ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Check for updates

Article submitted to journal

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

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

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

A systematic machine learning approach to quantifying coverage and representation bias in population estimates from mobile phone app data

Carmen Cabrera[1], Francisco Rowe[1]

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

Traditional data sources such as censuses and surveys are costly, infrequent, and often unavailable in crisis-affected regions. User location data derived from GPS-enabled mobile phone (MP) applications offer near-real-time, high-resolution insights into population distribution, but unequal access to and use of mobile technologies introduces biases that threaten representativeness. Existing bias assessments typically require demographic attributes, which are often unavailable, and focus on national-level estimates. We present a generalisable framework to measure and explain biases in aggregated MP app data without the need for individual-level demographic data. The framework quantifies coverage bias, which relates to the share of the population captured, at national and subnational levels, evaluates spatial heterogeneity and clustering, and models the geographic variation of bias as a function of area-based covariates using explainable machine learning. We illustrate the framework using four MP app datasets for the UK, aligned with the 2021 national census. We find that MP data consistently achieve higher coverage than major national surveys, though bias varies spatially and by data source. Multi-application datasets generally reduce (?) coverage bias relative to single-application sources. X emerges as a consistent and important factor in determining the local magnitude of coverage bias.

THE ROYAL SOCIETY
PUBLISHING

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 [4], sense land use patterns [REF], predict socioeconomic levels [REF], define urban extents [REF], quantify tourism activity [REF] and estimate migration and population displacement [5].

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 [6]. 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 [7], but a smaller percentage actively use Facebook (70%) or Twitter (23%) [8]. 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 [9]. Unadjusted, 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 relative level of population coverage and bias of MPD sources to widely-used traditional surveys?
- To what extent, does the level of population bias vary across subnational areas and cluster in particular regions?
- How systematic is the association between larger population biases and over-representation of particular population subgroups, such as rural, more deprived and elderly populations?
- To what extent, are MP-based population data from multiple applications versus single applications associated with lower population 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 or around March 2021 to align as closely as possible with the dates of the most recent census in the area of study, hence enabling 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-app1, whose provider name cannot be disclosed due to a non-disclosure agreement, and Multi-app2, provided in its raw format by the company Locomizer. 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. Further details of access and processing for each data source are provided in the following subsections.

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.

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-app1	Multi app	GPS data from multiple apps	Second	First week, April 2021	GPS coordinates	Proprietary, from analytics company (not public)	No
Multi-app2	Multi app	GPS data from multiple apps (pre-processed)	Averaged over a month	November 2021	MSOA	Open access via GitHub (pre-processed)	Yes

Table 1. Summary description of mobile phone data sources.

(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 accounts in the UK, who have the location services setting activated. We take the number of unique accounts as a proxy for the number of unique users, although it could be the case that one user has more than one account. 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 several techniques which include small-count dropping for population counts under 10, addition of random noise and spatial smoothing using inverse distance-weighted averaging [10].

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 [11]. 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 [10].

We process the Facebook Population data by averaging daily values and aggregating them to the level of Local Authority Districts (LADs), to ensure temporal and spatial alignment with official census data. 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 tests 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) accounts in the UK, originally collected via the Twitter Academic API. Like in the data for Meta, we take the number of unique accounts as a proxy for the number of unique users. The data consists of monthly records for the location of unique Twitter accounts, 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 [12]). 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 [12].

(iii) Multi-app1

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. We consider the number of devices as a proxy for the number of unique users, although it could be the case that some users have more than one device. 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 [12,13]. 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-app2

Our analysis includes a second source of analysis-ready dataset of population counts. This dataset is openly-available on GitHub (https://t.ly/dzlzB)), and has been processed to identify the home location of users according to the methodology described in [13]. 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).

(v) Census data

We use 2021 UK Census, aggregated at the LAD level for two main purposes. First, we use resident population counts derived from the census as the denominator of the population coverage measure in the first stage of the methodology (see Methods section). The resident population count derived from the census is taken as the official benchmark for population counts, so we take these as the ground truth for resident populations and use it to compare with the population counts derived from each digital dataset. Second, we also draw on a set of area-based covariates from the census, covering demographic, socioeconomic, resource accessibility, mobility-related and geographic characteristics. These variables are detailed in Table 2, and are used as predictors in stage three of the methodology (see Methods section), to investigate and explain the factors that are most strongly associated with the magnitude and spatial variation of coverage bias in the digital trace data. This allows us to analyse representativeness bias in each source of digital data.

