Geographical Distribution of Biomedical Research in the USA and China

|  |  |  |
| --- | --- | --- |
| Yingjun Guan School of Information Sciences University of Illinois at Urbana-Champaign Champaign, IL 61820, USA yingjun2@illinois.edu | Jing Du School of Information Sciences University of Illinois at Urbana-Champaign Champaign, IL 61820, USA jingdu4@illinois.edu | Vetle I. Torvik School of Information Sciences University of Illinois at Urbana-Champaign Champaign, IL 61820, USA vtorvik@illinois.edu |

ABSTRACT[[1]](#footnote-2)

We analyze nearly 20 million geocoded PubMed articles with author affiliations. Using K-means clustering for the lower 48 US states and mainland China, we find that the average published paper is within a relatively short distance of a few centroids. These centroids have shifted very little over the past 30 years, and the distribution of distances to these centroids has not changed much either. The overall country centroids have gradually shifted south (about 0.2° for the USA and 1.7° for China), while the longitude has not moved significantly. These findings indicate that there are few large scientific hubs in the USA and China and the typical investigator is within geographical reach of one such hub. This study sets the stage for comparing the centralization of biomedical research at national and regional levels across the globe, and over time.

CCS CONCEPTS

• **Information Systems** → **Geographic information systems**

• **Applied computing** → **Digital libraries and archives**

KEYWORDS

Geocoding, author affiliation, bibliographic databases, clustering

ACM Reference format:

Y. Guan, J. Du, and V. I., Torvik. 2017. Geographical Distribution of Biomedical Research in the USA and China. In *Proceedings of WOSP 2017, June 19, 2017, Toronto, ON, Canada*, 8 pages.

DOI: 10.1145/3127526.3127534

1 INTRODUCTION

During the past three decades, there has been an explosive growth and geographical spread of the scientific literature[1]. To explore these changes and create an updated framework for innovation-oriented subjects and federal funds, there has been an intense research activity in studying the geographic distribution of scientific activities [2]. A variety of studies have examined city concentrations [3,4], countries [5,6,7], subject areas or journals [8,9], and innovation as reflected in patents[10,11,12] and industrial activity [13]. This research suggest that linkages between research affiliations are fostered by geographic proximity, and geographic distance is an obstructive factor in achieving collaborations. Therefore, it is of significance to analyze the geographical proximity of localities to regional hubs.

To analyze the geographical proximity and centroid movement of biomedical literature in USA, this study uses bibliometric methods to investigate the geocode of localities of 1988~2016 publications in PubMed, calculate the centroid of affiliation localities of the lower 48 states. The longitudes and latitudes of centroid and average distances between each locality from the centroid over the 30 years are observed. Clustering methods are also conducted to examine geographic proximity relative to centroid representations. and calculate the average distance from the hub of each cluster.

2 DATA

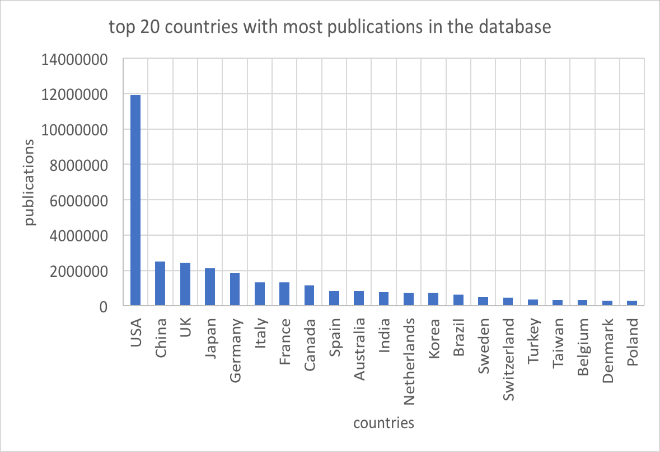
MapAffil[14] contains disambiguated and geocoded place names in 37 million affiliations from nearly 20 million PubMed articles published between 1867 to 2016. PubMed started indexing first-author affiliations in 1988 and all authors’ affiliations in 2014. Therefore, the availability of geospatial data surges in 1988 and again in 2014. MapAffil covers a significant portion affiliations missing from PubMed. These were harvested from external sources including PubMed Central, Microsoft Academic Graph (MAG), Astrophysics Data System (ADS), and NIH grants. The geographical data of each article is identified by MapAffil, which maps an author’s affiliation to its city and the corresponding city-center geocode (the longitude and latitude) across 227 countries and territories worldwide. MapAffil has a high overall performance and provides additional geo-linked data e.g., via US FIPS codes. It is worth noting that place name disambiguation is a non-trivial process complicated by low-precision of many relatively rare place names such as the ones listed in Table 1, and the ambiguity of highly frequent place names, such as the ones listed in Table 2. Some place names have low precision (when the name points to multiple places or things that are not place names) and low recall (when the place has variant names or sub-divisions). The top 20 most common countries are shown in Figure 1. Among 37 million affiliations, there are ~11 million in the USA (lower 48 states) and ~3 million in China (mainland) during the period of 1988-2016. For each publication, each city is counted once when multiple coauthors are from the same city. The large area and geographical diversity of both countries make it an interesting subject of study in terms of geographic distribution and clustering.

**Table 1:** A Sample of very low-precision place names

|  |  |  |
| --- | --- | --- |
| University, MS, USA | Harvard, MA, USA | Rome, NY, USA |
| Usa, Oita, Japan | Cambridge, WI, USA | Mayo, YT, Canada |
| Institute, WV, USA | Carolina, PR, USA | Sydney, NS, Canada |
| Center, CO, USA | Columbia, NJ, USA | Durham, CT, USA |
| York, NE, USA | Federal, NSW, Australia | King, NC, USA |
| London, KY, USA | Ontario, OR, USA | Melbourne, AR, USA |
| Boston, VA, USA | Denmark, SC, USA | Yale, MI, USA |
| Washington, TX, USA | Poland, ME, USA | Madison, GA, USA |
| Street, Somerset, UK | Hopkins, MN, USA | Wales, WI, USA |
| North, VA, USA | Rochester, MO, USA | Indiana, PA, USA |
| Paris, IL, USA | Florida, NY, USA | Athens, IN, USA |
| Oxford, IA, USA | Mexico, NY, USA | DE, USA; IN, USA |

**Table 2:** Ambiguity of the top most frequent place names

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| City | Prec. | Rec. | % of nation output | % of nation population |
| London, UK | 91.8 | 91.5 | 29.2 | 13.7 |
| New York, NY | 68.0 | 87.6 | 5.5 | 2.7 |
| Boston, MA | 98.7 | 93.3 | 5.1 | 0.2 |
| Paris, France | 99.0 | 68.7 | 39.5 | 18.5 |
| Tokyo, Japan | 94.7 | 97.4 | 20.7 | 10.2 |
| Beijing, China | 99.4 | 68.7 | 18.2 | 1.6 |
| Seoul, Korea | 97.1 | 99.3 | 48.8 | 19.8 |
| Baltimore, MD | 99.5 | 94.9 | 97.2 | 2.7 |
| Philadelphia, PA | 99.5 | 95.1 | 97.2 | 2.7 |
| Los Angeles, CA | 99.6 | 86.5 | 92.6 | 2.6 |



**Figure 1:** The t**op 20 countries in MapAffil.**

3 RESULTS AND DISCUSSION

3.1  Centroids and distances

*3.1.1 Geographical centroid.* For every affiliation in the corpus, the longitude and latitude of its city have been identified and recorded. Given the assumption of Euclidean Geometry, where the longitudes and latitudes are perpendicular to each other and form a plane, for the collections of cities in the Lower 48 states of United States, the central point can be calculated by averaging their latitude and longitude. Localities outside the “Lower 48” (the 48 states other than Alaska and Hawaii) will cause significant influence in calculating the centroid of United States because of the long distance in between. Only the affiliations inside the “Lower 48” are taken into the consideration while calculating the geographical centroid. Those states and territories will be analyzed separately in future work. Furthermore, the method of locating the geographical centroid and calculating the variability in the following section can be applied to not only the “Lower 48”, but also USA or other countries, territories and collections of areas.

