ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research, volume 177
International Conference on Education Reform and Management Science (ERMS 2018)

On Analysis and Prediction of Universities' Research Competitiveness Using Scientometrics Methods

Yan ZHANG[1, 2,*], Yan-shuo ZHU[1], Man WU[1], Xiao-mei ZhOU[1], Xiao-chuan YIN[1], Guang-xu LI[2], Yu-tao ZHAO[2] and Kang SUN[1]

[1]Qingdao University of Science and Technology Library, Qingdao 266061, China

[2]College of Electromechanical Engineering, Qingdao University of Science and Technology, Qingdao 266061, China

*Corresponding author

Keywords: Research competitiveness, Universities, Scientometrics.

Abstract. How universities make decisions based on scientific and technological information and allocate resources effectively is an important guarantee to promote the construction of "Double First-rate" in China. Based on Pearson correlation and least squares linear fitting, this paper compares, analyzes and predicts the competitiveness of scientific research of ESI universities in Shandong province. The results show that *A* University has an absolute advantage in quantity, quality and total citations; also, *A* University, *F* University and *B* University have better performance in citation impact than the rest; and Web of Science documents quantity and H-index have a strong correlation with total citations; *C* University, *E* University and *D* University would possibly outperform *B* University in annual publications in WoS in 2020. The results of this study will help the universities in Shandong to grasp the orientation and trend of the research competitiveness, and further optimize the policy guidance and scientific research management, thus effectively promote the development of universities.

基于科学计量学的高校科研竞争力对标分析与预测

张岩[1,2,*], 朱艳硕[1],吴曼[1], 周晓梅[1], 殷小川[1], 李光旭[2], 赵玉涛[2], 孙康[1]

[1]青岛科技大学图书馆，青岛，中国

[2]青岛科技大学机电工程学院，青岛，中国

[*]通讯作者

关键词：科研竞争力；大学；科学计量学

摘要： 高校如何依据科技情报科学决策、有效地配置资源，是国家推进"双一流"建设的重要保障。基于Pearson相关性分析和最小二乘线性拟合，对山东省ESI高校的科研竞争力指标进行对比、分析和预测。结果表明，A校在论文数量、质量和总被引次数具有绝对优势；A校、F校和B校的引文影响力优于其余高校；指标的Pearson相关分析结果表明,总被引频次和Web of Science(WoS)论文数、H指数具有极强的相关性；根据预测结果,至2020年C校、E校和D校的年发表WoS论文量可能赶超B校。研究结果有助于高校把握其科研竞争力水平和发展趋势、优化发展政策导向和科研管理，进而推进高校的发展和学科进步。

1．引言

随着我国《统筹推进世界一流大学和一流学科建设总体方案》的提出，为了与学科建设加速发展的需要相适应，包括C9高校在内的ESI（Essential Science Indicators，以下简称ESI）学科快速发展的高校无一例外的都已经利用科学计量学的统计和计算方法对科学活动的投入、产出和过程（如知识传播、交流网络的形成）进行定量分析，试图从中找出科学活动的规律性，助力高校从自身优势出发，为一流师资队伍的建设与引进、拔尖创新人才的培养、科学研究水平的提升、学科发展的规划和高校资源配置的优化提供有力的决策依据，从而提升高校科研创新力和国际竞争力。基于科学计量学对高校科研竞争力进行量化分析，有助于高校认清各学科的定位和差距，从而在“双一流”建设不进则退的大背景下，优化政策导向和科研管理、掌握竞争高校的科研进展和动态，以有限的资源来最大限度的支持和推动学校的发展和学科的进步。

