Measuring research output and collaboration in South Asian countries

*Ashraf Uddin and Vivek Kumar Singh**

This article presents a scientometric analysis of academic research output, growth trend, citation & impact, and research collaboration levels in the South Asian region. The analysis is done on several important parameters such as total research production, global share and rank, subject categories, citation impact, in and out-region citation patterns, and inter-country collaborations. The economic indicators relating to higher education and research for the countries in the region are correlated with the analytical results. It also analyses the research growth and maturity levels for the region. In summary, it tries to map the academic research status in the South Asian region, including details about the countries in the region.

Keywords: Citation analysis, knowledge creation, research collaboration, research impact, scientometrics.

DURING the last decades the South Asian region has started focusing more on the higher education and research sector. Here the term 'South Asia' refers to countries in the South Asian region that are part of the South Asian Association for Regional Cooperation (SAARC)[1], which was established in 1985 with Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka as members; Afghanistan joined as a member in 2007, making the number of countries in the association as eight. As outlined in the SAARC charter[2], its key objectives are to promote active collaboration and mutual assistance in the economic, social, cultural, technical and scientific fields among the member countries. The countries in the region have started paying more attention to the higher education and research sector during the last decades, as measured in terms of policies and funding. There has been a positive impact of these efforts; however, the South Asian countries continue to struggle to meet the demand of the 318 million 15–24-year-olds for higher education access[3]. The public funding in the higher education sector is still less compared to the needs. A detailed statistics of population, GDP, public funding for higher education and research, and gross enrollment in higher education is given in Table 1. There are large variations in population and GDP figures. However, all the countries spend less than 10% of their GDP on the higher education and research sector. There are also variations in facilities and quality of the institutions of higher learning in the region, which is also an important aspect that requires detailed analysis. It is in this context that we analyse the research production, its growth and impact in the South Asian region during the last 50 years. For this, we collected data for the last 50 years' publications indexed in *Scopus*[4]. We performed a detailed computational analysis of the data from the scientometric and network-theoretic viewpoints to mine useful inferences, such as total research output and growth pattern of the region during the last 50 years, citation counts and their impact, research maturity levels, research collaboration patterns, etc. Some of the past works[5–9] have attempted to do a somewhat similar scientometric analysis for a specific country and/or domain with a more focused area/domain. We have done a scientometric and network-theoretic analysis for the entire South Asian region, which requires much more effort for data collection, filtering and making inferences. We present a systematic, wider and detailed analysis of bibliographic data pertaining to the South Asian region.

Bibliographic data collection

We have collected bibliographic data from *Scopus*, which has more than 50 million records (as on August 2013). The collection pertains to the documents published from 1856 onwards and includes documents of different types, namely article, conference paper, review, letter, article in press, note, editorial, short survey, chapter, erratum and book. Since our primary aim was to perform the scientometric analysis of publications originating from institutions belonging to South Asia, we have filtered data corresponding only to the institutions belonging to the South Asian region. Our collection, thus, contains all the records in *Scopus* from the beginning till the year 2013, which originate from the institutions of the South Asian region. We got a total of 1,286,092 records, with at least

Ashraf Uddin and Vivek Kumar Singh are in the Department of Computer Science, South Asian University, Akbar Bhawan, Chanakyapuri, New Delhi 110 021, India.
*For correspondence. (e-mail: vivekks12@gmail.com)

Table 1. Economic and educational statistics of the South Asian region

Country	Population[a] (in 2012)	GDP[b] (in 2012; billion US$)	Higher education expenditure per student (% of GDP per capita)[c]	Gross enrollment in higher education (% of total population)[d]
Afghanistan	29,824,536	20.49	1.73 (1982)	3.74 (2011)
Bangladesh	155,000,000	116.35	2.233(2009)	13.15 (2011)
Bhutan	741,822	1.78	4.653 (2011)	9.43 (2012)
India	1,240,000,000	1858.74	3.16 (2011)	23.27 (2011)
Maldives	338,442	2.22	6.82 (2011)	13.18 (2008)
Nepal	27,474,377	18.96	4.72 (2010)	14.49 (2011)
Pakistan	179,000,000	225.14	2.13 (2012)	9.53 (2012)
Sri Lanka	20,328,000	59.42	1.72 (2012)	16.97 (2012)

[a] http://data.worldbank.org/indicator/SP.POP.TOTL
[b] http://data.worldbank.org/indicator/NY.GDP.MKTP.CD/countries?display=default
[c] http://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS/countries?display=default
[d] http://data.worldbank.org/indicator/SE.TER.ENRR/countries?display=default

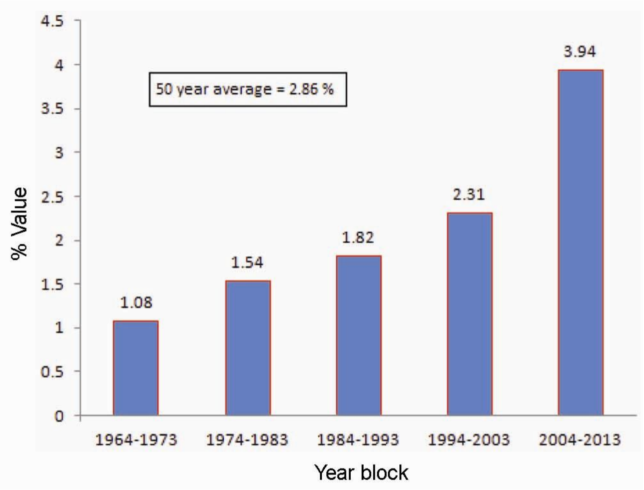

Figure 1. Research output share of the South Asian region.

one author affiliation to an institution belonging to one of the South Asian countries. Some duplicate entries were found in the process, which were removed. The resulting dataset comprised of records for 1,070,662 documents. Every record in the data contains 40 fields, with entries describing the basic meta data of documents such as authors, title, publication year, source title, abstract, references, citation, etc.

The data have been collected through an institution-based search for all the eight countries. The affiliation search interface of *Scopus* lists all the institutions of a country when searched with a specific country. The number of institutions from these countries listed in *Scopus* (as on February 2014) is: Afghanistan (1), Bangladesh (105), Bhutan (2), India (2315), Maldives (9), Nepal (63), Pakistan (210) and Sri Lanka (43). These figures are thus somewhat different from a country-wise search, which lists all records with the country name in the address part. For example, in case of India, a country-wise search results in 1,080,207 records as against 1,123,272 records returned through institution-based search, for the period 1964–2013. The former figure is obtained using the search query 'AFFILCOUNTRY (India)'. Thus, the country-wise search value is lesser than the institution-based search value for India. This is due to the fact that certain documents are the output of collaborative effort between multiple institutions (and hence are counted once for each collaborating institution). Similarly, for certain countries like Afghanistan and Maldives, country-wise search figures are higher than institution-based search figures (since not all documents have institutional affiliation mentioned).

Research output and share of South Asia

The eight South Asian countries taken together have 21% of the world's population. It may be relevant to measure both the absolute number and the share of research production of the region. We used the data for the last 50 years (1964–2013) to compute measures of the total research output of the region and the entire world. We found that a total of 1,042,846 records correspond to the South Asian region as against a total of approximately 44 million records for the entire world. Thus, approximately 2.86% of the research output of the world during the last 50 years is from the South Asian region. To do a deeper level of analysis, we divided the 50-year period into five blocks of 10 years each, with the resulting blocks as 1964–1973, 1974–1983, 1984–1993, 1994–2003 and 2004–2013. The main reason for this was to identify and measure the growth trend in research production of the region over the last 50 years. We found that a total of 20,645, 64,650, 109,342, 212,084 and 636,125 unique records correspond to the South Asian region for the five blocks respectively (in chronological order). When we compare this with the statistics for the entire world, it comes out to be 1.08%, 1.54%, 1.82%, 2.31% and 3.94%, share of total for the five blocks respectively (in chronological order). Figure 1 illustrates the research contribution of

Table 2. Research output of the South Asian region during the last five decades

Period/country	1964–1973	1974–1983	1984–1993	1994–2003	2004–2013	Total
Afghanistan	14	10	1	2	35 (1.300)	62
Bangladesh	70	551	1618	5282	20,730 (140.067)	28,251
Bhutan	0	0	0	6	37 (53.790)	43
India	20,937	65,282	109,195	227,641	700,217 (595.929)	1,123,272
Maldives	0	0	0	18	26 (82.415)	44
Nepal	1	50	163	1036	4,563 (174.172)	5,813
Pakistan	494	1,088	3197	10,819	69,783 (417.862)	85,381
Sri Lanka	115	543	1181	2425	7,391 (371.752)	11,655
Total	21,631	67,524	115,355	247,229	802,782	

Values in parenthesis indicate research output per million inhabitants.

the institutions in the South Asian region vis-à-vis the output for the entire world, for the five consecutive chronologically ordered blocks. The results show that the overall contribution from the South Asian region has witnessed a continuous upward trend from the first block (1964–73) to the second block (1974–83) and so on, with the most recent decade accounting for 3.94%. But overall the output share remains around 3%. Though the last 50 years have witnessed increased public funding in the sector, expansion of the higher education sector with private universities being set up, and a substantial increase in enrollment, the South Asian region is contributing only about 3% to the total research output of the world.

Country-wise research output and growth

The South Asian countries have different levels of infrastructure and provisions for higher education, which also reflects in the research production figures of individual countries. We have computed the country-wise research output for the South Asian region for the five consecutive blocks. There were some issues in this. The records do not have an explicit field for country. We have, therefore, parsed the author details field (which states the author affiliation) to identify the country for a research paper. Further, some (though very less) of the records did not have the exact value for the year field and so we discounted those entries. There are documents with multiple country affiliations (such as a collaborative paper from India and Bangladesh); so they are counted as entries for all the affiliating countries (resulting in a collaborative paper counted more than once in the output). Table 2 shows the country-wise research production statistics. It may be observed that the sum of these figures is higher than the total number of unique entries, due to repetition of some records for countries. The results show that Bhutan and Maldives did not have any research paper entry in the first three decades. Afghanistan's contribution was remarkably less for a long period, but has a upward trend now. India dominates the figures in terms of number of publications (accounting to approximately 90%). Pakistan, Bangladesh and Sri Lanka seem to be on a progressive path. The research output should also be measured with respect to population of the concerned country. Therefore, the second last column of the table shows (within brackets) research output figures per million inhabitants. This is computed by dividing the research output during the 2004–2013 block by the average population during this period. The average population for the period is computed by averaging the population figures of the countries during 2004 and 2013 (ref. 10). We have shown research output per million habitants only for the last block (2004–2013) since we do not have reliable population figures for the earlier period for all the countries. Table 3 shows the relative position of South Asian countries compared to rest of the world. The table shows data for the top 10 countries in terms of research output (during the period 1996–2012) and also the rank and data for countries in South Asia. Among the South Asian countries, only India figures in the top 10 list (at the tenth position). Other South Asian countries are placed from 46th to 191st rank.

Category-wise research output and growth

The *Scopus* database categorizes the publication records into four broad subject categories. These are physical science, health science, social science and life science. They are further classified into a more detailed categorization with 26 categories. Among these, physical science includes 10, health science has 5, social science has 6, and life science has 5 categories. In order to obtain an indicative picture of subject category-wise research strengths of countries in the South Asian region, we computed the statistics for research output of individual countries with respect to subject categories. Since the records in the dataset do not contain the category value for the documents, we had to learn the categories of these documents. In order to do so, we used the 'source' field in the record. The source field contains the journal name. For each record, we used this value and mapped it to journal categorization of SCImago[11]. However, we found that some

Table 3. Research output-based ranks of countries 1996–2012 (top 10 and South Asian countries)

Rank	Country	Documents	Citable documents	Citations	Self-citations	Citations per document	*H* index
1	The United States	7,063,329	6,672,307	129,540,193	62,480,425	20.45	1,380
2	China	2,680,395	2,655,272	11,253,119	6,127,507	6.17	385
3	United Kingdom	1,918,650	1,763,766	31,393,290	7,513,112	18.29	851
4	Germany	1,782,920	1,704,566	25,848,738	6,852,785	16.16	740
5	Japan	1,776,473	1,734,289	20,347,377	6,073,934	12.11	635
6	France	1,283,370	1,229,376	17,870,597	4,151,730	15.6	681
7	Canada	993,461	946,493	15,696,168	3,050,504	18.5	658
8	Italy	959,688	909,701	12,719,572	2,976,533	15.26	588
9	Spain	759,811	715,452	8,688,942	2,212,008	13.89	476
10	India	750,777	716,232	4,528,302	1,585,248	7.99	301
46	Pakistan	58,133	55,915	243,958	72,199	6.22	111
63	Bangladesh	19,481	19,037	115,329	22,662	8.37	97
78	Sri Lanka	8,239	7,853	61,175	6,285	9.91	86
90	Nepal	6,070	5,582	41,907	5,494	9.73	71
158	Afghanistan	485	441	2,088	241	5.38	21
176	Bhutan	295	290	1,360	173	6.55	18
191	Maldives	135	131	895	52	6.57	15

Data source: http://www.scimagojr.com/countryrank.php

Table 4. Subject category-wise research output in the South Asian countries (1964–2013)

Category	Afghanistan	Bangladesh	Bhutan	India	Maldives	Nepal	Pakistan	Total
Physical science	12	9,185	24	460,746	11	783	24,914	495,675
Health science	11	6,736	4	205,138	12	3,048	20,794	235,743
Social science	10	1,339	8	21,954	1	214	2,723	26,249
Life science	28	7,265	7	288,176	16	948	21,764	318,204
Total	61	24,525	43	976,014	40	4,993	70,195	

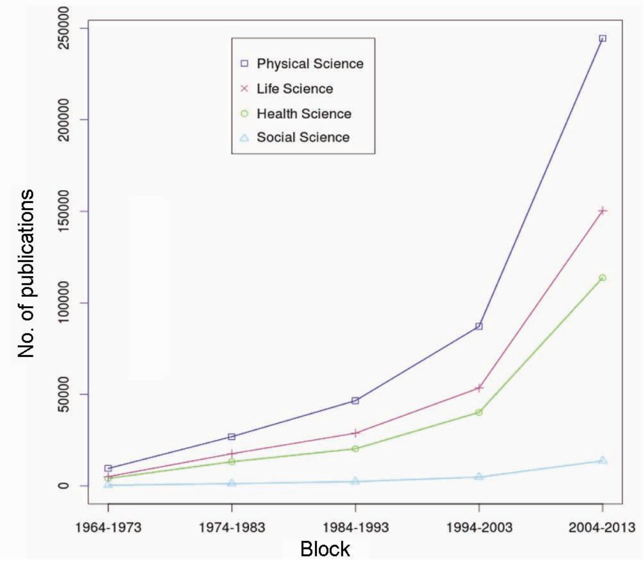

Figure 2. Subject category-wise research output of South Asia during 1964–2013.

of the document records do not have any value in the ‘source’ field of the data. These records could not be mapped to a category using the journal categorization. For these records, we extracted the ‘index keywords’ in the record and used a majority class machine learning classifier to learn appropriate categories. The training data of the keywords for each category were the keywords extracted from the records with known source value (and thus the category). Some of the records did not have even the ‘index keywords’ information and we excluded them from our final counts. In all, we have been able to identify broad subject category of a total of 910,927 documents out of the total 1,042,846 records available with us. Table 4 presents the actual figures for research output of all the South Asian countries in the four broad subject categories. The physical science category tops in the research output. On the other hand the research output in social science was the lowest. We also wanted to measure the category-wise growth of research output of the South Asian region and therefore mapped the category-wise research output data to the five chronological blocks. Figure 2 shows the plot of growth trend in research output for the South Asian region as a whole during the last 50 years. The figures correspond to five chronological blocks as used earlier. We also computed percentage increase for each chronological block and then averaged all these rates to obtain an average percentage growth. For example, if research output for a category is *A* in block 1964–73 and *B* in 1974–83, then the percentage growth during the period will be $((B - A)/A)*100$. A simple calculation obtains that for the four categories (physical science, life science, health science and social science, in order) the average percentage growth figures (over the previous block) for entire period are 90%, 97%, 77% and 97% respectively.

Measuring the impact of research

The previous analysis shows us that research output of the South Asian region has increased during the last 50 years. However, just measuring the absolute research output will be incomplete if we do not measure the impact of the research output. For this, we resorted to computing the citation counts of research papers published. The fact that a research paper is cited by other researchers indicates its impact and usefulness. Higher citation count for a research paper is a measure of its increased contribution and impact on the research in the corresponding area. The *Scopus* data we downloaded contain a 'citation count' field for every research paper, but we do not have the list of successive research papers citing a particular paper. We, therefore, mined the collected data and obtained the citation records for each paper during the entire study period (1964–2013) and also segregated them into the five chronological blocks. There are, however, some problems with the values available in the *Scopus* data. Some of the records for 'cited by' field are not updated. We have found some errors in the values for this field for some records. We reported it to the *Scopus* team, which accepted that some records may not be updated. In all erroneous cases, however, the value in the 'cited by' field is a bit less than the actual value. We have scanned the full *Scopus* record and updated the 'cited by' field wherever it was observed to be erroneous vis-à-vis our computation. In Figure 3, we plot the total citation counts for the research output from the South Asian region during the five chronological blocks. The results indicate that there is a substantial increase in citation counts of research papers produced in South Asia over the five blocks in the last 50 years.

In Figure 4, we plot two more computed values, the average citation per document (ACPD) and the average citation per document – time adjusted (ACPD$^{\mathrm{TA}}$). While ACPD is computed by dividing the total number of citations by the total number of research papers, ACPD$^{\mathrm{TA}}$ is computed with a time adjustment. It is obvious that older papers are likely to have more citations (as they have a wider time-span of being cited). New research papers, on the other hand, are less cited as they are new. In order to take the time factor into account, we have computed a time-adjusted value for ACPD. For this, the ACPD value for a block is divided by the total time-span (in terms of blocks) available for this to be cited. Thus the ACPD value of 3.18 for the block 1964–73 becomes 0.63 for the ACPD$^{\mathrm{TA}}$ measure after it is divided by 5. Similarly, for the block 1974–83, it is 0.73. Taking a closer look at the ACPD$^{\mathrm{TA}}$ values, we observe that in actual sense there is an increase in the impact of research output, since the time-adjusted ACPD shows an increasing trend against the decreasing nature of ACPD values.

Figure 5 presents a plot of research output and citation count on a year-wise basis, for the entire 50-year period. For this, the research output and citation count for all the 50 years are extracted from the data and plotted on a year-wise basis. There seems to be a correlated increasing relationship between research output and citation count. Though the citations are not increasing in proportion to the research output, there is a clear trend of growth in the time-adjusted ACPD values. Overall, we see an increase

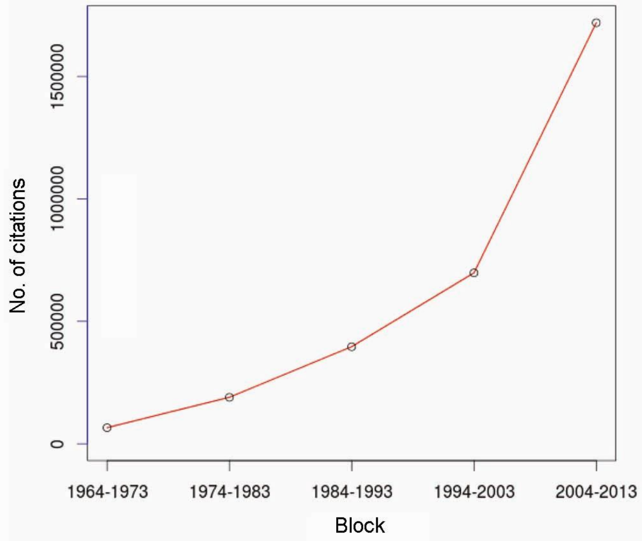

Figure 3. Citation counts of research output in South Asia during 1964–2013.

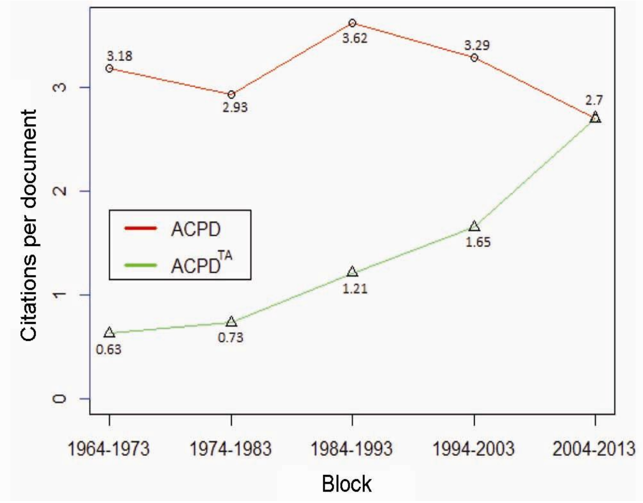

Figure 4. Average citation per document. ACPD, Average citation per document; TA, Time adjusted.

in the research output and impact of the region, though it is less when compared to the worldwide statistics.

Assessment of research maturity level

We wanted to measure the logical incremental continuum of research output in the South Asian region. For this, we have taken the citation data as the base and then performed a systematic mining to find out how many citations of each research paper correspond to a research paper produced within the South Asian region. For every research paper, we process its references. For each reference record of a given paper, we look for its corresponding records in *Scopus* to identify whether it is a paper from within the South Asian region (measured by author affiliation institution) or from outside the region. This has been a tedious and time-consuming process. For every record in *Scopus*, we had to process its reference list and then for each item in the reference set, a backward look-up in *Scopus* database needs to be done. The main idea was to observe how much of the new created knowledge in South Asia is based on the existing knowledge of the region. In other words, how much self-incremental knowledge building is happening in South Asia and to what extent do our researchers look to the knowledge created in rest of the world? We call this measure as research maturity level, though we understand that this nomenclature may be debatable. In terms of our use of the term, the measure 'research maturity' gives an idea about the utilization of existing in-house knowledge vis-à-vis knowledge from the outside; for a newly published research paper. In this sense, it also gives an indirect idea about the usefulness of the ongoing research activities in the region. Figure 6 presents an area under the curve (AUC) plot for the results. It shows how much proportion of citations originating from the South Asian region over the period of last 50 years (measured in 10-year blocks) is within the region or outside. In numerical terms, for the entire 50-year period, we find that approximately 17% of citations originating from publications of the South Asian region are to papers within the region. The AUC plot shows the proportion of citations within and out of the region during the five chronological blocks. It appears that there is an increasing trend of depending on and citing research from across the world. This may be understood from the fact that now the research community in the South Asian region has access to a wider and larger part of research production across the world and not only within the region. From a different viewpoint, it may also be inferred that the South Asian region has not yet been able to reach the quantitative and qualitative self-sufficiency level in research production and a lot more needs to be done in the higher education and research sector.

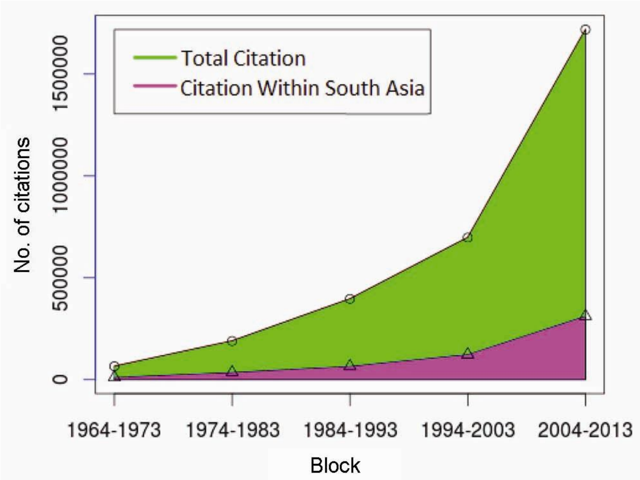

Figure 6. Citation proportions within and outside South Asia.

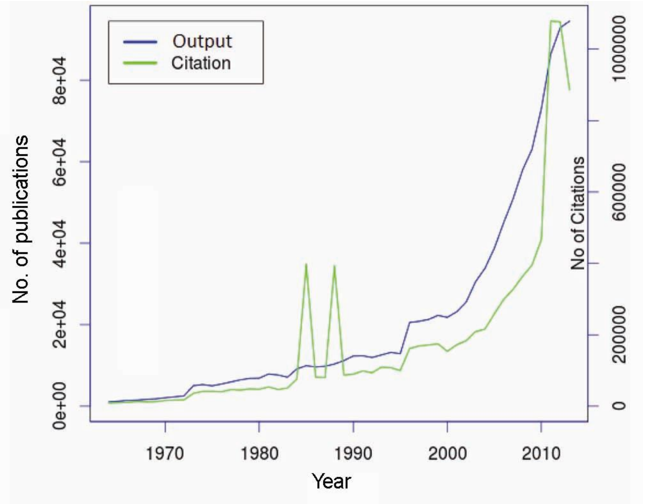

Figure 5. Research output and citation count (year-wise).

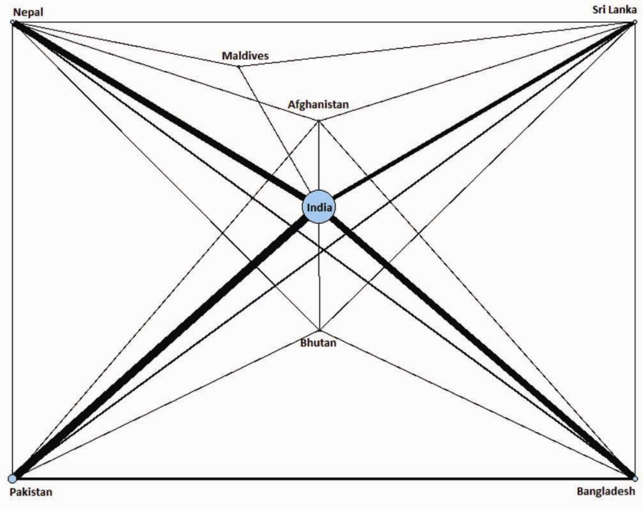

Figure 7. Research collaboration network of South Asian countries.

Table 5. Research collaboration statistics for South Asia (1964–2013)

Collaboration among South Asian authors	Collaboration between South Asian and external authors	Total collaborative research output	Total research output
3872 (2.24% of collaborative output)	169,079 (97.76% of collaborative output)	172,951 (16.58% of total)	1,042,846

Table 6. Top 20 countries involved in collaborative research with the South Asian region (1964–2013)

		Collaborative output	
Rank	Country	Absolute number	Percentage
1	The United States	48,145	20.21
2	United Kingdom	18,936	7.95
3	Germany	17,114	7.18
4	Japan	13,599	5.71
5	France	10,003	4.2
6	South Korea	8,998	3.78
7	Canada	8,864	3.72
8	Australia	7,264	3.05
9	China	7,232	3.04
10	Italy	6,383	2.68
11	Malaysia	5,186	2.18
12	Switzerland	4,468	1.88
13	The Netherlands	4,183	1.76
14	Spain	4,096	1.72
15	Sweden	3,874	1.63
16	Russian Federation	3,703	1.55
17	Saudi Arabia	3,573	1.5
18	Taiwan	3,548	1.49
19	Brazil	3,091	1.3
20	Poland	2,732	1.15

Research collaboration within and outside South Asia

Collaboration among the academic community is known to improve both the quality and quantity of research output. It promotes amalgamation of different sources of knowledge, different disciplines and different approaches. A recent study suggests that geopolitical location, cultural relations and language are major determining factors in collaboration and co-authorship[12]. Research groups in a geographical region develop as knowledge clusters in a particular area. Since our analysis is for the South Asian region, where many countries share geographical boundaries, we thought of measuring the research collaboration levels among them. For this, we extract all records from the dataset that have authors from at least two of the eight countries in the region. Figure 7 shows the collaboration network of the eight countries in the region. An edge in this graph represents a collaborative relationship between the two vertices (representing two countries). The thickness of an edge depends on the collaborative activity (so is the thickness of a node). The results show that India, in addition to having higher research production has also been in the forefront of research collaboration with other countries. One of the reasons for this is that India has contributed about 90% of the total research publication in the South Asian region during the last 50 years. We also observe that there has been at least some collaborative work in almost all country-pairs of the region, except few involving Afghanistan, Bhutan and Maldives.

However, when we measured the research collaboration of the South Asian region with the rest of the world, we found that the principles of geographical location for higher research collaboration do not hold. We computed that out of the total 1,042,846 research papers produced during 1964–2013 in the South Asian region, 172,951 papers (approximately 16.5%) involve research collaboration within at least two countries. When we look deeper into the collaborative research output, we find that only 3872 papers (approximately 2.2%) involve collaboration among the South Asian countries. A large number of collaborative papers (169,079 – approximately 97.7% of the total collaborative output) are the result of collaboration between a South Asian country and a country out of this region. Table 5 presents the actual figures for this result. We also found the major collaborating countries for the South Asian region. Table 6 presents the list of top 20 countries with which the South Asian region has research collaboration. The United States and United Kingdom top the list with approximately 20% and 8% share respectively, of the total collaborative output with South Asia.

Conclusion

We have done a scientometric analysis of research output, growth trends, citation, impact and collaboration behaviour for the South Asian region. The results present details about the total research output and global share of the South Asian region. We have also presented country-wise and subject category-wise details for the research output and the research impact & growth trends of the South Asian region for the 50-year period. The article also measures the research maturity level of the South Asian region by analysing the citation patterns within and outside the region. The research collaboration pattern among the South Asian countries and also that of South Asia with rest of the world is computed and analysed. The article presents an interesting and useful insight into the research output, growth, citations, impact and collaboration patterns of the South Asian region. The results can be used for academic analysis, policy decision or for formulating scientific and collaborative programmes to enhance the research capabilities and output of the South Asian region. The South Asian nations share a common history, geography and also have common problems. A positive growth in research effort and output and increased research cooperation among the countries in the region, habituating a substantial 21% of the world population, is the need of the hour.

1. http://en.wikipedia.org/wiki/South_Asia (accessed on 17 April 2014).
2. http://en.wikipedia.org/wiki/South_Asian_Association_for_Regional_Cooperation (accessed on 17 April 2014).
3. http://saarc-sec.org/SAARC-Charter/5/ (accessed on 17 April 2014).
4. Report on higher education in South Asia, *The Economist*, June 2013.
5. Zhou, P., Zhong, Y. and Yu, M., A bibliometric investigation on China–UK collaboration in food and agriculture. *Scientometrics*, 2013, **97**, 267–285.
6. Bilir, S., Gogus, S., Onal, O., Ozturkmen, N. D. and Yontan, T., Research performance of Turkish astronomers in the period of 1980–2010. *Scientometrics*, 2013, **97**, 477–489.
7. Kutlar, A., Kabasakal, A. and Ekici, M. S., Contributions of Turkish academicians supervising Ph D dissertations and their universities to economics: an evaluation of the 1990–2011 period. *Scientometrics*, 2013, **97**, 639–658.
8. Prathap, G., A bibliometric profile of *Current Science*. *Curr. Sci.*, 2014, **106**(7), 958–963.
9. Pathak, M. and Bharati, K. A., Botanical Survey of India (1971–2010): a scientometric analysis. *Curr. Sci.*, 2014, **106**(7), 964–971.
10. http://data.worldbank.org/indicator/SP.POP.TOTL (accessed on 18 April 2014).
11. SCImago, SJR – SCImago Journal and Country Rank, 2007; http://www.scimagojr.com (accessed 18 April 2014).
12. Schubert, A. and Glänzel, W., Cross-national preference in co-authorship, references and citations. *Scientometrics*, 2006, **69**(2), 409–428.

Received 27 April 2014; revised accepted 7 May 2014