

全球化国家科技创新能力综合评价指数统计模型*

——基于互联网科技统计视阈

程 豪

内容摘要：本文从互联网科技统计思维出发，旨在构建全球化背景下国家科技创新能力综合评价指数。在比较研究现有国家或地区科技创新能力相关指标体系的基础上，提出国家科技创新能力体系，并借助统计软件编程技术，实现网络爬虫技术、二阶因子模型、偏最小二乘估计方法和可视化展示的一体化操作，完成数据收集、模型估计和结果分析。研究表明，全球化国家科技创新能力综合评价指数统计模型能够客观反映各国科技创新能力及其关键影响因素，实现不同国家科技创新能力的横向比较和纵向动态分析，为统计指标研究、综合评价指数构建等科技统计领域提供一定借鉴。

关键词：互联网科技统计；国家科技创新能力；综合评价指数

中图分类号：C813 **文献标识码：**A **文章编号：**1004-7794(2020)06-0025-07

DOI: 10.13778/j.cnki.11-3705/c.2020.06.004

一、研究背景

（一）互联网科技统计的思维与概念

随着互联网技术的发展，人们的生产生活方式、社会的组织机构和内在关系均发生翻天覆地的变化。与此同时，数据生成、获取、储存、处理技术正在发生日新月异的变化。在互联网背景下，统计思维在数据处理和分析中发挥着重要作用。

作为统计及其交叉学科的重要领域之一，科技统计一直在国家发展、社会变革乃至全球创新发挥着重要的作用。在互联网背景下，科技统计思维与统计思维的主要区别在于数据对象和数据问题的不同。钱力等（2019）指出，科技数据不同于传统论文数据，也不同于一般意义上的网络及行业数据^[1]。它包括科技活动数据、科技成果数据以及互联网媒体科技资讯数据。（1）科技活

动数据包括科技实体数据与知识关系数据。其中，科技实体数据包括科技项目、学术会议、科技人才、科技机构、研究方法等，知识关系数据包括语义关系及计量关系等。（2）科技成果数据包括各分支学科内通过数据、资料、文献、报告、报道等形式承载的知识和数据。（3）互联网媒体科技资讯数据，特别是及时发布、频繁互动的APP数据。根据赵彦云（2016）对互联网统计概念的界定，本文将互联网科技统计定义为：面向统计方法和科技数据，利用互联网思想支撑科技统计领域的统计理论方法，是实现全球科技互联、互通、互动一体化量化的根本手段和科学方法^[2]。

综上所述，本文利用互联网科技统计思维，从基于数据技术的互联科技统计、基于数据社会标准的互通科技统计、基于数据主体的互动科技统计出

*基金项目：中国人民大学科学研究基金（中央高校基本科研业务费专项资金资助）项目“分位回归中缺失数据插补方法研究”（项目编号：16XNH102）。

发,编制全球化国家科技创新能力综合评价指数,构建统计模型并实现计算过程,实现了从传统小数据时代到DT(Data Technology)时代过渡进程中统计思维方式、理论方法和分析技术的重要突破。

(二) 现有国家科技创新能力相关指标体系比较研究

20世纪80年代中后期,弗里曼、伦德瓦尔等经济学家分别提出“国家创新体系”的不同概念。参考蔡跃洲(2012)的研究成果,“国家创新体系”可概括为一个国家内部创新活动相关主体以及相关的各种制度、政策,在推动创新活动过程中相互作用形成的网络体系^[3]。随着时间的推移,全球在推动国家创新能力建设过程中,逐步形成一系列关于国家创新能力测评的指标体系。其中,在全球范围具有一定权威性且引起广泛共识的国家创新能力评价指标体系主要包括以下8种:世界竞争力年鉴、欧洲创新联盟记分牌、全球创新记分牌、世界经济论坛创新能力指数、世界银行知识经济指数、全球创新指数、国家创新能力报

告和国家创新指数报告。具体内容如表1所示。

王智慧等(2015)指出上述部分指标体系存在的不足^[4]。比如,欧洲创新联盟记分牌没有很好的区分驱动要素与表征要素,投入指标可能与产出指标之间存在一定程度上的重复计算,欧洲创新联盟记分牌中指标多为均量指标,主要考虑比重,缺少绝对量的比较。世界经济论坛创新能力指数对创新产出的考察缺少对企业的考察。世界银行知识经济指数未将R&D经费纳入测量指标中。国家创新指数报告的创新环境相关指标大多需要通过主观性指标来测量,这一方面会造成某些主观性偏差,同时也使得数据收集的难度增加。

