分类号_____ 密级 _____
U D C _____ 编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于黄河干流径流量数据的函数型数据分析方法研究

研 究 生 姓 名：张 悦

指导教师姓名、职称：高海燕、教授

学 科、专 业 名 称：统计学、数理统计学

研 究 方 向：复杂数据分析

提 交 日 期：2024 年 6 月 5 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：张悦 签字日期：2024年6月3日

导师签名：高海燕 签字日期：2024年6月3日

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：张悦 签字日期：2024年6月3日

导师签名：高海燕 签字日期：2024年6月3日

Research on Functional Data Analysis Method based on Runoff Data of the Main Stream of The Yellow River

Candidate : Yue Zhang

Supervisor: Haiyan Gao

摘 要

随着互联网技术的迅猛发展，产生了大量复杂连续数据。然而，传统的多元统计分析方法在处理这些数据时存在一定的局限性。因此，函数型数据分析方法应运而生。本文旨在拓展函数型数据的微分方程分析方法的应用范围，提出函数型聚类方法以及构建函数型预测模型，并基于 2002-2022 年黄河干流径流量数据开展系统的函数型数据分析方法研究，以期从函数型视角研究径流量数据，全面分析黄河流域径流演变机理，为流域生态环境保护、洪涝灾害防治以及气候变化等研究提供新的理论方法和科学依据。

首先，通过研究径流量曲线、均值曲线和方差曲线，探索了径流量的统计特征，并利用相平面图和主微分分析方法揭示了其季节变动特征。研究结果显示 2002-2022 年黄河干流径流量呈增加趋势，周期性显著，在 2004、2008、2013、2017 以及 2021 年发生了突变，且季节特征明显。

其次，基于非负矩阵分解思想，引入函数型主成分和主微分，构建基于主成分和主微分的函数型聚类方法(PCPDFCM)，探究黄河干流径流量的时空分布特征。利用函数型主成分分析发现径流量主要在 2013 年前后和 2019-2021 年显著增大。从空间分布来看，黄河干流径流量整体呈自上而下逐渐减小的趋势，空间差异性明显。此外，采用 PCPDFCM 方法将黄河干流 12 个水文站聚为三类，且三类水文站的水文特征差异显著。进一步，利用 ArcGIS 软件进行可视化展示，直观呈现 12 个水文站径流量的差异性特征。

最后，结合函数型主成分分析与主微分分析的思想原理，构建基于函数型主微分与主成分的岭回归模型(FPDPCRR)，预测黄河干流径流量。具体地，在利用累积量斜率变化分析法和多元函数型回归模型探究黄河干流径流量影响因素的基础上，采用 FPDPCRR 方法预测径流量，结果表明所提方法的预测效果较好，与实际情况较为相符，从而为流域水资源综合高效调配和防洪减灾调度提供一定的理论依据。

关键词：函数型数据分析 聚类分析 预测方法 黄河流域 实测径流量

Abstract

As the rapid development of Internet technology, a large number of complex and continuous data are produced. However, the traditional multivariate statistical analysis method has some limitations when dealing with these data. Therefore, functional data analysis method came into being. Thesis aims to expand the application range of the differential equation analysis method of functional data, put forward functional clustering method and construct a functional prediction model, the functional data analysis method of the system is studied based on the runoff data of the main stream of the Yellow River from 2002 to 2022, and the runoff data is studied from the functional point of view, comprehensively analyze the runoff evolution mechanism of the Yellow River basin, and provide new theoretical methods and scientific basis for the study of ecological environment protection, flood prevention and climate change in the basin.

Firstly, by studying the runoff curve, mean curve and variance curve, the statistical characteristics of runoff are explored, and its seasonal variation characteristics are revealed by using phase plan and principal differential analysis. The results show that the runoff of the main stream of the Yellow River shows an increasing trend from 2002 to 2022, with obvious periodicity and abrupt changes in 2004, 2008, 2013, 2017 and 2021, with obvious seasonal characteristics.

Secondly, based on the idea of non-negative matrix decomposition, functional principal component and principal differential are introduced, and a functional clustering method (PCPDFCM) combining principal component and principal differential is constructed to explore the temporal and spatial distribution characteristics of runoff in the main stream of the Yellow River. Using functional principal component analysis, it is found that the runoff mainly increased significantly around 2013 and from 2019 to 2021. From the perspective of spatial distribution, the overall runoff of the main stream of the Yellow River is gradually decreasing from top to bottom, with obvious spatial differences. In addition, 12 hydrological stations in the main stream of the Yellow River are grouped into three categories by PCPDFCM method, and the hydrological characteristics of the three types of hydrological stations are significantly different. Furthermore, ArcGIS software is used for visual display, and the difference characteristics of runoff of 12 hydrological stations are presented intuitively.

Finally, combining the principle of functional principal component analysis and principal differential analysis, a ridge regression model (FPDPCRR) based on functional principal differential and principal component analysis is constructed to predict the runoff of the main stream of the Yellow River. Specifically, on the basis of exploring the influencing factors of runoff in the main stream of the Yellow River by

using cumulant slope change analysis method and multivariate functional regression model, FPDPCRR method is used to predict runoff. The results show that the prediction effect of the proposed method is good, which is consistent with the actual situation, thus providing a certain theoretical basis for comprehensive and efficient allocation of water resources and flood control and disaster reduction dispatching in the basin.

Keywords： Functional data analysis; Cluster analysis; Forecasting method; Yellow River basin; Measured runoff

目 录

1 绪论 1

1.1 研究背景 1

1.2 研究目的及意义 2

1.3 国内外研究现状 3

1.3.1 函数型数据分析研究现状 3

1.3.2 径流演变特征研究现状 4

1.4 研究内容及技术路线 5

1.5 创新点 6

2 基础理论与方法 7

2.1 离散数据的曲线拟合 7

2.2 主微分分析 7

2.3 函数型主成分分析 9

2.4 多元函数型主成分回归 9

2.5 非负矩阵分解 11

3 基于主微分分析的黄河干流径流量季节变动特征 12

3.1 研究区概况与数据来源 12

3.1.1 研究区概况 12

3.1.2 数据来源 13

3.2 径流量的统计特征 13

3.3 基于主微分分析的径流量季节变动特征 15

3.4 径流量季节变动的刻画 20

3.5 本章小结 21

4 基于函数型聚类分析的黄河干流径流量时空分布特征 22

4.1 函数型聚类分析方法的构建 22

4.1.1 目标函数 22

4.1.2 求解算法 24

4.1.3 收敛性分析 27
4.2 基于函数型主成分分析的径流量时空分布特征 28
4.3 基于函数型聚类分析的主要水文站径流量差异性分析 31
4.4 本章小结 34
5 基于函数型岭回归模型的黄河干流径流量预测 35
5.1 函数型岭回归模型的构建 35
5.2 黄河干流径流量变化的影响因素分析 37
5.2.1 黄河干流径流量变化的影响因素 37
5.2.2 气候变化和人类活动与径流量的响应关系 42
5.3 基于函数型岭回归模型的径流量预测 45
5.4 本章小结 47
6 总结与展望 48
6.1 总结 48
6.2 展望 48
参考文献 50
攻读硕士学位期间承担的科研任务及主要成果 55
致谢 56

1 绪论

1.1 研究背景

随着互联网技术的快速发展以及数据收集技术的不断进步，大量复杂连续数据应运而生，其中一些样本数据的外在表现形式虽是离散、稀疏的片段点集，但内在结构却呈现出连续、动态的函数曲线(或曲面)特征，如不同地区的多期温度与降雨量数据、黄河流域水文站的径流量数据、生物学中的基因表达数据、以及金融学中的股价走势数据等。对于如何高效处理这些高频连续数据并探究其内在规律性，函数型数据分析方法(Functional Data Analysis，FDA)(Ramsay 和 Dalzell，1991)应运而生。FDA 可以将这些随时间变化的观测数据拟合成一条条函数曲线，运用函数曲线的研究方法剖析数据，从而揭示数据的内在结构。在传统方法中，通常要求采样点一致、采样间隔一致等条件，而函数型数据经常以不规则采样的形式存在。因此，引入函数型数据分析方法来处理这些复杂的高频连续数据是至关重要的。FDA 的研究旨在开发适用于函数型数据的统计分析方法和模型，该方法可从函数型数据中提取有用的信息，并用于数据的描述、压缩、降维、聚类、分类、回归和预测等问题。目前，FDA 已被广泛应用于多个领域，如经济学、环境学、医学、生物学、工程学等，为解决实际问题提供了有效的工具和方法。通过函数型数据分析，可以更好地理解数据的演化过程、预测未来趋势，并作出相应地决策。

典型的 FDA 包括函数型回归分析、函数型聚类分析以及函数型预测分析等。聚类是一种根据给定样本数据的相似性或距离将其归并到几个类别的数据分析方法，目的是通过所得的类来发现数据的特点，揭示数据的模式和结构。聚类在大数据时代具有重要意义，其可帮助减少数据量、提取数据特征并降低数据分析的难度。然而，由于函数型数据维度高，对其进行聚类具有一定的挑战性。事实上，非负矩阵分解(Nonnegative Matrix Factorization，NMF)是一种常见的数据降维和特征提取技术，其可提取出数据的非负和稀疏表示。基于 NMF 的函数型聚类分析方法结合了 NMF 的优势和函数型数据聚类的需求，通过 NMF 得到函数型数据的基础特征和权重，从而实现对函数型数据的聚类和模式提取。该方法能够挖掘函数型数据的隐藏模式和结构，进而揭示数据的内在规律并进行更深入的

分析。此外，函数型回归分析模型在函数型数据的预测中具有举足轻重的作用，其利用变量间的关系进行预测插补，使得模型更具解释性。

党的十八大以来，习近平总书记多次深入沿黄河流域地区视察，并研究和分析黄河流域环境保护及开发存在的困难，明确提出了黄河流域生态环境保护和高质量开发目标。2022 年 10 月 30 日，十三届全国人大常委会第三十七次会议通过《中华人民共和国黄河保护法》，为统筹推进黄河流域生态保护和高质量发展提供法制保障。由此可见，国家对黄河流域生态及发展状况极度重视，已上升到法律层面。黄河是我国的第二大河流，发源于我国西部青藏高原的巴颜喀拉山脉，自西向东分别流经黄土高原、内蒙古高原、关中平原、最后流经华北平原注入渤海。由于人类活动频繁干扰和流域气候趋势的影响，与 20 世纪五六十年代相比，20 世纪九十年代以来主要径流量已经逐渐减少。水资源短缺已经成为制约沿黄河省区社会经济快速发展的主要瓶颈，综合开发利用水资源与黄河流域径流过程高度相关。目前学者们对于区域径流演变特征及影响因素与预测的研究都是在离散数据视角下探究径流的演变机理，未能完全充分挖掘径流资料的潜在信息以及波动特征。因此，为更加全面、系统地探究黄河流域径流量数据的季节变动特征和时空分布特征，以及准确预测未来径流的变化趋势，在已有 FDA 框架下，针对黄河干流径流量数据进行系统的函数型数据分析方法研究，对黄河流域的生态环境保护、洪涝灾害防治以及气候变化等具有重要的理论价值和现实意义。

1.2 研究目的及意义

基于函数型数据分析框架，针对径流资料数据，利用函数型数据的微分方程分析方法，构建函数型聚类分析方法，克服传统离散数据分析方法未能充分挖掘数据信息的不足，探究黄河干流径流量的季节变动特征和时空分布特征。此外，构建函数型岭回归模型，并通过径流量与影响因素之间的关系来预测一定时期内的径流量。既提出了函数型数据分析的新方法，又全面揭示了径流变化的时空特征、周期性变化以及可能存在的非线性关系等信息。为黄河流域水资源管理、生态环境保护与高质量发展提供理论分析方法，以及一定的决策依据与科学指导，具有重要的理论意义和现实价值。

1.3 国内外研究现状

1.3.1 函数型数据分析研究现状

1982 年加拿大学者 Ramsay 提出函数型数据的概念，并提出针对函数型数据的分析方法，开启函数型数据的相关研究。随着现代数据收集技术的提高，所收集到的数据具有动态特征，传统的数据分析方法容易导致估计偏差增大或丢失必要信息，因此，Ramsay 和 Dalzell(1991)提出部分函数型数据分析的方法，如函数型主成分分析(Functional Principal Component Analysis，FPCA)和函数型回归分析等，并运用这些方法处理加拿大温度和降水的关系，开启 FDA 实证应用的先河。随后，Ramsay 和 Silverman(2005)、Ferraty 和 Vieu(2006)等学者发表了多本相关著作，FDA 被应用于越来越多的学科。FDA 引入国内的时间较短，目前的研究主要集中于 FDA 的应用，而在方法模型的扩展与改进的研究相对较少。张崇岐(2006)讨论了离散数据拟合函数曲线的平滑方法，且详细地阐述了平滑过程中需注意的细节。严明义(2007)以统计学的角度介绍 FDA 的核心思想，为我国 FDA 的引入与发展奠定了基础。

FDA 是一种新型的非参数统计分析方法，其从动态随机过程的视角全面考虑问题，有利于准确反映和把握实际规律。相较于传统的水文气象统计方法，FDA 方法能够更好地量化研究对象随时间变化的过程，且在探索函数型数据的波动特征以及潜在变化模式时具有一定的优越性。例如，王德青等(2021)从连续、动态的视角出发，借助 FDA 思想构建函数型金融状况指数，探究自 2002 年以来中国金融整体形势的动态变化规律；朱冉等(2023)基于 FDA 回归算法构建电阻性电路电弧放电数学模型，并建立相应的本质安全能量依据。FDA 包括函数型主成分分析、主微分分析(Principal Differential Analysis，PDA)(Ramsay，1996)、函数型聚类分析(王德青等，2018)以及函数型回归分析(丁辉等，2018)等方法。函数型数据聚类分析是将函数对象划分为多个类，使得类内的对象具有相似的曲线变化模式，类间的对象具有不同的曲线变化模式。非负矩阵分解(Lee 和 Seung，1999)也应用于函数型聚类过程中，如 Dijana 等(2018)提出基于 NMF 的子空间聚类方法，高海燕等(2020)提出了基于 NMF 的函数型聚类方法。此外，函数型回归模型在预测问题中得到广泛应用，如苏栝芳等(2022)提出基于残差函数主成分的估

计方法预测股市开盘价，为高频经济金融数据的处理分析提供新视角，Oshinubi等(2022)基于 FPCA 分析法国 COVID-19 数据，并建立函数型线性回归模型来预测死亡人数。

1.3.2 径流演变特征研究现状

黄河流域径流变化的特征是水文研究领域中备受关注的热点问题，通过探究径流的变化规律和趋势，为提高水资源的利用效率以及促进黄河省区社会经济高速发展提供政策与建议。马柱国(2005)的研究表明，自 20 世纪 80 年代以来，黄河流域的径流量呈下降趋势，且年际变化趋势显著；Liu 等(2020)发现近几十年来黄河径流量和泥沙量急剧下降；苏贤保等(2021)通过灵活样本熵测度黄河上游径流复杂度变化特征，发现大多水文站的年、汛期和非汛期的径流量下降趋势显著。对于定量区分气候条件变化和人类活动对径流的影响程度，有累积量斜率变化率分析法、降水径流关系法、T-S 框架、SWAT 等方法。例如 Lei 等(2021)对潘阳湖流域进行趋势和跳跃变异检测，且采用线性回归模型对流域的年径流量和汛期、非汛期流量进行归因分析；张亚丽等(2022)运用累积斜率变化率比较法定量探究了广西北部湾入海流域径流的影响因素。此外，受众多因素的相互作用及影响，径流序列具有显著的非平稳、高维度和模糊性等复杂特性(梁浩等，2020；黄亚等，2020)，因此如何获得预测精度高且稳定的预测模型仍是当前水文水资源领域的研究重点。目前，国内外学者关于径流预测的方法有过程驱动模型和数据驱动模型。过程驱动模型有新安江模型和 SWAT 模型等，其是通过模拟产流过程和河流演变过程来预测径流过程的模型(王文和马俊，2005)，具有实际的物理意义。例如 Tucci 等(2003)基于分布式水文模型，利用降水数据预测未来几个月的河道流量；闻昕等(2022)通过建立基于多因素相似度的融雪径流预测模型，实现了 7 天预测期内日径流的滚动预测。而依托于统计学或其他人工智能技术的数据驱动模型通过对时间序列的分析建立更完备的模型，使得其在水文资料缺失或复杂非线性情况下具有更好的适用性。例如 Xie 等(2021)构建了长短期记忆神经网络，并利用 531 个流域的样本进行训练，模型精度显著提高；胡作龙和高鹏(2023)采用集合经验模态分解-支持向量机耦合模型预测吴旗站的月径流量，结果表明模型预测精度较高。

综上所述，目前有关函数型数据分析的研究主要集中在其实证应用方面，然而，与实证应用相比，关于函数型数据分析方法模型拓展和理论研究的数量相对较少。此外，对于区域径流演变特征及影响因素与预测的研究均采用离散数据分析方法，未能系统、充分地挖掘径流资料的潜在信息以及波动特征。因此，从函数型数据视角出发，探究一套系统完备的函数型数据分析方法，并将其应用于径流变化的内在规律及影响因素与预测研究中，既提出了函数型数据分析的新方法，又全面揭示了径流变化的时空特征、周期性变化以及可能存在的非线性关系等信息，为进一步解释径流变化机理提供新的思路和研究方法。

1.4 研究内容及技术路线

本文旨在拓展函数型数据的微分方程分析方法的应用范围，并构建函数型聚类分析方法和函数型岭回归模型，对黄河干流径流量进行深入探究。本文总共分为六部分，各部分的内容安排如下：

第一部分为绪论。简要介绍研究背景、目的、意义、现状、内容以及创新性等。

第二部分为基础理论与方法。对离散数据的曲线拟合、主微分分析、函数型主成分分析、多元函数型主成分回归和非负矩阵分解的基本思想以及原理进行说明。

第三部分为基于主微分分析的黄河干流径流量季节变动特征。首先通过径流量曲线、均值函数曲线和方差函数曲线探究径流量的统计特征；其次利用相平面图和主微分分析对其季节变动特征进行分析；最后利用傅里叶基函数刻画其季节变动。

第四部分为基于函数型聚类分析的黄河干流径流量时空分布特征。首先在非负矩阵分解框架下，引入函数型主成分和主微分，构建基于主成分和主微分的函数型聚类方法；其次利用函数型主成分分析探究径流量的时空分布特征；最后基于所提函数型聚类模型，对黄河干流 12 个主要水文站进行差异性分析。

第五部分为基于函数型岭回归模型的黄河干流径流量预测。首先结合函数型主成分分析与主微分分析的思想原理，考虑数据曲线及波动特征两个视角，构建基于函数型主微分与主成分的岭回归模型；其次通过累积量斜率变化分析法和函

数型线性回归模型分析黄河干流径流量变化的影响因素；最后利用所提函数型岭回归模型预测一定时期内径流量。

第六部分为本研究的结论与展望，对本文主要内容进行分析总结，并梳理未来可开展的研究工作。

本文的技术路线图如图 1.1 所示。

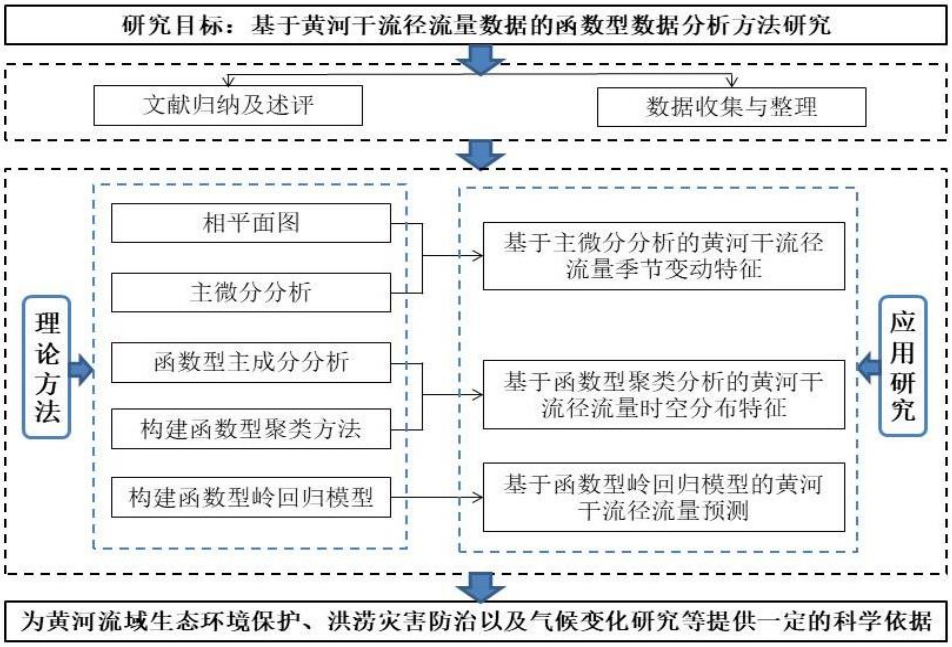


图 1.1 技术路线图

1.5 创新点

针对传统离散数据分析方法未能充分挖掘数据信息的问题，本文在函数型数据分析框架下，基于黄河干流径流量数据，拓宽主微分分析方法的应用，并提出基于主成分和主微分的函数型聚类方法和基于函数型主微分与主成分的岭回归模型，探究一套系统完备的函数型数据分析方法。主要创新点如下：

（1）函数型主成分分析旨在提取原始数据的绝大多数信息，主微分分析旨在提取原始数据的波动特征。本文在非负矩阵分解框架下，引入函数型主成分和主微分分析，构建基于主成分和主微分的函数型聚类方法，更加科学地探究了黄河干流径流量的时空分布特征；

