### 2.1.1 宏观经济指数的大数据监测预测

国内外经济环境、经济增长情况对于资源能源安全具有显著影响。在经济增长中考虑全球经济增长以及国内情况。经济增长的衡量可以使用宏观经济指数，来反映跨产业的、整体的经济运行情况。宏观经济指数是体现经济情况的一种方式，主要指标包括国民生产总值、通货膨胀与紧缩、投资指标、消费、金融、财政指标等，对于宏观经济调控起着重要的分析和参考作用。由于宏观经济指数可以在全球和国家等多个层面进行统计，因此对全球经济增长和国内情况的监测本质上是一致的，都是对宏观经济指数的监测。

通常情况下，政府统计依赖于传统的调查方式来建立宏观经济指数,这些宏观数据往往滞后于政府宏观调控政策制定或企业经营决策需求。（苏志, 张骐）这种滞后对于资源能源安全监测的影响也是重大的，时效性是风险监测的一个核心要求。因此,使用实时更新的互联网大数据相对于传统数据而言更能满足对于即时统计信息的需求。使用互联网大数据进行宏观经济指数监测预测的工作已经包含如下宏观经济指数：价格指数、失业率、消费指数。在本节中会逐一介绍。

此外，如果对经济增长趋势建立合适的指标体系（申红艳, 吴晨生, 扆铁梅,等），确定影响总体经济增长的因素和其权重，也可以通过对子因素的监测加以组合，形成对经济增长趋势的直接描述。所使用的指标体系可以是目前使用较多的，采用人工专家的方式构建。

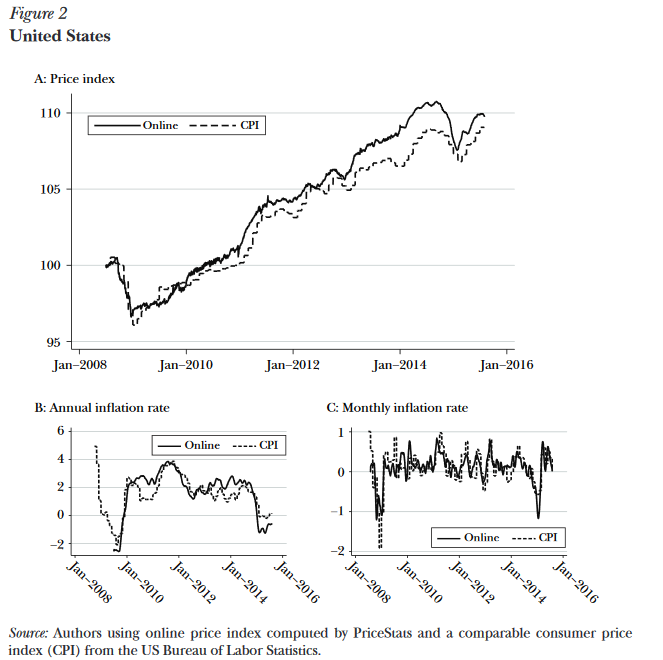
1）统计预测价格指数：

（1）MIT学者创建的Billion Prices Project(BPP)通过采集网络零售商品价格建立零售价格通货膨胀指数,为宏观价格和国际价格比较研究提供新数据来源。Billion Prices Project通过互联网中存在的价格信息构建多个国家的居民消费物价指数（CPI）。（Cavallo A, Rigobon R. The Billion Prices Project: Using Online ）文中使用多种方法来避免统计上的偏差，如数据的滞后性和汇率的影响。传统的CPI测算方式采用工作人员亲自访问商铺的形式得到物价信息，并且只测算选定的一篮子商品的价格。采样的不频繁和一篮子中物品的选择更新缓慢都导致难以对统计质量进行把控，也不利于新的商品添加。互联网数据在这个统计问题中具有天然的优势，一方面由于价格数据散落在数以千万的网页、网站中，信息量巨大，特别是对于在互联网上出现频繁的商品，一方面由于互联网数据的搜集工具——爬虫技术的普及使得数据的收集迅速、高效、精准，同时对于一篮子中商品组合变化也可以进行快速的调整。截止到2010年，该项目平均每天从50余个国家的300个零售商处采集超过5百万条价格信息，远远超过通过传统方式采集到的数据。

