Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00670-0 OPEN

Semi-automatic mapping of pre-census enumeration areas and population sampling frames

Sarchil Qader[1,2✉], Veronique Lefebvre[3], Andrew Tatem[1,3], Utz Pape[4], Kristen Himelein[4], Amy Ninneman[3], Linus Bengtsson[3] & Tomas Bird[3,5]

Enumeration Areas (EAs) are the operational geographic units for the collection and dissemination of census data and are often used as a national sampling frame for various types of surveys. In many poor or conflict-affected countries, EA demarcations are incomplete, outdated, or missing. Even for countries that are stable and prosperous, creating and updating EAs is one of the most challenging yet essential tasks in the preparation for a national census. Commonly, EAs are created by manually digitising small geographic units on high-resolution satellite imagery or physically walking the boundaries of units, both of which are highly time, cost, and labour intensive. In addition, creating EAs requires considering population and area size within each unit. This is an optimisation problem that can best be solved by a computer. Here, for the first time, we produce a semi-automatic mapping of pre-defined census EAs based on high-resolution gridded population and settlement datasets and using publicly available natural and administrative boundaries. We demonstrate the approach in generating rural EAs for Somalia where such mapping is not existent. In addition, we compare our automated approach against manually digitised EAs created in urban areas of Mogadishu and Hargeysa. Our semi-automatically generated EAs are consistent with standard EAs, including having identifiable boundaries for field teams to follow on the ground, and appropriate sizing and population for coverage by an enumerator. Furthermore, our semi-automated urban EAs have no gaps, in contrast, to manually drawn urban EAs. Our work shows the time, labour and cost-saving value of automated EA delineation and points to the potential for broadly available tools suitable for low-income and data-poor settings but applicable to potentially wider contexts.

[1] WorldPop, Geography and Environment, University of Southampton, Southampton SO17 1BJ, UK. [2] Natural Resources Department, College of Agricultural Engineering Science, University of Sulaimani, Sulaimani, Iraq. [3] Flowminder Foundation, Stockholm, Sweden. [4] World Bank, Washington, USA. [5] Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada. ✉email: S.Qader@soton.ac.uk

Introduction

Globally, population data count and distribution data in the form of census enumerations, are used for population segmentation, planning, and a myriad of other functions that support the government. EAs are the operational geographic units for the collection of census data (UN, 2007). The set of all EAs of a country constitutes a partition of that country, with EAs not overlapping with each other. In principle, EAs are designed such that each unit contains a similar population size and conforms to certain constraints imposed by the logistics of counting large numbers of people with limited resources. Regardless of being created manually or digitally, the design of EAs should take several criteria into account including (i) be mutually exclusive (none-overlapping) and exhaustive (cover the entire country), (ii) have boundaries that are easily identifiable on the ground, (iii) be consistence with the administrative boundary hierarchy, (iv) be compact without pockets or disjoint sections, (v) have populations of approximately equal size (vi) be small enough and accessible to be covered by an enumerator within the census period, (vii) be small and flexible enough to allow the widest range of tabulations for different statistical reporting units, (viii) be large enough to guarantee data privacy, (ix) be useful for other types of censuses and data collection activities, (x) be satisfactory to the needs of government departments and other data users (BUCEN, 1978; Unite, 2000). In some countries, census data can be outdated or incomplete in which EAs might not be available or need to be updated. Creating and updating digitised EAs in the preparation for a national census is among the most complex and massive peacetime exercises. However, with the availability of data on population density at a sufficiently high spatial resolution (e.g., 100 m), the process of designing EAs can be semi-automated, which could accelerate the census mapping process.

In some LMIC contexts, and particularly in conflicted affected areas, EAs are based on old population data or do not exist, a fact that has far-reaching consequences. In particular, the lack of properly defined EAs means that there is no nationally representative sampling frame. National sampling frames are used to draw representative samples from the population to understand the geographic distribution of population characteristics (Turner, 2003). Without sample weights in which an accurate national sampling frame is an essential input, survey data sampled from a country's population are likely to be inefficient, typically undersampling vulnerable populations (Thomson et al., 2012; Ellard-Gray et al., 2015). Therefore, the creation of EAs to form a sampling frame creates significant benefits to generate nationally representative data to allow evidence-based analysis informing governments and NGOs. Especially in resource constrained environments (like Fragile, Conflicts and Violence (FCV) settings), it is best to have an up-to-date population sampling frame because it increases efficiency and therefore reduces costs.

Historically, the creation of census EAs has been an expensive task solved by approaches such as physically walking to map EA boundaries, which can require intensive resources and multiple years to complete (Lu, 2009; Yacyshyn and Swanson, 2011). For instance, in Zambia, the 2010 census mapping exercise was expected to take about two years to be completed at a total cost of US $7 million (Yacyshyn and Swanson, 2011). Since the advent of Geographic Information Systems (GIS) and high-resolution satellite photography census cartographers in some countries have been able to manually digitise EA boundaries from satellite imagery, leading to better control of EA delineation. However, manual digitisation of EAs is still costly and labour-intensive. It is also prone to human error and can have poor accuracy (Alazab et al., 2009; Balinski et al., 2010; Cockings et al., 2011). In addition, in the manually digitise EA approach, it can be challenging to accommodate population and area constraints. Furthermore, EAs or sampling frames in countries with large population displacements and rapid urbanisation need regular updating, meaning that the manual digitisation process must be replicated. Combining automation and manual approaches could be a viable middle ground to accelerate this process.

Several tools are available for creating a population sampling frame from gridded population estimates such as GridSample (Thomson et al., 2017) and Geo-sampling tool (Cajka et al., 2018). Some approaches have used remote sensing data and GIS techniques to generate homogeneous regions and spatial sample design (Kumar, 2007; Lang et al., 2014; Wang et al., 2019). Other works have employed statistical region merging techniques to group areas into zones for a range of purposes including investigation of neighbourhood effects on health and release of census data (Cockings et al., 2011; Flowerdew et al., 2007; Haynes et al., 2007). Statistical region merging uses algorithms to segment an image into regions of similar intensity or colour (Nock and Nielsen, 2004). For instance, Folch and Spielman (2014) showed the advantage of applying the improved max-p algorithm on growing the irregular regions from census EAs. The max-p was also introduced as a new spatially constrained clustering problem in which a set of geographic areas will be clustered into the maximum number of homogeneous regions such that the value of a spatially extensive regional attribute is above a predefined threshold value (Duque et al., 2012). The challenge with such approaches is that the derived EAs may not align with landmarks such as roads and buildings, which are needed by enumerators on the ground. By contrast, the ArcGIS/AZTool toolkit, a region-merging tool, was developed to design new reporting geographic units using the 2006 census data and existing EAs as their bas unit (Martin and Lyndon, 2009). The AZTool approach has the advantage of using existing EAs which are assumed to align with known geographic features. However, in a lower and middle-income countries where existing EA data may be outdated or even paper-based, making them unsuitable for region-merging

While suitable EA-level data may not be available in many LMIC contexts, many new and freely available data sources have recently emerged, which can directly inform the creation of EAs. Two basic ingredients are needed. First, reliable spatial data on suitable infrastructure and environment elements are needed to inform boundaries of EAs. For this, global, road data from OSM are available. OSM data are ~83% complete and more than 40% of countries (including several in the developing world) have a fully mapped street network (Barrington–Leigh and Millard–Ball, 2017). As well, other features such as rivers, elevation and buildings can often be used to provide greater detail on fine-scale landscape features. The second required ingredient is an estimate of population density, to inform suitable sizing for EAs, such that they can be covered by enumerators in a suitable timeframe. For this element, global high-resolution population models such as WorldPop (WorldPop, 2019a) provide estimates of population density are available in many contexts and can be used to help inform the creation of EAs.

Here for the first time, we describe how these freely available data on population and georeferenced features can be combined to design a new full set of pre-defined census EAs, using the example of Somalia to demonstrate the process.

Methodology

We applied the semi-automated EA delineation process to build a national sampling frame for Somalia. This section describes our approach then details its application in the context of Somalia.

General approach. Our process automatically creates EAs within user-defined ranges of population size and area for an entire country, combining Our proposed automated process is based on a 'split and merge' methodology inspired from the field of image processing, specifically image segmentation using mathematical morphology (e.g., watershed and waterfall algorithms). The process first combines all sources of vector data to split the country into small sub-areas as small as possible that follow visible boundaries (e.g., roads, waterways), as well as available administrative boundaries, then progressively re-merge them so that they are as large as possible while respecting user-defined constraints on the area and population size. As well, the process defines a number of 'hard' boundaries across which regions cannot be merged, such as administrative units. As a consequence, the resulting EAs do not cut cross buildings and have boundaries that can be seen from the ground by enumerators. In addition, at each merging move, the compactness criterion is calculated and tested in order to merge with the neighbour region to avoid creating complex shapes. In regions where data on geographic and manmade features are too sparse to obtain small enough regions after the splitting stage (either non-existent e.g., desertic areas or unmapped), a quadtree algorithm is used to further split the country. In this case, not all automated EAs follow visible boundaries, however quadtree-derived areas can then be further subdivided through a manual process (Qader et al., 2020). A visual description of this process is given in Fig. 1.

Case study area: Somalia. Somalia is situated in the Horn of Africa with an official population estimated at 12.3 million in 2014, up from the 1975 estimate of 4.1 million (UNFPA, 2016). In 2012, the first nationwide Population Estimation Survey (PESS) took place. The Somalian Government, United Nations Population Fund (UNFPA), and United Nations Development Programme (UNDP) collaborated, prepared, and carried out this survey, aiming to use the PESS as a basis for a census. The PESS survey in 2014 estimated that 42% of the population was permanently settled in urban areas and 23% in rural areas, while 26% were nomadic people and 9% were internally displaced persons (IDPs) (UNFPA, 2014). The Somali population is rapidly increasing with almost 3% population growth per year and a high fertility rate of 6.26 children per woman, which is the fourth highest in the world (Gure et al., 2015).

However, the results of the PESS alone were not suitable for creating a nationally representative sampling frame, as the PESS only created EAs in urban areas. The risks associated with fieldwork and the lack of funding were just two hurdles faced by the PESS. As a consequence, significant displaced populations exist in parts of Somalia, without any official population information available. Rebuilding Somalia's statistical infrastructure and capacity is key in supporting resilience efforts and proposed a spatial analysis approach as an innovative way to create a new sampling frame, especially given the barriers in this context.

Data sources. To conduct this work, several datasets have been compiled and combined from various sources (Table 1).

Data pre-processing

Definition of urban and rural stratum within 18 pre-war regions. Defining major strata is the first step in generating EAs as the maximum population size and area constraints of EAs may need to vary in different strata, or adhere to existing administrative segmentation. For example, Urban versus Rural strata typically need very different constraints to account for the differing population densities and landscape features. In Somalia, urban strata were defined using the previous urban EAs from PESS 2014 (UNFPA, 2014). The previous urban EAs were dissolved using the dissolve tool in ArcGIS. The remaining area outside of the urban strata was considered rural. To define and compute the urban and rural strata for each Somalia pre-war region, urban and rural strata were intersected with the 18 pre-war regions administrative boundary. Based on expert opinion (Philip Rothberg, *personal communication*) the urban and rural strata in Banadir were merged and considered as urban since the region is almost urban.

National definitions of urban and rural areas vary significantly from one country to another, therefore, comparing these areas across national borders is difficult. Many countries rely on maximum population size to define urban areas whereas in other countries the urban areas are defined by administrative boundaries. For instance, the World Urbanisation Prospects (2018) uses minimum population size, either exclusively or in combination with other criteria or indicators to define urban and rural areas in 233 countries and areas (UN DESA P.D., 2018). In this project, regardless of population density, it was suggested by World Bank that the manual digitisation PESS 2014 urban areas must be used to define urban areas. The outline of the urban EAs is not always in line with the edge of the cities. In some cases, the outline of a city may not cover the entire city, or some densely populated areas were not considered as urban areas. In other cases, the outline of the cities included rural areas. The misclassification of urban and rural areas might have negative consequences on the construction of the EAs. It may either produce EAs with a small population size in urban areas or complicate EA creating in rural areas. The complication in rural areas could be because of reducing the flexibility in the merging process or further datasets may be required to split the area.

Settlement boundaries. Data on settlement locations and boundaries provide valuable data to inform EA delineation in the 'merge' part of the algorithm and are also used to create our refined 100 m population density estimates. Recent increased availability of high-resolution satellite imagery and high-power computing resources with adequate image processing algorithms have advanced the development of high-resolution human settlement layers (Vijayaraj et al., 2007; Florczyk et al., 2016; Roy et al., 2018). Examples of recently developed human settlement layers include Global Urban Footprint (GUF) (Esch et al., 2017), Global Human Settlement Layer (GHSL) (Pesaresi et al., 2013) and LandScan Settlement Layer (LandScan SL) (Cheriyadat et al., 2007). These datasets either are not available for Somalia or are incomplete for the country or are inappropriate for Somalia since the country is facing a constant regional instability and severe drought that forced substantial migration within the region. The existing available settlement datasets such as GUF and GHSL for Somalia at the time of this work were overlaid against high-resolution satellite imagery to assess their quality. It was found that these datasets miss many settlements across Somalia particularly in a rural context. In addition, during this work, several new datasets on settlements were available from a variety of sources (Table 1), each of which contains unique georeferenced settlements. We combined these diverse datasets into an improved high-resolution settlement map. Details on the processing done on each dataset can be found in the supplementary information Section 1. The resulting settlement map was used to improve population density predictions and EA delineation (Fig. 2a).

High-resolution population density estimates for Somalia. A high spatial resolution population map (100 m × 100 m, in this case) is necessary to estimate the population of each sub-region created

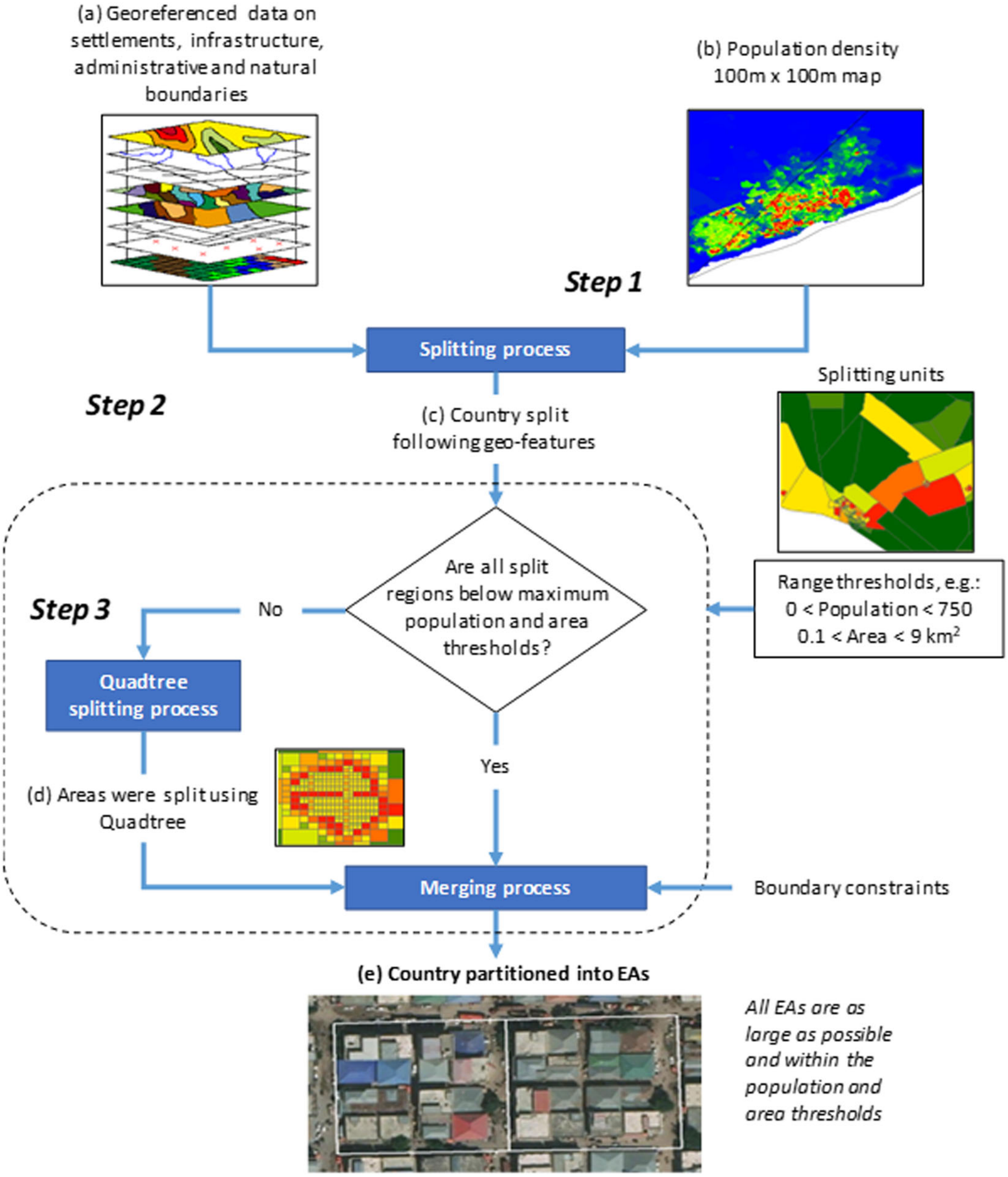
(a) Georeferenced data on settlements, infrastructure, administrative and natural boundaries
(b) Population density 100m x 100m map
Step 1
Splitting process
Step 2
(c) Country split following geo-features
Splitting units
Are all split regions below maximum population and area thresholds?
Range thresholds, e.g.: 0 < Population < 750 0.1 < Area < 9 km²
Step 3
No
Quadtree splitting process
(d) Areas were split using Quadtree
Yes
Merging process
Boundary constraints
(e) Country partitioned into EAs
All EAs are as large as possible and within the population and area thresholds

Fig. 1 Schematic diagram of our semi-automated EA delineation process. The first step is to split the territory into regions as small as possible using georeferenced features such as administrative and natural boundaries (e.g. rivers, terrain), settlement location and outlines, and road and path networks (**a**). We then compute the estimated population in each region thus obtained (using a very high spatial resolution population density map) (**b**) and check if all regions have a population and an area below given thresholds. If not, we further split regions using the quadtree algorithm, until all regions are below the population and area thresholds. We then merge regions so that they exceed the given minimum area threshold and until they are as close as possible to but remain below the maximum population and area thresholds. The merging process does not merge regions across a set of specified boundaries (e.g. administrative boundaries and large rivers). The result is a partition of the country into EAs that follow visible boundaries, that are not across obstacles or administrative boundaries, and that comply with given ranges of population size and area.

during the automated EA delineation process. Multiple high to moderate resolution global modelled population datasets are freely accessible to download including WorldPop (WorldPop, 2019), Global Rural-Urban Mapping Project, Version 1 (GRUMP) (CIESIN, 2011), Gridded Population of the World Version 4 (GPWv4) (CIESIN, 2017), Gridded Population of the World, United Nation (Azar et al., 2013) and Global Human Settlement Population Grid (GHS-POP) (JRC and CIESIN, 2015). However, none of these datasets on their own was sufficient for our purposes as they were created without the use of the PESS 2014 population data, or the final total population was not adjusted to match the PESS regional total. In addition, we had access to more recent datasets (high-resolution DigitalGlobe population estimates), which we wanted to use to ensure that our EA delineation of Somalia is based on the most up to date population estimates. Therefore, we produced a 100 m × 100 m population density map to calibrate our EA delineation. We give below an only succinct overview of the method employed as the generation of accurate population surfaces was beyond the scope of this paper and it does not influence the description of our novel approach for EA delineation. We used multiple data sources to create population surfaces for rural and urban areas, including information on building density, household density and population density (see

Table 1 Data used for somalia enumeration areas.

