编号：10 未经许可不得它用

国家信息中心2021年度青年人才基础研究项目成果

# 利用机器学习等大数据算法支撑宏观经济监测预测研究

部 门：大数据发展部

项目负责人：易成岐

目 录

[一、研究背景 8](#_Toc93926371)

[（一）时代背景：数字经济时代的到来及深刻影响 9](#_Toc93926372)

[（二）应用牵引：国内外政府部门的应用实践探索 11](#_Toc93926373)

[（三）问题导向：大数据时代宏观经济监测预测面临的困难 14](#_Toc93926374)

[二、研究综述 15](#_Toc93926375)

[（一）大数据在经济监测中的应用 15](#_Toc93926376)

[（二）大数据在经济预测中的应用 21](#_Toc93926377)

[（三）时序预测算法模型改进优化研究 26](#_Toc93926378)

[（四）大数据支撑宏观经济预测的四点启示 30](#_Toc93926379)

[三、研究目标和研究内容 40](#_Toc93926380)

[（一）研究目标 40](#_Toc93926381)

[（二）研究内容 40](#_Toc93926382)

[（三）研究框架 42](#_Toc93926383)

[四、研究内容之一：基于传统时序预测方法的宏观经济指标预测研究 42](#_Toc93926384)

[（一）研究思路 42](#_Toc93926385)

[（二）基于ARIMA的单变量预测 45](#_Toc93926386)

[（三）基于Prophet的单变量预测 49](#_Toc93926387)

[（四）基于LASSO的多变量预测 52](#_Toc93926388)

[（五）基于BSTS的多变量预测 55](#_Toc93926389)

[（六）实验结论 57](#_Toc93926390)

[五、研究内容之二：基于机器学习方法的宏观经济指标分类预测研究 58](#_Toc93926391)

[（一）研究思路 58](#_Toc93926392)

[（二）构建“三库” 59](#_Toc93926393)

[（三）“三库碰撞”数据实验 64](#_Toc93926395)

[六、研究内容之三：基于机器学习方法的宏观经济指标数值预测研究 73](#_Toc93926396)

[（一）融合机器学习算法和多重因素的白条鸡日均交易量预测 73](#_Toc93926397)

[（二）利用机器学习算法和广泛数据特征预测某市规模以上工业增加值 85](#_Toc93926398)

[七、研究内容之四：基于深度学习方法的宏观经济指标数值预测研究 97](#_Toc93926399)

[（一）研究思路 97](#_Toc93926400)

[（二）研究方法 97](#_Toc93926401)

[（二）搭建平台 99](#_Toc93926402)

[（三）数据实验 99](#_Toc93926403)

[（四）结论展望 106](#_Toc93926404)

[八、总结与展望 106](#_Toc93926405)

[（一）总结 106](#_Toc93926406)

[（二）展望 108](#_Toc93926407)

[参考文献 110](#_Toc93926408)

[附件：项目成果 123](#_Toc93926409)

利用机器学习等大数据算法支撑宏观经济 监测预测研究

摘 要

近年来，党中央、国务院高度重视运用大数据手段提高宏观经济分析研判水平。2015年8月，国务院印发的《促进大数据发展行动纲要》（国发〔2015〕50号）首次提出：“要建立运行平稳、安全高效的经济运行新机制，实现对经济运行更为准确的监测、分析、预测、预警，提高决策的针对性、科学性和时效性。”2018年2月，党的十九届三中全会再次指出：“综合运用大数据、云计算等技术手段，增强宏观调控前瞻性、针对性、协同性”。2020年5月，中共中央、国务院正式印发的《关于新时代加快完善社会主义市场经济体制的意见》对强化经济监测预测预警能力提出明确要求：“强化经济监测预测预警能力，充分利用大数据、人工智能等新技术，建立重大风险识别和预警机制”。在党中央、国务院的高度重视和坚强领导下，以及社会各界的广泛参与和大力推动下，利用大数据提升政府治理、宏观决策和产业转型水平，已经成为社会各界广泛共识。目前，国内外各级政府、学术界、产业界均广泛关注运用大数据、人工智能等新技术优化改进宏观经济监测预测的理论和实践。日益复杂多变的经济环境对这项工作提出了严峻挑战，也为大数据开展宏观经济监测预测研究范式的兴起提供了新的契机。

大数据是近十余年来学术界和产业界高度关注的新兴研究领域。1995年，Fayyad等学者在国际数据挖掘领域的顶级峰会——知识发现会议（KDD）上首次提出大数据的概念，并将其基本目标概括为两个方面：描述（Descriptive）与预测（Predictive）。前者是针对海量数据背后所隐藏的各种关联模式开展挖掘。维克托·迈尔·舍恩伯格等学者认为，由于大数据突破了传统样本采集方式的数据规模局限，而得以在很大程度上采用全样本海量数据开展分析，因此其可以大量使用相关性挖掘的方法，发现隐藏在海量数据背后的线索性信息。后者则是根据数据中潜在的模式进行预测。不同于传统统计学主要基于有限统计样本数据的预测方法，大数据可以帮助人们突破样本采集方法的局限，实现对全样本、全天候、全场景、全方位数据的采集，从而有助于提升对经济社会运行主体的预测能力。随着全社会范围内万物互联化、数据泛在化趋势不断兴起，大数据已经成为支撑宏观经济运行分析的重要手段。

利用机器学习等大数据算法支撑宏观经济监测预测研究是一个典型的多学科交叉研究领域。随着大数据的兴起，运用各种非传统统计数据开展宏观经济分析研判不仅成为经济学研究者关注的焦点问题，同时还吸引了包括信息科学、生物学、物理学、社会学等多学科研究者共同关注的热点领域。从该领域较为活跃的研究团队的学科背景看，经济学家是其中的重要力量，但绝不是唯一关注这一领域的学科群体。近年来，来自信息科学、生命科学、物理学、复杂科学等其他领域的研究者也开始大量关注宏观经济运行中的大数据分析。特别是，近年来美国经济研究局、经济分析局、劳工统计局、人口普查局等多个宏观经济部门均已探索深度利用大数据手段支撑辅助宏观经济分析。相比之下，我国在该领域实践尚处起步阶段，缺乏更为系统深入的利用机器学习、深度学习等新技术开展宏观经济监测预测实证研究成果。

在此背景下，课题组积极探索利用机器学习、深度学习等大数据算法开展宏观经济监测预测研究，尝试利用新技术应对宏观调控工作面临的新机遇、新模式、新挑战。主要工作及贡献包括：

**一是开展了基于传统时序预测方法的宏观经济指标预测研究。**依托国家信息中心大数据发展部所掌握的覆盖全国范围1.1亿家企业和个体工商户工商注册、就业招聘、招投标等数据的企业生产经营行为大数据平台，利用传统ARIMA和Prophet模型开展了单变量时序预测实验，利用传统LASSO和BSTS模型开展了多变量时序预测实验，分析发现无论单变量预测还是基于外部数据的多变量预测，传统时序预测模型均有明显的局限性，现有模型对指标的定性预测关注不足，其本质都是以趋势、季节、特殊事件以及数值间的距离为中心的定量预测，而在评估预测结果时，最常用的标准则是数值误差的大小。但在实际预测中，波动模式的预测相对简单，而精准的数值预测难度较大，特别是对于PMI这一类存在参照阈值的指标，如果通过预测值无法判断实际值未来是高于还是低于荣枯线，那么预测结果的价值将大打折扣。

**二是开展了基于机器学习方法的宏观经济指标分类预测研究。**针对上述传统方法存在突出的问题，本研究提出“三库碰撞”方法，尝试通过机器学习分类算法判断未来预测值的性质。课题组从国家统计局官方网站获取月度统计数据并整理构成监测指标库，共11类489个指标84034条数据，用以作为“三库碰撞”实验的预测目标。之后，从企业生产经营行为大数据平台提取出抽取6项企业数据，按地域（31个省份）、性质（国企、民企、外企）、行业分类（20个行业大类）进行多种数值变换处理，经过数据抽取维度、数据分类维度和数值加工处理维度等交叉匹配，最终形成共1296个数据特征，用作机器学习分类预测的数据特征库。实验结果发现，与监测指标的历史上升下降或荣枯背景分布概率相比，该方法对大部分经济指标T+1预测的准确率均有明显提升（准确率最高提升33%），从目前实验结果看，“三库碰撞”方式可适用于大部分经济指标的升降预测或荣枯预测，特别是，对于建筑材料及非金属矿类购进价格指数和制造业采购经理指数等部分经济指标，在使用该方法前提下随着训练时间的推移预测准确率能够持续提升。

**三是开展了基于机器学习方法的宏观经济指标数值预测研究。**针对大部分宏观经济指标历史数据较为缺乏、可获取的宏观经济指标影响因素数据较多等特点，可兼容自变量多、因变量少的机器学习预测方法为宏观经济指标数值预测研究提供了新的解决思路。课题组对融合机器学习算法与多重数据特征的数值型预测方法开展了研究。一方面是白条鸡日均交易量数值预测研究，该研究在数据特征选取过程中，兼顾鸡禽类产业、互联网舆情信息、网民需求意愿、白条鸡相关统计数据等多种趋势变化，综合考虑多方面因素对白条鸡交易量的影响，共提取了50个可能与白条鸡日均交易量预测有关的变量作为数据特征，包括市场主体特征30个、舆情信息特征2个、搜索意愿特征14个和统计数据特征4个。实验结果显示，除线性回归算法在预测效果稳定性表现不佳外，其余三个算法均具有预测效果的稳定性，且预测效果从优至劣依次为：GBRT、RandomForest、ElasticNet、LinearRegression，随着时间不断积累以及训练样本数量不断增加，白条鸡日均交易量预测效果不断提升。另一方面是某市规模以上工业增加值数值预测研究，课题组选定某市规模以上工业增加值数据作为因变量，选定某市统计数据、某市企业工商登记注册数据、某市互联网招聘数据作为自变量构建预测函数，采用平均绝对误差（MAE）和均方根误差（RMSE）进行效果评估，设计了数据特征、模型算法、训练样本、提前周期4组对比实验。实验结果表明利用高频数据和机器学习算法开展某市逐月统计数据预测表现出了最高的平均预测准确率，即利用机器学习和多重数据特征的方法对某市规模以上工业增加值的预测起到重要作用。

**四是开展了基于可解释性深度学习方法的宏观经济数值预测研究。**课题组选用了全球目前较为先进的N-BEATS深度学习模型开展宏观经济数值预测研究，通过对N-BEATS的模型原理、技术体系及平台架构进行研究剖析，我们搭建了基于N-BEATS深度学习技术的宏观经济监测预测实验平台，该预测平台采用三层次的多级信号编码，包括全连接神经网络层（FCNN）、预测块层和预测栈层，其中，全连接神经网络层实现了预测和拟合两方面的信息提取，分别用于时序预测和信息拟合的参数学习；预测块层之间通过残差数据流连接，使得预测块的拟合函数可对输入进行顺序过滤；预测栈层用于构成预测平台的顶层系统，每个栈被限制为不同的归纳函数，如趋势和周期函数。根据数据实验结果持续对N-BEATS模型进行参数调优，开展与目前较为流行的BVAR等传统时序预测模型对比实验。结果表明在经济极端环境下，如重大事件产生的波动或疫情动荡时，N-BEATS模型展现出了远超过经济学模型的性能。此外也为经济学知识的引入提供了可能，如将函数约束设置为某种特定的经济公式，可进一步提高N-BEATS模型预测准确率。

## 一、研究背景

近年来，党和国家高度重视运用大数据手段提高宏观经济分析研判水平。2015年发布的《促进大数据发展行动纲要》（国发〔2015〕50号）特别指出[[[1]](#footnote-1)]，要建立运行平稳、安全高效的经济运行新机制，实现对经济运行更为准确的监测、分析、预测、预警，提高决策的针对性、科学性和时效性。党的十九届三中全会[[[2]](#footnote-2)]提出要“综合运用大数据、云计算等技术手段，增强宏观调控前瞻性、针对性、协同性”；四中全会[[[3]](#footnote-3)]再次提出要“优化经济治理基础数据库。”在党和政府的高度重视和大力推动下，社会各界对大数据高度重视，促进大数据发展、以大数据提升政府治理、宏观决策和产业转型水平，有效释放数据红利，已经成为社会各界广泛共识。从全球范围来看，全球政府治理模式正在从传统的韦伯模式和新公共管理（NPM）模式过渡到数字治理（DEG）模式，其基本特征就是将大数据和数字化技术置于机构层级的核心位置，推动数字化的整体性政府建设，在决策模式上高度强调“使用数据来理解公民，并为政策制订提供依据”[[[4]](#footnote-4)]。充分发挥大数据技术优势，助力提升国家经济监测预测和宏观调控水平，已经成为大势所趋。本研究主要关注在大数据时代，经济学研究和实践所面临的机遇和挑战，并从理论建构和实践探索两方面展开论述，主要研究背景包括以下三方面。

### （一）时代背景：数字经济时代的到来及深刻影响

进入21世纪以来，随着以互联网、物联网、移动通信、移动计算等为代表的新一代信息技术在全球范围内的蓬勃发展，人类社会正在进入以数字化生产力为主要标志的全新历史阶段。2017年7月，习总书记在二十国集团领导人汉堡峰会就世界经济形势发表讲话时指出：“研究表明，全球95%的工商业同互联网密切相关，世界经济正在向数字化转型”[[[5]](#footnote-5)]。所谓数字经济，是以数据作为关键生产要素，以网络通信技术作为核心驱动力，以现代信息网络作为重要载体，重塑经济社会发展与治理模式的新型经济形态，主要包括数字产业化（数字经济基础部分）和产业数字化（数字经济融合部分）两个方面。数字经济是继农业经济、工业经济之后新的经济社会发展形态，对于国民经济各部门具有十分广泛的辐射带动效应，对提高我国经济效率、促进经济结构加速转变具有强大的驱动作用。从数字经济发展总体情况来看，据中国信通院《中国数字经济发展白皮书（2020年）》统计，2019年我国数字经济增加值规模达到35.8万亿元，占GDP比重达到36.2%。自2012年以来，我国数字经济增速一直高于GDP增速，对GDP增长的贡献率不断攀升。随着数字技术的持续发展与深化应用，数字经济已成为我国社会经济发展提质增效的新动能。

近年来，我国坚持以供给侧结构性改革为主线的产业转型升级战略，有力推进了重点行业数字化升级改造进程。随着互联网、大数据、人工智能等新一代信息技术与实体经济持续融合渗透，从2013年起，我国以产业数字化转型为主要特征的融合型数字经济常年保持20%以上的高速增长态势。2019年我国产业数字化增加值约为28.8万亿元，占GDP比重为29.0%，其中服务业、工业、农业数字经济渗透率分别为37.8%、19.5%和8.2%。面向制造业的数字经济蓬勃发展，以数字化研发工具的创新应用为突破口发展数字经济，是各行业中数字经济发展最具潜力的领域。以装备行业为例，行业数字化研发设计工具普及率达到79.8%，其中交通设备制造行业超过80%。以石化行业为代表的原材料行业以强化制造环节的智能化水平为着力点，在生产设备数字化率、数字化生产设备联网率和智能制造就绪率方面位于各行业领先水平。消费品行业基于互联网构建用户需求的精准采集、快速传导和实时响应的新能力，促进“互联网+”在消费品行业中的不断渗透和深化应用。重点行业的数字化转型取得突破性进展，逐渐成为驱动我国数字经济增长的重要引擎。

随着数字经济飞速发展，万物互联化、数据泛在化的大趋势日益明显。对于经济学而言，大数据不仅仅意味着对数据处理技术和处理能力的提升，更重要的是对散落在全社会各处的实时、交互、海量、离散、非结构化经济运行数据的采集、分析、处理、挖掘能力有了巨大提升，这对经济学研究方法影响十分深远。一方面，大数据已经对传统经济学研究方法提出了明确挑战，正如谷歌首席经济学家哈尔·瓦里安（Varian H R）所言，现代经济学要与数据打交道，而传统分析用的是样本等小数据，随着经济交流的日益频繁和技术水平的提高，数据越来越大，大数据的出现使得传统经济学分析方法在分析时显得捉襟见肘[[[6]](#footnote-6)]。另一方面，很多经济学家相信，大数据也将对经济学理论进步产生巨大推动作用。1944年，冯·诺伊曼在《博弈论和经济行为》中指出，17世纪的物理学大发展，其基础是几千年系统、科学的天文观察，特别是第谷的天文观测所积累的数据和知识。而“在经济科学中，没有任何类似事件发生。在物理学中，设想没有第谷而出现开普勒和牛顿是荒唐的，我们没有理由希望经济学的发展会比物理学的发展更容易。”[[[7]](#footnote-7)]经济学家陈禹则认为，“大数据的出现的确可以和第谷的工作相比拟。”[[[8]](#footnote-8)]因此，经济学抓住大数据发展的机遇，有可能早日迎来自身的“决定性的转折”。从这个意义上说，构建顺应于数字经济和大数据时代的经济分析框架势在必行。

### （二）应用牵引：国内外政府部门的应用实践探索

#### 1.国外的实践探索

2012年3月29日，美国白宫网站发布了《大数据研究和发展倡议》（Big Data Research and Development Initiative）[[[9]](#footnote-9)]，该倡议书中提到，大数据计划的实施，旨在帮助美国获得从海量复杂数据集中萃取知识的能力，借此提高国家应对急迫挑战的能力。这一计划类似于过去美国联邦政府在高性能计算和互联网等领域的战略性投入，旨在从海量数据信息获得、组织和收集知识所必需的工具和技能，提升国家在科学研究、教育和国家安全等领域的能力。该倡议书的发布，标志着大数据已经上升成为美国的国家战略。此后，主要西方发达国家均发布与大数据相关的宏观政策优化战略，如澳大利亚政府发布的《公共服务大数据战略》、法国政府印发的《数字化路线图》、英国政府发布的《把握数据带来的机遇：英国数据能力战略》等等[[[10]](#footnote-10)]。总体而言，目前西方发达国家政府运用大数据开展宏观经济监测预测的实践尚处在起步阶段，但也形成了一些成熟经验和模式。

**一是利用大数据技术丰富和提升传统统计手段。**美国经济研究局（NBER）2019年3月举办了题为“面向21世纪的经济统计大数据”（Big Data for 21st Century Economic Statistics）的专题研讨会。通过此次会议可以发现，美国宏观经济运行的多个相关部门都在积极探索利用大数据等新技术手段改进宏观经济统计分析方法，如美国经济分析局（BEA）[[[11]](#footnote-11)]尝试对比机器学习和替代性数据在经济预测中的效果，美国劳工统计局（BLS）[[[12]](#footnote-12)]尝试利用第三方数据、企业数据和Web抓取的零售商API数据来优化消费者价格指数（CPI）的数据采集，美国人口普查局[[[13]](#footnote-13)]利用机器学习和公共数据尝试自动化生成北美工业分类系统（NAICS）代码，等等。

**二是利用大数据技术优化宏观经济决策。**早在大数据技术诞生的初期，各国政府就十分关注大数据技术在公共政策领域的应用[[[14]](#footnote-14)]。如澳大利亚政府通过采集交通系统感应器的海量信息，构建了交通密度指标，并根据该指标观察出经济活动与交通密度之间的高度相关性，以此作为经济政策的评价工具之一[[[15]](#footnote-15)]。德国政府利用家庭电表反馈的海量数据制定具有可行性的激励政策以增加对智能电网基础设施的投资，调整传统的补贴方式和规制方式，根据需求优化电力资源配置[[[16]](#footnote-16)]。英国政府曾组织开展基于社交媒体的政策效果预测研究，通过在一系列社交平台中系统性地抓取数据，针对即将推出的经济政策变动在社交媒体中反应进行预测性分析[[[17]](#footnote-17)]。

**三是构建全国大集中的政府宏观决策数据中心。**在这方面，澳大利亚社会保障服务信息中心（Centrelink）和新加坡的“风险评估和水平扫描”系统（RAHS）是代表案例。Centrelink是澳大利亚联邦政府的一个政府机构，是公共事业部（Human Services portfolio）的六大机构之一，在联邦和各州都设有，其在堪培拉设有两个大型数据中心，与联邦和各州的税务、金融、警局等部门实现联网共享，并运用大数据分析技术构建了一系列围绕就业、社会福利、医保等领域的风险预测评估模型[[[18]](#footnote-18)]。新加坡建设的“风险评估和水平扫描”系统（RAHS）最初目的是应对恐怖主义和传染病，后来逐渐扩展到住房、交通、教育、安全等各个领域，其不仅可以帮助新加坡各级部门监测和识别恐怖袭击等风险事件，还可以帮助政府规划采购周期和预算、预测经济走势、制定移民政策和研究房地产市场。

#### 2.国内的实践探索

大数据同样引起了我国政府的高度关注。2015年，国务院正式印发《国务院关于印发促进大数据发展行动纲要的通知》（国发〔2015〕50号），标志着大数据正式上升为国家战略。党的十九大报告提出，要建设网络强国、数字中国、智慧社会，推动互联网、大数据、人工智能和实体经济深度融合，培育新增长点、形成新动能。2017年12月8日，习近平总书记在中共中央政治局就实施国家大数据战略进行第二次集体学习时指出，大数据是信息化发展的新阶段。他指出，随着信息技术和人类生产生活交汇融合，互联网快速普及，全球数据呈现爆发增长、海量集聚的特点，对经济发展、社会治理、国家管理、人民生活都产生了重大影响。在党和政府的高度重视和大力推动下，社会各界对大数据高度重视。据课题组不完全统计，近年来全国已经有25个省、自治区、直辖市设立了大数据局或大数据中心等职能部门，促进大数据发展、以大数据提升政府治理、宏观决策和产业转型水平，有效释放数据红利，已经成为社会各界广泛共识。据课题组不完全统计，国家发展改革委、国家税务总局、国家市场监管总局、国家统计局等多个中央部委均开展了不同层面的宏观经济大数据分析研判实践探索。

地方层面，海南、重庆、河北、深圳、杭州、青岛、某市等地方政府也由当地宏观部门牵头，在利用大数据开展宏观经济分析方面开展了大量有益探索。如2017年11月，依托国家发展改革委大数据中心重庆分中心建设的重庆经济社会发展大数据决策支持平台项目正式启动，平台归集了与重庆市经济社会运行相关的17亿条数据资源，并构建了产业地图、投资地图、消费地图、外贸地图、创新地图、人才地图等分析板块，得到中央政治局委员、重庆市委书记陈敏尔[[[19]](#footnote-19)]等领导同志表扬。河北省发改委[[[20]](#footnote-20)]提出加快“全委信息化系统整合及大数据建设”的工作任务，通过构建信息资源目录和大数据系统实现了河北省发展和改革委员会全委政务信息资源共享、建立完善的宏观经济数据采集渠道、宏观经济大数据融合分析及可视化展示等目标，并将时序算法、预测算法、相关性算法、聚类算法、影响因素等五类算法应用于宏观经济分析全生命周期。海南省发展改革委建设了海南省宏观经济决策大数据分析系统，其发布的《2018年海南省经济社会发展大数据分析报告》，从重点产业结构优化、招商引资成果、国际知名度提升、商标品牌建设、基层医疗卫生投入成效等几个方面，对习近平总书记“4·13”重要讲话和中央12号文件发布后海南省重点公共政策实施效果进行了全面评价，取得了良好效果[[[21]](#footnote-21)]。深圳市、青岛市发展改革委先后启动“智慧发改”工程建设，全面构建以大数据为核心的新型决策支撑体系。2016年，某市地税局启动了“某市税收发展指数”课题研究，基于地方税收大数据开展挖掘应用，课题组将反映经济运行的税收指数通过建指标、合成统一数值的方法，所形成的税收指数与统计局所发布的PPI数值相关性达70%左右[[[22]](#footnote-22)]。

### （三）问题导向：大数据时代宏观经济监测预测面临的困难

#### 1.时效难，依赖历史统计数据难以实时研判最新态势

现行的经济监测预测方法大多基于历史统计数据进行分析或预测未来的数据变化走势，但此方法往往受制于统计数据的时效性。传统统计数据主要是通过开展经济普查、抽样调查、重点调查、典型调查等手段，利用层层报送数据的方式汇总计算相应宏观经济指标，这种方法最明显的缺陷是具有很强的时滞性。在应对重大突发事件和高频度经济震荡时，传统经济监测预测方法很可能会导致政府决策部门后知后觉，容易造成决策失误等严重后果。

#### 2.细分难，统计样本的局限性容易导致分析粒度不足

统计样本的数量和样本获取的成本往往成反比，随着统计样本量的增加，统计过程中的时间成本、人力成本和资金成本等均会相应急速增高，因此统计样本的覆盖范围存在一定的局限性。由于传统经济监测预测方法高度依赖统计数据样本，一旦样本出现偏颇，一方面，很容易导致分析预测结果精度不足，另一方面，很难为细分行业及新兴行业的经济决策提供有效支撑。

#### 3.客观难，人为误报等因素可能会造成结果不够客观

传统统计数据的产生需要大量使用人工进行搜集、加工、上报及汇总，不仅人为参与的程度较深、耗费的时间周期较长，而且统计流程也较为复杂。由于人为参与了全过程中过多环节，无论是出于主观考虑或出现客观失误，都极易发生数据迟报、漏报、瞒报、误报等现象，此时往往会存在统计样本真实性偏差的隐患，如果使用这种存在偏差的数据进行经济监测预测，则会进一步扩大误差范围。

#### 4.精准难，经济监测预测尚未与最先进技术紧密结合

以经济学家主导的经济监测预测模型拥有相对成熟的经济理论作为支撑，主要包括结构化计量预测模型、非结构化时间序列预测模型和动态随机一般均衡预测模型三类，其能够较好地解释经济社会运行的关键影响因素。但当今世界，全球发展深层次矛盾突出，保护主义、单边主义思潮抬头，多边贸易体制受到冲击，我国经济也由高速增长阶段转向高质量发展阶段，全球经济整体发展环境面临诸多风险和不确定性，面对这一错综复杂的变化，传统计量方法的种种假设也常常不符合当下的经济现实。同时，传统经验理论往往与现实情况会产生一定偏差，传统经济监测预测模型略显乏力。而目前机器学习技术已经在诸多领域展现出超凡的预测能力和实用价值，但是却未能与经济监测预测深度结合。

不难发现，基于历史性统计数据去分析或预测未来的数据变化走势会存在时效难、细分难、客观难、精准难等诸多难题，大量理论方法和实践经验已经在一定程度上证明了利用高频数据和机器学习算法开展经济预测的可能性。

## 二、研究综述

### （一）大数据在经济监测中的应用

经济系统是由无数个政府机构、企业、个体消费者、媒体和各类社会组织共同组成的巨复杂系统。从经济学诞生之初，经济学家就致力于通过采集各类数据观察经济现象运行的内在规律，但这个问题从来就是困扰经济学家和统计学家的最大难题之一[[[23]](#footnote-23)]。传统经济学研究开展经济监测的主要手段就是开展经济普查，通过层层报送归集数据的方式汇总计算相应宏观经济指标，如国民生产总值（GDP）、通胀率、消费品物价指数、生产者价格指数、失业率等等[[[24]](#footnote-24)]。但正如前文所论述的，这种基于传统调查统计方式开展宏观经济监测的方法存在着数据时滞较长、成本居高不下、样本覆盖面较窄、无法进行细颗粒度分析等问题。在这种情况下，基于大数据手段改进经济监测效果，就成为学术界和政府决策者共同关注的热点问题。总体而言，大数据在经济监测中的应用方面主要包括四个大的方面。

#### 1.传统统计调查方法对大数据手段的接纳

近年来，大量文献开始关注大数据对于传统经济普查和政府统计的改进[[[25]](#footnote-25)][[[26]](#footnote-26)][[[27]](#footnote-27)][[[28]](#footnote-28)][[[29]](#footnote-29)][[[30]](#footnote-30)][[[31]](#footnote-31)][[[32]](#footnote-32)][[[33]](#footnote-33)]。有学者认为，GDP指标源于上世纪30年代，当时尚未形成政府利用数据和统计来改善系统性经济弊病的传统，所以GDP这样的指标就有助于决策者在绝望的时刻驾驭很多政策试验。但是今天，人们如果还使用现行指标，就会造成在一些国家阻止了政策的创新而非促进创新[[[34]](#footnote-34)]，因此国民经济核算体系有必要对自身做出一定的改进甚至改革[[[35]](#footnote-35)]。李红艳等[[[36]](#footnote-36)]提出新型国民经济核算体系的概念，并提出新体系应当实现对全社会经济活动进行全方位动态跟踪，对每一个产品（服务）从其产生到被消费（消亡）的全过程进行跟踪记录，对市场中每笔交易活动进行跟踪记录，对每个经济主体的每次经济活动和经济关联活动进行跟踪记录。在政府实践层面，早在2013年11月，国家统计局就与百度、阿里巴巴等11家企业签署了《大数据战略合作框架协议》，就建立大数据应用的统计标准，完善政府统计数据的内容、形式和实施步骤等达成合作。[[[37]](#footnote-37)]在2020年的全美经济普查中，美国普查局也计划采取“网络抓取技术”等多种模式进行数据收集，试图实现企业数据与普查、调查数据的全面融合，提高统计调查的效率和质量[[[38]](#footnote-38)]。

#### 2.基于大数据构建新型监测指标

随着互联网的飞速发展，人们逐渐具备了一些快速归集反映经济运行某一个剖面的全样本数据的手段和技术工具，从而可以构建观测经济运行某一个剖面的相对全样本新型监测指标。最著名的非传统经济指标之一就是克强指数[[[39]](#footnote-39)]。所谓克强指数（Li keqiang Index），是英国著名政经杂志《经济学人》在2010年推出的用于评估中国GDP增长量的指标，其源于李克强总理2007年任职辽宁省委书记时，喜欢通过耗电量、铁路货运量和贷款发放量三个指标分析当时辽宁省经济状况。该指数是三种经济指标：工业用电量新增、铁路货运量新增和银行中长期贷款新增的结合。自推出后，受到花旗银行在内的众多国际机构认可。再比如，最早由三一重工集团推出的挖掘机指数[[[40]](#footnote-40)]，通过对三一重工集团运营管理的遍布全国的重型机械上机载控制器、传感器和无线通讯模块所汇集的汇聚，描述各个经济区域的基础建设开工率，进而成为宏观经济部门分析固定资产投资进度的重要参考指标。此外，近年来被人们所广泛关注的各种非传统经济指标还有所谓榨菜指数、啤酒指数、男士内裤指数、床垫指数、用云量指数等等。这些指数的共同特征，是选择了经济社会运行的某一个环节，通过对相对全样本范围内数据的归集分析，从而推测经济运行的实际情况，并与传统统计指标相互印证。此外，在以互联网、大数据、分享经济等为代表的新经济领域，由于缺乏较为完善的统计指标体系，大量由大型互联网企业和社会组织提出的创新型指标不断出现。如阿里巴巴基于覆盖阿里电商平台数百万种商品数据构建的中国县域电商发展指数aEDI[[[41]](#footnote-41)]、阿里巴巴网购价格系列指数aSPI[[[42]](#footnote-42)]，财新智库和数联铭品（BBD）联合推出的新经济指数（NEI），国家信息中心大数据发展部近年来主持发布的中国大数据发展指数[[[43]](#footnote-43)]、分享经济发展指数[[[44]](#footnote-44)]等。

#### 3.大数据“现时预测”（Nowcasting）研究

近年来，很多研究者开始利用大数据高频、即时的特点，通过构建与传统经济监测指标具有高度关联性同步指标，提高经济监测的效率，也就是近年来十分热门的“现时预测”（Nowcasting）研究。第二类研究相比，这类研究所构建的经济指标往往是对标某个主流统计经济指标，并利用大数据即时获取的特点，利用大数据与统计数据的“时间差”优势比传统统计方法更早获取经济运行情况。从实际研究来看，从事这方面研究的学者最多，成果也最集中。举例来说，运用电力大数据开展宏观经济分析研判[[[45]](#footnote-45)][[[46]](#footnote-46)][[[47]](#footnote-47)]；利用夜间卫星灯光亮度数据对区域和国别经济生产率数据进行现时预测[[[48]](#footnote-48)][[[49]](#footnote-49)]；基于用户移动通话行为模式[[[50]](#footnote-50)][[[51]](#footnote-51)]和社交网络结构特征[[[52]](#footnote-52)][[[53]](#footnote-53)]现时预测区域经济活跃度；基于银行支付结算数据现时预测宏观经济增长态势[[[54]](#footnote-54)]；利用搜索指数数据提高对消费者信心指数[[[55]](#footnote-55)][[[56]](#footnote-56)][[[57]](#footnote-57)][[[58]](#footnote-58)][[[59]](#footnote-59)][[[60]](#footnote-60)][[[61]](#footnote-61)]和通胀增长情况[[[62]](#footnote-62)][[[63]](#footnote-63)]的现时预测水平；基于各类电商平台数据[[[64]](#footnote-64)][[[65]](#footnote-65)][[[66]](#footnote-66)]、搜索数据[[[67]](#footnote-67)][[[68]](#footnote-68)][[[69]](#footnote-69)]和媒体数据预测全社会商品价格波动情况；利用网络搜索数据[[[70]](#footnote-70)][[[71]](#footnote-71)][[[72]](#footnote-72)][[[73]](#footnote-73)][[[74]](#footnote-74)][[[75]](#footnote-75)][[[76]](#footnote-76)][[[77]](#footnote-77)][[[78]](#footnote-78)][[[79]](#footnote-79)][[[80]](#footnote-80)]、手机通话记录数据[[[81]](#footnote-81)][[[82]](#footnote-82)][[[83]](#footnote-83)][[[84]](#footnote-84)][[[85]](#footnote-85)]和在线招聘数据[[[86]](#footnote-86)][[[87]](#footnote-87)]现时预测失业率变化；利用搜索引擎数据[[[88]](#footnote-88)][[[89]](#footnote-89)][[[90]](#footnote-90)][[[91]](#footnote-91)][[[92]](#footnote-92)][[[93]](#footnote-93)][[[94]](#footnote-94)][[[95]](#footnote-95)][[[96]](#footnote-96)][[[97]](#footnote-97)]和新闻数据[[[98]](#footnote-98)]现时预测房地产市场走势，等等。

#### 4.基于大数据开展经济运行态势分析研判

在大数据时代，运用大数据进行产业和区域运行情况分析，提升信息分析效率是大势所趋[[[99]](#footnote-99)]，这方面研究大致可以分为几类。第一类是在微观层面上，针对各类企业运行情况的大数据分析监测。如Li等[[[100]](#footnote-100)]基于271 家中小型美国绿色食品和制造公司网站信息数据分析中小企业业绩受和政府、行业、学术界联系的影响。Daepp等[[[101]](#footnote-101)]分析了北美超过25000家上市公司，发现任何商业类型公司的生命半衰期大约都为10年。国内一些地方政府还利用各类政府业务数据，建立起区域企业运行情况监测大数据平台，对企业经营状况、财务状况、投资状况、产销能力、盈利能力、偿债能力、缴纳情况和能耗情况等进行全面监测[[[102]](#footnote-102)][[[103]](#footnote-103)]。第二类是在中观层面上，对区域产业发展情况进行分析。这方面研究也非常多，主要集中在对区域产业集聚路径[[[104]](#footnote-104)]、区域经济禀赋[[[105]](#footnote-105)]、经济均衡度[[[106]](#footnote-106)][[[107]](#footnote-107)][[[108]](#footnote-108)]、产业配套设施分布[[[109]](#footnote-109)]、产业空间分布[[[110]](#footnote-110)]、消费空间[[[111]](#footnote-111)]、经济活动多样性[[[112]](#footnote-112)]、功能区划定[[[113]](#footnote-113)][[[114]](#footnote-114)][[[115]](#footnote-115)][[[116]](#footnote-116)][[[117]](#footnote-117)][[[118]](#footnote-118)]、城市群网络结构[[[119]](#footnote-119)][[[120]](#footnote-120)][[[121]](#footnote-121)][[[122]](#footnote-122)][[[123]](#footnote-123)][[[124]](#footnote-124)][[[125]](#footnote-125)]、贸易网络[[[126]](#footnote-126)]等方面分析上。第三类是在宏观层面上，运用大数据手段开展宏观经济调控研究[[[127]](#footnote-127)]。如有研究者提出，应当运用大数据手段形成“全面+定向”宏观调控政策手段组合，全面提高宏观调控工具的精确度和效力，实现针对国民经济薄弱环节例如服务业、“三农”问题和小微企业精准发力，有的放矢的降低税率、费用、利息[[[128]](#footnote-128)]。

### （二）大数据在经济预测中的应用

经济预测是经济学最重要的职能之一，也是政府、企业、个人等进行决策的参考和依据[[[129]](#footnote-129)]。然而长期以来，人们对于经济学的预测理论一直存在质疑。哈耶克就曾做出论断，认为人类经济行为是不可预测的。他指出：“尽管我们拥有关于人类理智运行原理的知识，但主要事实依然是，我们不可能对导致个人在特定时间做了一件特定的事情之全部具体事实加以说明。对于我们来说，个人的个性仍然是一种非常独特的、难以计算的现象，我们有望通过一些从经验中发展出来的做法，比如赞扬和谴责，从可取的方向对它施加影响，但是对于它的具体行为，我们无法进行预测或控制，因为我们无法了解决定着它的全部具体事实”[[[130]](#footnote-130)]。

大数据时代的到来，使得人们对于经济学的预测从过去基于因果推断的预测理论范式转向基于概率论的预测理论范式。蔡跃洲认为，大数据对经济预测的影响涉及数据来源、预测方法、预测结果等几乎每一个环节，在某种程度上改变了常规经济预测所遵循的基本范式，他从经济预测的数据基础、指标生成方式和预测方法与工具等几个方面对大数据经济预测和传统经济预测进行了系统比较。随着大数据广泛获取经济社会主体各类行为数据能力的日益提升，将为测量经济社会主体预期和量化主体情绪提供了新的路径，有望渐进打开经济主体预期形成过程的“黑箱”[[[131]](#footnote-131)]，大大提高预测分析能力。李华杰等认为，基于大数据的预测方法在很多方面不是对传统经济学研究方法的颠覆，它更是一个对传统研究方法的补充[[[132]](#footnote-132)]。从现有研究成果来看，可以将大数据在经济学预测研究中的应用划分为构建先行指标、改进传统预测模型、构建预测模型和复杂性预测等几大类，后文将逐一详细论述。需要指出的是，很多文献将现时预测研究也纳入经济学大数据预测的范畴，但课题组认为其实质还是对当前发生的情况的即时分析，只不过利用统计方法相对滞后的“时间差”而实现了一定程度上的超前性，对其的介绍在上一小节已经有所展开，此处不再赘述。

#### 1.基于先行性指标的大数据预测

先行性指标预测具体又可以划分为三大类研究。

其一，是从人类经济行为链条的角度，通过对行为链条上一环节的分析来实现对下一环节行为的预测。如Liu等[[[133]](#footnote-133)]用PLSA 算法（Probability Latent Semantic Analysis）从博客内容数据中挖掘用户观点和情绪用以预测电影票房销售情况。Schneider等[[[134]](#footnote-134)]引入词袋模型（Bag-of-Words），基于Amazon.com的用户评论预测一周后的笔记本电脑的销售情况。Khadivi等[[[135]](#footnote-135)]通过分析Wikipedia使用数据预测夏威夷旅游需求。再比如，大量研究者基于股票论坛、微博、推特等自媒体公众言论情绪[[[136]](#footnote-136)][[[137]](#footnote-137)][[[138]](#footnote-138)][[[139]](#footnote-139)][[[140]](#footnote-140)][[[141]](#footnote-141)][[[142]](#footnote-142)][[[143]](#footnote-143)]，以及金融搜索数据[[[144]](#footnote-144)]判断人们对股市的心理预期，从而进一步预测股市运行走势。

其二是基于各种经济学理论构建能够超前反映经济运行走势的先行性指标。较为典型的如钱斌华[[[145]](#footnote-145)]基于职工人数变化率、总资产变化率、所有者权益变化率等指标构建某市税收发展先行指标；崔趁欣等[[[146]](#footnote-146)]基于河北省国库收入指数预测地区国民生产总值走势。张秋雁等[[[147]](#footnote-147)]基于电力消费情况预测经济景气度走势；曲延玲[[[148]](#footnote-148)]提出基于存贷款关键指标和社会融资结构变化等数据预测宏观经济走势；BRANDYN等[[[149]](#footnote-149)]基于纽约联储银行实时数据和动态因子模型预测GDP增速，等等。

其三是在海量非统计指标中析取具有一定先行性的指标，并合成为新的先行性指标。百度[[[150]](#footnote-150)]曾基于自身所掌握的海量搜索数据开发了百度经济指数预测，其基本原则就是在海量搜索数据中析取出与企业需求和用户经济行为信息相关的搜索指数，并合成为先行指数。LEVENBERG等[[[151]](#footnote-151)]提出了一种利用复合异质网络数据流来合成经济预测变量的方法，采用贝叶斯分类器组合模型对非农就业指数进行了高精度预测。SOBOLEVSKY等[[[152]](#footnote-152)]在分析西班牙个人银行转账记录数据的基础上，提出了35种个体经济行为量化指标，能够预测GDP、房屋价格、失业率、犯罪率、高等教育比例、生活成本和预期寿命等社会经济统计指标。IBM日本公司基于互联网新闻数据析取出制造业相关的480项经济指标，用于预测采购经理人指数[[[153]](#footnote-153)]。

#### 2.改进传统预测模型

清华大学刘涛雄等[[[154]](#footnote-154)]曾提出利用大数据优化改进传统统计预测模型的“两步法”，是这方面非常有代表性的理论方法。其基本步骤：首先，仅使用政府统计信息选择初步最优预测模型；其次，将互联网搜索行为加入选择的模型中，最终确定最优模型。Götz等[[[155]](#footnote-155)]利用Google 网络搜索数据结合偏最小二乘法、LASSO 方法等选择指标加入传统的桥梁方程模型（Bridge Equation Models）预测德国的GDP。Kholodilin等[[[156]](#footnote-156)]通过主成分分析法（PCA）来对Google Insights进行降维后整合进基于统计数据的美国私人消费增长率预测模型，发现可以提高20% 的预测准确率。AMURI等[[[157]](#footnote-157)]使用互联网求职指标（谷歌指数GI）扩展传统失业率预测模型，发现大多数州级预测和与专业预测者调查的比较中均有更好表现。ARTOLAS等[[[158]](#footnote-158)]通过增加相关Google搜索指数优化改进传统的ARIMA模型，发现其短期针对德国、英国和法国去西班牙旅游人数的预测精度提高了42%。许伟[[[159]](#footnote-159)]通过结合 Google 搜索数据和网络新闻情感，构建了基于网络情感和搜索行为的数据挖掘集成模型，在其中加入房地产价格指数时间序列的滞后项，利用支持向量回归SVR模型，实现了对房地产价格指数的更好预测。

#### 3.构建新的大数据预测模型

大数据在发现海量经济变量间关联关系和关联模式方面具有传统统计数据所无法比拟的优势。利用大数据的这一特性，很多研究者开始积极探索通过相关性分析挖掘建立经济变量间的联系，从而形成全新的经济学预测模型[[[160]](#footnote-160)]。在这方面被大量应用的是被统称为复杂性方法或非线性方法的复杂网络、非线性迭代、混沌分析等一系列方法。布莱恩.阿瑟详细探讨了资本市场中“泡沫和崩溃”、集群波动（clusted volatility）和突然渗透（sudden percolation）等三种非均衡现象的复杂性建模预测方法[[[161]](#footnote-161)]。王国成[[[162]](#footnote-162)]指出，通过利用大数据方法，对个体的冲动行为、情绪因子、羊群行为和不同主体对政策信息响应的不对称等有限理性或“非”理性行为进行分析，能够更合理地解释和预见股市诸多异象的微观成因、影响方式、演变路径及走势。此外，HIDALGO等[[[163]](#footnote-163)]和BUSTOS等[[[164]](#footnote-164)]提出解释经济增长和发展的新观点，发现所构建的“国家—产品”二部分网络的结构特征可以定量刻画国家经济的复杂性；经济复杂性与国家的收入水平非常相关，甚至可以用来预测国家未来的经济发展水平。高见等[[[165]](#footnote-165)]基于企业注册信息数据分析和定量刻画了中国区域经济复杂性，结果发现经济复杂性与人均GDP有很强正相关。CRISTELLI[[[166]](#footnote-166)]提出了一种数据驱动的非线性方法来预测国家经济发展趋势。TACCHELLA等[[[167]](#footnote-167)]提出了一种基于“国家—产品”矩阵的非线性迭代算法，同时刻画国家发展潜力和产品复杂性，所得结果比文献[[[168]](#footnote-168)]更好地解释不同国家的经济竞争力。

### （三）时序预测算法模型改进优化研究

#### 1.传统趋势预测方法

学术界对国民经济领域的统计指标预测研究大多采用传统趋势预测模型，包括指数平滑[[[169]](#footnote-169)]、自回归移动平均（ARIMA）[[[170]](#footnote-170)]和灰色预测（GM）[[[171]](#footnote-171)]等时间序列预测模型。例如：李志强等采用ARIMA模型预测内蒙古年度羊存栏量[[[172]](#footnote-172)]，王晓梅通过建立灰色理论GM(1,1)模型对我国主要畜产品年产量做出预测[[[173]](#footnote-173)]，林绍森等对常见的指数平滑、ARIMA和灰色预测模型在粮食产量预测中的效果进行比较[[[174]](#footnote-174)]。此外，如刘峰等构建的非平稳时间序列ARIMA模型能很好地基于白菜月价格的当前值和过去值预报未来白菜月价格趋势[[[175]](#footnote-175)]。虽然时间序列方法具有需要的特征数据种类少、计算简单等优势，但对历史数据量要求较高，且难以在预测中引入短期市场变化因素，无法及时反映产业运行、互联网信息、重要时间节点和重大突发事件等因素对指标预测造成的冲击。考虑到本课题预测对象——规模以上工业增加值具有历史数据较缺乏、可获取的影响因素数据较多的特点，自变量多、因变量少的机器学习预测方法为本课题研究提供了新的解决思路。

随着贝叶斯思想的不断发展，多元时间序列贝叶斯分析方面也已有一些文献发表，其中最为广泛传播的就是贝叶斯向量自回归（BVAR）模型。BVAR模型不仅能够克服一般联立方程模型需要严格区分内、外生变量的缺陷，而且在面临参数较多、时间序列长度较短、样本数据较少的问题时，会有良好的拟合效果。由于我国的金融市场数据时间跨度不长，相对于其他传统宏观经济预测模型，BVAR模型能够更精准的进行短期预测。但上述模型均是同频数据模型，由于GDP等宏观经济指标是按照季度统计，而像CPI、PPI等宏观经济指标是按月度统计，所以这种数据频率的差异在预测宏观经济变量中会造成研究范围受限。王飞（2011）以青海省为例建立了一个BVAR模型，样本内和样本外的预测误差都表明该模型的GDP预测效果优于向量自回归（VAR）、差分整合移动平均自回归（ARIMA）等回归模型。周建和况明（2015）构建BVAR模型对中国GDP预测误差进行分析，发现该模型优于现有文献中的其他方法。

#### 2.机器学习预测方法

目前，被广泛认定的机器学习定义是T. Mitchell于1997年提出的“计算机利用经验来改善系统自身的性能”。意指机器学习是计算机从数据中寻找内在规律，并用所学的经验来指导系统提升性能，从而对样本数据作出决策。机器学习不仅提供了新的分析工具，而且试图去解决传统计量方法所不能解决的问题。

目前，主流的机器学习算法主要包括线性回归[[[176]](#footnote-176)]、弹性网络回归（ElasticNet）[[[177]](#footnote-177)]、随机森林（RandomForest）[[[178]](#footnote-178)]和梯度提升回归树（GBRT）[[[179]](#footnote-179)]等，已成功应用于供应链、能源、卫生、交通等领域[[[180]](#footnote-180)][[[181]](#footnote-181)][[[182]](#footnote-182)][[[183]](#footnote-183)]，并体现出相对于传统时间序列预测方法的优势。如，董莉等基于网络搜索数据和ElasticNet建立的CPI及时预测模型，有效避免了时间序列和灰色预测等方法的时间滞后性[[[184]](#footnote-184)]。再如，康传利等的研究表明基于梯度提升回归树的旅游流量预测模型比指数平滑算法有更高的预测准确性[[[185]](#footnote-185)]。特别是近年来得到广泛应用的GBRT算法，在处理影响因素复杂且历史数据较少的预测中表现尤为突出，具有较高的预测精度[[[186]](#footnote-186)]。韩忠明等提出的早期电影票房预测模型对比实验结果显示，与随机森林模型、决策树模型和非线性回归模型相比，GBRT模型相对准确率最高[[[187]](#footnote-187)]。

监督机器学习方法，如回归树（regression trees）、LASSO、随机森林（random forest）、支持向量机（support vectir machines）、神经网络（neutral network）、深度学习（deep learning）等技术，其算法在预测上的优势，已被广泛地用于解决经济预测问题。例如，在回归数方面，Mullainathan和Spiess研究发现以回归树方法构建的房地产价格预测模型较传统方法具有较好的预测效果。[[[188]](#footnote-188)]葛浩发现利用提前两年的财务指标数据训练出的CART模型预测企业财务风险的效果最好。[[[189]](#footnote-189)]李颖基于GBRT集成模型构建了新冠疫情下宏观经济预测指标体系并得出了促进我国宏观经济发展的对策建议。[[[190]](#footnote-190)]在LASSO回归方面，贺平等人采用LASSO-logistic算法有效识别了特征变量与股票市场预期收益之间的复杂关系，解决了资产预测中的特征变量不全面、特征变量维度较高以及特征变量之间的高相关性导致的预测不稳定的问题。[[[191]](#footnote-191)]欧阳梦倩等人结合LASSO和核偏最小二乘回归研究了百度搜索指数对CPI的预测作用。[[[192]](#footnote-192)]胡清畅采用基于LASSO算法的自回归分部滞后模型（ADL）对数据进行拟合，研究表明充分使用政府统计数据与百度搜索指数能够使经济景气指数的拟合误差降低。[[[193]](#footnote-193)]陈梦根等人基于新常态经济发展背景研究了居民消费价格指数（CPI）的预测模型，发现LASSO回归分析的预测精度明显由于最小二乘回归与岭回归。[[[194]](#footnote-194)]在随机森林方面，Joy等人采用随机森林算法分析了发达经济体银行在货币危机爆发前的经济、金融和结构状况。[[[195]](#footnote-195)]高子宜将随机森林算法应用在股票分析研究上，随机森林模型可以有效消除系统性波动和行业波动对个股的影响，取得较好的预测效果。[[[196]](#footnote-196)]邹玉莹对票据转贴现利率预测进行理论和实证分析，随机森林对月利率具有较好的预测效果，商业银行可以采用随机森林模型对据转贴现周利率和月利率进行预测。[[[197]](#footnote-197)]在支持向量方面，屠星月等人基于EMD－SVM的农产品市场价格预测模型预测精度高，其为中国农产品价格短期预测提供了新的方法。[[[198]](#footnote-198)]肖雪梦建立了PLS-SVM模型预测了重庆市2013年的GDP值，证明了PLS-SVM模型在预测GDP数据上的有效性。[[[199]](#footnote-199)]在神经网络方面，何雁明等人研究表明BP神经网络在CPI指数上的预测具有较高的准确性。[[[200]](#footnote-200)]李杰研究了如何应用神经网络来对利率期限结构进行预测，结果表明，BPNN与WNN在利率期限结构预测上具有较好的预测效果。[[[201]](#footnote-201)]王永杰提出了PSO（粒子群优化算法）-BP神经网络算法，证明了PSO-BP神经网络用于经济预测的有效性。[[[202]](#footnote-202)]在深度学习方面，沈虹等人基于深度学习长短期记忆神经网络的有色金属期货市场预测，证实了LSTM模型是ARIMA模型的有效替代。[[[203]](#footnote-203)]李佳等人为了提高预测准确度，创新性地使用最新深度学习方法的独立循环神经网络（IndRNN）来提升预测效果，证明了IndRNN是当前较为有效的经济预测工具，在经济预测中具有较大应用价值。[[[204]](#footnote-204)]

#### 3.深度学习预测方法

通过对经济预测领域国内外最新研究成果跟踪调研发现，以机器学习和深度学习为代表的非线性模型近两年成为研究热点。郭庆春和何振芳（2014）使用改进的BP神经网络建立GDP预测模型，研究表明该模型具有较高的预测精度，并且还具有泛化能力强、训练速度快、稳定性好等优点。Choi和Lee（2018）使用深度学习LSTM模型来预测实时时间序列数据，结果表明该模型取得了已知各类模型中的最好预测效果。赵军豪等（2018）等使用融合微博情感分析和深度学习LSTM模型进行预测，发现该方法对数据抖动有良好的适应性，并具有很好的稳健性。特别是，2018年图灵奖获得者Yoshua Bengio教授近期发表了数篇有关利用深度学习开展时序预测的重磅论文，提出了一种用于可解释时间序列预测深度学习模型N-BEATS，其被证明了在多个数据集上具有良好的预测效果。

### （四）大数据支撑宏观经济预测的四点启示

过去十余年来，运用大数据手段对宏观经济监测预测开展实证分析，已经成为相关研究领域的热点话题。关于大数据在经济学中的应用前景，在经济学领域尚存在诸多争论。激进者认为，过去数百年来，受技术手段所限，包括经济学在内的社会科学只能基于随机采样方法归集统计数据，这“本身存在许多固有的缺陷”[[[205]](#footnote-205)]。而海量非干预数据的开放和使用会对经济学研究产生深刻影响[[[206]](#footnote-206)]，甚至“会从根本上改变我们理解世界的方式。很多旧有的习惯将被颠覆，很多旧有的制度将面临挑战”[[[207]](#footnote-207)]。观望者则认为，大数据方法更多体现在方法创新上，与传统经济学研究方法是互为补充和相互促进的关系，而非替代关系，大数据指数可以为传统统计调查提供佐证和补充[[[208]](#footnote-208)]。更有研究者指出，当前大数据在技术手段和理论范式方面尚存在诸多局限[[[209]](#footnote-209)]。结合对这些研究的回顾与梳理，课题组认为，研究大数据时代的经济学范式转变和实践路径时，应当注意以下四点基本原则：

#### 1.理性看待大数据在宏观经济分析中的应用

大数据作为一种新兴的技术手段，其所具有的及时性、精准性、客观性高等特点，使得人们可以构建面向个体、企业、事件、商品等微观主体数据模型，并运用离散数学和计算数学的方法进行中观层面的模式识别和规律总结，从而为宏观层面的决策提供依据，这对经济学发展意义十分重大。正如P·A·Attewell等人所指出的[[[210]](#footnote-210)]，以人的行为为基点，探讨深化经济研究与大数据及类似概念之间的内在关联，具有历史使然的规律性。但在看到大数据应用于宏观经济监测预测领域的诸多潜在优势的同时，也要承认，大数据并不是万能的，其由于技术手段和理论范式方面的诸多局限，也存在很多不足之处。正如汪毅霖所指出：“公平地说，传统经济学预测的准确性并没有人们印象中的那么差，大数据预测的效果也没有人们希望的那样好。”具体来说，大数据应用宏观经济研究目前尚存在以下问题：

一是大数据分析中经常会遇到的所谓“维数灾难”问题。在宏观经济模型中，由于大数据价值密度低的问题，往往需要通过大数据挖掘技术，获得足够多的数据信息，这包括多维度上的数据信息，因而解释变量会大大增加，因此研究中会出现高维数据问题，即因可得信息过多，往往会出现和被解释变量相关的解释变量数量过多，产生所谓的“维数灾难”[[[211]](#footnote-211)]。

二是大数据分析中大量应用的语义理解等新技术的不可靠性问题。宏观经济预测分析，离不开对大量非结构化数据，比如自媒体评论、新闻文本、政策文件、图像、视频等的处理。在处理这些信息时，大量所应用的技术发展还处在初级阶段，存在很多问题。比如在中文论坛中大量使用的网民口语化语言，对这些语言进行情绪分析、观点抽取、命名主体识别等自然语言处理时，其准确性往往会很低，难以像统计数据那样形成权威结论。

三是数据处理过程的主观倾向性问题。从方法论上说，大数据分析属于归纳性研究的范畴。本质而言，任何经验调查都依靠同一性、连续性和可测性的假设。这些假设不可能自己从经验数据中得出[[[212]](#footnote-212)]。大数据分析同样不可能完全从数据出发，而必然包含有一定的先验主观性。汪毅霖指出，大数据分析主张让数据自己发声，但大数据预测的程序中的两个关键节点——确定关键词和计算相关性——都具有很强的主观性。于是，大数据预测不仅要克服数据误差或者说数据的不确定性，也需要解决模型设定的不确定性[[[213]](#footnote-213)]。举例来说，当为了考察国外股票市场的情况而搜索“Bull”和“Bear”两个单词时，Google将会给我们反馈大量关于两种动物的信息[[[214]](#footnote-214)]。因此如果不对关键词依靠专家经验进行筛选判断，那么最终结果中会包含大量数据噪音。而既然在关键词和算法两个环节都具有经验数据之外的先验属性，我们就需要质疑某些大数据专家提出的可以不要理论的观点——关键词和算法模型的选择就是未来需要解决的重大理论问题[2]。

四是分析结果的不可解释性问题。由于大数据经济模型维数很高，且变量之间可能存在相关性，在对数据进行降维的过程中，会损失大量有效信息，从而使得最终分析结果只具有概率论意义，而很难从经济学理论上加以解释，也就更难以利用传统的计量模型来加以处理了[[[215]](#footnote-215)]。另一方面，很多大数据指标反映的是最终经济学现象，由于市场上发生的一切变化都可能对经济主体情绪和行为产生影响，而这些主体的变化又会反过来导致市场变化，因此很多大数据指标在解释某些经济现象时很难将其他变量的影响剥离开来[[[216]](#footnote-216)]。

五是数据噪声问题。噪声数据是指获取的对研究目的没有价值的信息。某种程度上说，大数据的大容量是一把双刃剑，数据体量大虽克服了抽样误差，但其中难免包含了大量噪音。由于现阶段大量大数据分析方法需要依靠自然语言处理方法、语言识别、图像识别等方法进行分析，中文的语义分析和情感分析相对于英文更加困难，以目前的技术手段，在数据清洗过程中很难将噪音完全消除，这难免会对大数据宏观经济分析的准确程度产生负面影响。比如，国家发展改革委大数据中心曾组织围绕“大众创业万众创新”（简称“双创”）对全国各省份舆论数据进行大数据分析[[[217]](#footnote-217)]，曾发现某省舆论声量非常大。进一步分析才发现，该省省当年度在大规模推动创建全国文明城市和国家卫生城市活动（同样也简称“双创”）；在国家层面分析筛选关键词时，由于分析师没有发现这一问题，最终分析结果中就把大量该省媒体和网民讨论的与“大众创业万众创新”无关的信息也包含进来。

六是大数据分析往往存在难以复制和重复验证的问题。与传统统计数据不同，在互联网时代，经济主体的每一个行为都能够转化为数据被记录下来，因此数据每天都在不断产生和累积，特别是网络数据，几乎是时时更新。这种数据的时效性一方面使得大数据分析方法和分析指数克服了传统方法的时滞性缺陷，但同时也使得基于此构建的大数据指数难以复制。这在一定程度上影响了宏观经济大数据分析的方法与理论被主流经济学界广泛认同。

在这种情况下，过度强调大数据的优势，甚至抛弃原有经济学理论框架而单纯依靠数据分析进行宏观经济分析研判，也会存在很大隐患。在此，课题组还是想引用冯·诺伊曼在《博弈论和经济行为》中所阐述的另一个观点：“用精确的方法处理根本就不明晰的概念和问题是毫无意义的。因此，要想用精确的方法处理经济问题，第一步是通过更细致的描述工作澄清人们对事物的认识。”因此，“在经济学的某些分支里，最有成效的工作也许是耐心的解释描述；事实上，在目前和今后一段时间内，这种工作将占经济学研究的绝大部分。”[[[218]](#footnote-218)]

#### 2.从经济学史的角度思考继承和超越传统经济学理论框架

从目前的研究成果来看，学者们更加倾向于将大数据纳入现有的宏观经济分析框架之中，将基于大数据的各种新型指标作为传统统计指标的完善或补充，并将其纳入传统的宏观经济预测模型。然而，课题组认为，要想在大数据时代实现经济学理论自身“决定性的转折”，仅仅将宏观经济大数据分析作为传统经济学理论框架的补充是远远不够的。从历史视角看，过去数百年间，经济学理论一直沿着均衡范式和演化范式两条路径前进。前者遵循社会物理学范式，视图把经济学建成类似于经典物理一样精密的科学体系，在数据分析方面，形成了一整套基于经济统计的计量经济学体系，并将当前的大数据非传统指标作为统计指标的有效补充，视图对现有的经济学均衡框架进行修补和完善；后者遵循社会生物学范式，则将经济社会系统视为一个动态演化的生态系统，在数据建模分析方面，则将复杂性科学、演化博弈论等现代数学方法和建模工具引入经济分析之中，这一路径下的经济学家同样视图将各种反映经济社会运行细节的大数据引入模型之中，视图提高对宏观经济走向的预测预警能力。

上述两种路径的研究方法尽管视角不同，但相比历史上任何一个时期而言，其所基于的微观数据基础却前所未有地实现了统一。从这个意义上说，经济学经过数百年的理论发展，沿着“均衡—物理学范式”和“演化—生物学范式”两条路径的研究方法体系面临着在大数据时代实现“殊途同归”的可能性。从我们的综述也可以看到，很多原本从属于新古典经济学流派的经济学家也开始尝试运用演化经济学中的主体建模（ABM）[[[219]](#footnote-219)]、复杂网络[[[220]](#footnote-220)]、非线性迭代[[[221]](#footnote-221)]等方法来分析宏观经济走势，尽管这类研究目前尚不多见，但这种学科共同体间相互融合的态势却已经初见端倪了。

当前，在理论层面，应当在继承和超越传统经济学理论框架的基础上，认真思考大数据在融合均衡与演化这两大经济学研究进路的基础上，构建宏观经济大数据分析的特有理论框架。这就要求我们要坚持历史思维，从经济学乃至整个社会科学大的学科发展脉络上思考大数据经济学的历史定位问题。汪丁丁[[[222]](#footnote-222)]曾指出，当前西方社会科学思想正从牛顿的机械论模式向达尔文的演化论模式实现整体迁移。在这一背景下，近几十年来在物理学、数学、地质学、化学等大量学科中，都存在一种学科视角的转变：从将世界视为高度有序的、机械的、可预见的、在某种程度上静态的，转变为将世界视为不断进化的、有机的、不可预测的、处于永远发展中的。虽然现在经济学相比其他学科稍微有点落在后面了，但是它终究是要追随时代精神的[[[223]](#footnote-223)]。因此从大的学科脉络上看，当前大数据在经济学中的大量应用，是这种学科范式大转变的具体体现。一个有意思的佐证，是近年来开展宏观经济大数据监测预测的学者中，大量来自交叉学科领域，如生物学、物理学、计算机科学、人工智能等领域。

从这个意义上说，宏观经济大数据监测预测研究，实际上是始于熊彼特、凡勃伦、哈耶克、沙克尔等经济学大师对经济学中诸如创新、中断、颠覆、在真正意义上的不确定性情况下的决策等问题的讨论在新的技术背景和时代背景下的延续与深化。如布莱恩.阿索所说，这种强调非均衡性和复杂性的观点，在经济学中其实已经有很长的历史了，它们并不是经济学中的新问题。所不同的只不过是，我们现在可以更加严格地对这些主题进行研究。我们拥有了更多、更强大的工具，包括更复杂的概率理论，我们还可以在严格的控制下进行计算机实验。因此，在继续探讨宏观经济大数据分析这一话题时，有必要首先从近代以来经济学大的学科发展脉络入手，对近代以来经济学研究中的数据传统的产生和演化进行一个系统回顾，为我们在理论层面研究大数据经济学的基本框架设定一个历史锚点。

#### 3.从科技哲学的高度重新思考大数据经济学的方法论问题

英国哲学家怀特海曾指出，如果科学不愿退化成一堆杂乱无章的特殊假说的话，就必须以哲学为基础，对自身的基础进行彻底的批判[[[224]](#footnote-224)]。在科学革命或经济学范式的重大变革时期，哲学总是起到关键性的先导作用。面向未来，应当坚持哲学视角，从哲学的角度思考和界定宏观经济大数据分析的本体论、认识论和方法论问题。

过去十几年间，特别是2008年金融危机以来，主流经济学正在面临前所未有的理论挑战。2009年，《经济学人》（The Economist）杂志的一篇文章严厉地指出，华尔街绝不是金融危机的唯一受害者，标准的新古典经济学也是，它已经随着金融的崩溃而崩溃了。布莱恩·阿瑟曾尖锐地批判道：“主流经济学的研究纲领，至少它的超理性版本，已经失败了。如果进行波普尔式（Popperian）的证伪检验，那么2008年的金融崩溃及随后几年世界经济的表现，已经不容置疑地证伪了这个研究纲领。没有人敢说，市场之所以在很短的时间内失去了一半的价值，是因为那些公司突然失去了一半的有用性，但公司一如既往还是那些公司。也没有人敢说，欧洲一些经济体的失业率高达20%，而且仍在上升，是因为劳动者的偏好突然完全改变了，因为人们仍然像以前一样想得到工作。”[[[225]](#footnote-225)]究其原因，则是主流经济学理论模型与现实的脱节性问题，正如人民大学高德步教授曾批判指出的[[[226]](#footnote-226)]：“小阿尔费雷德·马拉伯通过战后颇具影响的经济学家及其活动的研究，对上述问题做出了一个比较客观的评价。他指出：有不少经济学理论是在大学的象牙塔里构思出来的，这样的理论和建议，经常与经济的现实不符，当然不能凑效。现实经济不可能简单得用几条‘抛物线’就能客观描述。这种观点真可谓切中经济学的时弊。这些年来，经济学的发展，越来越‘学术化’，经过精密的数学运算和精心的雕琢，真的成了象牙塔中的‘精品’，遗憾的是许多‘精品’中看不中用。”

出现这种情况的原因很多，从方法论层面，长期以来，社会科学受制于技术方法的局限性，而只能依靠样本统计方法获取数据，导致其解释力和预见性收到根本性限制。19世纪以来，当面临针对大范围、复杂多变问题的研究时，社会科学一直缺乏有效技术手段，只能依赖于采样分析。这在一定程度上实际上是一种方法论的落后性，但社会科学据此发展出了一套完整的方法理论体系，并且被一代又一代的社会科学研究者所接受和习惯，甚至被视为“理所当然的限制”。哈耶克曾总结过所谓全量数据获取的“不可能定理”，他指出：“尽管我们拥有关于人类理智运行原理的知识，但主要事实依然是，我们不可能对导致个人在特定时间做了一件特定的事情之全部具体事实加以说明……以为我们可以掌握全部数据是‘荒谬的’”[[[227]](#footnote-227)]。建构于此的社会科学，不可避免地存在方法论层面的先天不足。正如维克托·迈尔·舍恩伯格指出的，社会科学经典的随机采样方法固然取得了很大的成功，但“这只是一条捷径，是在不可收集和分析全部数据的情况下的选择，它本身存在许多固有的缺陷”。

随着大数据时代的加速到来，万物互联化的大趋势越来越明显。相对乐观者认为，人们将越来越有可能构建出一个可以完全匹配或对应现实世界的数字“孪生”世界。徐晋将这种新场景归纳为“有限的绝对理性”的数据表达，或者说“数据理性”，是指在现代经济背景下，个人理性决策依赖于数据技术，并且个人理性可以通过数据进行模拟与表达。[[[228]](#footnote-228)]从未来视角看，大数据时代的全量数据获取成为现实，将对包括经济学在内的社会科学研究方法产生深远影响，有学者将其视为是经济学研究范式的一场革命，将彻底改变传统的统计学和计量经济学方法，推动经济学研究进入数据驱动范式时代[[[229]](#footnote-229)]。还有学者进一步指出，在大数据时代，经济学不同流派所坚持的结论会得到更好证实或证伪，最终经济学会越来越趋于形成一个统一的流派——数据经济学流派[[[230]](#footnote-230)]。激进者如克里斯·安德森（Chris Anderson）甚至提出“理论的终结”的口号，断言“数据的泛滥导致科学方法论失效了。”[[[231]](#footnote-231)]

当然，当前阶段大数据发展尚存在诸多技术瓶颈问题，过度强调大数据对于社会科学的颠覆性改变并无意义。但从科学哲学和学科范式转变的角度，大数据经济学相比主流经济学而言，在学科研究视角和方法论层面有着巨大差异性。这些差异在相关学者的研究中也被零散提及，比如不关注因果逻辑，而关注相关概率；不采用假设验证的统计思维，而更加重视对异常点、突变点和孤立点的分析；不仅仅关注经济总体体量规模变化，而更加关心细分领域的结构性变化，等等。在未来，确实应当从哲学层面重新思考大数据经济学的本体论、认识论和方法论体系构建问题，从而为实践提供比较完善的理论指引，这也是本课题接下来拟重点研究和解决的问题。

#### 4.在实践意义层面着重思考宏微观一体化的方法体系构建

21世纪首次诺贝尔经济学奖得主斯蒂格利茨教授曾经发表过一篇题为《经济学的又一个世纪》的文章，在文章中他谈到：20世纪的经济学患了“精神分裂症”，微观经济学和宏观经济学的脱节，这种脱节即表现为研究方法上的难以沟通，又反映出二者在意识形态上的分析和对立。据此他认为，21世纪经济学的发展目标应当包括两方面，一是为宏观经济学寻找微观基础，二是微观经济学应当从微观个体的行为推演出总量的含义。出现这种问题的根本原因，人类社会系统是典型的复杂系统，表现在社会成员之间存在差异和非线性的互动关系。个体行为不是独立的，而是相互影响的，社会在整体层面上“涌现”出的行为和目标与社会成员自身行为或目标不同。简单的线性加总不适用于分析和解释这样的社会系统，需要一种可以处理非线性关系的研究方法[[[232]](#footnote-232)]。

而大数据对于经济学理论而言的一个重要价值，就在于其有可能在技术层面真正实现宏观数据微观化和微观数据宏观化。在大数据背景下，经济学研究方法要想实现宏微观一体化的目标，有赖于三个方面的变革：

一是数据采集方法的变革。在大数据时代，通过运用各种高性能并发式数据采集分析技术，人们可以很方便地归集经济社会运行各方面的全样本数据，动辄形成几亿甚至几十亿的分析数据集，从而使得人们有可能采用一种完全不同的方法开展趋势分析和政策评估推演。

二是建模方法的变革。过去，在受限于样本统计数据的背景下，社会科学领域主要的分析量化方法是统计分析，或者说是基于方程的建模方法；近年来，基于“人工适应主体”（artificial adaptive agents）的建模方法[[[233]](#footnote-233)]或者叫做“基于主体建模”（Agent-based Modeling，ABM）的方法[[[234]](#footnote-234)]越来越受到研究者的关注。2009年诺贝尔奖得主，政治学家埃莉诺·奥斯特罗姆教授在2006年曾专门著文讨论该方法[[[235]](#footnote-235)]，并与其他研究者合作建立了一个基于主体建模方法的关于公共池塘的分析模型[[[236]](#footnote-236)]。基于主体建模是一种基于计算机的建模方法。它通过计算机模拟出一个虚拟的人类社会，并通过设置各种社会变量为社会过程建模，从而提供了一种解释社会科学问题的动态分析途径[[[237]](#footnote-237)]。

三是数学方法的发展。近几十年来，数学方法上最大的变革源自计算和计算机科学的崛起。格里高利·蔡廷（Gregory Chaitin）曾经指出，数学正在从连续的公式、微分方程、静态的结果转向离散的公式、组合推理及算法思维。他说：“计算机不仅是一种极其有用的技术，还是一种具有革命性意义的新数学，它带来了深刻的哲学后果，它揭示了一个新世界的面纱。”这对于经济学研究而言意义十分重大，人们可以基于采集的海量数据，通过构建个体、企业、事件、商品等主体模型的基础上，运用离散数学和计算数学的方法进行模式识别和规律总结。正如汪丁丁指出的：凡是可预期的行为，必定遵循某些行为模式（pattern）。现代数学的核心问题是模式识别（pattern recognition）问题——数学不再是关于数字的学问，那是古代的定义；数学也不是关于微积分的学问，它也不是关于抽象代数或拓扑或你指出的任何数学分支的学问；数学是模式识别的学问，所以数学主要是一种直觉[[[238]](#footnote-238)]。

与上述三个层面变革相对应，近年来国内外经济学界出现了诸如计算经济学（Agent-based Computational Economics，ACE）[[[239]](#footnote-239)]、复杂经济学[[[240]](#footnote-240)]等新提法，从本质上讲，这些提法与大数据经济学在理论和方法层面均是相通的。落地到政策层面而言，大数据经济学也好，复杂经济学或计算经济学也罢，都是为了规避传统主流经济学针对同质的理性主体开展建模分析后的政策偏差和滞后性问题，毕竟现实中宏观调控政策能否收到预期效果，不仅取决于市场基本运行状况和相关数据，更重要的决定性因素是微观主体行为的本质特点及响应方式[[[241]](#footnote-241)]。本研究落脚点，就是要探讨在方法论层面，如何进一步发挥大数据在宏微观一体化分析方面的技术优势，构建更加科学、完善的宏观经济分析研判方法体系，探索将宏观经济调控与微观经济运行有机地连接起来的实现路径[[[242]](#footnote-242)][[[243]](#footnote-243)]。

## 三、研究目标和研究内容

### （一）研究目标

本课题将以利用机器学习等大数据算法支撑宏观经济监测预测研究为主线。具体研究目标包括：**一方面要构建理论框架**，方法融合与学科交叉是未来科学研究的大趋势，本研究将充分吸收国内外经济预测的理论基础和模型方法精髓，构建利用机器学习、深度学习技术支撑宏观经济监测预测研究方法理论框架。**另一方面要聚焦成果应用**，鉴于预测传统统计指标对实操性要求较高，本研究将利用开源软件搭建实验平台，通过遵循行动研究范式，在推进过程中与有关专家充分互动，切实提升预测方法的科学性和有效性。

本课题将致力于突破三方面创新**：一是研究方法论的整合创新。**本研究将以客观数据和大数据技术为突破口，通过实证检验和方法改进，尝试突破现有方法边界，力争弥合传统经济指标预测与大数据方法之间的缝隙。**二是实践环境中的应用创新。**立足于实际问题，搭建利用机器学习、深度学习技术支撑宏观经济监测预测的实验平台，力求在实践环境中不断迭代优化应用方法创新。**三是研究成果上的形式创新。**除产出研究报告等“软成果”外，本研究拟沉淀数据实验平台、算法模型库、预测指标库等多项“硬成果”，力图为后续其他相关研究提供支撑参考。

### （二）研究内容

课题组积极探索利用机器学习、深度学习等大数据算法开展宏观经济监测预测研究，尝试利用新技术应对宏观调控工作面临的新机遇、新模式、新挑战。主要研究内容如图1所示。



图1 利用机器学习等大数据算法支撑宏观经济监测预测研究内容

具体包括：**一是开展基于传统时序预测方法的宏观经济指标预测研究**，依托国家信息中心大数据发展部所掌握的覆盖全国范围1.1亿家企业和个体工商户工商注册、就业招聘、招投标等数据的企业生产经营行为大数据平台，利用传统ARIMA和Prophet模型开展了单变量时序预测实验，利用传统LASSO和BSTS模型开展了多变量时序预测实验，分析单变量预测或多变量预测情况下，传统时序预测模型是否存在较为明显的局限性。**二是开展基于机器学习方法的宏观经济指标分类预测研究**，尝试通过NaiveBayse、LibSVM、AdaBoostM1、J48、RandomForest等机器学习分类算法判断未来预测值的性质，从国家统计局官方网站获取月度统计数据并整理构成监测指标库，用以作为宏观经济指标分类预测实验的预测目标。之后，从企业生产经营行为大数据平台提取出抽取多维数据特征，用作机器学习分类预测的数据特征库并开展宏观经济指标分类预测数据实验。**三是开展基于机器学习方法的宏观经济指标数值预测研究**，针对大部分宏观经济指标历史数据较为缺乏、可获取的宏观经济指标影响因素数据较多等特点，尝试应用可兼容自变量多、因变量少的机器学习预测方法开展宏观经济指标数值预测研究，拟融合LinearRegression、ElasticNet、RandomForest、GBRT等机器学习数值预测算法与多重数据特征开展研究。**四是开展基于可解释性深度学习方法的宏观经济数值预测研究**，拟选用了全球目前较为先进的N-BEATS深度学习模型开展宏观经济数值预测研究，通过对N-BEATS的模型原理、技术体系及平台架构进行研究剖析，搭建基于N-BEATS深度学习技术的宏观经济监测预测实验平台，根据数据实验结果持续对N-BEATS模型进行参数调优，开展与目前较为流行的BVAR等传统时序预测模型对比实验。

### （三）研究框架

本课题研究框架可概括为：**“1+2+3”，即一项重点工作、二维数据输入、三种有效资产**。其中，**一项重点工作**是指搭建利用机器学习、深度学习技术支撑宏观经济监测预测的实验平台，实现多源异构数据清洗去噪、特征提取及特征选择等功能，根据预测目标优化调整时序数据预测算法，并通过所生成的预测模型开展传统统计指标自动化预测；**二维数据输入**是指根据实际情况利用目前国家信息中心大数据发展部已拥有的大数据资源（维度之一）及国内外可月度更新的统计数据（维度之二）；**三种沉淀资产**是指以“万物皆数”为指导思想，不断沉淀经济监测预测各环节的过程性数据，利用技术手段加深数据提取的广度和深度，积极沉淀数据实验平台、算法模型库、预测指标库三方面有效资产。

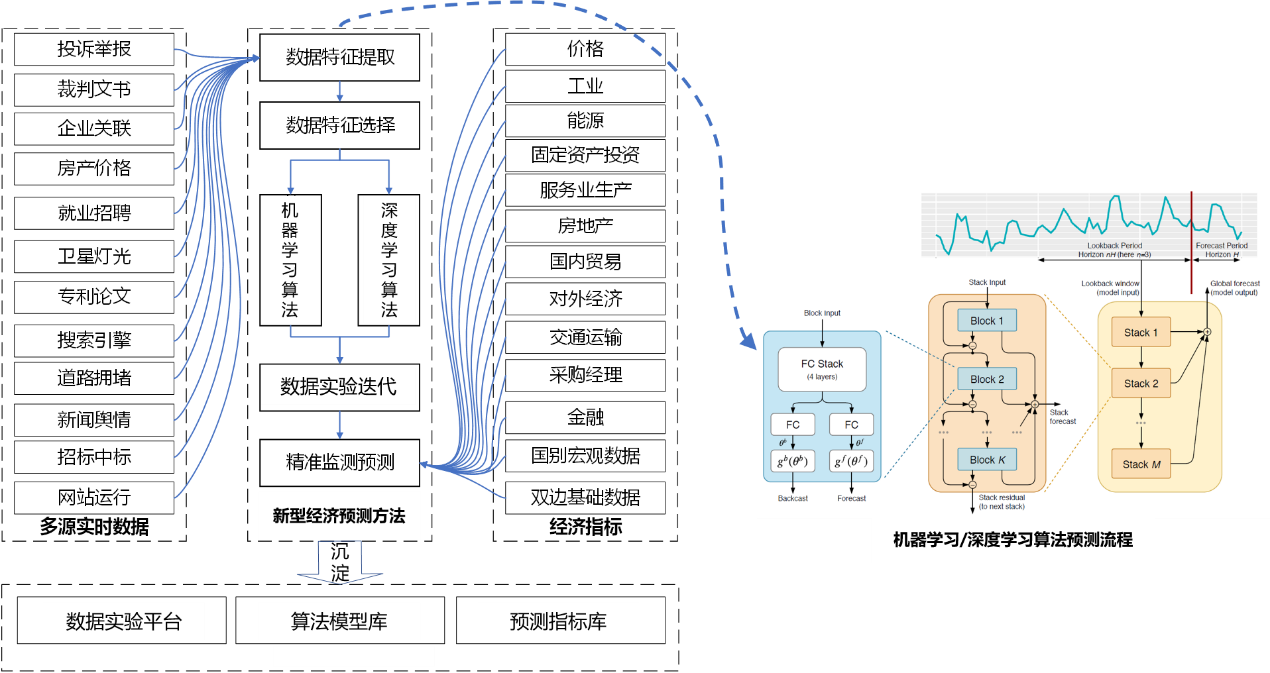


图2 利用机器学习等大数据算法支撑宏观经济监测预测研究框架

## 四、研究内容之一：基于传统时序预测方法的宏观经济指标预测研究

### （一）研究思路

课题组基于中心数据资源开发了Processor、Matcher和Predictor三个模块分别用于数据预处理、指标匹配和自动化预测。三个模块的工作流程如图3所示。本章主要简要介绍Processor和Matcher的工作原理、Predictor中四种典型的基于时间序列自动化预测方法和对实验中的发现进行总结。

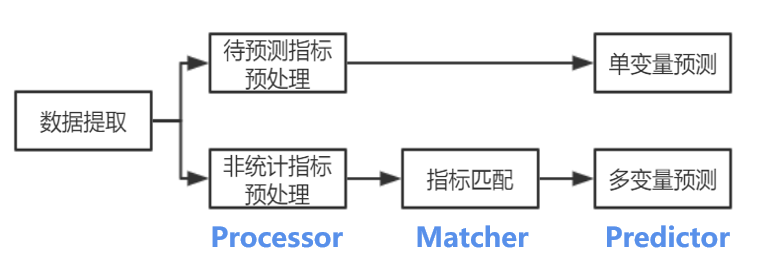


图3 基于时序的自动化预测流程

#### 1.数据预处理

通过对所提取数据的探索，课题组发现数据中存在以下几个问题：（1）时间序列的起止时间各异；（2）不同指标的年内期数不同，例如多数月度指标为12期，固定资产投资等指标为11期等；（3）部分指标缺失严重；（4）缺失值较少的指标中缺失的模式也存在差异。

针对以上问题，课题组开发了相应的子模块：（1）将指标按照起止时间进行自动分类；（2）将指标按照年内月份分布进行自动分类；（3）设置了可调整的阈值，可以自动剔除缺失值比例高于阈值的指标；（4）对缺失值较少的指标进行自动化插补。

完成以上预处理工作后，数据已基本可用。但数据库中导出的数据一般是长数据（如图4所示），不利于输入模型，因此需要将其变成宽数据并进行多维交叉。

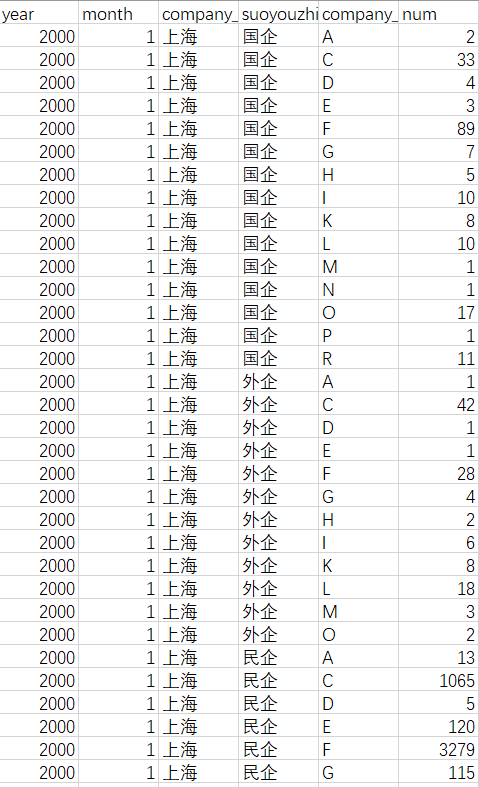


图4 原始数据示例

以图4中所示数据为例，其中包含了省份、所有制、行业三个类别属性，因此一共有7种交叉，即{省份}{所有制}{行业}{省份，所有制}{省份，行业}{所有制，行业}{省份，所有制，行业}。交叉后的宽数据如图5所示，图中“安徽\_A”表示安徽省A行业的数据，“安徽\_国企\_A”表示安徽省A行业国企的数据。