文章使用的数据源为零售商的价格数据。网络零售商的选择采用线上+线下的零售商为主，尽量不选择只有线上销售的零售商以保证数据可以涵盖更多的地区。在商品类型的选择上，尽可能选择原有CPI计算时采用的一篮子商品，而避免过多使用在互联网上过度展现的商品如CD、书籍等。该项目没有采用第三方网站提供的价格信息，而是直接从零售商的线上商城处获得价格，虽然增加了获取信息的难度，但是使得数据的采集范围和质量都更加的可控。数据完成采集后进行清洗，标准化，将来自零售商的商品种类纳入标准的商品分类，并使用相关统计方法对数据进行初步分析。对零售商的数据进行一段时间的考察，分析结果不符合要求的将不予以采用，达到数据质量控制的目的。该方法目前在25个国家都采集到了CPI中所占权重达到70%的商品种类。

文章使用的分析方法为传统的CPI计算方法，包括对所统计商品消费者支出权重的分配等。由于与政府发布的CPI指数具有相同的计算方法，因此可以直接进行比较，其差异仅仅来源于数据源。

该项目目前在CPI计算、通货膨胀率计算上取得了与官方数据十分接近的效果。以美国为例，在2008-2016年间，网络价格计算的CPI、年度及月度通货膨胀率与官方数据始终有很好的贴合性，仅有少数数据有较小的差异，具有实用价值。同时，网络价格计算的CPI还对突发事件有更迅速的反应，当突发事件影响到商品价格时，网络价格CPI的变化要提前官方CPI的变化，在美国某事件的分析中可以提前两个月。在中国、巴西、南非、英国、德国、日本等多个国家的实验中也可以验证网络价格CPI与官方CPI的具有很高的相近度。



网络价格数据在CPI计算中一个主要的问题是采样的零售商数目不够全面，同时采样到的商品种类也不够全面，比如很多种服务的价格在网络上是缺失的，在这一点上要逊于传统方法。另一个问题是目前网络价格中缺乏消费者实际支出的相关信息，这些信息目前只能通过政府发布的官方数据，消费支出的调查，以及相关应用产品获得。

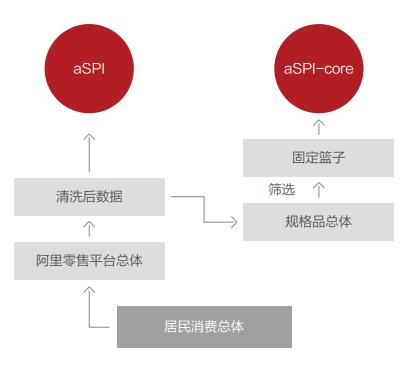
1. “阿里指数”是阿里巴巴构建的数据库。2012年11月26日，阿里指数正式上线，根据阿里巴巴网站每日运营的基本数据包括每天网站浏览量、每天浏览的人次、每天新增供求产品数、新增公司数和产品数这5项指标统计计算得出。其数据大部分是淘宝卖家来阿里巴巴采购留下的，它动态的反应了电子商务平台市场的发展情况，被许多跨国公司以及中国大部分中小企业看成中国电子商务市场的一个关键的晴雨表。（基于淘宝和阿里巴巴平台的电商业态分析）

**aSPI-core指数**

aSPI-core（alibaba Shopping Price Index-core，阿里巴巴网购核心商品价格指数是固定篮子价格指数，通过创新的筛选算法圈定阿里零售平台上近五百个基本分类下接近100000种核心商品作为固定“篮子”，每月追踪该篮子内商品和服务实际网购成交价格变化，以刻画网购主流商品和服务的一般价格波动，从而从网络零售渠道反映宏观物价走势。建立在严格可比的规格品成交价格基础上，数据误差较小，与通常应用的宏观物价指数在基本方法论上具有较高一致性。

