National Research Priorities in a Global Perspective:

A Bibliometric Analysis

ANJU CHAWLA
National Institute of Science, Technology & Development Studies
New Delhi- 110012 (India)
Email: Anjuc_2000@yahoo.com

Abstract

This study examines inter-country similarities and differences in the priority accorded to eight macro-fields of science (Clinical medicine, Biomedical research, Biology, Chemistry, Physics, Earth and space sciences, Engineering & technology and Mathematics). This study is based on the contribution of top 30 countries to the mainstream scientific literature indexed by the Institute for Scientific Information (ISI). Obviously, raw counts of publication are confounded by the size of research fields and the size of the countries. Hence an index of research priority was constructed for cross-national comparison... A novel graphic technique, viz. Parallel Coordinates, was used for visualizing the priority profiles of the countries. Cluster analysis and multi-dimensional scaling were used to construct a typology of countries based on similarities of their priority profiles. Implications of this study for science policy are briefly discussed

1. Introduction

The publication pattern of a nation is an indicator of its capacity and commitment to perform mainstream research in different fields of science. The output of publications in different fields is not a random event; it is the cumulative effect of resource allocation and policy decisions for different fields of science, taken in the past, whether implicitly or explicitly. In this paper, we examine the research portfolios of 50 countries, which had published at least 500 articles in mainstream journals, indexed in the ISI database. These countries account for approximately 99% of the world output of scientific literature.

2. Objectives

This study has two major objectives:

- To identify priorities and potential holes in the research portfolios of different countries.
- To construct a typology by classifying the countries into groups characterized by the similarities of their research profiles. Typologies provide parsimonious descriptions of the data, which are useful for further discussion and research.

3. *The Data*

The data on publication output of 50 most prolific countries in eight macro fields during 2001 were taken from the most recent "Science and Engineering Report (National Science Foundation, USA)". The macro fields are: Clinical Medicine (CLI), Biomedical research (BIM), Biology (BIO), Chemistry (CHM), Physics (PHY), Earth & Space (EAS), Engineering & Technology (ENT), and Mathematics (MAT). The names of the countries and their triliteral codes are given in the Appendix.

4. *Analysis and Results*

4.1. *Research priorities*

Since, the raw data on publication counts are confounded by the size of the countries and the size of the research fields, an index called "research priority index (*PI*) "was computed according to the following formula:

$$PI = \frac{n_{ij}/n_{i.}}{n_{.j}/n_{..}}$$

where

n_{ij} = the number of publications of country i in field j

$n_{i.}$ = the number of publications of country i in all fields

$n_{.j}$ = the number of publications of all countries in field j

$n_{..}$ = the number of publications of all countries in all fields

Here all refers to the comparison set of 50 countries. Note that index *PI* is identical to the activity index, proposed by Schubert and Braun[1.]

$PI = 100$ indicates average priority

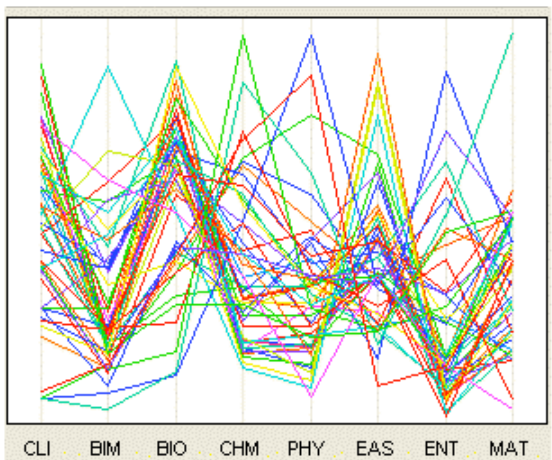
$PI < 100$ indicates less than average priority

$PI > 100$ indicates above average priority

There are wide variations among the countries in the emphasis given to different fields, depending upon their historical traditions, scientific capacity and socio-economic goals. The priority profile of a country can be represented by a point in an 8-dimensional Euclidean space, but the visualization of multidimensional data is difficult and non- intuitive. A novel graphical technique, viz. *Parallel Coordinates*[2] was used to visualize the priorities and potential holes in the research agenda of various countries In traditional Cartesian coordinates, all axes are mutually perpendicular. In Parallel coordinates, all axes are parallel to one another and equally spaced. A single horizontal line is drawn and a series of vertical axes, each representing a separate variable, are placed at equal distances along the line. The values of a given variable are represented on the vertical axis pertaining to that variable. The values on each of the *N* axes that correspond to an individual point in *N*-dimensional Euclidean space are connected by line segments between successive vertical axes. The result is a graph of line segments connected between axes to form polygonal lines across the entire representation. Each polygonal line of (*N*-1) segments represents a distinct point in the *N* -dimensional space.

