Jointly published by Elsevier Science Ltd, Oxford
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 37, No. 3 (1996) 401–416

THE FREQUENCIES OF OCCURRENCE OF SCIENTIFIC PAPERS WITH AUTHORS OF EACH INITIAL LETTER AND THEIR VARIATION WITH NATIONALITY

G. LEWISON

PRISM, The Wellcome Trust, 210 Euston Road, London NW1 2BE (England)

(Received July 16, 1996)

This paper introduces "alphabet spectra" which are the 26 frequencies of occurrence of scientific papers in a given sample with at least one author of each initial, A, B,...Z. The sum of these frequencies exceeds unity because of multiple authorships. Formulae are given relating this sum to the mean number of authors per paper in the sample. The method is applied to show the increase in this number over the last 15 years in different fields of science and for different countries. The "alphabet spectra" vary greatly depending on the nationality of the scientists concerned and can be compared to frequency absorption spectra for chemical elements or molecules. The spectra can be used to determine the national composition of a country's scientific authors and how this has changed with time.

Introduction

This investigation began with the need to solve a sampling problem but rapidly developed into a study in its own right. The original problem arose from a project[1] involving the definition of some biomedical subfields so that the relative output performance of different groups of scientists could be compared. The traditional method, whereby the area or subfield is defined by means of a set of specialist journals, is not satisfactory because it ignores the large number – often the majority – of publications in general journals. These can only be retrieved by means of a filter based on keywords. In some databases, these would be index terms, but often only titles are available and the filter would need to be based on title keywords. This is a practical procedure, but it depends on some individual expert or committee perusing a long list of papers and marking them "yes", "no" or "borderline", and so expressing their view on the boundaries of the chosen area of science.

The details of the procedure that would allow the size of a scientific subfield (i.e., the number of papers) to be determined are given elsewhere.[1] This investigation was concerned initially with the problem of sampling, so as to reduce the subfield expert's

0138–9130/96/US $ 15.00

task of marking to one of reasonable size, e.g., the provision of a set of a few hundred papers instead of one of several thousand or possibly tens of thousands. How could a given percentage of the records on a database be extracted by means of proprietary software that does not provide a random number facility or the possibility of a mathematically-structured sample (e.g., every 10th record)? Of course, the complete set of records could be downloaded into a spreadsheet and a random sample then taken, but this is a very cumbersome procedure and could incur high charges on some systems. What was needed was a simpler, readily available, system that would allow an unbiassed sample of specified size to be selected and which could be used either with on-line systems or with databases held on a CD-ROM.

Method

The solution that was found was to select papers from the overall population that had at least one author with a surname beginning with a given letter of the alphabet. Figure 1 shows the proportions, P, expressed as percentages, of the world total of papers that have at least one author with a surname or family name beginning with each of the 26 letters of the English alphabet in turn. It is based on 174763 papers in the *Science Citation Index (SCI)*, CD-ROM version, for January to March 1995. It shows that the sample size would vary between 29% for authors with names beginning with S(P_S=0.29) and 0.7% for those starting with X(P_X=0.007). Larger samples

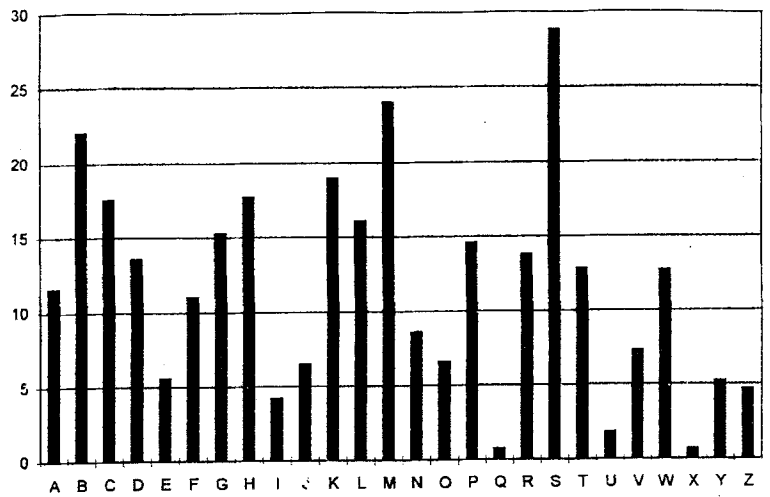

Fig. 1. Percentage of papers with an author with each initial letter, *SCI*, Jan to Mar 1995

could be obtained if two or more letters were combined by an "OR" operator, or smaller samples if they were combined by an "AND" operator. The latter procedure would usually be preferable to the use of a single unusual letter such as X or Q for the retrieval of small samples because there will be so few such authors that the sample will not be at all random.

The sum of the percentages shown in Fig. 1 is 303%, because most papers have several authors, and in fact the ratio, R, of this sum to 100% is an approximate value for the mean number of authors per paper. It is an under-estimate because papers will only be counted once even if there are two or more authors present with the same initial letter. Figure 2 shows the relationship between R and the mean number of authors per paper, $\underline{n}$, as determined by an analysis of nine groups of papers, typically each numbering several thousand (squares).

Although there is some scatter and the correlation is not perfect, the points mostly lie on or very close to the parabola:

$$\underline{n} = (2 + 5R + 2R^2)/9 \qquad (1)$$

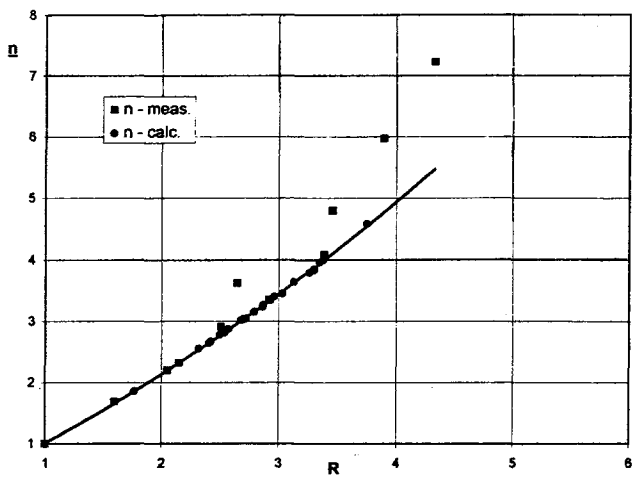

Fig. 2. Mean numbers of authors per paper, n, for different ratios, R, of sum of percentages of incidence of papers with authors of each initial letter to 100%

The relative under-representation of papers with authors whose names begin with popular letters such as S compared with papers with authors with uncommon initial letters such as E is endemic, for the reason given above, but the exact relationship depends on the distribution of the proportion of papers in the sample with 1, 2, 3...

authors. The relationship shown in Fig. 3 is typical, where the spots each represent one author's initial. The intercept of the curve (it is almost a straight line) is equal to the mean number of authors per paper, $\underline{n}$, and the initial slope is given theoretically by

$$-(n_2 + 3n_3 + 6n_4 + 10n_5 + 15n_6.....) \quad (2)$$

where n_i=the proportion of papers in the sample with just i authors and the coefficient of n_{i+1}=i+the coefficient of n_i. The actual mean slope, S, is plotted against the measured value of the mean number of authors per paper, $\underline{n}$, for seven sample groups of papers in Fig. 4. The points lie on or close to the parabola:

$$S = -1.32\,(\underline{n}-1) - 0.37\,(\underline{n}-1)^2 \quad (3)$$

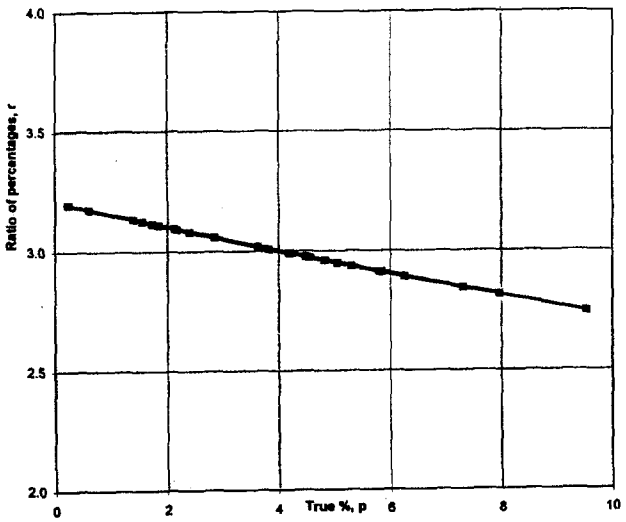

Fig. 3. Variation of ratio, r, of apparent percentages, P, of frequency of appearance of papers with authors with each initial to true percentages, p. (Mean authors per paper $\underline{n}$=3.21, slope S=–4.7)

The ratio, R, between the observed frequency of occurrence, P, of papers by an author with given initial and the proportion, p, of scientists with this initial in the population of authorships is given by

$$P = \underline{n}p + Sp^2 \quad (4)$$

and therefore

$$p = 0.5\,[-\underline{n} + (\underline{n}^2 + 4SP)^{0.5}]/S \qquad (5)$$

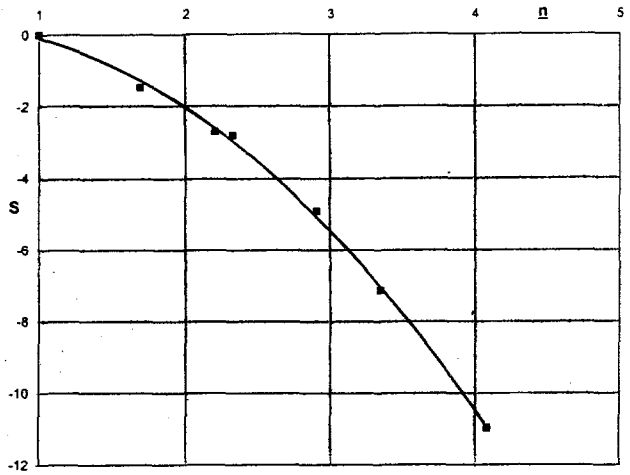

Fig. 4. Mean slope, S, of graph of ratio, r, of apparent %, P, against true %, p, of papers with authors of given initial

The calculation of $\underline{n}$ from the ratio, R, of the sums of the 26 individual gross frequencies of author initial occurrence, P, to 1 from Eq. (1) enables successively S and the 26 values of p to be determined for a given group of papers. Since the sum of the 26 values of p must equal unity, this affords a check on the accuracy of the original estimate of $\underline{n}$. In practice, the actual distribution of the proportion of papers in the group with 1, 2, 3... authors will vary and the consequence is that a slightly different value of $\underline{n}$ may yield a set of values p_A, p_B,... summing to unity.

The results of the application of this revised procedure to some 20 groups of papers taken from the *SCI* are shown in Fig. 2 as circles and they give a slightly different best-fit parabola:

$$\underline{n} = (1 + 7R + R^2)/9 \qquad (6)$$

This formula, rather than Eq. (1), is used in the following section as it is based on many more sample groups of papers. The value of $\underline{n}$ permits the conversion of a gross spectrum like Fig. 1 to a net spectrum in which the ordinates represent the actual percentage presence of authors with given initial in the relevant population of authorships. The result for the world population of scientists is shown in Fig. 5.

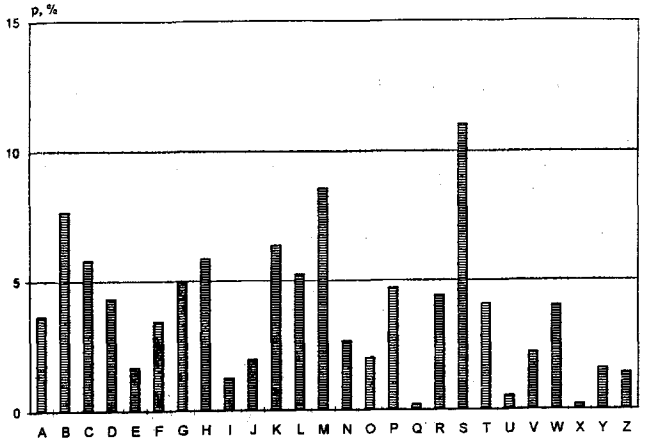

Fig. 5. Net alphabet spectrum (true percentage frequencies for papers with authors of each initial) for world papers in *SCI*, Jan to Mar 1995

Results

Authors per paper

Table 1 shows the mean number of authors per paper for five different disciplines in 1981, 1988 and 1995, calculated according to the procedure just described. The papers were retrieved from the *SCI* on the basis of address keywords, a method originally suggested by *de Bruin* and *Moed*[2] as a means of classifying papers by discipline. The mean number of authors per paper is highest in medicine and lowest in mathematics. The table shows that the number of authors per paper has increased in all disciplines over the 15-year period, but especially in biology, where the annual percentage rise (2.7% per year) is almost twice that in chemistry (just 1.6% per year).

There is also a variation in the number of authors per paper for papers from different countries. Table 2 shows the results of the analysis for the world and five countries (the UK, CN=China, JP=Japan, NL=Netherlands and ZA=South Africa) in 1981, 1988 and 1995. All show an increase in this number with time. The countries selected are not the extremes: in 1995 several European countries had a mean number of authors per paper greater than 5: Italy=5.6; Belgium and Switzerland=5.1. On the other hand India had a low number (2.95), suggesting either a low number of Indian authors per paper or a lack of international co-authorship.

Table 1
Variation of mean numbers of authors per paper with time for five scientific subjects, 1981–95, computed from differences between gross and net alphabet spectra for *SCI* papers

Subject	Address keywords	Mean number of authors/paper		
		1981	1988	1995
Mathematics	MATH, MATEMAT	1.68	1.86	2.18
Biology	ANIM, BIOL, BOT, PLANT, ZOOL	2.65	3.15	3.85
Physics	PHYS, FIS	2.81	3.23	3.83
Chemistry	CHEM, CHIM	3.02	3.34	3.79
Medicine	CLIN, HOP, HOSP, KLIN, MED, VET	3.40	3.96	4.59

Table 2
Variation of mean numbers of authors per paper with time for the world and for five countries, 1981–95, computed from differences between gross and net alphabet spectra for *SCI* papers

Code	Country name	Mean number of authors/paper		
		1981	1988	1995
UK	United Kingdom of GB and NI	2.55	3.02	3.64
CN	China (People's Republic)	2.91	3.48	4.08
JP	Japan	3.32	3.95	4.56
NL	Netherlands	3.16	4.18	4.77
ZA	South Africa	2.32	2.78	3.04
World	–	2.67	3.04	3.45

National distributions

When the distributions of the percentages of papers with an author with each initial letter are converted to net alphabet spectra for individual countries, some striking differences emerge. The spectrum for the UK is shown in Fig. 6. Comparison with the shape of the net alphabet spectrum for the world in Fig. 5 shows that certain letters such as W and J occur more frequently than the world average, but that letters such as I, K, U, V, X, Y and Z occur only about half as often as the world average. This

distinctive pattern is relatively constant over time although there is a tendency for the relative frequencies of occurrence gradually to approach unity as the amount of international co-authorship grows.

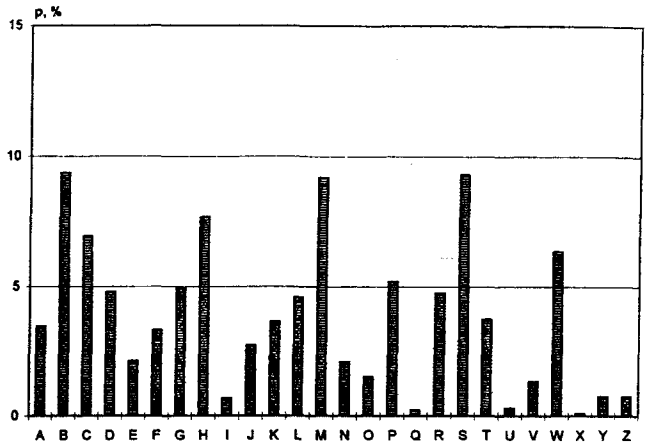

Fig. 6. Net alphabet spectrum (true percentage frequencies for papers with authors of each initial) for UK papers in *SCI*, Jan to Mar 1995

The net alphabet spectra for China, Japan, the Netherlands and the (scientifically) larger Hispanic countries of Latin America (Argentina, Chile, Mexico, Uruguay and Venezuela) are shown in Figs 7, 8, 9 and 10 respectively. The patterns are quite different from that of the UK. China shows a very high relative frequency of authors with names beginning with X, Q and Z. Japan has the highest occurrence ratios to the world average for names beginning with I, Y, U and O (all more than 3). The Netherlands has many names beginning with V, more than 5 times the world average. So each country has a distinctive pattern and the peaks at particular letters can be likened to the peaks of the ultraviolet absorption spectrum for chemical molecules. Now if a pure substance is analysed, the absorption peaks will be higher than for an impure one where they will be blurred. We might therefore expect that national alphabet spectra will become blurred because of immigration and because of international co-authorship.

