

Improving the accuracy of gridded population estimates in cities and slums to monitor SDG 11: Evidence from a simulation study in Namibia

Dana R. Thomson¹, Forrest R. Stevens², Robert Chen³, Gregory Yetman³, Alessandro Sorichetta⁴, Andrea E. Gaughan²

1. Faculty of Geo-Information Science & Earth Observation, University of Twente, Enschede Netherlands
 2. Department of Geography & Geosciences, University of Louisville, Louisville KY USA
 3. Center for International Earth Science Information Network (CIESIN), Columbia University, New York NY USA
 4. WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton UK
- * Corresponding author: dana.r.thomson@gmail.com

Highlights

- Slum and deprived area populations are under-represented in LMIC censuses
- Gridded population estimates derived from census can inherit census inaccuracies
- Gridded population model inputs can exacerbate inaccuracies in deprived urban areas
- Accuracy is greatly improved with building footprint metrics and slum training data
- Settlement and building typology layers might further improve model accuracy

Abstract (200 of 200 words)

People living in slums and other deprived areas in low- and middle-income country (LMIC) cities are under-represented in censuses, and subsequently in "top-down" gridded population estimates. Modelled gridded population data are a unique source of disaggregated population information to calculate local development indicators such as the Sustainable Development Goals (SDGs). This study evaluates if, and how, WorldPop-Global (WPG) -Unconstrained and -Constrained "top-down" datasets might be improved in a simulated realistic LMIC urban population by incorporating slum profile population counts into model training. We found that the WPG-Unconstrained model with or without slum training data grossly underestimated population in urban deprived areas while grossly overestimating population in rural areas. SDG 11.1.1, the percent of population living in slums, for example, was estimated to be 20% or less compared to a "true" value of 29.5%. The WPG-Constrained model, which included building auxiliary datasets, far more accurately estimated the population in all grid cells (including rural areas), and the inclusion of slum training data further improved estimates such that SDG 11.1.1 was estimated at 27.1% and 27.0%, respectively. Inclusion of building metrics and slum training data in "top-down" gridded population models can substantially improve grid cell-level accuracy in both urban and rural areas.

Keywords

LMIC, urban, deprivation, informal settlement, poverty, Global South

1. Introduction

The United Nation's 2030 Agenda for Sustainable Development outlines a wide-ranging, integrated set of 17 Sustainable Development Goals (SDGs) and 169 targets coupled with 231 unique indicators for measuring and monitoring progress toward sustainable global development [1]. SDG 11 focuses on sustainable development of cities and communities, including the upgrading and integration of slums, informal settlements, and other deprived areas into cities. SDG indicator 11.1.1 specifically measures the percent of urban populations living in slums, informal settlements, or inadequate housing [2]. Fundamental to SDG 11 is the ability to measure all populations in urban settings; however, those living in deprived areas, particularly in low- and middle-income countries (LMIC), are often grossly under-represented in official demographic datasets such as censuses and household surveys [3,4]. The average LMIC census omits half of residents in slums and informal settlements, though this varies widely with some censuses omitting nearly all slum-dwellers [3,4]. Missing population counts are crucial to SDG efforts because cities that are expected to grow the most over the next 30 years will see much of that growth concentrated in deprived areas [5].

Modelled gridded population datasets that are based on census data can inherit some census data limitations, but they also provide an important proxy for fine-scale population distribution information that otherwise would not exist and are crucial to ensure that no one is left behind, especially in countries with coarse and/or poor quality census data [6]. However, gridded population modelling assumptions and approaches can further mask the urban poorest, for example, by assuming that average population density in large administrative units reflect population density in smaller grid cells [7]. Although the ability to disaggregate population and demographic information at fine-scale is vital to measuring more than half of SDG indicators [8], and despite continued advances in the data and techniques for spatial disaggregation of population data, reasonable population estimates for most deprived urban areas are lacking globally [7].

1.1. Understanding available gridded population datasets

Nine multi-country gridded population datasets are available in LMICs: LandScan [9,10], WorldPop-Global-Unconstrained (hereafter WPG-Unconstrained) [11,12], WorldPop-Global-Constrained (hereafter WPG-Constrained) [13], HRSL [14], GRID3 [15,16], GHS-POP [17,18], WPE [19], GPW [20,21], and WorldPop-Peanut Butter [22] (Table 1). While more detailed summaries of these datasets are available elsewhere [6,7,23], we provide a visual glossary of terms to understand their key characteristics (Figure 1). Top-down gridded population datasets are derived by disaggregating census or other complete population counts to grid cells (i.e., GHS-POP, GPW, HRSL, LandScan, WPG-Constrained, WPG-Unconstrained, and WPE), while bottom-up estimates are derived from micro-census population counts or assumptions about the average population per building (i.e., GRID3, WP-Peanut Butter). Datasets are referred to as unconstrained when non-zero population estimates are provided for all grid cells including vast unsettled areas (i.e., GPW, WPG-Unconstrained), while constrained gridded population datasets have estimates in only those grid cells classified as settled or built-up (i.e., GHS-POP, GRID3, HRSL, LandScan, WPG-Constrained, WP-Peanut Butter, and WPE). An important distinction between gridded population datasets is the degree to which the outputs are modelled. Un-modelled gridded population datasets are created through direct disaggregation (i.e., GPW) or aggregation (i.e., WP-Peanut Butter) based on simple assumption and no auxiliary datasets. Lightly modelled gridded population datasets involve simple direct disaggregation, but make use of an auxiliary data such as land cover type to vary the population disaggregation (i.e., GHS-POP, HRSL) (Table 1, Figure 1). Highly modelled datasets are created with a range of geo-statistical models or geographic algorithms that use multiple auxiliary datasets, and often include a measure of model accuracy (i.e., WPG-Constrained, WPG-Unconstrained) or output uncertainty (i.e., GRID3, LandScan, and WPE) (Table 1, Figure 1).

Table 1. Summary of multi-country gridded population datasets available in low- and middle-income countries

Dataset	Model Approach	Output Scale (at Equator)	Citations
HRSL	Top-down, Constrained, Lightly modelled	~30x30 m	[14]
WPG-Unconstrained	Top-down, Unconstrained, Highly modelled	~100x100 m	[11,12]
WPG-Constrained	Top-down, Constrained, Highly modelled	~100x100 m	[12,13]
WP-Peanut Butter	Bottom-up, Constrained, Un-modelled	~100x100 m	[22]
GRID3	Bottom-up, Constrained, Highly modelled	~100x100 m	[15,16]
WPE	Top-down, Constrained, Highly modelled	162x462 m	[19]
GHS-POP	Top-down, Constrained, Lightly modelled	~250x250 m	[17,18]
GPW	Top-down, Unconstrained, Un-modelled	~1x1 km	[20,21]
LandScan	Top-down, Constrained, Highly modelled	~1x1 km	[9,10]

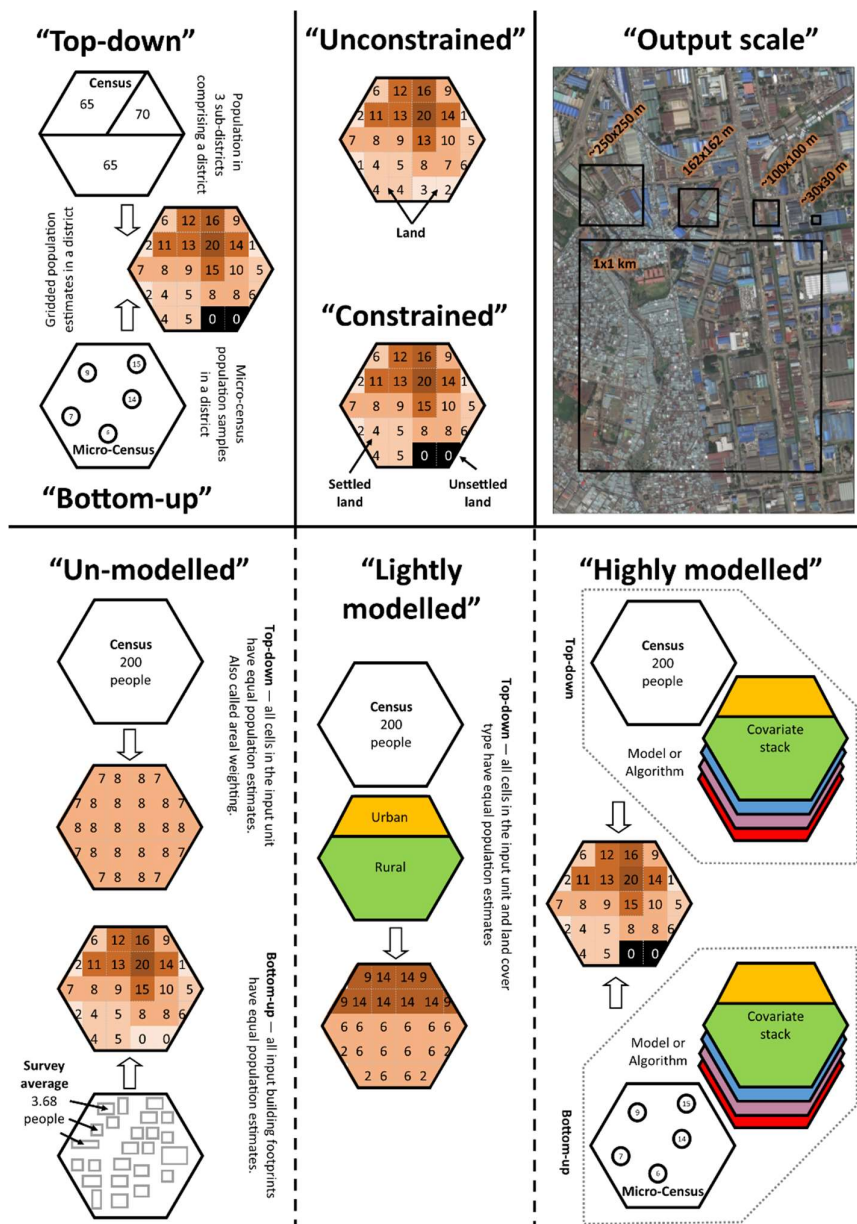


Figure 1. Visual glossary of terms used to describe gridded population datasets

1.2. Gridded population accuracy in deprived urban areas

Previous work evaluated the accuracy of the above nine gridded population datasets in Nigerian and Kenyan slums as compared to mapped and field referenced data by Slum Dwellers International (SDI) [7]. Findings from that study highlight that the most accurate gridded population datasets only predicted 39% of field-referenced slum residents [7]. There were potentially two reasons for these systematic underestimates of slum dwellers across datasets. First, the implicit assumption in all top-down gridded population models is that average population densities in larger administrative units represent population densities in much smaller grid cells [24]. Second, most gridded products assessed in that study did not include an auxiliary dataset, such as building density, which correlates with variation in urban population density at the scale of the gridded output [7]. The exception was WPG-Constrained which uses cell-level building footprint metrics (e.g. building density, total building edge length) to estimate population counts at the grid cell-level in most African countries [13].

The evaluation of gridded population dataset accuracy within cities, particularly within urban deprived areas or grid cells, is challenging and generally only performed in high-income countries where highly detailed reference data are available [e.g., 13,14]. An exception is an analysis which evaluated grid cell-level accuracy of WPG-Unconstrained gridded population using a realistic simulated, geo-referenced population in Khomas, Namibia [3]. The simulation analysis showed that even when “true” population counts were input to the model, 0.5 to 1 non-slum dweller and 0.75 to 1.5 slum dwellers were omitted for every 1 person accurately estimated in an urban grid cell [3]. The systematic underestimation of slum populations raised ethical questions about the use of gridded population estimates in development applications such as SDG reporting. How much value does a gridded population dataset, like WPG-Unconstrained, offer for SDG 11.1.1 monitoring if it masks – and therefore reinforces marginalisation of – the urban poorest?

The aim of this study is to evaluate if, and how, gridded population datasets might be improved to overcome systematic under-estimates of populations in deprived areas in a typical low- and middle-income country urban environment. This study focuses on WorldPop open-access data products because they have been widely used in development practice for over a decade, often in urban settings [23,27–30]. It specifically evaluates grid cell-level accuracy in the WPG-Unconstrained and WPG-Constrained datasets, and the effect of supplementing each model with SDI-like slum population counts during model training. This is done by using the earlier referenced simulated “true” population in Khomas, Namibia [3] as the reference population and to calculate the population density inputs to all models.

2. Methods

2.1 Setting

This study utilised an existing 2016 simulated “true” population in Khomas, Namibia [3], which is described in detail in section 2.3.1. Khomas is one of 13 Regions (1st-level administrative units) in Namibia, and more than 95% of the population in Khomas lives in Windhoek, Namibia’s capital (Figure 2). Windhoek’s population and footprint have grown rapidly since 1990 when the country gained independence because many rural populations were able to move freely and legally to cities for the first time [31]. Between the 2001 and 2011 censuses, the population in Windhoek grew by a staggering 37% [32].

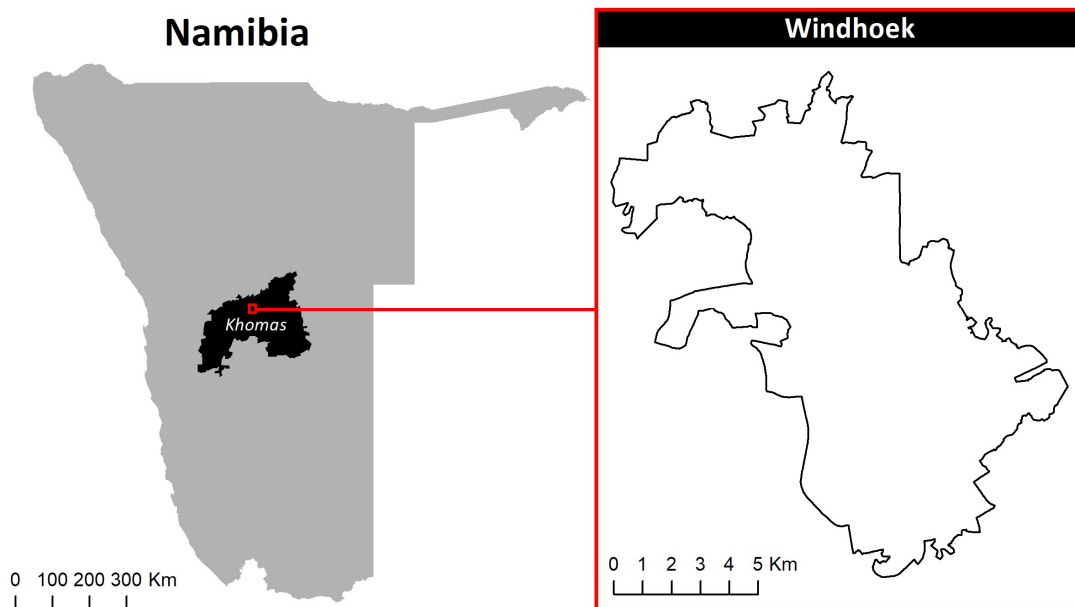


Figure 2. The location of Khomas region and Windhoek in Namibia

2.2 Analysis

Four gridded population datasets generated with four different models were compared: (1) WPG-Unconstrained standard model, (2) WPG-Unconstrained model plus slum training data, (3) WPG-Constrained standard model, and (4) WPG-Constrained model plus slum training data. WPG models are based on a Random Forest machine-learning algorithm which builds a series of “trees” that characterise the relationship between population density and a set of covariates [12]. WPG estimates are generated in three steps. First, for each input administrative unit, the corresponding average population density (people per hectare) and average covariate values are calculated, and used to generate the series of model “trees,” each with a set of parameters that define the relationship between population density and covariates at the administrative unit level. In the WPG-Unconstrained model, average population density is calculated from the entire areas of the administrative unit while in the WPG-Constrained model, average population density is calculated with the land area that is classified as settled only [33]. Second, gridded datasets representing the same set of covariates used to train the Random Forest model and the “trees” of parameters are used to generate multiple population density predictions in $\sim 100 \times 100 \text{m}$ grid cells [12]. The WPG-Unconstrained model contains 23 globally-available covariates [34], and the WPG-Constrained model in Africa contains an additional 10 covariates related to building footprint presence and patterns [35]. The hundreds of cell-level population predictions are averaged to create a final layer of cell-level population density. Third, because the averaged grid cell values do not sum to the administrative unit input populations values, the Random Forest output layer is used as a weights layer to disaggregate input administrative population totals into grid cells (see [36] for a visual and details).

During the first phase when the relationships between population density and covariate values are being defined at the administrative unit level, it is possible, and even advisable, to incorporate more spatially detailed information about population distribution, especially when the input administrative units are very coarse (geographically much larger than the $\sim 100 \times 100 \text{m}$ target grid cells) [12]. This is because the Random Forest model can only predict population densities during the second phase (to $\sim 100 \times 100 \text{m}$ grid cells) based on the results of the model fitting at the

administrative unit level (first phase). Finer-scale administrative units from a similar, neighbouring country are often used as supplemental training data to ensure the model includes finer and more realistic population densities relevant at the grid cell-level. Here, we test the use of a few slums as supplemental training data in the 2nd and 4th models because (a) slum are geographically smaller than constituencies, and (b) are likely to contain high population densities which, in theory, would enable the Random Forest model to predict more realistic population densities in high-density grid cells.

To assess the magnitude of cell-level error, we calculated root mean square error (RMSE), which penalises large errors, and to assess the direction of cell-level error, we calculated Bias. Both of these measure can be sensitive to population size (such that errors are larger in cells with larger populations), so we also calculate Normalised RMSE (NRMSE) and Relative Bias which adjust the statistic by average cell-level population. In error calculations, y_i is the “true” population in cell i , \hat{y}_i is the gridded population estimate in cell i , and n is the number of cells.

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

$$NRMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \div \frac{\sum_{i=1}^n (y_i)}{n} \times 100 \quad (2)$$

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

$$Relative\ Bias = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)}{n} \div \frac{\sum_{i=1}^n (y_i)}{n} \times 100 \quad (4)$$

2.3 Data

2.3.1 Reference: Population. The simulated “true” population was derived from 2011 census 20% microdata [37] and 2016 building point locations manually digitised using Maxar and SPOT imagery (40cm). The simulated households were joined to digitised building point locations based on probability surfaces of different household types. The input datasets and methodology for the simulated “true” population are detailed elsewhere and available for download and use [3,38]. Note that in the 2011 Namibian census, 89,438 households were recorded in Khomas, and the simulated population put this number at 97,667 households in 2016 (1.8% annual growth rate) [3], which is smaller than the projected growth rate estimated by the Namibia Statistics Agency (> 4%) [39]. This means that actual population in 2016 was likely higher than this simulated “true” population, and likely led to conservative results. However, for the purposes of this analysis, we treat the simulated population as the “true” location of individuals and households in 2016 (Figure 3).

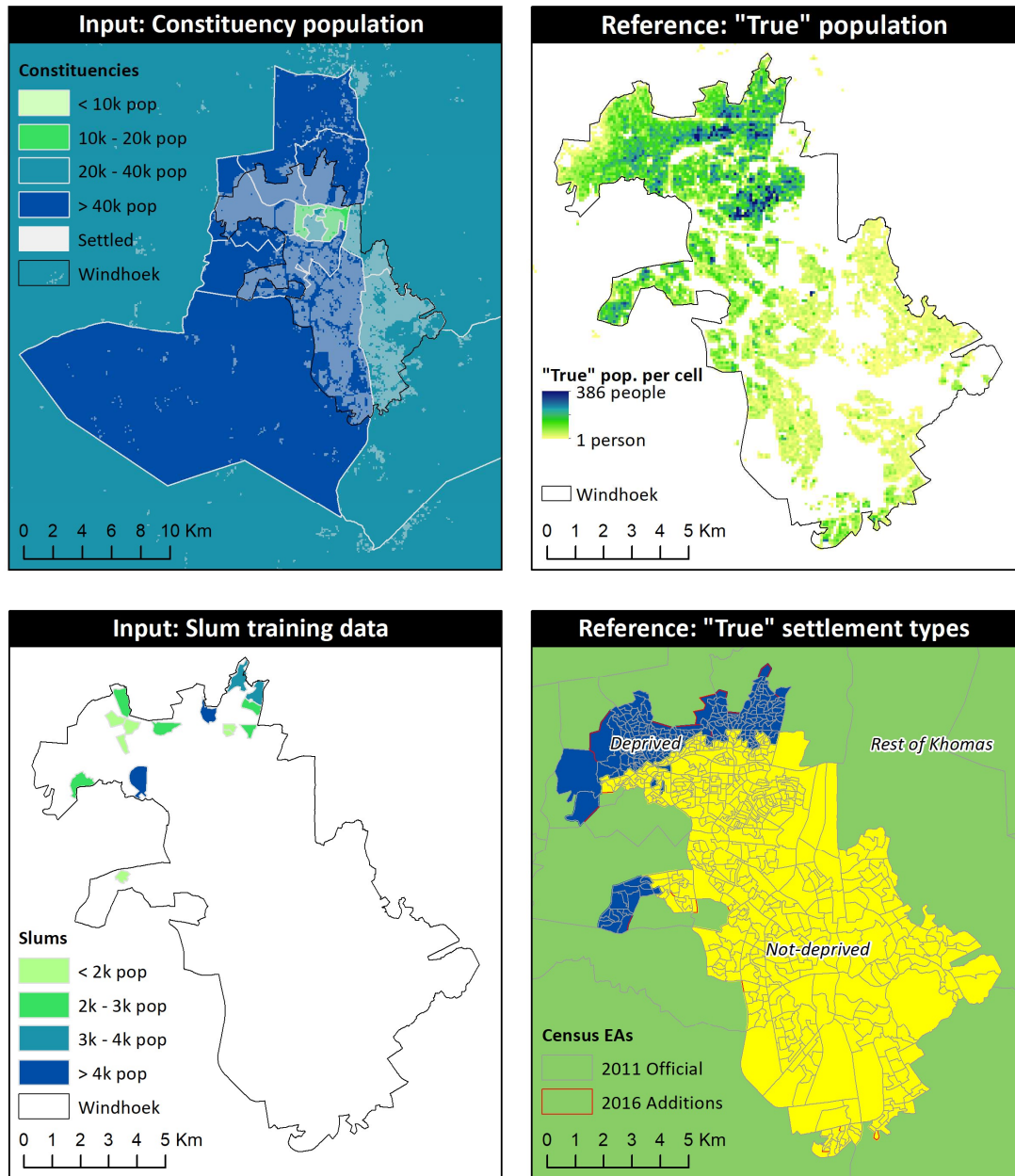


Figure 3. Model input population and KYC slum datasets, as well as the simulated “true” reference population and reference deprived/not-deprived areas in Windhoek

2.3.2 Reference: deprived, not-deprived and other areas. Starting with the Namibia Statistics Agency boundary file of 2011 census enumeration areas (EAs) [40], 24 EAs were manually added to Windhoek’s periphery to cover urban expansion through 2016, resulting in a total of 922 EAs across Khomas [3] (Figure 3). Each EA was then visually inspected over 2016 very high resolution Maxar and SPOT (40cm) imagery, and EAs were classified as Windhoek-Deprived, Windhoek-Not-Deprived, or Rest of Khomas (which included sparsely populated rural areas and two towns) (Figure 3).

2.3.3 Model input: Population. The model input population was simply the “true” simulated population aggregated from household point locations to constituencies (2nd-level administrative units). While Namibia is one of several countries that provides EA-level population counts to gridded

population modellers, a large number of LMICs (for example Nigeria, Cameroon, Chad, Iraq, Uzbekistan, Argentina, or India) do not share similarly disaggregated population counts [41]. In most LMICs, gridded population estimates are derived from coarse (and often outdated or inaccurate) census population counts, thus Namibia's constituency boundaries were used to mimic this common situation. The "true" reference population was used to derive input population counts in constituencies located in Khomas, while 2016 constituency UN population projections [11] were used for all other Namibian constituencies to (i) eliminate the effect of outdated or inaccurate census data on model outputs, and (ii) isolate the effect of the models used to generate gridded population estimates. As described above, WPG-Unconstrained model inputs reflected average population densities across the entire area of constituencies, while WPG-Constrained model inputs reflected population densities in the settled area only of constituencies as indicated in the Figure 3 upper-left panel with white shading [33].

2.3.4 Model input: Slum training data. We traced all 11 SDI slum settlement boundaries from the Know Your City (KYC) campaign website [42] and created four additional boundaries in ArcGIS in similar areas of small disorganised buildings to represent a total of 15 slum training communities (Figure 3). As with other model input data, the simulated "true" population was summed within slum boundaries to create the corresponding population counts and densities.

2.3.5 Model input: Predictive covariates. All 23 standard WorldPop-Global covariates were included in both the WPG-Unconstrained and -Constrained models, with zonal statistics generated by constituency (and slums, as applicable) [34] (Table 2). WPG-Constrained models additional included 10 covariates derived from Maxar/Esri building footprints in 100x100m grid cells [35], which is a defining characteristics of this modelling approach [33]. See appendix Figure A1 for a map of each covariate in Windhoek, and Figure A2 for importance scores for the most predictive covariates in each model.

Table 2. Summary of datasets used in the analysis

Name	Description (Year)	Original scale	Original source
Reference data			
pop	Simulated “true” population (2016)	Vector, <30 m	Thomson, et al. [3]
ea_deprived	Census Enumeration Areas classified as “deprived” or “not-deprived” in Windhoek, and “other” for the rest of Khomas	Vector, <30 m	Namibia Statistics Agency [40]
Input population data			
constituency	Constituency boundaries (2011) with total simulated “true” population	Vector, <30 m	Namibia Statistics Agency [40]
kyc	Modified Know Your City campaign slum boundaries (2016-18) with total simulated “true” population	Vector, <30 m	Slum Dwellers International. [42]
Input predictive covariates			
cciwat	Distance to ESA-CCI-LC inland waterbodies (2000-12)	4.5” (~150 m)	ESA CCI [43]
011	Distance to cultivated areas (2015)	9” (~300 m)	ESA CCI – LC [44]
040	Distance to woody areas (2015)	9” (~300 m)	ESA CCI – LC [44]
130	Distance to cultivated areas (2015)	9” (~300 m)	ESA CCI – LC [44]
140	Distance to herbaceous areas (2015)	9” (~300 m)	ESA CCI – LC [44]
150	Distance to sparse vegetation areas (2015)	9” (~300 m)	ESA CCI – LC [44]
160	Distance to aquatic vegetation areas (2015)	9” (~300 m)	ESA CCI – LC [44]
190	Distance to urban areas (2015)	9” (~300 m)	ESA CCI – LC [44]
200	Distance to bare areas (2015)	9” (~300 m)	ESA CCI – LC [44]
tt50k	Resampled travel time to cities of 50,000+ (2000)	30” (~900 m)	Weiss, et al. [45]
ghsl	Distance to urban area (2012)	1.26” (~38 m)	Pesaresi, et al. [46]
road	Distance to OSM major roads (2016)	Vector, <30 m	OpenStreetMap [47]
intsec	Distance to OSM major road intersections (2016)	Vector, <30 m	OpenStreetMap [47]
waterway	Distance to OSM major waterways (2016)	Vector, <30 m	OpenStreetMap [47]
slope	SRTM-based slope (2000)	3” (~90 m)	de Ferranti [48,49]
topo	SRTM-based elevation (2000)	3” (~90 m)	de Ferranti [48,49]
viirs	Resampled VIIRS night-time lights (2012-2016)	30” (~900 m)	NOAA [50]
wdpa	Distance to IUCN nature reserve (2000-17)	30” (~900 m)	UNEP-WCMS & IUCN [51]
bsgme	Distance to built settlement expansion (2016)	3” (~90 m)	Nieves, et al. [52]
prec	Average total annual precipitation (1970-2000)	30” (~900 m)	Fick and Hijmans [53]
coast	Distance to open-water coastline (2000-20)	3” (~90 m)	CIESIN [20]
b_count	Total building count (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_density	Total building density (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_cvarea	Coefficient of variance in building area (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_cvlength	Coefficient of variance in building edge length (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_imageyr	Year of satellite imagery (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_meanarea	Mean area of buildings (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_meanlength	Mean length of building edges (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_totarea	Total area of buildings (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_totlength	Total length of building edges (2020)	Vector, <30 m	Maxar/Esocopia [35]
b_urban	Classification of buildings as urban vs rural (2020)	Vector, <30 m	Maxar/Esocopia [35]

3. Results

The distribution of model input population densities varied widely between the WPG-Unconstrained (Model 1) and -Constrained (Model 3) due to massive differences in the total constituency area (Model 1) versus total constituency area classified as settled (Model 3). The presence of vast unsettled areas across Namibia meant that the majority of population densities calculated across full constituencies were below 1 person per hectare, with a handful of values registering up to 123 people per hectare (Figure 4). Conversely, because WPG-Constrained densities were calculated from settled areas only, all population densities were at least 2 people per hectare, with the bulk of estimates between 3 and 6 people per hectare, and a maximum of 134 people per hectare (Figure 4). The incorporation of slum training data into models extended the distributions of possible

population densities to a maximum of 177 people per hectare (Models 2 and 4), and provided a more complete range of values to the Random Forest model to assign to ~100x100m grid cells (Figure 4).