（2）针对响应变量存在缺失的函数型径流量数据，基于函数型主成分分析和主微分分析的思想，构建基于函数型主微分与主成分的岭回归模型，有效地预

测一定时期的径流量，且该模型的解释性较好；

（3）从实际应用角度，本文从函数型数据新视角出发，利用较为系统完备的函数型数据分析方法揭示了黄河干流径流演变特征，拓展了函数型数据分析方法的应用。

2 基础理论与方法

2.1 离散数据的曲线拟合

函数型数据分析的关键是在离散时间点$\{t_j\}$观察一个连续可微的过程

$$y_j = x(t_j) + \varepsilon_j, \tag{2.1}$$

其中，$\{y_j\}$为观测值序列，$x(t)$为潜在可微连续函数，ε_j为不可观测的误差成分。为保证拟合函数$x(t)$精确表示观测值$\{y_j\}$，通常使用最小二乘准则刻画拟合函数的准确性；同时也要求拟合函数是匀滑的，因此将导数平方的积分作为粗糙惩罚项，以刻画函数的光滑程度。因此，考虑以下拟合方程：

$$\begin{aligned} \mathrm{PENSSE}_{\lambda}(x \mid y) &= \sum_{j} \left[y_j - x\left(t_j\right)\right]^2 + \lambda \,\mathrm{PEN}_k(x), \\ \mathrm{PEN}_k(x) &= \int \left\{D^k x(s)\right\}^2 \,\mathrm{d}s = \left\|D^k x\right\|^2, \end{aligned} \tag{2.2}$$

其中，D^k表示k阶导数，λ为修匀参数，可通过广义交叉验证准则(Generalized Cross-Validation，GCV)得到。通常通过最小化式(2.2)，可获得拟合函数$x(t)$。

2.2 主微分分析

在实际问题中，为得到更为准确的拟合函数，利用某些具有特殊结构的函数构造粗糙惩罚项。对于函数$x(t)$，假设存在线性微分算子

$$L = \beta_0 I + \cdots + \beta_{m-1} D^{m-1} + D^m. \tag{2.3}$$

于是，相应的微分方程为

$$D^m x = -\beta_0(t)x - \beta_1(t)Dx - \cdots - \beta_{m-1}(t)D^{m-1}x, \tag{2.4}$$

其中，I为单位算子，β_j为权重系数函数，可以是常数或时间的函数。式(2.4)也可表示为零化的线性微分算子$Lx = 0$，其可分析径流量的变化模式。为求解最优拟合函数，将Lx平方的积分代替式(2.2)中的$\mathrm{PEN}_k(x)$，即最小化式(2.5)来求解拟

合函数$x(t)$

$$\mathrm{PENSSE}_{\lambda}(x \mid y)=\sum_{j}\left[y_{j}-x\left(t_{j}\right)\right]^{2}+\lambda \int[Lx(t)]^{2}dt. \tag{2.5}$$

在水文数据的研究中，从输入输出的视角出发，以探寻水文系统的内在运行规律。设函数观测对象$x_i, i=1,\cdots,N$及与其对应的协变量函数$f_i, i=1,\cdots,N$，通过x_i和f_i来确定下式

$$\begin{aligned} Lx &= \beta_0(t)x+\cdots+\beta_{m-1}(t)D^{m-1}x+D^{m}x \\ &= f(t). \end{aligned} \tag{2.6}$$

当系统不存在强迫函数时，即$f(t)=0$，此时目的在于寻求一个线性微分算子以满足线性微分方程$Lx_i=0, i=1,\cdots,N$。为此，需要估计m个权重系数函数$\beta_j, j=0,\cdots,m-1$，从而得到微分算子L。若要求微分算子L满足$Lx_i=0$，则可将Lx_i看作下式的残差

$$D^{m}x_i=-\beta_0(t)x_i-\beta_1(t)Dx_i-\cdots-\beta_{m-1}(t)D^{m-1}x_i,\ i=1,\cdots,N. \tag{2.7}$$

基于残差函数的范数平方和，构建如下最小二乘准则

$$\begin{aligned} \mathrm{SSE}_{PDA}(L \mid x) &= \sum_{i=1}^{N}\int[Lx_i(t)]^{2}dt \\ &= \sum_{i=1}^{N}\|Lx_i\|^{2}. \end{aligned} \tag{2.8}$$

当系统的强迫函数f_i与输出函数x_i同时被观测到时，求解非齐次方程$Lx_i=f_i, i=1,\cdots,N$。此时，最小二乘准则变为

$$\begin{aligned} \mathrm{SSE}_{PDA}(L \mid x,f) &= \sum_{i=1}^{N}\int[Lx_i(t)-f_i(t)]^{2}dt \\ &= \sum_{i=1}^{N}\|Lx_i-f_i\|^{2}. \end{aligned} \tag{2.9}$$

为研究方便起见，本文仅考虑齐次情况，即强迫函数为 0 的情形。给定N个函数型数据的观测值$x_i(t), i=1,\cdots,N$，利用式(2.8)估计系数$\beta_j(t)$，通过逐点最小化法，得到其最小二乘解为

$$\hat{\beta}(t)=[Z(t)'Z(t)]^{-1}Z(t)'\omega(t), \tag{2.10}$$

式(2.10)中$\beta(t)=(\beta_0(t),\cdots,\beta_{m-1}(t))'$，$\omega(t)=(\omega_1(t),\cdots,\omega_N(t))'$，其第$i$行为$\omega_i(t)=D^{m}x_i(t)$，$Z(t)$为$N\times m$阶矩阵，其第$i$行为$z_i(t)=\{-x_i(t),\cdots,-D^{m}x_i(t)\}$。关于主微分分析更详细的介绍可参看文献(Ramsay 和 Silverman，2005)。

2.3 函数型主成分分析

FPCA 的基本思想是将函数型数据表示为一组基函数的线性组合，通过线性投影将高维的函数型数据降维到低维空间中。具体而言，FPCA 通过求解特征函数和特征值，找到函数空间中最重要的成分，即主成分，以解释数据中最显著的变异性。而在传统多元统计分析中，主成分分析是利用数据的样本方差-协方差矩阵的特征值进行分解，以寻找数据的最大变化方向，该方向由主成分的系数构成的向量描述，称为主成分的权向量。在 FDA 中，类似于多元统计中的主成分权向量，相对应的是主成分权函数，记为$f(s)$，其中s在一个区间T中变化，且$f(s)$平方可积。考虑$L^2(T)$中的随机函数X用于描述各函数型数据变动轨迹，其中$L^2(T)$表示封闭时间间隔上平方可积函数的希尔伯特空间，则第i个样本$x_i(s), i=1,2,\cdots,n$(经过中心化处理)的主成分得分定义为

$$\xi_i = \int_T f(t)x_i(t)dt,\ i=1,2,\cdots,n.$$

类似于多元统计分析中主成分的研究思路，第一主成分权函数$f_1(s)$通过求解如下优化问题得到

$$\begin{aligned} &\max\ \frac{1}{n}\sum_{i=1}^{n}\xi_{i1}^2, \\ &\text{s.t.}\ \int_T f_1(t)^2dt = 1, \end{aligned}$$

其中，$\xi_{i1}=\int_T f_1(t)x_i(t)dt$为第$i$个样本曲线$x_i(s)$的第一主成分得分。

同理，可以求得第j个函数主成分，其权函数$f_j(s)$满足如下优化问题：

$$\begin{aligned} &\max\ \frac{1}{n}\sum_{i=1}^{n}\xi_{ij}^2, \\ &\text{s.t.}\ \int_T f_j(t)^2dt = 1, \\ &\int_T f_j(t)f_1(t)dt = \cdots = \int_T f_j(t)f_{j-1}(t)dt, \end{aligned}$$

其中，$\xi_{ij}=\int_T f_j(t)x_i(t)dt$为第$i$个样本曲线$x_i(s)$的第$j$主成分得分。

2.4 多元函数型主成分回归

多元函数型线性回归模型是利用J个函数型预测变量$X=(X_1,\cdots,X_J)'$来

估计函数型响应Y。例如一个来自(X,Y)的随机样本$\{(x_i,y_i):i=1,\cdots,n\}$，其中$x_i=(x_{i1},\cdots,x_{iJ})'$，假设所有函数变量都在区间$T$上平方可积函数的希尔伯特空间$L^2(T)$上取值，通常内积定义为$< f,g >=\int_T f(t)g(t)dt,\forall t\in T$。函数型线性回归模型表示为

$$y_i(t)=\alpha(t)+\sum_{j=1}^{J}\int_T x_{ij}(s)\beta_j(s,t)ds+\varepsilon_i(t),\ i=1,\cdots,n, \tag{2.11}$$

其中，$\alpha(t)$为截距函数，$\beta_j(s,t)$为J个系数函数，$\varepsilon_i(t)$是独立的误差函数。式(2.11)可写为矩阵形式

$$y_i(t)=\alpha(t)+\int_T x_i(s)'\beta(s,t)ds+\varepsilon_i(t),\ i=1,\cdots,n,$$

其中，$x_i(s)=(x_{i1}(s),\cdots,x_{iJ}(s))'$，$\beta(s,t)=(\beta_1(s,t),\cdots,\beta_J(s,t))'$。

由于多元函数型线性回归模型易受多重共线性的影响，从而导致参数估计精度下降，因此 Acal(2021)提出了主成分多元函数型线性回归模型。函数型预测因子和函数型响应的主成分分解分别为

$$\begin{aligned} x_{ij}(t)&=\bar{x}_j(t)+\sum_{l=1}^{n-1}\xi_{il}^{x_j}(t)f_l^{x_j}(t),\\ y_i(t)&=\bar{y}(t)+\sum_{l=1}^{n-1}\xi_{il}^{y}(t)f_l^{y}(t), \end{aligned} \tag{2.12}$$

其中，权重函数$f_l^{x_j}$和f_l^y分别是$x_{ij}(t)$和$y_i(t)$的样本协方差算子的特征函数。主成分得分$\xi_{il}(t)$是中心化的不相关标量变量，其最大方差由与其权重函数相关的特征值$Var(\xi_{il}^{x_j})=\lambda_l^{x_j}$，$Var(\xi_{il}^{y})=\lambda_l^{y}$给出。结合式(2.12)的主成分分解，式(2.11)可化为下面线性回归模型

$$\xi_{ik}^{y}=\sum_{j=1}^{J}\sum_{l=1}^{n-1}b_{1kl}^{x_j}\xi_{il}^{x_j}+\epsilon_{ik},\ i=1,\cdots,n;\ k=1,\cdots,n-1. \tag{2.13}$$

通过截断每个主成分分解，得到以下函数型响应的多元函数型主成分回归模型(Multivariate Functional Principal Component Regression Model，MFPCR)

$$\begin{aligned} \hat{y}_i(t)&=\bar{y}(t)+\sum_{k=1}^{K}\hat{\xi}_{ik}^{y}f_k^{y}(t)\\ &=\bar{y}(t)+\sum_{k=1}^{K}(\sum_{j=1}^{J}\sum_{l\in L_{kj}}\hat{b}_{1kl}^{x_j}\xi_{il}^{x_j})f_k^{y}(t), \end{aligned} \tag{2.14}$$

其中，$\hat{b}_{1kl}^{x_j}$为回归系数b_{1kl}的线性最小二乘估计。

2.5 非负矩阵分解

非负矩阵分解(NMF)是一种应用广泛的矩阵因式分解方法。NMF 既能利用基向量的线性组合表示原始矩阵中所呈现的样本特征，还可以实现高维数据的有效降维。NMF 的目标是寻找两个非负的低秩矩阵，通过这两个矩阵的乘积来近似原始数据，其基本原理如图 2.1 所示。

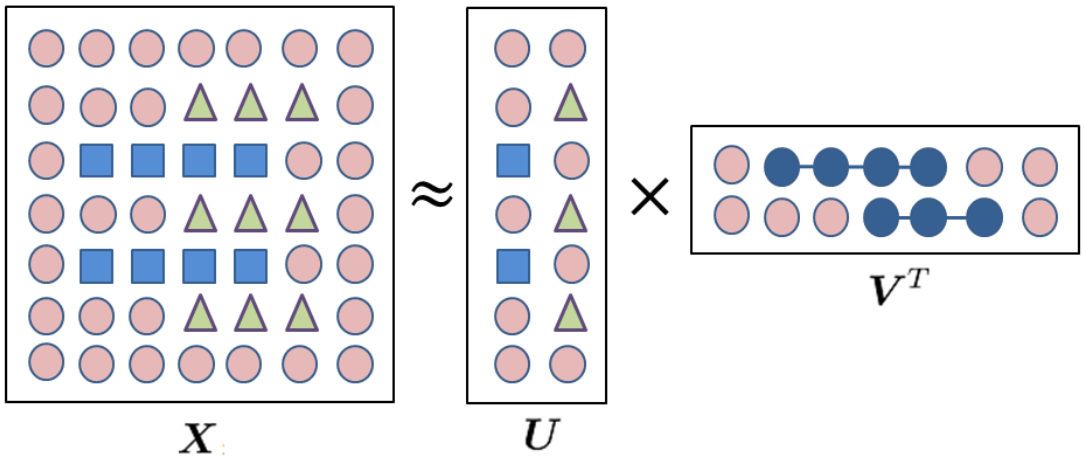


图 2.1 非负矩阵分解的基本原理

对于给定的非负数据矩阵$\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\in\mathbb{R}^{d\times n}$，其中$\boldsymbol{x}_i\in\mathbb{R}^d, i=1,\cdots,n$为第$i$个样本列向量。NMF 利用两个低秩非负矩阵的乘积近似$\boldsymbol{X}$，其目标函数为

$$\min_{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0}\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}^T\|_F^2=\sum_{j=1}^{n}\|(\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}_j^T)\|_2^2, \tag{2.15}$$

其中，$\boldsymbol{U}=(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)\in\mathbb{R}^{d\times k}$为基矩阵，$\boldsymbol{V}=(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n)\in\mathbb{R}^{n\times k}$为系数矩阵。通过乘法迭代，得到下面迭代更新算法

$$\begin{aligned}\boldsymbol{U}_{ij}&\leftarrow\boldsymbol{U}_{ij}\frac{(\boldsymbol{X}\boldsymbol{V})_{ij}}{(\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{V})_{ij}},\\ \boldsymbol{V}_{ij}&\leftarrow\boldsymbol{V}_{ij}\frac{(\boldsymbol{U}^T\boldsymbol{X})_{ij}}{(\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{V}^T)_{ij}}.\end{aligned} \tag{2.16}$$

此外，NMF 也具有聚类的特性，在聚类过程中，基矩阵$\boldsymbol{U}$中的各个基向量都代表了某一个指定类的最重要的特征。

3 基于主微分分析的黄河干流径流量季节变动特征

3.1 研究区概况与数据来源

3.1.1 研究区概况

黄河是中国第二长河流，其发源于青海省巴颜喀拉山脉，流经青海、四川、甘肃等 9 个省份，最终注入渤海，流域面积为$79.5 \times 104km^2$。在黄河干流上设置有兰州站、秦安站、梁家河站等多个水文站，用于监测和记录黄河干流的水位、流量、水温、水质等水文数据。由于地形复杂，黄河流域各水资源区的气候条件存在显著差异，不同区域的水文特征、径流规律、水资源状况等均有所不同。图 3.1 展示了黄河干流的地理位置及主要水文站的分布情况，各水文站基本情况如表 3.1 所示。

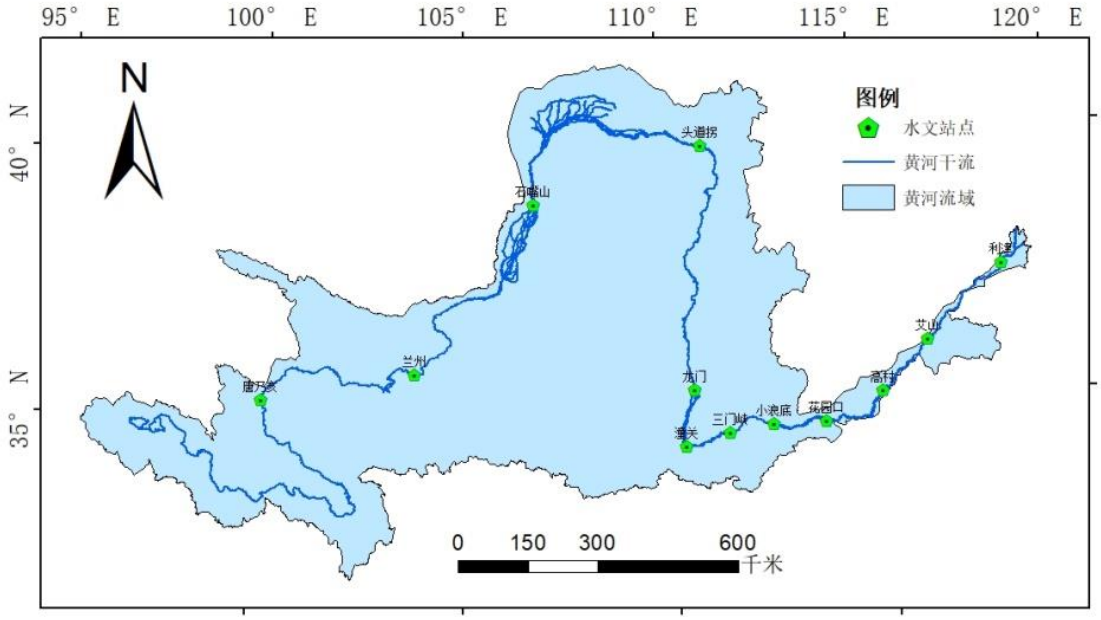


图 3.1 研究区地理位置及水文站分布

表 3.1 黄河干流主要水文站基本情况

区间	水文站	控制面积/10^4km^2	距河口距离/10^2km	经度	纬度
黄河上游	唐乃亥	12.20	39.11	$100^\circ09'$	$35^\circ30'$
	兰州	22.26	33.45	$103^\circ49'$	$36^\circ04'$
	石嘴山	30.91	26.65	$106^\circ47'$	$39^\circ15'$
	头道拐	36.79	20.02	$111^\circ04'$	$40^\circ16'$

续表 3.1 黄河干流主要水文站基本情况

区间	水文站	控制面积/10^4km^2	距河口距离/10^2km	经度	纬度
黄河中游	龙门	49.76	12.69	$110^{\circ}35'$	$35^{\circ}40'$
	潼关	68.21	11.38	$110^{\circ}18'$	$34^{\circ}37'$
	三门峡	68.84	9.50	$111^{\circ}25'$	$34^{\circ}51'$
	小浪底	69.42	8.14	$112^{\circ}41'$	$34^{\circ}92'$
	花园口	73.00	6.82	$113^{\circ}40'$	$34^{\circ}54'$
黄河下游	高村	73.41	5.13	$115^{\circ}05'$	$35^{\circ}26'$
	艾山	73.91	3.40	$116^{\circ}18'$	$36^{\circ}16'$
	利津	75.19	0.79	$118^{\circ}18'$	$37^{\circ}32'$

3.1.2 数据来源

考虑黄河干流水文站分布情况以及数据的准确性和完整性因素，本文沿程自上而下选取 2002-2022 年黄河流域干流具有代表性的 12 个水文站，即唐乃亥、兰州、石嘴山、头道拐、龙门、潼关、三门峡、小浪底、花园口、高村、艾山、利津，为探究黄河流域径流的变化特征及其与气候变化和人类活动之间的响应关系提供新思路和新方法。所用实测径流量数据和降水量数据来源于黄河水利委员会[①]，气温数据来源于欧盟及欧洲中期天气预报中心等组织发布的 ERA5-Land 数据集[②]，论文中所用地图均来源于中国科学院资源环境科学与数据中心[③]。

3.2 径流量的统计特征

为探究黄河干流实测径流量的变化特征，首先对 2002-2022 年 12 个水文站 21 年的实测径流量序列进行曲线拟合并修匀。本文将各年的每个月份看作节点，采用 255 条 5 阶 B-样条基函数来拟合函数曲线。同时，选取修匀参数$\lambda = 0.1$对拟合曲线进行修匀处理。图 3.2 为匀滑的黄河干流实测径流量拟合曲线图，其中实线为 12 个水文站的实测径流量曲线，虚线为趋势线，该图反映了实测径流量随时间变化的趋势及重要特征。可以看出，在 2002-2022 年间，12 个水文站实测径流量虽存在显著差异，但总体上变化趋势相似，呈上升趋势。具体地，实测径流量的变化并非始终上升，而是先上升后下降、再上升再下降的循环上升过程，

[①]http://www.yrcc.gov.cn/
[②]https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-landmonthly-means?tab=overview
[③]https://www.resdc.cn/

