ANNALS OF GIS, 2018
https://doi.org/10.1080/19475683.2018.1471518

Taylor & Francis
Taylor & Francis Group

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

Yihong Yuan, Guixing Wei and Yongmei Lu

Department of Geography, Texas State University, San Marcos, TX, USA

ABSTRACT

Researchers have utilized location-based social media (LBSM) as potential resources to characterize daily mobility patterns and social perceptions of place. 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. However, these issues have not been investigated sufficiently for LBSM users. This research aims to quantitatively examine the sampling biases of a Chinese microblogging site, Weibo, which is functionally similar to Twitter. The analysis focuses on investigating the bias in gender groups, and how this bias varies/autocorrelates in different provinces of China. The results indicate that in general, women are more likely to use Weibo in China. We also detected a strong regional pattern for Weibo gender ratios. The results provide valuable input in quantifying demographic biases in Weibo, and the methodology can be applied to other LBSM to analyse sample biases. This study also offers a data preprocessing strategy to identify potential research questions in sociology, regional science, and gender studies.

ARTICLE HISTORY
Received 24 November 2017
Accepted 27 April 2018

KEYWORDS
Location-based social media (LBSM); sampling bias; gender studies; spatial autocorrelation; big (geo) data

1. Introduction

In recent decades, the development of Information and Communication Technologies (ICTs, such as the Internet and Web 2.0) has provided more flexibility and capability regarding where, when, and how people connect to each other (De Souza e Silva 2007; Chow, Lin, and Chan 2011). A series of Social Network Sites (SNS), such as Facebook and Twitter, have allowed worldwide users to communicate, socialize, and share their daily lives. Meanwhile, the widespread use of smartphones, which are equipped with sensors that allow users to instantly locate themselves, has brought another crucial aspect to this development: location. Researchers have defined location-based social media (LBSM) as 'Social Network Sites that include location information' (Roick and Heuser 2013).

Unlike traditional travel surveys or actively collected Global Positioning System (GPS) logs (Scholtz and Lu 2014), LBSM data sets often cover a large sample size and can easily be accessed through application programming interfaces (APIs) in standard formats, and therefore can be utilized as potential resources to characterize daily mobility patterns and social perceptions of place (Malleson and Birkin 2014; Barbier et al. 2012). The ability to accurately process such massive data sets brings new challenges to the big data era. Compared to geo-referenced mobile phone data such as Call Detailed Records (CDRs), LBSM data often have more reliable spatial accuracy (5–10 m from built-in GPS devices versus 100–1000 m from cellular towers). Another advantage of LBSM data is the potential to extract subjects' background information (e.g. demographic information like age and gender), which tends to be extremely difficult to obtain from cell phone carriers due to privacy issues (Calabrese, Ferrari, and Blondel 2015). Therefore, crowd-sourced LBSM tends to provide faster and more detailed contextual data than traditional sources (Barbier et al. 2012).

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. Researchers in sociology and public relations have addressed the need to validate social media data for both personal usage (e.g. information subscription) and authority usage (e.g. emergency planning) (Zamri, Darson, and Wahab 2014; Westerman, Spence, and Van der Heide 2014; Poorthuis and Zook 2017). Previous studies also focused on inferring LBSM demographic attributes from a text-

CONTACT Yihong Yuan yuan@txstate.edu

mining perspective (Zhong et al. 2015; Sloan et al. 2015). However, the demographic bias of LBSM data and its influence on the quality of derived mobility patterns have not been thoroughly studied, which inevitably affect the reliability of mobility studies based on these data sets (Burger et al. 2011). For instance, each social media platform has certain characteristics and affordances – things that it allows and makes easy versus things that are difficult to accomplish (Rutherford et al. 2013; Golub and Jackson 2010). This helps shape behaviour as well as the user group (e.g. race, gender, age) on social media sites (Tufekci 2014). For example, more than 50% of Twitter users are between the ages of 16 and 34 years (Business Insider 2014). Instagram, on the other hand, particularly attracts adults between the ages of 18 and 29, women, and urban dwellers (Forbes 2014). However, there is very limited research on investigating how the sampling bias in demographic groups varies or autocorrelates in space.

Realizing the necessity to evaluate the sampling biases and representativeness of LBSM in modelling human mobility (Cho, Myers, and Leskovec 2011; Hasan, Zhan, and Ukkusuri 2013), this research takes an initial step by examining the gender biases of a Chinese microblogging site, Weibo, which is one of the most popular Chinese social-networking websites and functionally similar to Twitter. Even though it is common knowledge that user profiles in any social media platform are inevitably biased, this study does not stop at confirming the fact that 'almost all LBSM sites have a biased user group.' Instead, we want to go one step further by addressing how this data representativeness issue is distributed spatially. That is, the demographic biases of LBSM users naturally manifest into a source of geographic biases in LBSM, which originates from the tendency of geographic features relating to each other in space, violating the assumption of independent observations required in classical statistics (Griffith 2003).

As exploratory research, this study focuses on one of the most basic demographic factors: gender. In addition to the basic question, '*Is there an unbalanced gender representativeness in Weibo (especially for users posting their locations)?*,' we also aim to investigate two follow-up research questions: 1) '*Does this under-representation/over-representation show a spatially autocorrelated pattern following the first law of geography?*' and 2) '*Does this under-representation/over-representation show a statistically significant correlation with basic social economic factors that can be used to explain gender inequality?*' These questions are essential for assessing LBSM data quality and soundness of experimental design. Hence, this paper contributes from the following two perspectives: 1) empirically, we examine the variation of LBSM gender biases in Chinese provinces and how these variations autocorrelate spatially, and 2) methodologically, we demonstrate how these autocorrelated patterns can be utilized as a powerful data-mining tool for crucial social topics, such as gender inequality and segregation (Yuan and Wei 2016).

This paper is organized as follows. Section 2 synthesizes related studies in the areas of LBSM and data quality. Section 3 illustrates the fundamental research design, including the Weibo data set and the methodology. Section 4 presents the data analyses and results and discusses various aspects of the output in detail. We conclude this research and present directions for future work in Section 5.

2. Related work

2.1. Analysing user behaviours from LBSM

The continued development of social networking websites (SNS) such as Twitter and Facebook provides ever-increasing opportunities to explore activity patterns of individuals within diverse geographic environments, social statuses, and cultural backgrounds (Wu et al. 2014; Liu et al. 2014b; Noulas, Mascolo, and Frias-Martinez 2013). LBSM has attracted users worldwide and allowed them to share their whereabouts at daily, weekly, and long-term temporal scales, making LBSM particularly suitable for modelling individual activity patterns such as activity scheduling, social network structure, location prediction, etc. These technologies also lead to a fundamental change in how citizens may contribute crowdsourcing data in the decision-making process of urban planning (Crampton et al. 2013; Elwood 2006).

Despite potential data-quality issues, previous studies have demonstrated the effectiveness of LBSM data in analysing activity behaviours and constructing mobility models (Musolesi, Hailes, and Mascolo 2004; Cho, Myers, and Leskovec 2011). For instance, Gao, Tang, and Liu (2012) investigated the role of social correlation in users' check-in behaviour to improve the accuracy of location prediction. Hasan, Zhan, and Ukkusuri (2013) analysed the timing distribution of visiting different places depending on the activity category of individual users. Researchers also investigated how LBSM data sets can be utilized to parameterize the traditional mobility models under demographic controlling factors (Gao and Liu 2015; Noulas et al. 2012).

In addition to the analysis of individual activity patterns and space, LBSM data provide a great opportunity

for investigating how human mobility patterns are shaped by urban environments from an aggregated perspective and how the latter (urban-oriented studies) may be managed or designed to better suit the needs of the former (individual-oriented studies) (Bawa-Cavia 2011; Cranshaw et al. 2012). As noted by Roick and Heuser (2013), the continuously shared location information through LBSM services can be used to analyse urban structures, clusters, and dynamics (Lu 2000). The concept of social sensing – using social media as a data source to study cities and societies – has provided useful information for urban planners and policymakers (Liu et al. 2015).

Because Twitter, Facebook, and several other SNS are not directly accessible in mainland China, previous studies have used Weibo as a primary data source to analyse LBSM user activity patterns and provide useful input for urban planners and policymakers in China (Zhang et al. 2016; Zhang et al. 2012; Liu, Dong, and Gu 2014a). For instance, Liu, Dong, and Gu (2014a) demonstrated how to use geotagged Weibo posts as data to analyse the distribution of air pollution topics in China. Guan et al. (2014) analysed user behaviour patterns on Weibo during hot events. However, many of these studies did not address the potential sampling biases and data-quality issues of Weibo, which is crucial for evaluating the results of quantitative studies based on such data sets.

2.2. Data-quality issues in LBSM

In spite of the widespread use of LBSM as a major source of big (geo) data, there have been drawbacks in quantifying its quality issues and justifying the usage of such data in certain applications (Elwood and Leszczynski 2011; Harvey 2013; Kitchin 2013). As discussed in Goodchild (2013), big (geo)data are often assembled from various data sources that lack consistent quality control, which inevitably brings extra challenges to analysing such data. To help better understand big data quality, researchers have proposed four big V's to determine the characteristics of big data: Volume (massive amount of data available), Velocity (how fast data is being generated), Variety (big data is a collection of data sets that are in different formats), and Veracity (the degree of accuracy and uncertainty of data) (IBM 2015). These uncertainty issues may come from various sources, from the data set to be mined, the process of mining such data, or applying uncertain knowledge to new data sets (Xia 2005; Mislove et al. 2012).

The gender representativeness issue addressed in this study can be considered a subset of the 'Veracity' issue of the four big 'V's, which addresses the incompleteness and bias that originate from the data set itself. As discussed in Section 1, there is insufficient study on how basic geographic laws, such as Tobler's First Law of Geography (Tobler 2004; Tobler 1979), may influence how these biases manifest spatially; hence, it is crucial to assess the reliability of LBSM data for mobility analysis (Spielman 2014; Veregin 1999), including but not limited to the following.

- Data quantity and resolution: Limited location sampling resolution is an inevitable issue in LBSM (users may check-in once per day or even less, and check-in data are generated at a different speed for various user groups). However, in practice, the appropriate data size and sampling resolution are often determined arbitrarily when using LBSM to analyse activity patterns. There has yet to be a systematic study on how a user's activity space changes upon collected sample size and temporal resolution of LBSM. Although in general, larger sample sizes can provide more location information for a certain user, researchers often seek a 'reasonable' sampling size and resolution, which achieves a balance between the details of information and computation efficiency/collection cost. In addition, researchers in computer science and engineering have also focused on the real-time processing of social media data, such as semantic labelling of fast-generated social media posts in a real-time flow (Trinh Minh Tri and Gatica-Perez 2014).
- Data completeness, sampling bias, and population representativeness: Obviously, users of LBSM are not a randomly selected population (Golub and Jackson 2010; Rutherford et al. 2013; Crooks et al. 2013). Pinterest, for instance, particularly attracts women between the ages of 25 and 34 with average household incomes of $100,000 (Carnegie Mellon University 2014). Researchers have addressed the representativeness issue of LBSM in recent studies, which is related to the concept of 'racialized cyberspace' in cultural geography and political science. Previous studies investigated whether the large spread of new media reinforced or mitigated racial and ethnic stereotypes (Fekete 2015; Zook and Graham 2007). For example, Zickuhr (2013) found that age, gender, and race have substantial impact on people's usage of social media, where young people, women, and minorities show a greater percentage of usage. However, her results indicated that gender did not affect the rate of whether an account is

location-enabled. These studies naturally raised concerns that LBSM data sets may not be accurate, objective, or representative of the entire population (Kitchin 2014). Studies by Leszczynski and Crampton (2016) and Hawelka et al. (2014) also discussed the validity of utilizing geotagged tweets from a selective and self-selecting population. Other biases include the 'tyranny of the loud,' where a small but active group of users generates a large amount of records, which distorts the representativeness of the data set (Rzeszewski 2018). As discussed in Section 2.1, current studies on LBSM spatio-temporal patterns mainly focus on analysing user preferences (i.e. where and when geotagged posts are more likely to be generated) (Hasan, Zhan, and Ukkusuri 2013), as well as their implications for urban and transportation studies (Cho, Myers, and Leskovec 2011; Bawa-Cavia 2011; Roick and Heuser 2013; Zhang et al. 2016; Liben-Nowell et al. 2005; Rzeszewski 2018). Many of these studies, however, do not focus on identifying how user demographic biases distribute spatially, which is crucial for understanding the influence of LBSM data biases on the quality of derived human mobility.

Even though the complexity of the geography world makes it virtually impossible for researchers to draw a random sample, and most geography experiments and surveys are 'natural' and uncontrolled (Goodchild 2013), it is still beneficial for researchers to understand how biased these limited samples are. Questions like 'Are females more likely to use LBSM in more urbanized areas?' and 'Do industrialized large cities have a less biased sample on LBSM than smaller cities?' can help researchers evaluate the soundness of their research design when utilizing such data sets (Longley, Adnan, and Lansley 2015).

- Data accuracy and consistency: Studies have concentrated on identifying spam and untrustworthy posts on social media sites (DeBarr and Wechsler 2010; Saini 2014; Guo and Chen 2014). However, there have been very limited solutions for detecting fake/suspicious location check-ins. In certain cases, it is even challenging for researchers to identify whether a social media post is from a human being or is automatically generated by an algorithm (i.e. 'social bots'). This brings extra challenge to human mobility studies (Crampton et al. 2013; Tsou et al. 2015). Another issue to take into consideration is the accuracy of mobile GPS. Applications like Foursquare allow users to check-in to receive points and/or rewards when they are within the vicinity of a certain location. Another category of accuracy checking applies to aggregated LBSM data in urban studies (i.e. cross-validating urban clusters and estimation accuracy with other data sources (e.g. census)).

Although this study focuses on exploring gender bias, it is worth noting that gender studies related to LBSM are not limited to imbalanced gender representation or data-quality issues. Several studies also analysed other gender-specific behaviours on social media and the implications for online communications, such as how gender differences relate to different language styles and vocabularies (Schwartz et al. 2013; Ye et al. 2018). For example, Ye et al. (2018) concluded that females are more likely to use emotional and positive hashtags while posting photos on Instagram. These gender-specific language styles also provide quantified evidence to predict gender information for users with incomplete profiles (Burger et al. 2011; Argamon et al. 2003; Longley and Adnan 2016; Mislove et al. 2012). Besides language preferences, researchers also studied the social network structures of men and women and identified gender differences in constructing new connections and maintaining existing connections on social media (Marwick 2013; Mazman and Usluel 2011; van Oosten, Vandenbosch, and Peter 2017). Mazman and Usluel (2011) concluded that even though men are more likely to make new connections, women are actually more inclined to maintain existing relationships using social media. In Volkovich et al. (2014), the authors found a tendency of gender segregation on social media, where people with the same gender are much more likely to connect on social media; however, users with a large social circle tend to make more connections with users of the opposite gender.

As discussed in Section 1, this research aims to address LBSM data quality from completeness and representativeness perspectives. We aim to analyse the spatial distribution of gender bias in Chinese provinces and how this bias may relate to certain socio-economic factors, such as the male-to-female sex ratio at birth (SRB).

3. Research design

3.1. Data set

Weibo was launched in 2009 by its parent company, Sina Corporation, and soon became one of the most influential and popular microblogging/social networking sites in China. By 2015, Weibo already had 222 million subscribers and 100 million daily users, and this number continues to grow rapidly (Sina Corp

2013–2017). Because of its widespread usage and popularity among the public in China, we chose Weibo as our analysis test bed. The data set used in this research was acquired from the official Weibo APIs in JavaScript Object Notation (JSON) format. Weibo users can choose to publish their age, gender, education, employment, and more detailed background information in their public profiles. Previous data collections show around 20%–30% of Weibo users have their demographic information publicly available.[1] We retrieved this information directly from Weibo user profiles through Weibo APIs. The location of Weibo users is acquired through built-in GPS modules of smartphones. Originally, we collected data from over 4.3 million users who checked-in their locations at least once between March 2015 and November 2015. The subset utilized in this research covers around 0.24 million Weibo users who reported their date of birth, gender, and current residential city, which is approximately 5.67% of the over 4.3 million users whose data we collected. This number is lower than the aforementioned 20%–30%, as many users with a public profile may have incomplete information (e.g. reported gender information, but no current residential city, or vice versa). Within the selected sample set, the percentage of male and female users is 33.37% male to 66.63% female. It is important to note that we use reported residential city information from user profiles instead of the coordinates of individual posts, because this study focuses on user background instead of on individual posts.

Note that Weibo also allows an individual or an entity (e.g. organizations, companies, governmental agencies, etc.) to register for a verified account, where the creator needs to submit government-issued documents to verify their information. An organizational account can also choose a gender during the registration process. In our sample set, only 2.8% of the accounts are verified, with a mix of individual (e.g. celebrities) and organizational accounts. Because Sina Corp does not currently provide a good method to differentiate between individual and organizational accounts, and semantic analysis is not the focus of this study, we did not eliminate the small percentage of organizational accounts.

Table 1 shows a few sample user profile records. (Only data fields related to this research are displayed.)

Besides Weibo data, this research also utilizes province-level census data acquired from the National Bureau of Statistics of China as background information used to verify population demographics, as well as basic socio-economic data in different provinces (National Bureau of Statistics of China 2010). National census data is collected every 10 years in China, and the statistics used in this study are from the most recent data collection campaign in 2010.

Table 1. Example data records.

User ID	Gender	Date of Birth	Province	City
*3453******	Female	1979–01-12	Beijing	Beijing
*2185******	Male	1990–10-15	Shanxi	Taiyuan

3.2. Methodology

As mentioned in Section 1, many existing studies have utilized LBSM data to study human mobility, but few have addressed how biases of such data or the populations they represent distribute spatially, as well as the potential data quality issues they may reveal. This study examines the sampling bias – the sample representing specific population groups in terms of gender, age, and geography, using the following three steps.

- Step 1. Preprocess data

The 0.24 million Weibo users were grouped based on the 'current province/city' information of their profiles. We calculated the percentage of users by gender (male:female) groups. As defined by the American Psychological Association (APA), 'gender refers to the attitudes, feelings, and behaviors that a given culture associates with a person's biological sex.' (APA 2015, 3). A person's gender identity – the inherent sense of being a male, female, or an alternative gender (e.g. genderqueer, gender non-conforming, boygirl, ladyboi) – may or may not be consistent with a person's sex assigned at birth (APA 2015). So, although the concept of gender goes beyond being a binary variable, Weibo's user profiles only have two options for the gender field, 'female' or 'male.' Based on that Weibo limitation, gender in this study is considered a binary variable.

- Step 2. Compare Weibo gender data with census data

To better illustrate the under-/over-representation of demographic groups in Weibo, we define a normalized M:F ratio $(M{:}F)_N$ as follows:

$$(M:F)_N = \frac{(M:F)_W}{(M:F)_C} \tag{1}$$

where $(M{:}F)_W$ and $(M{:}F)_C$ indicate the male-to-female ratios in the Weibo data and census data, respectively. $(M{:}F)_N < 1$ shows an over-representation of females in Weibo. In the scenario where $(M{:}F)_N > 1$, female users are under-represented in Weibo data. The smaller the $(M{:}F)_N$ is, the better the females are represented in Weibo data.

- Step 3. Analyse spatial autocorrelation and regional patterns

The $(M:F)_N$ index defined in step 2 is used to analyse the regional pattern of under-/over-representation in demographic groups on Weibo. We use classic spatial autocorrelation analyses to explore the spatial patterns of demographic representation in Chinese provinces. Specifically, we use the Getis-Ord General G method for the hotspot analysis (Ripley 2004) and the grouping analysis method for the detection of natural groups considering spatial constraints, which is a process to cluster regions by applying a connectivity graph (minimum spanning tree) to find natural groupings (Esri 2015). Although researchers have proposed various methods to quantify spatial autocorrelation (Griffith 1988; Gelfand 2010), we chose the Getis-Ord General G analysis because of its ability to differentiate between the clusters of high values and low values. The grouping analysis aims to explore finer regional patterns by clustering the provinces based on their $(M:F)_N$ and spatial adjacency.

4. Analysis and results

Table 2 and Figure 1 illustrate the normalized ratio of male to female users in Chinese provinces, provincial-level cities, special administrative units (e.g. Hong Kong and Macau), and users outside of China ('overseas'), as well as the $(M:F)_N$. Because gender is considered as a binary variable in this study, the analysis and interpretation in this section focus on female users only.

As shown in Table 2 and Figure 1, the majority of provinces exhibit an M:F ratio lower than 0.65, indicating that there are almost twice as many female users posting their locations as male users. It is noted that the average $(M:F)_W$ of the whole data set is 0.5. In contrast, the official census data indicates an opposite trend, where most provinces in China have a slightly higher male than female population ($(M:F)_C > 1$). The

Table 2. Male to female ratio (M:F ratio) from Weibo data and census data.

	M:F (Weibo)	M:F (Census)		M:F (Weibo)	M:F (Census)
Beijing	0.605	1.068	Guangdong	0.546	1.09
Tianjin	0.499	1.145	Guangxi	0.487	1.083
Hebei	0.543	1.028	Hainan	0.501	1.109
Shanxi	0.475	1.056	Chongqing	0.422	1.024
Inner Mongolia	0.468	1.081	Sichuan	0.473	1.031
Liaoning	0.47	1.025	Guizhou	0.56	1.069
Jilin	0.456	1.027	Yunnan	0.541	1.078
Heilongjiang	0.44	1.032	Tibet	0.757	1.057
Shanghai	0.506	1.062	Shaanxi	0.567	1.069
Jiangsu	0.532	1.015	Gansu	0.779	1.044
Zhejiang	0.45	1.057	Qinghai	0.745	1.074
Anhui	0.585	1.034	Ningxia	0.612	1.051
Fujian	0.562	1.06	Xinjiang	0.486	1.053
Jiangxi	0.539	1.075	Taiwan	0.965	0.998
Shandong	0.581	1.023	Hong Kong	0.706	1.070
Henan	0.594	1.021	Macau	1.753	0.946
Hubei	0.57	1.056	Overseas	0.818	N/A
Hunan	0.43	1.058			

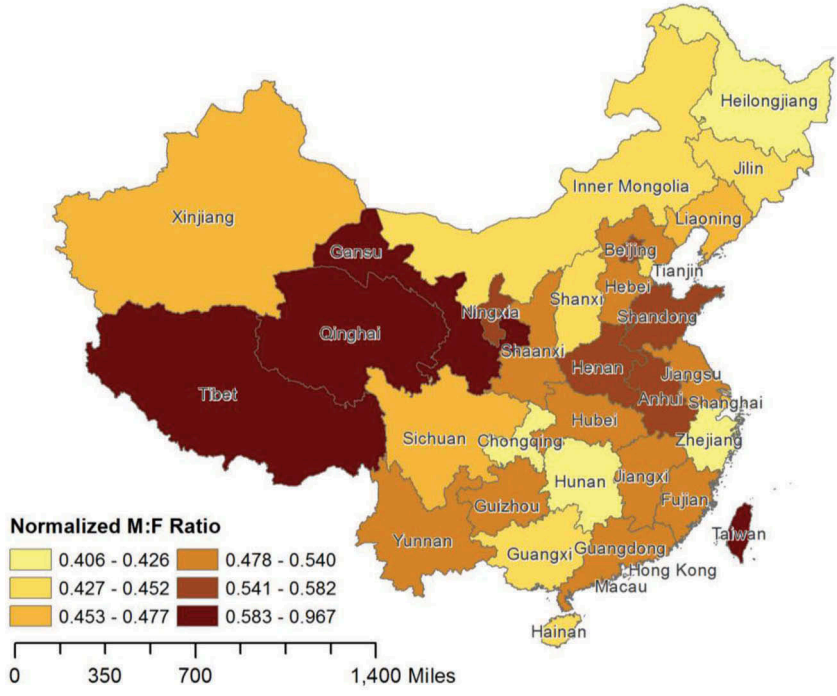

Figure 1. $(M:F)_N$ (normalized M:F ratio) by province.

correlation of M:F ratios between Weibo data and census statistics is not significant. On the other hand, the 2014–2015 statistics published by Sina Corporation (the parent company of Weibo) indicate that the M:F ratio of active users (not limited to those who post their locations) decreased from 1.56 to 1 (Sina Corp 2013–2017), indicating a trend of increasing percentage of female active users on the site from 2014 to 2015. Table 3 shows the M:F ratios published by Sina Corp from 2013 to 2017. As can be seen, the M:F ratios fluctuate from year to year and do not show a clear increasing or decreasing trend.

Additionally, 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) (Sina Corp 2011). Our case study further confirmed that the M:F ratio continued to decrease to 0.50 in 2015 among users who checked in locations on Weibo, which implies that more and more females are actively using the check-in feature of Weibo in China.

Table 3. Male to female ratio (M:F ratio) from Sina Corp Yearly Report (Sina Corp 2013–2017).

Year	2013	2014	2015	2016	2017
M:F Ratio	1.00	1.56	1.00	1.25	1.29

Figure 1 also demonstrates the spatial distribution of LBSM gender biases among Chinese provinces. For example, the western provinces (Tibet, Qinghai, and Gansu) show a cluster of high $(M{:}F)_N$, meaning that female LBSM users in these provinces are less likely to identify their locations on LBSM compared to provinces with a low $(M{:}F)_N$.

Figure 2 shows the results of a Getis-Ord General G analysis. We adopt the Euclidean distance to measure the distance between the centres of each province. The clustering of high/low values is based on the commonly used inverse distance method. The result confirms a hotspot of $(M{:}F)_N$ in northwestern China, where women are less likely to share their locations on Weibo. In addition, we can also observe two moderate cold spots of $(M{:}F)_N$ ($1<\text{GiZscore}\leq 2$) in northeast and southwest China. Note that outlier provinces that have distinct patterns from their adjacent provinces may not be reflected in the General-G analysis. For example, in Figure 1, Xinjiang province in northwest China has a much lower $(M{:}F)_N$ than the surrounding provinces (i.e. Tibet and Qinghai); however, the General-G analysis did not pick up this isolated cold spot.

Because hotspot analysis mainly focuses on clustered extreme values, we also conducted a spatially constrained *K*-means clustering analysis to explore regional patterns in China (Esri 2015). We grouped Chinese provinces into clusters based on their adjacency and the

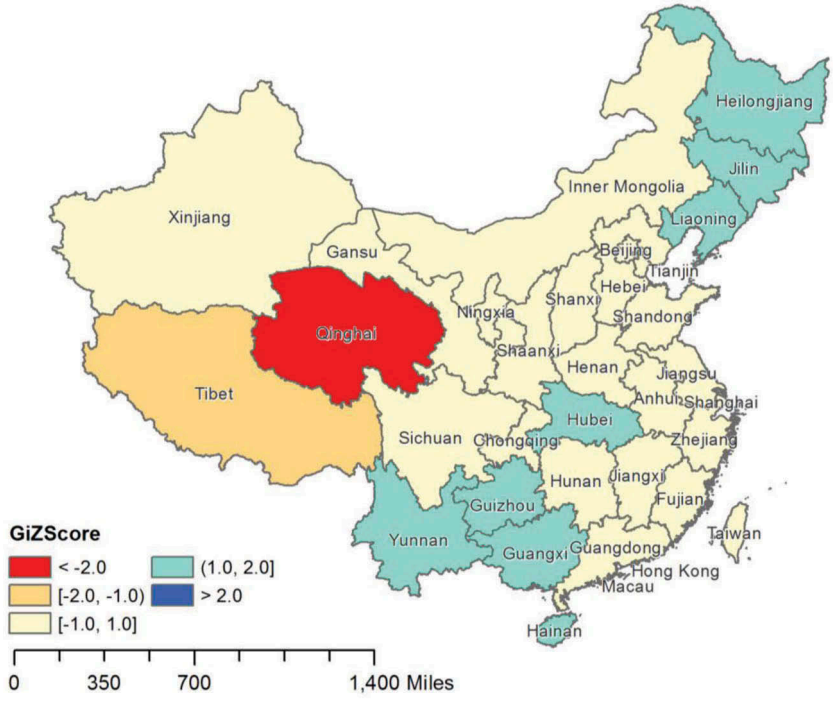

Figure 2. Getis-Ord General G analysis of $(M{:}F)_N$; GiZ Score < -1.0 indicates cold spots, GiZ Score >1.0 indicates hotspots. GiZ Score $\in [-1.0, 1.0]$ indicates that the feature is not a significant cold spot or hotspot.

$(M:F)_N$ index (c.f. Figure 3). The distance method is also Euclidean distance, and we only consider contiguity along edges. The number of clusters is determined based on the Pseudo *F*-statistics (Esri 2015). Here, we adopt 10 clusters for the analysis.

As can be seen, the normalized Weibo gender ratios of China provinces exhibit a clear regional pattern, which can help propose hypotheses in cultural geography and gender studies in China. Some examples are as follows.

- The three western provinces (Tibet, Qinghai, and Gansu) form a cluster of high $(M:F)_N$. These regions are generally perceived as 'under-developed' areas in China with low gross domestic product (GDP) values, which potentially affects women's openness in reporting their geographic locations in new media such as LBSM. A similar result was also reflected in Zhao, Yang, and Hao (2016), where the authors investigated the percentage of female principle investigators (PIs) funded by the Chinese National Science Foundation by province. Tibet, Qinghai, and Gansu are among the lowest, which further confirmed gender inequality in the sciences in these three provinces.
- Northeast China (Heilongjiang, Jilin, and Liaoning) as well as Inner Mongolia and Shanxi form a cluster of low $(M:F)_N$, whereas Central China (Shandong, Henan, and Anhui) forms a cluster of high $(M:F)_N$. In gender studies, male-to-female SRB is a commonly used indicator to reflect women's social status (Pani and Pani 2010; Edwards and Roces 2009; Poston et al. 1997). The SRB data used in this study is also from the 2010 census data collection. Previous studies showed that the northeastern provinces (Heilongjiang, Jilin, and Liaoning) have a low SRB that is below the first quartile (25%) of all Chinese provinces, but the central provinces (Shandong, Henan, and Anhui) are all among the provinces with the most imbalanced gender ratios (Poston et al. 1997; Wang, Leung, and Handayani 2006). This indicates that the traditional Chinese birth preference for sons is weaker in Northeast China, but much stronger in Shandong, Henan, and Anhui, which also demonstrates a possible correlation between the social status of women and their usage of LBSM. To test this hypothesis, we conducted a geographically weighted (GWR) analysis exploring the correlation between SRB and the normalized M:F ratio from Weibo. Unlike traditional ordinary least square (OLS) regression, GWR generates a separate regression equation for every feature analysed in a sample data set to address spatial variation. Therefore, it improves modelling accuracy and ameliorates residual errors by mitigating spatial-autocorrelation (Fotheringham, Brunsdon, and Charlton 2002; Di Ciaccio, Coli, and Angulo

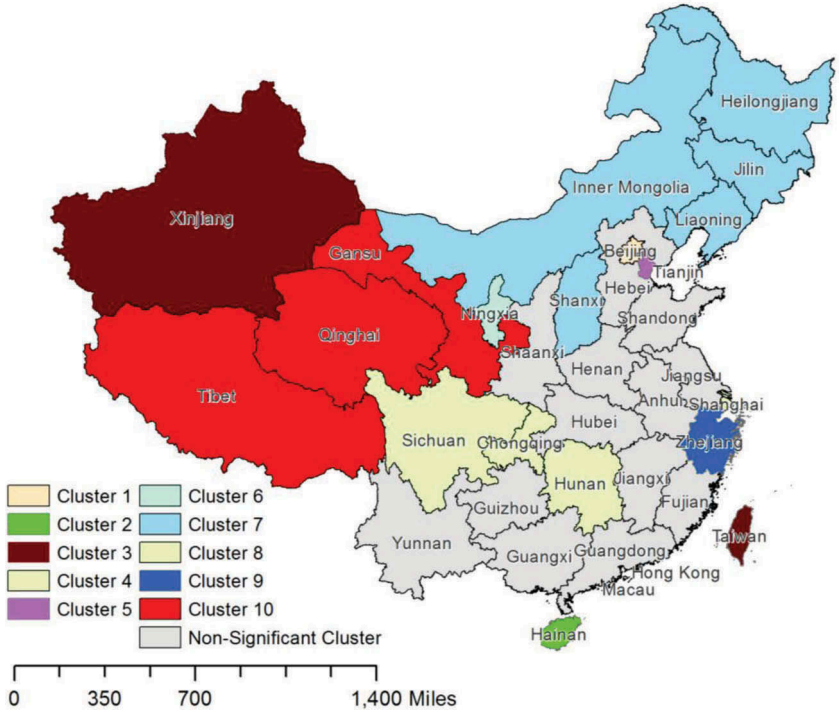

Figure 3. Grouping analysis of Chinese provinces based on $(M:F)_N$ index.

Ibañez 2012). Figure 4 shows the distribution of local R^2 for Chinese provinces. As can be seen, Shandong, Henan, and Anhui show the highest local R^2 between 0.6 and 0.8 (i.e. the strongest correlation between SRB and the normalized M:F ratio on Weibo). Several provinces/provincial-level regions in Southwest China (e.g. Sichuan and Chongqing) also show a moderate R^2 between 0.4 and 0.5. These are also the provinces with the lowest SRB and culturally more female-dominant in families and households. This further confirms our hypothesis that a higher female social status correlates positively with their location-sharing behaviour on Weibo. Note that GWR takes input from adjacent provinces to construct a regression model; hence, it did not fully capture the low cluster of normalized M:F ratio in northeast China. For example, Heilongjiang has only two adjacent provinces, making it challenging to construct a statistically significant model.

- The two special administrative units (Hong Kong and Macau) and Taiwan demonstrate very different behaviour from mainland China, with higher M:F ratios in general (Hong Kong: 0.70641; Macau: 1.752874; Taiwan: 0.965445). Macau is the only study area that has more reported male than female LBSM users on Weibo. The behavioural differences and the openness of women to LBSM in Hong Kong, Macau, 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.

This study aims to inspect the spatial pattern of gender biases from Weibo data, as well as provide a data-mining strategy for LBSM-related studies. The methodology used in this research is valuable for pattern recognition, interest identification, and hypotheses formulation in multiple areas, such as sociology, gender studies, urban planning, and cultural geography. This was best demonstrated by the concept of 'social sensing' proposed in Liu et al. (2015), where the authors argued that big (geo) data are powerful sensor tools for monitoring social activities in the age of instant access. For the researchers in the LBSM field, this study is valuable for demonstrating the potential data-quality issues and demographic biases in such data sets, which is crucial for designing a sound experiment and/or exploring geo-temporal factors causing these biases.

5. Discussion and conclusion

As discussed in Section 1, the motivation of this research is to explore the gender biases in Weibo users, as well as investigate the spatial autocorrelation

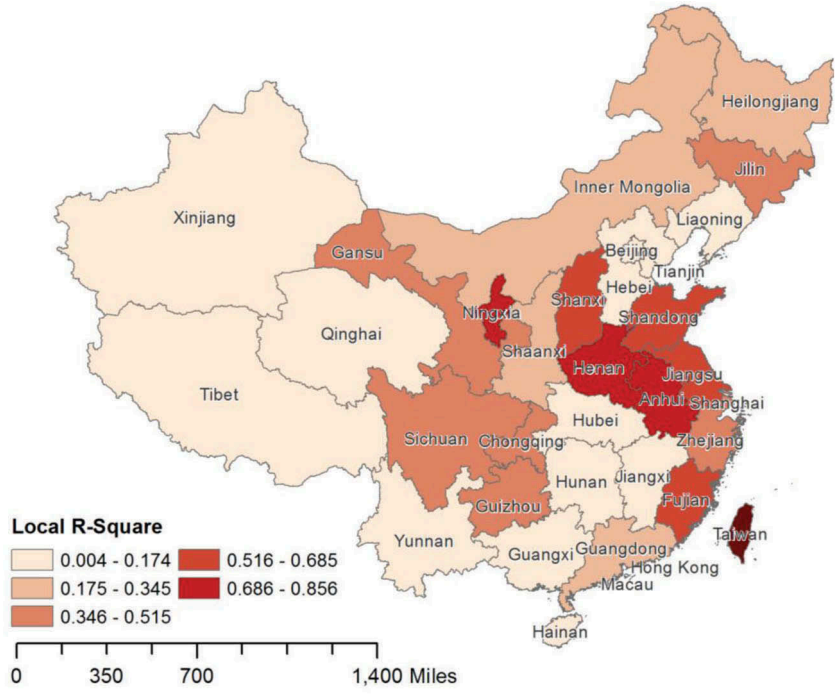

Figure 4. Local correlation coefficient (R^2) of a geographically weighted regression.

of such biases among Chinese provinces and the derived regional patterns. In summary, this study quantified the gender biases of Weibo users from various empirical perspectives.

- The results indicate that, in general, among the users who reported their age, gender, and current residential city, female users are more likely to report their locations on Weibo in China. This is consistent with previous findings on gender preferences on other LBSM sites, where researchers identified that females are more likely to report their geographic information on social media (Haffner et al. 2018). This verification of sampling statistics provides valuable input for various application fields, such as the personalization of LBSM user experiences. LBSM provide a rich data source for analysing demographic patterns at various spatial scales, such as investigating the variations inside different neighbourhoods of an urban system. The results provide valuable input for quantifying demographic bias in LBSM and investigating how this bias varies spatially. We also detected a strong regional pattern for LBSM gender ratios in different provinces, which further verified the conclusion of many regional studies that Chinese provinces are well bounded by diverse cultural backgrounds. The regional pattern of Weibo gender biases is potentially a synergistic result of multiple socio-economic factors, including but not limited to the social status of women, average GDP and income, etc. In this study, we confirmed the local correlation between gender ratio at birth and the normalized M:F ratio from Weibo.
- The methods and results of this study provide valuable input for various applications, including but not limited to: 1) quantifying and reducing demographic biases in LBSM data to achieve a more accurate result; 2) providing a data-mining strategy for social topics, such as gender inequality; and 3) reconfirming and validating regional patterns from other sociology studies. For example, it is well known that certain Chinese provinces are grouped or bounded by similar cultural backgrounds; however, it has been challenging to quantify such patterns in social studies before big (geo)data became available.

It is important to highlight that although this study generated valuable insights for investigating the biases of LBSM across geographies, there are several aspects that may be further addressed in future studies.

- Potential modifiable areal unit problem (MAUP): It is worth noting that the MAUP may affect the analysis results (Horner and Murray 2002; Jelinski and Wu 1996). MAUP refers to a source of statistical bias that can radically affect statistical hypothesis tests when point-based measures (e.g. population density) are aggregated into districts. As an exemplary study, we have used the provincial-level scale (consistent with the census data published by the Chinese census bureau). For example, Figure 5 depicts a point density analysis

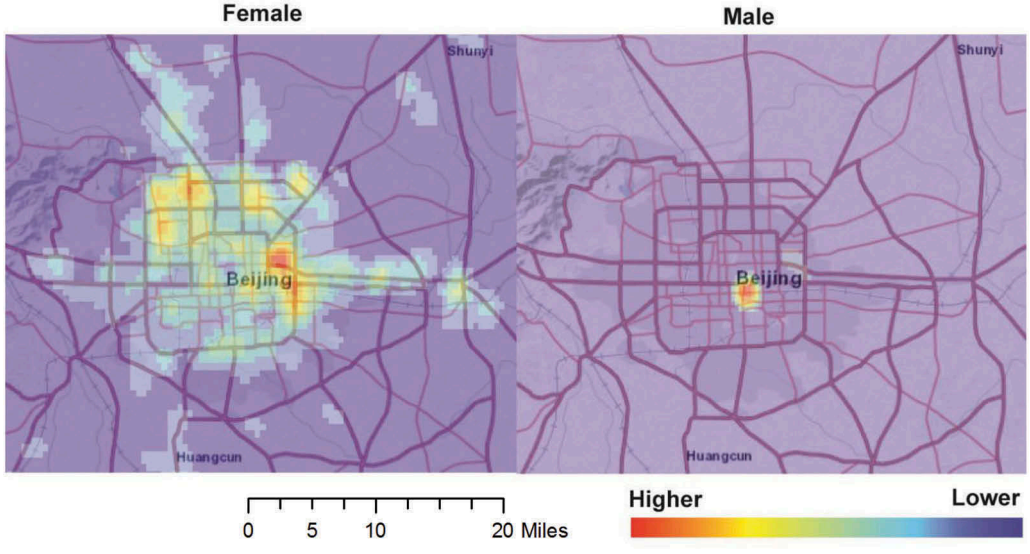

Figure 5. Point density distribution of male and female users in Beijing.

of male and female users in Beijing. As can be seen, the two gender groups demonstrate distinct patterns, where females are more spread out over the entire city with a few clusters in the southwest and the east side of the city, and male users clustered at the city centre. This may cause substantial sampling biases if the study area is only limited to a subregion of Beijing. Future studies can focus on exploring how spatial scales and MAUP impact our results (i.e. perform cross-comparisons of results at the city or sub-city levels).

- Accuracy of self-reported data: This study is based on self-reported data from user profiles. It is possible for users to report false data when registering for a Weibo account, which inevitably affects the results of this study. However, the gender/age ratio derived from this study is consistent with the trend indicated in the official demographic statistics published by the Sina Corporation (the parent company of Weibo) (Sina Corp 2013–2017), which supports the usability of user profile data despite its potential accuracy problem.
- Weibo API protocol sampling issue: In practice, the sampling strategy of LBSM access protocol is often controlled by the data vendor (González-Bailón et al. 2014). Weibo, for example, only makes unfiltered 'firehose' data available for selected business partners and leaves the sampling strategy and protocol of its publicly available Search and Streaming APIs in a black box. This research does not evaluate the effect of API sampling strategy because: 1) the firehose data are generally not available to the public; and 2) the volume of geotagged posts from such APIs was shown to be spatially representative and was close to the complete set, especially if a geographic bounding box was used (Morstatter et al. 2013).
- Creation of a more synthetic model for sampling bias: In addition to SRB, we also attempted to correlate the sampling bias indicator $(M{:}F)_N$ with various other socio-economic factors, such as the average income, urbanization ratio, and GDP of provinces, but none of the factors in isolation was significantly correlated with $(M{:}F)_N$. This is potentially due to the limited number of provinces that demonstrate a connection between $(M{:}F)_N$ and these explanatory variables to even construct a GWR regression model. In addition, this suggests that even though the sampling bias reveals a clear regional pattern, the underlying cause of how this bias varies spatially is a multifaceted synthetic effect, which may relate to various aspects of daily life, such as cultural background, economics, education, employment, government spending, etc., and these effects can be further explored in sociological studies.
- Correlation of demographic biases with activity space/trajectory analysis: This study is an initial attempt to quantify the data-quality issue of LBSM and how this issue manifests geographically. Our next step is to correlate the demographic information with specific individual activity space and trajectory indicators. Although previous studies have attempted to classify neighbourhoods (such as the Livehoods project (Cranshaw et al. 2012) and/or extract activity anchor points (e.g. 'home' and 'work') from LBSM (Qu and Zhang 2013), there is a lack of understanding of the morphology (e.g. shape, size) of activity space from such user-contributed data sets (Malleson and Birkin 2014), as well as the correlation of these measurements with user demographics.

The methods and models can be applied to other LBSM data sets (e.g. Twitter or Foursquare) to test their robustness. Even though demographics may not be directly available on certain SNS such as Twitter, it is potentially obtainable through semantic analysis based on previous studies. We will further extend this analysis to other demographic factors such as age, employment, and education level. In this study, we used only gender ratios to measure the overall number of Weibo users in each demographic group; future studies may measure more detailed aspects of *how* these demographic groups are using Weibo, such as the frequency and time of check-ins. Owing to the nature of self-reported data, it is possible for users to falsify or spoof their profile information; future research can cross-compare user profile information by analysing the content of their Weibo posts and the structure of their social network. This study only included users who reported their age, gender, and residential city, which can potentially introduce additional biases into the sample set. In addition, LBSM as an input to the analysis of human mobility has the potential to transform research in diverse fields, including geography, transportation, planning, and economics, and this study provides a reference for verifying LBSM user sampling biases when using such data in human mobility studies.

Note

1. http://www.datatang.com/data/46324.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

APA. 2015. "Key Terms and Concepts in Understanding Gender Diversity and Sexual Orientation Among Students." https://www.apa.org/pi/lgbt/programs/safe-supportive/lgbt/key-terms.pdf

Argamon, S., M. Koppel, J. Fine, and A. Shimoni. 2003. "Gender, Genre, and Writing Style in Formal Written Texts." *Text - Interdisciplinary Journal for the Study of Discourse* 23: 321–346.

Barbier, G., R. Zafarani, H. Gao, G. Fung, and H. Liu. 2012. "Maximizing Benefits from Crowdsourced Data." *Computational and Mathematical Organization Theory* 18: 257–279. doi:10.1007/s10588-012-9121-2.

Bawa-Cavia, A. 2011. "Sensing the Urban: Using Location-Based Social Network Data in Urban Analysis." In *The First Workshop on Pervasive Urban Applications (PURBA)*. San Francisco, CA.

Burger, J. D., J. Henderson, G. Kim, and G. Zarrella. 2011. "Discriminating Gender on Twitter." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1301–1309. Edinburgh, United Kingdom: Association for Computational Linguistics.

Business Insider. 2014. "This Chart Reveals the Age Distribution at Every Major Social Network." http://www.businessinsider.com/age-distribution-of-facebook-twitter-instagram-2014-11#ixzz3jtx8Nlao

Calabrese, F., L. Ferrari, and V. D. Blondel. 2015. "Urban Sensing Using Mobile Phone Network Data: A Survey of Research." *ACM Computing Surveys* 47: 1–20.

Carnegie Mellon University. 2014. "Using Social Media for Large Behavioral Studies Is Fast and Cheap, but Fraught with Biases and Distortion." https://www.cmu.edu/news/stories/archives/2014/december/december1_socialmediadatabiased.html

Cho, E., S. A. Myers, and J. Leskovec. 2011. "Friendship and Mobility: User Movement in Location-Based Social Networks." In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1082–1090. San Diego, California, USA: ACM.

Chow, T. E., Y. Lin, and W. D. Chan. 2011. "The Development of a Web-Based Demographic Data Extraction Tool for Population Monitoring." *Transactions in GIS* 15: 479–494. doi:10.1111/j.1467-9671.2011.01274.x.

Crampton, J. W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, and M. Zook. 2013. "Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb." *Cartography and Geographic Information Science* 40: 130–139. doi:10.1080/15230406.2013.777137.

Cranshaw, J., R. Schwarts, J. Hong, and N. Sadeh. 2012. "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City." In *The sixth International AAAI Conference oon Webpages and Social Media*. Dublin, Ireland.

Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski. 2013. "#Earthquake: Twitter as a Distributed Sensor System." *Transactions in GIS* 17: 124–147. doi:10.1111/tgis.2013.17.issue-1.

De Souza e Silva, A. 2007. "Mobile Phones and Places: The Use of Mobile Technologies in Brazil." In *Societies and Cities in the Age of Instant Access*, edited by H. J. Miller, 295–310. Dortdrecht, The Netherlands: Springer.

DeBarr, D., and H. Wechsler. 2010. "Using Social Network Analysis for Spam Detection." *Advances in Social Computing, Proceedings* 6007: 62–69.

Di Ciaccio, A., M. Coli, and J. M. Angulo Ibañez. 2012. *Advanced Statistical Methods for the Analysis of Large Data-Sets*. Berlin; New York: Springer.

Edwards, L. P., and M. Roces. 2009. *Women in Asia: Critical Concepts in Asian Studies*. Milton Park, Abingdon, Oxon; New York: Routledge.

Elwood, S. 2006. "Critical Issues in Participatory GIS: Deconstructions, Reconstructions, and New Research Directions." *Transactions in GIS* 10: 693–708. doi:10.1111/tgis.2006.10.issue-5.

Elwood, S., and A. Leszczynski. 2011. "Privacy, Reconsidered: New Representations, Data Practices, and the Geoweb." *Geoforum* 42: 6–15. doi:10.1016/j.geoforum.2010.08.003.

Esri. 2015. "ArcGIS Pro - Grouping Analysis." http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/grouping-analysis.htm

Fekete, E. 2015. "Race and (Online) Sites of Consumption." *Geographical Review* 105: 472–491. doi:10.1111/j.1931-0846.2015.12106.x.

Forbes. 2014. "Scientists Warn About Bias In The Facebook And Twitter Data Used In Millions Of Studies." http://www.forbes.com/sites/bridaineparnell/2014/11/27/scientists-warn-about-bias-in-the-facebook-and-twitter-data-used-in-millions-of-studies

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, England; Hoboken, NJ, USA: Wiley.

Gao, H., and H. Liu. 2015. *Mining Human Mobility in Location-Based Social Networks*. San Rafael, CA: Morgan & Claypool Publisher.

Gao, H., J. Tang, and H. Liu. 2012. "Exploring Social-Historical Ties on Location-Based Social Networks." In *6th International AAAI Conference on Weblogs and Social Media*, 114–121. Dublin, Ireland.

Gelfand, A. E. 2010. *Handbook of Spatial Statistics*. Boca Raton: CRC Press.

Golub, B., and M. O. Jackson. 2010. "Naive Learning in Social Networks and the Wisdom of Crowds." *American Economic Journal-Microeconomics* 2: 112–149. doi:10.1257/mic.2.1.112.

González-Bailón, S., N. Wang, A. Rivero, J. Borge-Holthoeferd, and Y. Moreno. 2014. "Assessing the Bias in Samples of Large Online Networks." *Social Networks* 38: 16–27. doi:10.1016/j.socnet.2014.01.004.

Goodchild, M. F. 2013. "The Quality of Big (Geo)Data." *Dialogues in Human Geography* 3: 280–284. doi:10.1177/2043820613513392.

Griffith, D. 2003. *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization*. Berlin: Springer Science & Business Media.

Griffith, D. A. 1988. *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*. Dordrecht; Boston: Kluwer Academic Publishers.

Guan, W., H. Gao, M. Yang, Y. Li, H. Ma, W. Qian, Z. Cao, and X. Yang. 2014. "Analyzing User Behavior of the Micro-Blogging Website Sina Weibo during Hot Social Events." *Physica A: Statistical Mechanics and Its Applications* 395: 340–351. doi:10.1016/j.physa.2013.09.059.

Guo, D., and C. Chen. 2014. "Detecting Non-Personal and Spam Users on Geo-Tagged Twitter Network." *Transactions in GIS* 18: 370–384. doi:10.1111/tgis.2014.18.issue-3.

Haffner, M., A. J. Mathews, E. Fekete, and G. A. Finchum. 2018. "Location-Based Social Media Behavior and Perception: Views of University Students." *Geographical Review* 108: 203–224. doi:10.1111/gere.2018.108.issue-2.

Harvey, F. 2013. "To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information." In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, edited by D. Sui, S. Elwood, and M. Goodchild, 31–42. Dordrecht: Springer Netherlands.

Hasan, S., X. Zhan, and S. V. Ukkusuri. 2013. "Understanding Urban Human Activity and Mobility Patterns Using Large-Scale Location-Based Data from Online Social Media." In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 1–8. Chicago, Illinois: ACM.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. 2014. "Geo-Located Twitter as Proxy for Global Mobility Patterns." *Cartography and Geographic Information Science* 41: 260–271. doi:10.1080/15230406.2014.890072.

Horner, M. W., and A. T. Murray. 2002. "Excess Commuting and the Modifiable Areal Unit Problem." *Urban Studies* 39: 131–139. doi:10.1080/00420980220099113.

IBM. 2015. "The Four V's of Big Data." http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Jelinski, D. E., and J. Wu. 1996. "The Modifiable Areal Unit Problem and Implications for Landscape Ecology." *Landscape Ecology* 11: 129–140. doi:10.1007/BF02447512.

Kitchin, R. 2013. "Big Data and Human Geography." *Dialogues in Human Geography* 3: 262–267. doi:10.1177/2043820613513388.

Kitchin, R. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles, California: SAGE Publications.

Leszczynski, A., and J. Crampton. 2016. "Introduction: Spatial Big Data and Everyday Life." *Big Data & Society* 3: 1–6. doi:10.1177/2053951716661366.

Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. 2005. "Geographic Routing in Social Networks." *Proceedings of the National Academy of Sciences of the United States of America* 102: 11623–11628.

Liu, H., Z. Q. Dong, and H. Y. Gu. 2014a. "Microblogging as a Social Sensing Tool." *2014 IEEE 11th International Conference on Networking, Sensing and Control (Icnsc)*, 513–517.

Liu, Y., X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi. 2015. "Social Sensing: A New Approach to Understanding Our Socioeconomic Environments." *Annals of the Association of American Geographers* 105: 512–530. doi:10.1080/00045608.2015.1018773.

Liu, Y., Z. W. Sui, C. G. Kang, and Y. Gao. 2014b. "Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data." *PLoS ONE* 9: e86026. doi:10.1371/journal.pone.0086026.

Longley, P. A., and M. Adnan. 2016. "Geo-Temporal Twitter Demographics." *International Journal of Geographical Information Science* 30: 369–389. doi:10.1080/13658816.2015.1089441.

Longley, P. A., M. Adnan, and G. Lansley. 2015. "The Geotemporal Demographics of Twitter Usage." *Environment and Planning A* 47: 465–484. doi:10.1068/a130122p.

Lu, Y. 2000. "Spatial Cluster Analysis of Point Data: Location Quotients versus Kernel Density." In *2000 University Consortium of Geographic Information Science (UCGIS) Summer Assembly Graduate Papers*. Portland, Oregon.

Malleson, N., and M. Birkin. 2014. "New Insights into Individual Activity Spaces Using Crowd-Sourced Big Data." In *ASE Bigdata/Socialcom/Cybersecurity Conference*. Stanford, CA.

Marwick, A. 2013. "Gender, Sexuality and Social Media." In *Routledge Handbook of Social Media*, edited by T. Senft and J. Hunsinger, 59–75. New York: Routledge.

Mazman, G., and Y. Usluel. 2011. "Gender Differences in Using Social Networks." *Turkish Online Journal of Educational Technology* 10: 133–139.

Mislove, A., S. Lehmann, -Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. 2012. "Understanding the Demographics of Twitter Users." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 554–557. Association for the Advancement of Artificial Intelligence.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." In *Proceedings of ICWSM*, 9. Association for the Advancement of Artificial Intelligence.

Musolesi, M., S. Hailes, and C. Mascolo. 2004. "An Ad Hoc Mobility Model Founded on Social Network Theory." In *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, 20–24. Venice, Italy: ACM.

National Bureau of Statistics of China. 2010. "2010 Population Census of China." http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm

Noulas, A., C. Mascolo, and E. Frias-Martinez. 2013. "Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments." In *2013 IEEE 14th International Conference on Mobile Data Management (Mdm 2013), Vol 1*, 167–176.

Noulas, A., S. Scellato, R. Lambiotte, M. Pontil, and C. Masscolo. 2012. "A Tale of Many Cities: Universal Patterns in Urban Mobility." *PLoS ONE* 7. doi:10.1371/annotation/ca85bf7a-7922-47d5-8bfb-bcdf25af8c72.

Pani, S. P., and N. Pani. 2010. *Essays on Contemporary Gender Issues*. New Delhi: Hirmoli Press.

Poorthuis, A., and M. Zook. 2017. "Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media." *Journal of Urban Technology* 24: 115–135. doi:10.1080/10630732.2017.1335153.

Poston, D. L., B. Gu, P. P. Liu, and T. McDaniel. 1997. "Son Preference and the Sex Ratio at Birth in China: A Provincial Level Analysis." *Social Biology* 44: 55–76.

Qu, Y., and J. Zhang. 2013. "Regularly Visited Patches in Human Mobility." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 395–398. Paris, France: ACM.

Ripley, B. D. 2004. *Spatial Statistics*. Hoboken, N.J.: Wiley-Interscience.

Roick, O., and S. Heuser. 2013. "Location Based Social Networks - Definition, Current State of the Art and Research Agenda." *Transactions in GIS* 17: 763–784.

Rutherford, A., M. Cebrian, S. Dsouza, E. Moro, A. Pentland, and I. Rahwan. 2013. "Limits of Social Mobilization." *Proceedings of the National Academy of Sciences* 110: 6281–6286.

Rzeszewski, M. 2018. "Geosocial Capta in Geographical Research – A Critical Analysis." *Cartography and Geographic Information Science* 45: 18–30.

Saini, J. S. 2014. "A Study of Spam Detection Algorithm on Social Media Networks." *Computational Intelligence, Cyber Security and Computational Models* 246: 195–202.

Scholtz, R. W., and Y. Lu. 2014. "Detection of Dynamic Activity Patterns at a Collective Level from Large-Volume Trajectory Data." *International Journal of Geographical Information Science* 28: 946–963. doi:10.1080/13658816.2013.869819.

Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, et al. 2013. "Personality, Gender, and Age in the Language of Socia Media: The Open-Vocabulary Approach." *PLoS ONE* 8: e73791. doi:10.1371/journal.pone.0073791.

Sina Corp. 2011. "Chinese SNS and Weibo User Behavior Research." http://it.sohu.com/20110519/n280623821.shtml

Sina Corp. 2013–2017. "Weibo User Background Report." http://data.weibo.com/report/

Sloan, L., J. Morgan, P. Burnap, and M. Williams. 2015. "Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data." *PLoS ONE* 10: e0115545.

Spielman, S. E. 2014. "Spatial Collective Intelligence? Credibility, Accuracy, and Volunteered Geographic Information." *Cartography and Geographic Information Science* 41: 115–124. doi:10.1080/15230406.2013.874200.

Tobler, W. 1979. "Cellular Geography." In *Philosophy in Geography*, edited by S. Gale and G. Olsson, 379–386. Dordrecht: Reidel.

Tobler, W. 2004. "On the First Law of Geography: A Reply." *Annals of the Association of American Geographers* 94: 304–310. doi:10.1111/j.1467-8306.2004.09402009.x.

Trinh Minh Tri, D., and D. Gatica-Perez. 2014. "The Places of Our Lives: Visiting Patterns and Automatic Labeling from Longitudinal Smartphone Data." *IEEE Transactions on Mobile Computing* 13: 638–648. doi:10.1109/TMC.2013.19.

Tsou, M.-H., C.-T. Jung, C. Allen, J.-A. Yang, J.-M. Gawron, B. H. Spitzberg, and S. Han. 2015. "Social Media Analytics and Research Test-Bed (SMART Dashboard)." In *Proceedings of the 2015 International Conference on Social Media & Society*, 1–7. Toronto, Ontario, Canada: ACM.

Tufekci, Z. 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, edited by E. Adar, P. Resnick, M. De Choudhury, B. Hogan, A. H. Oh, 505-514. Palo Alto, CA: The AAAI Press.

van Oosten, J. M. F., L. Vandenbosch, and J. Peter. 2017. "Gender Roles on Social Networking Sites: Investigating Reciprocal Relationships between Dutch Adolescents' Hypermasculinity and Hyperfemininity and Sexy Online Self-Presentations." *Journal of Children and Media* 11: 147–166. doi:10.1080/17482798.2017.1304970.

Veregin, H. 1999. "Data Quality Parameters." In *Geographical Information Systems*, edited by P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, 188–189. New York: NY: Wiley.

Volkovich, Y., D. Laniado, K. E. Kappler, and A. Kaltenbrunner. 2014. "Gender Patterns in a Large Online Social Network." In *Social Informatics*, edited by L. M. Aiello and D. McFarland, 139–150. Cham: Springer International Publishing.

Wang, C., A. Leung, and S. W. Handayani. 2006. "China: Research Report on Gender Gaps and Poverty Reduction. Worldbank.

Westerman, D., P. R. Spence, and B. Van der Heide. 2014. "Social Media as Information Source: Recency of Updates and Credibility of Information." *Journal of Computer-Mediated Communication* 19: 171–183. doi:10.1111/jcc4.12041.

Wu, L., Y. Zhi, Z. W. Sui, and Y. Liu. 2014. "Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data." *PLoS ONE* 9: e97010. doi:10.1371/journal.pone.0097010.

Xia, Y. 2005. "Integrating Uncertainty in Data Mining." *Department of Computer Science*, 165. Los Angeles: University of California.

Ye, Z., N. H. Hashim, F. Baghirov, and J. Murphy. 2018. "Gender Differences in Instagram Hashtag Use." *Journal of Hospitality Marketing & Management* 27: 386–404. doi:10.1080/19368623.2018.1382415.

Yuan, Y., and G. Wei. 2016. "Evaluating Demographic Representativeness of Location-Based Social Media: A Case Study of Weibo." In *Annual Conference of the Association of European Geo-Information Laboratories*, 1–2. Helsinki, Finnland.

Zamri, M. H., M. D. Darson, and A. M. F. Wahab. 2014. "Social Media: Credibility, Popularity and Its Benefits Towards Events' Awareness." In *Theory and Practice in Hospitality and Tourism Research*, edited by S. M. Radzi, M. F. Saiful Bakhtiar, Z. Mohi, M. S. M. Zahari, N. Sumarjan, C. T. Chik, and F. I. Anuar, 317–321. London: CRC Press.

Zhang, K., Q. Yu, K. Lei, and K. Xu. 2012. "Characterizing Tweeting Behaviors of Sina Weibo Users via Public Data Streaming." In *13th International Conference, WAIM 2012*, edited by. H. Gao, L. Lim, W. Wang, C. Li, and L. Chen. Harbin, China.

Zhang, W., B. Derudder, J. Wang, W. Shen, and F. Witlox. 2016. "Using Location-Based Social Media to Chart the Patterns of People Moving between Cities: The Case of Weibo-Users in the Yangtze River Delta." *Journal of Urban Technology* 23: 91–111. doi:10.1080/10630732.2016.1177259.

Zhao, Y., X. Yang, and L. Hao. 2016. "Differentiations of Subject Distribution and Regional Distribution of Chinese Female Social Science Talents: Based on National Social Science Fund Projects 2000-2015." *Arid Land Geography* 39: 1350–1357.

Zhong, Y., N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. 2015. "You are Where You Go: Inferring Demographic Attributes from Location Check-Ins." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 295–304. Shanghai, China: ACM.

Zickuhr, K. 2013. "Location-Based Services." In *Pew Research Center's Internet & American Life Project*. Washington, D.C.: Pew Research Center. http://www.pewinternet.org/2013/09/12/location-based-services-2/

Zook, M. A., and M. Graham. 2007. "The Creative Reconstruction of the Internet: Google and the Privatization of Cyberspace and DigiPlace." *Geoforum* 38: 1322–1343. doi:10.1016/j.geoforum.2007.05.004.