基于混合模型的地下水埋深时空预测研究

——以民勤绿洲为例

张仲荣　闫浩文　著

科学出版社

需要全本请访问 www.ertongbook.com

基于混合模型的地下水埋深时空预测研究

——以民勤绿洲为例

张仲荣　闫浩文　著

科学出版社
北京

内 容 简 介

本书以民勤绿洲地下水埋深时空预测为例，介绍了时空预测研究中常遇到的缺失数据修复、多序列预测、时空插值三大问题的最新混合建模成果，是时空数据分析与建模的新尝试。

本书是关于地下水埋深时空动态变化研究的论著，主要介绍地下水埋深空间分布预测的统计代理方法。模拟精度问题是基于代理模型的时空预测研究中一个非常重要问题。构建混合模型是提升时空预测精度的有效策略，基于时空统计理论、人工智能、机器学习的混合建模是当前研究的热点。

本书具有较强的理论性、系统性和实践性，可作为水文水资源、生态环境、水文地质等专业的本科生、研究生的教学参考书，也可供水文水资源及相关领域的科学人员参考。

图书在版编目(CIP)数据

基于混合模型的地下水埋深时空预测研究：以民勤绿洲为例/张仲荣，闫浩文著．—北京：科学出版社．2019.4

ISBN 978-7-03-060862-8

Ⅰ. ①基… Ⅱ. ①张… ②闫… Ⅲ. ①地下水–预测–研究 Ⅳ. ①P641.2

中国版本图书馆 CIP 数据核字（2019）第 049304 号

责任编辑：刘浩旻 韩 鹏 姜德君 / 责任校对：张小霞

责任印制：吴兆东 / 封面设计：铭轩堂

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

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 4 月第 一 版 开本：720×1000 B5

2019 年 4 月第一次印刷 印张：10 插页：2

字数：200 000

定价：98.00 元

（如有印装质量问题，我社负责调换）

前　言

民勤绿洲自然条件恶劣，水资源匮乏，生态环境脆弱。近几十年来，地下水系统功能严重退化，生态环境问题频发。如何以民勤绿洲为典型区域，科学、准确、及时地掌握干旱、半干旱区域内水资源量成为该类地区水资源研究中的一个关键问题。

区域地下水埋深是随地点和时间的变化而变化的时空变量，区域地下水埋深时空预测的目的是揭示未来地下水空间分布及其随时间的动态变化规律，为未来动态人工模拟提供依据。

时空数据分析为区域地下水埋深时空预测代理模型的构建提供了重要的理论支持。时空预测是区域变量时空数据分析的基本问题，近年来，已经成为国内外研究的热点。时空预测就是基于时空序列预测理论建立数学模型，在时域预测的基础上建立未来预测面，揭示未来地下水埋深空间连续分布及其时空变化，预测地下水埋深未来的分布。

地下水埋深的时空预测不同于传统的时间序列预测和简单的空间预测。如何提高精度是预测研究的核心问题。目前常用的模型对不同监测序列间复杂的时间和空间影响考虑不足，对非线性考虑不足，对不同的方法确定的参数带来的模型预测精准性考虑甚少。混合预测是将不同的模型，不同的时间序列分析理论混合、融合或集成一个预测模型，只有一个预测结果。混合预测模型是目前提高预测精度的重要而且前沿的研究方法。当前，基于非线性理论、多元时间序列分析理论、参数的启发式优化算法等结合的混合预测模型是提高预测精度的主要技术手段。

运用正确、科学的方法对民勤绿洲地下水资源进行研究，完成对民勤县境内地下水埋深的模拟，建立研究区域地下水空间预报体系，对区域地下水问题的解决具有重要的理论意义和实用价值。

本书在针对民勤绿洲地下水埋深时空预测研究中遇到的三个主要问题如下。

1）缺失数据的修复问题

在长期的监测过程中，受人为原因和水文地质等原因的影响，存在两种类型的缺失数据：一是同一站点，间断检测造成的间断式缺失数据；二是监测井废

弃，换位新建造成的截断式缺失数据。传统研究往往是将换位前后的序列直接拼接。水文地质条件的差异 、土壤植被的差异等，使监测井换位前后的检测值产生差异，忽视这种差异，可能会改变原序列的固有连续性和时空相关性，从而使预测结果产生较大误差。研究由于监测井关停及新建产生的截断式长期缺失序列的修复，是一项极具挑战的工作。

