Continuous, weighted Lorenz theory and applications to the study of fractional relative impact factors

Continuous, weighted Lorenz theory and applications to the study of fractional relative impact factors

by

L. Egghe

Limburgs Universitair Centrum (LUC), Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@luc.ac.be

ABSTRACT

This paper introduces weighted Lorenz curves of a continuous variable, extending the discrete theory as well as the non-weighted continuous model.

Using publication scores (in function of time) as the weights and citation scores (in function of time) as the dependent variables, we can construct an "impact Lorenz curve" in which one can read the value of any fractional impact factor, i.e. an impact factor measured at the time

---

[1] Permanent address

Key words and phrases: continuous weighted Lorenz curve, fractional impact factor
Acknowledgement: The author is grateful to prof. Dr. R. Rousseau for stimulating discussions on the topic of this paper and for drawing his attention to the paper Sombatsompop, Markpin and Premkamolnetr (2004). The author is also grateful to Mrs. A. Kuppens for delivering the publication and citation data of the journal "Journal of near Infrared Spectroscopy".

that a certain fraction of the citations is obtained or measured at the time a certain fraction of the publications is obtained.

General properties of such Lorenz curves are studied and special results are obtained in case the citation age curve and publication growth curve are exponential functions. If g is the growth rate and c is the aging rate we show that $\frac{\ln c}{\ln\left(\frac{1}{g}\right)}$ determines the impact Lorenz curve and also we show that any two situations give rise to two non-intersecting (except in (0,0) and (1,1)) Lorenz curves. This means that, for two situations, if one fractional impact factor is larger than the other one, the same is true for all the other fractional impact factors. We show, by counterexample that this is not so for "classical" impact factors, where one goes back to fixed time periods.

The paper also presents methods to determine the rates c and g from practical data and examples are given.

I. Introduction

Impact factors, as introduced by Garfield and Sher and the Institute for Scientific Information (ISI) in the sixties (see e.g. Garfield and Sher (1963), Garfield (1972, 1979a) or Egghe and Rousseau (1990)) are the subject of many research debates. In this paper we will not deal with the debate on the applicability of impact factors (and citation analysis in general) to the evaluation of scientific research (e.g. in comparison with peer review). For this, see e.g. Garfield (1979b, 1983) or Egghe and Rousseau (1990).

In this paper we are involved in the mathematical aspects of impact factors, i.e. in the comparison of the different mathematical forms of the impact factor. Let us give its historical definition and some possible variants. ISI uses the so-called two-year synchronous impact factor for its source journals in its products, e.g. in the JCR (Journal Citation Reports®). The two-year impact factor of a journal, denoted $IF(2)$, can be defined as

$$IF(2) = \frac{c(1) + c(2)}{p(1) + p(2)} \qquad (1)$$

where at time $t = 0$ (e.g. this year but any other year is also possible) $c(i), i = 1,2,$ denotes the number of citations given by source journals in year 0 to articles of this journal published i years ago and where $p(i), i = 1,2,$ denotes the number of articles published in this journal i years ago. The "Garfield" impact factor is then $IF(2)$ where $t = 0$ is the year of publication of the JCR (hence each year, new $IF(2)$s for the so-called source journals, i.e. journals covered by ISI, are published). However, although not published in the JCR, $IF(2)$ can be calculated for any journal and even scientific field. In the latter case it is important to distinguish between the field considered as a set of journals or as a set of articles in these journals (in the latter case one considers the field as a "meta journal"). In this paper we do not go into this problem. For this see e.g. Egghe and Rousseau (1996a,b) or Rousseau (1988).

It is clear that the above definition of $IF(2)$ can be (and is) subject of many discussions, even when we restrict ourselves – as indicated above – to the mathematical aspects. There are two major points.

(i) The limitation to 2 years in IF(2)

It is clear that, when ISI wanted to define "an impact factor" one had to choose the time period that one wants to go back. ISI choose for $t = 2$, a rather short period. There are suggestions that this choice was made based on commercial arguments (Dierick and Rousseau (1988)). In any case, even without any verification, the experienced informetrician knows that $t = 2$ cannot be the "ideal" time period to go back for every journal or scientific field. It is clear that, say in fields where the aging is small (i.e. when relatively old work is still cited), IFs calculated over longer time periods will be higher than $IF(2)$. Let us first define what we mean by a synchronous impact factor calculated over other time periods. For $t = 1,2,3,4,...$ we define (cf. Rousseau (1988), Ingwersen, Larsen, Rousseau and Russell (2001)):

$$IF(t) = \frac{c(1) + c(2) + ... + c(t)}{p(1) + p(2) + ... + p(t)}$$

$$IF(t) = \frac{\sum_{i=1}^{t} c(i)}{\sum_{i=1}^{t} p(i)} \qquad (2)$$

where $c(i)$ and $p(i)$ are as above (but now i ranges in the set $\{1, 2, ..., t\}$).

$IF(t)$ is a synchronous impact factor since the period in which citations are given is fixed and the "target" period is variable. If the reverse is true (citing period variable and target period fixed) we speak of a diachronous impact factor (cf. Stinson (1981), Stinson and Lancaster (1987), being studies of diachronous versus synchronous obsolescence and more recently Ingwersen, Larsen, Rousseau and Russell (2001)).

In this paper we will limit ourselves to synchronous impact factors but the results can easily be extended to the diachronous ones.

In Rousseau (1988) it is shown by experiment that, for pure mathematics journals and for pure mathematics as a field, $IF(4)$ is often larger than $IF(2)$. This is not surprising following our intuition described above: pure mathematics, in its citation behavior, follows the trends of the social sciences (or humanities) where relatively older work is still very much used.

A similar result was found in Rousseau, Jin, Yang and Liu (2001) where $IF(3)$ was found to be larger than $IF(2)$ in a general model based on ISI's database, but was found not to be so (in most cases) based on data of the CSCD (Chinese Science Citation Database).

$IF(t)$, as function of the discrete variable t, was studied in Rousseau (1988) where it was shown that $IF(t)$ – if it attains a maximum – it will be in a value $t_0 > t_1$ where t_1 is the value in which the function

$$\alpha(t) = \frac{c(t)}{p(t)} \qquad (3)$$

attains its maximum.

In Egghe (1988) the model (2) was extended to the case that time t is a continuous variable: $t \in \mathbb{R}^+$. Now $c(t)$ denotes the density of citations at time t and $p(t)$ denotes the density of articles at time t, i.e. $c(t)$ divided by the total number of citations C and $p(t)$ divided by the total number of articles P, are probability densities. Hence we have that

$$C = \int_0^{\infty} c(t')dt' \tag{4}$$

and

$$P = \int_0^{\infty} p(t')dt' \tag{5}$$

are, respectively, the total number of citations and the total number of articles. Definition (2) can now be adapted as follows: for every $t \in \mathbb{R}^+$ define

$$IF(t) = \frac{\int_0^t c(t')dt'}{\int_0^t p(t')dt'} \tag{6}$$

as the continuous t impact factor. The above result of Rousseau on the "retarted" maximum of $IF(t)$ with respect to $\alpha(t)$ (for discrete t) has been proved in Egghe (1988) to be also correct for continuous t. In Egghe (1988) it is also shown that $IF'(t)$ has the same sign as $\alpha(t)$- $IF(t)$ which has as a consequence that $IF(t)$ decreases in t if $\alpha(t) < IF(t)$ for all $t \in \mathbb{R}^+$. This is for instance the case if $\alpha(t) = \frac{c(t)}{p(t)}$ strictly decreases in t since then, by (3) and (6), $\alpha(t) < IF(t)$ for all t obviously. The proofs are given in the Appendix for the sake of completeness.

The condition

$$\alpha(t) = \frac{c(t)}{p(t)} \text{ strictly decreases in } t \qquad (7)$$

is "logical" from a citation point of view: it expresses the "property" that an average article is less cited the older it is. This is clearly true in almost all cases except for a starting period (indicated above as $[0, t_0]$ for $IF(t)$ and $[0, t_1]$ for $\alpha(t)$) in which "one has to study and understand an article and one has to write a new paper in which the former article has been used (cited)". In this connection we speak of delay times (cf. Egghe and Rousseau (2000)).

These remarks show that IFs based on one single time period for all journals and fields (e.g. $t = 2$ as in the JCR) have a major drawback: a single $IF(t^*)$ can "measure" a journal or a field at a point t* where the function $IF(t)$ is increasing, has its maximum or where the function $IF(t)$ is decreasing (see the vertical lines in Fig. 1). For two different fields or journals the corresponding intervals of increase and decrease can be different and hence so is the time t_0 at which $IF(t)$ reaches its maximum. This is also illustrated in Fig. 1.

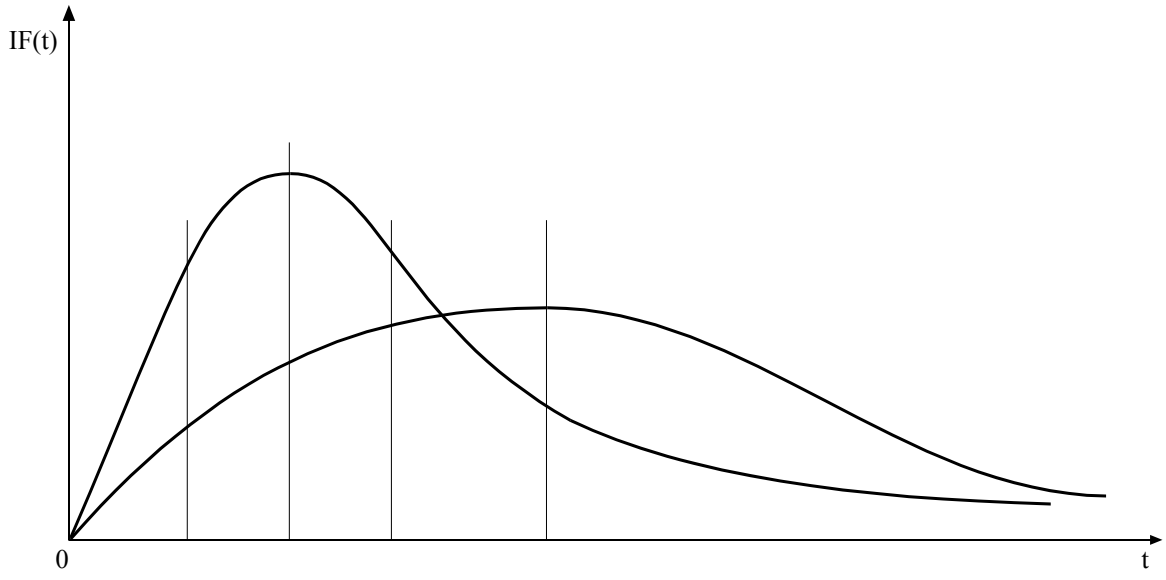


Fig. 1 IF-curves for 2 different fields or journals.