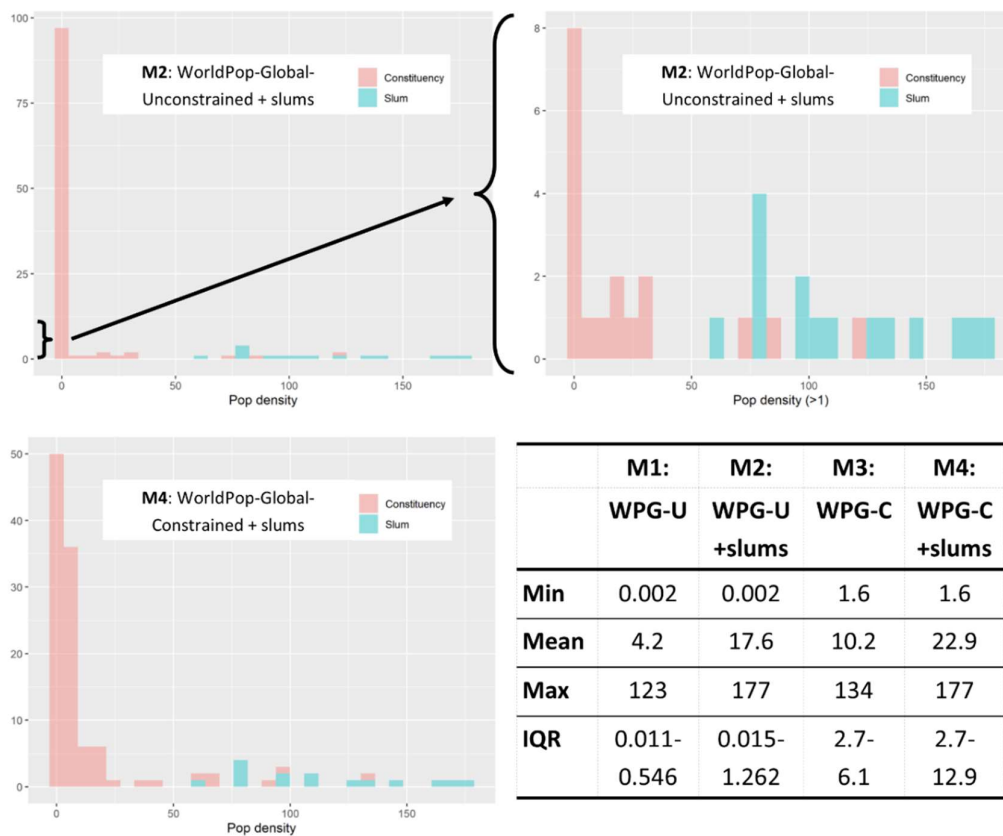


Figure 4. Distribution of input population density values for each model tested. *Constituencies have complete national coverage, while the sample of 15 slums are located in Windhoek only.*

The impact of unconstrained versus constrained population density inputs, and the addition of slum training data were both strongly apparent in the modelled outputs. Model 1 (WPG-Unconstrained) resulted in highly smoothed estimates across settled and unsettled areas that ranged from less than one person to 112 people in the most populous grid cell (Figure 5). Adding slum training data to the Unconstrained model (Model 2) slightly concentrated population estimates over Windhoek but did not increase the most populous grid cell value (Figure 5). Model 3 (WPG-Constrained) estimated zero population in grid cells with no buildings such that the estimated population was concentrated in settled grid cells, but still highly smoothed; the most populous grid cell was only estimated to have 121 people (Figure 5). The addition of slum training data to the WPG-Constrained model (Model 4) resulted in the most variable output and higher population counts, though the maximum estimate of 131 people in a grid cell was roughly a third of the “true” maximum grid cell population (386 people) (Figure 5). All models incorrectly attributed population to a large unpopulated industrial zone near the city centre (Figure 5).

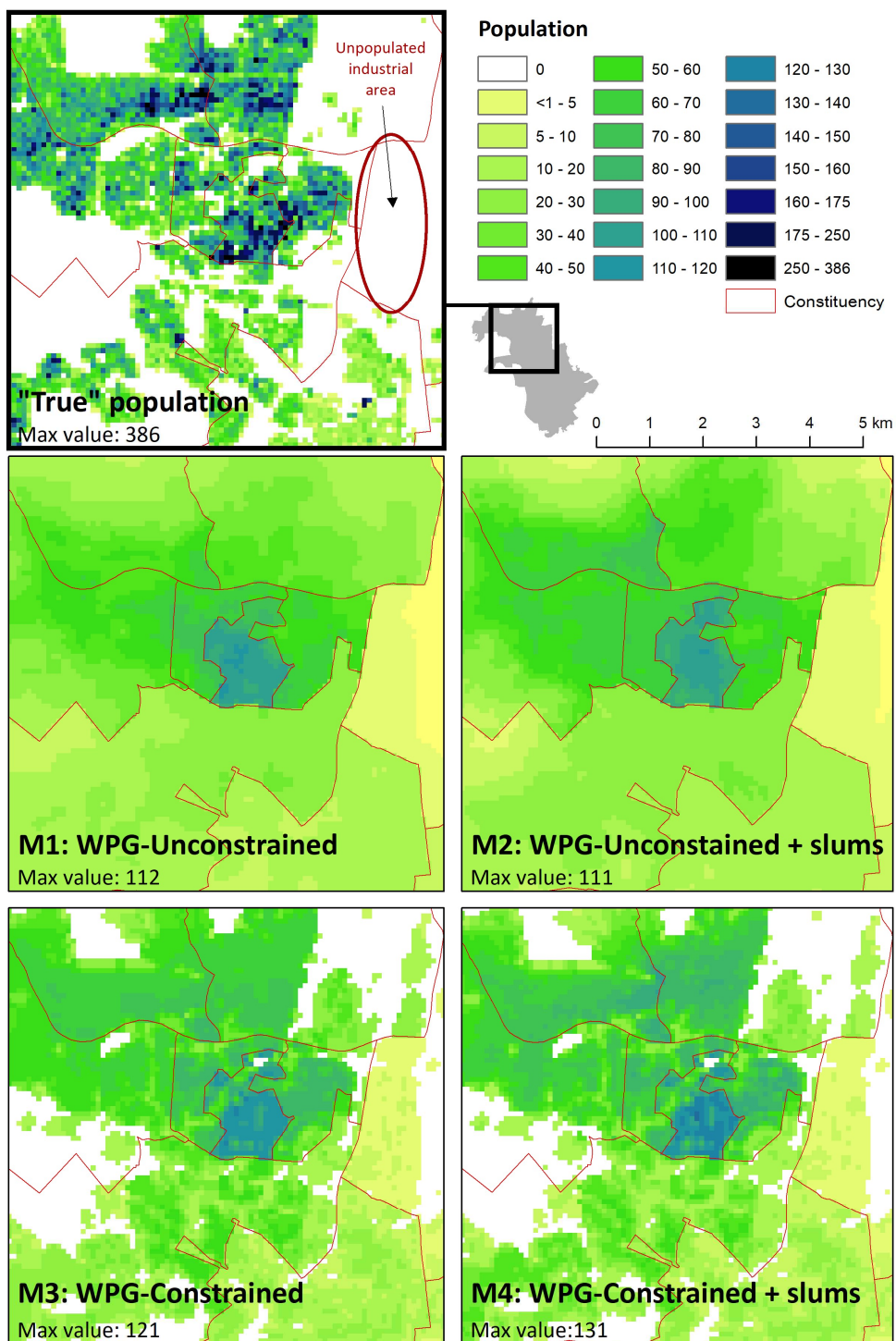


Figure 5. Visual comparison of model population estimates in a northern section of Windhoek, Namibia

Next, differences between each of the four estimates and the “true” population were calculated. Figure 6 shows the distribution of these differences in all grid cells with one or more people (omitting grid cells with <1 person), and visualised separately for Windhoek not-deprived, Windhoek deprived, and the rest of Khomas. Unsurprisingly, the differences in the rest of Khomas were far more variable in unconstrained models (Models 1 and 2) compared to constrained models (Models 3

and 4) because population was often estimated in unsettled grid cells in unconstrained models. Within Windhoek, the standard WPG-Unconstrained (Model 1) slightly underestimated the population in not-deprived areas (Bias: -3.8), and consistently and severely underestimated the population in deprived areas (Bias: -26.4) (Figure 6). Adding slum training data (Model 2) only marginally improved estimates in Windhoek's not-deprived (Bias: -2.2) and deprived areas (Bias: -22.7) (Figure 6). The WPG-Constrained model (Model 3) was more promising with reduced rates of error in Windhoek's deprived (RMSE: 37) and not-deprived (RMSE: 26) areas, and a better balance in bias across not-deprived (Bias: -0.9) and deprived (Bias: -6.3) areas (Figure 6). Again, adding slum training data (Model 4) had the effect of slightly improving the errors and bias in all areas (Figure 6). The marked improvement in models 3 and 4 compared to 1 and 2 were driven by constrained input population densities, and by building covariates, in particular building density which aligned with population density (see Appendix Figure A2).

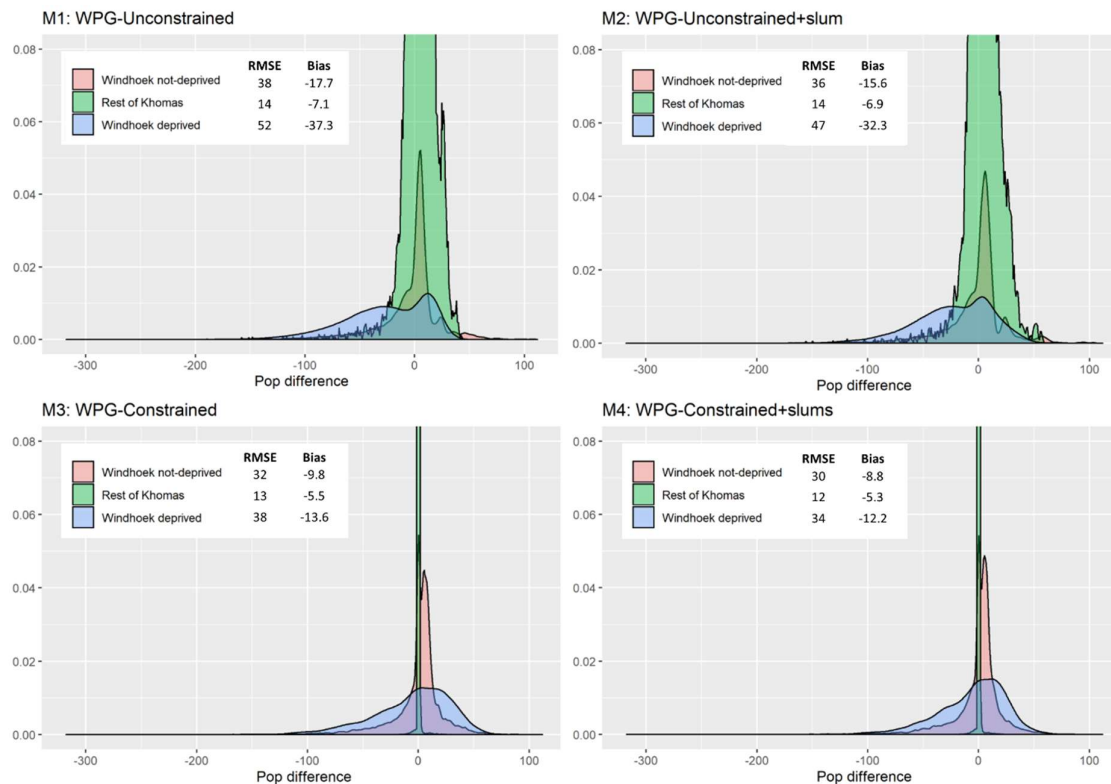


Figure 6. Histograms of differences between predicted and “true” population by cell in Windhoek, and unadjusted RSME and Bias statistics. Graphs and statistics represent only grid cells with 1+ estimated people, and are presented by settlement type. The tops of histograms are not shown for readability.

Given the sensitivity of these statistics to population size, Table 3 reports RMSE and Bias adjusted by average grid cell population to represent error and bias per person in each model, allowing for a fairer comparison across settlement types (i.e., deprived, not-deprived and rest of Khomas). We further report the percent of population inaccurately misallocated to “truly” unpopulated grid cells, and the percent of Windhoek’s population living in deprived areas (SDG 11.1.1). Findings are consistent with the unadjusted error and bias statistics reported in Figure 6, and Model 4 continued to offer the best balance of low error and bias across all settlement types (Table 3). For accuracy statistics across all cells, including cells in the rest of Khomas located in vast desert areas, see Appendix Table A1.

Table 3. Accuracy per person of WorldPop-Global gridded population models in Khomas, Namibia

Model	% pop Misalloc.	Cells estimated to have 1+ people		SDG 11.1.1 Windhoek ¹
		NRMSE	RBias	
M1: WPG-Unconstrained				
Windhoek not-deprived	31.0	73	-9.4	
Windhoek deprived	13.5	74	-41.6	19.2
Rest of Khomas	95.1	132	71.2	
M2: WPG-Unconstrained + slums				
Windhoek not-deprived	31.3	71	-5.4	
Windhoek deprived	10.6	67	-35.8	20.0
Rest of Khomas	93.1	144	69.8	
M3: WPG-Constrained				
Windhoek not-deprived	21.4	65	-2.1	
Windhoek deprived	10.7	58	-9.9	27.1
Rest of Khomas	77.4	126	25.7	
M4: WPG-Constrained + slums				
Windhoek not-deprived	19.3	61	-1.7	
Windhoek deprived	8.5	51	-9.6	27.0
Rest of Khomas	75.0	115	22.4	

¹ The “true” percent of population living in slums, informal settlements, or other deprived areas in Windhoek (SDG 11.1.1) according to the “true” simulated population is 29.5%.

In practical terms, the tendency of WPG-Unconstrained models (Models 1 and 2) to severely underestimate population in deprived areas meant that SDG 11.1.1 in Windhoek was grossly underestimated. According to the “true” simulated population in Windhoek, the actual percent of population living in slums, informal settlements, or inadequate housing (SDG 11.1.1) was 29.5% in 2016, though outputs from Model 1 put this figure at just 19.2%, and Model 2 at 20.0% (Table 1). Model 3 (27.1%) and Model 4 (27.0%) produced reasonably accurate estimates and could, therefore, plausibly be used to measure SDG 11.1.1 in cities with maps of deprived versus not-deprived areas (Table 1). For a map of model differences from the “true” simulated population in deprived versus not-deprived areas of Windhoek, see Appendix Figure A3.

4. Discussion

Using a simulated “true” population in a real-world location, grid cell-level accuracy in four gridded population models were compared across three settlement types (urban not-deprived, urban deprived, and rural). Findings show that the recently released WPG-Constrained model, which includes 10 building footprint auxiliary datasets derived from Maxar/ECOPIA, performed well on its own across diverse urban and rural settings, and even better with a few slum training locations. In real-world settings, SDI and similar data could be used to identify slum boundaries and realistic population counts. The WPG-Constrained model, however, was imperfect with slight underestimation of urban deprived populations and slight overestimation of rural populations at the grid cell-level. While users should always remain cautious about interpreting population counts in small areas due to the accumulation of model errors and uncertainty at finer scales, our reasonably accurate SDG 11.1.1 results inspire confidence in the WPG-Constrained model with slum training data in terms of the *relative* distribution of population within urban and rural areas. For example, we would feel comfortable using these model outputs as a household survey sample frame (stratified by urban/rural) [23] or for SDG 11 reporting (within urban areas).

These findings add nuance to previous findings that gridded population datasets tend to vastly underestimate populations in slums [7]. This analysis differed from the previous study by controlling for the accuracy of the input population datasets, using simulated “true” population counts rather

than real-world census data which contain unquantifiable inaccuracies. Our results suggest that the addition of building auxiliary data and slum training data can sufficiently improve modelled estimates in urban and rural areas for SDG reporting, if the input population data are relatively accurate. It is worth reiterating that the input population data in the four models were aggregated to constituencies (2nd administrative level), which bodes well for the applicability of these findings across LMIC settings where published census data are often highly aggregated. While countries with the greatest need for gridded population estimates are likely to have the most outdated or inaccurate census data, we can be confident that a gridded population model that includes building auxiliary data and slum training data produces the most accurate, disaggregated top-down estimates possible.

We posit two additional datasets could improve the accuracy of top-down gridded population data. First, inclusion of settlement classification layer that accurately distinguishes urban deprived from not-deprived areas. We are not aware of any such global datasets at fine geographic scale that could be used for this purpose, yet; however, the IDEAMAPS Network is working toward developing such a layer [54], and several algorithms have been published and tested in small areas that could, in theory, be scaled across cities. Jochem and Tatem (2021), for example, published an R package that uses building footprints to classify settlement types, though it has only been tested in Europe and imperfectly distinguishes between urban settlements types [55]. The World Resources Institute released Python code to distinguish urban land use types, including “residential informal” land use, from Sentinel-II imagery and demonstrated its application in India and Mexico, though substantial bespoke training data were required [56,57]. If, and when, a routine accurate map of deprived urban areas becomes available across cities, population distribution modellers might consider dividing a country by settlement type before modelling (e.g. urban-deprived, urban-not-deprived, rural), given differences in population densities across these settlements types.

A contributor to error in Windhoek’s not-deprived areas was the estimation of people in grid cells containing non-residential industrial buildings. Across all four models, hundreds of non-residential cells in an industrial zone near downtown Windhoek were allocated dozens of people each (see figure 5). Among the building covariates included in the WPG-Constrained model, not a single one distinguished between residential and non-residential buildings, though the algorithms and technologies to extract such information are being developed [58]. Inclusion of a building residential/non-residential dataset would likely improve estimates, especially in urban not-deprived areas, by further constraining settled *residential* areas. In addition to exclusion of industrial buildings, this type of dataset should exclude airports, government buildings, universities, hospitals, military bases, and commercial centres, all of which contribute to inaccurately allocate population in gridded population model outputs, creating practical challenges for data users in the field [59], and potentially skewing development indicators in urban areas [30].

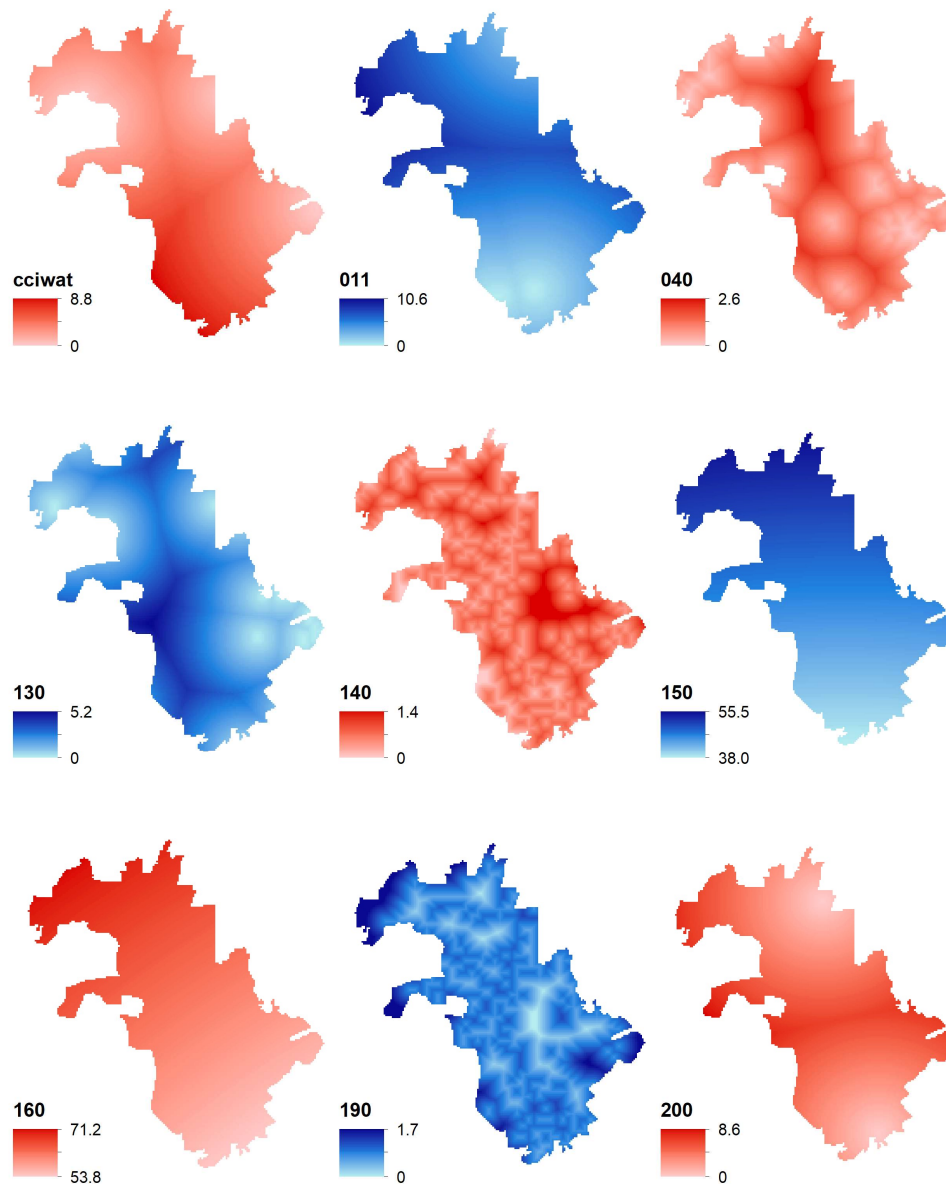
These findings are particularly relevant for understanding the accuracy of top-down gridded population estimates in grid cells and small-area in LMIC cities at a recent point in time. Numerous types of users will be interested in findings at this fine scale in LMIC contexts, including governments and organisations involved with planning and development. We specifically focus on users involved with monitoring of SDG 11.1.1; however, these results are applicable to other SDG and development indicators calculated within LMIC cities. Other gridded population datasets might be better suited to other types of use cases; for example, GHS-POP is appropriate for longitudinal analyses at more aggregated scales [17], LandScan is often appropriate for day-time regional-scale disaster response when people are likely to be away from home [10], and GPWv4.11 is recommended as an input to climate and other global research models [21]. We echo the findings from other studies that indicate that the different gridded population datasets are better suited for different use cases, and users need to be informed about which model outputs are most relevant for their aims [6,60].

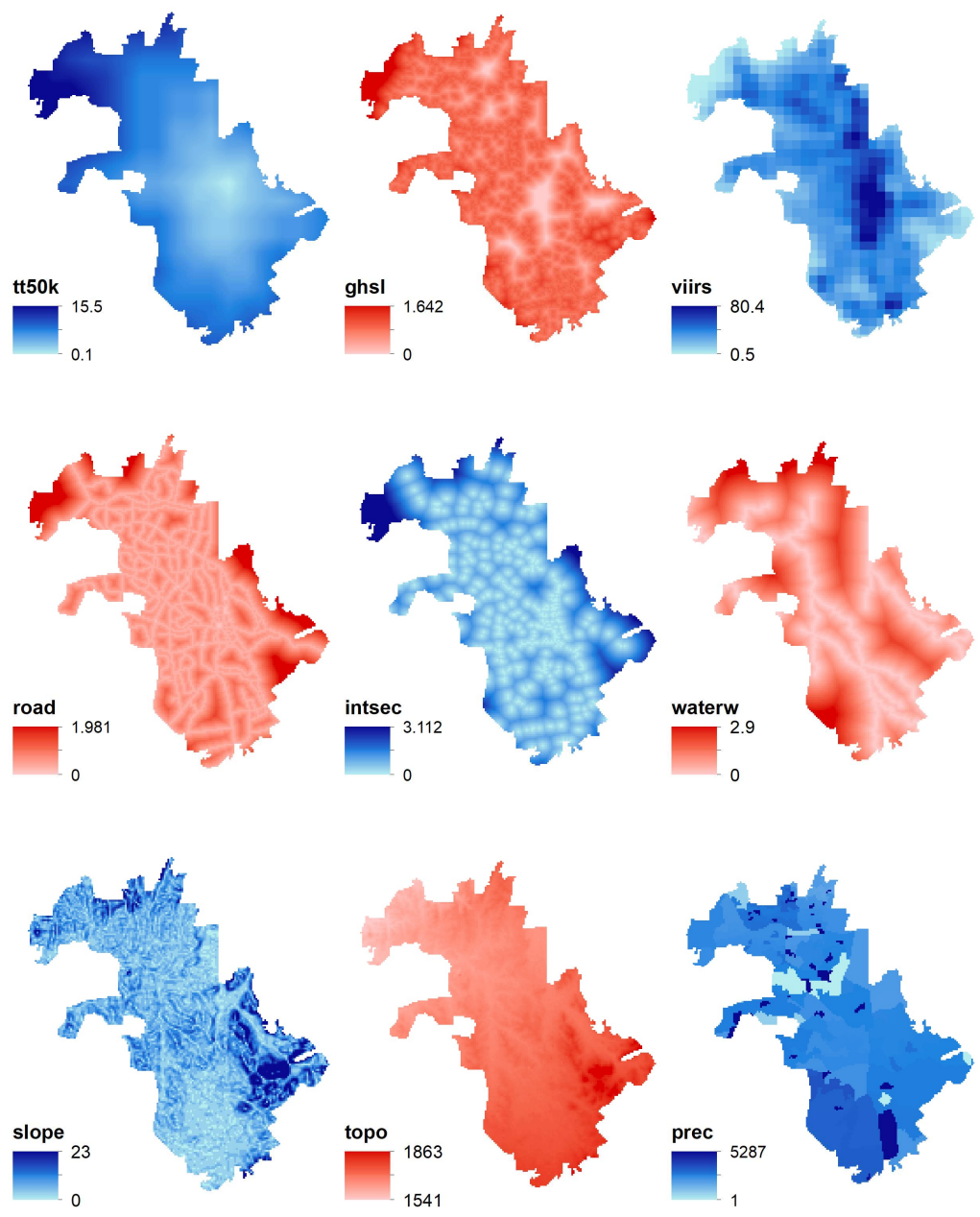
5. Conclusion

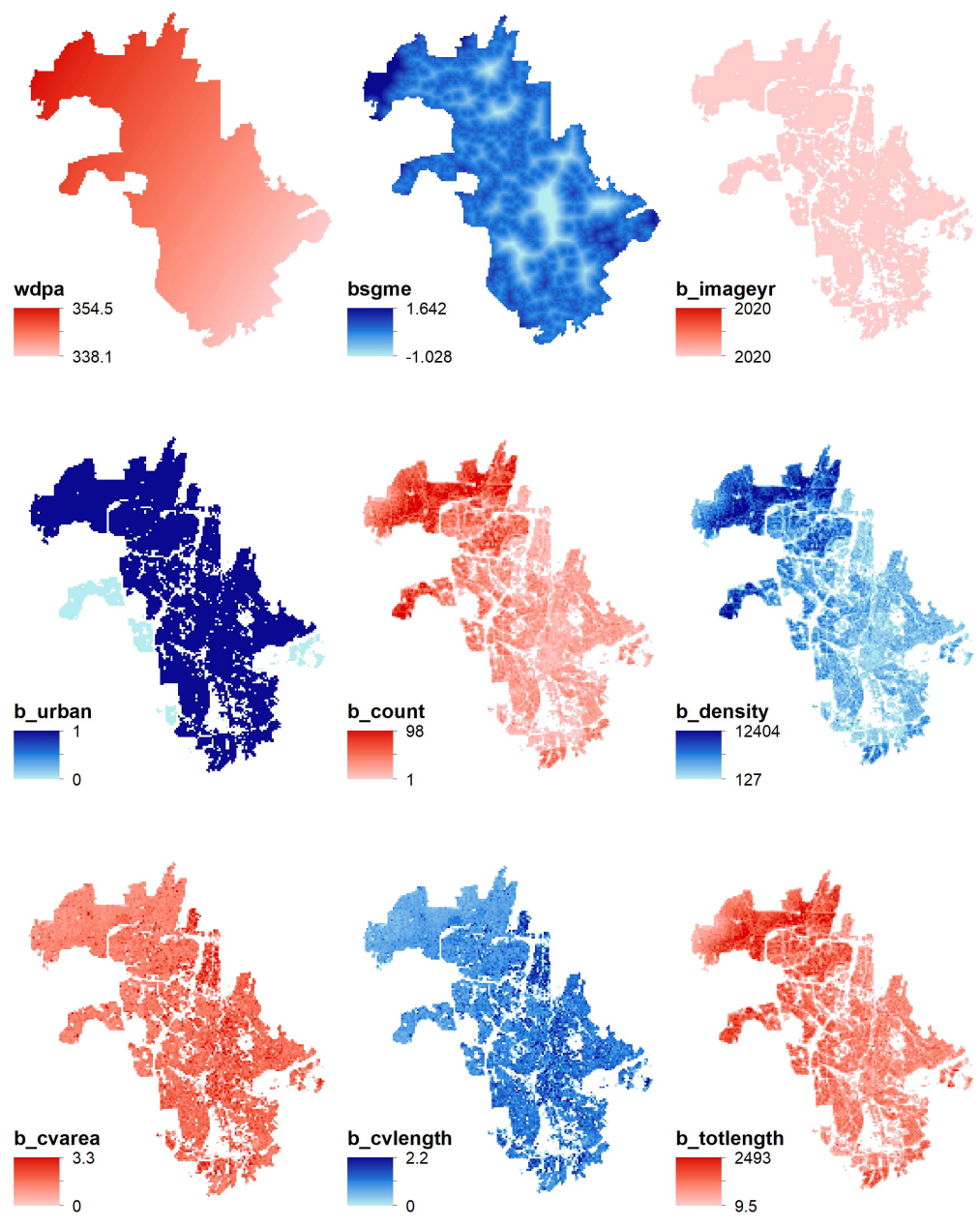
Using slum training data in top-down gridded population modelling improves urban population estimates, particularly in deprived areas, which are systematically underrepresented in LMIC population data (including censuses, surveys, and gridded population estimates). The WorldPop-Global-Constrained models with, and without, slum training data produced fairly accurate SDG 11.1.1 estimates for Windhoek, though the inclusion of slum training data reduced errors and bias across urban slum, urban non-slum, and rural areas. This analysis puts us closer to accurately estimating SDG indicators from gridded population data, but additional model improvements are still needed, such as inclusion of slum training data (e.g., from SDI community profiles), deprived area boundaries, and building typology information.

Appendix

Figure A1. Distribution of all spatial covariates used in Random Forest models in Windhoek, Namibia







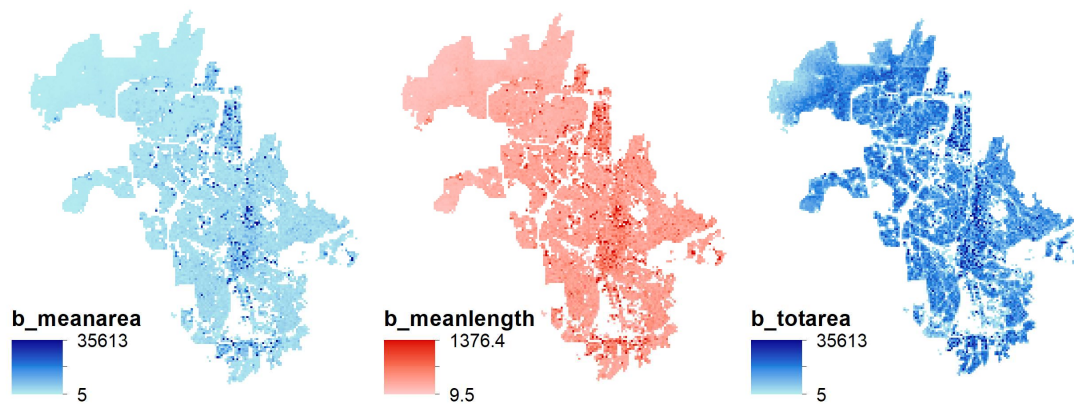


Figure A2. Importance Scores (national constituency inputs + 15 Windhoek slums)
 Across all WPG models, VIIRS night-time lights intensity was a strong predictor of population density. In the WPG-Constrained models (M3 and M4), building counts were, by far, the most predictive of population density.

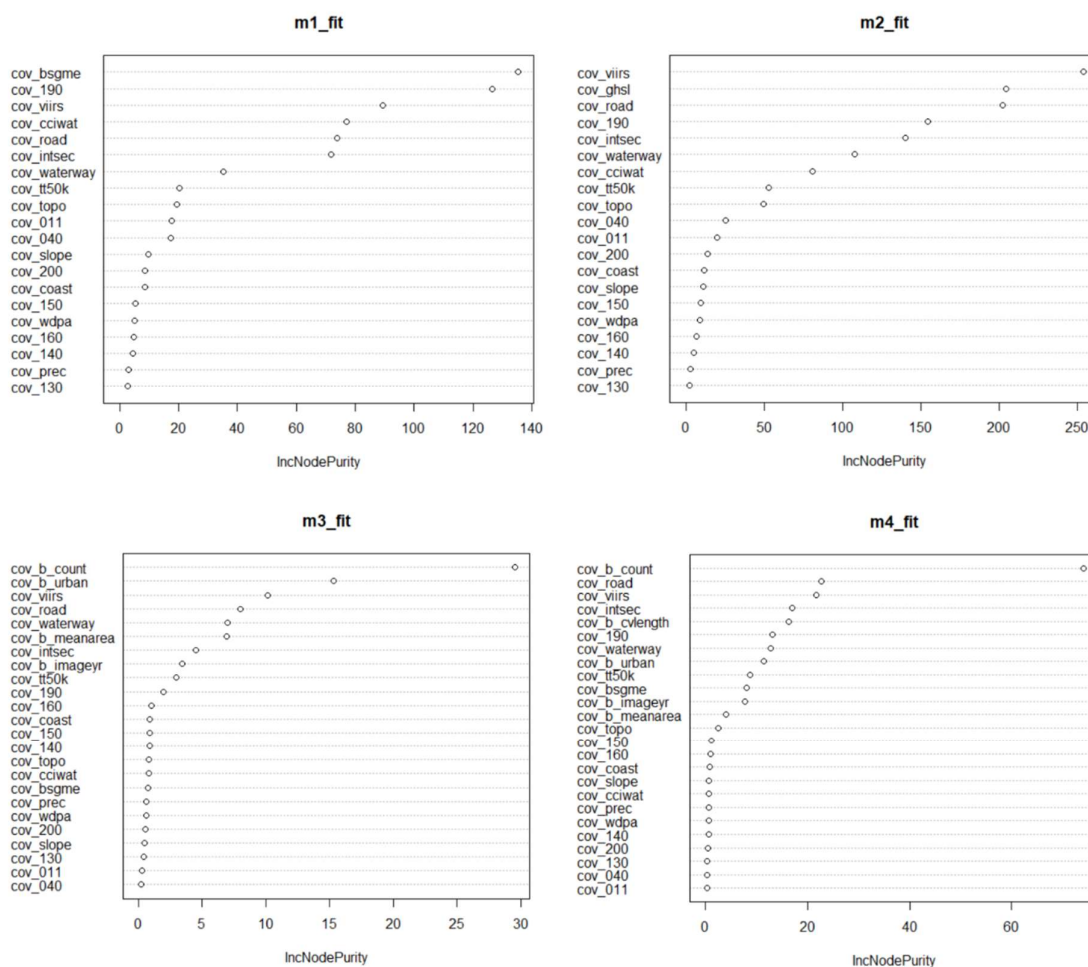
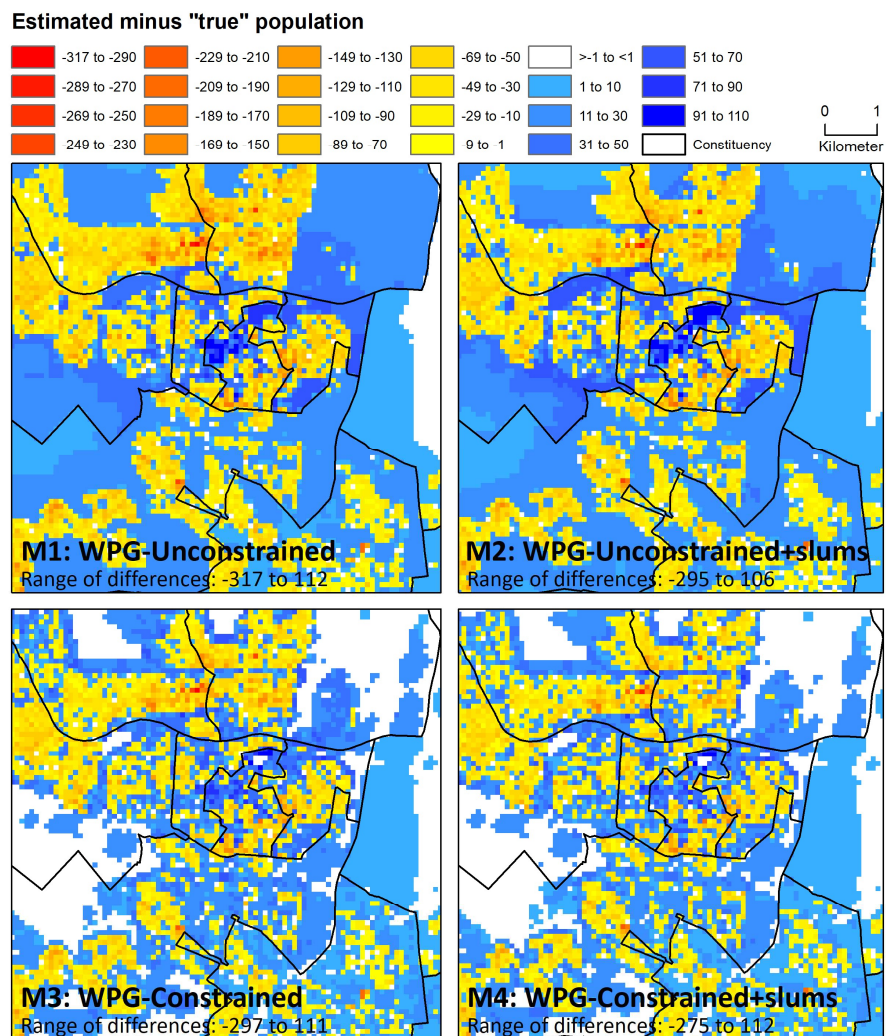


Table A1. Accuracy per person of WorldPop-Global gridded population models in Khomas, Namibia

Model	All cells	
	NRMSE	RBias
M1: WPG-Unconstrained		
Windhoek Not-deprived	57	-5.7
Windhoek Deprived	34	-8.9
Rest of Khomas	8	0.2
M2: WPG-Unconstrained + slums		
Windhoek Not-deprived	56	-3.3
Windhoek Deprived	31	-7.6
Rest of Khomas	7	0.2
M3: WPG-Constrained		
Windhoek Not-deprived	48	-1.1
Windhoek Deprived	26	-2.0
Rest of Khomas	6	0
M4: WPG-Constrained + slums		
Windhoek Not-deprived	45	-0.9
Windhoek Deprived	23	-2.0
Rest of Khomas	5	0

Figure A3. Visual comparison of modelled population over/under-estimates in a northern section of Windhoek, Namibia



References

1. Pradhan P. Antagonists to meeting the 2030 Agenda. *Nat Sustain.* 2019;2: 171–172. doi:10.1038/s41893-019-0248-8
2. United Nations Department of Economic and Social Affairs (UN-DESA). Sustainable Development Goals. In: Sustain Development Knowledge Platform. 2018 [cited 13 Jun 2021]. Available: <https://sustainabledevelopment.un.org/sdgs>
3. [dataset] Thomson DR, Leasure DR, Bird TJ, Tzavidis N, Tatem AJ. How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia. *Preprints.* 2021; 1–38. doi:10.20944/preprints202102.0492.v2
4. Carr-Hill R. Missing millions and measuring development progress. *World Dev.* 2013;46: 30–44. doi:10.1016/j.worlddev.2012.12.017
5. UN Department of Economic and Social Affairs (UN-DESA). World Urbanization Prospects: The 2018 Revision. 2018 [cited 13 Jun 2021]. Available: <https://population.un.org/wup/DataQuery/>
6. 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: 1385–1409. doi:10.5194/essd-2019-82
7. Thomson DR, Gaughan AE, Stevens FR, Yetman G, Elias P, Chen R. Evaluating the accuracy of gridded population estimates in slums: A case study in Nigeria and Kenya. *Urban Sci.* 2021;5: 48. doi:10.3390/urbansci5020048
8. Qiu Y, Zhao X, Fan D, Li S. Geospatial disaggregation of population data in supporting SDG assessments: A case study from Deqing County, China. *ISPRS Int J Geo-Information.* 2019;8: 356. doi:10.3390/ijgi8080356
9. [dataset] Oak Ridge National Laboratories. LandScan Data Availability. 2017 [cited 2 Feb 2017]. Available: www.ornl.gov/sci/landscan/landscan_data_avail.shtml
10. Dobson JE, Bright 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;66: 849–857.
11. [dataset] WorldPop. Population Counts 2000-2020 UN-Adjusted Unconstrained 100m. 2020 [cited 13 Jun 2021]. doi:10.5258/SOTON/WP00660
12. 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: e0107042. doi:10.1371/journal.pone.0107042
13. [dataset] Bondarenko M, Kerr D, Sorichetta A, Tatem AJ. Population Counts 2000-2020 UN-Adjusted Constrained 100m. In: WorldPop, University of Southampton, UK. 2020 [cited 13 Jun 2021]. doi:10.5258/SOTON/WP00685
14. [dataset] Facebook Connectivity Lab, CIESIN—Columbia University. High Resolution Settlement Layer (HRSL). 2016 [cited 13 Jun 2021]. Available: <https://data.humdata.org/dataset/highresolutionpopulationdensitymaps>
15. [dataset] CIESIN, UNFPA, WorldPop, Flowminder. Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). 2018 [cited 21 Jun 2021]. Available: www.grid3.org
16. 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: 24173–24179. doi:10.1073/pnas.1913050117
17. Pesaresi M, Ehrlich D, Florczyk AJ, Freire S, Julea A, Kemper T, et al. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. *Ispra Italy: European Commission Joint Research Centre;* 2016. doi:10.2788/253582