经过对上述指标体系中共性部分的研究,以避免上述不足和局限为原则,本文旨在从全球化视阈出发,构建国家科技创新能力指标体系。

二、国家科技创新能力指标体系的构建

(一) 构建依据

经研究发现,当前评价国家创新能力的主要指标体系中共同出现的共性指标仅涉及科技人力

表1 主要国家创新能力指标体系

名称	来源	首发年份	概述
世界竞争力年鉴	瑞士洛桑国际管理发展学院	1989	瑞士洛桑国际管理发展学院通过统计数据(硬指标)以及针对企业管理人员的问卷调查结果(软指标)对不同经济体的国家竞争力进行评价和排序,共采用了超过300个竞争力指标,指标体系包括3个层级。
欧洲创新联盟记分牌	欧盟创新政策研究中心	2000	根据欧盟委员会的要求,2000年制定了《欧洲创新记分牌》,旨在为监测欧盟高技术创新和技术进步方面提供可操作的指标体系。经2000年至2009年不断完善,于2010年更名为欧洲创新联盟记分牌。该指标体系有3个领域8个大项25个指标组成。
全球创新记分牌	欧盟	2006	全球创新记分牌的选择样本超出欧洲范围,因跨国数据收集比较困难,仅用了24个指标中的12个二级指标,从创新驱动、知识创造、扩散、应用和知识产权5个方面对国家创新能力进行评估,2008年进一步修订将5个支柱因素合并为3个:公司活力和产出、人力资源以及基础设施和吸收能力。
世界经济论坛创新能力指数	世界经济论坛	2010	该指数认为创新是未来经济发展的主要推动力量,与传统的生产资料投入和组织制度均紧密的关联。指标体系中共有5个一级指标(制度环境、人力资本和社会参与、管制和法律框架、R&D情况、信息通信技术的应用)、12个二级指标和61个三级指标构成。
世界银行知识经济指数	世界银行	2012 (非首发时间)	致力于帮助样本国家建立利用知识经济增强国家竞争力、促进经济和国民财富增长的实施战略,使样本国家可在世界知识经济中处于有利地位。指标体系包括经济激励与政治体制、教育与人力资源、信息基础设施和创新系统共4个维度,每个维度各对应3个二级指标。
全球创新指数	欧洲工商管理学院	2007	编制的全球创新指数始于3个重要的动机:创新在增长战略中处于核心地位;创新的范围更加广泛;创新者和新生代企业家备受鼓舞。2012年形成概念框架,其中包括5个投入指数和2个产出指数共84个变量。
国家创新能力报告	中国科协发展研究中心	2009	包括3个维度、7大要素、21个三级指标对国家创新能力进行评价。创新投入要素维度包括物质资本、科技人力资源和政策环境;创新产出包括知识产权和应用绩效;创新潜能包括战略储备和发展趋势。
国家创新指数报告	中国科学技术发展战略研究院	2013 (非首发时间)	在国内外关于国家竞争力和创新评价的研究基础上建立的评价指标体系,从创新资源、企业创新、知识创造、创新绩效和创新环境5个维度,30个二级指标进行的评价。

资源(含教育发展)、科研经费投入(含国家和企业层面)、专利申请和转让、科技论文发表4个方面。为了避免现有国家创新能力评价体系存在的不足,兼顾跨国数据收集的难度,保证指标体系在全球绝大多数国家和地区的适用性,本文在保留科技人力资源、科研经费投入两个维度基础上,将专利申请和转让与科技论文发表合并为科技成果产出,引入经济社会环境,并深入考虑和选择支撑上述4个维度的指标。具体原则和依据包括:(1)保证尽可能多的国家或地区具备该指标数据;(2)指标数量尽可能少,所带信息避免冗余重复;(3)充分考虑现有国家创新能力评价指标体系共有的指标;(4)考虑全球跨国数据收集难度;(5)保证该指标体系适用性广泛,简洁易懂且足以说明问题。

(二) 指标体系

基于现有指标体系及已有相关研究成果,兼顾全球各国相关指标数据资源,国家科技创新能力(National Science and Technology Innovation Capacity, NSTIC)指标主要测评的是经济社会环境(Economic and Social Environment, ESE)、科技人力资源(Scientific and Technological Human Resources, STHR)、科研经费投入(Research Funding Investment, RFI)和科技成果产出(Output of Scientific and Technological Achievements, OSTA)4个维度。其中,经济社会环境主要从人口总数和GDP总量两个指标进行测度。科技人力资源要涵盖对教育发展方面的测度指标,主要从R&D研究人员和高等院校入学率两个指标进行测度。科研经费投入既要体现国家整体层面的经费支出情况,也要包括企业在其中的表现。因此,科研经费投入主要从研发支出占GDP的比例和企业研发占总研发支出比重两个指标进行测度,科技成果产出主要从科技期刊文章和专利申请量两个指标进行测度。综上所述,初步形成国家科技创新能力三级指标体系,指标体系结构及符号说明见表2。

表2 指标体系结构及符号说明

一级指标	符号	二级指标	符号	三级指标	符号
国家 科技 创新 能力	NSTIC	经济社会环境	ESE	人口总数 GDP 总量	ESE1 ESE2
		科技人力资源	STHR	R&D 研究人员 高等院校入学率	STHR1 STHR2
		科研经费投入	RFI	研发支出占 GDP 的比例 企业研发占总研发支出比重	RFI1 RFI2
		科技成果产出	OSTA	科技期刊文章 专利申请量	OSTA1 OSTA2

三、国家科技创新能力综合评价指标的编制

(一) 网络爬虫技术获取数据

1. 网络爬虫的一般流程。

在互联网背景下,网络成为大量信息的载体,如何及时有效地提取并利用这些信息成为巨大的挑战。此时,人工填录数据的方式已无法满足数据体量巨大、获取要求时效的需求。孙易冰等(2014)指出,网络爬虫技术从指定的初始网页出发,通过分析网页内的标记结构,获取指向其他页面的超级链接,然后通过既定的搜索策略选择下一个要访问的站点,以获取数据^[5]。

2. 国家科技创新能力评价指标数据爬取。

国家科技创新能力指标数据主要来自世界银行指标数据^①和全球创新指数 2017 报告数据^②。通常,研究人员会采用人工下载和采集的方式进行摘录和整理,需要耗费较大的人力和物力。通过 R 或 Python 等统计软件则可以实现数据爬取操作,提高操作效率。根据表 2 中 8 个指标的数据来源和媒介,这里主要借助网页数据文件爬取和 PDF 数据读取两部分操作。具体来说,来自世界银行的指标数据需要借助统计软件爬取网页中的数据文件,而来自全球创新指数 2017 的报告数据需要借助网页 PDF 数据读取相关操作。

需要说明的是,不同数据库涉及的国家 and 地区名称及代码不尽相同,在处理时需要注意统一。本文以世界银行中给定的国家和地区名称代码为参照,进行统一。鉴于不同变量数据量级差异较大,且存在部分数据缺失的现象,因此需要考虑两方面

①世界银行指标数据网址: <https://data.worldbank.org.cn/indicator>。

②全球创新指数 2017 报告数据网址: https://www.wipo.int/global_innovation_index/en/。

的数据预处理问题:数据标准化问题和缺失数据处理问题。经考虑,这里选择最常见的最小值-最大值标准化法 $((x-\min)/(\max-\min))$ 和均值插补法。

(二) 统计模型与估计方法

1. 以二阶因子模型为基础模型。

根据国家科技创新能力指标体系,选择二阶因子模型计算国家科技创新能力指数。显然,在该模型背景下,8个三级指标为可测变量,3个二级指标和1个一级指标为潜变量。其中,由于二级指标可以看作一级指标的3个方面,因此二级指标称为一阶因子,一级指标称为二阶因子。综上所述,国家科技创新能力综合评价指数二阶因子模型的表达形式如下:

$$x_{ih} = \lambda_{ih}\xi_i + \varepsilon_{ih} \quad (1)$$

$$\xi_i = \beta_i\eta + \delta_i \quad (2)$$

式(1)为测量模型,它反映的是可测变量 x_{ih} (8个三级指标)与一阶因子 ξ_i (3个二级指标)间的关系。 λ_{ih} 是载荷系数,表示一阶因子 ξ_i 对可测变量 x_{ih} 的影响。 ε_{ih} 为第 i 个一阶因子 ξ_i 中第 h 个可测变量 x_{ih} 的测量误差,均值为0,方差为 δ_{ih}^2 ,且与一阶因子 ξ_i 不相关。

