ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
Mobile phone data, Human mobility, Explainable AI, Spatial analysis

Keywords:
Mobile phone data, Human mobility, Explainable AI, Spatial analysis

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

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 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 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

We propose a systematic framework to measure and explain biases in population count data derived from mobile phones (MPs). We use four datasets to illustrate this framework, collected in

Data Source	Type	Form of data collection	Finest temporal resolution	Temporal coverage	Finest spatial resolution	Access method	Free at time of access
Facebook Population	Single app	GPS from app users with location services enabled	8-hour windows	March 2021	Bing Tiles level 13	Restricted access via Meta Data for Good	Yes
Twitter (X)	Single app	Geotags and IP-based location via Academic API (pre-processed)	Month	March 2021	Local Authority District	Open access via GitHub (pre-processed)	Yes
Multi-app, Source 1	Multi app	GPS data from multiple apps	Second	First week, April 2021	GPS coordinates	Proprietary, from analytics company (not public)	No
Multi-app, Source 2	Multi app	GPS data from multiple apps (pre-processed)	Averaged over a month	November 2021	MSOA	Open access via GitHub (pre-processed)	Yes

Figure 1. Table 1. Summary description of data sources derived from mobile phone apps used in the article.

or around March 2021 to align as closely as possible with the dates of the most recent census in the area of study to enable direct and temporally consistent comparisons. We focus on aggregated population counts, which are commonly used in mobility research, as a privacy-preserving and ethically responsible data format. The datasets include sources derived from a single MP application (Meta and Twitter/X) as well as from multiple MP applications, each capturing distinct user groups through different data generation mechanisms. These differences allow us to assess how source characteristics influence population coverage and representativeness. The multi-application sources are referred to as Multi-app Source 1, whose provider name cannot be disclosed due to a non-disclosure agreement, and Multi-app Source 2, provided in its raw format by the company Locomizer. Details of access and processing for each data source are provided in the following subsections. Table 1 summarises the main characteristics of each dataset, including the source type, form of data collections, temporal granularity, temporal coverage, spatial resolution, access method and data acquisition cost.

It is important to note that, while Twitter/X is not exclusively accessed via mobile devices and its location data are not always collected via GPS, it has nonetheless been widely used in population and mobility research for its ability to capture patterns at high spatio-temporal resolution and across broad geographic areas. Additionally, the Twitter Academic API is no longer available for free data collection, limiting access to new data. Despite these limitations, we include Twitter/X in our analysis as a representative single-application data source. Archived datasets, such as the one used in this study or the Harvard Geotweet Archive (https://gis.harvard.edu/data) continue to support population and mobility research.

(i) Meta

We use the Facebook Population dataset created by Meta and accessed through their Data for Good Initiative (https://dataforgood.facebook.com). This consists of anonymised aggregate location data from Facebook app users in the UK, who have the location services setting activated on their smartphone. We selected data entries covering March 2021, the month when the most recent UK Census was carried out. Prior to releasing the datasets, Meta ensures privacy and anonymity by removing personal information and applying privacy-preserving techniques, including small-count dropping for population counts under 10, addition of random noise and spatial smoothing using inverse distance-weighted averaging [8].

The dataset includes the number of active Facebook app users, aggregated into three daily 8-hour time windows (i.e. 00:00-08:00, 08:00-16:00 and 16:00- 00:00). To approximate the resident population, we focus on the time window corresponding to nighttime hours (00:00–08:00), when users are more likely to be at home. For the study area, this time window yields an average of 4.2 million daily user records. Spatially, the Facebook Population data is aggregated according to the Bing Maps Tile System [9]. In this study, we use data aggregated at Bing tile level 13, which corresponds to a spatial resolution of approximately 4.9 $\times$ 4.9 km at the Equator [8].

To enable comparison with UK census data, we process the Facebook Population data by averaging daily values over March 2021 and aggregating them to the level of Local Authority Districts (LADs). This harmonisation ensures temporal and spatial alignment with official census boundaries. In the Supplementary Information, we test alternative processing strategies, including averaging over a single week in March and reversing the order of spatial and temporal aggregation. These sensitivity checks confirm that our main findings are robust to variations in the data processing workflow.

(ii) Twitter

We use an anonymised, analysis-ready dataset of active X (previously Twitter) users in the UK, originally collected via the Twitter Academic API. The data consists of monthly counts of active users, spatially aggregated across the UK, and is openly available at `https://github.com/c-zhong-ucl-ac-uk/Twitter-Internal-Migration`. Geolocation is obtained either directly from geotagged tweets or through manual geocoding using bounding boxes provided by the API, based on the IP address of the posting device (for methodological details, see [10]). The full dataset includes approximately 161 million tweets from February 2019 to December 2021. For this study, we restrict the analysis to March 2021 to align with the timing of the 2021 UK Census, during which 125,637 user home locations were identified. Home locations are assigned to Local Authority Districts (LADs) using a frequency-based detection algorithm, further described in [10].

(iii) Multi-app Source 1

We sourced data from a location analytics company that collects GPS data from approximately 26% of smartphones in the UK. The raw data consist of anonymised device-level GPS traces collected via a range of smartphone applications, where users have explicitly granted location-sharing permissions. The dataset spans a 7-day period corresponding to the first week of April 2021 and includes 443,553,155 GPS records. Although the dataset does not perfectly align with the official 2021 UK Census date, the temporal proximity ensures a high degree of comparability.

To infer the place of residence of users, we apply a commonly used rule-based classification method, following approaches outlined in [10,11]. Specifically, the place of residence associated with a device is defined as the location with the highest number of GPS records recorded during nighttime hours (10 PM–6 AM). To be classified as a residence, a location must account for more than 50% of the device nighttime records. Furthermore, the number of nighttime records during the observation period must be at least 2. For comparability across data sources, all identified residence locations are aggregated to the level of Local Authority Districts (LADs). Using this method, we detect 1,536,922 home locations.

(iv) Multi-app Source 2

Our analysis includes a second source of analysis-ready 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 according to the methodology described in [11]. 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 dataset covers the entire UK for November 2021 and includes inferred home and work locations for 630,946 users.

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).

Group	Variable
Demographic	Residents
	Households
	UK (%)
	Female (%)
	Age bands (%): 0-4, 5-9, 10-14, ..., 85-89
Socioeconomic	Not deprived (% households)
	Non-white (%)
	Bad health (%)
	Severe disability (%)
	Socioeconomic Classification [SEC] (%): higher managerial, lower managerial, intermediate, small employers, lower supervisory, semi-routine, routine, no work, students
	Qualifications (%): no qualifications, level 4 qualifications
Housing	No car (% households)
	No central heating (% households)
	Owned (% households)
Geography	Longitude
	Latitude

Figure 2. Table 2. List of covariates used in the analysis, grouped by thematic category. Variables are derived from the 2021 UK Census and aggregated to the Local Authority District (LAD) level, except for geographic coordinates, which correspond to the centroid of each LAD.

(v) Other data

In addition to the mobile phone app data sources described above, we use resident population counts from the 2021 UK Census, aggregated at the LAD level. These counts serve as the ground truth reference for comparing population estimates derived from each digital dataset. We also draw on a set of area-based covariates from the census, covering demographic, socioeconomic, and housing characteristics, along with the geographic coordinates of each LAD centroid. These variables are used to investigate and explain the contextual factors most strongly associated with the magnitude and spatial variation of bias in the digital trace data. The full list of covariates is provided in Table 2.

3. Methods

In the first stage, we introduce a population coverage metric to quantify coverage bias within a given spatial unit. This metric is first computed at the national level for each of the four MP data sources examined in this study. To contextualise the magnitude of bias in MP data, we also compute the same metric for several established national surveys. This enables us to directly compare the extent of coverage bias introduced by digital and traditional data sources.

In the second stage, we calculate the coverage metric for subnational spatial units and anlyse its geographic variation. We map the resulting patterns to identify areas of overrepresentation and underrepresentation and perform statistical tests to assess spatial clustering using measures of spatial autocorrelation. This analysis allows us to detect the presence of systematic spatial patterns in coverage bias across different regions. In the third stage, we use explainable machine learning to model the spatial variation in coverage bias as a function of demographic and socioeconomic covariates derived from the 2021 UK Census. This modelling approach allows us to model the magnitude of bias across areas and, importantly, to quantify the relative importance of each covariate, so it is possible to identify the population characteristics (e.g. age structure, income levels, educational attainment) that are most strongly associated with overrepresentation or underrepresentation in the different sources of MP app data. Figure 1 provides an overview of the methodological workflow, which includes data acquisition, bias measurement, comparative analysis with national surveys, spatial analysis, and bias explanation through modelling.

(i) Measuring coverage bias

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:

$$c_i = \frac{P_i^D}{P_i} \times 100, \tag{3.1}$$

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:

$$e_i = 100 - c_i \tag{3.2}$$

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 analyse the magnitude and spatial distribution of coverage bias across multiple sources of data.

(ii) Spatial patterns of bias

(iii) Explainable machine learning

We used explainable machine learning to identify the key predictors of population bias and how these the importance of these predictors varies across geographical areas. Existing evidence based on social media suggests that population location data from digital platforms are biased over-representing urban, wealthy and young-adult populations **[REF]**. We therefore modelled our measure of population bias from Equation~3.2 as a function of key area-level attributes reflecting geographical differences in engagement and access to digital technology across demographic, socioeconomic, household, housing and location factors. **TABLE XX** reports the set of predictors included in our analysis. We used data from the 2021 census for England and Wales to measure these predictors.

We used an eXtreme Gradient Boosting (XGBoost) algorithm. XGBoost is an ensemble that combines outputs from multiple models to produce a single prediction and represents an efficient and scalable adaptation of the gradient boosting machine algorithm proposed by [12]. It utilises gradient descent to improve model performance, and decision trees are built iteratively, with each tree built to minimise the error residuals of a preceding iteration. XGBoost has been optimised for scalability and computational efficiency, providing high predictive accuracy with limited training time [13,14]. XGBoost has also become one of the most widely-used off-the-shelf machine learning models in applied settings because of its built-in regularization that mitigates overfitting, sparsity-aware tree construction and parallelisation efficiency [13]. It can accommodate nonlinearities and is robust to multicollinearity [13]. We fitted the following XGBoost regression model.

$$\widehat{e}_i = \sum_{m=1}^{M} f_m\left(D_i, S_i, H_i, U_i, L_i\right), \quad f_m \in \mathcal{F} \tag{3.3}$$

e_i is our measure of population bias. f_m denotes an individual regression tree from the boosted ensemble $\mathcal{F}$ and M is the total number of trees. The input variables D, S, H, U, L represent key demographic, socioeconomic, housing, household, and locational attributes of area i, respectively. The model iteratively learns the contribution of each feature to the prediction of the bias indicator e_i, allowing for complex, nonlinear interactions.

To implement Equation~3.3, we randomly split the data into training (80%) and testing (20%) sets to ensure robust model evaluation. We used 10-fold cross validation to train models and performed grid search over learning rates, tree depths, subsample ratios, and regularisation penalties to identify optimal hyperparameters. We applied regularisation penalties including L1 (Lasso) and L2 (Ridge) terms to penalise overly complex trees, promote feature sparsity, improve model generalisation and mitigate multicollinearity among predictors. XGBoost's tree-based structure additionally handles multicollinearity by hierarchically selecting the most informative splits [13]. We then fitted a final model on the full training set using these tuned settings of optimal parameters and evaluated on the held-out test set. We evaluated models based on the number of trees minimising the root mean squared error (RMSE), the convergence of training and test error, and difference between predicted and observed values.

4. Results

(a) Varying extent of bias across data sources

As MP 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 MP 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 [15]. 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 3.2. The figure highlights the remarkable ability of MP 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 MP data, highlighting its potential to inform comprehensive empirical analyses.

While the findings in Figure 3 demonstrate the potential of MP 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

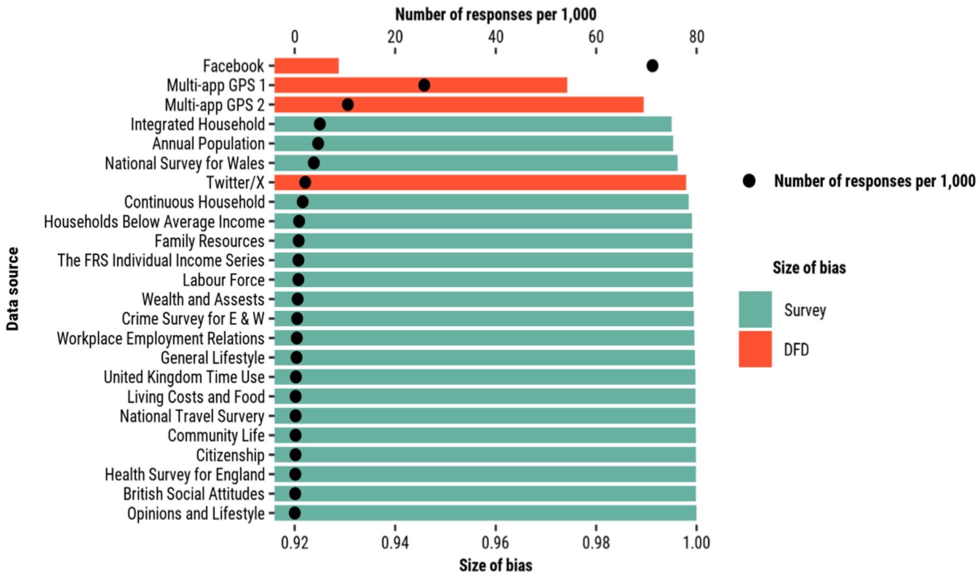


Figure 3. Size of bias and population coverage (per 1,000 population) by data source.

interest [16]. 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 by surveying the whole population.

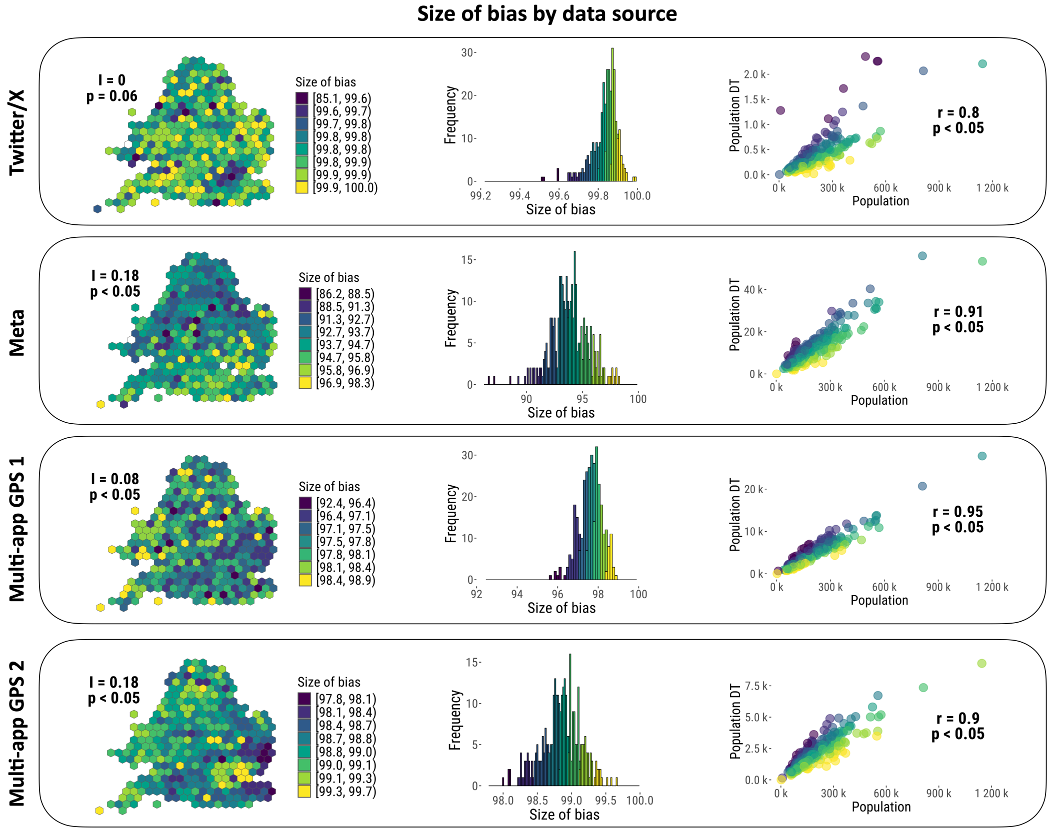
With MP data, achieving statistical representativeness is even more challenging. Unlike survey data, which is actively collected using structured sampling methods, MP data is generated passively as a byproduct of digital interactions, transactions, or device usage, without any control over who is included 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 MP data. This is a necessary first step to understand which population groups might be overrepresented or underrepresented in different sources of MP 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

Next, we take advantage of the detailed geographic information in the MP 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 MP 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 [10] 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

5. Discussion

6. 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. 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. Microsoft. 2023 Bing Maps Tile System. Accessed on 7 August 2025.
10. 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**.
11. 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.
12. Friedman JH. 2001 Greedy function approximation: A gradient boosting machine.. *The Annals of Statistics* **29**.

13. Chen T, Guestrin C. 2016 XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794.
14. Nielsen D. 2016 Tree boosting with xgboost-why does xgboost win" every" machine learning competition?. Master's thesis NTNU.
15. UK Data Service UK surveys - UK Data Service. https://ukdataservice.ac.uk/help/data-types/uk-surveys/. [Accessed 17-02-2025].
16. Cochran W. 1977 *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley.