arXiv:2302.05720v1 [cs.DL] 11 Feb 2023

科学计量学指标的 Gintropic 缩放

Tam´as S. Bir´o[a,b,c], Andr´as Telcs[b], M´at´e J´ozsa[a], Zolt´an N´eda[a]

[a]巴贝什-博雅伊大学物理系，克卢日-纳波卡，罗马尼亚[b]维格纳物理研究中心，布达佩斯，匈牙利 c 复杂科学中心、维也纳医科大学，奥地利

摘要

科学家生产力和影响力最常用的指标是出版物总数（N_{pub}）、引文总数（N_{cit}）和赫希（h）指数。自赫希于 2005 年发表开创性论文以来，h 指数能否被视为独立于 N_{pub} 和 N_{cit}的指标，一直备受争论。本文利用研究人员所著论文引文分布的帕累托形式，探讨 h、N_{pub} 和 N_{cit}之间的标度关系。分析将基尼指数作为引用分布的不平等度量，并纳入近期提出的不平等核——基尼熵（与熵核相似）。我们发现了一个以引文总数为自变量的 h 值新上界，并通过从 Google Scholar 收集的大规模数据加以验证。我们的分析还显示，根据作者出版物所得引文计算的个体化基尼指数峰值约为 0.8，远高于通常社会经济不平等所具有的数值。

关键词：标度、H 指数、数据挖掘、洛伦兹曲线、科学计量学

1. 引言

为了表征研究人员或机构的科学生产力，人们以已发表的研究论文／书籍作为代理指标，而这些成果的影响力则通过引文来衡量。众所周知，生产力与影响力并非完全相互独立，因为没有生产力就不可能产生影响力。为了构建一个同时衡量生产力和影响力的简单指标，阿根廷物理学家 Jorge Hirsch 提出了一种简单的标量指标，如今称为“赫希指数”，或简称“h 指数” [2]。不过，有人可能只有一篇论文，却获得了诺贝尔奖；另一些人可能发表了数千篇引文表现较弱或平平的论文，但在赫希指数方面仍超过前者。因此，赫希指数本身并不足以作为科学绩效的充分指标。在科学计量学领域，一个由来已久的问题是：三个最简单、最知名的指标——出版物数量（N_{pub}）、引文总数（N_{cit}）和“h指数”（h）——是否通过标度定律在统计上相互关联 [3]。

在一篇近期论文中，Siudem 等人 [4] 基于第一性原理构建了一个增长模型：即“优先连接”和“富者愈富”，也就是马太效应。他们的模型与经验数据高度吻合。他们认为，科学影响力有三个维度。具体而言，他们将赫希指数视为一种广泛使用的指标，但认为它不足以准确衡量所研究的研究人员、期刊或研究所的科学成就。基于马太效应的科学计量学动态建模可追溯至 Schubert [27]。Barab´asi 将“富者愈富”原则应用于网络演化，并将其重新命名为优先连接。Siudem [4] 在其模型中结合了这两种机制，并将马太效应开启与关闭的概率作为第三个因素。事实上，即使不进行任何建模，也有人得出了类似的观察结果（参见 [5] 及其中的参考文献）。

偏好转换概率在建模中被用作第三个因素，但它并非真正的“科学计量学”指标。尽管如此，Sidum模型颇具说服力，并且与多个科学领域的经验数据高度吻合。Siudem [4] 还作为副产品表明，单一指标，尤其是赫希指数，不足以准确衡量科学成就。

在后续论文（[8, 9, 12] ）中，进一步提出了对赫希指数及其他科学指标进行三变量回归的出色改进。

下文将提出一个理论框架，把赫希指数作为第三项科学计量指标，与生产力（出版物）和影响力（引文）并列，用于描述科学成就。更多

具体而言，我们借鉴基尼指数 [14] 和基尼熵 [15], 等不平等度量的基本概念，寻找有关赫希指数标度的统计证据，并使用从 Google Scholar 收集的数据验证这一方法。

科学引文及其统计数据一直是科学计量学领域深入研究的对象 [16]。h指数表示一位作者拥有的出版物数量 h，使得其中每篇出版物至少有 h 条引文。赫希的研究成果 [2] 如今在 Google Scholar 上获得了超过 13,000 条引文，这体现了它受到科学界的关注。关于 h 指数，人们提出了许多批评，也发现了它与其他指标之间的联系，这一主题由此发展成科学计量学中的一个独立分支（综述见 [17, 18]）。除 h指数外，科学计量学还使用许多其他指数来描述科学产出。这里可以举出一些广为人知的指数，但不求面面俱到，例如“R指数”、“m指数”、“i-10指数”、“g指数”等 [19]。如今人们已认识到，没有任何指数是完美的；即使在某个分支领域内，也不应根据这些粗略的数字给研究人员贴标签 [20]。尽管如此，科学计量指标仍是一种客观工具，可补充更个性化但有时也带有主观性的评价。为了进一步简化将粗略数字用作评价工具的过程，理解这些指标彼此独立还是相互关联至关重要。

如果不将研究范围限定在狭窄的科学子领域内，那么已发表研究论文的总数、引文数、独立引文或施引论文的数量、h指数，甚至将这些指标按作者数量进行标准化后得到的数值，都不是可靠的科学计量学测度。然而，从纯粹的统计研究角度来看，考察涵盖多个科学领域的大量研究人员群体是有意义的。这类研究关注科学领域发表习惯的宏观、粗粒度图景，从纯粹的统计视角来看是合理的。人们早已推测，N_{cit} 和 h 指数在统计上存在关联 [3]。研究人员、期刊或机构所发表出版物的论文数量和引文数量的分布存在普适模式，这一事实已得到确认 [21, 22], ，表明人们可以进一步探寻主要科学计量学指标与生产力指标之间的统计联系。在以下各节中，我们将在简要介绍研究所采用的测度后，从不平等的角度分析引文，以揭示这类联系。

2. 基尼指数、累积分布与基尼熵

基尼指数由科拉多·基尼提出，旨在用一个介于零和一之间的单一指标，概括财富不平等的信息 [14]。其定义很简单：两个数据点之差的绝对值的平均数，除以它们之和的期望值。对于定义在 $x \geq 0$ 上的给定概率密度函数（概率密度函数），$\rho(x)$，它可以写成如下积分之比：

$$G \equiv \frac{\langle\, |x-y| \,\rangle}{\langle\, |x+y| \,\rangle} = \frac{1}{2\,\langle x\rangle} \int_0^\infty dx \int_0^\infty dy\; |x-y|\, \rho(x)\, \rho = \frac{1}{\langle x\rangle} \int_0^\infty dx \int_x^\infty (y-x)\rho(x)\rho(y). \tag{1}$$

基变量的期望值用尖括号表示：$\langle \cdot \rangle$。

这个定义有几种等价形式。使用

$$\overline{C}(x) \equiv \int_x^\infty \rho(y)\, dy, \tag{2}$$

and

$$\overline{F}(x) \equiv \frac{1}{\langle x\rangle} \int_x^\infty y\, \rho(y)\, dy, \tag{3}$$

可得

$$G = \int_0^\infty \rho(x) \left[\overline{F}(x) - \frac{x}{\langle x\rangle} \overline{C}(x) \right] dx \tag{4}$$

通过一些直接的数学推导，还可以得到基尼指数的其他表达式，这些表达式体现了面积的几何意义：

$$G = \int_0^\infty dx\, \overline{F}(x)\,\rho(x) - \int_0^\infty dx\, \frac{x}{\langle x\rangle}\overline{C}(x)\,\rho(x) = \int_0^1 \overline{F}\, d\overline{C} \;-\; \int_0^1 \overline{C}\, d\overline{F}. \tag{5}$$

基尼指数 c　　也可以通过累积人口来表示，其形式具有低端与高端对称性：

$$G \;=\; \frac{1}{\langle x\rangle}\int_0^\infty \overline{C}(x)\,\Big(1 - \overline{C}(x)\Big)\, dx. \tag{6}$$

请注意，对于标度后的概率密度函数，$\rho(x) = \frac{1}{\langle x\rangle} f\left(\frac{x}{\langle x\rangle}\right)$，C、F 和 G 不直接依赖于 ⟨x⟩。由概率密度函数构建的 F − C 映射，即 $\rho(x)$，就是洛伦兹曲线 [23], ，如图1所示。

基尼熵 [15, 24] 的构建旨在财富与收入分配研究中计算基尼指数时，突出其类熵性质。它基于洛伦兹曲线几何，该几何描述了由累积人口轴和累积财富轴构成的数据平面中的特定区域。其定义如下：

$$\sigma = \overline{F} - \overline{C}, \tag{7}$$

其几何图示见图1。根据洛伦兹曲线几何，我们得到：

$$\int_0^1 \sigma\, d\overline{C} \;=\; \int_0^1 \sigma\, d\overline{F} \;=\; \frac{1}{2}G. \tag{8}$$

这里回顾基尼熵的一般类熵性质。

1. 基尼熵始终非负：
   $\sigma = \overline{F} - \overline{C} = C - F \geq 0$. This is obvious from inspecting the integral $\sigma = \int_x^\infty \left(\frac{y}{\langle x\rangle} - 1\right)\rho(y)dy = \int_0^x \left(1 - \frac{y}{\langle x\rangle}\right)\rho(y)dy \geq 0$. One can take the first form for $y \geq x \geq \langle x\rangle$ and the second form for the opposite case.
2. $\sigma(x)$ 在 $x = \langle x\rangle$ 处达到最大值。考察 $d\sigma(x)/dx = (1 - x/\langle x\rangle)\rho(x)$，$\rho \geq 0$，即可直接确定 $\sigma(x)$ 取得最大值的位置。
3. $\sigma(x)$ 是一条凹曲线。考虑到 $d\sigma\, dC = x\langle x\rangle^{-1}$，我们得到：

$$\frac{d^2\sigma}{d\overline{C}^2} = \frac{1}{\langle x\rangle}\frac{dx}{d\overline{C}} = -\frac{1}{\langle x\rangle\,\rho(x)} \;<\; 0. \tag{9}$$

4. 对于许多在实际中具有重要意义的概率密度函数，$\sigma(C)$ 公式看起来像是基于 $\rho(x)$ [15] 的熵项。

3. Hirsch 指数缩放

3.1. 一般考量

赫希指数（或简称“h”指数） [2], 被推荐用于评估个人、期刊和机构的科学影响力，其定义为至少获得同等数量引文的出版物数量：

$$h \;=\; N(x \geq h) \;=\; N_{\text{pub}}\,\overline{C}(h). \tag{10}$$

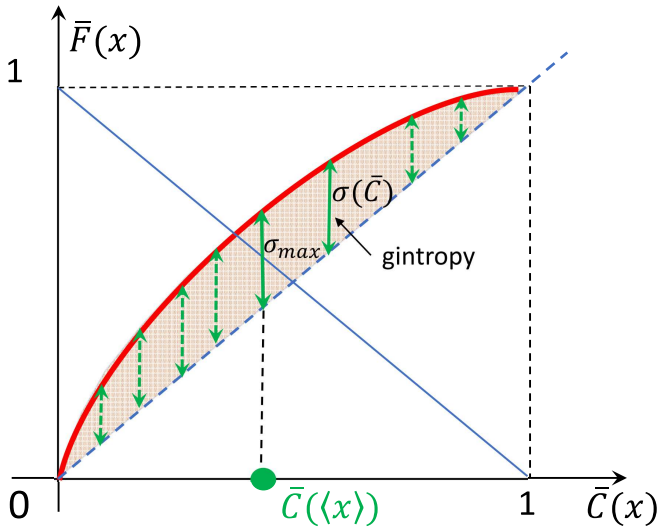

图1：累计引文数 F 与被引次数至少为 x 的出版物的累积比例 C 的关系图——这也称为洛伦兹曲线。此图还直观地展示了基尼熵 σ 和基尼指数 G： σ 是洛伦兹曲线与对角线之间的距离，G/2 是两条线之间的面积。

对于给定的概率密度函数 $\rho(x)$，赫希指数 h 是上述方程的解，在大多数常见情况下，该解是超越的。

某些分布具有特殊性。只要它们体现了某种潜在的标度性质，我们就预期赫希指数也会相应地受到某种标度关系的影响。问题在于，如何构造才能使这种标度关系显现出来。本文旨在构建赫希指数 h、出版物总数 N_{pub}以及这些出版物的引文数 N_{cit}之间的标度关系。注意， $\langle x \rangle = N_{cit}/N_{pub}$。

在风险分析中，典型的累积函数形式如下

$$\bar{C}(x) = e^{-H(x)}. \tag{11}$$

在这种情况下，概率密度函数 $\rho(x) = -dC/dx$ 变为

$$\rho(x) = H'(x)\, e^{-H(x)}. \tag{12}$$

这里，H(x) 称为累积风险，而其导数 H′(x) 称为风险率 [29]。

用于拟合统计数据的概率密度函数通常有两个常见参数，分别表示自变量的平移和缩放。这两个参数可以与均值和变异联系起来。作为公式，可以采用（或猜测）一个以自变量的线性形式为输入的普适函数。假设

$$H(x) = f(ax + b) \tag{13}$$

对于一个无参数函数，概率密度函数会自动满足归一化条件。参数 a 和 b 可以与期望值 $\langle x \rangle$ 联系起来，这样剩下的问题就只包含一个参数。总可以选择 f (x)，使得 $a > 0$。

使用公式 (13) 及其逆变换，

$$x \equiv \frac{f^{-1}(H) - b}{a}, \tag{14}$$

可确定以下普适值

$$\kappa(b) \equiv a\langle x \rangle + b = \int_{f(b)}^{\infty} f^{-1}(H) e^{-H}\, dH. \tag{15}$$

累积风险以及所有其他与概率密度函数相关的函数，都仅由 b 参数化，并且是 x/⟨x⟩ 的函数。因此，这种标度性质具有相当的普遍性。

$$H(x) = f\left([\kappa(b) - b]\frac{x}{\langle x\rangle} + b\right). \tag{16}$$

总之，一类非常重要的概率密度函数呈现出标度

$$\rho(x) = \frac{1}{\langle x\rangle}\, p\left(\frac{x}{\langle x\rangle}\right). \tag{17}$$

在这些情况下，累积积分 C 和 F，因此也包括基尼熵 $\sigma \equiv$ F −C，仅取决于相对于期望值的比值 x/⟨x⟩，以及参数 b。因此

$$\overline{C}_b(h) = c_b\left(\frac{h}{\langle x\rangle}\right). \tag{18}$$

知道 当 $\langle x\rangle = N_{cit}/N_{pub}$ 时，赫希指数是利用累积分数：

$$\frac{h}{N_{pub}} = \mathrm{e}^{-f\left([\kappa(b)-b]\frac{h}{N_{pub}}\frac{N_{pub}^2}{N_{cit}}+b\right)}, \tag{19}$$

因为

$$\frac{h}{\langle x\rangle} = \frac{h}{N_{pub}}\frac{N_{pub}^2}{N_{cit}}. \tag{20}$$

由此可知，N_{cit}、N_{pub} 与 h 指数之间存在标度关系，该关系基于普适函数 f、参数 b 及其特殊积分 κ(b)：

$$\frac{\sqrt{N_{cit}}}{N_{pub}} = \sqrt{[\kappa(b) - b]\frac{h/N_{pub}}{f^{-1}[-\ln(h/N_{pub})] - b}}. \tag{21}$$

通过多种方式可以得到 h指数的其他标度关系。这些关系并非独立于上述结果。我们推导出，h/⟨x⟩ 满足一个一致性关系

$$\frac{h}{\langle x\rangle} = \frac{N_{\mathrm{pub}}^2}{N_{\mathrm{cit}}}\, c_b\left(\frac{h}{\langle x\rangle}\right) \tag{22}$$

对于固定的 b 值，该比值的解取决于单一参数[1]

$$\lambda = \frac{N_{\mathrm{pub}}^2}{N_{\mathrm{cit}}}. \tag{23}$$

我们注意到，$\lambda\langle x\rangle = N_{\mathrm{pub}}$，因为 $\langle x\rangle = N_{\mathrm{cit}}/N_{\mathrm{pub}}$ 是所研究分布的平均引文数。现在比较赫希指数关系的两种形式，即方程（19）和（22）

$$\frac{h}{\langle x\rangle} = \lambda c_b\left(\frac{h}{\langle x\rangle}\right), \qquad \frac{h}{N_{\mathrm{pub}}} = c_b\left(\lambda\frac{h}{N_{\mathrm{pub}}}\right), \tag{24}$$

可以看出，h/⟨x⟩ 和 h/N_{pub} 都是单一参数组合 λ 的函数。

将上述两个标度定律相乘，可得已知形式：

$$\frac{h^2}{N_{\mathrm{cit}}} = \lambda c_b\left(\frac{h}{\langle x\rangle}\right)^2 = \frac{h}{\langle x\rangle}c_b\left(\frac{h}{\langle x\rangle}\right). \tag{25}$$

结果是，对于常数参数 b，比值 h^2/N_{cit} 仅取决于比值 h/⟨x⟩，与基尼熵的情况一样。由于基尼熵存在最大值，因此也可以由此推导出 h^2/N_{cit} 比值的不等式。根据 [3], ，对于大多数数据集

$$\sqrt{N_{\mathrm{cit}}} \approx 2h \tag{26}$$

[1]方程（21）显示了 $1/\sqrt{\lambda}$ 随 N_{pub}变化的情况。

3.2. 帕累托分布的标度

标度 Tsallis-Pareto 分布 [25] 是满足式（16）所隐含形式的分布的一种特例。文献 [22]证明了这一形式适用于个人引文分布。其概率密度函数如下：

$$\rho(x) = a\,b\,(1+ax)^{-b-1} \tag{27}$$

帕累托情形下的尾部累积积分为：

$$\overline{C}(x) \;=\; (1+ax)^{-b} \tag{28}$$

$$\overline{F}(x) \;=\; \frac{(1+a\,b\,x)}{a}\,(1+a\,x)^{-b} \tag{29}$$

而基尼熵为

$$\sigma(x) \;=\; ab\,x\,(1+ax)^{-b}. \tag{30}$$

基尼 指数也可以针对 Tsallis-Pareto 分布进行计算。通过简单的数学推导可得

$$G_{TP} = \frac{b}{2b-1}, \tag{31}$$

这似乎是刻画引文分布的概率密度函数的一般规律。

对于 Tsallis-Pareto 分布，(25) 标度关系写为：

$$\frac{h^2}{N_{\rm cit}} \;=\; \frac{h}{\langle x\rangle}c_b\left(\frac{h}{\langle x\rangle}\right) = \;\frac{h}{\langle x\rangle}\overline{C}_b(h) \tag{32}$$

注意到 $1/\langle x\rangle = a(b-1)$，并结合 (29) 和 (30)，我们得到

$$\frac{h^2}{N_{\rm cit}} \;=\; \frac{b-1}{b}\,\sigma(h). \tag{33}$$

现在，[15] 中已证明基尼熵存在最大值，因此

$$\frac{h^2}{N_{\rm cit}} \le (1-1/b)\,\max(\sigma). \tag{34}$$

剩下的工作是求出标度 Tsallis-Pareto 分布的基尼熵最大值。考虑到当 $x = \langle x\rangle$ 时，基尼熵达到最大值，可得：

$$\max(\sigma) = \sigma(\langle x\rangle) = \frac{b}{b-1}\left(1+\frac{1}{b-1}\right)^{-b}. \tag{35}$$

利用方程 (34)，我们得出：

$$\frac{h^2}{N_{\rm cit}} \le \left(1+\frac{1}{b-1}\right)^{-b}. \tag{36}$$

根据 h 指数的定义，N_{cit}、N_{pub} 与 h 之间的标度关系为：

$$\frac{h}{N_{pub}} = \overline{C}(h) \;=\; \left(1+\frac{h\,N_{pub}}{(b-1)\,N_{cit}}\right)^{-b} \tag{37}$$

这使得式 (21) 所提出的标度呈现出如下特殊形式

$$\frac{\sqrt{N_{cit}}}{N_{pub}} = \sqrt{\frac{h/N_{pub}}{(b-1)\left[(h/N_{pub})^{-1/b}-1\right]}} \tag{38}$$

4. 在 Google Scholar 数据上的测试

根据我们之前的研究 [22], , 引文分布与脸书分享的分布相似，并遵循共同的 Tsallis-Pareto 分布进行标度，其中 b = 1.4。我们认为，这种现象的原因在于引文中的优先动态以及出版物数量（脸书帖子）的指数增长。本文对大样本中个体科学家的赫希指数（h）、引文数量（N_{cit}）与出版物总数（N_{pub}）之间的关系进行了数值研究。我们还旨在验证关于引文分布概率密度呈 Tsallis-Pareto 形状的假设，并重新考察标度指数 b 的普适性。

数据通过网络爬虫机器人收集，该机器人对 Google Scholar 网站进行映射。

在收集引文时，我们从一位高度连接的作者（Prof. H.E. Stanley）开始，并递归地映射其合著者网络。通过这种方式，我们收集了 44 360 名研究人员的相关数据，其中 $N_{cit} \geq 10000$ 和 $N_{pub} \geq 100$。设置这些限制是为了获得足够的数据来构建概率密度函数。对于每位研究人员，Tsallis-Pareto拟合均自动完成，覆盖 $b \in (1, 8)$ 值，并以 $a\Delta b = 0.025$ 为步长。我们在此寻找能够使平均相对差值平方达到最小的 b 值：

$$s = \frac{1}{W}\sum_{i}^{W}\left(\frac{\log[\overline{C}_i^{exp}] - \log[\overline{C}_i^{theo}]}{\log[\overline{C}_i^{exp}]}\right)^2, \tag{39}$$

这里的 W 表示用于量化该分布的实验点数量。我们使用累积分布函数 C(x)，以获得更加平滑的实验数据分布。进行对数变换是为了准确拟合分布的尾部，同时确保所有数据点具有相同的数量级。

图 2 首先展示了我们随机选取的几位引文数量相差很大的研究人员，说明使用我们的方法确定的 b 指数时，Tsallis-Pareto 概率密度函数对其引文的适用性。这里，我们通过考察相对于均值的数值，并将$\rho($x/⟨x⟩$)$绘制为 x/⟨x⟩ 的函数，将引文分布的概率密度函数归一化。

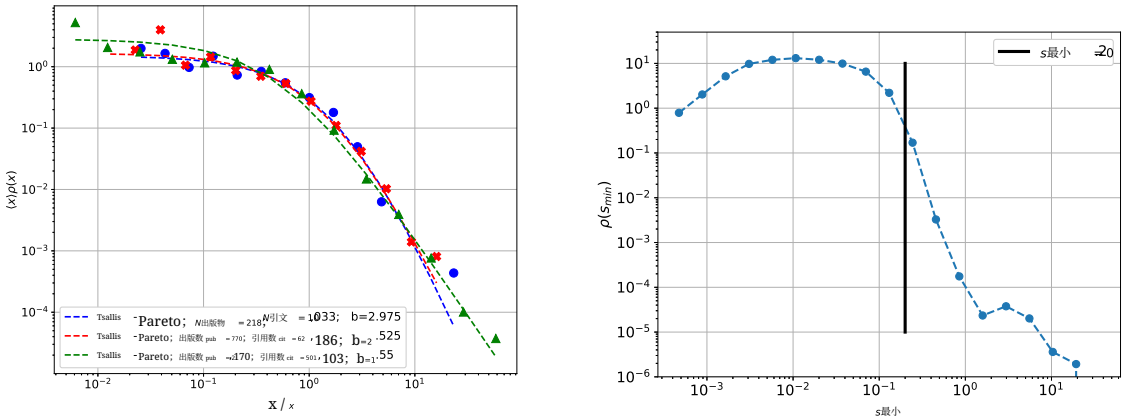

图 2：Tsallis-Pareto 形式对引用分布的有效性。上图展示了三位 N_{cit} 数量差异很大的研究人员的情况。随着统计数据变得更加充分，Tsallis-Pareto拟合也会有所改善。下图展示了 s_{min} 参数分布的概率密度函数，表明了 Tsallis-Pareto拟合对不同研究人员的拟合优度。垂直线表示本研究所接受的最大限值。

拟合优度可用最小 s 值表示，记作 s_{min}。所得 s_{min} 值的分布在图 2 下方以双对数坐标轴展示。

从图中可以看出，该分布具有长尾。因此，我们将上限截断值设为 $s_{min} = 0.2$，并在后续研究中忽略 Tsallis-Pareto拟合可靠性较低的少数情况。

此外，我们也忽略了最佳拟合超出所设 b = 8 限值的情况。在最初研究的 44360 名研究人员中，最终保留了 43656 条记录。

这些研究人员拟合的 b 值分布如图 5 所示，其中我们再次使用了双对数坐标轴。该分布在 b_{max} = 1.32 附近有一个尖锐峰值，并表明 ⟨b⟩ = 1.58。最大值与平均值之间的区间与我们在先前假设[22]中提出的标度指数吻合良好。

利用通过筛选条件的研究人员数据，我们现在可以研究上一节所述标度关系的有效性。

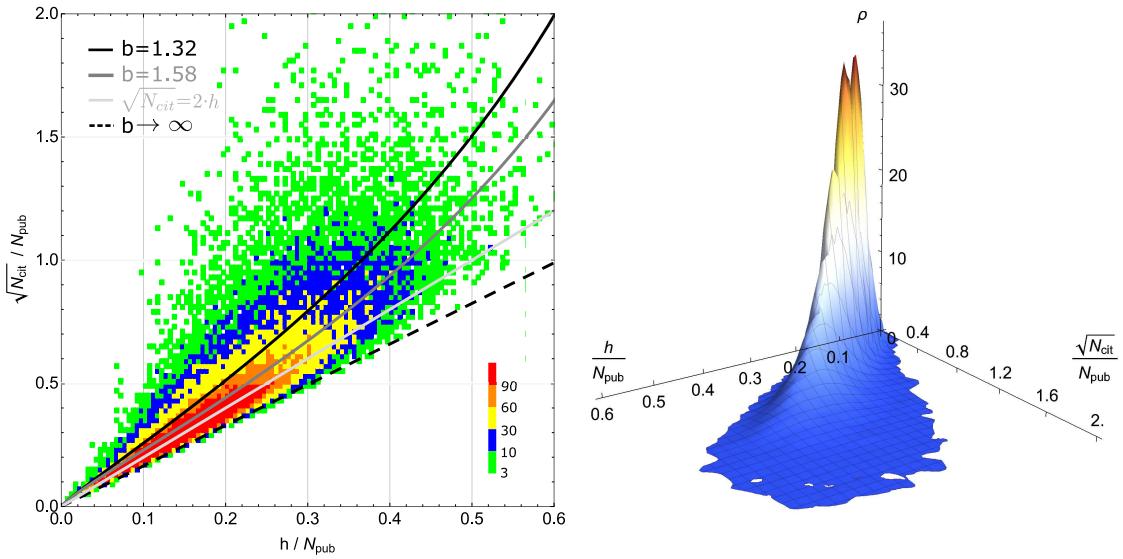

图 3：h、N_{cit} 和 N_{pub}之间标度关系的有效性与局限性。在二维和三维表示中，我们展示了坐标为（$\sqrt{N_{cit}}/N_{pub}$，h/N_{pub}）的点的分布。对于二维表示，使用图例中给出的颜色编码来表示各点的密度。不同的线表示各种 b 帕累托指数对应的标度关系，以及通常被接受的 $\sqrt{N_{cit}} = 2h$ 趋势。

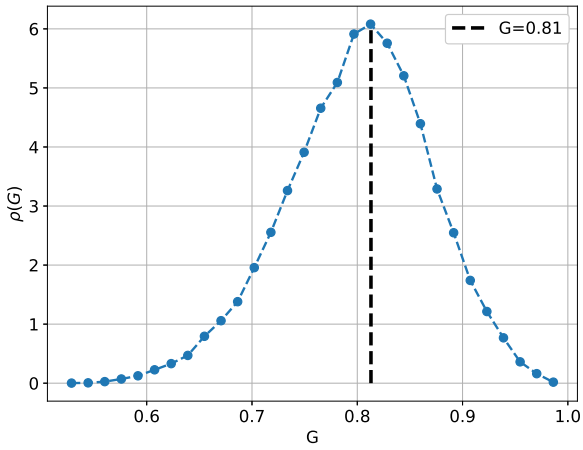

图 4：所研究研究人员的个体化基尼系数值分布

我们首先来检验公式（38）所示的 N_{cit}、N_{pub} 与 h 指数之间的标度关系。为展示这种标度关系，我们绘制了所有 43656 条记录中 $\sqrt{N_{cit}}/N_{pub}$ 随 h/N_{pub} 变化的情况。图 3 中标出了这些点的密度

图中通过两种方法呈现了密度。上图采用二维密度图，下图则采用三维平滑直方图表示。从密度图可以清楚看出，对于较小的 h/N^{pub} 值，$\sqrt{N_{cit}} \propto h$，这证实了公式 26 所表达的标度关系有效。此前提出的标度关系 [3]、$\sqrt{N_{cit}} = 2h$ 对我们的绝大多数结果都适用。此图还表明，只要考虑 $b = \infty$ 极限，由基尼熵最大值推导出的界限便适用。由此可得一个普遍有效的标度关系，并为给定 h 值时引文最小数量提供了一个有趣的界限。

$$\sqrt{N_{cit}} \geq h \lim_{b\to\infty}(1 - 1/b)^{-b/2} = h\sqrt{e} \tag{40}$$

这一条件 $N_{cit} \geq eh^2$, 确实比显而易见的 $N_{cit} \geq h^2$, 强得多，也比先前提出的 $N_{cit} = 4h^2$, 弱；后者仅适用于主要趋势。三维图（图 3b）更有力地表明，数据偏离主要趋势的范围实际上很窄。

正如我们先前提到的，Tsallis-Pareto 型分布意味着基尼系数满足：$G > 1/2$。对于所有选定研究人员的出版物，我们计算了其所获引文的个体化基尼指数。正如预期，我们发现该指数呈现较宽的分布，峰值位于 $G = 0.81$，对应于 $b \approx 1.3$，表明引文数量存在很大差异。图 4 所示的概率密度函数呈现出相当宽的分布，而我们发现最初设定的下界 $G = 1/2$ 是成立的。

5. 讨论与结论

科学绩效的维度不仅具有理论意义。科学指标的收集与分析成本高昂且复杂。减少所需指标的数量，吸引了科学计量学研究的关注（例如参见 [5] 及其参考文献）。有两种方法：统计方法和建模方法（再次参见 [5, 4]）。本文采用前一种方法，特别关注将赫希指数纳入指标集合。难点在于，赫希指数的定义归结为一个无法显式求解的函数方程。已有若干研究采用参数假设，以便为赫希指数找到良好的近似值。其中包括 Glanzel 在 [6], 提出的帕累托分布、Bertoli-Barsotti 和 Lando 在 [7], 提出的几何分布，以及 [10] 采用的威布尔分布（另见综述 [11]）。按照（10）和（12）计算 H 的超越方程数值解，见[8]。

我们的首要目标是获得 h 指数的一般非参数标度关系。为此，并为了利用基尼熵与引文分布的生存描述在形式上的类比及其良好性质，我们首先简要回顾了基尼指数和基尼熵。随后，借助 gintropy 形式体系推导出了非常一般但隐含的 h 指数标度关系（21）。接着，我们采用引文遵循 Tsallis-Pareto 分布这一假设，获得了更为明确的表达式。位置参数的选择以及对论证有效性的检验，是基于 Google Scholar 数据完成的。除了推导赫希指数的标度关系外，本文还为 h 指数给出了一个新的、更严格的上界（40）。

我们在先前研究中揭示的普适引文模式 [21, 22], 在此通过一项基于从 Google Scholar 收集数据的统计分析研究得到了证实。在统计分析中不局限于所研究的领域时，我们发现，拟合得到的帕累托指数分布接近先前提出的 $b = 1.4$ 值（最大值位于 $b_{max} = 1.32$，均值位于 $\langle b\rangle = 1.58$），但同时也呈现出重尾（图 5）。

这一发现表明，对于个别作者出版物的引文计算得出的个体基尼指数应满足 $G > 1/2$。这确实与 Google Scholar 中观察到的数据一致。不过，令人惊讶的是，我们发现，考察所有相关作者的这些个体 G 值，其分布近似于正态分布，并在很高的数值 $G_{max} \approx 0.81$ 处达到峰值。这表明，研究人员的出版物所获引文数量存在显著不平等。此外，研究人员撰写的文章的引文分布呈帕累托–齐利斯型形式，由此可得出若干有趣的关联和极限，涉及出版物总数、引文总数和 h 指数。由分布函数的帕累托型形状推导出的关系式 (38)，与从 Google Scholar 收集的数据一致。从图 3 的数据三维表示中可以看出，数据点的分布非常集中，因此这种标度关系确实应当成立。然而，令人惊讶的是，仔细查看图 3 后可以发现，先前提出的标度关系 [3] $\sqrt{N_{cit}} = 2h$ 比

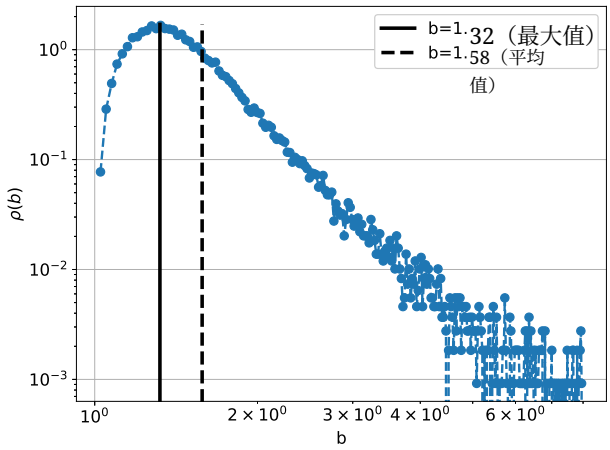

图 5：通过设定的筛选条件的 43656 名研究人员的拟合的 b 值分布。

使用 b_{max} or $\langle b \rangle$ 得出的结果更贴合该分布的峰值。基尼熵作为近期提出的一种新型不平等度量 [15], , 已证明有助于确定 h 关于 N_{cit} 的合理极限。通过这一论证得出的界限 $N_{cit} \geq e\, h^2$, , 得到了所用 Google Scholar 数据的有力印证。

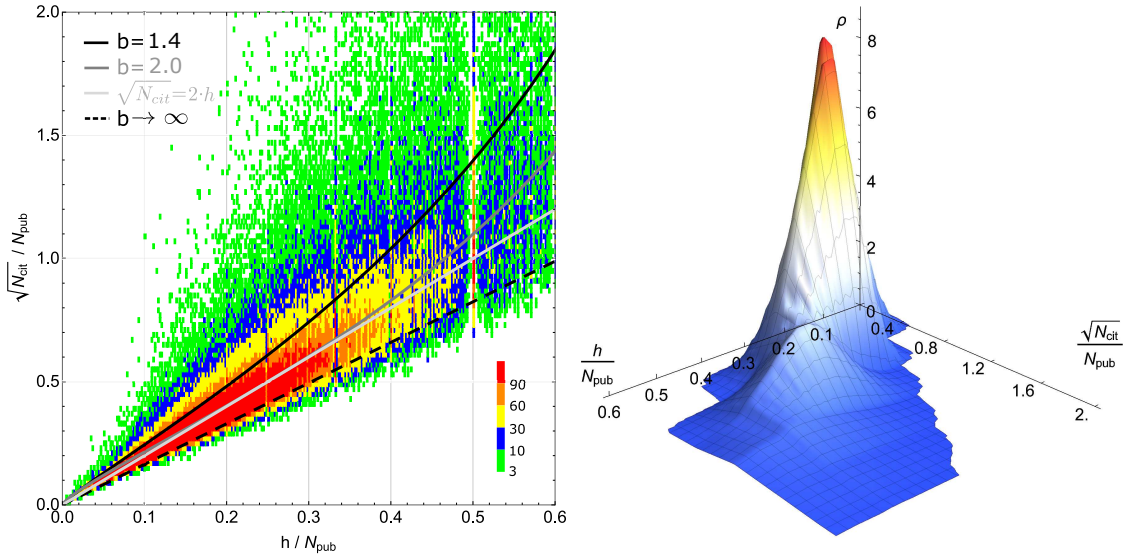

图6：在考虑 $N_{pub} \geq 10$ 和 $N_{cit} \geq 100$的限值时，二维和三维表示中坐标为 ($\sqrt{N_{cit}}/N_{pub}$, h/N_{pub}) 的点的分布。可与图 3 对照

最后，应注意，我们的统计分析仅针对生产力较高（$N_{pub} \geq 100$）且引文数量较多（$N_{cit} \geq 10000$）的研究人员进行。我们的初始假设本就基于这样一个事实：引文的分布函数可以用 Tsallis-Pareto 形式描述。为了了解我们的结果适用范围的限制，值得考虑将其推广到生产力和影响力低得多的研究人员。考虑到限制条件 $N_{pub} \geq 10$ 和 $N_{cit} \geq 100$，并据此将通过我们的爬虫程序递归映射的研究人员中超过 96% 纳入统计，现将与图 3 对应的结果呈现在图6中。扩展统计表明，点 ($\sqrt{N_{cit}}/N_{pub}$, h/N_{pub}) 的分布更加

分散，但标度 $\sqrt{N_{cit}} = 2h$ 仍遵循主要趋势。限制条件 $\sqrt{N_{cit}} \geq \sqrt{e}h$ 不太明显，有 4.31% 的记录违反了该限制。相比之下，在引文数量较多的研究人员中（$N_{pub} \geq 100$ 和 $N_{cit} \geq 10000$），只有 0.16% 的研究人员违反了这一限制。所有这些结果都证实了我们最初的工作假设：提出的限制条件和标度基于引文的帕累托型分布，并且是在可以构建此类分布函数的假设下推导出来的。因此，对于大多数出版物和引文数量较少的研究人员，应谨慎看待这些结果。

当然，也可以利用 Web of Science 或 Scopus 等科学性更强的数据库进行更恰当的数据分析；不过，为了克服这些数据库严格的用户政策，此类研究应采用更复杂的方法。这正是我们在此使用免费且可能不那么严谨的 Google Scholar 数据的主要原因。