2）缺乏高效、高精度时空多序列预测建模的问题

区域地下水埋深的时域预测，是在多个监测点长期监测得到的时空序列基础上预测实现的。提高时空预测精度的关键，就是要提高各测点监测序列的预测精度和空间插值的精度。传统的预测，主要采用时间序列分析理论，建立基于神经网络和机器学习等模型的方法。这些方法不仅耗费工作量大，而且具有对多监测点、大区域的多序列时空预测适应性差，预测精度不高等问题。

3）空间插值方法存在精度不高、通用性差的问题

时空预测的最终目的是可视化地展示未来的地下水埋深空间格局。以往的空间插值方法如曲面拟合法、样条插值法、克里格插值法等虽然成功地刻画了地下水埋深的空间格局，但这些方法只考虑了已知监测点测定值对待估值的空间影响，而没有考虑历史数据对待估值的影响；时空克里格插值法虽然考虑了时空相关性，但是其关键的时空协方差函数通用性差，且精度不高。

针对地下水埋深时空序列非平稳非线性本质，将最小二乘支持向量机（least squares support vector machine，LSSVM）、广义回归神经网络（generalized regression neural network，GRNN）等非线性分析方法与新型人工智能参数优化、网格搜索（grid search，GS）、交叉验证（cross validation，CV）、自组织神经网络映射（self-organizing neural network mapping，SOM）聚类、小波消噪（wavelet noise reduction，WNK）、时空克里格（space time kriging，STK）等方法相结合，对研究区地下水埋深时空预测进行深入研究，主要的研究工作可以归纳为如下几个方面。

1）缺失数据时空修复混合模型研究

为了得到更加科学、可靠的研究区地下水埋深时空数据集，提出了基于自组织神经网络映射聚类与果蝇优化最小二乘支持向量机时空修复混合模型（SOM-FLSSVM），应用于时空序列的修复。该模型不仅考虑了每一类缺失数据监测站点的时间影响因素，而且考虑了空间因素，充分利用缺失数据监测站点与其相邻站点的信息进行数据的插补工作。实验证明，提出的时空预测混合模型相比其他的一些经典的数据插补模型，在精度方面有很大的提高。

需要全本请访问 www.ertongbook.com

2）多序列混合预测模型研究

在考虑时空序列非线性、非平稳特性的基础上，针对目标序列值会受历史数据和邻近序列时空影响的特点，建立了基于小波分析和广义回归神经网络的多序列混合预测模型。首先将每个监测序列进行小波消噪；其次建立以同期及历史若干期相关站点监测数据为输入变量的，考虑空间影响的多输入多输出时空序列GRNN模型；最后，引入网格搜索和交叉验证算法对模型参数进行优化。实验结果表明，与其他已有模型相比，本书提出的混合模型，提高了预测精度，同时也展示了小波变换在地下水埋深预测研究中的优势。应用该模型得到了十二期的地下水埋深时域预测结果。

3）空间插值混合模型研究

针对时空克里格插值中协方差函数通用性差且精度不高的问题，引入广义回归神经网络自适应拟合时空克里格插值变异函数，建立广义回归神经网络时空插值混合模型（GRNN-STK），绘制了研究区地下水埋深不同年份的空间分布图。通过分布图的对比及相应检测站点具体影响因素的分析，揭示研究区地下水埋深的空间格局演化过程。实验表明，与普通克里格插值精度相比，该模型有很大的提高，为建立研究区域可靠的地下水空间预报体系，做了重要的前提工作，能够为相关部门采取高效而可靠的地下水管理与合理开发决策，提供参考依据。

本书创新点在于针对研究中遇到的监测数据部分缺失，缺乏高效、高精度时空多序列预测模型，空间分布插值算法精度不高、通用性差等问题，提出了三个新的混合模型。鉴于地下水埋深时空序列的非线性本质，将分类算法、时空相关性、支持向量机、参数优化算法相结合，构建了时空数据修复混合模型；将消噪理论与广义回归神经网络相结合，构建了非线性多序列时域混合预测模型；将广义回归神经网络与时空克里格插值相结合，构建了时空插值混合模型。数据实验验证了模型的精度和有效性。

本书工作是对代理模型高精度时空模拟研究的探索，旨在为快捷有效地开展空间动态模拟预报提供科学方法。

全书共分七章。其中，第1章为绪论，主要是介绍地下水埋深时空预测，对民勤绿洲地下水埋深时空预测研究的目的背景、研究的意义，地下水埋深模拟的国内外研究现状和不足，地下水埋深时域预测方法研究进展，空间插值研究进展，地下水埋深时空混合建模研究进展，地下水埋深时空预测混合建模研究的内容、技术路线和组织情况。第2章为时空预测混合建模理论与方法，主要介绍本

