*Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht*

*Scientometrics, Vol. 81, No. 2 (2009) 493–498*
DOI: 10.1007/s11192-008-2169-6

# An investigation on mathematical models of the h-index

FRED Y. YE [a,b]

[a] *Dept. Information Resources Management, Zhejiang University, Hangzhou, P. R. China*
[b] *Institution of Scientific and Technical Information of China, Beijing, P. R. China*

Based on two large data samples from ISI databases, the author evaluated the Hirsch model, the Egghe–Rousseau model, and the Glänzel–Schubert model of the h-index. The results support the Glänzel–Schubert model as a better estimation of the h-index at both journal and institution levels. If $h_c$, $h_p$ and $h_{pc}$ stand for the Hirsch estimation, Egghe–Rousseau estimation, and Glänzel–Schubert estimation, respectively, then an inequality $h_p < h \sim h_{pc} < h_c$ holds in most cases.

## Introduction

Since Hirsch introduced the idea of the h-index [HIRSCH, 2005], mathematical models of the h-index have been discussed by Egghe and Rousseau [EGGHE & ROUSSEAU, 2006; EGGHE, 2007; YE & ROUSSEAU, 2008], Glänzel and Schubert [GLÄNZEL, 2006; SCHUBERT & GLÄNZEL, 2007, CSAJBÓK & AL., 2007] and Burrell [BURRELL, 2007], among others. In these papers, Hirsch's original model, the Egghe–Rousseau model, and the Glänzel–Schubert model are three representatives. Here, we provide an empirical investigation on those h-index models both at journal and institution levels, based on two large data samples.

## Methodology

When we define the general h-index as, 'an information source has its index $h$ if $h$ of its $N$ produced outputs have at least h citations each and others ($N$–$h$) outputs have no more than $h$ citations each', then all assignees, authors, research groups, institutions, journals, and countries have their h-indices.

In Hirsch's original paper [HIRSCH, 2005], Hirsch proposed his mathematical model for the h-index, in which the h-index only links total citations ($C$) and measures with

$$h = \sqrt{\frac{C}{a}} \ , \tag{1}$$

where $a$ is a constant ranging between 3 and 5.

Received October 22, 2008; Published online April 16, 2009

*Address for correspondence*:
FRED Y. YE
E-mail: yye@zju.edu.cn

*0138–9130/US $ 20.00*

Egghe and Rousseau derived the Egghe-Rousseau model (Egghe and Rousseau, 2006), with the h-index linking only total source publications (*P*), using

$$h = P^{1/\alpha} \ , \tag{2}$$

where α>1 is Lotka's exponent.

Glänzel and Schubert set up the Glänzel–Schubert model [GLÄNZEL, 2006; SCHUBERT & GLÄNZEL, 2007; CSAJBÓK & AL., 2007] with the formula

$$h = cP^{1/3}(CPP)^{2/3} \ , \tag{3}$$

in which *CPP=C/P* denotes citations per publication (for journals, *C/P* is associated with the Impact Factor, IF), and *c* is a constant.

Using each of those models in turn, we can estimate the h-index with the following formulae, when α=2 and *a*=5.

$$h_c \sim \sqrt{C/5} \tag{4}$$

$$h_p \sim \sqrt{P} \tag{5}$$

$$h_{pc} \sim cP^{1/3}(C/P)^{2/3} \tag{6}$$

Let us call $h_p$, $h_c$ and $h_{pc}$ the Hirsch estimation, the Egghe-Rousseau estimation, and the Glänzel-Schubert estimation, respectively, of the h-index and check them using actual data.

## Data

In the database ISI Web of Science (WoS), we can search the h-index of a journal via publication name, and of an institution via its address. Meanwhile, in the database ISI Essential Science Indicators (ESI), we have selected total publications (*P*), citations (*C*), and citations per publication (*CPP*) in a time window of 10–11 years. When we choose the time span 1997–2007 in WoS to correspond to ESI 1997–2007, we can obtain comparable data. The double databases, WoS and ESI, coming from same ISI source, enable us to construct comparable data sets for the investigation. While searching h-indices from WoS are applied to be real h-indices, the ESI data are applied to be corresponding computing ones.

The data of journal titles and institutions constitutes two large samples, in which the top 200 journals and institutions are chosen according to their citation ranking in ESI. Using the same journal titles and institutions, we search WoS and obtain the h-indices by sorting "Times Cited". For simplifying data collection, different spellings of institutions are ignored, as their main spellings contribute their main h-indices, so that we still get the referable results on a statistical level. (Readers who need the original data may email the author.)

## Results

We can compare the searching h-indices from WoS with the computing h-indices via formulas (4), (5) and (6), using data from ESI. In Table 1, below, the example data and results of the top 10 journals and institutions are listed according to their h-index ranking. (For titles with identical h-indices, we rank according to $h_{pc}$. In computing $h_{pc}$, we set $c$=0.9 for journals and c=1 for institutions, and record the corresponding $h_{pc}$ as $0.9h_{pc}$ for journals and $h_{pc}$ for institutions).

Table 1 Examples of retrieved and computed h-indices: The top 10 journals and institutions

| Journal | $P$ | $C$ | $CPP$ | $h$ | $h_p$ | $h_c$ | $0.9h_{pc}$ |
|---|---|---|---|---|---|---|---|
| *Nature* | 11,274 | 1,337,209 | 118.61 | 487 | 106.18 | 517.15 | 487.17 |
| *Science* | 10,404 | 1,263,175 | 121.41 | 476 | 102.00 | 502.63 | 481.74 |
| *N Engl J Med* | 3,879 | 569,640 | 146.85 | 374 | 62.28 | 337.53 | 393.61 |
| *Cell* | 3,824 | 552,923 | 144.59 | 350 | 61.84 | 332.54 | 387.71 |
| *Proc Nat Acad Sci Usa* | 31,437 | 1,485,447 | 47.25 | 315 | 177.30 | 545.06 | 371.24 |
| *Lancet* | 7,320 | 438,190 | 59.86 | 286 | 85.56 | 296.04 | 267.41 |
| *J Biol Chem* | 59,611 | 1,864,004 | 31.27 | 264 | 244.15 | 610.57 | 348.96 |
| *Jama-J Am Med Assn* | 4,076 | 340,127 | 83.45 | 264 | 63.84 | 260.82 | 274.54 |
| *Nat Genet* | 2,244 | 254,603 | 113.46 | 259 | 47.37 | 225.66 | 276.15 |
| *Circulation* | 10,271 | 495,513 | 48.24 | 250 | 101.35 | 314.81 | 259.25 |
| Institution | $P$ | $C$ | $CPP$ | $h$ | $h_p$ | $h_c$ | $h_{pc}$ |
| Harvard Univ | 95,457 | 2,651,015 | 27.77 | 330 | 308.96 | 728.15 | 419.10 |
| Nih | 7,800 | 232,109 | 29.76 | 323 | 88.32 | 215.46 | 190.45 |
| Stanford Univ | 49,363 | 1,131,732 | 22.93 | 307 | 222.18 | 475.76 | 296.08 |
| Univ Calif San Francisco | 36,621 | 981,823 | 26.81 | 299 | 191.37 | 443.13 | 297.47 |
| Johns Hopkins Univ | 53,594 | 1,211,258 | 22.60 | 297 | 231.50 | 492.19 | 301.38 |
| Univ Washington | 55,003 | 1,131,765 | 20.58 | 297 | 234.53 | 475.77 | 285.60 |
| Mit | 36,315 | 814,312 | 22.42 | 291 | 190.56 | 403.56 | 263.30 |
| Univ Calif San Diego | 41,318 | 920,778 | 22.29 | 290 | 203.27 | 429.13 | 273.81 |
| Brigham & Womens Hosp | 14,940 | 482,231 | 32.28 | 289 | 122.23 | 310.56 | 249.69 |
| Max Planck Society | 72,087 | 1,346,597 | 18.68 | 284 | 268.49 | 518.96 | 293.00 |

From Table 1, we see that all $h_{pc}$ correlate with the searching h-indices better, while almost all $h_p$ form the bottom and all $h_c$ leap to the top.

## Discussion and analysis

For observation in larger data sample, we can broaden the top 10 to the top 100 journals and institutions for their h-index fittings as shown in Figures 1 and 2.

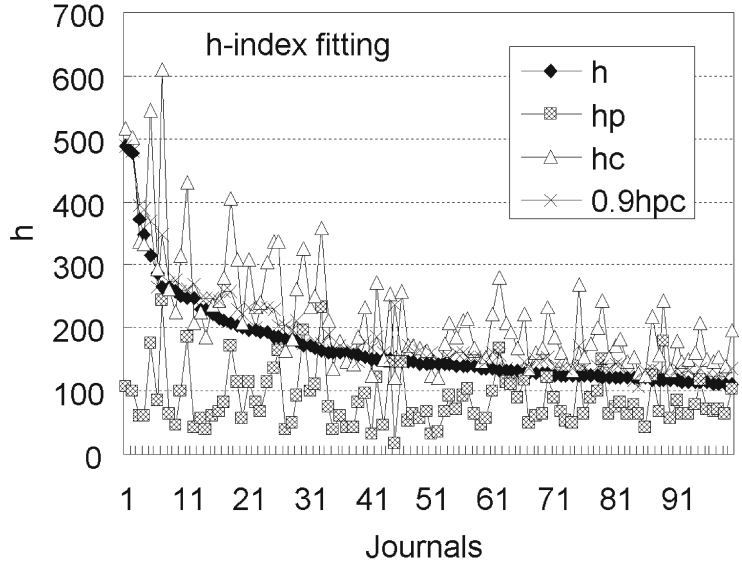


Figure 1. Fitting h-index to the top 100 journals

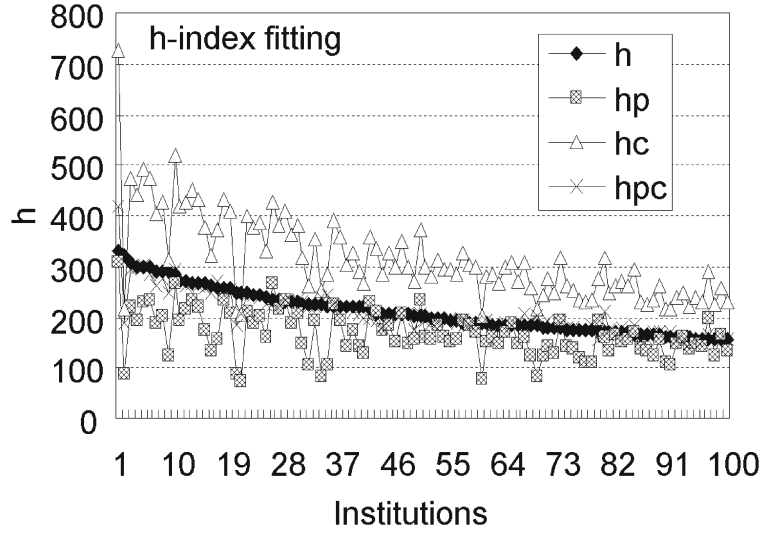


Figure 2. Fitting h-index to the top 100 institutions

It can be clearly seen that . the Glänzel-Schubert estimation gives the closest fit..

The above results support the Glänzel-Schubert model as a better estimation of the h-index at both journal and institution level. Because we know that the h-index relates both publications and citations, it is realistic and correct to support the Glänzel-Schubert model. The fitting results also show that there is an inequality for most cases, as follows, at both journal and institution levels

$$h_p < h \sim h_{pc} < h_c \tag{7}$$

Meanwhile, we can compute the Pearson correlation coefficients of the top 100 data columns, using SPSS, shown in Table 2 (for journals) and Table 3 (for institutions).