为了促进国内大学的学科发展和竞争力提升，相关学者基于科学计量学对大学及其科研竞争力开展了研究。国外相关学者从不同的角度和层面研究了针对国家和科研机构的科研影响力评价指标[1-3]。现有研究大都基于ESI数据，对学科的引文分析多通过引文分析评价国家、地区或某机构当前或总结过去的发展情况，或者尝试研究基于引文的更为客观的科研影响力评价方法，然而仅基于不同机构的Toppaper与ESI前1%学科排名和数量等指标进行竞争力评价有失偏颇，并且对指标的相关性、对不同数据源的差异性关系以及对未来发展的预测研究较少。基于科研产出对科研活动进行测度，是评估一个机构、国家或地区科技发展水平的视角之一。科学论文是科研产出的主要形式，也是文献计量分析方法进行科研评价的主要依据。清华大学教育研究所的李越等人的研究认为，世界一流大学的学术基准是科研经费、SCI、SSCI论文数量、在Nature和Science上发表的论文数量、教师中院士的人数、诺贝尔奖获得者和学术声誉。在目前的科技评价与测度中，科技论文尤其是SCI论文作为评价科技人员开展科学研究的系统总结，仍然是衡量高校学科科技竞争力的主要统计指标的基础。由此可见，在更为广泛数据范围内，利用更加客观的评价指标进行科研竞争力的分析、评价、跟踪和预测，对于“双一流”建设的推进极为重要。

本文基于科学计量学的方法，通过InCites数据平台对山东省ESI高校的竞争力情况进行对标分析，利用最小二乘法线性拟合方法对2006-2020年各校的WoS（Web of Science）论文发展趋势进行预测，为山东高校的科研竞争力评价和学科发展与管理提供方法参考和决策依据。

2．数据来源

2016年，山东省共有十所高校（包括部属院校和省属院校）进入ESI学科排名前1%，分别为：A校、B校、C校、D校、E校、F校、G校、H校、I校和J校。本文以这十所高校作为研究对象，以InCites数据平台中2006-2015年以山东省十所ESI高校为署名单位的Article和Review类型的论文作为基础对标数据进行分析。数据检索时间为2017年2月19日（InCites数据平台更新时间为2017年1月14日）。

3．科研竞争力的对标分析与预测

3.1 对比指标的选取

为了全面的比较山东省十所ESI高校的科研竞争力水平，从科学计量学的角度，选取论文产出量、总被引频次、学科规范化的引文影响力（Category Normalized Citation Impact，简称CNCI）、被引百分比、高被引论文百分比、被引次数排名前 1% 的论文百分比、被引次数排名前10%的论文百分比、H指数（H-index）、国际合作论文百分比以及论文影响力（Citation Impact）等指标对十所高校的科研竞争力进行对比分析。

其中，CNCI通过文献的实际被引次数除以同文献类型、同出版年、同学科领域文献的期望被引次数获得。对于一篇只被划归为一个学科领域的论文，其CNCI值的计算如式（1）所示：

$$CNCI = c / e_{ftd} \tag{1}$$

式中c为被引次数，e_{ftd}为同文献类型、同出版年、同学科领域文献的期望被引次数。

H指数由美国物理学家Jorge Hirsch在2005年提出，用以量化科研人员作为独立个体的研究成果，其原始定义为一名科学家发表的N_p篇论文中有h篇论文，每篇至少被引h次、而其余N_p-h篇论文每篇被引均小于或等于h次[4-5]。

一组文献的引文影响力指标的计算则是通过使用该组文献的引文总数除以总文献数量n得到的，展现了该组文献中某一篇文献的平均引用次数

$$CI_n = \frac{\sum_{i=1}^{n} Citation_i}{n} \tag{2}$$

引文影响力作为文献计量学指标被广泛应用于科研绩效评价过程中，尽管该指标忽略了科研产出的总体数量，仅从引文角度计算了平均质量。但是由于引文影响力可以被应用于所有的组织层面（作者、机构、国家/区域、科研领域或期刊），因此能够较为全面的衡量一所高校整体的科研绩效水平。

3.2 2006-2015十年WoS论文统计数据对比分析

通过Web of Science检索了山东省10所ESI高校的文献数据及其相关指标，如表1所示。其中，A校以31244篇论文表现出巨大的体量优势；B校和C校均未突破10000篇。从图1给出的十所高校2006-2015年WoS论文发展趋势可以看出，A校、B校、C校、D校和E校从2011年开始历年发表论文数量不断上升，而其余五所高校十年间论文产出量相对较为稳定。

