INTERNATIONAL JOURNAL OF DIGITAL EARTH
2020, VOL. 13, NO. 1, 61–77
https://doi.org/10.1080/17538947.2018.1548656

Enhanced data and methods for improving open and free global population grids: putting 'leaving no one behind' into practice

Sergio Freire [a,b], Marcello Schiavina [a], Aneta J. Florczyk [a], Kytt MacManus[c], Martino Pesaresi [a], Christina Corbane [a], Olena Borkovska[c], Jane Mills[c], Linda Pistolesi[c], John Squires[c] and Richard Sliuzas[b]

[a]European Commission, Joint Research Centre (JRC), Ispra, Italy; [b]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, Netherlands; [c]Columbia University CIESIN Geoscience, Palisades, NY, USA

ABSTRACT

Data on global population distribution are a strategic resource currently in high demand in an age of new Development Agendas that call for universal inclusiveness of people. However, quality, detail, and age of census data varies significantly by country and suffers from shortcomings that propagate to derived population grids and their applications. In this work, the improved capabilities of recent remote sensing-derived global settlement data to detect and mitigate major discrepancies with census data is explored. Open layers mapping built-up presence were used to revise census units deemed as 'unpopulated' and to harmonize population distribution along coastlines. Automated procedures to detect and mitigate these anomalies, while minimizing changes to census geometry, preserving the regional distribution of population, and the overall counts were developed, tested, and applied. The two procedures employed for the detection of deficiencies in global census data obtained high rates of true positives, after verification and validation. Results also show that the targeted anomalies were significantly mitigated and are encouraging for further uses of free and open geospatial data derived from remote sensing in complementing and improving conventional sources of fundamental population statistics.

ARTICLE HISTORY
Received 28 June 2018
Accepted 11 November 2018

KEYWORDS
Population statistics; census; built-up areas; GPW; GHSL

1. Introduction

Accurate geospatial data on global population distribution and characteristics are increasingly required and relied upon for analysis and modelling in an expanding range of disciplines (Gaughan et al. 2014; Wardrop et al. 2018). These data are also crucial in the frame of the Digital Earth perspective in the path towards decision-making (Shupeng and van Genderen 2008). In the context of evidence-based assessment for policy support, the recent wave of post-2015 international development agreements (including the Sendai Framework for Disaster Risk Reduction 2015–2030 (adopted March 2015); United Nations (UN) 2030 Agenda for Sustainable Development (SDGs, September 2015); COP 21 Paris Agreement on Climate Change (November 2016); UN New Urban Agenda (December 2016)) places great demands and responsibility on geospatial data, and in particular on that related to population.

CONTACT Sergio Freire sergio.freire@ec.europa.eu European Commission, Joint Research Centre (JRC), Via E. Fermi, 2749, Ispra, VA 21027, Italy

Public information about the characteristics and spatial distribution of population is a strategic resource supporting the monitoring and implementation of international frameworks. Information about population density per uniform spatial sampling schemas (e.g. grid cells) are necessary for modelling critical aspects of human-environment interaction, such as: Exposure (to hazards, pollutants); Access (to resources, services, facilities); and Impacts, in both perspectives of (i) the impact of human activities on the planet and (ii) impact of the environment on the people living on the Earth's surface (natural disasters, environmental change) (UNISDR 2015; UNDESA 2016; UNECOSOC 2016). The Sustainable Development Goals (SDG) indicators are 'action oriented, global in nature and universally applicable' and should be themselves *sustainable* and comparable across space and time (UNECOSOC 2016, 7). Ideally, the indicators supporting the international development agenda should be based on geospatial population data that is up-to-date, sufficiently detailed, accurate, consistent, cost-effective (i.e. sustainable), transparent (i.e. using clear methods), and accessible to all. These development agreements, their targets and monitoring needs (i.e. indicators) present an opportunity not only for producing more data, but also for improving existing datasets. The reliability of international statistics is periodically called into question (e.g. Wolff, Chong, and Auffhammer 2011), and the recent surge in initiatives aimed at producing more accurate and detailed statistics (e.g. by UN (UNDP 2018), World Bank (WB 2018), Bill & Melinda Gates Foundation and UK Department of International Development) are a realization that statistical capacity must be strengthened (UNECOSOC 2016).

Regular grids (raster or vector) are now well established and widely used spatial structures to model and report population attributes, and their advantages are largely recognized (Deichmann, Balk, and Yetman 2001). Several global gridded population datasets are currently produced. They are differentiated by the underlying population concept (e.g. 'resident' vs. 'ambient' population), gridding method, and distribution policy (for reviews and recent developments see Linard and Tatem 2012; Stevens et al. 2015; Freire et al. 2016). For a given spatial unit, population distribution grids can be produced by disaggregating (e.g. gridding) population counts (top-down approach) or by estimating that count at the grid cell level through combining sampling with ancillary data (bottom-up method) (Wardrop et al. 2018). Global population grids aiming to support policies and international agreements in global forums are typically based on available statistics whose national totals match, or are adjusted to match, those used in UN population estimates (e.g. UNDESA 2015). UN estimates are used as a standard in order to produce more accurate and harmonized datasets for countries where census data are considered to be less reliable, and because of the extensive work done by the UN Population Division to adjust and correct census data post enumeration. Examples of such grids include Gridded Population of the World (GPW) (Deichmann, Balk, and Yetman 2001; Doxsey-Whitfield et al. 2015) the Global Rural-Urban Mapping Project (GRUMP) (Balk et al. 2006), and the Global Human Settlement Population (GHS-POP) (Freire et al. 2016).

For population grids that are generated by disaggregating available statistics (top-down) while preserving their volume, output accuracy (with respect to ground truth data) is largely dependent on the spatial detail and quality of input geospatial layers. In addition to their spatial detail, population census-like layers vary widely in quality concerning their geometry, attributes, reliability, and currency. Incidentally and unfortunately, the quality and reliability of statistics' are particularly low in many developing countries (Tatem et al. 2007), at whom the post-2015 development agenda is especially directed.

Production of global grids relies on population sources that are somewhat heterogeneous in respect to these characteristics. These sources include provisional or final censuses originating in National Statistical Offices (NSOs) as well as estimates and projections provided by Non-governmental (NGOs) and other organizations. Despite valuable efforts to collect, integrate, and improve global census data obtained from disparate sources (e.g. GPW), some important requirements remain unsolved. Addressing the universal inclusiveness requisite associated with the UN 'leaving no one behind' imperative of the new sustainable development agenda (UNDESA 2016) requires that all people are counted and accounted for in the place they live. However, census enumerations

often do not meet this requirement (even sometimes by design as in *de facto* census), and any statistical sampling inherently lacks information about marginal and uncounted populations.

Perhaps surprising and unknown to some users, available population statistics also suffer from error and uncertainty that affect even the most basic of demographic variables (i.e. total population). While uncertainty is difficult to assess and communicate, error often propagates unimpeded to downstream analyses and applications. The main deficiencies and shortcomings affecting geospatial population statistics can be summarized in the following basic taxonomy:

1. Issues affecting geography (census reporting units):
   a) Lack of spatial detail (i.e. too coarse or generalized, related to spatial precision)
   b) Low spatial accuracy (i.e. units partially displaced or completely misplaced)
2. Issues affecting attributes (population counts in present case)
   a) Undercounting and/or underreporting (i.e. underenumeration)
   b) Over-counting and/or over-reporting (i.e. overenumeration)

These issues have been unnoticed or unaddressed due to the lack of external accuracy assessment related to a shortage in independent compatible high-resolution reference data. Notable exceptions are the works of Hay et al. (2005) and Mondal and Tatem (2012) that discuss deficiencies in the representation of population distribution along coastlines, and Linard and Tatem (2012) that address these in frame of infectious disease research. In contrast to remote sensing-based maps where validation is expected to rely on independent reference data, typically collected from ground truthing (see Congalton and Green 2009), quality assessment of population grids is usually limited to internal consistency of model performance (i.e. population data used for validation is the same that was used for modelling, just finer). Also, the relative coarse resolution of global grids produced until recently (i.e. ~1–5 km) mitigated the impact of some deficiencies that now become apparent as cell sizes approach ~100 m.

