文献与数据学报
第2卷 第2期 2020年6月

一种基于学科差异的ESI学科潜力值校准方法*

李婷婷 魏园婷
（西安理工大学图书馆，西安 710048）

摘 要：[目的/意义]“双一流”高校建设方案与实施办法颁布以及第四轮学科评估将ESI指标纳入评估体系后，多数高校图书馆都开展了基于WoS平台、ESI和InCites数据库的学科数据分析评价工作。在实际工作中因为数据库间的数据不同步问题导致分析结果存在一定误差，影响学科分析结果的参考价值。[方法/过程]应用Kruskal-Wallis H检验ESI数据库22个学科间机构ESI/InCites被引频次比值数据，验证了ESI学科之间的差异，参考ESI学科差异计算相应学科的潜力值校准系数，设计学科潜力值校准方法。[结论/结果]数据分析与实例验证结果表明，本方法能够帮助决策层更加直观、准确地衡量机构尚未进入ESI前1%的学科与ESI阈值的相对位置，对于了解机构学科现状、合理调整学科政策等具有实际应用价值。

关键词：ESI 学科差异 学科潜力值 校准
分类号：G353.1
DOI：10.31193/SSAP.J.ISSN.2096-6695.2020.02.11

自2015年起，“双一流”高校建设方案与实施办法的颁布对高校图书馆的学科信息服务提出了更高要求，高校图书馆面临着新的发展机遇和挑战。2016年第四次学科评估将Essential Science Indicators（简称ESI）指标纳入评估体系后，ESI越来越受到重视。2019年对42所“双一流”大学图书馆的调研显示，学科数据分析评价已经逐步形成一定的业务规模与模式流程，39所设置相关业务的高校图书馆，都开展了依托Web of Science（简称WoS）平台、ESI和InCites数据进行数据分析的实践[1]。ESI与InCites数据库具有数据标准、指标完备、易使用、易对标等特点，不少机构将自身在ESI的发展纳入学科建设规划。如何能够准确预判优势学科与潜力学科的发展成为新的研究热点，对于未进入ESI的学科潜力值的校准则具有关键的战略意义。

1 研究背景

ESI与InCites数据库主要存在三大不同：第一，数据库基础数据覆盖范围不同：ESI数据库仅统

* 本文系2018年度陕西省教育厅人文社会科学专项基金“双一流背景下高校学科论文产出分析支持学科发展决策研究”（项目编号：18JK0529）、西安理工大学科研基金“西安理工大学ESI高被引论文检索分析研究”（项目编号：602-451418003）成果之一。
[作者简介]李婷婷，女，馆员，硕士，研究方向为学科服务、数据服务，Email：tsgltt@xaut.edu.cn；魏园婷，女，助理馆员，硕士，研究方向为数据服务、用户研究，Email：442054724@qq.com。

计 SCIE 和 SSCI 两个数据库中近 11 年来以 Article、Review 和 Datapaper 为文献格式的论文及引文篇数作为各项分析指标的数据基础，而 InCites 数据库则涵盖了 WoS 核心合集中的七大引文索引数据库中 1980 年至今的更多文献格式（例如：会议论文、会议摘要等）的论文及引文篇数作为统计分析的数据基础；第二，数据库更新时间不同：ESI 数据库每两个月更新一次，而 InCites 数据库则每月更新一次，而且在同一时间点看 InCites 数据库比 ESI 数据库覆盖了更多新的论文数据；第三，ESI 数据库只提供基于基础数据得出的各项分析指标结果数据列表，而 InCites 数据库则提供基础数据以及多项分析功能。

尽管 ESI 与 InCites 两个数据库存在的数据不同步问题导致同一时间在两个数据库中提取的机构学科数据并不一致，但对机构 ESI 学科指标进行深入分析与预测所采用的基础数据仍需来源于 InCites 数据库，所以将两个数据库的数值进行比对分析并建立预测模型则显得尤为重要。