其中上升幅度大于下降幅度，具有一定的周期性。因此，黄河干流实测径流量具有明显的趋势性和周期性。

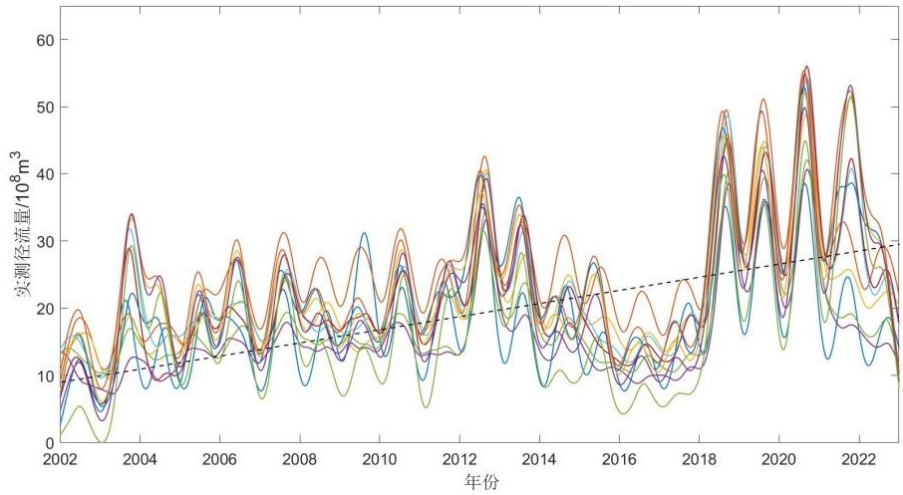
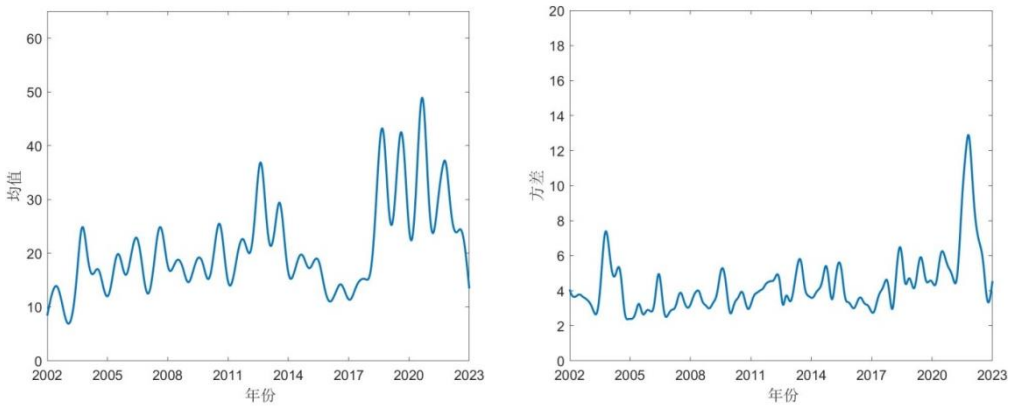


图 3.2　2002-2022 年黄河干流 12 个水文站实测径流量拟合曲线



(a)　　(b)

图 3.3　实测径流量拟合曲线的均值、方差函数曲线

为进一步探究黄河干流实测径流量的统计特征，绘制实测径流量拟合曲线的均值函数和方差函数曲线，如图 3.3 所示，可以看出，实测径流量的均值和方差均具波动增加趋势，且方差的变化相对平稳。在图 3.3(a)中，可将 2002-2022 年黄河干流实测径流量大致分为三个阶段，第一阶段为 2002-2012 年，此阶段实测径流量均值波动幅度较小，呈增加趋势；第二阶段为 2013-2017 年，此阶段实测径流量均值处于大幅度的减少状态；2018-2022 年为第三阶段，实测径流量均值在这一阶段波动幅度较大，且此阶段径流量均值明显高于前两个阶段。同时，实测径流量均值函数于 2002-2008 年、2009-2016 年以及 2017-2022 年三个时间段呈先上升后下降趋势，具有显著的周期性，变化周期约 5-7 年。此外，实测径流

量在 2004、2008、2013、2017 和 2021 年前后发生明显的升降变化，说明实测径流量曲线在这些时间点发生突变。图 3.3(b)中实测径流量的方差函数呈现波动状态，且其在样本区间内具有增加趋势，这说明近年来受全球气候变暖和人类活动的影响，黄河流域径流变化愈发频繁。

3.3 基于主微分分析的径流量季节变动特征

由黄河干流实测径流量的统计特征知，实测径流量的趋势性、突变性和周期性显著。进一步，基于函数型数据视角，探究实测径流量函数曲线的导数信息，以获得实测径流量变化更多的潜在信息。由于图 3.2 中实测径流量序列的拟合曲线较为光滑，故存在一阶导数和二阶导数，分别代表实测径流量序列的速度、加速度变化曲线。从图 3.4 可以看出，实测径流量曲线的速度和加速度曲线基本处于规则的波动状态，且加速度变化曲线波动幅度较大。在 2002-2014 年，实测径流量的速度和加速度变化趋势基本一致，均处于较小幅度的上下波动状态；2015-2017 年期间，实测径流量的速度与加速度均处于基本平缓的波动状态；2018 年之后，实测径流量的速度与加速度具有较大幅度的波动。

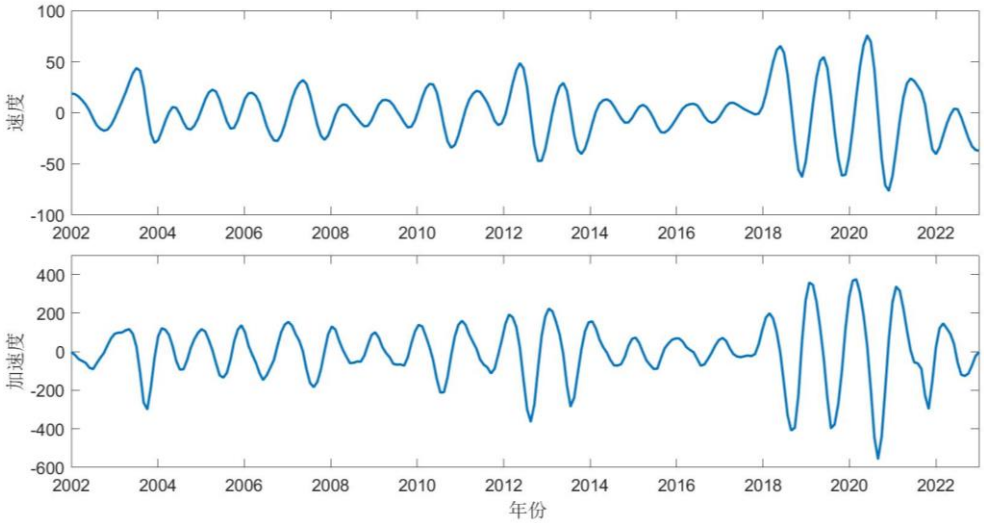


图 3.4 实测径流量拟合曲线的一阶、二阶导函数

接下来探索黄河干流实测径流量序列的动态变化模式。首先通过绘制 2002 年、2012 年、2022 年和突变年份实测径流量的相平面图，探究实测径流量变化速度与加速度的关系、动能与势能交替变化的规律。在图 3.5 中，一年中的 1-12 月分别用 1、2、3、4、5、6、7、8、9、O、N、D 表示，两条相互垂直的虚线

的交点为绝对零点，此点处实测径流量的速度和加速度都为零，意味着实测径流量在此点处运动的动能和势能为零。曲线越靠近绝对零点，实测径流量曲线波动的持久性和变化幅度就越弱。由图 3.5 可以看出，各个年份实测径流量的相平面图基本都呈圆圈状，说明样本空间内黄河干流实测径流量具有上升、下降再上升的波动周期，季节特征显著。

图 3.5(a)中，2002 年实测径流量的相平面图为一个不闭合的圈，实测径流量速度自 1 月起开始下降，至 9 月中旬达到最小值，之后持续增大。2002 年 1-7 月实测径流量变化的速度为正、加速度为负，且 7 月势能最小，说明该段时间实测径流量处于缓慢的增长状态，为其春、夏季变化特点。自 8 月开始实测径流量速度变为负值，至 9 月上旬达速度最小值，为秋季变化特征。冬季实测径流量曲线的加速度为正。因此，2022 年实测径流量具有春、夏、秋、冬季不同的变化特征。此外，2002 年、2004 年、2012 年以及 2022 年中四个季节分别分布于以绝对零点为中心的四个不同的象限中，且各个象限中速度、加速度具有相应的增加或减少趋势。这意味着大多年份实测径流量的变动具有较强的季节性特征。

图 3.5(c)中，2008 年实测径流量的相平面图由一个半圈和一个小圈组成，半圈自 1 月份开始，途径 2、3 月，至 4 月中旬结束，在此阶段实测径流量加速度处于基本不变状态，意味着实测径流量在 2008 年春季变化幅度较小。自 4 月下旬起速度增加、加速度减小，至 6 月下旬速度达最大值、加速度为 0，此时对应于 2008 年实测径流量变化曲线的“波峰”位置，之后实测径流量速度持续下降，至年末稍有回升。进一步，2017 年实测径流量的相平面图与 2008 年类似，意味着 2008 年和 2017 年实测径流量变化特征相似。

由图 3.5 (e)和图 3.5 (g)可知，2013 年和 2021 年实测径流量的相平面图基本位于垂直虚线的左侧，说明这两年实测径流量曲线处于减少趋势。进一步，以图 3.5 (e)为例，探索这两年实测径流量曲线的动态变化特征。将两条垂直虚线看作平面直角坐标系，1-5 月位于第二象限，速度为负、加速度为正且逐渐减小，这说明实测径流量在春、夏季呈下降趋势；6-10 月位于第三象限，速度、加速度均为负，且自 6 月起运动轨迹远离绝对零点，因此 6 月开始实测径流量变化幅度增大。10-12 月相平面曲线又回到第二象限。因此，2013 年和 2021 年实测径流量的季节变动特征也较为明显。

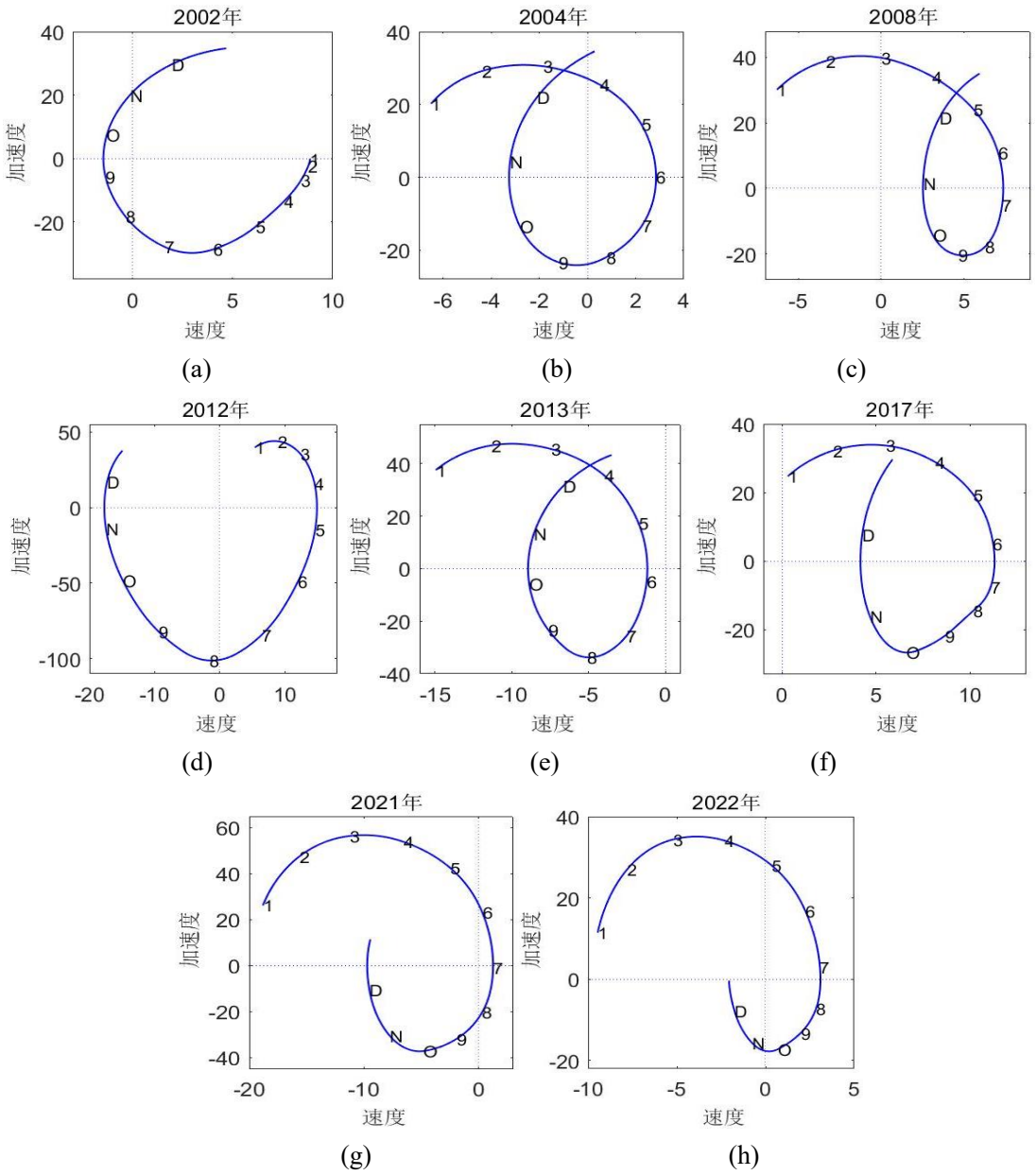


图 3.5 相平面图

综上可知，相平面图可视化了黄河干流实测径流量随时间变化的季节变动特征。受地形影响，黄河流域的气候差异较大，在兰州水文站以上，由于黄河主要产流区位于青藏高原的湿润、半湿润、半干旱地区，因此黄河径流主要来源于降水和冰雪融水；在兰州至龙门水文站，该区域降水量少、蒸发大，导致断面径流较小；而作为亚湿润区的龙门至花园口区域，降水量较大，进而径流量也较大。因此，黄河流域实测径流量的季节性变化显著，由于夏秋降雨等因素，易导致洪灾，而春冬季气候干燥寒冷，造成水源匮乏，径流年内分配极不均匀。根据水利部黄河水利委员会颁布的《黄河水资源公报》显示，与 1956-2000 年相比，近 21

年来黄河干流主要水文站实测径流量的年均值偏小。然而，在 2002-2022 年间，由于降雨量增多，使得黄河流域径流量总体上呈增多趋势。

下面对黄河干流实测径流量数据进行主微分分析，选取三阶线性微分方程

$$Lx_i = \beta_0(t)x_i + \beta_1(t)Dx_i + \beta_2(t)D^2x_i + D^3x_i = 0, \tag{3.1}$$

其中，弹性系数$\beta_0(t)$反映了在位置x处施加于系统的位置相关力；$\beta_1(t)$与速度成正比，反映整个系统的速度；$\beta_2(t)$与加速度成正比，反映整个系统的加速度(剡亮亮，2013)。

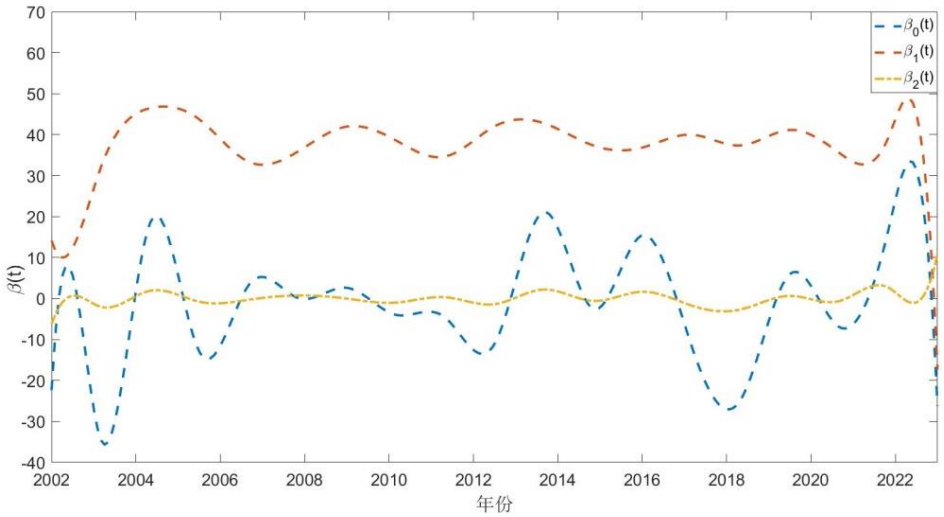


图 3.6 权重函数的图像

图 3.6 为权重函数的变化曲线。可以看出，权重函数$\beta_0(t)$、$\beta_1(t)$和$\beta_2(t)$均具一定的波动性，且$\beta_0(t)$的变化幅度最大，$\beta_2(t)$的变化相对平缓。2002 年以来$\beta_0(t)$和$\beta_2(t)$均以 0 为中心上下波动，$\beta_1(t)$在样本区间内基本全为正值，且$\beta_1(t)$与系统的速度成正比，这说明自改革开放以来，政府高度关注黄河流域的保护与发展并采取措施，如建设龙羊峡、小浪底等水利枢纽工程，推进退耕还林还草、植树造林、河道治理、汛期监测预警等，有效地解决了河道萎缩和黄河断流的难题，黄河流域的生态情况也明显好转。此外，通过$\beta_0(t)$、$\beta_1(t)$和$\beta_2(t)$的值可以得到任意时间黄河干流实测径流量所满足的微分方程，并通过线性微分算子的表达式可以直观观测到实测径流量的位置，即相应梯度的变化。

下面绘制 12 个水文站实测径流量的观测值与拟合曲线的对比图，验证 PDA 方法的有效性。图 3.7 为基于微分方程的拟合图，图中实线表示微分方程解的拟合曲线，圆圈表示实测径流量原始观测值，可以看出，原始数据基本分布在拟合曲线周围，说明微分方程的拟合效果较好。因此，运用 PDA 方法研究实测径流

量序列，不仅能较好地拟合曲线，还可以探索其导数等潜在信息，具有相当好的实际应用价值。

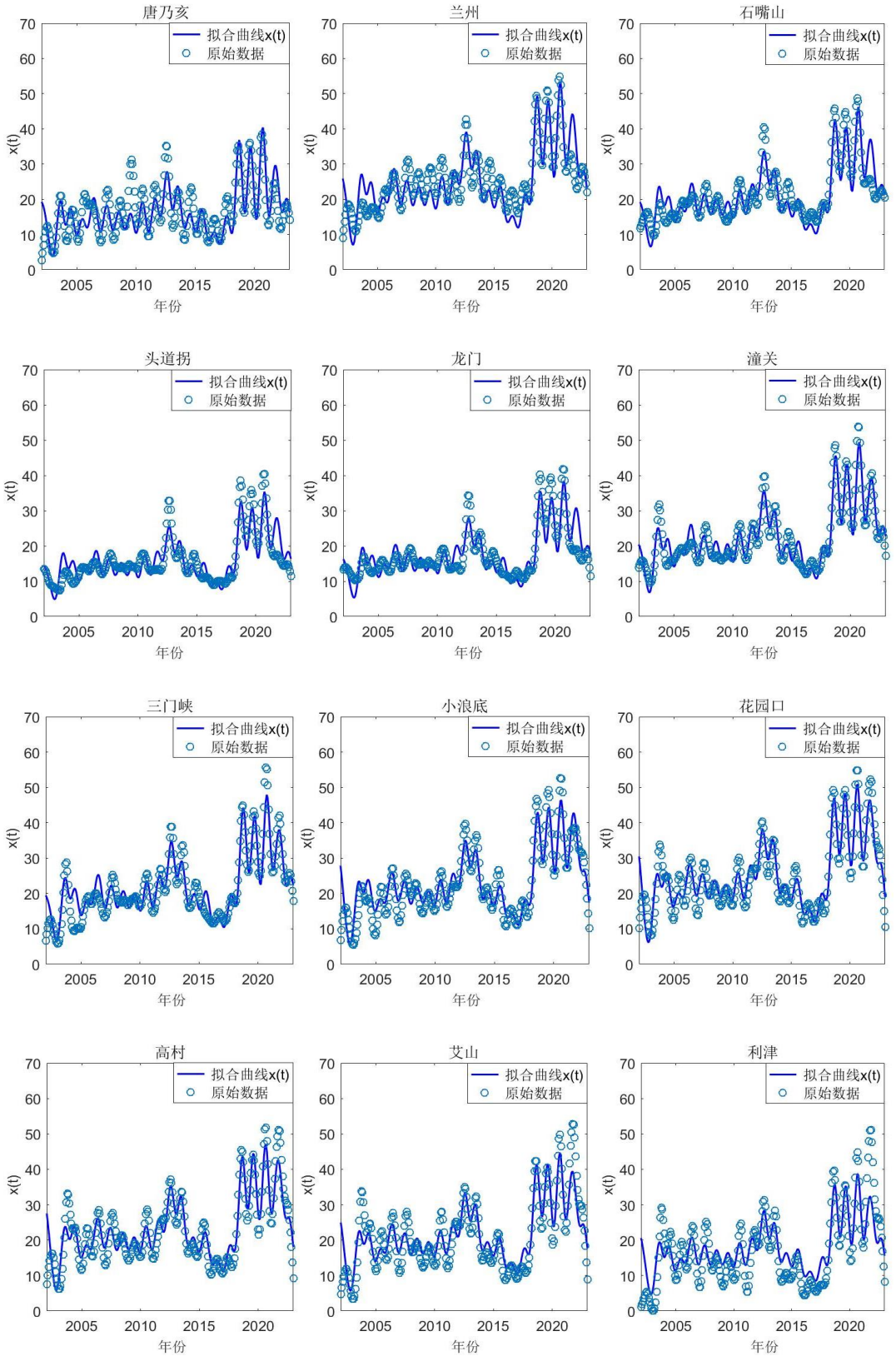


图 3.7　基于微分方程的拟和图

3.4 径流量季节变动的刻画

基于主微分分析方法，从原始实测径流量中分离出重要的季节变动并对其进行刻画。为方便起见，将黄河干流 12 个代表水文站实测径流量的均值看作实测径流量的原始数据。首先拟合匀滑曲线$g(t)$，其能够捕捉到实测径流量的长期变动趋势；其次通过将实测径流量的拟合函数$x(t)$和长期变动趋势曲线$g(t)$作差，得到季节变动成分和误差成分；最后拟合实测径流量的季节变动成分。实测径流量的季节变动曲线为

$$s(t) = x(t) - g(t), \tag{3.2}$$

其中，$x(t)$由实测径流量的原始月度数据拟合，$g(t)$由实测径流量的年度数据拟合。图 4.1 为$x(t)$与$g(t)$拟合曲线。

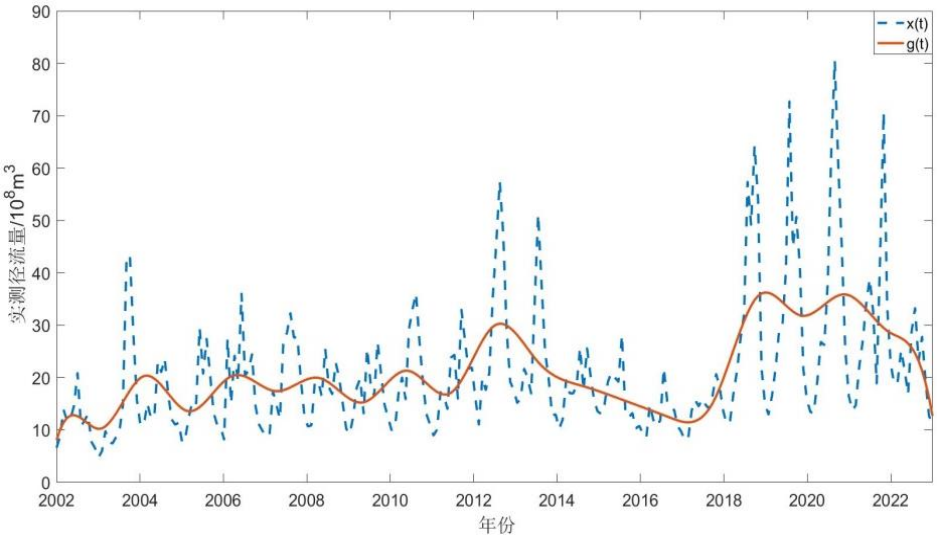


图 4.1 $g(x)$与$x(t)$的图像曲线图

实测径流量的季节变动曲线$s(t)$如图 4.2 所示，可以看出，实测径流量的季节性变动极不稳定，在实测径流量波动比较小的年份，其季节变动较为平缓；反之在实测径流量波动较大的年份，其季节变动较为剧烈。例如在 2002-2017 年间，实测径流量波动幅度较小，其季节变动幅度也较小，而 2018-2022 年间实测径流量大起大落，其对应的季节变动幅度也大。图 4.3 为采用傅里叶基函数拟合的季节变动曲线，其中圆圈为各个年份每月的季节变动的数值点，实线为季节变动曲线$s(t)$的估计$\hat{s}(t)$。可以看出，季节变动曲线的拟合效果较好，较好地呈现了黄河干流实测径流量的季节变动特征，基本刻画了实测径流量的季节变动规律。

