YIHONG YUAN
Department of Geography
Texas State University, San Marcos
Email: yuan@txstate.edu

Yihong Yuan received her Ph.D. in Geography and M.A in Statistics from the University of California, Santa Barbara. She is currently an Assistant Professor of Geography at Texas State University (Texas State).

Before joining Texas State, she worked as a visiting researcher at ETH Zurich and as a senior data scientist in the private sector. Yuan's research focuses on big geo-data analytics and spatial-temporal knowledge discovery. She is particularly interested in extracting human activity patterns from multiple data sources, such as telecommunication devices and social networking sites. She has published more than 30 papers in peer-reviewed journals and conference proceedings. She also serves as an expert reviewer for multiple academic journals, including *Nature* (*Human Behavior*), *Annals of the American Geographers,* and *the International Journal of Geographic Information Science*.

The Missing Parts from Location-Based Social Media: Who, Where, When, and What

Researchers have defined location-based social media (LBSM) as "Social Network Sites (SNS) that include location information"(Roick and Heuser 2013). LBSM has been widely used as potential resources to characterize social perceptions of places and to model human activities in various applications. However, like other types of big geo-data, LBSM data have various data quality issues, such as accuracy, precision, completeness, and representativeness. Different LBSM tend to attract certain population groups and support the sharing of particular content, making them limited in data representation (Golub and Jackson 2010). In other words, biased sampling (e.g., demographically, spatially, temporally, and semantically) naturally leads to data representativeness issues. If LBSM data are applied to decision-making in city services, understanding the sampling biases of such data is critical for implementing better policies or management practices. This study discusses the representativeness of LBSM data and their impacts on human activity modeling from sociodemographic, spatiotemporal, and semantic perspectives. The main objective is providing a framework to examine the sampling biases of LBSM and their limitations when applied to city services.

First, LBSM users are not a random sample in terms of their social, economic and demographic background (Golub and Jackson 2010). Naturally, such demographic biases may impact the reliability of applying LBSM to urban services. Previous research either conducted user surveys or harvested user profiles or posts to infer their demographics (Longley and Adnan 2016). Despite the challenges in

mitigating LBSM demographic biases, it still provides a valuable data source for smart city applications. Salganik (2018) pointed out that even though social scientists are more used to probabilistic random samples from a well-defined population, nonrepresentative data can still provide valuable insights, especially in the exploratory stage of outlier patterns and causations. Therefore, if city officials were to rely on nonrepresentative social media data to engage a broader audience in urban planning and infrastructure renovation, it would be important to identify suitable research questions. For example, it is feasible to answer questions like "Are there abnormal spatial clusters in the city during a musical festival?" based on LBSM data; however, questions like "What is the average number of people impacted by Hurricane Harvey in each county?" requires more representative data and cannot be answered solely based on social media data.

Second, studies have demonstrated that check-in data tend to cluster in certain areas and times, causing an biased profiling of activities across space and through time (Sloan et al. 2015). For example, it is more likely for users to post during holidays and at special events. Bawa-Cavia (2011) identified social hubs (i.e., where social media users are more likely to generate a high density of activities) in London, New York, and Paris. Hence, there are inherent biases and representativeness issues in the spatio-temporal data acquired from LBSM. The "where" and "when" challenges are beyond simple sampling biases. Locations and time stamps from LBSM data may have different levels of accuracy; space-based geotagged posts with precise x- and y-coordinates are more accurate than place-based posts using a descriptive of or reference to a loosely-defined location. For example, "Houston" can refer to its downtown area, the centroid of the city, or anywhere within the city limit as determined by that social media platform.

Third, semantic biases from LBSM are also worth noting. The content of social media is closely related to the functionalities and characteristics of each SNS platform. Inevitably, various biases exist when conducting sentiment analysis, public opinion collection, and topic extraction from such datasets. Instead of expressing opinions on public matters such as traffic, politics, or urban planning, social media users are more willing to publicly discuss topics related to their personal life (e.g., leisure activities) (Lansley and Longley 2016). Hence, if policy makers aim to collect opinions on city services, it is crucial to understand the nature, popularity, and associated sentiment of various topics on social media.

It is important to note that the above limitations are often inseparable. For example, SNS tend to attract young people, who have their own preferred check-in locations and topics to discuss on social media. In other words, user-sampling bias is the foundation of social media biases. In the meantime, activities conducted by these users distribute unevenly across space and time. Furthermore, these unevenly distributed activities also have different likelihoods of being posted to social media. Therefore, we should consider the missing parts of LBSM in a synergistic way when developing city services and other geospatial applications.

Despite a lack of solutions to fully address or quantify these deficiencies of LBSM data, there are several ways to mitigate the potential problem. First, LBSM data can always be supplemented or

corroborated by other data sources, such as census data and survey data, to improve the representativeness of LBSM samples. Second, it is important to identify target user groups from SNS data. Although user sample biases are inevitable, researchers can still extract the most representative groups on different social media sites and design their research objectives according to the user groups available. Third, due to the low spatio-temporal sampling resolution of LBSM data, it is necessary to re-evaluate the validity of classic mobility models, measurements, and algorithms when applied to such datasets.

References

Salganik, M. J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Longley, P. A., and M. Adnan. 2016. Geo-Temporal Twitter Demographics. *International Journal of Geographical Information Science* 30(2): 369–389.

Lansley, G., and P. A. Longley. 2016. The Geography of Twitter Topics in London. *Computers Environment and Urban Systems* 58: 85–96.

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(3): e0115545.

Roick, O., and S. Heuser. 2013. Location Based Social Networks - Definition, Current State of the Art and Research Agenda. *Transactions in GI* 17(5): 763–784.

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.

Golub, B., and M. O. Jackson. 2010. Naive Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal-Microeconomics* 2(1): 112–149.