从文献看来，较成熟的基于科研产出数据分析的 ESI 潜力学科测算与预测研究从 2011 年开始出现，典型的有：2012 年中国农业大学陈仕吉等[2]提出了“学科欠缺度指标 Pd”，该方法基于学科总被引次数、发文篇数和篇均被引频次等指标衡量机构学科与 ESI 差距；2013 年东华大学董政娥等[3]提出了“学科比重 Qi”指标，该方法通过计算 InCites 中机构某学科的被引频次与 ESI 中对应学科基准线的比值来测算 ESI 潜力学科；2016 年清华大学管翠中等[4]提出的采用潜力学科被引频次与 ESI 学科阈值两者历史曲线拟合预测的方法；2017 年南京医科大学汪莉[5]提出了根据“学科领域百分位”字段计算学科 EV 值进行潜力学科预测的方法；2018 年中国科学院大学侯志江[6]提出了通过计算机构 ESI 与 InCites 学科排位进行潜力学科发展态势预测的方法；2018 年中国地质大学程建萍等[7]提出了“学科误差修正因子 Qi”指标，该方法通过计算每个学科排序后 100 位机构在 ESI 与 InCites 数据库中被引次数的比值 Xi，再分学科加权平均得到修正因子 Qi。

综上，目前已有根据学科总被引次数差距、学科领域百分位、学科排位等不同角度进行计算的 ESI 学科潜力值指标，其中较为主流的方法是通过机构学科被引频次数差距来进行预测。早年的方法比较简单，近年则做了一些补充研究，本文也是在两个数据库机构学科被引频次差距理论的基础上进一步实践总结得出的方法。其中 2018 年程建萍等提出的“学科误差修正因子 Qi”与本研究有相似之处，但具体有两点疑问尚未解答：第一，文中提及在日常工作中常看到 ESI 数据的学科差异比较大，但学科差异是怎样的，如何验证，文中并未详说；第二，文中提及对学科内各机构的 Xi 值进行加权平均得到修正因子 Qi，但权重分配方案并未详说。

本研究尽量选取相同时间段从 ESI 与 InCites 两个数据库提取机构论文被引频次数据作为研究基础数据，首先根据各学科内机构 ESI/InCites 论文被引频次比值数据验证 ESI 数据库所划分的 22 个学科间的差异性，然后分别计算 ESI 各学科的潜力值校准系数，并进行筛选验证。

2　ESI 学科差异性验证

在科学研究过程中，各学科间由于研究习惯、研究行为、论文发表方式和渠道、技术更新速度等的不同，导致各个学科科研成果论文发表以及施引文献的文献类型比例有所不同，例如：计算机科学、工程学等学科更倾向于会议交流；人文社会科学相关学科更倾向于图书资料等。

由于存在 ESI 数据库与 InCites 数据库数据不同步问题以及学科差异问题，在研究机构各个

ESI 学科的时候，从 InCites 数据库导出的机构各学科数据与相应 ESI 各学科数据的差距是不一致的，需要将数据不同步和学科差异问题结合考虑，不能一概而论。本文通过统计学方法对不同 ESI 学科的修正因子进行分别计算，首先需对 ESI 学科的差异性进行统计学检验。

2.1 数据说明

为了尽量避免 ESI 数据库与 InCites 数据库数据不同步问题导致的数据计算误差，本研究在数据选取时特别注意了两个问题，即数据库时间窗和数据库文献类型的问题。首先，遴选 ESI 与 InCites 两个数据库时间窗相近的机构被引频次数据作为基础数据，ESI 数据：2019 年 7 月 11 日更新，数据时间窗为 2009.1.1–2019.4.30；InCites 数据：2019 年 5 月 31 日更新，数据时间窗为 2009.1.1–2019.5.3。其次，将 InCites 数据库导出数据的文献类型限定为 ESI 数据相应文献类型，即“Article”、“Review”和“Data Paper”。通过以上前期准备可以尽量规避两个数据库中提取的数据误差。

2.2 数据处理

2.2.1 有效数据组筛选

本研究以 ESI 数据库 22 个学科分类中的所有机构作为基础数据，经与 InCites 数据库相应学科机构名列表进行对比，剔除由于缩写不规范、不完整、在 InCites 数据库中无对应机构的数据，剔除后剩余 19094 组对照机构数据。

将所有有效机构的 ESI/InCites 被引频次比值计算出来作为计算校准系数的基础数据，对其进行 K–S 正态检验（表 1），结果显示校准系数基础数据的 P 值远小于 0.05，所以校准系数基础数据符合正态分布规律，故可以利用拉依达准则（3σ 准则）剔除异常数据。

表 1 校准系数基础数据正态性检验结果

	统计量	df	P 值
比值	0.464	19094	0.000

拉依达准则认为当实验数据值总体服从正态分布时，其中出现大于 $\mu+3\sigma$ 或小于 $\mu-3\sigma$ 数据值的概率很小（μ 与 σ 分别表示正态总体的数学期望和标准差），小于 0.3%。则可认为数据值几乎全部集中在 $[\mu-3\sigma,\mu+3\sigma]$ 区间内，超出此范围的数据即被认为是异常数据，应当剔除。本数据均值 $\mu=0.9573$，标准差 $\sigma=0.7187$，根据拉依达准则可将机构 ESI/InCites 被引频次比值落在 [–1.1988，3.1134] 区间外的数据组剔除。经过筛选，剔除 3σ 区间以外的 10 个 ESI 学科中的 11 个异常机构数据组，如表 2 所列。剩余的 19083 组有效数据将作为验证 ESI 数据 22 个学科间差异性和计算 ESI 学科潜力值校准系数的基础数据。

2.2.2 ESI 学科差异性验证

根据数据库介绍可知，在 ESI 数据库中每篇论文仅会被分入唯一的 ESI 学科，故机构的论文将被不重复的分入不同的 ESI 学科，可以认为每个学科分组下的机构被引表现是独立的，可采用非参数检验中 Kruskal–Wallis H 检验判断各 ESI 学科分组间的机构 ESI/InCites 被引频次比值有无差异。Kruskal–Wallis H 检验是一种广泛使用的非参数检验方法，其基本思路是，首先对所有样本合并后按升序排列得出每个数据的秩，然后对各组样本求平均秩，若平均秩相差很大，则认为两组样本分别所属的总体有显著差异[8]。

表 2　3σ 区间外的特异值数据

学科	机构名	ESI 被引次数	InCites 被引次数	比值
地球科学	UNIVERSITE DE BREST	22567	465	48.5312
地球科学	ENGLISH HERITAGE	8033	202	39.7673
工程学	LG ELECTRONICS	2658	2	1329.0000
化学	UNIVERSITE DE BREST	14729	3191	4.6158
环境生态学	UNIVERSITE DE BREST	9085	687	13.2242
交叉学科	CAIRO UNIVERSITY	2843	100	28.4300
临床医学	UNIVERSITE DE BREST	14241	207	68.7971
农业科学	UNIVERSITE DE BREST	3391	884	3.8360
物理学	IVANE JAVAKHISHVILI TBILISI STATE UNIVERSITY	24145	3596	6.7144
一般社会科学	NATURE CONSERVANCY	1713	352	4.8665
植物与动物科学	UNIVERSITE DE BREST	8565	398	21.5201

本研究在 K–W 检验前对 ESI 学科进行编码，用 1= 材料科学，2= 地球科学，…，以此类推。编码后，以 ESI 学科为分组变量、学科内相应机构 ESI/InCites 被引频次比值为检验变量，进行 K–W 检验，结果如表 3 所示，数据的渐近显著性水平为 0.000，远小于 0.05；进一步通过 K–W 检验中的 ESI 学科组分别进行成对比较，结果显示，除六组学科之间的显著性水平大于 0.05 外（表 4），其余成对检验结果均小于 0.05，说明除个别学科两两之间的校准系数基本不存在差异外，不同 ESI 学科间的 ESI/InCites 被引频次比值是存在差异的。

表 3　学科样本 K–W 检验结果

	比值
卡方	11363.762
df	21
渐近显著性	0.000

表 4　学科样本成对比较结果

学科 – 学科	调整显著性
经济与商学 – 数学	0.053
计算机科学 – 工程学	0.066
经济与商学 – 物理学	0.104
工程学 – 经济与商学	0.290
一般社会科学 – 空间科学	1.000
物理学 – 数学	1.000

