计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):214−219 [doi: 10.15888/j.cnki.csa.006406]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

全国信息化水平测度指标体系修正与分析①

刘一粟[1], 沙晋明[1], 金　彪[1,2]

[1](福建师范大学 地理科学学院, 福州 350007)
[2](福建师范大学 软件学院, 福州 350108)

摘　要: 信息化水平测度是一项复杂、技术含量颇高的工作, 主要问题集中在信息化水平测度指标体系指标选取与权重确定两方面. 本文立足于现有信息化水平测度研究, 综合地理学与统计学等方面的知识, 通过词云分析、变异系数与相关系数、因子分析与客观赋权法等技术手段, 在对现有指标体系进行修正的基础上建立一套科学合理的信息化水平测度指标体系, 并对其结果进行了拟合分析, 说明了该指标体系的合理性与科学性.

关键词: 信息化; 词云分析; 相关系数; 客观赋权法; 拟合

引用格式: 刘一粟,沙晋明,金彪.全国信息化水平测度指标体系修正与分析.计算机系统应用,2018,27(6):214–219. http://www.c-s-a.org.cn/1003-3254/6406.html

Revision and Analysis of National Informatization Level Measurement Index System

LIU Yi-Su[1], SHA Jin-Ming[1], JIN Biao[1,2]

[1](College of Geographical Sciences, Fujian Normal University, Fuzhou 350007, China)
[2](Faculty of Software, Fujian Normal University, Fuzhou 350108, China)

Abstract: Estimating the level of informatization is a kind of complex, high technical content work, the problems mainly are the selection of indicators and the determination of weights in the informatization level measurement indicator system. Based on the existing informatization level measurement research, comprehensive geography and statistics, and other aspects of knowledge, through the word cloud analysis, coefficient of variation and correlation coefficient, factor analysis and objective weighting, and other technical means, a set of scientific and reasonable informatization level measure index system is established on the basis of amending the existing index system. The result is fitted and analyzed, which shows that the index system is reasonable and scientific.

Key words: informatization; word cloud analysis; coefficient of variation; objective weighting method; fitting

1　引言

21 世纪以来, 新一代信息化测度持续发展, 其理论与实践都有了新的进步. 现有信息化水平测度体系多为多级指标, 指标选取与各级指标权重的确定是一个不可忽视的问题. 2015 年, 欧盟设计了一套测度欧洲数字经济与社会进步的指标体系——欧洲数字经济与社会进步指数 (DESI), 跟随社会发展, 测度范围进行了调整和扩展[1]. 马岩等使用层次分析法和专家打分法 (德尔菲法) 确定指标权重[2], 使用专家打分法并通过增加专家与问卷发放数量来提高准确度[3,4]; 马增林等使用波拉特法测度黑龙江农业信息化水平, 实际测度使用三个信息部门比重类指标[5]; 朱婕、岳毅蒙等使用熵权法确定指标权重, 而指标选取是通过综合现有研究, 总结分类得出[6,7]; 灰色关联动态分析法可以得出影响信息化水平发展指标的重要性排序及其时空动态变化[8]; 模糊综合评价法是一种确定指标权重的有效方法, 其

① 基金项目: 国家自然科学基金青年科学基金(61402109); 科技部对欧合作专项(247608); 福建省青年基金创新项目(2015J05120); 福建省教育厅A类项目 (JA15116)
收稿时间: 2017-10-27; 修改时间: 2017-11-14; 采用时间: 2017-11-17; csa 在线出版时间: 2018-05-28

中选择基准指标是关键环节[9]; 国家信息中心的全国信息社会发展指标指标考虑全面, 分级科学, 曾应用在全球信息社会发展水平测度上[10].

新一代信息化水平测度指标体系在调整、扩展测度范围的同时, 指标选取与权重确定仍是一个对测度结果科学性、准确性有直接影响的重要因素. 广泛搜集已有指标体系, 综合选取指标从广度上保证了指标的全面性; 使用专家打分法并增加调查问卷发放数量, 能直接、便捷的借鉴已有经验, 但在指标保留与删除、具体权重确定等在客观性上有所缺失, 亟需定量的数理方法参与到这一过程中来, 以得出合理可靠、科学严谨的信息化水平测度指标体系.

国家信息中心隶属于国家发展和改革委员会, 科研经验丰富, 指标框架设计合理, 在数据获取方面有得天独厚的条件, 颇具权威性, 因此考虑借鉴此课题的指标选取与整体框架, 综合十套信息化水平测度指标体系, 进一步进行修正, 在具体指标选取、权重确定中采用更多的数理方法, 使相关指标及其权重的确定更具有说服力.

2 信息化水平指标体系修正

2.1 数据与研究路线

信息化水平测度指标体系的修正需要大量、准确的数据作为支撑. 为保证研究的科学性与准确性, 本文选取国家权威部门发布的统计数据, 主要包括《中国统计年鉴 2016》[11]、《中国科技年鉴 2016》[12]与《中国信息年鉴 2016》[13].

为使不同指标数据均具可比性与同趋化, 需进行数据标准化. 考虑到统计数据分布特征, 采用 max-min 标准化方法得出原统计数据的正向标准化数据, 公式如下:

$$x'_{ik}=\frac{x_{ik}-\min\limits_{1\le i\le n}x_{ik}}{\max\limits_{1\le i\le n}x_{ik}-\min\limits_{1\le i\le n}x_{ik}} \quad (1)$$

式中, x_{ik} 为第 k 地区第 i 个指标的统计值, n 为指标个数.

本文以现有诸多信息化水平测度指标体系为基础, 利用词云分析、相关系数、变异系数递进式筛选指标, 指标分类后进行类别内部因子分析, 确定指标体系, 使用 3 种客观赋权法确定各指标权重, 最后计算各地区综合排名、得分并对整体过程进行总结分析.

2.2 指标综合与词云分析

综合现有十套信息化测度指标体系, 得到 186 个指标, 利用词云分析提取关键词并计算其出现频率, 能直观显示出 186 个指标中被频繁提及的指标. 根据谷尼舆情图悦 picdata.cn 热词分析工具分析得出热词图词频与权重图、关键词词频表.

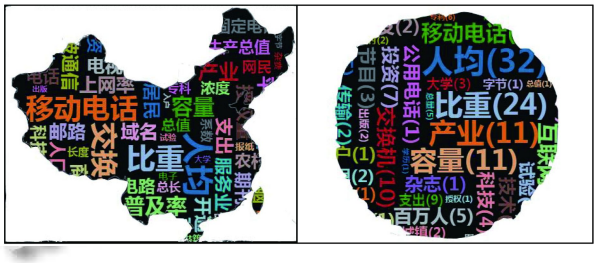

图 1 热词词频与权重图

关键词词频显示, 人均、比重这一类次词频最高, 说明大多体系都包含了比值类相对指标. 以具有权威性的国家信息中心发布的信息社会评测指标为主, 综合以上词频图, 从十份信息化指标体系的 186 个指标中初步选取 42 个指标 (见表 1), 并从以上年鉴中提取、计算出这些指标在全国 31 个省市区 (不包括港、澳、台) 的具体值.

表 1 关键词词频表

词频	关键词
32	人均
24	比重
11	移动电话、容量、产业、
10	交换机、互联网
9	支出、经费、人数、计算机、技术
8	普及率、人口
7	投资
6	邮电业、拥有率、覆盖率、专利、广播
5	光缆、百万人、第三产业、电视机、在校、长度、总量、教育
4	用户数、宽带、长途、居民、电话、设施、科技
3	拥有量、长途电话、增加值、生产总值、电视节、就业、节目、报纸、图书、点数、学生、电话机、消费、百分比、从业、大学生、专业

2.3 初步指标筛选

统计数据不同指标间可能具有较强的相关性, 与其他指标相关性较大的即视为冗余指标, 可通过相关系数的计算予以剔除. 计算 42 个指标间的相关系数, 第 i 个指标和第 j 个指标的相关系数 r_{ij} 的计算公式:

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{p}(x_{ki}-\overline{x_i})\sum_{k=1}^{p}(x_{kj}-\overline{x_j})^2}} \quad (2)$$

式中, k 为所考虑地区; i、j 为不同指标 (相同指标相关系数为 1); p 为研究单元数量 (本文中 m=31).

相关系数说明指标间差异性, 变异系数可说明指标内部数据的离散程度, 一般认为, 离散程度过小的指标, 对不同地区间差异性的描述意义不大[14–16]. 对经相关系数分析后剩余的 34 个指标进行变异系数分析, 以进一步简化指标体系:

$$v_k=\frac{s_k}{\overline{x_k}} \quad (3)$$

式中, s_k 表示 x_{ik} 的样本标准差, k 表示具体指标; $\overline{x_k}$表示 k 指标在 i 单元具体值的算数平均值.

表 2 词云分析指标选取结果

编号	指标名称	编号	指标名称
*N*1	人均地区生产总值 (元)	*N*22	普通高等学校 (本专科) 在校学生数 (人)
*N*2	高技术产业企业数 (个)	*N*23	大专及以上文化程度人口 (人)
*N*3	移动电话用户数 (户)	*N*24	普通高等学校数量 (所)
*N*4	移动电话基站 (万个)	*N*25	R&D 从业人员 (人)
*N*5	移动电话交换机容量 (万户)	*N*26	国内专利申请受理数 (件)
*N*6	固定长途电话交换机容量 (路端)	*N*27	三种专利申请授权数 (件)
*N*7	局用交换机容量 (万门)	*N*28	国内有效专利数 (件)
*N*8	第三产业增加值 (亿元)	*N*29	有线广播电视实际用户数 (万户)
*N*9	互联网普及率	*N*30	图书、报刊、杂志总印张数 (亿)
*N*10	IPv4 地址数 (万个)	*N*31	订销报纸、杂志累计数 (万份)
*N*11	网站数 (万个)	*N*32	快递 (万件)
*N*12	互联网上网人数 (万人)	*N*33	快递业务收入 (万元)
*N*13	域名数 (万个)	*N*34	教育经费 (万元)
*N*14	每百人使用计算机 (台)	*N*35	公共图书馆总流通人次 (万人次)
*N*15	各地区网上零售额 (亿元)	*N*36	电子商务销售额 (亿元)
*N*16	邮电业务总量 (亿元)	*N*37	信息技术服务收入 (万元)
*N*17	广播节目综合人口覆盖率	*N*38	微波实有站 (座)
*N*18	电视节目综合人口覆盖率	*N*39	专业公共卫生机构 (所)
*N*18	光缆线路长度 (公里)	*N*40	亿元以上商品交易市场数 (个)
*N*20	互联网宽带接入端口数 (万个)	*N*41	技术市场成交额 (万元)
*N*21	互联网宽带接入用户数 (万户)	*N*42	邮政业就业人员数 (人)

综合考虑变异系数与相关系数, 相关系数大于 0.8 说明两组数据相关性强、大于 0.9 说明两组数据相关性极强. 计算 42 个指标内部两两相关系数、每个指标与其他 42 个指标相关系数范围, 进而分别统计 41 个指标中, 与目标指标相关系数大于 0.8、0.9 的个数, 用 *Co*1、*Co*2 表示, 此结果越大, 说明该指标越能被其他指标说明, 即其冗余性越高, 考虑予以删除. 变异系数度量总体相对变异性, 作为一个无量纲数可以表征总体内部离散性. 变异系数过小 (本文取 0.15), 说明该指标在研究区内的区分度较小, 考虑删除指标. 综合变异系数 (*Cv*) 与相关系数的结果, 删除指标如表 3 所示.

表 3 初步删除指标一览表

编号	*Cv*	*Co*1	*Co*2	编号	*Cv*	*Co*1	*Co*2
*N*34	0.71	28	12	*N*35	1.07	20	6
*N*25	1.04	26	8	*N*30	0.79	20	6
*N*16	0.99	25	14	*N*2	1.46	19	7
*N*8	0.86	24	10	*N*28	1.37	19	10
*N*21	0.83	24	10	*N*3	0.77	19	9
*N*4	0.72	24	10	*N*26	1.25	18	6
*N*29	0.75	23	10	*N*12	0.75	18	9
*N*20	0.74	23	9	*N*23	0.61	17	6
*N*27	1.35	20	8				

2.4 分类-因子分析

参照国家信息中心所制定的信息社会评价指标体系, 将剩余 25 个指标分为 4 类, 在组内分别进行因子分析, 以进一步简化指标.

(1) *KMO* 检验

标准化后的数据能否进行因子分析需先进行 *KMO* 检验:

$$KMO = \frac{N}{M+N} \tag{4}$$

M: 所有变量两两之间 (不包括变量与自身) 的偏相关系数的平方和;

X 和 *Y* 的偏相关系数: *X* 和 *Z* 线性回归得到的残差 R_Y 与 *Y* 和 *Z* 线性回归得到的残差 R_Y 之间的简单相关系数, *Z* 代表其他所有的变量[17];

N: 所有变量两两之间 (不包括变量与自身) 相关系数的平方和.

当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时, *KMO* 值接近 1. *KMO* 值越接近于 1, 说明变量间的相关性越强, 原有变量越适合作因子分析[18,19]; 反之亦然. 对四类指标分别进行检验的结果如表 4.

表 4 分类别检验表

类别	*KMO* 值	*Bartlett* 检验
第一类	0.761	159
第二类	0.724	300
第三类	0.653	144
第四类	0.686	120

4 类 *KMO* 值均大于 0.6, 适宜进行因子分析.

(2) 因子分析

因子分析可在 SPSS 中进行, 对其结果进行整理分析四类指标分别提取两个主成分, 都可表达原数据 85% 以上的信息率, 旋转成份载荷矩阵各因子贡献率在 0.9 以上的多个指标能表示出原数据绝大部分的信息[20], 其余指标对整体贡献过小, 相当于冗余信息, 删除这一类指标对于指标体系整体的简洁、高效具有重要意义, 故以载荷矩阵因子贡献率 0.9 作为阈值进一步筛选指标.

3 客观赋权与综合分析

客观赋权法是根据数据特点进行赋权, 排除了人工干扰, 能够得出各指标科学、准确的权重[21], 常用的客观赋权法有标准离差法、CRITIC 法与熵权法.

3.1 标准离差法

指标标准差越大, 说明其指标值的变异程度越大, 提供的信息量越大, 在综合评价中所起的作用越大, 则其权重也越大, 反之亦然[22]. 利用标准差计算权重的公式为:

$$w_j = \frac{\delta_j}{\sum_{j=1}^{n} \delta_j} \tag{5}$$

式中, w_j 表示 *j* 指标在指标体系中的权重; δ_j 表示 x_i 的标准差; x_i 表示 *j* 指标在各研究单元的具体值; $\bar{x}$表示 *j* 指标具体值的算数平均值; *m* 表示研究单元数量 (本文中 *m*=31); *n* 表示指标个数.

表 5 因子分析保留指标

指标	成分矩阵贡献率
人均地区生产总值 (元)	0.941
普通高等学校 (本专科) 在校学生数 (人)	0.924
普通高等学校数量 (所)	0.952
快递 (万件)	0.956
快递业务收入 (万元)	0.951
IPv4 地址数 (万个)	0.953
每百人使用计算机 (台)	0.95
电视节目综合人口覆盖率	0.924
光缆线路长度 (公里)	0.913

3.2 Critic 法

基于指标相关性的指标权重确定方法 (criteria importance through inter-criteria correlation) 由 Diakoulaki 提出[23], 其中对比强度表示同一个指标各个评价方案之间取值差异的大小, 标准差越大, 不同方案之间取值差异越大; 评价指标之间的冲突性以指标间的相关性为基础, 两个指标之间相关性越强, 冲突性越弱, 第 *j* 个指标与其他指标冲突性的量化公式为:

$$C_j = \delta_j \sum_{i=1}^{n} \left(1 - r_{ij}\right) \tag{6}$$

式中, C_j 表示 *j* 指标包含的信息量; δ_j 表示公式 (5) 中的计算结果; r_{ij} 表示指标 *i*、*j* 间的相关系数, 具体计算参照公式 (2); *n* 表示指标个数.

C_j 越大, *j* 指标包含的信息量越大, 该指标的相对重要性也就越大, 相应权重为:

$$W_j = \frac{C_j}{\sum_{j=1}^{n} C_j} \tag{7}$$

式中, W_j 表示 *j* 指标在指标体系中的权重; C_j 表示公式 (6) 计算结果; *n* 表示指标个数.

3.3 熵权法

熵权法是目前社会学、地理学、信息论各学科常

用的一种客观赋权法, 指标信息熵与变异程度呈负相关关系, 信息熵越小, 变异程度越大, 包含的信息量越大, 对综合评价的影响越大, 反之亦然[24]. 熵值计算公式为:

$$E_j = -(\ln m)^{-1} \sum_{i=1}^{m} P_{ij} \ln P_{ij} \tag{8}$$

$$P_{ij} = \frac{d_{ij}}{\sum_{i=1}^{m} d_{ij}} \tag{9}$$

式中, m 表示研究单元数量 (本文中 m=31); n 表示指标个数; d_{ij} 表示 j 指标标准化后的具体值; $P_{ij}=0$时, $\lim\limits_{d_{ij}\to 0} P_{ij}\ln P_{ij}=0$;

$$W_j = \frac{1-E_j}{n-\sum_{j=1}^{n} E_j} \tag{10}$$

式中, E_j 为公式 (8) 中的计算结果; n 表示指标个数.

3 种客观赋权法得出综合得分与排名情况如图 2 和图 3.

由相关系数按顺序计算熵权法与标准离差法、熵权法与 CRITIC 法、标准离差法与 CRITIC 的得分、排名折线图的拟合度, 可得以上曲线的拟合程度.

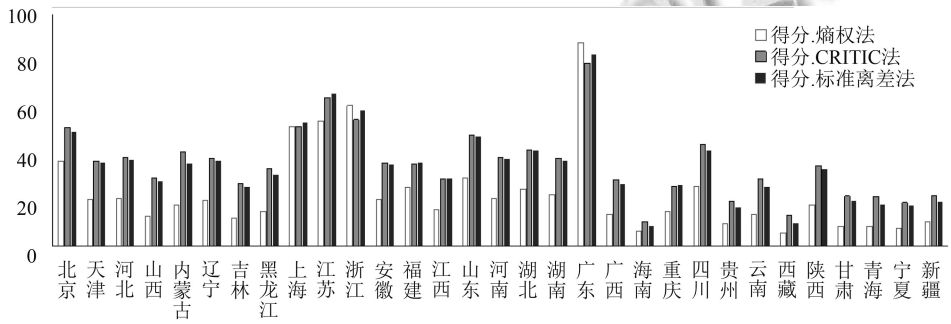

图 2　全国信息化水平得分图

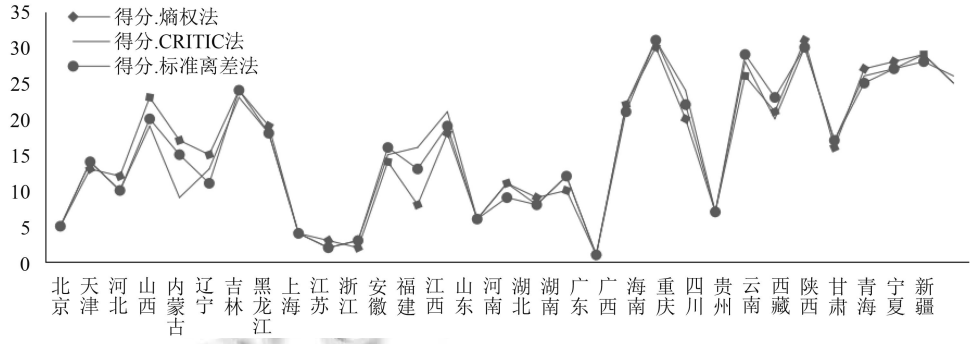

图 3　全国信息化水平排名图

表 6　结果拟合程度表

拟合变量	得分 1	得分 2	得分 3	排名 1	排名 2	排名 3
拟合度 (%)	92.9	94.7	99.6	96	98.4	97.9

可知, 拟合度全部在 90% 以上, 最高达 99.6%, 说明客观赋权法所得权重与结果较为科学准确, 能通过统计数据刻画全国 31 个省市区的信息化发展水平.

4　总结与展望

本文立足于 2015 年国家统计数据及现有十套信息化水平测度指标体系, 针对信息化测度指标选取与权重确定两大关键环节进行了修正与重建. 收集现有指标或者依据自身经验判断直接筛选指标, 存在较大主观性, 针对这一问题, 文中采用了词云分析的方法, 通过关键词词频筛选指标, 使指标选取结果更为客观, 进而使用递进式方法继续筛选指标, 利用相关系数的范围删除冗余指标, 变异系数衡量指标内部差异, 在因子分析中以对载荷成分矩阵贡献率作为测度指标, 在分类的基础上选取能表达原有指标体系至少 90% 信息的指标, 得出了简洁高效、可靠合理的指标体系; 权

重确定方面, 针对现有赋权方法主观性与难以说明指标内部信息的问题, 文中采用了客观赋权法, 充分挖掘数据的内部联系与意义, 且使用三种客观赋权法相互比较, 结果拟合度较高也能说明文中得出指标体系的合理性. 因此, 本文对于将数理方法与已有经验相结合进行信息化测度指标体系修正有重要意义.

本研究后续将以目前得到的信息化水平测度指标体系为起点, 丰富从现有指标体系得出的指标库, 扩展研究的时间尺度, 加强数理方法与已有经验的结合, 进一步完善信息化水平测度指标体系的修正与分析.

参考文献

1 European Commission: DESI2015: The digital economy and society index. https://ec.europa.eu/digital-agenda/en/digital-economy-and-society-index-desi. [2015-04-07].

2 马岩, 孙红蕾, 郑建明. 流动空间视角下新型城镇信息化水平测度实证分析. 图书馆论坛, 2017, 37(5): 18–26.

3 苏君华, 孙建军. 全国及各省市信息化水平测度. 情报科学, 2005, 23(6): 817–822.

4 杨洋. 安徽省区域信息化水平测度及其对经济增长影响的实证研究[硕士学位论文]. 合肥: 合肥工业大学, 2015.

5 马增林, 王天一, 张云峰, 等. 黑龙江省农业信息化水平测度分析. 中国集体经济, 2017, (33): 22–24. [doi: 10.3969/j.issn.1008-1283.2017.33.012]

6 朱婕. 江苏省新型城镇化和信息化协调发展测度研究[硕士学位论文]. 南京: 南京大学, 2017.

7 岳毅蒙, 李江涛. 基于改进熵权法的智能手机评价模型. 计算机系统应用, 2017, 26(4): 203–206. [doi: 10.15888/j.cnki.csa.005651]

8 李燚, 丁生喜, 任海静. 基于灰色关联分析法的青海省信息化与区域经济发展分析. 价值工程, 2017, 36(30): 55–58.

9 Yang YP, Shan N. Evaluation of shallow groundwater quality in Haikou based on fuzzy comprehensive evaluation method. Ground Water, 2017, 39(4): 20–22, 59.

10 国家信息中心. 中国信息社会发展报告 2015. 北京: 国家信息中心, 2015.

11 国家统计局. 2016 中国统计年鉴. 北京: 中国统计出版社, 2016.

12 国家统计局社会科技和文化产业统计司, 科学技术部创新发展司. 2016 中国科技统计年鉴. 北京: 中国统计出版社, 2016.

13 国家信息中心. 中国信息年鉴. 北京: 《中国信息年鉴》期刊社, 2016.

14 陈勇, 杨未未. 信息化水平测度方法研究. 科技情报开发与经济, 2009, 19(6): 90–92.

15 许慧玲. 信息化水平测度及对区域经济增长影响研究[博士学位论文]. 南京: 南京农业大学, 2008.

16 李美洲, 韩兆洲. 信息化水平测度——以广东省为例. 科技管理研究, 2007, (7): 172–175.

17 陈小磊, 郑建明, 万里鹏. 信息化水平测度指标体系理论研究述评. 图书情报知识, 2006, (5): 65–70.

18 刘文云, 葛敬民. 国内外信息化水平测度理论研究比较. 情报理论与实践, 2004, 27(2): 144–147.

19 郑丽琳. 信息化水平测度研究综述. 合作经济与科技, 2005, (2S): 60–61.

20 王爱兰, 张俊山. 评美国与日本学者关于信息化水平测度的理论与方法——兼论我国国家信息化水平测度指标体系的完善. 图书情报工作, 2005, 49(1): 117–120, 137.

21 颜惠琴, 牛万红, 韩惠丽. 基于主成分分析构建指标权重的客观赋权法. 济南大学学报 (自然科学版), 2017, 31(6): 519–523.

22 杨宇. 多指标综合评价中赋权方法评析. 统计与决策, 2006, (7): 17–19.

23 梁海丽, 于洪彬. 我国信息化水平指数测度研究. 情报资料工作, 1999, (4): 4–8.

24 于伟, 张鹏. 我国信息化水平的空间不均衡、极化特征和收敛性研究. 山东财经大学学报, 2016, 28(5): 92–99.