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:
Carmen Cabrera
e-mail: C.Cabrera@liverpool.ac.uk

A systematic machine learning approach to quantifying the spatial extent of bias in human population data from mobile phones

Carmen Cabrera[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 (FR)

Traditional data streams, such as the census and surveys have been the primary official source to provide a comprehensive representation of national populations in countries worldwide. However, fast-paced societal changes and emergency disasters, such as climate-induced hazards and COVID-19 have tested and accentuated weaknesses in traditional data systems [1]. Traditional data systems often provide data in infrequent and coarse temporal and geographical resolutions [2]. Generally they are expensive to maintain and operate, and are slow taking months or years since they data are collected to their release [2]. Data collection from climate- or conflict-impacted areas is generally unfeasible because of restrictions due to high levels of insecurity and risk [3]. Yet, fast-paced societal changes require high frequency, granular and up-to-date information to support real-time planning, policy and decision making.

At the same time, we have seen the confluence of two diverging trends in data availability. On the one hand, growing evidence of declining survey response rates across many countries over the last 20 years is accumulating **[REF]**. Dwindling numbers in surveys can represent distorted picture of society **[REF]**. On the other hand, significant advances in sensor technology, computational power, storage and digital network platforms have unleashed a data revolution producing large trails of digital trace data **[REF]**. These data are now routinely collected and stored. They offer spatially granular, frequent and instant information to capture and understand human activities at unprecedentedly high resolution and scale, with the potential to produce real-time actionable intelligence to support decision making **[REF]**. Hence, national statistical offices are actively seeking to integrate these data into their national data infrastructure **[REF]**.

Mobile phone data (MPD) collected via GPS- and IP-based technology have become a prominent source of nontraditional data to monitor population changes. Increasing usage of mobile services on smartphones and wearable devices have resulted in the generation of large volumes of geospatial data, offering novel opportunities to advance understanding of spatial human behaviour, and thus revolutionise research, business and government decision making and practices [2]. MPD are now a core component of the digital economy, creating new market opportunities for data intelligence businesses, such as Cuebiq/Spectus, Safegraph and Locomizer. They have been used to create critical evidence to support policy making, prominently during the COVID-19 pandemic. In research, MPD have been used to develop innovative approach to infer mode of transport [REF], monitor footfall changes [REF], profile daily mobility signatures [REF], sense land use patterns [REF], predict socioeconomic levels [REF], define urban extents [REF], quantify tourism activity [REF] and estimate migration and population displacement [REF].

However, the use of MPD present major epistemological, methodological and ethical challenges [2]. A key unresolved challenge is potential biases in MPAD compromising 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 mobile phone applications used to collect MPD [4]. Only a fraction of the population in a geographical area owns a smartphone, and even an smaller share actively uses 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 [5], but a smaller percentage actively use Facebook (70%) or Twitter (23%) [6]. Additionally, biases emerge from differences in the access and use of digital technology across population subgroups reflecting socioeconomic and demographic disparities. For instance, wealthy, young and urban populations generally have greater access and more intensively use of mobile phone applications, and therefore tend to be over-represented in MPD [REF].

The use of biased MPD can thus have major practical and societal implications. If used uncorrected, MPD reproduce selective patterns of smartphone ownership and application usage, rendering inaccurate or distorted representations of human population activity. Such representations disproportionately reflect behaviours of younger, urban and higher-income users while underrepresenting marginalised or less-connected groups. Distorted representations based on biased MPD can thus misguide decision making, policy and planning interventions, and thus amplify existing socio-economic disparities. In practice, existing applications of MPD often use

uncorrected population statistics derived from MPD and have thus been constrained to offer a partial picture for a limited segment of the overall population. Such data can only afford to provide rough signals about the spatial distribution of (e.g. spatial concentration), trends (e.g. increasing) and changes (e.g. low to high) in populations [7]. They have cannot provide a full representation of the overall population.

Efforts have been made to measure and assess biases in aggregate population counts from digital data sources. Existing analyses typically measure the extent of bias measuring the system-wide difference in the representation of population counts from digital platforms and censuses. To estimate the representation of digital data sources, the penetration rate is computed as the active user base of a digital platform over the census resident population. Existing analyses have thus been able to established systematic gender, age and socio-economic biases in population data obtained via API (or Application Programming Interface) from social media platforms, such as Facebook and Twitter/X. However, this approach requires information on the demographic and socio-economic attributes of the collected sample and has focused on estimating biases at the country level. Yet, these attributes are rarely available from mobile phone network operators. In practice, they are generally unavailable for MPD, and biases may vary widely across subnational areas. What is missing is an systematic approach to measure biases in population counts from digital platforms, when population attributes are unknown, and quantify the geographic variability in the extent of biases in these data.

To address this gap, this paper aims to establish a standardised approach to empirically measure the extent of biases in population data derived from digital platforms, and identify their key underlying contextual factors across subnational areas. We seek to address the following research questions:

- What is the comparative extent of population coverage of digital sources relative to widely-used traditional surveys?
- How systematic is the association between larger biases and the over-representation of rural, more deprived, child and elderly populations?
- To what extent, are population data assembled from multiple applications versus single applications associated with lower bias?

Our approach proposes a statistical indicator of population coverage to measure the geographic extent of bias, and uses explainable machine learning to identify key contextual factors contributing to spatial variations in the extent of bias. Biases in digital trace data can emerge from multiple sources, such as algorithmic changes, device duplication and geographic location accuracy [REF]. We do not intend to identify these individual sources of error. We focus on quantifying the extent of "cumulative'' bias; that is, the resulting bias from the accumulation of these error sources. We use data collected from single and multiple mobile phone apps, and compare their results. As outlined above, we test the extent to which biases can be mitigated by leveraging information from multiple apps encompassing a more diverse user population. Specifically, we use two single-app (i.e. Facebook and Twitter/X) and two multi-app providers (i.e. Locomizer and a European provider). We focus on the use of aggregated population counts as this has become a common ethical and privacy-preserving practice for companies to provide access to highly sensitive data for social good.

Our study makes two key contributions. * Methodological contribution i.e. what we hope to achieve with our approach / quality assessment framework ideas + start setting standards of good practice in the use of MPD.
* Substantive contribution - systematic evidence identifying key predictor of biases + do we find evidence of lower biases / greater population coverage for multi-app better than single app?

2. Data and methods

[NOTE: I think that we need a paragraph describing and providing an overview of the methodological strategy, including both data and methods. Two points are particularly crucial to connect: (1) The use of data from March 2021 for our assessment against census data; and (2) the use of multiple data sources. We need to describe the idea of single- and multiple-sourced app data. I wonder if we should include a table listing their general attributes: advantages and limitations in terms of their temporal and spatial coverage and resolution. This may not be the place for the table but would be good to consider for the book if we compared GPS data to other sources.]

(a) Data (CC)

(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 [8] . 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).

Spatially, the Facebook Population dataset is 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 [8].

We process Facebook Population data to enable comparison with UK census data. Specifically, we take the average of daily Facebook Population data over March 2021, the census month, and aggregate it into UK Local Authority Districts (LADs) to align with the census data. This approach is used to generate the figures and results in the body paper. In the Supplementary Information, we test alternative approaches, including averaging over a single week in March 2021, or performing the spatial aggregation before temporal averaging. Our findings indicate that the results remain robust regardless of the chosen approach

(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 [9] 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 [9] .

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: source 1

We sourced data from 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 close.

We process the data to estimate users' place of residence based on a commonly used rule-based classification (e.g. [9, **(author?)** [10]]), which assumes that the residence of a device owner corresponds to the location with the highest number of GPS records during night hours (7 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 during the period of study. To ensure consistency when comparing with other data sources, we aggregate these residence locations at the Local Authority District (LAD) level.

(iv) Multi-app GPS data: source 2

Our analysis includes a second source of analysis-ready GPS location data, which is openly-available on GitHub (https://t.ly/dzlzB)). This dataset has already been processed to identify the home location of users accoring to the methodology described in [10]. The raw data is collected by a UK-based data service company, which licenses mobile GPS data from 200 smartphone apps and applies pre-processing methods to ensure user privacy and anonymity. The full dataset covers the UK in November 2021. While this period does not exactly coincide with the 2021 UK Census, the difference of less than a year is considered sufficiently close for our analysis.

