JISSI
The International Journal of Scientometrics and Informetrics

Volume 1 Number 3-4 September - December 1995

Guest Editors :

Hildrun Kretschmer
Wolfgang Glänzel

Special Issue

In This Issue

Selected Papers from the Proceedings of the

Fourth International Conference on Bibliometrics, Informetrics and Scientometrics, Berlin
11-15 September 1993

BIS BRZARK INFORMATION SYSTEMS

JISSI The International Journal of Scientometrics and Informetrics

International Society for Scientometrics and Informetrics (ISSI) has been officially founded on 5 October 1994 in Utrecht, The Netherlands.

Membership information of ISSI is included in this issue of JISSI.

ISSI Membership Application Form

FAMILY NAME :..... ..... MALE/FEMALE*.....

FIRST NAME(s) :..... TITLE(s) :..... .....

.....

INSTITUTION:..... FUNCTION:..... .....

....

ADDRESS: ....

.....

.....

.....

PHONE : ..... EMAIL : .....

TELEFAX : .....

MEMBERSHIP FEE : $20.00
$ 5.00 (for people from developing countries)

CHECK ENCLOSED/AMOUNT REMITTED TO : *) ABN ALN 2A
ABN-AMRO Bank Utrecht
acc.no. 55.57.54.855.STW

*) please cross out which does not apply

Please return the completed form to : ISSI Administration
c/o Technology Foundation(STW)
PO Box 3021
3502 GA UTRECHT
Netherlands

JISSI The International Journal of Scientometrics and Informetrics

Volume 1 Number 3-4 Berlin Conference (1993) Special Issue September-December 1995

Editorial Correspondence to :
Subir K Sen
Honorary Executive Editor
85 (old) Devi Nivas Road
P.O. - Motijhil
Calcutta 700 074, India
Fax : + 91-33-551 2180

Business Correspondence to :
Brzark Information Systems (P) Ltd.
112, Humayun Pur
(Near NCC Office)
Safdarjung Enclave,
New Delhi 110 029, India

Subscription Information :
Subscription rates include
air mail surcharge
For detailed subscription
information see inside
Reduced price
(two-third of the net price)
for advanced subscription
for first three years together
Subscriptions may start from
any time of the year

Payments may be made through :
Demand Draft/ Cheque/Money Order/
Credit Card
payable to
Brzark Information Systems (P) Ltd.
For further enquiry and
for proforma invoice
write to the Publisher

Advertisements :
JISSI offers publication of
selected advertisements from
interested parties.
For rates and other conditions
please contact the publisher

Publishers :
Prabhat Pandey
Bina Pandey

Cover design :
Bibhas Dutta

Price : Single Copy :
US $ 55.00 / Rs. 700.00

Introductory Note From The Exec. Editor

This special issue of JISSI contains the last instalment of selected papers from the Fourth International Conference on Scientometrics, Informetrics and Bibliometrics in 1993.

Publication of this journal from a third-world country like India should be considered as a significant but natural event.

Publication of scholarly technical journals is now in a critical stage. Future of journals is being debated all over the world. As yet, the doom's day for scholarly journal is not within sight. Journals will continue in publication, in whateverforms they may be. There will be on line journals and there will be journals on telefax.It is difficult to envisage individual issues of a highly specialized technical journal on CD-ROM.

The situation is peculiar. In respect of production and utilization of research communications the gap between the North and the South or between so called Developed and Developing countries is increasing despite serious concern of many.

The so called international journals published from developed countries and by the big publishing houses are already out of reach of most of the scholars and libraries of the developing (third-world) countries. The first reason is the high subscription price. Most of the third world countries have utterly incompatible exchange rates of their currencies with those of developed countries. Second reason is the supply channel. In most cases the airmail surcharge is also very high.

Again there is scarcely any provision for harnessing information from or subscribing to online databases or journals in machine readable format for a scientist or a student in a developing country. There is lack of equipments and lack of money.

On the other hand there a re too many journals which need to be consulted by a scientist. The over-all cost of purchasing journals has become tremendous.

At the other side of the game, the scientists from developing countries and non-English-major countries are finding it more and more difficult to get their papers published in truly international good quality journals (excepting a few) not because the contents are not good but because of many other peripheral reasons. Major two reasons are poor quality of language or expression and inability of utilising sophisticated tools, instruments, equipments, printing or graphic facilities etc.

The only possible wayouts seem to be to have new journals on every emerging specialized subject area published from the developing countries. Such journals should primarily be in the hard copy format and they should also be made available in machine readable form and on line.

As it happens, many of the developing countries now have expertise in modern and futuristic publishing and growing facilities for online connections.

This is what exactly we are aiming at with JISSI.

Instead of searching in a dozen or more journals sciento-informetricians may now look for their materials in two or three journals including JISSI.

JISSI is presently being published in hard copy format only. From 1997 we shall be able to provide JISSI on floppies also.

To speed up publication and coverage in secondary services we have already taken necessary steps to receive submissions in machine readable form, get them refereed and edited on floppies, proof corrected by the authors and sending the content pages and abstracts to the secondaryservices through speediest possible channels. But, we shall continue in receiving typed manuscripts and processing them in the best possible manner. We like to assure our prospective authors and peers and subscribers and readers that we shall publish the best quality materials only. And we shall continue in helping authors from non-English-major countries by language editing of their manuscripts by experts.

Subir K Sen

Preface

The International Conference on Bibliometrics, Informetrics and Scientometrics held in Berlin, September 11-15, 1993, and organized under the auspices of the Minister of Science and Research of the Land Berlin, Senator Manfred Erhardt, was the fourth in a series of increasingly prominent biennial conferences. Previous meetings were held in Belgium, 1987, in Canada, 1989, and in India, 1991. The fifth conference will take place in Chicago, 1995, and it is already planned to organize the sixth conference in Israel in 1997.

The Berlin Conference was the largest Bibliometrics meeting so far. About 180 scientists from 32 countries attended the conference. In all, 145 papers were presented as lectures or posters. The conference, however, proved to be a notable event in more respects. For the first time bibliometricians from East and West, North and South took the opportunity for personal contacts and for sharing their scientific views with colleagues from almost all regions of the world.

The program covered many topics of our field. Of course, research evaluation was one of the focal points of the meeting. The great number of "evaluative" studies also reflects the increasing demand for such research results in science policy. A part from applied studies (evaluation of research in selected fields and/or special regions, countries or institutions; analysis of scientific collaboration), a series of methodological papers were presented marking the advances in recent "basic" research in scientometrics. Several theoretical papers addresed topics like information measures, library circulation models, and generalizations of classical bibliometric laws (Bradford's law, regularities of growth and obsolescence of scientific literature, etc.) and proved the quality and importance of informetric research in our field.

The Berlin Conference was held in memory of Derek de Solla Price, one of the pioneers and founders of scientometrics, who died in 1983. In the inaugural address by the honorary chairman of the conference, Eugene Garfield, and in the plenary session commemorating Derek de Solla Price, his personality and the impact of his work on present bibliometric research were stressed. A report on the history of the Price Medal, which is awarded by the journal *Scientometrics* since 1984, and a review of the scientific work of the Price awardees were given.

The plenary session "Bridging the gaps between Bibliometrics, Scientometrics and Informetrics" was a further high-light of the conference. The plenary lectures showed the great concern bibliometricians have for the development and future of their field. In this context the foundation of the International Society for Scientometrics and Informetrics (ISSI) during the conference can be seen, too. The society aims at encouraging communication in our field, coordinating organizational tasks and helping to establish scientometrics and informetrics as a scientific discipline.

We would like to thank all the referees and other persons who have helped in reviewing the some 110 papers submitted for publication in the conference proceedings. The selected papers are published in four separate proceedings volumes. Three of them are special issues of the international journals *Scientometrics, Research Evaluation,* and *International Journal of Science and Science of Science*. The fourth volume, focussed on informetric topics, is now being published in parts by the present journal, *JISSI : The International Journal of Scientometrics and Informetrics.*

Wolfgang Glänzel
Hildrun Kretschmer

Guest Editors
(Editors of the Conference Proceedings)

Message of Greeting
from
the Senator of Science and Research,
Prof. Dr. Manfred Erhardt,
to the 4th World Congress on Bibliometrics, Informetrics and Scientometrics
from 11 to 15 September 1993

Ladies and gentlemen,

I welcome most cordially all participants of your 4th World Congress to Berlin, the old and the new German capital.

I was happy to assume the patronage of your meeting and I pay tribute to the excellence of the founder of your young scientific discipline, Derek de Solla Price, the tenth anniversary of whose death you are commemorating.

I am pleased that you have chosen our city as the venue of your congress for it enables Berlin once again to fulfil its new geopolitical function as a central European metropolis : we are the host and meeting place of scientists from the North and the South, from the East and the West.

My special greetings go to all who are able, after the fall of the Wall, to pursue their international contacts free from external and internal constraints at long last.

About 150,000 students are enrolled at 17 tertiary education institutions in the city of science, Berlin, among them three universities with three teaching hospitals, four colleges of art and nine special technical colleges. There exist, in addition, more than 70 extra-university research institutions.

Berlin is also a city with a rich cultural life and a surrounding countryside worth seeing. I hope you will have time to enjoy both outside your work at the meeting.

I wish your congress every success and I wish all of you a pleasant stay in our city.

Yours sincerely,

Signature
Professor Dr. Manfred Erhardt

Berlin, 30 August 1993

Where Linguistics, Physical Measurement and Social Measurement Converge

David Andrich
School of Education, Murdoch University, Western Australia 6150, Australia

Classification of entities is central to thinking and language, to applied problem solving and to scientific research. The paper is concerned with ordered classes which reflect more or less of some property, and which are used often when no measuring instrument is available. Ordered classes mirror measurement in that they are envisaged to partition a continuum into adjacent classes, and the entities of classification are presumed to be located at points on a continuum. In the prototype of measurement, the continuum is partitioned into equal classes, termed units, and the estimate of the location of the entity is taken simply as a count of the number of units it exceeds from the origin, where the precision of the estimate increases as the size of the unit decreases. Ordered classes also differ from measurements in that their number is finite, they are not presumed to be equal, and uncertainty of classification is built immediately into any model which characterises the classification process. While classes are defined conceptually and operationally, it is known from linguistics that classes are also defined relatively, in that if one of the classes in a semantic space is added or removed, the relative meanings of the other classes change. The paper shows that the probabilistic Rasch model for ordered classifications, which is known to provide invariant estimates of location of entities with respect to different instruments and therefore with respect to different partitions of the continuum when the model holds, also has the surprising features that it is both sensitive to the kind of relative definition of classes understood in linguistics and that it specialises fully as the prototype of measurement when the classes form equal units. In particular, the relative probability of classification in a particular class depends on the size of all classes in the system, and not just the one class and the adjacent classes on either side of it, and in the case of equal classes as in measurement, the best estimate of the location of the entity is simply the count of the classes from the origin and the precision of the estimate increases inversely as the size of the class.

1. Introduction and Summary

Central in everyday language, in formal activities involving applied problem solving, and in scientific research, is classification. By *classification* is meant that classes are conceptualised so that different entities can be assigned to one of the classes. In everyday language this tends to be carried out implicitly, while in scientific work it is carried out very explicitly. This paper is concerned with a particular kind of classification system in which the classes have some kind of ordinal structure among them and in which the classification of an entity in a particular class involves some uncertainty that needs to be characterised formally.

Ordered classification systems are common in the social sciences, and in particular, where measurement of the kind found in the physical sciences would be desirable, but no measuring instrument exists. This paper brings together three lines of study, first, the relationships among ordered classes in a semantic space from a linguistic perspective, second, a probabilistic model of social measurement for ordered classes which belongs to the set of Rasch models[1], third, physical measurement and the structure of models in natural science, and shows how the three converge to be consistent and mutually reinforcing.

The rest of the paper is structured as follows. Section 2 briefly analyses the linguistic relationship among classes, central to which is that classes are defined not only operationally, but also in terms of each other; Section 3 briefly constructs

the Rasch model of social measurement for classifying entities into ordered classes, central to which is that the locations of entities are invariant under different partitions of the continuum; Section 4 briefly articulates the prototype of measurement, central to which is the partitioning of a continuum into mutually exclusive contiguous classes of equal size from a natural origin[a] . To anticipate the argument in the paper, the cornerstone of the relationship among these three lines of study is that the Rasch model for ordered classes both specialises to the case of physical measurement and has the property that location of an entity in any class depends on the operation of all classes. Section 5 summarises this relationship and Section 6 makes it concrete with an analysis of a real data set using the model.

2. Linguistic Classes

Though this paper is not about a theory of classes, some comments are required to clarify the case with which it is concerned. First, the minimum number of meaningful classes is two – a class that contains certain entities and the complementary class that excludes them. Second, and by extension, classes imply both *comparisons* and *contrasts* in the relevant context. Thus entities that belong to mutually exclusive classes, and therefore by definition are different, also have some relevant features in common. For example, if books are defined according to whether they belong to the class of fiction or the complementary class of non-fiction, the entities have the common feature that they are books.[b] The relevant common features, background and context for the contrasts implied by the different classes is termed generally the *semantic space*. Third, classes defined within some semantic space may be related to each other nominally or ordinally, the distinction being that in the former the classes are different in *kind*, while in the latter they are different in *degree*. This paper is concerned with ordered classifications, where the relationship among the classes implies *more or less* of some property, such as more or less correct in a specified sense, more or less relevant to some requirment, and so on. Table 3 shows an example of a set of four ordered classes, termed *inadequate setting, discrete setting, integrated setting,* and *integrated and manipulated setting,* according to which writing samples from students were assessed. While this case is specific, it is presented as prototypic of the kind of classification system with which this paper is concerned.

Classes themselves are usually assigned names, which may be either single words or phrases, and are termed *linguistic* entities. For example, the term *integrated setting* of Table 3 names and represents a property of an essay, and the word *sheep* names and represents a kind of animal. In order to judge whether or not any entity belongs to a class, their conceptual definitions must be made operational, and this is central to applied problem solving and to empirical research in that they direct data collection. It might be expected, especially in scientific work, that the operational definition of classes provides a unique and invariant classification process. However, classes which partition some semantic space are not just defined operationally, but, according to Saussure[2], also relatively. Thus Saussure articulated that linguistic entities in general did not have meanings only in terms of some object that they represented, but that the meanings of words were *defined also in terms of each other*. The former Saussure referred to as the *signification* of the linguistic entity, and the latter he referred to as its *value*.

> Within the same language, all words used to express related ideas limit each other reciprocally; synonyms like French *redouter* 'dread', *criandre* '*fear*' and *avoir peur* 'be afraid' have value only through their opposition: if *redoubter* did not exist, all its content would go to its competitors... . The value of just any term is accordingly defined by its environment; it is impossible to fix even the word signifying 'sun' without first considering its surroundings: in some languages it is not possible to say 'sit in the sun'... (Saussure, 1959[2], p.116).

If this principle of linguistic value is universal, then definitions of all classes should obey the

[a] By *measurement* is meant *fundamental measurement*, sometimes also referred to as additive or conjoint measurement (Wright, 1985; Duncan, 1984), and does not refer to levels of measurement as expressed in Stevens(1945).

[b] For example, in general one would not make an issue of the feature that shoes and books form mutually exclusive classes, simply because in general shoes and books have nothing relevant in common.

principle. This means that within some system of classes, no matter how rigorously defined operationally, if one class were eliminated or added, then it would change the relative operation of the other classes. How they change is an empirical question.

The point to emerge from this analysis is that the operation of classes is in some sense less objective and determinate than might initially be considered: because the operational definition of each class depends in part on the definition of all other classes, the probability of an entity being assigned to each class depends on the definition of all classes, and not just that class. The tasks of the rest of the paper are first to demonstrate that though apparently lacking one kind of objectivity, this feature is consistent with the Rasch model for ordered classifications so that another kind of invariance and objectivity of the location of an entity nevertheless prevails, and second, to lend further credibility to this kind of objectivity, to show that it is entirely consistent with physical measurement.

3. The Rasch Model for Ordered Classifications

The prototype of measurement is the location of an entiy on a real line partitioned by thresholds sufficiently fine that their own width can be ignored, and the use of an *instrument* to locate the entity on this line. Thus in principle, the continuum is divided into mutually exclusive contiguous classes of equal size, the size being called a *unit.* Then the observed location of the entity, usually termed a *measurement,* is the count of the number of classes exceeded from the origin. A distinction is made between the hypothesised location of the entity, and the observed measurements directed towards identifying the location. It is understood that instruments have operating ranges, but in principle, the measurement of an entity is not taken to be a function of the operating range of any one instrument – if an entity is outside the range of an object, then another instrument is sought.

Although the number of categories is finite, ordered classification systems in the social sciences mirror measurement in that a latent continuum is postulated and the classes are located on the continuum. In elementary treatments, the prototype of measurement is followed closely, and the location of the entity is taken to be the count of the category, from the first, in which the entity is located. In more advanced treatments, the sizes of the categories or the intervals are not assumed equal, but are estimated, and a probabilistic element is formalised immediately to the classification process.

The two major models for ordered classifications are those based on the work of Thurstone[3] and Rasch[1], who expressed similar requirements of an invariance data had to meet before they permitted measurement[4]. However, unlike Thurstone, who used populations, Rasch expressed his requirements in terms of a single individual and a single instrument:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; ...(A)
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; ...(Rasch, 1961, p.322). (B)

The mathematical counterpart of these requirements for discrete variables can be expressed in a number of forms [1,5,6], but the one convenient in this exposition is expressed multiplicatively:

$$\Pr\{x_{pi}\} = \frac{1}{\gamma_{pi}}\left(\xi_p^x \Big/ \prod_{k=1}^{x}\omega_{ki}\right),$$

$$\xi_p > 0, \omega_{ki} > 0, \omega_{0i} \equiv 1 \qquad (1)$$

where ξ_p is the location of entity p, ω_{ki}, k=1,m_i are the locations of m_i ordered thresholds of instrument i which partition the continuum into m_i+1 ordered classes ($\omega_{0i} \equiv 1$ for completeness), X_{pi} , $x_{pi} \in\{0,1,2,...m_i\}$ is an integer random variable denoting the $m+1$ successive classes, and $\gamma_{pi} = \sum_{x=0}^{m_i}\left(\xi_p^x \Big/ \prod_{k=1}^{x}\omega_{ki}\right)$ is a normalising factor. Figures 1 and 2, which show the probabilities of classification of a set of essays graded using the classifications of Table 3 as a function of ξ_p , and the more symmetric ln ξ_p , illustrate the operation of Equation (1).

To appreciate how the definitions of classes come to have a relative influence on each other, the model is derived briefly from dichotomous classifications at each threshold k. Consider that the continuum is partitioned by two thresholds ω_{1i} and ω_{2i} where $\omega_{2i} > \omega_{1i}$ and that in the first instance, decisions at these thresholds are made *independently*. If y_{pki} =1 indicates that the classification for entity p with instrument i at threshold k is in the positive direction, and y_{pki} =0 is its complement, then the unique model that satisfies conditions A and B [7,5,8,9] takes the form

$$\Pr\{y_{pki}=0;\ \omega_{ki}\}=\frac{1}{\eta_{pki}};\quad \Pr\{y_{pki}=1;\omega_{ki}\}=\frac{1}{\eta_{pki}}\left(\xi_p/\omega_{ki}\right), \qquad (2)$$

where $\eta_{pki}=1+\left(\xi_p/\omega_{ki}\right)$ is a normalising factor. The probabilities of all possible outcomes are shown in Table 1.

However, in order for there to be a response in only one of three ordered classifications, the responses cannot remain independent, and in particular, if $\omega_{2i} > \omega_{1i}$, then the outcome (0,1), in which the object is classified above the second threshold and below the first one simultaneously, cannot occur. Accordingly, the only legitimate outcomes are those that conform to the Guttman pattern, and if the probabilities are renormalised within the Guttman patterns, and these are successively characterised by the random variable X_{pi} ,$x_{pi} \in \{0,1,2\}$, then the probabilities are as shown in Table 2. These can be generalized to any number of thresholds and identified with Equation (1).

Table 1 : Probabilities of the joint outcomes (y1i, y2i) at thresholds 1 and 2 when considered independently

Outcome	Probabilities
$\{(y_{1i}, y_{2i})\}$	$Pr\{(y_{1i}, y_{2i})\}$
(0, 0)	$(1/\eta_{p1i})(1/\eta_{p2i})$
(1, 0)	$(\xi_p/\omega_{1i})/\eta_{p1i}\eta_{p2i}$
(0, 1)	$(\xi_p/\omega_{2i})/\eta_{p1i}\eta_{p2i}$
(1, 1)	$(\xi_p^2/\omega_{1i}\omega_{2i})/\eta_{p1i}\eta_{p2i}$

Table 2 : Probabilities of the single outcomes (x_{pi}) when the outcomes at the thresholds are constrained to the Guttman pattern

Outcome		Probabilities
$\{(y_{1i}, y_{2i})\}$	$= x_{pi}$	$Pr\{x_{pi}\}$
(0, 0)	= 0	$1.1/\gamma_{pi}$
(1, 0)	= 1	$(\xi_p/\omega_{1i})1/\gamma_{pi}$
(1, 1)	= 2	$(\xi_p^2/\omega_{1i}\omega_{2i})/\gamma_{pi}$
$\gamma_{pi}=\sum_{x=0}^{2}\xi_p^x/\prod_{k=0}^{x}\omega_{ki}$		$Pr\{x_{pi}\}=\frac{1}{\gamma_{pi}}\xi_p^x/\prod_{k=0}^{x}\omega_{ki}$

The key step in this construction is that when the independent outcomes at each threshold are brought together and constrained to the Guttman pattern, the denominator in Equation (1) contains all thresholds, and therefore the probability of being located in any one of the classes becomes a function of the sizes of all classes, and not just one particular class. Further, if two adjacent categories x and $x+1$ are combined to give category x' in Equation (1), then $\Pr\{x'_{pi}\} = \Pr\{x_{pi}\} + \Pr\{(x+1)_{pi}\}$ cannot be expressed in the form of Equation (1) [10,5,6,11.]. These effects are consistent with the relationship between the value and signification of a linguistic entity discussed in Section 2, and they are elaborated in Section 5.

4. Physical Measurement

The function of mesurement in science is central to the formulation of quantitative scientific laws and furthermore, fundamental measurement seems to ensure that these laws take on a simple multiplicative structure [12,13].

> ...virtually all the laws of physics can be expressed as multiplications and divisions of measurement (Ramsay, 1975 [13], p 258).
>
> ...throughout the gigantic range of physical knowledge, numerical laws assume a remarkably simple form, provided fundamental measurement has taken place (Ramsay, [13], p262).

The reason that laws take on such a structure may be that counts of units, which involve successive additions of the unit, are expressed as multiplications. This relates to the property of fundamental measurement that entities may be amalgamated or concatenated, and then the new entity so formed behaves as a single entity with a value equal to the sum of the values of original entities[14]. Measurement of mass exemplifies fundamental measurement: two objects may be amalgamated, and then within Newton's laws of motion, the new object behaves like a single object whose mass is the sum of the masses of the individual objects. It is evident that Equtaion (1) has the multiplicative structure of parameters contained in the laws of physics, and it is shown later in the Section that, when specialised to the protoype of measurement, it immediately characterises concatenation.

Though central to the formulation of physical laws, it is understood that measurement inevitably contains error. In expressions of deterministic theories, these errors are considered sufficiently small relative to the variation in the locations of entities measured that they are ignored. Nevertheless, on replications of measurements, the observed locations of the entity on the continuum are expected to be different, the variation depending on the size of the unit and control of the instrumentation; then these differences in observed locations manifest themselves as different measurements, Equation(1) immediately specifies the random component which represents this error, and it is shown later in the Section that, when specialised to the prototype of measurement, as the size of the unit decreases and the implied precision increases, so the variance of these measurements also decreases.

Suppose now that as in the prototype of measurement, the thresholds ω_{ki} are *equally spaced,* and that the first threshold is at a distance ω_i from the origin of 0. Then ω_i is the unit of instrument i, and the locations of the successive thresholds are given by $\omega_{1i} = 1\omega_i$, $\omega_{2i} = 2\omega_i$, $\omega_{3i} = 3\omega_i$, ... $\omega_{xi} = x\omega_i$, ($\omega_{0i} \equiv 1$). Substituting these values for the thresholds in Equation (1), and upon simplification,

$$\Pr\{x_{pi}\} = \frac{1}{\gamma_{pi}}\left\{\left(\xi_p/\omega_i\right)^x / x!\right\},$$

$$x_{pi} \in \{0,1,2,...m\}, \qquad (3)$$

Where $\gamma_{pi} = \sum_{x=0}^{m}\left\{\left(\xi_p/\omega_i\right)^x / x!\right\}$. If, as in the prototype of measurement, the number of categories is not finite so that in principle $x_{pi} \in \{0,1,2,.....\infty\}$, then $\gamma_{pi} = \sum_{x=0}^{\infty}\left\{\left(\xi_p/\omega_i\right)^x / x!\right\} = exp\{\xi_p/\omega_i\}$ and Equation (3) reduces to the Poisson distribution

$$\Pr\{x_{pi}\} = e^{-\left(\xi_p/\omega_i\right)}\left(\xi_p/\omega_i\right)^x / x!\},$$

$$x_{pi} \in \{0,1,2,...\infty\}. \qquad (4)$$

Thus the distribution of responses is Poisson whose parameter is the *ratio of the loation of the entity* ξ_p and *the size of the unit* ω_i, and therefore the expected value and variance are given by

$$E\left[X_{pi}\right]=V\left[X_{pi}\right]=\xi_p/\omega_i \qquad (5)$$

From this result it is now shown that the model characterises concatenation and that as the unit decreases in size, the variance of the replicated measurements decreases proportionately.

Concatenation. It is well known that the sum of two Poisson distributions is again Poisson whose parameter is the sum of the parameters of the original distributions. Therefore, if two entities p and q are measured using the same instrument i, the probability of the sum of the measurements $y_{(p+q)i}=x_{pi}+x_{qi}$ is given by

$$\Pr\left\{y_{(p+q)i}\right\}=e^{-(\xi_p+\xi_q)/\omega_i}\times$$

$$\left(\left(\xi_p+\xi_q\right)/\omega_i\right)^y/y!,\ y_{(p+q)i}\in\{0,1,2,....\infty\} \qquad (6)$$

with $E\left[Y_{(p+q)i}\right]=\left(\xi_p+\xi_q\right)/\omega_i$.

Thus if two entities are concatenated, the sum of the measurements of the two entities has the same distribution, and therefore the same expected value, as that of an entity whose parameter is the sum of the parameters of the original entities.

Precision. Suppose an instrument n with the unit size $\omega_n=\omega_i/n$ is used[c] to measure the same entity p. Then according to the expectation from Equation (5),

$$E\left[X_{pn}\right]=\xi_p/\omega_n=\xi_p/(\omega_i/n)=n\xi_p/\omega_i \qquad (7)$$

indicating that in the smaller units, the location of the entity is n times that in the original units, as expected (e.g. a measurement in millimetres has 10 times the value of a measurement in centimetres). In the original units of X_{pi}, in which $X_{pn}=X_{pi}/n$ (a measurement of x mm $=x/10$ cm),

$$E\left[X_{pn}\right]=E\left[X_{pi}/n\right]=\frac{1}{n}n\xi_p/\omega_i=\xi^p/\omega_i \qquad (8)$$

showing that the measurements in the unit (1/n)th of the original unit, but *expressed* in the original unit, retains the same expected value. The essential distinction here between obtaining a measurement in a particular unit (reflected by Equation (7)), and then expressing this measurement in a different unit (reflected by Equation (8)). The latter step, *being after the data are collected,* does not change the parameter of the Poisson distribution, only the units in which the observations are expressed, while the former, which pertains to *the data collection,* affects the parameter in terms of the size of the unit in which the data are collected.

Next, according to the variance in Equation (5),

$$V\left[X_{pn}\right]=\xi_p/\omega_n=\xi_p/(\omega_i/n)=n\xi_p/\omega_i \qquad (9)$$

indicating that in the new units the variance in proportionately larger. Expressed in original units however,

$$V\left[X_{pn}\right]=V\left[X_{pi}/n\right]=\frac{1}{n^2}n\xi_p/\omega_i=\frac{\xi_p/\omega_i}{n} \qquad (10)$$

showing that the measurement is made in a unit (1/n)th of the original unit, then the variance of the measurement reduces proportionately by 1/n. This means that as the size of the unit is made smaller, so the variance of measurements is made smaller, and a sufficiently small unit may be reduced sufficiently, and therefore the precision of measurement increased sufficiently, so that the error in any particular measurement may be ignored for certain practical or theoretical purposes, as in deterministic laws of physics[d]. Further, as the parameter of the Poisson increases, it tends to the normal distribution, so as the unit decreases, the distribution of the errors of measurement tend to the normal, which is consistent with what is expected in physical measurement.

Thus the model of Equation (1) specialises to

[c] To be concrete, the reader might take throughout that ω_i is one centimetre and that $n=10$ so that $\omega_n=1$ millimetre.

[d] Note that as the new unit ω_n becomes smaller relative to the original unit w_i, then $V[X_{pn}]=v\ \xi_p/\omega_i$ in the new unit becomes larger. Thus as the unit becomes smaller, and the estimate of ξ_p more precise, so the uncertainty of any particular outcome in this unit becomes greater. In the limit, infinite prescision leads to infinite uncertainty in the infinitely small units, which is also remainiscent of results in physics at the quantum level.

fundamental measurement found in physics: (1) the expression is multiplicative in terms of the location of the entity and the size of the unit, (2) it characterises concatenation, and (3) when expressed in a constant unit, the size of the unit and the variance of measurements decreases proportionately and tends to the normal distribution. While the model specialises literally to fundamental measurement in the case of equal units, its general form provides a generalised form of concatenation, termed simultaneous or additive conjoint measurement[15,16].

5. Specifically Objective Comparisons

It is now possible to consider more closely the kind of invariance and objectivity implied by Equation (1). To begin the appreciation, the special case of Equation (4) when the units are equal is particularly instructive. If two entities p and q are measured by the same instrument i, then according to the expectation in Equation(5)

$$E\left[X_{pi}\right]=\xi_p/\omega_i\ ;\ E\left[X_{qi}\right]=\xi_q/\omega_i$$

from which the ratio

$$E\left[X_{pi}\right]/E\left[X_{qi}\right]=\left(\xi_p/\omega_i\right)/\left(\xi_q/\omega_i\right)=\xi_p/\xi_q$$

is independent of the instrument i. Thus the *comparison* of the locations of p and q is independent of the instrument i, and of the size of the unit of the instrument - thus p and q can be compared using any instrument which belongs to the same class. However, the precision of the comparisons will depend on the size of the units. Likewise, if a single entity p is measured by two instruments i and j, then again according to the expectation in Equation (5),

$$E\left[X_{pi}\right]=\xi_p/\omega_i\ ;\ E\left[X_{pj}\right]=\xi_p/\omega_j$$

from which the ratio

$$E\left[X_{pi}\right]/E\left[X_{pj}\right]=\left(\xi_p/\omega_i\right)/\left(\xi_p/\omega_j\right)=\omega_j/\omega_i$$

is independent of the location of entity p. Thus the units of the instruments can be compared independently of the locations of the entities, and so they can be compared using any entities that belong to the same class.

In the general form of Equation (1), the corresponding expressions are a little more complex. Consider the comparison of two entities p and q: let the subspace $S\left(x_{pi},x_{qi}\right)=\left\{\left(x_{pi},x_{qi}\right),\left(x_{qi},x_{pi}\right)\right\}$ for $x_{pi}\neq x_{qi},x_{pi},x_{qi}\in\left\{0,1,2,\ldots m_i\right\}$ be a subspace of all possible pairs of responses, then it can be shown readily that

$$\Pr\left\{\left(x_{pi},x_{qi}\right)\middle|\left(x_{pi},x_{qi}\right)\right\}=\frac{1}{\gamma_{pqi}}\xi_p^{x_{pi}}\xi_q^{x_{qi}}\ , \qquad (11)$$

where $\gamma_{pqi}=\xi_p^{x_{pi}}\xi_q^{x_{qi}}+\xi_p^{x_{qi}}\xi_q^{x_{pi}}$, which is independent of the instrument i. Equation (11) can be used to estimate the relative locations of entities, though usually the locations of many entities are estimated using the locations of the thresholds of the instruments once these have been estimated independently of the parameters of the entities from the following equation. For entity p, let $r_p=x_{pi}+x_{pj}$, then

$$\Pr\left\{\left(x_{pi},x_{pj}\right)\middle|r_p\right\}=\frac{1}{\gamma_{rpij}}\prod_{k=0}^{x_{pi}}\frac{1}{\omega_{ki}}\prod_{k=0}^{x_{pj}}\frac{1}{\omega_{kj}} \qquad (12)$$

where $\gamma_{rpij}=\sum_{\max(0,r-m_j)}^{\min(r,m_i)}\prod_{k=0}^{x_{pi}}\frac{1}{\omega_{ki}}\prod_{k=0}^{x_{pj}}\frac{1}{\omega_{kj}}$,

is independent of the location of the entity, and can be used as a basis for estimating the locations of the thresholds of the correspondig instruments, which is equivalent to estimating the different partitions of the continuum.

Now consider again the implication of the invariance of the locations of entities with respect to different units. Every observation x_{pi} can be used to estimate ξ_p, itself being immediately an estimate in the case of typical measurement with equal units. The above analysis implies that *on the average,* the entity would be located at the same *point* independently of the size of the classes, and in the case of measurement, this estimate is literally the average of repeated measurements and is independent of the unit of the instrument. The stress on the invariance of *location* is important

because any particular observation x_{pi} also indicates the particular interval into which the entity is classified, and it has been argued in the literature that, rather than invariance of location, the *probabilties* of the observations being in particular intervals should be invariant under different partitions of the continuum. This proposition is central to the case made by Jansen and Roskam, who articulate the *joining assumption* as follows[11]:

> The probability of subject's responding with category j or category k is equal to the probability of respondig with category h if category h replaces the categories j and k. (Jansen and Roskam, 1986,p. 73).

As already indicated, the Rasch model of Equation (1) does not have this property, and therefore neither does the special case of the Poisson distribution in Equation (4), which has been identified here with physical measurement. This special case illustrates relatively readily why the joining assumption cannot hold even in physical measurement – when the continuum is partitioned into smaller units, the variance of the observed distribution changes – it becomes smaller, and so the sum of the probabilities of outcomes in two adjacent classes will not be the same as the probability of an outcome in the larger class composed of the two classes. If it did, no increase in precision would follow from a reduction in the sizes of the units. Another manifestation of the difference between the invariance of locations of Equation (1) and Jansen and Roskam's joining assumption under different partitions of the continuum is that the former implies that the entities can be concatenated while the latter implies that it is the probabilities of the observations that can be concentrated. These concatenations are incompatible.

6. An Example

Although a standard text book perspective is that measurement is an independent and objective procedure from which theories are deduced, Kuhn[12] points that

> In text books, the numbers that result from measurements usually appear as the archetypes of the 'irreducible and stubborn facts' to which the scientist must, by struggle, make his theories conform. By scientific practice, as seen through the journal literature, the scientist often seems rather to be struggling with the facts, trying to force them to conformity with a theory he does not doubt. Quantitative facts cease to seem simply the 'given'. They must be fought for and with, and in this fight the theory with which they are to be compared proves the most potent weapon. Often scientists cannot get numbers that compare well with theory until they know what numbers they should be making nature yield (Kuhn, 1961, p. 193).

Further, he stresses that

> *The road from scientific law to scientific measurement can rarely be traveled in the reverse direction* (Kuhn, 1961, p. 219, emphasis in original).

This implies that the data collected are driven by the theory, and not the other way round. Because the theoretical implications of this paper are unusual, it has proven difficult to find existing data either in general or in the scientometric literature which illustrate these implications. The data used in this paper, which concern the grading of essays, were collected deliberately in order to put the theory to an empirical test, and therefore they are presented illustratively. However, the implcations for data collection in scientometrics is immediately relevant. For example, in the rating of the prestige of journals in some field, the implications are that the probability of classification in any category will depend not only on the operational definition of that category, but also on the definition and number of all of the other categories. For example, a system of rating for journals in some field may involve the following four categories: Category A - *defines the present state of the art;* Category B - *has important influence;* Category C- *reports relevant matters;* Category D- *minor importance,* or the following 5 categories in which the first category is a new category and all of the other categories are identical: Category A1- *creates new directions;* Category A2 - *defines the present state of the art;* Category B - *has important influence;* Category C - *reports relevant matters;* Category D- *is of minor importance.* According to predictions from both linguistic and measurement

theory, the probability that any journal will be classified into any one of the common categories will necessarily be different in the two systems. Further, if data were collected on two different occasions with different numbers of categories, pooling categories in order to make comparisons is not justified. On the surface, this may appear arbitrary – however, if the data conform to the model of equation (1), then the estimates of the relative locations of the journal on a continuum of prestige will be invariant under the two systems of data collection, and so comparisons of relative locations on two different occasions would be meanigful and independent of the classification systems. A more obvious and immediate example concerns the process of peer review of articles for journal publication. These invariably require reviewers to allocate the paper into one of three or four categories ranging from a category such as *publish as is* to *reject*. The implication of the argument in this paper is that if an extra category were added or eliminated from any system, then even if the categories common to the two systems were defined identically, the probability of a paper being classified in a category depends on the definition and number of all categories. Theoretical and empirical studies on these classification systems may lead to a better understanding, and operation, of recommendations based on peer review.

The example is provided by Harris[17], who had eight essays of narrative writing produced by Year 7 and 8 students (ages approximately 11 and 12) in Western Australian schools graded according to the classes shown in Table 3. These essays were produced as part of a larger project on monitoring writing standards, and were chosen because they had been found relatively easy to assess and because they covered a wide range in quality. As can be seen, the classes are defined hierarchically with respect to one feature of the writing, the setting, and the class scored 3 was explicitly defined to be a subset of class scored 2. The graders were students in a class in educational measurement at Murdoch University in Western Australia. The results presented are those from 28 graders who graded the same essays first using all four classes, and then using only the first three classes. In both sets of gradings, the first three classes were defined identically, as shown in Table 3, and the graders had the definitions of the classes in front of them when they graded the essays. The order of the essays was distributed randomly within graders on each occasion, so that different graders read the essays in different order, and they read the essays in different order when judging them with respect to the three classes and with respect to the four classes. For the purpose of analysis according to the model of Equation (1), the 8 essays graded by the 28 graders were treated as (28)(8) = 224 essay/grader contacts, and each contact was made twice, once using the four classes and once using the three classes. Thus the data were pairs of 224 classifications, each governed by the same parameter $\xi_{(ge)} = \xi_p$ where '(ge)' refers to the combination of grader g and essay e. In estimating the locations of the thresholds of the two sets of classifications, these parameters were eliminated according to Equation (12).

Table 3 : Operational definitions of ordered classes for judging essays

0	***Inadequate setting :*** Insufficient or irrelevant information given for the story. Or, sufficient elements may be given, but they are simply listed from the task statement, and not linked or logically organised.
1	***Discrete setting*** : Discrete setting as an introduction, with some details which also show some linkage and organisation. May have an additional element to those listed which is relevant to the story.
2	***Integrated setting :*** There is a setting which, rather than simply being at the beginning, is introduced throughout the story.
3	***Integrated and manipulated setting :*** In addition to the setting being introduced throughout the story, pertinent information is woven or integrated so that this integration contributes to the story.

Reprinted with permission from Harris, 1992, p. 49.

To obtain an orientation to the data, Table 4 shows the overall distributions of the classifications. The proportion of responses in the first class are similar, but it is evident that the addition of the fourth class drew classifications away from both the second and the third classes from the three-class set. The sum of the proportions in the third and fourth classes of the four-class set is .31, .10 greater than the proportion in class three of the three-class set. This is statistically different - $\chi^2 = 15.86$, df= 2, $p<<0.01$. Table 5 shows the values of the estimates of the threshold parameters in the multiplicative metric. It is evident that while the first threshold is located at approximately the same distance from the origin, the second threshold is further from the origin in the three-class set than in the four-class set, indicating that it spans a greater proportion of the continuum in the former set. Figures 1 and 2 show the probabilities of the distributions in each of the classes for each of the sets as a function of the location of the essay on the continuum. It is perhaps of some interest that for the four-class set, the distance between the first and second thresholds (1.64) is quite close to the distance between the second and third thresholds (1.43).

Table 4 : Distribution of classifications

Classes	0	1	2	3
Four class set	.18	.51	.24	.07
Three class set	.17	.62	.21	

Table 5 : Locations of the thresholds

Thresholds	1	2	3
Four class set	.31	1.95	3.38
Three class set	.26	2.39	

Table 6 : Observed and expected frequencies given a total score across the two sets of classifications

Total Score	Response Pairs $\{x_{pi}, x_{pj}\}$ Given the Total Score			Total Frequency
0	{0, 0}			
	66 (66)			66
1	{1, 0}	{0, 1}		
	17 (16.34)	19(19.66)		36
2	{2, 0}	{1, 1}	{0, 2}	
	2(6.04)	44(45.32)	6(0.64)	52
3	{3, 0}	{2, 1}	{1, 2}	
	4(1.91)	16(24. 89)	9(2.20)	29
4	{3, 1}		{1, 2}	
	5(16.45)		14(2.54)	19
5	{0, 0}			
	22{22}			22

Expected frequencies are in parentheses

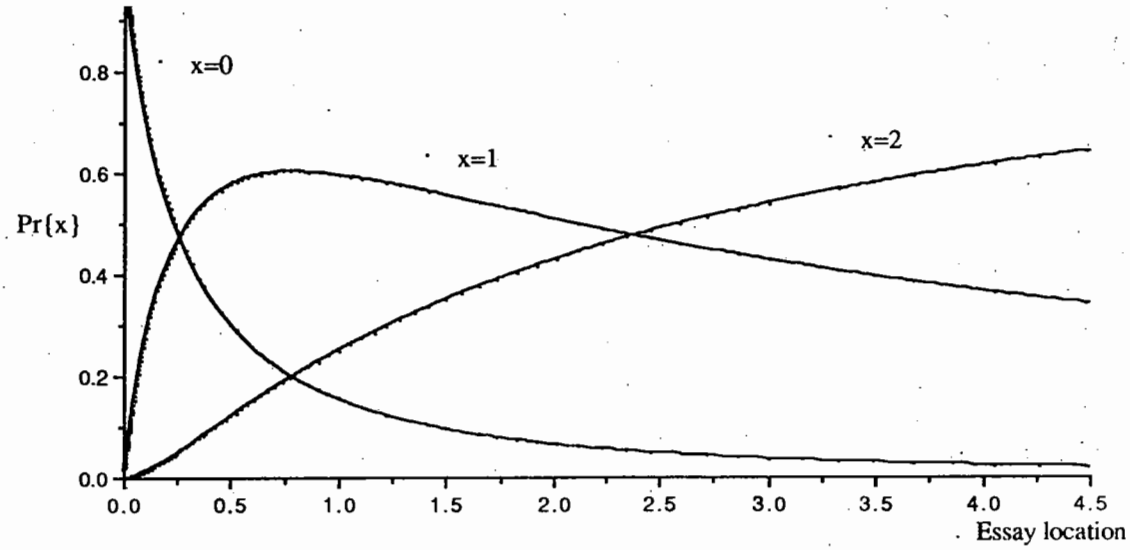


Figure 1a: Probability of classification into one of three ordered classes as a function of essay location expressed in the multiplicative metric $\Pr\{x_{pi}\} = \frac{1}{\gamma_{pi}} \xi^x / \prod_{k=0}^{x} \omega_{ki}$

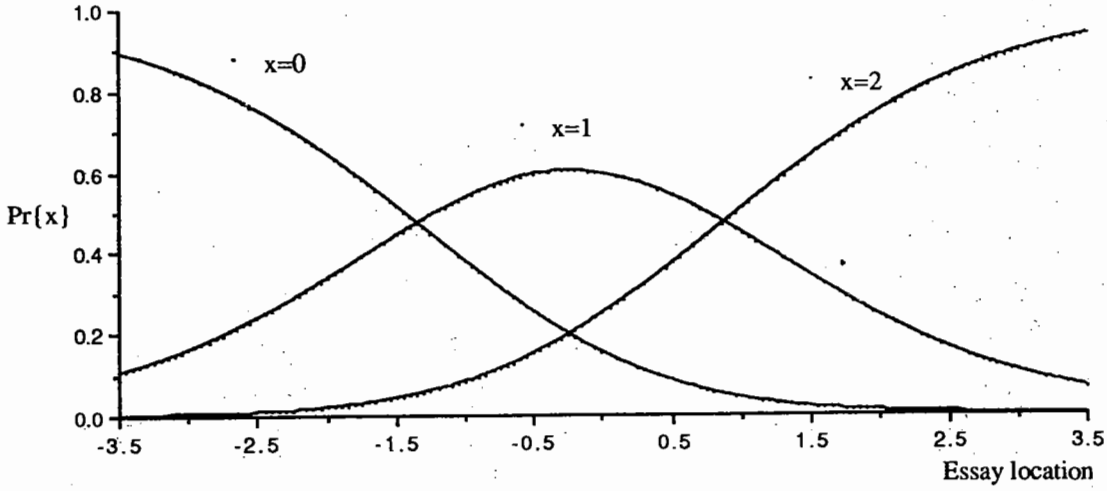


Figure 1b: Probability of classification into one of three ordered classes as a function of essay location expressed in the additive metric $\Pr\{x_{pi}\} = \frac{1}{\gamma_{pi}} \exp[x(\ln \xi) - \sum_{k=0}^{x} \ln \omega_{ki}]$

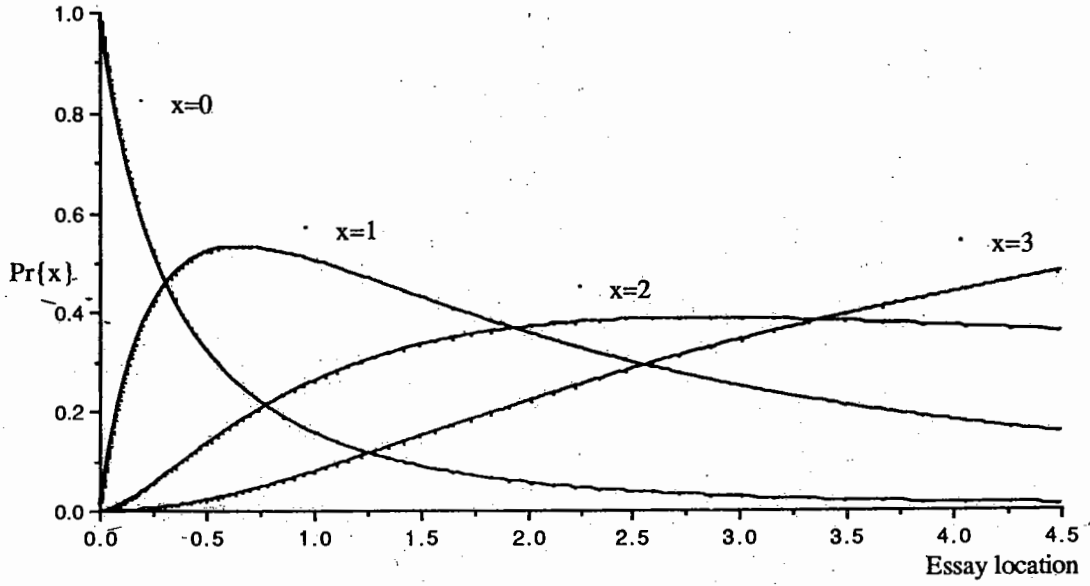


Figure 2a: Probability of classification into one of four ordered classes as a function of essay location expressed in the multiplicative metric $\Pr\{x_{pi}\} = \frac{1}{\gamma_{pi}} \xi^x / \prod_{k=0}^{x} \omega_{ki}$

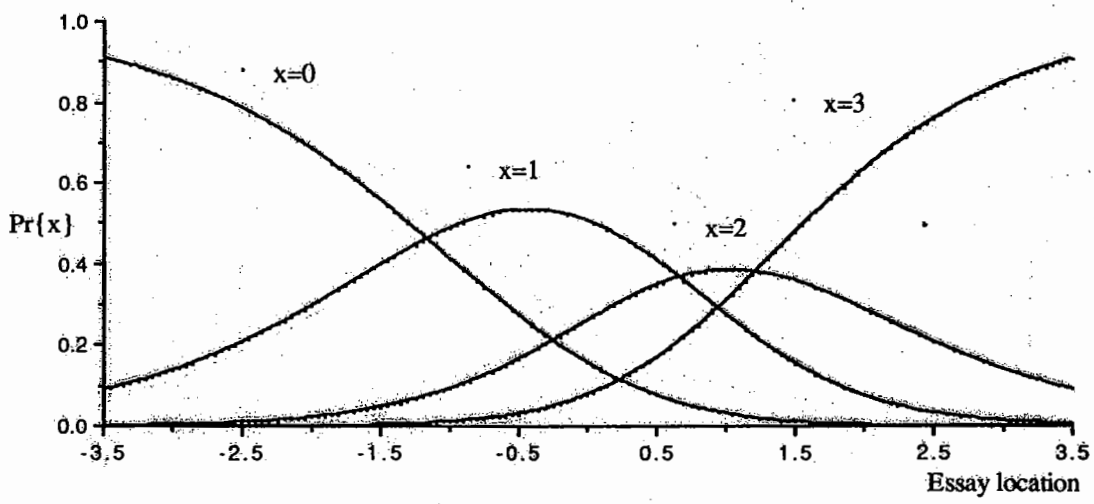


Figure 2b: Probability of classification into one of three ordered classes as a function of essay location expressed in the additive metric $\Pr\{x_{pi}\} = \frac{1}{\gamma_{pi}} \exp[x(\ln \xi) - \sum_{k=0}^{x} \ln \omega_{ki}]$

Even though the probabilities of being classified in a class, particularly in classes two and three, is affected by the additon or exclusion of the fourth class, to the degree that the data fit the model, to that degree the *location* of the essays on the continuum will be invariant under the two sets of classifications. To check on the degree to which the data fit the model, Table 6 shows the observed and expected values conditional on the total score for each essay across the two classifications[e] . Unfortunately, 4 of the 10 cells available for comparison have expected values well less than 5, making a chi-square statistic meaningless. However, some descriptive comments may be made. Taking the four-class set to be the first in each ordered pair as in Table 6, for a total score of 1, the distribution across (1,0) and (0,1) is as expected. Likewise for a score of 2, the number of (1,1) cases is as expected, although the number of (0,2) cases is greater than expected, and that of (2,0) is less than expected. The data follow the model less closely for a total score of 3, though the pattern is in the expected direction, with the (2,1) the most frequent observed and expected pair. The extreme pairs (1,2) and (3,0) are both greater than expected acccording to the model. For the total score of 4, there is the greatest departure from the model - here the pair (2,2) is much greater than expected, and the pair (3,1) much less than expected.

The data collection is of course open to empirical study. In this example the essays were graded using the four-class set followed immediately by the grading using the three-class set and some interference is likely, and further, the graders were not experienced or especially trained to grade such essays. Accounting for such factors is relevant in professional classification and in research studies. For the reasons presented in this paper, namely that the model provides a convergence of linguistics, social measurement and physical measurement, it is suggested that empirical studies relevant to different situations can be directed to understanding the conditions conducive to making classifications conform to the model of Equation (1). This would not only provide the invariance implied by Equation (1) when the classifications do conform to the model, something required in professional classifications and in research, but in the process of achieving conformity to the model, a greater understanding of classification processes in these situations would follow. Attaining measurements in science is often as important in providing understanding of the relevant processes as the uses subsequently made of the measurements, and this same feature can apply to the social sciences, particularly in scientometrics, where the science itself is studied scientifically.

[e] From Equation (12), the distribution of classifications given the total score is a function only of the parameters of the classes, and independent of the parameters of the essays, and the extreme total scores of 0 and 5 cannot be used in the test of fit because there is only one way of obtaining the

Acknowledgments

This paper was supported in part by a grant from the Australian Research Council. John Harris gave permission to use the data of the example.

References

1. Rasch, G. On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. IV, 321-334. Berkelcy CA : University of California Press, 1961.
2. Saussure, F. *Course in General Linguistics*. London: Peter Owen Ltd., 1959.
3. Edwards, A.L. & Thurstone, L.L. An interal consistency check for scale values determined by the method of successive integers. *Psychometrika*, 17,169-180, 1952.
4. Andrich, D. *Rasch models for measurement*. Sage Universty Papers Series on Quantitative Applications in the Social Sciences. Beverley Hills: Sage Publications, 1988.
5. Andersen, E.B. Sufficient statistics and latest trait models *Psychometrika*, 42, 69-81, 1977
6. Andrich, D. A rating formulation for ordered response categories. *Psychometrika*, 43, 357-374, 1978.
7. Rasch, G. *Probabilistic models for some intelligence and attainment tests.* (Copenhagen: Danish Institute for Educational Research), expanded edition (1980) with foreward and afterword by B.D. Wright. Chicago: The University of Chicago Press, 1960.
8. Fischer, G.H. On the existence and uniqueness of maximum-liklihood estimates in the Rasch model. *Psychometrika*, 46,1,59-77, 1981.

9. Wright, B.D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J: Educational Testing Service, 1968.
10. Rasch, G. An individualistic approach to item analysis. In *Readings in Mathematical Social Science.* P.F. Lazarsfeld and N.W. Henry, (Eds.) Chicago: Science Research Associates, 89-108,1966.
11. Jansen P.G.W. & Roskam. E.E. Latent trait models and dichotomization of graded responses. *Psychometrika*, 51 (1), 69-91, 1986.
12. Kuhn, T.S. The function of measurement in modern physical science. *Isis*, 52, (part 2) 161-193. Reproduced in T.S. Kuhn, *The Essential Tension* (1977). Chicago: The University of Chicago Press, 1961.
13. Ramsay, J.O. Review of Foundations of measurement, Vol. I, by D.H. Krantz, R.D. Luce, P. Suppes, A. Tverskey. *Psychometrika*, 40, 257-62, 1975.
14. Wright, B.D. Campbell concatenation for mental testing. *Special Interest Group: Rasch Measurement,* 2(1) 3-4, 1988.
15. Luce, R.D. & Tukey, J.W. Simultaneous conjoint measurement: A new type of fundamental measurement, *Journal of Mathematical Psychology*, 1, 1-27, 1964.
16. Wright, B.D. Additivity in Psychological Measurement. In E.E. Roskam (Ed.) *Measurement and Personality Assessment.* Selected papers, XXIII International Congress of Psychology, Volume 8, 101-111,1985.
17. Harris, J. *Consequences for social measurement of collapsing adjacent categories with three or more ordered categories.* Unpublished Master of Education Dissertation, Murdoch University, Western Australia, 1992.

Regression Structure of a Book Circulation Model Incorporating Loan Periods

Quentin L. Burrell and Michael R. Fenton
Statistical Laboratory, Department of Mathematics,
University of Manchester, Oxford Road,
Manchester, M13 9PL, U.K.

Keywords : Informetrics; mixed Poisson process; renewal process; correlation structure; regression.

In a recent paper (Burrell & Fenton, 1994) the authors enhanced the well-known gamma-Poisson model for library book circulations by incorporating an additional parameter to allow for the loan period. In this report the investigation is extended to study the bivariate structure of the model. In particular, the regression function is derived and is shown to have the approximately linear form found in many empirical studies, dating at least as far back as Morse (1968).

1. Introduction

Mathematical models to describe the circulations (i.e. checked-out borrowings) of library stock seek to emulate at least some of the general features of frequency-of-circulation (FOC) distributions which have been reported in many empirical studies. As more of these features are built into a model then, almost inevitably, the more complex the model becomes so that each improved approximation to reality is gained at a (mathematical) price.

In the case of FOC distributions, an early noted feature [see, e.g. Morse (1968), Montgomery et al. (1976), Burrell (1980) and Burrell and Cane (1982)] was that if such a distribution for data collected over a relatively short period of time (typically one academic year) is plotted on a log-linear scale (i.e. the logarithm of the number of items circulated against the number of circulations) then the resulting graph is approximately linear. (Almost always, the point for the number of items not circulated at all constitutes a departure from this linearity. The reasons for this, and arguments for ignoring the zero category in FOC distributions, have been well rehearsed in the literature and so will not be entered into again here.) This empirical fact suggests that the appropriate form for the FOC distribution is geometric.

However, for certain data sets, especially those collected over more extended periods of time, the log-linear FOC plot exhibits a more obviously curvilinear form which is much better described by a negative binomial distribution (NBD). [Ravichandra Rao (1980, 1982) compared the relative merits of a wide range of standard discrete distributions and found that the NBD had the best overall performance. More recently suggested alternatives have been the beta-binomial distribution (BBD), by Gelman and Sichel (1987), and the generalized NBD, by Ajiferuke and Tague (1990).]

More importantly, if the FOC distribution of a given collection is observed over an extending period of time then the overall shape of the log-linear plot changes in a systematic fashion. This is clearly revealed by graphs such as those of Bulick et al. (1976) and Kent et al. (1979; Fig. 9, p.41). [A similar phenomenon in cumulative bibliographies has been noted by Burrell (1992) and Oluić-Vuković (1992).] The inevitable implication is that models for library circulations

must be time-dependent, i.e. that one needs not merely a probability distribution to describe data collected over a single time period but a stochastic process to describe the (probabilistic) evolution of the system over time. [A non-technical discussion and some simple illustrative theoretical models were given by Burrell (1980, 1982)]. For the BBD model '*time*' is incorporated in a sometwhat artificial manner, while there seems to be no time-dependent form of the generalized NBD. The ordinary NBD, on the other hand, arises quite naturally as the single-period form of a stochastic process known as the gamma-Poisson (GP) process (see Burrell and Cane, 1982). This process has achieved considerable success in modelling library circulation on data, as in Burrell (1980, 1985, 1990c) and Brownsey and Burrell (1986).

One particularly striking feature of the GP process is its linear regression property. In the library circulation context this can be described as follows. Suppose that, for each item in the collection, the numbers of circulations in each of two consecutive time periods (typically two academic years) are noted, giving rise to a bivariate FOC distribution. Then for each r = 0,1,2,..., the mean number of circulations during the second period of those items circulating r times during the first is determined and plotted against r. This graph, of the regression function of the number of circulations during the second period given the number during the first, is linear for the GP process, see e.g. Burrell (1990b). That the observed regression function is, at least approximately, linear has been reported by Morse (1968), Chen (1976), Burrell and Cane (1982) and Burrell (1990a).

The linearity of regression was in fact used as a built-in assumption in the first reasonably successful model of library circulations originally proposed by Morse (1968). [For a discussion of the relative merits of Morse's Markov model and the GP model see Burrell (1986) and Tague and Ajiferuke (1987).]

The final empirical feature to be mentioned again relates to FOC distributions over a single time period but now focusing on the upper tail of the distribution. This concerns those items with large numbers of circulations for which the log-linear plot shows a rapid tailing-off as can be seen graphically in the University of Pittsburgh data of Kent et al. (1979; Fig. 9, p. 41) and the University of Sussex data of Burrell (1985; Fig. 2). Such tailing-off cannot be modelled by the NBD which tends, therefore, to overestimate the tail of FOC distributions. Indeed, Gelman and Sichel (1987) were able to use this tailing-off as an argument against a whole class of processes, including the GP, while at the same time promoting the case of the BBD. The generalised NBD of Ajiferuke and Tague (1990) has also been successfully employed in modelling this aspect of library circulations.

In order to overcome this, the major defect of the GP process, Burrell and Fenton (1994) proposed a modification which results in a form which outperforms the BBD and the generalized NBD for single-period data sets. As it is a stochastic model, there remains a need to consider its bivariate form and, in particular, its regression structure.

First, though, in the next section we give a brief description of the model.

2. The GPL Model

The important new element of the Burrell and Fenton (1994) model is the incorporation of the loan period into the standard GP model, thus acknowledging the fact that there are times when an item is off the shelf and hence not available for loan. Thus, each time an item is borrowed, this reduces the remaining time when it is available for loan in the period and the more often it is loaned the greater is the reduction. The effect of a loan period is therefore to depress the number of actual loans as compared to the number of potential loans if the item were always immediately available for loan.

In the gamma-Poisson with loans (GPL) model it is assumed that requests arrive for each book according to some stochastic process; if the item is available then it is loaned, while if it is already out on loan then this request (or potential loan) is lost. The situation can be illustrated as in **Figure 1** (taken from Burrell and Fenton, 1994). More precisely, the GPL model is based on the following :-

Assumption 1. Requests for a particular book arrive as a Poisson process.

Assumption 2. The rate λ of the request process varies over the collection of books according to a gamma distribution with probability density function

$$f(\lambda) = \frac{\beta^{-\nu}}{\Gamma(\nu)} \lambda^{\nu-1} e^{-\lambda/\beta}, \quad \lambda \geq 0$$

Assumption 3. The loan period is a constant, τ.

To be specific, we shall suppose in what follows that the unit of time (= period of observation) is an academic year and write

X = number of loans (of a typical book) during the period of observation.

[Implicitly, all books are available at the start of the year and X gives the number of times a book is taken out during the year rather than the number of completed loans (= returns).] .

The basic result for this GPL model is the following.

Theorem 1. (Burrell and Fenton, 1994).

$P(X = 0) = q_0$

$P(X = r) = q_r - q_{r-1}, \quad r = 1, 2, \ldots, R - 1$

$$P(X = R) = \begin{cases} 0 & \text{if } \tau^{-1} \text{ is an integer} \\ 1 - q_{R-1} & \text{otherwise} \end{cases},$$

where

$$q_r = \sum_{k=0}^{r} \binom{k+\nu-1}{k} \frac{[\beta(1-r\tau)]^k}{[1+\beta(1-r\tau)]^{\nu+k}},$$

$r = 0, 1, \ldots, R-1$

and $R = \left[\tau^{-1}\right] + 1$ = maximum number of loans possible during the year.

The success of this GPL model in fitting FOC distributions over a single time period is amply illustrated in Burrell and Fenton (1994), particularly its improvement over the GP model in fitting the upper tail. Note that it is a three-parameter model and hence in fitting to observed data each of β, ν and τ are to be estimated.

3. The Bivariate FOC Distribution and Regression Structure

Suppose now that loans are observed over two successive time periods both, for specificity, assumed to be academic years. (The case where the time periods have different lengths can be similarly handled, although notationally more inconvenient. For the same reasons we ignore the effects of ageing of library material, see Burrell, 1986). In order to investigate the bivariate structure, details of which may be found in the **Appendix,** we make one further assumption, as follows.

Assumption 4. There is an annual end-of-year recall of all books currently on loan.

Although the major effect of this assumption is to simplify much of the mathematics, it is also a true reflection of actual practice in many libraries and hence not too restrictive. With the help of **Assumption 4** we can prove the following.

Theorem 2. For the GPL model with **Assumptions 1-4**, if $X^{(1)}$, $X^{(2)}$ denote the number of loans of a single book during the first and second years respectively, then the bivariate distribution of $(X^{(1)}, X^{(2)})$ is given by

$$P(X^{(1)} = i, X^{(2)} = j) = q(i,j) - q(i,j-1) - q(i-1,j) + q(i-1, j-1)$$

for $i,j = 0, 1, \ldots, R$, where

$$q(m,n) = \begin{cases} \dfrac{\beta^{-\nu}}{\Gamma(\nu)} \displaystyle\sum_{k=0}^{m} \sum_{\ell=0}^{n} \frac{\Gamma(k+\ell+\nu)}{k!\ell!} \frac{(1-m\tau)^k (1-n\tau)^\ell}{\left[2-(m+n)\tau+\beta^{-1}\right]^{k+\ell+\nu}}, & \text{if } m \geq 0,\ n \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Proof. For stochastic processes constructed via mixtures of Poisson processes, distributional results are usually best derived by conditioning on the mixing parameter, in this case the rate of the

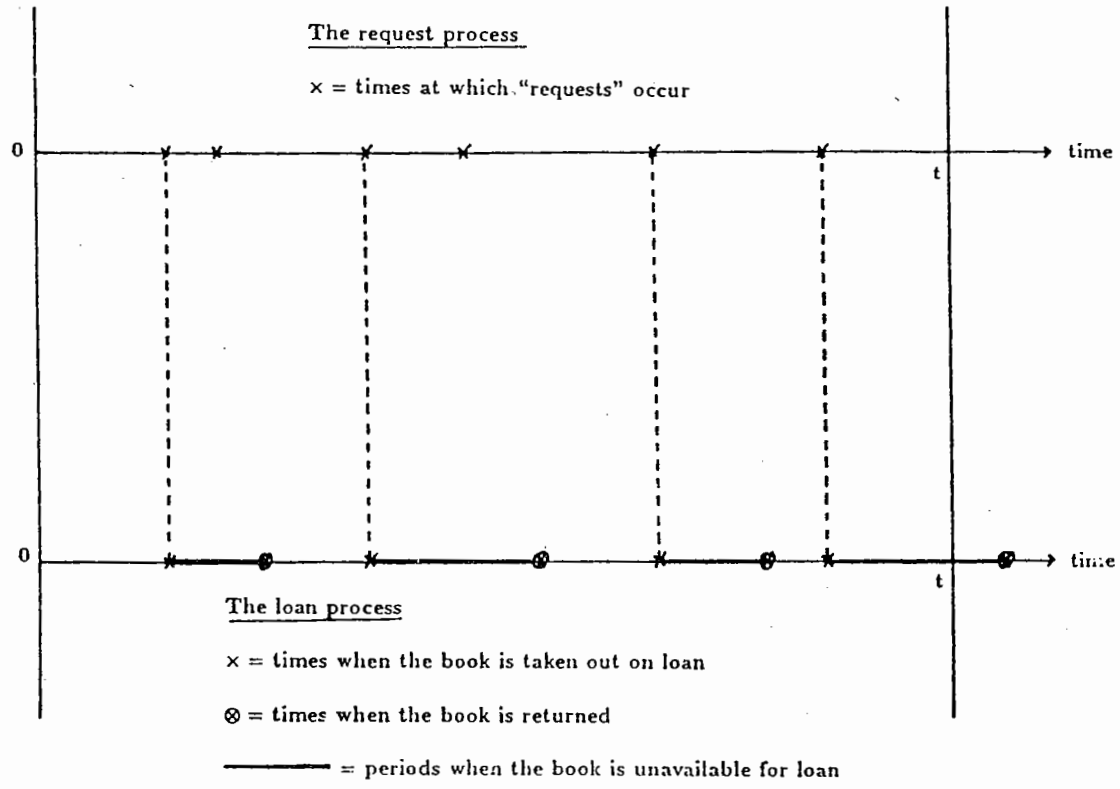


Figure 1. The request and loan processes for a single book.

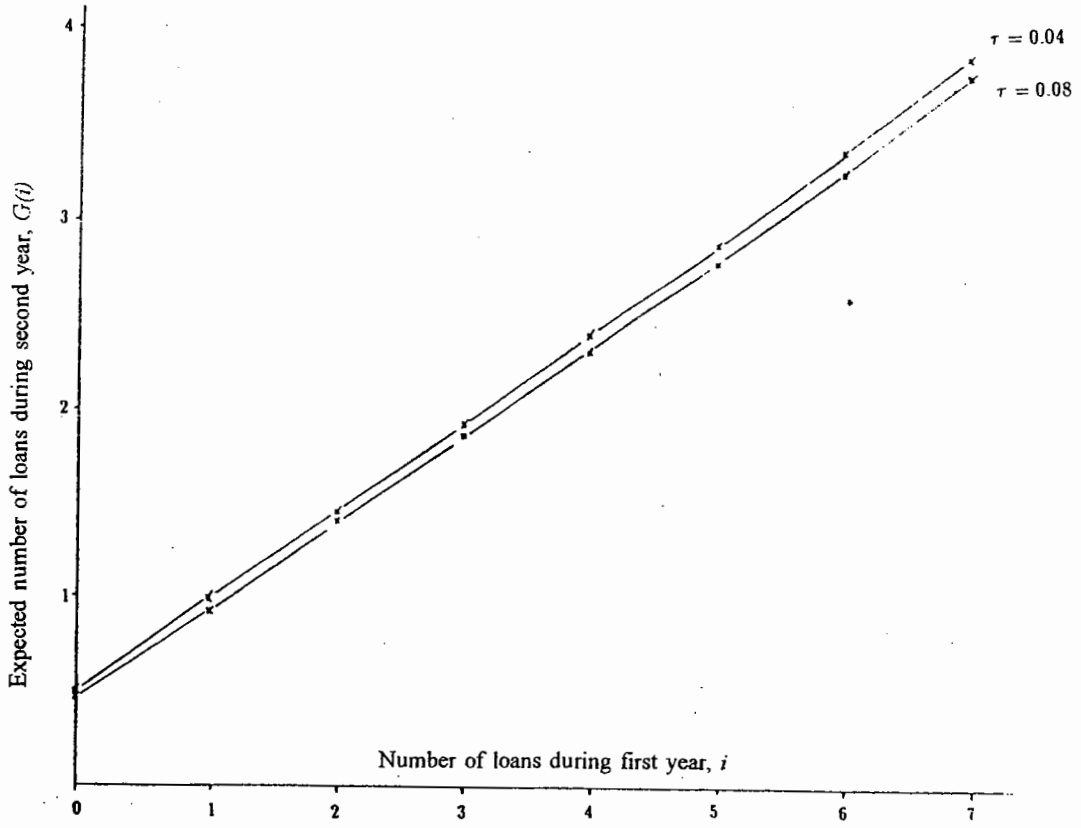


Figure 2. The regression function for the GPL model with $\nu = \beta = 1$

request process. One can then make use of the fact that the underlying process has stationary and independent increments. The actual derivation of the above is rather long and somewhat technical and so is deferred to the **Appendix**.

For the regression structure we need the conditional distribution of the number of loans during the second year given the number during the first. This is provided by the following.

Corollary. Under the conditions of **Theorem 2,**

$$P(X^{(2)} = j | X^{(1)} = i) = \frac{q(i,j) - q(i,j-1) - q(i-1,j) + q(i-1,j-1)}{q_i - q_{i-1}}$$

for $i,j = 0,1, \ldots, R,$

where $q(m,n)$ is as in **Theorem 2,** q_r is as in **Theorem 1** and $q_{-1}=0$, $q_R = 1$. [Note: If τ^{-1} is an integer then the apppropriate range extends only to $R-1$.]

Rather than the full bivariate and conditional distributions, we are here concerned with the regression function of $X^{(2)}$ on $X^{(1)}$, i.e. the (conditional) mean number of circulations during the second year given the number during the first. Writing G for the regression function, we therefore need

$$G(i) = E\left[X^{(2)} | X^{(1)} = i\right] = \sum_{j=0}^{R} jP\left(X^{(2)} = j | X^{(1)} = i\right)$$

for $i = 0,1,\ldots,R.$

For purposes of computation it is more convenient to use the alternative representation, whose derivation may be found in the **Appendix**, given in the following **Theorem**.

Theorem 3. Under the conditions and notation of **Theorem 2**, the regression function of $X^{(2)}$ on $X^{(1)}$ is given by

$$G(i) = (R+1) - \frac{Q_i - Q_{i-1}}{q_i - q_{i-1}}$$

where $Q_i = \sum_{j=0}^{R} q(i,j)$ for $i = 0,1,\ldots,R.$

The above neat, compact, formula for the regression function is most useful for purposes of computation, its application being straightforward, if somewhat tedious.

Example : $\nu = 1$, $\beta = 1$.

Choosing the gamma index to be $\nu = 1$ is equivalent to assuming an exponential mixing distribution and to the frequency distribution of requests (or the FOC distribution in the case where the loan period is zero) being geometric. Taking $\beta = 1$ corresponds to a choice of time scale which simplifies calculations. This example was given for purposes of illustration by Burrell and Fenton (1994) and revealed, as expected, a FOC distribution tailing off in the higher frequencies, with the effect being more marked the longer the loan period.

For computation of the regression function, note that in this case the relevant quantities to be evaluated become

$$q_n = 1 - \left(\frac{1-n\tau}{2-n\tau}\right)^{n+1}$$

and

$$Q_n = \sum_{m=0}^{R} q(m,n) = \sum_{m=0}^{R} \sum_{k=0}^{m} \sum_{\ell=0}^{n} \frac{(k+\ell)!}{k!\ell!} \frac{(1-m\tau)^k (1-n\tau)^\ell}{\left[3-(m+n)\tau\right]^{k+\ell+1}}$$

for $n = 0,1,\ldots,R.$

The graph of the resulting regression function is given in **Figure 2** for various values of τ.

As was to be hoped, and expected, the plots are almost exactly linear. Certainly, any departures from linearity are ones which it would be difficult to detect in a practical context : for the smaller frequencies the departures are exceedingly small, for the larger ones the data would be very sparse.

4. Concluding Remarks

To the extent that informetrics is concerned with the mathematical modelling of library and information systems, an important consideration is that the models should as far as possible re-

flect the known features of the system under investigation. The more closely we consider the details of these features, almost inevitably the more complex is the required model. Nowadays, with the aid of computers, investigation of fairly complicated systems can be relatively straightforward even if the theoretical models are analytically daunting. This is the case here.

For the practitioner, the important factors are that the basic ideas behind the model can be explained in non-technical terms and that the model "works". Our preliminary findings suggest that, in the present case, the latter requirement is satisfied; the reader must decide if the former is also!

It is hoped that more detailed investigations -- including the fitting of the bivariate model to empirical data -- will be reported in due course.

Acknowledgement

During part of the period of this research, Michael Fenton received financial support from the Science and Engineering Research Council, to which body he is grateful.

Appendix

Much of the notation and certain of the results are adapted from Burrell and Fenton (1994), referred to simply as B and F in the following.

Denote by $X^{(1)}$, $X^{(2)}$ the number of loans of a single book during the first and second years respectively. As always in the construction of processes via mixtures, distributional results are most easily derived by conditioning on the mixing parameter, in this case the rate of the underlying request process. Thus for the bivariate distribution of $(X^{(1)}, X^{(2)})$ we have first of all that, given the request rate $\wedge$, the numbers of loans during the two periods are determined by the Poisson request process. The annual recall assumption together with the lack-of-memory and independence of increments of a Poisson process then ensure that $X^{(1)}$ and $X^{(2)}$ are conditionally independent, given $\wedge$. Hence

$$P\left(X^{(1)}=i, X^{(2)}=j\right)$$

$$= E_{\wedge} P\left(X^{(1)}=i, X^{(2)}=j|\wedge\right)$$

$$= E_{\wedge} P\left(X^{(1)}=i|\wedge\right) P\left(X^{(2)}=j|\wedge\right) \qquad \text{(A1)}$$

Now stationarity of increments of the request process, together with the annual recall assumption, implies that $X^{(1)}$ and $X^{(2)}$ both have the same conditional distribution, given $\wedge$. Hence, writing $\left\{W_t^{\lambda}; t \geq 0\right\}$ for a Poisson process of rate λ, from the above remark and B and F (Lemma 3) we have

$$P\left(X^{(1)}=r|\wedge=\lambda\right)$$

$$= P\left(X^{(2)}=r|\wedge=\lambda\right)$$

$$= P\left(W^{\lambda}_{1-r\tau} \leq r \leq W^{\lambda}_{1-(r-1)\tau}\right)$$

$$= P\left(W^{\lambda}_{1-r\tau} \leq r\right) - P\left(W^{\lambda}_{1-(r-1)\tau} \leq r-1\right)$$

$$= q_r(\lambda) - q_{r-1}(\lambda) \text{ for } r=0,1,\ldots,R \qquad \text{(A2)}$$

where R denotes the maximum possible number of loans in one year and we have written

$$q_r(\lambda) = P\left(W^{\lambda}_{1-r\tau} \leq r\right)$$

$$= e^{-\lambda(1-r\tau)} \sum_{k=0}^{r} \frac{\left[\lambda(1-r\tau)\right]^k}{k!} \text{ for } r=1,2,\ldots,R \qquad \text{(A3)}$$

and $q_{-1}(\lambda) = 0$

Thus from (A1) we have

$$P\left(X^{(1)}=i, X^{(2)}=j\right)$$

$$= E_{\wedge}\left[\left(q_i(\wedge)-q_{i-1}(\wedge)\right)\left(q_j(\wedge)-q_{j-1}(\wedge)\right)\right] \qquad \text{(A4)}$$

Upon expanding the RHS of (A4) we find four terms of the form $E_{\wedge}\left[q_m(\wedge)q_n(\wedge)\right]$ and, using (A3) :-

$$E_{\wedge}\left[q_m(\wedge)q_n(\wedge)\right]$$

$$=\sum_{k=0}^{m}\sum_{\ell=0}^{n}E_{\wedge}\left[e^{-\wedge(1-m\tau)}\frac{\left[\wedge(1-m\tau)\right]^k}{k!}e^{-\wedge(1-n\tau)}\frac{\left[\wedge(1-n\tau)\right]^\ell}{\ell!}\right]$$

$$=\sum_{k=0}^{m}\sum_{\ell=0}^{n}\frac{(1-m\tau)^k}{k!}\frac{(1-n\tau)^\ell}{\ell!}E_{\wedge}\left[e^{-\wedge[2-(m+n)\tau]}\wedge^{k+\ell}\right] \qquad \text{(A5)}$$

Now making use of **Assumption 2** of the GPL model

$$E_{\wedge}\left[e^{-\wedge[2-(m+n)\tau]}\wedge^{k+l}\right]$$

$$=\int_0^\infty e^{-\lambda[2-(m+n)\tau]}\lambda^{k+\ell}f(\lambda)d\lambda$$

$$=\frac{\beta^{-\nu}}{\Gamma(\nu)}\int_0^\infty \lambda^{k+\ell+\nu-1}\exp\left\{-\lambda\left[2-(m+n)\tau+\beta^{-1}\right]\right\}d\lambda$$

$$=\frac{\beta^{-\nu}}{\Gamma(\nu)}\frac{\Gamma(k+\ell+\nu)}{\left[2-(m+n)\tau+\beta^{-1}\right]^{k+\ell+\nu}}$$

so that (A5) becomes

$$E_{\wedge}\left[q_m(\wedge)q_n(\wedge)\right]$$

$$=\beta^{-\nu}\sum_{k=0}^{m}\sum_{\ell=0}^{n}\frac{\Gamma(k+\ell+\nu)}{k!\ell!\Gamma(\nu)}\frac{(1-m\tau)^k(l-n\tau)^\ell}{\left[2-(m+n)\tau+\beta^{-1}\right]^{k+\ell+\nu}}$$

$$=q(m,n) \text{ say} \qquad \text{(A6)}$$

Substituting back into (A4) gives the bivariate distribution as

$$P\left(X^{(1)}=i,X^{(2)}=j\right)=q(i,j)-q(i,j-1)-q(i-1,j)+q(i-1,j-1) \qquad \text{(A7)}$$

Remark. Although the expression for *q(m,n)* given in (A6) may book awkward, its form becomes more transparent if we write

$$x = 1 - m\tau,\ y = 1 - n\tau,\ z = \beta^{-1}$$

since then it becomes

$$q(m,n)=\sum_{k=0}^{m}\sum_{\ell=0}^{n}\frac{\Gamma(k+\ell+\nu)}{k!\ell!\Gamma(\nu)}\frac{x^k y^\ell z^\nu}{(x+y+z)^{k+\ell+\nu}} \qquad \text{(A8)}$$

so that the individual terms are of a trinomial form. For purposes of computation (A8) can be rearranged to give.

$$q(m,n)=\left(\frac{z}{x+y+z}\right)^{\nu}\frac{1}{\Gamma(\nu)}\sum_{k=0}^{m}\left\{\frac{1}{k!}\left(\frac{x}{x+y+z}\right)^{k}\sum_{\ell=0}^{n}\frac{\Gamma(k+\ell+\nu)}{\ell!}\left(\frac{y}{x+y+z}\right)^{\ell}\right\}$$

Proof of Theorem 3.

By definition

$$G(i)=\sum_{j=0}^{R} jP\left(X^{(2)}=j\,|\,X^{(1)}=i\right)$$

$$=\sum_{j=0}^{R} jP\left(X^{(1)}=i,X^{(2)}=j\right)/P\left(X^{(1)}=i\right)$$

So that, from **Theorems 1** and **2**,

$$(q_i-q_{i-1})G(i)$$

$$=\sum_{j=0}^{R} j\left[q(i,j)-q(i,j-1)-q(i-1,j)+q(i-1,j-1)\right]$$

$$=\sum_{j=0}^{R} jq(i,j)-\sum_{j=0}^{R-1}(j+1)q(i,j)-\sum_{j=0}^{R} jq(i-1,j)+\sum_{j=0}^{R-1}(j+1)q(i-1,j)$$

$$=Rq(i,R)-\sum_{j=0}^{R-1}q(i,j)-Rq(i-1,R)+\sum_{j=0}^{R-1}q(i-1,j)$$

$$=(R+1)\left[q(i,R)-q(i-1,R)\right]-\left(\sum_{j=0}^{R}q(i,j)-\sum_{j=0}^{R}q(i-1,j)\right) \qquad \text{(A9)}$$

Now for any m, n we have

$$P\left(X^{(1)}\le m,X^{(2)}\le n\right)=E_{\wedge}P\left(X^{(1)}\le m,X^{(2)}\le n|\wedge\right)=E_{\wedge}P\left(X^{(1)}\le m|\wedge\right)P\left(X^{(2)}\le n|\wedge\right)$$

$$=E_{\wedge}P\left(\hat{W}_{1-m\tau}\le m\right)P\left(\hat{W}_{1-n\tau}\le n\right)=E_{\wedge}\left[q_m(\wedge)q_n(\wedge)\right]=q(m,n)$$

Hence, in particular

$$q(m,R)=P\left(X^{(1)}\le m,X^{(2)}\le R\right)=P\left(X^{(1)}\le m\right)=q_m$$

Substituting into (A9), and writing

$$Q_m=\sum_{n=0}^{R}q(m,n)$$

we find

$$(q_i-q_{i-1})G(i)=(R+1)(q_i-q_{i-1})-(Q_i-Q_{i-1})$$

so that

$$G(i)=(R+1)-\frac{Q_i-Q_{i-1}}{q_i-q_{i-1}}$$

as required.

References

1. Ajiferuke, I. & Tague, J.M. (1990). A model for the full circulation data. In L. Egghe & R. Rousseau (eds.) *Informetrics 89/90 : selection of papers submitted for the second International Conference on Bibliometrics, Scientometrics and Informetrics* (pp. 1-16). Amsterdam : Elsevier.
2. Brownsey, K.W.R. & Burrell, Q.L. (1986. Library circulation distritutions : some observations on the PLR sample. *Journal of Documentation*, **42**, 22-45.
3. Bulick, S., Montgomery, K.L., Fetterman, J. & Kent, A. (1976). Use of library materials in terms of age. *Journal of the American Society for Information Science*, **27**, 175-178.
4. Burrell, Q.L. (1980). A simple stochastic model for library loans. *Journal of Documentation*, **36**, 115-132.
5. Burrell, Q.L. (1982). Alternative models for library circulation data. *Journal of Documentation*, **38**, 1-13.
6. Burrell, Q.L. (1985). A note on ageing in a library circulation model. *Journal of Documentation*, **41**, 100-115.
7. Burrell, Q.L. (1986). A second note on ageing in a library circulation model : the correlation structure. *Journal of Documentation*, **42**, 114 - 128.
8. Burrell, Q.L. (1990a). Correlation structure of library circulation data: a case study. Part 1: the empirical view. *Library Science with a slant to Documentation and Information Studies*, **27**, 213-236.
9. Burrell, Q.L. (1990b). Correlation structure of library circulation data: a case study. Part 2: theoretical aspects. *Library Science with a slant to Documentation and Information Studies*, **27**, 237-252.
10. Burrell, Q.L. (1990c). Using the gamma-Poisson model to predict library circulations. *Journal of the American Society for Information Science*, **41**, 164-170.
11. Burrell, Q.L. (1992). The dynamic nature of bibliometric processes: a case study. In: I.K. Ravichandra Rao (ed.) *Informetrics 91* (pp 97-129). Bangalore: Ranganathan Endowment for Library Science.
12. Burrell, Q.L. & Cane, V.R. (1982). The analysis of library data. (With discussion.) *Journal of the Royal Statistical Society, Series A*, **145**, 439-471.
13. Burrell, Q.L. & Fenton, M.R. (1993). Yes, the GIGP really does work — and is workable! *Journal of the American Society for Information Science*,**44**, 61-69.
14. Burrell, Q.L. & Fenton, M.R. (1994). A model for library book circulations incorporating loan periods. *Journal of the American Society for Information Science*, **45** 101-116.
15. Chen, C.-C. (1976). *Applications of operations research models to libraries: a case study of the use of monographs in the Francis A. Countway Library of Medicine*. Cambridge, Mass.: MIT Press.
16. Gelman, E. & Sichel, H.S. (1987). Library book circulation and the betabinomial distribution. *Journal of the American Society for Information Science*, **38**, 4-12.
17. Kent, A. et al. (1979). *Use of library materials: the University of Pittsburgh study*. New York: Marcel Dekker.
18. Montgomery, K.L., Bulick, S., Fetterman, J. & Kent, A. (1976). Cost-benefit model of library acquisitions in terms of use: Progress report. *Journal of the American Society for Information Science*,**27**, 73-74.
19. Morse, P.M. (1968). *Library effectiveness: a systems approach*. Cambridge, MASS: MIT Press.
20. Oluic-Vukovic, V. (1992). Journal productivity distribution: Quantitative study of dynamic behaviour. *Journal of the American Society for Information Science*,**43**, 412-421.
21. Ravichandra Rao, I.K. (1980). The distribution of scientific productivity and social change. *Journal of the American Society for Information Science*, **31**, 111-122.
22. Ravichandra Rao, I.K. (1982). Document and user distributions. *Library Science with a slant to Documentation*, **19**, 69-96.
23. Tague, J. & Ajiferuke, I. (1987). The Markov and the mixed-Poisson models of library circulations compared. *Journal of Documentation*, **43**, 212-231.

Confidence Intervals and Sample Sizes for Multinomial Data

L. Egghe
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

N. Veraverbeke
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium.

This paper surveys the existing literature on confidence intervals for multinomial data. Multinomial situations (fractions) are very common as is illustrated in medicine and information sciences. Only conservative results exist. We make a comparison of several results and select the best confidence intervals. We show that, in general, the simplest methods perform best (such as the methods derived from the central limit theorem and the Bonferroni inequality). We also show that the method of Fitzpatrick and Scott cannot be the best for all fractions involved (at least if there are more than two classes). The sample sizes are derived from the formulae of the confidence intervals. Fixed length intervals as well as proportional lengths are considered. It is found that for only estimating the highest proportions in an accurate way, the sample sizes are reasonable and well-applicable. The paper closes with improved confidence intervals in the case they are dependent on the proportions and with some practical conclusions.

1. Introduction

It is remarkable how many statistical experiments are of the following type : a sample of N individuals or objects is observed and each is classified in one and only one of a finite number of mutually exclusive classes (cells, categories), say $C_1,.....,C_k (k \geq 2)$. The outcomes of such an experiment are $n_1,.....,n_k$ where $n_i (i = 1,.....,k)$ is the number of times that class C_i occurs in the sample. Note that $n_1 + ... + n_k = N$

The statistical problem is to make inference on the proportions (fractions) $p_1,.....,p_k$ where $p_i (i = 1,...,k)$ is the probability of falling in class C_i.

A constraint for these unknown parameters is of course that $p_1 + ... + p_k = 1$.

Examples are easy to find in exact, medical and social sciences :

(1) What are the fractions of library visits classified by sex and a number of age intervals ?

(2) What are the fractions of books in a library with a certain number of authors (and posssibly further divided according to age, language, or the fact that they are in the automated system or not)?

(3) What are the fractions of people with a certain blood type ?

(4) What are the numbers of patients with a certain disease status ?

(5) What are the proportions of men whose occupation falls in a certain number of job categories ?

(6) What are the proportions in the different progeny types of a Mendel experiment with plants ?

(7) What will be the scores of the different political parties in the next elections ?

The numbers of classes (k) can be as smalll as 2 but can also be rather large. In example (1) we have 20 classes if we select 10 age intervals. In example (2), we could easily go above 100 classes.

The simplest case is that with two classes for a so called dichotomous situation. The two classes of the experiment are often arbitrarily denoted by

[1] Permanent Address

success and failure (which stands for yes-no, boy-girl, zero-one, etc ...). Denote by p the probability of success. It is well known that in a sample of size N, the probability of having say n times success (and N - n times failure) is given by the binomial probability

$$\binom{N}{n} p^n (1-p)^{N-n} \qquad (1)$$

If in a sample of size N, the number of observed successes equals n, then the maximum likelihood estimate for the unknown probability p is the observed class frequency

$$\hat{p} = \frac{n}{N} \qquad (2)$$

Generalization to k classes $(k \geq 3$ is as follows. If $p_1, \ldots, p_k$ denote the probabilities of the k classes, then the probability that in a sample of size N, the 1st class occurs n_1 times, the 2nd occurs n_2 times, ..., the k-th occurs n_k times $(n_1 + \ldots + n_k = N)$ is given by the multinomial formula

$$\frac{N!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} \qquad (3)$$

If $n_1, \ldots, n_k$ are the observed number of times that each of the k classes occurs in a sample of size N, then the maximum likelihood estimates for the unknown class probabilities $p_1, \ldots, p_k$ are the observed class frequencies

$$\hat{p}_i = \frac{n_i}{N} (i = 1, \ldots, k). \qquad (4)$$

2. Simultaneous Confidence Intervals for Proportions

First we recall the standard types of confidence intervals of 1 single proportion p. Since the standard deviation of $\hat{p}$ equals $\sqrt{\frac{p(1-p)}{N}}$, the normal approximation to the binomial distribution says that, for N large,

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}}$$

is approximately standard normal (5)

There are several ways to derive from this fact approximate confidence intervals for p.

The first is to use that (5) remains time if the denominator is replaced by the estimator $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. This provides the approximate $100(1-\alpha)\%$ confidence interval for p of the form

$$\left[\hat{p} - z(\alpha/2) \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \hat{p} + z(\alpha/2) \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \right] \qquad (6)$$

where $z(\alpha/2)$ satisfies $P(Z > z(\alpha/2)) = \alpha/2$ with Z the standard normal random variable.

A second way is a bit more sophisticated : from (5) it follows that, for large N,

$$P\left(\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}} \right)^2 \leq z^2(\alpha/2) \right) \approx 1 - \alpha$$

or equivalently, and writing $z = z(\alpha/2)$,

$$P\left((\hat{p} - p)^2 \leq z^2 \frac{p(1-p)}{N} \right) \approx 1 - \alpha$$

or

$$P\left((N + z^2) p^2 - (2N\hat{p} + z^2) p + N\hat{p}^2 \leq 0 \right) \approx 1 - \alpha$$

The two roots of this quadratic are easily found and this gives the approximate $100(1-\alpha)\%$ confidence interval for p :

$$\left[\frac{z^2 + 2N\hat{p} - \sqrt{z^2 \left(z^2 + 4N\hat{p}(1-\hat{p}) \right)}}{2(N + z^2)}, \right.$$

$$\left.\frac{z^2+2N\hat{p}+\sqrt{z^2\left(z^2+4N\hat{p}\left(1-\hat{p}\right)\right)}}{2\left(N+z^2\right)}\right] \qquad (7)$$

Also note the well known fact that $z=z(\alpha/2)=\chi_1^2(\alpha)$ where $\chi_1^2(\alpha)$ satisfies $P(T>x_1^2(\alpha))=\alpha$ with T a random variable with a $\chi_1^2(\alpha)$ distribution.

A third very simple approximate confidence interval for p is obtained by replacing $\hat{p}\left(1-\hat{p}\right)$ in (6) by its maximal value 1/4. This gives the following 'conservative' interval

$$\left[\hat{p}-z(\alpha/2)\frac{1}{2\sqrt{N}},\hat{p}+z(\alpha/2)\frac{1}{2\sqrt{N}}\right] \qquad (8)$$

When dealing with k proportions $p_1, ..., p_k$ as outlined above, one could apply one of the above confidence intervals, for every i, yielding for every $\hat{p}_i$ an interval I_i with the property that $P(I_i)\approx 1-\alpha$. However, these k results **do not** combine to yield a simultaneous result with confidence level $100(1-\alpha)\%$. Indeed, the occurence of I_1 does not imply the occurence of I_2, since both have probability of $100(1-\alpha)\%$, and so on for the other I_i. So, to have the $I_i (i=1, ..., k)$ jointly (i.e. k simultaneous confidence intervals), we must have information on the probability that $\bigcap_{i=1}^{k} I_i$ will occur. This is what is really needed since we wish to make joint statements about the k proportions $p_1, ..., p_k$ (as e. g. in prediction tables on the occasion of elections, where k political parties are involved).

A simple method for constructing simultaneous confidence intervals for $p_1, ..., p_k$ is to use the above confidence intervals for each of the pi and to combine them in one statement. The Bonferroni inequality then gives a lower bound for the probability of the joint statement in terms of the probabilities of the individual statements. This elementary inequality says that for any events $E_1, ..., E_k$:

$$P\left(\bigcap_{i=1}^{k} E_i\right)\geq 1-\sum_{i=1}^{k}\left[1-P(E_i)\right] \qquad (9)$$

Applying this idea to the confidence intervals of the form (6) gives a first type of simultaneous confidence intervals:

The k events (i=1, ..., k)

$$A_i=\left\{\hat{p}_i-z(\alpha/2k)\sqrt{\frac{\hat{p}_i(1-\hat{p})}{N}}\leq p_i\right.$$

$$\left.\leq \hat{p}_i+z(\alpha/2k)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}}\right\} \qquad (10)$$

satisfy

$$\lim_{N\to\infty} P\left(\bigcap_{i=1}^{k} A_i\right)\geq 1-\alpha \qquad (11)$$

Indeed, each of the A_i constitutes an approximate $100(1-\alpha/k)\%$ confidence interval, so that for i=1, ..., k

$$\lim_{N\to\infty} P(A_i)=1-\frac{\alpha}{k}$$

and

$$\lim_{N\to\infty} P\left(\bigcap_{i=1}^{k} A_i\right)\geq 1-\sum_{i=1}^{k}\frac{\alpha}{k}=1-\alpha$$

See also Goodman (1965) [4].

In exactly the same way we obtain from the intervals (7) and Bonferroni's inequality:

The k events (i=1, ..., k)

$$B_i = \left\{ \frac{z^2(\alpha/2k) + 2n_i - \sqrt{z^2(\alpha/2k)\left(z^2(\alpha/2k) + 4n_i\left(1 - \frac{n_i}{N}\right)\right)}}{2\left(N + z^2(\alpha/2k)\right)} \le p_i \right.$$

$$\left. \le \frac{z^2(\alpha/2k) + 2n_i + \sqrt{z^2(\alpha/2k)\left(z^2(\alpha/2k) + 4n_i\left(1 - \frac{n_i}{N}\right)\right)}}{2\left(N + z^2(\alpha/2k)\right)} \right\} \qquad (12)$$

satisfy:

$$\lim_{N\to\alpha} P\left(\bigcap_{i=1}^{k} B_i\right) \ge 1 - \alpha. \qquad (13)$$

See Goodman (1965) [4] and Quesenberry and Hurst (1964) [6].

A third type of simultaneous confidence intervals is obtained by combining k intervals of the form (8). Fitzpatrick and Scott (1987)[3] proved the following:

The k events (i=1, ..., k)

$$C_i = \left\{ \hat{p}_i - z(\alpha/4)\frac{1}{2\sqrt{N}} \le p_i \le \hat{p}_i + z(\alpha/4)\frac{1}{2\sqrt{N}} \right\} \qquad (14)$$

satisfy

$$\lim_{N\to\infty} P\left(\bigcap_{i=1}^{k} C_i\right) \ge 1 - \alpha, \text{ if } \alpha \le 0.032$$

or,

$$\ge 6\Phi\left(\frac{3z(\alpha/4)}{\sqrt{8}}\right) - 5, \text{ if } 0.032 \le \alpha \le 0.300 \qquad (15)$$

(here Φ denotes the standard normal distribution function).

This formula reduces to

$$\lim_{N\to\infty} P\left(\bigcap_{i=1}^{k}\left\{|\hat{p}_i - p_i| \le \frac{1.13}{\sqrt{N}}\right\}\right) \ge 0.95 \qquad (16)$$

for $\alpha = 0.05$ and to

$$\lim_{N\to\infty} P\left(\bigcap_{i=1}^{k}\left\{|\hat{p}_i - p_i| \le \frac{1.40}{\sqrt{N}}\right\}\right) \ge 0.99 \qquad (17)$$

for $\alpha = 0.01$.

The confidence intervals A_i, B_i and C_i above are all based on the central limit result for the $\hat{p}_i$ together with Bonferroni's inequality. Intervals of a different type have been derived by Huschens (1990) [5]. These are not asymptotic and rely on well known finite sample inequalities for the probabilities $P\left(|\hat{p}_i - p_i| > c\right)$, together with Bonferroni's inequality. For instance, since the exact distribution of $n_i = N\hat{p}_i$ is binomial with parameters N and p_i, we have by Markov's inequality

$$P(|\hat{p}_i - p_i| > c) = P(|N\hat{p}_i - Np| > cN) \le \frac{p_i(1-p_i)}{c^2 N}$$

Hence, by Bonferroni's inequality (9),

$$P\left(\bigcap_{i=1}^{k}\left\{|\hat{p}_i - p_i| \le c\right\}\right) \ge 1 - \frac{1}{c^2 N}\sum_{i=1}^{k} p_i(1-p_i).$$

Furthermore, using for instance that $\sum p_i(1-p_i) \le 1$ or that $\sum p_i(1-p_i) \le k/4$ gives, respectively

$$P\left(\bigcap_{i=1}^{k}\left\{|\hat{p}_i - p_i| \le c\right\}\right) \ge 1 - \frac{1}{Nc^2} \qquad (18)$$

or

$$P\left(\bigcap_{i=1}^{k}\left\{|\hat{p}_i - p_i| \le c\right\}\right) \ge 1 - \frac{k}{4Nc^2} \quad (19)$$

If, instead of the Markov inequality, we use an exponential type inequality of Hoeffding (see e. g., Serfling (1980), p. 75 [7]),

$$P\left(|\hat{p}_i - p_i| \ge c\right) \le 2e^{-2Nc^2} \quad (20)$$

we obtain by Bonferroni's inequality

$$P\left(\bigcap_{i=1}^{k}\left\{|\hat{p}_i - p_i| \le c\right\}\right) \ge 1 - 2ke^{-2Nc^2} \quad (21)$$

The formulae (18) - (19) - (21) immediately give the gollowing statements:

$$\text{For } D_i = \left\{\hat{p}_i - \frac{1}{\sqrt{N\alpha}} \le \hat{p}_i \le \hat{p}_i + \frac{1}{\sqrt{N\alpha}}\right\},$$

$$i = 1,\ldots,k \text{ we have } P\left(\bigcap_{i=1}^{k} D_i\right) \ge 1 - \alpha. \quad (22)$$

$$\text{For } E_i = \left\{\hat{p}_i - \frac{1}{2}\sqrt{\frac{k}{N\alpha}} \le p_i \le \hat{p}_i + \frac{1}{2}\sqrt{\frac{k}{N\alpha}}\right\},$$

$$i = 1,\ldots,k \text{ we have } P\left(\bigcap_{i=1}^{k} E_i\right) \ge 1 - \alpha. \quad (23)$$

$$\text{For } F_i = \left\{\hat{p}_i - \sqrt{\frac{\log\left(\frac{2k}{\alpha}\right)}{2N}} \le p_i \le \hat{p}_i + \sqrt{\frac{\log\left(\frac{2k}{\alpha}\right)}{2N}}\right\},$$

$$i = 1,\ldots,k \text{ we have } P\left(\bigcap_{i=1}^{k} F_i\right) \ge 1 - \alpha \quad (24)$$

Based on the above analysis, we have the following confidence intervals to examine : (I): formulae (10)-(11), (II):formulae (12)-(13), (III):formulae (14)-(15) (or(16),(17)), (IV a,b,c): formulae (22)-(23)-(24).

3. Examples (for $\alpha = 0.05$)

3.1. An Example with Four Classes, N=1176

We reuse the example in Fitzpatrick and Scott (1987) but for $\alpha = 0.05$ which is more common confidence level than $\alpha = 0.$, as used in their paper. The data are

i	n_i	p_i
1	458	0.39
2	412	0.35
3	212	0.18
4	94	0.08

In libraries these could be estimated fractions of books in the library with 1, 2, 3, or more than 3 authors.

According to (I), we find $p_1 \in [0.354, 0.426]$, $p_2 \in [0.315, 0.385]$, $p_3 \in [0.152, 0.208]$, $p_4 \in [0.060, 0.100]$ with probability at least 0.95. The absolute length (*AL*) resp. the relative length (*RL*) of these intervals is $\left(RL = AL / \hat{p}_i\right)$:

	1	2	3	4
AL	0.072	0.070	0.056	0.040
RL	0.185	0.200	<u>0.311</u>	<u>0.500</u>

According to (II), we find $p_1 \in [0.354, 0.426]$, $p_2 \in [0.316, 0.386]$, $p_3 \in [0.154, 0.210]$, $p_4 \in [0.062, 0.102]$ with probability at least 0.95. Now

	1	2	3	4
AL	0.072	0.070	0.056	0.040
RL	0.185	0.200	<u>0.311</u>	<u>0.500</u>

According to (III), we find (use formula (16)): $p_1 \in [0.357, 0.423]$, $p_2 \in [0.317, 0.383]$, $p_3 \in [0.147, 0.213]$, $p_4 \in [0.047, 0.113]$, with probability at least 0.95. Now

	1	2	3	4
AL	0.066	0.066	0.066	0.066
RL	<u>0.169</u>	<u>0.189</u>	0.367	0.825

Finally, we examine the intervals in case(IV). (IVa)

p_1 [0.26, 0.52], p_2 [0.22, 0.48], p_3 [0.05, 0.31], p_4 [0.00, 0.21] with probability at least 0.95, and we have

	1	2	3	4
AL	0.26	0.26	0.26	0.21
RL	0.667	0.743	1.444	2.625

This happens to be also the case for (IVb), since k=4.

(IVc)

p_1 [0.344, 0.436], p_2 [0.304, 0.396], p_3 [0.134, 0.226], p_4 [0.034, 0.126] with probability at least 0.95. We have

	1	2	3	4
AL	0.092	0.092	0.092	0.092
RL	0.236	0.263	0.511	1.15

The performance of the cases (IVa, b) is very bad and it will become worse in the next examples (where N is lower and k is higher). We, therefore will not consider the cases (IVa, b) anymore. (IVc) is also very bad. The underlined scores are the best.

3.2. An Example with Four classes, N=300

Interpreting the above data for N=300, we find (we only present the AL and RL results) :

(I)	1	2	3	4
AL	0.140	0.138	0.110	0.078
RL	0.359	0.394	<u>0.611</u>	<u>0.975</u>

(II)	1	2	3	4
AL	0.142	0.138	0.112	0.081
RL	0.364	0.394	0.622	1.013

(III)	1	2	3	4
AL	0.130	0.130	0.130	0.130
RL	<u>0.333</u>	<u>0.371</u>	0.722	1.625

(IV)	1	2	3	4
AL	0.184	0.184	0.184	0.172
RL	0.472	0.526	1.022	2.15

3.3. An Example with Sixteen Classes, N=324

Let us consider the population of the visitors of a library, divided according to sex and age (8 categories : 1: ≤ 10 years, 2:11–20 years, 3:21–30 years, 4:31–40 years, 5: 41–50 years, 6: 51–60 years, 7: 61–70 years, 8: >70 years old). The sampled data are:

	1	2	3	4	5	6	7	8	totals
M	11	21	39	35	22	20	10	10	168
F	10	19	35	40	24	15	7	6	156
totals	21	40	74	75	46	35	17	16	324=N

The $\hat{p}_i$ are the follows

	1	2	3	4	5	6	7	8
M	0.034	0.065	0.120	0.108	0.068	0.061	0.031	0.031
F	0.031	0.059	0.108	0.123	0.074	0.046	0.022	0.019

We have, for (I)

AL	1	2	3	4	5	6	7	8
M	0.060	0.080	0.106	0.102	0.082	0.078	0.056	0.056
F	0.056	0.078	0.102	0.108	0.086	0.068	0.046	0.041

RL	1	2	3	4	5	6	7	8
M	1.765	1.231	0.883	0.944	1.206	1.279	1.806	1.806
F	1.806	1.322	0.944	0.878	1.162	1.478	2.091	2.158

For (II), we have

AL	1	2	3	4	5	6	7	8
M	0.066	0.086	0.110	0.106	0.088	0.084	0.064	0.064
F	0.064	0.082	0.106	0.112	0.090	0.076	0.056	0.052

RL	1	2	3	4	5	6	7	8
M	1.941	1.323	0.917	0.981	1.294	1.377	2.065	2.065
F	2.065	1.390	0.981	0.911	1.216	1.652	2.545	2.737

For (III) we have

AL	1	2	3	4	5	6	7	8
M	0.097	0.126	0.126	0.126	0.126	0.124	0.094	0.094
F	0.094	0.122	0.126	0.126	0.126	0.109	0.085	0.082

RL	1	2	3	4	5	6	7	8
M	2.853	1.938	1.050	1.167	1.853	2.033	3.032	3.032
F	3.032	2.068	1.167	1.024	1.703	2.370	3.864	4.316

Finally, for (IV) we have

AL	1	2	3	4	5	6	7	8
M	0.134	0.165	0.2	0.2	0.168	0.161	0.131	0.131
F	0.131	0.159	0.2	0.2	0.174	0.146	0.122	0.119

RL	1	2	3	4	5	6	7	8
M	3.941	2.538	1.667	1.852	2.471	2.639	4.226	4.226
F	4.226	2.695	1.852	1.626	2.351	3.174	5.545	6.263

From these results it is clear that (IV) performs very badly and hence will not be used further on. (III) only performs best in case k is small and for the larger $\hat{p}_i$'s. Case (I) is best in all the other cases. So we strongly advise to use the intervals (10) in almost all cases. Only try also the Fitzpatrick-Scott intervals when k is low and for the larger $\hat{p}_i$'s. In these cases we will still be in need of (10) as the next result shows.

4. A Drawback of the Fitzpatrick-Scott Confidence Intervals

We restrict our attention to $\alpha = 0.05$. In order for the confidence intervals in (16) to be smallest for all $\hat{p}_i$ (hence smaller than the intervals in (10)), we have the condition (by (16) and (10)):

$$\frac{1.13}{\sqrt{N}} \le z(0.05/2k)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}},$$

for all $i=1, \ldots, k$. This yields the condition

$$\hat{p}_i - \hat{p}_i^2 \ge \frac{1.2769}{z^2(0.05/2k)}. \qquad (25)$$

Now the equation

$$-x^2 + x - \frac{1.2769}{z^2(0.05/2k)} = 0$$

has two real roots when $z^2(0.05/2k) \ge 5.1076$. But $z^2 < 5.1076$ implies $z < 2.26 = z(0.0119)$.

So $z^2(0.05/2k) < 5.1076$ only for $\frac{0.05}{2k} > 0.0199$, which implies $k < 2.1$. Since $k \ge 3$ we hence have two real roots. These are positive:

$$\frac{1}{2} \mp \frac{1}{2}\sqrt{1 - \frac{5.1076}{z^2(0.05/2k)}},$$

one above 0.5 and one below 0.5. Hence we have the condition

$$\frac{1}{2}\left(1-\sqrt{1-\frac{5.1076}{z^2(0.05/2k)}}\right)\leq \hat{p}_i$$

$$\leq \frac{1}{2}\left(1+\sqrt{1-\frac{5.1076}{z^2(0.05/2k)}}\right). \qquad (26)$$

So if the left hand side of (26) is larger than 1/k then the necessarily there exist $i=1, \ldots, k$ for which $\hat{p}_i$ does not satisfy (26) and hence the intervals in (10) are then smaller than the intervals in (16). We now show that this is always the case, for any $k \geq 3$. Indeed, the condition

$$\frac{1}{2}\left(1-\sqrt{1-\frac{5.1076}{z^2(0.05/2k)}}\right)>\frac{1}{k}$$

is equivalent with

$$z^2(0.05/2k)<\frac{5.1076}{\frac{4}{k}\left(1-\frac{1}{k}\right)}. \qquad (27)$$

This is true as the next table shows

k	LHS of (27)	RHS of (27)
3	5.736	5.746
4	6.250	6.810
5	6.631	7.981
6	6.954	9.194
7	7.290	10.428
8	7.508	11.675
9	7.673	12.929
10	7.896	14.188

An exact proof now follows: using Feller (1968) [2], p. 175, we note that

$$x(1-\Phi(x))\leq \frac{e^{-x^2/2}}{\sqrt{2\pi}} \qquad (28)$$

where Φ denotes the cumulative standard normal distribution.

Since $z(\alpha)=\Phi^{-1}(1-\alpha)$ we have to show that

$$\Phi^{-1}\left(1-\frac{0.05}{2k}\right)\leq \sqrt{\frac{5.1076}{\frac{4}{k}\left(1-\frac{1}{k}\right)}},$$

or equivalently :

$$e^{-\frac{0.63845}{x(1-x)}}\leq 0.0708\sqrt{\frac{x}{1-x}}, \qquad (29)$$

where $x=\frac{1}{k}$, $k=3,4\ldots$ This can be seen to be true using elementary calculus.

Hence, there are always intervals in (16) that are larger than the intervals in (10). The same can be proved for all $\alpha \leq 0.032$ (based on (15)). Now condition (29) becomes

$$e^{-\frac{C(\alpha)}{x(1-x)}}\leq \alpha\sqrt{\frac{x}{1-x}},$$

were $C(\alpha)$ is a positive function of α. This again, can be seen using elementary calculus. Hence in all these cases it is so that there are always intervals of the Fitzpatrick-Scott type that are larger than the ones in (10)

Conclusion

Always use the intervals in (10) - (11). Also check the intervals in (14) - (15) in the case of few classes and then only for the larger p_i's.

5. Sample Sizes for Multinomial Proportions

From the above formulae for large sample simultaneous confidence intervals for multinomial proportions, we want to determine the minimal sample size N for which these k intervals have specified lengths.

Let L_i be the length of an interval around $\hat{p}_i$. Then one can express that, either

$$L_i \leq \beta \qquad (30)$$

or

$$L_i \leq \beta \hat{p}_i \qquad (31)$$

for a fixed $\beta \in]0, 1[$, for all $i=1, \ldots, k$. In the first case one requires that all intervals are shorter than a fixed absolute length. In the second case one requires that all intervals are shorter than a fixed

proportion of their midpoints (see also the study of Tortora (1978) [9]).

We will now deduce some minimal values of the sample size N in order to have (30) or (31) in the cases (I), (II) and (III). As mentioned above, the case (I) is the most important one. In this case one obtains, for (30):

$$2z(\alpha/2k)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}} \leq \beta$$

or

$$N \geq \frac{4z^2(\alpha/2k)\hat{p}_i\left(1-\hat{p}\right)}{\beta^2} \qquad (32)$$

for all i=1, ..., k.
Since

$$\hat{p}_i\left(1-\hat{p}_i\right) \leq \frac{1}{4}$$

for all $\hat{p}$ [0, 1] we can use the following practical (conservative) rule of thumb:

$$N \geq \frac{z^2(\alpha/2k)}{\beta^2} \qquad (33)$$

Other (more intricate) "worst case" studies (also related to determining the maximal α) can be found in Thompson (1987)[8] and Angers (1989)[1]. The latter paper discusses the two possible approaches: using $\hat{p}_i$, hence involving a two-stage sampling procedure or applying the "worst case" approach. In case (31) is required we find

$$N \geq \frac{4z^2(\alpha/2k)\left(1-\hat{p}_i\right)}{\beta^2\,\hat{p}_i} \qquad (34)$$

for all i=1, ..., k.
For case (II), the formulae become rather intricate. For (30) one finds, in the worst case, the simple formula

$$N \geq \frac{z^2(\alpha/2k)}{\beta^2} - z^2(\alpha/2k). \qquad (35)$$

Note that this almost equals the value in (33). For case (III), one has for (30)

$$N \geq \frac{z^2(\alpha/4)}{\beta^2} \qquad (36)$$

and for (31):

$$N \geq \frac{z^2(\alpha/4)}{\beta^2\hat{p}_1^2} \qquad (37)$$

for all i = 1, ..., k.

Examples

Let us investigate the formulae (33), (35) and (36) which represent the sample sizes under condition (30), not involving the values $\hat{p}_i$. We will use the previous examples.

Example 1 : $\alpha = 0.05$, $k = 4$, $\beta = 0.04$ (this represents a 10% value of the largest fraction $\hat{p}_i = 0.39$). (33) becomes $N \geq 3899$, (35) becomes $N \geq 3893$ and (36) becomes $N \geq 3136$.

Example 2 : $\alpha = 0.05$, $k = 4$, $\beta = 0.008$ (this represents a 10% value of the smallest fraction $\hat{p}_i = 0.08$). We have now respectively $N \geq 97469$, $N \geq 97463$ and $N \geq 78400$

Example 3: $\alpha = 0.05$, $k = 16$, $\beta = 0.0$. We have now respectively $N \geq 87025$, $N \geq 87017$, $N \geq 50176$

As one can see, the values in the last two examples are very high. The conclusion is that in order to estimate the small porportions rather accurately, one needs very high sample sizes. If one only wants to estimate the largest proportions in an accurate way, much smaller sample sizes suffice (cf. example 1).

Now let us compare formula (32) with the "rule of thumb" (33). To see the reduction of N in (32) one must calculate

$$\max_{i=1,\ldots,k} 4\hat{p}_i\left(1-\hat{p}_i\right).$$

For the three examples above this is 0.9516, 0.9516 and 0.4315 respectively. Conslusion: it pays off to use (32) in case the largest $\hat{p}_i$ is still rather small (here 0.123). This will often be the case when k is high.

6. An Improvement of $\hat{p}_i$ - Dependent Confidence Intervals Based on Sample Size Results

Formulae (32), (34) and the ones that can be derived from the intricate formulae (12) and (13) have to be based on two stage sampling. First take an initial sample to estimate $\hat{p}_i$, ..., $\hat{p}_k$ and then determine the definite sample size by applying (e. g.) (32) or (34). Let us fix the ideas on formula (34) (but our argument works equally well for all other formulae involving $\hat{p}_i$, i=1, ..., k). One must take

$$N \geq \max_{i=1,\ldots,k} \frac{4z^2(\alpha/2k)\left(1-\hat{p}_i\right)}{\beta^2 \hat{p}_i} \tag{38}$$

in order to have the length o all confidence intervals inferior to $\beta\hat{p}_i (i=1,\ldots,k)$ There is a certain inefficiency in this since for k - 1 intervals, we are probably over sampling. We sample in our "optimal" way if the right hand side of (38) is independent of i = 1, ..., k. This is not so, but can be achieved when allowing for variable α_i instead of $\frac{\alpha}{2k}$. The method is as follows: take

$$N = \max_{i=1,\ldots,k} \frac{4z^2(\alpha/2k)\left(1-\hat{p}_i\right)}{\beta^2 \hat{p}_i}. \tag{39}$$

We then look for confidence intervals of the type

$$\left[\hat{p}_i - z(\alpha_i)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}},\ \hat{p}_i + z(\alpha_i)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}}\right] \tag{40}$$

such that, for all i =1, ..., k

$$N = \frac{4z^2(\alpha_i)\left(1-\hat{p}_i\right)}{\beta^2 \hat{p}_i} \tag{41}$$

hence for these intervals we have

$$2z(\alpha_i)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}} = \beta\hat{p}_i, \tag{42}$$

for all i =1, ..., k, hence they are all of the form

$$\left[\hat{p}_i - \beta\hat{p}_i, \hat{p}_i + \beta\hat{p}_i\right]. \tag{43}$$

Now (41) implies

$$z(\alpha_i) = \sqrt{\frac{N\beta^2 \hat{p}_i}{4\left(1-\hat{p}_i\right)}}$$

and, since we choose N as in (39), we have now that $\alpha_i \leq \frac{\alpha}{2k}$ (and < for most i) so that

$$\sum_{i=1}^{k} \alpha_i < \frac{\alpha}{2}$$

This yields $100\left(1-2\sum_{i=1}^{k}\alpha_i\right)$% simultaneous confidence intervals better(and usually much better) than $100(1-\alpha)$%. Note that our intervals are now of the form (43), hence have lengths proportional to $\hat{p}_i$.

This approach is similar to the one in Thompson (1987) [8] but in a "reverse" way : there, also $(\alpha_1,\ldots,\alpha_k)$ is considered but, if for some N, $\sum_{i=1}^{k}\alpha_i < \alpha$, the procedure is repeated with smaller N, until the smallest N is found for which $\sum_{i=1}^{k}\alpha_i \leq \alpha$. In short, one keeps α and lowers N; here we keep N and lower α.

Example : $\alpha = 0.05, \beta = 0.5$ (which is very accurate for these conservative intervals (see the previous section) and the example of the previous section:

i	$\hat{p}_i$
1	0.39
2	0.35
3	0.18
4	0.08

The maximum value of $\frac{4z^2(\alpha/2k)\left(1-\hat{p}_i\right)}{\beta^2 \hat{p}_i}$ is 1150.

Now require

$$\frac{4z^2(\alpha_i)\left(1-\hat{p}_i\right)}{\beta^2\hat{p}_i}=1150$$

for all i=1,2,3,4. This yields

$$z(\alpha_1)=6.779$$

$$z(\alpha_2)=6.221$$

$$z(\alpha_3)=3.972$$

$$z(\alpha_4)=2.500 \text{ (this had to be so!)}$$

This yields α_1 and $\alpha_2 << 0.000005$, $\alpha_3 < 0.00005$, and $\alpha_4 = 0.0062$. We have confidence intervals $A_1=[0.2925,0.4875]$ for p_1, $A_2=[0.2625,0.4375]$ for p_2, $A_3=[0.135,0.225]$ for p_3 and $A_4=[0.06,0.1]$ for p_4.

These intervals satisfy, using (6) and (2)

$$P\left(\bigcap_{i=1}^{4}A_i\right)\geq 1-2\sum_{i=1}^{4}\alpha_i$$

$$>1-0.0125\,.$$

Hence the confidence is more than $100(1-\alpha')\%$ with $\alpha'=0.01252$; we started from $\alpha=0.05$ So this is a substantial improvement, whilst we have not changed the sample size N.

7 Practical Techniques - Summary

Estimating k multiple fractions (as e.g. the several examples in section 1) can be done by establishing simultaneous confidence intervals. Suppose a sample of size N yields estimates for the k fractions as $\hat{p}_1,\hat{p}_2,\ldots,\hat{p}_k$. Then the k intervals

$$\left[\hat{p}_i-z(\alpha/2k)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}},\right.$$

$$\left.\hat{p}_i+z(\alpha/2k)\sqrt{\frac{\hat{p}_i\left(1-\hat{p}_i\right)}{N}}\right]$$

are simultaneously valid for a confidence level of at least $100(1-\alpha)\%$. These should be used in any case. N should be increased if the obtained intervals are found to be too large.

In case some intervals remains unsatisfactory large, one can try the following remedies:

1. In case there are only a few classes and for the larger $\hat{p}_i$'s : try also the intervals (16) or (17), which are sometimes smaller, for some i.

2. Leave small proportions $\hat{p}_i$ out of consideration. They are, usually, not so important, and leaving them out reduces k and hence the size of all the other simultaneous confidence intervals.

Acknowledgement :

The authors are grateful to H. J., Voorbij (Royal Library, The Hague, The Netherlands) for mentioning the problem of sampling multinomial data)

References

1. Angers, C. (1989). Note on quick simultaneous confidence intervals for multinomial proportions. *The American Statistician* 43(2), 91. Correction in 44 (1), 65.
2. Feller, W. (1968). *An introduction to probability theory and its application.* Vol. 1, 3rd edition, Wiley, New York.
3. Fitzpatrick, S. and A. Scott (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association* 82(399), 875-878.
4. Goodman, L. A. (1965) An simultaneous confidence intervals for multinomial proportions. *Technometrics* 7(2), 247-254.
5. Huschens, S. (1990). Necessary sample sizes for categorical data. *Statistical Papers* 31, 47-53.
6. Quesenberry, C. P. and D. C. Hurst (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* 6(2), 191-195.
7. Serfling, R. J. (1980). Approximation theorems of mathematical statistics. John Wiley, New York.
8. Thompson, S. K. (1987). Sample size for estimating multinomial proportions. *The Americal Statistician* 41(1), 42-46.
9. Tortora , R. D. (1978). A note on sample size estimation for multinomial populations. *The American Statistician* 32(3), 100-102.

Indian Abstracting and Indexing Services in Science and Technology: An Analysis

Debal C Kar and P Bhattacharya
Tata Energy Research Institute, 102 Jor Bagh,
New Delhi - 110003, India

The present study makes a census and critically analysis of the Abstracting and Indexing (A & I) services in Science and Technology (S & T) published from India since 1943. The objective of the study is to help knowing what services are existing in a particular discipline which may serve as the basic source of information to facilitate research activities in S & T. This article analyses these services in respect of periodicity, nature of publication, chronological development, subject coverage, document type included for services, sponsorship, arrangement of entries etc. A total of 187 services have been identified in S & T, out of which 106 are abstracting and 81 are indexing services.

1. Introduction

Abstracting and indexing (A & I) services are the most important secondary publications. They not only keep users abreast of the current progress in their subject field of specialization or interest but also serve as record of research output for posterity. Whereas both abstracting services and indexing services serve as a tool to identify, select and locate a document for use, the abstracting services give much more information to facilitate the selection process. The need for such services are felt more in the field of Science and Technology (S & T) and hence these occupy an unique position in science communication system so far as current awareness services and retrospective search services for scientists and technologists are concerned. An attempt has been made in this endeavour to study the A & I services from different point of view.

2. Objective

The objective of the present study is to make a census of the A & I services in S & T, published from India, to help knowing what services are existing in particular discipline which may serve as the basic source of information to facilitate research activities; to analyze these services in respect to periodicity, nature of publication, chronological development, subject coverage etc; to identify the weaker and stronger disciplines on the basis of analysis, so suggestion for future programme can be made.

3. General Review

Survey of the A & I services in India by the Survey and Planning of Scientific Research Unit, CSIR[3], recorded more than 60 A & I periodicals issued by the special libraries and documentation centres in India. The compilers attempted to make the list exhaustive as far as possible. A large number of these services were issued in mimeographed form and they were generally known as local documentation list. It was unfortunate that though some of those lists were valuable for their design and coverage, very little attention had been given to standardization and bibliographical control for development of coordinated services throughout India.

Mr R Pal [9] had also traced a total of 60 A & I periodicals. However, his study was concentrated on a particular periodical named Indian Science Abstract. Ms S Relan [10] had established a total number of 150 Indian A & I services. But it took account of current awareness list also. Chakravarty [2] had also conducted a study of International A &

I services in S & T, but it did not cover Indian A & I services. Earlier studies in this area were also conducted by Bhattacharya [1], Guha [4,5], Kesharwani [6], Kumar [7] and Neelameghan [8]. But their fields were limited to particular S & T discipline, i.e, Kesharwani covered only environmental science; Neelameghan had emphasized on medical sciences. The present study has covered whole gamut of A & I services in S & T published from India.

4. Subject Coverage

A total of 187 number of services has been identified, out of which 106 are abstracting services and 81 are indexing services, but very few of them can be considered as a national level services. The services taken together cover a wide spectrum of subjects (Table 1). The stress is decidedly on Engineering and Technology (E & T) topics which between them share as many as 144 services (about 77%) out of the total services (187) which have been covered.

Table 1. Subject wise distribution of the services covered

Subject covered	No. of Services	% of services
Agricultural Science and related discipline	9	4.8
Engineering & Technology	144	77.0
Science & Technology	21	11.2
Biological Science	5	2.7
Medical Science & related disciplines	8	4.3
Total	**187**	**100.0**

5. Analysis

5.1 Type of Documents

It has been found that primary periodicals has been taken for inclusion in maximum number of abstracting services (Table 2). About 50% of total services are covering periodical literature. The other type of source documents are mostly primary and secondary services, followed by technical reports, theses, patents, conference proceedings, newspaper clippings and standards.

Table 2. Services according to type of documents included

Sources included	Number of services
1. Periodicals	120
2. Other primary & secondary source	75
3. Indian primary & secondary sources	10
4. Scientific & technical reports	7
5. Theses	5
6. Patents	4
7. Conference proceedings	4
8. Newspaper clippings	3
9. Standards & Specifications	2

5.2 Chronological Analysis

It was revealed in the chronological analysis (Table 3 and Fig.1) of the services which started during the period from 1943 to 1990, the maximum number of A & I services started during the period of 1971-75 (Total No.54) followed by 1981-85 (Total No.37). Table 3 also shows the subject covered of the services started during the particular period and their trends.

Table 3. Chronological analysis of the services and subjects covered

Year of starting	No. of services	%	Subjects covered
1943-60	19	10.2	Agriculture; Irrigation; Medical sciences; Railway & highway engineering; Textile industry.
1961-65	14	7.5	Building industry; Defence science; Fisheries etc.
1966-70	28	15.0	Communication; Instrumentation; Glass & Ceramic industries; Military & Naval engineering; Textile industry etc.
1971-80	54	28.9	Chemical Technology; Electronics; Industrial management; Instrumentation; Leather industry; Agriculture; Metallurgy; Mining & mineral dressing; Public health engineering; Nuclear technology etc.
1976-80	21	11.2	Chemical technology; Botany; Electrial engineering; Food sciences; Metallurgy; Pharmaceuticals; Oceanography; Space flight engineering etc.
1981-85	37	19.8	Agriculture; Medical sciences; Non-conventional sources of energy; Petroleum technology; Rural development etc.
1986-90	14	7.5	Medical sciences; Texicology; Metallurgy; Military science; Environmental science; Agriculture etc.

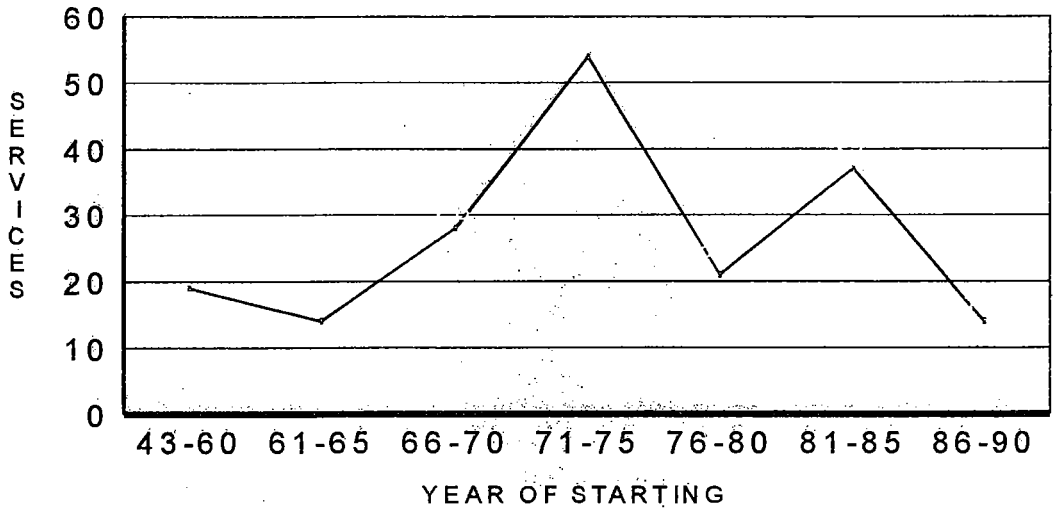


Fig.1 Chronological analysis: indexing & abstracting services

A study was made to find out the subject areas of the services started during peak periods 1971-75 and 1981-85 (Table 4).

Table 4: Subjectwise analyis of the services started publishing during the periods 1971-75 and 1981-85.

Subject, discipline	Total No. of Services 1971-75	1981-85
Agricultural Sciences	6	4
Cement Industry	2	-
Computer Sciences	1	-1
Electronics	4	3
Energy Sources	1	10
Engineering & Technology (General)	7	-
Environmental Sciences	-	7
Medical Sciences	-	4
Military Sciences	7	-
Mining Engineering	2	2
Occupational Health	3	-
Petroleum Sciences	2	5
Science & Technology (General)	11	-
Space Sciences	8	-
Town Planning	-	11
Total	**54**	**37**

An observation on Table 4 indicates interesting trend in the field of some specific disciplines. During the years 1971-75 apart from general S&T the major number of services appeared in the field of military science, space science, engineering and agriculture and a need was felt for secondary (A&I) services in these areas. This was the period when more emphasis were laid on defence activities after the Indo-Pak war.

During the period 1981-85 the emphasis was given on various kinds of energy including alternative sources of energy, environmental sciences, agricultural science, medical science and petroleum technology. These disciplines got a new boost during this period due to worldwide change in economic orders. International movement for global environmental protection started during this period. Analysis of the sponsoring agencies shows that during 1971-75 period most of the services were brought out by defence establishment and R&D labs, Council of Scientific and Industrial Research Labs (CSIR), Indian Council of Agricultural Research (ICAR). During 1981-85 period, Oil & Natural Gas Commission (ONGC), National Medical Library (NML). Tata Energy Research Institute (TERI), Indian School of Mines, Indian Bureau of Mines, Electronics and Radar Development establishment have started bringing out the services.

5.3 Sponsorship

The services were analyzed according to the sponsoring body. It reveals that laboratories and research institutions under the CSIR publishes maximum number of A&I services. (Table 5 and Fig.2)

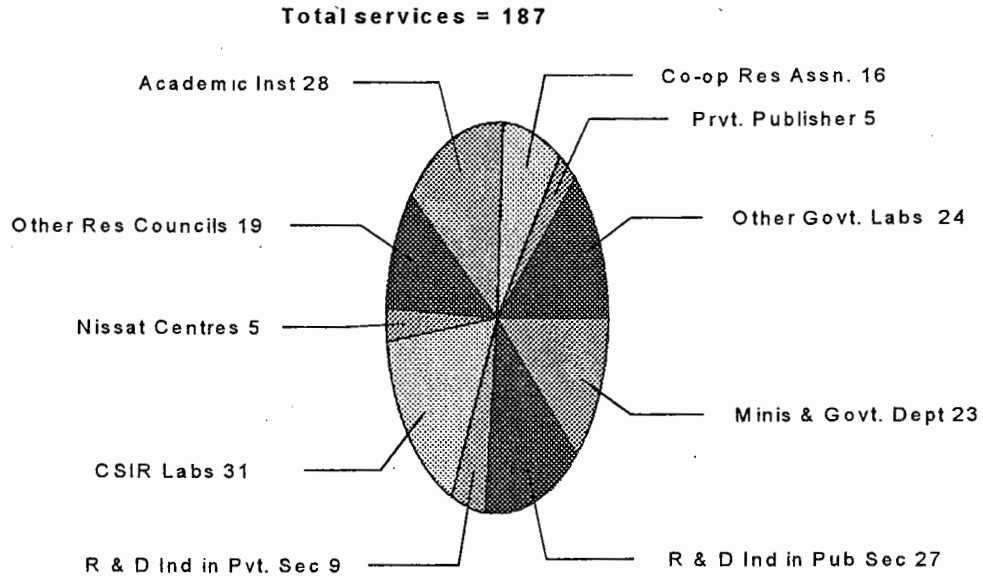


Figure 2. Sponsoring agencies : distribution of services

Table 5. Sponsoring agencies of the services

Agencies for publication	Number of Services	% of services
Ministries & Govt. Dept.	23	12.3
R & D Industries in Public Sector	27	14.4
R & D Industries in Private Sector	9	4.8
Laboratories & Research Centres CSIR	31	16.6
Other Laboratories under Government	24	12.8
Private Publishers	5	2.7
Cooperative Research Association	16	8.6
University & Academic Institute	28	15.0
Council of Research other than CSIR	19	10.2
NISSAT Sectoral Centres	5	2.7
Total	**187**	**100**

5.4 Periodicity

Table 6, Figure 3 and Figure 4 show the trend of periodicity of the A&I services. Out of 106 abstracting services, 44 services are published quarterly and 35 are monthly services whereas in case of indexing services 49 services are published monthly.

Table 6. Publication. periodicity of the A & I service

Periodicity	No. of Services Abstracting	Indexing
Weekly Service	1	2
Fortnightly Service	4	4
Monthly Service	35	49
Biomonthly Service	13	7
Quarterly Service	44	8
Half Yearly Service	4	1
Annual Service	2	3
Irregular Service	3	7
	106	**81**

5.5 Arrangement of Entries

The entries are arranged under broad subject headings alphabetically in most cases (155 nos.). Only one service is arranged by Zoological nomenclature (Bibliography of Indian Zoology). 25 Services are arranged by classified using Universal Decimal Classification (UDC) scheme and 5 services using Colon Classification (CC) scheme. Arrangement of entries under broad subject by majority of services indicated that these services are designed more for a current awareness function rather than specific search function.

5.6 Form of Publication :

Analysis of services shows that in case of

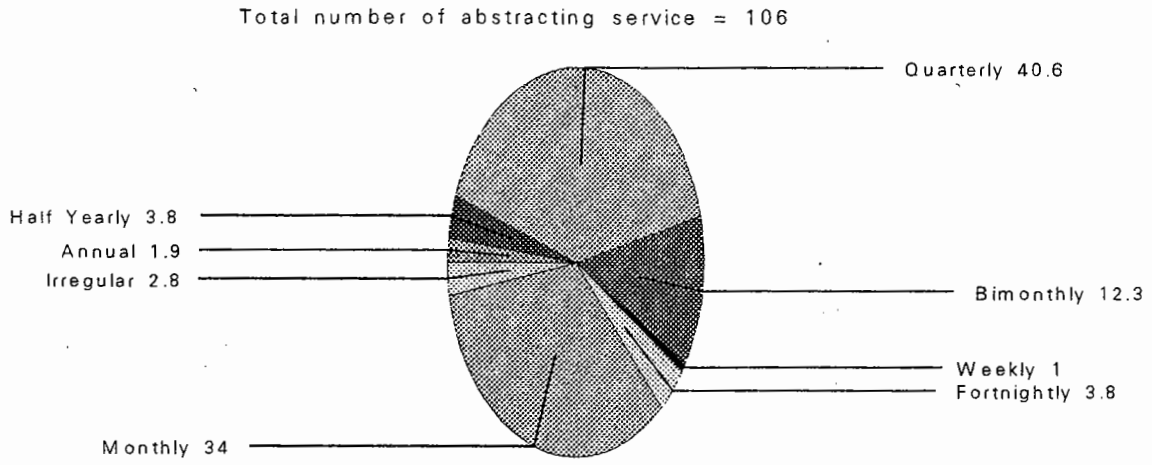


Figure 3. Periodicity of the abstracting services

abstracting services (106), the number of services in printed form (55) and non printed form (51) are almost same, whereas in case of indexing services (81) most of them are in non printed form (67), printed indexing services are only 14 numbers, which shows that the indexing services are designed mostly to perform a temporary function - a feature of current awareness service.

5.7 Geographical Distribution

The Geographical wise distribution of the services shows that Delhi city publishes the maximum number of services in S&T (Table 7 and Fig.5). The services brought out by different organizations in Delhi shows that 27 organizations are responsible for bringing out 41 services. Some organizations (viz. NML, TERI etc.) bring out more than one service. The concentration of services at Delhi is due to the fact that Delhi being the national capital, majority of national level research institutions, societies, government departments are located at Delhi. 83 services (45%) are located only at five major cities e.g. Delhi, Bangalore, Bombay, Calcutta, Madras and rest 104 (55%) are located at different part of the country.

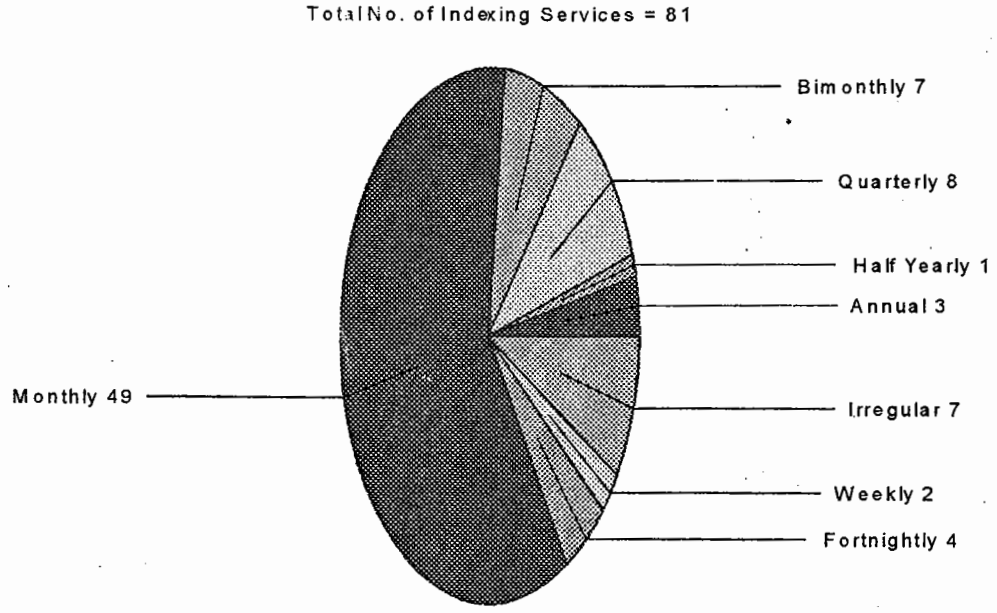


Figure 4. Periodicity of indexing services

Table 7. Distribution of services among major cities

	Delhi	Bangalore	Bombay	Calcutta	Madras	Others	Total
No. of services	41	16	12	8	6	104	187
% of services	21.9	8.5	6.4	4.2	3.2	55.6	100

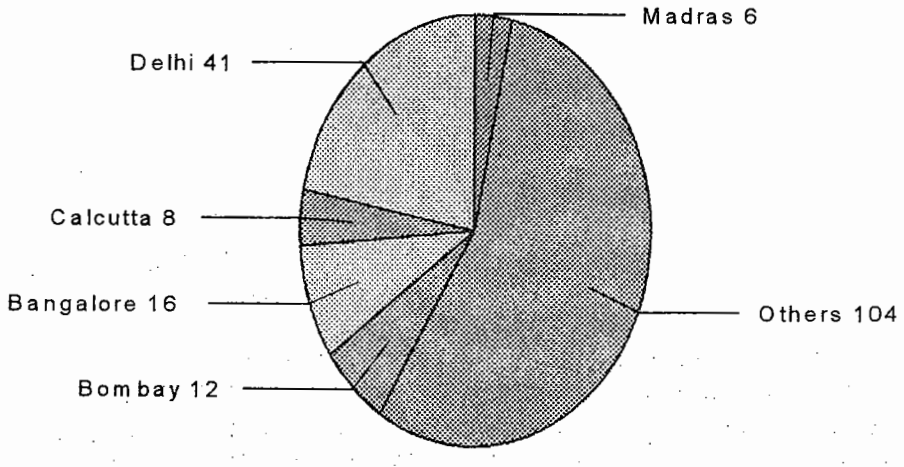


Figure 5. Distribution of services

6. Observation

It has been found that in most cases subject specialists are involved in the publication of abstracting services whereas librarians and information specialists are involved in the publication of indexing services. Less than 50% of the total services are priced whereas the rest are available on exchange/ complementary/request basis to other institutions.

7. Study of Selected Indian A&I Services

The objective of this study is to find out the coverage of the A&I services for their selectivity i.e. whether these services cover all the literature published in a particular periodical or certain amount of selectiveness is there. Normally the services which have more circulation may be considered as useful to the users. The services which have a circulation of a minimum of 200 have been considered for the purpose of this study.

A particular volume of each of the services was selected for the study. The list of journals covered by these services was checked and one particular title was chosen by selecting a well known primary journal to find out the total number of articles published by them in a particular volume (Table 8). The contents of the particular volume of the journal was noted from the contents page. Each title was compared with the abstracting services to assess its coverage. Table 8 shows the selectivity coverage of the Indian A & I services. All the services are selective, whereas Indian Geoscience Abstracts covers almost all articles in their services.

8. Conclusion

From the above analysis of the services, it is evident that the stress is decidedly on engineering and technological areas. The number of services is limited as far as pure science subjects are concerned. In applied sciences, like metallurgy, space flight engineering, textile industry, petroleum engineering and technology, mining and mineral dressing, pharmaceuticals, electronics, glass and ceramic industries, cement industries, chemical technology are quite rich in abstracting services, but fields like physics, chemistry, zoology, environmental sciences, oceanography, forestry, civil

Table 8. Selectivity of the coverage of the Indian A & I sercives

Service & Vol no.	Selected Journal for comparing with 1	Total no. of articles in the journal at 2	Total no. of articles covered by the service at 1	% of coverage
1	2	3	4	5
1. Food Technology Abstracts Vol. 22, 1987-Vol. 23, 1988 (Monthly)	Journal of Food Science & Technology Vol. 24, 1987 -Vol. 25, 1988	144	56	39
2. Indian Energy Abstracts Vol.7 1988 & Vol.8(1) 1989 (Quarterly)	Indian Forester Vol.114, 1988-89 (Monthly)	180	26	14.4
3. Indian Geo-Science Abstracts 1975 (Annual)	Journal of Geological Society of India Vol. 16, 1975 (Quarterly)	44	41	93.18
4. Medicinal and Aromatic Plants Abstracts Vol 10, 1988(Biomonthly)	Indian Drugs Vol. 24, 1987-Vol. 25,1988 (Monthly)	140	42	30
5. Paryavaran AbstractsVol. 5, 1988-89 (Quarterly)	Indian Journal of Environmental Protection Vol. 7, 1987-88(Monthly)	60	36	60
6. Indian Science Abstract Vol.24, 1988-89 (Fortnightly)	Journal of Geological Society of India (Quarterly) Vol. 31 & 32, 1988	150	100	66

engineering and railway engineering are lagging behind as far as number of services are concerned. Various services which are being published from Central Leather Research Institute (CLRI), Defence Research Development Organization (DRDO) labs, Central Food Technological Research Institute (CFTRI), Ahmedabad Textile Industry's Research Association (ATIRA), Publication and Information Directorate (PID), CSIR, Ministry of Environment, Forest and Wild Life and laboratories and establishments under Space Commission occupy a unique position in their subject coverage.

In the case of national level services it has been found that in most cases coverage of primary periodicals in the secondary services is selective. On an average 40-50% of the primary journal articles are covered in the secondary abstracting service. It has been found that during 1971-75 the maximum number of A&I services had started publishing. Computerization of A&I services has started getting momentum from 1980's.

The study shows that most of the services, even at national level, are selective. As such there is not a single mechanism for the total bibliographical control of S&T literature in general and specific subject literature in particular. If these services have to meet the objective of archival function, it is necessary that the coverage of these services should be broad based.

References

1. Bhattacharya, K: *A survey of abstracting and indexing services available in India and their usefulness for the scientific community at national and international level.* Sixth IASLIC Symposium on Local Documentation lists and their usefulness at national level, Trivandrum. 1965. Part II, 67-84.
2. Chakravorty, A R. *A survey of indexes of abstracting/ indexing services in science & technology.* Unpublished project report submitted to INSDOC for the award of "associateship in documentation and reprography". 1975.
3. CSIR. *Scientific societies in India* (CSIR Survey report no. 3). 1965.
4. Guha, B. *Documentation and information services, techniques and systems.* 1984, Calcutta: World Press Private Limited. 1984.
5. Guha,B. Indian current awareness services. *UNESCO bulletin for libraries,* 22(2),1968; 73-81.
6. Kesharwani, S.K. *Abstracting indexing and current awareness services in environmental science.* Gurgaon: Indian Documentation Service. 1978.
7. Kumar,A. *Indexing and abstracting in scince and technology. Indian Libraries.* 23(1); 1968; 46-55.
8. Neelameghan, A. Abstracting services in medical science. *Annals of Library Science.*2; 1955; 85-96.
9. Pal, R: *Scientific development and abstracting services in India with special reference to Indian Science Abstracts.* Unpublished project report submitted INSDOC for the award of assiciateship in information science. 1985.
10. Relan, S. *Documentation list in science and technology.* Unpublished project report submitted to INSDOC for the award of associateship in information science. 1985.

Constructing Linear Measures from Counts of Qualitative Observations

John Michael Linacre
MESA Psychometric Laboratory, University of Chicago,
5835 S. Kimbark Avenue, Chicago IL 60637, USA

Both disorder and linearity are essential to science. Empirically developed laws demonstrate that less than ideal counts can reflect ideal underlying structures. From the premise that the most understandable ideal structure is linear, Rasch deduced that the mechanism that constructs linear measures from ordered qualitative observations, such as counts, is the logistic ogive. Rasch models are available for dichotomies, Poisson counts and rating scales. These models transform counts into linear measures. The degree to which any particular data cooperate in constructing measures is gauged by quality-control fit statistics. The method is demonstrated with a set of scientometric indicators.

1. Replication and Disorder

Science depends on replication. It was not when Roentgen[11] made the accidental discovery of a "peculiar black line" that science advanced, for "many such researchers had found their photographic plates in the lab unaccountably spoiled." It was only after Roentgen "put the new effect through its paces" and took the time to "check the facts," that X-rays became a scientific advance.

Roentgen replicated the original phenomenon. Replication combines duplication with alteration. Certain conditions are maintained. Certain conditions are different. The identical conditions maintain order. The changed conditions introduce disorder. The more that the essential nature of the original phenomenon, its "information", continues in the face of increasing disorder, the more robust is that information. "On the side of pure speculative theory I suggest that information is measured ... by the order it produces out of disorder. But order of what ? The answer seems to be that each piece of information has value insofar as it relates to the order of other information, and that what we see in mapping [scientific papers onto a knowledge plane] is this basic order."[13]

Thus disorder is necessary to identify and quantify useful information, and useful information permits the establishment of meaningful locations, i.e., useful measurement. Though location on a plane in two-dimensions can be more instructive than location on a line in a single dimension, the common requirement for utility is equal-interval, linear measurement.

2. Linearity and Measurement

The struggle to wrest useful order from the hodgepodge of data that inundates the scientist devolves ultimately into a quest for linear measurement. If we cannot measure a phenomenon in equal-interval units, there is grave doubt that we understand it in any scientific sense. "Length" and "Love" are cases in point. "Indeed, it is a character of all the higher laws of Nature to assume the form of precise quantitative statement."[9]

Linearity requires that measures be obtained along a line in the same sense that measures of length are conceptualized to occur along one line. Further, the measures must be reportable in units, e.g., meters, such that one more unit (meter) implies the same increment wherever along the line the unit is observed. Such measures then have the properties expected of numbers for which the usual rules of arithmetic and statistical analysis are meaningful. The average of a set of numbers can always be obtained, even when the numbers are merely assigned arbitrarily for identification.

For any average to have a well-defined meaning, however, it must be located in the middle of both the numbers and their implied meanings. This requires that the units above the average must be both numerically and substantively equal in size to the units below the average, i.e., the numbers must be linear measures.

Linearity, not mere numerosity, is what nearly all widely used statistical procedures assume of their data. The computation of meaningful means, standard deviations, and least-squares estimates requires linearity in the underlying data. Once linear measures have been constructed, regression analysis, ANOVA and even simple plots have a foundation. Without linear data, it is not known whether the regression coefficients are explaining the underlying phenomenon or merely the non-linearity of the counts. Without linearity, it cannot be determined whether "ceiling" and "floor" effects are substantive findings or merely artifacts of the way the counts were made.

The empirical approach is to collect whatever there is in the way of data and then to hunt for some description that summarizes those data in a simple form. Bradford's Law of Scattering demonstrates equal-interval measurement, linearity, over the central part of its range. Bradford obtained this linearity by arbitrary manipulation of the data. Garfield observes that "Bradford's, Zipf's, Lotka's and Pareto's laws were independently formulated to explain disparate phenomena. It seems more than coincidence that they should closely resemble each other. One wonders if they are not all governed by a single underlying principle."[6] As Bookstein[1] contends, the existence of Bradford's law proves that a more general form of regularity must be present in the data, of which Bradford's law is a special case, or else Bradford could not have discovered his law.

The theory-based approach is to develop the criteria that must be satisfied in order for linear measures to be constructed, and then to learn how to assemble data that satisfy those criteria. The theory-based approach speeds the scientific process because it leads to the elimination of superficially attractive, but fundamentally unproductive, statistical manipulations. It also permits investigators to locate coherent subsets of data within otherwise incoherent sets.

3. Counting : A Qualitative Decision

Linear measurement begins with counting. Linear measures, e.g., grams, are ideals only imperfectly realized in practice. Counts are empirical absolutes. The observer can count ten apples on a tree, or ten papers by an author. Clearly, not all apples are equally large and crisp, nor all papers are of equal value (p. 40).[12] The first aspect of counting to be investigated is thus qualitative. Are the objects counted sufficiently alike to be considered interchangeable in the counting process? Stricter rules lead to more discriminating counting processes. Is a rotten apple counted as an apple? Is a corrected version of an earlier paper counted as another paper?

Air lines have discovered that there is a convenient, approximate, but useful and accurate enough conversion from passenger count to passenger weight. Passenger count is an ordinal, discrete integer summarizing passengers with a wide range of characteristics. Passenger weight is an interval-scale amount that idealizes one aspect of the passengers. Once the qualitative decision has been made as to who to include in the count, the particular details of each passenger become irrelevant.

The essential element in making scientific inferences from counts is that the same mechanism converts counts to linear measures regardless of the couting rule employed. Thus counts of green apples, red apples, big apples or ripe apples are all converted to linear weights by the same mathematical process. Only the conversion factors differ.

4. Linear Measurement : An Idealization

A single dichotomous observation of a variable, such as the presence or absence of a particular author among the list of names heading a paper, is not a linear measure of the proclivity to authorship of that author. Yet an accumulation of such dichotomous observations into a count summarizes all the information that exists about that author's proclivity. A count is often treated as, or assumed to be, a linear measure. But the mere fact that counts exhibit numerosity and can be written as linear numbers does not give them linear properties. Table 1 contrasts the properties encountered in counts with those required of linear measures. Counts often approximate linearity over a central range closely enough for researchers to

Table 1. Counts contrasted withlinear measures

Characteristic	Counts are:	Linear measures are:
Meaning :	ordinal ranks on hoped - for variables	linear positions on explicitly defined variables
Continuity :	discrete integers with gaps between numbers undefined	continuous, with all values defined
Additivity	in numerical form, but with not-linear meaning, so unsuited to satistics like means, linear regression, plots and statistical operations	numerical and equal-interval in both form and meaning, and so support the usual arithmetical
Precision :	unknown or ill-defined in precision, but often mistaken for exact ``truth''	known to be estimates of unattainable ideal measures, modelled to have well-defined standard errors
Conceptualization :	concrete descriptions of events in a limited historical context	abstract, general idealizations for inference in all similar contexts
Occurrence :	accidental, irreproducible results of a historical process, chosen for pragmatic reasons	deliberately constructed, via a reproducible measurement process, to have ideal properties
Quality control :	all equal in status as historical truths	evaluated as to the degree to which each observation supports measurement construction and inference

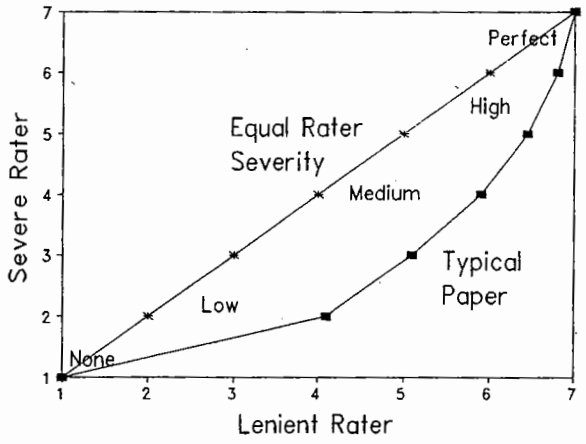
ignore their deficiencies. Nevertheless this lack of linearity has usually proved fatal to meaningful theory development.

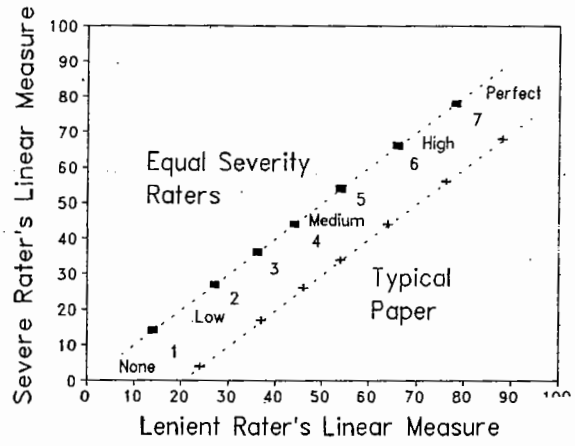
Consider journal articles rated on a number of criteria by two experts. One expert applies the criteria leniently and loosely, the other severely and strictly. The ratings for each artic e are recorded as counted levels up a rating sc.le. The nonlinearity of these counts is sh>wn in the Lorenz-like curve in Figure 1A where the equivalent numerical category counts awarded by the lenient and severe raters are plotted. These counts may approximate linearity over their central ranges, but provide only weak support for inference from the counts of the lenient rater to those of the severe rater. Counts by both raters must be linearized into measures and then equated, as shown in Figure 1B, before inferences based on the lenient rater's counts can be applied with confidence to the severe rater's counts.

Figure 1C depicts the general ogival relationship between measures and counts. The step from counts to linear measures is one that empiricists find by accident. The logistic ogive or "autocatalytic curve" is widely used to describe physical processes. "We are led to suggest a second basic law of the analysis of science : all the apparently exponential laws of growth must ultimately be logistic." (p. 30)[12] It is not some physical mechanism that connects the counts with the measures, however. It is a mathematical property that only the logistic transformation possesses.

5. Rasch Measurement Models

Rasch[14] constructed a measurement theory that contains the necessary and sufficient criteria for the conversion of counts of qualitatively similar objects into quantitatively linear measures. This theory capitalizes on the disorder present in observations due to the stochastic element that exists in all empirical counting operations. Rasch started with the premise that the linear measure of an object, e.g., an author's proclivity to publish in a special field, must be independent of the measurement agent, e.g., any particular journal, and *vice-versa*. Each object has a single linear mea-





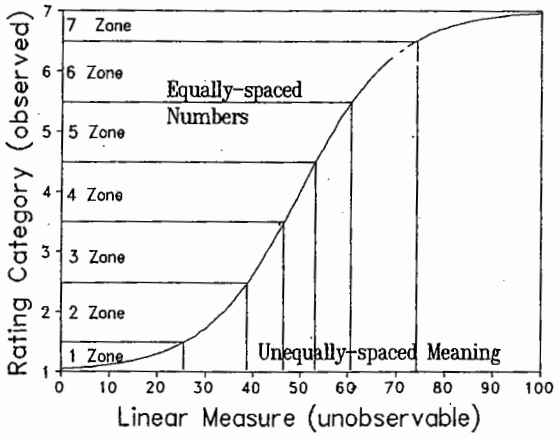


Figure 1. Non-linearity of counted data. A, non-linear relationship between observed numerical category for raters with different severities. B, linear relationship between linearized measures corresponding to rating categories of different raters. C, relationship of numerical rating category to linear measures.

sure. Each agent has a single linear measure, its calibration, in the same frame of reference. When measures and calibrations are hypothesized to interact to produce stochastic, qualitatively ordered data, a necessary and sufficient measurement model can be derived for any type of count.

The essential transformation is always the logistic, log-odds transformation.[17] The basic model for dichotomous, true/false, present/absent, 1/0 observations is

$$\log\left(\frac{p_{ni}}{1-P_{ni}}\right) = B_n - D_i \qquad (1)$$

Object n is parameterized to have linear measure B_n. Agent i is parametrized to have linear calibration D_i. These interact stochastically to produce a datum with probability P_{ni} of being observed to be 1.

When the data are Poisson counts of independent events, the comparison of interest for the interaction between object n and agent i is the probability of observing a count of j, P_{nij}, against the probability of observing count of $j-1$, P_{nij-1}.

$$\log\left(\frac{P_{nij}}{P_{nij-1}}\right) = B_n - D_i - F_i \log(j) \qquad (2)$$

F_i is the conversion factor from the logarithm of the count to the general linear scale for the particular agent of investigation, i. This formulation supports most empirical laws based on logarithms of counts.

Likert ratings are observations on a rating scale or some other qualitatively ordered series of categories. Even though given numerical lables, the observations are ordinal. Each higher rating denotes qualitatively more of what is being observed. A measurement model is required to discover how much more, quantitatively, is implied by each qualitatively higher rating. This measurement model is

$$\log\left(\frac{P_{nij}}{P_{nij-1}}\right) = B_n - D_i - Fj \qquad (3)$$

The inevitably uneven spacing of the rating categories is parameterized by F_j, the extra "difficulty" overcome in order to be observed in

Table 2. Annual average of publication counts by field (from Braun et al.[2]) with ordinal ratings. The most unexpected ratings are marked by "*".

Field:	Clinical Medicine		Biomedical Research		Biology		Chemistry		Physics		Earth & Space Science		Engineering		Mathematics		Total Rated
Country	Avge	Rated	Avge	Rated	Avge	Rated	Avge	Rated	Avge	Rated	Avge	Rated	Avge	Rated	Avge	Rated	
IND	825	5*	1181	7	836	7	1048	7	1323	7	245	6	743	7	118	7	53
AUS	2195	7	688	6	856	7	751	6	490	5	392	7	375	6	139	7	51
NLD	1092	6	709	6	345	6	674	6	659	6	174	6	225	6	104	7	49
ITA	1489	6	714	6	149	5	1163	7	798	6	208	6	267	6	81	6	48
SWE	2161	7	794	6	196	5	468	5	385	5	87	5	233	6	60	6	45
ISR	952	5	536	5	257	6	313	5	486	5	114	5	240	6	145	7	44
SWI	1536	6	583	5	89	4	594	6	758	6	87	5	321	6	78	6	44
POL	458	4	477	5	106	4	686	6	458	5	51	4	316	6	140	7	41
GDR	432	4	373	5	92	4	748	6	335	5	61	4	156	5	113	7	40
BEL	716	5	447	5	104	4	402	5	325	5	67	4	139	5	57	6	39
CZE	424	4	400	5	138	5	902	7	252	4	104	5	70	4	27	5	39
DEN	1315	6	308	4	132	5	218	4	306	5	56	4	58	3	56	6	37
AUT	704	5	149	3	58	4	205	4	182	4	26	3	92	4	53	6	33
HUN	398	4	307	4	92	4	461	5	137	4	13	2	105	4	71	6	33
NOR	726	5	188	4	97	4	190	4	109	4	53	4	51	3	30	5	33
FIN	801	5	232	4	62	4	104	3	114	4	27	3	96	4	26	5	32
NZL	462	4	144	3	366	6	127	4	59	3	119	5	57	3	19	4	32
SAF	873	5	125	3	102	4	150	4	78	3	86	5	88	4	15	4	32
SPA	257	3	249	4	109	4	406	5	145	4	30	3	41	3	18	4	30
BRA	167	3	151	3	104	4	83	3	196	4	33	3	46	3	39	5	28
YUG	105	3	124	3	21	3	210	4	147	4	11	2	56	3	16	4	26
ARG	237	3	142	3	46	4	93	3	71	3	26	3	28	3	10	3	25
EGY	137	3	67	2	70	4	181	4	47	3	10	2	92	4	4	2	24
GRE	145	3	46	2	30	3	62	3	84	3	32	3	46	3	19	4	24
ROM	62	2	63	2	7	2	195	4	218	4	6	2	38	3	32	5	24
IRE	193	3	47	2	70	4	71	3	52	3	12	2	14	2	21	4	23
BUL	69	2	62	2	14	3	129	4	77	3	10	2	33	3	7	3	22
NIG	177	3	56	2	130	5	29	2	17	2	14	2	22	2	13	4	22
MEX	173	3	60	2	33	3	28	2	86	3	18	2	21	2	9	3	20
CHL	208	3	40	2	29	3	23	2	8	1	54	4*	5	1	2	1	17
PRC	7	1	2	1	2	1	4	1	17	2	1	1	6	1	5	2	10

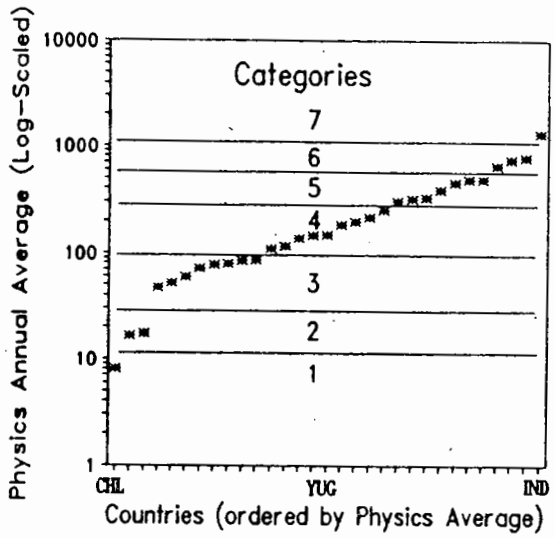


Figure 2. Assignment of categories to country publication activity for physics.

category j, relative to category j-1. The count-to-measure transformation has the appearance of a logistic ogive, but its precise mathematical form depends on the behavior of the rating scale categories. This model underlies Figure 1.

6. Measurement and Quality Control

A major difference between research in the social sciences and that in other sciences is quality control. Most social scientists attempt to include in their analysis every observed datum. They perform global fit tests in which the data is treated as sacrosanct so that it is the hypothesized theoretical model that is accepted or rejected *in toto*. Physicists, on the other hand, take many observations, rejecting those that seem dubious, and keeping only those of satisfactory quality.[3] In bibliometrics, the need to verify the quality of data is well established : "As we have repeatedly done in the past, may we respectfully caution against the serious implication ... that quantitative data [i.e., counts] can be used without considered (not rote) qualitative judgments."[4]

Since the Rasch model embodies the requirement for linear measurement, rejection of the model implies abandonment of the attempt to construct linear measures. Accordingly, Rasch measurement rejects the "global fit" approach to model validation in favor of data validation through quality control. Indeed, detailed examination of the fit of the data to the measurement model is vital. Here it is essential to apply Maier's tongue-in-cheek law : "If facts do not conform to theory, they must be disposed of."[5] For linear measures to be obtained, the counts must fit a Rasch model. For each analysis, the decisive question is not "Does the theoretical model fit the empirical data?", but "Does a relevant subset of the data fit the model usefully enough to support linear measurement?"

Equality of unit size in measurement is a theoretical ideal. This ideal is never realized in practice in any part of science, though it is closely approximated in many empirical measuring instruments by means of careful construction. The quality of a measuring instrument is often determined by how closely it can be made to approach the theoretical ideal of measurement.

Rasch models provide quality-control fit statistics at all levels of analysis. Individual data points can be assessed for their likelihood. Objects and agents, e.g., authors and journals, can be scrutinized for their adherence to the general behavior of objects and agents. Groups of objects and agents can be investigated for distinguishing characteristics.

The nature of the particular lack of fit motivates the remedial action. Idiosyncratic, non-cooperative data are isolated for diagnostic investigation. Are they data entry errors? Are they unusual circumstances? Data-to-model misfit prompts deeper questions : do the data manifest quantities of one predominant underlying varibale, or are they a heterogeneous collection of observations? Can homogeneous subsets be identified and measures for these constructed? Irremedial lack of fit signifies that those counts must be rejected as not useful for linear measurement.

7. A Methodological Example

An indication of the scientific production of different countries is the number of scientific publications each produces. The annual average publication counts by scientific field for 31 countries from 1976 to 1980 is reported by Braun et al.[2] This is shown in Table 2. From these data, it is desired to construct a linear measurement system on which to locate the publication activity of each country.

The average publication rate may be a Poisson-distributed quantity, but it is not clear with what level of precision each reported number has

been obtained. Nor is it known in what way economic or other factors are perturbing the publication rate. Accordingly, though the reported averages might seem to provide the basis for the measurement process, their precise numerical values may be misleading.

From the measurement perspective, the "averages" might better be taken as qualitative, rather than quantitative. For instance, though the average number of clinical medicine publications for Belgium, at 716, is higher than that for Austria, 704, this small relative difference can hardly be taken to indicate a higher level of activity in Belgium than Austria. In any particular year, Austria may well publish more than Belgium. Australia, however, at 2,195, is clearly publishing at a higher level. In order to construct measures, the identification of qualitatively different publication levels is more important than the precise volume of publication within those levels. A first step, therefore, is to identify qualitatively higher levels of activity and to label them ordinally in ascending sequence for each of the fields.

Initial qualitative, ordinal ratings are assigned by inspection of the distribution of counts. Figure 2 shows the categorization for physics averages. Each apparently qualitative higher level of publication activity is labelled with a higher ordinal category number. Table 2 also shows the categories assigned within each field.

These category definitions can be further refined as analysis proceeds, in order to obtain rating levels within each field that are more coherent with the hypothesis that each field is an independent indicator of one overall scientific publication activity measure for each country. For instance, after inspection of Figure 2, it could be argued that category 4 is a composite of two categories, a high one and a low one. On the other hand, it could be argued that categories 5 and 6 do not represent qualitatively different performance levels. These alternative categorizations can be analyzed with a view to discovering which categorization best supports measurement construction. Since the goal is to characterize each country with a measure of publication activity that is robust against the stochastic aspects of data, the measures should also be robust against minor recategorizations of those data. Consequently, if different, but reasonable categorizations lead to very different measures, then the robustness of the measurement system is suspect, and its value as a basis for inference is doubtful.

The Rasch measurement model specified for these ordinally-advancing qualitative categories is:

$$\log\left(\frac{P_{nij}}{P_{nij-1}}\right) = B_n - D_{ij} \qquad (4)$$

where

$\mathbf{P}_{nij}$ is the probability of observing category j when country n is rated on field i.

$\mathbf{P}_{nij-1}$ is the probability of observing category j-1 when country n is rated on field i.

$\mathbf{B}_n$ is a measure of the overall scientific publication activity in country n.

$\mathbf{D}_{ij}$ is the additional difficulty of publishing at a level of category j in field i relative to category j-1.

Application of this model to the data in Table 2, by means of the BIGSTEPS[16] program, permits the construction of the linear measurement system shown in Figure 3. Each country has been located on a line according to its overall publication activity measure. The unit of measurement is the *logit* (log-odds unit).

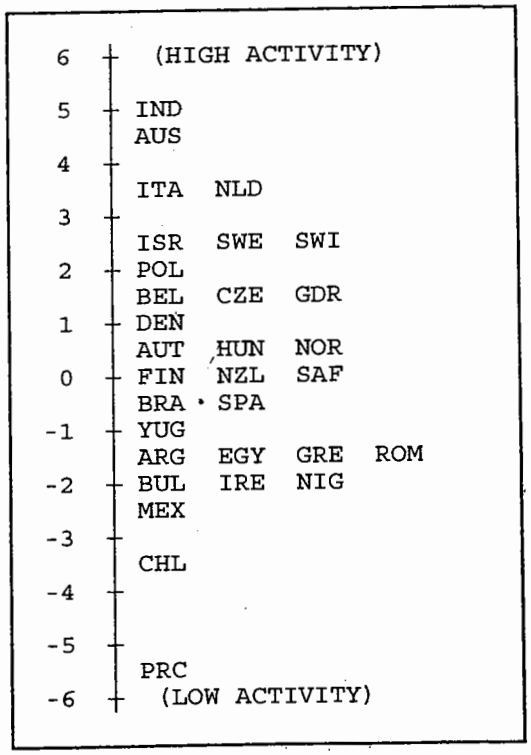


Figure 3. Overall publication activity of countries summarized as linear measures. The metric is in logits, such that a one unit difference has the same meaning everywhere.

These measures are ideal for further analysis manipulation. They have the equal-interval properties expected of numbers used in the usual statistical procedures : means, linear regression.

Within this frame of reference, instances of unexpectedly high or low publication activity can be identified for further diagnosis and analysis. The most unexpectedly high activity for any country within a field, based on its overall activity measure, is Chile's in Earth Science. Chile has an average of 54, rated 4. Chile's expected level is a rating of 2, corresponding to a publication average of about 12. The most unexpectedly low activity is for India on Clinical Medicine. It has an average of 825, rated 5. A rating of 7, with an average of about 2,200 is expected. Through these anomalous ratings, special features of countries and fields can be investigated, and those areas in which the measures are useful as a basis for inference delimited.

8. Conclusion

Rasch[14] developed and expanded his measurement methodology between 1951 and 1959. Early work included linear measurement of oral misreadings, readings speed and math ability. This lead to many applications of Rasch measurement in educational and psychological testing. Though initially applied to dichotomous multiple-choice tests, the model is now employed, in its rating scale form, for all types of counted data. A feature that has increased its utility is the robustness of the model against missing data.

The ready availability of software, e.g., BIGSTEPS[16] and *Facets*[10], has promoted the application of Rasch measurement to many fields including quality of paint finishes [15], performance of wine-tasters, measurement of evolutionary development in fossils, and medical patient evalution.[7,8] The facility with which Rasch theory constructs linear measures from ordinal counts indicates that Rasch methodology can be a useful aid in the endeavor of building a firm and level foundation from bibliometric counts which are necessarily uncertain and uneven.

References

1. Bookstein,A. Informetric Distributions, Part 1 : Unifid Overview, *Journal of the American Society for Information Science,* 41 (1990), 5, 368-375.
2. Braun, T. W. Glänzel, A. Schubert. *Scientometric Indictors.* World Scientific, Singapore, 1985, Table 162.
3. Dorsey, N.E. The Velocity of Light, *Transactions of the American Philosophical Society,* 34 (1944), 1-110.
4. Garfield, E. Citation measures used as an objective estimate of creativity, *Essays of an Information Scientist,* Vol 1, 1962-1973, ISI Press, Philadelphia, 1977, 120-121.
5. Garfield, E. Humor in scientific journals and Journals of Scientific Humor, *Essays of an Information Scientist,* Vol 2, 1974-1976, ISI Press, Philadelphia, 1977, 664-670.
6. Garfield, E. Bradford's Law and Related Statistical Patterns, *Essays of an Information Scientist,* Vol 4, 1979-1980, ISI Press, Philadelphia, 1981, 476-483.
7. Granger, C.V., B.B. Hamilton, J.M. Linacre, A.W. Heinemann, B.D. Weight. Performance profiles of the Functional Independence Measure, *American Journal of Physical Medicine and Rehabilitation,* 72 (1993), 2, 84-89.
8. Haley, S.M., L.H. Ludlow. Applicability of the Hierarchical Scales of the Tufts Assessment of Motor-Performance for School-Aged Children and Adults with Disabilities, *Physical Therapy* 72 (1992), 3, 191-202.
9. Herschel, J.F.W. *Preliminary discourse on the study of natural philosophy*, Longman, Brown, Green & Longmans, London, 1851, Art. 116.
10. Linacre, J.M. *Facets many-facet Rasch analysis computer program*, MESA Press, Chicago, 1987.
11. De Solla Price, D. *Science since Babylon*, Yale University Press, New Haven, 1961, 77-78.
12. De Solla Price, D., *Little Science, Big Science,* Columbia University Press, New York, 1963.
13. De Solla Price, D. Foreword to E. Garfield, *Essays of an Information Scientist*, Vol 3, 1977-1978, ISI Press, Philadelphia, 1980, vii.
14. Rasch, G. *Probabilistic models for some intelligence and attainment tests*, University of Chicago Press, 1980.
15. Rehfeldt, T. K. Judgement of stain resistance, *Journal of Coatings Technology,* 62 (1990), 790, 53-58.
16. Wright, B.D., J.M. Linacre. *BIGSTEPS Rasch analysis computer program*, MESA Press, Chicago, 1991.
17. Wright, B.D., G. N. Masters. *Rating Scale Analysis*, MESA Press, Chicago, 1982.

Bibliographic Knowledge And Its Application in Retrieval Processes

H. Peter Ohly
Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany

Bibliographic databases are biased or moduled by scientific network patterns that are the producing sources. Modeling this latent scientific community should be possible by incorporating its manifold ties in databases on scientific outcome. Thus, we not only have a means of describing this community, but also of improving the retrieval in these information banks. As it includes ideas of social network analysis, we can speak of a "socially enriched" retrieval as an alternative or additional means of statistical or linguistic analysis of bibliographic documents.
The Social Science Information Center in Bonn (IZ) is developing the knowledge-based system AKCESS (Assistance by Knowledge-based Context Evaluation in Social Science Retrieval), which reconstructs links between first hand information to related documents. The total aggregation of the information scattered over different documents allows a new judgement of the "relevant" information and inclusion of otherwise lost information. Such a search is by definition not restricted to documents, but relevant for all kinds of information imbedded in the data set (persons, institutions, scientific communities, concepts, etc.).

1. Introduction : Retrieval in Bibliographic Databases and its Problems

Information retrieval systems (IRS) were developed to store large sets of data (rsp. documents) and to make them re-available for different problems. It is of special importance to these systems that the data contained in the documents reflect the original documented material in the most complete and authentic manner possible. Even a single retrieved document should be interpretable with respect to potential intentions of the user. In order to select adequate documents from the database for a wide range of queries as efficiently as possible, information retrieval is supported as a rule by the representation of substantial questions through document strings such as keyword searching, Boolean expressions, and distance operators. A good retrieval result expresses itself in the greatest possible number of appropriate documents (see Ohly 1991)[16].

From the information seeker's point of view this is not very satisfying. Research field novices will certainly be overloaded with very specific or peripheral articles, whereas domain experts are not willing to browse through a large number of retrieval results containing only common knowledge. Redundancy by the same author or by a conform scientific community is another issue that is not tolerable for an information seeker simply looking for the best answer to a precise question. On the other hand, value of individual works is only found when diachronous or synchronous comparison with other outcomes of the same research group is taken into consideration. Additionally, intimate knowledge terminology of the subject and its terminological control within the documentation process cannot be presupposed by the user (for solutions see: O'Neill and Morris 1988 [17]; Croft and Thompson 1987 [3]; Fox 1987 [5]).

Databases represent nothing more than a more or less arbitrary selection of all possible publications. The result is an even more serious distortion of the ideal answer. The definition of the scope might exclude relevant publications of non-related publishing houses. Often, only certain core journals are included. Large specialized databases are often only a combination of small document collections of data manufacturers of very restricted interests implying peaks and slopes for the *representativity* of the subject field. Often publi-

cations are only adequately documented regarding quality and occurence after more than one year. Therefore, the more up-to-date the information is, the more poorly it can be found by straight forward access. This holds especially if the subject of interest is not yet established within a certain field.

Respectively, the extent to which a database favours certain subjects and disfavours others within an information search must be considered. Further, every available knowledge on the overall position of the author as well as further background should be taken into consideration together with the content information. Cross references to other authors are as important as the development of an author in the past in making inferences on the involvement of this author in the actual discussion.

If such an additional knowledge is processed together with an information search, the output of relevant groups or schools of research can be retrieved instead of the isolated works of individual scientists. In addition, false interpretation, e.g., by homosemi, should be avoidable. Thus, each retrieval process ideally requires a specific bibliometric analysis that revises the retrieval strategy and leads in particular to a re-evaluation of the results (Herfurth and Ohly 1989)[9]. The following outlines an approach which analyzes network information on persons concerned with research activities (e.g. publications) and which is partially incorporated into an analysis system for retrieval results.

2. Multiple Networks of Scientific Communities in Bibliographic Databases

According to professional standards, we find the following information that are needed to identify the object of reference in bibliographic data banks :

author-s
title - subtitle
higher bibliographic unit
(journal, compilation)
publication place
publishing house
publication year
pages

We often find additional descriptive categories in the field of documentation :

address / affiliation
thematic keywords (e.g. from the title)
conceptual descriptors
(according to a thesaurus)
short content abstract, etc.

These descriptive fields can be expanded if non-literary document types such as projects, maps, patents, and so on, are stored. Here it is important that these fields contain enough indications of mutuality among documents under formal but also substantial consideration. These must be suitable for the reconstruction of real or cognitive patterns of similarity.

The persons mentioned in the documents (authors, editors, head, staff) enable an interlink among all those projects in which they participated (see Fig. 1). Undirected, and therefore equivalent for both participants, are links concerning co-authors, co-editors, etc., indicating a functional and personal relationship. This implies that these persons know each other and are concerned with the same object, causing mental proximity or completion.

Links over diverse levels (author-editor, staff-head, etc.) are valued differently and have to be given a direction. This kind of relationship also expresses organizational and personal proximity, but has to be valued more ambiguously with respect to the conceptual dimension. Only the detailed examination of the actual content profile allows the assumption of conformity and transfer It does express social distance in the sense that the supervising person obviously plays a more administrtive, general role. It is also advisable to examine the relations according to differnt criteria, such as stabilization through further relationships, absolute positional indicators (academic degree), or involvement in many projects or diverse subjects.

In addition to the personal characteristics mentioned above which only describe functional equivalent positions in the scientific community ("is head", "is editor"), there are also vague person links. They describe a certain social distance together with the indication of a personal proximity, but they cannot be seen as an equivalent to a concept transfer, examples being links between compilers of collective editions and authors, intro-

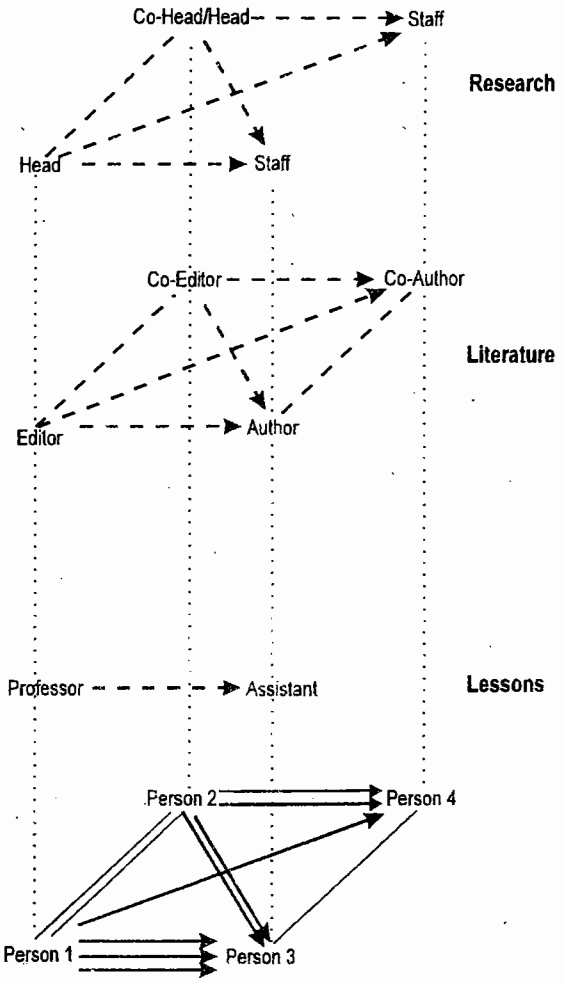


Figure 1. Multiple network patterns in bibliographic databases

duction authors and other contributors, institution heads and staff, sponsors and customers, etc. Often the connection may be established only via other persons, the editor or the head of the (same) project.

In contrast to personal relations, the institutional affiliations are not so much a situaton of face-as-face contact as a kind of action frame for the persons concerned. At least for research descriptions, but also often for literature information, institutions can be determined which supervise the cited work. Thus, a network is constituted to all other persons that belong to that institution and to the projects performed by these persons. Strictly speaking, this means that all projects are directed under the supervision of this institution, but in a broader sense, all research projects carried out by these persons have to be included as a kind of personal resource context, mediated by this institution.

Since institutions for their part are again affiliated with other institutions (co-institutions), higher level networks are formed, whereas multiple affiliations of one person can generate networks indicating a kind of "brain drain". But explicit binds are given by projects that are performed by different institutions in cooperation. Unsymmetrical relations can be demonstrated concerning the role of institutions (e.g. customer-compiler) if such information is embedded in the descriptions of the documents.

In the same sense that commissioner and sponsor express more about general resources than of concept proximity, publishing houses and journals are a non-binding infrastructural, but specialized frame that should be taken into consideration for ex-ante hypothesis or ex-post confirmation. Who publishes under a certain publishing house or in a special journal having a distinct profile, not only renders a corresponding evaluation, but also provides for an appropriate distribution in the professional world.

Besides personal relations, ideal relationships (mental climate) are important for the location of themes and persons. If the same concept is used in the same semantic, geographic and time space, looking for additional ties is of particular value (cf. de Haan 1993)[7]. Specific terms are of importance in creating a common idea, as homonymous, out-of-field usage is no longer possible. The methodological and theoretical scientific background of one particular work or scientist is of the same importance as a contentual and conceptual relationship. Similarly, scientific background, methodological paradigms and research design are an indicator of a certain school tradition, showing an equal knowledge standard implying communication and transfer or even personal acquaintance. Regional and historic proximity are not sufficient for the examination of diffusion processes. A common origin as well as divergent processes can be revealed by backward or forward chains over time and space. This, however, does not exclude the possibility of an eventual convergence.

3. The Modeling of Scientific Networks in an Expert System

The distinguishing characteristic of "expert" systems lies in storing knowledge for the solving of a certain kind of task and in taking data of the problem field into account during the electronic data processing, corresponding to the more general term "knowledge-based" systems (e.g. Wachsmuth/Meyer-Fujara 1991) amd [26]. Along with the general inferential solving strategy, the processing of the data then depends on the quality and usefulness of the stored knowledge (cp. Gebhardt 1985). Since the working procedure or the application field of such systems are not mechanically fixed, feedback with sufficiently good specialists must be included in their usage. As a result they are often designated as "assistant" systems

Standard models in knowledge processing are representations as (production) rules or as frames (see Puppe 1990). In the case of rules, the implicit conclusions can be derived by given assumptions (forward chaining), or the necessary causes of the desired aims can be proved (backward chaining). In particular, virtual intermediate goals can be created which an be summed up to an abstract, qualitative statement. The second kind of modeling, frame representation, assumes more static and coherent classification problems. Nevertheless, rules and frames are mutually transferable. This representation gains its potential by the fact that each problem part can point to other objects, making the construction of new objects originating out of concrete data (instances) possible. Highly economic, but also relatively inflexible, is the option to inherit characteristics (is-a relation), by which the denomination of the most general case is also valid for more specific generic subclasses. Other possibilities of modeling are semantic networks, which represent knowledge according to its multiple relations (see Reimer 1991)[19]. The propagation of these networks on the shortest path then given the most efficient link between the given information that is to be connected. This and similar techniques such as the constraint approach (the search problem is reduced to the only possible solutions by its restricting parameters), case-based reasoning (assigning new cases to known prototypes), or neural nets (diagnosed input and intended output are traced back to one another by optimal calculations) will not be discussed here in further detail.

Modeling is relatively independent (in contrast to the application of so-called expert system shells) if one uses a declarative program language such as LISP or PROLOG. The latter is based on first order predicate logic. When it is coded in so-called clauses (laws or acts), knowledge can be sufficiently analyzed to its implicit conclusions by logic simplification, which judges "not existent" as "not true" in the database (closed world assumption), and which accepts only unequivocal sentences. This proof propagation (unification) is part of the PROLOG software. Here, the order of processing is of special importance, since the first completely proved solution is no longer scrutinized. PROLOG offers very many different possibilities of changing the proof process. The proof chain (trace) is also protocoled in a manner that allows the result to be analyzed with regard to the solving path (explanation). Although this language is very similar to the production rule model, frame-like (or more generally object-oriented) structures or semantic nets, etc., can also be modeled (see Sterling and Shapiro 1986)[25].

The so-called knowledge base, which is used to direct the fact base, contains generally true rules and generally true facts. If subject knowledge which is extracted from literature or empirical analyses is incorporated, statements on the aggregation, propagation and evaluation of networks will be part of the rule base. Facts, such as the value of journals, information on the real conjunctions between institutions, etc., are a permanent resistant part of the knowledge base and can be provided for before processing or be computed by analyzing the processed data. In the same way, rules can be inducted by empirical correlations during the processing.

Knowledge on networks can be implemented in the following sequential steps :

Generating absolute and relational variables suitable to characterizing an object (e.g. person) in its social relations,

Constructing the ties between these objects by different networks up to the desired depth, e.g. by information on the nodes and links passed,

Processing the connections found with respect to the specific query.
Repeated iterative generation of relational characteristics of the nth degree, depending on the desired reflexivity.

Since logical inference also allows for recursion, it is not difficult to accept all degrees of relations, such as master / pupil or head / staff relations, that reproduce themselves. If not only a mere segmentation into separated groups out of the set of objects is of interest, the recursion on egalitarian and transfer-specific relations should be confined, or the additional distances by the inclusion of periferal nodes should be computed by a non-linear algorithm. Possibly a confinement to triplets (triad census) should be considered (cp. Hummel and Sodeur 1991) [11].

Whereas it can be relatively easy to determine the final points of networks (objects with only one relation or with relations in only one direction) by logical analysis, the denomination of centroids and similar cluster-analytical centres (cf. Scott 1991) [23] is more adequate for statistical processing systems or else it requires more complicated programming, e.g. with secondary files.

4. Retrieval Improvement by the AKCESS System

The AKCESS system, developed at the Informationszentrum Sozialwissenschaften Bonn (IZ), aims at the retrieval of dominant persons with respect to given topics. FORIS (social science research projects), SOLIS (social science literature), LEHRE (sociological university lessons) are used as databases. The system prototype is programmed in PROLOG and therefore contains the necessary knowledge for processing in a logical clause manner (Herfurth, Mutschke and Ohly 1992) [9].

The "relevant" persons are found after different steps. In a first approach, all documents with some concept relation to the query are retrieved by a rough selection. All important information of these documents is stored in PROLOG as facts and then is logically analyzed (provisionally) in offline mode. The user specifies his query by determinating and weighting the requested combination of concepts and selecting the appropriate terms offered by the system to describe these concepts. On the basis of the quantity and completeness of the concepts per documents and per person, a set of scientists is found which more or less corresponds to the query subject. Also the subject can be refined by definitions of historic periods that are processed under consideration of overlapping and included time intervals. In the future, the scope and current importance of the activities of the candidates will be also considered.

The most significant difference to a problem solution by conventional boolean retrieval is, that not a set of documents (for totally open usage by the inquirer) is provided for, but that the most relevant persons (in decreasing rank order) are listed. Interpretations, comparisons and combinations of the original information are not only left to the user, but are made available automatically by the system or assisted as much as possible..

The task of network options in this system is to identify wrong or unusual specified concepts in the documents and to relocate them. Thus, a certain person and a specific concept are well associated if this personal concept system remains more or less stable within different person and institution contexts, and the concept under consideration does not change dramatically by its different term contexts. Personal and conceptual clusters can reinforce each other if the co-persons represent the same concepts and the co-concepts are significant for the same persons (see Figure. 2). Instead of shrinking the research result, these clusters can also be used to broaden the fund of information if appropriate persons could not be found directly (see Figure 3). On these grounds a vast publication of a group of authors that are generally known for a certain topic can be stated to be of the same content. If and how profit from these context clusters should be taken must be decided by the user.

As a first approach, the social relations are analyzed with respect to the bibliographic context of the person under consideration. Thus, a publication in a compiled edition is amplified, if this compilation itself is a core representation of the field. Furthermore, reinforcement is given if the editor of this compilation is a relevant expert regarding his personal profile. Then, if an author is simultaneously an editor in his domain, he will also be more important if he writes a book instead

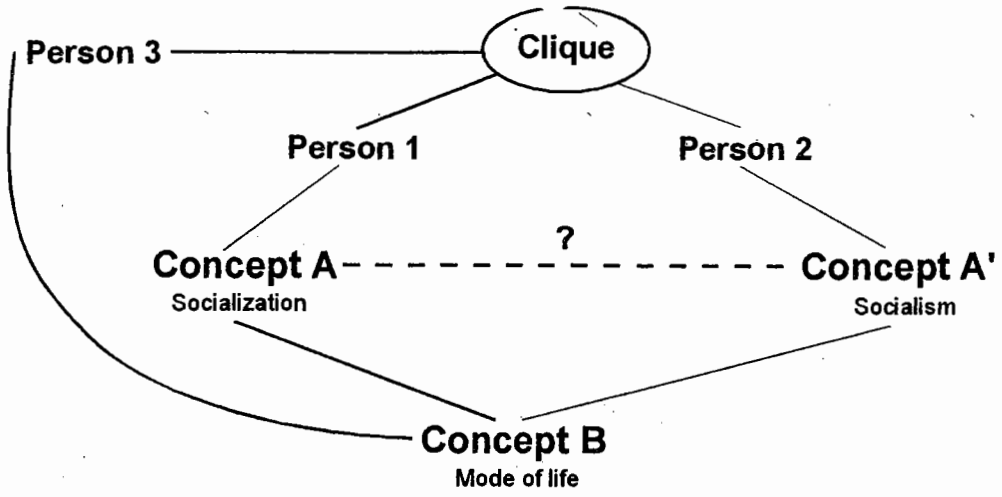


Figure 2. Evaluation of concepts by person context

of an journal article. Whatever the default values of the system will be for these cases can be changed by the user.

A network module has been developed separately as a prototype which identifies all non-overlapping networks and determines the individual position of one member in relation to all other members (Mutschke 1994)[15]. At the moment, equivalent asymmetric network generators are the editor function of a compilation and the management function (head) of a research project, each in relation to other involved staff. Further co-functions are included as symmetric cross relations (with respect to editors, authors, staff). An appropriate combination of the different networks still has to be found especially if contradictions between the networks exist. Respectively, a simple additive model will be offered as supplementation to the user for individual weighting (e.g. co-relations are stronger than role-different relations; head relations are stronger than editor relations).

In a first attempt, the global network component provides for those persons that are in the highest or most central position in each network, depending on information like "is a head", "is a publisher", and "is a co-author"). In addition the number of connected persons is computed. Thus, this analysis is suitable for the determination of segments of the tied persons within the set of persons who are relevant to a certain subject (e.g. more egalitarian groups vs. more hierarchically organized groups). Still unsolved are questions concerning the depth of distance and internal contradiction, the optimal determination of the "star" position, and the combination of hierarchical position with thematic competence.

The local network analysis computes the distance and the qualified chain of nodes of all person-pairs in the net of persons under consideration. Here the manager function (head) and the editor function are counted as one distance unit, whereas co-functions are not seen as distance but as equivalent branching. Thus, densities and structures around one person can be analyzed as an alternative to or completion of the thematic retrieval process (see as demonstration : Mutschke 1993). How competing paths between two persons should be treated dynamically, as well as the qualitative value of distances and the kind of nodes in combination with the thematic relevance of the person nodes passed, are questions still to be considered.

5. Possibilities for Theory Building

Whereas the simulation of dynamic macroprocess or microprocess has found its place in the social sciences for the generation of hypothesis and the verification of basic epistemologies such as exchange-theoretical behaviour, the application of knowledge-based approaches, except for

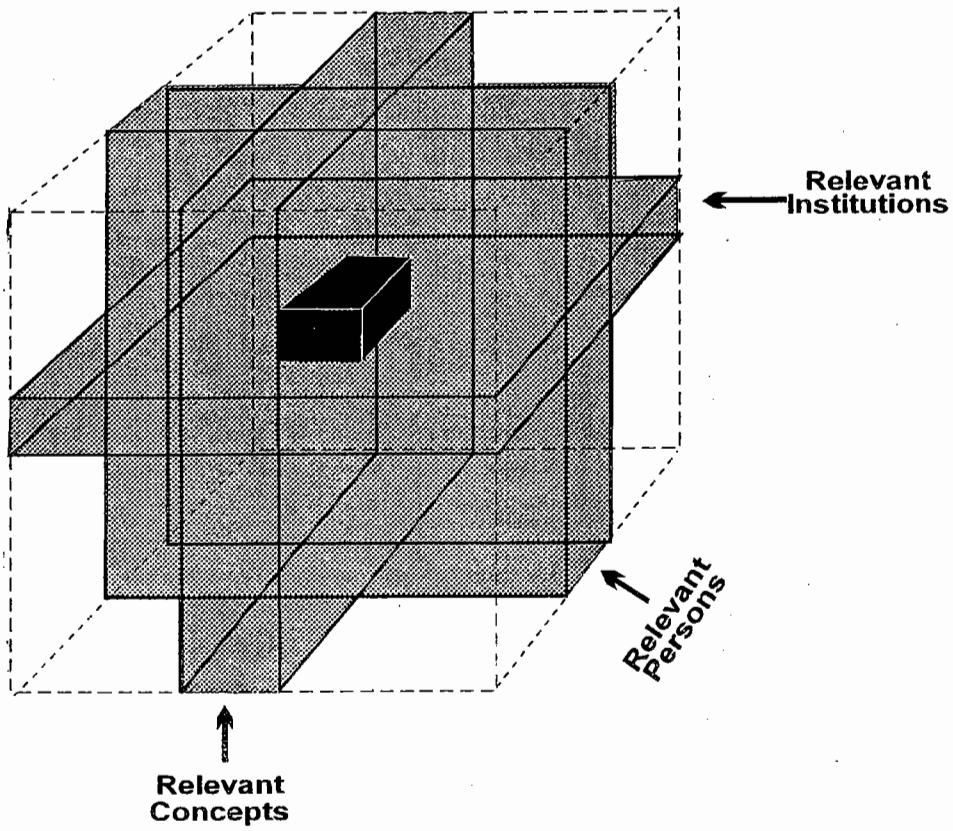


Figure 3. Broader and narrower definitions of a concept-person-institution set.

planning decisions, is very poorly developed (cf. Schnell 1992)[21]. The induction of laws from empirical findings is seriously in discussion in the natural sciences (Slezak 1989; Branningan 1989). It is evident that this cannot be expected for the social sciences, when even the rationalistic approach of the last century has not succeeded in finding social laws independent of space and time.

It is a different situation if modeling, whether with simulation or knowledge engineering, is considered a possibility of encoding theoretical, verbal models forming a non-contradictory representation and an intersubjective reconstruction (e.g. Schrodt 1988[23]; see : Manhart 1991[13]). If there is corresponding acceptance, this results in the side-effect that these models with real or simulated data can be analyzed for possible consequences, and eventually even the integration of different models should be possible (Klüver 1991)[12]. The advantage of knowledge engineering that applies declarative languages consists in the possibility of isolating the theoretical potential from the rest of the program and in the easy processing of non-numeric values (Brent 1986[2]; Sallach 1989[20]; Haas 1990[8]). That non-numeric values can be easily processed is an important argument against those who prefer the subjective interpretation of individual social scientists rather than automated algorithms as an answer to social questions, even if computing rules for vague formulations do not go far beyond heuristic approaches.

Just as a knowledge-based support in ex-ante hypothesis formulation and ex-post integration of empirical theory fragments can improve the strategy of a researcher (cf. Faulbaum 1986)[4], the inclusion of theoretical cognition into practical operating decision systems should be a foundation of and therefore produce an optimization of the results. In the case of systems like AKCESS, which processes assumptions on the importance of scientific communities, scientifically based decision support is gained by adequate inclusion of theoretical findings from bibliometric studies and

science research studies. Here the used presumptions, lacking available studies or avoiding analogy problems, can by all means be extracted by empirical analysis of the databases used for AKCESS. This can be justified since AKCESS uses only small thematic partitions and since the database is continuously supplemented. Finally, it appears reasonable that only that information is acquired that leads to an improvement of the search answer (to be judged by an real expert), particularly since the interplay of inferences from knowledge bases (e.g. in product rule style) is not predictable and also cannot be prognosed by multivariate analysis techniques (e.g. discriminant analysis). In logical analysis, every evidence which the formerly vague decision constraints counts cumulatively; that is, it is usually not statistically relativated. Ideally, a combination of statistical and logical procedure will be preferred, the one more for structural, relational effects, the other more for individual, absolute characteristics. Moreover, a theory of logical contradictions in social relations and group structures has to be developed in the field of network analysis. In contrast to statistical procedures, they should not result in assimilation of position differences.

On the other hand, the conclusion can be drawn that theoretical knowledge which contributes to an improved solution in a practical problem has a certain practical relevance and therefore gains a kind of validation, regardless of the fact that this knowledge must be integrated without contradictions into the model. If it becomes evident that relations in the field of scientific production result in a concept consonance, that relevant persons can be found more easily, then the related consonance theories are validated without the necessity of specifying statistical causal relations.

References

1. Brannigan, Augustine. (1989). Artificial intelligence and the attributional model of scientific discovery. *Social Studies of Science*, 19, 601 - 613.
2. Brent, Edward E. (1986). Knowledge-based systems : A qualitative formalism. *Qualitative Sociology*, 9, 256-282.
3. Croft, W., Thompson, R. (1987). I3R : A new approach to the design of document retrieval systems. *Journal of the ASIS*, 38, 389-404.
4. Faulbaum, Frank. (1986). *Very soft modelling* (ZUMA-Arbeitsbericht 86-04). Mannheim : ZUMA e. V.
5. Fox, E. (1987). Development of the CODER system : A testbed for artificial intelligence methods in information retrieval systems. *Information Processing and Management*, 23, 341-366.
6. Gebhardt, F. (1985). Querverbindungen zwischen Information-Retrieval-und Expertensystemen. *Nachrichten für Dokumentation*, 36, 255-263.
7. De Haan, J. (1993). Group formation in Dutch sociology. Utrecht University. (Paper presented at the *3rd European Conference on Social Network Analysis*, June 1993, Munic).
8. Haas, J. (1990). Treatment of uncertainty in social science expert systems. Czap, H; Nedobity, W. : TKE'90: Terminology and knowledge engeneering 1. Frankfürt a.M. : INDEKS, 62-76.
9. Herfurth, M.; Ohly, H. (1990). Von bibliographischen Datenbanken zu Wissensbanken. Deutsche Gesellschaft für Dokumentation : Deutscher Dokumentartag 1989. Informationsmethoden : Neue Ansätze und Techniken. 4-6. Oktober, Bremen (DOK-2). Frankfurt / M. : DGD, 405-418.
10. Herfurth, M.; Mutschke, P.; Ohly, H. (1992). Inference from bibliographic facts : A social network approach between front-ends and text comprehension. Zimmermann et al. : Mensch und Maschine : Informationelle Schnittstellen der Kommunikation. Proceedings des ISI '92, Nov. 1992, Saarbrücken. Konstanz: Universitätsverlag, 200-207.
11. Hummel, H.; Sodeur, W. (1991). Modelle des Wandels sozialer Beziehungen in triadischen Umgebungen. Esser, H.; Troitsch, K. G. : Modellierung sozialer Prozesse (Sozialwissenschaftliche Tagungsberichte 2). Bonn : IZ Sozialwissenschaften, 695-733.
12. Klüver, J. (1991). Formale Rekonstruktion und vergleichennde Rahmung soziologischer Theorien. *Zeitschrift fur Soziologie* 20, 209-222.
13. Manhart, Klaus. (1991). KI-Modellierung in den Sozialwissenschaften. *KI Künstliche Intelligenz*, 5 (2), 32-40.
14. Mutschke, P. (1993). AKCESS-Dossier "Ausländerfeindlichkeit". IZ Sozialwissenschaften : Im Focus : Ausländerfeindlichkeit. Ein Überblick über Forschung und Literatur aus sozialwissenschaftlicher Perspektive. IZ : Bonn.
15. Mutschke, P. (1994). Processing scientific networks in bibliographic databases. Bock, H.; Lenski, W.; Richter, M. : Information systems and data analysis. Studies in classification, data analysis, and knowledge organization 4. Heidelberg : Springer (in print).

16. Ohly, H. Peter. (1991). Conceptual information retrieval by knowledge-based programming techniques. International Classification, 3. Frankfurt/M. : INDEKS, 148-152.
17. O'Neill, M.; Morris, A. (1988). Database and expert systems - the way forward. Online Information 88. *12th international online information meeting. Proceedings*. Oxford 1988, 279-290.
18. Puppe, F. (1988). Einführung in Expertensysteme. Berlin
19. Reimer, U. (1991). Einführung in die Wissensrepräsentation : Netzartige und Stukturierte Repräsentationsformate. Stuttgart.
20. Sallach, David L. (1989) . Toward an expert support system for social science. In MacCrank, L., *Databases in the humanities and social sciences 4. Proceedings of the ICDBHSS '87*, Montgomery (571-577). Medford, NJ : Learned information.
21. Schnell, Rainer (1992). Artificial intelligence, computer simulation and theory construction in the social science. FAUI BAUM, Frank : Softstat ' 91. *Advances in statistical software 3. (Proceedings 6th Conference on the Scientific Use of Statistical Software)*. Stuttgart: Gustav Fischer, 335-342.
22. Schrodt, P. (1988). PWORLD : A precedent-based global simulation. *Social Science Computer Review*, 7, 27-42.
23. Scott, J. (1991). Social network analysis : A handbook. London : Sage.
24. Slezak, P. (1989). Scientific discovery by computer as empirical refutation of the strong programme. *Social Studies of Science*, 19, 563-600.
25. Sterling, L.; Shapiro, E. (1986). *The art of PROLOG. Advanced programming techniques*. Cambridge : MIT Press.
26. Wachsmuth, I.; Meyer-Fujara, J. (1991). Wissensmodellierung und Expertensysteme. *Kursunterlage zur KIFS-91 der Gesellschaft für Informatik FB 1*, Günne, März 1991.

FILS: A NEW JOURNAL IN INFORMATION & LIBRARY SCIENCE.

Frontiers of Information and Library Science (FILS) is a refereed professional journal in English to be issued in the months of June and Deccember every year.

Its scope includes well articulated experience of practice in information centres and libraries around the world; the results of empirical research in any aspect of information and library science; and personal experience in research and development activities especially in S&T presented in a form directly consumable by non-specialists. Others are guest editorials featuring a leading information and library professional; "man of science and technology" featuring interview with a leading scientist or technologist; short communications of relevant completed and ongoing research and development activities; forthcoming academic and professional events; book reviews; general and personal news; information about education and training opportunities; new products and services, etc.

Call for Articles

Practitioners in information centres and libraries; researchers in any aspect of information and library science and experienced scientists and technologists are invited to submit original articles in English language for consideration for publication in **FILS.** Three copies of each article should be submitted; typed double spaced throughout on one side of A4-size paper. Such articles should not exceed 20 pages and should include an informative abstract of not more than 150 words. It is the responsibility of authors to ensure that their submissions are not published elsewhere in substantially similar form. A modest honorarium in cash will be paid to the author of a lead article in each issue of the journal.

Bibliographic Style : All references should be keyed into the text with superior numbers, and listed numerically with full and accurate citations at the end of the article. Styles of citations are exemplified below :

a. Monographs:

(i) Martin, Murray S. *Academic library budgets*. Freenwich, Connecticut : JAI Press, 1993. p. 93. *(Foundations in library and information science, vol. 28).*

(b) These Dissertations:

(ii) Nzotta, Briggs C. *The career and mobility of librarians in Nigeria.* Ph.D thesis, Loughborough University of Technology, 1981. p. 120.

(c) Journal Articles:

(iii) Obaka, Daniel N. The challenge of medical librarianship in Nigeria. *International Library Review,* 17(1), 1985, p. 52.

(d) Papers in conference proceedings:

(iv) Adeyemi, Nat M. New technology and developing countries : the Nigerian experience. In : Brown, K.A. ed. *The challenge of information technology.* Proceedings of the Forty-First FID Congress held in Hong Kong, 13-16 September 1982. Amsterdam : North-Holland, 1983. p. 240.

e. Unpublished papers:

(v) Greaves, Monica A. *The BLS programme at Ibadan : an analysis of the present situation and future prospects.* Paper presented at staff seminar, Department of Library Studies, University of Ibadan, April 1985. 14pp. (Unpublished).

Correspondence :

Dr. Ken M. C. Nweke, Editor-in-Chief, FILS
University of Nigeria, Department of Library Science,
P.O.Box 3169, Nsukka, Nigeria. Fax : +234-042-770-644.

An Analysis of a Bivariate Distribution of Circulation Data An Informetric Approach

I. K. Ravichandra Rao
Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile, Mysore Road
Bangalore 560 059, India

Circulation records are analyzed to study the distribution of transaction. It has been observed that the negative binomial hardly fits the circulation data in a special library. Analysis of the number of times a book was used and its idle time shows that no bivariate probability distribution can be used to explain the data. However, bivariate negative binomial can be used as an approximate model to explain the data. Distributions of X, the number of time the book is used, given that the last date of circulation was Y years ago (i. e., f(X|Y)) follows a negative binomial distribution. The observed distribution the result obtained by this model ($f(x|y) = a+b/x+c/x^2$). The results obtained by this model are much better than the generalized bibliometric model i. e., $f(x|y) = ax^b$.

1. Introduction

Trueswell in 1964 argued that "statistic of last circulation date has potential in the preparation of a decision rule for a quantitative method of thinning the stacks" **(9)** Since then, many have worked in the area of circulation data analysis. For instance Morse [3] in 1968 has shown that the distribution of number of times a book is issued follows a Poisson Distribution. Ravichandra Rao [4, 5, 6] and Burrell [1] have shown for many sets of data that the distribution of transactions follows a negative binomial distribution. Sichel [7] suggested a Generalized-Inverse Gaussian-Poisson (GIGP) distribution to explain empirical distribution of transactions. Recently, Leemans et al [2] based on data collected from Flemish public libraries, showed how the negative binomial distribution can be used as a trend distribution for library circulation data. Ravichandra Rao [4, 5, 6] has also argued that the phenomenon of transaction of documents is a manifestation of the success-breeds-success phenomenon:

> "It is only those documents which are circulated that are likely to be circulated again and again" . .... (A)

However, in practice, in addition, the following is also equally important., as observed by Trueswell [9]:

> "time gap between the date of last circulation and the current date".

In this paper., this time gap is referred to as the idle time of a book. Thus, the idle time of a book refers to the "period" in which the book was not issued out. Normally, for those books which have fairly a good demand or which are issued out frequently, it is expected that the idle time of book is very low in addition to its frequent usage. For example, let us consider two different cases of book usage: a book might have been used 20 times, 20 years ago (the idle time is 20 years). Another book might have been used only once, a week ago (the idle time is only one week). It is quite likely that chances of the latter book to be in circulation in the near future are much higher than the chances of the former, the reason being that the last circulation was 20 years ago. In this context, the statement (A) mentioned earlier can be rewritten as:

> "It is only those documents which are circulated recently that are likely to be circulatd again and again" .....(B)

While the statement (A) can be explained by a negative binomial model, there is no such mathematical model to explain the statement (B).

2. Objectives

Although it has been argued that the negative binomial distribution fits fairly well for many data sets of circulation data, there are several examples where the negative binomial hardly fits the observed circulation distribution. This may be partly due to the fact that the data come from a heterogeneous population and from a population where the distribution is highly positively skewed with a long tail. For example, for the data given in column (6) of Table 1, the negative binomial hardly fits. In this paper, it is observed that if the data come from a population, where statement B is more relevant (in the context of a special library) than statement A. negative binomial hardly fits

It is in this context that an attempt has been made to study the conditional distribution of transactions, f(X= x|Y = y),where

Table 1. Distribution of transactions in specialist library

No. of times borrowed	Number of books in Lib. Sc.& related topic	Math. Science	Economics & related topic	Other subjects	Entire collection
(1)	(2)	(3)	(4)	(5)	(6)
0	1373(1359)	1637(1732)	945(949)	1340(1337)	5295(5574)
1	272(291)	641(559)	202(194)	204(226)	1419(1181)
2	168(160)	332(306)	103(103)	125(96)	729(662)
3	97(105)	193(190)	60(66)	48(49)	398(393)
4	86(75)	114(125)	55(46)	19(27)	274(269)
5	56(57)	89(85)	35(33)	15(15)	195(192)
6	44(44)	63(59)	22(25)	5(9)	134(142)
7	32(34)	29(42)	12(19)	5(6)	78(106)
8	21(27)	20(30)	16(15)	5(3)	64(81)
9	24(22)	21(21)	8(11)	1(2)	54(63)
10	14(18)	16(15)	5(9)	2(1)	37(49)
11	12(15)	9(11)	6(7)	2(1)	29(38)
12	15(12)	6(8)	7(6)	2(1)	30(30)
13	8(10)	4(6)	5(5)	-	17(24)
14	9(9)	9(4)	4(4)	-	22(19)
15	11(7)	2(3)	3(3)	-	16(15)
16	5(6)	4(2)	4((2)	-	13(12)
17	2(5)	1(2)	0(2)	-	3(10)
18	9(4)	0(1)	3(1)	-	12(8)
19	5(4)	2(1)	2(1)	-	9(6)
20	4(3)	2(0)	2(1)	-	8(5)
21	1(3)	-	2(1)	-	3(4)
22	3(2)	-	2(1)	-	5(3)
23	3(2)	-	1(0)	-	4(3)
24	4(2)	-	-	-	4(2)
25	1(1)	-	1(0)	-	2(2)
26	-	3	1(0)	-	4(2)
27	3(1)	-	-	-	3(1)
28	2(1)	-	-	-	2(1)
30	1(1)	-	-	-	1(1)
51	1(0)	-	-	-	1(0)
Total	2286	3197	1508	1873	8864
$\bar{X}$	1.8320	1.4048	1.4171	0.5494	1.3364
σ_X	3.9578	2.4648	3.1324	1.2455	2.9026

Note: The data, mentioned withing the parentheses in Col (2) to Col (6) refer to the estimated frequencies under the assumption that data follow a negative binomial distribution. The parameters of the distribution are computed using the moment's method. The maximum likelihood method didn't yield better result than the moment's method.

X: the stochastic variable denoting the number of times a book is issued out.

Y: the stochastic variable denoting the difference between the date of last circulation and the current date (in years); this difference will be referred to as the idle time of the book

In other words, the main objective of the study is to re-examine the distributions of transactions, in particular, taking idle time into consideration.

3. Data Collection

The following data for each of the books were collected from a special library:

i) due date for the last circulation (d_1)

ii) number of times a book was circulated.

The due date slips, in each of the books observed, show only the due dates, The procedure to compute the idle time (in years) is that :

a) the difference between the current date and due date is computed; let it be d days

b) 30 days are subtracted from d, since in the library where this study was conducted, the books are issued out for a maximum period of 30 days. The difference between d and 30 gives the idle time (in days) of the book; i.e., the difference between the last date of transaction and current date.

To compute the idle time, in terms of months.

Table 2. A bivariate distribution of circulation data
(x: No. of times a book was issued; y: Idle time of the book)

x \ y	1	2	3	4	5	6	7	8	9	10	11	12	
	17	20	Total										
1	241	194	207	138	78	52	116	93	67	99	89	43	1	1	1419
2	218	131	95	52	39	28	33	37	34	29	30	3	-	-	729
3	149	78	62	33	19	12	14	9	8	7	3	4	-	-	398
4	113	58	24	14	14	14	10	13	4	5	2	3	-	-	274
5	96	39	18	7	4	4	8	6	6	4	3	-	-	-	195
6	74	24	10	6	4	5	1	5	2	1	1	1	-	-	134
7	48	13	5	6	2	2	1	1	-	-	-	-	-	-	78
8	35	9	7	6	1	1	1	2	1	-	-	1	-	-	64
9	31	17	3	1	-	1	-	1							54
10	30	3	1	-	1	-	1	-							36
11	22	3	1	1	-	2	-	-							29
12	16	9	2	2	-	1	-	-							30
13	9	3	3	1	-	-	-	1							17
14	17	5	-	-	-	-									22
15	12	2	-	-	1	1									16
16	10	1	1	-	1										13
17	3	-	-	-	-										3
18	7	3	1	1	-										12
19	7	1	-	1	-										9
20	4	-	-	4	-										8
21	1	-	1	-	1										3
22	4	1	-												5
23	3	-	1												4
24	4	-													4
25	1	-													1
26	3	1													4
27	3														3
28	2														2
30	1														1
51	1														1
Total	1165	595	442	273	165	123	185	169	122	146	122	55	1	1	3570

it is divided by 30 and further to calculate in terms of years. it is divided by 12. Hence, idle time (y in years of the book is calculated as

y= (((current date - d_1) - 30 days)/30)/12.

y is rounded off to the nearest year. Summary of the data is given in Table 1 and Table 2.

4. Data Analysis

The data have been analyzed in relation to the number of times a book was issued. Table 1 gives the distributions of transactions in a special library for library science and related topics, mathematical sciences, economics and related topics and "other subjects". The Table 1 also gives the data for the entire collection, for all the subjects together (in column (6)).

Attempts were made to fit several probability distributions to identify a suitable one to explain the distribution of transactions. Based on the Kolmogorov-Smirnov test, it is observed that the negative binomial distribution fits the data in the field of library science, mathematical sciences, economics and in "other subjects". This observation is similar to that of the earlier observations by Ravichandra Rao [4, 5, 6,] Burrell [1] and also by Leemans et al [2]. However, it does not fit the data for the "entire collection".

4.1. *Bivariate Distribution*

Let f (x, y) be the number of documents which were circulated x times and the idle time of the book was y years. In order to study how far y influences x, the data pertaining to f(x, y) are analyzed. The different values of f(x, y) are given in Table 2.

Attempts were made to fit three bivariate distributions(negative binomial, log-normal and exponential) to the data given in Table 2. It is observed that none of the three bivariate distribution fits the empirical bivariate distribution. A fairly good result was however obtained only for the bivariate negative binomial distribution. The following recursive relations were used to compute the parameters of the bivariate negative binomial distribution.[5]

$$f(1, 1) = (1+g_0 + g_1 - \beta_2)^{-k}$$

$$f(u, 1) = \frac{(g_0 - \beta_2)(K+u-2) f(u-1,1)}{(1+g_0+g_1-\beta)(u-1)}$$

u =2. 3. 4. . . . . .

$$f(u,v) = \frac{(g_1 - \beta_2)(K+v-2) f(1,v-1)}{(1+g_0+g_1-\beta_2)(v-1)}$$

v = 2, 3, 4, ...

$$f(u,v) = \frac{(g_1-\beta_2)(K+u+v-3) f(u,v-1) + \beta_2 (K+u-2) f(u-1,v-1)}{(1+g_0+g_1-\beta_2)(u-1)}$$

$u, v \geq 2$.

The values of the constants are computed using the moments methods[8]. In the above equation u and v are defined as x + 1 and y +1 respectively. The values of the constants are

g_0	= 2.462024	g_1	= 2.997117	correlation coefficient
K	= 0.941705	β_2	= -11.044113	$r_{(x,y)}$ = -0.284078.
$\bar{u}$	= 3.3185	$\bar{v}$	= 3.8224	
σ_u	= 3.7868985	σ_v	= 3.2081926	

The estimated values of f (x, y) are given in Table 3. Marginal distributions clearly suggest that the bivariate negative binomial does not yield good result. However, the result is much better than the result obtained for the other two bivariate distributions. (log-normal and exponential).

Table 3. Estimated value of f(x,y) (under the assumption that data follow a bivariate negative binomial distribution)

x \ y	1	2	3	4	5	6	7	8	9	10	11	12	Total
1	241	182	142	112	88	70	56	44	35	28	22	18	1038
2	175	95	71	53	40	30	40	30	23	17	13	10	708
3	131	95	71	53	40	30	23	17	13	10	7	8	496
4	99	70	51	38	28	20	15	11	8	6	4	3	353
5	75	52	37	27	19	13	9	7	5	3	2	1	250
6	58	39	27	19	13	9	6	4	2	1	1		170
7	44	29	20	13	9	6	4	2	1	1			129
8	34	22	14	9	6	4	2	1					92
9	26	16	10	6	4	3	1						66
10	20	12	8	4	3	1							48
11	15	9	5	3	1								33
12	12	7	4	2	1								26
13	9	5	3	1									18
14	7	4	2	1									14
15	5	3	1	1									10
16	4	2	1										7
17	3	2	1										6
18	2	1											3
19	2	1											3
20	1	1											2
21	1												1
22	1												1
23	1												
Total	956	682	496	465	271	202	151	113	85	66	49	38	3484

4.2. Conditional Distributions

Attemps were made to fit the negetave binomial distribution to each of the distributions given in columns 1 to 12 of Table 2. It is thus observed that the distributions of transactions - x, the number of times a book is issued for given a y (the idle time of the book), follow a negative binomial distribution. The results are given in Tables 4 to 6. In Table 6, "m. e." refers to the case where moment's method is used to estimate the parameters; "m. l. e" (maximum likelihood estimate) refers to the cases where maximum likelihood methods is used to estimate the parameters. The "zero cell" refers to the cases where the relative frequency (f(0), i. e., f(x) for x=0) is used to estimate k, for known p, i. e. by equating p^k to f(0). Where the m. l. e. could not be computed, the m. e. was used.

The probability distribution function of the negative binomial is given by

$$p(x) = \frac{(k+x-2)!}{(k-1)!(x-1)!} p^k q^{x-1}$$

$x = 1. 2. 3, \ldots .$

$0 \leq p \leq 1, q = 1 - p$

The mean and variance of the distribution are given by $(kq/p) + 1$ and kq/p^2 respectively. The values of the mean, variance, p, and k are given in Table 5.

4.3 Regression Analysis

Regression analysis is yet another approach to identify a suitable model to predict the number of documents which were transacted x times for given y. Attempts have been made here to identify a suitable model to explain the transaction distribution. The following models were tried;

I $z = a + bx$ (linear model)

II $z = a + b/x$ (linear : inverse relation)

III $z = a + b/x + c/x^2$ (quadratic)

IV $z = ax^b$ (non-linear: generalized bibliometric model; Lotka's model is a special case of this model)

Parameters were computed using the least square method and these are given in Table 7. The values of $R^{**}2$ in Table 8, suggest that the model III (quadratic equation: $z = a + b/z + c/x^2$) explains

Table 4. Expected frequencies under the assumption that the conditional distribution follows a negative binomial distribution

x \ y	1	2	3	4	5	6	7	8	9	10	11	12
1	295	214	233	113	92	54	108	87	69	115	92	34
2	185	112	77	45	27	25	38	34	28	12	23	10
3	135	73	43	28	15	15	18	18	13	6	8	5
4	104	51	27	20	9	10	9	11	6	3	3	3
5	82	37	18	14	6	6	5	7	3	2	1	1
6	65	27	12	11	4	4	3	4	1	2	1	1
7	52	20	9	8	3	3	2	3	1	-	-	-
8	42	15	6	7	2	2	1	2	1	-	-	-
9	35	11	5	5	2	1	1	1				
10	28	8	3	3	1	1		1				
11	23	6	2	3		1		1				
12	19	5	2	2								
13	16	4	1	1								
14	13	3	1	1								
15	11	2	0	3								
16	9	2	0									
17	7	1	-									
18	6	1	-									
19	5	1	-									
20	4	-	-									
21	3	-										
22	3											
23	2											
24	2											
25	2											
26	1											
27	1											
28	1											
30	1											
51	0											

Table 5. Value of mean, and variance, p and k

y	Mean	Variance	p	k
1	4.9082	24.3829	0.160285	0.746005
2	3.4017	10.9866	0.218602	0.671888
3	2.5090	6.9169	0.218083	0.420887
4	2.6813	10.6860	0.157339	0.479218
5	2.4061	6.6775	0.210556	0.375038
6	2.6829	6.0529	0.277989	0.647961
7	1.8108	2.0345	0.398535	0.584568
8	2.1361	3.6442	0.311754	0.568293
9	1.8443	1.7261	0.490380	0.807561
10	1.7055	4.7557	0.148343	0.122882
11	1.4609	0.8266	0.557631	0.563839
12	1.5818	1.8433	0.315638	0.427523

Table 6. Kolmogorov – Smirnov statistic

Y	Kolmogorov-Smirnov	Max D_α by the method m.e./m.l.e.	Max D_α by the method zero cell
1	0.0479629*	0.5643	0.06522
2	0.0557545	0.03397	0.04266
3	0.0646886	0.05846	0.6595
4	0.098652	0.13445	0.06359
5	0.1058758	0.08477	0.0853
6	0.122627	0.02262	0.03008
7	0.0999891	0.04298	0.02484
8	0.1046153	0.05193	0.02484
9	0.1231286	0.03393	0.04329
10	0.1125544	0.03325	0.08106
11	0.1202081	0.03083	0.043
12	0.1833823	0.17103	0.05966
	0.17312*	0.05643	--.---
	0.022761***	0.11184	0.04259

*α = 0.01; H_0 is rejected

**H_0 : Rejected, the data refer to the entire collection. (column 6 of Table) (x = 0,1,2,. . .)

***H_0 : Rejected, the data refer to the entire collection (Col (6) of Table 1) (x = 1, 2).

Table 7. Values of the constants of different models

y	Model II $z = a + b/x$		Model III $z = a + b/x + c/x^2$			Model IV $z = ax^b$	
	a	b	a	b	c	a	b
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	-1.7382	305.8072	-27.71375	657.9676	-386.2439	6.91877	-1.7948
2	-10.0337	222.7009	-21.10524	33.6134	-120.9125	6.19742	-1.9147
3	-18.2223	221.8105	-13.81068	179.3964	43.4301	5.72729	-1.9925
4	-12.6457	142.8783	-4.92433	72.7065	71.0007	4.89722	-1.6865
5	-7.5401	85.3125	-6.23335	74.6111	10.5924	4.45326	-1.6970
6	-4.4636	57.5147	-5.32526	64.1240	-6.4013	4.26130	-1.6845
7	-19.8682	126.6031	-3.75233	26.6881	93.6881	5.08572	-2.3825
8	-12.8468	10.0232	-5.18573	49.4219	49.3261	4.71980	-1.9652
9	-11.7791	79.3957	-11.52367	77.9718	1.2529	4.46946	-2.0332
1	-17.8948	108.9413	-1.07168	3.3071	97.2995	3.95312	-1.5072
1	-23.5776	109.9859	-9.21792	37.6075	61.1307	4.55739	-2.5176
1	-9.6925	47.6442	5.91723	37.9452	74.7207	3.20365	-1.6849

Note : (1) The values of y in column (s) refer to the different conditional distrubutions in column (1) to (12) of Table 2.
*(2) values of the constants of model I are not given since R ** 2 are very low as may be seen in Table 8.*

Table 8. Values of R2**

y	$z = a + bx$	$z = a + b/x$	$z = a + b/x + c/x^2$	$z = ax^b$
(1)	(2)	(3)	(4)	(5)
1	0.443556	0.866178	0.987623	0.879609
2	0.467005	0.960643	0.983514	0.90125
3	0.345559	0.983741	0.992303	0.847208
4	0.324789	0.973514	0.992303	0.847208
5	0.357659	0.962068	0.983226	0.905678
6	0.515110	0.977234	0.978082	0.902964
7	0.524152	0.968745	0.969589	0.921246
8	0.439823	0.971475	0.986233	0.917309
9	0.619848	0.969581	0.969602	0.939028
10	0.158664	0.937342	0.995143	0.810973
11	0.636744	0.971522	0.985063	0.855912
12	0.376776	0.866192	0.986431	0.855912

Note: In column 1. the value of 'y' refer to the different conditional distributions in column 1 to 12 of Table 2.

Table 9. Estimated values of f(x/y) for different values of y, for model III

x \ y	1	2	3	4	5	6	7	8	9	10	11	12
1	244	197	209	129	78	52	115	94	68	100	90	43
2	205	118	87	49	34	25	34	32	28	25	25	6
3	139	78	51	27	20	15	15	17	15	11	10	2
4	113	60	34	18	13	10	9	10	8	6	4	1
5	88	42	24	12	9	6	5	3	4	2	1	2
6	71	32	17	9	6	5	3	4	2	3	*	2
7	58	25	13	7	5	4	2	3	*	*		
8	48	19	9	5	3	3	1	2				
9	41	15	7	4	1	2		1				
10	34	12	5	2	*	0		*				
11	28	9	3	2	*	*						
12	24	6	1	1	*	*						
13	21	4	0	*								
14	17	2	*	*								
15	14	1	*	*								
16	12	*	*									
17	10	*	*									
18	8	*										
19	6	*										
20	4	*										
21	3											
22	1											
23	0											
24 to 51	*											

*Note: * Refers to negetive value*

Table 10: A summary of book transactions

Subject	Number of books which were used at least once	Number of books which were not used even once
Mathematical sciences	1560 (48.80%)	1637 (51.20%)
Library science and related topics	913(39.40%)	1373(62.60%)
Economics and related topics	563(37.33%)	945(62.67%)
Other subjects	533(28.46%	1340(71.54%)
Total collection	3570(40.27%)	5295(59.73%)

Table 11 : Distributions of idle time of the book

Percentage of books which were used at least once	Idle time after the last issue (in years)
10	0.23
20	0.5
30	1.0
40	1.5
50	2.3
60	3.0
70	4.3
80	6.6
90	8.8
95	10.0
98	10.8
99	11.2
100	19.3

the transaction data much better than any of the other three models. The generalized bibliometric model fits equally well and the value of b is very close to -2, the case of Lotka's function. The estimated values of z(i. e. for $f(x|y)$) for model III are given i n Table 9.

4.4. Concentration of Documents

Table 10 shows that the number of books which were issued out (at least once over a period of more than twenty years) was only 3570, which is about 40.20% of the total book collection. This gives an impression that the collection is hardly being utilized. But, this may be due to the fact that:

i) The collection policy might not be in consonance with users' needs; and /or
ii) It is quite likely that users are not aware of the availability and accessibility of books.
iii) The most of the books might have been used only withing the library, but not issued out!

On the basis of further analyses of the data (by subject), it was observed that

i) In mathematical sciences, nearly 50% of the books were issued out; and
ii) In library science, economics and related topics, 40% of the books were issued out.
iii) In other disciplines, only 29% of the books were issued out.

Data in Tables 1,9 and 10 suggest that

i) 60% of the books were not issued out;
ii) 16% of the books were issued out only once;
iii) 8.2% of the books were issued out twice;

iv) 6.3% of the books were issued out five times;
v) 4.3% of the books were issued out six times;
vi) Only one percent of the books were issued out more than 15 times.

In general, while only very few books are being circulated again and again, most of the books are hardly being borrowed. (Most of them, in fact, are not issued out at all).

On the other hand, Table 11 gives the data on the distribution of time gaps between the date of last circulation and the current date (idle time), From Table 11, it may be observed that the median idle time of a book is 2.3 years; that is 50% of the books were issued at least once in the last 2.3 years.

In the special library, wherein this study was conducted, if an automated circulation system is introduced, it may be enough to prepare the machine readable records for the 60% of the active collection (for those books which were issued out at least once in the last three years). For the other 40% of the collection (for those books which were not used even once in the last three years). the machine readable records, may be prepared as and when it is required .

This approach to circulation analysis further indicates that a very small portion of the library's holding accounts for a larger fraction of the circulation transactions. The statistics of last circulation date have thus the "potential in the quantitative method of thinning the stacks"[9]. This statistics also has the additional advantage of reflecting user requirements and does not reflect the subjective judgment of individuals so often used in stack thinning.

It may thus be possible to define the core collection, based on this type of study, which satisfies 90% of the user requirements. Other questions which one may answer from this type of study are:

i) Question of multiple copy needs;
ii) Physical organization of documents not used frequently/recently;
iii) Size of a library holding.

This study indicates that the conditional distribution of transactions

1. follows a negative binomial distribution
2. can be explained by ā quadratic model, using a regression analyis
3. can equally be explained by a generalized bibliometric model; it is very close to Lotka's model.

Further, data suggest that over a period of time, number of frequently used books increases. This can be seen from Table 2. as y increases from column 1 to 12, the tail of the distribution decreases. It implies that the core size consisting of the frequently used documents is increasing over a period of time.

References

1. Burrell, Q. L. and Cane, V. R. (1982), The analysis or library data. *Journal of the Royal Statistical Society*, SEries A. **145**. 439-71.
2. Leemans, Marrie-Jeanne; Maes, Marleen; Rousseasu, R; and Ruts, Christel (1992). The negative binomial distribution as a trend distribution for circulation data in Flemish public libraries. *Scientometrics*. (1) 48-57.
3. Morse, P. M. (1968). *Library effectiveness*. MIT Press. Cambridge. Mass.
4. Ravichandra Rao, I. K (1980). Distribution of scientific Productivity and social change. *Journal of American Society for Information Science*. **31** (2) 111-3.
5. Ravichandra Rao, I. K. (1981) *Documents and user distribution in transaction records of Canadian University Libraries*. Ph. D. Thesis. Faculty of Graduate Studies. School of Library and Information Science; the University of Western Ontario, London, Ontario, London, Ontario, Canada.
6. RavichandraRao, I; K. (1982). Documents and users distribution *Library Science with a Slant to Documentation* **19**. 69-96.
7. Sichel, H. S. (1985). A bibliometric distribution which really works. *Journal of the Americal Society for Information Science*. **36** (5); 314-21, 1985.
8. Subramaniam, Kocherlakota and Subramaniam, Kathleen. On the information of the parameters in the bivariate negative binomial distribution *Journal of Royal Statistical Society*. Series B, **35**(1); 131-146, 1973.
9. Trueswell Richard W. A quantitative measure of user circulation requirements and its possible effect on stack thinning and multiple copy determination *American Documentation*; **16**; 20-25. 1965.

Some Elementary Characterizations for a Class of Skew Distributions with Applications to Scientific Productivity Analysis

Shan Shi and He Jiajun
Document and Information Management Department, Liberal Arts College,
Shanghai University, San Men Road, Shanghai 200434, China

Some elementary characterization for a class of skew distributions are presented. An equivalent relation between the Waring distribution and the truncated first moment with a particular linear form is established. The Yule distribution is shown to exhibit the important invariance property that the distribution of observed productivity, suitably truncated ,coincides with the true distribution if and only if the distribution is of the Yule form. An illustrative example is provided.

1. Introduction and Summary

The Waring distribution has an important place in distribution analysis of scientific productivity.

A special case, named the Yule distribution, of the Waring distribution was derived in the classic paper of Price (1976). Authors such as Simon (1955), Price (1976), Tague (1981), and Schubert and Glänzel (1984) have asserted law - like adherence to a Waring-Yule-type model. Moreover, experience as evidenced by an impressive variety of data sets, supports the assertion that, a Waring-Yule-type distribution provides a satisfactory model of publication productivity in different scientific fields.

In Section 2, the (shifted) Waring distributior is exactly characterized by the expectation of the distribution truncated from the left at m. Let the latter be denoted by E_{m+1}. An equivalent relation between the (shifted) Waring distribution and E_{m+1} with a particular linear form of m+1 is established.

Truncation is specially important in cases where the relation between the fraction of sources having m + 1 or more items and the fraction of items yielded by the sources having m + 1 or more items is studied. (A familiar model has been first developed by the Glänzel et al (1984)).

Let ρ and θ be as follows :

ρ = ρ(m + 1) = the fraction of sources having m + 1 or more items (i.e. , the fraction of authors having m + 1 or more papers).

θ = θ (m + 1) = the fraction of items yielded by the sources having m + 1 or more items (i.e., the fraction of papers produced by the authors having m + 1 or more papers).

Authors such as Egghe (1986) and Burrell (1985) have studied the relation between ρ and θ, and discussed the "80 / 20-effect" derived from them. θ / ρ may be regarded as the ratio of output to input. From the definition of E_{m+1}, we have

$$\frac{\theta}{\rho} = \frac{\sum_{k=m+1}^{\infty} kp(X=k)}{\sum_{k=1}^{\infty} kp(X=k)} \cdot \frac{1}{\sum_{k=m+1}^{\infty} p(X=k)}$$

$$= \frac{E(X \mid X \geq M+1)}{E(X \mid X \geq 1)} = E_{m+1} / E_1.$$

so the conclusion in Section 2 that E_{m+1} having a particular linear form of m + 1 is equivalent to the Waring distribution will give a insight into the strength of the "output/input effect."

In Section 3, observed or reported productivity Y is assumed to relate to true productivity X according to the equation

Y = [UX]

where [UX] denotes the integral part of UX and U is a uniform (0,1) variable distributed independently of X, and under the assumption and the other conditions it is proved that the distribution of Y truncated from the left at m coincides with the distribution of X if and only if X has the Yule distribution truncated from the left at m (m is a non-negative integer). [A special case where m = 1 has been examined by Krishnaji (1970).]

Characterization by under-reported data is specially important in cases where a suitable distribution governing the true values on the basis of a sample of under-reported data is selected and used in prediction problems.

Publication data bases, such as a bibliography covering a certain span of years on a particular subject and the Science Citation Index (SCI), have been succesfully used in scientific productivity studies. But this approach, however, has its own problem. No publication data base can be free from geographic and language biases.

Data reported in a publication data base represents, however, only a fraction of the total scientific productivity (in our terminology : it is an understatement of true scientific productivity). A true distribution on the basis of under-reported publication data will provide a new indicator to measure scientific productivity potential.

In Section 4 a new index derived from the true distribution, named the coefficient of the potential of scientific productivity potential, is presented for scientific productivity analysis. We provide an illustrative example. Utilizing the previous results, we study the scientific potential and true productivity distributions in the information science field of China with respect to the different time spans and periods.

2. Waring Distribution and Its Characteristic Property

2.1. The Waring Distribution

It is well known that the Waring distribution of random variable X is given by the following

$$p(X=k)=\alpha\frac{\beta^{(k)}}{(\alpha+\beta)^{(k+1)}}$$

$$=\frac{\alpha\,\Gamma(\beta+k)\Gamma(\alpha+\beta)}{\Gamma(\beta)\,\Gamma(\alpha+\beta+k+1)},k=0,1,2\ldots, \quad (1)$$

where $\alpha>0$, $\beta>0$, and $a^{(b)}=\Gamma(a+b)/\Gamma(a), a>0, b>0$, and $\Gamma(.)$ is the gamma-function.

Now we consider a more general form of (1). Let the distribution be shifted ι units to the right and ι be a integer not smaller than 0. The resulting distribution is given by

$$p_\iota(X=k)=\alpha\frac{\beta^{k-\iota}}{(\alpha+\beta)^{(k-\iota+1)}},\quad k\iota=\iota,\ \iota+1,\ldots \quad (2)$$

The shifted Waring distribution is denoted as $p_\iota(X=k)$ in (2). It is obvious that $p_0(X=k)$ in (2), i.e., $p_\iota(X=k)$ for $\iota=0$, is $p(X=k)$ in (1).

2.2. Truncated Distribution and Conditional Expectation

Let X be a random variable defined on the set of non-negative integers : { ι, $\iota+1$, ..., k, ... } where $\iota\geq 0$. Then the following conditional distribution is called the distribution truncated from the left at m :

$$p(X=k|X\geq m+1)=\frac{P(X=k)}{\sum_{k=m+1}^{\infty}p(X=k)},$$

$$k=m+1,m+2,\ldots, \quad (3)$$

where m is a non-negative integer not smaller than ι, and the conditional expectation of X, given $X\geq m+1$, written as $E(X\mid X\geq m+1)$, is defined by

$$E(X|X\geq m+1)=\sum_{k=m+1}^{\infty}kp(X=k|X\geq m+1) \quad (4)$$

2.3. A Characteristic Property of the Waring Distribution

The following theorem will display the important property of the Waring distribution which exactly characterizes the distribution by the conditional expectation $E(X\mid X\geq m+1)$. The theorem has been partly proved by Glänzel et al (1984). Also they have generalized it to the Waring Distribution with three parameters, i.e., the generalized Waring distribution. But the ι - shifted case was

not dealt with in their study. We prove it in a more intuitive way and include the proof here for the sake of completeness. However, the characterization model based on the truncated moments for the discrete skew distributions was first developed by Glanzel et al.

THEOREM 1. Assume α to be greater than one. Also suppose that $\beta > 0$, and $m \geq \iota$, both m and ι are non-negative integers, and $k = \iota, \iota+1, \ldots$, then

$$p_\iota(X=k) = \frac{\alpha\beta^{(k-\iota)}}{(\alpha+\beta)^{(k-\iota+1)}}$$

if and only if

$$E(X|X \geq m+1) = \frac{\alpha(m+1)-\iota}{\alpha-1} + \frac{\beta}{\alpha-1} \qquad (6)$$

PROOF. Necessity. Assume that $p_\iota(X=k)$ has the form in (2), it is easy to check that

$$p_\iota(X=k) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta(1-\theta)^{k-\iota}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

which says that $p_\iota(X=k)$ is a geometric-beta compound distribution. The probability generating functions (p.g.f.) of the Waring distribution is given by

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \frac{\theta^\alpha(1-\theta)^{\beta-1}s^\iota}{1-(1-\theta)s}d\theta \qquad (7)$$

Hence we have the relation.

$$g(s) = \sum_{k=m+1}^{\infty} p_\iota(X=k)s^k$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \frac{\theta^\alpha(1-\theta)^{m-\iota+\beta}s^{m+1}}{1-(1-\theta)s}d\theta \qquad (8)$$

From (8), we have

$$\sum_{k=m+1}^{\infty} p_\iota(X=k) = g(1) == \frac{\beta^{(m+1-\iota)}}{(\alpha+\beta)^{m+1-\iota}} \qquad (9)$$

and

$$\sum_{k=m+1}^{\infty} kp_\iota(X=k) = \frac{d}{ds}g(s)|_{s=1}$$

$$= \frac{\beta^{(m+1-\iota)}}{(\alpha+\beta)^{(m+1-\iota)}}\left[\frac{\alpha(m+1)-\iota}{\alpha-1} + \frac{\beta}{\alpha-1}\right] \qquad (10)$$

Substituting (9) into (5) and then (10) into (4), we obtain

$$E(X|X \geq m+1) = \frac{\alpha(m+1)-\iota}{\alpha-1} + \frac{\beta}{\alpha-1}.$$

Sufficiency. Assume that $E(X|X \geq m+1) = \frac{\alpha(m+1)-\iota}{\alpha-1} + \frac{\beta}{\alpha-1}$, because α has been assumed to be greater than one, $E(X|X \geq m+1)$ exists. From the relation :

$$E(X|X \geq m)\sum_{k=m}^{\infty} p(x=k) - E(X|X \geq m+1)$$

$$\sum_{k=m+1}^{\infty} p(X=k) = mp(X=m),$$

the following will thus hold for $k \geq \iota, m \geq \iota$,

$$\frac{\alpha m-\iota+\beta}{\alpha-1}\sum_{k=m}^{\infty} p(X=k) - \frac{\alpha(m+1)-\iota+\beta}{\alpha-1}$$

$$\sum_{k=m+1}^{\infty} p(X=k) = mp(x=m) \qquad (11)$$

Let

$$U = p(X=m) \text{ and } V = \sum_{k=m+1}^{\infty} p(X=k),$$

we have

$$U(\alpha-1)m = (U+V)(\alpha m-\iota+\beta) - V\left[\alpha(m+1)-\iota+\beta\right]. \qquad (12)$$

From (12) the ratio of U to V is

$$\frac{U}{V} = \frac{p(X=m)}{\sum_{k=m+1}^{\infty} p(x=k)} = \frac{\alpha}{m-\iota+\beta} \qquad (13)$$

Then we have

$$\sum_{k=m+1}^{\infty} p(X=k) = \frac{m-\iota+\beta}{\alpha}p(X=m) \qquad (14)$$

and

$$\sum_{k=m}^{\infty} p(X=k) = \frac{\alpha+\beta+m-\iota}{\alpha}p(X=m). \qquad (15)$$

It thus follows that

$$\sum_{k=m+1}^{\infty} p(X=k) = \frac{\beta^{(m-\iota+1)}}{(\alpha+\beta)^{m-\iota+1}}$$

which has been given in (9). This result says that X follows the (shifted) Waring distrubution.

3. Yule Distribution and its Charactristic Property

3.1. A Model for Under-Reported Data

Let X be a random variable such that p (X = k), k = 1, 2, ... Then, as found by Bissinger (1964) a new distribution can be defined by

$$q(Y=k) = \sum_{i=k+1}^{\infty} \frac{p(X=i)}{i}, k = 0,1,2,.... \quad (16)$$

One may legitimately expect that the observed or reported magnitudes are under-statement of the true values of the variable. Assume that X is unobservable, but is related to an observable Y according to the equation :

$$[UX] = Y \quad (17)$$

where [UX] denotes the integral part of UX, U is a random variable distributed independently of X and uniformly in (0,1). Then (16) gives precisely the distribution of [UX] for

$$p([UX]=k) = p(k \le UX < k+1)$$

$$= \sum_{i=k+1}^{\infty} p(X=i)p\left(\frac{k}{i} \le U < \frac{k+1}{i}\right)$$

$$= q(Y=k), \quad k = 0,1,2,....., \quad (18)$$

(18) and the probability generating function of q(Y = k) were discussed by Krishnaji (1970).

We consider a deferent model which was introduced by Rao (1964) : Here an observation may be somewhat destroyed. If the probability of an original observation X be P (X) and the chance of k survivors out of is S (k,X), then the probability of observing Y = k is

$$q(Y=k) = \sum_{i=k}^{\infty} p(X=i)s(k,X=i), \quad k=0,1,2,... \quad (19)$$

Let where p (x = 0) = 0, s (x,x) = 0, and s (k, X = i) = 1 / i, (19) will reduce to (16).

From (16) it follows that

$$q(Y=k) - q(Y=k+1) = \frac{p(X=k+1)}{k+1},$$

$$k=0,1,2,... \quad (20)$$

3.2. The left-Truncated Yule Distribution

The Waring distribution

$$p(X=k) = \alpha\frac{\beta^{(k)}}{(\alpha+\beta)^{(k+1)}} \quad k=0, 1, 2. ...,$$

in the special case where β = 1 reduces to

$$p(X=k) = \alpha\frac{k!}{(\alpha+1)^{k+1}} = \alpha\beta(\alpha+1,k+1) \quad (21)$$

(β (. , .) denotes the beta-function), which is called the Yule distribution.

Let $\iota = 0$ and β = 1, the formula in (9) will reduce to

$$\sum_{k=m+1}^{\infty}\alpha\frac{k!}{(\alpha+1)^{(k+1)}} = \sum_{k=m+1}^{\infty}p(X=k) = \frac{(m+1)}{(\alpha+1)^{(m+1)}} \quad (22)$$

From (22) the Yule distribution truncated from the left at m, i.e., the left-truncated Yule distribution , is given by

$$p(X=k|X \ge m+1) = \alpha\frac{(2+m)^{(k-m-1)}}{(\alpha+2+m)^{(k-m)}}$$

$$k=m+1, m+2, ..., \quad (23)$$

where m is a non-negative integer.

3.3. Characterization by Under-Reported Data

In the 18th century, Waring showed that the function x / (x - a), x > a > 0, can be expanded in the following way :

$$\frac{x}{x-a} = \sum_{i=0}^{\infty}\frac{a^{(i)}}{(x+1)^{(i)}}$$

$$= 1 + \frac{a}{x+1} + \frac{a(a+1)}{(x+1)(x+2)} + ... \quad (24)$$

Let a non-negative integer m be given and [UX] denote the integral part of UX. We have the following theorem.

THEOREM 2. Let X be a positive integral-valued random variable not smaller than m + 1. Let U be a random variable distributed independently of X and uniformly in (0,1), and Y = [UX]. Then

$$q(Y=k|Y\geq m+1)=p(X=k|X\geq m+1)$$

for $K = m+1, m+2, \ldots\ldots$, and $\sum_{k=0}^{m} q(Y=k) = 1-(m+1)q(Y=m) > 0,$

if and only if X follows the Yule distribution truncated from the left at m.

PROOF. *Necessity*. Substituting (23) into (16) and using (24) we get

$$q(Y=k)=\begin{cases}\dfrac{\alpha}{(1+m)(\alpha+2+m)}\sum_{i=0}^{\infty}\dfrac{(1+m)^{(i)}}{(\alpha+3+m)^{(i)}}=\dfrac{\alpha}{(\alpha+1)(m+1)} & \text{for } k=0,1,2,\ldots m;\\ \alpha\dfrac{(2+m)^{(k-m-1)}}{(\alpha+2+m)^{(k-m+1)}}\sum_{i=0}^{\infty}\dfrac{(k+1)^{(i)}}{(\alpha+3+m)^{(i)}}=\alpha\dfrac{(2+m)^{(k-m-1)}}{(\alpha+2+m)^{(k-m)}}\cdot\dfrac{1}{\alpha+1} & \text{for } k=m+1,\ m+2,\ \ldots\ldots\end{cases} \tag{25}$$

From (25) it may be seen that $1/(\alpha+1) = 1-(m+1)\,q(Y=m)$. Let $q(Y=m) = q_m$, we have

$$q(Y=k)=\alpha\frac{(2+m)^{(k-m-1)}}{(\alpha+2+m)^{(k-m)}}\cdot\frac{1}{\alpha+1}$$

$$=\left[1-(m+1)q_m\right]p(X=k|X\geq m+1) \tag{26}$$

(see (23)) for $k=m+1, m+2 \ldots$, and

$$1-\sum_{k=0}^{\infty}q(Y=k)=\sum_{k=m+1}^{\infty}q(Y=k)$$

$$=1-(m+1)q_m=1/(\alpha+1), \tag{27}$$

We thus have $1-(m+1)\,q_m > 0$, and

$$q(Y=k|Y\geq m+1)=p(X=k|X\geq m+1).$$

Sufficiency. Now Y=[UX] follows the distribution given by (16) we further assume

$$q(Y=k|Y\geq m+1)=p(x=k|X\geq m+1)$$

and

$$\sum_{k-m+1}^{\infty}q(Y=k)=1-(m+1)q_m>0 \tag{28}$$

From (28) and (16) it is easy to show that

$$1-(m+1)q_m<1. \tag{29}$$

We have

$$q(Y=k)=\left[1-(m+1)q_m\right]p(X=k|X\geq m+1),$$

$$k=m+1, m+2, \ldots, \tag{30}$$

Since

$$q(Y=k)-q(Y=k+1)=\frac{p(X=k+1|X\geq m+1)}{k+1} \tag{31}$$

(see (20)), substituting (30) into (31) yields

$$p(X=k|X\geq m+1)$$

$$=\frac{k}{k+1+\alpha}p(X=k-1|X\geq m+1)$$

where

$$\alpha=\left[1-(m+1)q_m\right]^{-1}- \tag{32}$$

From (28) and (29) it follows that $\alpha>0$. Then we have

$$p(X=k\mid X\geq m+1)=\frac{k}{k+1+\alpha}\cdot\frac{k-1}{k+\alpha}\cdot\ldots$$

$$\cdot\frac{m+2}{m+3+\alpha}p(X=m+1|X\geq m+1). \tag{33}$$

Let $k=m+1$ in (30) and $k=m$ in (31), we get

$$p(X=m+1\mid X\geq m+1)=\frac{(\alpha+1)(m+1)q_m}{m+1+\alpha+1} \tag{34}$$

From (32) it is easy to check that

$$p(X=m+1\mid X\geq m+1)=\frac{\alpha}{m+2+\alpha}. \tag{35}$$

Substituting (35) into (33), we finally obtain

$$p(X=k|X\geq m+1)=\alpha\frac{(2+m)^{(k-m-1)}}{(\alpha+2+m)^{(k-m)}}$$

which is the Yule distribution truncated from the left at m.

The case m=0 has been examined by Krishnaji (1970).

So long as these reporting errors tend to influence all the true values of the variable in the

way that the observation error variable is a uniform (0,1) one, even if for the m- truncated case, the Yule-type model will be preserved. This sturdiness gives us new insight into the meaning of the Yule-type model and additional confidence that the model provides at least a good first-approximation description of the productivity data.

4. Applications

4.1. A Potential Analysis of Scientific Productivity

Let the expectation of a distribution truncated from the left at m be denoted by E_{m+1}, namely, E_{m+1} is the conditional expectation E (X|X > m + 1). Let e_{m+1} be the sample mean of the m +1– truncated sample given by

$$e_{m+1} = \sum_{k=m+1}^{\infty} kn_k \Big/ \sum_{k=m+1}^{\infty} n_k$$

where n_k is the kth absolute frequency, namely, the number of authors having k papers, e_{m+1} can be used as an estimator of E_{m+1}.

Let the observable productivity Y have the Yule distribution (the case β = 1 of the Waring distribution with i = 0 in (2)), we have (see (6))

$$E(Y|Y \geq m+1) = \frac{\alpha(m+1)}{\alpha-1} + \frac{1}{\alpha+1} \quad m = 0,1,2...,$$

Also when Y follows the Yule distribution defined as

$$q(Y=k) = \alpha\beta(\alpha+1, k+1), \quad k = 0,\ 1,\ 2,...,$$

we have

$$E(Y) = E(Y|Y \geq 0) = \frac{1}{\alpha-1} \qquad (37)$$

Combining (36) and (37), we have the following linear formula

$$E(Y|Y \geq h) = \frac{\alpha}{\alpha-1}h + \frac{1}{\alpha-1} \qquad (38)$$

where h is a non-negative integer. The series e_1, e_2, ...can be calculated directly from the observable data and an estimator, α_t, of α can be obtained by some proper methods (see 4.2 for details).

Now we assume that the observable productivity Y is related to true productivity X through the relation Y = [UX] where U is distributed uniformly in (0,1) and independently of X and [.] denotes the integral part of " . ". From Theorem II, the distribution of Y is related to the distribution of X through the following equation

$$\frac{q(Y=k)}{1-q(Y=0)} = P(X=k|X \geq 1).$$

We have ,

$$\frac{q(Y=k)}{1-q(Y=0)} = \alpha(\alpha+1)\beta(\alpha+1, k+1), \ k = 1,\ 2,..., \qquad (40)$$

(40) can thus be used to estimate the potential of scientific productivity (relative frequence of "authors" having true productivity X). The ratio of P (X = k | X ≥ 1) to q (Y = k) is α + 1, an estimator, α_t, of *a* can thus be defined as the *coefficient of the potential of scientific productivity* (CPSP).

4.2. An Analysis of Research Potential of Information Science in China

We present an empirical dataset about the productivity distributions of information science in China. The dataset includes an actual distribution of author productivity among five major Chinese journals in information science field for the years 1987 - 1992. The five journals are IS, JIS, PIS, RIS and TSI. The complete count, crediting all collaborating authors, is used.

Let the n observable variables be ranked in an increasing order, i.e., $y_1 < y_2 < ... < y_n$, where $y_1 = 1$ usually and y_n presents the maximum value of the productivity variable y_i. The sample mean of y_i-truncated sample is denoted as e_i. Plotting e_i versus y_i it will be found that the sample step — curve is asymptonic to a straight line. The theoretical reasons for this have been sought previously (see (38)).

From (38) the parameter α can be estimated through the following equation

$$e_i = \frac{\alpha}{\alpha-1}y_i + \frac{1}{\alpha-1}.$$

The estimated value based upon e_i is denoted as α_i, i.e.,

$$\alpha_i = \frac{e_i+1}{e_i-y_i}, \qquad i = 1, ... n-1 \qquad (41)$$

(note $e_n - y_n = 0$, therefore α_n is invalid and that i = n is omitted). An estimator, α_t, of α, i.e., the coefficient of the potential of scientific productivity (CPSP), is taken as

$$CPSP = \left(\prod_{i=1}^{n-i} \alpha_i \right)^{\frac{1}{n-1}} \quad (42)$$

which is the geometric mean of α_i's.

For our dataset the sample step-curve consists of two identifiable regions : a large linear portion containing the low, moderate and prolific authors, and a very short dropping tail of prolific contributors. The latter region is characterized by large and discontinuous values of the variable y_i and lack of ties for productivity scores. Correspondently, the α_i whose associated y_i is in the latter region would be higher than the α_i that its y_i is in the previous region through the estimation procedure (41). For estimating CPSP, (42) would make that those α_i 's whose associated y_i's are in the previous region outweigh the other α_i's whose y_i's are in the latter one by smoothing the α_i's with the geometric mean.

It is necessary that the number of the invisible authors (the authors who produced zero papers) should be estimated from the observable data in order to predict the true distribution of scientific productivity. The number of the invisible authors is estimated by

$$n_0 = (\alpha_t + 2) n_1 \quad (43)$$

which is derived from the recursion relation

$$q(Y = k+1) = \frac{k+1}{\alpha + k + 2} q(Y = k), \text{k=0,1,2,...} \quad (44)$$

where Y has the Yule distribution.

In our statistical analysis, the six-year period 1987-1992 is divided into two comparative parts, i.e., the first three years 1987-1989 and the second three years 1990-1992. Chinese potential of scientific productivity in information science is analysed by our statistical model with the deferent time spans (three years and six years) and periods.

Also the productive manpower that was estiamted by the above procedure is contained in the first three tables(1,2,3). It should be mentioned that a familiar method to estimate publication potential was given by Schubert et al (1986) although in a deferent setting. Obviously, our model is akin to theirs.

Finally, the pairs (y_i, e_i) (i = 1,2,...,n) were measured by the sample correlation coefficient r. The results are presented in Table 4. All the values of r are fairly high (more than 0.95).

Table 1. Author productivity distribution of information science in China 1987 - 1989

i	y_i	# of authors	# of expected authors	# of expected true authors
0	0	(5862) *	5856.85	------
1	1	861	860.29	4997
2	2	200	220.36	1280
3	3	73	75.05	436
4	4	30	30.60	177
5	5	23	14.16	82
6	6	5	7.19	42
7	7	7	3.93	23
8	8	5	2.27	15
9	9	2	1.38	8
10	10	1	0.87	5
11	12	2	0.38	2
12	13	2	0.26	1.5
13	15	1	0.13	0.8
14	18	1	0.05	0.3
α = CPSP				4.80799
# of observable authors				1213
# of invisible authors				5862
productive manpower				7075

* it was estimated by (43), i, e., 5862 ≈ (4. 80799+2). 861

Table 2. Author productivity distribution of information science in China 1990 - 1992

i	y_i	# of authors	# of expected authors	# of expected true authors
0	0	(7005) *	6988.60	--------
1	1	935	932.74	6056
2	2	193	219.66	1426
3	3	60	69.42	451
4	4	34	26.46	172
5	5	10	11.51	75
6	6	12	5.53	36
7	7	3	2.8	18
8	8	4	1.58	10
9	9	2	0.92	6
10	10	2	0.55	3.6
11	15	1	0.07	0.5
α = CPSP				5.4925
# of observable authors				1256
# of invisible authors				7005
productive manpower				8261

* it was estimated by (43), i. e., 7005≈(4, 80799+2).935

Table 3. Author productivity distribution of information science in China 1987 - 1992

i	y_i	# of authors	# of expected authors	# of expected true authors
0	0	(9995)*	10043.07	--------
1	1	1482	1489.26	8555
2	2	370	384.61	2207
3	3	132	131.95.	758
4	4	59	54.16	311
5	5	35	25.20	145
6	6	28	12.87	74
7	7	15	7.07	40
8	8	10	4.11	24
9	9	9	2.51	14
10	10	7	1.59	9
11	11	5	1.04	6
12	12	2	0.70	4
13	13	4	0.49	3
14	15	2	0.25	1.44
15	16	1	0.18	1.03
16	17	1	0.13	0.75
17	19	1	0.08	0.46
18	20	2	0.06	0.34
19	24	1	0.02	0.11
α = CPSP				4.7442
# of observable authors				2166
# of invisible authors				9995
productive manpower				12161

* it was estimated by (43), i.e., $9995 \approx (4.7442+2).1482$

Table 4.Sample correlation coefficient of observable productivity values and means of truncated samples

Period	r
1987-1989	0.9545
1990-1992	0.9906
1987-1992	0.9869

Acknowlegement

The authors wish to thank Dr. Kretschmer for inviting them to the Berlin Conference on Informetrics where most of this work was presented and Dr. Glänzel for his encouragement of introducing their studies on characterization by truncated moment to the authors. One of the authors should like to thank Dr. Bao of theEconomics Department, Karlsruhe University, the author visited Karlsruhe, for his advice and kind hospitality.

References

1. Bissinger, H.B. A type-resisting distribution generated from considerations of an inventory model, In : *Classical and contagious discrete distributions*, Statistical Publishing Society, Calcutta, 1964, 15-17.
2. Burrell, Q.L. The 80 / 20 rule : Library Iore or statistical law ? *Journal of Documentation*, **41**,1985, 24-39.
3. Chen, W.C. On the weak form of Zipf's law, *Journal of Applied Probability*, **17**,1980, 611 - 622.
4. Egghe, L. On the 80 / 20 rule, *Scientometrics*, **19** , 1986, 55-68.
5. Glänzel, W.; A. Telcs and A. Schubert, Characterization by truncated moments and its application to Pearson-type distributions, *Zeitschrift für Warscheinlichkeitstheorie und Verwandte Gebiele*, **66**,1984, 173-183.
6. Hartley, M.L. and N.S.Revanker. On the estimation of the Pareto law from under-report data, *Journal of Econometrics*, **2**, 1974, 327-341.
7. Hill, B.M. The rank-frequency form of Zipf's law, *Journal of the American Statistical Association*, **89**,1974, 1017-1026.
8. Ijiri,Y.and H.A.Simon. *Skew distributions and the sizes of business firms*, North-Holland. Amsterdam,1977.
9. Irwin, J.O. The generalized Waring distribution; I, II, III. *Journal of the Royal Statistical Society*, Series **A, 138**,1975, 18-31, 204-227, 374-384.

10. Krishnaji, N. A characteristic property of the Yule distribution, *Sankhya* Series **A,32**,1970, 343-346.
11. Krishnaji, N. Characterization of the Pareto distribution through a model of under-reported incomes, *Econometrika*, **38**,1970, 251-255.
12. Price, D. A general theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science*, **27** 1976, 292-306.
13. Rao, C.R. On discrete distributions arising out of methods of ascertainment, In : *Classical and Contagious Discrete Distributions* : Statistical Pulishing Society, Calcutta, 1964, 320-332.
14. Schubert, A.and W. Glänzel. A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, **6**, 1984, 149-167.
15. Schubert, A. and A Telcs. Publication potential - an indicator of scientific strength for gross-national comparisons, *Scientometrics*, **9**, 1986, 231-239.
16. Shan, S. and T.M. Wang. The Simon model and its expansion (in Chinese), *Information Science*, 14 ,1993.
17. Simom, H.A. On a class of skew distribution functions, *Biometrika*, **42**, 1955,425-440.
18. Tague, J. The success-breeds-success phenomenon and bibliometric processes, *Journal of the American Society for Information Science*, **32**, 1981, 280-286.

With sense of deep loss and grief, we had to accept the news of sudden demise of Prof Dr H S Sichel. Despite his advanced age and ill health, Prof Sichel very gladly agreed to be on the Editorial board of JISSI. He proposed to help JISSI in all possible ways. We publish here the letter of Mrs Sichel as an obituary.

From : Mrs Hilda Sichel
24/10 Zalman Shazar
Rishon Lezion 75700
Israel.

Dated 15.09.1995

Subir K Sen
Hony Executive Editor
JISSI

Dear Mr Sen

It is with great sadness that I have to write and inform you that my dear husband Herbert Simon Sichel died on the evening of the 4th September.

He had been suffering from lung and heart problems for nearly three years. He had a spell in hospital in July of this year. Even though greatly debilitated physically, his mental powers were as strong as ever. His death was sudden. He was 79 years old and would have celebrated his 80th on the 8/11.

He will be a great loss to the field of Applied Statistics.

Yours sincerely,

Sd/Hilda Sichel

How Strongly Deterministic Is Chaotic Behavior in Computer Mediated Network Communication ?

Herbert Snyder
School of Library and Information Science, Indiana University, Bloomington, IN 47404, USA

Douglas Kurtze
Department of Physics, North Dakota State University, Fargo, ND 58102, USA

The work is part of an ongoing series of projects that examines the use of chaos theory in modeling time series data generated by computer mediated communication (CMC). The study regards the time series data generated from a CMC discussion group as the sum of two components. One is a deterministic "signal" which presumably obeys some unknown nonlinear dynamics. The other component is truly random "noise." The study's overall goal is to assess the relative importance of these two components, using techniques devised by Procaccia and Grassberger for studying chaos in time series data. Analysis of the time series indicates that approximately 70 - 80% of the variance in the data can be accounted for by deterministic chaos. Implications for future research using chaos and CMC are discussed

1. Introduction

Previous work in the field of bibliometrics has explored the utility of using nonlinear dynamics, or chaos theory to describe mathematically patterns in scholarly communication such as cocitation clustering [1], frequency of publication of journal articles [2], and postings to scholarly computer bulletin boards [3]. It remains unclear whether chaos analysis is a suitable methodology for investigating scholarly communication. Among the criticisms raised against chaos as a bibliometric tool are :

(a) Chaos analysis may be too sensitive to apply to data generated by behavior as disorderly as human communication. Chaos analysis is usually applied to systems that behave relatively simply (e.g., biological populations, crystal growth, etc.). There is a concern, however, that chaos analysis may indicate chaotic behavior is present when applied to human-generated data not because there is underlying deterministic chaos, but because the data are extremely complex.

(b) Even if scholarly communication exhibits chaotic behavior, it is unclear how much variance in the phenomena can be accounted for by deterministic chaos. Chaos may be a major factor in explaining scholarly communication or a minor factor overlaid with large amounts of noise. The presence of chaotic behavior does not indicate how strongly or completely it can explain the phenomenon in which it is found.

(c) Chaos may not offer any predictive power. If scholarly communication is governed by extremely complex chaotic processes it may not be possible to identify them with finite amounts of data or it may not be possible to extract the deterministic mechanism that describes the communication behavior. In either case there is no practical difference between random behavior and chaos.

The current study is part of an ongoing, incremental examination of these points within the context of computer mediated communication (CMC).

Earlier work by the authors in the field on CMC established that a time series of postings to an electronic bulletin board exhibited chaotic behavior[4]. In an effort to address concerns of

sensitivity in disorderly data, the authors subjected randomly generated data to chaos analysis and determined that randomly generated data with parameters similar to those of chaotic human-generated data did not behave chaotically[5]. The current research advances this line of inquiry by investigating the extent to which chaos can account for variance in CMC.

To explore the utility of chaos analysis as a bibliometric tool within the arena of CMC, it is necessary to determine how much variance in CMC time series data can be accounted for by an underlying chaotic mechanism. Since noise cannot be removed from the date, this was done by adding progressively larger amounts of noise and reanalyzing the time series until chaotic behavior was no longer exhibited. The details of this technique and the findings of the study are reported below.

2. Methodology

2.1 Sample Selection

Computer mediated discussion groups typically deal with a specific area of interest such as philosophy or computer-use. Participants in the discussion groups read and respond to messages which have been previously posted on the network; single messages frequently generate a number of other messages in reply.

An examination of the discussion groups available on the USENET system was made. There were three criteria for selection : 1) scholarly or technical discussion subject, 2) availability of discussion group archives (i.e., copies of all messages sent to the discussion group), and 3) continuous use of the group for at least 5 years. The discussion group which was finally selected was IBM-PC Digest. IBM-PC Digest deals with issues concerning IBM personal computer use. The group draws its participants almost exclusively from academic or research institutions, and has been in operation continuously since 1982. To the best of the authors' knowledge, this places it among the longest continuously-running groups on the USENET system. Complete archives were available for all traffic on the network, and copies of these were obtained for the period January, 1982-November, 1989.

All messages sent to the system have a standard heading which includes the date and time of transmission. A computer program was written which identified the date and counted the number of messages posted each day, including an extra day for leap years and days for which there were zero postings. The resulting data was in the form of a time-series of 2707 separate days, each paired with the total number of messages sent that day.

2.2 Investigating Deterministic Chaos in Time-Series Data

We regard the data as the sum of two components. One is a deterministic "signal" which presumably obeys some unknown nonlinear dynamics. This signal is itself chaotic, yet if the dynamics can be unravelled the signal should be predictable at least for short times. The other component is truly random "noise", which is not predictable at all, even in principle. Our overall goal is to assess the relative importance of these two components.

A time series obtained from a purely deterministic, nonlinear dynamical system can be " chaotic", resembling a purely random time series so closely that the eye cannot distinguish between the two. The two can be distinguished, however, by calculating their respective fractal dimensions. To see how this is done, suppose the dynamics produce each new datum x in the time series using only the information contained in the m-most recent data :

$$(1) \qquad x_{\{i\}} = f\left(x_{\{i-1\}}, x_{\{i-2\}}, \ldots, x_{\{i-m\}}\right)$$

Another way to say this involves arranging the data into m-tuples of consecutive data, of the form $x_{\{i\}} = f\left(x_{\{i-1\}}, x_{\{i-2\}}, \ldots, x_{\{i-m\}}\right)$. Then each m-tuple contains all the information needed to compute the next m-tuple :

$$(2) \quad \left(x_{\{i-1\}}, x_{\{i-2\}}, \ldots, x_{\{i-m\}}\right) \rightarrow \left(x_{\{i\}}, x_{\{i-1\}}, \ldots, x_{\{i-m+1\}}\right),$$

where the new datum $x_{\{i\}}$ is given by the equation (1) above.

Geometrically, then, the dynamics take each point in the m-dimensional space of m-tuples of consecutive data to another point in that space. Note that the m-tuples overlap; this is part of the definition of their dynamics. If the underlying

function f uses only the m most recent data (or fewer) to determine the next, then each m-tuple contains all the information needed to determine its successor. Chaotic deterministic dynamics characteristically concentrate the m-tuples on a subset of the full m-dimensional space, called an attractor. Typically, the attractor is a fractal, with a dimension d which is less than m, and which is not an integer. On the other hand, if the data are actually random (independent and identically distributed), then the m-tuples will be distributed randomly in their m-dimensional space, and will not concentrate on any lower-dimensional subspace.

It is thus possible to distinguish a purely deterministic, chaotic time series from a random one by calculating the fractal dimension of the figure formed by the m-tuples of data points. For random data, the dimension will be m, while for chaotic, deterministic data, it will be less than m. One small complication in this prescription is the fact that m, the number of data points which the putative nonlinear dynamics actually needs to calculate the next point in the time series, is unknown. This is actually an advantage, however. If we choose a value for m which is too high, then the nonlinear dynamics actually uses fewer than m values to compute the next datum, and so the dimension of the attractor is virtually unaffected. Random data will always give a dimension equal to m, whatever value of m we have chosen.

The standard method for calculating the fractal dimension of an attractor from a time series of points on the attractor is the correlation algorithm of Grassberger and Procaccia[6] Conceptually, this counts the average number of m-tuples which lie within a distance r of a given m-tuple. The volume of a fractal of dimension d lying within a distace r of a given point on the fractal is proportional to r^d, so we can estimate the dimension d by fitting the dependence of the counts on r by a power law. The calculation is actually done by computing the correlation sum.

$$(3) \qquad C(r,N) \approx \left(1/N^2\right) \text{ [number of pairs } (i,j) \text{ for which } | \, |v_i| - |v_j| \, | \leq r],$$

where N is the total number of m-tuples, then plotting log (C) as a function of log(r), and fitting a straight line through as large a section of the plot as possible. The slope of the line is then an estimate of d.

In a previous investigation, we carried out this program for the raw data, obtaining estimates of the dimension d of the attractor for various choices of the embedding dimension m. There are several technical difficulties with the procedure, mostly having to do with the fact that an N of 2707 is actually a very small sample for these techniques. Another difficulty arose because the data themselves are small integers. We used the l_∞ definition of the distance between two m-tuples, namely the absolute value of the greatest difference between corresponding elements of the two. Since the data are small, the greatest distance between any two m-tuples comes out to be 31, so that the distance r only ranges from 1 to 31, a range of $1\frac{1}{2}$ decades. As a result, the correlation plot was well approximated by a straight line over less than a decade of r. In order to increase the range of r, we used data for consecutive three-day periods instead of data for consecutive days. This increases the size of the data and hence the range of possible r values, at the cost of decreasing the size N of the data set ; the choice of three-day intervals, rather than, e.g., four-or two-day intervals, represents a compromise.

Until we have actually identified which part of the time series is random, it is clearly impossible to subtract off the random component and isolate the deterministic, chaotic part. What we can do, however, is to add random noise to the data. By adding successively larger random components to the data, we obtain new time series with larger random components than the original data. We then analyze these new time series to determine the effect of the added noise on the estimates of the fractal dimensioin of the attractor.

We have repeated our calculations applying the Grassberger-Procaccia algorithm to "spoiled" data sets obtained by adding varying amounts of random noise to the original data set. The random noise is obtained by generating a set of independent, gaussian distributed random numbers with mean zero and standard deviatiion equal to some prescribed multiple of the standard deviation of the original data, which is 10.26. For each spoiled data set we apply the correlation algorithm to estimate the apparent dimension of the attractor.

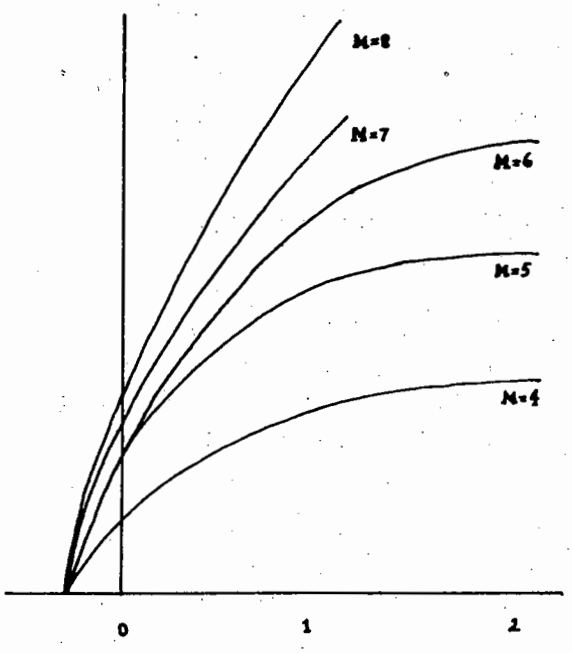


Figure 1. Plots od the estimated of the attractor dimension *vs.* strength of added noise for network data contaminated with Gaussian random noise. Plots are given for embedding dimensions *m* running from 4 through 8. The curves appear to intersect at a noise strength of -0.2 to -0.3 sigma, where sigma is the standard deviation of the original dataset.

3. Results

The graph of successive additions of noise versus changes in the calculated dimension of the attractor (*d*) can be seen in Figure 1. As noted earlier, the dimensions were calculated after adding sucesively larger amounts of noise, represented in the graph as fractions of the standard deviation of the original data (sigma). The dimension of the attractor can be seen to increase significantly for the first 0.2 sigmas of noise and flatten out and appproach *m* for large (> 1.0 sigma) values of noise.

Extrapolating backwards from the contaminated data (the equivalent of subtracting the noise), the curves intersect at a point in the range of -.2 to -.3 sigma. A variation of .1 sigma is probably the most accurate estimate of the amount of noise we can achieve using the graph techique. The curves become extremely steep (probably vertical) prior to intersecting, and it is generally difficult to get more accurate estimates from correlation plots.

What the results indicate, however, is that from twenty to thirty percent of the variation in the data is the result of noise. Conversely, seventy to eighty percent of the variation in the data can be accounted for by deterministic chaos. Even allowing for the higher estimate of noise, this indicates a strong deterministic signal. The strength of the signal is probably stronger than the experimental results exhibit since it is reasonable to assume that the time series data already contained some noise prior to "contamination".

4. Conclusion

Using deterministic chaos as a techique for describing time series data generated by CMC requires that a variety of concerns about chaos be addressed. These are the appropriateness of the technique for investigating human communication, the amount of variance the technique can account for, and the predictive power of deterministic chaos. Prior research by the authors has addressed the appropriateness of using chaos for investigating human communication; determining the variance accounted for is the aim of the current research.

The results indicate that a large part of the variance in CMC time series data may be described by a deterministic chaos mechanism. In particular, the relative importance of the deterministic signal is significantly higher than that of the noise component of the signal. It may still be problematic to use deterministic chaos for predicting the behavior of CMC systems, given the difficulty of extracting deterministic mechanisms from a time series. However, to the extent that the data are random, there is no predictive power. To the extent that data may be described as chaotic, there is an underlying order which has the possibility of being exploited.

The predictive power of systems which behave chaotically is generally limited to short periods. Beyond relatively short time horizons (e.g., 3-7 days for weather systems) the system becomes so dependent on the cumulative effects of small differences in its initial conditions that prediction becomes impractical. However, for many systems there is value in being able to predict behavior for comparatively short times. In the case of CMC, even several days of predictive power has utility in areas such as allocating bandwidth capacity for communication networks.

Current research in the field of deterministic chaos indicates that it may be possible to estimate deterministic mechanisms from time series data, thus making prediction possible for systems that exhibit chaotic behavior[7]. Future research to extract the deterministic mechanism from CMC time series data is planned.

References

1. Van Raan, A. (1991) Fractal Geometry of Information Space as Represented by Co-Citation Clustering, *Scientometrics*, 20 pp. 439-450.
2. Tabah, A. (1992) Nonlinear Dynamics and the Growth of Literature, *Information Processing and Management*, 28 pp. 61-74.
3. Kurtze, D., H. Snyder, G. Newby (1992) An Investigation of Chaos in Computer Mediated Communication, *Informetrics-91*, Sarada Ranganathan Endowment for Library Science, Bangalore, India, pp. 332-342.
4. *Ibid*, pp. 332-342.
5. Snyder, H., D. Kurtze, (1992) An Examination of the Utility of Non-linear Dynamic Techiniques for Analyzing Human Information Behaviors, *Proceedings of the 55th Annual Meeting of the American Society for Information Science*, Lear ned Information, Medford, NJ, pp. 98-100.
6. Grassberger, P. ,I. Procaccia (1983), Measuring the Strangeness of Strange Attractors, *Physica D*, 9 pp. 189-208.
7. Kaplan, D., L. Glass (1992) *Physical Review Letters*, 68 p. 427.

Detecting the Formation Orbit of Virtual Subject Knowledge Images

Bor-sheng Tsai
Library and Information Science Program, Wayne State University,
106 Kresge Library, Detroit, Michigan 48202, USA.

In a daily information professional's service activities, something is commonly and constantly encountered : what is pondered in a researcher's mind may not be always easily matched with what is recorded in virtual files. There always abides certain discrepancies or distances between the mostly predetermined *mind-sets* and usually stochastic *record-sets*. These discrepancies or distances urgently need to be efficiently and effectively balanced. Our two major foci, therefore, must be concentrated on the "mind-set" and the "record-set". To achieve this objective, a conceptual model for detecting the scholarly communications within a particular subject knowledge field was designed. The result of this detection reveals the formation of the orbit that regulates the virtual image projected by this special subject knowledge field. The significance of this approach is to aid subject information system analysts, designers, and managers : 1) see the relative distance between the two sets of models, namely, the conceptual model and the data model; 2) relocate optimal paths and layers for subject information researchers to reach their targets; 3) reconstruct a subject infosphere, and 4) study the ecological progression / regression of this subject infosphere.

1. Introduction

Instantaneously locating a path to an intelligence zone where problems will be resolved and questions answered has been the major endeavor of information scientists in multi-disciplines. One of the keys in finding the path is to understand the organization of the memory of a particular subject knowledge field. This key is in developing a knowledge road map which is the atlas of knowledge reality that can help navigate the information researchers while exploring a virtual subject knowledge space.

1.1 Conceptual Terminologies

The delineation of the following concept clusters is helpful in building a hypothetical model for this study. The conceptual model of *instruction "photon" movement* prepares a foundation for exploring and deploying animation-supported subject knowledge constellations. It links the Citation Analysis and Epidemic Model together and allows the detection of a special scientific communication network through a long term observation.

1.1.1 Instruction "Photon" Movement : Light Source / Spectrum / Polarization

This is the act / practice of directing / giving orders by way of the quantum of electro-magnetic energy to achieve a specific goal with strategic objectives or tactical maneuvers [1]. Such a practice will result in the display of a series of automatic insturctions, which can be broadly / narrowly redistributed / polarized (according to the frequencies of request / demand within the span of the subject knowledge field) through parallel processing, via multiple intermediaries / channels / media, or from mirror imaginary quadrants / coordinates.

1.1.2. Intelligence Zone : Object Frame Focus Loci

A set of specially pre-organized local focal points (local foci, or simply loci) which permit users to conduct post-coordination at will. Such loci sets could include text strings, pixels, highlighted keywords, field names, matrix cells, hypertext tilde (~) connections, mathematical for-

mulas, IF-THEN / DO-WHILE rules, animation frameworks, video slides, film strips, bitmap drawings, musical codes, virtual objects / files, voice / e-mail message segments, Internet telepaths, etc. These loci should sufficiently reflect the magnetic points of the centrality and locality of a particular subject domain. The conditions of centrality and locality will be the attributes that help coordinate the retrieval functions as well as measure the effectiveness of the operations conducted within the infosphere. (See also Section 1.1.7.)

1.1.3. Info / Geo-metrical Optics : Filter / ReferenceAngle / Convergence / Divergence

This is an information study work which applies geometrical optics (reflection and refraction) to detect the incoming parallel information sources, and the reference angles for discriminating and redistributing source information to the target destination.

1.1.4. Image Casting : Projection / Mirror Reflection / Blocking / Shadowing

The mirror image is a sketch / duplicate or direct reflection / indirect refraction of the original information source. By passing instruction 'photons' through a proper filtering device (which must always be controlled within a certain scanning / reference angle for loci), certain components of the original flavor / projection are *filtered, blocked or removed* with the application of virtual (not substantial) borrowing / combing / mirroring / migrating (or mutating / distorting) processes. The shadow of the mirror image is thus formed.

1.1.5. Subject-object Information Coordination: Cut, Bond (intersect / overlap / paste), and Reunion

Loci are identified as the **middle agents** which can **bond** together desired *instruction sources* (released and projected from the information objects) and *concept images* (filtered and reappeared on the subject knowledge screen) such that they may be optimally matched in greatest proportion.

1.1.6. Informapping : Animation-supported Subject Knowledge Constellations

By applying infomapping metrics and animated cartooning techniques, the image of a particular subject knowledge related scholarly communication network can be captured and represented in map forms and in time series for continuous scrutiny by subject researchers.

1.1.7. Infosphere : Virtual Subject Knowledge Image / Field / Space

The Infosphere is a mirror reflection of subject-object coordinated knowledge domain which is composed of influential information forces and potential drives. These forces and drives are constantly moulding a series of maps regarding scholarly communication and directing the interactions among human, machine, system, organization, and environment where values are defined by a scholarly or professional community or group with a special interest in mind. The properties, structures and relationships of the components within this infosphere can be studied by way of identification of loci, application of linear regresssion, measurement of reference angles, mapping of virtual subject knowledge images in layers, and detection of information orbit formation. (See also *Section 1.1.2.*)

2. Hypothetical Models

A great portion of this report is devoted to the techniques in constructing an animation-supported subject knowledge constellation which is the base for the detection of instruction photon movement (its convergence, intelligence, and divergence). Through periodical review on the transformation of constellations, and with threshold control for filtering or focusing elements in constellations, major constituents can be identified. These identified constituents make possible the determination of a group of loci. The continuity of the virtual subject knowledge images in animation series and the frequency of co-occurrence of loci enable the finding of an intelligence circuitry to define the boundary of image projection. These findings serve as bases for detecting the orbit formation of subject knowledge images.

In order to detect the orbit formation of these subject knowledge images, developed is a set of mathematical models, which are derived from geometrical optics and the ellipse, and are linked to the Epidemic Model. It helps detect the behavioral patterns of instruction photon movement, and even-

tually determines the relative "infosphere" of the investigated virtual subject knowledge image, namely, a mirror reflection of subject-object coordinated knowledge domain. The closing of distances between loci in an elliptical orbit reduces its eccentricity and thereby induces the possibility of forming near-circular orbits due to the concentration of foci. These hypothetical models are detailed in the following sections.

2.1. Epidemic Model (for diversity measurement)

Goffman's Epidemic Model can be generally described as : 1) $dY / dt = -B * X * Y + M$, and 2) $dX / dt = B * X * Y - R * X$, where X denotes the number of connected elements (infectives), and Y denotes the number of disconnected elements (susceptibles) of total population N at an arbitrary threshold T. "dt" denotes the time interval. B is the rate of effective contact (connection / infection). R is the rate of removal of elements from N. M is the rate of influx of new elements into N. For an epidemic to occur, as Goffman indicates, "it is necessary that $dX / dt > 0$, that is, that the rate of change of infectives with respect to time be positive. Whence, $Y > (R / B) = P$ constitutes a threshold condition, that is, for an epidemic outbreak to take place at time t, the susceptible population Y must exceed the threshold P at t". As the epidemic reached its highest peak, " the number of susceptibles remaining in the population N when the epidemic has spent itself is precisely at the threshold (P)." [2]

2.2. Infor / Geo-metrical Optics (for identifying focus, object and image)

Borrowing the laws of reflection and refraction from geometrical optics, information processing gains insight for better understanding and henceforth controlling image processing. For example, Ruelle's study points out that our visual system constantly performs visual data compression and processing beginning at the level of the retina and ending at the visual cortex with interpreted images as inside information representation of outside physical reality[3]. In a recent research, C.T. Liu has cited Chinese optics to support his explanation on *shadow formation* [4]. Graham and Sivin also focus their works on the ancient Chinese Mohist optics, and summarize eight optical propositions for *image formation* [5]. By applying optical image processing, a harmonic wave such as Y * cost (see *Section 2.3* : Elliptical Model) could be identified. This type of coherent wave (instruction photon movement) "contains both phase and amplitude information"[6]. As is commonly understood, an image can be located by controlling the relationships among light source, object, projector, lens, focus, image, and screen. The relationships are illlustrated in Figure 1.

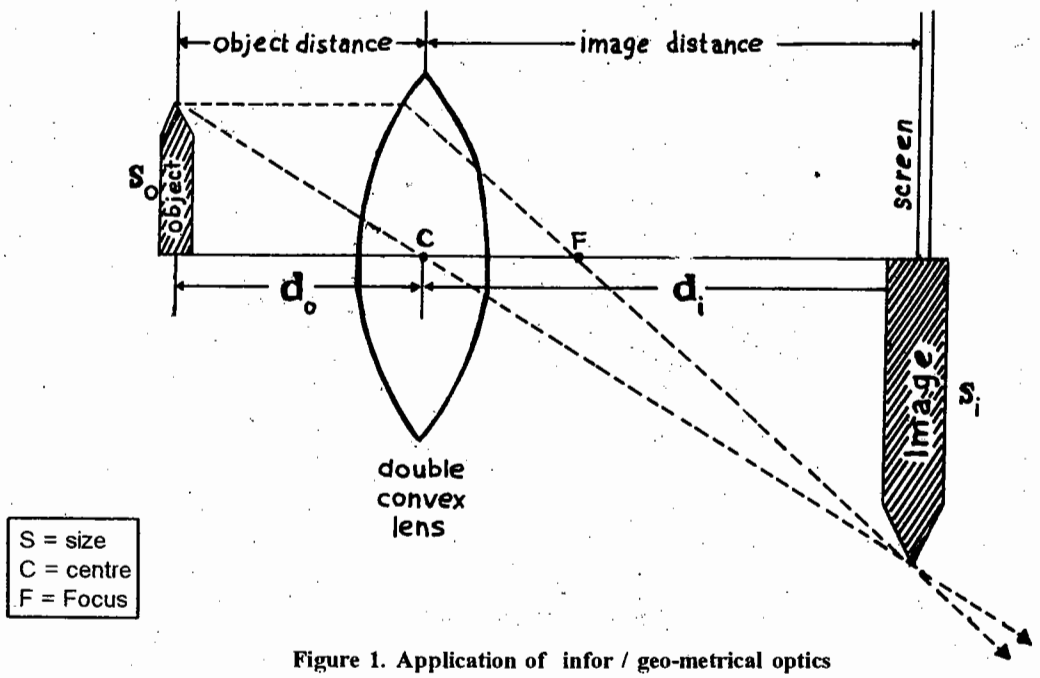


Figure 1. Application of infor / geo-metrical optics

The relationships can be represented by the following formulas :

1) $1 / f = 1 / d_o + 1 / d_i$

where d_o is the object distance,
d_i is the image distance, and
f is the focal length.
("/" means "divided by".)

Note : The above equation may be transformed into: $f = (d_o * d_i) / (d_o + d_i)$.

2) $s_i / s_o = d_i / d_o$

where s_i is the image size,
s_o is the object size,
d_i is the image distance, and
d_o is the object distance.
("/" means "divided by".)

3) $i = r$

(when REFLECTION is applied, similar to the state of equilibrium in the Epidemic Model : BXY=RX, which implies Y=R/B = P), where i denotes the measure-ment of incidence angle, and r denotes the measurement of reflection angle.

4) $P = R / B$

(when REFRACTION is integrated with the Epidemic Model : BXY > RX, which implies Y is > but ≈ P), where B denotes the rate of acceleration of instruction, R denotes the rate of deceleration of instruction, P is the index (threshold ratio) of refraction.

("/" means "divided by", and "≈" means "not equal to".)

Note : See *Section 2.1.* Epidemic Model on P = R / B.

2.3. Elliptical Model (for bi-focal approach)

By tracing all the points of an object that passed along an elliptical orbit, a constant can be found. This constant is a given distance which is the sum of each point's distances to two foci located on a directed axis within the orbit. To show this elliptical orbit, the graph is illustrated as follows.

The ellipse is represented mathematically as $(X^2 / a^2) + (Y^2 / b^2) = 1$, where $b^2 = a^2 - c^2$, with the two foci F (c,0) and F' (-c,0), and the constant N (=2*a) as the sum of the focal radii. ("/" means "divided by," and '*" means "multiplied by".) This equation can reinterpret the mathematical representation in the Epidemic Model dX / dt = B * X * Y -- R * X. That is, that dX / dt is similar to FA, OA (=BF) is similar to B * X * Y, and OF is similar to R * X. Moreover, OA = BF, FA = OA -- OF = (AA' - FF') / 2, OA = AA' / 2 = OF + FA, BF = SQR ($OF^2 + OB^2$) = SQR (OA^2) = OA = AA' / 2; and, OF = FF' / 2. Therefore, FA = OA (or BF) - OF, FA = (AA' - FF') / 2. This is very similar to (N - P) / 2 in the stage when dX / dt (similar to FA) is very close to zero (most diversified and out of focus), given that N = X + Y, and Y is very close to threshold value P (=R / B) (See *Section 2.1* Epidemic Model). Besides, the eccentricity may equate FF' / AA' = OF / OA = OF / BF = RX / BXY = R / BY = cosT, which implies that P = R / B = Y * cosT, where T is the degree measurement of the angle formed. ("SQR" means "squared root.") This Elliptical Model is further detailed in the next Section 2.4. (Infospherics).

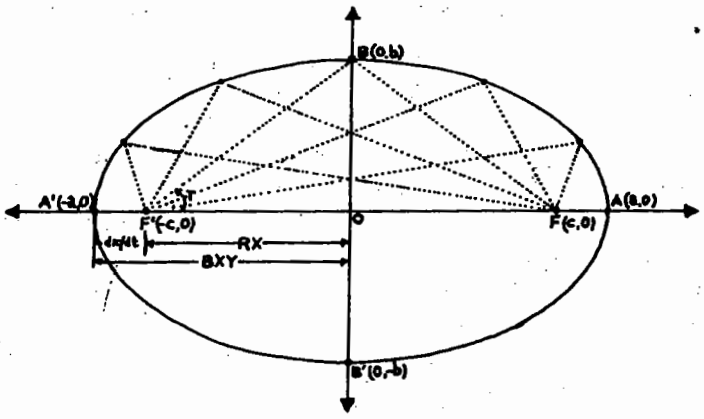


Figure 2 : Elliptical orbit

2.4. Infospherics :Recombination of Info / Geometrical Optics, Elliptical Model, and Epidemic Model (for building multiple virtual subject knowledge shells)

As defined in section *1.1.7.*, the "infosphere" concept and technique are developed to support subject researchers with a mirror reflection of subject-object coordinated knowledge image. Since the traditional subject-oriented approach is to deal with attributes of the object, it becomes more characteristic and time dependent. In the mean time, the current object-oriented approach is to meet the needs for efficient interoperation among multi-access points. Therefore, the organization of these objects must be more purpose / condition independent. The vague line between subject and object-oriented approaches has twisted many researchers' mind-sets who are basically subject-oriented when seeking subject-related objects. These objects are, at the present moment, mostly oriented by subject headings to reflect / refract their contents. In depicting the spatial logic and object representation / mapping, Jeansoulin indicates that for gathering natural spatial objects, the contents define (or reflect) the container (or boundary), while in the collection of cultural spatial objects, the contents are gathered within an artificial container [7]. Henceforth, the ideal automatic instruction process must provide an automatic clustering capability that enables any subject researcher to manipulate the interested source instructions (subject contents) at his / her own will and pace. However, the process still has to be conducted within the subject knowledge boundary. For these reasons, the preparation for easy access and identification and relocation of the desired loci are the most difficult and challenging tasks. Nevertheless, through the integration of the above three models, and with the help from citation data collection and analysis, it is possible to reconstruct an infosphere by identifying loci, tracing regressive lines, measuring reference angles, mapping virtual subject knowledge images in parallel, or in layers and detecting the property, structure, and frequency of communication behaviors which regulate the infospherical orbits.

Three states, which the formation of the subject infosphere may go through are : 1) Definite State, 2) Orbital State, and 3) Circular State. Similar stages were illlustrated by Prigogine and Stengers[8]. They are elaborated as follows.

2.4.1. Definite (reflective / refractive) State : Linearization / Simplification

As was known, the shortest (original) distance (effort / force) between two focal points A and A' is a tight straight line segment AA'. The Definite State represents this situation. There is no room (or no need / use of effort / force) for further stretching. The components (or traffics) in this communication channel are within this straight line. This state usually occurs in the very primitive phase of a developing system.

2.4.2. Orbital (Elliptical) State : polarization diversification

As time goes by, when the two focal points are brought closer (co-excited) to each other by a third influential point, given that the original maximal distance (2*a) has remained constant, the space which allows for communication may be stretched outward. This third point, which is directed by the original force, may be located outside the original line segment AA'. The continued maximal stretch of the controlled points will reach their "sphere" of influence. An infosphere is thus first formed in an elliptical shape. (See previous Figure 2 and Figure 3 below).

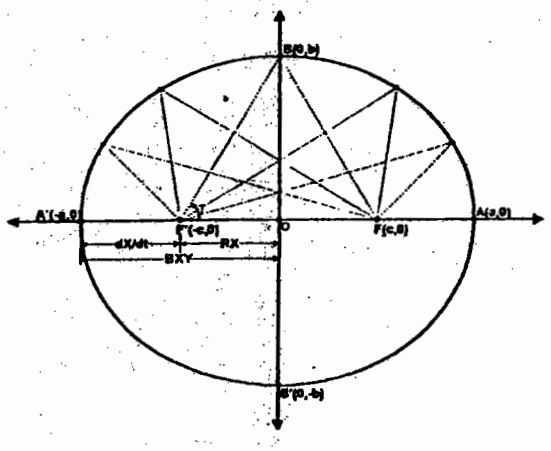


Figure 3 : Semi-elliptical orbit

2.4.3. Circular (infospherical) State Globalization / Unification

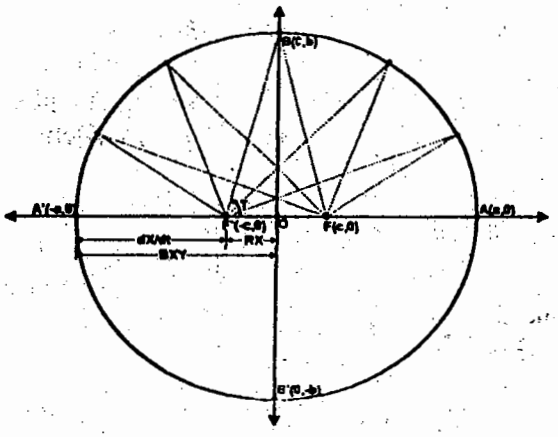


Figure 4 : Near-circular orbit

As the two focal points (F and F') are getting closer and closer, they eventually overlap into one single point O. The ellipse is thus transformed into a circle. At this moment in time, OF = OF' = 0, that is, FF' = 0, and FA = OA = FB = OB. There is no further need for any kind of discrimination and focus any more. As the process progresses, OF< (OA/2), that is, FF'< (AA' / 2), the foci (F and F') are departing from each other and away from the original (once uniquely overlapped focal) point O. The elliptical shape has resumed. If OF = FA = OA / 2, that is, FF' = AA' / 2, the foci must be the mid-points of OA and OA', respectively (See Figure 4). As time continues, OF > (OA / 2), that is, FF' > (AA' / 2), the foci are closer to the points A and A', and farther away from the original point O. At the time when OF = OA, that is, FF' = AA', and, F and A, and F' and A' are respectively overlapped to form segment AA', the foci may reach their farthest limitation. The ellipse is finally flattened down and returned to a line segment. The Definite State has come back to stage again. The states may go on and on and the information cycle toward re-discrimination and re-harmonization continues.

3.1. Citation-based and Cartoon-supported Info--Mapping Technique

By applying citation analysis and through a long period of observation, a set of loci (ninety-nine synthetic information elements-authors) dwelling in a particular subject knowledge domain (Nutrition and Dietetics) may be identified[9]. These loci were then plotted on the Machine-Readable Mapping Format (MARM) based on the information mapping technique developed by the author, and its eight key mapping components include: scale, loci, bondage, direction, frequency, relative distance, boundary, and threshold (See Figure 5) [10].

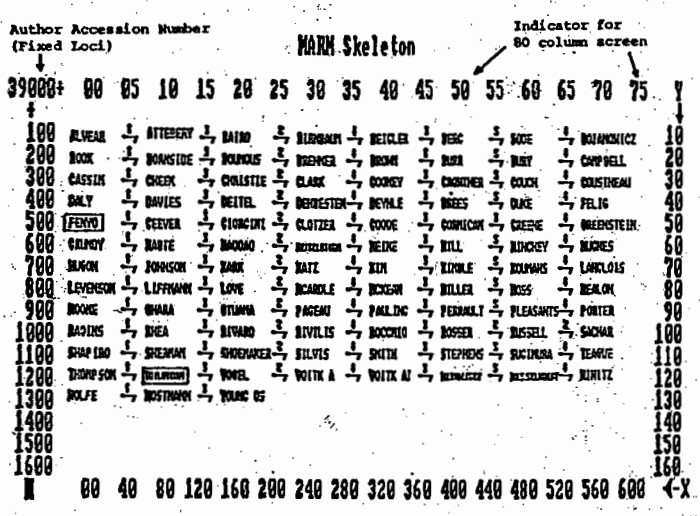


Figure 5 : 2-D machine readable mapping framework

As Drewery referred, animation might be used to assist adaptive motion which could occur " when the control processor uses information about the objects and their environment to control the objects' movements, " and "attributes of the objects can be updated and examined" to enhance adaptive motion[11]. The "animating space" concept may be injected here, which is "the process of panning and zooming around and into a large two-dimensional static image" before we can look at a dynamic map[12]. With the application of animated cartoons, the above mapping framework can be transformed into an animated format. (See Figure 6). The original machine readable information mapping technique is thus advanced to dynamic object-oriented presentation of maps in continuing series of frames that indicate periodical movements of those *focused* elements in a special subject field. "The adoption of an animation to navigaiton map-making is an innovative approach, which speeds up the process of drawing design, deposit of name string patterns, line connections, arrow directions, etc. The most efficient and effective contribution of an animation programming is the map continuum shown in chronological or topological order"[13].

3.2. Detecting the Orbit of Infosphere

A detection of the orbit formation of a particular subject infosphere must start from its early stages of development and follow up and focus on its periodical cycle thereafter. To conduct periodical observation and review, the hypothetical models in Section 2 are applied. The result is a collection of maps, which is the main source of clues for analyzing the behavioral and structural relationships among scholarly communications, and for drawing various layers of configurations, which may start from the very inner core elements to the outward boundary elements. Logistically, the detection of the orbit formation of the infosphere may follow the five Formation Periods : 1) Pre-formation, 2)In-Formation, 3) Trans-Formation, 4) Uni-Formation, and 5) Re- / Con- / De-Formation[14]. These five periods are similar to the "abstract place called 'phase space' " where the "attractors" dwell. To understand this space requires a map[15]. The following series of graphs may show the periodical transformation of the involved loci. (See Figure 7. and refer to section *2.4.* : Infospherics)

YEAR = T > 1							
A1	C1	F1	H1	L1	O1	S1	V1
A2	C2	F2	H2	L2	O2	S2	V2
	C3	F3	H3	L3		S3	V3
	C4		H4	L4		S4	
B1	C5		H5		P1	S5	
B2	C6	G1	H6		P2	S6	V1
B3	C7	G2	H7	M1	P3	S7	V2
B4	C8	G3	H8	M2	P4	S8	V3
B5	C9	G4		M3	P5		V4
B6		G5		M4			V5
B7		G6	J1			T1	V6
B8	D1	G7			R1	T2	
B9	D2	G8		N1	R2	T3	
B10	D3		K1	N2	R3		Y1
B11	D4		K2		R4		
B12	D5		K3		R5		
B13	D6		K4		R6		
	D7		K5		R7		

Figure 6 : Animation-based mapping framework

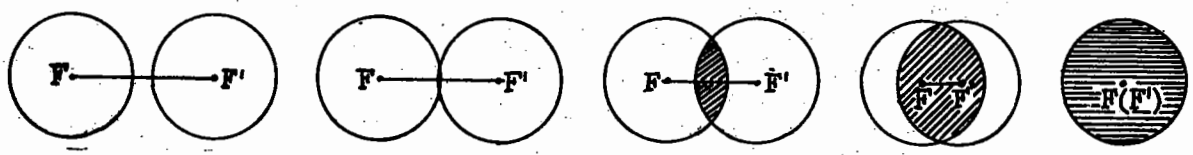


Figure 7 : Series of periodical transformation of loci

Table G. General data*

Year	DER	PAP	Cd PAP	Cg PAP	In Cd & Cg PAP	CA	AU	Cd AU	Cg AU	Cd & Cg AU	JL	Cd JL	Cg JL	Cd & Cg JL
1964	8	1	8	12	9	3	3	5	6	10	1	3	4	7
1965	13	2	10	21	12	8	8	5	13	17	2	3	10	12
1966	11	2	8	17	8	4	4	4	11	15	2	4	8	12
1967	17	2	5	11	6	7	7	2	11	13	2	4	10	14
1968	10	1	7	16	7	3	3	5	14	19	1	4	13	16
1969	59	5	15	37	14	11	11	8	25	33	5	7	23	27
1970	56	10	15	39	19	23	26	9	24	31	7	8	21	27
1971	99	21	32	114	41	43	56	16	45	56	16	17	28	36
1972	115	28	39	102	48	58	61	23	48	66	21	23	39	50
1973	161	37	56	162	74	92	99	37	71	103	26	30	56	68
1974	160	37	62	183	73	91	97	39	108	138	27	30	75	87
1975	166	51	105	317	-	100	108	72	123	193	27	51	83	110
1976	169	64	100	287	-	136	155	72	153	211	37	42	91	110

DER = Descriptors, PAP = Papers, Cd = Cited, Cg = Citing,
In = Inside**, CA = Coauthors, AU = Authors, JL = Journals.

* Data collected from the SCI®

** "Inside" refers to those papers, authors or journals, which appeared in the original database.

Table 1 : Dynamics of authors

Year	N	X	Y	dX	BXY	RX	B	R	P	H
1964	-	3	-	-	-	-	-	-	-	-
1965	-	8	0	5	5	0	-	.0000	-	-
1966	8	4	4	-4	1	5	.3083	.6250	2	2.9864
1967	8	7	1	3	7	4	.4063	1.0000	2	2.7694
1968	8	3	5	-4	3	7	.4405	1.0000	2	2.8649
1969	8	11	24	8	8	0	.5595	.0000	0	4.0000
1970	35	26	50	15	22	8	.0833	.7273	9	13.1345
1971	76	56	46	30	40	19	.0308	.7308	24	26.1364
1972	102	61	-	5	52	50	.0203	.8929	44	29.0074
1973	-	99	78	38	75	43	-	.7049	-	-
1974	177	97	12	-2	82	87	.0106	.8788	83	47.0472
1975	109	108	-	11	96	90	.0807	.9278	11	48.7517
1976	-	155	-	47	122	89	-	.8241	-	-

N : estimated total population
X : the number of connected elements (infectives)
Y : the number of disconnected elements (susceptibles)
dX : the rate of change of infectives (connected elements)
BXY : the number of incoming members
RX : the number of outgoing members
B : the rate of effective contact (connection / infection)
R : the rate of removal of elements from N
P : the changing threshold condition balanced by the rate R to the rate B (P = R/B)
H : the axis of symmetry of the represented parabolic curve from the Epidemic Model

4. Results

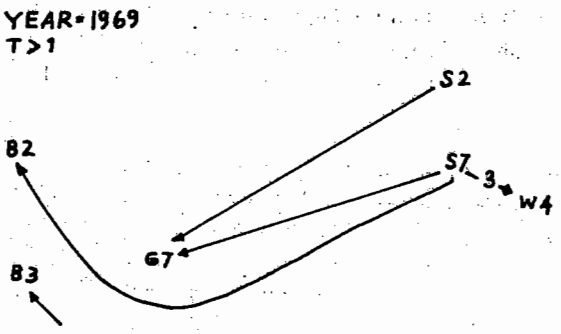
4.1. Measurement of Diversity Effects

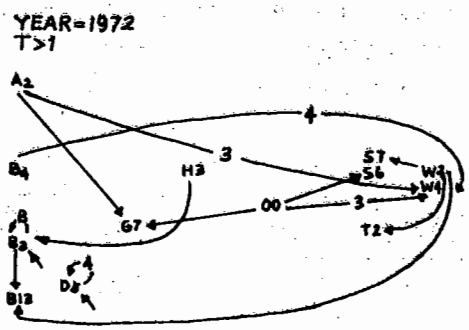
As Table 1 (dynamics of authors) shows, that several consecutive years observed demonstrated similar amount of dynamics. These years were clustered to represent the authors' collective contributions in four progressive periods: 1) 1964-69; 2) 1970-73; 3) 1974-75; and 4) after 1976. More evidences from the periodical reviews (provided in the next few sections) support this categorization.

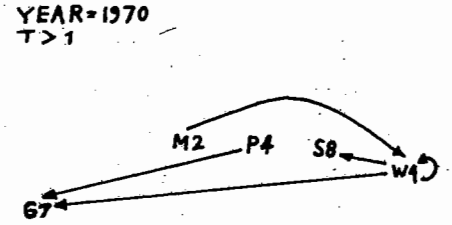
4.2. Identification of Objects, Foci and Images

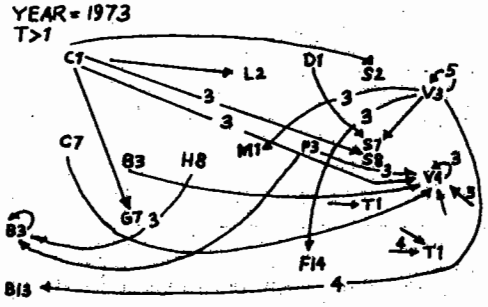
The following map series show that the authors in a particular discipline intercommunicate among themselves in a persistent way. To conduct a better observation on this phenomenon, we need to apply animation techniques. The results are shown in the following map series.

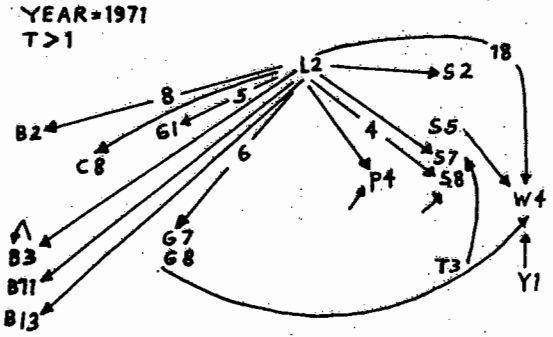
By animating these maps in chronological and / or topological sequences, those who / which had passed the pre-set filter and had shown their

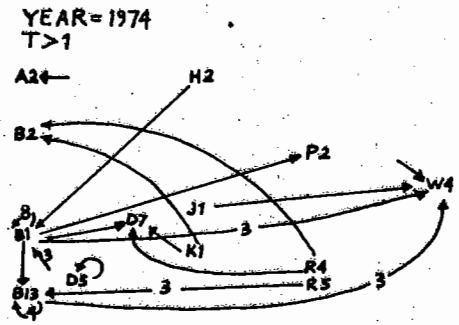












Map: Movement of synthetic information elements (Authors)

frequent co-occurrence in clusters, may be identified as : 1) the main **objects** (e.g., published documents) for "quasi object distance measurement" $\mathbf{d_o}$ (e.g., *cited documents*); 2) the main **foci f** (e.g., *cited authors*); and 3) the main **images** (e.g., contributed authors) for "quasi image distance measurement" $\mathbf{d_i}$ (e.g., *citing authors* or *citing documents*). As for the measurement of the quasi sizes of objects s_o and images s_i, they may be respectively represented by *the number of documents published annually*, and *the number of authors who contributed these documents annually*. Theconfigurations, which help the measurement and determination of quasi distances and sizes of objects and images, were the ingredients for testing the following formulas.

By plotting the data collected from the analyses of maps, the formulas : $\mathbf{1/f = 1/d_o + 1/d_i}$, and $\mathbf{s_i/s_o = d_i/d_o}$ were tested individually. Both results were shown in Table 2. As for $\mathbf{d_i}$, there were two sets of data used to reflect the number of citing authors or the number of citing documents. As the results showed, both sets corresponded well with the formula tested.

Based on the above experiments, we may reinterpret the above formulas as : 1) 1 / the total number of annual cited authors $\approx$ 1 / the total number of annual cited documents + 1 / the total

Table 2 : Infor. / geo-metric optical process using citation data

Year	1/f	≈	1/d_o	+	1/d_i		s_i	/	s_o	≈	d_i	/	d_o
1964	1/5		1/8		1/6	(or 1/12)	3		1		12		18
1965	1/5		1/10		1/13	(or 1/21)	8		2		21		10
1966	1/4		1/8		1/11	(or 1/17)	4		2		17		8
1967	1/2		1/5		1/11	(or 1/11)	7		2		11		5
1968	1/5		1/7		1/14	(or 1/16)	3		1		16		7
1969	1/8		1/15		1/25	(or 1/37)	11		5		37		15
1970	1/9		1/15		1/24	(or 1/39)	26		10		39		15
1971	1/16		1/32		1/45	(or 1/114)	56		21		114		32
1972	1/23		1/39		1/48	(or 1/102)	61		28		102		39
1973	1/37		1/56		1/71	(or 1/162)	99		37		162		56
1974	1/39		1/62		1/108	(or 1/183)	97		37		183		62
1975	1/72		1/105		1/123	(or 1/317)	108		51		317		105
1976	1/72		1/100		1/153	(or 1/287)	155		64		287		100

The above data can be translated into the following.

Year	1/f	≈	1/d_o	+	1/d_i		s_i / s_o	≈	d_i / d_o
1964	.2000		.2917			(or .2083)	3.0000		1.5000
1965	.2000		.1769			(or .1476)	4.0000		2.1000
1966	.2500		.2159			(or ..1838)	2.0000		2.1250
1967	.5000		.2909			(or .2909)	3.5000		2.2000
1968	.2000		.2143			(or .2054)	3.0000		2.2857
1969	.1250		.1067			(or .0937)	2.2000		2.4667
1970	.1111		.1084			(or .0923)	2.6000		2.6000
1971	.0625		.0535			(or .0401)	2.6667		3.5625
1972	.0435		.0465			(or .0354)	2.1786		2.6154
1973	.0270		.0319			(or .0241)	2.6757		2.8929
1974	.0256		.0254			(or .0216)	2.6216		2.9516
1975	.0139		.0176			(or .0127)	2.1176		3.0190
1976	.0139		.0165			(or .0135)	2.4219		2.8700

number of annual citing authors (or 1 / the total number of annual citing documents); and, 2) the ratio of the total number of annual contributed authors to the total number of annual published documents is close to the ratio of the total number of annual citing documents to the total number of annual cited documents (Note : "/" means "divided by," and "1/" means " the reciprocal of"). Since the above equation $1 / f \approx 1 / d_o + 1 / d_i$ may be transformed into : $f \approx (d_o * d_i) / (d_o + d_i)$, its interpretation above may be rephrased as : "the total number of annual cited authors ≈ [(the total number of annual cited documents * the total number of annual citing authors (or the total number of annual citing documents)] / [the total number of annual cited documents + the total number of annual citing authors (or the total number of annual citing documents)].

4.3. Bi-focal Approach

From Table 1 (Dynamics of authors), four growth patterns were distinguished. Further clues sought and gathered were : dynamics of journals, descriptors, and inside authors or inside journals citation networks. ("Inside" marks members of the citation networks focused on : citing authors or citing journals appeared in the original database, the authors-citing-authors, or journals-citing-journals) [See Table 3, 4, 5 and 6]. These clues indicate similar growth patterns which led to an establish-

Table 3 : Dynamics of journals

Year	N	X	Y	dX	BXY	RX	B	R	P	H
1964	-	1	-	-	-	-	-	-	-	-
1965	-	2	0	1	2	1	-	1.0000	-	-
1966	2	2	-	0	1	1	1.5000	.5000	0	.8333
1967	-	2	-	0	2	2	-	1.0000	-	-
1968	-	1	1	-1	1	2	-	1.0000	-	-
1969	2	5	1	4	5	1	4.5000	1.0000	0	.8889
1970	6	7	-	2	6	4	.9500	.8000	1	2.5789
1971	-	16	8	9	15	6	-	.8571	-	-
1972	24	21	21	5	16	11	.1270	.6875	5	9.2933
1973	42	26	95	5	17	12	.0381	.5714	15	13.5013
1974	121	27	-	1	20	19	.0081	.7308	90	15.3889
1975	-	27	-	0	21	21	-	.7778	-	-
1976	-	37	-	10	26	16	-	.5926	-	-

Table 4 : Dynamics of descriptors

Year	N	X	Y	dX	BXY	RX	B	R	P	H
1964	-	8	-	-	-	-	-	-	-	-
1965	-	13	4	5	8	3	-	.3750	-	-
1966	17	11	3	-2	6	8	.1077	.6154	6	5.64
1967	14	17	4	6	14	8	.4056	.7273	2	6.10
1968	21	10	8	-7	9	16	.1239	.9412	8	6.70
1969	18	59	7	49	52	3	.6672	.3000	0	8.78
1970	66	56	6	-3	38	41	.0930	.6949	7	29.26
1971	62	99	58	43	69	26	.1960	.4643	2	29.82
1972	157	115	-	16	70	54	.0122	.5455	45	56.14
1973	-	161	77	46	82	36	-	.3130	-	-
1974	238	160	22	-1	72	73	.0058	.4534	78	79.91
1975	182	166	-	6	75	69	.0215	.4313	20	80.97
1976	-	169	-	3	79	76	-	.4578	-	-

Table 5 : Dynamics of inside authors *

Year	N	X	Y	dX	BXY	RX	B	R	P	H
1964	-	2	-	-	-	-	-	-	-	-
1965	-	2	-	0	1	1	-	.5000	-	-
1966	-	0	-	-2	0	2	-	1.0000	-	-
1967	-	1	-	1	1	0	-	-	-	-
1968	-	3	-	2	3	1	-	1.0000	-	-
1969	7	6	4	3	6	3	.5000	1.0000	2	2.5000
1970	39	11	33	5	11	6	.0556	1.0000	18	10.5072
1971	17	15	6	4	11	7	.1667	.6364	4	6.5192
1972	-	22	-	7	16	9	-.0167	.6000	-	-
1973	127	28	105	6	22	16	.0095	.7273	77	25.2211
1974	50	26	22	-2	22	24	.0357	.8571	24	12.9958
1975	33	32	7	6	29	23	.1648	.8846	5	13.8161
1976	-	48	-	16	39	23	-.0172	.7188	-	-

* Citing authors appeared in the original database and the authors-citing-authors-network.

Table 6 : Dynamics of inside journals *

Year	N	X	Y	dX	BXY	RX	B	R	P	H
1964	-	1	-	-	-	-	-	-	-	-
1965	-	5	-	4	4	0	-	.0000	-	-
1966	5	3	0	-2	1	3	.9500	.6000	1	2.1842
1967	6	5	3	2	2	0	.2333	.0000	0	3.0000
1968	23	5	18	0	3	3	.0333	.6000	18	2.4910
1969	-	13	-	8	11	3	-	.6000	-	-
1970	16	11	3	-2	7	9	.2077	.6923	3	6.3334
1971	16	16	5	5	10	5	.1153	.4545	2	6.7736
1972	-	26	-	7	19	9	-.0557	.5625	-	-
1973	32	31	6	5	12	7	.0726	.2692	4	14.1468
1974	-	37	-	6	17	11	-.0174	.3548	-	-
1975	68	36	31	-1	17	18	.0148	.4865	33	17.5642
1976	73	43	37	7	17	10	.0126	.2778	22	25.6484

* Citing journals appeared in the original database and the journals-citing-journals-network.

ment of four chronological dissections. These dissections are the bases for the configuration of Stage-Period Table of Synthetic Information Elements (Authors).

4.4. Reconstruction of Multiple Virtual Subject Knowledge Shells

The continuous detection on the formation process of this subject knowledge field from multiple aspects such as authors, journals, descriptors, authors or journals citation networks, etc., [See Table 1 - 6] results in the configuration of a Stage-Period Table of Synthetic Information Elements (basically focused on Authors) (See Table 7 below). This Stage-Period Table may be used as the backbone for constructing a subject infosphere with multiple subject knowledge shells (such as authors, journals, documents, descriptors, citing authors / journals, etc.). (See Figure 8 : Subject infosphere)

5. Discussion

The advantages of this approach include : 1) providing information researchers with abilities for holistic view, local review, or historical overview on a particular subject knowledge field; 2)

Table 7. Stage-period table of synthetic information elements (authors)

P	1976	IVe					98,99		
	-	IVd	83	84-88	89	90-93	94	95-97	
	-	IVc		73				74-82	
	-	IVb						67-72	
	1974	IVa						64-66	
E	1974	IIIf	55	56-59	60,61	62,63			
R	-	IIIe		42,43	44-47	48-50	51,52	53,54	
	-	IIId		28,29	30-34	35-38	39-41		
I	-	IIIc		17-20	21	22,23	24-26	27	
	-	IIIb			15	16	11-14		
O	1969	IIIa				10			
	1969	IIc				9			
D	-	IIb				8			
	1966	IIa			6	7			
S	1965	Id			5				
	-	Ic		4					
	-	Ib	2	3					
	1956	Ia	1						
				75 Connected Elements					*24 Neutral Elements
				G R O U P S					

* 24 neutral elements are those independent and isolated authors who are totally disconnected from the citation networks.

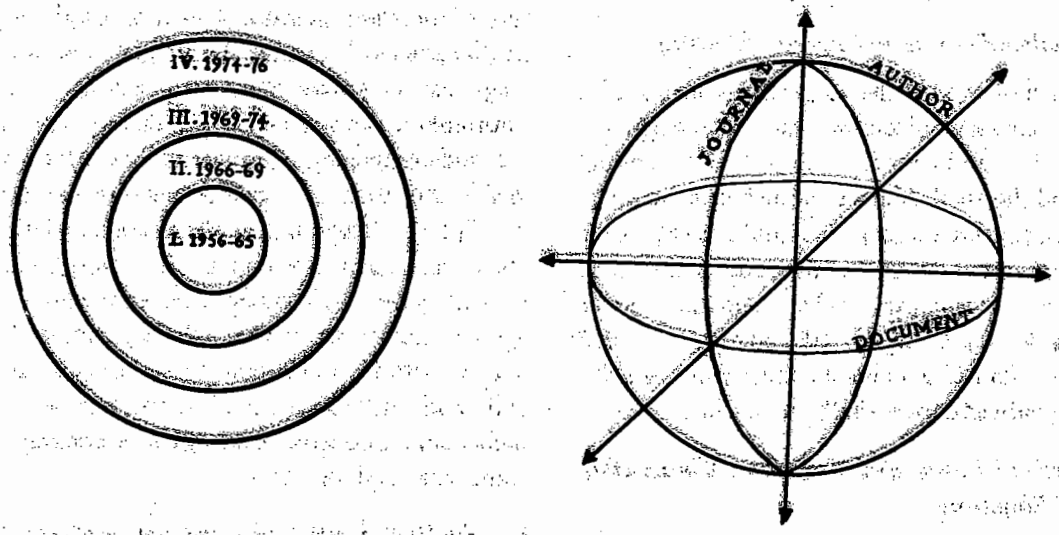


Figure 8 : Subject infosphere

updating network information on scholarly communications within a particular subject field; and 3) demonstrating a general communication model and the animation-supported computer programming techniques for constructing any special subject information navigation system. Some of the strong points found in this study deserve more attentions as follows.

5.1. Automatic Clustering and Categorization

One of the less harmonious areas in information studies is on the automatic clustering and categorization which seem arbitrary and hence less recognizable. Chaos science may be helpful in this case. As has been observed in Tabah and Saber's study on chaotic structure, one of the characteristics of "chaotic attractors" is " that successive points in the plot can diverge exponentially, yet stay within certain confines, namely the limit cycle. Furthermore, there is a class of attractors that delineate the confines of system behavior. These are called strange attractors"[16] .With the demonstrations of the above models and experiments, control over an ambiguous population within a particular scientific community can be exerted. This approach greatly enhances navigation programming, since major leaderships and the interrelationships among leaders-followers within the same scientific discipline were identifiable and traceable through applications of citation analyses, epidemic theory, info / geo-metrical optics, and practice of elliptical model.

5.2. Synchronization and Expert Shelling

As demonstrated, the major components of a specific scientific discipline can be recognized through the application of info / geo-metrical optics and elliptical model as published documents (s_o), cited documents (d_o), cited authors (**f**), citing authors or citing documents (d_i), and contributed authors (s_i), in regards to the formulas : **(time interval)** $1 / f \approx 1 / d_o + 1 / d_i$, and, $s_i / s_o \approx d_i / d_o$. These identified elements can be utilized for building subject expert shells in parallel.

5.3. Physical Laws and Cognitive Computing and Mapping

Balancing mind, energy and matter and showing its feasibility are among the endeavors of information studies, particularly in biological information processing with regards to molecular computing, cellular automaton, and neural net connectivity[17]. The theories in optics and ellipse we borrowed from physical science for this study have proved to be effective. The outcome of this study encourages tremendously the advanced study in cognitive computing. Nonetheless, any effective user interface system design must seriously and sufficiently elaborate on possible cyberspace to be involved in the human-computer interaction. We must recognize that the creation of a good techinical interface for cyberspace, or virtual cultural information, may require "a fusion of graphic artistry, programming, psychology ... "[18]. One of the immediate effects of utilizing infospherics is that it allows us to accurately construct a set of "fuzzy cognitive maps" that prescribe the scholarly communication relationships for a concerned scientific community. This series of fuzzy cognitive maps can reveal casual reasoning processes. In other words, "the fuzziness allows hazy degrees of causality between hazy causal objects (concepts)", and "their graph structure allows systematic causal propagation, in particular forward and backward chaining, and it allows knowledge bases to be grown by connecting different fuzzy cognitive maps."[19]. Van Raan and Tijssen have acknowledged the practical applications and epistemological values provided by *maps of science.* They treasure the maps of science not only as useful "descriptive tools" but also as "an interface between 'objective' literature-based data and 'subjective' accounts of experts". They regard bibliometric maps as "cognitive patterns resembling stored information in neural nets," which may represent" the self-organizing character of scientific activities in the form of a neural network-like structure."[20]. Currently, the information professions are constructing types of artificial worlds using virtual reality tools to represent information, images or other virtual effects for users to manipulate. Carefully conducting behavioural detection and quality filtration on information systems and their communication networks is becoming more significant and inevitable.

6. Further Study and Concluding Remarks

Further development and practice on this theory

and methods for information detection are discussed as follows.

6.1. Ecological Progression / Regression of Subject Infosphere

A subject infosphere is a new concept in bringing constructive information navigation forces to the information science professions. The methodology that infospherics provides may be advanced into the realm of "information ecology", where information scientists are concentrating on the progression or regression of the subject discipline and the objects this discipline are currently utilizing, as well as the amount of images projected within the subject infosphere. S.E. Palmer has similar proposal which states that the ratings of subjective goodness can be predicted by a model including a scaling constant, a weight (or salience) for each requested local orientation, and a maximal radius of symmetry for local or global orientation[21]. On the other hand, the neural net technology is particularly promising to allow us in capturing the intelligence from a "computer-based neural net paradigm"[22]. This task can be achieved by recognizing recursive patterns and using iterated function systems' frameworks for encoding and generating a large class of fractals or a real-time animation, taking into account of the large amount of computation and memory bytes consumption[23]. With the assistance of emerging information techonology, the concepts and techniques of infospherics can be utilized in assisting local area or online database searching, as well as in navigating Internet dial-up services. The observation on the networks' behaviors (telepaths) is another broad and urgent field where the infospheric approach can contribute.

References

1. Tsai. B. The development of theory and practice of cognitive computing-based automatic instruction. Paper contributed to the 1993 *International Symposium on the Development of Theory and Practice of Library and Information Science*, Wuhan, China, May 21-25, 1993.
2. Goffman, W.and K.S. Warren, *Scientific Information Systems and the Principle of Selectivity,* Praeger Publishers, New York, USA, 1980. 30-31.
3. Ruelle, D. *Chance and Chaos,* Princeton University Press, New Jersey, USA, 1991, 157.
4. Liu, C.T. The obstruction ("A_i") concept and the paradigm of contraposition in Chinese optics, *Thought and Words (Journal of the Humanities and Social Sciences)*, March, 1991, **30**(1) 123-144.
5. Graham, A.C. and N. Sivin, 1973, A systematic approach to the Mohist optics, In : *Chinese Science,* S. Nakayama, N. Sivin (Eds.). The MIT Press, Massachusetts USA, 105-152.
6. Eshaghian, M.M., S.H. Lee, and M.E. Shaaban, Parallel image computing with optical technology, In: *Parallel Architectures and Algorithms for Image Understanding,* V.K. Prasanna Kumar (Ed.), Academic Press, Inc., New York, USA, 1991, 29-56.
7. Jeansoulin, R. Basic tools for spatial logic, *Cognitiva 90*, Proceedings of the Third Cognitiva Symposium, Madrid, Spain, November 20-23, 1990, 329-336.
8. Prigogine, I. and I. Stengers, *Order Out of Chaos : Man's New Dialogue with Nature,* Bantam Books, New York, USA, 1984, 131-176.
9. Tsai, B. *The Behavioral and Structural Analysis of a Special Subject Literature,*, Ph. D. Dissertation, Case Western Reserve University, Cleveland, Ohio, USA, 1987, 251.
10. Tsai, B. Mapping metrics for subject-object coordinated field recognition and representation. Paper presented to the Third International Conference on Informetrics, Bangalore India, August 9-12, 1991. Paper published in the *Journal of Library Science with a Slant to Documentation*,**30,** September, 1993.
11. Drewery, K. and J. Tsotsos, Goal directed animation using English motion commands. *Proceedings of Graphics Interface and Vision Interface*, Vancouver, B. C,. Canada, May 26-30, 1986, 131-135.
12. Dorling, D. Stretching space and splicing time : from cartographic animation to interactive visualization. *Cartography and Geographic Information Systems,* **19**(4) 1992, 215-227.
13. Tsai, B. Formulas for constructing an animation-supported navigation system. Paper presented to the 1993 American Society for Information Science Mid-Year Meeting, Knoxville, Tennessee, USA, May 23-27, 1993. Published by the *Journal of Library and Information Science*, April, 1994.

14. See Ref. 9. 38-41, 93-94.
15. Briggs, J. and F.D. Peat, *Turbulent Mirror*, Harper & Row, Publishers, Inc., New York, USA, 1989, 31.
16. Tabah, A.N. and A.J. Saber, Chaotic structures in informetrics, In : *Informetrics 89 / 90*, L. Egghe, R. Rousseau (Eds.), Elsevier Science Publishers, Amsterdam, 1990, 281-289.
17. Hameroff, S.R. *Ultimate Computing : Biomolecular Consciousness and Nano Technology*, Elsevier Science Publishing Company, Inc., New York, USA, 1987, 26-53.
18. Romkey, J. Whither cyberspace ? *Journal of the American Society for Information Science*, September, 1991, 42(8), 618-620.
19. Kosko,B. Fuzzy cognitive maps, *International Journal of Man-Machine Studies*, 24 (1986) 65-75.
20. Van Raan, A.F.J. and, R.L.W. Tijssen, The neural net of neural network research : an exercise in bibliometric mapping, *Scientometrics*, 26(1) (1993) 169-192.
21. Palmer, S.E., Goodness, Gestalt, groups, and garner : local symmetry subgroups as a theory of figural goodness, In : *The Perception of Structure*, G.R. Lockhead, J.R. Pomerantz (Eds.), American Psychological Association, Washington, D.C. (USA), 1991, 37-38.
22. Kurzweil, R. Another formula for intelligence: the neural net paradigm, *Library Journal*, October 15, 1992, 47-48.
23. Stark, J. Iterated function systems as neural networks, *Neural Networks*, 4(5) (1991) 679-690.

A Model for the Optimum Allocation of Information Resources within A Library Network

Qiaoqiao Zhang
CAB International, Wallingford, Oxon, OX10 8DE, UK

This paper presents a mathematical model for optimising the allocation of information resources in a library network environment. Emphasis is placed on the major stages of a modelling process, including the identification and definition of performance criteria, and the formulation of objective functions. Analytical solutions of one objective function are illustrated. The model, as an objective tool, can help network managers determine optimum title coverage, the degree of title overlap between libraries, and the level of duplication of titles within libraries. This model is primarily intended for less developed countries like China but it is also applicable to the developed world.

1. Introduction

The immediate and most powerful impetus to alter the idea of local self-sufficiency in information resources may be the present recession, and inflation. With competing pressures for money, libraries are often categorised as a non-essential resource and find their budgets reduced. Adding to the dilemma, the so-called "Information Explosion" has not abated, and user demand for information only seems to increase. At the other end of the spectrum, it is becoming evident that the great storehouses of information resources, such as libraries, are decaying at an ever increasing rate.

Resource sharing is an effective solution to this dilemma. Logically, the first stage of resource sharing should lie with the acquisition of information resources.

1.1. Definitions and Concepts

In the broadest sense, *information resources* are collections of data, documents, databases or any kind of data, regardless of medium.

A *Library Network* comprises two or more libraries and / or other organisations engaged in a common pattern of information exchange, through communication, for some functional purpose. A network usually consists of a formal arrangement whereby materials, information, and services provided by a variety of libraries and / or other organisations are made available to all potential users (Kent, 1978)[7].

The following terminology has frequently been used to discuss the concepts of *Information Resource Allocation* :

(1) Co-operative Acquisition
(2) Co-operative (Co-ordinated) Collection Development
(3) Information Resource Allocation (Distribution)

The broad term *Co-operative Acquisition* refers to joint action in acquiring and utilising information resources. Beginning with the preliminary stage of selection, organisations come together in a network to purchase material, resulting in joint ownership and / or use (Atherton, 1977) [1].

Co-operative Collection Development may be defined as the process by which a group of libraries develop individual and / or joint collections of resources within an overall plan for the accomplishment of certain goals (Fiels, 1986) [6].

Optimum Information Resources Allocation is one of the goals of *Co-operative Collection Development.* In this paper, the terminology *Information Resource Allocation* is used.

Information Resource Allocation in a library network refers to the mechanisms by which dispersed and uncoordinated information resources are organised and redistributed in a planned way so that a unified network information resource system can be established to satisfy the information needs of every library of the network (Zhang, 1990) [16,].

1.2. Objectives of Optimum Information Resource Allocation

Optimum Information Resource Allocation in a network environment usually sets out to achieve the following goals :

(1) Developing a comprehensive and large network collection on a particular subject to satisfy various types of users in various locations.

(2) Providing all users with access to a wider range of materials and services, and increasing user request fill capacity at different levels.

(3) Reducing duplication of less-used resources, sharing the expense of expensive items and avoiding omission in acquisition.

(4) Allowing individual libraries increased specialisation in meeting the primary needs of local users.

Comprehensive and largest possible pooled collections of information resources permit a high degree of access by network users. Unnecessary duplication, a waste of limited budgets, is avoided between libraries and within a library.

When devising any information resource allocation programme it is important to ask the following questions : what is the largest collection size the network can possibly afford to acquire bearing in mind budget constraints? Which is more cost-effective, purchasing, or obtaining the less-used and expensive items via Inter-Library Loan (ILL) or Document Delivery (DD) from outside of the network (Khalil, 1993)[8] (Pitt & Kraft, 1974)[11]? Is it desirable and convenient for local users if budgets are spent on purchasing unique titles, resulting in too many ILL or DD transactions and delays in satisfying requests in the network? What is a reasonable queue length for popular items in local libraries?

2. Optimum Information Resource Allocation-Mathematical Modelling

2.1. Defining Performance Criteria

In any mathematical modelling process the objectives need to be quantified. Performance criteria usually reflect the objectives to be fulfilled and the problems to be solved. Therefore, the identification and definition of performance criteria is the first stage of modelling (Williams, 1976) [14] (Rouse, 1979) [13].

In reviewing the objectives of *Optimum Information Resource Allocation,* the following four performance criteria can be identified to describe how well allocation policies will operate :

(1) Network Material Accessibility.

(2) Cost-effectiveness of ILL or DD from outside of the Network.

(3) Geographical Accessibility.

(4) Local Availability.

2.1.1. Network Material Accessibility

This can be defined as the capability of a network to offer its information resources for use in a variety of ways, e.g. borrowing, retrieving and consulting.

The major concerns in this objective criterion is to determine 1) the optimum total network collection size, i.e. how many titles the network will acquire and store; 2) the optimum collection allocation among the popularity classes, i.e. how many titles in each popularity class a network should acquire and store; and 3) the optimum collection allocation among the libraries, i.e. how many items each library should acquire and store.

2.1.2. Cost-effectiveness of ILL or DD from outside the network

This can be defined as the relative costs of purchasing compared with borrowing, taking into account such facts as the purchase price, the real costs of borrowing, and probable frequency of use.

Due to the rapid growth of publications, inflation and tightened budgets, networks are unable to purchase all of the titles requested, making ILLs inevitable. For some less-used and expensive items,

it might be cheaper to rely on ILLs or DD than to purchase them. Thus, in a sense, outside ILLs or DD become a cost-effective option for the network. Such mechanisms as usage study, citation study and operation research etc. will be helpful in deciding what groups, or more specifically, what items fall into the category of 'not to be purchased' by the network (Brookes, 1970) [2] (Zhang, 1986)[5]. The network will therefore satisfy requests for these items from sources outside the network.

2.1.3. Geographical Accessibility

This can be expressed as the distance covered to satisfy an ILL request within the network. In other words, it can be defined as the ability to satisfy a request locally and, if inaccessible, via ILL within the network with the least possible delay. Here the *distance* does not necessarily mean actual physical distance. It is assumed that actual differences in distance between two libraries do not cause a significant difference in delivery time within a region. We can therefore ignore geographical distance and concentrate only on relay times.

If one assumes that no union catalogue is available in the system concerned, ILL then becomes an N-body transaction (Duggan, 1969) [4]. In such a type of transaction, relay times become a major factor, which is related to the number of available copies of the same title and relay routes etc. In an N-body transaction environment, increased numbers of copies of the same title in the network (excluding duplicates in the same library) result in higher fill rates and less relay times from the ILL requesters' point of view, and a greater degree of satisfaction and convenience for local users. Therefore, *Geographical Accessibility* can be used to work out the optimum degree of overlap between libraries. In other words, determining the degree of overlap for the different popularity classes of titles between libraries will bring about a cost-effective solution with reasonable relay times (Egghe & Rousseau, 1990).[5]

2.1.4. Local Availability

This can be defined as the probability of a user's request being satisfied by local stock.

A library circulation system is a queuing system. When several users require the same title at the same time and the number of requests exceeds the number of copies available, or the copies are already on loan, some users will have to wait. It is undesirable to keep users waiting for long periods and thus strategies are needed to reduce waiting times. One efficient strategy is to increase the number of copies of popular titles in local libraries. This will reduce the waiting times and improve the probability of satisfaction. Therefore, *Local Availability* can be used to work out an optimum solution regarding duplication rates for different popularity classes of titles within local libraries.

2.2. Formulation of the Objective Functions

An *Objective Function* can be formulated for each performance criterion.

We first assume that the network has J member libraries and there are I title classes to be acquired but L title classes to be borrowed via ILL or DD from outside of the network. The title classes are divided according to the number of requests, i.e. popularity. Therefore, there are corresponding I and L request classes. In the formulae, *Title* refers to unique titles whereas *Item* includes duplicated copies.

2.2.1. Network Material Accessibility

In mathematical terms, *Material Accessibility* can be defined as the total number of requests for titles acccessible in the network expressed as a percentage of the total number of requests on the titles universally available.

$$\frac{1}{Tu_t}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} I_{ij}U_{ij}\right)\times 100\% \qquad (1)$$

The formulation of this criterion takes account of the number of requests for different classes from the users, in other words, the popularity of the titles. It is obvious that the more popular a title is, the more it will be requested, and the heavier will be its use, and vice versa. Therefore, the more popular titles should have acquisition priority.

Here, we intend to maximise the objective in order to achieve an optimum solution. It is usually the case that due to budget constraints, libraries are not able to purchase as many titles as they

would wish. To guarantee a minimum level of satisfaction, both at the network level and local library level, and to keep a balance between the libraries, there should be some minimum title requirement (I_n) for the network and minimum item requirement (I^L_j) for each library. Again there should be minimum title requirement (I^L_i) and maximum title restriction (I^U_i) for each title class to keep a balance between classes. The problem can be formulated as follows :

$$Max\left\{\left[\frac{1}{Tu_t}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} I_{ij}u_{ij}\right)\right]\times 100\%\right\} \qquad (2)$$

s. t.

$$\left[\sum_{i=1}^{I}\sum_{j=1}^{J} I_{ij} - \left(\sum_{i=1}^{I}\sum_{j=1}^{J} I_{ij}\right)\times R\%\right] \geq I_n$$

(Total Title Constraints) (3)

$$I_i^L \leq \left[\sum_{j=1}^{J} I_{ij} - \left(\sum_{j=1}^{J} I_{ij}\right)\times R_i\%\right] \leq I_i^u$$

(Titles of Each Class) (4)

$$\sum_{i=1}^{I} I_{ij} \geq I_j^L$$

(Title of Each Library) (5)

$$\sum_{i=1}^{I} cI_{ij} \leq B_j$$

(Budget Constraints in Each Library) (6)

$$\sum_{j=1}^{J} cI_{ij} \leq B_i$$

(Budget Constraints in Each Class) (7)

$$I_{ij} \geq 0 \quad i=1, ..., I; j=1,..., J \qquad (8)$$

(see Appendix I for an explanation of the symbols used)

2.2.2. Cost-effectiveness of ILL or DD from outside of the Network

Assuming that the network will attempt to meet all the demands of its users, the budget will have to cover the following costs (Buckland, 1975) [3].

(1) the cost of purchasing titles (including overlapping and duplicate copies).
(2) ILL or DD costs for items that the network fails to purchase (due to budget restrictions).

What is the balance between purchasing titles by the network and ILL or DD from outside of the network that will minimise the sum of these costs without reducing the user satisfaction rate? The acquisition policy could be defined as purchasing as many titles as possible until reliance on ILL or DD becomes cheaper. Therefore, the *Cost-effectiveness of ILL or DD from outside the Network* can be formulated as the proportion of requests satisfied from outside of the network at a cost lower than that of the network purchasing these items itself :

$$\left(\frac{1}{Tu_t}\sum_{l=1}^{L} L_l u_l\right)\times 100\% \qquad (9)$$

However, this is subject to the following constraints :

$$\sum_{i=1}^{I} I_i C_{ti} + \sum_{l=1}^{L} L_l C_{tl} \leq C$$ (Total Costs for Purchasing & ILL or DD) (10)

and $C_{tl} \leq C_{ti}$ (ILL or DD Costs less than Purchase Costs for Specific Title) (11)

(See Appendix I for an explanation of the symbols used).

2.2.3. Geographical Accessibility & Local Availability

When formulating these two objective criteria, queuing theory is applied. Therefore, it is necessary to give a brief description of the queuing system.

Here, we only consider the case of complete balking. In such a case, every user, if he does not find the title available, gives up and does not try again. The fraction of the year an item is not on the shelf is given by the expression (Morse, 1969)[10] (Egghe & Rousseau)[5] :

$$R\,(1/\mu) \qquad (12)$$

As there are on the average λ users a year who want to borrow the item, there are λ (R/μ) who do

not find the item on the shelf and give up. Consequently, there are $\lambda - \lambda R/\mu$ who can and will borrow it. This number is by definition equal to R. This reasoning yields :

$$R = \lambda - \lambda (R / \mu) \quad (13)$$

Thus the probability that an item will not be available is

$$P_0 = \frac{\lambda}{\lambda + \mu} \quad (14)$$

and the probability of that an item is available is (Morse, 1969)[10] (Egghe & Rousseau, 1990)[5]

$$P_1 = 1 - P_0 = \frac{\mu}{\lambda + \mu} \quad (15)$$

Often, queuing systems are categorised using a notation of the form (x/y/z) : (u/v/w) where each of these six symbols is defined as follows :

1. x = inter-arrival time distribution
2. y = service time distribution
3. z = number of parallel servers
4. u = service discipline
5. v = maximum number of users allowed
6. w = size of user population (Kleinrock, 1975) [9].

For the following two formulations, we assume that class i requests enter the network at library j at an average rate of λ_{ij} (requests/unit time). Here we do not distinguish the local requests from ILL or DD requests. The average return rate of library j for class i is μ_{ij}. Here, $1/\mu_{ij}$ is the mean loan period. Assuming that the title is borrowed continuously, λ_{ij} / μ_{ij} is the mean number of users who can read a class i title at a given time. It is also assumed that at library j there are an average n copies of titles in class i.

Also, if we assume that users arrive randomly and the time at which the next user arrives is unrelated to the time when the previous user arrived, then x = M, where M stands for Markov and denotes arrivals that follow a Poisson Process. y = G for service times with a general distribution (or y = M for exponentially distributed services times), and v and w = ∞. The service discipline u = FCFS, i.e. first come-first served, is adopted.

2.2.3.1. Geographical Accessibility

First, we assume that the distance between the request library and resource library is d_{ij}. Before it is satisfied by library k, the request may have been relayed t times. We also assume that the request will be sent to the nearest library first, then relayed to the next library if the first requested library fails to satisfy the request. The relay direction is clockwise. Since difference in mailing times are not significant in a region, we assume that the distance between any two close (neighbouring) libraries is the same. Therefore, the number of times an item is relayed (t) becomes a major concern (Figure 1.).

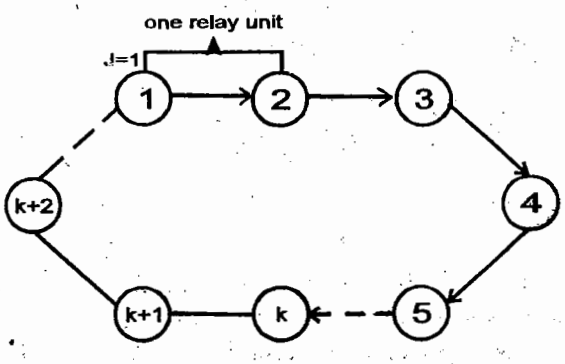


Figure 1. ILL Relay Route

There exist two cases in the N-body transaction : in the first case, when an i class request is made to library j, the library fails to satisfy the request due to either inaccessibility or unavailability, and the library relays the request to another library. By inaccessibility, we mean that the title requested is not collected by the library to which the request was made. By unavailability, we mean that the title is in the library's collection but it is on loan at the moment the request is made. The request is most likely to be satisfied immediately only when the probability of availability is high. In this case, we can only consider the probability of availability. Thus, the probability of satisfaction at the kth library is

$$P'_{ik} = \frac{\mu_{ik}}{\lambda_{ik} + \mu_{ik}} \quad (16)$$

and the probability of availability within the network thus is

$$P^{I}_{ik} = P_{i1} \times P_{i2} \times \ldots, P_{i,k-1} \times \left(\frac{\mu_{ik}}{\lambda_{ik} + \mu_{ik}} \right) \quad (17)$$

In the second case, we only consider the probability of accessibility. Once a class i title is accessible at library k, the request will eventually be satisfied since it joins the waiting queue. Therefore, we can ignore the probability of availability. In this case, the probability of accessibility at the kth library is

$$P_{ik}^{2} = P_{i1} \times P_{i2} \times \ldots, P_{i,k-1} \times P_{ik}'' \qquad (18)$$

If there are m copies of title i available in the network, the probability of accessibility at the first library is

$$m / (J - 1) \qquad (19)$$

at the kth library it is

$$P_{ijk}'' = \left(1 - \frac{m}{J-1}\right)\left(1 - \frac{m}{J-2}\right) \times \ldots . \times \left(1 - \frac{m}{J-k+2}\right) \times \frac{m}{J-k+1} \qquad (20)$$

For convenience, we assume that the j th library starts a class i request with probability, P'_{ij}. Here, we only take the second case into account, i.e. only considering the probability of accessibility. We further assume that the resource distribution is uniform. By uniform, we mean a distribution where all J libraries have an equal value of P'_{ij}, i.e. m/J, and the probability of accessibility at the kth library is P^{*}_{ik}.

With a uniform resource distribution, the appropriate policy is to send a request to the nearest library, and if not satisfied there, to refer it to the next nearest library, and so on until the request is finally satisfied (Rouse, 1976). It is obvious that the probability of satisfaction increases as the number of copies of the same title in class i within the network increases, i.e. more libraries own the title. Here, multiple copies owned by the same library have been assumed to have the same effect as does one, i.e. same probability for the library to satisfy the request. Therefore, only one copy of the title will be counted for each library.

Based on the above assumptions, therefore, the *Expected Distance* for a request on a class i title to be satisfied can be expressed as

$$D_i = \frac{\sum_{j=1}^{J}\left[\sum_{k=1}^{k} d_{jk} P_{ik}''\right] P_{ij}' \lambda_{ij}}{\sum_{j=1}^{J} \lambda_{ij}} \qquad (21)$$

The *Expected Distance* to satisfy a request for any classes of titles is given by

$$D = \frac{\sum_{i=1}^{I} D_i \lambda_i}{\sum_{i=1}^{I} \lambda_i} \qquad (22)$$

Minimisation of the *Expected Distance* can be achieved by integer programming (Egghe & Rousseau, 1990)[5] (Appendix II).

2.2.3.2. Local Availability

This is defined as the probability of a user's request being satisfied by local stock. To avoid complications, an opposite measure, *Probability of Dissatisfaction,* is used when formulating a queuing system.

According to Morse's formula (Morse, 1969)[10] (Egghe & Rousseau, 1990)[5] we can assume that when one copy of a class i title is circulated at library j , the copy is not on the shelf (not available) for a part of the year (R_{ij} / μ_{ij}), where R_{ij} is the circulation rate per year and $1 / \mu_{ij}$ is the mean circulation time (the loan period). During this time users have come looking for the title and, having found it missing, give up and leave. The probability of dissatisfaction is thus

$$P_{ij} = \frac{\lambda_{ij}}{\lambda_{ij} + \mu_{ij}} \qquad (23)$$

What happens when a second copy of a class i title is made available at library j for circulation? Since arrivals of prospective users are at random and since the two copies circulate independently of each other, there will still be times when both copies are out of the library and some potential users will still be dissappointed. The probability of dissatisfaction for a class i title at library j when two copies circulate is :

$$P_{ij}^2 = \frac{\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^2}{2\left[1+\frac{\lambda_{ij}}{\mu_{ij}}+\frac{1}{2}\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^2\right]} \qquad (24)$$

What happens if the average number of copies for each title of class i is n at library j ? We can infer the probability of dissatisfaction from the above formula and it will be (Morse, 1969)[10]

$$P_{ij}^n = \frac{\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^n}{n!(1+E_n)} \qquad (25)$$

where

$$E_n = \sum_{n=1}^{N} \frac{1}{n!}\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^n \qquad (26)$$

And the average *Probability of Dissatisfaction* for each title in class i in the network when an average of n copies for each library are circulated is :

$$P_j^n = \frac{1}{\lambda_i}\sum_{j=1}^{J} P_{ij}^n \lambda_{ij} \qquad (27)$$

An integer programming model can be formulated to minimise the *Probability of Dissatisfaction* (Appendix II) (Egghe & Rousseau, 1990)[5].

3. Discussion and Conclusion

3.1. Relationship of the Performance Criteria

The four criteria mentioned above can be categorised as network *information resource allocation* effectiveness measures since they can be measured in terms of how well the allocation policies will operate and satisfy the demand placed by the network users.

On reviewing the four performance criteria, the first is concerned with acquiring more unique titles; the second with purchasing or borrowing less—used and expensive items via outside ILLs or DDs, and the last two with acquiring more copies. Each of these criteria has its own cost-effectiveness thus they compete with one another for the network budgets. The overall cost-effectiveness of an *Information Resource Allocation* programme can be expressed as the relationship between the level of overall performance (effectiveness) and the costs involved in achieving this level. Figure 2. shows the relationship. Several alternative methods may be used to obtain a particular performance level. Usually, the least expensive alternative is the most cost-effective.

3.2. Analytical Solutions and Trade-off

The individual objective functions, as mathematical expressions of individual performance (effectiveness) criteria, can be a useful objective tool for cost analysis and trade-off. The above objective functions can serve two main purposes:

(1) Helping to determine the cost requirements of each objective (activity) by setting desired and reasonable goals (objective measures), thus achieving analytical solutions.

(2) Then taking the allocation results (after cost-effective-benefit analysis and trade-off between cost requirement and budget level) to achieve optimum solutions to these effectiveness measures (Zhang, 1990)[17].

It is usually the case that different analytical solutions can be obtained by making different assumptions. We can choose the solution which reflects the desired objective measures.

Here, we can give an illustrative example of analytical solutions of *Geographical Accessibility*. When formulating this performance criteria, we assumed that the distance for class i request from library j being satisfied at library k is d_{jk}. As the distance between two close libraries is one unit of relay times, d_{jk} is actually number of relay times between library j and library k. Again, since we assumed that the resources distribution is an uniform distribution this implies that each library has

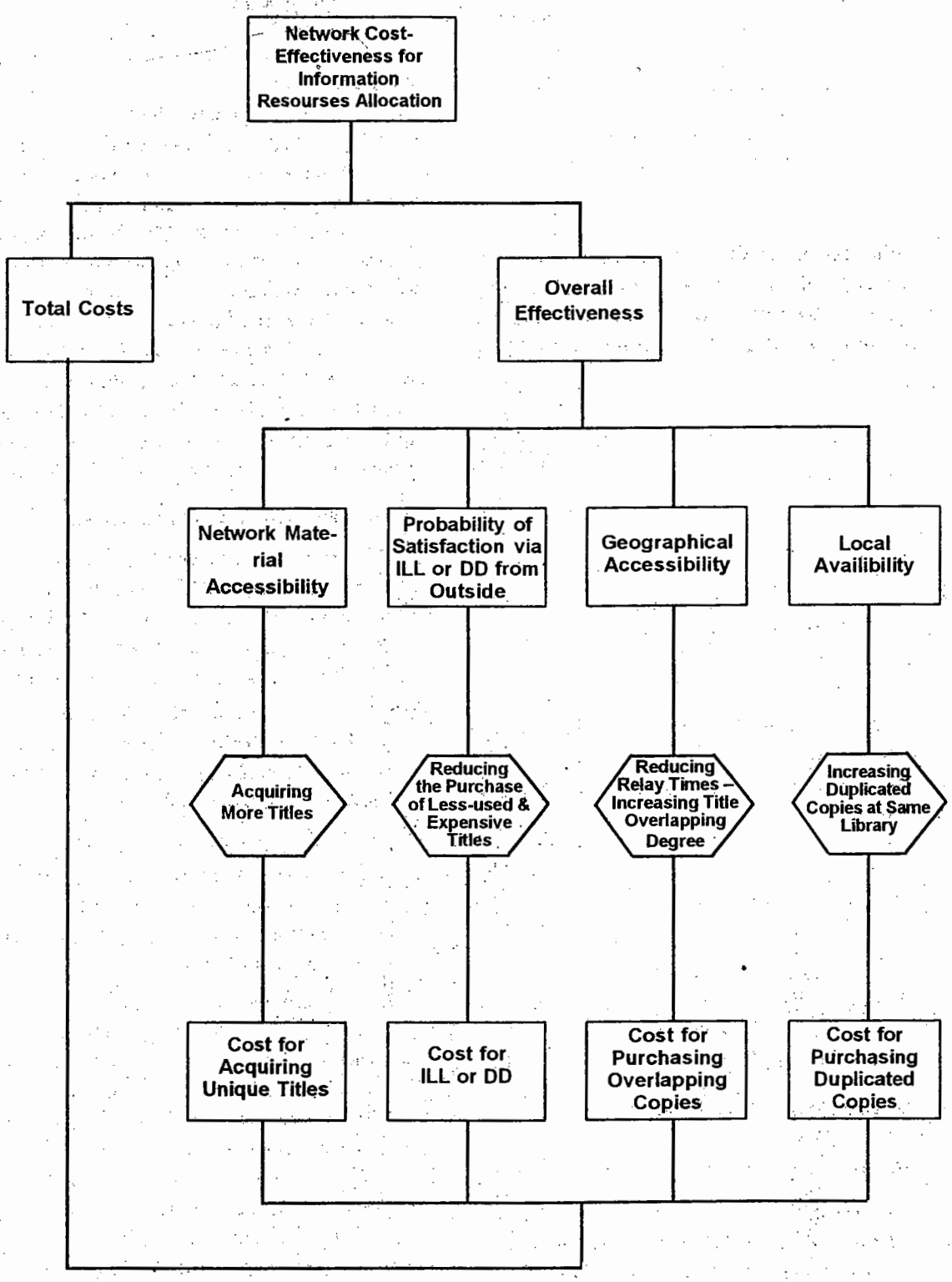


Figure 2. Relationship of four performance criteria

equal probability of initialising a class *i* request, which is related to the number of participating libraries and number of copies of a particular title. Therefore, the products of d_{jk} and P''_{ik} and the sum of products are important parameters of the *Expected Distance* (see Appendix III). Here, we assume that the number of participating libraries is 28. For the purpose of illustration and analysis, the number of copies (*m*) of a particular title given are increased by five each time. Multiplying the probability, P''_{ik} by the corresponding relay times first, and then summing them up, we thus obtain average relay times under different acquisition decisions. It can be seen clearly from the table that at a low basis, i.e. number of copies increased from one to five or from five to ten, the average relay times reduce dramatically. While at a higher basis, i.e. number of copies increased from twenty to twenty-five, the reduction of the average relay times is not so distinguished.

A trade-off consideration of costs, number of request (Popularity Class) and average relay times is a process of cost-effective analysis. By multiplying number of request with the average relay times, we can decide desired or reasonable targets for different popularity classes. By referring back to the corresponding overlap degree, we then obtain the desired number of copies (see Appendix IV). For example, if we decide that the desired product of average relay times and number of request should be around 5.00 plus or minus, we then can reach a decision of 25 copies for the popularity class 4.5 plus (i.e. average number of request per library), 20 copies for class 3.5-4.5 and 5 copies for class 0.5-1.5 etc. Multiplying number of copies by number of titles in a particular popularity class, we then obtain the total number of items in this class.

We then can consider the costs associated with different acquisition decisions. The cost involved in this performance criteria can be divided into two main types, i.e. cost of purchasing and cost of processing of interlibrary loan requests (mainly relaying). It is obvious that the cost of purchasing increases but the processing cost decreases as the number of copies increases. Finally, we can work out the total cost for this class. However, the cost is subject to the budget constraints. The decision can be altered according to the budget constraints.

3.3. Conclusion

Although what has been proposed is a purely theoretical model, it has the potential to be used for practical purposes.

The four mathematical expressions can indicate very useful trade-offs, allowing network administrators to work out a reasonable allocation programme with :

(1) optimum network collection size;

(2) optimum rules for purchasing or borrowing;

(3) optimum allocation of limited funds between collection size and duplicate copies both among libraries and within libraries; and

(4) optimum allocation of collection among network libraries and popularity classes.

Acknowledgements

I would like to thank Professor S.E. Robertson for his patient supervision and helpful guidance. Also the late Professor B.C. Brookes for his valuable advice and encouragement through my PhD research, from which this paper has been derived. I would also like to thank Professor R. Rousseau and Professor H. Eto for their encouragement and useful comments during the preparation and revision of this paper.

Appendix I - Components of Formulae

I is the number of title classes to be purchased

L is the number of title classes not to be purchased — can be borrowed more cheaply via ILL or DD from outside of the network

J is the number of participating libraries in the network

T is the total number of titles universally available

u_t is the average number of requests for titles universally available

u_i is the average number of requests for titles which are not going to be purchased

l_{ij} is the number of titles in class i to be allocated to library j

u_{ij} is the average number of requests for titles in class i from library j

$\sum_{i=1}^{I}\sum_{j=1}^{J} I_{ij}$ is the total number of items in the network, including overlaps between libraries

$\left(\sum_{j=1}^{J} I_{ij}\right) \times R_i\%$ is the number of overlapped items in class i

$\left(\sum_{i=1}^{I}\sum_{j=1}^{J} I_{ij}\right) \times R\%$ is the total number of overlapped items in each class in the network

$\sum_{i=1}^{I} I_{ij}$ is the number of items library j is going to be allocated

is the number of titles which are going to be acquired in class i

R% is the average overlap rate of any classes of items between libraries

I_n^L is the minimum title requirement for the network, i. e., title constraint for network

$R_i\%$ is the average overlap rate of class i between Libraries

I_i^L is the lower-level title constraint in class i

I_i^U is the upper-level title constraint in class i

I_j^L is the minimum title requirement in library j

$\sum_{l=1}^{L} L_1$ is the number of titles not to be purchased —can be borrowed more cheaply via ILL

c is the average cost of each title irrespective of classes

B_j is the budget constraint of library j for pubchasing titles

B_i is the budget constraint for each class of titles

$\sum_{i=1}^{I} I_i C_{ti}$ is the total costs for purchasing titles

$\sum_{l=1}^{L} L_1 C_{t1}$ is the total costs of ILL or DD from outside of the network

C_{ti} is the unit cost of purchasing one item in class i

C_{tl} is the unit cost of ILL or DD per item

C is the total cost of purchasing and ILL or DD charges

λ_i is the number of requests for titles in class i in the network

λ_{ij} is the number of requests for titles in class i at library j

λ_{ik} is the number of requests for class i titles being satisfied at library K

$\sum_{j=1}^{J} \lambda_{ij}$ is the sum of i class requests at library j

$\sum_{i=1}^{I} \lambda_i$ is the sum of requests for I classes of titles in the network

μ_{ij} is the average return rate in library j for class i titles

μ_{ik} is the average return rate in library k for class i titles

n is the average number of copies of class i titles

$K_{,}$ is the number of libraties to which the request has been relayed before it is satisfied (at kth library)

$P^{1}_{i\,k}$ is the probability of a class i title requested by library j being available within the network (ie, satisfied at the kth library)

P_{ik}^{2} is the probability of class i title requested by library j being accessible within the network (ie. satisfied at the kth library)

P_{ik}'' is the probability of class i requests by library j being accessible in the kth library

d_{jk} is the distance between the request library j and resource library k

P'_{ij} is the probability of class i requests being initiated at library j

P'_{ik} is the probability of class i requests being initiated at library j being satisfied by library k

D_i is the expected distance for a request on class i titles to be satisfied

D is the expected distance for a request on any classes of titles to be satisfied

P_{ij}^{1} is the probability of dissatisfaction for class i titles at library j when a single copy circulates

P_{ij}^{2} is the probability of dissatisfaction for class i titles at library j when two duplicated copies circulate

P_{ij}^{n} is the probability of dissatisfaction for class i at library j when n duplicated copies circulate

Appendix II: Minimization of *Expected Distance and Probability of Dissatisfaction*

Minimization of Expected Distance

When one copy of a title in any class is circulated in the network, the *Expected Distance* is D^1, two copies D^2 and m copies D^m.

The minimization of the *Expected Distance* to satisfy a request of any class can be expressed as:

$$\text{Min}\{D^1x_1 + D^2x_2 + \ldots + D^m x_m\} \qquad (29)$$

s. t.

$$x_1 + x_2 + \ldots + x_m = 1$$
$$x_k = \{0,1\}$$
$$k = 1,2,3,\ldots,m$$
$$C^1x_1 + C^2x_2 + \ldots C^m x_m \le B$$

where

C^1 is the unit cost of one copy of a title in any class

C^2 is the cost of two copies of a title in any class

C^m is the cost of m copies of a title in any class

$x_1, \ldots, x_m$ are zero-one integer variables

Minimization of Probability of Dissatisfaction

The minimization of the *Probability of Dissatisfaction* can be expressed as:

$$\text{Min}\left\{P_i^1x_1 + P_i^2x_2 + \ldots + P_i^n x_n\right\} \qquad (30)$$

s. t.

$$x_1 + x_2 + \ldots + x_n = 1$$
$$x_k = \{0,1\}$$
$$k = 1,2,3,\ldots,n$$
$$C_i^1x_1 + C_i^2x_2 + \ldots C_i^n x_n \le B_i$$

where

P_i^1 is the average *Probability of Dissatisfaction* when one copy of each title in class i is circulated

P_i^2 is the average *Probability of Dissatisfaction* when two copies of each title in class i are circulated

P_i^n is the average *Probability of Dissatisfaction* when n copies of each title in class i are circulated

C_i^1 is the unit cost of one copy of any title in class i

C_i^2 is the cost of two copies of any title in class i

C_i^n is the cost of n copies of any title in class i

$x_1, \ldots, x_n$ are zero-one integer variable

Appendix III – Analytical solutions for geographical accessibility

J=1

K	d_{ij}	when m=1		when m=5		when m=10	
		P''_{ik}	$P''_{ik} \times d_{ij}$	P''_{ik}	$P''_{ik} \times d_{ij}$	P''_{ik}	$P''_{ik} \times d_{ij}$
1	0						
2	1	0.037	0.037	0.185	0.185	0.370	0.370
3	2	0.037	0.074	0.157	0.304	0.242	0.484
4	3	0.037	0.111	0.132	0.396	0.158	0.474
5	4	0.037	0.148	0.110	0.440	0.096	0.384
6	5	0.037	0.185	0.090	0.450	0.058	0.290
7	6	0.037	0.222	0.074	0.444	0.035	0.210
8	7	0.037	0.259	0.060	0.420	0.020	0.140
9	8	0.037	0.296	0.048	0.384	0.011	0.088
10	9	0.037	0.333	0.038	0.342	0.005	0.045
11	10	0.037	0.370	0.029	0.319	0.003	0.030
12	11	0.037	0.407	0.023	0.276	0.001	0.011
13	12	0.037	0.444	0.017	0.221	0.001	0.012
14	13	0.037	0.481	0.012	0.156		
15	14	0.037	0.518	0.009	0.126		
16	15	0.037	0.555	0.006	0.090		
17	16	0.037	0.592	0.004	0.064		
18	17	0.037	0.629	0.003	0.051		
19	18	0.037	0.666	0.002	0.036		
20	19	0.037	0.703	0.001	0.019		
21	20	0.037	0.740				
22	21	0.037	0.777				
23	22	0.037	0.814				
24	23	0.037	0.851				
25	24	0.037	0.888				
26	25	0.037	0.925				
27	26	0.037	0.962				
28	27	0.037	0.999				
Average Relay Times			13.986		4.723		2.538

J=1 (cont.)

K	d_{ij}	when m=15		when m=20		when m=25	
		d_{ij}	$P''_{ik} \times d_{ij}$	d_{ij}	$P''_{ik} \times d_{ij}$	d_{ij}	$P''_{ik} \times d_{ij}$
1	0						
2	1	0.556	0.556	0.741	0.741	0.926	0.926
3	2	0.256	0.584	0.199	0.398	0.071	0.142
4	3	0.113	0.339	0.084	0.252	0.003	0.009
5	4	0.047	0.188	0.010	0.040		
6	5	0.018	0.090	0.002	0.010		
7	6	0.003	0.018				
8	7	0.002	0.014				
9	8	0.001	0.008				
10	9						
11	10						
12	11						
13	12						
14	13						
15	14						
16	15						
17	16						
18	17						
19	18						
20	19						
21	20						
22	21						
23	22						
24	23						
25	24						
26	25						
27	26						
28	27						
Average Relay Times			1.717		1.441		1.077

Appendix IV– Product of average relay times and number of requests (popularity class)

λ	when m=1 R=13.99	when m=5 R=4.75	when m=10 R=2.54	when m=15 R=1.72	when m=20 R=1.44	when m=25 R=1.07
> 4.5	69.95	23.60	12.72	8.68	7.20	5.35*
3.5 – 4.5	55.96	18.88	10.16	6.88	5.76*	4.28
2.5 – 3.5	41.97	14.16	7.62	5.16*	4.32	3.21
1.5 – 2.5	27.98	9.44	5.08*	3.44	2.88	2.14
0.5 – 1.5	13.99	4.72*	2.54	1.72	1.44	1.07
0 – 0.5	0	0	0	0	0	0
*		4.72*	5.08*	5.16*	5.76*	5.35*

* Desired product of average relay times and number of requests

References

1 Atherton, P. *Handbook for Informtion System and Services,* Paris, UNESCO, 1977.

2. Brookes, B.C. The design of cost-effective hierarchical information systems, *Information Storage and Retrieval,* **6** (2) (1970), 127-136.

3. Buckland, M.K. *Book Availability and Library User,* New York, Toronto, Oxford, Sydney, Braunschweig : Pergamon Press Inc., 1975.

4. Duggan, M. Library network analysis and planning (LIB-NAT), *Journal of Library Automation,* **2**(3), (1969), 157-175.

5. Egghe, L. and Rousseau, R. *Introduction to Informetrics,* Amsterdam : Elsevier 1990.

6. Fiels, K.M. Co-ordinated collection development in a multitype environment : promise and challenge, *Collection Building,* 7(2), (1986) 26-31.

7. Kent, A. Network Anatomy and Network Objectives. In: *The Structure and Governance of Library Networks.* THE 1978 CONFERENCE, Pittsburgh, Pennsylvania. Proceedings. ed. by National Commission on Libraries and Information Science. University of Pittsburgh, Marcel Dekker, Inc., New York and Basel, 1978.

8. Khalil, M. Document delivery : a better option?, *Library Journal* , **118** (2) (1993), 43-45.

9. Kleinrock,L. *Queueing Systems. Volume I : Theory,* John Wiley & Sons, New York, London, Sydney, 1975.

10. Morse, P.M. *Library Effectiveness : a systems approach,* Cambridge, Massachusettes and London, England : The M.I.T. Press, 1969.

11. Pitt, W.B. and D.H. Kraft. Buy or copy? A library operations research model, *Information Storage and Retrieval,* **10** (1974), 331-341.

12. Rouse, W.B. A library network model, *Journal of the American Society for Information Science* , **15**(2) (1976) 88-98.

13. Rouse, W.B. Tutorial : mathematical modelling of library systems, *Journal of the American Society for Information Science,* **30** (4) (1979), 181-191.

14. Williams, J.G. Performance criteria and evaluation for a library resource sharing network, In : *Conference on Resource Sharing in Libraries, Pittsburgh,* 1976. Proceedings. New York : Marcel Dekker, Inc. 255-277.

15. Zhang, Q. *Obsolescence and Bradford Distribution of Rice LIterature,* M.Sc. Thesis, The City University, London, UK, 1986.

16. Zhang, Q. *Agricultural Libraries and Information Centres in China : co-operation. resource-sharing and networking,* PhD Thesis, The City University, London UK, 1990.

17. Zhang, Q. Improving the accessibility and availability of information in the agricultural library and information system of China, In : *Proceedings of the VIIIth World Congress of IAALD, May 27-30.* Budapest, IAALD Quarterly Bulletin, **37**(1-2) (1990), 55-58.

Information Diffusion of Terms as Found Through Various Types of Documentary Sources

Subir K Sen *

Department of Library Science, University of Calcutta, Asutosh Building, Calcutta 700 073, India

A. Seal

I I T Library, New Delhi 110016, India

In the present study a methodology of studying terminological diffusion has been suggested and used with seven technical terms. The process of diffusion occurs in two planes. One is in the vertical plane where the seepage in various types of documents from research journals to general language dictionaries and newspapers has been studied. The other is in the horizontal plane where the recurrence (or frequency of occurrence) of the term in the secondary services has been studied. These two together give the idea of life history of the term in the process of becoming public knowledge and gaining literary warrant. This in itself may indicate the cultural and informational status of a term for inclusion in an information processing system eg. glossary, thesaurus, IRS etc.

1. Introduction

The present paper is a preliminary report on an empirical study of diffusion of concepts as seen through the creation and recurrence of use of technical terms over time.The main objectives had been to try to formalize a method of studying information diffusion of terms and pave the way for further research and conceptualization. It was also the intention to understand the difficulties involved.

Technical terms can be considered as carriers of information representing the key concepts. Coining or creation of a term can be considered as an innovation : terminological innovation. Terminological diffusion can therefore provide a means of studying diffusion and spread of information. Even if small ,there have been some published material dealing with diffusion of information or idea [2,4,5,6,7,8,9].But none of them addressed specifically to the problem of terminological diffusion and its relation to informetrics.

Some confusions also prevail in using the term "diffusion" in the context of spread of information or idea [3]. No literature survey is presented here, which may be done in a future review. The terminological confusion and the meanings of diffusion as used in different contexts or in other subjects have been discussed elsewhere [11].

Following the career of a term can be of much interest. This can provide for an indication or estimation of the `cultural' and `informational' status of a term which can be utilized for inclusion in a glossary or thesaurus or an information retrieval system. A pioneering preliminary study on the careers of terminologies was done by Peritz in 1984[10]. The present study is both extension of and modification over Peritz's study.

2. Terminological Clarification

In this paper we have used the term `diffusion' in the framework of a specific technical meaning : the phenomenon or process of new information (new knowledge, in the form of a term representing a concept of idea) becoming public knowledge (public information).

The term public knowledge can again create confusion. We are very much specific about its use here. Following Wilson[13] we define public knowledge as the common stock of articulated knowledge. Anything which is public knowledge

*Permanent Address: 85 Devi Nivas Road, Calcutta 700 074

is usually a piece of information about which most of the members of a community are aware of, or otherwise, the information is available to the members of the community in a usable form. In this work Wilson's statement that '... encyclopedias and other works of reference in which public knowledge is formulated' has been a supporting force.

We can visualise the process of diffusion of information comprehensively. The phenomenon is prominent in science and technology rather than other traditional subjects. Borgman[1] and also Sen and Chatterjee[12] have considered diffusion studies as a subfield of bibliometrics. Borgman[1] writes - 'Bibliometrics may be used to trace the evolution of an idea within and across disciplines. At the earliest stages of diffusion the idea is linked with the document in which it was first presented, thus allowing tracing through citations. As an idea diffuses further, it may become disassociated from its bibliographic origins, thus requiring tracing through terminology'.

In science and technology someone has usually to coin a term for a new discovery or invention. Most commonly such a coinage is done by the discoverer or the inventor; occasionally, however, this may be done by others. In many cases some conceptually new development may prevail for sometime without having a name. Again, in exceptional cases, (but not very uncommon in science and technology) a name or a term may be coined much before the actual discovery or description of the new development e.g. the names of the elementary particles like electron, meson, neutrino etc were coined much before actual discovery of the particles.

3. Four Phases of Diffusion

Once a term is coined, it gets associated not only with a concept, but also with one or more context. First stage of diffusion of the information starts with the use of the term along with the associated concept and context by others.

In science and technology, a new concept is described in a primary publication such as a research journal or a patent application. It may however be announced at a meeting or a press conference. The new terms therefore, make their first appearances in such media as are called primary publications or primary information sources.

Then, there should be further primary publications with the term being used in these publications. In the first phase of diffusion, the new term would repeatedly appear in primary publications with varying temporal gaps. In some cases, the term may also appear in popular press. But, it should also appear in primary publications if it is a technical term. In this early period, the term may undergo various types of modifications in respect of spelling, definition etc.

Sometimes there are more than one term coined for the same concept. All these terms are put into a competition for acceptance. Many terms may be discarded and are not used further. Their lives are started and finished during this first phase of diffusion process.

If a term is accepted by the peers and continues to be used, we may say that the associated concept has been granted literary warrant. When this phase is attained, the term is found in the titles of articles and in indexes of the secondary periodicals. The concept, and so the term starts getting places in the progress reports, reviews, year books, anthologies etc. They are then reccognised as key-words or thesaurus words or access points by the actual practioner and also by the professional indexer, database manager and information scientist. This is the second phase.

In the third phase of diffusion, the term should be incorporated in the subject encyclopedias, in the subject glossaries or the subject dictionaries, in popular periodicals and in news papers. At this phase, such a term is almost always defined or explained whenever used in a non-technical medium. Usually during this phase, the concept and the term starts to be included in subject heading lists and classification schemes.

In the last or fourth phase, the term makes appearance in text books, in general dictionaries and also in popular press without being explained or defined. They may also be used by the public in their regular conversations. When this stage is attained, one may say that the term and hence the the concept has become public knowledge.

Although the sequence of these four phases or stages is ideal, in many instances, the sequence is jeopardised. Again the term public knowledge may

mean several 'states' of knowledge. One should consider various levels of public knowledge if one may think in terms of the community in which the concept becomes familiar (under rational consensus) [14]. This community may be an international group of specialists or the students,scholars and persons trained in a particular field or an educated intelligentia or common folk in general. We may talk of public knowledge among physicists or among those persons who read newspapers. We may illustrate by taking a few examples. The term 'quark' or its associated concept is public knowledge in the community of physicists, the term bibliometrics is common knowledge for a very small group of persons only, but the terms like radar or AIDS or vaccine are known to almost everybody. Again the status of a term is to a large extent connected to the language to which it belongs. In this study only the English language technical terms have been considered.

Seven terms were selected. They are AIDS, Pheromone, Holography, Laser, Gluon, Parton and Quark.

4. Two Types of Diffusion

Terminological diffusion as information diffusion can be studied in two ways. Firstly as a vertical flow or seepage - by finding out how a term spreads at different levels of communicati channels and how it appears in documents to be used more by general public. The other aspect is the horizontal flow which would indica e the use and acceptance of the term within the professional community or the community of specialists. This can be measured or estimated by counting the frequency of occurrence of a term in the secondary services over time. This second aspect is much ealier to study than the first.

5. Methodology

In studying the horizontal diffusion we have counted annual frequencies of occurrence of a term from the indexes of the relevant secondary services as far as practicable - firstly from the abstracting periodical of the subject field in which the term belongs, then from an abstracting periodical in a related field. We had also counted frequencies from the Permuterm Subject Index of SCI®in all cases. But we have not presented this data here as we have found that the data collected manually (required for pre–1980) and those collected mechanically largely differ. Only for AIDS (which was coined in 1982, SCI®data have been provided. To find out popular appeal and public importance of a term we searched through The New york Times (NYT)index also. Except for AIDS, and insignificantly for holography, no frequency table was possible for NYT.

The horizontal diffusion when observed against time, indicates growth of primary literature on the topic which the term represents. Most of the growth studies of literature have followed this method. However, what these studies have not so far noted is that this is also an indication of information diffusion. Moreover, frequency counts from newspaper index have given an extra dimension to diffusion aspect. On the other hand, studying the vertical diffusion is much more difficult and complex. It is very difficult to discover when a technical term makes its first appearance in a particular type of document.

All the seven terms chosen had been coined within last thirty-five years. The point of genesis of each term has been pin-pointedly identified. This involved identification of the source document where the term first appeared, the person who coined the term, and the date of the source document. It might have some gestation period before its appearance and previous history of informal circulation which have been ignored. Then, the possible channels eg. indexes of the primary or secondary periodicals, annual reviews, different editions of subject and general dictionaries and encyclopedias and ultimately subject heading lists were identified and located.

We have identified 17 different types of information sources besides the source of first appearance of the term to follow the vertical spread from the most specialised to the most popular. These source types are listed below:

1. First appearance of the term - date and source.
2. Appearance in the annual index of the secondary service in the relevant subject field in which the term belongs.
3. Appearance of the term in annual indexes of secondary services of related subject areas.

4. Appearance of the term in annual permuterm index of Science Citation Index.
5. Appearance of the term in newspaper indexes.
6. Appearance of the term in indexes of annual reviews or progress of relevant subject field.
7. Appearance of the term in index of specialised subject year book, if there is any.
8. Appearance of the term in a year book of a broad subject field, if any.
9. Appearance of the term in the index of a general year book.
10. Appearance of the term in specialised subject encyclopaedia.
11. Appearance of the term in broad subject encyclopaedia.
12. Appearance of the term in general encyclopaedia.
13. Appearance of the term in a specialised subject dictionary.
14. Appearance of the term in broad subject dictionary.
15. Appearance of the term in general dictionary
16. Appearance of the term in a small desk or pocket dictionary and or bi-lingual dictionary.
17. Appearance of the term in text book.
18. Appearance of the term in subject heading list and/or subject catalogue.

For each term, all these types of sources were searched and the date of first appearance in each case was noted.This gave an idea in quantitative terms of life history and line of spread of each term. Searching the terms in the dictionaries and encyclopaedias was troublesome. This might become misleading if not done cautiously and systematically. Not all the dictionaries and encyclopaedias can suitably reflect the diffusion process as these do not have policies of regular revision.

One has therefore to select a number of source dictionaries and encyclopaedias with publication history of revised editions at regular intervals, which are available commonly and have sufficient publication life to cover the period from genesis of the term up to date. This is also a reason why we could not study the cases of the terms which are quite old. Another, practical situation creating problem to searching should be discussed. The libraries as a general practice keep and display only the recent editions of these sources. Earlier editions in many cases are disposed off.

For these reasons we tried the sources selectively. Different editions of the McGraw Hill Encyclopaedia of Science and Technology and its year books, Encyclopaedia Britannica and its year books, Encyclopaedia Americana and its year books (rarer occasions), Chamber's Dictionarry of Science & Technology, Oxford English Dictionary, Shorter Oxford Dictionary were searched. In case of non-availability of a term in these sources other general or broad subject encyclopaedias and dictionaries were consulted. For subject headings, different editions of Sears' List of Subject Heading and/or Library of Congress Subject Heading List were consulted. For specialized narrow subject dictionaries and encyclopaedias we could not follow any pattern.

Although, occurrence of a technical term and its frequent recurrence in a newspaper (as reflected through the index of NYT in this study) indicates the ultimate status of the term towards public knowledge which may be considered as the most popular level. Yet we have put it for an early position before the annual reviews, encyclopaedias and dictionaries, just after the secondary services. The reason is that newspapers usually report a major conceptual advance much before the reviews and reference tools do.

As the last level of diffusion channel we have taken the subject heading list. Inclusion in such a list occurs usually after the term gains 'literary warrant'.

6. Results and Discussions

Table 01 gives the complete scenario of vertical diffusion of the terms (AIDS, Gluon, Holography, Laser, Parton, Pheromone, Quark). The first row gives the year of first coinage or first appearance in a documentary source and is indicated as the Zero year. The next seventeen rows (numbered 2 to 18) give the time lag in years between the first appearance (Zero year) and the appearance in the correspondig category of documentary source - in the sequence shown above.

The term AIDS first appeared in *Morbidity and Mortality Weekly Report* (US) on 24 September 1982. Gluon appeared in *Physics Bulletin* of 3 December 1971. Holography was in *Physics Let-*

ters of 15 December 1964, Laser as an acronym or abbreviation had its first appearance on 8 July 1960 in the *New York Times* and not in any technical journal. Parton, pheromone and quark appeared in *New Scientist*, a general science magazine of 26 June 1969, in *Nature* in 1959 and in *Physics Letters* in 1964 respectively.

Among the seven terms five are from physics, three of them from highly technical basic research field, namely elementary particle physics. They are conceptual notions. The other two physical terms refer to two mechanisms and machines with wide applications.

The other two terms have been taken from life sciences and biomedicine. Pheromone refers to a chemical substance in the animal kingdom. The other term, AIDS refers to a deadly disease.

The earliest of the term, pheromone was coined in 1959 and the lastest, AIDS in 1982. The two of the terms laser and AIDS are abbreviations of some larger descriptive phrases. Laser has become an acronym. But status of AIDS an acronym is not clear. It is not written as Aids.

By going through the *Collins Cobuild English Language Dictionary* which has been published in 1987,only three terms from our list have been found out. These terms are AIDS as an abbreviation, holograph and hologram but not holography; and laser as an acronym. In respect of technical terminology the dictionary says that it includes

Table 01. Time gap in years for appearance in different types of documents after first appearance (zero year - first row) for each term

No.	AIDS	Gluon	Holography	Laser	Parton	Pheromone	Quark
1	0 (1982)	0(1971)	0(1964)	0 (1960)	0(1969)	0(1959)	0(1964)
2	1	1	1	1	2	0	1
3	-	1	2	1	1	0	0
4	0	1	1	5	1	6	1
5	0	10	3	0	1	15	1
6	2	-	7	2	-	17	7
7	3	-	7	-	-	-	-
8	3	13	2	2	1	8	6
9	1	10	7	13	6	11	11
10	10	11	10	3	5	11	10
11	5	12	4	6	8	17	7
12	4	16	9	13	9	14	7
13	4	15	7	12	9	13	7
14	5	14	10	7	15	12	11
15	5	11	6	6	15	7	6
16	9	13	4	14	-	-	10
17	3	14	3	4	-	10	3
18	4	16	11	5	17	16	11

words which are technical in origin but which are of regular use in the general vocabulary.

Taking quark, parton and gluon together as falling in the same narrow speciality of theoretical elementary particle physics, we may judge their respective status and diffusion characteristics. The term quark was coined in 1963-64 by the famous Nobel laureate M. Gellmann taking it from the novel Finnegans Wake by James Joyce. The term parton was coined by equally famous R. P. Feynman, another Nobel laureate. Although the term parton appeared in print by about 1968-69, it was in use in informal discussions from 1965-66.The term gluon was derived from glué and was coined probably by Gellmann. In a thesaurus, the relations among these terms should be as follows : Quark may be considered as an NT (Narrower Term) - it means quark can be considered as one type of parton and parton can be taken as a Broader Term (or BT) of quark. Gluon should be considered as RT (a conceptually Related Term) of quark, it may be taken as an RT as well as an NT of parton.

It could be assumed that these terms representing some esoteric concepts should have taken much more time to diffuse. But judging their importance in particle physics, the terms should be heavily used in specialised documents of high energy physics. Except for quark the case was not so (Tables G1, P1, Q1) as is revealed from annual indexes of *Physics Abstracts* (PA).

The term 'quark' appeared in the `Advances in Nuclear Physics' in 1971, but the other two terms have not got a place in that periodical. The most notable for our purpose is the fact that all the three have been included in subject heading lists; it took sixteen years for gluon, seventeen years for parton and eleven years for quark.

Going over to the terms 'holography' and 'laser', we find that the terms became very quickly diffused. Development of holography has been very much depended on the development of laser. Holography has not been developed to the stage of such universal use as laser. Till now we do not have viable holographic devices in miniature forms. Thus, diffusion of holography lagged far behind that of laser.

The term 'laser' appeared in the New York Times Index before it appeared in any secondary service. This also signifies the rapidity of diffusion of the term.

Going through the terms in biological science we find that the term 'pheromone' was very quickly accepted and used in the specialised techniccal literature. But its spread in general literature of science and literature outside the scope of the specialised subject field has been slow.

A study of diffusion process of pheromone reveals some interesting modes of diffusion. The term pheromone appeared in *Chemical Abstracts* (CA) and *Biological Abstracts* (BA) in the same year of first coinage. After that there were gaps of six years and two years in BA and CA respectively for its recurrence. Pheromone appeared in *New York Times Index* in 1991 after 15 years of its coinage.

The case of AIDS is unique. As the term is an acronym used for a devastating disease having global impact the diffusion of term has been very quick. Within months of its first appearance, it was reported in *The New York Times* (NYT) and other ser is a term of physical science and AIDS is a term in medical science having grave consequences for and interrelation with public health.

Table A 1. Frequency of occurrence of the term AIDS in Biological Abstracts

Year	No.of occurrence
1983	8
1984	15
1985	200
1986	326
1987	344
1988	210
1989	332
1990	565

Table A2. Frequency of occurrence of the term AIDS in Science Citation Index

Year	No. of occurrence
1982	102
1983	401
1984	578
1985	996
1986	1307
1987	1603
1988	1708
1989	1359
1990	1491

Table A 3. Frequency of occurrence of the term AIDS in New York Time Index

Year	No. of occurrence
1982	3
1983	128
1984	70
1985	390
1986	392
1987	1002
1988	454
1989	395
1990	563

The horizontal diffusion of AIDS gives (Table A1, 2, 3) a peculiar pattern as revealed through three different types of secondary sources.There is no entry of the term in 1982 in Biological Abstracts. There are eight and fifteen entries respectively in 1983 and 1984. In 1985, the number becomes 200, going to 565 in 1990. The term appeared three times in 1982 itself in The New York Times (NYT). In 1983, the number is 128. In 1987 NYT index registers 1002 occurrences. Between 1985 and 1990, in other years, the number is in between 390 and 563. SCI registers 102 occurrences in its index in 1982. There is then a steady growth culminating in 1988 to 1708 occurrences. In the next two years the number falls a little. These show a very fast diffusion and growth of both specialised and general literature. The years around 1985 was significant for a number of major publications. It is also to be noted that items of specialised literature had much less relevance for *Biological Abstracts.*

Unfortunately, we could not study the growth or diffusion through the most relevant secondary services in this case - *AIDS and Virology Abstracts or Index Medicus* which were not readily available.

Frequencies of occurrences of pheromone are significant. In BA and CA continuous growth is prominent with some jump increment at intervals. The increasing trend is maintained as the global coverage for BA and CA gives the picture of growth due to inclusion of lesser journals (Table PH 1,2).

Frequencies of occurrences of Gluon (G 1,2) and Parton (P1,2) follow a pattern. Gluons are supporting particles to quarks. As such chances of appearance of the term 'Gluon' in the titles of articles and in indexes of journals or in the indexes of secondary services is less than that of quark (Q 1,2). The terms quark and gluon might also appear together as in quark-gluon model.

Table PH 1. Frequency of occurrence of the term Pheromone in Biological Abstracts

Year	No. of occurrence
1959	1
1960	0
1961	0
1962	0
1963	0
1964	0
1965	0
1966	0
1967	20
1968	23
1969	26
1970	57
1971	81
1972	64
1973	145
1974	174
1975	158
1976	181
1977	173
1978	240
1979	235
1980	240
1981	257
1982	272
1983	300
1984	363
1985	392
1986	352
1987	287
1988	353
1989	380
1990	419

The terms gluon, parton and quark do not appear in the subject Index of CA from 1971 onwards. They however appeared in the *Chemical Substance Index.* The fequency table (G 2, P 2, Q 2) for them show occurrences of these terms from the Chemical Substance Index. For this reason the frequencies have become higher for CA than for PA.

As already discussed realization of practical holography has been very much dependent on the development of laser technology. It may be noted that the theoretical principle of laser was discovered even before 1920. But no idea or concept of such an artefact let alone a name was forthcoming till fiftees. The theory of holography was enunciated by Denis Gabor in 1948 and the term was coined at that time.

Table PH 2. Frequency of occurrence of the term Pheromone in Chemical Abstracts

Year	No. of occurrence
1959	1
1960	0
1961	0
1962	2
1963	2
1964	3
1965	1
1966	4
1967	14
1968	24
1969	25
1970	11
1971	61
1972	60
1973	92
1974	104
1975	76
1976	100
1977	204
1978	221
1979	235
1980	249
1981	255
1982	279
1983	293
1984	301
1985	352
1986	325
1987	335
1988	315
1989	297
1990	309

Table G 1. Frequency of the term Gluon in Physics Abstract

Year	No. of occurrence
1972	2
1973	1
1974	1
1975	4
1976	23
1977	22
1978	22
1979	42
1980	25
1981	57
1982	26
1983	23
1984	55
1985	53
1986	50
1987	56
1988	65
1989	79
1990	121
1991	114

Table G 2. Frequency of the term Gluon in Chemical Abstracts

Year	No. of occurrence
1972	2
1973	27
1974	35
1975	55
1976	61
1977	86
1978	87
1979	214
1980	320
1981	354
1982	279
1983	427
1984	465
1985	502
1986	509
1987	496
1988	485
1989	480
1990	479

The tables (L 1,2,3) show that within a short span of time publication on laser proliferated in a very large way. This is quite expected because laser has multifarious applications in many fields. This prompted publication of a separate secondary service for laser - *Journal of Laser Abstracts* (JCLA).

Table P 1. Frequency of occurrence of the term Parton in Physics Abstracts

Year	No. of occurrence
1971	2
1972	4
1973	129
1974	131
1975	105
1976	108
1977	92
1978	147
1979	144
1980	124
1981	203
1982	179
1983	195
1984	183
1985	173
1986	202
1987	223
1988	193
1989	176
1990	150
1991	188

Table P 2. Frequency of occurrence of the term Parton in Chemical Abstracts

Year	No. of occurrence
1970	2
1971	14
1972	80
1973	130
1974	119
1975	130
1976	140
1977	101
1978	151
1979	116
1980	171
1981	162
1982	193
1983	201
1984	179
1985	175
1986	225
1987	209
1988	182
1989	178
1990	190

Table Q 1. Frequency of occurrence of the term Quark in Physics Abstracts

Year	No. of occurrence
1965	14
1966	14
1967	89
1968	133
1969	71
1970	90
1971	89
1972	98
1973	227
1974	266
1975	186
1976	154
1977	158
1978	533
1979	960
1980	1315
1981	1401
1982	1529
1983	1637
1984	1499
1985	1401
1986	157
1987	145
1988	184
1989	174
1990	160

Table Q 2. Frequency of occurrence of the term Quark in Chemical Abstracts

Year	No. of occurrence
1964	2
1965	27
1966	25
1967	214
1968	85
1969	86
1970	164
1971	213
1972	182
1973	211
1974	255
1975	463
1976	251
1977	236
1978	202
1979	196
1980	190
1981	189
1982	170
1983	210
1984	202
1985	169
1986	152
1987	162
1988	?
1989	132
1990	121

Table L 1. Frequency of occurrence of the term Laser in Physics Abstracts

Year	No. of occurrence
1961	1
1962	6
1963	4
1964	42
1965	630
1966	961
1967	1226
1968	1669
1969	1470
1970	1475
1971	1490
1972	1503
1973	1532
1974	1540
1975	1572
1976	1595
1977	1601
1978	1612
1979	1622
1980	1629
1981	1635
1982	1642
1983	1653
1984	1675
1985	1683
1986	1695
1987	1715
1988	1727
1989	1736
1990	1756

Table L 2. Frequency of occurrence of the term Laser in Journal of Current Laser Abstracts

Year	No. of occurrence
1968	2852
1969	3422
1970	3373
1971	3705
1972	4074
1973	4005
1974	4198
1975	4912
1976	4598
1977	4403
1978	4417
1979	4446
1980	4624
1981	4276
1982	4147
1983	4107
1984	4618
1985	4544
1986	4654
1987	4715

Table L 3. Frequency of occurrence of the term Laser in Chemical Abstracts

Year	No. of occurrence
1961	10
1962	40
1963	176
1964	289
1965	694
1966	1131
1967	918
1968	1057
1969	994
1970	982
1971	1049
1972	1075
1973	1096
1974	1125
1975	1122
1976	1159
1977	1230
1978	1301
1979	1392
1980	1431
1981	1489
1982	1535
1983	1575
1984	1641
1985	1669
1986	1725
1987	1757
1988	1825
1989	1792
1990	1845

Table H 1. Frequency of occurrence of the term Holography in Physics Abstract

Year	No. of occurrence
1965	2
1966	23
1967	33
1968	204
1969	250
1970	469
1971	577
1972	592
1973	601
1974	620
1975	355
1976	402
1977	310
1978	341
1979	352
1980	375
1981	389
1982	426
1983	439
1984	427
1985	435
1986	442
1987	452
1988	459
1989	462
1990	479
1991	383

Table H 2. Frequency of occurrence of the term Holography in Chemical Abstracts

Year	No. of occurrence
1966	4
1967	25
1968	11
1969	20
1970	41
1971	59
1972	104
1973	107
1974	108
1975	120
1976	159
1977	162
1978	175
1979	182
1980	185
1981	187
1982	188
1983	189
1984	192
1985	201
1986	210
1987	219
1988	224
1989	229
1990	230

Table H 3. Frequency of occurrence of the term Holography in New York Times Index

Year	No. of occurrence
1967	3
1968	1
1969	2
1970	4
1971	2
1972	3
1973	2
1974	2
1975	6
1976	2
1977	2
1978	0
1979	1
1980	3
1981	1
1982	1
1983	0
1984	5
1585	3
1986	2
1987	4
1988	1
1989	3
1990	3
1991	3

Entries in JCLA are more than 2.5 times that in PA and about 3 times that in CA.Laser has become a household word in the advanced countries. The horizontal diffusion is in tune with the vertical diffusion in case of laser. In contrast, frequencies of occurrence of the term holography in secondary services are less. The frquencies in PA and CA have somewhat similar trends. PA has much more entries than CA (Table H 1,2).

6. Concluding Remarks

The method adopted here either for vertical or for horizontal diffusion cannot be taken as 'very precise' in quantitative terms.

As has already been pointed out, unless all the publications of a certain category (say general desk dictionaries) in a given period can be searched, the extent of diffusion cannot be determined specifically. But this is not usually possible. Moreover, due to gap in publication dates, two different terms having different diffusivity may appear in the same edition simultaneously. Different sources (except the secondary services) may or may not include a topic at a certain period according to policy or other reasons. So the picture we get may be fuzzy.

In case of horizontal diffusion, counting the occurrences from indexes of different secondary services would give different pictures. Those indexes prepared from titles would give a largely different number than those prepared otherwise. Similarly coverage of periodicals by the concerned services would affect the frequency of recurrence. But if taken together a definitive picture evolves out. Moreover, by comparing the frequencies of occurrences of a term in different secondary services we can surely gain insights, not only about the diffusion of the term, growth of literature and spread of the concepts but also about the nature of the services themselves.

Coadic[4] presented a theoretical model of heuristic diffusion of scientific ideas which is related to our horizontal diffusion reflecting the growth of literature which is different from epidemic model of Goffmann[7].

Study of diffusion processes of technical terms following the methodology suggested here may lead to an understanding of the process of information dynamics and flow. Such studies may also shed light to the social relevance of an idea or a concept as also the social or cultural status vis-a-vis academic or technical status.

In the more practical level such a study would indicate the time when a term (keyword) should be included in a thesaurus or classification scheme or IR system or an index of a secondary service.

S K S has been able to develop a means of finding out a sort of an index for information or terminological diffusion which may have some consequences in understanding the process and using the methodology adopted in this study. That finding will be reported in future.

References

1. Borgman, Christine L. (1989) Bibliometrics and schol arly communication : editor's introduction .*Communications Research* **16**(8) 583-99.
2. Chatman, E.A. (1986) Diffusion theory : a review and test of a conceptual model of information diffusion. *JASIS* .**37** (6), 377-386.
3. Le Coadic, Y.F. (1986) Modelling the communication, distribution transmission or transfer of scientific information *Journal Inf. Sci* **13,** 143-48.
4. Le Coadic, Y.F. (1979) The heuristic diffusion of scientific ideas : a mathematical and sociological approach. *Int. Forum Inf. Doc.* **4** (2) 7-11.
5. Dahling, R.L. (1962) Shannon's information theory : the spread of an idea (in *Studies of innovation and of communication to the public.* Standard, (CA) : Institute for Communication Research).
6. Feder, Gershon and G.T. Omara.(1982)On information and innovation diffusion: a Bayesian approach. *American J. Agricultural Economics.* **64,** 145-47.
7. Goffmann, W. . (1970) A general theory of communica tion (in Tefko Saracevic (ed.) *Introduction to information science.* New York : R. R. Bowker), 726-747.
8. Goffmann, W. (1966) Mathematical approach to the spread of scientific ideas - the history of mast cell research. *Nature,* **212,** (reprinted in T. Saracevic (ed) *Introduction to information science.* New York : R. R. Bowker, 1970, 65-69.
9. Paisley,W.(1989) *Communication in the communica tion sciences.* (in B. Dervin and M. Voigt, eds. Progress in the communication Sciences 5, Newbury Park, CA: Sage) 1-43.
10. Peritz, Bluma C. (1984) On the careers of terminologies; the case of bibliometrics. *Libri* **34**(3),233-242.
11. Sen, Subir K. Knowledge, information, data and all that a conceptual spadework of information dynamics. *Iaslic Bulletin* (to be published).
12. Sen, Subir K. and S.K. Chatterjee (1990) An introduction to researches in bibliometrics. Part 1 background and perspective. *Iaslic Bulletin* **35**(3), 105-117.
13. Wilson, Patrick.(1977) *Public knowledge, private ignorance; towards a library and information policy.* Westport (CT) : Green wood Press.
14. Ziman, J.M. (1968) *Public knowledge.* London : Cambridge University Press.

CALL FOR PAPERS

SIXTH INTERNATIONAL CONFERENCE ON SCIENTOMETRICS AND INFORMETRICS

To be Held at

THE ISRAEL ACADEMY OF SCIENCES AND HUMANITIES

JERUSALEM

JUNE 16-19, 1997

SPONSORED BY
THE INTERNATIONAL SOCIETY FOR
SCIENTOMETRICS AND INFORMETRICS

This is the sixth in a series of increasingly successful biennial conferences. The previous conferences were held at the ***Limburgs Universitair Centrum, Diepenbeek, Belgium;*** at the ***University of Western Ontario, Canada;*** at the ***Indian Statistical Institute, Bangalore, India;*** at the ***Association for Science Studies, Berlin, Germany,*** and ***Rosary College, River Forest, Illinois.***

SCOPE

The scope can be broadly defined as those topics which treat in quantitative fashion the creation, flow, dissemination, and use of scholarly or substantive information.

Representative, but by no means inclusive, topics are : informetric "laws" and distributions; mathematical models of communication; citation analyses; theory of document, text and information retrieval; information and productivity; the quantitative sociology of science and of other substantive information-based activities; study of aging and dispersion; application of informetrics to file design, data compression, etc.; informetric applications to policy analysis, R&D management, etc.; ...

LOCATION

The conference will be hosted by the School of Library, Archive and Information Studies of the Hebrew University of Jerusalem. It will be held at The Israel Academy of Sciences and Humanities.

Jerusalem is warm but pleasant in June. The conference venue is situated in the most attractive part of town, next door to the President's Residence and the Jerusalem Theatre. The Ben-Gurion International Airport is a 35 minutes drive from Jerusalem.

We look forward to seeing you all in Jerusalem in June 1997.

International Society for Scientometrics and Informetrics (ISSI)

From 11 to 15 September 1993, the 'Fourth International Conference on Bibliometrics,Informetrics and Scientometrics' was held in Berlin, Germany. The meeting, which was dedicated to the memory of Derek de Solla Price, the founding father of our field of research, has been a great success. The fact that it was attended by 189 participants from 33 countries shows that this emerging scientific field is gaining importance and worldwide recognition. In order to further stimulate this development, the program committee has decided to found the International Society for Scientometrics and Informetrics (ISSI).

After a period of preparations, the new Society was officially founded on 5 October 1994 in Utrecht, The Netherlands.

ISSI's goals are the advancement of theory, method and explanation of the following areas :

1. Quantitative studies of :

- scientific, technological and other scholarly and substantive information;
- the science of science and technology, social sciences, arts and humanities;
- generation, diffusion and use of information;
- information systems, including libraries, archives and databases.

2. Mathematical, statistical and computational modelling and analysis of information processes.

In order to achieve them, the organization directs its activities at:

1. Communication and exchange of professional information;
2. Improving standards, theory and practice in all areas of the discipline;
3. Education and training;
4. Enhancing the public perception of the discipline.

In recognition of her expertise and dedication, Dr. Hildrun Kretschmer, organizer of the Berlin conference, has been chosen to be the first president of ISSI.

Secretary-treasurer is Dr. C. le Pair, Technology Foundation (STW),P.O. Box 3021, 3502 GA Utrecht, The Netherlands.

Meanwhile, plans for the other International Conferences took shape.

In the following page you will find an ISSI-membership application form.

ISSI Membership Information

If you are active in the field of Scientometrics or Informetrics, you can become a member of the newly-founded Society. For a membership fee of $ 20.-, you will recieve early information on activities in the field through regularly appearing newsletters.

In view of the fact that many researchers in our field live in developing countries, the ISSI Board has decided on a special, reduced fee of $ 5.— for people from those countries.

JISSI is issued quarterly. Correspondence regarding advertishment and subscription should be made to the publisher. Payments may be sent by Cheque/Bank Draft/Money order.
For cheques outside Delhi but from within India a sum of Rs 15/- should be added. For cheques out side India a sum of U.S. dollar 2.00 or equivalent should be added.
Claims for missing number(s) should be made within four months of the date of issue of such number to the Publisher.

Cheques/Demand Drafts should be drawn in favour of **'Brzark Information Systems (P) Ltd'.**

The order form may please be filled up and sent to :

Brzark Information Systems (P) Ltd,
112, Humayun Pur (Near N.C.C. Office)
Safdarjung Enclave. New Delhi - 110029.

Subscription Rates (All rates include airmail surcharges)

Rates for Canada,USA,Europe,Japan,Singapore,Australia, New Zealand & Middle East:

INSTITUTIONS (including LIBRARIES)		
Annual	.....	US $200
For first three years	.....	US $400
INDIVIDUALS		
Annual	.....	US $100
For first three years	.....	US $200
ISSI MEMBERS		
Annual	.....	US $75
For first three years	.....	US $150

Special Offer for Other (Developing) Countries:

INSTITUTIONS		
Annual	.....	US $125
For first three years	.....	US $250
INDIVIDUALS		
Annual	.....	US $50
For first three years	.....	US $100
ISSI MEMBERS		
Annual	.....	US $30
For first three years	.....	US $60

Rates for Indian Subcontinent:

INSTITUTIONS		
Annual	.....	Rs2500
For first three years	.....	Rs5000
INDIVIDUALS		
Annual	.....	Rs1000
For first three years	.....	Rs2000
ISSI MEMBERS		
Annual	.....	Rs750
For first three years	.....	Rs1500

JISSI BUSINESS REPLY FORM

We Accept Payment by Credit Cards

I/We want to subscribe JISSI for one year / for three years*.

Name : .....

Postal Address : .....

.....

.....

Postal Code : .....

Country : .....

Telephone : .....

Fax : .....

E Mail : .....

Whether member ISSI * : Yes / No

I/We enclose the sum of.....

Drawn in favour of 'Brzark Information Systems (P) Ltd.'

Reference and date of DD/Cheque/MO* : .....

Charge my credit card VISA / MASTER / DINERS*

No.....

Expiry date.....

Signature

*Please cross out which do not apply/add US $ 2-for cheques outside India

(Please ask for proforma invoice/bill if required)

Send to

BRZARK INFORMATION SYSTEMS (P) LTD,
112, HUMAYUN PUR (NEAR N.C.C. OFFICE)
SAFDARJUNG ENCLAVE, NEW DELHI-110 029
TELEFAX / TELEPHONE + 91-11-688 2366