# K-MEANS CLUSTERING FOR ANALYZING PRODUCTIVITY IN LIGHT OF R & D SPILLOVER

R. S. Kamath[1] and R. K. Kamat[2]

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

***ABSTRACT***

*The differences between countries go far beyond the physical and territorial aspects. Hence, for analytical purposes, it is essential to classify countries in groups based on some of their attributes. Investment in Research and Development (R&D) influences innovations which in turn stimulates growth of a country. In this context the productivity of the R&D expenditure is analysed pragmatically. Present study aims to discover impact of R&D expenditure on its productivity in terms of number of journal articles published, patent applications filed and trademark applications registered. A more significant analysis by means of designing prominent clusters of countries by applying unsupervised learning has been presented. In this division, percentage of Gross Domestic Product (GDP) spending on R&D and its productivity are considered.*

***KEYWORDS***

*R&D productivity; data mining; clustering; unsupervised learning*

## 1. INTRODUCTION

R&D productivity is a key indicator of a development of a country. There exists a direct relationship between research and overall development of a nation. Spending on research and development is vital for the advancement in science and technology in addition to social and economic development [9, 12]. Scientific writing in terms of research publication is essential components of academic excellence. Present communication aimed to compare the impact of R&D expenditure on journal articles published, patent applications filed and trademark applications registered.

Igor Prodan has presented a model which depicts dependency of the number of patent applications on R&D expenditure [1]. Research confirms the positive correlation exists between patent applications and R&D expenditure, R&D investment generates patent applications with a delay, which varies from country to country and the quantity of patent applications in developed countries depends more on R&D expenditure in the business sector than on R&D gross domestic expenditure. Meo and Usmani have reported impact of R&D spending on research publications, patents and high technology exports among 47 European countries [2]. This study collected the information regarding per capita GDP, R&D expenditure, number of universities, scientific journals, technology exports and number of patents. The main source for information for this study was Web of Science, World Bank, Thomson Reuters and SCImago/Scopus. This research concluded that, expenditure on R&D, scientific indexed journals and research publications are the most significant contributing factors towards a knowledge economy which in turn give a boost to patent applications, high technology exports and ultimately GDP.

DOI: 10.5121/ijitmc.2016.4204 

Janodia has compared Research and development spending and patents of India among SAARC and BRICS countries [3]. This study reveals that it is essential to increase in R&D expenditure by the Government of India which encourages research leading to innovation, increasing patenting and larger number of publications. The performance in terms of R&D expenditure and patents is strong among SAARC countries, whereas it is miserable among the countries of BRICS. Yet another paper by Dietmar reported the effect of R&D spillovers on R&D spending and its productivity in German manufacturing firms [4]. The result of panel estimation technique suggested that spillovers affect industries in a heterogeneous manner. High-technology industries spillovers have a increasing productivity effect in adding to encouraging R&D investment.

In the backdrop of the literature portrayed above, the present paper reports clustering of countries based on their R&D expenditure and its productivity. Data Mining promotes distinct tools and algorithms for analyze the data patterns [6]. We have explored efficiency of using machine learning algorithms for designing prominent clusters of countries based on R&D expenditure and its productivity. This paper explains a data mining process for investigating the relationship between the same using WEKA a popular open source free suite [5]. In this process, many criteria, such as R&D expenditure, journal articles, patents and trademark statistics are considered [10, 11]. These datasets have been taken from Knoema a free to use web based public and open data platform. As a broad goal, authors intended in extraction of the hidden knowledge from these datasets and designed clusters of countries based on R&D and its productivity.

## 2. DATASET DESCRIPTION AND DATA EXPLORATION

The datasets used in the present communication have been taken from Knoema (http://knoema.com) a free to use web based public and open data platform launched for the purpose of statistical and infographics analysis. The input dataset containing the numeric values of attributes such as R&D expenditure in terms of percentage of GDP, number of Scientific and technical journal articles, number of Patent applications, number of Trademark applications for 115 countries for the year 2011.

