How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia

Dana R. Thomson[1],*, Douglas R. Leasure[2], Tomas Bird[2,#], Nikos Tzavidis[1], and Andrew J. Tatem[2]

1. Department of Social Statistics and Demography, University of Southampton, Southampton UK
2. WorldPop, Geography and Environmental Science, University of Southampton, Southampton, UK

\#   Current Address: NorthWest Atlantic Fisheries Centre, Department of Fisheries and Oceans, St. John's, Canada
*   Corresponding author, Email: dana.r.thomson@gmail.com (DRT)

Abstract

Disaggregated population counts are needed to calculate health, economic, and development indicators in Low- and Middle-Income Countries (LMICs), especially in settings of rapid urbanisation. Censuses are often outdated and inaccurate in LMIC settings, and rarely disaggregated at fine geographic scale. Modelled gridded population datasets derived from census data have become widely used by development researchers and practitioners. These datasets are evaluated for accuracy at the spatial scale of the input data which is often much courser (e.g. administrative units) than the neighbourhood or cell-level scale of many applications.

We simulate a realistic "true" 2016 population in Khomas, Namibia, a majority urban region, and introduce realistic levels of outdatedness (over 15 years) and inaccuracy in slum, non-slum, and rural areas. We aggregate these simulated realistic populations by census and administrative boundaries (to mimic census data), and generate 32 gridded population datasets that are typical of a LMIC setting using WorldPop-Global-Unconstrained gridded population approach. We evaluate the cell-level accuracy of these simulated datasets using the original "true" population as a reference.

In our simulation, we found large cell-level errors, particularly in slum cells, driven by the use of average population densities in large areal units to determine cell-level population densities. Age, accuracy, and aggregation of the input data also played a role in these errors. We suggest incorporating finer-scale training data into gridded population models generally, and WorldPop-Global-Unconstrained in particular (e.g., from routine household surveys or slum community population counts), and use of new building footprint datasets as a covariate to improve cell-level accuracy. It is important to measure accuracy of gridded population datasets at spatial scales more consistent with how the data are being applied, especially if they are to be used for monitoring key development indicators at neighbourhood scales with relevance to small dense deprived areas within larger administrative units.

Key Words

LMIC, Global South, indicator, Random Forrest

Introduction

Small area population counts, especially in low- and middle-income countries (LMICs), provide essential denominators for health, economic, and development indicators [1]. For example, small area population counts are used to calculate vaccination coverage rates [2], understand health service utilisation [3], and estimate infection rates of malaria, COVID-19, and many other health conditions [4]. Censuses are generally collected every ten years, though one in ten LMICs has not held a census in the last 15 years [5], and some national censuses have poor data quality due to negligence (e.g., [6,7]) or deliberate mis-counting of sub-populations for political purposes (e.g., [8–10]). Due to increasing rates of mobility and urbanisation worldwide, especially in African and Asian cities where 90% of global population growth is expected in the next 30 years [11], the urban poorest are increasingly difficult to count as more people take-up residence in informal settlements or atypical housing locations (e.g., a shop) [12].

In the absence of updated, fine-scale census data, many policy-makers, researchers, and service providers have turned to gridded population estimates as a source of population counts in their work. Gridded population data provide estimates of the total population in small grid cells, and are derived with geo-statistical methods using population counts and spatial datasets [13]. Gridded population data are viewed by data producers and users as meeting a global development challenge to "leave no one off the map" and thus leave no one behind [14]. "Top-down" gridded population estimates have been available for roughly 15 years and disaggregate census or other complete population counts in areal units (e.g., 3rd-, 4th-, or 5th-level administrative units) to grid cells (e.g., 30x30m, 100x100m, 1x1km) [12]. The simplest models assume a uniform distribution of population within areal units (i.e., GPW [15,16], GHS-POP [17,18], HRSL [19]), while the most complex models use spatial covariates to inform spatial disaggregation of the areal unit population to grid cells (i.e., WorldPop-Global [20,21], LandScan [22,23], WPE [24]). Most models aim to reflect the average night-time residential population distribution, though LandScan reflects "ambient" population, the average between night-time residential and daytime commuter populations [22]. To estimate gridded population figures beyond the year of the last census; birth, migration, and death rates are used to project new population totals by areal unit [25]. "Bottom-up" gridded population estimates are derived from micro-census population counts in a sample of areas, or from assumptions about the average household size, and have only recently been developed [26,27]. Most gridded population datasets use a settlement layer to "constrain" population estimates to settled grid cells. Exceptions include Gridded Population of the World (GPW) in which the areal unit population is simply divided equally among cells [15], and WorldPop-Global-Unconstrained which uses a complex model to produce disaggregation weights for all land areas, and results in very small estimates of a person (e.g., 0.00001 people) in unsettled grid cells [20]. Read papers by Leyk and colleagues (2019) and Thomson and colleagues (2020) for detailed descriptions and comparisons of these gridded population datasets [12,13].

The accuracy of "top-down" gridded population data is generally calculated at the scale of the input population areal units because these are the finest-scale population counts available to the data producers. A number of factors contribute to gridded population model accuracy including: (1) the modelling algorithm itself, (2) inaccuracy of the input population data, (3) the geographic scale of the input population data (e.g., census tracts versus districts), (4) the age, accuracy, completeness, and type of ancillary data, (5) the nature of the relationship between ancillary data and population density, and (6) the geographic scale of the output grid. Of these, the two strongest predictors of accuracy (at the scale of areal units) in top-down gridded population models are the resolution and age of the data [28]. Among top-down gridded population datasets, the WorldPop-Global-Unconstrained Random Forest model is among the best documented and most accurate gridded population models available [20,29] with publicly available model code [30] and pre-processed model covariates [31,32] enabling reproducibility and evaluation. WorldPop-Global-Unconstrained and its preceding data products (AfriPop, AsiaPop, and AmeriPop) result in estimates for all

land areas; however, a new WorldPop-Global-Constrained dataset was recently published limiting population estimates to cells with buildings or built-up features [33].

Accuracy of gridded population data at the scale of input areal units is not ideal, as users of these datasets tend to need population counts at finer geographic scales [13]. Compounding this issue is that users often turn to gridded population estimates when census counts are excessively outdated or untrustworthy [12]. The accuracy of gridded population data at the scale of output grid cell is largely unknown due to the challenges of finding reliable population counts to use as a reference. Where extremely fine-scale population counts are available for the entire population, gridded population estimates are not needed. Given that many gridded population estimates are derived from outdated or inaccurate census data, and use of gridded population data is most common when census data are outdated or inaccurate, it is imperative to understand if, and how, census inaccuracies propagate through gridded population data. Few accuracy assessments have ever been performed on gridded population data at the cell-level, and those that exist are generally from high-income countries (e.g., [34,35]).

To evaluate cell-level accuracy of gridded population data, actual population counts are needed for each grid cell or in finer-scale units such as household point locations. Few censuses in LMICs collect household latitude-longitude coordinates, and where they do, the data are extremely sensitive and difficult to obtain. Furthermore, even the best census data might be problematic because vulnerable sub-populations including homeless and nomadic populations are supposed to be counted separately in special enumerations, though under-resourced statistical offices are often not able to perform these counts [36], and some censuses do not include certain refugee or internally displaced populations [37]. To ensure that this analysis of cell-level accuracy did not exclude the urban poorest and other hidden populations, we chose to simulate a realistic population in a LMIC setting. It was important that the simulated population was located in a real-world location so that actual covariate datasets – with their own imperfections – could be used to generate realistic gridded population datasets. We adapted methods outlined by Thomson and colleagues (2018) for simulating a geo-located realistic household population, and added classification of urban households by slum/non-slum area in a final step to focus this analysis on dynamic, complex LMICs cities where inaccuracies in gridded data are likely to propagate [38].

This paper describes how we evaluated the cell-level accuracy of 32 simulated 100x100m WorldPop-Global-Unconstrained gridded population datasets which reflect realistic scenarios of census (1) outdatedness, (2) inaccuracy, and (3) aggregation in an urban LMIC setting.

Methods

Setting

We chose to simulate a population in Khomas, Namibia – in which the vast majority of residents reside in Windhoek, the capital – because the government has produced numerous high-quality population datasets [39], and Windhoek's population is incredibly dynamic (Figure 1). Namibia, like some other countries that inherited colonial boundaries, placed restrictions on freedom of movement until independence in 1990 [40]. After independence, vast numbers of the population migrated to Windhoek, exaggerating rural-to-urban migration patterns observed globally during this time period [41,42]. Windhoek is also a destination for immigrants from neighbouring countries including financially unstable Zimbabwe [42,43]. The population of the Windhoek metropolitan area grew by a staggering 37% between the 2001 and 2011 censuses [39], with much of that growth in informal settlements [40].

Figure 1. Location of Khomas region in Namibia, and of constituencies in Windhoek area

Simulation Overview

To simulate realistic gridded population datasets for Khomas, Namibia, we first simulated a "true" 2016 population geo-located to realistic manually-generated household point locations; introduced realistic outdatedness by removing households in 2011, 2006, and 2001; introduced realistic inaccuracies among urban-slum, urban-non-slum, and rural sub-populations; and finally aggregated these 16 simulated population scenarios into two geographic areal units (census EA and constituency) to generate 32 realistic WorldPop-Global-Unconstrained 100x100m gridded population datasets. This workflow is summarised in Figure 2 and detailed below.

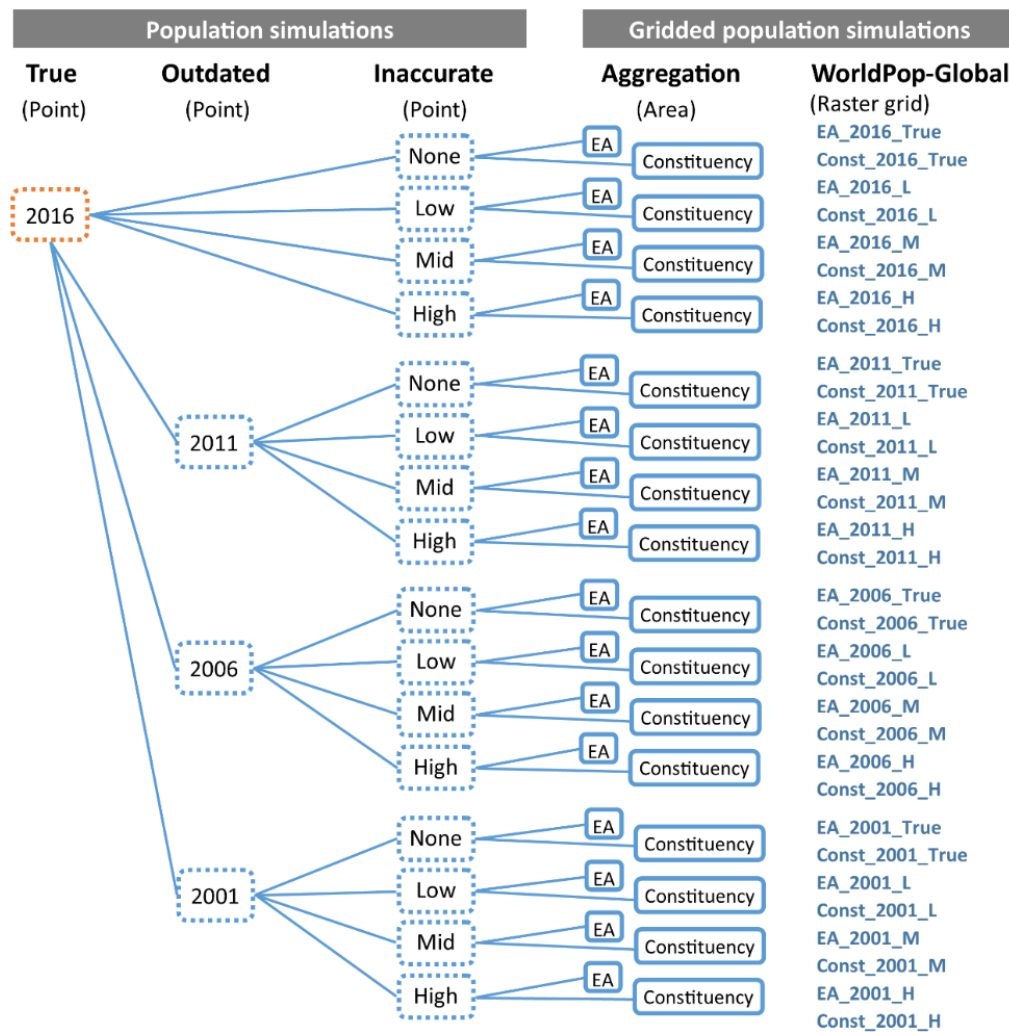



Figure 2. Summary of the population and gridded population simulation workflow

1) Simulate a realistic population geo-located to realistic building point locations, 2) simulate three periods of outdatedness by removing households at point locations not present on satellite imagery in earlier years, 3) simulate low/middle/high census inaccuracy by removing points at random from rural, urban-slum, and urban-non-slum household types, 4) aggregate to 922 census enumeration areas (EAs) and 10 constituencies (admin-2), 5) generate 100x100m gridded population datasets in raster grid format using WorldPop-Global-Unconstrained approach and WorldPop-Global spatial covariates.

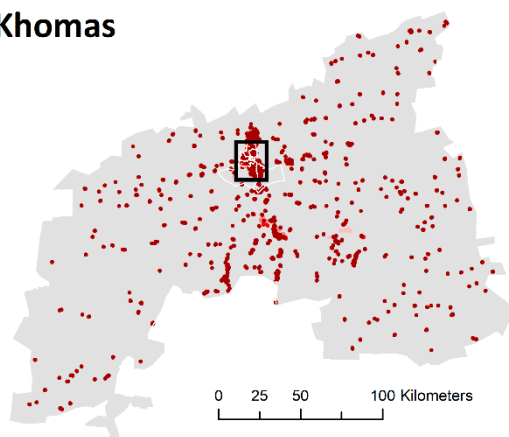
Simulating a "true" 2016 population geo-located to household latitude-longitude points

To simulate a realistic population in Khomas, Namibia, we used all of the same population inputs and spatial auxiliary datasets as Thomson and colleagues (2018) [38]. Broadly, this involved the creation of three datasets - modelled surfaces of household types, manually digitised building point locations, and synthetic (simulated) households - then assigned synthetic households to point locations based on the household type probability surfaces. Household types were defined from Namibia 2013 Demographic and Health Survey (DHS) data using k-means analysis with variables that were also present in the Namibia 2011 census (e.g., improved sanitation facilities, gender of head of household). Next, probability surfaces of these household types were created using a Random Forrest model and spatial covariates to interpolate the likelihood of a given household type across Namibia between DHS survey locations [38]. The probability surfaces of "urban poor" and "urban non-poor" household types were manually adjusted due to high misclassification. These adjustments were made by manually assigning the proportion of households in each census enumeration area (EA) that appeared to be located in areas of small disorganised buildings based on visual inspection of 30m Quickbird satellite imagery. Separately, we modelled a synthetic population of individuals nested within

households across Khomas from Namibia 2011 census microdata using an iterative proportional fitting model and conditional annealing [44]. A third set of data, building point locations, were manually digitised from 2014-2016 30cm Quickbird imagery in ArcGIS 10. To link synthetic households with building locations, we calculated the most likely household type of each synthetic household using k-means analysis scores, and iteratively assigned households to building point locations based on the probability of each household type at a given building point. Finally, using the manually classified EAs (with our estimated portion of urban poor households), we classified all urban households as being located in either a slum or non-slum area. All of these steps are detailed in Supplement 1 and the paper by Thomson and colleagues (2018) [38]. This simulated population is meant to represent a realistic "true" reference population for 2016.

Simulating realistic outdatedness of Khomas census population

To simulate population outdatedness in Khomas, we imported the above 2016 "true" population household point locations into Google Earth, and used the software's historical Maxar and SPOT imagery (40cm) to flag all buildings that were not present in 2011, 2006, and 2001 imagery. The oldest imagery available at 40cm resolution in Google Earth was from 2004, so we used some judgement to flag buildings that looked recently built in 2004 (e.g., bare fresh soil) and assumed they were not present in 2001. During this exercise, we ensured that the number of household coordinates in each constituency matched the number of households reported in the 2001 and 2011 Population and Housing Census final reports to ensure that both patterns and degree of outdatedness were realistic [39] (Figure 3). The simulated population is provided in Supplement 2 and is comparable to the Oshikoto, Namibia 2016 simulated population created by Thomson and colleagues [38].

Khomas

Households	2016	2011	2006	2001
Tobias Hainyeko	12,756	12,428	10,486	8,872
Katutura Central	5,182	5,096	4,948	4,072
Katutura East	3,824	3,756	3,659	3,165
Khomasdal	11,684	10,471	7,302	5,770
Soweto	3,470	3,377	3,167	2,553
Samora Machel	16,718	13,250	7,573	6,598
Windhoek East	7,532	7,089	6,451	5,620
Windhoek Rural	7,256	6,330	5,415	4,961
Windhoek West	13,947	13,837	12,590	9,991
Moses //Garoeb	15,298	13,804	10,315	6,978
Khomas	**97,667**	**89,438**	**71,906**	**58,580**

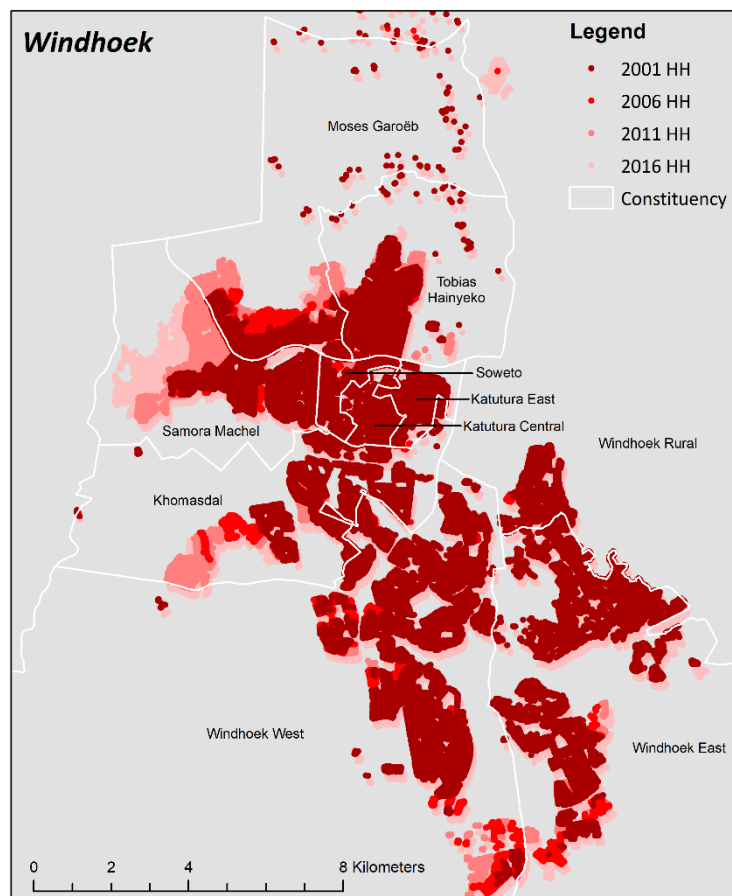



Figure 3. Household point locations in Khomas, Namibia by presence in 2016, 2011, 2006, and 2001

Simulating realistic levels of under-count inaccuracy in censuses

To identify realistic levels of under-counts among urban-slum, urban-non-slum, and rural populations in LMIC censuses, we reviewed the scientific and grey literature. The review included census post enumeration surveys (PESs) in 108 LMICs listed by the UN Statistical Division Census Programme website [5], and a systematic search in PubMed and Scopus of articles published between January 1, 1990 and February 28, 2017 using the following search criteria: "census AND (listing OR enumerat* OR count OR coverage OR miss*) AND (nomad* OR pastoral* OR refugee OR displaced OR migrant OR slum OR poorest OR unregistered OR homeless OR [street] sleeper OR pavement [dweller] OR floating)". The first wave of the literature search resulted in 459 unique articles, of which co-author DRT screened all titles and abstracts. Of 72 potentially eligible articles from LMICs, DRT reviewed the full-text, and kept five which reported a census under-count. In a second wave, we used Google Scholar to identify the top 20 "cited by" and top 20 "related" articles for each of the five articles identified in the first wave. The second wave resulted in 334 unique articles, of which 49 had potentially relevant titles or abstracts. After a full-text review of these articles, we found eight additional reported census under-counts. Together, census under-counts in LMICs were collated from 10 PESs [45–54], and 13 articles [7,55–66] (Figure 4). The average census under-counts were 46% in urban-slum populations, 6% in urban-non-slum populations, and 7% in rural populations (Table 2, see Supplement 3 for details).

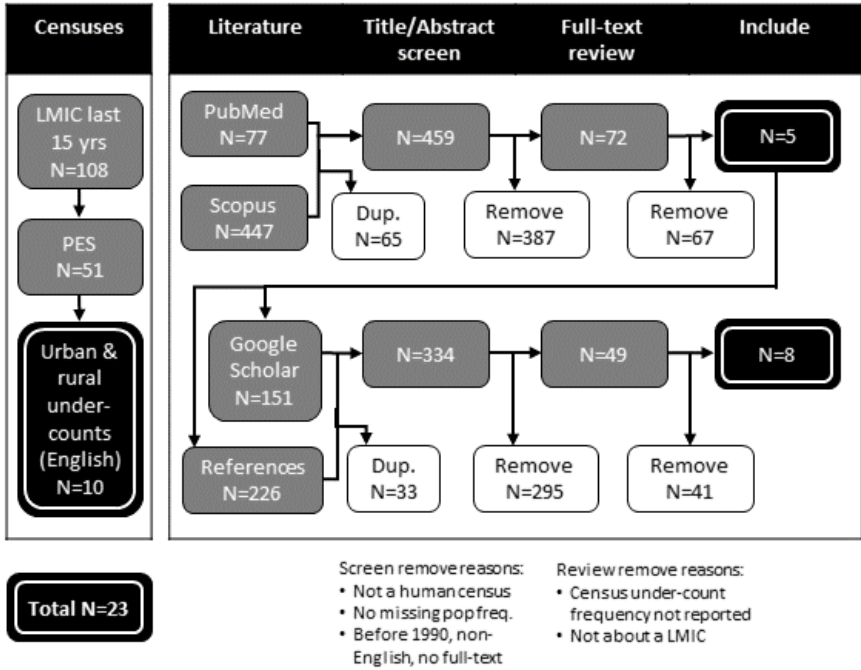



Figure 4. Search terms and process used in the census under-count literature review

Based on these findings, we simulated three levels of census inaccuracy: low inaccuracy was considered to be missing 2% of rural and urban-non-slum households, and 10% of urban-slum households; medium inaccuracy was considered to be missing 5% of rural and urban-non-slum households, and 30% of urban-slum households; and finally, high inaccuracy was classified as missing 10% of rural and urban-non-slum households, and 60% of urban-slum households (Table 1). We applied the inaccuracy rates at random within rural, urban-slum, and urban-non-slum households such that there was no spatial pattern inherent to the simulated under-counts. This exercise resulted in one "true" and 15 simulated outdated-inaccurate populations to generate realistic gridded population datasets that reflect typical gridded population estimates currently available across LMICs (Table 2).

Table 1. Range and average percent of population missing from LMIC censuses based on literature review

Location	Literature review findings			Simulated inaccuracy		
	Minimum	Average	Maximum	Low	Medium	High
Urban-slum	5%	46%	100%	10%	30%	60%
Urban-non-slum	2%	6%	15%	2%	5%	10%
Rural	2%	7%	13%	2%	5%	10%

Table 2. Number of households simulated in the "true" population and 15 realistic scenarios of census outdatedness and inaccuracy, by household type

*Low inaccuracy: missing 2% rural and urban-non-slum households, and 10% urban-slum households. **Medium inaccuracy**: missing 5% rural and urban-non-slum households, and 30% urban-slum households. **High inaccuracy**: missing 10% rural and urban-non-slum households, and 60% urban-slum households.*

Year	No inaccuracy	Low inaccuracy	Medium inaccuracy	High inaccuracy
2016 (current)				
Urban slum	35,001	31,500	24,500	14,000
Urban non-slum	57,843	56,677	54,942	52,073
Rural	4,823	4,735	4,590	4,326
2011 (5 years old)				
Urban slum	28,583	25,724	20,008	11,433
Urban non-slum	55,680	54,566	52,895	50,122
Rural	5,175	5,071	4,917	4,647
2006 (10 years old)				
Urban slum	18,018	16,216	12,612	7,207
Urban non-slum	49,742	48,747	47,258	44,769
Rural	4,146	4,063	3,935	3,730
2001 (15 years old)				
Urban slum	13,149	11,834	9,204	5,259
Urban non-slum	41,700	40,866	39,612	37,514
Rural	3,731	3,656	3,547	3,373

Simulating realistic gridded population datasets

To simulate realistic gridded population datasets, we aggregated each of the simulated household populations to EA or constituency (second-level administrative unit) boundaries, and applied the WorldPop-Global-Unconstrained modelling technique (for a total of 32 datasets). We applied the WorldPop-Global-Unconstrained model in three phases as described in their method publication [20] (Figure 5, Table 3). In the first phase (A), a non-parametric Random Forest ensemble machine-learning algorithm grows a "forest" of decision trees for each input unit (EA or constituency) [67]. Each Random Forest tree is a model of the potential relationships between multiple auxiliary covariates and census population counts. In the Random Forest modelling workflow, this is where model uncertainty is calculated – at the scale of the input population areal unit. In the second phase (B), all of the covariates are prepared in 100x100m cells. In this phase, the split values of each classification tree developed in phase A are used to parameterize corresponding regression models to predict population density within 100x100m cells [20]. For each cell, the predicted population values from all regression models are averaged to make a single population estimate, though these population estimates are not pycnophylactic, meaning that estimates in cells do not necessarily sum to the original areal unit population. Thus the WorldPop-Global-Unconstrained workflow involves a third phase (C) outside of the Random Forest model to normalize cell-level predicted population densities to preserve census input population counts [20].

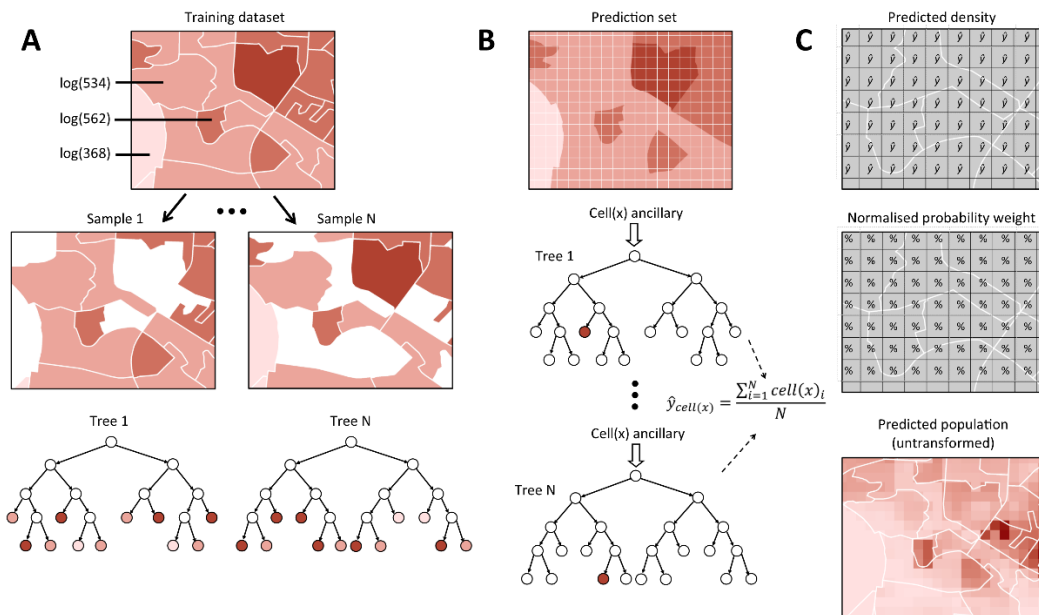



Figure 5. Overview of WorldPop-Global Random Forest Modelling Workflow

A) Each decision tree in the ensemble is built upon a random bootstrap sample of the log-transformed population and ancillary data by administrative unit. B) Population density prediction for each cell $y_{cell(x)}$ is based on an average of the individual trees. C) Predicted cell densities are normalized by administrative unit and used to dasymetrically disaggregate log-transformed administrative unit population, then transformed to predict population per cell.

Table 3. Covariate data sources for Random Forest gridded population estimates

OSM: OpenStreetMap; VIIRS: Visible Infrared Imaging Radiometer Suite; DMSP-OLS: Defence Meteorological Satellite Program Operational Linescan System; ESA-CCI-LC: European Space Agency Climate Change Initiative Land Cover; UNEP-WSMS: UN Environment Programme World Conservation Monitoring Centre; IUCN: International Union for Conservation of Nature; NOAA: US National Oceanic and Atmospheric Administration; CIESIN: Center for International Earth Science Information Network; DLR EOC: German Aerospace Center Earth Observation Center

Name	Description (Year)	Original scale	Original source
cov_road	Distance to OSM major roads (2016)	Vector, <30 m	OpenStreetMap [68]
cov_intsec	Distance to OSM major road intersections (2016)	Vector, <30 m	OpenStreetMap [68]
cov_waterw	Distance to OSM major waterways (2016)	Vector, <30 m	OpenStreetMap [68]
cov_wdpa	Distance to IUCN nature reserve (2000-17)	30" (~900 m)	UNEP-WCMS & IUCN [69]
cov_viirs	Resampled VIIRS night-time lights (2012-2016)	30" (~900 m)	NOAA [70]
cov_dmsp	Resampled DMSP-OLS night-time lights (2011)	30" (~900 m)	NOAA & Zhang, et al. [71,72]
cov_tt50k	Resampled travel time to cities of 50,000+ (2000)	30" (~900 m)	Weiss, et al. [73]
cov_001	Distance to cultivated areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_040	Distance to woody areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_130	Distance to cultivated areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_140	Distance to herbaceous areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_150	Distance to sparse vegetation areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_160	Distance to aquatic vegetation areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_190	Distance to urban areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_200	Distance to bare areas (2015)	9" (~300 m)	ESA CCI – LC [74]
cov_cciwat	Distance to ESA-CCI-LC inland waterbodies (2000-12)	4.5" (~150 m)	ESA CCI [75]
cov_slope	SRTM-based slope (2000)	3" (~90 m)	de Ferranti [76,77]
cov_topo	SRTM-based elevation (2000)	3" (~90 m)	de Ferranti [76,77]
cov_coast	Distance to open-water coastline (2000-20)	3" (~90 m)	CIESIN [78]
cov_ghsl	Distance to urban area (2012)	1.26" (~38 m)	Pesaresi, et al. [79]
cov_guf	Distance to settlement built-up areas (2012)	2.8" (~84 m)	DLR EOC [80]
cov_bsgme	Distance to built settlement expansion (2016)	3" (~90 m)	Nieves, et al. [81]
cov_prec	Average total annual precipitation (1970-2000)	30" (~900 m)	Fick and Hijmans [82]
cov_temp	Average annual temperature (1970-2000)	30" (~900 m)	Fick and Hijmans [82]

Analysing cell-level accuracy

To empirically measure cell-level accuracy of the 32 gridded population datasets, we compared each cell-level estimate against the "true" 2016 population count in that cell, then calculated root mean square error (RMSE), a measure of error magnitude that penalises large errors. This was performed on 100x100m cells, and then cells were aggregated and assessed for accuracy at 200x200m, 300x300m, and so on up to 1x1km. To compare RMSE across cells of different geographic sizes, we normalised the statistic by average population (equation 1) and by area (equation 2). The former represents RMSE of population counts expressed as a portion of the population [83], while the latter represents RMSE of population density per hectare (100x100m unit) [84]. We evaluated RMSE in urban-slum, urban-non-slum, and rural cells separately. In the calculation of RMSE, y_i is the "true" population count in cell i, $\hat{y}_i$ is the gridded population estimate in cell i, D_i is the "true" population density per hectare, $\widehat{D}_i$ is the estimated population density per hectare, and n is the number of grid cells.

$$Pop-adj\ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \div \frac{\sum_{i=1}^{n}(y_i)}{n} \tag{1}$$

$$Area-adj\ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{D}_i - D_i)^2}{n}} \tag{2}$$

To better understand the mechanics of the WorldPop-Global-Unconstrained model and workflow, we calculated bias, a measure of error direction and magnitude, for the two gridded population datasets derived from "true" population counts to tease out accuracy related to the model and covariate datasets alone. Bias (equation 3) reveals to what extent cell-level estimates are systematically under- or over-estimated, and reflects over/under-counts in cells of different sizes that a user might encounter in the field. Relative bias (equation 4) refers to bias normalised by the average "true" population which enables comparisons across grid scales. As above, bias and relative bias were assessed in 100x100m cells as well as cell sizes that ranged up to 1x1km, and separately in urban versus rural areas.

$$Bias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n} \tag{3}$$

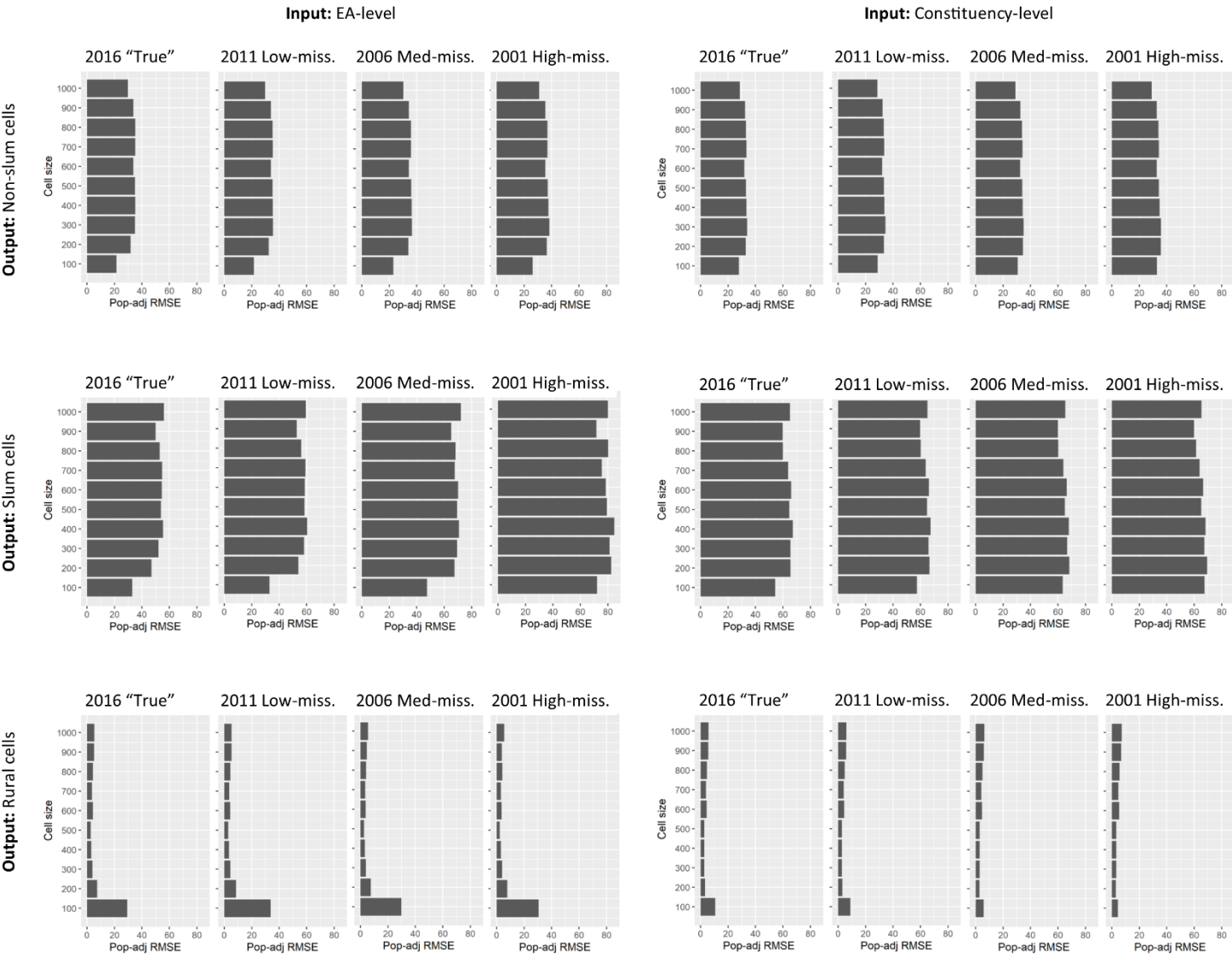
$$Relative\ bias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n} \Big/ \frac{\sum_{i=1}^{n}(y_i)}{n} \tag{4}$$

To assess the degree to which non-zero population estimates in the WorldPop-Global-Unconstrained dataset resulted in misallocation of population, a third statistic was calculated counting the entire modelled population in Khomas that was misallocated to cells which were unsettled according to the "true" population. For all statistics, we excluded gridded population cell-level estimates of less than 1 person to avoid millions of near-zero cell-level estimates in unsettled areas of Khomas (located outside of Windhoek) from dominating the accuracy assessments.

Results

Neither measure of RMSE differed substantially across the simulated outdated-inaccurate census scenarios (Figures 6 and 7). Furthermore, errors only slightly decreased when the input data were aggregated to EA (finer) rather than constituency (coarser) (Figures 6 and 7). The major driver of RMSE in cells was urban versus rural location, with further difference between urban-slum and urban-non-slum. In urban cells,

Figure 6. Population-adjusted Root mean square error (RMSE) according to input population aggregation, a selection of scenarios, and cell size

1
2

3
4   **Figure 7.** Population density Root mean square error (RMSE) per hectare according to input population aggregation, a selection of scenarios, and cell size

population-adjusted RMSE were substantially smaller than rural cells (Figure 6), but much larger per hectare due to larger population numbers (Figure 7). In urban areas, RMSE per hectare was lowest in 100x100m cells (slum range: 32-72, non-slum range: 21-33), while in rural areas, RMSE per hectare was lowest in cells 300x300m to 500x500m (rural range: 2-54) (Figure 7). Results for select scenarios are presented in Figure 6 ranging from the "true" 2016 population to the most outdated (2001) and inaccurate (missing 10% to 60%) population, though tables of all results are provided in Supplement 5.

Assessment of bias in the two gridded population datasets derived only from "true" 2016 population counts revealed systematic and substantial under-estimates of populations in urban-slum and urban-non-slum cells due solely to the aggregation-level of the input population data and modelling approach, not inaccuracies in the input data (Tables 4 and 5). For example, the average 300x300m urban-slum cell under-estimated the population by more than 350 people (EA-level input) up to 500 people per cell (constituency-level input). For comparison, the average 300x300m non-slum cell was under-estimated by 165 people (constituency-level input) to 187 people (EA-level input), while the average rural cell of the same size was over-estimated by 3 people (constituency-level input) to 14 people (EA-level input) (Table 4). When adjusted for population, the results indicate that for every person estimated in an urban non-slum cell, nearly a whole person is omitted; and for every person estimated in an urban slum cell, one to one and a half people are omitted (Table 5).

Table 4. Bias in gridded population estimates derived from "true" population counts, by output grid cell size and urban/rural location (in cells >=1 estimated person)

Cell size	EA-level input			Constituency-level input		
	Non-slum	Slum	Rural	Non-slum	Slum	Rural
100	0	0	20	-4	-34	7
200	-71	-135	18	-64	-212	6
300	-187	-353	14	-165	-498	3
400	-346	-678	8	-303	-929	-1
500	-549	-1029	3	-483	-1401	-8
600	-769	-1480	-22	-672	-2080	-34
700	-1094	-2114	-33	-981	-2747	-51
800	-1410	-2692	-72	-1247	-3359	-90
900	-1728	-3215	-126	-1576	-4437	-152
1000	-1928	-4453	-126	-1770	-5834	-167

Table 5. Population-adjusted bias in gridded population estimates derived from "true" population counts, by output grid cell size and urban/rural location (in cells >=1 estimated person)

Cell size	EA-level input			Constituency-level input		
	Non-slum	Slum	Rural	Non-slum	Slum	Rural
100	0.00	0.00	3.36	-0.10	-0.64	1.28
200	-0.58	-0.74	1.90	-0.52	-1.16	0.67
300	-0.76	-0.94	1.07	-0.67	-1.32	0.23
400	-0.85	-1.04	0.52	-0.74	-1.43	-0.08
500	-0.92	-1.07	0.15	-0.81	-1.46	-0.40
600	-0.96	-1.09	-0.77	-0.84	-1.53	-1.21
700	-0.99	-1.15	-1.02	-0.89	-1.50	-1.59
800	-1.03	-1.21	-1.70	-0.91	-1.51	-2.12
900	-1.00	-1.14	-2.41	-0.91	-1.58	-2.92
1000	-1.05	-1.09	-2.45	-0.96	-1.43	-3.25

Table 6 summarises the percent of the estimated population misallocated to "true" unsettled cells. For this analysis, no cells in the estimated population were excluded. Roughly 20% (EA-level input) or 10% (constituency-level input) of the population was misallocated to unsettled 100x100m cells (Table 6). However, as cells were aggregated, the percent of misallocated population dropped precipitously. For example, at 300x300m, approximately 2% (EA-level input) or 1% (constituency-level input) of Khomas's population was misallocated to unsettled cells. This indicated that most of the population was disaggregated to unsettled cells within, or near to, settlements. The rates of misallocation were similar when cells with less than one person were excluded (not reported).

Table 6. Percent of the overall population that is misallocated to unsettled cells (no exclusion), by aggregation level of the input data and output grid cell size

Grid cell size (m²)	EA-Level Input	Constituency-Level Input
100	20.8 %	12.5 %
200	5.0 %	2.6 %
300	2.2 %	1.0 %
400	1.3 %	0.5 %
500	0.8 %	0.3 %
600	0.6 %	0.2 %
700	0.4 %	0.1 %
800	0.3 %	0.1 %
900	0.3 %	0.1 %
1000	0.2 %	0.1 %

Discussion

This is among the first accuracy assessments of a top-down gridded population model at the grid-cell level, and the first that we know of in a LMIC setting. By developing a simulated realistic population and several scenarios of the population with realistic levels of outdatedness and inaccuracy, we were able to evaluate the accuracy of a gridded population modelling approach, as well as assess the impact of outdated-inaccurate census inputs on estimates. In this paper, we evaluated just one of several gridded population models – WorldPop-Global-Unconstrained. In this analysis of one simulated population in the setting of Khomas, Namibia using a WorldPop-Global-Unconstrained population model, cell-level inaccuracies between urban versus rural areas dominated the results. There was limited evidence in this context that outdated or inaccurate census data played a major role in cell-level inaccuracy of gridded population estimates. Here we address three potential sources of the cell-level inaccuracies observed.

The first issue is specific to the modelling approach of WorldPop-Global-Unconstrained relating to the exclusion of input administrative units with zero population and use of a log-transformation on those population counts before use of the Random Forest model. While this procedure ensures that population counts are normally distributed during modelling, it also means that unpopulated cells are assigned a very small fraction of a person [20]. A possible concern is that non-zero population estimates across millions of unsettled cells could result in a sizable portion of the population being misallocated. Our analysis of misallocation, however, indicates that this phenomenon played only a minor role in cell-level inaccuracies, if at all. Table 6 demonstrates that even in this context of vast unsettled areas, only a small portion of

Khomas's population was misallocated to cells far from actual settlements. Nearly all of the population was estimated to be in cells within 200 to 300 metres of the "true" population.

Most global gridded population producers constrain estimates to settled cells as defined with a settlement layer (e.g. LandScan [22,85], GHP-POP [17,18], HRSL [19], GRID3 [26,86], WPE [24]). Until recently, these settlement layers tended to be relatively coarse (e.g. GHS-BUILT 1x1km [87]) and/or had a tendency to omit areas with few sparse buildings (e.g. GUF [80]) which could result in under-estimation of the population in rural areas and over-allocation of the population in urban areas. However, new free very high resolution Sentinel-2 imagery, and major leaps in computing power for extracting building footprints and other features from imagery, have enabled development of several new detailed settlement layers in the last one or two years (e.g., GHS-BUILT-S2 [88], Maxar/Ecopia [89]). Recently, WorldPop-Global produced a constrained global gridded population estimate for 2020 that uses the same input population and covariate datasets as its unconstrained model plus several building footprint metrics (in Africa), and then masks all 100x100m cells without building footprints (in Africa) or built settlement (rest of the world) [33], eliminating the issue of non-zero population estimates in unsettled cells.

The second potential source of inaccuracy relates to covariate resolution and the relationship of covariates with population density. This issue seems to have contributed more substantially to errors in this analysis, particularly within the city of Windhoek. A number of the Random Forest model covariates, such a land cover type and night-time lights, had an original resolution substantially larger than 100x100m which could have resulted in a "halo" effect around settlements, causing populations to be disaggregated to cells near a settlement, but not directly over it. Table 5 provides evidence of this; the accuracy of the estimated population distribution, and correct allocation of population to settled cells, both performed well when the estimated population was aggregated to 300x300m or larger. Other covariates, such as distance to roads and intersection locations were available at very fine spatial resolution and thus were precise at the 100x100m scale. Although they are good indicators of a settlement, they are not necessarily good indicators of higher or lower population density within a settlement. The lack of fine-scale covariates associated with population density within cities and towns likely explains a portion of the cell-level error observed in Khomas's urban population. Other issues that might further decrease local spatial accuracy are temporal miss-match of covariates [13] and covariate spatial autocorrelation [90]. With the recent release of several building footprint datasets (e.g., Maxar/Ecopia in most of Africa [89], Bing in Tanzania and Uganda [91]), several new covariate layers have been created by the WorldPop team including number of buildings and total area of buildings in 100x100m cells [92]. Building footprints are likely associated with population density within settlements and have a finer spatial resolution than 100x100m, making it a potentially powerful covariate to differentiate low and high population density within urban areas in any gridded population model. The WorldPop team, among other gridded population producers, is currently working to test and incorporate building footprint datasets into gridded population models.

The third potential source of cell-level inaccuracies is use of average population densities from large administrative units to estimate population density in much smaller grid cells. This is known as the ecological fallacy [93], and probably played the largest role in cell-level inaccuracies, especially within urban areas. Population densities are used by the Random Forest model to establish relationships between covariates and population *density* (total population divided by total area), not population totals. Even with perfect covariates and exclusion of unsettled areas, this would mean that cells with high "true" population counts are likely to be severely underestimated because the geographic size of input population units are larger (and population densities are smaller) than the output grid cells. When this happens, population counts that are not allocated to the densest cells will instead be allocated to other less dense cells in the same input areal unit. Tables 4 and 5 provide strong evidence of this issue with the population in urban cells, especially urban-slum cells, systematically underestimated regardless of cell size.

Although these results apply only to the WorldPop-Global-Unconstrained model, we can speculate about how these results might apply to other gridded population datasets. Most top-down gridded population datasets use average population densities from large input areal units in some way to populate smaller grid cells, and are thus likely subject to similar errors linked with the ecological fallacy. The High Resolution Settlement Layer (HRSL), for example, uses uniform areal disaggregation of the population from input units (e.g., EA) to 30x30m grid cells which contain a building footprint [19], and the Global Human Settlement GHS-POP dataset takes a similar approach disaggregating input populations into 250x250m cells that are classified as settled [17,18]. This problem of aggregate averages being used in smaller grid cells is accentuated when input areal units are geographically large because average densities tend to be smaller.

In these cases, WorldPop-Global incorporates training data from a neighbouring country that has finer-scale input population counts [20]. Our analysis showed, however, that even when relatively small geographic units (census EAs) were used as the input population area unit, urban slum and non-slum cell-level errors were still substantial, and cell-level accuracy with EA-level input was only marginally improved compared to constituency-level input (Figure 7). This suggests that finer-scale training data (e.g., closer to 100x100m) should be incorporated during the model training phase, particularly from high-density urban areas, to ensure that the WorldPop Random Forest model contains sufficiently large population density values to assign to urban cells. Fine-scale training datasets might come from existing household survey enumerations (e.g., World Bank Living Standards Measurement Surveys), or slum community profiles such as those published on the Know Your City Campaign website [94]. Even if fine-scale densities are only available for a small sample of locations, they would provide the Random Forest model with more accurate maximum population values at the scale of 100x100m during model training. The final Random Forrest model only disaggregates the population counts provided with the full coverage administrative units (e.g. EAs); however, the population densities derived during model training would likely improve cell-level accuracy within urban areas.

This analysis of WorldPop-Global-Unconstrained data raises broader questions about the cell-level accuracy of gridded population estimates in urban areas, especially the densest parts of cities such as in slums, informal settlements, and neighbourhoods with high-rise apartment buildings [95–97]. New datasets derived from very high resolution satellite imagery, in particular building footprints, are a promising new covariate to reduce the "halo" effect of populations misallocated nearby, but not directly over, the highest density cells. More work will be needed to improve building footprint datasets by distinguishing residential and non-residential buildings to avoid population being misallocated to business districts, factories, universities, airports, and other non-residential cells [98,99]. These two steps – use of building footprint covariates and finer-scale training data - stand to improve cell-level accuracy of gridded population datasets derived from complex models, including all WorldPop-Global datasets as well as LandScan [22,23], WPE [24], and GRID3 [26,86]. Gridded population datasets that do not vary (weight) population densities within areal units (e.g., HRSL [19], GHS-POP [17,18], GPW [15,16]) should be used cautiously within urban areas, as cell-level inaccuracies are likely to be high.

CONCLUSIONS

Global gridded population data initiatives aim to fill a gap in available disaggregated and current population counts to ensure that everyone is counted and that all needs are met in development initiatives. However, many gridded population datasets are not evaluated for accuracy at the small spatial scales. This analysis of one simulated population in one setting revealed substantial and systematic under-estimation of population in slums. Further analyses of other gridded population datasets are needed across diverse settings. However, if severe under-estimates in slums and other high-density urban areas are widespread, this means that gridded population datasets might unintentionally reinforce marginalisation of the urban poorest by omitting them from maps and population counts. We offer two suggestions to address this challenge with

inclusion of finer-scale training data and new building footprints data as a covariate. Given the increased use of gridded population datasets for monitoring health and development outcomes in small areas, it is imperative that gridded population datasets area assessed for cell-level accuracy and improved where possible.

Acknowledgements

We would like to thank Drs. Angela Luna Hernandez and Ryan Engstrom for their feedback on an earlier version of this work.

Funding

Dana R. Thomson was funded by the Economic and Social Research Council grant number ES/5500161/1.

References

1. UN-Habitat. World Cities Report 2020: The Value of Sustainable Urbanization. Nairobi Kenya: UN-Habitat; 2020.
2. Utazi CE, Wagai J, Pannell O, Cutts FT, Rhoda DA, Ferrari MJ, et al. Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: Analysis of recent household surveys. Vaccine. 2020;38(14):3062–71.
3. Ruktanonchai CW, Ruktanonchai NW, Nove A, Lopes S, Pezzulo C, Bosco C, et al. Equality in maternal and newborn health: Modelling geographic disparities in utilisation of care in five East African countries. PLoS One. 2016;11(8):e0162006.
4. Cutts FT, Ferrari MJ, Krause LK, Tatem AJ, Mosser JF. Vaccination strategies for measles control and elimination: time to strengthen local initiatives. BMC Med. 2021;19(1):1–8.
5. United Nations Statistics Division (UNSD). 2020 world population and housing census programme [Internet]. Census dates for all countries. 2019 [cited 2020 Jan 13]. Available from: https://unstats.un.org/unsd/demographic-social/census/censusdates/
6. Bekele S. The accuracy of demographic data in the Ethiopian Censuses. East Afr Soc Sci Res Rev. 2017;33(1):15–38.
7. Carr-Hill R. Missing millions and measuring development progress. World Dev. 2013;46:30–44.
8. Ahonsi BA. Deliberate falsification and census-data in Nigeria. Afr Aff (Lond). 1988 Oct;87(349):553–62.
9. Okolo A. The Nigerian Census: Problems and prospects. Am Stat. 1999;53(4):321–5.
10. Yin S. Objections surface over Nigerian census results [Internet]. Population Reference Bureau. 2007 [cited 2020 Jan 13]. p. 1–3. Available from: www.prb.org/Publications/Articles/2007/ObjectionsOverNigerianCensus.aspx
11. UN Department of Economic and Social Affairs (UN-DESA). World Urbanization Prospects: The 2019 Revision [Internet]. 2019 [cited 2020 Jan 13]. Available from: https://population.un.org/wup/DataQuery/
12. Thomson DR, Rhoda DA, Tatem AJ, Castro MC, Foundation F, Place C, et al. Gridded population survey sampling: A systematic scoping review of the field and strategic research agenda. Int J Health Geogr. 2020;19:34.
13. Leyk S, Gaughan AE, Adamo SB, de Sherbinin A, Balk D, Freire S, et al. Allocating people to pixels: A review of large-scale gridded population data products and their fitness for use. Earth Syst Sci Data Discuss. 2019;11:1385–409.
14. POPGRID Data Collaborative. Leaving no one off the map: A guide for gridded population data for sustainable development. New York NY USA; 2020.
15. Doxsey-Whitfield E, MacManus K, Adamo SB, Pistolesi L, Squires J, Borkovska O, et al. Taking advantage of the improved availability of census data: A first look at the Gridded Population of the

World, Version 4. Pap Appl Geogr. 2015 Jul 3;1(3):226–34.

16. Center for International Earth Science Information Network (CIESIN), Columbia University. Gridded Population of the World v4 [Internet]. 2016 [cited 2017 Feb 2]. Available from: http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/sets/browse

17. Pesaresi M, Ehrlich D, Florczyk AJ, Freire S, Julea A, Kemper T, et al. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. Ispra Italy: European Commission Joint Research Centre; 2016. 67 p.

18. European Commission JRC. Global human settlement population model (GHS-POP) [Internet]. 2020 [cited 2020 Oct 27]. Available from: https://ghsl.jrc.ec.europa.eu/data.php

19. Facebook Connectivity Lab, CIESIN. High Resolution Settlement Layer (HRSL) [Internet]. 2016 [cited 2020 Oct 27]. Available from: https://data.humdata.org/dataset/highresolutionpopulationdensitymaps

20. Stevens FR, Gaughan AE, Linard C, Tatem AJ. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS One. 2015;10(2):e0107042.

21. WorldPop. WorldPop 2000-2020 UN-Adjusted Unconstrained 100m [Internet]. 2020 [cited 2020 Oct 27]. Available from: www.worldpop.org/geodata/listing?id=69

22. Dobson JE, Brlght EA, Coleman PR, Worley BA, Bright EA, Coleman PR, et al. LandScan: A global population database for estimating populations at risk. Photogramm Eng Remote Sensing. 2000 Jul;66(7):849–57.

23. Oak Ridge National Laboratories. LandScan Data Availability [Internet]. 2017 [cited 2017 Feb 2]. Available from: http://www.ornl.gov/sci/landscan/landscan_data_avail.shtml

24. Frye C, Nordstrand E, Wright DJ, Terborgh C, Foust J. Using classified and unclassified land cover data to estimate the footprint of human settlement. Data Sci J. 2018;17:1–12.

25. Long JF, McMillen DB. A survey of census bureau population projection methods. Clim Change. 1987;11:141–77.

26. Leasure DR, Jochem WC, Weber EM, Seaman V, Tatem AJ. National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. Proc Natl Acad Sci U S A. 2020;117(39):24173–9.

27. Leasure DR, Dooley CA, Bondarenko M, Tatem AJ. peanutButter: An R package to produce rapid-response gridded population estimates from building footprints, version 0.3.0 [Internet]. 2020 [cited 2020 Oct 27]. Available from: https://apps.worldpop.org/peanutButter/

28. Hay S, Noor A, Nelson A, Tatem A. The accuracy of human population maps for public health application. Trop Med Int Heal. 2005 Oct;10:1073–86.

29. Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ. High resolution population distribution maps for Southeast Asia in 2010 and 2015. PLoS One. 2013;8(2):e55882.

30. Bondarenko M, Nieves JJ, Stevens FR, Gaughan AE, Tatem A, Sorichetta A. wpgpRFPMS: Random Forests population modelling R scripts, version 0.1.0. Southampton UK; 2020.

31. Lloyd CT, Chamberlain H, Kerr D, Yetman G, Pistolesi L, Stevens FR, et al. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. Big Earth Data. 2019;3(2):108–39.

32. WorldPop. WorldPop-Global Namibia covariates [Internet]. 2018 [cited 2020 Jan 13]. Available from: ftp://ftp.worldpop.org.uk/GIS/Covariates/Global_2000_2020/NAM/

33. WorldPop. Top-down estimation modelling: constrained vs unconstrained [Internet]. 2020 [cited 2020 Oct 27]. Available from: www.worldpop.org/methods/top_down_constrained_vs_unconstrained

34. Archila Bustos MF, Hall O, Niedomysl T, Ernstson U. A pixel level evaluation of five multitemporal global gridded population datasets: A case study in Sweden, 1990–2015. Popul Environ. 2020;42(2):255–77.

35. Calka B, Bielecka E. GHS-POP accuracy assessment: Poland and Portugal case study. Remote Sens. 2020;12(7):1105.

36.   United Nations Statistics Division (UNSD). Report on the results of a survey on census methods used by countries in the 2010 census round. New York NY USA; 2010. (Working paper). Report No.: UNSD/DSSB/1.

37.   Cobham A. Uncounted: Power, inequalities and the post-2015 data revolution. Development. 2014;57(3–4):320–37.

38.   Thomson DR, Kools L, Jochem WC. Linking synthetic populations to household geolocations: A demonstration in Namibia. Data. 2018;3(3):30.

39.   Namibia Statistics Agency (NSA). Namibia 2011 Population and Housing Census main report. Windhoek Namibia; 2011.

40.   Newaya TP. Rapid urbanization and its influence on the growth of informal settlements in Windhoek, Namibia [dissertation]. Cape Peninsula University of Technology; 2010.

41.   Lai S, Erbach-Schoenberg E zu, Pezzulo C, Ruktanonchai NW, Sorichetta A, Steele J, et al. Exploring the use of mobile phone data for national migration statistics. Palgrave Commun. 2019;5(1).

42.   International Organization for Migration (IOM). Migration in Namibia. Geneva Switzerland; 2015.

43.   WorldPop. Africa 1km internal migration flows [Internet]. 2016 [cited 2020 Jan 13]. Available from: https://www.worldpop.org/geodata/summary?id=1281

44.   Alfons A, Kraft S, Templ M, Filzmoser P. Simulation of close-to-reality population data for household surveys with application to EU-SILC. Stat Methods Appl. 2011;20(3):383–407.

45.   Oliveira LC de S, Freitas MPS de, Dias MRML, Nascimento CMF, Mattos E da S, Junior JJAR. Censo Demográfico 2000 - Pesquisa de avaliação da cobertura da coleta. Rio de Janeiro Brazil; 2003.

46.   Korale RBM. Post Enumeration Survey 2001 [Nepal Population Census] Draft Report. Kathmandu Nepal; 2002.

47.   Maro R. Post Enumeration Survey Tanzania Experience [Internet]. Workshop on the 2010 World programme on population and housing censuses: census evaluation and post enumeration surveys, for English-speaking African countries. 2009 [cited 2020 Jan 13]. p. 12. Available from: https://unstats.un.org/unsd/demographic/meetings/wshops/Ethiopia_14_Sept_09/Country_Present ations/Tanzania.ppt

48.   Uganda Bureau of Statistics (UBS). Post enumeration survey: 2002 Uganda populaiton and housing census. Entebbe Uganda; 2005.

49.   Ghana Statistical Service (GSS). 2010 Population and Housing Census Post Enumeration Survey Report. Accra Ghana; 2012.

50.   Central Statistical Office (CSO). [Zambia] 2010 Census of Population and Housing Post Enumeration Survey (PES). Lusaka Zambia; 2013.

51.   Bangladesh Institute of Development Studies (BIDS). Report of the post enumeration check (PEC) of the [Bangaldesh] Population and Housing Census, 2011. Dhaka Bangladesh; 2012.

52.   National Statistical Commission (NSC). Census of India 2011: Report on post enumeration survey. New Delhi India; 2014.

53.   Statistics South Africa (SSA). Census 2011 Post-enumeration survey. Pretoria South Africa; 2012.

54.   National Institute of Statistics of Rwanda (NISR). Post enumeration survey report: Fourth Population and Housing Census, Rwanda, 2012. Kigali Rwanda; 2010.

55.   Agarwal S. The state of urban health in India: Comparing the poorest quartile to the rest of the urban population in selected states and cities. Environ Urban. 2011;23(1):13–28.

56.   Sabry S. How poverty is underestimated in Greater Cairo, Egypt. Environ Urban. 2010;22(2):523–41.

57.   Stark L, Rubenstein BL, Pak K, Taing R, Yu G, Kosal S, et al. Estimating the size of the homeless adolescent population across seven cities in Cambodia. BMC Med Res Methodol. 2017;17:1–8.

58.   Treiman DJ, Mason WM, Lu Y, Pan Y, Qi Y, Song S. Observations on the design and implementation of sample surveys in China. Los Angeles CA USA; 2005. Report No.: CCPR-006-05.

59.   Carr-hill ROY. Improving population and poverty estimates with citizen surveys: Evidence from East Africa. World Dev. 2017;93:249–59.

60.   Ebenstein A, Zhao Y. Tracking rural-to-urban migration in China: Lessons from the 2005 inter-census

population survey. Popul Stud (NY). 2015;69(3):337–53.

61. Gidado SO, Nguku PJ, Ndadilnasiya Waziri M, Ohuabunwo C, Etsano A, Mahmud MZ, et al. Polio field census and vaccination of underserved populations Northern Nigeria, 2012-2013. Morb Mortal Wkly Rep. 2013;62(33):653–7.

62. Gurgel RQ, Da Fonseca JDC, Neyra-Castañeda D, Gill G V., Cuevas LE. Capture-recapture to estimate the number of street children in a city in Brazil. Arch Dis Child. 2004;89:222–4.

63. Jiang Q, Li X, Sánchez-barricarte JJ. Data Uncertainties in China's Population. Asian Soc Sci. 2015;11(13):200–5.

64. Karanja I. An enumeration and mapping of informal settlements in Kisumu, Kenya, implemented by their inhabitants. Environ Urban. 2010;22(1):217–39.

65. Kronenfeld DA. Afghan refugees in Pakistan: Not all refugees, not always in Pakistan, not necessarily Afghan? J Refug Stud. 2008;21(1):43–63.

66. Lucci P, Bhatkal T, Khan A. Are we underestimating urban poverty? World Dev. 2018;103:297–310.

67. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

68. OpenStreetMap contributors. OpenStreetMap base data [Internet]. 2000 [cited 2020 Jan 13]. Available from: www.openstreetmap.org

69. UNEP-WCMS, IUCN. World database on protected areas & Global database on protected areas management effectiveness [Internet]. UN Environment Programme-World Conservation Monitoring Centre & International Union for Conservation of Nature. 2016 [cited 2020 Jan 13]. Available from: https://www.protectedplanet.net/

70. [USA] National Oceanic and Atmospheric Administration (NOAA). VIIRS nighttime lights [Internet]. 2012 [cited 2020 Jan 13]. Available from: https://maps.ngdc.noaa.gov/viewers/VIIRS_DNB_nighttime_imagery/index.html

71. [USA] National Oceanic and Atmospheric Administration (NOAA). Version 4 DMSP-OLS Nighttime Lights Time Series [Internet]. 2017 [cited 2020 Jan 13]. Available from: www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

72. Zhang Q, Pandey B, Seto KC. A Robust Method to Generate a Consistent Time Series From DMSP / OLS Nighttime Light Data. IEEE Trans Geosci Remote Sens. 2016;54(10):1–11.

73. Weiss D, Nelson A, Gibson H, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature. 2018;553(7688):333–6.

74. European Space Agency - Climate Change Initiative (ESA-CCI). Land Cover CCI Product - Annual LC maps from 2000 to 2015 (v2.0.7) [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://maps.elie.ucl.ac.be/CCI/viewer/

75. European Space Agency - Climate Change Initiative (ESA-CCI). Land cover CCI product - MERIS Waterbody product v4.0 (150 m) [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://maps.elie.ucl.ac.be/CCI/viewer/

76. de Ferranti J. Digital elevation data - Viewfinder panoramas [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://www.viewfinderpanoramas.org/dem3.html

77. de Ferranti J. Digital elevation data: SRTM void fill - Viewfinder panoramas [Internet]. 2017 [cited 2020 Jan 13]. Available from: www.viewfinderPanoramas.org/voidfill.html

78. Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4.11 (GPWv4.11) [Internet]. 2018 [cited 2020 Oct 27]. Available from: https://doi.org/10.7927/H4F47M65

79. European Commission. Global human settlement city model (GHS-SMOD) [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://ghsl.jrc.ec.europa.eu/faq.php

80. DLR Earth Observation Center. Global Urban Footprint (GUF) [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-11725/20508_read-47944/

81. Nieves JJ, Sorichetta A, Linard C, Bondarenko M, Steele JE, Stevens FR, et al. Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night. Comput Environ Urban Syst. 2020;80(May 2019):101444.

82.   Fick SE, Hijmans RJ. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 2017;37(12):4302–15.

83.   Gregory IN. An evaluation of the accuracy of the areal interpolation of data for the analysis of long-term change in England and Wales. In: GeoComputation. Greenwich UK; 2000.

84.   Bozheva AM, Petrov AN, Sugumaran R. The effect of spatial resolution of remotely sensed data in dasymetric mapping of residential areas. GIScience Remote Sens. 2005;42(2):113–30.

85.   Oak Ridge National Laboratories (ORNL). LandScan documentation [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://web.ornl.gov/sci/landscan/landscan_documentation.shtml

86.   CIESIN, UNFPA, WorldPop, Flowminder. Geo-Referenced Infrastructure and Demographic Data for Development (GRID3) [Internet]. 2018 [cited 2020 Jan 13]. Available from: www.grid3.org

87.   EC-JRC. GHS-BUILT [Internet]. 2019 [cited 2021 Feb 16]. Available from: https://ghsl.jrc.ec.europa.eu/ghs_bu2019.php

88.   Corbane C, Sabo F, Politis P, Syrris V. HS-BUILT-S2 R2020A - GHS built-up grid, derived from Sentinel-2 global image composite for reference year 2018 using Convolutional Neural Networks (GHS-S2Net). European Commission, Joint Research Centre (JRC); 2020.

89.   Maxar. Satellite Imagery [Internet]. 2019 [cited 2021 Apr 2]. Available from: www.maxar.com/products/satellite-imagery

90.   Sinha P, Gaughan AE, Stevens FR, Nieves JJ, Sorichetta A, Tatem AJ. Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. Comput Environ Urban Syst. 2019;75:132–45.

91.   Microsoft. Building Footprints [Internet]. AI for Humanitarian Action program. 2020 [cited 2020 Dec 16]. Available from: www.microsoft.com/en-us/maps/building-footprints

92.   WorldPop. WorldPop Open Population Repository - Buildings [Internet]. 2020 [cited 2020 Dec 16]. Available from: https://wopr.worldpop.org/?/Buildings

93.   Selvin HC. Durkheim's suicide and problems of empirical research. Am J Sociol. 1958;63(6):607–19.

94.   SDI. Know Your City [Internet]. Slum/Shack Dwellers International Dwellers International. 2016 [cited 2020 Mar 1]. Available from: http://knowyourcity.info/explore-our-data/

95.   Nuissl H, Heinrichs D. Slums: Perspectives on the definition, the appraisal and the management of an urban phenomenon. J Geogr Soc Berlin. 2013;144(2):105–16.

96.   Ezeh A, Oyebode O, Satterthwaite D, Chen Y, Ndugwa R, Sartori J, et al. The history, geography, and sociology of slums and the health problems of people who live in slums. Lancet. 2017;389:547–58.

97.   Mahabir R, Croitoru A, Crooks A, Agouris P, Stefanidis A. A Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities. Urban Sci. 2018;2(1):8.

98.   Sturrock HJW, Woolheater K, Bennett AF, Andrade-Pacheco R, Midekisa A. Predicting residential structures from open source remotely enumerated data using machine learning. PLoS One. 2018;13(9):1–10.

99.   Lloyd CT, Sturrock HJW, Leasure DR, Jochem WC, Lázár AN, Tatem AJ. Using GIS and machine learning to classify residential status of urban buildings in low and middle income settings. Remote Sens. 2020;12(23):1–20.

Supporting information

S1 File. Simulating a population in Khomas, Namibia

S2 File. Simulated population in Khomas, Namibia

S3 Table. Percent of population missing from LMIC censuses by source

S4 Figure. Exclusion of grid cells with less than one estimated person in accuracy analysis

S5 Table. Root Mean Square Error (RMSE) statistics for all scenarios

Supplement 1. Simulating population in Khomas, Namibia

Supplement to Thomson DR, Leasure DR, Bird T, Tzavidis N, Tatem AJ. 2021. How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia.

The simulation in Khomas, Namibia followed the same steps outlined by Thomson and colleagues (2018)[1] for a simulated population in Oshikoto, Namibia:

(1) Use of a supervised clustering k-means algorithm to define realistic and distinct types of households in Khomas, Namibia based on eight variables in the 2013 Demographic and Health Survey (DHS) (Table S1.1, A) that were also present in a 20% census microdata sample (Table S1.1, B): urban, improved toilet, improved water source, sufficient sleeping space, durable structure, non-solid fuel for cooking, whether the head of household had any formal education, and whether there were any children under age five. A dendrogram showing the Euclidean distance between each pair of child clusters and their parent cluster in the k-means analysis indicated a sensible cut-off value of 1.0 to define four easy-to-interpret household types: urban poor, urban non-poor, rural poor, rural non-poor (Figure S1.1).

Cluster	rural	noedu	any_u5	unimp. toilet	unimp. water	inadeq. structure	limited space	solid fuel	Household type label
Type 1	0.00	0.02	0.30	0.30	0.01	0.08	0.10	0.01	Urban non-poor
Type 2	0.00	0.10	0.35	0.96	0.01	0.54	0.09	0.24	Urban poor
Type 3	1.00	0.33	0.40	1.00	0.13	0.53	0.14	1.00	Rural poor
Type 4	1.00	0.04	0.21	0.43	0.00	0.08	0.00	0.00	Rural non-poor
Khomas	0.06	0.05	0.31	0.50	0.01	0.21	0.09	0.09	

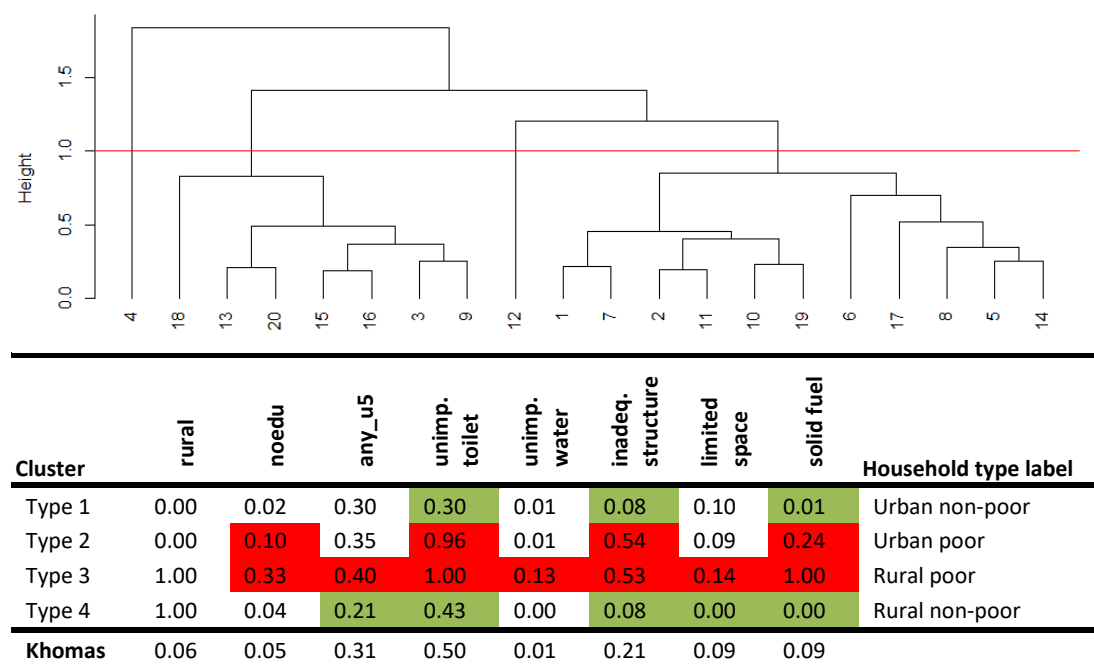



Figure S1.1. Dendrogram & k-mean scores of unique household types in Khomas, Namibia based on 2013 DHS

(2) Steps 2 and 3 involve prediction of household type probability surfaces. Although we only care about the household type probabilities in Khomas, we model probability surfaces for all of Namibia due to the limited number of 2013 DHS primary sampling units (PSUs) in Khomas (53 PSUs Khomas, 550 PSUs Namibia) available to train a model. Thus, in step 2, we processed 19 spatial auxiliary datasets available from free, public sources into 100x100m raster cells across all of Namibia, then calculated the average value within a 2km buffer from each cell (2km because the DHS randomly geo-displaces urban cluster coordinates by up to 2km) (Table S1.1).

---

[1] Thomson DR, Kools L, Jochem WC. 2018. Linking Synthetic Populations to Household Geolocations: A Demonstration in Namibia. *Data* 3(3), 30; DOI:10.3390/data3030030.

Table S1.1. Data sources for simulated population in Khomas, Namibia

Short name	Long name	Source, original unit	Output unit
Population			
dhs_hh	Individual recode file summarized by household	2013 Demographic and Health Survey [A]	region
dhs_geo	Geo-displaced cluster coordinates	2013 Demographic and Health Survey [A]	coordinate (cluster)
census_housing, census_person	20% microdata census sample	2011 Namibia Statistics Agency [B]	constituency
census_report	Final census report	2011 Namibia Statistics Agency [C]	constituency
Used to generate new spatial data			
Imagery_2014	High resolution satellite imagery	2014-2016 Maxar (DigitalGlobe) Quickbird imagery, 30cm [D]	Coordinate (2016 household)
Imagery_2004	High resolution satellite imagery	2004-2013 Maxar (DigitialGlobe) SPOT imagery, 40cm [D]	Coordinate (2001, 2006, 2011 household)
census_ea	2011 Census EA & constituency boundaries	2011 Namibia Statistics Agency [E]	EA, constituency
Auxiliary data			
ccilc_dst011_2012	Dist to land-cover: Cultivated terrestrial lands	2008-2012 GlobCover, 300m [F]	100m
ccilc_dst040_2012	Dist to land-cover: Woody / Trees	2008-2012 GlobCover, 300m [F]	100m
ccilc_dst130_2012	Dist to land-cover: Shrubs	2008-2012 GlobCover, 300m [F]	100m
ccilc_dst140_2012	Dist to land-cover: Herbaceous	2008-2012 GlobCover, 300m [F]	100m
ccilc_dst150_2012	Dist to land-cover: Other vegetation	2008-2012 GlobCover, 300m [F]	100m
ccilc_dst190_2012	Dist to land-cover: Urban	2008-2012 GlobCover, 300m [F]	100m
ccilc_dst200_2012	Dist to land-cover: Bare	2008-2012 GlobCover, 300m [F]	100m
cciwat_dst	Dist to water bodies	2000 OSM [G]	100m
dmsp_2011	Night-time lights intensity	2012 Suomi VIIRS, 500m [H]	100m
gpw4coast_dst	Dist to coastline	GPWv4, 1km [I]	100m
osmint_dst	Dist to road intersections	2000 OSM [G]	100m
osmriv_dst	Dist to major water ways	2000 OSM [G]	100m
slope	Slope	2000 HydroSHEDS, 100m [J]	100m
topo	Elevation	2000 HydroSHEDS, 100m [J]	100m
tt50k_2000	Travel time to populated places	2000 JRC-EC [K]	100m
urbpx_prp_1_2012	Proportion of urban pixels within 1 cell radius	2009 Modis [L,M]; Global Human Settlement City Model, 1km [N]	100m
hfacilities_dst	Dist to health centre or hospital	2001 UN-OCHA [O]	100m
schools_dst	Dist to primary/secondary school	2001 UN-OCHA [P]	100m
npp_2012	Annual net primary productivity	2010 MODIS, 1km [Q]	100m

A. ICF International. 2020. Available datasets. https://dhsprogram.com/data/available-datasets.cfm
B. NSA. 2013. Namibia 2011 Population and Housing Census version 1.0. https://nsa.org.na/microdata1/index.php/catalog/19
C. NSA. 2011. Namibia Population and Housing Census 2011 main report. http://www.nsa.org.na/files/downloads/Namibia 2011 Population and Housing Census Main Report.pdf
D. Maxar. 2019. Satellite Imagery. www.digitalglobe.com/products/satellite-imagery
E. NSA. 2011. 2011 Census EA boundaries. https://nsa.org.na/page/gis-data-requests/
F. European Space Agency. 2012. GlobCover. www.esa-landcover-cci.org/?q=node/158
G. OpenStreetMap contributors. 2000. OpenStreetMap base data. www.openstreetmap.org
H. NOAA. 2012. VIIRS nighttime lights. https://maps.ngdc.noaa.gov/viewers/VIIRS_DNB_nighttime_imagery/index.html
I. CIESIN. 2018. Gridded Population of the World, Version 4.11 (GPWv4.11). DOI:10.7927/H4F47M65
J. Lehner B, Verdin K, Jarvis A. 2006. HydroSHEDS technical documentation. www.worldwildlife.org/freshwater/pubs/HydroSHEDS_TechDoc_v10.pdf
K. Nelson A. 2008. Travel time to major cities: A global map of accessibility. https://forobs.jrc.ec.europa.eu/products/gam/
L. Schneider A, Friedl MA, Potere D. 2009. A new map of global urban extent from MODIS satellite data. Environ Res Lett;4:1–11. DOI: 10.2307/2346830.
M. Schneider A, Friedl MA, Potere D. 2010. Mapping global urban areas using MODIS 500-m data: New methods and datasets based on "urban ecoregions." Remote Sens Environ;114:1733–46. DOI:10.1016/j.rse.2010.03.003.
N. European Commission. 2017. Global human settlement city model (GHS-SMOD). http://ghsl.jrc.ec.europa.eu/faq.php
O. UN-OCHA-ROSA. 2001. Namibia health facilities. HDX. https://data.humdata.org/organization/ocha-rosa
P. UN-OCHA-ROSA. 2001. Namibia education facilities. HDX. https://data.humdata.org/organization/ocha-rosa
Q. Steven W. R, Ramakrishna R. N, Faith Ann H, et al. 2004. A continuous satellite-derived measure of global terrestrial primary production. Bioscience;54(6):547–60. DOI:10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2

(3) In step 3, we calculated the main type of household in each 2013 DHS primary sampling unit (PSU) (550 nationally) based on k-means groups defined in Khomas (step 1), and joined the 2km averaged auxiliary data values (step 2) to each PSU point. The distribution of PSU main household type across Namibia was: 185 (34%) urban non-poor, 82 (15%) urban poor, 249 (45%) rural poor, and 34 (6%) rural non-poor. We used these 550 PSU household types as training data, and the average 2km covariate values in a Random Forest machine classification model to predict a probability surface for each household type in each 100x100m cell in Namibia. This model performed well for urban non-poor households (14.6% misclassification) and rural poor households (7.6% misclassification), though classification error was high in areas comprised of mostly urban poor households (58.5% misclassification) and rural non-poor households (76.5% misclassification) (Table S1.2). Errors within urban areas were expected because auxiliary data 2km buffers can mask disparities between neighbourhoods. Although expected, poor performance of the model for urban poor households was problematic and addressed in the next step. Misclassification of rural non-poor households was also not surprising given the small size of this population, though this problem was ignored because non-poor rural households comprised a very small portion of the population in Khomas (<1%).

Table S1.2. Random Forest confusion matrix for average household type in 550 DHS clusters in the Khomas, Namibia simulation

	Type 1 – Urban non-poor	**Type 2** – Urban poor	**Type 3** – Rural poor	**Type 4** – Rural non-poor	**Classification Error**
Type 1 – Urban non-poor	158	23	3	1	0.146
Type 2 – Urban poor	40	34	7	1	0.585
Type 3 – Rural poor	8	3	230	8	0.076
Type 4 – Rural non-poor	4	0	22	8	0.765

(4) To improve the accuracy of the urban household probability layers in Khomas, we created an urban poor/non-poor weights layer by manually assigning each census EA with the portion of population that appeared to be located in a slum or informal settlement in 2016 based on visual inspection of 30cm Quickbird satellite imagery. Before beginning this process, we split large EAs at the periphery of Windhoek to create new EAs for areas that had undergone urban expansion since the 2011 census boundaries were drawn (total of 922 EAs). Rural EAs had a null probability in this step. The poor/non-poor weights layers were multiplied by the predicted household probability surfaces (step 3) to produce final 100x100m household probability surfaces (Figure S1.2).

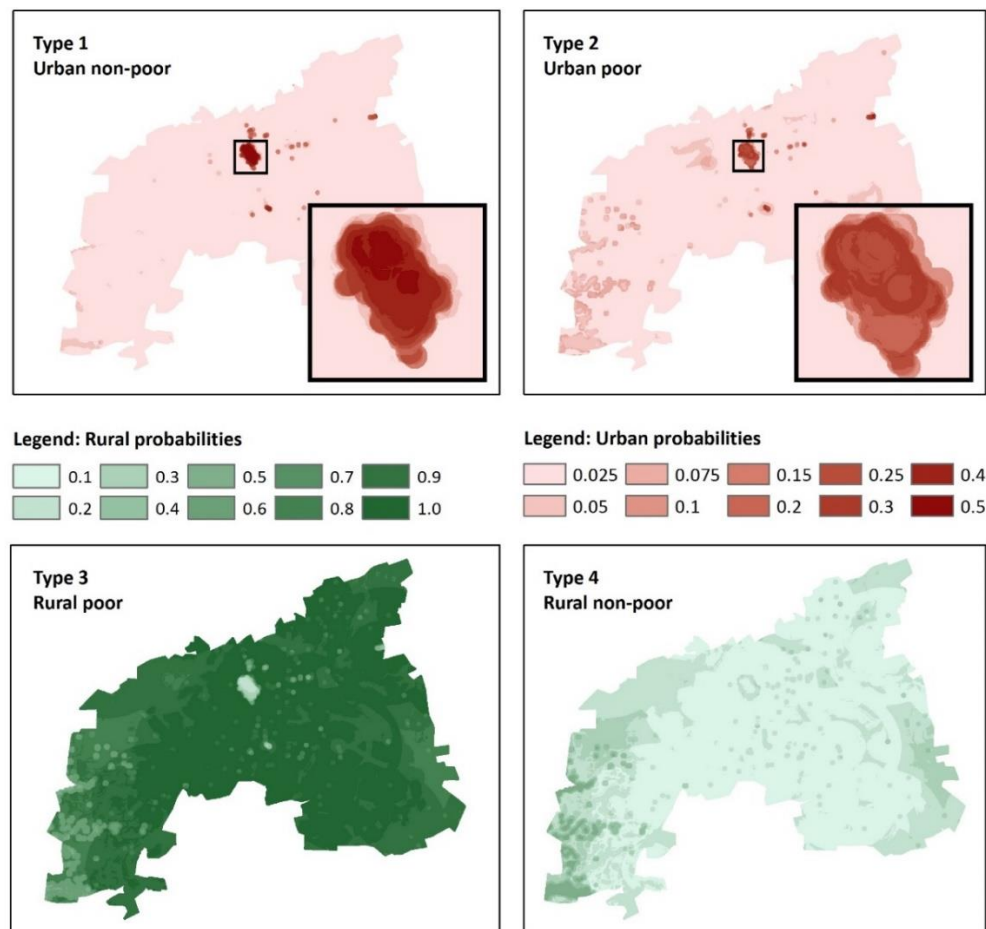



Figure S1.2. Household type probability surfaces (steps 1-4)
in Khomas, Namibia population simulation

(5) In step 5, we manually digitized building locations across Khomas using 2014-2016 high-resolution (30cm) Quickbird imagery in ArcGIS 10. Subjective judgement was required; for example, deciding not to digitize some buildings on main streets in densely populated areas where shops and offices seemed likely. In areas of dense settlement, some points were duplicated to represent more than one household in the same building. A total of 97,667 household points were digitized in 2016. As a benchmark, we exported points to Google Earth and used 2011 Maxar and SPOT (40cm) imagery to identify buildings that were missing in 2011, and ensured that the reduced number of points matched constituency household counts in the 2011 census (Table S1.1, C).

(6) In step 6, we simulated a population of realistic households in Khomas using iterative proportional fitting (IPF) with combinatorial optimisation in the R *simPop* package [2] (Table S1.3). IPF starts by defining a basic household structure to ensure the synthetic population is realistic. We defined household structure with household size, urban/rural residence, and age and sex of household head at the household-level; and age, sex, and relationship (to head) at the individual-level. Inputs to the model were the 2011 Census 20% microdata sample, as well as urban and rural household sizes, and constituency population by age, sex, and relationship based on the 2011 census report (Table S1.1, C). The IPF model selects random samples of records from the microdata with replacement until each of the household structure targets per constituency are met.

---

[2] Templ M, Meindl B, Kowarik A, et al. 2017. Simulation of synthetic complex data: The R package simPop. J Stat Softw;79(10):1–38. www.jstatsoft.org/v79/i10/

Table S1.3. Iterative proportional fitting of household structure
in Khomas, Namibia simulation by constituency

	Tobias Hainyeko	Katutura Central	Katutura East	Khomasdal North	Soweto	Samora Machel	Windhoek East	Windhoek Rural	Windhoek West	Moses Garoëb
N	60553	30868	24078	60465	19570	80036	27309	30028	62588	62807
HH Size										
Average	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
Residence										
Urban	100%	100%	100%	100%	100%	100%	100%	26%	100%	100%
Rural	0%	0%	0%	0%	0%	0%	0%	74%	0%	0%
Relationship										
Head	27%	21%	20%	24%	22%	26%	34%	30%	28%	30%
Spouse	10%	6%	5%	9%	6%	8%	18%	13%	13%	9%
Child	26%	27%	27%	31%	25%	27%	28%	28%	29%	23%
Grandchild	4%	8%	12%	4%	10%	6%	1%	7%	2%	5%
Extended	29%	31%	29%	26%	31%	28%	12%	14%	20%	29%
Other	5%	8%	7%	6%	5%	5%	8%	7%	8%	5%
Sex										
Female	45%	55%	56%	53%	53%	52%	51%	46%	53%	47%
Male	55%	45%	44%	47%	47%	48%	49%	54%	47%	53%
Age										
<1	4%	2%	3%	3%	2%	3%	2%	3%	2%	4%
1 - 4	9%	8%	9%	8%	7%	9%	7%	9%	7%	9%
5 - 9	9%	10%	10%	9%	9%	8%	6%	10%	7%	8%
10 - 14	8%	10%	10%	10%	9%	9%	6%	10%	8%	6%
15 - 19	8%	11%	11%	11%	11%	10%	8%	9%	11%	7%
20 - 24	15%	12%	13%	14%	17%	15%	8%	9%	15%	14%
25 - 29	14%	12%	10%	10%	12%	14%	9%	8%	10%	15%
30 - 34	11%	10%	8%	9%	9%	11%	9%	7%	9%	13%
35 - 39	9%	7%	7%	8%	6%	7%	9%	7%	7%	11%
40 - 44	6%	5%	5%	6%	4%	5%	9%	7%	6%	6%
45 - 49	4%	4%	4%	5%	3%	4%	6%	5%	5%	4%
50 - 54	2%	3%	3%	3%	4%	2%	6%	5%	4%	2%
55 - 59	1%	2%	2%	2%	3%	2%	5%	3%	3%	1%
60 - 64	1%	1%	2%	1%	1%	1%	3%	3%	2%	1%
65 - 74	0%	1%	2%	1%	1%	1%	5%	4%	2%	0%
75+	0%	1%	1%	1%	0%	0%	2%	2%	1%	0%

Next, using the R *simPop* package, we added household and individual characteristics present in the 20% microdata census dataset (toilet, water, structure, space, fuel, education) to the simulated dataset using a multinomial logistic regression technique and conditional annealing (Table S1.4**Error! Reference source not found.**). This treated age, sex, relationship, household size, and urban/rural residence as predictors, and each of the household characteristic as a conditional outcome.

Table S1.4. Multinomial logistic regression output of household characteristics
in Khomas, Namibia simulation by constituency

	Tobias Hainyeko	Katutura Central	Katutura East	Khomasdal North	Soweto	Samora Machel	Windhoek East	Windhoek Rural	Windhoek West	Moses Garoëb
N (individuals)	60553	30868	24078	60465	19570	80036	27309	30028	62588	62807
Water										
Improved	100%	100%	100%	100%	100%	100%	100%	96%	100%	100%
Unimproved	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%
Toilet										
Improved	25%	58%	67%	76%	69%	44%	97%	52%	94%	24%
Unimproved	75%	42%	33%	24%	31%	56%	3%	48%	6%	76%
Floor										
Durable	44%	97%	99%	88%	96%	72%	96%	80%	98%	44%
Non-durable	56%	3%	1%	12%	4%	28%	4%	20%	2%	56%
Space										
Adequate	81%	64%	64%	78%	74%	74%	96%	75%	93%	81%
Inadequate	19%	36%	36%	22%	26%	26%	4%	25%	7%	19%
Fuel										
Non-solid	87%	99%	97%	93%	99%	94%	100%	50%	100%	92%
Solid	13%	1%	3%	7%	1%	6%	0%	50%	0%	8%
HH Head Education										
No formal	24%	20%	21%	18%	16%	21%	14%	30%	14%	24%
Some primary	22%	20%	19%	19%	17%	18%	10%	24%	12%	20%
Primary	37%	38%	35%	32%	32%	36%	14%	28%	18%	38%
Secondary	15%	19%	20%	22%	26%	21%	33%	12%	32%	18%
Tertiary	2%	3%	5%	9%	8%	4%	29%	6%	24%	1%

We confirmed that there were not major differences between the distributions of characteristics in the 20% microdata and simulated dataset (all differences were less than +/- 0.002). Confident that the simulated household and individual characteristics were realistic, we calculated the most likely household type for each household based on variable factor weights created in the k-means analysis in step 1.

The 2011 census microdata sample was provided with a weight of approximately five for each observation to scale the 20% microdata sample to the total population in 2011. We calibrated the simulation to create an extra 20% of households to ensure there were enough simulated households to assign to 2016 point locations; left over simulated households were discarded in step 7. This resulted in 122,079 simulated households in Khomas before assignment to point locations.

(7) In step 7, we joined the re-weighted household type probabilities created in step 4 to the household latitude-longitude coordinates created in step 5. For each latitude-longitude coordinate created for 2016 household point locations, we randomly sampled a simulated household created in step 6 from the corresponding constituency and urban/rural strata based on the probabilities of household types at each coordinate. We repeated assignment of simulated households to coordinate point locations until all coordinates were assigned a simulated household, and then discarded the extra unassigned simulated households for a total of 97,667 simulated households located at realistic coordinate locations in Khomas for 2016.

(8) In step 8, we used the 2013 DHS records in Khomas (n=931 households) to develop multinomial models in R to simulate the same three individual and household outcomes as Thomson and colleagues (2018): household wealth quintile (five ordinal categories), woman's use of modern contraception (binary in women age 15 to 49), and child's receipt of 3rd DPT vaccination (binary in children under five) (Table S1.5). We used a multinomial model to calculate associations between each outcome and household-level covariates in the 2013 DHS dataset, and applied coefficients to the simulated dataset to predict wealth quintile, modern contraceptive use, and receipt of 3rd DTP vaccine for each household, woman 15 to 49, and child under five, respectively.

Table S1.5. Multinomial model coefficients and fit statistics for three outcomes in the 2013 DHS for Khomas, Namibia

Predictor	Household wealth quintile (ref=poorest)				Women 15-49 use of modern contraception	Child <5 DPT3 vaccination coverage
	poorer	middle	richer	richest		
Rural	0.479	0.773*	2.299***	2.061***	-0.227**	2.334***
HH Head						
15-29	(ref.)	(ref.)	(ref.)	(ref.)		
30-49	-11.595***	-11.222***	-11.581***	-10.890***		
50+	-9.957***	-9.171***	-8.901***	-7.715***		
HH Head Female	1.003***	0.778**	0.929**	0.333		
Age						
15 – 19					-1.290***	
20 – 24					-0.111**	
25 - 29					0.208***	
30 – 34					(ref.)	
35 – 39					0.030	
40 - 44					0.123**	
45 - 49					-0.023	
Child age 1 – 4						0.795***
Female						-0.188***
HH Head						
No education	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)
Some primary	0.133	-0.133	0.121	0.166	0.562***	0.680***
Primary	1.459***	2.243***	2.401***	3.216***	-0.038	0.447***
Secondary	0.466	1.651***	2.675***	4.092***	0.023	0.258
Tertiary	4.844***	6.455***	7.491***	9.515***	-0.259***	0.667***
Water Unimproved	-1.262*	0.429	-106.655	-0.169	-0.023	11.129
Toilet Unimproved	-23.935***	-26.157***	-28.908***	-30.603***	-0.018	0.021
Space Inadequate	-0.771**	-1.652***	-0.292	-1.216***	0.028	0.293***
Floor Non-durable	-21.756***	-22.962***	-24.338***	-26.003***	0.297***	0.748***
Fuel Solid	-19.316***	-20.937***	-23.301***	-105.303***	-0.197**	-0.621***
Constant	77.205***	80.003***	82.729***	82.498***	0.446***	-0.250
AIC	30,400				27,470	6,344
Note: *p<0.1; **p<0.05; ***p<0.01						

(9) To check the realism of this dataset, we compared the distribution of simulated household and individual outcomes (summarised by census enumeration areas - EAs) to households and individuals measured in the 2013 DHS (summarised by primary sampling units – PSUs) in Figure S1.3. The distribution of household characteristics appeared to be consistent between the simulated and DHS populations. However, individual characteristics were less consistent, and more heaped around the mean in the simulated dataset (Figure S1.3). This may have occurred because there were more observations per unit (EA vs PSU) in the simulated dataset, and more census units (922 EAs) compared to the 2013 DHS dataset (53 PSUs). Due to these inconsistencies, we only report household-level outcomes in the simulated dataset.

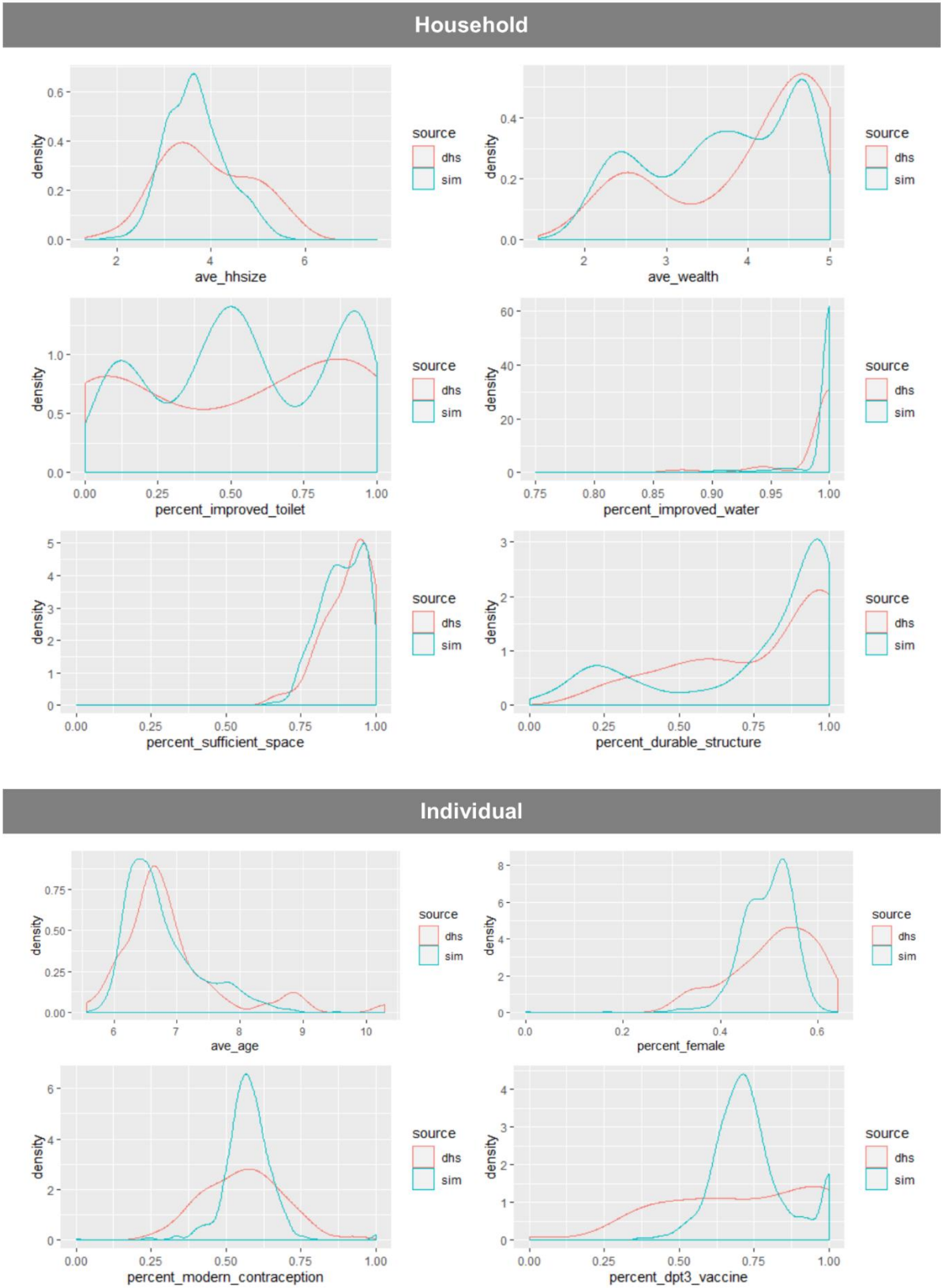



Figure S1.3. Comparison of household and individual outcomes by 2013 Namibia DHS cluster (n=53) and simulated population EA (n=922) in Khomas, Namibia

Supplement 2. Simulated Population in Khomas, Namibia [csv]

Supplement 3. Literature Review Results

Supplement to Thomson DR, Leasure DR, Bird T, Tzavidis N, Tatem AJ. 2021. How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia.

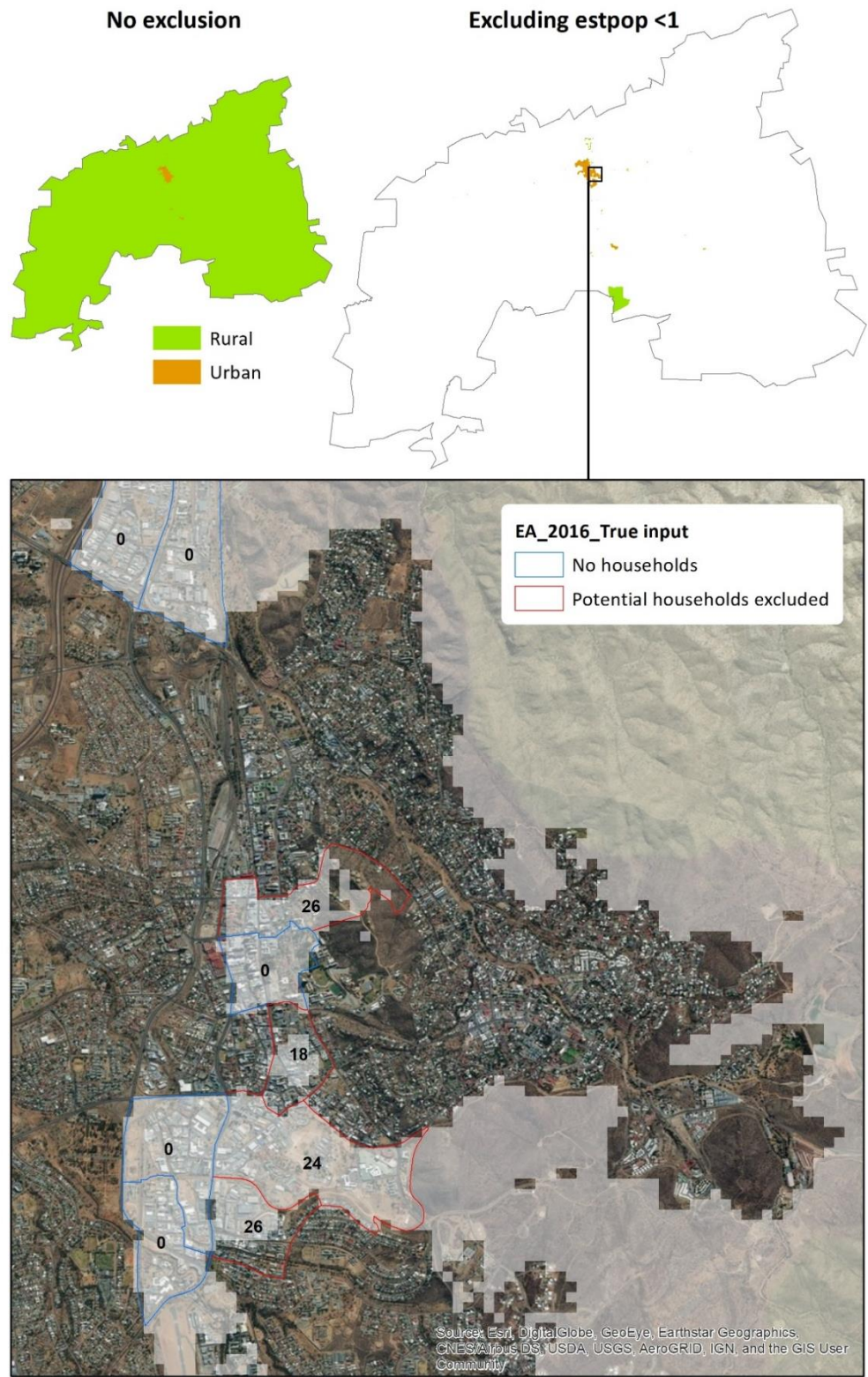
Table S3.1. Percent of population missing from LMIC censuses by source

Location & Housing Type	Agarwal (2011) India	Carr-Hill (2013) LMICs	Carr-Hill (2017) Tanzania, Kenya, Uganda	Ebenstein (2015) China	Gidado (2013) Nigeria	Gurgel (2003) Brazil (kids only)	Jiang (2015) China	Karanja (2010) Kenya	Kronefeld (2008) Afghan-istan	Lucci (2018) Kenya	Sabry (2010) Egypt
Urban slum											
Permanent/semi-permanent (not mobile)	50%	5-13%	17-51%					21%		18, 21, 38, 59%	45%
Permanent/semi-permanent (mobile)				33-61%					44%		
Homeless		100%		33-61%		64%					
Urban non-slum											
Permanent							2%				
Rural											
Permanent/semi-permanent							2%				
Remote					12%						
Nomad					12%						
Institutional											
Hospital / care home		<1%									
Prison		0.09									
Refugee camp		10-15%									
Citation	[55]	[7]	[59]	[60]	[61]	[62]	[63]	[64]	[65]	[66]	[56]

Location & Housing Type	Stark (2017) Cambodia (kids only)	Treiman (2005) China	PES (2000) Brazil	PES (2001) Nepal	PES (2002) Tanzania	PES (2002) Uganda	PES (2010) Ghana	PES (2010) Zambia	PES (2010) Rwanda	PES (2011) Bangladesh	PES (2011) India	PES (2011) South Africa
Urban slum												
Permanent/semi-permanent (not mobile)												
Permanent/semi-permanent (mobile)		50%										
Homeless	80-96%	50%										
Urban non-slum												
Permanent			2-4%	12%	7%	12%	3%	5%	2%	5%	3%	15%
Rural												
Permanent/semi-permanent			4-11%	5%	7%	5%	2%	10%	2%	4%	2%	13%
Remote												
Nomad												
Institutional												
Hospital / care home												
Prison												
Refugee camp												
Citation	[57]	[58]	[45]	[46]	[47]	[48]	[49]	[50]	[54]	[51]	[52]	[53]

Supplement 4. Excluding Cells for Accuracy Analysis

Supplement to Thomson DR, Leasure DR, Bird T, Tzavidis N, Tatem AJ. 2021. How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia.

Figure S4.1. Exclusion of grid cells with less than one estimated person in accuracy analysis

Supplement 5. Root Mean Square Error (RMSE) Statistics for all scenarios

Supplement to Thomson DR, Leasure DR, Bird T, Tzavidis N, Tatem AJ. 2021. How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia.

Population Density RMSE per hectare, EA-level input data, excluding cells with estimated population <1 (adjusts for area)																
Cell size	2016_True	2016_L	2016_M	2016_H	2011_True	2011_L	2011_M	2011_H	2006_True	2006_L	2006_M	2006_H	2001_True	2001_L	2001_M	2001_H
(metres) **All**																
100	27.1	29.3	25.8	29.3	28.0	29.7	33.1	30.2	31.1	30.6	28.6	30.8	30.6	29.8	29.5	31.2
200	24.7	25.3	25.8	27.2	25.5	26.0	26.9	27.7	25.6	25.9	26.3	27.4	26.2	26.4	26.9	27.7
300	26.5	26.7	27.1	27.8	27.0	27.2	27.6	28.3	26.7	26.9	27.2	27.8	27.2	27.3	27.7	28.2
400	26.8	27.0	27.2	27.7	27.3	27.4	27.6	28.1	27.1	27.3	27.4	28.0	27.5	27.6	27.8	28.2
500	26.4	26.6	26.8	27.2	26.9	27.0	27.1	27.5	26.7	26.8	27.0	27.3	26.9	27.0	27.2	27.4
600	26.0	26.1	26.2	26.6	26.4	26.6	26.6	26.9	26.3	26.4	26.5	26.9	26.6	26.6	26.9	27.1
700	25.7	25.7	25.9	26.2	26.2	26.2	26.3	26.6	26.0	26.0	26.2	26.4	26.5	26.6	26.6	26.7
800	25.3	25.3	25.3	25.7	25.6	25.7	25.8	26.0	26.0	26.2	26.3	26.4	26.3	26.3	26.4	26.6
900	24.3	24.3	24.5	24.8	24.7	24.7	24.8	25.1	24.8	24.9	25.0	25.1	25.2	25.2	25.3	25.5
1000	24.0	24.0	24.3	24.6	24.4	24.4	24.5	24.7	24.6	24.7	24.7	24.8	25.1	25.1	25.1	25.3
Rural																
100	29.2	32.9	26.7	29.9	31.0	33.7	38.3	30.5	34.9	34.0	29.9	31.0	33.2	31.7	30.2	30.6
200	7.4	8.3	6.9	7.6	7.9	8.5	9.6	7.8	8.8	8.5	7.6	7.9	8.3	7.9	7.6	7.7
300	4.0	4.3	3.8	4.1	4.2	4.4	4.8	4.2	4.4	4.3	3.9	4.1	4.2	4.0	3.9	4.0
400	3.2	3.3	3.1	3.2	3.3	3.4	3.5	3.3	3.3	3.3	3.2	3.2	3.2	3.1	3.1	3.1
500	2.8	2.9	2.8	2.9	2.8	2.9	2.9	2.9	2.5	2.5	2.5	2.5	2.3	2.3	2.2	2.2
600	4.2	4.3	4.3	4.3	4.2	4.3	4.3	4.3	3.8	3.8	3.8	3.9	3.6	3.6	3.6	3.7
700	3.6	3.7	3.7	3.7	3.6	3.6	3.6	3.6	3.3	3.3	3.3	3.4	3.0	3.0	3.0	3.0
800	4.3	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.1	4.1	4.1	4.2	4.1	4.1	4.1	4.1
900	5.2	5.2	5.2	5.3	5.2	5.2	5.3	5.3	4.5	4.5	4.5	4.6	3.8	3.8	3.8	3.8
1000	5.3	5.3	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4
Urban																
100	24.4	24.3	24.7	28.5	24.0	24.2	25.4	29.8	25.1	25.4	26.7	30.5	26.4	26.9	28.6	32.2
200	35.6	36.3	37.4	39.5	36.9	37.5	38.6	40.5	37.7	38.2	39.2	41.1	39.4	39.8	40.8	42.2
300	39.3	39.6	40.3	41.5	40.3	40.5	41.1	42.3	40.7	41.0	41.5	42.7	41.9	42.1	42.9	43.8
400	40.3	40.5	40.8	41.6	41.1	41.2	41.5	42.5	41.6	41.8	42.2	43.1	42.8	42.9	43.4	44.1
500	39.7	39.9	40.2	40.9	40.4	40.5	40.8	41.4	40.8	40.9	41.4	41.9	41.6	41.7	42.4	42.6
600	39.5	39.7	39.8	40.5	40.0	40.3	40.4	41.0	40.5	40.7	40.9	41.6	41.5	41.5	42.0	42.4
700	39.7	39.8	39.8	40.5	40.4	40.5	40.5	40.9	40.5	40.7	40.8	41.2	41.6	41.8	42.0	42.1
800	39.3	39.3	39.4	40.0	39.7	39.7	40.0	40.3	40.5	40.9	41.2	41.5	41.5	41.5	42.0	42.5
900	37.6	37.6	37.8	38.4	38.1	38.2	38.2	38.8	38.7	38.9	39.2	39.5	39.8	39.9	40.2	40.7
1000	37.2	37.2	37.4	37.7	37.6	37.7	37.7	38.2	38.1	38.3	38.4	38.6	39.3	39.3	39.3	39.7

Cell size (metres)	2016_True	2016_L	2016_M	2016_H	2011_True	2011_L	2011_M	2011_H	2006_True	2006_L	2006_M	2006_H	2001_True	2001_L	2001_M	2001_H
Population Density RMSE per hectare, Constituency-level input data, excluding cells with estimated population <1 (adjusts for area)																
All																
100	24.5	24.6	25.3	26.5	24.7	25.2	25.8	27.1	26.4	26.7	27.1	28.9	28.4	28.5	29.2	30.5
200	27.7	28.0	28.4	29.2	28.0	28.4	28.8	29.7	29.1	29.4	29.6	30.9	30.8	30.7	31.2	32.3
300	28.1	28.3	28.6	29.3	28.3	28.6	29.0	29.7	29.2	29.4	29.5	30.6	30.6	30.4	30.8	31.8
400	28.1	28.3	28.5	29.0	28.3	28.6	28.8	29.4	29.0	29.2	29.3	30.3	30.2	30.1	30.5	31.4
500	27.6	27.8	28.0	28.4	27.7	28.0	28.3	28.7	28.4	28.5	28.7	29.6	29.5	29.4	29.8	30.7
600	27.3	27.5	27.6	28.1	27.5	27.6	27.8	28.4	28.0	28.2	28.2	29.1	29.0	28.9	29.2	30.2
700	27.1	27.4	27.5	27.8	27.3	27.6	27.7	28.1	27.8	27.9	28.0	28.7	28.7	28.6	29.0	30.0
800	26.6	26.7	26.8	27.4	26.6	27.0	27.3	27.8	27.4	27.6	27.8	28.4	28.4	28.3	28.5	29.4
900	25.8	26.1	26.2	26.7	26.0	26.1	26.5	27.0	26.6	26.8	26.9	27.8	27.6	27.5	27.6	28.5
1000	25.8	26.2	26.0	26.5	25.9	26.1	26.4	26.8	26.5	26.6	26.7	27.6	27.4	27.4	27.7	28.5
Rural																
100	10.7	9.8	8.6	6.7	9.3	9.0	7.7	6.2	7.2	6.8	5.9	5.1	6.2	5.8	5.2	4.5
200	3.2	3.1	2.9	2.7	3.0	3.0	2.8	2.7	2.8	2.8	2.7	2.7	2.8	2.8	2.8	2.9
300	2.7	2.8	2.7	2.8	2.7	2.8	2.8	2.9	2.8	2.9	2.9	3.0	3.0	3.0	3.1	3.3
400	2.7	2.8	2.8	2.9	2.7	2.8	2.8	2.9	2.9	2.9	2.9	3.1	3.1	3.0	3.1	3.3
500	2.7	2.7	2.7	2.8	2.7	2.7	2.8	2.9	2.8	2.8	2.8	3.0	3.0	3.0	3.1	3.2
600	4.4	4.5	4.5	4.6	4.4	4.5	4.6	4.7	4.6	4.6	4.6	4.9	4.9	4.8	5.0	5.3
700	4.0	4.0	4.0	4.1	4.0	4.1	4.1	4.2	4.1	4.2	4.1	4.3	4.4	4.3	4.4	4.7
800	4.6	4.7	4.7	4.9	4.7	4.8	4.9	5.0	4.9	5.0	5.0	5.2	5.2	5.2	5.3	5.6
900	5.6	5.7	5.8	6.0	5.7	5.7	5.9	6.1	5.9	6.0	6.1	6.5	6.4	6.3	6.4	6.8
1000	5.9	6.0	6.0	6.2	5.9	6.0	6.2	6.4	6.2	6.2	6.3	6.7	6.6	6.6	6.8	7.2
Urban																
100	33.2	33.3	34.8	36.6	33.8	34.5	35.4	37.0	36.3	36.6	37.5	39.2	38.4	38.8	39.4	40.2
200	39.3	39.4	40.2	40.9	39.7	40.1	40.4	41.1	40.8	41.0	41.3	42.1	42.0	42.2	42.3	42.6
300	40.1	40.1	40.7	41.0	40.4	40.6	40.8	41.1	41.0	41.1	41.3	41.8	41.8	41.8	41.9	42.1
400	40.1	40.1	40.6	40.8	40.4	40.6	40.7	40.9	40.8	40.9	41.1	41.4	41.4	41.4	41.5	41.6
500	39.3	39.3	39.7	39.8	39.4	39.6	39.8	39.8	39.8	39.9	40.2	40.4	40.4	40.4	40.4	40.5
600	39.1	39.2	39.4	39.6	39.2	39.4	39.5	39.7	39.7	39.7	39.8	40.0	40.0	40.0	40.0	40.1
700	39.1	39.1	39.4	39.5	39.3	39.4	39.4	39.5	39.6	39.7	39.8	39.8	39.9	39.9	39.9	40.0
800	38.2	38.2	38.5	38.6	38.4	38.5	38.6	38.6	38.6	38.8	39.0	39.0	39.0	39.0	39.0	39.2
900	37.4	37.3	37.5	37.6	37.4	37.5	37.5	37.6	37.6	37.6	37.6	37.8	37.8	37.8	37.8	37.8
1000	37.0	37.0	37.0	37.1	37.0	37.0	37.1	37.1	37.1	37.2	37.4	37.4	37.4	37.4	37.4	37.4

Cell size (metres)	2016_True	2016_L	2016_M	2016_H	2011_True	2011_L	2011_M	2011_H	2006_True	2006_L	2006_M	2006_H	2001_True	2001_L	2001_M	2001_H
Population-Adjusted RMSE, EA-level input data, excluding cells with estimated population <1 (no adjustment for area)
All																
100	0.829	0.896	0.79	0.896	0.858	0.911	1.013	0.925	0.952	0.936	0.874	0.943	0.936	0.912	0.904	0.957
200	1.275	1.303	1.329	1.403	1.314	1.338	1.386	1.43	1.319	1.333	1.355	1.412	1.352	1.362	1.387	1.43
300	1.976	1.993	2.023	2.078	2.018	2.033	2.064	2.111	1.996	2.009	2.029	2.078	2.033	2.041	2.068	2.107
400	2.609	2.626	2.65	2.698	2.66	2.671	2.691	2.738	2.643	2.654	2.673	2.723	2.68	2.692	2.708	2.747
500	3.251	3.267	3.294	3.349	3.311	3.316	3.335	3.39	3.288	3.298	3.326	3.357	3.315	3.322	3.348	3.367
600	3.871	3.884	3.897	3.958	3.936	3.959	3.967	4.013	3.911	3.925	3.941	4.003	3.965	3.968	4.003	4.038
700	4.483	4.484	4.515	4.567	4.569	4.573	4.594	4.643	4.532	4.544	4.567	4.614	4.621	4.637	4.642	4.658
800	5.116	5.121	5.134	5.212	5.183	5.204	5.218	5.274	5.274	5.301	5.331	5.338	5.322	5.331	5.354	5.378
900	5.74	5.759	5.796	5.875	5.834	5.845	5.867	5.95	5.858	5.885	5.916	5.949	5.961	5.964	5.994	6.026
1000	6.305	6.319	6.396	6.468	6.409	6.422	6.452	6.504	6.474	6.487	6.495	6.513	6.602	6.603	6.607	6.647
Rural																
100	5.044	5.68	4.599	5.155	5.352	5.818	6.61	5.257	6.024	5.858	5.155	5.349	5.727	5.462	5.203	5.273
200	3.208	3.591	2.96	3.293	3.394	3.671	4.151	3.363	3.783	3.688	3.272	3.401	3.582	3.422	3.271	3.311
300	2.794	3.025	2.684	2.882	2.924	3.095	3.386	2.939	3.067	3	2.752	2.846	2.917	2.822	2.737	2.774
400	3.11	3.236	3.071	3.201	3.202	3.306	3.476	3.25	3.289	3.266	3.152	3.201	3.154	3.09	3.051	3.09
500	3.699	3.817	3.694	3.801	3.737	3.793	3.904	3.785	3.38	3.366	3.284	3.253	3.043	3.005	2.978	2.9
600	5.386	5.432	5.415	5.511	5.379	5.417	5.463	5.458	4.881	4.888	4.87	4.926	4.614	4.601	4.619	4.667
700	5.521	5.55	5.572	5.631	5.415	5.43	5.488	5.508	5.049	5.054	5.061	5.106	4.543	4.55	4.53	4.537
800	6.549	6.566	6.579	6.664	6.602	6.649	6.676	6.696	6.231	6.246	6.256	6.268	6.124	6.135	6.142	6.173
900	8.03	8.075	8.137	8.195	8.133	8.109	8.158	8.26	7.014	7.05	7.058	7.085	5.845	5.843	5.849	5.854
1000	10.22	10.255	10.429	10.554	10.42	10.453	10.556	10.594	10.536	10.539	10.544	10.558	10.432	10.435	10.441	10.485
Urban																
100	0.566	0.565	0.574	0.662	0.556	0.561	0.591	0.693	0.582	0.589	0.619	0.708	0.613	0.626	0.664	0.747
200	1.027	1.046	1.08	1.139	1.064	1.081	1.113	1.169	1.087	1.102	1.131	1.184	1.136	1.149	1.177	1.217
300	1.262	1.275	1.294	1.333	1.295	1.303	1.322	1.36	1.308	1.317	1.335	1.373	1.348	1.353	1.379	1.407
400	1.379	1.387	1.398	1.425	1.407	1.411	1.422	1.453	1.426	1.433	1.445	1.477	1.464	1.47	1.487	1.509
500	1.45	1.459	1.469	1.494	1.475	1.479	1.491	1.514	1.491	1.496	1.515	1.533	1.519	1.525	1.55	1.558
600	1.509	1.513	1.517	1.546	1.528	1.539	1.543	1.565	1.544	1.551	1.559	1.585	1.582	1.584	1.603	1.618
700	1.541	1.542	1.546	1.569	1.568	1.57	1.573	1.587	1.573	1.578	1.583	1.599	1.616	1.62	1.63	1.635
800	1.605	1.606	1.609	1.634	1.622	1.624	1.634	1.647	1.656	1.671	1.683	1.697	1.696	1.696	1.718	1.738
900	1.54	1.541	1.548	1.571	1.562	1.562	1.564	1.588	1.585	1.591	1.604	1.618	1.631	1.632	1.645	1.666
1000	1.586	1.586	1.594	1.609	1.605	1.606	1.609	1.631	1.625	1.633	1.639	1.648	1.676	1.677	1.678	1.694

Cell size (metres)	2016_True	2016_L	2016_M	2016_H	2011_True	2011_L	2011_M	2011_H	2006_True	2006_L	2006_M	2006_H	2001_True	2001_L	2001_M	2001_H
All																
100	0.752	0.754	0.776	0.813	0.757	0.773	0.789	0.829	0.809	0.817	0.83	0.886	0.871	0.873	0.894	0.934
200	1.427	1.441	1.463	1.506	1.442	1.462	1.484	1.531	1.502	1.516	1.525	1.593	1.586	1.583	1.608	1.664
300	2.099	2.116	2.137	2.185	2.114	2.137	2.165	2.216	2.182	2.197	2.205	2.283	2.282	2.269	2.304	2.377
400	2.734	2.759	2.774	2.826	2.756	2.782	2.808	2.862	2.826	2.844	2.851	2.95	2.94	2.928	2.966	3.056
500	3.391	3.421	3.441	3.495	3.413	3.445	3.478	3.537	3.491	3.513	3.528	3.645	3.628	3.62	3.669	3.779
600	4.065	4.102	4.109	4.182	4.091	4.11	4.147	4.235	4.168	4.195	4.199	4.335	4.323	4.3	4.354	4.506
700	4.737	4.778	4.796	4.853	4.775	4.817	4.834	4.905	4.859	4.874	4.882	5.013	5.012	4.998	5.066	5.233
800	5.384	5.406	5.435	5.559	5.388	5.466	5.527	5.623	5.543	5.6	5.623	5.76	5.751	5.73	5.777	5.965
900	6.097	6.172	6.194	6.323	6.155	6.184	6.263	6.385	6.293	6.353	6.366	6.572	6.526	6.506	6.54	6.749
1000	6.787	6.876	6.83	6.964	6.815	6.867	6.947	7.059	6.962	6.999	7.028	7.26	7.2	7.201	7.292	7.483
Rural																
100	1.839	1.698	1.484	1.153	1.608	1.549	1.329	1.073	1.248	1.168	1.013	0.878	1.077	0.994	0.901	0.77
200	1.402	1.353	1.265	1.176	1.309	1.295	1.224	1.177	1.211	1.194	1.15	1.18	1.218	1.19	1.191	1.234
300	1.913	1.932	1.922	1.975	1.915	1.934	1.963	2.027	1.989	2.005	2.001	2.12	2.118	2.088	2.149	2.29
400	2.682	2.734	2.727	2.817	2.708	2.744	2.788	2.887	2.821	2.849	2.853	3.038	3.006	2.981	3.061	3.266
500	3.527	3.59	3.601	3.71	3.566	3.612	3.675	3.802	3.708	3.742	3.756	3.999	3.956	3.938	4.051	4.303
600	5.597	5.702	5.692	5.872	5.664	5.697	5.795	6.014	5.83	5.899	5.91	6.274	6.228	6.149	6.32	6.802
700	5.995	6.112	6.113	6.26	6.074	6.164	6.213	6.385	6.257	6.284	6.283	6.567	6.613	6.578	6.693	7.163
800	7.015	7.154	7.114	7.424	7.078	7.19	7.34	7.598	7.381	7.495	7.517	7.895	7.861	7.802	7.938	8.485
900	8.684	8.913	8.95	9.294	8.843	8.92	9.144	9.477	9.206	9.382	9.421	10.017	9.869	9.805	9.913	10.61
1000	11.48	11.769	11.615	12.063	11.563	11.737	12.003	12.393	12.064	12.159	12.206	13.044	12.817	12.808	13.158	13.928
Urban																
100	0.771	0.774	0.808	0.849	0.786	0.801	0.821	0.859	0.843	0.85	0.87	0.91	0.892	0.901	0.915	0.933
200	1.135	1.137	1.158	1.179	1.145	1.155	1.166	1.186	1.177	1.182	1.192	1.214	1.211	1.216	1.221	1.229
300	1.29	1.291	1.307	1.318	1.299	1.305	1.311	1.322	1.317	1.32	1.329	1.344	1.343	1.345	1.348	1.353
400	1.374	1.373	1.389	1.398	1.384	1.389	1.393	1.4	1.397	1.4	1.407	1.418	1.417	1.418	1.42	1.424
500	1.436	1.437	1.45	1.454	1.441	1.448	1.453	1.455	1.453	1.458	1.468	1.476	1.475	1.476	1.477	1.48
600	1.494	1.494	1.503	1.512	1.497	1.502	1.506	1.515	1.513	1.516	1.517	1.526	1.525	1.528	1.527	1.53
700	1.517	1.515	1.528	1.53	1.524	1.528	1.529	1.534	1.535	1.539	1.546	1.544	1.547	1.547	1.548	1.552
800	1.563	1.563	1.572	1.578	1.572	1.572	1.576	1.578	1.577	1.585	1.594	1.595	1.595	1.595	1.596	1.601
900	1.532	1.528	1.534	1.539	1.533	1.533	1.534	1.54	1.539	1.539	1.54	1.546	1.546	1.546	1.546	1.547
1000	1.579	1.579	1.58	1.581	1.579	1.58	1.58	1.582	1.581	1.586	1.593	1.594	1.594	1.594	1.594	1.595

Population-Adjusted RMSE, Constituency-level input data, excluding cells with estimated population <1 (no adjustment of area)