Based on Fig. 1, how can the two fields or journals be compared with respect to IF ? Certainly, no vertical line (fixed t) measures the "right" $IF(t)$ for both fields or journals and using $IF(t_0)$, where the IF-curve reaches its maximum requires the calculation of IF for different t-values (and dependent on each field or journal). This clearly illustrates the problems with $IF(t)$. The next subsection also indicates another problem that is present in IF: the fact that IF is not normalized with respect to P and C.

(ii) The not-normalized nature of all IF(t)-values

From the definition of $IF(t)$ (formula (2) or (6)) it is clear that $IF(t)$ is normalized with respect to the "size" of the journal in the sense that the number of citations in the period $\{1,2,...,t\}$ or $[0,t]$ is divided by the number of publications in that period. However there is no normalization for the total number of citations (C) or publications (P). Here "total" can mean different things. Usually C and P refer to the field totals in the period under study (see further) but, for a journal, it can also refer to the total number of citations to this journal and the total number of publications in this journal (again, see further).

There should at least be a normalization with respect to $\frac{C}{P}$, the average number of citations per article in the field or the journal. Indeed, for a field, the value $\frac{C}{P}$ can be very high or very low. The effect of this non-normalized aspect of any $IF(t)$ is e.g. clear when considering the Subject Category Listings in the JCR. Top impact factors $IF(2)$ of journals in e.g. biochemistry are about 15 times larger than top impact factors $IF(2)$ of journals in mathematics which makes direct comparison of IFs impossible, also in this context. Some science evaluators use percentiles (quartiles) to weight journals in such different Subject Category Listings, a method that is applicable in any field. E.g. the author's main university (LUC) uses this method for about a decade (see Rousseau and Smeyers (2000)). We can refer to Pudovkin and Garfield (2004) for a (late) promotion of this methodology. However, there is an easy way to overcome the problem mentioned in this part (ii): the introduction of the so-called "relative impact factor (RIF)". Such a RIF was first introduced by Braun, Glänzel and Schubert in a number of publications in the eighties: Schubert, Glänzel and Braun (1983),

Braun, Glänzel and Schubert (1985), Schubert, Glänzel and Braun (1986) and Braun, Glänzel and Schubert (1989) (see also Egghe and Rousseau (2003)). The simple definition is as follows: for any journal, belonging to a field (e.g. expressed by a set of journals, e.g. ISI's Subject Category Listing), let IF be any impact factor as defined above and let C and P denote the total number of citations, respectively publications (to be defined in an exact way in the sequel). Then define the relative impact factor

$$RIF = \frac{IF}{\mu} \tag{8}$$

where

$$\mu = \frac{C}{P} \tag{9}$$

denotes the average number of citations per publication.

Examples:

1. Let a field (e.g. in the Subject Category Listing in JCR) consist of N journals. Let the i^{th} journal $(i = 1,...,N)$ have C_i citations and P_i publications (as e.g. calculated for a 2-year impact factor, but other values than 2 are equally possible). Hence $IF_i = \frac{C_i}{P_i}$ is the impact factor of this i^{th} journal. Its relative impact factor RIF_i (see e.g. Egghe and Rousseau (2003)) is then

$$RIF_i = \frac{IF_i}{\mu} \tag{10}$$

where

$$\mu = \frac{C}{P} = \frac{\sum_{j=1}^{N} C_j}{\sum_{j=1}^{N} P_j} \tag{11}$$

is the average number of citations per publication in the whole field.

In Egghe and Rousseau (2003) one promotes the idea of publishing in ISI's publications (e.g. the JCR), besides the two-year impact factors $IF_i(2)$ of each journal i in a field, also the $RIF_i(2)$: such RIF_is are directly comparable also over the different fields (see also Braun, Glänzel and Schubert (1985), p. 47) and can be published as a second column next to the $IF_i(2)$ column in the Subject Category Listing: indeed, the multiplication of each IF_i by $\frac{P}{C}$ (a constant) does not change the order of the journals according to IF_i in the Subject Category Listing.

2. For a single journal (or even a field considered as a meta-journal) one can normalise the IF with respect to the total citation and publication volume of the journal. So let $IF(t)$ be as in (2), i.e. the total number of citations to this journal to 1,2,…,t years ago divided by the total number of publications in this journal in these years. Normalizing over all citations (in year 0) to all years $i = 1,2,3,...$ ago of this journal and with respect to all articles in this journal 1,2,3,… years ago we can hence define

$$RIF(t) = \frac{\sum_{i=1}^{t} c(i)}{\sum_{i=1}^{t} p(i)} \Big/ \frac{\sum_{i=1}^{\infty} c(i)}{\sum_{i=1}^{\infty} p(i)} \qquad (12)$$

Using formula (6) this is

$$RIF(t) = \frac{\frac{\int_0^t c(t')dt'}{\int_0^\infty c(t')dt'}}{\frac{\int_0^t p(t')dt'}{\int_0^\infty p(t')dt'}} \qquad (13)$$

This formula will play an important role in this paper as will become clear in the sequel. As far as we know is the interpretation of (12) and (13) as a RIF new. It must be emphasized that Example 2 is different from Example 1 in that only one journal is involved in Example 2 where normalization is done with respect to "all years", while in Example 1 normalization is done with respect to "all journals" in a field, keeping the time period fixed. The above definition (12), (13) can serve to overcome the two problems (i) and (ii), discussed above, as the following heuristic argument shows.

Formulae (12) and (13) clearly normalise each journal: each numerator and denominator is limited to the interval $[0,1]$, for every $t \in \mathbb{R}^+$. To overcome the time problem as discussed in (i) we can simply say that, in theory, we should have at our disposal all $IF(t)$s (or rather $RIF(t)$s) for all $t \in \mathbb{R}^+$. In other words, we do not only need $RIF(t)$ as numbers but the curve

$$t \to RIF(t) \qquad (14)$$

for $t \in \mathbb{R}^+$. If, however, one looks at the numerator and denominator of (12) and (13) separately we see that the numerator gives a fraction of citations to this journal while the denominator gives a fraction of publications in this journal. So, instead of (14) we could look for a curve

$$\text{fraction of publications} \to \text{fraction of citations} \qquad (15)$$

hence a curve that belongs to the unit square $[0,1]x[0,1]$. This directly links these ideas to the Lorenz curves which give cumulative fractions of items (e.g. citations) in

function of cumulative fractions of sources (e.g. publications). We refer to Egghe (2002, 2004, 2005) and references therein for detailed treatments of Lorenz theories.

In the next section we will briefly overview Lorenz theory and we will introduce weighted Lorenz curves of a continuous variable. As an application of this we introduce the impact Lorenz curves in which impact factors, dependent on fraction of publications or on fraction of citations (cf. (15)), can be calculated. These ideas are linked with ideas developed in Sombatsompop, Markpin and Premkamolnetr (2004) where "median" impact factors are introduced. General properties of such impact Lorenz curves are proved.

In the third section we explicitely calculate the functional form of such impact Lorenz curves in the special (basic case) that we have an exponential aging curve for citations and an exponential growth curve for publications. We show that the quotient of the logarithms of these rates are completely determining the impact Lorenz curve and applications are given: we will show that, whenever we have two such curves (e.g. for two journals) these Lorenz curves are never intersecting (except in (0,0) and (1,1)), an important conclusion in Lorenz theory. This gives, as a consequence that, given two such curves, all impact factors dependent on fractions (of publications or of citations) of one of such situations are larger than all impact factors dependent on fractions of the other situation. We also show by counterexample that this is not true (even when using the simple exponential models) for impact factors calculated with a fixed time period.

In the fourth section we present four methods to calculate the above mentioned quotient of the logarithms of the aging and growth rates (hence determining the impact Lorenz curves) and examples are given.

II. Weighted Lorenz curves of a continuous variable

We first give a brief overview of unweighted Lorenz curves of a discrete or of a continuous variable. We refer the reader to Egghe (2002, 2004, 2005), Egghe and Rousseau (1990) for more details.

II.1 Lorenz curves of a discrete variable

Let the vector $X = \{x_1, x_2, ..., x_N\}$ be given. For an easy argument we suppose that all x_i are positive (or zero) and that we have ordered the vector X decreasingly. For each $i = 1, ..., N$, x_i can be considered as the production (or money earned) of the i^{th} employee in a company but could also be considered as the number of articles in the i^{th} journal (given N journals in a field) or could even be considered as the number of citations to (or from) the i^{th} article (in a set of N articles). The Lorenz curve of X, denoted L_X, is the polygonal curve connecting (0,0) consecutively with the points

$$\left(\frac{i}{N}, \sum_{j=1}^{i} a_j\right) \tag{16}$$

where

$$a_j = \frac{x_j}{\sum_{k=1}^{N} x_k} \tag{17}$$

Hence L_X connects the points with abscissa: normalized cumulative number of sources (up to i, $i = 1, ..., N$) and with ordinate: normalized cumulative number of items in these sources. Fig. 2 illustrates this construction. We obtain a concave curve since the x_i are decreasing.

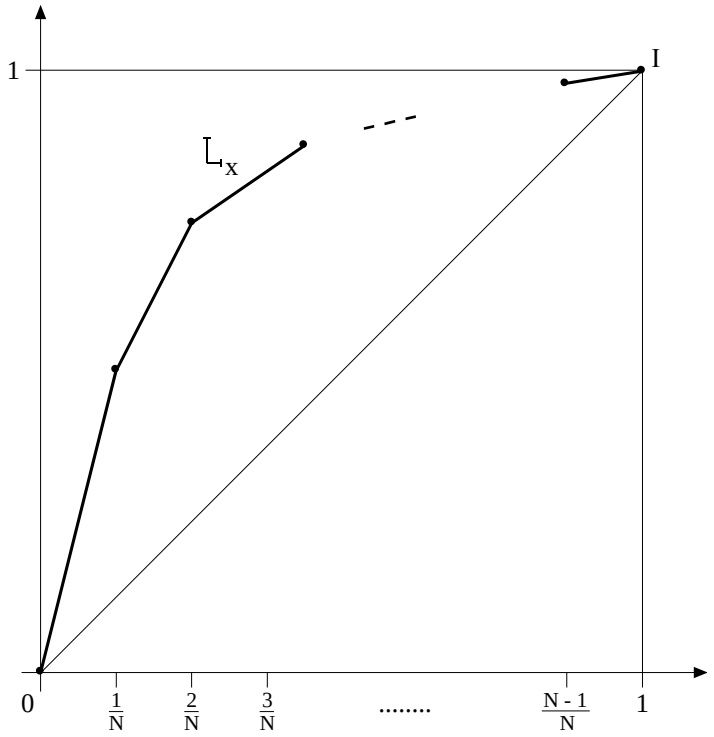


Fig. 2 A general discrete unweighted Lorenz curve.