18. [dataset] European Commission Joint Research Centre (EC-JRC). Global human settlement population model (GHS-POP). 2020 [cited 13 Jun 2021]. Available: <https://ghsl.jrc.ec.europa.eu/data.php>
19. Frye C, Nordstrand E, Wright DJ, Terborgh C, Foust J. Using classified and unclassified land cover data to estimate the footprint of human settlement. *Data Sci J*. 2018;17: 1–12. doi:10.5334/dsj-2018-020
20. [dataset] Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4.11 (GPWv4.11). 2018 [cited 13 Jun 2021]. Available: <https://doi.org/10.7927/H4F47M65>
21. Doxsey-Whitfield E, MacManus K, Adamo SB, Pistolessi L, Squires J, Borkovska O, et al. Taking advantage of the improved availability of census data: A first look at the Gridded Population of the World, Version 4. *Pap Appl Geogr*. 2015;1: 226–234. doi:10.1080/23754931.2015.1014272
22. 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. 2020 [cited 13 Jun 2021]. doi:10.5258/SOTON/WP00681
23. Thomson DR, Rhoda DA, Tatem AJ, Castro MC. Gridded population survey sampling: A systematic scoping review of the field and strategic research agenda. *Int J Health Geogr*. 2020;19: 34. doi:10.1186/s12942-020-00230-4
24. Thorndike E. On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. *Am J Psychol*. 1939;52: 122–4. doi:10.2307/1416673
25. 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: 255–277. doi:10.1007/s11111-020-00360-8
26. Calka B, Bielecka E. GHS-POP accuracy assessment: Poland and Portugal case study. *Remote Sens*. 2020;12: 1105. doi:10.3390/rs12071105
27. Hay S, Noor A, Nelson A, Tatem A. The accuracy of human population maps for public health application. *Trop Med Int Heal*. 2005;10: 1073–1086. doi:10.1111/j.1365-3156.2005.01487.x
28. Bharti N, Djibo A, Tatem AJ, Grenfell BT, Ferrari MJ. Measuring populations to improve vaccination coverage. *Sci Rep*. 2016;6: 34541. doi:10.1038/srep34541
29. Wigley AS, Tejedor-Garavito N, Alegana V, Carioli A, Ruktanonchai CW, Pezzulo C, et al. Measuring the availability and geographical accessibility of maternal health services across sub-Saharan Africa. *BMC Med*. 2020;18: 1–10. doi:10.1186/s12916-020-01707-6
30. Cutts FT, Dansereau E, Ferrari MJ, Hanson M, McCarthy KA, Metcalf CJE, et al. Using models to shape measles control and elimination strategies in low- and middle-income countries: A review of recent applications. *Vaccine*. 2020;38: 979–992. doi:10.1016/j.vaccine.2019.11.020
31. Newaya TP. Rapid urbanization and its influence on the growth of informal settlements in Windhoek, Namibia [dissertation]. Cape Peninsula University of Technology. 2010. Available: <http://digitalknowledge.cput.ac.za/xmlui/handle/11189/340>
32. NSA. Namibia Population and Housing Census 2011 main report. Windhoek Namibia; 2011. Available: [www.nsa.org.na/files/downloads/Namibia 2011 Population and Housing Census Main Report.pdf](http://www.nsa.org.na/files/downloads/Namibia%202011%20Population%20and%20Housing%20Census%20Main%20Report.pdf)
33. WorldPop. Top-down estimation modelling: Constrained vs unconstrained. 2020 [cited 13 Jun 2021]. Available: www.worldpop.org/methods/top_down_constrained_vs_unconstrained
34. Lloyd CT, Chamberlain H, Kerr D, Yetman G, Pistolessi L, Stevens FR, et al. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data*. 2019;3: 108–139. doi:10.1080/20964471.2019.1625151
35. [dataset] Dooley CA, Leasure DR, Boo G, Tatem AJ. Gridded maps of building patterns throughout sub-Saharan Africa, version 2.0. In: WorldPop, University of Southampton, UK. 2021 [cited 13 Jun 2021]. doi:10.5258/SOTON/WP00712

36. Sorichetta A, Hornby GM, Stevens FR, Gaughan AE, Linard C, Tatem AJ. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci Data*. 2015;2: 150045. doi:10.1038/sdata.2015.45
37. [dataset] Namibia Statistics Agency (NSA). Namibia 2011 Population and Housing Census [PUMS dataset] version 1.0. Windhoek Namibia; 2013. Available: <https://nsa.org.na/microdata1/index.php/catalog/19>
38. Thomson DR, Kools L, Jochem WC. Linking synthetic populations to household geolocations: A demonstration in Namibia. *Data*. 2018;3: 30. doi:10.3390/data3030030
39. Namibia Statistics Agency (NSA). Namibia Population Projections 2011-2041. Windhoek Namibia; 2014. Available: <https://cms.my.na/assets/documents/p19dn4fhgp14t5ns24g4p6r1c401.pdf>
40. [dataset] NSA. 2011 Census EA boundaries. In: Digital Namibia. 2011 [cited 13 Jan 2020]. Available: <https://nsa.org.na/page/gis-data-requests/>
41. Center for International Earth Science Information Network - CIESIN - Columbia University. Country-level Information and Sources Revision 11. 2018 [cited 17 May 2021]. Available: <https://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-country-level-summary-rev11.xlsx>
42. [dataset] Slum/Shack Dwellers International (SDI). Know Your City. 2016 [cited 13 Jun 2021]. Available: <http://knowyourcity.info/explore-our-data/>
43. [dataset] European Space Agency - Climate Change Initiative (ESA-CCI). Land cover CCI product - MERIS Waterbody product v4.0 (150 m). 2017 [cited 13 Jan 2020]. Available: <http://maps.elie.ucl.ac.be/CCI/viewer/>
44. [dataset] European Space Agency - Climate Change Initiative (ESA-CCI). Land Cover CCI Product - Annual LC maps from 2000 to 2015 (v2.0.7). 2017 [cited 13 Jan 2020]. Available: <http://maps.elie.ucl.ac.be/CCI/viewer/>
45. [dataset] 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: 333–6. doi:10.1038/nature25181
46. [dataset] European Commission. Global human settlement city model (GHS-SMOD). 2017 [cited 13 Jan 2020]. Available: <http://ghsl.jrc.ec.europa.eu/faq.php>
47. [dataset] OpenStreetMap contributors. OpenStreetMap base data. 2000 [cited 13 Jan 2020]. Available: www.openstreetmap.org
48. [dataset] de Ferranti J. Digital elevation data - Viewfinder panoramas. 2017 [cited 13 Jan 2020]. Available: www.viewfinderpanoramas.org/dem3.html
49. [dataset] de Ferranti J. Digital elevation data: SRTM void fill - Viewfinder panoramas. 2017 [cited 13 Jan 2020]. Available: www.viewfinderpanoramas.org/voidfill.html
50. [dataset] [USA] National Oceanic and Atmospheric Administration (NOAA). VIIRS nighttime lights. 2012 [cited 13 Jan 2020]. Available: https://maps.ngdc.noaa.gov/viewers/VIIRS_DNB_nighttime_imagery/index.html
51. [dataset] UNEP-WCMS, IUCN. World database on protected areas & Global database on protected areas management effectiveness. In: UN Environment Programme-World Conservation Monitoring Centre & International Union for Conservation of Nature. 2016 [cited 13 Jan 2020]. Available: www.protectedplanet.net/
52. [dataset] Nieves JJ, Sorichetta A, Linard C, Bondarenko M, Steele JE, Stevens FR, et al. Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night. *Comput Environ Urban Syst*. 2020;80: 101444. doi:10.1016/j.compenvurbsys.2019.101444
53. [dataset] Fick SE, Hijmans RJ. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol*. 2017;37: 4302–4315. doi:10.1002/joc.5086

54. Thomson DR, Kuffer M, Boo G, Hati B, Grippa T, Elsey H, et al. Need for an integrated deprived area “slum” mapping system (IDEAMAPS) in low- and middle-income countries (LMICs). *Soc Sci.* 2020;9: 80. doi:10.3390/socsci9050080
55. Jochem WC, Tatem AJ. Tools for mapping multi-scale settlement patterns of building footprints: An introduction to the R package foot. *PLoS One.* 2021;16: e0247535. doi:10.1371/journal.pone.0247535
56. Kerins P. WRI: Urban Land Use. In: GitHub [Internet]. 2020 [cited 18 May 2021]. Available: <https://github.com/wri/UrbanLandUse/releases/tag/v0.2.0>
57. Kerins P, Guzder-Williams B, Mackres E, Rashid T, Pietraszkiewicz E. Mapping Urban Land Use in India and Mexico using Remote Sensing and Machine Learning. Washington DC USA; 2021. doi:10.46830/writn.20.00048
58. Lloyd CT, Sturrock HJW, Leasure DR, Jochem WC, Lázár AN, Tatem AJ. Using GIS and machine learning to classify residential status of urban buildings in low and middle income settings. *Remote Sens.* 2020;12: 1–20. doi:10.3390/rs12233847
59. Thomson DR, 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.* 2021;98: 111–129. doi:10.1007/s11524-020-00485-z
60. POPGRID Data Collaborative. Leaving no one off the map: A guide for gridded population data for sustainable development. New York NY USA; 2020. Available: www.popgrid.org/sites/default/files/documents/popgrid-executive-summary-final.pdf