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

Through PubMed, a worldwide database on biomedical research, we analyze nearly 19.2 million geocoded PubMed articles from 1867 to 2016. The United States and China are the two countries with the most publications in record. By visualizing the distribution of the publications in both countries and conducting statistical analysis, it can be concluded that both USA and China have their unbalanced development in western and eastern parts. The national centroid, the mean of all national observations, of both countries locates in the state of Illinois and the province of Shandong, respectively. With the time development, the national centroid of all publications is moving gradually southwards in both countries (0.2 degree in USA and 1.7 degree in China), while the longitude has no obvious moving tendency. Using K-means clustering method, the publications inside the “lower 48” US states and the mainland China are clustered with different K values. And in the US, when K = 4, the average distance from each publication to its closest centroid is around 100 miles; and when the k value increase to 7, the distance is less than 50 miles. The top cities with the most publications are also studied: They are always or are close to the centroid of capacity-intensitive clusters. The top cities in China have a much larger increase rate than that of top cities in USA. 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.

KEYWORDS

geographical distribution, affiliation, average distance, clustering

1 INTRODUCTION

During the past three decades, there has been an explosive growth and geographical spread of the medical literature [1]. For instance, from 1983–1984 to 2013–2014, the number of participating countries in scientific literature on headache increased from 26 to 67 [2]. 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. Some previous researchers used patent data to identify the trend of scientific activities, and examine the movement of geographical patterns [3][4]. Their studies focus on innovation activities, but the patent data set is relatively small (2000~5000) compared to datasets of scholarly publications. Some other studies used publication datasets to analyze the geographic distribution of scientific activities and examine the potential movement of city concentration [5-11]. These researches either analyzed the geographic distribution at a city level [5] [7-10], or at a country level [6][11], but few of them studies the agglomeration and geographic proximity between neighbor cities, which have more similarities in economics, environments and advantage to form collaborations. Previous researches argue that linkages between research affiliations are strongly fostered by geographic proximity [12], and geographic distance is an obstructive factor in achieving collaborations [13]. Therefore, it is of significance to analyze the geographical proximity of research affiliation and estimate the distance from the hubs as a foundation for building frequent academic collaboration between affiliations.

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 in the United States and the mainland China. The longitudes and latitudes of centroid and average distances between each locality from the centroid over the 29 years are observed. Clustering methods are also conducted to examine geographic proximity and calculate the average distance from the hub of each cluster.

2 DATA

We analyze the geographical data of biomedical publications using a data set of 19.4 million PubMed papers published from 1867 to 2017. The date of the publications can go back to 1867 when PubMedCentral is included in PubMed [14]. PubMed started indexing first-author affiliations in 1988 and all authors’ affiliations in 2014 [14]. Therefore, the availability of geospatial data surges in 1988 and again in 2014. Note that our data source (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 [14], 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. The top 20 most common countries are listed in Figure 1.

It can be concluded from Figure 1 that nearly 12 million (32%) authorships are from USA, including domestic and international collaborations. China, the country with the second most publications in the database, is also worth researching, because China shares a similarity size in geography with the US, but differs a lot in the publication numbers (2.9 million) and the population density, which might lead to a difference in the following analysis. The large area and geographical diversity of both countries make it an interesting subject of study in terms of geographic distribution and clustering.

In this paper, the study focuses on the publications from 1988 through 2016 in the lower 48 states in the United States and the mainland China, followed by a comparison at the end.

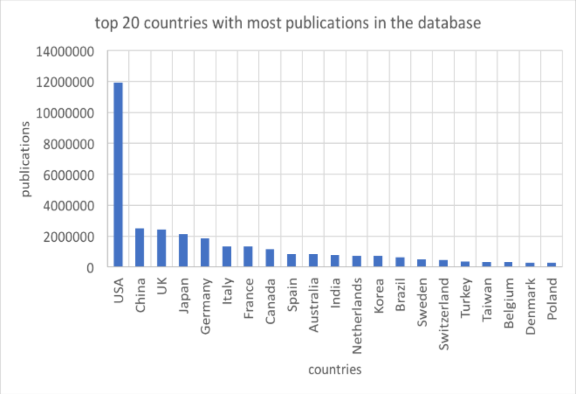


Fig 1. Top 20 countries with most authorships in the database

3 METHODS

3.1 Temporal and geospatial distribution of all the scientific activities in USA and China

By observing the scientific activities inside the United States, we can locate the affiliations not only in the lower 48 states of United States, but also in the State of Alaska, Hawaii, and some territories of the United States, for instance Guam, American Samoa, Puerto Rico, Virgin Islands, and so on. Similar to the United States, Apart from the mainland, China also consists of Taiwan, Hong Kong, Macaw and some other islands. Due to the political independence and geographical isolation of these areas, we only take the mainland into consideration.

In this section, the statistical descriptions will be provided, namely the distribution of longitude and latitude. The quantity and the density of localities of affiliations are all over the United States (both inside and outside the "Lower 48 states") and the mainland China. Within the recent 29 years (1988 to 2016), there is a considerable change in not only the quantity but also the geospatial distribution of the research publications.

In this paper, locality is the unit to be analyzed: each affiliation will get counted to the centroid of its city (or the suburb). And if multiple authors from the same city contributes to the same publication, considering the phenomenon of co-authorship, the locality will be counted as one. For instance, if one publication has one author from an affiliation in city A and three co-authors from different affiliations in city B, then, city A and B will both be counted as one in calculation for participating the research activity.

3.2 Centroid and average distance of the affiliations

3.2.1 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 and mainland China, the central point of each can be calculated by averaging their latitude and longitude.

Noticing that the 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. In this section, 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. Only the mainland China is taken into consideration for the same reason.

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" and the mainland China, but also other countries, territories and collections of areas.

3.2.2 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 Euclid'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. We use different methods to calculate the circumference of the earth, and find that Euclid distance has the largest error, the Great Circle Distance also has a moderate error, slightly larger than the Vincenty distance, while the Vincenty distance provides an approximate result to the real value.

Since USA and China are both giant countries with large span in latitude and longitude, 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 r_i is the distance from an individual locality to the centroid, n is the total number of localities in each calculation, $\bar{r}$ is the average distance.

$$\bar{r} = \sqrt{\frac{1}{n} * \sum_i r_i^2}$$

3.2.3 *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.

3.3 The clustering of the "Lower 48" scientific activities

3.3.1 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 remaining stable. 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" and mainland China will be clustered by using K-Means method to demonstrate whether there are some city concentrations or active area in the United States and China. A change over years will also be provided to manifest the development.

3.3.2 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 and mainland China.

4 Results

4.1 Temporal distribution of the scientific activities

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

The period starts from 1988 because the PubMed officially regulated the recording of locality since 1988 and the data earlier may contain some bias in spatial analysis.

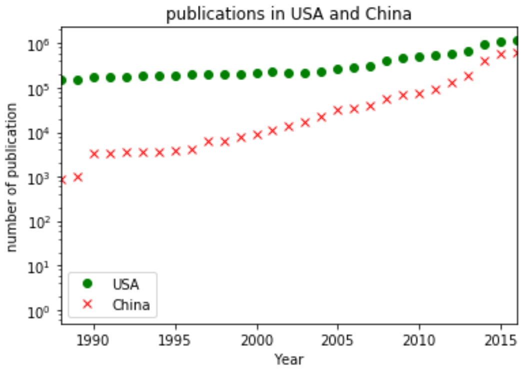


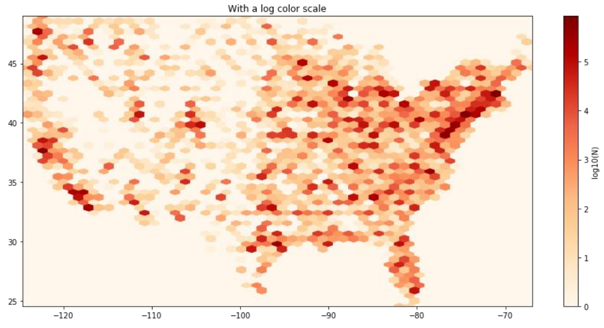
Fig 2. Number of publications in USA, 1988~2016

As shown in Figure 2, the quantity of the publications in the US and China keeps increasing from 1988 to 2016. Since publications in China is about 100 times less than that of US in the 1990s, the scale of y axis has been set logarithm to moderate the enormous difference between two countries' publication numbers in the 1990s. The trend shows a relatively clear linear characteristic, indicating an exponential tendency in the publication number of both countries.