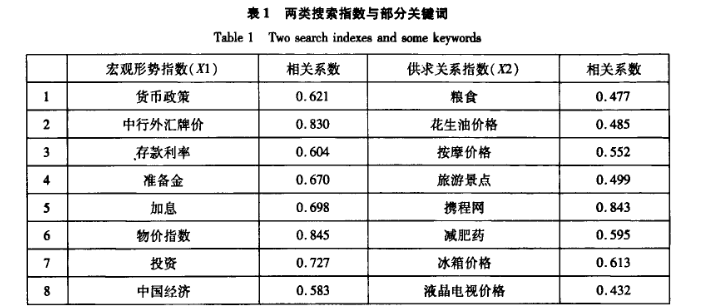
**aSPI指数**

阿里巴巴全网网购价格指数（aSPI）建立在叶子类目每月加权成交均价变动的基础上，以上月成交份额为权重所计算得到的价格指数，用以反映全网总体网购支出价格水平的变化。它同时包含了商品层面一般价格变动和消费者在叶子类目下消费结构变动的信息。 由于现代社会产品更新换代频率加快，许多产品的销售及价格具有生命周期特征，使得固定篮子指数aSPI-core在长期有可能会低估消费支出成本的上升趋势。以生活费用理论为基础的阿里巴巴全网网购价格指数（aSPI），在理论上能克服固定篮子指数面临的这一局限。该价格指数衡量的是消费者获得效用不变条件下，由产品相对价格变动引发消费数量上的替代效应，所带来的最小支出水平的变动。



1. 网络搜索数据与CPI相关性

文章（网络搜索数据与CPI的相关性研究）研究了Google搜索数据与CPI的相关性。数据选择上采用谷歌搜索解析，关键词搜索总量，2004-2009年的月度数据。表为部分搜索关键词。



分析方法采用回归模型，并使用平稳性检验、协整检验等。实验效果验证Google搜索数据与CPI的确有显著的相关性，模型拟合度达到0.978，预测绝对误差为0. 48。宏观形势搜索指数和供求关系搜索指数相对于CPI的先行周期分别为五个月和两个月；同时模型具有很强的时效性，比国家统计局的数据发布提前一个月左右。同时，搜索数据还对CPI及通货膨胀情况具有一定的预测作用。

1. 预测失业率、就业情况

失业率作为一个重要的宏观经济指标，一直受到研究者广泛关注。失业率的预测是网络搜索数据应用于社会经济研究的一个较早的课题。许多研究都表明谷歌趋势中与就业相关的词条查询或招聘查询指数可预测德国、以色列、土耳其、意大利、美国的失业率趋势。而且，部分模型的效果要优于基于专业预测人士调查的失业率预测模型。下表为部分代表性研究的总结。

|  |  |  |  |
| --- | --- | --- | --- |
| 相关研究 | 数据源 | 方法 | 效果 |
| Xu W, Li Z, Chen Q（2012） | 2004-2011谷歌搜索数据与美国失业率，500个失业相关词 | 神经网络 | 对现有模型的预测性能有提升 |
| Askitas N, Zimmermann K F（2009） | 搜索数据与德国失业率 | 传统经济学回归模型加入搜索数据 | 建立了某些关键词的搜关注度与失业率的关联 |
| Suhoy T（2009） | 搜索数据，以色列 | 传统经济学回归模型加入搜索数据 | 预测长期和短期的失业初请人数，绝对误差分别降低 15.74% 和 12.90% |
| Choi H, Varian H（2012） | Google 中有关“工作”和“福 利和失业”类别搜索指数，美国的周失业率 | 传统回归模型加入搜索数据 | 模型的拟合度有较大的提高 |
| Francesco D（2009） | 失业相关的Google 搜索指数、意大利的季度失业率 | 基于较小的样本建立工作搜索指数、将该指数加入传统预测模型 | 模型的预测效果显著高于传统模型，采样频低仍有较好效果 |
| 彭赓, 苏亚军, 李娜（2012） | Google 推荐的关键词搜索数据 | 改进的逐步回归方法，分层建立预测模型，因果关系检验及有效性检验 | 模型的拟合优度分别达到 0.930、0.935、0.936，三期预测值的 MAPE 分别为 1.20%、 0.89%、0.57%。 |

再利用该方法进行失业率预测时，搜索数据关键词的选取是一个重要问题。由于面对的都是海量的搜索数据和关键词，真正具有预测价值的关键词却是需要甄别与筛选的。其主流的方法可以总结为三种。（彭赓, 苏亚军, 李娜，2012）

第一种是采取技术取词法。即利用高性能、大规模的计算设备将一切可能的关键词都纳入到研究范围内。然后计算与主题之间相关性选出核心关键词。例如 Ginsberg 等人利用 800 余台高速计算机在 2003 年~2008 年间 5 000 万个最为常用的搜索词中选择出 45 个与 CDC 发布的流感病人就诊量数据（研究课题）相关性最高的关键词作为预测关键词的来源。

第二种是经验取词法。即由作者运用主观经验确定关键词。例如Askitas N, Zimmermann K F（2009）在网络搜索与失业率相关性时，认为与劳动局或失业机关机构、失业率、人事顾问和德国比较流行的几个职业搜索引擎四类关键词的搜索量将出现变化。因而以这四类关键词为核心合成搜索指标。

第三种是范围取词法。 即先确定一个选词的范围，然后在范围内进行精选。最初的词范围可以是某些工具提供的相关类目关键词，也可以是各个类目中较为重要的关键词。精选的准则通常是度量与主题之间的相关程度。该方法可以看成是首先使用经验法选取词的范围，再用技术取词法选取最终的关键词，是两者的一种折中。例如 Konstantin 在研究网络搜索与美国个人消费增长率的相关性时，首先收集了 Google 提供 的 27 个分类中的前十大搜索词，然后分别做出相关性分析后剔除与个人消费不相关的词，利用剩下 220 个与消费相关的词合成一个指标 。Xu W, Li Z, Chen Q（2012）从 Google Trends 中与失业相关的分类中收集了 500 个左右的关键词作为原始关键词，从这 500 个关键词中找出相关系数大于 0.65 的 108 个关键词。然后利用神经网络方法，从这 108 个关键词中又筛选出少数几个关键词进行拟合。（彭赓, 苏亚军, 李娜，2012）则是先选一个和失业相关的关键词作为初始关键词，由该关键词出发得到Google Trends 推荐的与其热门相关 10 个关键词。进一步搜索这 10 个关键词，得到Google Trends 推荐出其热门相关的大约 100 个关键词。如此往复，剔除重复的关键词后，这些关键词构成基础关键词库，约有403个关键词。

技术选词法精度较高，但是容易受到资源的限制而难以复制应用。直接取词法虽然降低了工作量但是主观性较强，降低了学术研究的科学性。范围取词法可以看为一种折衷。

1. 统计预测消费指数

消费者信心指数（Consumer Confidence Index ，CCI）是反映消费者信心强弱的指标，是综合反映并量化消费者对当前经济形势评价和对经济前景、收入水平、收入预期以及消费心理状态的主观感受，是预测经济走势和消费趋向的一个先行指标，是监测经济周期变化不可缺少的依据。

国际上通行的消费指数编制做法，对消费者信心（或情绪）调查采用的是[问卷调查法](https://baike.baidu.com/item/%E9%97%AE%E5%8D%B7%E8%B0%83%E6%9F%A5%E6%B3%95" \t "https://baike.baidu.com/item/_blank)。问卷的设计紧密围绕以下几个方面内容：经济发展形势、家庭收入和就业、[物价水平](https://baike.baidu.com/item/%E7%89%A9%E4%BB%B7%E6%B0%B4%E5%B9%B3" \t "https://baike.baidu.com/item/_blank)、消费或[购买意愿](https://baike.baidu.com/item/%E8%B4%AD%E4%B9%B0%E6%84%8F%E6%84%BF" \t "https://baike.baidu.com/item/_blank)。每一方面由两类问题构成：对现状的看法和对未来的预期。

（Vosen S, Schmidt T，2011）使用互联网大数据预测美国总体个人消费水平变化,无论使用样本内还是样本外的预测，其精度都比基于调查数据构建的两种消费者信心指数预测精度要高。其具体使用的数据为使用Google Insights获得的56种与消费相关的关键词与搜索量，关键词包括耐用消费品、非耐用消费品以及服务类的细分商品。分析方法为计量经济学中的VAR模型。



