投资研究中的大数据分析趋势及应用

规划研究部　李芮

宏观研究通常被视作大类资产配置的起点：以美林时钟和风险平价为代表的资产配置模型，建立了宏观经济状态和各类资产表现的理论和逻辑关联，从而奠定了当前投资研究中“自上而下”分析的基础。然而，对于投资者而言，即便明确了宏观经济走势和大类资产表现之间的关系，也并不意味着能够轻松完成资产配置的工作，因为现代经济具有高度复杂性，其趋势和拐点均难以直观呈现在某一特定表征上。尤其在近些年来，随着技术革新和产业演进，新模式、新业态层出不穷，导致传统统计方法所搜集的信息量正在逐步缩减和偏离。如此一来，投资者如果希望更加精准及时地判断宏观经济的未来走势，就必须修正传统的统计方法，加入新的统计参数和变量，建立更加完整的分析框架及相应数据库。在这一方面，近年来蓬勃发展的大数据分析无疑值得关注。

一、传统宏观经济分析与大数据分析

传统宏观经济分析建立在凯恩斯提出的总需求框架上，一般以 GDP 增长核算为基础，通过将总需求分解到消费、投资、出口等多个驱动因素上，并区分代表“量”的实际产出

和代表“价”的货币指标，来对经济活动做出整体性的刻画和分析。这一分析框架在逻辑体系上似乎无懈可击，但具体到统计手段上却存在几个较为明显的缺陷：一是受制于人力物力条件限制，传统统计方法几乎完全建立在抽样基础上，尽管在中心极限定理假设下，大样本群体的参数会无限接近总体，但由于现实中许多变量的并不完全为正态分布，抽样统计必定会牺牲掉总体的部分信息；二是传统指标的涵盖范围越来越难以适应当前经济结构和形态的快速变化，例如随着网络线上消费的兴起，传统的商贸零售销售等指标可能无法准确描述居民整体消费情况；三是数据的时效性较低，传统经济指标一般采取问卷和自愿填报的方式经自下而上汇总而来，过程存在较长时滞，一些指标发布延时多达数月之久；四是传统经济统计存在明显的激励不相容问题，上级统计部门严重依赖下级统计部门的统计结果，却无法完全控制下级统计部门或经济主体报送数据的质量，导致数据失真情况时有发生。上述缺陷汇聚到一起，使得当前传统经济统计的结果出现不全面、不准确、不及时的问题。

值得庆幸的是，大数据分析的兴起可以有效克服上述传统经济统计的缺陷。关于大数据的界定目前尚没有一致观点，但一般公认，大数据的特点可以用“4V”即大量化（Volume）、高速化（Velocity）、多样化（Va riety）和价值化（Value）来

概括[①]；换而言之，大数据的特点绝不仅仅限于“量大”，还包括高频、非结构化和高关联度等等。与之相应，大数据分析则着眼于将过去的抽样分析变为总体分析、让过去事后汇总信息变为实时传送信息、将过去非标准化、非结构化信息变为标准化、结构化数据，从而大大提高数据分析的全面性、准确性和及时性。更重要的是，大数据分析承认现实世界的高度复杂性和混杂性，不再尝试建立不同指标间的因果联系，而是更注重考察指标间的相关关系，这种观念更符合现代经济活动的本质。与此同时，以互联网为代表的新经济快速崛起，一方面让大数据分析具备更多的应用场景，另一方面也为大数据分析的应用提供了更多的技术和工具支持。因此，从任何角度来讲，大数据分析都应该成为未来投资研究的重点关注领域之一[②]。

二、宏观大数据分析的主要内容和发展现状

大数据概念出现的时间虽然不长，但在近些年经历了突飞猛进的进步发展，在宏观经济分析领域也得到了较为广泛的应用。从目前情况来看，宏观领域应用比较成熟的大数据指标主要源自以下几个类别：[③]

（一）网络消费数据。网上购物被戏称为中国的新“四

[①] 参见 [英] 维克托·迈尔-舍恩伯格，[英] 肯尼思·库克耶 著，盛杨燕，周涛 译，大数据时代，浙江人民出版社，2013.

[②] 大数据分析在投资领域的另一重要应用领域是分析微观层面上的客户偏好和行为，继而进行相应的产品开发和咨询服务，该领域的应用本文不做探讨。

[③] 需要特别指出的是，一些研究将过去几年在宏观经济中已经得到广泛应用的用电量、货运量、高炉开工率等指标视为大数据，但我们认为这些指标虽然具有高频的特征，但仍为标准的结构化数据，且统计方式仍基于抽样和自下而上汇总，因此并不符合真正意义上的“大数据”分析。

大发明”，正深刻改变着中国居民的生活。从现有的社会消费品零售总额数据来看，网上零售额占零售总额的比例已经达到 20%以上，并且趋势上占比还在快速提升。网络消费领域的大数据分析，既可以在宏观层面了解总需求中的消费以及物价的变化情况；也可以在微观层面用于消费者的行为分析和相应品牌公司的研究。由于网络消费数据的巨大分析价值，以阿里巴巴、京东等为代表的网络零售巨头，在多年的运营中已经积累了大量销售数据，发展出了大数据分析的成熟技术手段，并定期发布关于网络消费的若干代表性指标，如阿里巴巴的“新消费指数系列”、京东的“中国线上消费指数”和“新华▪京东中国线上消费平衡指数”等等。

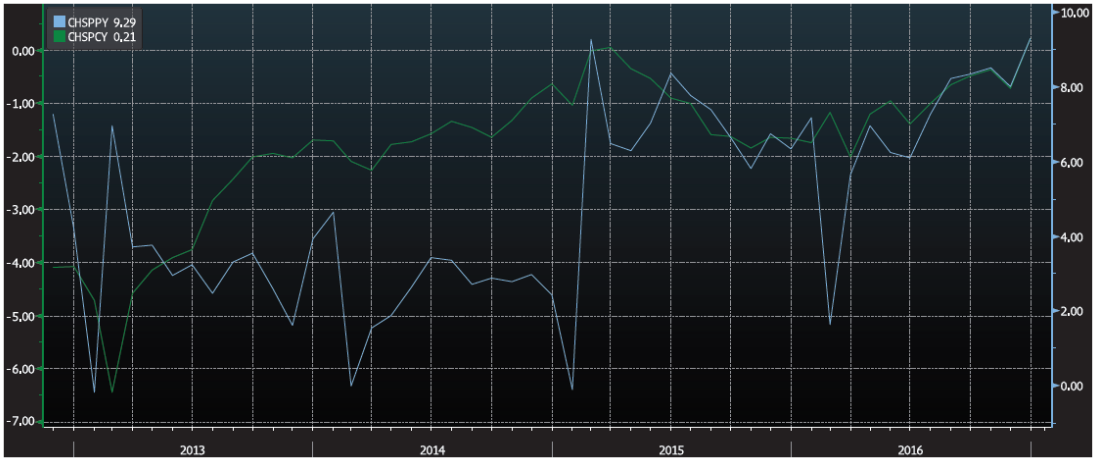
图 1 阿里品质消费指数（2012-2017）

数据来源：国家统计局、阿里研究院

上述大数据分析指标对于深化关于消费的理解具有极大帮助，以阿里巴巴发布的“阿里品质消费指数”为例，可以鲜明地体现出从 2012 年到 2017 年出现的“消费升级”趋势，这无疑能够很好地解释股票市场中消费龙头的优异表现（图

1）。同样由阿里巴巴发布的“阿里巴巴消费价格指数”，则能够比国家统计局的 CPI 数据更加及时和高频地反映出消费价格变化的情况，从而能够佐证研究者对物价变动的判断。

图 2 阿里巴巴消费价格指数（2013-2016）



数据来源：Bloomberg

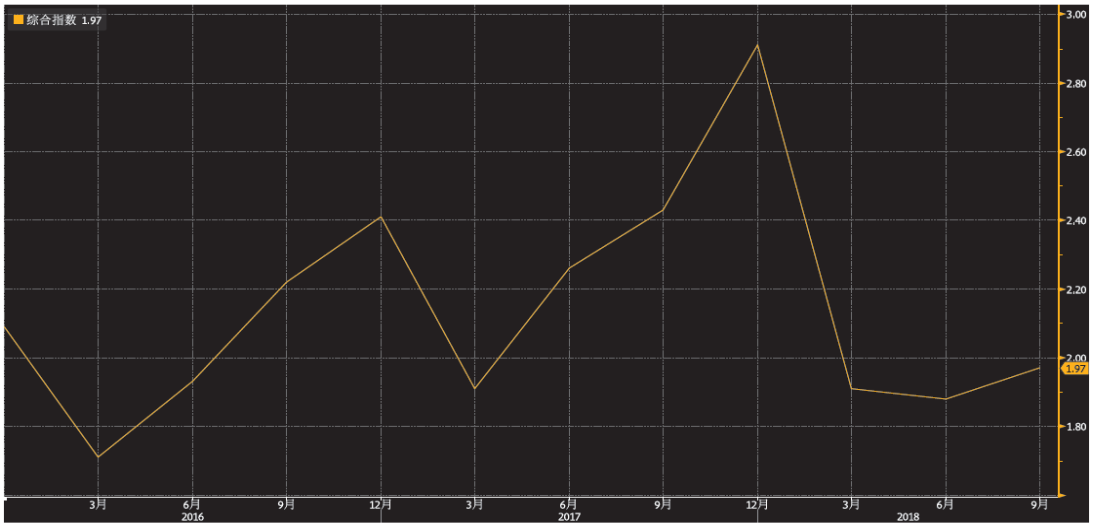
除了社会零售品消费之外，近年来房屋和二手车辆等资产也越来越倾向于采取线上交易的模式，目前相关公司也开始尝试编制和发布相应的大数据分析数据，这无疑将进一步完善我们对于消费活动的监测。

（二）互联网搜索信息。除了网络消费以外，互联网的影响已经渗入现代经济活动的方方面面。互联网搜索引擎作为当前网络用户寻求信息的起点，可以折射出大量关于经济活动的潜在信息。国际上已经有一些学者研究了谷歌趋势(Google Trends)如何可以用来为预测当前经济变量服务，发现失业和相关福利的搜索可以提高对于失业救济首次申请时间的预测[④]。而在国内，随着网络使用频次的提高，不少

[④] 参见 N. Askitas, K. F. Zimmermann, “Google Econometrics and Unemployment Forecasting”, Applied

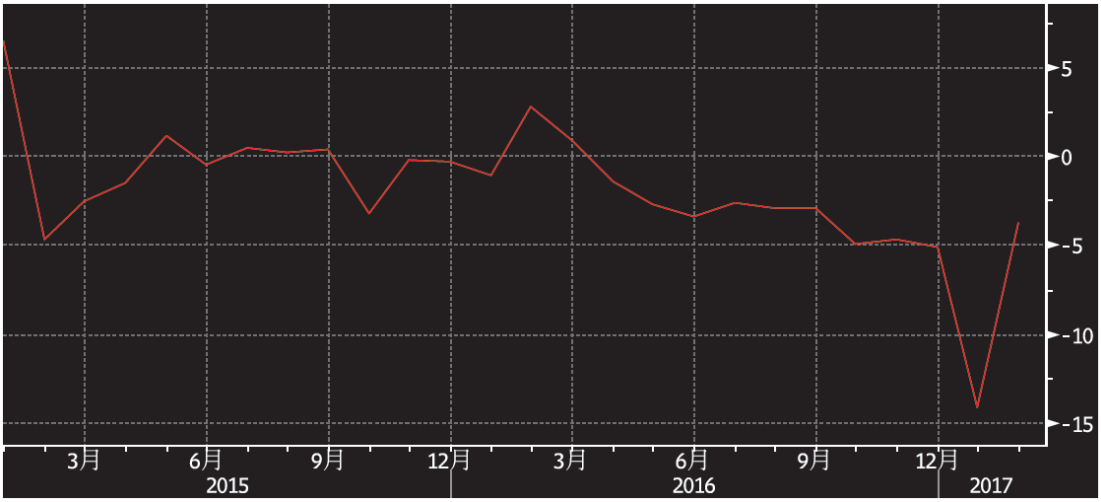
公司已经根据网络搜索情况编制了相应的大数据分析指数，其中比较具有代表性的包括智联招聘发布的“智联招聘就业指数”和百度公司的“百度就业指数”（图 3 和图 4）。

图 3 智联招聘就业指数（2016-2018）



数据来源：Bloomberg

图 4 百度就业指数（2015-2017）



数据来源：Bloomberg

（三）卫星数据。众所周知，卫星图像在地理勘探、气候预测和军事侦查等领域已经具有不可取代的作用。不过近

些年来，研究者却逐渐发掘出卫星图像中包含的另外一些信息——譬如夜间灯光、工业企业排放、交通运输等等——的经济价值。譬如，不少研究都发现，基于夜间灯光对经济增长的估算可以显著提高 GDP 测算的准确性[⑤]。在投资领域，以贝莱德（Blackrock）为代表的国际大型资产管理机构已经开始使用卫星图像中包含的建筑、交通等信息来分析判断行业和公司经营情况并直接为股票投资提供参考。

（四）社交媒体信息。随着社交媒体成为个人表达意见和观点的主要渠道，社交媒体信息对于大众主观情绪的分析具有愈发重要的意义。传统的媒体分析早已在政治学和公共政策研究中大量应用，例如通过分析官媒对某一话题发表的文章数量和关键词频率进行统计，来判断政策意图和走向。然而，网络社交媒体的规模和联结性远远超出了传统媒体分析的能力范围，大量高频和非结构化信息只能通过大数据的提取和处理方法才能转化为可以量化分析的数据。由于社交媒体数据的分析需要综合运用网络爬虫、社会网络分析和文本分析等多种技术手段，掌握起来难度极高，因此目前其前沿应用还主要局限在学术研究中[⑥]。但可以肯定的是，社交媒体信息在揭示投资者情绪和行为上的应用潜力巨大，未来还需要开发更有效的方法去收集和处理相关数据。

---

⑤ 徐康宁,陈丰龙,刘修岩.中国经济增长的真实性:基于全球夜间灯光数据的检验.经济研究,2015,50(9):17-29

⑥ 对于中国社交媒体数据分析的范例可以参见 King G , Pan J , Roberts M . How Censorship in China Allows Government Criticism but Silences Collective Expression[J]. American Political Science Review, 2013, 107(2):326-343.

综上所述，目前宏观经济领域的大数据分析仍处于萌芽阶段，但已经形成了一些成果，如果按照宏观经济的维度对现有指标进行重新排列，可以形成一个简单的指标体系：

表 1 大数据分析指标体系

维度		指标	来源	备注
总需求	消费	阿里消费指数	阿里研究院	未公开
		京东中国消费指数	京东研究院 Bloomberg	
价格	零售品价格	阿里巴巴消费价格指数	阿里研究院	
	房地产价格	房产交易数据指数	贝壳研究院	
实体经济活动	就业	智联招聘就业指数	Bloomberg	
		百度就业指数	Bloomberg	
	城市灯光	卫星数据	美国海洋与气象局（NOAA）、谷歌地图	需自行加工处理
	物流	卫星数据		
	工业排放	卫星数据		
情绪	投资者情绪	社交媒体信息	网络社交媒体	需自行加工处理
	分析师预测	社交媒体信息	网络社交媒体	需自行加工处理

资料来源：作者总结

从上表可以看出，目前在宏观经济各个维度大数据分析的发展进程不尽相同。由于网络购物的高度发达和互联网巨头企业的巨额投入，互联网消费和价格领域的大数据分析已经比较成熟，亦形成了不少相应指数。相比之下，刻画实体经济活动卫星图像数据和反映情绪的相应大数据分析还主要停留在学术研究阶段，或者仅由少数拥有强大信息技术能力的资产管理机构使用，关于这方面的大数据分析还处于起步阶段。

三、总结与启示

总体来看，大数据分析正对宏观经济研究带来近乎颠覆性的冲击变化，包括从依靠传统统计数据向依靠互联网数据的转变;从关注总量和滞后指标向关注结构和领先指标的转变;从中长期监测预测向实时监测预测的转变。不过，大数据分析作为新兴事物，在普及过程中必定会遭遇不少挑战和问题，需要运用者予以高度重视。

一是处理好大数据分析和传统宏观分析的关系。需要明确的是，大数据分析依然是对传统统计分析方法的补充而非替代，在当前数据获取、处理和分析能力有限的情景下，传统方法仍旧有着无法取代的价值。大数据虽然有及时性的优势，但过于高频的数据可能反而会对研究者造成干扰。举例而言，消费数据会因为消费习惯、节假日、基数效应等因素而出现明显波动，如果仅仅通过大数据分析，可能会导致研究者过于关注短期变化而误判长期趋势。因此，需要综合运用传统宏观经济分析和大数据分析，将长期研判和高频验证相结合，形成对宏观经济运行更综合、更全面的理解。

二是积极开发和运用大数据分析的技术方法。大数据的搜集、处理和分析应用需要技术手段乃至组织架构的革新，其最重要的底层技术基础是统一的数据标准制定。由于标准不一的同类数据之间无法进行合并和比较，因此规范一致的数据标准是对大数据进行预处理、清洗和分类的首要前提。

同时，大数据通常涉及将非结构化数据转存为结构化数据，需要建立固定的规则，并积极应用云计算和机器学习等手段，这都对组织的技术能力提出了更高更新的要求。

三是谨慎应对当前大数据分析中存在的缺陷和问题。大数据时代带来的不仅是益处，也有弊病和挑战。如何处理包括隐私保护、信息安全和数据过载等问题，是每个希望享受大数据福利的机构都必须同时解决的课题。

诚如有言，数据之于21世纪，就如同石油之于20世纪。对于投资者而言，认识和重视大数据时代带来的变化，掌握并充分挖掘数据背后的价值，是在未来竞争中抢占先机的重要途径。