*3.1.2 Variability calculation*. With the geocode (longitude and latitude) of geographical centroid acquired, the average distance from all cities to the centroid can be calculated. There are different methods to calculate the distance: The Eucilid’s distance is based on the assumption of plane space; the Great Circle Distance is treating the Earth as a perfect globe and the distance as an arc; the Vincenty distance is based on the assumption of the Earth being an oblate spheroid. In this section, Vincenty distance is selected because of its accuracy to the real situation. Then the distance from each city to the geographical centroid can be calculated by applying the following equation, where is the distance from an individual locality to the centroid, *n* is the total number of localities in each calculation, is the average distance.

3.1.3 *The change of centroid and corresponding average distance* *over time.* In this section, for each year, the migration of the geographical centroid and the average distance will be calculated and analyzed from 1988 to 2016.

3.2 Clustering

*3.2.1 K-Means clustering.* K-Means is a common clustering method in machine learning. In this method, all the data get clustered into k clusters with the k-value predefined by the researcher. First, a group of randomly selected k initial centroids are set, and all the points in the database got clustered to their closest centroid; Then the initial centroids move towards the real centroids of the current clusters, and afterwards, all the data got clustered again with the newly established centroids. The iteration goes until all members are stably clustered in k groups, with the centroids being exactly their centroids. To eliminate the influence of initial centroids, the K-Means clustering will be repeated for several times to check the robustness of the final clustering.In this section, the affiliation cities within the “Lower 48” will be clustered by using K-Means method to demonstrate whether there are some city concentrations or active area in the United States. A change over years will also be provided to manifest the development.

*3.2.2 Determining the number of clusters, k.* The number of clusters, k, has salient significance in the clustering results. The value of k also influence the variability within clusters. When the number of cluster increases, there will be more centroids allocated for all the data points, and the variability, namely the average distance from each point to its nearest centroid will decrease. A threshold on the change of variability can be set before clustering, and the value k is decided when the increase of k does not cause a larger influence than the threshold.

In this section, the effects of the number of clusters will be studied and the average distance within clusters will be estimated. Therefore, the number of clusters and geographical distance can be analyzed for the affiliation localities on the “Lower 48” of United States.

4 Results

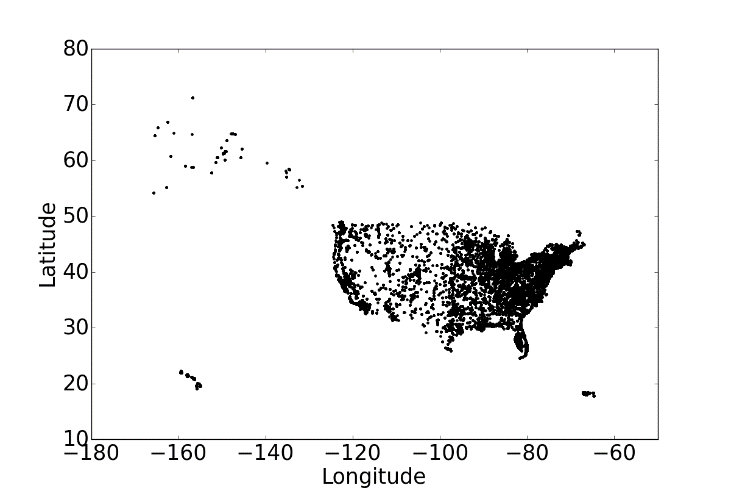
4.1 Geospatial distribution of all the scientific activities

For all research activities in the United States, especially the “Lower 48” area, there are millions of publications get born. There is also an increasing momentum from 1988 to 2016.



**Fig. 3.** **Number of publications in USA, 1867~2017**.

As shown in the Figure 3, the quantity of the US publication keeps increasing year by year. Note that the y-axis is shown on a log-scalearithm in y axis, the number of publication. The trend shows a relatively clear linear characteristic.