To ensure consistency across datasets, we further process the data by aggregating it spatially from the Middle Layer Super Output Area (MSOA) level to the Local Authority District Level (LAD).

(b) Methods

Our proposed methodology consists of two 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.

The first stage of our methodology seeks to quantify 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.

The second stage seeks to quantify 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) Identifying the key predictors 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) Varying extent of bias across data sources

As digital trace data becomes increasingly accessible, it opens up new avenues for studying human behaviours with remarkable temporal and spatial precision, extensive geographic coverage, and near real-time access. However, the potential presence of biases can undermine the validity of the data to deliver statistically representative evidence.

In this section, we focus on quantifying the biases in multiple sources of digital trace data that arise due to the extent of population coverage, i.e. the proportion of the total population captured in the dataset. In Figure , we contextualise these findings by comparing them with various traditional datasets, particularly key UK surveys available through the UK Data Service [11]. 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 population coverage of the dataset; and at the bottom, a measure of bias in terms of this coverage, as defined in equation

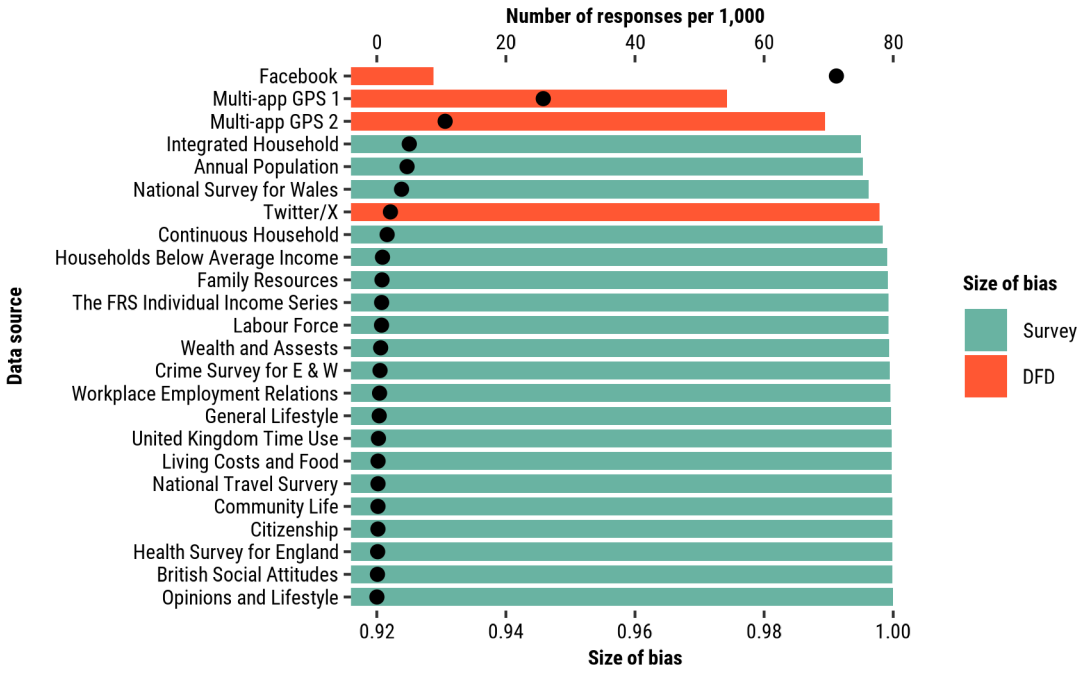


Figure 1. Size of bias and population coverage (per 1,000 population) by data source.

2.2. The figure highlights the remarkable ability of digital trace data to capture a larger share of the total population 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 size of bias is generally lower for digital trace data, highlighting its potential to inform comprehensive empirical analyses.

While the findings in Figure 1 demonstrate the potential of digital trace data compared to traditional data sources, high population coverage alone does not ensure the data is representative of different population groups.

In surveys, specific strategies are usually implemented during the data generation process to improve the statistical representativeness of the sample. For example, sampling techniques such as stratified sampling or cluster sampling can be applied so that the sample reflects the broader population of interest. After sampling, if certain groups remain under-represented, responses can be adjusted using post-stratification techniques. However, even when these strategies are applied, there is no guarantee that the survey will be fully representative of the broader population of interest [12]. This is because representativeness can only be achieved with respect to a finite set of attributes (e.g. age, gender, income levels, location, etc.). Ensuring perfect representativeness would only be possible either by surveying the whole population.

With digital trace data, achieving statistical representativeness is even more challenging. Unlike survey data, which is actively collected using structured sampling methods, digital trace data is generated passively as a byproduct of online interactions, transactions, or device usage, without any control over who is included in the dataset. Furthermore, by the time this data reaches researchers or analysts, it is often anonymised, and does not contain demographic identifiers. As a result, it is not possible to apply the standard post-stratification weighting techniques that are typically used to adjust survey or census data for improved representativeness.

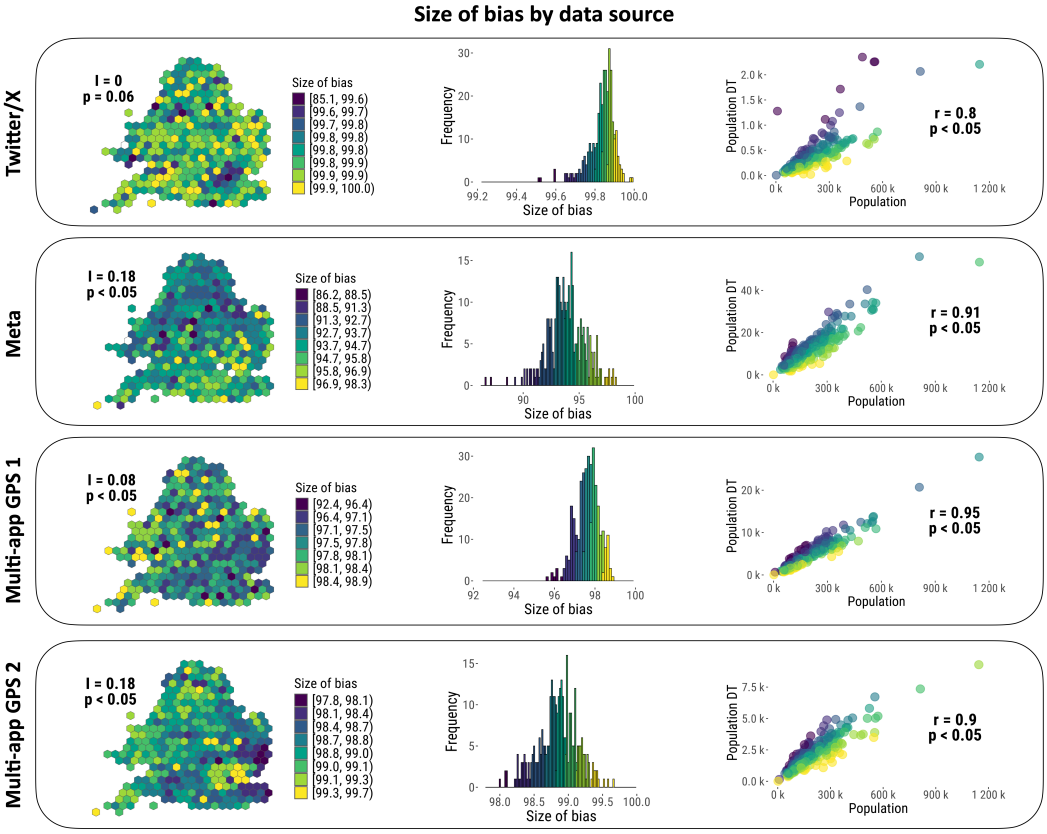
We argue that, even though we do not always have specific demographic information of the individuals captured through digital trace data, we can infer some of these characteristics by

leveraging the spatio-temporal granularity of digital trace data. We argue that this is a necessary first step to understand which population groups might be under or over-represented in different sources of digital trace data. This information is necessary to later adjust the data so that it is more representative of the population of interest.

(b) The spatial distribution of biases (CC)

Next, we take advantage of the detailed geographic information in the digital trace data to analyse bias at smaller, more localised spatial levels. This helps us understand how well different geographic areas are represented in the datasets. Since local populations vary in their socioeconomic characteristics, we can use the degree of bias at these smaller scales in the next step of our analysis, to determine which population attributes are most associated with underrepresentation in the data.

