

# **SUBNATIONAL POPULATION ESTIMATES COMPONENTS FOR THE SPATIAL DATA REPOSITORY**

Robert M. Leddy, Jr.  
Population Geography Staff  
Population Division  
U.S. Census Bureau

Version 4  
September 21, 2016

This document pertains to the U.S. Census Bureau products that were uploaded onto the Spatial Data Repository (SDR) beginning in 2016, consisting of Excel and Unicode tab-delimited text files containing 2000-2020 population estimates for subnational areas in PEPFAR-partnering countries, and shapefiles of those countries together with metadata. Some fields in the files differ from those in the Census Bureau products uploaded onto the SDR prior to 2016. Descriptions of the fields in the pre-2016 products are given in the Data Dictionaries contained in their Excel files, and the metadata accompanying their shapefiles are contained in Microsoft Word documents. Products loaded prior to 2015 contain population data for 2000 to 2015, and those data are joined to their shapefiles as attribute tables. The Census Bureau's goal is to replace these earlier products as more recent population data for each country become available.

## **Contents in this Document**

1. Subnational Boundary Files with Population Estimates for the Spatial Data Repository
  2. Data in Country Shapefiles and Tables
  3. Population Estimates: 2000-2020
  4. Shapefiles
  5. Tables
  6. Metadata
  7. Standards
  8. Updates, Upgrades, Additions
- Data Dictionary

## **1. Subnational Boundary Files with Population Estimates for the Spatial Data Repository**

The U.S. Census Bureau has uploaded into the Spatial Data Repository subnational 5-year age group population estimates by sex for 2000-2020 for partnering countries in

the President's Emergency Plan for AIDS Relief (PEPFAR) program. There are for each country the following **components**:

- **Shapefiles** (ArcGIS) of the country's subnational administrative divisions with U.S. Federal Geographic Data Committee-compliant **metadata** in the North American Profile of ISO19115 2003 style
- Stand-alone **Excel table** with subnational population estimates for 2000-2020
- **Unicode tab-delimited text file** version of the Excel table that can be joined to the shapefile

Files of the individual countries contain data for first-, second-, and/or third-order administrative divisions based on the availability of census data and shapefile boundaries.

## **2. Data in Country Shapefiles and Tables**

- **Subnational Division Names.** These consist of the names of the country's administrative divisions of the first order (provinces, states), second order (districts, counties), and third order (communes, townships), based on data availability. Names and spellings comply with the U.S. Board on Geographic Names (BGN) standards. When BGN spellings are not available, spellings from the official census are used.
- **Subnational Population Estimates.** These are for the 5-year age groups from ages 0 to 79 and the open-ended groups of 80-years-and-over, all by sex, for the years 2000 to 2020.
- **Subnational Area Codes.** These consist of the U.S. Government Geopolitical Entities, Names, and Codes (GENC) Standards, the U.S. Government Federal Information Processing System (FIPS) codes, and, where available, official statistical area codes from the country's national statistical office. The GENC Standards and the FIPS codes are for countries and first-order administrative divisions only.
- **Comment.** Alternate area names and other relevant information.
- **Linkage: Tables to Shapefiles.** The data from the Unicode text file that can be joined (linked) to the shapefile via the field "GEO\_MATCH" common to both, consisting of a unique concatenation of administrative names, usually without diacritics, for each administrative division.

A **Data Dictionary** describing the fields in the Excel and Unicode text files in more detail is appended at the end of this document. It is also provided as a separate worksheet in each country's Excel file.

### **3. Population Estimates: 2000-2020**

- For each subnational area, population estimates (and projections) for the **5-year age groups by sex** between ages 0 to 79 and for the open-ended age groups of 80 years and older, also by sex, are provided for 2000 to 2020.
- The estimates are based on population data from the country's most recent available **census**.
- Using the census as a base, the 2000-2020 estimates are calculated using the **Iterative Proportional Fitting** method. When population data for the subnational areas are available from the two latest censuses, a **trend method** such as the Shift-Share or Logistic Growth Rate is combined with the Iterative Proportional Fitting method to calculate the estimates based on intercensal trends. The estimates are consistent with the country's **national midyear 2000-2020 5-year age and sex group estimates and projections** in the U.S. Census Bureau's *International Data Base*, available at <http://www.census.gov/population/international/data/idb/informationGateway.php>.
- **Descriptions** of the Iterative Proportional Fitting, Shift-Share, and Logistic Growth Rate methods are provided in the document "Methods for Calculating 5-Year Age Group Population Estimates by Sex for Subnational Areas," available at the following web addresses: (1) from the SDR, [http://spatialdata.dhsprogram.com/references/subntl\\_pop\\_est\\_methods\\_pgs\\_uscb\\_sep16.pdf](http://spatialdata.dhsprogram.com/references/subntl_pop_est_methods_pgs_uscb_sep16.pdf); (2) from the U.S. Census Bureau, <http://www.census.gov/population/international/data/mapping/spatialdata.html>.

### **4. Shapefiles**

- **Shapefiles by Administrative Divisions.** For a country, a separate shapefile is provided for each of the first three orders of administrative divisions for which boundaries and population data are available. The shapefile consists of two fields: "AREA\_NAME" and "GEO\_MATCH." The former contains the subnational area name from the BGN or census, and the latter an identifier that is identical to that in the Unicode text file's "GEO\_MATCH" field, which enables linkage of the text file to the shapefile (Section 2).
- **Shapefile Sources.** The U.S. Census Bureau has drawn upon various open sources for the shapefiles, some of which follow:
  - National Statistical Offices or Mapping Agencies
  - *Global Administrative Areas (GADM)*, available at <http://www.gadm.org>
  - *United Nations DevInfo Digital Map Library*, available at [http://www.devinfo.org/di\\_digital\\_map\\_library.html](http://www.devinfo.org/di_digital_map_library.html)

- *United Nations Second Administrative Level Boundary (SALB) project*, available at <http://www.unsalb.org/> (accessible until 2014)

The Census Bureau replaces the attribute fields in the original shapefiles with the “AREA\_NAME” and “GEO\_MATCH” fields.

- **User Restrictions.** Restrictions apply to how shapefiles from some open sources may be used and are specified in the shapefile originators’ websites. For example, the DevInfo, GADM, and SALB shapefiles cannot be used for commercial purposes. Users would have to consult the shapefile metadata (Section 6) and the websites for more information.
- **Conformity to the U.S. Department of State’s International Boundaries.** The U.S. Census Bureau fitted the shapefiles to the international boundaries from the “Large Scale International Boundary” (LSIB) and “World Vector Shoreline” (WVS) data sets, which are from the U.S. Department of State (DOS), Office of the Geographer and Global Issues. The data sets are available at the DOS website <http://geonode.state.gov/> and for versions from 2011-2014, the DOS Humanitarian Information Unit website <https://hiu.state.gov/data/data.aspx>. The LSIB-WVS boundaries are seamless.
- **Attribute Data.** The Unicode text file that can be joined to the shapefile via the “GEO\_MATCH” field contains the country’s subnational 5-year age and open-ended group population estimates by sex for the years 2000 to 2020, along with subnational area names, codes, and comments described in Section 2. The data in the text file are identical to those in the stand-alone Excel table.
- **Currency and Completeness.** The U.S. Census Bureau seeks to link the subnational population data to the most current and complete shapefiles of those subnational areas obtainable. The ideal shapefile is one containing subnational areas that correspond one-to-one to those reported in a country’s most recent census.
- **Retrofitting.** Some subnational areas in a country shapefile may differ from those reported in the country’s most recent census. Usually this is because the shapefile is of a vintage that predates the formations of new subnational areas; polygons are still in place for preexisting areas. Records with the age and sex group population estimates are added to the tables for linkage to the “older” polygons. Such estimates are usually aggregates of the estimates for the “newer” subnational areas that fit the “older” polygons.
- **Disclaimer.** The U.S. Census Bureau, which maintains the shapefiles in the World Geodetic System (WGS) 1984 Geographic Coordinate System, cannot guarantee the positional accuracy of their boundaries. Other than fitting the subnational boundaries to the LSIB-WVS international boundaries, the Census Bureau has modified the former only in rare instances.

## 5. Tables

- **Comprehensive List of Population Estimates at all Administrative Levels.** Records for a country’s subnational divisions and their 2000-2020 age-group population estimates by sex are contained in a stand-alone table in an Excel file (workbook). The records for the lower-order divisions are nested in those for their next higher-order divisions. The Unicode text file, the file that can actually be joined to the shapefile, is created from the Excel file. Among the records are those that can be directly linked to the country shapefile polygons via the “GEO\_MATCH” field, some being for multiple subnational areas with the aggregates of their population estimates (Sections 2, 4).
- **Section 508 Compliance.** The Excel table is formatted for accessibility in compliance with **Section 508** of the **U.S. Government’s Federal Rehabilitation Act**. The Unicode text file is not formatted for Section 508 compliance.
- **Data Dictionary.** A Data Dictionary describing the fields in the tables is provided as a separate tab in the Excel file. The Data Dictionary is also appended at the end of this document.

## 6. Metadata

- Metadata in the **U.S. Federal Geographic Data Committee-compliant North American Profile of ISO19115 2003** style are included with each shapefile.
- The metadata include information for the following: the **subnational administrative divisions** within the country and supplemental background pertaining to that country; **citations** for the census data, the shapefile, and the U.S. Census Bureau’s *International Data Base*; **spatial characteristics** of the shapefile; any **restrictions** with using the shapefile; **methods** for calculating the subnational 5-year age and sex group population estimates; and **contact information**.

## 7. Standards

As noted previously, the data that the U.S. Census Bureau is providing in the Spatial Data Repository adhere to the following standards:

- **Geographic Area Names.** Spellings from the U.S. Board on Geographic Names (BGN), available on the National Geospatial-Intelligence Agency’s *GEOnet Names Server* website <http://earth-info.nga.mil/gns/html/index.html>, are used for the subnational areas. When a BGN spelling is unavailable, the spelling from the country’s census is used.
- **International Boundaries.** As discussed in Section 4, the international boundaries from the shapefiles were fitted to the boundaries from the U.S.

Department of State’s “Large Scale International Boundary” and “World Vector Shoreline” data sets.

- **Tables.** As noted in Item 5, the Excel tables are formatted in compliance with Section 508 of the U.S. Government’s Federal Rehabilitation Act.
- **Metadata.** The metadata in the shapefiles adhere to the Federal Geographic Data Committee-compliant North American Profile of ISO19115 2003 style (Section 6).

## **8. Updates, Upgrades, Additions**

The U.S. Census Bureau uploads new, expanded, or revised data into the Spatial Data Repository by country on a flow basis. Subnational age and sex group population estimates for a country may be calculated or re-calculated for any of the following reasons:

- A new national census with subnational 5-year age group populations by sex was released.
- The national estimates and projections for a country in the U.S. Census Bureau’s *International Data Base* were revised.
- A country shapefile became available with new subnational boundaries or with boundaries of the divisions of the next lower administrative order that correspond with most recent census data.
- A combination of any of the above.
- The country was added as a partner to the PEPFAR community.

## DATA DICTIONARY

The following are descriptions of the fields in the Excel and Unicode text files.

<u>FIELD NAME</u>	<u>DESCRIPTION</u>
AREA_NAME	Name of the country or subnational administrative division
GEO_MATCH	Matching field for joining the data from the Unicode tab-delimited text file derived from this Excel file to the shapefile polygon, with a unique identifier for each record. This is usually a concatenation of all administrative division names separated by an underscore (“_”), beginning with the country name and ending with the division name at its lowest subnational administrative order. Diacritics are usually not included. For an area that shares a polygon with one or more other areas, the identifier is tagged with "(part)," indicating that the area is part of that polygon.
GEO_LABEL	Concatenation of all administrative division names separated by underscores (“_”), beginning with the country name and ending with the division name for the lowest subnational administrative order (level). Names may include diacritics. The field serves as one for name sorting and ordering.
COUNTRY	Country name
ADM1_NAME	First-order administrative division name (ADM1, e.g., province or state)
ADM2_NAME	Second-order administrative division name (ADM2, e.g., district or county)
ADM3_NAME	Third-order administrative division name (ADM3, e.g., commune or township)
ADM_LEVEL	<p>“0” -- country level            “1” – ADM1 level            “2” -- ADM2 level            “3” -- ADM3 level</p> <p>The code for a polygon representing multiple subnational administrative divisions includes the number of those divisions in parentheses. The code for a polygon containing 3 second-order administrative divisions is “2(3).”</p> <p>The code for a polygon containing divisions at different subnational administrative levels is shown with the levels separated by a forward slash (“/”). For a polygon containing 1 second-order administrative division and 2 third-order administrative divisions, the code is “2/3(2).”</p>
COMMENT	Any notes about a country or subnational division, which may include former names or alternate names or spellings
GENC	U.S. Government Geopolitical Entities, Names, and Codes (GENC) Standard (for country and first-order administrative divisions only)

FIPS	U.S. Government Federal Information Processing System (FIPS) code (for country and first-order administrative divisions only)
NSO_CODE	Official code for a subnational administrative division provided by the country's statistical agency (where available), usually in the national census
BTOTL_YYYY	Total population estimate (both sexes) for the year "YYYY"
B0004_YYYY	Population estimate of all people (both sexes) between ages 0 and 4 for the year "YYYY"
B0509_YYYY	Population estimate of all people (both sexes) between ages 5 and 9 for the year "YYYY"
...	...
B7579_YYYY	Population estimate of all people (both sexes) between ages 75 and 79 for the year "YYYY"
B80PL_YYYY	Population estimate of all people (both sexes) at the ages of 80 and over for the year "YYYY"
MTOTL_YYYY	Total male population estimate for the year "YYYY"
M0004_YYYY	Population estimate of males between ages 0 and 4 for the year "YYYY"
M0509_YYYY	Population estimate of males between ages 5 and 9 for the year "YYYY"
...	...
M7579_YYYY	Population estimate of males between ages 75 and 79 for the year "YYYY"
M80PL_YYYY	Population estimate of males at the ages of 80 and over for the year "YYYY"
FTOTL_YYYY	Total female population estimate for the year "YYYY"
F0004_YYYY	Population estimate of females between ages 0 and 4 for the year "YYYY"
F0509_YYYY	Population estimate of females between ages 5 and 9 for the year "YYYY"
...	...
F7579_YYYY	Population estimate of females between ages 75 and 79 for the year "YYYY"
F80PL_YYYY	Population estimate of females at the ages of 80 and over for the year "YYYY"