3. Methods

We introduce a framework to measure and explain biases systematically in population count data derived from mobile phones (MPs). This framework consists of three stages. In the first stage, we introduce a metric to quantify bias related to population coverage within a given geographic area. We refer to this as “population coverage bias”. In the second stage, we calculate the population coverage bias metric for different data sources and subnational geographic areas. We then analyse its variability and assess whether it is evenly distributed across geographies. Detecting such patterns is important for understanding the limits of data applicability, and to assess whether spatial dependencies should be considered in the third stage of the analysis. In the third stage, we analyse representativeness bias. We do this by using explainable machine learning to model the 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 population 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 by dataset D (sample size) in a geographic area i, denoted as P_i^D, to the total

Dimension	Variable	Bias explanation	Description
Demographic	Household size	Large households may share devices, lowering per-capita ownership and digital visibility. Shared phones reduce stability of inferred home locations.	Households with ≥6 people (share of all households).
	Female	Gendered differences in app use and daily routines can alter detection of trips and stays. If women carry or use devices differently, inferred home location can be biased.	Residents who are female (% of populatior
	Age bands	Smartphone ownership and location-service use vary sharply by age. Older groups show lower adoption and usage intensity, leading to under-representation.	Population by 10-year bands: 0–9, 10–19, 70+.
Socioeconomic	Socioeconomic classification	NS-SEC correlates with income and job conditions, shaping handset quality, data plans, and app usage. Lower classes face affordability constraints and intermittent connectivity.	Population by NS-SEC categories.
	Qualifications	Education relates to digital literacy and intensity of technology use. Lower qualifications tend to associate with lower adoption of location-enabled services.	Population by highest qualification (e.g., n Level 4+).
Resource accessibility	Deprivation	Higher deprivation implies barriers to devices and data plans, reducing digital visibility. Prepaid SIM churn may degrade longitudinal tracking of home locations.	Households not deprived in any dimensior
	Car ownership	No-car households proxy lower material resources and different mobility modes. They often have lower device affordability and different app ecosystems, affecting capture.	Households with no car (%).
	Home ownership	Owners have higher income stability and better digital access, improving persistent device presence and home inference. Renters show higher SIM churn and instability in traces.	Owner-occupied households (%).
	Central heating	Lack of central heating is a deprivation proxy linked to lower digital access. Housing quality can also affect indoor signal conditions, increasing missingness.	Households without central heating (%).
Mobility	Non-UK born	Foreign operator SIMs, language settings, and app ecosystems can reduce inclusion in domestic panels. Roaming and multi-SIM use complicate home/work inference.	Residents born outside the UK (%).
	Recent migrant	Short residence duration implies unstable addresses and higher SIM turnover, lowering longitudinal representativeness. Temporary housing can degrade location inference.	Residents in UK <2 years (% of population)
	Commuting	Work-from-home reduces observable trips and alters diurnal patterns used for home/work inference. Mode and route choices also affect signal continuity and coverage.	Residents working mainly or only from hor (%).
Geographic	Population density	Dense urban areas have better coverage but more multipath and building interference, distorting stay detection. Sparse areas suffer coverage gaps, undercounting movements.	Usual residents per km².
	Rural	Rural coverage is patchier and devices may be offline longer, increasing under-representation. Longer mast distances reduce spatial precision of traces.	Population living in rural areas (%).

Table 2. Model variable description, expected influence and description.

local population of the same area, P_i. Formally, the coverage c_i for area i is given by:

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

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 for area 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 distribution of coverage bias across multiple sources of data and geographic areas.

(ii) Identifying spatial patterns of bias

For each data source, we compute the coverage bias metric at the subnational level and examine its geographic variation. This stage has two main objectives. First, to assess whether bias is evenly distributed across geographies. Second, to determine whether spatial effects are sufficiently strong to consider them in the subsequent methodological stage through the inclusion of spatial lag terms in the explainable machine learning model.

To evaluate the variability of bias across geographies, we first conducted exploratory analyses using thematic maps and histograms. To formally test for spatial clustering, we calculate Moran's I statistic for each dataset. Because Moran's I is sensitive to the definition of spatial relationships, we evaluate four alternative spatial weighting schemes: 1) queen neighbourhood, 2) k-nearest neighbours, 3) distance band (set by algorithm), and 4) distance band (set by user) [14]. Comparing results across these schemes enables us to assess the robustness of clustering patterns to different definitions of the spatial weight matrix. In the main body of the paper, we report Moran's I values obtained using scheme 1, as it produces the highest statistic across datasets when statistically significant, thereby providing the most conservative test for the presence of spatial clustering. Results for the other schemes are provided in the Supplementary Information. For each dataset, we compute the range of Moran's I values across the four spatial weighting schemes (maximum minus minimum). The largest such range observed across all datasets is 0.286. This indicates that, while the exact values of Moran's I vary somewhat with the choice of weighting scheme, the differences are small and the resulting statistics remain relatively close.

Finally, to assess whether bias is associated with population size, we examine the relationship between population counts from digital data sources and census population counts. If bias varies systematically with population size, we would expect departures from proportionality. Conversely, an approximately linear relationship through the origin with a stable slope would suggest that bias is largely independent of population size. To test this, we generate scatterplots comparing the two sources and quantify the strength of the relationship using Pearson's correlation coefficient.

(iii) Analysing representativeness bias with explainable machine learning

We used explainable machine learning to identify the key predictors of population bias and how 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, resource accessibility, mobility and geographic factors. Table 2 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 [15]. 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 [16,17]. 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 [16]. It can accommodate nonlinearities and is robust to multicollinearity [16]. We fitted the following XGBoost regression model.

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

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 [16]. 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

Next, we illustrate our proposed methodological framework on four sources of digital data derived from mobile phone (MP) apps . As described in the Data section 2, these sources include data for the UK, collected in or around March 2021 to align as closely as possible with the reference date of the most recent national census.

(a) The extent of population coverage bias varies across data sources

Population coverage bias was first computed at the national level for each MP data source. To contextualise these results, we compared the MP sources with several widely used traditional datasets, including major UK surveys available through the UK Data Service [18]. Figure 1 presents these comparisons across data sources, showing on the top x-axis the population coverage (expressed as the number of respondents or subjects per 1,000 people) and on the bottom x-axis the corresponding measure of coverage bias.

Figure 1 shows differences between MP data and traditional survey sources. UK surveys typically achieve coverage of only a few individuals per 1,000 population, whereas MP datasets can capture more observations. At the same time, MP data exhibit comparatively lower national-level coverage bias, suggesting that in terms of raw size and breadth, digital trace data have strong potential to support large-scale empirical analyses.

Despite the higher population coverage in MP data, both MP data and traditional surveys offer complementary strengths and weaknesses. Traditional surveys are designed to achieve statistical representativeness, usually, at the national level. Sampling techniques such as stratified or cluster sampling are applied and, if imbalances remain, further post-stratification adjustments are implemented [19]. Yet, representativeness can only ever be guaranteed with respect to a finite set of attributes (e.g., age, gender, income, region) [20]. Moreover, survey data usually provides a snapshot at one point in time, and offers little visibility into how coverage and bias vary geographically or dynamically over time.

By contrast, MP data is not collected with sampling strategies and usually lacks demographic identifiers, which makes it difficult to apply post-stratification adjustments. It is also generated passively, as a byproduct of digital interactions, without any guarantee of inclusion for particular groups. However, the much broader poulation coverage and fine spatio-temporal resolution of MP datasets provide information that is usually unavailable through traditional surveys cannot. They allow us to track how representation varies across regions and through time, and this is particularly valuable for studying dynamic social processes.

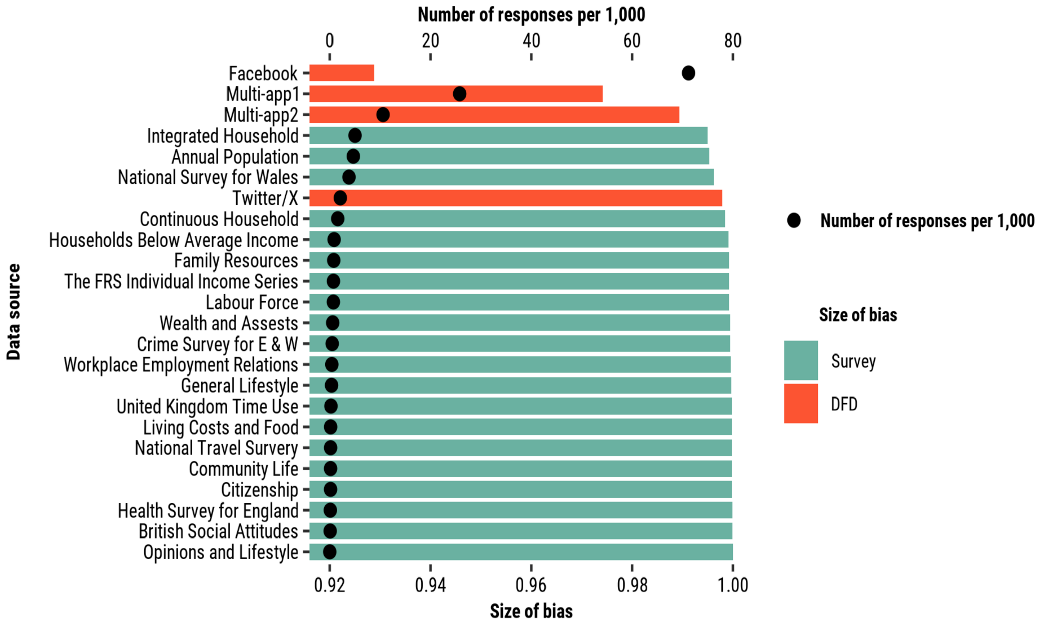

Figure 1. Size of coverage bias (bottom x-axis) and population coverage per 1,000 population (top-x-axis) by data source.