Table 1. Details of R&D attributes

| **Attribute** | **Description** |
|---|---|
| R&D expenditure, % of GDP | Current and capital expenditures both public and private on creative work undertaken systematically to increase knowledge and the use of knowledge for new applications. R&D covers basic research, applied research, and experimental development. |
| Scientific and technical journal articles | 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. |
| Patent applications, residents | The number of worldwide patent applications filed through the Patent Cooperation Treaty procedure or with a national patent office. |
| Trademark applications, direct resident | The number of trademark applications to register a trademark with a national or regional Intellectual Property office. |

Thus dataset for the study consists of R&D details of 115 countries representing expenditure and productivity. To preserve the semantics of the clusters, all the values used in this example are real statistics of the countries. Snapshot of dataset shown in figure 1. Table 1 explains details of these attributes and corresponding statistical analysis is given in Table 2. Dataset chosen for the study is analysed through radar diagrams. Figure 2(a-d) gives radar charts of 115 countries for R&D expenditure in terms of percentage of GDP, number of journal articles, number of Patent applications, number of Trademark applications respectively.

| | publication | Patent | Trademark | GDP | | publication | Patent | Trademark | GDP | | publication | Patent | Trademark | GDP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Albania | 27 | 3 | 274 | 0.2 | Greece | 4,534 | 721 | 4,065 | 0.7 | Norway | 4,777 | 1,122 | 3,411 | 1.7 |
| Algeria | 599 | 94 | 2,294 | 0.1 | Guatemala | 22 | 4 | 3854 | 0.0 | Pakistan | 1,268 | 56 | 14,003 | 0.3 |
| Argentina | 3,863 | 735 | 61121 | 0.6 | Hungary | 2,289 | 662 | 3772 | 1.2 | Panama | 67 | 92 | 4,167 | 0.2 |
| Armenia | 185 | 121 | 1,102 | 0.3 | Iceland | 258 | 50 | 691 | 2.6 | Paraguay | 21 | 21 | 569 | 0.1 |
| Australia | 20,603 | 2,383 | 40,150 | 2.4 | India | 22,481 | 8841 | 176386 | 0.8 | Paraguay | 9 | 75 | 13251 | 0.1 |
| Austria | 5,103 | 2,154 | 5,693 | 2.8 | Indonesia | 533 | 533 | 50,653 | 0.1 | Peru | 162 | 46 | 15231 | 0.1 |
| Azerbaijan | 149 | 193 | 1,915 | 0.2 | Iran, Islamic | 8,176 | 11529 | 26,825 | 0.7 | Philippines | 241 | 39 | 10,572 | 0.1 |
| Bahrain | 40 | 1 | 269 | 0.5 | Ireland | 3,186 | 567 | 1,485 | 1.7 | Poland | 7,564 | 186 | 14,252 | 0.8 |
| Bangladesh | 291 | 37 | 8,632 | 0.2 | Israel | 6,096 | 1,360 | 2,509 | 4.0 | Portugal | 4,621 | 3,879 | 15,616 | 1.5 |
| Barbados | 16 | 1 | 142 | 0.3 | Italy | 26,503 | 8,794 | 37001 | 1.3 | Qatar | 111 | 571 | 656 | 0.5 |
| Belarus | 342 | 1,725 | 3,649 | 0.7 | Jamaica | 51 | 20 | 1,119 | 0.7 | Romania | 1,626 | 49 | 8,389 | 0.5 |
| Belgium | 7,484 | 636 | 21,129 | 2.2 | Japan | 47,106 | 287,580 | 84,671 | 3.4 | Russian Fed | 14,151 | 1,424 | 33,252 | 1.1 |
| Bhutan | 8 | 3 | 8 | 0.6 | Jordan | 342 | 40 | 2,298 | 0.4 | Senegal | 79 | 26,495 | 873 | 0.5 |
| Bolivia | 47 | 52 | 2321 | 0.2 | Kazakhstan | 87 | 1,415 | 1,890 | 0.2 | Serbia | 1,269 | 40 | 1,133 | 0.8 |
| Bosnia and H | 54 | 43 | 243 | 0 | Kenya | 290 | 135 | 2,488 | 1 | Seychelles | 6 | 118 | 67 | 0.3 |
| Botswana | 50 | 1 | 329 | 0.1 | Korea, Dem. | 4 | 7,956 | 769 | 4 | Singapore | 4,543 | 180 | 4,236 | 2.2 |
| Brazil | 13,148 | 4,695 | 122,671 | 1.2 | Korea, Rep. | 25,593 | 138,034 | 112,576 | 4.0 | Slovak Repu | 1,099 | 1,056 | 2,332 | 0.7 |
| Bulgaria | 650 | 262 | 4058 | 0.6 | Kyrgyz Repu | 17 | 124 | 180 | 0.2 | Slovenia | 1,239 | 224 | 2 | 2.5 |
| Burkina Fasc | 53 | 2 | 34 | 0.2 | Latvia | 204 | 173 | 1,267 | 0.7 | South Africa | 3,125 | 470 | 19,522 | 0.8 |
| Cambodia | 33 | 1 | 903 | 0.6 | Lebanon | 251 | 102 | 1,884 | 0.4 | Spain | 22,910 | 656 | 42,748 | 1.4 |
| Canada | 30115 | 4,754 | 21,337 | 1.8 | Lithuania | 457 | 93 | 2,756 | 0.9 | Sri Lanka | 130 | 3,430 | 200 | 0.2 |
| Chile | 1,979 | 339 | 25,254 | 0.4 | Luxembourg | 85 | 85 | 5,814 | 1.4 | Sweden | 9,473 | 194 | 9,290 | 3.4 |
| China | 89,894 | 415,829 | 1,273,827 | 1.8 | Macao SAR, | 4 | 4 | 476 | 0.0 | Switzerland | 10,019 | 2,004 | 432 | 2.2 |
| Colombia | 727 | 183 | 16,976 | 0.2 | Macedonia, | 77 | 37 | 917 | 0.2 | Tajikistan | 18 | 1,597 | 161 | 0.1 |
| Costa Rica | 106 | 14 | 6,759 | 0.5 | Madagascar | 33 | 3 | 621 | 0.1 | Tanzania | 121 | 4 | 98 | 0.5 |
| Croatia | 1,289 | 230 | 1461 | 0.8 | Malaysia | 2,092 | 1,076 | 13001 | 1.1 | Thailand | 2,304 | 121 | 23,457 | 0.3 |
| Cuba | 224 | 62 | 256 | 0.3 | Malta | 46 | 9 | 423 | 0.7 | The Gambia | 3 | 927 | 105 | 0.1 |
| Cyprus | 211 | 4 | 646 | 0.5 | Mexico | 4128 | 1,065 | 71,091 | 0.4 | Tunisia | 1,016 | 47 | 1,675 | 1.1 |
| Czech Repub | 4127 | 783 | 8091 | 1.6 | Moldova | 76 | 97 | 1300 | 0.4 | Turkey | 8,328 | 137 | 103748 | 0.9 |
| Denmark | 6,071 | 1574 | 3060 | 3.0 | Monaco | 20 | 6 | 388 | 0 | Uganda | 158 | 3,885 | 563 | 0.6 |
| Ecuador | 60 | 4 | 8851 | 0.2 | Mongolia | 25 | 110 | 2542 | 0.3 | Ukraine | 1,727 | 6 | 16,836 | 0.7 |
| Egypt, Arab | 2,515 | 618 | 439 | 0.4 | Montenegro | 28 | 20 | 94 | 0.4 | United Arab | 324 | 2,649 | 3,208 | 0.5 |
| Estonia | 514 | 62 | 888 | 2.4 | Morocco | 386 | 169 | 5490 | 0.7 | United King | 46,035 | 26 | 31,253 | 1.8 |
| Ethiopia | 170 | 1 | 42 | 0.2 | Myanmar | 9 | 42 | 4007 | 0.1 | United State | 208610 | 15,343 | 256,775 | 2.8 |
| Finland | 4,878 | 1650 | 3,403 | 3.8 | Nepal | 64 | 8 | 2,204 | 0.3 | Uruguay | 290 | 247,750 | 2,458 | 0.4 |
| France | 31,686 | 14,655 | 85,713 | 2.2 | Netherlands | 15,508 | 2,585 | 1023 | 2.0 | Venezuela | 302 | 20 | 11,066 | 0.4 |
| Georgia | 118 | 138 | 841 | 0.2 | New Zealan | 3,472 | 1,501 | 8,632 | 1.3 | Vietnam | 432 | 282 | 22,376 | 0.5 |
| Germany | 46,259 | 46,986 | 60,606 | 2.9 | Nigeria | 439 | 64 | 20,560 | 1 | Yemen Rep. | 33 | 300 | 2,191 | 0.6 |

Figure 1. Snapshot of dataset

Table 2. Statistical Analysis of R&D attributes

| | R&D expenditure, % of GDP | Journal Articles | Patent Applications | Trademark Applications |
|---|---|---|---|---|
| **Minimum** | 0 | 3 | 1 | 2 |
| **Maximum** | 4 | 208610 | 415829 | 1273827 |
| **Mean** | 0.957 | 7016.791 | 11236.261 | 27336.983 |
| **Standard Deviation** | 0.999 | 22657.798 | 53425.323 | 122676.98 |

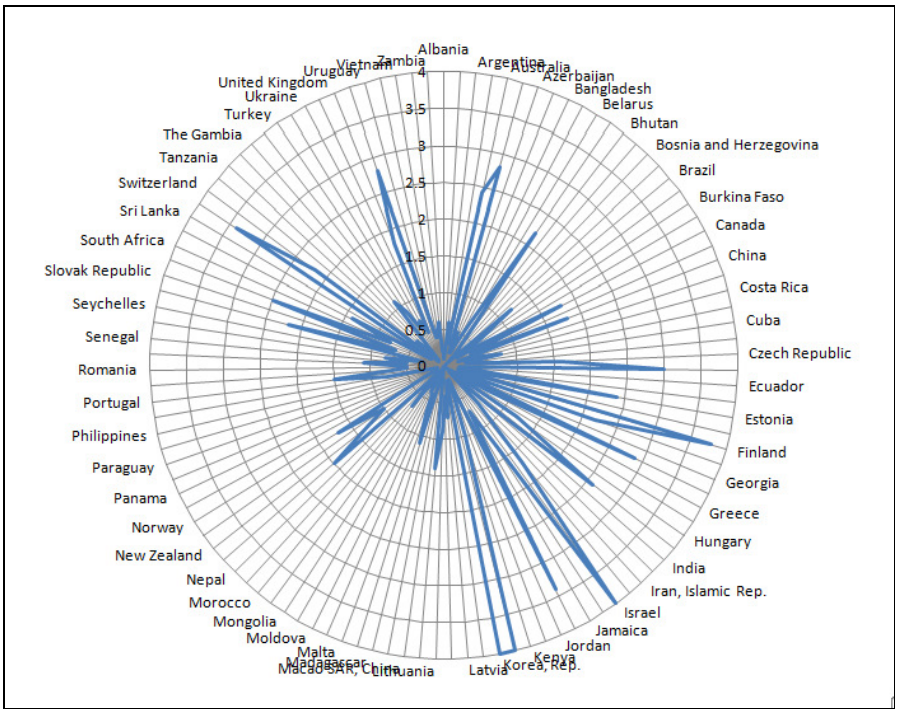


Figure 2(a). Radar chart for countries' R&D expenditure in terms of percentage of GDP

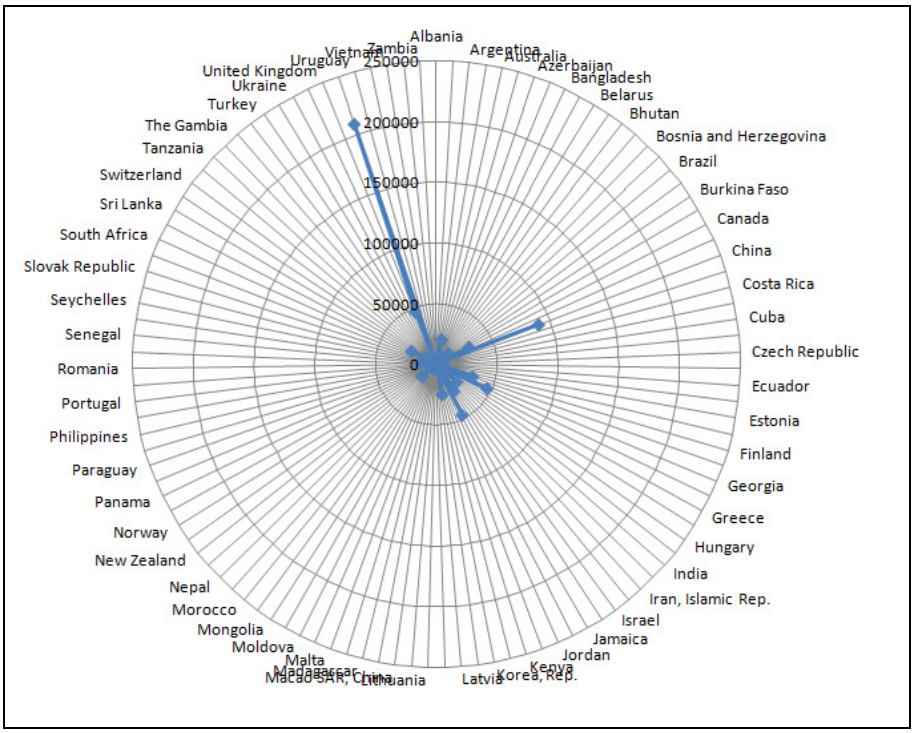


Figure 2(b). Radar chart for countries' number of Scientific and technical journal articles

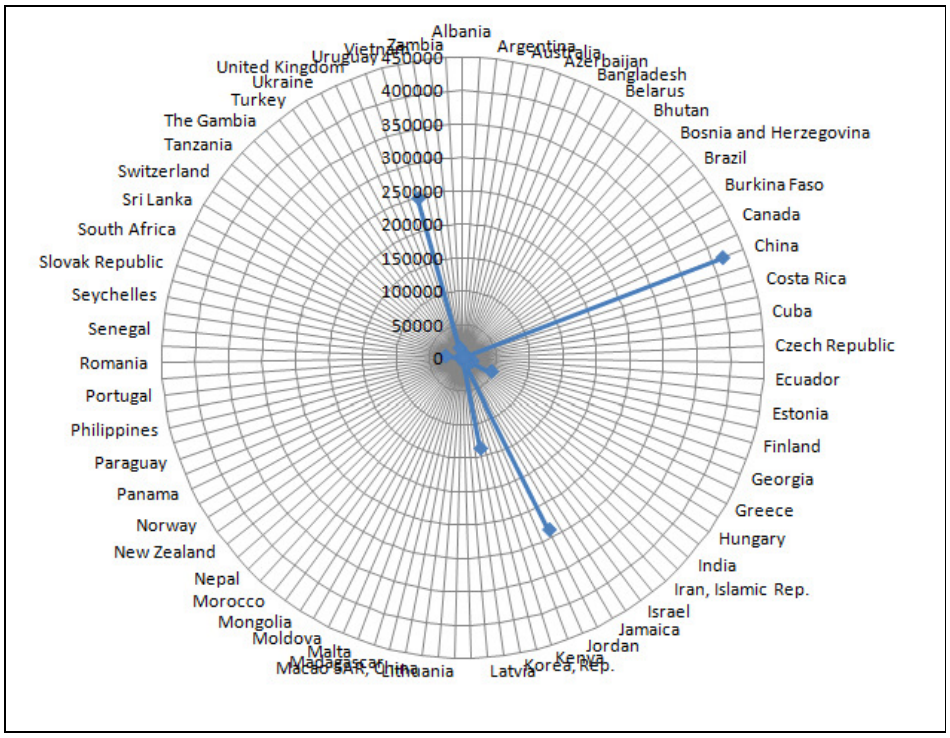


Figure 2(c). Radar chart for countries' number of Patent applications, residents

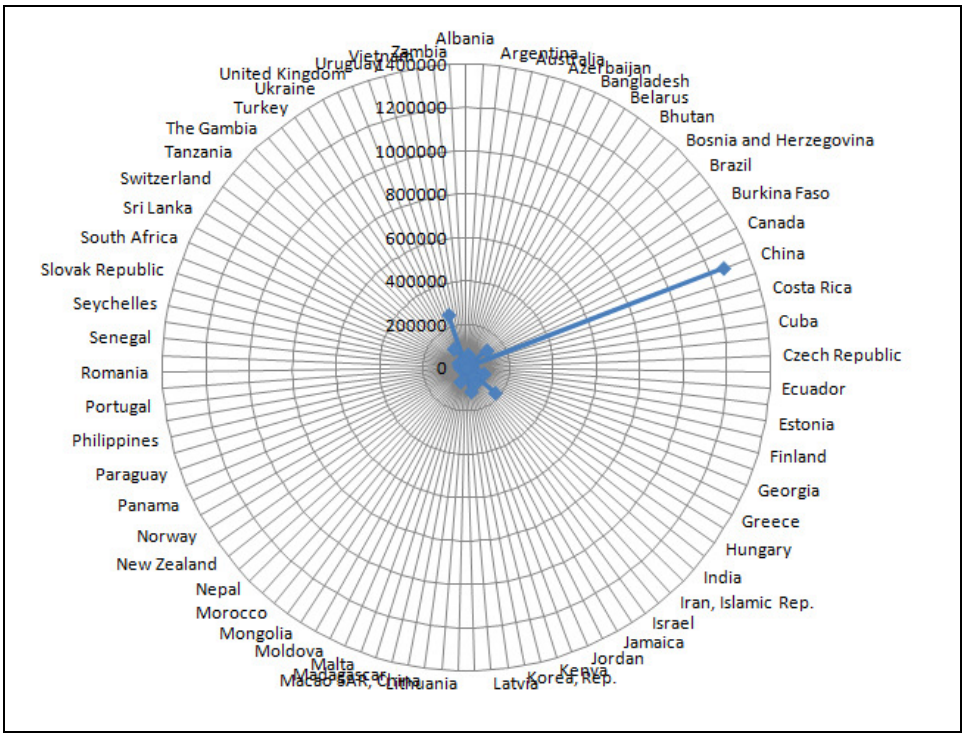


Figure 2(d). Radar chart for countries' number of Trademark applications, direct resident

## 3. K-MEANS CLUSTERING: THEORETICAL CONSIDERATIONS

K-Means is a simple unsupervised learning algorithm for cluster design and analysis. The aim of this algorithm is to find the best split of N entities in to K groups, so that the total distance between the members of group and its corresponding centroid, representative of the group, is minimized. Thus the goal is to partition the N entities into K sets $S_i$, i=1, 2, ..., K in order to minimize the sum of squares error within cluster [7]. This error is defined as:

$$E = \sum_{i=1}^{K} \sum P \subset Ci \, |p \quad mi|^2 \qquad (1)$$

Where E is the sum of the square error for all points in the data set; p is the point in data space representing a given object; and the mean of cluster $S_i$ is mi. In other words, for each data item in each cluster, the distance from the event to its cluster center is squared, and the distances are summed.

In this technique, clusters are dependent on the choice of the initial cluster centroids. Randomly K data items are selected as initial cluster centers followed by the distances of all points are calculated by Euclidean distance formula. Data items having less distance to centroids are moved to the appropriate cluster. This process is continued until no more alterations occur in clusters. The Figure 3 shows basic K-mean clustering algorithm [7].

INPUT: Number of desired clusters K

Students Records D= {$d_1$, $d_2$...$d_n$}

OUTPUT: A set of K clusters

Steps:

1. Randomly select k data objects from data set D as initial centers.
2. Repeat;
   a. Calculate the distance between each data object $d_i$ (1<= i<=n) and all k clusters Cj(1 <= j<=k) and assign data object $d_i$ to the nearest cluster.
   b. Calculate for each cluster Cj(1 <= j<=k) recalculate the cluster center.

   Until no change in the center of clusters.
3. Time complexity of K-mean Clustering is represented by

   O(nkt)

   Where n is the number of objects, k is the number of clusters and t is the number of iterations.

Figure 3. K-means Clustering Algorithm

## 4. CLUSTER DESIGN AND ANALYSIS

In regard to the scenario mentioned in the introduction, to analyze countries similarity and assign them to the clusters, the R&D attributes taken into account. The k-means algorithm is a technique for grouping entities according to the similarity of their attributes [6]. As the presenting problem consists of dividing countries into similar groups, it is plausible that K-means can be applied to this task. As observed in Figure 4 three clusters are created to classify datasets in to three categories.

The implementation of K-means generated three clusters, consisting of 9, 83 and 23 countries. Corresponding details are tabulated in Table 3. Figure 5 gives scatter chart of cluster density generated in Weka. Analyzing the cluster means, we can relate each group with each of the three classes of countries:

- Cluster 0 formed by countries has highest R&D expenditure, Patent applications and Trademark applications and medium in Journal Publications
- Cluster 1 formed by countries has lowest R & D expenditure as well as lowest in Journal Publications and Trademark Applications. But they have medium number of Patents
- Cluster 2 formed by other countries has medium R & D expenditure as well as medium in Trademark Applications. These countries have highest Journal Publications and lowest in Patents