致谢

本研究得到 UEFSCDI 资助，资助合同编号为 PN-III-P4-ID-PCE-2020-0647。T.S.B. 感谢 NKFIH 通过匈牙利国家实验室计划资助相关研究，项目编号为 2022-2.1.1-NL-2022-00002；该计划除其他指标外，还要求提高引文数量和出版物方面的生产力。M.J. 的工作得到匈牙利 Collegium Talentum计划的资助。

利益冲突

作者声明不存在利益冲突。

作者贡献

TSB 和 ZN 负责概念构思。TSB 和 AT 负责分析。MJ 和 ZN 负责数据挖掘与处理。TSB、ZN 和 AT 对论文初稿的贡献相同。

参考文献

[1] A. Schubert, & W. Gl¨ **6**, anzel, 149 167 1984 《偏态分布类的动态考察：一个具有科学计量学应用的模型》。科学计量学 - ( )。[2] J.E. Hirsch, 《量化个人科学产出的指数》。PNAS 102, 16569-16572 (2005)。[3] W. Gl¨ **67**, anzel, 315 321 2006 《论 h 指数——一种衡量出版活动与引用影响力的新指标的数学方法》。科学计量学 - ( )。[4] G. Siudem, B. ˙Zogała-Siudem, A. Cena 和 M. Gagolewski, 《科学影响力的三个维度》。PNAS 117, 13896-13900 (2020)。[5] G. Prathap, 《致编辑的信：对 Gagolewski 等人论文〈奥卡姆引用影响力指数〉的评论》，科学计量学 127, 6051 6054 2022- ( )。[6] W. Gl¨anzel, 《统计学在 h 指数相关领域的一些新文献计量学应用》。科学计量学 77, 187-196 (2008)。[7] L. Bertoli-Barsotti 和 T. Lando, 《h指数公式》。信息计量学杂志 9, 762-776 (2015)。[8] L. Bertoli-Barsotti 和 T. Lando, 《h指数与其他简单引用指标之间关系的理论模型》。科学计量学 111, 1415-1448 (2017)。[9] L. Bertoli-Barsotti 和 T. Lando, 《平均排名和平均规模如何决定广义洛伦兹曲线：及其在引用分析中的应用》。信息计量学杂志 13, 387-396 (2019)。[10] L. Bertoli-Barsotti 和 T. Lando, 《h指数作为若干基本统计量的近似精确函数》。科学计量学 113, 1209–1228 (2017)。[11] L. Egghe 和 R. Rousseau, 《h指数形式体系》。科学计量学 126, 6137-6145 (2021)。[12] M. Gagolewski, B. ˙Zogała-Siudem, G. Siudem 和 A. Cena, 《奥卡姆引用影响力指数》。科学计量学 127, 2829-2845 (2022)。[13] A.L. Barab´asi 和 R. Albert, 《随机网络中标度的出现》。Science 286, 509-512 (1999)。[14] C. Gini, 《论特征的集中度与变异性的度量》。Atti del Reale Istituto Veneto di Scienze, (Lettere ed Arti. A.A., 1914), pp. 1203-1248。[15] T.S. Bir´o 和 Z. N´eda, 《基尼熵：基于基尼指数的熵推广》，Entropy 22, 879 (2020)。[16] J. Mingers 和 L. Leydesdorff, 《科学计量学理论与实践综述》。European Journal of Operational Research 246, 1-19 (2015)。[17] S. Alonso、F.J. Cabrerizo、E. Herrera-Viedma 和 F. Herrera, 《h指数：聚焦其变体、计算方法及其在不同科学领域标准化的综述》。Journal of Infometrics 3, 273-289 (2009)。[18] A. Bihari, S. Tripathi 和 A. Deepak, 《关于 h 指数及其替代指数的综述》，Journal of Information Science 0, 0165551521101447 (2021)。[19] L. Bornmann, R. Mutz 和 H.D. Daniel, 《是否存在比 h 指数更适用于评估用途的指标？使用生物医学数据比较 h 指数的九种不同变体》。Journal of the American Society for Information Science and Technology 59, 830-837 (2008)。[20] A. M. C. Sengor, 《科学计量学如何扼杀科学》。GSA today 24, 44-45 (2014)。[21] K. Barcza 和 A. Telcs, 《帕累托型出版模式意味着帕累托型赫希指数》。科学计量学 81, 513 (2009)。[22] Z. N´eda, L. Varga, T.S. Bir´o, 《科学与脸书：相同的流行度定律！》PLOS ONE 12, 0179656 (2017)。

[23] M.O. Lorenz, 《财富集中度测量方法》。美国统计学会出版物 9, 209-219 (1905)。[24] T.S. Bir´o、A. Telcs、M. J´ozsa 和 Z. N´eda, 《f-Gintropy：基于基尼指数的熵距离排序》。熵 24, 407 (2022)。[25] https://en.wikipedia.org/wiki/Generalized Pareto distribution [26] W. Gl¨anzel、A. Telcs 和 A. Schubert, 《基于截断矩的刻画及其在皮尔逊型分布中的应用》。概率论及相关领域杂志 66, 173-183 (1984)。[27] A. Schubert 和 W. Gl¨anzel, **6**, 149–167, 1984 《偏态分布类的动态考察：一个具有科学计量学应用的模型》。科学计量学 - ( )。[28] A. Telcs、W. Gl¨anzel 和 A. Schubert, 《利用截断期望对一类偏态分布进行刻画和统计检验》。数学社会科学 10, 169-178 (1985)。[29] https://en.wikipedia.org/wiki/Survival analysis [30] C. Tsallis, 《非广延统计力学导论：探索复杂世界》。（施普林格，2009），第 41 页。