Figure **??** shows the geographic variation of the size of bias at the Local Authority District (LAD) level. Each row in the figure corresponds to each of the digital trace datasets analysed here. Within each row, we include: i) an hexagonal cartogram for the size of bias in each LAD, representing the LADs as hexagons of equal size to simplify the visualisation while maintaining relative positions; with this cartogram, we report Moran's I as a measure of spatial autocorrelation and its associated p-value, ii) a histogram of the size of bias, showing the distribution of values across LADs, iii) a scatter plot of the population covered by the digital trace data vs. the actual population of each LAD; with the scatter plot, we include the Pearson correlation coefficient and its associated p-value.



Examining the spatial variation in bias size, we observe distinct patterns across the DT datasets considered. These varied spatial patterns likely stem from differences in the demographic composition of users for each technology. Factors such as age, socioeconomic status, digital literacy, and regional preferences for certain platforms or devices may contribute to these

variations. Bias tends to display stronger spatial patterns for Meta data and the second source of multi-app GPS data, with lower bias in the North of England and Wales. In contrast, Twitter/X data and the first souce of multi-app GPS data follow more mixed patterns, as demonstrated by the values of Moran's I closer to zero. Twitter/X data generally exhibits high bias, except in London, the South East, and a few isolated areas. Similarly, bias in first source of multi-app data tends to be lower in the South and South East.

Turning to the histograms, we observe that bias size is highest for Twitter/X data, with all values exceeding 99.5 except for a single outlier, the City of London. This outlier likely arises due to the unique demographic and occupational characteristics of the area. While relatively few people reside in the City of London, it hosts a large number of workers, including temporary professionals, who may be staying in hotels. The home-detection algorithm in [9] used to generate the Twitter/X data used here might classify the workplace or temporary accommodations of City of London workers as their primary residences, leading to an anomalously low bias measurement. Following Twitter/X, the second source of multi-app GPS data exhibits the next highest bias values. In contrast, the first source of multi-app data shows lower bias, while Meta data has the lowest overall bias. Notably, Meta data also displays the widest distribution of bias values, indicating greater variability across different locations.

The scatter plots show a high linear correlation between the population covered by the digital trace data and the actual population of each LAD, as demonstrated by the Pearson coefficient, all above 0.8. This suggests that, on average, the actual population in the LADs is not an indicator of the size of bias in DT data, as the population coverage c_i remains the consistent regardless of P_i . This could be a result of the fact that the biases in the data are not driven by the number of people, but rather by other their demographic characteristics such as age, income or educational level. In the next section, we explore the variability of demographic attributes of local populations as possible determinants of the size of bias.

(c) Explaining biases (FR)

4. Discussion (FR)

5. Conclusion (CC)

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. Green M, Pollock FD, Rowe F. 2021 pp. 423–429. In *New Forms of Data and New Forms of Opportunities to Monitor and Tackle a Pandemic*, pp. 423–429. Springer International Publishing.
2. Rowe F. 2023 pp. 42 – 47. In *9.: Big data*, pp. 42 – 47. Cheltenham, UK: Edward Elgar Publishing.
3. Iradukunda R, Rowe F, Pietrostefani E. 2025 Producing population-level estimates of internal displacement in Ukraine using GPS mobile phone data. .

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. Rowe F, Neville R, González-Leonardo M. 2022 Sensing Population Displacement from Ukraine Using Facebook Data: Potential Impacts and Settlement Areas. *OSF Preprints*. Submitted.
8. Maas P. 2019 Facebook Disaster Maps. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
9. Wang Y, Zhong C, Gao Q, Cabrera-Arnau C. 2022 Understanding internal migration in the UK before and during the COVID-19 pandemic using twitter data. *Urban Informatics* **1**.
10. Zhong C, Sari Aslam N, Wang Y, Zhou Z, Enaya A. 2024 Anonymised human location data for urban mobility research. UCL Centre for Advanced Spatial Analysis - Working Papers Series - Paper 240.
11. UK Data Service UK surveys - UK Data Service. https://ukdataservice.ac.uk/help/data-types/uk-surveys/. [Accessed 17-02-2025].
12. Cochran W. 1977 *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley.