徐晓飞 著

DASHUJU SHIJIAOXIA

大数据视角下

宏观经济预测的技术与方法研究

HONGGUAN JINGJI YUCEDE
JISHU YU FANGFA YANJIU

中国财经出版传媒集团
中国财政经济出版社

本书为教育部人文社会科学研究青年基金项目“大数据视角下宏观经济预测的技术与方法研究”（项目批准号15YJC790119）的最终研究成果。本成果亦受北京语言大学校级科研项目（中央高校基本科研业务专项资金）资助，项目编号为16YBB20。

大数据视角下宏观经济预测的技术与方法研究

◎徐晓飞/著

中国财经出版传媒集团
中国财政经济出版社

图书在版编目（CIP）数据

大数据视角下宏观经济预测的技术与方法研究/徐晓飞著．—北京：中国财政经济出版社，2017.8

ISBN 978-7-5095-7496-6

Ⅰ.①大… Ⅱ.①徐… Ⅲ.①宏观经济分析-经济预测-研究 Ⅳ.①F015

中国版本图书馆 CIP 数据核字（2017）第 115464 号

责任编辑：彭 波 段 钢　　责任校对：张 凡

封面设计：王 颖

中国财政经济出版社 出版

URL：http：//www.cfeph.cn

E-mail：cfeph@cfeph.cn

社址：北京市海淀区阜成路甲 28 号 邮政编码：100142

营销中心电话：88190406 北京财经书店电话：64033436 84041336

北京财经印刷厂印刷 各地新华书店经销

787×1092 毫米 16 开 10.25 印张 200 000 字

2017 年 8 月第 1 版 2017 年 8 月北京第 1 次印刷

定价：58.00 元

ISBN 978-7-5095-7496-6

前　言

本书为教育部人文社会科学研究青年基金项目“大数据视角下宏观经济预测的技术与方法研究”（项目批准号15YJC790119）的最终研究成果。本成果亦受北京语言大学校级科研项目（中央高校基本科研业务专项资金）资助，项目编号为16YBB20。

随着互联网的普及，人类创造的在线信息总量正以空前的速度爆炸性增长，“大数据”时代已经到来。在这个时代，纷繁复杂的数据实时可得，整个社会经济产生了根本的变化，大数据对宏观经济分析和预测具有革命性的意义。就经济分析而言，大数据时代到底为我们带来什么呢？借鉴 Einav 和 Levin（2013）的概括，至少有三点是很重要的。一是大量数据的实时可得。如 Internet 上的大量信息是实时的，移动互联网和物联网的发展导致每个人随时随地都可能在制造数据。经济模型应充分利用数据的实时性，提高分析或预测的时效性。二是可得数据是海量的。正如 Mayer－Schönberger 和 Cukier（2013）所述，传统统计学处理的主要是样本，而在大数据时代，你能得到的数据可能就是总体本身，如就物价而言，电子商务网站成交的每一笔商品价格都记录在案。大数据其量之大超过一般计量经济学软件所能处理的范围，同时解释变量增加会导致高维数据中的“维数灾难”，解决这些问题需要新的分析方法和工具。而这些方法已在信息学科领域得到长足发展，现代经济模型应该充分利用大数据处理方法为经济分析服务，如机器学习、云计算等。三是数据的非结构化。数据的来

源和形式都十分多样化。如互联网信息包含文本、图片、影音等多种形式，甚至看似杂乱无章，这些信息中到底哪些包含我们所需要的信息？在大数据被广泛应用的今天，能否以及如何利用大数据对宏观经济进行预测成为经济学研究的一个新领域。本书旨在跟随大数据的时代步伐，寻求利用在线信息进行宏观经济预测的有效方法，为宏观经济政策的准确制定提供帮助。

本书首先对大数据与宏观经济预测的相关文献作了梳理。大数据在宏观经济分析应用中最活跃也是最重要的四个领域为经济数据挖掘、经济预测、经济分析技术和经济政策，针对这四个领域本书分别进行述评。尤其对大数据背景下宏观经济数据挖掘的主要信息来源和经济预测方法、将机器学习等大数据分析技术引入经济分析、利用LASSO等方法解决“维数灾难”、大数据对经济政策制定的影响等方面的现有研究进行了深入调研和评析。

本书对将大数据应用于宏观经济预测的技术与方法进行了探讨。着重针对利用大数据进行宏观经济预测的基本思路、结构化数据与非结构数据、“降维”等关键点进行探讨。通过实证检验，提出在构建基于大数据的宏观经济预测模型时，应该区分结构化数据和非结构化信息两类数据来源，并提出合理运用两类数据的“两步法”，实证表明两步法有明显的优点。研究表明，非结构化新兴大数据可以帮助预测宏观经济，但依赖适当模型选择方法。非结构化新兴大数据不是对现有统计数据的替代，而是补充。“两步法”能有效地改进预测效果，即先使用质量高的结构化数据选择初步最优预测模型，在此基础上将新兴非结构化数据加入模型中，最终确定最优模型。即“两步法”提出先穷尽使用结构化数据，再加入非结构化信息进行模型挑选，这样可以减少犯错误的概率，改进模型挑选效果。

在实证分析上，本书选取政府统计指标作为结构化数据的代表，将互联网搜索行为有关指标作为非结构化信息的代表。首先研究通过多种模型对宏观经济总量和分量进行预测。在经济总量预测时，比较

多种模型如何更好地预测 GDP。在经济分量预测时，本书的经济预测分量包括五个：消费分量（数据上以“社会消费品零售总额”计量）、投资分量（数据上以“固定资产投资完成额”计量）、出口分量（数据上以“出口总额”计量）、进口分量（数据上以“进口总额”计量）、政府财政支出（数据上以“政府财政支出额”计量）。政府统计指标数据来源于中华人民共和国统计局，以 2005～2015 年的政府统计月度数据为基础，挑选出与宏观经济紧密相关的 12 个指标，包括消费价格指数、社会消费品零售总额等，经过整理、计算生成季度数据。互联网搜索行为数据来源于 2006～2015 年的百度指数网站的相关百度搜索指数。百度搜索指数的计算是以网民在百度的搜索量为数据基础，以关键词为统计对象，分析并计算出各个关键词在百度网页搜索中搜索频次的加权和。本书共选取 85 个百度搜索指数用来衡量互联网搜索行为。将此 85 个百度搜索指数分成五类，根据和宏观经济的联系，这五类分别为消费、投资、净出口、政府购买和就业。根据网民搜索与宏观经济的关联，分别挑选和确定代表性的搜索词，搜集相应的百度搜索指数，确定每类信息搜索的词语与变量的数量。其次，同样利用两类信息对行业风险和宏观经济变量进行预测分析。研究结果表明，互联网搜索行为可以帮助预测宏观经济状况，但必须依赖适当模型选择方法。搜索行为数据不是对现有统计数据的替代而是补充。选择结构化数据与非结构化信息变量的正确方法是“两步法”。首先，仅使用政府统计信息选择初步最优预测模型；其次，将互联网搜索行为加入选择的模型中，最终确定最优模型。通过对经济总量和分量的预测模型研究，得到主要结论有：一是通过比较多个宏观经济预测模型发现，对于宏观经济预测而言，如果仅使用互联网搜索行为，预测效果并不理想；但如果在政府统计变量的基础上，增加互联网搜索行为变量则可以帮助改进预测。其背后的机理是：一方面，新兴非结构化大数据信息往往包含了大量的噪音，从信息质量而言，相对于传统的统计数据具有明显劣势，但并不构成对传

统统计数据的替代；另一方面，新兴非结构化大数据往往包括了传统统计调查数据所没有的其他信息，如最新的实时信息，因而是对统计数据的有益补充。二是合理处理两类信息的“两步法”，指在首先充分使用结构化数据挑选模型的基础上，再加入非结构化信息进行变量挑选。这背后的机理在于，“两步法”保证了先对质量更好的统计数据的充分应用，同时发挥噪音较大的在线信息的有益补充作用。如果不加区分的将两类数据放在一起降维，则更可能将有用的统计指标剔除，从而降低了预测效果。三是研究同时表明，如果方法得当，就宏观经济预测而言，充分利用非结构化信息，特别是在线信息，可以提高预测的效果。因此，今后宏观经济预测应该更充分的利用在线数据等新的信息来源，提高宏观经济预测和政策反应的时效性与准确性。两类信息综合利用与“两步法”的模型变量挑选方法不仅在宏观经济预测中有重要的应用价值，也可将其推广到诸如公共卫生、公共安全等利用大数据预测的其他方面。

本书同时对我国在大数据背景下进行宏观经济分析预测有重要的政策启示，建议政府要加大扶持力度、搭建平台、及早建立基于大数据的宏观经济分析与预测体系，提高宏观经济政策的时效性和科学性。

作　者

2017 年 3 月

目　录

第一篇　研究基础篇

第二篇　实证研究篇

第一篇
研究基础篇

大数据视角下
宏观经济预测的
技术与方法研究
Chapter 1

第1章 绪 论

第一节　研究背景与研究意义

一、研究背景

随着互联网的普及，人类创造的信息总量正以空前的速度爆炸性增长，人类社会进入了一个以“PB”[①] 为单位的数据信息新时代，人们惊呼大数据时代已经来临。大数据并非一个确切的概念。最初，这个概念是指需要处理的信息量过大，已经超出了一般电脑在处理数据时所能使用的内存量，因此工程师们必须改进处理数据的工具，例如谷歌的 MapReduce 和开源 Hadoop 平台，这些技术使得人们可处理的数据量大大增加。

目前一般认为，大数据的典型特点可以用“4V”即 Volume、Velocity、Variety 和 Value 来概括。一是数据体量巨大（Volume）。据估计，人类至今生产的所有印刷材料的数据量大概是 200PB，而历史上全人类说过的所有的话的数据量大约是 5EB（1EB = 1024PB）。当前互联网上的数据以每年 50% 左右的速度增长，目前人类 90% 以上的数据都是最近几年产生的；到 2015 年，世界上存储的数据量远远超过 1 ZB（等于 2^{70} 字节，约 10 亿 TB）。二是处理速度快（Velocity）。在如此海量数据面前，处理数据的效率就是企业的生命。社交媒介、移动设备、网上交易和网络设备更新的速度非常快，巨大数据流会导致传统数据分析的软硬件被淘汰，产生从快速生成数据中实时获取价值的专门技术和数据分析系统。三是数据类型繁多（Variety）。构成大数据的信息类型有不同来源，包括网络日志、音

① 1PB 等于 250 字节，即 1024TB，1TB 为 1024GB。

频、视频、图片、地理位置信息等。其中大概只有约10%属于结构化数据，适合整齐的进入相关数据库的行和列，其余90%是非结构化数据。四是价值密度低（Value）。价值密度的高低与数据总量大小成反比。如一部1小时视频有用数据可能仅有一两秒。如何通过强大的机器算法更迅速完成数据的价值“提纯”变得十分重要，也是数据挖掘的关键（参见 Mayer－Schönberger & Cukier，2013）。

近年来人类对大数据特别是非结构化甚至看似杂乱无章的海量数据的分析能力已大大加强，其关键是机器学习（Machine Learning）算法的迅速发展。简单地说，机器学习就是让计算机经过“训练”在输入变量和输出变量间建立起某种“最佳”的匹配关系。所谓“训练”指把输入和输出信息都已知的样本输入计算机，然后根据一定的算法，由计算机建立起由输入变量预测输出变量的方法。机器学习的主要算法包括线性模型、拓展的线性模型、决策树（Decision Tree）、支持向量机（Support Vector Machine）、人工神经网络（Artificial Neural Network）、自组织映射网络（Self－Organizing Map）、遗传算法（Genetic Algorithm），等等，并仍在蓬勃发展。机器学习已经在图像识别、语音识别、自然语言处理、智能机器人的诸多领域取得巨大成功，是当前进行数据挖掘和大数据分析的基本手段。（见 Mitchell 1997，Hastie et. al. 2009）

就经济分析而言，大数据时代到底为我们带来什么呢？借鉴 Einav 和 Levin（2013）的概括，至少有三点是很重要的。一是大量数据的实时可得。如 Internet 上的大量信息是实时的，移动互联网和物联网的发展导致每个人随时随地都可能在制造数据。经济模型应充分利用数据的实时性，提高分析或预测的时效性。二是可得数据是海量的。正如 Mayer－Schönberger 和 Cukier（2013）所述，传统统计学处理的主要是样本，而在大数据时代，你能得到的数据可能就是总体本身，例如，就物价而言，电子商务网站成交每一笔商品价格都记录在案。大数据其量之大超过一般计量经济学软件所能处理的范围，同时

解释变量增加会导致高维数据中的“维数灾难”，解决这些问题需要新的分析方法和工具。而这些方法已在信息学科领域得到长足发展，现代经济模型应该充分利用大数据处理方法为经济分析服务，如机器学习、云计算等。三是数据的非结构化。数据的来源和形式都十分多样化。如互联网信息包含文本、图片、影音等多种形式，甚至看似杂乱无章，这些信息中到底哪些包含我们所需要的信息？经济分析如何充分利用数据挖掘技术，将这些非结构化信息转化为经济模型所能利用的形式？能否建立基于大数据的宏观经济预测模型，并应用于中国经济？这些都是需要解决的问题。

二、研究意义

利用大数据能够快速、精准地对宏观经济进行预测，进而更快捷更准确地判断宏观经济的基本态势。随着当代经济的迅猛发展，对于决策者来讲需要及时考虑是否需要尽快调整政策，对于市场特别是金融市场而言希望尽快产生尽可能正确的判断和预期，这对宏观经济预测的速度和准确性提出了挑战。目前大家对宏观经济的判断依赖于各种统计调查系统发布的统计数据，如季度 GDP、CPI、投资、进出口、PMI（制造业采购经理人指数）等。除 PMI 每月初发布上月数据外，其他都有相当滞后，如 GDP 数据需要一个多月以后才能统计出上个季度数据。各国央行执行货币政策面临的最大困难之一在于宏观经济的数据滞后太多，而基于这些数据再进行货币政策调整往往不能对症下药，甚至被认为助长了宏观经济波动。大数据时代大量有关人类活动实时数据的产生，为我们更快捷的估测宏观经济变量提供了可能。例如，人们的网上搜索行为中也可能包含许多和宏观经济有关的信息，搜“招聘”的频率可能和失业人数有关，百度指数和 Google Trends 每日更新关于这些搜索行为的数据（见图 1.1，其中上部为百度指数提供的“招聘”搜索指数，下部为 GDP 季度同比增长率，可

以看出在2012年上半年明显处于“招聘”搜索高峰阶段，对应阶段GDP增长率有明显下滑）。各种电子商务网站、论坛都包含了大量和宏观经济形势相关的内容，并且是实时更新。实际上，宏观经济分析早就面临着大量实时数据，如股票、期货等金融市场的数据是实时的，银行间拆借市场利率每天发布。但更新越快意味着数据量越大，面对大数据，宏观经济分析依靠传统方法无法收集和挖掘这些数据，传统计量经济学工具也无法处理海量数据。正因为这样，过去方法无法做到对宏观经济进行预测，经济学家们往往更关注对宏观经济变量未来走势的预测，如预测下一季度GDP等，并发展出了大量预测方法（Clements & Hendry，2011对此作了全面介绍）。大数据技术与方法的发展则为进行宏观经济预测提供了新的条件。

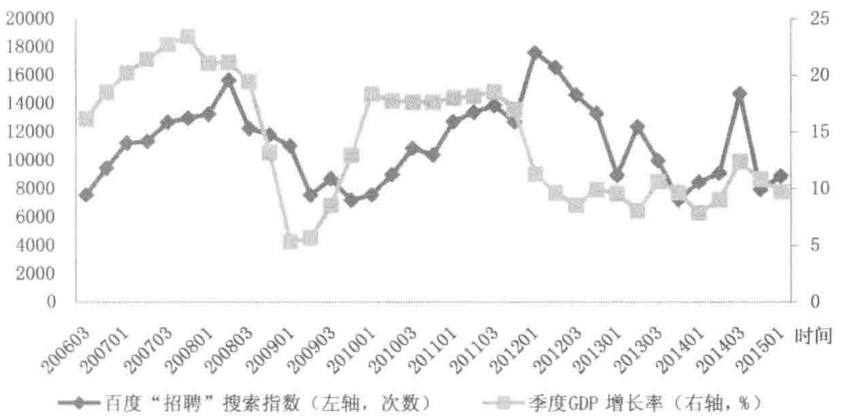

图1.1 百度“招聘”搜索指数与季度GDP增长率

国际上利用大数据对宏观经济预测方面的研究尚处于萌芽状态。代表性的是Google首席经济学家斯坦福大学教授Varian最近的一些工作。Choi和Varian（2009a，2012）介绍google trends如何可以用来为预测当前经济变量服务，Choi和Varian（2009b）描述了如何用搜索引擎的数据来预测领取失业保险的有关情况。但这些研究只是局限于google trends数据的利用，而google trends数据在浩瀚互联网数据中无异于九牛一毛。MIT的BPP项目等研究也还只是针对经济的局