4.2 Geospatial distribution of the scientific activities

When collecting all the publication affiliations in the United States, apart from the affiliations from "Lower 48", those from Alaska, Hawaii and other territories are also provided. However, 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 be 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, as explained in 3.1.

Similarly, although we have the data of Hong Kong, Macaw, and other islands, when analyzing publications in China, only mainland is taken into consideration. Indeed, this helps a lot in the following heatmap and the clustering section. Figure 4 shows the density map of the publications on the "Lower 48" and the mainland China.



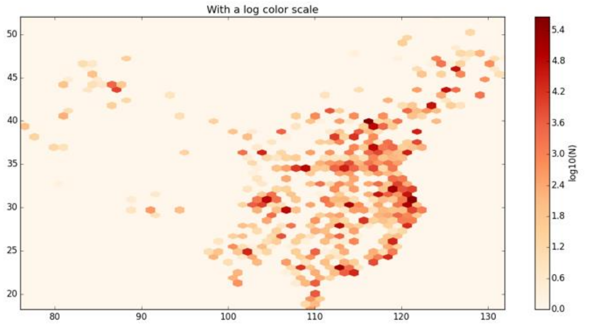


Fig 3. Density Map of publications in "Lower 48" of USA (up) and the mainland China (down) from 1988 to 2016.

In Figure 3, as the color bar shows at the right side, the darker the color is, the place has the more density of the publication affiliation. To better emphasize the difference of densities, the color bar is plotted in logarithm.

From the density map, it can be observed that in the US, 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.

And in China, the high-density areas are on the eastern coast, Beijing, Shanghai, Wuhan for instance. There's an obvious unbalance when comparing the west to the east. For most of the places in the west, especially southwest China, the publication activities are not as frequent as other areas, with only a few, or even none there.

4.3 Geospatial centroid of the scientific activities

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 and mainland China. The data and the movement in each year is calculated and plotted in Figure 4.

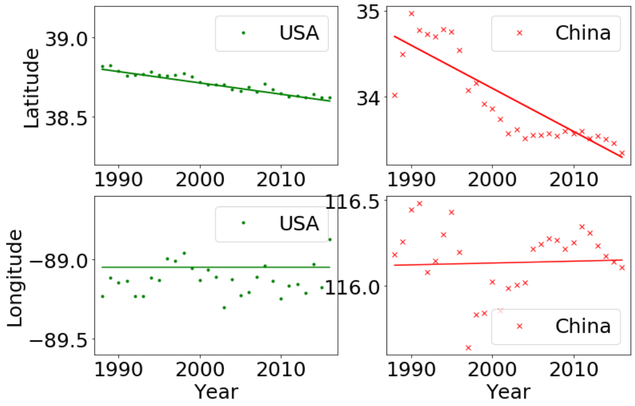


Fig 4. The change of average latitude, averaged longitude over time from 1988 to 2016

Figure 4 shows the trend of latitudes and longitudes over time. The latitude of USA's centroid 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, the average longitude keeps stable throughout the 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.

In mainland China, the latitude average also gets smaller, with a more significant degree, around 1.7 degree (to the south). There is no oblivious tendency in the change of longitude. In this case, we can conclude that the centroid in both countries are moving southward. However, although there might be some changes in specific year in west-east direction, no obvious trend can get concluded.

Figure 5 demonstrates the centroid with all the scientific activities. The star indicates the centroid, whose location is (-89.2, 38.7) for USA, geographically seated in the state of Illinois and (116.2, 34.7). For China, the centroid geographically seated in the province of Shandong. The arrow in the figure indicates that for both countries the centroid is moving towards south during decades (1988~2016).

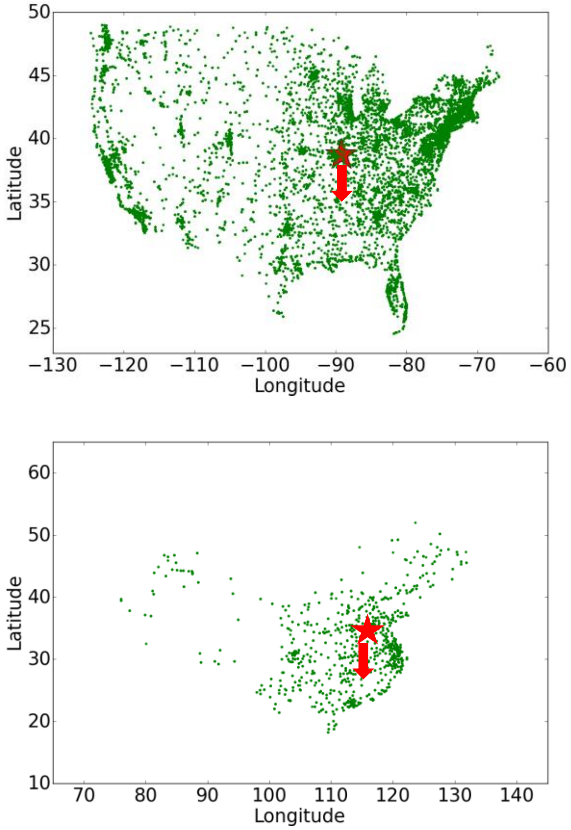


Fig 5. Geospatial movement of centroid in USA.

4.4 The clustering of the scientific activities

Instead of treating the “Lower 48” and the mainland China as one cluster respectively, we also study whether dividing the areas into many clusters may help better understand the research activities distribution. In the field of machine learning, K-Means is a most common method to divide the points into k clusters, as introduced in section 3.3.

Take the US. Lower 48 as an example. Figure 6 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 decrease slightly.

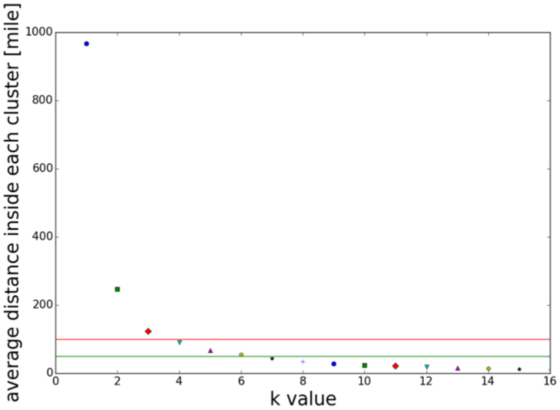


Fig 6. Average distance vs different k values in clustering

The red line in Figure 6 is around 100 miles’ variability, and the green line 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 connection and collaboration. We suggest that it is reasonable to set k = 7 to conduct clustering and estimate the variability of geographic proximities.

Figure 7 demonstrates the clustering results. With more clusters, the lower variability gives the researchers better environment to communicate and collaborate.

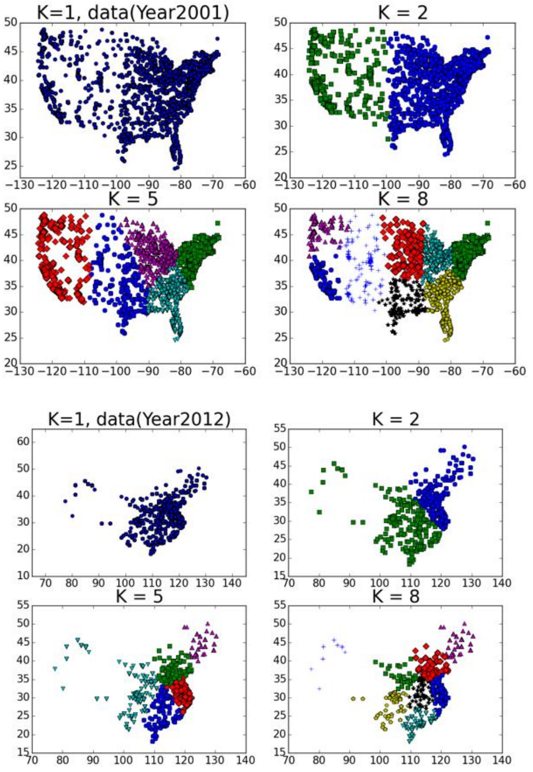


Fig 7. Visualizations of clustering results with different Ks

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.

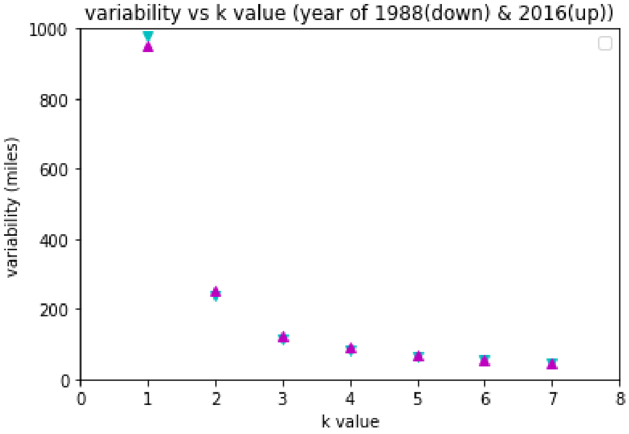


Fig 8. Comparison of variability vs k values in the year of 1988 and 2016.

As shown in Figure 8, 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 over time.

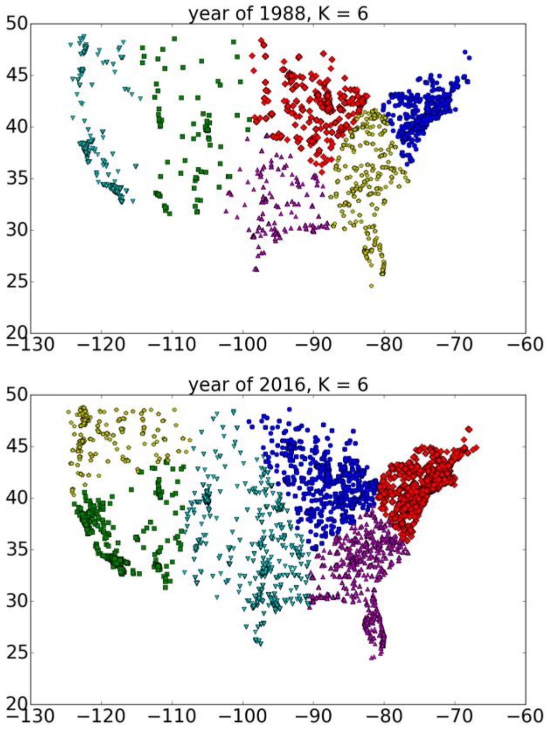


Fig 9. Comparison of clustering results in the year of 1988 and 2016. (k=6)

As analyzed before, Figure 9 shows the clustering in the year of 1988 and 2016. Regardless of the difference in number of publication, we can name the 6 clusters of the Lower 48 area in the United States as: Northeastern Area (1 cluster), South Eastern Area (1 cluster), Mid-West Area (1 cluster), Mid-South Area (1 cluster), Western Coast Area (2 clusters). In the development throughout the three decades, the distribution of eastern sections keeps stable, however, there shows some adjustment in western coast.

4.5 Top cities and explanation of clustering

According to the analysis in the city level, the top cities with the most publications are listed for further research.

The title and location of top cities are worth studying because the clustering results indicate that those capacity-intensive clusters are usually formed around top cities, as shown in Figure 10. If these cities can get proved to be the cluster centroids, or closely related to the centroids, resources can be provided for these cities to form hubs in the area around. The K-Means method can also be improved to set these cities as initial centroids.

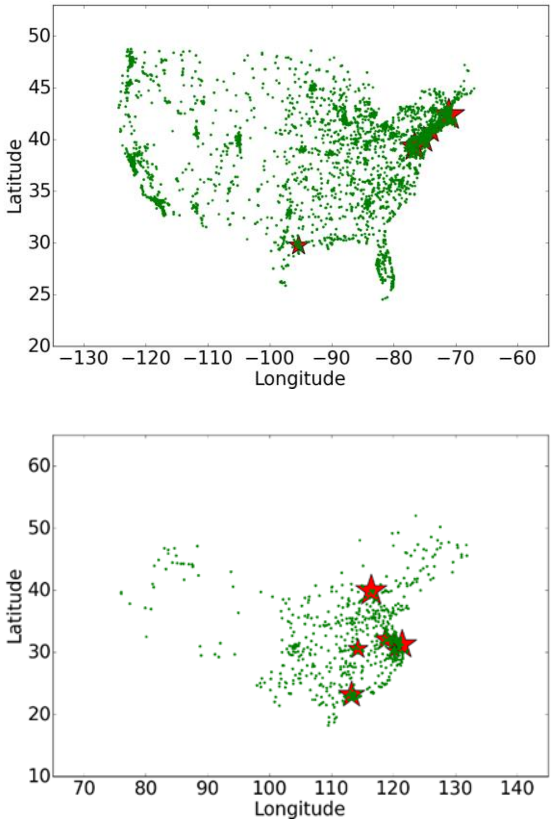


Fig 10. Top cities with most publications in USA and China

Figure 11 displays the comparison between top five cities with the most publication in USA and China from 1988 to 2016. We can observe that with a logarithmic scale, both USA and China have a considerable growth in biomedical literature in the past three decades. However, as shown in Figure 11, the top cities in China have a much larger increase rate than that of USA. The number of publications of top ten cities with most publications in USA and China are listed in Table 1 and Table 2 respectively. In 1988, the average number of publications of top 10 cities in USA is about 64 times of that of top 10 cities in China. Nevertheless, in 2016, the average of top 10 cities in China (36289) is larger than that of USA (34198). The growth rate of China’s top ten cites is 9.64~10.19 in each decade, while the growth rate of USA’s top ten cities is 0.36~1.78. Note that the first two growth rate are calculated within 1988~1998 and 1998~2008, while the last growth rate is calculated from 2008~2016. The growth rates reveal that biomedical literature in China is growing rapidly and its top cities’ publications are expected to be comparable to USA’s top cities in the future.

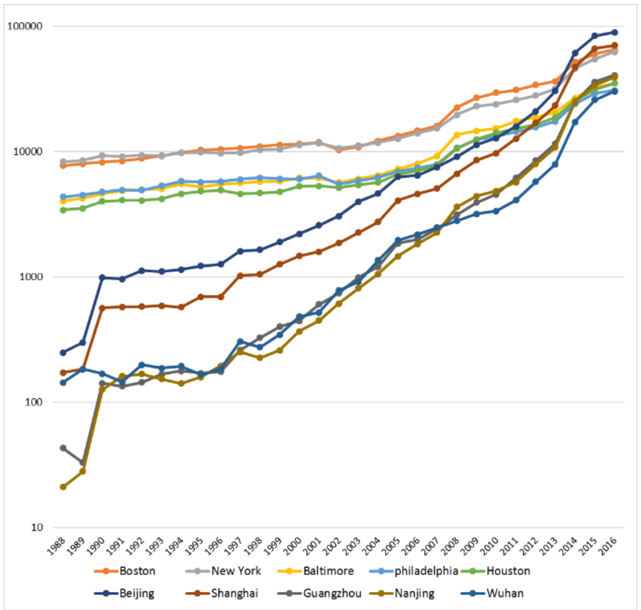


Fig 11. Comparison between top five cities with the most publication in USA and China, from 1988 to 2016

Table 1. Publication numbers and growth rate of ten top cities with most publications in USA

City	1988	1998	2008	2016
Boston	7709	10993 (0.43)	22502 (1.05)	65418 (1.91)
New York	8284	10318 (0.25)	19715 (0.91)	62696 (2.18)
Baltimore	4032	5783 (0.43)	13629 (1.36)	35232 (1.59)
Philadelphia	4366	6216 (0.42)	10708 (0.72)	30878 (1.88)
Houston	3433	4679 (0.36)	10593 (1.26)	35198 (2.37)
Bethesda	4921	5382 (0.09)	12249 (1.28)	21524 (0.76)
Chicago	3893	4800 (0.23)	9348 (0.95)	27466 (1.94)
Seattle	3083	4216 (0.37)	8841 (1.10)	22609 (1.56)
San Francisco	3357	3910 (0.16)	7235 (0.85)	18257 (1.52)
Atlanta	1652	3033 (0.84)	7199 (1.37)	22704 (2.15)
Average	4473	5933 (0.36)	12202 (1.08)	34198 (1.78)

Table 2. Publication numbers and growth rate of ten top cities with most publications in China

City	1988	1998	2008	2016
Beijing	249	1651 (5.63)	9125 (4.53)	89605 (8.82)
Shanghai	172	1049 (5.10)	6657 (5.35)	70393 (9.57)
Guangzhou	43	326 (6.58)	3143 (8.64)	40796 (11.98)
Nanjing	21	226 (9.76)	3620 (15.02)	39367 (9.87)
Wuhan	143	275 (0.92)	2804 (9.02)	30250 (9.79)
Hangzhou	10	111 (10.10)	2890 (25.04)	23350 (7.08)
Xi'an	13	193 (13.85)	1550 (7.03)	18683 (11.05)
Chengdu	9	319 (34.44)	1887 (4.92)	15557 (7.24)
Tianjin	18	125 (5.94)	1385 (10.08)	17635 (11.73)
Chongqing	26	131 (4.04)	1096 (7.37)	17251 (14.74)
Average	70	441 (9.64)	3416 (9.72)	36289 (10.19)

5 Discussion

For the future work, first, rather than just inside the United States and China, the same method will get applied to the data all over the world. For example, Europe and South America can be the next target to get analyzed. Furthermore, a world-wide distribution on research activities will be analyzed. Second, the world-wide movement (both inter-national and intra-national) will get analyzed. Especially at an era of frequent transportation all over the world, it is meaningful to analyze the moving trend of scientific activities. Third, more underlying factors (politics, economics, population factors, etc.) on the movement will get analyzed more as an explanation for the future work.

Acknowledgements

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] P. S. Pagel and J. A. Hudetz, “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, vol. 27, no. 2, pp. 208–212, Apr. 2013.

[2] C. Robert, C. S. Wilson, R. B. Lipton, and C.-D. Arreto, “Growth of Headache Research: A 1983–2014 bibliometric study,” Cephalalgia, p. 0333102416678636, Nov. 2016.

[3] F. Liu and Y. Sun, “A comparison of the spatial distribution of innovative activities in China and the U.S.,” Technological Forecasting and Social Change, vol. 76, no. 6, pp. 797–805, Jul. 2009.

[4] Grossetti, D. Eckert, Y. Gingras, L. Jégou, V. Larivière, and B. Milard, “Cities and the geographical deconcentration of scientific activity: A multilevel analysis of publications (1987–2007),” Urban Studies, vol. 51, no. 10, pp. 2219–2234, Aug. 2014.

[5] M. Grossetti, D. Eckert, Y. Gingras, L. Jégou, V. Larivière, and B. Milard, “Cities and the geographical deconcentration of scientific activity: A multilevel analysis of publications (1987–2007),” Urban Stud., vol. 51, no. 10, pp. 2219–2234, Aug. 2014.

[6] M.-H. Huang, H.-W. Chang, and D.-Z. Chen, “The trend of concentration in scientific research and technological innovation: A reduction of the predominant role of the U.S. in world research & technology,” J. Informetr., vol. 6, no. 4, pp. 457–468, Oct. 2012.

[7] T. Chiarini, V. P. Oliveira, do C. e S. Neto, and F. Chaves, “Spatial distribution of scientific activities: An exploratory analysis of Brazil, 2000–10,” Sci. Public Policy, vol. 41, no. 5, pp. 625–640, Oct. 2014.

[8] L. Leydesdorff and O. Persson, “Mapping the geography of science: Distribution patterns and networks of relations among cities and institutes,” J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 8, pp. 1622–1634, Aug. 2010.

[9] M. P. Devereux, R. Griffith, and H. Simpson, “The geographic distribution of production activity in the UK,” Reg. Sci. Urban Econ., vol. 34, no. 5, pp. 533–564, Sep. 2004.

[10] L. Bornmann and L. Leydesdorff, “Which cities produce more excellent papers than can be expected? A new mapping approach, using Google Maps, based on statistical significance testing,” J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 10, pp. 1954–1962, Oct. 2011.

[11] P. Zhou, B. Thijs, and W. Glänzel, “Regional analysis on Chinese scientific output,” Scientometrics, vol. 81, no. 3, pp. 839–857, Apr. 2009.

[12] R. Garcia, V. Araujo, and S. Mascarini, “The Role of Geographic Proximity for University-Industry Linkages in Brazil: An Emprical Analysis,” Australasian Journal of Regional Studies, vol. 19, no. 3, pp. 433–455, Jun. 2013.

[13] W. Hong and Y.-S. Su, “The effect of institutional proximity in non-local university–industry collaborations: An analysis based on Chinese patent data,” Research Policy, vol. 42, no. 2, pp. 454–464, Mar. 2013.

[14] V. I. Torvik, “MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide,” -Lib Mag., vol. 21, no. 11/12, Nov. 2015.