doi:10.3969/j.issn.1672-626x.2021.05.003

利用网络价格信息改进CPI编制与国际经验借鉴研究

李 倩[1],周 迪[2],李 丽[1]
(1. 湖北经济学院 信息管理与统计学院,武汉 430205;2. 广东外语外贸大学 数学与统计学院,广州 510006)

摘 要:大数据时代的到来给政府统计工作带来前所未有的历史机遇和重要挑战,作为其中之一的消费价格指数(CPI)编制可谓首当其冲,实现CPI与时代接轨已成为当务之急。本文主要研究利用网络价格信息改进CPI编制问题。在借鉴国际经验的基础上,首先探讨如何对网络价格数据进行收集与整理,包括零售商网站的选取、网络价格数据收集方法、网络价格数据的收集过程及数据整理;其次分析基于网络数据的价格指数编制面临的挑战;接着介绍单独基于网络抓取数据和网络数据纳入传统CPI统计范围的价格指数编制方法;然后总结基于网络数据的价格指数相关实证结果;最后是研究展望。本文研究为我国国家统计局推进网络价格在CPI统计中的应用提供一些参考。

关键词:爬虫技术;网络价格数据;CPI;国际经验
中图分类号:F222.3;C813 **文献标志码**:A **文章编号**:1672-626X(2021)05-0044-12

一、引言

消费价格指数(CPI)是衡量经济发展的重要指标,自1925年以来,CPI编制的国家标准不断更新。2004年国际组织编制的《消费物价指数手册:理论与实践》从理论上对CPI进行了全面阐释,成为各国统计机构编制CPI的重要指导手册。但从实际应用过程看,还需针对不同国家的具体情况给出具有实务性的操作指导,为此国际组织于2009年联合颁布了《CPI编制实用指南》,该指南主要侧重实际问题的解决,是对《消费物价指数手册:理论与实践》的补充。为了更好地满足国民经济核算要求,中国国家统计局于2000年开始启用与国际接轨的CPI编制方法,但中国CPI的编制要求与国际规范相比仍存在较大差距。

大数据时代的到来给政府统计工作带来前所未有的历史机遇和重要挑战,作为其中之一的CPI编制可谓首当其冲,实现CPI与时代接轨已成为当务之急。有些国家已将收集的网络价格纳入官方CPI统计中,如2014年瑞典CPI中家用电子设备有17%的价格数据来自于网络,服装和鞋类为10%,图书和媒体为38%,交通服务费有很大比例来源于网络;美国CPI的9%是通过网络价格计算的;加拿大CPI的5%~10%是通过手工收集网络价格计算的;挪威CPI的18%是通过软件收集网络价格计算的;荷兰CPI中纳入了服装类网络价

收稿日期:2021-06-06
基金项目:国家社会科学基金青年项目(17CTJ013);国家自然科学基金项目(61903130)
作者简介:李倩(1986-),女,山东济宁人,湖北经济学院信息管理与统计学院讲师,经济学博士,研究方向为经济统计、人口统计、价格理论与实践。

格等[1~2]。此外，英国等国家网络价格数据正处于试验研究阶段，尚未正式纳入CPI编制过程中，单独编制基于网络抓取数据的CPI，试验范围仅限于部分商品。为推进我国网络价格在CPI统计中的应用，2013年11月国家统计局与阿里巴巴、百度等11家企业签署了大数据战略合作框架协议；2015年1月国家统计局沈阳调查总队积极参与沈阳地区电视机、洗衣机、电脑和手机商品网购价格的调查及环比指数的测算试点工作；2015年以来，北京调查总队积极探索在CPI统计中通过人工定期浏览电商网站等方式开展网络采价；2015年浙江省针对电视机、空调、热水器、电脑、手机等商品在苏宁易购、京东商城等网络销售平台中进行互联网人工采价试点工作。越来越多的网络公司或研究机构利用网络数据即时生产、发布类似的指数，如麻省理工学院计算的每日网上价格指数、阿里研究院推出的阿里巴巴全网网购价格指数（aSPI）和网购核心商品价格指数（aSPI-core）、清华大学项目团队编制并实时发布的基于互联网在线数据的居民消费价格指数（iCPI）①等。

我国统计学界较早关注的是如何利用扫描数据改进CPI编制[3~4]，利用网络价格改进CPI编制的研究还很少，只有少部分学者进行了相关方面的研究，例如基于CPI统计方法的研究[5~7]和基于CPI编制、公布及数据质量的相关研究[8~10]。在借鉴国际经验的基础上，本文的结构安排如下：首先是网络价格数据的收集与整理研究；其次是基于网络数据的价格指数编制面临的挑战分析；再次是基于网络数据的价格指数编制方法和相关实证结果梳理；最后是研究展望。本文的研究价值在于可以为我国国家统计局推进网络价格在CPI统计中的应用提供参考。

二、网络价格数据的收集与整理

零售商数量很多，既包括只在线上销售的纯在线零售商（如eBay、亚马逊等），又包括线上线下均销售的多渠道零售商（如沃尔玛、苏宁易购等）。虽然网上购物越来越受欢迎，但网上购物者并不一定代表典型的消费者，同样网络上的价格可能与实体店价格不同。在收集网络价格数据时怎么选取代表性零售商网站？怎么收集零售商网站上的网络价格数据？怎么整理收集的网络价格数据？这是本部分主要讨论的内容。

（一）网络价格数据的收集

1. 零售商网站的选取

通常从市场份额排名靠前的零售商网站上收集网络价格数据，这类零售商集中了绝大多数的零售交易，成为“代表性”的数据来源。Alberto Cavallo（2017）[11]对10个国家56家大型多渠道零售商的网站和实体店同时收集的价格进行大规模比较，选取的零售商都进入了各自国家市场份额排名前20位的零售商名单。余芳东（2018）[2]总结了利用网络抓取数据编制CPI的实践做法，其中荷兰统计局选择网上销售量大、线上和线下均有交易的服装零售商店网站作为抓取价格数据的目标网站，挪威统计局每日从在挪威注册且销售规模大的4家在线商店上自动抓取价格数据和相关信息，英国统计局每天从占市场销售比重较大的3个超市网站（特易购、森斯伯瑞、维特罗斯）抓取CPI采价目录中食品、非酒精饮料、酒精饮料三类35种食品价格数据。

2. 网络价格数据收集方法

大数据处理过程主要包括收集、预处理、存储及管理、分析及挖掘、展现和应用，目前大数据应用领域比较典型的有商业智能、公共服务、政府决策等领域。

目前主要有两种方式收集网络价格信息，一种是人工收集网络代表规格品价格，另一种是网络爬虫技术自动收集海量价格数据。人工收集方式中调查员通常从网站上反复复制粘贴各代表规格品的价格相关信息，并截取图片以保证收集信息的真实、可靠。这种收集方式较为繁琐，费时费力，容易出现人为差错。

网络爬虫技术是指从网上自动提取数据的技术，包括脚本编写方法和“点击”方法。脚本编写方法要求研究人员具有使用Python和PHP等语言编程的能力，网络爬虫程序根据预先定义的条件，系统地下载从起点到达的所有网络资源。“点击”方法（如Import.io）需要较少的编程技能，用户可以简单地用鼠标告诉“爬虫器”他们想从网页上收集的信息，爬虫程序遍历网络站点，并从与设置的参数类似的页面中提取信息，将数据结构化为行和列存储在云服务器上，以便下载和加载。然后将数据加载到合适的软件中进行分析、计算和存储，每天同一时间自动收集数据。与开源脚本语言相比，这些工具是“闭源”的，能够使用户得到更好的支持，更加依赖提供工具的公司，灵活性较差。网络爬虫技术能及时、低成本地收集大量数据，但不如人工收集严格，很难控制收集的准确性，面临标签挑战。从互联网上自动提取数据是为统计目的收集价格的新方法，为了利用这些数据有必要解决各种问题，首先是网站结构变化频繁问题，每个网站使用不同方式存储信息，当网站结构发生变化时需要对相应的网络爬虫重新编程；其次是从网站中频繁提取大量数据的合法性问题，这取决于抓取的数据类型、访问和复制的信息量、访问对页面所有者的系统和数据的负面影响程度；在某些情况下，网站管理员还可能在网站上设置屏蔽机制，以阻止使用网络爬虫。网络爬虫技术在解析页面复杂、网站改版频繁、网络阻塞等情况下存在一定的局限性。

在价格收集过程中不同国家使用不同的爬虫技术，如德国和意大利将web抓取软件（iMacros）与java编程结合起来，输入、选择、删除和存储价格数据；荷兰使用r软件建立自己的网页抓取框架；英国使用Python软件编写自己的网页抓取程序。

3. 网络价格数据的收集过程

针对不同的研究目的，研究人员收集不同的网络价格数据。

为了研究基于网络抓取数据的CPI，Radzikowski和Mietanka（2016）[12]主要从比价网站上收集了3000多个销售点的价格数据，有些数据是实时更新的，有些是定期更新的（至少每月更新一次），有些数据仅在某些商品价格发生变化时更新（如宣布新电价时）。从比价网站上收集的数据能确保在线CPI不受某一零售商及其定价策略的影响，考虑了动态变化的市场环境，对于不在比价网站上列出的代表品，从其专门网站（如石油价格行业门户网站）上收集。英国统计局从2014年6月到2015年6月每天上午5点从占市场销售比重较大的特易购、森斯伯瑞、维特罗斯3个超市网站上抓取食品、非酒精饮料、酒精饮料三类35种食品价格数据，根据超市网站上展示的商品数量，每天收集约6500笔价格数据，数据量远大于传统的价格数据收集量。

为了比较网站价格与实体店价格的相似性，Alberto Cavallo（2017）[11]从2014年12月至2016年3月在全球56家多渠道零售商共收集24000多个产品的38000个线上线下可匹配价格，数据覆盖范围主要集中在美国，有17家零售商和大约40%的观察结果，但在中国的数据只有两家零售商。

Hull等（2017）[13]总结了瑞典为了调查网上销售的水果和蔬菜价格能否提高短期通胀预测的准确性开展的一项小规模试点研究。该研究创建了一个自动在线数据收集流程，每天从瑞典零售商收集一些选定的水果和蔬菜的在线价格数据。所有数据收集任务都在Linux虚拟专用服务器（VPS）上执行，服务器每天按顺序执行三个脚本，第一个脚本访问4家大型杂货零售商的网站，从所有与水果和蔬菜相关的页面中提取代码，然后在代码中标识所有产品价格和名称，并保存在.csv文件中，原始代码也以.txt格式保存90天，以便纠正以后发现的错误，然后该脚本使用正则表达式过滤数据，创建只包含目标水果和蔬菜的第二个.csv文件。爬取完数据后，服务器执行第二个脚本，将过滤后的数据与过去的数据合并。最后，服务器执行第三个脚本，检查错误。Powell等（2018）[14]使用两个数据集探究更频繁的月度综合CPI预测目标实现情况，第一个数据集包括英国3家大型超市网站的33种商品类别的每日网络价格，历时约14个月；第二个数据集包含了相同产品类别的分类CPI值，以及对综合CPI有贡献的更多数据。

为了探索网络价格纳入CPI统计，荷兰统计局每日抓取3家服装零售商网站的服装类价格数据，从每个

网站上抓取的数据框架至少包括商店网址、商品类型、商品具体名称、简要规格说明和价格数据5项基本信息。意大利统计局通过网络爬取消费者物价调和指数（HICP）中“消费者电子产品”（商品）和“机票”（服务）信息来探讨网络价格爬虫技术，一是定期收集消费者电子产品信息，每个产品平均选择18家左右的商店收集网络价格；二是从16家低成本航空公司网站和3家机票销售网站（Opodo、Travelprice和Edreams）进行机票价格的数据搜集，网站上每月登记的基本票价超过960种，但只收集传统航空公司的机票信息，两名专家进行这项机票数据收集工作，每人每月工作约15小时，为期三天。Kjersti和Leiv（2016）[15]使用Import.io软件从专注于消费者电子产品和个人护理产品领域的四个主要电商网站爬取数据，在一年多的时间里，每天爬取大约60种不同消费品的4300份价格观察报告。

（二）网络价格数据的整理

由于网站的格式、描述和产品分类等形式多样，因此需要将抓取的网络价格原始数据进行整理，以便进行分析和指数测算。数据清洗和处理工作量较大，大约占整个项目时间的50%~80%。网络抓取数据不如人工采价严格，快速收集的大量数据准确性难以控制，特别是对商品无法准确分类，经常出现分类错误，还需要进行人工检查，结合项目描述中关键的数值信息有助于商品准确分类。研究团队根据网络爬虫技术每日自动抓取的数据集文件信息，进行数据检查，检验通过之后方可进入指数编制过程。

三、基于网络数据的价格指数编制挑战

新的数据源在质量和效率方面都有可能改进官方价格统计，将新的数据源集成到价格统计中并不简单，需要处理多方面的挑战。

一是使用爬虫技术成本效益分析。在开始探索爬虫技术之前需要进行充分的成本效益分析，必须投入相当多的资源，以便能够成功地使用它，即使软件本身可能不需要任何编码技巧。二是网络抓取数据的网站问题。每个网站都有一个特定的结构，可能随时更改，导致网站抓取技术不断变化。三是网络价格与位置对应问题。在传统的价格收集调查中，选择市场中最受欢迎的门店或零售营业额最高的门店进行价格收集，使价格数据能够代表该地区的大部分消费者，但在网络商店的价格收集中，需要在全国各地进行大规模的调查，花费巨大。因此，要将这些价格纳入CPI，还需要制定一些替代方案。四是价格收集的频率问题。在标准调查中，价格是在一周中某一天的高峰时段收集的，选择高峰时段是为了获得大多数消费者支付的价格。在线商店的价格变化非常频繁，甚至按小时计算，在这种情况下，确定数据收集的时间点变得非常困难。五是产品匹配问题。基于网络抓取数据的主要问题包括产品分类和指数聚合，在传统的价格收集中，价格收集者可以很容易识别产品是否相同，而当前的匹配方法无法识别描述更改。由于数据量大，不匹配的产品很难找到可比较的替代品，这就限制了某些指标的代表性和样本量。六是法律问题。经常从网站上提取大量数据合法吗？从某企业的网站提取数据需要许可吗？这取决于正在抓取的数据类型、访问和复制的信息量以及访问对页面所有者的系统和数据使用的负面影响程度。需要考虑的一个重要问题是网络抓取是否可能违反网络站点的使用条款。当我们访问并停留在一个特定的网站时，经常同意根据其条款使用该网站，但一个网站上允许的内容可能在另一个网站上被禁止，且在许多情况下网站上根本没有任何可用的使用条款。大多数网站都强调其网站上的所有信息都受到版权法的保护，未经网站所有者的明确同意，不应下载或复制数据。然而，《挪威统计法》明确规定，国家统计局有义务提供必要的资料以编制官方统计数字，在法律上有权收集资料，无需通知资料拥有人。奥地利没有任何法律程序涉及网络抓取的可接受性。但在其他欧洲国家如德国已经有关于在线数据库所有者权利的法庭判决，以防止网络抓取者系统地复制内容。

四、基于网络数据的价格指数编制方法研究

价格指数编制方法存在差异，同样的数据在不同的计算方法下会产生不同的指数结果。基于网络数据的价格指数编制方法研究主要集中在两个方面：一是单独基于网络抓取数据的价格指数编制方法；二是网络数据纳入传统CPI统计范围的价格指数编制方法。与传统的CPI数据不同，网络价格数据没有商品支出权数，一般按日收集，频率更高，数量更大。由于商品网站上展示的产品更新换代快，报告期与基期的产品匹配度低，时间上同质可比性差，并且由于数据量大，不匹配的产品很难找到可比较的替代品，这就限制了某些指标的代表性和样本量。因此，研究基于网络数据的价格指数编制方法十分必要，可以加深我们对价格行为的理解。下面分析几种适合于高频率和高容量数据的方法，以探讨应用于网络抓取数据的最适当方法。

（一）单独基于网络抓取数据的价格指数编制方法

1. 选取链式加权指数法计算aSPI

以生活费用理论为基础的aSPI指数体系不仅包括价格指数系列，还包括实物交易量指数系列。价格指数反映一定时期内网络零售商品一般价格变化，实物交易量指数反映一定时期内网络零售交易实物量的一般变化。价格与实物交易量指数系列除总体指数外，还包括食品、衣着等九个基本分类指数。aSPI建立在叶子类目每月加权成交均价基础上，采用链式指数算法，以反映全网总体网购支出价格水平的变化。链式加权具体实施可采用间接法和直接法两种方法。

（1）间接法

间接法先计算相邻时期共有最细类目平均价格的平均值，利用平均值计算相邻两期共有最细类目的交易额，交易额之比即为可比价格的不变类目交易物量增长率。基于基期价格计算的基期交易额，乘以此比率，就得到可比价格的当期交易额。将根据当期价格计算的当期交易额与可比价格的当期交易额相比，就得到当期网络零售交易额的价格平减指数。这是一种先计算实际交易物量，再计算物价指数的间接方法。具体公式如下：

相邻两期共有最细类目：$J_t = K_t \cap K_{t-1}$

最细类目成交均价：$\bar{p}_{jt} = \dfrac{\sum_{i_{jt}} p_{i_{jt}}}{I_{jt}}, \forall j \in J_t$

相邻两期平均价格：$\tilde{p}_{jt} = \dfrac{\bar{p}_{j,t-1} + \bar{p}_{jt}}{2}, \forall j \in J_t$

物量指数：$QX^I_{[t-1,t]} = \dfrac{\sum_j \tilde{p}_{jt} * I_{jt}}{\sum_j \tilde{p}_{jt} * I_{j,t-1}}, \forall j \in J_t$

物价指数：$IX^I_{[t-1,t]} = \dfrac{\sum_j \bar{p}_{jt} * I_{jt}}{QX^I_{[t-1,t]} * \sum_j \bar{p}_{j,t-1} * I_{j,t-1}}, \forall j \in J_t$

以 $t=0$ 为基期的指数：$IX^I_{[0,t]} = \prod_{s=1}^{t} IX^I_{[s-1,s]}$

其中，K_t 表示 t 期淘宝网后台最细类目集合，i_{jt} 为类目 j 在 t 期的第 i 笔交易，I_{jt} 为类目 j 在 t 期的总交易笔数，$QX^I_{[t-1,t]}$ 为间接法计算的以 $t-1$ 为基期的 t 期不变类目物量指数，$IX^I_{[t-1,t]}$ 为间接法计算的以 $t-1$ 为基期的 t 期不变类目物价指数，$IX^I_{[0,t]}$ 为间接法计算的以 $t=0$ 为基期的 t 期不变类目物价指数。

（2）直接法

直接法也是先计算相邻时期共有最细类目平均价格的平均值，同时还计算共有最细类目在两个时期的成交量与成交份额。在此基础上，计算拉氏（Laspeyres）与帕式（Paasche）指数。作为对通用的拉氏与帕式指数的额外改进，汤式（Tornqvist）指数法也可在这一步一道实施，为应用者提供更多的选择[②]。具体公式如下：

最细类目价格指数：$IX_{jt}=\frac{\bar{p}_{jt}}{\bar{p}_{j,t-1}},\forall j\in J_t$

最细类目成交份额：$w_{jt}^{L}=\frac{\bar{p}_{jt}I_{jt}}{\sum_j \bar{p}_{jt}I_{jt}},\forall j\in J_t$

交易笔数固定在当期、价格固定在上期时的成交份额：$w_{jt}^{P}=\frac{\bar{p}_{j,t-1}I_{jt}}{\sum_j \bar{p}_{j,t-1}I_{jt}},\forall j\in J_t$

Laspeyres物价指数：$IX_{[t-1,t]}^{L}=\sum_j IX_{jt}*w_{j,t-1}^{L},\forall j\in J_t$

Paasche物价指数：$IX_{[t-1,t]}^{P}=\sum_j IX_{jt}*w_{j,t-1}^{P},\forall j\in J_t$

Tornqvist物价指数：$IX_{[t-1,t]}^{T}=\prod_{i=1}^{n} IX_{jt}^{\frac{w_{j,t-1}^{L}+w_{jt}^{L}}{2}}$

以 $t=0$ 为基期的指数：$IX_{[0,t]}^{LorPorT}=\prod_{s=1}^{t} IX_{[s-1,s]}^{LorPorT}$

其中，IX_{jt} 表示以 $t-1$ 为基期的 t 期最细类目 j 的价格指数，w_{jt}^{L} 为最细类目 j 的当期交易份额，w_{jt}^{P} 为假定当期交易笔数与上期价格情况下的交易份额占比。

2. 固定基期Jevons指数（Fixed Based Jevons Index）

固定基期Jevons指数将基期固定在数据集中的第1期，并选取所有时期的共有产品进行计算。具体公式如下：

$$P_{FBJ}^{0,t}=\prod_{j\in S^*}\left(\frac{p_j^t}{p_j^0}\right)^{\frac{1}{n^*}}$$

其中，p_j^t 为产品 j 在时期 t 的价格，S^* 为所有期共有产品集合，n^* 为 S^* 中产品的数量。

3. 链式双边Jevons指数（Chained Bilateral Jevons Indices）

该指数首先计算 i 期相对于 $i-1$ 期的Jevons指数，然后将该指数序列连乘得到。公式定义如下：

$$P_{CJ}^{0,t}=\prod_{i=1}^{t}P_J^{i-1,i}=\prod_{i=1}^{t}\left(\prod_{j\in S^{i-1,i}}\frac{p_j^i}{p_j^{i-1}}\right)^{\frac{1}{n^{i-1,i}}}$$

其中，$P_J^{i-1,i}$ 为第 i 期相对于 $i-1$ 期的Jevons指数，p_j^i 为产品 j 在时期 i 的价格，$S^{i-1,i}$ 为 i 期和 $i-1$ 期共有的产品集合，$n^{i-1,i}$ 为 $S^{i-1,i}$ 中产品的数量。

4. 单位价值指数（Unit Value Index）

单位价值指数定义为时期0和时期 t 两个不匹配产品集均值之比，具体公式如下：

$$P_{UV}^{0,t}=\frac{\left(\prod_{j\in S^t}p_j^t\right)^{\frac{1}{n^t}}}{\left(\prod_{j\in S^0}p_j^0\right)^{\frac{1}{n^0}}}$$

其中，S^0 为时期0的产品集，n^0 为 S^0 中的产品数量，S^t 为时期 t 的产品集，n^t 为 S^t 中的产品数量。

5. GEKS指数族(GEKS Family of Indices)

GEKS指数族是一组指数,下面分别介绍其中的GEKS-J指数、RYGEKS-J指数、ITRYGEKS指数、IntGEKS-J指数。

(1)GEKS-J指数

GEKS-J指数是一个多边指数,使用两个时间段之间的全路径计算。以时期0为基期的时期 t 的GEKS-J价格指数是以每一个中间点($i=1,\cdots,t-1$)为连接的时期 t 相对于时期0的链式Jevons价格指数的几何平均值。出现在时期 i 并且出现在时期0或时期 t 的产品包含在指数中。具体公式如下:

$$P_{GEKSJ}^{0,t}=\prod_{i=0}^{t}(P_J^{0,i}P_J^{i,t})^{\frac{1}{t+1}}$$

(2)滚动年份的GEKS链式指数(RYGEKS-J指数)

GEKS-J指数测算中当有新时期的数据时需要不断修正前期数据,为了克服这个缺点,Ivancic等(2011)[16]提出了RYGEKS-J指数。RYGEKS-J指数计算过程是假设初始窗口包含的数据是0至 t 期的数据,根据初始窗口计算第一个GEKS指数。当使用新时期数据时,窗口包含的数据变成了1到 $t+1$ 时期的数据,根据此窗口数据计算第二个GEKS指数,依次类推。具体公式如下:

$$P_{RYGEKS-J}^{0,t}=\begin{cases}\prod_{i=0}^{t}(P_J^{0,i}P_J^{i,t})^{\frac{1}{t+1}},t<d\\ \prod_{i=0}^{d-1}(P_J^{0,i}P_J^{i,d-1})^{\frac{1}{d}}\prod_{k=d}^{t}(\prod_{i=k-d+1}^{k}(P_J^{k-1,i}P_J^{i,k})^{\frac{1}{d}})\end{cases}$$

其中,窗口长度d选择的是13个月。

(3)特征虚拟Tornqvist-RYGEKS指数(ITRYGEKS指数)

RYGEKS指数中忽视了质量变化的影响,因此需要进行质量调整。De Haan和Krsinich(2012)[17]提出了以估算的Tornqvist作为RYGEKS指数的基础,其中估算的Tornqvist指数是特征调整的Tornqvist指数,新产品或消失产品的价格分别使用当前或基期的特征回归来估算,特征回归假设产品的价格由一组k个特征决定。估算的Tornqvist指数定义如下:

$$P_{TT}^{0,t}=\prod_{j\in S^{0,t}}(\frac{p_j^t}{p_j^0})^{\frac{w_j^0+w_j^t}{2}}\prod_{j\in S_{N(0)}^t}(\frac{p_j^t}{\overrightarrow{p_J^0}})^{\frac{w_j^0}{2}}\prod_{j\in S_{D(t)}^0}(\frac{\overrightarrow{p_J^t}}{p_j^0})^{\frac{w_j^t}{2}}$$

其中, w_j^0 为产品 j 在时期0的支出份额, w_j^t 为产品 j 在时期 t 的支出份额, $\overrightarrow{p_J^t}$ 为缺失产品在时期 t 的估计价格, $S^{0,t}$ 为在两期同时观察到的产品集, $S_{N(0)}^t$ 为时期 t 观察到而时期0观察不到的产品集, $S_{D(t)}^0$ 为时期0观察到而时期 t 观察不到的产品集。De Haan和Krsinich(2012)[17]提出了三种计算 $\overrightarrow{p_J^t}$ 的方法,具体如下:

A. 线性特征方法

每期使用回归模型估计特征参数,具体公式如下:

$$\overrightarrow{p_J^t}=\exp(\overrightarrow{\alpha^t}+\sum_{k=1}^{K}\overrightarrow{\beta_k^t}z_{jk})$$

其中, $\overrightarrow{\alpha^t}$ 为截距项, $\overrightarrow{\beta_k^t}$ 为特征 k 对价格的影响程度, z_{jk} 为产品 j 的特征 k 的值。

B. 加权时间虚拟特征方法

该模型假定特征参数不随时间变化,引入虚拟变量 D_j^t ,具体公式如下:

$$\overrightarrow{p_J^t}=\exp(\vec{\alpha}+\overrightarrow{\delta^t}D_j^t+\sum_{k=1}^{K}\overrightarrow{\beta_k}z_{jk})$$

其中, $\overrightarrow{\delta^t}$ 表示特定时间参数估计。

C. 加权时间产品虚拟方法

该方法中当详细的产品特征信息不可用时引入一个虚拟变量 D_j，具体公式如下：

$$\overrightarrow{p_j^t} = \exp(\vec{\alpha} + \vec{\delta^t} D_j^t + \sum_{j=1}^{N-1} \overrightarrow{\gamma_j} D_j)$$

其中，$\overrightarrow{\gamma_j}$ 为特定虚拟产品的参数估计值，第 N 个产品作为参考产品。该方法认为对消费者来说不同产品的质量是不同的，这是一个合理假设，因为潜在特征的数量很大并且不是所有的特征都可见。

以上三种方法都以支出额为权重，使用加权最小二乘估计。

(4)交叉GEKS-J指数(IntGEKS-J指数)

IntGEKS-J指数指数用于处理较长窗口长度下RYGEKS的明显变平问题。该方法仅包含在时期0、i 和 t 共有的产品集，用 $S^{0,i,t}$ 表示。具体公式如下：

$$P_{IntGEKSJ}^{0,t} = \prod_{i=0}^{t} (P_{J,j\in S^{0,i,t}}^{0,i} P_{J,j\in S^{0,i,t}}^{i,t})^{\frac{1}{t+1}}$$

如果没有产品变动(产品进出库)，IntGEKS-J就降低为标准GEKS-J。IntGEKS-J要求产品在更长时间内出现，导致比标准GEKS-J更有可能“失败”。

6. 固定效应窗口拼接指数(FEWS)

固定效果窗口拼接产生一个不可修改的并且完全质量调整的价格指数，在详细的产品规格水平上有纵向价格和数量信息。该方法基于固定效应指数，定义如下：

$$P_{FE}^{0,t} = \frac{\prod_{j\in S^t} (p_j^t)^{\frac{1}{n^t}}}{\prod_{j\in S^0} (p_j^0)^{\frac{1}{n^0}}} \exp(\overline{\hat{\gamma}^0} - \overline{\hat{\gamma}^t})$$

其中，$\overline{\hat{\gamma}^0}$ 为时期0固定效应回归系数的估计均值。使用固定效果回归克服了时间虚拟ITRYGEKS的一些缺点。就像RYGEKS-J一样，在初始估计窗口之后，新序列被拼接到当前序列上，用于后续的周期，这称为窗口拼接。窗口拼接本质上使用的是估计窗口期间的价格移动，而不是最近期间的价格移动，需要在当期指数质量与长期指数质量之间进行权衡。从长期来看，FEWS方法将消除由于没有对新产品和正在消失的产品的隐含价格变动进行调整而产生的任何系统性偏差。该方法的完整描述见Krsinich(2016)[18]。

7. 大型数据集聚类价格指数(CLIP)

CLIP是国家统计局最近开发的一种价格指数，该指数将产品分组到集群中，并随着时间的推移追踪这些集群。在基期产品根据特征进行集群，随着时间的推移集群根据同一规则形成，但是形成集群的产品可能会随着时间的推移而变化，从而导致产品的波动。先对两个时期集群的几何平均值作比，为每个集群建立一个单位值指数，然后使用基期集群大小对其进行聚合。具体公式如下：

$$P_{CLIP}^{0,t} = \frac{\sum_k |C_{k,0}| \dfrac{(\prod_{j\in k,t} p_j^t)^{\frac{1}{|C_{k,t}|}}}{(\prod_{j\in k,0} p_j^t)^{\frac{1}{|C_{k,0}|}}}}{\sum_k |C_{k,0}|}$$

其中，$C_{k,0}$ 为时期0时的集群 k，$C_{k,t}$ 为时期 t 时的集群 k，$|C_{k,0}|$ 为时期0时集群 k 的大小。该方法详情见Metcalfe等(2016)[19]。

（二）网络数据纳入传统CPI统计范围的价格指数编制方法

1. 加权几何平均数方法（加权GM方法）

价格指数由分别计算的线下市场相对价格几何平均数和线上市场相对价格几何平均数加权得到。具体公式如下：

$$G^{off}=(\prod_{j=1}^{n}\frac{P_{ij}^{c}}{P_{ij}^{0}})^{\frac{1}{n}}$$

$$G^{on}=(\prod_{k=1}^{m}\frac{P_{ik}^{c}}{P_{ik}^{0}})^{\frac{1}{m}}$$

$$w_i=w_i^{off}+w_i^{on}$$

$$\text{产品 } i \text{ 价格指数}=[(G^{off})^{w_i^{off}}\times(G^{on})^{w_i^{on}}]^{\frac{1}{w_i^{off}+w_i^{on}}}$$

其中，n 为抽取的线下市场的数量，m 为抽取的线上市场的数量。P_{ij}^{c} 为当期从第 j 个线下市场收集的第 i 个产品的价格，c 表示当期，0表示基期，G^{off} 为线下市场相对价格的几何平均数。同理，P_{ik}^{c} 为当期从第 k 个线上市场收集的第 i 个产品的价格，G^{on} 为线上市场相对价格的几何平均数。w_i 为第 i 个产品的支出份额即第 i 个产品的支出在总支出中所占的份额，w_i^{off} 为线下市场第 i 个产品的权重，w_i^{on} 为线上市场第 i 个产品的权重。

2. 利用网络价格指数修正同期CPI方法

利用网络商品价格指数修正同期CPI，具体方法如下：

$$P=W_{off}P_{off}+W_{on}P_{on}$$

其中，W_{off} 表示实体店社会消费品零售额占比，W_{on} 表示网络社会消费品零售总额占比，P_{off} 为根据传统实体店调查数据计算得到的CPI指数，P_{on} 为根据电商平台交易计算的消费品价格指数[4]。

基于网络抓取数据编制的价格指数在数据收集技术、采价点、采价时间、采集数据量以及抽样范围等方面不同于传统发布的CPI。将网络数据纳入传统CPI编制过程尚处于探索阶段，相关研究比较少，目前主要考虑从数据范围和数源途径上纳入传统CPI。将线上线下价格指数融合可以借鉴模型平均法。模型平均法以其稳健性好、遗失有用信息少等诸多优点成为目前统计学和计量经济学界研究的热门问题，在经济、金融、生物、医学等领域有着广泛的应用前景。模型平均法主要分为频率模型平均（FMA）和贝叶斯模型平均（BMA）两大类，权重选择是模型平均理论研究中最重要的问题。学者对基于FMA的权重选择进行了大量研究：Buckland等（1997）[20]根据信息准则权重提出了S-AIC和S-BIC方法；Hjort和Claeskens（2003）[21]提出了S-FIC方法；Hansen（2007）[22]基于最小化Mallows准则提出了MMA估计；Liang等（2011）[23]提出OPT方法，同时证明OPT估计是渐进最优的；为解决存在异方差的线性模型平均问题，Hansen和Racine（2012）[24]提出JMA方法；Gao等（2016）[25]提出基于删组交叉验证的LsoMA方法；Zhu等（2017）[26]提出基于马氏距离的MMMA方法等。模型平均法将成为线上线下价格指数融合方法研究的一个方向。

由于网络价格具有数据规模大、更新速度快、种类繁多等特征，传统的价格指数编制方法存在许多不足，如链式价格指数一般存在链式漂移、权重缺失等问题。而GEKS指数族能够有效解决以上不足，在大数据背景下应用前景广阔。在此基础上对于集群产品，运用CLIP编制价格指数也是一个好的选择。网络数据纳入传统CPI统计的价格指数编制方法研究较少，一般采用线上线下价格指数加权平均，模型平均法将成为指数融合方面一个好的研究方向。

五、基于网络数据的价格指数相关实证结果

关于网络价格指数与传统CPI之间关系的研究主要有以下观点：

一是不同学者关于网络价格指数与传统CPI之间变动趋势的研究结果不同。刘发跃和马丁丑（2015）[27]将aSPI和CPI分别作为线上和线下价格指标，研究发现线上价格指数普遍高于线下价格指数，并且波动更大。Metcalfe等（2016）[19]针对食品、非酒精饮料和酒精饮料开发了web抓取CPI，研究发现这一指数与公布的CPI数据有类似的长期趋势，但在价格走势上有所不同。Alberto Cavallo（2017）[11]通过对10个国家56家大型多渠道零售商的网站和实体店同时收集的价格进行大规模比较发现，在大约72%的情况下，价格水平是相同的，价格变化不是同步的，但有相似的频率和平均大小。余芳东（2018）[2]研究发现基于网络抓取数据的CPI与基于商店采价数据的CPI有着类似的变动趋势，基于网络抓取数据的CPI趋势拐点要比传统发布的CPI提前1个月。Radzikowski和Mietanka（2016）[12]认为在线消费者价格指数与传统的通胀衡量方法具有互补性。田涛和周薇薇（2017）[28]通过对aSPI指数及其各分类商品价格指数与国家统计局公布的CPI历史数据关联关系定量分析，发现线上线下商品价格之间存在稳定均衡的关系。

二是认为网络价格指数对传统CPI具有良好的预测能力。Hull等（2017）[13]通过研究瑞典一些在线零售商收集选定的水果和蔬菜的销售价格，发现日数据信息可以提高短期通货膨胀预测的精度。方匡南和曾武雄（2018）[29]通过研究aSPI和基于传统编制方法的官方CPI之间的关系，发现阿里网购价格指数与官方CPI之间是周期匹配的，阿里网购价格指数对CPI具有一定的预警和预测能力。Powell等（2018）[14]研究的模型揭示了不同产品类别之间动态行为的不同级别，能够在产品类别特定的CPI发布之前立即对其进行良好的预测，并且认为高频率的月度综合CPI预测是一个可以实现的目标。

此外，价格之间的关系可能因参考时期不同而不同；尽管线上和线下价格在年度基础上遵循相似的趋势，但在某些方面，月度指数存在显著差异；不同的采购渠道之间的价格变动可能有很大的差异，与实体店相比，在线商店的定价策略可能有很大不同等。

总之，对于网络价格指数与传统CPI之间的关系，不同学者基于不同研究基础在两者变动趋势、波动幅度等方面得出的结论有所不同，且认为网络价格指数对传统CPI有良好的预警和预测能力。

六、研究展望

（一）需要更好的方法对数据进行分类

深入研究无监督机器学习技术和有监督机器学习技术，以提高准确性和效率。无监督机器学习技术不需要人工创建训练数据集，无监督机器学习的两个关键例子是k均值聚类和主成分分析（PCA），它们可以用于从数据中推断结构。有监督机器学习技术需要一个训练数据集，该训练数据集用于训练分类算法，经过训练的算法可以用来对不可见数据进行分类，有监督机器学习技术的例子有逻辑回归、神经网络或支持向量机，这些技术可以根据价格的特点对价格进行系统的分类。这可以与无监督机器学习一起使用。此外，需要对网络爬虫器进行编辑，以收集零售商的产品代码，使用这些代码和产品描述提高匹配的质量，并利用更多的可用数据。

（二）探索更好的将高频数据编制成价格指数的方法

继续探索编制高频指数的方法，研究如何将网络抓取价格与专业价格收集者选择的具有代表性的价格相结合来计算价格指数。从官方统计机构的角度来看，使用在线数据是非常有前途的，最有希望的方法是某种形式的混合方法。从在线数据中提取的高频实时指标可以校正利用扫描仪数据或传统的现场采集数

据等出现的问题。将价格指数编制方法与中国实践更多地结合是未来的一个研究方向。

（三）更及时地公布新指数

目前，国家统计局在月后13号左右公布月度传统CPI，季度、年度则延至月后20号左右，公布滞后。为提高价格指数的时效性，不论是网络价格指数、传统CPI还是两者融合指数，当商品价格信息或属性信息发生变动时，都应及时更新价格指数。基于此，应进一步探讨更及时发布价格指数的方法。

注　释：

① 互联网在线数据的居民消费价格指数（iCPI）项目组成立于2015年9月，由清华大学社会科学学院经济学研究所的刘涛雄教授、汤珂教授与清华大学计算机系的许斌教授联合指导，团队运用大数据的理念和技术手段，采集来自电商平台、价格信息网站等的商品价格数据，设计和编制了一套基于互联网在线大数据的居民消费价格指数，可实现每日于网站（www.bdecon.com）可视化发布，并且可以在CEIC数据库下载，数列编码是422327377。

② 我国国家统计局测算官方CPI时采用的是链式“拉氏”公式，官方CPI测算方法可参考国家统计局的走进CPI专题（http://www.stats.gov.cn/ztjc/tjzs/zjcpi/index_1.html）。

参考文献：

[1] 易冰，赵子东，刘洪波.CPI中人工采集网络价格的实践与思考[J].中国统计，2014，（9）：9-10.

[2] 余芳东.国外网络抓取数据在CPI统计中的应用实践[J].调研世界，2018，（7）：3-6.

[3] 陈相成，乔晗.扫描数据支持下CPI编制方法研究[J].统计研究，2013，（1）：23-30.

[4] 陈梦根，刘浩.大数据对CPI统计的影响及方法改进研究[J].统计与信息论坛，2015，30（6）：8-13.

[5] 李平.对我国现行CPI统计方法的思考及完善意见[J].价格理论与实践，2007，（3）：56-57.

[6] 宋晨.我国现行居民消费价格指数编制方法的改进研究[D].北京：中国石油大学，2009.

[7] 许涤龙，谢敏.CPI编制方法的国际比较[J].中国统计，2008，（7）：28.

[8] 高艳云.中美CPI数据质量的比较分析——基于国际货币基金组织的DQAF框架[J].统计研究，2008，（11）：51-56.

[9] 高艳云. CPI 编制及公布的国际比较[J].统计研究，2009，（9）：15-20.

[10] 石刚. 提高 CPI 数据质量的编制技术研究评述[J].统计研究，2012，（5）：105-112.

[11] Alberto Cavallo. Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers[J]. American Economic Review，2017，107（1）：283-303.

[12] Radzikowski B，Mietanka A. Online CASE CPI[C].First International Conference on Advanced Research Methods and Analytics，2016.

[13] Hull I，Löf M，Tibblin M. Price Information Collected Online and Short-term Inflation Forecasts[C].IFC-Bank Indonesia Satellite Seminar on“Big Data”at the ISI Regional Statistics Conference，2017.

[14] Powell B，Nason G，Elliott D，et al. Tracking and Modelling Prices Using Web-scraped Price Microdata：towards Automated Daily Consumer Price Index Forecasting[J]. Journal of the Royal Statistical Society：Series A（Statistics in Society），2018，181（3）：737-756.

[15] Kjersti N H，Leiv T S R. Keeping Up with the Modern Consumer-Online Data in Price Statistics[J]. Conference of Nordic Statisticians Stockholm，2016，（8）：22-24.

[16] Ivancic L，Diewert W E，Fox K J.Scanner Data，Time Aggregation and the Construction of Price Indexes[J]. Journal of Econometrics，2011，161（1）：24-35.

[17] De Haan J，Krsinich F. The Treatment of Unmatched Items in Rolling Year GEKS Prices Indexes：Evidence from New Zealand Scanner Data[C]. Meeting of Groups of Experts on Consumer Price Indices Organized Jointly by UNECE and ILO at the United Nations Palais des Nations，Geneva Switzerland，2012.

[18] Krsinich F.The FEWS Index：Fixed Effects with a Window Splice[J]. Journal of Official Statistics，2016，32（2）：375-404.

[19] Metcalfe L，Breton R，et al. Research Indices Using Web Scraped Price Data：Clustering Large Datasets into Price Indices（CLIP）[C]. Office for National Statistics of UK，2016.

[20] Buckland S T, Burnham K P, Augustin N H. Model Selection: An Integral Part of Inference [J]. Biometrics, 1997, 53(2): 603-618.

[21] Hjort N L, Claeskens G. Frequentist Model Average Estimators[J]. Journal of the American Statistical Association, 2003, 98(464): 879-899.

[22] Hansen B E. Least Squares Model Averaging[J]. Econometrica, 2007, 75(4): 1175-1189.

[23] Liang H, Zou G, Wan A T K, et al. Optimal Weight Choice for Frequentist Model Average Estimators[J]. Journal of the American Statistical Association, 2011, 106(495): 1053-1066.

[24] Hansen B E, Racine J S. Jackknife Model Averaging[J]. Journal of Econometrics, 2012, 167(1): 38-46.

[25] Gao Y, Zhang X, Wang S, et al. Model Averaging Based on Leave-subject-out Cross-validation[J]. Journal of Econometrics, 2016, 192(1): 139-151.

[26] Zhu R, Zou G, Zhang X. Model Averaging for Multivariate Multiple Regression Models[J].Statistics, 2017, 52(1): 1-23.

[27] 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，(20)：29-32.

[28] 田涛，周薇薇.大数据背景下线上商品价格变动对CPI的影响[J].统计与决策，2017，(13)：34-38.

[29] 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，(2)：28-35.

（责任编辑：彭晶晶）