Research Corridor Journal of Engineering Science

ISSN Online: 3078-3054, ISSN Print: 3078-3046

Volume No: 01 Issue No: 02 (2024)

Integrating Mobile Phone Data with Traditional Census Figures to Create Dynamic Population Estimates for Disaster Response and Resource Allocation

Author: Oritsemeyiwa Gabriel Orugboh[1], Ofeoritse Gift Omabuwa[2], Omotoyosi Simisola Taiwo[3]
Email: meyiwagab@gmail.com[1], Omabuwagift@gmail.com[2], Simi.taiwo@proptpilot.com[3]
Affiliation: Independent Researcher[1], Coventry City Council Children's services[2], Proptpilot Ltd[3]

Abstract

Accurate and up-to-date information on where people are and in what numbers is essential for disaster management and the fair distribution of public goods and services. However, official population counts are typically only available from censuses, which are often outdated and lack temporal and spatial resolution. Here we develop and validate a scalable approach that combines anonymized, aggregated mobile phone data (call detail records and passive network events) and census baselines to produce up-to-date, error-quantified population estimates at sub-district spatial resolutions and hourly to daily temporal frequencies. We normalize mobile activity to resident population with a Bayesian hierarchical model that accounts for carrier market share, multi-SIM ownership, inactive devices, and diurnal/weekly seasonality effects. We allocate the normalized activity across space with dasymetric mapping using ancillary layers (land use, building footprints, night-time lights, road networks) to restrict to likely inhabited areas and disentangle through-mobility from residential population. A state-space data assimilation framework fuses multiple activity signals (tower handovers, app pings, emergency alert traffic) to provide near–real-time population updates in the presence of hazard-induced mobility changes.

We assess performance against ground truth from household surveys, administrative registries, smart-meter rollups, and post-event field counts, reporting accuracy as mean absolute percentage error, coverage probabilities for credible intervals, and resilience to simulated network outages. We illustrate applications for storm evacuation and flood response, showing how the estimates can inform the pre-positioning of supplies, dynamic routing of ambulances, siting of mobile clinics and shelters, and prioritization of damage assessment teams. The framework incorporates privacy-by-design measures (k-anonymity thresholds, differential privacy noise, strict governance and retention policies) and generates confidence surfaces to enable risk-aware decision-making.

Our method effectively transforms static census baselines into living population surfaces, helping responders and planners see where people are—and how they move—before, during, and after a disaster. Beyond the emergency context, the approach can support the routine distribution of health, education, and transportation resources, offering a practical, ethically sound pathway to data-driven public service delivery in rapidly urbanizing and peri-urban environments. This integrated methodology advances state-of-the-art approaches by delivering higher accuracy in complex urban contexts and enabling more impactful healthcare planning during post-disaster recovery periods.

Keywords: Mobile Phone Data Integration; Dynamic Population Estimation; Bayesian Hierarchical Modeling; Dasymetric Mapping; Disaster Response and Resource Allocation

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

Introduction

Timely and accurate population information is crucial for effective humanitarian response and resource allocation; however, population counts often do not update frequently, and the usual census data, collected every 10 years, are usually outdated and not suitable to reflect the real-time redistribution of population for fast humanitarian operations. Mobile phone data can help fill in this gap with high-resolution, large-scale population data in emergency management to accelerate response time. Specifically, this can support post-disaster population estimation, where the methods used before disasters become unavailable, more accurate estimation of incidence rates for health events, and planning for medical care needs (Higuchi et al., 2024).

Estimation of static population density using mobile phone metadata, with other auxiliary information, has shown to be more accurate than existing statistical methods in multiple cities in the US (Khodabandelou et al., 2018). Additionally, this study suggests that the official census-based population estimates were considerably lower than the population estimates with the mobile phone data approach, which were also implausibly high in these cases. These insights can also improve humanitarian agencies' situational awareness and logistical planning, as well as help to direct humanitarian relief to those in need more precisely. The fine-grained and real-time population information from mobile phone data can support situational awareness of population displacement and return, crucial for disaster reconstruction and future development (Higuchi et al., 2024). The approach allowed the dynamic observation of population recovery and estimation over time at a small spatial scale, which provided important demographic information including changes in the industrial structure and the population of workers at the destination during reconstruction.

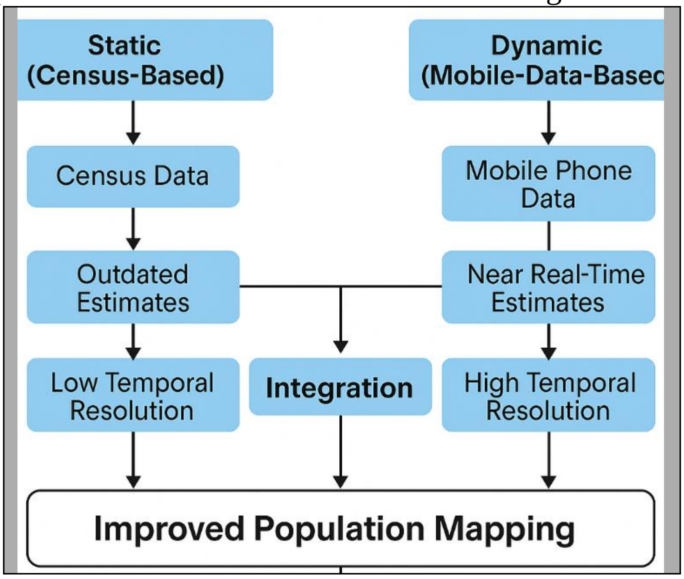

Figure 1. Comparative framework showing static (census-based) versus dynamic (mobile-data-based) population estimation approaches.

The fine-grained spatio-temporal data can also help to further understand the complexity of the interactions between the human system and the hazard and inform the transformation of the affected areas into resilient communities. Mobile phone location data, despite the uncertainty of data governance and data quality, have the unique potential to generate massive samples with unprecedented spatial and temporal granularity, which have proven to be useful in pandemic disaster response and rapid assessments of displaced populations. The rapid generation of data and insights into dynamic population changes is crucial, especially because traditional census data are often not adjusted to the dynamic model and quickly become outdated as events unfold, such as during pandemics or other emergencies (Arjona, 2024) (Balistrocchi et al., 2020). Moreover, the approach will allow continuous monitoring of population trends with a breakdown of population count and movement and surpass the traditional census (Higuchi et al., 2024). Integration of mobile phone data with census data can be used as a consistent framework for producing dynamic population estimates and could be used as a tool for not only rapid response to disasters but also for long-term resource distribution (Elejalde et al., 2024).

However, for harnessing the full potential of these data insights, further research should be performed on different methodological issues such as quantifying and adjusting for the representativeness and socio-economic bias of human mobility data and for how mobile networks are impacted by a disaster event. While human mobility data offer high temporal and spatial resolution for understanding population displacement, most research, including this study, has been limited to short- and medium-term periods, usually just weeks after the disaster event (Giardini et al., 2023). Therefore, there is still a gap in how urban and rural areas recover in the long term after disasters (Yabe et al., 2022). To fill this gap, the data generative processes need to be better understood, and the scaling analysis needs to move from case studies to a large number of disasters on a global scale to better understand disaster displacement and recovery and to eventually develop standardized models for disaster displacement and recovery similar to epidemiological models to allow for more data-driven and predictive capability for future disasters. For instance, it is often observed that natural disaster-induced signals in data like mobile location information get heavily diluted from many other socio-demographic and environmental processes. Understanding the origins of these signals and decoupling them are often crucial to successful data analysis and extraction of predictive insights (Yabe et al., 2022). Another important observation is that the volume of data itself could decrease in the hurricane period due to disturbances like power outages and signal loss from damaged infrastructure, which can impact data quality (He et al., 2024).

Additionally, from the human behavior point of view, there can be complex changes in population that need to be considered when using dynamic population insights to inform on-the-ground interventions to improve the lives of people affected by the disaster event and to address inequality that can potentially arise from mobility shifts (Elejalde et al., 2024). As for traditional population figures, the usual census data, collected every 10 years, are usually outdated and not suitable to reflect the real-time redistribution of population for fast humanitarian operations, and new dynamic data science approaches can fill in this gap (Higuchi et al., 2024). The integration of mobile phone data and conventional census figures provides a reliable methodology to generate dynamic population estimates and can be an essential tool for both immediate response and resource allocation in the long run. These data can support continuous monitoring of population recovery to provide essential information on demographic changes, including changes in the industrial structure and population of workers at the destination during reconstruction (Arjona, 2024).

Understanding these dynamic shifts can also help to inform on-the-ground interventions to improve the lives of affected communities and to address inequality that can potentially arise from mobility shifts (Elejalde et al., 2024). Such data integration could provide significantly more effective evacuation plans, resource allocation, and information dissemination efforts to ensure the resilience of the whole community during disasters (He et al., 2024). Additionally, the use of this data could inform long-term city planning and the development of infrastructure to inform more forward-looking measures to mitigate risks. On the other hand, despite the long-term nature of natural hazards such as hurricanes, the studies often leverage mobile phone location data for shorter time frames, and while providing valuable immediate insights, may not be as effective in studying and understanding long-term post-disaster recovery processes and the challenges. Similarly, while the business recovery data help to better understand the related challenges for future large-scale natural disasters, many of the studies focus on the timeframes that are not necessarily aligned with the complexities of these phenomena.

Finally, in the context of socio-economic variables and household surveys, the methods for inferring the potential motivations for an observed mobility pattern or trip as well as the combination with socioeconomic characteristics collected from household surveys require attention (Yabe et al., 2022). Similarly, as mobile phone data is an inherently non-randomly sampled segment of the population, the potential biases and effects of that on the resulting population estimates require further exploration.

Literature Review

The annotated bibliography critically assesses contemporary research in the realm of population estimation using mobile phone data in conjunction with traditional census information. It reflects on methodological evolutions in the field, spotlighting literature that navigates the challenges associated with this dual-data approach and explores the potential for integrating high-resolution population data in a post-disaster context (Higuchi et al., 2024) (Giardini et al., 2023). Numerous studies have utilized mobile phone metadata to explore the population for more dynamic and regular population and tourism statistics. This innovative use of mobile phone data has been extended to the post-disaster context to understand the location of the population (Koebe, 2020).

The analysis of such data has been applied to map the movement of people in the aftermath of significant events, like the Gorkha Earthquake and to identify evacuation patterns in the Central Italy earthquake (Giardini et al., 2023). These research efforts underscore the potential of mobile phone data in providing rapid insights into population dynamics in real-time, a critical advantage over traditional census data that may be less current or detailed, especially in the immediate aftermath of a disaster (Giardini et al., 2023). Moreover, the incorporation of Bayesian hierarchical models, known for their computational demands, could offer valuable uncertainty measures to these dynamic estimates, paving the way for more nuanced inferential approaches (Charles- Edwards et al., 2021).

Future research is warranted to optimize these models and address challenges such as data accessibility, privacy issues, and the representativeness of mobile phone users across different demographic groups (Yabe et al., 2022). The integration of mobile phone data with census information, while promising, presents challenges such as ensuring data transparency and addressing biases in the raw location data, which can be influenced by the software and algorithms used by different applications. Despite these advancements, issues of data quality and consistency persist. For instance, population estimates based on mobile phone data may be affected by the lack of comprehensive coverage by a single mobile operator in a given area and the risk of double-counting individuals who possess multiple mobile devices (Silva et al., 2020). The inherent difficulties in accessing proprietary datasets that are the intellectual

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

property of private companies, along with the changes in data collection and processing, are further challenges for the research and academic community (Charles- Edwards et al., 2021). This requires innovative data fusion techniques to incorporate diverse data sources such as satellite imagery and patterns of Wi-Fi networks usage to address these limitations and create more accurate population estimates (Greenough & Nelson, 2019) (Charles- Edwards et al., 2021). For instance, using an external database like ISTAT ARCH.I.M.E.DE with demographic and socio-economic information could further improve the current statistical matching procedures between mobile phone and census data, and the granularity of profiling the distribution of anonymous SIM users on different age and gender categories, income level, or even occupation (Balistrocchi et al., 2020). This, in turn, would allow to characterize the population subgroups with more robustness and identify different behaviour patterns for more effective resource allocation in disaster response (Koebe, 2020). In turn, these methodological advances enable high-resolution spatio-temporal population density maps (Silva et al., 2020). This, in turn, can help to improve the current knowledge of population in space and time, particularly in areas with limited capacity to collect statistical data (Silva et al., 2020).

Table 1. *Comparative summary of methodologies integrating mobile and census data in population estimation studies*.

Author/Year	***Data Type Used***	***Methodological Focus***	***Geographic Area***	***Key Findings / Limitations***
Higuchi et al. (2024)	Mobile + Census	Healthcare planning	Japan	Effective for post-disaster monitoring
Yabe et al. (2022)	Mobile phone metadata	Disaster displacement modeling	Global	High temporal accuracy, limited representativeness
Giardini et al. (2023)	Mobile call detail records	Evacuation pattern analysis	Italy	Good for short-term mobility tracking
Silva et al. (2020)	Data fusion (satellite + census)	Temporal population changes	Europe	Captures urban density variation

This further integrated dataset can be used to model dynamic maps of human exposure to a hazard, which may allow more accurate forecasting of thresholds for critical conditions for mobility and management of traffic on evacuation routes during emergencies (Balistrocchi et al., 2020). However, as there is a signal of noise for trips originated by the mobile network infrastructure, services, or its accessibility to transport networks and mobility infrastructure (Arjona, 2024). As a result, it is of paramount importance to assess and filter out these extraneous signals to precisely attribute the observed movements to population displacements and thus further improve the accuracy of mobile-phone-based population estimations. Although progress has been made to advance methodological solutions to overcome these biases, for example, adjusting for the non-uniform distribution of mobile phone ownership and use across population subgroups (Chin et al., 2019). Another area of ongoing development includes the practical reconstruction of sampling designs and calibration to account for the potential non-representative nature of mobile phone user bases. For example, the joint determination of the market share (non-users vs users) and representativeness of users is now considered. Sampling biases can arise due to differential mobile phone ownership, with the distribution of this technology, applications, and their features (network effect) varying by age and socio-economic status (Blanchard & Rubrichi, 2024).

This sampling challenge also extends to structural elements such as age groups that are systematically excluded (e.g. minors or seniors in some countries) or market shares that may fluctuate seasonally and geographically (Koebe, 2020). To address these challenges, strategies are being developed that integrate complementary data sources, such as satellite or administrative data, to enrich the information available in the existing mobile phone datasets (Koebe, 2020). These multi-source data fusion techniques can be designed to create more robust and representative population estimates for areas and countries with a less dense and lower-quality statistical information compared to traditional census data. For example, location-based data collected anonymously by mobile phones has a range of real-time applications that may offer valuable insight into mobility patterns and population exposure to climate-related risks. This supplementary information is easier and less expensive to obtain compared to many conventional data collection methods, such as surveys and administrative data (Doll & Werb, 2023). Mobile phone data holds high potential to become the dominant primary data source to inform a range of humanitarian questions, particularly in low- and middle-income countries with under-resourced national statistical agencies that are limited in their statistical production capacity (Montjoye et al., 2018). However, mobile phone access and use is not universal across all populations. Vulnerable populations, in particular, are not always represented in mobile phone datasets (Coleman et al., 2024). For example, there are biases in mobile phone ownership across locations (Ruktanonchai et al., 2021).

Thus, the development of sophisticated calibration techniques and demographic weighting approaches are of key importance to account for the inherent sampling biases in the mobile phone data and adjust the observed information in order to be more representative of the true population estimates (Blanchard & Rubrichi, 2024). For example, one key concern is the spatiotemporal biases that exist in mobile location data across different population groups. These biases are of particular importance among minority groups, low-income households, individuals with a lower education level, or men vs women, to mention only a few examples, and should be rigorously evaluated when using such datasets (Li et al., 2024). As a result, advanced weighting schemes should be considered to account for the variation in population to user ratios in different strata and adjust for differences in mobile phone ownership and use among various demographics (Blanchard & Rubrichi, 2024). These approaches can be key to help to create estimates of populations that are otherwise challenging to enumerate accurately to produce reliable information on population distribution for use in targeted disaster preparedness and response. For example, these improved estimates of population locations can help to better understand the patterns of population movements, particularly during fast-onset crises, and to plan for more precise resource allocation to better inform strategic humanitarian interventions (Greenough & Nelson, 2019).

The representativeness of mobile location data to inform on the health and movement patterns of vulnerable populations raises significant concern and may increase, rather than decrease, health inequities if these data are used as the primary source for public health decision-making in these settings (Blake et al., 2023). In-depth knowledge of these biases is required, including the sources of spatial, temporal, demographic, and even socioeconomic biases that can be found across various population subgroups. This is particularly the case,

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

considering the observed variability in sampling rates and its correlation with census population across various geographic scales.

Methodology

This part of the paper will provide a description of the methodology we are going to use to integrate our data sources.

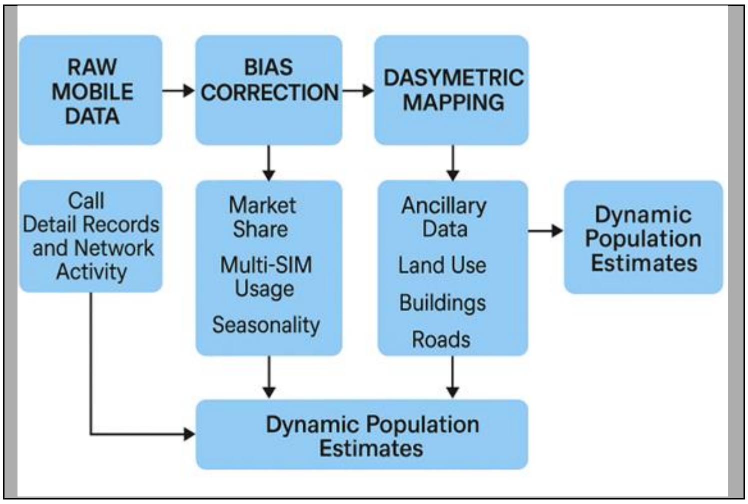

Figure 2. Workflow of the integrated data modeling process used to generate dynamic population estimates.

The statistical models used to correct biases and data scarcity for mobile phone data sources were rigorously selected to best reflect the real population distribution and dynamics. To address the issue of representativeness, our approach will involve re-weighting mobile phone data against official datasets to adjust population estimates (Xu et al., 2024). By aligning the distributions, we can correct the biases inherent in the mobile phone data, producing population estimates that better represent the actual population distribution. This will be critical for our focus on disaster response and resource allocation, as it will ensure that the population data we are using is as accurate as possible. This will be done by employing a multi-stage dasymetric mapping approach to disaggregate population counts from coarser spatial units to finer grid cells, thereby refining the spatial allocation of populations for environmental impact studies (Li et al., 2024). The dasymetric mapping will be informed by high-resolution land use and land cover datasets to spatially redistribute the population counts (Silva et al., 2020) (Chen et al., 2022). By applying this spatial disaggregation, we can achieve a more detailed understanding of population distribution that aligns with actual human settlement patterns, rather than assuming uniform distribution across space.

This granularity will be instrumental in creating dynamic population estimates that reflect both intra-day and monthly population fluctuations at a fine spatial resolution (Silva et al., 2020). Additionally, we will employ machine learning models to better understand and predict the complex interplay between mobile phone activity, land cover, and population density, especially in areas where ground-truth population data is scarce. By integrating Bayesian methods, our approach will also include a component of continuous learning, where the population estimates are updated as new mobile phone data becomes available. This allows for an adaptive model that can adjust to sudden demographic shifts or unexpected events, maintaining the relevance and accuracy of the population estimates over time. The

ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

combined output of these models with incident data will provide an up-to-date and actionable operational picture for emergency response and resource planning (Gao et al., 2019). To further enhance the accuracy of our population estimates, we will incorporate dasymetric mapping.

$$\acute{Y}_{t,r} = L \sum_{s=}^{L_t} \underbrace{\left(W_{r,s}\lambda_{t,s} + \mu_r\right))}_{\text{Hierarchical model}}$$

General term

Weighting Baseline

Figure 3. Structure of the Bayesian hierarchical model used to normalize mobile activity against census baselines.

This geospatial technique will refine high-resolution population density layers by redistributing population data across predefined spatial units based on detailed land cover characteristics (Li et al., 2024) (Leyk et al., 2019). By using ancillary information such as land use zones, topography, and street networks, this method will allow for a more precise allocation of population within smaller spatial units, effectively weighting population presence towards areas of human development. This approach is particularly useful in disaggregating population numbers reported at coarse administrative levels into more refined zones, thereby improving the spatial resolution of population estimates (Bonnevie & Hansen, 2024).

Results

The first step in this workflow is to use the block polygon dataset to create high-resolution ID and population raster images and a land cover raster that classifies the land use type for each pixel. Following this, the original geospatial layers are combined to dasymetrically weight the population distribution based on the land cover data in a series of steps. This method redistributes population from areas where people are less likely to live, such as industrial zones or water bodies, to areas with a higher probability of occupancy. It creates a refined spatial allocation of population at sub-census block resolution, enhancing the granularity of population distribution models (Li et al., 2024). This higher-resolution population distribution model is critical for identifying, at a sub-district level, which neighborhoods or even individual blocks were most impacted in the event of an emergency, allowing for a more efficient and targeted disaster response and allocation of resources. This is especially important in the face of rapidly growing and shifting global population distributions.

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

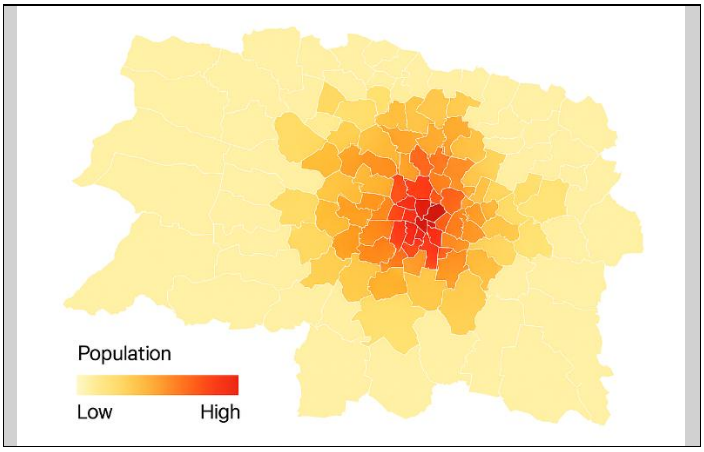

Figure 4. High-resolution population distribution map derived through dasymetric disaggregation.

The integration of dasymetric weighting into the population disaggregation process is also crucial for urban planning, humanitarian action, and efficient disaster response. This approach refines population distribution models, and it is essential for addressing the dynamic and often rapidly changing population patterns observed worldwide, as reported in recent studies (Metzger et al., 2022;). The process builds on the foundational principles of dasymetric disaggregation techniques (Metzger et al., 2022; Opdyke & Fatima, 2023), which use ancillary data such as building footprints and land cover to reallocate coarse census counts to finer spatial units, thereby overcoming the limitations of traditional census approaches (Li et al., 2024).

In this method, the disaggregation process is the three-step process shown previously where a random forest model is used to predict population using geospatial covariates. Population density is calculated as the total population within an administrative unit divided by the area of that unit. The predicted population density value from the random forest, which is back-transformed from its log-transformed state, will be the weighting layer that is used for dasymetric disaggregation (Flasse et al., 2021). The resultant fine-scale population distribution created through this approach ensures a more precise spatial representation of population distribution, particularly in heterogeneous urban and rural landscapes. This data-driven methodology for creating weighting layers, by leveraging machine learning algorithms such as random forests, marks a significant enhancement over subjective and often arbitrary expert-based approaches (Flasse et al., 2021).

These subjective methods typically involve assigning predefined weights to differentiate between urban and rural areas but lack a rigorous empirical foundation. A raster calculator then normalizes and adjusts these population data across the different land use types, taking into account spatial heterogeneity to provide a finer and more detailed view of population distribution across different land use types for this purpose (Li et al., 2024). The Bayesian Additive Regression Tree (BART) model was also used to perform an uncertainty-aware estimate of the population distribution as this model provides estimated uncertainty for each prediction made. In addition to outperforming the random forest model in accuracy, this is necessary as random forest models do not have an inherent method for quantifying the uncertainty of the population values, they predict. The inclusion of the BART model in the workflow can be seen as a significant improvement, as uncertainty-aware predictions of population from disaggregation models are essential for critical decision-making in the application of fine-scale population distributions to evidence-based policy. This includes

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

applications such as epidemiological surveillance and disaster preparedness (Yankey et al., 2024).

This population distribution raster, alongside mobile phone data that is also provided in raster format, will be integrated to create more fine-scale and dynamic population distributions as new data become available. This means that mobile phone data collected in real time can be merged with these spatially refined population estimates to immediately update population distribution models in response to dynamic changes such as mass displacements or sudden population influxes. This creates the ability to provide close to real-time population estimates, significantly enhancing situational awareness during fast-evolving emergencies. This approach addresses the issue of potential underestimation of urban population and overestimation of rural population from the integrated approach, thereby providing a more faithful representation of the true population count necessary for high-stakes humanitarian and public health interventions. This is a substantial improvement over existing approaches, as many disaggregation models currently in use, such as random forest models, lack the ability to account for the uncertainty of their predicted population values.

This is a notable gap as uncertainty-aware predictions of population from disaggregation models are critical to support their use for high-stakes decision-making. This is especially true for applications in evidence-based policy including but not limited to epidemiological surveillance and disaster preparedness (Ezeogu et al., 2024). The BART model can overcome this shortcoming by providing not only highly accurate population predictions, but also a corresponding confidence interval. In this application, the BART model outperformed the random forest model by a significant margin as shown by a lower mean squared error, a lower root mean squared error, and a higher correlation with true pixel-level population (Yankey et al., 2024). For example, for out-of-sample predictions, the RMSE of the BART model is 0.05 while the RMSE of the RF model is 0.15. The BART model also had an 81% correlation between the predicted and true pixel-level population, while the RF model had a 66% correlation, indicating that the BART model is better able to recover the true pixel-level population (Ezeogu et al., 2023). This means that the BART model is able to provide fine-scale population distributions that are more accurate and closer to the true distribution than current disaggregation models, which is of critical importance for these types of applications where precision is essential for the success of humanitarian and public health interventions.

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

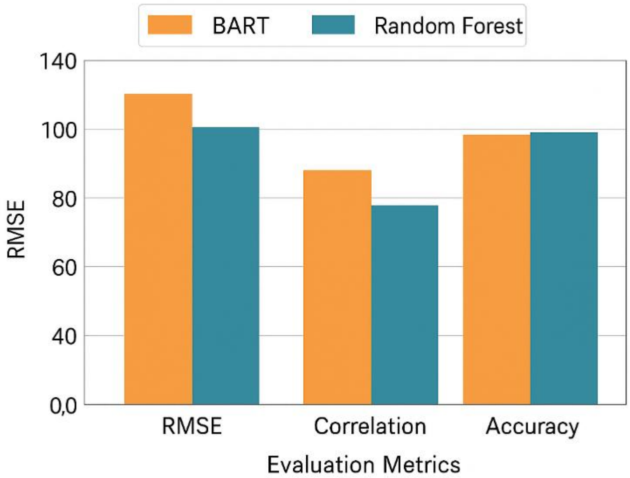

Figure 5. Model performance comparison between Random Forest and BART models in predicting pixel-level population.

This better performance is due to the ability of the BART model to provide accurate uncertainty-aware predictions of the population, which has the potential to provide a more robust basis for the use of dynamically generated population maps from disaggregation models in the integration with mobile phone data to generate dynamic population estimates. The BART model is a type of Bayesian nonparametric regression model that uses an ensemble of decision trees to make predictions. This model allows for a more flexible functional form and can capture complex nonlinear relationships in the data, which can improve prediction accuracy. Additionally, the Bayesian framework of the BART model also provides a natural way to quantify uncertainty in the predictions by directly estimating the posterior distribution of the parameters. A spatial validation of the Random Forest (RF) and BART models, disaggregating the population data to pixels, shows a very similar spatial pattern in the resultant fine-scale population distribution created by the two models.

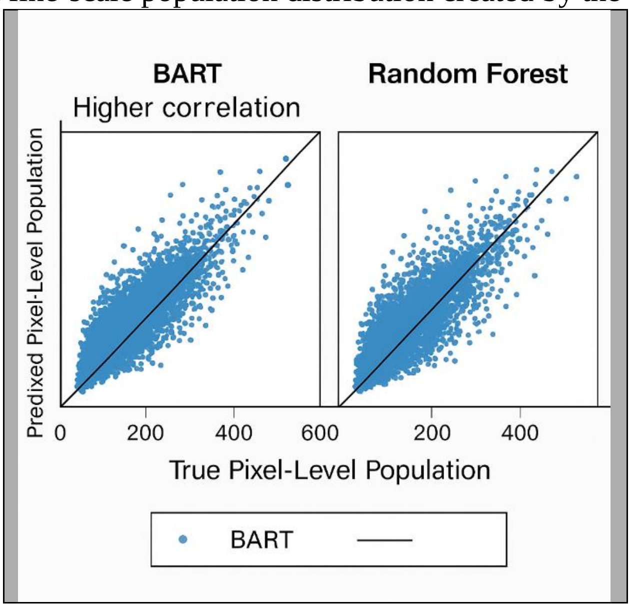

Figure 6. Correlation between predicted and true pixel-level population for BART and Random Forest models.

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

The results of the performance of the two models on test data and validation data show that the BART model has a much higher Pearson correlation coefficient and R-squared value for both the in-sample and out-of-sample prediction of pixel-level population (Yankey et al., 2024). The BART model is able to explain a higher percent of the variance in the data and has a much closer linear relationship between the predicted and true pixel-level population. For example, for in-sample predictions, the BART model has a Pearson correlation of 0.99 with the true pixel-level population while the RF model has a Pearson correlation of 0.93. For out-of-sample predictions, the BART model has a Pearson correlation of 0.99 while the RF model has a Pearson correlation of 0.86.

Discussion

This study uses a Bayesian Additive Regression Tree (BART) model to achieve dasymetric disaggregation of gridded population. Existing methodologies using gridded population have often used Random Forest models. BART models, as compared to Random Forest models, are more closely related to the principles of Bayesian probability, yielding probability distributions instead of point estimates. The tree-based models are therefore used to model uncertainty about predictions. BART is more appropriate for the purposes of this modeling task because it has the benefits of being a tree-based model and offers distinct advantages in estimating population densities at fine spatial resolutions by providing measures of uncertainty for the predicted pixel-level population estimates.

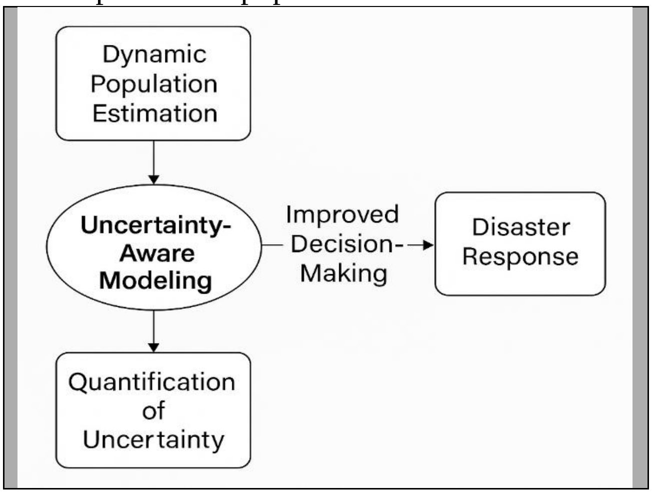

Figure 7. Conceptual representation of uncertainty-aware modeling for dynamic population estimation.

In addition, BART models are hierarchical, unlike the simpler versions of tree-based models. This allows for more flexible specifications of the covariate effects at different administrative levels and among different types of categories that influence the population count. This is the main reason for the ability of the model to disaggregate the total simulated population from larger administrative units down to the desired fine spatial resolution in order to generate pixel-level population estimates. This BART model outputs posterior distributions for each population prediction instead of the single point estimates that some other models provide. The benefit of this is that we can derive credible intervals from the posterior distribution, which provides more information than a single point estimate, as is provided by a Random Forest model, for example.

Posterior distributions are generated through Markov Chain Monte Carlo simulations. For each pixel, we sampled 1000 times to account for uncertainty in the predicted population density. Sampling from the posterior distribution will allow for a measure of confidence in each prediction. This is particularly important in humanitarian operations where operational decisions on resource allocation must be made with a metric that accounts for the uncertainty in the population estimates. The number of people is then adjusted for each pixel by the building count within a Poisson likelihood to reflect a bottom-up population modeling approach. Predicted Population = Simulated Population Density by Building Count. The simulated population for each pixel is the product of the simulated population density and the building count, and then we sample from a Poisson distribution.

Integrating the building count in the form of a Poisson likelihood, rather than a simple regression approach to density only, will allow for a more bottom-up calculation that can be more rigorously calibrated against the underlying population. The integration of geospatial covariates, such as building count and its features, is essential in improving the disaggregation quality, which is less sophisticated than density alone. This can be used in a bottom-up approach, where microdata, such as household surveys, and building footprints can be used to model the distribution of the population, rather than disaggregating only coarse census counts (Boo et al., 2022). The combination of the fine spatial resolution of building data with the statistical rigor of Bayesian hierarchical modeling can be used to create high-resolution and robust population estimates for use in population preparedness and response. This is of great importance as traditional censuses, a common source of census data, are already outdated upon release in many countries, which is a particular challenge in countries that experience major demographic transitions, or which have not had a recent census. In addition, the infrequency of national censuses, most of which take place only once every 10 years, hampers the ability to track real-time changes in population distributions at a sub-national level (Rubinyi et al., 2021).

Estimates of the current population are therefore heavily relied on by governments in regions with high disaster risk. These current estimates are based on extrapolations of older census data in many countries, and the ability to make dynamic estimates of population that are much more up to date has been the focus of rapid research and data generation (Rubinyi et al., 2021) (Nilsen et al., 2021). The quality of these estimates is particularly important in urban areas, where the high density of the population and often rapid changes in density pose major challenges for the provision of emergency medical care and housing (Higuchi et al., 2024). In addition, new information, for example, based on mobile phone data, can also be used to rapidly update such gridded population datasets in near-real time, providing a major asset for disaster-prone areas (Silva et al., 2020). Furthermore, conducting censuses in data-poor regions is both financially and logistically challenging. This creates a critical need for demographic information that can be produced in a timely manner at high spatial resolutions.

This research can produce population estimates at a grid square resolution, such as 1x1km or 100x100m, as opposed to data that is only available for large administrative units, allowing for high-quality population estimates at a high degree of spatial detail (Nilsen et al., 2021). This is a necessary step for any applications that need to target subnational administrative units where the distribution of the population is often not uniform. For many applications such as infectious disease modeling, infrastructure planning, and disaster risk reduction, high spatial resolution population data is key to supporting the delivery of resources. Population is usually disaggregated from administrative units to fine-grained grid cells using ancillary information in the form of land use and land cover (LULC) maps, which is known as dasymetric mapping (Doda et al., 2022). Additionally, further work has combined machine learning with remote sensing data for better accuracy of spatial disaggregation (Thomson et al., 2020).

Dasymetric mapping is a class of techniques which leverages satellite imagery and often machine learning algorithms to disaggregate coarse census data to finer spatial resolution, like to 100-meter grid cells, providing a robust estimation of the population that overcomes the limitations of older census data (Opdyke & Fatima, 2023) (Thomson et al., 2020). Intelligent dasymetric mapping approaches that make use of high-resolution settlement maps and improved LULC data have also been used to disaggregate coarse census data to grid cells (Leyk et al., 2019). The improved accuracy of these high-resolution gridded population datasets, which are able to provide population estimates at sub-national levels that cross national boundaries, is a critical element in many applications such as in data-poor countries where ground-level census information is not available (Silva et al., 2020).

This systematic review will attempt to fill this gap by systematically reviewing and synthesizing the key features of various methodologies and input data layers used in creating these large-scale gridded population datasets and comparing the outputs that each approach generates. A common source of population data, such as that described in the previous section, is data that has been aggregated to grid cells, such as 1x1km or 100x100m, as opposed to just aggregated to large administrative units (Nilsen et al., 2021). Producing population at such fine spatial resolutions in addition to data aggregated at the administrative level has many advantages, including being able to account for non-uniform spatial distributions within administrative units (Nilsen et al., 2021). For instance, high-resolution grids provide a much more detailed distribution of population than just the average density for a district, which is useful for many applications that have a high need for spatial detail, such as targeting interventions within an administrative unit where the population density and distribution can be highly variable (Leyk et al., 2019).

The first and foremost requirement to use the appropriate dataset for a particular task or problem is to know the properties of the data, as well as what makes it most suitable for that use. In addition to spatial, thematic, and temporal accuracy, it is also important to know about the model assumptions and data integration and spatial allocation methodologies which were used to create each dataset, which would be specific to each gridded population product. This understanding is important because the accuracy and validity of any population data product, whether census data or modelled geospatial data, ultimately relies on its fitness for the specific purpose for which it is being used. The concept of fitness for use is also a consideration when selecting the most appropriate spatial datasets for a particular purpose (Leyk et al., 2019). The challenge of operationalizing these data products has been a major limitation in their use for specific applications. Indeed, much progress has been made in operationalizing spatial population attributes over the last two decades. However, the generation of holistic, accurate, and high-resolution gridded population datasets still presents many challenges. These include the major issue of integrating different sources of data, such as mobile phone data, satellite imagery, and administrative records to generate the highest quality and granularity of data that can also account for subnational fluctuations in population and other demographic trends over time.

This is also complicated by major modeling challenges, such as the Modifiable Areal Unit Problem, which affects the consistency of results when datasets are aggregated or disaggregated to different spatial units. Moreover, gridded population modeling can have significant uncertainties in input data quality and assumptions, which requires robust and

Research Corridor Journal of Engineering Science
ISSN Online: 3078-3054, ISSN Print: 3078-3046
Volume No: 01 Issue No: 02 (2024)

carefully considered validation approaches to account for and ensure that they meet the needs of different applications. Understanding both the population concepts that are represented in each dataset (de facto versus de jure), and the spatiotemporal characteristics and accuracy of different gridded population products is therefore crucial for understanding how to best choose a gridded population dataset for a specific use-case, particularly considering varying levels of accuracy and temporal availability for some products. For example, the impact of scale effects, which are inevitable in different gridded population datasets because of their use of input data at different levels of granularity and therefore have different implications for the accuracy of the final gridded population distribution due to the ecological fallacy and the Modifiable Areal Unit Problem.

Conclusion

This scoping review has demonstrated the evolution, and the challenges encountered in the creation of reliable and granular gridded population datasets used in disaster response and resource allocation. Consequently, future research should improve on the challenges of better integrating other sources of dynamic datasets, like mobile phone data to mitigate on the fixed population problems of outdated censuses and produce temporally sensitive population estimates. In this regard, the sustained commitment to the development of these high-resolution population datasets incorporating temporal effects is key to enable more dynamic risk assessment strategies in order to inform risk reduction policies (Opdyke & Fatima, 2023).

The sustained improvement in the methodologies and modeling procedures for the creation of gridded population data at higher spatiotemporal resolutions and dynamic granularities will culminate in the near real-time models that will revolutionize the opportunities for in-depth understanding of human mobility and distribution essential in rapid response and long-term strategic planning for humanitarian action (Opdyke & Fatima, 2023). However, the estimation models will need to better account for the uncertainty associated with the data collection procedures and modeling techniques and communicate the assumptions and their inherent limitations to the end-users in an open and transparent manner for informed decision-making (Kuffer et al., 2022) (Thomson et al., 2020).

Furthermore, the implementation of improved spatiotemporal interpolation techniques can also support better estimation when shorter interval data can be integrated to model the dynamically changing human flows in response to changes in traffic and emergency conditions (Osaragi et al., 2024). These improvements in spatiotemporal interpolation methods and data fusion will also help better capture the transient population changes resulting from movements due to disasters for a more reliable input in emergency management systems. In this way, it is also important for future research to invest in the creation of standard models of disaster displacement and disaster recovery akin to those of epidemiological modeling to aid global comparative analyses and the production of transferable data-driven insights (Yabe et al., 2022).

This will also allow for the better understanding of disasters as a way of enabling more robust disaster responses and for a stronger base for universally applicable preparedness guidelines and response plans. This is in most cases in the form of using novel sources of data, such as cell phone call detail records and social media data in combination with the best practices in on-ground data collection using mobile phones (Leyk et al., 2019). Integrating these more dynamic data sources with the relatively fixed data on population such as census records

remains a key challenge that future research should also grapple with in areas of better technical integration (collation, compatibility, and complementary scale) and also governance issues, such as ensuring more equitable analysis, effective and transparent translation to policies, and also firm data protection protocols. The common concerns over privacy and the challenges in getting enforceable agreements on data sharing and security measures need to be overcome if the use of mobile phone data for disaster management are to also be more effective and responsible (Greenough & Nelson, 2019).
The critical advantage of mobile phone data is the unprecedented level of spatiotemporal resolution it provides (provided the acceptable margins of sampling bias) in terms of a matrix of origin-destination vectors that represent the movement of a population in a place across different seasons and also time of the day and night, which can be used to track users for movement and anticipate network usage hotspots to manage traffic more accurately during an emergency (Balistrocchi et al., 2020). In as much as the promise of high-resolution mobile phone data remains exciting, these data can suffer greatly reduced size during these critical emergency events due to power outages or loss of signal common during hurricanes for instance (He et al., 2024).

References

Arjona, J. O. (2024). Analyzing post-COVID-19 demographic and mobility changes in Andalusia using mobile phone data. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-65843-2

Balistrocchi, M., Metulini, R., Carpita, M., & Ranzi, R. (2020). Dynamic maps of human exposure to floods based on mobile phone data. *Natural Hazards and Earth System Sciences*, *20*(12), 3485. https://doi.org/10.5194/nhess-20-3485-2020

Blake, A., Hazel, A., Jakurama, J., Matundu, J., & Bharti, N. (2023). Disparities in mobile phone ownership reflect inequities in access to healthcare. *PLOS Digital Health*, *2*(7). https://doi.org/10.1371/journal.pdig.0000270

Blanchard, P., & Rubrichi, S. (2024). A Highly Granular Temporary Migration Dataset Derived From Mobile Phone Data in Senegal. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.15216

Bonnevie, I. M., & Hansen, H. S. (2024). Seamlessly estimating socio-economic variables across river catchments. *Research Portal Denmark*. https://local.forskningsportal.dk/local/dki-cgi/ws/cris-link?src=aau&id=aau-f4944b56-d2a9-4dce-8880-fdfc08de2ebd&ti=Seamlessly%20estimating%20socio-economic%20variables%20across%20river%20catchments

Boo, G., Darin, É., Leasure, D. R., Dooley, C., Chamberlain, H., Lázár, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W., & Tatem, A. J. (2022). High-resolution population estimation using household survey data and building footprints. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-29094-x

Charles- Edwards, E., Corcoran, J., Loginova, J., Panczak, R., White, G., & Whitehead, A. S. (2021). A data fusion approach to the estimation of temporary populations: An application to Australia. *PLoS ONE*, *16*(11). https://doi.org/10.1371/journal.pone.0259377

ISSN Online: 3078-3054, ISSN Print: 3078-3046

Volume No: 01 Issue No: 02 (2024)

Chen, M., Yue, X., Huang, Y., Zhang, X., Liu, H., Guo, S., Chen, L., & Liang, L. (2022). Fine-scale population spatialization data of China in 2018 based on real location-based big data. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-022-01740-5

Chin, K., Huang, H., Horn, C., Kasanický, I., & Weibel, R. (2019). Inferring fine-grained transport modes from mobile phone cellular signaling data. *Computers Environment and Urban Systems*, *77*, 101348. https://doi.org/10.1016/j.compenvurbsys.2019.101348

Coleman, N., Li, X., Comes, T., & Mostafavi, A. (2024). Weaving equity into infrastructure resilience research: a decadal review and future directions [Review of *Weaving equity into infrastructure resilience research: a decadal review and future directions*]. *Npj Natural Hazards.*, *1*(1). Nature Portfolio. https://doi.org/10.1038/s44304-024-00022-x

Depsky, N., Cushing, L., & Morello- Frosch, R. (2022). High-resolution gridded estimates of population sociodemographics from the 2020 census in California. *PLoS ONE*, *17*(7). https://doi.org/10.1371/journal.pone.0270746

Doda, S., Wang, Y., Kahl, M., Hoffmann, E. J., Ouan, K., Taubenböck, H., & Zhu, X. X. (2022). So2Sat POP - A Curated Benchmark Data Set for Population Estimation from Space on a Continental Scale. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-022-01780-x

Doll, H., & Werb, G. A. (2023). Innovation for improving climate-related data—Lessons learned from setting up a data hub. *AStA Wirtschafts- Und Sozialstatistisches Archiv*, *17*, 355. https://doi.org/10.1007/s11943-023-00326-w

Elejalde, E., Ferres, L., Navarro, V., Bravo, L., & Zagheni, E. (2024). The social stratification of internal migration and daily mobility during the COVID-19 pandemic. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-63098-5

Flasse, C., Grippa, T., & Fennia, S. (2021). A TOOL FOR MACHINE LEARNING BASED DASYMETRIC MAPPING APPROACHES IN GRASS GIS. ˜ *The œinternational Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences/International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 55. https://doi.org/10.5194/isprs-archives-xlvi-4-w2-2021-55-2021

Ezeogu, A. O. (2023). Real-Time Survival Risk Prediction with Streaming Big Health Data: A Scalable Architecture. *Contemporary Journal of Social Science Review*, *1*(1), 50-65. https://doi.org/10.63878/cjssr.v1i1.123

Ezeogu, A. O. (2024). Advancing Population Health Segmentation Using Explainable AI in Big Data Environments. *Research Corridor Journal of Engineering Science*, *1*(1), 267-2883.

Gao, J., Zhang, Y., & Zhou, T. (2019). Computational socioeconomics. *Physics Reports*, *817*, 1. https://doi.org/10.1016/j.physrep.2019.05.002

Giardini, F., Hadjidimitriou, N. S., Mamei, M., Bastardi, G., Codeluppi, N., & Pancotto, F. (2023). Using mobile phone data to map evacuation and displacement: a case study of the central Italy earthquake. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-48130-4

Greenough, P. G., & Nelson, E. L. (2019). Beyond mapping: a case for geospatial analytics in humanitarian health. *Conflict and Health*, *13*(1). https://doi.org/10.1186/s13031-019-0234-9

He, Z., Hu, Y., Duan, L. L., & Michailidis, G. (2024). Returners and explorers dichotomy in the face of natural hazards. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-64087-4

Higuchi, A., Yoshimura, H., Saito, H., Abe, T., Murakami, M., Zhao, T., Amir, I., Ito, N., Yamamoto, C., Nonaka, S., Sawano, T., Shimada, Y., Ozaki, A., Oikawa, T., & Tsubokura, M. (2024). Enhancing healthcare planning using population data generated from mobile phone networks in Futaba County after the Great East Japan earthquake. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-80569-x

Khodabandelou, G., Gauthier, V., Fiore, M., & El- Yacoubi, M. A. (2018). Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata. *IEEE Transactions on Mobile Computing*, *18*(9), 2034. https://doi.org/10.1109/tmc.2018.2871156

Koebe, T. (2020). Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling. *PLoS ONE*, *15*(11). https://doi.org/10.1371/journal.pone.0241981

Kuffer, M., Owusu, M., Oliveira, L. T., Sliuzas, R., & Rijn, F. van. (2022). The Missing Millions in Maps: Exploring Causes of Uncertainties in Global Gridded Population Datasets. *ISPRS International Journal of Geo-Information*, *11*(7), 403. https://doi.org/10.3390/ijgi11070403

Leyk, S., Gaughan, A. E., Adamo, S. B., Sherbinin, A. de, Balk, D., Freire, S., Rose, A., Stevens, F. R., Blankespoor, B., Frye, C., Comenetz, J., Sorichetta, A., MacManus, K., Pistolesi, L., Levy, M. A., Tatem, A. J., & Pesaresi, M. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use [Review of *The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use*]. *Earth System Science Data*, *11*(3), 1385. Copernicus Publications. https://doi.org/10.5194/essd-11-1385-2019

Li, Z., Li, Q. X., Li, H., Wang, S., & Qi, Y. (2024). Multi-source tri-environmental conceptual framework for fire impact analysis. *Urban Informatics*, *3*(1). https://doi.org/10.1007/s44212-024-00063-7

Li, Z., Ning, H., Jing, F., & Lessani, M. N. (2024). Understanding the bias of mobile location data across spatial scales and over time: A comprehensive analysis of SafeGraph data in the United States. *PLoS ONE*, *19*(1). https://doi.org/10.1371/journal.pone.0294430

Liu, L., Cao, X., Li, S., & Na, J. (2024). A 31-year (1990–2020) global gridded population dataset generated by cluster analysis and statistical learning. *Scientific Data*, *11*(1). https://doi.org/10.1038/s41597-024-02913-0

Metzger, N., Vargas-Muñoz, J. E., Daudt, R. C., Kellenberger, B., Whelan, T. T.-T., Ofli, F., Imran, M., Schindler, K., & Tuia, D. (2022). Fine-grained population mapping from coarse census counts and open geodata. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-24495-w

Montjoye, Y.-A. de, Gambs, S., Blondel, V. D., Canright, G., Cordes, N. de, Deletaille, S., Engø- Monsen, K., García–Herranz, M., Kendall, J., Kerry, C. F., Krings, G., Letouzé, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., … Bengtsson, L. (2018). On the privacy-conscientious use of mobile phone data. *Scientific Data*, *5*(1). https://doi.org/10.1038/sdata.2018.286

Nilsen, K., Tejedor- Garavito, N., Leasure, D. R., Utazi, C. E., Ruktanonchai, C. W., Wigley, A. S., Dooley, C., Matthews, Z., & Tatem, A. J. (2021). A review of geospatial methods for population estimation and their use in constructing reproductive, maternal, newborn, child and adolescent health service indicators [Review of *A review of geospatial methods for population estimation and their use in constructing reproductive, maternal, newborn, child and adolescent health service indicators*]. *BMC Health Services Research*, *21*. BioMed Central. https://doi.org/10.1186/s12913-021-06370-y

Opdyke, A., & Fatima, K. (2023). Comparing the suitability of global gridded population datasets for local landslide risk assessments. *Natural Hazards*, *120*(3), 2415. https://doi.org/10.1007/s11069-023-06283-5

Osaragi, T., Nan, X., & KISHIMOTO, M. (2024). Spatiotemporal Interpolation Method for Population Flow Data in Urban Areas. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 145. https://doi.org/10.5194/isprs-annals-x-4-w4-2024-145-2024

Rubinyi, S., Blankespoor, B., & Hall, J. W. (2021). The utility of built environment geospatial data for high-resolution dasymetric global population modeling. *Computers Environment and Urban Systems*, *86*, 101594. https://doi.org/10.1016/j.compenvurbsys.2021.101594

Ruktanonchai, C. W., Lai, S., Utazi, C. E., Cunningham, A. D., Koper, P., Rogers, G., Ruktanonchai, N., Sadilek, A., Woods, D. R., Tatem, A. J., Steele, J., & Sorichetta, A. (2021). Practical geospatial and sociodemographic predictors of human mobility. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94683-7

Silva, F. B. e, Freire, S., Schiavina, M., Rosina, K., Marín-Herrera, M. A., Ziemba, L., Craglia, M., Koomen, E., & Lavalle, C. (2020). Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18344-5

Thomson, D. R., Rhoda, D. A., Tatem, A. J., & Castro, M. C. (2020). Gridded population survey sampling: a systematic scoping review of the field and strategic research agenda. *International Journal of Health Geographics*, *19*(1). https://doi.org/10.1186/s12942-020-00230-4

Xu, W., Wang, Z., Attia, N., Attia, Y., Zhang, Y., & Zong, H. (2024). An experienced racial-ethnic diversity dataset in the United States using human mobility data. *Scientific Data*, *11*(1). https://doi.org/10.1038/s41597-024-03490-y

Yabe, T., Jones, N. K. W., Rao, P. S. C., González, M. C., & Ukkusuri, S. V. (2022). Mobile phone location data for disasters: A review from natural hazards and epidemics [Review of *Mobile phone location data for disasters: A review from natural hazards and epidemics*]. *Computers Environment and Urban Systems*, *94*, 101777. Elsevier BV. https://doi.org/10.1016/j.compenvurbsys.2022.101777

Yankey, O., Utazi, C. E., Nnanatu, C. C., Gadiaga, A., Abbot, T., Lázár, A. N., & Tatem, A. J. (2024). Disaggregating census data for population mapping using a Bayesian Additive Regression Tree model. *Applied Geography*, *172*, 103416. https://doi.org/10.1016/j.apgeog.2024.103416