 doi:10.20944/preprints202102.0492.v1

# How accurate are WorldPop-Global 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 affiliation: NorthWest Atlantic Fisheries Centre, Department of Fisheries and Oceans, St. John's, Canada

* Corresponding author, Email: dana.r.thomson@gmail.com

## 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 Censuses are often outdated and inaccurate in LMIC settings, and rarely disaggregated at fine geographic scale.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; however, none of these datasets have been evaluated for accuracy of population estimates at the grid cell-level. This is because the finest-scale population figures generally available to data producers are those input into gridded population models and disaggregated to smaller grid cells (e.g., 100x100m).

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 then 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's gridded population approach. We evaluate the cell-level accuracy of these simulated WorldPop-Global datasets, using the original "true" population as a reference.

In our simulation, we found large cell-level errors, particularly in urban cells, driven by WorldPop-Global's use of average population densities in large areal units to determine cell-level population densities. Age, accuracy, and aggregation of the input data did play a primary role in these errors. We suggest incorporating finer-scale training data into gridded population models generally, and WorldPop-Global in particular (e.g., from simulated populations, routine household surveys, or slum community profiles), and use of new building footprint datasets as a covariate to improve cell-level accuracy of gridded population data. It is important to measure cell-level accuracy of all gridded population datasets, especially if they are to be used for monitoring key development indicators.

## Key Words

LMIC, Global South, indicator, Random Forrest

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

# 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 from geo-statistical models using population counts and spatial datasets [13]. “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 [14,15], GHS-POP [16,17], HRSL [18]), while the most complex models use spatial covariates to inform spatial disaggregation of the areal unit population to grid cells (i.e., WorldPop-Global [19,20], LandScan [21,22], WPE [23]). 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 [21]. 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 [24]. “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 [25,26]. 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 [14], 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 [19]. 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 [27]. Among top-down gridded population datasets, the WorldPop-Global Random Forest model is among the best documented and most accurate gridded population models available [19,28] with publicly available model code [29] and pre-processed model covariates [30,31] enabling reproducibility and evaluation. WorldPop-Global and its preceding data products (AfriPop, AsiaPop, and AmeriPop) are unconstrained; however, a new WorldPop-Global-constrained dataset was recently published [32].

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, 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., [33]).

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 [34], and some censuses do not include certain refugee or internally displaced populations [35]. 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 [36].

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.

# Materials and 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 [37], 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 [38]. After independence, vast numbers of the population migrated to Windhoek, exaggerating rural-to-urban migration patterns observed globally during this time period [39,40]. Windhoek is also a destination for immigrants from neighbouring countries including financially unstable Zimbabwe [40,41]. The population of the Windhoek metropolitan area grew by a staggering 37% between the 2001 and 2011 censuses [37], with much of that growth in informal settlements [38].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

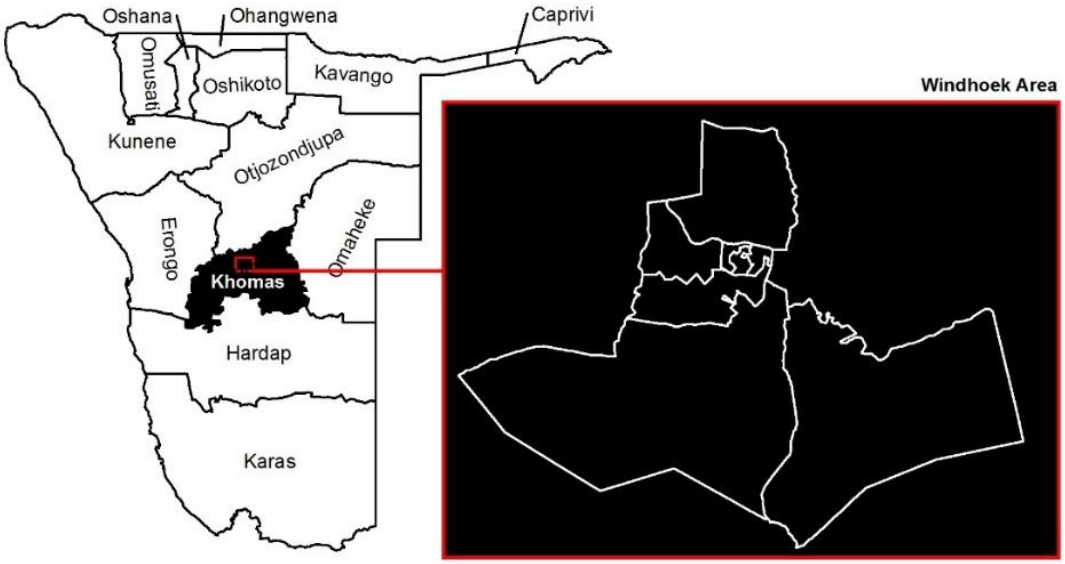



**Figure 1.** Location of Khomas region in Namibia, and of constituencies in Windhoek area

## 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

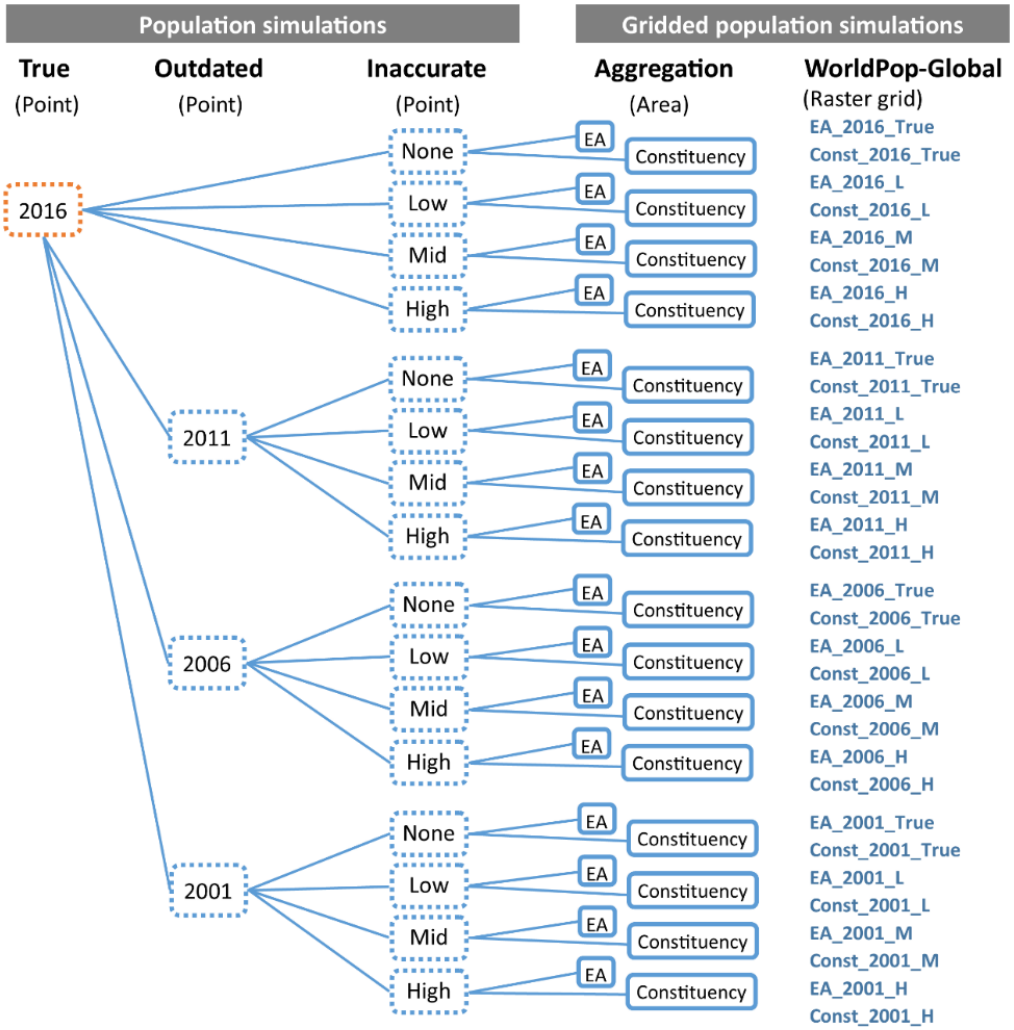



**Figure 2.** Summary of the population and gridded population simulation workflow

***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**

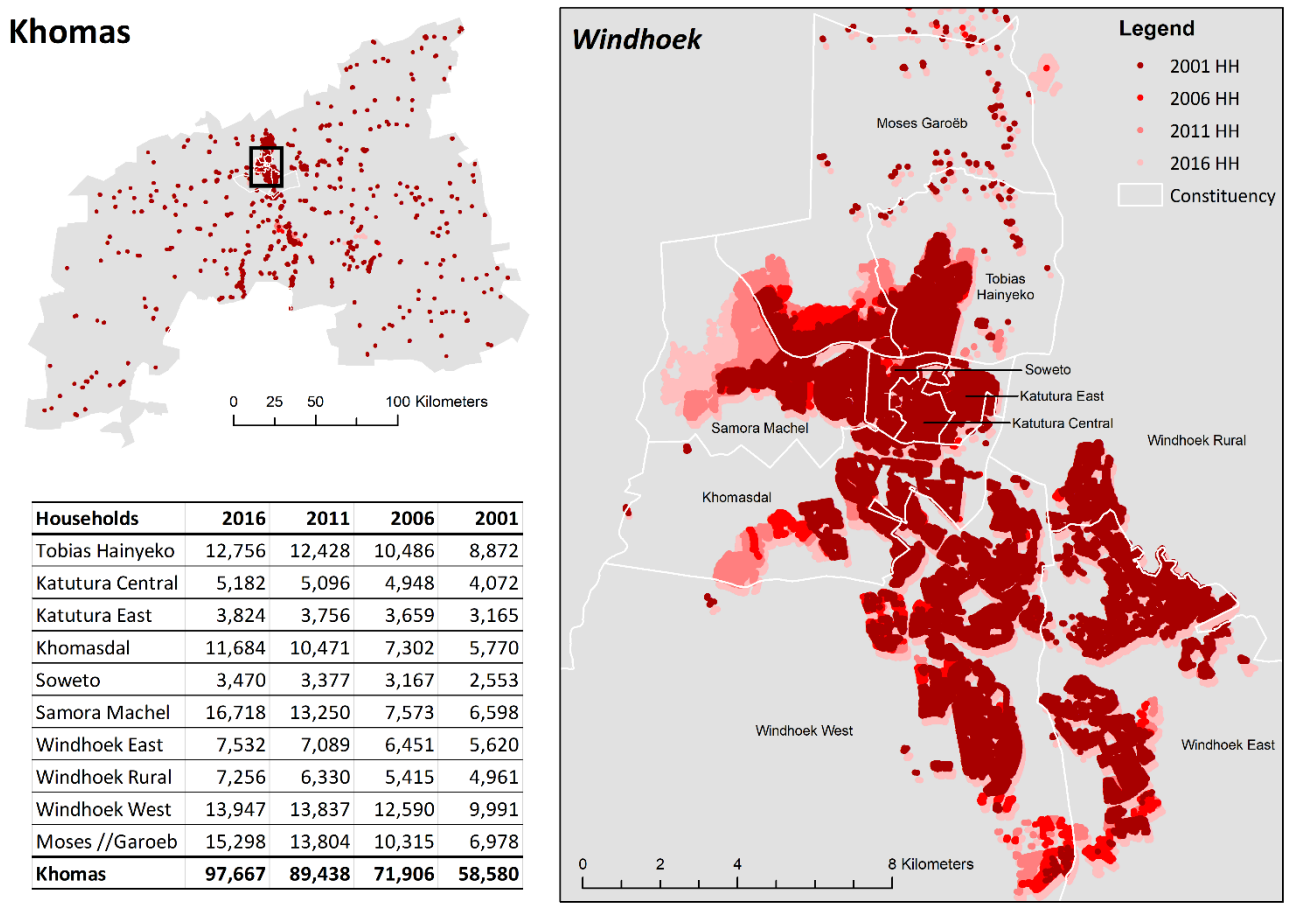
To simulate a realistic population in Khomas, Namibia, we used all of the same population inputs and spatial auxiliary datasets as Thomson and colleagues (2018) [36]. Broadly, this involved the creation of three datasets (modelled surfaces of household types, manually digitised building point locations, and synthetic (simulated) households), then assigned synthetic households to point locations based on the household type probability surfaces. Household types were defined from the 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 [36]. The probability surfaces of "urban poor" and "urban non-poor" household types were manually adjusted due to high misclassification. 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 imagery. Separately, we modelled a synthetic population of individuals nested within

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

households across Khomas from Namibia 2011 census microdata using an iterative proportional fitting model and conditional annealing [42]. A third set of data, building point locations, were manually digitised from 2014-2016 30cm Quickbird imagery in ArcGIS 10. To link 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) [36]. 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 [37] (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 [36].




| 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

**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 that eight additional articles reported census under-counts. Together, census under-counts in LMICs were collated from 10 PESs [43–52], and 13 articles [7,53–64] (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 1, see Supplement 3 for details).

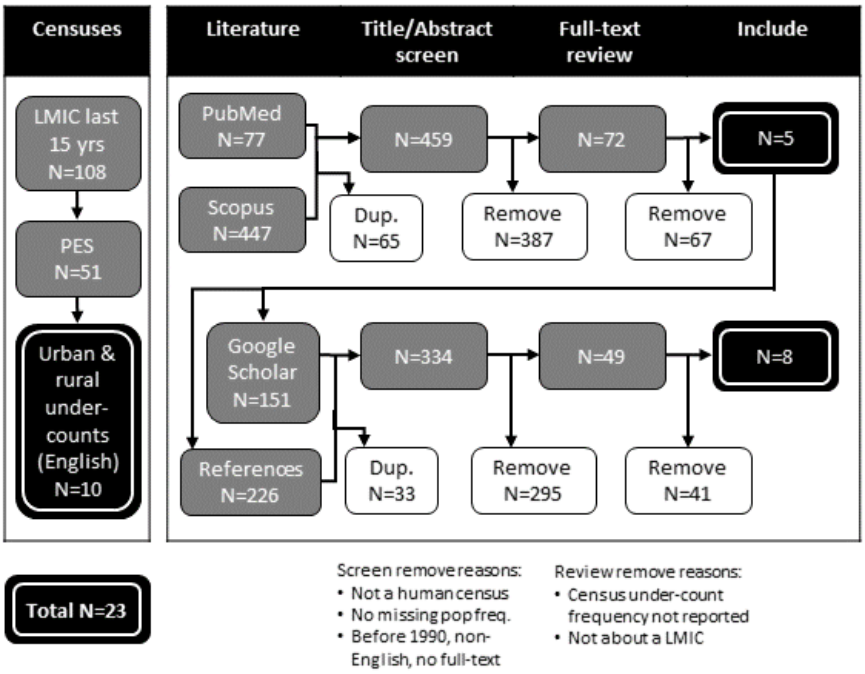



**Figure 4.** Search terms and process used in the census under-count literature review

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 [19] (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) [65]. 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 [19]. 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 [19].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

