THE DUALITY OF INFORMETRIC SYSTEMS WITH APPLICATIONS TO THE EMPIRICAL LAWS

Leo Egghe

Dissertation submitted for the degree of
Doctor of Philosophy

Department of Information Science

The City University
St. John Street
London EC1V4PB

April 1989

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I want to thank profs. B.C. Brookes and R. Rousseau for their continuous support and interest in my work. They were always prepared for long discussions (by phone, in the office or even at home) concerning informetrics.

I also thank prof. S.E. Robertson for his increasing interest in the topic of the thesis and for his agreement to read and comment on this work. Furthermore his help in the administrative organisation is very much appreciated.

I thank the universities of Limburg (LUC), Antwerp (UIA) and Amsterdam for their interest in informetrics and for their financial support.

Financial support was also appreciated from the NFWO (Belgian National Science Foundation) and the British Council. I thank Miss L. Bull for her ever continuing efforts to support these projects.

I thank, finally, my family for giving me the time to prepare this thesis, being a project that consumed many evenings.

DECLARATION

I GRANT POWERS OF DISCRETION TO THE UNIVERSITY LIBRARIAN TO ALLOW THIS THESIS TO BE COPIED IN WHOLE OR IN PART WITHOUT FURTHER REFERENCE TO ME. THIS PERMISSION COVERS ONLY SINGLE COPIES MADE FOR STUDY PURPOSES, SUBJECT TO NORMAL CONDITIONS OF ACKNOWLEDGEMENT.

ABSTRACT

This thesis is mainly concerned with the dual fundaments of informetrics. After an intuitive introductory chapter, we study, in a broad informetric context, general information production processes (IPP) (both discrete and continuous ones) and duality principles (between the sources and the items) in them in an exact and formalistic way. Classical informetrics evolves from this study as an example of a purely dual situation. The general duality technique is also able to recover new informetric laws (including a modelling of the Groos droop) that are easy to fit in practice.

We present also parameter relations and classifications of some informetric laws, using only exact mathematical techniques. The most interesting features here are : the study of the group-free version of Bradford's law, the generalised Leimkuhler law and the place of Zipf's (or Pareto's) law in this context. Also the derivation of some formulae for some parameters, appearing in the group-dependent version of Bradford's law, is non-trivial and very useful in the sequel : they are e.g. basic tools in the fitting of the "nuclear" part of the Leimkuhler graph, even if a Groos droop is apparent, a result that has nice applications. Also the generalised Lotka and Leimkuhler functions are fitted.

The thesis is rounded off by a summary of the results.

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviations or symbols	Meaning	Section
IPP	Information Production Process	I.4, II.1.1, II.1.2
f	Lotka's function or generalisation	I.5.1. II.4.1
α	Lotka's exponent	I.5.1
T	total number of sources	I.5.1
g	Zipf, Mandelbrot function	I.5.3 I.5.4
R	Leimkuhler's function	I.5.5
r_o, y_o, k, p	parameters in Bradford's law (group-dependent)	I.5.6
$\mathbb{N}$	positive entire numbers	I.5.6
A	total number of items	I.5.6
(S,I,V)	continuous IPP	II.1.1
S	the source set	II.1.1
I	the item set	II.1.1
V	function from S into I	II.1.1
[x,y]	the interval starting in x and ending in y (x and y included)	II.1.1
(S,I,i)	discrete IPP	II.1.2
i	function from S into the subsets of I	II.1.2
$\subset$	subset of	II.1.2
$\in$	element of	II.1.2
#	number of elements in	II.1.2
$\prec$	the order on S resp. on I	II.1.2
V^{-1}, ρ^{-1}	the inverse function of V resp. ρ	II.2.1 II.4.1
U	dual function of V	II.2.1

σ	$\sigma(i) = U'(i)$	II.2.1
U', V'	derivatives	II.2.1
ρ	$\rho(r) = V'(r)$	II.2.1
$\square$	End of a proof	II.2.1.1
$[x]$	largest entire number smaller than or equal to x	II.3.2
K	Bradford factor (group-free)	II.3.3.2
$\int_a^b$	The integral with lower bound a and upper bound b	II.4.1
$\int_a^\infty$	The improper integral with lower bound a	II.4.2.1
$[1,\infty[$	The real numbers larger than or equal to 1	II.4.2.1
log	logarithm (Naperian)	II.4.3.1
$\min_i$	minimum over all i	II.4.3.2
$\lim_{\alpha \to 2} \sigma(i)$	the limit of $\sigma(i)$ for α going to 2	II.4.3.2
$\bar{X}$	Closure of the set X	III.1.2
$\bigcup_{\ell \in \mathbb{N}} A_\ell$	The union of the sets $A_1, A_2, A_3, \ldots$	III.1.2
$\approx$	approximation	III.2
y_m	number of items in the most productive source	III.2
$\sum_{j=1}^{y_m} \frac{C}{j}$	sum of $\frac{C}{j}$ for $j = 1,2,\ldots,y_m$	III.3.1
γ	number of Euler $\gamma \approx 0.5772\ldots$	III.3.1
$\sum_{j=1}^{\infty} \frac{1}{j^2}$	the series $\lim_{k\to\infty} \sum_{j=1}^{k} \frac{1}{j^2} = \frac{\pi^2}{6}$	III.3.1
m(i)	the number of items in the most productive source in the i^{th} group (right or left)	III.3.2 III.3.4
μ	average production	III.3.3
f_p	$f_p(k) = \frac{k}{\mu}$	III.3.3

$\lim_{k \to 0^{>}}$	limit for k going to zero, $k > 0$	III.3.3
$\lim_{k \to 1^{<}}$	limit for k going to 1, $k < 1$	III.3.3
$\alpha(i)$	fraction of the items, belonging to sources with production m(i) that belong to the i^{th} group (right to left)	III.3.4
Δ	difference	III.3.4
r_1, y_o, k_1, p	parameters in Bradford's law (graphical, group-dependent)	III.4.1
Σ	$\Sigma(i) = \int_0^i \sigma(i') \, di'$	III.4.1
K_1	Bradford factor (graphical, group-free)	III.4.1
R_1	Brookes' function or the function of Weber-Fechner	III.4.1
$\lim_{p \to \infty} r_o$	limit of r_o for p going to infinity	IV.1.1
ρ_o	cut-off rank	IV.3
r'	final cut-off rank	IV.3.1
$\hat{T}$	$\hat{T} = r'$	IV.3.1
$\hat{A}$	the number of items in the truncated IPP	IV.3.1
$\hat{\mu}$	average production of the "missed" sources	IV.4.1
ζ	zèta function	IV.5

CHAPTER I : INTRODUCTION

I.1. The background of the present study

Studies of many aspects of the social sciences need, first, relevant data and then, secondly, an analysis of these data, applying the most appropriate quantitative, analytical techniques available for the purpose. In this sense, such studies run parallel with analogous studies in the physical sciences. But, compared with the analytical techniques of the physical sciences, the available techniques applicable to the social sciences are weak, primitive and incoherent.

That there are indeed many differences between the analytical techniques available for the physical and social sciences has been pointed out, for instance, in the work of S.D. Haitun (1982 a, b, c and 1983). He noted the dichotomy between the type of statistics that is needed : the well-known Gaussian statistics for the physical sciences and the far less well-known "Zipfian" statistics, involving distribution functions without finite moments (or with, at most, one finite moment : the mean), which is the type of statistics that seems to be appropriate for the social sciences.

Though some valuable work of applying Gaussian techniques to Zipfian distributions has been done, (Sichel, 1986) and (Burrell, 1988), these techniques usually become very difficult and do not reach far enough. They furthermore deal only with frequency distributions; in the social sciences one needs also techniques to analyse ranks, which require, obviously, more detailed aspects of statistics.

In conclusion, one can say that the close relation between mathematics/statistics found in the physical sciences does not exist in the social sciences. Moreover, applications of mathematics/statistics to the social sciences, are not comprehensively systematic.

So, as a mathematician, I see a field of enquiry, inviting an attempt to provide a mathematical framework which is at present lacking.

I.2. The general approach

An approach to the application of mathematics to the empirical sciences has been propounded by Stefan Körner (1969). He suggests that three steps are needed :

1. Inexact empirical concepts have to be replaced by exact mathematical concepts.
2. Exact conclusions are then deduced from these mathematical concepts.
3. The exact mathematical conclusions are then replaced by empirical concepts.

As a model of this type has never been developed for the social sciences, the present work is the first of its kind. However, we must clarify the terminology.

I.3. Sources of the empirical laws : terminology

Throughout the literature one finds the terms : bibliometrics, scientometrics, informetrics, econometrics, sociometrics, quantitative linguistics, and so on. It is not clear what the exact definitions of the above subjects must be (for a review on this problem, see e.g. (Egghe, 1988b)). Clearly, there is an overlap between fields and certainly between bibliometrics and scientometrics. Scientometrics has been used mostly in Eastern Europe and the term bibliometrics may be considered as its Western equivalent. The term informetrics is the most recent and to my idea, is the most general (see (Brookes, 1984))

I therefore adopt the term informetrics as the generic term for all the above (and possibly other) disciplines. The other terms will be used whenever they are linked with a

historically recognisable regularity such as, for example, Pareto's law in econometrics. The mathematical model to be developed in this work, is basically independent of the different "-metrics". Applications to specific "-metrics" will however be given, whenever possible.

It is then within this general informetric framework that we will consider Information Production Processes :

I.4. Information Production Processes (IPP). Sources and items

In this work we will use the notions "information production process" (IPP) in which there are two kinds of entities : the sources and the items produced by these sources. Exact definitions follow in the next chapter. Let us give some examples.

1. In econometrics we can give the example of a group of workers or employees and study their productivity (Theil, 1967). Productivity can be measured in several ways : as quantity (numbers of produced items), as quality, or in terms of profits (number of pounds earned in a certain time period). In this example, the choice of the term "production" is quite clear. In a more general way the next examples can also be considered as information production processes.

2. In demography one considers cities and villages in conjunction with their populations.

3. In linguistics one considers words (as entities or "types" - as is often used in linguistics (Herdan, 1960)) and their occurrences (or "tokens" in linguistic terms) in a given text (book, article, ...) (Zipf, 1949).

4. In bibliometrics one can study books in a library and the number of times they are borrowed, say, in a year (Burrell and Cane, 1982).

5. One can also study a group of researchers and the number of publications they produce, say in a ten-year period (Lotka, 1926).

6. Still in bibliometrics, one can consider a bibliography (on a specified topic), in which the contributing journals produce papers (Bradford, 1934).

7. Papers themselves can be considered as sources rather than as items in the previous case. Thus, a set of papers can be considered together with the citations they receive within a fixed time period (Garfield, 1983). In this connection one has the "cited" relationship between papers. An interesting point to make here is the well-known fact that another example can be constructed when "cited" is changed into "citing". An equivalent way of expressing this is : retain the term "cited" but interchange the two sets of papers. This is a first indication of what this work is mainly concerned with.

In all of the above examples one has an IPP consisting of sources that produce items. Hence these terms can be used as generic terms in the theory to be developed.

I.5. Empirical laws and corresponding mathematical functions

The regularity that is the simplest to be introduced is the law of Lotka.

I.5.1. The law of Lotka

In 1926, A.J. Lotka (1926) studied a 10-year Cumulative Index of authors listed in Chemical Abstracts (1907-1916) and Auerbach's Geschichtstafeln der Physik (1910) was also examined.

He found the following regularity : if f(j) denotes the number of authors with j publications, then

$$f(j) = \frac{C}{j^{\alpha}} \qquad (I.1)$$

where $\alpha \approx 2$, but not necessarily $\alpha = 2$.

If $\alpha = 2$, then

$$C = \frac{6}{\pi^2} T \approx 0.6079\ T\ , \qquad (I.2)$$

where T denotes the total number of authors. Function (I.1) will be called the Lotka function, as it expresses the law of Lotka.

The other empirical laws all relate to rankings of the IPP.

I.5.2. A ranking (intuitively)

In the sequel, we suppose the following ranking on the sources of an IPP : the most productive source receives rank 1, then the second rank is for the second most productive source, and so on : the last rank (T) is for the source with the least production; ties are broken arbitrarily (see also the next chapter for a more accurate description).

I.5.3. The laws of Zipf and Mandelbrot

Formulated originally in linguistics, Zipf's law can be expressed thus (Zipf, 1949) : Order the words in a text in decreasing order of occurrence in this text. Then the product of the rank r of a word and the number of times j it is used in the text is a constant for that text :

$$r.j = E \qquad (I.3)$$

or, putting $j = g(r)$:

$$g(r) = \frac{E}{r}\ . \qquad (I.4)$$

More generally one can formulate the following general Zipf function :

$$g(r) = \frac{F}{r^{\beta}} \qquad (I.5)$$

where F and β are constants.

From the same context, but with an expression different from (I.5) is the law of Mandelbrot (Mandelbrot, 1954 and 1977).

$$g(r) = \frac{G}{(1 + Hr)^{\beta'}} \quad , \qquad (I.6)$$

where G, H and β' are constants.

I.5.4. The law of Pareto

This law is formulated in econometrics (Theil, 1967). It states that the number h(j) of workers with an income larger than or equal to j is

$$h(j) = \frac{L}{j^{\gamma}} \quad , \qquad (I.7)$$

where L and γ are constants. As is obvious in combining I.5.3 and the above (with an obvious unification of the terminology), we see

$$r = h(j) = \frac{L}{j^{\gamma}}$$

or

$$j = \frac{L^{1/\gamma}}{r^{1/\gamma}}$$

and hence

$$g(r) = \frac{L^{1/\gamma}}{r^{1/\gamma}} \qquad (I.8)$$

In conclusion, the Pareto function and the Zipf function are identical though their respective laws apply to different contexts. This kind of identity is another issue to be considered in the general context of informetrics.

I.5.5. The law of Leimkuhler

Consider a bibliography of papers on a specific topic, published in journals. Using the order of I.5.2 and denoting by F(x) the cumulative fraction of papers in the journals of rank 1,2,...,r, where $x = \frac{r}{T}$, the cumulative fraction of the journals, we have :

$$F(x) = \frac{\log (1 + \delta x)}{\log (1 + \delta)} , \qquad (I.9)$$

where δ is a constant (Leimkuhler, 1967). In the sequel we will work with the function R(r) = F(x).A (A = the total number of papers and r = x.T. We hence have the following Leimkuhler function (equivalent to formula (I.9)) : Let R(r) denote the cumulative number of items in the journals of rank 1,2,...,r. Then

$$R(r) = a \log (1 + br) , \qquad (I.10)$$

where a and b are constants. A similar relation was already observed by Bradford (1934).

I.5.6. The law of Bradford

The most intriguing of all the empirical laws is that of Bradford (Bradford, 1934) based on observations of bibliographies on Applied Geophysics, 1928-1931 (incl.) and Lubrication, 1931-june 1933.

We present it here in its original definition which is, as far as I can see, clear enough. We must remark however that some informetrists have been confused by its formulation, giving rise to what is now known as the "verbal" and the "graphical" formulation of Bradford's law (which are not exactly equivalent). For the difference between these laws, see the third chapter. We present here the original "verbal" version. It states :
Order the journals in decreasing order of the number of papers (in this bibliography) they contain. If the journals are subdivided into p groups (according to the above order) such that each group of journals contains the same number

y_o of papers in this bibliography, then there exist r_o and $k > 1$ such that the first group has r_o journals, the second has $r_o k$ journals, the third has $r_o k^2$ journals and so on, until the last (p^{th}) group, contains $r_o k^{p-1}$ journals.

Otherwise stated, if p is a given positive integer (denoted $p \in \mathbb{N}$), then there exists $r_o \in \mathbb{N}$ and $k > 1$ (a real number) such that the first (most productive) r_o journals produce $y_o = \frac{A}{p}$ (A = total number of papers) papers, the next $r_o k$ journals produce again y_o papers, the next $r_o k^2$ journals also produce y_o papers, and so on, until the last (least productive) $r_o k^{p-1}$ journals producing again y_o papers.

One aspect of this formulation is the kind of symmetry between the journals and the papers. If we represent the bibliography and the order on it by a straight line (or better, an axis with coordinates the ranks of the journals),

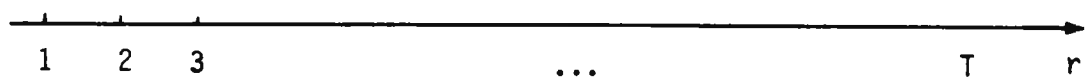


Fig.I.1 : The rank-axis

then we feel intuitively that, when going from left to right, the "visibility" of the journals is changed into the "visibility" of the papers. We continue this heuristic approach in the next paragraph.

I.6. Heuristic approach to duality

In this paragraph we will analyse the information production processes (IPP), introduced in I.4, together with some of their historical regularities, introduced in I.5. The analysis here will remain heuristic (reflecting

the way in which the formal concepts developed in the author's mind!) and it is our hope that, in this way, the formal theory, to be developed in the next chapters, will become clearer and convincing.

A. From I.4 it is clear that, in IPP's, we deal with two sets : the set S of sources and the set I of items. Indeed, unifying I.4 we can make the following Table I.1 :

Subject	Sources	Items
econometrics	workers	their salary
demography	cities	their inhabitants
linguistics	words	their occurrence in a text
bibliometrics	books	their borrowings
bibliometrics and research policy	researchers	their publications
bibliometrics	journals	papers in them (on a fixed topic)
bibliometrics and research policy	papers	the citations they receive
bibliometrics and research policy	papers	the citations they give (i.e. the references)

Table I.1 : Examples of information production processes

I would call a 1-dimensional informetric study, any study dealing with the sources or the items separately (i.e. not linked with each other).

Examples :

1. The numbers of books in a library.
2. The numbers of circulations in a library.
3. The total number of publications in geography (say in a year).
4. The total number of researchers in mathematics in Belgium.

Such data can be very interesting, especially in the connection with evolution in time. Many publications have resulted from such studies.

However, as is clear from the above table and also from I.5, one can also develop what I would like to call 2-dimensional study of informetrics. This is, a study of the quantitative properties of the sources vs. the items and/or vice-versa.

Examples :

1. The average number of circulations of books in a library.
2. The historical laws, described in I.5.
3. The evolution of the average number of citations per publication in physics.

So, central to this 2-dimensional setting is the source-item-relationship. Again, there are many publications dealing with this kind of problem. Our main focus in this work is the theoretical basis of such a situation.

The law of Lotka, as introduced in I.5.1, is certainly an example of a model in 2-dimensional informetrics. But, as compared with the other laws described in I.5, it is a more primitive law : indeed, for this law we do not need to introduce an ordering on the set of sources, as is indeed needed in all the other laws I.5.3 - I.5.5 (even in Pareto's law, the order is implicitely assumed). Expressed differently, from any of the settings I.5.3 - I.5.6 we can derive the form of the I.5.1-setting, but not conversely. These matters are well-known : we talk here about the difference between frequency distributions and rank-order distributions.

So, a "complete" 2-dimensional informetric study requires not only the set of sources and the set of items, but needs also a device function expressing what items are produced by what sources. We hence have the system

(S,I,i)

where S is the set of sources, I is the set of items and i is the device function. This formalism will be studied more systematically in the next chapter, where detailed definitions will be given.

B. Another aspect of informetric studies is the problem of explaining certain regularities (e.g. the laws in I.5). It is not easy to say what "explaining" means. Let us content ourselves here by considering as a contribution to an explanation, every logical argument leading to one or another form of regularity. In this sense, several explanations of the laws described in I.5 (and others) (hence of 2-dimensional informetrics) have been given : by Price (1976), Bookstein (1984), Mandelbrot (1977).

The best known of these explanations is the so-called "success-breeds-success" principle expressing simply that the more items a source has already published, then the greater is the probability that this source will produce another item. This principle leads to a form of Lotka law and is therefore valuable.

In this work, however, we start differently and, in a way, also from a more elementary point of view in our approach to 2-dimensional informetrics. All the known explanations start from an assumed property (principle, axiom, ...) between sources and items and there is nothing wrong with that : if we want to prove a certain law, we must assume some property! Later on in our approach we will do the same (f.i. in order to prove that Bradford's law belongs to a class of laws with special properties). But we start by pointing out that two different approaches to every problem in 2-dimensional informetrics are possible : the one looking at (sources, items) in this order and the other looking at (items, sources) in the reverse order. We call this the duality principle as applied to informetrics.

So, our first claim is not a property that can be proved or denied (such as the success-breeds-success principle) but a more elementary procedure that needs no proof : the duality aspect of 2-dimensional informetrics.

In other words, this duality aspect points only to a procedure : every time we look at a function or property relating (sources, items) in this order, we also consider the corresponding function or property, relating (items, sources) in this order. We might even use both functions at the same time. The fact that we have, in all these problems, the disposition of a pair of functions might lead to new results and in fact does so as will be shown in the next chapter.

After this, we can formulate a property from which we might try to derive a known law.

This approach is new, although the concept of duality - between types and tokens in linguistics - was already formulated in (Herdan, 1960), but no mathematical exploration of this principle followed. It makes no initial assumptions. Also, since new, applicable results will be obtained, this approach is productive.

This dual approach to informetrics can also be compared with the duality principle in geometry. In geometry one considers the duality between straight lines and points. Also in geometry, this principle is just a procedure and not something that has to be proved in itself. The branch of geometry devoted to these duality aspects is called projective geometry. Every time one obtains a theorem proving a property between points and lines (in this order), one can formulate the dual theorem by interchanging of the words lines and points. This is the duality procedure. After this action, one still needs to prove the dual theorem thus obtained, since duality in itself cannot and does not have to be proved.

I.7. Other approaches

In this work we restrict ourselves to 2-dimensional informetrics since, in this framework, we find many

interesting problems. No doubt, in the future, a 3-dimensional approach would be possible.

Examples :

1. Journals have papers and these papers are written by authors.
2. Journals have papers and papers receive (or have) citations.
3. Papers have references but do also receive citations.

These examples conform, schematically with the following diagrams and graphs :

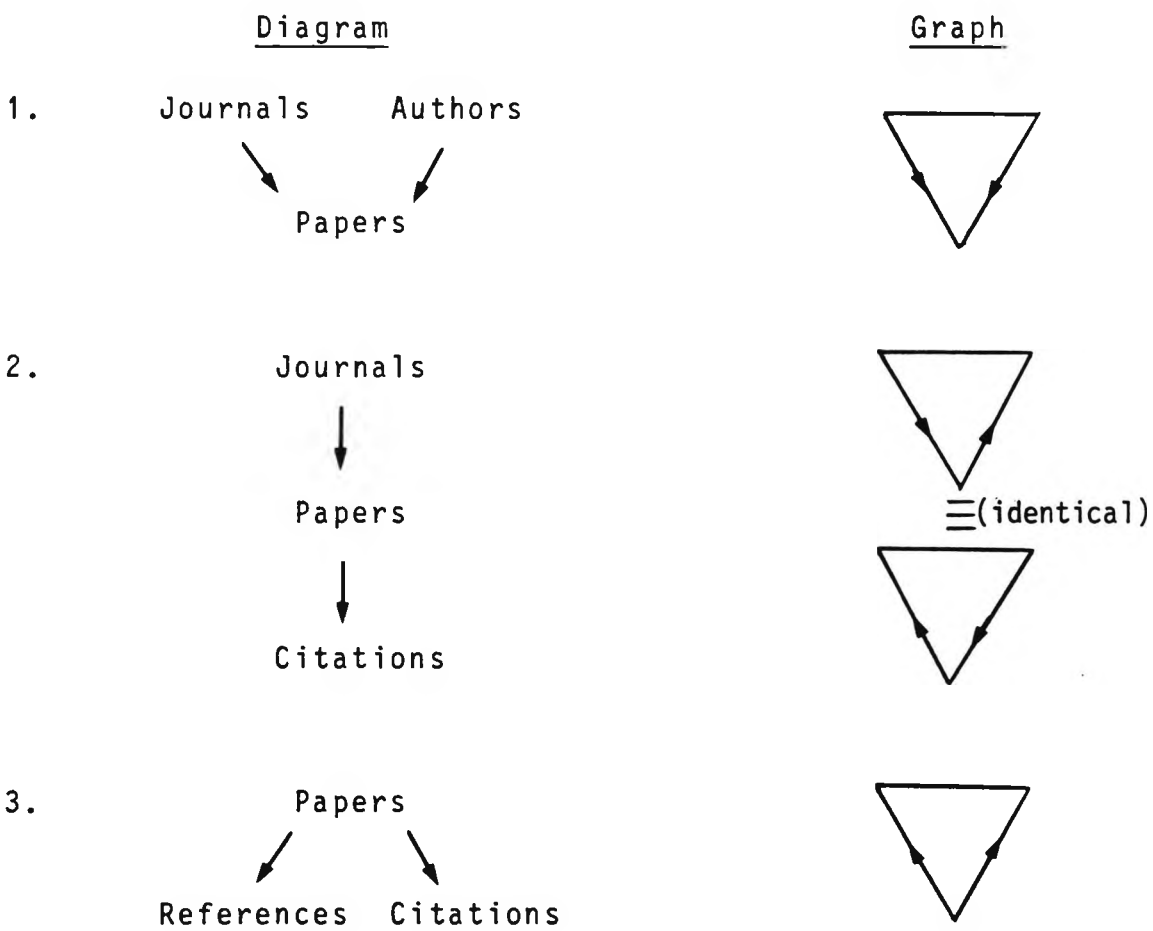


Fig.I.2 : Schemes of three dimensional informetrics.

One can even think of 4-dimensional (or even higher dimensional) informetrics. This is not an easy problem and can be the subject of several research projects.

Another fundamental viewpoint of informetrics, different from the above, is the "many-to-many" relationship, as was pointed out to me by S. Robertson. It is in contrast with the above techniques in that it studies (e.g. in the 2-dimensional version) many sources versus many items. We have here the study of the relationship between a set S_1 and a set I_1 such that S_1 is a subset of the source set S, I_1 is a subset of the item set I and where the "device" function now is

$$f : S_1 \rightarrow I_1 \qquad (\text{or } \leftarrow)$$

An example is offered by
I_1 = a set of index words
S_1 = the set of papers
f = the function "saying" that the sources in S_1 have the index words in I_1.

In our approach, S_1 is a singleton, but many singletons may be linked with the same item set I_1.

I.8. Overview of the results

In the next chapter we will formalise the above heuristic approach giving operational definitions of "Information Production Processes" (IPP's) and by defining duality in them. This will be done for continuous as well as for discrete processes.

It will be proved that Bradford's law is the only (known) informetric law that has a pertinent property : the IPP and its dual IPP have the same informetric calculus. In this connection, the Bradford function for continuous IPP's is defined in a group-free way, yielding also a group-independent Bradford factor.

Using the duality technique again we find the new laws of Bradford and Leimkuhler, equivalent with the general

laws of Lotka

$$f(j) = \frac{C}{j^{\alpha}} \qquad (I.1)$$

($\alpha > 1$).

Hence this chapter is mainly concerned with duality and the derivation of informetric laws from it. I would like to call this : first order informetrics.

The next (third) chapter deals then with second order informetrics : classification of informetric laws (showing also the special place of Zipf's (Pareto's) law amongst the informetric laws and hence of linguistics (econometrics) w.r.t. informetrics), and proving methods (again based on the duality formalism) for calculating parameters of the several informetric laws dealt with in the previous chapter. For example, an explicit formula for the Bradford factor is proved (using again a duality argument).

In the fourth chapter, these methods form the basis for calculating Leimkuhler's function, given a table of practical data. In connection with the often encountered deviation from Leimkuhler's curve (the so-called Groos droop, see (Groos, 1967)) we provide finer fitting methods for the first part of a Leimkuhler curve (before the Groos droop appears). This in turn has applications in the determination of core-collections and of the completion of bibliographies.

We provide also methods for calculating the general Lotka function (I.1), given a table of raw data and we show that our methods - though simpler than some of Nicholls and Tague - provide comparable fits. Furthermore, we fit the generalised Leimkuhler function (shown in the previous chapter) by the same methods and calculations as are needed in fitting Lotka's function.

The last chapter is devoted to a general summary of the most important results, that are discussed in this work.

Informetrics can be developed in several different ways. Other distribution functions can be considered and other methods of investigation are thinkable. This work has the aim of presenting one viewpoint of informetrics and tries to be consistent within this viewpoint. Apart from consistency, we also aim to be as accurate as possible in the mathematics used. Approximations - when necessary - are clearly defined and explained logically : they become part of the theory rather than being weak points within it. Concerning the mathematical results and formulae we make use of, the reader is referred to (Apostol, 1974), (De Lillo, 1982) or (Gradshtein and Ryzhik, 1965).

CHAPTER II : DUALITY IN INFORMATION PRODUCTION PROCESSES

II.1. Definition of Information Production Processes (IPP)

We distinguise between continuous and discrete IPP's. Discrete IPP's are exact models of practical situations such as for instance the examples in section I.4, but continuous IPP's are close enough models for large discrete IPP's and certainly contain (in the subset sense) all the discrete IPP's. Continuous IPP's also give more insight in both the dual theory of IPP's and also in Bradford's law (coming up). In general, continuous IPP's are mathematically easier to handle than discrete IPP's.

II.1.1. Continuous IPP's

A continuous IPP is a triple of the form

$$(S,I,V) \tag{II.1}$$

where $S = [0,T]$ (the closed interval starting in 0 and ending in T), $I = [0,A]$ and where V is a strictly increasing differentiable function

$$V : S \to I \tag{II.2}$$

such that $V(0) = 0$ and $V(T) = A$.
The elements of S are called sources; the elements of I are called items. In the sequel we will always consider $V(r)$ (for every $r \in S \setminus \{0\}$) to be the cumulative number of items in all the sources $s \in [T-r,T]$ (taking $[T-r,T]$ rather than $[0,r]$ for technical reasons, to become clear in section II.2.1). Hence, V is an integral of a certain density function, to be introduced in section II.2.1.

II.1.2. Discrete IPP's

A discrete IPP is a triple of the form

$$(S,I,i) \qquad (II.3)$$

where S and I are countable sets and where, for every $s \in S$, $i(s) \subset I$. The elements of S are called sources; those of I items.

Example :
Every example in section I.4 is an example of a discrete IPP, where - e.g. in the case of a bibliography -, S is the set of journals, I is the set of papers and i is the "device" function indicating, for every $s \in S$, the articles $i(s) \subset T$ that are published in s. A definition of the form (II.1), (II.2) is also possible here but for practical reasons, the above is preferred.

S is linearly ordered as follows : For every $s,s' \in S$,

$$s \prec s' \text{ if and only if } \#i(s) > \#i(s') \text{ or } s = s' \qquad (II.4)$$

(# denotes "the number of elements in "); i.e. we order the sources in decreasing order of the number of corresponding items. In this way, $S,\prec$ is partially ordered. Ties are broken arbitrarily in order to obtain a linear order (i.e. for every $s,s' \in S$: $s \preccurlyeq s'$ or $s' \preccurlyeq s$).

If we order the sets $i(s) \subset I$ (for every $s \in S$) in an arbitrary linear way, then the order $\prec$ on S induces an order on I, which is also linear. This order on I we will also denote by $\prec$ since confusion with $\prec$ on S is not possible.

The proof of the assertion (*) on p 19 is as follows.

Lemma: Let (S, I, V) be an arbitrary continuous IPI and (I, S, U) its dual. Then, for every $r \in [0, T]$ and $i \in [0, A]$.

$$V(r) = i \iff U(A-i) = T - r \qquad (\dashv$$

Proof: $\Rightarrow$ If $V(r) = i$, then $i = V^{-1}(r)$. Applying (II.6) to $i' = A - i$ yields

$$U(A-i) = T - V^{-1}(A-(A-i))$$
$$= T - r$$

$\Leftarrow$ this proof is similar to the above. □

Corollary: The bidual of an IPI is the IPI itself: If (I, S, U) is the dual of (S, I, V) and if (S, I, W) is the dual of (I, S, U), then

$$(S, I, V) = (S, I, W).$$

Proof: Applying $(\dashv)$ to the systems V, U, resp. U, W yields

$$V(r) = i \iff U(A-i) = T - r$$

$$U(A-i) = T - r \iff W(T-(T-r)) = A-(A-i)$$
$$\iff W(r) = i$$

Hence

$$V(r) = i \iff W(r) = i$$

Hence

$$V = W$$

and so

$$(S, I, V) = (S, I, W)$$

□

II.2. Duality in IPP's

II.2.1. Duality in continuous IPP's

Let

$$(S,I,V) = ([0,T],[0,A],V)$$

be an arbitrary continuous IPP.

The dual IPP of the IPP (S,I,V) is defined to be the IPP

$$(I,S,U) = ([0,A],[0,T],U) \ , \tag{II.5}$$

where

$$U(i) = T - V^{-1}(A-i) \tag{II.6}$$

(here V^{-1} denotes the inverse function of V). It is easy to see that the dual IPP of the IPP (I,S,U) is again the IPP (S,I,V).

We also define

$$\sigma(i) = U'(i) \tag{II.7}$$

for every $i \in I$ and

$$\rho(r) = V'(r) \tag{II.8}$$

for every $r \in S$
(Here U' resp. V' denote the derivative of U resp. V).

Note that, since $V(0) = 0$, $V(r) = \int_0^r \rho(r')\,dr'$ (by (II.8)). From (II.6) it also follows that $U(0) = 0$; hence by (II.7), $U(i) = \int_0^i \sigma(i')\,di'$, for every $r \in [0,T]$ and $i \in [0,A]$.

When expressed as a function of i in the IPP (S,I,V), hence $i = V(r)$, $\rho(r)$ becomes :

$$\rho(i) = V'(V^{-1}(i)) \tag{II.9}$$

We have the following results :

Lemma II.2.1.1 :

$$\sigma(i) = \frac{1}{\rho(A-i)} \quad , \tag{II.10}$$

for every $i \in I$.

Proof :
For every $i \in I$, we have, using (II.6) :

$$U'(i) = \frac{dU}{di}(i)$$

$$= \frac{1}{V'(V^{-1}(A-i))}$$

$$= \frac{1}{\rho(A-i)} \quad ,$$

by (II.9). Hence (II.7) gives

$$\sigma(i) = \frac{1}{\rho(A-i)} \quad ,$$

for every $i \in I$. Note that $\rho \neq 0$ everywhere since V is strictly increasing. □

Corollary II.2.1.2 :
In the IPP (I,S,U) we have :
1. $\rho(i)$ is the density function of the items, in the point $A-i \in I$.
2. $\sigma(i)$ is the density function of the sources, in the point $i \in I$.

Proof :
This follows readily from (II.9), resp. (II.10) and the definition of U and V. □

Alternatively (and equivalently), the functions ρ and σ could have been used as defining functions of a continuous IPP and its dual.

From now on we consider only IPP's with increasing function $\rho > 0$ (so $\sigma > 0$ also is in increasing, by lemma II.2.1.1). This supposition is natural and does not obstruct our general approach; it reduces to an ordering of the set S in the same way as introduced in section II.1.2 (but now for the continuous setting). We further also assume ρ (hence also σ) to be continuous functions.

The functions σ and ρ each introduce a coordinate system on $[0,A] \times [0,T]$, different for σ and ρ. So whenever we use a coordinate (i,r) we have to specify wether it belongs to the σ-system (IPP (I,S,U) : $U(i) = r$) or to the ρ-system (IPP (S,I,V) : $V(r) = i$). We have that (i,r) is a coordinate in the σ-system if and only if $(A-i,T-r)$ is a coordinate in the ρ-system.