The priority profiles of these countries are depicted (in the format of parallel coordinates) in Fig. 1.

Fig.1 Parallel plot of research priorities



It can be easily seen that there are considerable variations among the countries in the emphasis given to different fields. Inter-country differences in the priority accorded to Biomedical research are much greater than those accorded to Chemistry, Physics, Engineering and Technology, Mathematics, Earth & Space sciences, Clinical medicine and Biology in that order.

4.2. Construction of Typology

The countries were classified into homogeneous groups based on the similarities of their research priorities. Clustering algorithms in popular statistical packages (SPSS, SYSTAT, STATISTICA, MINITAB, etc) suffer from certain important limitations[3]. They do not provide any guidance as to:

- How to determine the optimal number of clusters?
- How to judge whether an object has been properly assigned to a particular cluster or not?
- How to distinguish a good cluster from a bad cluster?
- How to judge the quality of cluster structure? Is it real or only an artifact of the clustering algorithm?

In this paper, we have used a clustering algorithm PAM (Partition around Mediods).implemented in WinIDAMS[4] , NCSS[5] and SPLUS[6]. The algorithm is well described in Kauffman and

Rousseaw[7.] PAM has several advantages over the well known *k*-means clustering algorithm

- It is more robust, because it minimizes the sum of dissimilarities instead of the sum of squared Euclidean distances.

- It provides a novel graphical display, the *silhouette plot*, which provides key information for deciding the optimal number of clusters and also for judging the quality of cluster structure.

A silhouette measures how well an object has been classified by comparing its dissimilarity within its cluster to its dissimilarity with its nearest neighbour. It is computed as follows:

Consider an object $i \in$ Cluster A.
Let a (i) = Average dissimilarity of i to all other objects in A.
Let b (I) = Average dissimilarity of I to all objects in the neighbouring cluster B
Silhouette $S(i)$ is computed by the following formula:

$$: S(i) = \frac{[b(i) - a(i)]}{\text{Max}\,[a(i), b(i)]}$$

$S(i)$ ranges between −1 and + 1. Silhouette value close to +1 indicates that the object has been well classified. Silhouette value close to zero means that the object has been arbitrarily classified; in other words it lies between clusters A and B. Silhouette value close to −1 implies that the object has been misclassified. The silhouette plot shows which objects lie well within the cluster and which ones are arbitrarily or wrongly classified. A useful summary statistic is the average value of S across all objects. This is called Silhouette coefficient (SC), which summarizes how well the cluster structure fits the data. An easy way to select the appropriate number of clusters is to choose that number of clusters which maximizes the average silhouette: Rousseaw[8] has suggested the following thumb rule for interpreting the silhouette coefficient.

Range of SC	*Interpretation*
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial.
< 0.25	No substantial structure has been found

A series of cluster analyses were performed with the number of clusters ranging from 2 to 10. The results are summarized in Table 1. It can be easily seen that the 2- cluster solution yields the highest value of silhouette coefficient, but that classification would be rather too broad for subsequent elaboration and interpretation. Hence, we have opted for the 7-cluster solution, which yielded the next highest value of silhouette coefficient.

Table 1: Silhouette coefficients for different cluster solutions

Number of clusters	Average silhouette (Silhouette coefficient)	Number of clusters	Average silhouette (Silhouette coefficient)
2	0.585902	7	0.467726
3	0.396241	8	0.444091
4	0.324119	9	0.351605
5	0.418655	10	0.344704
6	0.455940		

The assignment of countries to different clusters and quality of cluster membership and cluster structure is indicated in Table 2.