式(2)为结构模型,它反映的是一阶因子 ξ_i (3个二级指标)与二阶因子 η (1个一级指标)间的关系。 β_i 是路径系数,表示二阶因子 η 对一阶因子 ξ_i 的影响。 δ_i 为第 i 个一阶因子 ξ_i 的测量误差,均值为0,方差为 δ_i^2 。

2. 不同偏最小二乘估计方法。

考虑到指标数据的真实分布情况,国家科技创新能力综合评价指数二阶因子模型的估计方法采用偏最小二乘估计方法,对数据分布没有特别严格的要求。根据 Tenenhaus et al(2005), Leguina

et al(2015)和 Cheng et al(2018)的研究,该估计方法的迭代过程如下^[6-8]。

第一步,标准化潜变量的外部估计,即利用可测变量的线性组合对潜变量进行逼近。

第二步,标准化潜变量的内部估计,即利用潜变量间的某种数学关系,对潜变量的外部估计值进行调整的过程。

第三步,权重估计。显变量权重的确定需要区分两种类型的测量模型:反映型模型和构成型模型。模型中,各个可测变量都反映对应的潜变量,即通过潜变量指向可测变量,这类测量模型称为反映型模型。反映型模型权重估计为: $\omega = \text{Cov}(X, Z) / \text{Var}(Z)$ 。但是,有些情况可测变量并不总是反映潜变量,即通过可测变量指向潜变量,称为构成型模型。构成型模型权重估计为: $\omega = (X'X)^{-1}X'Z$ 。其中, X 是可测变量, Z 是潜变量的内部估计。

初始的权重可以任意赋值,然后进行上述迭代计算,直到收敛为止。常用的收敛判断标准为:相邻两次的权重估计值相差小于 10^{-5} 。

对于二阶因子模型,常用的偏最小二乘估计方法包括重复指标法、两步法和混合法。这3种方法的具体解释与内容比较见表3。

四、国家科技创新能力综合评价指数统计模型估计结果解析

(一) 基于参数估计结果的国家科技创新能力影响因素分析

在避免指标变量间相互独立的强假定条件下,偏最小二乘-二阶因子模型充分挖掘和利用了变量间客观存在的关系。实际上,在外部估计过程中,偏最小二乘估计算法包括可测变量与一阶因子、一阶因子与二阶因子间相关系数的计算部分,通过

表3 不同偏最小二乘估计方法的解释与比较

序号	方法名称	可测变量的分配	适用情况	不足
方法1	重复指标法	将分配到低阶潜在变量上的所有可测变量,再全部重新分配到高阶潜在变量上。	要求所有可测指标都是反映型	可能导致参数估计的有偏性
方法2	两步法	第1步:计算各个低阶潜在变量所含的可测变量的主成分,作为低阶潜在变量的取值。 第2步:根据结构模型和低阶潜在变量的取值,完成估计。	可以处理可测变量是影响型的情况	在计算低阶潜在变量的得分时,没有考虑高阶潜在变量
方法3	混合法	随机的将一半可测变量分配给它们所反映的低阶潜在变量,其余的分配给高阶潜在变量。	要求所有可测指标都是反映型	分配的随机性会导致不同的估计结果

相关系数的计算及符号的判断,确定每一步迭代中的一阶因子和二阶因子的外部估计。经过不断的内部调整和外部调整,更新内生权重和外生权重,反复迭代,得到收敛的数值计算结果。借助上述3种偏最小二乘估计方法,国家科技创新能力综合评价指数统计模型的估计结果如表4所示。

需要说明的是,表4中,载荷系数一栏方法3有3条路径($ESE \rightarrow ESE1$ 、 $STHR \rightarrow STHR1$ 和 $OSTA \rightarrow OSTA1$)对应的系数没有取值,这是因为方法3随机将 $ESE1$ 、 $STHR1$ 和 $OSTA1$ 这3个指标分配给二阶因子 $NSTIC$,所以它们没有建议与一阶因子的直接关系, $NSTIC \rightarrow ESE1$ 、 $NSTIC \rightarrow STHR1$ 和 $NSTIC \rightarrow OSTA1$ 对应的系数估计结果分别为0.41,0.82和0.67。

不难发现,上述3种方法的估计结果表现出较强的一致性。根据上述3种方法估计的路径系数可知,对于国家科技创新能力,科研经费投入的影响均表现为最大(分别为0.94,0.82和0.72),科技人力资源和科技成果产出影响居中,经济社会环境的影响最小(0.33,0.59和0.12)。根据方法1和方法2中载荷系数的估计结果可知,按照可测指标的效应大小由高到低排序,经济社会环境中的可测变量分别为人口总数(0.98和0.90)和GDP总量(0.28和0.53),科技人力资源中的可测变量分别为R&D研究人员(0.90和0.87)和高等院校入学率(0.75和0.79),科研经费投入中的可测变量分别为企业研发占总研发支出比重(0.95和0.88)和研发支出占GDP的比例(0.78和0.87),科技成果产出中的可测变量分别为科技期刊文章(0.99和0.96)和专利申请量(0.89

和0.94)。由此说明,个人总数对经济社会环境的影响大于GDP,R&D研究人员对科技人力资源的影响大于高等院校入学率。企业研发占总研发支出比重对科研经费投入的影响大于研发支出占GDP的比例。科技期刊文章对科技成果产出的影响大于专利申请量。

(二) 基于因子得分的国家科技创新能力排序

作为国家科技创新能力综合评价指数统计模型的主要估计结果,因子得分包括二阶因子得分和一阶因子得分两部分。其中,二阶因子得分可理解为综合评价指数的取值,用于评价国家科技创新能力顺序。下面以二阶因子得分估计结果为例,分别给出3种估计方法下国家科技创新能力综合评价指数排序结果,如图1至图3所示(国家或地区名称代码参照世界银行给定代码)。实际上,因子得分的绝对数值实际意义不大,而由其决定的不同国家或地区在该指标体系上的相对位置或相对段位才具有更大的参考价值。因此,图1至图3将全部国家或地区按照四分位法分成4段,并且给出4段内排序情况。

五、构建全球化国家科技创新能力综合评价指数统计模型的意义

(一) 实现互联网科技统计视阈下指数构建的一体化操作

从信息技术到数据技术的发展过程反映出从信息化到互联网再到物联网的第四次工业革命发展趋势。在科技统计领域,指标体系研究和评价指标构建是非常重要的研究方向。对于国家科技创新能力评价指标体系,存在指标数量众多、人

表4 国家科技创新能力综合评价指数统计模型参数估计结果

路径	路径系数			路径	载荷系数		
	方法1	方法2	方法3		方法1	方法2	方法3
$NSTIC \rightarrow ESE$	0.33	0.59	0.12	$ESE \rightarrow ESE1$	0.98	0.90	-
				$ESE \rightarrow ESE2$	0.28	0.53	1.00
$NSTIC \rightarrow STHR$	0.77	0.61	0.36	$STHR \rightarrow STHR1$	0.90	0.87	-
				$STHR \rightarrow STHR2$	0.75	0.79	1.00
$NSTIC \rightarrow RFI$	0.94	0.82	0.72	$RFI \rightarrow RFI1$	0.78	0.87	0.82
				$RFI \rightarrow RFI2$	0.95	0.88	0.92
$NSTIC \rightarrow OSTA$	0.59	0.76	0.49	$OSTA \rightarrow OSTA1$	0.99	0.96	-
				$OSTA \rightarrow OSTA2$	0.89	0.94	1.00

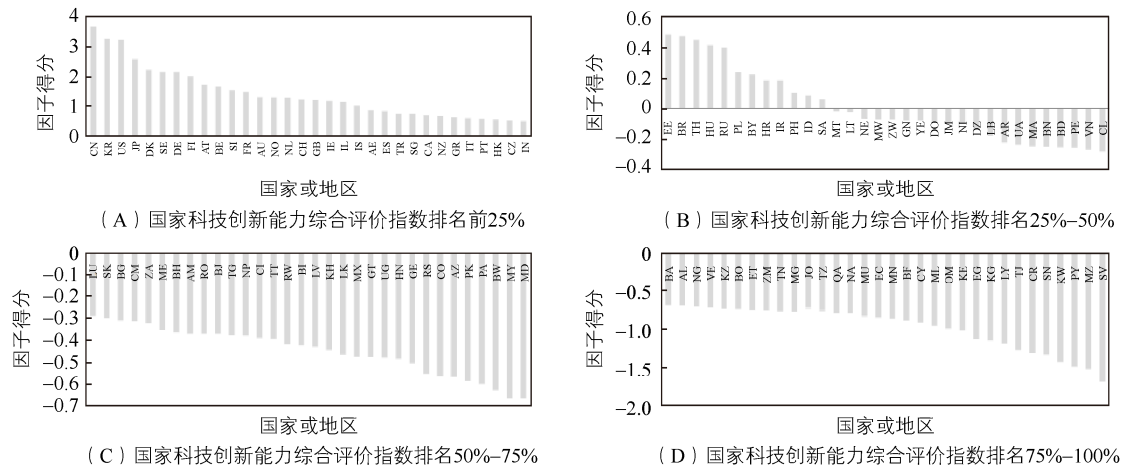


图1 基于方法1的国家科技创新能力综合评价指数计算结果

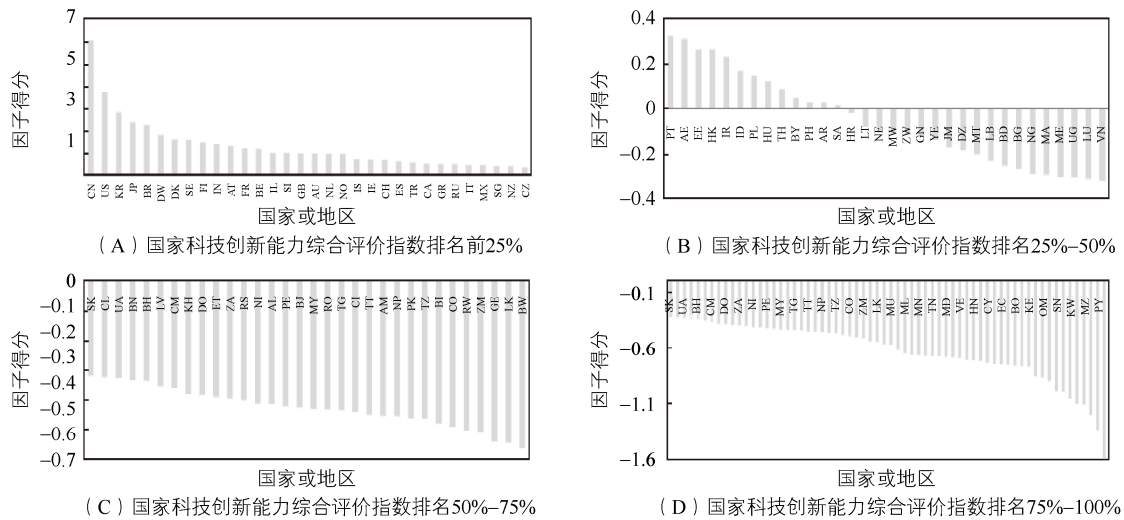


图2 基于方法2的国家科技创新能力综合评价指数计算结果

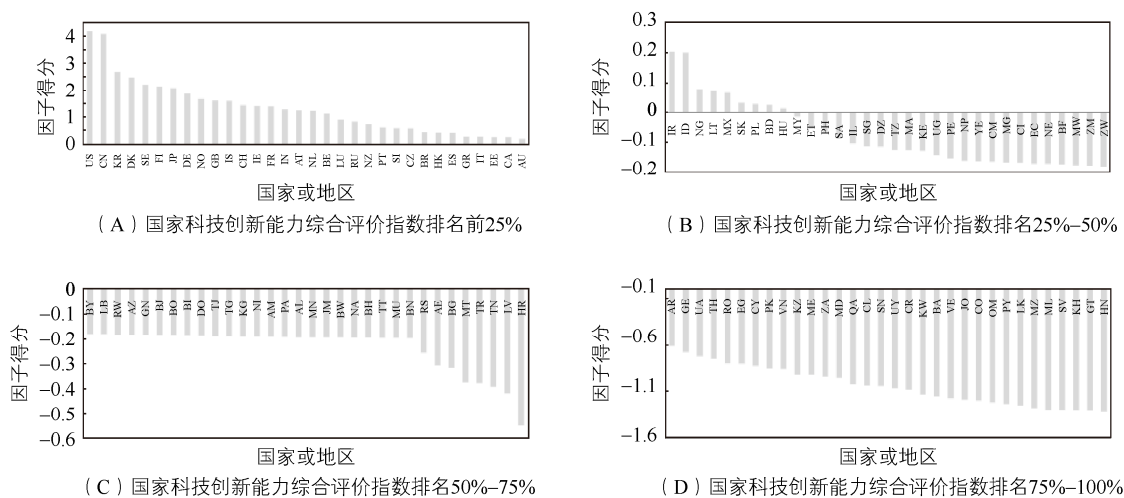


图3 基于方法3的国家科技创新能力综合评价指数计算结果

工操作过多、各环节衔接不够、方法技术落后等诸多不足。(1) 评价国家科技创新能力并不意味着指标数量越多越好。通常,过多的指标和过于庞杂的指标体系往往导致信息冗余重复,适用国家范围受限,工作量繁重等不足。