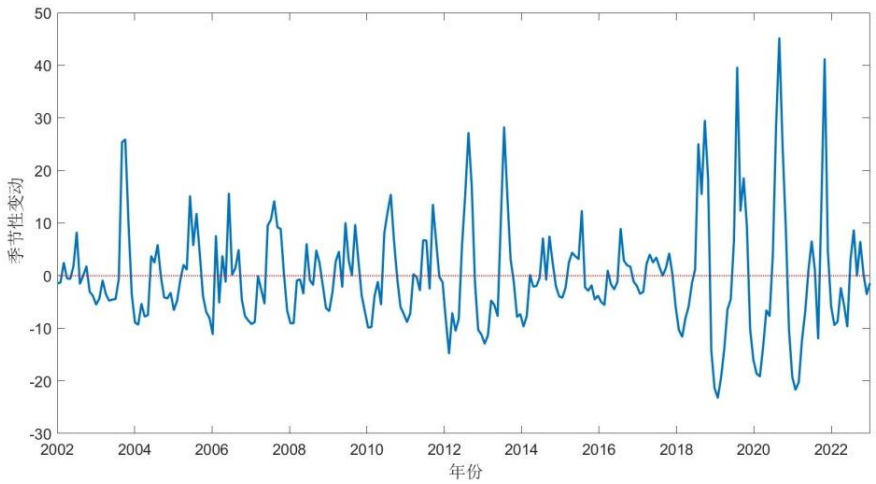


图 4.2 季节变动函数曲线图

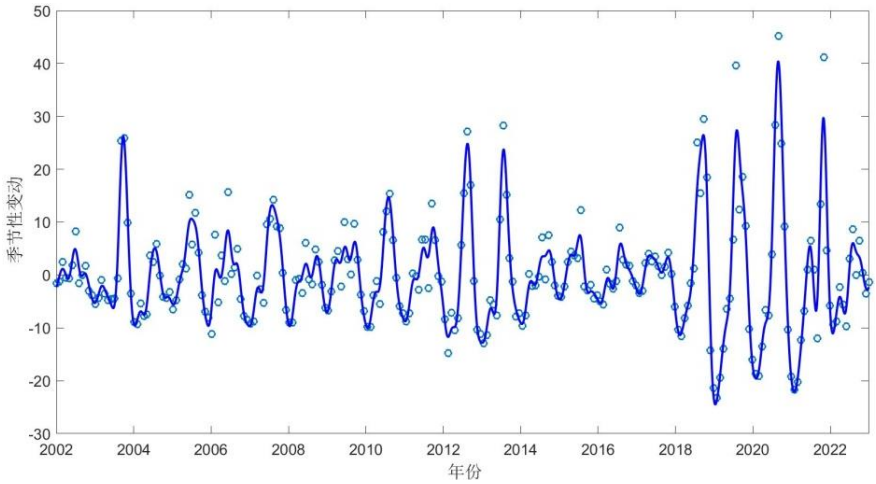


图 4.3 实测径流量季节变动函数的拟合曲线图

3.5 本章小结

本部分基于相平面图、主微分分析、基函数拟合法探究黄河干流径流量的趋势性、突变性、周期性，以及其季节变动特征。首先，通过对径流量序列进行曲线拟合、修匀，并引入其均值函数和方差函数，掌握径流量的趋势性、突变性以及周期性；其次，通过其导数信息绘制径流量的相平面图，以探究黄河干流径流量变化速度与加速度的关系、动能与势能交替变化的规律，并分析径流量的突变特征；此外，对径流量数据进行主微分分析，通过微分方程来刻画径流量的动态演变规律以及波动特征；最后，从原始径流量数据中分离出重要的季节变动并运用基函数拟合法对其进行刻画。

4 基于函数型聚类分析的黄河干流径流量时空分布特征

本节充分考虑径流量数据的曲线特征，以及径流量受水文站地势、流域面积及气候等因素影响其变化速度，基于非负矩阵分解思想，引入矩阵的$L_{2,1}$范数，结合主微分分析和主成分分析的思想，构建一种函数型聚类分析方法，进而给出迭代更新算法以优化该目标模型，最后将其应用于黄河干流 12 个主要水文站的聚类中，以期为水文数据的合理聚类提供方法参考。

4.1 函数型聚类分析方法的构建

4.1.1 目标函数

考虑到径流量数据的函数型特性，基于多视角学习方法思想，结合函数型数据自身特征信息与其波动特征信息，构建函数型聚类模型。同时，该模型能够统一处理函数型数据的生成和聚类特征的提取。下面为所构造的优化问题框架

$$\min \ f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{\theta}) + g(\boldsymbol{\theta}) + \mathrm{PEN}(\boldsymbol{\theta}), \tag{4.1}$$

其中，$\boldsymbol{\theta}$为待估参数，$f_1(\boldsymbol{\theta})$、$f_2(\boldsymbol{\theta})$分别描述主成分、主微分所代表的数据信息，$g(\boldsymbol{\theta})$代表函数型数据聚类结果的总差异，惩罚函数$\mathrm{PEN}(\boldsymbol{\theta})$以防止过拟合。

一般地，利用式(2.1)进行曲线拟合，假设某个既定空间的一组基底函数$\{\phi_{i1}, \phi_{i2}, \cdots\}$可将曲线$x_i(t)$线性表出为

$$x_i(t) = \sum_{i=1}^{\infty} \alpha_{il}\phi_{il}(t). \tag{4.2}$$

将有限性的线性组合逼近$x_i(t)$，有

$$x_i(t) \approx \sum_{l=1}^{L} \alpha_{il}\phi_{il}(t) = \boldsymbol{\alpha}_i{}^T \boldsymbol{\phi}_i(t), \tag{4.3}$$

其中，$\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iL})^T$，$\boldsymbol{\phi}_i(t) = (\phi_{i1}(t), \phi_{i2}(t), \cdots, \phi_{iL}(t))^T$。式(4.3)的矩阵形式为

$$\boldsymbol{X} \approx \boldsymbol{\Phi}\boldsymbol{A},$$

其中，$\boldsymbol{\Phi} = (\phi_i, \phi_2, \cdots, \phi_m)^T$，$\boldsymbol{A} = (\alpha_i, \alpha_2, \cdots, \alpha_N)$。如果曲线拟合过程中采用相同的基函数，那么$\boldsymbol{X}(t)$中曲线间的差异完全取决于系数矩阵$\boldsymbol{A}$。式(2.1)的矩阵

形式为

$$\boldsymbol{Y}=\boldsymbol{\Phi A}+\boldsymbol{E}, \tag{4.4}$$

其中$\boldsymbol{E}=(\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_1,\cdots,\boldsymbol{\epsilon}_N)$，$\boldsymbol{\epsilon}_i=(\epsilon_{i1},\epsilon_{i2},\cdots,\epsilon_{im})^T$。$\boldsymbol{Y}=\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_n\}$为离散点的集合。最小化式(4.4)中的目标函数，有

$$\|\boldsymbol{Y}-\boldsymbol{\Phi A}\|_F^2+\beta\mathrm{PEN}(\boldsymbol{A})$$

可得到$\boldsymbol{A}$的估计。其中，$\|\cdot\|_F^2$为矩阵的 Frobenius 范数，β为调节参数。

鉴于 NMF 的聚类特性(Ding 等，2005)，本节将在 NMF 框架下进行聚类。然而，传统 NMF 算法的目标函数常以L_2范数作为度量准则，当数据中存在噪声或异常值时，易放大目标函数中异常值的作用，导致聚类性能变差。针对这一问题，在目标函数中引入$L_{2,1}$范数，一定程度上降低异常值在目标函数中的作用，使得聚类方法更加精准。事实上，系数矩阵$\boldsymbol{A}$决定了函数曲线的差异，则利用 NMF 进行函数型数据聚类的过程中可仅对系数矩阵$\boldsymbol{A}$进行展开。此外，径流数据还具有函数型特性，从函数型主成分分析和主微分分析两个视角出发，既考虑了函数型径流量数据本身的绝大多信息，又考虑了其波动特征等潜在信息。为此，针对函数型径流量数据，在基于$L_{2,1}$范数的函数型非负矩阵分解聚类算法中引入函数型主成分和主微分得分，即式(4.1)中的$f_1(\boldsymbol{\theta})+f_2(\boldsymbol{\theta})$为

$$\begin{aligned}&f_1(\boldsymbol{\theta})+f_2(\boldsymbol{\theta})=w_1\|\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1\|_{2,1}+w_2\|\boldsymbol{Y}-\boldsymbol{\Phi}_2\boldsymbol{U}_2\boldsymbol{V}_2\|_{2,1},\\&\text{s.t. }\boldsymbol{V}_v\geq 0,\ \ \boldsymbol{U}_v\geq 0,\ v=1,2,\end{aligned} \tag{4.5}$$

其中，w_1、w_2分别为函数型主成分部分和主微分部分的权重系数，$\|\cdot\|_{2,1}$为矩阵的$L_{2,1}$范数，$\boldsymbol{Y}$为原始观测值，$\boldsymbol{\Phi}_1$、$\boldsymbol{\Phi}_2$均为基函数，如 B 样条基函数、傅里叶基函数和多项式基函数等，$\boldsymbol{U}_v$为基矩阵，系数矩阵$\boldsymbol{V}_1$和$\boldsymbol{V}_2$分别表示在主成分视角和主微分视角的函数型数据聚类结果。为统一两个视角下的聚类结果，构造以下损失函数

$$D(\boldsymbol{V}_v,\boldsymbol{V}^*)=\sum_{v=1}^{2}(\lambda_v\|\boldsymbol{V}_v-\boldsymbol{V}^*\|_F^2), \tag{4.6}$$

其中，λ_v为调节参数，$\boldsymbol{V}^*=(v_{ik})_{n\times k}\in\{\boldsymbol{0},\boldsymbol{1}\}^{n\times k}$表示聚类结果，$\boldsymbol{1}$为元素全为 1 的列向量。另外，加入惩罚项以防止过拟合

$$\mathrm{PEN}(\boldsymbol{\theta})=\sum_{v=1}^{2}(\alpha\|\boldsymbol{V}_v\|_F^2+\beta\|\boldsymbol{U}_v\|_F^2), \tag{4.7}$$

其中，调节参数α和β用来调整两视角间的相对权重。

结合式(4.5)、式(4.6)和式(4.7)，构建本文所提的基于主成分和主微分的函数型聚类方法(Functional Clustering Model based on Principal Component and Principal Differential，PCPDFCM)，其目标函数如下

$$\begin{aligned}\mathcal{O} = \min\ \{&w_1\|\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\|_{2,1}+w_2\|\boldsymbol{Y}-\boldsymbol{\Phi}_2\boldsymbol{U}_2\boldsymbol{V}_2^T\|_{2,1}\\&+\sum_{v=1}^{2}(\lambda_v\|\boldsymbol{V}_v-\boldsymbol{V}^*\|_F^2+\alpha\|\boldsymbol{V}_v\|_F^2+\beta\|\boldsymbol{U}_v\|_F^2)\},\\&\text{s.t. } \boldsymbol{V}_v\geq 0,\ \ \boldsymbol{U}_v\geq 0,\ \ \boldsymbol{V}^*\geq 0,\end{aligned} \tag{4.8}$$

其中，w_1、w_2分别为函数型主成分部分和主微分部分的权重系数，$\|\cdot\|_{2,1}$为矩阵的$L_{2,1}$范数，$\boldsymbol{Y}$为原始观测值，$\boldsymbol{\Phi}_v$为基函数，$\boldsymbol{U}_v$为基矩阵，系数矩阵$\boldsymbol{V}_1$、$\boldsymbol{V}_2$分别表示在主成分、主微分视角下的聚类结果，$\boldsymbol{V}^*=(v_{ik})_{n\times k}\in\{\boldsymbol{0},\boldsymbol{1}\}^{n\times k}$表示在两个视角下的最终聚类结果。

4.1.2 求解算法

目标函数式(4.8)同时针对待估参数矩阵$\boldsymbol{U}_1$、$\boldsymbol{V}_1$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$是非凸函数，不易得到全局最优解。因此，采用乘性迭代方法(Liang 等，2020)和 KKT 互补松弛条件以施加非负性约束，给出获得局部最优解的交替迭代算法。

（1）固定$\boldsymbol{V}^*$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$，更新$\boldsymbol{U}_1$和$\boldsymbol{V}_1$

固定$\boldsymbol{V}^*$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$，对于主成分视角，目标函数式(4.8)可简化为

$$\begin{aligned}&\min\ w_1\|\boldsymbol{Y}-\Phi_1\boldsymbol{U}_1\boldsymbol{V}_1^T\|_{2,1}+\lambda_1\|\boldsymbol{V}_1-\boldsymbol{V}^*\|_F^2+\alpha\|\boldsymbol{V}_1\|_F^2+\beta\|\boldsymbol{U}_1\|_F^2\},\\&\text{s.t. }\ \boldsymbol{V}_1\geq 0,\ \ \boldsymbol{U}_1\geq 0,\ \ \boldsymbol{V}^*\geq 0.\end{aligned}$$

令$\boldsymbol{U}_1\geq 0$、$\boldsymbol{V}_1\geq 0$和$\boldsymbol{V}^*>0$的拉格朗日乘子矩阵分别为$\boldsymbol{\Lambda}_1$、$\boldsymbol{\Gamma}_1$和$\boldsymbol{\Theta}_1$，相应的拉格朗日函数为

$$\begin{aligned}L_1 =&w_1\|\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\|_{2,1}+\lambda_1\|\boldsymbol{V}_1-\boldsymbol{V}^*\|_F^2+\alpha\|\boldsymbol{V}_1\|_F^2\\&+\beta\|\boldsymbol{U}_1\|_F^2-\mathrm{tr}(\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T)-\mathrm{tr}(\boldsymbol{\Gamma}_1\boldsymbol{V}_1^T)-\mathrm{tr}(\boldsymbol{\Theta}_1\boldsymbol{V}^{*T}).\end{aligned} \tag{4.9}$$

①保持$\boldsymbol{V}_1$不变，更新$\boldsymbol{U}_1$。

L_1关于$\boldsymbol{U}_1$求偏导，并令$\frac{\partial L_1}{\partial \boldsymbol{U}_1}=0$，则

$$\begin{aligned}\frac{\partial L_1}{\partial \boldsymbol{V}_1}&=\frac{\partial[w_1\|\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\|_{2,1}+\beta\|\boldsymbol{U}_1\|_F^2-\mathrm{tr}(\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T)]}{\partial\boldsymbol{U}_1}\\&=\frac{\partial\mathrm{tr}[w_1(\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T)\boldsymbol{D}(\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T)^T+\beta\boldsymbol{U}_1^T\boldsymbol{U}_1-\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T]}{\partial\boldsymbol{U}_1}\\&=-2w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1+2w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+2\beta\boldsymbol{U}_1-\boldsymbol{\Lambda}_1\\&=0,\end{aligned}$$

其中$\boldsymbol{D}$为对角矩阵，其主要对角元素为

$$\boldsymbol{D}_{kk}=\sqrt{\frac{1}{\sum_{i=1}^{m}(\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T)_{ik}^2}},$$

从而

$$\boldsymbol{\Lambda}_1=-2w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1+2w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+2\beta\boldsymbol{U}_1.$$

非负约束使得 KKT 条件$\boldsymbol{\Lambda}_1\odot\boldsymbol{U}=0$成立，即满足

$$(-w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1+w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+\beta\boldsymbol{U}_1)_{ij}\odot\boldsymbol{U}_{1,ij}=0,$$

其中$\odot$为 Hadamard 积。对j求和，依据矩阵乘法，有

$$\begin{aligned}&\sum_j[(-w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1+w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+\beta\boldsymbol{U}_1)]_{ij}\odot\boldsymbol{U}_{1,ij}\\=&(-w_1\boldsymbol{U}_1^T\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1+w_1\boldsymbol{U}_1^T\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+\beta\boldsymbol{U}_1^T\boldsymbol{U}_1)_{ii}\\=&0,\end{aligned}$$

从而可得$\boldsymbol{U}_1$的更新规则为

$$\boldsymbol{U}_{1,ij}\leftarrow\boldsymbol{U}_{1,ij}\sqrt{\frac{w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1}{w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+\beta\boldsymbol{U}_1}}.\tag{4.10}$$

②固定$\boldsymbol{U}_1$，更新$\boldsymbol{V}_1$。

L_1关于$\boldsymbol{V}_1$求偏导，并令$\frac{\partial L_1}{\partial\boldsymbol{V}_1}=0$，则

$$\begin{aligned}\frac{\partial L_1}{\partial\boldsymbol{V}_1}=&\frac{\partial[w_1\|\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\|_{2,1}+\lambda_1\|\boldsymbol{V}_1-\boldsymbol{V}^*\|_F^2+\alpha\|\boldsymbol{V}_1\|_F^2-\mathrm{tr}(\boldsymbol{\Gamma}_1\boldsymbol{V}_1^T)]}{\partial\boldsymbol{V}_1}\\=&\frac{\partial\mathrm{tr}[w_1(\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T)\boldsymbol{D}(\boldsymbol{Y}-\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T)^T]}{\partial\boldsymbol{V}_1}+\frac{\partial\mathrm{tr}[\lambda_1(\boldsymbol{V}_1-\boldsymbol{V}^*)(\boldsymbol{V}_1-\boldsymbol{V}^*)^T]}{\partial\boldsymbol{V}_1}\\&+\frac{\partial\mathrm{tr}[\alpha\boldsymbol{V}_1^T\boldsymbol{V}_1-\boldsymbol{\Gamma}_1\boldsymbol{V}_1^T]}{\partial\boldsymbol{V}_1}\\=&-2w_1\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{\Phi}_1\boldsymbol{U}_1+2w_1\boldsymbol{\Phi}_1^T\boldsymbol{U}_1^T\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1\boldsymbol{\Phi}_1+2\lambda_1(\boldsymbol{V}_1-\boldsymbol{V}^*)+2\alpha\boldsymbol{V}_1-\boldsymbol{\Gamma}_1\\=&0,\end{aligned}$$

从而

$$\boldsymbol{\Gamma}_1=-2w_1\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{\Phi}_1\boldsymbol{U}_1+2w_1\boldsymbol{\Phi}_1^T\boldsymbol{U}_1^T\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1\boldsymbol{\Phi}_1+2\lambda_1(\boldsymbol{V}_1-\boldsymbol{V}^*)+2\alpha\boldsymbol{V}_1.$$

非负约束使得 KKT 条件$\boldsymbol{\Gamma}_1\odot\boldsymbol{V}=0$成立，即满足

$$(-w_1\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{\Phi}_1\boldsymbol{U}_1+w_1\boldsymbol{\Phi}_1^T\boldsymbol{U}_1^T\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1\boldsymbol{\Phi}_1+\lambda_1(\boldsymbol{V}_1-\boldsymbol{V}^*)+\alpha\boldsymbol{V}_1)_{ij}\odot\boldsymbol{V}_{1,ij}=0,$$

其中$\odot$为 Hadamard 积。对j求和，依据矩阵乘法，有

$$\begin{aligned}&\sum_j[(-w_1\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{\Phi}_1\boldsymbol{U}_1+w_1\boldsymbol{\Phi}_1^T\boldsymbol{U}_1^T\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1\boldsymbol{\Phi}_1+\lambda_1(\boldsymbol{V}_1-\boldsymbol{V}^*)+\alpha\boldsymbol{V}_1)_{ij}]_{ij}\odot\boldsymbol{V}_{1,ij}\\&=(-w_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{\Phi}_1\boldsymbol{U}_1+w_1\boldsymbol{V}_1^T\boldsymbol{\Phi}_1^T\boldsymbol{U}_1^T\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1\boldsymbol{\Phi}_1+\lambda_1\boldsymbol{V}_1^T(\boldsymbol{V}_1-\boldsymbol{V}^*)+\alpha\boldsymbol{V}_1^T\boldsymbol{V}_1)_{ii}\\&=0,\end{aligned}$$

进而，可得$\boldsymbol{V}_1$的更新规则为

$$\boldsymbol{V}_{1,ij} \leftarrow \boldsymbol{V}_{1,ij}\sqrt{\frac{w_1\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{\Phi}_1\boldsymbol{U}_1}{w_1\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1^T\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1+\lambda_1(\boldsymbol{V}_1-\boldsymbol{V}^*)+\alpha\boldsymbol{V}_1}}. \tag{4.11}$$

（2）固定$\boldsymbol{V}^*$、$\boldsymbol{U}_1$和$\boldsymbol{V}_1$，更新$\boldsymbol{U}_2$和$\boldsymbol{V}_2$

固定$\boldsymbol{V}^*$、$\boldsymbol{U}_1$和$\boldsymbol{V}_1$，对于主微分视角，目标函数式(4.8)可简化为：

$$\begin{aligned}&\min\ w_2\|\boldsymbol{Y}-\Phi_2\boldsymbol{U}_2\boldsymbol{V}_2^T\|_{2,1}+\lambda_2\|\boldsymbol{V}_2-\boldsymbol{V}^*\|_F^2+\alpha\|\boldsymbol{V}_2\|_F^2+\beta\|\boldsymbol{U}_2\|_F^2\},\\&\text{s.t.}\ \ \boldsymbol{V}_2\geq 0,\ \ \boldsymbol{U}_2\geq 0,\ \ \boldsymbol{V}^*\geq 0.\end{aligned}$$

令$\boldsymbol{U}_2\geq 0$、$\boldsymbol{V}_2\geq 0$和$\boldsymbol{V}^*>0$的拉格朗日乘子矩阵分别为$\boldsymbol{\Lambda}_2$、$\boldsymbol{\Gamma}_2$和$\boldsymbol{\Theta}_2$，则相应的拉格朗日函数为

$$\begin{aligned}L_2=&w_2\|\boldsymbol{Y}-\boldsymbol{\Phi}_2\boldsymbol{U}_2\boldsymbol{V}_2^T\|_{2,1}+\lambda_2\|\boldsymbol{V}_2-\boldsymbol{V}^*\|_F^2+\alpha\|\boldsymbol{V}_2\|_F^2\\&+\beta\|\boldsymbol{U}_2\|_F^2-\mathrm{tr}(\boldsymbol{\Lambda}_2\boldsymbol{U}_2^T)-\mathrm{tr}(\boldsymbol{\Gamma}_2\boldsymbol{V}_2^T)-\mathrm{tr}(\boldsymbol{\Theta}_2\boldsymbol{V}^{*T}).\end{aligned} \tag{4.12}$$

与$\boldsymbol{U}_1$和$\boldsymbol{V}_1$同理：

①当$\boldsymbol{V}_2$不变时，$\boldsymbol{U}_2$的更新公式为

$$\boldsymbol{U}_{2,ij} \leftarrow \boldsymbol{U}_{2,ij}\sqrt{\frac{w_2\boldsymbol{\Phi}_2^T\boldsymbol{Y}\boldsymbol{G}\boldsymbol{V}_2}{w_2\boldsymbol{\Phi}_2^T\boldsymbol{\Phi}_2\boldsymbol{U}_2\boldsymbol{V}_2^T\boldsymbol{G}\boldsymbol{V}_2+\beta\boldsymbol{U}_2}}, \tag{4.13}$$

其中$\boldsymbol{G}$为对角矩阵，其主要对角元素为

$$\boldsymbol{G}_{kk}=\sqrt{\frac{1}{\sum_{i=1}^m(\boldsymbol{Y}-\boldsymbol{\Phi}_2\boldsymbol{U}_2\boldsymbol{V}_2^T)_{ik}^2}}.$$

②当$\boldsymbol{U}_2$不变时，$\boldsymbol{V}_2$的更新公式为

$$\boldsymbol{V}_{2,ij} \leftarrow \boldsymbol{V}_{2,ij}\sqrt{\frac{w_2\boldsymbol{G}\boldsymbol{Y}^T\boldsymbol{\Phi}_2\boldsymbol{U}_2}{w_2\boldsymbol{G}\boldsymbol{V}_2\boldsymbol{U}_2^T\boldsymbol{\Phi}_2^T\boldsymbol{\Phi}_2\boldsymbol{U}_2+\lambda_2(\boldsymbol{V}_2-\boldsymbol{V}^*)+\alpha\boldsymbol{V}_2}}. \tag{4.14}$$

（3）固定$\boldsymbol{U}_1$、$\boldsymbol{V}_1$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$，更新$\boldsymbol{V}^*$

当保持$\boldsymbol{U}_1$、$\boldsymbol{V}_1$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$不变时，令$\boldsymbol{V}^*\geq 0$的拉格朗日乘子矩阵为$\boldsymbol{\Theta}$，则相应的拉格朗日函数为

$$L_3=\sum_{v=1}^2\lambda_v\|\boldsymbol{V}_v-\boldsymbol{V}^*\|_F^2-\mathrm{tr}(\sum_{v=1}^2\boldsymbol{\Theta}\boldsymbol{V}^{*T}). \tag{4.15}$$

L_3关于$\boldsymbol{V}^*$求偏导，并令$\frac{\partial L_3}{\partial \boldsymbol{V}^*}=0$，则

$$\begin{aligned}\frac{\partial L_3}{\partial \boldsymbol{V}^*}&=\frac{\partial\mathrm{tr}[\sum_{v=1}^2\lambda_v(\boldsymbol{V}_v^T\boldsymbol{V}_v-2\boldsymbol{V}^{*T}\boldsymbol{V}_v+\boldsymbol{V}^{*T}\boldsymbol{V}^*)-(\boldsymbol{\Theta}\boldsymbol{V}^{*T})]}{\partial\boldsymbol{V}^*}\\&=-2\sum_{v=1}^2\lambda_v\boldsymbol{V}_v+\sum_{v=1}^2 2\lambda_v\boldsymbol{V}^*-\boldsymbol{\Theta}\\&=0,\end{aligned}$$

从而

$$\boldsymbol{\Theta} = 2\sum_{v=1}^{2}\lambda_v(\boldsymbol{V}^* - \boldsymbol{V}_v).$$

利用 KKT 条件$\boldsymbol{\Theta} \odot \boldsymbol{V}^* = 0$，得到$\boldsymbol{V}^*$的更新规则为

$$\boldsymbol{V}^* = \frac{\sum_{v=1}^{2}\lambda_v\boldsymbol{V}_v}{\sum_{v=1}^{2}\lambda_v}. \tag{4.16}$$

综上分析，依次利用式(4.10)、式(4.11)、式(4.13)、式(4.14)和式(4.16)分别更新$\boldsymbol{U}_1$、$\boldsymbol{V}_1$、$\boldsymbol{U}_2$、$\boldsymbol{V}_2$和$\boldsymbol{V}^*$，实现基于主成分和主微分的函数型聚类方法(PCPDFCM)。具体过程归纳如表 4.1 所示。

表 4.1　PCPDFCM 方法框架

算法 1　基于主成分和主微分的函数型聚类方法(PCPDFCM)

输入：原始数据矩阵$\boldsymbol{Y}$、基函数$\boldsymbol{\Phi}_1$、$\boldsymbol{\Phi}_2$、类别数K，权重系数w_1和w_2，参数α和β

过程：

1：初始化：$\boldsymbol{V}_1^0$、$\boldsymbol{V}_2^0$和$\boldsymbol{V}^{*0}$;
2：**for** $t = 1, 2, \cdots$最大更新迭代次数
3：　固定$\boldsymbol{V}_1^{t-1}$、$\boldsymbol{U}_2^{t-1}$、$\boldsymbol{V}_2^{t-1}$和$\boldsymbol{V}^{*t-1}$，根据式(4.10)更新$\boldsymbol{U}_1^{t-1}$;
4：　固定$\boldsymbol{U}_1^{t-1}$、$\boldsymbol{U}_2^{t-1}$、$\boldsymbol{V}_2^{t-1}$和$\boldsymbol{V}^{*t-1}$，根据式(4.11)更新$\boldsymbol{V}_1^{t-1}$;
5：　固定$\boldsymbol{V}_1^{t-1}$、$\boldsymbol{U}_1^{t-1}$、$\boldsymbol{V}_1^{t-1}$和$\boldsymbol{V}^{*t-1}$，根据式(4.13)更新$\boldsymbol{U}_2^{t-1}$;
6：　固定$\boldsymbol{U}_2^{t-1}$、$\boldsymbol{U}_1^{t-1}$、$\boldsymbol{V}_1^{t-1}$和$\boldsymbol{V}^{*t-1}$，根据式(4.14)更新$\boldsymbol{V}_2^{t-1}$;
7：　固定$\boldsymbol{V}_1^{t-1}$、$\boldsymbol{U}_2^{t-1}$、$\boldsymbol{V}_2^{t-1}$和$\boldsymbol{U}_2^{t-1}$，根据式(4.16)更新$\boldsymbol{V}^{*t-1}$;
8：　**if**　式(4.8)收敛
9：　　**break**
10：　**end if**
11：**end for**

输出：$\boldsymbol{V}_1^t$、$\boldsymbol{U}_2^t$、$\boldsymbol{V}_2^t$、$\boldsymbol{U}_2^t$和$\boldsymbol{V}^{*t}$，簇划分$\mathcal{C} = \{C_1, C_2 \cdots, C_K\}$

4.1.3 收敛性分析

由于目标函数式(4.8)关于$\boldsymbol{U}_1$、$\boldsymbol{V}_1$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$是联合非凸的，表 4.1 中的算法 1 无法保证获得全局最优解。事实上，可以证明算法 1 是局部收敛的。

定理 1　目标函数式(4.8)分别在更新公式(4.10)、(4.11)、(4.13)和(4.14)下是单调递减的。

证明　利用标准辅助函数法(Lee 和 Seung，2001)证明更新公式(4.10)、(4.11)、(4.13)和(4.14)的收敛性。在目标函数式(4.8)中，剔除无关项，保留与$\boldsymbol{U}_1$有关的项，有

$$L(\boldsymbol{U}_1) := \mathrm{tr}(-2w_1\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1^T\boldsymbol{\Phi}_1^T + w_1\boldsymbol{\Phi}_1\boldsymbol{U}_1\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1\boldsymbol{U}_1^T\boldsymbol{\Phi}_1^T) + \beta\mathrm{tr}(\boldsymbol{U}_1^T\boldsymbol{U}_1).$$

构造$L(\boldsymbol{U}_1)$的辅助函数$G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)$ (Ding 等，2006)

$$\begin{aligned}G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)=&\sum_{i,j}\frac{(w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1^t\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1)(i,j)\boldsymbol{U}_1(i,j)^2}{\boldsymbol{U}_1^t(i,j)}+\beta\sum_{i,j}\frac{\boldsymbol{U}_1^t(i,j)\boldsymbol{U}_1(i,j)^2}{\boldsymbol{U}_1^t(i,j)}\\&-2\sum_{i,j}(w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1)(i,j)\boldsymbol{U}_1^t(i,j)(1+\log\frac{\boldsymbol{U}_1(i,j)}{\boldsymbol{U}_1^t(i,j)}),\end{aligned}\tag{4.17}$$

则满足条件

$$G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)=L(\boldsymbol{U}_1),\ \ G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)\geq L(\boldsymbol{U}_1).$$

如果取$\boldsymbol{U}_1^{t+1}$，使得

$$\boldsymbol{U}_1^{t+1}=\arg\min_t G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)\tag{4.18}$$

成立，易知$L(\boldsymbol{U}_1)$是单调递减的，即

$$L(\boldsymbol{U}_1^{t+1})\leq G(\boldsymbol{U}_1^{t+1},\boldsymbol{U}_1^t)\leq G(\boldsymbol{U}_1^t,\boldsymbol{U}_1^t)\leq L(\boldsymbol{U}_1^t).$$

根据式(4.18)，使辅助函数式(4.17)达到最小，求解$\boldsymbol{U}_1^{t+1}$。对式(4.17)关于$\boldsymbol{U}_{1,ij}$求偏导，得

$$\frac{\partial G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)}{\partial\boldsymbol{U}_{1,ij}}=\frac{(2w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1^t\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1)_{ij}\boldsymbol{U}_{1,ij}}{\boldsymbol{U}_{1,ij}^t}+2\beta\boldsymbol{U}_{1,ij}-(2w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1)_{ij}\frac{\boldsymbol{U}_{1,ij}^t}{\boldsymbol{U}_{1,ij}}.$$

令$\frac{\partial G(\boldsymbol{U}_1,\boldsymbol{U}_1^t)}{\partial\boldsymbol{U}_{1,ij}}=0$，则有

$$(w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1)_{ij}\frac{\boldsymbol{U}_{1,ij}^t}{\boldsymbol{U}_{1,ij}^{t+1}}=(\frac{(w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1^t\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1)_{ij}}{\boldsymbol{U}_{1,ij}^t}+\beta)\boldsymbol{U}_{1,ij}^{t+1},$$

进一步，得到

$$\boldsymbol{U}_{1,ij}^{t+1}\leftarrow\boldsymbol{U}_{1,ij}^t\sqrt{\frac{(w_1\boldsymbol{\Phi}_1^T\boldsymbol{Y}\boldsymbol{D}\boldsymbol{V}_1)_{ij}}{(w_1\boldsymbol{\Phi}_1^T\boldsymbol{\Phi}_1\boldsymbol{U}_1^t\boldsymbol{V}_1^T\boldsymbol{D}\boldsymbol{V}_1+\beta\boldsymbol{U}_1^t)_{ij}}},$$

上式即为$\boldsymbol{U}_1$的更新公式(4.10)。类似地，可分别证明$\boldsymbol{V}_1$、$\boldsymbol{U}_2$和$\boldsymbol{V}_2$的更新公式(4.11)、(4.13)和(4.14)。

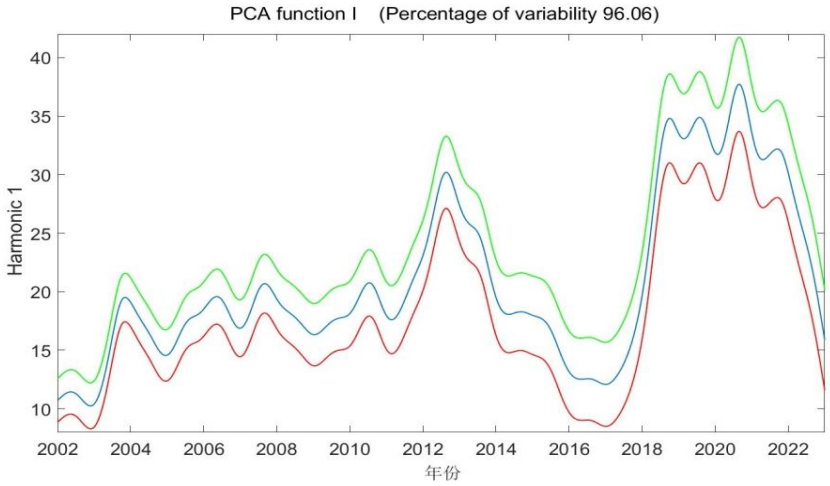
4.2 基于函数型主成分分析的径流量时空分布特征

由于$x_i(t),i=1,2,\cdots,12$作为“原始数据”具备无限维的空间特征，故利用函数型主成分分析对其进行降维处理，进一步明确黄河干流实测径流量的变化模式。表 4.2 为 FPCA 的贡献率及累计贡献率，可以看出，前两个主成分的方差累计贡献率达到 99.72%，基本上能够解释原始数据的全部信息，故选取前两个函数型主成分探究黄河干流实测径流量的整体变化特征。

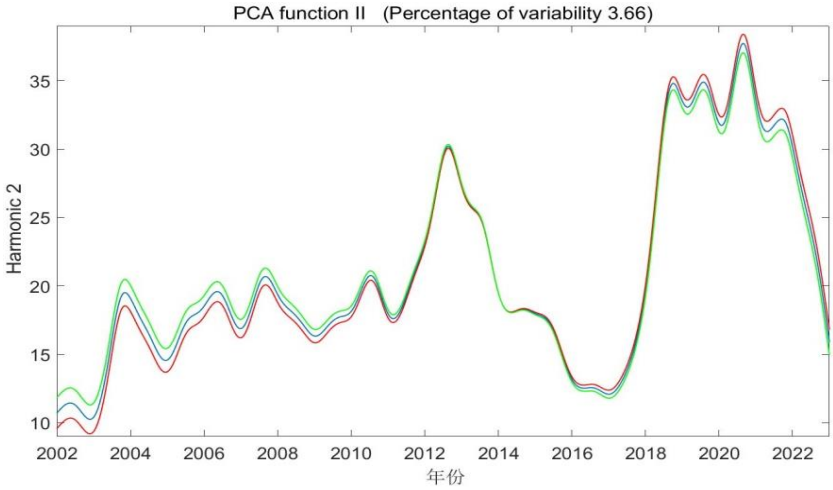
表 4.2　函数型主成分分析的贡献率和累计贡献率

函数型主成分	第一函数主成分	第二函数主成分
贡献率	96.06%	3.66%
累计贡献率	96.06%	99.72%

图 4.4 给出了第一、二主成分函数对均值函数的扰动图，反映各个主成分所代表的变化模式对黄河干流实测径流量变化的影响。第一主成分的权函数在 2013 年前后、2019-2021 年两个阶段对 2002-2022 年黄河干流实测径流量的变化有显著影响。由表 4.2 可以看出，第一主成分反映实测径流量变化的主要模式，从而这两个时期的变化情况是导致黄河干流实测径流量变动的最主要方式。尽管第二主成分只解释了总体变化程度的 3.66%，但第二主成分的权函数在 2004 年期间对黄河干流实测径流量的变化产生了影响。



(a)第一主成分



(b)第二主成分

图 4.4　第一、第二主成分函数对均值函数的扰动图

根据 12 个水文站的第一、二主成分得分，以第一主成分得分为横坐标、第二主成分得分为纵坐标，画出二维平面图，如图 4.5 所示。若第一主成分得分为正，表明该水文站在 2013 年前后、2019-2021 年期间的实测径流量大于均值，得分越高，实测径流量越大，相反，若第一主成分得分为负，表明该水文站 2013 年前后、2019-2021 年期间的实测径流量小于均值，得分越低径流量越小。同理，若第二主成分得分为正，2003 年期间的实测径流量高于均值，得分越高，实测径流量越大。

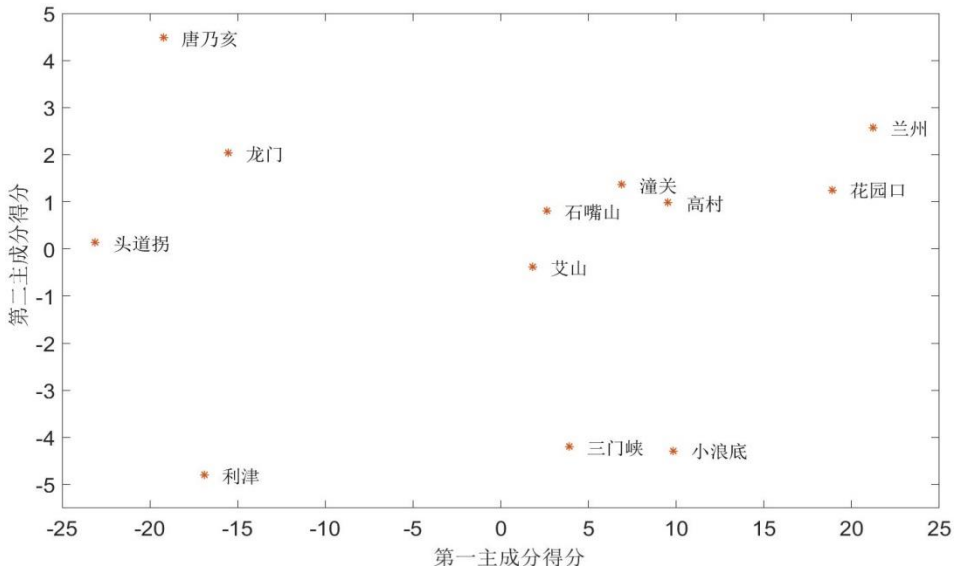


图 4.5 第一、第二主成分得分图

图 4.5 中，左下角区域的水文站(利津)的第一、第二主成分得分均为负，表明该水文站的实测径流量低于全部水文站的均值，其位于黄河干流的最下游区域，是黄河干流主要水文站实测径流量的最小值；右上角区域的兰州、花园口、潼关、高村和石嘴山等五个水文站，其第一、第二主成分得分均为正，且值相对较大，表明这些水文站的实测径流量在 2002-2022 年期间较高，是黄河干流主要水文站实测径流量较大的水文站；右下角区域的三门峡、艾山和小浪底，第一主成分为正，第二主成分为负，三门峡和小浪底位于黄河干流中游区域，这些水文站的实测径流量在 2013 年前后、2019-2021 年期间高于其他水文站；左上角区域的水文站为位于黄河干流上中游的唐乃亥、龙门和头道拐，其第一主成分为负，第二主成分为正，表明该水文站在 2003 年期间的实测径流量高于其他水文站。

综上，通过函数型主成分分析探究黄河干流各水文站不同年份径流量的时空分布特征，结果表明各水文站的实测径流量变化主要发生在 2013 年前后、

2019-2021 年，在此期间实测径流量显著增大；其次利用主成分得分分析 12 个水文站的径流量，结果表明 2002-2022 年期间，位于上游的兰州站实测径流量最高，而位于下游的利津站实测径流量最低。从空间分布上看，黄河干流实测径流量整体自上而下逐渐减小，空间差异性明显。

4.3 基于函数型聚类分析的主要水文站径流量差异性分析

基于函数型主成分分析和主微分分析的结果研究黄河干流水文站实测径流量的变化特征。由图 3.2 中黄河干流 12 个水文站的实测径流量拟合曲线知，在 2002-2022 年间，12 个水文站实测径流量总体上趋势一致，呈现上下波动的循环趋势，但也存在显著差异。为了进一步探究不同水文站实测径流量之间的异同，利用基于主成分和主微分的函数型聚类方法(PCPDFCM)对黄河干流 12 个水文站实测径流量进行函数型聚类分析。一般地，主成分主要体现原始数据的绝大多数信息，而主微分反映了实测径流量的波动信息，因此采用 PCPDFCM 方法分析 12 个水文站的实测径流量的变化规律及异同。根据各水文站径流量变化特征的走势，将 12 个水文站聚为三类，具体聚类结果如表 4.3 所示。

表 4.3 聚类结果展示

类别	水文站
第一类	唐乃亥、兰州、石嘴山、头道拐、龙门、潼关、三门峡
第二类	小浪底、花园口、高村
第三类	艾山、利津

各类中心变化特征的比较如图 4.6 所示，在整个样本区间上，第一类水文站的类中心最高，第三类次之，而第二类的类中心最低。导致径流量差异化的因素很多，例如各水文站距离入海口的远近程度、气候变化、人为因素以及国家出台的相关政策等。第一类水文站为唐乃亥、兰州、石嘴山、头道拐、龙门、潼关和三门峡，其位于黄河流域上游和中游地区，由于 1999 年国家实施的“退耕还林草”政策以及其地理优势，该类水文站的径流量较高；第二类水文站为位于黄河流域中游和下游的小浪底、花园口和高村，其径流量变化幅度较为平缓，该地区人口相对较少，人类活动强度低，气候变化对径流量的影响较大；黄河流域最下游的两个水文站为第三类水文站，该区域降水充沛，水资源开发利用和水利工程建设等人类活动较为强烈，其实测径流量较低。

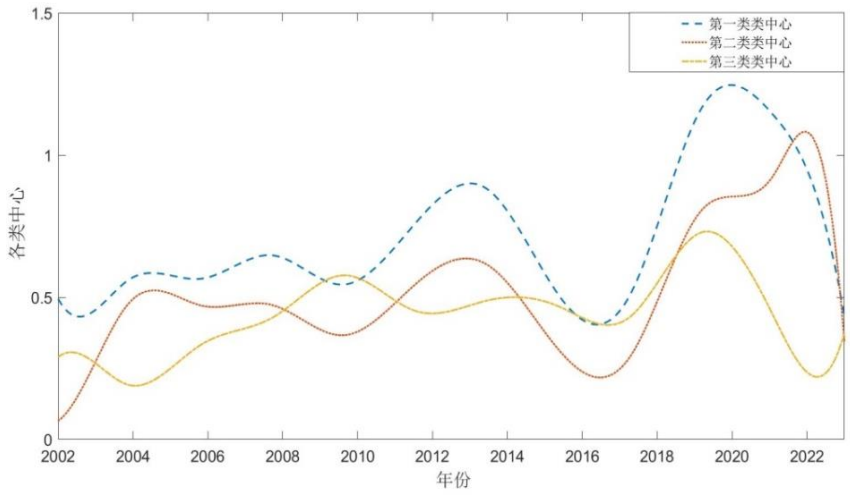


图 4.6 各类中心比较图

聚类结果的空间分布如图 4.7 所示，其中红色三角形代表第一类地区，紫色圆形代表第二类地区，绿色五角星代表第三类地区。整体来看，黄河流域干流 12 个主要水文站可分为三类：第一类为位于黄河流域上游和中游的水文站，上游地区以高原地区、崎岖山地和冰川湖泊供水为主，而中游则以平原地区、温带大陆性气候和灌溉农业为主，其实测径流量变化幅度较大；第二类水文站位于黄河流域中下游地区，该区域夏秋季暴雨较多，沙源丰富，洪峰流量大，此外，该地水库、水阀等水利工程的建设对径流量也具一定的影响；第三类水文站位于黄河流域最下游区域，其实测径流量变化特征相对较为平缓。总的来说，三类水文站在空间分布上特征明显，其水文特征具有显著差异，此研究结果也符合黄河流域上、中、下游地区划分。

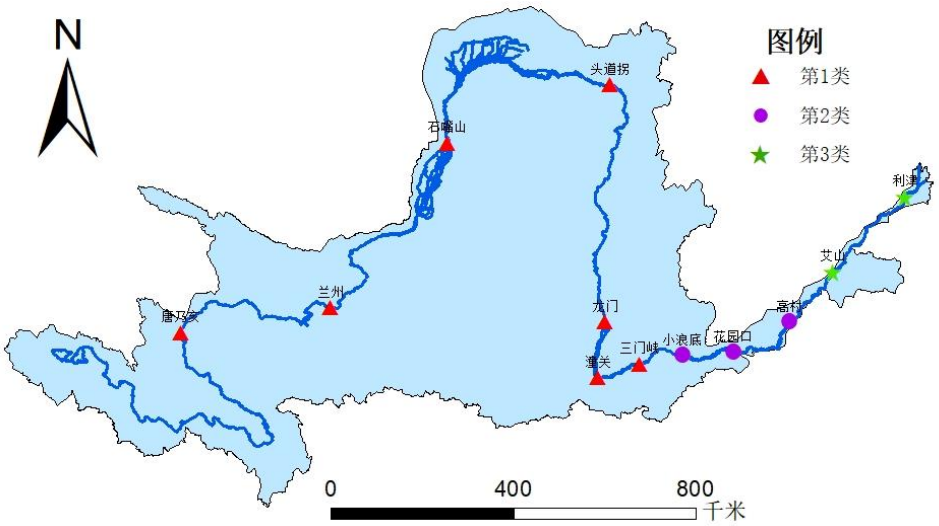


图 4.7 聚类结果空间分布

此外，黄河上游地区主要以农业耗水为主，应以农业节水为重点，统筹考虑粮食安全问题和水资源条件问题，坚定以水定地的观念，对水土开发的规模和强度进行合理确定，推进现代化改造，并分区域推广节水灌溉技术，提高灌溉用水率；黄河中下游区域在工业、生活以及生态环境的耗水较大，应开发节水技术并迅速完成工程体系，进而提高水资源的承载能力，通过对供需矛盾的调节，控制人们水的用量，采用生态项目减轻水的消耗，恢复黄河流域受损的河湖生态环境，维持河流系统健康。

进一步，验证本文所提 PCPDFCM 的聚类效果。基于 2002-2022 年黄河干流 12 个主要水文站的实测径流量数据，在固定参数设置的情况下，将 PCPDFCM 方法与函数型聚类两步法(TA)(Abraham，2003)和函数型聚类一步法(FCOF)(黄恒君等，2019)等算法进行对比。选取聚类精度(Accuracy)、聚类纯度(Purity)和兰德指数(Rand Index，RI)等聚类评价指标，具体计算公式如下

$$\mathrm{Purity}(\Omega, C) = \frac{1}{N}\sum_{k=1}^{K}\max_{j}(\omega_k \cap c_j),$$

$$\mathrm{Accuracy} = \frac{N_{cor}}{N},$$

$$\mathrm{RI} = \frac{a+b}{C_N^2},$$

其中，N为样本量，$\Omega = \{\omega_1, , \omega_K\}$为实际聚类效果，第$k$个聚类簇为$\omega_k$，$C = \{c_1, \cdots, c_K\}$为真实类别，$c_j(j = 1, 2, \cdots, K)$为真实类别的第$j$类，$K$为聚类数，$N_{cor}$为正确聚类的样本个数。真实类别与实际聚类结果中都属于同一类的元素对数为a，真实类别与实际聚类结果中都不属于同一类的元素对数为b，C_N^2为数据集中两两可以组成的对数。上述三类指标均在$[0, 1]$区间取值，且值越大，意味着聚类效果越好。表 4.4 为对径流量数据进行 50 次实验所得到的三类评价指标的均值。

表 4.4 径流量数据的聚类评价结果

评价指标 / 方法	聚类纯度 (Purity)	聚类精度 (Accuracy)	兰德指数 (Rand Index)
TA	0.786	0.786	0.268
FCOF	0.756	0.756	0.238
PCPDFCM	**0.807**	**0.807**	**0.269**

注：粗体表示比较结果为优。

由表 4.4 可以看出，本文所提 PCPDFCM 方法的三类评价指标均高于另外两

种聚类算法，意味着基于拟合曲线本身以及曲线波动特征两个视角构建的PCPDFCM 方法具有较好的聚类效果，能够为黄河流域水文研究提供一定的理论基础。

4.4 本章小结

考虑到函数型径流量数据曲线本身的特征，以及其受水文站地势、流域面积及气候等因素影响，基于非负矩阵分解思想，引入函数型主成分和主微分，构建基于主成分和主微分的函数型聚类方法(PCPDFCM)；然后，利用函数型主成分分析探究黄河干流各水文站不同年份径流量的时空分布特征；最后基于所提PCPDFCM 方法对黄河干流主要水文站的径流量聚类，研究各水文站不同年代径流量的异同情况，并通过 ArcGIS 软件对黄河干流径流量的空间分布特征进行可视化展示。

5 基于函数型岭回归模型的黄河干流径流量预测

5.1 函数型岭回归模型的构建

精准开展径流量预测，为流域水资源综合高效调配和防洪减灾调度提供一定的理论依据。考虑到函数型主成分代表函数曲线的绝大部分信息，而函数型主微分旨在刻画函数曲线的主要变化特征，如曲线的变化趋势、梯度以及曲率等特征。同时考虑主微分与主成分两个视角，将有助于提高函数型回归模型的预测精度。因此，构建基于函数型主微分与主成分的岭回归模型(FPDPCRR)，在两个互补视角中估计预测值，并自加权的调节视角权重，获得最终预测结果。

类似于 FPCA 中主成分得分，定义函数型主微分得分(Jang 和 Lim，2021)。$x_i(k)$的第k个 PDA 得分定义为

$$S_{ik}=\sum_{j=1}^{T}(x_i(t_j)-\bar{x}(t_j))\Phi_k(t_j),\ i=1,\cdots,N,k=1,\cdots,m, \tag{5.1}$$

其中，$\bar{x}(t)$为均值曲线，特征函数$\Phi_k(t_j)_{k=1}^m$通过求解微分方程来估计(Dalla 等，2014)。由式(2.10)计算得到$\hat{\beta}(t)$，再利用特征多项式$\lambda^m+\hat{\beta}_{m-1}\lambda^{m-1}+\cdots+\hat{\beta}_1\lambda+\hat{\beta}_0=0$得到复根$\{\hat{\lambda}_k\}_{k=1}^m$，从而特征函数的估计为$\hat{\Phi}_1(t)=e^{\hat{\lambda}_1 t},\cdots,\hat{\Phi}_K(t)=e^{\hat{\lambda}_m t}$。

对主微分得分的定义式(5.1)变形处理，得到函数型预测因子和函数型响应的主微分分解为

$$\begin{aligned}x_{ij}(t)&=\bar{x}_j(t)+\sum_{l=1}^{n-1}S_{il}^{x_j}(t)(\Phi_l^{x_j}(t))^{-1},\\ y_i(t)&=\bar{y}(t)+\sum_{l=1}^{n-1}S_{il}^{y}(t)(\Phi_l^{y}(t))^{-1}.\end{aligned} \tag{5.2}$$

式(5.2)中给出的主微分分解将式(2.11)转化为如下线性回归模型

$$S_{ik}^{y}=\sum_{j=1}^{J}\sum_{l=1}^{n-1}b_{2kl}^{x_j}S_{il}^{x_j}+\epsilon_{ik},\quad i=1,\cdots,n;k=1,\cdots,n-1. \tag{5.3}$$

受函数型响应的多元函数型主成分回归模型(Acal 等，2021)启发，引入函数型主微分分析结果，FPDPCRR 方法整合两个视角的估计值，以产生最终预测结果为

$$\begin{aligned}\hat{y}_i(t) &= \bar{y}(t) + w_1 \sum_{k=1}^{K} \hat{\xi}_{ik}^{y} f_k^{y}(t) + w_2 \sum_{k=1}^{K} \hat{S}_{ik}^{y} (\Phi_k^{y}(t))^{-1} \\ &= \bar{y}(t) + w_1 \sum_{k=1}^{K} (\sum_{j=1}^{J} \sum_{l \in L_{kj}} \hat{b}_{1kl}^{x_j} \xi_{il}^{x_j}) f_k^{y}(t) \\ &\quad + w_2 \sum_{k=1}^{K} (\sum_{j=1}^{J} \sum_{l \in L_{kj}} \hat{b}_{2kl}^{x_j} S_{il}^{x_j}) (\Phi_k^{y}(t))^{-1},\end{aligned} \tag{5.4}$$

其中，变量$\hat{y}_i(t)$表示预测值。$\hat{b}_{2kl}^{x_j}$为回归系数b_{2kl}的线性最小二乘估计，w_1、w_2为分配给两个视角的权重，且满足$w_1 + w_2 = 1$。

FPDPCRR 方法的损失函数为

$$\begin{aligned}\text{loss function} = &\sum_{i=1}^{n} [y_i(t) - (\bar{y}(t) + w_1 \sum_{k=1}^{K} \hat{\xi}_{ik}^{y} f_k^{y}(t) + w_2 \sum_{k=1}^{K} \hat{S}_{ik}^{y} (\Phi_k^{y}(t))^{-1})]^2 \\ &+ \gamma(w_1^2 + w_2^2),\end{aligned}$$

其中，γ是正则化参数，l_2正则化项有助于提高模型的泛化能力。

考虑当所有预测变量X_j都能完整观测到，只有响应变量Y存在缺失的情况。假设样本中前n个响应变量均具观测值，而缺少最后m个值，即所有响应变量具有n条完整的观测曲线和m条存在缺失的观测值。为估计存在缺失的响应曲线，用响应变量和预测变量完整的n个样本曲线估计式(2.13)中的参数b_{1kl}和式(5.3)中的b_{2kl}；计算预测因子的主成分得分$\{\xi_{il}^{x_j} : i = n+1, \cdots, n+m, l = 1, \cdots, n-1\}$以及由式(5.1)给出的预测因子的主微分得分$\{S_{il}^{x_j} : i = n+1, \cdots, n+m, l = 1, \cdots, n-1\}$，并将其代入式(5.4)，用以估计缺失的响应曲线$\{y_i^{miss}(t) : i = n+1, \cdots, n+m\}$。则估计的 FPDPCRR 方法可用于预测测试样本上新的响应值Y，并准确解释预测值和响应变量之间的关系。此外，还可以通过估计回归模型式(5.4)，预测未来区间上的响应变量，即已知区间$[0, T]$中的预测变量$(X_1(t), \cdots, X_J(t), Y(t))$，预测未来区间$[T, T+k]$上的响应变量$Y(t)$。

特别地，式(5.4)中当$\omega_2 = 0$时，模型退化为函数型响应的多元函数型主成分回归模型(MFPCR)；而当$\omega_1 = 0$时，模型退化为函数型响应的多元函数型主微分回归模型(Multivariate Functional Principal Differential Regression Model，MFPDR)

$$\begin{aligned}\hat{y}_i(t) &= \bar{y}(t) + \sum_{k=1}^{K} \hat{S}_{ik}^{y} (\Phi_k^{y}(t))^{-1} \\ &= \bar{y}(t) + \sum_{k=1}^{K} (\sum_{j=1}^{J} \sum_{l \in L_{kj}} \hat{b}_{2kl}^{x_j} S_{il}^{x_j}) (\Phi_k^{y}(t))^{-1}.\end{aligned}$$

然而，由于 MFPCR 方法仅考虑函数曲线的绝大多数信息，而 MFPDR 方法仅考虑函数曲线的导数等波动信息，理论上它们的预测插补性能次于 FPDPCRR 方法。为此，下面运用实例检验来证明上述推断。

5.2 黄河干流径流量变化的影响因素分析

5.2.1 黄河干流径流量变化的影响因素

2020 年 3 月 18 日中国气象局气象宣传与科普中心发布的《气候变化和人类活动影响黄河流域水循环》中提到，气候变化与人类活动是影响黄河流域径流量的两大关键因素。在全球变暖背景下，流域内气候发生改变，进而导致径流量也发生变化。此外，城市化、退耕还林、农田灌溉等人类活动也是影响黄河流域径流量变化的重要原因。下面利用累积量斜率变化分析法(赵益平等，2019)定量计算气候变化和人类活动对径流变化的影响。因此，下面对累积量斜率变化分析法做简单介绍。

首先，利用前文径流量突变特征，将实测径流量依据时间顺序划分为d个阶段，其中设第一个时段为基准期(A 时段)，剩余$d-1$个时段为影响期(B 时段)。利用线性方程

$$y = St + a \tag{5.5}$$

建立累积径流量、累积降水量和累积气温(y)与年份(t)的关系式。式(5.5)中S为斜率，a为截距。

其次，分别定义累积径流量、累积降水量和累积气温的斜率变化率R_{SR}、R_{SP}和R_{ST}，即

$$\begin{aligned} R_{SR} &= \frac{S_{Rb} - S_{Ra}}{S_{Ra}} \times 100\%, \\ R_{SP} &= \frac{S_{Pb} - S_{Pa}}{S_{Pa}} \times 100\%, \\ R_{ST} &= \frac{S_{Tb} - S_{Ta}}{S_{Ta}} \times 100\%, \end{aligned} \tag{5.6}$$

其中，S_{Ra}、S_{Pa}、S_{Ta}和S_{Rb}、S_{Pb}、S_{Tb}分别为 a 和 b 时段累积径流量、累积降水量、累积气温与年份的关系式中的斜率。

事实上，降水和气温与径流量分别呈现正、负相关关系，则降水、气温分别

对径流量变化的贡献率C_P、C_T定义为

$$\begin{aligned} C_P &= \frac{R_{SP}}{R_{SR}} \times 100\%, \\ C_T &= \frac{R_{ST}}{R_{SR}} \times 100\%. \end{aligned} \tag{5.7}$$

综合水、热等因素，可知气候变化对径流量变化的贡献率C_C，即

$$C_C = C_P + C_T. \tag{5.8}$$

一般地，人类活动和气候变化对径流量变化的影响较大，其他因素可忽略。则人类活动对径流量变化的贡献率C_H为

$$C_H = 1 - C_C. \tag{5.9}$$

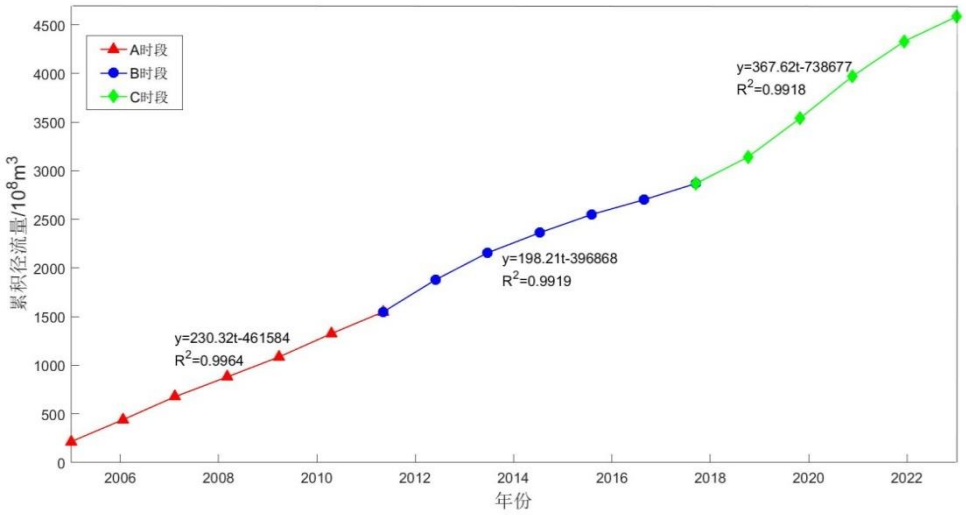
根据上文中实测径流量的突变性和周期性等特征，将黄河干流年实测径流量序列分为三个时段，以 2005-2011 年(A 时段)为基准期，2012-2017 年(B 时段)和 2017-2022 年(C 时段)为影响期。基准期内气候变化为影响径流量变化的主要因素，而随着人类活动的加剧，影响期内径流量受人类活动和气候变化的双重影响。表 5.1 为三个时段的实测径流量、降水量和气温的均值和均值变化率，可以看出径流量和降水量均在 2018 年前后发生显著变化，而气温在 2012 年前后发生显著变化。

表 5.1　三个时段黄河干流年均实测径流量、降水量和气温的均值及其变化率

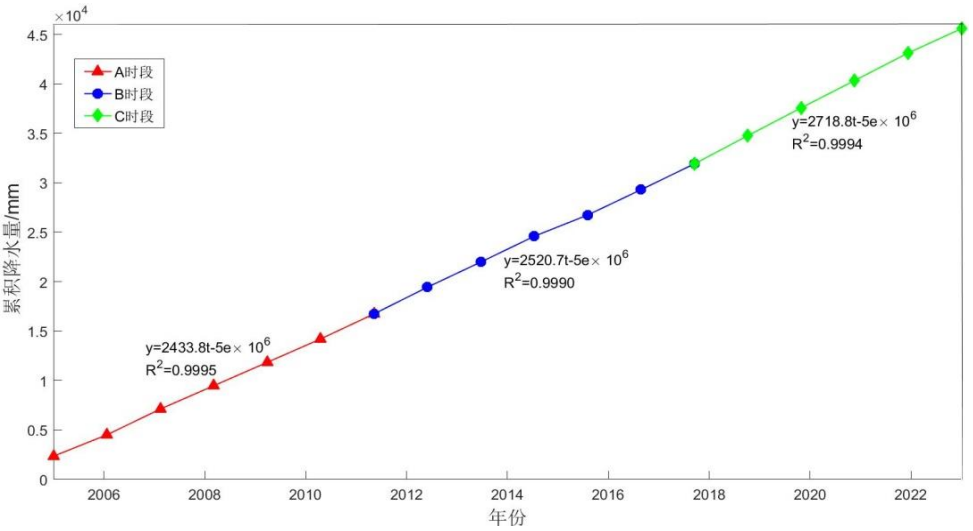
水文变量	均值$10^8 m^3 \cdot mm^{-1} \cdot ^\circ C^{-1}$			均值变化率/%	
	A 时段	B 时段	C 时段	2012 年	2018 年
实测径流量	221.38	219.84	342.96	-0.7	56
降水量	2389.84	2530.26	2734.57	5.9	8.1
气温	11.45	11.77	11.98	2.3	1.8

实测径流量、降水量、气温的累积曲线如图 5.1 所示，对累积曲线的三个时段建立累积量与年份的线性关系，图 5.1 中展示了具体的拟合方程。

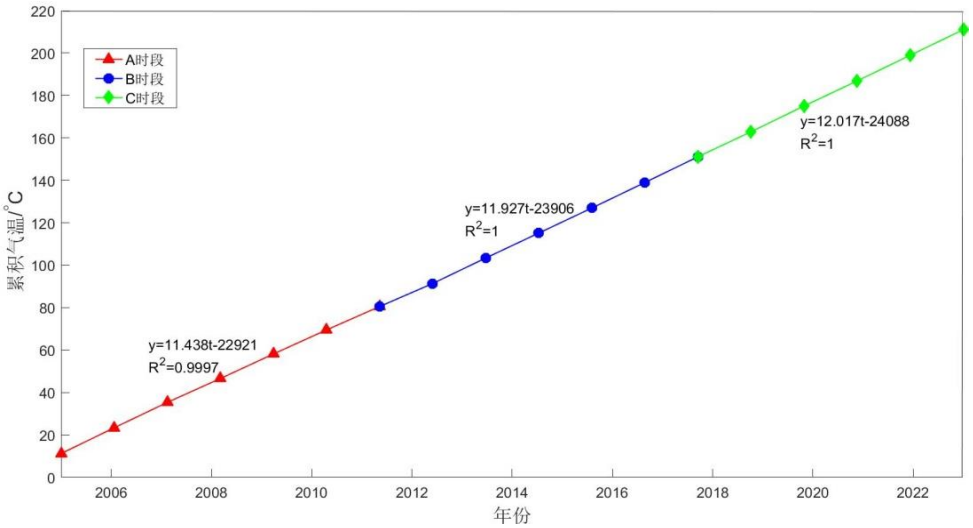
表 5.2 为由式(5.6)计算的各累积量在影响期的斜率变化率(R_{SR}、R_{SP}、R_{ST})。可以看出，三个序列在 B 时段和 C 时段的斜率变化率均有明显的变化特征。其中，B 时段累积径流量的斜率变化率为负，而 C 时段其斜率变化率为正，说明径流量具有先下降后上升的趋势；而累积降水量和气温在 B 时段和 C 时段的斜率变化率均为正，且 B 时段小于 C 时段，说明降水量和气温均有持续上升的趋势。



(a)累积径流量



(b)累积降水量



(c)累积气温

图 5.1 实测径流量、降水量、气温的累积曲线

表 5.2 各影响期累积径流量、累积降水量、累积气温序列斜率变化率

时段		A 时段	B 时段	C 时段
累积径流量	斜率(单位：$10^8m^3/a$)	230.32	198.21	367.62
	变化量(单位：$10^8m^3/a$)		-32.05	137.30
	变化率(单位：%)		-14.91	59.61
累积降水量	斜率(单位：mm/a)	2433.80	2520.70	2718.80
	变化量(单位：mm/a)		86.90	285.00
	变化率(单位：%)		3.60	11.70
累积气温	斜率(单位：$°C/a$)	11.44	11.93	12.02
	变化量(单位：$°C/a$)		0.49	0.58
	变化率(单位：%)		4.28	5.07

由式(5.7)、式(5.8)和式(5.9)计算得到的气候变化和人类活动对径流量变化的贡献率如表 5.3 所示。可以看出，气候变化对径流量的贡献率由 B 时段的 52.85%下降到 C 时段的 28.14%，而人类活动对径流量的贡献率由 B 时段的 47.15%上升到 C 时段的 71.86%。这表明近年来人类活动对黄河干流径流量的影响愈加剧烈。

表 5.3 气候变化和人类活动对径流量变化的贡献率

计算方案	仅考虑降水/%		综合考虑降水和气温/%	
	B 时段	C 时段	B 时段	C 时段
降水	24.14	19.63	24.14	19.63
气温			28.71	8.51
气候变化	24.14	19.63	52.85	28.14
人类活动	75.86	80.37	47.15	71.86

此外，2005-2022 年黄河干流主要水文站年均天然径流量与实测径流量如图 5.2 所示，其中天然径流量=实测径流量+地表水耗水量+蓄变量。图 5.2 表明，黄河干流天然径流量自上而下呈上升趋势，而实测径流量无显著增大，甚至在兰州和利津两处呈下降趋势。此外，天然径流量与实测径流量的差值(地表水耗水量+蓄变量)自兰州站开始逐渐增大，至利津水文站时，实测径流量仅约占天然径流量的三分之一。2005-2022 年各水文站平均耗水量和蓄变量如表 5.4 所示。表 5.4 中，黄河干流各水文站自上而下耗水量不断增加，位于上游的兰州站最少，为 $24.94\times10^8m^3$；而位于下游的利津站耗水量最大，高达$298.72\times10^8m^3$。耗水量的逐渐增大导致天然径流量与实测径流量的差值逐渐增大。

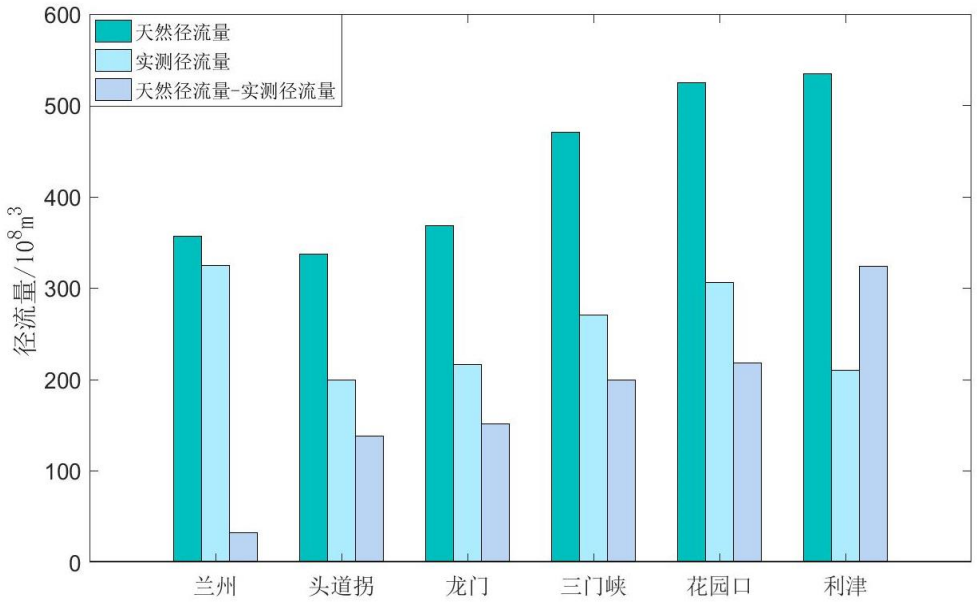


图 5.2 2005-2022 年天然径流量与实测径流量对比

表 5.4 2005-2022 年各站断面以上平均耗水量

	兰州	头道拐	龙门	三门峡	花园口	利津
耗水量/10^8m^3	24.94	131.90	141.74	184.96	202.87	298.72
蓄变量/10^8m^3	17.16	0.94	17.68	16.49	22.86	21.77

考虑图 5.3、图 5.4 中各分区不同领域平均耗水情况，分析人类活动如何影响径流量。从图 5.3 可以直观地看出，2005-2019 年各分区的平均耗水量最高为花园口以下区域、兰州至头道拐段次之。不同领域中耗水最大项均为农田灌溉，花园口以下区域多年平均农田灌溉耗水量高达$96.768 \times 10^8m^3$，这与河南、山东的农业发展密切相关；兰州至头道拐段的农田灌溉耗水量为$82.702 \times 10^8m^3$，其宁夏和内蒙古灌溉区属于中国农业耗水大区。同时，兰州至头道拐段的林牧渔畜耗水量在各分区中居于最大，约为$8.228 \times 10^8m^3$；对于工业耗水量而言，花园口以下区域最多，为$11.306 \times 10^8m^3$，且该区域的城镇公共、居民生活以及生态环境耗水量在各分区中均为最大值，分别为$2.06 \times 10^8m^3$、$6.608 \times 10^8m^3$和$8.396 \times 10^8m^3$。2020-2022 年各区各行业的平均耗水量如图 5.4 所示。图 5.4 表明，近两年平均耗水量排名前二的两个分区依次为兰州—头道拐段和花园口以下，分别达到了$28.48 \times 10^8m^3$和$29.34 \times 10^8m^3$。与 2014-2019 各分区不同领域平均耗水情况类似，不同领域中耗水最多的均为农业，兰州至头道拐段农业耗水量高达$91.08 \times 10^8m^3$，花园口以下区域为$60.79 \times 10^8m^3$，同时，花园口以下区域的工业、生活以及生态环境的耗水量最大，分别为$11.025 \times 10^8m^3$、$21.295 \times 10^8m^3$和

$24.23 \times 10^8 m^3$。

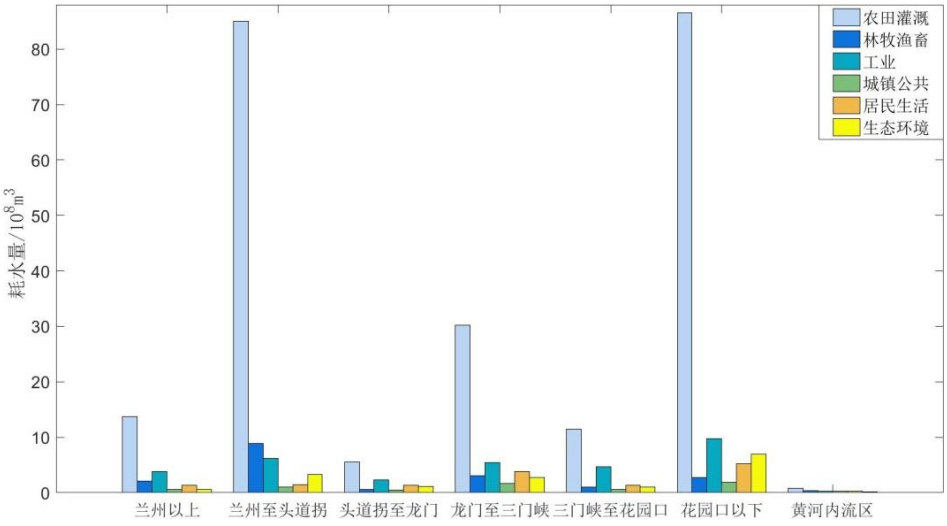


图 5.3 2005-2019 年黄河流域各区各行业平均耗水情况

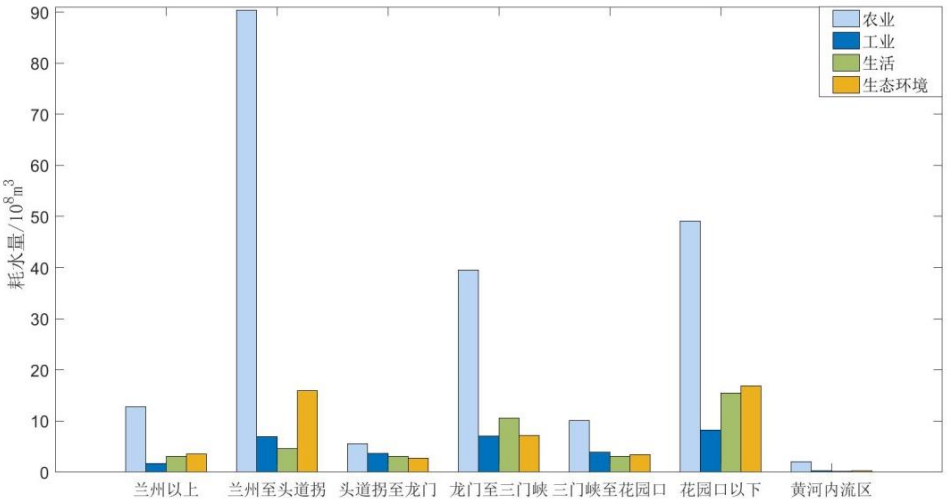


图 5.4 2020-2022 年黄河流域各区各行业平均耗水情况

因此，工农业耗水、土地利用变化等人类活动中，耗水量增加对径流量影响较大，土地利用变化对径流量变化的影响较小。此外，文献(Han 等，2023)表明，在气候变化方面，降水和气温对径流量变化有一定的影响，降水增加使得径流量增加，同时，气温升高致使蒸发量增大，从而导致径流量减少。

5.2.2 气候变化和人类活动与径流量的响应关系

为进一步考察实测径流量与气候变化和人类活动之间的关系，充分考虑水文数据的函数变化特征，利用多元函数型线性模型，估计降水、气温以及耗水量与

实测径流量之间的响应关系，自变量选用兰州、头道拐、龙门、三门峡、花园口、利津等六个代表水文站的降水量($x_1(t)$)、气温($x_2(t)$)、耗水量($x_3(t)$)，实测径流量变化曲线作为响应变量，建立实测径流量与降水、气温和耗水量之间的多元函数型线性模型

$$y_i(t) = \alpha_0(t) + \sum_{j=1}^{3} \int_0^T x_{ij}(s)\alpha_j(s,t)ds + \epsilon_i(t),\ i = 1, \cdots, 6, \tag{5.10}$$

其中，$\alpha_0(t)$为截距函数，$\alpha_j(t)$为待估回归系数函数，$\epsilon(t)$为误差项。样本曲线$\epsilon(t)$的变动区间为$[0,T]$，样本量为 6。估计模型公式(5.10)中截距函数$\alpha_0(t)$和斜率函数$\alpha_1(t)$、$\alpha_2(t)$以及$\alpha_3(t)$的估计曲线，如图 5.5 所示。

从图 5.5 可以看出降水量、气温以及耗水量对实测径流量的影响效应，在不同的时间点体现出不同的特征。事实上，截距函数$\alpha_0(t)$的趋势模拟了 6 个水文站实测径流量的平均变化趋势，表明实测径流量在 2005-2006 年春季下降，之后至 2012 年末有所增加，2013-2016 年迅速下降至低谷，2016-2019 年又稳步上升，2019 年之后又开始下降。回归系数函数$\alpha_1(t)$、$\alpha_2(t)$和$\alpha_3(t)$分别表示实测径流量受降水量、气温以及耗水量的影响，从$\alpha_1(t)$、$\alpha_2(t)$和$\alpha_3(t)$的波动幅度可以看出，气候变化和人类活动对实测径流量的影响较大。$\alpha_1(t)$为正值，表明降水量对实测径流量具有正向的影响，降水量与实测径流量的变化趋势基本一致，为黄河干流实测径流量的主要影响因素；而$\alpha_2(t)$和$\alpha_3(t)$在样本区间上为负值，表明气温和耗水量对实测径流量具有负向影响。2005-2007 年$\alpha_1(t)$呈下降趋势，之后又上升至 2011 年，$\alpha_2(t)$和$\alpha_3(t)$与$\alpha_1(t)$变化趋势相反；2011-2015 年$\alpha_1(t)$呈下降趋势且在 2015 年降至低谷，而$\alpha_2(t)$和$\alpha_3(t)$呈增长趋势且其在 2015 年达绝对值最小，表明这一时期降水量、气温以及耗水量对实测径流量的影响效应较小；2015-2020 年$\alpha_1(t)$呈增加趋势，$\alpha_2(t)$和$\alpha_3(t)$呈下降趋势，且在 2020 年$\alpha_1(t)$、$\alpha_2(t)$和$\alpha_3(t)$的绝对值最大，意味着此阶段实测径流量受降水量的正向影响以及气温和耗水量的负向影响效应较大。因此，可以得出，实测径流量受到气候变化和人类活动的显著影响，且降水量对其具有正向影响，而气温和耗水量对其有负向影响。

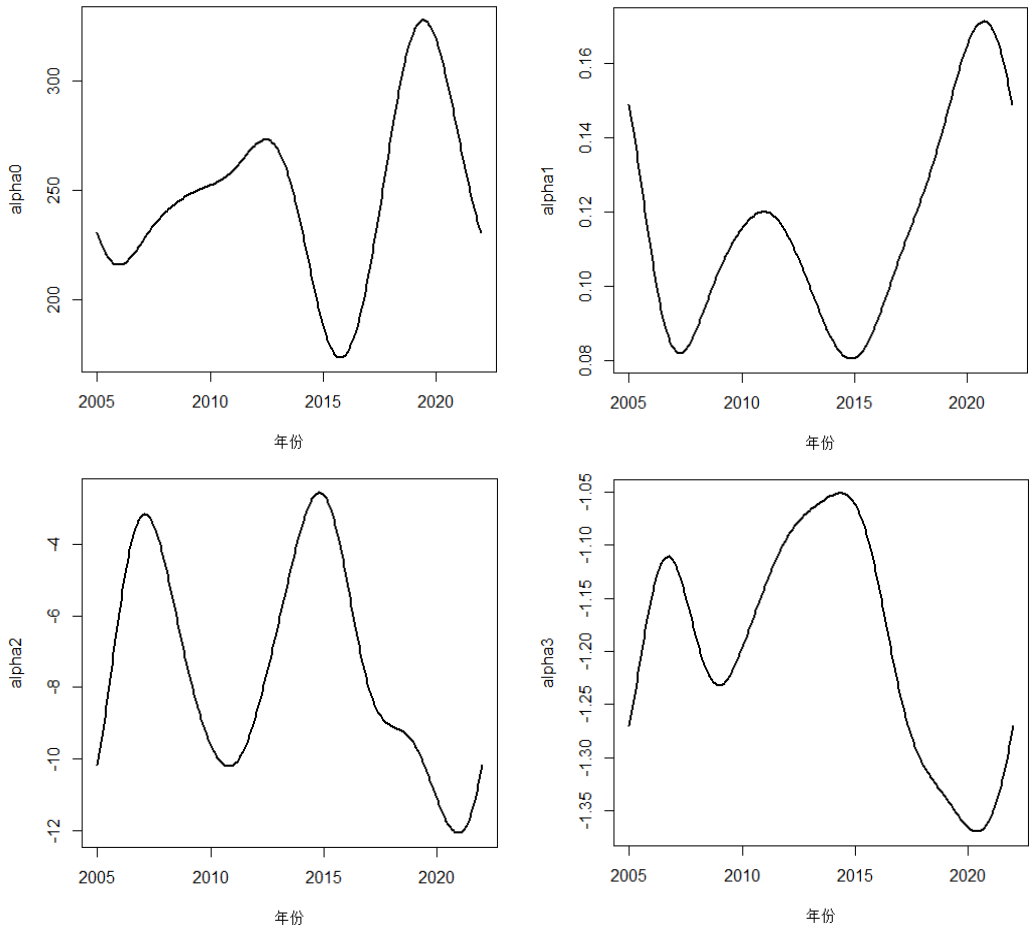


图 5.5 $\alpha_0(t)$、$\alpha_1(t)$、$\alpha_2(t)$和$\alpha_3(t)$的估计曲线

表 5.5 模型调整的判定系数R^2

年份	调整的R^2	年份	调整的R^2	年份	调整的R^2
2005	0.7861671	2011	0.8414571	2017	0.8820893
2006	0.6937146	2012	0.8139283	2018	0.9022508
2007	0.7712426	2013	0.8154313	2019	0.8780305
2008	0.8605297	2014	0.8129338	2020	0.8590318
2009	0.8887355	2015	0.8005590	2021	0.8533534
2010	0.8746862	2016	0.8273565	2022	0.7861671

对上述建立的函数型线性回归模型进行检验，以确定降水量、气温以及耗水量与实测径流量之间的关系是否显著。实际上，表 5.5 中模型的调整的判定系数R^2接近于 1，说明模型的解释力度强。因此，若想控制黄河干流实测径流量，必须以气候变化(如降水量、气温)以及人类活动的耗水量为切入点；同时，以降水量、气温以及耗水量为预测因子预测实测径流量，模型预测精度较高，可为黄河流域水资源的合理配置及利用提供理论基础。

5.3 基于函数型岭回归模型的径流量预测

通过上文对径流量变化的影响因素研究，并充分考虑黄河干流径流资料和气象资料的函数型特征，以函数型数据的角度出发，从数据曲线特征及其波动特征信息两个互补视角中估计，动态地学习分配给两视角不同的权重，以预测一定时期内的径流量。下面基于 FPDPCRR 方法处理实测径流量的缺失插补预测问题。具体地，首先运用 4 个完整数据的水文站估计 FPDPCRR 方法，然后对两个不完整数据的水文站的实测径流量进行插补预测。

首先，对 6 个水文站的实测径流量、降水量、气温和耗水量 4 个变量进行曲线拟合并修匀，分别表示为$Y(t)$、$X_1(t)$、$X_2(t)$和$X_3(t)$。本文选取 4 阶 B-样条基函数拟合曲线，修匀参数取$\lambda = 1e-3$。

其次，估计 4 个函数型变量的每个函数型主成分，结果表明第一主成分解释了 4 个变量中的绝大多信息，即$Y(t)$、$X_1(t)$、$X_2(t)$和$X_3(t)$分别为 77.9%、99.8%、99.7%和 99.8%。