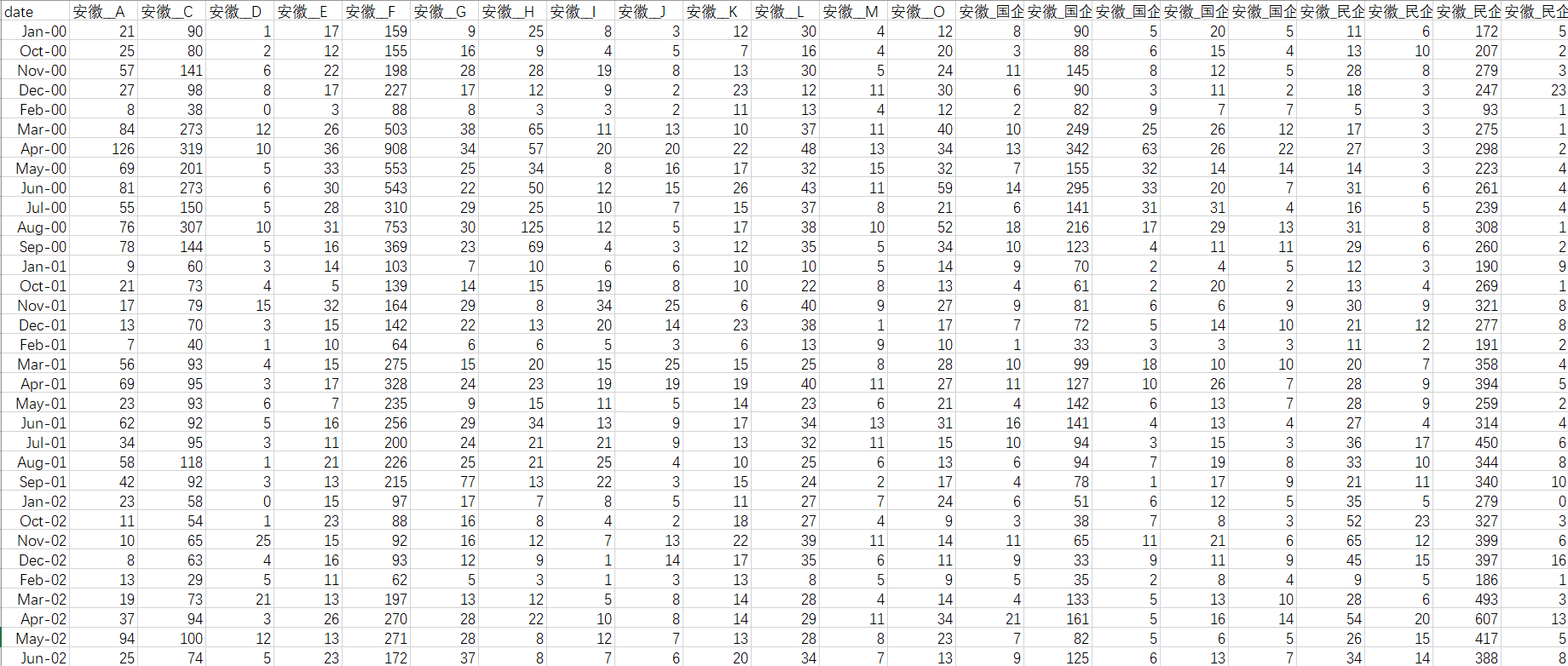


图5 处理后的数据示例

此外，课题组还在Processor中加入了数据变换功能，可以对数据进行求增速、幂、移动平均值等操作。

#### 2.数据匹配

完成以上工作后，即可进行单变量指标预测，而在多变量预测中，提前筛选出自变量有利于提高预测模型的针对性和运行效率，因此课题组开发了Matcher模块，用以自动化筛选出与待预测指标高度相关的外部指标。

Matcher模块的核心是计算时间序列间的相似度，课题组在综合Pearson、Spearman、Kendall等相关系数计算方法的基础上，设计了一套高精度的匹配算法，所得结果如图6所示。x和y分别表示匹配出的自变量和因变量，delay表示x相对于y的提前期数，后4列是相关系数和一元回归的结果。



图6 数据匹配结果示例

接下来，本章将以非制造业采购经理指数中的从业人员指数（以下简称“从业人员指数”）为例，介绍Predictor模块中四种典型的基于时间序列自动化预测方法。

### （二）基于ARIMA的单变量预测

#### 1.ARIMA模型

ARIMA是经典的时间序列模型，广泛应用于对各类时间序列。它的基本思想是：待预测的时间序列是由某个随机过程生成的，如果生成序列的随机过程不随时间变化，则该随机过程的结构可以被确切地刻画和描述，因此可以利用序列过去的观察值，来外推出序列的未来值。在ARIMA模型中，序列的未来值被表示成滞后项和随机干扰项的当期及滞后期的线性函数，即模型的一般形式如公式4.1所示：

 式（4.1）

ARIMA模型具有扎实的理论基础在建模时充分考虑历史数据和序列自身的规律，具有建模方法简单、操作步骤规范等优点，非常适合对单变量时间序列进行短期预测，并在一些领域取得了较好的预测效果。

ARIMA模型的建模过程可以分为以下四个步骤：

（1）时间序列的平稳性检验。通常采用ADF或PP检验方法，对原始序列进行单位根检验。如果序列不满足平稳性条件，可以通过差分变换或者对数差分变换，将非平稳时间序列转化为平稳时间序列，然后对平稳时间序列构建模型。

（2）确定模型的阶数。通过借助一些能够描述序列特征的统计量，如自相关系数和偏自相关系数，初步识别模型的可能形式，然后根据AIC等定阶准则，从可供选择的模型中选择一个最佳模型。

（3）参数估计与诊断检验。包括检验模型参数的显著性，模型本身的有效性以及检验残差序列是否为白噪声序列。如果模型通过检验，则模型设定基本正确，否则，必须重新确定模型的形式，并诊断检验，直至得到设定正确的模型形式。

（4）用建立的ARIMA模型进行预测。

#### 2.数据实验

ARIMA模型需要确定三个参数，即自回归阶数（p）、差分阶数（d）和移动平均阶数（q）。以从业人员指数为例，课题组首先采用传统方法，以人工的方式进行了参数确定。

该指数如图7所示，初步判断其具有季节性，趋势性较弱，可能符合平稳性要求。

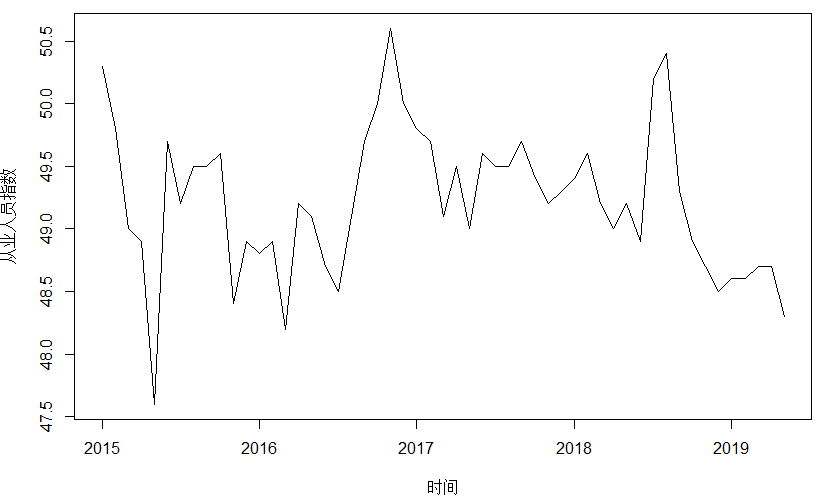


图7 从业人员指数

表1中的ADF检验结果显示，三类ADF检验中，有一种检验（drift）在5%的水平下是显著的，说明可以认为从业人员指数为平稳时间序列。虽然三个检验中只要有一个通过即可认为时间序列是平稳的，但三个结果的不一致性也说明了平稳性判断的难度。在许多经典的时间序列分析方法中，平稳性常常是重要前提，需要深入研究。而在最新的时间序列模型中，如下文将实验的Prophet和BSTS，都不再将时间序列的平稳性作为重要前提，因此此处不再对从业人员指数的平稳性进行深究，接受该序列平稳的结论，将ARIMA模型的参数d确定为0。

表1 平稳性检验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 检验模式 | t值 | 1%临界值 | 5%临界值 | 10%临界值 | 结论 |
| trend | -3.23 | -4.04 | -3.45 | -3.15 | 平稳 |
| drift | -3.27 | -3.51 | -2.89 | -2.58 |
| none | -0.52 | -2.6 | -1.95 | -1.61 |

因为从业人员指数是平稳时间序列，因此可直接通过偏自相关系数（PACF）和自相关系数（ACF）来确定ARIMA模型的p和q两个参数。从下图不难看出，PACF和ACF分别在滞后1期和滞后2期出现截尾，因此可使用p=1和q=2作为ARIMA模型的参数。

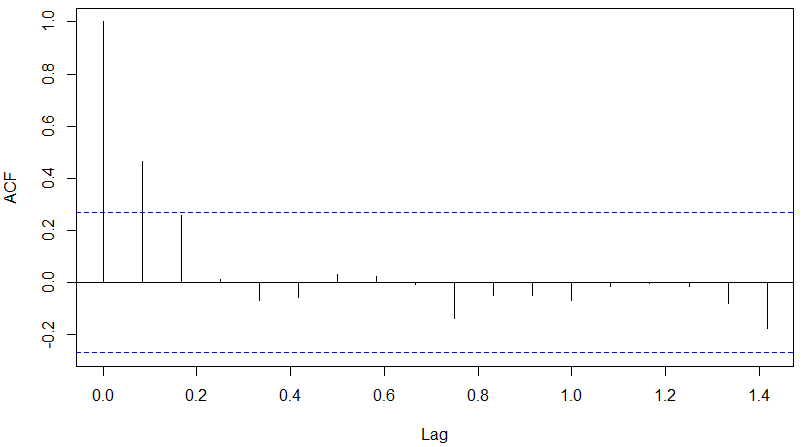
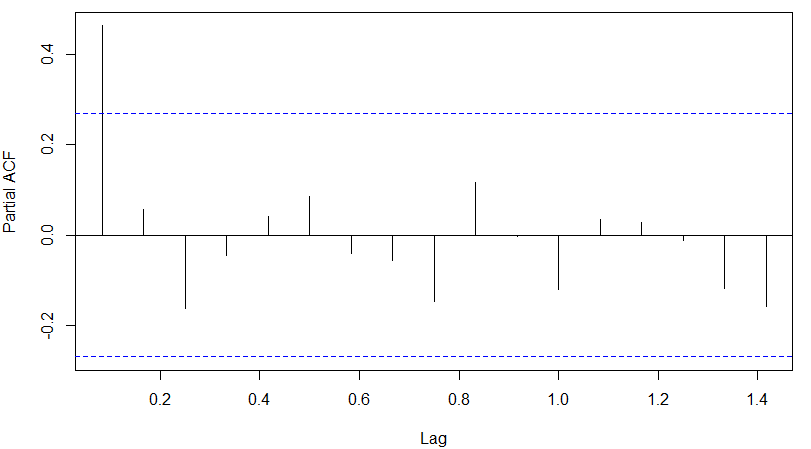


图8 PACF和ACF图

ARIMA(p=1,d=0,q=2)的估计值与检验结果如下表所示，从表2中不难看出，除截距项以外的其它系数均不显著。

表2 人工调参的ARIMA模型检验

|  | Estimate | Std.Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| ar1 | 0.3554 | 0.2854 | 1.2453 | 0.213 |
| ma1 | 0.123 | 0.2775 | 0.4432 | 0.6576 |
| ma2 | 0.2378 | 0.1543 | 1.5414 | 0.1232 |
| intercept | 49.2251 | 0.1423 | 345.8677 | 0 |
| sigma^2 estimated as 0.2497531  log likelihood = -38.64263 aic = 87.28525 | | | | |

从以上实验结果不难发现，ARIMA模型的参数确定是十分困难的，在实践中常常需要针对每个具体的时间序列进行深入分析，并通过经验积累提高判断的准确性。为了解决这一难题，有研究者提出了自动化计算ARIMA模型参数的方法，在实践中，这种方法常常具有较好的拟合优度，但是有时候会加大模型的解释难度。

课题组也用自动化方法进行了实验，实验结果如表3所示。自动化方法计算出的p和d与人工方法的结果相同，但自动化方法发现q=0时模型的拟合效果更好。从表中可以看出，所有系数均显著，同时模型的AIC也有所下降。

表3 自动计算的ARIMA模型检验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | t value | Pr(>|t|) |
| ar1 | 0.5103 | 0.1246 | 4.096 | 0 |
| intercept | 49.2238 | 0.14 | 351.628 | 0 |
| sigma^2 estimated as 0.2690088  log likelihood = -39.5404 aic = 85.0808 | | | | |

自动化模型的预测结果如下图所示。从图中可以看出，ARIMA模型的拟合值在形态上与实际值十分接近，但拟合值全部在50以下，而实际值则有多个点在50以上。预测方面，预测值整体略高于实际值，前三个月的预测值与实际结果较为接近，之后则出现了较大的偏差。

图9 自动化ARIMA模型结果

### （三）基于Prophet的单变量预测

以ARIMA为代表的一系列经典模型，都要求使用者对其所分析的时间序列有较深的理解，因此需要消耗较多的时间和人力。对于规律明显、更新频度较低的少数时间序列的分析，这并不成问题，但是如果需要分析的时间序列较多，尤其是其中有大量不平稳序列时，使用经典模型常常受到较多制约。

近年来，随着数据的不断丰富和算力的提升，研究者提出了一批分析方法，用于弥补经典时间序列模型的不足。这类模型最主要的特点是对数据要求更少、自动化程度更高，更多地通过迭代、交叉验证等方法寻找参数的最优解。接下来，本文将介绍其中的三个典型代表，并用其进行实验。

#### 1.Prophet模型

Prophet模型是2017年由Facebook开源的一项技术。不同于传统的时间序列预测模型，Prophet模型在实质上是对时间序列数据的曲线拟合，同时对数据中的节日效应和趋势变化点具有出色的适应能力，尤其对缺失值、趋势的转变和大量的异常值的鲁棒性极强，目前主要应用于Facebook社交网站的流量预测。

Prophet本身是一个基于自加性模型的预测时间序列数据的模型，模型整体由growth（趋势项）、seasonality（周期项）、holidays（节日项）３部分叠加组成，基本形式如下：

 式（4.2）

其中，趋势项是整个Prophet模型的核心组件，包含不同程度的假设和调节光滑度的参数，用于拟合时间序列中的非周期性变化，从数据中选择突变点检测趋势走向。基本趋势项是一个逻辑函数：

 式（4.3）

其中， 表示模型容量，表示增长率，代表偏移量，随着的增加，趋于。

是周期项，它用傅里叶级数近似表达周期性分量，具体表达式如下：

 式（4.4）

其中，表示某个固定的周期，表示期望在模型中使用周期的个数。

是节日项，将每个节假日在不同时刻下的影响视作独立模型，为每个独立模型设置了一个虚拟变量。节假日模型可表示为：

 式（4.5）

其中，表示窗口期中的节假日对预测值的影响；表示第个虚拟变量，若时间变量属于虚拟变量，则虚拟变量值为１，否则为０；表示节假日，表示窗口期中包含的时间；是误差项并且服从正态分布，表示模型未预测到的波动。

#### 2.数据实验

在数据实验中，课题组发现，相对于ARIMA模型，Prophet模型既有更高的自由度，又有更高的自动化程度。

更高的自由度是指Prophet模型有更多的参数可以选择，这一点最集中地体现在了对突变点（changepoint）的设置上。从ARIMA的拟合结果中，不难发现，由于使用了自回归和移动平均，拟合结果的方差明显小于实际值，实际值中的突变点（例如两个峰值）在拟合值中表现的并不明显，这可能导致无法预测出未来的突变点，错过提前预警的机会。而在Prophet模型中有一系列参数可以用来解决这一问题。

更高的自动化程度是指，虽然Prophet有更多的参数，但是所有参数都无需手动调节，包括突变点参数在内的核心参数都可以由模型计算得出。课题组首先尝试了在不进行任何调参的情况下使用该模型，结果如下图所示。

图10 自动化Prophet模型结果

从图中可以看出，在不进行任何调参的情况下，Prophet模型的拟合和预测效果都优于ARIMA模型。拟合效果方面，Prophet模型更好地拟合了实际值的波动，其方差更接近于实际值，尤其是在各个突变点上，Prophet模型的优势表现得更加突出。预测效果方面，Prophet模型虽然也高估了实际值，但是其预测出了后两个月的下降过程，整体表现更好。

课题组还尝试了人工调整模型的参数，其中重点调整了突变点的相关参数。通过大量的实验发现，人工参与虽然能使结果有一定的提升，但是效果十分有限。下图是经过反复尝试之后得到的最优结果，可以看出虽然历史数据的拟合效果有了一定的提升，但是预测效果变化很小，甚至有可能出现了过拟合。通过查阅资料发现，当突变点过多或过少时，人工调参会更有价值，一般情况下使用模型自动算出的结果即可。由此可见，Prophet模型具有较好的自动化预测效果。

图11 人工调参的Prophet模型结果

### （四）基于LASSO的多变量预测

之前两节的实验表明，即使只有待预测的时间序列这一个变量，现有模型也能够进行自动化的短期预测，但是单变量时间序列预测有着明显的局限。因为其本质上是外推法在时间序列领域的应用，而外推法要求已知信息中包含尽可能多的规律，即时间序列的趋势性、季节性、周期性和特殊事件规律等都必须尽可能完整地包含在历史数据中。这就对数据提出了较高的要求，一方面要求数据有较强的规律性且规律应保持不变，另一方面要求历史数据足够长，以包含尽可能多的规律。

在实际的预测任务中，很多待预测指标都无法满足以上两个条件，此时引入更多的变量提高预测效果就显得十分必要。接下来两节，课题组将介绍两种能够用于自动化预测方法，首先是Tibshirani于1996年提出、目前已经相对成熟的LASSO模型，然后是Scott和Varian于2014年提出的贝叶斯结构时间序列模型（BSTS）。

#### 1.LASSO模型

LASSO（Least Absolute Shrinkage and Selection Operator）是一种系数压缩、变量选择的方法，即在回归方程中加入惩罚项使一些影响较小或没有影响的自变量系数趋近于零，实现对模型的系数压缩和变量选择。

一般线性回归模型为

 式（4.6）

使用最小二乘估计得到最小化模型的目标函数为

 式（4.7）

其中， 是数据矩阵，而是由标签组成的列向量。 对于线性回归，最优解为

 式（4.8）

但是，当样本的数目比每个样本的维度还要小时，矩阵就不是满秩的，会变得不可逆，会有无穷多个解，这样的数据容易过拟合。

解决过拟合最常用的方法是正则化，如岭回归添加一个正则项

 式（4.9）

直观地来看，添加正则项会使模型的解偏向于二范数值较小的。最小化式（4.9）等价于如下问题

 式（4.10）

其中，为常数，与对应，其闭式解的形式如下

 式（4.11）

由以上的推导可知，岭回归能够由的大小决定在某一程度上通过限制条件，实现对模型空间的限制，从而避免过拟合。但岭回归不具有产生稀疏解的能力，得到的系数仍然需要数据中的所有特征才能计算预测结果。

引入稀疏性的方法是用正则项代替，得到如下目标函数

 式（4.12）

最小化式（4.12）等价于如下问题

 式（4.13）

该问题称为 LASSO 估计，其闭式解的形式如下

 式（4.14）

由式（4.14）知，LASSO将OLS得到的系数的绝对值进行衰减，如果小于零，就将其变为0，即将系数稀疏化。

#### 2.数据实验

课题组首先在不进行参数调整的情况下，使用LASSO模型进行了自动化预测，使用的自变量是与从业人员指数提前五期高度相关的172个外部变量，预测结果如下图所示。从图中可以看出，在拟合部分，LASSO相较于Prophet有更好的表现，但是在加入自变之后，最近5个月的预测效果反而不如Prophet模型。由此，课题组推断这可能由于以下两种原因：（1）模型出现了过拟合，导致历史数据拟合很好但预测效果不佳；（2）引入的外部数据近几个月出现了系统性变化，对预测结果产生了影响。

图12 自动化LASSO模型结果

通过模型拆解和数据回溯，课题组发现以上两个问题都存在。一方面，默认值的LASSO模型虽然对172个自变量进行压缩，但是最后仍保留了33个，因为因变量训练数据只有53期，模型很可能出现了过拟合。但更重要的是，由于本次实验所用的外部数据范围较小，此次匹配得到的172个变量均出自企业数据库，但是受2019年2月春节影响，企业数据库2月的数据出现了系统性大幅变动，具体如下图所示。由于自变量提前5期，且多数自变量的回归系数为负数，所以导致预测值7月前后的结果出现较大偏差。

图13 自动化LASSO模型的自变量

对于以上问题，在传统分析中，需要人工挑选变量，并加入代表新年这一特殊事件影响的控制变量，通过反复调整以提高预测效果。由于本研究的主题是自动化预测，因此课题组更倾向于增加对模型参数的普适性限制，而非利用先验知识改变模型内容。但在反复实验后，课题组发现，由于LASSO模型并不是专门为解决时间序列问题而开发，并没有解决特殊事件影响的参数，因此参数调整的效果并不理想。

### （五）基于BSTS的多变量预测

#### 1.BSTS模型

为解决“即时预测”模型遇到的“维数灾难”、“伪回归”等问题，Scott和Varian在Harvey提出的结构时间序列模型的基础上构建了贝叶斯结构时间序列模型，用来解决大数据背景下的经济时间序列“即时预测”问题。一方面，该模型通过引入结构时间序列模型克服了平稳序列的限制，此外钉板回归的应用使得该模型可以相对稳定和自动化的对解释变量进行选择。结构时间序列模型是时间序列数据的状态空间模型，它可由如下两个方程来定义：

 式（4.15）

 式（4.16）

等式(4.15)是观测方程，它将观测数据与潜在的维状态向量联系起来。等式(4.16)是状态转移方程，它定义了潜在状态如何随时间变化。上述两式中是一个维输出向量，是一个的转移矩阵，是的控制矩阵，是标准化的观测误差，其噪音变量是， 是一个维的系统误差项，包含一个状态扩散矩阵，且。

状态空间模型模块化的特性给建模者提供了极大的灵活性，让他们可以根据需要选择模型的趋势项、季节项、回归项和潜在的其他状态成分来进行建模。通过在“基本结构模型”中加入一个回归项，上述模型可写成如下形式：

 式(4.17)

式（4.17）中为高斯随机噪声的独立成分。为水平趋势项，为“斜率”趋势项，为季节项。模型的待估参数为和的方差。

模型的求解过程基于三种贝叶斯技术：卡尔曼滤波、钉板回归和贝叶斯模型平均。其中卡尔曼滤波用来将铁路旅客发送量时间序列进行分解，得到模型的趋势项和季节项；钉板回归用于解释变量的筛选，其求解基于马尔科夫链蒙特卡洛采样算法；贝叶斯模型平均是一种集成方法，用于提高模型的预测精度。

#### 2.数据实验

BSTS模型自动化程度较高，几乎所有参数都是通过迭代求解获得。但是相较于LASSO等通用模型，BSTS作为时间序列专用预测模型，其允许使用者在其中加入趋势、季节项等时间序列特征作为变量，同时这些特征由模型自动计算得到。课题组在模型中设定了加入趋势项和季节项，并将匹配得到的172个指标作为备选自变量加入模型，三者的拟合结果如图14所示。从图中可以看出趋势项较好地反映了原始数据的总体变化，季节项也符合数据年内变化规律，回归项则对突变点有较好地体现。

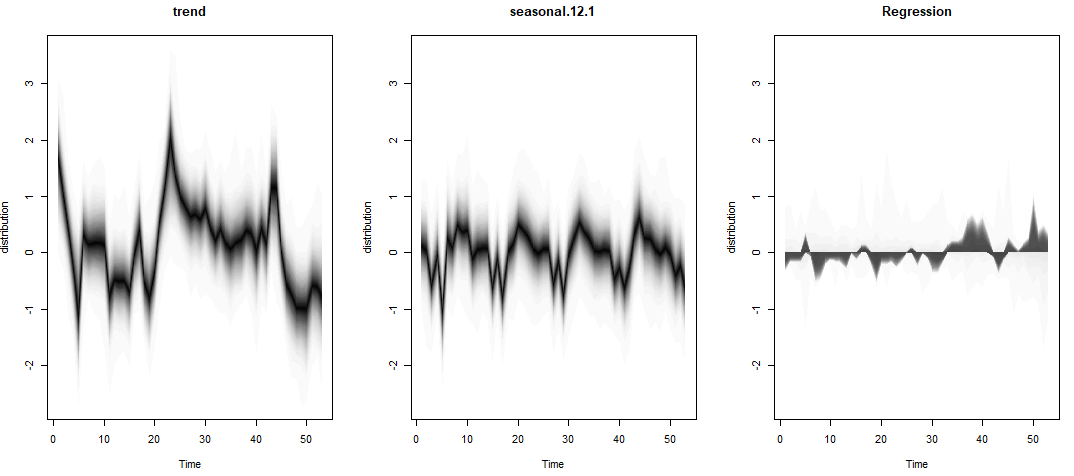


图14 BSTS模型分解结果

从最终结果来看，BSTS模型表现整体优于LASSO模型，尤其是在预测部分。在预测部分，虽然加入了受到春节影响而出现系统性波动的外部变量，但由于同时考虑了季节性和趋势性，基于春节前后数据的预测结果并未出现较大偏差，短期波动模式和预测值都与实际情况较为接近，误差小于LASSO结果模型。由此可见，BSTS模型具有较好的自动化预测效果。

图15 BSTS模型结果展示

### （六）实验结论

通过对现有的时间序列预测模型的调研与实验，课题组初步得出如下结论：

（1）对于时间序列预测，目前已有一些经过检验的可用模型，这些模型可以作为基础方法纳入自动化预测的模型库中，由系统根据数据情况和任务需求调用。除了上文中提到的四种模型，课题组已将ARCH、GARCH、岭回归、ElasticNet、SVR等主流模型加入模型库。

（2）无论单变量预测还是基于外部数据的多变量预测，都有明显的局限性。在实际使用过程中，可以综合使用多种不同的模型，通过集成学习的方法提高预测精度。

（3）基于外部数据的多变量预测对外部数据质量十分敏感。如果外部数据出现系统性偏差，可能对结果产生严重影响。对于这一问题，除了综合考虑多种因素外，还可以通过增加数据检测模块、引入不同来源的数据等方法减少外部数据本身对预测结果产生的影响。

（4）现有模型对指标的定性预测关注不足。在以上四类典型模型的介绍中不难发现，它们本质上都是以趋势、季节、特殊事件以及数值间的距离为中心的定量预测，而在评估预测结果时，最常用的标准则是数值误差的大小。但在实际预测中，波动模式的预测相对简单，而精准的数值预测难度较大，常常出现这样的情况：时间序列上各个点之间的相对位置得到了准确预测，但是预测出的数值却不够精准，例如预测结果与实际值可能是两条平行的折线。对于不存在参照阈值（如荣枯线）的指标，这种结果已经具有一定的参考价值。但是对PMI这一类存在参照阈值的指标，如果通过预测值无法判断实际值未来是高于还是低于荣枯线，那么预测结果的价值将大打折扣。因此，课题组将研究的重心转向了定性预测，即通过机器学习分类算法判断未来值的性质。

## 五、研究内容之二：基于机器学习方法的宏观经济指标分类预测研究

### （一）研究思路

围绕基于机器学习方法的宏观经济指标分类预测研究，课题组提出了一种完全数据驱动化的经济监测预测方法（简称“三库碰撞”），其总体思路如图16所示，主要思路是既要减少人工假设和人为经验的控制程度，也要从多源细粒度数据中抽取大量的数据特征，不断尝试数据特征的组合方式，与监测预测目标结果进行自动化碰撞，并根据监测预测效果不断优化调整数据特征及相应算法，最终由机器自动化选择一种最优这组合从而达到经济精准监测预测效果。



图16 “三库碰撞”总体思路

### （二）构建“三库”

目前，“三库碰撞”即利用高频数据和机器学习算法预测统计指标大致分为三个准备阶段，分别是构建算法模型库、构建监测指标库和构建数据特征库。

1.构建算法模型库

常用的算法模型主要分为回归算法、正则化算法、基于实例和记忆算法、降维算法、深度学习算法、关联规则算法、集成学习算法、贝叶斯算法、决策树算法、聚类算法、神经网络算法等，具体详见表4。

表4 算法模型库

|  |  |
| --- | --- |
| 算法分类 | 算法名称 |
| 回归算法 | 普通最小二乘回归（OLSR） |
| 线性回归 |
| Logistic回归 |
| 逐步回归 |
| 多变量自适应回归样条曲线（MARS） |
| 局部估计散射平滑（LOESS） |
| 折刀回归 |
| 正则化算法 | 岭回归 |
| 最少的绝对收缩和选择算子（LASSO） |
| 弹性网 |
| 最小角度回归（LARS） |
| 基于实例和记忆算法 | k-最近邻（kNN） |
| 学习矢量量化（LVQ） |
| 自组织映射（SOM） |
| 本地加权学习（LWL） |
| 降维算法 | 主成分分析（PCA） |
| 主成分回归（PCR） |
| 偏最小二乘回归（PLSR） |
| Sammon映射 |
| 多维度缩放（MDS） |
| 投影追求 |
| 判别分析（LDA，MDA，QDA，FDA） |
| 深度学习算法 | 深度玻尔兹曼机器（DBM） |
| 深信仰网络（DBN） |
| 卷积神经网络（CNN） |
| 堆叠的自动编码器 |
| RNN |
| 关联规则算法 | Apriori |
| Eclat |
| FP-Growth |
| 集成学习算法 | Logit Boost（Boosting） |
| 自举聚合（Bagging） |
| AdaBoost |
| 堆叠泛化（混合） |
| 梯度增压机（GBM） |
| 梯度增强回归树（GBRT） |
| 随机森林 |
| 贝叶斯算法 | 朴素贝叶斯 |
| 高斯朴素贝叶斯 |
| 多项式朴素贝叶斯 |
| 平均一依赖估计量（AODE） |
| 贝叶斯信仰网络（BBN） |
| 贝叶斯网络（BN） |
| 隐马尔可夫模型 |
| 条件随机字段（CRF） |
| 决策树算法 | 分类和回归树（CART） |
| 迭代Dickotomiser 3（ID3） |
| C4.5和C5.0 |
| 卡方自动交互检测（CHAID） |
| 决策树桩 |
| M5 |
| 有条件的决策树 |
| 聚类算法 | 单连接群集 |
| K-均值 |
| K-中位数 |
| 预期最大化（EM） |
| 分层聚类 |
| 模糊聚类 |
| DBSCAN |
| OPTICS算法 |
| 非负矩阵分解 |
| 潜在狄利克雷分配（LDA） |
| 神经网络算法 | 自组织映射 |
| 感知 |
| 径向基函数网络（RBFN） |
| 反向传播 |
| 自动编码 |
| Hopfield网络 |
| 玻尔兹曼机器 |
| 限制玻尔兹曼机器 |
| Spiking神经网络 |
| 学习矢量量化（LVQ） |
| 其它算法 | 支持向量机（SVM） |
| 进化算法 |
| 归纳逻辑编程（ILP） |
| 强化学习 |
| ANOVA |
| 信息模糊Netowkr（干扰素） |

#### 2.构建监测指标库

为便于开展“三库碰撞”有关实验，课题组从国家统计局官方网站获取月度统计数据并整理构成监测指标库，如表5共11类489个指标84034条数据，用以作为“三库碰撞”实验的预测目标。

表5 监测指标库

|  |  |  |  |
| --- | --- | --- | --- |
| 类目名称 | 下属指标数 | 时间跨度 | 总条数 |
| 财政 | 6 | 2000年1月-2019年4月 | 1320 |
| 采购经理指数 | 23 | 2005年1月-2019年5月 | 3739 |
| 房地产 | 84 | 2000年2月-2019年5月 | 17786 |
| 工业（增加值、价格指数、主要产品产量） | 128 | 2000年1月-2019年5月 | 22424 |
| 固定资产投资 | 20 | 2012年2月-2019年5月 | 1620 |
| 国内贸易 | 24 | 2000年1月-2019年5月 | 3544 |
| 价格指数 | 52 | 1995年1月-2019年4月 | 10259 |
| 交通运输 | 88 | 2005年1月-2019年4月 | 15104 |
| 金融 | 6 | 2000年1月-2019年4月 | 1392 |
| 能源 | 16 | 2000年1月-2019年5月 | 3300 |
| 邮电通信 | 42 | 2002年1月-2019年4月 | 3546 |

#### 3.构建数据特征库

从国家信息中心大数据发展部所掌握的覆盖全国范围1.1亿家企业和个体工商户工商注册、就业招聘、招投标等数据的企业生产经营行为大数据平台中抽取6项企业数据[[[244]](#footnote-244)]，按地域（31个省份）、性质（国企、民企、外企）、行业分类（20个行业大类[[[245]](#footnote-245)]）进行多种数值变换处理[[[246]](#footnote-246)]，经过数据抽取维度、数据分类维度和数值加工处理维度等交叉匹配，最终形成共1296个数据特征。数据时间跨度为2000年1月~2019年7月，共235个月数据，数据特征库建设逻辑如表6所示。

表6 数据特征库建设逻辑

|  |  |  |
| --- | --- | --- |
| 数据抽取维度 | 分类维度 | 数值加工处理维度 |
| 企业增量 | 分地域 | 原始值 |
| 企业存量 | 分性质 | 标准差 |
| 企业注销数量 | 分行业 | 同比 |
| 企业吊销数量 |  | 熵 |
| 企业平均存续时长（月） |  |  |
| 企业登记注册资本比例 |  |  |

### （三）“三库碰撞”数据实验

#### 1.经济指标预测结果

采用“三库碰撞”方法开展经济指标预测的实验结果如表7和表8所示，不难发现，与监测指标的历史上升下降或荣枯背景分布概率相比，该方法对大部分经济指标T+1预测的准确率均有明显提升，从目前实验结果看，“三库碰撞”方式可适用于大部分经济指标的升降预测或荣枯预测。

表7 经济指标预测结果（预测升降）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 监测指标名称 | 算法1（NaiveBayse） | 算法2（LibSVM） | 算法3（AdaBoostM1） | 算法4（J48） | 算法5（RandomForest） | 历史背景分布概率 | 准确率提升（最高） | 准确率提升（最低） |
| 1 | 建筑材料及非金属矿类购进价格指数(上年同月=100) | 81.39% | 85.28% | 77.06% | 82.25% | 81.39% | 51.95% | 33.33% | 25.11% |
| 2 | 国家财政支出(不含债务还本)当期值(亿元) | 87.45% | 83.98% | 74.03% | 79.22% | 82.25% | 55.84% | 31.60% | 18.18% |
| 3 | 房地产施工面积累计增长(%) | 80.17% | 72.41% | 62.50% | 76.72% | 77.16% | 50.65% | 29.52% | 11.85% |
| 4 | PMI制造业采购经理指数 | 78.88% | 81.03% | 79.31% | 83.19% | 79.31% | 54.07% | 29.12% | 24.81% |
| 5 | 白酒（折65度，商品量）产量当期值(万千升) | 85.34% | 82.76% | 70.26% | 84.05% | 87.07% | 58.08% | 28.99% | 12.18% |
| 6 | 房地产开发计划总投资累计增长(%) | 79.74% | 76.29% | 67.24% | 75.86% | 78.45% | 51.57% | 28.17% | 15.67% |
| 7 | 有色金属材料和电线类购进价格指数(上年同月=100) | 79.65% | 72.73% | 70.56% | 72.73% | 77.92% | 52.38% | 27.27% | 18.18% |
| 8 | 房地产竣工面积累计增长(%) | 78.72% | 76.60% | 76.60% | 74.47% | 74.47% | 52.81% | 25.91% | 21.65% |
| 9 | 国家财政收入当期值(亿元) | 84.85% | 86.15% | 78.79% | 82.25% | 86.15% | 61.47% | 24.68% | 17.32% |
| 10 | 黑色金属材料类购进价格指数(上年同月=100) | 78.36% | 74.46% | 74.46% | 72.29% | 77.92% | 54.11% | 24.24% | 18.18% |
| 11 | 商品房销售面积累计增长(%) | 74.57% | 70.69% | 68.53% | 71.12% | 71.12% | 50.65% | 23.92% | 17.89% |
| 12 | 房地产投资累计增长(%) | 74.57% | 71.98% | 70.26% | 71.12% | 75.43% | 52.38% | 23.05% | 17.88% |
| 13 | 化工原料类购进价格指数(上年同月=100) | 74.46% | 72.29% | 68.83% | 71.43% | 72.73% | 51.52% | 22.94% | 17.32% |
| 14 | 木材及纸浆类购进价格指数(上年同月=100) | 75.76% | 66.67% | 68.83% | 72.73% | 75.76% | 52.81% | 22.94% | 13.85% |
| 15 | 原盐产量当期值(万吨) | 79.65% | 81.82% | 75.76% | 73.16% | 81.39% | 59.13% | 22.69% | 14.03% |
| 16 | 纺织原料类购进价格指数(上年同月=100) | 77.06% | 73.59% | 73.16% | 77.06% | 78.79% | 56.28% | 22.51% | 16.88% |
| 17 | 工业生产者出厂价格指数(上年同月=100) | 77.06% | 74.03% | 74.03% | 75.76% | 75.32% | 54.98% | 22.08% | 19.05% |
| 18 | 燃料、动力类购进价格指数(上年同月=100) | 72.73% | 72.29% | 71.43% | 71.43% | 71.43% | 52.38% | 20.35% | 19.05% |
| 19 | 农副产品类购进价格指数(上年同月=100) | 72.29% | 69.26% | 69.26% | 70.13% | 72.29% | 52.38% | 19.91% | 16.88% |
| 20 | 生产资料工业生产者出厂价格指数(上年同月=100) | 74.46% | 71.00% | 64.94% | 72.73% | 74.46% | 54.55% | 19.91% | 10.39% |
| 21 | 房地产业土地购置面积累计增长(%) | 66.38% | 69.83% | 66.81% | 69.40% | 71.12% | 51.95% | 19.17% | 14.43% |
| 22 | 房地产住宅投资累计增长(%) | 69.40% | 68.97% | 63.36% | 69.40% | 69.83% | 51.08% | 18.75% | 12.28% |
| 23 | 货币(M1)供应量同比增长(%) | 68.83% | 68.83% | 68.83% | 68.83% | 69.70% | 51.08% | 18.61% | 17.75% |
| 24 | 生铁产量当期值(万吨) | 75.86% | 73.71% | 72.41% | 64.66% | 74.57% | 57.58% | 18.29% | 7.08% |
| 25 | 工业生产者购进价格指数(上年同月=100) | 71.43% | 67.53% | 66.23% | 71.86% | 72.29% | 54.11% | 18.18% | 12.12% |
| 26 | 房地产开发新增固定资产投资累计增长(%) | 71.12% | 69.40% | 68.97% | 70.69% | 72.41% | 54.26% | 18.15% | 14.71% |
| 27 | 工业增加值累计增长(%) | 74.03% | 74.89% | 70.56% | 71.00% | 75.76% | 58.70% | 17.06% | 11.87% |
| 28 | 生铁产量累计增长(%) | 76.29% | 75.43% | 68.97% | 71.98% | 74.57% | 60.17% | 16.12% | 8.79% |
| 29 | 生活资料工业生产者出厂价格指数(上年同月=100) | 77.49% | 73.59% | 73.16% | 74.46% | 76.62% | 61.90% | 15.58% | 11.26% |
| 30 | 水泥产量累计增长(%) | 72.41% | 71.55% | 71.55% | 71.12% | 71.12% | 56.71% | 14.84% | 14.41% |
| 31 | 纯碱（碳酸钠）产量累计增长(%) | 71.98% | 68.97% | 63.79% | 66.38% | 68.10% | 57.14% | 14.84% | 6.65% |
| 32 | 其它工业原材料及半成品类购进价格指数(上年同月=100) | 70.13% | 69.70% | 67.53% | 67.10% | 67.97% | 55.41% | 14.72% | 11.69% |
| 33 | 房地产商业营业用房投资累计增长(%) | 62.93% | 60.78% | 60.78% | 65.09% | 60.78% | 50.65% | 14.44% | 10.13% |
| 34 | 国家财政收入累计增长(%) | 65.80% | 67.97% | 67.10% | 66.67% | 68.83% | 53.68% | 14.29% | 12.12% |
| 35 | 工业增加值同比增长(%) | 67.53% | 67.53% | 67.53% | 67.53% | 67.53% | 53.48% | 14.05% | 14.05% |
| 36 | 其它房地产投资累计增长(%) | 68.97% | 68.97% | 60.34% | 68.97% | 68.97% | 54.98% | 13.99% | 5.37% |
| 37 | 水泥产量当期值(万吨) | 79.31% | 75.00% | 75.86% | 77.16% | 79.31% | 65.37% | 13.94% | 9.63% |
| 38 | 精炼铜（电解铜）产量同比增长(%) | 65.09% | 65.09% | 65.09% | 64.22% | 65.09% | 51.53% | 13.56% | 12.70% |
| 39 | 白酒（折65度，商品量）产量同比增长(%) | 62.07% | 62.07% | 61.21% | 62.07% | 62.07% | 50.22% | 11.85% | 10.99% |
| 40 | 精炼铜（电解铜）产量当期值(万吨) | 72.41% | 70.26% | 68.97% | 68.53% | 67.67% | 61.14% | 11.28% | 6.54% |
| 41 | 国家财政收入累计值(亿元) | 96.54% | 98.27% | 97.84% | 96.54% | 97.40% | 88.31% | 9.96% | 8.23% |
| 42 | 国家财政支出(不含债务还本)累计值(亿元) | 96.54% | 98.27% | 97.84% | 96.54% | 97.40% | 88.31% | 9.96% | 8.23% |
| 43 | 房地产办公楼投资累计增长(%) | 59.91% | 58.62% | 57.33% | 56.90% | 59.91% | 50.22% | 9.70% | 6.68% |
| 44 | 铁矿石原矿产量当期值(万吨) | 71.43% | 67.53% | 70.13% | 64.50% | 66.23% | 61.74% | 9.69% | 2.76% |
| 45 | 水泥产量同比增长(%) | 65.95% | 64.22% | 59.05% | 65.52% | 65.95% | 56.28% | 9.67% | 2.77% |
| 46 | CPI居民消费价格指数（上年同月=100） | 67.67% | 66.81% | 61.64% | 66.81% | 67.67% | 58.19% | 9.48% | 3.45% |
| 47 | 办公楼施工面积累计值(万平方米) | 99.57% | 99.14% | 97.41% | 98.28% | 98.71% | 91.03% | 8.54% | 6.38% |
| 48 | 办公楼竣工面积累计值(万平方米) | 100.00% | 100.00% | 98.28% | 94.83% | 99.14% | 91.48% | 8.52% | 3.35% |
| 49 | 房地产开发计划总投资累计值(亿元) | 97.41% | 98.28% | 95.26% | 92.67% | 97.84% | 90.13% | 8.14% | 2.54% |
| 50 | 办公楼新开工施工面积累计值(万平方米) | 99.57% | 99.57% | 99.14% | 93.97% | 99.57% | 91.48% | 8.09% | 7.66% |
| 51 | 办公楼销售面积累计值(万平方米) | 99.57% | 99.57% | 99.14% | 93.97% | 99.57% | 91.48% | 8.09% | 2.49% |
| 52 | 房地产开发新增固定资产投资累计值(亿元) | 99.57% | 99.57% | 97.41% | 94.83% | 99.57% | 91.48% | 8.09% | 3.35% |
| 53 | 精炼铜（电解铜）产量累计增长(%) | 58.62% | 58.19% | 56.03% | 57.76% | 58.62% | 50.66% | 7.97% | 5.38% |
| 54 | 铁矿石原矿产量累计值(万吨) | 99.57% | 99.13% | 98.70% | 98.27% | 99.13% | 91.74% | 7.83% | 6.53% |
| 55 | 原盐产量累计值(万吨) | 99.57% | 99.13% | 98.70% | 98.27% | 99.13% | 91.74% | 7.83% | 6.53% |
| 56 | 房地产投资累计值(亿元) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 57 | 房地产住宅投资累计值(亿元) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 58 | 房地产办公楼投资累计值(亿元) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 59 | 房地产商业营业用房投资累计值(亿元) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 60 | 其它房地产投资累计值(亿元) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 61 | 房地产施工面积累计值(万平方米) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 62 | 房地产竣工面积累计值(万平方米) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.93% |
| 63 | 商品房销售面积累计值(万平方米) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 64 | 商品住宅施工面积累计值(万平方米) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 65 | 纯碱（碳酸钠）产量累计值(万吨) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 66 | 生铁产量累计值(万吨) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 67 | 水泥产量累计值(万吨) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 7.79% | 6.50% |
| 68 | 房地产业土地购置面积累计值(万平方米) | 99.57% | 99.14% | 99.14% | 98.28% | 98.71% | 91.77% | 6.93% | 6.50% |
| 69 | 铁矿石原矿产量累计增长(%) | 64.07% | 63.64% | 64.50% | 63.64% | 63.20% | 58.26% | 6.24% | 4.94% |
| 70 | 纯碱（碳酸钠）产量同比增长(%) | 61.21% | 62.07% | 56.90% | 58.19% | 59.48% | 55.84% | 6.22% | 1.05% |
| 71 | 货币(M1)供应量期末值(亿元) | 77.92% | 78.79% | 76.19% | 78.79% | 81.82% | 75.76% | 6.06% | 0.43% |
| 72 | 纯碱（碳酸钠）产量当期值(万吨) | 67.24% | 67.67% | 67.24% | 67.67% | 67.67% | 62.34% | 5.33% | 5.33% |
| 73 | 白酒（折65度，商品量）产量累计值(万千升) | 96.98% | 94.83% | 94.83% | 96.12% | 96.98% | 91.70% | 5.28% | 3.12% |
| 74 | 精炼铜（电解铜）产量累计值(万吨) | 96.98% | 94.83% | 94.83% | 96.12% | 96.98% | 91.70% | 5.28% | 3.12% |
| 75 | 原盐产量同比增长(%) | 61.47% | 61.47% | 61.47% | 61.47% | 61.47% | 56.96% | 4.52% | 4.52% |
| 76 | 货币和准货币(M2)供应量期末值(亿元) | 92.21% | 89.61% | 90.04% | 88.31% | 92.21% | 87.88% | 4.33% | 0.43% |
| 77 | 生铁产量同比增长(%) | 59.48% | 59.48% | 59.48% | 59.48% | 59.48% | 55.84% | 3.64% | 3.64% |
| 78 | 白酒（折65度，商品量）产量累计增长(%) | 61.21% | 61.21% | 61.21% | 61.21% | 61.21% | 57.64% | 3.56% | 3.56% |
| 79 | 原盐产量累计增长(%) | 54.11% | 52.38% | 54.11% | 52.38% | 51.52% | 53.91% | 0.20% | -2.40% |
| 80 | 货币和准货币(M2)供应量同比增长(%) | 50.65% | 49.78% | 50.65% | 50.65% | 49.78% | 50.65% | 0.00% | -0.87% |
| 81 | 国家财政支出(不含债务还本)累计增长(%) | 52.38% | 50.22% | 52.38% | 52.38% | 52.38% | 52.38% | 0.00% | -2.16% |
| 82 | 铁矿石原矿产量同比增长(%) | 56.71% | 56.71% | 56.71% | 56.71% | 56.71% | 56.96% | -0.25% | -0.25% |

表8 经济指标预测结果（预测荣枯）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 统计指标类别 | 统计指标名称 | 算法1（NaiveBayse） | 算法2（LibSVM） | 算法3（AdaBoostM1） | 算法4（J48） | 算法5（RandomForest） | 历史背景分布概率 | 准确率提升（最高） | 准确率提升（最低） |
| 制造业采购经理指数 | 制造业采购经理指数 | 97.66% | 92.98% | 92.98% | 90.64% | 95.32% | 86.55% | 11.11% | 4.09% |
| 生产指数(%) | 98.83% | 98.83% | 98.83% | 96.49% | 99.42% | 96.49% | 2.92% | 0.00% |
| 新订单指数(%) | 95.91% | 92.40% | 91.23% | 90.06% | 94.74% | 88.89% | 7.02% | 1.17% |
| 采购量指数(%) | 92.98% | 92.40% | 93.57% | 90.06% | 94.74% | 85.38% | 9.36% | 4.68% |
| 进口指数(%) | 87.13% | 87.72% | 81.87% | 81.29% | 87.72% | 50.29% | 37.43% | 30.99% |
| 从业人员指数(%) | 94.74% | 94.74% | 93.57% | 91.81% | 95.91% | 60.23% | 35.67% | 31.58% |
| 非制造业采购经理指数 | 新订单指数(%) | 95.27% | 93.24% | 95.27% | 95.27% | 95.95% | 91.22% | 4.73% | 2.03% |
| 销售价格指数(%) | 84.46% | 84.46% | 81.76% | 81.08% | 87.84% | 61.49% | 26.35% | 19.59% |
| 从业人员指数(%) | 94.59% | 94.59% | 95.27% | 91.22% | 97.97% | 62.84% | 35.14% | 28.38% |

#### 2.经济指标时序预测变化趋势分析

为了分析随着时间推移即训练样本不断增加情况下，利用“三库碰撞”方法开展经济指标预测效果的变化趋势，本研究将实验样本进行了切片处理，共分为10个时间切片。分析发现该方法适用于部分经济指标，即随着时间的推移预测准确率能够不断提升，其中，建筑材料及非金属矿类购进价格指数和制造业采购经理指数的时序预测结果如图17、图18所示。

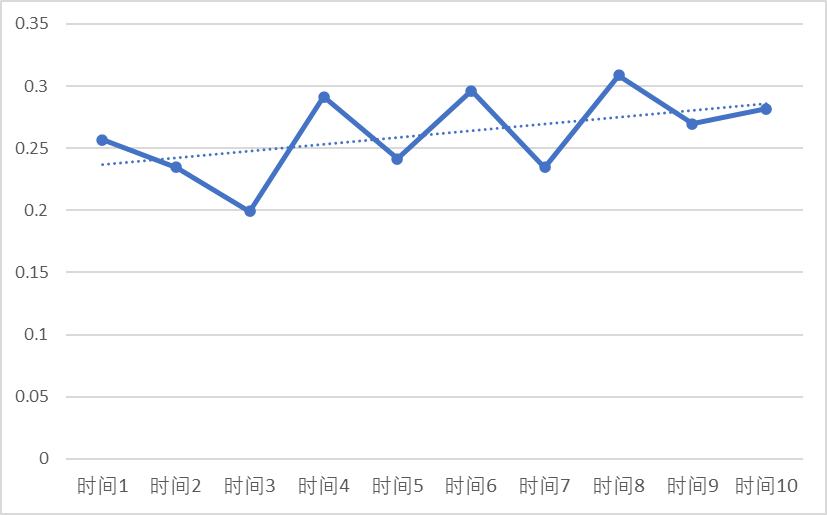


图17 建筑材料及非金属矿类购进价格指数时序预测结果

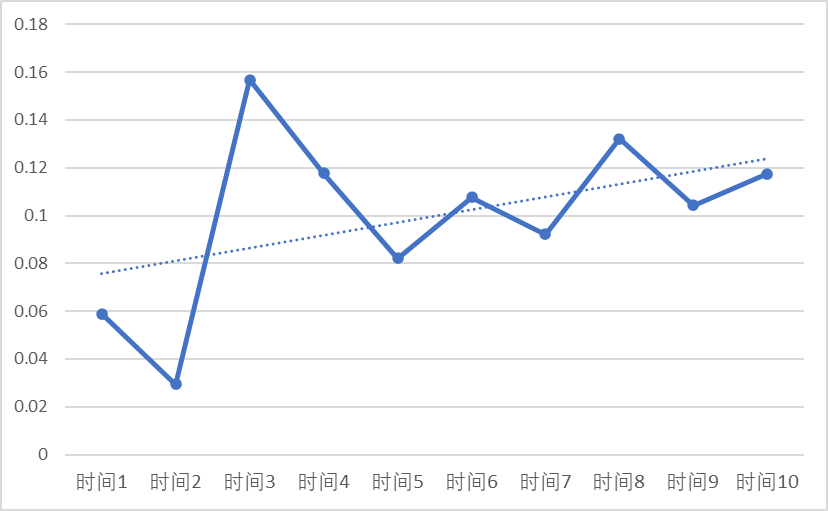


图18  制造业采购经理指数时序预测结果

## 六、研究内容之三：基于机器学习方法的宏观经济指标数值预测研究

近年来机器学习数值预测成为国内外学者的研究热点之一，由于其在处理多变量、多时相数据时表现出高精度、高时效等优势。目前，主流的机器学习数值预测算法主要包括线性回归、弹性网络回归（ElasticNet）、随机森林（RandomForest）和梯度提升回归树（GBRT）等。

### （一）融合机器学习算法和多重因素的白条鸡日均交易量预测

有鉴于此，课题组以白条鸡为例，综合考虑市场主体、舆情信息、搜索意愿和相关统计等诸多影响因素的动态变化趋势，利用主流的机器学习数值预测算法创新性提出一种白条鸡日均交易量预测方法，量化评估畜禽市场变化。目前，学术界对国民经济和农业生产领域的预测研究大多采用传统趋势预测模型，虽然时间序列方法具有需要的特征数据种类少、计算简单等优势，但对历史数据量要求较高，且难以在预测中引入短期市场变化因素，无法及时反映节假日和重大突发事件等对畜禽行业造成的冲击。同时，考虑到本研究预测对象——白条鸡交易量具有历史数据较缺乏、可获取的影响因素数据较多的特点，自变量多、因变量少的机器学习预测方法为本研究提供了新的解决思路。

总体来看，目前对畜禽肉产品交易量或产量预测的研究较少，尤其缺乏综合考虑短期影响因素的动态预测方法研究。在影响因素较多、预测值历史数据较少的情况下，机器学习预测算法相较传统方法表现更佳。因此，本研究提出一种融合机器学习算法和广泛特征数据的预测方法，尝试对白条鸡日均交易量做出短期预测，并分析预测效果。考虑到市场同时受到新型冠状肺炎疫情影响，在特征选取过程中也纳入了相关数据。

#### 1.研究思路

融合机器学习算法和多重因素的白条鸡日均交易量预测方法总体思路如下图所示，其核心思想是能够综合考虑中国农历新年等短期变化因素对白条鸡交易量的影响，采集并抽取鸡禽类产业、互联网舆情信息、网民需求意愿和相关统计数据等多源数据中可量化数据特征，利用主流机器学习算法预测白条鸡交易量，对比不同算法的预测效果，探索随着时间增长、训练样本增加时各算法的迭代滚动效果，同时分析各算法达到稳定预测所需要的训练样本数量和提前期数，总体思路图如图19所示。

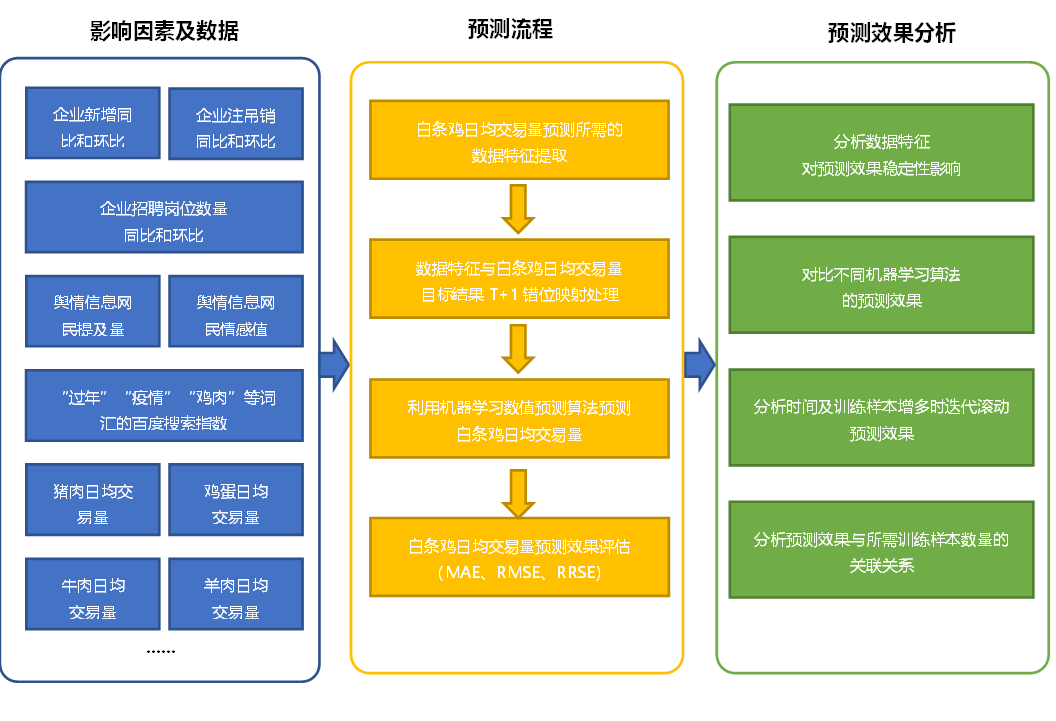


图19 融合机器学习算法和多重因素的白条鸡日均交易量预测方法

其中，白条鸡日均交易量预测函数表示为:

 式（6.1）

公式（6.1）中，Φt代表白条鸡在第t周的每日交易量，φt-1代表鸡禽类上下游行业市场主体及招聘岗位在第t-1周的变化数据，ξt-1代表第t-1周的相关舆情数据，δt-1代表互联网网民在第t-1周的搜索意愿数据，ηt-1代表其他渠道公开发布的第t-1周相关统计数据，函数f拟采用机器学习数值型预测算法。

对白条鸡的日均交易量开展预测前，需要预先设定预测算法所需的数据特征，特征选取的优劣会直接影响预测算法的性能。由于当前及近期白条鸡交易行为已经受到农历新年和新冠肺炎疫情影响，因此，本研究在数据特征选取过程中，兼顾鸡禽类产业、互联网舆情信息、网民需求意愿、白条鸡相关统计数据等多种趋势变化，综合考虑多方面因素对白条鸡交易量的影响。本研究共提取了50个可能与白条鸡日均交易量预测有关的变量作为数据特征，包括市场主体特征30个、舆情信息特征2个、搜索意愿特征14个和统计数据特征4个，并且利用常见的统计方法对每一类特征进行了细分、量化及逐周提取，白条鸡日均交易量预测特征如下表所示。

①市场主体特征

市场主体特征是指鸡禽类相关上下游产业的企业和个体工商户等市场主体变化及招聘岗位数量变化情况。主要包括：白条鸡相关鸡鸭等家禽养殖类、鸡鸭等家禽饲料类、鸡鸭等家禽屠宰加工类、鸡苗种鸡类、生产禽药类企业和个体工商户的注销及吊销数量、新增数量和招聘岗位数量，并进行相应的同比和环比变换处理。

②舆情信息特征

舆情信息特征是指网民对鸡肉等相关信息提及量及情感值变化情况。主要包括：网民提及鸡肉等相关舆情信息的发帖数量和网民的对应情感值等。

③搜索意愿特征

搜索意愿特征是指网民对鸡肉、过年和疫情等相关主题的搜索意愿程度。主要包括：“过年”“鸡肉”“鸡肉价格”“饲料”“扫福”“在线办公”“返乡”“年货”“拜年”“疾病”“买菜”“疫情”“抢票”“鸡苗”等词汇的百度指数结果。

④统计数据特征

统计数据特征是指与白条鸡日均交易量有关的其他统计数据情况如表9所示。主要包括：农业农村部信息中心发布的猪肉日均交易量、鸡蛋日均交易量、牛肉日均交易量、羊肉日均交易量等。

表9 白条鸡日均交易量预测特征（逐周）

|  |  |  |
| --- | --- | --- |
| 市场主体特征 | | |
| F1 | BREEDING\_ADD\_YOY | 鸡鸭等家禽养殖企业和个体工商户新增数量同比值 |
| F2 | BREEDING\_CANCEL\_REVOKE\_YOY | 鸡鸭等家禽养殖企业和个体工商户注销及吊销数量同比值 |
| F3 | BREEDING\_RECRUIT\_YOY | 鸡鸭等家禽养殖企业和个体工商户招聘岗位数量同比值 |
| F4 | FEED\_ADD\_YOY | 鸡鸭等家禽饲料企业和个体工商户新增数量同比值 |
| F5 | FEED\_CANCEL\_REVOKE\_YOY | 鸡鸭等家禽饲料企业和个体工商户注销及吊销数量同比值 |
| F6 | FEED\_RECRUIT\_YOY | 鸡鸭等家禽饲料企业和个体工商户招聘岗位数量同比值 |
| F7 | SLAUGHTER\_ADD\_YOY | 鸡鸭等家禽屠宰加工企业和个体工商户新增数量同比值 |
| F8 | SLAUGHTER\_CANCEL\_REVOKE\_YOY | 鸡鸭等家禽屠宰加工企业和个体工商户注销及吊销数量同比值 |
| F9 | SLAUGHTER\_RECRUIT\_YOY | 鸡鸭等家禽屠宰加工企业和个体工商户招聘岗位数量同比值 |
| F10 | CHICK\_ADD\_YOY | 鸡苗种鸡企业和个体工商户新增数量同比值 |
| F11 | CHICK\_CANCEL\_REVOKE\_YOY | 鸡苗种鸡企业和个体工商户注销及吊销数量同比值 |
| F12 | CHICK\_RECRUIT\_YOY | 鸡苗种鸡企业和个体工商户招聘岗位数量同比值 |
| F13 | MEDICINE\_ADD\_YOY | 生产禽药企业和个体工商户新增数量同比值 |
| F14 | MEDICINE\_CANCEL\_REVOKE\_YOY | 生产禽药企业和个体工商户注销及吊销数量同比值 |
| F15 | MEDICINE\_RECRUIT\_YOY | 生产禽药企业和个体工商户招聘岗位数量同比值 |
| F16 | BREEDING\_ADD\_QOQ | 鸡鸭等家禽养殖企业和个体工商户新增数量环比值 |
| F17 | BREEDING\_CANCEL\_REVOKE\_QOQ | 鸡鸭等家禽养殖企业和个体工商户注销及吊销数量环比值 |
| F18 | BREEDING\_RECRUIT\_QOQ | 鸡鸭等家禽养殖企业和个体工商户招聘岗位数量环比值 |
| F19 | FEED\_ADD\_QOQ | 鸡鸭等家禽饲料企业和个体工商户新增数量环比值 |
| F20 | FEED\_CANCEL\_REVOKE\_QOQ | 鸡鸭等家禽饲料企业和个体工商户注销及吊销数量环比值 |
| F21 | FEED\_RECRUIT\_QOQ | 鸡鸭等家禽饲料企业和个体工商户招聘岗位数量环比值 |
| F22 | SLAUGHTER\_ADD\_QOQ | 鸡鸭等家禽屠宰加工企业和个体工商户新增数量环比值 |
| F23 | SLAUGHTER\_CANCEL\_REVOKE\_QOQ | 鸡鸭等家禽屠宰加工企业和个体工商户注销及吊销数量环比值 |
| F24 | SLAUGHTER\_RECRUIT\_QOQ | 鸡鸭等家禽屠宰加工企业和个体工商户招聘岗位数量环比值 |
| F25 | CHICK\_ADD\_QOQ | 鸡苗种鸡企业和个体工商户新增数量环比值 |
| F26 | CHICK\_CANCEL\_REVOKE\_QOQ | 鸡苗种鸡企业和个体工商户注销及吊销数量环比值 |
| F27 | CHICK\_RECRUIT\_QOQ | 鸡苗种鸡企业和个体工商户招聘岗位数量环比值 |
| F28 | MEDICINE\_ADD\_QOQ | 生产禽药企业和个体工商户新增数量环比值 |
| F29 | MEDICINE\_CANCEL\_REVOKE\_QOQ | 生产禽药企业和个体工商户注销及吊销数量环比值 |
| F30 | MEDICINE\_RECRUIT\_QOQ | 生产禽药企业和个体工商户招聘岗位数量环比值 |
| 舆情信息特征 | | |
| F31 | CHICKEN\_NUMS | 网民提及鸡肉等相关舆情信息数量 |
| F32 | CHICKEN\_ EMOTION | 网民提及鸡肉等相关舆情信息情感值 |
| 搜索意愿特征 | | |
| F33 | SEARCH\_ SPRING\_FESTIVAL | “过年”一词百度指数结果 |
| F34 | SEARCH\_ CHICKEN | “鸡肉”一词百度指数结果 |
| F35 | SEARCH\_ CHICKEN\_PRICE | “鸡肉价格”一词百度指数结果 |
| F36 | SEARCH\_ FEED | “饲料”一词百度指数结果 |
| F37 | SEARCH\_ BLESS | “扫福”一词百度指数结果 |
| F38 | SEARCH\_ ONLINE\_OFFICE | “在线办公”一词百度指数结果 |
| F39 | SEARCH\_RETURN | “返乡”一词百度指数结果 |
| F40 | SEARCH\_NECESSITIES | “年货”一词百度指数结果 |
| F41 | SEARCH\_GREETINGS | “拜年”一词百度指数结果 |
| F42 | SEARCH\_DISEASE | “疾病”一词百度指数结果 |
| F43 | SEARCH\_VEGETABLES | “买菜”一词百度指数结果 |
| F44 | SEARCH\_EPIDEMIC | “疫情”一词百度指数结果 |
| F45 | SEARCH\_TICKET | “抢票”一词百度指数结果 |
| F46 | SEARCH\_ CHICK | “鸡苗”一词百度指数结果 |
| 统计数据特征 | | |
| F47 | PORK\_NUMS | 猪肉日均交易量（统计口径） |
| F48 | EGG\_NUMS | 鸡蛋日均交易量（统计口径） |
| F49 | BEEF\_NUMS | 牛肉日均交易量（统计口径） |
| F50 | MUTTON\_NUMS | 羊肉日均交易量（统计口径） |

#### 2.应用机器学习算法介绍

当前主流的机器学习算法包括线性回归（Linear Regression）、弹性网络回归（Elastic Net）、随机森林（Random Forest）和梯度提升回归树（GBRT）等，下面分别对其原理作简要介绍，便于在后续对比实验中应用。

线性回归假设目标值与特征之间线性相关，即满足一个多元一次方程。通过构建损失函数，来求解损失函数最小时的参数。求解方式有两种：一是基于均方误差最小化的最小二乘法；二是对自变量进行不断的更新，使得目标函数不断逼近最小值的过程的梯度提升法。该算法简单易操作，但在自变量较多时容易过度拟合。

弹性网络回归为了减小过度拟合的程度，用最小二乘法最小化损失函数，并加入权值向量中各元素绝对值之和（L1），与权值向量中各元素平方和的根（L2）正则化18，是一种结合岭回归和Lasso算法的组合模型。该模型收敛速度较快，尤其适用于在多个特征间存在较强相关性时保持稳定性，但在特征选择时会降低原始数据所包含的信息维度。

随机森林由多棵决策树构成，且森林中的每一棵决策树之间没有关联，模型的最终输出由森林中的每一棵决策树共同决定。该模型被广泛应用于分类和回归问题上，主要取决于随机森林的每颗CART树是分类树还是回归树19。如果是回归树，则CART树是回归树，采用的原则是最小均方差。该模型的优点是较好地解决了决策树过拟合和精确度低的不足，适用于高维数据预测，但具有运算效率低的缺点。

梯度提升算法的主要思路：利用最速下降的近似方法，即利用损失函数的负梯度在当前模型的值，作为回归问题中提升树算法的残差的近似值（伪残差），拟合一个回归树20。在回归问题中，这称为梯度提升回归树，分类问题则称为梯度提升决策树，具体公式如（6.2）所示。给定训练集，梯度提升模型的目标是找到一个函数，使映射到的损失函数到达最小：

式（6.2）

通过多次迭代，最终预测函数为迭代过程中所有子回归树的线性组合。该算法的优点为非线性变换比较多、表达能力强，不需要做复杂的特征工程和特征变换，并且能够防止过拟合。

#### 3.实验数据与评估指标

①数据描述

为了有效开展白条鸡日均交易量预测方法实验，本研究抓取了农业农村部信息中心官方网站上公开发布的一年来白条鸡日均交易量（逐周）数据作为预测目标结果，数据时间范围为2019年2月25日（周一）至2020年3月1日（周日），共53周。同时，为了有效提取数据特征，同步抓取了四个方面数据：

（1）2018年2月25日至2020年2月23日（104周）期间鸡鸭等家禽养殖类、鸡鸭等家禽饲料类、鸡鸭等家禽屠宰加工类、鸡苗种鸡类、生产禽药类企业和个体工商户登记注册数据共522,811条，以及这些市场主体发布的招聘岗位数量共1,147,948条。其中，选取两年作为该数据周期是为了便于同比计算处理。

（2）2019年2月25日至2020年2月23日（52周）期间1,072,805条网民提及鸡肉等相关信息的发帖数据。

（3）2019年2月25日至2020年2月23日（52周）期间“过年”“鸡肉”“疫情”“抢票”等14个关键词的百度指数数据。

（4）2019年2月25日至2020年2月23日（52周）期间农业农村部信息中心官方网站上公开发布的猪肉、鸡蛋、牛肉、羊肉日均交易量统计数据。

此外，为了模拟真实预测场景，本研究将最终提取的50项数据特征与白条鸡日均交易量目标结果进行了T+1错位映射处理，即利用第T周的数据特征映射第T+1周的目标值（共映射形成52周对应数据），以便开展后续预测实验。

②评估指标选取

本研究选取了数值型预测领域较为常用的评估指标进行数据实验效果评估，其中，对于评估预测目标值不变的情况，采用平均绝对误差（MAE）和均方根误差（RMSE）进行效果评估，对于评估预测目标值随着时序不断变化的情况，采用相对平方根误差（RRSE）进行效果评估，其中三种指标数值越小则代表预测效果越好。

——平均绝对误差（MAE）

MAE能更好地反映预测值误差的实际情况，具体公式如（6.3）所示，其中yi和yi’分别代表真实值和预测值。

 式（6.3）

——均方根误差（RMSE）

RMSE同样可以来衡量真实值和预测值之间的偏差，但RMSE面对异常值情况时更为敏感，具体公式如（6.4）所示。

 式（6.4）

——相对平方根误差（RRSE）

RRSE更适用于评估预测目标随着评估流程不同或时序演变而不断变化的情况，具体公式如（6.5）所示。

 式（6.5）

#### 4.数据实验

①数据特征对预测效果稳定性影响

为了证明数据特征对白条鸡日均交易量预测效果的稳定性，即预测效果不会随着算法的不同和数据集的小部分缺失而导致大幅度波动，本研究选取了4种前文所述常见的数值型预测算法对实验数据进行测试，即LinearRegression、ElasticNet、RandomForest和GBRT，各个模型均采用默认参数设置。同时，从全部实验样本中随机抽样5次，每次抽样在上次抽样样本量基础上剔除掉一周数据样本，最终形成52周、51周、50周、49周、48周共5个数据样本，并分别采用十折交叉验证方法测试4种预测算法效果，实验结果如图20所示。

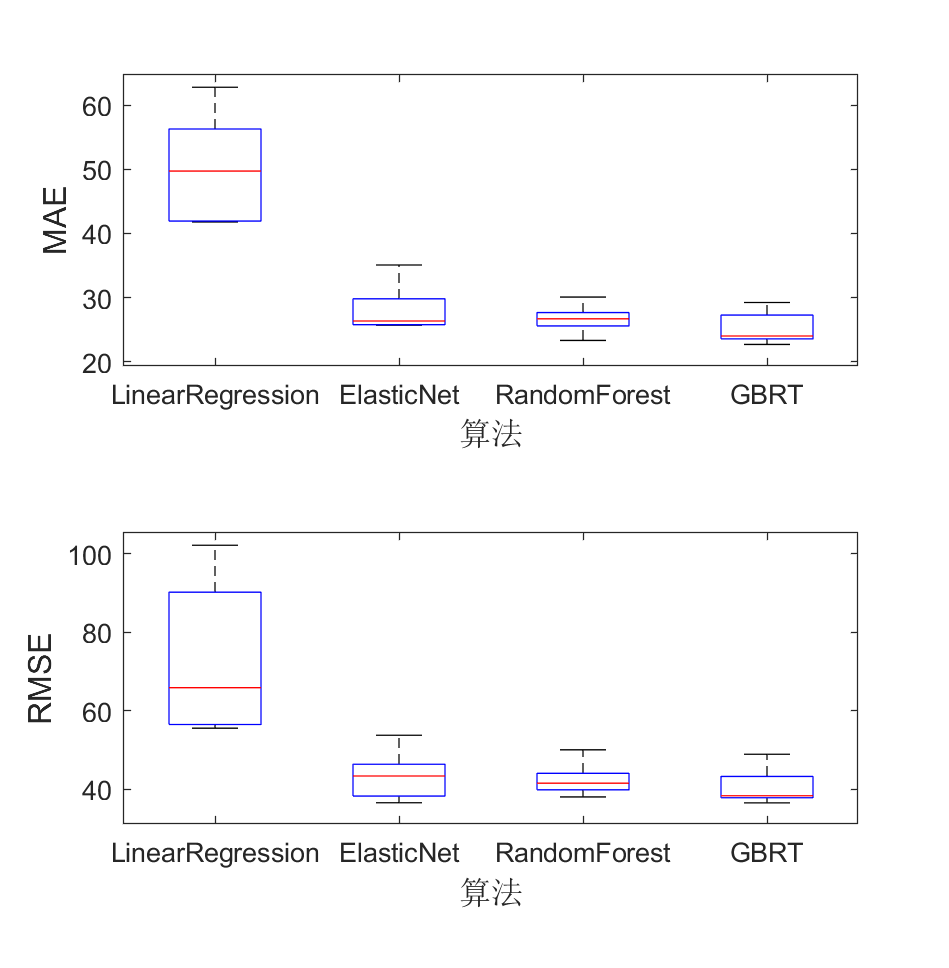


图20 随机抽样数据集对预测效果稳定性的对比结果

从上图可以看出，无论从MAE还是RMSE评估指标看，除线性回归算法在预测效果稳定性表现不佳外，其余三个算法均证明了预测效果的稳定性，而线性回归算法的实验效果不佳主要是因为白条鸡日均交易量预测的数据特征相对较多，而预测样本量相对较少，容易造成过拟合现象。

②不同算法的预测效果分析

为了对比分析不同算法对白条鸡日均交易量的预测效果，本研究分别对比了上述5组数据实验中MAE和RMSE的均值表现，如图21所示。可以看出，从MAE和RMSE评估结果看，预测效果从优至劣排序分别为：GBRT > RandomForest > ElasticNet > LinearRegression，其中，MAE分别为：25.30188、26.67456、28.21178、50.06592，RMSE分别为：40.72684、42.47114、43.32624、73.35752。

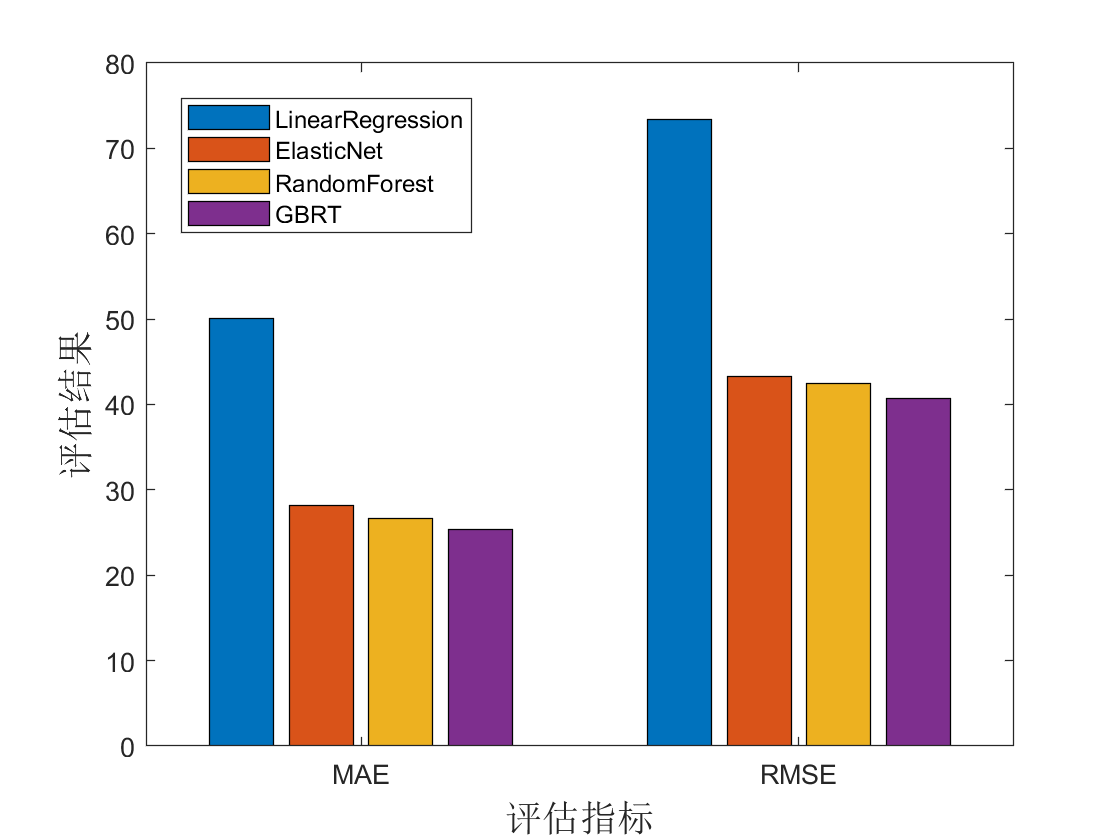


图21 不同算法的预测效果对比分析

③迭代滚动预测效果分析

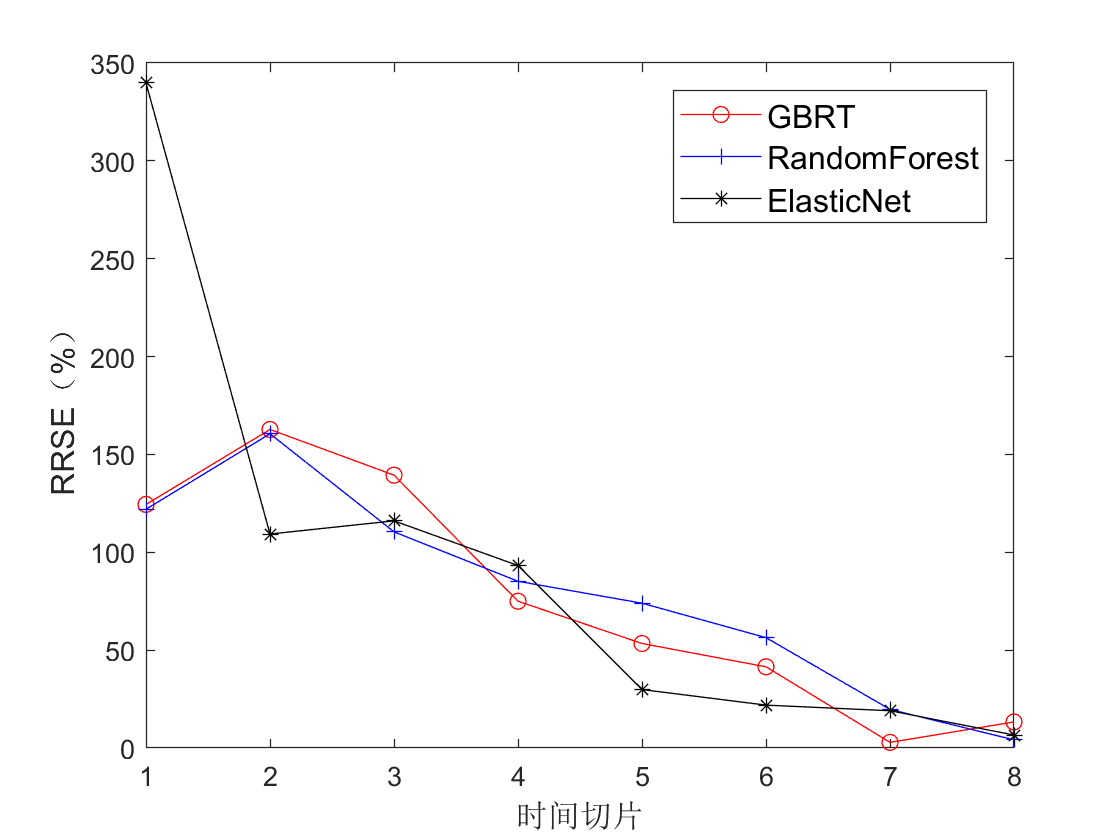
为了分析随着时间推移以及训练样本不断增多时白条鸡日均交易量预测效果的变化情况，本研究将52周实验样本进行了切片处理（具体数据集划分方式如表9所示），并采用迭代滚动预测方式进行实验，由于该种实验方式会导致预测目标值不断变化，因此该部分采用RRSE指标进行评估。

表9  迭代滚动预测实验数据集划分方式

|  |  |  |
| --- | --- | --- |
| 时间切片 | 训练集时间跨度 | 测试集时间跨度 |
| 1 | 第1-44周 | 第45周 |
| 2 | 第1-45周 | 第46周 |
| 3 | 第1-46周 | 第47周 |
| 4 | 第1-47周 | 第48周 |
| 5 | 第1-48周 | 第49周 |
| 6 | 第1-49周 | 第50周 |
| 7 | 第1-50周 | 第51周 |
| 8 | 第1-51周 | 第52周 |

从下图不难发现，从RRSE指标结果看，GBRT、RandomForest和ElasticNet三种算法的RRSE结果呈现出波动下降趋势，可以表明随着时间不断积累以及训练样本数量不断增加，白条鸡日均交易量预测效果不断提升，以GBRT为例，8个时间切片的RRSE分别为124.4436%、162.7494%、139.4145%、74.9973%、53.3966%、41.4741%、2.8766%、13.3162%。

此外，ElasticNet算法在早期预测时波动性较大，而RandomForest算法在前4期的预测效果要优于GBRT算法，但随着时间推移GBRT算法在后期预测效果更佳，具体结果如图22所示。

****

**图22 不同算法的预测效果对比分析**

④预测所需训练样本和提前期数分析

为了分析白条鸡日均交易量预测效果与所需训练样本数量的关联关系，本研究将52周实验样本进行了不同策略的切片处理（具体数据集划分方式如表10所示），即选定同一预测目标，分析所需的样本数量，由于该种实验方式对预测目标值不变，因此该部分采用MAE指标进行评估。

表10 预测效果与所需训练样本数量分析实验数据集划分方式

|  |  |  |
| --- | --- | --- |
| 时间切片 | 训练集时间跨度 | 测试集时间跨度 |
| 1 | 第1-44周 | 第52周 |
| 2 | 第1-45周 | 第52周 |
| 3 | 第1-46周 | 第52周 |
| 4 | 第1-47周 | 第52周 |
| 5 | 第1-48周 | 第52周 |
| 6 | 第1-49周 | 第52周 |
| 7 | 第1-50周 | 第52周 |
| 8 | 第1-51周 | 第52周 |

图23表明，GBRT、RandomForest和ElasticNet三种算法的MAE结果呈现出明显下降趋势，说明训练样本数量越多则对白条鸡日均交易量的预测效果越好，此外，当时间切片为6之后，三种算法的预测效果均趋于稳定，可以表明利用第1至49周数据进行训练便可通过第52周的数据特征预测第53周的白条鸡日均交易量，说明该方法具备提前3期的预测效果，以GBRT为例，8个时间切片的MAE分别为120.8459、114.4232、116.8219、107.348、59.8887、11.7666、13.6099、16.8542。

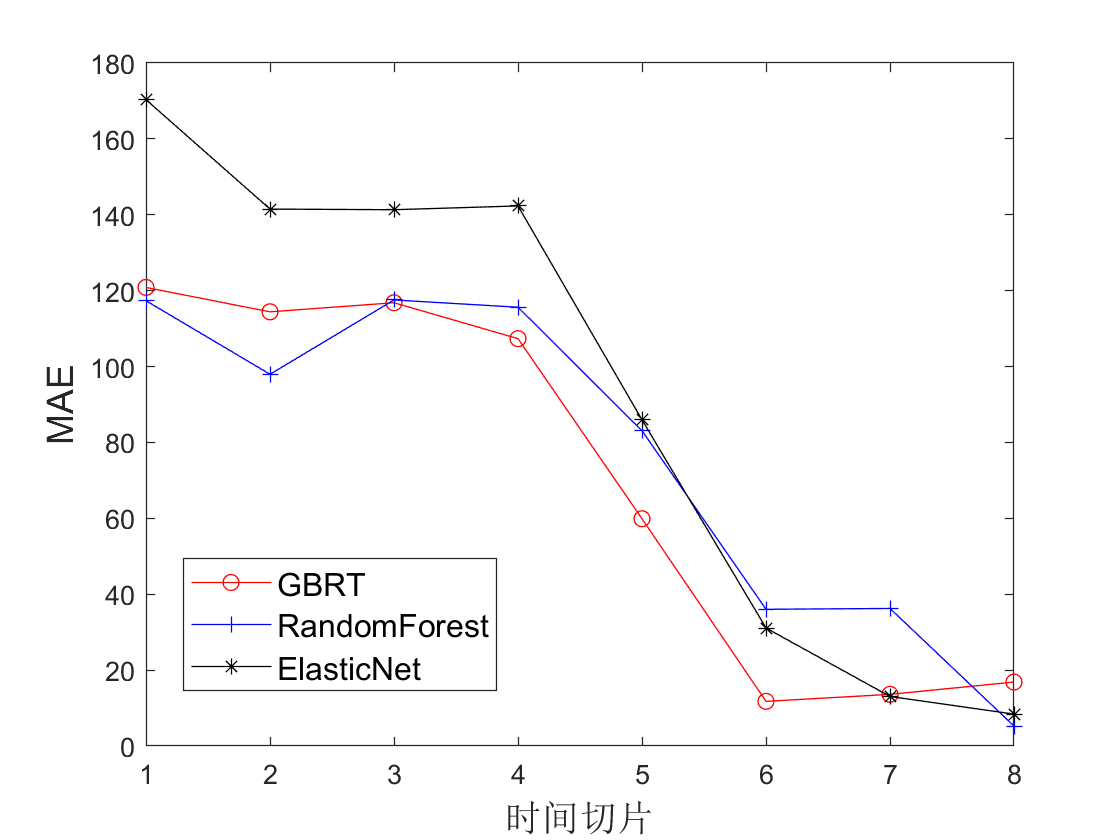


图23 预测所需训练样本和提前期数分析

#### 5.实验结论

本研究以白条鸡日均交易量为例，综合考虑农历新年假期和新冠肺炎疫情背景下诸多因素对白条鸡交易量的动态影响，探索提出一种基于机器学习的白条鸡日均交易量预测方法，对比分析了LinearRegression、ElasticNet、RandomForest和GBRT四种主流算法的预测效果。实验结果显示，除线性回归算法在预测效果稳定性表现不佳外，其余三个算法均具有预测效果的稳定性，且预测效果从优至劣依次为：GBRT、RandomForest、ElasticNet、LinearRegression（MAE分别为：25.30188、26.67456、28.21178、50.06592，RMSE分别为：40.72684、42.47114、43.32624、73.35752）；随着时间不断积累以及训练样本数量不断增加，白条鸡日均交易量预测效果不断提升（以GBRT为例，8个时间切片的RRSE分别为124.4436%、162.7494%、139.4145%、74.9973%、53.3966%、41.4741%、2.8766%、13.3162%）；GBRT、RandomForest和ElasticNet算法具备提前3期的稳定预测效果（以GBRT为例，8个时间切片的MAE分别为120.8459、114.4232、116.8219、107.348、59.8887、11.7666、13.6099、16.8542）。

### （二）利用机器学习算法和广泛数据特征预测某市规模以上工业增加值

目前利用高频数据和机器学习算法对规模以上工业增加值等统计指标预测仍相对研究较少，尤其缺乏综合考虑短期影响因素的动态预测方法研究。在影响因素较多、预测值历史数据较少的情况下，机器学习预测算法相较传统方法表现更佳。因此，本部分拟利用融合机器学习算法和广泛特征数据的预测方法，尝试对某市规模以上工业增加值做出预测，并分析预测效果。

#### 1.研究思路

（1）构建某市统计数据预测函数

考虑到因变量数据的时间范围、更新频率、数据完整性等因素，建议选定某市规模以上工业增加值数据作为因变量，验证预测方法的可行性和预测结果的有效性。拟考虑的解释变量主要包括：（1）某市统计数据（如固定资产投资、金融机构人民币存款余额、工业用电、价格等）；（2）某市企业工商登记注册数据（如各行业新增企业数量及同比环比、各行业注销吊销企业数量及同比环比等）；（3）某市互联网招聘数据（如各行业平均招聘岗位数量及同比环比、各行业平均招聘薪酬及同比环比等）等。数据预测函数可表示为:

 式（6.6）

其中，某一因变量在t月的实际数据，、、、等代表相关解释变量在t-k月的数据特征，供给函数f主要考虑采用数值型预测机器学习相关算法。

（2）数值型预测机器学习算法选型

拟定可供选择的数值型预测机器学习算法如表11所示。

表11 数值型预测机器学习算法

|  |  |  |
| --- | --- | --- |
| 序号 | 算法名称 | 算法特点 |
| 1 | 线性回归（Linear Regression） | 线性回归假设目标值与特征之间线性相关，即满足一个多元一次方程。通过构建损失函数，来求解损失函数最小时的参数。求解方式有两种：一是基于均方误差最小化的最小二乘法；二是对自变量进行不断的更新，使得目标函数不断逼近最小值的过程的梯度提升法。该算法简单易操作，但在自变量较多时容易过度拟合。 |
| 2 | 弹性网络回归（Elastic Net） | 弹性网络回归为了减小过度拟合的程度，用最小二乘法最小化损失函数，并加入权值向量中各元素绝对值之和（L1），与权值向量中各元素平方和的根（L2）正则化，是一种结合岭回归和Lasso算法的组合模型。该模型收敛速度较快，尤其适用于在多个特征间存在较强相关性时保持稳定性，但在特征选择时会降低原始数据所包含的信息维度。 |
| 3 | 随机森林回归（Random Forest） | 随机森林由多棵决策树构成，且森林中的每一棵决策树之间没有关联，模型的最终输出由森林中的每一棵决策树共同决定。该模型被广泛应用于分类和回归问题上，主要取决于随机森林的每颗CART树是分类树还是回归树。如果是回归树，则CART树是回归树，采用的原则是最小均方差。该模型的优点是较好地解决了决策树过拟合和精确度低的不足，适用于高维数据预测，但具有运算效率低的缺点。 |
| 4 | 和梯度提升回归树（GBRT） | 梯度提升算法利用最速下降的近似方法，即利用损失函数的负梯度在当前模型的值，作为回归问题中提升树算法的残差的近似值（伪残差），拟合一个回归树。同时也可以使用梯度下降的思想来理解这一过程：若将当前模型的输出值看作一个变量，要使得损失函数最小化，实际上是一个一维搜索问题，从当前模型的输出值出发，以负梯度方向来搜索最优值。对于新引入决策树划分的每个区域，分别计算其搜索步长，将新决策树在每个区域的预测值与搜索步长合并在一起，便等效于上面的每个区域重新计算输出值。梯度提升回归树的优点主要包括非线性变换比较多、表达能力强，不需要做复杂的特征工程和特征变换，防止过拟合。 |
| 5 | 支持向量机回归（Support Vector Regression） | 支持向量机(SVM)本身是针对二分类问题提出的，而SVR（支持向量回归）是SVM（支持向量机）中的一个重要的应用分支。SVR回归与SVM分类的区别在于，SVR的样本点最终只有一类，它所寻求的最优超平面不是SVM那样使两类或多类样本点分的“最开”，而是使所有的样本点离着超平面的总偏差最小。传统的回归方法当且仅当回归f(x)完全等于y时才认为是预测正确，需计算其损失；而支持向量回归(SVR)则认为只要是f(x)与y偏离程度不要太大，既可认为预测正确，不用计算损失。 |
| 6 | 高斯过程回归（Gaussian Processes） | 高斯过程回归的思想是从无数的函数中找到任何符合该组测试数据的函数集，根据对一组数据集的事先的了解（即先验信息），不断地缩小所找函数集的范围，最终通过贝叶斯法则的方法与高斯分布的性质算出所找到函数集的分布，根据求得函数集的分布来对后续的数据进行预测。 |

（3）评估指标选定

拟选取数值型预测领域较为常用的评估指标进行数据实验效果评估，其中，对于评估预测目标值不变的情况，采用平均绝对误差（MAE）和均方根误差（RMSE）进行效果评估。

MAE能更好地反映预测值误差的实际情况，具体公式如式（6.7）所示，其中yi和yi’分别代表真实值和预测值。

 式（6.7）

RMSE同样可以来衡量真实值和预测值之间的偏差，但RMSE面对异常值情况时更为敏感，具体公式如（6.8）所示。

 式（6.8）

#### 2.实验设计

为有效评估利用上述方法预测规模以上工业增加值的实际效果，本部分从数据特征、模型算法、训练样本等方面出发，共设计了四组对比实验：

实验一：分析数据特征对预测效果稳定性影响。即保持数据特征不变，对实验数据集进行5次随机抽样，并分别对抽样后的数据集应用6种不同数值型预测算法，重点评估多次实验下的MAE和RMSE预测误差值是否存在较大差异，验证预测效果是否稳定。

实验二：对比分析不同算法的预测效果。在实验一基础上，侧重对比6种数值型预测算法在5次随机抽样数据集上预测误差的均值表现，即对比分析哪种算法的MAE和RMSE预测误差均值最小。

实验三：对比分析数据特征组合对预测效果。侧重对比市场主体特征、统计特征、网络招聘特征3大类数据特征组合方式对预测误差MAE的影响，共设计了7种特征组合的对比实验。

实验四：对比分析训练样本提前周期对预测效果的影响。拟探究规模以上工业增加值预测效果与所需训练样本提前周期的关联关系，本研究将提前1期、提前2期、提前3期、提前4期、提前5期的数据特征分别利用RandomForest算法进行训练学习，并对预测误差MAE进行了对比。

实验五：对比分析特征选择对预测效果的影响。拟分析哪些数据特征对某市规模以上工业增加值预测起到更为重要的作用，即探究数据特征对预测效果的贡献程度，拟采用CfsSubset和HarmonySearch相结合的特征选择算法开展了对比实验，即对比只应用自动化筛选后的数据特征和应用全部数据特征的预测效果差异。

#### 3.数据实验

（1）数据特征对预测稳定性影响分析

为验证上述的数据特征对规模以上工业增加值预测效果的稳定性，即在保持数据特征不变的前提下，预测效果不会随着算法的不同和训练集的小部分缺失而导致大幅度波动。本部分首先对完整数据集（07年1月至20年8月逐月数据）进行5次随机抽样操作，每次抽样在完整数据样本量基础上剔随机除掉1个月数据样本，最终形成5个抽样实验数据集，并分别采用LinearRegression、RandomForest、GaussianProcesses（采用RBF径向基函数）、GBRT、ElasticNet和SVM共6种算法测试规模以上工业增加值预测误差范围，实验结果如图24所示。

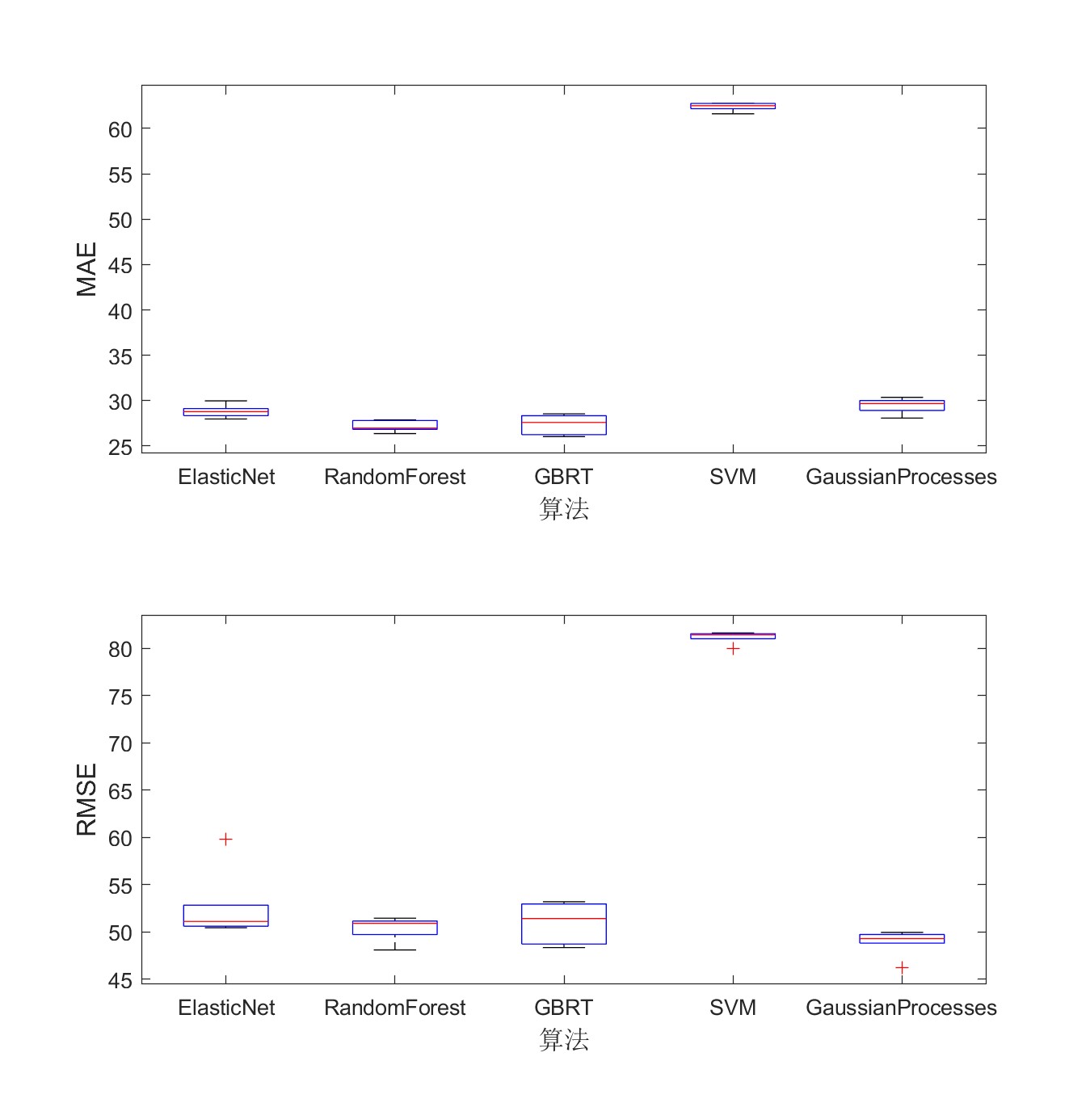


图24 随机抽样数据集对预测效果稳定性的对比结果

其中，由于数据特征相对较多，而预测样本量相对较少，致使LinearRegression算法过拟合现象较为严重，预测结果稳定性表现十分不佳（MAE大致在93.68左右，RMSE大致在181.66左右），因此未列入对比图中。

从图25可以看出，无论从MAE还是RMSE评估指标看，其余5个算法均表现出了预测效果的稳定性，MAE和RMSE数值波动范围不大。即ElasticNet预测结果的MAE最高最低分别为29.98和27.98，RMSE最高最低分别为59.82和50.45；RandomForest的MAE最高最低分别为27.88和26.37，RMSE最高最低分别为51.46和48.11；GBRT预测结果的MAE最高最低分别为28.55和26.04，RMSE最高最低分别为53.20和48.36；SVM的MAE最高最低分别为62.77和61.61，RMSE最高最低分别为81.62和79.95；GaussianProcesses的MAE最高最低分别为30.37和28.08，RMSE最高最低分别为49.97和46.25。

（2）不同算法对预测效果准确性分析

在上述实验基础上，为了进一步对比分析哪种算法对规模以上工业增加值的预测效果最优，即预测误差值最小，本部分分别对比了上述5组随机抽样数据实验中MAE和RMSE的均值表现，如图25所示。

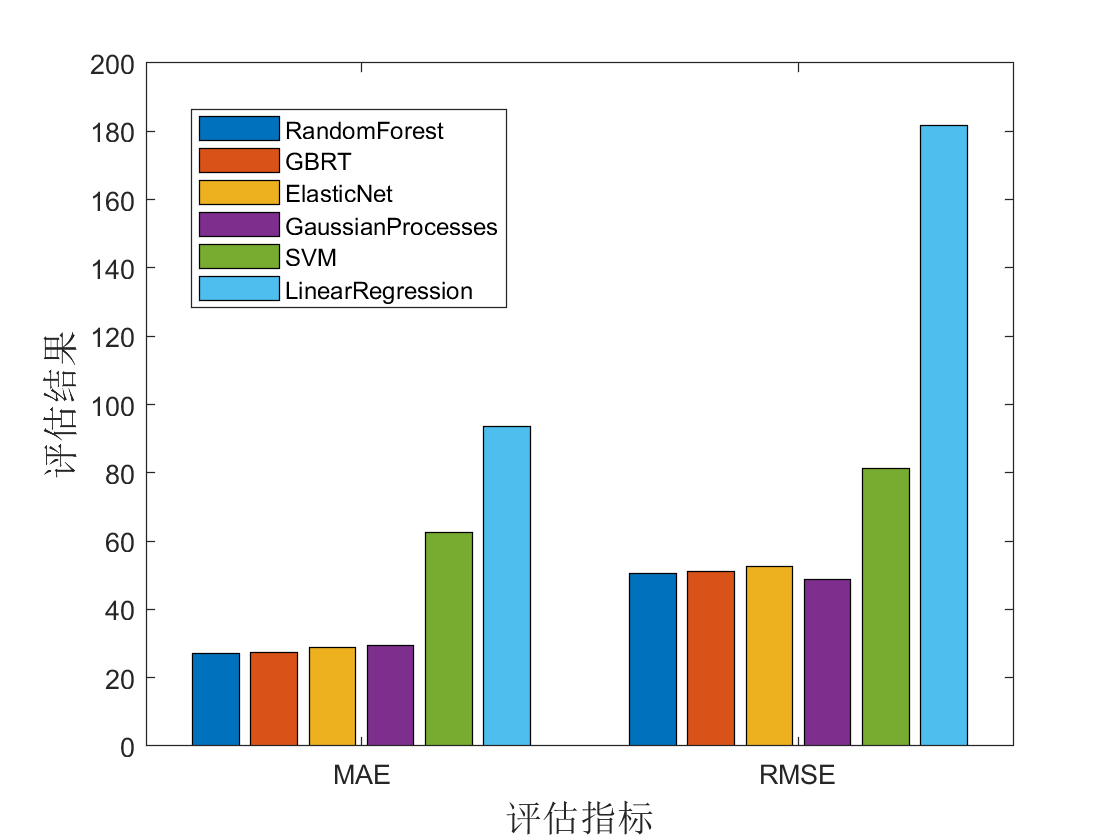


图25 不同算法的预测效果对比分析

可以看出，从MAE和RMSE评估结果看，预测效果从优至劣排序分别为：RandomForest > GBRT > ElasticNet > GaussianProcesses > SVM > LinearRegression，其中，上述四种算法的MAE值分别为：27.14、27.40、28.84、29.46、62.38、93.68，RMSE值分别为：50.39、51.02、52.66、48.90、81.16、181.66，RandomForest算法均表现出了最高的平均预测准确率，即平均绝对误差为27.14，平均相对误差= 平均绝对误差/平均规上工业增加值 = 27.14/210.28 = 12.91%。

（3）数据特征组合对预测效果的影响分析

为分析从不同数据来源提取的数据特征组合对预测效果的影响，本研究针对市场主体特征、统计特征、网络招聘特征3大类数据特征等共设计了7组数据特征组合对比实验，分别为：①全部特征（266个特征），②市场主体特征（162个特征），③统计特征（50个特征），④网络招聘特征（54个特征），⑤市场主体特征+网络招聘特征（216个特征），⑥统计特征+市场主体特征（212个特征），⑦统计特征+网络招聘特征（104个特征），其中，该实验采用的算法为上述实验效果最优的RandomForest算法，实验结果如图26所示。

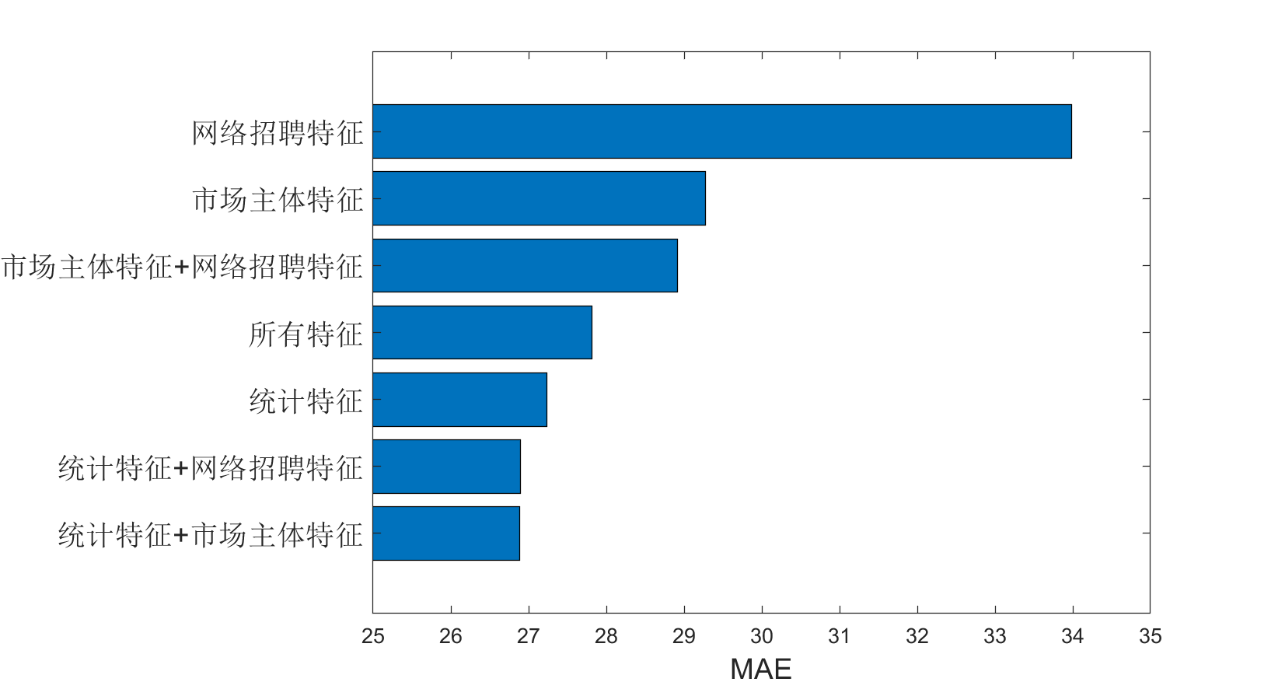


图26 数据特征类别组合对预测效果的影响分析

从图26可以发现，并不是数据特征越多则预测准确性就会越好，也不是从单一数据来源提取的数据特征会导致预测准确性越优，而是采用适当的特征组合方式会有效提升预测的准确性。预测效果从优至劣排序分别为：统计特征+市场主体特征（212个特征）>统计特征+网络招聘特征（104个特征）>统计特征（50个特征）>全部特征（266个特征）>市场主体特征+网络招聘特征（216个特征）>市场主体特征（162个特征）>网络招聘特征（54个特征），其中，上述几种特征组合的MAE值分别为：26.88、26.89、27.23、27.82、28.91、29.28、33.99。

（4）训练样本提前周期对预测效果的影响分析

为分析规模以上工业增加值预测效果与所需训练样本提前周期的关联关系，本研究将提前1期（预测下个月结果）、提前2期（预测2个月后结果）、提前3期（预测3个月后结果）、提前4期（预测4个月后结果）、提前5期（预测5个月后结果）的数据特征分别利用RandomForest算法进行训练学习，并对预测准确性进行了对比分析，实验结果如图27所示。

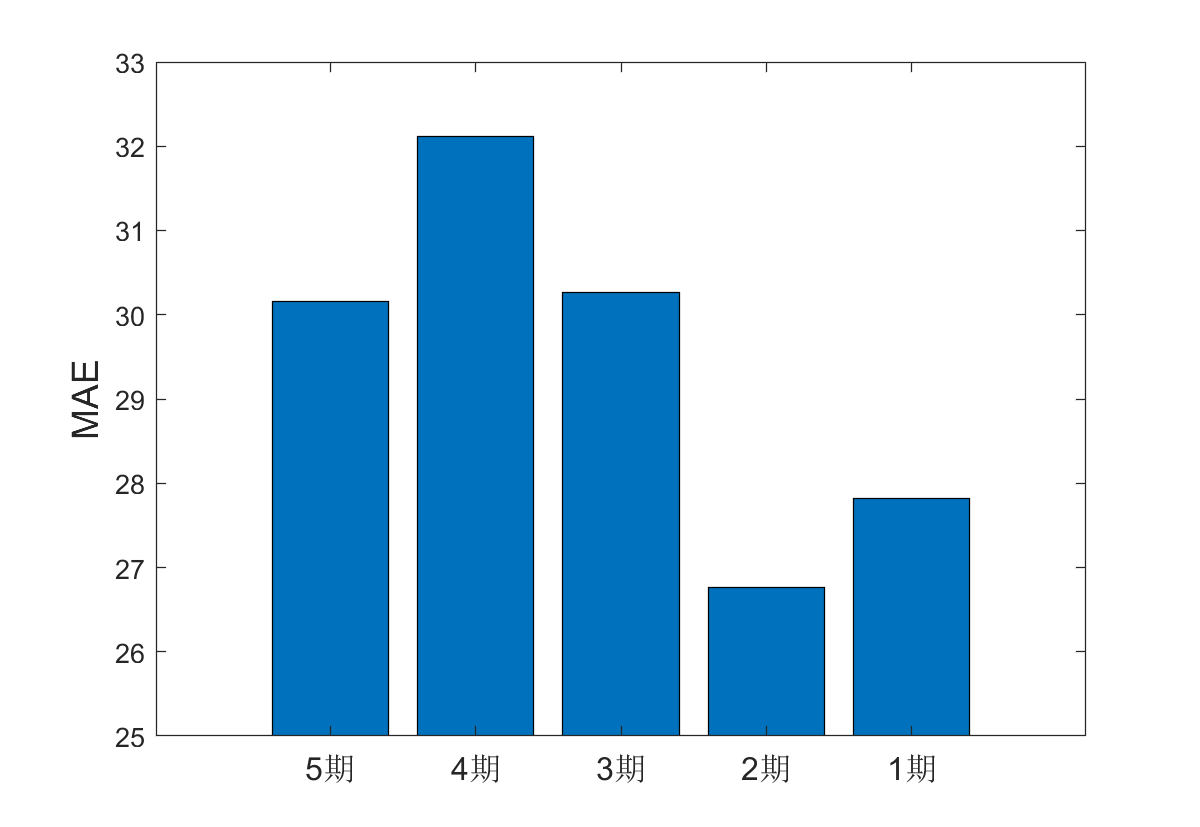


图27 训练样本提前周期对预测效果的影响分析

不难看出，提前1期和2期的预测效果较优，MAE均低于28（提前1期MAE为27.82，提前2期MAE为26.76），但是提前3期以上预测准确性相对较差，MAE均高于30，其中，提前3期MAE为30.27，提前4期MAE为32.11，提前5期MAE为30.15。

（5）特征选择对预测效果影响分析

为分析哪些数据特征对某市规模以上工业增加值预测起到更为重要的作用，即探究数据特征对预测效果的贡献程度，本部分采用CfsSubset和HarmonySearch相结合的特征选择算法开展了对比实验。其中，CfsSubset（基于相关性的特征子集选择）可以根据属性子集中每一个特征之间的关联性以及各个特征的预测能力进行评估；HarmonySearch（和声搜索）类似于遗传算法对生物进化的模仿、模拟退火算法对物理退火的模拟以及粒子群优化算法对鸟群的模仿等，和声算法模拟了音乐演奏的原理，其基本思想是通过反复调整记忆库中的解变量，使函数值随着迭代次数的增加不断收敛，从而来完成优化。

为验证特征选择的有效性，本研究基于CfsSubset和HarmonySearch相结合的特征选择算法，从266个数据特征中自动化选择出57个数据特征用做对比实验，特征选择结果详见表6-5所示。其中，统计特征共11项，分别是存贷-存款余额、存贷-贷款余额、价格-工业品出厂价格指数月度、价格-原材料购进价格指数累计、价格-化工原料类月度、价格-建筑材料及非金属矿类累计、价格-纺织原料类累计、消费-限额以上社会消费品零售总额累计、商品销售-限额以上商品销售总额当月、商品销售-限额以上商品销售总额累计、交通-水路货物运输总量累计值，其余46项均为市场主体特征和网络招聘特征。

表13 特征选择结果

|  |  |
| --- | --- |
| 序号 | 特征名称 |
| 1 | 存贷-存款余额 |
| 2 | 存贷-贷款余额 |
| 3 | 价格-工业品出厂价格指数月度 |
| 4 | 价格-原材料购进价格指数累计 |
| 5 | 价格-化工原料类月度 |
| 6 | 价格-建筑材料及非金属矿类累计 |
| 7 | 价格-纺织原料类累计 |
| 8 | 消费-限额以上社会消费品零售总额累计 |
| 9 | 商品销售-限额以上商品销售总额当月 |
| 10 | 商品销售-限额以上商品销售总额累计 |
| 11 | 交通-水路货物运输总量累计值 |
| 12 | 新增注销企业数环比-农、林、牧、渔业 |
| 13 | 企业存量数环比-制造业 |
| 14 | 新增企业数-制造业 |
| 15 | 新增企业数同比-制造业 |
| 16 | 新增企业数环比-制造业 |
| 17 | 新增注销企业数同比-制造业 |
| 18 | 企业存量数环比-电力、热力、燃气及水生产和供应业 |
| 19 | 新增注销企业数-电力、热力、燃气及水生产和供应业 |
| 20 | 新增招聘岗位数-电力、热力、燃气及水生产和供应业 |
| 21 | 新增招聘岗位数同比-电力、热力、燃气及水生产和供应业 |
| 22 | 新增招聘岗位数环比-电力、热力、燃气及水生产和供应业 |
| 23 | 企业存量数-建筑业 |
| 24 | 企业存量数同比-建筑业 |
| 25 | 新增注销企业数同比-建筑业 |
| 26 | 新增招聘岗位数环比-建筑业 |
| 27 | 新增招聘岗位数同比-批发和零售业 |
| 28 | 新增注销企业数同比-交通运输、仓储和邮政业 |
| 29 | 新增招聘岗位数环比-交通运输、仓储和邮政业 |
| 30 | 新增企业数同比-住宿和餐饮业 |
| 31 | 新增企业数环比-住宿和餐饮业 |
| 32 | 新增注销企业数环比-信息传输、软件和信息技术服务业 |
| 33 | 新增企业数-金融业 |
| 34 | 新增招聘岗位数同比-金融业 |
| 35 | 企业存量数同比-房地产业 |
| 36 | 新增企业数-房地产业 |
| 37 | 新增企业数同比-房地产业 |
| 38 | 企业存量数环比-租赁和商务服务业 |
| 39 | 企业存量数同比-科学研究和技术服务业 |
| 40 | 新增企业数同比-科学研究和技术服务业 |
| 41 | 新增注销企业数-科学研究和技术服务业 |
| 42 | 新增招聘岗位数同比-科学研究和技术服务业 |
| 43 | 新增招聘岗位数环比-科学研究和技术服务业 |
| 44 | 新增企业数同比-水利、环境和公共设施管理业 |
| 45 | 企业存量数环比-居民服务、修理和其他服务业 |
| 46 | 新增企业数-居民服务、修理和其他服务业 |
| 47 | 新增招聘岗位数同比-居民服务、修理和其他服务业 |
| 48 | 企业存量数环比-教育 |
| 49 | 新增企业数同比-教育 |
| 50 | 企业存量数环比-卫生和社会工作 |
| 51 | 新增企业数-卫生和社会工作 |
| 52 | 新增企业数同比-卫生和社会工作 |
| 53 | 企业存量数-文化、体育和娱乐业 |
| 54 | 企业存量数同比-文化、体育和娱乐业 |
| 55 | 新增企业数-文化、体育和娱乐业 |
| 56 | 新增企业数环比-文化、体育和娱乐业 |
| 57 | 新增招聘岗位数-文化、体育和娱乐业 |

实验结果如图28所示，不难发现，基于CfsSubset和HarmonySearch相结合的特征选择算法，只应用自动化筛选后的57个数据特征开展预测MAE误差值（27.11），低于应用266个全部特征的预测结果（27.82），说明有效的特征选择对预测效果能够起到正面影响，表13中的数据特征能够对某市规模以上工业增加值预测起到更为重要的作用。

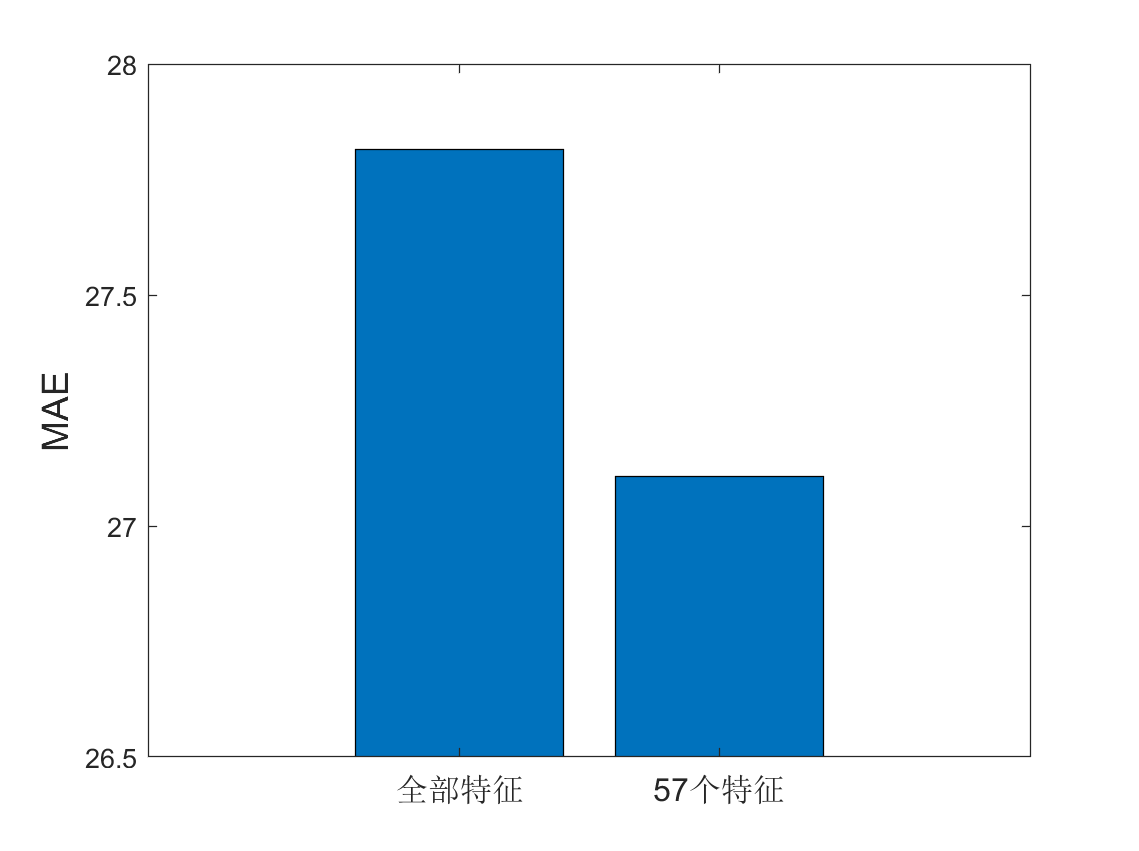


图28 特征选择对预测效果影响分析

#### 4.实验结论

本部分主要结论包括：

一是以规模以上工业增加值为例，验证了利用高频数据和机器学习算法开展某市逐月统计数据预测方法研究方法的可行性和有效性，其中，RandomForest算法表现出了最高的平均预测准确率，即平均绝对误差为27.14，平均相对误差为12.91%。

二是分析了数据特征对预测稳定性影响，分析发现LinearRegression算法出现了过拟合现象，但RandomForest、GaussianProcesses、GBRT、ElasticNet和SVM等5种算法均表现出预测效果的稳定性。

三是分析了不同算法对预测效果准确性，预测效果从优至劣排序分别为：RandomForest > GBRT > ElasticNet > GaussianProcesses > SVM > LinearRegression，其中，MAE值分别为：27.14、27.40、28.84、29.46、62.38、93.68。

四是分析了数据特征组合对预测效果的影响，发现采用适当的特征组合方式会提升预测的准确性。数据特征组合从优至劣排序分别为：统计特征+市场主体特征（MAE为26.88）>统计特征+网络招聘特征（MAE为26.89）>统计特征（MAE为27.23）>全部特征（MAE为27.82）>市场主体特征+网络招聘特征（MAE为28.91）>市场主体特征（MAE为29.28）>网络招聘特征（MAE为33.99）。即在一定程度上，结合某市固定资产投资等历史经济统计数据及企业工商登记注册历史数据可较好地预测某市规模以上工业增加值。

五是分析了训练样本提前周期对预测效果的影响，分析发现本方法对提前1期（MAE为27.82）和提前2期（MAE为26.76）的数据进行训练的预测效果较优，对提前3期以上预测准确性相对较差（MAE均高于30）。即运用历史统计数据、市场主体数据、网络招聘数据在一定程度上可提前反映测下个月或两个月后的规模以上工业增加值变化趋势，而对三个月以后的规模以上工业增加值的影响较小。

六是对比分析特征选择对预测效果的影响，分析发现基于CfsSubset和HarmonySearch相结合的特征选择算法，只应用自动化筛选后的57个数据特征开展预测MAE误差值（27.11）低于应用266个全部特征的预测结果（27.82），说明有效的特征选择对预测效果能够起到正面影响，表13中的数据特征能够对某市规模以上工业增加值预测起到更为重要的作用。

本研究虽尚有不足之处，但期望能够对利用高频数据和机器学习算法预测统计数据方法的研究提供有益借鉴，下一步建议在拓展数据获取渠道、完善特征提取方法、探寻特征选择最优组合、加强深度学习算法研究等方面不断继续深化。

## 七、研究内容之四：基于深度学习方法的宏观经济指标数值预测研究

### （一）研究思路

基于N-BEATS的宏观经济指标数值预测方法采用三层次的多级信号编码，包括全连接神经网络层（FCNN），预测块层和预测栈层，如图29所示。全连接神经网络层实现了预测和拟合两方面的信息提取，分别用于时序预测和信息拟合的参数学习。预测块层之间通过残差数据流连接，使得预测块的拟合函数可对输入进行顺序过滤。预测栈层用于构成预测平台的顶层系统，每个栈被限制为不同的归纳函数，如趋势和周期函数。

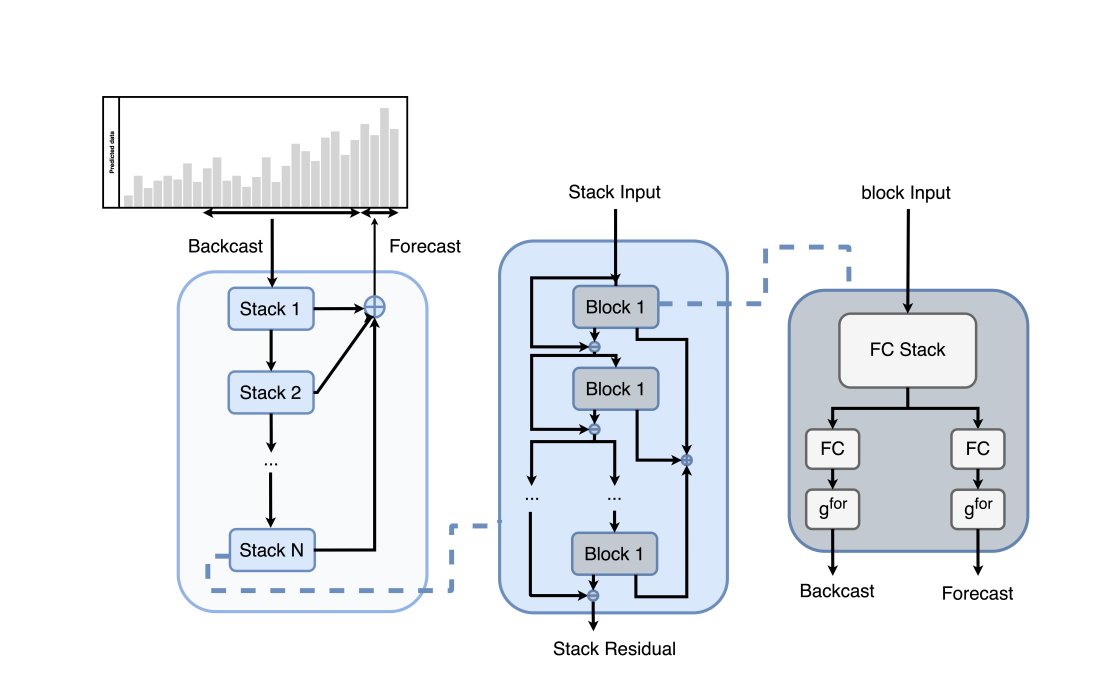
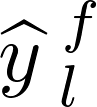
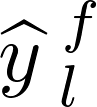
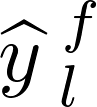


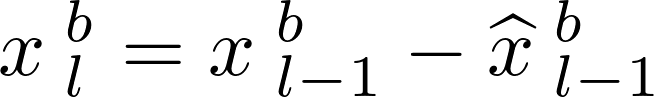
图29 基于深度学习技术的宏观经济监测预测实验平台原理图

### （二）研究方法

本小节旨在提供必要的符号以描述预测模型：X表征被预测的宏观经济数据时间序列，如GDP和PMI。在全连接神经网络层中，预测和拟合两类信息序列输出分别计作Xf和Xb，信息长度分别为H和L。具体而言，Xb的长度是拟合模型中的自回归延迟项，而Xf的长度表征预测的范围，在本平台定为1。

在全连接神经网络层中，每一个预测块根据是否包括外部信息输入接收一或两组输入，并产生两组输出，以第l个块为例，输入为wpsoffice和 El-1 ，输出为 和 。输入数据经过堆叠的全连接神经网络层处理，产生了扩展参数/private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.IxiZSkwpsoffice和 /private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.imDROEwpsoffice的前向拟合和后项预测。下一步是根据l个块计算得出的基向量 vl的线性变换得到转换参数  和 。

在预测块层中采用了残差连接实现了对  和 的后处理工作。前者通过残差相减实现了被处理信号的顺序分解，而后者则整合了当前函数的预测子项，如方程（7.1）与（7.2）所示。

 式（7.1）

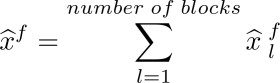
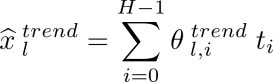
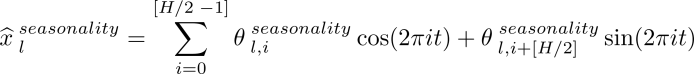
 式（7.2）

表14 参数定义

|  |  |  |
| --- | --- | --- |
| 预测块层和预测栈层的变量 | | |
| X | 被预测的宏观经济参数序列 |  |
| Xf | X的预测窗口向量 | 预测窗口长度为 H. |
| Xb | X的拟合窗口向量 | 拟合窗口长度为 L |
| E | 外部变量序列 |  |
| hl | 第l个预测块的隐藏层 | /private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.QJjGwWwpsoffice  /private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.cYOdhUwpsoffice |
| θl | 对 hl 扩展参数的预测 | wpsoffice  wpsoffice |
| vl | 第l个预测块的基向量 |  |
| /private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.pwYnITwpsoffice | 第 l个预测块的转换参数 | /private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.oLsoitwpsoffice  /private/var/folders/t4/98xjf63j577g5p8p6j31chnr0000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.EcZoUTwpsoffice |
| T | 预测时间序列 | T = [t 0，t 1，t2，...，tH-2，tH-1] |

分块预测机制使得模型的可解释成为了可能。由于预测平台的应用对象是宏观经济预测，模型的可解释基本函数分解和经济可分析变得非常重要。预测平台提供了约束分解的设计，将栈函数转化为可解释的形式[[[247]](#footnote-247)]如多项式函数和调和函数[[[248]](#footnote-248)]，如方程（7.3 7.4）所描述的。逆向分解的设计让宏观经济预测可以分别从发展趋势，经济周期，外部影响等角度提供经济变化的解释。

 式(7.3)

 式(7.4)

### （二）搭建平台

通过对论文中N-BEATS的模型原理、技术体系及平台架构进行研究剖析，初步评估出基于深度学习技术的宏观经济监测预测实验平台所需的最低配置及深度学习技术选型方案。项目组目前已租用了一台内嵌GPU的云服务设备作为基础实验硬件（操作系统：Linux Ubuntu，GPU型号：RTX2080Ti，CPU型号：E5，内存：10G，硬盘：20G），该云平台目前支持一定样本量的数据实验，后续拟视实验数据规模情况再对硬件设备升级扩容。此外，项目组已成功搭建由Facebook人工智能研究院发布的开源深度学习基础平台PyTorch，用以支持N-BEATS模型复现和数据实验。

### （三）数据实验

#### 1.GDP预测对比

在本实验，我们关注传统经济模型BVAR和预测平台对GDP数据的预测。模型的训练数据为中国每个季度的GDP数据，BVAR和预测平台的时间窗口均设为12，即通过累计3年的GDP变化信息进行预测。图30显示了预测平台和BVAR模型的预测和真实GDP曲线，预测平台在对宏观经济发展趋势和周期都更具优势。具体而言，预测平台相对于传统经济模型经济增长速度、变化拐点预测精度上都有显著提高。除此之外，在经济极端环境下，重大波动、疫情动荡时，预测平台展现出了远超过经济学模型的性能，如2020和2021年的GDP预测。

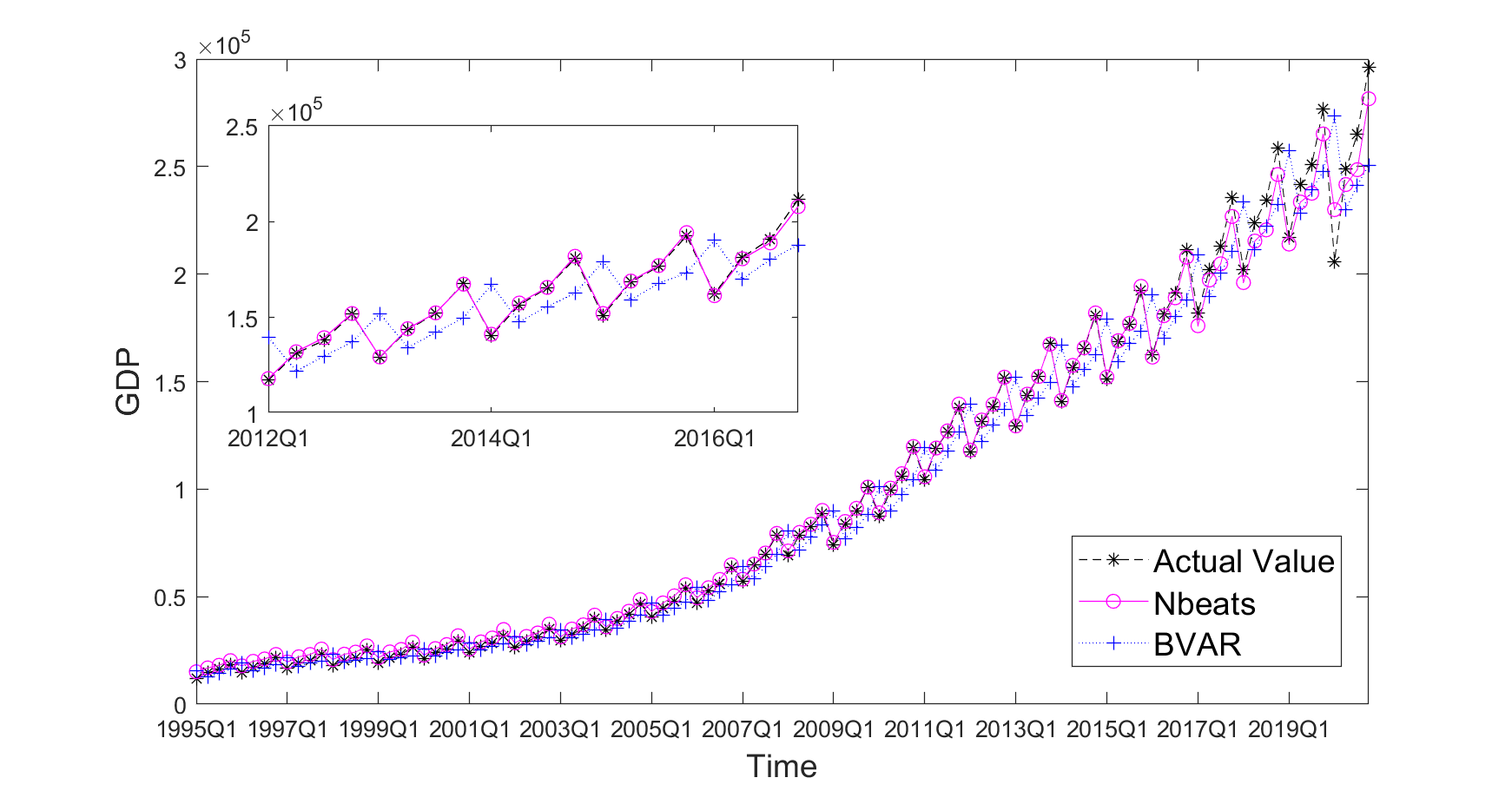


图30 预测平台与BVAR对GDP数据的预测曲线对比

预测平台的预测曲线呈现了突出的性能，为了在统计上量化预测的精度，SMAPE、MAPE和MASE三种误差被测算，如方程（7.5）-（7.7）所示。图31展示了三种误差衡量方法的箱式图绘制，预测平台不但具有更低的误差期望，同时其预测的精度相对于传统经济学方案体现了更高的稳定性。预测平台可以在GDP预测的平均误差和稳定性上，相对于BVAR模型，具有影响力的传统宏观经济预测模型，均保持突出优势。

|  |  |
| --- | --- |
| wpsoffice | 式（7.5） |
| wpsoffice | 式（7.6） |
| wpsoffice | 式（7.7） |

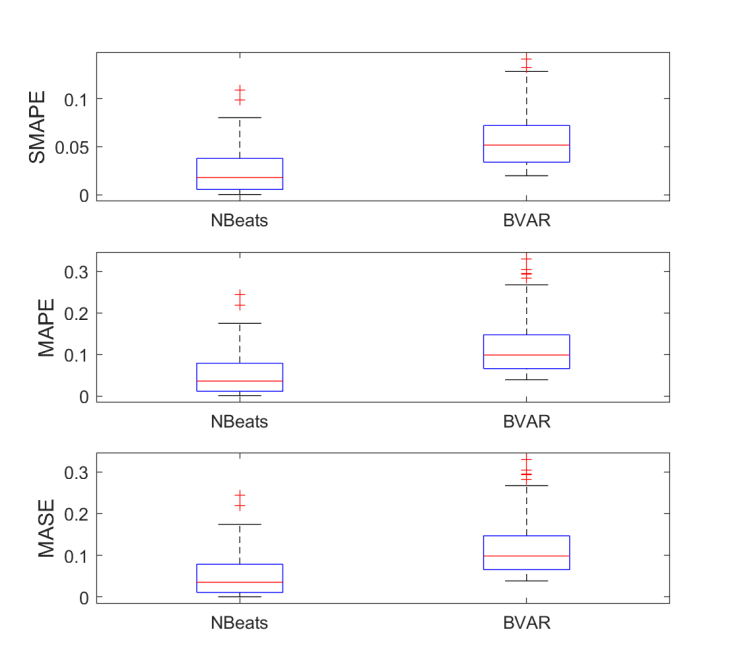


图31 预测平台与BVAR对GDP数据的预测误差箱式图

#### 2.结构优化分析

在本实验中，预测平台针对不同预测数据属性的结构优化被测试。预测平台可以通过不同的函数约束提供不同拟合方式。最终的预测结果通过预测栈函数的串联累加得到。而通过调整不同预测栈函数的构成和组合顺序，预测平台可适应不同被预测数据的信息特征。在GDP预测实验中，预测平台采用的是S-T-G的模型结构组合（S表示周期调谐函数，T表示趋势多项式函数，G表示深度神经网络），而通过结构优化可以进一步提高预测精度。

图32展示了预测平台不同的模型结构在GDP预测中的性能基于SMAPE的误差箱式图被比较。结构优化可以带来预测准确度和稳定性的提升，或两者兼升，如TG结构对GDP预测的优化。需要被强调的是，虽然预测平台是基于深度学习模型，但模型的深度并不一定与预测性能呈正相关，这个结论可以被“STGG并未实现最优性能”所得出。

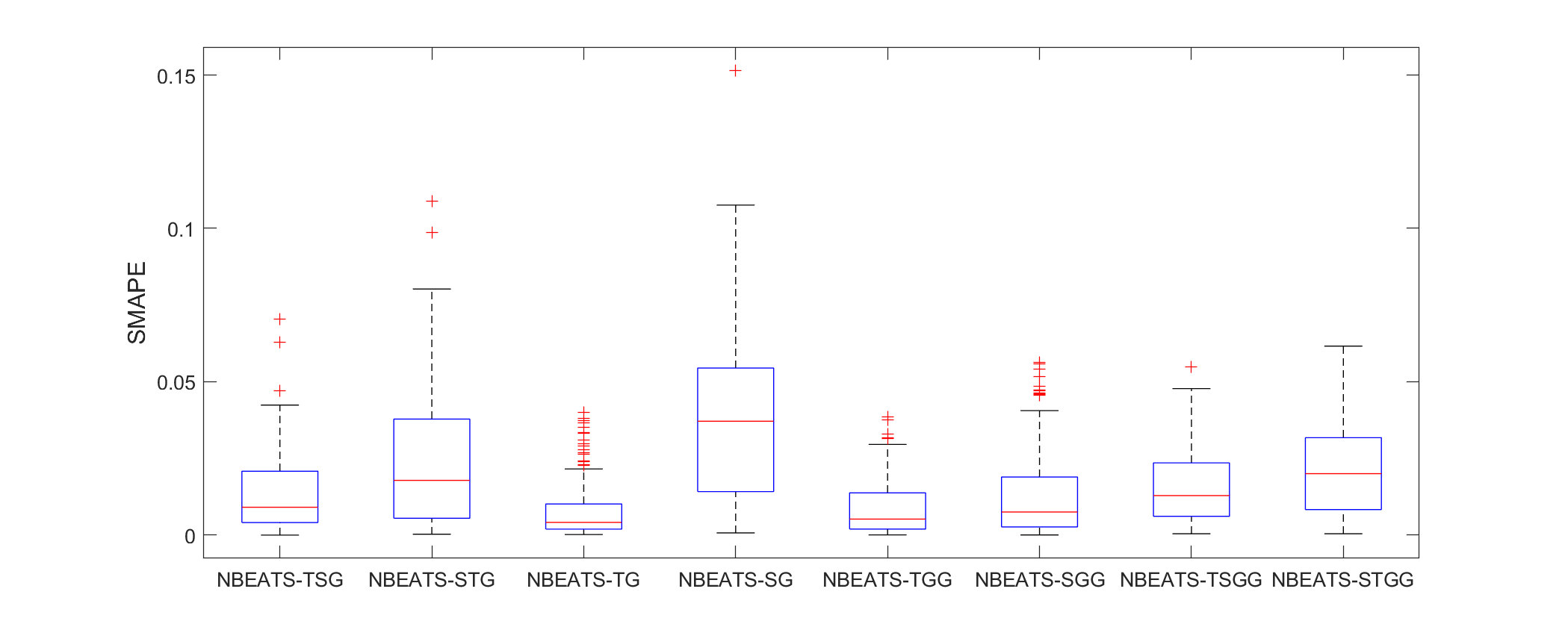


图32 预测平台不同模型结构对GDP预测的性能对比

#### 3.预测频率模型稳健性分析

预测频率在实际任务中需要根据实际被调整，因而预测结果需要对不同的频率需求保持较高的准确性，即具有高预测频率模型稳健性。预测误差通常将降低当更多可用的数据可被用于预测，而现实需求可能仅仅能提供较小的滑动时间窗口，信息损失对预测性能的影响程度是模型结构的重要指标。

图33通过将预测时间窗口从15个数据单位缩减为3个数据单位，观察预测结果误差变化情况。不同模型展现出了不同的预测频率模型稳健性，但即使是将数据滑动窗口缩减到3，预测平台也比窗口为12的BVAR模型少一倍的预测误差。换言之，预测平台仅仅需要四分之一的数据信息，就可以取得比BVAR模型优一倍的预测性能。另外一个发现是滑动窗口在6和12时呈现了两个误差的波谷，这两处性能的提升可能是由于GDP数据时间跨度为半年/一年的整倍数，包含了完整的GDP变化周期信息。

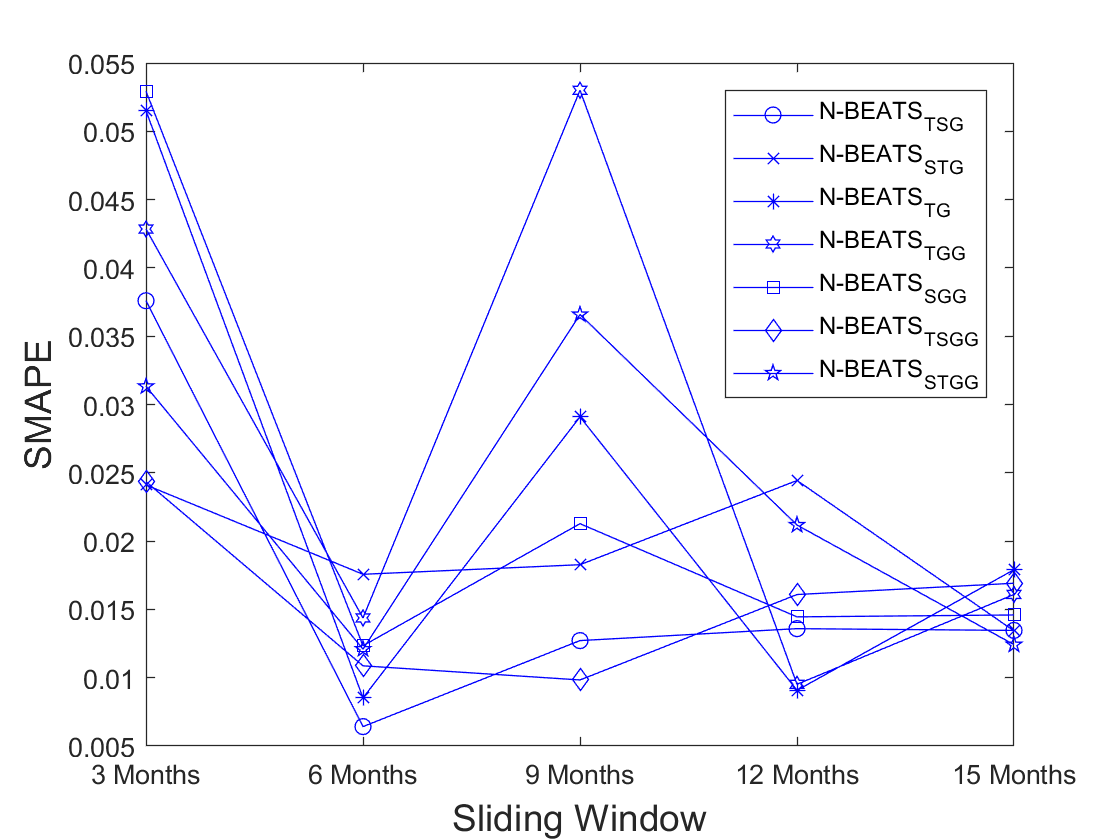


图33 不同结构预测频率模型稳健性对比

#### 4.指标扩展模型稳健性分析

我们的预测平台旨在建立纯数据驱动的经济预测方法，本实验将对除GDP外其他的宏观经济指标进行预测，检验预测平台对于不同宏观经济指标的稳健型。我们选择了两个典型的指标，采购经理总数（PMI）和发电量（ELEC）作为测试。采购经理指数（PMI），是面对全国企业采购经理的调查指数，体现企业对经济的乐观程度。发电量（ELEC）是反映地区工业生产活跃度以及工厂开工率的重要指标。这两个指标分别对应宏观经济中的经济信心和工业基础量，具有典型意义。

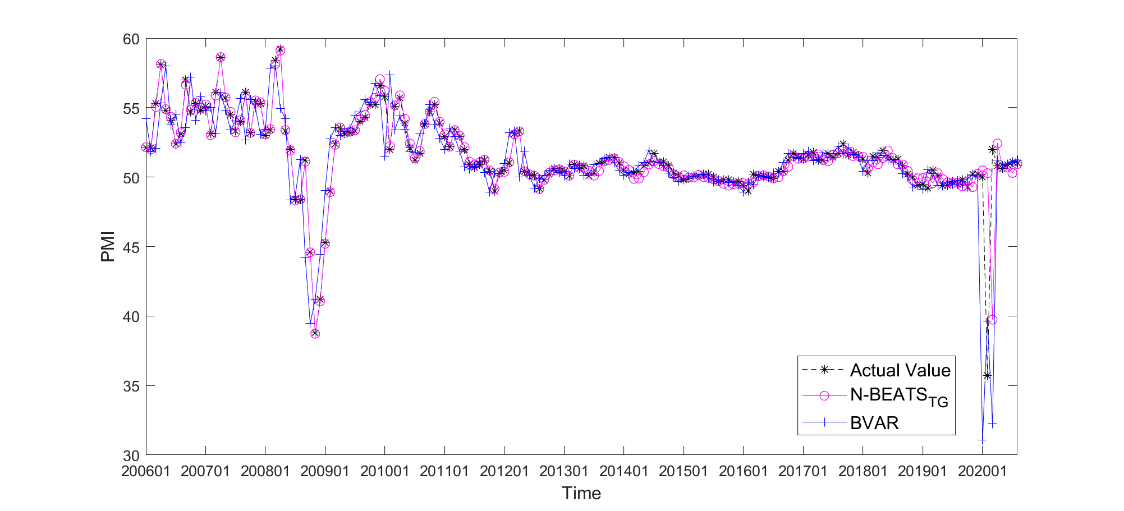


图34 预测平台与BVAR对PMI数据的预测曲线对比

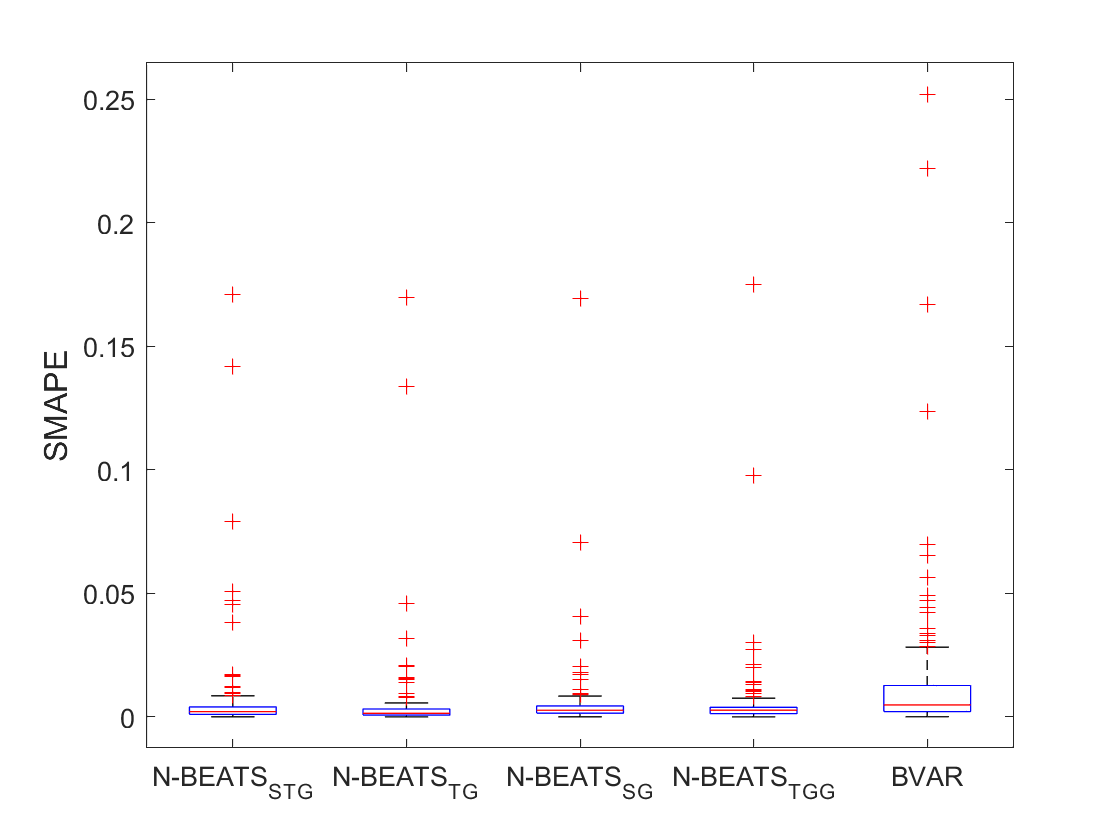


图34 预测平台与BVAR对PMI数据的预测误差箱式图

由图34可以看出，预测平台与BVAR均可对PMI进行较好的预测。预测平台对于实时数据更敏感，BVAR的预测相较于实际数据有明显的滞后。另外在2020年，受疫情影响，经济乐观程度收到重大波动，预测平台对于这类突发异常的分析和对于后续演进的预测显著优于传统经济学方法。

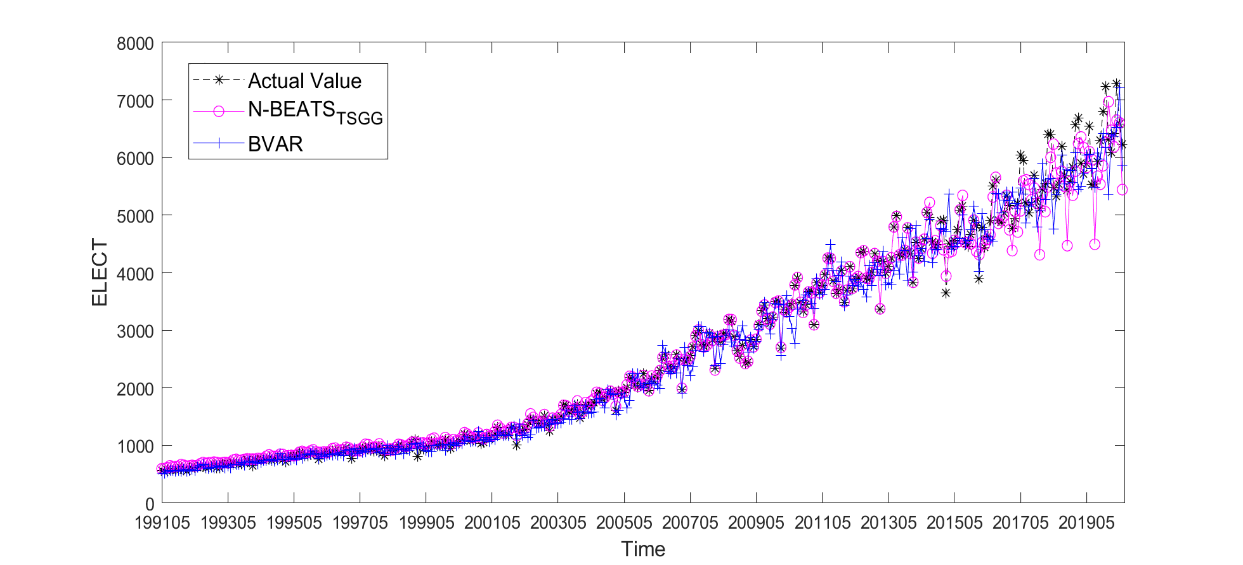


图35 预测平台与BVAR对ELEC数据的预测曲线对比

图35展示了预测平台与BVAR对发电量（ELEC）的预测曲线。从中可以看出，当ELEC呈现较高稳定时，两种方法都可以实现相对高的准确率，且预测平台更优。但是当近年来指标发生非规律波动时，两种方法的预测都出现了失真。图36显示了预测平台多种不同结构对ELEC的预测误差。预测平台的预测更具稳定性，异常点更少，且TSGG结构的预测误差期望小于BVAR。

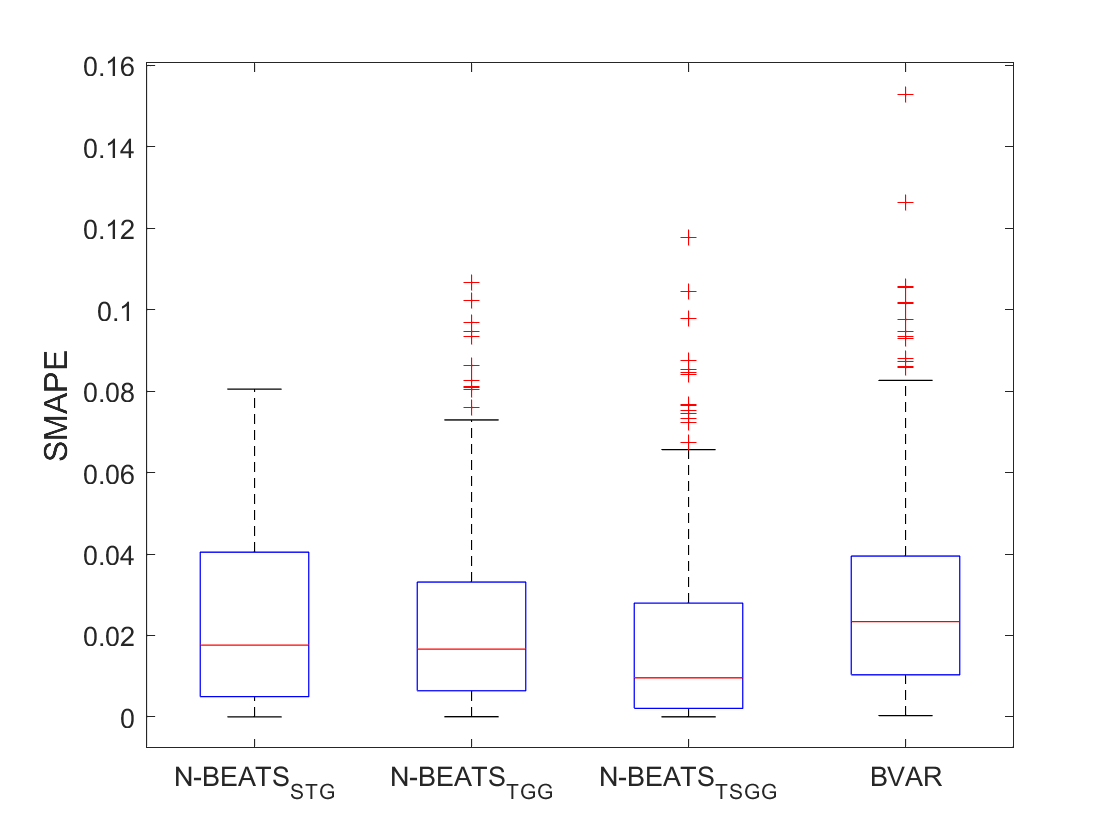


图36 预测平台与BVAR对ELEC数据的预测误差箱式图

图36显示了不同预测平台模型结构对于PMI进行预测的误差情况。预测平台对于PMI预测的误差均值与BVAR相似，但是稳定性显著突出。面对PMI这类波动大的指数，BVAR的预测高误差异常点明显多于预测平台。

#### 5.可解释性分析

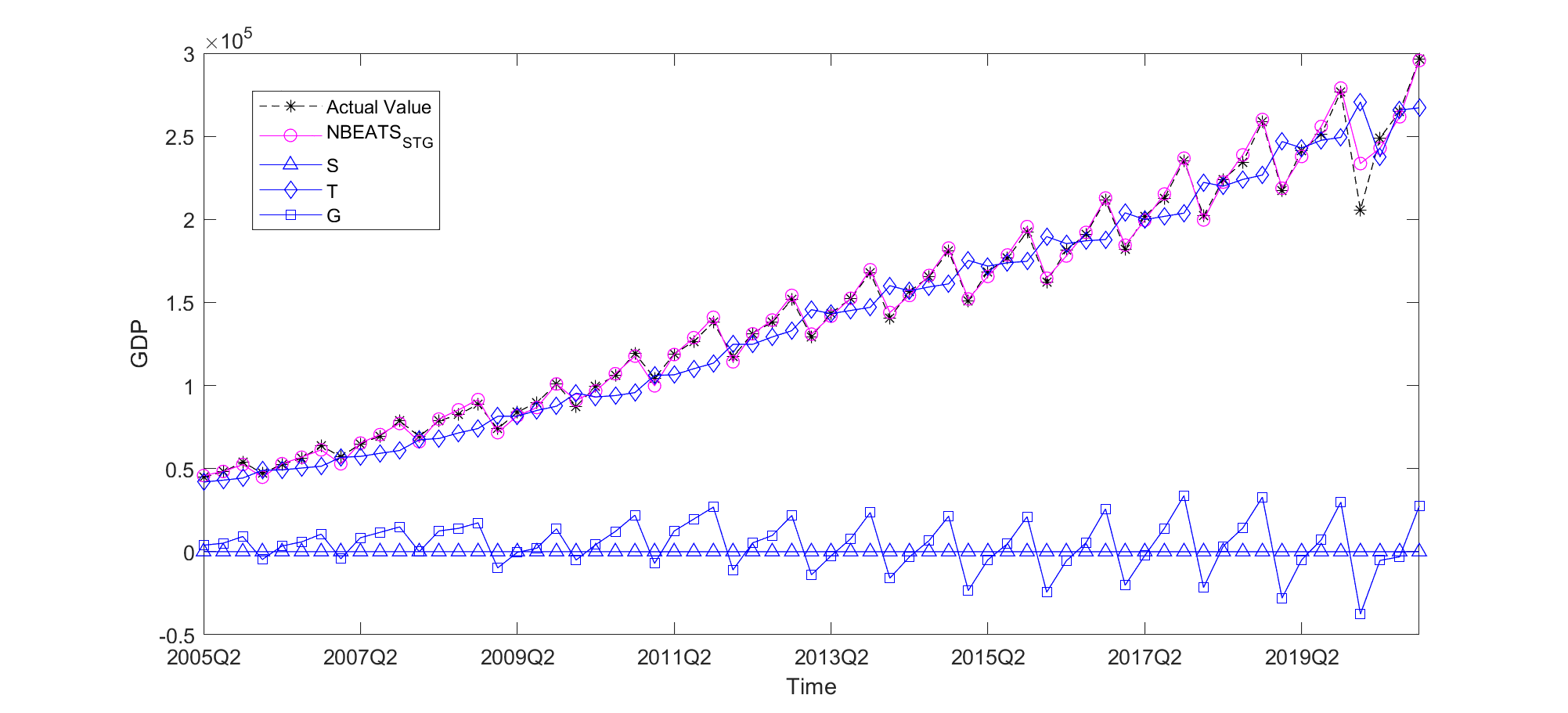


图37 预测平台逆向分解与可解释性

纯数据驱动的深度学习模型往往是“黑箱”，考虑到宏观经济领域的应用需求，预测平台开发了预测的可解释功能。

可解释功能的开发基于预测平台的预测栈模型结构。通过将预测栈约束为可解释函数，并在训练完成后实现模型分解以实现可解释性，如图37所示。T、S、G即为预测栈的约束函数，T是用趋势的方法拟合经济变化，S是用周期的方法拟合，G是采用深度神经网络。

### （四）结论展望

这项工作主要有四方面的意义：首先在工程上，预测平台预测曲线可分解为STG意味着，我们给深度学习网络加了“约束”，即，必须采用多项式函数拟合、而这些约束往往是具有可解释性的（如周期函数和多项式函数）。第二，T和S的存在为模型和经济的对接实现了接口，如，根据拟合出的S模型，提取S的参数，可以反推宏观经济变化的周期，根据拟合出的T模型，可以反推国家经济变化的主体趋势。第三，预测平台可以通过逆向分解分析宏观经济的主导项，如增长趋势和经济周期分别在国家经济变化的中起到了多少的作用。第四，预测平台还为经济学知识的引入提供了可能，如将函数约束设置为某种特定的经济公式，不断提高预测平台预测准确率。

## 八、总结与展望

### （一）总结

**课题组围绕以利用机器学习等大数据算法支撑宏观经济监测预测研究为主线。：一方面构建了理论框架**，方法融合与学科交叉是未来科学研究的大趋势，本研究充分吸收国内外经济预测的理论基础和模型方法精髓，构建了利用机器学习、深度学习技术支撑宏观经济监测预测研究方法理论框架。**另一方面聚焦了成果应用**，鉴于预测传统统计指标对实操性要求较高，本研究利用开源软件搭建实验平台，通过遵循行动研究范式，在推进过程中与有关专家充分互动，切实提升了预测方法的科学性和有效性。本研究主要贡献包括：

**一是开展了基于传统时序预测方法的宏观经济指标预测研究**，依托国家信息中心大数据发展部所掌握的覆盖全国范围1.1亿家企业和个体工商户工商注册、就业招聘、招投标等数据的企业生产经营行为大数据平台，利用传统ARIMA和Prophet模型开展了单变量时序预测实验，利用传统LASSO和BSTS模型开展了多变量时序预测实验，分析发现无论单变量预测还是基于外部数据的多变量预测，传统时序预测模型均有明显的局限性，现有模型对指标的定性预测关注不足，其本质都是以趋势、季节、特殊事件以及数值间的距离为中心的定量预测，而在评估预测结果时，最常用的标准则是数值误差的大小。但在实际预测中，波动模式的预测相对简单，而精准的数值预测难度较大，特别是对于PMI这一类存在参照阈值的指标，如果通过预测值无法判断实际值未来是高于还是低于荣枯线，那么预测结果的价值将大打折扣。

**二是开展了基于机器学习方法的宏观经济指标分类预测研究**，针对上述传统方法存在突出的问题，本研究提出“三库碰撞”方法，尝试通过机器学习分类算法判断未来预测值的性质。课题组从国家统计局官方网站获取月度统计数据并整理构成监测指标库，共11类489个指标84034条数据，用以作为“三库碰撞”实验的预测目标。之后，从企业生产经营行为大数据平台提取出抽取6项企业数据，按地域（31个省份）、性质（国企、民企、外企）、行业分类（20个行业大类）进行多种数值变换处理，经过数据抽取维度、数据分类维度和数值加工处理维度等交叉匹配，最终形成共1296个数据特征，用作机器学习分类预测的数据特征库。实验结果发现，与监测指标的历史上升下降或荣枯背景分布概率相比，该方法对大部分经济指标T+1预测的准确率均有明显提升（准确率最高提升33%），从目前实验结果看，“三库碰撞”方式可适用于大部分经济指标的升降预测或荣枯预测，特别是，对于建筑材料及非金属矿类购进价格指数和制造业采购经理指数等部分经济指标，在使用该方法前提下随着训练时间的推移预测准确率能够持续提升。

**三是开展了基于机器学习方法的宏观经济指标数值预测研究**，针对大部分宏观经济指标历史数据较为缺乏、可获取的宏观经济指标影响因素数据较多等特点，可兼容自变量多、因变量少的机器学习预测方法为宏观经济指标数值预测研究提供了新的解决思路。课题组对融合机器学习算法与多重数据特征的数值型预测方法开展了研究。一方面是白条鸡日均交易量数值预测研究，该研究在数据特征选取过程中，兼顾鸡禽类产业、互联网舆情信息、网民需求意愿、白条鸡相关统计数据等多种趋势变化，综合考虑多方面因素对白条鸡交易量的影响，共提取了50个可能与白条鸡日均交易量预测有关的变量作为数据特征，包括市场主体特征30个、舆情信息特征2个、搜索意愿特征14个和统计数据特征4个。实验结果显示，除线性回归算法在预测效果稳定性表现不佳外，其余三个算法均具有预测效果的稳定性，且预测效果从优至劣依次为：GBRT、RandomForest、ElasticNet、LinearRegression，随着时间不断积累以及训练样本数量不断增加，白条鸡日均交易量预测效果不断提升。另一方面是某市规模以上工业增加值数值预测研究，课题组选定某市规模以上工业增加值数据作为因变量，选定某市统计数据、某市企业工商登记注册数据、某市互联网招聘数据作为自变量构建预测函数，采用平均绝对误差（MAE）和均方根误差（RMSE）进行效果评估，设计了数据特征、模型算法、训练样本、提前周期4组对比实验。实验结果表明利用高频数据和机器学习算法开展某市逐月统计数据预测表现出了最高的平均预测准确率，即利用机器学习和多重数据特征的方法对某市规模以上工业增加值的预测起到重要作用。

**四是开展了基于可解释性深度学习方法的宏观经济数值预测研究**，课题组选用了全球目前较为先进的N-BEATS深度学习模型开展宏观经济数值预测研究，通过对N-BEATS的模型原理、技术体系及平台架构进行研究剖析，我们搭建了基于N-BEATS深度学习技术的宏观经济监测预测实验平台，该预测平台采用三层次的多级信号编码，包括全连接神经网络层（FCNN）、预测块层和预测栈层，其中，全连接神经网络层实现了预测和拟合两方面的信息提取，分别用于时序预测和信息拟合的参数学习；预测块层之间通过残差数据流连接，使得预测块的拟合函数可对输入进行顺序过滤；预测栈层用于构成预测平台的顶层系统，每个栈被限制为不同的归纳函数，如趋势和周期函数。根据数据实验结果持续对N-BEATS模型进行参数调优，开展与目前较为流行的BVAR等传统时序预测模型对比实验。结果表明在经济极端环境下，如重大事件产生的波动或疫情动荡时，N-BEATS模型展现出了远超过经济学模型的性能。此外也为经济学知识的引入提供了可能，如将函数约束设置为某种特定的经济公式，可进一步提高N-BEATS模型预测准确率。

### （二）展望

未来，应吸收借鉴美国、澳大利亚、新加坡等国通过归集数据开展宏观决策的成功经验，充分发挥我国独特的制度优势，有效归集和开发利用全社会范围内经济运行相关数据，全面应用机器学习、深度学习等新技术，建设集宏中微观经济运行分析于一体的“国家经济大脑”，助力实现国家治理体系和治理能力现代化。

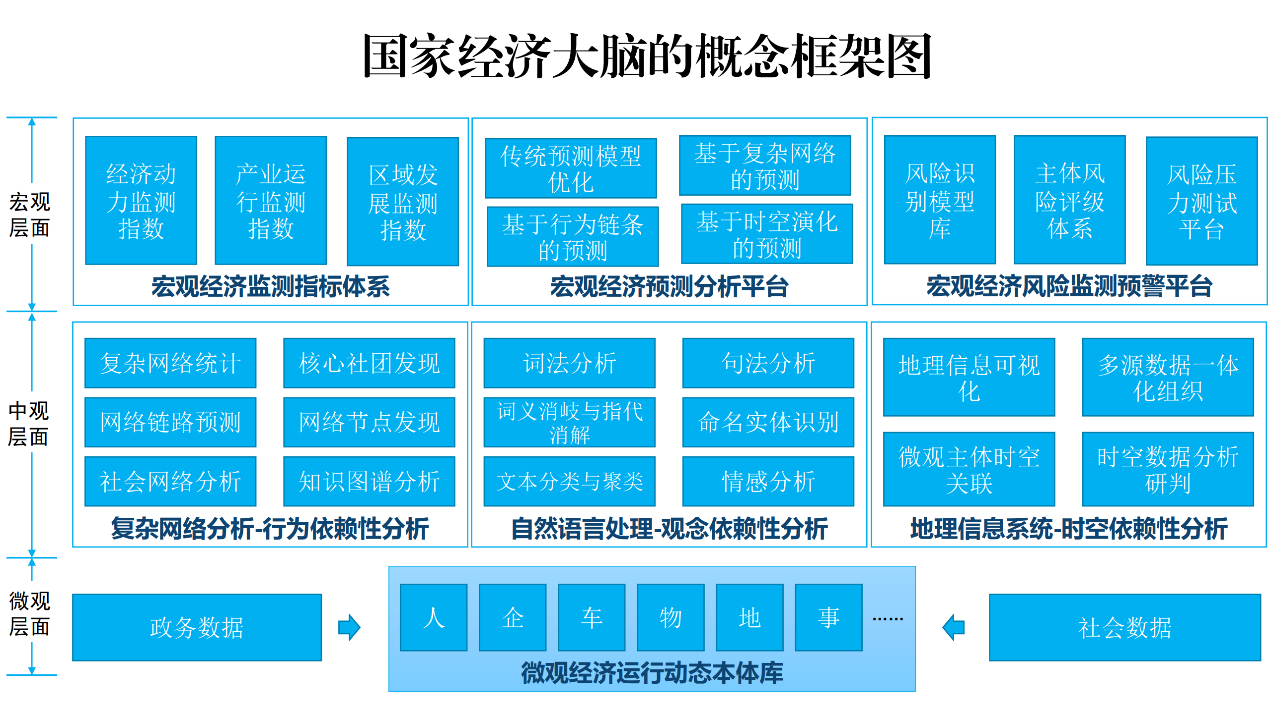


图38 国家经济大脑的总体框架

“国家经济大脑”概念框架如图38所示，一是在国家经济大脑建设的微观层面，应当着眼于利用大数据手段快速构建领域本体和通用本体，形成对海量微观主体行为演变和关系网络的快速挖掘能力。二是构建国家经济大脑的中观系统，要整合复杂网络、自然语言理解和时空分析三大类算法模型，将其作为模拟仿真中观经济现象的技术支撑。三是基于微观层面经济动态本体和中观层面经济涌现仿真分析平台的坚实基础，在宏观层面，可以进一步围绕经济运行和重大风险防范等需求，依托机器学习、深度学习等技术构建经济运行监测、经济预测和风险识别“三位一体”的宏观经济监测预测大数据平台。

## 参考文献

1. 国务院关于印发促进大数据发展行动纲要的通知[EB/OL].http://www.gov.cn/zhengce/content/2015-09/05/content\_10137.htm, 2015-09.
2. 中共中央关于深化党和国家机构改革的决定[EB/OL].http://www.gov.cn/xinwen/2018-03/04/content\_5270704.htm, 2018-03.
3. 中共中央关于坚持和完善中国特色社会主义制度 推进国家治理体系和治理能力现代化若干重大问题的决定[EB/OL]. http://www.gov.cn/zhengce/2019-11/05/content\_5449023.htm, 2019-11.
4. Clarke A, Margetts H. Governments and citizens getting to know each other? Open, closed, and big data in public management reform[J]. Policy & Internet, 2014, 6(4): 393-417.
5. 习近平. 坚持开放包容 推动联动增长——在二十国集团领导人汉堡峰会上关于世界经济形势的讲话.[EB/OL]. http://jhsjk.people.cn/article/29391615, 2017-07.
6. Varian H R. Big data: New tricks for econometrics[J]. Journal of Economic Perspectives, 2014, 28(2): 3-28.
7. 冯.诺依曼. 数学在科学和社会中的作用[M]. 大连：大连理工大学出版社. 2009: 58.
8. 陈禹, 方美琪. 复杂性研究视角中的经济系统[M]. 北京：商务印书馆. 2015: 52-53.
9. Big Data is a Big Deal[EB/OL][2012-04-20]. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012-03.
10. 张勇进,王璟璇. 主要发达国家大数据政策比较研究[J]. 中国行政管理. 2014(12): 114-117.
11. Chen J C, Dunn A, Hood K, et al. Off to the races: A comparison of machine learning and alternative data for predicting economic indicators[M]. Big Data for 21st Century Economic Statistics. University of Chicago Press, 2019.
12. Konny C G, Williams B K, Friedman D M. Big Data in the US Consumer Price Index: Experiences and Plans[M]. Big Data for 21st Century Economic Statistics. University of Chicago Press, 2019.
13. Cuffe J, Bhattacharjee S, Etudo U, et al. Using Public Data to Generate Industrial Classification Codes[M]. Big Data for 21st Century Economic Statistics. University of Chicago Press, 2019.
14. 王建冬, 童楠楠, 易成岐. 大数据时代公共政策评估的变革：理论、方法与实践[M]. 北京：社科文献出版社. 2019: 54-55.
15. Vintar M. Information Polity: The International Journal of Goverment and Democracy in the Information Age[M]. IOS Press, 2006.
16. Hundt R. From Carbon to Clean—How to Attract Investment in Smart Grid Infrastructures[M]. Broadband Networks, Smart Grids and Climate Change. Springer, New York, NY, 2013: 39-45.
17. Clarke A, Margetts H. Governments and citizens getting to know each other? Open, closed, and big data in public management reform[J]. Policy & Internet, 2014, 6(4): 393-417.
18. Milakovich M E. Anticipatory government: Integrating big data for smaller government[J]. Internet, politics, policy 2012: Big data, big challenges, 2012.
19. 陈敏尔：让智能化为经济赋能为生活添彩[EB/OL] http://difang.gmw.cn/cq/2019-07/03/content\_32968465.htm, 2019-07.
20. 牛文胜, 董晓博, 吴又奎. 政务信息资源目录及宏观经济大数据系统的建设及应用[J]. 价值工程, 2018, 37(20): 251-255.
21. 《2018年海南省经济社会发展大数据分析报告》出炉. http://www.hainan.gov.cn/hainan/tingju/201901/dbebcc4ad3624e4b812057a1b6712309.shtml,2019-01.
22. 钱斌华. 税收大数据预测经济走势的某市经验[J]. 某市日报, 2018: 03-15.
23. Einav L, Levin J. Economics in the age of big data[J]. Science, 2014, 346(6210).
24. Balasubramanian N, Sivadasan J. What happens when firms patent? New evidence from US economic census data[J]. The Review of Economics and Statistics, 2011, 93(1): 126-146.
25. 马建堂. 大数据在政府统计中的探索与应用[M]. 北京：中国统计出版社, 2013.
26. 徐先华. 大数据时代的经济普查数据分析与探究[J]. 时代金融, 2018(11): 49-50.
27. 杜成琳. 大数据时代我国经济统计现状及趋势研究[J]. 知识经济, 2018(14): 14+16.
28. 张维扬. 大数据时代我国经济统计现状及趋势研究[J]. 现代经济信息, 2018(01): 104.
29. 李勇, 罗良清, 张敏, 李禹锋. 大数据时代我国经济统计现状及趋势研究[J]. 中国统计, 2016(12): 21-23.
30. 吴毅. 大数据环境下城市经济统计调查分析及改进措施[J]. 经贸实践, 2017(24): 71.
31. 王怡博, 方颀卓. 浅析云数据分析在经济统计中的应用[J]. 中国市场, 2018(03): 189-190.
32. 祝旭. 经济统计中大数据分析的应用研究[J]. 无线互联科技, 2018, 15(17): 110-111.
33. 何海. 大数据在区域经济统计中的应用研究[J]. 现代营销(下旬刊), 2018(01): 102.
34. 解路英. 大数据使“定制”新的经济指标成为可能[N]. 经济参考报, 2014-07-11(005).
35. 李世伟. 大数据环境下国民经济核算体系的改进[J]. 市场经济与价格, 2015(07): 10-14.
36. 李红艳, 汪涛. 大数据时代背景下的新型国民经济核算体系研究[J]. 经济视角(上), 2013(08): 24-26.
37. 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报(人文社会科学), 2017, 17(03): 54-59.
38. 刘涛雄, 徐晓飞. 大数据与宏观经济分析研究综述[J]. 国外理论动态, 2015(01): 57-64.
39. Economist. Keqiang ker-ching: How China’s next prime minister keeps tabs on its economy[EB/OL]. http://www.economist.com/node/17681868, 2010-12.
40. 骆兆华. 经济孰冷孰热 “挖掘机指数”有的聊[N]. 中国城乡金融报,2016-01-15(B03).
41. 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报(人文社会科学), 2017, 17(03): 54-59.
42. 陈沁, 沈明高, 沈艳. 财智BBD中国新经济指数技术报告[EB/OL].http://www.nsd.edu.cn/teachers/professorNews/2016/0304/ 25596.html, 2016-03.
43. 国家信息中心发布《中国大数据发展报告(2017)》[EB/OL]. <http://www.sic.gov.cn/News/79/7727.htm,> 2017-02.
44. 国家信息中心发布《中国共享经济发展年度报告（2019）》[EB/OL]. <http://www.sic.gov.cn/News/79/9907.htm,> 2019-03
45. 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究:前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104.
46. 张秋雁, 宋强. 基于用电大数据的经济分析平台设计研究[J]. 电力大数据, 2017, 20(08): 6-9.
47. 王凯军, 龙厚印, 吴良良, 石清. 基于电力大数据的产业结构调整及经济指标研究[J]. 经济研究导刊, 2017(25): 38-39.
48. Doll C N H, Muller J P, Morley J G. Mapping regional economic activity from night-time light satellite imagery[J]. Ecological Economics, 2006, 57(1): 75-92.
49. [] Elvidge C D, Baugh K E, Kihn E A, et al. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption[J]. International Journal of Remote Sensing, 1997, 18(6): 1373-1379.
50. ŠĆEPANOVIĆ S, MISHKOVSKI I, HUI P, et al. Mobile phone call data as a regional socio-economic proxy indicator[J]. PLoS ONE, 2015, 10(4): e0124160.
51. EAGLE N, MACY M, CLAXTON R. Network diversity and economic development[J]. Science, 2010, 328(5981): 1029-1031.
52. HOLZBAUER B O, SZYMANSKI B K, NGUYEN T, et al. Social ties as predictors of economic development[M]// WIERZBICKI A, BRANDES U, SCHWEITZER F, et al. Advances in Network Science. Switzerland: Springer International Publishing, 2016: 178-185.
53. Liu J H, Wang J, Shao J, et al. Online social activity reflects economic status[J]. Physica A: Statistical Mechanics and its Applications, 2016, 457: 581-589.
54. 曲延玲. 基于大数据构建经济金融深度监测分析体系[J]. 金融电子化, 2017(05): 40-41.
55. Kholodilin K A, Podstawski M, Siliverstovs B, et al. Google searches as a means of improving the nowcasts of key macroeconomic variables[J]. 2009.
56. Della Penna N, Huang H. Constructing consumer sentiment index for US using Google searches[R]. 2010.
57. Suhoy T. Monthly assessments of private consumption[R]. Bank of Israel, 2010.
58. Vosen S . A monthly consumption indicator for Germany based on Internet search query data[J]. Applied Economics Letters, 2012, 19(7): 683-687.
59. 李映桥. 网络消费者信心与宏观经济波动关系研究[D].吉林大学,2017.
60. 孙毅, 吕本富, 陈航, 薛添. 基于网络搜索行为的消费者信心指数构建及应用研究[J]. 管理评论, 2014, 26(10): 117-125.
61. 刘伟江, 李映桥. 网络消费者信心指数和经济增长的动态相关性研究[J]. 财贸研究, 2017, 28(05): 1-10.
62. 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报（人文社会科学）, 2017, 17(03): 54-59.
63. 孙毅, 吕本富, 陈航, 等. 大数据视角的通胀预期测度与应用研究[J]. 管理世界, 2014, 4: 171-172.
64. Cavallo A , Rigobon R . The Billion Prices Project: Using Online Prices for Measurement and Research[J]. The Journal of Economic Perspectives, 2016: 151-178.
65. 袁铭. 基于网购搜索量的CPI及时预测模型[J]. 统计与信息论坛, 2015, 30(4): 20-24.
66. Müller R, Herren H M, Röthlisberger S, et al. Recent developments in the Swiss CPI: scanner data, telecommunications and health price collection[C]//9th Ottawa Group Meeting on Prices 14th to 16th May, London. 2006.
67. Guzman G. Internet search behavior as an economic forecasting tool: The case of inflation expectations[J]. Journal of economic and social measurement, 2011, 36(3): 119-167.
68. 徐映梅, 高一铭. 基于互联网大数据的CPI舆情指数构建与应用——以百度指数为例[J]. 数量经济技术经济研究, 2017, 34(1): 94-112.
69. 张崇, 吕本富, 彭赓等. 网络搜索数据与CPI的相关性研究[J]. 管理科学学报, 2012: 50-59+ 70.
70. Ettredge M, Gerdes J, Karuga G. Using web-based search data to predict macroeconomic statistics[J]. Communications of the ACM, 2005, 48(11): 87-92.
71. Askitas N, Zimmermann K F. Google Econometrics and Unemployment Forecasting[J]. Applied Economics Quarterly, 2009, 55(2): 107.
72. D'Amuri F. Predicting unemployment in short samples with internet job search query data[R]. University Library of Munich, Germany, 2009.
73. Suhoy T. Monthly assessments of private consumption[R]. Bank of Israel, 2010.
74. McLaren N, Shanbhogue R. Using internet search data as economic indicators[J]. Bank of England Quarterly Bulletin, 2011 (2011): Q2.
75. Choi H, Varian H. Predicting the present with Google Trends[J]. Economic record, 2012, 88: 2-9.
76. Barreira N, Godinho P, Melo P. Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends[J]. NETNOMICS: Economic Research and Electronic Networking, 2013, 14(3): 129-165.
77. Pavlicek J, Kristoufek L. Can Google searches help nowcast and forecast unemployment rates in the Visegrad Group countries?[J]. arXiv preprint arXiv:1408.6639, 2014.
78. Su Z. Chinese online unemployment-related searches and macroeconomic indicators[J]. Frontiers of Economics in China, 2014, 9(4): 573-605.
79. Pavlicek J, Kristoufek L. Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries[J]. PloS one, 2015, 10(5): e0127084.
80. 王勇, 董恒新. 大数据背景下中国季度失业率的预测研究------基于网络搜索数据的分析[J]. 系统科学与数学, 2017, 37(2): 460-472.
81. 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(04): 625-633.
82. 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究：前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104.
83. Toole J L, Lin Y R, Muehlegger E, et al. Tracking employment shocks using mobile phone data[J]. Journal of The Royal Society Interface, 2015, 12(107): 20150185.
84. Llorente A, Garcia-Herranz M, Cebrian M, et al. Social media fingerprints of unemployment[J]. PloS one, 2015, 10(5): e0128692.
85. PAPPALARDO L, VANHOOF M, GABRIELLI L, et al. Estimating economic development with mobile phone data[EB/OL]. (2016-05-30). http://www.cisstat.com/ BigData/CIS-BigData\_08\_Eng%20%20IT%20Luca%20Pa ppalardo%20Et%20Al%20Estimating%20Economic%20D evelopment.pdf.
86. 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报（人文社会科学）, 2017, 17(03): 54-59.
87. 王智, 余桂南, 赵路. 基于大数据的我国新经济服务业人才需求景气指数研究[J]. 2016年全国统计建模大赛论文集, 2016.
88. Webb G K. Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures[J]. Issues in Information Systems, 2009, 10(2): 82.
89. Wu L, Brynjolfsson E. The future of prediction: How Google searches foreshadow housing prices and sales[M]//Economic analysis of the digital economy. University of Chicago Press, 2015: 89-118.
90. Kulkarni R, Haynes K E, Stough R R, et al. Forecasting housing prices with Google econometrics[J]. GMU School of public policy research paper, 2009 (2009-10).
91. Widgrén J. Predicting Housing Prices with Google Searches in Finland[R]. The Research Institute of the Finnish Economy, 2016.
92. 董倩, 孙娜娜, 李伟. 基于网络搜索数据的房地产价格预测[J]. 统计研究, 2014, 31(10): 81-88.
93. 洪涛, 厉伟. 基于网络搜索数据的住房价格预期与实际价格波动分析[J]. 统计与信息论坛, 2015: 49-53.
94. Choi H, Varian H. Using search engine data for nowcasting—an illustration[J]. Actes des Rencontrees Economiques, 2011: 535-538.
95. Wu L, Brynjolfsson E. The future of prediction: how Google searches foreshadow housing prices and quantities[J]. ICIS 2009 Proceedings, 2009: 147.
96. Webb G K. Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures[J]. Issues in Information Systems, 2009, 10(2): 82.
97. McLaren N, Shanbhogue R. Using internet search data as economic indicators[J]. Bank of England Quarterly Bulletin, 2011: Q2.
98. 霍琳, 尚维, 徐山鹰. 房地产开源舆情指数构建与政策影响研究[J]. 信息系统学报, 2013, 2: 57-66.
99. 王钧超. 大数据时代产业经济信息分析及在宏观决策中的应用[D]. 中国地质大学（北京）, 2016.
100. Li Y, Arora S, Youtie J, et al. Using web mining to explore Triple Helix influences on growth in small and mid-size firms[J]. Technovation, 2018, 76: 3-14.
101. Daepp M I G, Hamilton M J, West G B, et al. The mortality of companies[J]. Journal of The Royal Society Interface, 2015, 12(106): 20150120.
102. 毛红卫, 阮伯虎, 吴铭宇等. 诸暨市大数据经济运行监测与研究[J]. 民营科技, 2018(02): 182-185.
103. 王展. 大数据经济运行系统的研究[D]. 郑州大学, 2017.
104. Neffke F, Henning M, Boschma R. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions[J]. Economic geography, 2011, 87(3): 237-265.
105. Lin J Y. New structural economics: A framework for rethinking development and policy[M]. The World Bank, 2012.
106. 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(04): 625-633.
107. Hartmann D, Guevara M R, Jara-Figueroa C, et al. Linking economic complexity, institutions, and income inequality[J]. World development, 2017, 93: 75-93.
108. Salesses P, Schechtner K, Hidalgo C A. The collaborative image of the city: mapping the inequality of urban perception[J]. PloS one, 2013, 8(7): e68400.
109. Hidalgo C A, Castañer E E. Do we need another coffee house? The amenity space and the evolution of neighborhoods[J]. arXiv preprint arXiv:1509.02868, 2015.
110. GAO Jian. Quantifying local industry structure of China[EB/OL]. (2015-11-18). http://gaocn.net/project. html#industry.
111. Louail T, Lenormand M, Arias J M, et al. Crowdsourcing the Robin Hood effect in cities[J]. Applied network science, 2017, 2(1): 1-13.
112. Youn H, Bettencourt L M A, Lobo J, et al. Scaling and universality in urban economic diversification[J]. Journal of The Royal Society Interface, 2016, 13(114): 20150937.
113. 杨振山, 龙瀛. 大数据对人文—经济地理学研究的促进与局限[J]. 地理科学进展, 2015 (4): 410-417.
114. Nelson A C, Duncan J B. Growth management principles and practices[M]. Routledge, 1995.
115. Tan K C, San Lim H, MatJafri M Z, et al. Landsat data to evaluate urban expansion and determine land use/land cover changes in Penang Island, Malaysia[J]. Environmental Earth Sciences, 2010, 60(7): 1509-1521.
116. Long Y, Liu X. Featured graphic. How mixed is Beijing, China? A visual exploration of mixed land use[J]. Environment and Planning A, 2013, 45(12): 2797-2798.
117. 秦萧, 甄峰, 熊丽芳, 等. 2013. 大数据时代城市时空间行为研究方法[J]. 地理科学进展, 32(9): 1352-1451.
118. Long Y, Han H, Tu Y, et al. Evaluating the effectiveness of urban growth boundaries using human mobility and activity records[J]. Cities, 2015, 46: 76-84.
119. 种照辉, 覃成林, 叶信岳. 城市群经济网络与经济增长——基于大数据与网络分析方法的研究[J]. 统计研究, 2018, 35(01): 13-21.
120. 赵渺希, 吴康, 刘行健,等. 城市网络的一种算法及其实证比较[J]. 地理学报, 2014, 69(2):169-183.
121. Liu Y, Sui Z, Kang C, et al. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data[J]. PloS one, 2014, 9(1): e86026.
122. 刘正兵, 刘静玉, 何孝沛, 王发曾. 中原经济区城市空间联系及其网络格局分析——基于城际客运流[J].经济地理, 2014, 34(07): 58-66.
123. 侯赟慧, 刘志彪, 岳中刚. 长三角区域经济一体化进程的社会网络分析[J]. 中国软科学, 2009(12): 90-101.
124. 陈彦光, 刘继生. 基于引力模型的城市空间互相关和功率谱分析[J]. 地理研究, 2002, 21(6): 742-751.
125. 李涛, 周锐, 苏海龙等. 长三角区域经济一体化水平的测度:以关系型大数据为基础[C]// 2015中国城市规划年会. 2015.
126. Hidalgo C A, Klinger B, Barabási A L, et al. The product space conditions the development of nations[J]. Science, 2007, 317(5837): 482-487.
127. 何大安, 杨益均. 大数据时代政府宏观调控的思维模式[J]. 学术月刊, 2018, 50(05): 68-77.
128. 罗慧如. 大数据时代的宏观经济调控及政策优化[D]. 广东财经大学, 2016.
129. 蔡跃洲. 大数据改变经济预测范式[N]. 中国社会科学报, 2015-12-09(004).
130. （英）弗里德里希·冯·哈耶克. 哈耶克文选：哈耶克论文演讲集[M]. 南京：江苏人民出版社. 2000: 452.
131. 黄燕芬, 张超. 大数据情绪指数与经济学研究: 现状、问题与展望[J]. 教学与研究, 2018(05): 40-50.
132. 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究：前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104.
133. Liu Y, Huang X, An A, et al. ARSA: a sentiment-aware model for predicting sales performance using blogs[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007: 607-614.
134. Schneider M J, Gupta S. Forecasting sales of new and existing products using consumer reviews: A random projections approach[J]. International Journal of Forecasting, 2016, 32(2): 243-256.
135. Khadivi P, Ramakrishnan N. Wikipedia in the tourism industry: forecasting demand and modeling usage behavior[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016: 4016-4021.
136. Bagnoli M, Beneish M D, Watts S G. Whisper forecasts of quarterly earnings per share[J]. Journal of Accounting and Economics, 1999, 28(1): 27-50.
137. Tumarkin R, Whitelaw R F. News or noise? Internet postings and stock prices[J]. Financial Analysts Journal, 2001, 57(3): 41-51.
138. 林振兴. 网络讨论、投资者情绪与IPO抑价[J]. 山西财经大学学报, 2011, 33(2): 23-29.
139. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of computational science, 2011, 2(1): 1-8.
140. Oh C, Sheng O. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement[J]. 2011.
141. O'Connor B, Balasubramanyan R, Routledge B, et al. From tweets to polls: Linking text sentiment to public opinion time series[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2010, 4(1).
142. Zhou S, Shi X, Sun Y, et al. Stock market prediction using heat of related keywords on micro blog[J]. Journal of Software Engineering and Applications, 2013, 6(3B): 37.
143. Das S R, Chen M Y. Yahoo! for Amazon: Sentiment extraction from small talk on the web[J]. Management science, 2007, 53(9): 1375-1388.
144. Moat H S, Curme C, Stanley H E, et al. Anticipating stock market movements with Google and Wikipedia[M]//Nonlinear phenomena in complex systems: From nano to macro scale. Springer, Dordrecht, 2014: 47-59.
145. 钱斌华. 税收大数据预测经济走势的某市经验[N]. 某市日报, 2018-03-15(011).
146. 崔趁欣, 李岩. 大数据时代国库收入与经济增长耦合性分析及应用前景——以河北省为例[J]. 河北金融, 2018(03): 7-13.
147. 张秋雁, 宋强. 基于用电大数据的经济分析平台设计研究[J]. 电力大数据, 2017, 20(08): 6-9.
148. 曲延玲. 基于大数据构建经济金融深度监测分析体系[J]. 金融电子化, 2017(05): 40-41.
149. Bok B, Caratelli D, Giannone D, et al. Macroeconomic nowcasting and forecasting with big data[J]. Annual Review of Economics, 2018, 10: 615-643.
150. http://di.baidu.com/solution/finance?castk=LTE%3D
151. Levenberg A, Simpson E, Roberts S, et al. Economic Prediction using heterogeneous data streams from the World Wide Web[C]//Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE), Proceedings of ECML/PKDD Workshop. 2013.
152. Sobolevsky S, Massaro E, Bojic I, et al. Predicting regional economic indices using big data of individual bank card transactions[C]//2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017: 1313-1318.
153. 俞立平. 大数据经济学的概念、框架与学科定位研究[J]. 统计与信息论坛, 2015, 30(06): 3-7.
154. 刘涛雄, 徐晓飞. 互联网搜索行为能帮助我们预测宏观经济吗?[J]. 经济研究, 2015, 50(12): 68-83.
155. Götz T B, Knetsch T A. Google data in bridge equation models for German GDP[J]. International Journal of Forecasting, 2019, 35(1): 45-66.
156. Kholodilin K A, Podstawski M, Siliverstovs B, et al. Google searches as a means of improving the nowcasts of key macroeconomic variables[J]. 2009.
157. D'Amuri F, Marcucci J. 'Google it!'Forecasting the US unemployment rate with a Google job search index[R]. Nota di Lavoro, 2010.
158. Artola C, Pinto F, de Pedraza García P. Can internet searches forecast tourism inflows?[J]. International Journal of Manpower, 2015, 36(1): 103-116.
159. 许 伟. 基于网络大数据的社会经济监测预警研究[M]. 北京：科学出版社, 2016.
160. 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究：前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104.
161. （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018.
162. 王国成. 行为大数据与计算实验——探索经济研究新方法[J]. 天津社会科学, 2016(03): 86-92.
163. Hidalgo C A, Hausmann R. The building blocks of economic complexity[J]. Proceedings of the national academy of sciences, 2009, 106(26): 10570-10575.
164. Bustos S, Gomez C, Hausmann R, et al. The dynamics of nestedness predicts the evolution of industrial ecosystems[J]. PloS one, 2012, 7(11): e49393.
165. GAO Jian. Modeling local economy complexity[EB/OL]. http://gaocn.net/project.html #complexity, 2015-11-18.
166. Cristelli M, Tacchella A, Pietronero L. The heterogeneous dynamics of economic complexity[J]. PloS one, 2015, 10(2): e0117174.
167. Tacchella A, Cristelli M, Caldarelli G, et al. A new metrics for countries' fitness and products' complexity[J]. Scientific reports, 2012, 2(1): 1-7.
168. Hidalgo C A, Hausmann R. The building blocks of economic complexity[J]. Proceedings of the national academy of sciences, 2009, 106(26): 10570-10575.
169. 潘迪特, 李昌琪. 时间序列及系统分析与应用[M]. 北京：机械工业出版社, 1988.
170. 李一智. 经济预测技术[M]. 北京：清华大学出版社, 1991.
171. 邓聚龙. 灰色系统 社会·经济[M]. 北京：国防工业出版社, 1985.
172. 李志强, 白文斌, 张亚丽, 等. 基于ARIMA模型的内蒙古羊产业分析与预测[J]. 山西农业科学, 2011, 39(07): 729-732+743.
173. 王晓梅. 灰色理论GM(1,1)模型在我国畜产品产量预测中的应用[J]. 安徽农业科学, 2007(01): 7-8.
174. 林绍森, 唐永金. 三种模型对我国粮食产量预测效果的比较[J]. 统计与决策, 2007(07): 39-40.
175. 刘峰,王儒敬,李传席. ARIMA模型在农产品价格预测中的应用[J]. 计算机工程与应用, 2009, 45(25): 238-239+248.
176. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. New York: Springer Science & Business Media, 2009.
177. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net[J]. Journal of the Royal Statistical Society Series B, Statistical Methodology. 2005, 67 (2):301-320.
178. Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees [M]. CRC Press,1984.
179. Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.
180. 谢坤, 容钰添, 胡奉平, 陈桓, 姚小龙. 基于数据集成的随机森林算法[J]. 计算机工程, 2020, 46(12): 290-298.
181. 林霞, 刘宗尚, 高宇等. 基于机器学习的产油量主控因素分析[J]. 信息系统工程, 2019(12): 94-97+99.
182. Ayaru L, Ypsilantis P P, Nanapragasam A, et al. Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting[J]. PloS one, 2015, 10(7): e0132485.
183. 张棪, 曹健. 面向大数据分析的决策树算法[J]. 计算机科学, 2016, 43(S1):374-379+383.
184. 董莉, 彭凯越, 唐晓彬. 大数据背景下的CPI实时预测研究[J]. 调研世界, 2017(08): 51-54.
185. 康传利, 顾峻峰, 刘兆威. 梯度提升回归树的旅游流量预测模型[J]. 数学的实践与认识, 2019, 49(15): 251-261.
186. 巩晓文, 凤思苑, 崔壮等. 基于SVGD分类预测的梯度提升机与随机森林的性能比较[J]. 中国卫生统计, 2019, 36(05): 674-677.
187. 韩忠明, 原碧鸿, 陈炎等. 一个有效的基于GBRT的早期电影票房预测模型[J]. 计算机应用研究, 2018, 35(02): 410-416.
188. 黄乃静,于明哲.机器学习对经济学研究的影响研究进展[J].经济学动态,2018(07):115-129.
189. 葛浩. 基于CART决策树模型的上市企业财务风险预警研究[D].上海财经大学,2020.
190. 李颖. 基于GBRT集成模型的新冠疫情对我国宏观经济影响研究[D].重庆工商大学,2021.DOI:10.27713/d.cnki.gcqgs.2021.000402.
191. 贺平,兰伟,丁月.我国股票市场可以预测吗?——基于组合LASSO-logistic方法的视角[J/OL].统计研究:1-15, 2021-12-24.
192. 欧阳梦倩,周先波,朱君梅.大数据时代下使用互联网搜索量预测CPI——基于LASSO和核偏最小二乘的联合使用[J].金融学季刊,2020,14(02):112-136.
193. 胡清畅. 互联网搜索行为对宏观经济景气指数预测的影响研究[D].暨南大学,2019.
194. 陈梦根,任桃萍.新常态经济的CPI预测模型——构建与实证比较[J].调研世界,2020(02):3-8.
195. 刘晓鑫,景祥云,叶骏骅,金博,毕文祥.关于大数据在中央银行监管应用的文献综述[J].吉林金融研究,2020(10):6-9+18.
196. 高子宜. 基于随机森林的股价走势预测研究[D].中国政法大学,2021.
197. 邹玉莹. 基于机器学习的票据转贴现利率预测研究[D].江西财经大学,2020.
198. 屠星月,于辉辉,郭承坤,阮怀军,陈英义.基于EMD-SVM的农产品市场价格短期预测模型[J].湖北农业科学,2015,54(19):4903-4908.
199. 肖雪梦. 基于PLS-SVM的重庆市经济发展状况研究[D].重庆大学,2014.
200. 何雁明,黄邱婧,郑其敏.大数据时代的数量经济模型研究——以BP神经网络的中国CPI预测为例[J].西部金融,2021(02):27-32.
201. 李杰. 基于神经网络的利率期限结构组合预测研究[D].北京化工大学,2018.
202. 王永杰. 人工神经网络算法在GDP和CPI中的预测应用[D].中北大学,2017.
203. 沈虹,李旭,潘琪.基于深度学习长短期记忆神经网络的有色金属期货市场预测研究[J].南京理工大学学报,2021,45(03):366-374.
204. 李佳,黄之豪,王佳慧.基于独立循环神经网络方法的GDP预测[J].统计与决策,2020,36(14):24-28.
205. 维克托·迈尔·舍恩伯格, 周涛. 大数据时代 生活,工作与思维的大变革[J]. 人力资源管理, 2013, 03: 176-176.
206. Mokyr J. Intellectuals and the rise of the modern economy[J]. Science, 2015, 349(6244): 141-142.
207. 陈争平. 大数据时代与经济史计量研究[J]. 社会科学文摘, 2017(02): 30-32.
208. 黄燕芬, 张超. 大数据情绪指数与经济学研究：现状、问题与展望[J]. 教学与研究, 2018(05): 40-50.
209. 汪毅霖. 大数据预测与大数据时代的经济学预测[J]. 学术界, 2016(11): 77-90+325.
210. Attewell P, Monaghan D, Kwong D. Data mining for the social sciences: An introduction[M]. University of California Press, 2015.
211. 张涛, 刘宽斌. “大数据”在宏观经济预测分析中的应用[J]. 社会科学文摘, 2018(08): 47-49.
212. （英）杰弗里·M·霍奇逊. 经济学是如何忘记历史的:社会科学中的历史特性问题[M]. 北京：中国人民大学出版社. 2008: 11.
213. 汪毅霖. 大数据预测与大数据时代的经济学预测[J]. 学术界, 2016(11): 77-90+325.
214. Mao H, Counts S, Bollen J. Quantifying the effects of online bullishness on international financial markets[R]. ECB Statistics Paper, 2015.
215. 张涛, 刘宽斌. “大数据”在宏观经济预测分析中的应用[J]. 社会科学文摘, 2018(08): 47-49.
216. 黄燕芬, 张超. 大数据情绪指数与经济学研究：现状、问题与展望[J]. 教学与研究, 2018(05): 40-50.
217. 王建冬, 童楠楠, 易成岐. 大数据时代公共政策评估的变革：理论、方法与实践[M]. 北京：社科文献出版社. 2019: 120-138.
218. 冯·诺依曼. 数学在科学和社会中的作用[M]. 大连：大连理工大学出版社. 2009: 54-58.
219. （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018.
220. Bustos S, Gomez C, Hausmann R, et al. The dynamics of nestedness predicts the evolution of industrial ecosystems[J]. PloS one, 2012, 7(11): e49393.
221. Tacchella A, Cristelli M, Caldarelli G, et al. A new metrics for countries' fitness and products' complexity[J]. Scientific reports, 2012, 2(1): 1-7.
222. 汪丁丁. 理解“涌现秩序”(推荐序一)//（美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018
223. 布莱恩.阿瑟访谈录（中文版序）//（美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018.
224. （英）A.N.怀特海. 科学与近代世界[M]. 商务印书馆. 1959.
225. （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018.
226. 何新. 反主流经济学[M]. 沈阳：万卷出版公司. 2013: 263.
227. （英）弗里德里希·冯·哈耶克. 哈耶克文选：哈耶克论文演讲集[M]. 南京：江苏人民出版社. 2000: 450-452.
228. （英）维克托・迈尔-舍恩伯格, 肯尼思・库克耶等. 大数据时代：生活、工作与思维的大变革[M]. 浙江人民出版社, 2013: 2.
229. 徐晋. 大数据经济学[M]. 上海：上海交通大学出版社. 2014: 96.
230. 宋圭武. 大数据时代背景下的经济学[J]. 发展, 2014(07): 5.
231. Anderson C. The end of theory: The data deluge makes the scientific method obsolete[J]. Wired magazine,2008, 16(7): 16-07.
232. 黄璜. 社会科学研究中“基于主体建模”方法评述[J]. 国外社会科学, 2010(05): 40-47.
233. Holland J H, Miller J H. Artificial adaptive agents in economic theory[J]. The American economic review, 1991, 81(2): 365-370.
234. （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018.
235. Janssen M A, Ostrom E. Empirically based, agent-based models[J]. Ecology and society, 2006, 11(2).
236. Janssen M A, Ostrom E. Adoption of a new regulation for the governance of common-pool resources by a heterogeneous population[J]. Inequality, Cooperation, and Environmental Sustainability, 2006: 60-96.
237. 黄璜. 社会科学研究中“基于主体建模”方法评述[J]. 国外社会科学, 2010(05): 40-47.
238. 汪丁丁. 行为经济学讲义：演化论的视角[M]. 上海：上海人民出版社. 2011: 230.
239. Tesfatsion L. Agent-based computational economics: Growing economies from the bottom up[J]. Artificial life, 2002, 8(1): 55-82.
240. （美）布莱恩·阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018.
241. 王国成. 行为大数据与计算实验——探索经济研究新方法[J]. 天津社会科学, 2016(03): 86-92.
242. 蔡永鸿, 唐斯琪, 于娟. 大数据时代宏观经济与微观企业的行为联接[J]. 中国市场, 2015(08): 88-89.
243. 赵怡君. 大数据时代宏观经济与微观企业的行为联接探讨[J]. 全国流通经济, 2018(08): 75-76.
244. Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. STL: A seasonal- trend decomposition procedure based on Loess (with discussion). Journal of Official Statistics, 6: 3–73, 1990.
245. U.S. Census Bureau. Reference manual for the X-13ARIMA-SEATS Program, version 1.0 URL http://www.census.gov/ts/x13as/docX13AS.pdf, 2013.

## 附件：项目成果

**1、搭建实验平台1套：**搭建基于机器学习（Weka）和深度学习（PyTorch）框架的宏观经济预测实验平台，为本项目数据实验及后续开展研究奠定平台基础。

**2、沉淀算法模型库1套：**除ARIMA、Prophet、LASSO、BSTS等传统时序预测模型外，本研究还积累沉淀了线性回归、弹性网络回归（Elastic Net）、随机森林回归（Random Forest）、梯度提升回归树（GBRT）、支持向量机回归（Support Vector Regression）、高斯过程回归（Gaussian Processes）等机器学习数值型预测算法模型以及最新N-BEATS深度学习时序预测算法模型，能够为后续开展研究奠定算法模型基础。

**3、出版中文专著1部：**于施洋, 王建冬, 易成岐. 宏观经济大数据分析[M]. 北京：社会科学文献出版社, 2021.

**4、发布国家发展改革委官网新闻1篇：**“《宏观经济大数据分析》正式出版”（2021年10月26日发布）。

**5、发布中心公众号新闻1篇：**“《宏观经济大数据分析》正式出版”（2021年10月28日发布，阅读量2078）。

**6、投稿人工智能领域顶级期刊SCI论文1篇：**Chengqi Yi, Xuanzheng Wang, Changwang Li, Xinan Xu, Jiandong Wang, Shiyang Yu. EcoForecast: An interpretable data-driven approach for short-term macroeconomic forecasting using N-BEATS neural network[J]. Knowledge-Based Systems. （影响因子8.04）。

**7、推动工程项目1项：**依托本研究成果设计的“宏中微观一体化大数据分析系统”被纳入“粤港澳大湾区大数据中心（一期）项目”，该项目总投资2.4亿元，已完成公开招投标。

**8、培养实习生4名：**共有来自清华大学、首都经济贸易大学、英国牛津大学、英国华威大学等4名实习生（2名博士生、2名硕士生）参与到本研究中，全程参与了研究思路讨论、数据实验设计、结题报告撰写等有关工作，该项目对培养实习生强化利用机器学习等大数据算法支撑宏观经济监测预测研究能力和兴趣点亦有贡献。

|  |  |
| --- | --- |
| **项目组成员** | |
| 王建冬 | 大数据发展部，处长 |
| 郭明军 | 大数据发展部，副处长 |
| 童楠楠 | 大数据发展部，副处长 |
| 黄倩倩 | 大数据发展部，干部 |
| 李慧颖 | 大数据发展部，干部 |
| 陈 东 | 大数据发展部，干部 |
| 窦 悦 | 大数据发展部，干部 |
| 赵 正 | 大数据发展部，干部 |
| 马 骁 | 大数据发展部，干部 |
| 聂 磊 | 北京外国语大学，讲师 |
| 李昌旺 | 大数据发展部，干部 |
| 王璇铮 | 粤港澳大湾区大数据研究院 |
| 徐溪楠 | 英国牛津大学统计系硕士生，干部 |
| 马慧洁 | 首都经济贸易大学财政税务学院，博士生 |
| 刘苏瑶 | 英国华威大学数据科学专业，硕士生 |

1. [] 国务院关于印发促进大数据发展行动纲要的通知[EB/OL].http://www.gov.cn/zhengce/content/2015-09/05/content\_10137.htm, 2015-09. [↑](#footnote-ref-1)
2. [] 中共中央关于深化党和国家机构改革的决定[EB/OL].http://www.gov.cn/xinwen/2018-03/04/content\_5270704.htm, 2018-03. [↑](#footnote-ref-2)
3. [] 中共中央关于坚持和完善中国特色社会主义制度 推进国家治理体系和治理能力现代化若干重大问题的决定[EB/OL]. http://www.gov.cn/zhengce/2019-11/05/content\_5449023.htm, 2019-11. [↑](#footnote-ref-3)
4. [] Clarke A, Margetts H. Governments and citizens getting to know each other? Open, closed, and big data in public management reform[J]. Policy & Internet, 2014, 6(4): 393-417. [↑](#footnote-ref-4)
5. [5] 习近平. 坚持开放包容 推动联动增长——在二十国集团领导人汉堡峰会上关于世界经济形势的讲话.[EB/OL]. http://jhsjk.people.cn/article/29391615, 2017-07. [↑](#footnote-ref-5)
6. [] Varian H R. Big data: New tricks for econometrics[J]. Journal of Economic Perspectives, 2014, 28(2): 3-28. [↑](#footnote-ref-6)
7. [] 冯.诺依曼. 数学在科学和社会中的作用[M]. 大连：大连理工大学出版社. 2009: 58. [↑](#footnote-ref-7)
8. [] 陈禹, 方美琪. 复杂性研究视角中的经济系统[M]. 北京：商务印书馆. 2015: 52-53. [↑](#footnote-ref-8)
9. [] Big Data is a Big Deal[EB/OL][2012-04-20]. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012-03. [↑](#footnote-ref-9)
10. [] 张勇进,王璟璇. 主要发达国家大数据政策比较研究[J]. 中国行政管理. 2014(12): 114-117. [↑](#footnote-ref-10)
11. [] Chen J C, Dunn A, Hood K, et al. Off to the races: A comparison of machine learning and alternative data for predicting economic indicators[M]. Big Data for 21st Century Economic Statistics. University of Chicago Press, 2019. [↑](#footnote-ref-11)
12. [] Konny C G, Williams B K, Friedman D M. Big Data in the US Consumer Price Index: Experiences and Plans[M]. Big Data for 21st Century Economic Statistics. University of Chicago Press, 2019. [↑](#footnote-ref-12)
13. [] Cuffe J, Bhattacharjee S, Etudo U, et al. Using Public Data to Generate Industrial Classification Codes[M]. Big Data for 21st Century Economic Statistics. University of Chicago Press, 2019. [↑](#footnote-ref-13)
14. [] 王建冬, 童楠楠, 易成岐. 大数据时代公共政策评估的变革：理论、方法与实践[M]. 北京：社科文献出版社. 2019: 54-55. [↑](#footnote-ref-14)
15. [] Vintar M. Information Polity: The International Journal of Goverment and Democracy in the Information Age[M]. IOS Press, 2006. [↑](#footnote-ref-15)
16. [] Hundt R. From Carbon to Clean—How to Attract Investment in Smart Grid Infrastructures[M]. Broadband Networks, Smart Grids and Climate Change. Springer, New York, NY, 2013: 39-45. [↑](#footnote-ref-16)
17. [] Clarke A, Margetts H. Governments and citizens getting to know each other? Open, closed, and big data in public management reform[J]. Policy & Internet, 2014, 6(4): 393-417. [↑](#footnote-ref-17)
18. [] Milakovich M E. Anticipatory government: Integrating big data for smaller government[J]. Internet, politics, policy 2012: Big data, big challenges, 2012. [↑](#footnote-ref-18)
19. [] 陈敏尔：让智能化为经济赋能为生活添彩[EB/OL] http://difang.gmw.cn/cq/2019-07/03/content\_32968465.htm, 2019-07. [↑](#footnote-ref-19)
20. [] 牛文胜, 董晓博, 吴又奎. 政务信息资源目录及宏观经济大数据系统的建设及应用[J]. 价值工程, 2018, 37(20): 251-255. [↑](#footnote-ref-20)
21. [] 《2018年海南省经济社会发展大数据分析报告》出炉. http://www.hainan.gov.cn/hainan/tingju/201901/dbebcc4ad3624e4b812057a1b6712309.shtml,2019-01. [↑](#footnote-ref-21)
22. [] 钱斌华. 税收大数据预测经济走势的某市经验[J]. 某市日报, 2018: 03-15. [↑](#footnote-ref-22)
23. [] Einav L, Levin J. Economics in the age of big data[J]. Science, 2014, 346(6210). [↑](#footnote-ref-23)
24. [] Balasubramanian N, Sivadasan J. What happens when firms patent? New evidence from US economic census data[J]. The Review of Economics and Statistics, 2011, 93(1): 126-146. [↑](#footnote-ref-24)
25. [] 马建堂. 大数据在政府统计中的探索与应用[M]. 北京：中国统计出版社, 2013. [↑](#footnote-ref-25)
26. [] 徐先华. 大数据时代的经济普查数据分析与探究[J]. 时代金融, 2018(11): 49-50. [↑](#footnote-ref-26)
27. [] 杜成琳. 大数据时代我国经济统计现状及趋势研究[J]. 知识经济, 2018(14): 14+16. [↑](#footnote-ref-27)
28. [] 张维扬. 大数据时代我国经济统计现状及趋势研究[J]. 现代经济信息, 2018(01): 104. [↑](#footnote-ref-28)
29. [] 李勇, 罗良清, 张敏, 李禹锋. 大数据时代我国经济统计现状及趋势研究[J]. 中国统计, 2016(12): 21-23. [↑](#footnote-ref-29)
30. [] 吴毅. 大数据环境下城市经济统计调查分析及改进措施[J]. 经贸实践, 2017(24): 71. [↑](#footnote-ref-30)
31. [] 王怡博, 方颀卓. 浅析云数据分析在经济统计中的应用[J]. 中国市场, 2018(03): 189-190. [↑](#footnote-ref-31)
32. [] 祝旭. 经济统计中大数据分析的应用研究[J]. 无线互联科技, 2018, 15(17): 110-111. [↑](#footnote-ref-32)
33. [] 何海. 大数据在区域经济统计中的应用研究[J]. 现代营销(下旬刊), 2018(01): 102. [↑](#footnote-ref-33)
34. [] 解路英. 大数据使“定制”新的经济指标成为可能[N]. 经济参考报, 2014-07-11(005). [↑](#footnote-ref-34)
35. [] 李世伟. 大数据环境下国民经济核算体系的改进[J]. 市场经济与价格, 2015(07): 10-14. [↑](#footnote-ref-35)
36. [] 李红艳, 汪涛. 大数据时代背景下的新型国民经济核算体系研究[J]. 经济视角(上), 2013(08): 24-26. [↑](#footnote-ref-36)
37. [] 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报(人文社会科学), 2017, 17(03): 54-59. [↑](#footnote-ref-37)
38. [] 刘涛雄, 徐晓飞. 大数据与宏观经济分析研究综述[J]. 国外理论动态, 2015(01): 57-64. [↑](#footnote-ref-38)
39. [] Economist. Keqiang ker-ching: How China’s next prime minister keeps tabs on its economy[EB/OL]. http://www.economist.com/node/17681868, 2010-12. [↑](#footnote-ref-39)
40. [] 骆兆华. 经济孰冷孰热 “挖掘机指数”有的聊[N]. 中国城乡金融报,2016-01-15(B03). [↑](#footnote-ref-40)
41. [] 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报(人文社会科学), 2017, 17(03): 54-59. [↑](#footnote-ref-41)
42. [] 陈沁, 沈明高, 沈艳. 财智BBD中国新经济指数技术报告[EB/OL].http://www.nsd.edu.cn/teachers/professorNews/2016/0304/ 25596.html, 2016-03. [↑](#footnote-ref-42)
43. [] 国家信息中心发布《中国大数据发展报告(2017)》[EB/OL]. <http://www.sic.gov.cn/News/79/7727.htm,> 2017-02. [↑](#footnote-ref-43)
44. [] 国家信息中心发布《中国共享经济发展年度报告（2019）》[EB/OL]. <http://www.sic.gov.cn/News/79/9907.htm,> 2019-03 [↑](#footnote-ref-44)
45. [] 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究:前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104. [↑](#footnote-ref-45)
46. [] 张秋雁, 宋强. 基于用电大数据的经济分析平台设计研究[J]. 电力大数据, 2017, 20(08): 6-9. [↑](#footnote-ref-46)
47. [] 王凯军, 龙厚印, 吴良良, 石清. 基于电力大数据的产业结构调整及经济指标研究[J]. 经济研究导刊, 2017(25): 38-39. [↑](#footnote-ref-47)
48. [] Doll C N H, Muller J P, Morley J G. Mapping regional economic activity from night-time light satellite imagery[J]. Ecological Economics, 2006, 57(1): 75-92. [↑](#footnote-ref-48)
49. [] Elvidge C D, Baugh K E, Kihn E A, et al. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption[J]. International Journal of Remote Sensing, 1997, 18(6): 1373-1379. [↑](#footnote-ref-49)
50. [] ŠĆEPANOVIĆ S, MISHKOVSKI I, HUI P, et al. Mobile phone call data as a regional socio-economic proxy indicator[J]. PLoS ONE, 2015, 10(4): e0124160. [↑](#footnote-ref-50)
51. [] EAGLE N, MACY M, CLAXTON R. Network diversity and economic development[J]. Science, 2010, 328(5981): 1029-1031. [↑](#footnote-ref-51)
52. [] HOLZBAUER B O, SZYMANSKI B K, NGUYEN T, et al. Social ties as predictors of economic development[M]// WIERZBICKI A, BRANDES U, SCHWEITZER F, et al. Advances in Network Science. Switzerland: Springer International Publishing, 2016: 178-185. [↑](#footnote-ref-52)
53. [] Liu J H, Wang J, Shao J, et al. Online social activity reflects economic status[J]. Physica A: Statistical Mechanics and its Applications, 2016, 457: 581-589. [↑](#footnote-ref-53)
54. [] 曲延玲. 基于大数据构建经济金融深度监测分析体系[J]. 金融电子化, 2017(05): 40-41. [↑](#footnote-ref-54)
55. [] Kholodilin K A, Podstawski M, Siliverstovs B, et al. Google searches as a means of improving the nowcasts of key macroeconomic variables[J]. 2009. [↑](#footnote-ref-55)
56. [] Della Penna N, Huang H. Constructing consumer sentiment index for US using Google searches[R]. 2010. [↑](#footnote-ref-56)
57. [] Suhoy T. Monthly assessments of private consumption[R]. Bank of Israel, 2010. [↑](#footnote-ref-57)
58. [] Vosen S . A monthly consumption indicator for Germany based on Internet search query data[J]. Applied Economics Letters, 2012, 19(7): 683-687. [↑](#footnote-ref-58)
59. [] 李映桥. 网络消费者信心与宏观经济波动关系研究[D].吉林大学,2017. [↑](#footnote-ref-59)
60. [] 孙毅, 吕本富, 陈航, 薛添. 基于网络搜索行为的消费者信心指数构建及应用研究[J]. 管理评论, 2014, 26(10): 117-125. [↑](#footnote-ref-60)
61. [] 刘伟江, 李映桥. 网络消费者信心指数和经济增长的动态相关性研究[J]. 财贸研究, 2017, 28(05): 1-10. [↑](#footnote-ref-61)
62. [] 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报（人文社会科学）, 2017, 17(03): 54-59. [↑](#footnote-ref-62)
63. [] 孙毅, 吕本富, 陈航, 等. 大数据视角的通胀预期测度与应用研究[J]. 管理世界, 2014, 4: 171-172. [↑](#footnote-ref-63)
64. [] Cavallo A , Rigobon R . The Billion Prices Project: Using Online Prices for Measurement and Research[J]. The Journal of Economic Perspectives, 2016: 151-178. [↑](#footnote-ref-64)
65. []袁铭. 基于网购搜索量的CPI及时预测模型[J]. 统计与信息论坛, 2015, 30(4): 20-24. [↑](#footnote-ref-65)
66. [] Müller R, Herren H M, Röthlisberger S, et al. Recent developments in the Swiss CPI: scanner data, telecommunications and health price collection[C]//9th Ottawa Group Meeting on Prices 14th to 16th May, London. 2006. [↑](#footnote-ref-66)
67. [] Guzman G. Internet search behavior as an economic forecasting tool: The case of inflation expectations[J]. Journal of economic and social measurement, 2011, 36(3): 119-167. [↑](#footnote-ref-67)
68. [] 徐映梅, 高一铭. 基于互联网大数据的CPI舆情指数构建与应用——以百度指数为例[J]. 数量经济技术经济研究, 2017, 34(1): 94-112. [↑](#footnote-ref-68)
69. [] 张崇, 吕本富, 彭赓等. 网络搜索数据与CPI的相关性研究[J]. 管理科学学报, 2012: 50-59+ 70. [↑](#footnote-ref-69)
70. [] Ettredge M, Gerdes J, Karuga G. Using web-based search data to predict macroeconomic statistics[J]. Communications of the ACM, 2005, 48(11): 87-92. [↑](#footnote-ref-70)
71. [] Askitas N, Zimmermann K F. Google Econometrics and Unemployment Forecasting[J]. Applied Economics Quarterly, 2009, 55(2): 107. [↑](#footnote-ref-71)
72. []D'Amuri F. Predicting unemployment in short samples with internet job search query data[R]. University Library of Munich, Germany, 2009. [↑](#footnote-ref-72)
73. [] Suhoy T. Monthly assessments of private consumption[R]. Bank of Israel, 2010. [↑](#footnote-ref-73)
74. [] McLaren N, Shanbhogue R. Using internet search data as economic indicators[J]. Bank of England Quarterly Bulletin, 2011 (2011): Q2. [↑](#footnote-ref-74)
75. [] Choi H, Varian H. Predicting the present with Google Trends[J]. Economic record, 2012, 88: 2-9. [↑](#footnote-ref-75)
76. [] Barreira N, Godinho P, Melo P. Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends[J]. NETNOMICS: Economic Research and Electronic Networking, 2013, 14(3): 129-165. [↑](#footnote-ref-76)
77. [] Pavlicek J, Kristoufek L. Can Google searches help nowcast and forecast unemployment rates in the Visegrad Group countries?[J]. arXiv preprint arXiv:1408.6639, 2014. [↑](#footnote-ref-77)
78. [] Su Z. Chinese online unemployment-related searches and macroeconomic indicators[J]. Frontiers of Economics in China, 2014, 9(4): 573-605. [↑](#footnote-ref-78)
79. [] Pavlicek J, Kristoufek L. Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries[J]. PloS one, 2015, 10(5): e0127084. [↑](#footnote-ref-79)
80. [] 王勇, 董恒新. 大数据背景下中国季度失业率的预测研究------基于网络搜索数据的分析[J]. 系统科学与数学, 2017, 37(2): 460-472. [↑](#footnote-ref-80)
81. [] 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(04): 625-633. [↑](#footnote-ref-81)
82. [] 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究：前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104. [↑](#footnote-ref-82)
83. [] Toole J L, Lin Y R, Muehlegger E, et al. Tracking employment shocks using mobile phone data[J]. Journal of The Royal Society Interface, 2015, 12(107): 20150185. [↑](#footnote-ref-83)
84. [] Llorente A, Garcia-Herranz M, Cebrian M, et al. Social media fingerprints of unemployment[J]. PloS one, 2015, 10(5): e0128692. [↑](#footnote-ref-84)
85. [] PAPPALARDO L, VANHOOF M, GABRIELLI L, et al. Estimating economic development with mobile phone data[EB/OL]. (2016-05-30). http://www.cisstat.com/ BigData/CIS-BigData\_08\_Eng%20%20IT%20Luca%20Pa ppalardo%20Et%20Al%20Estimating%20Economic%20D evelopment.pdf. [↑](#footnote-ref-85)
86. [] 明文彪. 大数据背景下宏观经济分析及对浙江的建议[J]. 浙江树人大学学报（人文社会科学）, 2017, 17(03): 54-59. [↑](#footnote-ref-86)
87. [] 王智, 余桂南, 赵路. 基于大数据的我国新经济服务业人才需求景气指数研究[J]. 2016年全国统计建模大赛论文集, 2016. [↑](#footnote-ref-87)
88. [] Webb G K. Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures[J]. Issues in Information Systems, 2009, 10(2): 82. [↑](#footnote-ref-88)
89. [] Wu L, Brynjolfsson E. The future of prediction: How Google searches foreshadow housing prices and sales[M]//Economic analysis of the digital economy. University of Chicago Press, 2015: 89-118. [↑](#footnote-ref-89)
90. [] Kulkarni R, Haynes K E, Stough R R, et al. Forecasting housing prices with Google econometrics[J]. GMU School of public policy research paper, 2009 (2009-10). [↑](#footnote-ref-90)
91. [] Widgrén J. Predicting Housing Prices with Google Searches in Finland[R]. The Research Institute of the Finnish Economy, 2016. [↑](#footnote-ref-91)
92. [] 董倩, 孙娜娜, 李伟. 基于网络搜索数据的房地产价格预测[J]. 统计研究, 2014, 31(10): 81-88. [↑](#footnote-ref-92)
93. [] 洪涛, 厉伟. 基于网络搜索数据的住房价格预期与实际价格波动分析[J]. 统计与信息论坛, 2015: 49-53. [↑](#footnote-ref-93)
94. [] Choi H, Varian H. Using search engine data for nowcasting—an illustration[J]. Actes des Rencontrees Economiques, 2011: 535-538. [↑](#footnote-ref-94)
95. [] Wu L, Brynjolfsson E. The future of prediction: how Google searches foreshadow housing prices and quantities[J]. ICIS 2009 Proceedings, 2009: 147. [↑](#footnote-ref-95)
96. [] Webb G K. Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures[J]. Issues in Information Systems, 2009, 10(2): 82. [↑](#footnote-ref-96)
97. [] McLaren N, Shanbhogue R. Using internet search data as economic indicators[J]. Bank of England Quarterly Bulletin, 2011: Q2. [↑](#footnote-ref-97)
98. [] 霍琳, 尚维, 徐山鹰. 房地产开源舆情指数构建与政策影响研究[J]. 信息系统学报, 2013, 2: 57-66. [↑](#footnote-ref-98)
99. [] 王钧超. 大数据时代产业经济信息分析及在宏观决策中的应用[D]. 中国地质大学（北京）, 2016. [↑](#footnote-ref-99)
100. [] Li Y, Arora S, Youtie J, et al. Using web mining to explore Triple Helix influences on growth in small and mid-size firms[J]. Technovation, 2018, 76: 3-14. [↑](#footnote-ref-100)
101. [] Daepp M I G, Hamilton M J, West G B, et al. The mortality of companies[J]. Journal of The Royal Society Interface, 2015, 12(106): 20150120. [↑](#footnote-ref-101)
102. [] 毛红卫, 阮伯虎, 吴铭宇等. 诸暨市大数据经济运行监测与研究[J]. 民营科技, 2018(02): 182-185. [↑](#footnote-ref-102)
103. [] 王展. 大数据经济运行系统的研究[D]. 郑州大学, 2017. [↑](#footnote-ref-103)
104. [] Neffke F, Henning M, Boschma R. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions[J]. Economic geography, 2011, 87(3): 237-265. [↑](#footnote-ref-104)
105. [] Lin J Y. New structural economics: A framework for rethinking development and policy[M]. The World Bank, 2012. [↑](#footnote-ref-105)
106. [] 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(04): 625-633. [↑](#footnote-ref-106)
107. [] Hartmann D, Guevara M R, Jara-Figueroa C, et al. Linking economic complexity, institutions, and income inequality[J]. World development, 2017, 93: 75-93. [↑](#footnote-ref-107)
108. [] Salesses P, Schechtner K, Hidalgo C A. The collaborative image of the city: mapping the inequality of urban perception[J]. PloS one, 2013, 8(7): e68400. [↑](#footnote-ref-108)
109. [] Hidalgo C A, Castañer E E. Do we need another coffee house? The amenity space and the evolution of neighborhoods[J]. arXiv preprint arXiv:1509.02868, 2015. [↑](#footnote-ref-109)
110. [] GAO Jian. Quantifying local industry structure of China[EB/OL]. (2015-11-18). http://gaocn.net/project. html#industry. [↑](#footnote-ref-110)
111. [] Louail T, Lenormand M, Arias J M, et al. Crowdsourcing the Robin Hood effect in cities[J]. Applied network science, 2017, 2(1): 1-13. [↑](#footnote-ref-111)
112. [] Youn H, Bettencourt L M A, Lobo J, et al. Scaling and universality in urban economic diversification[J]. Journal of The Royal Society Interface, 2016, 13(114): 20150937. [↑](#footnote-ref-112)
113. [] 杨振山, 龙瀛. 大数据对人文—经济地理学研究的促进与局限[J]. 地理科学进展, 2015 (4): 410-417. [↑](#footnote-ref-113)
114. [] Nelson A C, Duncan J B. Growth management principles and practices[M]. Routledge, 1995. [↑](#footnote-ref-114)
115. [] Tan K C, San Lim H, MatJafri M Z, et al. Landsat data to evaluate urban expansion and determine land use/land cover changes in Penang Island, Malaysia[J]. Environmental Earth Sciences, 2010, 60(7): 1509-1521. [↑](#footnote-ref-115)
116. [] Long Y, Liu X. Featured graphic. How mixed is Beijing, China? A visual exploration of mixed land use[J]. Environment and Planning A, 2013, 45(12): 2797-2798. [↑](#footnote-ref-116)
117. [] 秦萧, 甄峰, 熊丽芳, 等. 2013. 大数据时代城市时空间行为研究方法[J]. 地理科学进展, 32(9): 1352-1451. [↑](#footnote-ref-117)
118. [] Long Y, Han H, Tu Y, et al. Evaluating the effectiveness of urban growth boundaries using human mobility and activity records[J]. Cities, 2015, 46: 76-84. [↑](#footnote-ref-118)
119. [] 种照辉, 覃成林, 叶信岳. 城市群经济网络与经济增长——基于大数据与网络分析方法的研究[J]. 统计研究, 2018, 35(01): 13-21. [↑](#footnote-ref-119)
120. [] 赵渺希, 吴康, 刘行健,等. 城市网络的一种算法及其实证比较[J]. 地理学报, 2014, 69(2):169-183. [↑](#footnote-ref-120)
121. [] Liu Y, Sui Z, Kang C, et al. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data[J]. PloS one, 2014, 9(1): e86026. [↑](#footnote-ref-121)
122. [] 刘正兵, 刘静玉, 何孝沛, 王发曾. 中原经济区城市空间联系及其网络格局分析——基于城际客运流[J].经济地理, 2014, 34(07): 58-66. [↑](#footnote-ref-122)
123. [] 侯赟慧, 刘志彪, 岳中刚. 长三角区域经济一体化进程的社会网络分析[J]. 中国软科学, 2009(12): 90-101. [↑](#footnote-ref-123)
124. [] 陈彦光, 刘继生. 基于引力模型的城市空间互相关和功率谱分析[J]. 地理研究, 2002, 21(6): 742-751. [↑](#footnote-ref-124)
125. [] 李涛, 周锐, 苏海龙等. 长三角区域经济一体化水平的测度:以关系型大数据为基础[C]// 2015中国城市规划年会. 2015. [↑](#footnote-ref-125)
126. [] Hidalgo C A, Klinger B, Barabási A L, et al. The product space conditions the development of nations[J]. Science, 2007, 317(5837): 482-487. [↑](#footnote-ref-126)
127. [] 何大安, 杨益均. 大数据时代政府宏观调控的思维模式[J]. 学术月刊, 2018, 50(05): 68-77. [↑](#footnote-ref-127)
128. [] 罗慧如. 大数据时代的宏观经济调控及政策优化[D]. 广东财经大学, 2016. [↑](#footnote-ref-128)
129. [] 蔡跃洲. 大数据改变经济预测范式[N]. 中国社会科学报, 2015-12-09(004). [↑](#footnote-ref-129)
130. [] （英）弗里德里希·冯·哈耶克. 哈耶克文选：哈耶克论文演讲集[M]. 南京：江苏人民出版社. 2000: 452. [↑](#footnote-ref-130)
131. [] 黄燕芬, 张超. 大数据情绪指数与经济学研究: 现状、问题与展望[J]. 教学与研究, 2018(05): 40-50. [↑](#footnote-ref-131)
132. [] 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究：前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104. [↑](#footnote-ref-132)
133. [] Liu Y, Huang X, An A, et al. ARSA: a sentiment-aware model for predicting sales performance using blogs[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007: 607-614. [↑](#footnote-ref-133)
134. [] Schneider M J, Gupta S. Forecasting sales of new and existing products using consumer reviews: A random projections approach[J]. International Journal of Forecasting, 2016, 32(2): 243-256. [↑](#footnote-ref-134)
135. [] Khadivi P, Ramakrishnan N. Wikipedia in the tourism industry: forecasting demand and modeling usage behavior[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016: 4016-4021. [↑](#footnote-ref-135)
136. [] Bagnoli M, Beneish M D, Watts S G. Whisper forecasts of quarterly earnings per share[J]. Journal of Accounting and Economics, 1999, 28(1): 27-50. [↑](#footnote-ref-136)
137. [] Tumarkin R, Whitelaw R F. News or noise? Internet postings and stock prices[J]. Financial Analysts Journal, 2001, 57(3): 41-51. [↑](#footnote-ref-137)
138. [] 林振兴. 网络讨论、投资者情绪与IPO抑价[J]. 山西财经大学学报, 2011, 33(2): 23-29. [↑](#footnote-ref-138)
139. [] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of computational science, 2011, 2(1): 1-8. [↑](#footnote-ref-139)
140. [] Oh C, Sheng O. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement[J]. 2011. [↑](#footnote-ref-140)
141. [] O'Connor B, Balasubramanyan R, Routledge B, et al. From tweets to polls: Linking text sentiment to public opinion time series[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2010, 4(1). [↑](#footnote-ref-141)
142. [] Zhou S, Shi X, Sun Y, et al. Stock market prediction using heat of related keywords on micro blog[J]. Journal of Software Engineering and Applications, 2013, 6(3B): 37. [↑](#footnote-ref-142)
143. [] Das S R, Chen M Y. Yahoo! for Amazon: Sentiment extraction from small talk on the web[J]. Management science, 2007, 53(9): 1375-1388. [↑](#footnote-ref-143)
144. [] Moat H S, Curme C, Stanley H E, et al. Anticipating stock market movements with Google and Wikipedia[M]//Nonlinear phenomena in complex systems: From nano to macro scale. Springer, Dordrecht, 2014: 47-59. [↑](#footnote-ref-144)
145. [] 钱斌华. 税收大数据预测经济走势的某市经验[N]. 某市日报, 2018-03-15(011). [↑](#footnote-ref-145)
146. [] 崔趁欣, 李岩. 大数据时代国库收入与经济增长耦合性分析及应用前景——以河北省为例[J]. 河北金融, 2018(03): 7-13. [↑](#footnote-ref-146)
147. [] 张秋雁, 宋强. 基于用电大数据的经济分析平台设计研究[J]. 电力大数据, 2017, 20(08): 6-9. [↑](#footnote-ref-147)
148. [] 曲延玲. 基于大数据构建经济金融深度监测分析体系[J]. 金融电子化, 2017(05): 40-41. [↑](#footnote-ref-148)
149. [] Bok B, Caratelli D, Giannone D, et al. Macroeconomic nowcasting and forecasting with big data[J]. Annual Review of Economics, 2018, 10: 615-643. [↑](#footnote-ref-149)
150. [] http://di.baidu.com/solution/finance?castk=LTE%3D [↑](#footnote-ref-150)
151. [] Levenberg A, Simpson E, Roberts S, et al. Economic Prediction using heterogeneous data streams from the World Wide Web[C]//Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE), Proceedings of ECML/PKDD Workshop. 2013. [↑](#footnote-ref-151)
152. [] Sobolevsky S, Massaro E, Bojic I, et al. Predicting regional economic indices using big data of individual bank card transactions[C]//2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017: 1313-1318. [↑](#footnote-ref-152)
153. [] 俞立平. 大数据经济学的概念、框架与学科定位研究[J]. 统计与信息论坛, 2015, 30(06): 3-7. [↑](#footnote-ref-153)
154. [] 刘涛雄, 徐晓飞. 互联网搜索行为能帮助我们预测宏观经济吗?[J]. 经济研究, 2015, 50(12): 68-83. [↑](#footnote-ref-154)
155. [] Götz T B, Knetsch T A. Google data in bridge equation models for German GDP[J]. International Journal of Forecasting, 2019, 35(1): 45-66. [↑](#footnote-ref-155)
156. [] Kholodilin K A, Podstawski M, Siliverstovs B, et al. Google searches as a means of improving the nowcasts of key macroeconomic variables[J]. 2009. [↑](#footnote-ref-156)
157. [] D'Amuri F, Marcucci J. 'Google it!'Forecasting the US unemployment rate with a Google job search index[R]. Nota di Lavoro, 2010. [↑](#footnote-ref-157)
158. [] Artola C, Pinto F, de Pedraza García P. Can internet searches forecast tourism inflows?[J]. International Journal of Manpower, 2015, 36(1): 103-116. [↑](#footnote-ref-158)
159. [] 许 伟. 基于网络大数据的社会经济监测预警研究[M]. 北京：科学出版社, 2016. [↑](#footnote-ref-159)
160. [] 李华杰, 史丹, 马丽梅. 基于大数据方法的经济研究：前沿进展与研究综述[J]. 经济学家, 2018(06): 96-104. [↑](#footnote-ref-160)
161. [] （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018. [↑](#footnote-ref-161)
162. [] 王国成. 行为大数据与计算实验——探索经济研究新方法[J]. 天津社会科学, 2016(03): 86-92. [↑](#footnote-ref-162)
163. [] Hidalgo C A, Hausmann R. The building blocks of economic complexity[J]. Proceedings of the national academy of sciences, 2009, 106(26): 10570-10575. [↑](#footnote-ref-163)
164. [] Bustos S, Gomez C, Hausmann R, et al. The dynamics of nestedness predicts the evolution of industrial ecosystems[J]. PloS one, 2012, 7(11): e49393. [↑](#footnote-ref-164)
165. [] GAO Jian. Modeling local economy complexity[EB/OL]. http://gaocn.net/project.html #complexity, 2015-11-18. [↑](#footnote-ref-165)
166. [] Cristelli M, Tacchella A, Pietronero L. The heterogeneous dynamics of economic complexity[J]. PloS one, 2015, 10(2): e0117174. [↑](#footnote-ref-166)
167. [] Tacchella A, Cristelli M, Caldarelli G, et al. A new metrics for countries' fitness and products' complexity[J]. Scientific reports, 2012, 2(1): 1-7. [↑](#footnote-ref-167)
168. [] Hidalgo C A, Hausmann R. The building blocks of economic complexity[J]. Proceedings of the national academy of sciences, 2009, 106(26): 10570-10575. [↑](#footnote-ref-168)
169. [] 潘迪特, 李昌琪. 时间序列及系统分析与应用[M]. 北京：机械工业出版社, 1988. [↑](#footnote-ref-169)
170. [] 李一智. 经济预测技术[M]. 北京：清华大学出版社, 1991. [↑](#footnote-ref-170)
171. [] 邓聚龙. 灰色系统 社会·经济[M]. 北京：国防工业出版社, 1985. [↑](#footnote-ref-171)
172. [] 李志强, 白文斌, 张亚丽, 等. 基于ARIMA模型的内蒙古羊产业分析与预测[J]. 山西农业科学, 2011, 39(07): 729-732+743. [↑](#footnote-ref-172)
173. [] 王晓梅. 灰色理论GM(1,1)模型在我国畜产品产量预测中的应用[J]. 安徽农业科学, 2007(01): 7-8. [↑](#footnote-ref-173)
174. [] 林绍森, 唐永金. 三种模型对我国粮食产量预测效果的比较[J]. 统计与决策, 2007(07): 39-40. [↑](#footnote-ref-174)
175. [] 刘峰,王儒敬,李传席. ARIMA模型在农产品价格预测中的应用[J]. 计算机工程与应用, 2009, 45(25): 238-239+248. [↑](#footnote-ref-175)
176. [] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. New York: Springer Science & Business Media, 2009. [↑](#footnote-ref-176)
177. [] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net[J]. Journal of the Royal Statistical Society Series B, Statistical Methodology. 2005, 67 (2):301-320. [↑](#footnote-ref-177)
178. [] Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees [M]. CRC Press,1984. [↑](#footnote-ref-178)
179. [] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232. [↑](#footnote-ref-179)
180. [] 谢坤, 容钰添, 胡奉平, 陈桓, 姚小龙. 基于数据集成的随机森林算法[J]. 计算机工程, 2020, 46(12): 290-298. [↑](#footnote-ref-180)
181. [] 林霞, 刘宗尚, 高宇等. 基于机器学习的产油量主控因素分析[J]. 信息系统工程, 2019(12): 94-97+99. [↑](#footnote-ref-181)
182. [] Ayaru L, Ypsilantis P P, Nanapragasam A, et al. Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting[J]. PloS one, 2015, 10(7): e0132485. [↑](#footnote-ref-182)
183. [] 张棪, 曹健. 面向大数据分析的决策树算法[J]. 计算机科学, 2016, 43(S1):374-379+383. [↑](#footnote-ref-183)
184. [] 董莉, 彭凯越, 唐晓彬. 大数据背景下的CPI实时预测研究[J]. 调研世界, 2017(08): 51-54. [↑](#footnote-ref-184)
185. [] 康传利, 顾峻峰, 刘兆威. 梯度提升回归树的旅游流量预测模型[J]. 数学的实践与认识, 2019, 49(15): 251-261. [↑](#footnote-ref-185)
186. [] 巩晓文, 凤思苑, 崔壮等. 基于SVGD分类预测的梯度提升机与随机森林的性能比较[J]. 中国卫生统计, 2019, 36(05): 674-677. [↑](#footnote-ref-186)
187. [] 韩忠明, 原碧鸿, 陈炎等. 一个有效的基于GBRT的早期电影票房预测模型[J]. 计算机应用研究, 2018, 35(02): 410-416. [↑](#footnote-ref-187)
188. [] 黄乃静,于明哲.机器学习对经济学研究的影响研究进展[J].经济学动态,2018(07):115-129. [↑](#footnote-ref-188)
189. [] 葛浩. 基于CART决策树模型的上市企业财务风险预警研究[D].上海财经大学,2020. [↑](#footnote-ref-189)
190. [] 李颖. 基于GBRT集成模型的新冠疫情对我国宏观经济影响研究[D].重庆工商大学,2021.DOI:10.27713/d.cnki.gcqgs.2021.000402. [↑](#footnote-ref-190)
191. [] 贺平,兰伟,丁月.我国股票市场可以预测吗?——基于组合LASSO-logistic方法的视角[J/OL].统计研究:1-15, 2021-12-24. [↑](#footnote-ref-191)
192. [] 欧阳梦倩,周先波,朱君梅.大数据时代下使用互联网搜索量预测CPI——基于LASSO和核偏最小二乘的联合使用[J].金融学季刊,2020,14(02):112-136. [↑](#footnote-ref-192)
193. [] 胡清畅. 互联网搜索行为对宏观经济景气指数预测的影响研究[D].暨南大学,2019. [↑](#footnote-ref-193)
194. [] 陈梦根,任桃萍.新常态经济的CPI预测模型——构建与实证比较[J].调研世界,2020(02):3-8. [↑](#footnote-ref-194)
195. [] 刘晓鑫,景祥云,叶骏骅,金博,毕文祥.关于大数据在中央银行监管应用的文献综述[J].吉林金融研究,2020(10):6-9+18. [↑](#footnote-ref-195)
196. [] 高子宜. 基于随机森林的股价走势预测研究[D].中国政法大学,2021. [↑](#footnote-ref-196)
197. [] 邹玉莹. 基于机器学习的票据转贴现利率预测研究[D].江西财经大学,2020. [↑](#footnote-ref-197)
198. [] 屠星月,于辉辉,郭承坤,阮怀军,陈英义.基于EMD-SVM的农产品市场价格短期预测模型[J].湖北农业科学,2015,54(19):4903-4908. [↑](#footnote-ref-198)
199. [] 肖雪梦. 基于PLS-SVM的重庆市经济发展状况研究[D].重庆大学,2014. [↑](#footnote-ref-199)
200. [] 何雁明,黄邱婧,郑其敏.大数据时代的数量经济模型研究——以BP神经网络的中国CPI预测为例[J].西部金融,2021(02):27-32. [↑](#footnote-ref-200)
201. [] 李杰. 基于神经网络的利率期限结构组合预测研究[D].北京化工大学,2018. [↑](#footnote-ref-201)
202. [] 王永杰. 人工神经网络算法在GDP和CPI中的预测应用[D].中北大学,2017. [↑](#footnote-ref-202)
203. [] 沈虹,李旭,潘琪.基于深度学习长短期记忆神经网络的有色金属期货市场预测研究[J].南京理工大学学报,2021,45(03):366-374. [↑](#footnote-ref-203)
204. [] 李佳,黄之豪,王佳慧.基于独立循环神经网络方法的GDP预测[J].统计与决策,2020,36(14):24-28. [↑](#footnote-ref-204)
205. [] 维克托·迈尔·舍恩伯格, 周涛. 大数据时代 生活,工作与思维的大变革[J]. 人力资源管理, 2013, 03: 176-176. [↑](#footnote-ref-205)
206. [] Mokyr J. Intellectuals and the rise of the modern economy[J]. Science, 2015, 349(6244): 141-142. [↑](#footnote-ref-206)
207. [] 陈争平. 大数据时代与经济史计量研究[J]. 社会科学文摘, 2017(02): 30-32. [↑](#footnote-ref-207)
208. [] 黄燕芬, 张超. 大数据情绪指数与经济学研究：现状、问题与展望[J]. 教学与研究, 2018(05): 40-50. [↑](#footnote-ref-208)
209. [] 汪毅霖. 大数据预测与大数据时代的经济学预测[J]. 学术界, 2016(11): 77-90+325. [↑](#footnote-ref-209)
210. [] Attewell P, Monaghan D, Kwong D. Data mining for the social sciences: An introduction[M]. University of California Press, 2015. [↑](#footnote-ref-210)
211. [] 张涛, 刘宽斌. “大数据”在宏观经济预测分析中的应用[J]. 社会科学文摘, 2018(08): 47-49. [↑](#footnote-ref-211)
212. [] （英）杰弗里·M·霍奇逊. 经济学是如何忘记历史的:社会科学中的历史特性问题[M]. 北京：中国人民大学出版社. 2008: 11. [↑](#footnote-ref-212)
213. [] 汪毅霖. 大数据预测与大数据时代的经济学预测[J]. 学术界, 2016(11): 77-90+325. [↑](#footnote-ref-213)
214. [] Mao H, Counts S, Bollen J. Quantifying the effects of online bullishness on international financial markets[R]. ECB Statistics Paper, 2015. [↑](#footnote-ref-214)
215. [] 张涛, 刘宽斌. “大数据”在宏观经济预测分析中的应用[J]. 社会科学文摘, 2018(08): 47-49. [↑](#footnote-ref-215)
216. [] 黄燕芬, 张超. 大数据情绪指数与经济学研究：现状、问题与展望[J]. 教学与研究, 2018(05): 40-50. [↑](#footnote-ref-216)
217. [] 王建冬, 童楠楠, 易成岐. 大数据时代公共政策评估的变革：理论、方法与实践[M]. 北京：社科文献出版社. 2019: 120-138. [↑](#footnote-ref-217)
218. [] 冯·诺依曼. 数学在科学和社会中的作用[M]. 大连：大连理工大学出版社. 2009: 54-58. [↑](#footnote-ref-218)
219. [] （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018. [↑](#footnote-ref-219)
220. [] Bustos S, Gomez C, Hausmann R, et al. The dynamics of nestedness predicts the evolution of industrial ecosystems[J]. PloS one, 2012, 7(11): e49393. [↑](#footnote-ref-220)
221. [] Tacchella A, Cristelli M, Caldarelli G, et al. A new metrics for countries' fitness and products' complexity[J]. Scientific reports, 2012, 2(1): 1-7. [↑](#footnote-ref-221)
222. [] 汪丁丁. 理解“涌现秩序”(推荐序一)//（美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018 [↑](#footnote-ref-222)
223. [] 布莱恩.阿瑟访谈录（中文版序）//（美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018. [↑](#footnote-ref-223)
224. [] （英）A.N.怀特海. 科学与近代世界[M]. 商务印书馆. 1959. [↑](#footnote-ref-224)
225. [] （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018. [↑](#footnote-ref-225)
226. [] 何新. 反主流经济学[M]. 沈阳：万卷出版公司. 2013: 263. [↑](#footnote-ref-226)
227. [] （英）弗里德里希·冯·哈耶克. 哈耶克文选：哈耶克论文演讲集[M]. 南京：江苏人民出版社. 2000: 450-452. [↑](#footnote-ref-227)
228. [] （英）维克托・迈尔-舍恩伯格, 肯尼思・库克耶等. 大数据时代：生活、工作与思维的大变革[M]. 浙江人民出版社, 2013: 2. [↑](#footnote-ref-228)
229. [] 徐晋. 大数据经济学[M]. 上海：上海交通大学出版社. 2014: 96. [↑](#footnote-ref-229)
230. [] 宋圭武. 大数据时代背景下的经济学[J]. 发展, 2014(07): 5. [↑](#footnote-ref-230)
231. [] Anderson C. The end of theory: The data deluge makes the scientific method obsolete[J]. Wired magazine,2008, 16(7): 16-07. [↑](#footnote-ref-231)
232. [] 黄璜. 社会科学研究中“基于主体建模”方法评述[J]. 国外社会科学, 2010(05): 40-47. [↑](#footnote-ref-232)
233. [] Holland J H, Miller J H. Artificial adaptive agents in economic theory[J]. The American economic review, 1991, 81(2): 365-370. [↑](#footnote-ref-233)
234. [] （美）布莱恩.阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018. [↑](#footnote-ref-234)
235. [] Janssen M A, Ostrom E. Empirically based, agent-based models[J]. Ecology and society, 2006, 11(2). [↑](#footnote-ref-235)
236. [] Janssen M A, Ostrom E. Adoption of a new regulation for the governance of common-pool resources by a heterogeneous population[J]. Inequality, Cooperation, and Environmental Sustainability, 2006: 60-96. [↑](#footnote-ref-236)
237. [] 黄璜. 社会科学研究中“基于主体建模”方法评述[J]. 国外社会科学, 2010(05): 40-47. [↑](#footnote-ref-237)
238. [] 汪丁丁. 行为经济学讲义：演化论的视角[M]. 上海：上海人民出版社. 2011: 230. [↑](#footnote-ref-238)
239. [] Tesfatsion L. Agent-based computational economics: Growing economies from the bottom up[J]. Artificial life, 2002, 8(1): 55-82. [↑](#footnote-ref-239)
240. [] （美）布莱恩·阿瑟. 复杂经济学：经济思想的新框架[M]. 杭州：浙江人民出版社. 2018. [↑](#footnote-ref-240)
241. [] 王国成. 行为大数据与计算实验——探索经济研究新方法[J]. 天津社会科学, 2016(03): 86-92. [↑](#footnote-ref-241)
242. [] 蔡永鸿, 唐斯琪, 于娟. 大数据时代宏观经济与微观企业的行为联接[J]. 中国市场, 2015(08): 88-89. [↑](#footnote-ref-242)
243. [] 赵怡君. 大数据时代宏观经济与微观企业的行为联接探讨[J]. 全国流通经济, 2018(08): 75-76. [↑](#footnote-ref-243)
244. ————————————————————

     [] 数据抽取维度包括企业增量、企业存量、企业注销数量、企业吊销数量、企业平均存续时长（月）、企业登记注册资本比例等。 [↑](#footnote-ref-244)
245. [] 行业大类包括农、林、牧、渔业，采矿业，制造业，电力、热力、燃气及水生产和供应业，建筑业，批发和零售业，交通运输、仓储和邮政业，住宿和餐饮业，信息传输、软件和信息技术服务业，金融业，房地产业，租赁和商务服务业，科学研究和技术服务业，水利、环境和公共设施管理业，居民服务、修理和其他服务业，教育，卫生和社会工作，文化、体育和娱乐业，公共管理、社会保障和社会组织，国际组织等。 [↑](#footnote-ref-245)
246. [] 数值加工处理方式主要包括原始值、标准差、同比、熵值等。 [↑](#footnote-ref-246)
247. [] 40.Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. STL: A seasonal- trend decomposition procedure based on Loess (with discussion). Journal of Official Statistics, 6: 3–73, 1990. [↑](#footnote-ref-247)
248. [] U.S. Census Bureau. Reference manual for the X-13ARIMA-SEATS Program, version 1.0 URL http://www.census.gov/ts/x13as/docX13AS.pdf, 2013. [↑](#footnote-ref-248)