人工智能合成数据与工商管理研究的融合：回顾与展望*

□ 彭正银　赵琰钰　李佳雷　王鑫阳

领域编辑推荐语：

合成数据正在成为人工智能时代破解“数据荒”、推动管理研究方法创新的重要工具。该文立足工商管理学科前沿，系统梳理合成数据的概念内涵、技术演进与应用脉络，运用 BERTopic 模型对 UTD24 和 FT50 期刊相关文献进行主题挖掘，归纳其在研究方法与实践应用中的核心议题，并对其方法论价值、适用边界与潜在风险进行了较为审慎的辨析。该文有助于拓展国内工商管理研究的方法视野，并推动学界形成关于合成数据应用的规范性讨论。

——张光磊

摘　要：人工智能驱动合成数据技术不断突破，有效缓解数智化时代的“数据荒”。合成数据通过优化研究设计与方法，嵌入多领域科学研究的创新过程，尤其在工商管理学科领域中，其为相关研究方法提供了先验证性条件，并缓解了部分研究中数据短缺问题。国际工商管理领域顶级期刊已广泛利用合成数据开展研究，但国内对合成数据的关注却十分不足。本文综合应用机器学习和知识图谱方法，开展系统文献分析。首先，本文界定合成数据基本概念，阐述其生成方式，分析研究现状与演进脉络。其次，精选 1993—2025 年发表于 UTD24 和 FT50 期刊上的相关主题文献，运用 BERTopic 模型对 133 篇英文文献进行文本分析，结合编码技巧凝练得到合成数据嵌入工商管理领域的两个核心议题：研究方法和实践应用。最后，从合成数据治理、学科领域应用和研究方法创新三个方面，展望合成数据在工商管理领域的研究方向。本研究为国内团队加入该领域国际对话提供应用框架，为进一步深化合成数据在工商管理研究中的应用提供方法论基础，具有重要的理论与实践意义。

关键词：工商管理；研究方法；合成数据；系统综述；主题模型

*　本文受国家社会科学基金重大项目“平台企业治理研究”（21&ZD135）和教育部人文社会科学青年课题“资源编排与共享发展理念融通视角下的共享经济平台可持续成长机制研究”（23YJC630080）资助。作者感谢审稿专家、领域编辑、主编为本文提供的专业性、建设性的指导意见，感谢编辑部老师的耐心、细致、理解与支持。

一、引言

人工智能合成数据（后文简称“合成数据”，Synthetic Data）技术是借助深度学习中的生成对抗网络等模型算法（Marano et al.，2024），基于对真实数据统计特征的学习（El Emam et al.，2020），生成保留原始数据模式或统计特征（Bellovin et al.，2019）的结构化、非结构化（Achuthan et al.，2022）或半结构化数据集的技术。这种高效且灵活的数据生成技术有助于破解当下一系列发展困境。数据作为数智化时代的关键要素，既推动经济社会发展，更为重大科学问题的突破提供了基础性支撑。当前各领域对于高质量数据源、数据处理和分析技术的需求与日俱增，兼之人们对于个人隐私、商业机密、知识专利等保护意识的增强，出现“不愿开放、不敢开放、不会开放”的高质量数据“数据荒”困境（端利涛和李海舰，2025）。作为社会科学中的重要分支，工商管理领域中的市场营销、财务会计等研究方向需依托大量真实数据，才能够对消费者行为偏好、市场定价、销售预测、成本控制及投资预算等问题展开准确分析。因此，数据也成为获取科研创新和科技竞争先发优势的基石（佟泽华等，2025），为相关研究的顺利推进提供了不可或缺的支撑。但目前数据“供给不足”的问题，制约了工商管理领域中相关学科的创新进程与成果价值的充分释放。2025 年，《国务院关于深入实施“人工智能 +”行动的意见》擘画了利用人工智能技术驱动科学技术进步、创新哲学社会科学研究方法的战略规划，明确提出要“打造开放共享的高质量科学数据集，提升跨模态复杂科学数据处理水平”。以大语言模型为代表的生成式人工智能技术快速发展，降低了知识获取与利用的门槛，在一定程度上消解了传统的信息与知识壁垒，对数据的获取、处理、管理、生成等带来了全新的变革，推动合成数据技术的演化与升级，为破解上述难题、推动国家战略实施提供抓手。

当前工商管理领域中相关研究数据主要依赖两类来源：其一为直接采用专业数据平台已整合的标准化数据，如 CSMAR 数据库、RESSET 金融研究数据库等；其二则是研究人员基于具体研究需要，借助 Python、八爪鱼等数据采集工具或软件，从相关网站或企业年报中自主抓取结构化或非结构化数据。这两种方式均建立在数据已公开披露且易于获取的基础上。然而，若企业或网站未公开相关数据，或虽公开但禁止采集，则研究人员往往难以针对相应问题开展实证分析，从而影响研究进程。合成数据在改进科学研究方法中的巨大潜力吸引了学者们的广泛关注，推动了交叉学科领域的研究进展。特别是 2014 年以来，随着生成对抗网络（Generative Adversarial Networks，GANs）和变换模型（Transformers）等深度学习技术的发展，由人工智能驱动的合成数据技术逐步得到工商管理领域国际主流期刊的关注，这有助于突破工商管理研究领域局限，更推动合成数据在工商管理中的应用方式与价值逐步成为该领域的重要议题，并带动相关发文量持续增长。可惜的是，在中美两国领衔全球人工智能技术发展的今天，国内工商管理领域对合成数据的关注严重不足，尚处于初步探索阶段。回顾

CNKI 文献数据库中的相关文献，当前国内对于合成数据的研究主要集中于计算机科学和法学相关领域，其中法学多聚焦于合成数据治理的法律性问题，因此缺少对合成数据在管理学尤其是工商管理领域的应用场景与研究价值的系统分析。

基于此，本文从三个层面展开研究。第一层面，本文阐释了合成数据的基本特征与生成方式，总结具体应用方式与场景，借助 CiteSpace 软件对现有研究进行系统回顾，厘清发展脉络。第二层面，采用 BERTopic 模型和人工复核方式对发表于 UTD24 和 FT50 期刊上的 133 篇合成数据主题相关文献进行机器学习文本分析，凝练出合成数据与工商管理研究融合的两个重要研究主题，识别相关潜在使用风险，述评主要研究结论。第三层面，基于 BERTopic 模型分析结果，从合成数据治理、学科领域应用和研究方法创新三个方面，对合成数据未来研究做出展望。本文潜在的边际贡献如下：第一，厘清合成数据的概念特征、生成方式、演进阶段及发展脉络，构建国内关于合成数据的现有分析思路。第二，综合应用机器学习和知识图谱分析方法，将 BERTopic 模型和人工复核相结合，扎根于工商管理研究领域，识别出合成数据与工商管理领域研究融合的两大核心议题。第三，通过融合机器学习、文本挖掘、文献对话和理论探索等研究手段，实现技术与理论、应用相结合，进而讨论合成数据在工商管理领域中的具体应用方式，尝试提出合成数据在工商管理研究方法中的使用流程以及评估指标，为国内团队加入该领域国际对话提供应用框架，为进一步深化合成数据在工商管理研究中的应用提供方法论基础。

二、概念界定与文献检索

（一）合成数据概念界定

合成数据尽管目前已被应用于学术研究和商业活动中，但学界对于合成数据的概念尚未达成共识。当前学者们从生成技术和数据来源两个视角对合成数据的概念进行界定。从生成技术视角看，有学者将合成数据视为应用多重插补方法所生成的数据集，其主要作为一种原始数据的替代，用于解决数据缺失问题（Keller et al.，2016）。另有学者认为，合成数据是深度学习模型中生成对抗网络所生成的代表实际样本特征的数据集（Marano et al.，2024）。从数据来源视角看，合成数据是基于真实的非结构化数据（如图像、文本、音频或视频）以及真实的结构化数据（如表格数据）创建的数据集（Achuthan et al.，2022）。

通过对不同观点的分析，可以发现合成数据具有以下几个明显特征：一是基于真实数据及其统计特征；二是需通过算法、模型等特定技术生成（张涛，2023）；三是数据形态多样。基于此，本文将合成数据定义为：旨在替代真实数据，实现隐私保护、模型训练、数据分析等任务，由特定生成模型、算法等技术创建的非真实数据。这些特定生成模型和算法等技术由真实数据训练而成，因此，这一非真实数据同样具备真实数据的统计特征。

合成数据生成过程如图 1 所示。合成数据借助生成对抗网络（GANs）、变分自编码器（VAEs）等深度学习技术，学习真实数据的内

在关联和分布特征，生成的数据既具备真实数据统计特征，又能实现数据脱敏，这有效解决了隐私保护与数据利用之间的矛盾（Jordon et al.，2022）。因此，合成数据当前多被应用于计算机、金融等数据需求大且保密性强的相关领域，对于工商管理领域中的应用较少。

