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 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 (FR)

Technological advances in computational power, storage and digital network platforms have unleashed a data revolution producing large trails of digital trace data containing location information. These data have revolutionised research and business activities offering novel opportunities to understand human behaviour and processes [1]. Digital trace data offer high spatial granularity, geographic coverage, high temporal frequency and instant information to capture and understand human activities at unprecedentedly high resolution and scale, and produce actionable intelligence in real time to support rapid decision making. Digital trace data have been used for a range of applications, such monitoring footfall changes [REF], inferring mobility signatures [REF], sensing land use patterns [REF], predicting socioeconomic levels [REF], defining urban extents [REF] and estimating population displacement [REF].

However, the use of digital trace data present major epistemological, methodological and ethical challenges [1]. A key unresolved challenge is the potential presence of biases in digital trace data to compromise their statistical representativeness and perpetuate social injustice [REF]. Biases reflect societal digital and socioeconomic inequalities. Biases emerge from differences in the access and use of the particular digital technology used to collect data, such as mobile phone applications [2]. Only a fraction of the population in an area owns a mobile phone, and even an smaller share actively use a specific mobile phone app. In the UK, for example, 98% of the adult population have a mobile phone and 92% of this population use a smartphone [3], but a smaller percentage actively use Facebook (70%) or Twitter (23%) [4]. Additionally, biases emerge from differences in the access and use of digital technology across population subgroups according to their socioeconomic and demographic profile. For instance, wealthy and urban populations tend to be over-represented in mobile phone data [REF] and of digital social platforms, such as Facebook [REF] and Twitter/X [REF].

The use of biased digital trace data can thus have major practical and societal implications.

As a result, population statistics derived from digital trace data cannot provide population-level representation. They can only offer rough signals about the spatial distribution of (e.g. spatial concentration), trends (e.g. increasing) and changes (e.g. low to high) in populations [5].

and amplify socioeconomic disparities

Gaps:

Aim: We propose an approach to measure biases -

Research questions:

Contribution:

Structure:

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. [6], [7]). 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. [8], [9], [10]). 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 [11]. This approach has primarily been used in Ecology to obtain representative population distributions of animal species [12], 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 [13]. 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 (CCA)

(i) Facebook

We use anonymised aggregate location data from Facebook app users who have the location services setting turned on on their smartphone for the UK, covering March 2021, the month when the 2021 UK Census was carried out. We use the Facebook Population dataset created by Meta and accessed through their Data for Good Initiative (https://dataforgood.facebook.com). Prior to releasing the datasets, Meta ensures privacy and anonymity by removing personal information and applying privacy-preserving techniques [@Maas19]. Small-count dropping is one of these techniques. A data entry is removed if the population or movement count for an area is lower than 10. The removal of small counts may mean that population counts in small sparsely populated areas are not captured. A second technique consists in adding a small undisclosed amount of random noise to ensure that it is not possible to ascertain precise, true counts for sparsely populated locations. Third, spatial smoothing using inverse distance-weighted averaging is also applied is applied to produce a smooth population count surface.

The Facebook Population dataset offers information on the number of active Facebook users in a spatial unit at a given point in time. The data is temporally aggregated into three daily 8-hour time windows (i.e. 00:00-08:00, 08:00-16:00 and 16:00- 00:00). In this work, we are interested in capturing resident population, so we consider only data corresponding to the time window corresponding to the night-time hours (00:00-08:00).

The Facebook Population dataset is spatially aggregated according to the Bing Maps Tile System developed by Microsoft (Microsoft). The Tile System is a geospatial indexing system that partitions the world into tile cells in a hierarchical way, comprising 23 different levels of detail (Microsoft). At the lowest level of detail (Level 1), the world is divided into four tiles with a coarse spatial resolution. At each successive level, the resolution increases by a factor of two. The data that we used are spatially aggregated into Bing tile levels 13. That is about 4.9 x 4.9km at the Equator [@Maas19]. In the next steps, we compare the Facebook Population data with UK census data. To facilitate this comparison, we aggregate the data into administrative units, specifically UK Local Authority Districts (LADs).

(ii) Twitter

We use an anonymised, openly available, analysis-ready dataset of active Twitter users in the UK. The data is derived from X (previously Twitter) in the form of monthly active user counts residing across the UK geography. The dataset is based on tweets from UK users [@wang2022] collected via the Twitter Academic API. These tweets are either geolocated at the time of posting or manually geocoded using a bounding box provided by the Twitter Academic API, based on the IP address of the posting device. The full dataset includes 161 million tweets from February 2019 to December 2021; however, we focus on data from March 2021 to align with the 2021 UK Census. Users' Local Authority District (LAD) of residence is identified using a frequency-based home-location algorithm. Further details on the dataset's methodology can be found in [@wang2022].

While the X Academic API is no longer available to download data, but existing and future projects offer an opportunity for research based on X data. Global repositories of historical geolocated tweet data are accessed through the Internet Archive (1996) and Harvard Geotweet Archive (https://gis.harvard.edu/data). Despite these limitations, we consider X data as it remains a key source of historical digital trace data.

(iii) Multi-app GPS data

The data used in this study were sourced by a data analytics company that collects GPS location data from around 26% of smartphones in the UK. The raw data is collected for individual anonymised devices, from numerous smartphone applications where the users have explicitly granted location-sharing permissions. The full dataset covers 7 days corresponding to the first week of April for the UK, including X GPS records and X unique devices. While the dates covered by dataset do not exactly coincide with the 2021 UK Census dates, the alignment is very close.

We process the data to estimate users' place of residence by assuming that the residence of a device owner corresponds to the location with the highest number of GPS records during night hours (11 PM – 7 AM). For a location to be classified as a residence, it must account for more than 50% of recorded nighttime locations and be visited at least twice. To ensure consistency in our analysis when comparing with other data sources, we aggregate these residence locations at the Local Authority District (LAD) level.

(b) Methods

In this section, we outline our proposed methodology, which consists of two interconnected stages aimed at quantifying two types of biases: coverage biases and representational biases. Coverage biases relate to the sample size of the dataset and refer to the proportion of the population covered in the dataset. Representational biases, arise from the demographic and socioeconomic characteristics of the users who generate the digital trace data through specific technologies.

In the first stage of our methodology, we focus on quantifying coverage biases by examining the variations in coverage across different spatial units. We leverage the spatial granularity of digital trace data to analyse coverage biases at more localised spatial scales. This allows us to identify the extent to which different regions are represented within the datasets, revealing any potential underrepresentation or overrepresentation in specific locations.

In the second stage, we focus on quantifying representational biases. To do this, we leverage the spatial heterogeneity of coverage biases and model this variation in terms of demographic and socioeconomic variables that characterise local populations. This analysis allows us to identify which specific demogrpahic and socioeconomic population attributes, such as average income, education level or age composition, are more likely to be associated with higher values of coverage bias, thus highlighting which population groups tend to be underrepresented in different sources of digital trace data.

(i) Measuring coverage bias (CCA)

We define a metric to quantify the magnitude of coverage bias in each subnational area. This metric is based on the population coverage of the dataset, which we compute as the ratio of the population captured (sample size) by dataset D, denoted as 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 dataset, and i denotes each subnational area. The resulting ratio c_i is assumed to take values between 0 and 100, with 100 representing full population coverage. If users have multiple accounts, the ratio can exceed 100, since the total sample size could be greater than the local population of area i.

We then define the size of bias e_i as:

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

In this case, a value of $e_i = 0$ indicates a lack of coverage bias, which corresponds to full population coverage ($c_i = 100$). We use this bias indicator to analysis the magnitude and spatial distribution of coverage bias across multiple sources of digital trace data.

(ii) Assessing the driving factors of bias (FR)

We seek to understand the association between the size of bias and 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) Contextualising coverage bias in digital trace data

As digital trace data becomes increasingly accessible, it opens up new avenues for studying human behaviors with remarkable temporal and spatial precision, extensive geographic coverage, and near real-time access. However, a critical challenge remains: the potential presence of biases that can undermine the validity of the data to deliver statistically representative evidence. In particular, we first focus on the coverage bias, which refers to the proportion of the total population represented in the dataset.

In this section, we quantify coverage bias across multiple sources of digital trace data and contextualize our findings by comparing them with coverage bias observed in various traditional data sources. Particularly, in Figure X, we compare multiple sources of digital trace data explored in this paper with traditional survey-based datasets, particularly key UK surveys available through the UK Data Service []. On the x-axis, we represent two variables: at the top, the sample size of the dataset, expressed as the number of respondents or subjects per 1,000 people, which reflects the level of coverage of the dataset; and at the bottom, a measure of coverage bias as defined in equation X. The figure highlights the remarkable ability of digital trace data to capture a larger number of "respondents" compared to traditional surveys, thanks to the automated, passive nature of data collection on digital platforms. This contrasts with the manual recruitment and data collection processes required for surveys. As a result, the coverage bias is generally lower for digital trace data, highlighting its potential to inform comprehensive empirical analyses.
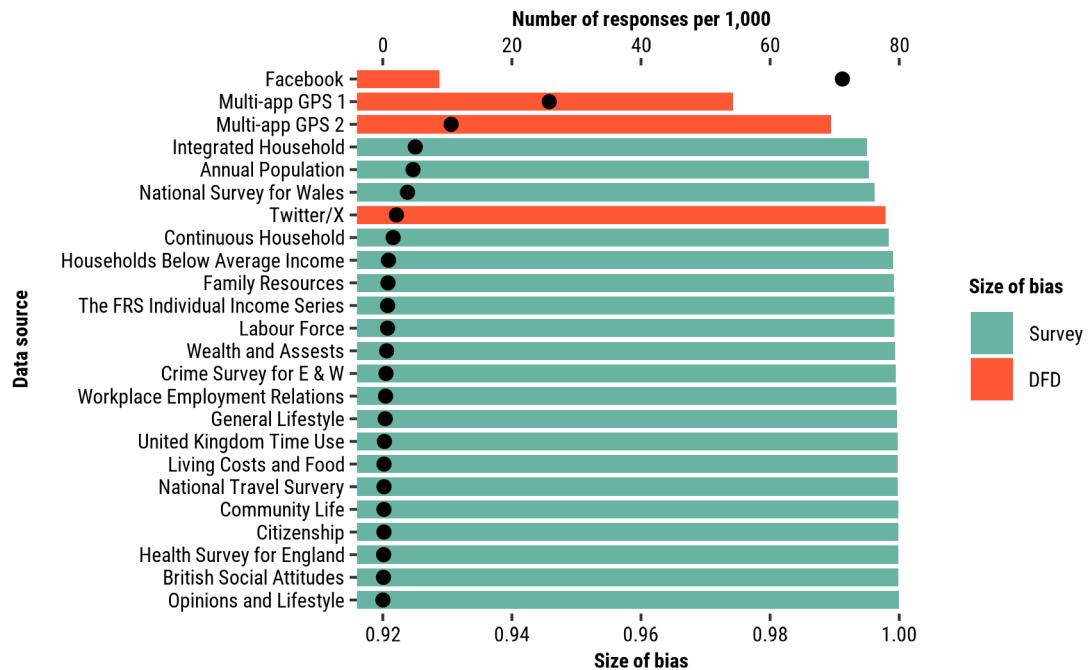
While the findings in Figure X suggest the advantages of digital trace data compared to traditional data sources, high population coverage alone does not ensure representativenness. Unlike surveys, where data collection follows a carefully designed sampling strategy to ensure statistical representativeness, digital trace data are generated passively, without such protocols in place.

(b) The spatial distribution of coverage bias (CCA)

Next, we leverage the spatial granularity of the digital trace data sources considered here to analyse coverage biases at more localised spatial scales. This allows us to identify the extent to which different geographies are represented within the datasets, revealing any potential underrepresentation or overrepresentation in specific locations. - CCA continue HERE

Size of bias by data source
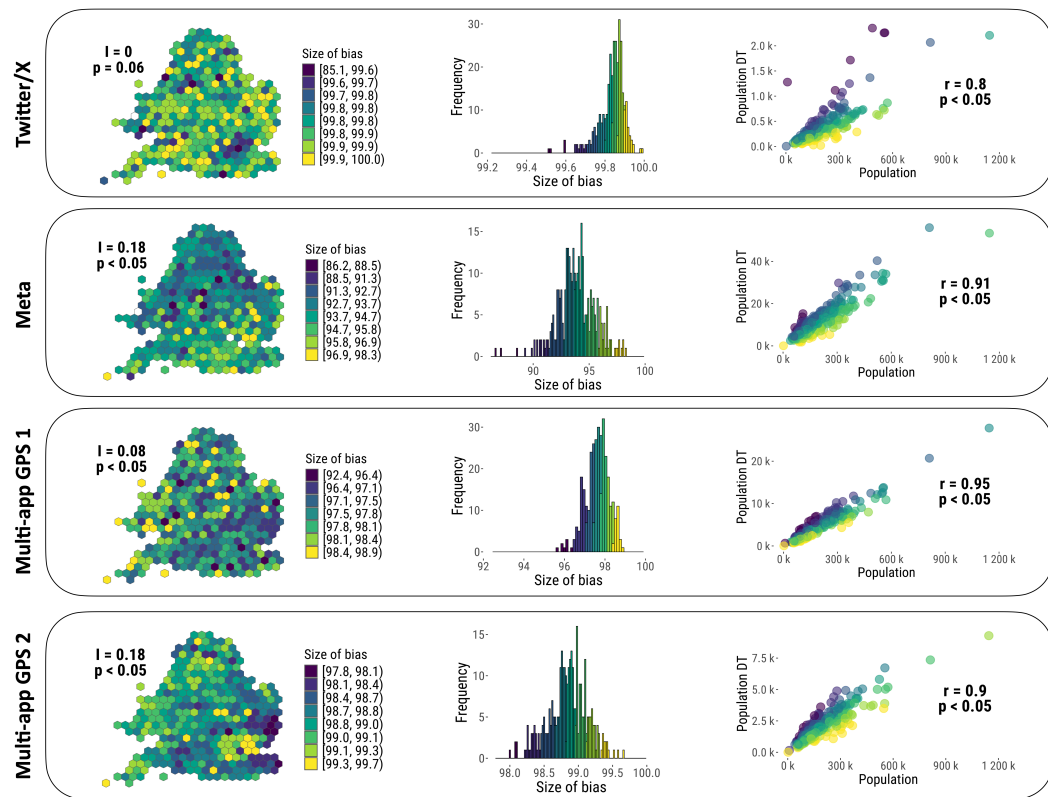
(c) Explaining biases (FR)

4. Discussion (FR)

5. Conclusion (CCA)

References

1. Rowe F. 2023 pp. 42 – 47. In *9.: Big data*, pp. 42 – 47. Cheltenham, UK: Edward Elgar Publishing.
2. 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.
3. Ofcom. 2023 Communications Market Report 2023. Accessed: Nov 2023.

4. Statista. 2024 Social Media & User-Generated Content - Market Overview. Accessed: 2024-11-14.

5. Rowe F, Neville R, González-Leonardo M. 2022 Sensing Population Displacement from Ukraine Using Facebook Data: Potential Impacts and Settlement Areas. *OSF Preprints*. Submitted.

6. Yildiz D, Holland JA, Vitali A, Munson J, Tinati R. 2017 Using Twitter data for demographic research. *Demographic Research* **37**, 1477–1514.

7. 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.

8. Rodriguez-Carrion A, Garcia-Rubio C, Campo C. 2018 Detecting and Reducing Biases in Cellular-Based Mobility Data Sets. *Entropy* **20**.

9. Schlosser F, Sekara V, Brockmann D, Garcia-Herranz M. 2021 Biases in human mobility data impact epidemic modeling. .

10. 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.

11. 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.

12. Zizka A, Antonelli A, Silvestro D. 2021 sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography* **44**, 25–32.

13. 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.