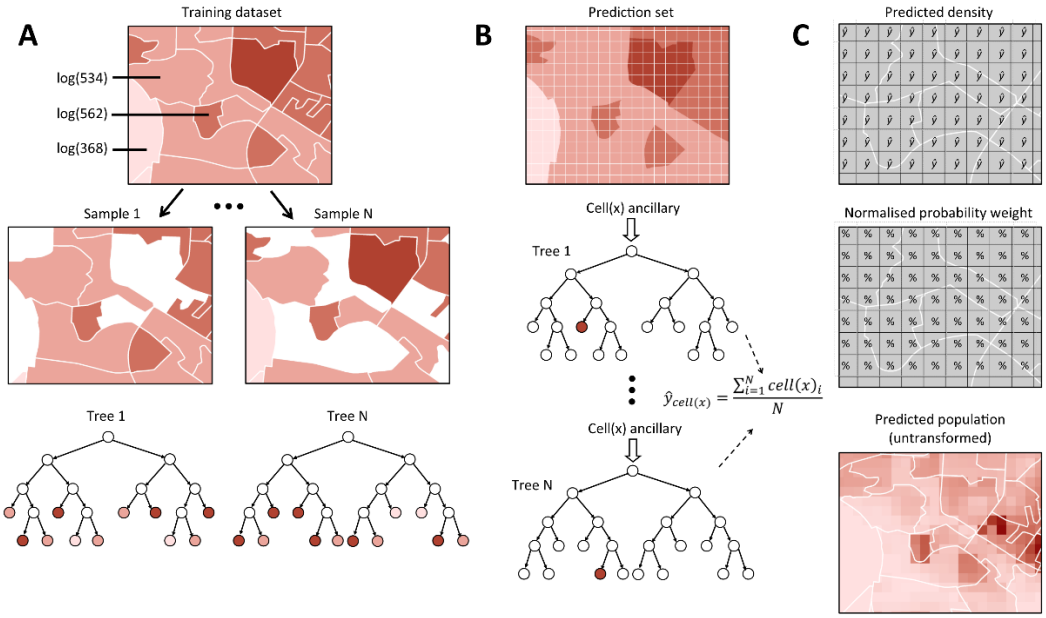



**Figure 5.** Overview of WorldPop-Global Random Forest Modelling Workflow

***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 [66] |
| cov_intsec | Distance to OSM major road intersections (2016) | Vector, <30 m | OpenStreetMap [66] |
| cov_waterw | Distance to OSM major waterways (2016) | Vector, <30 m | OpenStreetMap [66] |
| cov_wdpa | Distance to IUCN nature reserve (2000-17) | 30” (~900 m) | UNEP-WCMS & IUCN [67] |
| cov_viirs | Resampled VIIRS night-time lights (2012-2016) | 30” (~900 m) | NOAA [68] |
| cov_dmsp | Resampled DMSP-OLS night-time lights (2011) | 30” (~900 m) | NOAA & Zhang, et al. [69,70] |
| cov_tt50k | Resampled travel time to cities of 50,000+ (2000) | 30” (~900 m) | Weiss, et al. [71] |
| cov_001 | Distance to cultivated areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_040 | Distance to woody areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_130 | Distance to cultivated areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_140 | Distance to herbaceous areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_150 | Distance to sparse vegetation areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_160 | Distance to aquatic vegetation areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_190 | Distance to urban areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_200 | Distance to bare areas (2015) | 9” (~300 m) | ESA CCI – LC [72] |
| cov_cciwat | Distance to ESA-CCI-LC inland waterbodies (2000-12) | 4.5” (~150 m) | ESA CCI [73] |
| cov_slope | SRTM-based slope (2000) | 3” (~90 m) | de Ferranti [74,75] |
| cov_topo | SRTM-based elevation (2000) | 3” (~90 m) | de Ferranti [74,75] |
| cov_coast | Distance to open-water coastline (2000-20) | 3” (~90 m) | CIESIN [76] |
| cov_ghsl | Distance to urban area (2012) | 1.26” (~38 m) | Pesaresi, et al. [77] |
| cov_guf | Distance to settlement built-up areas (2012) | 2.8” (~84 m) | DLR EOC [78] |
| cov_bsgme | Distance to built settlement expansion (2016) | 3” (~90 m) | Nieves, et al. [79] |
| cov_prec | Average total annual precipitation (1970-2000) | 30” (~900 m) | Fick and Hijmans [80] |
| cov_temp | Average annual temperature (1970-2000) | 30” (~900 m) | Fick and Hijmans [80] |

## 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 area to represent RMSE per hectare (100x100m unit). We also evaluated RMSE in urban versus 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 $i$, and $n$ is the number of grid cells.

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

To better understand the mechanics of the WorldPop-Global model and workflow, two additional statistics were calculated. Bias, a measure of error direction and magnitude, was calculated for the two gridded population datasets derived from "true" population counts to tease out accuracy related to the model and covariate datasets alone. Bias reveals whether cell-level estimates are systematically under- or over-estimated. As above, bias was assessed in 100x100m cells as well as cell sizes that ranged up to 1x1km, and separately in urban versus rural areas.

$$Bias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n} \quad (2)$$

To assess the degree to which WorldPop-Global's non-zero population estimates resulted in misallocation of population, a third analysis was performed counting the entire modelled population in Khomas that was misallocated to cells which were unsettled according to the "true" population. Millions of near-zero cell-level estimates in Khomas located outside of Windhoek severely skewed these statistics and made them appear artificially accurate, therefore gridded population cell-level estimates of less than 1 person were excluded (see Supplement 4 for a visual).

# Results

RMSE in grid cells did not differ substantially across the simulated outdated-inaccurate census scenarios (Figures 6). Furthermore, errors only slightly decreased when the input data were aggregated to EA rather than constituency. The major driver of RMSE in cells was urban location; errors in urban cells were substantially larger than in rural cells (Figure 6). In urban areas, RMSE per hectare was lowest in 100x100m cells ranging from 25 to 35, while in rural areas, RMSE per hectare was lowest in cells 300x300m to 500x500m with an approximate value of 2 (Figure 6). 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 scenario results are provided in Supplement 5.

Assessment of bias in the two gridded population datasets derived from "true" 2016 population counts revealed systematic and substantial under-estimates of populations in urban cells, and less dramatic over-estimates of populations in rural cells (Table 4). For example, the average 300x300m urban cell underestimated the population by more than 200 people, while the average 300x300m rural cell was over-estimated by just 3 (constituency-level input) to 14 (EA-level input) people (Table 4).

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

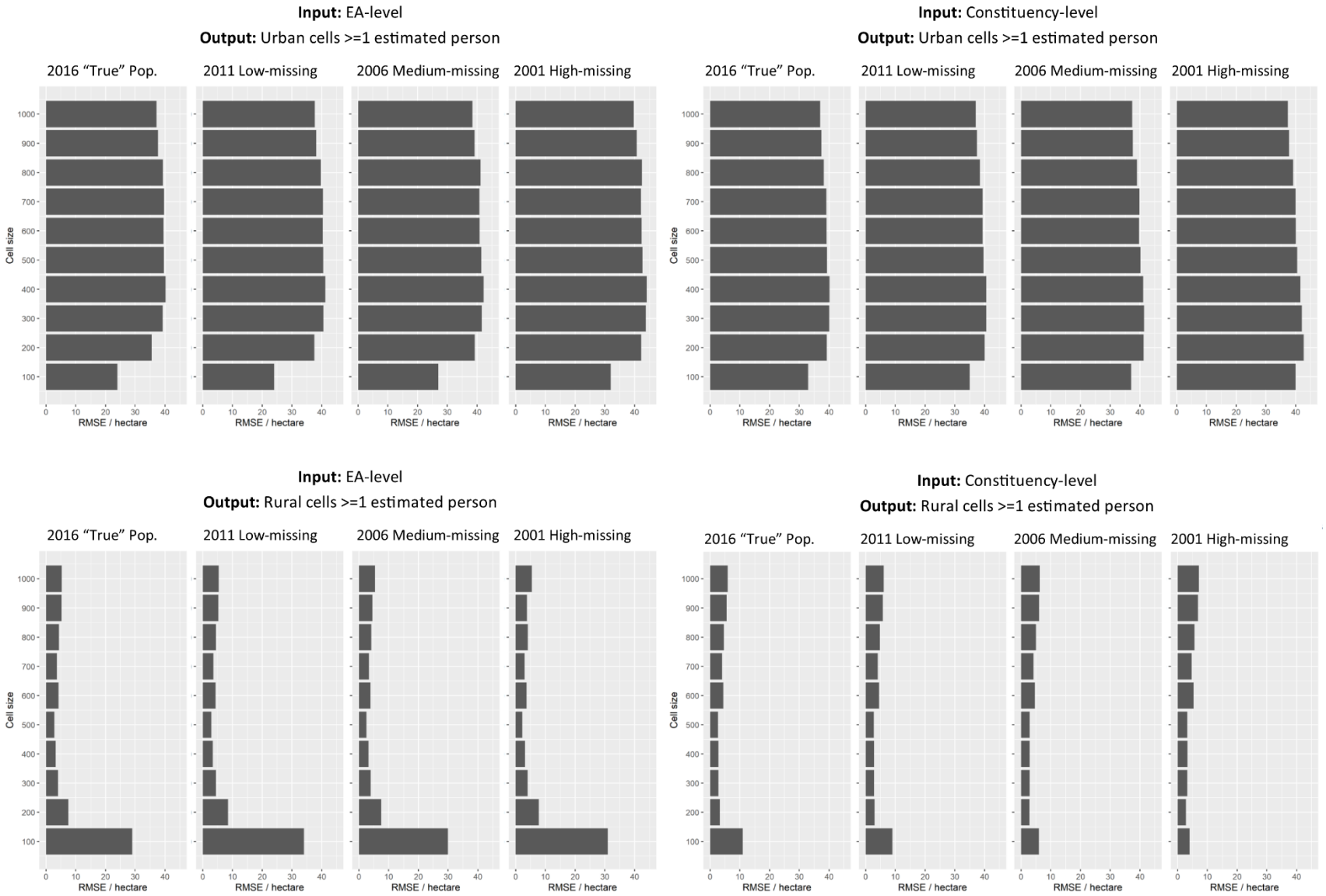



**Figure 6.** Root mean square error (RMSE) per hectare according to input population aggregation, a selection of scenarios, and cell size

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

**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)

| | EA-level input | | Constituency-level input | |
|---|---|---|---|---|
| Cell size | Rural | Urban | Rural | Urban |
| 100 | 20 | 0 | 7 | -8 |
| 200 | 18 | -85 | 6 | -86 |
| 300 | 14 | -223 | 3 | -214 |
| 400 | 8 | -416 | -1 | -394 |
| 500 | 3 | -650 | -8 | -616 |
| 600 | -22 | -933 | -34 | -891 |
| 700 | -33 | -1,293 | -51 | -1,229 |
| 800 | -72 | -1,664 | -90 | -1,556 |
| 900 | -126 | -2,026 | -152 | -1,974 |
| 1000 | -126 | -2,476 | -167 | -2,421 |

Table 5 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 5). 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 within, or near to, settlements. The rates of misallocation were similar across grid cell sizes when cells with less than one person were excluded (not reported).

**Table 5.** 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 ($m^2$) | EA-Level Input Population Misallocation | Constituency-Level Input Population Misallocation |
|---|---|---|
| 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 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 – however, a simulated realistic population could be used to assess the accuracy of other gridded population datasets. In the WorldPop-Global-unconstrained model of Khomas, Namibia, cell-level inaccuracies in 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 use of a log-transformation on input administrative 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 cells with zero population are assigned a very small fraction of a person when log-population is untransformed [19]. A concern is that non-zero population estimates across millions of unsettled cells might add up to 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. Table 5 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 [21,81], GHP-POP [16,17], HRSL [18], GRID3 [25,82], WPE [23]). Until recently, these settlement layers tended to be relatively coarse (e.g. GHS-BUILT 1x1km [83]) and/or had a tendency to omit areas with few sparse buildings (e.g. GUF [78]) 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 [84], Maxar/Ecopia [85]). 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, but masks all 100x100m cells without building footprints (in Africa) or built settlement (rest of the world), eliminating the issue of non-zero population estimates in unsettled cells [32].

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 [86]. With the recent release of several building

footprint datasets (e.g., Maxar/Ecopia in most of Africa [85], Bing in Tanzania and Uganda [87]), several new covariate layers have been created by the WorldPop team including number of buildings and total area of buildings in 100x100m cells [88]. 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 [89], 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. Table 4 provides strong evidence of this issue with the population in urban cells systematically underestimated regardless of cell size.

Most top-down gridded population datasets use averaged population densities from input areal units in some way to populate smaller grid cells, and are subject to similar error. 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 [18], and the Global Human Settlement GHS-POP dataset takes a similar approach disaggregating input populations into 250x250m cells that are classified as settled [16,17]. The problem of aggregate averages is accentuated when input areal units are geographically large because average densities tend to be smaller across larger spaces. In these cases, WorldPop-Global will incorporate training data from a neighbouring country that has finer-scale input population counts [19]. Our analysis showed, however, that even when relatively small geographic units (census EAs) were used as the input population area unit, urban cell-level errors were still substantial, and cell-level accuracy was only marginally improved with EA-level inputs compared to constituency-level inputs (Figure 7). This suggests that much finer-scale training data (e.g., 100x100m) should be incorporated into models, particularly from high-density urban areas, to ensure that the WorldPop Random Forest model contains sufficiently large population density values to assign to cells. Fine-scale training datasets might be simulated, like the dataset that we created here, or they might come from existing geo-referenced household survey enumerations (e.g., World Bank Living Standards Measurement Surveys or Demographic and Health Surveys), or slum community profiles such as those published on the Know Your City Campaign website [90]. 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, and likely substantially improve cell-level accuracy within urban areas.

This analysis of WorldPop-Global-unconstrained data raises broader questions about the cell-level accuracy of all 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 [91–93]. 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. These two steps – use of building footprint covariates and finer-scale training data - stand to improve cell-level accuracy of all top-down and bottom-up gridded population

datasets derived from complex models, including all WorldPop datasets as well as LandScan [21,22], WPE [23], and GRID3 [25,82]. Gridded population datasets that do not vary (weight) population densities within areal units (e.g., HRSL [18], GHS-POP [16,17], GPW [14,15]) should be used cautiously within urban areas, as cell-level inaccuracies are likely to be high. Given the increased use of gridded population datasets for monitoring health and development outcomes in small areas, it is imperative that all gridded population datasets area assessed for cell-level accuracy.

## 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 [Internet]. Nairobi Kenya: UN-Habitat; 2020. 377 p. Available from: https://unhabitat.org/sites/default/files/2020/10/wcr_2020_report.pdf

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 [Internet]. 2020;38(14):3062–71. Available from: https://doi.org/10.1016/j.vaccine.2020.02.070

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. 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 [Internet]. 2017;33(1):15–38. Available from: https://muse.jhu.edu/article/653072/pdf

7. Carr-Hill R. Missing millions and measuring development progress. World Dev [Internet]. 2013;46:30–44. Available from: http://dx.doi.org/10.1016/j.worlddev.2012.12.017

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: http://www.prb.org/Publications/Articles/2007/ObjectionsOverNigerianCensus.aspx

11. 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 [Internet]. 2020;19:34. Available from: https://doi.org/10.1186/s12942-020-00230-4

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. 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 [Internet]. 2015 Jul 3;1(3):226–34. Available from: http://dx.doi.org/10.1080/23754931.2015.1014272

15. 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

16. 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 [Internet]. Ispra Italy: European Commission Joint Research Centre; 2016. 67 p. Available from: http://publications.jrc.ec.europa.eu/repository/handle/JRC97705

17. 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

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

19. 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.

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

21. 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.

22. 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

23. 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 [Internet]. 2018;17:1–12. Available from: http://datascience.codata.org/articles/10.5334/dsj-2018-020/

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

25. 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.

26. 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/

27. Hay S, Noor A, Nelson A, Tatem A. The accuracy of human population maps for public health application. Trop Med Int Heal [Internet]. 2005 Oct [cited 2020 Jan 13];10:1073–86. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3173846&tool=pmcentrez&rendertype=abstract

28. 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.

29. Bondarenko M, Nieves JJ, Stevens FR, Gaughan AE, Tatem A, Sorichetta A. wpgpRFPMS: Random Forests population modelling R scripts, version 0.1.0 [Internet]. Southampton UK; 2020. Available from: doi.org/10.5258/SOTON/WP00665

30. 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 [Internet]. 2019;3(2):108–39. Available from: https://doi.org/10.1080/20964471.2019.1625151

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

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

33. 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.

34. UNSD. Report on the results of a survey on census methods used by countries in the 2010 census round [Internet]. New York NY USA; 2010. (Working paper). Report No.: UNSD/DSSB/1. Available from: http://unstats.un.org/unsd/census2010.htm

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

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

37. NSA. Namibia 2011 Population and Housing Census main report [Internet]. Windhoek Namibia; 2011. Available from: http://www.nsa.org.na/files/downloads/Namibia 2011 Population and Housing Census Main Report.pdf

38. Newaya TP. Rapid urbanization and its influence on the growth of informal settlements in Windhoek, Namibia [dissertation] [Internet]. Cape Peninsula University of Technology; 2010. Available from: http://digitalknowledge.cput.ac.za/xmlui/handle/11189/340

39. 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 [Internet]. 2019;5(1). Available from: http://dx.doi.org/10.1057/s41599-019-0242-9

40. IOM. Migration in Namibia [Internet]. Geneva Switzerland; 2015. Available from: https://publications.iom.int/system/files/pdf/mp_namibia_for_web_14june2016.pdf

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

42. 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.

43. 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 [Internet]. Rio de Janeiro Brazil; 2003. Available from: https://biblioteca.ibge.gov.br/biblioteca-catalogo.html?id=21402&view=detalhes

44. Korale RBM. Post Enumeration Survey 2001 [Nepal Population Census] Draft Report [Internet]. Kathmandu Nepal; 2002. Available from:

http://202.45.144.7/cbsgov/nada/index.php/catalog/42/download/546

45. 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_Presentations/Tanzania.ppt

46. UBS. Post enumeration survey: 2002 Uganda populaiton and housing census [Internet]. Entebbe Uganda; 2005. Available from: https://unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10682.aspx

47. GSS. [Ghana] 2010 Population and Housing Census Post Enumeration Survey Report [Internet]. Accra Ghana; 2012. Available from: http/www.statsghana.gov.gh/../2010phc/010_Population_and Housing

48. CSO. [Zambia] 2010 Census of Population and Housing Post Enumeration Survey (PES) [Internet]. Lusaka Zambia; 2013. Available from: http://www.census.gov.ph/content/population-mountain-province-showed-increase14-thousand-results-2010-census-population-and

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

50. NSC. Census of India 2011: Report on post enumeration survey [Internet]. New Delhi India; 2014. Available from: http://censusindia.gov.in/2011Census/pes/pes_highlights.html

51. SSA. [South Africa] Census 2011 Post-enumeration survey [Internet]. Pretoria South Africa; 2012. Available from: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/485/download/8289

52. NISR. Post enumeration survey report: Fourth Population and Housing Census, Rwanda, 2012 [Internet]. Kigali Rwanda; 2010. Available from: http://www.statistics.gov.rw/publication/rphc4-post-enumeration-survey

53. 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.

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

55. 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 [Internet]. 2017;17:1–8. Available from: http://dx.doi.org/10.1186/s12874-017-0293-9

56. Treiman DJ, Mason WM, Lu Y, Pan Y, Qi Y, Song S. Observations on the design and implementation of sample surveys in China [Internet]. Los Angeles CA USA; 2005. Report No.: CCPR-006-05. Available from: https://www.semanticscholar.org/paper/Observations-on-the-Design-and-Implementation-of-in-Treiman-Mason/615f1059443a3be862bcfa3aa46253ef109ad4e8

57. Carr-hill ROY. Improving population and poverty estimates with citizen surveys: Evidence from East Africa. World Dev [Internet]. 2017;93:249–59. Available from: http://dx.doi.org/10.1016/j.worlddev.2016.12.017

58. Ebenstein A, Zhao Y. Tracking rural-to-urban migration in China: Lessons from the 2005 inter-census population survey. Popul Stud (NY) [Internet]. 2015;69(3):337–53. Available from: http://dx.doi.org/10.1080/00324728.2015.1065342

59. 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 [Internet]. 2013;62(33):653–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23965827

60. 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.

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

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

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

64. Lucci P, Bhatkal T, Khan A. Are we underestimating urban poverty? World Dev [Internet]. 2018;103:297–310. Available from: https://doi.org/10.1016/j.worlddev.2017.10.022

65. Breiman L. Random forests. Mach Learn [Internet]. 2001;45:5–32. Available from: https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf

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

67. 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/

68. NOAA. VIIRS nighttime lights [Internet]. 2012 [cited 2020 Jan 13]. Available from: https://maps.ngdc.noaa.gov/viewers/VIIRS_DNB_nighttime_imagery/index.html

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

70. 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 [Internet]. 2016;54(10):1–11. Available from: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=7490418&contentType=Journals+%26+Magazines

71. 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.

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

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

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

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

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

77. 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

78. 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/

79. 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 [Internet]. 2020;80(May 2019):101444. Available from: https://doi.org/10.1016/j.compenvurbsys.2019.101444

80. 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.

81. ORNL. LandScan documentation [Internet]. 2017 [cited 2020 Jan 13]. Available from: http://web.ornl.gov/sci/landscan/landscan_documentation.shtml

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

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

84. 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.

85. Maxar. Satellite Imagery [Internet]. Image Library. 2019 [cited 2020 Jan 13]. Available from: https://www.digitalglobe.com/products/satellite-imagery

86. 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 [Internet]. 2019;75:132–45. Available from: https://doi.org/10.1016/j.compenvurbsys.2019.01.006

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

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

89. Selvin HC. Durkheim's suicide and problems of empirical research. Am J Sociol [Internet]. 1958;63(6):607–19. Available from: www.jstor.org/stable/2772991

90. Slum Dwellers International. Know Your City [Internet]. 2016 [cited 2020 Mar 1]. Available from:

http://knowyourcity.info/explore-our-data/

91. 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.

92. 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 [Internet]. 2017;389:547–58. Available from: http://dx.doi.org/10.1016/S0140-6736(16)31650-6

93. 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 [Internet]. 2018;2(1):8. Available from: http://www.mdpi.com/2413-8851/2/1/8

94. Thomson DRDR, Bhattarai R, Khanal S, Manandhar S, Dhungel R, Gajurel S, et al. Addressing unintentional exclusion of vulnerable and mobile households in traditional surveys in Kathmandu, Dhaka, and Hanoi: A mixed-methods feasibility study. J Urban Heal [Internet]. 2020; Available from: https://www.preprints.org/manuscript/201910.0320/v2

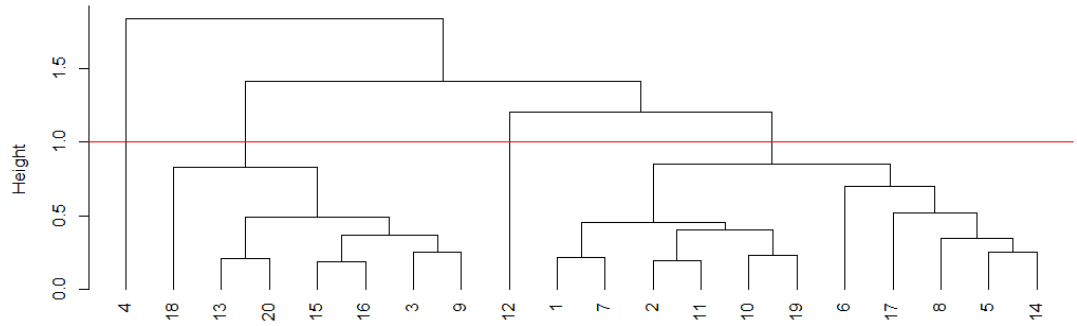
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

# 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 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

**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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

(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).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

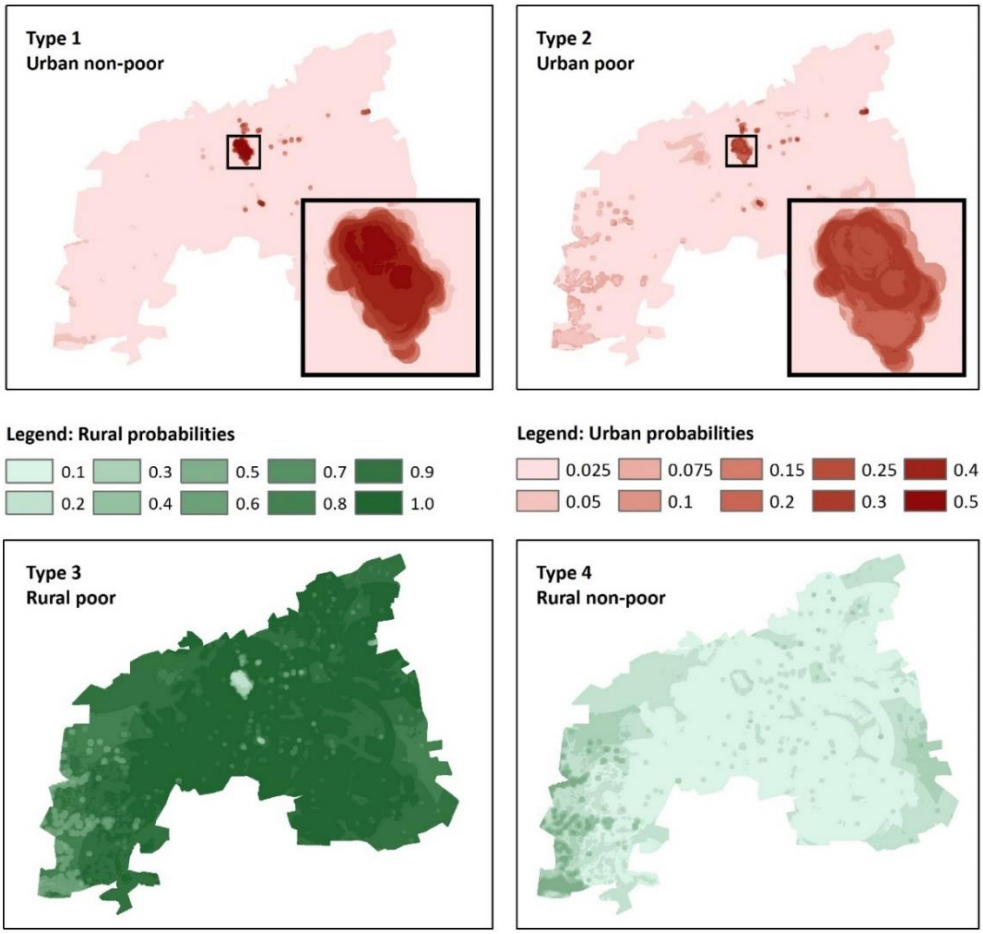



**Figure S1.2.** Household type probability surfaces (steps 1-4) in Khomas, Namibia population simulation

(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/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

**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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

**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

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

(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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

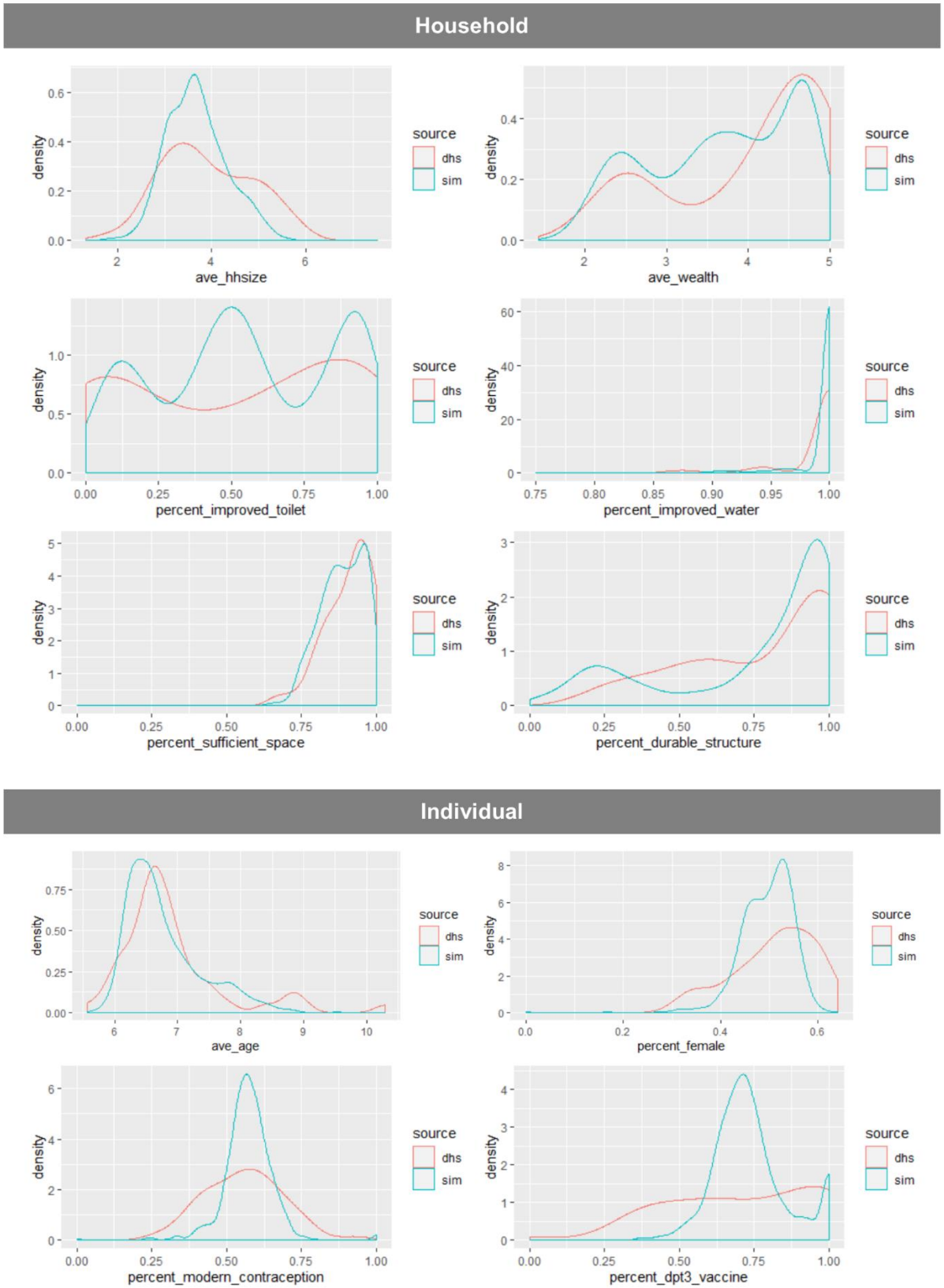



**Figure S1.3.** Comparison of household and individual outcomes by 2013 Namibia DHS cluster (n=53) and simulated population EA (n=922) in Khomas, Namibia

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

# Supplement 3. Literature Review Results

*Supplement to Thomson DR, Leasure DR, Bird T, Tzavidis N, Tatem AJ. 2021. How accurate are WorldPop-Global 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 | [53] | [7] | [57] | [58] | [59] | [60] | [61] | [62] | [63] | [64] | [54] |

*continued…*

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

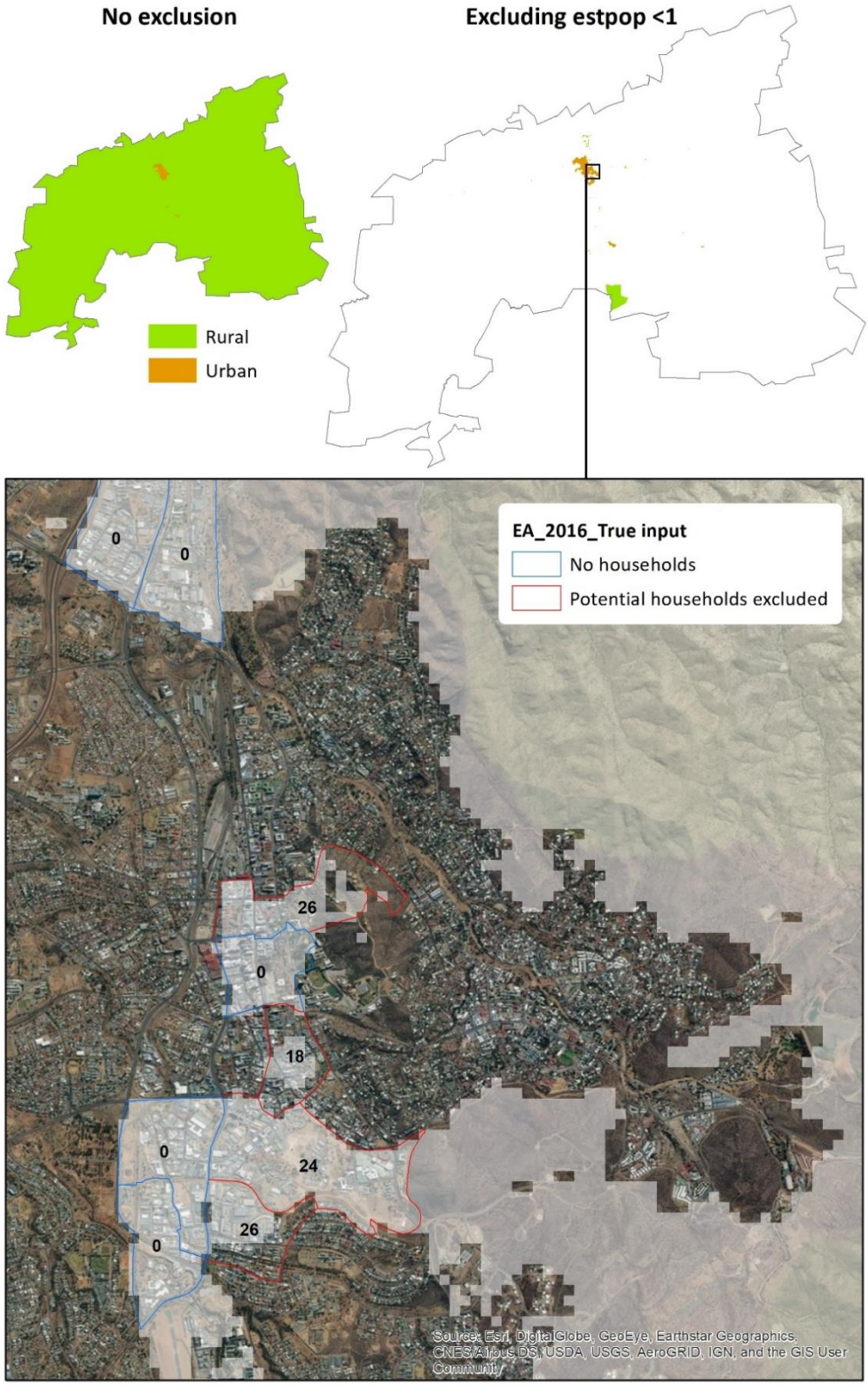
| Location & Housing Type | Stark (2017) Cambodia (kids only) | Treiman (2005) China | PES (2000) Brazil | PES (2001) Nepal | PES (2002) Tanzania | PES (2002) Uganda | PES (2010) Ghana | PES (2010) Zambia | PES (2010) Rwanda | PES (2011) Bangladesh | PES (2011) India | PES (2011) South Africa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | [55] | [56] | [43] | [44] | [45] | [46] | [47] | [48] | [52] | [49] | [50] | [51] |

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

# 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 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



*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

# 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 gridded population data at the cell-level?: A simulation analysis in urban Namibia.*

| **RMSE**, EA-level input data, excluding cells with estimated population <1 (no adjustment 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 | 29 | 26 | 29 | 28 | 30 | 33 | 30 | 31 | 31 | 29 | 31 | 31 | 30 | 30 | 31 |
| 200 | 99 | 101 | 103 | 109 | 102 | 104 | 108 | 111 | 102 | 103 | 105 | 110 | 105 | 106 | 108 | 111 |
| 300 | 238 | 240 | 244 | 250 | 243 | 245 | 249 | 254 | 241 | 242 | 245 | 250 | 245 | 246 | 249 | 254 |
| 400 | 429 | 431 | 435 | 443 | 437 | 439 | 442 | 450 | 434 | 436 | 439 | 447 | 440 | 442 | 445 | 451 |
| 500 | 660 | 664 | 669 | 680 | 673 | 674 | 678 | 689 | 668 | 670 | 676 | 682 | 673 | 675 | 680 | 684 |
| 600 | 936 | 939 | 942 | 956 | 951 | 957 | 959 | 970 | 945 | 949 | 952 | 967 | 958 | 959 | 967 | 976 |
| 700 | 1258 | 1258 | 1267 | 1282 | 1282 | 1283 | 1289 | 1303 | 1272 | 1275 | 1281 | 1295 | 1297 | 1301 | 1303 | 1307 |
| 800 | 1617 | 1618 | 1622 | 1647 | 1638 | 1644 | 1649 | 1666 | 1666 | 1675 | 1685 | 1687 | 1682 | 1684 | 1692 | 1699 |
| 900 | 1965 | 1971 | 1984 | 2011 | 1997 | 2001 | 2008 | 2036 | 2005 | 2014 | 2025 | 2036 | 2040 | 2041 | 2052 | 2063 |
| 1000 | 2398 | 2403 | 2433 | 2460 | 2437 | 2443 | 2454 | 2474 | 2462 | 2467 | 2470 | 2477 | 2511 | 2512 | 2513 | 2528 |
| | **Rural** | | | | | | | | | | | | | | | |
| 100 | 29 | 33 | 27 | 30 | 31 | 34 | 38 | 30 | 35 | 34 | 30 | 31 | 33 | 32 | 30 | 31 |
| 200 | 30 | 33 | 27 | 30 | 31 | 34 | 38 | 31 | 35 | 34 | 30 | 32 | 33 | 32 | 30 | 31 |
| 300 | 36 | 39 | 34 | 37 | 38 | 40 | 43 | 38 | 39 | 39 | 35 | 37 | 37 | 36 | 35 | 36 |
| 400 | 51 | 53 | 50 | 52 | 52 | 54 | 56 | 53 | 53 | 53 | 51 | 52 | 51 | 50 | 50 | 50 |
| 500 | 70 | 72 | 70 | 72 | 70 | 72 | 74 | 71 | 64 | 63 | 62 | 61 | 57 | 57 | 56 | 55 |
| 600 | 152 | 154 | 153 | 156 | 152 | 153 | 154 | 154 | 138 | 138 | 138 | 139 | 130 | 130 | 131 | 132 |
| 700 | 179 | 180 | 180 | 182 | 175 | 176 | 178 | 178 | 163 | 164 | 164 | 165 | 147 | 147 | 147 | 147 |
| 800 | 278 | 278 | 279 | 283 | 280 | 282 | 283 | 284 | 264 | 265 | 265 | 266 | 260 | 260 | 260 | 262 |
| 900 | 419 | 421 | 425 | 428 | 424 | 423 | 426 | 431 | 366 | 368 | 368 | 370 | 305 | 305 | 305 | 305 |
| 1000 | 525 | 527 | 536 | 542 | 535 | 537 | 542 | 544 | 541 | 541 | 542 | 542 | 536 | 536 | 536 | 539 |
| | **Urban** | | | | | | | | | | | | | | | |
| 100 | 24 | 24 | 25 | 29 | 24 | 24 | 25 | 30 | 25 | 25 | 27 | 30 | 26 | 27 | 29 | 32 |
| 200 | 142 | 145 | 150 | 158 | 148 | 150 | 154 | 162 | 151 | 153 | 157 | 164 | 158 | 159 | 163 | 169 |
| 300 | 353 | 357 | 362 | 373 | 362 | 365 | 370 | 381 | 366 | 369 | 374 | 384 | 377 | 379 | 386 | 394 |
| 400 | 644 | 648 | 653 | 666 | 658 | 659 | 665 | 679 | 666 | 670 | 675 | 690 | 684 | 687 | 695 | 705 |
| 500 | 992 | 998 | 1005 | 1022 | 1009 | 1012 | 1020 | 1035 | 1020 | 1023 | 1036 | 1048 | 1039 | 1043 | 1060 | 1066 |
| 600 | 1423 | 1428 | 1432 | 1459 | 1442 | 1452 | 1456 | 1477 | 1457 | 1464 | 1471 | 1496 | 1493 | 1494 | 1513 | 1526 |
| 700 | 1946 | 1948 | 1952 | 1982 | 1981 | 1983 | 1986 | 2005 | 1987 | 1993 | 2000 | 2020 | 2041 | 2047 | 2059 | 2065 |
| 800 | 2513 | 2515 | 2519 | 2558 | 2540 | 2542 | 2558 | 2578 | 2593 | 2617 | 2635 | 2657 | 2655 | 2656 | 2689 | 2721 |
| 900 | 3048 | 3049 | 3063 | 3109 | 3089 | 3091 | 3094 | 3141 | 3135 | 3147 | 3173 | 3200 | 3227 | 3228 | 3255 | 3296 |
| 1000 | 3718 | 3720 | 3738 | 3773 | 3763 | 3765 | 3773 | 3824 | 3811 | 3829 | 3844 | 3865 | 3931 | 3932 | 3934 | 3973 |

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

**RMSE**, Constituency-level input data, excluding cells with estimated population <1 (no adjustment of 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 | 25 | 25 | 25 | 27 | 25 | 25 | 26 | 27 | 26 | 27 | 27 | 29 | 28 | 29 | 29 | 30 |
| 200 | 111 | 112 | 114 | 117 | 112 | 113 | 115 | 119 | 117 | 118 | 118 | 124 | 123 | 123 | 125 | 129 |
| 300 | 253 | 255 | 258 | 263 | 255 | 258 | 261 | 267 | 263 | 265 | 266 | 275 | 275 | 273 | 278 | 286 |
| 400 | 449 | 453 | 456 | 464 | 453 | 457 | 461 | 470 | 464 | 467 | 468 | 485 | 483 | 481 | 487 | 502 |
| 500 | 689 | 695 | 699 | 710 | 693 | 700 | 707 | 719 | 709 | 714 | 717 | 741 | 737 | 736 | 745 | 768 |
| 600 | 982 | 991 | 993 | 1011 | 989 | 993 | 1002 | 1023 | 1007 | 1014 | 1015 | 1048 | 1045 | 1039 | 1052 | 1089 |
| 700 | 1329 | 1341 | 1346 | 1362 | 1340 | 1352 | 1356 | 1376 | 1363 | 1368 | 1370 | 1407 | 1406 | 1402 | 1421 | 1468 |
| 800 | 1701 | 1708 | 1717 | 1756 | 1702 | 1727 | 1746 | 1777 | 1752 | 1769 | 1777 | 1820 | 1817 | 1810 | 1825 | 1885 |
| 900 | 2087 | 2112 | 2120 | 2164 | 2106 | 2117 | 2144 | 2186 | 2154 | 2174 | 2179 | 2249 | 2234 | 2227 | 2239 | 2310 |
| 1000 | 2581 | 2615 | 2598 | 2648 | 2592 | 2612 | 2642 | 2685 | 2648 | 2662 | 2673 | 2761 | 2738 | 2739 | 2773 | 2846 |
| | **Rural** | | | | | | | | | | | | | | | |
| 100 | 11 | 10 | 9 | 7 | 9 | 9 | 8 | 6 | 7 | 7 | 6 | 5 | 6 | 6 | 5 | 4 |
| 200 | 13 | 13 | 12 | 11 | 12 | 12 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 300 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 | 27 | 27 | 27 | 28 | 29 |
| 400 | 44 | 44 | 44 | 46 | 44 | 45 | 45 | 47 | 46 | 46 | 46 | 49 | 49 | 48 | 50 | 53 |
| 500 | 66 | 68 | 68 | 70 | 67 | 68 | 69 | 72 | 70 | 71 | 71 | 75 | 75 | 74 | 76 | 81 |
| 600 | 158 | 161 | 161 | 166 | 160 | 161 | 164 | 170 | 165 | 167 | 167 | 177 | 176 | 174 | 179 | 192 |
| 700 | 194 | 198 | 198 | 203 | 197 | 199 | 201 | 207 | 202 | 203 | 203 | 212 | 214 | 213 | 217 | 232 |
| 800 | 297 | 303 | 302 | 315 | 300 | 305 | 311 | 322 | 313 | 318 | 319 | 335 | 333 | 331 | 337 | 360 |
| 900 | 453 | 465 | 467 | 485 | 461 | 465 | 477 | 494 | 480 | 490 | 492 | 523 | 515 | 512 | 517 | 554 |
| 1000 | 590 | 605 | 597 | 620 | 594 | 603 | 617 | 637 | 620 | 625 | 627 | 670 | 658 | 658 | 676 | 715 |
| | **Urban** | | | | | | | | | | | | | | | |
| 100 | 33 | 33 | 35 | 37 | 34 | 35 | 35 | 37 | 36 | 37 | 37 | 39 | 38 | 39 | 39 | 40 |
| 200 | 157 | 158 | 161 | 164 | 159 | 160 | 162 | 164 | 163 | 164 | 165 | 168 | 168 | 169 | 169 | 171 |
| 300 | 361 | 361 | 366 | 369 | 364 | 365 | 367 | 370 | 369 | 370 | 372 | 376 | 376 | 377 | 377 | 379 |
| 400 | 642 | 641 | 649 | 653 | 647 | 649 | 651 | 654 | 653 | 654 | 657 | 663 | 662 | 663 | 663 | 665 |
| 500 | 982 | 983 | 992 | 995 | 986 | 991 | 994 | 996 | 994 | 998 | 1004 | 1009 | 1009 | 1009 | 1010 | 1012 |
| 600 | 1409 | 1410 | 1418 | 1426 | 1413 | 1417 | 1421 | 1429 | 1428 | 1430 | 1431 | 1439 | 1439 | 1442 | 1440 | 1444 |
| 700 | 1917 | 1914 | 1930 | 1933 | 1925 | 1930 | 1931 | 1937 | 1939 | 1944 | 1953 | 1951 | 1954 | 1955 | 1955 | 1961 |
| 800 | 2447 | 2448 | 2462 | 2470 | 2460 | 2461 | 2468 | 2471 | 2469 | 2482 | 2496 | 2498 | 2497 | 2498 | 2498 | 2506 |
| 900 | 3031 | 3022 | 3034 | 3045 | 3033 | 3034 | 3035 | 3046 | 3045 | 3045 | 3047 | 3058 | 3058 | 3058 | 3059 | 3061 |
| 1000 | 3702 | 3702 | 3705 | 3708 | 3703 | 3704 | 3706 | 3708 | 3707 | 3719 | 3735 | 3737 | 3736 | 3737 | 3738 | 3739 |

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

| | **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 | 27.0 | 29.0 | 26.0 | 29.0 | 28.0 | 30.0 | 33.0 | 30.0 | 31.0 | 31.0 | 29.0 | 31.0 | 31.0 | 30.0 | 30.0 | 31.0 |
| 200 | 24.8 | 25.3 | 25.8 | 27.3 | 25.5 | 26.0 | 27.0 | 27.8 | 25.5 | 25.8 | 26.3 | 27.5 | 26.3 | 26.5 | 27.0 | 27.8 |
| 300 | 26.4 | 26.7 | 27.1 | 27.8 | 27.0 | 27.2 | 27.7 | 28.2 | 26.8 | 26.9 | 27.2 | 27.8 | 27.2 | 27.3 | 27.7 | 28.2 |
| 400 | 26.8 | 26.9 | 27.2 | 27.7 | 27.3 | 27.4 | 27.6 | 28.1 | 27.1 | 27.3 | 27.4 | 27.9 | 27.5 | 27.6 | 27.8 | 28.2 |
| 500 | 26.4 | 26.6 | 26.8 | 27.2 | 26.9 | 27.0 | 27.1 | 27.6 | 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 | 26.4 | 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 | 26.1 | 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 | 26.5 |
| 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 | 29.0 | 33.0 | 27.0 | 30.0 | 31.0 | 34.0 | 38.0 | 30.0 | 35.0 | 34.0 | 30.0 | 31.0 | 33.0 | 32.0 | 30.0 | 31.0 |
| 200 | 7.5 | 8.3 | 6.8 | 7.5 | 7.8 | 8.5 | 9.5 | 7.8 | 8.8 | 8.5 | 7.5 | 8.0 | 8.3 | 8.0 | 7.5 | 7.8 |
| 300 | 4.0 | 4.3 | 3.8 | 4.1 | 4.2 | 4.4 | 4.8 | 4.2 | 4.3 | 4.3 | 3.9 | 4.1 | 4.1 | 4.0 | 3.9 | 4.0 |
| 400 | 3.2 | 3.3 | 3.1 | 3.3 | 3.3 | 3.4 | 3.5 | 3.3 | 3.3 | 3.3 | 3.2 | 3.3 | 3.2 | 3.1 | 3.1 | 3.1 |
| 500 | 2.8 | 2.9 | 2.8 | 2.9 | 2.8 | 2.9 | 3.0 | 2.8 | 2.6 | 2.5 | 2.5 | 2.4 | 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 | 3.7 | 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 | 4.3 | 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 | 24.0 | 24.0 | 25.0 | 29.0 | 24.0 | 24.0 | 25.0 | 30.0 | 25.0 | 25.0 | 27.0 | 30.0 | 26.0 | 27.0 | 29.0 | 32.0 |
| 200 | 35.5 | 36.3 | 37.5 | 39.5 | 37.0 | 37.5 | 38.5 | 40.5 | 37.8 | 38.3 | 39.3 | 41.0 | 39.5 | 39.8 | 40.8 | 42.3 |
| 300 | 39.2 | 39.7 | 40.2 | 41.4 | 40.2 | 40.6 | 41.1 | 42.3 | 40.7 | 41.0 | 41.6 | 42.7 | 41.9 | 42.1 | 42.9 | 43.8 |
| 400 | 40.3 | 40.5 | 40.8 | 41.6 | 41.1 | 41.2 | 41.6 | 42.4 | 41.6 | 41.9 | 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 | 40.1 | 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 | 40.4 | 40.4 | 40.5 | 40.5 | 40.9 | 40.6 | 40.7 | 40.8 | 41.2 | 41.7 | 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 | 38.7 | 39.3 | 39.3 | 39.3 | 39.7 |

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0492.v1

| | **RMSE per hectare**, Constituency-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 | 25.0 | 25.0 | 25.0 | 27.0 | 25.0 | 25.0 | 26.0 | 27.0 | 26.0 | 27.0 | 27.0 | 29.0 | 28.0 | 29.0 | 29.0 | 30.0 |
| 200 | 27.8 | 28.0 | 28.5 | 29.3 | 28.0 | 28.3 | 28.8 | 29.8 | 29.3 | 29.5 | 29.5 | 31.0 | 30.8 | 30.8 | 31.3 | 32.3 |
| 300 | 28.1 | 28.3 | 28.7 | 29.2 | 28.3 | 28.7 | 29.0 | 29.7 | 29.2 | 29.4 | 29.6 | 30.6 | 30.6 | 30.3 | 30.9 | 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 | 30.4 | 31.4 |
| 500 | 27.6 | 27.8 | 28.0 | 28.4 | 27.7 | 28.0 | 28.3 | 28.8 | 28.4 | 28.6 | 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 | 30.3 |
| 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 | 29.5 |
| 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 | 26.9 | 26.5 | 26.6 | 26.7 | 27.6 | 27.4 | 27.4 | 27.7 | 28.5 |
| | **Rural** | | | | | | | | | | | | | | | |
| 100 | 11.0 | 10.0 | 9.0 | 7.0 | 9.0 | 9.0 | 8.0 | 6.0 | 7.0 | 7.0 | 6.0 | 5.0 | 6.0 | 6.0 | 5.0 | 4.0 |
| 200 | 3.3 | 3.3 | 3.0 | 2.8 | 3.0 | 3.0 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 |
| 300 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.9 | 2.9 | 2.9 | 2.9 | 3.0 | 3.0 | 3.0 | 3.1 | 3.2 |
| 400 | 2.8 | 2.8 | 2.8 | 2.9 | 2.8 | 2.8 | 2.8 | 2.9 | 2.9 | 2.9 | 2.9 | 3.1 | 3.1 | 3.0 | 3.1 | 3.3 |
| 500 | 2.6 | 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 | 3.0 | 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 | 4.1 | 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 | 6.1 | 6.0 | 6.2 | 5.9 | 6.0 | 6.2 | 6.4 | 6.2 | 6.3 | 6.3 | 6.7 | 6.6 | 6.6 | 6.8 | 7.2 |
| | **Urban** | | | | | | | | | | | | | | | |
| 100 | 33.0 | 33.0 | 35.0 | 37.0 | 34.0 | 35.0 | 35.0 | 37.0 | 36.0 | 37.0 | 37.0 | 39.0 | 38.0 | 39.0 | 39.0 | 40.0 |
| 200 | 39.3 | 39.5 | 40.3 | 41.0 | 39.8 | 40.0 | 40.5 | 41.0 | 40.8 | 41.0 | 41.3 | 42.0 | 42.0 | 42.3 | 42.3 | 42.8 |
| 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 | 41.9 | 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 | 41.4 | 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 | 39.3 | 39.4 | 39.5 | 39.7 | 39.7 | 39.7 | 39.8 | 40.0 | 40.0 | 40.1 | 40.0 | 40.1 |
| 700 | 39.1 | 39.1 | 39.4 | 39.4 | 39.3 | 39.4 | 39.4 | 39.5 | 39.6 | 39.7 | 39.9 | 39.8 | 39.9 | 39.9 | 39.9 | 40.0 |
| 800 | 38.2 | 38.3 | 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 | 37.1 | 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 |