Dataset	Data used	Source	Date
Road data (OSM)	Lines	www.openstreetmap.org	2016
Waterway (OSM)	Lines		2016
River (OSM)	Lines		2016
Residential area (OSM)	Points and polygons		2016
Building (OSM)	Polygons		2016
'places', 'hamlet', and 'villages' (OSM)	Points		2016
Waterbody (OSM)	Polygons		2016
DLR Global Urban Footprint (GUF)	Binary raster	(Esch et al., 2017)	2016
BMGF/DigitalGlobe (DG) population estimates	Scatter points	(BMGF/DigitalGlobe)*	2015
World Bank/Flowminder/WorldPop building counts from Google Satellite imagery	Polygons	(World Bank/Flowminder/WorldPop)*	2016
UNFPA/PESS urban Enumeration Areas	Polygons	(World Bank)*	2014
UNFPA/PESS urban Enumeration Areas (EAs) household number	# households per EA.	(World Bank)*	2014
BMGF/DigitalGlobe (DG) settlement outlines for North Somalia	Polygons	(BMGF/DigitalGlobe)*	2015
UNFPA/PESS rural population estimates	Points	(World Bank)*	2014
Pre-war regions boundary	Polygons	OCHA, HDX (https://data.humdata.org/dataset/somalia-administrative-boundaries)	2016

The data that are marked with asterisks (*) cannot be shared
OSM an open street map, *BMGF* Bill and Melinda Gates Foundation, *UNFPA* United Nations Population Fund.

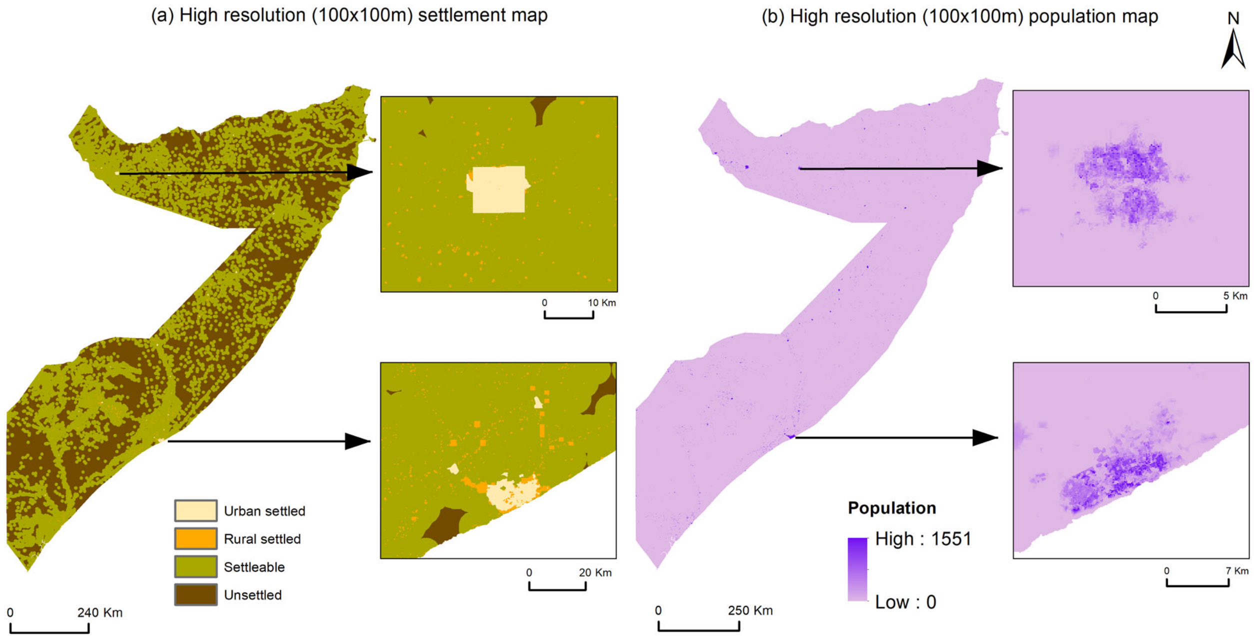

Fig. 2 Delineated urban, rural settled, settleable, and unsettled areas, high gridded population estimates, in Somalia. The high-resolution settlement map (100 m × 100 m) (**a**) represents the spatial distribution of different settlement types across Somalia and it has been obtained from transformation and compilation of multiple data sources. **b** High resolution population density map provides total population estimates per 100 m^2. The detail on the process done on each dataset to generate both products can be found in the supplementary information section 1 and 2.

supplementary information Table S2, S3 and S5). We used the World Bank survey (World Bank, 2017) to estimate a median number of people per building and per household to approximate population density from data on building and household densities (see supplementary information Table S2). In places lacking data but identified as settled we modelled population density based on the distribution of population estimates in similar settlements. We then set population density to zero in locations known to be not settled, and to a low value in locations that could be settled but for which we have no data (around known settlements). Finally, we rescaled the population density map thus obtained using the PESS 2014 regional totals (UNFPA, 2014) (Fig. 2b). The detail on the processing done on each dataset to generate high resolution gridded population data can be found in the supplementary information Section 2.

Split and merge algorithm for the creation of Enumeration Areas (EAs)

Splitting process. The aim of the splitting process is to partition the country into regions that are as small as possible so that the subsequent merging process has enough flexibility to combine them into optimal EAs. We used three steps to do this.

Step 1–Splitting based on geo-referenced features to create regions with tangible boundaries:. The country was split using road data, rural settlement boundaries, waterway and river data and administrative boundaries from OpenStreetMap (OSM), using the feature to polygons tool in ArcGIS. These datasets where either lines or polygons, whose geometry will be used to create area features, were the input features to the tool. From the merging of these features, each small "closed" area became a feature in the output feature class (here called 'Primary Units' (PU). This first step results in a set of fully contiguous units that are much smaller than the target EA size, with no gaps or islands and with all regions delineated by georeferenced features. If the road data is complete, then the process ensures that no building will be cut.

Step 2–Estimate population and area. With the primary units feature set defined, we were able to then compute the population size for each primary unit using the high-resolution gridded population datasets in R software.

Optional step 3–Splitting based on Quadtree algorithm:. Some areas do not contain sufficient line data to create small enough regions to meet the 9 km^2 area threshold or the 750 person population threshold—this was particularly true in sparse and non-populated areas, which remained larger than 9 km^2, in spite of having a population below 750. We further split any areas still larger than 9 km^2 or containing more than 750 people after steps 1 into square grid cells using a quadtree algorithm (Finkel and Bentley, 1974). A quadtree is a tree data structure in which each internal node in the underlying tree has exactly four children (Wanderer, 2017). This approach is commonly employed to partition a two-dimension space by recursively decomposing it into four equal quadrants or regions (Feng and Watanabe, 2015). Here, the algorithm splits the area and population into successively smaller quadrants by checking whether the content of each split is smaller than prescribed values (e.g., population >750 and area <9 km^2). Following this step, all shapes produced were smaller than 9 km^2 and contained fewer than 750 people (Fig. 3). For a more detailed discussion on the Quadtree approach, see Qader et al. (2020).

Merging process. When all split regions have population size and area smaller than the requested thresholds, the regions are then merged until they match constraints designed to facilitate on-the-ground logistics of enumeration (Table 2). The merging process tries to obtain regions as close to the target population of 650 as possible while keeping below the threshold of 750 and area within 0.1 to 9 km^2. At the same time, the algorithm ensures that merged areas do not cut across obstacles or administrative borders. With a lower priority, the process also tries to produce shapes that are as compact as possible. By construction, the boundaries of the EAs resulting from the merging process will also follow geo-referenced features (or square sides, if quadtree squares are needed).

To complete the merging step, we used the Automated Zone-design Tool (AZTool, (Martin, 2002), which is based on Openshaw's (1977) Automated Zoning Procedure (AZP), originally developed by the Office for National Statistics (ONS) for the 2001 census in England and Wales (Cockings et al., 2011; Martin and Lyndon, 2009). The software is written in VB.NET and no GIS software is required to run AZTool. However, data preparation and visualisation of the results require GIS software. Before employing the AZTool, the ESRI (shapefiles) containing polygon data must be converted to .aat and .pat files in the format required by AZTool. This process can be done using AZTImporter. The AZTool and AZTImporter are freely available at https://www.geodata.soton.ac.uk/software/AZTool/.

AZTool iteratively combines and recombines sets of geographic areas to generate larger zones optimised to meet a set of pre-defined user-specified constraints. Such specified constraints include population threshold (Min, Max and target) and compactness of the shape (i.e., avoiding difficult shapes such as snake-like or donut shapes). In our case, The merging process takes as inputs: (1) the primary unit features defined in the split process, (2) ranges of target population and area values for EAs (Table 2), (3) the gridded population density dataset to compute the population for the re-merged region at each step, and (4) a set of specified boundaries across which regions should not be merged (e.g., large rivers, delineation of urban and rural strata, administrative boundaries).

The AZTool was originally designed for contexts in which good data on existing household locations and EAs are available. Given the potential for large uninhabited areas and a paucity of road and other data in many regions of Somalia, we modified the AZTool process to include a constraint on the maximum and minimum area. Since this is a computationally intensive process when applied to large areas, the method was applied to 18 pre-war regions separately. Separation of the country into 18 units does not influence the output results as the algorithm would normally be constrained to keep EAs within these 18 units.

Computation of EA probability of selection. The split and merge algorithm results in a partition of the country into regions that satisfy the definition of an EA. We calculate the probability of selection for each EA proportional to the population within each pre-War region. This probability was defined according to the expected population within each EA divided by the total population in the regional stratum, recalling that the country was divided into 18 pre-war regions and further subdivided into urban/rural strata.

$$P\left(EA_{ij}\right) = \frac{EApop_{ij}}{\sum_{i=1}^{n} EApop_{ij}} \tag{1}$$

so that $\sum P(EA_{ij}) = 1$ for each j ($j = 1\ldots18$).

$P(EA_{ij})$ is a probability of selection for an EA in a specific stratum (Urban or rural) and pre-war regions (e.g., rural Bari), i is an EA number, j is the regional stratum type. $EApop_{ij}$ is the population within an EA in a specific regional stratum, n is the number of EAs within a regional stratum.

Comparison between Manual PESS urban EAs and our semi-automated approach Urban EAs. In UNFPA's 2014 PESS, no Rural EAs were created, and the urban EAs were manually digitised based on high-resolution Google Earth Imagery. We compared the results of our semi-automated methodology against those manually digitised EAs in urban areas in both Mogadishu and Hargeysa cities. Since the boundary of PESS urban EAs and their household size are confidential, we were unable to publish a complete comparison. Instead, we compared the distribution of population size between automatically and manually generated urban EAs. In the AZTool, maximum, minimum and target population should be defined. For this exercise, we have set the maximum urban population as 2000 people, minimum as 150 and the preferred target is 1000 people per an EA different to the

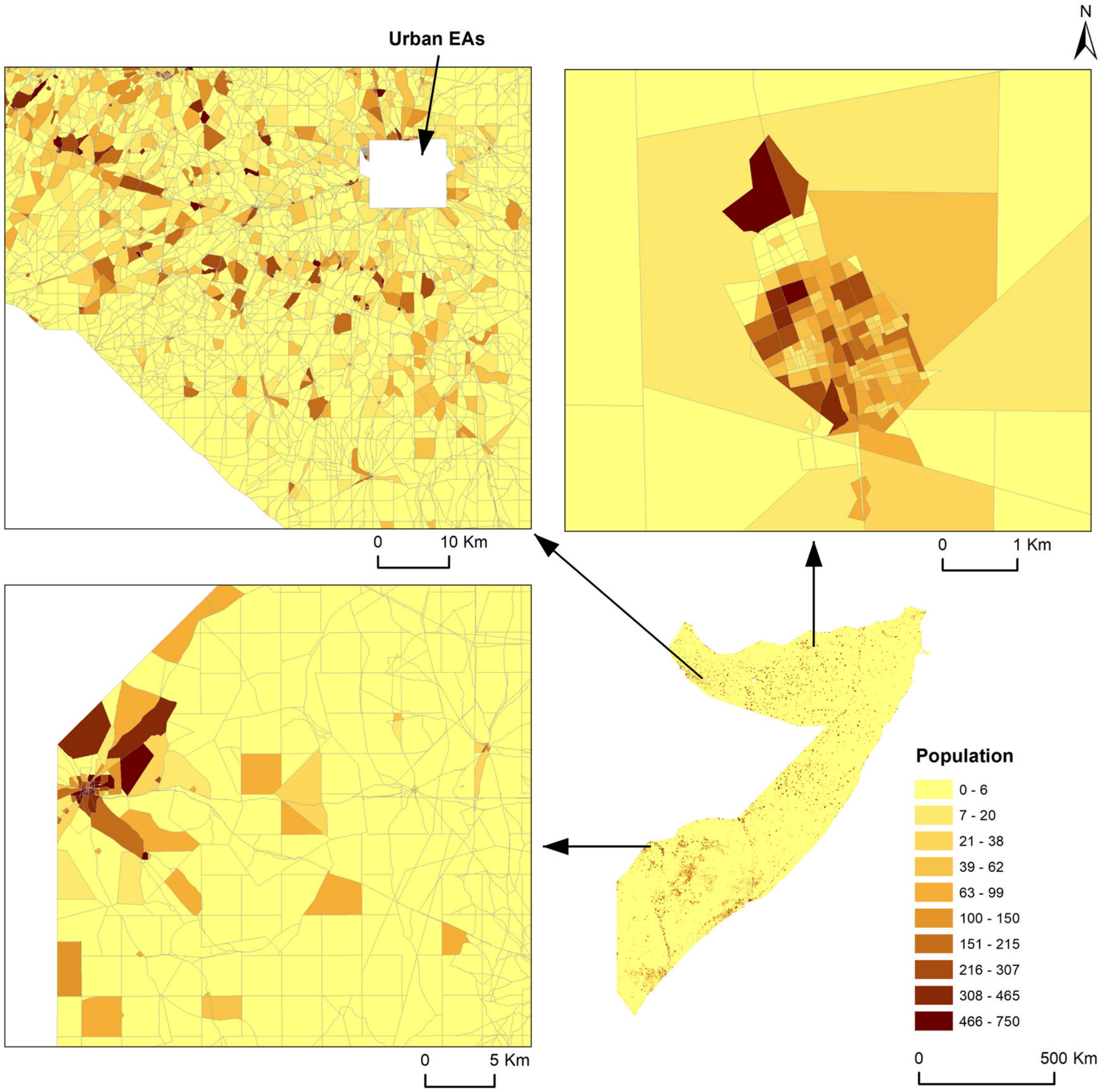

Fig. 3 Primary units (PU) and population sizes in rural areas generated from the splitting process. After the splitting process, the rural territory in Somalia was subdivided into small geographic units and each unit has total population estimates computed from high gridded population dataset.

Table 2 Illustrates the criteria were set in AZTool to generate sensible EAs in Somalia's rural and urban (Mogadishu and Hargeysa cities) areas.

Criteria	Hard constraints		Soft constraints		On or off
	Rural	Urban	Rural	Urban	
Shape compactness			Yes	Yes	On
Population	Min = 0 Max = 750	Min = 150 Max = 2000	Target = 650	Target = 1000	On
Area	Min = 1 km^2 Max = 9 km^2	Min = 0.002 km^2 Max = 4 km^2	Target = 8.8km^2	Target = 2 km^2	On
Donuts	Yes	Yes			On

rural setting criteria (Table 2). The target population size per EA is based on the population and area that an enumerator could reasonably cover in a day. Usually, population size per EA is larger in urban areas compared to the rural since the population density is higher. Similarly, for the purpose of this comparison, the total population for EAs used in the 2014 PESS were estimated using updated WorldPop gridded population data.

Results: Somalia case study

Rural EAs

Description of results from our split and merge algorithm. A total of 253,833 polygons were created in Somalia's rural areas after the splitting process. The region merging technique resulted in 113,367 rural EAs, with population size ranging from 0 to 750 and a maximum area of 9 km^2 (Fig. 4a). In addition, the

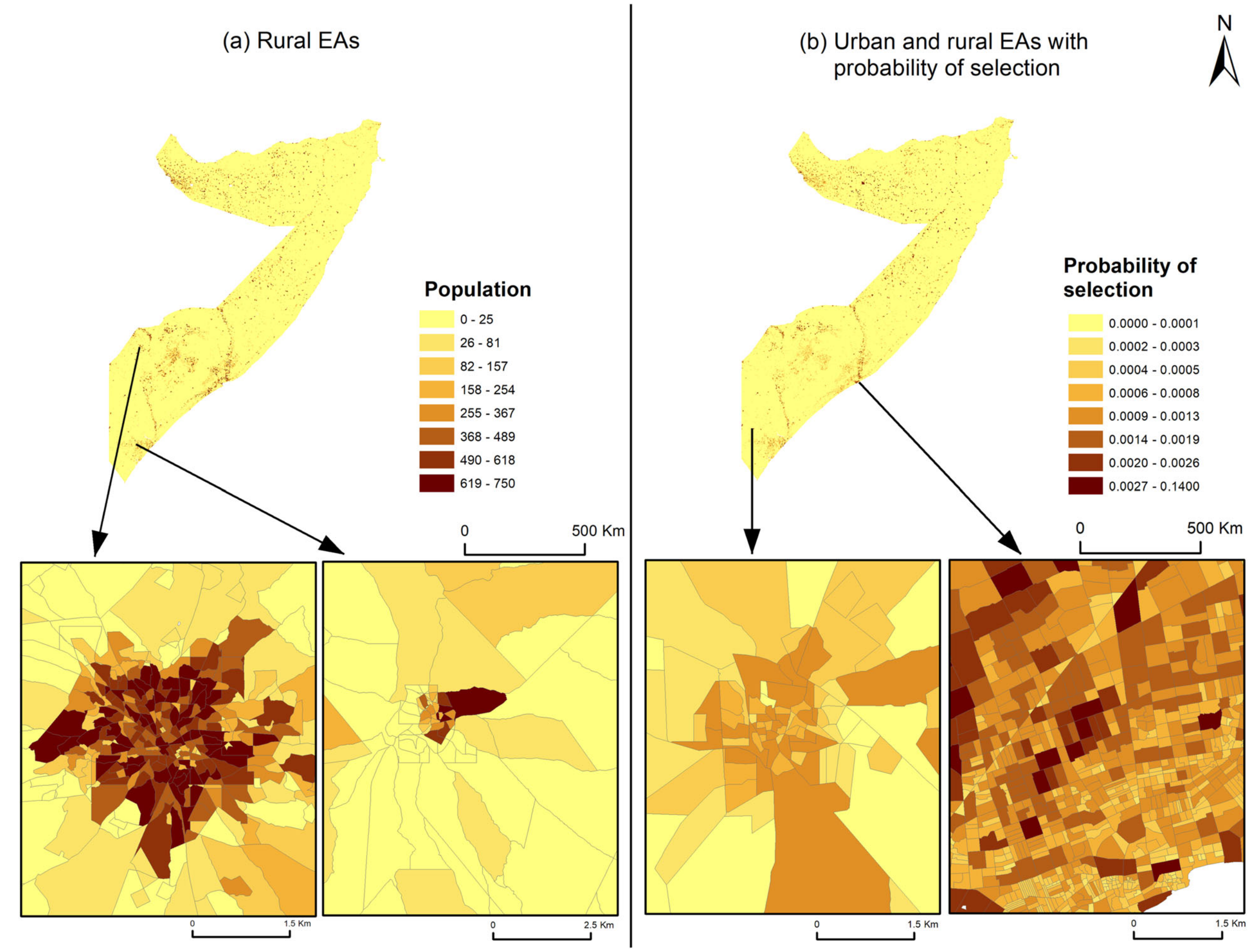

Fig. 4 The population size per regional strata. a Rural EAs in Somalia with the population size after the merging process, **b** Urban and rural EAs with their probability of selections proportional to the population size per regional strata. This product was generated after merging the splitting units until they match constraints designed to facilitate on the ground logistics of enumeration.

probability of EA selection proportional to the population size was also computed for each EA. Furthermore, the probabilities of selection were summed up in each regional stratum type and are equal to 1 (Fig. 4b).

Outlines of some generated Rural EAs are overlaid on high-resolution Google Earth imagery in Fig. 5, showing that EA boundaries conform well to natural boundaries in populated areas. Three categories of EAs boundaries can be seen. The first category is in towns or highly populated areas where EA boundaries are well-matched with logical ground natural boundaries such as roads (Fig. 5a). The second category consists of very sparsely populated or unpopulated areas, where roads and natural boundaries are still present (Fig. 5b). Finally, the third category represents very sparsely populated or unpopulated areas, where natural boundaries do not contribute to EA shapes (Fig. 5c). In addition, it can be seen from Fig. 5d, e that the rural EA boundaries conform to stratum boundaries (urban and rural).

Urban EAs

Results from split and merge algorithm. The semi-automated approach was applied to both Mogadishu and Hargeysa cities. Figure 6 presents the results obtained from overlaid our semi-automated urban EAs boundaries on high-resolution Google Earth imagery. Figure 6a shows examples of EAs boundaries in Hargeysa city while Fig. 6b is illustrating the boundary of urban EAs in Mogadishu city. Importantly the boundary of EAs perfectly matches the natural demarcation on the ground particularly roads, reflecting the good quality of road data available for these areas. In addition, generated urban EA boundaries are nested within urban strata (Fig. 6b).

Comparison with PESS manual urban EAs. We compared the estimated urban population size (based on updated WorldPop gridded population data) in our semi-automated EAs to the manually digitised PESS 2014 EAs (Fig. 7). In urban areas across Somalia, there were 1380 PESS EAs, while our process generated 1775. In terms of usability characteristics, for the most part, PESS urban EAs follow roads well but we have found some examples where this is not the case as some may bisect buildings, likely due to changes in building layouts since the construction of the PESS EA dataset. The population size per PESS urban EA in Mogadishu, (based on high resolution gridded population datasets), ranges from zero to 17,000 and the area ranges from 5 m^2 to 7 km^2. The zero values in population size and small area likely indicate the presence of gaps in the datasets. The large population and area size for some of the EAs indicate that the EAs may not be practical for a surveyor in the urban context as it may either cover a high-populated area or cover a large space. In the semi-automated process, these constraints can be tuned based on user requirements.

The population size in the automated EAs ranged from 150 to 2000. The area constraints ranged from 2000 m^2 to around 4 km^2.

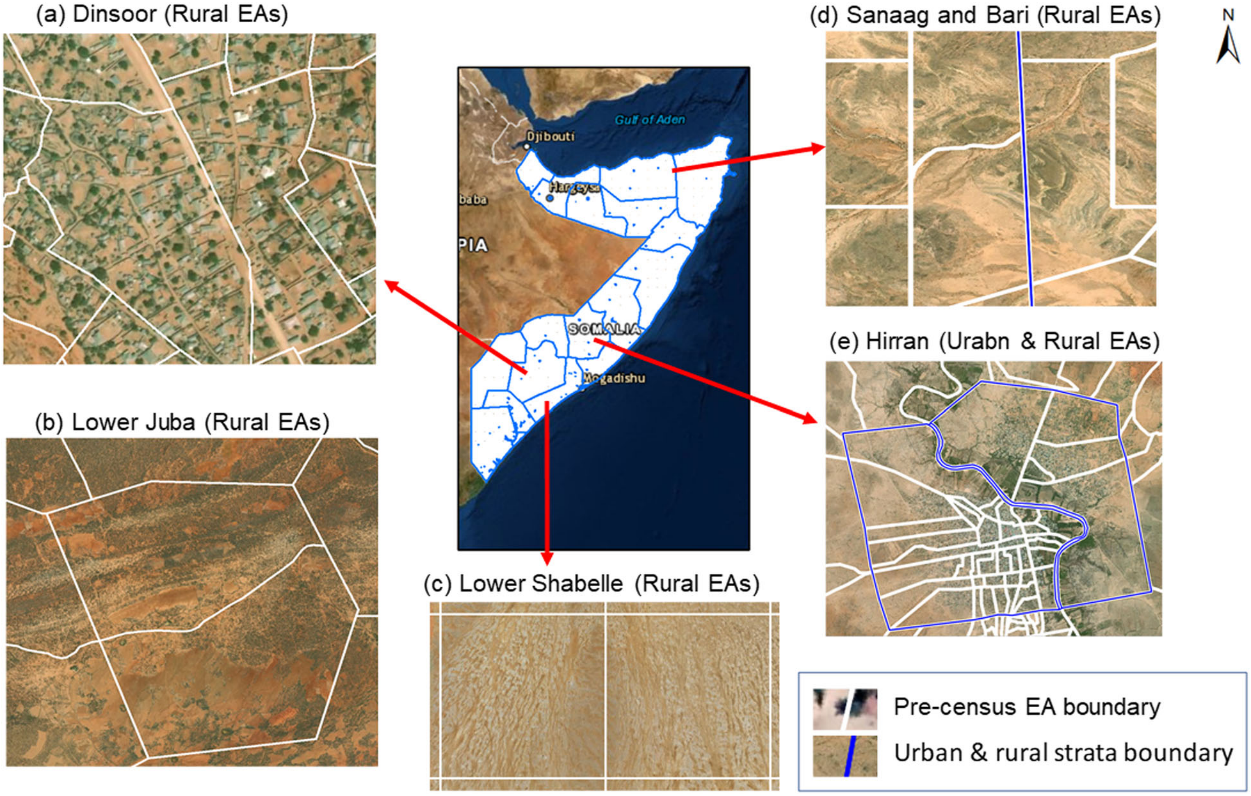
(a) Dinsoor (Rural EAs)
(b) Lower Juba (Rural EAs)
(c) Lower Shabelle (Rural EAs)
(d) Sanaag and Bari (Rural EAs)
(e) Hirran (Urabn & Rural EAs)
N
Djibouti
Gulf of Aden
Hargeysa
SOMALIA
Mogadishu
PIA
baba
Pre-census EA boundary
Urban & rural strata boundary

Fig. 5 Outlines of rural EAs in different locations over Somalia generated by the split-merge algorithm. In the given examples, the generated rural EAs are overlaid on high resolution Google Earth imagery. Rural EA boundaries are well matched with logical ground boundaries in highly populated areas (**a**) whereas this might be different in other areas with respect to data availability and ruralness (**b**, **c**). In addition, the EAs are nested within urban and rural strata (**d**, **e**).

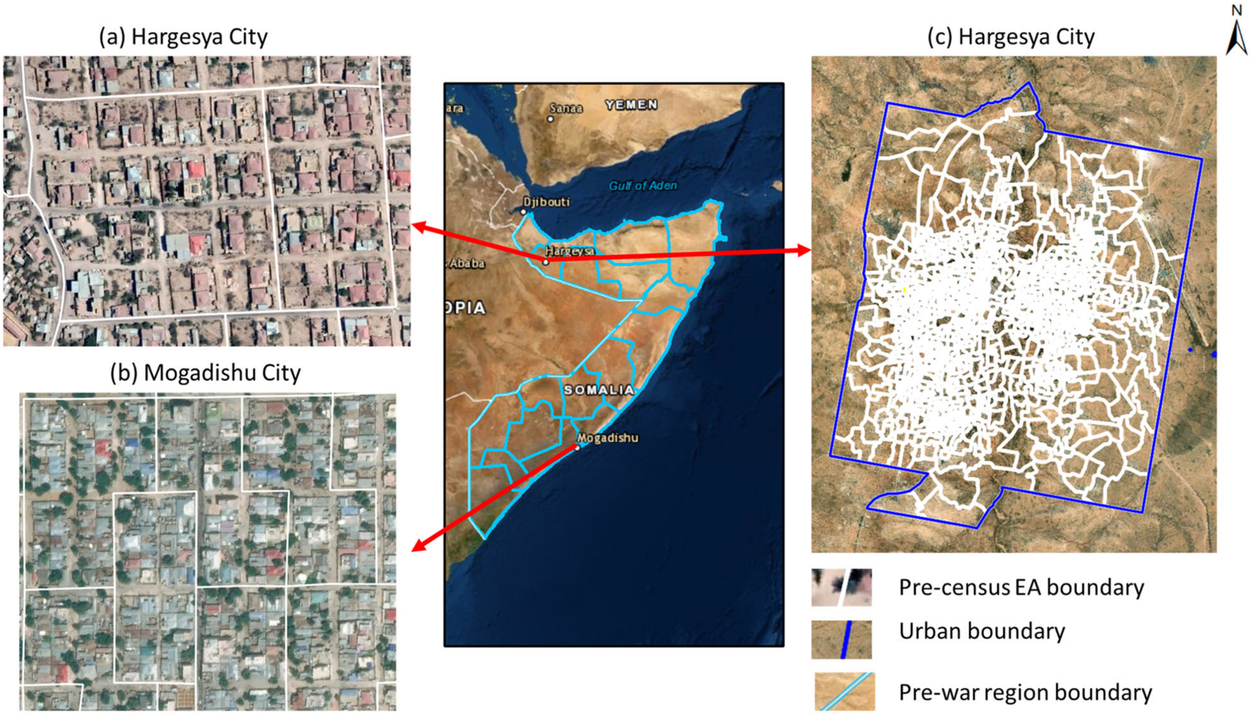

Fig. 6 Outlines of semi-automated urban EAs in different locations over Hargeysa and Mogadishu in Somalia. a,**b** The automated urban EA boundaries are matching well with ground roads and they are nested within urban strata (**c**).

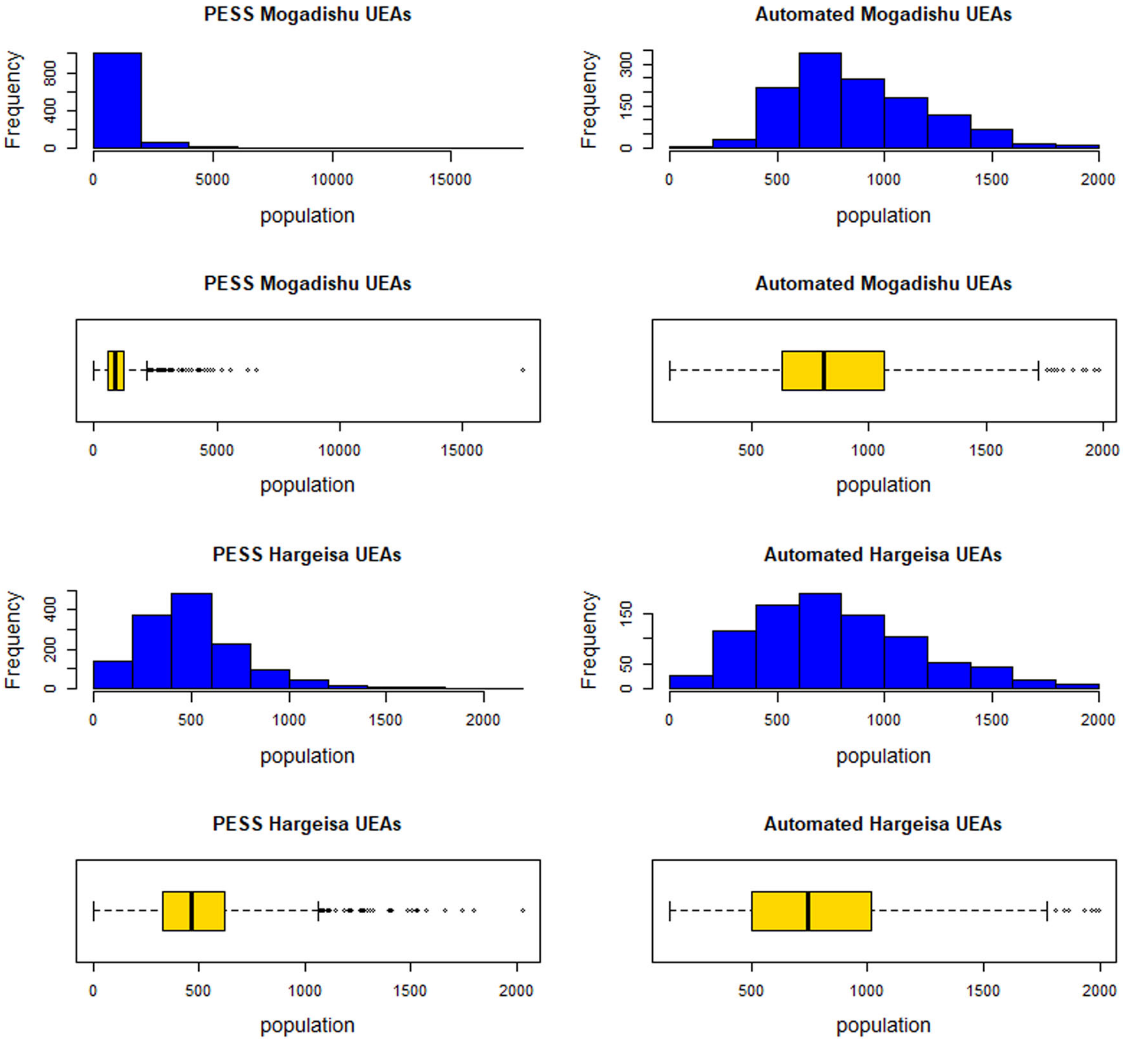

Fig. 7 Histogram distribution of population within urban PESS 2014's EAs and automated urban EAs in Hargeysa and Mogadishu, Somalia, based on high resolution gridded population estimates. For the comparison, we assume that the maximum population could be 2000 people per EA, but the preferred target is 1000 people per EA. The population size per PESS urban EA in Mogadishu ranges from 0 to 17,000 and area ranges from 5 m^2 to 7 km^2. In both cities, the population size in the automated EAs ranged from 150 to 2000 and the area constraints ranged from 2000 m^2 to 4 km^2.

Discussion

This work represents the first attempt to generate an automatic mapping of pre-census enumeration areas and a population sampling frame where the boundaries are based on vector data for roads, settlements and other physical features. Our algorithm simplifies the complex and time-consuming processes of EA delineation that is crucial in the preparation of the national census. The process of EA delineation is a crucial early step in census preparation as it helps in budgeting and allocating materials, logistical and personnel requirements, and is a key factor in the success of census.

While many countries will try to re-use existing census demarcations for political or logistic reasons, typically EAs need to be updated when a new census is taking place. In particular, for some countries transitioning from paper-based census maps, new EAs will need to be created in places that have never been enumerated, either due to inaccessibility or growth of new settlement areas. Our proposed approach for the first time facilitates this process by leveraging increasingly available geospatial data and computing power to facilitate the EA delineation in a time-effective and cost-effective manner.

Assuming adequate datasets such as high-resolution gridded population data (such as WorldPop, which has global population estimates at 100 m from 2000 to 2020) and good coverage digital boundaries (e.g., roads, river, waterways, etc) are available for a country, our results show that the semi-automatic approach can produce similar results to manual approaches. Once data were acquired, our digitisation approach took a matter of weeks for a team of two people, including data harmonisation, running the algorithm and checking the results. By contrast, a traditional manual digitisation approach can take months for a large staff. Ground-based approaches require even more effort and often impose a significant risk to on-the-ground mapping staff. In addition, the task of accounting for appropriate population size within EAs is much more easily handled algorithmically and provides more consistent results versus manual digitisation. By considering recent data on boundaries, roads and other features the algorithmic process also creates EAs that can account for features that may not be readily observable from satellite imagery, thus making them more practical for fieldwork. As well, EAs can be updated as new data becomes available. Finally, because each EA is already associated with an estimation of population size, these EAs can be easily turned into a nationally representative sampling frame with a probability of selection.

In this work, semi-automatic EA datasets were produced for Somalia that span both rural and urban contexts, and which showed significant promise relative to pre-existing EAs. Our EA boundaries are mutually exclusive (non-overlapping) and exhaustive (cover the entire country). The outlines of the EA boundaries are in line with ground feature boundaries and easily identifiable in the urban and populated areas (Figs. 5a and 6b, c). All the generated EA boundaries are nested within administrative boundaries and they are not crossing through urban and rural stratum (Figs. 5d, e and 6c). The EAs are compact and free of pockets or disjoint sections. The population size and area were set to thresholds such that each EA can be small enough and accessible to be covered by an enumerator within the census period (e.g., max population in a rural area was 750 and area was set to 9 km^2). In addition, with respect to the population and area constraints, the created EAs are large enough to guarantee data privacy. Furthermore, the EAs are flexible enough to allow the widest range of tabulations for different statistical reporting units.

The sampling frame derived from our EA delineation process is reproducible and complete, making it for other data collection activities such as nationally representative household surveys, which typically rely on census data as a sampling frame. While a complete, recent census makes the best sampling frame, sampling frames based on outdated or inaccurate census data are the norm in many countries around the world, resulting in potentially biased sampling (Turner, 2003). Because our EAs can be based on the most recently available data and has probability of selection, the developed sampling frame can generate a sampling frame that better reflects the population of interest.

One limitation of the method we have presented is that it relies on quality of high-resolution population estimates, as well as good quality data on roads and other boundaries. This study was motivated by the context of Somalia, which presents an extreme challenge in terms of the availability of good population data and results circular problem: to get new EA delineations for the census, we need a good population estimate, for which a census would be very helpful. For the population data, we overcame these challenges through extensive manipulations of available population and settlement data to arrive at a gridded population estimate that was, to the best of our abilities, the most reliable gridded population estimate we could find. At the same time, we recognise that such manipulations are not a desirable feature of an easily replicable method for delineating Enumeration areas and highlight that these manipulations were only necessary for the present context. It should also be noted that population estimates are inherently wrong and will likely contain various biases depending on their provenance, originating data and the methods used to derive them. As well, numerous gridded population data are increasingly becoming available globally. It is therefore vital that the inherent biases and inaccuracies of such datasets be identified and acknowledged in methods such as ours. As a counterpoint, we also add that even new census-based population data can be inaccurate through the deliberate or accidental exclusion of certain populations. Furthermore, national populations themselves are inherently variable, especially in fragile contexts. In such cases, modelled population estimates tailor-made for such contexts may provide more useful results than old census results that contain uninhabited settlements or ignore newly settled areas. Whether or not to base trust on modelled population estimates must, therefore, come down to the needs of the census or survey in question and the quality of all available data sources.

