

Article

Linking synthetic populations to household geolocations: a demonstration in Namibia

Dana R Thomson ^{1,2,3,*,†}, Lieke Kools ^{4,†} and Warren C Jochem ^{1,2}

¹ Flowminder Foundation, Stockholm Sweden
² WorldPop, Department of Geography and Environment, University of Southampton, Southampton UK
³ Department of Social Statistics, University of Southampton, Southampton UK
⁴ Department of Economics, Leiden University, Leiden The Netherlands; l.kools@law.leidenuniv.nl
[†] These authors contributed equally
^{*} Correspondence: dana.thomson@flowminder.org; Tel.: +44 238 202 6000

Abstract: Whether evaluating gridded population dataset estimates (e.g. WorldPop, LandScan) or household survey sample designs, a population census linked to residential locations are needed. Geolocated census microdata data, however, are almost never available and are thus best simulated. In this paper, we simulate a close-to-reality population of individuals nested in households geolocated to realistic building locations. Using the R simPop package and ArcGIS, multiple realizations of a geolocated synthetic population are derived from the Namibia 2011 census 20% microdata sample, Namibia census enumeration area boundaries, Namibia 2013 Demographic and Health Survey (DHS), and dozens of publicly available spatial datasets. Realistic household latitude-longitude coordinates are manually generated based on public satellite imagery. Simulated households are linked to latitude-longitude coordinates by identifying distinct household types with multivariate *kmeans* analysis, and modelling a probability surface for each household type using Random Forest machine learning methods. We simulate five realizations of a synthetic population in Namibia's Oshikoto region, including demographic, socioeconomic and outcome characteristics at the level of household, woman, and child. Comparison of variables in the synthetic population were made with 2011 census 20% sample and 2013 DHS data by primary sampling unit/enumeration area. We found that synthetic population variable distributions matched observed observations and followed expected spatial patterns. We outline a novel process to simulate a close-to-reality microdata census geolocated to realistic building locations in a low- or middle-income country setting to support spatial demographic research and survey methodological development while avoiding disclosure risk of individuals.

Dataset: Supplement 1

Dataset License: CC-BY-4.0

Keywords: simulation; census; simpop; LMIC

1. Introduction

The ideal resource to evaluate the accuracy of gridded population datasets and certain household survey methodologies would be a complete set of individual records from a population linked to location of residence, though this is generally not available. Gridded population datasets model counts of human population in small grid cells, often based on census data and spatial covariates such as land cover type [1–4]. Various gridded population datasets have evaluated accuracy of population counts at the geographic scale of input census data [3–5], and other analyses have evaluated whether cells were accurately classified as populated or not populated [6], however accuracy of population count per grid cell has not been evaluated because it requires a geo-located microdata census (thus negating the need for a population model). In the realm of household surveys, evaluation of sample variability, measurement error, and missing values due to sample design requires a close-to-reality census of microdata to perform statistical simulations of repeated samples of households [7].

Although microdata are commonly made publicly available as census samples [8] or household survey samples [9], full census microdata are almost never publicly released to protect the anonymity of respondents. A more realistic option for researchers to obtain a dataset of all household observations and associated characteristics in a population, is to simulate it, and recent advances in generating synthetic populations have made this approach a viable alternative [10]. Synthetic population datasets also have the advantage over actual census data that multiple scenarios can be generated to test outcomes in potential future populations.

Previous work to simulate or reconstruct synthetic human populations has explored multiple methods. Most commonly, small area estimates of populations and socio-demographic characteristics are created by expanding or reweighting observations from a survey of individuals to meet totals and marginal distributions in more aggregated areal units. Iterative proportional fitting (IPF) is often used to incrementally improve the fit of a joint probability distribution of person- or household-level attributes (e.g. from a household survey) subject to known joint probabilities of attributes (e.g. from an aggregated census) [11,12]. Combinatorial optimisation procedures such as simulated annealing (SA) [13] or quota sampling [14] can also be used to prevent sub-optimal combinations of attributes in the simulated dataset. Templ and colleagues discuss a model-based approach to simulation of individual or household attributes with regression models, which they implement in an open-source software [15]. Agent-based models (ABMs) can also produce a realistic count of individuals, or “agents”, along with key attributes and relationships [16,17]. Some ABMs have also incorporated space into agent interactions, or produce outputs allocated to semi-realistic spaces such as a city [18].

Despite the advances in simulation methods, a lack of geographic specificity is a problem to most previous studies. The simulated populations are often only allocated to small output areas, such as census enumeration areas (EAs). While small area units are sufficient for many studies, they do not allow for local-scale analyses of health, education, and demographics. Some attempts have been made to associate simulated households to random points in space or along roads [19,20]. There is a growing demand for such spatially-disaggregated population datasets, particularly in low- and middle-income countries (LMIC) to plan projects and monitor progress toward the Sustainable Development Goals [21] which has led to novel techniques for producing gridded populations [3], [22] and other high spatial-resolution maps of sociodemographic characteristics interpolated from cluster survey locations [23–25]. However, it is difficult to assess the accuracy of these techniques in the absence of reliable population data at an equally fine spatial resolution.

The aim of this paper is to simulate a close-to-reality population of individuals nested within households and then to geo-locate this synthetic population to realistic building locations in a LMIC context. Our approach uses two commonly available population datasets (a census microdataset and a household survey) as well as openly available geospatial data to enable replication in other areas. This work was motivated by a need for a population dataset that could be used to develop and evaluate household survey methodologies in general, and gridded population survey methodologies in particular (e.g. GridSample [26]), though georeferenced population datasets will be useful for many applications. The synthetic population has to be located in both a real-world context to take

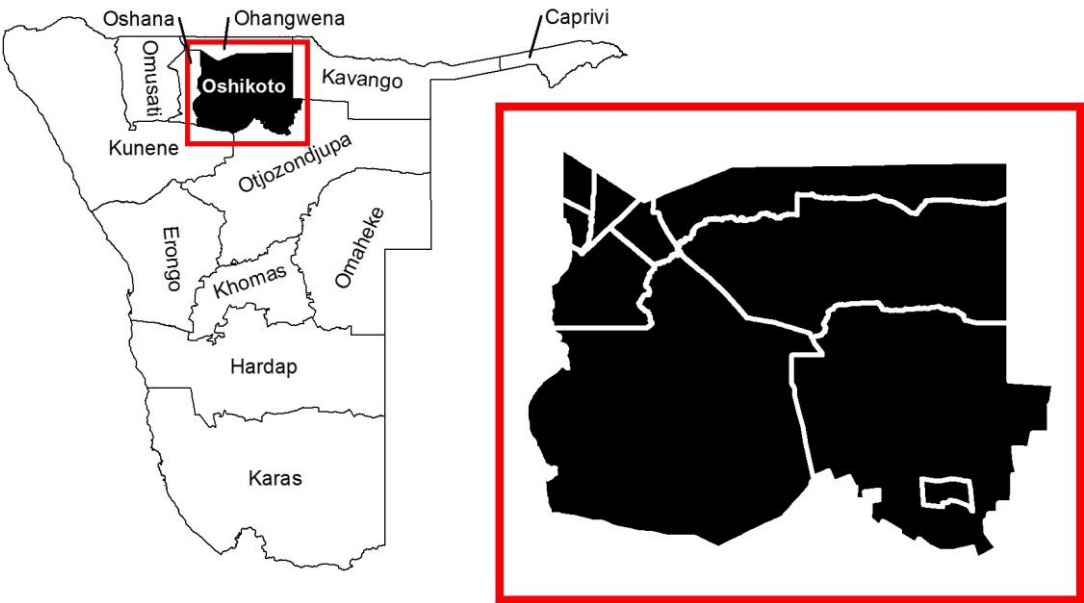
advantage of the realistic spatial covariates used in gridded population modelling, and at or below the same geographic scale as the gridded population data (~100 metre by 100 metre grid cells). The use of realistic, rather than randomly generated, latitude-longitude coordinates to represent home locations, however, raises new ethical questions for population simulations. We discuss how we approached these issues while openly releasing the code and simulated datasets from our case study in Namibia.

