数学建模在社会调查数据分析中的应用

李林佳

山东建筑大学理学院，山东济南，250102；

摘要：随着大数据时代的到来，社会调查数据呈现出体量大、维度高、结构复杂等特点，传统统计方法在处理此类数据时面临效率低、解释力弱等挑战。数学建模凭借其严谨的逻辑框架与强大的抽象能力，为社会调查数据分析提供了新范式。本文系统探讨数学建模在问卷设计优化、样本代表性评估、变量关系挖掘及预测推演等环节中的具体应用，提出“基于机制驱动与数据驱动融合的混合建模”新观点，强调将社会学理论嵌入数学模型以提升解释力与政策指导价值。通过典型案例分析，验证了该方法在人口流动、公共舆情、教育公平等议题中的有效性。研究表明，数学建模不仅能提升社会调查数据的分析精度，更能揭示隐藏的社会运行机制，为科学决策提供坚实支撑。

关键词：数学建模；社会调查；混合建模；机制驱动

DOI：10.69979/3029-2735.26.03.103

引言

社会调查是理解社会现象、制定公共政策的重要工具。然而，面对日益复杂的社会系统和海量异构数据，传统描述性统计与简单回归分析已难以满足深度洞察的需求。数学建模作为连接现实问题与定量分析的桥梁，近年来在社会科学领域展现出巨大潜力。不同于纯数据驱动的机器学习方法，数学建模强调对社会过程内在逻辑的刻画，能够整合先验知识与实证数据，实现“可解释性”与“预测性”的统一。当前，如何将社会学理论有效融入数学模型，避免“黑箱化”倾向，成为学界关注焦点。本文旨在系统梳理数学建模在社会调查数据分析中的应用场景，提出机制驱动与数据驱动相结合的混合建模范式，并通过实证案例论证其科学性与实用性，为提升社会科学研究的量化水平提供新思路。

1 数学建模在社会调查中的基础作用

社会调查数据具有鲜明的非实验性特征，通常源于自然状态下的观察与记录，难以控制混杂变量，且普遍存在高噪声、缺失值和测量误差等问题。同时，这类数据常呈现多层级结构，如个体嵌套于家庭、社区或区域之中，变量间关系错综复杂，动态演化明显。传统统计方法，如线性回归或卡方检验，往往基于强假设前提，例如变量间线性关联、误差独立同分布等，在面对非线性交互、时间依赖或空间异质性时表现乏力，难以捕捉社会现象背后的深层机制。相比之下，数学建模通过构建形式化的抽象框架，能够清晰表达变量间的逻辑结构与作用路径，不仅支持对现有数据的拟合分析，还可进行情景模拟与反事实推演，为政策评估提供“如果—那么”式的量化依据。在具体工具层面，微分方程适用于刻画连续动态过程，如舆情传播或人口迁移的速率变化；马尔可夫链擅长描述状态转移概率，可用于分析职业流动或健康状态演变；网络模型则能揭示个体间关系结构对集体行为的影响，如信息扩散或社会支持的形成；贝叶斯层次模型则有效处理多层嵌套数据，整合先验知识与样本信息，在小样本或不均衡数据下仍保持稳健估计。这些方法共同构成了面向复杂社会系统的分析工具箱，显著拓展了社会调查数据的解释深度与应用广度。

2 机制驱动与数据驱动融合：一种新范式

2.1 纯数据驱动模型的解释性困境

当前社会科学中广泛应用的机器学习与深度学习方法，虽在预测精度上表现突出，却普遍面临“黑箱”问题。这类纯数据驱动模型依赖大量观测数据自动挖掘变量关联，缺乏对社会过程内在逻辑的显性表达，难以回答“为何如此”或“机制何在”等根本性问题。社会现象本质上由制度、文化、认知与互动等多重机制共同塑造，若仅关注统计相关而忽视因果结构，极易导致模型泛化能力弱、政策建议空洞甚至误导。例如，在分析公众对某项政策的支持度时，仅用随机森林识别关键变量，无法揭示态度形成的动态路径或群体间差异的根源。

这种解释性缺失不仅削弱了研究的理论价值，也限制了其在公共治理中的实际应用。社会科学研究的核心使命在于理解机制、指导实践，而非止步于模式识别，因此亟需一种能融合理论洞察与数据实证的新建模范式。

2.2 机制驱动与数据驱动融合的建模框架

为突破上述局限，提出“机制驱动+数据驱动”的混合建模框架。该框架以社会学、经济学或心理学等领域的成熟理论为基础，先构建反映社会运行逻辑的结构化模型，再利用调查数据对模型参数进行校准、优化与验证。机制驱动部分确保模型具备清晰的因果链条与可解释的变量关系，如将 SIR（易感-感染-恢复）传染病模型迁移至舆情传播研究，将“感染”类比为观点采纳，“恢复”视为态度固化，从而刻画信息扩散的动力学过程。数据驱动部分则通过最大似然估计、贝叶斯推断或机器学习辅助拟合，使模型更贴合现实数据分布。这种融合并非简单叠加，而是形成“理论引导建模—数据反馈修正—机制再阐释”的闭环。模型既保留了理论的结构性优势，又吸纳了数据的经验支撑，有效避免了过度简化或过度拟合的两极风险，为复杂社会问题提供兼具科学性与实用性的分析工具。

2.3 青年就业意愿变迁的 Logistic 反馈模型

以青年群体就业意愿随时间变化为例，传统横截面回归难以捕捉其非线性演化特征。采用带反馈机制的 Logistic 增长模型，将就业意愿视为受社会环境、政策信号与同辈效应共同影响的动态变量。模型设定基础增长速率反映宏观就业形势，引入负反馈项模拟“意愿饱和”效应——当多数人已形成明确意向后，新增信息对整体意愿的边际影响递减；同时加入正向激励项，代表政府补贴或职业培训等干预措施的放大作用。利用多年份全国青年追踪调查数据，通过非线性最小二乘法估计参数，发现模型拟合优度显著优于线性趋势模型。更重要的是，反事实模拟显示，若提前半年实施技能培训政策，青年稳定就业意愿峰值可提升 12%，且持续时间延长。该案例表明，融合机制假设与实证数据的模型不仅能准确描述趋势，还能量化政策时序与强度的影响，为精准施策提供依据。

2.4 社区信任演化的多主体仿真建模

社区信任作为社会资本的核心维度，其形成依赖于个体互动、声誉积累与制度保障。基于格兰诺维特的“嵌入性”理论，构建多主体模型（Agent-Based Model, ABM），将居民设为具有记忆、偏好与学习能力的智能体，其信任决策依据过往合作经验、邻居行为及社区规范。初始网络结构由真实社区调查数据生成，信任更新规则嵌入社会网络理论中的“强弱连接”效应。模型运行后，可观察到信任水平随互动频率与公平感知呈非线性增长，并在遭遇外部冲击（如诈骗事件）后呈现差异化恢复能力。通过调整“社区调解机制”参数，发现引入第三方协调可使信任重建速度提升 40%。该仿真不仅复现了调查中观察到的信任分布格局，还揭示了微观互动如何涌现出宏观秩序。此类机制导向的建模方式，使抽象理论具象化，为基层治理创新提供了可测试、可迭代的数字实验平台。

3 关键应用场景与实证分析

3.1 问卷设计优化中的信息熵建模

社会调查的起点是科学的问卷设计，而冗余或低效题项不仅增加受访者负担，还可能引入测量误差。信息熵模型为题项筛选提供了量化依据。该方法将每个问题的回答分布视为信息源，计算其香农熵值——熵越高，说明选项分布越均匀，所携带的信息量越大；反之则趋于无效或引导性过强。在某省民生满意度调查预测试中，对初稿 58 个题项进行熵值评估，剔除熵值低于 0.3 的 12 个题项（如“您是否支持政府工作？”这类高度趋同问题），保留高熵题项构建最终问卷。后续主调查数据显示，精简后问卷的信度（Cronbach’s α）从 0.76 提升至 0.83，因子结构更清晰。这一过程表明，数学建模可将主观经验转化为客观标准，使问卷设计从“经验导向”转向“信息效率导向”，显著提升数据质量与分析效力。

3.2 非随机样本的偏差校正建模

现实中的社会调查常因拒访、覆盖不全或便利抽样导致样本非随机，进而产生选择性偏差。逆概率加权（IPW）结合响应面模型为此提供了解决路径。该方法先基于可观测协变量（如年龄、教育、区域）拟合受访者参与调查的概率模型（通常采用 Logistic 回归或广义加性模型），再以该概率的倒数作为权重，对样本进行重新加权。在一项农民工城市融入研究中，原始样本中高学历群体占比偏低。通过构建响应面模型估计各子群体的响应概率，并实施 IPW 调整后，关键变量（如社保

参保率、子女入学率）的估计值与官方统计数据的偏差从 18%降至 5%以内。此策略有效还原了目标总体的分布特征，使推断更具代表性。建模在此不仅是技术修正，更是对调查局限性的主动回应，保障了研究结论的外部效度。

3.3 潜变量识别的结构方程建模

许多核心社会概念如“社会资本”“主观幸福感”或“制度信任”无法直接观测，需通过多个指标间接测量。结构方程模型（SEM）通过整合测量模型与结构模型，实现对潜变量的识别与关系检验。在一项关于社区治理效能的研究中，研究者以“邻里互动频率”“互助行为”“组织参与度”等 6 个观测变量构建社会资本的测量模型，再将其作为自变量纳入对“居民公共事务参与意愿”的结构路径分析。模型拟合指标（CFI=0.94, RMSEA=0.05）良好，结果显示社会资本每提升 1 个标准差，参与意愿上升 0.38 个单位，且该效应在控制人口学变量后依然显著。SEM 不仅验证了理论假设，还量化了不可见构念的影响强度，使抽象概念获得可操作化的实证支撑。这种建模方式强化了社会科学研究的概念严谨性与因果推断能力。

3.4 教育公平政策的系统动力学仿真

面向长期性、系统性社会议题，静态模型难以捕捉反馈循环与延迟效应。以教育公平为例，构建系统动力学（System Dynamics）模型，将财政投入、师资流动、家庭期望、升学率等要素纳入统一框架，设定存量（如优质教师数量）与流量（如教师流失率）关系。模型利用十年省级面板数据校准参数，模拟三种政策情景：增加硬件投入、提高乡村教师津贴、实施城乡教师轮岗。仿真结果显示，单纯增加硬件投入在五年内对升学差距影响微弱；而教师轮岗政策虽初期成本高，但十年后可使城乡高中升学率差距缩小 22%。该模型揭示了教育公平改善的关键杠杆点在于“人”而非“物”。此类动态仿真使决策者能在虚拟环境中预演政策后果，避免资源错配。建模的价值在此体现为前瞻性治理能力的生成，而非仅对现状的描述。

4 挑战与应对策略

数学建模在社会调查中的应用面临多重现实约束。调查数据常受回忆偏差、社会期许效应或测量误差影响，导致模型输入存在系统性偏误，进而削弱结论可靠性。即使采用先进算法，若原始数据失真，模型输出亦难逃“垃圾进、垃圾出”的困境。模型本身亦需在复杂性与可解释性之间寻求平衡——过度简化可能忽略关键机制，而高度复杂的模型又易沦为难以理解的黑箱，丧失社会科学所需的理论对话能力。应对这些挑战，依赖单一学科视角远远不够，必须推动社会学家、统计学家与计算科学家深度协作：前者提供机制假设与问题定义，后两者负责模型构建与算法实现，形成知识互补的建模共同体。在此基础上，倡导“透明建模”成为必要规范，即完整公开模型的基本假设、变量定义、代码实现及敏感性分析结果，允许同行复现与检验。唯有如此，数学建模才能在保持科学严谨的同时，真正服务于对社会现实的深刻理解与有效干预。

5 结束语

数学建模正深刻改变社会调查数据分析的范式。本文提出的“机制驱动与数据驱动融合”路径，不仅回应了当前社会科学量化研究中“重算法轻理论”的倾向，更通过将社会运行逻辑内嵌于数学结构，实现了从“相关性发现”向“因果机制揭示”的跃迁。实践表明，在人口迁移、公共健康、社会治理等领域，此类混合模型能有效支撑前瞻性政策设计。然而，建模绝非万能钥匙——其有效性高度依赖于对社会情境的深刻理解与对数据局限的清醒认知。未来研究应进一步推动建模工具的普及化与标准化，同时强化伦理审查，防止模型滥用导致的决策偏差。唯有坚持“问题导向、理论支撑、数据验证、政策落地”的闭环逻辑，数学建模才能真正成为连接社会现实与科学治理的坚实桥梁，为构建更加公平、高效、韧性的社会系统贡献智慧力量。

参考文献

[1]赵庆利,车参仪,陈怀广.数学建模驱动大学生创新能力培养的实践与调查[J].教育教学论坛, 2024(47):177-180.

[2]王文发,武忠远,许淳.数学建模活动在创新教育背景下的探索与实践[J].科教导刊, 2015(05Z):2.DOI:CNKI:SUN:KJDK.0.2015-05-015.