书在区域地下水埋深时空预测研究中运用的研究方法：时空序列预测理论，自组织映射神经网络，最小二乘支持向量机，广义回归神经网络等非线性分析方法；模型参数的果蝇优化算法，网格搜索法；以及建模中重要的交叉验证，小波消噪等基本理论方法。第 3 章为研究区概况，主要介绍研究区的地理位置，气温、地形地貌、工程地质、土壤、自然灾害、社会经济和生态等基本情况，降水、蒸发、地表水、地下水资源量、水资源的利用现状、土地资源状况的水资源概况，以及研究的数据来源等。第 4 章为地下水埋深时空缺失数据修复，主要介绍自组织特征映射网络分类的果蝇优化最小二乘支持向量机时空修复模型的建模步骤及示意图；传统经典参考模型的介绍，实验对比及模型精度说明；运用所构建的模型对民勤地下水埋深时空缺失数据修补，实际修补过程及修补结果。第 5 章为民勤地下水埋深时空序列时域预测，主要介绍多序列小波消噪广义回归神经网络混合模型构建；多序列小波消噪广义回归神经网络混合模型在研究区地下水埋深时空序列预测中的应用；模型效果评价指标；模型比较；结果分析与讨论。第 6 章为研究区地下水埋深时空插值，主要介绍基于广义回归神经网络自适应拟合时空变异函数及广义回归神经网络时空插值混合模型的构建；对比模型介绍；精度比较与结果分析。第 7 章为结论与讨论，介绍主要创新点，不足与展望。

本书的研究工作受到国家重点研发计划（No. 2017YFB0504203），“万人计划”科技创新领军人才（闫浩文）项目（052-152023），兰州交通大学优秀平台支持项目（201806）和甘肃省自然科学基金项目（1610RJZA057）等资助。

本书由笔者和闫浩文教授共同组织撰写，笔者负责统稿并审定全稿。第 1 ~ 6 章由笔者撰写，第 7 章由闫浩文教授撰写。在本书的编辑、整理、校对和出版过程中，除了感谢闫浩文教授高屋建瓴的指导，还必须特别感谢兰州大学数学与统计学院李维德教授的帮助和指导。感谢民勤县水务局在数据收集方面给予的大力支持。兰州交通大学测绘与地理信息学院禄晓敏博士和兰州交通大学数理学院硕士生刘宝成、程丽娟、盛秀梅对本书初稿进行了认真阅读和文字修改，在此一并表示感谢。

由于笔者水平有限，书中不妥之处实属难免，恳请读者不吝赐教。

张仲荣

2019 年 1 月于兰州

目　录

前言
第 1 章　绪论 …… 1
1.1　研究目的和意义 …… 1
1.2　地下水时空预测 …… 2
1.3　研究进展 …… 5
1.4　研究方案及技术路线 …… 17
1.5　本章小结 …… 19
第 2 章　时空预测混合建模理论与方法 …… 20
2.1　时空序列预测理论 …… 20
2.2　人工神经网络 …… 37
2.3　支持向量机 …… 45
2.4　最小二乘支持向量机 …… 47
2.5　模型参数优化 …… 48
2.6　欧氏距离鉴别不完整序列理论 …… 58
2.7　交叉验证算法 …… 59
2.8　小波分析 …… 60
2.9　时空克里格插值 …… 63
2.10　本章小结 …… 64
第 3 章　研究区概况 …… 65
3.1　研究区基本情况 …… 66
3.2　水资源概况 …… 75
3.3　本研究应用数据说明 …… 84
3.4　本章小结 …… 88
第 4 章　地下水埋深时空缺失数据修复 …… 89
4.1　SOM-FLSSVM 模型的构建 …… 90
4.2　时空混合模型 SOM-FLSSVM 说明 …… 90
4.3　民勤地下水埋深时空缺失数据修补 …… 94
4.4　对比模型 …… 98