2. Methods

2.1. Setting

Namibia is selected for the simulation because population varies widely from low-to-high density, and the 2011 Namibia census meets the UN recommendations for high quality census data [27]. We select Oshikoto, one of Namibia’s 13 regions in northern Namibia to demonstrate the simulation methods discussed here because it presents a rich microcosm of conditions and population types (Figure 1). Oshikoto covers an area of 38,653 square kilometres, and is home to roughly 182,000 people [28]. The region has an unpopulated desert in the southwest, rural settled agriculture area in the north, rural area comprised mostly of a nomadic population in the southeast, and two cities comprised of planned and unplanned neighbourhoods. Oshikoto is comprised of 10 administrative sub-regions called constituencies, for which there are published census population and household totals.

Figure 1. Map of Oshikoto Region, Namibia and Oshikoto’s 10 constituency boundaries



2.2. Data

All input data are publicly available including the 20% microdata sample from the 2011 Namibia Population and Housing Census, available by request from the Namibia NSA [29]; 2011 Namibia census enumeration area boundaries, provided by request from the Namibia NSA [30]; 2013 Namibia Demographic and Health Survey (DHS) recode files and geo-displaced cluster coordinates, available by request from ICF International [31]; high-resolution (30cm) satellite imagery available through ESRI via ArcGIS 10.5 [32]; and multiple spatial data layers such as land cover type, nighttime lights intensity, and health facility locations all summarized in Table 1 and described elsewhere [33].

119 **Table 1.** Data sources for simulated population

Short name	Long name	Source, original unit	Output unit
Population			
dhs_hh	Individual recode file summarized by household	2013 Demographic and Health Survey [31]	region
dhs_geo	Geo-displaced cluster coordinates	2013 Demographic and Health Survey [31]	coordinate (cluster)
census_housing, census_person	20% census microdata sample	2011 National Statistics Agency [29]	constituency
census_report	Final census report	2011 National Statistics Agency [28]	constituency
Used to generate new spatial data			
imagery	High resolution satellite imagery	2014-2016 DigitalGlobe Quickbird imagery, 50cm [32]	Coordinate (household)
census_ea	2011 Census EA boundaries	2011 Namibia Statistics Agency [30]	EA
Spatial covariates			
ccilc_dst011_2012	Distance to land-cover: Cultivated terrestrial lands	2008-2012 GlobCover, 300m [34]	100m
ccilc_dst040_2012	Distance to land-cover: Woody / Trees	2008-2012 GlobCover, 300m [34]	100m
ccilc_dst130_2012	Distance to land-cover: Shrubs	2008-2012 GlobCover, 300m [34]	100m
ccilc_dst140_2012	Distance to land-cover: Herbaceous	2008-2012 GlobCover, 300m [34]	100m
ccilc_dst150_2012	Distance to land-cover: Other terrestrial vegetation	2008-2012 GlobCover, 300m [34]	100m
ccilc_dst190_2012	Distance to land-cover: Urban	2008-2012 GlobCover, 300m [34]	100m
ccilc_dst200_2012	Distance to land-cover: Bare	2008-2012 GlobCover, 300m [34]	100m
cciwat_dst	Distance to water bodies	2000 OSM [35]	100m
dmsp_2011	Nighttime lights intensity	2012 Suomi VIIRS, 500m [36]	100m
gpw4coast_dst	Distance to coastline	GPWv4, 1km [37]	100m
osmint_dst	Distance to road intersections	2000 OSM [35]	100m
osmriv_dst	Distance to major water ways	2000 OSM [35]	100m
slope	Slope	2000 HydroSHEDS, 100m [38]	100m
topo	Elevation	2000 HydroSHEDS, 100m [38]	100m
tt50k_2000	Travel time to populated places	2000 JRC-EC	100m
urbpx_prp_1_2012	Proportion of urban pixels with 1 cell radius	2009 Modis [39,40] & Global Human Settlement City Model [41], 1km	100m
hfacilities_dst	Distance to health centre or hospital	2001 UN-OCHA [42]	100m
schools_dst	Distance to primary or secondary school	2001 UN-OCHA [43]	100m
npp_2012	Annual net primary productivity	2010 MODIS, 1km [44]	100m

The 2011 Namibia 20% census microdata sample is comprised of 36,137 individuals in 7,536 conventional households [28], and the DHS survey sample is comprised of 3,316 individuals in 705 households located in 38 primary sampling units (PSUs) [31] (Table 2). In addition to the variables age, sex, relationship, and household size used to simulate household membership configurations, six covariates, common to both the DHS and census microdata, are simulated to support modelling of household type and prediction of outcome variables (Table 2). Four of these covariates are often used to operationalize the UN-Habitat definition of a “slum household”: lack of improved toilet, lack of improved water source, inadequate space defined as three or more people per sleeping room, and unimproved structure defined as having an earthen or wood floor [45]. Other characteristics include urban versus rural location, use of solid fuel for cooking, whether the head of household has no formal education, and whether there are any children under age five in the household.

While the microdata provides a large, systematic sample reflecting the distribution of characteristics in the population, it is not a complete census and cannot be linked to local geographic positions (in this case, below the constituency level). The DHS survey on the other hand, provides geographic coordinates, albeit displaced, for each PSU allowing us to explore spatial variation in the population. The method developed here leverages the strengths of each dataset and takes advantage of variables common to both datasets in order to link a simulated population to geographic positions.

Table 2. Size of Namibia 2011 20% Census Microdata Sample and 2013 DHS Sample, by sub-group

Variable Name	Category	20% Census	DHS	DHS
		<i>unweighted</i> n (%)	<i>unweighted</i> n (%)	<i>weighted</i> n (%)
Households	Oshikoto (N)	7,475	705	817
<i>urban_rural</i>	Urban	1,167 (15.6)	113 (16.0)	139 (17.1)
	Rural	6,308 (84.4)	592 (84.0)	678 (82.9)
<i>structure</i>	Durable floor	2,910 (38.9)	281 (39.8)	340 (41.6)
	Non-durable floor	4,551 (60.9)	422 (59.9)	475 (58.1)
	Missing/unknown	14 (0.2)	2 (0.3)	2 (0.3)
<i>fuel</i>	Non-solid fuel	1,217 (16.3)	141 (20.0)	182 (22.3)
	Solid fuel	6,253 (83.6)	562 (79.7)	633 (77.4)
	Missing/unknown	5 (0.1)	2 (0.3)	2 (0.3)
<i>water</i>	Improved water	5,388 (72.1)	589 (83.6)	688 (84.2)
	Unimproved water	2,045 (27.3)	72 (10.2)	80 (9.8)
	Missing/unknown	42 (0.6)	44 (6.2)	49 (7.0)
<i>toilet</i>	Improved toilet	1,955 (26.1)	207 (29.4)	258 (31.6)
	Unimproved toilet	5,491 (73.5)	492 (69.8)	553 (67.6)
	Missing/unknown	29 (0.4)	6 (1.0)	6 (0.8)
<i>space</i>	Adequate space	6,529 (87.3)	619 (87.8)	717 (87.7)
	Inadequate space	946 (12.7)	82 (11.6)	95 (11.6)
	Missing/unknown	0 (0.0)	4 (0.6)	6 (0.7)
<i>noedu</i>	Head household– any education	5,797 (77.6)	581 (82.4)	677 (82.8)
	Head household– no education	1,528 (20.4)	111 (15.7)	125 (15.3)
	Missing/unknown	150 (2.0)	13 (1.9)	15 (1.9)

<i>any_u5</i>	No child under age 5	4,267 (57.1)	405 (57.5)	478 (58.5)
	Any child under age 5	3,208 (42.9)	300 (42.5)	340 (41.5)
Individuals	Oshikoto (N)	36,137	3,316	3,576
<i>relationship</i>	Head	7,475 (20.7)	705 (22.5)	817 (22.9)
	Spouse	2,391 (6.6)	218 (7.0)	250 (7.0)
	Child	10,394 (28.8)	785 (25.0)	888 (24.8)
	Grandchild	8,635 (23.9)	591 (18.9)	660 (18.5)
	Extended	5,519 (15.3)	622 (19.8)	713 (19.9)
	Other	1,723 (4.8)	215 (6.9)	247 (6.9)
<i>sex</i>	Female	18,814 (52.1)	1,669 (53.2)	1,899 (53.1)
	Male	17,323 (47.9)	1,467 (46.8)	1,677 (46.9)
<i>age</i>	0	1,136 (3.1)	87 (2.8)	99 (2.8)
	1-4	3,968 (11.0)	364 (11.6)	414 (11.6)
	5-9	4,514 (12.5)	404 (12.9)	461 (12.9)
	10-14	4,895 (13.6)	389 (12.4)	435 (12.2)
	15-19	4,643 (12.9)	385 (12.3)	433 (12.1)
	20-24	3,284 (9.1)	280 (8.9)	323 (9.0)
	25-29	2,391 (6.6)	213 (6.8)	245 (6.9)
	30-34	1,912 (5.3)	195 (6.2)	230 (6.4)
	35-39	1,756 (4.9)	161 (5.1)	193 (5.4)
	40-44	1,371 (3.8)	106 (3.4)	120 (3.4)
	45-49	1,341 (3.7)	118 (3.8)	139 (3.9)
	50-54	968 (2.7)	102 (3.3)	118 (3.3)
	55-59	872 (2.4)	68 (2.2)	76 (2.1)
	60-64	802 (2.2)	71 (2.3)	79 (2.2)
	65-74	1,105 (3.1)	98 (3.1)	107 (3.0)
	75+	1,177 (3.3)	95 (3.0)	104 (2.9)

2.3. Simulation

We generate realistic household membership with realistic household point location and demographic and social characteristics in the following three phases. In Phase A, we define household types and then predict the spatial distribution of the types in Oshikoto using DHS data, spatial covariates, and visual inspection of satellite imagery. The output is a probability surface for each household type. In phase B, we generate the synthetic population using a census microdata sample and assign the population to household point locations using the household type probability surfaces generated in phase A. Phase C involves prediction of additional population characteristics in each household. The code is written in R [46] and spatial data are generated in ArcGIS [47]. Each phase is summarized in Figure 2 and described below. Five realizations of the simulated population (Supplement 1), the code (Supplement 2), and interim output (Supplement 3) is provided.

Figure 2. Simulation workflow with steps 1 through 8 organized in three phases. Green indicates original dataset, and orange indicates derived dataset.

Create 5km spatial covariates

Create Urban Weights

159

...continued

Phase B: Generate synthetic population, assign to household locations

Coverage: Oshikoto

5

Simulate Household Characteristics & Type

Input: **census_housing** (N=7,536 households),
census_person (N=37,767 individuals),
census_report (N=10 constituencies),
hh_types (N=7 types)

Method: multinomial logistic regression [R],
apply kmeans from **A** to calculate hh type [R]

Output: **hh_obs** (N=39,162 households).
Example variables:

- urban_rural**
- constituency**
- sex_hhhead**
- age_hhhead**
- toilet**
- water**
- structure**
- space**
- fuel**
- hh type**

Coverage: Oshikoto

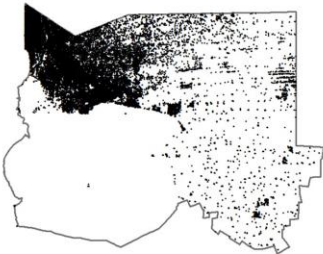
6

Create Building Locations

Input: **imagery** (30cm resolution),
census_report (N=10 constituencies)

Method: Digitize household points to match
census totals [ArcGIS]

Output: **hhpt** (N=37,298 households)



Coverage: Oshikoto

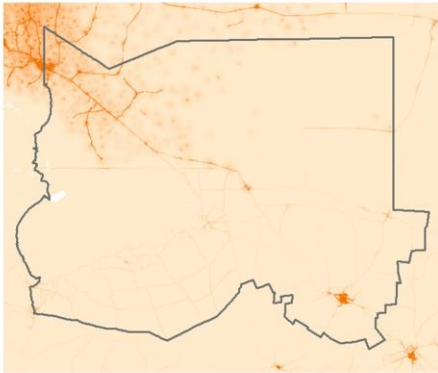
7

Assign Households to Locations

Input: **hh type_prob**, **urban_reweight**, **hhpt**

Method: multiply **hh type_prob** and
urban_reweight to get hh type probabilities
per cell [R], join hh type probabilities to
hhpt [R], for each **hh_obs** created in step B,
sample and assign one **hhpt** based on
joined hh type probability [R]

Output: **hh_obs_pts** (N=37,298 households)



HH ID	EA	hh type	X	Y
1					
2					
....					
37,298					

Phase C: Predict additional population characteristics

Coverage: Oshikoto

8

Simulate outcome variables

Input: **dhs_hh** (N=705 households), **hh_obs_pts**
(N=37,298 households)

Method: multinomial regression on DHS [R],
predicted probability on synthetic pop [R]

Output: **hh_obs_pts_out** (N=37,298 households).
New variables:

- Household wealth index*
- Women's use of modern contraception*
- Child's (under five) DPT3 vaccination*

HH ID	X	Y	Wealth
1				
2				
....				
37,298				

2.3.1. Phase A. Predict spatial distribution of household types

- Using the DHS dataset, we first define realistic and distinct types of households present in Oshikoto based on the 2013 DHS data of 705 households. We use the *kmeans* function in R [46] to generate a large number of clusters (k=20) from eight household demographic and social variables common to both the DHS and census microdata (*urban_rural*, *noedu*, *any_u5*, *toilet*, *water*, *structure*, *space*, *fuel*). K-means is a form of unsupervised clustering which seeks to partition observations into groups by minimising the within group sum of squares. We then utilize the

output dendrogram visualizing the hierarchically clustered k-means centroids to choose a smaller number of statistically distinct household types (long Euclidean distance between parent and child clusters in the dendrogram) that are easily interpretable. In the case of Namibia 2013 DHS, 7 household types are identified. To interpret and label household types, we consider whether the household type values are above, below, or near the Oshikoto average (Table 3). We save the k-means centroids and hierarchical clustering cut-off points to classify household types in other datasets in steps 3 and 5.

Table 3. Average prevalence of variables and label for each k-means household type cluster. Red indicates that the value is above the Oshikoto average (less desirable), and green indicates the value is below the Oshikoto average (desirable)

Cluster	urban_rural	noedu	any_u5	toilet	water	structure	space	fuel	Household type label
Type 1	0.00	0.00	0.04	0.06	0.00	0.00	0.00	0.00	Urban rich
Type 2	0.00	0.19	0.07	0.85	0.06	0.47	0.32	0.80	Urban poor
Type 3	1.00	0.05	0.12	0.55	0.00	0.00	0.04	0.10	Rural rich
Type 4	1.00	0.12	0.06	0.46	0.07	0.39	0.09	0.79	Rural middle
Type 5	1.00	.012	0.11	0.81	0.04	0.45	0.01	0.97	Rural middle (lack fuel)
Type 6	1.00	.012	0.16	0.92	0.49	0.83	0.06	0.96	Rural poor (lack water)
Type 7	1.00	0.22	0.13	0.91	0.09	0.83	0.04	0.98	Rural poor (lack education)
Oshikoto	0.84	.016	0.12	0.77	0.11	0.60	0.07	0.79	

- Second, we process 19 spatial covariates from free, public data sources including land cover types, night time light intensity, and health facility locations (see Table 1). These datasets are available for the whole region, enabling predictive mapping, and are shown to be related to population density [3], [48]. We convert each covariate into a 100 meter by 100 meter raster, and then for each cell, calculate the minimum, maximum, and average values within a five kilometre buffer using WGS84 geographic projection. This five kilometre moving window is used because the DHS data used to fit models in the next step are randomly geo-displaced up to five kilometres in rural areas. Further, the average covariate value within a five kilometre buffer of a displaced DHS PSU location is closer to the real, non-displaced, unpublished covariate value than the published, displaced covariate value [49,50]. Although DHS PSU coordinates are only displaced up to two kilometres in urban areas, a five kilometre buffer is used for all PSUs, and urban probability surfaces are improved manually in step 4.
- Third, using the 2013 DHS data for all of Namibia (N=550 clusters) and household types created in step 1, we calculate the most common household type for each PSU using the k-means centroids and cut-off points. Next, we extract the five kilometre averaged spatial covariates created in step 2 to each DHS PSU location, resulting in 550 observations of household type linked to (19 X 3) 57 spatial covariates. In this step 3 we find a relationship between household type and spatial covariates in order to predict household types over the whole region. To do this, we use a Random Forest model – a non-parametric ensemble machine-learning algorithm that grows a “forest” of decision trees during the modelling process [3] – to model this relationship and predict a 100 meter by 100 meter probability surface for each household type across Namibia.
- Fourth, we manually create household type probabilities for urban EAs. This step is necessary because initial tests found that the household type probability model generated in step 3 could not adequately distinguish household types within urban areas. This was expected given the displacement of the DHS PSU locations and the summary of geospatial covariate data which are

essentially identical across urban household types. Without step 4, simulated households of different socioeconomic types would be evenly spatially integrated in urban areas, which is unrealistic. Poor and rich households are often segregated in urban areas worldwide [51], and visual inspection of satellite imagery indicates that socioeconomic segregation is present in Oshikoto's urban areas as well. From Step 1, we label the two urban household types as poor and rich, then manually assign proportion of households that we judge to be rich versus poor within each EA based on satellite imagery, such that the probabilities sum to 1. These manually created EA-level urban household type probabilities are multiplied by the predicted household type probability surfaces created in step 3 to create the final 100 meter by 100 meter household type probability surfaces.

2.3.2. Phase B. Generate synthetic population, assign household locations

5. Fifth, we simulate a population of realistic households in Oshikoto using the 20% census microdata sample and multinomial logistic regression techniques proposed by Alfons and colleagues (2011) and operationalized by Templ and colleagues (2017) in the R simPop package [7,15]. In this approach, we first calculate the proportion of households to simulate per household-size, per stratum (defined by constituency and urban/rural boundary). Second, we select random resamples from the microdata until the number of target households are reached in each household size and strata. Third, demographic characteristics of the household members (*age, sex, relationship*) are replicated from the microdata. Fourth, we add household socioeconomic characteristics to the simulated dataset (*education, toilet, water, structure, space, fuel*) using multinomial regression. This allows for simulation of combinations of demographic characteristics that exist in the population but are not present in the census microdata. For each simulated household, we assign the household type by selecting the class from step 1 with the smallest distance (i.e. most similar) between each household record and the k-means centroids.
6. The census microdata sample is provided with a weight equal to five for nearly all conventional households. We recalibrate these weights to the total number of households per constituency in the 2011 census [28]. However this process can lead to too few observations in some constituency-urban/rural strata, and too many observations in other strata. Therefore, we increase the weights to simulate an extra 5% of households from which a random selection of households is assigned to latitude-longitude coordinates in step 7.
7. Seventh, we join reweighted household type probabilities (100 metre X 100 metre grid) created in step 4 to the household latitude-longitude coordinates created in step 6. Finally, for each household simulated in step 5, we randomly sample one latitude-longitude coordinate within the constituency-urban/rural strata based on the probability of household type. We repeat the assignments until all coordinates are assigned a simulated household, and then discard the extra 5% unassigned simulated households.

2.3.2. Phase C. Predict additional population characteristics, generalize locations

8. In step 8, we use the 2013 DHS records in Oshikoto (N=705 households) to develop multinomial models of socioeconomic and health outcome variables. We store the coefficients of each model and apply them to our simulated dataset to predict outcomes in each simulated household. The three simulated outcome variables represent different prevalence levels and patterns of dispersion in the population. These outcome variables represent children under age five, women of reproductive age, and households in order to support within household clustering analyses. The outcome variables are: household wealth (expressed in quintiles), women's use of modern contraception (approximately 50% in Namibia and Oshikoto), and child's receipt of 3rd DPT vaccination (approximately 90% in Namibia and Oshikoto) [52]. Multinomial models are used for both multi-category and binary outcomes

$$\Pr(Y_i = K - 1) = \frac{e^{\beta_{K-1} \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_{K-1} \cdot X_i}} \quad (1)$$

where K is the number of categories in the outcome variable, Y_i is the outcome value for individual i , and X_i is a matrix of covariate values belonging to individual i . Model coefficients are applied to covariates of the 37,298 households in the simulated dataset to predict outcome values.

2.4. Assessment

We conduct global assessments to evaluate whether each of the five realizations of the simulated population are realistic overall, and local assessment to evaluate whether the realizations are realistic at an EA-level. In the global assessment, we aggregate the DHS records to PSU and the simulated census records to EA, and graphically compare the distributions of simulated covariates and outcomes. We also map simulated census records by EA to visually inspect the spatial distributions across Oshikoto. In the local assessment, DHS data are averaged by PSU and compared to the distribution from repeated samples simulating a set of survey respondents. For each of 10,000 simulations, a random EA is selected within 5km of each DHS PSU coordinate, then the same number of households as the observed DHS cluster are drawn from the simulated population. The characteristics are averaged from the sampled EAs and compared to the observed DHS data.

2.5. Ethics

Before releasing our simulated data, we closely reviewed papers about privacy of synthetic population data including a paper by Alfons and Templ (2010) who calculated disclosure risk of close-to-reality synthetic data generated with the simPop [R package] algorithm used in this analysis [53]. The authors found extremely low risk of disclosure for five worst case scenarios and concluded that simulations “implemented in simPop are confidential and can be distributed to the public” [53]. Any additional risk in our study due to linking simulated records to realistic building locations is negligible due to random spatial components in the analysis, and as a result of beginning with a random sample of the original census microdata in Phase B. Any match between characteristics in a simulation realization of a household at a given building location and a real-world household at that same location is purely by chance.

The main risk in this analysis is misinterpretation and/or misuse of the synthetic population data by users (e.g. believing that the simulated data are from actual households and treating real-world household members, or their communities, with stigma). To minimize misinterpretation, we release five realizations of the synthetic population and label each dataset as “synthetic”. To further minimize the risk of maltreatment of real-world people in the case that these data are misinterpreted, we only simulated commonly mapped variables which have been interpolated with real-world survey data to 1 km² grid square by the MeasureDHS project [54].

This analysis and public release of simulated data was reviewed by the University of Southampton Ethics Review Committee (#41006).

3. Results

Demographic and socioeconomic characteristics of the five simulated populations in Oshikoto (Table 4) were consistent with the 2013 DHS and 20% census distributions presented in Table 2.

Table 4. Demographic and socioeconomic characteristics of five realizations of the synthetic population

Variable	Category	pop_1 (%)	pop_2 (%)	pop_3 (%)	pop_4 (%)	pop_5 (%)
Households	Oshikoto (N)	37,298	37,298	37,298	37,298	37,298
<i>urban_rural</i>	Urban	84.3	84.3	84.3	84.3	84.3
	Rural	15.7	15.7	15.7	15.7	15.7
<i>structure</i>	Durable floor	38.6	38.7	38.6	38.5	37.9
	Non-durable floor	61.4	61.3	61.4	61.5	62.1
<i>fuel</i>	Non-solid fuel	16.2	16.4	16.0	16.0	15.9
	Solid fuel	83.8	83.6	84.0	84.0	84.1
<i>water</i>	Improved water	73.2	73.2	72.9	73.1	72.7
	Unimproved water	26.8	26.8	27.1	26.9	27.3
<i>toilet</i>	Improved toilet	20.1	20.1	19.9	19.7	19.5
	Unimproved toilet	79.9	79.9	80.1	80.3	80.5
<i>space</i>	Adequate space	92.5	92.2	92.3	92.5	92.3
	Inadequate space	7.5	7.8	8.7	7.5	7.7
<i>noedu</i>	Head household - any education	70.8	70.5	70.5	70.8	70.9
	Head household - no education	29.2	29.5	29.5	29.2	29.1
<i>any_u5</i>	No child under age 5	57.4	57.0	56.8	57.1	57.0
	Any child under age 5	42.6	43.0	43.2	42.9	43.0
Individuals	Oshikoto (N)	179,931	179,854	180,233	180,164	180,111
<i>relationship</i>	Head	20.7	20.7	20.7	20.7	20.7
	Spouse	6.6	6.6	6.5	6.6	6.6
	Child	28.8	28.8	28.7	28.9	28.8
	Grandchild	23.8	24.0	23.9	23.8	23.8
	Extended	15.1	15.1	15.2	15.0	15.3
	Other	4.9	4.8	5.0	4.9	4.8
<i>sex</i>	Female	52.2	52.0	51.9	51.8	52.0
	Male	47.8	48.0	48.1	48.2	48.0
<i>age</i>	0	3.1	3.1	3.2	3.1	3.2
	1-4	10.9	11.1	11.1	10.9	10.9
	5-9	12.7	12.6	12.5	12.4	12.7
	10-14	13.6	13.6	13.6	13.7	13.6
	15-19	12.9	12.9	12.7	13.0	12.9
	20-24	9.0	9.0	9.1	9.1	9.0
	25-29	6.7	6.6	6.6	6.6	6.6
	30-34	5.2	5.3	5.3	5.2	5.3
	35-39	4.9	4.9	5.0	4.9	4.9
	40-44	3.8	3.8	3.7	3.9	3.8
	45-49	3.7	3.8	3.8	3.8	3.7
	50-54	2.7	2.7	2.7	2.7	2.7
	55-59	2.4	2.4	2.4	2.4	2.4
	60-64	2.2	2.2	2.2	2.2	2.2

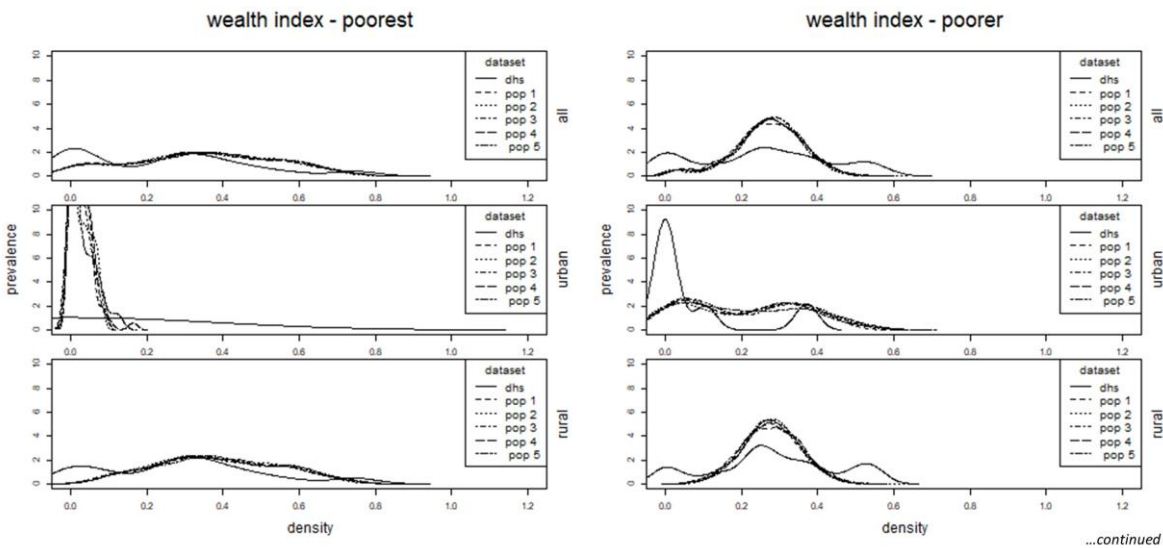
65-74	3.1	3.1	3.1	3.0	3.0
75+	3.2	3.1	3.2	3.2	3.2

294

295 The distribution of the three outcomes were heaped in the 2013 DHS dataset, perhaps due to
296 small sample size. In the global assessment of the simulated population by PSU/EA in Oshikoto,
297 Namibia, the distributions of households per wealth quintile, contraceptive use among reproductive
298 age women, and percent children who received 3rd DPT vaccination were consistent between the
299 2013 DHS PSUs and the synthetic population EAs in all five realizations of the population (Figure 3).
300 A key difference is that the Oshikoto synthetic populations distribute more households in the lowest
301 wealth quintile, while the DHS measured a greater percent of Oshikoto households in the second
302 lowest wealth quintile.

303 Maps showing simulated household wealth by EA followed expected spatial patterns with
304 higher wealth in planned urban neighbourhoods and large rural towns, and lowest household wealth
305 in remote rural areas (Figure 4, realization 1). Similarly, higher rates of contraceptive use were located
306 in urban EAs, and wealthier rural EAs, as expected. Namibia has greater DTP3 vaccination coverage
307 in rural, rather than urban, populations, which is atypical of LMICs [52]. This atypical pattern is
308 reflected in the maps of DPT3 vaccination coverage among one of the simulated population.

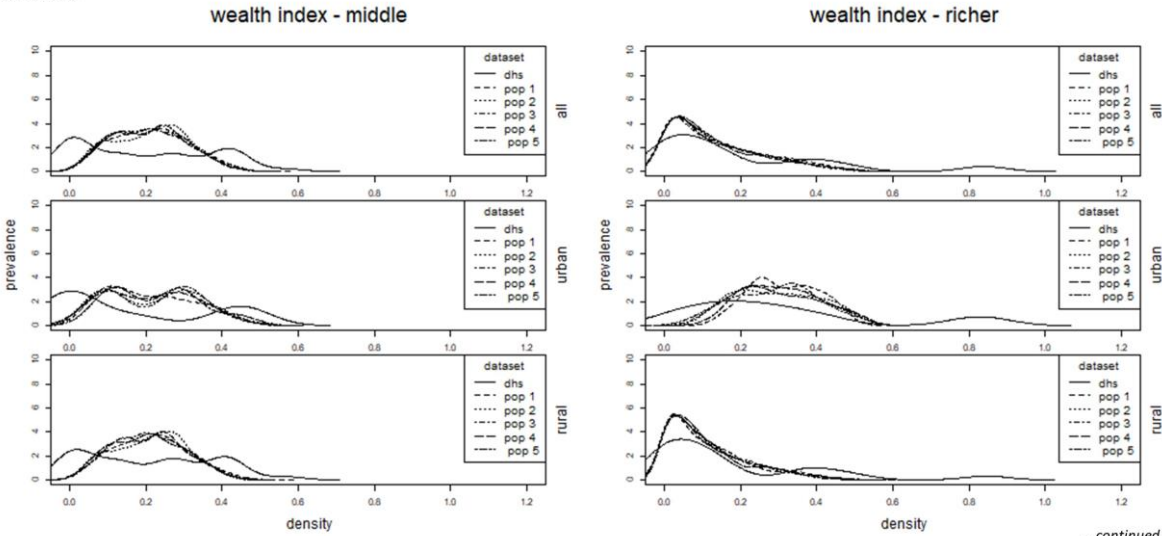
309 **Figure 3.** Comparison of outcome variables in the 2013 Namibia DHS (Oshikoto region only)
310 (solid line) and five synthetic population realizations (dotted lines)



311

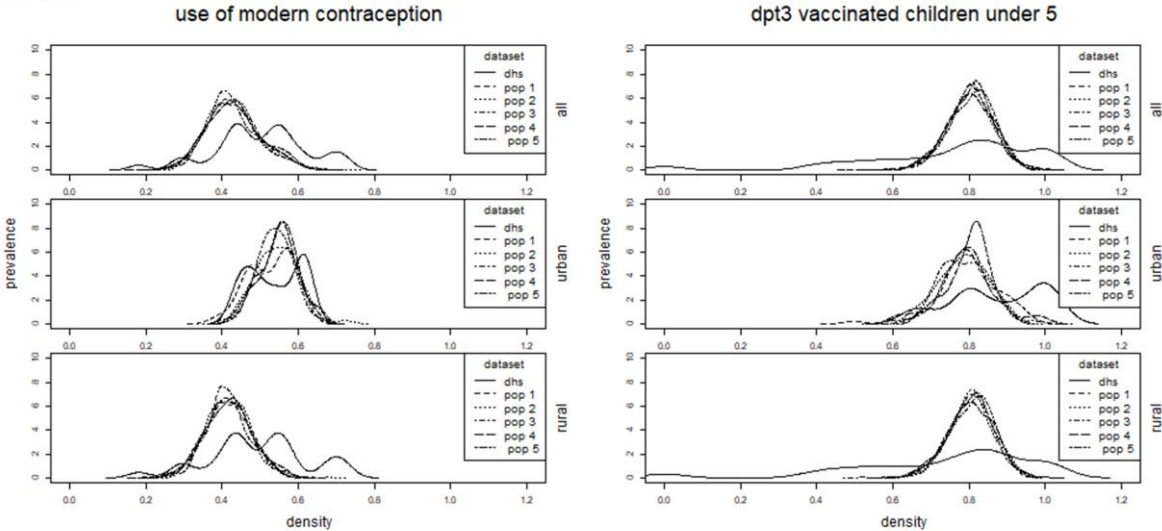
...continued

continued...



