 和热传播抛物线方程,这些方程可以被用来预测和模拟各种环境或控制下研究对象个例的分布情景(王劲峰等, 2006)。时空动力学方法的缺点是当时空过程的机制和主导因素不甚清楚时,无法进行建模。时空相关统计方法又叫时空序列建模方法,首先对已知的时空序列观测样本数据建立时空相关-自相关函数,得到时空序列观测样本数据的分布模型,然后使用时空序列数据分布模型估

算未观测时空位置数据值(Cheng et al., 2011; 韩卫国等, 2007; 周国法等, 1999; Pfeifer et al., 1981a, 1981b, 1980a, 1980b, 1980c, 1980d; Martin et al., 1975; Cliff et al., 1975)。STARMA 模型是典型的时空相关统计方法。由于 STARMA 模型充分考虑了时空自相关性等地理时空序列数据的特征, 因此较适合于分析和处理地理时空数据。时空插值方法又可分为约减和扩展法。约减法将时间维和空间维分开考虑, 首先对每个样本点在时间维上建立内插函数, 再将时间插值函数用于空间插值中, 从而得到时空插值结果。扩展法将时间维和空间维统一考虑, 将空间插值方法扩展到时空高维空间中进行插值的方法(Revesz et al., 2002; Antonic et al., 2001)。但是, 空间内插方法不仅忽略了时空观测样本数据对插值结果的影响权重, 而且空间样本数量的大小和观测时间的长短对时空插值结果的稳定性和准确性都会产生很大的波动。时空地统计学方法(space-time geostatistic), 即时空信息统计法, 是将地统计学方法扩展并应用到时空领域中, 当空间样本数量较大, 而观测时间相对较短时, 时空 Kriging 方法的估计结果会产生奇异(Chiles et al., 1999; Goovaerts, 1997; Buxton et al., 1996; 侯景儒等, 1995; Rouhani et al., 1990)。状态空间和随机理论方法首先将时空观测样本映射到状态空间中, 然后在状态空间中使用随机理论构建状态转换或状态相关模型, 进而使用转换或相关模型估计未观测时空位置的数值(Wikle et al., 2001; Brown et al., 1993)。这种方法较适于处理有较强时效性的时空数据。卡尔曼滤波也可以用于时空动态建模(Goodall et al., 1994)。卡尔曼滤波方法是以最小均方误差作为估计的最佳准则, 来寻求一套递推估计的算法, 其基本思想是: 采用状态空间模型, 利用前一时刻的估计值和当前时刻的观测值来更新对状态变量的估计, 求当前时刻的估计值。卡尔曼滤波是解决状态空间模型估计与预测的有力工具之一, 不需存储历史数据, 能够进行实时的预测, 但较适合于处理具有非常复杂的动态结构的时空数据(Huang et al., 1996)。层次贝叶斯方法是运用贝叶斯统计进行的一种时空建模方法, 层次贝叶斯统计不同于一般的统计方法, 其不仅利用模型信息和数据信息, 而且还能充分利用先验信息(朱庆杰等, 2002)。层次贝叶斯方法的缺点是很难描述数据的时空依赖关系和协方差结构, 很难准确地模拟出空间变异特征(Stroud et al., 2001)。近年来, 随着神经网络、支持向量机等智能计算方法被引入时空数据处理领域, 该方法又可分为基于支持向量机的方法、基于 Volterra 系统的方法、基于神经网络的方法、基于滤波的方法、基于元胞自动机的方法和模糊逻辑等方法(Cheng et al., 2008, 2007; Wang et al., 2007; 黎夏等, 2007; Xia et al., 2006; 徐薇等, 2004; 张显峰等, 2001)。基于智能计算的时空建模方法亦是未来发展的一个重要研究方向。

时空数据挖掘方法是在有效地组织时空数据库的基础上, 采用各种数据分析工具、技术和模型来发现时空数据库中隐含的信息和规律(Cheng et al., 2009, 2008; 吴忠东等, 2004; Yao, 2003; 陈捷等, 2002)。时空数据挖掘方法出现的时间还不久, 主要是以经典的数据挖掘理论研究为基础, 同时还受到时空数据表示和存取方式的限制(符海芳等, 2003)。时空数据挖掘的任务包括时空关联规则挖掘(Mennis et al., 2005)、时空异常模式探测(Liu et al., 2011; Cheng et al., 2006)、时空聚类(Kalnis et al., 2005)、时空预测和趋势分析(Cheng et al., 2008; Xu, 2004)等。时空关联规则是传统数据挖掘中关联规则的时空扩展, 即空间上某地 A 有某一事件在某一时刻发生, 在一定条件下导致该时刻或若干时刻后另一事件在该地 A 或另一处发生(苏奋振等, 2004)。时空关联规则形如:

$A > B[s\%, c\%]$, 其中, A 和 B 分别为时空谓词的集合; $s\%$ 为规则的支持度; $c\%$ 为规则的置信度(李光强等, 2010; Verhein et al., 2006)。对时空关联规则的挖掘有两种方法: 一种是在空间数据挖掘系统中加入时间变量; 另一种是在时态挖掘系统中结合空间分析(Roddick et al., 1999)。时空异常模式指在挖掘到的关联规则或模式中可能出现一类异常模式, 这类模式很可能是一类重要的隐含信息, 代表地理现象或地理过程的特殊性。时空异常模式探测分为两类: 一类是时空分治法, 另一类是时空综合法。时空分治法是分别从时间、空间邻域中使用传统的异常判别法检测时间、空间上均表现为异常的数据, 即时空异常数据(Liu et al., 2011; Cheng et al., 2006; Sun et al., 2005)。时空综合法是在探测过程中将时间和空间综合进行考虑(Birant et al., 2006; Adam et al., 2004)。时空聚类是按一定距离或相似度在时空数据集中标识出聚类或稠密分布的时空区域, 将时空数据分成一系列相互区分的簇, 簇内成员的属性值尽可能相同, 而不同簇间的属性值尽可能不同, 从而发现时空数据集的时空分布规律和典型时空模式(Deng et al., 2011; Kalnis et al., 2005)。时空趋势预测是根据对象的空间属性, 分析某些非空间属性随着时间和空间对象间关系变化而变化的规律(Cheng et al., 2008; Wang et al., 2007; Xu et al., 2004)。目前, 时空数据挖掘已广泛应用于移动位置服务、遥感信息处理、精细农业、无线通信、智能交通、传染病防治、气象等诸多领域, 各个行业领域对时空数据挖掘技术的巨大需求也推动了时空数据挖掘理论和方法不断发展和完善(王劲峰等, 2006)。然而, 地理空间域的复杂性决定了绝大多数地理时空数据集呈现出时空自相关和异构性、时空异质性等特征, 为时空数据挖掘在地理学中的发展提出了一系列挑战性的问题(Miller et al., 2001), 这也是时空数据挖掘应用于地理学领域首要解决的问题。

此外, 由于时空信息大多是多源、复杂、交互综合作用的结果, 要实现基于时空信息复合过程的科学预报, 不仅需要实现多源复杂因子的分解, 更需要将多种分解模式重新组织融合, 形成一个新的整体模式。王劲峰等(2006)提出了针对无主导机制地学过程模拟与预测的时空结构自适应模型, 或称时空信息多模式融合模型, 其基本思想是将时空信息进行分解, 然后用统计、时空耦合、神经网络、支持向量机等方法将各部分进行融合。因此, 时空信息多模式融合模型是将相关、时空耦合、神经网络、时间周期谱分解等多个模型融合, 形成新的多源复杂时空信息预测整体模型, 它充分发挥各个模型的优势, 不仅有利于模型的改进和完善, 而且可以提供更准确的科学预测, 从而使其具有更为广阔的应用领域。

当然, 上述理论和方法都不是孤立的, 时空建模时经常要综合运用它们。目前, 将时空相关统计学方法和智能计算方法结合是当前时空信息建模的发展趋势(王劲峰等, 2006)。因此, 时空序列建模要充分借鉴和吸收成功的智能计算等方法, 研究发展新的时空序列模型, 以适应当前地理学发展的需要, 推动时空数据分析和建模方法的发展。

1.3 时空序列分析建模的应用

时空模型在传染病学、气候、环境监测、农产品价格、公共卫生等领域有着广泛的应用, 其成功应用在很大程度上影响着时空数据分析和建模方法的发展, 因此, 有必要简要总结时空模型在各个领域的典型应用成果。

在传染病学领域,王劲峰等(2006)运用时空动力学模型对传染病传播机制进行了研究,并运用传染病时空传播模型 SSIR,以北京市 SARS 数据为例,结合北京市地理信息系统,基于数据驱动和模型驱动的理论和技术,利用热点分析、空间过程分析和因子识别等数据探索分析方法,根据传染病的多维传播特性,同时利用遗传规划和模拟退火相结合的算法对模型求解,从模型直接求取 SARS 流行病学参数,结果揭示了北京市人群在两个空间尺度上的流动接触过程,为人-人接触性传染病的预防和控制策略制定提供了重要依据。Waller 等(1997)研究了层次贝叶斯模型在疾病分布图中的应用,疾病分布图是传染病学研究中必不可少的工具之一,其主要目的是通过直观地呈现疾病分布的地理变异,形成新的病因学假设,即论证某种疾病是由具有空间分布的某些危险因素引起的。基于层次贝叶斯框架的模型能够分析和处理不同尺度下大数据量的时空数据集。

在气候领域中,华盛顿大学的 Kooperberg 等(1995)提出了一种针对温度数据时空分解的预测性振荡模式(predictive oscillation patterns, PROPS)方法用于反映时空变化。PROPS 方法首先在空间上将信息分解到多个空间尺度或空间模式上,然后再刻画各尺度或模式上信息的时域变化情况。该方法以北半球 445 个网格点 47 年的温度数据为实验数据进行验证,其中一半的数据用于模型的估计,采用协方差矩阵(covariance matrix)和相关形式对模型的参数进行估计,另一半数据用于模型的检验,并与用于空间模式分解的主成分分析(principal component analysis)和用于时域分解的主振荡模式(principal oscillation patterns)进行比较。PROPS 方法在预测精度和误差上均有所改善。

在降雨量时空建模中,Sanso 等(1999)利用二次多项式拟合其空间趋势,利用谱分解拟合其时域的季节趋势,在去除空间趋势和时间趋势后,假设降雨量的时空分布符合多维正态分布,并且对分布参数估计时考虑不确定因素,利用贝叶斯方法估计参数的不确定性考虑到参数的后验分布中,从而提高预测精度。Stroud 等(2001)提出了非平稳时空数据的状态空间框架,使得状态空间模型可以处理非平稳空间过程和时空相关,并以热带降雨量和北大西洋海平面温度为例对模型进行验证,实验表明所提出的框架是可行的。

在环境监测方面,Tonellato(1998)研究了有空间相关性的多个固定站点的时间序列,提出了自相关状态空间模型,并用于风力预测。Iaco 等(2001)对 52 个监测站 10 个月的 NO_2 平均浓度进行建模时,首先将其看作多个时间序列的集合,每个空间位置对应一个时间序列,分别拟合各时间序列的趋势和季节分量,然后对时空残差利用时空变异函数建模,时空残差建模采用积和模型(product-sum model)。Huang 等(2004)在对臭氧浓度进行时空建模时,考虑到风速、风向对浓度的影响,基于空间异质性和时间非平稳性的考虑,采用时空非分离的协方差形式,将一个时空过程分解为一个确定性的均值过程和一个随机的时空变异两部分,并用状态方程对平稳时空数据建模。

在农产品价格的时空建模方面,Roehner(1999)对法国 1825~1913 年 500 个市场的小麦价格进行了研究,首先对价格时空相关函数在不同时间窗口下与空间距离的关系及在不同空间距离下与时间步长的关系进行了探索,表明小麦价格波动具有空间和时间依赖性,然后基于空间套汇假设(spatial arbitrage assumption)和波传播(wave propagation)思想对小麦价格建立了随机偏微分方程,并从随机偏微分方程导出了时空相关函数的形式,对方程参数进行了估计,最后对方程的预测结果进行了经济学解释。

在公共卫生领域,斯坦福大学的 Buckeridge 等(2003)提出了一个通用的监测信息时空异常探测建模的分析框架,将探测任务分解为 5 个子任务:①指定探测目的;②定义数据结构;③组织观测数据;④预测期望数据;⑤比较观测和预测数据。Buckeridge 提出的时空异常探测建模框架主要特色在于其建模是基于知识构建,认为在各个子任务中具体模型和方法的选择应充分利用相关领域知识,并且对知识建模。然而,Buckeridge 等也说明目前还没有进行相应的实例分析和领域知识建模研究。

此外,越来越多的研究表明,STARMA 模型是建模地理时空现象的有效工具。在社会经济(Pfeifer et al., 1990, 1980b)、交通(韩卫国, 2007; Kamarianakis et al., 2005)、遥感影像处理(Crespo et al., 2007)、房地产价格(Elhorst, 2001; Pace et al., 2000, 1998)、水文地理(Cressie et al., 1997; Deutsch et al., 1986)等地理学及其相关领域都得到了广泛的应用,特别是在时空过程机理不清楚、多因素时空数据不全的情况下,单因素的时空序列模型发挥的作用更为重要。例如,韩卫国等(2007)利用 STARMA 模型建立路口间交通流的时空相关关系,用于区域交通流的短时预测和时空分析,采用长安街及其沿线路口监测站点的区域交通流量作为试验数据,验证了该模型在交通流的短时预测和时空分析中的可行性。Pfeifer 和 Deutsch(1980b)应用 STARMA 模型对美国 25 主要产粮区 1925~1964 年农业机械化比例进行了时空预测,实验结果与 Casetti 和 Semple(1969)提出的时空逻辑模型和 Cliff 和 Ord(1975)提出的线性时空相关模型进行比较显示,STARMA 模型均好于时空逻辑和线性时空相关模型,表明了 STARMA 模型是一种有效的格数据预测工具。Crespo 等(2007)用 STAR 模型对高低压等高线遥感影像序列进行了预测,由于遥感影像序列数据量巨大,Crespo 首先将序列影像进行压缩,得到压缩影像后再建立时空自相关(STAR)模型进行压缩影像序列的预测,然后将压缩影像进行还原,得到的预测结果显示了时空自相关模型具有很好的短时预测能力。Pace 等(2000, 1998)用时空自相关模型成功地对房地产价格数据时空误差进行了拟合,并对时空最邻近房屋价格进行了估计,实验表明,时空自相关模型通过建模时间和空间延迟能够成功地捕获房地产价格在时间和空间上变化,对比研究显示时空自相关模型的预测精度优于传统的指示预测模型。Martin 和 Oeppen(1975)用 STARMA 模型对英格兰东部地区和爱尔兰地区畜牧业产量进行了时空预测,显示出其在预测空间面状(离散分布)数据的有效性。另外,Deutsch 和 Ramos(1986)的研究结果也显示出了 STARMA 模型在水文地理学方面有成功的应用。