Significant improvements to the quality of population grids are to be gained by addressing some of these shortcomings. Detection and mitigation of deficiencies in population statistics require independent, reliable, higher-resolution data. Remote sensing imagery and methods have been evolving towards constituting a more detailed, objective and independent data source on human presence on the Earth surface. The combination of new cost-effective, automated and fully replicable data classification methods (e.g. machine learning) with the synoptic capacities of satellite Earth Observation imagery, made accessible in a public, full open-and-free frame, can contribute to fill information gaps and supplement existing statistics by mitigating some major shortcomings in population data. This is especially true in poor, remote, unsafe, disputed, very large, and/or highly dynamic areas of the globe where conventional data gathering and updating is challenging.

Current methods and processing capacity allow for global mapping of built-up areas and settlements with unprecedented spatio-temporal detail and accuracy – in essence capturing the local scale with global coverage, finally starting to fulfil a long-standing promise of remote sensing technology. Making these datasets available open and free helps to increase access, promotes transparency, and ensures accountability of the information produced. Global, consistent and updated geospatial data such as that made openly available in the framework of the Global Human Settlement Layer (GHSL) (Pesaresi and Ehrlich 2009) are already providing an effective contribution and improving the disaggregation of census data into derived population grids (Freire et al. 2015; Linard et al. 2017; Nieves et al. 2017). However, it remains to be tested in large scale if such remotely sensed data can also assist in assessing and mitigating major deficiencies present in geospatial population statistics.

This article addresses shortcomings present in global collections of census data, by introducing, applying, and discussing novel procedures aimed at investigating and detecting some of those major anomalies. It then demonstrates mitigation of some inconsistencies using high-resolution geospatial data derived from remote sensing. The focus is on deficiencies in mapping coastlines and declared unpopulated areas, as extreme instances of deficiencies 1a) and 1b) and 2a) in the taxonomy

mentioned above. Making use of open and free high-resolution global settlement layers derived from contemporary satellite imagery, the approaches are illustrated with GHSL in the frame of producing GHSL-based new global population grids (release 2018).

2. Materials and methods

Two main types of geospatial data were used in this work: vector-based census data reporting on estimated total population counts for the target year of 2015 (http://sedac.ciesin.columbia.edu/gpw), and raster layers reporting on built-up presence for 2013/2014 derived from Landsat image collections in the frame of the GHSL project (http://ghslsys.jrc.ec.europa.eu/index.php). Some aspects of these two datasets are compared and combined, with remote sensing data being used to identify anomalies in the geospatial census data. These anomalies are mitigated with the support of the remote sensing-derived data after verification and validation.

2.1. Data sets

2.1.1. Geospatial data on census population

As a source of census population data, a database assembled by the Center for International Earth Science Information Network (CIESIN) in the frame of the Gridded Population of the World project (Tobler et al. 1997) was used. For more than two decades, GPW has been collecting, combining, and harmonizing available population census and estimates into what is considered the most complete, detailed and coherent census-based geodatabase available globally. This database is periodically updated with more recent and improved data, with GPW employing clear and transparent methods to create open and free residence-based population grids for different reference years (Doxsey-Whitfield et al. 2015). Despite these efforts, the GPW databases are subject to availability and quality of source (national) population statistics, and therefore inherit their gaps and shortcomings, as reliability and currency of population data is quite heterogeneous among countries (for GPW metadata see CIESIN 2017a).

The GPW data used consisted in country-based layers (one for each of 241 countries) of census and administrative polygons containing estimated residential population for the GHSL target years 1975-1990-2000-2015, adjusted to country-level estimates of UN World Population Prospects 2015 (UNDESA 2015). Due to the development of the work and the updating schedule of GPW data, two versions from the same release were used: GPWv4, released in 2016 (CIESIN 2016), was used for detection of inconsistencies along coastlines; GPWv4.10, available in late 2017 (CIESIN 2017b), was used for revision of unpopulated areas. GPWv4.10 is a revision of GPWv4 with boundary or population updates for 64 countries. More details about these data can be found here: http://sedac.ciesin.columbia.edu/data/collection/gpw-v4.

2.1.2. Built-up areas from remote sensing

In the frame of the Global Human Settlement Layer (GHSL) project, global built-up (BU) areas were recently mapped with unprecedented spatial detail, consistency, and temporal coverage (Pesaresi et al. 2016). The GHSL has developed and employed novel and automated approaches to produce a time series of raster layers reporting on the presence of building structures, defined as 'all constructions above ground intended for human or animal sheltering or for the production of economic goods' (Pesaresi et al. 2013). These data were derived from Landsat image collections spanning four periods: 1975, 1990, 2000 and 2013–2014 and are made available open and free. Quality assurance was conducted to validate these data against a heterogeneous set of available layers mapping building footprints (Pesaresi et al. 2016). More recently, independent spatio-temporal quality assessment of the GHSL built-up time series was performed for the USA showing very encouraging accuracy that generally increases over time (Leyk et al. 2018). Most relevant for the present work, these built-up areas exhibited strong correlation with population distribution and density, and suitability

for population disaggregation and modelling (Freire et al. 2015, 2016; Linard et al. 2017; Nieves et al. 2017).

In this effort, global layers mapping built-up areas for the latest epoch (2013–2014), at spatial resolutions of 38 and 250 m, were used as an indication of the presence of human settlements. Due to the development timeline of this work and the updating schedule of GHSL products, data from different releases were used: the GHS-BUILT data released in 2015 (GHS_BUILT_LDSMT_GLOBE_R2015B) were used for combination with census data from GPW for initial flagging and systematic detection of discrepancies (Pesaresi et al. 2015a); while the latest and improved GHS-BUILT data (GHS_BUILT_LDSMT_GLOBE_R2018A), produced in Fall 2017 (Corbane et al. 2017), to be publicly released in Fall 2018, were used for population disaggregation and creation of final population grids. All geospatial layers were projected to World Mollweide projection (EPSG 54009), the equal-area projection adopted for production of the GHSL global population grids. This is also important for the processing undertaken and described here, which is mostly based in quantification and comparison of surfaces.

2.2. Revision of 'unpopulated' units

An extreme case of issue 2a) of the taxonomy of census anomalies proposed in Section 1. is the reporting of areas as 'unpopulated' or 'uninhabited' or otherwise as containing no population. During the early stages of production of GHSL global population grids, it was observed that some units declared as having no population in census data in fact contained settlements and significant areas of built-up according to the GHS-BUILT data. Consequently a spatial analysis procedure was designed to (i) check for the presence of (resident) population in census units where no population was reported, and (ii) mitigate these issues while minimizing changes to the input census data (conservative approach). Units deemed as 'uninhabited' or otherwise unpopulated in the census data were critically assessed for the presence of significant residential population, based on ancillary data (e.g. coordinates of populated places) and very high resolution (VHR) imagery. Such census units were then selected based on their size and the extent of built-up surface in 2014 as an indication of the likely presence of settlements. Due to the experimental nature of the work, a sequential approach was adopted in order to inspect, control, and calibrate the assumptions made at each step. Adopting a conservative approach was important to quickly narrow the selection to the main issues and limit the rate of false positives.

For selection of potentially problematic census units, a simple semi-heuristic rule-based approach was followed. From the set of census units in GPWv4.10 having no population (238,029 units), most non-residential zones were filtered out by selecting polygons larger than 300 ha (3 km^2). This threshold was adopted because it is the mean size of non-residential built-up patches (i.e. industrial, commercial, and service facilities) in the CORINE Land Cover (CLC) 2012 vector map, v18.5.1 (EEA 2016). CLC2012 currently covers 39 countries and 5.8M km^2 in Europe, and this threshold may not be the best one to use in other contexts.

From these 47,508 census units larger than 300 ha, units were selected that contained a total surface of built-up in 2014 (according to GHS-BUILT R2015B) greater than 10 ha, to increase confidence in presence of significant resident populations. As the resulting set of 2,717 units was still too large a number for visual inspection with available resources, and after confirming it would include a high rate of false positives in countries that have both detailed and accurate censuses, countries deemed, based on our in-depth empirical knowledge of the data, as having reliable census (14 of the 45 countries in set) were excluded. This led to the identification of potentially problematic census units in 31 countries, to be visually inspected for verification and validation of the approach. These units were inspected using VHR imagery from web mapping services (Esri World Imagery service through the ArcMap application).

After collecting sufficient visual evidence that some units were inhabited (i.e. presence of residential- type buildings with signs of habitation), that anomaly had to be mitigated. A decision was taken

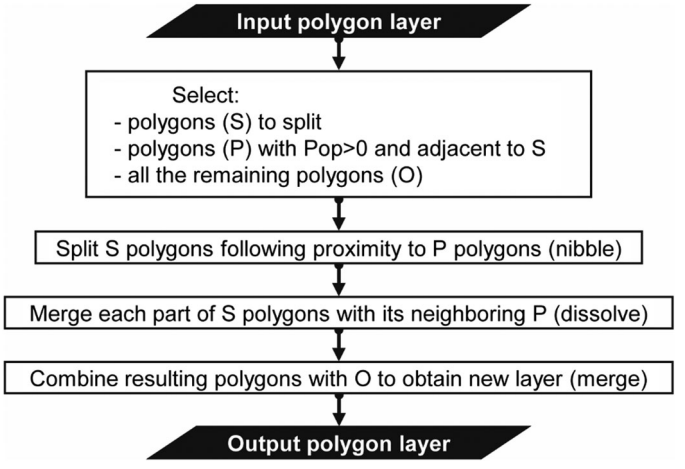

Figure 1. Flow chart of the 'split and merge' approach.

to adopt the most conservative approach that would imply minimal changes to the GPW's geospatial census dataset and still ensure that some population would be accounted in those areas, to be then disaggregated to the mapped BU areas in the creation of GHS-POP grids. To implement this, a 'smart' geoprocessing approach was devised and automated (hereafter called 'split and merge') to assign population to those previously unpopulated areas without altering the boundaries and the total population of the upper administrative level. The approach involves splitting and merging the confirmed problematic polygons, based on geographical proximity to those ones adjacent and containing population, and merging the split parts to the latter.

Figure 1 shows the main steps involved in the application of the 'split and merge' approach. This method consists in spatially dividing the problematic polygons according to a proximity rule accounting for adjacent populated polygons belonging to the same administrative unit. This task is performed by (i) re-projecting the polygons to an equidistant projection, more suitable for measuring distances (i.e. World Equidistant Cylindrical), (ii) rasterizing to a sufficiently fine resolution (i.e. a trade-off between accuracy and computational demand) with an all-touched approach, (iii) applying a nearest-neighbour nibbling technique (only considering populated neighbouring units), (iv) polygonising the resulting part and finally (v) re-projecting back to the original projection. The polygons obtained are then cleaned of geometric artefacts by clipping them to the original boundary. Each generated part of the problematic polygon is then dissolved with the related populated neighbouring unit, in this way 'populating' the whole original problematic area.

2.3. Harmonization of population and settlement data along coastlines

Seashores and waterfronts can be especially intense and dynamic zones, due both to natural processes and their attractiveness for settlement and building of infrastructure (EEA 2006; McGranahan, Balk, and Anderson 2007; Freire, Santos, and Tenedório 2009). These strong and fast dynamics contribute to making census or administrative geometries outdated or inaccurate. Moreover, the geospatial census data are not produced with globally harmonized technical specifications, thus including different nominal scales and spatial tolerance characteristics. The characteristics of the GHSL layer mapping built-up areas (i.e. seamless coverage, uniform technical specs, decametric-scale spatial detail, currency) have revealed discrepancies with GPW along coastlines (including inland water bodies). Most of these consisted in the coastline from GPW being inland respect to that actually detected by GHSL, implying in practice that no population would be disaggregated to those mapped

built-up areas in GHSL that were water bodies in GPW (i.e. spatial mismatch between source and target zone). This would cause changes to the population density and extent of coastal settlements or omit them entirely in the resulting population grids.

Therefore, a systematic procedure was developed aiming at identifying globally the main inconsistencies between the two datasets along coastlines, and efforts were undertaken to reconcile them. This procedure was aimed at harmonizing population and settlement data along coastlines, but does not intend to harmonize the delineation of water bodies themselves. The high-resolution (38 m) GHSL layer on built-up areas for 2014 (GHS-BUILT from R2015B), available at the time of work, was used to detect the significant human presence (i.e. BU) beyond GPWv4 censuses coastlines. The process of detecting inconsistencies included the following three main steps: (i) creation of a layer depicting the built-up areas outside of the GPW landmass, (ii) identification of potentially inhabited patches containing built-up, and (iii) visual inspection and validation of the patches. It is well known that full automatic detection of built-up areas from satellite imagery is prone to commission errors in sandy or rocky landscapes, and, as a result, part of the built-up area pixels in coastal areas might be false positives. Therefore, a visual inspection was required. It should be noted that the term coastline here refers to inland water bodies as well as seashores, and the level of detail at which coasts are outlined varies widely across countries (e.g. the inland water bodies in the USA are depicted with high detail while in most other countries only the large ones are outlined with low detail).

During step (i), first a water mask derived from the country polygons of GPWv4 data was rasterized at approximately 38 m, and only within the data domain of the GHSL layer as defined by the GHSL Datamask product (Pesaresi et al. 2015b). Then, this coastal water layer was intersected with the GHS-BUILT layer, producing the built-up areas not overlapping with any GPW country polygons. In total, these areas have amounted to 6,142 km^2 of surface. The procedure for selecting and validating potentially inconsistent patches included three steps: (1) vectorizing connected cells classified as built-up surface (using 4-connectivity rule) into individual polygons (nearly 800,000 patches, with a mean surface of 7700 m^2, corresponding to 5/6 contiguous pixels), (2) selecting all patches larger than 1 km^2 (287 selected), and (3) visual inspection using GIS software and very high resolution (VHR) satellite imagery from web mapping service (Google Maps) to confirm presence of buildings. The 1 km^2 size threshold was chosen to balance importance and potential impact of inconsistencies with resources available for verification and validation. Figure 2 shows examples of built-up patches located beyond the coastline of GPW in Japan, Tunisia, China, and Romania, assessed during visual inspection. For the selected patches (those larger than 1 km^2) in which visual inspection confirmed presence of built-up surfaces, mitigation consisted in reconciling the outline of GPW census units with built-up patches by manually extending the units on the coast to include these areas.

After conducting automatic detection and during visual inspection of the main coastal discrepancies between GPWv4 and GHSL, systematic or additional inconsistences were identified. These systematic inconsistencies between the more recent and improved GHS-BUILT P2018 and GPW4.10 layers were still present in some countries: Japan, Ukraine, Switzerland, and France. Therefore, an automated approach to mitigate these issues that appear to be caused by the coarser scale (low spatial detail) and/or positional errors (poor georeferencing) of census reporting units was developed and applied. For inconsistences detected in lakes (i.e. Switzerland and France) the 'split and merge' approach (described in Section 2.1) was directly applied to polygons representing lakes in order to split them and dissolve each part with the nearest neighbour polygon. In coastal areas, such as in Ukraine and Japan, the GPW coastline was first extended seawards using a 2-km external buffer. Next, this buffer was split according to proximity to the original census units by using the split and merge approach, and then assigned each part of the buffer to a census polygon (see Figure 8). In this case, the whole set of census polygons was used, without selecting only the populated areas.

3. Results and discussion

3.1 Revision of unpopulated units

The implementation of the semi-automated rule-based procedure to the initial 238,029 census units deemed as uninhabited in GPW geospatial census data resulted in the selection of 128 units in 31 countries as potentially containing significant resident population (Figure 3).

All of these 128 units were validated visually with VHR imagery from web mapping services (Esri/Bing, Google maps). This was both to proceed with confidence in such a sensitive issue, and to assess the validity of the semi-heuristic rule-based approach used for selection of problematic units. The validation with VHR imagery revealed the presence of significant residential areas having clear signs of habitation in 76 census units, belonging to 19 countries (Table 1), resulting in a success rate of the automated procedure of around 59% (of polygons). These 76 problematic units have a combined area of 297,090 km^2, of which 624 km^2 are reported by GHSL to be covered by building structures in 2014. If this built-up surface is accurate and all of it is dedicated to residential function,

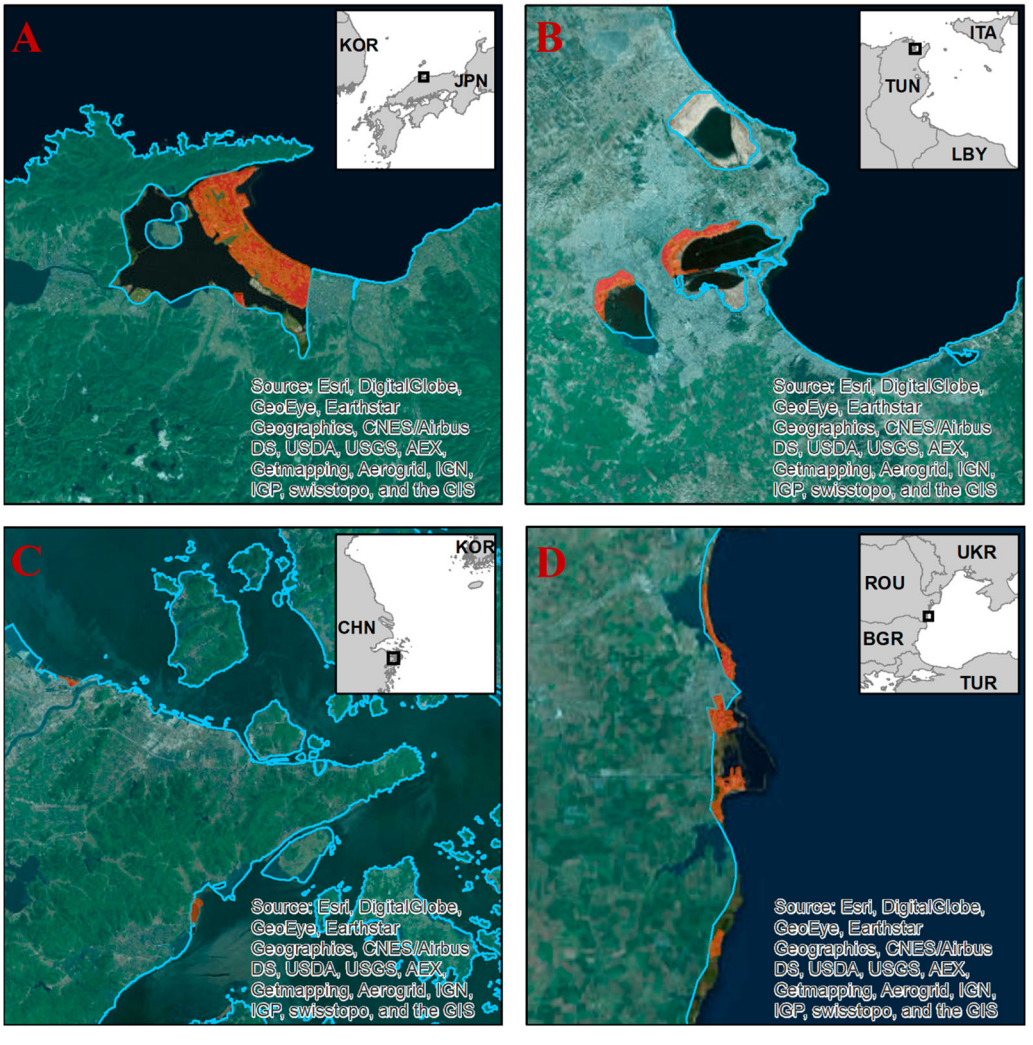

Figure 2. Examples of detected discrepancies (patches larger than 1 km^2) between GPWv4 (blue line) and GHS-BUILT 2014 (orange) along the coasts of (A) Japan (JPN), (B) Tunisia (TUN), (C) China (CHN), and (D) Romania (ROU).

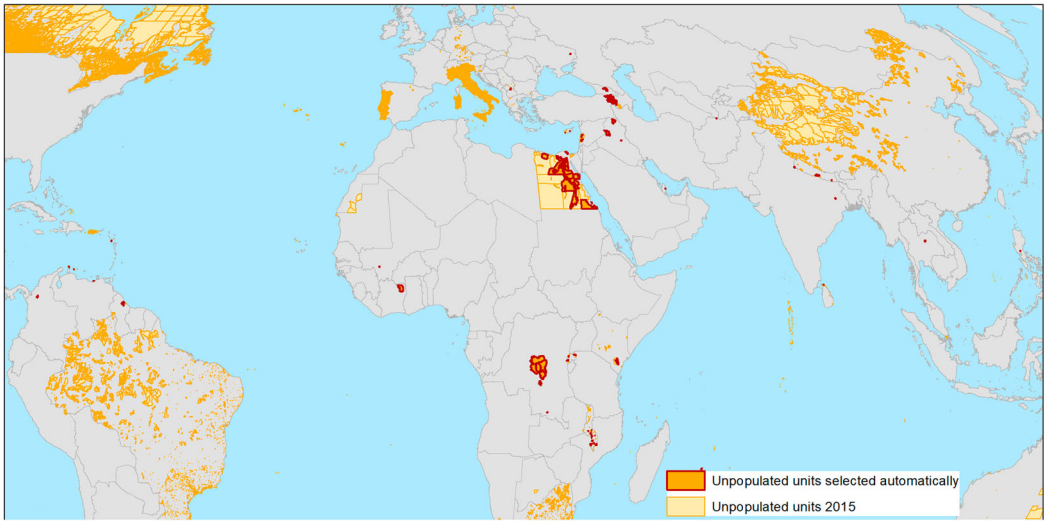

Figure 3. Global distribution of census units without population in GPWv4.10 data and those flagged based on the automated procedure developed (USA not considered). Note that existence of unpopulated units is merely a feature of census design and not of the landscape.

using a low occupancy rate of 100 m^2 of built-up surface per person would put a conservative estimate of under-reported population in these units at more than 6 million people. Some recent estimates put the value of 'missing' (i.e. undercounted) people in urban slums at 369 million worldwide (Carr-Hill 2017).

Mitigation of these anomalies was carried out using the 'split and merge' approach, which was applied to a total of 58 units covering about 296,670 km^2, in the following countries: Afghanistan, Armenia, Democratic Republic of the Congo (DRC), Colombia, Cyprus, Egypt, Georgia, Guyana, Iraq, Lebanon, Mali, Malawi, Nepal, Rwanda, Thailand, and Ukraine. Among them, the Democratic Republic of the Congo and Egypt are the most problematic, with the surface of incorrectly labelled

Table 1. List of countries in which problematic polygons were selected.

Country	*N*	Area (km^2)	BU area 2014 (km^2)
Afghanistan	1	76.7	0.1
Armenia	1	53.2	2.0
Democratic Republic of the Congo	7	99,575.9	154.9
Colombia	1	514.1	0.1
Cyprus	1	4.4	0.4
Egypt	16	184,254.7	283.2
Georgia	2	2005.2	1.8
Guyana	1	918.6	0.7
India	1	116.4	70.3
Iraq	4	8459.9	65.7
Lebanon	4	56.1	0.7
Mali	2	53.5	14.0
Malawi	14	121.1	4.7
Nepal	1	163.9	1.7
Philippines	1	86.6	0.9
Rwanda	1	37.7	0.2
Serbia	16	214.5	7.5
Thailand	1	306.4	2.1
Ukraine	1	71.1	12.6
Total	76	297,090	623.7

Note: For each country the number of polygons (*N*), the sum of the area of such polygons and the total built-up (BU) surface accounted by GHSL built-up layer in 2014 is reported.

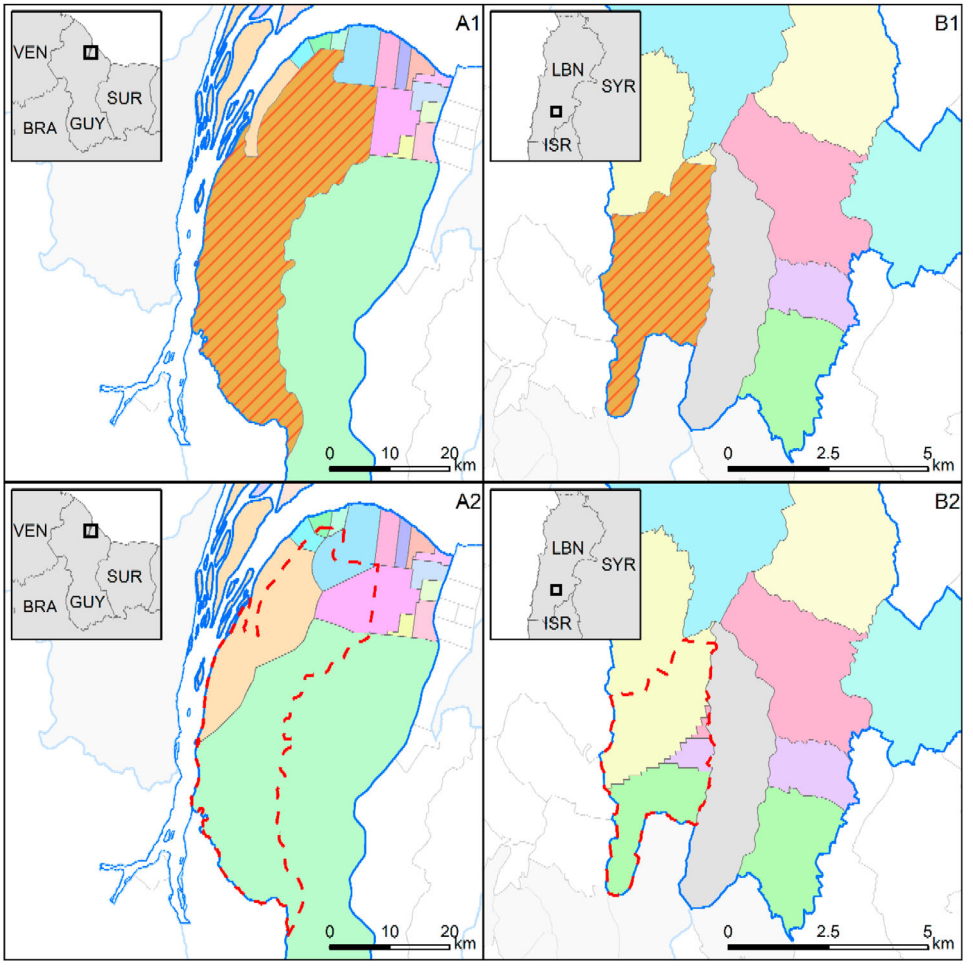

Figure 4. Examples of application of the 'split and merge' approach to census units incorrectly deemed as unpopulated in (A) Guyana and (B) Lebanon, before (1) and after (2) the procedure is applied. Orange filling and red dashed boundary represent the original problematic polygon; blue solid line delineates borders of upper administrative level; grey solid polygons are correctly declared unpopulated polygons; solid coloured areas are the resulting polygons; areas outside the processed administrative unit are shaded.

'unpopulated' units reaching 99,576 km^2 and 184,255 km^2, respectively. In Egypt some problems seem to be caused by the coarse spatial detail and spatial inaccuracy of census polygons.

Figure 4 shows the results of the application of the 'split and merge' approach in Guyana and Lebanon. In Guyana, the declared unpopulated polygon (Figure 4(A1)) is split into seven parts and is then dissolved with adjacent polygons belonging to the same administrative level (Figure 4(A2)). In Lebanon the setting is more complex: the target polygon (Figure 4(B1)) is split into four parts which are then dissolved with the neighbouring administrative units, three of which are to the East of a unit (in grey) correctly declared as uninhabited. The resulting polygons, preserving their respective population counts, replace the problematic one in the census database that is used as one of the inputs (source zones) in the production of population grids (Figure 4(B2)).

Ultimate results and benefits for population disaggregation are illustrated in Figure 5, for the same areas in Guyana and Lebanon, showing that final population distribution grids for year 2015 include populated cells in areas previously deemed as uninhabited.

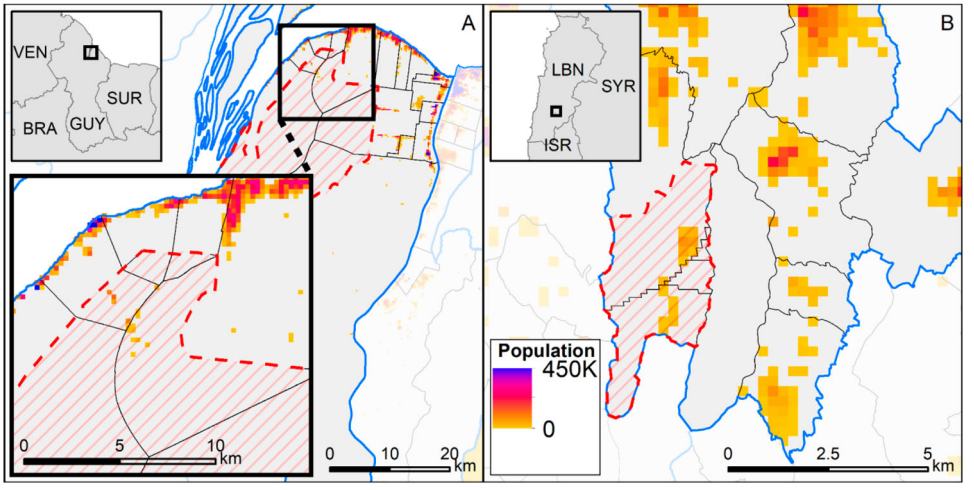

Figure 5. Illustration of resulting 250-m population grids for 2015 in (A) Guyana and (B) Lebanon.

The procedure for mitigating and 're-populating' units previously unpopulated was implemented while limiting changes to the original geospatial census data. While in practice a new census geometry was created, the approach effectively minimizes changes to the source census database by:

(a) Minimizing modifications to the original geometry,
(b) Retaining the census hierarchy,
(c) Maintaining the regional distribution of population,
(d) Preserving the overall population counts.

The adopted approach is conservative: it decreases the population density in the zones corresponding to the original neighbouring populated units by expanding their area, while preserving their population. This approach is most suitable for the case when the population in the problematic unit was enumerated but re-assigned to other neighbouring units on reporting, thus retaining an accurate country population. However, if these populations were not enumerated to begin with, the country total will be inherently incorrect, which constitutes a more challenging case. Mitigation of the latter anomaly would require estimation of population actually residing in the unit using a bottom-up approach (see Wardrop et al. 2018 for options and discussion). Moreover, sources of population counts are notoriously old and outdated for some countries (e.g. Lebanon, DRC). It is possible that some units became populated or were settled after the census or estimates were conducted, as may be the case with refugee camps in Lebanon. It is also conceivable that some units, at the time the enumeration or estimates were produced, included settlements which were uninhabited (for a number of reasons, including war or other crises). Some areas are obviously densely populated (e.g. Sadr City in Baghdad, Iraq); their omission is most probably due to political and/or religious conflicts.

Problems of coverage and completeness in censuses are not new and their causes may be multiple and varied, including practical and political reasons (Carr-Hill 2013), and would certainly deserve more investigation that is beyond the scope of this paper. While there is some research reporting the general problem of undercounts (see Carr-Hill 2017 for summary), especially of particular groups, there is much less on the causes for complete and flagrant omission of population in large areas.

In any case, the process of projecting population counts for a recent target year and subsequent adjustment to UN estimates, as done by GPW, does not 'populate' a census unit originally devoid of resident population because the adjustment is accomplished at the national level.

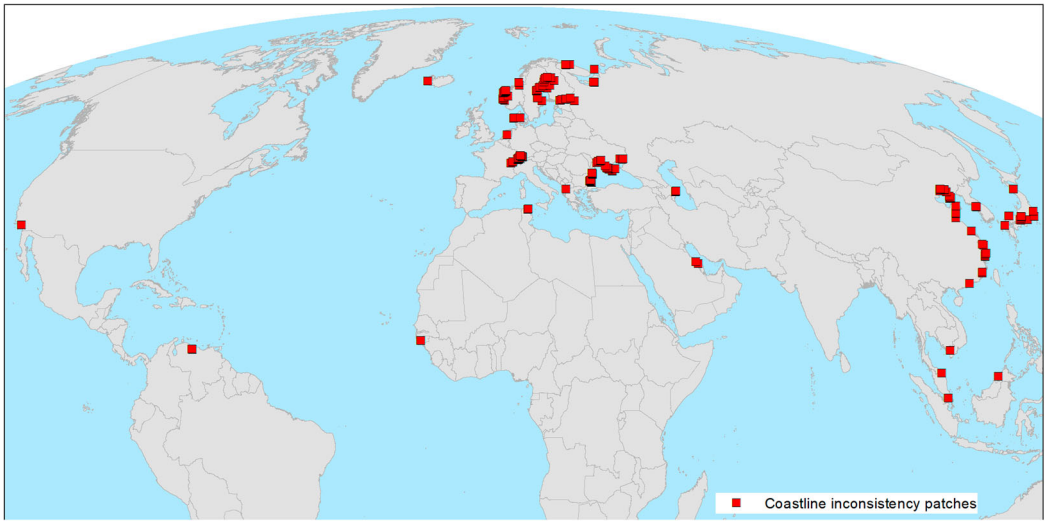

Figure 6. Visually confirmed coastline discrepancy patches larger than 1 km^2.

3.2. Harmonization of population and settlement data along coastlines

The semi-automated procedure to detect discrepancies between census data in GPWv4 and GHSL along coastlines (including inland water bodies) resulted in the detection of 287 patches having at least 1 km^2 built-up area based on GHSL BU 2014 (totalling nearly 760 km^2). The systematic visual inspection using VHR imagery revealed that 197 patches (containing 591 km^2 or 77% of the inspected built-up area) had correctly identified areas with significant built-up structures, located in 25 countries across the globe. In the remaining 90 patches, built-up structures could not be identified or their presence was not significant. Figure 6 shows the location of the 197 patches where built-up presence was confirmed, showing issues clustered in Europe and in E Asia.

As reported in Section 2.3, two procedures were followed for mitigating the confirmed discrepancies along coastlines, one manual and another automated. Initially the manual procedure was adopted, but the automated 'split and merge' procedure became more expeditious when the issue was found to be more systematic than initially assessed, as in case of Japan and the coast of the Ukraine.

Manual harmonization was carried out in the following countries: Albania, Austria, Azerbaijan, Bulgaria, Bahrain, Germany, Denmark, Dubai, Finland, Guinea-Bissau, Iceland, Republic of Korea, Malaysia, Netherlands, Norway, Romania, Russia, Singapore, Sweden, Tunisia, USA, Venezuela, and Viet Nam. Figure 7 illustrates the outcome of manual harmonization of coastline in Russia. The automated 'split and merge' approach was employed to modify and reconcile coastlines in Switzerland, France, Ukraine and Japan. Figure 8 shows the automated harmonization of coastlines applied along a stretch of Black Sea coast in Ukraine.

Although the detected anomalies may be significant for some applications such as the present one, and relevant from the perspective of data consistency, not all necessarily constitute errors in the original source data. Both in the case of coastline deficiencies and for units declared as unpopulated, the discrepancies may be due to anachronism between data sources. Coastal zones are typically attractive and dynamic areas from a settlement perspective, and geospatial data on administrative divisions used to support population estimates may not reflect the current land extents and settlement expansion obtained by land reclamation. Significant settlement expansion through land reclamation has been taking place in Asia (especially in China) and in the Middle East and it is important that time series of population grids capture and represent that process.

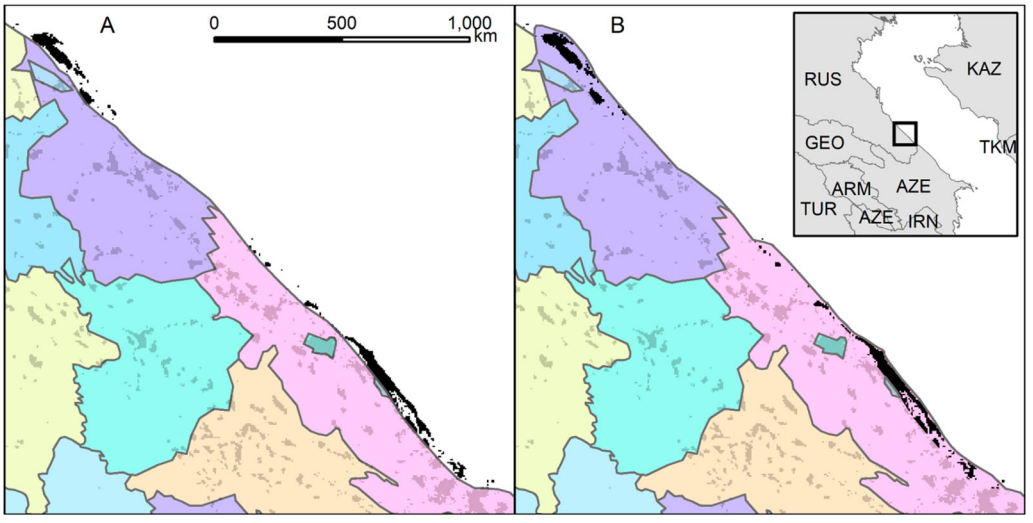

Figure 7. Example illustrating situation (A) before and (B) after manual harmonization of a stretch of Caspian Sea coastline in Russia. Solid black lines enclose original census units; black pixels denote mapped built-up areas (shaded: already within census boundaries; solid: outside original census areas).

Also, settlements can extend over water bodies such as floating settlements thus being located beyond the formal coastline. These can and are captured by GHSL, as no land mask is imposed *a priori* to limit detection of built-up. Instead of adopting the classical approach rooted in topography (e.g. Anderson et al. 1976) of following a hierarchical top-down procedure in abstracting the Earth surface into categories, GHSL adopts a people-centric approach that departs from the detection of human presence, and only then classifies phenomena into higher abstraction levels such as settlement classes (Pesaresi and Ehrlich 2009) (for explanation of GHSL family of products and analysis

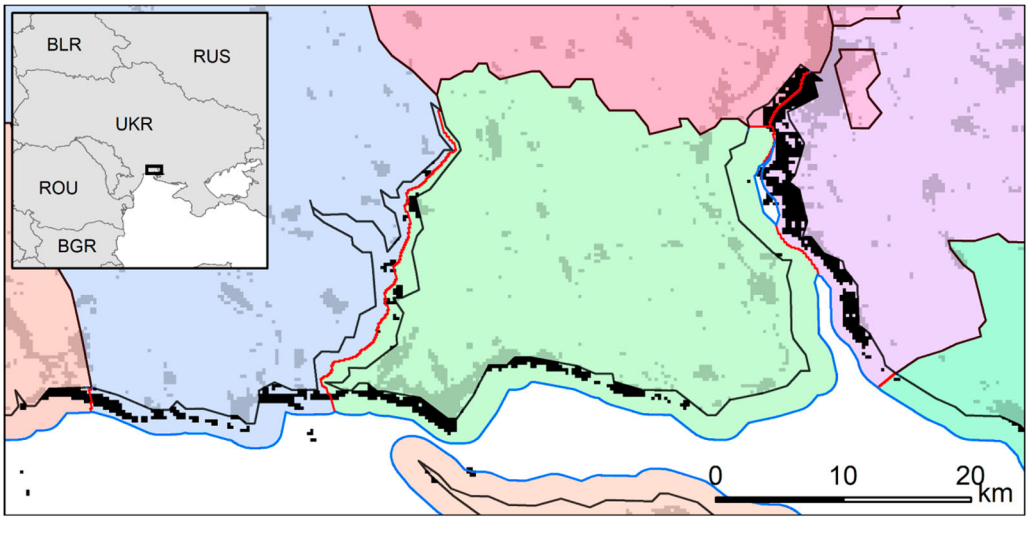

Figure 8. Example of automated approach to mitigate inconsistencies along the coastline of Ukraine. This example shows the generation of a 2-km buffer that was split using the 'split and merge' approach. Solid black lines enclose original census units; solid blue line represents the extended boundary into the sea through 2-km buffer; solid red lines represent the splitting of buffer into parts assigned to the nearest neighbouring unit; black pixels represent mapped built-up areas (shaded: already within census boundaries; solid: outside original census areas).

see Melchiorri et al. 2018). This approach may be a better match for the census paradigm, which aims to first identify where all people are so they can be surveyed and characterized. In this capacity, perhaps GHSL mapping of built-up could assist in planning census campaigns, especially in remote areas of countries having dynamic population trends that have not conducted a census in a long time.

Regarding the many discrepancies observed along coastlines, it is possible that these originate in differences in scale and level of detail among datasets, and at times can be introduced by the coarser detail of the database of Global Administrative Areas (GADM 2011) when boundaries are adjusted to a global framework in the GPW workflow.

Finally, it is important to note that the built-up structures detected and mapped could include abandoned or seasonal settlements, or otherwise non-residential facilities (e.g. ports, industrial areas). Ideally, a combination of census' spatial detail and attributes or ancillary data would allow further discrimination of land use according to application needs.

In practice, the two addressed anomalies can be regarded as an expression of the same underlying issue, i.e. complete omission of population counts (due to undercounting/under-reporting or poor census geography) against contradicting evidence. It has been shown that a similar approach can be used to automatically mitigate systematic issues of both types.

4. Conclusions

Global population distribution grids are increasingly required and relied upon for many applications. These datasets are mostly produced by gridding or disaggregating counts from official population census and estimates. Despite the continuous advancements in population gridding methods, these usually do not address ex-ante issues affecting input population statistics. Availability, quality and detail of national data reporting on population is highly heterogeneous, and the best available global geospatial census data is not immune to shortcomings. These propagate to derived population distribution grids and may negatively impact their applications, in an age of new Development Agendas whose effective monitoring relies heavily on accurate and reliable geospatial population data and calls for universal inclusiveness of people. Exploring the improved capacities of recent global settlement data derived from remote sensing, we have sought to improve the mapping of population distribution by addressing some of these shortcomings.

New automated procedures to detect and mitigate major discrepancies and anomalies occurring in geospatial census data were developed, tested and implemented, while minimizing changes to the original data. Global and consistent remote sensing-derived data reporting on built-up presence was used to revise census units deemed as 'unpopulated' and to harmonize population distribution along coastlines. The two procedures employed for the detection of deficiencies in global geospatial census data obtained high rates of true positives, after validation and confirmation. The results also show that the targeted anomalies were significantly mitigated and that the baseline census database has improved, potentially benefitting other uses of the same statistical base. These outcomes are encouraging for further uses of free and open geoinformation derived from remote sensing. However, it must be recognized that assessing and monitoring progress of the new Development Agendas involves measuring many other population characteristics beyond population counts, for which proper census and surveys will remain invaluable sources, and that even accurate total population counts will ultimately continue to rely on high-quality enumerations and estimates.

This work illustrates the value and possible contribution of detailed, updated, and independent remote sensing data to complement and improve conventional sources of fundamental population statistics. While we acknowledge that the proposed procedures reduce the independency in the production of the involved variables (population and built-up distribution), which may be desirable for subsequent combination and production of population grids, these processes contribute in other ways by closing data gaps, improving data quality, and will ultimately benefit global mapping of population and its downstream applications.

Future developments will focus on improving the automated detection and mitigation of deficiencies present in geospatial census data, including addressing additional anomalies (e.g. over-reporting of population and extremely high population densities). The multi-temporal nature of the data involved further increases the complexity of issues and adds challenges that should also be tackled, such as the estimation of population counts across epochs. For mitigation of areas incorrectly labelled as unpopulated, alternatives to the 'split & merge' approach should be explored, namely contextual bottom-up approaches that take into consideration the local patterns of settlement.

Acknowledgements

The term country also refers, as appropriate, to territories or areas, and does not imply recognition of borders or legal status by the European Commission.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work has been carried out in the frame of the institutional work programme of the Joint Research Centre (JRC, European Commission) and supported by the administrative arrangement no. 33994 between the JRC and the Directorate General for Regional and Urban Policies (DG REGIO, European Commission). Work from partners on GPW4.10 and improvements described herein are funded by NASA under contract NNG08HZ11C for the continued operation of the Socioeconomic Data and Applications Center (SEDAC) at Center for International Earth Science Information Network (CIESIN) at Columbia University.

ORCID

Sergio Freire http://orcid.org/0000-0003-2282-701X
Marcello Schiavina http://orcid.org/0000-0003-3399-3400
Aneta J. Florczyk http://orcid.org/0000-0001-8912-1500
Martino Pesaresi http://orcid.org/0000-0003-0620-439X
Christina Corbane http://orcid.org/0000-0002-2670-1302

References

Anderson, J. R., E. E. Hardy, J. T. Roach, and R. E. Witmer. 1976. *A Land Use and Land Cover Classification System for Use With Remote Sensing Data*. US Geological Survey, Professional Paper No. 964.

Balk, D. L., U. Deichmann, G. Yetman, F. Pozzi, S. I. Hay, and A. Nelson. 2006. "Determining Global Population Distribution: Methods, Applications and Data." *Advances in Parasitology* 62: 119–156.

Carr-Hill, R. 2013. "Missing Millions and Measuring Development Progress." *World Development* 46: 30–44.

Carr-Hill, R. 2017. "Improving Population and Poverty Estimates with Citizen Surveys: Evidence from East Africa." *World Development* 93: 249–259.

CIESIN (Center for International Earth Science Information Network). 2016. *Gridded Population of the World, Version 4 (GPWv4)*. Palisades, NY: Columbia University. http://www.ciesin.columbia.edu/data/gpw-v4.

CIESIN (Center for International Earth Science Information Network). 2017a. *Documentation for the Gridded Population of the World, Version 4 (GPWv4), Revision 10 Data Sets*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), Columbia University. doi:10.7927/H4B56GPT.

CIESIN (Center for International Earth Science Information Network). 2017b. *Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 10*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), Columbia University. doi:10.7927/H4PG1PPM.

Congalton, R. G., and K. Green. 2009. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. 2nd ed. Boca Raton, FL: CRC/Lewis Press.

Corbane, C., Martino Pesaresi, Panagiotis Politis, Vasileios Syrris, Aneta J. Florczyk, Pierre Soille, Luca Maffenini, et al. 2017. "Big Earth Data Analytics on Sentinel-1 and Landsat Imagery in Support to Global Human Settlements Mapping." *Big Earth Data*, 1–27. doi:10.1080/20964471.2017.1397899.

Deichmann, U., D. Balk, and G. Yetman. 2001. *Transforming Population Data for Interdisciplinary Usages: From Census to Grid*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), CIESIN, Columbia University.

Doxsey-Whitfield, E., K. MacManus, S. B. Adamo, L. Pistolesi, J. Squires, O. Borkovska, and S. R. Baptista. 2015. "Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4." *Papers in Applied Geography* 1 (3): 226–234. doi:10.1080/23754931.2015.1014272.

EEA (European Environment Agency). 2006. *The Changing Faces of Europe's Coastal Areas*. Report No 6/2006, 107 p. Copenhagen: European Environment Agency.

EEA (European Environment Agency). 2016. *Corine Land Cover (CLC) 2012, Version 18.5.1*. https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012?tab=download.

Freire, S., T. Kemper, M. Pesaresi, A. J. Florczyk, and V. Syrris. 2015. "Combining GHSL and GPW to Improve Global Population Mapping." 2015 IEEE International Geoscience & Remote Sensing Symposium (IGARSS), 2541–2543, Milan. doi:10.1109/IGARSS.2015.7326329.

Freire, S., K. MacManus, M. Pesaresi, E. Doxsey-Whitfield, and J. Mills. 2016. "Development of New Open and Free Multi-temporal Global Population Grids at 250 m Resolution." Proceedings of the 19th AGILE Conference on Geographic Information Science, June 14–17, Helsinki.

Freire, S., T. Santos, and J. A. Tenedório. 2009. "Recent Urbanization and Land Use/Land Cover Change in Portugal – the Influence of Coastline and Coastal Urban Centers." *Journal of Coastal Research SI* 56: 1499–1503.

GADM (Global Administrative Areas). 2011. *Version 2.0*. http://gadm.org.

Gaughan, A. E., F. R. Stevens, C. Linard, N. N. Patel, and A. J. Tatem. 2014. "Exploring Nationally and Regionally Defined Models for Large Area Population Mapping." *International Journal of Digital Earth*. doi:10.1080/17538947.2014.965761.

Hay, S. I., A. M. Noor, A. Nelson, and A. J. Tatem. 2005. "The Accuracy of Human Population Maps for Public Health Application." *Tropical Medicine & International Health* 10 (10): 1073–1086. doi:10.1111/j.1365-3156.2005.01487.x.

Leyk, S., J. Uhl, D. Balk, and B. Jones. 2018. "Assessing the Accuracy of Multi-temporal Built-up Land Layers Across Rural-urban Trajectories in the United States." *Remote Sensing of Environment* 204: 898–917.

Linard, C., C. W. Kabaria, M. Gilbert, A. J. Tatem, A. E. Gaughan, F. R. Stevens, A. Sorichetta, A. M. Noor, and R. W. Snow. 2017. "Modelling Changing Population Distributions: An Example of the Kenyan Coast, 1979–2009." *International Journal of Digital Earth* 10 (10): 1017–1029.

Linard, C., and A. Tatem. 2012. "Large-scale Spatial Population Databases in Infectious Disease Research." *International Journal of Health Geographics* 11: 7. doi:10.1186/1476-072X-11-7.

McGranahan, G., D. Balk, and B. Anderson. 2007. "The Rising Tide: Assessing the Risks of Climate Change and Human Settlements in Low Elevation Coastal Zones." *Environment and Urbanization* 19: 17–37.

Melchiorri, M., A. J. Florczyk, S. Freire, M. Schiavina, M. Pesaresi, and T. Kemper. 2018. "Unveiling 25 Years of Planetary Urbanization with Remote Sensing: Perspectives from the Global Human Settlement Layer." *Remote Sensing* 10 (5): 768.

Mondal, P., and A. J. Tatem. 2012. "Uncertainties in Measuring Populations Potentially Impacted by Sea Level Rise and Coastal Flooding." *PLoS ONE* 7 (10): e48191. doi:10.1371/journal.pone.0048191.

Nieves, J. J., F. R. Stevens, A. E. Gaughan, C. Linard, A. Sorichetta, G. Hornby, N. N. Patel, and A. J. Tatem. 2017. "Examining the Correlates and Drivers of Human Population Distributions Across Low- and Middle-Income Countries." *Journal of the Royal Society Interface* 14: 20170401. doi:10.1098/rsif.2017.0401.

Pesaresi, M., and D. Ehrlich. 2009. "A Methodology to Quantify Built-up Structures from Optical VHR Imagery." In *Global Mapping of Human Settlement Experiences, Datasets, and Prospects, ch. 3*, edited by P. Gamba and M. Herold, 27–58. Boca Raton, FL: CRC Press.

Pesaresi, M., D. Ehrlich, S. Ferri, A. Florczyk, S. Freire, S. Halkia, M. J. Andreea, T. Kemper, P. Soille, and V. Syrris. 2016. *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014. EUR – Scientific and Technical Research Reports*. Publications Office of the European Union. JRC97705. http://publications.jrc.ec.europa.eu/repository/handle/111111111/40182.

Pesaresi, M., D. Ehrlich, A. J. Florczyk, S. Freire, A. Julea, T. Kemper, P. Soille, and V. Syrris. 2015a. *GHS Built-up Grid, Derived from Landsat, Multitemporal (1975, 1990, 2000, 2014)*. European Commission, Joint Research Centre (JRC) [Dataset] PID. http://data.europa.eu/89h/jrc-ghsl-ghs_built_ldsmt_globe_r2015b.

Pesaresi, M., D. Ehrlich, A. J. Florczyk, S. Freire, A. Julea, T. Kemper, P. Soille, and V. Syrris. 2015b. *GHS Built-up Datamask Grid Derived from Landsat, Multitemporal (1975, 1990, 2000, 2014)*. European Commission, Joint Research Centre (JRC) [Dataset] PID. http://data.europa.eu/89h/jrc-ghsl-ghs_built_ldsmtdm_globe_r2015b.

Pesaresi, M., G. Huadong, X. Blaes, D. Ehrlich, S. Ferri, L. Gueguen, M. Halkia, et al. 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.

Shupeng, C., and J. van Genderen. 2008. "Digital Earth in Support of Global Change Research." *International Journal of Digital Earth* 1 (1): 43–65. doi:10.1080/17538940701782510.

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem. 2015. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely Sensed and Ancillary Data." *PLoS ONE* 10 (2): e0107042. doi:10.1371/journal.pone.0107042.

Tatem, A. J., A. M. Noor, C. von Hagen, A. Di Gregorio, and S. I. Hay. 2007. "High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa." *PLoS ONE* 2 (12): e1298.

Tobler, W., U. Deichmann, J. Gottsegen, and K. Maloy. 1997. "World Population in a Grid of Spherical Quadrilaterals." *International Journal of Population Geography* 3: 203–225.

UNDESA (United Nations Department of Economic and Social Affairs). 2015. *Department of Economic and Social Affairs, Population Division. World Population Prospects: The 2015 Revision*. New York: United Nations. http://esa.un.org/unpd/wpp/DVD/.

UNDESA (United Nations Department of Economic and Social Affairs). 2016. *Sustainable Development Goals*. Department of Economic and Social Affairs. Accessed March 2018. https://sustainabledevelopment.un.org/sdgs.

UNDP (United Nations Development Programme). 2018. *Strengthening the National Statistical System*. Accessed June 2018. http://www.md.undp.org/content/moldova/en/home/projects/strengthening-the-national-statistical-system-.html.

UN ECOSOC. 2016. *Report of the Inter-agency and Expert Group on Sustainable Development Goal Indicators*. UN Economic and Social Council Statistical Commission 47th Session, E/CN.3/2016/2/Rev.1.

UNISDR (United Nations International Strategy for Disaster Reduction). 2015. *Sendai Framework for Disaster Risk Reduction 2015–2030*. Accessed February 2017. http://www.wcdrr.org/uploads/Sendai_Framework_for_Disaster_Risk_Reduction_2015-2030.pdf.

Wardrop, N. A., W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. J. Tatem. 2018. "Spatially Disaggregated Population Estimates in the Absence of National Population and Housing Census Data." Proceedings of the National Academy of Sciences, March. doi:10.1073/pnas.1715305115.

Wolff, H., H. Chong, and M. Auffhammer. 2011. "Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index." *Economic Journal* 121 (553): 843–870.

World Bank. 2018. *Improving Statistical Accuracy in Mongolia for Evidence-based Policy Making*. Accessed June 2018. http://www.worldbank.org/en/results/2017/05/12/improving-statistical-accuracy-in-mongolia-for-evidence-based-policy-making.