**Fig. 4. Spatial distribution of publications in USA and its territories.**

When collecting all the publication affiliations in the United States, we can find not only the affiliations from “Lower 48”, but also those from Alaska, Hawaii and other territories. The Lower 48 has a majority of the publications, and in the rest of the paper, only the Lower 48 area will be analyzed (It is recommended to analyzed separately if by interest). In the figure, each dot indicates the city of the affiliation. In this paper, the city of the affiliation is the unit to be analyzed. Considering the phenomenon of co-authorship, if multiple authors from the same affiliation contributes to the same publication, the affiliation will be counted as one. For example, if one publication has one author from affiliation A and three co-authors from affiliation B, then, affiliation A and B will be counted as one in calculation for participating the research activity.

The density map of the publications on the “Lower 48” is also shown as below:



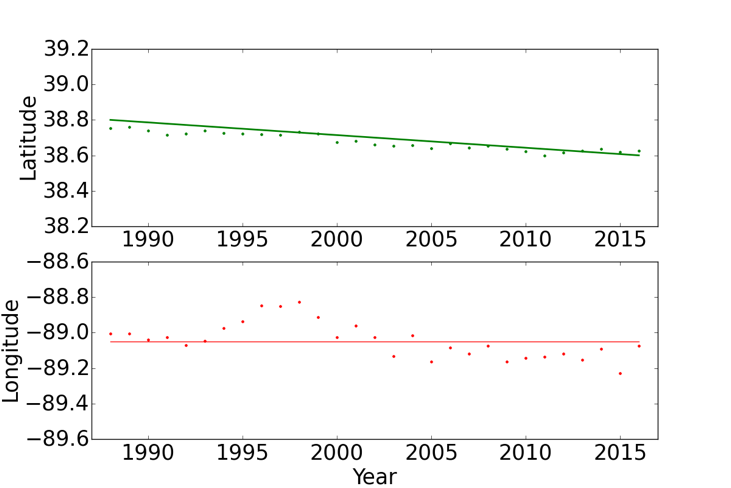
Fig. 5. Density Map of publications in “Lower 48” of USA of all the time from 1988 to 2016.

In this figure, as the color bar shows at the right side, the darker the color is, the place enjoys the more density of the publication affiliation. To better emphasize the difference of densities, the color bar is plotted in logarithm.

From the map, it can be observed that there are some high density areas in the northeastern coast, around Boston, and southwestern coast, around Los Angeles. In the West mountain area, around Nebraska, the publication activities are not as frequent as other areas.

4.2 Centroid and average distance of the affiliations

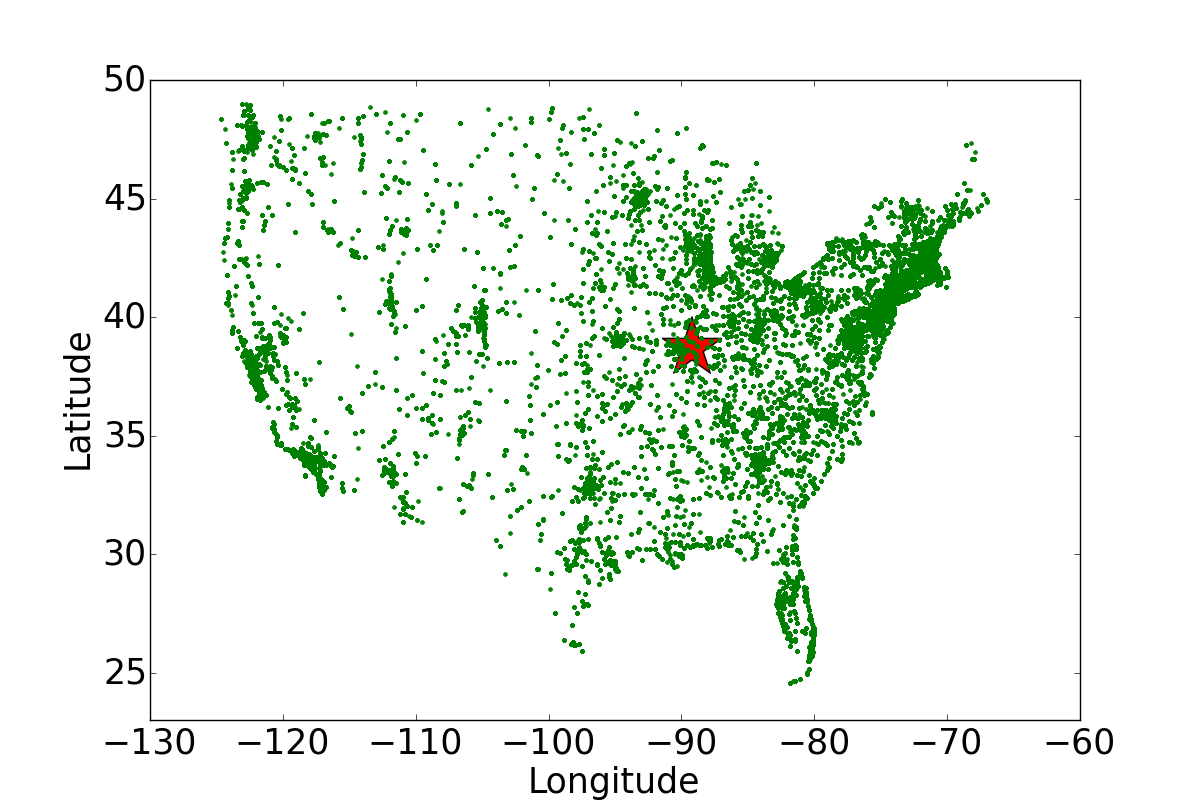
With an assumption of Euclid’s plane, this paper takes the mean of all the longitudes and latitudes to estimate the geographical centroids of the cities of affiliations within the Lower 48 as a unity. The data and the change in each year is calculated and plotted in the following figure.



