Evaluating Sampling Bias in Geotagged Social Media Data

Ruowei Liu and Angela Yao

ABSTRACT:

In recent years, there has been growing interest in geotagged social media data. Twitter is one of the most popular geotagged social media datasets that has been applied in all kinds of cross-disciplinary research. Much of the existing research is built on the following assumption: a large amount of the data indicates a large sample coverage, so the data can represent the entire population or possible sampling bias can be dismissed. However, this assumption is not true. Therefore, despite the success of such research, the sampling biases of such data are still under-investigated.

Sampling bias is a misrepresentation of the population when some members of the population have a lower or higher sampling probability than others. Existing articles have discussed how to understand the sampling biases in geotagged social media data. However, those discussions are often on an ad-hoc basis and most of them do not examine the sampling biases in a theoretical framework for a more comprehensive understanding.

In response, this research aims to offer several contributions to this widely discussed and urgent research question about the sampling bias of geotagged social media data. First, the research studies the nature of sampling biases in geotagged social media data and proposes a conceptual framework for it. Considering the characteristics of location-based social media data, sampling biases can be found with respect to the spatial, temporal, and demographic dimensions. Therefore, the conceptual framework that represents the sampling bias in geotagged social media data captures the three dimensions and the interactions among them.

Second, the research attempts to not only examine the sampling biases in geotagged social media data but also identify explanatory factors that cause the existence of the sampling biases. Third, the research aims to develop bias correction methods to mitigate the sampling bias in geotagged social media data. This contributes toward improvement of the data quality of geotagged social media big data, which is essential for future research based on it.

Social media has already been one of the primary ways for people to express their opinion. As sampling bias exists, voices from certain groups of people may not be heard. Therefore, the importance of understanding the sampling bias of geotagged social media data should not be ignored. Foreseeable is the increasing prosperity of the research on this topic.

KEYWORDS: *geotagged social media data, Twitter, sampling bias, spatial, temporal, demographic*

Ruowei Liu, Ph.D. Candidate, Department of Geography, University of Georgia, Athens, GA, 30602

Xiaobai (Angela) Yao, Professor, Department of Geography, University of Georgia, Athens, GA, 30602