...continued

continued...

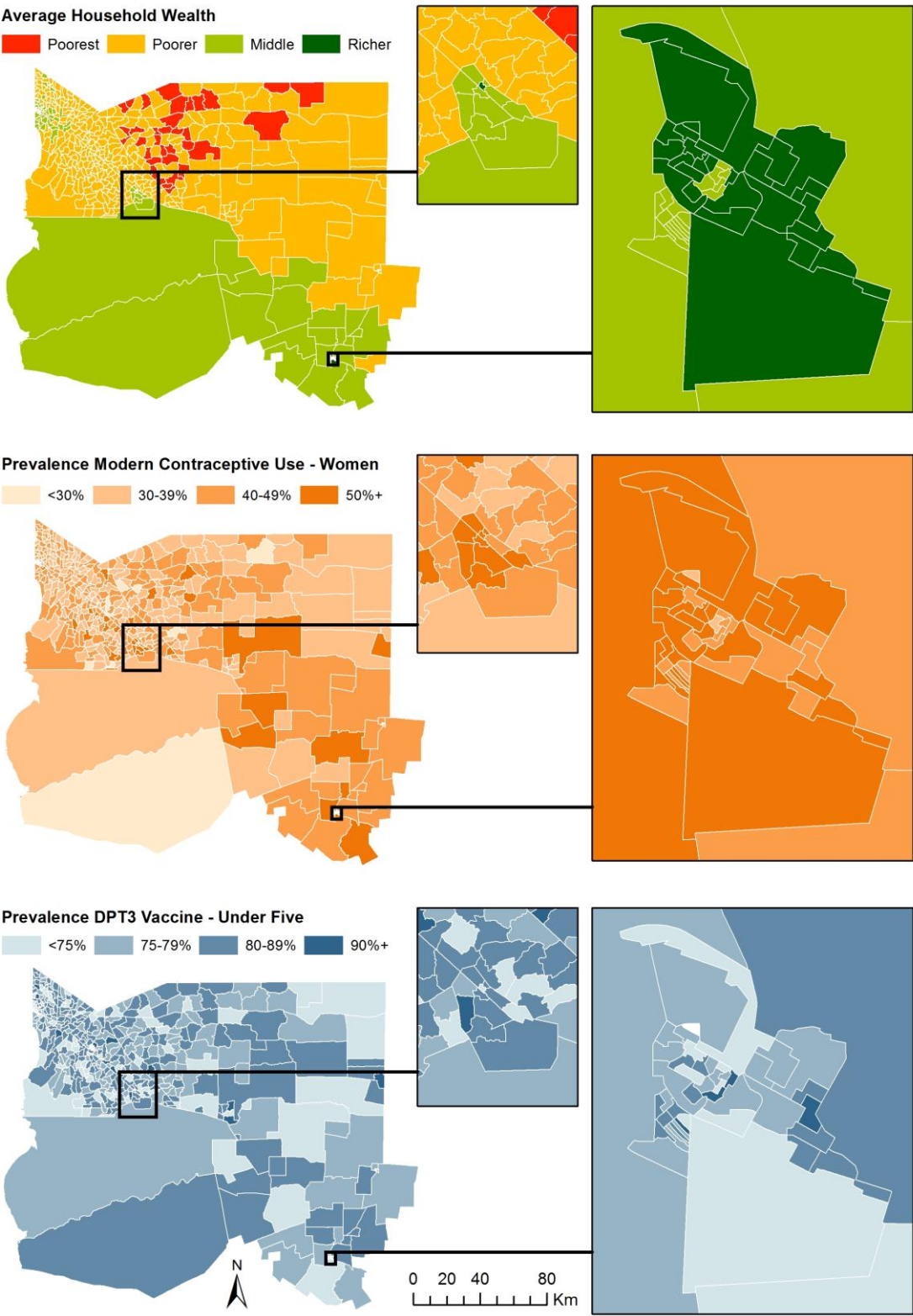


312

313

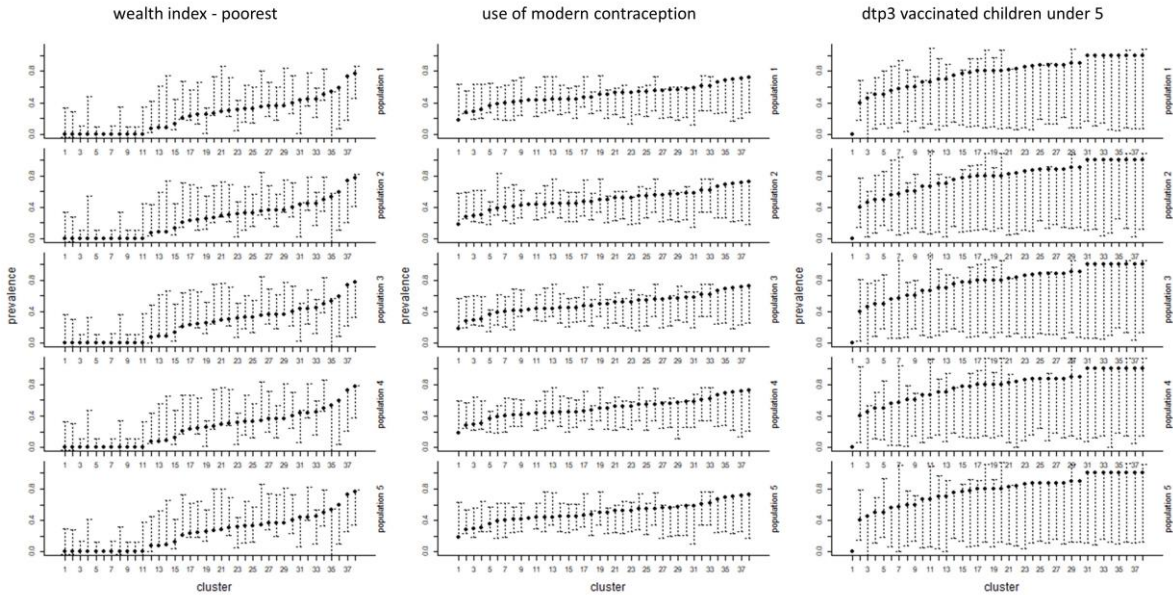
314

Figure 4. Maps of outcome variables by EA in one simulated population (synth_pop_1) of Oshikoto, Namibia



In the local EA-level assessment, we found that DHS estimates for each of the 38 Oshikoto clusters fell within the 95% confidence interval of repeated random simulated samples from the simulated population EAs near to the DHS PSU. This implies that the observed DHS results could potentially have been drawn from the synthetic population.

Figure 5. 2013 Namibia DHS PSU-level estimates of outcome variables versus the distribution of 100 samples selected in EAs located within 5 kilometres of DHS PSU, for five synthetic population realizations



4. Discussion

Close-to-reality simulated populations are needed to answer questions at the forefront of spatial demographic research and survey methodological development while reducing disclosure risks of releasing high spatial resolution census data. We outline a novel process to simulate multiple realizations of a population linked to realistic latitude-longitude coordinates in a LMIC setting. Our approach uses the strengths of two commonly available population datasets – household surveys and census microdata samples. We also draw together computational methods in microsimulation of individuals and households and high-resolution mapping of household characteristics that uses geospatial data. The result is a full enumeration of a synthetic population with household relations and characteristics, linked to realistic locations. The simulated population was assessed and found to be realistic in terms of socioeconomic and health outcomes at both regional and local (community) levels. We released the code and five realizations of the simulated population to encourage additional simulations of close-to-reality populations to realistic latitude-longitude coordinates, and to support development of household surveys and gridded population survey sample frames for LMICs.

