# 基于大数据的税收经济分析和预测探索

**谢波峰**
中国人民大学金融与财税电子化研究所，北京 100872

**摘要**
在现有的税收经济预测理论和实践当中，大数据的出现为理论的完善和实践的发展提供了契机。总结了目前税收经济领域的主要研究进展，归纳概括了大数据在税收经济分析和预测中的特点和主要特征，并针对我国税收经济分析和预测的大数据应用提出了若干设想和展望。

**关键词**
税收经济分析；税收预测；大数据

中图分类号：F810.42 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017026

# *Exploratory research on big data application of analysis and forecasting in economics of tax*

**XIE Bofeng**
School of Finance, Renmin University of China, Beijing 100872, China

***Abstract***
Based on the current forecasting theory and practice in the economics of tax, big data applications give chance for the improvement and development of theory and practice. The main progress and characters in this field were summarized. The main progress and characters of economic analysis and forecasting with big data in tax field were summarized, and some advices about the future development were given.

***Key words***
analysis of economics of tax, forecasting of tax, big data

# 1 大数据在经济和税收领域分析和预测现状

## 1.1 国外大数据在经济研究中的典型案例

随着大数据在经济研究中关注度的不断提高，若干经典案例已经成为了大数据应用的范例，具体介绍如下。

Piketty和Saez[1]通过使用美国国内税务局（Internal Revenue Service，IRS）的长期数据，研究分配的不平等现象。近些年中国经济学家熟知的法国经济学家皮卡蒂（Piketty）以其著作《二十一世纪新资本论》成为了世界经济圈的明星，该著作的核心支撑是通过对IRS数十年的不同收入等级纳税人的大数据进行挖掘得到的，这一研究对应用大数据进行包括税收经济在内的经济研究的价值给予了充分的证明。

通过大数据计算价格指数。例如在美国，Cavallo A等人[2]推出了著名的万亿商品价格计划（billion prices projects，BPP）项目，通过50多个国家在线网站的数据，计算零售价格的变动，由此代替传统计算居民消费价格指数（consumer price index，CPI）的办法。这一类似做法在一些拥有商业网络的公司并不鲜见，例如VISA、MASTER等信用卡平台。

结合网站搜索数据进行所谓的实时预测（now-casting）。最著名的例子就是Google公司首席经济学家Varian H R等通过Google搜索指数[3]，对当前的经济情况进行预测，由于这种预测比传统的经济预测方法更为及时地反映了当期的经济情况，所以被称为“实时预测”。

研究长序列的儿童教育效果。Chetty R等人[4]使用了近20年的IRS、纽约市政教育部门的数据，研究儿童教育的长期效应，该研究指出更好的教育大约可以带来四分之一的收入增加。

除此之外，Levin J等人[5]在电子商务领域通过大数据研究消费者的行为范式，也是大数据提供的传统研究方法无法达到的研究。

对于这些大数据在经济分析中的应用，Varian H R等人[6]指出：现代经济学要与数据打交道，而传统分析用的是样本等小数据，随着经济交流的日益频繁和技术水平的提高，数据越来越大，大数据的出现使得传统经济学分析方法在分析时显得捉襟见肘。

近年来，随着机器学习热点的形成，Athey S[7]等经济学家开始进一步探索和归纳机器学习在经济分析中的应用。

## 1.2 国内大数据在经济研究中的应用现状

### 1.2.1 基于大数据的经济指数编制

（1）行业经济指数

从国内的研究来看，在Google公司等大数据应用先锋的启发下，百度、腾讯、阿里巴巴等互联网企业纷纷启动相应的基于大数据的行业经济指数。例如基于搜索情况，推出全行业、中小企业、大企业点击量景气指数等大数据指标，通过与全行业、中小企业、大企业消费量景气指标进行对比，发现对应的景气指标之间具有一定的联动性。

除了互联网企业通过互联网数据形成行业经济指数之外，拥有大数据的传统企业也推出了一些行业经济指数，例如中国银联和清华大学数据科学研究院在2015年提出的“清数银联智策指数”，是基于银联卡结算的数据形成的行业指数，中国银联

拥有超过50亿张借记卡和信用卡、超过8亿持卡人和1 500余万商户的消费数据。基于中国银联借记卡、信用卡的消费数据，推出了房地产业指数、高端酒店指数及餐饮服务指数等。

（2）新兴的统计数据

通过新的角度、新的方式，产生新的统计数据，也成为了大数据时代的发展方向，包括对新兴经济的统计、网络CPI统计以及基于在线购物平台的消费指数等，典型的代表有新经济指数（new economy index，NEI）、基于互联网在线数据的居民消费价格指数（internet-based consumer price index，iCPI）、淘宝网络零售价格指数（internet shopping price index，iSPI）等。

● NEI：针对传统经济指数形成的分析和预测方法在新经济领域中往往有些无力。最近，被李克强总理多次提及的BBD新经济指数在这一方面做出了相应的尝试。NEI指标体系首先梳理了新经济行业，包括制造业与服务业等9个类别，分别考察高端劳动投入、优质资本投入与科技与创新投入（分别代表了新经济发展的劳动力投入、资本投入、科技与创新投入）等一级指标和11个二级指标的情况。针对这些指标，需要基于大数据理念，采集大量的非传统信息。例如，根据2015年发布的信息显示，为了生成NEI指标，采集了以下信息：企业网络招聘的信息（5 200万条）、网上公示的新成立企业工商登记信息（270万条）、风险投资数据 、网上的招标投标数据（370万条）、三板上市数据、各类专利数据等。

● iCPI：2016年11月，清华大学经济研究所推出了iCPI指数，类似于上文提及的国外BPP项目，对我国的居民消费价格进行实时性的在线搜集和在线分析。

● iSPI：网络零售价格是概括网络零售交易商品一般价格水平的指标。它建立在淘宝交易平台汇聚和实时积累的海量交易行为数据基础之上。目前，淘宝网是国内最主要的网络零售交易平台，基于淘宝网的iSPI可以大体反映国内网络零售渠道的一般物价变动。

### 1.2.2 国内基于大数据的经济分析和预测研究

近年来，国内在基于大数据的经济分析和预测方面也有不少的研究成果。中国学者通过学习和总结大数据在经济应用中的发展[8]，基于中国数据，将对宏观经济的CPI和失业率等重要变量的预测，与新方法、新指标进行结合[9,10]。在实践方面，对宏观经济监测预测[11]也进行了相应的探索，并且在模型算法上进行了较新的尝试[12]。基于网络数据进行算法改造，将实时预报和短期预测进行结合[13]。

## 1.3 税收经济分析和预测

作为经济分析重要组成部分的税收经济分析，既有一般性的特点，又有相对独立的一些特色。

### 1.3.1 国外发展情况

由于管理上的优势，税收一直以来就拥有丰富的数据，以美国纳税人遵从测算项目（taxpayer compliance measurement program，TCMP）和国民研究计划（national research program，NRP）项目最为著名。这两个调查都是采取分层抽样的方式，从个人所得税申报表、公司所得税申报表和S类公司（符合特定条件的小企业）申报表中随机选取一定数量的数据，用于美国纳税人的遵从情况研究。

随着税务管理信息化的深入，税收日常管理中积累了海量数据，这些数据可

以弥补TCMP和NRP等随机样本数据的缺点，例如需要大量的人力、物力进行纳税检查，并有一定的时滞，因此有些国家税务管理当局提出了应用税收管理数据（operational data）的设想。税收管理数据不仅包括纳税人正常申报给税务管理部门的各种数据，还包括一定时期内税务部门运用各种税收检查手段，对纳税人进行检查的数据。应用税收管理数据的优点是可以在具体税务管理活动中产生大量样本，而不必专门花费大量时间和金钱进行调查。

拥有这些数据的税务管理部门早就开始探索应用。Debarr D[14]的研究提到美国国内税务局早在2004年左右，就尝试通过关系型挖掘技术发现海量纳税人数据中的遵从信息，针对高收入人群非法避税的现象，应用了支持向量机的数据挖掘方法，使用2000—2003年纳税年度的数据，该模型识别了不少高收入纳税人不遵从的税收案例，每个案例都涉及上百万美元的偷逃税额。目前，各国税务部门正在积极探索相关方面的研究，在经济合作与发展组织（Organization for Economic Cooperation and Development，OECD）的国家中，德国、荷兰、瑞典等国家的税务部门设计了网络爬虫，收集相关的电子商务交易数据，整理来自社交媒体的有用信息，包括博客、社交网络、合作性质的维基百科类/互动类论坛、简易信息聚合（really simple syndication，RSS）和小插件、微博（推特）等多种形式，用于税务管理和税收政策制定[15]。

### 1.3.2 国内现状

（1）税收分析和预测的理论研究

国内对税收收入预测的研究也有不少优秀的成果，包括对相关的税收收入能力估算的探讨、对各种预测方法的研究，不仅包括了各种传统计量模型的应用，系统动力学、神经网络等新方法的引入，还涉及税收滚动预测实践经济的总结，并且也吸取了不少国际经验。

在这些研究和实践的基础上，近年来，也开始探索大数据的应用，例如构建微观数据体系的应用[16]、在纳税评估体系建设中的应用[17]、基于大数据的算法改进[18]。总体来看，基于大数据的税收分析和预测越来越得到重视，有着广阔的发展空间。

（2）税务领域大数据的发展

值得一提的是，目前我国税务领域积累了数量可观的大数据，为进一步的研究奠定了良好的数据基础。

第一，税务管理数据。由于我国纳税人数量众多，涉税事项各种各样，在税务管理和税收政策应用中，大数据的形成已成良好态势。根据2010年前后金税三期立项需求的不完全统计，当时全国国税部门的数据量约为17 901 GB，地税部门数据量约为13 068 GB，并且还在以378 GB/月的速度递增。据了解，目前苏州市国家税务局已经建立了税务系统首个基于Hadoop的大数据体系。

自2015年4月1日起，随着增值税发票升级版的正式运行，所有存量纳税人将不断纳入发票管理系统，并且在2015年12月1日全国推广电子发票。据不完全估计，在营业税改增值税（以下简称营改增）推广前，全国增值税专用发票数量超过2.4亿张，因此可以预期，随着营改增的全面推广，将形成基于电子发票的税务管理大数据。

第二，税收调查数据。除了在税务管理信息系统中通过重点税源管理分析系统（taxation resource analysing system，TRAS）对重点税源企业进行监控之外，我国税务系统还借鉴国外的税收微观数据库建设经验。为了进一步掌握税源情况，国

家税务总局大概从2006年开始开展全国税收调查，对税收经济情况进行了深入了解。全国税收调查覆盖了国民经济所有行业的纳税人，包括优惠政策等重要事项的数据，是其他数据（包括规模以上工业企业数据或经济普查数据）等力不能及的。

## 2 基于大数据的税收经济分析和预测概况

第1节对大数据在经济和税收领域分析预测现状进行了介绍，本节将对税收经济领域大数据应用特征、研究方法和理论进一步概括。

### 2.1 更长、更宽、更活的序列

（1）变量变“长”

在前文提到的大数据典型案例中，大数据技术应用在税收经济研究中，不仅可以具有更多的观测样本，还可以回溯到更远的过去，在一个更长的时间序列中观察税收经济变量的长期变化规律，由于时间拉长，往往比较短时空中呈现的微弱变化更加明显，增强了识别能力。

（2）变量变“宽”

在前文提到的大数据应用典型案例中，Varian H R等将互联网搜索情况纳入分析预测考虑的变量，即根据使用Google浏览器进行搜索的情况，将搜索关键词分类，形成上千个关键词，对这些关键词的搜索次数进行标准化、指数化，用以研究和预测某一方面的发展变化。许多基于大数据的研究还将变量选择范围拓宽到传统领域之外，并提出了不少新颖的变量。

（3）粒度变“活”

一直以来，宏观数据与微观数据之间的衔接对应，总是不尽如人意，这一状态，大数据时代或许有所改观，正如有些研究提出的宏观、微观数据的“混搭”探索一样，在大数据的支撑下，或许可以在多个不同数据颗粒度下进行分析，获得更多的了解和发现。

### 2.2 新的研究方法

在大数据的支撑下，经济分析方法和模型将出现新的进展。以Varian H R在Google大数据上的预测应用为例，可以发现，根据预测的目的，可以将搜索指数与传统模型进行结合。例如，在对汽车销售量的预测中，通过在传统的时间序列模型中加入搜索指数，可改善模型的预测效果，如式（1）所示：

$$\lg(y_t)=2.312+0.114\cdot\lg(y_{t-1})+0.709\cdot\lg(y_{t-12})+0.006\cdot x_t^{(1)} \quad (1)$$

其中，$x_t$是加入传统时序模型的$t$期的搜索指数，$t$-1、$t$-12分别是滞后1期和滞后12期的数据，$x_t^{(1)}$是对应的时序月份中的第一周的搜索指数，$y_t$是$t$期的预测值。从预测判断的错误百分比（PE）、平均绝对错误（MAE）等指标来看，加入了Google趋势搜索指数的模型，在样本预测期内，比传统模型小3个百分点左右，可以认为对预测结果有一定的改善。

除了在数据上引入大数据的理念之外，在模型方面，现有的研究也做了积极的尝试，例如引入现有的理论方法进行大数据情况下的模型优化。在一般的模型表述形式中，如式（2）所示：

$$y_t=\alpha+\beta x_t+\varepsilon_t \quad (2)$$

其中，$y_t$是估测变量，$\varepsilon_t$是估测偏差。针对所谓的“胖数据”，向量$x_t$中维度数量远大于$t$，由于向量$\beta$大部分元素为0，所以可以通过所谓的Spike-and-Slab变量选择回归。

近年来，随着机器学习热点的形成，Athey S等人[7]进一步开始探索机器学习在经济分析中的应用。机器学习是一个广义的名词，而在狭义的定义上，机器学习则可以分为有监督学习和无监督学习，具体包括CART、Boosting、Bagging、Bootstrap、Lasso等方法。传统的计量经济学模型是指定某个模型后，通过其他备选模型来检验其顽健性。与计量经济学方法相比，许多机器学习方法是通过交叉检验选择模型的，即机器学习反复在部分数据上估计模型，在另一部分数据上检验模型，然后通过复杂性惩罚项，找到最合适的模型。这种特点被概括为所谓的施加约束性（regularization）和系统性的模型选择（systematic model selection），在更长序列、更宽变量、更多粒度选择的大数据经济分析环境下，或许将会变成经济学实证分析的标配。

另外，虽然目前普遍认为机器学习只注重基于相关关系的分析和预测，对于因果推断还很不足，但部分学者已经开始了探索，例如Kolesar M等人[19]在理论层面的分析，还有一些研究者将机器学习应用到因果推断中，通过改变目标函数，将需要估计的因果效果通过定义一个算法来优化，从而能够测量因果关系。比如，Athey S[7]将随机森林模型改造后用于估计个体处理效应，通过个体特征的模型，可以呈现某个具有特定特征的群体是如何对价格进行变化的。

### 2.3 新的理论发现

在新数据、新方法的支持下，在包括税收经济在内的经济理论方面，或许会有一些新的发现，例如争论许久的税收经济理论问题（如一致性减税好还是结构性减税好、流转税制好还是所得税制好等）可能会出现一丝曙光。另外，在预测理论上的一些难题（如是否存在能够应对突变的一般性理论模型和估计方法）也将有所突破。

在基于大数据的经济分析和预测方面，新的变量、新的方法和新的理论这3个特征之间应该存在着互相推进、齿轮联动的有机关系，在某一方面的推进，必将推动其他两个方面的进展。

## 3 基于大数据的税收分析和预测体系

在国内外的研究基础上，将基于大数据的税收分析和预测体系分解成了以下两大目标：第一，涉税大数据的获取和处理，依托大数据技术，抓取、清洗、分析税收分析和预测所需的涉税信息；第二，通过传统方法和创新方法的结合，开展税收分析和预测分析，以提高分析预测工作的科学化和智能化水平。

围绕这两大目标，构建了相应的总体方案，如图1所示。

### 3.1 税收分析和预测大数据库的形成

税收分析和预测大数据库的形成首先应该关注的问题是综合不同来源税收数据，形成微观税收大数据体系，具体而言，包括以下工作。

（1）非税务部门、非传统渠道（主要是网络分布数据）的数据收集、清洗

大数据时代税收分析和预测所需的数据不仅仅限于税务管理部门已有的信息系统获得的征管和财务数据，还需要整合更多来源的信息，包括其他政府管理部门等第三方提供的结构化数据和其他来源的数据，尤其是动态发展中的来自于信息网络互动中的涉税信息，以实现大数据时代所

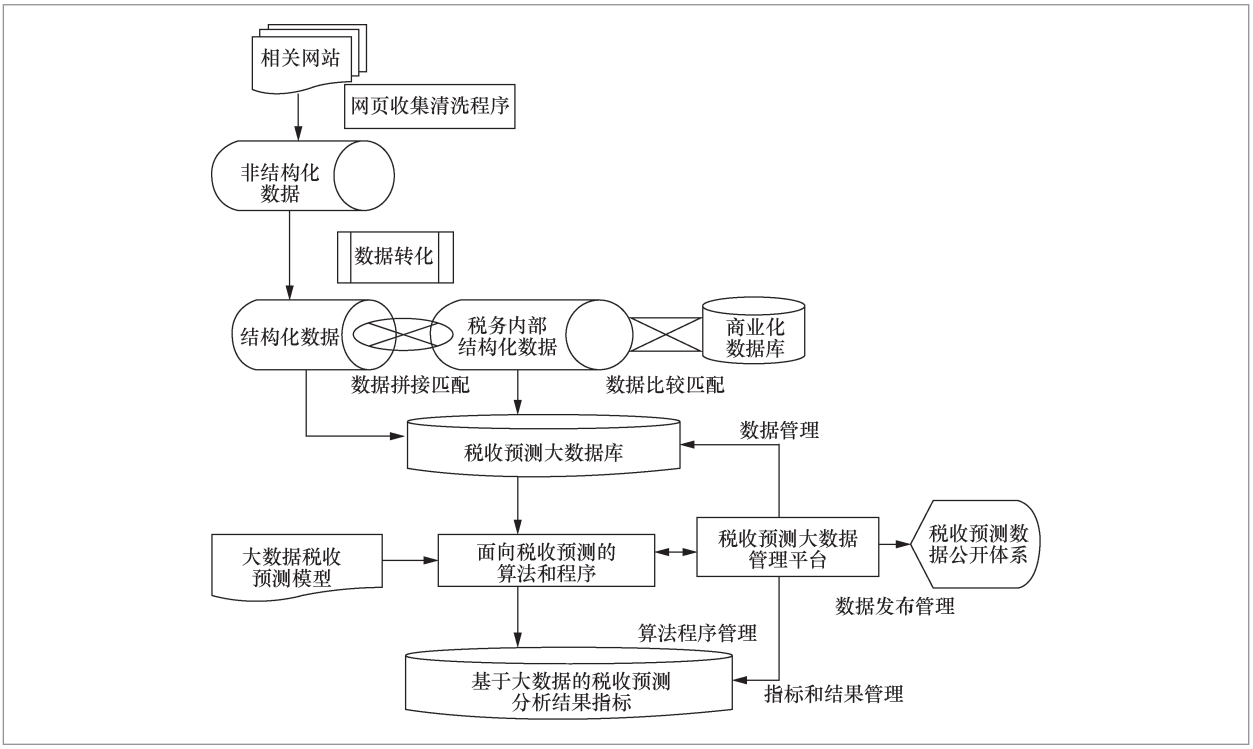


图1 基于大数据的税收分析和预测体系示意

谓“数据互联”的要求。典型的非税务部门非传统渠道数据具有分散式、非结构化等特点，对于这些数据需要考虑数据的有效收集机制，并对收集的数据进行清洗。

（2）非结构化数据的结构化转换

根据税收分析和预测的需要，按照工作中不同数据粒度的要求，对非结构化数据进行结构化转换。

（3）不同来源涉税数据的匹配

不同来源的数据必然存在不匹配问题，解决简单的不匹配问题不仅仅是统一个体在不同系统中的标识符（如税务管理中的纳税人识别号与商业数据库中企业代码不同），或者消除部分个体。实际上，解决数据匹配问题的思路应该借鉴经济研究过程中通过数据的弥补和插值来平滑缺失数据的做法，尽可能地拼接不同来源的数据，以形成相对完整的数据信息全图，并将最终呈现大数据模式下的涉税数据的完整信息链。

### 3.2 面向大数据的税收分析和预测理论及模型构建

本节将研究大数据时代数据分析的理论，根据不同市场（包括传统市场和新兴电子商务领域）的税收经济机制，研究重要的税收政策和管理问题的模型化问题。具体工作包括以下几方面。

（1）大数据的数据分析理论

针对大数据的数据分析理论需要进行创新探索，而不仅仅局限于相对模糊的智能学习方法等分析理论。基于大数据的税收分析理论的实质是多维变量空间的数据分析理论，可以简单描述如下：

$$y_i=f(x_{i1},x_{i2},x_{i3},\cdots,x_{in}) \qquad (3)$$

其中，各变量$x_1,x_2,x_3,\cdots,x_n$可以构成$n$维空间，$y$是这个$n$维空间的一个函数值，这个映射形成的多维空间样本点为（$y_i$，$x_{i1},x_{i2},x_{i3},\cdots,x_{in}$），$i$为不同的多维样本点的

个数。其中的关键科学问题是将若干个多维空间样本点($y_i$, $x_{i1}$,$x_{i2}$,$x_{i3}$,…,$x_{in}$) 构成的小样本群适用的已有传统分析理论一般化，以适用于全部多维空间样本点中，并且建立具有理论统一性的方法解释、有效的估计方法、科学的模型判断体系。

（2）面向大数据的税收分析和预测模型和算法构建

本部分的工作是应用大数据分析理论，在大数据空间中，对关键税收管理和税收政策问题选择算法和建立分析模型。针对税收管理和税收政策现有理论基础的不同情况，采用适用的算法建立模型。对于具有坚实理论基础的税收分析和预测问题，通过应用回归模型、决策树、逻辑回归、关联规则等算法建立模型。而对于理论上尚处于探索的问题，则采用聚类算法、随机森林、神经网络、深度学习等算法开展模型库建设。通过以上思路，形成面向税收分析和预测的算法和程序，结合预测大数据库，得出分析和预测模型，为应用奠定科学基础。

### 3.3 大数据环境中税收分析和预测应用

建立基于大数据的税收数据应用体系是非常必要的，通过这个体系更加能够体现和挖掘税收大数据的价值。在税收分析和预测数据管理平台的支撑下，包括以下两项主要的应用。

- 基于大数据的税收分析和预测结果指标。将分析预测模型和算法用于大数据，产生税收分析和预测结果指标。
- 税收预测结果公开体系。在大数据税收分析和预测结果指标的基础上，经过分析加工，建立反映税务系统、政府税门、社会公众等不同群体关注的税收分析和预测结果，体现不同群体的关注重点，形成分析数据公开、分析模型透明、结论可重复的税收分析和预测公开体系。

### 3.4 税收经济分析和预测的重点和难点

实现上述框架体系涉及的重点工作如下。

- 探索和形成基于大数据的若干税源变量，包括建立若干重点行业点击量景气指数、重点行业关键词指数序列、重点产品和服务点击量景气指数（可以考虑在卷烟、酒、纺织品、煤炭、原油、成品油、电力、化工产品、机械运输设备、钢坯钢材、有色金属、建材、服务皮革、电气器材、电信设备等领域及金融服务、建筑、租赁和商务服务业、住宿餐饮业、房地产业等重点行业中，选择数据基础好、可行性较强的若干行业），抓取在线交易商品和服务信息，生成涉税居民消费品价格指数。
- 基于大数据的税源数据，整合税务内部大数据，建立重点行业税收分析和预测模型，目前的设想是在国家税务总局分析预测模型基础上，加入以上构建的大数据变量，探索应用与现在大数据匹配可行的新模型和新方法，形成基于大数据的税收收入分析和预测模型。
- 推出重点行业税收预测的公开体系。在以上工作的基础上，根据税务部门、政府管理、纳税人、其他经济分析部门等不同群体需求，尝试推出部分重点行业重要指标的税收分析和预测，以利于相关部门和企业的决策。

从目前各行各业的大数据实践以及笔者的前期探索来看，以上设想工作的展开预计会存在以下难点。

- 税收分析预测和相关大数据源的确定。由于不同部门工作需求的不同，税收分析预测所需的相关税源数据可能并不存在，需要通过系统性调查研究确定税收收入分析领域与哪些互联网大数据联系比较

紧密，来源于哪些互联网数据源（搜索、微博、特定网站等），根据可行性确定相应的获取方案。

● 多源数据的融合。来源于不同部门的数据必然存在数据口径不同的问题，甚至随着时间的变化，各部门自身的数据也会发生口径差异，而税收经济分析必须融合多部门的数据，因此工作中必然碰到税务部门数据与大数据来源数据的融合问题。

● 基于大数据的模型构建。对于大数据而言，传统的税收经济分析模型已经碰到困难，探索“胖数据”等特征的大数据模型成为应用大数据发现更多税收经济规律的核心所在。

## 4 结束语

本文通过梳理大数据在经济和税收领域的国内外发展情况，概括了基于大数据的税收经济分析和预测的3个主要特征，并尝试进一步构建基于大数据的税收分析和预测框架体系，为推进该领域的发展做出初步的探索。

## 参考文献:


[1] CHETTY R, FRIEDMAN J N, FRIEDMAN J N, et al. Measuring the impacts of teachers i: evaluating bias in teacher value-added estimates[J]. The American Economic Review, 2013, 104(9): 2593-2632.

[2] CAVALLO A, RIGOBON R. The billion prices project: using online prices for measurement and research[J]. Journal of Economic Perspectives, 2016, 30(2): 151-178.

[3] VARIAN H R, CHOI H. Predicting the present with Google trends[J]. Economic Record, 2009(88): 2-9.

[4] CHETTY R, FRIEDMAN J N, FRIEDMAN J N, et al. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates[J]. The American Economic Review, 2013, 104(9): 2593-2632.

[5] EINAV L, LEVIN J. Economics in the age of big data[J]. Science, 2014, 346(6210): 1243089.

[6] VARIAN H R. Big data: new tricks for econometrics[J]. Journal of Economic Perspectives, 2014, 28(2): 3-28.

[7] ATHEY S. Machine learning and causal inference for policy evaluation[C]//The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015, Sydney, NSW, Australia. New York: ACM Press, 2015.

[8] 刘涛雄，徐晓飞. 大数据与宏观经济分析研究综述[J]. 国外理论动态，2015(1): 57-64.
LIU T X, XU X F. The research of big data and macroeconomic analysis[J]. Foreign Theoretical Trends, 2015(1): 57-64.

[9] 张崇，吕本富，彭庚，等. 网络搜索数据与CPI的相关性研究[J]. 科学管理学报，2012(7): 50-59.
ZHANG C, LV B F, PENG G, et al. A study on correlation between web search data and CPI[J]. Journal of Management Sciences in China, 2012(7): 50-59.

[10] 袁庆玉，彭赓，刘颖，等. 基于网络关键词搜索数据的汽车销量预测研究[J]. 管理学家：学术版，2011(1): 12-24.
YUAN Q Y, PENG G, LIU Y, et al. A prediction study on the amount of automotive sales based on web search data[J]. Management Scientist: Academic Edition, 2011(1): 12-24.

[11] 陈龙，王建冬，窦悦. 基于互联网大数据的宏观经济监测预测研究：理论与方法[J]. 电子政务，2016(1): 18-25.
CHEN L, WANG J D, DOU Y. The research on macroeconomic forecast based on internet big data: theory and method [J]. E-Government, 2016(1): 18-25.

[12] 彭庚，苏亚军，李娜. 失业率预测研究——基于网络搜索数据及改进的逐步回归模型[J].

现代科学管理, 2013(12): 40.
PENG G, SU Y J, LI N. The unemployment rate forecasting: the stepwise regression model based on web search data[J]. Modern Management Science, 2013(12): 40.
[13] 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(3): 4-6.
LIU H, LIU J Q. The real time and short term prediction of China's macro economy: an empirical study based on mixed data prediction model[J]. Economic Research Journal, 2011(3): 4-6.
[14] DEBARR D, HARWOOD M. Relational mining for compliance risk[R]. Washington: Internal Revenue Service, 2004.
[15] OECD. Tax administration in OECD and selected non-OECD countries: comparative information series(2010)[R]. Paris: OECD, 2011.
[16] 谢波峰. 大数据时代微观数据体系的构建[J]. 税务研究, 2015(1): 92-95.
XIE B F. The construction of micro data system in the era of big data[J]. Taxation Research, 2015(1): 92-95.
[17] 刘尚希, 孙静, 王亚军. 大数据思维在纳税评估选案建模中的应用[J]. 税务研究, 2015(10): 7-11.
LIU S X, SUN J, WANG Y J. Big data thinking in the application of tax assessment system modeling[J]. Taxation Research, 2015(10): 7-11.
[18] 孙存一, 王敏. 基于大数据算法的纳税遵从风险的行为轨迹与动机[J]. 税务研究, 2015(10): 16-20.
SUN C Y, WANG M. The behavior track and motivation of tax compliance risk based on big data algorithm[J]. Taxation Research, 2015(10): 16-20.
[19] KOLESAR M, CHETTY R, FRIEDMAN J N, et al. Identification and inference with many invalid instruments[J]. Journal of Business & Economic Statistics, 2015, 33(4): 474-484.

作者简介

**谢波峰**(1976-),男,中国人民大学财政金融学院副教授,中国人民大学金融与财税电子化研究所执行所长,国家电子商务示范城市专家委员会成员,上海市数据科学重点实验室数据科学家,近年来主要研究方向为大数据与互联网对财税金融管理的影响、基于微观数据的税收政策分析、企业税收风险管理。

**收稿日期:** 2017-02-22
**基金项目:**国家自然科学基金资助项目(No.71373267)
**Foundation Item:** The National Natural Science Foundation of China(No.71373267)