II.2 Lorenz curves of a continuous variable

Now the "production" vector $X = (x_i)_{i=1,...,N}$ in the discrete variable $i \in \{1,...,N\}$ is replaced by a positive decreasing integrable function h of a continuous variable $x \in [a,b]$ (a,b being the minimal, respectively the maximal possible value of x, but a can be $-\infty$ and b can be $+\infty$). Now the Lorenz curve, denoted $L(h)$, of such a situation is given by the points

$$\left(\frac{x-a}{b-a}, \frac{\int_a^x h(x')dx'}{\int_a^b h(x')dx'} \right), \tag{18}$$

a concavely increasing function connecting (0,0) with (1,1). Fig. 3 illustrates this.

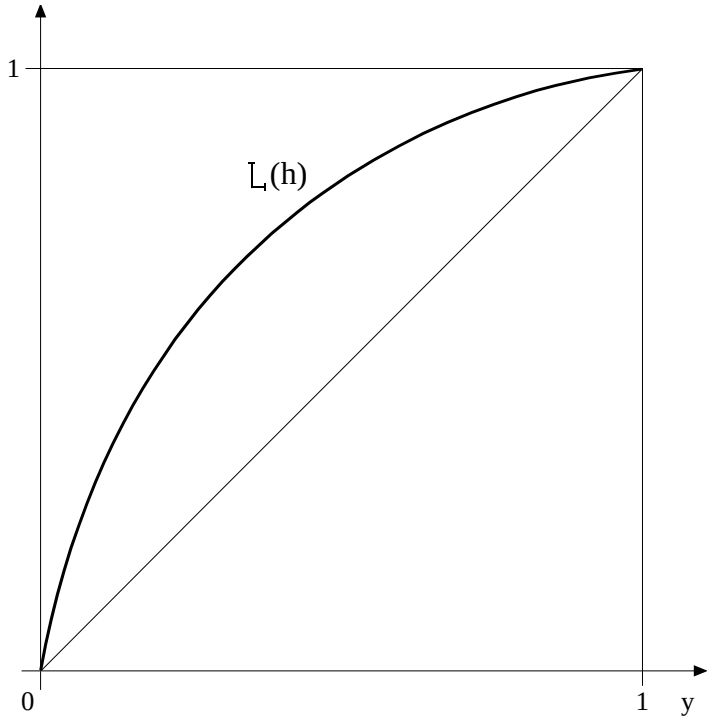


Fig. 3 A general continuous unweighted Lorenz curve.

In other words, putting

$$y = \frac{x - a}{b - a} \in [0,1] \tag{19}$$

hence

$$x = y(b - a) + a, \tag{20}$$

the Lorenz curve of the function h is the function $L(h)$, where

$$L(h)(y) = \frac{\int_a^{y(b-a)+a} h(x')dx'}{\int_a^b h(x')dx'} \tag{21}$$

II.3 Weighted Lorenz curves of a discrete variable

Here the uniform distribution $\left(\frac{1}{N},...,\frac{1}{N}\right)$ (N times) as abscissa in Subsection II.1 is replaced by a weight vector $W = (w_1,...,w_N)$, where $w_i \geq 0$ for each $i = 1,...,N$ and where

$$\sum_{i=1}^{N} w_i = 1 \tag{22}$$

The weighted Lorenz curve of a vector $X = (x_1,...,x_N)$, weighted with the vector $W = (w_1,...,w_N)$ is constructed as follows. We first rearrange X so that

$$\frac{x_1}{w_1} \geq \frac{x_2}{w_2} \geq ... \geq \frac{x_N}{w_N} \tag{23}$$

We then connect (0,0) with the consecutive points

$$\left(\sum_{j=1}^{i} w_j, \sum_{j=1}^{i} a_j\right), \tag{24}$$

where a_j is as in (17). Due to (23) we again obtain a concavely increasing curve, called the weighted Lorenz curve of X (w.r.t. X) and denoted as $L_{X,W}$. Fig. 4 illustrates this.

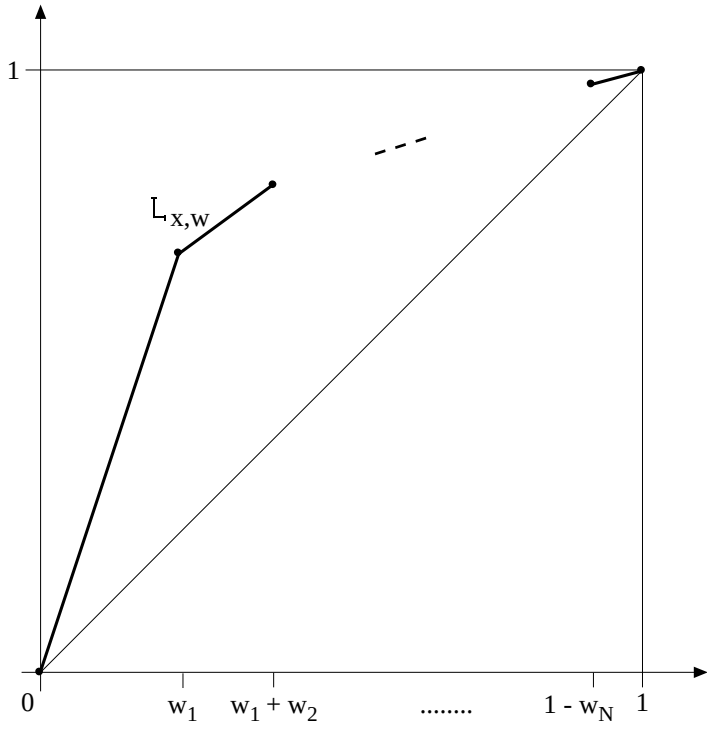


Fig. 4 A general discrete weighted Lorenz curve.

For an application of weighted Lorenz curves we refer the reader to Egghe and Rousseau (2001).

II.4 Weighted Lorenz curves of a continuous variable

Here the continuous uniform distribution $\frac{1}{b-a}$ on $[a,b]$ of Subsection II.2 is replaced by a general distribution of the form, for $x \in [a,b]$ and $v \geq 0$ integrable on $[a,b]$:

$$w(x) = \frac{v(x)}{\int_a^b v(x')dx'} \qquad (25)$$

(hence $\int_a^b w(x')dx' = 1$, comparable with (22)). The weighted Lorenz curve of h (weighted by v), denoted as $L(h,v)$ is the curve

$$\left(\int_a^x w(x')dx', \int_a^x g(x')dx'\right), \tag{26}$$

where g is the normalized form of h:

$$g(x) = \frac{h(x)}{\int_a^b h(x')dx'} \tag{27}$$

Here, as in Subsection II.3 (23), we suppose that the function

$$\alpha(x) = \frac{h(x)}{v(x)} \tag{28}$$

decreases in x. This then leads to a concavely increasing curve $L(h,v)$ between (0,0) and (1,1). This will be proved now.

Proposition II.4.1:

$L(h,v)$ is a continuous concave increasing function between (0,0) and (1,1).

Proof:

That (0,0), (1,1)Î $L(h,v)$ is clear from (25), (26) and (27). Now the function $y \to L(h,v)(y)$ has

$$y = y(x) = \frac{\int_a^x v(x')dx'}{\int_a^b v(x')dx'} \tag{29}$$

as independent variable and

$$L(h,v)(y) = \frac{\int_a^x h(x')dx'}{\int_a^b h(x')dx'} \tag{30}$$

as dependent variable. Hence the function $L(h,v)$ is continuous. We have

$$\frac{dL(h,v)}{dy} = \frac{dL(h,v)}{dx}\frac{dx}{dy}$$

$$= \frac{h(x)}{\int_a^b h(x')dx'}\frac{d\varphi^{-1}(y)}{dy} \tag{31}$$

by (30) and for the function $y = \varphi(x)$ as in (29). Hence

$$\frac{d\varphi^{-1}(y)}{dy} = \frac{1}{\varphi'(\varphi^{-1}(y))}$$

$$= \frac{1}{\frac{v(x)}{\int_a^b v(x')dx'}}$$

$$= \frac{\int_a^b v(x')dx'}{v(x)} \tag{32}$$

using (29). Hence, putting (32) in (31) we have, as a variable of x:

$$\frac{dL(h,v)}{dy} = \frac{\int_a^b v(x')dx'}{\int_a^b h(x')dx'}\frac{h(x)}{v(x)} \tag{33}$$

Since (33) is positive, $L(h,v)$ increases and since $\alpha = \frac{h}{v}$ decreases (by (28)) we have that $L(h,v)$ is a concavely increasing function of y (i.e. the same graph as in Fig. 3). ~

II.5 Application of weighted Lorenz curves of a continuous variable: Impact Lorenz curves

Suppose we take $t = 0$ as the present time. Let us fix a journal. In continuous time t to the past, let $c(t)$ be the density of citations from time 0 to time t to this journal and let $p(t)$ be the density of publications at time t (again t refers to the past). This means that, for every $t_1, t_2 \geq 0$, $t_1 < t_2$,

$$\int_{t_1}^{t_2} c(t')dt'$$

denotes the number of citations to this journal, given at t = 0 (say by a group of journals that is fixed) and

$$\int_{t_1}^{t_2} p(t')dt'$$

denotes the number of articles in this journal in the time period $[t_1, t_2]$.

Definition II.5.1:

The Impact Lorenz curve of a journal is the Lorenz curve $L(h,v)$, as defined in Subsection II.4 for the functions h = c and v = p and for a = 0 and $b = +\infty$ (in this case, as indicated in Subsection II.2, we use the interval $[0,+\infty]$). Of course an extension to other values of a and b is possible but we do not need this in this paper. As needed in Subsection II.4 (cf. (28)) we need here to suppose that (we change the variable x into t = time here)

$$\alpha(t) = \frac{c(t)}{p(t)} \tag{34}$$

decreases in t. As discussed in the introduction, this is a natural requirement, stating that the citation density per article decreases in time (to the past): this is always so, not taking into account a short initial increase of α due to the fact that there is an intial time period in which young articles are studied and new results are published based on these articles (which then receive a citation). In this connection we can speak of delay times (cf. Egghe and Rousseau (2000)).

What is the use of the impact Lorenz curve $L(c,p)$?

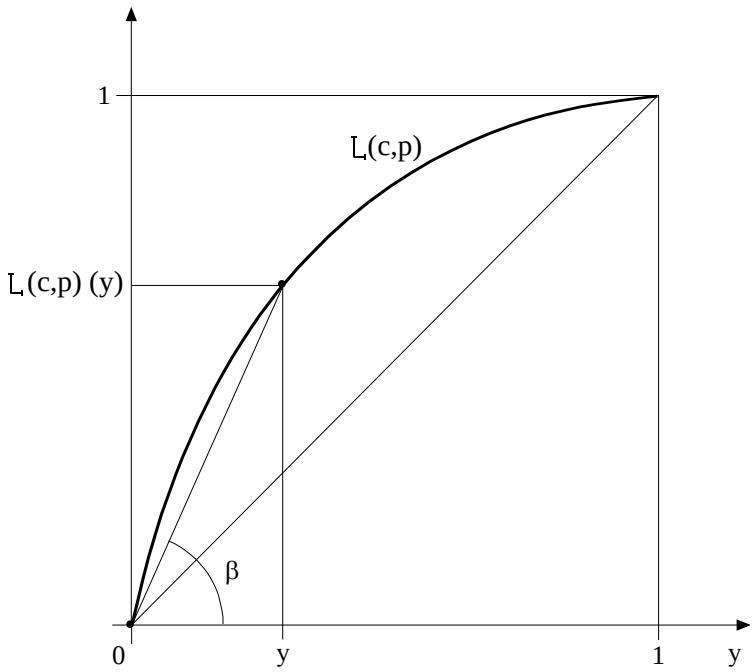


Fig. 5 A general impact Lorenz curve, showing a fractional publication RIF.

First of all note that $t = 0 \Rightarrow y = 0$ and $t = +\infty \Rightarrow y = 1$. Both results follow from (29) using that $a = 0$ and $b = +\infty$ and the notation $x = t$. We have, denoting

$$P = \int_0^{\infty} p(t')dt' \qquad (35)$$

$$C = \int_0^{\infty} c(t')dt' \tag{36}$$

the total number of publications in the journal, respectively citations to this journal (given at time $t = 0$) that

$$L(c,p)(y) = \frac{\int_0^{t} c(t')dt'}{C} \tag{37}$$

with

$$y = \frac{\int_0^{t} p(t')dt'}{P} \tag{38}$$

using (29) and (30). Using (33) gives

$$\frac{dL(c,p)}{dy} = \frac{P}{C}\frac{c(t)}{p(t)} \tag{39}$$

as function of time t (relation (38)).

So we have that (using that $t = 0 \Rightarrow y = 0$ and $t = +\infty \Rightarrow y = 1$)

$$\frac{dL(c,p)}{dy}(0) = \frac{P}{C}\frac{c(0)}{p(0)} \tag{40}$$

and that

$$\frac{dL(c,p)}{dy}(1) = \frac{P}{C}\lim_{t\to+\infty}\frac{c(t)}{p(t)}. \tag{41}$$

Each point $(y, L(c,p)(y))$ on $L(c,p)$ determines an angle β as indicated in Fig. 5. It follows that, by (37) and (38)

$$\mathrm{tg}\beta = \frac{L(c,p)(y)}{y} = \frac{P}{C}\frac{\int_0^t c(t')dt'}{\int_0^t p(t')dt'} \tag{42}$$

So we have that, by (13),

$$\mathrm{tg}\beta = \frac{L(c,p)(y)}{y} = \mathrm{RIF}(t), \tag{43}$$

the relative impact factor of this journal at time t, where y and t are related as in (38). Hence the impact Lorenz curves comprises relative impact factors and comprises all relative impact factors for all $t \in [0, +\infty]$ (again since $t = 0 \Rightarrow y = 0$ and $t = +\infty \Rightarrow y = 1$). Hence the knowledge of this curve solves both problems (i) and (ii) of impact factors discussed in Section I.

The following remark is important. By (43) we have indeed that all $\mathrm{RIF}(t)$ are contained in $L(c,p)$ but in an "implicite" way: in $L(c,p)$ we directly read impact factors RIF but in the variable y which then relates to $\mathrm{RIF}(t)$ for t given by (38). So we can as well say that we have found relative impact factors RIF as being dependent on y, for all $y \in [0,1]$. We can denote these as $\mathrm{RIF}[y]$.

Definition II.5.2:

The impact factors $\mathrm{RIF}[y]$ are defined as the fractional publication relative impact factors of the journal, meaning that it is the relative impact factor $\mathrm{RIF}(t)$ at time t (into the past) at which the journal has 100y% of its publications. As said above, y and t are related by (38).

We have the easy proposition:

Proposition II.5.3:

$$\{RIF(t) \| t \in [0,+\infty]\} = \{RIF[y] \| y \in [0,1]\} \qquad (44)$$

Proof:

Let $t \in [0,+\infty]$. Then

$$RIF(t) = RIF[y]$$

with y the unique value determined by (38). Since obviously $y \in [0,1]$ we have proved that

$$\{RIF(t) \| t \in [0,+\infty]\} \subseteq \{RIF[y] \| y \in [0,1]\}.$$

For the other inclusion, let $y \in [0,1]$. Since the function

$$t \to \int_0^t p(t')dt'$$

is a strictly increasing function, it is injective. Hence there exists a unique value $t \in [0,+\infty]$ such that (38) is valid. Hence, for this value of t, by (43),

$$RIF[y] = RIF(t). \qquad \sim$$

It is clear that all fractional publication relative impact factors are found, graphically by drawing, as in Fig. 5, a vertical line at abscissa y and then calculate $tg\beta$ as indicated.

Analogously we could draw a horizontal line (say at ordinate η) (see Fig. 6), again determining a point on $L(c,p)$.

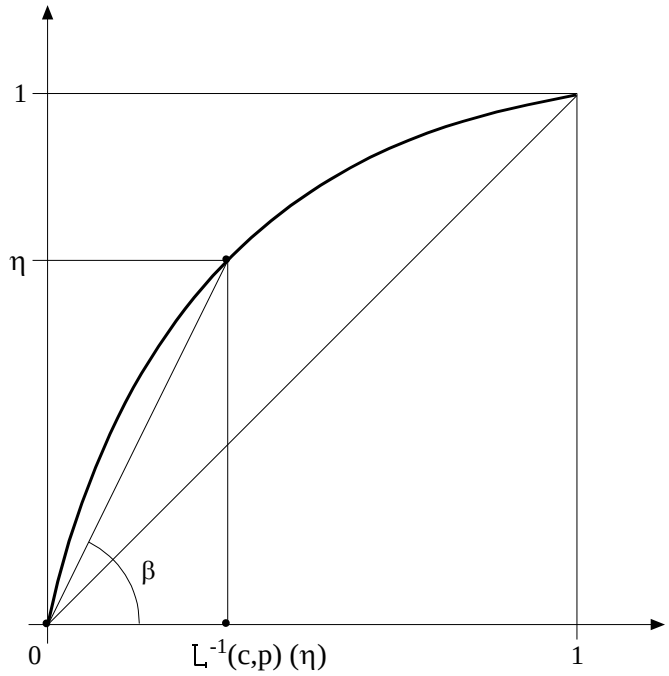


Fig. 6 A general Lorenz curve, showing a fractional citation RIF.

In the same way as above this determines an abscissa, being $L^{-1}(c,p)(\eta)$ and an angle β as indicated in Fig. 6. We have now

$$\mathrm{tg}\beta = \frac{\eta}{L^{-1}(c,p)(\eta)} = \mathrm{RIF}(t), \tag{45}$$

the relative impact factor of this journal at time t, where η and t are related as in (37) (η being the value expressed in (37)). We can denote these as $\mathrm{RIF}\{\eta\}$. So in $L(c,p)$ we also find all relative impact factors in the following sense.

Definition II.5.4:

The impact factors $\mathrm{RIF}\{\eta\}$ are defined as the fractional citation relative impact factors of the journal, meaning that it is the relative impact factor $\mathrm{RIF}(t)$ at time t (into the past) at which the journal has received $100\eta\%$ of its citations. As said above, η and t are related as in (37).

We again have the following easy proposition.

Proposition II.5.5:

$$\{RIF(t) \| t \in [0,+\infty]\} = \{RIF\{\eta\} \| \eta \in [0,1]\} \tag{46}$$

Proof:

Let $t \in [0,+\infty]$. Then

$$RIF(t) = RIF\{\eta\}$$

with η the unique value determined by (37):

$$\eta = \frac{\int_0^t c(t')dt'}{C} \tag{47}$$

Since obviously $\eta \in [0,1]$, we have proved that

$$\{RIF(t) \| t \in [0,+\infty]\} \subseteq \{RIF\{\eta\} \| \eta \in [0,1]\}.$$

For the other inclusion, let $\eta \in [0,1]$. Since

$$t \to \int_0^t c(t')dt'$$

is a strictly increasing function, it is injective. Hence there exists a unique value $t \in [0,+\infty]$ such that (47) is valid. Hence, by (37) and (43) we have that

$$RIF\{\eta\} = RIF(t). \qquad \sim$$

In Sombatsompop, Markpin and Premkamolnetr (2004) an impact factor is defined at time t (into the past) such that a journal has received 50% of its citations. In other words, they defined the non-normalized impact factor

$$IF\left(\frac{1}{2}\right) = \frac{C}{P} RIF\left(\frac{1}{2}\right) \tag{48}$$

which was clarified in Rousseau (2004). In fact it was the introduction of such a new type of impact factor, based on fractions (of citations) that led the present author to introduce impact factors in the framework of Lorenz curves.

Lorenz curves measure the degree of inequality between the numbers $L(c,p)(y)$ for $y \in [0,1]$ (see Egghe (2002, 2004, 2005): the higher the Lorenz curve the more concentrated (unequal) these numbers are). Suppose we have two journals with impact Lorenz curves L_1, respectively L_2. We have the following trivial proposition.

Proposition II.5.6:
The following assertions are equivalent:

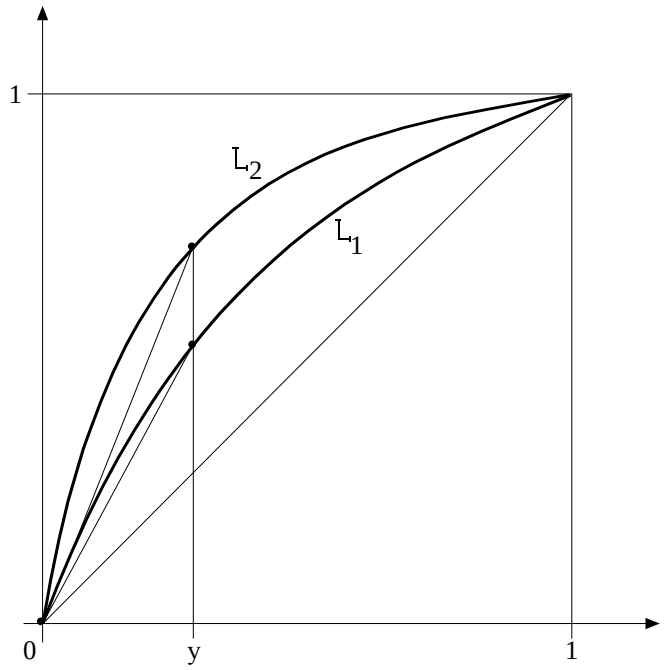
(i) $L_1 < L_2$ (meaning $L_1 \leq L_2$ and $L_1 \neq L_2$ except in (0,0) and (1,1))

(ii) $RIF_1[y] < RIF_2[y], \quad \forall y \in]0,1[$

(iii) $RIF_1\{\eta\} < RIF_2\{\eta\}, \quad \forall \eta \in]0,1[$.

The same is true when < is replaced by =.

Proof:
This is clear by graphical inspection of Figs. 7a,b. ~



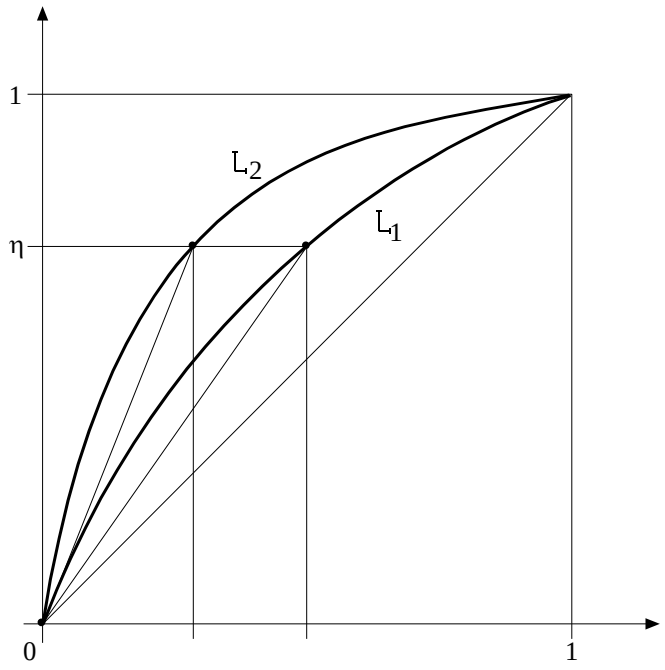


Fig. 7a,b Graphical proof of Proposition II.5.6.

III. Impact Lorenz curves for exponential aging curves (for citations) and exponential growth curves (for publications)

In this section we will shed more light on the explicite form of the impact Lorenz curves. This is very important if we want to use them in science evaluation and science policy studies as an improvement of the simple use of fixed time impact factors (e.g. IF(2)). In order to obtain basic results we will suppose the simplest functional relations for the functions c(t) and p(t), introduced in Subsection II.5: the exponential aging function for c(t) and the same for p(t) (with different parameters of course), since an exponential growth function (hence with time going into the future) becomes an exponential aging curve (hence decreasing) when time t goes to the past (starting from the present $t = 0$) as is the case in this paper.

Hence we suppose that the functions p(t) and c(t) of Subsection II.5 have the following form:

$$p(t) = p_0 p^t \tag{47}$$

$$c(t) = c_0 c^t \tag{48}$$

for all $t \geq 0$, where $p_0 = p(0)$ and $c_0 = c(0)$ are initial conditions parameters and where p and c are aging parameters such that $0 < p, c < 1$ since the functions p(t) and c(t) must be decreasing. Due to the condition that α in (34) must decrease we hence suppose that $c < p$.

III.1 Impact Lorenz curves and properties

Based on formulae (35), (36), (37) and (38), we obtain the following formulae

$$y = \frac{\int_0^t p(t')dt'}{\int_0^\infty p(t')dt'} = 1 - p^t \tag{49}$$

$$L(c,p)(y) = \frac{\int_0^t c(t')dt'}{\int_0^{\infty} c(t')dt'} = 1 - c^t \tag{50}$$

Of course, in order to really obtain $L(c,p)$ in function of y we must substitute (49) in (50): by (49) we have

$$t = \frac{\ln(1-y)}{\ln p} \tag{51}$$

Hence

$$L(c,p)(y) = 1 - c^{\frac{\ln(1-y)}{\ln p}} \tag{52}$$

or, in a form that is easier to interpret:

$$L(c,p)(y) = 1 - (1-y)^{\frac{\ln c}{\ln p}} \tag{53}$$

This important formula shows that $L(c,p)$ is only dependent on one parameter, namely

$$\frac{\ln c}{\ln p} \tag{54}$$

and not on both c and p and certainly not on p_0 or c_0. This, in turn has an important consequence: the fact that, for any two situations (e.g. any two journals) the respective Lorenz curves are never intersecting (except, of course in (0,0) and (1,1) since this are points on any Lorenz curve). The following theorem is stating this explicitely.

Theorem III.1.1:

For any two situations, where we have p_1, c_1 as aging parameters in the first case and p_2, c_2 as aging parameters in the second case, we have (denoting by L_1 respectively L_2 the impact Lorenz curves in situation 1 and 2).

(i) $$L_1 < L_2 \Leftrightarrow \frac{\ln c_1}{\ln p_1} < \frac{\ln c_2}{\ln p_2} \tag{55}$$

(ii) $$L_1 > L_2 \Leftrightarrow \frac{\ln c_1}{\ln p_1} > \frac{\ln c_2}{\ln p_2} \tag{56}$$

(iii) $$L_1 = L_2 \Leftrightarrow \frac{\ln c_1}{\ln p_1} = \frac{\ln c_2}{\ln p_2} \tag{57}$$

Proof:

By (53) it is clear that only (i) needs to be proved. Now, by (53)

$$L_1 < L_2$$

$$\Leftrightarrow (1-y)^{\frac{\ln c_1}{\ln p_1}} > (1-y)^{\frac{\ln c_2}{\ln p_2}}$$

$$\Leftrightarrow e^{\ln(1-y)^{\frac{\ln c_1}{\ln p_1}}} > e^{\ln(1-y)^{\frac{\ln c_2}{\ln p_2}}}$$

$$\Leftrightarrow \ln(1-y)\frac{\ln c_1}{\ln p_1} > \ln(1-y)\frac{\ln c_2}{\ln p_2}$$

$$\Leftrightarrow \frac{\ln c_1}{\ln p_1} < \frac{\ln c_2}{\ln p_2}$$

since $1-y \in]0,1[$. ~

Corollary III.1.2:

Let us have two situations as above.

(i) Suppose there exists an $y_0 \in]0,1[$ such that $RIF_1[y_0] < RIF_2[y_0]$. Then $RIF_1[y] < RIF_2[y], \forall y \in]0,1[$.

(ii) Suppose there exists an $\eta_0 \in]0,1[$ such that $RIF_1\{\eta_0\} < RIF_2\{\eta_0\}$. Then $RIF_1\{\eta\} < RIF_2\{\eta\}, \forall \eta \in]0,1[$.

Proof:

This follows trivially from Theorem III.1.1 and Proposition II.5.6: if (in case of (i)) $RIF_1[y_0] < RIF_2[y_0]$ for a certain $y_0 \in]0,1[$ then $L_1 \not> L_2$ by Proposition II.5.6 and obviously $L_1 \neq L_2$ also. According to Theorem III.1, we have that (ii) and (iii) in this theorem are not possible. Hence only (i) in this theorem is possible: $L_1 < L_2$ which implies, by Proposition II.5.6 that $RIF_1[y] < RIF_2[y], \forall y \in]0,1[$. The same proof goes for (ii). ~

Corollary III.1.3:

Suppose (i) or (ii) in the above Corollary III.1.2 is valid (in fact they are equivalent). Then we have that the inequality (concentration) between the values

$$\{yRIF_1[y] \| y \in [0,1]\} = \{L_1(c_1, p_1)(y) \| y \in [0,1]\}$$

is smaller than the inequality (concentration) between the values

$$\{yRIF_2[y] \| y \in [0,1]\} = \{L_2(c_2, p_2)(y) \| y \in [0,1]\}$$

Proof:

This follows immediately from the fact that $L_1 < L_2$ and this follows from Proposition II.5.6.

~

The rest of this Subsection III.1 is devoted to the construction of examples showing that Corollary III.1.2 is false for RIFs calculated with fixed times t instead of using fixed values of y or of η.

Construction III.1.4:

We will construct examples of 2 situations, where we have two time values $t_1, t_2 \geq 0$ such that

$$RIF_1(t_1) < RIF_2(t_1) \tag{58}$$

and

$$RIF_1(t_2) > RIF_2(t_2) \tag{59}$$

, hence clearly showing that (i) nor (ii) in Corollary III.1.2 are true for RIFs in fixed time values t. We are able to present such examples even in the simple case of exponential aging as used in this section. We will also present the methodology with which we arrived at such examples. Let us fix $t \geq 0$. According to (49) and (50) and also using (43) we have that, for one situation with aging parameters p, c, $RIF(t)$ is given by

$$RIF(t) = \frac{1 - c^t}{1 - p^t} \tag{60}$$

For two such situations (with parameters p_1, c_1 and p_2, c_2 respectively) we hence have that the function

$$f(t) = RIF_2(t) - RIF_1(t)$$

$$f(t) = \frac{1 - c_2^t}{1 - p_2^t} - \frac{1 - c_1^t}{1 - p_1^t} \tag{61}$$

is the key function for which we must find examples such that $f(t_1) > 0$ and $f(t_2) < 0$ (according to (58) and (59)).

Note first that

$$f(0) = \lim_{x \to 0^+} \left(\frac{1 - c_2^t}{1 - p_2^t} - \frac{1 - c_1^t}{1 - p_1^t} \right)$$

$$f(0) = \frac{\ln c_2}{\ln p_2} - \frac{\ln c_1}{\ln p_1} \tag{62}$$

(use de l'Hôspital's rule) and that

$$\lim_{x \to +\infty} f(t) = 0$$

Formula (62) is interesting since, based on Theorem III.1.1, the value $f(0)$ determines whether $L_1 < L_2$, $L_1 > L_2$ or $L_1 = L_2$ (where $L_i = L_i(c_i, p_i)$, $i = 1,2$, the impact Lorenz curves for the 2 situations). In other words, using Theorem III.1.1 and Proposition II.5.6 we have the following result.

Proposition III.1.4.1:

(i) $f(0) = 0$ iff

$L_1 = L_2$ iff

$RIF_1[y] = RIF_2[y], \ \forall y \in]0,1[$

iff

$RIF_1\{\eta\} = RIF_2\{\eta\}, \ \forall \eta \in]0,1[$

(ii) $f(0) > 0$ iff

$L_1 < L_2$ iff

$RIF_1[y] < RIF_2[y], \ \forall y \in]0,1[$

iff

$RIF_1\{\eta\} < RIF_2\{\eta\}, \ \forall \eta \in]0,1[$

and similarly for $f(0) < 0$.

We will understand how to arrange for (58), (59) in the case (ii) of the above proposition if we have understood the case (i): $f(0) = 0$. Hence this means that, by (62)

$$\frac{\ln c_1}{\ln p_1} = \frac{\ln c_2}{\ln p_2} \tag{63}$$

and that all $RIF_1[y] = RIF_2[y]$ and that all $RIF_1\{\eta\} = RIF_2\{\eta\}$, according to Proposition III.1.4.1.

Now equality (63) is valid if and only if there exists a $k \neq 0$ such that

$$c_2 = c_1^k \tag{64}$$

$$p_2 = p_1^k \tag{65}$$

Then f in (61) has the form

$$f(t) = \frac{1 - c_1^{kt}}{1 - p_1^{kt}} - \frac{1 - c_1^t}{1 - p_1^t} \tag{66}$$

which is positive if $k < 1$ and negative if $k > 1$. This is seen as follows. For every $0 < c < p < 1$, let $g(t)$ be the function

$$g(t) = \frac{1 - c^t}{1 - p^t}. \tag{67}$$

This function decreases in t as follows from the proof in the Appendix, noting that

$$g(t) = \frac{\ln c \int_0^t c^{t'} dt'}{\ln p \int_0^t p^{t'} dt'}$$

and that $\frac{\ln c}{\ln p} > 0$. An alternative proof is given by the fact that

$$g(t) = \frac{L(c,p)(y)}{y}$$

using (49), (50) and by the fact that the function $L(c,p)$ is concave (using that $\frac{c(t)}{p(t)}$ decreases since $0 < c < p < 1$).

The graph of f can be depicted as in Fig. 8, according to the values $k > 1$ or $k < 1$. Two concrete examples:

(i) $c_1 = 0.5 < p_1 < 0.7$, $k = 2$. Then $c_2 = c_1^k = 0.25 < p_2 = p_1^k = 0.49$. Now

$$f(1) = \frac{0.75}{0.51} - \frac{0.5}{0.3} = -0.196 < 0$$

(ii) $c_1 = 0.5 < p_1 < 0.7$, $k = \frac{1}{2}$. Then $c_2 = \sqrt{c_1} = 0.707 < p_2 = \sqrt{p_1} = 0.837$. Now

$$f(1) = \frac{0.293}{0.163} - \frac{0.5}{0.3} = 0.131 > 0.$$

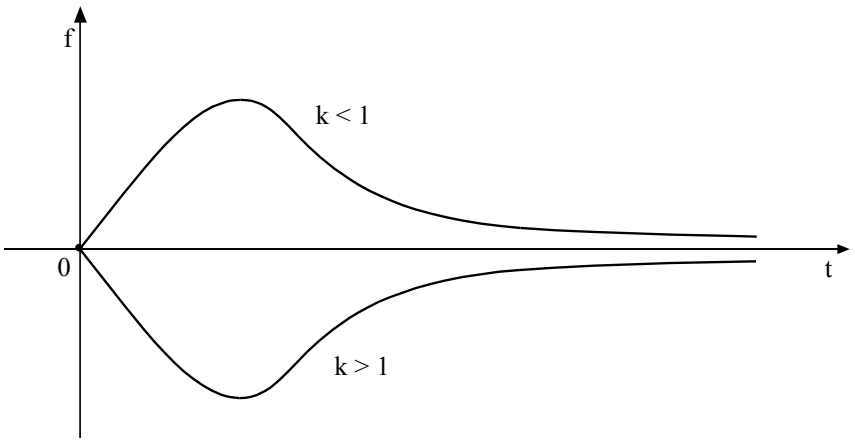


Fig. 8 Graphs of f for $k > 1$ and $k < 1$.

We hence have examples where all fractional relative impact factors in the 2 situations are equal $(L_1 = L_2)$ but where <u>or</u> $IF_1(t) < IF_2(t)$ for all $t > 0$, <u>or</u> $IF_1(t) > IF_2(t)$ for all $t > 0$. It must be clear that this, although interesting in itself, does not provide an example as described in (58) and (59). However, the knowledge of this case (63) gives us insight in how to construct examples showing that (58) and (59) can be true e.g. in the case that $L_1 < L_2$, i.e. all fractional impact factors in the first case are smaller than all fractional impact factors in the second case. Let $L_1 < L_2$, hence

$$1 < \frac{\ln c_1}{\ln p_1} < \frac{\ln c_2}{\ln p_2} \tag{68}$$

by Theorem III.1.1 (and since $0 < c_1 < p_1 < 1$). This means that $f(0) > 0$ instead of $f(0) = 0$ in Fig. 8. The trick will be to increase the beginning of the curve of f for $k > 1$ so that it intersects the t-axis and hence (58) <u>and</u> (59) become possible.

Fig. 9 illustrates this reasoning.

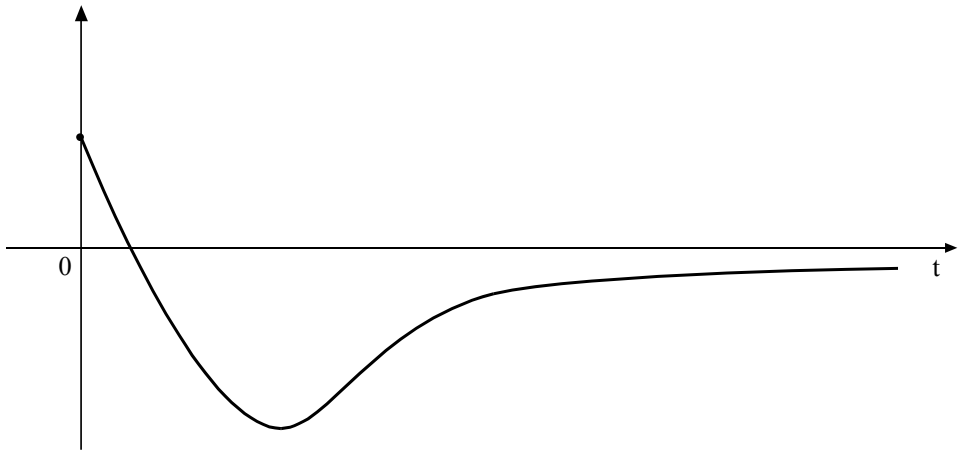


Fig. 9 Visualization of a function f (derived from Fig. 8) that can be positive as well as negative $(k > 1)$.

The choice of the parameters $0 < c_1 < p_1 < 1$ and $0 < c_2 < p_2 < 1$ can be made as follows. Since (68) is valid we have, for a certain $\varepsilon > 0$

$$\frac{\ln c_2}{\ln p_2} = \frac{\ln c_1}{\ln p_1} + \varepsilon$$

$$\frac{\ln c_2}{\ln p_2} = \frac{\ln c_1 + \varepsilon \ln p_1}{\ln p_1} \tag{69}$$

Hence, we can choose $k \neq 0$ freely so that

$$\ln c_2 = k \ln c_1 + k\varepsilon \ln p_1 \tag{70}$$

$$\ln p_2 = k \ln p_1 \tag{71}$$

hence

$$f(t) = \frac{1 - e^{t \ln c_2}}{1 - e^{t \ln p_2}} - \frac{1 - e^{t \ln c_1}}{1 - e^{t \ln p_1}}$$

$$f(t) = \frac{1 - e^{tk \ln c_1} e^{tk\varepsilon \ln p_1}}{1 - e^{tk \ln p_1}} - \frac{1 - e^{t \ln c_1}}{1 - e^{t \ln p_1}} \tag{72}$$

Taking $k > 1$ yields $f(t) < 0$ in (72) if the factor $e^{tk\varepsilon \ln p_1}$ is not "too disturbing" (based on our knowledge of the case (63)). Let us give two pertinent examples.

Examples III.1.4.2:

(i) $c_1 = 0.5 < p_1 = 0.7$, $k = 2$, $\varepsilon = 0.01$. Then we have the values

$$c_2 = c_1^2 p_1^{0.02} = 0.248223$$

$$p_2 = p_1^2 = 0.49$$

Note that $\frac{\ln c_1}{\ln p_1} < \frac{\ln c_2}{\ln p_2}$, hence $L_1 < L_2$. Now

$$f(0.1) = 1.8891795 - 1.9112185 < 0$$

$$f\left(\frac{1}{4}\right) = 1.8008638 - 1.8650319 < 0$$

$$f(1) = 1.4740726 - 1.6666667 < 0$$

$$f(2) = 1.23488 - 1.4705882 < 0$$

$$f(4) = 1.0571459 - 1.233715 < 0$$

but we can have $f(t) > 0$ but only for very small values of t !

$$f(0.01) = 1.9467368 - 1.9400937 > 0.$$

This is a strange example that all fractional relative impact factors in the first case are smaller than all fractional relative impact factors in the second case but that "almost all" ((certainly for all $t \geq 0.1$) relative impact factors $RIF_1(t)$ are larger than $RIF_2(t)$ (since $f(t) = RIF_2(t) - RIF_1(t)$) !

Let us give a last example where the order between $RIF_1(2)$, $RIF_2(2)$ and $RIF_1(4)$, $RIF_2(4)$ is reversed.

(ii) $c_1 = 0.5 < p_1 = 0.6$, $k = 2$, $\varepsilon = 1$. Now

$$c_2 = c_1^2 p_1^2 = 0.09$$

$$p_2 = p_1^2 = 0.36.$$

We have

$$f(2) = 1.139591 - 1.2 > 0$$

and

$$f(4) = 1.0170164 - 1.077091 < 0.$$

Hence

$$RIF_2(2) > RIF_1(2)$$

and

$$RIF_2(4) < RIF_1(4).$$

Note again that, since $L_1 < L_2$ that all fractional relative impact factors in the first case are smaller than all fractional relative impact factors in the second case !

III.2 Explicite formulae for RIF[y] and RIF{0} for all y,00]0,1[

Having established the explicite form of impact Lorenz curves in case of exponential aging and growth we can now easily determine explicite formulae for the fractional relative impact factors $RIF[y]$ and $RIF\{\eta\}$, for all $y, \eta \in]0,1[$. As with the impact Lorenz curve itself we have that they only depend on $\frac{\ln c}{\ln p}$ as will be seen now.

By formula (43) and Definition II.5.2 we have, for every $y \in]0,1[$:

$$RIF[y] = \frac{L(c,p)(y)}{y}$$

$$RIF[y] = \frac{1-(1-y)^{\frac{\ln c}{\ln p}}}{y} \tag{73}$$

using (53). In the same way, by formula (45) and Definition II.5.4 we have, for every $\eta \in]0,1[$:

$$RIF\{\eta\} = \frac{\eta}{L^{-1}(c,p)(\eta)} \tag{74}$$

Let now

$$y = L^{-1}(c,p)(\eta)$$

Then

$$\eta = L(c,p)(y)$$

$$= 1-(1-y)^{\frac{\ln c}{\ln p}}$$

using (53). Hence

$$y = 1-(1-\eta)^{\frac{\ln p}{\ln c}}.$$

This yields in (74):

$$RIF\{\eta\} = \frac{\eta}{1-(1-\eta)^{\frac{\ln p}{\ln c}}} \tag{75}$$

showing again that all we need is an estimate of the parameter $\frac{\ln c}{\ln p}$. This will be executed in Section IV.

III.3 Calculation of the impact Gini index and the impact coefficient of variation

We refer to Egghe (2002, 2004, 2005) for the definition of the Gini index and the square of the coefficient of variation: if $L(y)$ denotes a Lorenz curve we define the Gini index G as

$$G = 2\{\text{area under } L\} - 1 \tag{76}$$

and the square of the coefficient of variation, V^2, as

$$V^2 = \int_0^1 (L'(y))^2 dy - 1 \tag{77}$$

Hence we have, using (53)

$$G = 2\int_0^1 \left[1 - (1-y)^{\frac{\ln c}{\ln p}}\right] dy - 1$$

$$G = 1 - \frac{2}{\frac{\ln c}{\ln p} + 1} \tag{78}$$

(note that $\frac{\ln c}{\ln p} > 1$ since $0 < c < p < 1$). Alternatively we have

$$G = \frac{\frac{\ln c}{\ln p} - 1}{\frac{\ln c}{\ln p} + 1} \tag{79}$$

For V^2 we have, since

$$L'(y) = \frac{\ln c}{\ln p}(1-y)^{\frac{\ln c}{\ln p}-1}$$

that

$$V^2 = \int_0^1 \left(\frac{\ln c}{\ln p}\right)^2 (1-y)^{\frac{2\ln c}{\ln p}-2}\,dy - 1$$

$$V^2 = \frac{\left(\frac{\ln c}{\ln p}-1\right)^2}{2\frac{\ln c}{\ln p}-1} \tag{80}$$

Alternatively we have

$$V = \frac{\frac{\ln c}{\ln p}-1}{\sqrt{2\frac{\ln c}{\ln p}-1}} \tag{81}$$

We leave it as an open problem to interpret these formulae in the framework of (relative) impact factors.

From the above it is clear that it is important to have methods to calculate $\frac{\ln c}{\ln p}$ in any practical situation, even if the data are only approximately exponentially shaped (e.g. where the aging curves show a short initial increase as explained in the Introduction). Indeed, the value $\frac{\ln c}{\ln p}$ determines the properties of the tails of the general distributions $c(t)$ and $p(t)$, the Lorenz curve $L(c,p)$ and the fractional relative impact factors $RIF[y]$ and $RIF\{\eta\}$ for all $y, \eta \in]0,1[$. The estimation of $\frac{\ln c}{\ln p}$ will be done in the next section.

IV. Estimation of $\frac{\ln c}{\ln p}$ in theory and practise

We will develop 4 theoretical models to calculate $\frac{\ln c}{\ln p}$ in the next subsection. In Subsection IV.2 we will test these models on a theoretical as well as on a practical example.

IV.1 Theoretical models to determine $\frac{\ln c}{\ln p}$

IV.1.1 Brookes' method

The following method goes back to Brookes (1970, 1971) for the estimation of the aging rate of an exponentially decreasing function. We also refer to Egghe and Rousseau (1990) but repeat the method here for the sake of completeness (and since the argument is short). We will explain the method on the function $c(t) = c_0 c^t$ (formula (48)) but it works as well on the function $p(t) = p_0 p^t$ (formula (47)).

For any $t_0 \in \mathbb{R}^+$ (we will specify this later on), define

$$k = \int_{t_0}^{\infty} c(t')dt' \tag{82}$$

$$l = \int_{0}^{t_0} c(t')dt' \tag{83}$$

Then

$$k = \int_{t_0}^{\infty} c_0 c^{t'} dt'$$

$$= c^{t_0} \int_{t_0}^{\infty} c_0 c^{t'-t_0} dt'$$

$$= c^{t_0} \int_0^{\infty} c_0 c^{t''} dt''$$

$$k = c^{t_0} (k + l)$$

Hence

$$c = \left(\frac{k}{k + l} \right)^{\frac{1}{t_0}} \tag{84}$$

In practise (see the next subsection) it is best to calculate this for t_0 » the median of the data. This is to make the method stable: taking t_0 too small gives problems with the (in practise) initial increase of the curve $c(t)$ (cf. the Introduction); taking t_0 too large gives too few data for the part $[t_0, \infty[$.

Of course, the same method gives for

$$k^* = \int_{t_0^*}^{\infty} p(t') dt' \tag{85}$$

$$l^* = \int_0^{t_0^*} p(t') dt' \tag{86}$$

that

$$p = \left(\frac{k^*}{k^* + l^*} \right)^{\frac{1}{t_0^*}} \tag{87}$$

Formulae (84) and (87) then yield $\frac{\ln c}{\ln p}$.

Note that in (84), $k + l = C$ (formula (36)) and in (87), $k^* + l^* = P$ (formula (35)). C follows from the JCR Cited journal data but P can be hard to find and one is supposed to go back to volume 1 of the journal which can be difficult. This is the disadvantage of this (and also the following – see Subsection IV.1.2) method. In methods 3 and 4 to come (Subsection IV.1.3 and IV.1.4), the knowledge of C and P is not required.

IV.1.2 Second method

Using (49) and (50) we derive:

$$t \ln p = \ln(1 - y)$$

$$t \ln c = \ln(1 - L(c,p)(y)),$$

for every $t > 0$. So

$$\frac{\ln c}{\ln p} = \frac{\ln(1 - L(c,p)(y))}{\ln(1 - y)} \tag{88}$$

But, according to (49) and (50), y is the fraction of publications in the period $[0,t]$ and $L(c,p)(y)$ is the fraction of citations to the period $[0,t]$. Any $t > 0$ can be taken but for stability reasons it is best to take t as large as possible. If one uses the JCR one can go until $t = 10$. So the following formula can be used, when we have practical data (hence for discrete $t = 1,2,...$):

$$\frac{\ln c}{\ln p} = \frac{\ln\left(1 - \frac{\#\text{citations to } 1,2,...,t \text{ years back}}{C}\right)}{\ln\left(1 - \frac{\#\text{publications } 1,2,...,t \text{ years back}}{P}\right)} \tag{89}$$

As said above, the disadvantage of the method is that we need to know C and (especially) P.

IV.1.3 Quick and Dirty method

By (47) and (48) we have

$$\frac{\ln\left(\frac{c(t)}{c_0}\right)}{\ln\left(\frac{p(t)}{p_0}\right)} = \frac{\ln c}{\ln p} \tag{90}$$

quite simply, for every $t > 0$. Note that we do not need C or P here. The method is perfect but dependent on a one year score $c(t)$ and $p(t)$. This yields values of $\frac{\ln c}{\ln p}$ which are heavily dependent on the used year, which might result (for practical data) in heavily fluctuating results.

IV.1.4 Method using IF(2), IF(4) and IF(8)

Using (49) and (50) we have that the (non-relative) fixed year impact factors $IF(2)$, $IF(4)$ and $IF(8)$ are given by

$$IF(2) = \frac{C(1 - c^2)}{P(1 - p^2)} \tag{91}$$

$$IF(4) = \frac{C(1 - c^4)}{P(1 - p^4)} \tag{92}$$

$$IF(8) = \frac{C(1 - c^8)}{P(1 - p^8)} \tag{93}$$

Hence

$$\gamma =: \frac{IF(4)}{IF(2)} = \frac{1 + c^2}{1 + p^2} \tag{94}$$

$$\delta =: \frac{IF(8)}{IF(4)} = \frac{1+ c^4}{1+ p^4} \tag{95}$$

Formulae (94) and (95) constitute a system of equations in p and c that can be solved as follows. From (94) we have

$$c^2 = \gamma + \gamma p^2 - 1 \tag{96}$$

This can be put in (95)

$$\frac{1+ c^4}{1+ p^4} = \delta = \frac{1+ (\gamma + \gamma p^2 - 1)^2}{1+ p^4}$$

yielding the equation in the fourth degree:

$$p^4(\delta - \gamma^2) + p^2(2\gamma - 2\gamma^2) - 2 + 2\gamma + \delta - \gamma^2 = 0 \tag{97}$$

This equation can be solved numerically for p. Then (94) (or (95)) yields c: see formula (96), taking the positive square root since $c > 0$.

This method does not need P or C and uses only $IF(2)$, $IF(4)$ and $IF(8)$. $IF(2)$ can be read in the JCR and $IF(4)$ and $IF(8)$ can be determined using the JCR.

We will now test these 4 methods on a theoretical set of data and on a practical one.

IV.2 Examples of the calculation of $\frac{\ln c}{\ln p}$

IV.2.1 Theoretical example

We will calculate theoretical citation and publication data based on the formulae

$$c(t) = 100(0.6)^t \tag{98}$$

$$p(t) = 50(0.8)^t \tag{99}$$

Note that $c = 0.6 < p = 0.8$ as requested by the model. The calculated data are rounded off to the nearest entire number and we stop when we reach the value 0 for both functions $c(t)$ and $p(t)$ - see Table 1.

Table 1. Theoretical data based on (98) and (99).

t	p(t)	c(t)
0	50	100
1	40	60
2	32	36
3	26	22
4	20	13
5	16	8
6	13	5
7	10	3
8	8	2
9	7	1
10	5	1
11	4	0
12	3	0
13	3	0
14	2	0
15	2	0
16	1	0
17	1	0
18	1	0
19	1	0
20	1	0
21	0	0

We have $P = 246,\ C = 251$. We will now calculate $\frac{\ln c}{\ln p}$ according to the 4 methods presented in the previous section.

(i) Brookes method

Since $\frac{P}{2} = 123$ we take $t_0^* = 3$ in (87) since then $k^* = 124$, close to $\frac{P}{2}$. We find

$$p = \left(\frac{124}{246}\right)^{\frac{1}{3}} = 0.796$$

(close to the given 0.8). Since $\frac{C}{2} = 125.5$ we take $t_0 = 2$ in (84) since then $k = 91$, the closest we can get to $\frac{C}{2}$ (we want to combine at least 2 years, for reasons of stability). Now

$$c = \left(\frac{91}{251}\right)^{\frac{1}{2}} = 0.602$$

(close to the given 0.6). We find $\frac{\ln c}{\ln p} = 2.224$, close to the theoretical $\frac{\ln(0.6)}{\ln(0.8)} = 2.289$.

(ii) Second method

Formula (89) yields

$$\frac{\ln c}{\ln p} = \frac{\ln\left(1 - \frac{160}{251}\right)}{\ln\left(1 - \frac{90}{246}\right)} = 2.228$$

, again close to the theoretical 2.289.

(iii) Quick and Dirty method

This method works perfectly because this method is the way the data of Table 1 are calculated. E.g. for $t = 2$

$$\frac{\ln c}{\ln p} = \frac{\ln\left(\frac{36}{100}\right)}{\ln\left(\frac{32}{50}\right)} = 2.289 .$$

(iv) Method using IF(2), IF(4) and IF(8)

Since

$$IF(2) = \frac{100 + 60}{50 + 40} = \frac{160}{90}$$

$$IF(4) = \frac{100 + 60 + 36 + 22}{50 + 40 + 32 + 26} = \frac{218}{148}$$

$$IF(8) = \frac{100 + 60 + \ldots + 5 + 3}{50 + 40 + \ldots + 13 + 10} = \frac{247}{207}$$

we find that

$$\gamma = \frac{IF(4)}{IF(2)} = 0.8285473$$

$$\delta = \frac{IF(8)}{IF(4)} = 0.8100873$$

This gives the following equation of degree 4 (based on (97))

$$0.1235967p^4 + 0.2841133p^2 - 0.2193087 = 0$$

which has the solution $p = 0.781$ (close to the theoretical 0.8). Then (96) yields

$$c = \sqrt{\gamma(1 + p^2) - 1} = 0.578,$$

close to the theoretical 0.6. We find

$$\frac{\ln c}{\ln p} = 2.219,$$

close to the theoretical 2.289.

IV.2.2 Practical example

The next practical example will show that, due to the fact that there are deviations of (47) and/or (48) (the latter certainly for smaller t), the methods (iii) and (iv) (quick and dirty method respectively the method using $IF(2)$, $IF(4)$ and $IF(8)$) are unstable, since they use data based on one (or only a few) data point(s). On the other hand, methods (i) and (ii) (if we take t large enough) work well and give similar values for $\frac{\ln c}{\ln p}$. This also shows that the "old Brookes algorithm" is very stable, both in the determination of c and (new application) of p.

We have taken the journal "Journal of near Infrared Spectroscopy" for which 1993 is the year that volume 1 of this journal was published. We take the reference year 2003 as our $t = 0$. Hence $t = 1$ is the year 2002 and so on until $t = 10$ for the year 1993. The publication data $(p(t))$ are determined by simply counting the number of articles in every volume while the (synchronous) citation data $(c(t))$ are determined using the JCR of 2003. The data are as in Table 2.

Table 2. Publication and citation data for "Journal of near Infrared Spectroscopy" for the citing year 2003 $(t = 0)$.

t	0	1	2	3	4	5	6	7	8	9	10
p(t)	42	31	27	27	26	43	21	21	23	22	21
c(t)	6	28	51	55	50	67	11	21	35	18	17

Note that the "Rest" column in the JCR Cited Journal List (the one we use here) exactly refers to the year 1993 since this represents volume 1 of this joural.

We have $P = 304$, $C = 359$. For Brookes' method we have $\frac{P}{2} = 152$ hence we use (for the determination of p): $t_0^* = 5$, $k^* = 304 - 153 = 151$

$$p = \left(\frac{151}{304}\right)^{\frac{1}{5}} = 0.869$$

using (87) and (for the determination of c): since $\frac{C}{2} = 179.5$ we take $t_0 = 5$, $k = 359 - 190 = 169$ and, using (84)

$$c = \left(\frac{169}{359}\right)^{\frac{1}{5}} = 0.860 .$$

Note that $c < p$ and we have

$$\frac{\ln c}{\ln p} = 1.074 .$$

The second method yields, based on Table 2 (and taking t not too small, e.g. $t = 5$) and (89)

$$\frac{\ln c}{\ln p} = \frac{\ln\left(1 - \frac{190}{359}\right)}{\ln\left(1 - \frac{153}{304}\right)} = 1.077$$

close to the value obtained with the first method.

Due to the irregular set of data in Table 2 the quick and dirty method does not yield a good estimate for $\frac{\ln c}{\ln p}$. It can also be seen that the values of $IF(2)$, $IF(4)$ and $IF(8)$ do not even yield a solution for equation (97).

Not with standing the irregularity of the data in Table 2 we can conclude that the models developed in this paper can be applied as long as we use stable methods (as (i) and (ii)) to determine $\frac{\ln c}{\ln p}$.

V. Conclusions

In this paper we introduced impact Lorenz curves yielding all fractional (according to citations or to publications) relative impact factors which have the advantage not to be dependent on field or journal size and also one is not limited to a fixed time period for the calculation of the impact factor. These impact Lorenz curves are examples of weighted Lorenz curves of a continuous variable which are introduced in this paper.

We then study these curves and impact factors in the case of exponential aging (for citations) and growth (for publications) curves. The impact Lorenz curve has the following functional form (c = aging rate $< p$ where $g = \frac{1}{p}$ is the growth rate)

$$L(y) = 1 - (1 - y)^{\frac{\ln c}{\ln p}} \qquad (100)$$

implying that for any two such Lorenz curves L_1 and L_2 we have $L_1 > L_2$, $L_1 < L_2$ or $L_1 = L_2$, i.e. apart from $(0,0)$ and $(1,1)$, these curves do not intersect if $L_1 \neq L_2$. This also implies that if, say $L_1 < L_2$, all fractional relative impact factors in the first situation are smaller than all fractional relative impact factors in the second situation and we show that this is not true for relative impact factors dependent on time t.

Since, as is clear from (100), the parameter $\frac{\ln c}{\ln p}$ is crucial in this theory, we present 4 methods to calculate this parameter, given practical data. Examples are presented.

Appendix

Proposition (Egghe (1988)):

Let $IF(t)$ be as in (6) :

$$IF(t) = \frac{\int_0^t c(t')dt'}{\int_0^t p(t')dt'} \tag{101}$$

Then $IF'(t)$ has the same sign as

$$\frac{c(t)}{p(t)} - IF(t) \tag{102}$$

Proof:

$$IF'(t) = \frac{\left(\int_0^t p(t')dt'\right)c(t) - p(t)\int_0^t c(t')dt'}{\left(\int_0^t p(t')dt'\right)^2}$$

which has the same sign as

$$c(t)\int_0^t p(t')dt' - p(t)\int_0^t c(t')dt'$$

hence the same sign as

$$\frac{c(t)}{p(t)} - IF(t)$$

by (101).

Corollary:

$IF(t)$ strictly decreases if

$$\frac{c(t)}{p(t)} < IF(t) \tag{103}$$

This is e.g. satisfied if $\frac{c(t)}{p(t)}$ is a strictly decreasing function of t.

Proof:

That (103) implies that $IF(t)$ strictly decreases follows directly from the above proposition. If $\frac{c(t)}{p(t)}$ decreases strictly in t then it is clear that

$$\frac{c(t)}{p(t)} < \frac{\int_0^t c(t')dt'}{\int_0^t p(t')dt'} = IF(t). \qquad \sim$$

References

Braun, T., Glänzel, W. and Schubert, A. (1985). Scientometric Indicators. A 32-Country comparative Evaluation of publishing Performance and citation Impact. World Scientific, Singapore.

Braun, T., Glänzel, W. and Schubert, A. (1989). An alternative quantitative approach to the assessment of national performance in basic research. In: The Evaluation of scientific Research. Proceedings of a Ciba Foundation Conference, 32-49, Wiley, Chichester.

Brookes, B.C. (1970). The growth, utility and obsolescence of scientific periodical literature. Journal of Documentation, 26, 283-294.

Brookes, B.C. (1971). Optimum P% library of scientific periodicals. Nature 232, 458-461.

Dierick, J. and Rousseau, R. (1988). De impactfactor voor tijdschriften: een parameter bij het bepalen van een – al dan niet defensief – collectiebeleid? In: Het oude en het nieuwe Boek. De oude en de nieuwe Bibliotheek. Liber Amicorum H.D.L. Vervliet (J. Van Borm en L. Simons, eds.), 593-601, Pelckmans, Kapellen (Belgium).

Egghe, L. (1988). Mathematical relations between impact factors and average number of citations. Information Processing and Management 24(5), 567-576.

Egghe, L. (2002). Construction of concentration measures for general Lorenz curves using Riemann-Stieltjes integrals. Mathematical and Computer Modelling 35, 1149-1163.

Egghe, L. (2004). Zipfian and Lotkaian continuous concentration theory. Journal of the American Society for Information Science and Technology, to appear.

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

Egghe, L. and Rousseau, R. (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, the Netherlands.

Egghe, L. and Rousseau, R. (1996a). Average and global impact of a set of journals. Scientometrics 36(1), 97-107.

Egghe, L. and Rousseau, R. (1996b). Averaging and globalising quotients of informetric and scientometric data. Journal of Information Science 22(3), 165-170.

Egghe, L. and Rousseau, R. (2000). Aging, obsolescence, impact, growth and utilization: definitions and relations. Journal of the American Society for Information Science 51(11), 1004-1017.

Egghe, L. and Rousseau, R. (2001). Symmetric and asymmetric theory of relative concentration and applications. Scientometrics 52(2), 261-290.

Egghe, L. and Rousseau, R. (2003). A general framework for relative impact indicators. The Canadian Journal of Information and Library Science/La Revue Canadienne des Sciences de l'Information et de Bibliothéconomie 27(1), 29-48.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science 178, 471-479, 1972. Reprinted in: Essays of an Information Scientist 1, ISI Press, Philadelphia (USA), 527-544.

Garfield, E. (1979a). Citation Indexing: its Theory and Application in Science, Technology and Humanities. Wiley, New York (USA). Also: second Edition, ISI Press, Philadelphia (USA), 1983.

Garfield, E. (1979b). Is citation analysis a legitimate evaluation tool? Scientometrics 1, 359-375.

Garfield, E. (1983). How to use citation analysis for faculty evaluations, and when is it relevant? Part 2. Current Contents, November 7. Reprinted in: Essays of an Information Scientist 6, ISI Press, Philadelphia (USA), 363-372, 1984.

Garfield, E. and Sher, I.H. (1963). Citation indexes in sociological and historical research. American Documentation, 14, 289-291.

Ingwersen, P., Larsen, B., Rousseau, R. and Russell, J. (2001). The publication-citation matrix and its derived quantities. Chinese Science Bulletin 46(6), 524-528.

Pudovkin, A.I. and Garfield, E. (2004). Rank-normalized impact factor: a way to compare journal performance across subject categories. Proceedings ASIST 2004, to appear.

Rousseau, R. (1988). Citation distribution of pure mathematics journals. In: Informetrics 87/88. Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval (L. Egghe and R. Rousseau, eds.), Diepenbeek (Belgium), 249-262, Elsevier, Amsterdam, the Netherlands.

Rousseau, R. (2004). Median and percentile impact factors: a set of new indicators. Preprint.

Rousseau, R., Jin, B., Yang, N. and Liu, X. (2001). Observations concerning the two- and three-year synchronous impact factor, based on the Chinese Science Citation Database. Journal of Documentation 57(3), 349-357.

Rousseau, R. and Smeyers, M. (2000). Output-financing at LUC. Scientometrics, 47(2), 379-387.

Schubert, A., Glänzel, W. and Braun, T. (1983). Relative citation rate: a new indicator for measuring the impact of publications. In: Proceedings of the First National Conference with International Participation on Scientometrics and Linguistics of Scientific Text (D. Tomov and L. Dimitrova, eds.). Varna (Bulgaria), 80-81, Varna.

Schubert, A., Glänzel, W. and Braun, T. (1986). Relative indicators of publication output and citation impact of European physics research. Czechoslovak Journal of Physics B36, 126-129.

Sombatsompop, N., Markpin, T. and Premkamolnetr, N. (2004). A modified method for calculating the impact factors of journals in ISI Journal Citation Reports: polymer science category in 1997-2001. Scientometrics 60, 217-235.

Stinson, E.R. (1981). Diachronous versus synchronous Study of Obsolescence. Ph. D. Thesis. University of Illinois.

Stinson, E.R. and Lancaster, F.W. (1987). Synchronous versus diachronous methods in the measurement of obsolescence by citation studies. Journal of Information Science 13, 65-74.