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 develop subsequent data adjustment strategies that can improve the representativeness of the data relative to the target population.

(b) The spatial variability of bias

To examine coverage bias at subnational levels, we leveraged the fine-grained geographic resolution of the MP app data sources. Analysing the distribution of coverage bias across geographies allowed us to identify uneven patterns of population coverage across areas and assess whether these patterns exhibit spatial clustering. These assessments were important both for evaluating the limitations of each dataset for further research and for deciding whether spatial dependence terms should be incorporated into the subsequent modelling stage @**??**#sec-explain).

Figure **?**](fig:bias-size) summarises these analyses for each dataset. Each row corresponds to a data source and includes three elements: A) a hexagonal cartogram showing the size of coverage bias in each LAD, alongside the Moran's I statistic and corresponding p-value as a measure of spatial autocorrelation; B) a histogram showing the distribution of coverage bias across LADs; and C) a scatter plot comparing population counts derived from each MP dataset with census population counts for the same LADs, together with Pearson's correlation coefficient and p-value.

Our spatial analysis revealed several patterns in the distribution of coverage bias which are consistent across the mobile phone (MP) datasets. First, all data sources exhibited noticeable geographic variability in coverage bias, as evidenced by the spread of values in their respective distributions, as displayed in the histograms. For example, bias in data from Meta ranges from 86.2 ($\approx 14\%$ population coverage) to 98.3 ($\approx 2\%$ population coverage). This variability highlights that for a given data source, the degree of representation can differ substantially between areas.

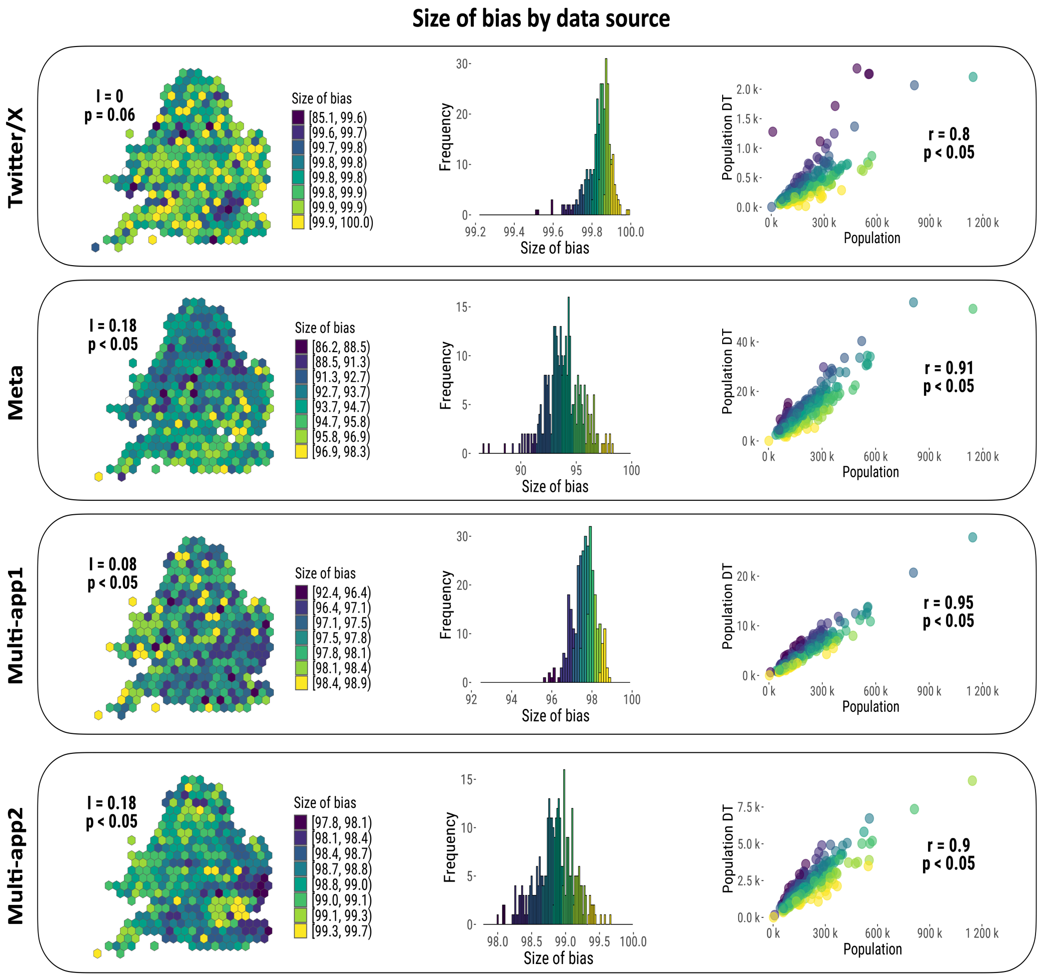

Figure 2. Extent and spatial distribution of population bias across local authority districts in the UK. A. Hexagon map of the size of population bias across local authority districts. I represents the Moran's I and p denotes the associated p-value. **B.** Histogram of the distribution of the size of population bias. **C.** Scatter plot of the relationship between population size and size of bias.

Second, despite the variability in coverage bias, the maps did not reveal strong geographic clustering patterns. Bias values fluctuate across longitude and latitude, and there are no strong north–south or east–west gradients. This observation was quantitatively supported by the Moran's I statistics, which are generally statistically significant, but close to zero. Their small magnitude indicated that spatial clustering is weak at the LAD scale. Consequently, we concluded that it is unnecessary to include spatial dependence terms (e.g. spatial lag) in the subsequent modelling stage of our framework.

Third, the variability in coverage bias was not explained by the absolute population size of each LAD. Scatter plots comparing population counts derived from MP data and from the census revealed strong linear relationships. This was quantitatively supported by Pearson correlation coefficients consistently close to one and statistically significant.

In terms of specific patterns of population coverage bias, we observed some differences across datasets. We found that Twitter/X data shows the higher values of population coverage bias across MP data sources. The majority of values of population coverage bias size for this source are above 99 ($\approx 1\%$ population coverage). An exception is the LAD for City of London, where population coverage bias is 85.1 ($\approx 15\%$ population coverage), where a higher number of home locations was detected. This pattern is likely related to the unique demographic and spatial context of the City of London. Although it has a very small resident population, it is a major centre of employment and commuting. Because home locations are inferred from the most frequently recorded nighttime position [12], even a relatively modest number of residents or temporary visitors can yield an unusually high apparent population coverage compared to larger LADs. By contrast, data from Meta displayed the lower values of population coverage bias, ranging from 86.2 ($\approx 14\%$ population coverage) to 98.3 ($\approx 2\%$ population coverage). Although Moran's I for Meta data is small ($I = 0.18$), the corresponding map suggested a slight gradient with LADs in Northern England displaying lower values of bias. Multi-app1 and Multi-app2 data showed intermediate levels of bias, but with differing distributions. Notably, data from Multi-app1 shows levels of population coverage bias as low as 92.4 for some LADs ($\approx 8\%$ population coverage), contrasting with the lowest levels of population coverage bias for Multi-app2, 97.8 ($\approx 2\%$ population coverage). Multi-app1 and Multip-app 2 also show different spatial patterns in the distribution of population coverage bias across LADs.

Taken together these findings suggested that there is spatial variability in coverage bias at the LAD level, but this variability is not explained by physical geographic location or population size. Instead, the observed patterns are more likely explained by other area-level characteristics, such as demographic or socioeconomic composition. We examined these factors in the third stage of our analysis.

(c) Key contextual features explain population biases

We assessed the contribution of key contextual features to explaining spatial variations in population biases across demographic, socioeconomic, resource accessibility, mobility and geographic domains. Figure 3 displays the relative importance of each individual model feature, representing the average absolute SHapley Additive exPlanations (SHAP) value per feature including in our XGBoost model described in Section iii **[REF]**. SHAP values are standardised by data source based on minimum and maximum scores to enable comparability. Figure 3 highlights features with standardised importance scores over 0.5 for individual data source.

The results reveal wide variability in the key predictors of population biases across digital data sources. Demographic and socioeconomic features appear as the most important predictors of population bias in Twitter/X data. Two features standout -the shares of population aged 20-29 and no car ownership- reflecting the over-representation of the young adult populations or population with access to resources on active Twitter/X users. In addition to demographic and socioeconomic features, resource accessibility and geographic attributes also display some of the largest contributions to explain population biases across Meta, Multi-app1 and Multi-app2 platforms. Coupled to the share of population in routine occupations and lacking central heating, geographical factors particularly population density and rurality report the highest SHAP averages, contributing to explain the spatial variability of Meta-derived population biases. Socioeconomic features standout as the most important predictors of population biases across both multi-app platforms, though differences exist. The population share with high or no qualification, and working in low skill occupations emerge as the most important features in explaining population biases in Multi-app1. This is in addition to the population share aged 60-69, residing in households lacking central heating and living in rural areas. For multi-app2-derived estimates, the share of student population, working in intermediate-level jobs, small employers, having a Level 4 qualification and living in large households score the highest average SHAP values. The variability in feature importance ranking across data sources reflects differences in the contextual features that contribute to explaining biases in their respective population estimates.

Expanding this evidence, Figure 4 depicts the way these contextual features contribute to increasing or reducing the extent of population bias across LADs. Figure 4.a shows the top 20 features based on their average SHAP value from the highest to the lowest. Figure 4.b displays top six of these features revealing how changes in feature values contribute to changes in population

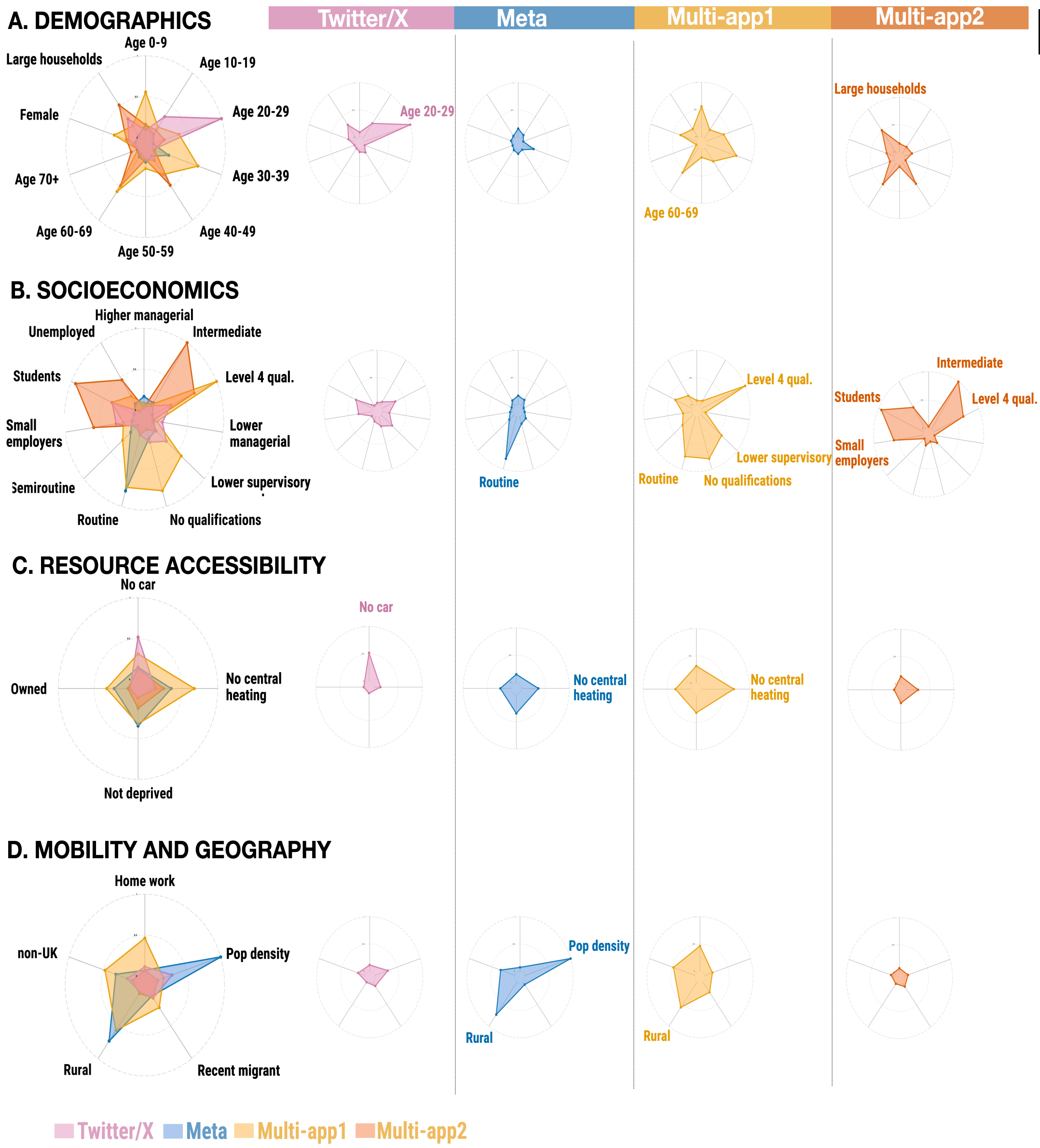

Figure 3. Radial charts illustrating the importance of, and contribution to, explaining differences in population biases across local authority districts. The importance is estimated based on SHAP feature importance scores, calculated as the average absolute SHAP value per feature using an XGBoost machine learning algorithm. Average absolute SHAP values were normalised across individual data using minimum and maximum scores, to ensure comparability across data sources. First column displays the estimates for all data sources and model features. Subsequent columns highlight features scoring SHAP values over 0.5 within a 0-1 range for individual data sources.

bias, with colour encoding feature values. For instance, the first column of plots show that areas with larger shares of population aged 20-29 and lacking central heating tend to have lower population bias compared to those with larger shares in Twitter/X-derived estimates. In contrast, areas with larger shares of population aged 10-19 are associated with higher Twitter/X-based population bias, potentially reflecting a limited number of active Twitter/X users in this age range. Of the features highlighted in Figure 3, Figure 4 indicates that population estimates

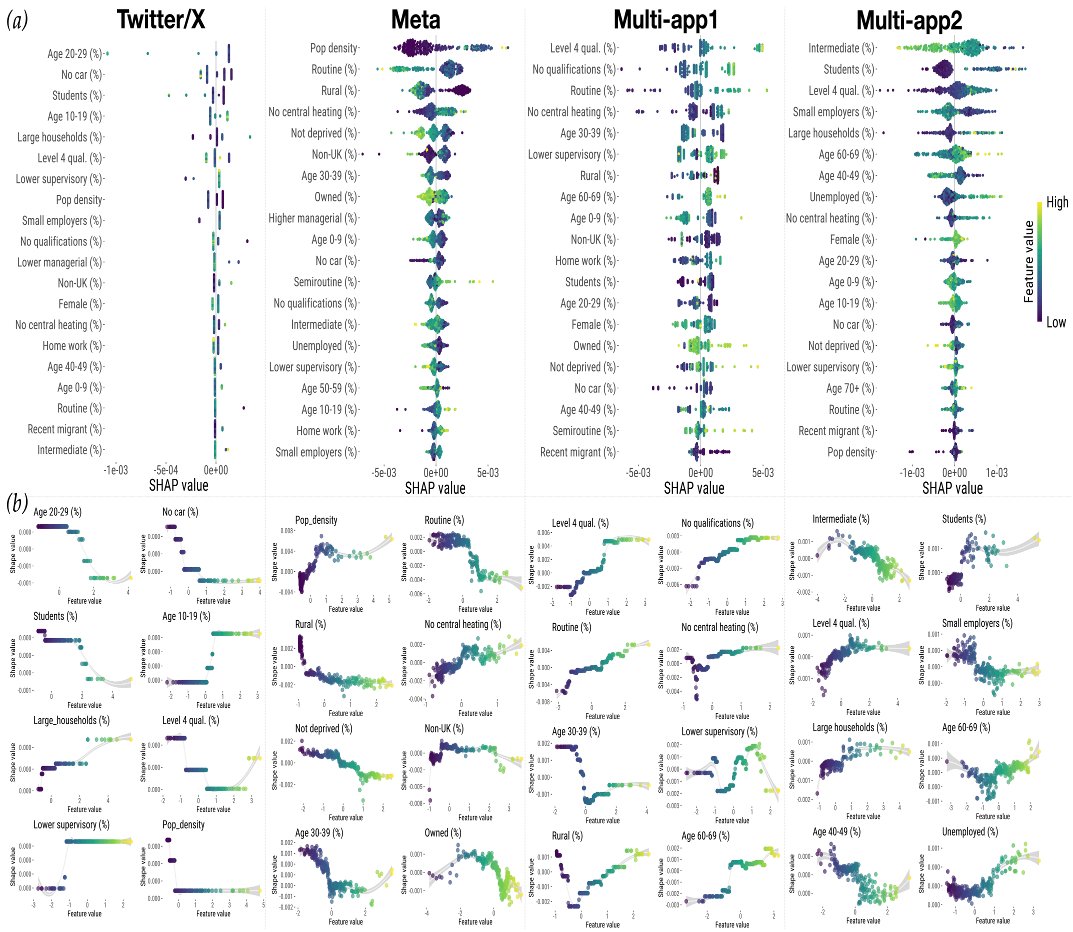

Figure 4. Top ranked model features contributing to explaining population biases across local authority districts from an XGboost model. (a) beeswarm plots of SHAP feature values displaying the relative importance and direction of influence of the top 20 contextual features on the extent of population bias. Features are ranked by their mean absolute SHAP value, with colours indicating feature values (low to high). (b) SHAP dependence plots for the top six features based on their mean absolute SHAP value, illustrating the marginal effect of variation in each predictor on population bias. Local polynomial regression modelling was used to represent local trends with 95% confidence intervals. Each point represents a local authority district.

derived from Meta tend to have larger biases in areas with higher population density and greater percentages of people lacking central heating. Operating in the opposite direction, larger shares of people working in routine-level jobs and living in rural areas display lower Meta-based population bias, reflecting greater engagement with Facebook among these communities. For multi-app1-derived estimates, biases are larger for areas with greater shares of population with Level 4 or no qualification, working in routine-level jobs, lacking central heating, living in rural communities and population aged 60-69. For multi-app2-derived estimates, biases are greater for areas with smaller shares of people working in intermediate-level jobs and self-employed in small businesses, but displaying larger percentages of students, people with Level 4 qualification and living in large households.

Figure 4 also reveals systematic complex nonlinear shapes in the association between contextual features and population bias. We identified three types of nonlinear relationships. First is curvilinear associations in the way of U-shape or inverse U shapes. These involve patterns of population biases decreasing at low values of contextual features, increasing at medium values and reducing again - or the reverse. A distinctive pattern of these associations is their curvature representing a reversal in the direction of the relationship between population bias and a contextual feature. Meta-based population density and Multi-app2-based intermediate feature estimates represent prominent curvilinear relationships. A second pattern takes the form of S-shaped associations. The distinctive feature of these associations is their multiple phase composition displaying a different pattern of population biases at low and high end values of the contextual feature relative to middle range feature values. Twitter/X-based 20-29 age estimates, for example, display high and unchanging population bias at low feature values, highly variable bias at mid values and low but increasing population bias at high values. A third pattern is threshold/stepwise associations representing sharp changes at a cut-off. The distinctive feature of this pattern is sharp "steps" or thresholds where the relationship between population bias and a contextual feature changes abruptly. Population biases would remain flat at low values and then jump and plateau at higher values. Our Multi-app1 estimates for Level 4 qualification represent a clear illustrative of this type of association displaying flat small population bias at low feature values but, then jump and remain high at higher feature values. These shapes do not appear to be data source specific and vary in unpredictable ways across variables.

5. Discussion

(a) Key findings

MPD have become a key data asset to understanding population dynamics in near real-time at high spatial and temporal resolution. Yet, biases have remained a major barrier eroding their trust and wider adoption. A key limitation to tackle this issue has been the lack of a standardised method to estimate and analyse coverage biases in MPD-derived population data. In this study, we sought to contribute a systematic and generalisable approach to measure and analyse coverage biases in MP-derived population data, when access to individual-level demographic attributes are lacking. Implementing this approach on UK data, we presented evidence showing that MDP tend to provide greater population coverage than widely used traditional surveys. Our findings also revealed that coverage biases vary markedly across data sources and subnational areas, that multi-app data sources does not necessarily offer broader coverage than single-app sources, and that demographic, socioeconomic and geographic features consistently explain a large share of spatial variation in coverage bias. Additionally, we identified distinctive nonlinear effects in the association between contextual features and coverage bias, suggesting that such biases cannot be addressed with simple linear adjustment.

(b) Implications

These findings carry wide-ranging implications for the use of MPD in research and official statistics. First, the framework we propose offers a replicable approach for evaluating the quality of MP-derived population counts when demographic information is unavailable, establishing a baseline for transparent quality assessment. Second, the evidence provided here underscores the need for source-specific adjustment strategies to mitigate data biases. While single-app datasets show biases predominantly related to a reduced set of factors -such as age and geographic features- multi-app sources require more complex calibration that accounts for variations across a wider range expanding from educational and occupational to age and household features. This finding suggests that improving the representativeness of statistics derived from multi-app sources will require more sophisticated reweighting and debiasing strategies than those from single-app data. More broadly, our study reinforces the view that digital trace data should not be treated as neutral reflections of population activity but rather as selective windows shaped by persistent digital divides. Recognising and adjusting for these divides is essential if MPD are to contribute effectively to decision-making in contexts such as crisis response, public health monitoring and infrastructure planning.

(c) Challenges and limitations

(d) Future directions

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. Cabrera-Arnau C, Zhong C, Batty M, Silva R, Kang SM. 2023 Inferring urban polycentricity from the variability in human mobility patterns. *Scientific Reports* **13**.
5. Rowe F, Cabrera-Arnau C, González-Leonardo M, Nasuto A, Neville R. 2024 Medium-term changes in the patterns of internal population movements in Latin American countries: effects of the COVID-19 pandemic. Población y desarrollo Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL).
6. 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.
7. Ofcom. 2023 Communications Market Report 2023. Accessed: Nov 2023.
8. Statista. 2024 Social Media & User-Generated Content - Market Overview. Accessed: 2024-11-14.
9. Rowe F, Neville R, González-Leonardo M. 2022 Sensing Population Displacement from Ukraine Using Facebook Data: Potential Impacts and Settlement Areas. *OSF Preprints*. Submitted.
10. Maas P. 2019 Facebook Disaster Maps. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
11. Microsoft. 2023 Bing Maps Tile System. Accessed on 7 August 2025.
12. 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**.
13. 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.
14. Rey S, Arribas-Bel D, Wolf LJ. 2023 Geographic Data Science with Python. .
15. Friedman JH. 2001 Greedy function approximation: A gradient boosting machine.. *The Annals of Statistics* **29**.
16. Chen T, Guestrin C. 2016 XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794.

17. Nielsen D. 2016 Tree boosting with xgboost-why does xgboost win" every" machine learning competition?. Master's thesis NTNU.
18. UK Data Service UK surveys - UK Data Service. https://ukdataservice.ac.uk/help/data-types/uk-surveys/. [Accessed 17-02-2025].
19. Lohr SL. 2021 Sampling. .
20. Cochran W. 1977 *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley.