基于多源异步混频 CPI 数据的预测方法研究

张 虎 沈寒蕾 夏 伦

【摘 要】研究目标:基于线上消费者价格指数和网络搜索价格指数预测 CPI。研究方法:在卷积神经网络(CNN)框架中融合 MADL_MIDAS 模型,建立异步混频卷积神经网络(AMCNN)模型,并通过选用 2016 年 1 月至 2019 年 12 月的数据验证该方法的有效性。研究发现:日度线上 CPI 及日度网络搜索指数属于 CPI 的领先指标,同时引入并保留原有数据特征有助于改进 CPI 预测精度,提高 CPI“拐点”捕捉能力。研究创新:揭示了高频日度线上 CPI 和网络搜索数据对低频月度 CPI 的预测能力,提出了一种融合神经网络与传统计量模型的异步混频数据处理方法。研究价值:预测 CPI 波动水平和“拐点”时,可辅助利用线上 CPI、网络搜索高频数据,结合 AMCNN 模型提高预测精度。AMCNN 模型可用于处理异步混频数据、探究变量间复杂不确定性(线性、非线性)关系,具有很强的适应性和扩展性,可应用于其他经济、金融领域,应用价值较高。

【关 键 词】AMCNN 模型;CPI 预测;网络搜索指数;线上 CPI;异步混频数据

【作者简介】张虎,沈寒蕾,中南财经政法大学统计与数学学院;夏伦,湖北经济学院信息管理与统计学院,湖北经济学院湖北数据与分析中心。

【原文出处】《数量经济技术经济研究》(京),2020.10.149~168

【基金项目】本文获得中央高校专项课题“基于混频数据的宏观经济数据质量改进”(412/31510000111)、2020 年学科建设专项资助项目“统计学科‘双一流’监测指标提升项”(31712011201)、2020 年一流学科重点建设项目“经济统计学产学融合培养模式与实践”(21123541830)、国家社科基金青年项目“基于空间统计分析视角的生产性服务业集聚与制造业转型升级研究”(17CTJ011)的资助。

一、问题的提出

消费者价格指数(Consumer Price Index,CPI)一直是宏观经济和经济统计领域关注的热点,是学术界研究最多的物价指数。CPI 能够反映国民生活质量、分析物价通胀或紧缩程度、刻画经济活跃状况、指导货币政策制定等。目前中国 CPI 的编制、抽样、计算以及公布主要由国家统计局负责,CPI 以“固定篮子指数”理论为编制基础,以“定人、定点、定时”为抽样原则,采用国际通行的链式拉氏公式逐级加权平均,最后以月为周期借助新闻发布会或官方网站公布。

近年来我国互联网发展迅速,截至 2019 年 6 月,中国网民规模为 8.54 亿人,互联网普及率达 61.2%。随着网民数量的迅速增长,网民搜索行为一定程度上体现出个体“认知系统”的变化。面对输入信息的“确定性”与“不确定性”,个体通过网络搜索行为进一步寻求信息的真实性。个体为满足自身特定的某一目标需求,通过网络搜索关键词反映其对信息的关注状态及独立思考,个体信息的不对称和特定的目标需求驱动其产生相关的搜索行为。网络搜索数据作为新的数据源,由于其客观性、实时性、高频性、代表性以及与相关经济变量的关联性,诸多学者在宏观经济指数研究中增加网络搜索数据提高预测精度。与此同时,随着电子商务

的迅速发展,网上购物成为居民常见的消费方式。据商务部统计,2013～2019年,中国网络购物规模快速增长,其间每年的销售额分别以59.4%、46.8%、36.8%、23.9%、29.6%、25.4%、16.5%的增长率增长,并呈现出网购消费份额不断提升的趋势,2019年全国网购销售额超过10万亿元。截至2019年6月,我国网络购物用户规模达6.39亿,占网民整体的74.8%,按照这种趋势未来线上消费将进一步发展。随着居民消费模式和消费习惯的改变,现行CPI测算存在一定的不足和局限性,具体表现在:第一,数据采集偏差。当前CPI包括线下抽取的超市、农贸市场等价格采集点以及少量线上价格采集点,样本规模较小。第二,逐层汇总偏差。当前CPI是国家统计局依据基层统计调查数据由县、市、省逐级多阶段多维度加权平均计算得到,中间环节过多,增加了数据统计误差。第三,权重设置偏差。当前CPI权重设置不透明,部分学者依据编制原理估算各类权重,发现食品烟酒类权重过高,教育、医疗以及居住类权重过低。第四,类目代表性偏差。目前CPI计算参数采用"五年一大调,一年一小调"的原则,容易导致样本代表性跟不上产品或服务的发展速度。线上CPI具有全面的测算范围、自动的测算流程、高频的数据发布等优势,保证了线上CPI的准确性、高频性以及时效性。因此,网络搜索数据以及线上CPI为当前CPI预测提供了新的思路。如何引入网络搜索数据和线上CPI提高CPI预测精度,如何充分利用网络搜索数据以及线上CPI的原始高频数据特征提高CPI预测时效性,这两点是本研究要解决的核心问题。

二、文献述评

近年来,基于大数据技术的CPI研究主要集中在两个方面。第一,基于网络搜索大数据(如百度搜索指数)的CPI预测;第二,基于网络购物大数据(如线上消费者物价指数)的线上线下CPI研究。其中,基于网络搜索大数据的CPI预测,代表性研究包括:张崇等(2012)采用2004年1月至2009年8月全国的CPI月同比数据与近300个关键词的Google搜索周数据合成月度宏观形势指数和供求关系指数,研究其与CPI的关系,分析发现二者之间存在较高的协整关系,且宏观形势指数和供求关系指数分别领先于官方CPI 5个月和两个月,网络搜索指数对CPI转折点也存在一定的预测能力。孙毅等(2014)探索网络搜索指数的有效合成方法,利用2006年6月到2012年12月与全国CPI月度数据相关系数大于0.5的14个关键词百度周指数,分别运用主成分和逐步回归进行指数合成,结果表明主成分分析法能够更好地预测CPI走势。随后,孙毅等(2014)进一步研究了网络搜索指数与通货膨胀的关系,发现二者存在长期协整关系且实际通货膨胀是网络搜索指数的格兰杰原因。董倩(2016)采用2013年11月至2015年11月北京市月同比CPI数据作为因变量,选取与雾霾经济相关的4个关键词百度搜索指数作为自变量,利用线性回归和支持向量机进行拟合,结果表明搜索指数提高了预测能力。董莉等(2017)利用2011年1月至2016年8月93个CPI构成商品及服务的关键词百度搜索指数,基于Elastic Net惩罚因子的分布滞后模型分别构建全国、城镇、乡村的CPI实时预测模型。徐映梅和高一铭(2017)利用2006年6月1日至2015年12月31日的"CPI""物价""价格""通货膨胀""涨价""降价""通货紧缩"7个关键词的电脑端日度百度搜索指数选取具有领先作用的关键词与2006年6月至2015年11月国家统计局公布的全国CPI月同比数据分别基于门限回归构建低频CPI指数、基于动态因子模型构建高频CPI指数分别预测CPI,结果表明CPI低频和高频舆情指数分别可领先官方CPI月同比数据40天、65天左右。刘宽斌和张涛(2018)利用2016年1月至2018年2月279个关键词的百度搜索日度指数构建4种不同的网络搜索指数,结合国家统计局公布的全国CPI月度环比数据,基于混频数据抽样模型(MIDAS)进行CPI短期预测及拐点检测,结果发现可领先官方公布CPI半个月获得较为准确的CPI数据。

基于网购大数据的CPI研究,学者们主要采用

阿里研究院构建的网购商品价格指数(iSPI、aSPI),研究线上 CPI 与线下 CPI 的关系。其中代表性研究包括:杜两省和刘发跃(2014)用 STAR 模型研究了 iSPI 和 CPI 在各类别上是否符合一价定理。刘发跃和马丁丑(2015)使用定性方法和 HP 滤波法,对月度同比 iSPI 和 CPI 在各类别的收敛性上进行了研究。米子川和姜天英(2016)利用 2012 年 1 月至 2016 年 1 月的 aSPI 指数和 CPI 指数的同比数据,依次采用协整、EMD 分解、Lasso 回归探究二者之间的同步性、相关性以及传导性,结果表明二者之间具有较好的同步性、一定的相关性且 aSPI 是 CPI 的先行指标。周薇薇和田涛(2016)基于 2011 年 2 月至 2015 年 9 月阿里研究院公布的刻画网络商品价格波动的月度环比物价指数 aSPI 和国家统计局发布的 CPI 月度环比数据,探究线上线下价格指数的关联性,研究发现二者之间长期具有相似的随机波动趋势,而短期波动存在相互同步影响。同年,田涛(2016)进一步利用动态相关模型研究了 aSPI 与 CPI 的月度环比价格指数波动溢出效应,且线上商品对线下商品的价格影响更大。韩胜娟和张敏(2017)选取 2012 年 1 月至 2016 年 12 月阿里研究院公布的 aSPI、aSPI - core 指数和国家统计局公布的 CPI 年同比价格指数定量探究阿里价格指数与官方价格指数的融合程度,结果发现二者的趋势性逐渐一致,但阿里价格指数由于数据来源的差异性其波动性更明显,价格变动更迅速。方匡南和曾武雄(2018)选取 2011 年 2 月至 2015 年 8 月 aSPI 和 CPI 的月环比数据和相应子类指数,采用 Hodrick - Prescott 滤波分析模型和交叉谱分析方法,研究二者之间变动在时间上的先后关系,研究发现二者存在 11.3 个月的匹配周期且 aSPI 领先 CPI 指数 1.02 个月。唐礼智和刘玉(2018)不仅探讨了月度环比 aSPI - core 和 CPI 之间的关联性,还就分类权重和分类市场两方面探讨了线上线下 CPI 的收敛性。

综上所述,首先,从研究视角来看,CPI 的预测大多基于网络搜索大数据进行研究,且主要集中在网络搜索大数据与线下 CPI 的同频数据研究,少有基于混频数据模型进行官方 CPI 预测。引入线上 CPI 的研究,主要研究线上 CPI 与官方 CPI 二者之间的相关性、差异性以及替代性,鲜有基于线上 CPI 对官方 CPI 进行预测的相关研究。尚无同时包含官方 CPI、网络搜索大数据、线上 CPI 三者的相关研究。本研究在已有研究的基础上,尝试同时引入高频网络搜索大数据和高频线上 CPI 数据基于混频数据模型进行官方 CPI 预测。另外,分别以官方 CPI、官方 CPI 与网络搜索大数据、官方 CPI 与线上 CPI 构建基准预测模型与本文提出的官方 CPI 预测模型的预测精度进行对比,探究同时结合 CPI、线上 CPI 以及网络搜索大数据是否能够更加全面、准确、及时地预测 CPI,为政府、企业及个人科学、准确作出宏观经济决策提供参考。其次,从研究方法来看,网络搜索大数据与线下 CPI 的研究大多采用传统计量经济模型,鲜有使用机器学习算法或混频数据抽样模型(MIDAS)进行相关研究。已有文献运用传统计量经济模型(如协整分析、主成分分析、线性回归分析、分布滞后模型、门限回归、动态因子模型等)或机器学习(如支持向量机、交叉谱分析等)研究 CPI,主要针对同频数据(如月度同比或环比的 CPI 数据、网络搜索数据、线上 CPI 数据等)展开研究。信息时代将产生更多高频数据(如日度网络搜索数据、日度线上 CPI 数据),为实时、精确的宏观经济研究提供了可能,但日度高频数据和以月为周期发布的 CPI 数据存在频率差异。若依然选用处理同频数据的传统计量经济模型或机器学习算法进行 CPI 研究,则需要以丢失高频数据信息为代价,将高频数据转化为低频数据,或者将低频数据转化为高频数据产生大量冗余信息。对于混频数据需要运用新的分析方法,以充分利用原始数据信息,得到更为精确的官方 CPI 预测结果。已有少量学者采用 MIDAS 或其改进模型针对同步混频数据(即高、低频数据的频率倍差保持不变,如高频月度数据与低频季度数据之间的频率倍差恒定为 3)进行 CPI 预测研究。但当高、低频数据的频率倍差或多变量高频数据之间的频率倍差并不完全同步时,

已有研究大多对这类异步混频数据(如日度数据与月度数据由于每月天数不完全相同存在异步混频现象)事先进行对齐处理,转化为同步混频数据,此时也损失了一定的数据信息。因此,已有文献尚未很好地解决异步混频数据的预测问题。本研究旨在更好地利用多源异步混频数据变量的原始信息,弥补目前研究的不足,同时实现变量间线性或非线性关系的探究,将传统混频数据模型与卷积神经网络结合构建异步混频卷积神经网络模型(Asynchronous Mixed Frequency Convolutional Neural Networks model,AMCNN),采用月度官方 CPI、日度网络搜索数据以及日度线上 CPI 进行官方 CPI 预测。AMCNN 模型不仅可以充分利用原始数据特征,满足多源异步混频数据的预测要求,而且由于神经网络结构的灵活性,能够丰富模型的结构,提高研究对象的预测精度。

基于现有研究本文拟做出以下改进:一是同时引入高频网络搜索大数据和高频线上 CPI 对官方 CPI 进行预测,一定程度上丰富了现有文献的研究视角,增强研究结果的合理性;二是构建 AMCNN 模型,与已有文献研究的优势在于充分利用了现有数据的原始特征,其灵活的模型结构扩展了已有文献中线性关系的研究,为类似研究问题提供参考;三是对比引入 CPI、线上 CPI 以及网络搜索指数不同变量组合下,分别运用同频数据模型(如 ADL)、传统混频数据模型(如 MIDAS)以及本文提出的 AMCNN 模型预测官方 CPI,研究发现同时引入 CPI、线上 CPI 以及网络搜索指数并运用 AMCNN 模型预测官方 CPI 或 CPI 子类别时预测误差均最小,除此之外,在不同模型的共同预测时间区域中,AMCNN 模型对 CPI“拐点”的捕捉能力不论是查全率、查准率还是综合准确率均优于同频数据模型 ADL 和传统混频数据模型 MIDAS,验证了本文选取的变量以及提出的方法具备科学性和准确性;四是采用参数自适应网格搜索方法确定最优参数估计结果,使用实证研究得到引入高频网络搜索大数据和线上 CPI 的最优数据周期和最优滞后阶数,研究发现提前 1 个月预测 CPI、CPI 自回归阶数为 1、高频日度 ICPI 和高频日度网络搜索指数滞后 30 ~ 45 天时,AMCNN 模型可提供比较精准的 CPI 预测结果,并且比国家统计局公布的时间提前 43 ~ 58 天,具备良好的时效性。本文后续结构安排如下:第一部分,问题的提出,结合当前现状,提出本文研究主题;第二部分,文献述评,总结当前 CPI 研究视角和研究方法,提出本文拟改进之处;第三部分提出本研究解决问题的思路和方法,详细阐述本文的研究思路,构建预测模型;第四部分为 CPI 预测的实证分析,包括数据来源、变量选取、数据预处理和实证结果分析;第五部分,本文的研究结论与展望。

三、CPI 预测模型构建

从中国社会经济发展现状出发,引入新的高频大数据资源预测 CPI,其一般性研究步骤为:数据获取、数据预处理、模型构建、结果分析。本文采用网络搜索大数据和线上 CPI 数据构建多变量异步混频卷积神经网络模型预测 CPI。

其整体研究思路如下:第一步,数据获取。主要包括国家统计局公布的官方 CPI 月度上月环比增长率数据、清华 iCPI 项目团队发布的线上 CPI 月度、日度不同周期的 ICPI 上月环比增长率数据以及移动端和电脑端整体的网络搜索日度数据。除此之外,还包括一些数据处理过程中涉及的相关数据如,网民规模数据、搜索引擎市场占有率数据。第二步,数据预处理。CPI 和线上 CPI 来自权威机构已整理好的数据,获取方式较为简单,数据质量比较高,数据处理方便。网络搜索大数据需利用网络爬虫技术编写相应的代码实现,同时由于数据采集过程中一些不稳定因素的存在,数据的处理过程比较复杂。第一,基于搜索引擎市场占有率数据确定本研究将选择的目标搜索引擎;第二,基于本文研究对象及研究目标确定网络搜索的初始关键词;第三,编写目标搜索引擎指定关键词在指定时间段全国移动端和电脑端整体的日搜索量数据;第四,对爬取数据进行缺失值检测与补充,异常值检测;第五,利用指定时间段网民数量变化以及目标搜索引擎市场占有率变化对爬取的日搜索量数据进行可

比性处理,消除由于其他因素导致的搜索量变动;第六,探索各初始关键词与官方 CPI 的时差相关系数,选取具有官方 CPI 领先相关性的关键词极其对应网络搜索数据作为进一步的研究变量;第七,选定合适的指数合成方法构建日度 CPI 网络搜索指数。第三步,预测模型构建。由于现有数据属于多变量时序数据且数据获取平台的差异、数据统计频率的不同、各变量间高低频频率倍差不等,各变量间不确定的线性或非线性关系,本文将构建多变量异步混频卷积神经网络模型,通过检测各特征值与标签值的混频倍差、滞后阶数自适应调整相关参数值,不限于变量间线性关系的探究,以预测精度作为主要评价指标。基于此构建的多变量异步混频卷积神经网络模型可根据不同需要实现对应的具体应用。第四步,实证分析。基于现有数据采用本文提出的预测模型进行官方 CPI 预测,同时在已有文献的基础上构建全面系统的基准模型进行结果的对比分析,对本文提出的多变量异步混频卷积神经网络模型的预测性能、泛化能力、稳健性进行验证和评估。

1. ISI 数据获取、预处理及合成

(1)搜索引擎选取。当前主流的搜索引擎主要有百度、谷歌、必应、360、搜狗等,本文将选取在中国市场占有率最高的搜索引擎作为网络搜索数据的数据来源。StatCounter[①],它是美国一家网站通讯流量监测机构,提供各种类型的统计报告以及网站流量统计服务。本文获取了中国 2016 年 1 月 1 日至 2019 年 12 月 31 日中国内陆搜索引擎市场占有率日度数据,主流搜索引擎市场占有率变化如图 1 所示。本文选取市场份额一直居于第一的百度统计的百度指数作为关键词搜索数据来源。