4.5 模型应用 …… 106
4.6 本章小结 …… 107
第 5 章 民勤地下水埋深时空序列时域预测 …… 109
5.1 M-WA-GRNN 混合模型介绍 …… 110
5.2 M-WA-GRNN 模型建立 …… 111
5.3 研究区时域预测模型的建立 …… 112
5.4 模型效果评价指标 …… 113
5.5 模型比较 …… 113
5.6 结果分析与讨论 …… 118
5.7 本章小结 …… 123
第 6 章 研究区地下水埋深时空插值 …… 124
6.1 模型构建 …… 124
6.2 GRNN-STK 模型 …… 124
6.3 对比模型 …… 126
6.4 精度比较与结果分析 …… 127
6.5 本章小结 …… 139
第 7 章 结论与讨论 …… 140
7.1 主要创新点 …… 140
7.2 不足与展望 …… 140
参考文献 …… 142
彩图

第1章 绪 论

1.1 研究目的和意义

水是人类赖以生存和发展必不可少的基础性自然资源，是自然生态环境中一个非常关键的因素。水为生态系统的物质传输和能量转换提供动力（马岚，2009）。长期以来，水资源已成为社会发展的“瓶颈”，水资源将是21世纪全世界普遍关注的战略资源。如何保障水资源、水环境和水生态安全，已成为21世纪中国乃至世界最重要的治理问题之一（王亚华，2007；魏玲玲，2014）。

20世纪80年代以来，中国开始了前所未有的社会变革和飞速的发展，人口的迅速增长及社会的变化导致对水资源的需求也日益增加（王亚华，2007）。水资源的污染、对水资源不合理的开采利用，以及城市化和工业化过程的快速推进，使人口与水资源短缺的矛盾也愈发明显。水资源的可持续发展，是国民经济和社会可持续发展的重要保障（马岚，2009）。加强水资源管制，对水资源合理开发和优化配置，对国民经济和社会可持续发展至关重要。

地下水是影响生态环境系统的一个重要因子，生态系统的自然平衡状态与地下水流量大小、埋深有很重要的关系。地下水是指储存于地球表面松散层孔隙或基岩裂隙中的水资源，地下水有分布广、水量稳定、受气候影响小、水质优良和不易受到污染等特点，是一种宝贵的自然资源。地球上水的体积大约有$1.36\times10^{9}km^{3}$，其中海洋占97.3%，而地下水仅占0.6%，尽管地下水所占比例很小，但是与人类生活有着密切的关系（张元禧、施鑫源，1998；钟华平等，2007）。

地下水已被世界上许多国家作为主要的饮用水源，特别是地表水相对贫乏的干旱、半干旱地区，地下水资源是重要的甚至是唯一的供水水源。地下水资源约占全球饮用水量的50%，占工业用水的40%，占灌溉用水的20%，我国地下水天然资源量多年平均为7837亿m^{3}，约占水资源总量的1/3，但人均水资源量也只有不到$500m^{3}$，相比国际规定的最低标准$1000m^{3}$相差甚远（张伟丽、阚燕，2011）。约占全国国土面积1/3的西北地区，地下水资源开采量每年达430亿m^{3}，开采资源量已占到全国的43%。我国约有70%的人口以地下

水为主要饮用水源，其中北方地区地下水利用率相对较高（张元禧、施鑫源，1998）。

我国西北地区地处干旱和半干旱地带，土地资源丰富，气候干燥，降水稀少，生态环境脆弱，属于水资源短缺型地区。随着城市化进程的加快及工农业的迅速发展，人类对水资源的开发量与需求量也与日俱增，西北地区水资源短缺的形势更为严峻（张斌、刘俊民，2012）。因此，建立高效、可靠的地下水空间预报体系，为地下水资源的合理开发利用提供重要的参考依据，对于地下水可持续利用是一项非常重要的研究课题。

石羊河坐落于我国西北地区河西走廊东部，是我国内陆河流域之一，也是甘肃省河西内陆河流域人口最集中、经济发展较快的地区。多年来，该流域水资源被过度开发，水资源急剧减少，严重威胁着流域区域的生态环境。目前，石羊河流域已是我国西北地区水资源开发利用程度最高，但同时也是供需矛盾最突出、生态环境问题最严重的地区（李宗礼，2007）。在石羊河源头地区，乱砍滥伐、开荒种地、超载放牧等，致使植被破坏、草场严重退化，造成水土流失，水源涵养功能急剧下降；在中游地区，工农业发展迅速，打井提水活动频繁，致使水资源被超额使用（张若琳，2006；Ma，2003a，2003b；Wang et al.，2002）。

民勤绿洲位于石羊河下游，被腾格里和巴丹吉林两大沙漠包围，自然条件恶劣，水资源匮乏，生态环境脆弱。“决不能让民勤成为第二个罗布泊”，是国务院原总理温家宝做出的批示。近年来，工农业迅速发展，水资源被大量开发和利用，致使地下水系统的功能严重衰退，生态平衡遭到破坏。生态系统失衡，则引发链条式的生态环境问题（李海涛等，2007；李小玉、肖笃宁，2005）。高效合理地利用地下水资源，以及保持生态环境和经济效益之间的统一协调是干旱、半干旱地区可持续发展的基本途径。民勤地下水位近年来的持续下降是导致该地区生态恶化的重要因素，因此，建立研究区域可靠的地下水空间预报体系，为相关部门科学管理与合理开发地下水的决策提供参考，对改善民勤县地下水生态环境有着重要的意义。

1.2　地下水时空预测

现实生活中很多问题的研究都可归结为对随时间和地点变化的时空变量的研究，如交通流量的时空变化、环境因素的时空变化、流行病的时空变化、地质灾害的时空变化、网络舆情的时空变化等（徐薇等，2004b）。近年来，随着全球

需要全本请访问 www.ertongbook.com

信息化的进一步发展，我们存储了大量来自社交媒体互动、实时传感器、地理空间信息和其他来源的数据，从而使得时空数据集的数量、体积、分辨率不断增加。时空数据由于集合了时间和空间上的数据信息，从而成为丰富的知识资源，正等待着我们进一步的挖掘。时空数据体积巨大，处理复杂，传统的统计学方法在处理这些数据集时遇到了前所未有的困难，所以寻找有效的时空数据挖掘方法具有十分重要的意义。

时空数据挖掘就是创造高效、可靠的技术，将大量的数据转化为有意义的信息知识的技术，也是处理时空数据集的主要技术（Cheng et al.，2000）。时空预测问题是时空数据挖掘的重要部分，它基于变量时间和空间相关信息的原理，运用经济学、统计学、地理学、神经网络、机器学习、人工智能等定性、定量方法，以及这些方法在引入时间和空间因素后的扩展方法，揭示具有时空特征的事物的未来发展趋势及运行规律（徐薇等，2005）。

我们的生活离不开水，而地下水是水循环过程中的一个重要组成部分，特别是在干旱和半干旱地区。因此，如何更好地对地下水进行监测并确保其可持续管理成为人们日益关注的问题（Yang et al.，2014b）。地下水埋深指标在衡量干旱、半干旱区地区地下水资源的过程中扮演着不可或缺的角色。地下水埋深时空预测就是通过一定的数学模型计算地下水未来预测面，揭示未来地下水埋深空间连续分布及其时空变化的技术方法。在研究建立地下水埋深时空预测预报体系过程中，主要存在以下几个问题。

1. 数据预处理（缺失数据修复）

由于研究区自身水文地质、自然状况及人为活动的影响，不同站点存在不同程度的监测数据的缺失。除了常见的数据缺失情形外，在地下水埋深的长期监测过程中，出现水位下降因地质因素无法获得监测数据时监测井的移位，造成长期监测序列的中断和新建监测井只有后半部分数据序列的情形，从而得到一个有数据缺失的时空数据集。完成缺失数据修复以获得完整的研究区时空数据集是开展研究工作的必要准备。

传统的修复方法处理时空数据集所需工作量大且精度不高，另外各种研究方法都有缺点，许多研究者仅仅聚焦于比较小的样本，或者仅仅研究了一个时间序列或者几个时间序列的数据修复，这使得这些方法在实际处理大的样本过程中仍存在大量烦琐的工作。另外，在许多研究中，大多数数据的缺失比率都保持在10%以下，在地下水埋深监测数据集中，由于复杂的水文条件、地理环境和人类活动的影响，一些序列的缺失率已经达到60%～80%，传统的数据修复方法难以

实现地下水数据集的高精度修复。另外，传统的数据修复方法几乎都没有考虑数据集中不同序列在时间和空间上的相关关系，从而在数据修复的过程中不能充分地利用相关序列的信息，以至于修复的精度不能很好地得到提高，并且在工作中会产生大量的时间消耗。

2. 动态模拟烦琐且精度低

确定性模型以地下水动力学为基础，根据不同的水文地质条件来确定动态模拟的数学模型，它可以反映出地下水系统的动力学机制。然而，该模型的弊端也是显而易见的。首先需设置许多假设条件，这就很难真实地模拟出实际的地下水水文地质条件，并且在获取含水层的参数信息方面需要耗费很高的成本。

地下水系统是一个复杂的动态系统。自身地理环境及人类活动等的影响，使得地下水埋深表现出复杂的非线性、多尺度和随机性特征。这给确定性模拟地下水埋深带来障碍。随机性模型是模拟地下水系统实用而有效的方法，特别是在缺乏水文地质条件资料或含水层参数资料不足的情况下，随机性模型是一个简单有效的方法，并能获得较好的精度。在以往的研究中，很多研究者往往只注重单条序列或几条序列的模型研究，传统的方法对于多维长时期观测序列处理，需要大量而烦琐的工作，而且这些模型往往在多维序列中很难保证预测精度，以至于难以开展科学研究。

3. 地下水埋深空间分布插值方法精度不高、通用性差

以往的空间插值方法有曲面拟合法、样条插值法、克里格插值法等，这些方法虽然成功地刻画出地下水埋深的空间格局，但是这些方法只考虑了已知监测点测定值对待估值的空间影响，没有考虑历史数据对待估值的时域影响。时空克里格插值虽然很好地考虑了时空相关性，但是时空协方差函数的建立是进行时空插值的关键，一般研究当中，变异函数选取模型有指数模型、线性模型、球形模型、高斯模型等；这些变异函数的实际选择是根据实验样本的半方差云图来选择最为接近其结构特征分布的变异函数模型，变异函数模型的选取是一个人为主观性抉择的过程，每种模型的参数确定又是一个比较复杂的过程，而且变异函数模型的最优选取也是一个很难解决的问题。针对变异函数的选取，有研究者提出用机器学习方法来拟合空间变异函数，但由于这些模型在建立过程中，要充分考虑自身的多个参数的影响，而自身多个参数的选择又是一个复杂而烦琐的计算过程，很大程度上影响了插值精度，给我们开展地下水埋深时空插值工作带来了很多困难。换言之，传统的分布插值方法存在精度不高、通用性差的问题。

混合预测是将不同的模型、不同的时间序列分析理论进行混合（融合）或集成（构成）一个新的统计预测模型，得到一个最终的预测结果。由于每个模型都有各自的优缺点，所以在对它们进行混合时，一个关键的问题是如何找出混合的切入点和结构，以达到取长补短的目的，所以在众多的研究当中，有学者建立了混合模型来提高模型预测的精度（车金星，2010）。许多学者已经提出了一些混合模型用于地下水埋深的研究，并获得了很好的结果，这是目前提高模型精度的前沿思路。通过建立相关的混合模型用于修复缺失数据、时空序列时域预测和空间插值，以揭示研究区地下水埋深时空变化的空间格局。

1.3 研究进展

地下水埋深时空预测的主要目的是实现地下水埋深未来时空动态的人工模拟。模拟是立足于研究区的多个监测点，以长期监测得到的监测时空序列数据集为基础，在时域对时空序列进行预测，然后在空间上插值来实现的。

本书基于有缺失的时空序列，以混合模型为主要方法，通过时空缺失数据修复，对时空序列进行时域预测，基于时空序列的时空插值来实现民勤绿洲地下水埋深未来动态的模拟研究。本节将对缺失数据修复、时域上的动态模拟、空间插值、地下水埋深时空预测及混合模型的研究进展做简要的综述。

1.3.1 时空序列缺失数据修复研究进展

准确、长期、有效的地下水埋深预测，对于开展地下水资源的开发、管理非常重要（朱长军等，2004）。科学的数据对于研究一个区域地下水的储量，以及地下水的持续开发利用是非常重要的。

由于监测设备的故障、研究区复杂的自然地理环境、人类不合理的开发利用等，长期的观测数据中产生许多缺失数据。缺失数据的存在给地下水埋深的研究带来很多困难，对有效运用相关理论进行分析研究造成障碍。因此，缺失数据的修复补全对于科学地开展地下水埋深的研究是一项非常有必要且有意义的研究工作。

数据缺失是许多科学领域广泛存在的一个问题。在前人研究中已经提出了一些用于缺失数据处理的方法，按照数据缺失的模式及缺失的机制选择相应的适合的方法，插值是用于处理缺失数据最常用的方法（Plaia and Bondì，2006）。已经

有许多研究者提出了各种时间序列缺失数据、面板缺失数据及其他形式的缺失数据处理方法（Malek et al.，2013；Young and Johnson，2015；Yozgatligil et al.，2013）。

在过去的研究中，Feng 等（2014）阐述了数学属性和逻辑属性是数据修复模型的两个重要核心关系。传统的缺失数据修复方法可粗略地划分为简单修复法（simple imputation，SI）、插值修复法（interpolation imputation，II）、回归修复法（regression imputation，RI）和近邻插值法（nearest neighbor imputation，NNI）（Chen and Shao，2011；Junninen et al.，2004；Schneider，2001）。近年来的研究显示，简单的修复方法直接将有缺失值的序列从原始数据中删除，其结果可以得到一个完整的数据集，但是失去了蕴含的一些重要信息，将对后面的数据分析结果产生影响（Cismondi et al.，2013）。插值法是运用一些插值算法，插补、替代缺失值。Linacre 用均值代替缺失值，Chen 等描述了一个简单适应随机插补方法，来减小均值或整体插值的方差（Chen et al.，2000；Linacre，1992）。Zainudin 等（2015）将三次 Berier 曲线和三次 Said-Ball 方法作为缺失值估计的工具，将插值技术进行了扩展。回归修复方法是普遍采用的方法，如果我们能够精确计算出适当的回归参数，将能获得很好的修复结果。Schneider（2001）测试了一个最大期望线性回归迭代分析算法用于数据缺失值修复。近邻插值法是很受欢迎的非参数修复算法之一，它能够弥补样本数据不能响应的信息（Chen and Shao，2011）。

这些传统的修复方法对时间序列缺失数据的修复起到了很好的作用，但是由于没有考虑缺失值空间因素的影响，修复精度难以满足时空数据集修复的精度要求。在修复工作中还会有大量的时间消耗。图 1.1 展示了时空数据集中缺失数据的修复方法。

随着互联网的发展，传感器技术的进步，大数据时代的到来，产生大量的数控数据集，如时空空气质量数据集、时空地理数据集、时空风速数据集等（Junninen et al.，2004；Poloczek et al.，2014；Yozgatligil et al.，2013）。时空缺失数据的修复研究显得异常迫切。已经有很多学者开展了时空缺失数据的修复研究（Feng et al.，2014；Poloczek et al.，2014）。有学者应用最小二乘法拟合经验正交函数，该方法也可以认为是一种优化修复方法，此外还有卡尔曼滤波法和最优平滑法（Kalman，1960；Kaplan et al.，1997；Smith et al.，1996）。2007 年，奇异谱分析和多元统计分析被用于修复几种不同类型的缺失数据值，然而，这些方法修复的精准性和可靠性很大程度上受数据缺失类型的影响（Kondrashov and Ghil，2006）。

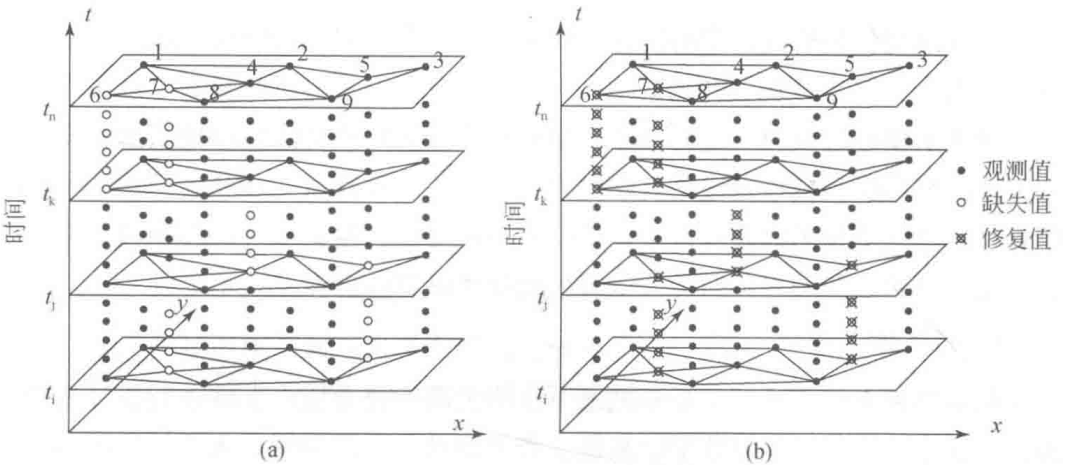


图 1.1 时空缺失数据修复示意图

(a) 数字 1~9 表示 9 个空间监测站点，四个平行四边形代表四个不同的时间点，黑圈代表数据缺失点；(b) 表示所有的缺失点数据已被修复（Zhang et al.，2017）

