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, Digital trace data, Population estimates, Coverage bias, Data representativeness, Location, Explainable AI, Machine learning, Spatial bias

**Author for correspondence:**

Carmen Cabrera, Francisco Rowe

e-mail: c.cabrera@liverpool.ac.uk; fcorowe@liverpool.ac.uk

# A systematic machine learning approach to assess biases in population data from mobile phone applications

Carmen Cabrera[1], Francisco Rowe[1]

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

Traditional sources of population data, such as censuses and surveys, are costly, infrequent, and often unavailable in crisis-affected regions. Mobile phone application data (MPD) offer near–real-time, high-resolution insights into population distribution, but their utility is undermined by unequal access to digital technologies, creating biases that threaten representativeness. Despite recognition of these issues, no standard framework exists to address such biases, limiting the reliability of MPD for research and policy. We develop and implement a systematic, replicable framework to quantify and explain population coverage bias in aggregated mobile phone application data without requiring individual-level attributes. The approach combines an indicator of population coverage bias with explainable machine learning to identify contextual drivers of spatial variation in bias. Using four datasets for the United Kingdom benchmarked against the 2021 census, we show that MPD achieve higher population coverage than national surveys, but biases persist across sources and subnational areas. Population coverage bias is strongly associated with demographic, socioeconomic, and geographic features, often in complex nonlinear ways. Contrary to common assumptions, multi-application datasets do not necessarily reduce bias compared to single-app sources. Our findings establish a foundation for bias assessment standards in MPD, offering practical tools for researchers, statistical agencies, and policymakers.



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 the 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 [4–6]. Dwindling numbers in surveys can represent distorted picture of society [6]. 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 [7]. 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 [2]. Hence, national statistical offices are actively seeking to integrate these data into their national data infrastructure [8,9].

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 approaches to infer mode of transport [10], monitor footfall changes [11,12], profile daily mobility signatures [13], sense mobility accessibility [14], predict socioeconomic levels [15,16], estimate income segregation [17], quantify tourism activity [18] and estimate migration [19,20] and population displacement [3,21,22].

However, the use of MPD present major epistemological, methodological and ethical challenges [2]. A key unresolved challenge is potential biases in MPD compromising their statistical representativeness and perpetuate social injustice [23]. Biases reflect societal digital and socioeconomic inequalities. Biases emerge from differences in the access and use of the MP applications used to collect MPD [23,24]. Only a fraction of the population in a geographical area owns a smartphone, and even an smaller share actively uses a specific MP app. In the UK, for example, 98% of the adult population have a MP and 92% of this population use a smartphone [25], but a smaller percentage actively use Facebook (70%) or Twitter (23%) [26]. 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 MP applications, and therefore tend to be over-represented in MPD [23,27,28].

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 [23,24]. 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 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 [29]. Unadjusted, they 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, compared to those from censuses [30–32]. 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 [30,32]. 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 [33]. 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 identify this as population coverage bias. We use data collected from single and multiple MP 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 an undisclosed 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. Our first contribution is methodological. We develop and demonstrate a systematic, replicable approach for assessing the quality of MP–derived population data when information on population attributes is unavailable. This approach quantifies population coverage bias at national and subnational scales, identifies the degree of spatial variation and clustering, and uses explainable machine learning to identify the contextual drivers of these biases. By establishing a simple yet robust and transparent indicator of coverage bias, our approach provides a practical tool to evaluate digital trace data prior to substantive analysis. In doing so, we respond directly to recent calls in the human mobility literature for greater transparency and standardisation in the processing and validation of digital trace data [33–35]. While prior work has emphasised the fragility of analytical choices and the absence of good practices in this field, our approach establishes a clear set of indicators and procedures that can be applied consistently across sources and geographies. By providing a privacy-preserving and transparent benchmark, our approach offers a foundation for emerging standards of good practice in the use of MPD, bridging the gap between innovative digital signals and the established norms of accountability that underpin official population statistics.

Our second contribution is substantive. Using our proposed approach, we present the first systematic cross-platform assessment of biases in MP app datasets for the UK. We show that coverage bias varies substantially across sources and geographies, and that demographic, socioeconomic and geographic features consistently explain much of this variation. This evidence provides robust insights into the structural roots of bias, such as the role of age composition, education, occupation and rurality in shaping representation. Importantly, we also demonstrate that MP multi-app datasets -often assumed to be more representative- do not inherently guarantee lower bias. Instead, they capture a broader population base at the expense of introducing more complex and multidimensional patterns of bias compared with single-app sources. These findings extend current debates on the quality of human mobility data by moving beyond descriptive comparisons to provide systematic, evidence-based assessment. They also directly address concerns raised by [34] that the lack of validation across datasets hinders the reliability of policy-relevant metrics. Our results highlight that improving representativeness requires source-specific adjustment strategies and cannot rely solely on aggregation across apps, thereby setting a new standard for empirical evaluation of MP–based population estimates.

## 2. Data

We propose a systematic framework to measure and explain biases in population count data derived from MPs. Figure 1 provides a digramatic representation of our proposed approach consisting of four stages: (1) measuring coverage biases; (2) assessing the comparative extent of bias against widely used survey data; (3) determining the level of spatial variability in bias; and, (4) identifying key contextual sources of biases across subnational areas. To illustrate our proposed framework, we draw on four MP datasets collected in March 2021 to align as closely as possible with the dates of the most recent census, and thus ensure temporal consistency. As explained below, the key principle here is that we use census data as our benchmark, so we assume that full representation is achieved when a source provides the same level of coverage as the census. 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. We conducted analysis at the common spatial framework, which is the Local Authority District (LAD) level given that data from Facebook and Twitter/X are available at aggregate geographical scales. 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 next.

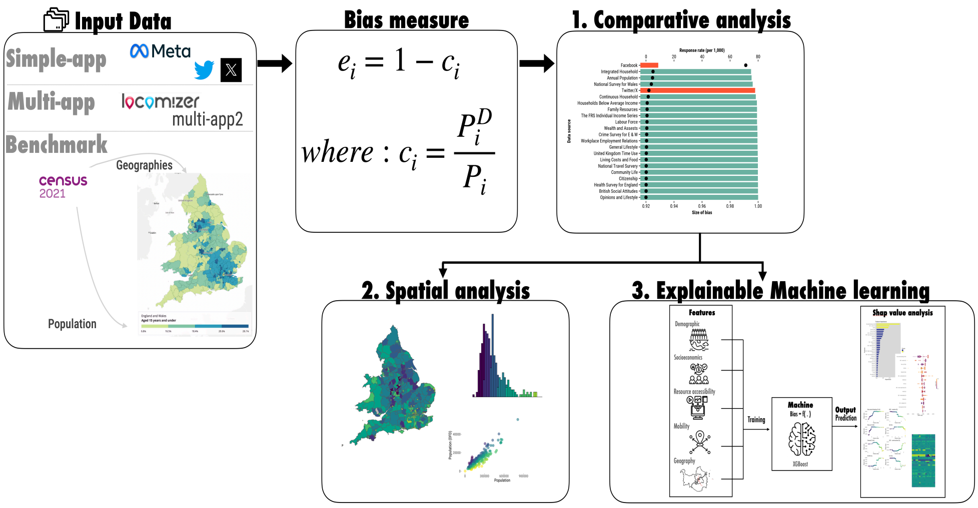


**Figure 1.** Systematic approach to measure and explain biases from MP-based population estimates.

| Data source | Type | Form of data collection | Finest temporal resolution | Temporal coverage | Finest spatial resolution | Access method | Free at time of access |
|---|---|---|---|---|---|---|---|
| Facebook | Single app | GPS from app users with location services enabled | 8-hour window | 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 | Month | March 2021 | Local Authority District | Open access via GitHub (pre-processed*) | Yes |
| Multi-app1 | Multi app | GPS data from multiple apps | Second | 1st week, April 2021 | GPS coordinates | Proprietary, from analytics company (not public) | No |
| Multi-app2 | Multi app | GPS data from multiple apps | Average over a month | November 2021 | MSOA | Open access via Zenodo (pre-processed*) | Yes |

* The original raw data was processed by other authors before use in this study. Specifically, data was cleaned, aggregated across space and time, and then made openly available. Further details are provided in the Data section.

**Table 1.** Summary description of mobile phone data sources.

## (a) Meta

We used the Facebook Population dataset created by Meta and accessed through their Data for Good Initiative[1]. This dataset provides anonymised aggregate location data from Facebook app accounts, who have the location services setting activated. We used data for the UK covering March 2021 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 techniques to remove population counts under 10, add random noise and spatial smoothing using inverse distance-weighted averaging [36].

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 [37]. In this study, we used data aggregated at Bing tile level 13, which corresponds to a spatial resolution of approximately 4.9 × 4.9 km at the Equator [36].

To integrate the Facebook and census data, we aggregated the Facebook Population data by averaging daily population counts to the level of LADs, to ensure temporal and spatial alignment

[1] https://dataforgood.facebook.com

with official census data. In Section 1 of the Supplementary Material (SM), we provide evidence testing alternative processing strategies, including averaging over a single week in March and reversing the order of spatial and temporal aggregation. These sensitivity provides evidence that our main findings are robust to variations in the strategy of spatial and temporal aggregations.

## (b) Twitter/X

We used an anonymised, analysis-ready dataset of active Twitter/X accounts in the UK, originally collected via the Twitter Academic API. We used unique accounts as a proxy for the number of unique users. The data consists of monthly records for the location of unique Twitter/X accounts, spatially aggregated across the UK, and is openly available[2]. 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 [38]). The full dataset includes approximately 161 million tweets from February 2019 to December 2021. For this study, we restricted the analysis to March 2021 to align with the timing of the 2021 UK Census, during which over 125 thousand user home locations were identified. Home locations were assigned to LADs using a frequency-based detection algorithm, further described in [38].

## (c) 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 considered 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 over 443.5 million 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 applied a commonly used rule-based classification method, following approaches outlined in [33,39]. 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 LADs. Using this method, we detected over 1.5 million home locations.

## (d) Multi-app2

Our analysis includes a second source of analysis-ready dataset of population counts. This dataset is openly-available on Zenodo[3], and was processed to identify the home location of users according to the methodology described in [39]. 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 over 630 thousand users. While this period does not exactly coincide with the 2021 UK Census, the difference of less than a year was 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 LAD.

[2]https://github.com/c-zhong-ucl-ac-uk/Twitter-Internal-Migration
[3]https://zenodo.org/records/13327082

## (e) Census data

We used 2021 UK census, aggregated at the LAD level for two main purposes. First, we used 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 was used as the official benchmark for population counts. The core assumption is that census data provide the "truth'' resident population count and so deviation from it indicates greater bias. We thus compare population counts derived from each digital dataset against those from the census. Second, we 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 Section iii. They are used as predictors in the final stage 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 MPs. This framework consists of three stages. We first introduce a metric to quantify bias related to population coverage within a given geographic area. We refer to this as "population coverage bias". Next, we compute the population coverage bias metric for different data sources and subnational geographic areas. We then analyse its variability and assess its unevenness distribution across geographies. Detecting such patterns is important for understanding the limits of data applicability, and to assess whether spatial dependencies should be considered in identifying the contextual sources of population bias. In a final stage, we analyse representativeness bias. We do this by using explainable machine learning to model the variation in population coverage bias as a function of key contextual features derived from the 2021 UK census. This modelling approach allows us to model the magnitude of bias across areas, and 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. As shown above, 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 defined a metric to quantify the magnitude of population 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 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 $i$. Though this is unusually in empirical applications.

We then define population coverage bias $e_i$ for area $i$ as:

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

A value of $e_i = 0$ indicates absence 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 population coverage bias

For each data source, we computed the coverage bias metric at the subnational level and examined its geographic variation. This stage has two main objectives. The first assesses the extent of geographical unevenness in bias across subnational areas. The second seeks to determine whether spatial effects are sufficiently strong to consider them in the subsequent methodological stage through the inclusion of spatial lag terms in our machine learning model.

To evaluate the variability of bias across geographies, we first analysed the spatial structure and concentration of bias coverage across subnational areas. To test for spatial clustering, we computed 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) optimal distance band, and 4) user-defined distance band [40]. 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 values of Moran's I 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 Section 2 of the SM. For each dataset, we computed the range of Moran's I values across the four spatial weighting schemes, taking the maximum minus the minimum value. 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 stable.

To assess the association between coverage bias and population size, we examined 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 generated scatterplots comparing the two sources and quantify the strength of the relationship using Pearson's correlation coefficient.

### (iii) Identifying sources of representativeness bias via explainable machine learning

We used explainable machine learning to identify the key predictors of population coverage 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 [23,27,28]. We therefore modelled our measure of population coverage bias from Equation 3.2 as a function of key area-level attributes reflecting contextual 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 [41]. 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 [42,43]. 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 [42]. It can accommodate nonlinearities and is robust to multicollinearity [42]. 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 coverage 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, resource accessibility, mobility and geographic attributes of area $i$, respectively. Table 2 lists and describes these features. Based on existing literature, Table 2 also describes the expected relationship with population coverage bias. Here it is relevant to highlight that representative subnational data on digital resource accessibility and engagement is not available in the UK. We proxied this via our census-derived measures of resource accessibility. The model iteratively learns the contribution of each feature to the prediction of the population coverage 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 [42]. 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 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 MPD 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 [44]. Figure 2 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 population coverage bias.

The results show considerable differences between MPD and traditional survey sources. UK surveys typically achieve coverage of less than 15 individuals per 1,000 population, whereas MP datasets provide numbers of as large as 22 to 70 people per 1,000 population, representing 1.5 to nearly 5 times greater coverage. In proportional terms, that equates to typical survey coverage of 1.5% of the population, while MPD capture between 2.2% and 7%, meaning a several-fold increase. At the same time, MPD exhibit comparatively lower national-level coverage bias as indicated by smaller bars. Overall, these results suggest that in terms of raw size and breadth, MDP have strong potential to support large-scale empirical analyses.

| 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 dimension |
| | 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.

Despite differences in population coverage, both MPD and traditional surveys offer complementary strengths and weaknesses. Traditional surveys are designed to achieve statistical representativeness at the national level. They rarely expand to provide robust levels of subnational representation. Sampling techniques such as stratified or cluster sampling are applied to survey data and, if imbalances remain, further post-stratification adjustments can be implemented [45]. Yet, representativeness of survey data can only be guaranteed with respect to a finite set of attributes (e.g., age, gender, income, region) [46]. Moreover, survey data usually provide a snapshot at one point in time, and offer little visibility into how coverage and bias vary geographically or dynamically over time.

By contrast, MPD are not collected with sampling strategies and usually lack demographic identifiers. As such, applying the similar post-stratification adjustments is challenging. Additionally, MPD are generated passively, as a by-product of digital interactions, without any guarantee of inclusion for particular groups. However, Figure 2 evidences, MPD provide a much broader population coverage and fine spatio-temporal resolution which is unavailable through traditional surveys. These attributes allow us to track how representation varies across regions

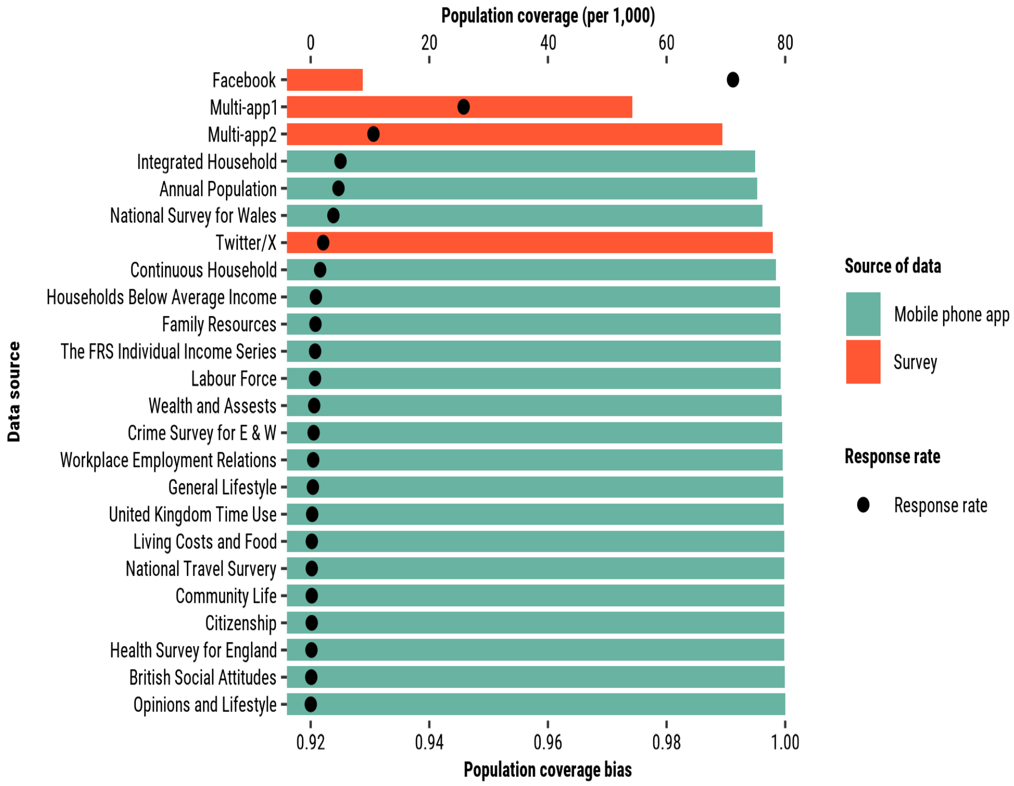


**Figure 2.** Size of population coverage bias (bottom x-axis) and population coverage per 1,000 population (top-x-axis) by data source.

and through time, and this is particularly valuable for studying dynamic social processes and render a richer representation of population dynamics.

We argue that we can infer some of these characteristics by leveraging the spatio-temporal granularity of MPD despite the absence of specific demographic information of the individuals from digital platforms. This is a necessary first step to understand which population groups might be over-represented or under-represented in different sources of MPD. Such information is necessary to develop subsequent data adjustment strategies that can improve the representativeness of the data relative to the target population.

## (b) Population coverage bias varies widely over space

To examine population coverage bias at subnational levels, we leveraged the fine-grained geographic resolution of the MPD. Analysing the distribution of population coverage bias across geographies allowed us to identify uneven patterns of population coverage across areas and assess whether these patterns exhibit spatial clustering. These analyses are important for evaluating the limitations for the use of each individual dataset in further research, and for deciding whether spatial dependence terms should be incorporated into the subsequent modelling stage of our framework.

Figure 3 summarises these analyses for each dataset. Each row corresponds to a data source and includes three elements. The first is a hexagonal cartogram displaying the size of population coverage bias in each LAD, coupled with the Moran's I statistic and corresponding p-value as a measure of spatial autocorrelation. The second is a histogram showing the distribution of

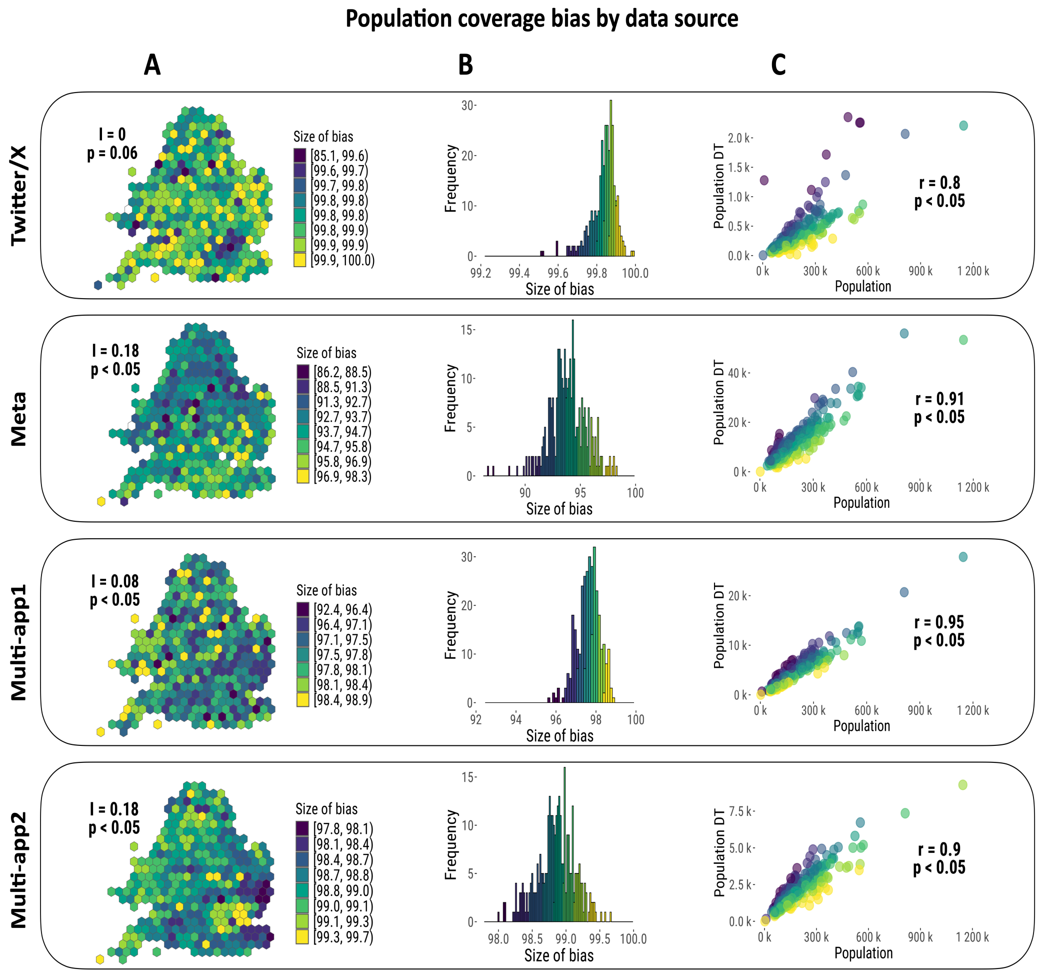


**Figure 3.** Extent and spatial distribution of population bias across local authority districts in the UK. A. Hexagon cartogram of the size of population bias across local authority districts. $I$ represents the Moran's I and $p$ denotes the associated p-value. B. Histogram of the distribution of the size of population bias. C. Scatter plot of the relationship between population size and size of bias.

population coverage bias across LADs. The third displays 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 three consistent patterns in the geographic distribution of population coverage bias which are consistent across MPD sources. First, all data sources exhibit noticeable geographic variability in population coverage bias, as evidenced by the spread of values across LADs in their respective distributions, as displayed in the cartograms and histograms. This result is consistent with recent findings for urban areas in the US [47].

Second, despite the variability in coverage bias, their geographic distribution does not display any strong patterns of geographic clustering around specific areas. Bias values fluctuate across longitude and latitude. Unexpectedly, there are no strong north–south or east–west gradients, as shown in the cartograms. 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. This finding indicates that we do not need to include a spatial dependence term (e.g. spatial lag) in the subsequent modelling stage of our framework. Yet, we warn that this may be different if our methodology is applied to other geographies.

Third, the variability in population coverage bias does not seem to a function of absolute population size. Instead, the association between MPD-derived and census population counts derived appears to be strongly linear, suggesting that population estimates from these two sources are proportionally similar and display similar geographic patterns. That is, LDAs with large census population counts also show large MPD-derived population counts. These findings were also quantitatively supported by Pearson correlation coefficients consistently close to one and statistically significant.

Despite these consistencies, we also observed differences across datasets. Focusing on the histograms, the range of values is larger for single-app than for multi-app data sources, indicating more variability in population coverage bias across LADs. Particularly, we found that Twitter/X data show the largest range of values: the lowest and highest across the four datasets, although it displays a higher concentration of population coverage bias at the higher end of the distribution of values. In particular, a large share of Twitter/X-based coverage bias scores are above 99 ($\approx 1\%$ population coverage). A notable exception is the LAD for City of London, where population coverage bias is low 85.1 ($\approx 15\%$ population coverage) since a large number of home locations (1,279) are detected relative to the resident population according to the Census (8,583). 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 employment and commuting hub. Because home locations are inferred from the most frequently recorded nighttime position, even a relatively modest number of residents or temporary visitors can yield an unusually high apparent population coverage compared to larger LADs.

Population coverage bias in data from Meta ranges from 86.2 ($\approx 14\%$ population coverage) to 98.3 ($\approx 2\%$ population coverage). Although the minimum value is higher than that observed for Twitter/X, the distribution is centred on lower values of bias, indicating that Meta data generally provide greater population coverage. The corresponding cartogram reveals a slight gradient, with lower bias estimates across LADs in Northern England. However, the associated Moran's I is small ($I = 0.18$) and statistically significant ($p < 0.05$), suggesting that there is some level of spatial autocorrelation or clustering, but this is quite minor.

In contrast, Multi-app1 and Multi-app2 data display a smaller range for the values of population coverage bias. This indicates that although population coverage bias for Multi-app 1 and 2 is generally higher than for Meta, it is also more uniform across areas. Particularly, Multi-app1 data show levels of population coverage bias as low as 92.4 for some LADs ($\approx 8\%$ population coverage), while Multi-app2 generally displays higher bias with the lowest being 97.8 ($\approx 2\%$ population coverage). Based on the associated cartograms and Moran's I, Multi-app1 and Multip-app 2 also show low levels of clustering in the spatial distribution of population coverage bias across LADs.

Taken together, these findings suggest that there is spatial variability in population coverage bias across LADs, but this variability does not seem to be significantly explained by physical geographic location or population size. Instead, the observed patterns are more likely explained by a wider range of area-level characteristics, including demographic or socioeconomic structures. We next examine the influence of these and a wider set of factors, to identify key contextual features shaping the observed levels of coverage bias. Such analysis can thus inform potential mitigation strategies to correct biases in a subsequent stage.

## (c) Key contextual features explain population coverage bias

We assessed the contribution of key contextual features to explaining spatial variations in population coverage bias across demographic, socioeconomic, resource accessibility, mobility and geographic domains. Figure 4 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. SHAP values are standardised by

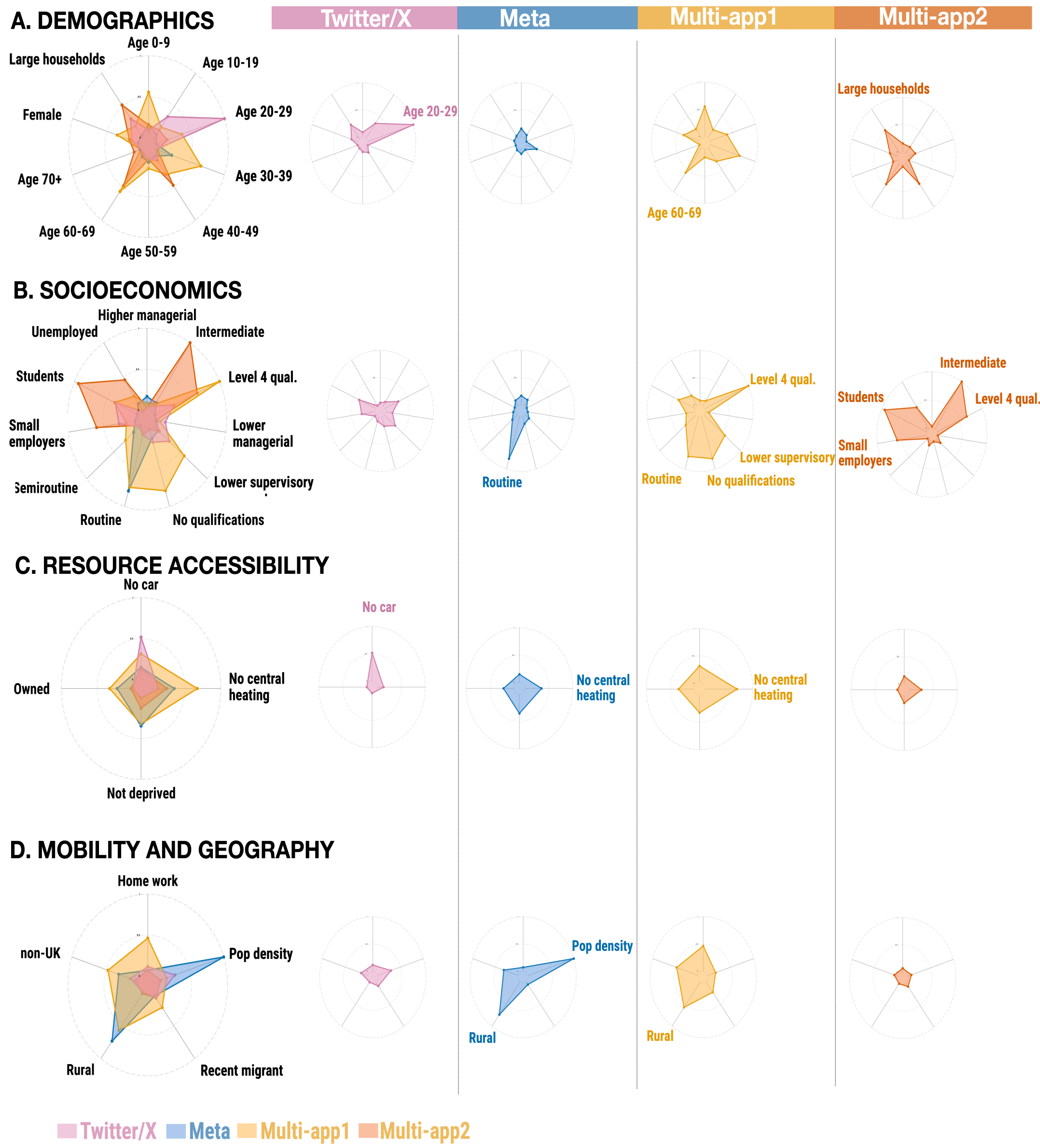


**Figure 4. Radial charts illustrating the importance of, and contribution to, explaining differences in population coverage bias across local authority districts.** The importance is estimated based on SHAP feature importance scores, calculated as the average absolute SHAP value per feature using an 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.

data source based on minimum and maximum scores to enable comparability. Figure 4 highlights features with standardised importance scores over 0.5 for individual data source.

The results reveal wide variability in the key predictors of population coverage bias across digital data sources. Demographic and socioeconomic features appear as the most important predictors of population coverage 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 coverage bias 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 coverage bias. Socioeconomic features standout as the most important predictors of population coverage bias 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 coverage bias 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 5 depicts the way these contextual features contribute to increasing or reducing the extent of population coverage bias across LADs. Figure 5.a shows the top 20 features based on their average SHAP value from the highest to the lowest. Figure 5.b displays top six of these features revealing how changes in feature values contribute to changes in population coverage 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 coverage 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 coverage bias, potentially reflecting a limited number of active Twitter/X users in this age range. Of the features highlighted in Figure 4, Figure 5 indicates that population estimates 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 coverage 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 5 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 coverage bias 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 coverage 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 coverage bias 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 coverage 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 coverage bias would

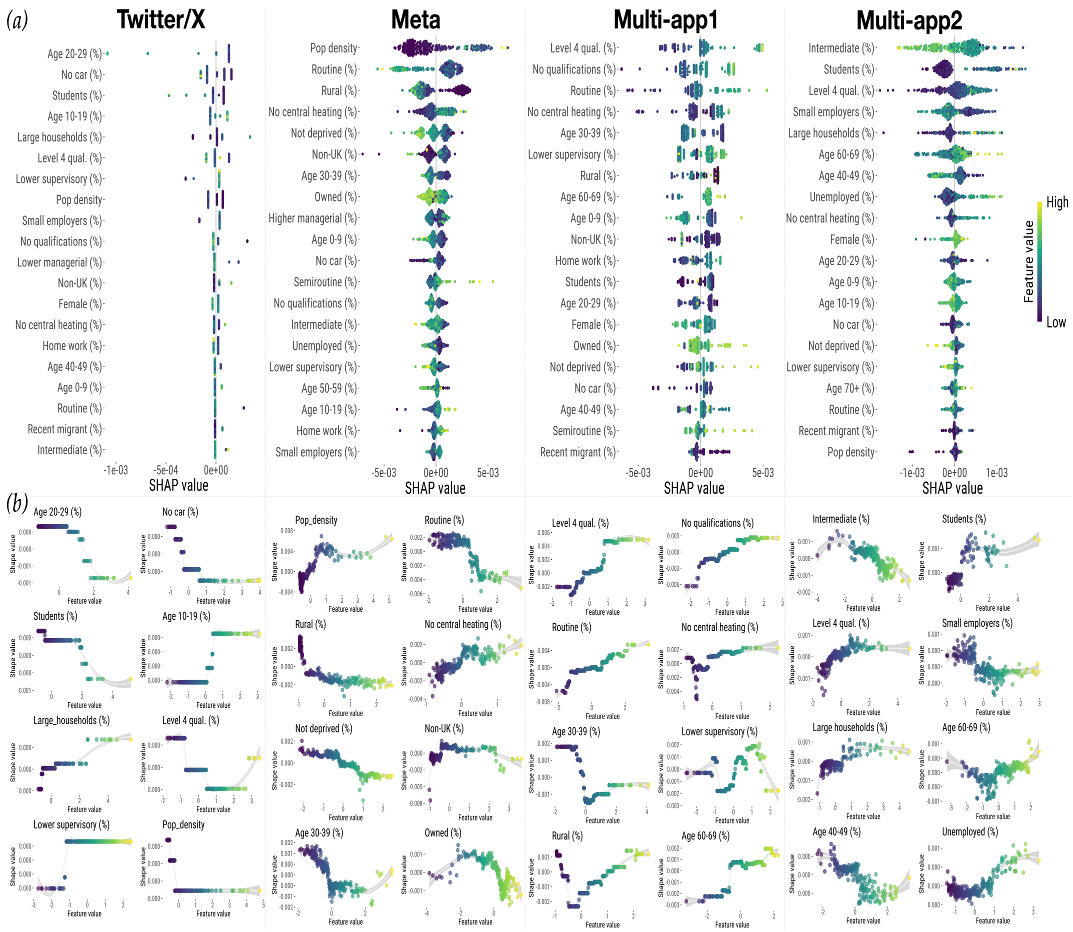


**Figure 5. Top ranked model features contributing to explaining population coverage bias across local authority districts from an XGboost model.** (a) beeswarm plots of SHAP feature values displaying the relative importance and direction of influence of the top 20 contextual features on the extent of population coverage 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 coverage bias. Local polynomial regression modelling was used to represent local trends with 95% confidence intervals. Each point represents a local authority district.

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

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 population coverage biases in MPD-derived population data. In this study, we sought to contribute a systematic and generalisable approach to measure and analyse population coverage bias 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 population coverage bias varies markedly across data sources and subnational areas, that multi-app data sources do 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 population 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.

These findings carry wide-ranging implications for the use of MPD in research and official statistics. First, the framework that we propose offers a replicable approach for evaluating the quality of MP-derived population counts when information on population characteristics is unavailable, establishing a baseline for transparent quality assessment. A key contribution is that it allows researchers and statistical agencies to move beyond anecdotal assumptions about representativeness and instead adopt a systematic, quantitative measure of bias that can be applied consistently across data sources and geographies. In practice, this means that MP datasets can be assessed before they are used in substantive analyses, enabling the identification of contexts where the data are reliable and those where caution is warranted. By being source-agnostic and privacy-preserving, the framework provides a common standard of evaluation that can be integrated into official workflows for producing population estimates, mobility indicators or crisis response statistics. In doing so, it helps bridge the gap between innovative but often opaque digital trace data and the established norms of transparency and accountability that underpin official statistics [33].

Second, the evidence provided here underscores the need for source-specific adjustment strategies to mitigate data biases. Single-app datasets, such as those derived from social media platforms, tend to display biases that are largely concentrated around a narrower set of characteristics, most notably age profiles, patterns of digital engagement, and certain geographic features. These biases, while still consequential, are comparatively easier to model and adjust for through established weighting or calibration techniques. Our evidence goes some way in validating the use of existing simple re-weighting scheme to mitigate biases in social media data based on age and sex structures. By contrast, multi-app datasets, which aggregate signals from a diverse set of platforms, capture broader segments of the population but at the cost of introducing more complex and multidimensional biases. These include systematic under- or over-representation linked to age and geography, and also to education, occupation, income and household structure. As a result, correcting multi-app data requires adjustment strategies that go beyond conventional post-stratification, demanding more sophisticated statistical tools that can simultaneously account for multiple intersecting dimensions of inequality. Such tools may include multilevel reweighting schemes, Bayesian hierarchical modelling [48] , data-fusion [10] or difference-in-difference approaches [49] that integrate auxiliary information from surveys or administrative records. Such distinction suggests that efforts to improve the representativeness of MP-derived statistics should be tailored to the nature of the data source. That is to recognise that multi-app datasets, while promising in their breadth, pose greater methodological challenges and cannot be assumed to be inherently less biased than single-app data.

Third, 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. Digital engagement is deeply patterned by age, gender, socioeconomic status and geography, meaning that the traces people leave behind are systematically unequal. Treating these data as representative without adjustment risks reproducing and amplifying existing inequalities in the evidence base used for public policy. By contrast, explicitly recognising the partial and selective nature of MPD allows researchers to design strategies to correct for these divides, improving both validity and fairness. This is particularly important in applications with direct policy relevance, such as crisis response, public health monitoring, transport planning and infrastructure investment, where decisions informed by biased data could inadvertently exclude or disadvantage already vulnerable groups. Ensuring that MPD are used critically and responsibly therefore requires integrating bias assessments and corrections into routine analytical practice, thereby maximising their potential to contribute to equitable decision-making.

Despite these contributions, our study has limitations that should be acknowledged. First, although we benchmarked against the 2021 UK census, not all countries have access to such high-quality and up-to-date reference data. In many regions, particularly in Africa censuses are infrequent, often collected only every five to ten years, and sometimes of variable quality [50]. To enhance the transferability of our framework, future work could draw on global gridded population datasets such as WorldPop [51], which provide continuous, spatially explicit population estimates that may serve as a useful alternative benchmark, and a large set of covariates to inform spatial analysis of bias source identification. Second, while our analysis is based on aggregated counts, this does not necessarily restrict the assessment of demographic-specific biases. MPD rarely contain direct information on individuals' characteristics, meaning that traditional demographic benchmarking is not feasible based on consistent information from MDP sources. Our approach therefore represents a key advantage, as it provides a systematic way to interrogate and inform adjustment methods for bias precisely in the absence of demographic attributes. Third, the reliance on area-based covariates introduces the risk of ecological fallacy, as heterogeneity within local units may be obscured. This highlights the need for more spatially granular analyses where data availability permits. Such analysis is possible if access to more granular spatial MPD is available. Finally, we examined four datasets. Further research is required to assess whether the patterns we document hold across additional providers, different time periods and diverse national contexts.

Future research should build on our proposed approach in at least three directions. First, applications in low- and middle-income countries, where mobile penetration is uneven, are crucial to assess how bias manifests in settings of greater digital inequality. Second, comparative work across datasets and providers is needed to identify which sources are most representative under different conditions and to establish standards of validation and robustness. Third, integrating our approach with emerging adjustment methods — such as data fusion with household surveys, imputation techniques and post-stratification weighting — offers a pathway to correct biases and improve representativeness. Advances in trusted research environments and synthetic data generation may also provide new avenues to interrogate raw data more transparently while safeguarding privacy. Taken together, these efforts will be vital to ensure that MPD can be harnessed as a timely, granular and adaptive complement to censuses and surveys, enhancing the capacity of researchers and policymakers to monitor population distributions in an era of accelerating social and environmental change.

## 6. Conclusion

This paper has introduced a systematic approach to quantify and explain population coverage and representativeness biases in mobile phone application data when information on population attributes is absent. Applying the framework to four datasets for the UK, we presented evidence that MPD can achieve higher overall coverage than traditional surveys, yet substantial biases persist across sources and subnational areas. We showed that these biases are systematically associated with demographic, socioeconomic and geographic characteristics, often in complex nonlinear ways, and that multi-app datasets, while broader in coverage, present more complex patterns of bias than single-app data. By establishing a transparent and replicable approach, our study provides a foundation for assessing the quality of digital trace data and offers guidance for designing adjustment strategies to improve their representativeness. In doing so, we take a step towards enabling MPD to complement censuses and surveys as reliable sources for research and policy. Realising this potential will require continued validation across diverse contexts, the integration of bias-adjustment techniques, and collaborative efforts between data providers, researchers and statistical agencies to ensure that the societal benefits of MPD are equitably realised.

Ethics. This research was conducted in accordance with the ethical standards of the University of Liverpool. All data used were anonymised and analysed at an aggregate level. No identifiable personal data were accessed or processed by the authors.

Data Accessibility. Please see data accessibility in the Data section. Code is openly available at https://github.com/de-bias/bias-detection.

Authors' Contributions. Both authors contributed equally.

Competing Interests. The authors declare that they have no conflicts of interest relevant to this work.

Funding. This work was supported by the Economic and Social Research Council (ESRC) through Smart Data Research UK programme [Project reference: ES/Y010787/1].

Disclaimer. Please provide disclaimer text here.

Acknowledgements. The authors acknowledge the support of UK ESRC and of project partners: the International Organization for Migration (IOM), Meta Platforms, Inc., and Prof. Esteban Moro.

## 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. De Heer W, De Leeuw E. 2002 Trends in household survey nonresponse: A longitudinal and international comparison. *Survey nonresponse* **41**, 41–54.
5. Stedman RC, Connelly NA, Heberlein TA, Decker DJ, Allred SB. 2019 The End of the (Research) World As We Know It? Understanding and Coping With Declining Response Rates to Mail Surveys. *Society 'I&' Natural Resources* **32**, 1139–1154.
6. Luiten A, Hox J, de Leeuw E. 2020 Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics* **36**, 469–487.
7. Kitchin R. 2014 *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.
8. United Nations Statistics Division. 2025 Mobile Phone Data Task Team. https://unstats.un.org/bigdata/task-teams/mobile-phone/index.cshtml. Accessed: 29 August 2025.
9. UK Research and Innovation. 2025 Smart Data Research UK (SDR-UK). https://www.sdruk.ukri.org/. Accessed: 29 August 2025.
10. Graells-Garrido E, Opitz D, Rowe F, Arriagada J. 2023 A data fusion approach with mobile phone data for updating travel survey-based mode split estimates. *Transportation Research Part C: Emerging Technologies* **155**, 104285.
11. Ballantyne P, Singleton A, Dolega L. 2022 Using unstable data from mobile phone applications to examine recent trajectories of retail centre recovery. *Urban Informatics* **1**.
12. Hunter RF, Garcia L, de Sa TH, Zapata-Diomedi B, Millett C, Woodcock J, Pentland A, Moro E. 2021 Effect of COVID-19 response policies on walking behavior in US cities. *Nature Communications* **12**.
13. 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**.
14. Graells-Garrido E, Serra-Burriel F, Rowe F, Cucchietti FM, Reyes P. 2021 A city of cities: Measuring how 15-minutes urban accessibility shapes human mobility in Barcelona. *PLOS ONE* **16**, e0250080.

15. Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E. 2011 pp. 377–388. In *Prediction of Socioeconomic Levels Using Cell Phone Records*, pp. 377–388. Springer Berlin Heidelberg.
16. Blumenstock J, Cadamuro G, On R. 2015 Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076.
17. Moro E, Calacci D, Dong X, Pentland A. 2021 Mobility patterns are associated with experienced income segregation in large US cities. *Nature Communications* **12**.
18. Raun J, Ahas R, Tiru M. 2016 Measuring tourism destinations using mobile tracking data. *Tourism Management* **57**, 202–212.
19. 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).
20. González-Leonardo M, Cabrera C, Neville R, Nasuto A, Rowe F. 2025 Beyond the Immediate Impacts of COVID-19 on Internal Population Movements in Mexico: Facebook Data Reveal Urban Decay and Slow Recovery—A Research Note. *Demography*.
21. Rowe F, Calafiore A, Arribas-Bel D, Samardzhiev K, Fleischmann M. 2022 Urban exodus? Understanding human mobility in Britain during the COVID-19 pandemic using Meta-Facebook data. *Population, Space and Place* **29**.
22. Cabrera C, González-Leonardo M, Nasuto A, Neville R, Rowe F. 2025 Sustained changes to urban mobility after COVID-19 amplified socio-economic inequalities in Latin America. .
23. 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.
24. Porter G, Hampshire K, Abane A, Munthali A, Robson E, Mashiri M, Tanle A. 2012 Youth, mobility and mobile phones in Africa: findings from a three-country study. *Information Technology for Development* **18**, 145–162.
25. Ofcom. 2023 Communications Market Report 2023. Accessed: Nov 2023.
26. Statista. 2024 Social Media & User-Generated Content - Market Overview. Accessed: 2024-11-14.
27. Blumenstock J, Eagle N. 2010 Mobile divides. *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development* pp. 1–10.
28. Schlosser F, Sekara V, Brockmann D, Garcia-Herranz M. 2021 Biases in human mobility data impact epidemic modeling. .
29. Rowe F, Neville R, González-Leonardo M. 2022 Sensing Population Displacement from Ukraine Using Facebook Data: Potential Impacts and Settlement Areas. *OSF Preprints*. Submitted.
30. Ribeiro FN, Benevenuto F, Zagheni E. 2020 How Biased is the Population of Facebook Users? Comparing the Demographics of Facebook Users with Census Data to Generate Correction Factors. In *Proceedings of the 12th ACM Conference on Web Science* WebSci '20 p. 325334 New York, NY, USA. Association for Computing Machinery.
31. Zagheni E, Weber I. 2015 Demographic research with non-representative internet data. *International Journal of Manpower* **36**, 13–25.
32. Gil-Clavel S, Zagheni E. 2019 Demographic Differentials in Facebook Usage around the World. *Proceedings of the International AAAI Conference on Web and Social Media* **13**, 647–650.
33. Rowe F, Magpantay E, Jalagonia M, Esko S, De Jesus E, Jansen R, Grum F, Blanchard PB, Lumala W. 2025 Mobile Phone Data for Migration Statistics (MPDMS) Guide. https://unstats.un.org/wiki/spaces/MPDMS/overview. Accessed: 29 August 2025.
34. Barreras F, Watts D. 2024 The exciting potential and daunting challenge of using GPS human-mobility data for epidemic modeling. *Nature Computational Science* **4**, 398–411.
35. UK Government. 2025 Population Movement Data: Unlocking Opportunities for a Modern Digital Government. https://www.gov.uk/government/publications/population-movement-data-unlocking-opportunities-for-a-modern-digital-government/population-movement-data-unlocking-opportunities-for-a-modern-digital-government. Accessed: 29 August 2025.

36. Maas P. 2019 Facebook Disaster Maps. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
37. Microsoft. 2023 Bing Maps Tile System. Accessed on 7 August 2025.
38. 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**.
39. 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.
40. Rey S, Arribas-Bel D, Wolf LJ. 2023 Geographic Data Science with Python. .
41. Friedman JH. 2001 Greedy function approximation: A gradient boosting machine.. *The Annals of Statistics* **29**.
42. Chen T, Guestrin C. 2016 XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794.
43. Nielsen D. 2016 Tree boosting with xgboost-why does xgboost win" every" machine learning competition?. Master's thesis NTNU.
44. UK Data Service UK surveys - UK Data Service. https://ukdataservice.ac.uk/help/data-types/uk-surveys/. [Accessed 17-02-2025].
45. Lohr SL. 2021 Sampling. .
46. Cochran W. 1977 *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley.
47. den Nijs K, Omodei E, Sekara V. 2025 Data Bias in Human Mobility is a Universal Phenomenon but is Highly Location-specific. .
48. Rampazzo F, Bijak J, Vitali A, Weber I, Zagheni E. 2021 A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom. *Demography* **58**, 2193–2218.
49. Zagheni E, Weber I. 2012 You are where you e-mail. *Proceedings of the 4th Annual ACM Web Science Conference* pp. 348–351.
50. Dindi PD, Stiegler N. 2025 Charting the future of censuses: Insights, lessons and recommendations for the 2030 round. *Statistical Journal of the IAOS* **41**, 402–420.
51. Tatem AJ. 2017 WorldPop, open data for spatial demography. *Scientific Data* **4**.