* 其余成对检验结果均小于 0.05

综上，进行机构某学科的潜力值校准时应考虑 ESI 学科间差异因素，需要根据不同学科分别计算学科潜力值校准系数。

3　机构 ESI 学科潜力值校准方法

3.1　定义

本研究通过对比相应数据时间窗内，某学科内各有效机构的 ESI 被引频次与 InCites 被引频

次比值来修正尚未入围 ESI 排名的机构学科潜力值。公式如下：

$$P=\frac{IC\times Q_i}{C_i}=\frac{IC\times AVERAGE\left(ICR_1,ICR_2,.....,ICR_n\right)}{C_i} \quad (1)$$

其中，P 为机构某学科的潜力值，IC 为机构某学科在 InCites 数据库中的被引频次数值，i 代表某个 ESI 学科，Q_i 为某 ESI 学科的潜力值校准系数，C_i 为某 ESI 学科的阈值数据，ICR 为某 ESI 学科内各有效机构 ESI/InCites 被引频次比值，n 为剔除异常比值数据后的各学科有效机构数。

根据 ESI 与 InCites 数据库数据范围来看，潜力值校准系数 Q_i 值一般小于 1，Q_i 值越接近 1，则代表该学科内各有效机构 ESI/InCites 被引频次越接近；Q_i 值越小，则代表该学科内各有效机构 ESI/InCites 被引频次差距越大。

3.2　学科潜力值校准系数

根据 Q_i 定义，分别对 ESI 各学科分组中有效机构的 ESI/InCites 被引频次比值求均值，分别得到 ESI 数据库 22 个学科的潜力值校准系数（表 5），发现除计算机科学与工程学两个学科潜力值校准系数较低外，其余 20 个学科的 Qi 值均在 0.9–1 之间，其中计算机科学为 0.7320，工程学为 0.8310，免疫学最高达到 0.9822。这说明计算机科学和工程学两个学科中，机构被引频次的 InCites 数据与 ESI 数据差距较大；而免疫学学科中机构被引频次的 InCites 数据与 ESI 数据差距较小。根据学科 Q_i 值即可以对尚未进入 ESI 的机构学科数据进行校准，计算出合理的、更贴近 ESI 实际情况的机构学科潜力值数据。

表 5　各 ESI 学科校准系数

ESI 学科	潜力值校准系数	ESI 学科	潜力值校准系数
计算机科学	0.7320	环境生态学	0.9616
工程学	0.8310	交叉学科	0.9617
经济与商学	0.9019	化学	0.9635
数学	0.9221	神经科学与行为科学	0.9650
物理学	0.9227	生物与生物化学	0.9655
一般社会科学	0.9243	植物与动物科学	0.9671
农业科学	0.9500	药理学和毒理学	0.9701
材料科学	0.9503	微生物学	0.9755
空间科学	0.9511	分子生物与遗传科学	0.9762
精神病学 / 心理学	0.9546	临床医学	0.9786
地球科学	0.9579	免疫学	0.9822

4　验　证

分别对本次数据（2019 年 7 月）中，机构学科被引频次在 ESI 相应学科阈值之上但尚未进入 ESI 高被引机构列表的中国大陆机构进行检验，共涉及 15 个学科组的 52 个机构。检验结果如表 6 所示，

大部分机构的 P 值经过修正后都回到了 100% 以下。计算机科学和工程学两个学科由于其机构被引频次的 InCites 数据与 ESI 数据差距较大，Q_i 值较低，因此机构 P 值的修正效果更加明显，所涉及的机构较多。计算机科学学科当期 ESI 阈值为 3373，所选取的 10 所机构的被引频次数值修正后都回落到阈值以下。工程学学科当期 ESI 阈值为 2574，所选取的 16 所机构，修正后仍有一所机构（北方工业大学）被引频次数值高于阈值，其他 15 所机构修正后被引频次数值都回落到阈值以下。免疫学 Q_i 值高，中科院 – 生物物理研究所免疫学的学科 P 值经过修正后仍大于 100%，与修正前 P 值比较接近。