**Fig. 6. The change of averaged latitude, averaged longitude over time from 1988 to 2016**

The figure above shows the trend of latitudes and longitudes over time. The latitude over years demonstrate a tendency of decreasing (moving south), and it moved around 0.2 degree which is around 20 miles. Although there are some fluctuation years in the longitudes development, generally speaking, the average longitude keeps stable throughout this three decades from 1988 to 2016. Averagely, the longitude of centroid of all time is around -89.0, with a deviation no more than 0.2 degree.

The figure below demonstrates the centroid together with all the affiliations. The star is where the centroid is, whose location is (-89.2, 38.7), geographically seated in the state of Illinois. The arrow in the figure indicates that the centroid is moving from north to south during decades.

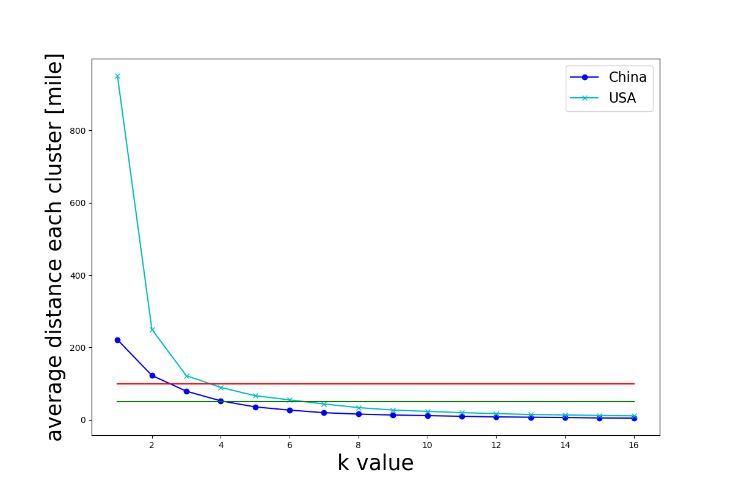
~~~~

**Fig. 7. Geospatial movement of centroid in USA.**

**4.3 The clustering of the US “Lower 48” scientific activities**

Instead of treating the “Lower 48” as one cluster, we can also study whether dividing the area into many clusters can help better understand the research distribution.

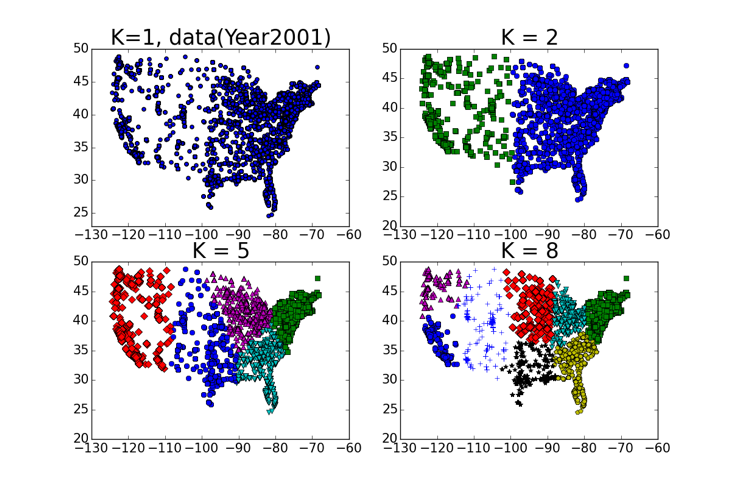
The figure below shows the relationship between k values and the average miles from each city to its closest centroid. It can be seen that the more clusters there are in dividing the area, the lower the variability (average distance from each point to its closest cluster centroid) will drop. After the k reaches 3 or 4, even the k value increases, the variability sustains. This is because most data will keep the same cluster even with one new centroid.



**Fig. 8. Average distance vs different k values in clustering.**

The red line in the figure above is around 100 miles’ variability, and the green line in the figure indicates 50 miles’ variability, which are reasonable distances for researchers to move frequently. In a hub or clusters with a radius of 50 to 100 miles, it is reasonable for researchers to form stable and easy connections; on the contrary, for example if the variability is around 1000 miles (almost half the length of United States), it would be impossible for researchers to keep close attach and collaboration.

The figure below demonstrates the clustering results. With more clusters, the lower variability gives the researchers better environment to communicate and collaborate.



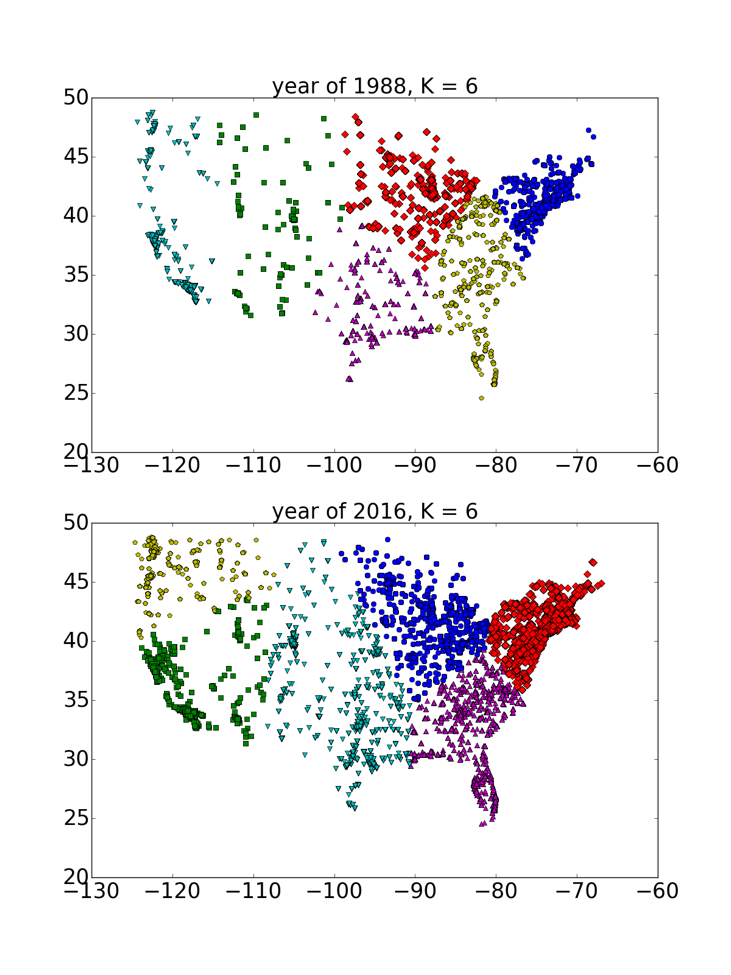
**Fig. 9. Visualizations of clustering results with different Ks**

Another interesting fact is that as the time develops from 1988 to 2016, although the quantity of the publications increases dramatically, the clustering conditions and the variability keeps almost the same!



**Fig. 10. Comparison of variability vs k values in the year of 1988 and 2016.**

As shown in the figure above, the down triangle markers stand for the data of year 1988 and the up ones stand for the data in year 2016. it can be noticed that there is some difference in the variability when k=1, and the average distance keeps almost the same when the k increases to 7. This phenomenon can be explained that the quantity within each cluster keeps increasing with a same velocity. The clustering itself changes very little.



As analyzed before, the figure above shows the real clustering in the year of 1988 and 2016. Regardless of the difference in number of publication, we can clearly name the 6 clusters of the Lower 48 area in the United States: Northeastern Area (1 cluster), South Eastern Area (1 cluster), Mid-West Area (1 cluster), Mid-South Area (1 cluster), west coaster Area (2 clusters).

ACKNOWLEDGMENTS

Research reported in this publication was supported in part by NIH National Institute on Aging P01AG039347. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

|  |  |
| --- | --- |
| [1] | M. Grossetti, D. Eckert, Y. Gingras, et al. 2014. Cities and the geographical deconcentration of scientific activity: A multilevel analysis of publications (1987–2007). Urban Studies, 51(10), 2219-2234. |
| [2] | K. Frenken, S. Hardeman, and J. Hoekman. 2009. Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics, 3(3), 222-232. |
| [3] | L. Leydesdorff and O. Persson. 2010. Mapping the geography of science: Distribution patterns and networks of relations among cities and institutes. Journal of the Association for Information Science and Technology, 61(8), 1622-1634. |
| [4] | L. Bornmann and L. Leydesdorff. 2011. Which cities produce more excellent papers than can be expected? A new mapping approach, using Google Maps, based on statistical significance testing. Journal of the Association for Information Science and Technology, 62(10), 1954-1962. |
| [5] | T. Chiarini, V.P. Oliveira, and F.C. do Couto e Silva Neto. 2013. Spatial distribution of scientific activities: An exploratory analysis of Brazil, 2000–10. Science and Public Policy, 41(5), 625-640. |
| [6] | P. Zhou, B. Thijs, and W. Glänzel. 2009. Regional analysis on Chinese scientific output. Scientometrics, 81(3), 839-857. |
| [7] | R. Garcia, V. Araujo, and S. Mascarini. 2013. The role of geographic proximity for university-industry linkages in Brazil: An emprical analysis. Australasian Journal of Regional Studies, 19(3), 433. |
| [8] | P.S. Pagel, and J.A. Hudetz. 2013. A bibliometric analysis of geographic publication variations in the Journal of Cardiothoracic and Vascular Anesthesia from 1990 to 2011. Journal of cardiothoracic and vascular anesthesia, 27(2), 208-212. |
| [9] | M. Palmblad and V.I. Torvik. 2017. Spatiotemporal analysis of tropical disease research combining Europe PMC and affiliation mapping web services. Tropical medicine and health, 45(1), 33. |
| [10] | F. Liu, and Y. Sun. 2009. A comparison of the spatial distribution of innovative activities in China and the US. Technological Forecasting and Social Change, 76(6), 797-805. |
| [11] | M.H. Huang, H.W. Chang, and D.Z. Chen. 2012. The trend of concentration in scientific research and technological innovation: A reduction of the predominant role of the US in world research & technology. Journal of Informetrics, 6(4), 457-468. |
| [12] | W. Hong, and Y.S. Su. 2013. The effect of institutional proximity in non-local university–industry collaborations: An analysis based on Chinese patent data. Research Policy, 42(2), 454-464. |
| [13] | M.P. Devereux, R. Griffith, and H. Simpson. 2004. The geographic distribution of production activity in the UK. Regional science and urban economics, 34(5), 533-564. |
| [14] | V.I. Torvik. 2015. MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. In D-Lib magazine: the magazine of the Digital Library Forum, 21(11-12). |
|  |  |

1. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

   WOSP 2017, June 19, 2017, Toronto, ON, Canada

   © 2017 Copyright is held by the owner/author(s). Publication rights licensed to ACM 978-1-4503-5388-5/17/06…$15.00

   DOI: 10.1145/3127526.3127534 [↑](#footnote-ref-2)