然后，基于主微分分析思想，选取二阶微分方程对$Y(t)$、$X_1(t)$、$X_2(t)$和$X_3(t)$进行主微分分析，并基于式(5.1)得到各个变量的两个主微分得分。之后，将 4 个完整数据的地区作为预测样本，分别考虑预测变量和响应变量之间的主微分得分、第一主成分得分的相关性。因此，根据每个预测变量的主微分得分和第一主成分，将函数型线性回归模型简化为主微分回归模型

$$\hat{S}_{ij}^{y} = \gamma_0 + S_{ij}^{x_1}\gamma_1^{y} + S_{ij}^{x_2}\gamma_2^{y} + S_{ij}^{x_3}\gamma_3^{y} + \epsilon_i^{y},$$

以及第一主成分回归模型

$$\hat{\xi}_{i1}^{y} = \gamma_0 + \xi_{i1}^{x_1}\gamma_1^{y} + \xi_{i1}^{x_2}\gamma_2^{y} + \xi_{i1}^{x_3}\gamma_3^{y} + \epsilon_i^{y},$$

其中$i = 1, \cdots, 6; j = 1, 2$。上述模型可以基于$X_1(t)$、$X_2(t)$和$X_3(t)$的主微分得分和第一主成分分别准确估计$Y(t)$的主微分得分和第一主成分，其决定系数见表 5.6。显然，第一主微分得分的解释性优于第二主微分得分。因此，选取第一主微分得分和第一主成分进行函数型数据预测。

表 5.6 $X_1(t)$、$X_2(t)$、$X_3(t)$**与**$Y(t)$**的决定系数**

变量	第一主微分得分	第二主微分得分	第一主成分得分
实测径流量	0.8940	0.8182	0.8211

根据 Karhunen-Loève 展开式$y(t) = \bar{y}(t) + \sum_{m=1}^{\infty}\xi_m f_m(t)$和 FPDPCRR 方法，

有

$$\hat{y}_i(t) = \bar{y}(t) + w_1\hat{\xi}_{i1}^{y} f_1^{y}(t) + w_2\hat{S}_{i2}^{y}(\Phi_2^{y}(t))^{-1},\ i = 1, \cdots, 6. \tag{5.11}$$

为展示所提 FPDPCRR 方法的插补预测效果，评价指标采用均方根误差(RMSE)、平均绝对误差(MAE)以及平均绝对百分比误差(MAPE)。具体公式为

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|,$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

其中，y_i为真实值，$\hat{y}_i$为数据预测值，n为样本的数量。为对比 FPDPCRR 方法、MFPDR 方法和 MFPCR 方法的插补预测性能，计算 4 个完整数据地区的观测曲线与预测曲线之间的 RMSE、MAE 和 MAPE 值，所得评价指标结果如表 5.7 所示。

表 5.7 实测径流量曲线的预测效果评估对比

水文站		兰州	头道拐	龙门	三门峡	平均值
RMSE	FPDPCRR	0.344224	0.219245	0.12839	0.138052	**0.207478**
	MFPDR	0.504261	0.389359	0.257501	0.353689	0.376203
	MFPCR	0.264771	0.218474	0.234861	0.15716	0.218816
MAE	FPDPCRR	0.047874	0.03062	0.018205	0.018104	**0.028701**
	MFPDR	0.049094	0.037286	0.024482	0.034867	0.036432
	MFPCR	0.059945	0.048494	0.049561	0.053514	0.052879
MAPE	FPDPCRR	0.016637	0.020694	0.011238	0.006444	**0.013753**
	MFPDR	0.156859	0.203943	0.118981	0.140598	0.155095
	MFPCR	0.005006	0.023033	0.023033	0.009676	0.014074

注：粗体表示比较结果为优。

由表 5.7 可知，MFPDR 方法的评价指标大于 FPDPCRR 方法和 MFPCR 方法，表明其插补预测能力较差。个别地区 MFPCR 方法的预测效果略优于 FPDPCRR 方法的，但从平均结果来看，FPDPCRR 方法三个评价指标的均值均小于 MFPCR 方法，其预测效果优于 MFPCR 方法。这意味着引入主微分，从数据曲线及其波动特征信息两个互补视角中估计，可进一步提升插补预测性能。总的来说，三个模型的插补预测性能：MFPDR 方法 < MFPCR 方法 < FPDPCRR

方法。

此外，分别绘制由 FPDPCRR 方法插补预测的两个数据缺失水文站(花园口和利津)的实测径流量的预测曲线、观测曲线以及其置信区间，分别如图 5.6 所示，图中红色实线为观测曲线，绿色实线为预测曲线，黑色虚线表示置信带。

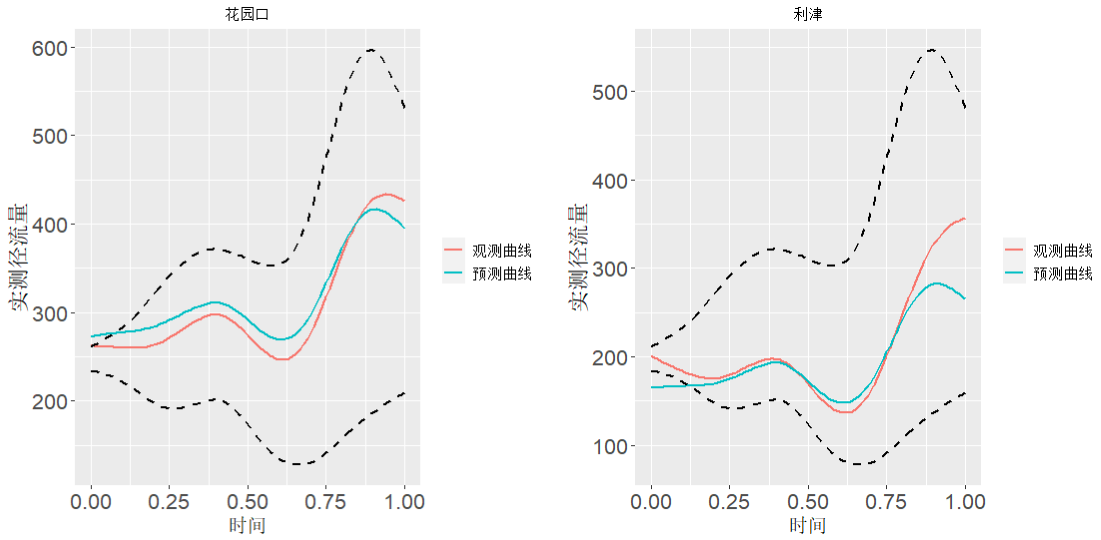


图 5.6 实测径流量的观测曲线、预测曲线及置信区间

考虑到 FPCA 能够代表原始数据的绝大多信息，PDA 既能反映曲线的变化趋势，又能挖掘数据的梯度、曲率等潜在信息，从这两个视角出发，针对响应变量存在缺失的函数型数据，构建 FPDPCRR 方法插补预测黄河干流实测径流量缺失问题。结果证实，与 MFPDR 方法和 MFPCR 方法相比，FPDPCRR 方法的插补预测效果最优。

5.4 本章小结

本章精准开展径流量预测，为流域水资源综合高效调配和防洪减灾调度提供一定的理论依据。首先，结合函数型主成分分析与主微分分析的思想原理，考虑数据曲线及波动特征两个视角，构建基于函数型主微分与主成分的岭回归模型(FPDPCRR)；其次，利用累积量斜率变化分析法定量评估径流变化的影响因素，并通过建立多元函数型线性回归模型估计气候变化和人类活动与径流量之间的响应关系，结果表明随着时间的变化，人类活动对黄河干流实测径流量的影响愈加强烈；最后，利用 FPDPCRR 方法预测径流量，结果表明所提 FPDPCRR 方法的预测效果较好。

6 总结与展望

6.1 总结

本文从函数型数据新视角出发，基于 2002-2022 年黄河干流径流量数据开展函数型数据分析方法研究，整体把握黄河流域径流变化趋势，为合理开发利用黄河流域水资源提供理论依据。主要有以下三项工作：

（1）基于拟合函数法、相平面图以及主微分分析研究黄河干流水文站实测径流量的演变特征，结果表明 2002-2022 年黄河干流实测径流量总体上呈增加趋势，周期性显著，且在 2004、2008、2013、2017 以及 2021 年发生突变。此外，微分方程能够充分刻画实测径流量的动态波动规律及季节变动特征。

（2）提出基于主成分和主微分的函数型聚类方法(PCPDFCM)，探究黄河干流径流量的时空分布特征。基于非负矩阵分解思想，引入函数型主成分和主微分，构建 PCPDFCM 方法；通过函数型主成分分析发现实测径流量主要在 2013 年前后和 2019-2021 年显著增大，从空间分布上看，黄河干流径流量整体自上而下逐渐减小，空间差异性明显；最后利用 PCPDFCM 方法将黄河干流 12 个水文站聚为三类，发现三类水文站在空间分布上特征明显，水文特征具有显著差异。进一步，利用 ArcGIS 软件可视化展示黄河干流 12 个水文站径流量的差异性特征。

（3）构建函数型岭回归模型，预测黄河干流径流量。首先结合函数型主成分分析与主微分分析的思想原理，构建基于函数型主微分与主成分的岭回归模型(FPDPCRR)；其次，利用累积量斜率变化分析法定量评估径流变化的影响因素，并建立气候变化和人类活动与径流量的函数型回归模型，发现随着时间的变化，人类活动对黄河干流径流量的影响愈加强烈；最后，利用 FPDPCRR 方法预测径流量，结果表明所提 FPDPCRR 方法的预测效果较好，可为流域水资源综合高效调配和防洪减灾调度提供一定的理论依据。

6.2 展望

尽管本文运用函数型数据分析方法对黄河干流径流演变特征、时空分布特征和预测进行了深入研究，但还是存在一些不足，未来可做一些改进：

（1）基于非负矩阵分解构建的 PCPDFCM 方法，应用过程中要求系数矩阵非负，应用范围局限，可考虑半非负矩阵分解，扩大应用范围；

（2）在分析径流量变化的影响因素时，仅考虑了气候变化中的降水量和气温以及人类活动中的耗水量，未对极端天气以及其他水利工程建设等做全面分析，存在一些不足，仍需进一步研究。

参考文献

[1] Abraham C. Unsupervised curve clustering using B-Splines［J］. Scandinavian Journal of Statistics. 2003,30(3): 581-595.

[2] Acal C, Escabias M, Aguilera A M, et al. COVID-19 data imputation by multiple Function-on-Function principal component regression[J]. Mathematics (Basel). 2021, 9(11): 1237.

[3] Avipsa R, Trisalyn N, Pavan T. Functional data analysis approach for mapping change in time series: A case study using bicycle ridership patterns[J]. Transportation Research Interdisciplinary Perspectives. 2023,17: 100752.

[4] Dalla Rosa M, Sangalli L M, Vantini S. Principal differential analysis of the Aneurisk65 data set[J]. Advances in Data Analysis and Classification. 2014,8(3): 287-302.

[5] Dijana T, Nino A F, Ivica K. A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering[J]. Pattern Recognition. 2018, 82:40-55.

[6] Ding C, He X F, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering[C] SIAM Internation Conference on Data Mining, Newport Beach, CA, 2005.

[7] Ding C H Q, Li T, Peng W, et al. Orthogonal nonnegative matrix T-Factorizations for clustering[J]. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006:126-135.

[8] Feng C, Gao M. An improved ARIMA method based on functional principal component analysis and bidirectional bootstrap and its application to stock price forecasting[J]. Academic Journal of Computing Information Science. 2022,5(10): 21-27.

[9] Greven S, Crainiceanu C, Caffo B, et al. Longitudinal functional principal component analysis[J]. Electronic Journal of Statistics. 2010, 4:1022-1054.

[10] Han Z, Zuo Q, Wang C, et al. Impacts of climate change on natural runoff in the Yellow River Basin of China during 1961-2020[J]. Water. 2023, 15(5): 929.

[11] Jang E, Lim Y. Classification via principal differential analysis[J]. Communications for Statistical Applications and Methods. 2021, 28(2): 135-150.

[12] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix

factorization[J]. Nature. 1999, 6755(401): 788-791.

[13] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]. International Conference on Neural Information Processing Systems. MIT Press, 2000.

[14] Lei X, Gao L, Wei J, et al. Contributions of climate change and human activities to runoff variations in the Poyang Lake Basin of China[J]. Physics and Chemistry of the Earth. Parts A/B/C. 2021, 123: 103019.

[15] Liang N, Yang Z, Li Z, et al. Semi-supervised multi-view clustering with graph-regularized partially shared non-negative matrix factorization[J]. Knowledge-Based Systems. 2020, 190:105185.

[16] Liang N ,Yang Z ,Li Z , et al. Multi-view clustering by non-negative matrix factorization with co-orthogonal constraints[J].Knowledge-Based Systems. 2020,194:105582-105582.

[17] Liu H, Wang Z, Ji G, et al. Quantifying the impacts of climate change and human activities on runoff in the Lancang River Basin based on the budyko hypothesis[J]. Water. 2020, 12(12):3501.

[18] Mengfei R, Yihe Y. Optimal estimation of large functional and longitudinal data by using functional linear mixed model[J].Mathematics.2022,10(22): 4322-4322.

[19] Qingting Q, Min L, Jinwu X. Dynamic prediction of multivariate functional data based on Functional Kernel Partial Least Squares[J].Journal of Process Control. 2022,116:273-285.

[20] Oshinubi K, Ibrahim F, Rachdi M, et al. Functional data analysis: Application to daily observation of COVID-19 prevalence in France[J]. AIMS Mathematics. 2022, 7(4):5347-5385.

[21] Ramsay J O. When the data are functions[J]. Psychometrika. 1982, 47(4): 379 -396.

[22] Ramsay J O, Dalzell C J. Some tools for functional data analysis[J]. Journal of the Royal Statistical Society. 1991, 53(3): 539-572.

[23] Ramsay J O. Principal differential analysis: data reduction by differential operators[J]. Journal of the Royal Statistical Society: Series B (Methodological). 1996, 58(3):495-508.

[24] Ramsay J O, Silverman B W. Functional data analysis[J]. International Encyclopedia of the Social and Behavioral Sciences. 2001, 50(2): 5822-5828.

[25] Ramsay J O, Silverman B W. Applied Functional Data Analysis: Methods and

Case Studies[M]. New York: Springer, 2002.

[26] Ramsay J O, Silverman B W. Functional Data Analysis[M]. Springer, 2005.

[27] Ferraty F, Vieu P. Nonparametric Functional Data Analysis: Theory and Practice [M]. New York: Springer, 2006: 112-118.

[28] Lundborg A R, Shah R D, Peters J. Conditional independence testing in Hilbert spaces with applications to functional data analysis[J].Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2022,84(5):1821-1850.

[29] Tucci C E M, Clarke R T, Collischonn W, et al. Long-term flow forecasts based on climate and hydrologic modeling: Uruguay River basin[J]. Water Resources Research. 2003, 39(7):1181.

[30] Vinué G, Epifanio I. Forecasting basketball players' performance using sparse functional data[J].Statistical Analysis and Data Mining: The ASA Data Science Journal. 2019,12(6):534-547.

[31] Xie K, Liu P, Zhang J, et al. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships[J]. Journal of Hydrology. 2021, 603(C):127043.

[32] 丁辉, 许文超, 朱汉兵, 等. 函数型数据回归分析综述[J]. 应用概率统计. 2018, 34(06): 630-654.

[33] 高海燕, 黄恒君, 王宇辰. 基于非负矩阵分解的函数型聚类算法 [J]. 统计研究. 2020, 37 (08): 91-103.

[34] 高海燕, 刘万金, 黄恒君. 鲁棒自适应对称非负矩阵分解聚类算法[J]. 计算机应用研究. 2023,40(04):1024-1029.

[35] 黄恒君, 高海燕, 张梦瑶. 函数型聚类分析: 基于距离的一步法框架[J]. 数理统计与管理. 2019,38(6) : 986-995.

[36] 黄亚, 易灵, 肖伟华, 等. 基于高斯过程回归模型的径流短期预测研究[J]. 水力发电. 2020, 46(12): 9-12.

[37] 胡作龙, 高鹏. 基于EEMD-SVM模型的北洛河上游径流预测[J]. 水土保持研究. 2023, 30(04): 98-102.

[38] 梁浩, 黄生志, 孟二浩, 等. 基于多种混合模型的径流预测研究[J]. 水利学报. 2020, 51(1): 112-125.

[39] 马骏, 王文. 若干水文预报方法综述[J]. 水利水电科技进展. 2005, 25(1): 56-60.

[40] 马柱国. 黄河径流量的历史演变规律及成因[J]. 地球物理学报. 2005, 48(6): 1270-1275.

[41] 苏贤保, 李勋贵, 王义鹏, 等. 多时间尺度下黄河上游径流复杂度变化特征研究[J]. 水资源与水工程学报. 2021, 32(5): 1-10.

[42] 苏栝芳, 李气芳, 陈美源. 基于残差函数主成分的相依函数型回归模型估计及金融应用[J]. 数量经济技术经济研究. 2022, 39(12): 195-213.

[43] 王德青, 刘宵, 王许, 等. 中国实时金融状况指数的另一种测度——基于函数型数据分析方法[J]. 金融发展研究. 2021(10): 14-22.

[44] 王德青, 朱建平, 刘晓葳, 等. 函数型数据聚类分析研究综述与展望[J]. 数理统计与管理. 2018, 37(01): 51-63.

[45] 闻昕, 陈然, 谭乔凤, 等. 基于多因素相似性的融雪径流预报方法研究[J]. 水力发电学报. 2022, 41(03): 46-59.

[46] 武祺然, 周力凯, 孙金金, 等.浙江省空气质量变化特征研究——基于函数型数据分析[J].山东大学学报(理学版). 2021,56(07):53-64.

[47] 于海超, 张扬, 马金珠, 等. 1969-2018年黄河实测径流与天然径流的变化[J]. 水土保持通报. 2020, 40(05): 1-7.

[48] 严明义. 函数性数据的统计分析: 思想、方法和应用[J]. 统计研究. 2007, 20(02): 87-94.

[49] 严明义, 杜鹏. 中国消费价格指数季节变动的函数性数据分析[J]. 统计与信息论坛. 2010, 25(08): 100-106.

[50] 严明义, 蒲泾泾, 严康. 函数性数据的微分方程分析方法及经济应用[J]. 统计与信息论坛. 2013, 28(8): 14-20.

[51] 剡亮亮. 基于函数性视角的经济数据分析——以主微分分析方法为例[J]. 统计与信息论坛. 2013, 28(01): 40-46.

[52] 姚晓红, 高海燕, 吕家奇, 等. 一种基于多视角学习的多元函数型聚类方法[J]. 数理统计与管理. 2022, 41(04): 689-702.

[53] 朱佳, 冯峥晖, 陈正宇. 基于函数型数据分析和广义分位数的PM2.5数据探究[J].数理统计与管理. 2021, 40(05): 771-784.

[54] 赵益平, 王文圣, 张丹, 等. 累积量斜率变化分析法及其在径流变化归因中的应用[J]. 水电能源科学. 2019, 37 (10): 17-20.

[55] 张崇岐, 赵娜, 孔丹. 函数数据分析新进展[J]. 广州大学学报(自然科学版), 2006, 12(03): 1-4.

[56] 张亚丽, 王栋华, 田义超, 等. 广西北部湾入海流域径流演变特征及其对气候变化和人类活动的响应[J]. 广西科学. 2022, 29(05): 971-983.

[57] 朱冉, 徐礼文, 孟庆海. 基于函数型数据分析算法的电阻性本质安全电路电弧放电模型[J]. 电工技术学报. 2023:1-14.

攻读硕士学位期间承担的科研任务及主要成果

已发表的论文：

[1]高海燕, **张悦**. 中国生产者价格指数季节变动特征与区域差异分析[J]. 统计与决策. 2023, 39 (19): 34-40.

[2]高海燕, **张悦**. 基于函数型主微分与主成分的岭回归模型与应用[J]. 兰州文理学院学报(自然科学版). 2023, 37 (06): 17-24.

[3]高海燕, 马文娟, 李唯欣, **张悦**. 稀疏空气质量函数型数据插补方法实证研究 [J]. 河北环境工程学院学报. 2023, 33 (05): 73-82.

[4]高海燕, **张悦**. 黄河干流实测径流量演变特征及影响因素分析[J]. 水土保持研究.(已录用)

参与科研项目：

（1）参与国家社会科学基金项目：大规模稀疏函数型数据修复方法与应用研究(19XTJ002)。

（2）参与完成甘肃省优秀研究生“创新之星”项目：基于现代多重插补的稀疏函数型数据修复方法研究(2022CXZX-701)，2022.6---2023.9，已结项。

竞赛获奖：

“黄河干流实测径流量演变特征及其与气候变化和人类活动的响应——基于函数型数据视角”荣获**第六届全国应用统计专业学位研究生案例大赛全国三等奖**，2023 年 10 月。

致谢

三载光阴如梭，即将完成硕士学业，在迈上新的人生阶段之际，感慨万千，借此机会向所有支持、帮助过我的人表达最诚挚的感谢！

学路坎坷，幸遇良师。感谢我的导师高海燕教授，感谢您的谆谆教诲和无私帮助。您的耐心指导、严谨态度和专业知识使我受益匪浅；您的鼓励和肯定激发了我不断探索知识的热情；您的言传身教将成为我人生中最宝贵的财富。

父持母暖，助我学成。感谢我的父母和家人，你们一直是我最坚强的后盾，给予了我无尽的理解、支持和鼓励。你们的默默付出是我前行路上最坚实的后盾，亦是我奋斗的动力。

人来人往，遇见皆缘。感谢我的同门和朋友们，谢谢你们在学术上和生活上的帮助和支持，你们的鼓励和陪伴使我克服了许多困难。

纸上得来终觉浅，绝知此事要躬行。硕士毕业，不是终点，乃是一个新的起点。愿将所学所获，化作实践之力，扬帆起航，勇往直前。愿我们的学术之路，能够如诗词中所言，绝知此事要躬行，不断努力，不断前行，不负韶华，不负自己。