综上所述,时空建模方法在诸多领域都有许多成功的应用,但由于大多数模型是针对具体应用建立的,并不都能反映地学数据的本质特征,真正能够用在地学及其相关领域的时空模型并不是很多。此外,由于 STARMA 模型充分考虑了时空自相关性等地理时空数据的特征,使它更适合于地理时空序列数据建模。因此,近年来,STARMA 模型在地学及其相关领域的应用相对较多,然而,其在建模非平稳和非线性时空序列方面的理论和应用研究较少(Openshaw, 1999)。经过对 STARMA 模型更深入的研究发现,其在某些方面还存在局限性,如残差分析显示对局部空间变异的解释不够等,因此,需要在这些方面做进一步探索研究。当前,海量的地理时空序列数据的迅速增长,迫切需要探索基于数据驱动的时空序列建模新方法(王劲峰等, 2006),而人工智能等新的智能计算方法发展起来以后,神经网络、支持向量机等基于数据驱动的方法为时空序列建模提供了新的手段。

因此,有必要在掌握时空序列建模原理的基础上构建新的时空序列模型,探索新的时空数据分析 and 建模方法是本研究的主要目标。

1.4 本书的主要研究内容及结构安排

本书主要研究了时空序列数据分析和建模的理论和方法。研究工作是在对时空序列数据的性质进行较为深入的分析 and 讨论之后,针对 STARMA 模型展开了较为系统的分析和研究,在掌握时空建模原理的基础上,结合智能计算方法,提出了时空序列混合框架和模型、时空序列神经网络模型(STANN)及时空序列支持向量相关模型(STSVR)等新的时空序列建模方法,并用实例对模型进行了较为全面的分析和评估。本书以东莞市 32 个镇 1994~2002 年国内生产总值(GDP)时空预测,我国 137 个气象站点 1951~2002 年年平均气温时空预测,以及伦敦交通路网数据时空预测为例,展开对时空序列建模的研究。具体研究内容包括以下几个方面:

(1) 对时空序列数据的性质进行了深入分析和讨论。时空序列数据与一般的时间序列数据和空间数据相比,具有海量、动态、高维、多尺度、非线性、时空相关和时空异质性等特征,并着重讨论了时空自相关和时空平稳性质。研究时空数据的性质对于时空序列分析和建模非常重要。

(2) 对 STARMA 模型进行了系统的分析和研究。首先讨论了自相关模型,包括时间自相关模型、空间自相关模型、时空自相关模型,然后分析了移动平均模型,包括时间移动平均模型、空间移动平均模型、时空移动平均模型,最后研究了 STARMA 模型建模过程。以东莞市 32 个镇 1994~2002 年 GDP 和伦敦交通路网数据时空预测为例,验证了模型预测空间面状和网状数据的有效性。

(3) 对时空异质性进行了研究,提出了时空序列混合模型及时空数据分析四阶段过程。混合模型的基本思想是:运用神经网络模型多层感知器算法提取非平稳时空序列中的大尺度或全局的时空非线性趋势,然后用 STARMA 建模小尺度或局部的随机时空变异。其中,高斯半变异函数被用于测度 STARMA 模型的空间权重。时空数据分析四阶段过程分别是时空数据准备、探索性时空数据分析(exploratory spatial and temporal data analysis, ESTDA)、时空数据训练和时空数据验证。运用提出的混合模型和四阶段过程,以我国 137 个气象站点 1951~2002 年年平均气温的时空预测为例,对提出的模型进行验证,表明模型预测空间点(或连续分布)数据的有效性,进一步对模型的残差分析结果表明,提出的模型在解释时空变异方面优于传统的 STARMA 模型。

(4) 深入研究了 STARMA 模型中的时空延迟算子,结合静态和动态神经元原理,提出了“时空”神经元模型,用“时空”神经元构造了一种新的动态神经网络模型——STANN 模型,并将空间邻接性或距离等先验知识嵌入到了 STANN 模型设计当中,提高了模型收敛速度。另外,通过在 STANN 模型中引入折扣最小二乘(discounting least square, DLS)误差评估准则,使其具有了非平稳时空序列建模的能力。并以东莞市 32 个镇 1994~2002 年的 GDP 和我国 137 个气象站点 1951~2002 年年平均气温为例,对提出的模型进行验证,表明 STANN 模型拟合与泛化性能优于 STARMA 模型,配对 t 检验结果显示 STANN 模型预测性能显著好于 STARMA 模型。