# Some Comments on Egghe's Derivation of the Impact Factor Distribution

Ludo Waltman and Nees Jan van Eck



| ERIM REPORT SERIES *RESEARCH IN MANAGEMENT* | |
|---|---|
| ERIM Report Series reference number | ERS-2009-016-LIS |
| Publication | March 2009 |
| Number of pages | 8 |
| Persistent paper URL | http://hdl.handle.net/1765/15184 |
| Email address corresponding author | lwaltman@ese.eur.nl |
| Address | Erasmus Research Institute of Management (ERIM)<br>RSM Erasmus University / Erasmus School of Economics<br>Erasmus Universiteit Rotterdam<br>P.O.Box 1738<br>3000 DR Rotterdam, The Netherlands<br>Phone: + 31 10 408 1182<br>Fax: + 31 10 408 9640<br>Email: info@erim.eur.nl<br>Internet: www.erim.eur.nl |

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website:
www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

# REPORT SERIES
## *RESEARCH IN MANAGEMENT*

| ABSTRACT AND KEYWORDS | |
|---|---|
| Abstract | In a recent paper, Egghe [Egghe, L. (in press). Mathematical derivation of the impact factor distribution. Journal of Informetrics] provides a mathematical analysis of the rankorder distribution of journal impact factors. We point out that Egghe's analysis relies on an unrealistic assumption, and we show that his analysis is not in agreement with empirical data. |
| Free Keywords | impact factor, distribution, rank-order distribution |
| Availability | The ERIM Report Series is distributed through the following platforms:<br>Academic Repository at Erasmus University (DEAR), DEAR ERIM Series Portal<br>Social Science Research Network (SSRN), SSRN ERIM Series Webpage<br>Research Papers in Economics (REPEC), REPEC ERIM Series Webpage |
| Classifications | The electronic versions of the papers in the ERIM report Series contain bibliographic metadata by the following classification systems:<br>Library of Congress Classification, (LCC) LCC Webpage<br>Journal of Economic Literature, (JEL), JEL Webpage<br>ACM Computing Classification System CCS Webpage<br>Inspec Classification scheme (ICS), ICS Webpage |

# Some comments on Egghe's derivation of the impact factor distribution

Ludo Waltman    Nees Jan van Eck

Econometric Institute, Erasmus School of Economics
Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
E-mail: {lwaltman,nvaneck}@ese.eur.nl

Centre for Science and Technology Studies, Leiden University
P.O. Box 905, 2300 AX Leiden, The Netherlands

**Abstract**

In a recent paper, Egghe [Egghe, L. (in press). Mathematical derivation of the impact factor distribution. *Journal of Informetrics*] provides a mathematical analysis of the rank-order distribution of journal impact factors. We point out that Egghe's analysis relies on an unrealistic assumption, and we show that his analysis is not in agreement with empirical data.

**Keywords**

Impact factor, distribution, rank-order distribution.

## 1 Introduction

In a recent paper in the *Journal of Informetrics*, Egghe (in press) provides a mathematical analysis of the rank-order distribution of journal impact factors (IFs). Egghe's aim is to give a theoretical explanation for the IF rank-order distributions that are shown in a paper by Mansilla, Köppen, Cocho, and Miramontes (2007). In this communication, we point out that Egghe's analysis relies on an unrealistic assumption. We also show that his analysis is not in agreement with empirical data. Based on our findings, we conclude that Egghe's explanation for the IF rank-order distributions shown by Mansilla et al. is unsatisfactory.

## 2 Summary of Egghe's analysis

Egghe interprets the IF of a journal as the average of a number of independent and identically distributed random variables. Each random variable represents the number of citations of one of the articles published in the journal. Using the central limit theorem, Egghe's interpretation implies that the IF of a journal is a random variable that is (approximately) normally distributed. Egghe also makes the assumption that for a given scientific field "each journal in this field can be considered as a random sample in the total population of all articles in the field". This assumption has the implication that the IFs of all journals in a field follow the same normal distribution.[1] Based on this result, Egghe studies the properties of two types of rank-order distributions, namely rank-order distributions of IFs and rank-order distributions of logarithms of IFs. Egghe proves that both types of rank-order distributions have an S-shape, that is, both types of rank-order distributions are first convexly decreasing and then concavely decreasing. The empirical results reported by Mansilla et al. (2007) (see also Althouse, West, Bergstrom, & Bergstrom, 2009) indicate that rank-order distributions of logarithms of IFs indeed have the S-shape predicted by Egghe.

