CORRELATING R & D EXPENDITURE AND SCHOLARLY PUBLICATION OUTPUT USING K-MEANS CLUSTERING

R. S. Kamath [1], R. K. Kamat [2] and S.M. Pujar [3]

[1] Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, India

[2] Department of Electronics, Shivaji University, Kolhapur, India

[3] Indira Gandhi Institute of Development Research, Gen AK Vaidya Marg, Goregaon East, Mumbai, India

ABSTRACT

History of humanity especially post renaissance era depicts the contribution of research and its output In terms of publication, patents and technology transfer paving the way for the societal prosperity. Scientific writing and research publication are fundamental components indicative of academic excellence and supposedly committed to the Research and Development (R&D) funding. Thus the budget allocation and expenditure thereof towards R&D is considered to be a vital parameter for the advancement in science and technology and also for social and economic well being. In view of the tall aspirations of society and government at large it becomes indispensable to investigate whether the scholastic output is going hand in hand with the R & D budget spillover or otherwise. We present in this communication a systematic clustering approach based on K-means algorithm to reveal the impact of R&D expenditure on the extent of research publications. Two independent sources of data, Research and development expenditure i.e. percentage of Gross Domestic Product (GDP) and Scientific and technical journal articles, are brought together in this comprehensive study. From an empirical perspective, present study found that there exist a positive linear correlation between R & D Expenditure and number of research publication.

KEYWORDS

R&D Expenditure, Publications, Correlation, Clustering, Dataset, Analysis

1. INTRODUCTION

The following meaningful quote by astronomer Carl Sagan "Somewhere, something incredible is waiting to be known," essentially portrays the psyche of a passionate researcher, how however fails to perform in the crunch of diminishing R & D budgetary allocation. It is without any doubt that the investment towards R&D influences innovations and in turn stimulates the growth of a country [11]. However the relationship between the percentage of GDP spent on the R&D vis-à-vis the quantity as well as quantity of scholarly journal articles published has been a topic of inquiry through the globe. There are number of studies reported in the past investigating the correlation of productivity in the academic research community in the light of the research expenditure [1, 2, 10].

Igor Prodan has reported the model which portrays influence of R & D expenditures on number of patent applications in selected OECD countries and central Europe [4]. Research confirms the positive correlation between the two. Meo and Usmani have presented impact of R&D spending on research publications, patents and high technology exports among 47 European countries [3]. This research collected the information regarding per capita GDP, R&D expenditure with the conclusion that R&D expenditure and research publications are the most significant contributing factors towards a knowledge economy.

Janodia has compared expenditure on R & D and patents of India among SAARC and BRICS countries [5]. The study revealed that it is essential to raise R&D expenditure to motivate research activities leading to innovation, increasing patenting and larger number of publications. Dietmar reported the effect of R&D spending on its productivity in German manufacturing firms [7]. The result of the study suggested that spillovers affect industries in a heterogeneous manner. Yet

DOI: 10.5121/ijitmc.2017.5101

another paper by Kamath and Kamat reported clustering of countries based on their R&D expenditure and its productivity [6]. Here analysis has done based on the number of research publications, patent applications and trademarks registered with reference to percentage of GDP spending on R&D.

In the present communication we report a systematic methodology with R&D datasets as input to the k-means algorithm to cluster the performers as well as the gloomy ones in the interest of the public policy at large. The impact of research recourses quantified in terms of R & D expenditure and the corresponding research output, as measured in terms of the quantity of journal articles have been taken as two independent data sets for the correlation purpose. In order to visualize the competitive picture we correlate R&D expenditure and journal publications on a global scale taking into consideration the individual country performance. We use R data mining tool for more significant analysis resulting into the clusters having varied performance.

2. DATASETS FOR PRESENT ANALYSIS

In this comprehensive study, two independent sources of data, Research and development expenditure i.e. % of GDP and Scientific and technical journal articles, are brought together for the assessment purpose. These datasets are available at Knoema (http://knoema.com) a free to use web based public and open data platform launched for the purpose of statistical analysis. This web based platform uses international open data platforms such as World Bank, IMF, UN and many more as sources to fetch the data. Expenditures for R&D available on this platform are current, up to date and include capital expenditures for both public and private sector on creative work undertaken systematically to increase knowledge and the use of knowledge for new applications. Knoema implies R&D that includes basic research, applied research, and experimental development. Scientific and technical journal articles refer to the number of scientific and engineering articles published in the following fields: physics, biology, chemistry, mathematics, clinical medicine, biomedical research, engineering and technology, and earth and space sciences. Snapshot of the dataset used in this study shown in figure 1.With such a voluminous availability of the data sets it would be really worthwhile to deice the trends and meaningful performance measures out of the same. Following section presents the methodology for devising the trends from the data sets.

R&D expenditure (% of GDP)							Scientific and technical journal articles						
	2011	2010	2009	2008	2005	2000		2011	2010	2009	2008	2005	2000
Argentina	0.6	0.6	0.6	0.5	0.5	0.4	Argentina	3,863	3,768	3,655	3,567	3,058	2,846
Armenia	0.3	0.2	0.3	0.2	0.3	0.2	Armenia	185	183	164	191	180	175
Austria	2.8	2.8	2.7	2.7	2.5	1.9	Austria	5,103	4,923	4,833	4,816	4,568	4,257
Azerbaijan	0.2	0.2	0.2	0.2	0.2	0.3	Azerbaijan	149	145	151	100	116	79
Belarus	0.7	0.7	0.6	0.7	0.7	0.7	Belarus	342	335	380	387	491	572
Belgium	2.2	2.1	2	2	1.8	2	Belgium	7,484	7,389	7,222	7,298	6,847	5,735
Brazil	1.2	1.2	1.2	1.1	1	1	Brazil	13,148	12,530	12,307	12,909	9,897	6,407
Bulgaria	0.6	0.6	0.5	0.5	0.5	0.5	Bulgaria	650	675	735	761	767	909
Burundi	0.1	0.1	0.1	0.2	0.1	0.1	Burundi	3	6	3	2	3	1
Canada	1.8	1.9	2	1.9	2	1.9	Canada	30112	29817	29017	28,637	25,862	22,701
China	1.8	1.8	1.7	1.5	1.3	0.9	China	89,894	79,991	74,034	65,301	41,604	18,479
Colombia	0.2	0.2	0.2	0.2	0.1	0.1	Colombia	727	692	608	575	401	332
Costa Rica	0.5	0.5	0.5	0.4	0.4	0.4	Costa Rica	106	102	98	107	106	80
Croatia	0.8	0.8	0.9	0.9	0.9	1.1	Croatia	1,289	1,247	1,164	1,188	953	678
Cuba	0.3	0.6	0.6	0.5	0.5	0.5	Cuba	224	195	222	253	261	284
Cyprus	0.5	0.5	0.5	0.4	0.4	0.2	Cyprus	211	212	195	169	91	59
Czech Republic	1.6	1.4	1.4	1.3	1.2	1.2	Czech Republic	4,127	4,164	3,949	3,936	3,172	2,483
Denmark	3	3	3.2	2.8	2.5	2.3	Denmark	6,071	5,639	5,307	5,304	5,048	4,883
Egypt, Arab Rep.	0.4	0.4	0.2	0.3	0.2	0.2	Egypt, Arab Rep.	2,515	2,431	2,247	2,019	1,658	1,433
El Salvador	0	0.1	0.1	0.1	0.1	0.1	El Salvador	9	5	6	3	6	3
Estonia	2.4	1.6	1.4	1.3	0.9	0.6	Estonia	514	527	518	477	439	328
Finland	3.8	3.9	3.9	3.7	3.5	3.3	Finland	4,878	4,869	4,952	5,113	4,813	4,844
France	2.2	2.2	2.3	2.1	2.1	2.2	France	31,686	31,368	31,757	31,983	30,340	31,427
Germany	2.9	2.8	2.8	2.7	2.5	2.5	Germany	46,259	45,338	45,017	44,915	44,194	43,510
Hungary	1.2	1.2	1.2	1	0.9	0.8	Hungary	2,289	2,221	2,399	2,554	2,619	2,358
Iceland	2.6	2.6	2.8	2.6	2.8	2.7	Iceland	258	274	260	233	206	150
India	0.8	0.8	0.8	0.8	0.8	0.7	India	22,481	20,882	19,924	18,988	14,635	10,276
Ireland	1.7	1.7	1.7	1.4	1.2	1.1	Ireland	3,186	3,056	2,800	2,662	2,120	1,581
Israel	4	4	4.2	4.4	4.3	4.2	Israel	6,096	6,139	6,306	6,609	6,322	6,290
Italy	1.3	1.3	1.3	1.2	1.1	1	Italy	26,503	26,348	26,770	26,854	24,663	21,409
Japan	3.4	3.3	3.4	3.5	3.3	3	Japan	47,106	47,043	49,632	51,842	55,527	57,101
Kazakhstan	0.2	0.2	0.2	0.2	0.3	0.2	Kazakhstan	87	83	99	77	96	113
Korea, Rep.	4	3.7	3.6	3.4	2.8	2.3	Korea, Rep.	25,593	24,106	22,280	21,091	16,396	9,572
Kuwait	0.1	0.1	0.1	0.1	0.1	0.1	Kuwait	202	192	214	240	234	238
Kyrgyz Republic	0.2	0.2	0.2	0.2	0.2	0.2	Kyrgyz Republic	17	17	15	17	15	18
Latvia	0.7	0.6	0.5	0.6	0.6	0.4	Latvia	204	134	162	161	134	150
Lithuania	0.9	0.8	0.8	0.8	0.8	0.6	Lithuania	457	356	388	515	406	262
Luxembourg	1.4	1.5	1.7	1.7	1.6	1.7	Luxembourg	204	149	137	111	59	40
Macedonia, FYR		0.2	0.2	0.2	0.2	0.4	Macedonia, FYR	77	59	57	46	43	56
Madagascar	0.1	0.1	0.1	0.1	0.2	0.1	Madagascar	33	41	35	38	34	29
Malaysia	1.1	1.1	1	0.8	0.6	0.5	Malaysia	2,092	1,608	1,351	951	615	460

Figure 1. Snapshot of dataset

3. ANALYSIS THROUGH CORRELATION AND CLUSTERING

In order to visualize meaningful performance index relating the R & D budget with the publication output we adopt the K-means clustering algorithm [8]. The data were analyzed by using MS-Excel and R data mining tool. In this study 76 countries' R&D expenditure and research publication for the years 2011, 2010, 2009, 2008, 2005 and 2000 were included. The correlation coefficient was calculated to find the imminence of relation between these two parameters. Correlation between R&D expenditure and quantity of research publications data demonstrated in table 1. All these 76 countries' correlation coefficients are shown as Radar chart in figure 2.

Table 1. Correlation of R&D expenditure and Research publications

Countries	Correlation Coef.	Countries	Corelation Coef.
Argentina	0.91	Latvia	0.43
Armenia	-0.39	Lithuania	0.77
Austria	0.94	Luxembourg	-0.70
Azerbaijan	-0.73	Macedonia, FYR	0.30
Belarus	0.20	Madagascar	-0.12
Belgium	0.38	Malaysia	0.95
Brazil	0.81	Malta	0.47
Bulgaria	-0.74	Mexico	0.83
Burundi	-0.29	Moldova	-0.58
Canada	-0.27	Mongolia	0.75
China	0.99	Nepal	0.31
Colombia	0.93	Netherlands	0.10
Costa Rica	0.23	New Zealand	0.67

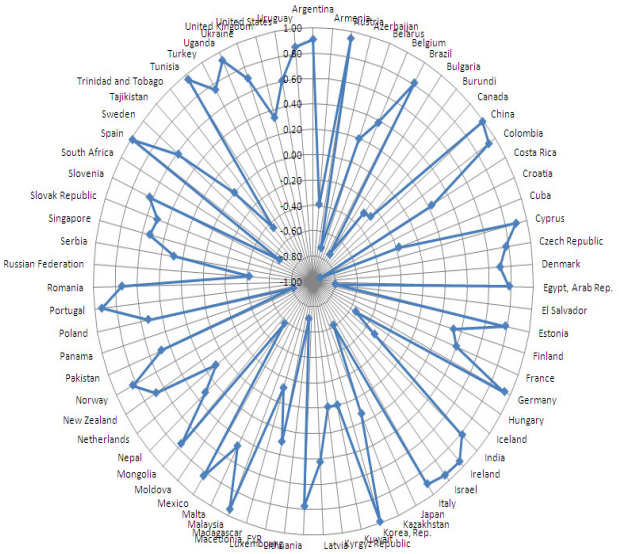

Figure 2. Radar chart representing correlation coefficients

This correlated data was further used for designing prominent clusters of countries by applying unsupervised learning. We undertake the same to uncover the relationship between R&D expenditure and research publications. R open source data mining tool has been used for clustering of countries based on their correlation coefficients [9]. The K-means algorithm is used here for grouping countries according to their similarity. Corresponding output of clustering using 'R' is shown in figure 3. Three prominent clusters have been devised implying positive, negative

or neutral correlation. K-means clustering resulted into three clusters of size 17, 33 and 21 given in table 2.

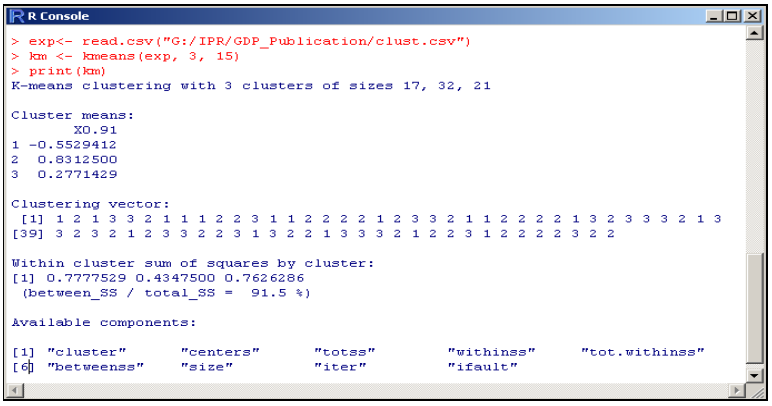

Figure 3. Snapshot of Result of clustering in R

Table 2: Clusters of countries based on correlation data

Cluster 1 - Negative Correlation	Cluster 2 - Positive Correlation		Cluster 3 - Weak Correlation
Armenia	Argentina	Malaysia	Belarus
Azerbaijan	Austria	Mexico	Belgium
Bulgaria	Brazil	Mongolia	Costa Rica
Burundi	China	New Zealand	Finland
Canada	Colombia	Norway	France
Croatia	Cyprus	Portugal	Kazakhstan
Cuba	Czech Republic	Romania	Kuwait
El Salvador	Denmark	Slovenia	Kyrgyz Republic
Hungary	Egypt, Arab Rep.	Spain	Latvia
Iceland	Estonia	Sweden	Macedonia, FYR
Japan	Germany	Tunisia	Madagascar
Luxembourg	India	Turkey	Malta
Moldova	Ireland	Uganda	Nepal
Panama	Israel	Ukraine	Netherlands
Russian Federation	Italy	United States	Pakistan
South Africa	Korea, Rep.	Uruguay	Poland
Trinidad and Tobago	Lithuania		Serbia
			Singapore
			Slovak Republic
			Tajikistan
			United Kingdom

4. INTRA CLUSTER ANALYSIS AND PERFORMANCE MEASURES

The countries belong to cluster 1 have non - linear relationship between R&D expenditure and Publication as their correlation coefficient is closer to -1. Negative correlation implies a completely inverse relationship between the R&D expenditure with number of publications. For instance Bulgaria and Japan belonging to this cluster with correlation coefficient -0.74 and -0.61 indicate gloomy performance in terms of publication output despite being the heavy investment in R & D shown in table 3.

Positive correlation i.e. correlation coefficient closer to +1 found in countries belonging to cluster2 reveals a linear relationship between R&D expenditure and Publication. Intra cluster analysis shows that as the R&D expenditure increases number of publications also increases in case of these countries. India, Argentina and Malaysia are the main countries belonging to this cluster with correlation coefficient 0.82, 0.91 and 0.95. Sample data of these countries belonging

to cluster 2 can be further analyzed so as to perceive the closeness of the publications versus the R & D spending given in table 4.

The weak correlation between R&D expenditure and number of publications is exhibited by countries belonging to cluster 3. The correlation coefficient is closer to 0 implies almost no relationship between R&D expenditure and number of publications. For example Kuwait and Serbia belong to this cluster with correlation coefficient 0.0 and 0.28 shown in table 5.

Table 3. Sample data of countries belonging to cluster

		2011	2010	2009	2008	2005	2000
Bulgaria	R&D Expenditure	0.6	0.6	0.5	0.5	0.5	0.5
	Publications	650	675	735	761	767	909
		2011	2010	2009	2008	2005	2000
Japan	R&D Expenditure	3.4	3.3	3.4	3.5	3.3	3
	Publications	47,106	47,043	49,632	51,842	55,527	57,101

Table 4. Sample data of countries belonging to cluster 2

		2011	2010	2009	2008	2005	2000
India	R&D Expenditure	0.8	0.8	0.8	0.8	0.8	0.7
	Publications	22,481	20,882	19,924	18,988	14,635	10,276
Argentina	R&D Expenditure	0.6	0.6	0.6	0.5	0.5	0.4
	Publications	3,863	3,768	3,655	3,567	3,058	2,846
		2011	2010	2009	2008	2005	2000
Malaysia	R&D Expenditure	1.1	1.1	1	0.8	0.6	0.5
	Publications	2,092	1,608	1,351	951	615	460

Table 5. Sample data of countries belonging to cluster 3

		2011	2010	2009	2008	2005	2000
Kuwait	R&D Expenditure	0.1	0.1	0.1	0.1	0.1	0.1
	Publications	202	192	214	240	234	238
		2011	2010	2009	2008	2005	2000
Serbia	R&D Expenditure	0.8	0.8	0.9	0.4	0.4	1
	Publications	1,269	1,165	1,173	1,043	976	943

5. CONCLUSIONS

Our investigation using the K-means clustering algorithm using R-mining is in fact one of its kinds systematic modus operandi to device the correlation between two independent data sets for perceiving the performance metrics for the benefit of the policy makers, scientific community and the society at large. Since the budget for the R & D is drawing from the tax payers, society ultimately becomes the major stakeholder of the research. Visualization and analysis technique such as the reported ones not only correlates independent data sets but also throws light on the preeminent trends and patterns that the stakeholders ought to know. Our investigation reveals that for 33 countries out of 76, there exists positive linear correlation between R & D Expenditure and number of research publications. The impact of research funding structure on quantity of journal articles is fundamental to an understanding of the functioning of research systems. This question is not of just aesthetic interest to scientific community. In times of limited funding there are pressures to implement policies that will affect the distribution of funds and the reported

framework will definitely ensure judicious distribution from the apex funding bodies like World Bank.

REFERENCES

[1] Jacob, Brian A and Lefgren (2011), "Lars, The impact of research grant funding on scientific productivity", Journal of Public Economics, 95, pp1168-1177.

[2] McAllister, Paul R and Wagner, Deborah Ann (1981), "Relationship between R&D expenditures and publication output for U.S. colleges and universities", Research in Higher Education, 15, 3-30.

[3] Meo SA and Usmani Adnan Mahmood (2014), "Impact of R&D expenditures on research publications, patents and high-tech exports among European countries", European Review for Medical and Pharmacological Sciences, 18, pp1-9.

[4] Prodan Igor, "Influence of Research and Development Expenditures on Number of Patent Applications: Selected Case Studies in OECD Countries and Central Europe", Applied Econometrics and International Development. AEID., 2005, Vol. 5-4.

[5] Manthan D. Janodia, "Research and development spending and patents: where does India stand among SAARC and BRICS", Current Science, 2015, 108.

[6] R.S. Kamath & R.K. Kamat (2016), "K-Means Clustering for Analyzing Productivity in Light of R & D Spillover", International Journal of Information Technology, Modeling and Computing (IJITMC), Vol.4, No. 2, pp55-64.

[7] Dietmar Harhoff, "R&D Spillovers", Technological Proximity and Productivity Growth – Evidence from German Panel Data

[8] Jiawei Han, Micheline Kamber and Jian Pei (2012), Data Minining Concepts and Techniques, Third Edition, Elsevier Inc.

[9] R.S.Kamath, R.K.Kamat (2016), Educational Data Mining with R and Rattle, River Publishers Series in Information Science and Technology, River Publishers.

[10] Bozeman, B., & Melkers, J. (1993), "Evaluating R & D impacts". Boston: Kluwer Academic.

[11] Gaillard, J. (2010) "Measuring Research and Development in Developing Countries: Main Characteristics and Implications for the Frascati Manual". Science Technology & Society, 15(1), pp77-111. doi:10.1177/097172180901500104

Authors

R.S. Kamath

Dr. R.S. Kamath is Associate Professor in the Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, India. She obtained her Bachelors and Masters in Computer Science from Mangalore University. She received her Ph.D. in Computer Science specialized in Computer Based Visualization from Shivaji University and completed the same in 2011. Dr. Kamath has to her credit 20 research papers published in reputed national and international journals and presented 11 papers in national conferences. She has completed two minor research funded by UGC. She is the author of two books and edited one seminar proceedings. The chapter entitled "Cost Effective 3D Stereo Visualization for Creative Learning – Virtual Reality in Education" is accepted for Encyclopedia of Information Science and Technology 4th Edition by IGI Global Publishing. Her areas of research interests are Artificial Intelligence, Virtual Reality, Soft Computing and Data Mining. She has immense skill of around twelve years in teaching and research.

R.K. Kamat

Dr. R.K. Kamat holds the position of Professor in the Department of Electronics and heads the Department of Computer Science of Shivaji University, Kolhapur. He is Director of Internal Quality Assurance (IQAC) of the Shivaji University, Kolhapur. He obtained his Bachalors, Masters and M.Phil in Electronics from the Shivaji University, Kolhapur. Dr. Kamat gained his Ph.D. specialized in Smart Sensors from Goa University, Goa. Professor Kamat has published over 70 plus papers in International journals of repute and presented equal number of papers at National and International Conferences. He has published 10 books though reputed publishing house such as Springer UK. He has published scholarly literature on the quality issues in higher education. Five students have been awarded Ph.D. under his guidance and 11 more are working for their Doctorate. He has been involved with various initiatives of the apex organization at international level such as IEEE USA and Engineering Education group of Central Quinsland Australia. Through these organizations he is playing key role in spreading the scholastic culture by organizing conferences at various international destinations. Through this network he has visited various countries. Dr. Kamat is a recipient of the Young Scientist Award under the fast track scheme of the Department of Science and Technology (DST) of Government of India.

S.M. Pujar

Dr. S. M. Pujar is Deputy Librarian in Indira Gandhi Institute of Development Research, Gen AK Vaidya Marg, Goregaon East, Mumbai. Dr. Pujar has published over 20 plus papers in reputed national and international journals and presented papers in conferences. His areas of research interests are Digital Libraries, Information Science, ICT applications for Libraries and Library Automation.