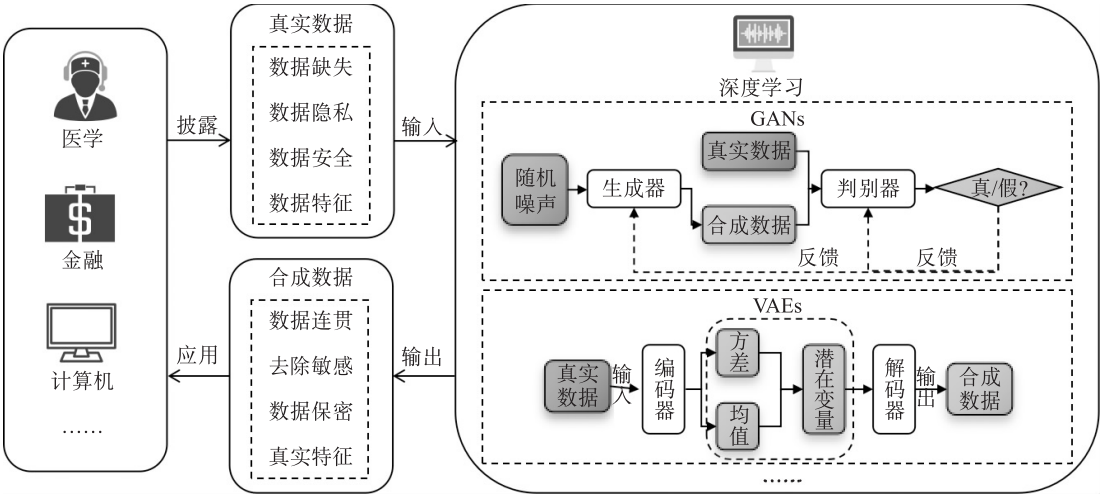

图 1　合成数据生成

资料来源：作者整理绘制。

合成数据最初生成方式主要基于概率统计方法，如高斯混合模型、协方差建模、联合分布拟合等方式对数据分布进行建模，并依据所建分布生成新的数据样本。然而，在面对复杂、高维数据时，该方法的性能往往面临限制（郭亚楠等，2026）。随着数字技术的不断发展，近年来，以生成对抗网络（GANs）和变分自编码器（VAEs）为主的深度学习不断展现出其在合成数据技术上的优势。生成对抗网络（GANs）（Goodfellow et al.，2014）由“生成器”和“判别器”组成，通过二者间的持续对抗训练，各自优化并提升能力，逐渐生成高保真的合成数据。其中，“生成器”通过持续学习真实数据的特征，生成符合真实数据统计特征的合成数据，“判别器”则准确地区分合成数据和真实数据，并在此过程中提升判别精度。

变分自编码器（VAEs）（Kingma & Welling，2013）则主要由“编码器”和“解码器”组成，其基本原理是“编码器”通过学习原始数据的潜在表示，找到潜在变量的分布特征，如方差、均值等，并形成潜在空间，“解码器”将从潜在空间中采集的潜在变量映射至数据空间，从而生成与原始数据具有相似统计分布特征的合成数据（叶英杰和李川，2025）。此外，还有一些包括计算图、自动微分、优化算法等深度学习框架组件，为设计、训练和部署深度学习模型提供了基础，这些技术的运用都能够生成与真实数据特征相似的合成数据。

（二）文献检索

本文主要聚焦于“工商管理”领域相关文献。依据教育部门发布的学科专业目录，“工商管理”属于一级学科，其包括会计学、企业管理（含财务管理、市场营销、人力资源管理）、旅游管理和技术经济与管理等方向。当前国际

管理学顶级期刊UTD24和FT50也覆盖了工商管理领域主要研究方向。因此，本文的文献研究范围主要聚焦于工商管理领域，这与本文的研究目的也相契合。

本文英文文献数据主要来源于Web of Science（后文简称“WOS”）数据平台中核心合集（Core Collection）数据库的SCI和SSCI两部分。鉴于合成数据最初由美国统计学家Rubin于1993年提出，因此时间跨度限定为1993—2025年，检索主题（Topic）为“synthetic data”（合成数据），共搜索到19669篇，进一步通过对管理学、商业经济等研究领域的筛选，最终获取254条符合主题的文献，导出“全纪录和引用的参考文献”信息，以此作为研究团队厘清合成数据在工商管理领域中发展情况的数据基础，文献筛选流程如图2所示。基于此，通过WOS平台的数据精炼功能进一步从UTD24和FT50中筛选出133篇文献，作为本研究分析的数据基础。本文对于中文文献的搜索方式主要通过中国知网，使用主题词为“合成数据”进行文献检索，时间为1993—2025年，检索范围包括《管理世界》《管理科学学报》等18本国内权威期刊，最终并未检索到相关文献。因此，本文对学科检索后的254条英文文献及引用文献开展知识图谱分析，借助CiteSpace工具进行文献共被引分析，把握合成数据在工商管理领域中的整体研究脉络。

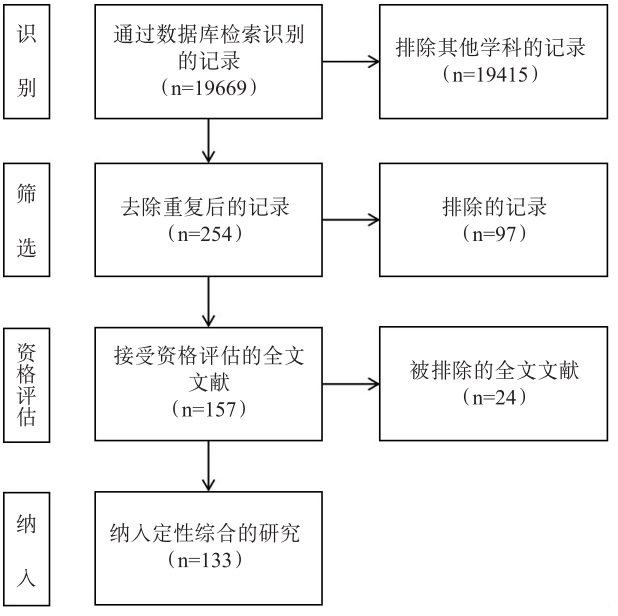

图2　文献筛选流程

资料来源：作者整理绘制。

（三）研究概况

1. 发文概况分析

合成数据在工商管理领域中的发文量整体呈上升趋势。如图3所示，2020年之后，合成数据的相关研究成果较之前年度有明显增加，尽管2023年出现波动，但整体发文量和发文趋势呈现增长状态。强化学习（Reinforcement Learning）、对抗性神经网络（Dueling Neural Networks）、灵巧机器人（Robot Dexterity）、差分隐私（Differential Privacy）、GPT－3、多技能

AI（Multi－skilled AI）、AI 数据生成（Synthetic Data for AI）、制作图像的 AI（AI that makes Images）等技术与合成数据的融合推动了合成数据研究进一步爆发[①]，带动了工商管理领域内学者对合成数据的使用和研究。

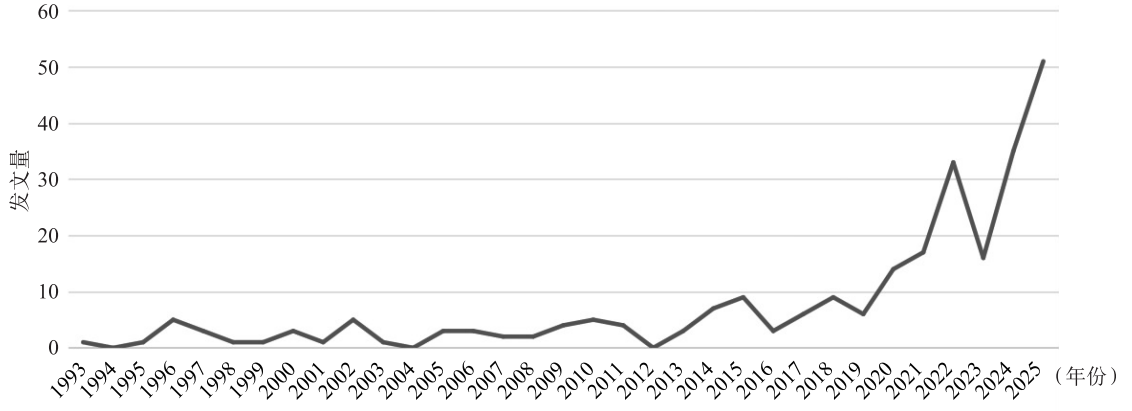

图 3　1993—2025 年应用合成数据的工商管理文献发文情况

资料来源：作者整理绘制。

关键词方面，“模型”与“算法”在合成数据的开发与应用流程中对数据质量起着关键作用。研究人员试图通过一种简单快捷的方式获取满意的合成数据，从而推动“模型”与“算法”成为该领域研究的核心焦点。同时，研究人员持续提出新颖的研究视角并优化数据生成方法，使得诸如“系统”“优化”等术语逐渐成为学术讨论中的常见词汇，且这些词汇在共现网络中也具有较高的中心度。随着计算机技术与人工智能领域的迅猛发展，合成数据的生成技术也进入了新的阶段，研究人员利用“深度学习”和“机器学习”等先进技术手段，生成出更加逼真且全面的合成数据，为研究与实践过程中高难度问题的猜想、验证与解释提供了新的可能，该术语也呈现出较高的中心度。此外，聚焦于合成数据在工商管理领域中的具体研究问题，使得“需求”“品牌选择”“行为”“市场”等词汇成为该研究方向的代表性关键词。国家发文方面，由于美国在统计学与计算机科学领域的应用实践较早，技术发展较为成熟，因此美国在该领域的研究发文时间也较早。尽管其他国家或地区的发文时间相较美国而言较晚，但随着对科研项目的资金、人才和技术的支持，部分国家或地区如丹麦、法国、德国等在发文的质量和重要性方面已超越美国，且在当前发文国家中，丹麦的研究价值重要性位居首位，其次是法国、智利、德国等。这一现象充分体现了不同国家在合成数据研究领域中研究价值、地位和影响力的持续提升。

2. 演进阶段及特征分析

本文借助 CiteSpace 软件对搜索到的文献进行分析，在此基础上，结合 Rubin（1993）系统地提出合成数据的概念和方法框架，以及整体发文情况，将合成数据的发展划分为三个演进

① 参见《麻省理工科技评论》2017—2023 年发布的“全球十大突破性技术”榜单，来源网址：https：//www. mittrchina. com/search.

