Geospatial Distribution of Biomedical Research in USA and China

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

School of Information Sciences
The iSchool at Illinois

Introduction

- Using disambiguated and geocoded place names of author affiliations in PubMed, we address the following questions:
  - What is 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 regions?

Geocoded Affiliation Data

- MapAffil[1] disambiguated and geocoded and place names in 37 million PubMed affiliations.

Table 1. **Sample of low-precision place names**

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

Table 2. **Ambiguity of top most frequent place names**

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
Baltimore, MD, USA	99.5	94.9	97.2	2.7
Philadelphia, PA, USA	99.5	95.1	97.2	2.7
Los Angeles, CA, USA	99.6	86.5	92.6	2.6

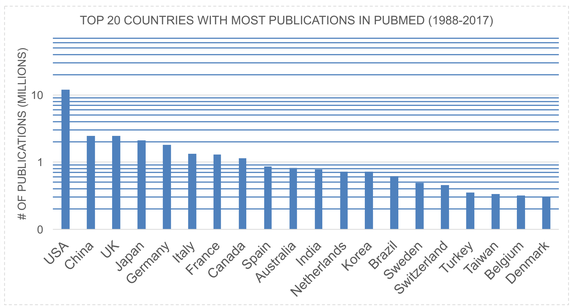

Figure 1. **Top 20 countries with most publications (1988 - 2017)**

Methods and Analysis

- **The geospatial distribution**

  Locality frequency calculation: 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.

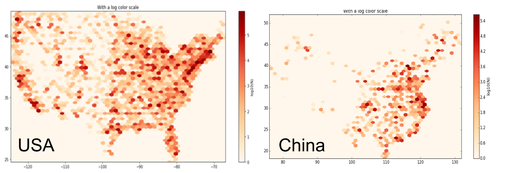

Figure 2. **The overall geospatial distribution of USA (left) and China (right)**

As expected, the geospatial distribution is uneven with several sparse and dense regions. The distribution is likely to mimic population density, which is analyzed through clustering analysis in the following sections.

- Centroid and average distance calculation

Geographical centroid: The centroid of a region or country is calculated by averaging the latitude and longitude of all the relevant localities. Here, both the overall (country-wide) centroid and multiple cluster centroids (regions) are identified and analyzed.

Average distance to a cluster centroid: The Rooted Mean Squared Vincenty Distance (RMSVD) is taken as shown below in Equation 1, where r_i is the Vincenty distance from an individual affiliation to the centroid, n is the total number of localities in each calculation, $\bar{r}$ is RMSVD:

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

Vincenty Distance: Vincenty Distance is the physical distance between two points based on the assumption that the Earth is an oblate spheroid. Note that the Vincenty distance approximates surface distance and differs from travel distance.

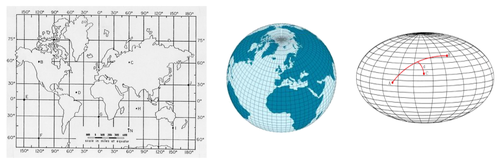

Figure 3. **Model of Euclidean plane, Great Circle and Vincenty model**

Centroid and Clustering Analysis

- **Country-wide centroid movement (1988 - 2016).**
  The overall US and Chinese centroids are located in state of Illinois and the province of Anhui, respectively, and have moved southward over time (Figure 4).

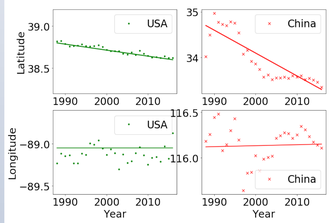

Figure 4. **Overall centroid movement over time.** In the time period from 1988 to 2016, the overall centroid (averaged latitude and averaged longitude) in both USA and China moved southward. For the US, the centroid move 0.2 degree southward (20 miles); and for China, the centroid moved around 1.7 degree (175 miles) southward.

- **K-Means Clustering**

The method aims to partition all publications into k clusters in which each publication belongs to the cluster with the nearest centroid. Clustering helps identify regions with high densities of research activity.

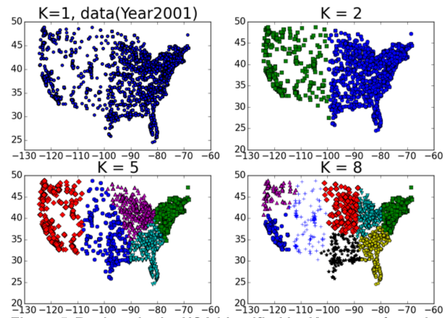

Figure 5. **Regions in the USA identified by K-means clustering.**

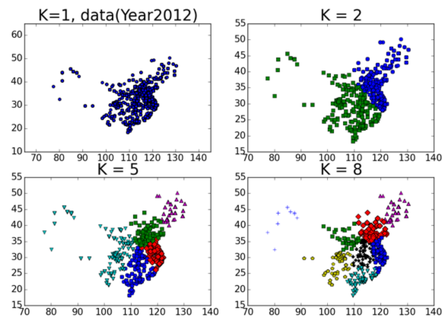

Figure 6. **Regions in China identified by K-means clustering.**

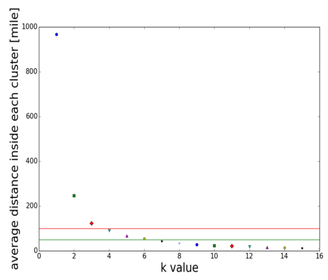

Figure 7. **The average distance (in miles) versus the number of clusters (k) in the USA.** As the cluster number increases, the average distance drops rapidly. When the number of clusters reaches 4, the distance is < 100 miles.

Top Cities in USA and China

In the USA, the top 13 cities produce nearly 40% of all publications, where Boston contributes the most (6.5%). In China, only 4 cities capture nearly 40% and there are two super cities: Beijing (18.2%) and Shanghai (13.0%).

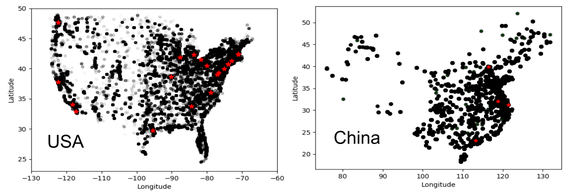

Figure 8. **The 13 top US cities (left) and 4 top Chinese cities（right).**

Table 3. **The top cities in the US and China during 1988 – 2017.**

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
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

References

[1] Torvik VI. MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. D-Lib Magazine 2015; 21(11/12).
[2] Guan, Y., Du, J. & Torvik, V. Geographical Distribution of Biomedical Research in the USA and China. Presented at the 6th International Workshop on Mining Scientific Publications, Toronto, Canada; June 2017.

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