Table 2 Pearson Correlation Matrix at the 0.01 level for journals

| Pearson(Sig.(2-tailed)) | $h$ | $h_p$ | $h_c$ | $0.9h_{pc}$ |
|---|---|---|---|---|
| $h$ | 1 | 0.124 (0.220)* | 0.740 (0.000) | 0.975 (0.000) |
| $h_p$ | 0.124 (0.220)* | 1 | 0.723(0.000) | 0.169 (0.093)* |
| $h_c$ | 0.740 (0.000) | 0.723(0.000) | 1 | 0.786(0.000) |
| $0.9h_{pc}$ | 0.975 (0.000) | 0.169(0.093)* | 0.786(0.000) | 1 |

In Table 1, the items marked * indicate no correlation. The others show significant correlations.

Table 3 Pearson Correlation Matrix at the 0.01 level for institutions

| **Pearson(Sig.(2-tailed))** | $\boldsymbol{h}$ | $\boldsymbol{h_p}$ | $\boldsymbol{h_c}$ | $\boldsymbol{h_{pc}}$ |
|---|---|---|---|---|
| $\boldsymbol{h}$ | 1 | 0.449 (0.000) | 0.756 (0.000) | 0.885 (0.000) |
| $\boldsymbol{h_p}$ | 0.449 (0.000) | 1 | 0.872(0.000) | 0.576 (0.000) |
| $\boldsymbol{h_c}$ | 0.756 (0.000) | 0.872(0.000) | 1 | 0.899(0.000) |
| $\boldsymbol{h_{pc}}$ | 0.885 (0.000) | 0.576 (0.000) | 0.899(0.000) | 1 |

Thus, all data suggest that the Glänzel-Schubert model is a better estimation of the h-index, so that it is feasible to apply the Glänzel-Schubert model to estimate the h-indices of countries and other information sources [CSAJBÓK & AL., 2007]. This is a reference system for further research.

## Conclusion

Employing the Hirsch model, the Egghe-Rousseau model, and the Glänzel-Schubert model respectively, we estimate the h-index as

$h_c \sim \sqrt{C/5}$,

$h_p \sim \sqrt{P}$ and

$h_{pc} \sim c(P)^{1/2}(C/P)^{2/3}$ ($c$=0.9 for journals and $c$=1 for institutions),

based on two large data samples from ISI databases. The results support the Glänzel-Schubert model as a better estimation of the h-index at both journal and institution levels, so that we can also apply Glänzel-Schubert model to estimate the h-indices of countries and other information sources. And an inequality for most cases is suggested by $h_p<h\sim h_{pc}<_{hc}$.

*

The author is grateful for the financial support of the National Natural Science Foundation of China (NSFC Grant No 70773101) and for Ms. Regina Entorf's kind review and English wording, as well as for anonymous referees' comments.

## References

BURRELL, Q. L. (2007), Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1 (1) : 16–25.

CSAJBÓK, E., BERHIDI, A., VASAS, L., SCHUBERT, A. (2007), Hirsch-index for countries based on Essential Science Indicators data. *Scientometrics,* 73 (1) : 91–117.

EGGHE, L. (2007), Dynamic h-index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, 58 (3) : 452–454.

EGGHE, L., ROUSSEAU, R. (2006), An informetric model for the Hirsch-index. *Scientometrics,* 69 (1) : 121–129.

GLÄNZEL, W. (2006), On the h-index – A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67 (2) : 315–321.

HIRSCH, J. E., An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA,* 2005, 102 (46) : 16569–16572.

SCHUBERT, A., GLÄNZEL, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics,* 1 (2) : 179–184.

YE, F. Y., ROUSSEAU, R. (2008), The power law model and total career h-index sequences. *Journal of Informetrics,* 4 : 288–297.