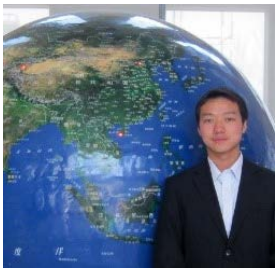
SONG GAO
Geospatial Data Science Lab
University of Wisconsin, Madison
Email: song.gao@wisc.edu



Song Gao is a tenure-track Assistant Professor at the Department of Geography, University of Wisconsin, Madison, where he directs the Geospatial Data Science Lab (GeoDS@UW-Madison). He got his Ph.D. in Geography from the University of California, Santa Barbara. He holds a M.S. degree at the Institute of Remote Sensing and GIS, Peking University, and a B.S. degree at Beijing Normal University, China. His main research interests include Big Geo-Data Analytics, Place-Based GIS, and Geospatial Semantics, as well interdisciplinary studies. He is the (co-)author of over 40 peer-review scientific articles that were published in prominent GIScience journals and conference proceedings. He is in the board of directors for AAG GI Systems and Science. He was the recipient for 2019 Microsoft AI for Earth Grant, 2018 & 2019 UW-Madison WARF Research Competition Grants, 2017 "Young Researcher Award" in GIScience selected by Austrian Academy of Sciences, 2017 International Cartographic Association Scholarship, 2016 National Award for Outstanding Chinese PhD Students Study Abroad, 2014 Cartography and Geographic Information Society Doctoral Scholarship Awards, Jack & Laura Dangermond Graduate Fellowship, and UCSB Geography Excellence in Research Award. He had industry internship experience at Esri and Apple Maps.

Toward Multi-Source Urban Sensing with Geospatial Data Science

The urbanization process is accelerating in world cities and attracting large-scale job opportunities, human flows, business, and social activities. With the rapid development of information and communication technologies (ICT), location-aware devices and sensor networks, the emergence of multi-source geospatial big data brings new opportunities to understand the rich semantics of space and place and associated human activities in urban areas using large-scale crowdsourcing data streams, such as geotagged social media posts, travel blogs, mobile phone data, smart card data from transportation, GPS-enabled ridesharing services, and so forth. While crowdsourced data offer promising opportunities, several internal challenges and limitations of crowdsourced data should be addressed for urban studies as follows.

First, although large volumes of content are contributed by millions of users every second, we may get a very sparse data matrix (e.g., Lee, Gao, and Goulias 2015) after slicing the user-generated content (UGC) into a fine spatiotemporal resolution (e.g., a city-block spatial unit with hourly temporal window), which is crucial in solving some urban problems such as transportation planning and traffic congestion control. The spatiotemporal data sparsity issue becomes more prominent in the regions with limited number of active users. Due to the reduced size of data volume, the uncertainty in each slice may increase when analyzing the data.

Second, a common concern about crowdsourcing refers to the lack of standardization for users in the data generation process, which causes poor data quality and low trustworthiness, as well as high uncertainty (Senaratne et al. 2017). Users produce geographic data based on their local knowledge and their perception of the place, which may vary across different users (Stephens 2013). And due to the vagueness and uncertainty in human conceptualization of location, space, and place, it is hard for users to express some geographic regions and spatial relations precisely (Montello et al. 2003; Goodchild and Li, 2012). Thus, an approach driven by data synthesis (Gao et al. 2017), combining UGC with a fuzzy-set theory informed approach (Wu, Wang, and Shi et al. 2019), and combining UGC with survey-based behavior approaches (Twaroch et al. 2019) have been proposed to address the abovementioned concerns. For instance, users may have different perceptions and cognitions for the same place, which can cause incorrect tagging behaviors for social media photos (Hollenstein and Purves, 2010).

The third issue concerns the representativeness, which refers to the degree to which UGC observation samples can represent the actual population. Existing studies have figured out that the information shared on social-media platforms usually follows a power-law distribution, indicating that only a small proportion of users contribute most of the content online (Kwak et al. 2010). Besides, the demographic bias in contributors also impedes the representativeness (Longley and Adnan 2016). For example, not all people in the real world use social media frequently. People who have limited access to the social media, such as the elderly group and the users in developing countries, may be less sampled by UGC.

In order to address these limitations and challenges, we propose to utilize multi-source data streams and theory-informed approaches to urban sensing (Liu et al. 2015; Janowicz et al. 2019). In traditional urban strategic planning or the classification results of remote sensing, many places in urban areas may be labelled as single-use land use type; however, these areas may in reality contain multiple functions and land uses. In order to capture citywide dynamics of both human activities and urban functions at finer resolutions, multi-source data are combined to over-come their own limitations and to enrich the understanding of urban spatial structure and neighborhood demographics. Both mobile phone data and taxi trajectories usually cover large numbers of users and contain rich location information but lack of place semantics. Social media data are sparsely distributed in space and time but contain rich content. By combining both mobile phone data and social media, it is possible to extract citizen's home-job locations and social activity dynamics more effectively in space and time in cities (Tu et al. 2017). Also, by the integration of mobilephone data and crowdsourced taxi trajectories, or the fusion of POI data and crowdsourced taxi trajectories, researchers have uncovered substantial differences between taxi trips and mobilephone-based human movements in terms of spatial distribution and distance-decay effects (Kang et al. 2013).

In the future, a number of multi-source data-fusion research areas that call for attention in urban sensing powered by geospatial data science. First, the data sampling and fusing resolution requirements in space and time need to be investigated among different sources to comprehensively understand human activities of different gender, age, and socioeconomic groups and place semantics

for intra-urban and inter-city human mobility modeling. Second, combining UGC and professional-generated content (PGC) or combining data-driven and knowledge-driven approaches may help solve urban problems such as traffic congestion and environmental pollution. Third, the development of data-sharing portals, methods, tools, and platforms for advancing geospatial data science. Last but not least, there is a need to increase the engagement of citizen science in addressing urban changes in responsive cities through data-smart governance (Goldsmith and Crawford 2014).

References

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J. A., McKenzie, G., Ju, Y., Adams, B., and Yan, B. (2017b). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science* 31(6): 1245–1271.

Goldsmith, S., and Crawford, S. (2014). *The responsive city: Engaging communities through data-smart governance*. John Wiley and Sons.

Goodchild, M. F., and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics* 1: 110–120.

Hollenstein, L., and Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 2010(1): 21–48.

Kang, C., Sobolevsky, S., Liu, Y., and Ratti, C. (2013, August). Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (pp. 1–8). ACM.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media*? Proceedings of the 19th International Conference on World Wide Web*, 591–600.

Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., and Gao, S. (2019). Using semantic signatures for social sensing in urban environments. In *Mobility patterns, big data and transport analytics* (pp. 31–54). Elsevier.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's downtown? Behavioral methods for determining referents of vague spatial queries. In *Spatial Cognition and Computation* 3(2–3): 185–204.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., and Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers* 105(3): 512–530.

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

Lee, J. H., Gao, S., and Goulias, K. (2015). Can Twitter data be used to validate travel demand models. In *IATBR 2015-WIND*.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., and Haklay, M. (Muki). (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science* (31): 139–167.

Tu, W., Cao, J., Yue, Y., Shaw, S. L., Zhou, M., Wang, Z., and Li, Q. (2017). Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science* 31(12): 2331–2358.

Twaroch, F. A., Brindley, P., Clough, P. D., Jones, C. B., Pasley, R. C., and Mansbridge, S. (2019). Investigating behavioural and computational approaches for defining imprecise regions. *Spatial Cognition and Computation* 19(2): 146–171.

Wu, X., Wang, J., Shi, L., Gao, Y., and Liu, Y. (2019). A fuzzy formal concept analysis-based approach to uncovering spatial hierarchies among vague places extracted from user-generated data. *International Journal of Geographical Information Science* 33(5): 991–1016.