阶段，如图 4 所示。通过对演进阶段的分析发现：合成数据的演进过程呈现出“应用浅层与框架碎片—内涵与技术逐渐丰富—技术与管理场景双重驱动”的递进演进特征。

阶段	第一阶段（1993~2002年）	第二阶段（2003~2014年）	第三阶段（2015~2025年）
背景	美国统计学家Rubin首次对合成数据概念、应用设想等方面进行描述	差分隐私、深度信念网络、变分自编码器、生成对抗网络等技术的产生	变分自编码器、生成对抗网络等技术推动合成数据的应用
特征	应用浅层与框架碎片	内涵与技术逐渐丰富	技术与管理场景双重驱动
应用	• 验证所提出的贝叶斯分层模型在恢复真实参数方面的表现； • 评估空间分割模型在数据生成过程是否具有空间结构时的表现	• 评估改进后的算法在不同行为假设下的表现； • 验证所提出的认知成本模型的可估计性； • 研究空间权重矩阵的选择对模型识别和参数估计精度的影响	• 设计并验证基于合成数据共享的金融数据生态系统，解决金融机构在隐私保护下的数据稀缺； • 介绍合成数据在健康经济学与结果研究中的潜在应用

图 4　合成数据发展演进阶段

资料来源：作者整理绘制。

第一阶段主要为合成数据在工商管理领域中应用的初步尝试。由于合成数据作为新型概念的提出，其需要拥有相关技术进行数据生成，因此多被应用于计算机领域。工商管理领域多采用实证、理论概念等相关研究，学者并未对合成数据有深刻的认识和研究，使得早期合成数据在工商管理领域内的研究未形成一个完整且系统的研究框架。尽管部分研究人员尝试通过合成数据丰富其研究内容，但主要利用合成数据对其提出的研究模型进行表现效果评估。如 Ter Hofstede 等（2002）在 *Marketing Science* 上发表了一篇利用生成的两类合成数据对文章中所提出的两种模型进行分析和检验，从而验证了空间先验信息在分割模型中的有效性和稳健性。

第二阶段随着计算机科学和人工智能相关研究的发展，在合成数据基本内涵形成的基础上，其生成技术不断更新，先后产生了差分隐私（Differential Privacy）、深度信念网络（DBNs）、变分自编码器（VAEs）、生成对抗网络（GANs）等生成技术。而在工商管理领域中，研究人员利用不同生成技术产生合成数据，评估模型、算法在不同场景下的表现，验证模型、算法的原理及其有效性，并探究算法参数的具体影响。如 Hauser 等（2014）在 *Management Science* 上发表的文章中基于真实的研究参数，模拟消费者在网站上的点击行为、形态切换等过程，通过使用合成数据，比较 WTM 算法和 HULB 算法。合成数据生成技术的优化逐步丰富了合成数据的内涵与生成方式，但其研究和应用场景却相对单一，在工商管理领域的其他研究场景中的应用则相对较少，尚未充分挖掘合成数据在工

商管理研究中的潜力。

第三阶段受变分自编码器（VAEs）、生成对抗网络（GANs）等技术推动，合成数据应用场景趋于广泛，并优化与拓展了合成数据的生成方法。为契合实际场景需求，研究人员持续优化并迭代数据生成模型，丰富了合成数据的应用类型。在技术与场景的双重驱动下，合成数据的生成质量不断提升、数据类型日趋多样，其应用场景也逐步从统计学、计算机科学等领域，向市场营销、技术经济与管理等更广泛的领域延伸扩展，如在消费者行为研究中，合成数据可用于模拟消费者的决策过程与消费策略（Lin et al.，2015），而在技术经济与管理领域，它则被用于定价问题的研究（Ciocan & Misic，2022）。

三、主题挖掘

本文对所收集到的论文题目、摘要和关键词进行了整理，形成原始数据集，并将其作为主题识别的主要分析依据（聂亚青等，2024）。主题识别主要通过 BERTopic 模型实现。此外，BERTopic 模型是由研究人员 Grootendorst 于 2022 年提出的一种主题模型。该模型基于已处理好数据，通过使用 BERT 嵌入提升语义理解，在文档主题提取前利用 UMAP 降维，保留数据结构，HDBSCAN 算法则对数据形成聚类。基于此，采用 c – TF – IDF 模型提取主题。相较于 LDA 等传统主题模型，BERTopic 在主题识别与文本分类等任务中展现出明显优势。其不仅能够自动确定最优主题数量（Egger & Yu，2022），还能生成语义更明确、更具可解释性的主题内容，有效克服了传统方法在密度聚类与中心采样机制之间存在的兼容性问题（曹树金和曹茹烨，2023）。

（一）数据采集与数据处理

为聚焦合成数据在工商管理领域的研究，本文基于 Web of Science 核心合集数据库，对 UTD24 和 FT50 两类顶级期刊列表进行了系统检索。经过逐层筛选与剔除，最终得到 133 篇有效文献作为本研究的样本数据。并将所获文献的标题、摘要与关键词等文本内容作为初始语料，由于文献主要为英文文献，因此，主要利用英文分词器 NLTK 分词对文献进行英文分词处理，通过分词处理、去停用词等一系列文本预处理方法，对初始语料进行清洗与标准化，使其符合后续模型分析要求。

（二）研究方法与主题生成

本文采用 BERTopic 模型和人工复核进行主题识别，主题识别分析过程如图 5 所示。BERTopic 模型的使用流程主要包括嵌入、降维、聚类和主题提取四个关键环节。嵌入主要通过 BERT 中的 Sentence Transformers 模型进行词嵌入处理，理解文本内容所包含的语义，对输入词汇的对应向量进行确认，如“bank”在不同语义情境中可翻译为银行、库存、河岸等，词嵌入处理有效避免了词意混淆。在降维部分采用 UMAP 算法，其针对词嵌入结果，可进行非线性处理，并在降维过程中保持语义相似文档的局部和全部结构。在聚类部分，采用 HDBSCAN 密度聚类算法，针对降维向量可自动识别不同聚类下的主题。在主题提取部分，主要利用 c – TF – IDF 模型提取不同主题下具有主题代表性的高频词汇，以及相关主题。最终从相关

