Evaluating demographic representativeness of location-based social media: a case study of Weibo

Yihong Yuan
Department of Geography
Texas State University
601 University Dr
San Marcos, TX, USA, 78666
yuan@txstate.edu

Guixing Wei
Department of Geography
Texas State University
601 University Dr
San Marcos, TX, USA, 78666
g_w38@txstate.edu

Abstract

Researchers have utilized location-based social media (LBSM) as potential resources to characterize daily mobility patterns and social perceptions of place. However, similar to other types of big data, LBSM data also have differential data quality issues such as accuracy, precision, temporal resolution, and sampling biases across various population groups. This research takes an initial step to examine the sampling biases of a Chinese microblogging site, Weibo (a Chinese social-networking website functionally similar to Twitter), as well as investigating the bias in genders and how this bias varies/auto-correlates in different provinces of China. The results indicate that in general, women are much more active on LBSM than men in China. We also detected a strong regional pattern for LBSM gender ratios. The results provide valuable input in quantifying demographic biases in LBSM. It also offers a data pre-processing strategy to identify potential research questions in social studies.

Keywords: Location-based social media (LBSM), Sampling biases, gender differences, spatial autocorrelation, big (geo)data.

1 Introduction

Researchers have defined location-based social media as "social network sites that include location information" [1]. Unlike traditional travel surveys or actively collected Global Positioning System (GPS) logs, location-based social media (LBSM) datasets often cover a large sample size and can easily be accessed through crowd-sourcing toolkits, and therefore can be utilized as potential resources to characterize daily mobility patterns and social perceptions of place [2, 3]. However, similar to other types of big data, LBSM data also have different data quality issues such as accuracy, precision, temporal resolution, and sampling biases across various population groups. However, the demographic bias of LBSM data and its influence on the quality of derived mobility patterns have not been thoroughly studied. Questions like, "Do we have unrepresentative demographic groups in underdeveloped areas?" are essential for assessing LBSM data quality and soundness of experimental design.

This research takes an initial step to examine the sampling biases of a Chinese microblogging site, Weibo (a Chinese social-networking website functionally similar to Twitter). Based on previous statistics, there is no doubt that all LBSM sites generate a biased user group, e.g., more than 50% of Twitter users are in the age group 16-34 [4]. However, we want to take one step further by addressing how this is distributed spatially.

2 Analysis and Preliminary Results

The dataset utilized in this research covers approximately 1 million Weibo users who checked-in their locations at least once between 03/2015 and 09/2015. As discussed in Section 1, this research examines the demographic biases of LBSM users and the spatial autocorrelation of sampling biases in LBSM users to reveal any geographic bias in different provinces.

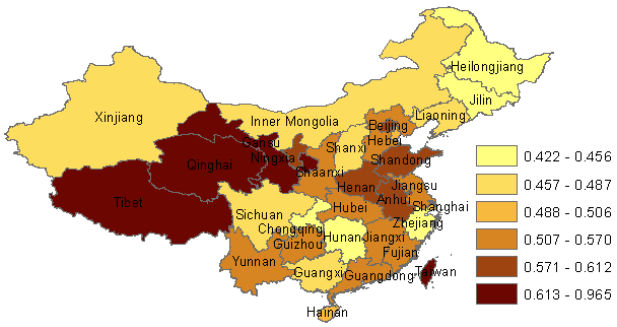
As an exploratory analysis, Table 1 and Figure 1 illustrate the ratio of male to female users in Chinese provinces, provincial-level cities, special administrative units (e.g., Hong Kong and Macaw), and users locate outside of China ("overseas").

Table 1: Male to female ratio (M:F ratio) from Weibo data and census data.

	M:F (Weibo check-in)	**M:F (Census)[5]**
Beijing	0.605:1	1.068:1

Tianjin	0.499:1	1.145:1
Hebei	0.543:1	1.028:1
Shanxi	0.475:1	1.056:1
Inner Mongolia	0.468:1	1.081:1
Liaoning	0.47:1	1.025:1
Jilin	0.456:1	1.027:1
Heilongjiang	0.44:1	1.032:1
Shanghai	0.506:1	1.062:1
Jiangsu	0.532:1	1.015:1
Zhejiang	0.45:1	1.057:1
Anhui	0.585:1	1.034:1
Fujian	0.562:1	1.06:1
Jiangxi	0.539:1	1.075:1
Shandong	0.581:1	1.023:1
Henan	0.594:1	1.021:1
Hubei	0.57:1	1.056:1
Hunan	0.43:1	1.058:1
Guangdong	0.546:1	1.09:1
Guangxi	0.487:1	1.083:1
Hainan	0.501:1	1.109:1
Chongqing	0.422:1	1.024:1
Sichuan	0.473:1	1.031:1
Guizhou	0.56:1	1.069:1
Yunnan	0.541:1	1.078:1
Tibet	0.757:1	1.057:1
Shaanxi	0.567:1	1.069:1
Gansu	0.779:1	1.044:1
Qinghai	0.745:1	1.074:1
Ningxia	0.612:1	1.051:1
Xinjiang	0.486:1	1.053:1
Taiwan	0.965:1	0.998:1
Hong Kong	0.706:1	1.070:1
Macaw	1.753:1	0.946:1
Overseas	0.818:1	N/A

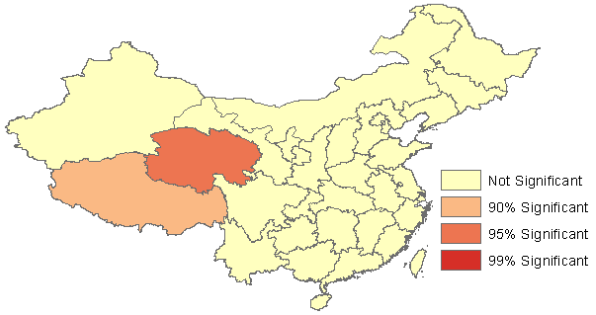
Figure 1: M:F ratio among Weibo users by province



As can be seen, the majority of provinces exhibit a M:F ratio lower than 0.65: 1 (the average M:F ratio of the whole dataset is 0.54:1), indicating that there are almost twice as many female users posting their locations as male users. However, the official census data indicates an opposite trend, where most study areas in China have more male than female population (M:F ratio >1:1). The correlation between M:F ratios from Weibo data and census statistics is not significant. Additionally, it is also distinct from the pattern exhibited by other LBSM sites. Foursquare, for instance, reports more male users than female users (60% male vs 40% female) [6]. On the other hand, Sina Corp (the parent company of Weibo) publishes yearly official statistics of Weibo users. The 2014-2015 statistics indicate that the M:F ratio of active users decreased from 1.56:1 to 1:1 [7], indicating a trend of increasing female active users on the site. However, a more detailed report from 2011 also indicates that, among all active users, the percentage of female users utilizing the location-based service (LBS) features (e.g., location check-in) is substantially higher than male users (M:F ratio = 0.71:1). Our case study further confirmed that the M:F ratio continued to decrease to 0.54:1 in 2015, which indicates more and more females are actively involved in LBSM usage in China.

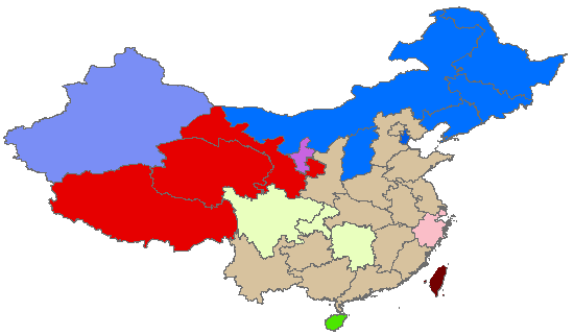
Figure 1 also demonstrates potential spatial autocorrelation among Chinese provinces. For example, the western provinces (Tibet, Qinghai, and Gansu) show a cluster of high M:F ratio, which indicates that female LBSM users in these provinces are less active. Figure 2 shows the results of a Getis-Ord General G analysis confirming this hotspot in northwestern China.

Figure 2: Getis-Ord General G (Hotspot) analysis.



We also conducted a spatial-constrained K-means clustering analysis to explore regional patterns (Figure 3, provinces in the same color indicate the same cluster). The number of clusters is determined based on the Pseudo F-statistics[8]. Here we adopt 10 clusters for the analysis.

Figure 3: Clustering analysis of Chinese provinces.



As can be seen, the M:F ratio of Chinese provinces exhibits a clear regional pattern, which can provide useful input for social studies (especially for cultural geography and gender issues in China). As demonstrated by previous studies, the societal status of women can be demonstrated by or correlated to multiple factors, including M:F ratio at birth, the employment rate of women, the industrial structure of the region, cultural influence, etc. [9]. The patterns demonstrated

in this study can help to propose hypotheses in social studies, for example:

- The three western provinces (Tibet, Qinghai, and Gansu) form a cluster of high M:F ratio. These places are generally perceived as "under-developed" areas in China with low gross domestic product (GDP) values, which potentially impacts women's openness in new media such as LBSM.
- Northeast China (Heilongjiang, Jilin, and Liaoning) as well as Inner Mongolia and Shanxi form a cluster of low M:F ratio. In these provinces, the traditional Chinese birth preference for sons is weaker, and there is a lower M:F birth rate, indicating a possible correlation between the social status of women and their usage of LBSM.
- Three provinces/provincial-level regions in Southwest China (Sichuan, Chongqing, and Hunan) form another low cluster. A national survey indicates that the employment rate of females is among the highest of all provinces in China [5].
- The two special administrative units (Hong Kong and Macaw) and Taiwan demonstrate very different behavior from mainland China, with higher M:F ratios in general (Hong Kong: 0.70641; Macaw: 1.752874; Taiwan: 0.965445). Macaw is the only study area that has more male than female LBSM users. The behavior differences and the openness of women to LBSM in Hong Kong, Macaw, and Taiwan on the one hand, and mainland China on the other, may relate to the different social regulations in these places.

However, these hypotheses have to be further tested and verified in social studies, which is beyond the scope of this research. For the researchers in the LBSM field, this study is valuable for demonstrating the potential data quality issue and demographic biases in such datasets, which is crucial for designing a sound experiment and/or exploring geo-temporal factors causing these biases.

3 Conclusion

This research investigated the spatial autocorrelation of gender biases in Weibo data. The results indicate that in general, female users are more active on LBSM than male users in China. We also detected a strong regional pattern for LBSM gender ratios. The results provide valuable input for quantifying demographic biases in LBSM. It also offers a data pre-processing strategy to identify potential research questions in social studies. The methods and models can be applied to other LBSM datasets (e.g., Twitter or Foursquare) to test its robustness. We will further extend this analysis to other demographic factors such as age and education level. Future research should look into patterns at different spatial scales (such as investigating the variations inside different neighborhoods of an urban system) to address potential modifiable areal unit problem (MAUP).

References

[1] O. Roick and S. Heuser, "Location Based Social Networks - Definition, Current State of the Art and Research Agenda," *Transactions in Gis*, vol. 17, pp. 763-784, Oct 2013.

[2] N. Malleson and M. Birkin, "New Insights into Individual Activity Spaces using Crowd-Sourced Big Data," presented at the ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford, CA, 2014.

[3] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu, "Maximizing benefits from crowdsourced data," *Computational and Mathematical Organization Theory*, vol. 18, pp. 257-279, 2012/09/01 2012.

[4] Business Insider, "This Chart Reveals The Age Distribution At Every Major Social Network," ed, 2014.

[5] National Bureau of Statistics of China, "2010 Population Census of China," ed, 2010.

[6] Brand on Gaille. (2015). *26 Great Foursquare Demographics*. Available: http://brandongaille.com/26-great-foursquare-demographics/

[7] Sina Corp. *Weibo user background report*. Available: http://data.weibo.com/report/

[8] I. ESRI. (2015). *ArcGIS Pro - Grouping analysis*. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/grouping-analysis.htm

[9] L. P. Edwards and M. Roces, *Women in Asia : critical concepts in Asian studies*. Milton Park, Abingdon, Oxon ; New York: Routledge, 2009.