One such question is whether one-stage sampling can result in precise and feasible household surveys compared to the classic two-stage sampling design. Nearly every nationally-representative multi-topic household survey implemented since the 1980s in LMICs has used a two-stage sampling design with census enumeration areas comprising the first-stage sample frame and a manual household listing comprising the second-stage sample frame [9]. This has proven to be an effective sample design when census EAs are the only available first-stage sample frame, maximizing statistical power while reducing field costs [55]–[57]. Two-stage sampling, however, requires that two field visits are made to each sampled household several months (or even years) apart, making it more likely that mobile and vulnerable households are excluded from the survey or fail to respond compared to stable long-term households [58]. This problem is of increasing concern in LMICs cities today as rates of urbanization and mobility increase [51], possibly leading to increased bias in standard two-stage household surveys. Gridded sampling frames open the door for one-stage

surveys, such that households are listed and interviewed on the same day, which can theoretically improve the accuracy of poor and vulnerable households in household surveys, however, one-stage sampling comes at the risk of increased design effect, requiring increased sample size. The use of close-to-reality simulated populations can be used to compare various sample designs under different realistic conditions of population distribution, mobility, and characteristics.

Another application of close-to-reality population simulations is the evaluation of gridded population dataset accuracy at the cell-level. Several gridded population datasets are generated at 100 metre by 100 metre scale from census data [3], [4]. Accuracy of these models is often performed at the geographic scale of the input census data, however accuracy is never evaluated at the grid cell-level. Microdata located to realistic household locations and aggregated to 100 metre X 100 metre grid cells provides a first opportunity for this kind of accuracy assessment.

One limitation of this work is that it relied on manually digitised building point locations and delineation of urban rich vs poor household locations. This data creation step was manageable for a subnational region but would require substantial time to scale nationally. It took one GIS analyst nearly one week of full-time work to generate building point locations in Oshikoto for this analysis. However, as coverage of publicly available sub-metre satellite imagery increases globally, so does automated feature extraction of individual buildings in LMICs [59], which is promising to help scale this simulation approach to larger geographic areas. Note that if feature extraction is used to generate building locations, additional information or researcher judgement may still be needed to identify multi-household building locations and to remove non-residential buildings. Machine learning techniques are showing promise in mapping neighbourhood types from very high resolution imagery [60] and other building datasets [61] which can also help address this limitation.

One might wonder why not generate random points for building locations within administrative areas near roads, or by using some other set of simple rules, as other researchers have done to simulate close-to-reality populations [19]. While this would permit certain types of analysis such as the comparison of one-stage and two-stage sampling, creation of random points for households within large administrative areas is not recommended if the simulated population will be used to evaluate accuracy of gridded population models, particularly gridded populations with real-world spatial covariates at fine geographic scale (e.g. 100 metre X 100 metre). There is a large amount of heterogeneity in human population distribution, and this must be reflected accurately at a very local level to be able to evaluate gridded population models on a cell-by-cell basis.

This novel method to simulate close-to-reality household records linked to realistic building locations in a LMIC stands to support development of more accurate household survey methods and gridded population datasets as household survey sample frames. These methods are feasible to implement in other LMIC settings and will become globally scalable as feature extraction methods evolve.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Data S1: Five realizations of a simulated, geo-located population in Oshikoto, Namibia, Code S2: R code to produce a simulated, geo-located population.

Author Contributions: Conceptualization, DRT, LK, and WCJ; Methodology, LK and WCJ; Formal Analysis, LK; Data Curation, DRT and LK; Writing-Original Draft Preparation, DRT; Writing-Review & Editing, LK, WCJ

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

- Doxsey-Whitfield, E.; MacManus, K.; Adamo, S. B.; Pistolesi, L.; Squires, J.; Borkovska, O.; Baptista, S. R. 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 (3), 226–234, DOI: 10.1080/23754931.2015.1014272
- Oak Ridge National Laboratories. LandScan Documentation. Available online: http://web.ornl.gov/sci/landscan/landscan_documentation.shtml (accessed Feb 6, 2007).
- Stevens, F. R.; Gaughan, A. E.; Linard, C.; Tatem, A. J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS One* 2015, 10 (2), e0107042, DOI:10.1371/journal.pone.0107042.
- Azar, D.; Graesser, J.; Engstrom, R.; Comenetz, J.; Leddy, R. M.; Schechtman, N. G.; Andrews, T. Spatial Refinement of Census Population Distribution Using Remotely Sensed Estimates of Impervious Surfaces in Haiti. *Int. J. Remote Sens.* 2010, 31 (21), 5635–5655, DOI:10.1080/01431161.2010.496799
- Hay, S. I.; Noor, A. M.; Nelson, A.; Tatem, A. J. The Accuracy of Human Population Maps for Public Health Application. *Trop. Med. Int. Heal.* 2005, 10 (10), 1073–1086.
- Tatem, A. J.; Noor, A. M.; Hay, S. I. Assessing the Accuracy of Satellite Derived Global and National Urban Maps in Kenya. *Remote Sens. Environ.* 2005, 96 (1), 87–97, DOI:10.1016/j.rse.2005.02.001.
- 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, DOI:10.1007/s10260-011-0163-2.
- Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.0 [Dataset]; Minneapolis, MN, USA, 2018.
- Global Health Data Exchange (GHDx). Available online: <http://ghdx.healthdata.org/> (accessed Feb 2, 2017).
- Tanton, R. A Review of Spatial Microsimulation Methods. *Int. J. Microsimulation* 2014, 7 (1), 4–25, DOI: 10.3153/jfscm.2010021.
- Birkin, M.; Clarke, M. The Generation of Individual and Household Incomes at the Small Area Level Using Synthesis. *Reg. Stud.* 1989, 23 (6), 535–548, DOI:10.1080/00343408912331345702.
- Birkin, M.; Clarke, M. SYNTHESIS: A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples. *Environ. Plan. A* 1988, 20, 1645–1671, DOI:10.1068/a201645.
- Ballas, D.; Kingston, R.; Stillwell, J.; Jin, J. Building a Spatial Microsimulation-Based Planning Support System for Local Policy Making. *Environ. Plan. A* 2007, 39, 2482–2499, DOI:10.1068/a38441.
- Farrell, N.; Morrissey, K.; O'Donoghue, C. Creating a Spatial Microsimulation Model of the Irish Local Economy. In *Spatial microsimulation: A Reference Guide for Users. Understanding Population Trends and Processes*, vol 6.; Tanton, R., Edwards, K., Eds.; Springer: Dordrecht, 2012; pp 105–125.
- Templ, M.; Meindl, B.; Kowarik, A.; Dupriez, O. Simulation of Synthetic Complex Data: The R package SimPop. *J. Stat. Softw.* 2017, 79 (10), 1–38, DOI:10.18637/jss.v079.i10.
- Macal, C. M. Everything You Need to Know About Agent-Based Modelling and Simulation. *J. Simul.* 2016, 10 (2), 144–156, DOI:10.1057/jos.2016.7.
- Chapuis, K.; Taillandier, P.; Renaud, M.; Drogoul, A. Gen*: A Generic Toolkit to Generate Spatially Explicit Synthetic Populations. *Int. J. Geogr. Inf. Sci.* 2018, 32 (6), 1–17, DOI:10.1080/13658816.2018.1440563.
- Heppenstall, A.; Malleon, N.; Crooks, A. "Space, the Final Frontier": How Good are Agent-Based Models at Simulating Individuals and Space in Cities? *Systems* 2016, 4 (1), 9, DOI:10.3390/systems4010009.
- Synthetic Populations and Ecosystems of the World (SPEW). Available online: <http://www.stat.cmu.edu/~spew/about/> (accessed May 15, 2018).
- Synthetic Household Population™. Available online: <https://www.rti.org/impact/synthpop> (accessed May 15, 2018).

21. SDG Indicators: Revised List of Global Sustainable Development Goal Indicators. Available online: <https://unstats.un.org/sdgs/indicators/indicators-list/> (accessed Sep 3, 2017).
22. Tatem, A. J. WorldPop, Open Data for Spatial Demography. *Sci. Data* 2017, 4, 170004, DOI: 10.1038/sdata.2017.4.
23. Bosco, C.; Alegana, V.; Bird, T.; Pezzulo, C.; Bengtsson, L.; Sorichetta, A.; Steele, J.; Hornby, G.; Ruktanonchai, C.; Ruktanonchai, N.; et al. Exploring the High-Resolution Mapping of Gender-Disaggregated Development Indicators. *J. R. Soc. Interface* 2017, 14 (129), 20160825, DOI:10.1098/rsif.2016.0825.
24. Alegana, V. A.; Atkinson, P. M.; Pezzulo, C.; Sorichetta, A.; Weiss, D.; Bird, T.; Erbach-Schoenberg, E.; Tatem, A. J. Fine Resolution Mapping of Population Age-Structures for Health and Development Applications. *J. R. Soc. Interface* 2015, 12, 1–11, DOI:10.1098/rsif.2016.0825.
25. Utazi, C. E.; Thorley, J.; Alegana, V. A.; Ferrari, M. J.; Takahashi, S.; Metcalf, C. J. E.; Lessler, J.; Tatem, A. J. High Resolution Age-Structured Mapping of Childhood Vaccination Coverage in Low and Middle Income Countries. *Vaccine* 2018, 36 (12), 1583–1591, DOI:10.1016/j.vaccine.2018.02.020.
26. Thomson, D. R.; Stevens, F. R.; Ruktanonchai, N. W.; Tatem, A. J.; Castro, M. C. GridSample: An R Package to Generate Household Survey Primary Sampling Units (PSUs) from Gridded Population Data. *Int. J. Health Geogr.* 2017, 16 (1), DOI:10.1186/s12942-017-0098-4.
27. 2020 World Population and Household Census Programme Census Dates for All Countries. Available online: <https://unstats.un.org/unsd/demographic/sources/census/censusdates.htm> (accessed Mar 3, 2017).
28. [Namibia] National Statistics Agency. Namibia Population and Housing Census 2011: Main Report; Government of Namibia: Windhoek, Namibia, 2011.
29. [Namibia] National Statistics Agency. Namibia 2011 Population and Housing Census [PUMS Dataset]. Version 1.0.; Windhoek, Namibia, 2013.
30. [Namibia] National Statistics Agency 2011 Census EA Boundaries. Available online: <https://digitalnamibia.nsa.org.na/> (accessed Feb 19, 2018).
31. ICF International Available Datasets. Available online: <https://dhsprogram.com/data/available-datasets.cfm> (accessed Nov 15, 2017).
32. DigitalGlobe Quickbird 50cm Imagery. Available online: <http://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9> (accessed Feb 1, 2018).
33. Lloyd, C. T.; Sorichetta, A.; Tatem, A. J. High Resolution Global Gridded Data for Use in Population Studies. *Nat. Sci. Data* 2017, 4, 1–17, DOI:10.1038/sdata.2017.1.
34. European Space Agency GlobCover. Available online: www.esa-landcover-cci.org/?q=node/158 (accessed Feb 19, 2017).
35. OpenStreetMap Base Data. Available online: www.openstreetmap.org (accessed Feb 19, 2017).
36. NOAA VIIRS Nighttime Lights. Available online: http://www.ngdc.noaa.gov/dmsp/data/viirs_fire/viirs_html/viirs_ntl.html (accessed Feb 19, 2017).
37. CIESIN Gridded Population of the World, Version 4 (GPWv4). Available online: <http://dx.doi.org/10.7927/H4F47M2C> (accessed Feb 19, 2017).
38. Lehner, B.; Verdin, K.; Jarvis, A. HydroSHEDS Technical Documentation; World Wildlife Fund: Washington DC, USA, 2006.
39. Schneider, A.; Friedl, M. A.; Potere, D. Mapping Global Urban Areas Using MODIS 500-m Data: New Methods and Datasets Based on “Urban Ecoregions.” *Remote Sens. Environ.* 2010, 114, 1733–1746, DOI:10.1016/j.rse.2010.03.003.
40. Schneider, A.; Friedl, M. A.; Potere, D. A New Map of Global Urban Extent from MODIS Satellite Data. *Environ. Res. Lett.* 2009, 4, 1–11, DOI:10.1088/1748-9326/4/4/044003.
41. European Commission. Global Human Settlement City Model (GHS-SMOD). Available online: <http://ghsl.jrc.ec.europa.eu/faq.php> (accessed Feb 6, 2017).
42. UN-OCHA-ROSA Namibia - Health Facilities. Available online: <https://data.humdata.org/organization/ocha-rosa> (accessed Feb 19, 2017).
43. UN-OCHA-ROSA Namibia - Education Facilities. Available online: <https://data.humdata.org/organization/ocha-rosa> (accessed Feb 19, 2017).
44. Steven W., R.; Ramakrishna R., N.; Faith Ann, H.; Maosheng, Z.; Matt, R.; Hirofumi, H. A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production. *Bioscience* 2004, 54 (6), 547–560.

45. Fink, G.; Günther, I.; Hill, K. Slum Residence and Child Health in Developing Countries. *Demography* 2014, 51 (4), 1175–1197, DOI:10.1007/s13524-014-0302-0.
46. R Core Team. R: Algorithm and Environment for Statistical Computing. R Core Team: Vienna, Austria, 2013.
47. ESRI. ArcGIS Release 10. Environmental Systems Research Institute: Redlands CA, USA, 2018.
48. Nieves, J. J.; Stevens, F. R.; Gaughan, A. E.; Linard, C.; Sorichetta, A.; Hornby, G.; Patel, N. N.; Tatem, A. J. Examining the Correlates and Drivers of Human Population Distributions across Low- and Middle-Income Countries. *J. R. Soc. Interface* 2017, 14 (137), 20170401, DOI:10.1098/rsif.2017.0401.
49. Burgert, C. R.; Zachary, B.; Colston, J. Incorporating Geographic Information into Demographic and Health Surveys: A Field Guide to GPS Data Collection; ICF International: Calverton, MD, 2013.
50. Perez-Heydrich, C.; Warren, J. L.; Burgert, C. R.; Emch, M. E. Influence of Demographic and Health Survey Point Displacements on Raster-Based Analyses. *Spat. Demogr.* 2016, 4 (2), 135–153, DOI:10.1007/s40980-015-0013-1.
51. UN Habitat. Urbanization and Development: Emerging Futures. World Cities Report 2016; United Nations Human Settlements Programme (UN-Habitat): Nairobi, Kenya, 2016.
52. [Namibia] Ministry of Health and Social Services (MoHSS); ICF International. Namibia Demographic and Health Survey 2013; ICF International: Windhoek, Namibia, and Rockville MD, USA, 2014.
53. Alfons, A.; Templ, M. Disclosure Risk of Synthetic Population Data with Application in the Case of EU-SILC. In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, vol 6344.; Domingo-Ferrer, J., Magkos, E., Eds.; Springer: Heidelberg, Germany, 2010; pp 174–186.
54. The Demographic and Health Surveys Program Modeled Surfaces. Available online: <https://spatialdata.dhsprogram.com/modeled-surfaces/> (accessed Apr 16, 2018).
55. United Nations Children's Fund (UNICEF). Multiple Indicator Cluster Surveys Round 4 (MICS4). Designing and Selecting the Sample; UNICEF: New York NY, USA, 2012.
56. United Nations (UN). Designing Household Survey Samples: Practical Guidelines. Studies in Methods Series F No.98; UN: New York NY, USA, 2005.
57. ICF International. Demographic and Health Survey Sampling and Household Listing Manual; ICF International: Calverton MD, USA, 2012.
58. Elsey, H.; Thomson, D. R.; Lin, R. Y.; Maharjan, U.; Agarwal, S.; Newell, J. Addressing Inequities in Urban Health: Do Decision-Makers Have the Data They Need? Report from the Urban Health Data Special Session at International Conference on Urban Health Dhaka 2015. *J. Urban Heal.* 2016, 93 (3), DOI:10.1007/s11524-016-0046-9
59. A Breakthrough in Building Footprint Extraction. Available online: <http://explore.digitalglobe.com/GBDX-Building-Footprints.html> (accessed May 15, 2018).
60. Graesser, J.; Cheriyaat, A.; Vatsavai, R. R.; Chandola, V.; Long, J.; Bright, E. Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2012, 5 (4), 1164–1176, DOI:10.1109/JSTARS.2012.2190383.
61. Jochem, W. C.; Bird, T. J.; Tatem, A. J. Identifying Residential Neighbourhood Types from Settlement Points in a Machine Learning Approach. *Comput. Environ. Urban Syst.* 2018, 69, 104–113, DOI:10.1016/j.compenvurbsys.2018.01.004.