Table 2 Assignment of countries to different clusters and quality of assignment

Country	Nearest neighbor	Average distance within	Average distance neighbor	Silhouette value	Silhouette bar
Cluster 1					
PRC	6	21.41	27.51	0.398880	|IIIIIIIIIIIIIIIIIIII
ROM	7	37.73	46.01	0.324180	|IIIIIIIIIIIIIIII
POL	4	22.42	25.41	0.211680	|IIIIIIIIIII
IRN	5	30.83	34.19	0.098100	|IIII
BGR	4	23.92	19.69	− 0.176505	|
Average		*27.26*	*30.56*	*0. 171268*	
Cluster 2					
ZAF	3	20.70	37.13	0.796680	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII
NZL	3	30.81	49.44	0.678420	|IIIIIIIIIIIIIIIIIII
AUS	3	17.39	27.01	0.641340	|IIIIIIIIIIIIIIIIII
NOR	3	20.71	28.51	0.492840	|IIIIIIIIIIIIII
ARG	3	21.58	27.85	0.404820	|IIIIIIIIIII
THA	3	28.00	33.36	0.404820	|IIIIIIII
MEX	3	24.88	28.71	0.240120	|IIIIIII
CHL	4	26.37	29.55	0.194040	|IIIII
DEN	3	21.30	20.35	−0.04493	|
CAN	3	21.23	18.94	−0.108247	|
Average		*23.30*	*30.09*	*0.358442*	
Cluster 3					
UK	2	14.50	28.02	0.868500	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
NLD	2	15.33	29.36	0.860220	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
SWE	2	15.12	28.31	0.838800	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
USA	2	15.49	28.63	0.826020	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
BEL	4	13.24	23.32	0.778320	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
CHE	4	16.08	27.14	0.734040	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII
AUT	4	16.22	26.08	0.680580	|IIIIIIIIIIIIIIIIIIIIIIIIIII
DEU	4	15.10	22.33	0.582300	|IIIIIIIIIIIIIIIIIIIIIII
TUR	5	21.27	31.21	0.573120	|IIIIIIIIIIIIIIIIIIIIII
ITA	4	17.44	22.61	0.411660	|IIIIIIIIIIIIIIII
FIN	2	17.79	22.21	0.358200	|IIIIIIIIIIIIII

JPN	5	20.45	23.73	0.248220	|IIIIIIIIIII
GRC	4	19.55	21.33	0.150840	|IIIIIII
ISR	4	21.80	23.76	0.149040	|IIIIII
BRA	4	21.69	21.37	−0.026820	|
IRL	2	24.00	23.37	−0.047160	|
Average		*17.82*	*25.17*	*0.499117*	
Cluster 4					
CZK	6	14.99	27.97	0.835380	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
HUN	7	170.84	280.76	0.683640	|IIIIIIIIIIIIIIIIIIIIIIIIII
42HUN	4	180.26	280.37	0.641520	|IIIIIIIIIIIIIIIIIIIIIII
POR	5	180.72	280.09	0.600300	|IIIIIIIIIIIIIIIIIIIIII
VEN	4	180.06	260.61	0.578520	|IIIIIIIIIIIIIIIIIIIII
ESP	3	160.51	220.72	0.491940	|IIIIIIIIIIIIIIIIIII
CRO	3	180.96	240.23	0.391500	|IIIIIIIIIIIIIIII
FRA	3	210.29	240.41	0.229500	|IIIIIIIIII
Average		*180.08*	*260.40*	*0.55654*	
Cluster 5					
EGY	6	150.88	360.17	0.976350	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IND	4	1500.88	2700.40	0.756540	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
Average		*1500.88*	*3100.79*	*0.866445*	
Cluster 6					
SGP	6	2300.16	4400.21	0.857160	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
YUG	7	1800.48	3300.38	0.803520	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
TWN	4	1800.60	2900.14	0.651240	|IIIIIIIIIIIIIIIIIIIIIIIIII
SLV	5	1800.57	25,24	0.475560	|IIIIIIIIIIIIIIIIIII
KOR	4	2000.90	2400.98	0.293760	|IIIIIIIIIII
SAU	3	2600.11	3000.18	0.293760	|IIIIIIIIIII
Average		*2000.97*	*3100.19*	*0.562500*	
Cluster 7					
UKR	1	3000.18	4200.87	0.532980	|IIIIIIIIIIIIIIIIIIIII
BRS	1	2600.63	3400.75	0. 420660	|IIIIIIIIIIIIIIII
RUS	1	2700.19	3100.82	0.420660	|IIIIIIIIIIIIIIII
Average		*2800.00*	*3600.48*	*0.458100*	

This table is self-explanatory. However, certain prominent features of the cluster structure are summarized below.

- Overall cluster structure is reasonable.
- *Cluster 1* is poorly defined; its silhouette coefficient (0. 171268) is quite low. Two countries, viz., Iran and Bulgaria are arbitrarily assigned to this cluster. Average distance of these countries within their own clusters is greater than that with their neighboring clusters

- *Cluster* 2 is weak, perhaps arbitrary. Two countries (Denmark and Canada) are arbitrarily assigned to this cluster.
- *Clusters 3, 4, 6* and *7* have more or less reasonable structures. Two countries (Brazil and Ireland) are arbitrarily assigned to their cluster.

Fields of emphasis and de-emphasis of different clusters can be visualized from the parallel plots depicted in Fig.2. Salient features of these clusters are described below:

- *Cluster* 1: Chemistry, Physics and Mathematics are prominent fields. Clinical medicine, biomedical research, biology and earth and space science are fields of relative de-emphasis
- *Cluster* 2: Biomedical research and earth and space are prominent areas of research
- *Cluste*r 3: Biomedical research and Physics receive relatively greater emphasis.
- *Cluster* 4: Prominent field of research are Biomedical research, Physics and Mathematics.
- *Cluster* 5*:* Chemistry and Engineering & Technology receive greater priority in this cluster.
- *Cluster 6:* High priority to Engineering& Technology; about average priority to all other fields.
- *Cluster 7:* High priority to Engineering& Technology and Physics.

4.3. Visualization of Cluster structure

Metric multidimensional scaling (MDS) algorithm was used to project the 8-dimensional data onto a

2-dimensional plot, The minimum stress value was equal to 0.206)). Stress can be reduced by increasing the dimensionality of projection or by using non-metric MDS (for ordinal data).which seeks to preserve rank order of objects and not inter-object distances in the high-dimensional space. Increasing the dimensionality of projection complicates the display and should be avoided unless the stress values are greater than the acceptable threshold (*viz*. 0.20). Moreover, the relationship between dissimilarities and inter-point distances in the MDS plot was found to be linear. Hence, we did not resort to non-metric MDS.

Figure 3 represents a two dimensional configuration of multivariate relations among the countries. In this figure, the countries are represented by circles of different colours to indicate the cluster to which they have been assigned, and of different size to indicate the quality of their assignment. The MDS plot more or less validates the cluster structure issued by PAM

5. Discussion

Comparative analysis of research priorities, particularly the identification of fields of research that need to be emphasized or downsized has important implications for science policy. Policy-makers are frequently confronted with such questions: What priorities are being given to different fields or subfields of science and how do they compare with other countries? This paper, though exploratory in nature, has attempted to address these questions. The methodological framework and analysis presented in this paper has also implications for identifying partners for bilateral or multilateral cooperation in science

Research priorities can also be assessed through input indicators like the distribution of scientific manpower among different fields or allocation of financial resources to different field of science. But the data on these indicators are not available for several countries Further, the data on financial resources, if available, are not amenable to direct comparison, since they are affected by the difference in the local.

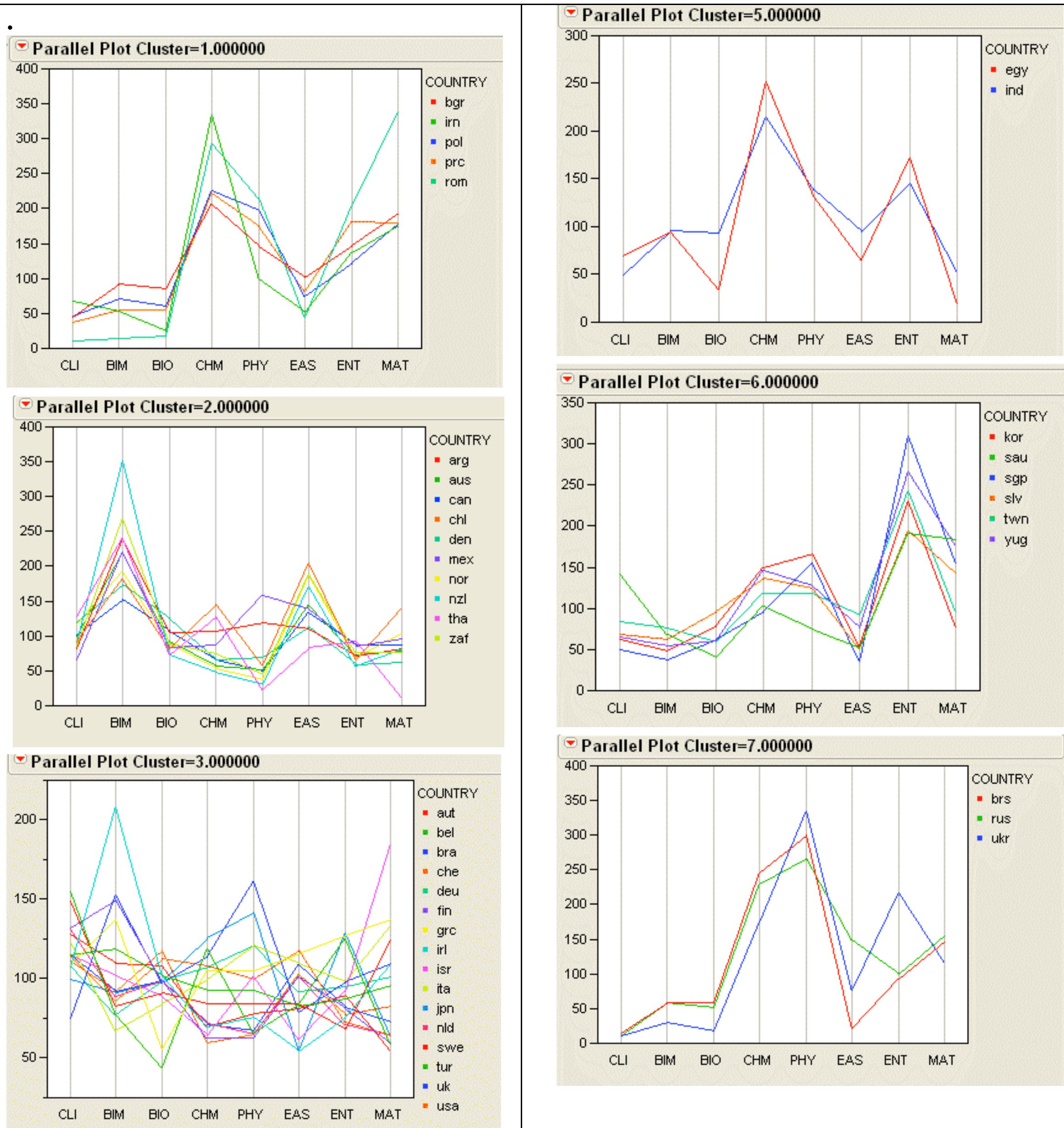


Fig.2 Parallel plots of research priorities of counties in different clusters

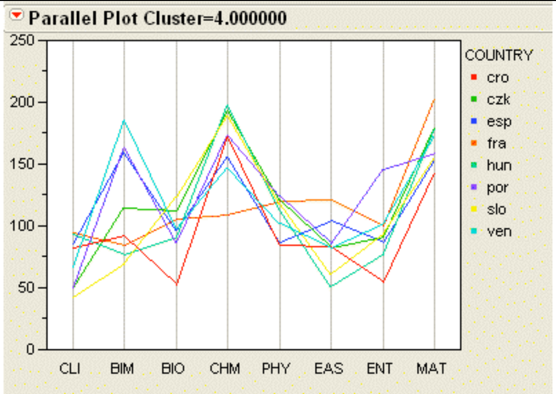
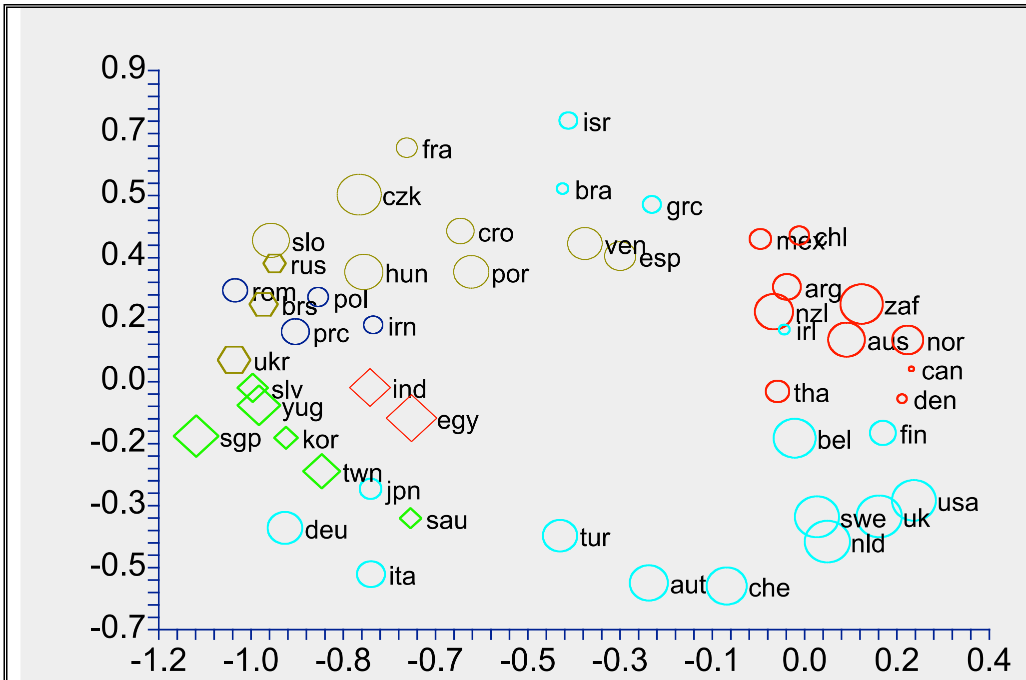

Parallel Plot Cluster=4.000000
250
200
150
100
50
0
CLI
BIM
BIO
CHM
PHY
EAS
ENT
MAT
COUNTRY
cro
czk
esp
fra
hun
por
slo
ven

Fig. 3 : MDS plot



- Countries located near a symbol of the same design and colour belong to the same cluster.
- Symbol size is proportional to silhouette width

1.

values of different countries[9]. On the other hand, the framework and methodology for cross-national comparison of research priorities do not suffer from these limitations. Bibliometric data can be collected fairly easily and are amenable to direct comparison[10]

References

1. Schubert and T. Braun, Relative indicators and relational charts for comparative assessment of publication output and citation impact, Scientometrics, 14, 1986.
2. A. Inselberg and B. Dimsdale , Parallel coordinates: A tool for visualizing multi-dimensional geometry, Visualization, San Francisco, CA., 1990.
3. P.S. Nagpaul (2001) *Guide to Advanced Data Analysis Using IDAMS Software*, http://www.unesco.org\Idams
4. WinIDAMS, Internatioally developed data analysis and management software package, UNESCO, Paris.
5. . J. HINTZE (2001) NCSS: Number Cruncher Statistical Systems, Kaysville. Utah
   SPLUS, Insightful Corporation, Seattle, Washington
6. L. Kaufman, P.J. Rousseuw (1990) Finding Groups in Data: An Introduction to Cluster Analysis, New York: John Wiley & Sons, Inc
7. P.J. Rousseeuw (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20: 53-65
8. P.S. Nagpaul and S. Bhattacharya, National Pattern of research output and priorities in physics: A bibliometric analysis, In P.S. Nagpaul et al. (Eds) Emerging trends in scientometrics, Allied Publishers, New Delhi (India), 1999.
9. B.R. Martin and J. Irvine, Assessing basic research: Some partial indicators of radio- astronomy, Research Policy, 12, 1983, 61-90,

Appendix

List of countries and their triliteral codes

United States	usa	Greece	grc

Japan	jpn	Norway	nor
United Kingdom	uk	Mexico	mex
Germany	deu	Argentina	arg
France	fra	New Zealand	nzl
Canada	can	Czech Republic	czk
Italy	ita	Singapore	sgp
China	prc	Hungary	hun
Russia	rus	South Africa	zaf
Spain	esp	Ukraine	ukr
Australia	aus	Portugal	por
Netherlands	nld	Ireland	irl
India	ind	Egypt	egy
South Korea	kor	Chile	chl
Sweden	swe	Romania	rom
Switzerland	che	Iran	irn
Taiwan	twn	Slovakia	slo
Brazil	bra	Slovenia	slv
Israel	isr	Bulgaria	bgr
Belgium	bel	Thailand	tha
Poland	pol	Croatia	cro
Finland	fin	Saudi Arabia	sau
Denmark	den	Ygoslavia	yug
Austria	aut	Venezuela	ven
Turkey	tur	Belarus	brs