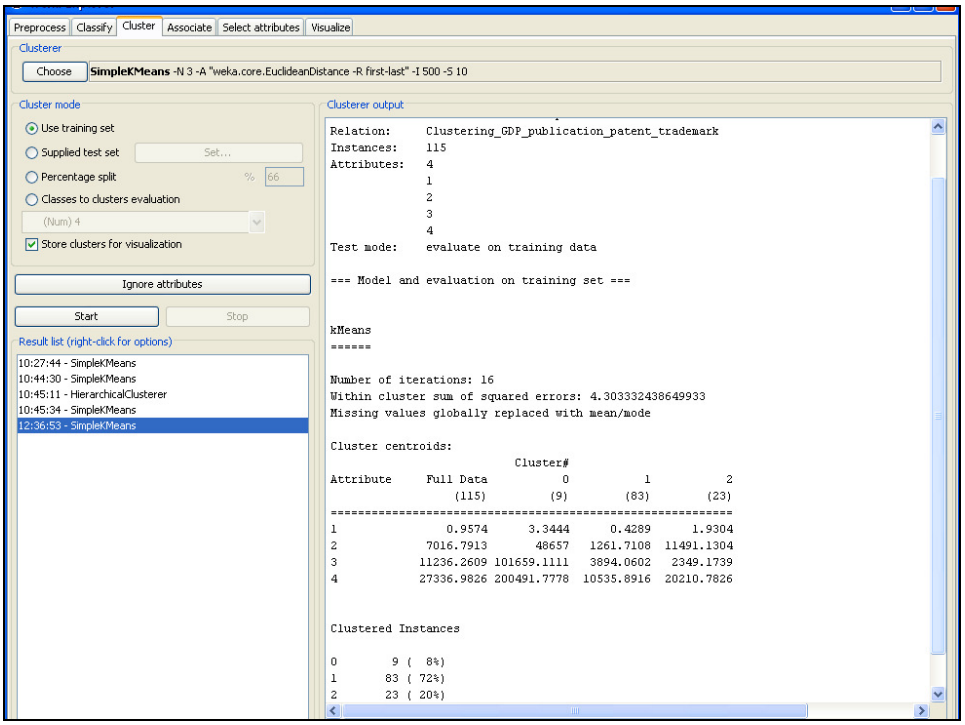


Figure 4. Clustering Results in Weka

Table 3. Clusters of countries

| **Cluster 0** | **Cluster 1** | | **Cluster 2** |
|---|---|---|---|
| China | Albania | Macao SAR, | Australia |
| Finland | Algeria | Macedonia, | Austria |
| Germany | Argentina | FYR | Belgium |
| Israel | Armenia | Madagascar | Brazil |
| Japan | Azerbaijan | Malaysia | Canada |
| Korea, Dem. Rep | Bahrain | Malta | Czech Republic |
| Korea, Rep. | Bangladesh | Mexico | Denmark |
| Sweden | Barbados | Moldova | Estonia |
| United States | Belarus | Monaco | France |
| | Bhutan | Mongolia | Hungary |
| | Bolivia | Montenegro | Iceland |
| | Bosnia and | Morocco | Ireland |
| | Herzegovina | Myanmar | Italy |
| | Botswana | Nepal | Luxembourg |
| | Bulgaria | Nigeria | Netherlands |
| | Burkina Faso | Pakistan | New Zealand |
| | Cambodia | Panama | Norway |
| | Chile | Paraguay | Portugal |
| | Colombia | Paraguay | Singapore |
| | Costa Rica | Peru | Slovenia |
| | Croatia | Philippines | Spain |
| | Cuba | Poland | Switzerland |
| | Cyprus | Qatar | United |
| | Ecuador | Romania | Kingdom |
| | Egypt, Arab Rep. | Russian | |
| | Ethiopia | Federation | |
| | Georgia | Senegal | |
| | Greece | Serbia | |
| | Guatemala | Seychelles | |
| | **India** | Slovak | |
| | Indonesia | Republic | |
| | Iran, Islamic Rep. | South Africa | |
| | Jamaica | Sri Lanka | |
| | Jordan | Tajikistan | |
| | Kazakhstan | Tanzania | |
| | Kenya | Thailand | |
| | Kyrgyz Republic | The Gambia | |
| | Latvia | Tunisia | |
| | Lebanon | Turkey | |
| | Lithuania | Uganda | |
| | | Ukraine | |
| | | United Arab | |
| | | Emirates | |
| | | Venezuela | |
| | | Vietnam | |
| | | Yemen Rep. | |
| | | Zambia | |

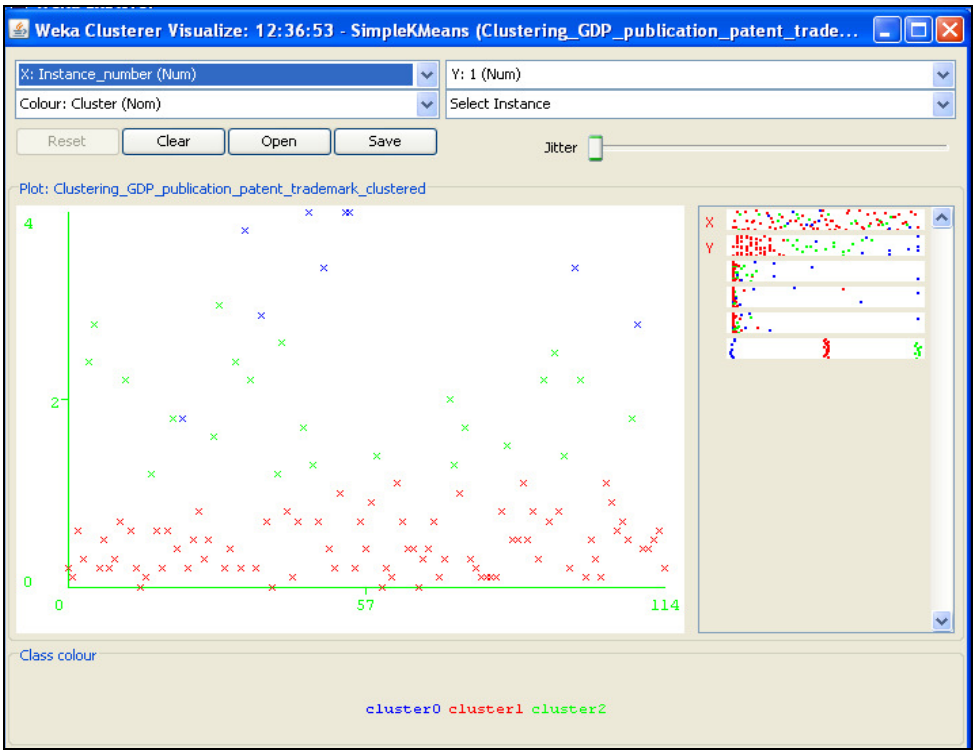


Figure 5. Cluster density

## 3. CONCLUSIONS

Our investigation using the K-means clustering algorithm in fact one of its kinds systematic modus operandi for perceiving the performance metrics for the benefit of the policy makers, scientific community and the society at large. This study has analyzed the number of research publications, patent applications and trademarks registered with reference to percentage of GDP spending on R&D. Unsupervised learning algorithm used for designing three clusters of countries based on these dataset. Countries belonging to cluster-0 should focus on increasing number of journal publications. Cluster 1 formed by countries must re-plan their R&D funds to motivate researchers in increasing research productivity.

## REFERENCES

[1] 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.

[2] Meo SA and Usmani Adnan Mahmood, Impact of R&D expenditures on research publications, patents and high-tech exports among European countries, European Review for Medical and Pharmacological Sciences, 2014, 18, 1-9.

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

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

[5] Cs.waikato.ac.nz, Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Retrieved 9th September 2015, from http://www.cs.waikato.ac.nz/ml/weka/

[6] Jiawei Han, Micheline Kamber and Jian Pei, Data Minining Concepts and Techniques, Third Edition, 2012 Elsevier Inc.
[7] R.S.Kamath, R.K.Kamat, Educational Data Mining with R and Rattle, River Publishers Series in Information Science and Technology, River Publishers, Netherland, 2016
[9] Jacob, Brian A and Lefgren, Lars, The impact of research grant funding on scientific productivity, Journal of Public Economics, 95, 2011, 1168-1177.
[10] McAllister, Paul R and Wagner, Deborah Ann, Relationship between R&D expenditures and publication output for U.S. colleges and universities, Research in Higher Education, 15, 1981, 3-30.
[11] Bozeman, B., & Melkers, J. Evaluating R & D impacts. Boston: Kluwer Academic. 1993.
[12] Gaillard, J. Measuring Research and Development in Developing Countries: Main Characteristics and Implications for the Frascati Manual. Science Technology & Society, 15(1), 2010, 77-111. doi:10.1177/097172180901500104