The functions ρ and σ are the central tools in our duality approach of continuous IPP's. We can also say that ρ plays the same role in (S,I,V) as σ does in (I,S,U), the dual.

II.2.2. Duality in discrete IPP's

Let (S,I,i) be an arbitrary discrete IPP. Let $p \in \mathbb{N}$ (the set of natural numbers) be fixed but arbitrary. Divide I into p equal sets, henceforth called groups (when there is divisibility problem we allow proportional fractions; in practice this gives rounding offs!). We number the groups as $1,2,\ldots,p$, following the order $\leq$ on I. Hence, because of the ordering on S, the average number $\sigma(i)$ of sources per item in group i, is an increasing function (cf. the analogy with the previous section). Hence, in group i, the average number $\psi(i) = \frac{1}{\sigma(i)}$ of items per source is decreasing. Consequently, the function

$$\rho(i) = \psi(p-i+1) = \frac{1}{\sigma(p-i+1)} \qquad \text{(II.11)}$$

is increasing, for $i = 1,2,\ldots,p$.

The functions ρ and σ, as introduced above, play the same role in discrete IPP's as the functions ρ and σ, introduced in II.2.1 for continuous IPP's (hence the same notation

although they are not the same functions; they cannot be confused since it will always be clear whether we work with a continuous or a discrete IPP).

Formula (II.11) above can also be interpreted as follows : since $\sigma(i)$ measures the average number of sources per item in group i following the order $\prec$ (hence from left to right ($L \to R$)), the number $\rho(i)$ measures the average number of items per source in group i following the reverse order (hence $R \to L$) (see Figs. II.1.a and b).

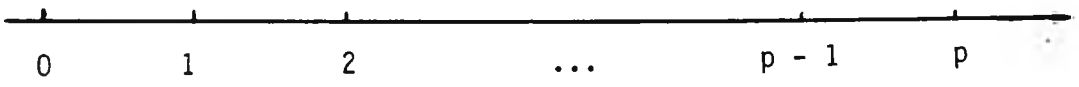


Fig.II.1.a

Fig.II.1.b
The group-axis ($L \to R$, resp. $R \to L$)

In this discrete format the duality approach can be illustrated clearly. A similar interpretation can be applied to the continuous setting.

II.3. The property of pure duality and classical informetrics

II.3.1. The property of pure duality

In the previous section we introduced the tools by which duality in IPP's can be studied : the functions ρ and σ (in the continuous or discrete setting). They are dual functions in the sense that they play the same role in the original resp. the dual situation. The following definition is therefore logical :

Definition II.3.1.1 :
Given an IPP (discrete or continuous), we say that we have the property of pure duality if there exists a constant $C > 0$ such that, for every i ($i=1,\ldots,p$, $p \in \mathbb{N}$ in the discrete case; $i \in I$ in the continuous case)

$$\sigma(i) = C.\rho(i) \qquad (II.12)$$

Otherwise stated, we have pure duality when the dual functions are proportionally the same.

What IPP's satisfy the property of pure duality?

II.3.2. Characterization of discrete IPP's that satisfy the pure duality property; the consequences for classical informetrics

We have the following easy result :

Theorem II.3.2.1 (Egghe, 1989a) :
Let (S,I,i) be any discrete IPP. Fix $p \in \mathbb{N}$. Then this IPP satisfies the pure duality property : i.e. there exists a constant $C > 0$ such that

$$\sigma(i) = C.\rho(i) \qquad (II.12)$$

for every $i = 1,2,\ldots,p$, if and only if

$$\sigma(i)\,\sigma(p-i+1) = C\ , \qquad (II.13)$$

for every $i = 1,2,\ldots,p$.

Proof :
In view of formula (II.11) it is clear that (II.12) is equivalent with (II.13). □

The main point here is that classical informetrics (represented by Bradford's law) satisfies the above property and, apart from some examples not encountered in practice, the Bradford law is the only one satisfying this pure duality property. First we will introduce Bradford's law for discrete IPP's.

Definition II.3.2.2 :
Given any discrete IPP, we say that this IPP satisfies the law of Bradford with p groups ($p \in \mathbb{N}$ fixed, but arbitrary), if we can divide the set I into p equal parts each containing $y_o > 0$ items such that, following the ordering $\prec$ on S, we have a corresponding number of sources equal to (respectively) :

$$r_o,\ r_o k,\ r_o k^2,\ \ldots,\ r_o k^{p-1} \qquad \text{(II.14)}$$

for a certain $r_o > 0$ and $k > 1$.
The number k is called the Bradford factor and is, of course, dependent on p : $k = k(p)$.

Theorem II.3.2.3 (Egghe, 1989a) :
If the discrete IPP satisfies Bradford's law with p groups, then this IPP satisfies the pure duality property : i.e. there exists a constant $C > 0$ such that

$$\sigma(i) = C.\rho(i) \qquad \text{(II.12)}$$

for every $i = 1,2,\ldots,p$.

Proof :
By definition of Bradford's law we have here

$$\sigma(i) = \frac{r_o}{y_o} k^{i-1} \qquad \text{(II.15a)}$$

for every $i = 1,2,\ldots,p$.

Hence

$$\sigma(p-i+1) = \frac{r_o}{y_o} k^{p-i} \qquad \text{(II.15b)}$$

Hence (II.15a) and (II.15b) combined yield

$$\sigma(i)\ \sigma(p-i+1) = \frac{r_o^2}{y_o^2} k^{p-1} ,$$

a constant (independent on i). By theorem II.3.2.1, this IPP satisfies the pure duality property. □

Note :
A slight generalization of (II.13) into

$$\sigma(i)\ \sigma(j) = C'.\sigma(i+j) \qquad \text{(II.16)}$$

for every i,j = 1,...,p, with C' a constant ((II.16) implies (II.13) by taking j = p-i+1) would imply that

$$\sigma(i) = A.k^{i-1} \qquad \text{(II.15c)}$$

for certain constants A and k and for i = 1,...,p, hence Bradford's law (as is easy to prove). Our property (II.13) however is weaker, since there exists a σ which satisfies (II.13) but which is not of the form (II.15c). Indeed, take for instance p even and, denoting [x] for the largest entire number smaller than or equal to x, define $\sigma(1) = [\frac{2}{p}]$, $\sigma(2) = [\frac{4}{p}]$, ..., $\sigma(\frac{p}{2}) = 1$ and for every $i = \frac{p}{2} + 1,...,p$

$$\sigma(i) = \frac{C}{\sigma(p-i+1)} \quad . \qquad \text{(II.17)}$$

(C : an arbitrary positive constant, as in (II.12) or (II.13)). Then (II.17) is also valid for all i. Hence (II.13) is satisfied, σ(i) > 0 for every i, σ is increasing, but σ is not of the form (II.15c).

We must stress however that Bradford's law is the only informetric law we know of that satisfies (II.13) and hence the property of pure duality. This shows the special place of IPP's in classical informetrics (bibliometrics) amongst other IPP's.

II.3.3. Characterisation of continuous IPP's that satisfy the pure duality property and consequences for classical informetrics

As in section II.3.2 we now have (see again (Egghe, 1989a)):

Theorem II.3.3.1 :
Let (S,I,V) be any continuous IPP. Then this IPP satisfies the pure duality property : i.e. there exists a constant $C > 0$ such that

$$\sigma(i) = C.\rho(i) \qquad (II.12)$$

for every $i \in I = [0,A]$, if and only if

$$\sigma(i)\,\sigma(A-i) = C$$

for every $i \in I$.

Proof :
This follows from (II.12) and lemma II.2.1.1 (formula (II.10)). □

This result and the previous section on Bradford's law, leads us to a new definition, which will prove to be very useful in the sequel : the group-free Bradford law for continuous IPP's (and corresponding Bradford function).

Definition II.3.3.2 (Egghe, 1989a) :
Let (S,I,V) be any continuous IPP. We say that this IPP satisfies the group-free law of Bradford if, for every $i \in I$,

$$\sigma(i) = M.K^{i} \qquad (II.18)$$

where $M > 0$ and $K > 1$ are constants.

Formula (II.18) is called the group-free Bradford function.

The number K is called the group-free Bradford factor and, of course, is independent of p in the previous section (p does not exist here!). This definition allows us to recognise Bradford's law as a function just like the other

informetric laws discussed in paragraph I.5. We furthermore have the following result :

Theorem II.3.3.3 (Egghe, 1989a) :
If the continuous IPP satisfies the group free Bradford function, then this IPP satisfies the pure duality property i.e. there exists a constant $C > 0$ such that

$$\sigma(i) = C.\rho(i) \qquad (II.12)$$

for every $i \in I$.

Proof :
Indeed

$$\sigma(i) = M.K^{i}$$

and hence

$$\sigma(A-i) = M.K^{A-i}$$

for every $i \in I$. Consequently

$$\sigma(i)\ \sigma(A-i) = M^2K^A$$

for every $i \in I$. Using theorem II.3.3.1 gives that this IPP satisfies the pure duality property. □

Note 1 :
The comments of section II.3.2 also apply here.
In addition we have the following interesting consequence. Suppose that we have a continuous IPP (S,I,V) (hence, in practice, a large discrete one). If this IPP satisfies Bradford's law then the informetric "calculus" ρ in (S,I,V) is the same as the informetric "calculus" σ in the dual IPP (I,S,U). This means for instance that, if (S,I,V) is a Bradfordian set of citation data (f.i. $S \rightarrow I$, where $\rightarrow$ is the relation "citing") then the "cited" set (I,S,U) satisfies the same informetric laws with the same proportional parameters.

Note 2 :
All definitions and results of section II.3.2 can also be given (and are also true) for continuous IPP's in an obvious way. The results of section II.3.3 are however typical for continuous IPP's. In the next chapter we will compare, for continuous IPP's, the group-dependent and the group-free law of Bradford and provide formulae, relating k(p) (for every $p \in \mathbb{N}$) and K.

The next paragraph gives another application of duality in IPP's.

II.4. General duality properties and applications to the laws of Lotka

In the previous paragraph we proved a first result on duality in IPP's, namely pure duality.

This paragraph deals with more general aspects of duality, valid for general continuous IPP's. We then apply these aspects to Lotka type laws (to be introduced in the sequel), to find conditions on the types of Lotka laws that are possible and on other laws that can be proved, based on Lotka's laws. The classical informetric laws come into this scenario but we also find the generalised Leimkuhler and Bradford laws that are linked with the general laws of Lotka. In the third chapter we will study their mutual interrelations and in the fourth chapter we will devote ourselves to the practical fittings of these laws.

II.4.1. Basic equations for σ and ρ, in general continuous IPP's

Let (S,I,V) be any continuous IPP with dual functions σ and ρ.

We introduce the following function :

$$f : [\rho(0),\rho(A)] \to \mathbb{R}^+ \quad \text{(the positive real numbers)}$$
$$j \to f(j)$$

where f(j) is defined to be the density function (w.r.t. the IPP (S,I,V)) of the number of sources in function of j. Hence, by definition, for every $i \in I$,

$$\int_{\rho(0)}^{\rho(i)} f(j)\, dj$$

denotes the cumulative number of sources for which $j \in [\rho(0),\rho(i)]$, equivalently on the coordinates (in I)

$$i' = \rho^{-1}(j) \in [0,i] \ .$$

This is, by corollary II.2.1.2 equal to :

$$\int_0^i \sigma(A-i')\, di'$$

Hence we have (alternatively to be used as the defining relation for f) :

Source - relationship

$$\int_0^i \sigma(A-i')\, di' = \int_{\rho(0)}^{\rho(i)} f(j)\, dj \qquad \text{(II.19)}$$

for every $i \in I$.

The integral equation (II.19) is difficult to handle because it is inversely retarded. Luckily we have that (II.19) is equivalent with the following easy integral equation :

Item - relationship

$$\int_{\rho(0)}^{\rho(i)} f(j)j\, dj = i \qquad \text{(II.20)}$$

for every $i \in I$.

Theorem II.4.1.1 (Egghe, 1989b) :
Equations (II.19) and (II.20) are equivalent.

Proof :
A. (II.20) is equivalent to

$$f(\rho(i))\ \rho(i)\ \rho'(i) = 1 \qquad \text{(II.21)}$$

Indeed, (II.20) implies (II.21) by differentiation.
From (II.21) we have

$$\int_0^i f(\rho(i'))\ \rho(i')\ \rho'(i')\ di' = i \qquad \text{(II.22)}$$

This gives

$$\int_{\rho(0)}^{\rho(i)} f(j)\ j\ dj = i \qquad \text{(II.20)}$$

using the transformation

$$j = \rho(i'). \qquad \text{(II.23)}$$

B. In the same way we can show that (II.19) is equivalent to

$$\sigma(A-i) = f(\rho(i))\ \rho'(i) \qquad \text{(II.24)}$$

C. Now (II.21) is equivalent to (II.24), using lemma II.2.1.1. Hence (II.19) and (II.20) are also equivalent. □

Consequently, whenever it is necessary, we can ignore Eqn. (II.19) and work with the system

$$\begin{cases} \rho(i) = \dfrac{1}{\sigma(A-i)} & \text{(II.10)} \\ \displaystyle\int_{\rho(0)}^{\rho(i)} f(j)\ j\ dj = i & \text{(II.20)} \end{cases}$$

for every $i \in I = [0,A]$. From now on we will also assume

$\rho(0) = 1$. This is not really necessary but we use it for convenience and since this is always true in practice. Hence we have the system

$$\begin{cases} \rho(i) = \dfrac{1}{\sigma(A-i)} \\ \displaystyle\int_1^{\rho(i)} f(j)j\,dj = i \end{cases} \qquad \text{(II.25)}$$

for every $i \in I = [0,A]$.

Note that from (II.21) (or (II.24)) it follows that f must be decreasing.

We now turn to a first application of this dual formalism.

II.4.2. Exclusion of certain laws of Lotka f

The next theorem is a result for general functions f (as defined in the previous section) that are continuous and strictly positive on the interval $[1,\infty[$. Considering f on the interval $[1,\infty[$ does not mean that we have sources with an unlimited number of items. We just assume the existence of the continuous function, being an extension of the original function. The function f is then, in practice, restricted to the interval $[1,\rho(A)]$.

Theorem II.4.2.1 (Egghe, 1989b) :
If f (restricted to $[1,\rho(A)]$) is the density function of the number of sources in $j \in [1,\rho(A)]$ in a general continuous IPP, and if f is continuous and strictly positive on $[1,\infty[$, then

$$A < \int_1^{\infty} f(j)\ j\ dj \qquad \text{(II.26)}$$

Proof :
From (II.20) we find that

$$A = \int_1^{\rho(A)} f(j)\ j\ dj \qquad \text{(II.27)}$$

Suppose that

$$\int_{\rho(A)}^{\infty} f(j)\ j\ dj = 0 \qquad \text{(II.28)}$$

Then the function $j \to f(j)\, j$ is zero almost everywhere on $[\rho(A),\infty[$ in the Lebesgue-sense. But f is continuous. Hence the function $j \to f(j)\, j$ is identically zero on $[\rho(A),\infty[$. Hence f(j) is zero on $[\rho(A),\infty[$, a contradiction. Hence

$$\int_{\rho(A)}^{\infty} f(j)\, j\, dj > 0 \,. \qquad \text{(II.29)}$$

(II.27) and (II.29) together yield (II.26). □

This result has an unexpected consequence for the Lotka functions :

Corollary II.4.2.2 (Egghe, 1989b) :
Suppose that (S,I,V) is a continuous IPP with function f (we define this function to be the general Lotka function, cf. section I.5.1)

$$f(j) = \frac{C}{j^{\alpha}} \qquad \text{(II.30)}$$

for every $j \in [1,\infty[$, where $\alpha > 1$. Then

$$\alpha < \frac{C}{A} + 2 \,. \qquad \text{(II.31)}$$

Proof :
From the previous theorem we see that

$$A < \int_{1}^{\infty} f(j)\, j\, dj \qquad \text{(II.32)}$$

Hence, upon integrating the function (II.30) (which obviously satisfies the requirements of the above theorem), we have :

a. If $\alpha < 2$, then (II.31) is automatically satisfied.

b. If $\alpha > 2$, then

$$\int_{1}^{\infty} f(j)\, j\, dj = \frac{C}{\alpha - 2} \qquad \text{(II.33)}$$

Hence (II.32) and (II.33) yield

$$A < \frac{C}{\alpha - 2} \,,$$

hence (II.31). □

This, in turn yields a further surprising.

Corollary II.4.2.3 (Egghe, 1989b) :
If (S,I,V) is as in the previous corollary, then $\alpha > 3$ implies :

$$f(1) = C > A$$

Proof :
Indeed, corollary II.4.2.2 yields

$$\alpha < \frac{C}{A} + 2 \quad . \tag{II.31}$$

Hence $\alpha > 3$ implies

$$f(1) = C > A \; . \quad \square \tag{II.34}$$

Note :
Although it is theoretically possible to have (II.34) (since f is a density function), the case $\alpha > 3$ is very likely to be excluded if the Lotka function (II.30) must fit a practical IPP. Indeed, in practical, discrete IPP's, $C = f(1)$ denotes the number of sources with one item and hence $C < A$.

In the literature we indeed find examples where $\alpha > 3$ (see e.g. (Pao, 1986)). They do not contradict the above remarks since the fittings are statistical and hence not based on a mathematical theory. Also practical data can differ from Lotka's function (random fluctuations). Furthermore, in most cases we do not know the complete IPP (usually missing the least productive sources) or we do not use the complete IPP (as in (Pao, 1986)) : in this case A is lower than in reality and hence, according to corollary II.4.2.3, $\alpha > 3$ is possible.

We can however conclude that, in general, $\alpha < 3$ will be more often encountered than $\alpha > 3$. The above theory is a first theoretical basis for it.

II.4.3. The Bradford and Leimkuhler type laws that are implied by the general law of Lotka $f(j) = \frac{C}{j^{\alpha}}$, $j \in [\rho(0),\rho(A)] = [1,\rho(A)]$.

II.4.3.1. The case $\alpha = 2$

This case is - in essence - known (cf. (Egghe, 1985)) but will be presented here in a new variant, namely based on the duality system (II.25). In the next chapter we will prove that the results developed here are the same, essentially, as those of (Egghe, 1985), but presented here more accurately.

Theorem :
Let (S,I,V) be any continuous IPP with Lotka function

$$f(j) = \frac{C}{j^2} \qquad (II.35)$$

($j \in [1,\rho(A)]$). Then

(i) This IPP satisfies the group-free version of Bradford's law (definition II.3.3.2).

(ii) This IPP conforms with the Leimkuhler law, to be defined now (cf. section I.5.5). In the IPP (I,S,U) : Let R(r) denote the cumulative number of items in the sources $s \in [0,r]$, for every $r \in [0,T]$ (Hence $R = U^{-1}$). Then

$$R(r) = a \log (1 + br) , \qquad (II.36)$$

where a and b are constants, and $r \in [0,T]$. Function (II.36) is called Leimkuhler's function.

Proof : Proof of (i)
From (II.20) and (II.35) we find, for every $i \in I$:

$$\int_1^{\rho(i)} \frac{C}{j}\, dj = i$$

Hence

$$C \log \rho(i) = i$$

(here log denotes the Naperian logarithm $\log_e$).

Using also (II.10) this gives :

$$-C \log (\sigma(A-i)) = i$$

We now use the transformation $A-i = i'$ yielding

$$C \log \sigma(i') = i' - A$$

Dropping the primes we can write

$$\sigma(i) = e^{-\frac{A}{C}} e^{\frac{i}{C}}$$

$$\sigma(i) = M.K^i \quad , \qquad \text{(II.37)}$$

for every $i \in I$, where M and K are constants. Note that $M > 0$ and $K > 1$. Hence we have found the group-free law of Bradford (II.18). Note that $\rho(i) = e^{i/C} = K^i$ for every $i \in I$.

Note also that, for every $i \in I$, $\rho(i) > 1$ and $\sigma(i) < 1$. This property is encountered every time that $V(r) > r$, $\forall r \in [0,T]$ (i.e. there are more items V(r) than sources r, $\forall r$) which is fairly evident in practice, and certainly so whenever we have Lotka's function (II.35). We did not put the condition $V(r) > r$, $\forall r \in [0,T]$ right from the beginning (in section II.1.1) since there was no need for it and since we wanted to be as general as possible (allowing for sources with production less than one). An analogous remark can be made for discrete IPP's.

Proof of (ii)

In our formalism, we clearly have, when R(r) = i, that

$$r = \int_0^i \sigma(i')\, di' \qquad \text{(II.38)}$$

Using (II.37) above, we have

$$r = \int_0^i \sigma(i')\, di' = \int_0^i M.K^{i'}\, di'$$

$$= \frac{M}{\log K}\,(K^i - 1)$$

Hence

$$R(r) = i = \frac{1}{\log K} \log\left(1 + r\,\frac{\log K}{M}\right), \qquad \text{(II.39)}$$

which is of the form

$$R(r) = a \log (1 + br), \qquad \text{(II.36)}$$

where

$$a = \frac{1}{\log K}, \qquad b = \frac{\log K}{M} \qquad \text{(II.40)}$$

and $r \in [0,T]$. □

II.4.3.2. The general case

Performing as in II.4.3.1 we can now construct the new functions $\sigma(i)$ (Bradford's function) and $\rho(i)$ that follow from the general Lotka function (II.30), and from it, the new function R(r), Leimkuhler's function.

Theorem (Egghe, 1989b) :
Let (S,I,V) be any continuous IPP with Lotka function

$$f(j) = \frac{C}{j^{\alpha}} \qquad \text{(II.41)}$$

($j \in [1,\rho(A)]$), where $\alpha \neq 2$ but $\alpha > 1$. Then

(i)

$$\rho(i) = \left(\frac{i(2-\alpha)}{C} + 1\right)^{\frac{1}{2-\alpha}} \qquad \text{(II.42)}$$

$$\sigma(i) = \left(\left(\frac{A(2-\alpha)}{C} + 1\right) - i\,\frac{2-\alpha}{C}\right)^{-\frac{1}{2-\alpha}} \qquad \text{(II.43)}$$

for every $i \in [0,A]$. Hence the general Bradford function, if $\alpha \neq 2$, is of the form

$$\sigma(i) = (A_1 + i\ A_2)^{A_3}\ , \qquad \text{(II.44)}$$

where A_1, A_2 and A_3 are constants.

(ii) In the IPP (I,S,U), let R(r) denote the cumulative number of items in the sources $s \in [0,r]$, for every $r \in [0,T]$. Then

$$R(r) = \frac{C}{2-\alpha}\left[\rho(A)^{2-\alpha} - \left(\rho(A)^{1-\alpha} + \frac{\alpha-1}{C}\ r\right)^{\frac{2-\alpha}{1-\alpha}}\right] \qquad \text{(II.45)}$$

for every $r \in [0,T]$, where $\rho(A)$ is as in (II.42) for $i = A$, the maximal density of items.

Proof : Proof of (i) :
(II.44) follows from (II.43) and (II.43) follows from (II.42), using (II.10). Hence we only have to show (II.42). From (II.20) it follows that

$$\int_1^{\rho(i)} \frac{C}{j^{\alpha-1}}\ dj = i$$

for every $i \in I$. Hence

$$\frac{C}{2-\alpha}\ (\rho(i)^{2-\alpha} - 1) = i$$

Consequently

$$\rho(i) = \left(\frac{i(2-\alpha)}{C} + 1\right)^{\frac{1}{2-\alpha}} \qquad \text{(II.46)}$$

under the condition that

$$\frac{i(2-\alpha)}{C} + 1 > 0 \qquad \text{(II.47)}$$

for every $i \in [0,A]$. To prove this, invoke corollary II.4.2.2, yielding, if $\alpha > 2$:

$$\frac{A}{C}\ (2-\alpha) + 1 > 0 \qquad \text{(II.48)}$$

a) If $\alpha < 2$ then

$$\frac{i(2-\alpha)}{C} + 1 > 0$$

always.

b) If $\alpha > 2$ then

$$\frac{A(2-\alpha)}{C} + 1 = \min_{i\in[0,A]} \left(\frac{i(2-\alpha)}{C} + 1\right) \qquad \text{(II.49)}$$

So, (II.49) and (II.48) imply

$$\frac{i(2-\alpha)}{C} + 1 > 0$$

for every $i \in [0,A]$.

In conclusion, (II.47) is satisfied for every $i \in [0,A]$ and every $\alpha \neq 2$; hence also (II.46).

Proof of (ii) :

We have

$$r = \int_0^i \sigma(i')\, di'$$

if $R(r) = i$.

Applying (II.44) we have

$$r = \int_0^i \sigma(i')\, di' = \frac{(A_1 + iA_2)^{1+A_3} - A_1^{1+A_3}}{A_2\,(1+A_3)}$$

Hence (using $R(r) = i$), we have

$$R(r) = \frac{1}{A_2}\left[(A_1^{1+A_3} + A_2(1+A_3)r)^{\frac{1}{1+A_3}} - A_1\right] \qquad \text{(II.50)}$$

We now interprete A_1, A_2 and A_3 by means of (II.43) and thus obtain for every $r \in [0,T]$:

$$R(r) = \frac{C}{2-\alpha}\left[\left(\frac{A(2-\alpha)}{C} + 1\right) - \left(\left(\frac{A(2-\alpha)}{C} + 1\right)^{\frac{1-\alpha}{2-\alpha}} - \frac{1-\alpha}{C} r\right)^{\frac{2-\alpha}{1-\alpha}}\right] \qquad (II.51)$$

But, using (II.42) we see that

$$\rho(A) = \left(\frac{A(2-\alpha)}{C} + 1\right)^{\frac{1}{2-\alpha}}$$

Hence (II.51) becomes :

$$R(r) = \frac{C}{2-\alpha}\left[\rho(A)^{2-\alpha} - \left(\rho(A)^{1-\alpha} - \frac{1-\alpha}{C} r\right)^{\frac{2-\alpha}{1-\alpha}}\right] \qquad (II.45)$$

completing the proof of this theorem. □

Note 1 :
The form of the function (II.45) was first derived by Rousseau (1988a) by other methods; the functions (II.42) and (II.43) however are new and hence, since (II.45) is derived from these formulas and also since our dual approach is new, the above proof of (II.45) is new.

Note 2 :
From formula (II.43) it follows that

$$\lim_{\alpha \to 2} \sigma(i)$$

is an exponential function of the form (II.18), hence the function $\sigma(i)$ for $\alpha = 2$ (Bradford's group-free version). Hence our theory for $\alpha \neq 2$ gives the classical Bradford function ($\alpha = 2$) as a limiting case (as it should).

This is the first time that Bradford's law for the general Lotka law (II.41) is proved. In (Egghe, 1985) we tried to put up a qualitative model for Bradford's law in case of formula (II.41). Although not perfect we predicted the next corollary (which we can now prove in an exact way!).

Corollary :
If the continuous IPP satisfies (II.41) then the corresponding law of Bradford $\sigma(i)$ satisfies

$\frac{\sigma'(i)}{\sigma(i)}$ increases with i if $\alpha < 2$

$\frac{\sigma'(i)}{\sigma(i)}$ decreases with i if $\alpha > 2$

$\frac{\sigma'(i)}{\sigma(i)}$ is constant if $\alpha = 2$

Proof :
Suppose that $\alpha \neq 2$. Formula (II.44) yields

$$\sigma'(i) = A_2 A_3 (A_1 + iA_2)^{A_3 - 1}$$

Hence

$$\frac{\sigma'(i)}{\sigma(i)} = \frac{A_2 A_3}{A_1 + iA_2}$$

Now, substituting the values of A_1, A_2 and A_3, in terms of A, C and α (using (II.43)) yields :

$$\frac{\sigma'(i)}{\sigma(i)} = \frac{1}{A(2-\alpha) + C - i(2-\alpha)} \qquad \text{(II.52)}$$

This is an increasing function if $\alpha < 2$ and a decreasing one if $\alpha > 2$. If $\alpha = 2$, the result is well-known : Formula (II.18) yields

$$\frac{\sigma'(i)}{\sigma(i)} = \log K \ , \qquad \text{(II.53)}$$

a constant. □

As is shown in (Rousseau, 1988a), the graph of the function R (formula (II.45)), in semilogarithmic (log r, R(r))-scale , shows an inflection point (in bibliometrics one calls this a Groos droop since Groos was the first to find such a "deviation" from the log-form, cf. (Groos, 1967)) for $\alpha < 2$ and has no inflection point for $\alpha > 2$ (as predicted also in (Egghe, 1985)). If there is a Groos droop ($\alpha < 2$), the inflection point is given by :

$$r_d = \frac{C}{2-\alpha} \left(\frac{A(2-\alpha)}{C} + 1\right)^{\frac{1-\alpha}{2-\alpha}} \qquad \text{(II.54)}$$

See section IV.2.8 for some basic notes on the Groos droop.

We now turn our attention to further links between the encountered informetric laws.

CHAPTER III : THE INFORMETRIC LAWS : CLASSIFICATION, APPROXIMATIONS AND PARAMETER DETERMINATION

This third chapter deals with relations between the informetric functions so far identified. We first classify these functions and then derive some relations between their parameters. "Legalised" approximations are introduced in a formal way.

III.1. Classification of informetric functions

We will classify the functions we have encountered so far. For the sake of completeness we repeat the formulae; for the meaning of them, we refer to the place were they have been introduced. We restrict ourselves to continuous IPP's $(S,I,V) = ([0,T],[0,A],V)$, with dual (I,S,U).

III.1.1. Informetric laws

1. The Lotka function (cf. I.5.1 and II.4.2.2)

$$f(j) = \frac{C}{j^{\alpha}} , \qquad \text{(III.1)}$$

where C and α are constants, $\alpha > 1$ and $j \in [1,\rho(A)] = [\rho(0),\rho(A)]$.

2. The Zipf or Pareto function (cf. I.5.3 and I.5.4). Consider the IPP (I,S,U) : let $g(r)$ denote the density of the number of items in $r \in [0,T]$. Then

$$g(r) = \frac{F}{(1+r)^{\beta}} \qquad \text{(III.2)}$$

where F and β are constants and $r \in [0,T]$. (In our framework : $r \in [0,T]$; so,in formula (III.2) our ranks start in 1 which is natural, but which will also be explained further on).

3. The Mandelbrot function (cf. I.5.3). Consider the IPP (I,S,U). Let g(r) denote the density of the number of items in r ∈ [0,T]. Then

$$g(r) = \frac{G}{(1+Hr)^{\beta'}} \quad , \qquad \text{(III.3)}$$

where G, H and β' are constants and r ∈ [0,T].
Note that g(r) = ρ(T-r) for every r ∈ [0,T].

4. The Leimkuhler function (cf. I.5.5 and II.4.3.1)

$$R(r) = a \log (1 + br) \quad , \qquad \text{(III.4)}$$

where a and b are constants and r ∈ [0,T].
Note that $R = U^{-1}$, the inverse function of U.

5. The generalised Leimkuhler function (cf. II.4.3.2)

$$R(r) = \frac{C}{2-\alpha} \left[\rho(A)^{2-\alpha} - \left(\rho(A)^{1-\alpha} + \frac{\alpha-1}{C} r\right)^{\frac{2-\alpha}{1-\alpha}}\right] \qquad \text{(III.5)}$$

with C, α and ρ(A) constants, r ∈ [0,T] and α ≠ 2.
They are the same as in 1.

6. Bradford's law (cf. I.5.6, II.3.2.2 and note 2 in section II.3.3). Fix p ∈ ℕ. We can divide the set I into p equal parts, each of length y_o such that the (with U) corresponding division in S has length respectively

$$r_o, r_o k, r_o k^2, \ldots, r_o k^{p-1} \qquad \text{(III.6)}$$

for a certain r_o and k > 1. This k, of course, is p-dependent k = k(p).

7. The group-free Bradford function (cf. II.3.3.2)

$$\sigma(i) = M.K^i \quad , \qquad \text{(III.7)}$$

where M and K are constants, K > 1, and i ∈ I = [0,A].

8. The generalised group-free Bradford function (cf. II.4.3.2).

$$\sigma(i) = \left(\left(\frac{A(2-\alpha)}{C} + 1\right) - i\,\frac{2-\alpha}{C}\right)^{-\frac{1}{2-\alpha}}, \qquad \text{(III.8)}$$

where A, C and α are constants, $\alpha \neq 2$ and $i \in I = [0,A]$. C and α are the same as in 5. and as in 1..

III.1.2. The informetric functions equivalent with Lotka's function with $\alpha = 2$

When we say "equivalent" functions, we mean equivalent in the mathematical sense. Of course, together with these equivalencies we will prove some relations between the parameters (the constants) in the respective functions. They will be very useful in the sequel.
Some proofs in the theorem below are partially in (Egghe, 1985 and 1989c) and (Rousseau, 1987a).

Theorem :
Let (S,I,V) be any continuous IPP. Then we have the following equivalencies :

(i) The IPP satisfies Lotka's function (III.1) with $\alpha = 2$.

(ii) The IPP satisfies Mandelbrot's function (III.3) with $\beta' = 1$.

(iii) The IPP satisfies Leimkuhler's function (III.4).

(iv) The IPP satisfies the law of Bradford (III.6), for every $p \in \mathbb{N}$.

(v) The IPP satisfies the group-free function of Bradford (III.7).

Assuming the validity of these equivalent functions, we have the following relations between the parameters :

$$a = \frac{y_0}{\log k} = \frac{1}{\log K} \qquad \text{(III.9)}$$

$$b = \frac{k-1}{r_0} = \frac{\log K}{M} \qquad \text{(III.10)}$$

Here y_0, k and r_0 form a valid Bradford triple as in (III.6)

(dependent on p) but a and b are independent of p (as are M and K)

$$G = \rho(A) = ab \tag{III.11}$$

$$H = \frac{\rho(A)}{C} = b \tag{III.12}$$

$$K = k(p)^{\frac{p}{A}} , \tag{III.13}$$

for every $p \in \mathbb{N}$. Here we denote $k = k(p)$.
Consequently, one has also

$$C = Aa \tag{III.14}$$

$$y_o = C \log k \tag{III.15}$$

$$r_o = \frac{C}{\rho(A)} (k-1) \tag{III.16}$$

Proof : Proof of the equivalence between (i) and (ii)

The proof is based on the general relation for $j \in [1,\rho(A)]$:

$$g^{-1}(j) = r(j) = \int_j^{\rho(A)} f(j')\, dj' , \tag{III.17}$$

which is intuitively clear but we will prove it now, in an exact way. Consider a valid triple (r,i,j) in (I,S,U) : i.e. $i \in [0,A]$, $r = U(i) \in [0,T]$ and $j \in [1,\rho(A)]$. From corollary II.2.1.2 we have that $j = \rho(A-i)$.

Using (II.7) and the notes following it, we have

$$r = U(i) = \int_0^i \sigma(i')\, di'$$

Hence, using the transformation $i'' = A - i'$:

$$r = -\int_{i''=A}^{i''=A-i} \sigma(A-i'')\, di'' = \int_{A-i}^{A} \sigma(A-i'')\, di''$$

$$= \int_{0}^{A} \sigma(A-i'')\, di'' - \int_{0}^{A-i} \sigma(A-i'')\, di'' .$$

Using (II.19) twice implies ($\rho(0) = 1$) :

$$r = \int_{1}^{\rho(A)} f(j)\, dj - \int_{1}^{\rho(A-i)} f(j)\, dj$$

$$r = \int_{\rho(A-i)}^{\rho(A)} f(j)\, dj$$

Hence, since $j = \rho(A-i)$ we find

$$r = \int_{j}^{\rho(A)} f(j')\, dj' \quad .$$

Since, by definition, $j = g(r)$ we also have $g^{-1}(j) = r$. Hence (III.17) is proved.

We now prove the different implications :

a) <u>(i) implies (ii)</u>

Since, assuming (III.1) with $\alpha = 2$,

$$g^{-1}(j) = r(j) = \int_{j}^{\rho(A)} \frac{C}{j'^2}\, dj' \qquad \text{(III.18)}$$

$$= C\left(\frac{1}{j} - \frac{1}{\rho(A)}\right)$$

we also have, putting again $j = g(r)$ and $r = r(j)$:

$$g(r) = \frac{\rho(A)}{1 + \frac{\rho(A)}{C}\, r} \qquad \text{(III.19)}$$

Hence, this implication is shown, together with the first half of the equalities in (III.11) and (III.12).

b) <u>(ii) implies (i)</u>

From (III.17) we have also

$$f(j) = -(g^{-1})'(j) = -r'(j) \qquad \text{(III.20)}$$

Assuming (III.3) this gives (with $\beta' = 1$) and again using that $j = g(r)$:

$$r = \frac{1}{H}\left(\frac{G}{j} - 1\right) \qquad \text{(III.21)}$$

Hence, (III.20) and (III.21) yield

$$f(j) = \frac{G}{H} \cdot \frac{1}{j^2} ,$$

completing this part of the proof.

Proof of the equivalence between (ii) and (iii)

This proof is based on the general defining relation (definition of R and g) :

$$R(r) = \int_0^r g(r')\, dr' \qquad \text{(III.22)}$$

a) (ii) implies (iii)
(III.22) together with (III.3) (for $\beta' = 1$) gives :

$$R(r) = \frac{G}{H} \log (1 + Hr) , \qquad \text{(III.23)}$$

for every $r \in [0,T]$, yielding Leimkuhler's law.

b) (iii) implies (ii)
(III.22) and (III.4) give :

$$g(r) = R'(r) = \frac{ab}{1 + br} \qquad \text{(III.24)}$$

for every $r \in [0,T]$. This also agrees with the second half of the equalities in (III.11) and (III.12).

Proof of the equivalence of (iii) and (iv)

a) (iii) implies (iv)
Let $p \in \mathbb{N}$ be fixed but arbitrary. Let $y_o = \frac{A}{p}$ and r_o be such that $R(r_o) = y_o$. Define $k > 1$ such that $R(r_o + r_o k) = 2\, y_o$. Using (III.4) we see that, if

$r = r_o + r_o k + \ldots + r_o k^{i-1}$ $(i = 2,\ldots,p)$ then

$$R(r) = i\, y_o \,. \qquad \text{(III.25)}$$

Indeed, from $R(r_o + r_o k) = 2\, y_o = 2R(r_o)$ we find $k = 1 + br_o$. So

$$r = r_o + r_o k + \ldots + r_o k^{i-1}$$

$$= r_o \frac{k^i - 1}{k - 1}$$

$$= \frac{(1 + br_o)^i - 1}{b} \,.$$

Hence

$$R(r) = i\, R(r_o) = i\, y_o \,,$$

for every $i = 2,\ldots,p$.
This relation is equivalent to the law of Bradford for p groups. A similar argument could be performed for every $p \in \mathbb{N}$. Hence (iv) is proved.

b) <u>(iv) implies (iii)</u>
This proof is not trivial and requires several steps.

A. If we show that the functions R and R^{-1} are differentiable, then they also must be continuous. The fact that they are differentiable follows from (III.22). Now $g(r) = \rho(T-r)$ for every $r \in [0,T]$ as follows from the definition of g (section III.1.1). Hence (III.22) becomes :

$$R(r) = \int_0^r \rho(T-r')\, dr'$$

for every $r \in [0,T]$. Hence $R'(r) = \rho(T-r)$, for every $r \in [0,T]$ and since

$$(R^{-1})'(i) = \frac{1}{R'(R^{-1}(i))}$$

$$= \frac{1}{p(T - R^{-1}(i))} \quad ,$$

the proof is finished, since $p > 1 > 0$ ~~(cf. II.4.3.1 - end of the proof of (i)).~~

B. Denote by A the set

$$A = \{r_o \frac{k^i - 1}{k - 1} \parallel (r_o, k, p^\ell) \text{ is a valid triple in Bradford's law, p fixed (take e.g. } p = 3) \ \ell \in \mathbb{N}, \text{ and } i = \frac{q}{p^\ell} A, \text{ where } q = 1,2,\ldots,p^\ell.\}$$

This set A is <u>dense</u> in $[0,T]$ (a set X is said to be dense in a set Y if every element of Y can be written as the limit of a sequence of elements of X).

<u>Proof</u> :
Since we have the validity of Bradford's law, for every $p \in \mathbb{N}$, we can consider Bradford situations for a number of groups respectively $p, p^2, p^3, \ldots$. In each case p^ℓ we have a division of the item set $[0,A]$ at the points

$$A_\ell = \{\frac{A}{p^\ell}, \frac{2A}{p^\ell}, \ldots, A\} \ ,$$

which is a subset of the divisions in the case $p^{\ell+1}$:

$$A_{\ell+1} = \{\frac{A}{p^{\ell+1}}, \frac{2A}{p^{\ell+1}}, \ldots, \frac{pA}{p^{\ell+1}} = \frac{A}{p^\ell}, \ldots, A\} \ .$$

By taking $\ell \in \mathbb{N}$ high enough we can make the length between two consecutive divisions as small as we wish. From the form of A_ℓ we see that $\bigcup_{\ell \in \mathbb{N}} A_\ell$ is dense in $[0,A]$. Now, as R^{-1} is continuous, we see that

$$[0,T] = R^{-1}([0,A])$$

$$= R^{-1}\left(\overline{\bigcup_{\ell \in \mathbb{N}} A_\ell}\right),$$

(where $\overline{\bigcup_{\ell \in \mathbb{N}} A_\ell}$ denotes the closure of the set $\bigcup_{\ell \in \mathbb{N}} A_\ell$)

$$[0,T] \subset \overline{R^{-1}\left(\bigcup_{\ell \in \mathbb{N}} A_\ell\right)}$$

But, as given by (iv),

$$R^{-1}\left(\bigcup_{\ell \in \mathbb{N}} A_\ell\right) = A$$

(since, for every i : $r_o + r_o k + \ldots + r_o k^{i-1} = r_o \frac{k^i - 1}{k - 1}$)

Hence A is dense in $[0,T]$.

C. Fix $p \in \mathbb{N}$ arbitrarily. We apply Bradford's law : we have $R(r) = i\, y_o$ for

$$r = r_o + r_o k + \ldots + r_o k^{i-1}$$

$$r = r_o \frac{k^i - 1}{k - 1},$$

where $i = 1,2,\ldots,p$.
Hence

$$r = r_o \left(\frac{k^{\frac{R(r)}{y_o}} - 1}{k-1}\right)$$

yielding

$$R(r) = \frac{y_o}{\log k} \log\left(1 + \left(\frac{k - 1}{r_o}\right) r\right), \qquad \text{(III.26)}$$

which is Leimkuhler's function for

$$r = r_o \left(\frac{k^i - 1}{k - 1}\right), \qquad \text{(III.27)}$$

$i = 1,2,\ldots,p$. Here we see that

$$a = \frac{y_o}{\log k}$$

$$b = \frac{k - 1}{r_o}$$

D. Let a_i resp. b_i be the above values when there are p^i divisions ($i = 1,2,3,\ldots$) ($p \in \mathbb{N}$ fixed; take e.g. $p = 3$). Then

$$\begin{cases} a_i = a_{i+1} \\ b_i = b_{i+1} \end{cases} \qquad \text{(III.28)}$$

for every $i = 1,2,\ldots$.

1st Proof :

Indeed, for every $i = 1,2,\ldots$ the divisions with p^{i+1} groups are a refinement of the divisions with p^i groups. So we have p^i common points. Select any two of them : r_1 and $r_2 \in S$, $r_1 \neq r_2$. Then

$$\begin{cases} R(r_1) = a_i \log (1 + b_i r_1) = a_{i+1} \log (1 + b_{i+1} r_1) \\ R(r_2) = a_i \log (1 + b_i r_2) = a_{i+1} \log (1 + b_{i+1} r_2) \end{cases} .$$

This system has only one solution :

$$\begin{cases} a_i = a_{i+1} \\ b_i = b_{i+1} \end{cases} \qquad \text{(III.28)}$$

2nd Proof :

We show it for (a_1,b_1) resp. (a_2,b_2). (III.28) follows then by induction.

Let (r_o,y_o,k,p) and (r'_o,y'_o,k',p^2) be the respective Bradford parameters. Then

$$\begin{cases} a_1 = \dfrac{y_o}{\log k}\,, & a_2 = \dfrac{y'_o}{\log k'} \\ b_1 = \dfrac{k-1}{r_o}\,, & b_2 = \dfrac{k'-1}{r'_o} \end{cases} \qquad \text{(III.29)}$$

according to C. But

$$y_o = p\, y'_o\,, \qquad \text{(III.30)}$$

obviously and so

$$r_o = r'_o + r'_o k' + \ldots + r'_o k'^{p-1}$$

$$r_o = r'_o \left(\frac{k'^p - 1}{k' - 1}\right) \qquad \text{(III.31)}$$

Also,

$$r'_o = \frac{T(k'-1)}{k'^{p^2} - 1}$$

so

$$r_o = T\,\frac{k'^p - 1}{k'^{p^2} - 1} \qquad \text{(III.32)}$$

But, as

$$r_o = T\,\frac{k-1}{k^p - 1} \qquad \text{(III.33)}$$

we see from (III.32) and (III.33) that

$$k = k'^p \qquad \text{(III.34)}$$

Now (III.30) and (III.34) give

$$a_1 = \frac{y_o}{\log k} = \frac{p y'_o}{\log k'^p} = \frac{y'_o}{\log k'} = a_2 \qquad \text{(III.35)}$$

and so it follows that

$$b_1 = b_2 \qquad \text{(III.36)}$$

~~(take one common point r in the p- and p^2-division).~~

As all a's, and all b's are equal, we have verified the validity of

$$R(r) = a \log (1 + br) \qquad \text{(III.37)}$$

in the points $r \in A$. This follows indeed from C and D.
E. Since R is continuous, since A is dense in [0,T] and since the function

$$r \to a \log (1 + br)$$

is already a continuous extension of R to [0,T], we can conclude that

$$R(r) = a \log (1 + br)$$

for every $r \in [0,T]$, where a and b are constants. We have also shown the first equalities in (III.9) and (III.10), where a and b are independent of p. Hence with these formulae and (III.11) and (III.12), also (III.14), (III.15) and (III.16) are shown.

Note :
From (III.9) it follows that

$$k(p)^p = \text{constant} , \qquad \text{(III.38)}$$

independent of p.

Proof :
Indeed : let k_1 correspond to a Bradford division into p_1 groups and k_2 correspond to a Bradford division into p_2 groups. Then, according to (III.9) and the reasoning of "(iv) implies (iii)" above we have

$$a = \frac{\frac{A}{p_1}}{\log k_1} = \frac{\frac{A}{p_2}}{\log k_2} \qquad \text{(III.39)}$$

Hence

$$\log k_1^{p_1} = \log k_2^{p_2} = \frac{A}{a} \qquad \text{(III.40)}$$

(III.40) yields now

$$k_1^{p_1} = k_2^{p_2}$$

hence

$$k(p)^p = \text{constant.}$$

We now show

Proof of the equivalence of (iv) and (v)

a) (iv) implies (v)

Let, first $i \in [0,A]$ be such that there is a q and $p \in \mathbb{N}$ such that $i = \frac{qA}{p}$.

Denote by S(i) :

$$S(i) = \int_0^i \sigma(i')\, di' \qquad \text{(III.41)}$$

We apply (iv) with $p \in \mathbb{N}$ as above. This yields p groups of respectively

$$r_o(p),\ r_o(p)k(p),\ \ldots,\ r_o(p)\,k(p)^{p-1}$$

sources, each containing $y_o(p) = \frac{A}{p}$ items. Hence

$$S(i) = r_o(p) + r_o(p)k(p) + \ldots + r_o(p)k(p)^{q-1}$$

$$= r_o(p)\,\frac{k(p)^q - 1}{k(p) - 1}$$

$$= \frac{r_o(p)}{k(p) - 1}\,(k(p)^{\frac{ip}{A}} - 1)$$

$$= \frac{r_o(p)}{k(p) - 1} \; p \log k(p) \; \frac{k(p)^{\frac{ip}{A}} - 1}{p \log k(p)}$$

$$= \frac{r_o(p)}{k(p) - 1} \; p \log k(p) \int_0^{\frac{i}{A}} (k(p)^p)^{i'} \, di'$$

$$= \frac{r_o(p)}{k(p) - 1} \; p \log k(p) \; \frac{1}{A} \int_0^{i} (k(p)^{\frac{p}{A}})^{i''} \, di'' \; , \quad \text{(III.42)}$$

using the transformation $i'' = A.i'$.
Furthermore,

$$r_o(p) + r_o(p)\, k(p) + \ldots + r_o(p)\, k(p)^{p-1} = T \; ,$$

the total number of sources. Hence

$$r_o(p) = \frac{T(k(p) - 1)}{k(p)^p - 1} \qquad \text{(III.43)}$$

Hence, if we put

$$M = \frac{r_o(p)}{k(p) - 1} \; p \log k(p) \; \frac{1}{A} \qquad \text{(III.44)}$$

then

$$M = \frac{\log (k(p)^p)}{k(p)^p - 1} \; \frac{T}{A}$$

Since we suppose (iv) we have also (III.38), concluding that M above is a constant. Put

$$K = k(p)^{\frac{p}{A}} \qquad \text{(III.45)}$$

Then we have that

$$S(i) = M \int_0^{i} K^{i'} \, di' \qquad \text{(III.46)}$$

for every $i = \frac{q}{p} A$, for a certain q, $p \in \mathbb{N}$, $q < p$. From (III.41) it follows that the function S is continuous. But since (III.46) is valid on a dense subset of $[0,A] = I$ and since the function

$$i \to M \int_0^i K^{i'}\, di'$$

is continuous on [0,A], we conclude that

$$S(i) = M \int_0^i K^{i'}\, di' \qquad \text{(III.47)}$$

for every $i \in I = [0,A]$. Compare this with (III.41) and differentiate. We then have (since σ is continuous) :

$$\sigma(i) = M.K^i$$

for every $i \in I$, showing (v).

b) <u>(v) implies (iv)</u>

Let $p \in \mathbb{N}$ be arbitrary. Define $y_o(p) = \frac{A}{p}$. Then, according to (v)

$$r_o(p) = S(y_o) = \int_0^{y_o} M\, K^{i'}\, di'$$

$$= \frac{M}{\log K} (K^{y_o} - 1)$$

Furthermore

$$S(2y_o) - S(y_o) = \int_{y_o}^{2y_o} M.K^{i'}\, di'$$

$$= \frac{M}{\log K} K^{y_o} (K^{y_o} - 1)$$

Hence we see that

$$S(2y_o) - S(y_o) = r_o(p)\, K^{y_o}$$

$$= r_o(p)\, K^{\frac{A}{p}}$$

In the same way we can show that, for every $q = 1,2,\ldots,p$:

$$S(qy_o) - S((q-1)y_o) = r_o(p)\, (K^{\frac{A}{p}})^{q-1}$$

Hence, if we put

$$k(p) = K^{\frac{A}{p}}$$

then

$$S(qy_o) - S((q-1)y_o) = r_o(p)\ k(p)^{q-1}$$

for every $q = 1,2,\ldots,p$, showing that the IPP satisfies Bradford's law for p groups.

From the above proofs, formula (III.13) is automatically proved :

$$K = k(p)^{\frac{p}{A}} \qquad \text{(III.13)}$$

for every $p \in \mathbb{N}$.

Proof of the equivalence of (iii) and (v)

The general relation between the Leimkuhler function R and the Bradford function σ is, for $i = R(r)$:

$$r = R^{-1}(i) = \int_0^i \sigma(i')\, di' \qquad \text{(III.48)}$$

for every $i \in [0,A]$.

a) (iii) implies (v)

Given that

$$R(r) = a \log (1 + br) \qquad \text{(III.4)}$$

for every $r \in [0,T]$, where a and b are constants, we find with (III.48) :

$$\sigma(i) = \frac{dR^{-1}(i)}{di} \qquad \text{(III.49)}$$

while

$$R^{-1}(i) = \frac{1}{b}\left(e^{\frac{i}{a}} - 1\right) = r \qquad \text{(III.50)}$$

(III.49) and (III.50) yield now

$$\sigma(i) = \frac{1}{ab}\, e^{\frac{i}{a}} = M.K^{i}\ , \qquad \text{(III.51)}$$

where

$$M = \frac{1}{ab} \text{ and } K = e^{\frac{1}{a}} \qquad \text{(III.52)}$$

b) <u>(v) implies (iii)</u>
Given that

$$\sigma(i) = M.K^{i} \qquad \text{(III.7)}$$

for every $i \in [0,A]$, where M and $K > 1$ are constants, we have by (III.48) that, with $i = R(r)$

$$r = R^{-1}(i) = \int_0^{R(r)} M.K^{i'}\, di'$$

$$r = \frac{M}{\log K}\left(K^{R(r)} - 1\right) \qquad \text{(III.53)}$$

Hence

$$R(r) = \frac{1}{\log K} \log\left(1 + r\,\frac{\log K}{M}\right)\ ,$$

which is of the form

$$R(r) = a \log (1 + br)$$

with

$$a = \frac{1}{\log K} \quad \text{and} \quad b = \frac{\log K}{M} \qquad \text{(III.54)}$$

(cf. also formulae (II.40)), which is in accordance with formulae (III.52). This shows also the second equalities in (III.9) and (III.10).

This completes the proof of this theorem. □

Remark :
The reader will have noticed that the proof of one of the above equivalencies is superfluous. The longest of these is the proof that (iii) is equivalent with (iv). However, this is the classic statement - seen often in the literature (explicitely or implicitely mentioned) - that the classical law of Bradford is equivalent with the law of Leimkuhler. We therefore provided a direct proof for it.

Corollary 1 :
If the continuous IPP satisfies Bradford's law for p groups ($p \in \mathbb{N}$), then the Bradford factor $k = k(p)$ has the value

$$k = \rho(A)^{\frac{1}{p}} \quad . \tag{III.55}$$

Proof :
Using (III.15) (valid for a fixed but arbitrary $p \in \mathbb{N}$) we see that

$$k = e^{\frac{A}{pC}} \tag{III.56}$$

But, using (II.20) we have

$$A = \int_1^{\rho(A)} j\, f(j)\, dj$$

$$A = \int_1^{\rho(A)} \frac{C}{j}\, dj$$

$$A = C \log \rho(A). \tag{III.57}$$

(III.56) and (III.57) now yield

$$k = \rho(A)^{\frac{1}{p}} \quad . \quad \square$$

This formula will be slightly adapted to discrete practical bibliographies, when fitting them to Bradford's law.

Corollary 2 :
If the continuous IPP satisfies Bradford's function group-free, then the continuous Bradford factor K has the value

$$K = \rho(A)^{\frac{1}{A}} \quad . \qquad \text{(III.58)}$$

Proof :
This follows readily from formulae (III.13) and (III.55); formula (III.55) can be used since, in the above theorem, (v) implies (iv). □

III.1.3. Functions equivalent to the general Lotka function

We have the following theorem (see partially (Egghe, 1985 and 1989b)) :

Theorem :
Let (S,I,V) be any continuous IPP. Then we have the following equivalencies :

(i) The IPP satisfies the Lotka function (III.1) (general α).
(ii) The IPP satisfies Mandelbrot's function (III.3) (general β').
(iii) The IPP satisfies the general Leimkuhler function (III.5).
(iv) The IPP satisfies the general group-free Bradford function (III.8).

Note :
Relations between the parameters can be proved as in the previous theorem but we omit them since we do not need them further on; this also simplifies the arguments.

Proof : Proof of the equivalence of (i) and (ii)

We use again the general relation (III.17) :

$$g^{-1}(j) = r(j) = \int_{j}^{\rho(A)} f(j')\, dj'$$

for $j \in [1, \rho(A)]$ (see previous section).

a) (i) implies (ii)

$$g^{-1}(j) = r(j) = \int_{j}^{\rho(A)} \frac{C}{j'^{\alpha}} \, dj'$$

$$= \frac{C}{1-\alpha} \left(\rho(A)^{1-\alpha} - j^{1-\alpha}\right)$$

from which follows that ($j = g(r)$)

$$g(r) = \frac{\rho(A)}{\left(1 + r \frac{\alpha - 1}{C \, \rho(A)^{1-\alpha}}\right)^{\frac{1}{\alpha-1}}} \qquad \text{(III.59)}$$

Hence

$$g(r) = \frac{G}{(1 + Hr)^{\beta'}} \qquad \text{(III.3)}$$

where $G = \rho(A)$, $H = \frac{\alpha - 1}{C \, \rho(A)^{1-\alpha}}$ and $\beta' = \frac{1}{\alpha - 1}$.

b) (ii) implies (i)

From

$$j = g(r) = \frac{G}{(1 + Hr)^{\beta'}}$$

one finds, using $f(j) = -r'(j)$ (as in (III.20)) that

$$f(j) = -\frac{1}{\beta' H} \left(\frac{G^{\frac{1}{\beta'}}}{j}\right)^{\frac{1+\beta}{\beta}} , \qquad \text{(III.60)}$$

being Lotka's law.

Proof of the equivalence of (i) and (iv)

a) (i) implies (iv)

This is proved in section II.4.3.2.

b) (iv) implies (i)

From Bradfords'law

$$\sigma(i) = (A_1 + iA_2)^{A_3}$$

for $i \in I$, and lemma II.2.1.1 one has that

$$\rho(i) = (A_1 + AA_2 - iA_2)^{-A_3} \qquad \text{(III.61)}$$

Consequently

$$\rho'(i) = A_2A_3(A_1 + AA_2 - iA_2)^{-A_3-1}$$

$$\rho'(i) = A_2A_3\ \rho(i)^{1+\frac{1}{A_3}} \qquad \text{(III.62)}$$

for every $i \in I$.

Using the duality relation (II.20) we see that

$$\int_1^{\rho(i)} f(j)\ j\ dj = i$$

and hence

$$f(\rho(i))\ \rho(i)\ \rho'(i) = 1 \qquad \text{(III.63)}$$

for every $i \in I$. Combining (III.62) and (III.63) we have

$$f(\rho(i)) = \frac{1}{A_2A_3\ \rho(i)^{2+\frac{1}{A_3}}} \qquad \text{(III.64)}$$

for every $i \in I$.

Hence also

$$f(\rho(A-i)) = \frac{1}{A_2A_3\ \rho(A-i)^{2+\frac{1}{A_3}}}$$

for every $i \in I$, since this is exactly the same as (III.64). Now $j = \rho(A-i)$ (corollary II.2.1.2; see also the previous section). Hence

$$f(j) = \frac{1}{A_2 A_3 \, j^{2 + \frac{1}{A_3}}}$$

for every $j \in [1, \rho(A)]$, which is the general Lotka function (III.1).

Proof of the equivalence of (iii) and (iv)

a) (iii) implies (iv)

Formula (III.5), written in general form (independent of Lotka's α), reads as

$$R(r) = B_1 (B_2 - (B_3 + B_4 r)^{B_5}) \qquad \text{(III.65)}$$

Now, for $i = R(r)$, we have

$$r = \int_0^i \sigma(i') \, di' \qquad \text{(III.66)}$$

(cf. (II.38)). Hence, combining (III.65) and (III.66) gives

$$i = B_1 (B_2 - (B_3 + B_4 \int_0^i \sigma(i') \, di')^{B_5}) \qquad \text{(III.67)}$$

This yields

$$\int_0^i \sigma(i') \, di' = \frac{(B_2 - \frac{i}{B_1})^{\frac{1}{B_5}} - B_3}{B_4}$$

and hence, differentiating :

$$\sigma(i) = - \frac{1}{B_1 B_4 B_5} (B_2 - \frac{i}{B_1})^{\frac{1}{B_5} - 1} \qquad \text{(III.68)}$$

which is of the form of the generalised group-free law of Bradford.

b) (iv) implies (iii)
This was shown in section II.4.3.2. □

Note :
In quantitive linguistics, Mandelbrot (1974) pointed out that $\frac{1}{\beta'}$ equals the fractal dimension D of the IPP (in this case, a text). Whether this is true (and how to understand this) in general IPP's is not clear for the moment. From the above proof it follows that $\alpha - 1 = D$. Suppose for the moment that $\frac{1}{\beta'} = \alpha - 1$ is the fractal dimension of a general IPP, then it is interesting to see, due to the results of section II.4.2 that $D < \frac{C}{A} + 1$ (cf. (II.31)) and most commonly $D < 2$ (cf. the note after corollary II.4.2.3). This result looks quite natural, an IPP being studied in a "2-dimensional" (dual) framework. In this connection also the assumption $\alpha > 1$ (i.e. $D > 0$) is clear.

So far, all calculations and theoretical developments are exact in the sense that they can be proved or worked out mathematically, with no approximations whatever. It must however be emphasized that - no matter how valuable continuous IPP's are for developing a theory - practical bibliographies are discrete but large. Results as above would never have been possible to be proved for discrete IPP's. But, in order to be able to apply the above results to practical bibliographies some approximations are in order. They are formulated in the next paragraph.

III.2. Informetric approximations

This section deals with the use of the symbol ≈ in further calculations. It is not easy to state "axiomatically" what is allowed and what is not, concerning approximations in informetrics in general and in informetric laws in special.

Basically, approximations are needed to cope with the fact that the above theory is for continuous IPP's, while

practical bibliographies are not : they are discrete but large. We therefore adopt the following acceptable principles :

(A_1) We may work with discrete sums, whenever we have been working with integrals in the continuous theory above. The reason why we have not done so from the beginning is technical and also for reasons of theoretical elegance. Some results even would have been impossible to prove in the discrete setting.

(A_2) $\rho(A)$, the maximal density of items, can be put equal to the number of items in the most productive source, on condition that there is only one such source. This quantity is henceforth denoted by

$$y_m = \rho(A) \qquad \text{(III.69)}$$

(A_3) y_m is large, in the absolute sense (i.e. when not in combination with other parameters).

All these principles do agree with all practical (i.e. not too small) bibliographies.

III.3. Relations between parameters of the classical informetric functions

We have adopted unique notations for the parameters that occur in the informetric functions, studied so far. So we do not repeat their meaning : they can be found in paragraph III.1. Let us just repeat Lotka's law :

$$f(j) = \frac{C}{j^{\alpha}} \qquad \text{(III.70)}$$

C and α are constants, $\alpha > 1$ and $j \in [1,\rho(A)] = [\rho(0),\rho(A)] = [1,y_m]$ (see (III.1) and (III.69)).

Most of our attention will be devoted to the very classical case $\alpha = 2$ since much is still to be done in this case. Nevertheless, we will also consider general $\alpha > 1$, subsequently.

III.3.1. The case $\alpha = 2$

We draw the reader's attention to the results in section III.1.2 and repeat the formulae that were obtained there (together with (III.69)) :

$$a = \frac{y_o}{\log k} \qquad (F_1)$$

$$b = \frac{k - 1}{r_o} \qquad (F_2)$$

$$G = y_m = ab \qquad (F_3)$$

$$H = \frac{y_m}{C} = b \qquad (F_4)$$

$$C = a \qquad (F_5)$$

$$y_o = C \log k \qquad (F_6)$$

$$r_o = \frac{C}{y_m} (k-1) \qquad (F_7)$$

$$K = k(p)^{\frac{p}{A}} \qquad (F_8)$$

We have also, following from the group-dependent Bradford formulation, that (for $p \in \mathbb{N}$ groups) :

$$A = y_o . p$$

and

$$T = r_o + r_o k + \ldots + r_o k^{p-1}$$

Hence

$$y_o = \frac{A}{p} \qquad (F_9)$$

and

$$r_o = \frac{T(k-1)}{k^p - 1} \qquad (F_{10})$$

This, together with (F_1) and (F_2) gives

$$a = \frac{\frac{A}{p}}{\log k}$$

$$a = \frac{1}{\log k^{p/A}}$$

$$a = \frac{1}{\log K} \qquad (F_{11})$$

and

$$b = \frac{\log K}{M} \qquad (F_{12})$$

(see formulae(II.40) or (III.9) and (III.10)).

That (F_{12}) also follows from (F_2) can be shown thus:

$$b = \frac{k - 1}{r_o}$$

$$= \frac{\frac{1}{A} \log k^p}{\frac{r_o}{k-1} \frac{\log k^p}{A}}$$

$$= \frac{\log K}{M} \quad ,$$

using formula (III.44). Hence all formulas (II.40), (III.9), (III.10) and (III.44) are in complete accord with each other.

Adapting the proof of corollary 1 in section III.1.2 to the discrete case we have, now using (cf. (A_1), (A_2) and (A_3)) :

$$\sum_{j=1}^{y_m} \frac{C}{j} \approx C(\log y_m + \gamma) \qquad \text{(III.71)}$$

(where γ is Euler's number, $\gamma \approx 0.5772...$), the following result :

$$k \approx (e^{\gamma} y_m)^{\frac{1}{p}} . \qquad (F_{13})$$

(F_8) and (F_{13}) imply

$$K \approx (e^{\gamma} y_m)^{\frac{1}{A}} \qquad (F_{14})$$

(see also (Egghe, 1986) and (Egghe, 1989c)).

A and T, the total number of items resp. of sources can be related to the above parameters as follows

$$A \approx \sum_{j=1}^{y_m} j \frac{C}{j^2}$$

(cf. (A_1), (A_2), (A_3)). Hence, since y_m is large (A_3) we have (see also (III.71)) :

$$A \approx C (\log y_m + \gamma) \qquad (F_{15})$$

and

$$T \approx \sum_{j=1}^{y_m} \frac{C}{j^2} .$$

Since $\sum_{j=1}^{\infty} \frac{1}{j^2}$ converges and by (A_3) we have

$$T \approx \sum_{j=1}^{\infty} \frac{C}{j^2} = C \frac{\pi^2}{6}$$

So

$$T \approx C \frac{\pi^2}{6} \qquad (F_{16})$$

Formulae (F_1) through (F_{12}) are not adaptable in this way (or do not need an adaptation!); so they are left as they are. The above set of formulae will be shown to be useful (f.i. for fitting purposes) in the sequel.

Note that the parameter C in Lotka's law is determined by formula (F_{16}), in the sense that C can be determined from the practical data, being

$$C \approx \frac{6}{\pi^2} T \qquad (III.72)$$

Note again that $C < T < A$ (cf. the note after corollary II.4.2.3).

If there is a need to express C in function of y_m, this has been done in (Allison et al., 1976), in (Egghe and Rousseau, 1986) and in (Egghe, 1987) in connection with Price's law. Following (Allison et al., 1976) there are at least two different ways of expressing the value of C in function of y_m :

(a) Expressing the fact that the (unique) most productive source stands for the "mathematical tail" of the function $f(j) = \frac{C}{j^2}$. Hence we then have that

$$\sum_{j=y_m}^{\infty} \frac{C}{j^2} \approx 1 \qquad (III.73)$$

Hence

$$C \approx \frac{1}{\sum_{j=y_m}^{\infty} \frac{1}{j^2}} \approx \frac{1}{\int_{y_m}^{\infty} \frac{dj}{j^2}} = y_m$$

$\left(\sum_{j=y_m}^{\infty} \frac{1}{j^2} \approx \int_{y_m}^{\infty} \frac{dj}{j^2} \right.$ since $\left. \sum_{j=1}^{\infty} \frac{1}{j^2} \right.$ converges and by (A_3)).

So

$$C \approx y_m \qquad \text{(III.74)}$$

(see section III.4 for more information on this case).

(b) Expressing the fact that the unique most productive source stands for itself and hence

$$1 = f(y_m) = \frac{C}{y_m^2} \qquad \text{(III.75)}$$

Consequently

$$C = y_m^2 \qquad \text{(III.76)}$$

Practical examples show that almost always $C \approx \frac{6}{\pi^2} T \in [y_m, y_m^2]$, which is not much of a property since the interval $[y_m, y_m^2]$ is indeed very wide! We refer the reader to the examples studied in the next chapter, to see that the above assertion is indeed true.

Therefore, as in (Allison et al., 1976) we can put

$$C = y_m^c \ , \qquad \text{(III.77)}$$

where $c \in [1,2]$. In this case we have the following formulae (indicated with PF since they are assuming the practical relation (III.77)) :

$$H = y_m^{1-c} \qquad (PF_1)$$

$$a = y_m^c \qquad (PF_2)$$

$$y_o = y_m^c \log k \qquad (PF_3)$$

$$r_o = y_m^{c-1} (k-1) \qquad (PF_4)$$

$$A \approx y_m^c \ (\log y_m + \gamma) \qquad (PF_5)$$

$$T \approx y_m^c \ \frac{\pi^2}{6} \qquad (PF_6)$$

III.3.2. The general case

In section III.1.3 we can read the following formulae (using (III.69) of course) (denoting GF for "General Formula") :

$$G = y_m \qquad (GF_1)$$

$$H = \frac{\alpha - 1}{C \ y_m^{1-\alpha}} \qquad (GF_2)$$

$$\beta' = \frac{1}{\alpha - 1} \qquad (GF_3)$$

and furthermore (see II.4.3.2), with

$$\sigma(i) = (A_1 + iA_2)^{A_3} \qquad (II.44)$$

and

$$\rho(i) = (1 - iA_2)^{-A_3} \qquad (II.43)$$

one has :

$$A_1 = \rho(A) = y_m = \frac{A(2-\alpha)}{C} + 1 \qquad (GF_4)$$

$$A_2 = - \frac{2 - \alpha}{C} \qquad (GF_5)$$

$$A_3 = - \frac{1}{2 - \alpha} \qquad (GF_6)$$

Also, for

$$R(r) = B_1 \ (B_2 - (B_3 + B_4 r)^{B_5}) \qquad (III.65)$$

one has (see also II.4.3.2) :

$$B_1 = \frac{C}{2-\alpha} \qquad (GF_7)$$

$$B_2 = y_m^{2-\alpha} \qquad (GF_8)$$

$$B_3 = y_m^{1-\alpha} \qquad (GF_9)$$

$$B_4 = \frac{\alpha - 1}{C} \qquad (GF_{10})$$

$$B_5 = \frac{2-\alpha}{1-\alpha} \qquad (GF_{11})$$

Of course, obvious interrelations between the A_i and B_j can be made.

For the rest of this section, we restrict ourselves to the case $\alpha = 2$, hence purely Bradfordian situations (cf. section III.1.2).

Firstly we investigate the relationship between the classical Bradford factor $k = k(p)$ and the average production μ (often claimed to be equal to k - wrong of course, but we will shed some new light on this claim).

Secondly, also in the case $\alpha = 2$, we will prove an explicit formula for $m(i)$ $(i = 1,\ldots,p)$, where $m(i)$ is the number of items in the most productive source in the $(p-i+1)^{th}$-Bradford group. Here we use duality aspects in discrete IPP's. This will have applications in fitting procedures, to be dealt with in the next chapter.

III.3.3. k as a function of μ

In (Goffman and Warren, 1969) as well as in (Yablonsky, 1980), the relationship between k and the average production is considered. In both publications

one considers a bibliography of journals (sources) containing papers (items). In the former publication one deals with the average number μ of papers per journal. In the latter however one uses the average production η of papers per author and they estimate $k \approx \eta$. We want to stress here that no k can be equal to any average (μ nor η), because of the p-dependence (see e.g. formula (F_{13})).

The relation between k and η is part of 3-dimensional informetrics (cf. I.7) and is not dealt with here. The relation between k and μ is as follows (see (Egghe, 1989c)) :

$$\mu = \frac{\sum_{j=1}^{y_m} j\, f(j)}{\sum_{j=1}^{y_m} f(j)} \qquad \text{(III.78)}$$

$$= \frac{\sum_{j=1}^{y_m} \frac{1}{j}}{\sum_{j=1}^{y_m} \frac{1}{j^2}}$$

$$\approx \frac{\log y_m + \gamma}{\frac{\pi^2}{6}}$$

Here we used (A_1), (A_2) and (A_3), together with the fact that $\sum_{j=1}^{\infty} \frac{1}{j^2}$ converges. So

$$\mu \approx \frac{6}{\pi^2} (\log y_m + \gamma) \qquad \text{(III.79)}$$

Using (F_{13}) we hence see that

$$\mu \approx \frac{6}{\pi^2} p \log k \qquad \text{(III.80)}$$

The graph of the function

$$f_p(k) = \frac{k}{\mu}$$

$$f_p(k) = \frac{\pi^2 k}{6\ p\ \log\ k} \qquad \text{(III.81)}$$

looks like the graph of Fig.III.1 (the exact form being, of course, dependent on p). We have the properties, for every $p \in \mathbb{N}$:

$$\lim_{\substack{k \to 0 \\ >}} f_p(k) = 0\ , \qquad \lim_{\substack{k \to 1 \\ <}} f_p(k) = -\infty \qquad \text{(III.82)}$$

(unimportant since in informetrics, $k > 1$)

$$\lim_{k \to +\infty} f_p(k) = +\infty \qquad \text{(III.83)}$$

$$\lim_{\substack{k \to 1 \\ >}} f_p(k) = +\infty \qquad \text{(III.84)}$$

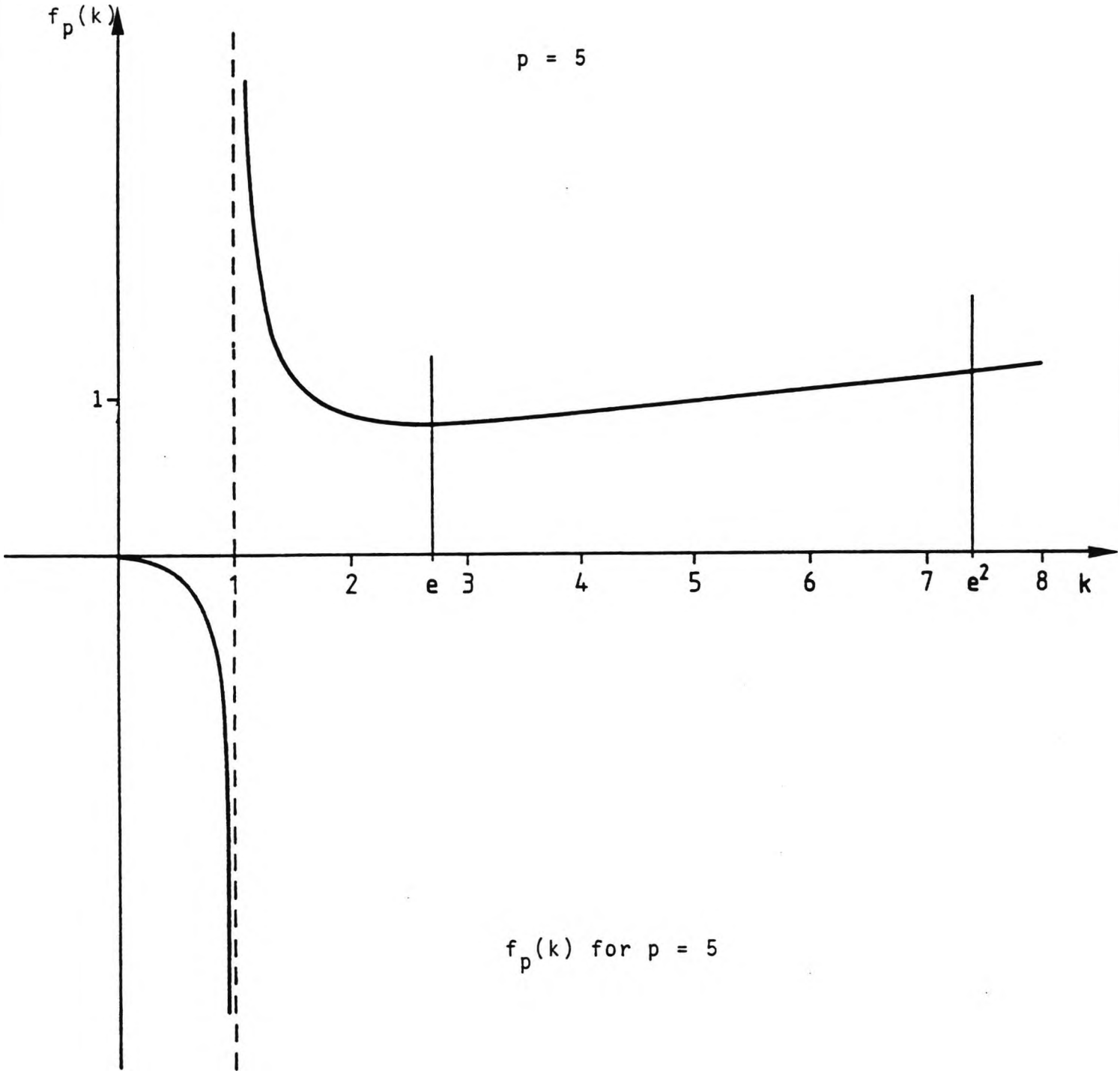
(a vertical asymptote in k=1) and

$$f_p(2) = f_p(4) = \frac{\pi^2}{3\ p\ \log\ 2} \qquad \text{(III.85)}$$

It is remarkable however that for a wide range of $k > 1$ the function f_p is almost horizontal, hence there is an approximately constant relationship between k and μ! The minimum is obtained in $k = e$ (independent of p) and there is an inflection in $k = e^2$. So in a wide range (see Fig.III.1) around $k = e$ we have an almost constant relation between μ and k :

$$f_p(e) = \frac{\pi^2 e}{6\ p} \qquad \text{(III.86)}$$

Hence, for $p \geq 5$ we always have that $k < \mu$ for a wide range of k's around $k = e$ while k might be $> \mu$ for



$f_p(k)$ for p = 5

Fig.III.1 : Graph of $\frac{k}{\mu}$ for p = 5

p = 3 or 4 or for p ≥ 5 and k close to 1 or k very large (not common in practice).

We have checked our theoretical findings, using some classical bibliographies : Applied Geophysics ((Bradford, 1934) or (Egghe, 1989d)), Lubrication ((Bradford, 1934) or (Egghe, 1989d)), ORSA ((Kendall, 1960) or (Egghe, 1989d)), Mast Cell ((Seley, 1968) or (Egghe, 1989d)), Schistosomiasis ((Warren and Newill, 1967) or (Egghe, 1989d)), Circulation data (Goffman and Morris, 1970), User's data (Goffman and Morris, 1970), Transplantation - Immunology (Goffman and Morris, 1970) and finally 6 bibliographies in (Aiyepeku, 1977) : Geography, USA-UK-France-Germany-data, USA-data, UK-data, France-data and Germany-data.

We have found a complete confirmation of the above results. See Table III.1.

Note 1 :
The Bradford data of Goffman and Warren on Mast Cell and Schistosomiasis (Table 3 in (Goffman and Warren, 1969)) are omitted for reasons to be given later.

Note 2 :
The reader can verify that Table III.1 is correct by simply checking the values of p, k and μ on the raw data. A method of calculating these values is given in chapter IV. Furthermore, there we can also find the calculations for k for the first 5 bibliographies (Applied Geophysics (p = 3 or p = 5), Lubrication (p = 3 or p = 7), ORSA (p = 4), Mast Cell (p = 13) and Schistosomiasis (p = 9). The calculations of the other values is carried out in exactly the same way.

bibliography	p	k	μ	confirmation
Applied Geophysics	3	5.49	4.09	Y : k > μ
Idem	5	2.78	4.09	Y : k < μ
Lubrication	3	3.40	2.41	Y : k > μ
Idem	7	1.69	2.41	Y : k < μ
ORSA	4	4.56	4.76	Y : k ≈ μ
Mast Cell	13	1.44	4.05	Y : k < μ
Schistosomiasis	9	2.03	5.70	Y : k < μ
Circulation data	8	1.4	2.36	Y : k < μ
User's data	8	1.4	2.22	Y : k < μ
Transplantation-Immunology	9	1.8	4.12	Y : k < μ
Geography	7	1.7	13.27	Y : k < μ
USA, UK, France, Germany data	6	1.8	5.50	Y : k < μ
USA data	5	2.3	4.05	Y : k < μ
UK data	7	2.4	4.23	Y : k < μ
France data	4	2.8	4.48	Y : k < μ
Germany data	6	1.9	3.20	Y : k < μ

Table III.1 : Verification of the relation between k and μ

III.3.4. The number of items in the most productive source in every Bradford group

Let us consider the case $\alpha = 2$ once more. We have here a pure Bradford IPP again. Consider the p Bradford groups ($p \in \mathbb{N}$ fixed but arbitrary). We might wonder how these groups are structured. For instance, where do the divisions (between one group and the following) occur?. Let us visualise the p groups as in Fig.III.2, numbering

Fig.III.2 : The Bradford groups

them from right to left (i.e. starting with the least productive sources).

This approach is identical to the dual approach for discrete IPP's in section II.2.2. Furthermore, in theorem II.3.2.3 we showed that the function $\rho(i)$, measuring the average number of items per source in group i (from right to left as above) is the same function (upon a constant) as the Bradford function $\sigma(i)$, i.e. an exponential one.

Analogous to the above but far more intricate is finding the exact place of the "cutting points" 1,2,...,p-1, in Fig.III.2 above : what sources are there and what is their production? As in chapter II, our dual approach will yield a solution, but it is more intricate to solve this problem. The problem itself is interesting and its solution will be applied in chapter IV.

Let i = 1,2,...,p. Denote by m(i) the number of items in the most productive source in group i (counted in the dual sense : from right to left). We suppose that we have a large discrete IPP for which Lotka's law

$$f(j) = \frac{C}{j^2} \qquad \text{(III.87)}$$

$j = 1,2,\ldots,y_m$, is valid (discrete law).

The cutting points i are such that every group has the same number of items (since they are the divisions of the Bradford groups). So, in general, these divisions do not coincide with some divisions $j = 1,2,\ldots,y_m$ in Lotka's law. Since, by definition, m(i) is the number of items in the most productive source in group i, this group ends in the Lotka category j = m(i) but, maybe, not all the sources with production m(i) are included or, what is the same, not all the items in all the sources with production m(i) are included. Let $\alpha(i)$ denote the fraction <u>of the items</u>, belonging to sources with production m(i), that belong to the i^{th} group. Hence $\alpha(i) \in]0,1]$ ($\alpha(i) \neq 0$, by definition of m(i)).

Lemma (Egghe, 1986) :

$$\alpha(i) = m(i)(i \log k - \sum_{j=1}^{m(i)-1} \frac{1}{j}) \qquad \text{(III.88)}$$

for every $i = 1,2,\ldots,p$.

Proof :
We will prove this formula by mathematical induction.

(a) $i = 1$

Since in every group (hence also in the first one) we have $y_o = \frac{A}{p}$ items and since $y_o = C \log k$ (formula (F_6)), we have, by definition of $\alpha(1)$, $m(1)$ and by (III.87) :

$$C \log k = y_o = C + \frac{C}{2^2}.2 + \frac{C}{3^2}.3 + \ldots + \frac{C}{(m(1)-1)^2}(m(1)-1)$$

$$+ \frac{C}{m(1)^2} m(1).\alpha(1) \qquad \text{(III.89)}$$

Hence

$$\log k = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{m(1)-1} + \frac{1}{m(1)} \alpha(1)$$

So

$$\alpha(1) = m(1)(\log k - \sum_{j=1}^{m(1)-1} \frac{1}{j}) \qquad \text{(III.90)}$$

showing (III.88) for $i = 1$.

(b) Given (III.88) for i, show (III.88) for i replaced by i+1

Now we have that the y_o items in group i+1 are composed of the items

- in the sources with production m(i) that are not in group i
- in the sources with production m(i)+1, m(i)+2,...,m(i+1)-1
- in the sources with production m(i+1), but only a fraction (of the items involved) $\alpha(i+1)$.

Expressed mathematically this gives

$$C \log k = y_o = \frac{C}{m(i)^2}\, m(i) - \frac{C}{m(i)^2}\, m(i)\, \alpha(i) + \frac{1}{(m(i)+1)^2}\,(m(i)+1)$$

$$+ \ldots + \frac{C}{(m(i+1)-1)^2}\,(m(i+1)-1) + \frac{C}{m(i+1)^2}\, m(i+1)\, \alpha(i+1) \qquad \text{(III.91)}$$

Using (III.88) for i, and putting this in (III.91) gives :

$$\log k = \frac{1}{m(i)} - \frac{1}{m(i)}\left[m(i)\left(i \log k - \sum_{j=1}^{m(i)-1} \frac{1}{j}\right)\right]$$

$$+ \frac{1}{m(i)+1} + \ldots + \frac{1}{m(i+1)-1} + \frac{1}{m(i+1)}\, \alpha(i+1)$$

$$= \sum_{j=1}^{m(i+1)-1} \frac{1}{j} - i \log k + \frac{\alpha(i+1)}{m(i+1)}$$

So this yields

$$\alpha(i+1) = m(i+1)\left[(i+1) \log k - \sum_{j=1}^{m(i+1)-1} \frac{1}{j}\right] .$$

(a) and (b) together show (III.88) for every i = 1,2,...,p. □

With the help of this lemma we can now prove the following useful result.

Theorem (Egghe, 1986) :
For every i = 1,2,...,p we have that

$$\frac{1}{i} \sum_{j=1}^{m(i)-1} \frac{1}{j} < \log k < \frac{1}{i} \sum_{j=1}^{m(i)} \frac{1}{j} \qquad \text{(III.92)}$$

and hence, for m(i) not too small,

$$m(i) \approx \frac{k^i}{e^\gamma} \approx 0.5615\, k^i \qquad \text{(III.93)}$$

Proof :

Group i has, by definition, sources with production

1. m(i-1); the fraction w.r.t. the items in the sources with production m(i-1) being $1 - \alpha(i-1) \in [0,1[$,
2. $m(i-1) + 1,\ldots,m(i) - 1$,
3. m(i); the fraction w.r.t. the items in the sources with production m(i) being $\alpha(i) \in]0,1]$.

Consequently, since there are $y_o = C \log k$ (F_6) items in every group, we can write the following equality (cf. also (III.91) with i replaced by i-1) :

$$C \log k = y_o = \frac{C}{m(i-1)} (1-\alpha(i-1)) + \frac{C}{m(i-1)+1} + \ldots + \frac{C}{m(i)-1} + \frac{C}{m(i)} \alpha(i) \qquad \text{(III.94)}$$

(III.88) together with (III.94) now yield :

$$\log k = \frac{1}{m(i-1)} [1 - m(i-1)((i-1) \log k - \sum_{j=1}^{m(i-1)-1} \frac{1}{j})] + \frac{1}{m(i-1)+1} + \ldots + \frac{1}{m(i)-1} + \frac{1}{m(i)} \alpha(i) .$$

Hence

$$\frac{\alpha(i)}{m(i)} = i \log k - \sum_{j=1}^{m(i)-1} \frac{1}{j} \qquad \text{(III.95)}$$

Since $\alpha(i) \in]0,1]$ we hence find (III.92), for every $i = 1,2,\ldots,p$.

If m(i) is not too small, we can use the following approximation :

$$\sum_{j=1}^{x} \frac{1}{j} \approx \log x + \gamma \qquad \text{(III.96)}$$

for x high.

This gives in (III.92), approximately :

$$\log (m(i)-1) + \gamma < \log k^i < \log m(i) + \gamma$$

Hence, since m(i) is large,

$$\log k^i - \gamma \approx \log m(i)$$

Finally we find

$$m(i) \approx \frac{k^i}{e^\gamma} \approx 0.5615\ k^i \ . \quad \square \qquad \text{(III.93)}$$

Corollary :

$$k \approx (e^\gamma\ y_m)^{\frac{1}{p}} \qquad (F_{13})$$

where k is the Bradford factor related to the division in p groups.

Proof :
Certainly $m(p) = y_m$ is high, supposing (A_3). Hence

$$y_m = m(p) \approx \frac{k^p}{e^\gamma} \ ,$$

from which (F_{13}) follows. $\square$

Note that the above proof is a second proof of this fact, the first one being given in section III.3.1, based on the results of section III.1.2. The first proof resulted from the relations that exist with other informetric laws; the second one is a proof completely within the Bradford framework. Neither proof is completely trivial.

The above results on m(i) will be re-used in the next chapter on fittings of Bradford's and Leimkuhler's law.

Note also that, in order to have formula (III.93) accurately, not i but only m(i) must be large. We have the following table :

n	log n	$\sum_{k=1}^{n} \frac{1}{k} - \gamma$	Δ
1	0	0.42	0.42
2	0.69	0.92	0.23
3	1.1	1.25	0.15
4	1.39	1.50	0.11
5	1.61	1.70	0.09
6	1.79	1.87	0.08
7	1.95	2.01	0.06
8	2.08	2.14	0.06
9	2.2	2.25	0.05
10	2.3	2.35	0.05

Table III.2 : $\sum_{i=1}^{n} \frac{1}{k}$ versus log n

So, already from m(i) $\geq$ 7 on, the difference is less than 3 % of the actual value, which can be reached already in the second or third Bradford group (counted in the dual way), in practical situations.

III.4. Further comments on the classification of certain informetric functions

It is not our intention to try to classify all kinds of informetric functions. Nevertheless it is not easy to see the exact place of some classical laws such as Pareto, Zipf, the graphical formulation of Bradford's law (see further) and Brookes' law (also called the Weber-Fechner law - see also further).

Without claiming the complete study of this problem we will present some ideas (both mathematical and philosophical) leading to the conclusion that the above mentioned laws together :

(1) are equivalent

(2) are part of our theory on continuous IPP's satisfying Lotka's law

but

(3) play also a separate role and this is found to be true in several ways, to be explained later on.

In order to be able to prove a correct result, all classical definitions are slightly modified in the sense that all ranks are lowered by 1; otherwise no relation between the laws will exist.

III.4.1. Definitions

For the sake of simplicity and unity we will always assume a continuous IPP (S,I,V) as in chapter II.

A. The graphical formulation of Bradford's law, group-dependent

Fix $p \in \mathbb{N}$. We say that our IPP satisfies the graphical formulation of Bradford's law (p-dependent) if we can divide the set I into p equal parts, each containing $y_0 > 0$ items such that, in (I,S,U), we have, for the first y_0 items, the first $r_1 - 1 > 0$ sources, for the first 2 y_0 items, the first $r_1 k_1 - 1$ sources ($k_1 > 1$), for the first 3 y_0 items, the first $r_1 k_1^2 - 1$ sources, and so on until : for the first (p-1) y_0 items, the first $r_1 k_1^{p-2} - 1$ sources and finally, the p y_0 = A items stand for $r_1 k_1^{p-1} - 1$ = T sources (see e.g. (Wilkinson, 1973), where the ranks are 1 higher).

B. The graphical formulation of Bradford's law, group-free

Let the function $\Sigma(i)$ denote the cumulative number of sources up to the coordinate $i \in I = [0,A]$ in (I,S,U). Then

$$\Sigma(i) = M_1 . K_1^i - 1 \qquad \text{(III.94)}$$

where M_1 and $K_1 > 1$ are constants.
Σ is called the group-free graphical Bradford function.
Note that in general IPP's :

$$\Sigma(i) = \int_0^i \sigma(i')\, di'$$

$$= r$$

$$= U(i) . \qquad \text{(III.95)}$$

C. Brookes' law or the law of Weber-Fechner

Let, in (I,S,U), $R_1(r)$ denote the cumulative number of items in the sources $s \in [0,r]$, for every $r \in [0,T]$. Then

$$R_1(r) = \alpha \log (\beta(1+r)) , \qquad \text{(III.96)}$$

where α and β are positive constants. R_1 is the corresponding Brookes (or Weber-Fechner) function.

D. Zipf's law or Pareto's law

In (I,S,U), let $g(r)$ denote the density of the numbers of items in $r \in [0,T]$. Then, for every $r \in [0,T]$

$$g(r) = \frac{F}{1+r} \qquad \text{(III.97)}$$

(cf. also (I.5) or (I.8)), where F is a constant (we restrict our attention to the power 1 in the denominator of (I.5) or (I.8)). g is called the Zipf (or Pareto) function.

Note :
In (III.96) as well as in (III.97) one uses 1+r instead of r. This is because these functions are not defined at 0. We could have used r (instead of 1+r) together with the interval S starting in 1 but, in view of our unified theory, we prefer to retain our framework : a continuous IPP of the form

$$(S,I,V) = ([0,T],[0,A],V) .$$

Both approaches are, however, equivalent.

Theorem III.4.2 (Egghe, 1988a) :
Let (S,I,V) be an arbitrary continuous IPP. Then the following assertions are equivalent :
(i) The IPP satisfies the graphical formulation of Bradford's law group-dependent, for every $p \in \mathbb{N}$, but with the relation $r_1 = k_1$.
(ii) The IPP satisfies the graphical group-free Bradford function, with $M_1 = 1$.
(iii) The IPP satisfies Brookes' function with $\beta = 1$.
(iv) The IPP satisfies Zipf's (or Pareto's) function.

In this case, and following the notations of III.4.1, we have the following relations between the parameters :

$$\alpha = \frac{1}{\log K_1} = F \qquad \text{(III.98)}$$

$$K_1 = k_1^{p/A} \qquad \text{(III.99)}$$

$$r_1 = K_1^{y_0} = k_1 \qquad \text{(III.100)}$$

Proof : Proof of the equivalence of (i) and (ii)

(a) (i) implies (ii)

Let first $i \in [0,A]$ be such that $i = \frac{qA}{p}$ where $q \leq p$, $q,p \in \mathbb{N}$, $q \geq 1$. By (i) we have, with p groups :

$$\Sigma(i) = r_1 k_1^{q-1} - 1 \qquad \text{(III.101)}$$

for a certain r_1, $k_1 > 1$, with $r_1 = k_1$

$$\Sigma(i) = r_1 k_1^{\frac{pi}{A}} - 1$$

$$\Sigma(i) = \frac{r_1}{k_1} (k_1^{\frac{p}{A}})^i - 1$$

$$\Sigma(i) = K_1^i - 1 \ , \qquad \text{(III.102)}$$

with

$$K_1 = k_1^{\frac{p}{A}} \ . \qquad \text{(III.99)}$$

Since Σ is continuous (cf. (III.95)) and since the function $i \to K_1^i$ is a continuous extension of Σ to $[0,A]$, we have that

$$\Sigma(i) = K_1^i - 1$$

for every $i \in [0,A]$. This is so because the set of the i's as considered above is dense in $[0,A]$.

(b) <u>(ii) implies (i)</u>

Let $p \in \mathbb{N}$ be arbitrary. Let $y_o = y_o(p) = \frac{A}{p}$ and $r_1 - 1 = r_1(p) - 1 = \Sigma(y_o) = K_1^{y_o} - 1$.

Then

$$\Sigma(2y_o) = K_1^{2y_o} - 1$$

$$= K_1^{y_o} . K_1^{y_o} - 1$$

and, more generally, for every $i = 2,\ldots,p$

$$\Sigma(iy_o) = K_1^{iy_o} - 1$$

$$= (K_1^{y_o})(K_1^{y_o})^{i-1} - 1 \qquad \text{(III.103)}$$

Hence, putting

$$r_1 = K_1^{y_o} = k_1 \qquad \text{(III.100)}$$

we have (i) for every $p \in \mathbb{N}$.

Proof of the equivalence of (ii) and (iii)

(a) (ii) implies (iii)

Since

$$\Sigma(i) = K_1^i - 1 \quad ,$$

for every $i \in [0,A]$, we have, using $i = R_1(r)$, and $\Sigma(i) = r$ (see (III.95))

$$r = K_1^{R_1(r)} - 1$$

Hence

$$R_1(r) = \frac{1}{\log K_1} \log (r+1) \quad , \qquad \text{(III.104)}$$

being Brookes' law, for every $r \in [0,T]$, but with $\beta = 1$. Here

$$\alpha = \frac{1}{\log K_1} \qquad \text{(III.98)}$$

(b) (iii) implies (ii)

Given

$$R_1(r) = \alpha \log (1+r)$$

for every $r \in [0,T]$, we have trivially :

$$R_1^{-1}(i) = r = e^{\frac{i}{\alpha}} - 1 \ .$$

Hence, using $\Sigma(i) = r$ again

$$\Sigma(i) = K_1^i - 1 \qquad \text{(III.105)}$$

for every $i \in [0,A]$. Here

$$K_1 = e^{\frac{1}{\alpha}} \qquad \text{(III.98)}$$

Proof of the equivalence of (iii) and (iv)

This proof is executed using the general formula (definition of R_1 and g) :

$$R_1(r) = \int_0^r g(r')\, dr' \qquad \text{(III.106)}$$

for every $r \in [0,T]$.

(a) (iii) implies (iv)

Since from the previous formula one also has

$$g(r) = R_1'(r) \ ,$$

we find, using $R_1(r) = \alpha \log (1+r)$,

$$g(r) = R_1'(r) = \frac{\alpha}{1+r} \ , \qquad \text{(III.107)}$$

being Zipf's law.

(b) (iv) implies (iii)

Now we have

$$R_1(r) = \int_0^r \frac{\alpha}{1+r'}\, dr'$$

$$R_1(r) = \alpha \log (1+r) \qquad \text{(III.108)}$$

Here we have the relation

$$\alpha = F \qquad \text{(III.98)}$$

This completes the proof of this theorem. □

III.4.3.

Now it is clear that the above group of laws is not completely separated from the groups we have considered earlier. Indeed, taking b = 1 or, what is the same, H = 1 in (III.12) gives us the laws we have encountered in this section. This reduces to the choice $C = y_m$ in Lotka's law (see (F_4), cf. also (III.74) and the (PF)-formulas with c = 1). This special case has already been noted (Rousseau, 1988b).

From the above, we hence also have that Zipf's law corresponds to a Bradford law (p-dependent) where

$$k = 1 + r_o \quad . \qquad \text{(III.109)}$$

Indeed, use (III.10), with b = 1.

From this we can draw the conclusion that we are dealing here with a highly concentrated situation in the sense that <u>or</u> r_o is small, which is a way of saying that (take p = 3 to fix the ideas) the core group of highly produced sources is small, <u>or</u> r_o is large but then, according to (III.109), k must be large and hence, the core group of r_o sources is nevertheless small w.r.t. the other groups $r_o k$, $r_o k^2$ and so on.

We conclude that linguistics (or econometrics) can be viewed as part of informetrics, but in practice there is a separation since

1. In most informetric examples (and we will present many in the next chapter on the fitting of the informetric laws) we have b < 1 and indeed : b ≪ 1.
2. In linguistics and econometrics one often finds b = 1.

III.4.4. Philosophical explanation for the special place of linguistical (econometrical) laws within the informetric laws

Informetric laws are mostly applied to bibliographies, or authors publishing research papers, while the law of Zipf originates from linguistics (measuring the number of times that a word occurs in a text).

In the first group there is a natural tendency in the most important sources to lower their number of items a little : the most prolific authors will not publish less important (but still publishable) work; in the same way, the most important journals in a research area will become more and more selective in accepting papers, and so on. This is not the case with the use of words in texts : the most heavily used words are words such as 'the', 'a', 'and', and so on : there is no limitation on these words for grammatical reasons! Synonyms are in use only for popular but not so heavily used words.

So this explains again why Zipf's law is a highly concentrated version of Mandelbrot's law.

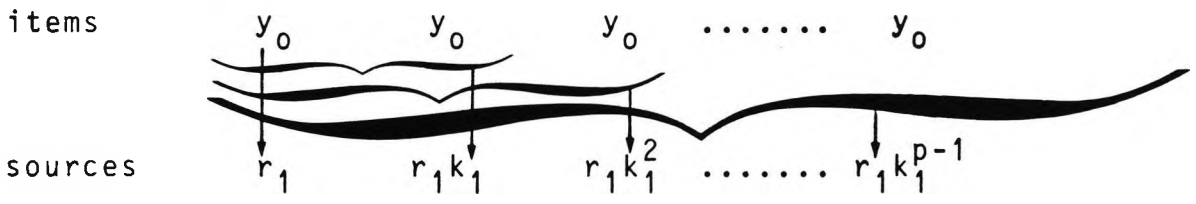
III.4.5. Solution of an apparent paradox

As a corollary of the above developments we see that, in the special case of b = 1, the graphical law of Bradford should be equivalent to the original "verbal" law of Bradford. This looks quite impossible. The verbal Bradford law for p groups yields r_o, k and $y_o = \frac{A}{p}$ such that (schematically) we have for every group :

items	y_o	y_o	y_o	.....	y_o
	↓	↓	↓		↓
sources	r_o	$r_o k$	$r_o k^2$	.....	$r_o k^{p-1}$

Situation I

Suppose this situation is also describable via the graphical law of Bradford with p groups in its classical formulation. Then each group still has y_o items. We now have r_1 and k_1 such that



Situation II

From this viewpoint, we never have the two situations occuring together. Indeed, to have both situations we need to have $r_1 = r_o$ and $r_1k_1 = r_o + r_ok$; hence

$$k_1 = 1 + k \qquad \text{(III.110)}$$

But then the third group contains

$$r_1k_1^2 = r_o\,(1+k)^2 \qquad \text{(III.111)}$$

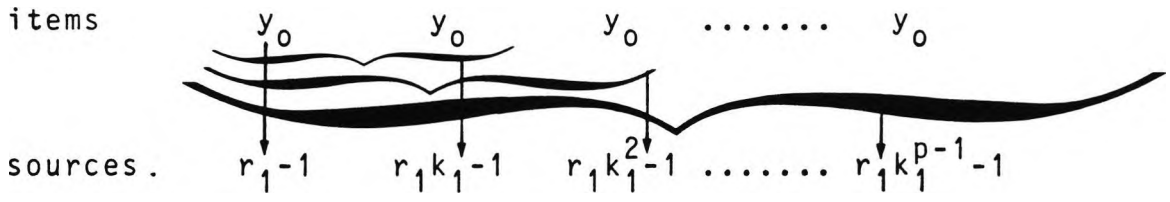
sources, while in the first case this group has r_ok^2 sources. Since both groups must be equal (since they are made that way) we conclude : The above situations are never equivalent. This is a new result.

This paradox (with theorem III.4.2) is solved as follows.

A. Group-dependent versions

To agree with our theory on both Bradford laws, we first note that theorem III.4.2 provides a way of changing situation II into situation II', which is the former situation but where all source ranks are reduced by 1.

To make this situation II' compatible with situation I it is necessary that $r_1 - 1 = r_o$, $r_1k_1 - 1 = r_o + r_ok$ and



Situation II'

$r_1k_1^2 - 1 = r_o + r_ok + r_ok^2$ (since $p \geq 3$ necessarily).
This yields

$$r_1 = 1 + r_o ,$$

$$k_1 = \frac{1 + r_o + r_ok}{1 + r_o} ,$$

$$k_1 = \frac{1 + r_o + r_ok + r_ok^2}{1 + r_o + r_ok} .$$

Equalising the last equations we find

$$k = 1 + r_o \qquad \text{(III.112)}$$

hence

$$b = 1 . \qquad \text{(III.113)}$$

Conversely, if b = 1 then it is easy to see that both situations I and II' are identical, when we take $r_1 - 1 = r_o$ and $k_1 = \frac{1 + r_o + r_ok}{1 + r_o} = k$ (since $k = 1 + r_o$). This solves our paradox in case A.

Although not strictly necessary (because of theorems III.4.2 and III.1.2) we will solve explicitly this paradox for the group-free versions :

B. Group-free versions

(1) If b = 1, then

$$\sigma(i) = M.K^i \qquad \text{(III.114)}$$

if and only if

$$\Sigma(i) = K^i - 1 \qquad \text{(III.115)}$$

for every $i \in [0,A]$, with $K > 1$ a constant.

Proof : only if
Since $b = 1$, we have by (III.10) that

$$M = \log K \qquad \text{(III.116)}$$

Furthermore

$$\Sigma(i) = \int_0^1 \sigma(i')\, di' \qquad \text{(III.95)}$$

according to (III.95).
Hence, combining (III.116) and (III.95), we find

$$\Sigma(i) = K^i - 1 \qquad \text{(III.115)}$$

for every $i \in [0,A]$.

if
From (III.95) it follows that

$$\sigma(i) = \Sigma'(i) \qquad \text{(III.116)}$$

for every $i \in [0,A]$. Hence $\Sigma(i) = K^i - 1$ implies $\sigma(i) = \log K.K^i = M.K^i$ (using $b = 1$ again), for every $i \in [0,A]$.

(2) If both σ and Σ are of the form

$$\sigma(i) = M.K^i \qquad \text{(III.114)}$$

$$\Sigma(i) = K_1^i - 1 \qquad \text{(III.115)}$$

for every $i \in [0,A]$, where M and $K, K_1 > 1$ are constants, then $b = 1$.

Proof :
Indeed

$$\Sigma(i) = \int_0^i \sigma(i')\, di'$$

by (III.95).
Hence

$$\Sigma(i) = \frac{M}{\log K} K^i - \frac{M}{\log K}$$

$$= K_1^i - 1 \qquad \text{(III.118)}$$

for every $i \in [0,A]$ only if

$$\frac{M}{\log K} = 1$$

and

$$K = K_1 \quad .$$

This gives, with (III.10) that $b = 1$. □

The above proofs in III.4.5 are a second proof (but yielding more insight) of the following corollary, which is in fact an immediate corollary of theorems III.4.2 and III.1.2 :

Corollary :
Zipf's (or Pareto's) function is the only function that agrees with both the verbal and the graphical form of Bradford's law (and this in the group-dependent as well as in the group-free version).

A final chapter deals with the practical applications of the results obtained so far and more specifically on the formulae that were proved.

CHAPTER IV : FITTING METHODS FOR INFORMETRIC LAWS

We will devote ourselves to the fitting of the group-dependent law of Bradford, Leimkuhler's function, the generalised Leimkuhler function (with general α) and Lotka's function (with general α).

These fittings, of course, are based on the results of the previous chapter. Extensive practical evidence of the value of the methods will be given. All methods can be applied directly to "raw" data as e.g. practical bibliographies (but also other IPP's are allowed).

These fittings do have applications (as will be indicated), but are also necessary to show the validity of the several formulae between parameters, we have proved so far. No doubt, the proved formulae are correct from a mathematical or theoretical informetric point of view, but we must be aware of possible problems arising from the fact that practical data differ (slightly) from the theoretical models. Only when our results are stable w.r.t. minor deviations in the data, will they be good and acceptable. This will be verified in the sequel.

IV.1. Fitting of the classical law of Bradford with p groups ($p \in \mathbb{N}$, $p \geq 3$)

IV.1.1. Methodology of fitting

In principle, choose any whole number p of Bradford groups that you want to obtain. Usually, take p between 4 and 10, but for large IPP's a choice larger than 10 may be appropriate (for smaller IPP's a choice for a large p gives rise to small Bradford groups and this might give some fluctuations in the Bradford groups).

p = 3 is allowed but gives not much of a "law", although Bradford himself adapted this value; cf. (Bradford,

1934).

An element in deciding what value of p to use is offered by the practical advantage of finding a value of r_o (the number of sources in the first Bradford group) which is a whole number (a formula for r_o is given after these lines). This is not really a requirement since one can always round of to the nearest whole number, but if the calculated r_o is close to a whole number, this reduces some initial fluctuations, when constructing the Bradford groups.

Once p is chosen, calculate

$$k = (e^{\gamma} y_m)^{1/p} = (1.781\ y_m)^{1/p} \qquad (F_{13})$$

and then

$$y_o = \frac{A}{p} \qquad (F_9)$$

and

$$r_o = \frac{T(k-1)}{k^p - 1} \qquad (F_{10})$$

Since A and T are obviously known from the raw data, r_o and y_o are easily calculated, once k is calculated by formula (F_{13}).

As said before, try several values of p in order to get a r_o that is close to a whole number. If this cannot be reached, use any p but take $[r_o]$, the largest whole number smaller than r_o (one could use $[r_o] + 1$ also, but if r_o is not close to a whole number, $[r_o] + 1$ is larger than r_o and this gives rise to an incomplete last Bradford group). For the calculations of $r_o k$, $r_o k^2$,..., we use the exact r_o and k (not rounded off) and $r_o k$, $r_o k^2$,..., themselves are rounded off in the usual way (we need a whole number of sources!).

A final remark : since formula (F_{10}) implies that

$$\lim_{p \to \infty} r_0 = 0 \qquad \text{(IV.1)}$$

(since $k > 1$), values of p for which $r_0 < 1$ are immediately excluded (r_0 must at least be 1!).

IV.1.2. Application to "Applied Geophysics"

This is one of the two classical bibliographies studied by Bradford himself in 1934, cf. (Bradford, 1934). The data are as follows :

# journals	corresponding # articles	r	R(r) (observed)
1	93	1	93
1	86	2	179
1	56	3	235
1	48	4	283
1	46	5	329
1	35	6	364
1	28	7	392
1	20	8	412
1	17	9	429
4	16	13	493
1	15	14	508
5	14	19	578
1	12	20	590
2	11	22	612
5	10	27	662
3	9	30	689
8	8	38	753
7	7	45	802
11	6	56	868
12	5	68	928
17	4	85	996
23	3	108	1065
49	2	157	1163
169	1	326	1332

Table IV.1 : Applied Geophysics, 1928-1931 (incl.)

Calculating Bradford's law with p = 3 does not involve much checking but, since Bradford himself considered this case, we will investigate also p = 3 with our methods. We find :

$$p = 3,\ k = (1.781\ y_m)^{1/3} = (1.781 \times 93)^{1/3} = 5.49$$

$$y_o = \frac{1332}{3} = 444 \quad \text{and} \quad r_o = \frac{326(k-1)}{k^3-1} = 8.93 \approx 9$$

The Bradford groups are

	# journals	# articles	k
1st group	$r_o = 8.93 \approx 9$	429	-
2nd group	$r_o k = 49.03 \approx 49$	449	5.44
3rd group	$r_o k^2 \approx 269$	446	5.49

Table IV.2 : Bradford's law for AG, p = 3

which is better than Bradford's original division, (Bradford, 1934) : he gets 9/59/258 journals yielding respectively 429/499/404 articles.

We now show that p = 3 can be changed into any reasonable number. For p = 5 we find $k = (1.781\ y_m)^{1/5} = 2.78$, $y_o = \frac{1332}{5} \approx 266$ and $r_o = \frac{326(k-1)}{k^5-1} = 3.52$. This value of r_o is far from a whole number but we nevertheless can work with it. We use $[r_o] = 3$. The Bradford groups are :

	# journals	# articles	k
1st group	$r_o = 3.53 \approx 3$	235	-
2nd group	$r_o k = 9.81 \approx 10$	258	3.33
3rd group	$r_o k^2 = 27.27 \approx 27$	274	2.70
4th group	$r_o k^3 = 75.81 \approx 76$	314	2.81
5th group	$r_o k^4 \approx 210$, exactly the last rank	251	2.76

Table IV.3 : Bradford's law for AG, p = 5

IV.1.3. Application to "Lubrication"

This is the second classical bibliography studied by Bradford, cf. (Bradford, 1934). The data are :

# journals	corresponding # articles	r	R(r) (observed)
1	22	1	22
1	18	2	40
1	15	3	55
2	13	5	81
2	10	7	101
1	9	8	110
3	8	11	134
3	7	14	155
1	6	15	161
7	5	22	196
2	4	24	204
13	3	37	243
25	2	62	293
102	1	164	395

Table IV.4 : Lubrication, 1931 - june 1933

Again, we consider the case p = 3 since Bradford himself considered this. We have $k = (1.781 \times 22)^{1/3} = 3.40$, $y_o = \frac{395}{3} = 131.67 \approx 132$ and $r_o = \frac{164(k-1)}{k^3-1} = 10.30$. Hence use $[r_o] = 10$. The groups are

	# journals	# articles	k
1st group	$r_o = 10.30 \approx 10$	126	-
2nd group	$r_o k = 35.02 \approx 35$	133	3.50
3rd group	$r_o k^2 \approx 119$, which is exactly the last rank in the bibliography	136	3.40

Table IV.5 : Bradford's law for L, p = 3

This is better than Bradford's original example, (Bradford, 1934) : he gets 8/29/127 journals yielding respectively 110/133/152 articles. For p = 4, 5 or 6 we

find a r_o not close to a whole number, so we do not use them. For p = 7 we find k = 1.69, y_o = 56 and r_o = 2.95 ≈ 3 (if r_o is close to a whole number, take this number (here 3) ; otherwise take $[r_o]$ even if $r_o > [r_o] + 0.5$. E.g. for 2.70 : choose 2 : if we take 3 in this case, the last Bradford group will be incomplete - this is not a problem with the present value 2.95). The Bradford groups are :

	# journals	# articles	k
1st group	$r_o = 2.95 \approx 3$	55	-
2nd group	$r_o k = 4.98 \approx 5$	55	1.67
3rd group	$r_o k^2 = 8.42 \approx 8$	56	1.60
4th group	$r_o k^3 = 14.23 \approx 14$	56	1.75
5th group	$r_o k^4 = 24.05 \approx 24$	55	1.71
6th group	$r_o k^5 = 40.64 \approx 41$	49	1.71
7th group	$r_o k^6 = 68.68 \approx 69$ which is exactly the last existing rank	69	1.68

Table IV.6 : Bradford's law for L, p = 7

This shows once more that any reasonable (i.e. related to the finiteness of the discrete IPP) value of p can be used, contrary to what was believed before.

IV.1.4. Application to "ORSA"

This bibliography in operations research was introduced by Kendall (1960) but the data can also be found in (Brookes, 1981). They are

# journals	corresponding # articles	r	R(r) (observed)
1	242	1	242
1	114	2	356
1	102	3	458
1	95	4	553
1	58	5	611
1	49	6	660
1	34	7	694
1	22	8	716
1	22	9	738
1	21	10	759
1	21	11	780
1	20	12	800
1	20	13	820
1	18	14	838
1	16	15	854
1	16	16	870
1	16	17	886
1	16	18	902
1	15	19	917
1	15	20	932
1	14	21	946
2	12	23	970
5	11	28	1025
3	10	31	1055
4	9	35	1091
8	8	43	1155
8	7	51	1211
6	6	57	1247
10	5	67	1297
17	4	84	1365
29	3	113	1452
54	2	167	1560
203	1	370	1763

Table IV.7 : ORSA

For $p = 4$, we find $k = (1.781 \times 242)^{0.25} = 4.56$, $y_o = \frac{1763}{4} = 441$ and $r_o = \frac{370\ (k-1)}{k^5 - 1} = 3.05 \approx 3$. We hence have the following Bradford groups

	# journals	# articles	k
1st group	$r_o = 3.05 \approx 3$	458	-
2nd group	$r_o k = 13.91 \approx 14$	428	4.67
3rd group	$r_o k^2 = 63.43 \approx 63$	463	4.50
4th group	$r_o k^3 = 289.24 \approx 289$ (1 article unused)	413	4.59

Table IV.8 : Bradford's law for ORSA, p = 4

IV.1.5. Application to "Mast Cell"

This bibliography was compiled by Selye (1968) for the period 1877 until early 1964, but the data can also be found in (Goffman and Warren, 1969) and (Goffman and Warren, 1980). They are

# journals	corresponding # articles	r	R(r) (observed)
1	66	1	66
1	58	2	124
1	57	3	181
1	55	4	236
1	53	5	289
1	46	6	335
1	40	7	375
2	38	9	451
1	37	10	488
1	35	11	523
1	34	12	557
1	32	13	589
1	31	14	620
1	30	15	650
1	28	16	678
1	27	17	705
2	23	19	751
1	22	20	773
2	21	22	815
2	20	24	855
2	19	26	893
2	18	28	929
1	17	29	946
1	16	30	962
3	15	33	1007
6	14	39	1091
3	13	42	1130
5	12	47	1190
8	11	55	1278
6	10	61	1338
11	9	72	1437
6	8	78	1485
8	7	86	1541
8	6	94	1589
16	5	110	1669
24	4	134	1765
35	3	169	1870
90	2	259	2050
328	1	587	2378

Table IV.9 : Mast Cell

In (Goffman and Warren, 1969) as well as in (Goffman and Warren, 1980), I was struck by the fact that only the first seven "Bradford groups" are presented, dealing only with 51,9 % of the articles (produced by only 8,7 % of the journals). I was even more surprised by the numbers themselves :

	# journals	# articles	k
1st group	3	181	-
2nd group	4	194	1.3
3rd group	5	182	1.2
4th group	6	171	1.2
5th group	8	165	1.3
6th group	11	170	1.3
7th group	14	171	1.2

Table IV.10 : "Bradford distribution" according to (Goffman and Warren, 1969 and 1980)

Indeed, according to our theory, to have $k \approx 1.25$ (the average of the multipliers in the above table), one has

$$1.25 = (1.781\ y_m)^{1/p}$$

So

$$p = \frac{\gamma + \log y_m}{\log 1.25} \approx 26$$

and hence

$$y_o = \frac{2378}{26} \approx 91$$

But in table IV.10 we have $y_m \approx 176$ (the average of the third column). Conversely, requiring $y_o = 176$ yields $\frac{2378}{y_o} \approx 13.5$ groups and so $k = (1.781\ y_m)^{1/13.5} \approx 1.424$

which differs considerably from k in table IV.10. So our theory proves, before checking in a direct way, that table IV.10 does not represent real Bradford groups. To convince the reader for 100 %, we will extend table IV.10 keeping $y_o \approx 176$, and watch what happens with k.

	# journals	# articles	k
8^{th} group	18	176	1.3
9^{th} group	27	179	1.5
10^{th} group	40	176	1.5
11^{th} group	70	175	1.8
12^{th} group	121	176	1.7
13^{th} group	262	262	2.2

Table IV.10bis : Completion of the Goffman-Warren table IV.10

What we predicted did come out : k had to increase and hence, table IV.10 and IV.10bis do not represent a Bradford analysis. I think that cutting at the seventh group is misleading and does not help to gain an insight into the mechanism of Bradford's law.

If we use also 13 groups, as Goffman-Warren did (we do not have to!), we find with our methods : p = 13, $k = (1.781 \times 66)^{1/13} = 1.44$, $y_o = 182.9 \approx 183$ and $r_o = \frac{587\ (k-1)}{k^{13} - 1} = 2.28$. Hence we have the following Bradford groups

	# journals	# articles	k
1st group	$r_o = 2.28 \approx 2$	124	-
2nd group	$r_o k = 3.28 \approx 3$	165	1.50
3rd group	$r_o k^2 = 4.72 \approx 5$	199	1.67
4th group	$r_o k^3 = 6.80 \approx 7$	217	1.40
5th group	$r_o k^4 = 9.79 \approx 10$	206	1.43
6th group	$r_o k^5 = 14.10 \approx 14$	206	1.40
7th group	$r_o k^6 = 20.30 \approx 20$	221	1.43
8th group	$r_o k^7 = 29.23 \approx 29$	227	1.45
9th group	$r_o k^8 = 42.09 \approx 42$	192	1.45
10th group	$r_o k^9 = 60.61 \approx 61$	162	1.45
11th group	$r_o k^{10} = 87.28 \approx 87$	153	1.43
12th group	$r_o k^{11} = 125.68 \approx 126$	126	1.45
13th group	$r_o k^{12} \approx 181$ (exactly the last rank in the bibliography)	181	1.44

Table IV.11 : Correct Bradford distribution for Mast Cell, p = 13

Furthermore, in the next section we will see that the "Mast Cell" bibliography does not fit perfectly Leimkuhler's function (nor Bradford's law either) and hence, in forming the Bradford groups we must expect some fluctuation in y_o. Table IV.11, although not perfect, is the best that can be done : the deviating values of y_o in the first group (124 instead of $y_o = 183$) is due to the rounding offs in the second column and is normal, because of the small number of journals involved while the values y_o in the tenth until the last group shows the deviation of the bibliography from Bradford's law (as will be seen more clearly later on). Such an analysis cannot be done for table IV.10 and IV.10bis : even if we cut the beginning and end groups away, we do not end up with a relatively constant k, and even if we keep only table IV.10 (the

table as shown by Goffman-Warren), an average value of $k = 1.25$ is not acceptable together with an average value of $y_o = 176$, according to the theory.

IV.1.6. Application to "Schistosomiasis"

Warren and Newill (1967) compiled the schistosomiasis bibliography (1852-1962), see also (Goffman and Warren, 1969 and 1980). The data are

# journals	corresponding # articles	r	R(r) (observed)
1	325	1	325
1	266	2	591
1	259	3	850
1	215	4	1065
1	211	5	1276
1	171	6	1447
1	159	7	1606
1	143	8	1749
1	137	9	1886
1	136	10	2022
1	118	11	2140
1	115	12	2255
1	112	13	2367
1	108	14	2475
2	105	16	2685
1	94	17	2779
1	90	18	2869
1	80	19	2949
1	74	20	3023
2	72	22	3167
2	70	24	3307
1	68	25	3375
1	66	26	3441
1	64	27	3505
1	56	28	3561
2	55	30	3671
2	51	32	3773
1	50	33	3823
1	47	34	3870
1	45	35	3915
1	44	36	3959
2	42	38	4043
1	41	39	4084
1	40	40	4124
2	39	42	4202
3	37	45	4313
1	36	46	4349
2	35	48	4419
1	34	49	4453

cont.

# journals	corresponding # articles	r	R(r) (observed)
1	33	50	4486
3	32	53	4582
3	31	56	4675
2	29	58	4733
5	28	63	4873
1	27	64	4900
1	26	65	4926
2	25	67	4976
3	24	70	5048
4	23	74	5140
2	22	76	5184
4	21	80	5268
3	20	83	5328
4	19	87	5404
10	18	97	5584
8	17	105	5720
10	16	115	5880
9	15	124	6015
10	14	134	6155
10	13	144	6285
6	12	150	6357
11	11	161	6478
14	10	175	6618
19	9	194	6789
29	8	223	7021
27	7	250	7210
44	6	294	7474
57	5	351	7759
76	4	427	8063
137	3	564	8474
266	2	830	9006
908	1	1738	9914

Table IV.12 : Schistosomiasis (1852-1962)

In (Goffman and Warren, 1969) again the first seven Bradford groups are presented, but in (Goffman and Warren, 1980) this time the complete Bradford distribution is shown, consisting of p = 16 groups. Now they succeeded in keeping y_0 constant (about 620) as well as k relatively constant (about 1.5, although at the end k increases). With our theory, we obtain with p = 16 (but we may take other values of course, if we want to), $k = (1.781 \times 325)^{1/16} = 1.49$, $y_0 = 619.6 \approx 620$: i.e., we re-calculate the Goffman-Warren values for y_0 and k; so we think that our method provides a rationale for calculating the Bradford groups. Previously, the only method that one could use was "trial

and error". We finally give the Bradford groups for p = 9, showing that p = 16 is not necessary : here we have $k = (1.781 \times 325)^{1/9} = 2.03$, $y_o = 1102$ and $r_o = \frac{1738\ (k-1)}{k^9 - 1} = 3.06 \approx 3$.

We obtain the following Bradford groups :

	# journals	# articles	k
1st group	$r_o = 3.06 \approx 3$	850	-
2nd group	$r_o k = 6.21 \approx 6$	1036	2.00
3th group	$r_o k^2 = 12.61 \approx 13$	1281	2.17
4th group	$r_o k^3 = 25.60 \approx 26$	1252	2.00
5th group	$r_o k^4 = 51.97 \approx 52$	1216	2.00
6th group	$r_o k^5 = 105.499 \approx 105$	1242	2.02
7th group	$r_o k^6 = 214.17 \approx 214$	1154	2.04
8th group	$r_o k^7 = 434.77 \approx 435$	999	2.03
9th group	$r_o k^8 \approx 883$ which is the second to last rank	883	2.03

Table IV.13 : Bradford's law for Schistosomiasis, p = 9

For the same reason why in (Goffman and Warren, 1980) k increases at the end (and y_o remains relatively constant) we find in table IV.13 that y_o decreases somewhat (with k constant). This is due (as in the preceding case) to the fact that the Schistosomiasis bibliography does not perfectly conform with a Bradford distribution. The reason for this is probably the fact that this bibliography ranges over a very long time-period 1852-1962 in which journals "come and go". This will be explained in more detail in section IV.2.8. The same note can be made for the Mast Cell bibliography.

So in this case, the Goffman-Warren Bradford groups are acceptable. Of course, in none of the previous works a real algorithm or rationale to calculate the Bradford groups

was given. We think that this section has provided this.

Note :
It is not clear what statistical test is needed to fit Bradford's law. As remarked to me by B.C. Brookes, a χ^2-test could be used (comparing the actual number of items in the groups with the expected "equal" number of items. However it is not clear if one tests the right thing here : the law of Bradford is an exponential function (σ in the continuous setting) that has been "reformed" in a group-dependent discrete way.

IV.2. Fitting of the Leimkuhler function for known IPP's

IV.2.1. Methodology of fitting

The Leimkuhler function

$$R(r) = a \log (1 + br) \qquad \text{(III.4)}$$

can be deduced from the law of Bradford by using the following exact formulae :

$$a = \frac{y_o}{\log k} \qquad (F_1)$$

$$b = \frac{k - 1}{r_o} \qquad (F_2)$$

Hence, in view of the methods developed earlier (cf. formulae (F_9), (F_{10}) and (F_{13})), it is extremely simple to calculate Leimkuhler's function (III.4). The situation here is even simpler since there is no need for a r_o close to a whole number.

The degree of closeness of the calculated function (III.4) with the observed cumulative distribution function is an indication of the value of the method developed earlier. We will investigate this for the bibliographies used in the previous paragraph as well as for two other bibliographies. Any IPP could have been checked, however.

IV.2.2. "Applied Geophysics"

Using some data from section IV.1.2, we have, for $p = 3$, $k = 5.49$, $r_o = 8.93$ (no need to round off now) and $y_o = 444$. Hence.

$$a = \frac{y_o}{\log k} = 260.7$$

$$b = \frac{k - 1}{r_o} = 0.503$$

Hence, the law of Leimkuhler for "Applied Geophysics" is

$$R(r) = 260.7 \log (1 + 0.503\ r) \qquad (IV.2)$$

See Fig.IV.1 (upper-curve), where plus-signs stand for the observed values and the crosses for the values according to the calculated function. It is obvious that the fit is very close. This can also be checked when comparing the table of observed R(r) (previous section) with the one of the calculated R(r) (from formula (IV.2)) and performing a Kolmogorov-Smirnov test for goodness of fit. The other marks (points) visualise the law of Leimkuhler as calculated by Brookes (1985). Our fit is better, although Brookes' fit is also very good. Brookes' method is one of fitting the beginning and ending points of the observed graph : this yields two equations from which a and b are calculated. This method is good but is an "ad hoc" method, not based on the theory of Bradford's law. Furthermore, the calculations needed to find the Leimkuhler function according to Brookes' method are more intricate (the equation for b for instance is transcendent and is numerically solved via an iteration process) than our simple formulae (which have also the advantage of being mathematically derived).

We finally remark here that the choice of p is arbitrary. Any other (reasonable) choice of p must lead to the same equation (IV.2). Take f.i. $p = 5$. Now we find $y_o = \frac{1332}{5} = 266.4$, $k = (e^{\gamma} y_m)^{0.2} = 2.778$ and $a = \frac{y_o}{\log k} = 260.7$. Furthermore, $r_o = \frac{T(k-1)}{k^5 - 1} = 3.525$. So $b = \frac{k - 1}{r_o} = 0.504$ yielding

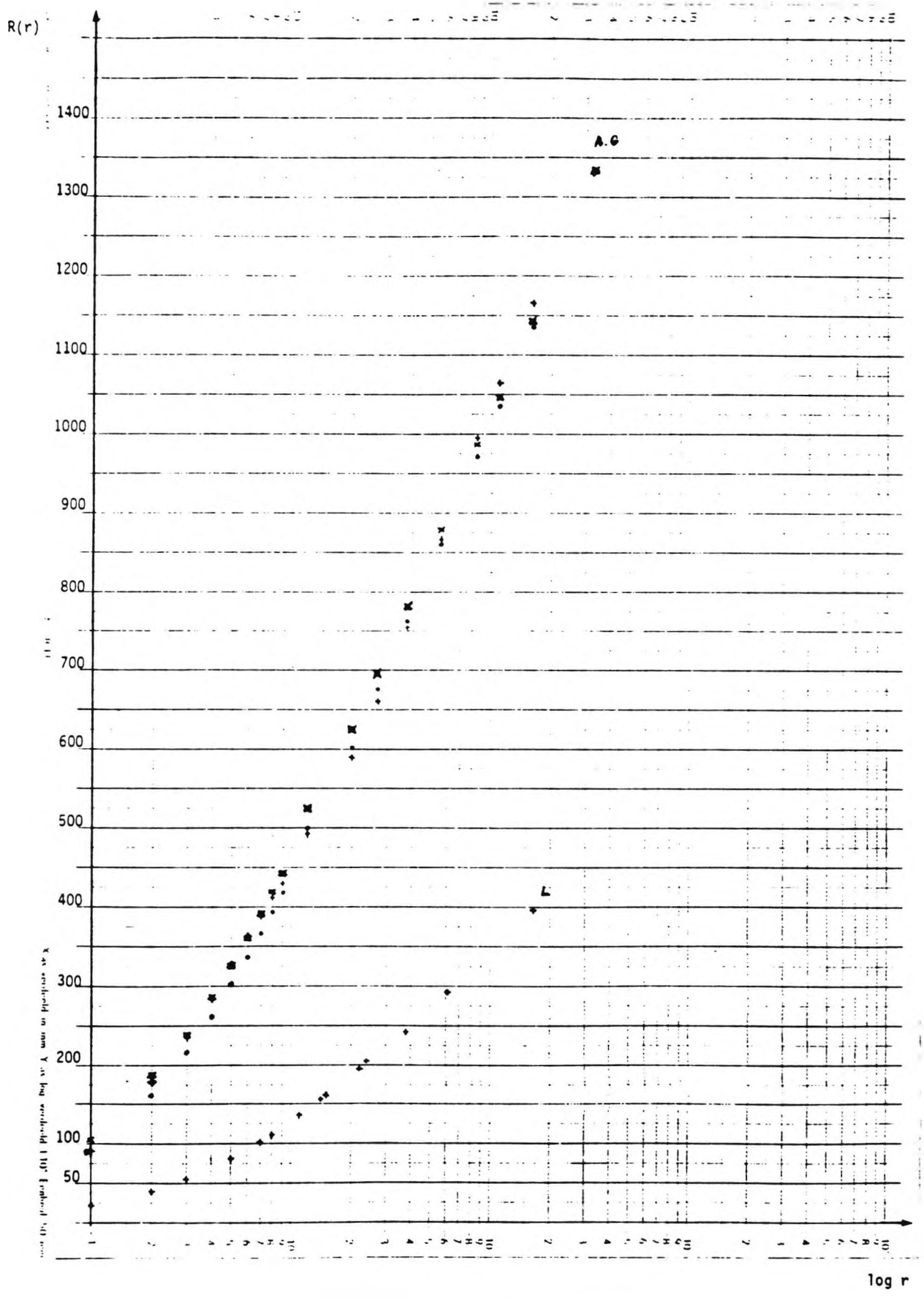


Fig.IV.1 : Fittings of AG and L

+ = Observed
• = Calculated (Brookes)
x = Calculated (Egghe)

$$R(r) = 260.7 \log (1 + 0.504\ r)$$

which is little different from formula (IV.2).

IV.2.3. "Lubrication"

From the data of section IV.1.3 we have $y_o = 131.67$, $k = 3.40$, $a = \frac{y_o}{\log k} = 107.7$, $r_o = 10.30$ and $b = \frac{k - 1}{r_o} = 0.233$. Hence, Leimkuhler's function for "Lubrication" is

$$R(r) = 107.7 \log (1 + 0.233\ r) \qquad (IV.3)$$

See Fig.IV.1 (lower curve), where we only showed the observed data since the points, as calculated by formula (IV.3) are not separable from the former ones (in fact the same is true for the data as calculated by Brookes (1985)). This remarkable fit is due to the fact that "Lubrication" is a perfect Bradfordian bibliography. This is of course a requirement for a good fit : if the data do not conform with the law of Bradford, their graph differs also from the Leimkuhler function and hence, a perfect fit cannot be obtained. We will encounter such bibliographies later on in this chapter. There, however, we will explain why Leimkuhler's function still is very important.

IV.2.4. "ORSA"

From section IV.1.4 we find, for $p = 4$ (but you make take any reasonable p to start with), that $a = 290.64$ and $b = 1.167$. Hence we find here

$$R(r) = 290.64 \log (1 + 1.167\ r) \qquad (IV.4)$$

See Fig.IV.2 for a visual view of the very good fit, and compare also with Brookes' fit, now based on a method of Wilkinson, (Wilkinson, 1978), but still using the method of fitting data points.

In conclusion we can say that, if the bibliography is Bradfordian (hence satisfies Leimkuhler's function), we

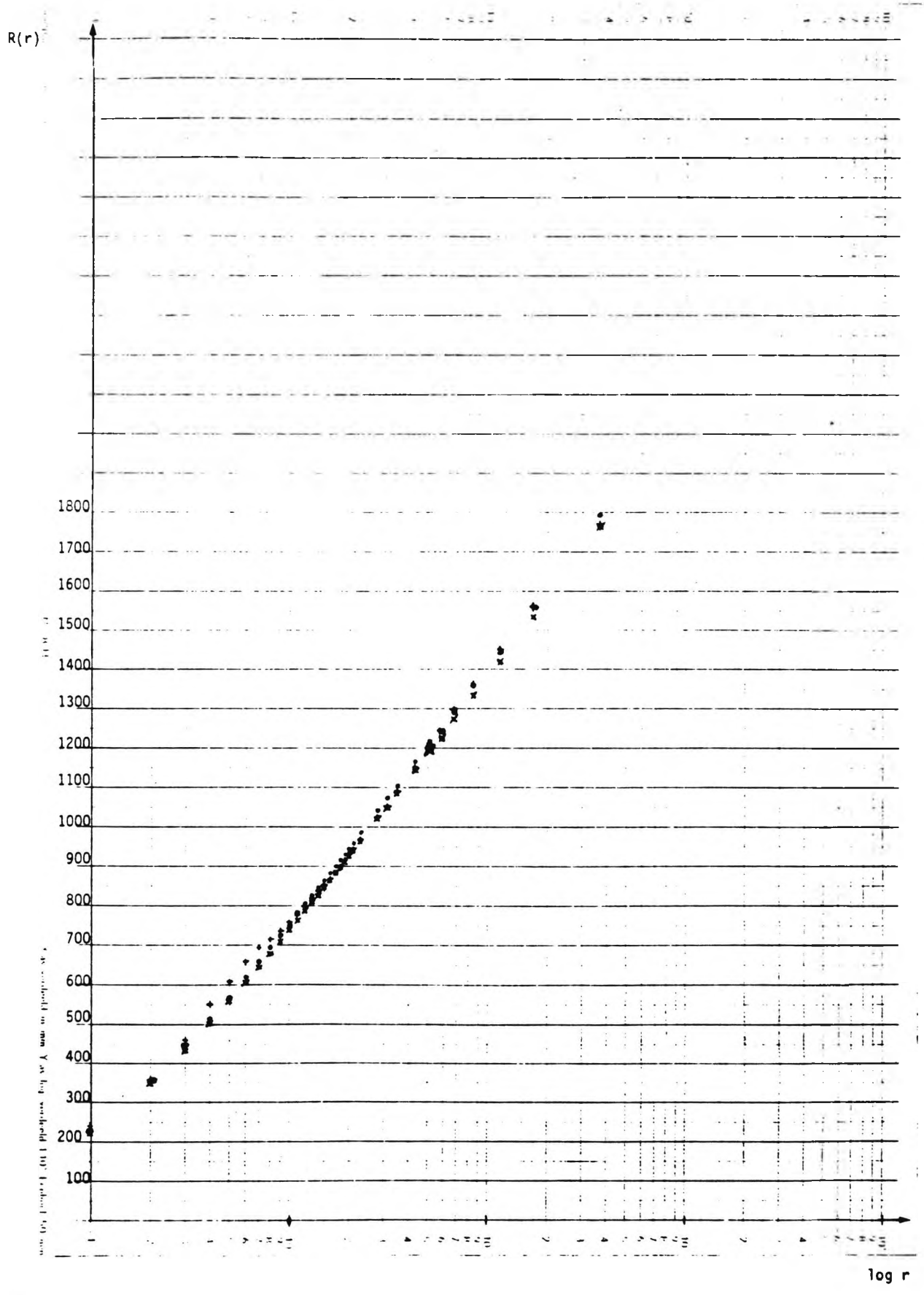


Fig.IV.2 : Fittings of ORSA

+ = Observed
• = Calculated (Brookes)
x = Calculated (Egghe)

find an almost exact fit by using our theory on Bradford's law.

In the next bibliographies we will encounter some deviations from Leimkuhler's function. These deviations, however, will introduce the most important aspect of this paragraph.

IV.2.5. "Mast Cell"

From section IV.1.5 we find for p = 13 and using one more decimal in k = 1.443, that $a = \frac{y_o}{\log k} = 498.7$ and $b = \frac{k-1}{r_o} = 0.1987$. So Leimkuhler's function for "Mast Cell" is

$$R(r) = 498.7 \log (1 + 0.1987\ r) \qquad (IV.5)$$

As we mentioned before, the method is very stable w.r.t. the original choice of p (f.i. for p = 5 we find r = 498.9 and b = 0.1984).

See Fig.IV.3 for a visual view of the observed and calculated points (disregard in this section the curve marked by dots •). The fit is very good, taking into account the "Groos droop" (see at the end of section II.4.3.2) which is in this case more a deflection starting about r = 150 and a revival about at r = 400. This mixed "ending" is also mirrored in table IV.11 (deflection in group 10, 11 and 12 and revival in group 13; this is not shown in tables IV.10 and IV.10bis).

IV.2.6. "Schistosomiasis"

From section IV.1.6 we find, for p = 9, and using again one more decimal in k = 2.027 that $a = \frac{y_o}{\log k} = 1559.7$ and $b = \frac{k-1}{r_o} = 0.3318$. So Leimkuhler's function for "Schistosomiasis" is

$$R(r) = 1559.7 \log (1 + 0.3318\ r) \qquad (IV.6)$$

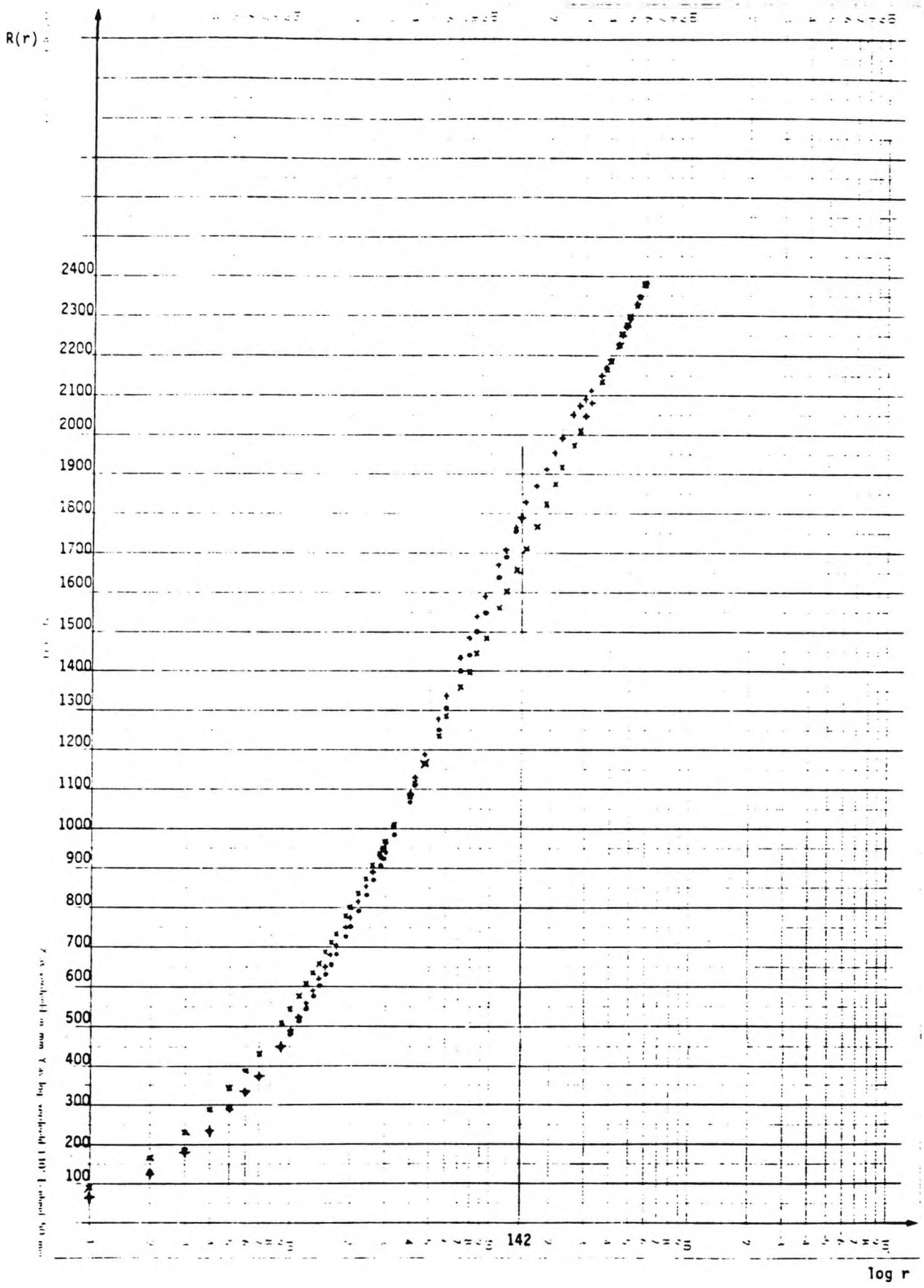
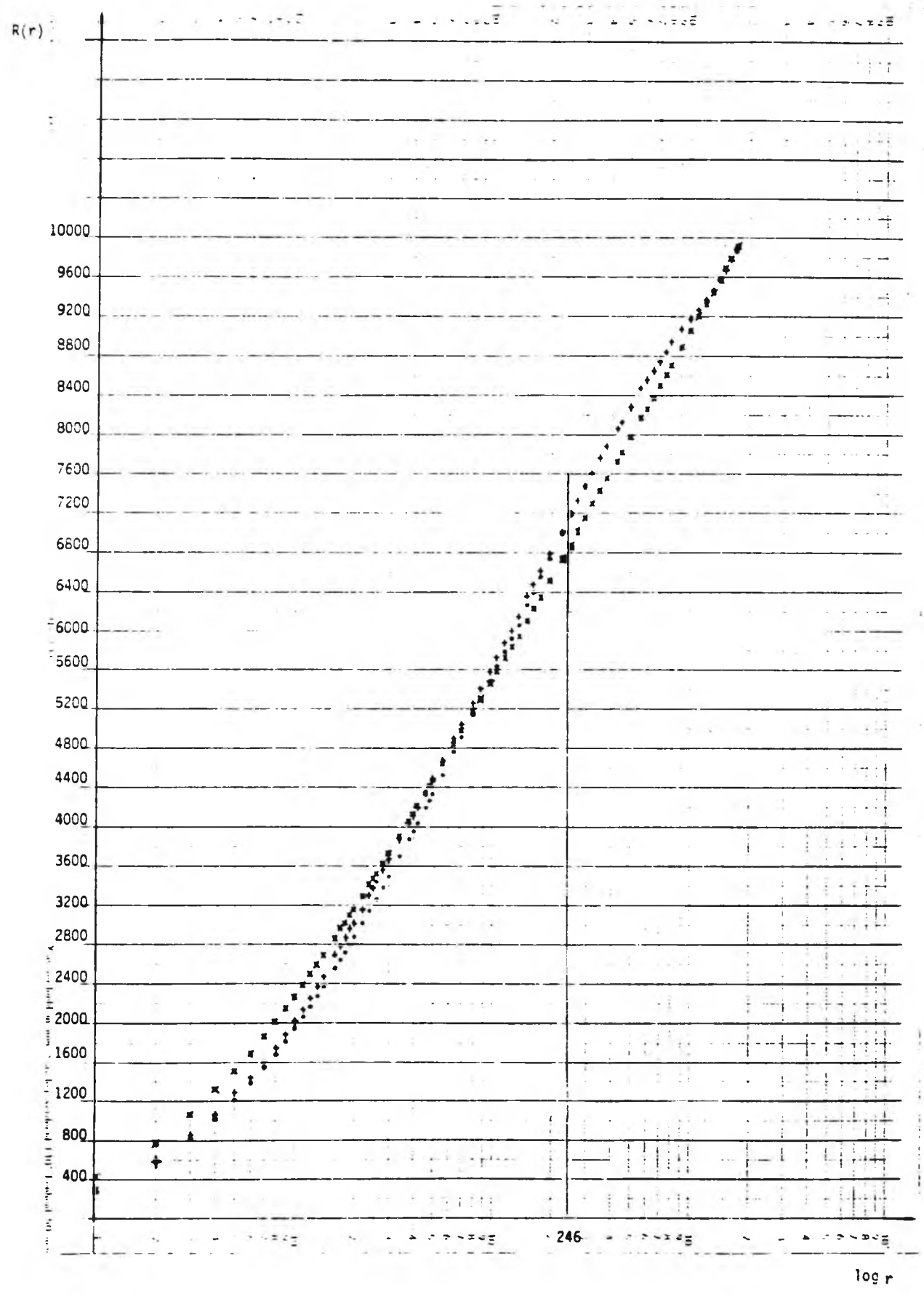


Fig.IV.3 : Fittings of Mast Cell

+ = Observed
x = Calculated (Egghe)
• = Calculated (Egghe, truncated)



+ = Observed
x = Calculated (Egghe)
• = Calculated (Egghe, truncated)

Fig.IV.4 : Fittings of Schistosomiasis

See Fig.IV.4 for a visual view of the observed and calculated points (disregard again in this section the curve marked by dots •). Again, the fit is very good, taking into account the "Groos droop" (cf. the deflection in table IV.13 which is hereby explained).

A bibliography with a large Groos droop now follows.

IV.2.6. Pope's bibliography

In (Pope, 1975), Pope introduces data of a bibliography on information science. They are

# journals	corresponding # articles	r	R(r) (observed)
1	261	1	261
1	259	2	520
1	220	3	740
1	211	4	951
1	205	5	1156
1	176	6	1332
1	168	7	1500
1	164	8	1664
1	155	9	1819
1	134	10	1953
2	120	12	2193
1	115	13	2308
1	105	14	2413
1	102	15	2515
1	96	16	2611
1	85	17	2696
1	80	18	2776
2	79	20	2934
1	78	21	3012
1	74	22	3086
1	64	23	3150
1	63	24	3213
2	60	26	3333
1	59	27	3392
1	53	28	3445
1	52	29	3497
2	51	31	3599
1	45	32	3644
1	44	33	3688
2	42	35	3772
1	40	36	3812
2	38	38	3888
1	36	39	3924

cont.

cont.

# journals	corresponding # articles	r	R(r) (observed)
2	33	41	3990
1	32	42	4022
5	31	47	4177
1	30	48	4207
1	29	49	4236
1	28	50	4264
1	27	51	4291
1	25	52	4316
3	24	55	4388
1	23	56	4411
6	22	62	4543
2	21	64	4585
5	20	69	4685
4	19	73	4761
8	18	81	4905
5	17	86	4990
3	16	89	5038
4	15	93	5098
7	14	100	5196
10	13	110	5326
9	12	119	5434
9	11	128	5533
7	10	135	5603
8	9	143	5675
12	8	155	5771
20	7	175	5911
14	6	189	5995
35	5	224	6170
45	4	269	6350
68	3	337	6554
140	2	477	6834
534	1	1011	7368

Table IV.14 : Pope's bibliography

In calculating Leimkuhler's function for Pope's bibliography, we take f.i. p = 5. Then

$y_o = \frac{7368}{y_o^5} = 1473.6$, $k = (1.781\ y_m)^{0.2} = 3.415$,

$a = \frac{y_o}{\log k} = 1199.8$, $r_o = \frac{1011\ (k-1)}{k^5 - 1} = 5.27$ and $b = \frac{k-1}{r_o} = 0.4584$. This gives

$$R(r) = 1199.8 \log (1 + 0.4584\ r) \qquad (IV.7)$$

See Fig.IV.5 for a comparison of the observed and calculated data. The fit is good, taking into account the large Groos droop which is permanently present from

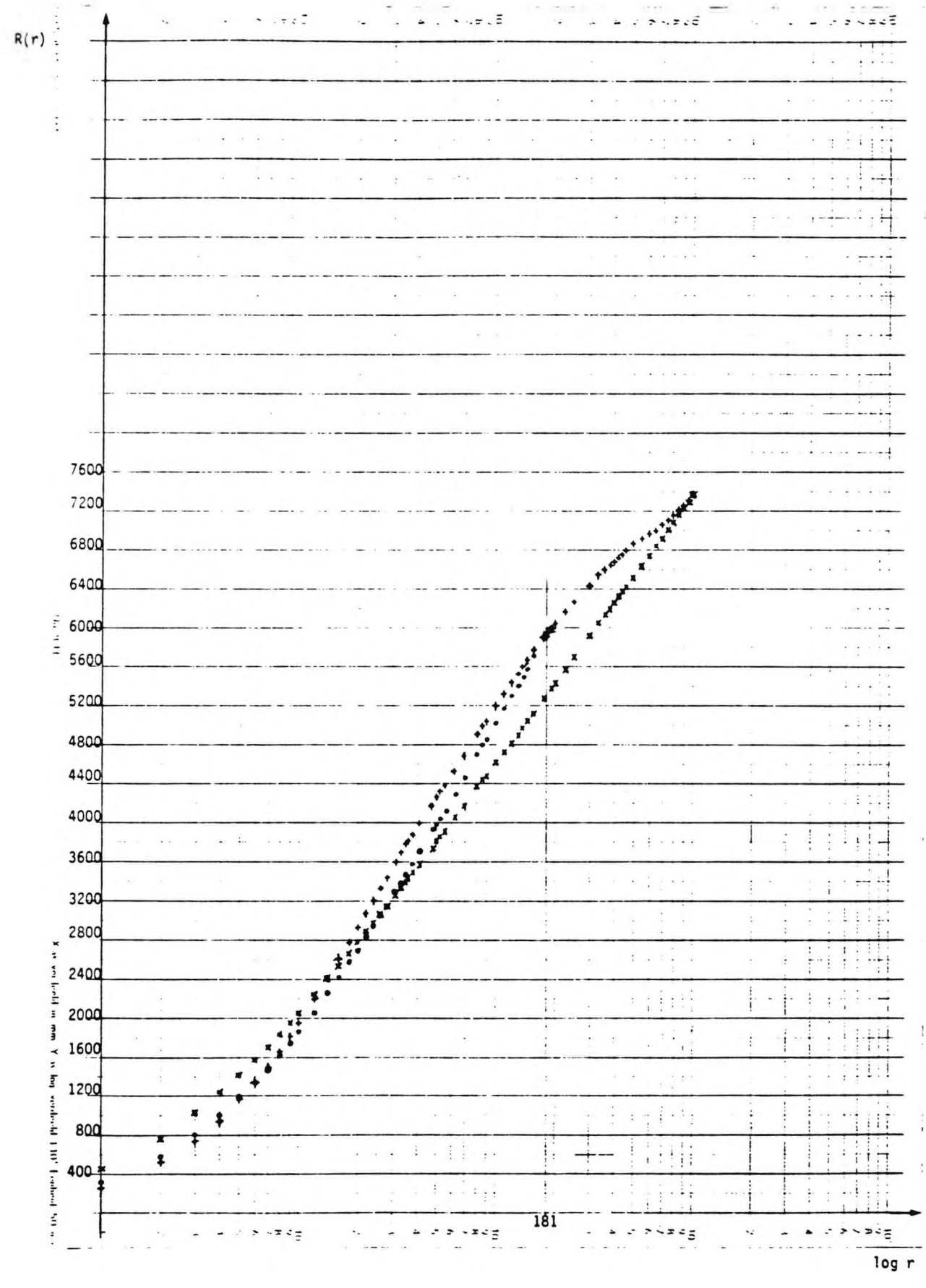


Fig.IV.5 : Fittings of Pope's bibliography

+ = Observed
x = Calculated (Egghe)
• = Calculated (Egghe,truncated)

r = 30 on but is strong from r = 200 on. Again disregard the dotted graph (•) in this section.

IV.2.7. Sachs' bibliography

The next bibliography (on statistical methods) is new and was published only recently by Sachs (1986). From this bibliography we calculated the following table (using only journals) :

# journals	corresponding # articles	r	R(r) (observed)
1	64	1	64
1	44	2	108
1	41	3	149
3	40	6	269
1	37	7	306
1	36	8	342
1	34	9	376
1	33	10	409
1	27	11	436
2	19	13	474
2	18	15	510
1	15	16	525
2	12	18	549
3	9	21	576
2	8	23	592
4	7	27	620
4	6	31	644
5	5	36	669
5	4	41	689
8	3	49	713
21	2	70	755
73	1	143	828

Table IV.15 : Sachs' bibliography

Leimkuhler's function for Sachs' bibliography follows by the following calculations (f.i. for p = 5) :

$k = (1.781\ y_m)^{0.2} = 2.579$, $y_o = 165.6$, $a = \frac{y_o}{\log k} = 174.8$,

$r_o = \frac{143\ (k-1)}{k^6 - 1} = 2.00$ and $b = \frac{k-1}{r_o} = 0.791$. This yields

$$R(r) = 174.8 \log (1 + 0.791\ r) \qquad \text{(IV.8)}$$

See Fig.IV.6 for a comparison between the observed data and the calculated curve, using (IV.8).

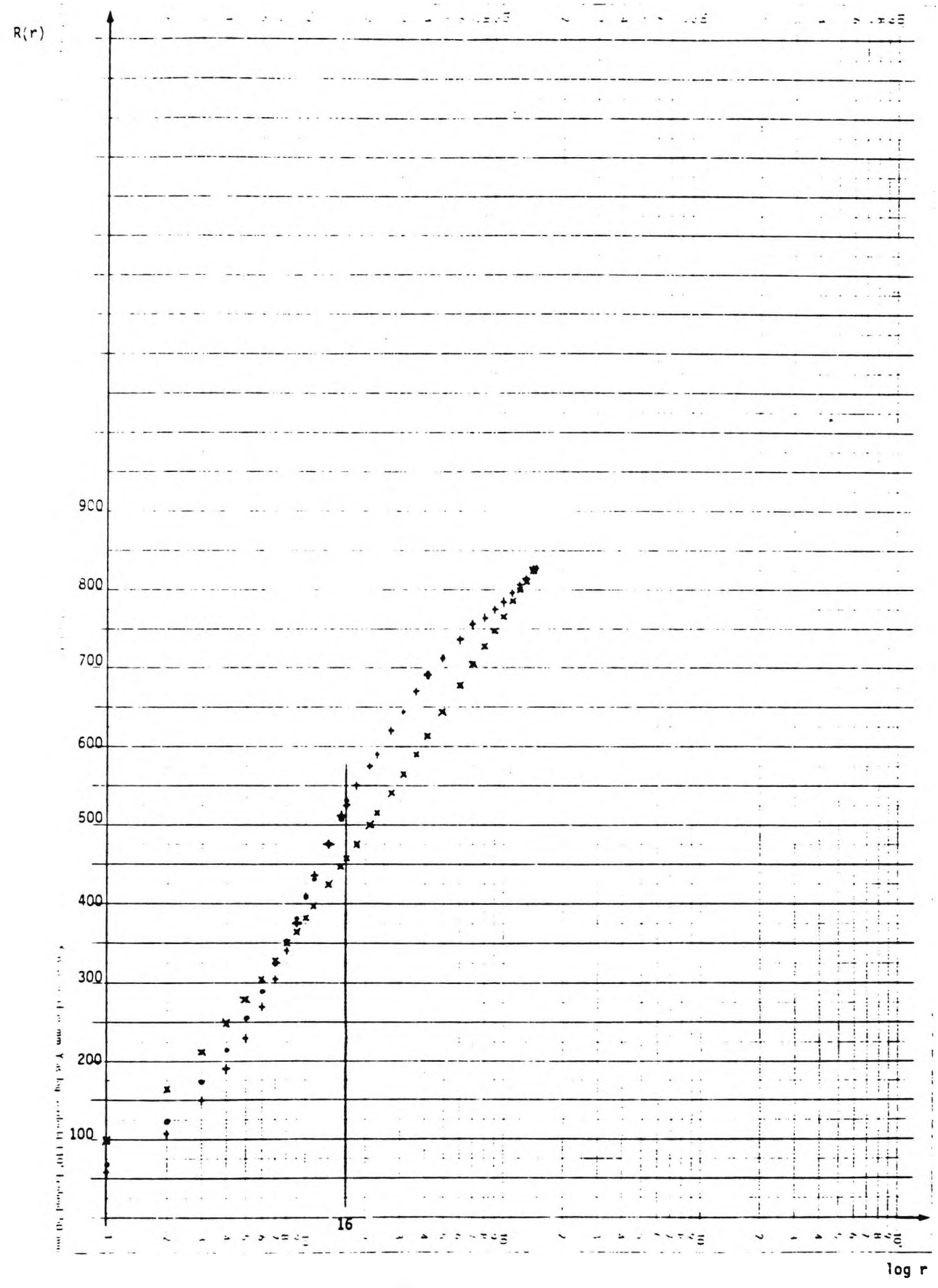


Fig.IV.6 : Fittings of Sachs' bibliography

+ = Observed
x = Calculated (Egghe)
• = Calculated (Egghe,truncated)

IV.2.8. Comments

The above method for calculating Leimkuhler's function is very good, at least for IPP's showing no "Groos droop" (cf. section II.4.3.2) : this is logical since Leimkuhler's function

$$R(r) = a \log (1 + br) \qquad \text{(III.4)}$$

does not involve a Groos droop. Indeed, the graph of (III.4) in semi-logarithmic scale (log r, R(r)) looks like in Fig.IV.7.

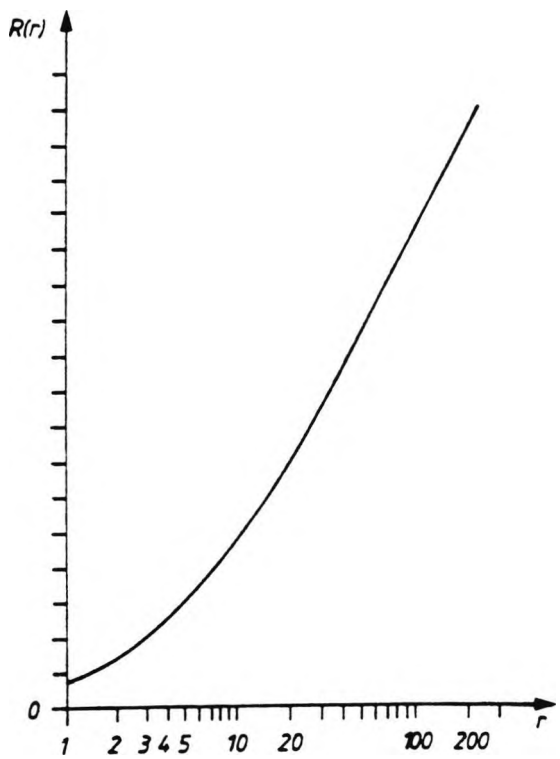


Fig.IV.7 : Leimkuhler's function expressed graphically

We have :

$$\lim_{r \to \infty} \frac{dR(r)}{d \log r} = \lim_{r \to \infty} \frac{abr}{1 + br} = a \quad , \qquad \text{(IV.9)}$$

a constant.

It is a well-known fact that most practical examples of IPP's show a Groos droop (small or large) : see e.g. the examples above, except the first three.

Other examples are (Aiyepeku, 1977), (Brookes, 1969), (Brookes, 1973), (Brookes, 1977),(Brown, 1977), (Drott, Mancall and Griffith, 1979), (Egghe, 1985), (Groos, 1967) (the one who invented the phenomenon, the term "droop" however being invented by B.C. Brookes), (Lipatov and Denisenko, 1986), (Praunlich and Kroll, 1978), (Saracevic and Perk, 1973), (Singleton, 1976), (Asai, 1981), (Avramescu , 1980), (Brookes, 1980) and (Haspers, 1976).

A Groos droop can be defined exactly, as the occurrence of an inflection point r_d in the curve of the function R(r) in semi-logarithmic scale :

$$\frac{d^2R}{d\log r^2}(r_d) = 0 \qquad \text{(IV.10)}$$

We then have a graph as in Fig.IV.8.

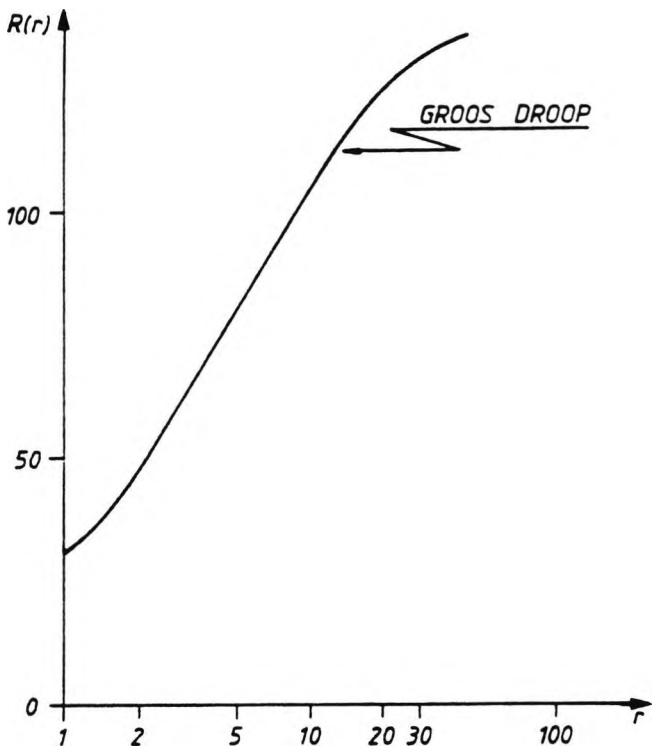


Fig.IV.8 : The Groos droop

This model is not included in Leimkuhler's function (III.4) (the case $\alpha = 2$) but it is included in the generalised Leimkuhler function (II.45) (or (III.65), together with (GF_7) until (GF_{11})), for $\alpha < 2$.

Two approaches are possible.

1. We try to model the Groos droop with function (III.65) (or any other well-fitting function). This droop is then explained in so far that (III.65) is explained. This reduces to the explanation of Lotka's law

$$f(j) = \frac{C}{j^{\alpha}} \qquad \text{(III.1)}$$

$j \in [\rho(0),\rho(A)] = [1,y_m]$. One reference to such an explanation can be found in (Bookstein, 1984). In an other context, in (Avramescu, 1980) one can also find an attempt of explanation. The fitting of function (III.65) will be done in the last part of this chapter.

2. We nevertheless accept Leimkuhler's function (III.4) as a good law to model certain "pure" informetric phenomena and then try to explain deviations from it, due to several reasons. In (Egghe and Rousseau, 1988), this last approach has been studied. There we encountered the following possible explanations for the Groos droop (from the "deviations"-point of view) :

a) Incompleteness of the IPP

One explanation is that our IPP (supposing it to be Bradfordian when complete) is in fact incomplete, due to selectivity or omission. This is very plausible, since, of course, the incompleteness occurs in the higher ranks and not in the lower ones : the less important a source is, in the compilation of an IPP the easier it is to miss a few of them, and that can explain the difference between Figs.IV.7 and IV.8. That this is certainly one of the facts that causes a Groos droop is illustrated by the bibliography on "statistical methods" as compiled by Sachs (1986). This publication has the aim of giving one or two pertinent references for every statistical method that Sachs included in the book. Hence, it was not the purpose to present a complete bibliography and hence we can expect a marked Groos droop here.

This was clearly the case as was seen in section IV.2.7. But incompleteness certainly is not the only reason for the occurrence of a Groos droop.

b) Merging of IPP's

Definition :
Suppose we have N IPP's. We suppose that when an item belongs to two or more of these N IPP's (assumed to be discrete), the corresponding sources are the same (meaning that we are considering only IPP's of the same type). Merging of these IPP's means that we join all items into one "big" IPP (i.e. we simply take the union of the N IPP's, considered as sets of items). Due to the above assumption, the merging of IPP's is again an IPP. Once this is done, we rearrange the sources in decreasing order of number of items that they have.

We can then formulate the following problem :
Suppose that we denote by $R_i(r)$ (resp. $R(r)$) the cumulative number of items in the sources of rank $1,\ldots,r$ in the i^{th} IPP ($i = 1,\ldots,N$) (resp. the merged IPP). What is then the relation between R and R_i ($i = 1,\ldots,N$)? More specifically, assuming all N IPP's to the Bradfordian (hence with R_i-curves as in Fig.IV.7), is then the merged IPP also Bradfordian?

In (Adenaike, 1982), (Aiyepeku, 1977) and (Sen, 1985) it is assumed that the answer to the above question is yes. In (Egghe and Rousseau, 1988) however we showed (using a merging-simulation package constructed by R. Philips) that the merging of Bradfordian IPP's does not necessarily yield a Bradfordian IPP and - what is more - yields sometimes a Groos droop. Explanations of this effect have also been given in (Egghe and Rousseau, 1988), solving at the same time a problem of Bonitz and Schmidt (see (Bonitz and Schmidt, 1982)).

A practical interpretation of "merging", in the area of bibliometry, is "interdisciplinarity" of a bibliography. In the bibliography made by Pope (see IV.2.6) on information science, a large Groos droop is noticed.

So far, this was always interpreted as an indication of the incompleteness of Pope's bibliography. In the light of our ideas, this large Groos droop might also be seen as a consequence of the fact that information science is very much an interdisciplinary subject.

Another interpretation of merging - still in bibliometrics (although generalisations are thinkable), but more in an "osmotic" sense, are the bibliographies that range over a very long time period. Indeed, when the time interval is large, journals change. Of course it is impossible to cut such a bibliography into a number of subbibliographies (according to the time period) such that the journals remain the same, but such a model can be used as a first approximation. The bibliography itself can then be considered as the merged bibliography of the sub-bibliographies, mentioned above. As such, as explained, a Groos droop can be expected, even if the bibliography is complete. A nice example of this is the Schistosomiasis bibliography, ranging over the period 1852-1962. It shows a Groos droop, although this bibliography is believed to be very complete (as mentioned to me by B.C. Brookes).

Note :
Prof. dr. I.K. Ravichandra Rao (DRTC, Indian Statistical Institute, Bangalore, India) remarked to me that another reason for the Groos droop in bibliometrics can be found in the fact that a subject is new and hence that the articles in this early state, are scattered over a lot journals (not directly devoted to the new subject). This can be true but needs further modelling. This aspect is, however, more belonging to the first approach, since special models are required.

In our second approach, i.e. accepting Leimkuhler's function (III.4) and considering the Groos droop as a deviation from it, the fit of our calculated model (III.4) to the practical data will be worse, the larger Groos droop we have. This does not mean that Leimkuhler's law is only

important for IPP's without a Groos droop (such as "Applied Geophysics", "Lubrication" and "ORSA"). Indeed, considering a Groos droop as being caused (completely or partially) by the incompleteness of the IPP, we can still be interested in the Leimkuhler fit (III.4) to the first part of the graph of the data, i.e. the part before the Groos droop : $r < r_d$. Knowing Leimkuhler's function for the (unknown) complete IPP can be important in order to deduce some properties of the complete IPP. What needs to be done is to "cut away" the Groos droop in the IPP and work with the truncated IPP. But, since our theory as developed here, was based on the complete IPP we must modify it in order to be applicable to incomplete IPP (truncating an IPP so as to get rid of the Groos droop leaves an IPP where the least productive sources might have a production larger than one, say 5 or 6; the theory as developed here uses the fact that the least productive sources have production one - so we have to modify our theory). This will be done in the next section and applied to all IPP's (bibliographies) studied so far, which show a Groos droop.

IV.3. Fitting of the Leimkuhler function for unknown IPP's

We will develop in this paragraph a theory and algorithm to calculate the Leimkuhler function for the complete (usually unknown) IPP. For this, we have to "cut off" the Groos droop at the rank ρ_o at which this droop becomes very explicite.

This can easily be done by visual inspection of the graph of observed data (R(r) in function of log r) : In most cases one can draw a line at a certain rank ρ_o such that for ranks $r < \rho_o$ we have almost no droop while for $r > \rho_o$ the droop is very apparent (see f.i. in Fig. IV.9 : rank ρ_o). Furthermore it must be stated that the lower the cut-off rank ρ_o, the more exact we work, from a mathematical point of view, but the fewer data we keep;

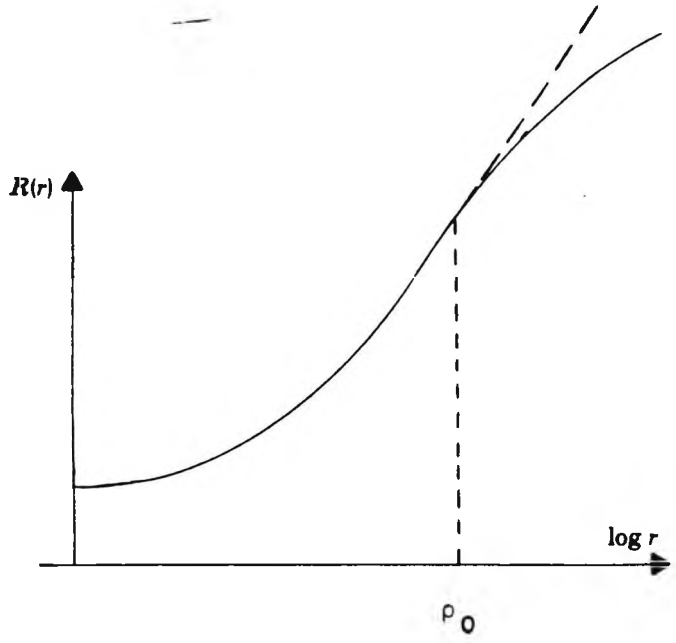


Fig.IV.9 : Cut-off rank

hence individual fluctuations enter into the calculations. Conversely, the higher the cut-off rank ρ_0, the more data are involved, but the Groos droop gives deviations in the calculations. So, anyway, the cut-off point ρ_0 must be chosen somewhere in the middle of the semilogarithmic graph.

We will give in the next section the methodology, then test it on a "perfect" example (to show that our method is "perfect") and then test it on practical bibliographies.

IV.3.1. Methodology (Egghe, 1989d)

- Choose a cut-off rank ρ_0 at which the Groos droop becomes apparent and check the production (the number of items) of the source at this rank, say n.

- Choose a number p of Bradford groups for the complete (unknown) IPP. Take p high enough (e.g. p = 10) so that we can "interpolate" until we reach rank $r = \rho_0$ (we will explain this further on).

- The Bradford factor for the complete IPP is determined as before :

$$k = (1.781\ y_m)^{1/p} \qquad (F_{13})$$

- We repeat a result that was proved in section III.3.4, and which can be applied here, very surprisingly.

Theorem :
Let q denote the number of the Bradford group counted from the last groups onward (i.e. 1 = the number of the last Bradford group, 2 = the number of the second to last Bradford group, etc.). Denote by n(q) the production (i.e. number of items) in the most productive source in this q^{th} (last) Bradford group. Then

$$n = n(q) = \frac{k^q}{e^\gamma} = \frac{k^q}{1.781} \qquad \text{(III.93)}$$

- We calculate q from n as follows : using formula (III.93) we have

$$q = \frac{\gamma + \log n}{\log k} \qquad \text{(IV.11)}$$

- Hence, the source on rank $r = \rho_0$ belongs to the $([q]+1)^{th}$ - last Bradford group ([q] = the largest whole number smaller than or equal to q; indeed q can be a decimal number!), where q is determined from the production n of the source on rank ρ_0 and by (IV.11).

- Since we need further on a whole number of groups, we will take our cut-off point a little lower in rank (not larger, in order to exclude the Groos droop). This means that we take the source with the highest rank in the $([q]+1)^{th}$-group. This is calculated using again formula (III.93) :

$$n' = \frac{k^{[q]+1}}{e^\gamma} \qquad \text{(IV.12)}$$

n' determines the final cut-off rank r'.

- What is left, after truncation at rank r', contains p - [q] - 1 Bradford groups.

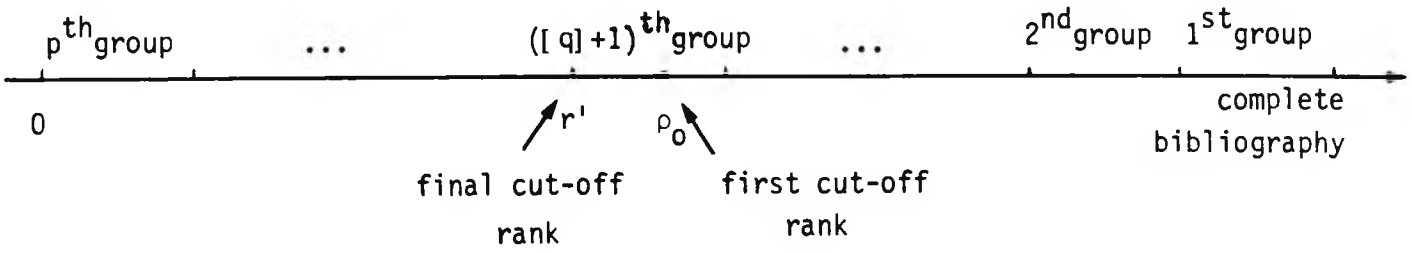


Fig.IV.10 : Geometry of Bradford groups

- We are now in a position to calculate all parameters for the Leimkuhler function for the complete Bradford distribution, based on our truncated one.
  The number of sources $r' = \hat{T}$ and the number $\hat{A}$ of items in the truncated IPP are of course immediately known from the table of observed data.

- Since $\hat{A}$ items are divided over $p - [q] - 1$ groups and since all groups (even for the complete IPP) contain y_o items, we have

$$y_o = \frac{\hat{A}}{p - [q] - 1} \qquad (IV.13)$$

- Since y_o and k for the complete IPP are now known, we have already

$$a = \frac{y_o}{\log k} \qquad (F_1)$$

- Since every Bradford group contains resp. $r_o, r_o k, r_o k^2, \ldots, r_o k^{p-1}$ sources, the truncated IPP contains

$$\hat{T} = r_o + r_o k + \ldots + r_o k^{p - [q] - 2}$$

  sources (since in the truncated IPP there are $p - [q] - 1$ groups). Hence r_o is also found :

$$r_o = \frac{\hat{T}}{1 + k + \ldots + k^{p - [q] - 2}}$$

$$r_o = \frac{\hat{T}(k-1)}{k^{p - [q] - 1} - 1} \qquad (IV.14)$$

- From this we finally have

$$b = \frac{k - 1}{r_o} \qquad (F_2)$$

and

$$R(r) = a \log (1 + br) \qquad \text{(III.4)}$$

representing now Leimkuhler's function for the complete unknown IPP.

Note :
In the above method one may take $\hat{T} = r'-1$. In fact, the source on rank r' is intermediate between the groups (calculating from the end) [q] + 1 and [q] + 2 (see Fig.IV.10) and it depends on the actual decimals to decide which case to take. Anyway, taking $\hat{T} = r'$ or r'-1 does not make a real difference in the calculations of a and b.

This algorithm will prove to be very accurate (see further) due to the "interpolation" technique.

IV.3.2. Application on a quasi-perfect example

Under perfect example we understand that we take data following from Lotka's function

$$f(n) = \frac{C}{n^2} , \qquad \text{(II.35)}$$

Indeed, as shown in section III.1.2, this function is mathematically equivalent with the classical Leimkuhler function (III.4), hence without a Groos droop. Take for instance the function

$$f(n) = \left(\frac{30}{n}\right)^2 \qquad \text{(IV.15)}$$

(but you may take any law of the form (II.35)).

Since we work only with whole numbers we have to round off the values of f(n) for the different n. Furthermore, we assume that $y_m = 30$. All these assumptions give rise to a distribution which is very much looking as a Leimkuhler one but deviating a little bit from it. Nevertheless we will see that the above described methodology

works fine. The "observed" data (conform with the above assumptions) are :

# sources	corresponding # articles	r	R(r) ("observed")
1	30	1	30
1	29	2	59
1	28	3	87
1	27	4	114
1	26	5	140
1	25	6	165
2	24	8	213
2	23	10	259
2	22	12	303
2	21	14	345
2	20	16	385
2	19	18	423
3	18	21	477
3	17	24	528
4	16	28	592
4	15	32	652
5	14	37	722
5	13	42	787
6	12	48	859
7	11	55	936
9	10	64	1026
11	9	75	1125
14	8	89	1237
18	7	107	1363
25	6	132	1513
36	5	168	1693
56	4	224	1917
100	3	324	2217
225	2	549	2667
900	1	1449	3567

Table IV.16 : Quasi-perfect example $f(n) = (\frac{30}{n})^2$

Suppose we cut (here articifially, since there is no Groos droop and furthermore we know the "exact" data) at $r = 70$. Hence, the production at this rank is $n = 9$. Take $p = 10$. So $k = (1.781 \times 30)^{0.1} = 1.49$. Formula (IV.11) gives

$$q = \frac{\gamma + \log n}{\log k} = 6.96$$

Hence $[q] = 6$ and $[q] + 1 = 7$.
Formula (IV.12) now gives

$$n' = \frac{k^7}{e^\gamma} = 9.15$$

So the final cut-off rank is (production 9 being between the ranks 65 and 75)

$$r' = 75 - 0.15\ (75 - 64)$$

$$r' \approx 73$$

So $\hat{T} = 73$ and $\hat{A} = 1125 - 18 = 1107$. So

$$y_o = \frac{1107}{3} = 369$$

$$a = \frac{y_o}{\log k} = 925.3$$

$$r_o = \frac{\hat{T}(k-1)}{k^3 - 1} = 15.5$$

$$b = \frac{k - 1}{r_o} = 0.0316$$

Hence Leimkuhler's function is

$$R(r) = 925.3 \log (1 + 0.0316\ r) \qquad \text{(IV.16)}$$

This, compared with the "observed" data, shows a very good fit as is seen from table IV.17 and is confirmed through a Kolmogorov-Smirnov test.

Note :
The attentive reader might remark that the above method also predicts the size of the complete IPP, namely the last calculated value of R(r) in the above table : 3558 which is very close to the (in practical cases unknown) 3567. However, the situation is not that simple, since, if we have an incomplete IPP, we have no idea of what the highest rank (here 1449) will be. Nevertheless, later on we will give a partial solution to this problem.

Let us now see what happens with real data : we will investigate the previously studied bibliographies which show a certain degree of Groos droop.

r	R(r) (original)	R(r) (calculated)
1	30	29
2	59	57
3	87	84
4	114	110
5	140	136
6	165	161
8	213	209
10	259	254
12	303	297
14	345	339
16	385	379
18	423	417
21	477	471
24	528	522
28	592	586
32	652	647
37	722	717
42	787	782
48	859	854
55	936	932
64	1026	1023
75	1125	1124
89	1237	1238
100	1314	1319
107	1363	1367
120	1441	1450
132	1513	1520
150	1603	1617
168	1693	1704
200	1821	1842
224	1917	1933
250	1995	2023
275	2079	2101
300	2145	2174
324	2217	2239
400	2369	2418
500	2569	2611
549	2667	2692
600	2718	2770
700	2818	2906
800	2918	3025
900	3018	3130
1000	3118	3224
1100	3218	3310
1200	3318	3388
1300	3418	3460
1400	2518	3527
1449	3567	3558

Table IV.17 : "Cut-off" fitting of the quasi-perfect example

IV.3.3. Application to "Mast Cell"

We refer again to Fig.IV.3. There, a visual inspection shows a small Groos droop. We estimate visually (but we do not have to be very exact here for the method to work!) that the droop really starts at a rank about $r = 170$. Hence $n = 3$ (we follow the data from section IV.1.5). Take $p = 10$. Hence $k = 1.611$. We find $q = \frac{\gamma + \log n}{\log k} = 3.51$. Hence $[q] = 3$ and $[q] + 1 = 4$. n' is determined by $n' = \frac{k^4}{e^\gamma} = 3.782$. Hence the final rank at which we cut off will be $r' = 169 - 0.782\ (169 - 134) \approx 142$. So $\hat{T} = 142$ and $\hat{A}$ is immediately determined from the data in Table IV.9 : $\hat{A} = 1789$. Hence $y_o = \frac{1789}{6} = 298.2$, $a = \frac{298.2}{\log k} = 625.3$, $r_o = \frac{\hat{T}(k-1)}{k^6-1} = 5.264$ and $b = \frac{k-1}{r_o} = 0.116$. This yields the following Leimkuhler function :

$$R(r) = 625.3 \log (1 + 0.116r) \qquad \text{(IV.17)}$$

When compared to formula (IV.5) (Leimkuhler's function for the global Mast Cell literature) we see that the value of a is larger for the truncated data. This is a requirement of course, since we have cut away the Groos droop. We refer to Fig.IV.3 (curve marked by dots •) for a comparison between the observed and calculated data (truncated for $r < 142$). The fit is almost perfect now and really follows the observed curve as long as the Groos droop is not present. Intuitively we see now that if we extrapolate formula (IV.17) for $r > 142$ we will follow the data of the completed (unknown) bibliography or, at least give an upper bound to it. But, as mentioned in the previous note, it is not clear what is the maximal rank to use. Obviously the present maximal rank is not high enough. This will be solved later on.

IV.3.4. Application to "Schistosomiasis"

Visual inspection of Fig.IV.4 shows that a cut not larger than rank $r = 450$ gets rid of most of the Groos droop.

So n = 4. For p = 10, we find k = 1.889. q is then $q = \frac{\gamma + \log n}{\log k} = 3.087$. So [q] = 3 and [q]+1 = 4. Then we find n' by $n' = \frac{k^4}{e^\gamma} = 7.15$. Hence (production 7 being for the sources between ranks 224 and 250), we finally take the rank

$$r' = 250 - 0.15\ (250-223)$$

$$r' \approx 246 = \hat{T}$$

From table IV.12 we now find $\hat{A} = 7210 - 28 = 7182$. Hence $y_o = \frac{7182}{6} = 1197$, $a = \frac{1197}{\log k} = 1881.9$, $r_o = \frac{\hat{T}(k-1)}{k^6 - 1} = 4.92$ and $b = \frac{k-1}{r_o} = 0.181$, yielding the function

$$R(r) = 1881.9 \log (1 + 0.181r) \qquad \text{(IV.18)}$$

See Fig.IV.4 (the dotted curve) and remark the very close fit.

IV.3.5. Application to Pope's bibliography

The Groos droop is very heavy here (see Fig.IV.5) and we cannot really cut it away completely. We propose to cut at about rank r = 185 (although better fits might be obtained when cutting at a rank 50 or so; we leave this exercise to the reader). Here n = 6. Take p = 10. Then k = 1.848. Now q = 3.86, hence [q] = 3 and [q]+1 = 4, n' = 6.55. So we finally take the rank
r' = 189 - 0.55 (189-175) ≈ 181. Hence $\hat{T}$ = 181 and so, by table IV.14, $\hat{A}$ = 5947. So $y_o = 991.2$, $a = 1614.0$, $r_o = 3.95$ and b = 0.215. This yields the function

$$R(r) = 1614 \log (1 + 0.215r) \qquad \text{(IV.19)}$$

We see from Fig.IV.5 (dotted line) that the fit is much better than when working with the complete bibliography. As said before, better fits are possible when cutting off earlier : indeed, as we can see in Fig.IV.5, the Groos droop is also present in the ranks before 181.

IV.3.6. Application to Sachs' bibliography

Amongst the bibliographies that we studied so far, the Sachs bibliography is the one with the largest Groos droop. So we have to cut early in the bibliography to eliminate as much as possible of this droop. We want to cut at r = 16. Hence n = 15. For p = 10 is k = 1.606. So q = 6.93 and hence [q] = 6, and [q]+1 = 7. So n' = 15.47. The only choice we have here is r' = $\hat{T}$ = 16 and hence $\hat{A}$ = 525. Since we had to cut early in the bibliography, there are not many sources attached to a certain production; in our case : there is only one source with a production 15 while the next more productive source has a production 18. So the "rule of three", as mentioned in the general methodology is not really applicable. Indeed, the choice of $\hat{T}$ = 16 gives a value of q' = 7.07 and not exactly [q]+1 = 7 (if there were a lot of sources this "fine-tuning" can always be done, as in all the previous bibliographies). We modify our formulae in this case to

$$y_o = \frac{\hat{A}}{p - q'} \qquad \text{(IV.20)}$$

(replacing formula (IV.13))
and

$$r_o = \frac{\hat{T}}{1 + k + \ldots + k^{p-q'-1}} \qquad \text{(IV.21)}$$

(replacing formula (IV.14)).
in order to correct for this little difficulty, because of a very early truncation (only then, is such a problem encountered).
Formulae (IV.20) and (IV.21), applied to our example, give :

$y_o = \frac{\hat{A}}{p - q'} = \frac{525}{2.93} = 179.2$ and $r_o = \frac{\hat{T}}{1 + k + k^{1.93}} = \frac{\hat{T}}{5.10} = 3.14.$

So $a = \frac{y_o}{\log k} = 378.2$ and $b = \frac{k - 1}{r_o} = 0.193$. Finally we found the function

$$R(r) = 378.2 \log (1 + 0.193r) \qquad \text{(IV.22)}$$

As we can see in Fig.IV.6, our generalized algorithm works very well and gives a very close fit.

IV.3.7. Application to Citation Data

R. Rousseau remarked to me that citation data do conform very well with Leimkuhler's function. He compiled the following data :

# journals	corresponding # citations	r	R(r) (observed)
1	3594	1	3594
1	3008	2	6602
1	2144	3	8746
1	1895	4	10641
1	1848	5	12489
1	1643	6	14132
1	1552	7	15684
1	1419	8	17103
1	1411	9	18514
1	1391	10	19905
1	1335	11	21240
1	1093	12	22333
1	1088	13	23421
1	1029	14	24450
1	1018	15	25468
1	957	16	26425
1	938	17	27363
1	854	18	28217
1	844	19	29061
1	819	20	29880
1	810	21	30690
1	688	22	31378
1	638	23	32016
1	629	24	32645
1	618	25	33263
1	590	26	33853
1	559	27	34412
1	524	28	34936
1	474	29	35410
1	427	30	35837
1	426	31	36263
1	413	32	36676
1	410	33	37086
1	408	34	37494
1	406	35	37900
1	395	36	38295
1	380	37	38675
2	374	39	39423
1	369	40	39792
1	367	41	40157
1	361	42	40520
1	347	43	40867
1	338	44	41205
1	330	45	41535
1	324	46	41859
1	323	47	42182
1	312	48	42494

cont.

cont.

# journals	corresponding # citations	r	R(r) (observed)
1	288	49	42782
1	280	50	43062
1	279	51	43341
1	278	52	43619
1	250	53	43869
1	249	54	44118
1	247	55	44365
1	245	56	44610
1	233	57	44843
2	223	59	45289
2	222	61	45733
1	209	62	45942
1	208	63	46150
1	200	64	46350
1	190	65	46540
1	189	66	46729
1	164	67	46893
1	161	68	47054
1	152	69	47206
1	141	70	47347
1	125	71	47472
1	107	72	47579
1	102	73	47681
1	91	74	47772
1	84	75	47856
1	67	76	47923
1	44	77	47967
1	41	78	48008
1	16	79	48024

Table IV.18 : Rousseau's Citation data : mathematics journals from SCI, 1985

See Fig.IV.11 (the curve drawn by plus signs +). Indeed, the observed curve ressembles very much a Leimkuhler function. Rousseau asked me if the above theory could be applied to table IV.18.

Note first that, apart from the few last journals in this bibliography, there is no Groos droop. This is due to the fact that a citation analysis deals with a very high number of items per source (i.e. citations per journal) and that the lower number of citations (the less important journals) cannot be determined (this is due to the way the Science Citation Index is constructed). In any case

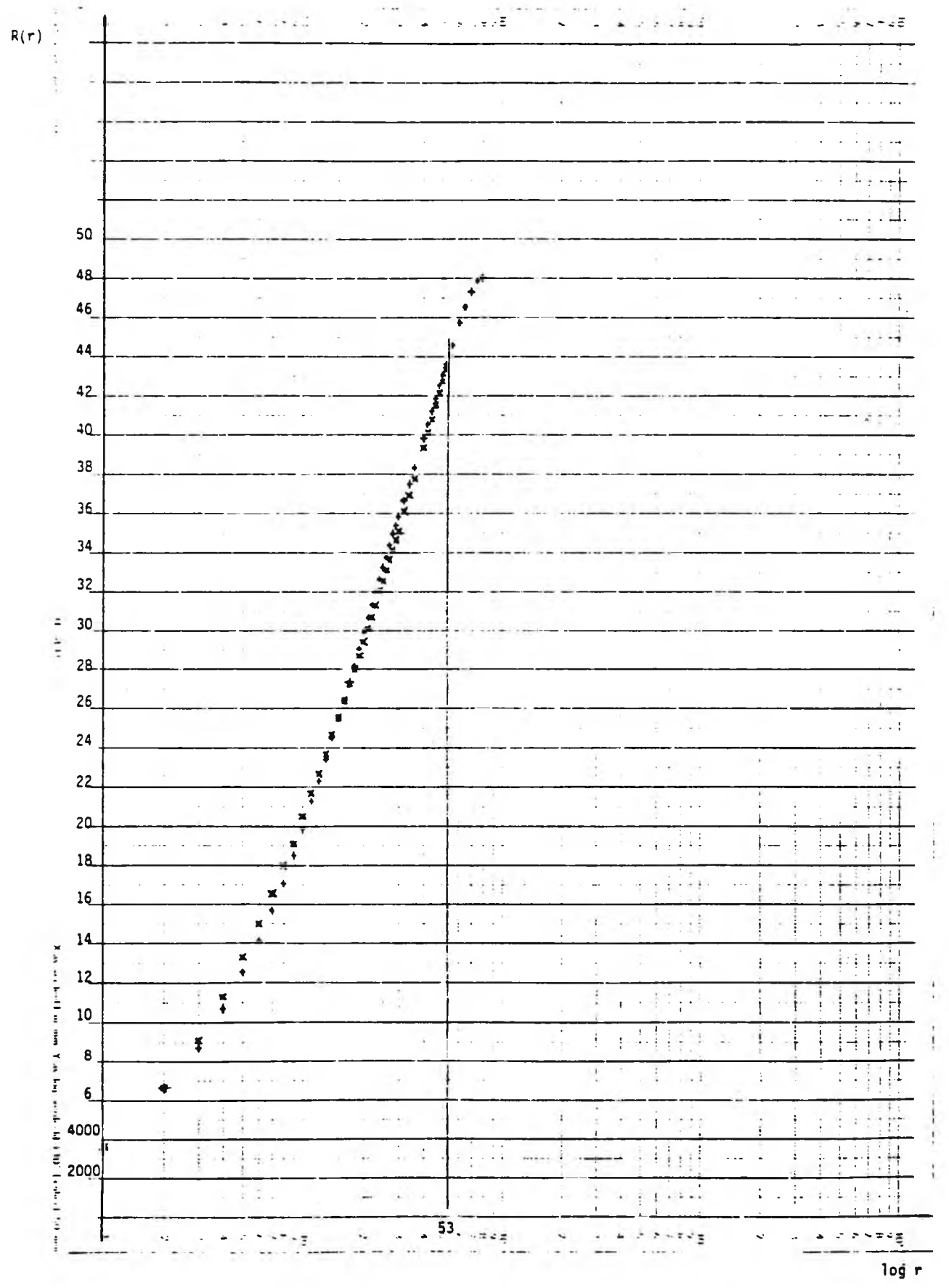


Fig.IV.11 : "Cut-off" fitting of Rousseau's data

+ = Observed

x = Calculated (Egghe)

it would be almost impossible to compile citation data up to the least important journals (number of citations equal to one). In Rousseau's data the least important journal still has 16 citations, the second to least important journal has even 41 citations.

Data, as appearing in table IV.18, can be handled by our methodology for unknown bibliographies, as explained in section IV.3.1. But, since it is typically the case for citation data, that (since the number of citations is high) we have not so many journals attached to a certain number of citations (usually even only one - see table IV.18) and furthermore, there can be a certain distance (larger than one) between two consecutive numbers of citations (f.i. the distance between the number of citations between ranks 75 and 76 still is 17), we have to apply the modified formulas (IV.20) and (IV.21) from section IV.3.6. As in IV.3.6, we will show that they also apply very well for Rousseau's data. We start now the calculation of the best Leimkuhler function for Rousseau's data.

Although small, there is a Groos droop, starting at about $r = 61$, hence $n = 222$. Put $p = 10$, so $k = (1.781 \times y_m)^{0.1} = 2.4023$ (since we deal here with higher data values we will keep a few more decimals than before). So $q = \frac{\gamma + \log n}{\log k} = 6.823$; hence $[q] = 6$. Therefore $n' = \frac{k^7}{e^\gamma} = 259.25$. Our only choice is $r' = 53 = \hat{T}$ (with $\hat{A} = 43869$) and $n' = 250$. Since, as mentioned before, we could not pick the exact source corresponding exactly to the lowest rank in the seventh last group we adapt the correction as explained in section IV.3.6. Hence $q' = \frac{\gamma + \log n'}{\log k} = 6.9586$. So $p' - q' = 3.0414$. Formula (IV.20) yields : $y_o = \frac{\hat{A}}{3.044} = 14424$. So $a = \frac{y_o}{\log k} = 16457.6$. Formula (IV.21) gives now

$$r_o = \frac{\hat{T}}{1 + k + k^{2.0414}} = 5.6464$$

and hence

$$b = \frac{k - 1}{r_o} = 0.248355$$

Leimkuhler's function is therefore

$$R(r) = 16457.6 \log (1 + 0.248355r) \qquad (IV.23)$$

A glance at Fig.IV.11 shows immediately that the fit is very close. Of course, if necessary, a truncation at r = 20 will produce even better fits (up to r = 20) and gives an even more exact form of Leimkuhler's function, being the underlying Leimkuhler function for the (unknown) complete bibliography of citation data.

General note :
More correct calculations can be done by keeping more decimals in the above calculations. We checked this problem but present here a degree of accuracy beyond which only small alterations occur.

We close by noting that Rousseau (1987b) was able to apply the above fitting methods (and, more particularly paragraph IV.3) to his theory on p-nuclei. This theory is based on the exact form of the Leimkuhler function and graph without a Groos droop; hence the above "truncated" fitting method could be applied.

IV.4. An upper estimation of the complete IPP from a given (incomplete) one

Finding a method to determine if an IPP is complete or not, and if it is not complete, giving an algorithm to calculate the size of the complete IPP has been studied in the past cf. (Brookes, 1969 and 1981).

If the complete IPP conforms with Leimkuhler's function but with a Groos droop-deviation, we have shown how to "cut away" the Groos droop and estimate the parameters of the underlying exact Leimkuhler function. If incompleteness

would be the only cause of the Groos droop we would hence have a perfect method to estimate the completion of an IPP, given incomplete data.

However, by the notes in section IV.2.8, incompleteness is not (necessarily) the only cause of the Groos droop : interdisciplinarity or long time periods are possibly also a reason.

So, in general, we can say that the fitted Leimkuhler function R, calculated in the previous paragraph IV.3, gives an upper bound of the complete IPP (if not an exact estimate of it, in case of unidisciplinarity).

Once the "exact" Leimkuhler function R(r) for the complete data has been determined, there remains the problem of where to cut this graph : i.e. if we know T = the total number of sources in the complete IPP, then of course A = R(T) = the total number of items in the complete IPP is known. But determining T is the hardest part since, talking in fractions of the total now, we know a large fraction of A but only a small fraction of T. In (Brookes, 1969) it is argued that we may stop whenever

$$R(r) - R(r-1) \approx 1 \qquad \text{(IV.24)}$$

since we do not allow increments of R(r) smaller than one. Although this requirement looks evident, there is no rationale for it. Furthermore, in (Brookes, 1981), algorithm (IV.24) is changed into : stop if

$$R(r) - R(r-1) = \frac{1}{b} \qquad \text{(IV.25)}$$

where b is the parameter appearing in $R(r) = a \log (1 + br)$. I do not see any rationale for this.

We may conclude that the completion of an IPP still is a problem.

Our method, as developed in the previous section, gives a trivial application in the solution of the above mentioned problem, at least as far as an upper bound is concerned.

IV.4.1. Methodology

Choose p not too small (take e.g. p = 10). In the previous section we calculated the number of items $\hat{A}$ corresponding to p - [q] - 1 Bradford groups, if p denotes the total number of Bradford groups of the complete IPP. We found

$$y_o = \frac{\hat{A}}{p - [q] - 1} \qquad (IV.13)$$

Since y_o is also the total number of items divided by p groups, we find simply

$$\frac{\hat{A}}{p - [q] - 1} = \frac{A}{p}$$

Hence

$$A = \frac{p}{p - [q] - 1} \hat{A} \qquad (IV.26)$$

which gives a simple, but exact solution to the calculation of A.

Note :
It is not even necessary to work with [q]+1. Also decimal q's can be used if necessary (see the examples).

But, using (F_{15}) and (F_{16})

$$A \approx C \ (\log y_m + \gamma) \qquad (F_{15})$$

and

$$T \approx C \frac{\pi^2}{6} \qquad (F_{16})$$

gives

$$\frac{6T}{\pi^2} = \frac{A}{\log y_m + \gamma}$$

So

$$T = \frac{\pi^2 A}{6\,(\log y_m + \gamma)} \qquad \text{(IV.27)}$$

Formulae (IV.26) and (IV.27) solve completely the problem of determining an upper bound for the size of the complete IPP. The average production of the sources that we have missed is given by

$$\hat{\mu} = \frac{A - \hat{A}}{T - \hat{T}} \quad , \qquad \text{(IV.28)}$$

in so far that incompleteness is the only cause of the Groos droop.

To show that the easy method as described by formulae (IV.26) and (IV.27) is good, we will apply it first on the "quasi-perfect" example of section IV.3.2.

IV.4.2. Application to the quasi-perfect example

We re-use the "quasi-perfect" example of section IV.3.2, (which is a complete IPP according to Lotka's function) in order to show that formulae (IV.26), (IV.27) and (IV.28) are very accurate. We have for p = 10, $k = (1.781 \times y_m)^{0.1} = 1.4886$. Suppose we know the above IPP only until about production n = 11. The ranks with n = 11 are 48, 49, 50, 51, 52, 53, 54 and 55. We take the average : $\hat{T} = r' = 51.5$ with a cumulative production $\hat{A} = 897.5$ (keep all decimals in order to be more accurate). Furthermore

$$q' = \frac{\gamma + \log n}{\log k} = 7.478$$

Hence

$$A = \frac{10\,\hat{A}}{10 - 7.478} = 3558$$

to be compared with the real value 3567 (a value which is unknown in practical examples). (IV.27) now yields :

$$T = \frac{\pi^2\ 3558}{6\ (\log\ 30 + \gamma)} = 1471$$

being close to the actual 1449 number of sources. Formula (IV.28) shows that

$$\hat{\mu} = \frac{A - \hat{A}}{T - \hat{T}} = 1.874$$

the average production of the "missed" sources.

We are now going to apply the method to the calculation of an upper bound of the completion of the bibliographies (that we have encountered so far) which show a Groos droop.

IV.4.3. Completion of practical bibliographies (upper estimates)

IV.4.3.1. Mast Cell

From section IV.3.3 and formulas (IV.26), (IV.27) and (IV.28) we find as upper estimates :

$$A = \frac{p}{p - q^r}\ \hat{A} = \frac{10}{6} \cdot 1789 = 2982$$

and

$$T = \frac{\pi^2 A}{6\ (\log\ y_m + \gamma)} = 1029$$

as compared to the actual size of the bibliography of resp. 2378 articles and 587 journals. Furthermore, the average number of articles per missed journal is estimated as

$$\hat{\mu} = 1.367$$

We may say that we have only missed journals with a low number of articles.

IV.4.3.2. Schistosomiasis

From section IV.3.4, we have

$$A = \frac{10}{6} . 7182 = 11970$$

and

$$T = \frac{\pi^2 . 11970}{6 (\log 325 + \gamma)} = 3095$$

to compare with the actual size of the bibliography of resp. 9914 articles and 1738 journals. Here

$$\hat{\mu} = 1.515$$

showing that the Schistosomiasis bibliography could be a bit less complete than Mast Cell. But we refer also to the comments in section IV.2.8.(b), concerning this bibliography.

IV.4.3.3. Pope

From section IV.3.5, we have A = 9912, T = 2655 and $\hat{\mu}$ = 1.547 while the actual size of Pope's bibliography is 7368 articles and 1011 journals.

IV.4.3.4. Sachs' bibliography

From section IV.3.6, we have A = 1792, T = 622 and $\hat{\mu}$ = 2.013 while the actual size of Sachs' bibliography (which was not meant to be complete!) is 828 articles and 143 journals. So this bibliography is very incomplete, the value of $\hat{\mu}$ being a measure for it.

IV.4.3.5. Rousseau's citation data

From paragraph IV.3.7 we have

$$A = \frac{10}{3.0414} . 43869 = 144239 \text{ (\# citations now)}$$

$$T = \frac{\pi^2 A}{6 (\log 3594 + \gamma)} = 27072$$

as compared to 48024 citations and 79 journals in the actual bibliography. Furthermore $\hat{\mu} = 3.72$.

Note :
A total of 27072 mathematical journals is not reasonable but completing Rousseau's bibliography (which is never possible in practise!) involves also non-mathematical journals which sometimes publish a mathematical paper (that receives then some citations). Anyway, the completion problem for citation data is, of course, less meaningful than it is for bibliographies consisting of journals, and articles in these journals.

IV.4.4. Estimating the number of missing sources in every category of production

In paragraph IV.4.1 we described a method to upper estimate A and T, the total number of items, resp. sources in the complete, unknown IPP. The missing sources are, of course, the least productive ones. But how to upper estimate how many sources are missing with 1 item, with 2 items, with 3 items, ...? This can be done by the following rationale : By formula (F_{15}) it follows that

$$C = \frac{A}{\log y_m + \gamma} \qquad (F_{15})$$

So

$$f(n) = \frac{C}{n^2} = \frac{A}{(\log y_m + \gamma)\, n^2} \qquad (IV.29)$$

Hence, the upper estimate for A, as given in the previous paragraphs, gives also the number f(n), an upper estimate of the number of sources (in the completed IPP) with n items (n = 1,2,3,...).

We examine the already used bibliographies.

1. Mast Cell

From IV.4.3.1 we deduce

$$f(n) = \frac{2982}{(\log 66 + \gamma)\, n^2} = \frac{625.6}{n^2}$$

This gives the following table, to be compared with the observed values.

n	f(n)	observed	Δ
1	626	328	298
2	156	90	66
3	70	35	35
4	39	24	15
5	25	16	9
6	17	8	9
7	13	8	5
8	10	6	4
9	8	11	-3

Table IV.19 : Upper estimation of Mast Cell

We notice, of course, the higher n, the more complete is the bibliography.

2. Schistosomiasis

From IV.4.3.2 we deduce

$$f(n) = \frac{1881.8}{n^2}$$

yielding the following table

n	f(n)	observed	Δ
1	1882	908	974
2	470	266	204
3	209	137	72
4	118	76	42
5	75	57	18
6	52	44	8
7	38	27	11
8	29	29	0
9	23	19	4
10	19	14	5

Table IV.20 : Upper estimation of Schistosomiasis

3. Pope

From IV.4.3.3 , we find

$$f(n) = \frac{1613.9}{n^2}$$

yielding the table

n	f(n)	observed	Δ
1	1614	534	1080
2	403	140	263
3	213	68	145
4	101	45	56
5	65	35	30
6	45	14	31
7	33	20	13
8	25	12	13
9	20	8	12
10	16	7	9
11	13	9	4
12	11	9	2
13	10	10	0

Table IV.21 : Upper estimation of Pope

4. Sachs' bibliography

From IV.4.3.4 , we find

$$f(n) = \frac{378.4}{n^2}$$

yielding the table

n	f(n)	observed	Δ
1	378	73	305
2	95	21	74
3	42	8	34
4	24	5	19
5	15	5	10
6	11	4	7
7	8	4	4
8	6	2	4
9	5	3	2

Table IV.22 : Upper estimation of Sachs

We now come to the final section of this chapter : the fitting of the Leimkuhler function, and, together with it, the fitting of the general Lotka function.

IV.5. Fitting of the generalised Leimkuhler and Lotka functions

This final section is devoted to the fitting of the general Leimkuhler function

$$R(r) = \frac{C}{2-\alpha}\left[y_m^{2-\alpha} - \left(y_m^{1-\alpha} - \frac{1-\alpha}{C}\, r\right)^{\frac{2-\alpha}{1-\alpha}}\right] \qquad \text{(IV.30)}$$

$r = 1,2,\ldots,T$
(see (III.5) or (III.65) together with (PF_7) until (PF_{11})).
Of course this gives immediately the problem of fitting Lotka's function ($\alpha > 1$) :

$$f(j) = \frac{C}{j^{\alpha}} \qquad \text{(IV.31)}$$

$j = 1,2,\ldots,y_m$ (see (III.1)).

Remark that, if $\alpha < 2$, the function (IV.30) has a Groos droop. Other functions have been proposed to fit this phenomenon but usually with complicated, unexplained distributions; see e.g. (Griffith, 1988) and (Sichel, 1986).

On fitting Lotka's function (for general α) there have been several papers : (Nicholls, 1986), (Nicholls, 1987), (Pao, 1982), (Pao, 1985), (Pao, 1986) and (Tague and Nicholls, 1987). In them a few methods to derive a good α have been constructed, some better than others.

Let us immediately remark that the whole problem of fitting (IV.30) and (IV.31) reduces to finding the "best" α. Once α determined, C follows as indicated below :

$$T = \sum_{j=1}^{y_m} f(j)$$

$$T = C \sum_{j=1}^{y_m} \frac{1}{j^{\alpha}}$$

where $\alpha > 1$. Since $\sum_{j=1}^{\infty} \frac{1}{j^{\alpha}}$ converges ($\alpha > 1$) we have

$$C \approx \frac{T}{\zeta(\alpha)} \quad , \qquad \text{(IV.32)}$$

where $\zeta(\alpha)$ denotes the classical zèta-function. Since T is known, C can then be determined from a table of $\zeta(\alpha)^{-1}$, appearing for instance in (Nicholls, 1987), but reproduced here since we need it further on : see Table IV.23.

Since also y_m is known, we now see that, once α is known, all parameters in (IV.30) and (IV.31) are known.

In the sequel, we will suffice by investigating whether some α and C that yield a well fitting Lotka function (IV.31) (to the practical data) will also yield a well fitting general Leimkuhler function. We will give three examples.

α	C|T	α	C|T	α	C|T	α	C|T	α	C|T
1.50	0.3828	1.90	0.5715	2.30	0.6981	2.70	0.7848	3.10	0.8450
1.51	0.3885	1.91	0.5753	2.31	0.7007	2.71	0.7866	3.11	0.8463
1.52	0.3942	1.92	0.5791	2.32	0.7033	2.72	0.7883	3.12	0.8475
1.53	0.3998	1.93	0.5828	2.33	0.7058	2.73	0.7901	3.13	0.8488
1.54	0.4054	1.94	0.5865	2.34	0.7083	2.74	0.7918	3.14	0.8500
1.55	0.4109	1.95	0.5902	2.35	0.7108	2.75	0.7935	3.15	0.8512
1.56	0.4163	1.96	0.5938	2.36	0.7133	2.76	0.7952	3.16	0.8524
1.57	0.4217	1.97	0.5974	2.37	0.7157	2.77	0.7969	3.17	0.8536
1.58	0.4270	1.98	0.6009	2.38	0.7181	2.78	0.7986	3.18	0.8547
1.59	0.4323	1.99	0.6044	2.39	0.7205	2.79	0.8003	3.19	0.8559
1.60	0.4375	2.00	0.6079	2.40	0.7229	2.80	0.8019	3.20	0.8571
1.61	0.4427	2.01	0.6114	2.41	0.7252	2.81	0.8035	3.21	0.8582
1.62	0.4478	2.02	0.6148	2.42	0.7276	2.82	0.8052	3.22	0.8593
1.63	0.4528	2.03	0.6182	2.43	0.7299	2.83	0.8068	3.23	0.8605
1.64	0.4578	2.04	0.6215	2.44	0.7322	2.84	0.8083	3.24	0.8616
1.65	0.4628	2.05	0.6249	2.45	0.7344	2.85	0.8099	3.25	0.8627
1.66	0.4677	2.06	0.6281	2.46	0.7367	2.86	0.8115	3.26	0.8638
1.67	0.4725	2.07	0.6314	2.47	0.7389	2.87	0.8130	3.27	0.8649
1.68	0.4773	2.08	0.6346	2.48	0.7411	2.88	0.8145	3.28	0.8660
1.69	0.4821	2.09	0.6378	2.49	0.7433	2.89	0.8161	3.29	0.8670
1.70	0.4868	2.10	0.6409	2.50	0.7454	2.90	0.8176	3.30	0.8681
1.71	0.4914	2.11	0.6441	2.51	0.7476	2.91	0.8191	3.31	0.8691
1.72	0.4961	2.12	0.6472	2.52	0.7497	2.92	0.8205	3.32	0.8702
1.73	0.5006	2.13	0.6502	2.53	0.7518	2.93	0.8220	3.33	0.8712
1.74	0.5051	2.14	0.6533	2.54	0.7539	2.94	0.8235	3.34	0.8723
1.75	0.5096	2.15	0.6563	2.55	0.7560	2.95	0.8249	3.35	0.8733
1.76	0.5140	2.16	0.6593	2.56	0.7580	2.96	0.8263	3.36	0.8743
1.77	0.5184	2.17	0.6622	2.57	0.7600	2.97	0.8277	3.37	0.8753
1.78	0.5227	2.18	0.6651	2.58	0.7620	2.98	0.8299	3.38	0.8763
1.79	0.5270	2.19	0.6680	2.59	0.7640	2.99	0.8305	3.39	0.8772
1.80	0.5313	2.20	0.6709	2.60	0.7660	3.00	0.8319	3.40	0.8782
1.81	0.5355	2.21	0.6737	2.61	0.7680	3.01	0.8333	3.41	0.8792
1.82	0.5397	2.22	0.6766	2.62	0.7699	3.02	0.8346	3.42	0.8801
1.83	0.5438	2.23	0.6793	2.63	0.7718	3.03	0.8360	3.43	0.8811
1.84	0.5479	2.24	0.6821	2.64	0.7737	3.04	0.8373	3.44	0.8820
1.85	0.5519	2.25	0.6848	2.65	0.7756	3.05	0.8386	3.45	0.8830
1.86	0.5559	2.26	0.6875	2.66	0.7775	3.06	0.8399	3.46	0.8839
1.87	0.5599	2.27	0.6902	2.67	0.7793	3.07	0.8412	3.47	0.8848
1.88	0.5638	2.28	0.6929	2.68	0.7811	3.08	0.8425	3.48	0.8857
1.89	0.5677	2.29	0.6955	2.69	0.7830	3.09	0.8438	3.49	0.8866

Table IV.23 : Table of $\frac{C}{T} = \frac{1}{\zeta(\alpha)}$ for $\alpha \in [1.50, 3.49]$ with increments of 0.01

α	C|T	α	C|T	α	C|T	α	C|T	α	C|T
1.50	0.3828	1.90	0.5715	2.30	0.6981	2.70	0.7848	3.10	0.8450
1.51	0.3885	1.91	0.5753	2.31	0.7007	2.71	0.7866	3.11	0.8463
1.52	0.3942	1.92	0.5791	2.32	0.7033	2.72	0.7883	3.12	0.8475
1.53	0.3998	1.93	0.5828	2.33	0.7058	2.73	0.7901	3.13	0.8488
1.54	0.4054	1.94	0.5865	2.34	0.7083	2.74	0.7918	3.14	0.8500
1.55	0.4109	1.95	0.5902	2.35	0.7108	2.75	0.7935	3.15	0.8512
1.56	0.4163	1.96	0.5938	2.36	0.7133	2.76	0.7952	3.16	0.8524
1.57	0.4217	1.97	0.5974	2.37	0.7157	2.77	0.7969	3.17	0.8536
1.58	0.4270	1.98	0.6009	2.38	0.7181	2.78	0.7986	3.18	0.8547
1.59	0.4323	1.99	0.6044	2.39	0.7205	2.79	0.8003	3.19	0.8559
1.60	0.4375	2.00	0.6079	2.40	0.7229	2.80	0.8019	3.20	0.8571
1.61	0.4427	2.01	0.6114	2.41	0.7252	2.81	0.8035	3.21	0.8582
1.62	0.4478	2.02	0.6148	2.42	0.7276	2.82	0.8052	3.22	0.8593
1.63	0.4528	2.03	0.6182	2.43	0.7299	2.83	0.8068	3.23	0.8605
1.64	0.4578	2.04	0.6215	2.44	0.7322	2.84	0.8083	3.24	0.8616
1.65	0.4628	2.05	0.6249	2.45	0.7344	2.85	0.8099	3.25	0.8627
1.66	0.4677	2.06	0.6281	2.46	0.7367	2.86	0.8115	3.26	0.8638
1.67	0.4725	2.07	0.6314	2.47	0.7389	2.87	0.8130	3.27	0.8649
1.68	0.4773	2.08	0.6346	2.48	0.7411	2.88	0.8145	3.28	0.8660
1.69	0.4821	2.09	0.6378	2.49	0.7433	2.89	0.8161	3.29	0.8670
1.70	0.4868	2.10	0.6409	2.50	0.7454	2.90	0.8176	3.30	0.8681
1.71	0.4914	2.11	0.6441	2.51	0.7476	2.91	0.8191	3.31	0.8691
1.72	0.4961	2.12	0.6472	2.52	0.7497	2.92	0.8205	3.32	0.8702
1.73	0.5006	2.13	0.6502	2.53	0.7518	2.93	0.8220	3.33	0.8712
1.74	0.5051	2.14	0.6533	2.54	0.7539	2.94	0.8235	3.34	0.8723
1.75	0.5096	2.15	0.6563	2.55	0.7560	2.95	0.8249	3.35	0.8733
1.76	0.5140	2.16	0.6593	2.56	0.7580	2.96	0.8263	3.36	0.8743
1.77	0.5184	2.17	0.6622	2.57	0.7600	2.97	0.8277	3.37	0.8753
1.78	0.5227	2.18	0.6651	2.58	0.7620	2.98	0.8299	3.38	0.8763
1.79	0.5270	2.19	0.6680	2.59	0.7640	2.99	0.8305	3.39	0.8772
1.80	0.5313	2.20	0.6709	2.60	0.7660	3.00	0.8319	3.40	0.8782
1.81	0.5355	2.21	0.6737	2.61	0.7680	3.01	0.8333	3.41	0.8792
1.82	0.5397	2.22	0.6766	2.62	0.7699	3.02	0.8346	3.42	0.8801
1.83	0.5438	2.23	0.6793	2.63	0.7718	3.03	0.8360	3.43	0.8811
1.84	0.5479	2.24	0.6821	2.64	0.7737	3.04	0.8373	3.44	0.8820
1.85	0.5519	2.25	0.6848	2.65	0.7756	3.05	0.8386	3.45	0.8830
1.86	0.5559	2.26	0.6875	2.66	0.7775	3.06	0.8399	3.46	0.8839
1.87	0.5599	2.27	0.6902	2.67	0.7793	3.07	0.8412	3.47	0.8848
1.88	0.5638	2.28	0.6929	2.68	0.7811	3.08	0.8425	3.48	0.8857
1.89	0.5677	2.29	0.6955	2.69	0.7830	3.09	0.8438	3.49	0.8866

Table IV.23 : Table of $\frac{C}{T} = \frac{1}{\zeta(\alpha)}$ for $\alpha \in [1.50, 3.49]$ with increments of 0.01

IV.5.1. Example : The Pao data on computational musicology

These data can be found in (Pao, 1979) and are as in the left part of Table IV.24.

r	R(r) observed	R(r) calculated
1	40	35.8
2	74	65.2
3	95	90.4
4	111	112.5
5	125	132.2
6	138	150.0
7	151	166.3
8	164	181.3
9	176	195.3
10	188	208.3
11	200	220.6
12	212	232.1
13	222	243.1
14	232	253.5
15	242	263.4
16	252	272.9
17	260	282.0
18	268	290.7
19	276	299.1
20	283	307.1
26	325	350.3
36	383	408.1
46	433	454.7
56	475	494.0
66	515	528.1
76	545	558.4
86	575	585.6
96	605	610.3
106	627	633.1
126	667	673.9
146	707	709.8
166	747	741.8
186	775	770.8
206	795	784.4
226	815	809.9
250	839	849.1
300	889	899.5
350	939	943.3
400	989	982.3
450	1039	1017.4
500	1089	1049.4

Table IV.24 : The Pao data on computational musicology

The maximum likelihood method in (Nicholls, 1986) gives $\alpha = 2.2000$ and $C/T = 0.6709$, giving a good fit of Lotka's function. The fit for Leimkuhler's function with this α and C is as in the right part of Table IV.24.

The fit is very good. Using the Kolmogorov-Smirnov test one has that the maximal relative deviation is

$$D_{max} = 0.0412$$

while the critical value (at the 5 % level) is approximately (incidentally) $\frac{1.36}{\sqrt{1089}} = 0.0412$. Hence, the model

$$R(r) = \frac{0.6709 \times 500}{-0.2} [40^{-0.2} - (40^{-1.2} + \frac{1.2}{0.6709 \times 500} r)^{0.1667}]$$

$$R(r) = -1667.25 [0.4782 - (0.0120 + 0.0036r)^{0.1667}] \qquad (IV.33)$$

is accepted.

IV.5.2. Example : The Murphy data

They can be found in (Murphy, 1973) but also in (Pao, 1986) or (Rao, 1980) - see Table IV.25. For these data, the least square method of Nicholls (1986) yields $\alpha = 2.104$ and $C/T = 0.6424$. Lotka's function is hereby well-fitted. With this α and C we also have a good fit to (IV.30).

Here $D_{max} = 0.0665$ but the 5 % critical value is approximately $\frac{1.36}{\sqrt{238}} = 0.0882$. Again we can accept our general Leimkuhler function :

$$R(r) = \frac{0.6424 \times 170}{-0.1047} [5^{-0.1047} - (5^{-1.1047} + \frac{1.1047}{0.6424 \times 170} r)^{0.0948}]$$

$$R(r) = -1043.05 [0.8449 - (0.1690 + 0.0101r)^{0.0948}] \qquad (IV.34)$$

r	R(r) observed	R(r) calculated
1	5	4.9
2	9	9.5
3	13	13.9
4	17	18.1
5	21	22.1
6	25	26.0
7	29	29.7
8	33	33.3
9	37	36.7
10	40	40.0
11	43	43.3
12	46	46.4
13	49	49.4
14	52	52.3
15	55	55.2
16	58	57.9
17	61	60.6
18	64	63.2
19	66	65.8
20	68	68.3
30	88	90.2
40	108	108.2
50	118	123.6
70	138	148.9
90	158	169.3
110	178	186.6
130	198	201.5
150	218	214.7
170	238	226.5

Table IV.25 : The Murphy data

IV.5.3. Example : The Radhakrishnan-Kerdizan data

They can be found in (Radhakrishnan and Kerdizan, 1979), see also (Pao, 1986) and Table IV.26.

In this case the Nicholls least squares method yields $\alpha = 3.4880$, $C/T = 0.8864$ and the maximum likelihood method (also of Nicholls) gives $\alpha = 3.4000$, $C/T = 0.8782$. Both methods give a fit to Lotka's function (IV.31) (although not splendid) but a very bad fit to Leimkuhler's function (IV.30). In this case I propose another simple method : Estimate C by

$$f(1) = C \qquad \text{(IV.35)}$$

Here this is (not indicated in Table IV.26) : 250.

r	R(r) observed	R(r) calculated
1	7	6.4
2	13	12.0
3	18	17.0
4	22	21.6
5	26	25.8
6	30	29.8
7	34	33.5
8	38	37.1
9	41	40.4
10	44	43.7
11	47	46.8
12	50	49.8
13	53	52.7
14	56	55.5
15	59	58.2
20	69	70.8
30	89	92.3
40	109	110.7
50	129	127.0
100	180	191.5
200	280	283.5
300	380	354.2
301	381	

Table IV.26 : The Radhakrishnan-Kerdizan data

This gives

$$\frac{C}{T} = \frac{250}{T} = \frac{250}{301} = 0.8306$$

and with this (using Table IV.23)

$$\alpha = 2.9907$$

With these values I do not only get a good fit of Leimkuhler's function but the fitted Lotka function

$$f(j) = \frac{250}{j^{2.9907}} \qquad \text{(IV.36)}$$

is better than Nicholls least square (LS) or maximum likelihood (ML) methods in (Nicholls, 1986). For Lotka's fitting I obtain $D_{max} = 0.0151$ which is smaller than Nicholls' fits :

LS : $D_{max} = 0.0367$
ML : $D_{max} = 0.0285$

For Leimkuhler's general function (IV.30) I obtain D_{max} : 0.086 (much better than Nicholls), which is at about the 1 % level. Hence we have, at the 1 % level a fit (contrary to Nicholls). We have here the law

$$R(r) = \frac{0.8378 \times 301}{-1.0442} \left[7^{-1.0442} - \left(7^{-2.0442} + \frac{2.0442}{0.8378 \times 301} r \right)^{0.5108} \right]$$

$$R(r) = -241.49 \left[0.1311 - (0.0187 + 0.0081r)^{0.5108} \right] \qquad \text{(IV.37)}$$

We need further investigation on the value of the above simple method.

IV.5.4. Conclusion

In general we can say that the new function (IV.30) is a valuable law and can easily be fitted. Further research is in order to determine what is the best fitting method; some preliminary calculations (such as in IV.5.3, but there have been other calculations) show that the simple method, described in IV.5.3 is not worse than the more sophisticated methods (least square or maximum likelihood) of Nicholls.

It is furthermore very well possible to fit (IV.30) and (IV.31) with the same set (α, C), a result which looks more evident than it is.

CHAPTER V : CONCLUDING COMMENTS AND SUMMARY OF THE RESULTS

In chapter I, I have explained that, having considered the great variety of techniques, mainly statistical, that have been applied in the past to empirical data within the field of informetrics, techniques which appear to tackle each problem of fitting in some different ad hoc way, I thought that it might be possible to device a mathematical framework which would accomodate all informetric problems of this kind. If this could be done, the present wide variety of analytical techniques would be replaced by one unified mathematical calculus.

In chapters II and III, this general mathematical framework was worked out by applying the 'duality principle' which enabled me to fit all the well-known empirical laws within the framework. It has to be remembered that informetric data are compiled from many sets of sources and their corresponding items - a long and tedious task (now aided by the computer of course). So the published results have to be accepted as 'given' - there is no chance of verifying them as there usually is in the physical sciences. So, when it comes to fitting mathematical formulae to empirical data of this kind, some degree of misfit must be expected.

However, the only practicable test which can be applied to the general mathematical framework is to test it in as many ways as possible against the data-sets already well-known, fitted by other techniques, in the hope of demonstrating that it provides fits as least as well as the various ad hoc techniques already published. The details of these tests are described in chapter IV. I claim that they demonstrate the success of the mathematical approach.

But I must also point out that as this is the first attempt of its kind, my approach has been exploratory.

I regard the present work not as an end in itself but as a beginning which suggests further lines of research. It may be possible to refine what has already been done by extending the duality approach from two-dimensional to three-dimensional informetrics as soon as I can find suitable sets of data to work on.

A further reservation I have to make concerns 'goodness of fit'. I have made some use of the Kolmogorov-Smirnov test when possible, but all the conventional methods of testing for goodness of fit are based on Gaussian statistics and I have doubts about their suitability for the Zipfian distributions of informetrics. This is however a very deep problem which has not yet been resolved.

I will now briefly repeat the most important results, giving always reference to the section number and/or the formula number.

V.1. Duality

The dual of the continuous IPP $(S,I,V) = ([0,T],[0,A],V)$ is the IPP

$$(I,S,U) = ([0,A],[0,T], U) \qquad (II.5)$$

where

$$U(i) = T - V^{-1}(A-i) \qquad (II.6)$$

Notation :

$$\sigma(i) = U'(i) \qquad (II.7)$$

$$\rho(r) = V'(r) \qquad (II.8)$$

for every $i \in I$ and $r \in S$.
One has

$$\sigma(i) = \frac{1}{\rho(A-i)} \qquad (II.10)$$

for every $i \in I$.

One has <u>pure duality</u> if

$$\sigma(i) = C.\rho(i) \qquad (II.12)$$

with $C > 0$ a constant, $i \in I$.

<u>Bradford's law</u> is the only known law that satisfies this property (section II.3.3). Analogous results are true for discrete IPP's (section II.3.2).

<u>Basic dual relations</u>

$$\text{(items)} \quad \int_1^{\rho(i)} f(j)\, j\, dj = i \qquad (II.20)$$

$$\text{(sources)} \quad \int_0^{i} \sigma(A-j)dj = \int_1^{\rho(i)} f(j)dj \qquad (II.19)$$

for every $i \in I = [0,A]$.

From this one finds, for Lotka's α :

$$\alpha < \frac{C}{A} + 2 \qquad (II.31)$$

and hence, in practice,

$$\alpha < 3 \qquad (II.34)$$

V.2. <u>Classification of informetric functions and formulae found as a consequence of it</u>

From the previous duality theory one finds

<u>If $\alpha = 2$</u> (cf. II.4.3.1, III.1.2 and III.3.1)

We have equivalency of the following laws :

1. Lotka's function ($\alpha = 2$) :

$$f(j) = \frac{C}{j^2} \quad , \qquad \text{(II.35)}$$

$j \in [1, \rho(A)] = [\rho(0), \rho(A)]$

2. The group-free Bradford function :

$$\sigma(i) = M.K^i \quad , \qquad \text{(II.18) or (III.7)}$$

$i \in I$.

3. The group-dependent version of Bradford's law (parameters r_o, y_o, k) (III.6)

4. Mandelbrot's function ($\beta' = 1$) :

$$g(r) = \frac{G}{1 + Hr} \quad , \qquad \text{(III.3)}$$

$r \in S$.

5. Leimkuhler's function :

$$R(r) = a \log (1 + br) \quad , \qquad \text{(II.36) or (III.4)}$$

$r \in S$.

In this case one has the relations (putting $\rho(A) = y_m$; cf. (A_2) in III.2) :

$$a = \frac{y_o}{\log k} = \frac{1}{\log k} \qquad (F_1) \text{ and } (F_{11})$$

$$b = \frac{k - 1}{r_o} = \frac{\log K}{M} \qquad (F_2) \text{ and } (F_{12})$$

$$G = y_m = ab \qquad (F_3)$$

$$H = \frac{y_m}{C} \qquad (F_4)$$

$$C = a \qquad (F_5)$$

$$y_o = C \log k \qquad (F_6)$$

$$r_o = \frac{C}{y_m}(k-1) \qquad (F_7)$$

$$K = k^{p/A} \qquad (F_8)$$

(k is in fact p-dependent : k = k(p))

$$y_o = \frac{A}{p} \qquad (F_9)$$

$$r_o = \frac{T(k-1)}{k^p - 1} \qquad (F_{10})$$

In the discrete case, one may use

$$\sum_{j=1}^{y_m} \frac{C}{j} \approx C(\log y_m + \gamma) \ , \qquad (III.71)$$

since y_m is high (cf. (A_3) in III.2).

$$k \approx (e^{\gamma} y_m)^{\frac{1}{p}} \ , \qquad (F_{13})$$

where $\gamma \approx 0.5772$ (Euler's number) and hence

$$K \approx (e^{\gamma} y_m)^{\frac{1}{A}} \qquad (F_{14})$$

Furthermore

$$A \approx C\ (\log y_m + \gamma) \qquad (F_{15})$$

$$T \approx C\ \frac{\pi^2}{6} \qquad (F_{16})$$

In case we put

$$C = y_m^c \qquad (III.77)$$

for a certain $c \in [1,2]$, one finds

$$H = y_m^{1-c} \qquad (PF_1)$$

$$a = y_m^c \qquad (PF_2)$$

$$y_o = y_m^c \log k \qquad (PF_3)$$

$$r_o = y_m^{c-1}(k-1) \qquad (PF_4)$$

$$A \approx y_m^c (\log y_m + \gamma) \qquad (PF_5)$$

$$T \approx y_m^c \frac{\pi^2}{6} \qquad (PF_6)$$

<u>If α is general (> 1)</u> (cf. II.4.3.2, III.1.3 and III.3.2)

We have equivalency of the following functions :

1. Lotka's function

$$f(j) = \frac{C}{j^{\alpha}} \quad , \qquad (II.41) \text{ or } (III.1)$$

$j \in [1, \rho(A)]$.

2. The general group-free Bradford function

$$\sigma(i) = (A_1 + iA_2)^{A_3} \quad , \qquad (II.44) \text{or } (III.8)$$

$i \in I$.

3. Mandelbrot's function

$$g(r) = \frac{G}{(1 + Hr)^{\beta'}} \quad , \qquad (III.3)$$

$r \in S$.

4. The generalised Leimkuhler function,

$$R(r) = B_1 (B_2 - (B_3 + B_4 r)^{B_5}) \qquad (II.45) \text{ or } (III.5) \text{ or } (III.65)$$

$r \in S$.

In this case one has the relations :

$$G = y_m \qquad (GF_1)$$

$$H = \frac{\alpha - 1}{C\ y_m^{1-\alpha}} \qquad (GF_2)$$

$$\beta' = \frac{1}{\alpha - 1} \qquad (GF_3)$$

$$A_1 = y_m = \frac{A(2-\alpha)}{C} + 1 \qquad (GF_4)$$

(using again (A_2) in III.2)

$$A_2 = -\frac{2 - \alpha}{C} \qquad (GF_5)$$

$$A_3 = -\frac{1}{2 - \alpha} \qquad (GF_6)$$

$$B_1 = \frac{C}{2 - \alpha} \qquad (GF_7)$$

$$B_2 = y_m^{2-\alpha} \qquad (GF_8)$$

$$B_3 = y_m^{1-\alpha} \qquad (GF_9)$$

$$B_4 = \frac{\alpha - 1}{C} \qquad (GF_{10})$$

$$B_5 = \frac{2 - \alpha}{1 - \alpha} \qquad (GF_{11})$$

V.3. Further formulae

VI.3.1. The average μ in function of the Bradford factor k

$$\mu \approx \frac{6p}{\pi^2} \log k \qquad (III.80)$$

were k = k(p) is the group-dependent factor.

The function

$$f_p(k) = \frac{k}{\mu} \qquad (III.81)$$

hence has the properties :

$$\lim_{k \to +\infty} f_p(k) = \lim_{k \to 1 \atop >} f_p(k) = +\infty \qquad \text{(III.83) and (III.84)}$$

$$f_p'(e) = 0$$

f_p is in a very long interval around e, nearly horizontal with value

$$f_p(e) = \frac{\pi^2 e}{6p} \qquad \text{(III.86)}$$

Consequently, for $p > 5$: $k < \mu$ almost always. k might be $>\mu$ for p = 3 or 4. Experiments agree with these findings.

V.3.2. The number of items m(i) in the most productive source in every Bradford group i (counted from right to left)

A duality argument gives

$$\frac{1}{i} \sum_{j=1}^{m(i)-1} \frac{1}{j} < \log k < \frac{1}{i} \sum_{j=1}^{m(i)} \frac{1}{j} \qquad \text{(III.92)}$$

Consequently

$$m(i) \approx \frac{k^i}{e^\gamma} \approx 0.5615\ k^i \qquad \text{(III.93)}$$

from $m(i) > 7$ on (i.e. i > 2 or 3 is enough in most cases).

These results have nice applications in fitting techniques further on.

V.4. Classifying Zipf's law

We have equivalency of the following informetric laws (cf. III.4.2) :

1. The graphical formulation of Bradford's law, group-dependent (parameters r_1, y_o, k_1) with $r_1 = k_1$.

2. The graphical formulation of Bradford's law, group-free

$$\Sigma(i) = M_1 . K_1^i - 1 \quad , \tag{III.94}$$

(where, in general :

$$\Sigma(i) = \int_0^i \sigma(i')\, di' = r) \tag{III.95}$$

for $i \in I$, with $M_1 = 1$.

3. Brookes' function

$$R_1(r) = \alpha \log (\beta(1+r)) \tag{III.96}$$

where $r \in S$, with $\beta = 1$.

4. Zipf's (Pareto's) function

$$g(r) = \frac{F}{1 + r} \tag{III.97}$$

$r \in S$.

In this case we have the relations :

$$\alpha = \frac{1}{\log K_1} = F \tag{III.98}$$

$$K_1 = k_1^{p/A} \tag{III.99}$$

$$r_1 = K_1^{y_o} = k_1 \tag{III.100}$$

We have also the corollary : Zipf's law (Pareto's law) is the only law that agrees with both the verbal and the graphical form of Bradford's law (and this in the group-dependent as well as in the group-free version).

Zipf's function represents a highly elitary situation.

V.5. Fittings

Fitting procedures have been carried out for
- Bradford's law (verbal, group-dependent)
- Leimkuhler (known IPP's)
- Leimkuhler (unknown IPP's)
- General Lotka together with general Leimkuhler.

V.5.1. Fitting of Bradford's law, p groups ($i \in \mathbb{N}$, $p \geq 3$)

The method is based on

$$k \approx (1.781 y_m)^{1/p} \qquad (F_{13})$$

$$y_o = \frac{A}{p} \qquad (F_9)$$

and

$$r_o = \frac{T(k-1)}{k^p - 1} \quad . \qquad (F_{10})$$

p can be chosen . Experiments have been carried out on several "classical" bibliographies, with good results. A mistake of Goffman-Warren has been corrected.

IV.5.2. Leimkuhler (known bibliographies)

Methodology (cf. IV.2)

$$R(r) = a \log (1 + br) \qquad (III.4)$$

with

$$a = \frac{y_o}{\log k} \qquad (F_1)$$

$$b = \frac{k - 1}{r_o} \qquad (F_2)$$

and k as in (F_{13}).

Although k = k(p), the method (to find (III.4)) is p - independent.

V.5.3. Leimkuhler (unknown bibliographies)

We refer here to the problem of practical IPP's, that show a Groos droop.

Groos droops occur e.g. if (cf. IV.2.8) :
- we have an incomplete IPP
- we have a merged IPP.

Since incompleteness can be a partial reason for the occurrence of a Groos droop, we can present upper estimates of the completed IPP.

Methodology (cf. IV.3) : dual technique and "cutting away" at the point q where the Groos droop becomes apparent. We interpolate until our cut-off point is at the connection [q]+1 of 2 Bradford groups. The first (non-affected) part is then modelled more exactly. If [q]+1 groups are cut away and $\hat{A}$ items resp. $\hat{T}$ sources are left in the first p - [q] - 1 groups, we have

$$y_o = \frac{\hat{A}}{p - [q] - 1} \qquad (IV.13)$$

$$a = \frac{y_o}{\log k} \qquad (F_1)$$

$$b = \frac{k - 1}{r_o} \qquad (F_2)$$

where

$$r_o = \frac{\hat{T}(k-1)}{k^{p-[q]-1} - 1} \qquad (IV.14)$$

This gives the "complete" function

$$R(r) = a \log (1 + br) \qquad (III.4)$$

Examples are given and, indeed, the fittings of the first part of the graphs are much better, when a Groos droop is present.

As a corollary we can upper estimate T and A (cf. (IV.4.1)) :

$$A = \frac{p}{p - [q] - 1}\,\hat{A} \qquad \text{(IV.26)}$$

and with this,

$$T = \frac{\pi^2 A}{6\,(\log y_m + \gamma)} \qquad \text{(IV.27)}$$

The average number of items per source, of the missed sources is

$$\hat{\mu} = \frac{A - \hat{A}}{T - \hat{T}} \qquad \text{(IV.28)}$$

Since

$$C = \frac{A}{\log y_m + \gamma} \qquad (F_{15})$$

we can hence also estimate the Lotka function

$$f(n) = \frac{C}{n^2} = \frac{A}{(\log y_m + \gamma)n^2} \qquad \text{(IV.29)}$$

for every $n \in \mathbb{N}$.

V.5.4. General Lotka together with general Leimkuhler (cf. (IV.5))

One has

$$f(j) = \frac{C}{j^{\alpha}} \qquad \text{(IV.31)}$$

with

$$C \approx \frac{T}{\zeta(\alpha)} \qquad \text{(IV.32)}$$

(table IV.23)
and

$$R(r) = \frac{C}{2 - \alpha}\left[y_m^{2-\alpha} - \left(y_m^{1-\alpha} - \frac{1 - \alpha}{C}\,r\right)^{\frac{2-\alpha}{1-\alpha}}\right] \qquad \text{(IV.30)}$$

The good Nicholls-fittings of (IV.31) yield also good fittings of (IV.30). A new, simple, ad hoc, method is presented, in case the Nicholls-fittings are not so good.

REFERENCES

Throughout the next references we will use the following abbreviations for the heaviest used journals :

JD = Journal of Documentation
JIS = Journal of Information Science
JASIS = Journal of the American Society for Information Science
SC = Scientometrics

ADENAIKE, B.O. (1982)
Bibliometric studies on a protein-rich crop-the cowpea. JIS 4 : 117-121.

AIYEPEKU, W.O. (1977)
The Bradford distribution theory : the compounding of Bradford periodical literatures and geography. JD 33(3) : 210-219.

ALLISON, P.D., DE SOLLA PRICE, D., GRIFFITH, B.C., MORAVCSIK, M.J. and STEWART, J.A. (1976)
Lotka's law : a problem in its interpretation and application. Social Studies of Science 6 : 269-276.

APOSTOL, T.M. (1974)
Mathematical analysis. (Addison-Wesley).

ASAI, I. (1981)
A general formulation of Bradford's distribution : the graph-oriented approach. JASIS 32(2) : 113-119.

AVRAMESCU, A. (1980)
Theoretical foundation of Bradford's law. International Forum on Information and Documentation 5(1) : 15-22.

Brookes, B. C. (1984),
Towards informetrics: Haitun, Laplace, Zipf,
Bradford and the Alvey programme.
J. Doc. 40(2), 120-143.

BONITZ, M. and SCHMIDT, P. (1982)
Transition from the macrolevel to the microlevel of information at rank distribution investigations of the report literature of an international information system. SC 4(4) : 283-295.

BOOKSTEIN, A. (1984)
Robustness properties of bibliometric distributions. Unpublished manuscript.

BRADFORD, S.C. (1934)
Sources of information on specific subjects. Engineering 137 : 85-88 (Reprinted in JIS 10(4) : 176-180, 1985).

BROOKES, B.C. (1969)
Bradford's law and the bibliography of Science. Nature 224 (dec.6) : 953-956.

BROOKES, B.C. ( )
Philosophy of Science. Nature

BROOKES, B.C. (1973)
Numerical methods of bibliographical analysis. Library Trends 22(1) : 18-43.

BROOKES, B.C. (1977)
Theory of the Bradford law. JD 33(3) : 180-209.

BROOKES, B.C. (1980)
Information space. L'Espace informatique. Canadian Journal of Information Science 5 : 199-211.

BROOKES, B.C. (1981)
A critical commentary on Leimkuhler's "exact" formulation of the Bradford law. JD 37(2) : 77-88.

BROOKES, B.C. (1985)
"Sources of information on specific subjects" by S.C. Bradford. JIS 10(4) : 173-180.

BROWN, P. (1977)
The distribution of articles in the literature. Australian Academic Research Libraries 8(1) : 26-32.

BURRELL, Q.L. (1988)
Modelling the Bradford phenomenon. JD 44(1) : 1-18.

BURRELL, Q.L. and CANE, V.R. (1982)
The analysis of library data (with discussion). Journal of the Royal Statistical Society A145(4) : 439-471.

DE LILLO (1982)
Advanced calculus with applications (New York : Mac Millan).

DROTT, M.C., MANCALL, J.C. and GRIFFITH, B.C. (1979)
Bradford's law and libraries : present applications - potential promise. ASLIB Proceedings 31(6) : 296-304.

EGGHE, L. (1985)
Consequences of Lotka's law for the law of Bradford. JD 41(3) : 173-189.

EGGHE, L. (1986)
The dual of Bradford's law. JASIS 37(4) : 246-255.

EGGHE, L. (1987)
An exact calculation of Price's law for the law of Lotka. SC (1-2) : 81-97.

EGGHE, L. (1988a)
On the classification of the classical bibliometric laws. JD 44(1) : 53-62.

EGGHE, L. (1988b)
Methodological aspects of bibliometrics. Library Science with a slant to Documentation and Information Studies 25(3) : 179-191.

EGGHE, L. (1989a)
Towards a dual theory of bibliometrics. Unpublished manuscript.

EGGHE, L. (1989b)
New Bradfordian laws equivalent with old Lotka laws, evolving from a source-item duality argument. Unpublished manuscript.

EGGHE, L. (1989c)
A note on different Bradford multiplicators. Unpublished manuscript.

EGGHE, L. (1989d)
Applications of the theory of Bradford's law to the calculation of Leimkuhler's law and to the completion of bibliographies. JASIS (to appear).

EGGHE, L. and ROUSSEAU, R. (1986)
A characterization of distributions which satisfy Price's law and consequences for the laws of Zipf and Mandelbrot. JIS 12 : 193-197.

EGGHE, L. and ROUSSEAU, R. (1988)
Reflections on a deflection : a note on different causes of the Groos droop. SC 14(5-6) : 493-511.

GARFIELD, E. (1983)
Citation Indexing : its theory and application in science, technology and humanities (Philadelphia : ISI-Press).

GOFFMAN, W. and MORRIS, T.G. (1970)
Bradford's law and library acquisitions. Nature 226 (june 6) : 922-923.

GOFFMAN, W. and WARREN, K.S. (1969)
Dispersion of papers among journals based on a mathematical analysis of two diverse medical literatures. Nature 221 (march 29) : 1205-1207.

GOFFMAN, W. and WARREN, K.S. (1980)
Scientific information systems and the principle of selectivity (New-York : Praeger).

GRADSHTEIN, I.S. and RYZHIK, I.M. (1965)
Tables of integrals, series and products. (New-York : Academic Press).

GRIFFITH, B.C. (1988)
Exact fits to large ranked, bibliometric distributions. JASIS 39(6) : 423-427.

GROOS, O.V. (1967)
Bradford's law and the Keenan-Atherton data. JD 19(1) : 48.

HAITUN, S.D. (1982a)
The stationary scientometric distributions. Part I. Different approximations. SC 4(1) : 5-25.

HAITUN, S.D. (1982b)
The stationary scientometric distributions. Part II. Non-Gaussian nature of scientific activities. SC 4(2) : 89-104.

HAITUN, S.D. (1982c)
The stationary scientometric distributions. Part III. Role of the Zipf distribution. SC 4(3) : 181-194.

HAITUN, S.D. (1983)
The "rank-distortion" effect and non-Gaussian nature of scientific activities. SC 5(6) : 375-395.

HASPERS, J.H. (1976)
The yield formula and Bradford's law. JASIS 27(5) : 281-287.

HERDAN, G. (1960)
Type-token mathematics. A textbook of mathematical linguistics ('s-Gravenhage : Mouton & Co).

KENDALL, M.G. (1960)
The biliography of operations research. Operations Research Quarterly 11 : 31-36.

KÖRNER, S. (1969)
Experience and theory : an essay in the philosophy of science (London : Routlegde and Kegan).

LEIMKUHLER, F.F. (1967)
The Bradford distribution. JD 23(3) : 197-207.

LIPOTOV, Y.S. and DENISENKO, L.V. (1986)
On the behaviour of information flows in multi-component polymer systems research. SC 9(5-6) : 197-208.

LOTKA, A.J. (1926)
The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16(1-2): 317-323.

MANDELBROT, B. (1954)
Structure formelle des textes et communication. Word 10(1) : 1-27.

MANDELBROT, B.B. (1977)
The fractal geometry of nature (New York : W.H. Freeman and Cy).

MURPHY, L.J. (1973)
Lotka's law in the Humanities? JASIS 24 : 461-462.

NICHOLLS, P.T. (1986)
Empirical validation of Lotka's law. Information Processing and Management 22(5) : 417-419.

NICHOLLS, P.T. (1987)
Estimation of Zipf parameters. JASIS 38 : 443-445.

PAO, M.L. (1979)
Bibliometrics and computational musicology. Collection Management 3 : 97-109.

PAO, M.L. (1982)
Lotka's test. Collection Management 4 : 111-124.

PAO, M.L. (1985)
Lotka's law : a testing procedure. Information Processing and Management 21(4) : 305-320.

PAO, M.L. (1986)
An empirical examination of Lotka's law. JASIS 37(1) : 26-33.

POPE, A. (1975)
Bradford's law and the periodical literature of information science. JASIS 26 : 207-213.

PRAUNLICH, P. and KROLL, M. (1978)
Bradford's distribution : a new formulation. JASIS 29 : 51-55.

PRICE, D. DE SOLLA (1976)
A general theory of bibliometric and other cumulative advantage processes. JASIS 27 : 292-306.

RADHAKRISHNAN, T. and KERDIZAN, R. (1979)
Lotka's law and computer science literature. JASIS 30 : 51-54.

RAO, RAVICHANDRA I.K. (1980)
The distribution of scientific productivity and social change. JASIS 31 : 111-121.

ROUSSEAU, R. (1987a)
Een vleugje bibliometrie : de equivalentie tussen de wetten van Bradford en Leimkuhler. Wiskunde en Onderwijs 13 : 71-78.

ROUSSEAU, R. (1987b)
The nuclear zone of a Leimkuhler curve. JD 43(4) : 322-333.

ROUSSEAU, R. (1988a)
Lotka's law and its Leimkuhler representation. Library Science with a slant to Documentation and Information Studies 25(3) : 150-178.

ROUSSEAU, R. (1988b)
Relations between continuous versions of bibliometric laws. Unpublished manuscript.

SACHS, L. (1986)
A guide to statistical methods and to the pertinent literature. Literatur zur angewandter Statistik (Heidelberg : Springer-Verlag).

SARACEVIC, T. and PERK, L.J. (1973)
Ascertaining activities in a subject area through bibliometric analysis. JASIS 24 : 120-134.

SELEY, H (1968)
The mast cells (London : Butterworth).

SEN, S.K. (1985)
Philosophy of bibliometry. Paper presented at the XVth IASLIC Conference, Bangalore, 1985.

SICHEL, H.S. (1986)
Word frequency distributions and type-token characteristics. Mathematical Scientist 11 : 45-72.

SINGLETON, A. (1976)
Journal ranking and selection : A review in physics. JD 32(4) : 258-289.

TAGUE, J. and NICHOLLS, P. (1987)
The maximal value of a Zipf size variable : sampling properties and relationship to other parameters. Information Processing and Management 23(3) : 155-170.

THEIL, H. (1967)
Economics and Information theory. (Amsterdam : North-Holland).

WARREN, K.S. and NEWILL, V.A. (1967)
Schistosomiasis, a bibliography of the world's literature from 1852-1962. (Cleveland : Western Reserve University Press).

WILKINSON, E.A. (1973)
The Bradford-Zipf distribution. OSTI-report #5172. (London : University College).

WILKINSON, E.A. (1978)
The ambiguity of Bradford's law. JD 28 : 122-130.

YABLONSKY, A.I. (1980)
On fundamental regularities of the distribution of scientific productivity. SC 2(1) : 3-34.

ZIPF, G.K. (1949)
Human behavior and the principle of least effort (Cambridge : Addison-Wesley) (Reprint, New York : Hafner, 1965).