表1 2006-2015年山东省十所ESI高校的科研产出指标对比数据

名称	WoS论文量	CNCI	总被引次数	被引百分比%	HCP %	1%论文%	10%论文 %	国际合作%	H指数	引文影响力
A校	**31234**	**0.97**	**342060**	85.2	0.81	0.94	10.31	22.58	**138**	**10.95**
B校	9006	0.92	92150	84.28	0.83	1.2	9.92	**24.42**	86	10.23
C校	8964	0.78	64588	78.55	0.71	0.87	8.14	19.24	71	7.21
D校	5007	0.93	45397	82.7	0.62	1.14	**11.64**	14.32	68	9.07
E校	4728	0.91	43870	82.09	1.08	**1.35**	9.05	20.81	71	9.28
F校	3952	0.84	42315	83.73	0.58	0.86	10.58	9.13	69	10.71
G校	3341	0.91	30235	**85.33**	0.45	0.57	10.06	13.77	57	9.05
H校	2956	0.71	24669	80.58	0.54	0.58	7.78	12.28	55	8.35
I校	2844	0.72	22062	79.85	0.42	1.16	7.67	6.33	55	7.76
J校	2505	0.92	20463	78	**1.24**	1.32	9.94	13.29	54	8.17

为了深入分析影响高校竞争力的关键因素，利用Pearson相关系数对表1中所示评价指标进行相关性分析，得出因素间的相关性。Pearson相关系数多被来衡量两个指标数据集合能够在多大程度上在一条线上，从而衡量定距变量间的线性关系。对于两个数据集合$\{x_i\}$和$\{y_i\}$其Pearson相关系数为

$$r=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i^2-(\sum x_i)^2}\sqrt{N\sum y_i^2-(\sum y_i)^2}} \tag{3}$$

相关系数r越接近于1或-1，相关度越强，相关系数越接近于0，相关度越弱。

对影响因素的相关分析矩阵进行计算，其结果如表2所示。可以看出总被引用次数和WoS论文数、H指数具有极强的相关性，而与被引次数排名前1%论文百分比相关性极小；CNCI和被引次数排名前10%论文百分比具有很强的相关性；引文影响力和被引百分比具有很强的相关性。为了验证Pearson相关系数的可靠性，利用Spearman相关系数和Kendall相关系数对以上结果进行验证，结果表明三种相关系数的计算结果具有较严格的一致性，如表3所示。

表2 高校的科研产出指标的相关分析矩阵计算结果

指标	WoS论文数	CNCI	总被引次数	被引百分比%	HCP %	1%论文%	10%论文 %	国际合作 %	H指数	引文影响力
WoS论文数	1	0.4352	0.9956	0.4243	0.1329	-0.036	0.2125	0.5903	0.9749	0.523
CNCI	0.4352	1	0.4634	0.5451	0.5247	0.3573	0.8171	0.578	0.5215	0.6088
总被引次数	0.9956	0.4634	1	0.4718	0.1334	-0.03	0.2512	0.5658	0.9788	0.5818
被引百分比%	0.4243	0.5451	0.4718	1	-0.277	-0.274	0.5716	0.3224	0.5292	0.8382
HCP%	0.1329	0.5247	0.1334	-0.277	1	0.6788	0.1715	0.5004	0.1776	0.0732
1%论文%	-0.036	0.3573	-0.03	-0.274	0.6788	1	0.1607	0.2255	0.0564	0.0058
10%论文%	0.2125	0.8171	0.2512	0.5716	0.1715	0.1607	1	0.1856	0.3145	0.6396
国际合作%	0.5903	0.578	0.5658	0.3224	0.5004	0.2255	0.1856	1	0.6575	0.355
H指数	0.9749	0.5215	0.9788	0.5292	0.1776	0.0564	0.3145	0.6575	1	0.6626
引文影响力	0.523	0.6088	0.5818	0.8382	0.0732	0.0058	0.6396	0.355	0.6626	1

表3 高校的科研产出指标的三种相关分析矩阵计算结果一致性对比结果

相关性	WoS论文数	CNCI	总被引次数	被引百分比%	HCP %	1%论文%	10%论文 %	国际合作 %	H指数	引文影响力
Pearson	1	0.4352	0.9956	0.4243	0.1329	-0.0356	0.2125	0.5903	0.9749	0.523
Spearman	1	0.5061	1	0.503	0.3091	0.0182	0.3333	0.8303	0.9512	0.5515
Kendall	1	0.3865	1	0.4222	0.3333	0.0222	0.2889	0.6444	0.8866	0.4667

由以上分析可知，WoS论文总量与总被引次数具有极强的正相关性，而总被引频次直接影响ESI全球综合排名，因此通过观测WoS论文总量可以在一定程度上间接预测机构的科研竞争力发展情况。