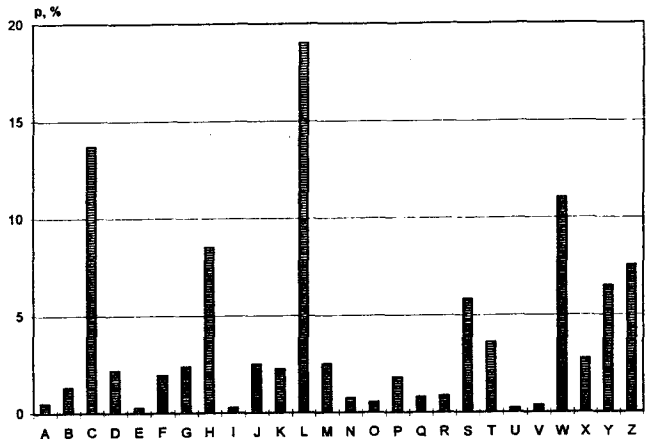

Fig. 7. Net alphabet spectrum (true percentage frequencies for papers with authors of each initial) for Chinese papers in *SCI*, Jan to Mar 1995

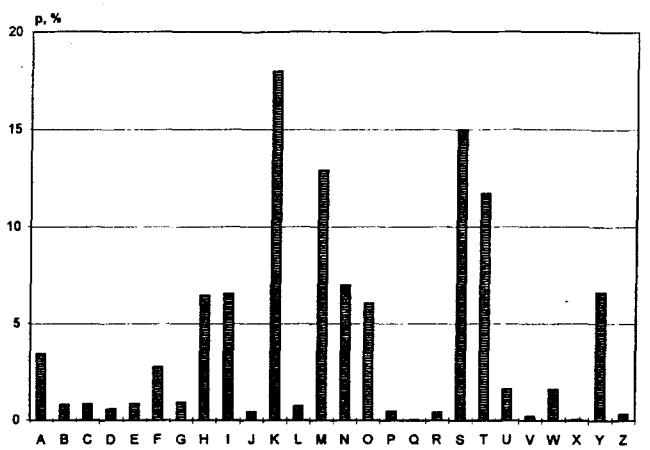

Fig. 8. Net alphabet spectrum (true percentage frequencies for papers with authors of each initial) for Japanese papers in *SCI*, Jan to Mar 1995

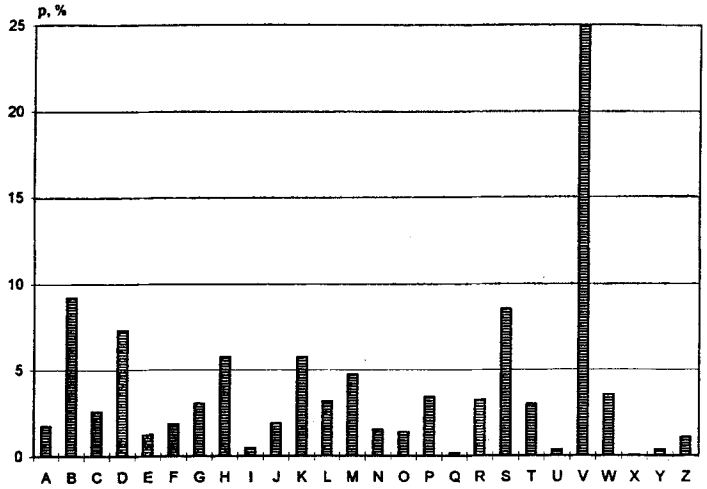

Fig. 9. Net alphabet spectrum (true percentage frequencies for papers with authors of each initial) for Netherlands papers in *SCI*, Jan to Mar 1995

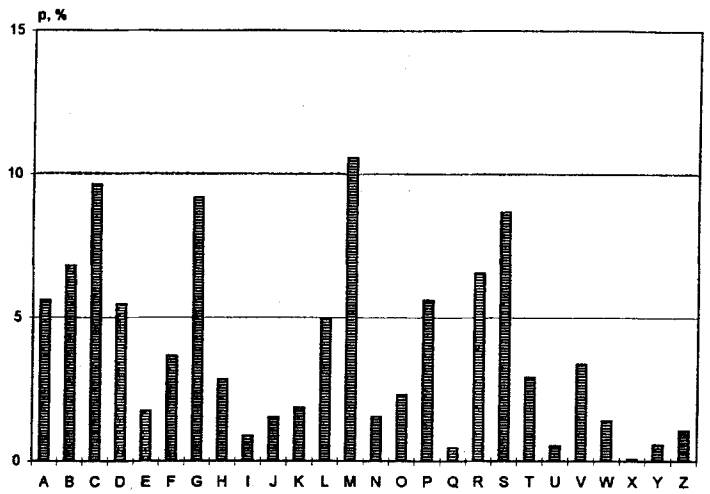

Fig. 10. Net alphabet spectrum (true percentage frequencies for papers with authors of each initial) for Latin American papers in *SCI*, Jan to Mar 1995

It is therefore possible to carry out, albeit in rather a crude way, an analysis of the national groups that make up a country's publishing scientists, just as if they were different chemical components of a mixture. The examples of Belgium (assumed to be a mixture of French and Dutch names) and South Africa (assumed to be a mixture of British and Dutch names) are interesting. The relative presence of the two components can be determined if the variances of the 26 individual differences between the p values for the "alphabet spectrum" of the mixture (BE; ZA) and those for estimated compositions based on different fractions of each of the two component "alphabet spectra" (FR=France and NL; UK and NL) are minimised. [For this purpose, papers were only retrieved from Belgium and South Africa that had no co-authorship addresses in France, NL or the UK, or in Germany or the US.] The best fits were obtained for BE with 40% FR and 60% NL, and for ZA with 80% UK and 20% NL, see Fig. 11. In both instances, the letter V acts as a decisive marker (or characteristic absorption frequency, by analogy) as it is so common in the NL and so rare everywhere else.

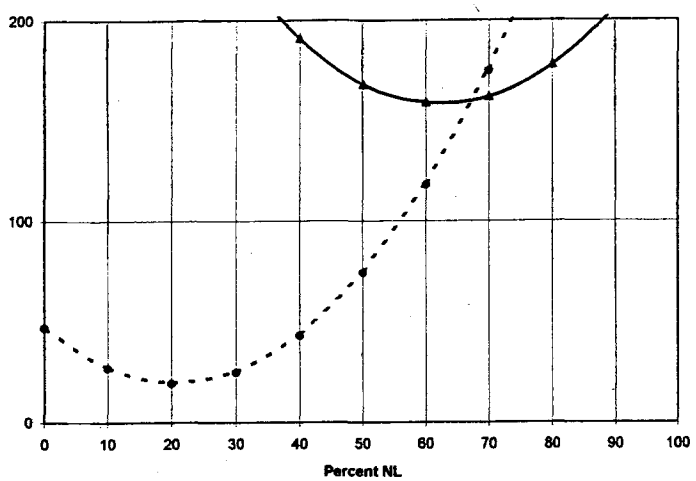

Fig. 11. Sum of variances for percentages of Belgian (triangles) and South African (circles) net alphabet spectral lines compared with FR+NL and UK+NL components

These percentages are not the proportions of scientists of each national origin but rather of authorships, i.e., an author is counted as many times as he/she has written a paper included in the sample. They will, however, approximate to the proportions of publishing scientists unless there is a bias in the *SCI*, such as an under-representation

of Afrikaans papers. The Belgian residual variance is much higher than that for South Africa, indicative of other national groups being present, whereas South Africa has had little addition to the British and Dutch groups.

A second exercise involved the determination of the national composition of authors from the three US west coast states, California, Oregon and Washington, in the three years 1981, 1988 and 1995, see Fig. 12. It might be expected that the composition of the population of publishing scientists would be similar to that of the UK, with an admixture of people of Latin American origin and, more recently, of Chinese and Japanese. The variance minimisation was first conducted between net alphabet spectra from the UK (Fig. 6) and Latin America (Fig. 10). This yielded a minimum for a Latin American presence rising from 8% to 14% over the period, Fig. 13a. For each optimum fit with respect to Latin American presence, the variance was again minimised with respect to the presence of Chinese authors (Fig. 7). This also appears to have risen over the period, from 2% to 7%; see the curves of Fig. 13b. Finally, for optimum values of Latin American and Chinese presence, the variance was minimised with respect to the presence of Japanese authors (Fig. 8); the results are shown in Fig. 13c and indicate that the Japanese presence has risen from 8% to 10% over the period. The resulting national compositions for the three west coast states in the years 1981, 1988 and 1995 are shown in Table 3.

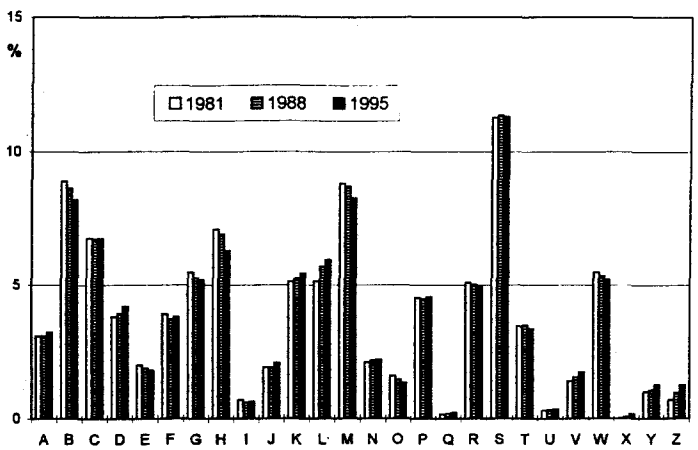

Fig. 12. Net alphabet spectra for three US west coast states (CA, OR, WA) in 1981, 1988 and 1995. (Note: papers co-authored with CN or JP excluded)

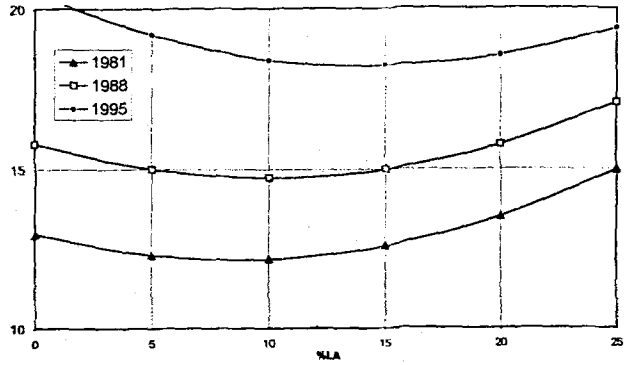

Fig. 13a. Sum of variances for percentages of US west coast states net alphabet spectral lines compared with varying %LA (remainder UK)

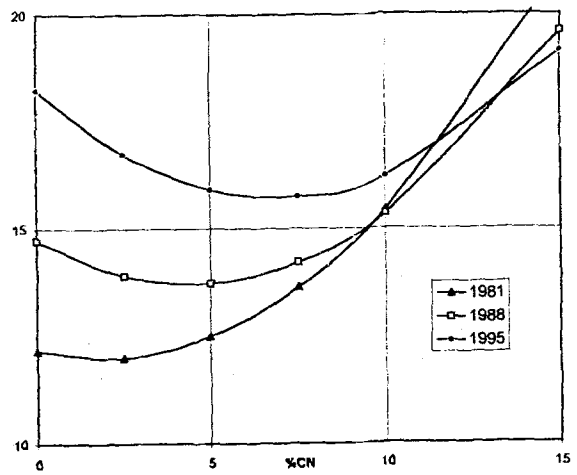

Fig. 13b. Sum of variances for percentages of US west coast states net alphabet spectral lines for optimum LA and varying %CN (remainder UK)

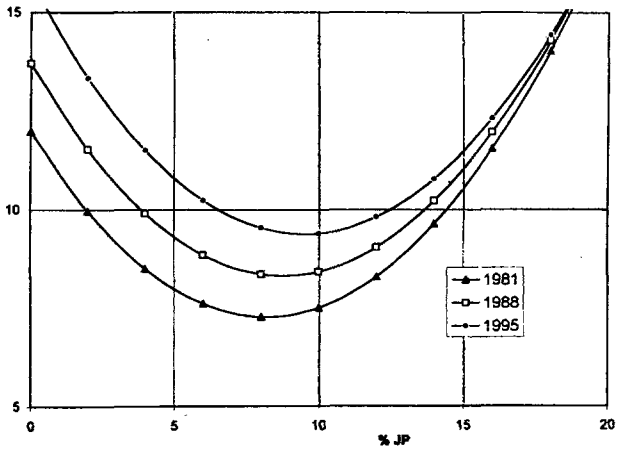

Fig. 13c. Sum of variances for percentages of US west coast states net alphabet spectral lines for optimum fit for LA and CN and varying %JP (remainder UK)

Table 3
Calculated national composition of publishing scientists from three US west coast states based on best fit of alphabet spectra

Year	1981	1988	1995
%UK	82	77	69
%LA	8	10	14
%CN	2	4	7
%JP	8	9	10
Total	100	100	100

It is notable that the minimum of the sum of the variances rises steadily from 1981 to 1995 in all three figures, indicating that there are increasing admixtures of publishing scientists of national origins other than the four groups discussed here. In this exercise, the papers from the west coast states co-authored with scientists with addresses in China (0.08%, rising to 0.73%) or Japan (0.54%, rising to 2.25%) were excluded. There were negligible numbers of papers (less than 0.4% in the 1990s) co-authored with Latin America. The alphabet spectra used for the variance minimisation were averages for those obtained from the three different years.

Discussion

The tendency for the mean number of authors per paper to increase with time has been described by many writers, notably *Smith*[3] and *White* et al.[4] for psychology; *Vlachý*[5] for physics; and *Price*[6] for chemistry. The present technique allows this number to be determined for a selected group of papers without the actual numbers of papers with 1,2,3... authors having to be counted. It seems to give results comparable with those obtained by the direct approach. For example, *Katz* et al.[7] at the University of Sussex found the mean numbers of authors on UK papers (articles, notes and reviews only, which would be expected to give a slightly higher value) to be 2.63 in 1981 and 3.34 in 1991, close to the values in the top line of Table 2. They also found that the rate of increase in the number of authors per paper for the UK was greatest in biology, which agrees with the result world-wide of Table 1.

The determination of national compositions on the basis of minimisation of the variances is less readily confirmed because there is a paucity of data on the national origins of publishing scientists. Within the US, there is currently a debate[8,9] on the merits or otherwise of allowing foreign-born scientists to settle there but it is clear that their numbers have been steadily increasing during recent years. The analysis by *Bouvier* and *Martin*[10] for 1995 suggested that Hispanics made up 2.8% of the scientific workforce, Chinese 2.5% and Japanese, 0.9%. These are lower by about half an order of magnitude than the ones suggested in Table 3. However they appear to underestimate the numbers of publishing scientists in the three west coast states with Chinese or Japanese names. Based on a scan of the authors of a 10% sample of the 1995 papers, 7.3% of the names looked to be Chinese in origin (standard error, 1.2%) and 1.8% of the names looked to be Japanese (s.e., 0.4%). The Chinese component of closely in line with that determined by the alphabet method for that year, but the Japanese component is much less, probably because of the admixture of Koreans and Vietnamese, whose alphabet spectra have not been investigated, but which may be similar to that of the Japanese. It is difficult from a visual scan to identify names of Latin American origin, so the difference between Bouvier's results and the present ones remains unresolved.

One possible way in which the alphabet spectra for a country could be confirmed would be through comparison with the distribution of names by initial letter in the general population. This could be approximated by the amount of space given to each letter in the telephone directory for the capital or other major cities, on the assumption that publishing scientists' names are similar to those of other people. A comparison between the net alphabet spectrum for the UK for 1995 and the London telephone directory, residential section, for the same year shows that the correlation between the

26 pairs of percentages is $r^2=0.99$. However this close correlation conceals a few striking differences, for example that the scientists whose names begin with X are almost all of Chinese origin whereas the London residents of the same initial are mostly Greek in origin.

This paper has provided details of a new method of analysis of scientific publications and has given some examples of how it can provide useful and interesting results. It can surely be refined and placed on a more systematic basis, and it can probably be applied to other areas of human activity where long lists of surnames are involved.

References

1. G. LEWISON, The definition of biomedical research subfields with title keywords and application to the analysis of research outputs, to be submitted to *Research Evaluation*.
2. R. E. DE BRUIN, H. F. MOED, Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications, *Scientometrics*, 26 (1993) 65–80.
3. M. SMITH, The trend toward multiple authorship in psychology, *American Psychologist*, 13 (1958) 596–599.
4. K. D. WHITE, Authorship patterns in psychology: national and international trends, *Bulletin of the Psychonomic Society*, 20(4) (1982) 190–192.
5. J. VLACHÝ, Physics journal in retrospect and comparisons, *Czechoslovak Journal of Physics*, 20B (1970) 501–526.
6. D. J. DE S. PRICE, *Little Science, Big Science*, Columbia University Press, 1963.
7. J. S. KATZ, D. HICKS, M. SHARP, B. R. MARTIN, N. LING, *The Bibliometric Evaluation of Sectoral Scientific Trends*, University of Sussex, Science Policy Research Unit, Report to Economic and Social Research Council, 1995.
8. R. FINN, Scientists' heated debate on immigration mirrors issues argued throughout US, *The Scientist*, 9, No. 23 (27 November 1995) 1, 8–9.
9. D. S. NORTH, *Soothing the Establishment: The Impact of Foreign-born Scientists and Engineers in America*, University Press of America, Lanham MD, 1995.