**Made available by Hasselt University Library in https://documentserver.uhasselt.be**

# Theoretical justification of the central area indices and the central interval indices

Peer-reviewed author version

EGGHE, Leo (2013) Theoretical justification of the central area indices and the central interval indices. In: SCIENTOMETRICS, 95(1), p. 25-34..

DOI: 10.1007/s11192-012-0803-9
Handle: http://hdl.handle.net/1942/14335

# Theoretical justification of the central area indices and the central interval indices

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]

and

Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

leo.egghe@uhasselt.be

---

## ABSTRACT

The central area indices and the central interval indices, as introduced in Dorta-González and Dorta-González, Scientometrics 88(3), 729-745, 2011, are studied from a theoretical point of view. They are defined in order to yield higher impact values of "selective" authors (i.e. authors with concentrated number of citations over their publications).

We show that this property is not valid for every citation distribution. However, if Zipf's law is adopted for the citation distribution, we can show that the central area indices and the central interval indices have indeed higher values for more selective authors.

---

[1] Permanent address

Key words and phrases: central area index, central interval index, *h*-index, Hirsch index.

# Introduction

The Hirsch index (or *h*-index) was introduced in Hirsch (2005) to measure the impact of a reasearcher's papers by means of their received citations. If we rank the papers of this researcher in decreasing order of their received citations then this researcher has *h*-index *h* if $r = h$ is the highest rank such that all the papers on ranks $1, 2, ..., h$ have at least *h* citations. This index was then later applied to other units such as journals, institutes, topics,…. For this see the review paper Egghe (2010) and the many references therein.

Also in Egghe (2010) (and references) the advantages and disadvantages of the *h*-index are described. One disadvantage is that, once a paper is in the *h*-core (i.e. once a paper belongs to the first *h* papers defining the *h*-index), it does not matter anymore how many citations this paper has received (as long as it is *h* or more). This has led researchers to define alternatives for the *h*-index that take more into account the actual number of citations to the most cited papers (e.g. the *g*-index (Egghe 2006)) and the *R*-index (Jin, Liang, Rousseau and Egghe (2007)).

Another approach is to define areas under the citation curve (e.g. of a researcher) so that, for papers within this area, all citations to these papers are counted. This is the approach followed in Dorta-González and Dorta-González (2011). Their exact definitions are as follows. Order the papers (e.g. of a researcher) in decreasing order of the number of received citations. Denote by *g(r)* the number of citations to the paper on rank *r* and let *h* be the *h*-index of this researcher. Then the "Central Interval Index" (CII) of radius $j = 1, ..., h-1$ is defined as

$$I_j = \sum_{r=h-j}^{h+j} g(r) \tag{1}$$

The "Central Area Index" (CAI) of radius *j* is defined as

$$A_j = (h-j)\,g(h-j) + \sum_{r=h-j+1}^{h+j} g(r) \tag{2}$$

also for $j = 1, ..., h-1$. They are depicted in Figs. 1 and 2 for a convexly decreasing citation function *g(r)*, in a continuous format.

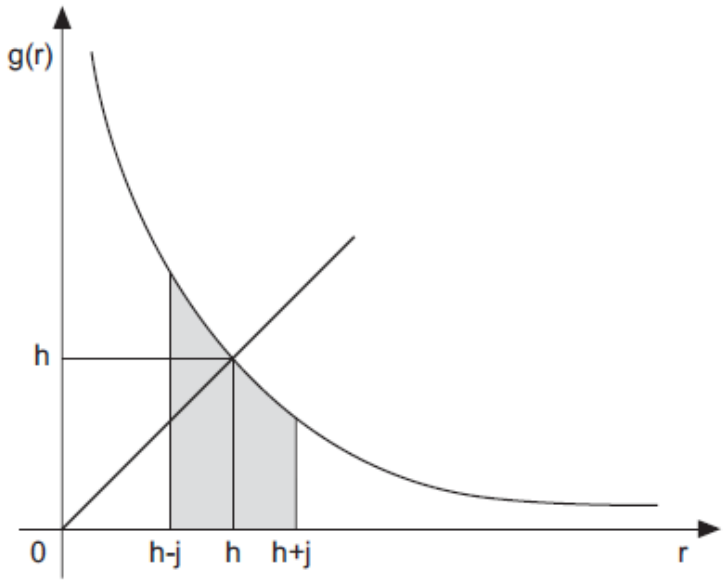


Fig. 1. The area under the curve *g(r)* for abscissae between $h-j$ and $h+j$ is the "Central Interval index" $I_j$ of radius *j*.

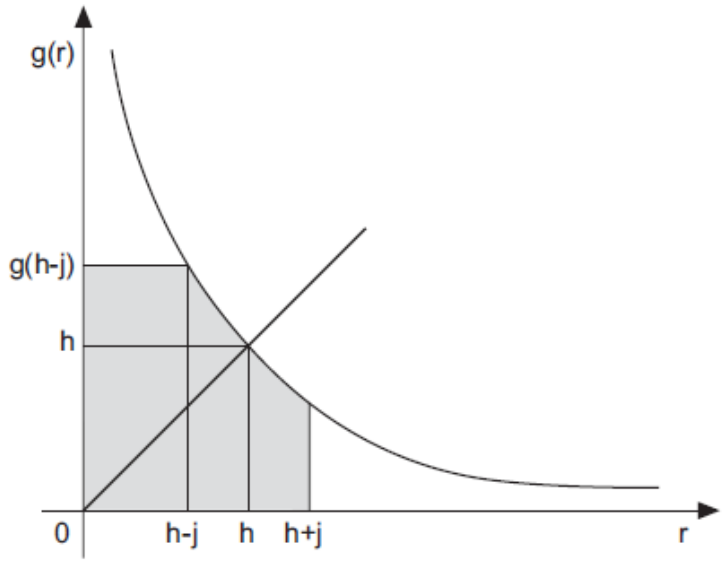


Fig. 2. The "Central Area Index" $A_j$ of radius $j$ is $I_j$ plus the area of the rectangle with abscissae in the interval $[0, h-j]$ and ordinates in the interval $[0, g(h-j)]$.

If we let $j = 1,...h-1$ in the discrete setting) increase, we include more papers and more citations to these papers than in the calculation of the *h*-index. It is claimed in Dorta-González and Dorta-González (2011) that, when comparing two researchers with the same *h*-index, the more selective researcher will receive the larger $I_j$ and $A_j$ values. What does this mean?

First of all, since they have the same *h*-index, their citation curves, $g_1(r)$ and $g_2(r)$ intersect in the point *(h,h)* (by definition of the *h*-index) – see Fig. 3.

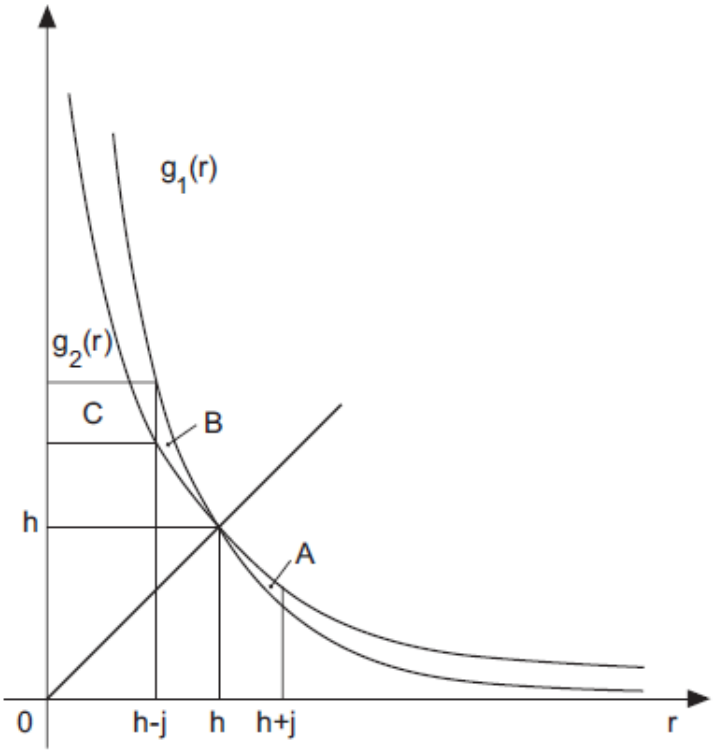


Fig. 3. Intersecting citation curves (in *(h,h)*) and consequences for the $I_j$ and $A_j$ indices.

As a consequence of this we have that one citation curve (say $g_1(r)$) is above the other one (say $g_2(r)$) for abscissae smaller than *h* while $g_1(r)$ is below $g_2(r)$ for abscissae larger than *h*. What consequences does this have on the calculation of the $I_j$ and $A_j$ indices for these two researchers? Denote by $I_j^{(1)}$ and $I_j^{(2)}$ the "Central Interval Indices" (CII) of researcher 1 and researcher 2 respectively and similarly we denote by $A_j^{(1)}$ and $A_j^{(2)}$ the "Central Area Indices" (CAI) of researcher 1 and researcher 2 respectively. Define (see Fig.

3) *A* to be the area between the two citation curves $g_1(r)$ and $g_2(r)$ for abscissae *r* between *h* and *h*+*j* ( $j = 1,...,h-1$ fixed). Similarly define *B* to be the area between the two citation curves $g_1(r)$ and $g_2(r)$ for abscissae *r* between *h*-*j* and *h*. Finally define *C* to be the area of the rectangle where the coordinates are $(x, y)$ where $x \le h - j$ and $g_2(h-j) \le y \le g_1(h-j)$.

Since $g_1(r)$ is above $g_2(r)$ for abscissae smaller than *h* and since $g_1(r)$ is below $g_2(r)$ for abscissae larger than *h* we have the following relations between $I_j^{(1)}$ and $I_j^{(2)}$ and between $A_j^{(1)}$ and $A_j^{(2)}$.

$$I_j^{(1)} = I_j^{(2)} + B - A \tag{3}$$

$$A_j^{(1)} = A_j^{(2)} + B - A + C \tag{4}$$

Due to the fact that one subtracts the value *A* it is not always so that $I_j^{(1)} > I_j^{(2)}$ and $A_j^{(1)} > A_j^{(2)}$, although this is implicitly assumed in Dorta-González and Dorta-González (2011) (see their Fig. 3, p. 734) and experimentally verified (in a heuristic way). This was exactly the reason for the introduction of the indices $I_j$ and $A_j$, complementing the *h*-index so that higher impact is given to more 'selective" cases, i.e. where the citations are more concentrated in the highly cited papers. A counterexample, however, for $I_j^{(1)} > I_j^{(2)}$ and $A_j^{(1)} > A_j^{(2)}$ is given by, given a situation as in Fig. 3, leaving $g_1(r)$ and $g_2(r)$ for $r > h$ and by moving $g_1(r)$ close to $g_2(r)$ for $r < h$. In this way the areas B and C can be made as small as needed while letting the area A constant so that (3) and (4) yield $I_j^{(1)} < I_j^{(2)}$ and $A_j^{(1)} < A_j^{(2)}$. Note, however, that, in the definition of Dorta-González and Dorta-González (2011), researcher 1 is more "selective" than researcher 2 since the citations to papers of researcher 1 are more concentrated in highly cited papers than is the case with researcher 2 (since the citation curve $g_1(r)$ is above $g_2(r)$ for $r < h$ and $g_1(r)$ is below $g_2(r)$ for $r > h$).

By the above counterexample, the CIIs $I_j$ and the CAIs $A_j$ do not give higher impact values to the more "selective" researchers. In terms of concentration theory (see e.g. Egghe (2005), Chapter III), the situation in Fig. 3 yields a more concentrated situation for researcher 1 in comparison with researcher 2: citations are more unequally distributed for researcher 1 in comparison with researcher 2. This would mean that, if we were constructing the Lorenz curves of these two researchers, the one of researcher 1 would be strictly above the Lorenz curve of researcher 2 (see Egghe (2005), Chapter III).

Does this mean that the indices $I_j$ and $A_j$ are bad impact measures? Not necessarily: if we can prove that the $I_j s$ and $A_j s$ increase for more selective cases (e.g. $g_1(r)$ is more selective than $g_2(r)$ in Fig. 3) using citation curves that occur in practice in informetrics, then we have proved that the indices $I_j$ and $A_j$ are good impact measures. The most classical informetrics theory is Lotkaian informetrics (see Egghe (2005), Chapters I, II). In this case we assume that the size-frequency curve $f(j)$ is the law of Lotka (Lotka (1926)): a decreasing power law: denote by $f(k)$ the number of papers (of a researcher) with *k* citations, then

$$f(k) = \frac{C}{k^{\alpha}} \tag{5}$$

where $C > 0$ and $\alpha > 1$ are parameters and $k \geq 1$ is the variable. This size-frequency function $f(k)$ is equivalent with the rank-frequency function $g(r)$ (as used above: the citation curve, i.e the number of citations to the paper on rank *r* ) by the following well-known proposition.

**Proposition 1 (Egghe):**

The following assertions are equivalent:

(i) The size-frequency function $f(k)$ is the law of Lotka

$$f(k) = \frac{C}{k^{\alpha}} \tag{5}$$

$C > 0,\ \alpha > 1,\ k \geq 1,$

(ii) The rank-frequency function $g(r)$ is Zipf's law

$$g(r) = \frac{B}{r^{\beta}} \tag{6}$$

$B > 0$, $\beta > 0$, $0 < r \leq T$, where *T* denotes the total number of papers. In addition, the parameters relate to each other as in (7), (8) and (9)

$$T = \frac{C}{\alpha - 1} \tag{7}$$

$$B = \left( \frac{C}{\alpha - 1} \right)^{\frac{1}{\alpha - 1}} = T^{\frac{1}{\alpha - 1}} \tag{8}$$

$$\beta = \frac{1}{\alpha - 1} \tag{9}$$

See Egghe (2005) (p. 134) or Egghe and Rousseau (2006), Appendix (p. 128-129) where a complete proof is given.

We also need the following result on the *h*-index in case of Lotkaian systems (5).

**Proposition 2 (Egghe and Rousseau)**

Let us have a Lotkaian system as in (5). Then the *h*-index of this system is given by formula (10)

$$h = T^{\frac{1}{\alpha}} \tag{10}$$

See Egghe and Rousseau (2006) for a proof (where also the above Proposition 1 is used).

In the next section we will study the CIIs and CAIs in the case that $g(r)$ is Zipf's law. There we will show that, independent of the parameters in Zipf's (or Lotka's) law, we always have

that $I_j^{(1)} > I_j^{(2)}$ and $A_j^{(1)} > A_j^{(2)}$ in case we have a situation as in Fig. 3. This shows that the CIIs and CAIs are useable impact measures (as shown in practise in Dorta-González and Dorta-González (2011)).

# The behavior of the CIIs and CAIs in case the citation curves are Zipfian

We work in the continuous setting, i.e. where variables *j, r* are real numbers.

In this setting the continuous versions of the definitions (1) of the CIIs and (2) of the CAIs are as follows. The "Central Interval Index" (CII) of radius *j*, $0 < j \le h$, is defined as

$$I_j = \int_{h-j}^{h+j} g(r)dr \tag{11}$$

, where *h* is the *h*-index of the system. The "Central Area Index" (CAI) of radius *j*, $0 < j \le h$, is defined as

$$A_j = (h-j)\,g(h-j) + \int_{h-j}^{h+j} g(r)dr \tag{12}$$

$$A_j = (h-j)\,g(h-j) + I_j \tag{13}$$

These measures, of course, correspond with the graphical definitions in Figs. 1 and 2. We have the following proposition.

**Proposition 3:**

For all *B*, $\beta > 0$, $\beta \neq 1$ we have, for all *j* such that $0 < j \le h$:

$$I_j = h^{\frac{\alpha}{\alpha-1}} \frac{\alpha-1}{\alpha-2}\left[(h+j)^{\frac{\alpha-2}{\alpha-1}} - (h-j)^{\frac{\alpha-2}{\alpha-1}}\right] \tag{14}$$

$$A_j = (h-j)\left(\frac{h^\alpha}{h-j}\right)^{\frac{1}{\alpha-1}} + h^{\frac{\alpha}{\alpha-1}}\frac{\alpha-1}{\alpha-2}\left[(h+j)^{\frac{\alpha-2}{\alpha-1}} - (h-j)^{\frac{\alpha-2}{\alpha-1}}\right] \tag{15}$$

**Proof:**

By (11) and (6) we have, for all *j* such that $0 < j \le h$,

$$\begin{aligned} I_j &= \frac{B}{1-\beta}\left[(h+j)^{1-\beta} - (h-j)^{1-\beta}\right] \\ &= T^{\frac{1}{\alpha-1}}\frac{\alpha-1}{\alpha-2}\left[(h+j)^{\frac{\alpha-2}{\alpha-1}} - (h-j)^{\frac{\alpha-2}{\alpha-1}}\right] \end{aligned}$$

using (8) and (9). Using (10) this gives (14). Now we use (6) to yield

$$(h-j)\,g(h-j) = (h-j)\frac{B}{(h-j)^\beta} \tag{16}$$

Using (8) and (9) yields that (16) equals

$$(h-j)\left(\frac{T}{h-j}\right)^{\frac{1}{\alpha-1}}$$

This, together with (10), (14) and (13) yields (15). □

Now let us go back to the situation in Fig.3. Let us denote

$$g_1(r) = \frac{B_1}{r^{\beta_1}} \tag{17}$$

and

$$g_2(r) = \frac{B_2}{r^{\beta_2}} \tag{18}$$

We have proved in Egghe (2011) that $\beta_1 > \beta_2$, hence, using (9) and denoting by $\alpha_1$, $\alpha_2$ the corresponding Lotka exponents, we have $\alpha_1 < \alpha_2$. We want to show that $I_j^{(1)} > I_j^{(2)}$ and $A_j^{(1)} > I_j^{(2)}$, for all *j*, $0 < j \le h$. In other words, we must show that $I_j$ and $A_j$ (in (14) and (15)) are decreasing functions of $\alpha$. This will be proved now.

**Proposition 4:**

For all *j*, $0 < j \le h$, the CIIs $I_j$ and the CAIs $A_j$ are decreasing functions of $\alpha$.

**Proof:**

Note that in (14) and (15), *h* is constant, since the curves $g_1(r)$ and $g_2(r)$ intersect in the point *(h,h)* and hence that they have the same *h*-index. So in the calculation of the dependence of $I_j$ and $A_j$ of $\alpha$ in (14) and (15), we have to keep *h* constant (due to (10) this means that, if $\alpha$ increases, *T* must decrease in order to keep *h* constant). This makes it clear how to take the derivative of $I_j$ and $A_j$ with the respect to $\alpha$, for all *j*

$$\begin{aligned}
&\frac{dI_j}{d\alpha} = I_j' = -h^{\frac{\alpha}{\alpha-1}} \frac{1}{(\alpha-1)(\alpha-2)} \cdot \\
&\cdot \Bigg[ (h+j)^{\frac{\alpha-2}{\alpha-1}} (\ln h) - (h-j)^{\frac{\alpha-2}{\alpha-1}} (\ln h) \\
&+ \frac{\alpha-1}{\alpha-2} \left( (h+j)^{\frac{\alpha-2}{\alpha-1}} - (h-j)^{\frac{\alpha-2}{\alpha-1}} \right) \\
&- \left( (h+j)^{\frac{\alpha-2}{\alpha-1}} \ln(h+j) - (h-j)^{\frac{\alpha-2}{\alpha-1}} \ln(h-j) \right) \Bigg]
\end{aligned} \tag{19}$$

(I) Let $\alpha > 2$.

Then, by (19), we have to prove that

$$\varphi(h+j) > \varphi(h-j) \tag{20}$$

where $\varphi(x)$ is defined as

$$\varphi(x) = x^{\frac{\alpha-2}{\alpha-1}}(\ln h) + \frac{\alpha-1}{\alpha-2} x^{\frac{\alpha-2}{\alpha-1}} - x^{\frac{\alpha-2}{\alpha-1}} \ln x \tag{21}$$

It turns out that the function $\varphi$ is not increasing in in *x*. So we strictly have to prove (20). We denote

$$h - j = \theta h \tag{22}$$

with $0 < \theta \leq 1$. Hence

$$h + j = (2-\theta)h \tag{23}$$

and hence we must prove that

$$\varphi(\theta h) < \varphi\big((2-\theta)h\big) \tag{24}$$

or

$$\begin{aligned} &(\theta h)^{\frac{\alpha-2}{\alpha-1}}\left[\ln h + \frac{\alpha-1}{\alpha-2} - \ln(\theta h)\right] \\ &< \big((2-\theta)h\big)^{\frac{\alpha-2}{\alpha-1}}\left[\ln h + \frac{\alpha-1}{\alpha-2} - \ln\big((2-\theta)h\big)\right] \end{aligned}$$

or

$$\begin{aligned} &(\theta h)^{\frac{\alpha-2}{\alpha-1}}\left(\frac{\alpha-1}{\alpha-2} - \ln\theta\right) \\ &< \big((2-\theta)h\big)^{\frac{\alpha-2}{\alpha-1}}\left(\frac{\alpha-1}{\alpha-2} - \ln(2-\theta)\right) \end{aligned} \tag{25}$$

It hence suffices (since $\theta h \leq (2-\theta)h$ since $0 < \theta \leq 1$) that the function

$$\psi(\theta)=(\theta h)^{\frac{\alpha-2}{\alpha-1}}\left(\frac{\alpha-1}{\alpha-2}-\ln\theta\right) \tag{26}$$

increases in $\theta$. But

$$\psi'(\theta)=\frac{\alpha-2}{\alpha-1}(\theta h)^{\frac{\alpha-2}{\alpha-1}-1}h\left(\frac{\alpha-1}{\alpha-2}-\ln\theta\right)-(\theta h)^{\frac{\alpha-2}{\alpha-1}}\frac{1}{\theta}$$

$$\psi'(\theta)=(\theta h)^{\frac{\alpha-2}{\alpha-1}-1}h\left(-\frac{\alpha-2}{\alpha-1}\ln\theta\right)\geq 0 \tag{27}$$

since $\alpha>2$ and $0<\theta\leq 1$.

(II) Let now $1<\alpha<2$

Then, by (19), we have to prove that

$$\varphi(h+j)<\varphi(h-j)$$

with $\varphi$ as in (21) or, using (21) and (22)

$$\varphi(\theta h)>\varphi\big((2-\theta)h\big) \tag{28}$$

or (same argument as in (25))

$$(\theta h)^{\frac{\alpha-2}{\alpha-1}}\left(\frac{\alpha-1}{\alpha-2}-\ln\theta\right)>\big((2-\theta)h\big)^{\frac{\alpha-2}{\alpha-1}}\left(\frac{\alpha-1}{\alpha-2}-\ln(2-\theta)\right) \tag{29}$$

So now we have to show that (26) decreases in $\theta$ (since $\theta h\leq(2-\theta)h$ since $0<\theta\leq 1$). But as in (27)

$$\psi'(\theta)=(\theta h)^{\frac{\alpha-2}{\alpha-1}-1}\left(-\frac{\alpha-2}{\alpha-1}\ln\theta\right)\leq 0$$

since $1<\alpha<2$ and $0<\theta\le 1$. This proves that CIIs $I_j$ decrease in $\alpha$. For the same result for the CAIs $A_j$ it suffices to show that the function

$$\xi(\alpha)=(h-j)\left(\frac{h^{\alpha}}{h-j}\right)^{\frac{1}{\alpha-1}} \tag{30}$$

decreases in $\alpha$ (for fixed *h*) (by (15) and the fact that $A_j=I_j+\xi(\alpha)$ and by the previous result). But

$$\begin{aligned}\frac{1}{h-j}\xi'(\alpha)&=\left(\frac{h^{\alpha}}{h-j}\right)^{\frac{1}{\alpha-1}}\ln\left(\frac{h^{\alpha}}{h-j}\right)\left(-\frac{1}{(\alpha-1)^2}\right)\\&+\frac{1}{\alpha-1}\left(\frac{h^{\alpha}}{h-j}\right)^{\frac{1}{\alpha-1}-1}\frac{1}{h-j}h^{\alpha}\ln h\\&=-\frac{1}{\alpha-1}\left(\frac{h^{\alpha}}{h-j}\right)^{\frac{1}{\alpha-1}}\left[\frac{1}{\alpha-1}\ln\left(\frac{h^{\alpha}}{h-j}\right)-\ln h\right]\end{aligned} \tag{31}$$

But

$$\frac{1}{\alpha-1}\ln\left(\frac{h^{\alpha}}{h-j}\right)-\ln h=\ln\left(\left(\frac{h^{\alpha}}{h-j}\right)^{\frac{1}{\alpha-1}}/h\right) \tag{32}$$

But $h-j<h$ so that

$$\frac{h^{\alpha}}{h-j}>h^{\alpha-1}$$

and hence (32)>0 so that (31) is >0 and hence (by (31)) $\xi(\alpha)$ (and hence $A_j$) decreases in $\alpha$ (since $\alpha>1$). This completes the proof of Proposition 4. □

Proposition 4 shows that in case of Fig.3, since in (17) and (18), $\beta_1 > \beta_2$ (hence $\alpha_1 < \alpha_2$ by (9)) that the $I_j$ values and $A_j$ values for $g_1(r)$ are larger than these values for $g_2(r)$ and hence the $I_j$ and $A_j$ values are higher in the more "selective" case (the curve $g_1(r)$). These are theoretical justifications for the use of the CIIs $I_j$ and the CAIs $A_j$ as a complement to the *h*-index, as was meant in Dorta-González and Dorta-González (2011).

# Conclusions and remarks

We have shown that the "Central Interval Indices" (CIIs) and the "Central Area Indices" (CAIs) do not always give higher impact values for more "selective" authors. However, if the citation curves follow the law of Zipf, then these indices give indeed higher impact values for more "selective" authors. In this sense, they are good complements to the *h*-index which is not the sensitive to highly cited papers.

Since there are many CIIs and CAIs (dependent on the radius $j = 1,..., h-1$) it is not easy to give interpretations to these values. Also, the calculation of these values is much more elaborate than the calculation of the *h*-index.

Since if *j* increases, we include more and more citations to highly cited papers, it is interesting to further study this *j*-dependence, both theoretically and empirically. Still, if $j \neq h$, we do not include the citations to the highest cited papers which can be considered as a disadvantage of these indices in comparison with e.g. the *g*-index and *R*-index.

# References

P. Dorta-González and M-I. Dorta-González (2011). Central indexes to the citation distribution: a complement to the *h*-index. Scientometrics 88(3), 729-745.

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford (UK).

L. Egghe (2006). Theory and practise of the *g*-index. Scientometrics 69(1), 131-152.

L. Egghe (2010). The Hirsch-index and related impact measures. Annual Review of Information Science and Technology, Volume 44 (B. Cronin, ed.), 65-114. Information Today, Inc., Medford, New Jersey, USA.

L. Egghe (2011). A disadvantage of *h*-type indices for comparing the citation impact of two researchers. Research Evaluation 20(4), 341-346.

L.Egghe and R. Rousseau (1984). An informetric model for the Hirsch-index. Scientometrics 69(1), 121-129.

J.E. Hirsch (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America 102(46), 16569-16572.

B.H. Jin, L.M. Liang, R. Rousseau and L. Egghe (2007). The *R*- and *AR*-indices: Complementing the *h*-index. Chinese Science Bulletin 52(6), 855-863.

A.J. Lotka (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16(12), 317-324.