Policy Research Working Paper 8972

A Novel Approach to the Automatic Designation of Predefined Census Enumeration Areas and Population Sampling Frames

A Case Study in Somalia

Sarchil Qader
Veronique Lefebvre
Amy Ninneman
Kristen Himelein
Utz Pape
Linus Bengtsson
Andy Tatem
Tomas Bird

WORLD BANK GROUP
Poverty and Equity Global Practice
August 2019

Policy Research Working Paper 8972

Abstract

Enumeration areas are the operational geographic units for the collection, dissemination, and analysis of census data and are often used as a national sampling frame for various types of surveys. Traditionally, enumeration areas are created by manually digitizing small geographic units on high-resolution satellite imagery or physically walking the boundaries of units, both of which are highly time, cost, and labor intensive. In addition, creating enumeration areas requires considering the size of the population and area within each unit. This is an optimization problem that can best be solved by a computer. This paper, for the first time, produces an automatic designation of predefined census enumeration areas based on high-resolution gridded population and settlement data sets and using publicly available natural and administrative boundaries. This automated approach is compared with manually digitized enumeration areas that were created in urban areas in Mogadishu and Hargeisa for the United Nations Population Estimation Survey for Somalia in 2014. The automatically generated enumeration areas are consistent with standard enumeration areas, including having identifiable boundaries to field teams on the ground, and appropriate sizing and population for coverage by an enumerator. Furthermore, the automated urban enumeration areas have no gaps. The paper extends this work to rural Somalia, for which no records exist of previous enumeration area demarcations. This work shows the time, labor, and cost-saving value of automated enumeration area delineation and points to the potential for broadly available tools that are suitable for low-income and data-poor settings but applicable to potentially wider contexts.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at upape@worldbank.org.

A Novel Approach to the Automatic Designation of Predefined Census Enumeration Areas and Population Sampling Frames: A Case Study in Somalia

Sarchil Qader
Veronique Lefebvre
Amy Ninneman
Kristen Himelein
Utz Pape
Linus Bengtsson
Andy Tatem
Tomas Bird

JEL: C55, C83, C80

Keywords: Census; Enumeration Areas; Sampling Frame; Somalia

1. Introduction

In most countries of the world, 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. In Low and Middle-Income Country (LMIC) settings, spatially referenced data on populations are crucial for targeting interventions, monitoring the impact of population growth, health and environmental applications, poverty mapping and transport and city planning (Hay et al. 2005; Balk et al. 2006; Tatem et al. 2007). Enumeration areas (EAs) are the operational geographic units for the collection of census data (United Nations Statistics Division 2010). 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. Given data on population density at a sufficiently high spatial resolution, the process of designing EAs can be automated, which could substantially accelerate census mapping.

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 provided by a national sampling frame, survey data sampled from a country's population are likely to be biased, typically under-sampling 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 fragile contexts where data are even more important due to the need of timely and evidence-based planning in volatile situations, up-to-date sampling frames are necessary.

However, defining new EAs is a challenging process that needs to consider a number of factors, including the quality of available data and the logistics of sampling; EAs need to make sense relative to the real-world distribution of people, buildings, and infrastructure. Good design of census EAs has a large impact on the efficiency of the census, a process which can cost on the order of hundreds of millions of dollars (Loots 2015). To better facilitate efficient census processes, EAs should lead to (i) better organization and management of the census itself and to (ii) small area statistics which meet common user needs (Martin 2002).

Historically the creation of census EAs has been an expensive problem solved by brute-force approaches such as physically walking to map EA boundaries, which can require thousands of staff and many 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 (United Nation Secretariat 2007). Since the advent of Geographic Information Systems (GIS) and high-resolution satellite photography, census cartographers in some countries have been able to manually digitize EA boundaries from satellite imagery, leading to better control of EA delineation. However, manual digitization of EAs is still cost-, time-, and labor-intensive. It is also prone to human error and can have poor accuracy (Alazab et al. 2009; Balinski et al. 2010; Cocking et al. 2011). In addition, poor quality or outdated satellite imagery can make it difficult or impossible to identify the boundaries or house blocks that align with actual house blocks and EA boundaries that are misaligned with real-world structures while natural boundaries can cause significant challenges in the field enumeration process. Furthermore, drawing or growing EAs manually with respect to population size and area constraints is challenging and can result in 'nonsensical' boundaries (e.g. a single structure split into different EAs). Finally, EAs or sampling frames in countries with large population

displacements and rapid urbanization need regular updating, meaning that the manual digitization process must be replicated.

Previously, automated zone design methods have been employed to group areas into zones for a range of purposes including investigation of neighborhood effects on health and release of census data (Cockings et al. 2011; Flowerdew et al. 2007; Haynes et al. 2007). To construct the created new zones some constraints were considered such as population size and building types. However, these works were conducted in developed countries where actual household data were available, and the method relied on previously-defined sub-units or EAs. In addition, in most cases, the created zone boundaries do not match well with the logical ground natural boundary. Several studies have also employed region merging techniques to derive the census enumeration areas. For instance, Folch & Spielman (2014) showed the advantage of applying the improved max-p algorithm on growing the irregular regions from census tracts. In New Zealand, the ArcGIS/AZTool toolkit, which is a region-merging tool, was developed to design new reporting geographic units using the 2006 census data as the main input, with significant improvements over traditional methods reported (Martin and Lyndon 2009). However, all these examples used existing EAs as their base unit and no works exist that have employed region merging techniques to generate new EAs in a country where an existing product is not available.

Despite the many potential advantages of automated EA generation, the approach has not gained significant uptake in LMIC contexts where the kinds of spatial data previously used in tools like the AZTool (such as census tracts) are unavailable or outdated. Fortunately, many new and freely available data sources have become available, which can directly inform the creation of EAs. For example, crowdsourced road and natural feature data, which can be used to define logical boundaries for EAs, are increasingly available in repositories such as Open Street Maps (OSM). Globally, OSM is ~83% complete; more than 40% of countries (including several in the developing world) have a fully mapped street network (Barrington-Leigh & Millard-Ball 2017). In addition, global high-resolution population models such as WorldPop (www.worldpop.org, 2019) provide estimates of population density that 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, or to update existing EAs, using Somalia to demonstrate the process. Our automated process is based on a methodology we termed ‘split and merge’, inspired from the field of image processing, specifically image segmentation using mathematical morphology (e.g. watershed and waterfall algorithms; Beucher 1994; Damiand & Resch 2003). The process first splits the country into regions as small as possible that follow visible boundaries (e.g. roads), then progressively re-merges them so that they are as large as possible while respecting given constraints on the area and population size, and so that they do not cross obstacles and administrative boundaries. This process ensures creating EAs so that their boundaries do not cut across buildings and can be seen easily from the ground by surveyors.

2. Methodology

At the request of the World Bank, we applied an automated EA delineation process to build a national sampling frame for Somalia. This section describes our approach and then details its application in the context of Somalia.

2.1 General approach

We developed a novel automated process to create EAs for an entire country that falls within given ranges of population size and area, that does not cut through buildings, that do not cross obstacles and administrative boundaries, and that follow tangible borders visible from the ground as much as possible.

Two kinds of data are needed in our process: i) vector data of roads, natural boundaries and/or settlement boundaries with which to split the country area into sub-units and ii) high-resolution (100 m) estimates of population density. Our process first combines all sources of vector data to split the country into small sub-areas, then merges them into units matching constraints such as population size or area, a process we termed the split and merge algorithm. Details of data requirements and processing are discussed in a later section. In regions where data on geographic and manmade features are too sparse to obtain small enough regions after the splitting stage (either nonexistent e.g. desertic areas or unmapped), a quadtree algorithm is used to further split the country. In this case, not all EAs follow visible boundaries.

In general, using our automated delineation method being developed into a tool we expect that a solution can always be found where no EA falls outside of the limits set. As long as space is partitioned in sufficiently small areas the region merging algorithm can have enough flexibility to create EAs that meet all the targets. Where there is a lack of geo-referenced features to obtain small enough areas, then the automated tool will use the quadtree algorithm to split. A flowchart of the process is given in Figure 1.

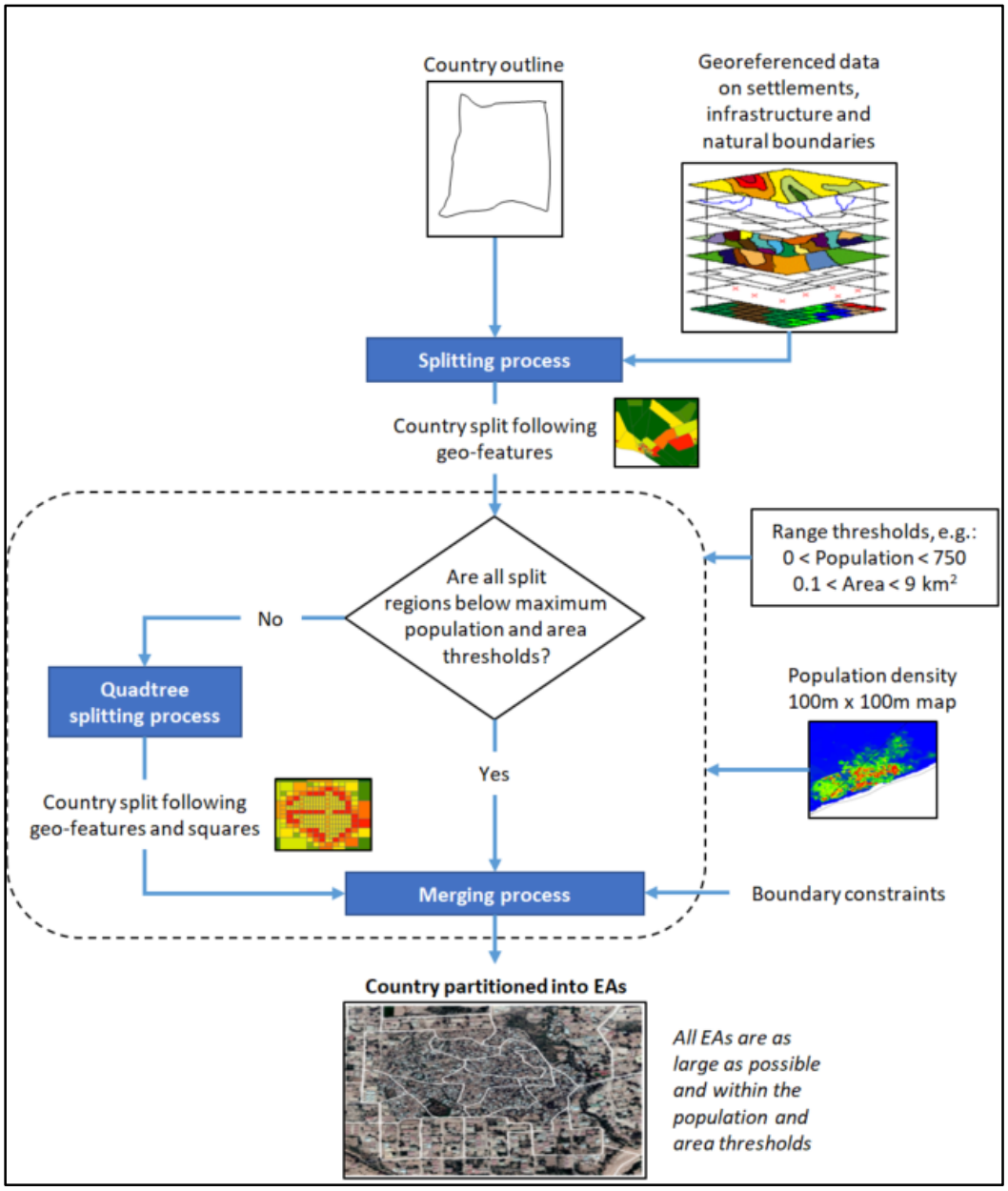

Figure 1. Schematic diagram of our automated EA delineation process. The first step is to split the territory into regions as small as possible using georeferenced features such as administrative and natural boundaries (e.g. rivers, terrain), settlement location and outlines, and road and path networks. We then compute the estimated population in each region thus obtained (using a very high spatial resolution population density map) 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.

2.2 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 with slightly more males (6.2 million) than females (6.1 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 created EAs in urban areas only. The risks associated with fieldwork and the lack of funding were just two hurdles this approach faced; significantly displaced populations exist in parts of Somalia, without any official population information available. Therefore, the World Bank recognized that re-building Somalia's statistical infrastructure and capacity was 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 (e.g. security risks, lack of funding).

2.3 Data sources

To conduct this work, several data sets have been compiled and combined from various resources (table 1).

Table 1. Data used for Somalia Enumeration Areas.

Source	Data description
Road Data (OSM)	Lines
Waterway (OSM)	Lines
River (OSM)	Lines
Residential area (OSM)	Part points and part polygons
Building (OSM)	polygons
'places', 'hamlet', and 'villages' (OSM)	Points
DLR Global Urban Footprint (GUF)	Binary raster
BMGF/DigitalGlobe (DG) population estimates	scatter points
World Bank/Flowminder/WorldPop building counts from Google Satellite imagery	polygons
UNFPA/PESS urban Enumeration Areas	Polygons
UNFPA/PESS urban Enumeration Areas (EAs)	# households per EA.

household number	
BMGF/DigitalGlobe (DG) settlement outlines for North Somalia	polygons
UNFPA/PESS rural population estimates	points
Pre-war regions boundary	polygons
Waterbody (OSM)	polygons

2.4 Data pre-processing

2.4.1 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 areas constraints of EAs may need to vary in different strata. For example, Urban versus Rural strata typically need very different constraints to account for the differing population densities. 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 regions, urban and rural strata were intersected with the 18 pre-war regions administrative boundary. Based on World Bank recommendations the urban and rural strata in Banadir were merged and considered as urban.

2.4.2 Settlement boundaries

Data on settlement locations and boundaries are needed to inform EA delineation in the 'merge' part of the algorithm and are also used to create our refined 100m population density estimates. Recent increased availability of high-resolution satellite imagery and high-power computing resources with adequate image processing algorithm have advanced the development of high-resolution human settlement layers (Vijayaraj et al. 2007; Florczyk et al. 2016; Roy Chowdhury 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 data sets either are not available for Somalia or are incomplete for the country or they might not specifically be trained for Somalia since the country is facing a constant regional instability and severe drought that forced substantial migration within the region. In addition, during this work, several new data sets on settlements were collected from a variety of sources (table 1), each of which georeferencing settlements absent from the other existing settlement layers. This has motivated the current work to generate an improved high-resolution settlement map by combining the datasets listed in table 1. Appendix 1 gives more detail on the processing done on each data set. The resulting settlement map was used to improve population density predictions and EA delineation.

2.4.3 High-resolution population density estimates for Somalia

A high spatial resolution map (100m x 100m) is necessary to estimate the population of each sub-region created during the automated EA delineation process. Multiple high to moderate resolution global modeled population datasets are freely accessible to download including WorldPop

(www.worldpop.org, 2018), Global Rural-Urban Mapping Project, Version 1 (GRUMP) (CIESIN et al. 2011), Gridded Population of the World Version 4 (GPWv4) (CIESIN, 2017), Gridded Population of the World, United Nation (GPW UNEP, 2006) and Global Human Settlement Population Grid (GHS-POP) (JRC and CIESIN 2015). However, none of these data sets on its own was sufficient for our purposes, as they were created without the use of the PESS 2014 data, or the final total population was not adjusted to match the PESS regional total. In addition, we had access to more recent data sets (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 novel 100m x 100m population density map to calibrate our EA delineation. We give below an only succinct overview of the method employed as it is not the object of the present paper, and it is not relevant to the description and results of our novel automated process for EA delineation, which can accept as input any gridded data set of sufficiently high resolution. Appendices 2 and 3 list the data sources that we used for urban and rural areas respectively, and the transformations we applied in order to obtain a 100m x 100m raster for each. Data sources include information on building density, household density and population density. We used the World Bank survey (UNFPA 2014) to estimate a median number of people per building and per household to approximate population density from data on building and household densities. In places lacking data but identified as settled, we modeled 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).

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

2.5.1 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 data sets 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 are complete, then the process ensures that no building will be cut.

For the areas that do not meet the criteria, Thiessen polygons based on settlement locations were created. However, for sparsely populated areas or vast featureless areas, with only a few geo-referenced settlements or residential areas, areas created from this approach were still larger than 9km^2.

Step 2 - Estimate population and area

With the PU feature set defined, we were able to then compute the population size for each PU using the high resolution gridded population data sets in the Zonal Statistics Tool in ArcGIS.

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, despite 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 and 2 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 & area < 9km^2). Following this step, all shapes produced were smaller than 9 km^2 and contained fewer than 750 people.

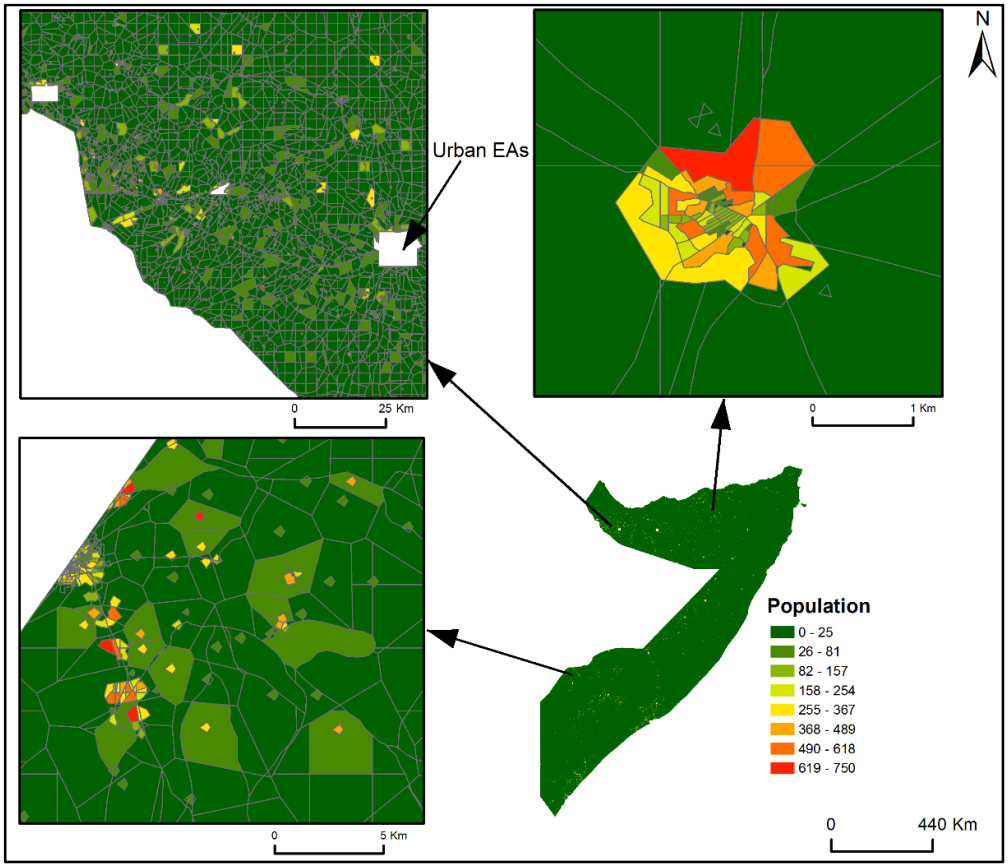

Figure 2. This figure shows the small irregular polygons in rural areas with population size after the splitting process.

2.5.2 Merging process

After splitting the area of interest (e.g. country) into small regions as small as possible and with population size and area smaller than the requested thresholds, the regions are merged until they match

the constraints given in table 2. The aim of the merging process is to obtain regions as close to the target as possible and with population below threshold (e.g. 750) and area within a specified range (e.g. 0.1 to 9km^2) while ensuring that merged areas do not cut across obstacles or administrative borders. With a lower priority, the process also tends 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).

The merging process takes as inputs: 1/ the PU 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 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). Since this is a computationally intensive process, the method was applied to every 18 pre-war regions separately.

To complete the merging step, we used the Automated Zone-design Tool (AZTool) since it is publicly available, most compatible compare to the other tools for the purpose of this study and user-friendly. AZTool was developed by Martin et. (2002) and is based on Openshaw's (1977) Automated Zoning Procedure (AZP). The AZP methodology was developed by the Office for National Statistics (ONS) for the 2001 census in England and Wales (Cocking et al 2011) and to revise census geography in New Zealand (Martin and Lyndon 2009). AZTool iteratively combines and recombines sets of geographic building blocks to generate larger zones optimized 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. Criteria and constraints used in the current study are reported in table 2. Compactness is a minor constraint used so that shapes 'tend' to be compact and avoid difficult shapes such as snake-like shapes. Donut is one output area surrounding another. The Area constraint was not originally included as a constraint in AZTool, however, we modified the .xml file to include the area constraint for the purpose of this study.

Table 2. Illustrates the criteria were set in AZTool to generate sensible EAs in Somalia rural area.

Criteria	**Hard Constraint**	**Soft Constraint**	**On or off**
Shape compactness		Yes	on
Population	Min=0 and Max=750	Target=650	on
Area	Min=0.1 km^2 and Max=9 km^2	Target=8.8 km^2	on
Donuts	Yes		on

2.6 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 population of each EA using the population density map and compute the probability of selection. The probability of sampling for each EA 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.

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

$PEA_{(1\ 1)} + PEA_{(2\ 1)} + PEA_{(3\ 1)} + \ldots + PEA_{(n\ j)} = 1$ for each stratum within a pre-war region

$P(EA_{ij})$ 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, $EApop_{ij}$ is the sum of EAs' population within a regional stratum, n is the number of EAs within a regional stratum.

2.7 Comparison between Manual PESS urban EAs and our automated approach Urban EAs

In UNFPA's 2014 PESS, no Rural EAs were created, and the urban EAs were manually digitized based on high-resolution Google Earth Imagery. We compared the results of our automated methodology against those manually digitized EAs in urban areas in both Mogadishu and Hargeisa 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 and EA area between automatically and manually generated Rural EAs. The population and area were also computed for PESS urban EAs based on our high-resolution gridded population data sets and the populations and area of EAs within Hargeisa and Mogadishu cities were compared.

3. Results: Somalia case study

3.1 Rural EAs

3.1.1 Description of results from our split and merge algorithm

A total of 253,833 polygons were created in Somalia rural areas after the splitting process. After applying the region merging technique in a rural area, 113,367 EAs were created, with population size ranging from 0 to 750 and a maximum area of 9 km2 (Figure 3a). In addition, the probability of EA selection proportional to the population size and area were also computed for each EA. Furthermore, the probabilities of selection were summed up in each regional stratum type and is equal to 1 (Figure 3b).

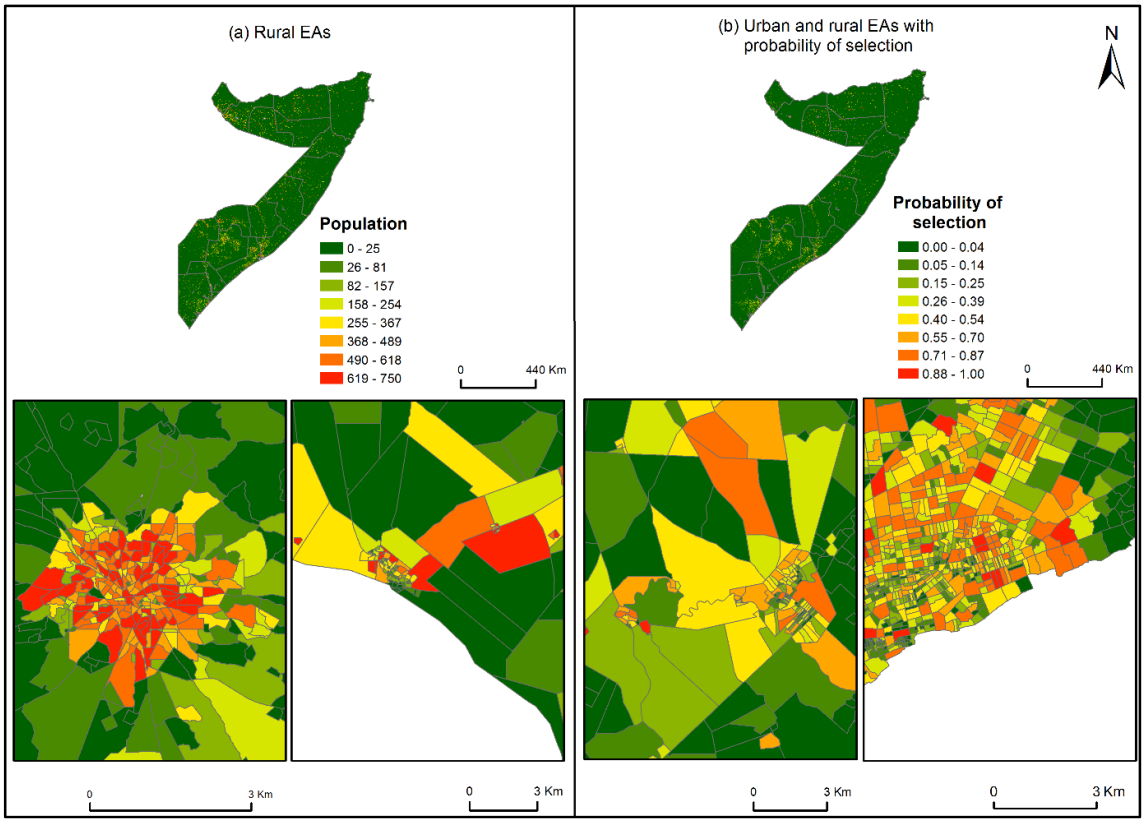

Figure 3. (a) Rural EAs in Somalia with the population size after the merging process, (b) Urban and rural EAs with their probability of selections proportional to the population size per regional strata.

Outlines of some generated Rural EAs are overlaid on high-resolution Google Earth imagery in figure (4), 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 boundary such as roads (Figure 4a, b, c, g, h). The second category can mostly be seen in very sparse or unpopulated areas, where roads and natural boundaries are still adhered to (Figure 4 d and f). Finally, the third category represents very sparsely populated or unpopulated areas, where natural boundaries do not contribute to EA shapes (Figure 4e).

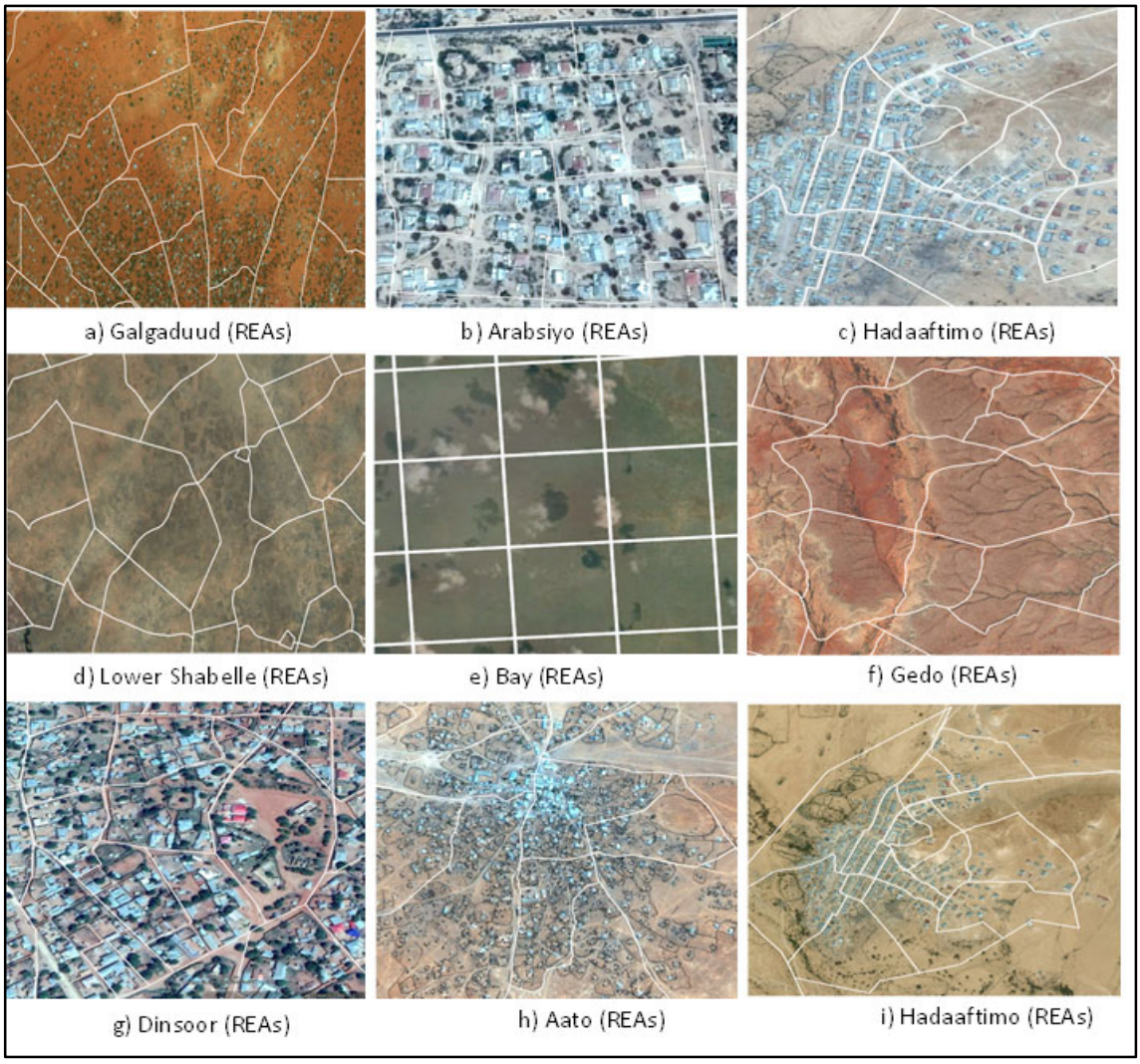

Figure 4. Outlines of rural EAs in different locations over Somalia generated by the split-merge algorithm.

3.2 Urban EAs

3.2.1 Description of results from split and merge algorithm

The automated approach was applied to both Mogadishu and Hargeisa cities. Figure 5 presents the results obtained from overlaying our automated urban EAs boundaries on high-resolution Google Earth imagery. Figure 5a, b, c, d, and e show the results for Hargeisa city while Figure 5f, g, h and i are 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.

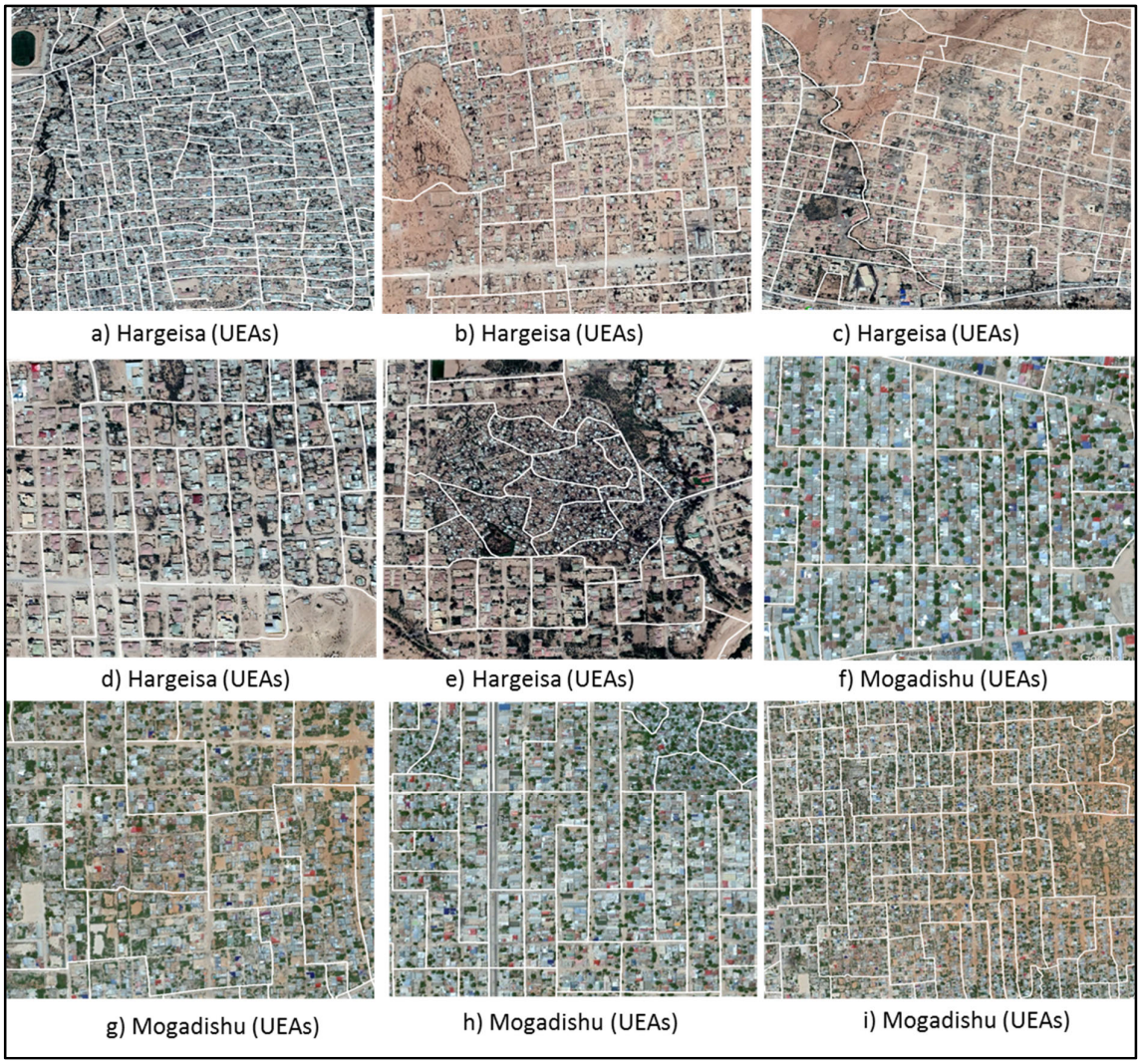

Figure 5. Outlines of our automated urban EAs in different locations over Hargeisa and Mogadishu in Somalia.

3.2.2 Comparison with PESS manual urban EAs

We compared our automated urban EAs to the manually digitized PESS 2014 urban EAs in terms of the population and area covered (Figure 6). The total number of PESS EAs is 1,380 and the total automated EAs is 1,775. If we consider the maximum urban population as 2,000 people per an EA and the preferred target is less 1,000 people per an EA, the automated EAs might have the same performance or even better in some cases compared to PESS urban EAs. While we cannot directly present the PESS urban EA data here, but we have noticed some issues when we have reviewed and generated the histograms. For the most part, PESS urban EAs follow roads well but we have found some examples where this is not the case as they cut the houses, likely due to changes in building layouts since the construction of the PESS EA data set. The minimum population size per PESS urban EA in Mogadishu, which was obtained from high gridded population data sets, ranges from zero to 17,000 and the minimum area ranges 5 m^2 to around 7,000,000 m^2. The zero values in population size and small area indicate the presence of gaps in the data sets. 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 automated process, these constraints can be tuned based on user requirements. The minimum population size for the automated EAs was 150 and the

maximum was 2,000. The area constraints were tuned as well, and it starts from 2,000 m^2 to around 4,000,000 m^2.

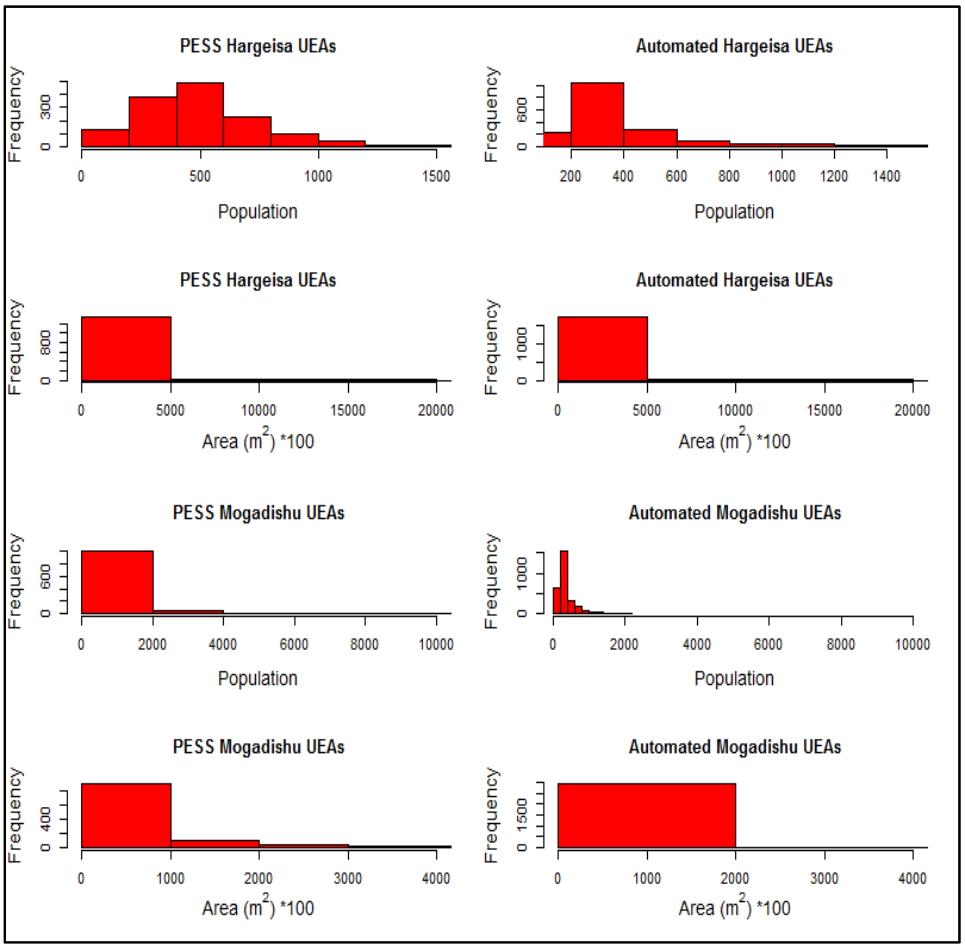

Figure 6. Histogram distribution of population and area for urban PESS 2014's EAs and automated urban EAs in Hargeisa and Mogadishu, Somalia. For the comparison, we assume the maximum population could be 2000 people per EA but the preferred target is less than 1000 people per EA.

4. Discussion

Our split and merge approach using freely available data on human population density and boundary data provides an automated practical approach to generating EAs, compared with the manual digitization of EA boundaries. This approach can be used to generate EAs automatically in countries where EAs are not existent or exist but need regular updating because of large population movement/changes or in the preparation of census and are inaccessible due to insecurity restrictions. The accuracy and comparison between both approaches might be varied from an area to another but if adequate data sets are available for a country, our automatic approach can produce a similar result and ensures all EAs will match the defined criteria, given the data the method has access to with a significant cost reduction. Using the methods shown in this paper, EAs for a country can be generated within ~2 weeks by a team of 2-3 GIS analysts if a good high-resolution population dataset and geospatial data on natural boundaries such as roads are available. Verification and validation, however, will need additional time. By contrast, adopting the traditional manual digitization to create EAs for a country can take years, hundreds of staff, a budget on the order of millions of dollars and often impose a significant risk to on-the-ground mapping staff. In addition, the task of accounting for appropriate

population size and area within EAs is much more easily handled algorithmically. By taking into account recent data on boundaries, roads and other features the algorithmic process also creates EAs that are more practical for fieldwork and can be updated as new data become 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.

One limitation of the method we have presented is that it relies on good quality data. Modeling population density is not the object of this work (Lloyd et al. 2017; Wardrop et al. 2018), however, reliable population estimates are crucial to use our method to obtain operational EAs. In addition, in Somalia, the method works well where OSM data sets and other ancillary data were available and of good quality - mostly in the highly populated areas or where volunteer OSM mapping effort has been highest. For instance, the boundaries of the first category (Figure 4a, b, c, g, h and i) are aligned with visible demarcations on the ground. This is since the used geospatial layers to automatize the EA delineation were derived from high-resolution satellite imagery. This pattern can be seen in the second category, but there is often a visible discrepancy to the actual natural boundaries (Figure 4 d and f) where natural boundary data appear not to exist. Finally, EA boundaries in the third category (Figure 4e) are devoid of boundary data and rely entirely on the quadtree algorithm to generate EAs. The lack of data or vast featureless areas is particularly problematic in rural areas because large stretches of land in Somalia are not or only sparsely populated. In addition, OSM road coverage does not exist in these areas. However, we note that the algorithm could easily be updated and re-run when new data become available. With the increasing prevalence of open-source data in development contexts (e.g. AidData.org, GRID3), these data gaps are likely to be filled before the next census.

Manual digitization of EAs may be still a preferable option in some cases, where no data on roads or natural boundaries are available, or where local context weighs significantly in terms of choosing EAs. For example, areas where security or gerrymandering may be important considerations will require more hands-on control of EAs to include or exclude specific houses or neighborhoods. In addition, manual digitization does not rely on ancillary data sets as long as high-resolution satellite imagery is available, the EAs can be manually digitized. However, manual digitization is prone to many errors including leaving gaps, unclosed EAs and greater irregularity in actual population within EAs. Importantly, intensive multi-resource involvement in manual EA delineation must be considered.

Although the actual household size per urban EA was one of the inputs of the population modeling, deviation still exists between actual population (based on HH size) and estimated population based on the modeling in manual urban EAs. This result may be explained by the fact that the modeling considered additional recent data sets besides the urban HH information to distribute the population within the grid cells (see appendixes 2 and 3). In addition, there were discrepancies between PESS 2014 HH number within urban EAs and regional population estimates. This means that, if HH size per EAs were considered to generate the total population per region, similar total population on 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 data set in the manual urban EAs (Figure 6) could be a result of their large spatial coverage. For instance, there is a case that an urban EA contains 1,500 HH (The HH were listed in the field based on PESS 2014).

Our use of a pre-existing tool that was not designed for the application provided us with valuable learning about the relevant features of EAs, as well as the challenges associated with the process and areas where the approach can be optimized. For example, AZTool was not developed to aggregate the multitude of small regions that were created by splitting the map using so many different data sets. The AZTool has a soft compactness constraint to produce EAs that are not oddly shaped (e.g. elongated or convoluted), however, the constraint is not often satisfied. This might be because the input data set

contains many irregular shapes and donuts which make it difficult for the tool to merge the neighboring EAs considering all given criteria. In addition, some region merging techniques are better than others at ensuring all final regions match the desired criteria. In some cases, techniques become stuck in configurations with no solution (e.g. with small unmerged areas along the map borders or surrounded by fully formed regions that cannot be merged more). Such errors could be eliminated by incorporating harder constraints or additional constraints that are more appropriate to the use of many small areas.

Our use of AZTool in the context of Somalia also pointed to potential extensions of the region-merging approach other criteria based on socio-economic variables such that socio-economic metrics are homogenous within the produced EAs. For instance, if the purpose of a survey is to collect data on poverty, the existence of high-resolution poverty maps (Tatem et al. 2014; Steele et al 2017) could be used to provide a weighting constraint which the tool could use to merge regions with the aim of ensuring that poverty levels with EAs were relatively homogeneous. This approach could help to ensure that selected samples are more representative, but still the priority is to have excellent population data, and if only low-resolution socio-economic data are available, the combination still can be made. In addition, a compactness metric based on the comparison of the longest length in a shape to the diameter of a circle of same area would provide a more robust selection criterion during the merging process and lead to more compact EAs, as it is insensitive to shape size, unlike the compactness metric used in the AZ tool. Furthermore, if available data have consistent and reliable labeling, the hierarchy approach will be adopted in terms of prioritizing the features to be merged in the merging process. By doing this, traversability will be accounted as well through ranking of split lines, and additional data sources such as the 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. Identified buildings could also help to inform the delineation of rural EAs, as well as providing spatial variables (building size and layout) that may be linked to socio-economic indicators.

Based on the experience in the present study, we 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. The tool has the capability of working directly with shapefiles and a wider. The tool will have the ability to consider several varieties of parameters to produce constructive EAs, as well as the integration of socio-economic variables as described above. It could also have broad development impact since it can update existing EAs with the condition if the old EA boundary should be kept or not or generate the new set of EAs and turning them into national sample frame within time, cost and quality constraints.

Conclusion

Somalia presents an example of a country where the need for basic spatial demographic data such as EAs is crucial to continued development, but where on-the-ground logistical constraints, security and data limitations remain significant barriers. Here, we show how a novel approach to creating predefined census EAs in such contexts where the census data are badly outdated and EA data are needed to progress development efforts. In addition, by highlighting existing freely-available, up-to-date high-resolution gridded population data and settlement maps, we hope to help remove a further barrier to broader uptake of this method. Future research will aim to improve the accessibility of the approach by providing user-friendly tools that will automate the data access, splitting and merging process according to user specifications. By presenting the approach in the challenging context of Somalia and providing the first up-to-date rural EAs in decades, we aim to demonstrate the feasibility of the approach and the potential for application in other contexts.

References

Alazab, M., Islam, M., & Venkatraman, S. (2009). Towards Automatic Image Segmentation Using Optimised Region Growing Technique. In A. Nicholson & X. Li (Eds.), Ai 2009: Advances in Artificial Intelligence, Proceedings (Vol. 5866, pp. 131-+).

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).

Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I., & Nelson, A. (2006). Determining global population distribution: Methods, applications and data. In S. I. Hay, A. Graham, & D. J. Rogers (Eds.), Advances in Parasitology, Vol 62: Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications (Vol. 62, pp. 119-156).

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

Beucher, S. (1994). Watershed, hierarchical segmentation and waterfall algorithm (Vol. 2).

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. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4VT1Q1H. Accessed 09 07 2018.

Center for International Earth Science Information Network - CIESIN - Columbia University. 2017. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 10. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).https://doi.org/10.7927/H4DZ068D. 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. http://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. Environment and Planning A, 43(10), 2399-2418. doi:10.1068/a43601

Damiand, G., & Resch, P. (2003). Split-and-merge algorithms defined on topological maps for 3D image segmentation. Graphical Models, 65(1-3), 149-167. doi:10.1016/s1524-0703(03)00009-2

Ellard-Gray, A., Jeffrey, N. K., Choubak, M., & Crann, S. E. (2015). Finding the Hidden Participant: Solutions for Recruiting Hidden, Hard-to-Reach, and Vulnerable Populations. International Journal of Qualitative Methods, 14. doi:10.1177/1609406915621420

Environmental Data Explorer: Gridded Population of the World. United Nations Environment Programme, Nairobi. 2006. http://geodata.grid.unep.ch/. Accessed 10 May 2018

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 Journal of Photogrammetry and Remote Sensing, 134, 30-42. doi: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

Feng, J., & Watanabe, T. (2015). Index and Query Methods in Road Networks Index and Query Methods in Road Networks (Vol. 29, pp. 1-161).

Finkel, R. A.; Bentley, J. L. (1974). "Quad Trees A Data Structure for Retrieval on Composite Keys". Acta Informatica. Springer-Verlag. 4: 1–9. doi:10.1007/bf00288933

Florczyk, A. J., Ferri, S., Syrris, V., Kemper, T., Halkia, M., Soille, P., & Pesaresi, M. (2016). A New European Settlement Map From Optical Remotely Sensed Data. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(5), 1978-1992. doi:10.1109/jstars.2015.2485662

Flowerdew, R., Feng, Z. Q., & Manley, D. (2007). Constructing data zones for Scottish Neighbourhood Statistics. Computers Environment and Urban Systems, 31(1), 76-90. doi:10.1016/j.compenvurbsys.2005.07.008

Folch, D. C., & Spielman, S. E. (2014). Identifying regions based on flexible user-defined constraints. International Journal of Geographical Information Science, 28(1), 164-184. doi:10.1080/13658816.2013.848986

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

Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M., & Snow, R. W. (2005). Urbanization, malaria transmission and disease burden in Africa. Nature Reviews Microbiology, 3(1), 81-90. doi:10.1038/nrmicro1069

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

Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. Scientific Data, 4. doi:10.1038/sdata.2017.1

Loots H. (2015). Demarcation of census enumeration areas for the 2016 population and housing census in Lesotho. Geomatics Indaba Proceedings 2015 – Stream 1 http://www.ee.co.za/wp-content/uploads/2015/08/Hennie-Loots.pdf

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

Martin R. and 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, D. J, (2002), Geography for the 2001 Census in England and Wales, Population Trends, 108, pp7-15.

Openshaw, S, (1977), A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling, Transactions of the Institute of British Geographers, NS 2, pp459-472.

Pesaresi, M., Guo, H. D., 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(5), 2102-2131. doi:10.1109/jstars.2013.2271445

Roy Chowdhury, P. K., Bhaduri, B. L., & McKee, J. J. (2018). Estimating urban areas: New insights from very high-resolution human settlement data. Remote Sensing Applications: Society and Environment, 10, 93-103. doi:10.1016/j.rsase.2018.03.002

San, D. K., & Turker, M. (2010). BUILDING EXTRACTION FROM HIGH RESOLUTION SATELLITE IMAGES USING HOUGH TRANSFORM. In K. Kajiwara, K. Muramatsu, N. Soyama, T. Endo, A. Ono, & S. Akatsuka (Eds.), Networking the World with Remote Sensing (Vol. 38, pp. 1063-1068).

Somalia National Development plan 2016. Somalia National Development plan (SNDP) – Towards recovery, democracy and prosperity 2017-2019. http://extwprlegs1.fao.org/docs/pdf/som169866.pdf

Steele, J. E., Sundsoy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., . . . Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. Journal of the Royal Society Interface, 14(127). doi:10.1098/rsif.2016.0690

Tatem AJ, Gething PW, Pezzulo C, Weiss D and Bhatt S, 2014, Development of High-Resolution Gridded Poverty Surfaces, Report for the Bill and Melinda Gates Foundation: http://www.worldpop.org.uk/resources/docs/Poverty-mapping-report.pdf

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa. Plos One, 2(12). doi:10.1371/journal.pone.0001298

Thomson, D. R., Hadley, M. B., Greenough, P. G., & Castro, M. C. (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. doi:10.1186/1471-2458-12-959

Turner A. G. (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

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). http://somalia.unfpa.org/sites/default/files/pub-pdf/Population-Estimation-Survey-of-Somalia-PESS-2013-2014.pdf

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

United Nations Statistics Division 2010. Definition of the National Census Geography. UNSD-CELADE Regional Workshop on Census Cartography for the 2010 Latin America's census round. https://www.cepal.org/celade/noticias/paginas/8/35368/pdfs/3UNSD.pdf

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., & Ieee. (2015). BUILDING DETECTION IN VERY HIGH RESOLUTION MULTISPECTRAL DATA WITH DEEP LEARNING FEATURES 2015 Ieee International Geoscience and Remote Sensing Symposium (pp. 1873-1876).

Vijayaraj, V., Bright, E. A., Bhaduri, B. L., & Ieee. (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 (pp. 278-281).

Wanderer, J. P. (2017). Analysis of Large and Complex Data. Anesthesia and Analgesia, 125(1), 345-345. doi:10.1213/ane.0000000000002127

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., . . . Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences of the United States of America, 115(14), 3529-3537. doi:10.1073/pnas.1715305115

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

Yacyshyn A. M. & Swanson D. A. (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

Appendix 1. Data sources used to create the 'rural settled areas' raster (100m)

Source	Data description	Transformation / raster creation 100m	Observation
UNFPA/PESS urban enumeration areas.	Polygons.	Urban EAs were rasterized at 100m.	-Matches satellite imagery but some urban areas are omitted. Rectangular shape around Northern cities instead o an outline (see Fig. A4.1: Hargeisa) -Needed to define if settled areas are rural or urban
UNFPA/PESS rural population estimates.	Points: centre of surveyed area (georeferenced either at settlement centre or at district centres in larger settlements) - No area size given.	Dilated circle radius 200m from point coordinates on a 100m raster grid to obtain settlement shape.	Large settlements have many recorde points, so dilating at 200m covers the settlement. Areas obtained mostly match settled areas from satellite imagery (circle may be too big or too small). Some regions are missing fron the dataset. If the recorded point is isolated then tl settlement shape is a lozenge of maximum 5 pixels length (500m).
OSM 'residential areas' (land use tag used by OSM)	Part points, part polygons. Volunteer-reported.	-Dilated circle radius 100m (i.e. cross) from point coordinates to obtain settlement shape. -Dilated circle radius 200m from polygons. -Merged both.	Areas obtained match settled areas from satellite imagery.
OSM 'buildings' (OSM category).	Centroid of polygons. Volunteer-reported.	Dilated circle radius 100m (i.e. cross) from point coordinates.	Areas obtained match parts of settled areas from satellite imagery.
OSM 'places': 'hamlets' and 'villages'. (OSM category).	Points. Volunteer-reported.	Dilated circle radius 200m from point coordinates to obtain settlement shape.	Areas obtained mostly match settled areas from satellite imagery (circle m be too big or too small). Possibly som outdated data or temporary or nomad settlements.
DLR GUF.	Binary raster.	Dilated circle 100m from settled pixels to fill in gaps.	Areas obtained match settled areas from satellite imagery.
Gates / Digital Globe (DG) population estimates.	Scattered points arranged on a sparse 50m grid.	Raster at 100m resolution was created with pixels with population estimate = 1 and the rest = 0.	Areas obtained match settled areas from satellite imagery. Dataset only covers the North of the country.
World Bank / Flowminder: Manual building count from Google Satellite imagery.	Polygons.	Polygons in which we observed at least 1 structure were rasterized.	Areas obtained match parts of settled areas from satellite imagery.

Appendix 2. Data sources used to obtain population estimates at 100m in urban areas.

Source	Data description	Transformation / raster creation 100m	Observation
UNFPA/PESS urban Enumeration areas.	Number of households per EA.	-Urban EAs were rasterized at 100m. -number of households in an EA was divided by the number of pixels in the EA.	-Number of household per pixel. -Pixels within a given EA have the same val (the average number c households within the EA).
Gates / Digitalglobe (DG) population estimates: Data derived from automated building delineation.	-Number of people given at points scattered on 50m resolution grid. -likely overestimated according to the data source (DG). -North of the country only. -does not fully cover the UNFPA/PESS urban EAs as these EAs sometimes include areas with no building.	-Population estimates falling on the same pixel on a 100m resolution grid were added up to create a raster of 100m resolution.	Number of people per pixel.
World Bank / Flowminder: Manual building count from Google Satellite imagery.	Polygons with the total number of buildings in each. Polygons may several km long.	-Polygons were rasterized at 100m. -the number of pixels in the block divided the number of counted building in the block.	-Number of buildings per pixel. -Pixels within a given polygon have the sam value (the average number of buildings within the polygon).
OSM 'buildings'.	Points: centroid of building polygons. -Number of buildings is often underestimated in dense urban areas -Not all buildings are delineated. -Volunteer-reported.	-Buildings falling on the same pixel on a 100m resolution grid were added up to create a raster of 100m resolution.	-Number of buildings per pixel.

Appendix 3. Data sources used to obtain population estimate at 100m in settled rural areas.

Source	Data description	Transformation / raster creation 100m	Observation
UNFPA/PESS rural household ground counts.	Household number georeferenced either at settlement centre or at district centres in larger settlements.	-Dilated circle radius 200m from point coordinates to obtain settlement shapes (there may be more than 1 survey points in the resulting settlement shapes). -We identified each settlement using connected component labelling. -For each settlement, we computed the number of households by summing the ground count associated with each survey point. -Instead of dividing by number of pixels in settlement, we scaled the number of household per pixels a settlement by the pixel distance to the settlement border: number of households in pixel = number of households in settlement * pixel distance to border / (sum of all pixel distance to border in settlement). -We placed highest values in settlement centres (around the coordinates given) as the settlement extent was not provided in the data	-Number of households per pixels. -Number of households per settlement match UNFPA data but not all pixels within a settlement have the same value: pixels most distant to settlement border (central pixels) have higher values
Gates / Digital Globe (DG) population estimates: Data derived from automated building delineation.	-Number of people given at points scattered on 50m resolution grid. -likely overestimated according to data source (DG). -North of country only.	-Population estimates falling on the same pixel on a 100m resolution grid were added up to create a raster of 100m resolution.	Number of people per pixel.
World Bank / Flowminder: Manual building count from Google Satellite imagery.	Polygons with a total number of buildings in each. Polygons may several km long.	-Polygons were rasterized at 100m. -the number of counted building in the block was divided by the number of pixels in the block.	-Number of buildings per pixel. -Pixels within a given polygon have the same value (the average number of buildings within the polygon).
OSM 'buildings'.	Points: centroid of building polygons. -Number of buildings is often underestimated in dense urban areas	-Buildings falling on the same pixel on a 100m resolution grid were added up to create a raster of 100m resolution.	-Number of buildings per pixel.