*data*

MDPI

*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, SE-11355 Stockholm, Sweden; w.c.jochem@soton.ac.uk
2 WorldPop, Department of Geography and Environment, University of Southampton, Southampton SO17 1BJ, UK
3 Department of Social Statistics, University of Southampton, Southampton SO17 1BJ, UK
4 Department of Economics, Leiden University, 2311 EZ Leiden, The Netherlands; l.kools@law.leidenuniv.nl
* Correspondence: dana.thomson@flowminder.org; Tel.: +44-238-202-6000
† These authors contributed equally to this work.

Received: 18 June 2018; Accepted: 7 August 2018; Published: 9 August 2018


**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 spatial covariates derived from publicly available 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 k-means 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.

**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 the 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 optimization 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 in 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 counties (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 static 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 layers derived from public data sources 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 and might be made dynamic with spatiotemporal modelling. 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 (approximately 100 m $\times$ 100 m 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 was selected for the simulation because the population varies widely from low-to-high density, and the 2011 Namibia census meets the UN recommendations for high-quality census data [27].

We selected 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 neighborhoods. Oshikoto is comprised of 10 administrative sub-regions called constituencies, for which there are published census population and household totals.

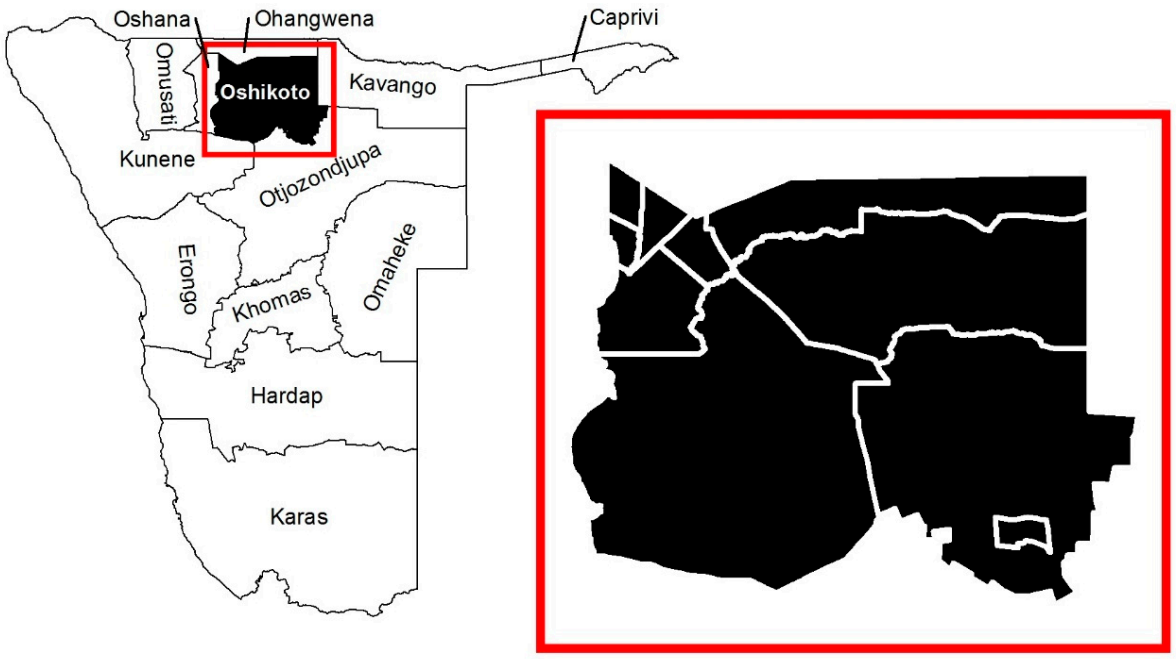



**Figure 1.** Map of Oshikoto Region, Namibia and Oshikoto's 10 constituency boundaries.

### *2.2. Data*

Input data included 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 (30 cm) satellite imagery available through ESRI via ArcGIS 10.5 [32]; and multiple spatial data layers derived from public sources such as land cover type, nighttime lights intensity, and health facility locations summarized in Table 1. Building off of earlier work [33], spatial covariates were processed as part of the "Global High Resolution Population Denominators" Project by the WorldPop team at University of Southampton and Center for International Earth Science Information Network at Columbia University, and are detailed in a forthcoming paper (Alessandro Sorichetta, personal communication, July 2018).

**Table 1.** Data sources used to generate a simulated population in Oshikoto, Namibia.

| Name | Description | 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 [32]; 50 cm | 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 | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| ccilc_dst040_2012 | Distance to land-cover: Woody/Trees | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| ccilc_dst130_2012 | Distance to land-cover: Shrubs | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| ccilc_dst140_2012 | Distance to land-cover: Herbaceous | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| ccilc_dst150_2012 | Distance to land-cover: Other terrestrial vegetation | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| ccilc_dst190_2012 | Distance to land-cover: Urban | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| ccilc_dst200_2012 | Distance to land-cover: Bare | 2012 ESA CCI annual LC maps v2.0.7 [34]; 10 arc seconds (≈300 m) * | 3 arc seconds (≈100 m) |
| cciwat_dst | Distance to water bodies | ESA CCI, Water bodies v4.0 [34]; 5 arc seconds (≈150 m) * | 3 arc seconds (≈100 m) |
| dmsp_2011 | Nighttime lights intensity | 2011 inter-calibrated version of the v4 DMSP-OLS Nighttime Lights Time Series [35]; 30 arc seconds (≈1 km) * | 3 arc seconds (≈100 m) |
| gpw4coast_dst | Distance to coastline | GPWv4 input administrative units [36]; 3 arc seconds (≈100 m) * | 3 arc seconds (≈100 m) |
| osmint_dst | Distance to road intersections | 2016 OSM highways [37] * | 3 arc seconds (≈100 m) |
| osmriv_dst | Distance to major water ways | 2016 OSM waterways [37] * | 3 arc seconds (≈100 m) |
| slope | Slope | 2000 Viewfinder Panoramas [38]; (≈100 m) * | 3 arc seconds (≈100 m) |
| topo | Elevation | 2000 Viewfinder Panoramas [38]; (≈100 m) * | 3 arc seconds (≈100 m) |
| tt50k_2000 | Travel time to populated places of 50,000 or more people | 2000 EC-JRC Travel time to major cities [39]; 30 arc seconds (≈1 km) * | 3 arc seconds (≈100 m) |
| urbpx_prp_1_2012 | Proportion of settlement pixels with a one cell radius | 2012 DLR Global Urban Footprint [40]; 0.4 arc seconds (≈12.5 m) & 2000 EC-JRC Global Human Settlement Layer [41]; 38 m * | 3 arc seconds (≈100 m) |
| hfacilities_dst | Distance to health center or hospital | 2001 UN-OCHA [42] | 3 arc seconds (≈100 m) |
| schools_dst | Distance to primary or secondary school | 2001 UN-OCHA [43] | 3 arc seconds (≈100 m) |
| npp_2012 | Annual net primary productivity | 2010 MODIS [44]; 30 arc seconds (≈1 km) | 3 arc seconds (≈100 m) |

* Spatial covariate was processed by the "Global High Resolution Population Denominators" Project.

The 2011 Namibia 20% census microdata sample is comprised of 36,137 individuals in 7536 conventional households selected at random from a complete census enumeration [28], and the DHS survey sample is comprised of 3316 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, were 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 leveraged the strengths of each dataset and took 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 Unweighted n (%) | DHS Unweighted n (%) | DHS Weighted n (%) |
|---|---|---|---|---|
| **Households** | **Oshikoto (N)** | **7475** | **705** | **817** |
| urban_rural | Urban | 1167 (15.6) | 113 (16.0) | 139 (17.1) |
| | Rural | 6308 (84.4) | 592 (84.0) | 678 (82.9) |
| structure | Durable floor | 2910 (38.9) | 281 (39.8) | 340 (41.6) |
| | Non-durable floor | 4551 (60.9) | 422 (59.9) | 475 (58.1) |
| | Missing/unknown | 14 (0.2) | 2 (0.3) | 2 (0.3) |
| fuel | Non-solid fuel | 1217 (16.3) | 141 (20.0) | 182 (22.3) |
| | Solid fuel | 6253 (83.6) | 562 (79.7) | 633 (77.4) |
| | Missing/unknown | 5 (0.1) | 2 (0.3) | 2 (0.3) |
| water | Improved water | 5388 (72.1) | 589 (83.6) | 688 (84.2) |
| | Unimproved water | 2045 (27.3) | 72 (10.2) | 80 (9.8) |
| | Missing/unknown | 42 (0.6) | 44 (6.2) | 49 (7.0) |
| toilet | Improved toilet | 1955 (26.1) | 207 (29.4) | 258 (31.6) |
| | Unimproved toilet | 5491 (73.5) | 492 (69.8) | 553 (67.6) |
| | Missing/unknown | 29 (0.4) | 6 (1.0) | 6 (0.8) |
| space | Adequate space | 6529 (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) |
| noedu | Head household—any education | 5797 (77.6) | 581 (82.4) | 677 (82.8) |
| | Head household—no education | 1528 (20.4) | 111 (15.7) | 125 (15.3) |
| | Missing/unknown | 150 (2.0) | 13 (1.9) | 15 (1.9) |
| any_u5 | No child under age 5 | 4267 (57.1) | 405 (57.5) | 478 (58.5) |
| | Any child under age 5 | 3208 (42.9) | 300 (42.5) | 340 (41.5) |
| **Individuals** | **Oshikoto (N)** | **36,137** | **3316** | **3576** |
| relationship | Head | 7475 (20.7) | 705 (22.5) | 817 (22.9) |
| | Spouse | 2391 (6.6) | 218 (7.0) | 250 (7.0) |
| | Child | 10,394 (28.8) | 785 (25.0) | 888 (24.8) |
| | Grandchild | 8635 (23.9) | 591 (18.9) | 660 (18.5) |
| | Extended | 5519 (15.3) | 622 (19.8) | 713 (19.9) |
| | Other | 1723 (4.8) | 215 (6.9) | 247 (6.9) |
| sex | Female | 18,814 (52.1) | 1669 (53.2) | 1899 (53.1) |
| | Male | 17,323 (47.9) | 1467 (46.8) | 1677 (46.9) |
| age | 0 | 1136 (3.1) | 87 (2.8) | 99 (2.8) |
| | 1–4 | 3968 (11.0) | 364 (11.6) | 414 (11.6) |
| | 5–9 | 4514 (12.5) | 404 (12.9) | 461 (12.9) |
| | 10–14 | 4895 (13.6) | 389 (12.4) | 435 (12.2) |
| | 15–19 | 4643 (12.9) | 385 (12.3) | 433 (12.1) |
| | 20–24 | 3284 (9.1) | 280 (8.9) | 323 (9.0) |
| | 25–29 | 2391 (6.6) | 213 (6.8) | 245 (6.9) |
| | 30–34 | 1912 (5.3) | 195 (6.2) | 230 (6.4) |
| | 35–39 | 1756 (4.9) | 161 (5.1) | 193 (5.4) |
| | 40–44 | 1371 (3.8) | 106 (3.4) | 120 (3.4) |
| | 45–49 | 1341 (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 | 1105 (3.1) | 98 (3.1) | 107 (3.0) |
| | 75+ | 1177 (3.3) | 95 (3.0) | 104 (2.9) |

### 2.3. Simulation

We generated realistic household membership with realistic household point location and demographic and social characteristics in the following three phases. In phase A, we defined household types and then predicted the spatial distribution of the types in Oshikoto using DHS data, spatial covariates, and visual inspection of satellite imagery. The output was a probability surface for each household type. In phase B, we generated the synthetic population using a census microdata sample and assigned the population to household point locations using the household type probability surfaces generated in phase A. Phase C involved prediction of additional population characteristics in each household. The code was written in R [46] and spatial data were 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.

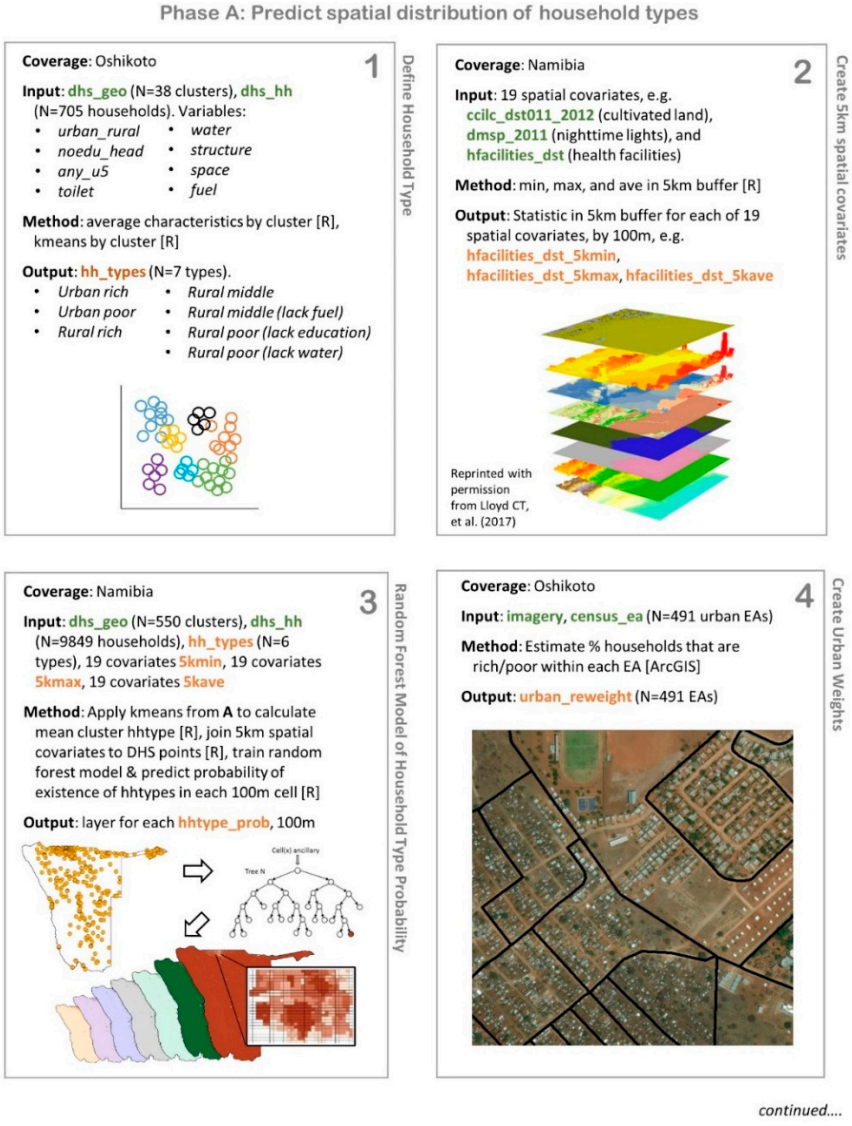



**Figure 2.** *Cont.*

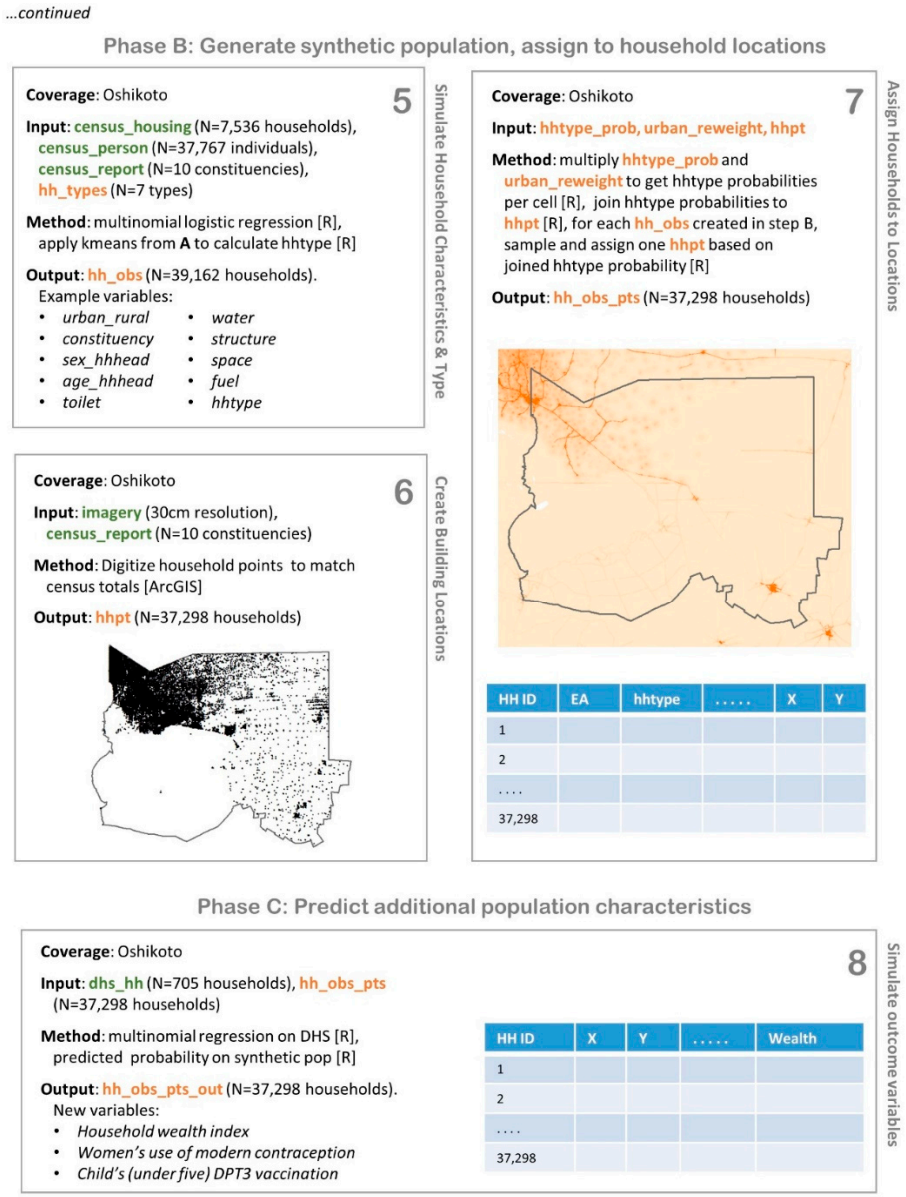



**Figure 2.** Simulation workflow with steps 1 through 8 organized in three phases. Green indicates original dataset, and orange indicates derived dataset.

### 2.3.1. Phase A: Predict Spatial Distribution of Household Types

Using the DHS dataset, we first defined realistic and distinct types of households present in Oshikoto based on the 2013 DHS data of 705 households. We used 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 minimizing the within group sum of squares. We then utilized 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 were easily interpretable. In the case of Namibia 2013 DHS, seven household types are identified. To interpret and label household types, we considered whether the household type values were above, below, or near the Oshikoto average (Table 3). We saved the k-means centroids and hierarchical clustering cut-off points to classify household types in other datasets in steps 3 and 5.

Second, we processed 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 were available for the whole region, enabling predictive mapping, and were shown to be related to population density [3,48]. We converted each covariate into a 100 m × 100 m raster, and then for each cell, calculated the minimum, maximum, and average values within a five kilometer buffer using WGS84 geographic projection. This five-kilometer moving window was used because the DHS data used to fit models in the next step were randomly geo-displaced up to five kilometers in rural areas. Further, the average covariate value within a five-kilometer buffer of a displaced DHS PSU location was closer to the real, non-displaced, unpublished covariate value than the published, displaced covariate value [49,50]. Although DHS PSU coordinates were only displaced up to two kilometers in urban areas, a five-kilometer buffer was used for all PSUs, and urban probability surfaces were 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 calculated the most common household type for each PSU using the k-means centroids and cut-off points. Next, we extracted the five-kilometer averaged spatial covariates created in step 2 to each DHS PSU location, resulting in 550 observations of household type linked to (19 × 3) 57 spatial covariates. In this step 3, we found a relationship between household type and buffered spatial covariates in order to predict household types over the whole region. To do this, we used 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 m by 100 m probability surface for each household type across Namibia.

Fourth, we manually created household type probabilities for urban EAs. This step was 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 was unrealistic. Poor and rich households are often segregated in urban areas worldwide [51], and visual inspection of satellite imagery indicates that socioeconomic segregation was present in Oshikoto's urban areas as well. From Step 1, we labeled the two urban household types as poor and rich, then manually assigned a proportion of households that we judged to be rich versus poor within each EA based on satellite imagery, such that the probabilities summed to 1. These manually created EA-level urban household type probabilities were multiplied by the predicted household type probability surfaces created in step 3 to create the final 100 m × 100 m household type probability surfaces.

**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 | 0.012 | 0.11 | 0.81 | 0.04 | 0.45 | 0.01 | 0.97 | Rural middle (lack fuel) |
| Type 6 | 1.00 | 0.012 | 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 | 0.016 | 0.12 | 0.77 | 0.11 | 0.60 | 0.07 | 0.79 | |

### 2.3.2. Phase B: Generate Synthetic Population and Assign Household Locations

Fifth, we simulated 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 calculated the proportion of households to simulate per household-size, per stratum (defined by constituency and urban/rural boundary). Second, we selected random resamples from the microdata until the number of target households was reached in each household size and strata. Third, demographic characteristics of the household members (*age*, *sex*, *relationship*) were replicated from the microdata. Fourth, we added household socioeconomic characteristics to the simulated dataset (*education*, *toilet*, *water*, *structure*, *space*, *fuel*) using a multinomial regression. This allowed for the simulation of combinations of demographic characteristics that existed in the population but were not present in the census microdata. For each simulated household, we assigned 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.

Sixth, the census microdata sample was provided with a weight equal to five for nearly all conventional households. We recalibrated these weights to the total number of households per constituency in the 2011 census [28]. However, this process could lead to too few observations in some constituency-urban/rural strata, and too many observations in other strata. Therefore, we increased the weights to simulate an extra 5% of households from which a random selection of households was assigned to latitude-longitude coordinates in step 7.

Seventh, we joined reweighted household type probabilities (100 m $\times$ 100 m grid cells) created in step 4 to the household latitude-longitude coordinates created in step 6. Finally, for each household simulated in step 5, we randomly sampled one latitude-longitude coordinate within the constituency-urban/rural strata based on the probability of household type. We repeated the assignments until all coordinates were assigned a simulated household, and then discarded the extra 5% unassigned simulated households.

#### 2.3.3. Phase C. Predict Additional Population Characteristics, Generalize Locations

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

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

where $K$ is the number of categories in the outcome variable, $Y_i$ is the outcome value for individual $i$, and $\boldsymbol{X}_i$ is a matrix of covariate values belonging to individual $i$. Model coefficients were applied to covariates of the 37,298 households in the simulated dataset to predict outcome values.

### *2.4. Assessment*

We conducted global assessments to evaluate whether each of the five realizations of the simulated population were realistic overall, and a local assessment to evaluate whether the realizations were realistic at an EA level. In the global assessment, we aggregated the DHS records to PSU and the simulated census records to EA, and graphically compared the distributions of simulated covariates and outcomes. We also mapped simulated census records by EA to visually inspect the spatial distributions across Oshikoto. In the local assessment, DHS data were 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 was selected within 5 km of each DHS PSU coordinate, then the same number of households as the observed DHS cluster were drawn from the simulated population. The characteristics were 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 km × 1 km grid cells 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.

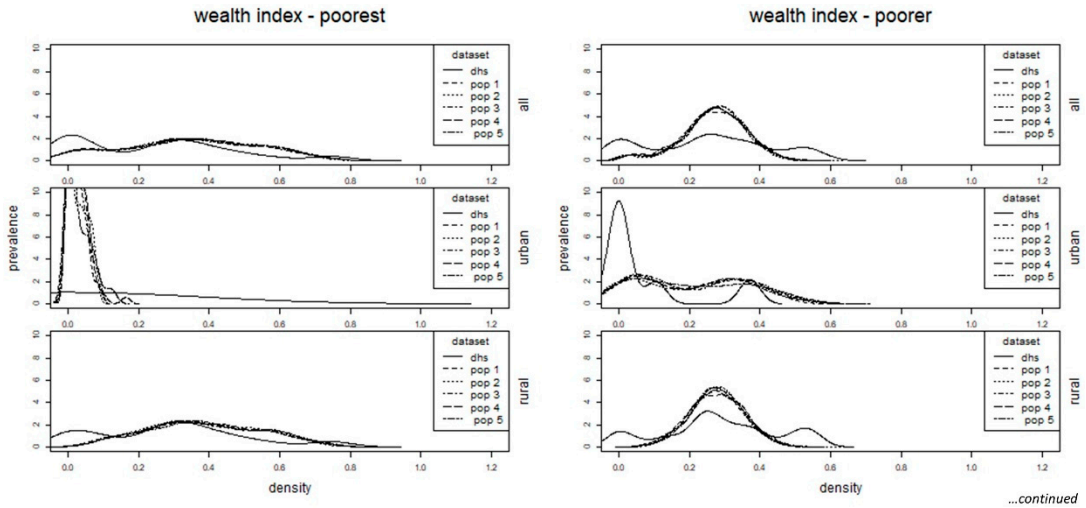
**Table 4.** Prevalence of demographic and socioeconomic characteristics in five realizations of the synthetic population in Oshikoto, Namibia.

| 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** |
| urban_rural | Urban | 84.3 | 84.3 | 84.3 | 84.3 | 84.3 |
| | Rural | 15.7 | 15.7 | 15.7 | 15.7 | 15.7 |
| structure | 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 |
| fuel | 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 |
| water | Improved water | 73.2 | 73.2 | 72.9 | 73.1 | 72.7 |
| | Unimproved water | 26.8 | 26.8 | 27.1 | 26.9 | 27.3 |
| toilet | Improved toilet | 20.1 | 20.1 | 19.9 | 19.7 | 19.5 |
| | Unimproved toilet | 79.9 | 79.9 | 80.1 | 80.3 | 80.5 |
| space | Adequate space | 92.5 | 92.2 | 92.3 | 92.5 | 92.3 |
| | Inadequate space | 7.5 | 7.8 | 8.7 | 7.5 | 7.7 |
| noedu | 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 |
| any_u5 | 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 |

**Table 4.** *Cont.*

| Variable | Category | pop_1 (%) | pop_2 (%) | pop_3 (%) | pop_4 (%) | pop_5 (%) |
|---|---|---|---|---|---|---|
| **Individuals** | **Oshikoto (N)** | **179,931** | **179,854** | **180,233** | **180,164** | **180,111** |
| relationship | 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 |
| sex | Female | 52.2 | 52.0 | 51.9 | 51.8 | 52.0 |
| | Male | 47.8 | 48.0 | 48.1 | 48.2 | 48.0 |
| age | 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 |

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




...continued

**Figure 3.** *Cont.*

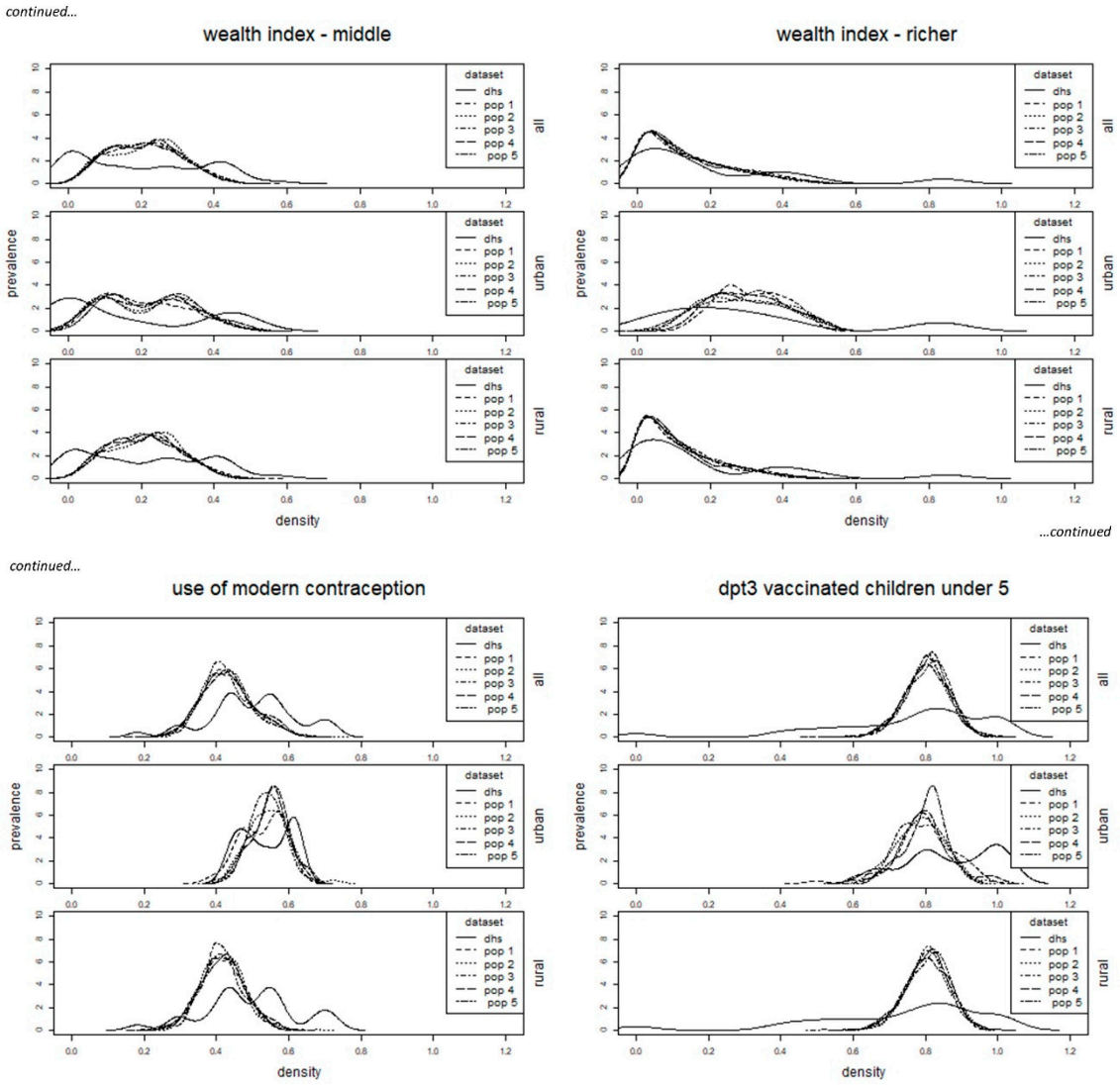



**Figure 3.** Comparison of prevalence (*y*-axis) distributions (*x*-axis) in the 2013 Namibia DHS for Oshikoto region (solid line) and five synthetic population realizations (dotted lines) across wealth categories, contraceptive use, and DTP3 vaccination.

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

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 (Figure 5). This implied that the observed DHS results could potentially have been drawn from the synthetic population.

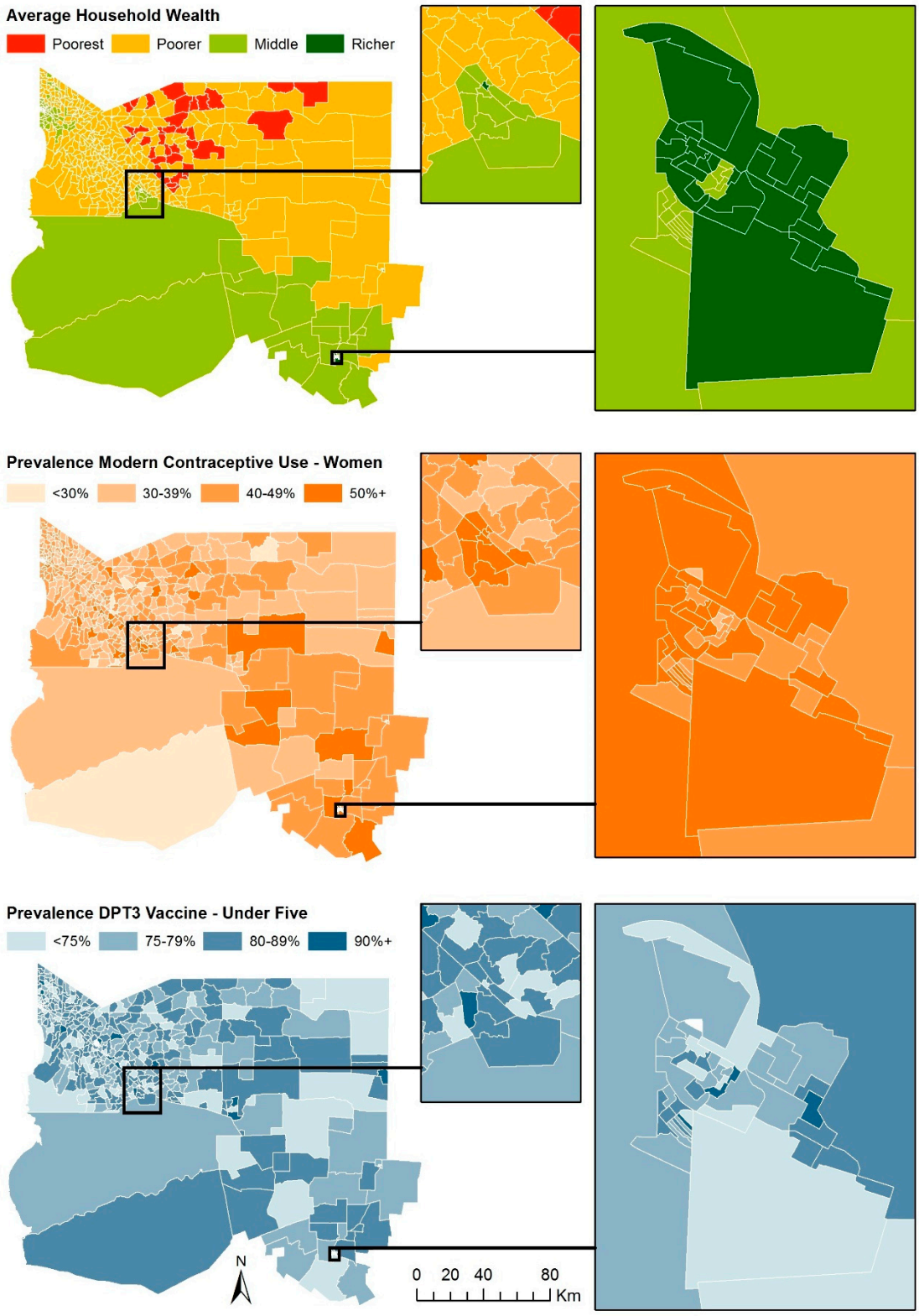



**Figure 4.** Maps of average household wealth level, conceptive use, and DPT3 vaccination by EA in one simulated population (synth_pop_1) in Oshikoto, Namibia.

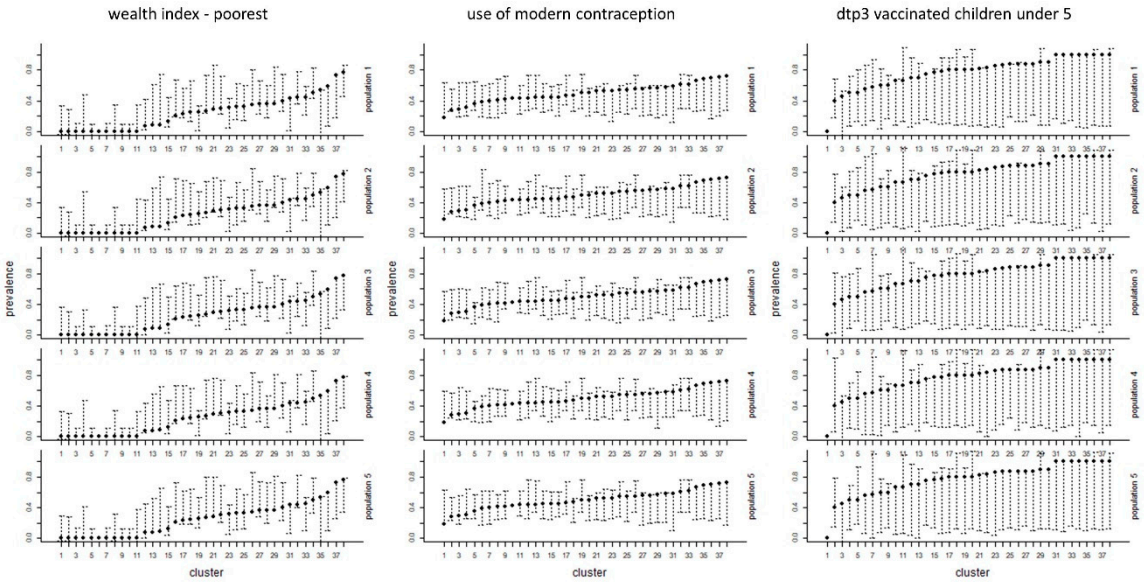



**Figure 5.** 2013 Namibia DHS PSU-level estimates (black dot) of average household wealth level, contraceptive use, and DTP3 vaccination versus the distribution of estimates from 100 samples (bar) selected in EAs located within 5 m of the DHS PSUs for five synthetic population realizations.

## 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 used the strengths of two commonly available population datasets, namely household surveys and census microdata samples. We also drew together computational methods in microsimulation of individuals and households with high-resolution mapping of household characteristics using geospatial data. The result was 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 m × 100 m 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 m × 100 m grid cells provides a first opportunity for this kind of accuracy assessment.

One limitation of this approach is it resulted in some spatial smoothing of household type probabilities due to the use of buffered covariates in the Random Forest model in step 3. The choice, however, was to either introduce substantial measurement error by training models on covariates at the location of geo-displaced DHS coordinates [50], or to reduce spatial precision in prediction of household type probabilities by using aggregated covariate values within a buffer region around the geo-displaced DHS coordinates. We opted for the latter approach because rural Oshikoto was sparsely populated, and thus we expected minimal impact of spatial imprecision of household type. Furthermore, we manually corrected the spatial distribution of urban household type probabilities during step 4 via manual inspection of satellite imagery and classification of census EAs. However, researchers applying these methods in more densely populated and/or heterogeneous settings might consider smaller buffers in urban area.

A related consideration when extending these methods is that Random Forest models cannot extrapolate beyond the range of the training data [3]. This could impact the accuracy of the prediction of household type maps. To ensure accuracy of predicted household types in step 3, the same geographic unit should be used in both training and prediction datasets (e.g., 5 km buffers), and the range of covariate values in the training data (e.g., all Namibia DHS clusters) must be similar or larger than the range in covariate values in the region of study (e.g., Oshikoto). We checked that training locations had a wider range of covariate values than the Oshikoto household locations (see Supplement 3).

A second limitation of this work is that it relied on manually digitized building point locations, and delineation of rich versus poor EAs in urban areas. Manual data creation 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 and to classify urban census EAs in Oshikoto for this analysis. However, as coverage of publicly available sub-meter 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. Furthermore, machine learning techniques are showing promise in mapping neighborhood types from very high-resolution imagery [60] and other building datasets [61], which can also help address this limitation.

A third limitation is potential errors introduced by temporal differences between datasets. The census and DHS datasets were collected two years apart, so major differences in population totals or demographic distributions were not expected; however, several covariates related to roads, travel time, facilities, and topography were more than a decade older and might not reflect most recent development. Furthermore, household point locations were digitized from more recent imagery, and thus might include new buildings not reflected in the spatial covariates. The predictive model may be improved with better temporal alignment to covariate data.

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 m × 100 m). 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 http://www.mdpi.com/2306-5729/3/3/30/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. Output S3: Interim graphs and maps used to assess accuracy of a simulated, geo-located population in Oshikoto, Namibia.

**Author Contributions:** Conceptualization, D.R.T., L.K., and W.C.J.; Methodology, L.K. and W.C.J.; Formal Analysis, L.K.; Data Curation, D.R.T. and L.K.; Writing—Original Draft Preparation, D.R.T.; Writing—Review & Editing, L.K., and W.C.J.

**Funding:** This research received no external funding.

**Acknowledgments:** We would like to thank Jeremiah J. Nieves for assembling and sharing the WorldPop geospatial datasets used in this study. The geospatial datasets were produced by David Kerr, Heather Chamberlain, Chris T. Lloyd, Maksym Bondarenko (WorldPop, University of Southampton), Gregory Yetman, and Linda Pistolesi (Center for International Earth Science Information Network, Columbia University) in the framework of the WorldPop "Global High Resolution Population Denominators" Project funded by the Bill & Melinda Gates Foundation (OPP1134076). We would also like to thank the anonymous reviewers for their helpful comments which helped to improve the manuscript.

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

## References

1. 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*, 226–234. [CrossRef]
2. Oak Ridge National Laboratories. LandScan Documentation. Available online: http://web.ornl.gov/sci/landscan/landscan_documentation.shtml (accessed on 6 February 2017).
3. 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*, e0107042. [CrossRef] [PubMed]
4. 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*, 5635–5655. [CrossRef]
5. Hay, S.I.; Noor, A.M.; Nelson, A.; Tatem, A.J. The accuracy of human population maps for public health application. *Trop. Med. Int. Health* **2005**, *10*, 1073–1086. [CrossRef] [PubMed]
6. 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*, 87–97. [CrossRef] [PubMed]
7. 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*, 383–407. [CrossRef]
8. Minnesota Population Center. *Integrated Public Use Microdata Series, International: Version 7.0 [Dataset]*; University of Minnesota: Minneapolis, MN, USA, 2018.
9. Global Health Data Exchange (GHDx). Available online: http://ghdx.healthdata.org/ (accessed on 2 February 2017).
10. Tanton, R. A review of spatial microsimulation methods. *Int. J. Microsimul.* **2014**, *7*, 4–25. [CrossRef]
11. Birkin, M.; Clarke, M. The generation of individual and household incomes at the small area level using synthesis. *Reg. Stud.* **1989**, *23*, 535–548. [CrossRef]

12. 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. [CrossRef]
13. 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. [CrossRef]
14. 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*; Tanton, R., Edwards, K., Eds.; Understanding Population Trends and Processes, volume 6; Springer: Dordrecht, The Netherlands, 2012; pp. 105–125.
15. Templ, M.; Meindl, B.; Kowarik, A.; Dupriez, O. Simulation of synthetic complex data: The R package simPop. *J. Stat. Softw.* **2017**, *79*, 1–38. [CrossRef]
16. Macal, C.M. Everything you need to know about agent-based modelling and simulation. *J. Simul.* **2016**, *10*, 144–156. [CrossRef]
17. 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*, 1–17. [CrossRef]
18. Heppenstall, A.; Malleson, N.; Crooks, A. Space, the final frontier: How good are agent-based models at simulating individuals and space in cities? *Systems* **2016**, *4*, 9. [CrossRef]
19. Synthetic Populations and Ecosystems of the World (SPEW). Available online: http://www.stat.cmu.edu/~spew/about/ (accessed on 15 May 2018).
20. Synthetic Household Population™. Available online: https://www.rti.org/impact/synthpop (accessed on 15 May 2018).
21. SDG Indicators: Revised List of Global Sustainable Development Goal Indicators. Available online: https://unstats.un.org/sdgs/indicators/indicators-list/ (accessed on 3 September 2017).
22. Tatem, A.J. WorldPop, open data for spatial demography. *Sci. Data* **2017**, *4*, 170004. [CrossRef] [PubMed]
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*, 20160825. [CrossRef] [PubMed]
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. [CrossRef] [PubMed]
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*, 1583–1591. [CrossRef] [PubMed]
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*, 25. [CrossRef] [PubMed]
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 on 3 March 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.; Government of Namibia: Windhoek, Namibia, 2013.
30. [Namibia] National Statistics Agency 2011 Census EA Boundaries. Available online: https://digitalnamibia.nsa.org.na/ (accessed on 19 February 2018).
31. ICF International Available Datasets. Available online: https://dhsprogram.com/data/available-datasets.cfm (accessed on 15 November 2017).
32. Digital Globe Quickbird 50 cm Imagery. Available online: http://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9 (accessed on 1 February 2018).
33. Lloyd, C.T.; Sorichetta, A.; Tatem, A.J. High resolution global gridded data for use in population studies. *Sci. Data* **2017**, *4*, 1–17. [CrossRef] [PubMed]
34. European Space Agency (ESA). Climate Change Initiative (CCI) Products. Available online: http://maps.elie.ucl.ac.be/CCI/viewer/download.php (accessed on 19 February 2017).
35. Zhang, Q.; Pandey, B.; Seto, K.C. A robust method to generate a consistent time series from DMSP/OLS nighttime light data. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5821–5831. [CrossRef]

36. CIESIN Gridded Population of the World, Version 4 (GPWv4). Country-Level Information and Sources Revision 10. Available online: http://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-country-level-summary-rev10.xlsx (accessed on 19 February 2017).
37. Open Street Map Base Data. Available online: www.openstreetmap.org (accessed on 19 February 2017).
38. de Ferranti, J. Digital Elevation Data: SRTM Void Fill. Available online: http://www.viewfinderPanoramas.org/voidfill.html (accessed on 19 February 2017).
39. Nelson, A. *Estimated Travel Time to the Nearest City of 50,000 or More People in Year 2000*; Global Environment Monitoring Unit—Joint Research Centre of the European Commission: Ispra, Italy, 2008. Available online: http://forobs.jrc.ec.europa.eu/products/gam/ (accessed on 8 August 2018).
40. Esch, T.; Marconcini, M.; Felbeir, A.; Roth, A.; Heldens, W.; Huber, M.; Schwinger, M.; Taubenböck, H.; Müller, A.; Dech, S. Urban footprint processor—Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 1617–1621. [CrossRef]
41. European Commission. Global Human Settlement Layer. Available online: http://ghsl.jrc.ec.europa.eu/faq.php (accessed on 6 February 2017).
42. UN-OCHA-ROSA Namibia—Health Facilities. Available online: https://data.humdata.org/dataset/namibia-health (accessed on 19 February 2017).
43. UN-OCHA-ROSA Namibia—Education Facilities. Available online: https://data.humdata.org/dataset/namibia-education-0 (accessed on 19 February 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*, 547–560.
45. Fink, G.; Günther, I.; Hill, K. Slum residence and child health in developing countries. *Demography* **2014**, *51*, 1175–1197. [CrossRef] [PubMed]
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*, 20170401. [CrossRef] [PubMed]
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, USA, 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*, 135–153. [CrossRef] [PubMed]
51. UN Habitat. Urbanization and development: Emerging futures. In *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; 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*; Domingo-Ferrer, J., Magkos, E., Eds.; Lecture Notes in Computer Science, volume 6344; 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 on 16 April 2018).
55. United Nations Children's Fund (UNICEF). Multiple indicator cluster surveys round 4 (MICS4). In *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 Health* **2016**, *93*, 526–537. [CrossRef] [PubMed]
59. A Breakthrough in Building Footprint Extraction. Available online: http://explore.digitalglobe.com/GBDX-Building-Footprints.html (accessed on 15 May 2018).

60. Graesser, J.; Cheriyadat, 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*, 1164–1176. [CrossRef]
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. [CrossRef] [PubMed]