表 6　对潜力值校准系数的验证

ESI 学科组	机构	InCites 数值	修正值	修正前 P 值	修正后 P 值
计算机科学	中科院 – 软件研究所	4443	3252	131.72%	96.42%
	苏州大学 **	4280	3133	126.89%	92.88%
	北京工业大学 **	4021	2943	119.21%	87.26%
	浙江工业大学 *	3959	2898	117.37%	85.92%
	大连海事大学	3935	2880	116.66%	85.40%
	重庆邮电大学	3646	2669	108.09%	79.12%
	吉林大学	3632	2659	107.68%	78.82%
	中科院 – 数学与系统科学研究院	3616	2647	107.20%	78.47%
	南开大学	3515	2573	104.21%	76.28%
	福州大学	3441	2519	102.02%	74.68%
工程学	北方工业大学 *	3326	2764	129.22%	107.38%
	中科院 – 沈阳自动化研究所 **	3092	2569	120.12%	99.82%
	上海师范大学 *	2819	2343	109.52%	91.01%
	中国农业科学院	2721	2261	105.71%	87.85%
	北京建筑大学 *	2705	2248	105.09%	87.33%
	中科院 – 南京土壤研究所 *	2691	2236	104.55%	86.88%
	东北财经大学 **	2679	2226	104.08%	86.49%
	中科院 – 青岛生物能源与过程研究所	2674	2222	103.89%	86.33%
	河南大学 *	2669	2218	103.69%	86.17%
	华南农业大学 **	2650	2202	102.95%	85.55%
	山东建筑大学	2639	2193	102.53%	85.20%
	西华师范大学 **	2601	2161	101.05%	83.97%
	烟台大学	2594	2156	100.78%	83.75%
	杭州师范大学	2591	2153	100.66%	83.65%
	西交利物浦大学	2585	2148	100.43%	83.46%
	湖南科技大学 **	2584	2147	100.39%	83.42%
经济与商学	浙江大学 *	4537	4092	108.10%	97.50%
	厦门大学 **	4222	3808	100.60%	90.73%
数学	天津工业大学	4289	3955	100.54%	92.71%
	南京师范大学	4273	3940	100.16%	92.36%
物理学	中国原子能科学研究院	22347	20620	107.44%	99.13%
	同济大学 *	21322	19674	102.51%	94.59%
	西北工业大学 **	20822	19212	100.11%	92.37%
一般社会科学	南京航空航天大学	1542	1425	105.04%	97.09%
农业科学	石河子大学 *	2277	2163	103.55%	98.37%
	清华大学 *	2250	2138	102.32%	97.20%
材料科学	中科院 – 上海技术物理研究所 *	6408	6090	101.31%	96.28%
	河北大学 *	6399	6081	101.17%	96.14%

续表

ESI 学科组	机构	InCites 数值	修正值	修正前 P 值	修正后 P 值
地球科学	西安电子科技大学 *	7177	6875	114.92%	110.09%
化学	西南科技大学 *	8229	7929	102.98%	99.22%
	上海理工大学 *	8085	7790	101.18%	97.48%
神经科学与行为科学	杭州师范大学	6434	6209	101.04%	97.50%
生物与生物化学	中科院 – 长春应用化学研究所	6487	6263	102.66%	99.12%
	中科院 – 遗传与发育生物学研究所	6435	6213	101.84%	98.32%
	北京林业大学 **	6336	6117	100.27%	96.81%
药理学和毒理学	青岛大学 *	3541	3435	103.06%	99.97%
	中国农业科学院 *	3484	3380	101.40%	98.37%
	广西医科大学 **	3455	3352	100.55%	97.55%
	南通大学 *	3445	3342	100.26%	97.26%
	华中农业大学 *	3437	3334	100.03%	97.04%
临床医学	北京航空航天大学	2661	2604	101.95%	99.77%
免疫学	中科院 – 生物物理研究所 *	5179	5068	103.19%	100.98%

注：* 表示 2019 年 9 月入围 ESI 前 1% 行列；** 表示 2019 年 11 月入围 ESI 前 1% 行列

笔者对这些机构学科进行追踪，看到北方工业大学工程学、西安电子科技大学地球科学和中科院—生物物理研究所免疫学（即 P 值修正后仍大于 100% 的机构学科）于之后一期（2019 年 9 月）入围 ESI 前 1% 行列；此外，P 值修正后接近 100% 或靠前的 28 所机构（涉及 9 个学科）也分别于 2019 年 9 月和 11 月入围 ESI 前 1% 行列。

5 结 论

基于 WoS 平台、ESI 和 InCites 数据库对机构各学科进行深度分析与评价，是高校图书馆利用资源和数据开展深层次学科服务的重要途径之一。本研究立足于日常工作中遇到的实际问题，采用统计学方法对 ESI 学科间差异进行验证，针对数据库间数据不同步这一实际问题，提出区别学科的 ESI 学科潜力值校准这一解决途径。

通过数据分析及实例验证，可以看到基于学科差异的 ESI 潜力值校准系数基本能够将科研机构在不同学科内的 InCites 数据库被引频次数值校准到更接近 ESI 数据库实际统计模式。本方法能够帮助决策层更加直观、准确地衡量机构尚未进入 ESI 前 1% 的学科与 ESI 阈值的相对位置，对于了解机构学科现状、合理调整学科政策等具有实际应用价值，能够为高校图书馆开展深度学科服务与情报服务提供便利。

本次统计数据选取由于时间原因，未能选择完全一致的时间窗，为了更加准确地进行实际测算应用，可选择两个数据库时间窗最为接近的 3 月份数据进行处理。同时，本文对于 ESI 学科差异未能进一步深入探究，后续将进一步关注 ESI 学科间的差异问题。

【参考文献】

［1］王雪莲，孔凡晶，刘万国，等．“双一流”大学图书馆学科数据分析工作调查研究［J］．图书馆学研

究 ,2019(10):68-74.

［2］陈仕吉 , 史丽文 , 左文革 . 科研机构潜势学科的识别方法与实证分析——以中国农业大学为例［J］. 情报杂志 ,2012,31(02):43-47.

［3］董政娥 , 陈惠兰 . 基于 ESI 和 InCites 数据库的东华大学学科发展预测［J］. 东华大学学报 (自然科学版),2013,39(05):689-694.

［4］管翠中 , 范爱红 , 贺维平 , 赵杰 , 孟颖 . 学术机构入围 ESI 前 1% 学科时间的曲线拟合预测方法研究——以清华大学为例［J］. 图书情报工作 ,2016,60(22):88-93.

［5］汪莉 . 基于 ESI 和 InCites 的高校潜力学科发展预测［J］. 情报杂志 ,2017,36(02):53-58.

［6］侯志江 . 基于 InCites 预测高校学科入围 ESI 前 1% 时间的方法研究［J］. 图书馆工作与研究 ,2018(04):37-45.

［7］程建萍 , 刘建辉 , 叶玫 . 基于 ESI 的潜力学科预测模型修正和实证分析［J］. 情报科学 ,2018,36(12):22-24+40.

［8］吴骏 .SPSS 统计分析从零开始学［M］. 北京 : 清华大学出版社 .2014:115-130.

A Calibration Method of Institutions' ESI Discipline Potential Value Based on Discipline Differences

LI Tingting　WEI Yuanting

(Xi'an University of Technology Library, Xi'an 710048, China)

Abstract:［**Purpose/significance**］After the "Double First-Class" university construction plan were promulgated, and the Fourth Discipline Assessment brought ESI index into system, most of university libraries have carried out discipline data analysis and evaluation based on WoS, ESI and InCites database. The data synchronization problems between databases will lead analysis results errors, which could affects the results' reference value.［**Method/process**］This article uses Kruskal Wallis H to test the institutions' ESI / InCites citation ratio data of ESI 22 discipline categories to verify the academic discipline differences, calculates calibration coefficients, and designes the calibration method of discipline potential values.［**Result/conclusion**］Through data analysis and case verification, it can be seen that this method could help decision-makers measure the relative position between institution and ESI threshold intuitively and accurately, and have practical value for finding out current institution situation and adjusting policies reasonably.

Keywords: ESI; Discipline difference; Discipline potential value; Calibration

（本文责编：周　霞）