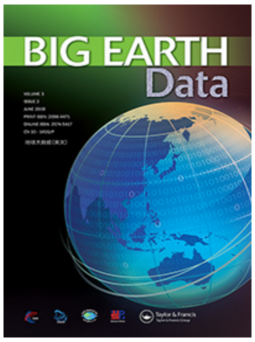

Taylor & Francis
Taylor & Francis Group

Big Earth Data

ISSN: 2096-4471 (Print) 2574-5417 (Online) Journal homepage: https://www.tandfonline.com/loi/tbed20

Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets

Christopher T. Lloyd, Heather Chamberlain, David Kerr, Greg Yetman, Linda Pistolesi, Forrest R. Stevens, Andrea E. Gaughan, Jeremiah J. Nieves, Graeme Hornby, Kytt MacManus, Parmanand Sinha, Maksym Bondarenko, Alessandro Sorichetta & Andrew J. Tatem

To cite this article: Christopher T. Lloyd, Heather Chamberlain, David Kerr, Greg Yetman, Linda Pistolesi, Forrest R. Stevens, Andrea E. Gaughan, Jeremiah J. Nieves, Graeme Hornby, Kytt MacManus, Parmanand Sinha, Maksym Bondarenko, Alessandro Sorichetta & Andrew J. Tatem (2019) Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets, Big Earth Data, 3:2, 108-139, DOI: 10.1080/20964471.2019.1625151

To link to this article: https://doi.org/10.1080/20964471.2019.1625151

Published online: 18 Jun 2019.

BIG EARTH DATA
2019, VOL. 3, NO. 2, 108–139
https://doi.org/10.1080/20964471.2019.1625151

DATA ARTICLE

OPEN ACCESS

Check for updates

Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets

Christopher T. Lloyd[a], Heather Chamberlain[a,b], David Kerr[a], Greg Yetman[c], Linda Pistolesi[c], Forrest R. Stevens[d], Andrea E. Gaughan[d], Jeremiah J. Nieves[a], Graeme Hornby[a,e], Kytt MacManus[c], Parmanand Sinha[d], Maksym Bondarenko[a], Alessandro Sorichetta[a] and Andrew J. Tatem[a,b]

[a]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, UK; [b]Flowminder Foundation, Stockholm, Sweden; [c]Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, USA; [d]Department of Geography and Geosciences, University of Louisville, Louisville, KY, USA; [e]GeoData, University of Southampton, Southampton, UK

ABSTRACT

Multi-temporal, globally consistent, high-resolution human population datasets provide consistent and comparable population distributions in support of mapping sub-national heterogeneities in health, wealth, and resource access, and monitoring change in these over time. The production of more reliable and spatially detailed population datasets is increasingly necessary due to the importance of improving metrics at sub-national and multi-temporal scales. This is in support of measurement and monitoring of UN Sustainable Development Goals and related agendas. In response to these agendas, a method has been developed to assemble and harmonise a unique, open access, archive of geospatial datasets. Datasets are provided as global, annual time series, where pertinent at the timescale of population analyses and where data is available, for use in the construction of population distribution layers. The archive includes sub-national census-based population estimates, matched to a geospatial layer denoting administrative unit boundaries, and a number of co-registered gridded geospatial factors that correlate strongly with population presence and density. Here, we describe these harmonised datasets and their limitations, along with the production workflow. Further, we demonstrate applications of the archive by producing multi-temporal gridded population outputs for Africa and using these to derive health and development metrics. The geospatial archive is available at https://doi.org/10.5258/SOTON/WP00650.

ARTICLE HISTORY
Received 15 March 2019
Accepted 25 May 2019

KEYWORDS
Human population; sub-national; global; spatial dataset; multi-temporal

1. Introduction

Human population mapping is fundamental in support of a broad range of applications by governments, non-governmental organisations, and private businesses. Detailed and up to date spatial datasets that accurately describe population distribution can support the planning and delivery of services (Langford, Higgs, Radcliffe, & White, 2008), election

CONTACT Christopher T. Lloyd C.T.Lloyd@soton.ac.uk

Supplemental data for this article can be assessed here.

mapping (Amos, McDonald, and Watkins 2017), estimation of populations at risk of infectious disease or hazards (Hay, Guerra, Tatem, Atkinson, & Snow, 2005; Linard, Alegana, Noor, Snow, & Tatem, 2010; Snow, Guerra, Noor, Myint, & Hay, 2005), and disaster relief operations (Bhaduri, Bright, Coleman, & Dobson, 2002; Nadim, Kjekstad, Peduzzi, Herold, & Jaedicke, 2006; Taramelli, Melelli, Pasqui, & Sorichetta, 2010).

Census data are typically made openly available only aggregated by large administrative areas as spatial (areal) units. Aggregation results in loss of spatial detail and is performed to protect confidentiality. It is possible to directly produce human population distribution maps from such data by linking counts to the appropriate boundaries. However, the use of large spatial areal units presents analytical challenges for population studies. Administrative unit boundaries are often unrelated to the demographic variables of interest, and in the physical world populations are not uniformly distributed within them (Sorichetta et al., 2015; Stevens, Gaughan, Linard, & Tatem, 2015). Such challenges make it difficult to compare the distribution of human populations over time and space in a consistent and methodological way.

In order to better characterise the distribution of populations and overcome the limitations of such aggregate data, much research has focused on creating alternative representations of the population as a continuous surface (Mennis, 2003). Such approaches use a variety of techniques to assign estimated population counts to grid cells, a topic discussed in more detail in Wardrop et al. (2018). There are two ways to approach modelling gridded population data, either a “top-down” or a “bottom-up” approach. The top-down modelling approaches (Azar, Engstrom, Graesser, & Comenetz, 2013; Stevens et al., 2015) are the most commonly used due to the availability of census and geospatialgeospatial covariate data.

A top-down approach relies on high quality and up to date census population counts or official estimates that are combined, or “aggregated”, into administrative units and linked to their digital boundaries. Subsequently, counts are redistributed (or “disaggregated”) into grid cells (i.e. pixels). A variety of techniques may be utilised to disaggregate, ranging from the simple through to the more statistically complex. The “areal-weighting” technique is a simple way to address the challenge of characterising the spatial variation of population within administrative units, taking (non-spatial) tabular counts of population (listed by administrative unit) and (spatial) administrative boundary data, and disaggregating population from census units into grid cells through the assumption that the population of a grid cell is an exclusive function of the land area within that pixel (Doxsey-Whitfield et al., 2015). The Gridded Population of the World (GPW) v4 dataset (CIESIN, 2016a) uses the areal-weighting technique (CIESIN, 2016b), detailing population count and density at 30 arc-second resolution (approximately 1 km resolution at the equator). The advantage of this simple disaggregation technique is that it does not incorporate more complex considerations. Output grids can, therefore, be used with other geographic information without endogeneity concerns. The major disadvantage is the inability to characterise spatial variations within the input geometry, especially in cases where the input administrative units are much larger than the spatial resolution of the output grid. Dasymetric mapping is a more complex technique that uses geospatial covariates (e.g. land cover) via a spatial weighting grid to more accurately distribute the population data assigned to selected administrative units. Dasymetric mapping has been shown to be the most accurate top-down approach to disaggregating census counts into gridded maps (Sorichetta et al., 2015; Stevens et al., 2015; Wardrop et al., 2018).

In comparison, "bottom-up" approaches (Checchi, Stewart, Palmer, & Grundy, 2013; Hillson et al., 2014, 2015; Tomás, Fonseca, Almeida, Leonardi, & Pereira, 2015; Wardrop et al., 2018; Weber et al., 2018) are a more recent development that take complete counts of population within small, defined areas (sometimes called "micro-census" surveys) and produce a gridded estimate of overall population through the prediction of population in (much larger) un-surveyed areas via the use of geospatial covariates and statistical modelling (Wardrop et al., 2018). Bottom-up approaches are difficult to implement at a global scale due to the resources required to collect data, and the storage and computational overhead. Bottom-up approaches are best applied to countries where census data are of poor quality, outdated, or non-existent.

Remotely sensed and other geospatial ancillary data can be used in population modelling in order to improve detail (Balk et al., 2006; Bhaduri et al., 2002; Tatem, Noor, von Hagen, Di Gregorio, & Hay, 2007). For example, the Global Rural–Urban Mapping Project (GRUMP) version 1 (Balk, Pozzi, Yetman, Deichmann, & Nelson, 2005; CIESIN, IFPRI, World Bank, & CIAT, 2011) build on GPW v3 (CIESIN and CIAT, 2005; Balk and Yetman, 2004), differentiating urban and rural areas by formulation of a mask via the combination of census data with remote-sensed nightlights data (Balk, Yetman, & de Sherbinin, 2010; CIESIN, 2005). Land cover data may be used to redistribute aggregated census counts in order to improve the accuracy of national scale gridded population data (Linard, Gilbert, & Tatem, 2011). Where settlement extents are used, e.g. GHS-POP (Freire, MacManus, Pesaresi, Doxsey-Whitfield, & Mills, 2016), population distribution datasets are generally more accurate than when simple areal weighting is used, as shown in previous studies (Gaughan, Stevens, Linard, Jia, & Tatem, 2013; Linard et al., 2010; Linard, Gilbert, Snow, Noor, & Tatem, 2012; Linard et al., 2011; Mennis & Hultgren, 2006; Tatem et al., 2007).

A wide range of factors are known to correlate with how humans distribute themselves on the landscape (Nieves et al., 2017). A larger number of covariates may be utilised in modelling in order to more effectively disaggregate census population counts within administrative units, and to better statistically describe population distribution (Lloyd, Sorichetta, & Tatem, 2017) – an approach used to produce the Landscan population datasets (ORNL 2010; Bhaduri, Bright, Coleman, & Urban, 2007; Dobson, Bright, Coleman, Durfee, & Worley, 2000). The Random Forest-based (RF) dasymetric model, a non-parametric ensemble approach (Breiman, 2001), is a further example used to produce WorldPop population datasets (Gaughan et al., 2016; Sorichetta et al., 2015; Stevens et al., 2015). The RF method, discussed in more detail later in this paper, incorporates census data and a wide range of ancillary datasets in a flexible estimation technique. Output suggests marked improvements in mapping accuracies over other "top-down" population mapping approaches, such as areal-weighting (Sorichetta et al., 2015; Stevens et al., 2015).

Due to lack of resources (financial and human) to carry out detailed censuses, fine spatial detail population count data are lacking for the present day and past decades in many countries (e.g. Afghanistan, Democratic Republic of Congo, Lebanon, Uzbekistan, in particular), thereby limiting applications linked to specific time periods or those measuring changes. Sub-national scale analyses related to population are beginning to utilise multi-temporal geospatial layers (Bennett & Smith, 2017a, 2017b). Multi-temporal geospatial layers are useful in providing globally consistent gridded population distribution datasets that can be used to support agendas aligned with Sustainable Development Goals (SDGs) (UN General Assembly, 2015). Aligned agendas are those such as the Institute for Health Metrics and Evaluation (IHME) Global Burden of Disease (GBD) studies (GBD, 2016; IHME, 2013, 2016; SDG Collaborators, 2017) or the Malaria Atlas Project (MAP, 2017; Bhatt et al., 2015; Cibulskis et al., 2016). The present situation of a lack of multi-temporal global modelled population data limits abilities to provide context to global multi-temporal disease prevalence mapping efforts and convert them to burden estimates.

In order to better support global high-resolution population mapping in the future, a set of methods have been developed here to assemble and harmonise a unique (in spatial and temporal scope), open access, archive of geospatial datasets. Datasets are provided as annual time series, where pertinent at the timescale of population analyses and where data are available. These can be used to construct consistent and comparable annual high-resolution global population distribution layers for the 2000–2020 period. The archive includes sub-national census-based population estimates, matched to gridded administrative boundaries, and a number of co-registered gridded geospatial factors that correlate strongly with population presence and density. The datasets described in this paper are mostly an assemblage of pre-existing datasets, created to provide researchers with easier access via considerable effort towards harmonisation.

A collection of harmonised geospatial layers has previously been developed for use in population studies (Lloyd et al., 2017), as an internal effort undertaken with the WorldPop programme. The collection described here demonstrates significant differences and advancements over that earlier work and is a significant cross-organisational collaboration between WorldPop and the Center for International Earth Science Information Network (CIESIN). The pre-existing datasets used to create the geospatial layers described in Lloyd et al. (2017) are almost entirely different to those discussed in this paper and are standardised solely to less accurate Global Administrative Areas version 2 (GADMv2) (GADM, 2015) country boundaries. In contrast, the newly assembled and harmonised layers, discussed here, mark a significant improvement by the inclusion of subnational census-based population estimates and by the utilisation of associated administrative boundaries. These are the same input data as previously used in the production of the GPWv4 gridded datasets (CIESIN, 2016a, 2016b; Doxsey-Whitfield et al., 2015). Further, the geospatial layers described in Lloyd et al. (2017) are mostly time invariant, therefore, not effectively facilitating the monitoring of change in population over time, whereas the layers described in this paper are provided as time series where relevant/available.

Here, we describe the production methods for the geospatial layers. A predominantly open source production environment is utilised, and a semi-automated workflow. We then present example applications of the geospatial layers, as harmonised gridded inputs to inform an RF model to provide spatially consistent gridded population outputs (Stevens et al., 2015). In particular, we use the workflow described by Gaughan et al. (2016) to compare population outputs for Africa at several time periods and demonstrate the potential usefulness of these high spatial resolution data in health and development metric applications.

2. Methods

To support the production of global maps of population distributions and demographics for the period 2000 to 2020, population counts (interpolated and forecast at sub-national level) are linked to spatially and temporally harmonised national and sub-national spatial data describing administrative unit extents, derived from GPWv4 (CIESIN (Center for International Earth Science Information Network, Columbia University), 2016a). A range of open access geospatial layers are collected and similarly harmonised, representing factors that correlate strongly with human population density (Nieves et al., 2017).

The time-invariant geospatialgeospatial layers produced as potential input grids for modelling population distribution are: Viewfinder Panoramas (SRTM based) topography (units in metres) for year 2000 (de Ferranti, 2017a); a slope layer derived from the topography (in degrees); pixel area (m^2), and coastline (binary, as land/open water pixels); OpenStreetMap (OSM) highway (major highway routes), highway intersection, and waterway locations (OSMF and Contributors, 2016); and WorldClim average global temperature (°C) and precipitation (mm) for 1970–2000 (Fick & Hijmans, 2017). The multi-temporal geospatialgeospatial layers (i.e. annual time series) produced are: DMSP-OLS version four night-time lights (2000–2011) composites (US NOAA, 2015; Zhang, Pandey, & Seto, 2016); VIIRS version 1 night-time lights (2012–2016) composites (US NOAA, 2017); ESA CCI annual global land cover for 2000–2015 (ESA CCI, 2017a); UNEP/IUCN World Database of Protected Areas for 2000–2017 (UNEP-WCMC and IUCN, 2017); and built settlement grids for 2000, 2012, and 2014, which combine the JRC Global Human Settlement Layer (Pesaresi et al., 2015) with the ESA CCI built settlement land-cover class and the DLR Global Urban Footprint (DLR EOC, 2016) dataset and which were extrapolated and interpolated into an annual time series as described in Nieves et al. (2018). The workflow for standardising and harmonising geospatialgeospatial layers is a significant development and expansion of methods discussed in Lloyd et al. (2017), and Lloyd (2017). Workflow is visualised diagrammatically in Figure 1. Source datasets are detailed in Table 1.

Categorical covariates are further each converted to binary grids (representing the feature of interest) as additional potential input grids for modelling, and from which derivative covariates can be produced if desired. Derivatives (such as datasets that indicate the distance to a given feature) increase covariate variability and therefore better captures the relationship with population density (e.g. urban core verses outskirts).

2.1. Source datasets

National (L0) and sub-national (L1) administrative unit boundary vector source material (CIESIN (Center for International Earth Science Information Network, Columbia University), 2016a) are rasterised by CIESIN, forming the base grids (i.e. mastergrids) of the archive. Regarding the sub-national L1 administrative units, it is important to highlight that even though hereafter these are simply referred to as L1, they actually represent the highest administrative unit level obtained for each country. The L1 data define administrative units per country for the entire globe, and are combined by CIESIN with ESA CCI – LC v4.0 inland waterbody raster data (ESA CCI, 2017b) to form one dataset. The L0 country identification (ID) global layer uses a three digit numerical ISO 3166 country code standard (ISO, 2017) that applies country/territory names/codes as designated by the United Nations, some of which are disputed. For disputed territories the intent is not to represent international boundaries, but rather to represent the source of the census administrative unit boundary and count estimate data. The L0

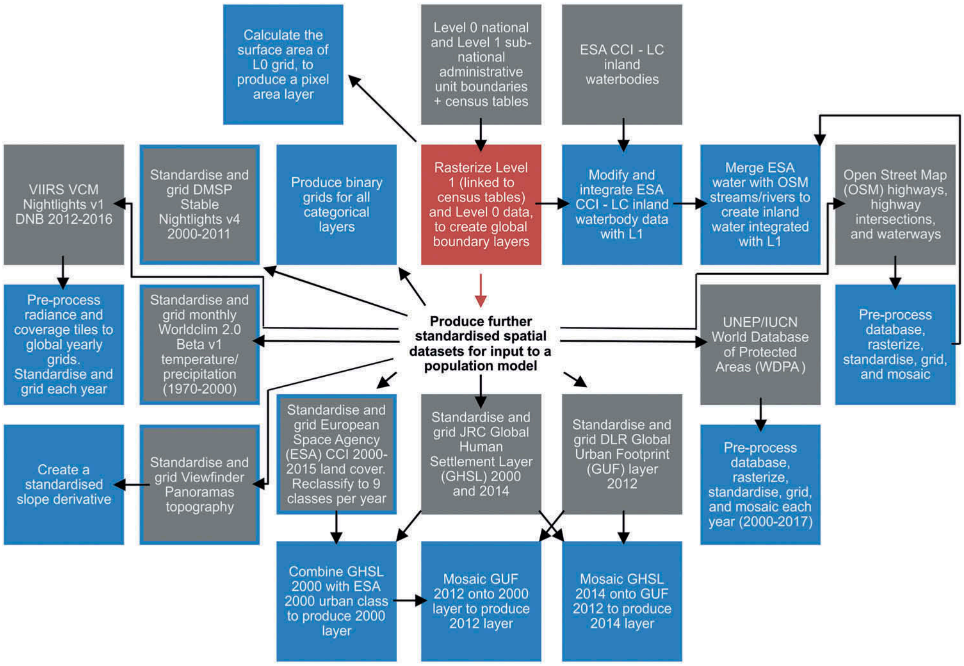
Calculate the surface area of L0 grid, to produce a pixel area layer
Level 0 national and Level 1 sub-national administrative unit boundaries + census tables
ESA CCI - LC inland waterbodies
VIIRS VCM Nightlights v1 DNB 2012-2016
Standardise and grid DMSP Stable Nightlights v4 2000-2011
Produce binary grids for all categorical layers
Rasterize Level 1 (linked to census tables) and Level 0 data, to create global boundary layers
Modify and integrate ESA CCI - LC inland waterbody data with L1
Merge ESA water with OSM streams/rivers to create inland water integrated with L1
Open Street Map (OSM) highways, highway intersections, and waterways
Pre-process radiance and coverage tiles to global yearly grids. Standardise and grid each year
Standardise and grid monthly Worldclim 2.0 Beta v1 temperature/ precipitation (1970-2000)
Produce further standardised spatial datasets for input to a population model
UNEP/IUCN World Database of Protected Areas (WDPA)
Pre-process database, rasterize, standardise, grid, and mosaic
Create a standardised slope derivative
Standardise and grid Viewfinder Panoramas topography
Standardise and grid European Space Agency (ESA) CCI 2000-2015 land cover. Reclassify to 9 classes per year
Standardise and grid JRC Global Human Settlement Layer (GHSL) 2000 and 2014
Standardise and grid DLR Global Urban Footprint (GUF) layer 2012
Pre-process database, rasterize, standardise, grid, and mosaic each year (2000-2017)
Combine GHSL 2000 with ESA 2000 urban class to produce 2000 layer
Mosaic GUF 2012 onto 2000 layer to produce 2012 layer
Mosaic GHSL 2014 onto GUF 2012 to produce 2014 layer

Figure 1. Flowchart of the workflow to produce standardised spatial datasets for potential input to a population model.

Production of base datasets is depicted in red, and source data in grey. Processes which directly lead to the production of further covariate output, for potential input to a population model, are represented in blue (or blue border as appropriate).

layer complements the L1 data via the use of common national boundaries. Other geospatial layers are spatio-temporally harmonised (standardised) to grid definitions and coastlines derived from administrative unit extents. All source datasets use a geographical coordinate system (GCS) with WGS 1984 datum (EPSG:4326), unless otherwise detailed.

L0 and L1 administrative unit boundary source material are gridded by CIESIN to 3 arc-second (0.00083333333 decimal degree) spatial resolution. CIESIN then resample and integrate ESA CCI – LC v4.0 4.5 arc-second (~150 m at the Equator) spatial resolution inland waterbody data with L1. Administrative unit boundary data take priority where there is complete overlap of units with waterbody data. Underlap between census and water is designated as water. The chosen cell size represents a middling spatial resolution to which source datasets (having various resolutions) can be rationalised, and offers reasonable storage and computational overhead at a global scale. Such overhead would increase significantly were a finer spatial resolution chosen instead.

Sub-national population count tables are interpolated and forecast annually by CIESIN from the year 2000 to 2020, using two census dates for most countries (circa 2000, and circa 2010) taken from GPWv4. The latest population census data have been collected during the 2010 round, between 2005 and 2014 (Doxsey-Whitfield et al., 2015). Countries conduct their censuses at different times. Hence, in order to interpolate and forecast, annual growth rates are used to adjust population counts to allow for global

Table 1. Source datasets used to produce geospatial raster layers for potential input to a population model.

Name	Acquisition year	Temporal variation	Source	Version, publication year	Data type	Spatial resolution	Format/ pixel type & depth	Spatial reference	Spatial coverage
National L0 and sub-national census L1 administrative boundaries	2005–2014	Time Invariant	Center for International Earth Science Information Network (CIESIN), Columbia University	GPW v4, 2016	Global population count and administrative boundaries, table and vector	Comparable to 3″ (~90 m)	ESRI polygon shapefiles	GCS WGS 1984	Global
Water bodies	2000–2012	Time Invariant	ESA (European Space Agency) CCI (Climate Change Initiative) – LC (Land Cover project)	v4.0, 2017	Inland water bodies, categorical raster	4.5″ (~150 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
Viewfinder Panoramas Topography	~2000	Time Invariant	de Ferranti, J.	28/11/17	Elevation, continuous raster	Typically 3″ (~90 m)	HGT tiles/ int16	GCS WGS 1984	Global
Open Street Map (OSM)	2016	Time Invariant	OpenStreetMap Foundation (OSMF) & Contributors	15/01/16	General mapping, categorical vector	Comparable to 1″ (~30 m)	PBF database	GCS WGS 1984	Global
WorldClim 2.0	1970–2000	Time Invariant	Fick, S.E. and Hijmans, R.J.	01/06/16	Monthly temperature and precipitation, continuous rasters	30″ (~900 m)	Geo-tiff/ flt32, int16	GCS WGS 1984	Global
DMSP-OLS Stable Nightlights	2000–2011	Time Series	US NOAA National Geophysical Data Center; Zhang et al.	v4, 2015; inter-calibrated, 2016	Annual night lights intensity, continuous rasters	30″ (~900 m)	Geo-tiff/ uint8	GCS WGS 1984	Between latitudes 75° North and 65° South
ViiRS Cloud Mask (VCM) Nightlights Day/ Night Band (DNB)	2012–2016	Time Series	US NOAA National Geophysical Data Center	v1, 2017	Monthly night lights intensity, continuous rasters	15″ (~450 m)	Geo-tiff tiles/ flt32,uint8	GCS WGS 1984	Between latitudes 75° North and 65° South
ESA CCI Land Cover	2000–2015	Time Series	ESA CCI – LC	v2.0.7, 2017	Annual land cover, categorical rasters	9″ (~300 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
World Database of Protected Areas (WDPA)	1819–2017	Time Series	UNEP-WCMC and IUCN	June 2017	Terrestrial and marine protected areas, categorical vector	Comparable to 30″ (~900 m)	ESRI geodatabase	GCS WGS 1984	Global
JRC Global Human Settlement Layer (GHSL)	2000, 2014	Time Series	Pesaresi, et al.	2015	Urban settlement, categorical rasters	1.26″ (~38 m)	Geo-tiff/ uint8	Spherical Mercator projection (EPSG:3857)	~85.06 degrees North and South latitude
Global Urban Footprint (GUF)	2012	Time Invariant	DLR EOC	2016	Urban settlement, categorical raster	2.8″ (~84 m)	Geo-tiff/ uint8	GCS WGS 1984	Global

Source datasets are here described. Data source, version, format, and spatial and temporal information are summarised.

comparison. Exponential growth rates have been calculated for each administrative unit by matching the total population from the latest census enumeration to those from a previous census enumeration. In cases where matching at the highest spatial detail is not possible between the two points in time (e.g. boundary changes), censuses have been matched and growth rates calculated at a less detailed administrative level (state/province or district), and applied to each unit (municipality) within that highest administrative level. For further detail see Doxsey-Whitfield et al. (2015) and CIESIN (2016b). The growth rate has been calculated using the following formula:

$$r = \frac{ln\left(\frac{P_2}{P_1}\right)}{t} \tag{1}$$

where r is the annualised growth rate, P_1 is the population count at the time of the earlier census, P_2 is the population count from the latest census, and t is the number of years between the two. Population estimates were then calculated for the target years as follows:

$$P_x = P_2 e^{rt} \tag{2}$$

where P_x is the population estimate in the target year x, and P_2, r, and t are as defined previously (CIESIN (Center for International Earth Science Information Network, Columbia University), 2016b).

National and sub-national administrative unit boundaries follow census cartography if available (CIESIN, 2018). When census cartography is not available, non-census boundaries are utilised by CIESIN if obtainable. This is in order that the full effective spatial resolution of the tabular census data may be utilised. A country is gridded at a coarser resolution only if one of the tabular census data or administrative unit boundaries are not available. Particularly for non-census boundaries, reconciliation with census data is a significant undertaking in both time and labour, discussed further in Doxsey-Whitfield et al. (2015). In order to ensure consistency between countries, administrative unit boundaries are aligned to a global framework in part based on the Global Administrative Areas version 2 (GADMv2) (GADM, 2015), sourced primarily from national governments and NGOs. GADM is utilised because it is openly available, consistent, and widely used in the research community (CIESIN (Center for International Earth Science Information Network, Columbia University), 2016b). In cases where the resolution of the administrative unit boundaries far exceed that of the GADM boundaries, the former are kept (Doxsey-Whitfield et al., 2015). Average census administrative unit resolution for highly developed regions is 936 arc-second (~31 km at the Equator), and 1764 arc-seconds (~59 km at the Equator) for less developed regions, calculated from de Sherbinin and Adamo (2015). Where country boundaries follow GADM the effective spatial resolution is comparable to that of census boundaries but varies in quality according to the original source material.

Topography data consists of the Viewfinder Panoramas dataset (de Ferranti, 2017a), which is primarily US NASA Shuttle Radar Topography Mission (SRTM) data (US NASA 2016) collected in the year 2000, has 3 arc-second (~100 m at the Equator) horizontal and 1 m vertical spatial resolution, and is amended and corrected by the dataset developer Jonathan de Ferranti (de Ferranti, 2017a). Viewfinder Panoramas data are provided filled and corrected from the best available alternative sources where SRTM data are unavailable (i.e. north of 60° 2′N and south of 56° S) or for some mountain and desert regions between these latitudes where there are voids and areas of phase unwrapping error (de Ferranti, 2017a, 2017b). Alternative sources are topographic maps, Landsat images, and ASTER GDEM data – sources that are much more accurate than the simple interpolation of SRTM data (de Ferranti 2017a).

OpenStreetMap (OSMF and Contributors, 2016) data (for January 2016) are global "voluntary geographic information" stored as a global database. OSM data have an effective resolution comparable with SRTM1 at 1 arc-second (~30 m at the Equator) but varies according to source data. OSM data use a system of nodes, ways, and relations to define points in space, linear features/area boundaries, and the way in which these attributes work together. Tags are used to categorise and label each attribute (OSMF, 2018a). The frequency of contributions by individual users will refine source data, as often can contributions from out of copyright maps (OSMF, 2018b) or contributions from professional cartographic organisations (OSMF, 2018c).

The WorldClim 2.0 Beta version 1 (Fick & Hijmans, 2017) global temperature (°C) and precipitation (mm) data are each provided as 12 30 arc-second (~1 km at the Equator) spatial resolution raster images representing average monthly climate data for the period 1970–2000.

DMSP-OLS version 4 stable night-time lights (2000–2011) annual composite time series (US NOAA, 2015) are light intensity data provided as raster layers with near global coverage between latitudes 75 degrees North and 65 degrees South. Source data have 30 arc-second (~1 km at the Equator) spatial resolution. For the years 2000–2007 (inclusive) data are available from two satellites, whereas for the years 2008–2011 (inclusive) data are available from one satellite. The stable composite product contains lights from cities, towns, and other sites with persistent lighting, including gas flares. Ephemeral events, such as fires are not included (US NOAA, 2015). An inter-calibrated version of the stable night-time lights annual composites (Zhang et al., 2016), which provides relative radiometric calibration and saturation correction is utilised to produce the global harmonised lights data for this study up to the year 2011. The unit of radiance employed in the standard uncalibrated DMSP data is a digital number ranging from 0 to 63. The inter-calibrated version of the data multiplies the digital number of the source by 100.

Similarly, for 2012 to 2016 (inclusive) we use VIIRS Cloud Mask (VCM) version 1 night-time lights Day/Night Band (DNB) monthly composite time series (US NOAA, 2017) light intensity data, which is provided as inter-calibrated tiled raster layers with near global coverage between latitudes 75 degrees North and 65 degrees South. Source data have 15 arc-second (~450 m at the Equator) spatial resolution. Twelve monthly average radiance composites are available for each year 2013–2016, and 9 for 2012 (April–December inclusive). Each monthly composite is divided into six tiles (75°N, 180°W; 75°N, 60°W; 75°N, 60°E; 0°N, 180°W; 0°N, 60°W; 0°N, 60°E). For each tile and each month, there is also a cloud-free observation raster that records how many cloud-free observations have been made by the satellite for each pixel within the average radiance image. These coverage files allow the end user to differentiate between no data pixels (i.e. in this case zero observations due to cloud cover) and pixels where observations were made but no lights were observed. The DNB VCM version of the data excludes data impacted by stray light, lightning, lunar illumination, and cloud cover. Version 1 data are not filtered to screen out lights from aurora, fires, boats, and other temporal lights (US NOAA, 2017). The unit of radiance employed by the VIIRS DNB data is nanoWatts/cm2/sr. Original radiance values are multiplied by 1E9 in the source data.

ESA CCI annual global land cover time series (2000–2015) version 2.0.7 (ESA (European Space Agency) CCI (Climate Change Initiative) – Land Cover project 2017, 2017a) classifies land use sub-categories for agriculture, forest, grassland, wetland, settlement, and other (including water) (ESA CCI, 2017c). Source data are provided as raster layers with global coverage, and a 9 arc-second (~300 m at the Equator) spatial resolution.

United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC) World Database of Protected Areas (WDPA), version June 2017 (UNEP-WCMC and IUCN, 2017) is the most comprehensive global database on terrestrial and marine protected areas (Chape, Harrison, Spalding, & Lysenko, 2005), comprising both spatial data (i.e. boundaries) and attribute data (i.e. descriptive information) for all protected areas from 1819 to 2017 (UNEP-WCMC, 2017). The International Union for Conservation of Nature (IUCN) Protected Area Management Categories, stored within the database, help classify protected areas based on their primary management objectives (Dudley, 2008). Effective resolution varies according to original source data (UNEP-WCMC, 2017; Visconti et al., 2013).

JRC Global Human Settlement Layer (GHSL) (Pesaresi et al., 2015) GHS BUILT LDS2000, and LDS2014, GLOBE R2016A 3857 38 grids detail built-up presence of settlement for years 2000 and 2014 respectively. Data are provided per year, each split into two rasters with cumulative near global coverage (~85.06 degrees North and South latitude), at 1.26 arc-second (~38 m at the Equator) spatial resolution, in Spherical Mercator projection (EPSG:3857) (GHSL, 2015).

The DLR Global Urban Footprint (GUF) (DLR EOC 2016) GUF28 v1 raster grid details the built-up presence of settlement for the year 2012. Data are provided with global coverage, and 2.8 arc-second (~84 m at the Equator) spatial resolution (Esch et al., 2017).

2.2. Production of datasets

The methods used to harmonise the datasets are here described. We subsequently use produced geospatial layers as input to an RF model (Stevens et al., 2015), using methods for temporal considerations described by Gaughan et al. (2016) to produce gridded population outputs and demonstrate applications for such data.

2.2.1. Processing software

Open source OSGEO4W64 geospatial Software (OSGF, 2017a) and the included geospatial Data Abstraction Library (GDAL) v2.1.3 package (OSGF, 2017b) are employed to produce archive datasets, using a Microsoft Windows 7, 64-bit operating system (OS). Occasionally proprietary ESRI ArcMap v10.3.1 and ArcInfo Workstation v9.3 GIS software (ESRI, 2016) are utilised where specific functionality is otherwise unavailable. Program code is implemented as windows batch script files within OSGEO4W64 at command line unless otherwise stated. Scripts and supporting “readme” files are available to download from Figshare (Lloyd, Chamberlain, Kerr, & Bondarenko, 2018). ESRI ArcMap v10.3.1 is employed to create the Level 0 and Level 1 tiled data. Python (v.3.6) (PSF, 2016) and the included Pandas module is used to interpolate the time series of population data, and to merge missing records.

Software used for initial OSM database processing is as described in Lloyd et al. (2017). Subsequent database access, filtering, and processing (on the Windows platform) are provided by QGIS 2.18.4 (QGIS project, 2017) and Spatialite v4.3.0a, including the Spatalite graphical user interface (GUI) 2.0.0 (Furieri, 2016) software. GDAL, and SAGA GIS 4.1.0 (SAGA 2017) command line utilities are used to convert to raster format and standardise the data. The processing of the UNEP/IUCN WDPA utilises PostgreSQL 9.1 (PostgreSQL GDG, 2016) and PostGIS 2.0 (PostGIS PSC, 2016) database software, and GDAL. The standardisation of output has utilised the IRIDIS 4 High-Performance Computing (HPC) Facility at the University of Southampton, using a Linux OS (Redhat 6) and GDAL version 1.10.1. Similarly, GHSL rasters are processed using GDAL on the HPC, owing to source spatial resolution and the associated computational overhead.

2.2.2. Viewfinder panoramas topography, and slope derivative

All Viewfinder Panoramas topography tiles are first mosaicked into one global image using GDAL utilities. The topography data is standardised to grid definition and coastlines. No data pixels at coastal edges (present due to inconsistencies in coastline location between topography and L0 data) are filled, to produce the global topography layer. A global slope layer is created from the topography data using GDAL.

2.2.3. L0/L1 derivatives

To create the pixel area grid, an ARC Macro Language (AML) script (modified from Santini, Taramelli, & Sorichetta, 2010) calculates the surface area of cells in a regularly spaced longitude-latitude (geographic) grid of the Earth's surface at 60 arc-second resolution, using ESRI ArcInfo (Arc) software. Our approach to the surface area calculation is based on the spherical approximation of the Earth's surface described by Santini et al. (2010). The production workflow is a refinement of that described in Lloyd et al. (2017). A binary grid of coastline is created from the L0 country data. Binary grids are created for all produced categorical covariates, for potential application in modelling.

ESA CCI – LC v4.0 inland waterbody data (modified by CIESIN) are extracted from the L1 data and mosaicked onto OSM "waterway" tagged polylines (streams and rivers) to provide a contiguous inland water dataset that is fully integrated with the L1 census unit data. A separate contiguous OSM inland water (streams, rivers, lakes, etc.) layer is also created as an alternative dataset.

2.2.4. OpenStreetMap (OSM)

After initial OSM database processing, relevant data are exported and converted into raster format. In common with the workflow of Lloyd et al. (2017), QGIS is utilised to modify each database table (i.e. point, line, and polygon) and to convert the database attributes of interest into spatialite tables (i.e. spatially enabled SQLite databases) in order to allow greater and faster manipulation of spatial data than would otherwise be possible if working directly with the source database. A classification field is added to each spatialite table if required in order to rank features (such as the priority of roads in the highway network) for later preservation as pixel values when tables are converted to raster format (e.g. higher priority roads take precedence). Assignment simplifies tagging so as to be manageable for display in raster format. Subsequently, attribute (tag) extraction from a given spatialite table, further processing specific to each subset (i.e. highways, waterways, etc.), and conversion to raster format can take place, using a combination of QGIS, GDAL (ogr2ogr utility, using sqlite SQL dialect) and SAGA GIS. For each subset, the relevant tagging filters utilised during extraction, any associated variants and/or misspellings of tags, as well as excluded tags, are detailed in the supplementary code. Particular attention has been paid to manual examination of OSM tagging in order to extract maximum information from OSM data across all subsets. For reasons of computational efficiency during the execution of certain algorithms (e.g. intersection), spatialite tables are tiled for processing before rasterisation and standardisation. Further, specific workflow for each OSM subset are here summarised, with further detail supplied in the supplementary material.

2.2.5. OSM highways

A highways layer with "highway" tags assigned a classification field "priority" value of 1–17 (footpath to motorway road classes) is created. Priority value assignment is detailed in Supplementary Table 1. A "bridge" and "tunnel" tagged (henceforth referred to as "links") layer is also created, only for those "inter-coast" highways situated over/under water (e.g. bridges and tunnels at estuaries, narrow sea ways, etc.). Such links are removed during the standardisation of highway rasters. Creation of the layer allows links to be restored after standardisation, so that coastal roads remain contiguous. Links are given an arbitrary priority value (of 30) to differentiate them from the rest of the road network. This part of the workflow follows that of Lloyd et al. (2017), albeit subsequently standardising to more accurate coastal boundaries.

Highway and links sets of spatialite tiles are each separately converted to a vector format using GDAL (ogr2ogr) for compatibility with the SAGA GIS rasterisation ("Shapes to grid") command line tool. Using GDAL, two copies of L0 raster tiles (at 100 m spatial resolution) are made, with country code values reset to an arbitrary value, and tile extents identical to the vector tiles. Onto these copies are rasterised the maximum priority attribute value that is apparent per 100 m pixel, for each set. Each set of raster tiles is mosaicked, and the highway mosaic standardised. The links mosaic is standardised so that only those features located offshore are retained. The two standardised layers are mosaicked together and background land values set to zero. A calculation is performed to produce the final highways layer for classes 8–30 (i.e., tertiary to motorway road classes, plus links). In addition, highway priority classes are each extracted individually (for classes 8–30) and rasterised separately using the same method in order to increase variability during modelling. Highway classes 8–30 are considered to be major highway routes, which are particularly well correlated with population density and so are significant for the purpose of modelling. Lesser classes are excluded.

2.2.6. OSM highway intersections

A highway intersection layer for road priorities of 8–17 (tertiary to motorway road classes) is created. These classes are selected because they represent only major highway routes and therefore provide only "significant" highway intersections. The inclusion of lower classes of the road via (for example) introduction of residential streets (class 7) into the layer would provide overly dense and potentially misleading intersection information after rasterisation (particularly in urban centres). Exclusion of higher classes would remove otherwise useful intersection information. Road intersection points are rasterised using the same process as described for highways (but using the simpler data/no data output value option). Prior processing to identify intersection uses the highway spatialite table and utilises ogr2ogr and associated PostGIS functions (PostGIS PSC, 2017a, 2017b).

Our approach to identifying intersection defines highways as having uniform road name, reference number, junction, and priority tags. Where one of these criteria change, an intersection will be found. Our approach identifies where highways cross bridges and tunnels. Geospatial utilities will identify intersections at such crossing points, where of course no such highway intersections exist in real life. Such false intersections are entirely removed by our technique. Further technical elucidation regarding the intersection method can be found in the supplementary material.

2.2.7. OSM waterways

A natural waterway layer is created. Three types of natural water attributes are separately extracted from the database. These attributes are “waterway” polylines (streams and rivers), riverbank polygons (where rivers, or similar, have a quantifiable width at source data resolution), and lake polygons (or similar). Canal waterways are included, despite being anthropogenic, because of their relevance to human population, transportation, and water supply. Filtered attributes are converted to three spatialite tables, extracted, rasterised, and standardised. Waterbodies are mosaicked onto riverbanks, and in turn onto waterways, using GDAL, to form one contiguous water layer.

2.2.8. Worldclim 2.0 beta version 1

2.2.8.1. Temperature. To create an average annual global temperature layer for the period 1970–2000, the 12 average monthly temperature rasters for the period are averaged using ESRI ArcMap Raster Calculator tool (ESRI, 2018a). The output raster is partially standardised (i.e. only spatial alignment, resolution, no data value) to a 1 km resample and reclassification (to “zero” value, ocean no data) of our coastline grid, using GDAL. In order to fill no data pixels at coastal edges, no data values in the partially standardised grid are modified to zero, converted to actual values, and output is summed with the 1 km coastline grid. Coastal areas in the modified grid are “nibbled” using the ESRI ArcMap Nibble tool (ESRI, 2018b), using the original partially standardised raster as a mask. Only data values are allowed to nibble into areas defined in the mask raster (ESRI, 2018b). Prior to use of the nibble tool, the values of each input grid are multiplied by one million (in order to preserve data precision), and then each grid is converted to integer format as a requirement of the tool. Output from the nibble tool is converted back to float format and the previous multiplication calculation reversed in order to restore original values. Output is then resampled to 100 m resolution using bilinear interpolation in GDAL and standardised to L0 coastlines. An ocean mask is applied and no data values asserted.

2.2.8.2. Precipitation. To create an average annual global precipitation layer for the period 1970–2000, the 12 average monthly precipitation rasters for the period are summed using GDAL. The output raster is partially standardised (i.e. only spatial alignment, resolution, no data value, data type) and then the same workflow as described for the Worldclim temperature grid is followed (where applicable).

2.2.9. DMSP-OLS version 4 stable night-time lights (2000-2011) annual composite time series

A time series of near global night-time lights annual composites is created for 2000–2011 using GDAL. Inter-calibrated annual composite input radiance rasters are averaged where data is available from two satellites (i.e., 2000–2007). The eight output grids and the grids representing 2008–2011 are subsequently standardised, being resampled to 100 m spatial resolution using nearest neighbour technique. Areas of no data coverage in polar regions are replaced with zero values, an ocean mask is applied, and the no data value asserted for each grid.

2.2.10. VIIRS cloud mask (VCM) version 1 night-time lights (2012-2016) Day/Night Band (DNB) annual composite time series

A time series of near global night-time lights annual composites is created for 2012–2016 using GDAL. For a specified input raster tile and year, annual average nightlights radiance values are calculated. Values in the 12 monthly average radiance input rasters per year are summed (or nine in the case of 2012). The equivalent "cloud-free observations" coverage rasters are converted to binary (to reflect which pixels have cloud-free observations and which have none, in any given month) and summed in order to identify no data pixels for each year. Using the output, a calculation is performed to eliminate from the summed radiance tiles, pixels with no recorded observations – and to attenuate summed radiance pixel values by the number of months for which night lights have been observed (rather than by the cumulative number of observations per year – monthly radiance input rasters are already averaged per month). By this method, tiles are created that display average radiance for each year. Radiance and coverage tiles are mosaicked per year, and a calculation performed on each annual radiance mosaic – utilising the annual coverages as masks in order to interpolate no data pixels using surrounding values. Output grids are then standardised as for DMSP night lights grids.

2.2.11. ESA CCI annual global land cover time series (2000-2015)

In order to create an annual global land cover time series for 2000–2015, land use subcategory classifications are extracted and simplified (to nine classes) for each annual input grid, for the efficiency of use in population analyses. GDAL is used throughout. The aggregated reclassifications can be found in Supplementary Table 2.

The output grids (containing individual aggregated classes) are summed to produce one raster (containing all aggregated classes) for each year and standardised as for the night lights grids (where applicable). Classes are individually extracted from the standardised grid for each year and converted to binary (1,0, no data) and single value (1, no data) stand-alone layers. Production of binary classes and simplification of the land cover grid increases variability for potential use in modelling. Built settlement datasets (or equivalent land cover classes) are particularly well correlated with population density and so are significant for the purpose of modelling (Nieves et al., 2017).

2.2.12. UNEP/IUCN world database of protected areas (WDPA)

To create an annual global time series, detailing the extent of (terrestrial/coastal/marine) protected areas from 2000–2017, source geodatabase polygons are dissolved based on year of designation and level of IUCN protection category. Some protected areas are represented by points in the database. Points are buffered by 70 m and the resulting circles used as a proxy for protected areas. This pre-processing has been undertaken by CIESIN. The remainder of processing is undertaken using GDAL. The database is imported into PostGIS using ogr2ogr. Geometry errors triggered by differing table rules (between ESRI and PostgreSQL) are rectified using the PostGIS command ST_MakeValid (PostGIS PSC, 2017a). Two integer classification columns are added to the table in order to facilitate SQL queries. One duplicates the marine code (0 = terrestrial; 1 = coastal; 2 = marine), and the other the ICUN protection category (1 = ICUN 1a and 1b; 0 = other categories).

Polygons are rasterised (using gdal_rasterise) incrementally on an annual basis from 2000–2017. The 2000 grid includes all prior years. For computational efficiency, years prior to 2000 are rasterised decadally up to 1960, and then annually until 2000 – and mosaicked. Subsequent years are each incrementally mosaicked onto the previous year. This process leads to four rasters being produced for each year from 2000–2017 – an ICUN category '1' and an ICUN category "others" raster for each of terrestrial and marine/coastal protected areas. In total, 72 output rasters, therefore, represent the 18 years. Subsequent processing is undertaken using the HPC. The 36 Terrestrial rasters are standardised and mosaicked onto the marine counterparts. Marine rasters are partially standardised (i.e. not to coastlines) as coastal protected areas straddle marine and terrestrial environments. The result is 36 protected areas rasters, two per year – each denoting an ICUN category.

2.2.13. JRC global human settlement layer (GHSL) & DLR global urban footprint composites for 2000, 2012, and 2014

In order to create a built settlement time series (2000, 2012, 2014), which can subsequently be extrapolated and interpolated annually as per work by Nieves et al. (2018), the GHSL built settlement grids for years 2000 and 2014 are produced. Each year is provided as two rasters. The two rasters are first joined and re-projected from Spherical Mercator projection (EPSG:3857) to geographical coordinate system (GCS) with WGS 1984 datum (EPSG:4326). Each yearly grid is then standardised, being resampled using nearest neighbour technique. Areas of no data coverage in polar regions are replaced with zero values, the value used to denote built settlement modified, and the no data value asserted for each grid. GUF built settlement data for the year 2012 are provided as a single raster. This grid is standardised as for GHSL (where applicable).

For the purpose of refining the accuracy of the built settlement grids, the GHSL 2000 layer is combined with the ESA CCI 2000 built landcover class. An ESA 2000 settlement pixel is only retained in the final year 2000 layer if further classified as a settlement pixel in GUF 2012. The 2012 and 2014 built settlement layers are, respectively, created by mosaicking GUF 2012 onto the year 2000 layer, and by mosaicking GHSL 2014 onto GUF 2012. The refined multi-temporal built settlement outputs are particularly well correlated with population density and so are significant for the purpose of modelling. The rationale for combining layers is that GHSL has large areas of densely built settlements that are missing due to imagery or atmospheric conditions at the time of collection. The ESA built settlement class is back-filtered using GUF (more accurate radar data) because the ESA data is more likely to have errors of commission due to roads, bare soil, etc. (owing to the nature of the satellite sensor). In back-filtering, a limit is placed on where ESA data is allowed to fill gaps in the GHSL.

3. Technical validation

Harmonised datasets produced for this paper have been obtained by processing input source data to produce consistent 3 arc-second outputs. Source data are validated by independent studies (Brigham, Gilbert, & Xu, 2011; Cao & Bai, 2014; CIESIN 2016c; ESA, 2017c; Esch et al., 2017; Fick & Hijmans, 2017; Henderson, Yeh, Gong, Elvidge, & Baugh, 2003; Hormann, 2018; Iwao, Nishida, Kinoshita, & Yamagata, 2006; Lloyd et al., 2017; Min, Gaba, Sarr, & Agalassou, 2013; Muck, Klotz, & Taubenbock, 2017; Pesaresi et al., 2016; Rabus, Eineder, Roth, & Bamler, 2003; Rodríguez et al., 2005; UNEP-WCMC, 2017; US NOAA, 2017; Varga & Bašić, 2015; Visconti et al., 2013). An exception is Open Street Map source data, which do not comply with standard quality assurance procedures (Haklay, Basiouka, Antoniou, & Ather, 2013) because OSM is "volunteered geographical information" provided by any number of individual contributors. However, OSM data have intrinsic quality assurance through analysis of the number of contributions for a given spatial unit. The assumption that as the number of contributors increase then so does the quality of the data is known as "Linus" Law'. Recent studies show that for OSM data this rule applies with regard to positional accuracy (Haklay et al., 2013). Whilst effective spatial resolution of OSM data is high, there is a lack of sufficiently standardised user tagging of attributes. This can cause inaccuracies and difficulties in map rendition (Lloyd et al., 2017). We provide the harmonised OSM data as a time-invariant layer in order to minimise issues common in volunteered geographic information relating to data completeness and heterogeneity. As of January 2016, OSM highway data are estimated to be globally ~83% complete with more than 40% of countries, including several in the developing world, having a fully mapped street network (Barrington-Leigh & Millard-Ball, 2017). Only the most significant road classes are included in the harmonised output, as these are likely the most complete globally. A further exception in terms of quality assurance is WDPA source data, which are subject to a series of quality checks and reformatting (by WDPA) to ensure that data standards are met. However, due to the inherent variability of data submitted by a wide range of providers with different capacity and resources to digitise protected area boundaries, issues with the accuracy of the WDPA should be expected (UNEP-WCMC, 2017). Discrepancies generated by such differences in resolution are discussed in Visconti et al. (2013).

4. Dataset value

The archive of harmonised geospatial layers is summarised in Table 2, with a visualised sample presented in Figure 2. Further, we present applications of the geospatial layers, as input to a Random Forest model (Gaughan et al., 2016; Stevens et al., 2015), and demonstrate the potential usefulness of these data in health and development metric applications.

A sample of harmonised gridded layers is visualised. (A) shows a plan view of L1 subnational administrative boundaries for the region surrounding the city of Hetauda, situated in the Makwanpur District of the Narayani Zone of southern Nepal, superimposed with harmonised 2014 built settlement (symbolised in black), and 2016 OSM roads (grey), and waterways (blue). (B) shows a pseudo-3D stack of grids for the same location, with (in ascending order) topography (ascending, blue-white), slope (green-red), L1 administrative units, protected conservation areas (2016), ViiRS nightlights (2016; green-white), ESA CCI reclassified land cover (2015; vegetation – green shades; waterbodies – blue; built settlement – black), built settlement (2014; black), and OSM layers (roads – orange; waterways – blue).

4.1. Random forest model

An RF-based dasymetric modelling approach is utilised to produce initial population count outputs. The approach is described in Stevens et al. (2015). We utilise the model to incorporate census data and a combination of the open access, remote-sensed and geospatial datasets discussed in this paper, in order to contribute to modelled dasymetric weights (Stevens et al., 2015). The RF model is used to generate a gridded prediction of population density at 3 arc-second spatial resolution (approximately 90 m resolution at the equator). This prediction layer is utilised as the weighting surface to perform dasymetric redistribution of census population counts to the pixel level all across a country in order to obtain the population distributions at a scale finer than the source subnational administrative units (Stevens et al., 2015).

4.2. Application 1: change in population at risk of p.falciparum malaria in Africa between 2000 and 2014

The Malaria Atlas Project (MAP) has used gridded population data as the denominator in malaria prevalence calculations for many years. It is this data and associated graphical output that is used in the (WHO World Malaria Report, 2015), produced annually. However, the MAP rely on a static denominator and some basic interpolation assumptions (Bhatt et al., 2015). We use the MAP modelled parasite prevalence rate of Plasmodium falciparum malaria (i.e. the proportion of the population with detectable parasites per year; Figure 3, Top) (Bhatt et al., 2015), as well as initial population count outputs from the previously published RF model (Stevens et al., 2015), in order to present the Log10 of change in country population (count) and the change in percentage of country population (Figure 3), at risk of Plasmodium falciparum malaria infection between 2000 and 2014 where prevalence is >10%. The table output can be found in Supplementary Table 3.

One of the Millennium Development Goals (MDGs) aimed to halt and begin to reverse the spread of malaria by 2015 (UN General Assembly, 2000). This target has been achieved – between 2000 and 2015, new cases in Africa fell by 42%, with mortality rates falling by 66% (WHO, 2015). However, progress has since stalled (WHO, 2017). Our output shows that, in many instances, country population at risk of malaria has fallen drastically between 2000 and 2014. Whilst a significant reduction in prevalence (where >10%) is apparent when the MAP data for the two time periods are compared (Figure 3, Top), the powerful combination of multi-temporal population data and malaria data for the same periods facilitates a very clear and detailed graphical (Figure 3, Bottom) and tabular (Supplementary Table 3) representation (and, therefore, understanding) of the change in actual country population count, and change in percentage of country population, at risk. It is clear that particularly good progress in risk reduction has been made in Gambia (a 68% reduction), Rwanda (71%), Senegal (64%), Guinea-Bissau (69%), Tanzania (52%), and Angola (50%), to name a few – but that there is still much work to do, with little to no progress made since 2000 in many other countries such as Ghana (in which nearly 100% of the population is still at risk), Mali (the same), Malawi (a 5% reduction, to nearly 95% risk), Mozambique (a 1% increase in risk since 2000, to 97%), and Nigeria (a 2% increase, to 96%). By using multi-temporal population data we can uncover trends about how

Table 2. Geospatial raster layers produced for potential input to a population model.

Name	Acquisition year	Temporal variation	Source	Version, publication year	Data type	Spatial resolution	Format/ pixel type & depth	Spatial reference	Spatial coverage
National L0 and sub-national census L1 administrative boundaries with integrated waterbodies	2005–2014/ 2000–2012	Time Invariant	Center for International Earth Science Information Network (CIESIN), Columbia University/ ESA CCI – LC	GPW v4, 2016/ v4.0 2017	Global population count and administrative boundaries, inland water bodies, table and categorical rasters	3″ (~90 m)	Geo-tiff/ uint16,uint32	GCS WGS 1984	Global
Pixel area	Derived from calculated Earth surface area grid and the country ID L0 layer				Pixel area, categorical rasters	3″ (~90 m)	Geo-tiff/ uint32	GCS WGS 1984	Global
Topography	~2000	Time Invariant	de Ferranti, J.	28/11/17	Elevation, continuous raster	3″ (~90 m)	Geo-tiff/ int16	GCS WGS 1984	Global
Slope	Derived from topography				Slope, continuous raster	3″ (~90 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
Open Street Map (OSM)	2016	Time Invariant	OpenStreetMap Foundation (OSMF) & Contributors	15/01/16	Highways, highway intersections, waterways, categorical rasters	3″ (~90 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
WorldClim 2.0	1970–2000	Time Invariant	Fick, S.E. and Hijmans, R.J.	01/06/16	Annual temperature and precipitation, continuous rasters	3″ (~90 m)	Geo-tiff/ flt32, flt32	GCS WGS 1984	Global
DMSP-OLS Stable Nightlights	2000–2011	Time Series	US NOAA National Geophysical Data Center; Zhang et al.	v4, 2015; inter-calibrated, 2016	Annual night lights intensity, continuous rasters	3″ (~90 m)	Geo-tiff/ int16	GCS WGS 1984	Between latitudes 75° North and 65° South

(*Continued*)

Table 2. (Continued).

Name	Acquisition year	Temporal variation	Source	Version, publication year	Data type	Spatial resolution	Format/ pixel type & depth	Spatial reference	Spatial coverage
ViiRS Cloud Mask (VCM) Nightlights Day/ Night Band (DNB)	2012–2016	Time Series	US NOAA National Geophysical Data Center	v1, 2017	Annual night lights intensity, continuous rasters	3″ (~90 m)	Geo-tiff/ flt32	GCS WGS 1984	Between latitudes 75° North and 65° South
ESA CCI Land Cover	2000–2015	Time Series	ESA CCI – LC	v2.0.7, 2017	Annual land cover, categorical rasters	3″ (~90 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
World Database of Protected Areas (WDPA)	2000–2017	Time Series	UNEP-WCMC and IUCN	June 2017	Terrestrial and marine protected areas, categorical rasters	3″ (~90 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
Urban Settlement	2000, 2012, 2014	Time Series	ESA CCI – LC / Pesaresi, et al. / DLR EOC	2017/ 2015/ 2016	Urban settlement, categorical rasters	3″ (~90 m)	Geo-tiff/ uint8	GCS WGS 1984	Global
Binary grids, for all categorical layers	-	-	-	-	Presence of features, categorical rasters	3″ (~90 m)	Geo-tiff/ uint8	GCS WGS 1984	Global

Potential population model input datasets are here described. Data source, version, format, and spatial and temporal information are summarised. See Methods section for production workflow.

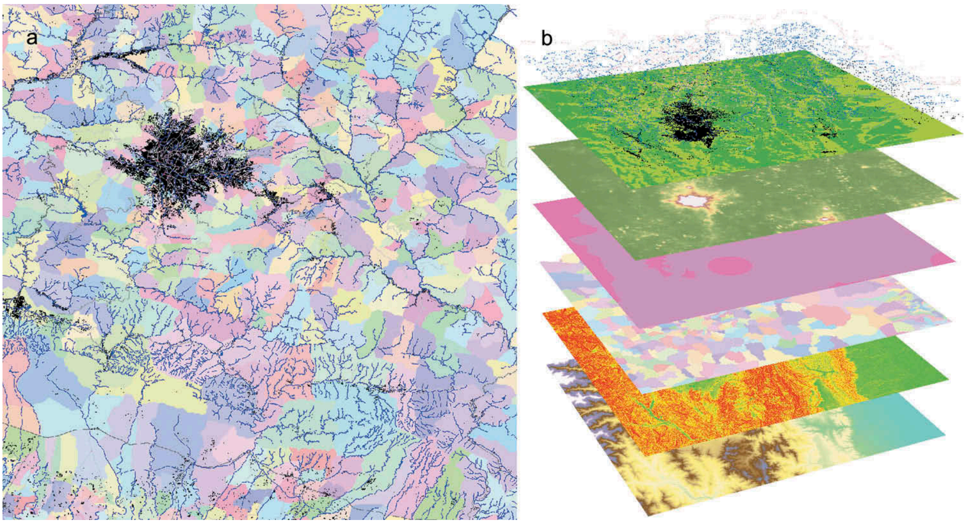

Figure 2. Geospatial raster layers produced for potential input to a population model.

in some countries the proportion and numbers at risk are increasing, despite general prevalence declines.

4.3. Application 2: change in population living in proximity to conflict in Africa between 2000 and 2014

Understanding the numbers impacted by conflict, and associated displacement trends, can be important for humanitarian relief contingency planning, as well as long term government policy. Conflicts are very geographically focussed and fluctuate a lot over time. Hence, there is a need for spatially detailed multi-temporal population data to obtain these metrics. We use the Armed Conflict Location & Event Data Project (ACLED (Armed Conflict Location & Event Data Project), 2018) disaggregated conflict and crisis mapping for Africa for years 2000, 2012 and 2014, and corresponding initial population count outputs from the previously published RF model (Stevens et al., 2015), to present the change in percentage of population living in proximity to conflict for each African region (North, East, Central, West, South) as defined by the UN Department of Economic and Social Affairs (2018). For each region and year, populations are considered to be proximal to a conflict where within a 9 × 9 km zone containing two or more conflict events. For the purpose of this example application of the multi-temporal data, zone size has been selected to represent a reasonable area within which people may be displaced as a result of a conflict event. The zones are displayed in Figure 4 (Top), per each region. Figure 4 (Bottom) depicts the percentage change over time of regional population living in proximity to conflict, per each region. The table output can be found in Supplementary Table 4. ACLED collects the dates, actors, types of violence, locations, and fatalities of all reported political violence and protest events across Africa, as well as elsewhere. Political violence and protest include events that occur within civil wars and periods of instability, public protest and regime breakdown (ACLED (Armed Conflict Location & Event Data Project), 2018).

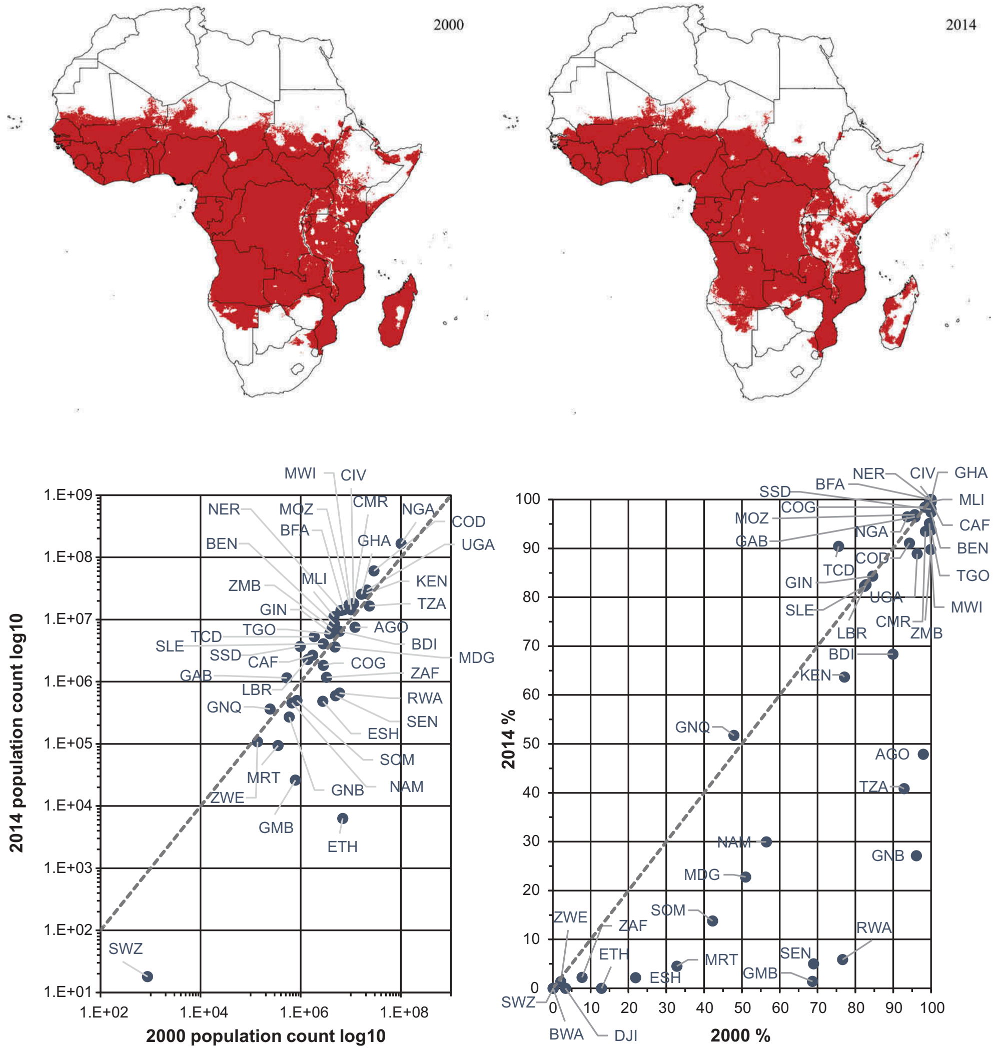

Figure 3. (Top) *Plasmodium falciparum* malaria prevalence rate, where >10%, for the years 2000 and 2014. (Bottom Left) Log10 of change in country population (count) at risk of *Plasmodium falciparum* malaria infection between 2000 and 2014, where prevalence is >10%. Countries (identified by ISO 3166 standard) below the trend line demonstrate a decrease in actual population count at risk of malaria infection between the respective years. (Bottom Right) Change in percentage of country population at risk of *Plasmodium falciparum* malaria infection between 2000 and 2014, where prevalence is >10%. Countries below the trend line demonstrate a decrease in the percentage of the total country population at risk of malaria infection between the respective years.

MDGs did not specifically mention conflict (e.g. civil wars, inter-state wars, and violence against civilians). A downward trend in the annual frequency of conflict in the world ended in the mid-2000s. Of 55 conflict-affected countries in 2015, 37 (67%) had met only two or fewer of the 15 MDGs (Norris, Dunning, & Malknecht, 2015). At least 20 of these countries are African (Themnér & Wallensteen, 2012). Even within otherwise stable countries, conflict-affected areas fared worse than areas with less or no conflict (MPSMRM (Ministère du Plan et Suivi de la Mise en œuvre de la Révolution de la Modernité), MSP (Ministère de la Santé Publique), and ICF

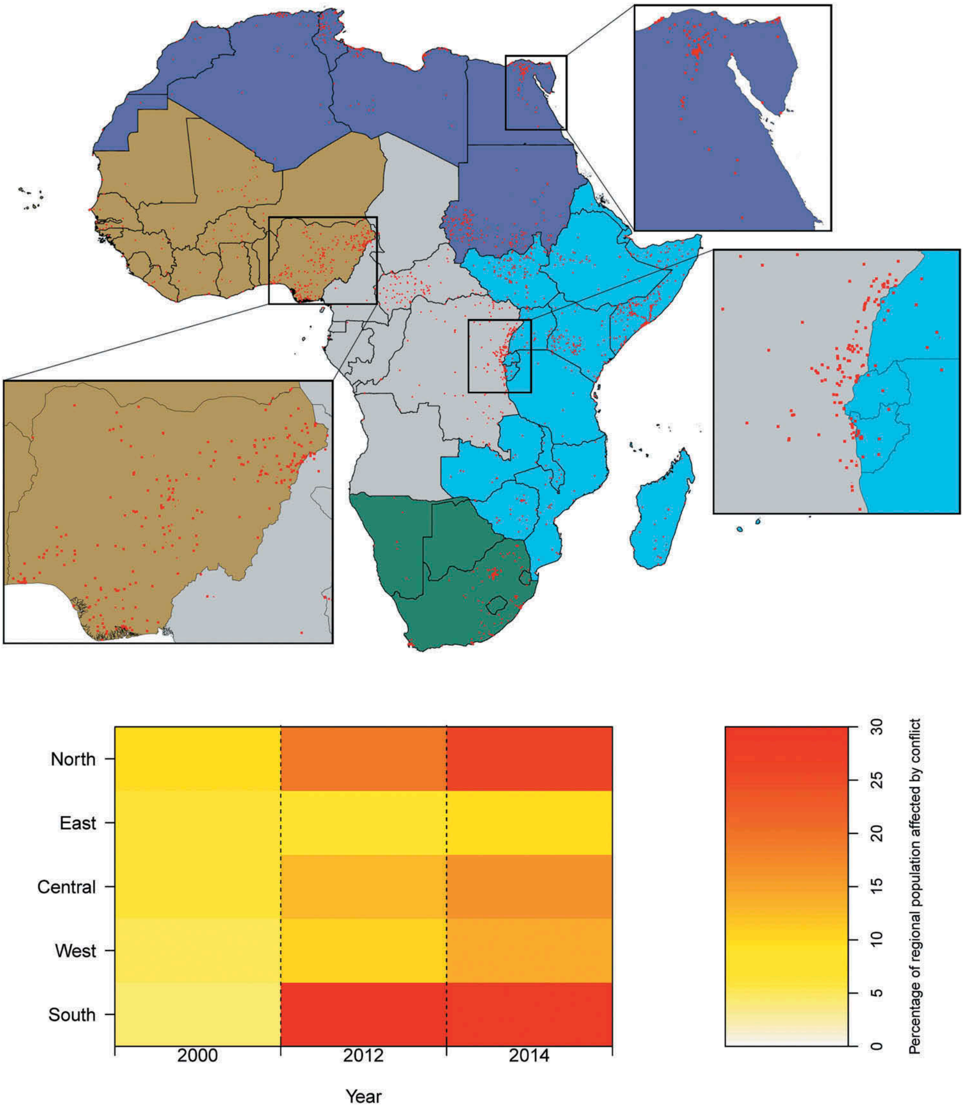

Figure 4. (Top) Zones (red dots) containing two or more conflict events in 2014, per each African region (Northern, Eastern, Central, Western, and Southern; depicted in purple, blue, grey, olive, and green, respectively). Break-out boxes show the same for Nigeria; The Nile, Egypt; and the eastern border of the Democratic Republic of the Congo. (Bottom) Change in the percentage of the regional population living in proximity to conflict, between 2000, 2012, and 2014, per each African region.

International, 2014). Our output demonstrates that the change in the population living in proximity to conflict, between 2000, 2012, and 2014, is in line with the accepted consensus that the conflict situation has deteriorated during this period. As is the case in the malaria application, the powerful combination of multi-temporal population data and conflict zone data (Figure 4, Top) facilitates a very clear/detailed graphical (Figure 4, Bottom) and tabular (Supplementary Table 4) representation/understanding of the change in the percentage of population living in proximity to conflict. It is clear that the percentage of those living in proximity to conflict in Africa in the year 2000 can be considered low, at between 4% and 10% in all regions, the highest being in the North. However, by 2012 this range is between 9% and 30% with the Northern and Southern regions particularly badly affected. This situation has deteriorated further by 2014, in all regions apart from the South, with a range of between 10% and 29% of population in proximity to conflict across Africa.

4.4. Summary

Global, harmonised, geospatial datasets are important for consistent and standardised inputs for any type of modelling or comparison effort. This can include any number of discipline-specific foci including health, ecology, climate, and so on. The output of the work described here provides a valuable resource for both applied and research oriented efforts where challenges with data access, quality and consistency are low. These data products are important in providing consistency in application across countries, in order to achieve or monitor progress towards a variety of Sustainable Development Goals (SDGs). Further, these data products are important in providing consistency in the application within countries, which is arguably as important. Monitoring progress towards SDG achievement at sub-national scales (via assessment of health and socio-economic development metrics) relies on the acquisition of ongoing spatially detailed sub-national scale data on population counts and distributions (Tatem, 2017). It is therefore important to improve the availability of and access to disaggregated data and statistics. There is a need to take urgent steps to improve the quality, coverage and availability of disaggregated data in order to target interventions and ensure that no one is left behind (UN General Assembly, 2015). The applied examples, detailed in sections 3.2 and 3.3, demonstrate the potential usefulness of multi-temporal gridded population data at the subnational level, for use in the monitoring of health and development metrics.

5. Usage notes

Future global high-resolution population mapping can use these unique, open access, geospatial datasets to construct consistent and comparable, freely available, and potentially age-structured, annual high-resolution global population distribution layers for the 2000–2020 period, perhaps using methods for temporal considerations described by Gaughan et al. (2016). Future methods can involve fine-tuning of covariates used as input to an RF (or other type of) model, utilising a covariate selection optimised per region, continent, or globally as per user requirements. Some users may wish to produce population distribution datasets avoiding the use of certain geospatial datasets discussed in this paper, in order to avoid any endogeneity within their own research. The geospatial datasets can perhaps be improved upon in the future via use of updated OSM data, the replacement of existing OSM layers (which are time invariant) with newly harmonised multi-temporal datasets where appropriate, and/or replacement of datasets with those of higher spatial resolution, as they become available. Similarly, other multi-temporal datasets which correlate well with population density would be valuable additions to the archive, as would updates to existing annual layers. Suitable additions for population analysis might include agricultural layers (seasonal variation may be needed for migration predictions), the location of conflict zones (which disperse population) or the location of major employers/industries in rural areas (which gather working population). Further, potential exists in terms of measuring the impact of the downscaling of covariate datasets upon population and built settlement growth models. Five out of 11 source datasets: ESA CCI land cover (9″), DMSP nightlights (30″), ViiRS nightlights (15″), WDPA (30″), and WorldClim 2.0 (30″) have been downsampled to make the spatial resolution common. In an RF-based dasymetric model, the impact of this downsampling on predicted values could be lower as the model is trained on the mean of the aggregated data. However, for a built settlement model, because the model is trained on disaggregated pixel values of the selected samples and covariates, the impact could be higher. If downsampled covariates have a higher importance in model training then the impact of downsampling on predicted values may also be analysed in terms of sample size and sampling strategy.

Limitations of the geospatial dataset gridding process include the potential for small islands to be absent from the country ID base grid because the islands are not present in source CIESIN data. This has the consequence that corresponding small island topographic or other spatial data are excluded from the harmonised geospatial layers. Further, where coastlines differ between L0 country ID and input topography/other spatial layers, coastal pixels (with a data value) may be removed from the output grid during harmonisation. When linking the census table to the L1 census unit raster, it has been found that a few administrative units are smaller than the resolution of the raster. In such instances, in order to preserve population counts, the administrative unit is removed from the table and the corresponding population count added to that belonging to the neighbouring (larger) administrative unit as defined by the respective pixel in the raster. Further to this, population estimate interpolation/forecast do not take into account natural disasters or similar events. This is justified by the global and temporal extent of the study.

Acknowledgments

This work forms part of the outputs of WorldPop (www.worldpop.org) and the Flowminder Foundation (www.flowminder.org). The funders had no role in study design, data collection and analysis, decision to publish, and preparation of the manuscript. The authors acknowledge the use of the IRIDIS High-Performance Computing (HPC) Facility, and associated support services at the University of Southampton, in the completion of specific aspects of this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Bill and Melinda Gates Foundation [OPP1134076; OPP1106427; OPP1032350; OPP1094793]; National Institute of Allergy and Infectious Diseases [U19AI089674]; Wellcome Trust [106866/Z/15/Z].

Data Availability Statement

The harmonised geospatial layers discussed in this paper are the product of the “Global High-Resolution Population Denominators Project” (WorldPop (www.worldpop.org) School of Geography and Environmental Science – University of Southampton, Department of Geography and Geosciences – University of Louisville, Département de Géographie – Université de Namur, & Center for International Earth Science Information Network (CIESIN) – Columbia University, 2018). Layers are being made publically available via the WorldPop FTP server (ftp://ftp.worldpop.org.uk/GIS/Covariates/Global_2000_2020/0_Mosaicked/), where licensing permits, reposited at the University of Southampton (https://doi.org/10.5258/SOTON/WP00650). Data is made available in the GeoTIFF format, with global coverage where source data permits.

Supporting information

The following supporting information is available as part of the online article: Supplementary material S1. Methods elucidation: OSM highways, highway intersections, ESA CCI land cover; Results: Additional tables.

ORCID

Christopher T. Lloyd http://orcid.org/0000-0001-7435-8230
Greg Yetman http://orcid.org/0000-0002-5270-6975
Forrest R. Stevens http://orcid.org/0000-0002-9328-3753
Jeremiah J. Nieves http://orcid.org/0000-0002-7423-1341
Graeme Hornby http://orcid.org/0000-0002-2833-8711
Alessandro Sorichetta http://orcid.org/0000-0002-3576-5826
Andrew J. Tatem http://orcid.org/0000-0002-7270-941X

References

ACLED (Armed Conflict Location & Event Data Project). (2018). *Armed conflict location & event data project* [Data set]. Retrieved March 2018, from https://www.acleddata.com/about-acled/

Amos, B., McDonald, M. P., & Watkins, R. (2017). When boundaries collide. *Public Opinion Quarterly*, *81*(S1), 385–400. doi: 10.1093/poq/nfx001

Azar, D., Engstrom, R., Graesser, J., & Comenetz, J. (2013). Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sensing of Environment*, *130*, 219–232. doi: 10.1016/j.rse.2012.11.022

Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I., & Nelson, A. (2006). Determining global population distribution: Methods, applications and data. *Advances in Parasitology*, *62*, 119–156. doi: 10.1016/S0065-308X(05)62004-0. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16647969

Balk, D. L., Pozzi, F., Yetman, G., Deichmann, U., & Nelson, A. (2005). *The distribution of people and the dimension of place: Methodologies to improve the global estimation of urban extents*. Paper presented at the Urban Remote Sensing Conference, Tempe, AZ.

Balk, D. L., & Yetman, G. (2004). *The global distribution of population: Evaluating the gains in resolution refinement*. Center for International Earth Science Information Network (CIESIN), Columbia University. Retrieved from http://sedac.ciesin.columbia.edu/downloads/docs/gpw-v3/gpw3_documentation_final.pdf

Balk, D. L., Yetman, G., & de Sherbinin, A. (2010, October 5–7). *Construction of gridded population and poverty data sets from different data sources*. Paper presented at the European Forum for Geography and Statistics, Tallinn, Estonia.

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PloS One*, *12*(8), e0180698. doi: 10.1371/journal.pone.0180698. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28797037

Bennett, M. M., & Smith, L. C. (2017a). Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sensing of Environment*, *192*, 176–197. doi: 10.1016/j.rse.2017.01.005

Bennett, M. M., & Smith, L. C. (2017b). Using multitemporal night-time lights data to compare regional development in Russia and China, 1992–2012. *International Journal of Remote Sensing*, *38*(21), 5962–5991. doi: 10.1080/01431161.2017.1312035

Bhaduri, B., Bright, E., Coleman, P. R., & Dobson, J. (2002). LandScan: Locating people is what matters. *Geoinformatics*, *5*(2), 34–37.

Bhaduri, B., Bright, E. A., Coleman, P. R., & Urban, M. L. (2007). LandScan USA: A high-resolution geo-spatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, *69*, 103–117.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., … Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, *526*(7572), 207–211. doi: 10.1038/nature15535. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26375008

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/a:1010933404324

Brigham, C., Gilbert, S., & Xu, Q. (2011). *Open geospatial data: An assessment of global boundary datasets. World Bank Institute*. Paper presented at the The Proceedings of GISRUK 2012, Lancaster.

Cao, C., & Bai, Y. (2014). Quantitative analysis of VIIRS DNB nightlight point source for light power estimation and stability monitoring. *Remote Sensing*, *6*(12), 11915–11935. doi: 10.3390/rs61211915

Chape, S., Harrison, J., Spalding, M., & Lysenko, I. (2005). Measuring the extent and effectiveness of protected areas as an indicator for meeting global biodiversity targets. *Philosophical Transactions of the Royal Society B Biological Science*, *360*(1454), 443–455. doi: 10.1098/rstb.2004.1592. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15814356

Checchi, F., Stewart, B. T., Palmer, J. J., & Grundy, C. (2013). Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations. *International Journal of Health Geographics*, *12*, 4. doi: 10.1186/1476-072X-12-4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23343099

Cibulskis, R. E., Alonso, P., Aponte, J., Aregawi, M., Barrette, A., Bergeron, L., … Williams, R. (2016). Malaria: Global progress 2000-2015 and future challenges. *Infectious Diseases of Poverty*, *5*(1), 61. doi: 10.1186/s40249-016-0151-8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27282148

CIESIN (Center for International Earth Science Information Network, Columbia University). (2005). Gridded population of the world (GPW) Version 3. GPW and GRUMP: A brief background, comparison, and history. Retrieved from http://sedac.ciesin.columbia.edu/data/collection/gpw-v3/about-us

CIESIN (Center for International Earth Science Information Network, Columbia University). (2016a). *Gridded population of the world, Version 4 (GPWv4)* [Data set]. Retrieved March 2017, from http://dx.doi.org/10.7927/H4HX19NJ

CIESIN (Center for International Earth Science Information Network, Columbia University). (2016b). Documentation for the Gridded Population of the World, Version 4 (GPWv4). doi: 10.7927/H4D50JX4. Retrieved from http://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-documentation.pdf

CIESIN (Center for International Earth Science Information Network, Columbia University). (2016c). Gridded Population of the World, Version 4 (GPWv4): Data Quality Indicators. doi: 10.7927/H49C6VBN. Retrieved from http://beta.sedac.ciesin.columbia.edu/data/set/gpw-v4-data-quality-indicators

CIESIN (Center for International Earth Science Information Network, Columbia University). (2018). Gridded population of the world (GPW), v4. National identifier grid. Country-level information and sources revision 11. Retrieved from https://sedac.ciesin.columbia.edu/binaries/web/sedac/collections/gpw-v4/gpw-v4-country-level-summary-rev11.xlsx

CIESIN (Center for International Earth Science Information Network, Columbia University), and CIAT (Centro Internacional de Agricultura Tropical). (2005). *Gridded population of the world, Version 3 (GPWv3)* [Data set]. doi: 10.7927/H4XK8CG2

CIESIN (Center for International Earth Science Information Network, Columbia University), IPFRI (International Food Policy Research Institute), The World Bank, & CIAT (Centro Internacional de Agricultura Tropical). (2011). *Global Rural-Urban Mapping Project (GRUMPv1)* [Data set]. Retrieved March 2018, from http://dx.doi.org/10.7927/H4R20Z93

de Ferranti, J. (2017a). *Digital elevation data. Viewfinder panoramas* [Data set]. Retrieved January 2017, from http://www.viewfinderPanoramas.org/dem3.html

de Ferranti, J. (2017b). Digital elevation data: SRTM void fill. Viewfinder panoramas. Retrieved from http://www.viewfinderPanoramas.org/voidfill.html

de Sherbinin, A., & Adamo, S. B. (2015, October 5–6). *CIESIN's experience in mapping population and poverty*. Paper presented at the United Nations Expert Group Meeting on strengthening the demographic evidence base for the post-2015 Development Agenda, New York.

DLR (German Aerospace Center) EOC (Earth Observation Center). (2016). *GUF28 (Global Urban Footprint) v1*. Retrieved March 2017, from http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-11725/20508_read-47944/

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., & Worley, B. A. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, *66*(7), 849–857. Retrieved from https://www.asprs.org/wp-content/uploads/pers/2000journal/july/2000_jul_849-857.pdf

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., & Baptista, S. R. (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

Dudley, N. (2008). *Guidelines for applying protected area management categories* (p. 86). Retrieved from https://www.iucn.org/sites/dev/files/import/downloads/iucn_assignment_1.pdf

ESA (European Space Agency) CCI (Climate Change Initiative) - Land Cover project 2017. (2017a). *Land Cover CCI Product - Annual LC maps from 2000 to 2015 (v2.0.7)* [Data set]. Retrieved June 2017, from http://maps.elie.ucl.ac.be/CCI/viewer/

ESA (European Space Agency) CCI (Climate Change Initiative) - Land Cover project 2017. (2017b). *Land cover CCI product - MERIS Waterbody product v4.0 (150 m)* [Data set]. Retrieved June 2017, from http://maps.elie.ucl.ac.be/CCI/viewer/.

ESA (European Space Agency) CCI (Climate Change Initiative) - Land Cover project 2017. (2017c). Land cover CCI product user guide. Version 2. D3.3 CCI-LC-PUGV2. Retrieved from https://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

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

ESRI (Environmental Systems Research Institute). (2016). ArcGIS. Software (Version 10.3.1). Retrieved from http://www.esri.com/software/arcgis

ESRI (Environmental Systems Research Institute). (2018a). Raster Calculator, ArcMap 10.3. ArcGIS for desktop arcmap spatial analyst toolbox map algebra toolset. Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/raster-calculator.htm

ESRI (Environmental Systems Research Institute). (2018b). Nibble, ArcMap 10.3. ArcGIS for desktop arcmap spatial analyst toolbox generalization toolset. Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/nibble.htm

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. doi: 10.1002/joc.5086

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., & Mills, J. (2016, June 14–17). *Development of new open and free multi-temporal global population grids at 250m resolution*. Paper presented at the 19th AGILE Conference on Geographic Information Science, Helsinki, Finland.

Furieri, A. (2016). Spatialite Software. The Gaia-SINS federated projects home-page. Retrieved from http://www.gaia-gis.it/gaia-sins/

GADM (Global ADMinistrative Areas). (2015). *GADM* [Data set]. Retrieved March 2017, from http://gadm.org/index.html

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., … Tatem, A. J. (2016). Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Scientific Data*, *3*, 160005. doi: 10.1038/sdata.2016.5. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26881418

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PloS One*, *8*(2), e55882. doi: 10.1371/journal.pone.0055882. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23418469

GBD 2016. SDG Collaborators. (2017). Measuring progress and projecting attainment on the basis of past trends of the health-related Sustainable Development Goals in 188 countries: An analysis from the Global Burden of Disease Study 2016. *Lancet*, *390*, 1423–1459. doi: 10.1016/S0140-6736(17)32336-X

GHSL (Global Human Settlement Layer). (2015). *GHS Built-Up Grid (LDS)* [Data set]. Retrieved from: http://ghsl.jrc.ec.europa.eu/ghs_bu.php

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. (2013). How many volunteers does it take to map an area well? The validity of linus' law to volunteered geographic information. *The Cartographic Journal*, *47*(4), 315–322. doi: 10.1179/000870410x12911304958827

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15608702

Henderson, M., Yeh, E. T., Gong, P., Elvidge, C., & Baugh, K. (2003). Validation of urban boundaries derived from global night-time satellite imagery. *International Journal of Remote Sensing*, *24*(3), 595–609. doi: 10.1080/01431160304982

Hillson, R., Alejandre, J. D., Jacobsen, K. H., Ansumana, R., Bockarie, A. S., Bangura, U., … Stenger, D. A. (2014). Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: A case study of Bo city, Sierra Leone. *PloS One*, *9*(11), e112241. doi: 10.1371/journal.pone.0112241. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25398101

Hillson, R., Alejandre, J. D., Jacobsen, K. H., Ansumana, R., Bockarie, A. S., Bangura, U., … Stenger, D. A. (2015). Stratified sampling of neighborhood sections for population estimation: A case study of Bo City, Sierra Lleone. *PloS One*, *10*(7), e0132850. doi: 10.1371/journal.pone.0132850 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26177479

Hormann, C. (2018, August). Evaluating void filling data for SRTM DEMs. Geo-Visualization: Geodata reviews. Retrieved from http://www.imagico.de/pov/earth_srtm.php

IHME (Institute for Health Metrics and Evaluation). (2013). The global burden of disease: Generating evidence, guiding policy. Retrieved from http://www.healthdata.org/sites/default/files/files/policy_report/2013/GBD_GeneratingEvidence/IHME_GBD_GeneratingEvidence_FullReport.pdf

IHME (Institute for Health Metrics and Evaluation). (2016). Rethinking development and health: Findings from the global burden of disease study. Retrieved from http://www.healthdata.org/sites/default/files/files/policy_report/GBD/2016/IHME_GBD2015_report.pdf

ISO (International Organization for Standardization). (2017). Country Codes - ISO 3166. Retrieved from https://www.iso.org/iso-3166-country-codes.html.

Iwao, K., Nishida, K., Kinoshita, T., & Yamagata, Y. (2006). Validating land cover maps with Degree Confluence Project information. *Geophysical Research Letters*, *33*(23). doi:10.1029/2006gl027768

Langford, M., Higgs, G., Radcliffe, J., & White, S. (2008). Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems*, *32*(1), 66–80. doi: 10.1016/j.compenvurbsys.2007.06.001

Linard, C., Alegana, V. A., Noor, A. M., Snow, R. W., & Tatem, A. J. (2010). A high resolution spatial population database of Somalia for disease risk mapping. *International Journal of Health Geographics*, *9*, 45. doi: 10.1186/1476-072X-9-45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20840751

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PloS One*, *7*(2), e31743. doi: 10.1371/journal.pone.0031743. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22363717

Linard, C., Gilbert, M., & Tatem, A. J. (2011). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*, *76*(5), 525–538. doi: 10.1007/s10708-010-9364-8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23576839

Lloyd, C. T. (2017). High resolution global gridded data for use in population studies. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4/W2*, 117–120. doi: 10.5194/isprs-archives-XLII-4-W2-117-2017 Retrieved from https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-4-W2/117/2017/

Lloyd, C. T., Chamberlain, H., Kerr, D., & Bondarenko, M. (2018). Global Spatio-temporally Harmonised Datasets for Producing High-resolution Population Denominators. doi: 10.6084/m9.figshare.7291250.v1. Retrieved from https://doi.org/10.6084/m9.figshare.7291250.v1

Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. *Scientific Data*, *4*, 170001. doi: 10.1038/sdata.2017.1 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28140386

MAP (Malaria Atlas Project). (2017). *Prevalence rate of Plasmodium falciparum malaria for years 2000-2015* [Data set]. Retrieved March 2018, from http://www.map.ox.ac.uk

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, *55*(1), 31–42. doi: 10.1111/0033-0124.10042

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, *33*(3), 179–194. doi: 10.1559/152304006779077309

Min, B., Gaba, K. M., Sarr, O. F., & Agalassou, A. (2013). Detection of rural electrification in Africa using DMSP-OLS night lights imagery. *International Journal of Remote Sensing*, *34*(22), 8118–8141. doi: 10.1080/01431161.2013.833358

MPSMRM (Ministère du Plan et Suivi de la Mise en œuvre de la Révolution de la Modernité), MSP (Ministère de la Santé Publique), & ICF International. (2014). Enquête démographique et de santé en République Démocratique du Congo 2013–2014. Retrieved from https://www.unicef.org/drcongo/french/00_-_00_-_DRC_DHS_2013-2014_FINAL_PDF_09-29-2014.pdf

Muck, M., Klotz, M., & Taubenbock, H. (2017, March). *Validation of the DLR Global Urban Footprint in rural areas: Acase study for Burkina Faso.* Paper presented at the Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates.

Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., & Jaedicke, C. (2006). Global landslide and avalanche hotspots. *Landslides*, *3*(2), 159–173. doi: 10.1007/s10346-006-0036-1

Nieves, J. J., Sorichetta, A., Linard, C., Bondarenko, M., Steele, J., Stevens, F., … Tatem, A. J. (2018). Modelling Built-Settlements between Remotely-Sensed Observations. *Preprints*. doi:10.20944/preprints201812.0250.v2. Retrieved from https://www.preprints.org/manuscript/201812.0250/v2

Nieves, J. J., Stevens, F. R., Gaughan, A. E., Linard, C., Sorichetta, A., Hornby, G., … Tatem, A. J. (2017). Examining the correlates and drivers of human population distributions across low- and middle-income countries. *Journal of the Royal Society, Interfacethe Royal Society*, *14*(137), 20170401. doi: 10.1098/rsif.2017.0401. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/29237823

Norris, J., Dunning, C., & Malknecht, A. (2015). Fragile progress: The record of the millennium development goals in states affected by conflict. Retrieved from https://resourcentre.savethechildren.net/node/9355/pdf/fragilestates-report_web.pdf?embed=1

ORNL (Oak Ridge National Laboratory). (2010). LandScan Data Availability. Retrieved from http://web.ornl.gov/sci/landscan/landscan_data_avail.shtml

OSGF (Open Source Geo-spatial Foundation). (2017a). OSGEO4W. OSGEO4W geo-spatial software. Retrieved from http://trac.osgeo.org/osgeo4w/

OSGF (Open Source Geo-spatial Foundation). (2017b). GDAL - Geo-spatial Data Abstraction Library. GDAL. Retrieved from http://www.gdal.org/

OSMF (OpenStreetMap Foundation). (2018a). Tags. Open street map. Retrieved from https://wiki.openstreetmap.org/wiki/Tags

OSMF (OpenStreetMap Foundation). (2018b). Accuracy. Open street map. Retrieved from http://wiki.openstreetmap.org/wiki/Accuracy

OSMF (OpenStreetMap Foundation). (2018c). Contributors. Open street map. Retrieved from https://wiki.openstreetmap.org/wiki/Contributors

OSMF (OpenStreetMap Foundation) and Contributors. (2016). "OpenStreetMap (OSM) January 2016. Planet OSM. Retrieved from http://planet.openstreetmap.org/; http://www.openstreetmap.org; http://www.opendatacommons.org; http://www.creativecommons.org

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A. J., Freire, S., Halkia, M., … Syrris, V. (2016). Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. doi: 10.2788/253582. Retrieved from https://publications.europa.eu/en/publication-detail/-/publication/6eedd1fe-e046-11e5-8fea-01aa75ed71a1/language-en

Pesaresi, M., Ehrlich, D., Florczyk, A. J., Freire, S., Julea, A., Kemper, T., … Syrris, V. (2015). *GHS built-up grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014)* [Data set]. Retrieved March 2017, from http://data.europa.eu/89h/jrc-ghsl-ghs_built_ldsmt_globe_r2015b

PostGIS PSC (Project Steering Committee). (2016). About PostGIS. PostGIS, spatial and geographic objects for postgreSQL (Version 2.0). Retrieved from http://postgis.net/

PostGIS PSC (Project Steering Committee). (2017a). PostGIS 2.4.4dev Manual - Chapter 8. PostGIS reference - 8.9. Spatial relationships and measurements. Retrieved from https://postgis.net/docs/reference.html#Spatial_Relationships_Measurements

PostGIS PSC (Project Steering Committee). (2017b). PostGIS 2.4.4dev Manual - Chapter 8. PostGIS reference - 8.11. Geometry processing. Retrieved from https://postgis.net/docs/reference.html#Geometry_Processing

PostgreSQL Global Development Group. (2016). About. PostgreSQL (Version 9.1). Retrieved from http://www.postgresql.org/about/.

PSF (Python Software Foundation). (2016). Python 3.6.0. Retrieved from https://www.python.org/downloads/release/python-360/

QGIS Project. (2017). QGIS. Downloads (Version 2.18.4). Retrieved from http://download.osgeo.org/qgis/

Rabus, B., Eineder, M., Roth, A., & Bamler, R. (2003). The shuttle radar topography mission—A new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, *57*(4), 241–262. doi: 10.1016/s0924-2716(02)00124-7

Rodríguez, E., Morris, C. S., Belz, J. E., Chapin, E. C., Martin, J. M., Daffer, W., & Hensley, S. (2005). An assessment of the SRTM topographic products. D-31639. Retrieved from https://www2.jpl.nasa.gov/srtm/SRTM_D31639.pdf

SAGA (System for Automated Geoscientific Analyses). (2017). Downloads (Version 4.1.0). Retrieved from http://www.saga-gis.org/en/index.html.

Santini, M., Taramelli, A., & Sorichetta, A. (2010). ASPHAA: A GIS-based algorithm to calculate cell area on a latitude-longitude (Geographic) Regular Grid. *Transactions in Gis*, *14*(3), 351–377. doi: 10.1111/j.1467-9671.2010.01200.x

Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., & Hay, S. I. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, *434*(7030), 214–217. doi: 10.1038/nature03342. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15759000

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data*, *2*, 150045. doi: 10.1038/sdata.2015.45. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26347245

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (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. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25689585

Taramelli, A., Melelli, L., Pasqui, M., & Sorichetta, A. (2010). Modelling risk hurricane elements in potentially affected areas by a GIS system. *Geomatics, Natural Hazards and Risk*, *1*(4), 349–373. doi: 10.1080/19475705.2010.532972

Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, *4*, 170004. doi: 10.1038/sdata.2017.4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28140397

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), e1298. doi: 10.1371/journal.pone.0001298. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18074022

Themnér, L., & Wallensteen, P. (2012). Armed Conflicts, 1946–2011. *Journal of Peace Research*, *49*(4), 565–575. doi: 10.1177/0022343312452421

Tomás, L., Fonseca, L., Almeida, C., Leonardi, F., & Pereira, M. (2015). Urban population estimation based on residential buildings volume using IKONOS-2 images and lidar data. *International Journal of Remote Sensing*, *37*(sup1), 1–28. doi: 10.1080/01431161.2015.1121301

UN (United Nations) Department of Economic and Social Affairs - Statistics Division. (2018). Methodology - Standard country or area codes for statistical use (M49) - geographic regions. United Nations. Retrieved from https://unstats.un.org/unsd/methodology/m49/

UN (United Nations) General Assembly. (2000). United nations millennium declaration. In *Resolution A/RES/55/2 Clause 19*. Retrieved from https://www.un.org/millennium/declaration/ares552e.htm

UN (United Nations) General Assembly. (2014). A/68/970 report of the open working group of the general assembly on sustainable development goals. Retrieved from http://www.un.org/ga/search/view_doc.asp?symbol=A/68/970&Lang=E

UN (United Nations) General Assembly. (2015). Transforming our world: The 2030 Agenda for Sustainable Development. In *Resolution A/RES/70/1* (pp. 12, 13, 27, 28, 32). Retrieved from https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf

UNEP-WCMC (United Nations Environment Programme-World Conservation Monitoring Centre). (2017). World database on protected areas user manual 1.5. Retrieved from http://wcmc.io/WDPA_Manual

UNEP-WCMC (United Nations Environment Programme-World Conservation Monitoring Centre) and IUCN (International Union for Conservation of Nature). (2017). *WDPA (World database on protected areas)/GD-PAME (Global database on protected areas management effectiveness)* [Data set]. Retrieved June 2017, from www.protectedplanet.net

US NASA (National Aeronautics and Space Administration). (2016). *Shuttle Radar Topography Mission. Jet Propulsion Laboratory. California Institute of Technology* [Data set]. Retrieved March 2017, from http://www2.jpl.nasa.gov/srtm/

US NOAA (National Oceanic and Atmospheric Administration) National Centers for Environmental Information. (2017). *VIIRS DNB cloud free composites. 2012-2016. Version 1 nighttimeday/night band composites* [Data set]. Retrieved March 2017, from https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

US NOAA (National Oceanic and Atmospheric Administration) National Geophysical Data Center/US Air Force Weather Agency. (2015). *Version 4 DMSP-OLS nighttime lights time series (1992–2013; Average visible, stable lights, and cloud free coverages)* [Data set]. Retrieved March 2017, from https://www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

Varga, M., & Bašić, T. (2015). Accuracy validation and comparison of global digital elevation models over Croatia. *International Journal of Remote Sensing*, *36*(1), 170–189. doi: 10.1080/01431161.2014.994720

Visconti, P., Di Marco, M., Alvarez-Romero, J. G., Januchowski-Hartley, S. R., Pressey, R. L., Weeks, R., & Rondinini, C. (2013). Effects of errors and gaps in spatial data sets on assessment of conservation progress. *Conservation Biology: The Journal of the Society for Conservation Biology*, *27*(5), 1000–1010. doi: 10.1111/cobi.12095. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23869663

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/29555739

Weber, E. M., Seaman, V. Y., Stewart, R. N., Bird, T. J., Tatem, A. J., McKee, J. J., … Reith, A. E. (2018). Census-independent population mapping in northern Nigeria. *Remote Sensing of Environment*, *204*, 786–798. doi: 10.1016/j.rse.2017.09.024. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29302127

WHO (World Health Organisation). (2015). World Malaria Report 2015. Retrieved from http://www.who.int/malaria/publications/world-malaria-report-2015/report/en/

WHO (World Health Organisation). (2017). World Malaria Report 2017. Retrieved from http://www.who.int/malaria/publications/world-malaria-report-2017/report/en/

WorldPop (www.worldpop.org) School of Geography and Environmental Science - University of Southampton, Department of Geography and Geosciences - University of Louisville, Département de Géographie - Université de Namur, & Center for International Earth Science Information Network (CIESIN) - Columbia University. (2018). *Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076)* [Data set]. Retrieved June 2019, from https://doi.org/10.5258/SOTON/WP00650

Zhang, Q. L., Pandey, B., & Seto, K. C. (2016). A robust method to generate a consistent time series from DMSP/OLS nighttime light data. *IEEE Transactions on Geoscience and Remote Sensing*, *54* (10), 5821–5831. doi: 10.1109/Tgrs.2016.2572724