(2)初始关键词选取。本文主要利用网络搜索数据和 ICPI 对 CPI 进行预测研究。关键词的选取应满足以下几条基本原则:第一,契合研究目的;第二,符合物价波动理论;第三,理论上与 CPI 具有相关性;第四,满足多数网民搜索习惯;第五,属于百度指数收录的关键词;第六,能够为研究提供数据信息。基于关键词初选原则,本文分别从宏观和微观层面考虑构建初始关键词池。

宏观层面的关键词,一是货币因素,如贷款利率、再贴现率等;二是输入性因素,如国际原油价格、外汇储备等;三是直观感受类,如 CPI、通货膨胀等。上述三方面的考虑符合物价波动理论及主观上与 CPI 直接相关的描述。微观层面的关键词一是从供需关系、成本因素出发,选取相关搜索词;二是保证理论上与 CPI 具有相关性,挑选构成 CPI 八大类食品烟酒类、衣着类、居住类、生活用品及服务类、交通和通信类、教育文化和娱乐类、医疗保健类、其他用品和服务类对应的关键词;最后基于关键词能否转变为研究相关的数据,主要从百度指数关键词收录检测以及对应关键词指定时间段的搜索量两方面筛选出最终的初始关键词。

(3)数据获取及预处理。百度指数[②]提供了关键词在指定时间段的搜索量折线图,数据无法直接下载,将通过编程提取相关搜索数据,然后得到消除网民规模变化及搜索引擎市场占有率波动影响后的搜索量数据 b_{it}。

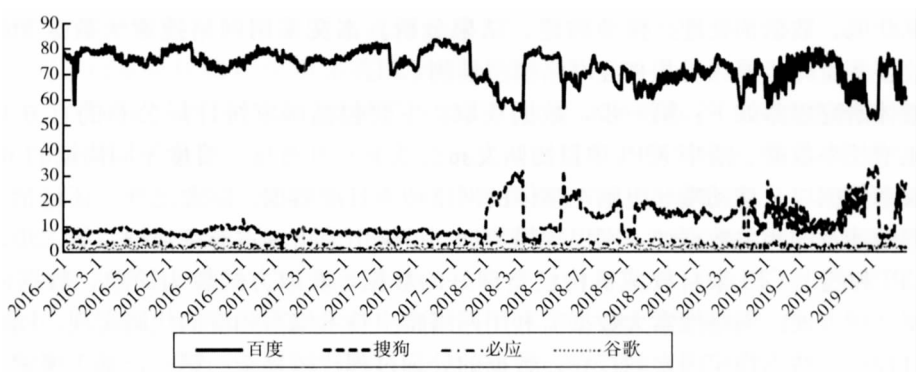

图 1 主流搜索引擎市场占有率

$$b_{it} = \log \frac{d_{it}}{p_t q_t} - \log \frac{d_{it-1}}{p_{t-1} q_{t-1}} \quad (1)$$

其中，d_{it}表示第 i 个关键词在时间 t 的搜索频数，p_t、q_t分别为时间 t 全国网民规模和百度搜索市场占有率。

与 CPI 有先行关系的变量能够在通货膨胀或紧缩前率先发生变化，有利于预测 CPI 变动趋向。本文将选取相对 CPI 具有先行关系的关键词合成网络搜索指数用于预测 CPI。时差相关系数通过计算两变量间的时序关系，验证经济时间序列先行、一致或滞后关系，如 X 指标领先、滞后或同步于 Y 指标。其数学表达式如下：

$$r_L = \frac{\sum_{t=1}^{n}(X_{t-L} - \overline{X})(Y_t - \overline{Y})}{\sqrt{\sum_{t}^{n} = 1(X_{t-L} - \overline{X})^2 (Y_t - \overline{Y})^2}} \quad -M_B \leqslant L \leqslant M_B \quad (2)$$

其中，X，Y 分别是两个时间序列，Y 为基准指标，L 为超前或滞后期，被称为时差或延迟数（$L=0$ 时表示不移动，表示同步；若表示超前，则对应的 $-M_B \leqslant L < 0$；若表示滞后，则对应的 $0 < L \leqslant M_B$）；n 为所取数据的个数；M_B表示最大延迟数，且 $M_B < n$。取不同的 L 值，分别代表不同的时差，并计算时差相关系数 r_L，取绝对值最大的值 r_{L_max}作为时差相关系数。时差相关系数 r_{L_max}对应的时差数 L 的正负，可判断被选指标与基准指标先行、同步或滞后的关系。它反映了被选择指标与基准指标的波动最接近，即为时差相关系数。常将超前相关性较强的指标选为预测性指标。

（4）搜索指数合成。由于关键词数量较多且部分关键词搜索数据之间存在较强的相关关系。本文将采用主成分分析方法，采用降维思想将多个关键词搜索数据用较少的综合搜索指标来代替，这样既涵盖了足够的信息，又可以得到相互间不相关的变量，更有利于官方 CPI 的预测分析。利用主成分分析方法提取出相关系数矩阵中取值大于 1 的特征值对应的主成分，进而构建不同频率的 CPI 网络搜索指数。

2. 异步混频卷积神经网络模型（AMCNN）构建动机

CPI 的变动在一定程度上反映了通货膨胀或紧缩程度，长期以来备受市场参与者、央行等部门的关注。CPI 的预测方法主要分为两类：理论驱动型和数据驱动型。理论驱动型预测结果依赖于特定的理论假设、经济理论和经验事实。数据驱动型较少依赖于经济理论，更注重用数据说话，是一种更加简单、预测精度较高的方法。本文提出的 CPI 预测模型灵感源于数据类型从同频时间序列推广到混频时间序列进而到异步混频时间序列，变量关系从因变量自相关到自变量自相关到二者均存在自相关，变量数量从一元到多元，模型形式从线性关系到非线性关系。最终提出满足多元非线性异步混频双自相关的 CPI 预测模型，其构建思路如图 2 所示。

CPI 的变动存在一定的“惯性”，具有时间上的滞后效应。即解释变量与被解释变量的因果联系不可能在短时间内完成，在这一过程中通常都存在时间滞后，即解释变量需要通过一段时间才能完全作用于被解释变量。由于经济活动的惯性、人们心理预期、技术、制度等原因，CPI 前期态势往往会延续到本期，从而形成 CPI 的当期变化同自身过去取值水平相关的情形。这种被解释变量受自身或其他变量（如线上 CPI、网络搜索指数）过去值影响的现象称为滞后效应。过去时期对当前被解释变量产生影响的变量，称为滞后变量。滞后变量分为滞后解释变量与滞后被解释变量。

当滞后变量为解释变量时，即被解释变量受解释变量当期值及其若干期滞后值的影响，模型形式如下：

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_5 X_{t-s} + \mu_t \quad (3)$$

具有这种滞后分布结构的模型称为分布滞后模型。其中，s 为滞后长度，根据滞后长度 s 取为有限和无限，模型分别称为有限分布滞后模型和无限分布滞后模型。在分布滞后模型中，各系数体现了解释变量的各滞后值对被解释变量的影响程度，即乘数效应。

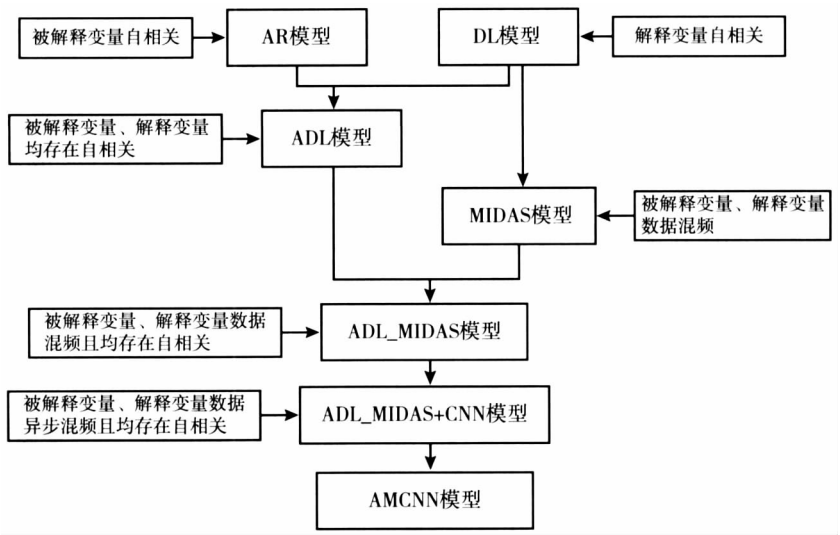

图 2 AMCNN 模型构建思路

当滞后变量为被解释变量时,即滞后变量模型的解释变量包括自变量 X 的当期值和被解释变量 Y 的若干期滞后值,模型形式如下:

$$Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + \cdots + \gamma_q Y_{t-q} + \mu_t \quad (4)$$

称这类模型为自回归模型,其中 q 称为自回归模型的阶数。

把滞后变量引入回归模型,这种回归模型称为滞后变量模型(ADL)。综合式(3)和式(4)可知,其一般形式为:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_s X_{t-s} + \gamma_1 Y_{t-1} + \cdots + \gamma_q Y_{t-q} + \mu_t \quad (5)$$

其中,s、q 分别为解释变量和被解释变量的滞后期长度。当 X 为多维外生变量时,各变量的滞后阶数可取不同数值。

实际建模时学者们发现传统时间序列计量模型的应用瓶颈,即要求所有数据变量同频。国家统计局 CPI 的发布周期为月度值,但较多影响 CPI 变量的频率高于月度,如日度线上 CPI、日度网络搜索指数等。传统建模方法,通常需要在建模之前将高频日度数据转化为与 CPI 相同的低频月度数据,然后进行参数估计与检验。这样会损失原有的高频数据信息,增加预测偏差。因此,在统计局 CPI 公布滞后 12 天左右的情况下,利用日度指标对月度 CPI 进行实时、超前预报具有重大现实意义。为解决基于不同频率变量所构建模型的参数估计和预测,弥补传统时间序列预测模型要求变量同频率的缺陷,Ghysels 等(2004)基于分布滞后模型式(3)提出了 MIDAS(Mixed Data Sampling, MIDAS)模型。一元混频数据回归预测模型 MIDAS 的模型形式如下:

$$Y_t = \alpha + \beta(\omega_0 X_t^m + \omega_1 X_{t-\frac{1}{m}}^m + \cdots + \omega_m X_{t-1}^m + \cdots + \omega_{2m} X_{t-2}^m + \cdots + \omega_{sm} X_{t-s}^m) + \mu_t \quad (6)$$

式(6)简记为:$Y_t = \alpha + \beta W(\theta, L) X_t^m + \mu_t$。其中,$Y_t$ 为第 t 期低频被解释变量,X_t^m 为高频解释变量,m 为高频数据与低频数据的频率倍差,α、β 为未知参数,为权重函数的参数,L 为延迟算子,s 同式(3)为解释变量滞后阶数,$\mu_t \sim N(0, \sigma^2)$ 为随机扰动项,$W(\theta, L) = \sum_{i=0}^{sm} \omega_i(\theta) L^{i/m}$ 且 $\sum_{i=0}^{sm} \omega_i(\theta) = 1$,$L^{i/m} X_t^m = X_{t-\frac{i}{m}}^m$。

由于 CPI 波动具有“惯性”,第 t 期的 CPI 与其自身的滞后项相关,因此,在式(6)一元 MIDAS 模型基础上引入滞后被解释变量,即 ADL - MIDAS。其模型形式如下:

$$(1 - \Phi(\gamma, L)) Y_t = \alpha + \beta W(\theta, L) X_t^m + \mu_t \quad (7)$$

其中,$\Phi(\gamma, L) = \sum_{i=1}^{q} \gamma_i L^i$,$q$ 为自回归滞后阶数。

CPI 波动不仅与自身过去数据相关还与其他变量相关,综合式(6)和式(7)可知多元混频数据自回

归预测模型,即 MADL – MIDAS。模型表达式如下:

$$(1-\Phi(\gamma,L))Y_t=\alpha+\beta_1 W_1(\theta^1,L)X_{1,t}^{m_1}+\beta_2 W_2(\theta^2,L)X_{2,t}^{m_2}+\cdots+\beta_K W_K(\theta^K,L)X_{K,t}^{m_K}+\mu_t \quad (8)$$

式(8)简记为:$(1-\Phi(\gamma,L))Y_t=\alpha+\sum_{k=1}^{K}\beta_k W_k(\theta^k,L_k)X_{k,t}^{m_k}+\mu_t$,$K$ 为解释变量个数。

近年来,MIDAS 被越来越多的学者用于宏观经济混频数据模型的研究与应用,且证明 MIDAS 能够很好地解决数据频率不同时的模型估计与预测问题。MIDAS 无须进行数据同频化处理,通过参数化的紧凑权重,可直接将不同频率变量引入同一模型中,并结合混频变量的样本数据,选取不同形式的权重函数和具体的滞后阶数,得到对应的参数估计值,实现基于高频数据的超前性对低频被解释变量做出预报。Ghysels 等(2007)提出阿尔蒙多项式函数(Almon)、指数阿尔蒙函数(Exponential Almon)、贝塔密度函数(Beta)、分段函数(Step Function),确定 4 种不同形式的权重函数 $w_i(\theta)$,将其代入式(7)可构建 MADL – MIDAS(A)、MADL – MIDAS(EA)、MADL – MIDAS(B)、MADL – MIDAS(S)共 4 种多元混频数据自回归预测模型。其权重函数具体形式在此不做赘述。

上述模型不论是基于同频时间序列样本数据的分布滞后模型、自回归模型,还是基于混频时间序列样本数据的 MADL – MIDAS 模型其参数估计都是建立在事先设定好模型和权重形式基础上完成的。相较于 ADL 模型,MADL – MIDAS 模型虽解决了变量间不同频率的问题,但对于不规则混频数据(低频被解释变量与高频解释变量各期对应的频率倍差不规则、不固定),其一般形式如下:

$$\begin{aligned}(1-\Phi(\gamma,L))Y_t=&\alpha+\beta_1(\omega_{1,0}X_{1,t}^{m_1,t}+\omega_{1,1}X_{t-\frac{1,t1}{m_1,t}}^{m_1,t1}+\cdots+\omega_{1,m_{1,t-1}}X_{t-1}^{m_1,t-1}+\cdots+\omega_{1,m_{1,t-1}+m_{1,t-2}}X_{t-2}^{m_1,t-2}+\cdots+\omega_{1,\sum_{l=0}^{s_1}m_{1,t-s_1}}X_{t-s_1}^{m_1,t-s_1})+\cdots+\beta_K(\omega_{K,0}X_{K,t}^{m_K,t}+\omega_{K,1}X_{t-\frac{1,t1}{m_K,t}}^{m_K,t1}+\cdots+\omega_{K,m_{K,t-1}}X_{t-1}^{m_K,t-1}+\cdots+\\&\omega_{K,m_{K,t-1}+m_{K,t-2}}X_{t-2}^{m_K,t-2}+\cdots+\omega_{K,\sum^{s_K}l=0m_{K,t-s_K}}X_{t-s_K}^{m_K,t-s_K})+\mu_t\end{aligned} \quad (9)$$

式(9)形式复杂,参数估计非常困难,而神经网络模型在模型形式和数据格式上更加灵活。基于上述背景,本文尝试将传统的 MADL – MIDAS 模型与神经网络模型相结合,构建 AMCNN 模型。在实现多源异步混频变量数据参数估计和预测过程中,保留原始时间序列样本数据的初始特征,充分挖掘变量间可能的形式,提高预测性能。

3. AMCNN 模型构建

AMCNN 模型的原理是参照 MADL – MIDAS 模型保留被解释变量和解释变量原始的时间序列特征,将原始多源混频变量按时间序列进行独热编码,将多变量混频数据用更加紧凑的方式表示出来,作为 AMCNN 模型的数据输入格式。利用卷积神经网络和适当的非线性激活函数探究变量间的非线性形式,不是变量间简单的线性关系预测,学习出的权重也不是简单的固定系数。

卷积层承担模型 AMCNN 中的大量计算,其参数由若干可学习的滤波器集合构成。卷积是两实值函数间的数学运算。设:$f(x)$,$g(x)$是定义域 R 上的两个可积函数,作积分:$\int_{-\infty}^{+\infty}f(\tau)g(x-\tau)d(\tau)$,可知对于任意的实数 x,该函数均可积。

$$b(x)=f(x)\otimes g(x) \quad (10)$$

式(10)称为函数 $f(x)$ 与 $g(x)$ 的卷积。其中,$b(x)$为输出函数;$\otimes$表示卷积运算;$f(x)$为备选输出,类似于原 MADL – MIDAS 模型输出;$g(x)$为卷积核函数。且 $b(x)$具有可交换性,即 $f(x)\otimes g(x)=g(x)\otimes f(x)$。

激活函数,是在线性模型的基础上引入非线性因素来增强模型的表达力。若卷积层不引入激活函数则每层的输出均为上层输入的线性函数,不论神经网络有多少层其输出均能表示为最初输入的线性组合,与仅含单个隐含层的效果相当,此时类似于多层感知机,无法探究变量间的非线性关系。因此,AMCNN 模型的卷积层需要引入非线性的激活函数 $\sigma(x)$,使得每层的输出不再是输入的线性组合,而可以逼近任意函数。即 $o(x)=b(x)\times\sigma(b(x))$。常见的激活函数多是分段线性(如 RELU、PRELU、LeakyRelu 等)和具有指数形状的非线

性函数(如 Sigmoid、Tanh 等)。本文选用 Sigmoid 函数作为激活函数。即 $\sigma(b(x)) = \frac{1}{1+e^{-b(x)}}$。

本文构建 AMCNN 模型的理论框架如下:设解释变量多元时间序列$(X(t_n), t_n - t_{n-1}) =: (x_n)_n \subset R^K$,被解释变量为 y_n。第 k 个解释变量的滞后阶数为 s_k,被解释变量的滞后阶数为 s_0,$x_n^{-s_k} = (x_{n-s})_{s=1}^{s_k}$,$y_n^{-s_0} = (x_{n-s})_{s=1}^{s_0}$。

$$\hat{y}_n = \sum_{k=0}^{K} \sum_{s=1}^{s_k} [F(x_n^{-s_k}) \otimes \sigma(H(x_n^{-s_k}))]_s \quad (11)$$

其中,$F(x_n^{-s_k}) = W \otimes [M(x_{n-s}) + x_{n-s}]_{s=1}^{s_k}$,$W \in R^{1 \times s_k}$,$M: R^{K+1} \to R$ 为多层感知器,则式(11)等价于:$y_n = \sum_{k=0}^{K} \sum_{s=1}^{s_k} [W_{k,s} \otimes [M(x_{n-s}) + x_{n-s}]_{s=1}^{s_k} \otimes \sigma(H_{k,s}(x_n^{-s_k}))]$。$M$、$H$ 为卷积神经网络的输出。为保证输入的时间序列长度与经全连接层输出的结果一致,AMCNN 中未使用池化层。为学习更多的非线性特征,选取 Softmax 作为代价函数。为保证与传统时间序列模型预测性能的可比性,选取均方根误差(RMSE)作为损失函数,模型训练时使用交叉验证法。AMCNN 模型的神经网络框架如图 3 所示。

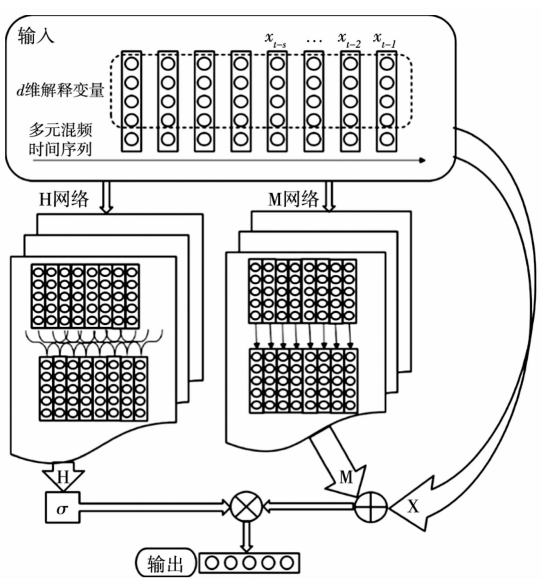

图 3 AMCNN 模型框架

为反映本文提出模型的 AMCNN 对 CPI 的预测性能,同时构建了式(5)同频时间序列模型,式(8)混频时间序列模型作为基准模型对 CPI 进行预测,比较不同模型的预测精度。上述模型中最佳滞后期数基于 BIC 准则来确定。为评估不同方法的预测性能,本文选用平方根误差 RMSE 评估预测精度。其表达式为:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{T}} \quad (12)$$

其中,y_t 和 $\hat{y}_t$ 分别为样本在时间 t 的真实值和预测值,T 为样本数量。RMSE 的值越小,表明预测效果越好。

四、CPI 预测实证分析

1. 数据来源与数据预处理

(1)变量选择与数据获取。本文涉及多种数据源,主要包括:国家统计局公布 CPI 数据、线上 CPI、百度指数、搜索引擎市场占有率、中国网民规模数据。数据时间范围为 2016 年 1 月 1 日至 2019 年 12 月 31 日。

第一,国家统计局公布的 CPI 数据。CPI 主要基于线下实体商品和服务及少量线上商品和服务价格产生,CPI 每 5 年进行一次基期轮换、目录调整和权数更新,2016 年 1 月开始一轮新的价格指数统计。由于本研究采用的线上 CPI 指数从 2016 年 1 月开始发布,且高频数据提供的是日度环比数据。因此,基于可比性考虑,本文选取 2016 年 1 月至 2019 年 12 月国家统计局公布的全国整体 CPI 月度环比数据作为样本数据,变量名为 CPI^{LM}③。

第二,线上 CPI。电子商务的发展,线上交易比重日益增加,部分机构参考 CPI 编制理论、利用大数据技术编制了线上消费者物价指数。目前影响力较大有三个:美国麻省理工的“十亿价格项目”(Billion Price Project,BPP)、中国阿里巴巴研究院的网购指数(iSPI,aSPI,aSPI - core)、清华大学的网上消费价格指数(Internet - based Consumer Price Index,ICPI)。其中,BPP 发布的中国价格指数仅覆盖生鲜食品和超市食品两类,无法全面覆盖中国 CPI 商品篮子,且相关数据并没有实时在网站公布;阿里巴巴线上价格指数(aSPI,aSPI - core)与官方 CPI 比子类别不同,数据来源单一,仅包含阿里巴巴平台数据,覆盖范围有限,且相关数据于 2017 年 1 月

后在其官网停止更新;2016 年 1 月,清华大学社会科学学院经济学研究所主导构建了 ICPI④,项目组定时、定网、定点采集 100 多个大型 B2C 在线购物平台中 ICPI 篮子商品和服务目录中的 8 个大类、45 个中类、262 个子类的线上价格信息,基于拉氏价格指数计算各类价格指数,并在官网实时公布主要结果。ICPI 覆盖范围包含传统 CPI 商品篮子,数据来源丰富,从 2016 年 1 月为基期开始以日、周、旬、月为周期均有数据发布。综上,ICPI 不论从指数编制的科学性、数据发布的实时性、数据收集的可行性上均较 BPP、aSPI 和 aSPI – core 更加符合中国实际。因此,本文线上 CPI 选取清华大学 iCPI 项目组发布的 2016 年 1 月 1 日至 2019 年 12 月 31 日的整体 ICPI 日度、月度环比数据为样本,变量名依次为 $ICPI^{HD}$、$ICPI^{LM}$。

第三,其他辅助数据。网络搜索数据采集时需借助搜索引擎市场占有率确定目标搜索引擎。搜索引擎市场占有率(Market Share of Search Engine, MSSE)主要统计指定区域指定搜索引擎的市场份额,本文获取 2016 年 1 月 1 日至 2019 年 12 月 31 日中国各搜索引擎的日度、月度市场份额为样本数据,变量名 $MSSE^{HD}$、$MSSE^{LM}$。另外,ISI 数据预处理过程中,为消除我国网民规模(Chinese Internet Users Number, CIUN)及百度搜索市场占有率(BMSSE)对百度指数搜索量的影响。本文采集了 CIUN、BMSSE,2016 年 1 月至 2019 年 12 月的月度数据,变量名分别为 $CIUN^{LM}$、$BMSSE^{LM}$。

第四,百度指数。百度指数(Baidu Index, BI)是海量网民基于百度搜索引擎的搜索行为产生的数据为基础的网络搜索数据分析平台。自 2006 年 7 月百度指数上线,便成为众多企业决策的重要依据、大量学术研究的补充数据。依据前文提出的关键词选取原则,本文百度指数的初始关键词如表 1 所示。主要包括宏观层面和微观层面的共 20 个初始关键词。本文采用爬虫技术抓取指定关键词在 2016 年 1 月 1 日至 2019 年 12 月 31 日的日度数据。各关键词的日度搜索量为移动端与电脑端搜索量之和。

(2)百度搜索指数合成。首先,将数据进行预处理后转化为日度、月度环比搜索指数,变量名依次为 BI^{HD}、BI^{LM}。然后,计算 BI^{LM} 与 CPI^{LM} 时差相关系数,结果如表 2 所示。根据表 2 数据选取领先于 CPI^{LM} 的关键词 CPI、通货膨胀、物价、GDP、猪肉价格、蔬菜价格、服装价格、房价、房租上涨、汽油价格作为最终关键词,包括月度数据和日度数据,分别命名为 $BI_1^{LM/HD},\cdots,BI_{10}^{LM/HD}$。最后,基于最终选取的 CPI 相关的关键词对应的百度指数 $BI_1^{LM/HD},\cdots,BI_{10}^{LM/HD}$,利用主成分分析选取特征值大于 1 的主成分,作为本文所使用的 CPI 网络搜索指数分别命名为 ISI_1^{HD},ISI_2^{HD}、ISI_3^{HD}、ISI_1^{LM}、ISI_2^{LM}、ISI_3^{LM}、ISI_4^{LM}。后文将利用 AMCNN 模型结合线上 CPI 和网络搜索指数的日度数据对 CPI 的月度数据进行预测。

2. AMCNN 模型的 CPI 预测结果分析

CPI 预测选取 2016 ~ 2018 年的样本数据作为训练集,2019 年的样本数据作为测试集。实验过程中借鉴 Bergstra 和 Bengio(2012)的研究成果,采用随机搜索的超参数搜索策略。基于随机搜索策略得到 AMCNN 模型的最优超参数取值组合如表 3 所示。

表 1　　网络搜索初始关键词

宏观层面	贷款利率、再贴现率、国际原油价格、外汇储备、CPI、通货膨胀、通货紧缩、物价、价格、GDP
微观层面	猪肉价格、蔬菜价格、白酒价格、服装价格、房价、房租上涨、空调价格、汽油价格、培训费、养老院价格

表 2　　最终关键词与 CPI^{LM} 的时差相关系数

关键词	CPI	通货膨胀	物价	GDP	猪肉价格	蔬菜价格	服装价格	房价	房租上涨	汽油上涨
变量名	BI_1^{LM}	BI_2^{LM}	BI_3^{LM}	BI_4^{LM}	BI_5^{LM}	BI_6^{LM}	BI_7^{LM}	BI_8^{LM}	BI_9^{LM}	BI_{10}^{LM}
领先阶数	3	2	2	4	2	3	3	5	1	5
相关系数	0.6988	0.3259	0.5973	0.4682	0.5915	0.4667	0.3383	0.4498	0.3910	0.2897

表 3　　本文 AMCNN 模型最优超参数取值

超参数	CPI 预测最优值	超参数	CPI 预测最优值
H 网络卷积层层数	3	M 网络卷积层层数	2
H 网络卷积核大小	5×1	M 网络卷积核大小	1×1
H 网络卷积核数量	8	M 网络卷积核数量	5
初始学习率	0.001	损失值波动容忍阈值	0.00035

CPI 预测过程中 CPI^{LM}、$ICPI^{HD}$、ISI_1^{HD}、ISI_2^{HD}、ISI_3^{HD}各变量滞后阶数依次为 q、s_1、s_2、s_3、s_4。被解释变量 CPI^{LM}自回归阶数 $q=0$、1、2、3,高频解释变量 $ICPI^{HD}$滞后阶数 $s_1=15$、30、45、60,百度搜索价格指数的滞后阶数 $s_2=s_3=s_4=15$、30、45、60。由于高频变量在预测中具有非完整周期适用性,CPI^{LM}提前预测步长 h 依次取 0.5、1、2,分别是基于当月前 15 天数据、上月底数据、前两月月底数据预测 CPI。训练集即 2019 年 CPI,预测结果对应 RMSE 如表 4 所示。

表 4　　AMCNN 模型样本外预测误差统计

s_2	$h=0.5$				$h=1$				$h=2$			
	s_1				s_1				s_1			
	15	30	45	60	15	30	45	60	15	30	45	60
$q=0$												
15	0.2157	0.2176	0.2013	0.2129	0.1952	0.2052	0.2012	0.2055	0.2187	0.1984	0.2090	0.2105
30	0.1990	0.1973	0.2054	0.2296	0.1962	0.1873	0.1892	0.1953	0.1928	0.2038	0.1946	0.2212
45	0.1986	0.1941	0.1938	0.2084	0.1947	0.1902	0.1886	0.1962	0.2025	0.1977	0.1989	0.2345
60	0.2167	0.2038	0.2199	0.2167	0.2086	0.1986	0.2008	0.2076	0.2094	0.2092	0.1995	0.2188
$q=1$												
15	0.1981	0.1867	0.1853	0.1861	0.1897	0.1853	0.1810	0.1855	0.1873	0.1986	0.1952	0.1980
30	0.1955	0.1803	0.1882	0.1873	0.1831	0.1724	0.1732	0.1894	0.1822	0.1894	0.1926	0.1943
45	0.1954	0.1879	0.1802	0.1928	0.1885	0.1730	0.1758	0.1832	0.1904	0.1769	0.1777	0.1986
60	0.1963	0.1878	0.1912	0.1914	0.1932	0.1853	0.1827	0.1872	0.1927	0.1872	0.1907	0.1995
$q=2$												
15	0.2062	0.1997	0.1963	0.2066	0.1927	0.1864	0.1911	0.1951	0.2113	0.2067	0.2087	0.2129
30	0.2038	0.1911	0.1937	0.2007	0.1912	0.1842	0.1868	0.1994	0.2027	0.1920	0.1895	0.1912
45	0.2078	0.1925	0.1924	0.2049	0.1901	0.1836	0.1872	0.1925	0.2035	0.1846	0.1974	0.2041
60	0.2108	0.2062	0.1983	0.2101	0.1929	0.1921	0.1953	0.1974	0.2089	0.2078	0.2048	0.2152
$q=3$												
15	0.2213	0.2235	0.2304	0.2420	0.2402	0.2591	0.2680	0.2587	0.2286	0.2379	0.3341	0.2587
30	0.2241	0.2104	0.2084	0.2149	0.2273	0.2006	0.2395	0.2390	0.2330	0.2222	0.2216	0.2371
45	0.2331	0.2073	0.2012	0.2394	0.2322	0.2583	0.2030	0.2201	0.2312	0.2268	0.2156	0.2547
60	0.2412	0.2272	0.1988	0.2399	0.2579	0.2451	0.2362	0.2068	0.2165	0.2135	0.2489	0.2444

由表4看出,从提前预测步长来看,在 q、s_1、s_2 条件相同的情况下,基于上月底数据($h=1$)预测本月CPI的RMSE整体较小;从CPI自回归阶数来看,在 h、s_1、s_2 条件相同的情况下,$q=1$ 时CPI预测误差RMSE整体较小;从高频解释变量 $ICPI^{HD}$、ISI_1^{HD}、ISI_2^{HD}、ISI_3^{HD} 的滞后阶数来看,在 h、q 条件相同时,s_1、s_2 取值组合为(30,30)、(30,45)、(45,30)、(45,45)时,CPI预测误差RMSE整体较小,考虑到待估参数随滞后阶数的增加而增多,为降低过拟合风险,选取 $s_1=30$、$s_2=30$。综上所述,按照RMSE最小的原则,本文选取的最优参数组合为 $h=1$、$q=1$、$s_1=30$、$s_2=30$。

进一步探讨最优参数组合下AMCNN模型对CPI的预测性能,CPI预测值与真实值如图4所示。由图4看出,CPI样本内拟合程度优于样本外预测。值得关注的是,AMCNN模型对于波动幅度较大时的CPI拟合性能更好。其原因在于,线上CPI和具有领先水平的网络搜索指数对于较大波动的反应更加明显。AMCNN结合ICPI、ISI能够更好地捕捉CPI“拐点”。

3. AMCNN模型的CPI预测性能对比分析

为验证AMCNN模型的有效性,本文将进行不同模型的对比分析。将AMCNN模型的CPI预测结果分别与同频时间序列ADL模型、混频时间序列模型MADL_MIDAS进行样本内、外预测结果比较分析(见表5)。其中,ADL模型各变量的滞后期基于BIC准则获得。MIDAS(A)、MIDAS(EA)、MIDAS(B)、MIDAS(S)分别表示MADL_MIDAS模型采用阿尔蒙多项式函数(Almon)、指数阿尔蒙函数(Exponential Almon)、贝塔密度函数(Beta)、分段函数(Step Function)4种不同权重函数下的预测模型,为满足可比性,模型中的参数与AMCNN模型最优参数取值一致。

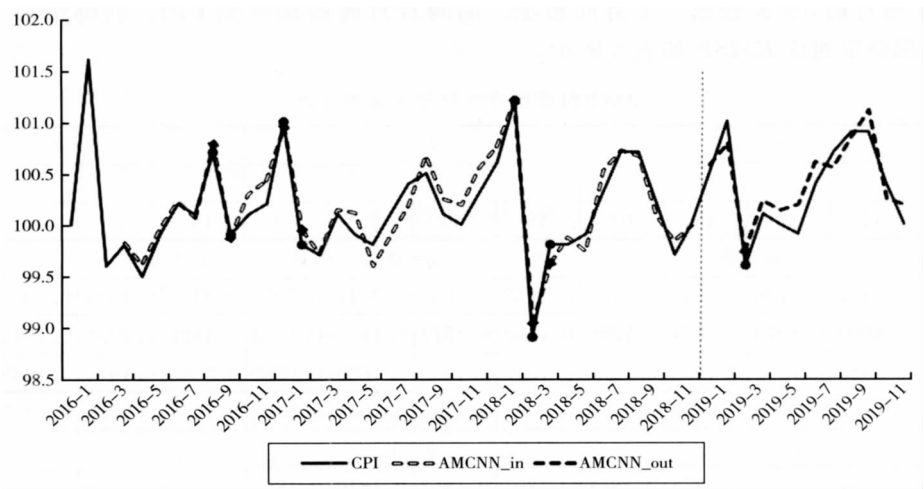

图4 CPI真实值及样本内、样本外预测值

表5 不同模型不同变量下CPI预测结果对比

模型	*CPI*、*ICPI*		*CPI*、*ISI*		*CPI*、*ICPI*、*ISI*	
	样本内	样本外	样本内	样本外	样本内	样本外
ADL	0.4522	0.3652	0.4861	0.7449	0.4377	0.6693
MIDAS(A)	0.3323	0.3658	0.3777	0.4266	0.3166	0.3424
MIDAS(EA)	0.3081	0.3247	0.3223	0.3469	0.3005	0.3216
MIDAS(B)	0.2832	0.2623	0.3017	0.3218	0.2657	0.3120
MIDAS(S)	0.3420	0.3802	0.3824	0.4281	0.3196	0.3488
AMCNN	0.1949	0.2024	0.2143	0.2284	0.1411	0.1724

由表5可知,从研究视角来看,ICPI和ISI对CPI均有一定程度的预测能力,同时引入CPI、ICPI、ISI三者预测性能最好。具体来看,ADL模型下CPI与ICPI组合的样本内、外预测误差为0.4522和0.3652,分别小于ADL模型下CPI与ISI组合下样本内外的预测误差0.4861和0.7449。此外,模型MIDAS(A)、MIDAS(EA)、MIDAS(B)、MIDAS(S)以及本文提出的AMCNN模型,在CPI与ICPI组合样本内、外的预测效果均比CPI与ISI组合更好,即线上CPI的预测能力优于百度搜索指数,表明客观的线上消费价格指数比主观的网络搜索指数(网络搜索关键词选取具有一定的主观性)具有更好的预测能力。另外,对比表5不同模型下样本内、外预测误差,结果发现不论是同频、混频还是异步混频模型,同时引入CPI、ICPI以及ISI的预测效果均优于单独引入ICPI或ISI。从研究方法来看,混频模型预测性能均优于同频模型,在MADL_MIDAS模型中,MIDAS(B)所在行的预测误差整体小于MIDAS(A)、MIDAS(EA)及MIDAS(S),即权重设为贝塔密度函数时预测误差更小。同等条件下,相较于MADL_MIDAS模型,AMCNN的预测效果更好,且同时引入CPI、ICPI、ISI三者,AMCNN模型样本内、外预测误差依次为0.1411和0.1724是所有结果中最小的,表明AMCNN模型与基准模型相比具有更好的预测能力。

4. CPI子类别预测结果分析

为验证AMCNN模型预测性能的稳健性,本文分别利用ADL、MIDAS(B)、AMCNN模型基于CPI、线上CPI、网络搜索指数对CPI各子类别进行预测。比较不同模型对不同权重和不同波动幅度CPI子类别的预测精度是否稳定。并利用方差分析判断不同模型对CPI子类别的预测性能是否有显著差异。

由表6可知,同时引入CPI、ICPI、ISI三者,AMCNN模型对CPI除生活用品及服务类、交通和通信类外,其他子类别的样本内、外预测误差均显著小于ADL和MIDAS(B)模型对应的样本内、外预测误差,尤其是食品烟酒类和居住类的样本内预测误差分别是0.1869、0.1712远远小于ADL模型的0.8941和0.8933、MIDAS(B)模型的0.5089、0.4930,样本外的预测误差仅为ADL模型的23%左右,MIDAS(B)模型的42%左右。针对CPI 8个子类别的预测,本文提出的AMCNN模型依然具有最佳的预测精度,表明在不同权重和不同波动幅度情况下,AMCNN模型表现稳健。

箱线图(见图5)清晰地展现了ADL、MIDAS(B)、AMCNN模型在预测CPI子类别样本内、外RMSE的差异。

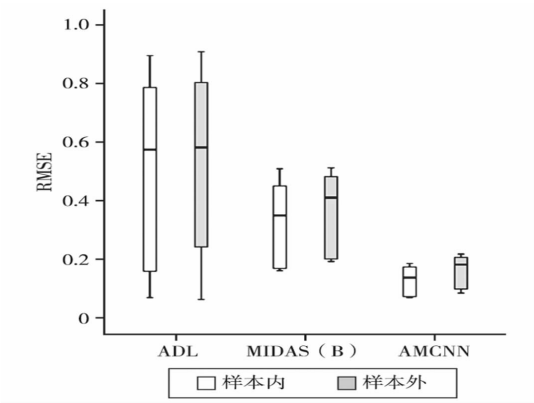

图5 不同模型CPI子类别样本内、外预测误差对比

表6 CPI子类别预测结果对比

CPI子类别	ADL		MIDAS(B)		AMCNN	
	样本内	样本外	样本内	样本外	样本内	样本外
食品烟酒类	0.8941	0.9004	0.5089	0.4761	0.1869	0.2126
衣着类	0.2327	0.3816	0.1688	0.1997	0.0714	0.1032
居住类	0.8933	0.9071	0.4930	0.5122	0.1712	0.2031
生活用品及服务类	0.0877	0.1052	0.1718	0.2039	0.0751	0.0959
交通和通信类	0.0705	0.0646	0.1621	0.1936	0.0729	0.0863
教育文化和娱乐类	0.5248	0.4726	0.3395	0.3933	0.1259	0.1704
医疗保健类	0.6238	0.7049	0.3613	0.4283	0.1518	0.1961
其他用品和服务类	0.6779	0.6911	0.4083	0.4892	0.1791	0.2189

表7中,方差分析多重比较结果显示样本内CPI各子类别在95%的置信水平下,MIDAS(B)模型上的平均预测误差显著低于ADL模型约0.2076,AMCNN模型上的平均预测误差显著低于ADL模型约0.4051,显著低于MIDAS(B)模型约0.1974。综上,进一步显示AMCNN模型的预测性能显著优于其他模型。

5. CPI“拐点”预测

在对宏观经济趋势进行分析的过程中,学者或决策者在关注预测精度的同时,更加注重宏观经济发展趋势和结构突变点。宏观经济分析中通常将其称之为“拐点”。由于国际形势日趋复杂,国际贸易受到诸多因素影响,加之国内原材料价格波动,以及货币政策等原因导致我国CPI“拐点”的出现较为普遍。为了更好地对“拐点”进行预测,首先对“拐点”的定义进行界定:若CPI第$t-1$期到第t期的变化趋势与第t期到第$t+1$期的变化趋势不同,即上月环比CPI数据中$(CPI_t^{LM}-100)\times(CPI_{t+1}^{LM}-100)<0$,则第$t$期称为CPI“拐点”。据此定义,2016年1月至2019年12月CPI共出现14个“拐点”,占总样本的29.17%。由图3可知AMCNN模型在CPI波动较大时表现出良好的拟合能力。下面我们将ADL、MIDAS(B)、AMCNN模型对“拐点”的预测进行对比(见图6),三个模型的共同预测区间为2016年6月至2019年12月,其中包含13个真实CPI“拐点”。结果显示,ADL模型共预测了16个“拐点”,准确捕捉了6个;MIDAS(B)模型共预测了15个“拐点”,准确捕捉了8个;AMCNN模型共预测了13个“拐点”,准确捕捉了10个。

表7　　不同模型样本内外预测性能差异性分析

模型比较		样本内			样本外		
		均值差	标准误	p值	均值差	标准误	p值
ADL	MIDAS(B)	0.2076*	0.0942	0.0389	.2227*	0.0843	0.0153
	AMCNN	0.4051*	0.0942	0.0003	.4239*	0.0843	0.0001
MIDAS(B)	ADL	-0.2076*	0.0942	0.0389	-.2227*	0.0843	0.0153
	AMCNN	0.1974*	0.0942	0.0485	.2012*	0.0843	0.0265
AMCNN	ADL	-0.4051*	0.0942	0.0003	-.4239*	0.0843	0.0001
	MIDAS(B)	-0.1974*	0.0942	0.0485	-.2012*	0.0843	0.0265

注:* 表示在5%的水平下显著。

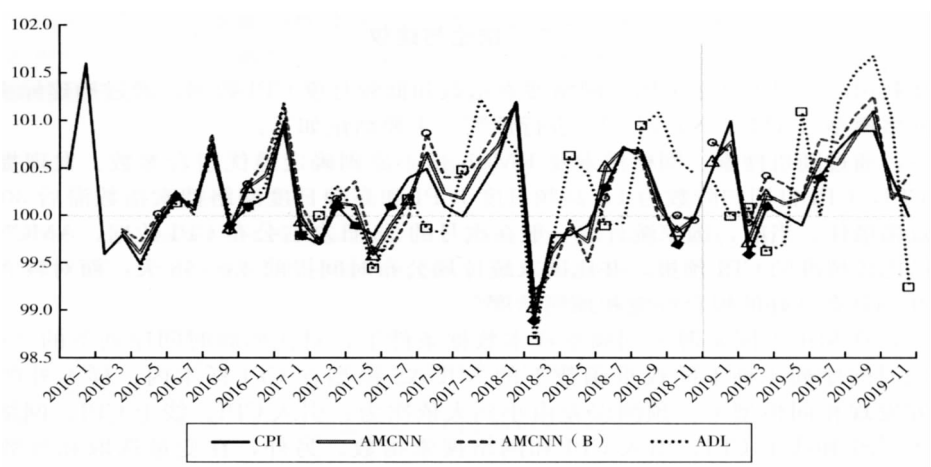

图6　真实CPI与不同模型样本内、外“拐点”捕捉能力对比

注:◆表示真实CPI的“拐点”,△、○、□分别表示AMCNN、MIDAS(B)、ADL模型在表5预测误差下CPI预测的“拐点”。

为进一步衡量 ADL、MIDAS(B)、AMCNN 模型对 CPI“拐点”的捕捉能力。分别从查全率(*recall*)、查准率(*precision*)及查全查准率(F_1)三方面评估其“拐点”预测水平,其表达式分别为:

$$recall = \frac{TP}{TP + FP} \tag{13}$$

$$precision = \frac{TP}{TP + FN} \tag{14}$$

$$F_1 = 2 \times \frac{1}{\frac{1}{recall} + \frac{1}{precision}} \tag{15}$$

其中,TP 表示 CPI 实际为“拐点”,预测也为“拐点”的数目;FP 表示 CPI 实际非“拐点”,预测为“拐点”的数目;FN 表示 CPI 实际为“拐点”,预测非“拐点”的数目。F_1 为查全率和查准率二者的调和平均,综合评估模型 CPI“拐点”预测的全面性和准确性。

表 8 显示,不论是查全率、查准率还是二者综合水平 AMCNN 模型均高于基准模型 ADL、MIDAS(B)。可见,利用高频线上消费价格指数、高频网络搜索指数结合 AMCNN 模型能更全面、准确地捕捉 CPI“拐点”。

表 8　不同模型 CPI“拐点”预测水平

	ADL	MIDAS(B)	AMCNN
查全率	0.4615	0.6154	0.7692
查准率	0.3750	0.5714	0.7692
查全查准率	0.4138	0.5926	0.7692

五、结论与建议

本文利用高频日度线上 CPI、网络搜索指数和低频月度 CPI 数据,通过构建异步混频卷积神经网络模型(AMCNN)对 CPI 进行预测。主要结论如下:

第一,通过随机搜索、网格搜索及 RMSE 最小原则确定最优组合参数。发现提前 1 个月预测 CPI,CPI 自回归阶数为 1,高频日度 ICPI 和高频日度网络搜索指数滞后 30 ~ 45 天时预测效果最佳。当前,国家统计局一般在次月的 13 日左右公布 CPI 数据。AMCNN 模型能够提供比较精准的 CPI 预报,并比国家统计局公布时间提前 43 ~ 58 天,面对较大幅度波动的 CPI 仍具有良好的拟合性能和预测精度。

第二,在利用不同来源不同频率样本数据条件下,对比同频时间序列下的 ADL 模型以及混频时间序列 4 种不同权重函数下的 MIDAS 模型在 CPI 样本内、样本外的预测精度。研究发现相同模型下,预测误差由小到大依次为:引入 CPI、线上 CPI、网络搜索指数,引入 CPI 和线上 CPI,引入 CPI 和网络搜索指数。另外,在变量选取和参数设置相同的条件下,混频模型优于同频模型(ADL),且本文提出的 AMCNN 模型优于传统混频模型(MIDAS),传统混频模型中权重函数为贝塔密度函数(MIDAS(B))时预测精度较高。

第三,为了进一步验证 AMCNN 模型对不同波动情况下的预测精度和稳健性,对 CPI 的 8 个子类别分别进行了预测,通过对比发现,AMCNN 模型的预测误差依然低于 ADL 模型和 MIDAS(B)模型,并通过方差分析证实了其显著性。AMCNN 模型除了在预测精度上具有优良表现外,在“拐点”的捕捉上也体现出优势,通过对比 ADL 和 MIDAS(B)模型发现,AMCNN 模型对“拐点”的识别能力在查全率、查准率和查全查准综合水平上均高于基准模型。

综上所述,从实际 CPI 编制、统计、计算的角度考虑,全面掌握全量的线上线下产品或服务价格,比传统抽样统计的样本代表性好、时间颗粒度细,能更加真实地反映 CPI 的变化特征,获得更具参考价值的信息,同时也能准确、及时反映当下商品消费价格的变化以及国民消费倾向,有助于决策层清晰地掌握物价波动情况,对政府有针对性地制定宏观经济政策奠定坚实基础。当然,目前线上 CPI 主要由企业制定一揽子商品,参考官方 CPI 编制方法确定权重,权威性受到一定质疑,随着大数据技术的不断发展、统计制度的进一步完善,建议政府层

面制定线上 CPI 核算体系,扩展线上商品核算种类,同时编制涵盖线上、线下商品的统一 CPI 核算方法,为提高 CPI 数据质量做好顶层设计。

本文提出一个具有较强可行性,基于多源异步混频 CPI 数据的预测框架。这个框架同样也可用于预测其他类型的宏观经济或金融指标,如 GDP 预测、股票收益率预测等。最终提出满足多元非线性异步混频双自相关的 CPI 预测模型。利用多源异步混频数据结合神经网络构建预测模型,不受变量频率倍差波动以及变量间不确定性(线性、非线性)关系的影响,具有很强的适应性和扩展性,应用价值较高。本文的研究也存在一定的局限性:第一,在利用网络搜索大数据时,搜索关键词的确定非常重要,具有一定的主观性,选取不同的关键词可能产生不同的预测结果;第二,由于涉及多来源的异步混频数据对数据采集能力要求较高,数据预处理能力要求较高。在数据可获取且数据质量较高的情况下,后续针对异步混频数据预测方法的研究可从以下三方面展开:第一,充分利用高频数据信息进一步优化模型对突发事件(如新型冠状病毒肺炎疫情)的预测能力;第二,采用贝叶斯方法推断小样本下的参数,减少神经网络模型对样本数量的限制;第三,探索神经网络的"黑箱"部分,针对经济统计相关研究在提高预测精度的同时保留模型参数的经济学解释。

注释:

①http://gs.statcounter.com/.

②http://index.baidu.com/v2/index.html#/.

③实证部分变量名解析:CPI 为官方公布的消费者价格指数(Consumer Price Index),ICPI 为线上消费者物价指数(Internet Consumer Price Index),ISI 为网络搜索指数(Internet Search Index)。H 表示高频数据(High Frequency Data),L 表示低频数据(Low Frequency Data),M 表示月度(Month)统计数据,D 表示日度(Day)统计数据。如 CPI^{LM} 表示消费者物价指数的低频月度数据,$ICPI^{HD}$ 表示线上消费者物价指数的高频日度数据。

④http://www.bdecon.com/homeIndex.

参考文献:

[1]Ghysels E., Santa-Clara P., Valkanov R., 2004, The MIDAS Touch: Mixed Data Sampling Regression Models[R], Working Paper, Anderson School of Management, UCLA.

[2]Ghysels E., Sinko A., Valkanov R., 2007, MIDAS Regressions: Further Results and New Directions[J], Econometric Reviews, 26(1), 53~90.

[3]Bergstra J., Bengio Y., 2012, Random Search for Hyper-Parameter Optimization[J], Journal of Machine Learning Research, 13(1), 281~305.

[4]张崇、吕本富、彭赓:《网络搜索数据与 CPI 的相关性研究》[J],《管理科学学报》2012 年第 7 期.

[5]孙毅、戴维、董纪昌、吕本富:《基于主成分分析的网络搜索数据合成方法研究》[J],《数学的实践与认识》2014 年第 21 期.

[6]孙毅、吕本富、陈航、薛添:《大数据视角的通胀预期测度与应用研究》[J],《管理世界》2014 年第 4 期.

[7]董倩:《基于网络搜索数据的雾霾经济与 CPI 相关性研究》[J],《调研世界》2016 年第 12 期.

[8]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 年第 8 期.

[9]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 年第 1 期.

[10]刘宽斌、张涛:《利用网络搜索大数据实现对 CPI 的短期预报及拐点预测——基于混频抽样数据模型的实证研究》[J],《当代财经》2018 年第 11 期.

[11]杜两省、刘发跃:《线上与线下,联动还是竞争?——基于 ISPI 和 CPI 的线上线下价格差异收敛性分析》[J],《投资研究》2014 年第 7 期.

[12]刘发跃、马丁丑:《网上与网下两类价格指数差异的收敛性分析》[J],《统计与决策》2015 年第 20 期.

[13]米子川、姜天英:《大数据指数是否可以替代统计

调查指数》[J],《统计研究》2016 年第 11 期.

[14]周薇薇、田涛:《大数据背景下电商发展对 CPI 的影响——基于线上线下价格波动同步性分析》[J],《商业研究》2016 年第 4 期.

[15]田涛:《电商发展对 CPI 的影响研究——基于大数据背景下线上线下价格波动的同步性分析》[J],《上海经济研究》2016 年第 3 期.

[16]韩胜娟、张敏:《大数据时代官方价格指数与非官方价格指数的融合——基于 aSPI 与 CPI、RPI 比较的视角》[J],《价格理论与实践》2017 年第 4 期.

[17]方匡南、曾武雄:《阿里网购价格指数与官方 CPI 的关系》[J],《统计与信息论坛》2018 年第 2 期.

[18]唐礼智、刘玉:《线上线下价格指数的互动:替代还是整合?》[J],《南京社会科学》2018 年第 2 期.

[19]刘涛雄、汤珂、姜婷凤、仉力:《一种基于在线大数据的高频 CPI 指数的设计及应用》[J],《数量经济技术经济研究》2019 年第 9 期.

[20]刘汉、刘金全:《中国宏观经济总量的实时预报与短期预测——基于混频数据预测模型的实证研究》[J],《经济研究》2011 年第 3 期.

Research on CPI Prediction Based on Multi – source Asynchronous Mixed Sampling Data

Zhang Hu　　Shen Hanlei　　Xia Lun

Abstract: Research Objectives: CPI is predicted based on internet consumer price index (ICPI) and internet search index (ISI). Research Methods: In the framework of convolution neural networks (CNN), MADL_MIDAS model is integrated, and the asynchronous mixed frequency convolution neural network (AMCNN) model is established. The data from January 2016 to December 2019 are used to verify the effectiveness of the method. Research Findings: High frequency daily data ICPI and ISI are the leading indicators of CPI. Adding all variables and retaining the original data characteristics will help to improve the CPI prediction accuracy and the ability of capture CPI's "change point". Research Innovations: This paper reveals the prediction ability of low frequency monthly CPI by ICPI and ISI of high frequency daily data, and proposes an asynchronous mixed sampling data processing method combining neural network and traditional econometric model. Research Value: When forecasting CPI fluctuation level and "change point", it can help to use ICPI and ISI, combined with AMCNN model to improve prediction performance. AMCNN model can be used to deal with asynchronous mixed sampling data and explore the complex uncertainty (linear or nonlinear) relationship between multi – variables. It has strong adaptability, scalability and high applicability. It can be applied to other economic and financial fields.

Key words: AMCNN; CPI prediction; internet search data; internet CPI; asynchronous mixed sampling data