ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
subject 1, subject 2, subject 3

Keywords:
one, two, optional, optional, optional

Author for correspondence:
C.Cabrera-Arnau
e-mail: C.Cabrera-Arnau@liverpool.ac.uk

An approach to quantifying the extent of bias in aggregated human population data extracted from digital platforms

Carmen Cabrera-Arnau[1], Francisco Rowe[1]

[1]Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Liverpool, United Kingdom.

The abstract text goes here. The abstract text goes here. The abstract text goes here. The abstract text goes here. The abstract text goes here. The abstract text goes here. The abstract text goes here. The abstract text goes here.

THE ROYAL SOCIETY
PUBLISHING

The main document should include:

Title (no more than 150 characters)

Author names and affiliations, and ORCID iDs where available

Abstract (no more than 200 words). (This will be used in reviewer invitation emails so please think about how to describe your work to make it easy for a potential reviewer to determine whether they would be suitable.)

All main manuscript text. Ensure that all figures, tables and any relevant supplementary materials are mentioned within the text References

Acknowledgements and funding statement (ensure that you have included grant numbers and the names of any funding providers)

Tables with captions Figure captions with credits where relevant

1. Introduction

Location data derived from DFs collected via digital technology, has created new opportunities for research, policy and decision making. These data offer high geographic and temporal granularity, extensive coverage and instant information to measure and transform our understanding of human mobility [1]. DF data (DFD) generation expands countries facilitating comparative analyses. Substantively, studies leveraging DFD have contributed to expanding existing theories, developing new explanations, adopting new analytical tools and infrastructures, and advancing new areas of research, such as computational social science and geographic data science [2]. Yet, these data also present major epistemological, methodological and ethical challenges [3].

A key unresolved limitation in the use of DFD is the potential presence of biases relating to its statistical representativeness. Two sources of biases are particularly prominent. First, biases emerge from differences in the access and use of the particular digital technology, such as mobile applications, used to collect data [4]. In the UK, for example, we know that 98% of the adult population have a mobile phone and 92% of this population use a smartphone [5], but a smaller percentage actively use Facebook (70%) or Twitter (23%) [6]. Second, biases can also emerge from differences in the access and usage of digital technologies across population groups. DF-derived mobility data from Twitter, for instance, display a young adult, male and urban user profile (e.g. [7], [8]). Differences in age, income and education have been found in Facebook-derived population counts [9]. As a result, DF-derived mobility data cannot be interpreted directly to provide a reliable estimate of population mobility levels [10]. They can only afford to offer rough signals about mobility patterns (e.g. spatial concentration), trends (e.g. increasing) and changes (e.g. low to high) [11].

Efforts have been made to correct these biases through two general approaches. A first general approach consists in adjusting DF-derived population counts from social media by developing correction factors (e.g. [12], [13]). Correction factors are often estimated as the ratio of active social media users to census population counts by demographic attributes (e.g. age). The principles are similar to survey post-stratification methods i.e. to make DF-derived population counts representative of the census populations. However, a key data requirement of this approach is on having data on population by attribute, but such data are generally unavailable from DFs. Only information on location, time and total active users is recorded. As such, this approach cannot be generalised to different DFD sources and geographical contexts, and when applied on total population counts, biases associated with demographic and socioeconomic user attributes are not corrected (e.g. [14], [15], [16]). A second approach uses a regression modelling approach. Intuitively this approach produces representative population counts by explicitly measuring and removing the sources of biases in the data [17]. This approach has primarily been used in Ecology to obtain representative population distributions of animal species [18], but it has not been used in the context of DFD. In recent work, the PI adopted a similar approach to correct multiple sources of biases in census data to produce bias-adjusted migration estimates [19]. DEBIAS builds on this work to develop a general framework and software package aiming to correct biases in origin-destination mobility counts derived from DFs in the absence of demographic and socioeconomic information on users of digital platforms.

2. Data and methods

(a) Data

(i) Facebook

(ii) Twitter

(iii) Other

(b) Methods

In this section, we present our proposed methodology, which has two primary aims: first, to quantify biases, and second, to identify the characteristics of local populations that increase their likelihood of being underrepresented in digital footprint data (DFD). This methodology serves as a general framework applicable to any digital technology that captures active user counts and operates on data aggregated into spatial and temporal units, aligning well with the structure of many DFD sources available to researchers.

The methodology unfolds in two interconnected stages, each corresponding to our aims. In the first stage, we develop a statistical indicator to quantify the magnitude of bias in each subnational area. This step is crucial for establishing a baseline understanding of bias levels, allowing us to pinpoint regions with significant underrepresentation. In the second stage, we analyse the association of these biases with demographic, socioeconomic, and geographic attributes at the area level. This analysis yields insights into the underlying characteristics contributing to disparities in the level of bias across areas, thereby addressing our second methodological aim.

(i) Bias indicator

First, we define a metric to quantify the magnitude of bias in each subnational area. We will do this by estimating the extent of population coverage of the digital technology used to collect the DFD (e.g. Facebook app). This will be computed as the ratio of the user population of the digital technology (P_i^D) to the total local population of an area (P_i). Formally, the coverage c_i is given by:

$$c_i = \frac{P_i^D}{P_i} \times 100, \tag{2.1}$$

where D identifies a given digital technology, and i denotes each subnational area. The ratio is assumed to take values between 0 and 100, with the latter representing full population coverage. The ratio can only take values greater than 1 if users have multiple accounts exceeding the total local population of an area.

We then define the size of bias e_i as:

$$e_i = 100 - c_i \tag{2.2}$$

in which case, $e_i = 0$ will indicate full population coverage or no bias. We will use this indicator to examine the magnitude and spatial distribution of DFD bias.

(ii) Machine learning

Second, we will seek to understand how DF coverage biases are associated with area-level demographic and socioeconomic attributes. To what extent different demographic and socioeconomic groups are represented in DFD? And how do these vary geographically and across digital platform? We will assess these questions by measuring the area-level association between our coverage indicator and key demographic and socioeconomic attributes. We will use a random forest to model our coverage indicator as a function of demographic and socioeconomic attributes. The outcomes will identify the most important area-level demographic and socioeconomic features to predict the coverage bias of a given digital technology. We will use this information to inform our models in WP-II.

eXtreme Gradient Boosting (XGBoost) is an efficient and scalable implementation of gradient boosting framework by [**? ?**].

3. Results

(a) Measuring the extent of biases

(b) Assessing the extent of biases in digital trace data

(c) Explaining biases

4. Discussion

5. Conclusion

Ethics. Please provide details on the ethics.

Data Accessibility. Please provide details on the data availability.

Authors' Contributions. Please provide details of author contributions here.

Competing Interests. Please declare any conflict of interest here.

Funding. Please provide details on funding

Disclaimer. Please provide disclaimer text here.

Acknowledgements. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

References

1. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, Nadai MD, Letouzé E, Salah AA, Benjamins R, Cattuto C, Colizza V, de Cordes N, Fraiberger SP, Koebe T, Lehmann S, Murillo J, Pentland A, Pham PN, Pivetta F, Saramäki J, Scarpino SV, Tizzoni M, Verhulst S, Vinck P. 2020 Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances* **6**, eabc0764.
2. Pappalardo L, Manley E, Sekara V, Alessandretti L. 2023 Future directions in human mobility science. *Nature Computational Science* **3**, 588–600.
3. Rowe F. 2023 pp. 42 – 47. In *9.: Big data*, pp. 42 – 47. Cheltenham, UK: Edward Elgar Publishing.
4. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO. 2013 The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface* **10**, 20120986.
5. Ofcom. 2023 Communications Market Report 2023. Accessed: Nov 2023.
6. Statista. 2024 Social Media & User-Generated Content - Market Overview. Accessed: 2024-11-14.
7. Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist J. 2021 Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media* **5**, 554–557.
8. Sloan L, Morgan J, Housley W, Williams M, Edwards A, Burnap P, Rana O. 2013 Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online* **18**, 74–84.
9. Ribeiro FN, Benevenuto F, Zagheni E. 2020 How Biased is the Population of Facebook Users? Comparing the Demographics of Facebook Users with Census Data to Generate Correction Factors. In *Proceedings of the 12th ACM Conference on Web Science* WebSci '20 p. 325334 New York, NY, USA. Association for Computing Machinery.
10. Cesare N, Lee H, McCormick T, Spiro E, Zagheni E. 2018 Promises and Pitfalls of Using Digital Traces for Demographic Research. *Demography* **55**, 1979–1999.
11. Rowe F, Neville R, González-Leonardo M. 2022 Sensing Population Displacement from Ukraine Using Facebook Data: Potential Impacts and Settlement Areas. *OSF Preprints*. Submitted.
12. Yildiz D, Holland JA, Vitali A, Munson J, Tinati R. 2017 Using Twitter data for demographic research. *Demographic Research* **37**, 1477–1514.

13. Hsiao Y, Fiorio L, Wakefield J, Zagheni E. 2024 Modeling the Bias of Digital Data: An Approach to Combining Digital With Official Statistics to Estimate and Predict Migration Trends. *Sociological Methods & Research* **53**, 1905–1943.
14. Rodriguez-Carrion A, Garcia-Rubio C, Campo C. 2018 Detecting and Reducing Biases in Cellular-Based Mobility Data Sets. *Entropy* **20**.
15. Schlosser F, Sekara V, Brockmann D, Garcia-Herranz M. 2021 Biases in human mobility data impact epidemic modeling. .
16. Chankyung Pak KC, Thorson K. 2022 Correcting Sample Selection Bias of Historical Digital Trace Data: Inverse Probability Weighting (IPW) and Type II Tobit Model. *Communication Methods and Measures* **16**, 134–155.
17. Kramer-Schadt S, Niedballa J, Pilgrim JD, Schröder B, Lindenborn J, Reinfelder V, Stillfried M, Heckmann I, Scharf AK, Augeri DM, Cheyne SM, Hearn AJ, Ross J, Macdonald DW, Mathai J, Eaton J, Marshall AJ, Semiadi G, Rustam R, Bernard H, Alfred R, Samejima H, Duckworth JW, Breitenmoser-Wuersten C, Belant JL, Hofer H, Wilting A. 2013 The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions* **19**, 1366–1379.
18. Zizka A, Antonelli A, Silvestro D. 2021 sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography* **44**, 25–32.
19. Aparicio Castro A, Winiowski A, Rowe F. 2023 A Bayesian approach to estimate annual bilateral migration flows for South America using census data. *Journal of the Royal Statistical Society Series A: Statistics in Society* **187**, 410–435.