

上海新金融研究院工作论文系列

**No. SFIWP0028**

**基于互联网搜索量的先导景气指数、需求预测及消费者购  
前调研行为——以汽车行业为例**

冯明 刘淳

2014年1月24日

说明：上海新金融研究院是为支持上海国际金融中心建设而成立的非营利性金融类专业学术研究机构，由中国金融四十人论坛举办，并与上海市黄浦区人民政府展开战略合作。

本工作论文是上海新金融研究院研究人员在工作期间形成的、尚未公开发表的研究成果，文中观点仅代表作者本人，不代表本研究院。未经书面同意，谢绝任何形式的转载和复制。

# 基于互联网搜索量的先导景气指数、需求预测及消费者购前调研行为

## ——以汽车行业为例<sup>①</sup>

冯明<sup>②</sup>，刘淳<sup>③</sup>

**摘要** 随着互联网的逐渐普及，越来越多的消费者开始在网上进行购前调研行为，而搜索引擎使得捕捉和追踪这些购前调研行为成为可能。本文以汽车行业为例，首次通过互联网搜索量频率数据设计并构建了中国汽车需求先导景气指数“GCAI”，并基于此对中国汽车消费者的购前调研行为进行了研究。实证分析发现该先导指数有较强的预测力，不仅可以提高预测精度，还可以增强预测的时效性。同时，我们发现搜索量可以度量消费者的“关注”，而且汽车行业消费者的购前网上调研行为是分阶段的，有明显的“U型”规律。

**关键词** 搜索引擎，需求预测，汽车，购前调研行为，先导指数

## 0 引言

商品的需求量是市场中一个非常重要的指标。准确和及时的需求预测，无论对企业进行生产、库存、定价等方面的内部决策，还是对消费者合理选择购买时机，以及宏观经济调控，都有着重要的参考作用。传统的需求预测模型一般是以历史销售量信息和对市场状况的估计为基础。而在互联网时代，消费者的一部分购买行为会转移到网上，比如购前通常会在网上进行信息搜索和浏览，购后会在网上进行商品评价等。消费者的这些网上行为与他们的真实需求关系密切。如果能提取这些网络记录中的有效信息，将可能提高预测能力，为市场参与者的决策提供更有效的信息支持。

链接网上行为与真实需求的是消费者的“关注”。在当今信息爆炸的时代，消费者的“关注”是非常稀缺和重要的资源（Kahneman, 1973）。尽管“关注”仅仅意味着购买的可能性而非必然性，但消费者的“关注”与最终购买之间的确存在着正向关系。因此现代市场营销学的核心目标之一就是研究如何获取或吸引消费者的关注。互联网的普及事实上创造了一种使消费者能更有效分配“关注”的机制，而互联网搜索引擎（如谷歌、百度等）的出现使得跟踪消费者“关注”成为可能。

本文研究的基本逻辑是：个人或家庭在进行大件商品的消费决策之前，往往会进行市场调研。在传统情况下，这种调研经常是通过逛商场、向亲朋好友打听等方式进行的。而随着计算机和互联网的普及，网络成为消费者进行购前调研的重要途径。而且，消费者基于互联网的购前调研往往是从谷歌、百度等搜索引擎开始的。例如，有购买“诺基亚”手机意向的消费者在购买之前通常会先通过百度等搜索引擎输入“诺基亚”、“E71<sup>④</sup>”或“诺基亚智能手机”等关键字进行搜索，根据搜索引擎的反馈结果，从而进入诺基亚官方网站、某手机零售商的网站、或者某网络“贴吧”进行更加详细的购前调研，并在此基础上做出是否购买的决定。

<sup>①</sup> 本研究得到国家自然科学基金资助项目（项目号为71202019和71232003），特此致谢。本文即将发表于《营销科学学报》，作者非常感谢两位匿名评审专家提出的宝贵建议。

<sup>②</sup> 冯明，通讯作者，金融学博士，清华大学经济管理学院，上海新金融研究院青年研究员。E-mail: [fengm.06@sem.tsinghua.edu.cn](mailto:fengm.06@sem.tsinghua.edu.cn)。

<sup>③</sup> 刘淳，清华大学经济管理学院金融系副教授，E-mail: [liuch@sem.tsinghua.edu.cn](mailto:liuch@sem.tsinghua.edu.cn)。

<sup>④</sup> “E71”是诺基亚手机的一种型号，此处仅为举例引用，无特别的含义。

最近几年出现了大量运用互联网搜索数据进行预测的应用研究。Cooper et al. (2005)、Polgreen et al. (2008)、Ginsberg et al. (2008)、Brownstein et al. (2009)、Corley et al. (2009)、Hulth et al. (2009)、Pelat et al. (2009)、Valdivia and Monge-Corella (2010)、Wilson and Brownstein (2009)等人的相继研究表明,互联网搜索数据可以在癌症、流行病传染等的预测中发挥作用。Ginsberg et al. (2008)发现互联网关键词搜索数据能比疾病预防控制中心提前一到两周预知流感爆发。

经济学领域最早的研究是 Ettredge et al. (2005),他的研究对象是美国的失业率。随后,Askitas and Zimmermann (2009)、Baker and Fradkin (2011)、D'Amuri and Marcucci (2010)、Suhoy (2009)等分别将互联网搜索量数据用于美国、德国、以色列的失业率预测。Guzman (2011)运用谷歌搜索量数据进行通货膨胀率的预测。McLaren and Shanbhoge (2011)的研究表明中央银行可以利用搜索量数据进行现时经济指标的估值。Preis et al. (2010)发现搜索量数据能很好地指标股票市场投资者的“关注”度,证实互联网搜索量和股票交易量之间存在相关性。Da et al. (2011)发现搜索量高的标的股票在接下来的两周内有更高的收益,同时更高的搜索量也意味着更高的IPO溢价。

运用互联网搜索数据进行需求预测的研究有:Vosen and Schmidt (2011)、Lindberg (2011)对零售行业的研究,Wu and Brynjolfsson (2009)对房地产销量的研究等。Penna and Huang (2010)利用谷歌数据构建了“消费者情绪指数”,并证实该指数与传统文献中通常使用的密歇根大学“消费者情绪指数”、世界大型企业联合会“消费者信心指数”高度相关。Choi和Varian (2012)对谷歌搜索量数据进行了细致的描述,并运用简单的统计模型展示了搜索量数据在汽车销售、失业率、旅游目的地等方面的预测能力。

国内学者运用互联网搜索数据进行的学术研究还不多:梁志峰(2010)利用谷歌的搜索量数据对湘潭地区的网络关注度进行了剖析,主要对一些关键词的搜索量进行了时间及空间上的比较。宋双杰等(2011)证实了谷歌搜索量数据对IPO市场热销程度、首日超额收益和长期表现都有很好的解释力和预测力。张崇等(2012)的研究发现:网络搜索数据与我国CPI之间存在协整关系,而且其构建的指数可以作为CPI的先导指标,时效性比国家统计局的数据发布提前一个月左右。

对于不同类别的产品,购物决策中的信息搜索行为是不同的。洪成一等(2003)、廖成林和史小娜(2012)分析了消费者在互联网上的搜索行为对购买决策的影响。在传统购物环境中,通常将产品分为日用品、选购品和特殊产品,而在网络购物环境中,通常将产品分为搜索产品、体验产品和信任产品(张莱和陈毅文,2006)。消费者通常对两类商品的搜索意愿相对更强:一是对于个人或家庭生活比较重要的大件消费品,如汽车、家电等;二是电子类商品,如电脑、手机、及其配件等。在本文的研究中,我们选取汽车作为研究对象。按照通常的分类,汽车属于特殊产品和体验产品类别,其特点是消费者在做出购买决策前,会进行大量的产品信息搜索和品牌比较活动(徐茵等,2010)。

本文通过互联网搜索量频率数据设计并构建了中国汽车需求先导景气指数“GCAI”,同时对中国汽车市场上消费者的购前调研行为进行了研究。研究发现,基于互联网搜索引擎的搜索量数据对汽车需求具有较强的预测能力,不仅能增强预测效率,而且能提高时效性。另外,通过对互联网搜索频率数据的深度考察,我们还发现中国消费者的购前网上调研行为可以分为两阶段,具有明显的“U型”规律。本文的贡献主要体现在以下三个方面:

首先,通过互联网搜索量数据对消费者购前调研行为的研究具有重要的理论意义。已有的文献往往局限于预测的工具价值,而忽略了对消费者行为的探讨,如Penna and Huang (2010),Ginsberg et al. (2008),Carrière-Swallow and Labbé (2013)等。通过借鉴Preis et al. (2010)、Da et al. (2011)在股票市场研究中的思路,我们发现在消费品市场,互联网搜索量不仅可以度量中小投资者的“关注”(attention),也可以很好的度量消费者的“关注”。尤其重要的是,本文通过对最优预测窗口的细致分析,得出的“U”型规律和两阶段购前调研行为,对市场营销领域的相关研究具有一定的启示意义。据我们所知,本文是第一篇用互联网搜索量数据对消费者购前调研行为进行数量化探讨的

研究。

第二，需求预测本身对于企业的生产、库存、营销等微观决策具有重要意义。而汽车行业又是一个非常重要而且独特的行业，其产业链长，涉及的相关上下游产业多，与宏观经济走势密切相关（刘小平（2012），东北财经大学宏观经济分析与预测课题组（2006））。本文设计和构造的“谷歌中国汽车指数”（GCAI），与目前普遍使用的其他预测指标相比，有更好的先导性和预测力。这个新指标的创立，不仅能用于微观层面的企业决策，而且对于宏观经济预警也有一定的帮助。另外，本文在使用互联网搜索数据预测需求方面提供了一个一般性的分析工具，可以为今后相关研究提供一个新的思路。

最后，中国的互联网市场正呈现着非常独特的特征：一方面，互联网普及率尚低。根据中国互联网络信息中心（CNNIC）测算，截至2011年11月，中国互联网用户总数约为5.05亿，互联网普及率为37.7%。另一方面，我国互联网市场飞速发展，通过网络进行购物的人数越来越多。根据艾瑞咨询发布的《2010-2011年中国搜索引擎用户行为调研报告》，53.3%的互联网用户会通过搜索引擎来了解商品、购物信息。已有文献多以发达国家为研究对象，本文以中国现在特有的互联网发展阶段为背景进行研究，是对相关文献有意义的补充。

本文接下来的基本结构为：首先设计和构造出一个中国汽车需求先导景气指数——“谷歌中国汽车指数”（Google China Automobile Index, GCAI）；第二部分通过时间序列计量模型对GCAI指数的预测效果进行研究；第三部分探讨最优预测窗口，同时对消费者的购前调研行为进行分析和解释；接下来是稳健性检验；最后是总结。

## 1 “先导景气指数”的构建

### 1.1 数据说明及处理

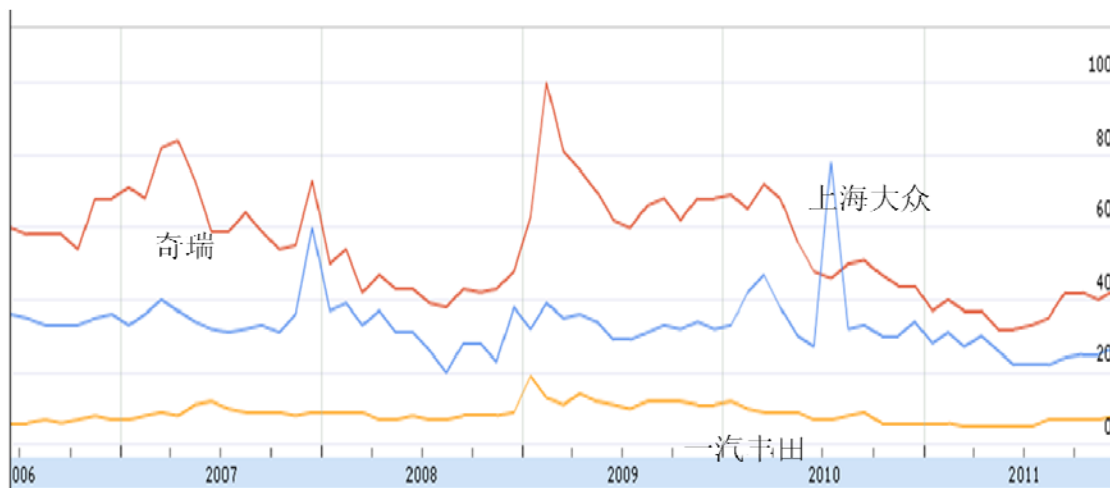
中国搜索引擎市场一直呈现百度和谷歌（google）“双寡头”格局。按照中国电子商务市场权威研究机构“艾瑞咨询”发布的《2011-2012年中国搜索引擎行业年度监测报告》，以本文样本时段的中间点2008年为例，谷歌占搜索请求量份额的20.7%，百度占73.2%，其余被搜搜、搜狗、雅虎等搜索引擎分享。由于百度的市场份额更大，理论上用其搜索频率数据能得到更好的结果，但我们在研究中最终采用了谷歌数据，主要原因在于百度数据未对公众开放，因此无法用来做研究。但百度公布了简单的图片格式的搜索量趋势图，通过对比图1（a）与图1（b）中谷歌与百度两个搜索引擎的搜索量趋势，可以看出二者在整体上相关性较强，其反映出的趋势是类似的。而且，根据艾瑞咨询发布的TGI指数显示<sup>①</sup>，谷歌的用户群与百度的用户群有一定的系统性区别——相对于所有网民而言，个人月收入低于2500元的人群更倾向使用百度，个人月收入高于2500元的人群更倾向于使用谷歌。由于购买汽车的群体更可能属于后一收入分组，因此，我们认为选用谷歌数据进行研究是合理的。从另外一个角度看，由于百度的市场份额更大，其给出的关键词相对搜索量数据的噪音应当更小，因而采用百度数据进行研究只可能强化本文所构建指数的预测效果，而不是相反。另外，尽管在样本期间内各个搜索引擎的市场份额会有所波动，但我们会在下文中说明，这并不对本文的分析构成影响。

本文使用的搜索量数据，为谷歌搜索引擎发布的“谷歌搜索解析”数据。该数据反映了在某一特定时间和区域内，相对于谷歌上执行的全部搜索量，某关键词被搜索的比例。公开数据为其0-100

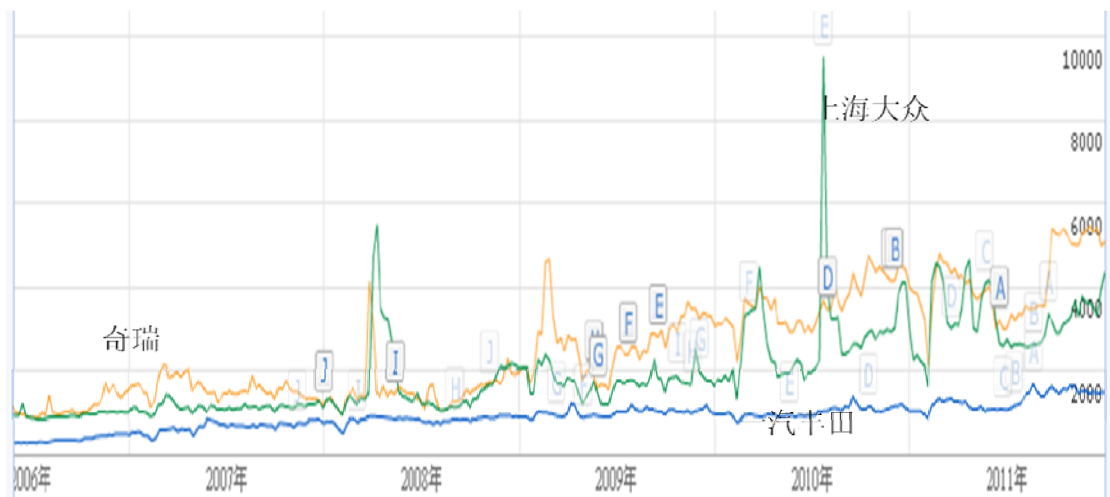
---

<sup>①</sup> TGI指数=目标用户群体在某网站的市场份额/所有网民在某网站的市场份额\*100。资料来源：《2010-2011年中国搜索引擎用户行为调研报告》。

的标准化处理后的结果。以“奇瑞”为例，当我们输入关键词“奇瑞”，并将地理位置参数设定为中国，将时间设定为2004年至今。这时“谷歌搜索解析”首先计算在这个时间内，中国国内针对“奇瑞”执行的搜索次数占总搜索量的百分比。然后将这个时间序列中最高数设定为100，其他数做同比例处理之后，返回的0-100之间的数就是谷歌公布的最终数据。下图1(a)是用以上方法返回的“上海大众”、“奇瑞”和“一汽丰田”三个关键词的搜索量随时间变化的趋势图。需要特别指明的是因为该数据反映的是相对意义的搜索量，这排除了因互联网的普及或谷歌市场份额变化等原因导致的对某一关键词搜索量变化的影响。因为2004年1月是谷歌开始发布“谷歌搜索解析”数据的时间点，而目前“谷歌搜索解析”提供以周为单位时间的搜索数据。因此本研究所使用的互联网数据为2004年1月-2012年2月的周度数据。



1 (a)



2 (b)

图1：搜索量趋势图：(a)为谷歌，(b)为百度  
来源：谷歌（Google）与百度网站截屏

汽车销售量数据来自于中国汽车工业协会，中国宏观经济景气指数来自于国家统计局中国经济景气监测中心网站。下图2描述了中国汽车销量和中国宏观经济景气指数的时间序列图。从该图可以看出，汽车销售量走势与宏观经济大环境有非常紧密的关系，经济的上涨和下跌往往伴随着同向的汽车销量变化，反映了汽车行业的顺周期特性。与宏观经济相比，汽车销售的波动要明显更大。

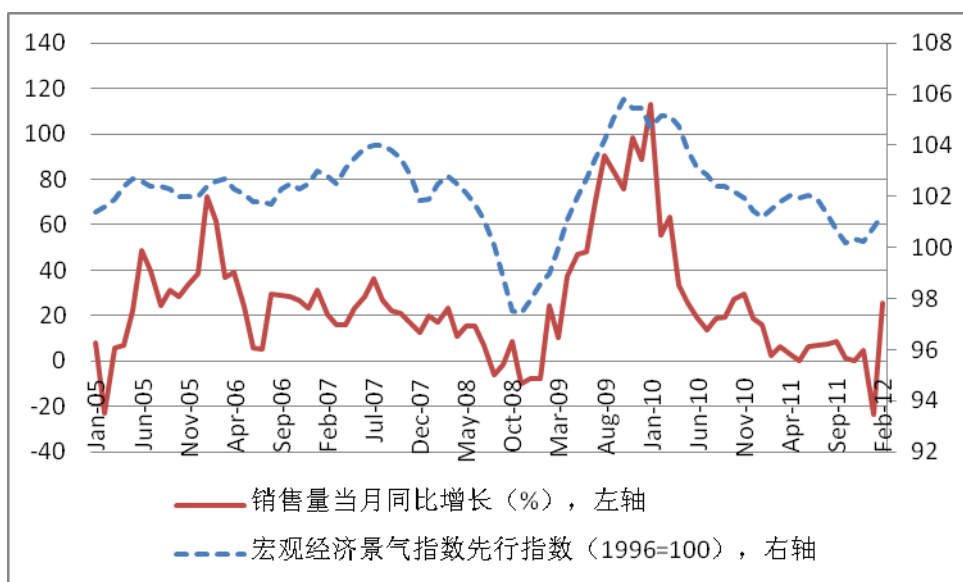


图 2: 汽车销售量与宏观经济景气指数

## 1.2 GCAI 指数构建

本部分中，我们分三个步骤构建出一个中国汽车需求先导景气指数——“谷歌中国汽车指数”（Google China Automobile Index, GCAI）。

第一步：选取数据样本，去除数据噪音。我们根据中国汽车工业协会发布的销量数据得到了 2012 年 2 月销量排名前十位的汽车品牌<sup>①</sup>，分别是：上海大众、东风日产、一汽大众、比亚迪、上海通用别克、北京现代、一汽丰田、奇瑞、东风悦达起亚和上海通用雪佛兰。这 10 个品牌的销量占当月乘用车总销售量的一半以上。由于“谷歌搜索解析”每次反馈的数据是随机抽样得到的，因而不同时间得到的数据会有略微不同；为了消除由此带来的噪音，我们于 2012 年 3 月到 4 月之间随机选取 16 天进行了数据检索，得到了 10 个关键词的 16 个搜索量序列（各序列的时间跨度为 2004 年 1 月-2012 年 2 月，频率为周度）。然后对各关键词 16 个样本点取平均值<sup>②</sup>，最终得到 10 个汽车品牌的搜索量历史数据。

第二步：周度数据转换为月度数据。由于谷歌提供的搜索量数据是周度的，而汽车销售量数据的频率最小是月，我们需要把周度数据转化为月度数据。由于月份和周不能完全对应，本文采取“去头留尾法”进行处理，即将下月初的半周数据加到本月<sup>③</sup>。尽管使用了部分的下月数据，但实际上，由于我国汽车销售量的数据发布机构是中国汽车工业协会，其数据来源于各汽车供应商上报的数据，统计和数据汇总工作带来了数据发布的滞后性。中国汽车工业协会一般于每月的 9-15 日发布上月的数据。因此，通过这种方法，我们得到的指数仍然是超前于此的，可以作为领先指数进行预测。

第三步，我们通过各品牌的搜索数据构造中国汽车市场先导景气指数 GCAI。考虑到各品牌的市场份额是动态变化的，本文参照 Carrière-Swallow and Labbé (2013) 的方法，利用递推样本构造动

<sup>①</sup> 在本文样本区间内，销售量排名前十位的汽车品牌存在略微的变化。因此该指数创建的另外一种办法是类似于编制股票成分指数的做法，定期将某些品牌纳入和去除。但由于分品牌历史数据不可通过公开途径获得，本文仅以该十种品牌作为元素进行指数构建。这种设计方法对于本文的核心结论影响不大。

<sup>②</sup> 我们计算了各时间点“信号噪声比”（Signal-to-Noise Ratio），发现其都在 11 以上。按照 Rose 标准，谷歌后台抽样引起的“噪音”并未对样本起到明显的干扰作用，用均值反映其变化趋势是合理的。

<sup>③</sup> 另一种可以将周度数据转换成月度数据的方法为“按日平均法”，即假定跨月各周的每天搜索量相同，将这些周的总搜索量除以 7，得到每天的搜索量。本月的总搜索量即是本月各天搜索量的加和。我们在研究中发现使用两种数据处理方法得到的结果相差不大，考虑到文章篇幅的限制，本文中只汇报“去头留尾法”得到的结果。

态权重的 GCAI 指数。具体步骤如下：首先对方程

$$q_t = \alpha + \sum_{i=1}^{10} \beta_i g_{i,t} + \varepsilon_t$$

进行回归。其中， $q_t$ 为汽车实际销售量的同比增长率， $g_{i,t}$  ( $i=1,2,\dots,10$ ) 为第*i*个汽车品牌在*t*时刻搜

索量的同比增长率。使用动态递归的样本进行估计<sup>①</sup>，得到其估计值  $E[\beta_i | \Omega_{t-1}]$ ，其中， $\Omega_{t-1}$ 是*t*-1

时刻的信息集。我们将  $GCAI_t$  定义为：

$$GCAI_t = \sum_{i=1}^{10} E[\beta_i | \Omega_{t-1}] \cdot g_{i,t}$$

该指数的构造反映了基于当前信息，通过搜索量预测出的销售量。通过以上步骤构造指数最大化地利用了某一时间点之前拥有的全部信息量，并且不依赖于除搜索数据和销量数据以外的其他数据来源，在实际中更方便和可行。图 3 描述了我们构造出的中国汽车市场先导景气指数 GCAI 与汽车销量之间的时间序列图。可以看出，两者之间确实存在较强的相关性。而且在大部分时间点上，GCAI 指数的变动趋势领先于汽车销售量的变动趋势。关于领先期数我们在下文第三部分会做更为详细的分析。

需要特别说明的是，消费者的购买行为是一系列复杂行为的集合，包括了需求认知、信息搜索、评价与选择、购买、购后行为等阶段，其中汽车消费者在“信息搜索”和“购后行为”两个阶段都可能进行互联网搜索行为。由于消费者在进行搜索时，不会向其他人公布其搜索的目的，所以理论上要区别两种搜索原因是非常困难的。这样以来，本文中构建的 GCAI 指数就包含两方面的搜索行为<sup>②</sup>：以  $GCAI_t$  为例，“*t*时刻后可能购买汽车的消费者在*t*时刻的购前调研搜索”和“*t*时刻前已购买汽车的消费者在*t*时刻的购后搜索”。其中，前者包含预测未来汽车需求（从而汽车销量）的信息，而后者与未来需求无关。虽然在已有的数据条件下我们无法将这两类搜索行为分解开，从而构建纯粹的“购前搜索量指数”；但“*t*时刻后可能购买汽车的消费者在*t*时刻的购前调研搜索”与“*t*时刻前已购买汽车的消费者在*t*时刻的需求后搜索”是相互独立的。“购前调研”反映的是消费者未来的需求，而“需求后搜索”受过去销售量的影响很大，过去汽车的销售量越大，则“需求后搜索”越多。因此在下文的回归方程中，我们将过去的销售量作为解释变量放入了方程中，以其作为“需求后搜索”的代理变量，包含了消费者这种事后搜索行为的信息。

<sup>①</sup> 估计最初 24 个点的  $\beta$  值时，使用 2005 年 1 月-2006 年 12 月的固定样本；2007 年 1 月后的  $\beta$  值由动态递推的样本进行估计。

<sup>②</sup> 作者感谢匿名评审人的这一提示。

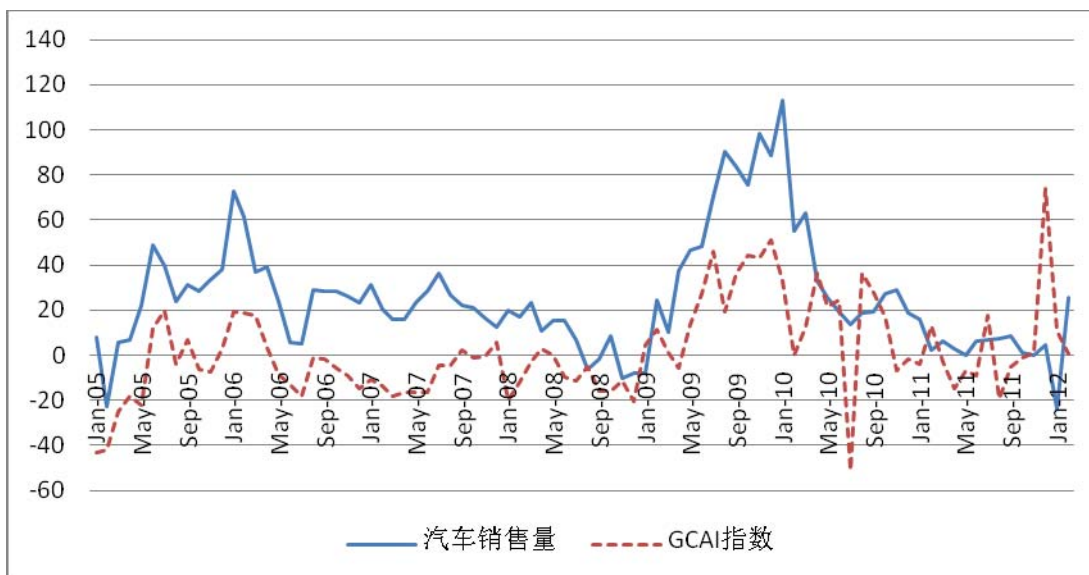


图 3: 汽车销售量与 GCAI 指数

## 2 实证分析及结论

为了分析 GCAI 指数对预测汽车需求的效果，我们选择了如下的四个基准模型：

$$(I.A) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \varepsilon_t$$

$$(I.B) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \delta y_{t-1} + \varepsilon_t$$

$$(I.C) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \sum_{q=0}^m \varepsilon_{t-q}$$

$$(I.D) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \delta y_{t-1} + \sum_{q=0}^m \varepsilon_{t-q}$$

其中， $q_t$ 为 $t$ 时期汽车销量的同比增长率， $y_t$ 为中国经济景气监测中心发布的宏观经济景气先行指数， $\varepsilon_t$ 为随即误差项。在这四个模型，第一个模型（IA）为产品需求预测文献和实务中使用最多的自回归模型（AR模型）。在第二个模型（IB）中，我们在（IA）模型的基础上，加入了表示经济当前及未来状态的“宏观经济景气指数”。中国宏观经济景气指数由国家统计局下属的中国经济景气监测中心编制，包括先行指数、一致指数、和滞后指数。考虑到我们的模型是以预测为目的的，此处选用先行指数。另外，为了更大限度地利用数据提供的信息，参照Carrière-Swallow and Labbé（2013）的研究，我们构建了移动平均自回归模型（ARMA）作为第三个基准模型（IC）。在模型（IC）中加入宏观经济景气先行指数得到基准模型四（ID）。通过AIC准则和SC准则，我们选择了AR(3)和ARMA(3,3)模型<sup>①</sup>。

<sup>①</sup> 变量  $q_t$  和  $y_t$  的 ADF 检验拒绝了其存在单位根的原假设，表明其为平稳序列。限于文章篇幅，我们略去了单位根检验及滞后阶数选择的过程。有兴趣的读者可向作者索取。



## 2.1 全样本估计

为了检验 GCAI 指数是否能够对预测汽车销量提供额外的信息，我们在以上四个基准模型 I.A-I.D 中分别加入上文构建的“GCAI 指数”作为解释变量，得到包含互联网搜索量信息的新模型 II.A-II.D:

$$(II.A) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \varphi^* GCAI_t + \varepsilon_t$$

$$(II.B) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \delta y_{t-1} + \varphi^* GCAI_t + \varepsilon_t$$

$$(II.C) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \varphi^* GCAI_t + \sum_{q=0}^m \varepsilon_{t-q}$$

$$(II.D) \quad q_t = \alpha_0 + \sum_{i=1}^n \alpha_i q_{t-i} + \delta y_{t-1} + \varphi^* GCAI_t + \sum_{q=0}^m \varepsilon_{t-q}$$

其中， $GCAI_t$  基于 t 时刻搜索量数据  $g_t$ ，以及 t-1 时刻信息集构建的动态权重  $E[\beta | \Omega_{t-1}]$ 。模型组 I 和 II 的估计结果如表 1。可以看出在所有的模型中，变量 GCAI 的系数都是显著为正的。这说明当互联网用户更多地通过互联网搜索相关关键词后，紧接着汽车的销售量会明显上升。谷歌搜索引擎的数据对汽车销售量有很好的解释力。另外，对于四个不同的模型设定 A-D，加入互联网搜索量后的新模型与基准模型相比，调整后的  $R^2$  更大，而均方误差更小，证实了加入的新变量 GCAI 提高了模型对数据的拟合能力。

表 1: 回归结果

	I.A	II.A	II.B	II.C	II.D	II.E	II.F	
<b>C</b>	27.47*** (8.79)	19.82** (7.77)	-790.84*** (251.08)	-645.61** (252.23)	24.33** (10.37)	18.05** (8.32)	-688.78** (285.82)	-630.81** (248.10)
<b>Y</b>			8.00*** (2.45)	6.51** (2.47)			7.00** (2.80)	6.36** (2.43)
<b>GCAI(-1)<sup>①</sup></b>		0.26*** (0.08)		0.25*** (0.08)		0.33*** (0.07)		0.31*** (0.07)
<b>AR(1)</b>	0.72*** (0.11)	0.74*** (0.12)	0.52*** (0.12)	0.58*** (0.12)	1.96*** (0.17)	0.3*** (0.1)	1.25*** (0.34)	0.25* (0.13)
<b>AR(2)</b>	0.32** (0.14)	0.20 (0.14)	0.31** (0.13)	0.20 (0.13)	-1.62*** (0.24)	-0.41*** (0.06)	0.17 (0.60)	-0.44*** (0.07)
<b>AR(3)</b>	-0.21* (0.11)	-0.14 (0.11)	-0.10 (0.12)	-0.06 (0.11)	0.57*** (0.12)	0.71*** (0.08)	-0.46 (0.28)	0.68*** (0.10)
<b>MA(1)</b>					-1.41*** (0.21)	0.52*** (0.16)	-0.77** (0.35)	0.44** (0.18)
<b>MA(2)</b>					1.24***	0.89***	-0.23	0.84***

<sup>①</sup> 模型组 II 中 GCAI 项可以是 t 期值，也可能是 t-i 期值，此处我们以拟合效果较好的滞后一期指数作为解释变量。关于滞后期的具体问题，下文会详细讨论。

					(0.25)	(0.09)	(0.48)	(0.10)
<b>MA(3)</b>					-0.25	-0.09	0.03	-0.18
					(0.21)	(0.14)	(0.20)	(0.16)
<b>样本数</b>	83	82	83	82	83	82	83	82
<b>Adjusted R<sup>2</sup></b>	0.70	0.73	0.72	0.75	0.75	0.77	0.72	0.78
<b>Sum squared resid</b>	15653.96	13605.68	14292.04	12650.64	12684.20	11441.67	13630.01	10538.68

注：（1）估计值下括号内的数为标准差；（2）“\*”，“\*\*”和“\*\*\*”分别表示估计参数在 10%，5% 和 1%水平下显著。

## 2.2 样本外预测

我们首先用 I 和 II 两组模型分别进行样本外预测，然后比较其的预测能力。用前四年的样本（2005 年 1 月-2008 年 12 月）进行样本内估计，用 2009 年 1 月-2012 年 2 月的数据进行样本外预测。为了衡量模型预测效果，我们采用了两个常用的统计量指标：误差均方根（Root Mean Squared Error, RMSE）和平均相对百分误差（Mean Absolute Percent Error, MAPE），定义如下：

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=T+1}^{T+n} (\hat{y}_t - y_t)^2}, \quad MAPE = 100 \frac{1}{n} \sum_{t=T+1}^{T+n} \left| \frac{\hat{y}_t - y_t}{y_t} \right|,$$

其中 T 表示模型估计部分所用的样本量，n 表示样本外预测期数， $\hat{y}_t$  表示预测值， $y_t$  为实际观测值。预测结果如表 2 所示。通过表 2 对比两组模型的 RMSE 和 MAPE，我们发现第 II 组模型的 RMSE 和 MAPE 都明显小于第一组的结果，表明加入互联网数据之后的模型有更好的预测效果。

表 2：样本外预测效果比较

	<b>I.A</b>	<b>II.A</b>	<b>I.B</b>	<b>II.B</b>	<b>I.C</b>	<b>II.C</b>	<b>I.D</b>	<b>II.D</b>
<b>RMSE</b>	34.91	30.03	23.20	21.64	37.00	29.48	21.45	21.10
<b>MAPE</b>	1096.30	863.44	834.60	671.84	963.58	815.29	792.76	646.63

图 4 以模型 I.B 与 II.B、I.D 与 II.D 为例，考察了包含互联网搜索信息前后模型的预测效果对比。从整体上看，第 II 组模型的样本外预测值与实际值拟合得更好。更重要的是，在很多销售量走势发生重大变化的时刻，如 2010 年 8 月、2011 年 3 月和 2011 年 11 月等关键点，传统的第 I 组模型几乎没有办法预测到销售量的趋势变化，但这种趋势性变化能够被我们包含搜索量信息的第 II 组模型比较好的预测到。这再次证实了我们创建的基于互联网搜索量信息的先行指数有很强的预测效果。

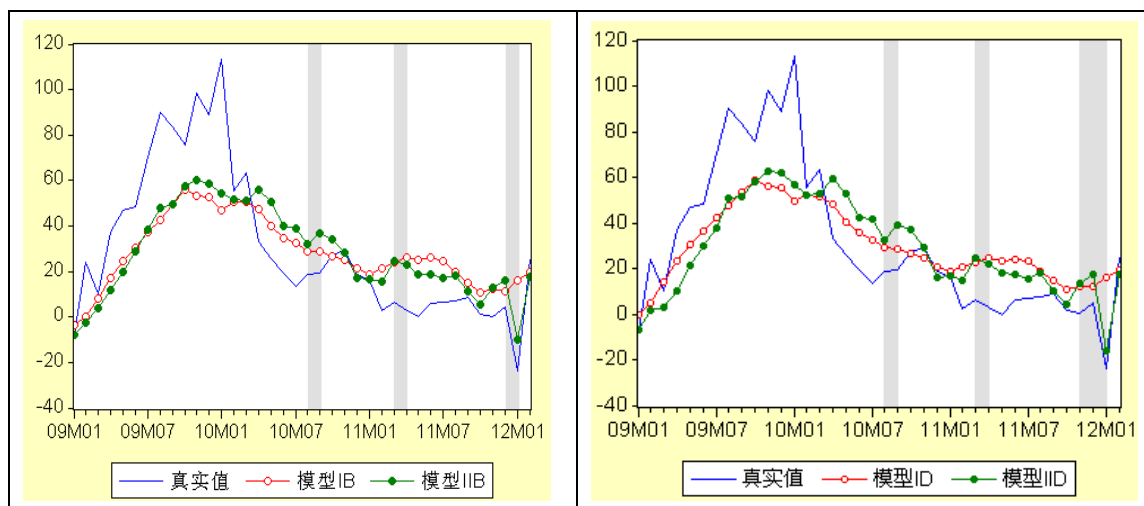


图 4：样本外预测

### 3 最优预测窗口及购前网上调研行为

#### 3.1 最优预测窗口

以上的分析说明我们构造的中国汽车业先导景气指数确实能为市场需求预测提供额外信息，提高预测效率。那么，在进行预测时，最优的预测周期是多长？为了解决这个问题，我们分别以某月汽车销售量作为被解释变量，以先导指数 GCAI 及其滞后项作为被解释变量，考察这两者之间的相关关系，结果如表 3 所示。可以看出随着滞后期数的增加，系数的估计值减小，而且显著性水平也有所降低：当滞后期为 1-4 期时，先导指数的系数在 1% 的水平下显著，意味着其能对预测需求起到作用，在 5 期以后显著性下降；到 7 期以后，先导指数的系数不显著。

表 3：GCAI 滞后各期回归结果

变量	系数	Adjusted R <sup>2</sup>
GCAI	0.70 ***	0.31
GCAI(-1)	0.61 ***	0.24
GCAI(-2)	0.59 ***	0.23
GCAI(-3)	0.54 ***	0.17
GCAI(-4)	0.38 ***	0.08
GCAI(-5)	0.30 **	0.04
GCAI(-6)	0.26 *	0.03
GCAI(-7)	0.20	0.01

注：“\*”，“\*\*”和“\*\*\*”分别表示估计参数在 10%，5%和 1%水平下显著。

谷歌提供的以周为频率的搜索数据使得我们可以就该问题进行更细致的分析。我们按照上文中方法重新构造按周平移的先导指数，表示为“GCAI<sub>n</sub>”，n=1, 2, …, 9 为先导的周数。这样，由于一月大概包含 4 周，指数 GCAI<sub>5</sub> 涵盖的周与指数 GCAI<sub>t</sub> 的一阶滞后项 GCAI(-1)涵盖的周是大致相同

的，指数  $GCAI_9$  涵盖的周与指数  $GCAI_t$  的二阶滞后项  $GCAI_{t-2}$  是大致相同的。

我们以某月汽车销售量作为被解释变量，以先导指数  $GCAI$ 、 $GCAI_1$ 、……、 $GCAI_9$  为解释变量，进行单变量回归，结果如表 4。与表 3 的结果一致，表 4 中  $GCAI$ 、 $GCAI_1$ 、……、 $GCAI_9$  的系数都是非常显著的。同时，我们还进一步发现：随着“n 周先导指数”中 n 的增大，其对销售量的预测效果先降低后提高，呈“U 型”变化趋势，如图 5 所示。当期和滞后一周的指数预测能力最强，其相关系数和  $R^2$  都是最大，表明本周和前一一周的网络搜索量对即将公布的本月销售量有最好的预测效果。随着滞后期的增加，预测能力开始下降，在 3-6 周时达到低谷，经调整的  $R^2$  在 12% 左右。之后预测力又开始上升，在 8-9 周时预测效果再次达到高潮，经调整的  $R^2$  达到了 30% 左右。说明本期的搜索量对大约 2 个月之后公布的汽车销售量也有很好的预测作用。

表 4：最优预测窗口

变量	系数	Adjusted R <sup>2</sup>
GCAI	0.70***	0.31
GCAI1	0.72***	0.36
GCAI2	0.63***	0.24
GCAI3	0.29***	0.11
GCAI4	0.30***	0.14
GCAI5	0.30***	0.13
GCAI6	0.35***	0.12
GCAI7	0.44***	0.19
GCAI8	0.59***	0.27
GCAI9	0.58***	0.29

注：“\*”，“\*\*”和“\*\*\*”分别表示估计参数在 10%，5%和 1%水平下显著。

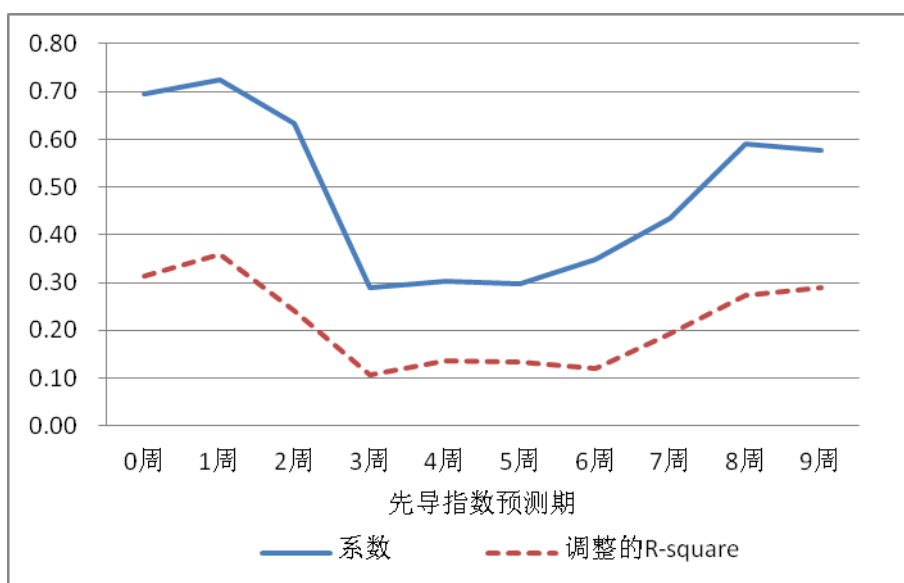


图 5：最优预测窗口

### 3.2 消费者购前网上调研行为

互联网的普及使得消费者在购买之前一般会在网上进行市场调研，而由于诸如汽车之类的大件消费品对于普通家庭来讲属于重大消费活动，因而其调研行为是更复杂的，而且是多阶段的：首先，在最开始的阶段，消费者仅仅是有买车的初步意向，消费者会先在互联网上进行一些浅层次的信息搜索和市场调研，我们称之为“意向调研”。第二，在消费者开始明确自己的需求，并进行更有针对性的网上调研，本文称之为“决策调研”。与最开始的阶段相比，这一阶段的网上调研行为更加具体，是针对某品牌、某型号车的调研和相互比较。图 5 中的“U 型”现象证实了这两阶段调研行为的存在。具体而言，GCAI8-GCAI9 表征的是第一阶段的调研行为（意向调研），GCAI-GCAI2 表征的是第二阶段的调研行为（决策调研）。

根据商品类型、消费习惯的不同、互联网普及度的差异等，两阶段网上调研行为的时间点可能会有所不同。一般而言，越是重大的商品，两阶段调研行为之间的间隔时间会越长，如汽车的间隔要大于笔记本电脑、笔记本电脑又会大于移动硬盘等。甚至我们可以猜测，一些小件电子产品的两阶段调研行为是合一的。另外，互联网越普及的市场，两阶段之间的间隔也会越小。表 4 显示，对于中国大陆市场上消费者购买汽车的调研行为，“意向调研”和“决策调研”之间大约有两个月的间隔。这两个月的时间可能是消费者在线下与经销商或有过使用经验的朋友进行沟通交流的时间，可能是融资所需的准备时间，也可能是家庭内部的讨论决策时间等等。

两阶段购前调研行为的存在表现在预测模型中就是最优预测窗口的 U 型规律——提前约 8 周的搜索量反应了“意向调研”的结果，而提前约 1 周的搜索量反应了“决策调研”的结果，他们都可以提高对市场需求的预测能力。