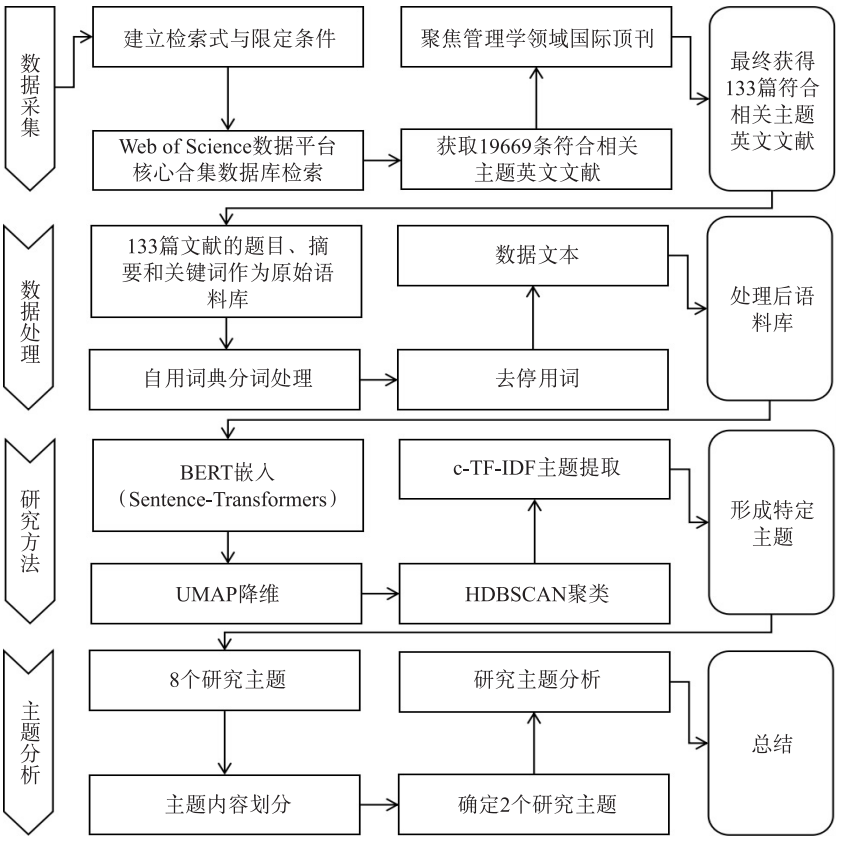

图 5　主题识别分析过程

资料来源：作者整理绘制。

文献中共提取出 8 个研究主题及主题特征词分布，如图 6 和图 7 所示。

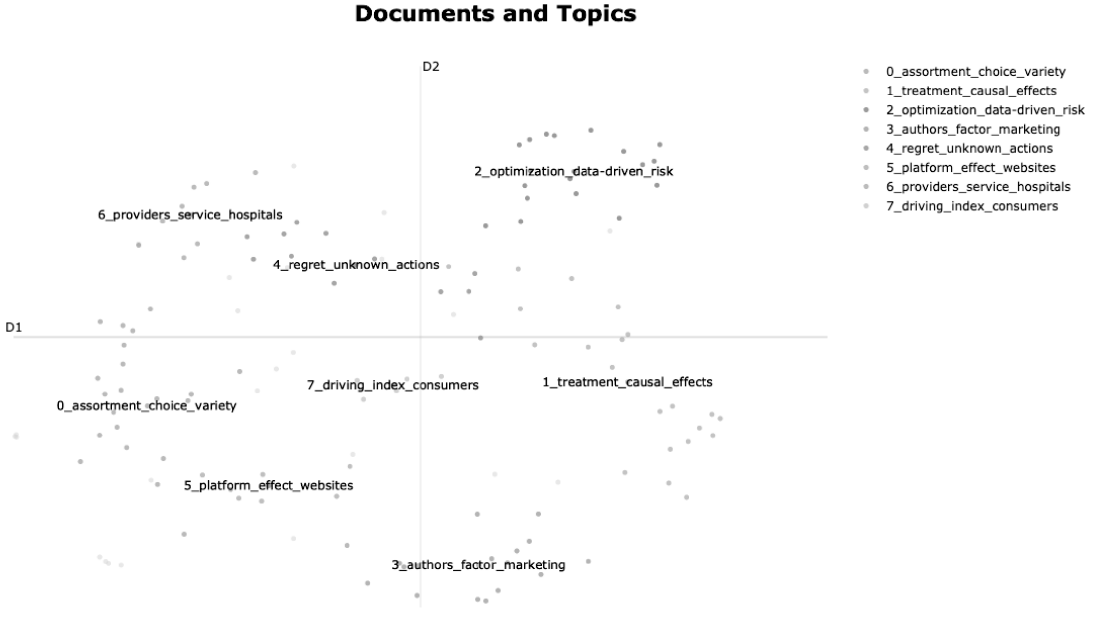

图 6　主题分布

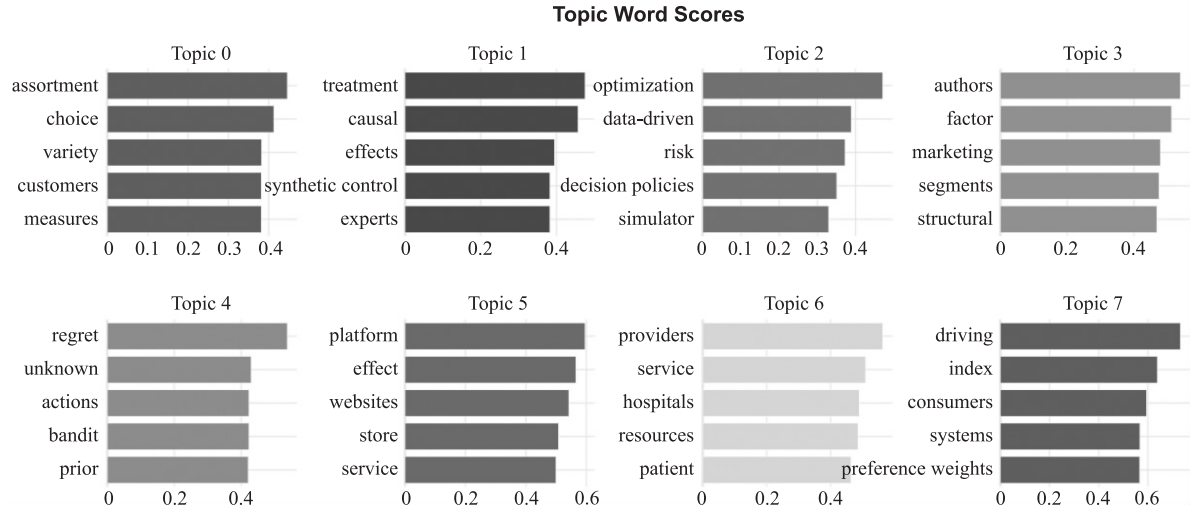

图 7　各主题特征词

基于 BERTopic 模型所提取的 8 个研究主题，本文再次通过人工根据 8 个研究主题的关键词以及文献分类情况进行复核，针对主题特征词和文献相似度的情况进行合并，通过对各研究主题及内容的二次分类和归纳，本文将 8 个研究主题提炼为两个主要研究主题，如表 1 所示。

表 1　　BERTopic 主题聚类结果

二级主题	一级主题	关键词	生成主题
主题一：研究方法	鲁棒优化与决策生成	Optimization，data－driven，risk，decision policies，simulator，generative，scenarios，exact，convex，training	主题 2
	因果推断与实验设计	Treatment，causal，effects，synthetic control，experts，estimator，learn，pay，control	主题 1
	市场空间细分	Authors，factor，marketing，segments，structural，heterogeneity，survey，missing，variables，respondents	主题 3
主题二：实践应用	消费者选择与决策	Assortment，choice，variety，customers，measures，assortments，pricing，prices，revenue	主题 0
	动态定价	Regret，unknown，actions，bandit，prior，achieve，bound，horizon，protection，stochastic	主题 4
	其他研究	Platform，effect，websites，store，growth，firm，public，retailers，preference weights，resources，patient，distributed，kidney	主题 5 + 主题 6 + 主题 7

资料来源：作者整理绘制。

基于对关键主题的识别，并结合工商管理领域的研究方法特性，本文构建了合成数据研究内容框架，如图 8 所示。该框架主要围绕以下两个核心主题展开探讨：合成数据如何在工商管理领域中运用是学者们日益关注的话题，在已有研究中，合成数据往往被用于与当前主

要研究方法相结合，因此本文确立主题一：“研究方法”。尽管研究人员将合成数据用于当前部分问题的研究，但合成数据能够被用于哪些研究也是学者们所关心的主要话题，由此引出研究主题二：“实践应用”。

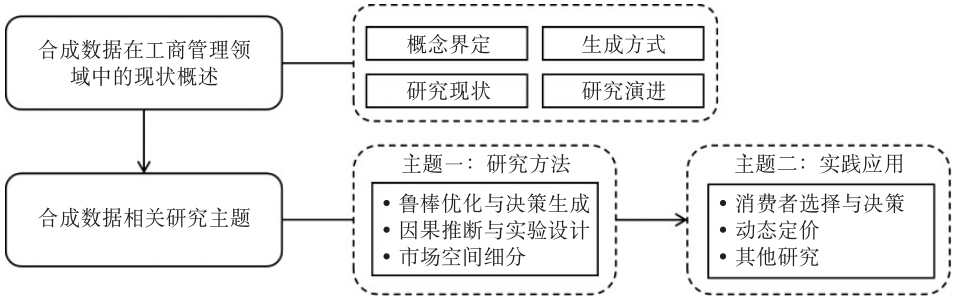

图 8　合成数据研究内容框架

资料来源：作者整理绘制。

四、研究主题

（一）研究方法

工商管理领域研究主题和内容多样，使得研究方法也具有多样性，同时也对数据质量提出了更高要求。由于合成数据具备真实数据的关键属性及特征，从而在研究应用中实现与真实数据相当的研究效果，为工商管理领域多样化的研究方法体系提供了有效的数据支撑与验证辅助。

1. 鲁棒优化与决策生成

鲁棒优化与其他优化方法相比，更多被使用在信息环境不完全或不确定性较强的研究问题中，因此是针对不确定性问题的一种优化方法，其主要通过考虑模型参数的不确定性，从而在预设的不确定性集合内构建各类场景下均具备稳健性的最优解（孟祝等，2025）。决策生成基于特定研究场景，通过构建决策模型，生成符合该场景需求的相应决策。

合成数据由研究人员基于原始数据集对生成器进行训练（Baena et al.，2020），使其能够学习到真实数据的内在特征。在此过程中，研究人员通过可行性判断来评估合成数据的质量，并将已生成的合成数据再次用于生成器的训练（Babier et al.，2024），合成数据使算法更具保障性，算法通过对数据的学习和运行，优化自身模型，从根本上来看，合成数据为算法营造一个安全区（Moon et al.，2007），从而优化生成决策模型。如 Babier 等（2024）在针对投资组合优化研究中，通过随机近似和松弛方法生成初始的可行与不可行决策，并将其作为初始数据，在 IPMAN 算法的每次迭代中，利用当前训练好的生成器为每个上下文生成决策，并用预言机标记后加入数据集，生成后的合成数据主要被用于扩充训练数据集，以改进分类器和生成器，从而在主动学习循环中不断提升模型性能，其中 IPMAN 能够快速生成高质量决策，并且对噪声具有鲁棒性。

合成数据为决策模型的有效运行及构建提供了重要支撑。一方面，合成数据可用于构建模型的控制组（Chesnes et al.，2017）或者实

验组（Shen et al.，2023），并借助合成控制法生成反事实对照组，进而评估政策（Lonati et al.，2018），通过实证方式检验决策的有效性及其差异。此外，研究人员还可利用合成数据模拟复杂系统中多变量的相互作用与动态演变，构建高度贴近现实情境的决策模型（Gourinchas & Parker，2002；Shi et al.，2023），并在此基础上开展分析与验证。另一方面，合成数据能够系统比较不同决策方法在效果与表现上的差异（Kallus & Zhou，2021），通过综合评估各方法的优劣，为研究人员在模型与方法选择上提供依据（Bijmolt & Wedel，1999；Kim et al.，2020）。如 Baena 等（2020）通过引入随机 1H2D 表格生成器获得合成数据，其中生成器主要按照 AMPL 模型语言的语法编写，并主要由行数、列数等参数所控制，通过改变参数生成不同规模和难度的合成数据，其主要被用于验证稳定化 Benders 的优越性，从而证明其所提出方法的有效性。

2. 因果推断与实验设计

因果推断作为一种研究方法，主要通过构建和分析因果模型，从数据中识别变量间的因果关系并估计因果效应，从而帮助研究人员揭示变量间的因果机制。实验设计通常作为研究活动的开端，其围绕已设定的研究目标，明确研究对象与方法，从而确保研究假说能够得到可靠验证。在以往的因果推断与实验设计中，研究主要依赖于真实数据的获取与分析，并通过对数据或模型的反复调试以实现预期目标。合成数据的引入则在认识论层面改变了这一逻辑。尽管合成数据并非源自真实观测，但其生成机制能够复现真实数据的关键统计特征。因此，基于既有的因果推断逻辑与实验设计，合成数据可为正式研究的开展提供先验推断与理论预演，为进一步开展相关研究提供先验证的基础，从而增强研究结论的稳健性与有效性。