Other sources of uncertainty might come from the quality of the digitised features that have been extracted mainly from OSM. OSM was chosen for this work because it has offered multiple publicly accessible comprehensive datasets for Somalia compare to other sources and the data are up to date as it constantly being updated by the subscribed users. In addition, in terms of quality assurance and to help to lead to a better quality of data, OSM has employed multiple automatic and manual approaches such as bug reporting, error detection, visualisation, monitoring, assistant, tag statistics and external compare. Furthermore, numerous studies have analysed the quality of OSA datasets in different contexts. For instance, an assessment in London showed that on average about 6 m of the position recorded by the Ordnance Survey (OS), and with approximately 80% overlap of motorway objects between OSM and OS datasets (Haklay, 2010). Although dramatic increases in volunteered information have substantially enhanced geographic data, it has also prompted concerns about its reliability, quality and overall value (Flanagin and Metzger, 2008). The OSM quality assessment results are heterogeneous if we compare the various areas investigated. For example, the spatial accuracy and completeness were generally good enough in developed countries (Wang et al., 2013; Graser et al., 2014), while in South Africa the rate at which data is generated varies in space and time (Siebritz et al., 2012). Therefore, lack of positional accuracy and incompleteness of OSM data in Somalia might have inherited inaccuracies into this work including boundaries might have cut through buildings. In addition, due to lack of feature coverage particularly in rural areas, insensible boundaries from the quadtree approach were incorporated to complete the national EA coverage.

With regard to infrastructure data, our results demonstrate the influence of data quality. For example, the boundaries in well-digitised urban areas (Figs. 5a, 6b and 6c) are aligned with visible demarcations on the ground. Areas with more sparse data (Fig. 5b, d) have lower spatial coverage data and there is often a visible discrepancy with the actual natural boundaries where data on these boundaries do not exist. Finally, boundaries with poor data (Fig. 5c) are devoid of boundary data and rely entirely on the quadtree algorithm to generate EAs. The lack of data is particularly problematic in rural areas because large stretches of land in Somalia are only sparsely populated. These results are in line with those of previous empirical studies indicated that urban areas are often better mapped than rural areas in OSM (Hagenauer and Helbich, 2012; Siebritz et al., 2012). However, the semi-automatic EA algorithm could easily be updated and re-run when new data becomes available. With the increasing prevalence of open-source data in development contexts (e.g., AidData.org, GRID[3]), these data gaps are increasingly being filled and made available.

Verification and validation are also needed in the case of manual and algorithmic EA delineation. Although the household size per urban EA was one of the inputs of the population modelling, deviation still exists between the actual population (based on HH size) and estimated population based on the modelling in manual urban EAs. This result may be explained by the fact that the modelled population data considered additional recent datasets besides the urban household information to distribute the population within the grid cells (See Supplementary information Table S3 and S5). In addition, there were discrepancies between PESS 2014 household numbers within urban EAs and regional population estimates. If household size per EAs were used to generate the total population per region, a similar total population in the region could not be achieved as it was reported in the PESS 2014, particularly in Gedo. In addition, the very high population size based on the gridded population dataset in the manual urban EAs (Fig. 7) could be a result of their large spatial coverage. For instance, one urban EA contains 1500 households (The households were listed in the field based on PESS 2014). The only way to understand these discrepancies is to conduct spot checks on population and other data used, as well as on the EAs generated. Indeed, a modern approach to delineation of enumeration areas should probably include some iterative approach to generating a first attempt set of areas, followed by random ground-based spot checking to determine whether certain populations are excluded from the frame. Such an approach could fit readily into existing census protocols and could be conducted on a small fraction of the areas normally covered by enumeration teams. Such approaches would also increase local trust in the results, as well as allowing the algorithm to better incorporate local context into their processes.

Because of the increasing prevalence of national data on infrastructure and populations, our method can be easily transferred and adopted in other countries. Furthermore, such datasets are becoming more accurate and detailed, facilitating more granular results. For instance, at the global scale, road data from OSM is ~83% complete (Barrington-Leigh and Millard-Ball, 2017). Similarly, Facebook has been running a project to perform AI-Assisted Road tracing within OSM and promise that such road data will be extracted for many countries around the world (Facebook AI-Assisted Road Tracing, 2020). In terms of potential sources of population data, high-resolution gridded population datasets are now available for all the countries from multiple sources. For instance, high-resolution population estimates were recently produced for Afghanistan, DRC and Nigeria, with the district-level predictive power of ~95% (United Nations Population Fund (UNFPA), 2019; WorldPop, 2019b; WorldPop, 2020). While we acknowledge that modelled population estimates will never be as good as actual census counts but they can be used where census data is unavailable or outdated.

Our application of a pre-existing tool identified areas where the approach can be optimised. For example, AZTool was not developed to aggregate the multitude of small regions that were created by splitting the map using so many different datasets. The soft compactness constraint used by AZTool aims to produce EAs that are not elongated or convoluted. This constraint is not often satisfied, possibly because the input data set contained many irregular shapes and donuts, making it difficult for the tool to merge the neighbouring EAs. We suggest incorporating a compactness metric based on the comparison of the longest length in a shape against the diameter of a circle of the same area, which would provide a more robust selection criterion during the merging process and lead to more compact EAs. This metric would be insensitive to shape size, unlike the compactness metric used in the AZ tool. In addition, some region merging techniques are better than others at ensuring all final regions match the desired criteria, suggesting another potential area for improvement. Furthermore, consistent and reliable labelling of data features would enable some features to be prioritised in the merging process. Such prioritisation would allow the algorithm to account for traversability through the ranking of split lines. In addition, additional data sources such as building delineation derived from satellite imagery (San and Turker, 2010; Vakalopoulou et al., 2015) would supplement road data, especially in rural areas and could contribute to more targeted EAs by clustering building locations and excluding large unsettled areas.

Finally, the usability of the present approach is limited to users of AZTool in ArcGIS software. Subsequent research will aim to develop a software tool that can be used to generate optimal and practical EAs with minimal user interaction and tailored to a wider set of needs. By incorporating a wider range of input datasets and shapefiles, as well as increased flexibility in the parameterisation of EAs based on factors described above, we hope that an open-source software package may have broad development impact in countries constrained by cost and accessibility considerations.

Conclusion

In many countries such as Somalia, the need for basic spatial sampling tools such as Enumeration areas are crucial to continued development, but on-the-ground logistical constraints, security and data limitations remain significant barriers. Here, we show how a novel approach to creating predefined census enumeration areas can overcome these barriers. In addition, by highlighting existing freely available, up-to-date high-resolution gridded population data and settlement maps we hope to help make the method more broadly accessible. Future research will aim to implement the approach in different countries such as the Democratic Republic of Congo, while improving the usability of the output areas through introducing more constraints and parameters, as well as merging algorithms. Ultimately, we aim to improve the accessibility of the approach by providing a user-friendly tool based on the approaches described in this paper.

Data availability

The digitised boundaries for roads, waterway, river, residential area, building delineation, waterbody and locations for places, hamlet and villages were obtained from Open Street Map (OSM) (www.openstreetmap.org).

Received: 7 February 2020; Accepted: 4 November 2020;
Published online: 04 January 2021

References

Alazab M., Islam M., Venkatraman S. (2009) Towards Automatic Image Segmentation Using Optimised Region Growing Technique. In: Nicholson A., Li X. (eds) AI 2009: Advances in Artificial Intelligence. AI 2009. Lecture Notes in Computer Science, vol 5866. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10439-8_14

Azar D, Engstrom R, Graesser J, Comenetz J (2013) Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. Remote Sens Environ. https://doi.org/10.1016/j.rse.2012.11.022

Balinski M, Johnston R, McLean I, Young P (2010) Drawing a new constituency map for the united kingdom: the parliamentary voting system and constituencies bill 2010. The British Academy, London

Barrington-Leigh C, Millard-Ball A (2017) The world's user-generated road map is more than 80% complete. PLoS ONE 12(8). https://doi.org/10.1371/journal.pone.0180698

Cajka J, Amer S, Ridenhour J, Allpress J (2018) Geo-sampling in developing nations. Int J Soc Res Methodol 21(6):729–746. https://doi.org/10.1080/13645579.2018.1484989

Center for International Earth Science Information Network-CIESIN-Columbia University, International Food Policy Research Institute - IFPRI, The World Bank, and Centro Internacional de Agricultura Tropical-CIAT (2011) Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Count Grid. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades . Accessed 09 July 2018

Center for International Earth Science Information Network-CIESIN-Columbia University (2017) Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 10. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades. Accessed 05 July 2018

Cheriyadat A, Bright E, Potere D, Bhaduri B (2007) Mapping of settlements in high-resolution satellite imagery using high performance computing. Geojournal 69:119–129. https://doi.org/10.1007/s10708-007-9101-0

Cockings S, Harfoot A, Martin D, Hornby D (2011) Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales. Environ Plan A 43 (10):2399–2418. https://doi.org/10.1068/a43601

Duque JC, Anselin L, Rey SJ (2012) The Max-P-regions problem*. J Region Sci 52 (3):397–419. https://doi.org/10.1111/j.1467-9787.2011.00743.x

Ellard-Gray A, Jeffrey NK, Choubak M, Crann, SE (2015) Finding the hidden participant: solutions for recruiting hidden, hard-to-reach, and vulnerable populations. Int J Qual Methods 14. https://doi.org/10.1177/1609406915621420

Esch T, Heldens W, Hirner A, Keil M, Marconcini M, Roth A, Strano E (2017) Breaking new ground in mapping human settlements from space-The Global Urban Footprint. Isprs J Photogr Remote Sens 134:30–42. https://doi.org/10.1016/j.isprsjprs.2017.10.012

European Commission, Joint Research Centre (JRC); Columbia University, Center for International Earth Science Information Network-CIESIN (2015): GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000,

2015) European Commission, Joint Research Centre (JRC) PID: http://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a

Facebook AI-Assisted Road Tracing (2020) OpenStreetMap Wiki, https://wiki.openstreetmap.org/w/index.php?title=Facebook_AI-Assisted_Road_Tracing&oldid=1957775.

Feng J, Watanabe T (2015) Index and Query Methods in Road Networks. In Index and Query Methods in Road Networks (Vol. 29, pp. 1–161). Berlin: Springer-Verlag Berlin

Finkel RA, Bentley JL (1974) "Quad trees a data structure for retrieval on composite keys". Acta Inform 4:1–9. https://doi.org/10.1007/bf00288933. Springer-Verlag

Flanagin AJ, Metzger MJ (2008) The credibility of volunteered geographic information. GeoJournal 72(3):137–148. https://doi.org/10.1007/s10708-008-9188-y

Florczyk AJ, Ferri S, Syrris V, Kemper T, Halkia M, Soille P, Pesaresi M (2016) A new european settlement map from optical remotely sensed data. IEEE J Selected Topics Appl Earth Observat Remote Sens 9(5):1978–1992. https://doi.org/10.1109/jstars.2015.2485662

Flowerdew R, Feng ZQ, Manley D (2007) Constructing data zones for Scottish neighbourhood statistics. Comput Environ Urban Syst 31(1):76–90. https://doi.org/10.1016/j.compenvurbsys.2005.07.008

Folch DC, Spielman SE (2014) Identifying regions based on flexible user-defined constraints. Int J Geogr Inform Sci 28(1):164–184. https://doi.org/10.1080/13658816.2013.848986

Graser A, Straub M, Dragaschnig M (2014) Towards an open source analysis toolbox for street network comparison: indicators. Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph. Transactions in GIS 18(4):510–526. https://doi.org/10.1111/tgis.12061

Gure F, Yusuf M, Foster AM (2015) Exploring Somali women's reproductive health knowledge and experiences: results from focus group discussions in Mogadishu. Reproduct Health Matters 23(46):136–144. https://doi.org/10.1016/j.rhm.2015.11.018

Hagenauer J, Helbich M (2012) Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. Int J Geogr Inform Sci 26(6):963–982. https://doi.org/10.1080/13658816.2011.619501

Haklay M (2010) How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. Environ Plan B 37(4):682–703. https://doi.org/10.1068/b35097

Haynes R, Daras K, Reading R, Jones A (2007) Modifiable neighbourhood units, zone design and residents perceptions. Health Place 13(4):812–825. https://doi.org/10.1016/j.healthplace.2007.01.002

Kumar N (2007) Spatial sampling design for a demographic and health survey. Population Res Policy Rev 26(5-6):581–599. https://doi.org/10.1007/s11113-007-9044-7

Lang S, Kienberger S, Tiede D, Hagenlocher M, Pernkopf L (2014) Geons-domain-specific regionalization of space. Cartogr Geogr Inform Sci 41(3):214–226. https://doi.org/10.1080/15230406.2014.902755

Lu X (2009) "Need a job? Apply to become a Census enumerator". Wise Bread: Living on a Small Budget. http://www.wisebread.com/need-a-job-apply-to-become-acensus-enumerator. Accessed Jan 2011

Martin R, Lyndon A (2009) Optimised geographies for data reporting: zone design tools for Census output geographies (Statistics New Zealand Working Paper No. 09–01). Wellington, Statistics New Zealand

Martin DJ (2002) Geography for the 2001 Census in England and Wales. Population Trends 108:7–15

Nock R, Nielsen F (2004) Statistical region merging. IEEE Trans Pattern Anal Machine Intelligence 26(11):1452–1458. https://doi.org/10.1109/tpami.2004.110

Openshaw S (1977) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. Trans Institute British Geogr 2:459–472

Pesaresi M, Guo HD, Blaes X, Ehrlich D, Ferri S, Gueguen L, Zanchetta L (2013) A global human settlement layer from optical HR/VHR RS data: concept and first results. IEEE J Selected Topics Appl Earth Observat Remote Sens 6 (5):2102–2131. https://doi.org/10.1109/jstars.2013.2271445

Qader SH, Lefebvre V, Tatem AJ, Pape U, Jochem W, Himelein K, Ninneman A, Wolburg P, Nunez-Chaim G, Bengtsson L, Bird T (2020) Using gridded population and quadtree sampling units to support survey sample design in low-income settings. Int J Health Geogr 19(1). https://doi.org/10.1186/s12942-020-00205-5

Roy Chowdhury PK, Bhaduri BL, McKee JJ (2018) Estimating urban areas: new insights from very high-resolution human settlement data. Remote Sens Applicat 10:93–103. https://doi.org/10.1016/j.rsase.2018.03.002

San DK, Turker M (2010) Building extraction from high resolution satellite images using hough transform. In: Kajiwara K, Muramatsu K, Soyama N, Endo T, Ono A, Akatsuka S (eds) Networking the World with Remote Sensing. ISPRS Tech Commiss, Japan. Vol. 38, pp. 1063–1068

Siebritz L, Sithole G, Zlatanova S (2012) Assessment of the homogeneity of volunteered geographic information in south africa. In Shortis M, Madden M (eds), Int Soc Photogrammetry & Remote Sensing; Hexagon; ESRI; RMIT Univ, Sch Math Geospatial Sci. Australia, Xxii Isprs Congress, Technical Commission Iv, Vol. 39-B4, pp. 553–558

Thomson DR, Hadley MB, Greenough PG, Castro MC (2012) Modelling strategic interventions in a population with a total fertility rate of 8.3: a cross-sectional study of Idjwi Island, DRC. BMC Public Health 12. https://doi.org/10.1186/1471-2458-12-959

Thomson DR., Stevens FR, Ruktanonchai NW, Tatem AJ, Castro MC (2017) GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data. Int J Health Geogr 16. https://doi.org/10.1186/s12942-017-0098-4

Turner AG (2003) Sampling frames and master samples. Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Surveys. UNITED NATIONS SECRETARIAT,ESA/STAT/AC.93/3. https://unstats.un.org/UNSD/demographic/meetings/egm/Sampling_1203/docs/no_3.pdf

U.S. Department of Commerce. Bur. of the Census (BUCEN) (1978) Mapping for Censuses and Surveys. International Statistical Programs Center

UN DESA PD (2018) World Urbanization Prospects 2018 [https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf]. Accessed 01 Jul 2020.

UNFPA (2016) Population Composition and Demographic Characteristics of the Somali People. http://analyticalreports.org/pdf/UNFPA_PESS_Vol_2.pdf

UNFPA, Federal Republic of Somalia (2014) Population Estimation Survey 2014 for the Pre-War Regions of Somalia (PESS). UNFPA, Nairobi

Unite N (2000) Handbook on geographic information systems and digital mapping. United Nations Publication, New York

United Nations Population Fund (UNFPA) (2019) New methodology: a hybrid census to generate spatially disaggregated population estimates. https://www.unfpa.org/resources/new-methodology-hybrid-census-generate-spatially-disaggregated-population-estimates. Accessed 17 Feb 2020

UNITED NATIONS SECRETARIAT (UNS) 2007. Report of the Sub-regional Workshop on Census Cartography and Management. ESA/STAT/AC.144/L.3

Vakalopoulou M, Karantzalos K, Komodakis N, Paragios N (2015) Building Detection In Very High Resolution Multispectral Data With Deep Learning Features 2015 Ieee International Geoscience and Remote Sensing Symposium. IEEE, pp. 1873–1876

Vijayaraj V, Bright EA, Bhaduri BL (2007) High resolution urban feature extraction for global population mapping using high performance computing Igarss: 2007 Ieee International Geoscience and Remote Sensing Symposium, Vols. 1-12: Sensing and Understanding Our Planet. IEEE, pp. 278–281

Wanderer JP (2017) Analysis of large and complex data. Anesth Analgesia 125 (1):345–345. https://doi.org/10.1213/ane.0000000000002127

Wang M, Li QQ, Hu QW, Zhou M (2013) Quality analysis of open street map data. In: Wu B, Guilbert E, Shi J (eds) 8th International Symposium on Spatial Data Quality, Vol. 40–2, pp. 155–158

Wang YJ, Jiang LL, Qi QW, Liu Y, Wang J (2019) Remote Sensing-Guided Sampling Design with Both Good Spatial Coverage and Feature Space Coverage for Accurate Farm Field-Level Soil Mapping. Remote Sensing, 11(16): https://doi.org/10.3390/rs11161946

World Bank (2017) Somali poverty profile: findings from wave 1 of the somali high frequency survey. World Bank, Washington, DC, p 2017

WorldPop (School of Geography and Environmental Science, University of Southampton) (2020) Bottom-up gridded population estimates for the Kinshasa, Kongo-Central, Kwango, Kwilu, and Mai-Ndombe provinces in the Democratic Republic of the Congo, version 1.0. https://doi.org/10.5258/SOTON/WP00658

WorldPop (School of Geography and Environmental Science, University of Southampton) (2019b) Bottom-up gridded population estimates for Nigeria, version 1.2. https://doi.org/10.5258/SOTON/WP00655

WorldPop Data (2019a) WorldPop, University of Southampton, Southampton, UK. 2019. http://www.worldpop.org.uk/data/data_sources. Accessed 10 Mar 2018

Yacyshyn AM, Swanson DA (2011) The Costs of Conducting a National Census: Rationale for Re-Designing Current Census Methodology in Canada and the United States. The 21st Annual Warren E. Kalbach Population Conference, at the University of Alberta, Edmonton, Canada

Acknowledgements

The World Bank and the Somalia Knowledge for Results Trust Fund of the Multi-Partner Fund provided funding for this project. "Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org".

Author contributions

SHQ was responsible for research design, data cleaning, implementation, analysis, interpretation, generating of the EAs, drafting and production of the final manuscript. VL was responsible for updating high gridded population data, settlement and preparation of final manuscript. AJT, PU, KH, AN, LB were responsible for interpretation and production of the final manuscript. TB was responsible for overall scientific management, drafting, interpretation and preparation of final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1057/s41599-020-00670-0.

Correspondence and requests for materials should be addressed to S.Q.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.