(2) 人工采集数据增加工作负担,降低工作效率且容易出现数据错误。本文通过网络爬虫技术获取指标数据,符合大数据时代的发展需求。

(3) 当前指标体系相关研究工作存在各个环节衔接紧密度不够,有待科研人员提高专业技术能力,实现从数据到指数构建再到统计模型估计完整过程。(4) 在指数计算和模型构建方面,当前绝大多数研究局限在层次分析法等由专家意见决定的主观赋权方法,缺乏新方法技术的支持。综上所述,本文从互联网科技统计思维出发,借助编程技术,实现网络爬虫技术、二阶因子模型和可视化结果的一体化操作,有效避免上述有待改进的不足和局限。

(二) 能够客观反映各国科技创新能力及其关键影响因素

在全球化视阈下,国家科技创新能力综合评价指数依托客观指标数据和科学的方法工具量化不同国家或地区科技创新能力。指标数据来自世界银行和全球创新指数报告的公开数据,数据来源具有一定的可靠性。指数统计模型清晰地测度因子间、因子与可测变量间的关系,同时,借助偏最小二乘算法完成估计,所有的数值计算结果完全依托于数据本身。而且,该方法本身具有的避免数据分布假定、采用客观权重、兼顾因素之间关系等优点,保证国家科技创新能力综合评价指数统计模型的计算可以得到基于客观指标数据的国家科技创新能力综合评价指数取值、二阶因子与一阶因子间路径系数以及一阶因子和可测变量间的载荷系数。

(三) 便于不同国家或地区科技创新能力的横向比较和纵向动态分析

对于不同国家或地区,国家科技创新能力综合评价指数基于同一套指标体系,采用同样的方法工具完成估计,因此可以比较各国科技创新能

力和影响因素。为考察不同区域(比如洲)或不同国家的国家科技创新能力差异,可以仿照全球化国家科技创新能力综合评价指数统计模型,构建不同区域的科技创新能力综合评价指数,横向比较不同区域科技创新能力。此外,可以持续进行不同年份下国家科技创新能力综合评价工作,完成纵向动态分析,以便及时发掘、追踪国家科技创新能力及其影响因素的变化情况,为相关部门及时制定调整措施、科技事业持续健康发展提供保证。

参考文献

- [1] 钱力, 谢靖, 常志军, 等. 基于科技大数据的智能知识服务体系研究设计[J]. 数据分析与知识发现, 2019(1): 4-14.
- [2] 赵彦云. 互联网统计研究[J]. 统计研究, 2016, 33(12): 4-10.
- [3] 蔡跃洲. 国家创新体系视角下的国家创新能力测度述评——理论方法、数据基础及中国实践[J]. 求是学刊, 2012, 39(1): 42-50.
- [4] 王智慧, 刘莉. 国家创新能力评价指标比较分析[J]. 科研管理, 2015, 36: 162-168.
- [5] 孙易冰, 赵子东, 刘洪波. 一种基于网络爬虫技术的价格指数计算模型[J]. 统计研究, 2014, 31(10): 74-80.
- [6] Tenenhaus M, Vinzi V E, Chatelin Y M, et al. PLS Path Modeling[J]. Computational Statistics & Data Analysis, 2005, 48(1): 159-205.
- [7] Leguina, Adrian. A Primer on Partial Least Squares Structural Equation Modeling (PLS-sem)[J]. International Journal of Research & Method in Education, 2015, 38(2): 220-221.
- [8] Cheng H, Yi D H, Si J S, et al. Establishment of Comprehensive Indicators in TCM Pectoral-Qi Case Report Based on Experts Diagnosis and Self-Test Technology[J]. Medicine, 2018, 97(7): e9916.

作者简介:

程豪, 男, 1989 年生, 山西人, 2017 年毕业于中国人民大学统计学院, 获经济学博士学位, 现为中国科协创新战略研究院助理研究员, 研究方向为科技统计指标综合评价研究、调查数据挖掘与统计机器学习。

(责任编辑: 黄煌)