合成数据在因果推断方法中的应用，通常作为验证模型以及算法有效性的关键手段。具体来看，合成数据可以对研究人员提出的模型（Ansari et al.，2000）以及创新性的计算方法（Dzyabura & Hauser，2011；Yang et al.，2010；Trusov，2013）进行验证，从而优化模型设计。包括开展模拟数值实验判断模型性能（Kim et al.，2013），或者结合真实数据进行实验并对比分析（Lin et al.，2015）。研究人员通过对合成数据的上述应用，可实现对模型和策略的效果、可计算性评估（Ping，1995；Ramaswamy et al.，1996），以及验证模型或方法的有效性等（Menon & Sarkar，2007；Bastani et al.，2022；Tsirtsis et al.，2024），从而增强所提模型与研究方法的认可度和可信度（Bertsimas & King，2016；Aswani et al.，2018）。如 Saghafian（2024）在保持与案例研究相似的复杂场景下，通过生成按时间顺序排列，但带有随机性的合成数据，以此来评估和比较 DAV - Learning、SAV - Learning 及其 BUC 版本在控制环境下的性能，并与案例研究的结果进行类比验证。此外，研究人员通过生成具有与真实数据相似特征的合成数据，模拟真实情况以开展研究与分析，并与真实数据形成比较，为开展真实数据研究奠定良好基础（Negahban et al.，2017）。如 Ciocan 和 Misic（2022）在研究定价问题中，基于已设定模型生成合成期权价格路径，以此建立具有不同参数的期权定价模型，从而生成相关

合成数据，并被用于在受控条件下系统性地评估和比较不同算法的性能，为开展真实数据研究提供性能基准和支撑。

合成数据的运用还体现在度量指标设计和参数敏感度分析上，Todorov（2010）利用合成数据构建了市场波动率风险溢价的度量指标，并将其用于分析市场情况，在原有度量指标的基础上拓宽了数据分析思路，提升了数据分析效率。研究人员还可通过调整模型参数，观察合成数据变化，以理解不同局部互动规则对全局动态的影响（Lomi & Larsen，1996）。之前研究中，研究人员往往依赖于实际采集的数据进行分析和模型训练，但这一过程通常耗费时间且成本高昂，尤其是在数据需求量大或数据稀缺的研究领域。相比之下，合成数据提供了一种高效且经济的解决方案，研究人员可以通过模型或算法生成大量符合特定分布和特征的数据，从而极大地加速了数据筛选收集的速度。同时，合成数据准确度相对较高，有助于降低实证分析过程中的出错概率，提升研究分析质量。

3. 市场空间细分

市场空间细分主要针对市场空间格局以及结构性问题展开研究，如国际市场划分、生命周期消费等，其多采用问卷调查、数据爬取等研究手段，并引入因果推断与鲁棒优化等分析模型。在消费者研究中，由于消费者行为数据多由商家或平台所掌握，因此，在收集数据的过程中较为困难。在已有研究中，研究人员通过采集评论区消费者评价内容进行文本分析，以达到同等的理想研究状态。然而，评论内容可能存在偏差，如用户更倾向于发表极端情感的评论，导致样本不均衡等问题（吕梦雅等，2024）。

合成数据需依据研究设计及其生成类型进行系统性整合。由于合成数据来源多样且生成模型各异，通常需结合研究需求，对合成数据进行交集、并集、分类等操作整合，筛选出结构、逻辑等合理的数据单元，在此基础上构建更为全面和更具针对性的数据集，如 Bai 等（2024）通过研究需求，将合成数据分为训练集、验证集和测试集，对机器学习进行评估。Menon 等（2022）依据研究设计对合成数据进行分级，作为后续试验基础。在此过程中，详细记录各数据集间的差异，包括数据分布、样本规模及内容特征等，从而为不同研究场景提供适配的数据支持（Alan & Browning，2010；Azizi et al.，2021）。此外，研究人员可借助合成数据建立新的生成模型、产生新的合成数据，以开展新的研究分析，尤其在隐私得到保护的条件下，能够模拟复杂多变的现实情境，进而持续优化决策过程与预测效果。Ter Hofstede 等（2002）为识别国际市场中的空间细分，通过生成两个数据集，分别对应细分的空间连续性和空间独立性情况，针对每个数据集又各自构建两个细分，除空间配置存在差异外，数据集共享相同的属性。在此基础上，通过对合成数据的分析，评估最具限制性的空间细分模型在有无空间模式条件下的性能差异。

（二）实践应用

合成数据在当前各领域的研究中，作为实验验证与辅助研究手段的重要价值日益凸显，旨在验证模型与方法的有效性并预测未来的研究趋势。具体而言，研究人员利用合成数据在模型验证、决策支持、估计与趋势推断等环节

中的辅助性作用，以帮助研究获得有效且理想的结果，同时还可进一步支持对后续研究方向的预估与探讨，从而为学者提供具有前瞻性的研究假设与长远推论。

1. 消费者选择与决策

消费者选择与决策研究聚焦于数字经济时代下，数字技术对消费者的心理、认知、决策行为等影响，如通过收集和分析消费者决策数据，研究智能算法（大数据杀熟、个性推荐等）对消费者的体验感受、购买决策等方面产生的影响（庞瑞芝和戚聿东，2025）。由于消费者相关数据获取方式复杂，因此部分研究者通过生成器合成消费者各项数据，以便开展消费者相关研究。

Derakhshan 等（2022）在研究在线平台对消费者选择的塑造时，通过使用合成数据开展数值研究，以补充研究的理论发展，即基于已有模型分析假设与测试环境，为获取估计结果，研究人员针对每个问题实例均生成了包含 100000 次查询的合成数据集。借助合成数据，研究人员比较了 w 有序排名、PTAS 算法以及两种启发式算法的市场份额表现，并将上述结果与最优解进行对比，进而量化各算法的实际性能。Dzyabura 和 Hauser（2019）在对消费者学习其偏好权重时的产品推荐研究中，通过合成数据实验进行概念验证，即检验当消费者学习他们的偏好时，所提出的推荐系统修改是否能提升其性能。

2. 动态定价

动态定价需基于市场变化、消费者特征等，设计稳健的定价方法或形成灵活的定价策略。在已有研究中，研究人员基于研究问题，通过历史真实数据生成消费者数据，对当前已有的定价模型或策略进行验证和对比，以便进一步选取适用于所提出的研究环境和研究假说的定价模型或策略，在此基础上，还可对模型进行优化和调整，使其达到研究目标。

Chen 等（2023）在基于交易数据的无模型组合定价研究中，主要对所提出的方法在合成数据集和真实数据集上进行数值研究，以评估其所提方法与经典模型方法的差异。用于评估近似算法性能中的合成数据主要基于历史数据进行随机生成；用于小样本量分析中的合成数据主要由 MNL 模型生成；用于模型误设分析中的合成数据则基于复杂的混合 Logit 模型生成。针对不同方法的有效性验证以及对比，合成数据为其构建了有效的测试环境与平台，使研究人员可在已有的真实情况下，量化不同方法的性能，为后续定价策略模型的选择以及策略生成提供数据基础。

3. 其他研究

合成数据的使用为丰富研究方法、拓展研究边界提供了新的可能。当前在工商管理领域，除被应用于消费者行为、定价策略等研究议题外，合成数据也逐步拓展至资源组合规划与策略优化等相关问题。如 Feng 等（2024）在研究中基于合成数据进行数值模拟，以评估其所提算法的性能。为构建既真实又可控的市场环境，研究人员采用特定生成模型，合成包括产品类型、消费者类型、费用与时间在内的多维数据，并在此基础上开展算法性能对比分析，通过多次模型运行与收益记录，该研究不仅验证了所提算法的有效性，还进一步比较了不同算法下的成本表现，从而为该类问题的研究提供了可借鉴的方法论基础。

（三）合成数据使用流程及评估指标

基于合成数据在工商管理领域中方法应用的分析，本文绘制了如图 9 所示的使用流程图以及所对应的相关文献，并从合成数据的可信度和可用度两方面提出具有可操作性的合成数据评估指标。

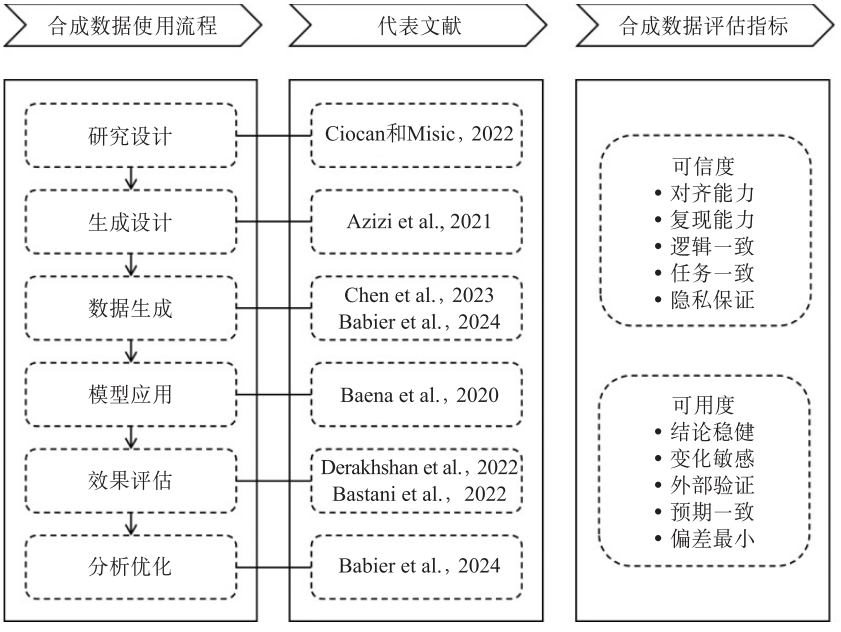

图 9　合成数据使用流程及评估指标

资料来源：作者整理绘制。

1. 使用流程

基于对合成数据在研究方法中应用情况的梳理及相关文献分析，本文归纳出合成数据的规范使用流程，包括研究设计、生成设计、数据生成、模型应用、效果评估与分析优化六个环节。具体而言：研究设计依据研究目标明确所需合成数据的特征与范围；生成设计确定合成数据的生成方法与技术路径；数据生成指通过特定模型生成符合要求的合成数据；模型应用将合成数据引入研究模型进行实证分析；效果评估侧重于利用合成数据对模型性能进行系统评价；分析优化则基于评估结果对研究模型进行迭代改进。

2. 评估指标

合成数据在方法研究中的适用条件已成为当前研究者关注的重要议题。数据可信是指数据产品在来源合规性、呈现合规性等方面的可信度（刘业政等，2024）。数据可用性是指在某一特定情景下数据使用者可以重用科学数据以开启或完成某项科研工作（黄国彬和郑霞，2019）。鉴于已有研究对数据可信度与可用度的探讨，本文从这两个维度构建合成数据的评估指标。考虑到合成数据在生成路径、自身特性及应用场景上与真实数据的差异，对其可信度与可用度的衡量主要基于与真实数据的对比分析以及在研究实践中的实际应用效果。

合成数据可信度是衡量合成数据准确性的核心维度，指其在多大程度上能够拥有真实数据的固有特征。其评估可从以下五个指标展开：对齐能力、复现能力、逻辑一致、任务一致及

隐私保证。其中，对齐能力衡量合成数据与真实数据在分布上的相似程度；复现能力考察合成数据能否再现真实数据所蕴含的结构性特征；逻辑一致关注合成数据的整体分布是否在统计意义上与真实数据相符；任务一致评估合成数据在多大程度上能够替代真实数据以完成特定的研究任务；隐私保证则指合成数据在保护原始信息不外泄的前提下，与真实数据保持统计特征的一致性。

合成数据可用度用于衡量合成数据在研究应用中的实际表现，具体指基于合成数据所获得的研究结果是否在统计意义上与真实数据的结果保持一致。该评估内容包含以下五个指标：结论稳健、变化敏感、外部验证、预期一致及偏差最小。结论稳健指基于不同版本合成数据所获研究结论的一致性程度；变化敏感衡量合成数据对模型或参数变化的响应程度；外部验证评估合成数据结论与真实数据结论的吻合度；预期一致检验基于合成数据的研究发现是否符合理论预期；偏差最小则关注由合成数据优化的模型在真实数据上的表现以及与真实数据结论的契合程度。

五、合成数据风险检视

合成数据在科研中的使用应遵循“数据向善，服务科研”原则，确保其技术研发与应用始终服务于科学研究、促进研究创新，并最大限度地防范与管控相关潜在风险。为此，需确保合成数据的质量可靠与偏差可控，建立健全合成数据使用规范与治理机制，并明确合成数据的主体责任，从而构建可信、可控、可持续的合成数据治理体系。

（一）数据质量

数据质量是决定科研工作能否顺利推进并取得有效成果的基础性保障。尽管当前合成数据在真实性上已高度接近真实数据，但在准确性与偏见控制方面仍然存在明显局限。由于真实数据的短缺导致合成数据技术在学习过程中，往往只能捕捉到全局性的数据特征，而难以充分反映数据内在的细微变化与动态关联。同时，合成数据不仅可能沿袭真实数据中既有的偏见，还可能在生成过程中引入新的偏差。因此，有必要对合成数据生成模型进行动态监测，根据其数据输出表现持续优化技术架构与训练策略，以逐步提升数据生成的准确性与完整性。

（二）数据安全

合成数据技术作为一种新兴的数据生成方法，能够有效应对当前科研数据资源短缺的挑战。然而，该技术也存在被滥用的风险，例如可能被不法分子用于训练恶意模型或生成虚假信息。合成数据通常基于真实数据构建，并保持其统计特征，这使得不法人员通过非法途径获取数据以及生成技术，用于训练诸如“GhostGPT”“WormGPT”等违规模型，进而生成高仿真欺诈邮件、钓鱼邮件等违法内容（Falade，2023）。在此过程中，不法分子借助相关技术生成适用于不同场景的违法数据，以满足其非法目的。因此，为在合法合规的框架下妥善使用并推广合成数据技术，必须建立系统的数据保障机制。首先，应从合成数据使用源头进行规范，强化数据调用与处理环节的实名认证机制，确保使用行为可溯源，数据泄露可追责；其次，

需加强对合成数据流动过程的动态监管与风险预警，通过对异常使用模式的实时监测，及时识别并干预潜在风险；最后，还应完善合成数据监督体系，持续跟踪合成数据流向和使用情况，分析研判违规使用合成数据的潜在路径，避免数据泄露。

（三）数据主体识别

合成数据重新定义了对数据供给、流通和使用全过程的理解。其主要由模型或算法生成的自标注信息，改变了数据的供给源头和预处理环节，涉及数据提供者、处理者及合成数据生成者等多方主体。在数据日益资产化的背景下，如何清晰界定合成数据的权利归属，已成为法律与治理领域亟待深入研究的议题。为此，可考虑依据合成数据的生成机理，对各参与主体包括真实数据提供者、合成数据生成者以及后续使用者（如模型训练与开发者）的贡献进行灵活的比例量化。具体可综合考量数据量大小、采集难易程度、使用场景、生成技术复杂度等多个维度，以此构建贡献度评价体系，并作为权属比例划分的依据。在此基础上，可结合现行相关法律法规，对合成数据的权利归属进行确认，从而在制度层面保障权属分配的合理性与可操作性。

六、结论与未来研究展望

（一）主要结论

随着生成式人工智能的快速发展，合成数据的概念内涵等得到不断丰富，合成数据也被赋予新的时代特征与使命。在此背景下，合成数据在工商管理领域中的应用成为学科研究热点议题。本文通过对合成数据相关英文文献的系统分析，总结了合成数据在工商管理研究中的应用与发展。具体而言，本文得出以下结论：（1）合成数据应用场景逐步多元。合成数据随着技术的不断优化，其研究与应用场景由单一学科向市场营销、技术经济与管理等多学科研究领域与应用场景拓展，同时其内涵在场景应用和技术发展过程中逐渐丰富；（2）当前合成数据的研究主题系统呈现了从方法到实践应用的逻辑关系。在研究方法阶段，阐明了合成数据与工商管理领域中研究方法的融合和影响，为后续研究应用提供基础；在实践应用阶段，展现了合成数据在具体研究情境中的实际效用，体现其方法论价值；（3）合成数据丰富并拓展了因果推断理论。一方面，合成数据提供了可操控的测试环境，可被用于工具变量法、双重差分法等多种识别策略的模拟验证，通过比较估计与预设真实因果效应之间的偏差，评估方法的有效性与稳健性；同时，合成数据也可在混杂因素环境中被用于辅助研究设计，检验不同控制策略的适用条件与效果，深化相关研究方法的验证基础。另一方面，合成数据不仅支持研究数据的填补、可复现、无隐私风险等，还在复杂模型验证、因果推断检验、鲁棒优化等方法优化中具有重要价值。合成数据可适用于理论模型构建、传统统计方法验证以及对管理实践的模拟等，为方法的可行性以及研究结果的实践性提供了有力支撑。

（二）未来研究展望

基于对现有文献研究主题的梳理与分析，本文从合成数据治理、学科领域应用和研究方法创新三个方面讨论未来工商管理学者运用合

成数据，植根科学问题，挖掘应用场景，变革研究范式与创新方法，深化合成数据在工商管理领域的应用深度与广度。

1. 合成数据风险识别与治理

（1）合成数据风险识别。在数字平台经济中，数据是核心资产，合成数据的应用可以为平台企业提供更加丰富、多样的数据资源，同时降低数据获取和使用成本，提高平台运营效率。但合成数据在其生成过程中可能存在放大真实数据中所隐含的偏见，以及数据投毒等问题，同时由于其“生成—处理—使用”环节的高度融合，导致合成数据的不同利益主体间责任边界难以清晰划分（霍敬裕，2026），也使得合成数据如何实现“隐私—效用”的动态平衡成为一项难题（汪焱梁，2026）。因而如何识别并规避合成数据在其生成过程中的风险，仍是当前企业和研究人员在合成数据研究与应用中面临的重要课题。

（2）合成数据治理与规制。合成数据作为真实数据的一种替代性方案，其关注度愈发高涨。由于合成数据主要通过特定模型、算法等技术生成，因而其与真实数据在生成逻辑上具有根本性区别，这也使得合成数据的治理思路不同于以往真实数据。有研究人员提出合成数据治理应由对数据来源合法性的关注，转向对合成数据自身财产利益的保护，以及确保社会各主体能够公平地获取合成数据的透明性（金耀，2026）。同时也有研究人员以制度供给、技术加持与协同共治为切入点，构建面向合成数据的软硬法体系，借助多种治理工具实现灵活治理，以及汇聚各方力量形成治理合力，从而最终建立起敏捷性治理模式（叶英杰和李川，2025）。然而当前有关合成数据治理与规制问题的研究多聚焦于法学视角，鲜有研究与实际的数据应用场景相结合，未来研究可从管理视角出发，基于合成数据的实际应用场景，进一步探讨合成数据的治理机制与策略。

2. 丰富合成数据学科领域应用场景

未来研究可对学科领域进行更深层次探讨。如在已有的工商管理研究成果与应用基础上，系统推进合成数据与管理学以及其他研究方向的交叉融合，深入挖掘合成数据驱动范式在多元管理情境下的应用场景与价值机制，从而推动合成数据赋能工商管理理论创新。

（1）基于合成数据的组织变革思考。根据 Epoch AI 的预测，训练人工智能模型所需要的数据将可能在 2028 年前后消失殆尽，尽管斯坦福大学在其最新发布的《2025 年人工智能指数报告》中表明，当前数据的实际使用时长可能超出预期，但训练数据面临枯竭已是一个既定的事实（汪焱梁，2026）。倘若合成数据技术衰退，那将可能会阻碍数据供给、流通和使用的链路，降低组织的资源优势，增大组织对外部资源的依赖。因此，未来研究可针对如何增强组织的自主创新能力，以及如何通过资源多样化策略降低对单一数据技术的依赖展开进一步讨论。

（2）拓展多学科领域研究应用场景。数据是开展科学研究的基础，尽管目前我国部分学科研究有所突破，但发展相较国外仍然缓慢，而合成数据由于其特有的生成方式，可通过学习不同类型数据的统计特征，能够实现数据多元化，缓解部分研究领域的数据不足、数据隐私保护等问题，例如，杜津宇和王洪亮（2023）

提出税收征管人工智能所涉主体范围最为广泛，且直接关系公民与法人的根本权益，对数据中性的要求更为严格，所以有必要将合成数据引入其中。同时，也有研究人员通过提出一种基于大语言模型构建合成数据的方法，以解决医疗领域虚假评论识别中数据稀缺问题（王嘉杰等，2025）。进而推动了合成数据与平台治理、医疗服务等领域的交叉研究，打破了数据学科壁垒，达到不同研究领域间的数据知识迁移，以及多学科间的效度验证，突破了学科研究间的理论与情景约束，从而找到数据所隐含的研究普适性，加强了研究领域的关联性。未来研究可针对不同学科的研究特点、实验情景等，利用合成数据对原有研究模型的设定、变量以及分析方法等进行优化与整合，从而拓展合成数据的多场景应用。

3. 基于现有研究方法的合成数据应用

为拓展合成数据在工商管理领域研究方法中的应用，未来研究可关注其在量表开发与效度评估、数据填补等现实应用场景中的实践。因此，需在现有研究方法基础上，借助合成数据，补充新的研究范式。

（1）量表开发与效度评估。在现有量表开发中，研究人员通常需邀请相关领域的专家学者、学生或企业参与前测，以审查题项的准确性与可读性（张洪等，2026；朱明洋等，2026）。合成数据的使用能够促进量表开发的合理性，即研究人员可借助大语言模型生成虚拟被访者（Trojacka et al.，2026），通过虚拟被访者对量表的回答，生成符合问卷要求的合成数据集，并保证不同维度下的数据集具有区分效度，进而根据所获取的数据结果对量表进行调整和修改。在未来量表开发中可基于已有量表开发模式引入合成数据的使用，节省量表开发成本，提升量表开发质量。在效度评估方面，研究的开展是基于对现实问题的思考引申至研究问题，并结合当前研究情况提出研究设想包括研究模型，通过已获取数据进而验证研究设想。但在实际研究过程中，往往研究结果与研究设想产生偏差，致使研究人员需对模型进行反复调整，以达到研究预期。合成数据则可对研究设想进行预先效度评价，如前文所述，可基于既有模型假设与测试环境，生成大规模合成数据集，用以系统对比不同算法并量化其性能表现（Derakhshan et al.，2022），即验证模型或方法的有效性等（Tsirtsis et al.，2024）。同时，通过对不同模型的评估，发掘其优势与不足，为研究人员在后续研究中的模型选择提供依据（Kim et al.，2020）。未来研究可在已有效度评估基础上进一步深化合成数据对于正式研究的先导验证作用，优化研究设计，完善研究逻辑。

（2）数据填补。数据完整与否直接影响研究结果的信度与效度，而数据缺失问题广泛存在于当前各项研究领域中。尽管研究人员试图通过剔除不完整样本的方式规避缺失样本对结果准确性的干扰，但这种方式仅在缺失比例较低时有效。基于真实数据所生成的合成数据，能够体现出真实数据的统计性特征，因而，其可被用于填补缺失数据。当前数据缺失类型主要有完全随机缺失、随机缺失和完全非随机缺失三类（Mahin et al.，2019），针对不同的数据缺失类型，研究人员通过对比不同数据填补方法，发现K近邻（KNN）、加权K近邻（WKNN）、

缺失森林（MF）和缺失森林与加权 K 近邻混合迭代（MFWKNN）四种数据填补方法优于传统数据填补方法，其中缺失森林与加权 K 近邻混合迭代（MFWKNN）算法在不同缺失率、不同评价准则下均能得到最优的插补结果和相对较优的稳定性，并且伴随缺失率的增大，这种优势愈发明显（郑智泉等，2023）。生成后合成数据仍需进行质量验证与评估，由研究人员对其进行信度与效度检验，重点核查数据是否已充分脱敏、是否存在潜在信息偏差等，以确保最终生成的合成数据能有效模拟真实数据的统计分布特征。因此，未来研究可利用已有的合成数据技术对缺失数据进行补充，提升研究数据完整性，降低研究误差。

接受编辑：张光磊
收稿时间：2025 年 9 月 15 日
接收时间：2026 年 7 月 20 日

作者简介

彭正银，天津财经大学商学院企业管理系教授，博士生导师，主要研究方向为网络组织与治理、数字平台治理。国家社科重大项目首席专家，国务院政府特殊津贴专家，教育部新世纪优秀人才，国家一流专业工商管理、国家一流课程《公司治理》负责人，天津市有突出贡献专家，天津市教学名师，天津市“网络组织与治理”创新团队带头人，天津市高校社科平台企业治理实验室负责人。现主持国家社科重大项目“平台企业治理研究”，主持完成国家自科基金项目 4 项、省部级项目 5 项，出版专著 5 部，主编参编教材 4 部，在《管理科学学报》《中国工业经济》《中国软科学》等学术期刊发表论文 60 多篇，获得天津市社会科学优秀成果一、二等奖 5 项。

赵琰钰，天津财经大学商学院博士研究生，研究领域为平台治理、数据治理等。

李佳雷（通讯作者，E - mail：lijialeioliver@163.com），天津财经大学企业管理专业博士后，大连交通大学经济管理学院硕士生导师。研究领域为智能原生组织创新，在《管理学季刊》《经济管理》等国内外刊物发表学术论文 10 余篇，3 篇论文被人大复印资料全文转载，出版专著《共享经济平台：组织模式、形成机制与发展理念》，主持教育部人文社会科学研究青年基金项目、省级和副省级项目 10 项。

王鑫阳，天津财经大学商学院硕士研究生，研究领域为网络组织与平台治理。

参考文献

［1］曹树金、曹茹烨：《基于研究主题和引文分析的信息资源管理学科发展探究》，《信息资源管理学报》，2023 年第 2 期。

［2］端利涛、李海舰：《数据供给、流通与应用逻辑——基于数据特性的视角》，《经济管理》，2025 年第 6 期。

［3］杜津宇、王洪亮：《税收征管数字化转型中的合成数据应用》，《税务研究》，2023 年第 7 期。

［4］郭亚楠、马瑞强、崔旭、杨皓然：《数据合成技术研究：从统计模型到深度学习模型系统综述》，《计算机工程与应用》，2026 年第 4 期。

［5］黄国彬、郑霞：《数据论文的内容规范性研究》，《图书情报工作》，2019 年第 22 期。

［6］霍敬裕：《合成数据的“合成”风险与法律规制》，《东方法学》，2026 年第 2 期。

［7］金耀：《人工智能模型训练的合成数据治理路径》，《法学论坛》，2026 年第 2 期。

［8］刘业政、宗兰芳、周芦娟、潘岳、袁昆：《数据要素流通交易的可信指标及测度方法》，《管理评论》，2024 年第 12 期。

［9］吕梦雅、王晓龙、李凯旋、孙梦梦、周莉莎、郭栋梁、赵静：《负面评论引导的高维多元数据可视分析系统》，《计算机辅助设计与图形学学报》，2024 年第 6 期。

［10］《麻省理工科技评论》：《全球十大突破性技术（2017—2023 年）》，《麻省理工科技评论》中文网，2017—2023 年，https：//www. mittrchina. com/search。

［11］孟祝、宋雪、张国伟、朱宁：《鲁棒优化在城市交通系统运营与管理中的应用》，《工程管理科技前沿》，2025 年第 6 期。

［12］聂亚青、吴庭璋、王若佳、翟兴：《基于 BERTopic 模型的健康信息学主题挖掘与发展演化研究》，《情报科学》，2024 年第 4 期。

［13］庞瑞芝、戚聿东：《数字经济时代中国产业经济学的自主学术体系建设》，《学术界》，2025 年第 2 期。

［14］佟泽华、许海云、孙杰、王雅洁、陈约冕：《科研大数据研究：基础框架体系及其关联逻辑》，《情报资料工作》，2025 年第 4 期。

［15］汪焱梁：《人工智能合成数据“隐私—效用”难题的动态平衡路径》，《大连理工大学学报（社会科学版）》，2026 年，网络首发。

［16］王嘉杰、李天逸、丁天、康乐乐：《基于大语言模型合成数据的在线医疗社区虚假评论识别研究》，《情报理论与实践》，2025 年第 10 期。

［17］叶英杰、李川：《人工智能模型训练中合成数据的应用风险及其治理路径》，《情报理论与实践》，2025 年第 6 期。

［18］张涛：《人工智能训练中合成数据的融贯性法律治理》，《计算机科学》，2025 年第 2 期。

［19］张洪、万兴瑶、邓朝华、刘海兵：《面向全场景智能服务的顾客价值共创行为——维度探索、量表开发与影响分析》，《管理世界》，2026 年第 2 期。

［20］郑智泉、陈妍、王孟孟、田维琦：《不同缺失率下的数据填补算法稳定性研究》，《统计与决策》，2023 年第 8 期。

［21］朱明洋、吴晓波、张玉利、张永强：《平台网络自主权研究：量表开发及对企业创新影响验证》，《管理世界》，2026 年第 1 期。

［22］Achuthan，S.，Chatterjee，R.，Kotnala，S.，Mohanty，A.，Bhattacharya，S.，Salgia，R.，& Kulkarni，P. 2022. Leveraging deep learning algorithms for synthetic data generation to design and analyze biological networks. *Journal of Biosciences*，47（3）.

［23］Alan，S.，& Browning，M. 2010. Estimating Intertemporal Allocation Parameters using Synthetic Residual Estimation. *Review of Economic Studies*，77（4）：1231 – 1261.

［24］Ansari，A.，Jedidi，K.，& Jagpal，S. 2000. A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science*，19（4）：328 – 347.

［25］Aswani，A.，Shen，Z.，& Siddiq，A. 2018. Inverse Optimization with Noisy Data. *Operations Research*，66（3）：870 – 892.

［26］Azizi，S.，Bozkir，C. D. C.，Trapp，A. C.，Kundakcioglu，O. E.，& Kurbanzade，A. K. 2021. Aid Allocation for Camp – Based and Urban Refugees with Uncertain Demand and Replenishments. *Production and Operations Management*，30（12）：4455 – 4474.

［27］Babier，A.，Chan，T. C. Y.，Diamant，A.，& Mahmood，R. 2024. Learning to Optimize Contextually Constrained Problems for Real – Time Decision Generation. *Management Science*，71（2）.

［28］Baena，D.，Castro，J.，& Frangioni，A. 2020.

Stabilized Benders Methods for Large – Scale Combinatorial Optimization, with Application to Data Privacy. *Management Science*, 66 (7): 3051 – 3068.

[29] Bai, Y., Feldman, J., Topaloglu, H., & Wagner, L. 2024. Assortment optimization under the multinomial logit model with utility – based rank cutoffs. *Operations Research*, 72 (4): 1453 – 1474.

[30] Bastani, H., Simchi – Levi, D., & Zhu, R. 2022. Meta Dynamic Pricing: Transfer Learning Across Experiments. *Management Science*, 68 (3): 1865 – 1881.

[31] Bellovin, S. M., Dutta, P. K., & Reitinger, N. 2019. Privacy and Synthetic Datasets. *Stanford Technology Law Review*, 22 (1): 1 – 52.

[32] Bertsimas, D., & King, A. 2016. OR Forum – An Algorithmic Approach to Linear Regression. *Operations Research*, 64 (1): 2 – 16.

[33] Bijmolt, T. H. A., & Wedel, M. 1999. A comparison of multidimensional scaling methods for perceptual mapping. *Journal of Marketing Research*, 36 (2): 277 – 285.

[34] Chen, N. Y., Cire, A. A., Hu, M., & Lagzi, S. 2023. Model – Free Assortment Pricing with Transaction Data. *Management Science*, 69 (10): 5830 – 5847.

[35] Chesnes, M., Dai, W., & Jin, G. Z. 2017. Banning foreign pharmacies from sponsored search: The online consumer response. *Marketing Science*, 36 (6): 879 – 907.

[36] Ciocan, D. F., & Misic, V. V. 2022. Interpretable Optimal Stopping. *Management Science*, 68 (3): 1616 – 1638.

[37] Derakhshan, M., Golrezaei, N., Manshadi, V., & Mirrokni, V. 2022. Product Ranking on Online Platforms. *Management Science*, 68 (6): 4024 – 4041.

[38] Dzyabura, D., & Hauser, J. R. 2019. Recommending products when consumers learn their preference weights. *Marketing Science*, 38 (3): 417 – 441.

[39] Dzyabura, D., & Hauser, J. R. 2011. Active machine learning for consideration heuristics. *Marketing Science*, 30 (5): 801 – 819.

[40] Egger R., & Yu J. 2022. A topic modeling comparison between LDA, NMF, Top2vec, and BERTopic to demystify Twitter posts. *Frontiers In Sociology*, 7.

[41] El Emam, K., Mosquera, L., & Hoptroff, R. 2020. Practical synthetic data generation: Balancing privacy and the broad availability of data. Sebastopol, CA: O'Reilly Media.

[42] Falade, P. V. 2023. Decoding the threat landscape: ChatGPT, FraudGPT, and WormGPT in social engineering attacks. arXiv preprint arXiv: 2310.05595.

[43] Feng, Y. D., Niazadeh, R., & Saberi, A. 2024. Technical Note – Near – Optimal Bayesian Online Assortment of Reusable Resources. *Operations Research*, 72 (5).

[44] Goodfellow, I. J., Pouget – Abadie, J, Mirza, M, Xu, B, Warde – Farley, D, Ozair, S, Courville, A, & Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27: 2672 – 2680.

[45] Gourinchas, P. O., & Parker, J. A. 2002. Consumption over the life cycle. *Econometrica*, 70 (1): 47 – 89.

[46] Grootendorst M. 2022. BERTopic: Neural Topic Modeling with a Class – based TF – IDF Procedure. arxiv preprint arxiv: 2203.05794.

[47] Hauser, J. R., Liberali, G., & Urban, G. L. 2014. Website Morphing 2.0: Switching Costs, Partial Exposure, Random Exit, and When to Morph. *Management Science*, 60 (6): 1594 – 1616.

[48] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. 2022. Synthetic Data—what, why and how? arXiv preprint arXiv: 2205.03257.

[49] Kallus, N., & Zhou, A. 2021. Minimax – Optimal Policy Learning Under Unobserved Confounding. *Management Science*, 67 (5): 2870 – 2890.

[50] Keller, S. A., Shipp, S., & Schroeder, A. 2016. Does big data change the privacy landscape? A review of the issues. *Annual Review of Statistics and Its Application*, 3: 161 – 180.

[51] Kim, S., Blanchard, S. J., DeSarbo, W. S., & Fong, D. K. H. 2013. Implementing managerial constraints in model – based segmentation: Extensions of Kim, Fong, and DeSarbo (2012) with an application to heterogeneous perceptions of service quality. *Journal of Marketing Research*, 50 (5): 664 – 673.

[52] Kim, S., Lee, C., & Gupta, S. 2020. Bayesian synthetic control methods. *Journal of Marketing Research*, 57 (5): 831 – 852.

[53] Kingma, D. P. & Welling, M. 2013. Auto – encoding variational bayes. arXiv preprint arXiv: 1312. 6114.

[54] Lin, S., Zhang, J., & Hauser, J. R. 2015. Learning from experience, simply. *Marketing Science*, 34 (1): 1 – 19.

[55] Lomi, A., & Larsen, E. R. 1996. Interacting locally and evolving globally: A computational approach to the dynamics of organizational populations. *Academy of Management Journal*, 39 (5), 1287 – 1321.

[56] Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. 2018. On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, 64: 19 – 40.

[57] Mahin, V., Mohammad, H. M., & Mehrdad, J. 2019. A Hybrid Bayesian Network and Tensor Factorization Approach for Missing Value Imputation to Improve Breast Cancer Recurrence Prediction. *Journal of King Saud University – Computer and Information Sciences*, 31 (2): 175 – 184.

[58] Marano, G. C., Rosso, M. M., Aloisio, A., & Cirrincione, G. 2024. Generative adversarial networks review in earthquake – related engineering fields. *Bulletin of Earthquake Engineering*, 22 (7): 3511 – 3562.

[59] Menon, S., & Sarkar, S. 2007. Minimizing information loss and preserving privacy. *Management Science*, 53 (1): 101 – 116.

[60] Menon, S., Ghoshal, A., & Sarkar, S. 2022. Modifying Transactional Databases to Hide Sensitive Association Rules. *Information Systems Research*. 33 (1): 152 – 178

[61] Moon, S., Kamakura, W. A., & Ledolter, J. 2007. Estimating promotion response when competitive promotions are unobservable. *Journal of Marketing Research*, 44 (3): 503 – 515.

[62] Negahban, S., Oh, S., & Shah, D. 2017. Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research*, 65 (1): 266 – 287.

[63] Ping, R. A. 1995. A parsimonious estimating technique for interaction and quadratic latent variables. *Journal of Marketing Research*, 32 (3): 336 – 347.

[64] Ramaswamy, V., Chatterjee, R., & Cohen, S. H. 1996. Joint segmentation on distinct interdependent bases with categorical data. *Journal of Marketing Research*, 33 (3): 337 – 350.

[65] Rubin, D. B. 1993. Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9: 461 – 468.

[66] Saghafian, S. 2024. Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach. *Management Science*, 70 (9): 5667 – 5690.

[67] Shen, D., Ding, P., Sekhon, J., & Yu, B. 2023. Same Root Different Leaves: Time Series and Cross – Sectional Methods in Panel Data. *Econometrica*, 91 (6): 2125 – 2154.

[68] Shi, P., Helm, J. E., Chen, C., Lim, J., Parker, R. P., Tinsley, T., & Cecil, J. 2023. Operations (management) warp speed: Rapid deployment of hospital-focused predictive/prescriptive analytics for the COVID-19 pandemic. *Production and Operations Management*, 32 (5): 1433-1452.

[69] Ter Hofstede, F., Wedel, M., & Steenkamp, J. B. E. M. 2002. Identifying spatial segments in international markets. *Marketing Science*, 21 (2): 160-177.

[70] Todorov, V. 2010. Variance risk-premium dynamics: The role of jumps. *The Review of Financial Studies*, 23 (1): 345-383.

[71] Trojacka, E., Przybek-Skrzypecka, J., Izdebska, J., Szaflik, J. P., Qazi, M. A., Azhar, A., & Skrzypecki, J. 2026. Using Large Language Models for In Silico Development and Simulation of a Patient-Reported Outcome Questionnaire for Cataract Surgery with Various Intraocular Lenses: A Pre-Validation Study. *Journal of Clinical Medicine*, 15 (1): 283.

[72] Trusov, M., Rand, W., & Joshi, Y. V. 2013. Improving prelaunch diffusion forecasts: Using synthetic networks as simulated priors. *Journal of Marketing Research*, 50 (6): 675-690.

[73] Tsirtsis, S., Tabibian, B., Khajehnejad, M., Singla, A., Schölkopf, B., & Gomez-Rodriguez, M. 2024. Optimal Decision Making Under Strategic Behavior. *Management Science*, 70 (12): 1-20.

[74] Yang, S., Zhao, Y., & Dhar, R. 2010. Modeling the underreporting bias in panel survey data. *Marketing Science*, 29 (3): 525-539.