As an illustration of Egghe's analysis, we consider the following hypothetical example. There are $1000$ journals in a certain scientific field. During a certain period of time, each of these journals has published $100$ articles. The number of citations of an article is a random variable. Since in total $100,000$ articles have been published, there are $100,000$ random variables. These random variables are assumed to be independent and identically distributed.[2] To examine how IFs are distributed in this example, we make use of computer simulation. For each of the $100,000$ articles, we determine the number of citations by a draw from a negative binomial distribution (e.g., Glänzel, 2009; Schubert & Glänzel, 1983) with mean $1$ and vari-

[1] In fact, the implication requires an additional assumption, namely the assumption that all journals in a field publish the same number of articles. However, this assumption seems less critical for Egghe's analysis.

[2] This is equivalent to Egghe's assumption that the articles published in a journal can be regarded as a random sample from the population of all articles published in a field.

ance $5/4$. (The choice of the distribution is insignificant. Other distributions could have been used as well.) We then calculate for each journal the average number of citations of the articles published in the journal. This yields the IF of the journal. The distribution of the IFs of all $1000$ journals is shown in Figure 1. As can be seen, the distribution is approximately normal. The rank-order distributions of the IFs and of the logarithms of the IFs are shown in Figures 2 and 3, respectively. Both rank-order distributions have an S-shape. The dashed lines in Figures 1, 2, and 3 indicate the average IF. In Figure 1, the dashed line coincides with the mean of the normal distribution. In Figure 2, the dashed line intersects the IF rank-order distribution approximately in its inflection point. We note that the rank-order distribution shown in Figure 3 has a similar shape as the rank-order distributions shown in Figures 2, 3, and 4 in the paper by Mansilla et al. (2007). At first sight, Egghe's analysis therefore appears to be in agreement with empirical data.

# 3 Comments on Egghe's analysis

Egghe's analysis depends crucially on the assumption that the articles published in a journal can be regarded as a random sample from the population of all articles published in a field. This is a rather unrealistic assumption. We all know that some journals have a significantly higher IF than others. Moreover, we also know that IFs are fairly stable over time, that is, most journals that have a relatively high (or low) IF in one year still have a relatively high (or low) IF a few years later. It is clear that this would not be the case if Egghe's assumption of random sampling of articles were true.

According to Egghe's analysis, the distribution of the IFs of the journals in a field is approximately normal (like in Figure 1). Egghe does not verify this empirically. In Figures 4, 5, and 6, we show IF distributions for the fields of physics, mathematics, and environmental science. The distributions are based on data from Popescu (2003). This is the same data as is used by Mansilla et al. (2007). It is easy to see that the data does not support Egghe's analysis. The distributions in Figures 4, 5, and 6 should approximate normal distributions with mean equal to the average IF (cf. Figure 1). This is clearly not the case.[3] In addition to physics, mathematics, and environmental science, there are nine other fields that are covered by Popescu's data. The

[3]The variance of the distributions in Figures 4, 5, and 6 is also much larger than what seems reasonable to expect based on Egghe's application of the central limit theorem.

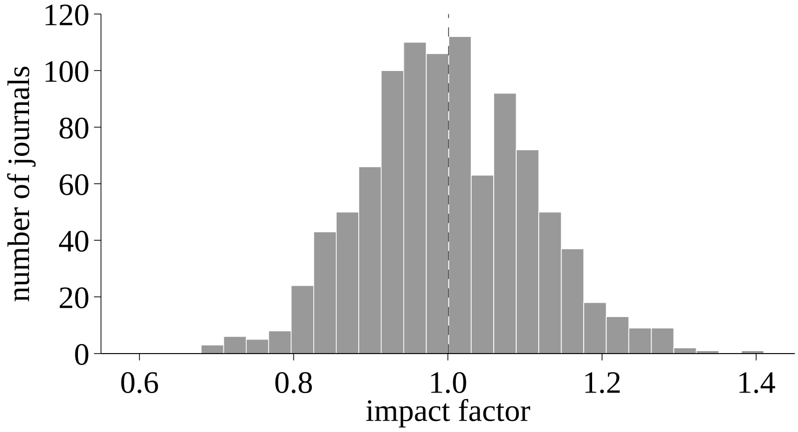


Figure 1: Distribution of the IFs of 1000 hypothetical journals.

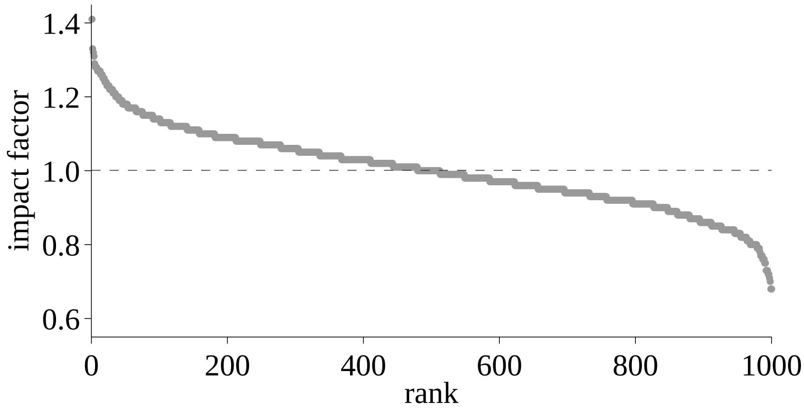


Figure 2: Rank-order distribution of the IFs of 1000 hypothetical journals.

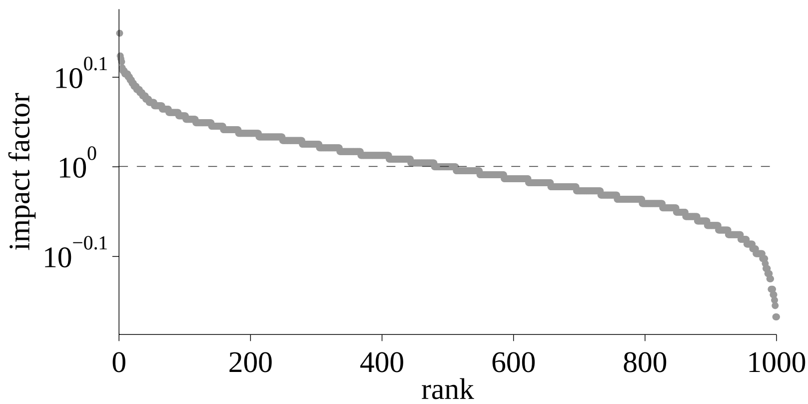


Figure 3: Rank-order distribution of the logarithms of the IFs of 1000 hypothetical journals.

IF distributions for these fields look similar to the distributions in Figures 4, 5, and 6 (most look similar to the distribution in Figure 4) and also do not support Egghe's analysis. For additional empirical evidence that IFs are not normally distributed, we refer to Beirlant, Glänzel, Carbonez, and Leemans (2007) and Schwartz and Lopez Hellin (1996).

Based on his analysis (in particular Theorem 3 in his paper), Egghe also claims that the rank-order distribution of the IFs of the journals in a field has an S-shape (like in Figure 2). Egghe does not provide empirical evidence for this claim. In Figures 7, 8, and 9, we show IF rank-order distributions for the fields of physics, mathematics, and environmental science. The distributions for the fields of physics and environmental science clearly do not have an S-shape. The distribution for the field of mathematics perhaps comes somewhat closer to an S-shape, but the location of the inflection point of the distribution does not correspond with Egghe's prediction (cf. Figure 2). The IF rank-order distributions for the nine other fields for which we have data all do not have an S-shape.

