Geographical Distribution of Biomedical Research in USA and China

Yingjun Guan, Jing Du, Vetle Torvik
School of Information Sciences, University of Illinois at Urbana-Champaign

School of Information Sciences
The iSchool at Illinois

Introduction

❑ **Research Questions** Using disambiguated and geocoded names of 19.4 millions author affiliations in PubMed, we address the following questions:

- What was the geospatial distribution of biomedical research in USA and China, and how has it changed over the past 30 years?
- To what degree are the publications concentrated in a few cities or in some regional hubs? Try analyze the clusters and measure of distances inside the publication clusters in both countries.

❖ ***Note:*** *In this research, USA refers to the lower 48 states and China stands for the mainland China.*

Data

❑ **Large but ambiguous database**
PubMed database offers more than 37million medical publications with affiliation and author information.

Figure 1. Examples of ambiguous address information

Table 1. The ambiguity phenomenon in names of world's largest cities

City	Precision	Recall	% of nation output	% of nation population
London, UK	91.8	91.5	29.2	13.7
New York, NY, USA	68.0	87.6	5.5	2.7
Boston, MA, USA	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	99.1	18.2	1.6
Seoul, Korea	97.1	99.3	48.8	19.8

Our research team exploits MapAffil, which maps author's affiliation to the its accurate city location. The top 20 most common countries are listed in Figure 2.

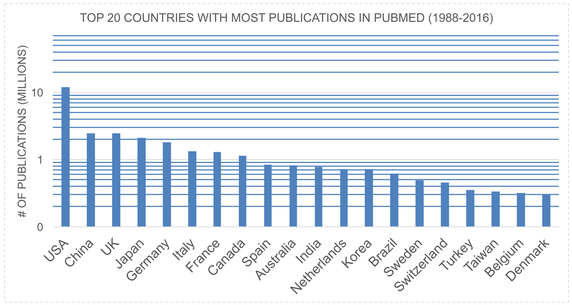


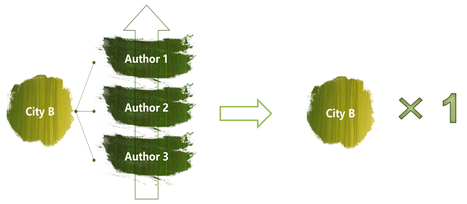
Figure 2. Top 20 countries with most authorships in the database

Method

❑ The spatial distribution of biomedical researches in USA and China, and its change over time.

Locality calculation: Each publication might have multiple affiliations involved, and each affiliation is counted towards its locality (city or suburb). When a publication connects to the same locality multiple times, it will only be counted once.

For example below, in a same paper, the affiliations of all three authors are City B. Then the locality of the publication is City B, whose count is only one.



❑ Centroid and average distance of the affiliations

Geographical centroid: The central point can be calculated by averaging the latitude and longitude of all the relevant localities. In this research both the overall centroid (treat as one cluster) and multiple cluster centroids are calculated and analyzed.

Measure of Distance: The Rooted Mean Squared Vincenty Distance (RMSVD) is taken as the measure of distance in clustering analysis. The average distance from each city to the geographical centroid can be calculated by Equation 1.

$$\bar{r} = \sqrt{\frac{1}{n} * \sum_i r_i^2} \qquad \textit{(Equation 1)}$$

Vincenty Distance: Vincenty Distance is based on the assumption of the Earth being an oblate spheroid. Compared with Euclidean plane and great circle assumptions, the error is smallest and neglect. [However, it should be noticed that the distance here is the theoretical distance rather than travel distance]

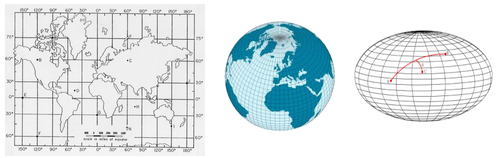


Figure 3. Model of Euclidean plane, great circle and vincenty model

❑ Clustering analysis in USA and China

K-Means clustering: the method aims to partition all observations into k clusters in which each observation belongs to the cluster with the nearest mean. As k increases, the RMSVD decreases. The method can help analyze research distributions.

Results

❑ **Geospatial distribution in USA and China, and overall centroid movement over time (1988 - 2016).**

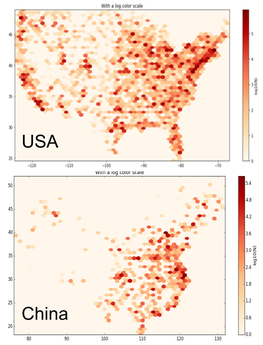


Figure 4. The overall geospatial distribution of USA (the lower 48) and China (mainland) are shown as left.
In the US, eastern and western coasts are the two main hubs of medical publication, the meddle part, like the the state of Nebraska and Utah, has fewer publications. The center of the lower 48 is in Illinois!
In China mainland, the eastern coast enjoy the most medical publications, there are obviously fewer publications in the middle and western sections. The center of the mainland of China is in Anhui Province.

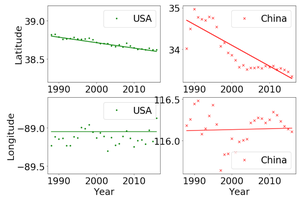


Figure 5. In the time period from 1988 to 2016, the overall centroid (averaged latitude and averaged longitude) in both USA and China moved southward.
For the US, the centroid slightly moved around 0.2 degree (20 miles); and for China, the centroid moved around 1.7 degree (175 miles).

❑ **Analysis of top cities in USA and China**

Top cities with most publications are analyzed in both countries. As shown in the figures and tables below, their behaviors are different: in USA, top 14 cities make up the 40% of all publications, where Boston contributes the most (6.5% of the publication); whereas in China, there are only four. Two super cities, Beijing and Shanghai, both contribute the majority (15.7% and 12.0%).

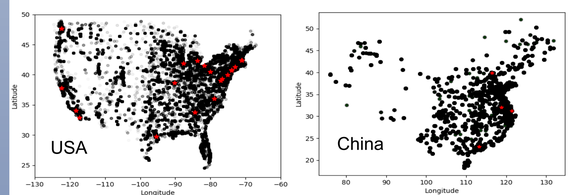


Figure 6. Top cities（contributing up to 40%）in both US and China have been marked above and listed in the table on the right.

Table 3. Information of top cities in both countries (latest data)

Country	ranking	name	publication%	accumulated%	Population%
USA	1	New York	6.5	6.5	2.6
USA	2	Boston	5.6	12.1	0.2
USA	3	Los Angeles	3.2	18.9	1.2
USA	4	Philadelphia	3.1	22.1	0.5
USA	5	Baltimore	3	25	0.2
USA	6	Chicago	2.8	27.8	0.8
USA	7	Houston	2.6	30.4	0.7
USA	8	Bethesda	2.4	32.9	0
USA	9	San Diego	2.3	35.2	0.4
USA	10	Seattle	2.1	37.3	0.2
USA	11	San Francisco	2.1	39.4	0.3
USA	12	St. Louis	1.9	41.3	0.1
USA	13	Durham	1.7	39.4	0.1
USA	14	Atlanta	1.7	41.1	0.1
China	1	Beijing	18.2	18.2	1.6
China	2	Shanghai	13	31.2	1.8
China	3	Guangzhou	6.2	33.8	1
China	4	Nanjing	5.9	39.8	0.7

❑ **Clustering analysis of the dataset**

In this paper, K-Means method is applied in clustering the dataset. By trying different parameters of k value.

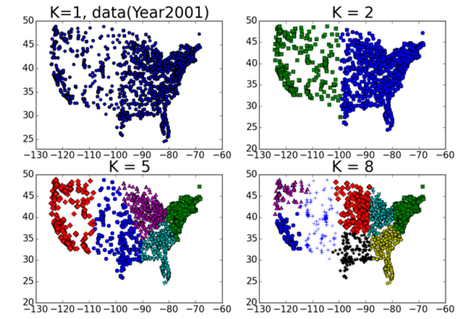


Figure 7. The medical publication distribution in USA by different clusters.

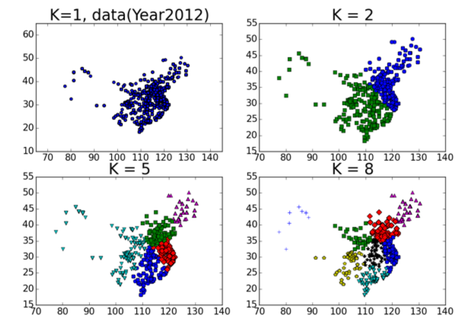


Figure 8. The medical publication distribution in China by different clusters.

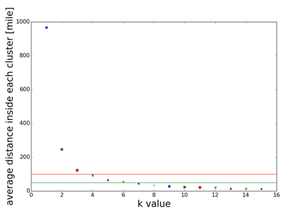


Figure 9. The change of average distance (in miles) versus the change of k values.

As the cluster number increases, the distance (RMSVD) drops. And when the number of clusters reaches 4, the average distance gets 100 miles, according to the USA dataset.

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.

ILLINOIS