分类号____________ 密级____________

UDC ____________ 编号 10741

LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目 基于多重插补的稀疏函数型数据修复方法研究

研 究 生 姓 名： 李唯欣

指导教师姓名、职称： 高海燕、教授

学 科、专 业 名 称： 统计学、应用统计硕士

研 究 方 向： 大数据分析

提 交 日 期： 2024年6月3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：李唯欣　　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

导师(校外)签名：＿＿＿＿　　签字日期：＿＿＿＿

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：李唯欣　　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

导师(校外)签名：＿＿＿＿　　签字日期：＿＿＿＿

Research on Sparse Functional Data Recovery Method Based on Multiple Imputation

Candidate : Weixin Li

Supervisor: Haiyan Gao

摘 要

大数据时代，随着科学技术的进步和数据收集储存能力的提升，数据结构变得复杂、形式变得多样。传统的结构化数据已经从简单的点数据扩展到区间数据、符号数据和函数型数据等。函数型数据是一类复杂的非线性结构数据，往往以函数(曲线)的形式呈现和储存。由于数据收集过程中，经常会出现数据缺失的情况，因此，针对缺失数据插补方法的研究成为国内外学者关注的重点。然而现有的传统插补方法并不适用于函数型数据，在数据修复过程中并没有考虑函数型数据的潜在信息。为了解决上述问题，本文首先引入类信息挖掘数据之间的相关性，提出一种融合类信息的函数型多重插补方法(Missforest Combining Class Information and PACE，CMFP)。同时，整合数据的横截面信息和纵向信息来推测缺失数据，提出一种基于横截面和纵向信息的函数型多重插补方法(Missforest Combining Gaussian Processes，MFGP)。本文的主要研究内容包括以下两部分：

（1）提出一种融合类信息的函数型多重插补方法(CMFP)。在函数型数据分析框架下，以缺失森林模型 MF 为基础，采用基于条件期望主成分分析的函数型插补方法 PACE 进行初始插补，并通过 K-means 聚类借助样本之间的相关性，给出了一种融合类信息的函数型多重插补方法。模拟数据插补实验结果表明，在不同缺失比例(5%~55%)下，该方法相较于 Hot.deck、均值插补、MF、PACE 等 7 种插补方法，能够保证插补的准确性和有效性。同时，针对股票数据的实例应用验证了该方法插补得到的数据符合实际情况和规律。

（2）提出一种基于横截面和纵向信息的函数型多重插补方法(MFGP)。将基于缺失森林模型 MF 的插补与基于高斯过程 GP 的预测相结合，有效整合函数型数据的横截面和纵向信息，进而提高插补精度。首先，应用 MF 对平面数据进行横截面插补。其次，利用 GP 进行纵向插补。然后，通过计算误差对插补结果进行加权结合。最后，模拟数据插补实验和股票数据实例分析结果表明：在不同缺失比例(5%~55%)下，相较于 Hot.deck、均值插补、MF、GP 等 7 种插补方法，MFGP 方法具有显著的插补优势，插补精度高。

关键词：函数型数据 缺失森林 多重插补 聚类 高斯过程

Abstract

In the era of big data, with the progress of science and technology and the improvement of data collection and storage capacity, the data structure has become complex and diverse. Traditional structured data has expanded from simple point data to interval data, symbolic data and functional data. Functional data is a kind of complex nonlinear structural data, which is often presented and stored in the form of functions (curves). Because data is often missing in the process of data collection, the research on imputation methods for missing data has become the focus of domestic and foreign scholars. However, the existing traditional imputation methods are not suitable for functional data, and the potential information of functional data is not considered in the process of data restoration. In order to solve the above problems, this thesis first introduces class information to mine correlation between sample data, and proposes a functional multiple imputation method based on class information (CMFP). At the same time, the missing data is inferred by integrating the cross-sectional information and longitudinal information of data, and a functional multiple imputation method (MFGP) based on cross-sectional and longitudinal information is proposed. The main research contents of this thesis include the following two parts:

(1) A functional multiple imputation method (CMFP) based on class information is proposed. Under the framework of functional data analysis,

based on the missforest model MF, a functional imputation method PACE based on conditional expectation principal component analysis is used for initial imputation, and a functional multiple imputation method integrating class information is given by K-means clustering with the help of correlation between samples. The experimental results of simulated data imputation show that this method can ensure the accuracy and effectiveness of imputation compared with seven imputation methods, such as Hot.deck, mean imputation, MF and PACE, under different missing ratios(5%~55%). At the same time, the application of stock data proves that the data imputed by this method accords with the actual situation and laws.

(2) A functional multiple imputation method (MFGP) based on cross-sectional and longitudinal information is proposed. Combining the imputation based on missforest method MF with the prediction based on Gaussian process GP, the cross-sectional and longitudinal information of functional data can be effectively integrated, and then the imputation accuracy can be improved. Firstly, MF is used to impute the cross section of plane data. Secondly, longitudinal imputation is carried out by using GP. Then, the imputation results are weighted and combined by calculation error. Finally, the simulation data imputation experiment and the analysis of stock data examples show that, under different missing ratios (5%~55%), compared with seven imputation methods such as Hot.deck, mean imputation, MF, HFI and GP, the MFGP method has significant imputation

advantages and high imputation accuracy.

Keywords： Functional data; Missforest; Multiple imputation; Clustering; Gaussian process

目 录

1 绪论 1

1.1 选题依据与意义 1

1.2 国内外研究现状 2

1.2.1 传统插补方法 2

1.2.2 函数型插补方法 4

1.3 研究内容与结构 5

1.4 本文的创新点 6

2 相关理论 7

2.1 缺失数据插补理论 7

2.1.1 缺失数据的产生机制 7

2.1.2 数据的缺失模式 7

2.2 缺失插补算法 9

2.2.1 热卡插补 9

2.2.2 均值插补 9

2.2.3 Missforest 插补 10

2.2.4 MICE 插补 11

2.2.5 SFI 和 HFI 插补 12

2.2.6 PACE 插补 13

2.2.7 高斯过程 14

2.3 K-means 聚类 15

2.4 缺失插补评价指标 15

2.5 本章小结 16

3 融合类信息的函数型多重插补方法 17

3.1 方法介绍 17

3.2 模拟实验 18

3.2.1 实验设计 18

3.2.2 最佳聚类数目确定 18

3.2.3 实验结果分析 19
3.3 实例应用 21
3.4 本章小结 24
4 基于横截面信息和纵向信息的函数型多重插补方法 26
4.1 方法介绍 26
4.2 模拟实验 27
4.2.1 实验设计 27
4.2.2 实验结果分析 28
4.3 实例应用 34
4.4 本章小结 40
5 总结及展望 42
5.1 总结 42
5.2 展望 42
参考文献 44
攻读硕士学位期间承担的科研任务及主要成果 49
后记 50
附录 51

1 绪 论

1.1 选题依据与意义

目前，数据已经成为众多行业的重要生产要素，随着大数据技术的发展，数据的收集和应用已然成为热点，建立在数据驱动方面的应用成为趋势，同时在数据处理和分析中，随着对数据采集程度的不断提高，出现了一种无穷维、具备连续特征的数据，称为函数型数据。它是一种基于时间或空间的数据类型，描述了某个变量随着时间或空间的推移而变化的规律。例如，股票价格、气温和心率都是函数型数据。如果将函数作为一个整体采用泛函工具进行分析，则称为函数型数据分析(Functional Data Analysis，FDA)(Ramsay&Silverman，2005)，FDA 对于分析无限维数据对象(如曲线、形状和图像)非常有用，它可以模拟在特定领域测量(如时间或空间)的函数之间的关系(Ferraty&Vieu，2006；Horváth&Kokoszka，2012；Kokoszka&Reimherr，2017)，因此其已经成为医学(Harezlak 等，2008；Gertheiss 等，2013)、经济(Ramsay&Ramsey，2002)、环境(黄恒君和漆威，2014)以及其他应用领域中处理大量且复杂数据类型的重要工具。

然而函数型数据往往是不能直接观测的，实际采集结果通常是具备曲线特征的离散采样点，并且数据缺失是较为常见的数据质量问题，其广泛存在于各个领域。例如，在环境监测和交通系统中，由于设备、电源、传输以及监测点增设或停运等原因，通常会导致监测结果存在大量的条状、块状缺失(黄恒君和漆威，2014；Chiou 等，2014)；在医学领域中，由于个体纵向数据搜集、测量成本过高等原因，通常会导致测量的数据是存在缺失的或是稀疏不规则的(James&Sugar，2003；Ciarleglio 等，2022)；在投资理财领域，股票数据经常由于停盘等因素导致其中存在缺失(钟宇航，2022)等等。由于标准的统计方法不能直接应用于缺失数据的统计分析，并且数据缺失往往会增加统计分析任务的复杂性、降低数据质量以及工作效率，严重时会导致统计结果出现偏倚，影响数据分析的质量和最终统计决策的稳定性。因此，如何处理数据缺失问题成为数据预处理过程中的一项重要任务。

1.2 国内外研究现状

目前针对数据缺失问题的处理，学者们已经提出一些插补方法，但是针对函数型数据缺失问题插补方法的研究还有限。下面从传统数据插补方法和函数型数据插补方法两个方面进行研究现状和发展趋势的梳理。

1.2.1 传统插补方法

（1）单一插补方法

单一插补方法包括热卡插补(Hot Deck Imputation，Hot.deck)(Bertsimas 等，2017)、K-近邻(K-Nearest Neighbor，KNN)(Rumaling 等，2020)、均值插补(Junninen&Niska，2004)、EM 插补(Expectation Maximization，EM)(张波，2022)、随机插补(金勇进，2001)、支持向量机(Jerez 等，2010)、贝叶斯模型(Mei 等，2015)等。这些方法可以处理小规模缺失数据，但不适用于解决稀疏、不规则数据的大规模缺失问题。当出现大规模缺失时，上述缺失数据插补方法可能失效。其中，Hot.deck 是在完整数据中找到一个与它相似的对象，然后利用这个相似的值进行填充，但该方法对于回归方程存在一定的局限；KNN 是一个理论上非常成熟的算法，具有易于理解且预测结果较为准确等优点。其思想是寻找 K 个最相似的“邻居”，并利用这些“邻居”的已知信息来填充矩阵中的缺失项。但该方法在数据缺失规模较大时，插补效果较差；均值插补是最常用的插补方法，用样本中已观测数据的均值作为缺失插补值，对数值属性和非数值属性的数据均能处理，其操作简单、快速，且适用范围广泛。但过度依赖于观测值，稳定性较差；EM 算法是基于正态假设的插补方法，对数据分布依赖性较强；支持向量机在选择参数时存在一定的缺陷，并且在训练集较大的情况下容易出现计算速度过慢等问题；神经网络的鲁棒性强，对缺失数据的敏感性不高，同时要求样本量足够大；贝叶斯模型只适用于小样本的缺失值插补。随着机器学习和深度学习方法的发展，越来越多结合机器学习的插补方法被提出，这些方法利用数据内部特征进行训练，通过构建预测模型，借助数据集的观测值对缺失值进行估算。例如混合插补法、K 最近邻法和模糊聚类法(Seaman 等，2012；Heinze 等，2013)。研究结果表明借助机器学习模型对数据集进行插补的方法效果较好(Moons 等，2006)，但对大量

的样本数据进行训练是基于深度学习和机器学习模型数据修复方法的前提，此类方法计算复杂，不易操作。

虽然上述单一插补方法在插补数据时具有一定的优势，但是使用该类方法得到的数据只有一个插补值，原始数据的不确定性并不能表现出来。而多重插补法则能弥补单一插补法的不足，将缺失数据的不确定性考虑在内，通过构造多个插补值来模拟一定条件下估计量的分布，从而通过随机抽样进行缺失插补(张妍，2018)。

（2）多重插补方法

为克服单一插补方法的不足，Rubin(1987)从贝叶斯统计的角度首次提出多重插补(Multiple Imputation，MI)算法，多重插补即多次填充缺失值以得到多个“完整”的数据集。这些完整数据集之间的可变性反映了插补方法引起的不确定性。多重插补方法灵活、通用，可适用于各种环境。针对带有缺失数据的线性模型，在误差分布未知或误差项不满足独立分布时，基于多重插补的缺失数据修复方法依然有效并且插补误差较小(Greg&Tanner，1991；Wei 等，2012)，且多重插补方法还可以应用于各类统计模型中，例如，Pan(2000)、Chen 和 Sun(2009)将多重插补方法应用于可加风险模型和 Cox 模型中解决区间缺失数据的参数估计问题；Wang 和 Feng(2012)将多重插补方法运用到 M 回归模型中；丁先文(2018)对响应变量存在缺失的线性分位数回归模型使用多重插补方法，因此可以看出多重插补方法具有一定的灵活性。同时，多重插补方法还可以应用于大多数领域的缺失数据插补中。在医学领域，Blazek 等(2021)运用多重插补方法恢复健康数据可能丢失的信息，在此基础上探索普通人群中成人高血压和肾脏疾病之间的关联；在金融风险管理领域，Zhao 等(2022)针对信用风险数据提出多重生成对抗性填充网络，该方法通过生成多抗填充网络对每个属性中的缺失数据进行填充，并通过加权平均来综合每个缺失值的多个结果，以此提高插补算法准确性；在抽样调查领域，王霄等(2020)利用聚类和排列组合等方法处理问卷，采用随机发放的策略进行数据采集，并运用多重插补方法对问卷采集过程中造成的数据缺失进行处理。张维群和段格格(2023)针对动态分层抽样中样本丢失的情况，运用多重插补技术对丢失样本变量进行预测。在多重插补的基础上，学者们还与其他模型相结合，进一步提高插补算法对缺失数据的修复能力(Han 等，2017；Faisal 等，2021)。

上述研究均验证了多重插补方法在缺失数据处理领域的可行性，由于多重插补为每个缺失值创建多个预测值，故相应的统计分析会将插补的不确定性考虑在内，从而产生更可靠的标准误差。简而言之，如果观测数据中关于缺失值的信息较少，使得插补的可变性增强，进一步导致分析中出现更高的标准误差。相反，如果观测到的数据对缺失值有较高的预测性，则多个插补数据集之间的插补结果将更加一致，从而产生更小、更可靠的标准误差。

1.2.2 函数型插补方法

函数型数据是一种基于时间或空间变化的复杂数据类型，传统的多重插补方法在插补函数型数据时，并没有考虑到数据中潜在的“函数”特性。针对这一情况，基于函数曲线的特征提出了混合效应模型(James 等，2000；Rice &Wu，2001)，进而插补稀疏不规则的缺失数据。然而，混合效应模型在参数估计较多时，对稀疏数据集的估计值可能存在高度可变性。对于函数型数据而言，插补时还可以利用数据的主成分信息，例如，Yao 等(2005)提出基于条件期望主成分分析(Principal Components Analysis through Conditional Expectation，PACE)的缺失函数型数据修复方法，并将其应用于 AIDS 病人的稀疏 CD4 数据以及酵母细胞的基因表达数据；Preda 等(2010)提出了一种非线性迭代偏最小二乘算法(NIPALS)，该方法是一种类似于雅可比的迭代算法，可用于估计有限随机向量的主成分分析的因素，由于主因素和主成分之间的对偶性，该算法可以适用于存在缺失的数据集；Crambes 等(2019)提出一种针对真实响应变量缺失，但函数型协变量完整的回归插补方法，该方法利用函数型数据的主成分估计回归模型的系数，并预测响应变量的缺失值。同时在模拟数据和空气质量数据上证明了该方法对缺失值具有较好的处理效果。并且，纵向数据作为函数型数据的一种，目前广泛存在于各个研究领域，国内外学者以纵向数据为背景展开的缺失值插补研究数不胜数(He 等，2011，Twisk 等，2013；公徐路等，2019；陈丽嫦等，2020)。同时，部分学者们进一步提出针对函数型数据的多重插补方法，例如，Rao 等(2020)提出用于修复函数型缺失数据的多重插补方法，并将该方法应用于稀疏的血压监测数据，以建立血压与吸烟复吸之间的函数关系；Ciarleglio 等(2022)通过对链式方程多重插补(Multiple Imputation by Chained Equations，MICE)进行扩展得到一种针对响应变

量为函数型数据的 fregMICE 算法，并将其应用在 EHR 数据中。

综上所述，通过对国内外相关文献进行梳理研究表明，虽然学者们从不同的角度对缺失数据进行了大量的研究，但针对函数型缺失数据的多重插补方法目前研究较少，并且利用数据函数信息的多重插补方法在处理缺失数据方面具有一定的优越性。因此，本文将利用函数型数据的类信息以及横截面信息和纵向信息对缺失数据进行插补，以此提高插补模型的性能。

1.3 研究内容与结构

论文的主要内容是在函数型数据分析框架下，构建基于多重插补方法的函数型数据插补方法，通过考虑类信息、纵向信息和横截面信息对插补效果的影响，提升方法的插补性能。论文总共分为五部分，各部分的研究内容如下。

第一部分为绪论。介绍选题的研究背景、国内外研究现状、研究内容和创新之处等。

第二部分为相关理论与原理。简要介绍数据的缺失机制、缺失模式、与本文工作相关的插补方法，以及缺失插补效果评价指标。

第三部分构建融合类信息的函数型多重插补方法。利用数据的类信息对函数型多重插补方法进行改进。首先给出方法框架，其次利用模拟数据进行实验，通过实验结果分析该方法的可行性，最后进行实例分析。

第四部分构建基于横截面信息和纵向信息的函数型多重插补方法。利用数据的横截面信息和纵向信息对函数型多重插补模型进行优化。首先介绍模型，其次通过模拟实验证明该方法的统计性能，最后进行实例分析。

第五部分为研究的总结与展望。对本文研究内容进行总结，并给出未来研究工作的展望。

本文的研究思路与技术路线如图 1.1 所示：

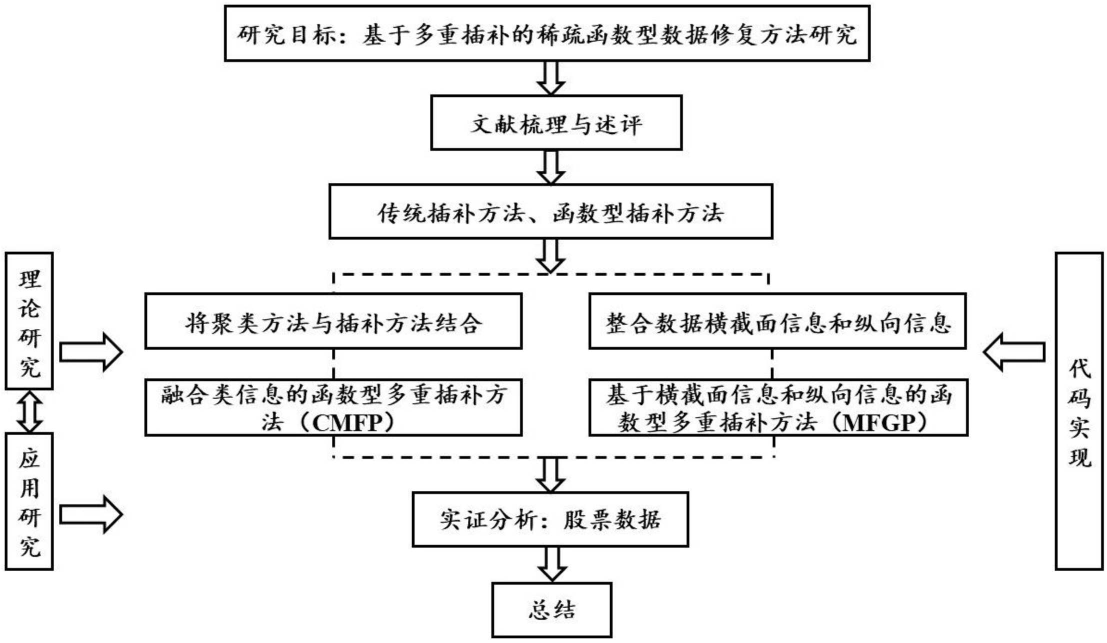


图 1.1　研究思路与技术路线图

1.4 本文的创新点

（1）提出了两类基于多重插补的函数型数据修复方法。以函数型数据为研究对象，在函数型数据分析的框架下研究数据缺失问题。考虑结合类信息、整合横截面信息和纵向信息，分别构建了融合类信息的函数型多重插补方法、基于横截面信息和纵向信息的函数型多重插补方法，提升插补模型的准确性和有效性。

（2）多种插补方法进行对比分析。分别通过均值插补、Hot.deck 等传统插补方法以及 SFI、HFI 等函数型插补方法与本文提出的方法进行比较研究。通过对比插补误差、相关系数、聚类准确度等来分析不同插补方法的优点及不足，并进一步应用到股票数据上进行插补分析。

2 相关理论

2.1 缺失数据插补理论

目前大部分的统计学方法，无论是传统的还是人工智能的，都建立在完整的高质量数据集基础之上。然而在数据收集和清洗过程中，由于人为失误、设备故障和数据传输错误等原因，不可避免的会造成数据缺失问题。数据缺失会导致分析过程中样本量减少、造成变量中信息丢失，进一步使得统计分析中的误差增大，统计结果出现偏倚。

2.1.1 缺失数据的产生机制

根据数据集中缺失部分与观测部分间的概率关系，缺失数据可分为三种常见的缺失机制，即完全随机缺失(Missing Completely at Random，MCAR)、随机缺失(Missing at Random，MAR)和非随机缺失(Not Missing at Random，NMAR)。MCAR 机制中数据缺失是完全随机的，不依赖于任何不完全变量或完全变量，即缺失值的分布既独立于观测部分又独立于缺失部分，不影响样本的无偏性，如联系方式存在缺失。MAR 机制中数据缺失是不完全随机的，即缺失值的分布依赖于观测部分，如财务数据缺失情况与企业的大小有关。NMAR 机制中缺失数据与观测部分和缺失部分均相关，如人群不愿意提供家庭收入。缺失机制会影响缺失数据插补方法的适用性，由于 MCAR 和 MAR 缺失机制中的缺失数据可以从被观测的部分推断出来，而 NMAR 很难通过统计分析的方式估算缺失部分，并且缺失数据基本是完全随机缺失和随机缺失的，故缺失插补方法通常是基于 MCAR 和 MAR 缺失机制的(Madley-dowd 等，2019)。

2.1.2 数据的缺失模式

在处理缺失数据时，除了要考虑数据的缺失机制，还要判断数据的缺失模式。判断数据的缺失模式有利于我们分析数据集中不同变量之间的关系。数据的缺失模式主要包括响应变量缺失、单变量缺失、多变量缺失、文件匹配、单调缺失、一般缺失 6 种缺失模式(申停波，2017)，如表 2.1 所示。表 2.1 中采用矩阵的形式

表示数据集，假设数据集矩阵Y是由n个观测、m个变量组成的$m \times n$矩阵，故可利用矩阵的特征对数据不同的缺失模式进行判断。

表 2.1 数据缺失模式

样本单位	A.响应缺失		B.单变量缺失					C.多变量缺失				
	变量		变量					变量				
	x	y	y_1	y_2	y_3	y_4	y_5	y_1	y_2	y_3	y_4	y_5
1	⊗	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙
2	⊗	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙
3	⊗	⊙	⊙	⊙	⊙	⊙	⊗	⊙	⊙	⊗	⊗	⊗
4	⊗	⊙	⊙	⊙	⊙	⊙	⊗	⊙	⊙	⊗	⊗	⊗
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	⊗	⊙	⊙	⊙	⊙	⊙	⊗	⊙	⊙	⊗	⊗	⊗

样本单位	D.文件匹配			E.单调缺失					F.一般缺失				
	变量			变量					变量				
	y_1	y_2	y_3	y_1	y_2	y_3	y_4	y_5	y_1	y_2	y_3	y_4	y_5
1	⊙	⊙	⊗	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊗	⊗	⊙
2	⊙	⊙	⊗	⊙	⊙	⊙	⊙	⊗	⊙	⊙	⊗	⊙	⊗
3	⊙	⊗	⊙	⊙	⊙	⊙	⊗	⊗	⊙	⊗	⊙	⊗	⊙
4	⊙	⊗	⊙	⊙	⊙	⊗	⊗	⊗	⊙	⊗	⊙	⊙	⊗
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	⊙	⊗	⊙	⊙	⊗	⊗	⊗	⊗	⊙	⊙	⊗	⊗	⊗

对于响应变量缺失模式，该模式表现为某个变量的全部数据缺失，通常会对该种缺失模式采取删除该变量的方法；对于单变量缺失模式，该模式表现为只缺失了一个变量的部分数据；对于多变量缺失模式，该模式与单变量缺失模式的表现类似，只是发生缺失的变量不少于两个，通常会对该种缺失模式采用借补的方式进行缺失值填充；对于文件匹配模式，该模式表现为某些变量出现比较多的数据缺失情况，因为上、下两部分数据源所得到的数据的变量相同，故该种缺失模式可以通过互相填充的方式进行补全；对于单调缺失模式，通常需要调整或舍弃部分数据才能保证该种缺失模式的单调性；对于一般缺失模式，该模式属于随机缺失模式，不存在任何规律，故该种缺失模式需要通过多重插补的方法进行处理，本篇论文后续两章中的模拟数据和实例数据，均属于该种缺失模式。

2.2 缺失插补算法

对于缺失值的处理，总体上可分为删除法和插补法两大类。删除法通过删除部分存在缺失的样本达到目标；插补法则通过寻找一个可能的值来代替缺失数据。虽然在大部分情况下，直接删除法便于操作、易于实现，但该方法在数据缺失较多时，会造成较大的信息缺失，因此该方法只适用于数据存在少量缺失的情况。相比之下，缺失插补可以在一定程度上减少数据损耗和信息缺失，然而缺失值插补相对复杂，该类方法首先要对数据的缺失机制进行假设，根据数据集中一些可用的样本信息，将样本中的缺失值替换成可能的估计值。目前一些传统的插补方法包括均值填充、热卡插补、冷卡插补，全信息极大似然等。

2.2.1 热卡插补

热卡插补法(Hot.deck)也称热平台插补法，该方法对于存在缺失数据的变量，在其余非缺失变量中找到一个与其最相似的变量，然后利用这个相似变量的值对缺失数据进行插补。其主要步骤如下：

步骤 1：获取与缺失数据变量最为相似的其他变量，根据不同问题选取不同的标准对相似进行界定，最常使用的相似判断标准为两变量之间相关系数的显著程度；

步骤 2：将所有样本按此前获取最为相似变量的取值大小排序；

步骤 3：缺失值可用排在其前面的最为相似变量的样本数据进行插补。

该方法概念简单，利用数据之间的关系进行缺失插补，但该方法会导致参数估计不稳定，在具体操作上时效性不强。

2.2.2 均值插补

均值插补法也称均值替换法，可对数值型数据和非数值型数据进行缺失插补。对于数值型数据，一般采用样本中所有观测数据的平均值作为其缺失值的插补值。其插补值的计算方程为：

$$\overline{y}_1 = \frac{\sum_{i=1}^{n} \beta_i y_i}{n_i}$$

其中，β_i为是否回答的描述符号表示，$\beta_i=1$代表“是”，$\beta_i=0$代表“否”，n_i表示个数。对于非数值型数据，缺失值可选择利用该变量中取值次数最多的值进行替换。

虽然该方法简单易行，但容易造成数据分布扭曲，进而导致估计是有偏的，同时该方法不适用于数据为偏态分布的情况。

2.2.3 Missforest 插补

Missforest(MF)是一种基于随机森林算法的非参数多重插补方法，这意味着它不会对函数形式做出明显的假设，而是尝试以最接近数据点的方式来估计函数。换句话说，它为每个变量建立一个随机森林模型，然后使用该模型来预测缺失值。该方法是一种高效、灵活且预测精度较高的缺失插补方法。该方法首先对缺失数据集进行初始插补，即对连续型变量选择该变量平均值对缺失数据进行插补，离散型变量选择该变量众数对数据进行插补，若众数存在多个相同值，则随机选择其中一个即可，同时将变量按照缺失值的数量升序排序，然后对每个特征变量使用缺失森林进行插补。MF 具体执行步骤如算法 1.1 所示，其中$\boldsymbol{X}_s$表示含有缺失的变量；$y_{obs}^{(s)}$表示$\boldsymbol{X}_s$的观测值；$y_{mis}^{(s)}$表示$\boldsymbol{X}_s$的缺失值；$y_{mis}^{(s)}$表示$\boldsymbol{X}_s$以外的观测值；$x_{mis}^{(s)}$表示$\boldsymbol{X}_s$的缺失值以外的其余观测值。终止条件γ定义为$\boldsymbol{X}_{new}^{imp}$和$\boldsymbol{X}_{old}^{imp}$之间的差异，如果$\gamma$增大，则达到终止条件，算法结束，输出插补结果$\boldsymbol{X}^{imp}$。$\boldsymbol{X}_{new}^{imp}$和$\boldsymbol{X}_{old}^{imp}$间的差异定义如下：

对连续型变量，N为连续型变量的集合，差异可定义为

$$\Delta N=\frac{\sum_{j\in N}(\boldsymbol{X}_{new}^{imp}-\boldsymbol{X}_{old}^{imp})^2}{\sum_{j\in N}(\boldsymbol{X}_{new}^{imp})^2}$$

对离散型变量，C为离散型变量的集合，差异可定义为

$$\Delta C=\frac{\sum_{j\in C}\sum_{i=1}^{n}I(\boldsymbol{X}_{(i,j)new}^{imp}\neq\boldsymbol{X}_{(i,j)old}^{imp})}{N_A}$$

其中，N_A为离散型变量缺失值的个数，n为样本数量。

算法 1.1：MF 插补算法

输入： $n\times p$ 的函数型缺失数据矩阵 $\boldsymbol{X}$，终止迭代条件 γ。

1. 对存在缺失值的变量运用该变量的平均值对缺失值进行插补；
2. 计算缺失数据集 $\boldsymbol{X}$ 中各个变量的缺失率，将缺失率从小到大排序，并将对应的变量存入向量 $\boldsymbol{m}$ 中；
3. 判断是否达到终止迭代条件 γ 和最大迭代次数，若达到则算法停止，输出插补后的矩阵 $\boldsymbol{X}_{new}^{imp}$，记为 $\boldsymbol{X}^{new}$，若没达到，则继续 4-6；
4. 存储先前插补后的矩阵，记为 $\boldsymbol{X}_{new}^{imp}$；
5. 对于 $s\in\boldsymbol{m}$，依次执行：

   ①训练随机森林模型：$y_{obs}^{(s)}\sim x_{obs}^{(s)}$；

   ②利用①训练好的随机森林模型输入 $x_{mis}^{(s)}$ 预测 $y_{mis}^{(s)}$ 值；

   ③使用 $y_{mis}^{(s)}$ 值更新插补后的矩阵，记为 $\boldsymbol{X}_{new}^{imp}$；
6. 更新 γ，最大迭代次数，返回 3。

输出： 插补后的矩阵 $\boldsymbol{X}^{imp}$。

2.2.4 MICE 插补

MICE 是一种链式方程多重插补方法，该方法可以为不同类型的特征变量分别建立插补模型，主要模型有预测匹配均值、逻辑回归和多项式逻辑回归等。其中缺失值被多次填充以创建完整的数据集，其假设缺失数据是随机缺失的。因此，插补过程中涉及来自其他观测列的信息。该方法考虑到了插补过程中的不确定性，且非常灵活，可以处理各种类型的变量以及各类复杂情况。其步骤如下：

步骤 1：所有缺失值都用常见的统计方法初始化，通过占位的方式(临时值)对缺失值进行填充；

步骤 2：逐列计算缺失比例，将缺失值最少的变量设置回缺少初始值；

步骤 3：缺失值最少的变量是回归或分类模型中的因变量，所有其他变量是回归模型中的自变量；

步骤 4：用回归模型的预测值替换变量中的缺失值，得到完整变量。当完整变量随后用作另一个变量回归模型中的自变量时，将同时使用该变量的观测值和预测值；

步骤 5：移动到下一个缺失值最少的变量，然后对每个缺失数据的变量重复步骤 2~4，每个变量的循环构成一个迭代或循环。重复 M 次，构成 M 个完整的

数据集。

随后使用标准的统计方法分析每个数据集，最后使用简单的规则将多次插补结果组合在一起以减少缺失数据的不确定性。大多数情况下，多重插补可获得比单一插补更好的结果。其插补流程如图 2.1 所示。

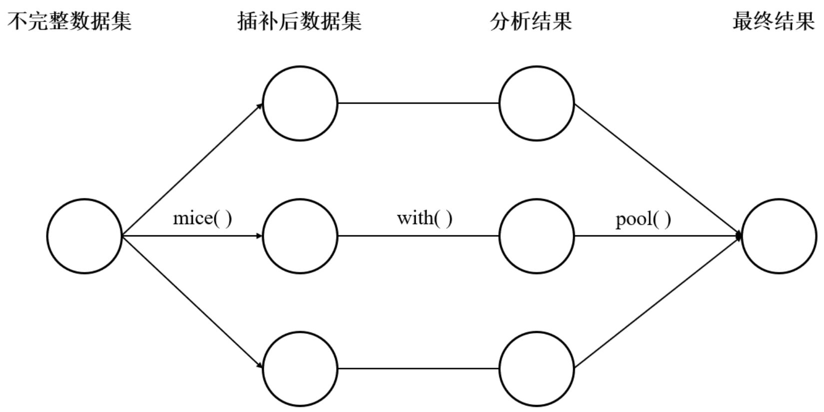


图 2.1　MICE 多重插补流程图

2.2.5 SFI 和 HFI 插补

从秩最小化的角度来理解，矩阵填充是在降维的框架下形成的(薛娇等，2022)，可以表示为：

$$\min_{M}\left\|P_{\Omega}(\boldsymbol{Y}-\boldsymbol{M})\right\|_{F}^{2}+\lambda\left\|\boldsymbol{M}\right\|_{*} \tag{2.1}$$

其中，$\boldsymbol{Y}$ 是观测数据矩阵，$\boldsymbol{M}$ 是待估矩阵，作为$\boldsymbol{Y}$ 的低秩近似。$P_{\Omega}(\cdot)$ 是投影算子，若$\boldsymbol{Y}$ 中的元素 y_{ij} 可观测，则 $P_{\Omega}(y_{ij})=y_{ij}$，否则，$P_{\Omega}(y_{ij})=0$。

设 $\boldsymbol{X}(t)=\left\{\boldsymbol{X}_1(t),\cdots,\boldsymbol{X}_{n_v}(t)\right\}$ 为多视角函数型数据，其中 $\boldsymbol{X}_V(t)=\left\{\boldsymbol{x}_1^v(t),\cdots,\boldsymbol{x}_n^v(t)\right\}$，第 $\boldsymbol{v}$ 个视角的曲线为 $\boldsymbol{x}_i^v(t)$，第 j 个离散观测值 $\tilde{y}_{ij}^v$ 由一般形式的回归模型生成：

$$\tilde{\boldsymbol{y}}_{ij}^{v}=\boldsymbol{x}_i^v(t_{ij})+\varepsilon_{ij}^{v} \tag{2.2}$$

其中，ε_{ij}^{v} 为随机误差项，$j=1,2,\cdots,m_i$ 为在区间 Γ 上离散点的数量。进一步地，$\boldsymbol{x}_i^v(t)$ 在有限维度下近似表示为：

$$\boldsymbol{x}_i^v(t) \approx \sum_{l=1}^{r} \alpha_{il}^v \phi_{il}^v(t) = \boldsymbol{\phi}_i^v(t)' \boldsymbol{\alpha}_i^v \tag{2.3}$$

上式中，$\boldsymbol{\phi}_i^v(t) = \left[\phi_{i1}^v(t), \phi_{i2}^v(t), \cdots, \phi_{ir}^v(t)\right]'$和$\boldsymbol{\alpha}_i^v = \left[\alpha_{i1}^v, \alpha_{i2}^v, \cdots, \alpha_{ir}^v\right]'$分别为既定空间中的一组基和待估系数形成的列向量。若采用矩阵形式，式(2.3)和式(2.2)分别表示为：

$$\tilde{\boldsymbol{X}}_v^+ \approx \boldsymbol{\Phi}_v \boldsymbol{A}_v \tag{2.4}$$

$$\tilde{\boldsymbol{Y}}_v^+ \approx \boldsymbol{\Phi}_v \boldsymbol{A}_v + \boldsymbol{E}_v \tag{2.5}$$

其中，$\boldsymbol{\Phi}_v = \left[\boldsymbol{\phi}_1^v(t), \boldsymbol{\phi}_2^v(t), \cdots, \boldsymbol{\phi}_{m_v}^v(t)\right]'$，$\boldsymbol{A}_v = \left[\boldsymbol{\alpha}_1^v, \boldsymbol{\alpha}_2^v, \cdots, \boldsymbol{\alpha}_n^v\right]'$，$\boldsymbol{E}_v = \left[\boldsymbol{\varepsilon}_1^v, \boldsymbol{\varepsilon}_2^v, \cdots, \boldsymbol{\varepsilon}_n^v\right]'$，$\boldsymbol{\varepsilon}_i^v = \left[\varepsilon_{i1}^v, \varepsilon_{i2}^v, \cdots, \varepsilon_{in}^v\right]$，$\tilde{\boldsymbol{Y}}_v^+$表示第$v$个视角离散数据点形成的矩阵，$\boldsymbol{Y} = \left\{\tilde{\boldsymbol{Y}}_1^+, \cdots, \tilde{\boldsymbol{Y}}_{n_v}^+\right\}$为所有视角的离散数据集，对于第$v$个视角，结合式(2.5)和式(2.1)，通过如下目标函数：

$$\left\|\boldsymbol{o} \odot (\tilde{\boldsymbol{Y}}_v^+ - \boldsymbol{\Phi}_v \boldsymbol{A}_v)\right\|_F^2 + \lambda_v Pen(\boldsymbol{A}_v)$$

即可得到$\boldsymbol{A}_v$的估计结果。其中，$\boldsymbol{o} \in \mathbb{R}^{m \times n}$与$\tilde{\boldsymbol{Y}}_v^+$同型的投影矩阵，即若$\tilde{\boldsymbol{Y}}_v^+$中的条目可测，则$o_{ij} = 1$；否则$o_{ij} = 0$。$Pen(\cdot)$为惩罚项，$\lambda_v$为旋钮参数。将式(2.1)和式(2.4)相结合，构造如下目标函数：

$$\arg\min_{\boldsymbol{A}} \left\|P_\Omega(\tilde{\boldsymbol{Y}} - \boldsymbol{\Phi}\boldsymbol{A})\right\|_F^2 + \lambda \left\|\boldsymbol{A}\right\|_* \tag{2.6}$$

针对式(2.6)的求解成为软函数型矩阵填充法(Soft Functional Impute，SFI)。进一步地，将式(2.6)惩罚项由核范数替换为l_0范数，即硬函数型矩阵填充法(Hard Functional Impute，HFI)，目标函数如下：

$$\arg\min_{\boldsymbol{A}} \left\|P_\Omega(\tilde{\boldsymbol{Y}} - \boldsymbol{\Phi}\boldsymbol{A})\right\|_F^2 + \lambda \left\|\boldsymbol{A}\right\|_0$$

2.2.6 PACE 插补

考虑$L^2(\tau)$中的随机函数$X(t)$用于描述样本曲线的变动轨迹，其中$L^2(\tau)$表示封闭时间间隔τ上平方可积函数的Hilbert空间。假设$X(t)$有均值函数$E(X(t) = \mu(t))$和协方差函数$C_X(t,s) = \mathrm{cov}(X(s), X(t))$，其表达形式如下：

$$X_{ij}(t) = X_i(t_{ij}) + \delta_{ij}$$

其中，$X_{ij}(t)=\left[x_{i1},\cdots,x_{im}\right]^T$ 为函数 X_i 的观测值，t 为时间点，i 为样本，δ 为观测误差。

PACE通过条件期望估计函数型主成分得分，其主成分展开式如下：

$$X_i(t)=\mu_X(t)+\sum_{j=1}^{\infty}\xi_{ij}v_j(t) \tag{2.7}$$

其中，$v_j(t)$ 是 C_X 的特征函数。主成分得分通过式(2.8)得到：

$$\xi_{ij}=\left\langle X_i-\mu_X,v_j\right\rangle \tag{2.8}$$

PACE利用函数型主成分分析降低数据的维度，便于计算和分析。然而PACE仍存在一些不足之处：第一，PACE没有考虑后续的模型是否适合，这导致估计模型参数时存在偏差；第二，PACE作为一种单一插补方法，在形成置信区间、预测区间或p值时，没有考虑插补的不确定性；第三，该方法的预测精度高度依赖数据的分布，在处理非线性数据时存在一定的局限性。

2.2.7 高斯过程

高斯过程(Gaussian Processes，GP)是无限多个服从高斯分布的随机变量的集合，从该集合中任意选取的有限个随机变量服从自身的联合高斯分布。也就是说高斯过程是一个无限维的高斯分布，如随便从中抽取三个维度 x_1,x_2,x_3，它们服从的是它们联合的三元高斯分布。该过程是一种非线性非参数的贝叶斯优化方法，该方法可作为一种预测模型。其预测结果由训练集中现有数据的加权线性组合生成，根据测试点到给定数据点的距离进行缩放(Ghassemzadeh 等，2004)。该方法不仅可以提供点预测，还可以对生成的预测的不确定性进行测量，进而减少误差(Gigl，2010)。

高斯过程通过定义均值函数和协方差函数来描述先验分布和训练数据之间的关系。对于缺失值，可以利用高斯过程的条件分布进行插补，通过预测未知点的属性进行插值。具体如下：

$$p(x_*\mid t_*,X_n^k,T_n^k)\sim N(u_*,\delta_*^2)$$

$$E[x_*]=k(t_*)^T(K(T_n^k,T_n^k)+\delta^2I_N)^{-1}X_n^k$$

$$Con[x_*]=k(t_*,t_*)-k(t_*)^T(K(T_n^k,T_n^k)+\delta^2I_N)^{-1}k(t_*)$$

其中，x_*为需要插补的样本，t_*为该插补点的时间，X_n^k为已有样本集合，T_n^k为已有样本时间点，n为n个点，k为k个高斯过程样本，$K(T_n^k,T_n^k)$表示协方差矩阵函数。通过高斯过程插补得到的数据可行性可由$u[x_*]$衡量：

$$u\left[x_*\right]=\begin{cases}0，对于真实值；\\ Con\left[x_*\right]>0，对于插补值。\end{cases}$$

2.3 K-means 聚类

聚类分析目前被广泛应用于生物和行为科学、市场以及医学研究等领域。最常见的聚类方法是 K 均值聚类分析。该方法首先选择 K 个中心点，把每个数据分配到离它最近的中心点；然后重复计算每类中的点到该中心点的平均距离，分配每个数据到它最近的中心点；最后重复计算和分配，直到观测值不再被分配或是达到最大迭代次数。该算法中损失函数可以定义为各个数据样本距离所属聚类中心的误差平方和，也就是说，该算法把样本分成 K 组并使样本到其指定聚类中心的误差平方和为最小。其具体算法步骤如下：

步骤 1：随机选取 K 个中心，记作：$\mu_1^0,\mu_2^0,\cdots,\mu_k^0$；

步骤 2：定义损失函数：$J(c,\mu)=\min\sum_{i=1}^{M}\left\|x_1-\mu_{c_i}\right\|^2$；

步骤 3：令$t=0,1,2,\cdots$为迭代步数，重复如下过程直至J收敛。

①对于每个样本x_i，将其划分到距离最近的中心类别：

$$c_i^t:=\arg\min_k\left\|x_i-\mu_k^t\right\|^2$$

②对于每个类别中心k，重新计算该类的位置中心：

$$\mu_k^{t+1}:=\arg\min_\mu\sum_{i:c_i^t=k}^{b}\left\|x_i-\mu\right\|^2$$

其中，x_i代表第 i 个样本，C_i是x_i所属的簇，μ_{C_I}代表簇对应的中心点，M 是样本总数。该算法聚类收敛快，算法思想简单易懂，可解释性强。

2.4 缺失插补评价指标

在处理缺失插补问题中，对插补方法性能的评估可利用不同的评价指标。对

于样本曲线 $x_i(t)$ 和估计曲线 $\hat{x}_i(t)(i=1,2,\cdots,n)$ ，本文主要使用的评价指标有：均方根误差(Root Mean Square Error，RMSE)、平均绝对误差(Mean Absolute Error，MAE)以及相关系数。

（1）均方根误差和平均绝对误差

RMSE 和 MAE 值的大小反映了插补值与真实值之间的误差，其值越小表示插补值与真实值越接近，插补效果越好，算法性能越高，其计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n_1}(\hat{x}_i(t)_i-x_i(t))^2}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n_1}\left|\hat{x}_i(t)_i-x_i(t)\right|$$

其中， n 为不完整变量中缺失值的个数。

（2）相关系数

相关系数是用来表明变量之间相关密切程度的，通常用 r 表示，取值范围为 $[-1,1]$，插补值与真实值的相关系数越大，则插补值的有效性越高。其计算公式如下：

$$r\left(x_i(t),\hat{x}_i(t)\right)=\frac{Cov\left(x_i(t),\hat{x}_i(t)\right)}{\sigma_{x_i}(t)\sigma_{\hat{x}_i}(t)}$$

其中， $Cov\left(x_i(t),\hat{x}_i(t)\right)=E\left((x_i(t)-\overline{x}_i(t))(\hat{x}_i(t)-\overline{\hat{x}}_i(t))\right)$ 。

2.5 本章小结

本章对缺失数据的产生机制、缺失模式、常见的缺失数据处理方法以及缺失插补评价指标进行了介绍。缺失数据的存在会降低实验的效率，也可能会给统计分析结果带来偏倚。因此，在选择缺失数据插补方法时，根据数据集的实际情况选择合适的处理方法对后续的数据分析工作尤为重要。每种插补方法都有不同的优缺点，因此对缺失数据插补这一领域的研究至今仍具有重要意义。

3 融合类信息的函数型多重插补方法

3.1 方法介绍

本章通过将函数型插补方法和多重插补方法相结合的方式克服单一插补方法的局限性，常用的多重插补方法有链式方程多重插补 MICE(van Buuren，2007)和缺失森林 MF(Stekhoven&Buhlmann，2012)。MICE 假定缺失值是随机缺失的(MAR)，需要对数据作出假设，并且需要针对不同类型的变量选取合适的插补方法，操作较为复杂，计算量较大。而 MF 是一种建立在随机森林算法基础上的非参数插补方法，不需要对缺失模式以及函数形式进行明确的假设，能够很好地处理具有复杂相互作用和非线性关系的数据。并且目前存在的函数型数据插补方法在处理缺失数据时并没有充分考虑样本曲线之间的相关性，例如空气质量监测站点在时空上具有一定的相关性，距离越近的监测站点在同一时间段观测到的数据相关性越强，所以样本之间的相关性是插补缺失值的一项重要依据(Zhang 等，2004)。

因此，本章以缺失森林方法 MF 框架为基础，首先利用基于条件期望主成分分析的函数型插补方法 PACE 对缺失数据进行初始插补，其次通过 K-means 聚类挖掘样本之间的相关性，并借助样本曲线的类内相似性提高插补精度，提出一种融合类信息的函数型多重插补方法 CMFP。CMFP 方法的执行步骤如算法 3.1 所示。

算法 3.1：CMFP 插补方法

输入： $n\times p$ 的函数型缺失数据矩阵 $\boldsymbol{X}$，终止迭代条件 γ。

1. 对缺失数据矩阵 $\boldsymbol{X}$ 利用 PACE 进行初始插补；
2. 计算初始插补后数据的函数型主成分得分，使用 K-means 对函数型主成分得分进行聚类，将样本划分为 $k(k=1,2,\cdots,N)$ 类，记最终聚类结果为 $\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_N\}$；
3. 对 $\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_N\}$ 分别使用 MF 方法，得到对应的插补结果 $\boldsymbol{C}_i(i=1,2,\cdots,N)$；
4. 将插补结果 $\boldsymbol{C}_i(i=1,2,\cdots,N)$ 按照缺失数据矩阵 X 中变量的顺序进行整合；

输出： 插补后的矩阵 $\boldsymbol{C}$。

3.2 模拟实验

3.2.1 实验设计

为验证 CMFP 方法在插补函数型数据时的有效性，构造模拟数据集，并在不同缺失比例下评价其插补性能。本章通过以下 3 个步骤完成实验设计。

步骤 1：随机构造一个 100*50 的模拟数据集(Rao&Reimherr，2021)。具体为：模拟 50 条独立同分布的曲线$\{X_1(t),\cdots,X_{50}(t)\}$，这些曲线服从均值为 0，协方差为$C_X(t,s)=\frac{\sigma^2}{\Gamma(v)2^{v-1}}\left(\frac{2v|t-s|}{\rho}\right)^v K_v\left(\frac{2v|t-s|}{\rho}\right)$的高斯分布，其中$K_v$是第二类的修正贝塞尔函数。设置参数$\rho=0.5$，$v=5/2$，$\sigma^2=1$，曲线在区间[0,1]上等距选取 100 个时间点进行估计。假设每一个观测点均包含一个均值为 0，方差为 3 的正态测量误差，最终得到函数型数据离散的观测矩阵。

步骤 2：随机生成含有缺失的数据集。为了验证 CMFP 方法在不同缺失比例下均较好的插补精度和插补效果，设置缺失率分别为 5%、15%、25%、35%、45%、55%。

步骤 3：对比方法与评价指标的确定。将 CMFP 方法与 Hot.deck、MF、均值插补、PACE、MFP、SFI、HFI 等 7 种方法进行插补性能的对比。评价指标采用 RMSE、MAE 以及相关系数。

3.2.2 最佳聚类数目确定

在CMFP方法中，聚类数目k会对插补结果产生一定的影响，将聚类数目设置为1~5，通过组内误差平方和(Sum of Squared Error，SSE)确定最佳聚类数目，不同聚类数目下SSE值如图3.1所示。随着聚类数目增多，每个类别中样本数量越来越少，距离越来越近，因此SSE值会随着聚类数目增多而减少，当SSE值下降较大存在“肘点”或减少较缓慢时，插补误差趋于平稳，此时应该停止聚类。从图3.1可以看出，当聚类数目为2时存在“肘点”，且当聚类数目大于等于3时，SSE值下降趋势平缓。进一步，通过RMSE和MAE两个评价指标确定最佳聚类数目，评价指标结果如表3.1所示。从表3.1可以看出，在任何缺失比例下，CMFP方法聚

类数目为2的插补误差均小于聚类数目为3的插补误差。因此，设定最佳聚类数目为2。

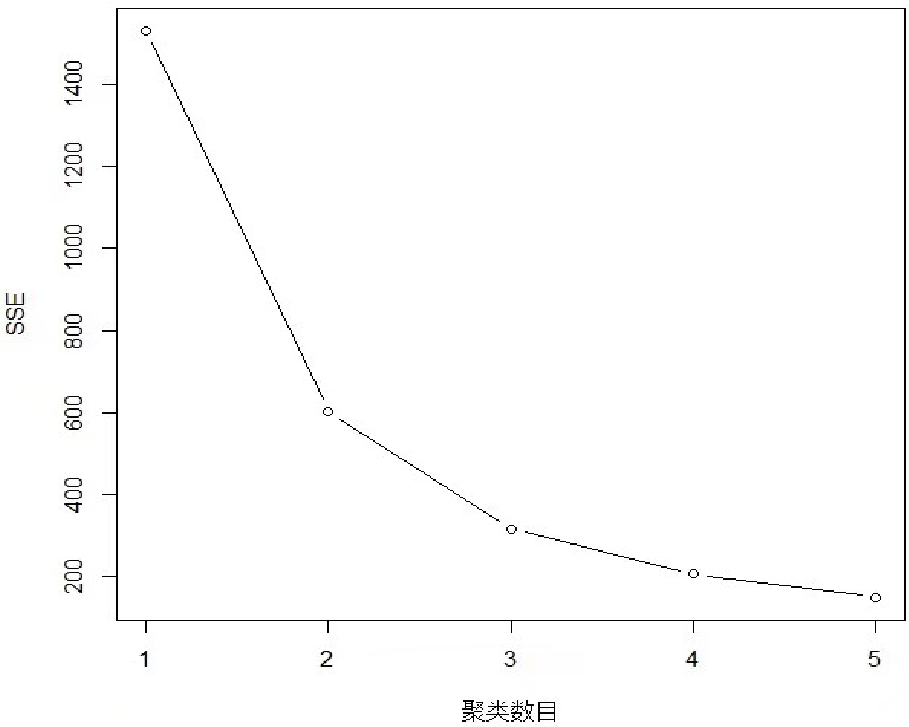


图3.1 不同聚类数目下SSE值

表3.1 不同聚类数目下CMFP方法评价指标

评价指标	k	缺失率(%)					
		5	15	25	35	45	55
RMSE	$k=2$	**0.1683**	**0.1900**	**0.2044**	**0.2231**	**0.2571**	**0.2937**
	$k=3$	0.1745	0.1973	0.2090	0.2328	0.2597	0.2964
MAE	$k=2$	**0.1219**	**0.1335**	**0.1437**	**0.1555**	**0.1805**	**0.2054**
	$k=3$	0.1249	0.1398	0.1461	0.1616	0.1814	0.2072

注：粗体表示比较结果为优，下同。

3.2.3 实验结果分析

利用CMFP方法与Hot.deck、MF、均值插补、PACE、MFP、SFI、HFI对不同缺失比例的数据集进行插补并评价其插补性能。实验主要通过R语言实现，计算

机环境为：Intel(R) Core(TM) i5-5200U CPU2.20 GHz，内存4GB，Windows10 64位操作系统。实验结果如表3.2和表3.3所示。从表3.2和表3.3可以看出，任何缺失比例下，Hot.deck、均值插补、PACE、SFI和HFI的插补误差均较大，其原因在于Hot.deck和均值插补作为传统多元插补方法，在插补函数型数据时并没有考虑到样本的曲线特征，而PACE、SFI和HFI作为单一函数型插补方法，插补有效性较差；MF、MFP和CMFP 3种方法均具有较小的插补误差，且较为接近。因此，为了直观地观察MF、MFP和CMFP 3种方法的插补效果，现将这3种方法的评价指标可视化，如图3.2和图3.3所示。从图3.2和图3.3可以看出，在不同的缺失率下，CMFP方法相较于MF和MFP方法，RMSE分别降低了1.08%~8.53%和1.27%~8.78%，MAE分别降低了0.82%~4.91%和0.61%~4.37%，CMFP方法插补性能在8种方法中表现最优。

表3.2　不同缺失率下RMSE评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	PACE	MF	MFP	SFI	HFI	CMFP
5	1.5710	0.9795	9.1217	0.1845	0.1840	0.9500	0.9500	**0.1683**
15	1.5169	0.9959	1.9185	0.1970	0.1952	0.9500	0.9500	**0.1900**
25	1.4877	1.0018	4.3235	0.2139	0.2139	0.9600	0.9500	**0.2044**
35	1.4512	1.0077	1.0589	0.2352	0.2347	0.9600	0.9500	**0.2231**
45	1.4410	1.0142	1.4739	0.2604	0.2599	0.9600	0.9500	**0.2571**
55	1.4585	1.0187	1.9122	0.3005	0.2969	0.9700	0.9500	**0.2937**

表3.3　不同缺失率下MAE评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	PACE	MF	MFP	SFI	HFI	CMFP
5	1.2473	0.7762	19.1889	0.1282	0.1250	0.7500	0.7500	**0.1219**
15	1.2071	0.7904	2.7008	0.1376	0.1376	0.7500	0.7400	**0.1335**
25	1.1887	0.7945	2.7358	0.1494	0.1493	0.7500	0.7500	**0.1437**
35	1.1611	0.8001	0.9785	0.1630	0.1626	0.7500	0.7500	**0.1555**
45	1.1629	0.8055	0.9553	0.1820	0.1816	0.7600	0.7500	**0.1805**
55	1.1659	0.8082	1.0979	0.2090	0.2071	0.7700	0.7500	**0.2054**

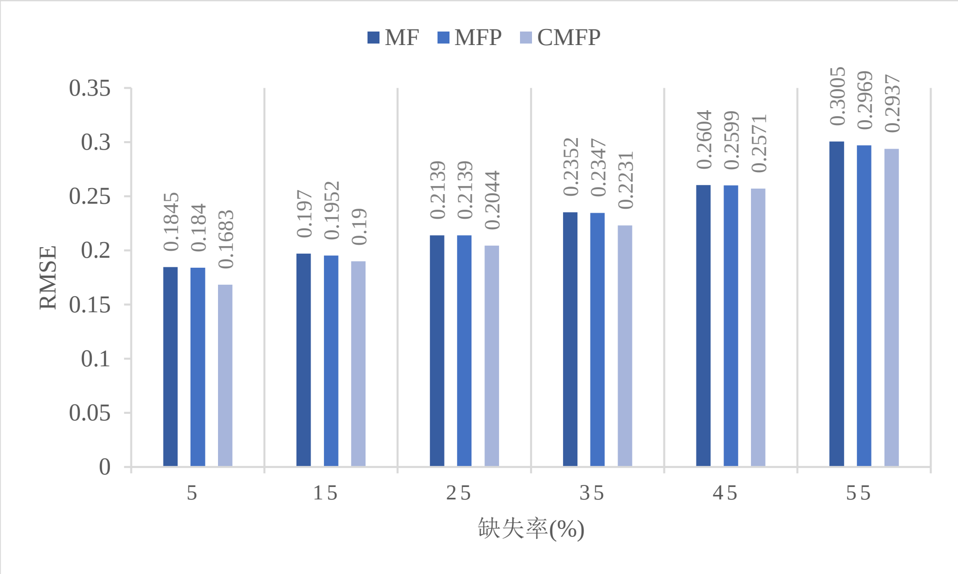


图3.2　不同缺失率下RMSE评价指标对比图

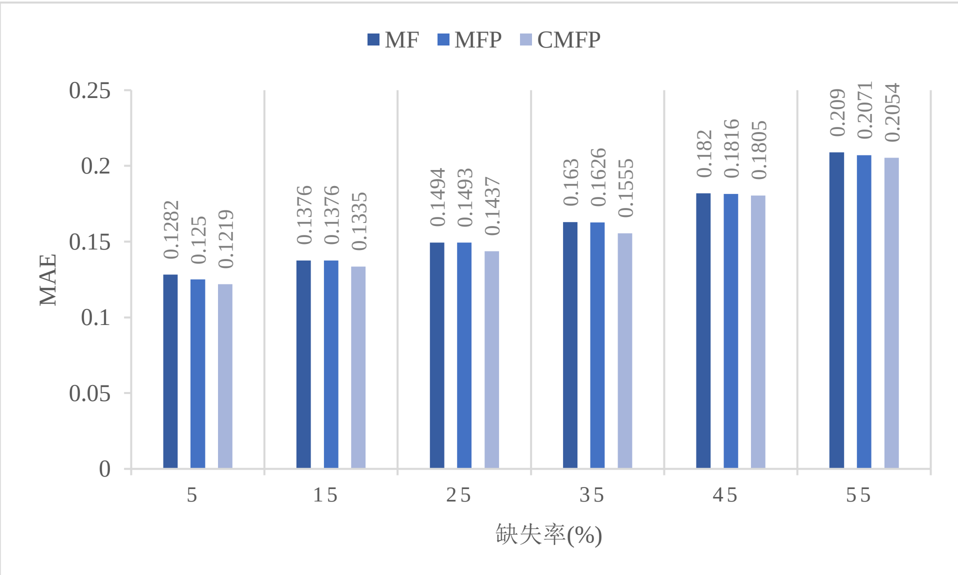


图3.3　不同缺失率下MAE评价指标对比图

3.3 实例应用

为进一步验证 CMFP 方法的实际应用效果，现将该方法应用于样本曲线变化波动较大的股票交易收盘价数据。由于股票分类中最常用的方法是按照股票行业进行分类，并且曾有学者指出行业分类的结果会影响股票的价格(张紫璇和段红梅，2020)，同时行业因素能解释股票收益率 26%波动率的结论，说明同行业

的股票数据之间具有一定的相关性，可按照行业对股票数据进行分类，因此按照2021年第3季度证监会上市公司行业分类结果，以2022年1月1日-2022年12月31日交易数据中的收盘价作为研究对象来验证CMFP方法的适用性。数据来源于东方财富choice，现选取其中4类股票中的24只股票，按其行业分类结果如表3.4所示，其中农业类5只、开采辅助活动类4只、货币金融服务类8只、综合类7只。

表3.4　股票按行业分类结果

门类名称及代码	行业大类代码	行业大类名称	上市公司代码	上市公司名称
农、林、牧、渔业(A)	01	农业	000998	隆平高科
			300087	荃银高科
			300972	万辰生物
			600371	万向德农
			600598	北大荒
采矿业(B)	11	开采辅助活动	002683	宏大爆破
			300191	潜能恒信
			603619	中曼石油
			603979	金诚信
金融业(J)	66	货币金融服务	601860	紫金银行
			601939	建设银行
			601988	中国银行
			601997	贵阳银行
			601998	中信银行
			603323	苏农银行
			600928	西安银行
			601077	渝农商行
综合(S)	90	综合	000551	创元科技
			000833	粤桂股份
			600212	江泉实业
			600603	广汇物流
			600620	天宸股份
			600673	东阳光
			600770	综艺股份

为更好的验证CMFP方法的有效性，选取的股票数据不含缺失值，因此对股票数据进行5%、25%、55%的随机缺失，并使用CMFP方法估计缺失数据。以农业类股票为例，将农业类5只股票完整数据的相关系数和在不同缺失比例下插补后数据的相关系数进行对比，结果如表3.5所示，如果完整数据与插补后数据的相关系数比较接近，则说明插补值符合原有数据规律，插补方法有效。

表 3.5　农业类股票皮尔逊相关系数

股票	缺失率(%)	隆平高科	荃银高科	万辰生物	万向德农	北大荒
隆平高科	0	1.0000	**0.8658**	0.8166	0.3112	0.2492
	5	1.0000	**0.8666**	0.8176	0.3136	0.2529
	25	1.0000	**0.8657**	0.8299	0.3440	0.2867
	55	1.0000	**0.8684**	0.9084	0.3397	0.1834
荃银高科	0	0.8658	1.0000	0.6228	0.5395	0.5156
	5	0.8666	1.0000	0.6285	0.5428	0.5210
	25	0.8657	1.0000	0.6436	0.5636	0.5452
	55	0.8684	1.0000	0.7234	0.5038	0.4089
万辰生物	0	0.8166	0.6228	1.0000	0.3740	0.2995
	5	0.8176	0.6285	1.0000	0.3727	0.3003
	25	0.8299	0.6436	1.0000	0.4015	0.3346
	55	0.9084	0.7234	1.0000	0.4279	0.2574
万向德农	0	0.3112	0.5395	0.3740	1.0000	0.7646
	5	0.3136	0.5428	0.3727	1.0000	0.7707
	25	0.3440	0.5636	0.4015	1.0000	0.8027
	55	0.3397	0.5038	0.4279	1.0000	0.8098
北大荒	0	0.2492	0.5156	0.2995	0.7646	1.0000
	5	0.2529	0.5210	0.3003	0.7707	1.0000
	25	0.2867	0.5452	0.3346	0.8027	1.0000
	55	0.1834	0.4089	0.2574	0.8098	1.0000

从表 3.5 可以看出，不论缺失比例如何，完整数据与插补后数据的相关关系基本保持一致，例如完整数据中隆平高科与荃银高科的相关系数是 0.8658，不同缺失比例下，利用 CMFP 方法插补后数据的相关系数分别为 0.8666、0.8657 和 0.8684，插补值符合原有数据的实际情况和潜在变化规律，故使用 CMFP 方法估计的缺失值合理有效。此外，为了进一步说明 CMFP 方法对后续统计分析的影响，将对不同缺失率下插补的完整数据进行 K-means 聚类，聚类结果如表 3.6 所示。

表 3.6　插补后数据聚类结果

缺失率(%)	类别	股票名称
5	第一类	隆平高科、荃银高科、万辰生物、万向德农、北大荒、宏大爆破、金诚信
	第二类	潜能恒信、中曼石油、紫金银行
	第三类	建设银行、中国银行、贵阳银行、中信银行、苏农银行、西安银行、渝农银行、创元科技
	第四类	粤桂股份、江泉实业、广汇物流、天宸股份、东阳光、综艺股份

续表 3.6

缺失率(%)	类别	股票名称
25	第一类	隆平高科、荃银高科、万辰生物、万向德农、北大荒、中曼石油
	第二类	宏大爆破、潜能恒信、金诚信
	第三类	紫金银行、建设银行、中国银行、贵阳银行、中信银行、苏农银行、西安银行、渝农银行
	第四类	创元科技、粤桂股份、江泉实业、广汇物流、天宸股份、东阳光、综艺股份
55	第一类	隆平高科、荃银高科、万辰生物、万向德农、北大荒、宏大爆破、金诚信
	第二类	潜能恒信、中曼石油、紫金银行
	第三类	建设银行、中国银行、贵阳银行、中信银行、苏农银行、西安银行、渝农银行、创元科技
	第四类	粤桂股份、江泉实业、广汇物流、天宸股份、东阳光、综艺股份

结合表 3.4 和表 3.6 可以看出，缺失率为 5%和 55%时，第一类股票与农业类股票相比，增加了宏大爆破和金诚信 2 只股票；第二类股票与开采辅助类活动类股票相比，有 3 只股票的变动，分别增加了紫金银行，减少了宏大爆破和金诚信 2 只股票；第三类股票与货币金融服务类股票相比，有 2 只股票的变动，分别增加了创元科技，减少了紫金银行；第四类股票与综合类股票相比，减少了 1 只股票，为创元科技。

缺失率为 25%时，第一类股票与农业类股票相比，增加了 1 只股票，为中曼石油；第二类股票与开采辅助类活动类股票相比，减少 1 只股票，为中曼石油；第三类股票与货币金融服务类股票相比以及第四类股票与综合类股票相比，均无变动。通过对比不同行业股票分类结果与不同缺失率下插补后聚类分析结果可以得到，插补后不同类别内股票数量及名称没有明显变动，说明插补值对聚类分析的影响较小。因此，CMFP 方法插补得到的数据符合实际情况，可以在一定程度上降低缺失值对后续统计分析和模型性能的影响。

3.4 本章小结

为了更高效地处理稀疏函数型数据修复问题，本章提出一种基于缺失森林模型的稀疏函数型数据多重插补方法。引入类信息挖掘数据之间的相关性，利用相似样本间的规律性进行插补，以提高插补性能。分别在模拟实验数据和实例数据上验证 CMFP 方法的插补效果，模拟实验结果表明，针对不同的缺失率(5%~55%)，

CFMP 方法相较于 Hot.deck、MF、均值插补、PACE、MFP、SFI、HFI 等 7 种插补方法，插补误差 RMSE 和 MAE 均有不同程度的降低，说明 CFMP 方法具有一定的插补优势，能够保证插补的有效性和准确性；股票交易收盘价数据的实证应用结果表明，CMFP 方法估计的缺失值符合原始数据的实际情况，进一步保障了后续研究的准确性。

4 基于横截面信息和纵向信息的函数型多重插补方法

4.1方法介绍

尽管目前存在很多插补方法可以对缺失数据进行处理,但大部分方法都是在横截面数据(在同一时间点上进行测量)的基础上实现的，不适用于函数型数据。例如，一些临床数据检测结果可能在几周到几个月之间(如血红蛋白)甚至每分钟(如血氧值)发生变化；部分环境科学数据(如气象数据、水资源数据等)可能会在不同的时间点上存在不同的测量结果。此外，这一类数据也会因为后续观测的时间间隔不同而存在较大的差异，往往会导致观测结果不一致。为解决传统插补方法的局限性，本章提出一种基于横截面信息和纵向信息的函数型多重插补方法(MFGP)，该方法利用函数型数据多个变量的信息对缺失进行插补。具体地，利用函数型数据的横截面信息和纵向信息推测缺失数据，将基于缺失森林模型 MF 的插补与基于高斯过程 GP 的插补相结合，通过整合二者的插补结果得到最终插补值。MFGP 方法具体步骤如下：

步骤 1：运用高斯过程 GP 对纵向信息数据进行插补。由于高斯过程减少了对潜在时间轨迹形式的限制，只假设满足局部性约束即可，并且在大部分日常数据中，较接近的时间点之间通常具有较为相似的观测值。因此高斯过程在插补单变量时间序列数据时具有较好的效果。

步骤2：利用缺失森林MF对横截面信息数据进行插补。MF通过随机森林算法建立预测模型，利用已有数据特征插补缺失值，并迭代多次提高插补准确性。该方法具有较好的插补效果和较高的插补精度。

步骤3：将分别借助纵向信息和横截面信息的插补结果进行自加权，权重通过RMSE和MAE和的反比进行赋值，如式(4.1)和式(4.2)所示。进而得到其最终的插补值，提高插补模型的性能。

$$w_1 = k \times (\varepsilon_{GP}^{RMSE} + \varepsilon_{GP}^{MAE}) \tag{4.1}$$

$$w_2 = k \times (\delta_{MF}^{RMSE} + \delta_{MF}^{MAE}) \tag{4.2}$$

其中，$k = \dfrac{1}{\delta_{MF}^{RMSE} + \delta_{MF}^{MAE} + \varepsilon_{GP}^{RMSE} + \varepsilon_{GP}^{MAE}}$，$w_1$表示使用MF插补结果的权重，$w_2$表示

使用GP插补结果的权重，ε_{GP}^{RMSE}表示使用GP插补的RMSE评价指标，ε_{GP}^{MAE}表示使用GP插补的MAE评价指标，δ_{MF}^{RMSE}表示使用MF插补的RMSE评价指标，δ_{MF}^{MAE}表示使用MF插补的MAE评价指标。

以股票数据为例，MFGP 方法示意图如图 4.1 所示。

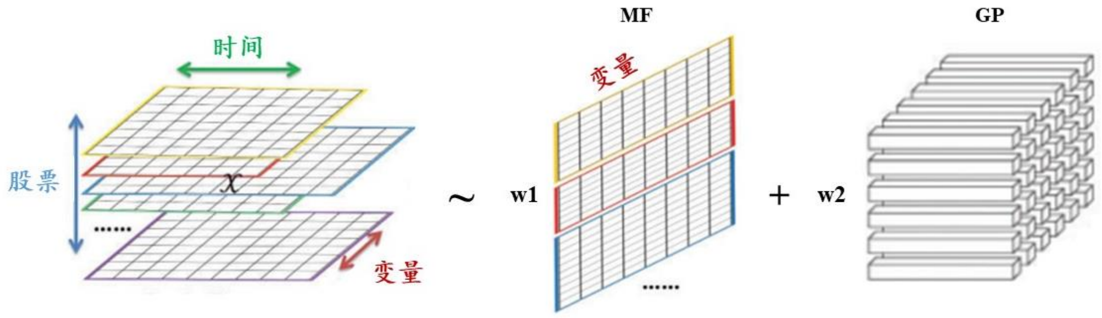


图 4.1　MFGP 方法示意图

4. 2 模拟实验

4. 2. 1 实验设计

为验证 MFGP 方法的插补性能，本章采用下述方法进行模拟实验，并通过对比不同变量在不同缺失率下的插补精度验证其插补性能。

步骤 1：生成模拟数据(薛娇等，2022)。MFGP 方法是利用数据的横截面信息和纵向信息进行插补，因此需要生成存在不同变量的函数型数据集，故本章实验部分生成变量数为 3 的实验数据。函数型数据由三角函数和多项式函数的线性组合生成：

$$\tilde{\boldsymbol{Y}}_{+}=\Phi \boldsymbol{A}+\boldsymbol{E}$$

其中，样本曲线数量$n=5$，时间$t\in[1,10]$，每条曲线的离散采样点为$m=101$，即$t=1,\ 1.1,\ 1.2,\cdots,\ \ j=1,2,\cdots,m$，根据式(2.2)有：

变量 1：取$\phi_j(t)=\left(1,\cos^2\left(\dfrac{t}{10}\right),\sin^2\left(\dfrac{t}{10}\right)\right)^T,\alpha_i=\left(\dfrac{21}{2}+t,\alpha_{i1},\alpha_{i2}\right)^T,r=3$。

变量 2：取 $\phi_j(t)=\left(1,\sin^2\left(\frac{t}{10}\right),\cos^2\left(\frac{t}{10}\right)\right)^T$, $\alpha_i=\left(\frac{21}{2}+t,\alpha_{i1},\alpha_{i2}\right)^T$, $r=3$ 。

变量 3：取 $\phi_j(t)=\left(1,\cos^2\left(\frac{t}{10}\right),\sin^2\left(\frac{t}{10}\right),\left(\frac{t}{10}\right)^2+\frac{t}{10}+1\right)^T$, $\alpha_i=\left(\frac{21}{2}+t,\alpha_{i1},\alpha_{i2},\alpha_{i3}\right)^T$, $r=4$ 。

其中 $\mathbf{\Phi}=\left(\phi_1(t),\phi_2(t),\cdots,\phi_m(t)\right)^T\in\mathbb{R}^{m\times r}$, $\boldsymbol{A}=\left(\alpha_1,\alpha_2,\cdots,\alpha_n\right)\in\mathbb{R}^{r\times n}$, $\boldsymbol{E}=\left(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\right)\in\mathbb{R}^{m\times n}$ ，可以得到每条拟合曲线 $x_i(t)(i=1,2,\cdots,n)$ ， $\alpha_{il}\sim N(1,2)(l=1,2,3)$ 。

步骤 2：针对模拟数据随机生成含有缺失的数据集。为验证 MFGP 方法在不同缺失比例下均有较好的插补效果，设置缺失率分别为 5%、15%、25%、35%、45%、55%。

步骤 3：对比方法与评价指标的确定。将 MFGP 方法与 Hot.deck、均值插补、SFI、HFI、MICE、MF 和 GP 等 7 种方法进行插补性能的对比。评价指标采用 RMSE、MAE 和相关系数。

4.2.2 实验结果分析

利用MFGP方法与Hot.deck、均值插补、SFI、HFI、MICE、MF和GP等7种方法，分别在5%、15%、25%、35%、45%、55%缺失比例的数据集上进行插补，并评价其插补性能，实验结果如表4.1~表4.6所示。为了更直观的观察8种方法之间插补性能的优劣，可视化实验结果如图4.2~图4.7所示。总体来看，在传统插补方法中，Hot.deck和均值插补作为单一插补方法，插补效果较差；MICE和MF为多重插补中常见的两种方法，MF作为非参数插补方法，插补性能优于MICE；SFI和HFI作为两种函数型插补方法，插补效果接近，但插补性能较差；而GP的插补误差与缺失率有关。

对于变量1，从表4.1、表4.2和图4.2、图4.3可以看出，除了在缺失率为55%的情况下，MFGP方法的RMSE和MAE评价指标较为逊色，其余任何缺失比例下，MFGP方法均具有较好的插补性能；对于变量2，由表4.3、表4.4和图4.4、图4.5呈现出，在缺失率为5%时，MFGP方法插补效果略差于MF，除此之外，任何缺失比例下，RMSE和MAE评价指标均为8中插补方法中最优；对于变量3，从表4.5、表4.6和图4.6、图4.7可以看出，缺失率较大时，MFGP方法在MAE评价指标上落

后于SFI和HFI。综合来看，MFGP方法虽然在缺失率较高的数据集上插补效果略差，但在缺失率较小的数据集上插补效果明显高于其他7种方法，具有显著的插补优势。

表 4.1 变量 1 不同缺失率下 RMSE 评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.6889	0.4719	0.5901	0.5901	0.6580	0.2585	0.2380	**0.2331**
15	0.7644	0.5101	0.5899	0.5963	0.7891	0.3679	0.7816	**0.3663**
25	0.8548	0.5399	0.5983	0.5983	0.6865	0.4397	0.2621	**0.2551**
35	0.8779	0.5628	0.5944	0.5944	0.9007	0.5294	0.3205	**0.3120**
45	0.9458	0.5701	0.5908	0.5908	0.7680	0.5069	0.7983	**0.4939**
55	0.8949	0.5889	0.5737	0.5737	1.0608	**0.4513**	1.4002	0.4973

表 4.2 变量 1 不同缺失率下 MAE 评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.4955	0.3459	0.4828	0.4828	0.5181	0.3670	0.3558	**0.3529**
15	0.5950	0.4107	0.4829	0.4871	0.4603	0.3650	0.4163	**0.3649**
25	0.6549	0.4332	0.4884	0.4884	0.5138	0.3565	0.4334	**0.3559**
35	0.6846	0.4590	0.4815	0.4815	0.5045	0.3808	0.4643	**0.3804**
45	0.7411	0.4674	0.4778	0.4778	0.5548	0.4964	0.4641	**0.4502**
55	0.7105	0.4881	**0.4635**	**0.4635**	0.5703	0.5038	0.4894	0.5036

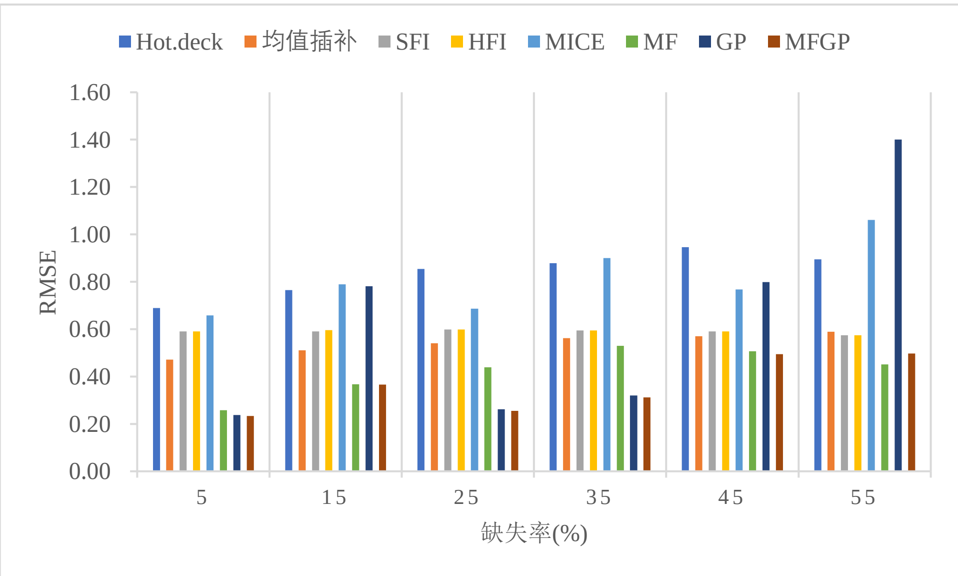


图 4.2 变量 1 不同缺失率下 RMSE 评价指标对比图

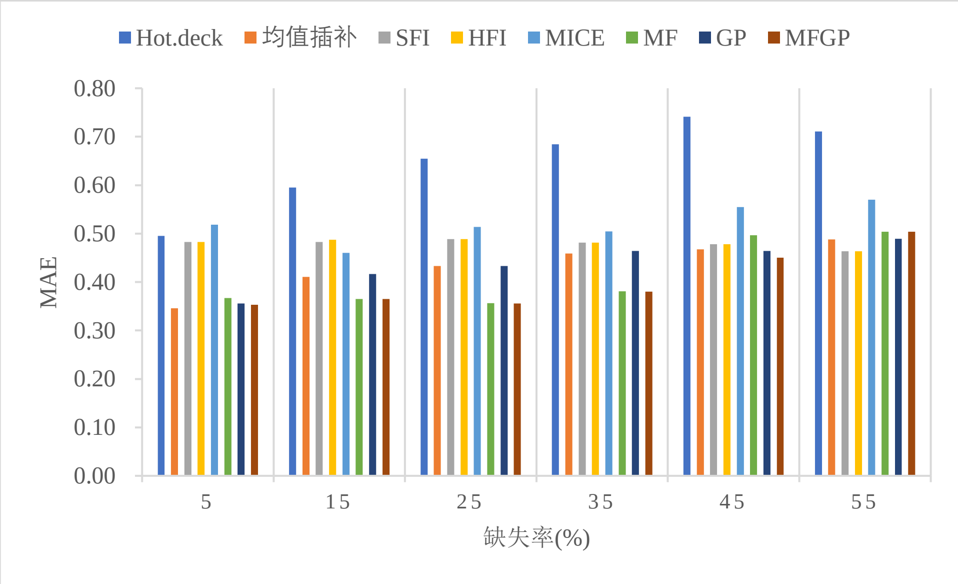


图 4.3　变量 1 不同缺失率下 MAE 评价指标对比图

表 4.3　变量 2 不同缺失率下 RMSE 评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.8405	0.6290	0.5471	0.5471	0.6499	**0.4547**	0.8650	0.4548
15	0.8237	0.5522	0.5449	0.5516	0.6682	0.3916	0.2797	**0.2772**
25	0.8368	0.5345	0.5570	0.5570	0.6623	0.5769	0.2240	**0.2239**
35	0.8202	0.5555	0.5611	0.5611	0.7192	0.4236	0.9101	**0.4233**
45	0.8229	0.5473	0.5646	0.5646	0.9963	0.4954	1.1641	**0.4703**
55	0.8767	0.5450	0.5816	0.5816	1.1152	0.4825	1.1487	**0.4808**

表 4.4　变量 2 不同缺失率下 MAE 评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.6580	0.5068	0.4479	0.4479	0.4069	0.3098	0.5047	**0.3096**
15	0.6616	0.4559	0.4459	0.4496	0.4125	0.4670	0.4559	**0.4534**
25	0.6767	0.4486	0.4554	0.4554	0.4885	0.3524	0.4486	**0.3517**
35	0.6524	0.4654	0.4613	0.4613	0.5244	0.3883	0.4666	**0.3878**
45	0.6510	0.4568	0.4681	0.4681	0.5111	0.4072	0.4911	**0.4064**
55	0.6964	0.4594	0.4777	0.4777	0.5416	0.4126	0.4608	**0.4125**

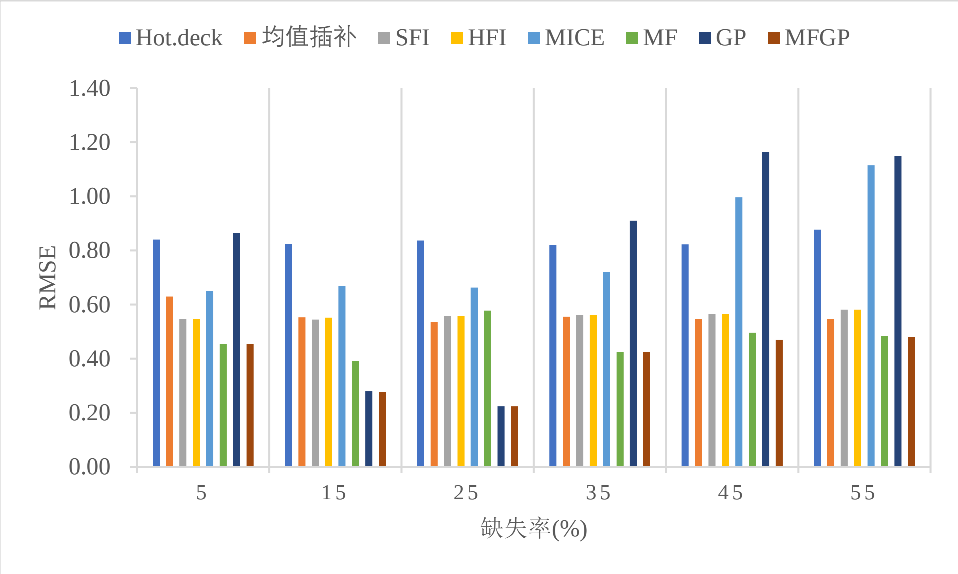


图 4.4　变量 2 不同缺失率下 RMSE 评价指标对比图

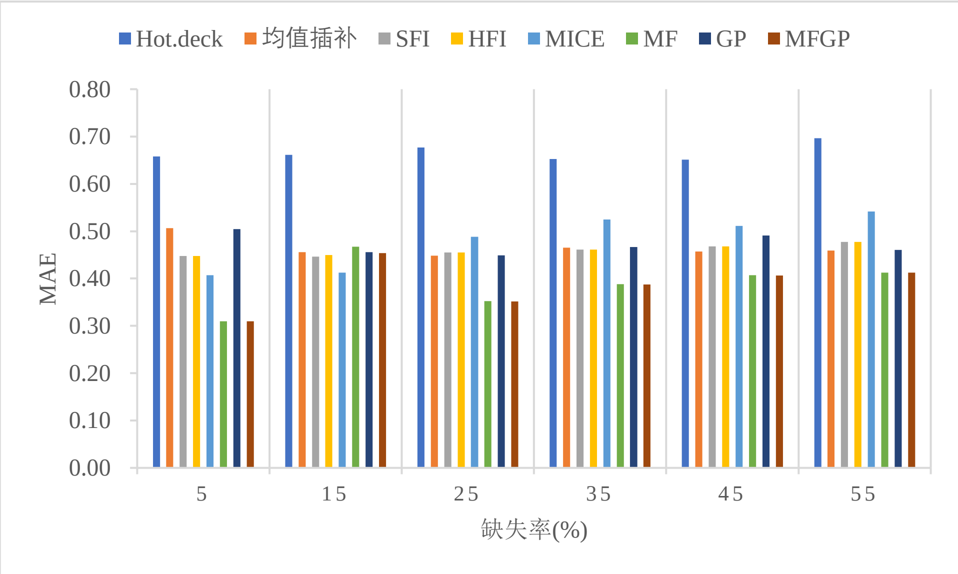


图 4.5　变量 2 不同缺失率下 MAE 评价指标对比图

表 4.5　变量 3 不同缺失率下 RMSE 评价指标

缺失率	方法							
(%)	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.7384	0.5084	0.5143	0.5143	0.7742	0.2976	0.0582	**0.0546**
15	0.7646	0.5397	0.5131	0.5108	0.4464	0.2277	0.1071	**0.1037**
25	0.8090	0.5600	0.4985	0.4985	0.5330	0.3261	0.3636	**0.2937**
35	0.7128	0.5805	0.4809	0.4809	0.5136	0.2858	0.8318	**0.2585**
45	0.6800	0.5725	0.4748	0.4748	0.9047	0.3265	0.3864	**0.3173**
55	0.6911	0.5715	0.4526	0.4526	0.7524	0.4256	0.3923	**0.3725**

表 4.6　变量 3 不同缺失率下 MAE 评价指标

缺失率 (%)	方法							
	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.5952	0.4490	0.4182	0.4182	0.4829	0.3176	0.4428	**0.3173**
15	0.5807	0.4525	0.4174	0.4150	0.5286	0.3544	0.4500	**0.3443**
25	0.5973	0.4567	0.4063	0.4063	0.4197	0.3495	0.4567	**0.3389**
35	0.5506	0.4590	0.3905	0.3905	0.4841	0.3732	0.4547	**0.3702**
45	0.5234	0.4535	**0.3819**	**0.3819**	0.5455	0.4386	0.4535	0.4542
55	0.5433	0.4549	**0.3565**	**0.3565**	0.5365	0.4802	0.4549	0.4379

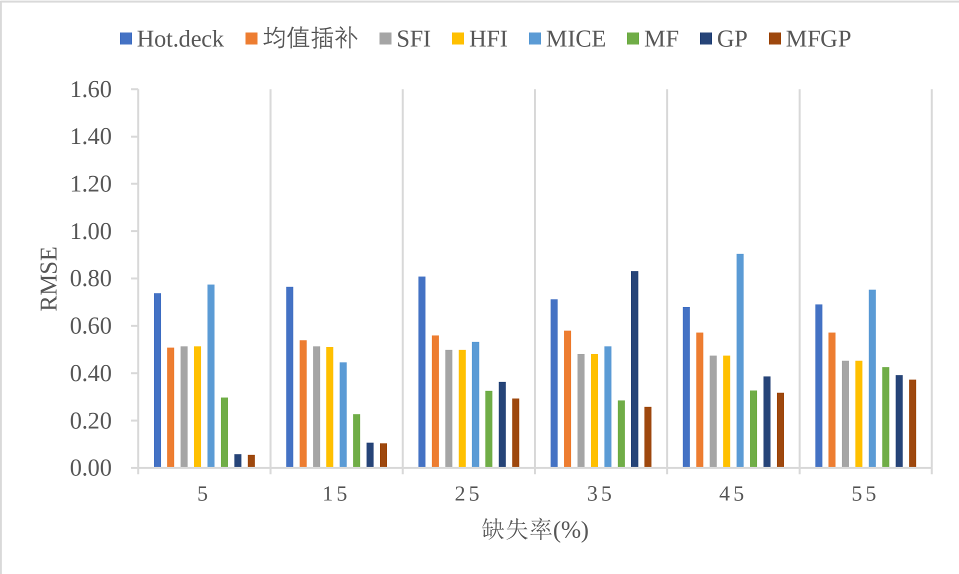


图 4.6　变量 3 不同缺失率下 RMSE 评价指标对比图

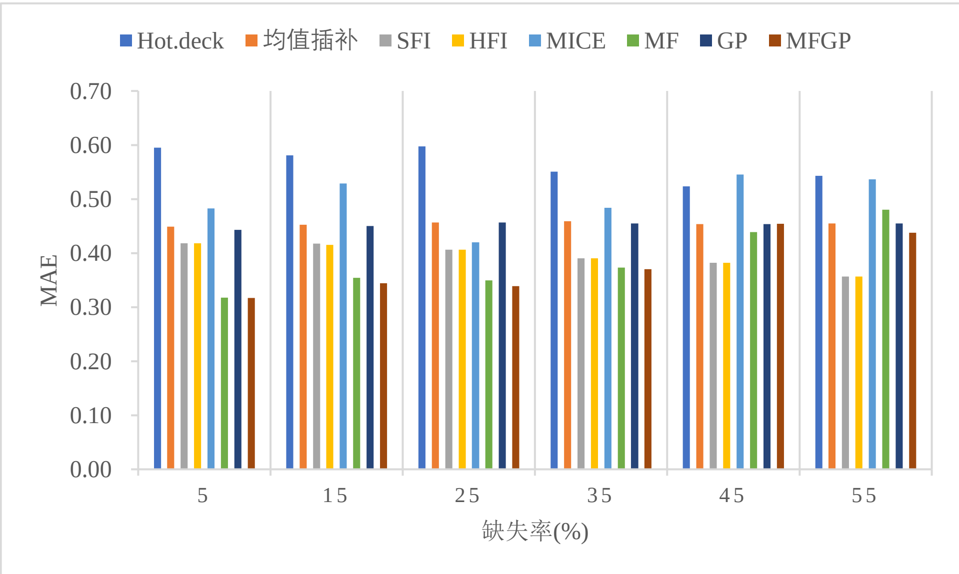


图 4.7　变量 3 不同缺失率下 MAE 评价指标对比图

同时，为验证 MFGP 方法的可靠性，本章对 8 种方法的评价指标随缺失率变化的趋势进行分析，结果如图 4.8~图 4.10 所示。从图中可以看到，相较于其他 7 种插补方法，MFGP 方法在变量 1 和变量 3 中的 RMSE 和 MAE 整体上随着缺失率的增大而增加，呈缓慢上升趋势；变量 2 的 RMSE 和 MAE 只有在缺失率为 15%和 25%时，存在稍大的涨落，但都在合理的范围之内。因此可以看出，MFGP 方法在插补数据时产生的插补误差具有一定的稳定性和鲁棒性，基本不会随着缺失率的变化以及不同变量的变化出现较大起伏，该方法在处理缺失数据时存在一定的优势。

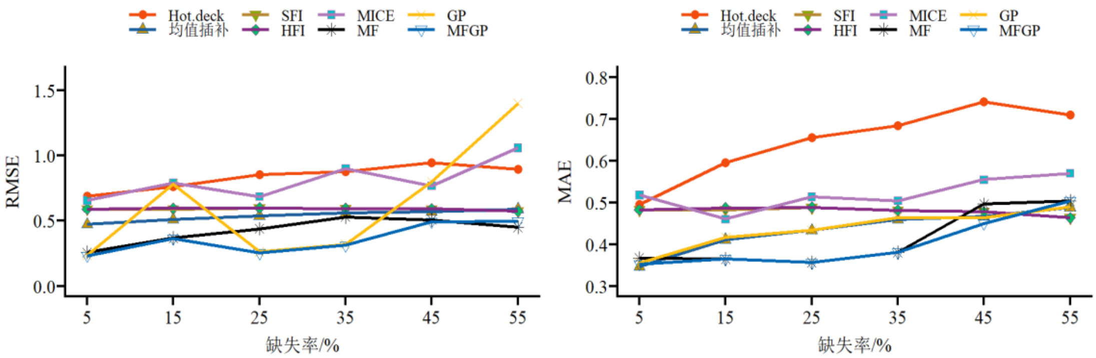


图 4.8　变量 1 不同缺失率下 RMSE 和 MAE 评价指标趋势图

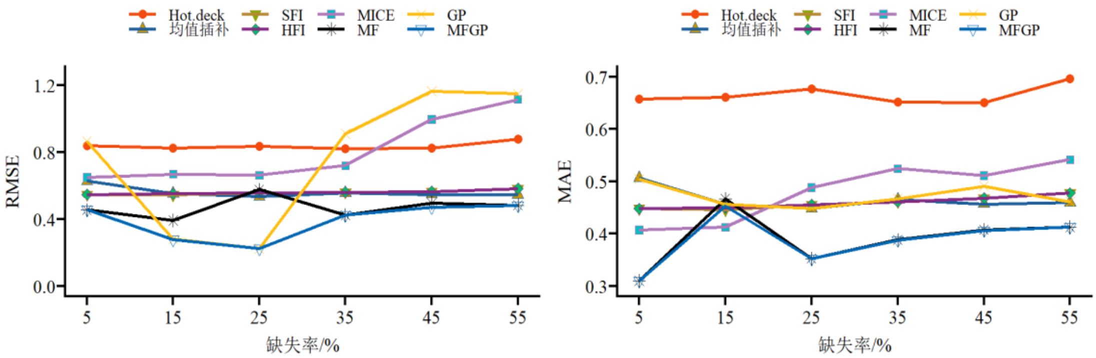


图 4.9　变量 2 不同缺失率下 RMSE 和 MAE 评价指标趋势图

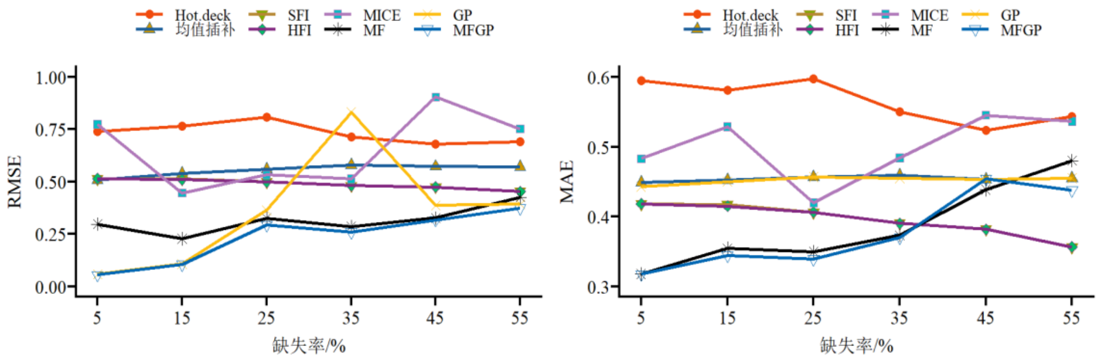


图 4.10 变量 3 不同缺失率下 RMSE 和 MAE 评价指标趋势图

4. 3 实例应用

为进一步验证 MFGP 方法的有效性和适用性，本章按照证监会上市公司行业分类结果随机选取文化、体育和娱乐类股票中华媒控股、中原传媒、华闻集团、中文在线、世纪天鸿 5 只股票，以其 2022 年 1 月 1 日-2022 年 12 月 31 日交易数据中的收盘价、开盘价和最高价 3 个变量作为研究对象来验证 MFGP 方法的可行性与准确性，数据来源于东方财富 choice。为更准确的研究 MFGP 方法是否适用于插补函数型数据，故首先选取的股票数据为完整的，其次针对收盘价进行不同程度的缺失，并采用 MFGP 方法与 Hot.deck、均值插补、SFI、HFI、MICE、MF、GP 等 7 种方法，分别在 5%、15%、25%、35%、45%，55%缺失比例的数据集上进行插补并评价其插补性能。其实验结果如表 4.7 和表 4.8 所示。从表 4.7 和表 4.8 可以看出，任何缺失比例下，Hot.deck、均值插补、SFI、HFI、MICE 插补误差较大，其原因在于 Hot.deck 和均值插补作为单一插补方法，在插补函数型数据时具有一定的缺陷，插补效果较差；SFI 和 HFI 虽是函数型插补方法，但作为单一插补方法，并没有考虑缺失值的不确定性和变异性，插补效果有限；MICE 和 MF 作为多重插补方法，虽克服了单一插补的局限，但对于函数型数据，并没有考虑不同信息对插补值的影响，插补效果并不理想；GP 和 MFGP 两种方法具有较小且较为接近的插补误差。由于本章提出的方法 MFGP 结合了 MF 和 GP 两种方法的优势，因此，将 MF、GP 和 MFGP 这 3 种方法的评价指标进行可视化，如图 4.11 和图 4.12 所示。图 4.11 和图 4.12 直观呈现出，MFGP 的插补误差最

小，其插补性能在 8 种方法中表现最优。

表 4.7　不同缺失率下 RMSE 评价指标

缺失率 (%)	方法							
	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	1.2467	0.8029	0.5113	0.5113	0.7010	0.3966	0.1262	**0.1210**
15	1.1047	0.8422	0.5066	0.5067	1.0577	0.4952	0.2346	**0.2273**
25	1.0646	0.8276	0.5144	0.5144	1.3931	0.5425	0.2977	**0.2893**
35	1.1347	0.8596	0.5062	0.5062	1.3444	0.6848	0.1013	**0.0935**
45	1.1497	0.9203	0.4899	0.4898	0.9274	1.0806	0.0975	**0.0857**
55	1.2135	0.9138	0.4646	0.4646	1.7633	0.4619	0.0746	**0.0693**

表 4.8　不同缺失率下 MAE 评价指标

缺失率 (%)	方法							
	Hot.deck	均值插补	SFI	HFI	MICE	MF	GP	MFGP
5	0.8120	0.5135	0.3301	0.3302	0.4248	0.1747	0.1215	**0.1204**
15	0.7725	0.5703	0.3253	0.3253	0.4492	0.2021	0.1081	**0.1068**
25	0.7326	0.5659	0.3577	0.3577	0.4848	0.3024	0.0955	**0.0946**
35	0.7335	0.5688	0.3233	0.3234	0.5019	0.3408	0.0921	**0.0912**
45	0.7431	0.5913	0.3143	0.3143	0.5700	0.3441	0.0926	**0.0921**
55	0.8962	0.5970	0.3021	0.3020	0.5923	0.4354	**0.0971**	0.0975

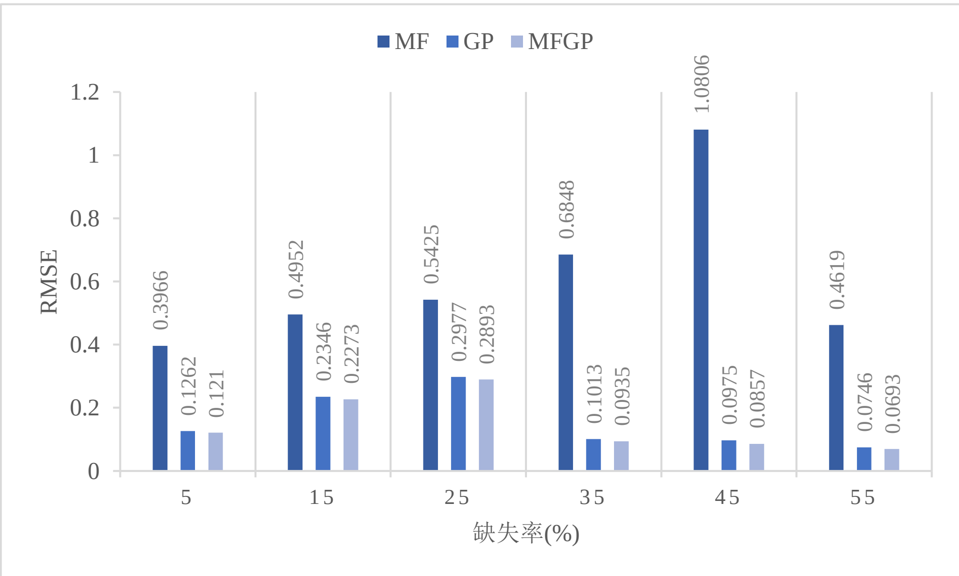


图 4.11　不同缺失率下 RMSE 评价指标对比图

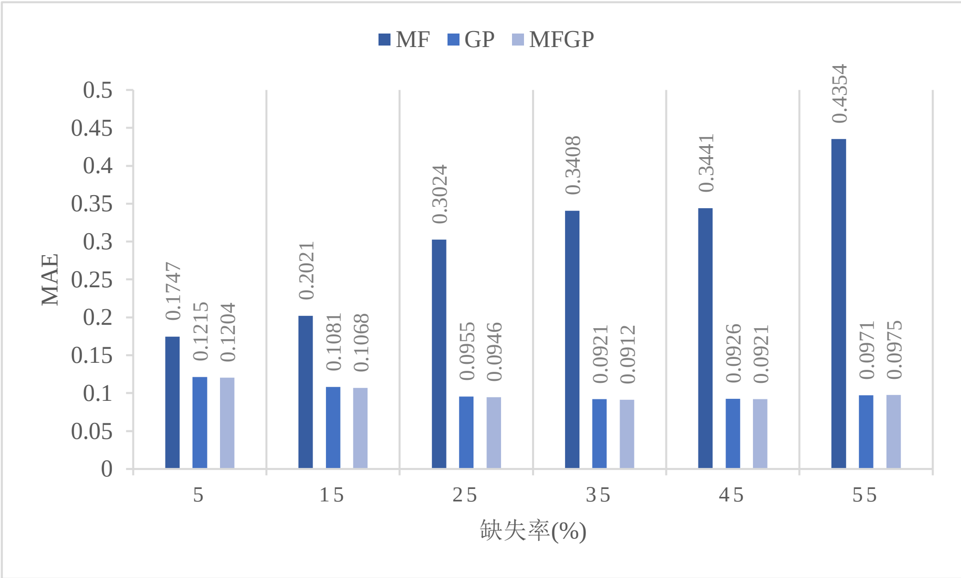
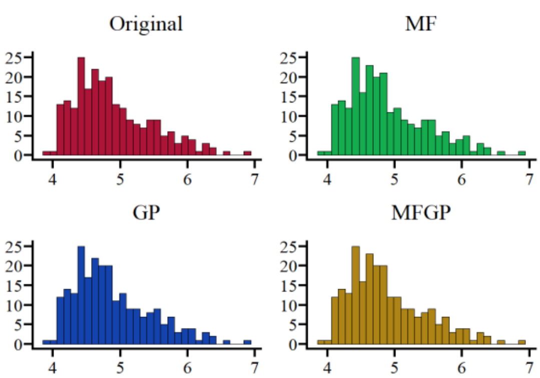
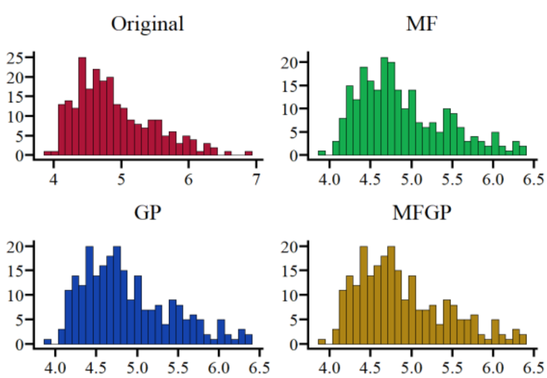


图 4.12　不同缺失率下 MAE 评价指标对比图

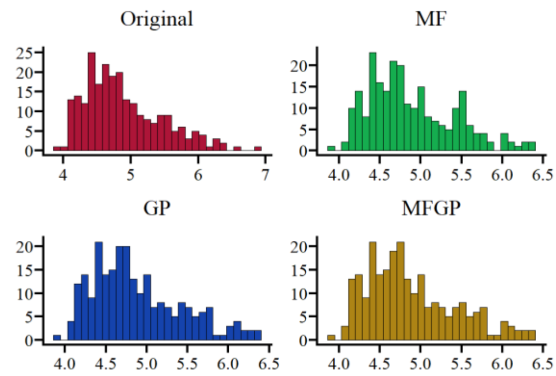
为对比数据插补前后的分布是否一致或相近，单从表格数据很难判断插补后数据对原数据的影响，因此，以华媒控股为例(其余股票见附录图 4.16~图 4.19)，选取插补误差较小的 MF、GP 和 MFGP 这 3 种方法，采用分布直方图进行验证，其结果如图 4.13 所示。图 4.13 中，对于不同缺失比例，红色(左上)为完整数据直方图；绿色(右上)为使用 MF 方法插补后数据直方图；蓝色(左下)为使用 GP 方法插补后数据直方图；黄色(右下)为使用 MFGP 方法插补后数据直方图。从图 4.13 中可以看出，在任何缺失比例下，MF 插补方法均对数据分布产生了较为严重的影响，GP 和 MFGP 方法对数据分布产生的影响较小，但 GP 相对于 MFGP 有较大插补误差。总体来看，MFGP 方法具有更高的插补性能。



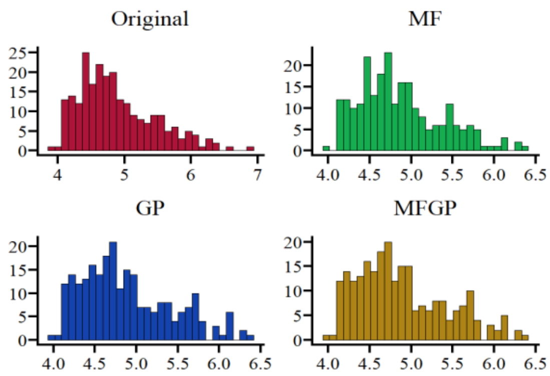
(a)5%



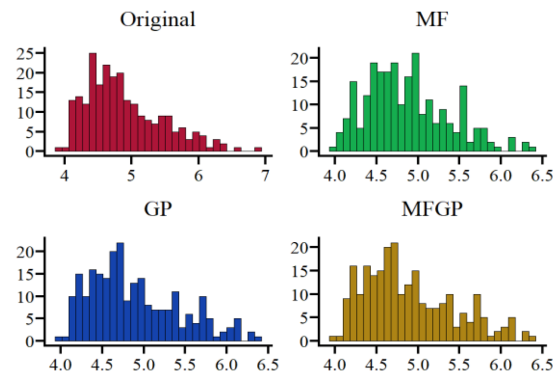
(b)15%



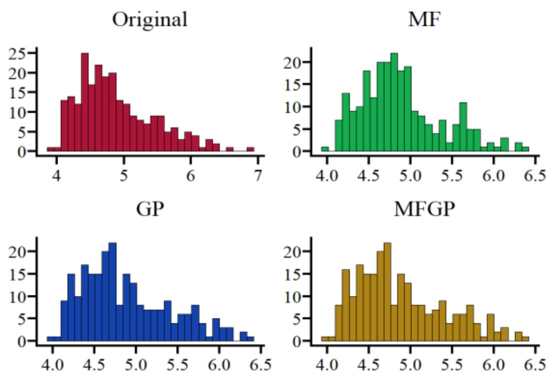
(c)25%



(d)35%



(e)45%



(f)55%

图 4.13 不同缺失率下华媒控股插补后数据分布直方图

进一步，为说明通过 MFGP 方法获得插补值的有效性和准确性，分别计算完整数据和在不同缺失比例下插补后数据的相关系数，结果如表 4.9 和图 4.14 所示。从表 4.9 中可以看出，在任何缺失比例下，完整数据与插补后数据的相关关系基本保持一致，例如完整数据中华媒控股与中原传媒的相关系数是 0.5415，不同缺失比例下，利用 MFGP 方法插补后数据的相关系数分别为 0.5320、0.5478、0.5404、0.5331、0.5317 和 0.5346，插补值符合原有数据的潜在变化规律，故 MFGP 方法估计的缺失值有效。在图 4.14 中，蓝色和从左下指向右上的斜杠表示单元格中两个变量呈正相关。相反，红色和从左上指向右下的斜杠表示变量呈负相关。颜色越深，饱和度越高，说明变量相关性越大。相关性接近于 0 的单元格基本无色，相关性大小由被填充的饼图块的大小来展示。因此从图 4.14 中可以看出，任何缺失比例下，插补前后的相关系数矩阵均与完整数据相关系数矩阵接近，说明通过使用 MFGP 方法得到的插补值有效性较高。

表 4.9 相关系数矩阵

股票	缺失率(%)	华媒控股	中原传媒	华闻集团	中文在线	世纪天鸿
华媒控股	0	1.0000	**0.5415**	0.8464	0.6809	-0.0013
	5	1.0000	**0.5320**	0.8444	0.6780	-0.0077
	15	1.0000	**0.5478**	0.8342	0.6945	0.0084
	25	1.0000	**0.5404**	0.8296	0.6948	0.0053
	35	1.0000	**0.5331**	0.8256	0.6910	-0.0005
	45	1.0000	**0.5317**	0.8245	0.6945	-0.0087
	55	1.0000	**0.5346**	0.8245	0.6954	-0.0042
中原传媒	0	0.5415	1.0000	0.3018	0.3405	0.5225
	5	0.5320	1.0000	0.2936	0.3312	0.5172
	15	0.5478	1.0000	0.2930	0.3290	0.5190
	25	0.5404	1.0000	0.2806	0.3213	0.5141
	35	0.5331	1.0000	0.2765	0.3275	0.5086
	45	0.5317	1.0000	0.2791	0.3287	0.4989
	55	0.5346	1.0000	0.2847	0.3291	0.4932
华闻集团	0	0.8464	0.3018	1.0000	0.7388	-0.3650
	5	0.6780	0.2936	1.0000	0.7317	-0.3698
	15	0.8342	0.2930	1.0000	0.7294	-0.3646
	25	0.8296	0.2806	1.0000	0.7315	-0.3810
	35	0.8256	0.2765	1.0000	0.7245	-0.3925
	45	0.8245	0.2791	1.0000	0.7253	-0.3929
	55	0.8245	0.2847	1.0000	0.7187	-0.3850
中文在线	0	0.6809	0.3405	0.7388	1.0000	-0.2853
	5	0.8444	0.3312	0.7317	1.0000	-0.2869
	15	0.6945	0.3290	0.7294	1.0000	-0.2775
	25	0.6948	0.3213	0.7315	1.0000	-0.2802
	35	0.6910	0.3275	0.7245	1.0000	-0.2815
	45	0.6945	0.3287	0.7253	1.0000	-0.2889
	55	0.6954	0.3291	0.7187	1.0000	-0.2836
世纪天鸿	0	-0.0013	0.5225	-0.3650	-0.2853	1.0000
	5	-0.0077	0.5172	-0.3698	-0.2869	1.0000
	15	0.0084	0.5190	-0.3646	-0.2775	1.0000
	25	0.0053	0.5141	-0.3810	-0.2802	1.0000
	35	-0.0005	0.5086	-0.3925	-0.2815	1.0000
	45	-0.0087	0.4989	-0.3929	-0.2889	1.0000
	55	-0.0042	0.4932	-0.3850	-0.2836	1.0000

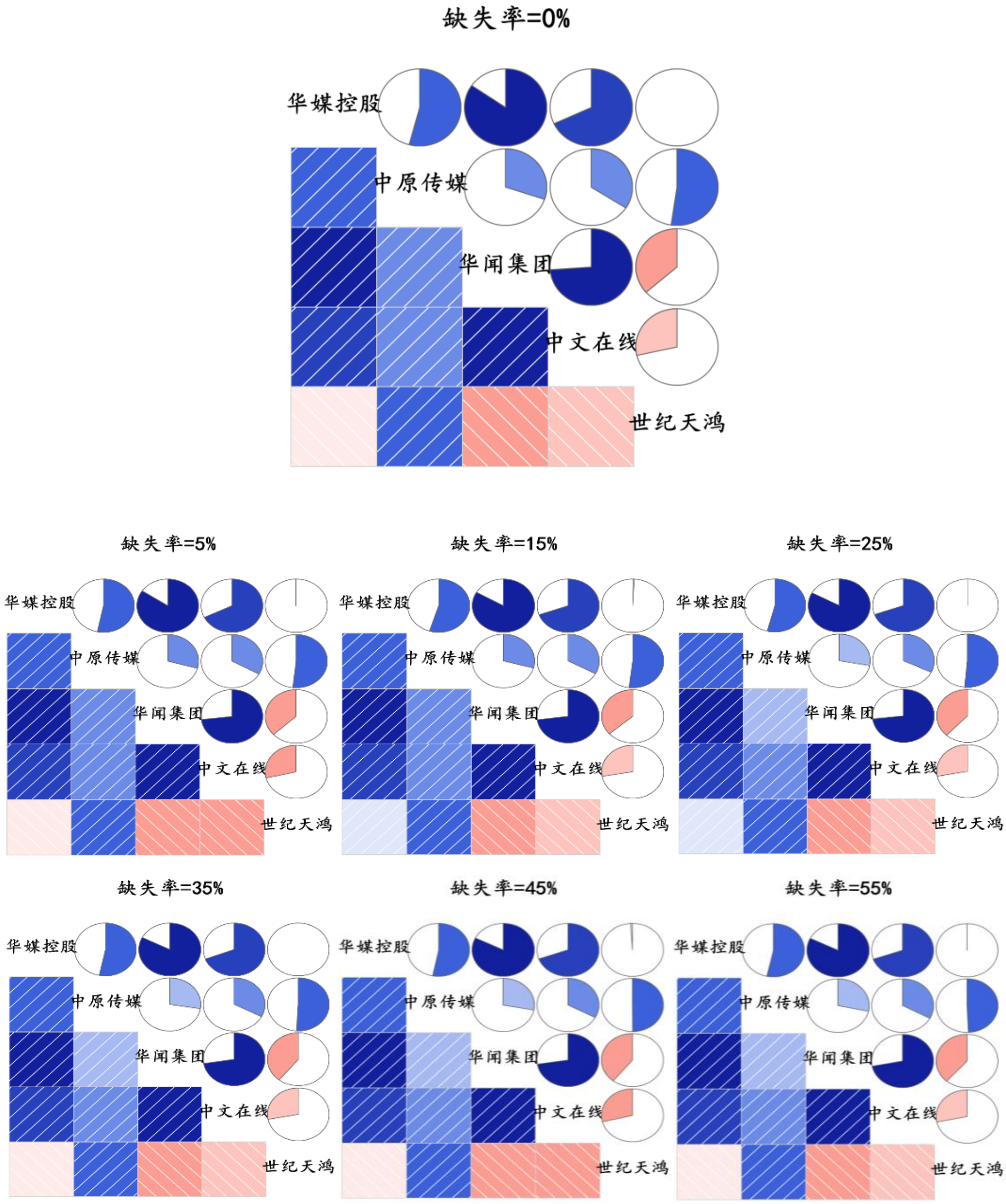


图 4.14 不同缺失率下插补后相关系数矩阵图

同时，以华媒控股为例绘制完整数据与不同缺失比例下插补后数据的散点图，如图 4.15 所示(其余股票见附录图 4.20~图 4.23)。从图 4.15 中可以看出，不同缺失比例下插补后得到的数据与完整数据的散点图分布基本一致，说明真实值与插补值之间的误差较小，即 MFGP 方法的插补性能较好，预测精度较高。

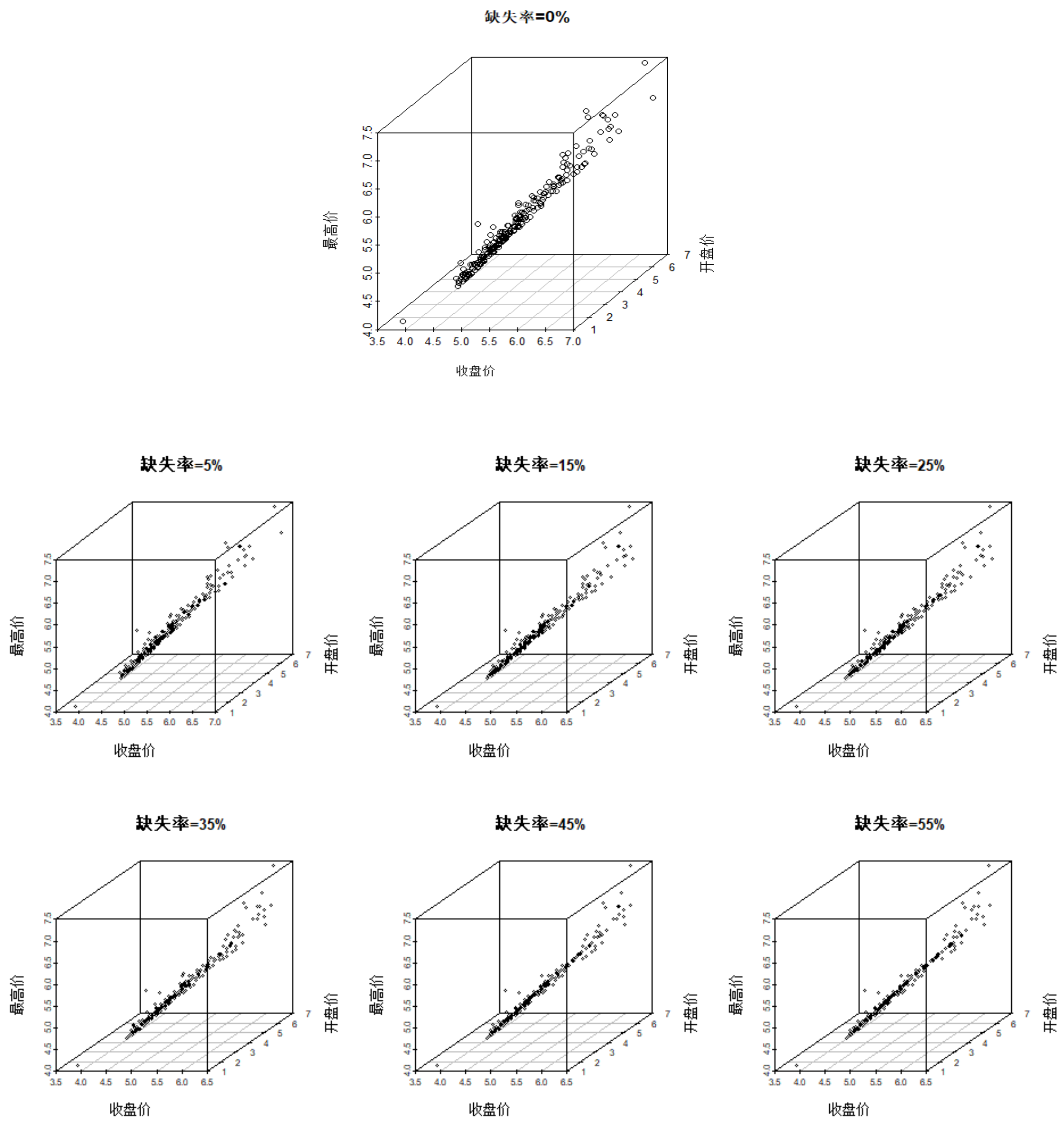


图 4.15 不同缺失率下华媒控股插补后散点图

4.4 本章小结

本章通过利用函数型数据的横截面信息和纵向信息，提出一种结合缺失森林插补模型和高斯过程预测模型的缺失插补方法，提高插补性能。通过模拟实验数据和实例数据对 MFGP 方法的插补效果进行验证，模拟实验结果表明，针对不同的缺失率(5%~55%)，MFGP 方法相较于 Hot.deck、均值插补、SFI、HFI、MICE、MF、GP 等 7 种插补方法，插补误差 RMSE 和 MAE 基本均保持在最低水平，说

明 MFGP 方法在插补函数型数据时具有一定的优势，能够保证插补的有效性和准确性；同时，通过股票数据对 MFGP 方法进行实例检验，结果表明，通过 MFGP 方法预测的缺失值符合原始数据的潜在规律，插补优势明显。

5 总结及展望

5.1 总结

随着函数型数据的大量出现，函数型数据插补问题得到了广泛关注。然而现有的函数型插补方法并没有充分考虑到原始数据的特征以及内部信息，以至于在处理函数型缺失数据时存在一定的局限，导致模型的适应能力较弱。针对上述问题，本文的主要研究内容和结论如下：

（1）首先引入样本类信息，并基于多重插补提出一种融合类信息的函数型多重插补方法(CMFP)。该方法通过借助样本相关性对缺失数据进行修复。首先采用模拟数据对 CMFP 方法的插补性能进行验证，其次利用 CMFP 方法对股票缺失数据进行实例应用分析。最终结果表明：相较于 Hot.deck、MF、均值插补、PACE、MFP 等 7 种方法，CMFP 方法具有显著的插补优势；同时，插补后得到的数据符合原始数据的潜在变化规律。

（2）提出一种基于横截面信息和纵向信息的函数型多重插补方法(MFGP)。通过结合缺失森林插补方法和高斯过程预测方法的优势，最大程度的利用函数型数据的样本信息，进而提高插补精度。MFGP 方法首先利用缺失森林对数据的横截面信息进行插补，其次利用高斯过程对数据的纵向信息进行预测，以此得到最终的插补值。分别通过模拟实验和实例分析表明：相较于其他传统插补方法，MFGP 方法具有较高的插补准确度，其插补可行性更高。

5.2 展望

尽管本文所提出的CMFP方法与MFGP方法在一定程度上克服了已有缺失值插补方法的局限，在数据插补任务中表现出较高的插补性能。但在实际数据中，由于其缺失机制和缺失模式情况复杂，且大多数据集存在噪声和异常。因此，在后续的工作中，我们将对以下方面进行研究：

（1）改进CMFP方法。在CMFP方法中，利用K-means聚类引入样本类信息，可考虑采用更适合样本数据的聚类方法代替，以此进一步提高算法的插补性能；改进MFGP方法。使其对于缺失率较大的数据集，有更稳定的插补误差，进一步提高插补方法的鲁棒性。同时，使改进的方法能适用于不同的缺失模式并克服异

常值和噪声的影响。

（2）CMFP方法和MFGP方法具有较高的可行性和插补适用性，可考虑应用于其他函数型数据的场景，如空气质量数据、医疗数据等。

参考文献

[1] Bertsimas D, Pawlowski C, Zhuo Y D. From predictive methods to missing dataimputation: an optimization approach[J]. Journal of Machine Learning Research, 2018,18(196): 1-39.

[2] Blazek K, van Zwieten A, Saglimbene V, et al. A practical guide to multiple imputation of missing data in nephrology[J]. Kidney International, 2021,99(1): 68-74.

[3] Chen L, Sun J. A multiple imputation approach to the analysis of current status data with the additive hazards model[J]. Communications in Statistics-Theory and Methods, 2009,38(7): 1009-1018.

[4] Chiou J M, Zhang Y C, Chen W H, et al. A functional data approach to missing value imputation and outlier detection for traffic flow data[J]. Transportmetrica B: Transport Dynamics, 2014,2(2): 106-129.

[5] Ciarleglio A, Petkova E, Harel O. Elucidating age and sex-dependent association between frontal EEG asymmetry and depression: an application of multiple imputation in functional regression[J]. Journal of the American Statistical Association, 2022,117(537): 12-26.

[6] Crambes C, Henchiri Y. Regression imputation in the functional linear model with missing values in the response[J]. Journal of Statistical Planning and Inference, 2019,201: 103-119.

[7] Faisal S, Tutz G. Multiple imputation using nearest neighbor methods[J]. Information Sciences, 2021,570: 500-516.

[8] Ferraty F, Vieu P. Nonparametric functional data analysis[M]. Springer, 2006.

[9] Gerg C G W, Tanner M A. Applications of multiple imputation to the analysis of censored regression data[J]. Biometrics, 1991,47(4): 1297-1309.

[10] Gertheiss J, Goldsmith J, Crainiceanu C, et al. Longitudinal scalar-on-functions regression with application to tractography data[J]. Biostatistics, 2013,14(3): 447-461.

[11] Ghassemzadeh S S, Jana R, Rice C, et al. Measurement and modeling of an ultra-

wide bandwidth indoor channel[J]. IEEE Transactions on Communications, 2004,52(10): 1786-1796.

[12] Gigl T. Low-complexity localization using standard-compliant UWB signals[D]. Austria: Graz University of Technology, 2010.

[13] Han W, Ji C, Chen Y, et al. Multiple imputation by chained equations for social data: 2nd International Conference on Computer Scienece and Technology(CST 2017)[C], 2017.

[14] Harezlak J, Wu M C, Wang M, et al. Biomarker discovery for arsenic exposure using functional data. Analysis and feature learning of mass spectrometry proteomic data[J]. The Journal of Proteome Research, 2008,7(1): 217-224.

[15] Heinze G, Ploner M, Beyea J. Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions[J]. Stat Med, 2013,32(29): 5062-5076.

[16] He Y, Yucel R, Raghunathan T E. A functional multiple imputation approach to incomplete longitudinal data[J]. Statistics in Medicine, 2011,30(10): 1137-1156.

[17] Horváth L, Kokoszka P. Inference for functional data with applications[M]. Springer Science & Business Media, 2012.

[18] James G M, Hastie T J, Sugar C A. Principal component models for sparse functional data[J]. Biometrika, 2000,87(3): 587-602.

[19] James G M, Sugar C A. Clustering for sparsely sampled functional data[J]. Journal of the American Statistical Association, 2003,98(462): 397-408.

[20] Jerez J M, Molina I, García-Laencina P J, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem[J]. Artificial Intelligence in Medicine, 2010,50(2): 105-115.

[21] Junninen H, Niska H, Tuppurainen K, et al. Methods for imputation of missing values in air quality data sets[J]. Atmospheric Environment, 2004,38(18): 2895-2907.

[22] Kokoszka P, Reimherr M. Introduction to functional data analysis[M]. CRC press, 2017.

[23] Madley-dowd P, Hughes R, Tilling K, et al. The proportion of missing data should

not be used to guide decisions on multiple imputation[J]. Journal of Clinical Epidemiology, 2019,110: 63-73.

[24] Mei P A, de Carvalho C C, Fraser S J, et al. Analysis of neoplastic lesions in magnetic resonance imaging using self-organizing maps[J]. J Neurol Sci, 2015,359(1-2): 78-83.

[25] Moons K G M, Donders R A R T, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred[J]. Journal of Clinical Epidemiology, 2006,59(10): 1092-1101.

[26] Pan W. A multiple imputation approach to Cox regression with interval-censored data[J]. Biometrics, 2000,56(1): 199-203.

[27] Preda C, Saporta G, Mbare M H. The NIPALS algorithm for missing functional data[J]. Revue Roumaine De Mathematiques Pures Et Appliquees, 2010, 55(4): 315-326.

[28] Ramsay J O, Ramsey J B. Functional data analysis of the dynamics of the monthly index of nondurable goods production[J]. Journal of Econometrics, 2002,107(1-2): 327-344.

[29] Ramsasy J O, Silverman B W. Functional data analysis[M]. 2nd. New York: Springer, 2005.

[30] Rao A R, Reimherr M. Modern multiple imputation with functional data[J]. Stat, 2021,10(1): e331.

[31] Rice J A, Wu C O. Nonparametric mixed effects models for unequally sampled noisy curves[J]. Biometrics, 2001,57(1): 253-259.

[32] Rubin D B. Multiple imputation for nonresponse in surveys[M]. New York: John Wiley and Sons,1987.

[33] Rumaling M I, Chee F P, Dayou J, et al. Missing value imputation for pm10 concentration in sabah using nearest neighbour method(nnm) and expectation-maximization(em) algorithm[J]. Asian Journal of Atmospheric Environment, 2020,14(1): 62-72.

[34] Seaman S R, Bartlett J W, White I R. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods[J].

BMC Med Res Methodol, 2012,12: 1-13.

[35] Stekhoven D J, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data[J]. Bioinformatics, 2012,28(1): 112-118.

[36] Twisk J, de Boer M, de Vente W, et al. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis[J]. Journal of Clinical Epidemiology, 2013,66(9): 1022-1028.

[37] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification[J]. Statistical Methods in Medical Research, 2007,16(3): 219-242.

[38] Wang H J, Feng X. Multiple imputation for M-regression with censored covariates[J]. Journal of the American Statistical Association, 2012,107(497): 194-204.

[39] Wei Y, Ma Y, Carroll R J. Multiple imputation in quantile regression[J]. Biometrika, 2012,99(2): 423-438.

[40] Yao F, Müller H G, Wang J L. Functional data analysis for sparse longitudinal data[J]. Journal of the American Statistical Association, 2005,100(470): 577-590.

[41] Zhang H S, Zhang Y, Li Z H, et al. Spatial-temporal traffic data analysis based on global data management using MAS[J]. IEEE Transactions on Intelligent Transportation Systems, 2004,5(4):267-275.

[42] Zhao F, Lu Y, Li X, et al. Multiple imputation method of missing credit risk assessment data based on generative adversarial networks[J]. Applied Soft Computing, 2022,126: 109273.

[43] 陈丽嫦，衡明莉，王骏，等. 定量纵向数据缺失值处理方法的模拟比较研究[J]. 中国卫生统计, 2020,37(3):384-388.

[44] 丁先文，陈建东，朱小芹. 带有缺失数据的分位数回归模型的参数估计[J]. 统计与决策, 2018,34(06): 65-67.

[45] 公徐路，李幸福. 响应变量缺失时纵向数据下部分线性模型的广义经验似然推断[J]. 统计与决策, 2019,35(14): 13-17.

[46] 黄恒君，漆威. 海量半结构化数据采集、存储及分析——基于实时空气质量

数据处理的实践[J]. 统计研究, 2014,31(05): 10-16.
[47] 金勇进. 缺失数据的插补调整[J]. 数理统计与管理, 2001(06): 47-53.
[48] 申停波. 缺失数据插补法的比较及实证分析[D]. 西南石油大学, 2017.
[49] 王霄, 王小宁, 刘允强. 问卷分割技术在市场调查中的应用及研究[J]. 统计理论与实践, 2020,(02): 17-25.
[50] 薛娇, 傅德印, 韩海波, 等. 基于多视角学习的非负函数型矩阵填充算法[J]. 统计与决策, 2022,38(07): 5-11.
[51] 张波, 宋国君. 大规模空气质量监测数据缺失处理方法实证研究[J]. 中国环境科学, 2022,42(05): 2078-2087.
[52] 张维群, 段格格. 基于多重插补的分层抽样估计方法与应用[J]. 统计与决策, 2023,39(02): 15-19.
[53] 张妍. 基于非参数多重插补的删失部分线性可加模型的分位数估计及其应用[D]. 浙江工商大学, 2018.
[54] 钟宇航. 基于广义Rescal分解的股票缺失值填充[D]. 西南财经大学, 2022.
[55] 张紫璇, 段红梅. 基于混沌游戏表示和自适应仿射传播聚类的股票板块分类[J]. 财会月刊, 2020(19): 152-155.

攻读硕士学位期间承担的科研任务及主要成果

发表或完成的论文：

[1] 高海燕,**李唯欣**,牛成英.基于矩阵填充的大型问卷调查数据缺失插补[J].湖北师范大学(自然科学版),2023,43(03):1-8.

[2] 高海燕,**李唯欣**,马文娟.基于缺失森林模型的稀疏函数型数据修复方法[J/OL].西华师范大学学报(自然科学版),2023,1-9.

[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,**李唯欣**.基于横截面和纵向信息的函数型多重插补方法[J].审稿中.

科研项目：

(1) **主持完成**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。

(2) 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)。

竞赛获奖：

(1) 第八届全国大学生统计建模大赛**甘肃赛区一等奖**：基于非负函数型数据填充方法的空气质量分析——以长三角地区为例，2022.08。

后　记

慕然回首，时光飞逝。三年忙碌充实的研究生生活即将画上句号。三年里，我得到了很多老师、同学和朋友的关怀和帮忙。在学位论文即将完成之际，我要向所有在此期间给予我支持、帮助和鼓励的人表示诚挚的谢意。

首先，特别感谢我的导师。从论文的选题、构思、撰写到最终定稿，都给了我悉心的指导和热情的帮助，使我的毕业论文能够顺利的完成。高海燕老师对工作的认真负责、对学术的钻研精神和严谨的学风，都是值得我终生学习的。

其次，感谢统计与数据科学学院的所有老师。正是他们的严格要求、无私教诲、才使我不断成长，不断提高自身的综合素养。饮水思其源，学成念吾师，谆谆教诲如春风，似润雨，永铭我心。谨以此向他们表示最诚挚的敬意和感谢。

再次，感谢我的父母和家人。父母不仅含辛茹苦的养育了我，还教会了我很多做人的道理，戒骄戒躁。正是他们的默默支持与鼓励，才使我在成长过程中少走了很多弯路，勇敢面对各种挫折。

最后，感谢我的同学们。同学们见证了我的成长，共同分享我的喜悦，倾听我的烦恼，帮我走出困境。不管是在学习中还是生活中，都给了我无私的帮助，让我感受到校园生活的美好。

“长风破浪会有时，直挂云帆济沧海”，用这句话作为这篇论文的一个结尾，也是一段生活的结束。期望自己能够继续少年时的梦想，勇往直前，永远无畏！愿大家都能成为想要成为的人！

附录

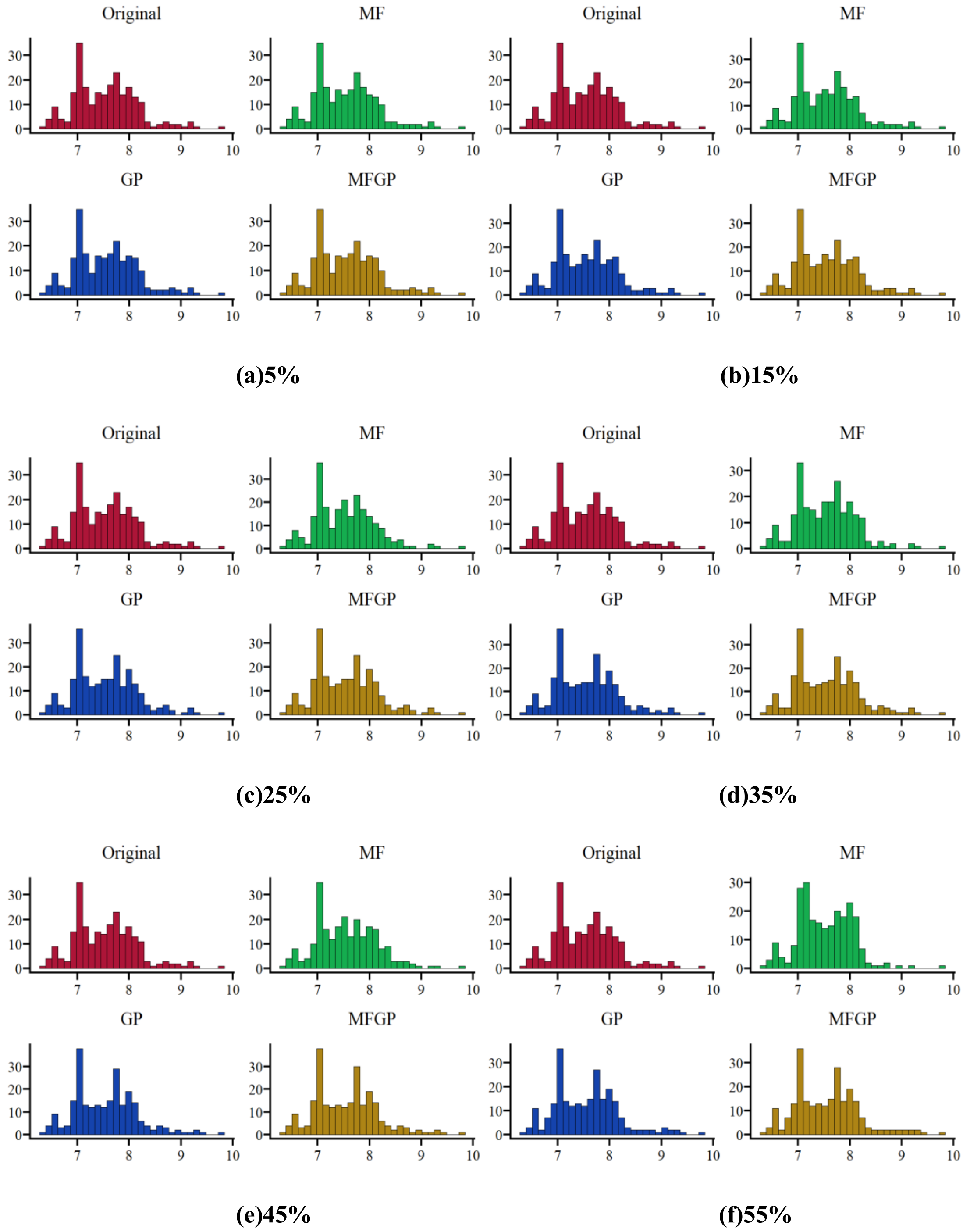


图 4.16　不同缺失率下中原传媒插补后分布直方图

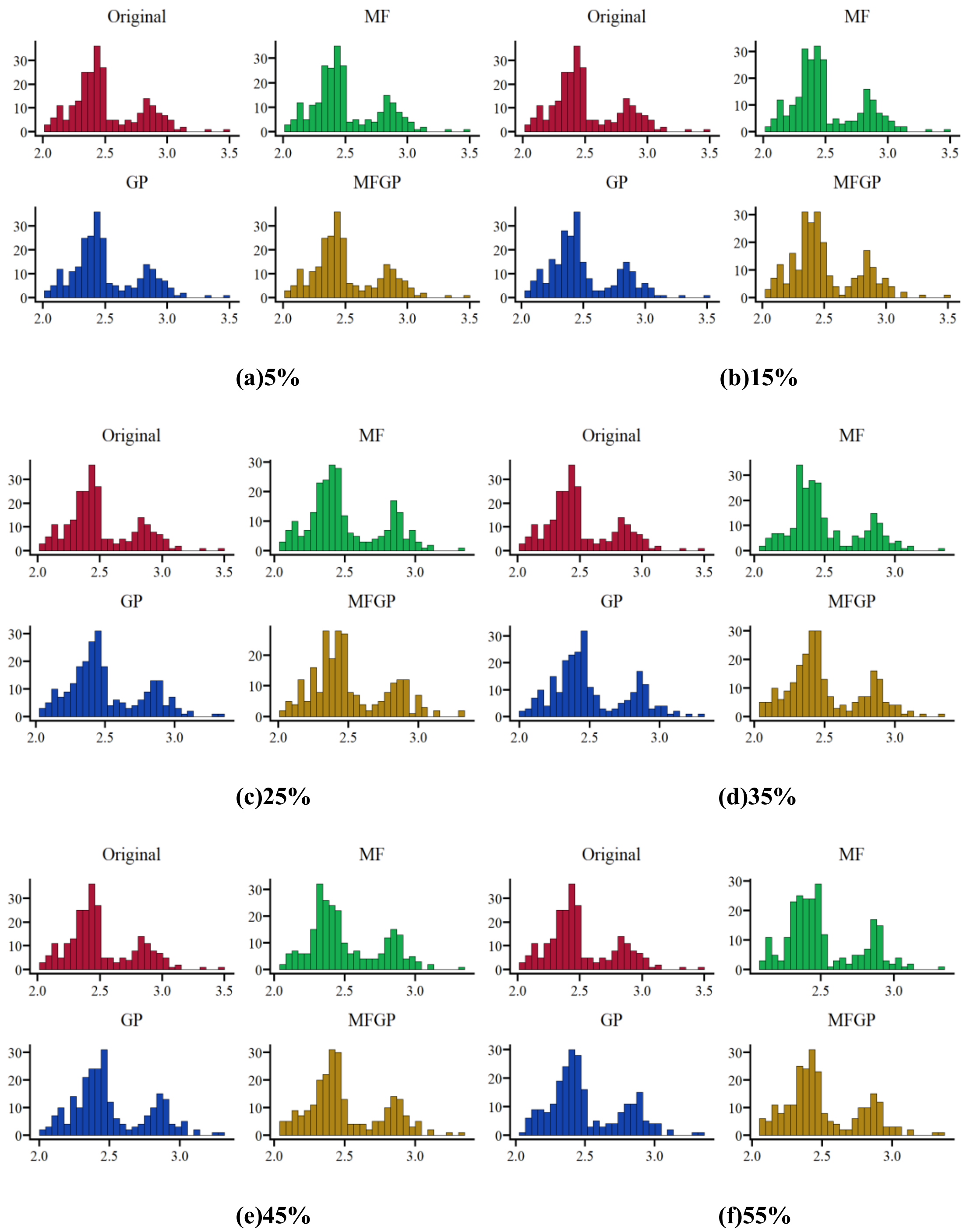
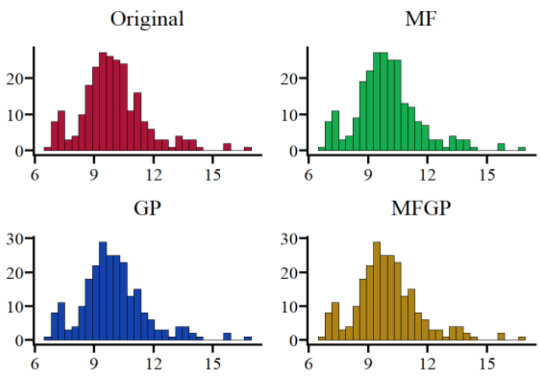
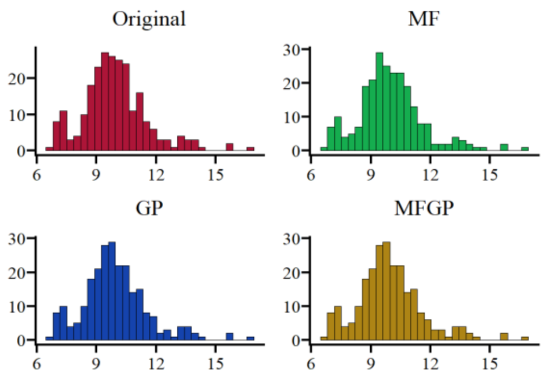


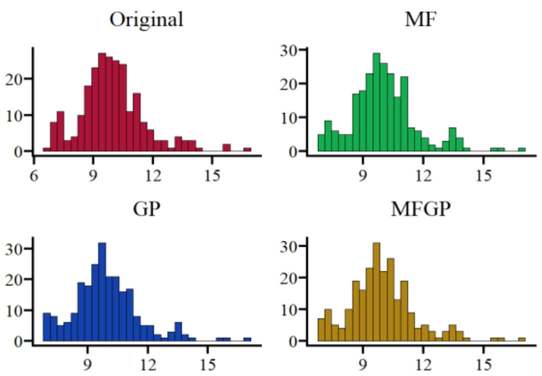
图 4.17　不同缺失率下华闻集团插补后分布直方图



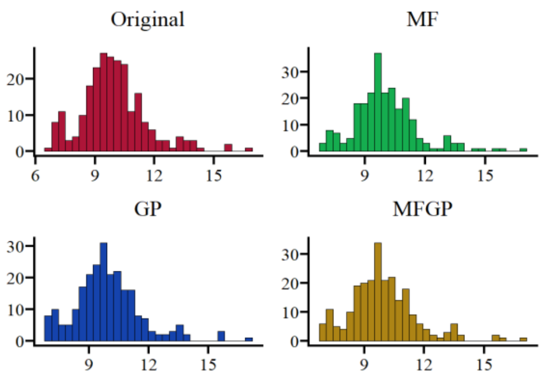
(a)5%



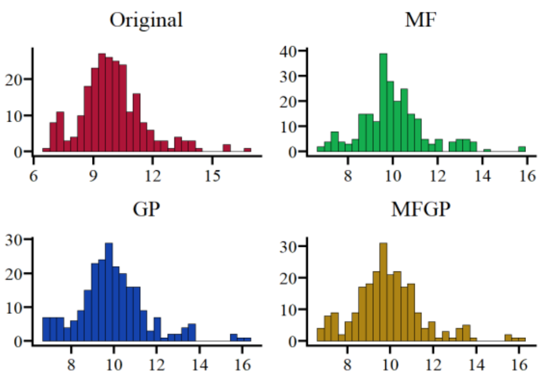
(b)15%



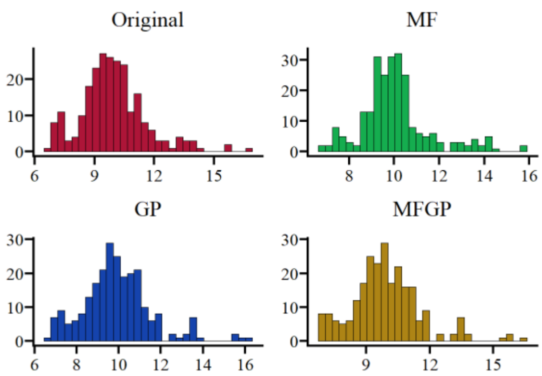
(c)25%



(d)35%



(e)45%



(f)55%

图 4.18　不同缺失率下中文在线插补后分布直方图

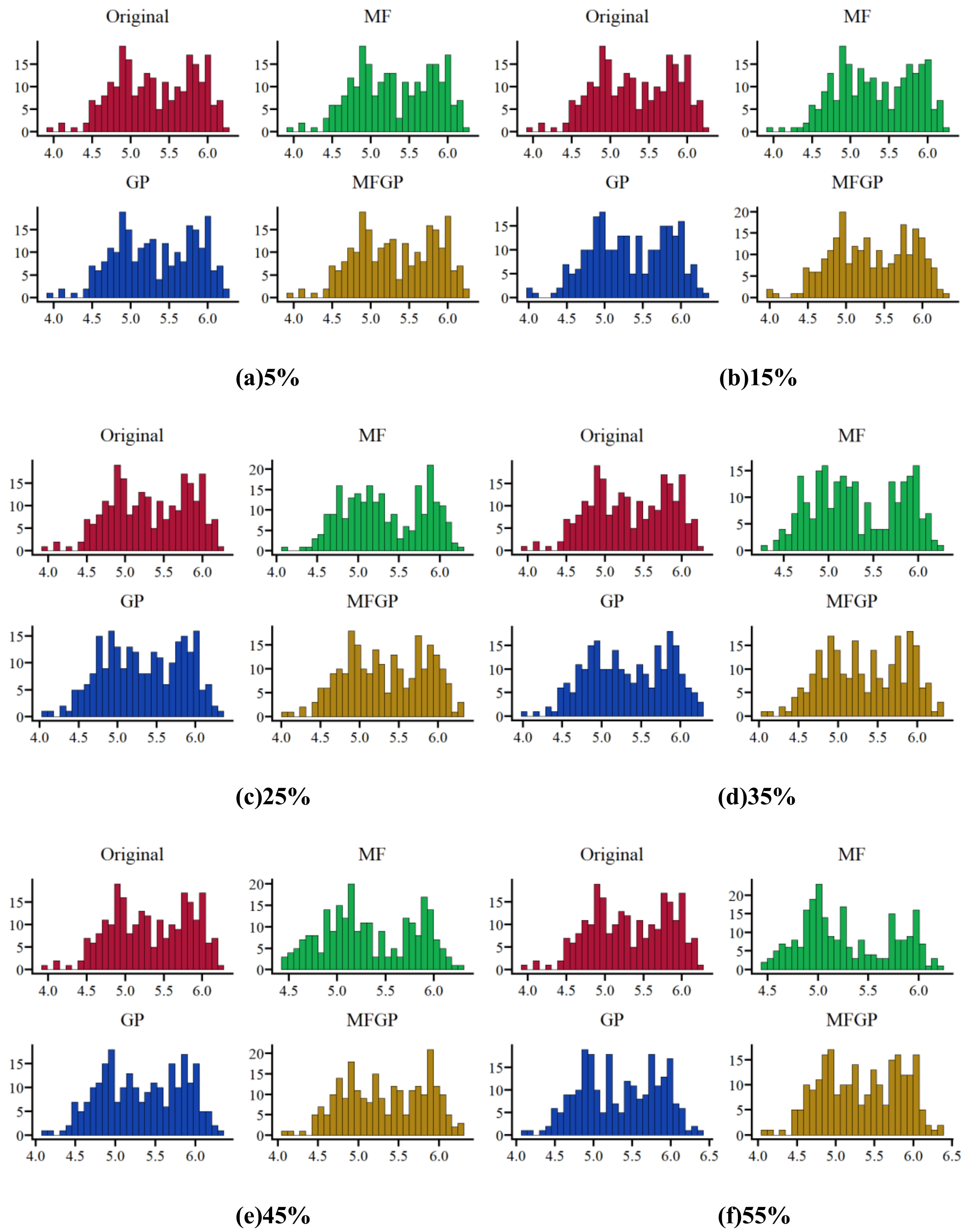


图 4.19　不同缺失率下世纪天鸿插补后分布直方图

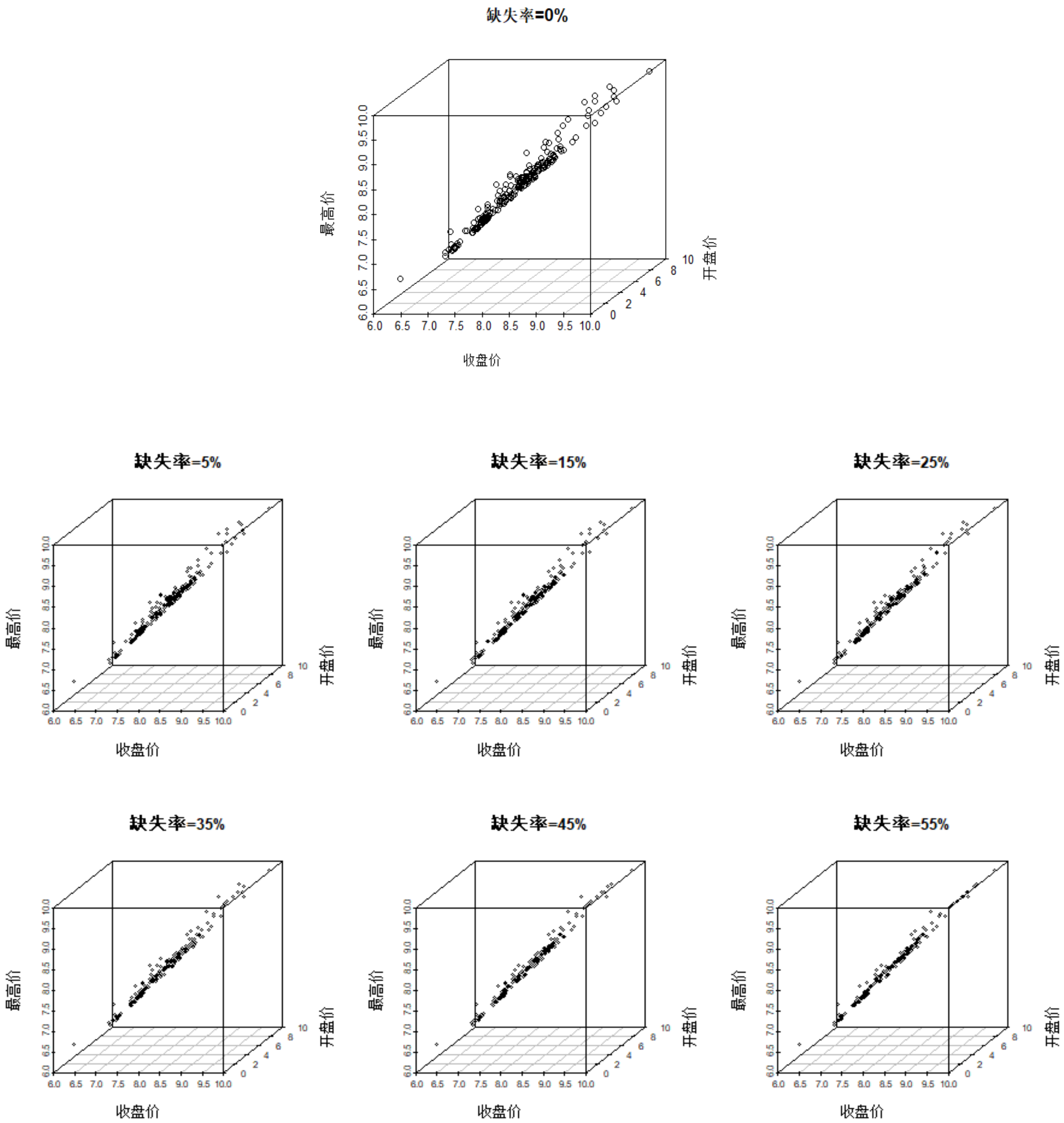
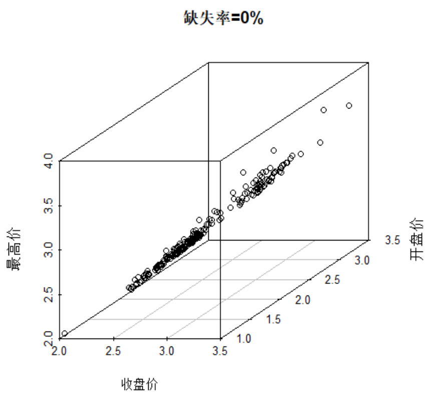


图 4.20　不同缺失率下中原传媒插补后散点图



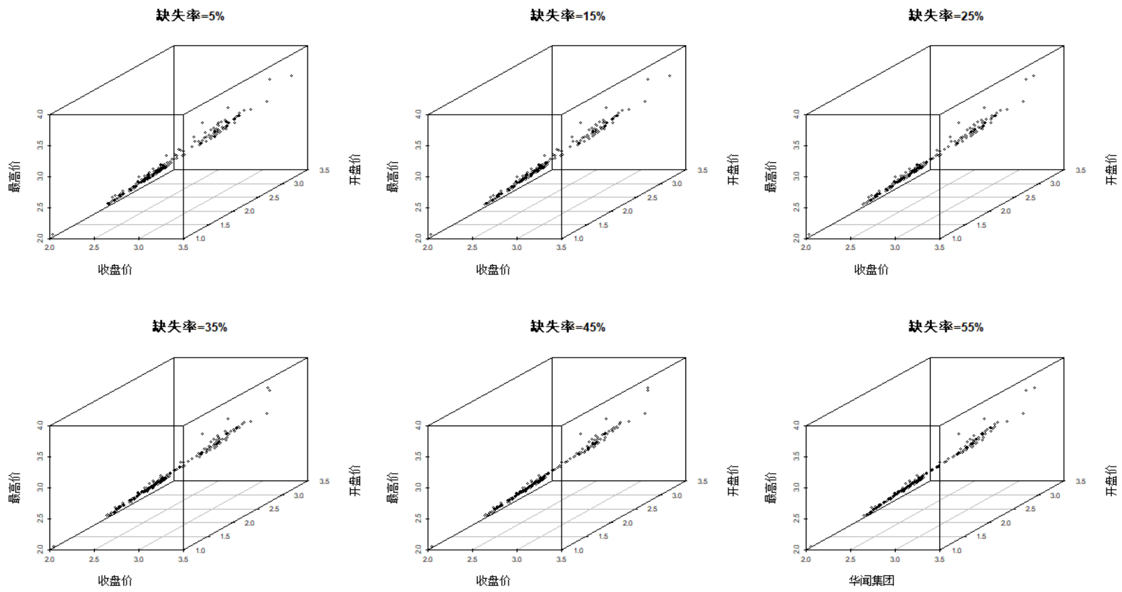


图 4.21　不同缺失率下华闻集团插补后散点图

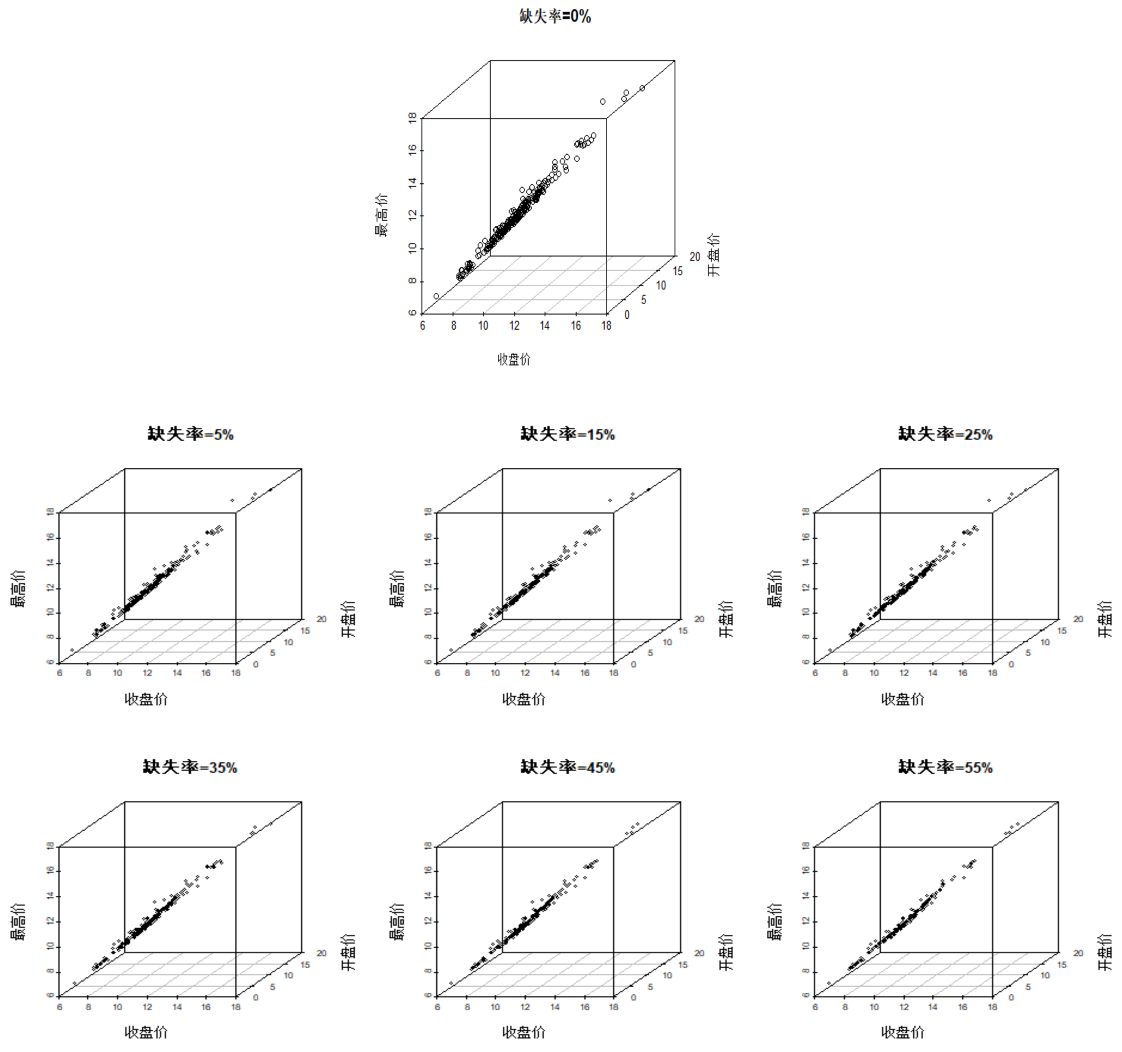


图 4.22　不同缺失率下中文在线插补后散点图

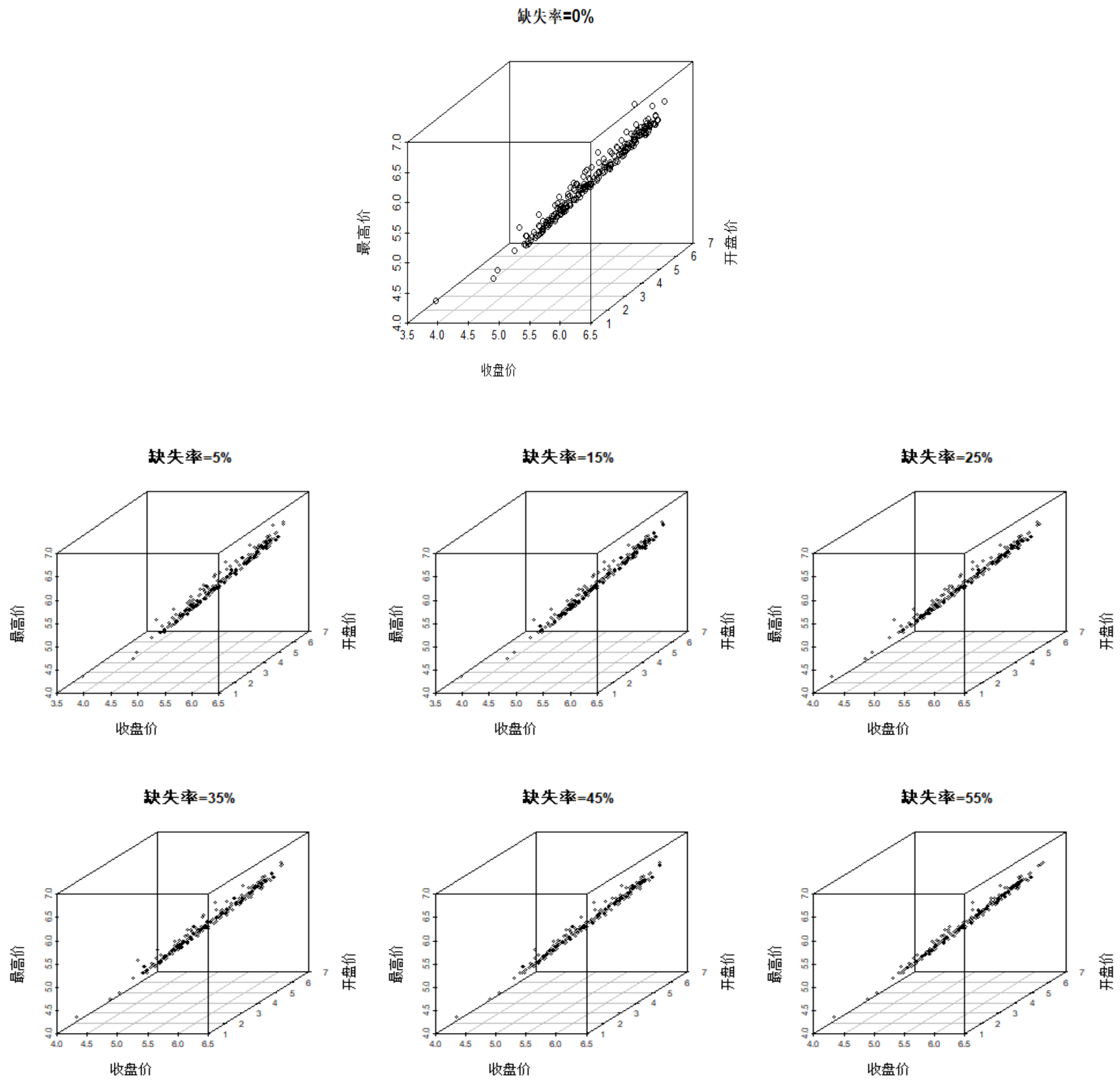


图 4.23　不同缺失率下世纪天鸿插补后散点图