HIGH RESOLUTION GLOBAL GRIDDED DATA FOR USE IN POPULATION STUDIES

C. T. Lloyd

WorldPop, Geography and Environment, University of Southampton, Highfield Campus, Southampton. SO17 1BJ. UK - C.T.Lloyd@soton.ac.uk

Commission IV, WG IV/4

KEY WORDS: WorldPop, human population, open access archive, high resolution, global, spatial dataset

ABSTRACT:

Open access geospatial data represent a range of metrics relevant to global human population mapping at fine spatial scales. Detailed and contemporary spatial datasets that accurately describe population distributions are vital in order to measure impacts of population growth, monitor change, and plan interventions. To construct such datasets the harmonisation of geospatial data layers is a prerequisite because layer specifications differ widely.

To this end the WorldPop Project is producing an open access archive of 3 and 30 arc-second (~100m and ~1km, respectively) resolution gridded data in a predominantly open source environment, using OSGEO4W utilities. Five tiled raster datasets form the basis of the archive: (i) Viewfinder Panoramas topography clipped to CIESIN national level administrative coastlines; (ii) a matching ISO 3166 country identification grid; (iii) country area; (iv) pixel area; (v) and slope layer. Further layers will include transport networks, landcover, urban extent, nightlights, climate, travel time to major cities, forest stand change, livestock densities, vegetation indices, and waterways. We here describe the base datasets and the production methodology in development. The alpha version of the archive can be downloaded both from the WorldPop Dataverse Repository and the WorldPop Project website. The improved and expanded beta version of the archive is in development for release next year, and will offer significantly improved standardisation of country boundaries, and inland water boundaries (forthcoming), to global census unit data.

1. INTRODUCTION

Detailed and contemporary spatial datasets that accurately describe human population distribution can support the measurement of the impacts of population growth, the monitoring of changes, environmental and health applications, and the planning of interventions (Tatem, Noor, von Hagen, Di Gregorio, & Hay, 2007). Spatial databases of human population have found use in disease burden estimation, epidemic modelling, resource allocation, disaster management, accessibility modelling, transport and city planning, poverty mapping and environmental impact assessment amongst others (Balk et al., 2006; Bhaduri, Bright, Coleman, & Dobson, 2002; Salvatore, Pozzi, Ataman, Huddleston, & Bloise, 2005; Hay, Guerra, Tatem, Atkinson, & Snow, 2005; Snow, Guerra, Noor, Myint, & Hay, 2005).

Previous population mapping work (Tatem et al., 2007; Gaughan, Stevens, Linard, Jia, & Tatem, 2013; Linard, Gilbert, Snow, Noor, & Tatem, 2012; Bhaduri, Bright, Coleman, & Urban, 2007; Dobson, Bright, Coleman, & Worley, 2000; Azar et al., 2010; Stevens, Gaughan, Linard, & Tatem, 2015; Alegana et al., 2015; Deville et al., 2014) has shown that incorporating multiple spatial datasets into population mapping approaches can improve accuracy. Consequently, to support population mapping applications in the future there is a need for standardised grid definitions, standardised (contiguous) country boundaries and coastlines, and covariate layers representing different time periods that match these and that are regularly updated - all created at fine spatial resolutions (Lloyd, Sorichetta, & Tatem, 2017). To begin to meet such needs, the WorldPop Project is producing a beta version open access archive of 3 and 30 arc-second spatial resolution gridded datasets. The 3 arc-second datasets are tiled (hereafter referred to as 100m tiled datasets). The 30 arc-second datasets are produced as global grids. A predominantly open source production environment is utilised, and a semi-automated workflow.

We here describe the five 'base' standardised 100m tiled datasets that have been generated, as well as the production methodology used to create them. Further, we summarise some of the additional layers that are being incorporated into the archive, to be used to construct covariates for population modelling. The methodology for base and further datasets will be explained in full in a forthcoming paper on the topic.

2. METHODS

Five 100m resolution datasets form the basis of the archive outlined here: topography, standardised, gridded, and clipped to country coastal boundaries; a slope layer derived from the topography; a country identification (ID) grid (to the ISO 3166 standard (ISO, 2015)); and a pixel area (m^2) and a country area (km^2) grid derived from this. The base topography, slope, country id, and country area grids are supplied as 100m tiles and 1 km resolution derivatives, the latter for convenience. Pixel area tiles are supplied as 100m tiles only. Additional spatial data layers that are incorporated into the archive are similarly standardised to match the grid definition and coastlines of the country ID base grid (Lloyd et al., 2017).

Country ID, country area, and pixel area grids provide essential basic metrics upon which to build population analyses. The construction of the slope layer is useful to human population studies because population densities tend to be much lower on steep slopes. Similarly, the construction of the topography layer is useful because population densities tend to be lowest at the highest elevations (Lloyd et al., 2017). Moreover, population densities are all related to landcover, infrastructure and climatic regimes, and therefore the additional layers are also potentially valuable inputs as covariates to population modelling efforts (Lloyd et al., 2017).

2.1 Source base datasets and archive formatting

The topography data consists of the Viewfinder Panoramas dataset (de Ferranti, 2015a), which is primarily US NASA Shuttle Radar Topography Mission (SRTM) data (US NASA, 2015) collected in the year 2000, with amendment and correction by the dataset developer, Jonathan de Ferranti (de Ferranti, 2015a). The country boundaries used are the national level 0 administrative boundary data produced by the Center for International Earth Science Information Network (CIESIN) (G. Yetman, personal communication, December 2016).

Viewfinder Panoramas data are provided as raster tiles in hgt format. Hgt is the raw SRTM digital elevation model data file format (US NASA, 2015). The CIESIN country ID data are provided as raster tiles in geo-tiff format, gridded to Viewfinder Panoramas tile extents, with cell values corresponding to the three digit numerical ISO 3166 country code standard. Both have 3 arc-second horizontal resolution. The former has 1m vertical resolution.

Viewfinder Panoramas tiles 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, 2015a; 2015b). Alternative sources are topographic maps, Landsat images, and ASTER GDEM data (de Ferranti, 2015a). These sources are much more accurate than simple interpolation of SRTM data (de Ferranti, 2015a). CIESIN country boundaries ideally follow population censuses, where these align between countries, but are otherwise aligned to a global framework based on the Global Administrative Areas version 2 (GADMv2). The former is preferable as census boundary data are usually highly accurate, whilst the latter is utilised selectively as an alternative because it is publically available and widely used in the research community (CIESIN, 2016).

Viewfinder Panoramas and CIESIN data are provided as 1,201 × 1,201 pixel tiles with frequent but irregular one pixel tile overlap, in geographical coordinate system (GCS) with WGS 1984 datum (EPSG:4326). These characteristics are maintained in output datasets, which otherwise utilise as efficient a data type/depth as is commensurate with the numerical values inherent in each dataset, with nodata taking the maximum value possible for the data type/depth, in geo-tiff format, with a 3 arc-second (i.e. 0.00083333333 decimal degree) cell size.

2.2 Incorporating further spatial datasets into archive

Further spatial datasets are currently being incorporated into the archive: Open Street Map (OSM) waterways, highways, railway network, railway stations, and airports (OSMF & Contributors, 2016); DMSP nightlights v4 1992–2013 (US NOAA, 2014); ESA CCI landcover 1.6.1 (ESA CCI, 2016); Travel time to major cities 2000 (Nelson, 2008); MERIS water bodies (G. Yetman, personal communication, March 2017); ViiRS nightlights 2012-2016 (US NOAA, 2017); Global Urban Footprint (DLR EOC, 2017); and Global Human Settlement Layer 1990-2014 (European Commission JRC, 2016); with additional datasets to follow.

2.3 Data processing software

Open source OSGEO4W64 Geospatial Software (OSGF, 2015a), the included Geospatial Data Abstraction Library (GDAL) v2.1.3 package (OSGF, 2015b), and (on occasion) proprietary ESRI ArcMap v10.3.1 and ArcInfo Workstation v9.3 GIS software (ESRI, 2016) are employed to produce output datasets, using a Microsoft Windows 7, 64 bit operating system (OS). The GDAL package is predominantly used due to better handling of large raster datasets and more effective semi-automation of workflow, and is preferred unless specific algorithms or functions are unavailable or significantly harder to implement than in the commercial proprietary software. The ‘nibble’ ArcGIS tool (ESRI, 2017) is an example of an algorithm of which functions are not wholly available elsewhere. The nibble tool is to be used in the production of some further spatial datasets. Further, ESRI proprietary software are employed in the first part of the country area gridding workflow (for practical reasons described later) but are otherwise not utilised in the production of base grids.

The processing of OSM data requires additional software. An Ubuntu Linux OS (14.04 LTS, Trusty Tahr) installation is utilised, with PostgreSQL 9.1 (PostgreSQL GDG, 2016) and PostGIS 2.0 (PostGIS PSC, 2016) database software from which spatial relational data can be exported. Osm2pgsql (OSMF, 2016) is an OSM specific software that is used to load OSM data into databases (Lloyd et al, 2017). Subsequent database access, processing, and filtering (on the Windows platform) is 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. QGIS, GDAL, and SAGA GIS 4.1.0 (SAGA, 2017) software are used to extract database attributes, and to convert to raster format for subsequent tiling and mosaicking as consistent with the workflow for other datasets.

The alpha version of the Archive (Lloyd et al, 2017) has been produced using a methodology that more frequently employs proprietary software (i.e. ESRI ArcMap). For production of the beta and subsequent versions there is an ongoing effort to redesign workflows to migrate as fully as possible to open source alternatives. This is for reasons of both better automation in batch scripting and for greater efficiency. As GDAL (in particular) and other open source software develops so this migration becomes more viable.

2.4 Production of base 100m tiled datasets

Program code is implemented as windows batch files within OSGEO4W64. Individual tile extent is specifically defined where necessary within relevant programing loops throughout the workflow, in order to ensure that consistent base tile extents are maintained and so that end products are aligned (Lloyd et al., 2017).

All Viewfinder Panoramas topography tiles are first batch converted to geo-tiff format for ease of processing in GDAL. A virtual raster table (VRT) is produced, and topography tiles mosaicked into one global image using gdalwarp. A calculation is performed to remove the few erroneous elevation data values that fall outside of the plausible global range of elevation (i.e. -450m to 8900m). To remove nodata pixels at coastal edges (due to inconsistencies in coastline location between topography and country data) and within the continental interior (due to the previous operation, or pre-existing voids), a calculation is made to produce a mask of the area to fill, using a mosaicked and reclassified global country ID grid. The gdal_fillnodata utility is then implemented. The topography data is clipped to CIESIN coastlines using a reclassified global country ID grid. Interpolation artefacts (-32768 value), generated by the fillnodata utility where there is a large expanse of zero values in the source data, are changed to zero values; and inland zero elevation values are reasserted where unnecessary but unavoidable inland interpolation of these values has taken place during the fillnodata operation. The global 100m resolution topography layer is then tiled.

A global slope layer at 100m resolution is created from the topography 100m mosaic using the gdaldem utility, prior to assertion of the correct nodata value and data type. The slope layer is then split into tiles (Lloyd et al., 2017).

To create the country 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). Prior to the calculation, country ID tiles are converted to ESRI grid raster format, aggregated (ESRI, 2015a) to 60 arc-second resolution, and mosaicked into a global grid. The AML script is run on the global grid. A calculation is run on the resulting cell area grid to convert cell area values within each 60 arc-second cell to that for a 3 arc-second cell size (the area of each cell is divided by 400). The grid is then resampled (using 'nearest neighbour' method in order to maintain cell values) to 3 arc-second cell size (Lloyd et al., 2017). This produces the pixel area global grid. GDAL is used to convert the grid to geo-tiff, clip to coastlines using a reclassified country ID grid, and tile the data. Subsequently, a zonal statistic (sum) calculation (ESRI, 2015b) is performed on the resampled output of the AML script (using ArcMap) with the global country ID layer as zone indicator. This creates a global output layer that expresses country area. After the adjustment of cell values from metres to kilometres the grid is tiled using GDAL (Lloyd et al., 2017).

3. CONCLUSION

The forthcoming beta version of the WorldPop Archive is a refinement of the previous alpha version in large part due to the new CIESIN country ID base gridding, which offers significantly improved standardisation of country boundaries (and inland water boundaries – forthcoming) to global census unit data. Further, the beta will offer a substantially expanded range of datasets that it is hoped will assist researchers by providing a uniform base upon which analysis of population distributions can be performed. Such analysis will in turn allow measurement of the impacts of population growth, the monitoring of changes, environmental and health applications, and the planning of interventions (Tatem et al., 2007).

4. ACKNOWLEDGEMENTS

C.T.L is supported by funding acquired by Prof. Andrew J. Tatem from the Bill & Melinda Gates Foundation (OPP1134076). This work forms part of the outputs of WorldPop (www.worldpop.org) and the Flowminder Foundation (www.flowminder.org). A.J.T is further supported by funding from NIH/NIAID (U19AI089674), the Bill & Melinda Gates Foundation (OPP1106427, 1032350, OPP1094793), the Clinton Health Access Initiative, National Institutes of Health, and a Wellcome Trust Sustaining Health Grant (106866/Z/15/Z). The funders had no role in study design, data collection and analysis, decision to publish, and preparation of the manuscript.

5. REFERENCES

Alegana, V. A. et al., 2015. Fine resolution mapping of population age-structures for health and development applications. *J. R. Soc. Interface*, 12, 20150073.

Azar, D. et al., 2010. Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. *Int. J. Remote Sens.*, 31, pp. 5635–5655.

Balk, D. L. et al., 2006. Determining global population distribution: methods, applications and data. *Adv. in Parasitology*, 62, pp. 119–156.

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

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

Center for International Earth Science Information Network (CIESIN), Columbia University, 2016. Documentation for the Gridded Population of the World, Version 4 (GPWv4). *US NASA Socioeconomic Data and Applications Center (SEDAC).* Retrieved from http://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-documentation.pdf (December 2016)

de Ferranti, J., 2015a. Digital Elevation Data. *Viewfinder Panoramas.* Retrieved from http://www.viewfinderPanoramas.org/dem3.html (November 2015)

de Ferranti, J., 2015b. Digital Elevation Data: SRTM Void Fill. *Viewfinder Panoramas*. Retrieved from http://www.viewfinderPanoramas.org/voidfill.html (November 2015)

Deville, P. et al., 2014. Dynamic population mapping using mobile phone data. *Proc. Natl Acad. Sci.*, 111, pp. 888–893.

DLR Earth Observation Center, 2017. Global Urban Footprint. *GUF Data and Access.* Retrieved from http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-11725/20508_read-47944/ (January 2017)

Dobson, J. E., Bright, E. A., Coleman, P. R. & Worley, B. A., 2000. LandScan: A global population database for estimating populations at risk. *Photogramm Eng. Remote Sens.*, 66, pp. 849–857.

ESRI, 2015a. Aggregate, 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/aggregate.htm (November 2015)

ESRI, 2015b. Zonal Statistics, ArcMap 10.3. *ArcGIS for Desktop ArcMap Spatial Analyst Toolbox Zonal Toolset.* Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/zonal-statistics.htm (November 2015)

ESRI, 2016. ArcGIS. *Software* (Version 10.3.1). Available from http://www.esri.com/software/arcgis (December 2016)

ESRI, 2017. 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 (May 2017)

European Commission Joint Research Centre, 2016. Global Human Settlement. GHS Built-Up Grid. *Download.* Retrieved

from http://ghslsys.jrc.ec.europa.eu/ghs_bu.php (November 2016)

European Space Agency (ESA) Climate Change Initiative (CCI), 2016. Global Land Cover v1.6.1. *Land Cover.* Retrieved from https://www.esa-landcover-cci.org/?q=node/169 (November 2016)

Furieri, A., 2016. Spatialite Software. *The Gaia-SINS Federated Projects Home-page*. Available from http://www.gaia-gis.it/gaia-sins/ (July 2016)

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, e55882.

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 Rev. Microbio.,* 3, pp. 81–90.

International Organization for Standardization (ISO), 2015. Country Codes—ISO 3166. *ISO Standards*. Retrieved from http://www.iso.org/iso/country_codes.html (November 2015)

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, e31743.

Lloyd, C. T., Sorichetta, A., & Tatem, A., 2017. High resolution global gridded data for use in population studies. *Scientific Data,* 4, 170001, pp. 1-17. http://dx.doi.org/10.1038/sdata.2017.1

Nelson, A. (European Commission Joint Research Centre Global Environment Monitoring Unit), 2008. Estimated Travel Time To The Nearest City Of 50,000 Or More People In Year 2000. *Joint Research Centre—The European Commission's In-house Science Service—Travel time to major cities: A global map of Accessibility.* Retrieved from http://forobs.jrc.ec.europa.eu/products/gam/ (October 2016)

OpenStreetMap Foundation (OSMF) & 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 (June 2016)

OpenStreetMap Foundation (OSMF), 2016. Osm2pgsql. *Open Street Map.* Retrieved from http://wiki.openstreetmap.org/wiki/Osm2pgsql (June 2016)

Open Source Geospatial Foundation, 2015a. OSGEO4W. *OSGEO4W Geospatial Software*. Available from http://trac.osgeo.org/osgeo4w/ (November 2015)

Open Source Geospatial Foundation, 2015b. GDAL—Geospatial Data Abstraction Library. *GDAL*. Retrieved from http://www.gdal.org/ (November 2015)

PostGIS Project Steering Committee (PSC), 2016. About PostGIS. *PostGIS, Spatial and Geographic Objects for PostgreSQL* (Version 2.0). Available from http://postgis.net/ (June 2016)

PostgreSQL Global Development Group, 2016. About. *PostgreSQL* (Version 9.1). Available from http://www.postgresql.org/about/ (June 2016)

QGIS project, 2017. QGIS. *Downloads* (Version 2.18.4). Available from http://download.osgeo.org/qgis/windows/ (May 2017)

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

Salvatore, M., Pozzi, F., Ataman, E., Huddleston, B. & Bloise, M., 2005. Mapping global urban and rural population distributions. Environment and Natural Resources Working Paper 24. *Food and Agri. Org. UN Corporate Document Repository.* Retrieved from http://www.fao.org/docrep/009/a0310e/a0310e00.htm (June 2016)

Santini, M., Taramelli, A. & Sorichetta, A., 2010. ASPHAA: A GIS-based algorithm to calculate cell area on a latitude-longitude (geographic) regular grid. *Trans. in GIS,* 14, pp. 351–377.

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, pp. 214–217.

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, e0107042.

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, e1298.

US NASA, 2015. Shuttle Radar Topography Mission. *Jet Propulsion Laboratory. California Institute of Technology*. Retrieved from http://www2.jpl.nasa.gov/srtm/ (December 2015)

US NOAA National Centers for Environmental Information, 2017. VIIRS DNB Cloud Free Composites. Version 1 Nighttime VIIRS Day/Night Band Composites. *Earth Observation Group*. Retrieved from https://ngdc.noaa.gov/eog/viirs/download_monthly.html (January 2017)

US NOAA National Geophysical Data Center/US Air Force Weather Agency, 2014. Version 4 DMSP-OLS Nighttime Lights Time Series (1992–2013; Average Visible, Stable Lights, & Cloud Free Coverages). *Earth Observation Group*. Retrieved from http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html (December 2015)