1. 预测经济增长

在宏观经济增长监测预测的模型建构方面，国内学者和研究人员进行了许多大 胆的创新和尝试，国内的研究成果就有国家信息中心的“中国宏观经济模型”、中国人民银行的“季度计量经济模型”以及厦门大学的“中国季度宏观经济模型”等。然而传统的宏观经济监测预测模型都是基于同频数据进行的，高频数据必须要降为低频数据。这样会造成数据信息的丢失，进而影响模型预测的准确性。而且国内现有的宏观经济监测预测模型以年度、季度模型为主，周期较长。因而有必要采用大数据的方法在时效性上进行改进。（陈龙, 王建冬等, 2016）

对于利用大数据进行经济增长预测，构建合适的经济增长判断模型，国内学者也提出了一些的看法。

国内学者的一个常用的研究框架是基于“克强指数”，该指数包含三个经 济指标，分别是“工业用电量新增”、“铁路货运量新增”和“银行中长期贷款新增”。“克强指数”源于时任辽宁省委书记的李克强总理会见美国驻华大使时表示，他喜欢通过耗电量、铁路货运量和贷款发放量三个指标来分析和预测辽宁省的经济运行情况。

（刘慧，2014）通过构建“克强指数”与经济增长的 VAR 和 VEC 模型，发现“克强指数”的三大指标与经济增长之间 存在长期均衡关系和短期调整机制。（申红艳, 吴晨生, 扆铁梅等，2014）把国内利用大数据进行宏观经济分析中预测经济增长的研究分为三类，与“克强指数”选取的指标相类似。一是用电量与经济增长。大多数研究表明，用电量，尤其是工业用电量与经济增长之间存在长期稳定的均衡关系和因果关系。而且用电量与经济增长之间的这种关系也得到国外学者研究的佐证。二是货运量与经济增长。研究发现，货运量，尤其是铁 5 路货运量与经济增长之间存在交替推拉作用的因果关系。三是银行贷款与经济增长。（刘恩猛, 汪波，2007）发现经济增长和贷款之间存在协整关系和双向因果关系。

“克强指数”也得到了国外研究者和相关机构的认可。2010 年，英国著名政经杂志《经济学人》将“克强指数”视为评估中国 GDP 增长的重要指标。《经济学人》杂志根据这三项指标构造了一个指数，并画出这个指数的时间序列曲线，冠名为“克强指数”。“克强指数”被一些国际机构所认可，如花旗银行就用它来对比工业企业利润，并认为它的解释力更强。

除此之外，国内也完成了一些利用大数据技术监测预测宏观经济增长的相关课题，取得了一定成果。2013 年 7 月全国人大财经委员会向中央提交了《企业发展和宏观经济发展关系分析》报告，该报告根据“企业发展工商 指数”预测了我国宏观经济将企稳回升的趋势，对经济增长采用大样本数据和相关技术进行了预测，而后来的经济发展形势证实了这一指数预测的准确性。“企业发展工商指数”涉及 10 个对宏观经济具有显著先行性的指标，可以提前 1～2 个季度预测宏观经济发展形势。

1. 预测GDP

吉林大学的刘汉和刘金全验证了混频数据抽样模型（MIDAS）对中国季度 GDP 的监测和预测能力。混频数据模型（MIDAS）可以利用混频数据，避免高频数据降为低频数据时的信息流失，提高了宏观经济监测预测的准确性。研究发现，出口是造成金融危机阶段中国经济增长减速的主要成因。而且，混频数据模型在短期预测中国宏观经济方面具有比较优势，在实时预报方面具有显著的可行性和时效性（刘汉、刘金全，2011）。

Liu T, Xu X, Fan F.（2016）采用了一个两步的模型对我国的GDP预测进行了探讨，首先仅使用来自于政府部门的结构化统计数据构建模型，选取最优的模型，其次将互联网的搜索行为数据加入前一步的最优模型，再进行挑选即可得到最优模型。