Male or Female? Evaluating the Gender Representativeness of Twitter

Ruowei Liu[1], Xiaobai (Angela) Yao[1], Chenxiao (Atlas) Guo[2]
[1]Department of Geography, University of Georgia
[2]Department of Geography, University of Wisconsin-Madison

University Consortium for GEOGRAPHIC INFORMATION SCIENCE

UNIVERSITY OF GEORGIA   WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

INTRODUCTION

- Background: Growing interests in location-based social media (LBSM) especially Twitter with many cross-disciplinary research
- Gap: Representational bias of LBSM is under-investigation; Can significantly affect modeling accuracy and potentially lead to unreliable findings
- Objective: Evaluate the representativeness of Twitter data by comparing demographics of active Twitter users with it of local population

RESEARCH QUESTION

1) What is the difference of gender between Twitter user and local population in Clarke County, Georgia at census tract level?
2) How the difference distributes across the county?

DATA

- Twitter Data: Geotagged tweets from December 2017 to April 2018 in Clarke County
- Demographic Data: Estimated male and female population at census tract level in Clarke County (2013-2017 ACS 5-year data)
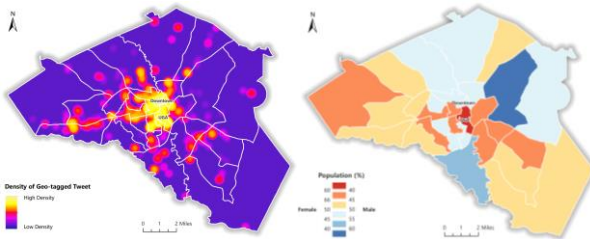

Figure 1. Density of Geotagged Tweets (Left) and Percentage of Male/Female Population (Right)

METHODOLOGY

Key Steps (workflow on the next column):

- Twitter User Gender Inference: Using a face recognition service (Face++ API) to infer gender
- Representativeness Calculation: $r_M = \frac{M_T}{M_{ACS}}$ where r_M is representativeness of male, M_T is percentage of male Twitter users, M_{ACS} is percentage of male population
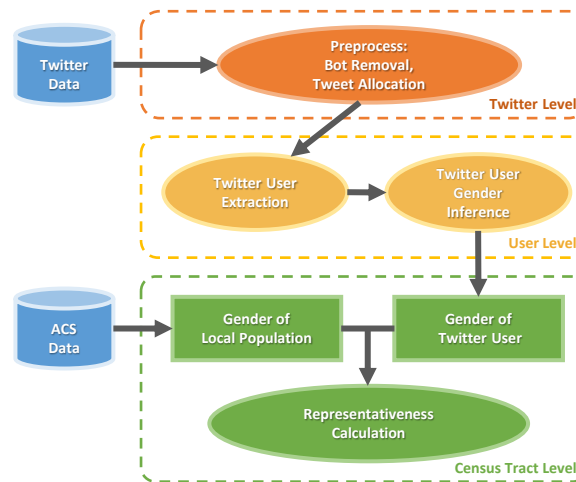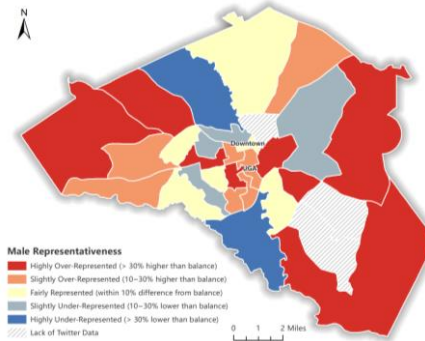

Figure 2. Workflow of This Study

RESULTS

Male Representativeness
- Highly Over-Represented (> 30% higher than balance)
- Slightly Over-Represented (10~30% higher than balance)
- Fairly Represented (within 10% difference from balance)
- Slightly Under-Represented (10~30% lower than balance)
- Highly Under-Represented (> 30% lower than balance)
- Lack of Twitter Data

Figure 3. Spatial Distribution of Male Representativeness

Female Representativeness
- Highly Over-Represented (> 30% higher than balance)
- Slightly Over-Represented (10~30% higher than balance)
- Fairly Represented (within 10% difference from balance)
- Slightly Under-Represented (10~30% lower than balance)
- Highly Under-Represented (> 30% lower than balance)
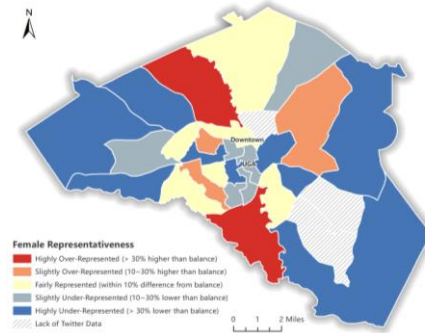- Lack of Twitter Data

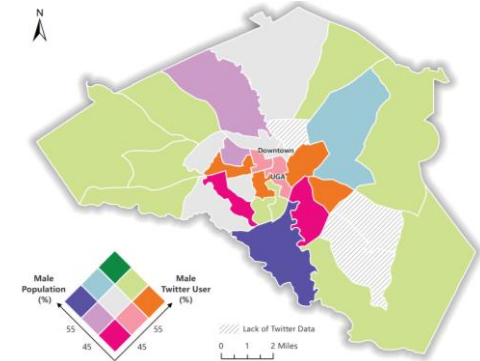Figure 4. Spatial Distribution of Female Representativeness

Figure 5. Spatial Relationship between Twitter Users and Local Population
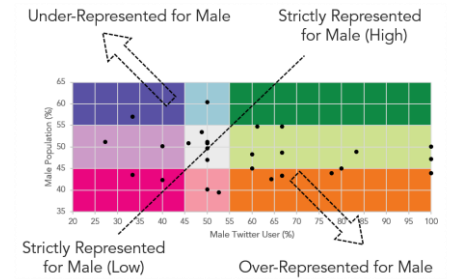

Figure 6. Relationship between Twitter Users and Local Population in Non-spatial Perspective

Spatial Characteristics:

- Male population in downtown and University of Georgia (UGA) are overrepresented on Twitter.
- Either male or female are overrepresented on Twitter in several census tracts in sub-urban areas.
- In the wide rural areas, there are generally more male than female Twitter users.

Non-spatial Characteristics:

- In more than half of the census tracts, male population are overrepresented on Twitter.
- No census tract with both high percentage of male Twitter users and population, but few for female.

REFERENCES

- Goodchild, M. F. (2013). The quality of big (geo) data. Dialogues in Human Geography, 3(3), 280-284.
- Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. Applied Geography, 70, 11-25.
- Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. ICWSM, 11, 5th.
- Yihong Yuan, Guixing Wei & Yongmei Lu (2018) Evaluating gender representativeness of location-based social media: a case study of Weibo, Annals of GIS, 24:3, 163-176.