# 4 Conclusion

We have pointed out that Egghe's analysis relies on the unrealistic assumption that the articles published in a journal can be regarded as a random sample from the population of all articles published in a field. We have also shown that Egghe's analysis is not in agreement with empirical data. Based on our findings, we conclude that Egghe does not give a satisfactory explanation for IF rank-order distributions such as those shown by Mansilla et al. (2007).

There is one remaining question: If Egghe's analysis is not correct, why does it appear to be in agreement with the IF rank-order distributions shown in Figures 2, 3, and 4 in the paper by Mansilla et al. (2007)? The answer to this question is twofold. First, there is only a partial agreement between Egghe's analysis (in particular Theorem 2 in his paper) and the distributions shown by Mansilla et al. The S-shape of the distributions is predicted correctly by Egghe, but his prediction of the location of the inflection point is not correct. Second, the S-shape predicted by Egghe can be obtained in many ways and does not require IFs to be normally distributed. Hence, Egghe's correct prediction of the S-shape does not imply that his analysis is correct. If, for example, one assumes IFs to be exponentially distributed (cf. Schwartz & Lopez Hellin, 1996), one also obtains an S-shape. (Moreover, under the assumption of exponentially distributed IFs,

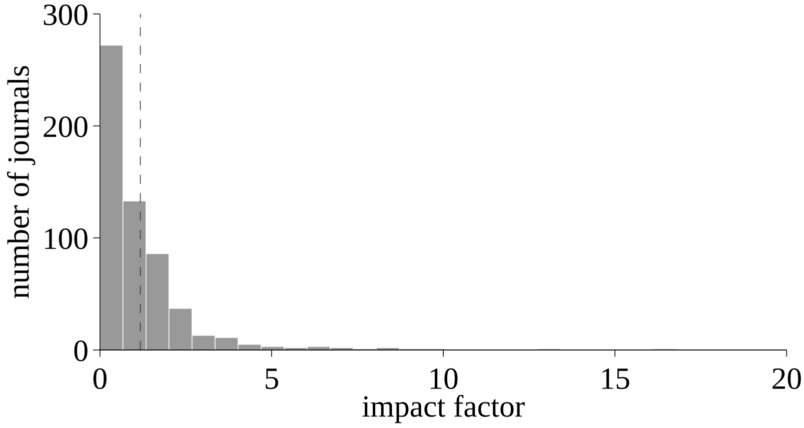


Figure 4: Distribution of the IFs of 574 physics journals.

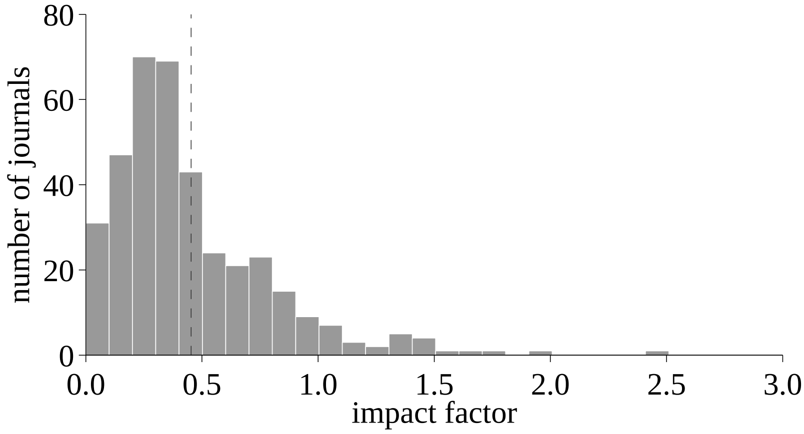


Figure 5: Distribution of the IFs of 378 mathematics journals.

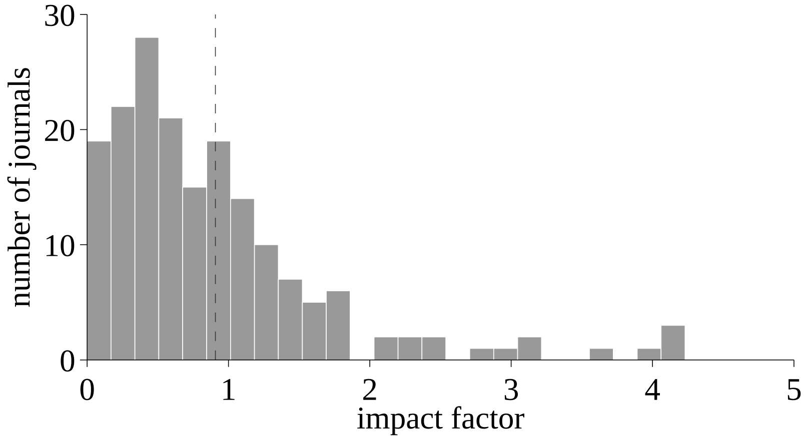


Figure 6: Distribution of the IFs of 181 environmental science journals.

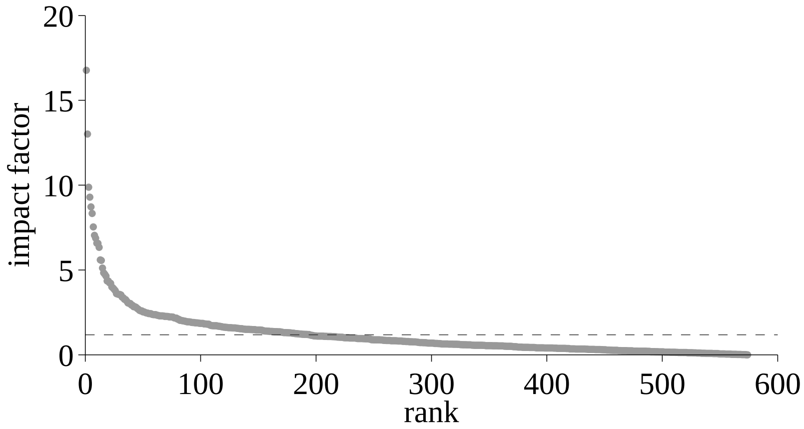


Figure 7: Rank-order distribution of the IFs of 574 physics journals.

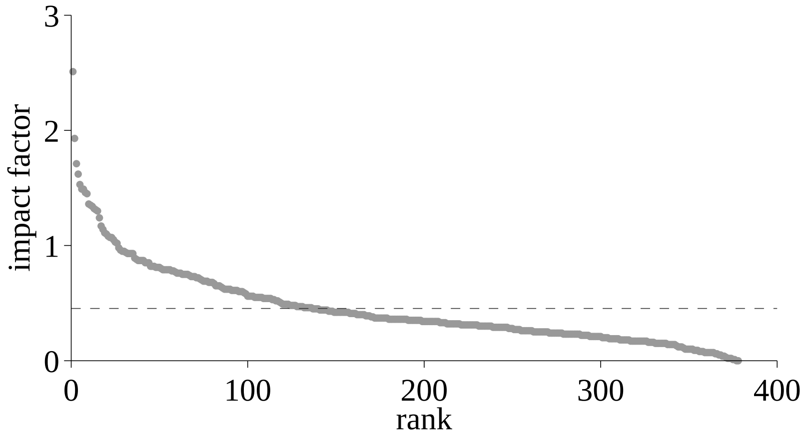


Figure 8: Rank-order distribution of the IFs of 378 mathematics journals.

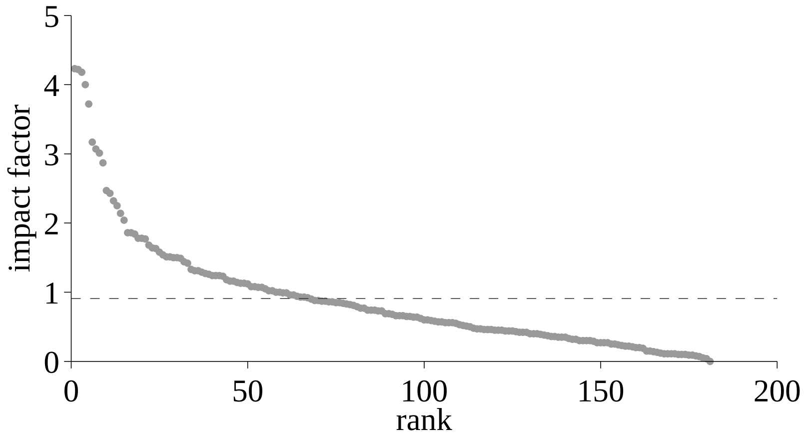


Figure 9: Rank-order distribution of the IFs of 181 environmental science journals.

the location of the inflection point is more in agreement with empirical data.)

# References

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, *60*(1), 27–34.

Beirlant, J., Glänzel, W., Carbonez, A., & Leemans, H. (2007). Scoring research output using statistical quantile plotting. *Journal of Informetrics*, *1*(3), 185–192.

Egghe, L. (in press). Mathematical derivation of the impact factor distribution. *Journal of Informetrics*.

Glänzel, W. (2009). The multi-dimensionality of journal impact. *Scientometrics*, *78*(2), 355–374.

Mansilla, R., Köppen, E., Cocho, G., & Miramontes, P. (2007). On the behavior of journal impact factor rank-order distribution. *Journal of Informetrics*, *1*(2), 155–160.

Popescu, I.-I. (2003). On a Zipf's law extension to impact factors. *Glottometrics*, *6*, 83–93.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, *5*(1), 59–74.

Schwartz, S., & Lopez Hellin, J. (1996). Measuring the impact of scientific publications. The case of the biomedical sciences. *Scientometrics*, *35*(1), 119–132.

## Publications in the Report Series Research* in Management

### ERIM Research Program: "Business Processes, Logistics and Information Systems"

2009

*How to Normalize Co-Occurrence Data? An Analysis of Some Well-Known Similarity Measures*
Nees Jan van Eck and Ludo Waltman
ERS-2009-001-LIS
http://hdl.handle.net/1765/14528

*Spare Parts Logistics and Installed Base Information*
Muhammad N. Jalil, Rob A. Zuidwijk, Moritz Fleischmann, and Jo A.E.E. van Nunen
ERS-2009-002-LIS
http://hdl.handle.net/1765/14529

*Open Location Management in Automated Warehousing Systems*
Yugang YU and René B.M. de Koster
ERS-2009-004-LIS
http://hdl.handle.net/1765/14615

*VOSviewer: A Computer Program for Bibliometric Mapping*
Nees Jan van Eck and Ludo Waltman
ERS-2009-005-LIS
http://hdl.handle.net/1765/14841

*Nash Game Model for Optimizing Market Strategies, Configuration of Platform Products in a Vendor Managed Inventory (VMI) Supply Chain for a Product Family*
Yugang Yu and George Q. Huang
ERS-2009-009-LIS
http://hdl.handle.net/1765/15029

*A Mathematical Analysis of the Long-run Behavior of Genetic Algorithms for Social Modeling*
Ludo Waltman and Nees Jan van Eck
ERS-2009-011-LIS
http://hdl.handle.net/1765/15181

*A Taxonomy of Bibliometric Performance Indicators Based on the Property of Consistency*
Ludo Waltman and Nees Jan van Eck
ERS-2009-014-LIS
http://hdl.handle.net/1765/15182

*A Stochastic Dynamic Programming Approach to Revenue Management in a Make-to-Stock Production System*
Rainer Quante, Moritz Fleischmann, and Herbert Meyr
ERS-2009-015-LIS
http://hdl.handle.net/1765/15183

*Some Comments on Egghe's Derivation of the Impact Factor Distribution*
Ludo Waltman and Nees Jan van Eck
ERS-2009-016-LIS
http://hdl.handle.net/1765/15184

---

* A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship