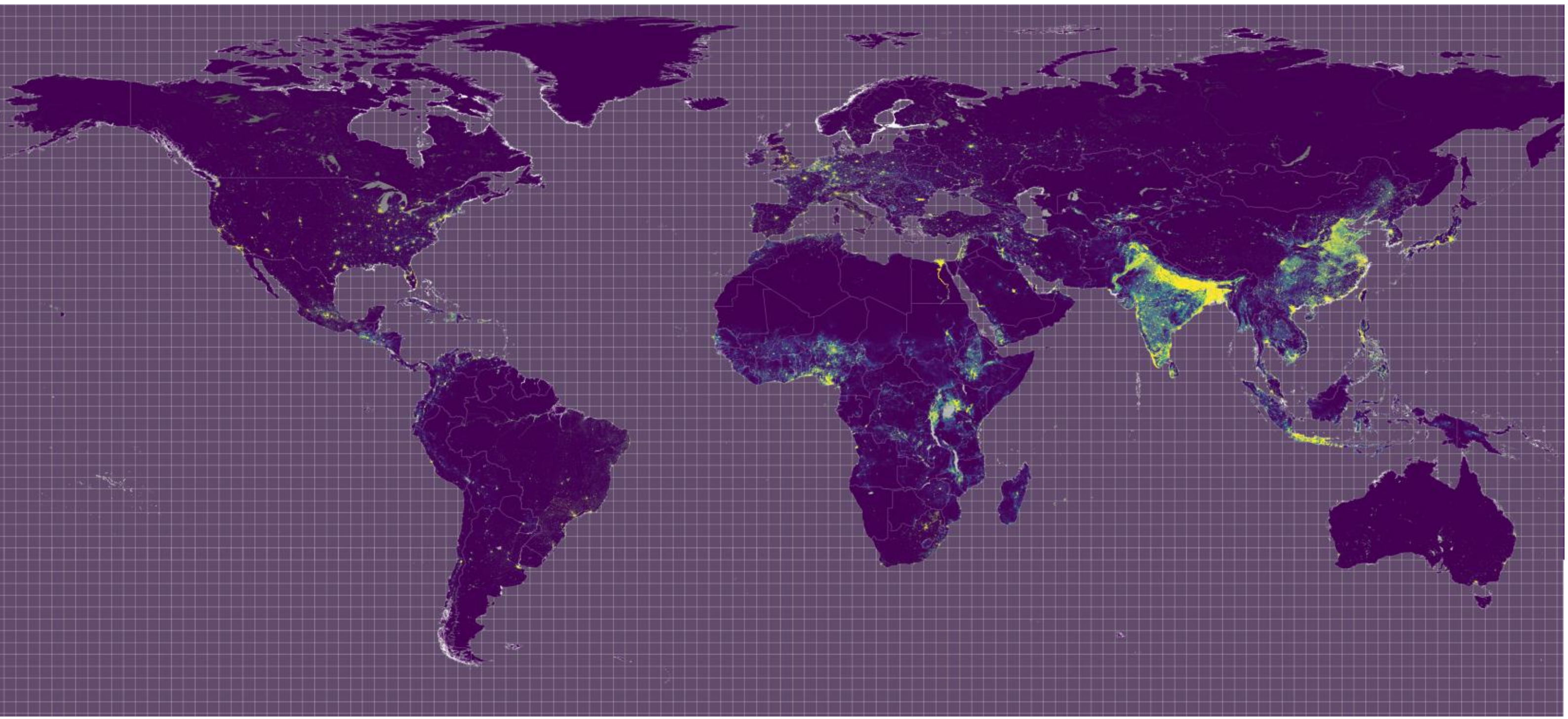
Global Demographic Data

Public release R2024B V1

worldpop.org
2025



This document outlines in brief how WorldPop's Global 2015-2030 gridded demographic datasets were produced. Please note that these are currently beta version datasets. Reviews, validation and testing are ongoing before production of final data. The construction of these data involve the assembly of millions of pieces of data, making it impossible to check every grid square estimate and we welcome feedback from the community where errors or anomalies are seen. This document will also remain live at present, with changes documented as they are made to the datasets, and new summaries or updates added as and when they occur. Updated versions of this document will be reflected in the version number.

These data are an open-access set of high-resolution gridded population datasets for 242 countries globally, providing age and sex-structured annual estimates for the period 2015-2030. These new 2015-2030 global demographic datasets exhibit marked improvements on the global datasets constructed in 2018 that represented the 2000-2020 period. Notably, refinements include the incorporation of the latest circa-2020 round of censuses, an updated and expanded geospatial covariate library, a new and improved inland water mask, improved approaches to projecting population numbers and age-sex breakdowns, alignment of national totals with the latest UN estimates, in addition to updated settlement mapping and growth modelling that incorporate building footprints mapped from satellite imagery.

These new global demographic datasets were constructed by WorldPop (www.worldpop.org), an interdisciplinary applied research programme that develops peer-reviewed research and replicable methods for the construction of open and high-resolution geospatial data on population distributions, demographics and dynamics. WorldPop produces many different forms of gridded population data, and readers are encouraged to review this page to be sure that these global data are the best for their needs before proceeding: https://www.worldpop.org/choosing-the-right-worldpop-population-data-for-you/.

Please note that this is a beta version dataset. Reviews, validation and testing are ongoing before production of a final dataset. The construction of these data involve the assembly of millions of pieces of data, making it impossible to check every grid square estimate and we welcome feedback from the community where errors or anomalies are seen.

For Pakistan datasets, we have been alerted to areas of unrealistic zero population estimates in the north of the country and are working to fix this as soon as possible.

Inputs

A 'top-down' method (https://www.worldpop.org/methods/populations/) to population disaggregation via a random forest (RF) dasymetric approach was implemented to construct this set of gridded population datasets (Stevens *et al.*, 2007). This 'top-down' RF method involves the identification, gathering and harmonisation of four principle data collections, each of which is described below.

Population data and administrative boundaries

Subnational census-based and official projection population counts and age/sex breakdowns at as small an administrative unit level as available at the time of data assembly were collected, alongside matching official administrative unit boundaries for each country. Population and boundary data were gathered for the two most-recent census rounds (c. 2010 and c. 2020). Where official census-based data were unavailable, population data from alternative sources were adopted. Predominantly, these alternative data were used for countries in conflict or where censuses had not been undertaken for more than a decade, and were sourced from either: (i) U.S. Census Bureau subnational population projections, or (ii) United Nations Common Operational Datasets (CODs) on population statistics. Both data sources are constructed to support humanitarian relief, health, development and disaster planning. Users should be aware that for many resource-poor countries, WorldPop co-develops 'bottom-up' gridded population estimates with governments and UN agencies, and these data may be more appropriate to user needs than the data outlined here – this page can help with making this decision: https://www.worldpop.org/choosing-the-right-worldpop-population-data-for-you/.

Population data for each country were processed to match corresponding official digital administrative boundary datasets. Administrative boundary data and population were further harmonised via a process involving the "nesting" of administrative units; this step aggregates sub-national administrative units and their corresponding population counts to ensure uniformity of administrative unit boundaries between timepoints for each country. These data were recoded adding a "GID" primary key field; this field enables the datasets to be joined.

For each country, the two timepoints of population data were interpolated and projected, providing a full set of annual population projections for the period 2010-2030. These projections and interpolations were undertaken using different types of approaches, depending on the data situation for each country. The projections were aligned at a national level with the estimated country totals set out in the 2024 Revision of the World Population Prospects produced by the Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat (UN, 2024). Population data were not modelled for nine countries or territories, due to classification as uninhabited or with static population totals (Table 1). Full census data metadata can be accessed via: https://data.worldpop.org/repo/prj/Global_2015_2030/R2024B/doc/census/global2_census_data_sources_v1.xlsx.

ISOAlpha	Country or Territory Name
ATF	Kerguelen Islands
BVT	Bouvet Island
CPT	Clipperton Island
HMD	Heard Island and McDonald Islands
IOT	British Indian Ocean Territory
SGS	South Georgia and the Sandwich Islands
SPR	Spratly Islands
UMI	Baker Island
VAT	Vatican City

Table 1. Countries or territories that were not modelled due to classification as uninhabited or with static population totals.

Mastergrid

Prior to the harmonisation of covariates and national administrative unit boundaries, a base grid was constructed. This base grid, known as the "mastergrid", is used for spatial alignment and resolution during harmonisation of all geospatial datasets used in the construction of the modelled population estimates. The mastergrid is constructed of a grid of 3 arc-seconds resolution (~100m at the equator) covering all major landmasses between 84°N and 60°S. Terrestrial areas are integrated via the aggregation of all land cover classes from ESA WorldCover 10m v200 (Zanaga *et al.*, 2022) and conversion to 100m output resolution; coastline pixels with greater than 75% water share are omitted in this stage. Oceans and major inland waterbodies, such as the Black Sea and Caspian Sea, are recoded as NoData.

The terrestrial surface area is further split into constituent countries using the Large Scale International Boundaries (LSIB) Dataset v11.3 (OGGI, 2024), which reflects U.S. Government policy on international boundary alignment, political recognition and dispute status. The boundaries form 41 ambiguous small areas which are not clearly assigned to a country. In these cases, Global Mosaiced National Boundaries produced by WorldPop (WorldPop, 2018), the GPWv4 National Identifier Grid (CIESIN 2016) and boundaries related to the latest national census were used as source for delineation of countries. Southern Asia contributed the largest proportion of these small areas by sub-region. Finally, each country was given a unique three-digit numerical code in line with ISO-3166.

Covariates

Human population density is known to be highly influenced and correlated with a variety of environmental and physical phenomena, each known to influence the spatial distribution of population presence and densities (Nieves et al., 2017, Dobson et al., 2000, Nagle et al., 2014). A set of geospatial datasets representing these phenomena were assembled for use as covariates in modelling population distributions at high spatial resolution.

These factors are classified into two distinct categories: firstly, continuous variables such as topographic elevation and slope (Cohen & Small, 1998, Schumacher et al., 2000), climatic variables (Small & Cohen, 2004) and intensity of night-time lights (Briggs et al., 2007, Stathakis & Baltas, 2018). Secondly, categorical variables, including land cover type (Gaughan *et al.*, 2013, Linard *et al.*, 2011) and the presence or absence of settlements and urban areas (Tatem *et al.*, 2007), roads (Tatem *et al.*, 2007), waterbodies and waterways (Kummu et al., 2011), and protected areas (McDonald *et al.*, 2009). Therefore, a collection of the most up-to-date (at the time of production), global geospatial raster and vector datasets available were identified, gathered and processed into a uniform set of default covariates used for model fitting and prediction (Table 2). All source datasets are open-access, providing near global coverage at sufficient spatial and temporal scales.

To generate a uniform set of default geospatial covariates used for model fitting and prediction, the raw datasets as available from data providers were cleaned and converted to ensure format accessibility for further spatial processing. All raster datasets were reprojected to WGS1984 datum, resampled to 3 arc-second resolution and harmonised according to the global mastergrid. Where vector datasets were obtained, features were rasterised prior to reprojection, through resampling and harmonisation. Categorical variables, i.e. those indicating presence or absence or a phenomena such as land cover class, were converted into binary raster covariates, and subsequently processed to generate continuous "distance to" raster covariates, which typically result in more accurate outputs in population modelling than categorical data. Full covariate metadata can be accessed via: https://data.worldpop.org/repo/prj/Global_2015_2030/R2024B/doc/covariates/global2_covariates_metadata_v1.xlsx.

Built settlement growth model

A built settlement growth model was developed for input into the RF model, incorporating data from several sources. Global Human Settlement Layer (GHSL) data were adopted as the base for the model, specifically the BUILT-S (Pesaresi & Politis, 2023) and BUILT-V (Pesaresi & Politis, 2023) products. These two products were gathered at the available temporal resolution of 5-year intervals for the study period: 2015, 2020, 2025 and 2030; both products are available at 100m spatial resolution. Secondarily, two vector datasets depicting building footprints were collected. Google Open Buildings

V3 Polygons (Sirko, 2021) and Microsoft Building Footprints (Microsoft, 2023) were collected, rasterised and combined with the GHSL base settlement layer to enable refined mapping of buildings where these were missed in GHSL.

An integrated approach of random forest and dasymetric modelling with subnational data was used to produce annual 100m resolution binary built settlement datasets from these three input data sources (similar to Nieves *et al.*, 2020). Moreover, a historical high-resolution binary settlement mask was implemented within this methodology to more accurately interpolate between timepoints. This mask was obtained from the World Settlement Footprint product, a 10m resolution binary mask of global human settlement extents, derived from Sentinel-1 radar and Sentinel-2 imagery (https://urban-tep.eu/#!pages/dataservices). This addition marks an improvement to the integrated approach outlined in Nieves *et al.* (2020).

Modelling methods

A top-down method to population disaggregation via a random forest (RF) dasymetric approach was implemented to construct the set of gridded population datasets, broken down by sex and age classes, for 242 countries and territories (Stevens et al., 2007).

A RF algorithm was implemented to generate a gridded population density weighting layer at 3 arc-second resolution (approximately 100m resolution at the equator); this prediction layer was then used to perform dasymetric disaggregation of population counts from administrative units into target grid cells at country level (Stevens et al., 2007). RF is a predictive, non-linear, and non-parametric ensemble learning approach that generates a large set of decision tree models and aggregates their predictions (Breiman, 2001). Decision trees are independently generated by bagging (*i.e.*, by sampling the entire dataset with replacement) (Breiman, 1996), and typically two thirds of samples are used to train the trees (known as the 'bagged' sample). Each node of each decision tree is split according to an iterative method in which, at each node, the optimal splitting method is used (Breiman, 2001). After all regression trees have been constructed, the outputs of all tree predictions are aggregated by calculating either their mode or average, contingent on whether the trees are utilised for classification or regression, to produce a final classification decision (Liaw and Wiener, 2002). The remaining third of unsampled data, known as 'out-of-bag' (OOB), are used to perform the internal cross-validation technique to accurately estimate the prediction error of the RF model (Breiman, 2001). This is achieved by averaging all mean squared errors calculated using the OOB data. The RF approach is robust to overfitting (Breiman, 2001), and its predictive accuracy is not very sensitive to the three parameters to be specified for model fitting (Liaw and Wiener, 2002), explicitly, (i) the number of observations in the terminal nodes of each tree, (ii) the number of trees in the forest, and (iii) the number of covariates to be randomly selected at each node.

The RF-based dasymetric population mapping approach was used to produce gridded population distribution datasets for 242 countries and territories. This approach consists of using a RF algorithm to generate gridded population density estimates that are subsequently used, as a weighting layer, to dasymetrically disaggregate population counts from administrative units into grid cells (Mennis, 2003).

RF model fitting was undertaken by generating 500 trees, and assigning the number of observations in the terminal nodes equal to one. Following RF model fitting, population density was predicted using a reduced selection of covariates. For each target grid cell, the average of all decision tree predictions was designated to the cell as the estimated population density value. Where there were insufficient observations (*i.e.* insufficient administrative unit population counts) to fit a RF model for a given country, an additional country with similar characteristics was selected, and utilised to fit an appropriate RF model for predicting population density at the grid cell level (Gaughan et al., 2014). Subsequently, in both scenarios, dasymetric disaggregation of the administrative unit-based population counts was undertaken using the population density weighting layer (Mennis, 2003), thereby generating two gridded population datasets of estimated number of people per grid cell.

All tasks described above were performed using the popRF package in R (Bondarenko M. et al., 2021). The popRF package functionalises the RF-informed dasymetric population modelling procedure (Stevens et al., 2007) within a single programming language framework, and is publicly available, open source, and environment agnostic (Nieves et al., 2021). This package has been parallelised where possible to achieve efficient prediction and geoprocessing over large extents, supporting functions that have applied utility beyond simply performing disaggregative population modelling (Nieves et al., 2021).

Assumptions and uncertainties

It is important to note that there are a number of limitations, caveats and uncertainties inherent in the modelling approach used to generate these gridded population datasets.

The administrative unit-linked input census and estimate population data represents a global mosaic of data sources across geography and time. The substantial differences between countries in numbers of datasets, administrative unit levels, types of data and quality of data mean that output gridded estimates vary in accuracy between countries and years. Across time, different demographic projection methods were adopted depending on the need to interpolate between timepoints of data, or make use of age/sex-structured datasets, or whether projections into the future were required. For some countries where recent censuses have not been undertaken, or where census data were not available, official estimates, US census bureau subnational projections or UN common operational datasets on population statistics were used, and these bring substantial uncertainties with them.

The administrative unit-level of input population data were not identical across each country. The use of finer administrative units for a given country is known to engender higher accuracy of corresponding gridded population, provided that the input demographic data is accurate. This difference is likely to contribute some variance in the corresponding RF models for each country. This effect is further magnified for countries in which insufficient administrative units were available to fit a country specific RF model. In these circumstances, the country was “paired” with another country or “grouped” with a set of countries during modelling. Some modelling variance is likely due to the influence of the paired or grouped countries.

Two timepoints of census data were gathered and used as inputs for the population modelling process where possible. For timepoints between or beyond these years, population counts are modelled and aligned to the UN World Population Prospects 2024 edition baseline national totals. Therefore, it is expected that the uncertainty will increase for population datasets representing reference years further away from the input population dataset timepoints. This same phenomenon is expected to be exhibited for the built settlement growth model, in which several timepoints of input settlement data are implemented and interpolated between.

For consistency, all datasets were produced using a fixed set of geospatial covariates available on a global basis. Therefore, a limited selection of factors considered to be related to population presence and densities in each country have been considered globally. This represents a trade-off in the production of generalisable models, in which the accuracy of gridded population datasets for some countries could be improved by considering additional, locally specific factors.

Additionally, the official census-based population data may or may not have captured changes caused by rapid onset events responsible for sudden fluctuations of population numbers at the administrative unit level (e.g. forced displacements due to natural disasters or conflict). Likewise, the produced gridded population datasets constructed by these inputs do not account for future rapid onset events or season/intra-annual population mobility between administrative units.

Adjustment to match the UN's 2024 Revision of World Population Prospects were implemented during the demographic modelling stage (UN, 2024). Projection and interpolation of input population sources is harmonised according to these baseline totals. Although the WPP population estimates and projections are underpinned by analyses of historical demographic trends (UN 2024), it is important to note the assumption that they represent the best-available projections at national levels. If there is a need to match national totals to alternative estimates other than those from the WPP 2024 edition, this is possible to do and can be discussed with the WorldPop team.

A default set of geospatial covariates were gathered and processed for RF model fitting and prediction. The choice of these geospatial factors was based on literature reviews of phenomena considered to be highly related with population densities and distributions (Nieves *et al.*, 2017, Dobson *et al.*, 2000, Nagle *et al.*, 2014). The methods presented here therefore make the assumption that the default set of covariates represent a comprehensive set of environmental and physical factors that are associated with population distribution.

Works Cited

Briggs, D. J., Gulliver, J., Fecht, D. & Vienneau, D. M. Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sens. Environ.* **108**, 451–466, https://doi.org/10.1016/j.rse.2006.11.020 (2007).

Bondarenko M., Nieves J.J., Forrest R.S., Andrea E.G., Jochem C., Kerr D., and Sorichetta A. (2021): popRF: Random Forest-informed Population Disaggregation R package, https://www.worldpop.org/sdi/poprf_v1_0_0/ , DOI:10.5258/SOTON/WP00715

Cohen, J. E. & Small, C. Hypsographic demography: the distribution of human population by altitude. *Proc. Natl. Acad. Sci* **95**, 14009–14014, https://doi.org/10.1073/pnas.95.24.14009 (1998).

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C. & Worley, B. A. LandScan: a global population database for estimating populations at risk. *Photogramm. Eng. Rem. S.* **66**, 849–857 (2000).

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P. & Tatem, A. J. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **8**, e55882, https://doi.org/10.1371/journal.pone.0055882 (2013).

Kummu, M., de Moel, H., Ward, P. J. & Varis, O. How close do we live to water? A global analysis of population distance to freshwater bodies. *PloS one* **6**, e20578, https://doi.org/10.1371/journal.pone.0020578 (2011).

Linard, C., Gilbert, M. & Tatem, A. J. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJ* **76**, 525–538, https://doi.org/10.1007/s10708-010-9364-8 (2011).

McDonald, R. I. *et al*. Urban effects, distance, and protected areas in an urbanizing world. *Landsc. Urban Plan.* **93**, 63–75, https://doi.org/10.1016/j.landurbplan.2009.06.002 (2009).

Microsoft. Microsoft Building Footprints. https://github.com/microsoft/GlobalMLBuildingFootprints (2023).

Nagle, N. N., Butterfield, B. P., Leyk, S. & Spielman, S. Dasymetric modelling and uncertainty. *Ann. Assoc. Am. Geogr* **104**, 80–95, https://doi.org/10.1080/00045608.2013.843439 (2014).

Nieves, J. J. et al. Examining the correlates and drivers of human population distributions across low- and middle-income countries. *J. R. Soc. Interface* **14**, 20170401, https://doi.org/10.1098/rsif.2017.0401 (2017).

Nieves, J. J. *et al.* Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night. Computers, *Environment and Urban Systems*, **80**, 101444, https://doi.org/10.1016/j.compenvurbsys.2019.101444 (2020).

Office of the Geographer and Global Issues (OGGI) , U.S. Department of State. Large Scale International Boundaries (LSIB) dataset v11.3.1. (ed U.S. Department of State Office of the Geographer and Global Issues) doi:10.5281/zenodo.7254221 (2024).

Pesaresi M. & Politis P. GHS-BUILT-V R2023A - GHS built-up volume grids derived from joint assessment of Sentinel2, Landsat, and global DEM data, multitemporal (1975-2030). European Commission, Joint Research Centre (JRC). doi:10.2905/AB2F107A-03CD-47A3-85E5-139D8EC63283 (2023).

Pesaresi M. & Politis P. GHS-BUILT-S R2023A - GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030). European Commission, Joint Research Centre (JRC). doi:10.2905/9F06F36F-4B11-47EC-ABB0-4F8B7B1D72EA (2023).

Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS ONE* **10**, e0107042, https://doi.org/10.1371/journal.pone.0107042 (2007).

Schumacher, J. V., Redmond, R. L., Hart, M. M. & Jensen, M. E. Mapping patterns of human use and potential resource conflicts on public lands. *Environ. Monit. Assess.* **64**, 127–137, https://doi.org/10.1007/978-94-011-4343-1_12 (2000).

Stathakis, D. & Baltas, P. Seasonal population estimates based on night-time lights. *Comput Environ. Urban Syst.* **68**, 133–141, https://doi.org/10.1016/j.compenvurbsys.2017.12.001 (2018).

Small, C. & Cohen, J. E. Continental physiography, climate, and the global distribution of human population. *Curr. Anthropol.* **45**, 269–277, https://doi.org/10.1086/382255 (2004).

United Nations, Department of Economic and Social Affairs, Population Division (2024). World Population Prospects 2024: Methodology of the United Nations population estimates and projections. UN DESA/POP/2024/DC/NO. 10, July 2024. https://population.un.org/wpp/Publications/Files/WPP2024_Methodology.pdf

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A. & Hay, S. I. High resolution population maps for low-income nations: combining land cover and census in East Africa. *PloS one* **2**, e1298, https://doi.org/10.1371/journal.pone.0001298 (2007).

Zanaga, D. *et al.* ESA WorldCover 10 m 2021 v200. doi:10.5281/zenodo.7254221 (2022). Sirko, W. *et al.* Continental-scale building detection from high resolution satellite imagery. arXiv:2107.12283 (2021).