CUTOFF 是一种新颖的时空数据插补方法，通过与 k 近邻（k nearest neighbor，KNN）、奇异值分解、多元插补法和随机森林方法比较，体现了 CUTOFF 在时空数据集中应用的优越性。KNN 回归作为一个地理插补方法已被成功地应用到了分速时空数据集（Feng et al.，2014；Poloczek and Kramer，2014）。Carvalho 等（2016）也提出了一个时空模型并应用到了日常降雨数据的修复，并对比证明该模型优于普通克里格模型。

为了克服单一处理方法的缺点，研究者又提出了混合模型修复方法。Narravula 和 Vadlamani（2011）提出了一个新颖的软件计算法，进行两个阶段软件的计算。Aydilek 和 Arslan（2013）结合模糊 c-均值、支持向量机回归和遗传算法建立了一个混合缺失值插补模型。Tian 等（2014）在聚类灰色系统理论和熵函数基础上，首先对完整数据做分类，然后在各类中开展数据修复工作。Shukur 和 Lee（2015）提出自回归（auto regression，AR）与神经网络混合数据修复模型 AR-ANN，在风速数据集的实验中，通过比较证明了该方法的可靠性。

综合对比，混合模型是提高修复精度的重要工具。但是，由于地下水自身的特点，上述方法难以获得较高的精度，并且会产生大量的时间消耗。

1.3.2 地下水埋深模拟研究进展

地下水动态一般指地下水系统相关属性随时间和空间的变化情况。按照研究

目的建立地下水系统数学模型，模拟地下水相关属性的变化被称为动态模拟。水位与水量的变化是研究的中心内容，而地下水动态模拟是研究地下水系统演化的重要组成部分（廖梓龙，2013）。

地下水模拟是按照一定的数学模型在计算机上用数值法模拟地下水运动状态的方法，数学模型模拟方法是其研究的主要方式（薛禹群，1980）。近几十年来，随着计算机科学技术的发展及研究理论的不断完善，地下水动态模拟的研究方法已经比较成熟，我们将其归类为随机性模型和确定性模型。

1. 随机性模型

随机性地下水数学模型是指把地下水相关属性视为随机事件进行描述的数学模型。其中，组合、模糊数学、灰色、指数滑动、时间序列、人工神经网络、频谱分析和回归模型这些都是比较常用的随机模型。

Patle 等（2015）在对哈里亚纳邦区实行季风期地下水埋深时间序列建模及预测研究时，用 Mann-Kendall 非参数秩次检验测试 Sen's 斜率来鉴别在前季风和后季风时期地下水埋深的趋势，实验结果显示，1974～2010 年，研究区域的地下水埋深有显著的减小。Gong 等（2016）考虑地表水对地下水的影响，应用支持向量机、人工神经网络和自适应模糊推理系统（adaptive neuro fuzzy inference system，ANFIS）对地下水埋深进行预测，实验结果显示，这三种方法对地下水埋深近三个月的预测是有效的。Maiti 和 Tiwari（2014）对用于地下水位波动预测的按比例共轭梯度优化的人工神经网络、按比例共轭梯度优化的贝叶斯神经网络、ANFIS 这三种现代软件计算技术的优点和缺点进行了分析。Emamgholizadeh 等（2014）研究了人工神经网络、ANFIS 在伊朗 Bastam 平原地下水埋深预测中的潜力，实验结果表明，这些方法都能够准确地预测研究区地下水埋深。但是，在利用人工神经网络和多元线性回归方法对日本的 17 个站点进行预测时，将所有重要的输入变量（降雨量、环境温度、河流水位、11 个季节虚拟变量和降雨量、环境温度、河流水位、地下水埋深）的影响考虑在内，进行拟合优度分析，实验结果显示，对这 17 个站点用人工神经网络模型预测的效果要比用多元线性回归预测的效果好（Sahoo and Jha，2013）。Mogaji 等（2014）应用地理信息系统（geographic information system，GIS）有序加权回归方法，研究分析了马来西亚 Perak 州南部地下水预测模型，结果显示，在该研究区域内，地理信息系统有序加权回归方法在决策评价地下水污染的潜在映射的复杂性方面拥有巨大的潜能。Nourani 等（2015）利用聚类技术的自组织映射网络识别地下水埋深数据的空间聚类，并结合前馈神经网络对地下水埋深进行一期或多期预测。结

需要全本请访问 www.ertongbook.com