3.3 2018-2020年WOS论文趋势预测

在科学统计研究中，通常从一组统计数据通过最小二乘法和回归分析等方法来求得自变量和因变量之间的一个近似解模型，并通过该模型进行预测。最小二乘法是解决数据拟合问题的最重要的方法之一。

假设给定数据点$(x_i,y_i)(i=0,1,\cdots,m)$，$\phi$为所有次数不超过$n(n\le m)$的多项式构成的函数类，现求

$$p_n(x)=\sum_{k=0}^{n}a_k x^k \in \phi \tag{4}$$

使得

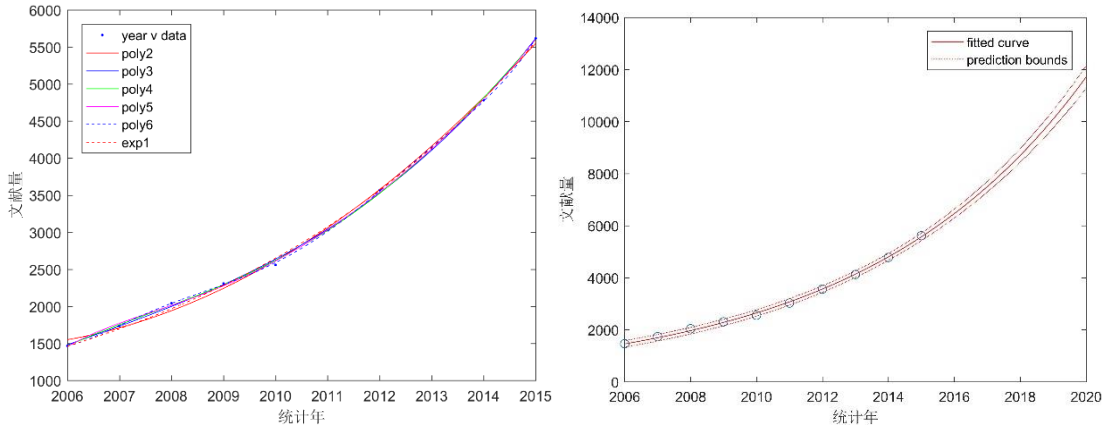
$$I=\sum_{i=0}^{m}\left[p_n(x_i)-y_i\right]^2=\sum_{i=0}^{m}\left(\sum_{k=0}^{n}a_k x_i^k-y_i\right)^2=\min \tag{5}$$

称为多项式拟合，满足上式的 $p_n(x)$ 即为最小二乘拟合多项式。

通过对山东ESI高校的WoS论文数据的最小二乘拟合得到其发展规律的模型，可以在一定程度上预测未来一段时间山东高校的发展方向。首先，对A校2006-2015年WoS论文数据进行了统计、最小二乘拟合多项式拟合并对2016-2020年的发展趋势进行了预测，经过2次至6次多项式线性模型的拟合效果比较，得出如下3次多项式线性模型

$$p(x)=p_1*x^3+p_2*x^2+p_3*x+p_4 \tag{6}$$

其中，$p_1=75.54\ (22.96,\ 128.1)$，$p_2=325.6\ (283.6,\ 367.6)$，$p_3=1225\ (1134,\ 1316)$，$p_4=2830\ (2780,\ 2881)$。该模型调整自由度以后的残差平方为 $AdjustedRsquare=0.9974$，均方根误差为 $rmse=69.6845$，说明模型具有较好的拟合效果。



(a)最小二乘拟合结果　　(b)预测结果

图1 A校WoS论文发展趋势预测

图1给出了A校WoS论文数据的拟合和趋势预测结果，图1（a）是利用指数函数和2次至6次多项式线性拟合的结果，通过拟合误差分析得到图1（b）所示拟合效果最好的指数模型的预测曲线和预测区间。通过指数拟合可以得出在95%置信区间约束下，A校WoS论文随时间的增长的指数模型如下

$$p(x)\ =\ a*exp(b*x) \tag{7}$$

其中，a=4.835e-127，置信区间为(-4.038e-126, 5.005e-126)，b=0.1486，置信区间为(0.144, 0.1533)。从预测结果可以看出，A校近几年的在科研方面具有较为稳定而迅速的发展，十年来WoS论文量呈指数增长，且有望在2020年达到12000篇。

同样，通过拟合分析与预测可以得出其余9所高校的WoS论文发展趋势。对山东ESI高校的WoS论文发展趋势预测结果数据如表4所示，从2020年WoS论文预测量可以看出，A校将继续保持绝对优势，C校、E校和D校有超过B校的可能性。

表4 山东ESI高校的WoS论文发展趋势预测结果

机构名称	2020 年 WoS 论文预测量	预测模型	R-Square 值	WoS 论文稳定性

A校	12000	$p(x) = a*exp(b*x)$	0.9974	较稳定
B校	2500	$p(x) = p_1*x^3 + p_2*x^2 + p_3*x + p_4$	0.9835	较稳定
C校*	5200	$p(x) = a*exp(b*x)$	0.9855	较稳定
D校	2800	$p(x) = a*exp(b*x)$	0.9966	较稳定
E校	3500	$p(x) = a*exp(b*x)$	0.9661	一般稳定
F校	1000	$p(x) = p_1*x^3 + p_2*x^2 + p_3*x + p_4$	0.8821	不稳定
G校	1200	$p(x) = p_1*x^3 + p_2*x^2 + p_3*x + p_4$	0.9918	较稳定
J校	450	$p(x) = p_1*x^3 + p_2*x^2 + p_3*x + p_4$	0.8913	不稳定
I校	350	$p(x) = p_1*x^3 + p_2*x^2 + p_3*x + p_4$	0.1027	较不稳定
H校	550	$p(x) = p_1*x^3 + p_2*x^2 + p_3*x + p_4$	0.9134	一般稳定

表4中，*R-Square*为模型的确定系数，其值由预测数据与原始数据均值之差的平方和除以原始数据和均值之差的平方和确定

$$R-Square = \sum_{i=1}^{n} \frac{w_i(\hat{x}_i - \bar{x}_i)^2}{w_i(x_i - \bar{x}_i)^2} \tag{8}$$

*R-Square*的值域为[0, 1]，越接近1表明模型的变量对数据的表征能力越强，模型对数据的拟合程度越好。

从表4可以看出，A校、D校和G校WoS论文数比较稳定，利用线性方法拟合的模型R-Square值较高，预测结果在当前条件下可信度较高；相比之下部分高校数据由于受到扰动因素的影响呈现出较强的非线性特征，如I校，因此不适合利用线性拟合方法。值得注意的是，当拟合多项式的次数较高时，虽然其均方根误差相对较小，但其正规方程组多为病态方程因此不适用于预测。

4．结束语

在“双一流”建设背景下，高校如何利用大数据挖掘、信息处理和科学计量学方法获得科技情报，使得决策者掌握学校和学科所处的位置以及学科发展趋势，使学校发展的决策更具科学性、资源分配更加合理、有效对科研竞争力发展尤为重要。

基于计量学统计了山东省具有ESI前1%学科的高校的WoS论文量及其影响力指标，对其进行了关键指标对比和分析。利用最小二乘线性拟合法对2016-2020年的WoS论文产出量进行了预测，对预测结果进行了分析和讨论。通过对WoS论文的竞争力相关指标的分析、评价和预测，有助于山东高校了解其发展的现状、优势与不足，为“双一流”建设提供情报和决策依据，促进学科建设的内涵发展，进而提高高校的科学发展和科研竞争力。

References

[1] D. Aksnes, J. Schneider, M. Gunnarsson, Ranking national research systems by citation indicators a comparative analysis using whole and fractionalised counting methods, J. of Informetrics. 6 (2012) 36-43.

[2] L. Bornmann, H. Schier, W. Marx, H. Daniel, What factors determine citation counts of publications in chemistry besides their quality? J. of Informetrics. 6 (2012) 11-18.

[3] T. F. Cova, A. C. Pais, S. J. Formosinho, Iberian universities: a characterisation from ESI rankings, Scientometrics. 94 (2013) 1239-1251.

[4] J. Hirsch, An index to quantify an individual’s scientific research output, Proc. of the National Academy of Sciences of the United states of America. 102 (2005) 165-169.

[5] E. Csajbók, A. Berhidi, L. Vasas, A. Schubert, Hirsch-index for countries based on essential science indicators data, Scientometrics. 73 (2007) 91-117.