### 3.3 搜索量对不同类型汽车的预测效果<sup>①</sup>

以上研究说明了互联网搜索量对汽车行业整体具有预测能力。那么由于不同层次消费者的购前调研行为很可能存在差别，互联网搜索量的预测效果对于不同类别的汽车品牌有差异吗？就此，我们补充搜集了 10 个汽车品牌从 2006 年 1 月-2012 年 3 月销售量的时间序列数据。同时，参考各汽车品牌旗下的主打车型市场售价<sup>②</sup>，我们将样本中的 10 个汽车品牌分为两类：中端品牌和低端品牌。其中，低端品牌包括比亚迪和奇瑞，其他 8 个为中端品牌。

为了对不同类型汽车品牌搜索量和销量之间的关系进行区别研究，我们分别以各汽车品牌销售量同比增长率为被解释变量、以该品牌关键词在互联网上搜索量的同比增长率为解释变量进行回归，结果如表 5。可以看出：除了上海大众、北京现代、和奇瑞三个品牌之外，其它 7 个品牌的搜索量都对实际销售量有显著的预测作用。其中，预测效果最强的是起亚、雪佛兰，其次是别克和比亚迪，最后是丰田、一汽大众和东风日产。我们希望在未来的研究中对这一部分进行拓展研究。

表 5：搜索量对不同品牌汽车销量的预测效果

	汽车品牌	Beta 值	标准差
中端品牌	起亚	1.1708	0.2549***
	雪佛兰	1.1093	0.2760***

<sup>①</sup> 作者特别感谢专业主编和匿名审稿人的这一建议。

<sup>②</sup> 我们根据销量数据列出了样本期间各汽车品牌销量占前四位的车型（销量前四位的车型占品牌全部销量的比例最低为东风日产 69.22%，最高为别克 88.07%），然后参考这些车型的指导价格对汽车企业进行分类。其中，指导价格数据来源为“新浪汽车”2013 年 4 月 1 日的指导价。

	别克	0.7894	0.1169***
	丰田	0.3415	0.1903*
	东风日产	0.2872	0.0861***
	一汽大众	0.2841	0.1138**
	北京现代	0.2044	0.7414
	上海大众	-0.1276	0.1454
低端品牌	比亚迪	0.7543	0.2150***
	奇瑞	0.0570	0.1312

注：(1) “\*”，“\*\*”和“\*\*\*”分别表示估计参数在 10%，5%和 1%水平下显著。

## 4 稳健性检验

中国是一个发展中国家，市场和制度都变化很快，但不论对汽车销售市场而言还是从中国互联网的发展角度来看，2009 年都是一个重要的时间节点。首先，2008 年的全球金融危机造成了国内宏观经济的波动，给包括汽车行业的许多行业造成了结构性影响，从图 2 中可以明显看出，2008 年下半年之后汽车销售量的波动率迅速加剧。同时，出于扩大内需和节能减排的双重目的，中国政府在 2009 年 1 月开始实施汽车购置优惠政策，具体措施包括部分车型的购车补贴、小排量汽车减征车辆购置税等。这些政策短期内显著地刺激了汽车销售市场，同年 10 月，中国汽车销售量超过 1000 万辆，成为继美国和日本之后第三个汽车销量过千万的国家。另外，2009 年也是中国互联网发展进程中重要的节点，2009 年网民数量增长了近 9000 万，首次有超过四分之一的中国家庭开始使用互联网<sup>①</sup>。基于这些原因，我们以 2009 年 1 月作为断点，对其后的样本进行实证分析，对上文的主要结论进行稳健性检验。

对子样本估计得到的结果见表 6。与之前的结论一致，第 II 组模型与第一组模型中相对应模型相比，调整后的  $R^2$  更大，且均方误差更小，都说明了第 II 组模型的拟合效果要优于第 I 组模型。另外，如图 6 所示，子样本的最优预测窗口分析也发现了类似的 U 型规律。

表 6：稳健性检验-回归结果

	IA	IIA	IB	IIB	IC	IIC	ID	IID
<b>C</b>	35.26**	26.91**	-994.58	-361.96	26.91**	38.58**	-1393.45*	-916.39
	(16.87)	(12.32)	(598.15)	(655.83)	(12.32)	(17.83)	(657.82)	(573.19)
<b>Y</b>			10.00*	3.78			13.81**	9.21
			(5.84)	(6.37)			(6.52)	(5.56)
<b>GCAI(-1)</b>		0.30***		0.30***	0.30***			0.29***
		(0.08)		(0.09)	(0.08)			(0.07)
<b>AR(1)</b>	0.63***	0.74***	0.38**	0.64***	0.74***	-0.90***	0.44	-0.45**
	(0.17)	(0.16)	(0.18)	(0.18)	(0.16)	(0.22)	(0.38)	(0.21)
<b>AR(2)</b>	0.62***	0.48**	0.56**	0.47**	0.48**	0.63***	0.64*	0.31
	(0.20)	(0.19)	(0.21)	(0.19)	(0.19)	(0.16)	(0.37)	(0.23)
<b>AR(3)</b>	-0.41**	-0.41**	-0.17	-0.33*	-0.41**	0.65***	-0.23	0.44**

<sup>①</sup> 数据来源：《第 29 次中国互联网络发展状况统计报告》，中国互联网络信息中心(CNNIC)，2012 年 1 月。

	(0.17)	(0.16)	(0.19)	(0.17)	(0.16)	(0.21)	(0.36)	(0.18)
<b>MA(1)</b>						1.56***	-0.30	1.32***
						(0.22)	(0.37)	(0.27)
<b>MA(2)</b>						1.00***	-0.41	1.17***
						(0.28)	(0.34)	(0.39)
<b>MA(3)</b>						0.44**	-0.29	0.15
						(0.17)	(0.39)	(0.23)
<b>样本数</b>	38	38	38	38	38	38	38	38
<b>Adjusted R-squared</b>	0.77	0.83	0.77	0.82	0.76	0.82	0.79	0.88
<b>Sum squared resid</b>	8746.49	6332.42	8211.91	6305.16	8225.49	6065.39	6782.85	3919.99

注：（1）估计值下括号内的数为标准差；（2）“\*”，“\*\*”和“\*\*\*”分别表示估计参数在 10%，5% 和 1%水平下显著。

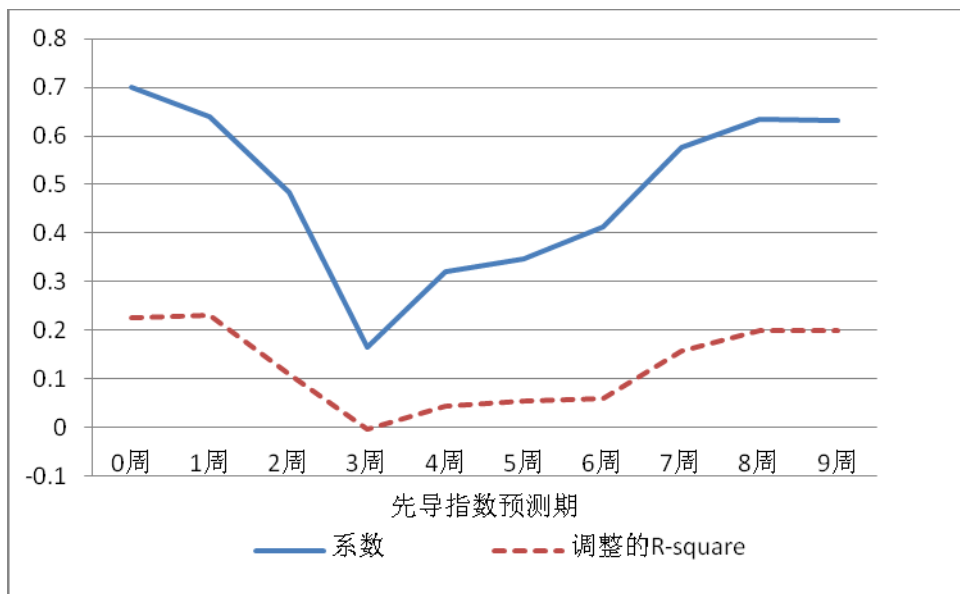


图 6：稳健性检验-最优预测窗口与 U 型规律

## 5 研究结论与启示

互联网的逐渐普及使得更多的消费者开始在网上进行购前调研行为，而搜索引擎使得捕捉和追踪网上购前调研行为成为可能。本文通过谷歌搜索引擎提供的数据，首次对中国汽车行业消费者的网上购前调研行为进行了研究。

我们首先通过搜索量频率数据设计和构造了中国汽车需求先导景气指数——“谷歌中国汽车指数 GCAI”，并通过样本内估计和样本外预测的方法证明该指数不仅能提高预测能力，为企业的预测和决策提供重要的额外信息；而且可以使企业更早地对未来需求作出预测，前移预测窗口。另外，通过对互联网搜索频率数据的研究显示，汽车行业消费者的购前网上调研行为是分阶段的，有明显的“U型”规律。

## 参考文献

- 东北财经大学宏观经济分析与预测课题组. 2006. 构建多维框架景气指数系统的初步尝试[J]. 数量经济与技术经济研究, (7): 49-57.
- 洪成一,朴宰秀,黄春华. 2003. 互联网信息搜索意图对消费者行为的影响分析[J]. 国际商务(对外经济贸易大学学报), 6: 57-61.
- 梁志峰. 2010. 基于 Google 趋势分析的区域网络关注度研究——以湘潭为例[J]. 湖南科技大学学报(社会科学版), 13 (5): 41-48.
- 廖成林,史小娜. 2012. 搜索引擎对网络购买意愿的影响研究[J]. 江苏商论, (5), 47-49.
- 刘小平. 2012. 中国汽车市场历年销量与主要宏观经济指标相关性分析[J]. 上海汽车, (5): 39-41.
- 宋双杰,曹晖,杨坤. 2011. 投资者关注与 IPO 异象——来自网络搜索量的经验证据[J]. 经济研究, 增 1 期: 145-155.
- 徐茵,王高,赵平. 2010. 顾客价值的生成与影响机制-对北京家用轿车市场的实证研究 [J]. 营销科学学报, 6(1): 1-12.
- 张崇,吕本富,彭赓,刘颖. 2012. 网络搜索数据与 CPI 的相关性研究[J]. 管理科学学报, 15(7), 50-59.
- 张莱,陈毅文. 2006. 产品类别与网上购物决策过程的关系[J]. 心理科学进展,14(3), 433-437.
- Askatas N, Zimmermann K F. 2009. Google Econometrics and Unemployment Forecasting [J]. Applied Economics Quarterly, 55 (2): 107-120.
- Baker S, Fradkin A. 2011. What Drives Job Search? Evidence from Google Search Data. SIEPR Discussion Paper No. 10-020.
- Brownstein J S, Freifeld C C, Madoff LC. 2009. Digital Disease Detection—Harnessing the Web for Public Health Surveillance[J]. New England Journal of Medicine, 360(21): 2153-2157.
- Carrière-Swallow Y, Labbé F. 2013. Nowcasting with Google Trends in an Emerging Market [J]. Journal of Forecasting, 32(4):289-298.
- Choi H, Varian H. 2012. Predicting the present with google trends[J]. Economic Record, 88(s1): 2-9.
- Cooper C P, Mallon K P, Leadbetter S, Pollack L A, Peipins L A. 2005. Cancer Internet Search Activity on a Major Search Engine, United States 2001–2003[J]. Journal of Medical Internet Research, 7(3).
- Corley C, Mikler A R, Singh K P, Cook D J. 2009. Monitoring Influenza Trends through Mining Social Media. In BIOCAMP: 340-346.
- Da Zhi, Engelberg J, Gao P. 2011. In search of attention [J]. Journal of Finance, 66(5): 1461–1499.
- D'Amuri F, Marcucci J. 2010. Google it! Forecasting the US Unemployment Rate with a Google Job Search Index, (No. 2010, 31). Nota di lavoro//Fondazione Eni Enrico Mattei: Global challenges.
- Ettredge M, Gerdes J, Karuga G. 2005. Using Web-based Search Data to Predict Macroeconomic Statistics[J]. Communications of the ACM, 48 (11), 87–92.
- Ginsberg J, Matthew M H, Patel R S, Brammer L, Smolinski M S, Brilliant L. 2008. Detecting influenza epidemics using search engine query data [J]. Nature, 457(7232): 1012–1014.
- Guzman G. 2011. Internet search behavior as an economic forecasting tool: The case of inflation expectations[J]. Journal of Economic and Social Measurement, 36(3): 119-167.
- Hulth, A, Rydevik G, Linde A. 2009. Web Queries as a Source for Syndromic Surveillance, PLoS ONE, 4(2), e4378.
- Kahneman D. 1973. Attention and Effort [M]. Englewood Cliffs. NJ: Prentice-Hall.
- Lindberg F. 2011. Nowcasting Swedish Retail Sales with Google Search Query Data (Master's thesis),



- Stockholm University.
- McLaren N, Shanbhoge R. 2011. Using Internet Search Data as Economic Indicators[J]. Bank of England Quarterly Bulletin, 2011(Q2).
- Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A J. 2009. More Diseases Tracked by Using Google Trends[J]. Emerg Infectious Diseases, 15(8): 1327–8.
- Penna N D, Huang H. 2010. Constructing Consumer Sentiment Index for U.S. Using Google Searches. University of Alberta, Department of Economics Working Papers, No.2009-26.
- Polgreen P M, Chen Y, Pennock D M, Nelson F D. 2008. Using Internet Searches for Influenza Surveillance[J]. Clinical Infectious Diseases, 47(11), 1443–8.
- Preis T, Reith D, Stanley H E. 2010, Complex Dynamics of Our Economic Life on Different Scales: Insights from Search Engine Query Data[J], Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1933), 5707-5719.
- Suhoy T. 2009, Query Indices and a 2008 Downturn: Israeli Data, Technical report, Bank of Israel.
- Valdivia A, Monge-Corella S. 2010. Diseases tracked by using Google trends[J]. Spain Emerg Infect Dis, 16(1): 168.
- Vosen S, Schmidt T. 2011. Forecasting private consumption: survey-based indicators vs. Google trends[J]. Journal of Forecasting, 30(6): 565-578.
- Wilson K, Brownstein J S. 2009. Early detection of disease outbreaks using the Internet[J]. Canadian Medical Association Journal, 180(8): 829-831.
- Wu L, Brynjolfsson E. 2009. The future of prediction: How Google searches foreshadow housing prices and sales. Available at SSRN 2022293.

**Pre-purchase Research Behaviors, Leading Climate Index and Demand Forecasting Based on  
Internet Search Data: the Case of Automobile Industry in China**

Feng Ming, Liu Chun

(School of Economics and Management, Tsinghua University)

**Abstract:** With the increasing popularity of Internet, many consumers search the information of an item on line before buying it. The search engines like Google and Baidu provide vehicle to capture and track the pre-purchase research behaviors. We design and construct a leading climate index of China's automobile demand based on the internet search volume data, namely “Google China Automobile Index”. Our index has strong forecasting power on the market demand. Meanwhile, we find that consumers’ pre-purchase research behavior has two stages, which is named as “U-shape” rule in this paper.

**Key words:** Search engine, Demand forecasting, Automobile, Pre-purchase research behavior, Leading index