United Nations | Department of Economic and Social Affairs

Introduction: *World Population Prospects 2021 Upgrade* - Towards a more open and reproducible WPP

United Nations Expert Group Meeting on Methods for the World Population Prospects 2021 and Beyond

Session 1: Monday 6 April 2020

Roadmap toward a more open and reproducible WPP

- Evolution of previous and current revisions for WPP estimates:
  - Reconstruct internally coherent time series of population and demographic components since 1950
  - 5x5 cohort-component data model + post-facto 1x1 interpolations for selected outputs
  - Incorporate the demographic impact of the HIV/AIDS epidemic
  - More systematic compilation of empirical data since 1950 into internal SQL database (DemoData)
  - Greater interactive internal validation of WPP estimates with country data
  - More systematic documentation of data used/considered and generation of metadata (country “short notes”)
  - Use of probabilistic projection methods for fertility and mortality components in addition to deterministic projection scenarios

Roadmap toward a more open and reproducible WPP

- Challenges and pending issues/requests for WPP estimates:
  - More transparent and open access to all underlying country empirical data – not just the metadata documentation of the data sources and estimation methods used **-> SQL database + front-ends (DemoData and DataPortal)**
  - Greater documentation and explanations of the various methods used to derive demographic estimates for each demographic components and the reconciliation with population estimates **-> WPP method protocol**
  - More replicable set of criteria/decisions used to derive estimates that can be more easily communicated and potentially replicated, or updated as new data become available
  - Better capacity to use annual time series (upon data availability and reliability)
  - Better capacity to use single age data (upon availability and reliability)
  - More efficient capacity to update and revise country estimates as new data become available and with greater resilience to staff turnover and staff shortage

Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER)

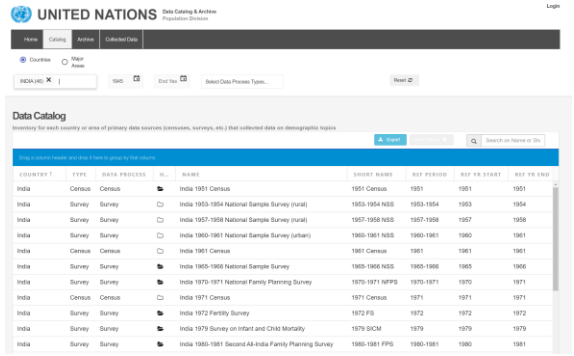
Item #	Checklist item (completed in blue text with hyperlinks, pending in red)	Status
Objectives and funding		
1	Define the indicator(s), populations (including age, sex, and geographic entities), and time period(s) for which estimates were made.	✓
2	List the funding sources for the work.	
Data Inputs		
For all data inputs from multiple sources that are synthesized as part of the study:		
3	Describe how the data were identified and how the data were accessed.	✓
4	Specify the inclusion and exclusion criteria. Identify all ad-hoc exclusions.	2021
5	Provide information on all included data sources and their main characteristics. For each data source used, report reference information or contact name/institution, population represented, data collection method, year(s) of data collection, sex and age range, diagnostic criteria or measurement method, and sample size, as relevant.	✓
6	Identify and describe any categories of input data that have potentially important biases (e.g., based on characteristics listed in item 5).	2021
For data inputs that contribute to the analysis but were not synthesized as part of the study:		
7	Describe and give sources for any other data inputs.	✓
For all data inputs:		
8	**Provide all data inputs in a file format from which data can be efficiently extracted (e.g., a spreadsheet rather than a PDF), including all relevant meta-data listed in item 5. For any data inputs that cannot be shared because of ethical or legal reasons, such as third-party ownership, provide a contact name or the name of the institution that retains the right to the data.**	2021
Data analysis		
9	Provide a conceptual overview of the data analysis method. A diagram may be helpful.	✓
10	**Provide a detailed description of all steps of the analysis, including mathematical formulae. This description should cover, as relevant, data cleaning, data pre-processing, data adjustments and weighting of data sources, and mathematical or statistical model(s).**	2021
11	Describe how candidate models were evaluated and how the final model(s) were selected.	2021
12	Provide the results of an evaluation of model performance, if done, as well as the results of any relevant sensitivity analysis.	2021
13	Describe methods for calculating uncertainty of the estimates. State which sources of uncertainty were, and were not, accounted for in the uncertainty analysis.	
14	State how analytic or statistical source code used to generate estimates can be accessed.	2021
Results and Discussion		
15	Provide published estimates in a file format from which data can be efficiently extracted.	✓
16	Report a quantitative measure of the uncertainty of the estimates (e.g. uncertainty intervals).	
17	Interpret results in light of existing evidence. If updating a previous set of estimates, describe the reasons for changes in estimates.	✓
18	Discuss limitations of the estimates. Include a discussion of any modelling assumptions or data limitations that affect interpretation of the estimates.	

Roadmap toward a more open and reproducible WPP

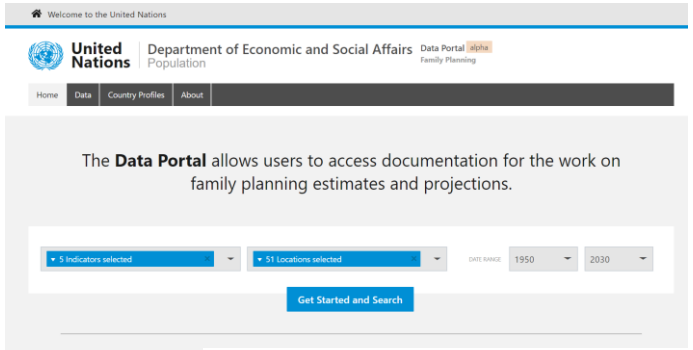
- Long term goals for WPP 2025 and beyond:
  - More open access to both methods, underlying data and analytical steps
  - More collaborative with potential greater country engagement/interactions
- Short term goals for WPP 2021 (planned release May/June 2021)
  - Upgrade production system into 1x1
  - Streamline/harmonize steps used to prepare country data and WPP estimates
  - Provide access to both WPP estimates and underlying empirical data for key demographic indicators -> Data Portal + Demo Data + Data Archive
  - More GATHER compliant
- Constraints and limitations
  - Limited time between upgrade of methods, IT infrastructure and data system
  - Fixed/limited resources (staff, skills, computing, budget, etc.)

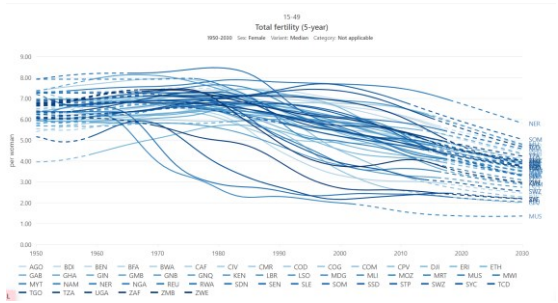
DataCatalog, DataArchive and DemoData

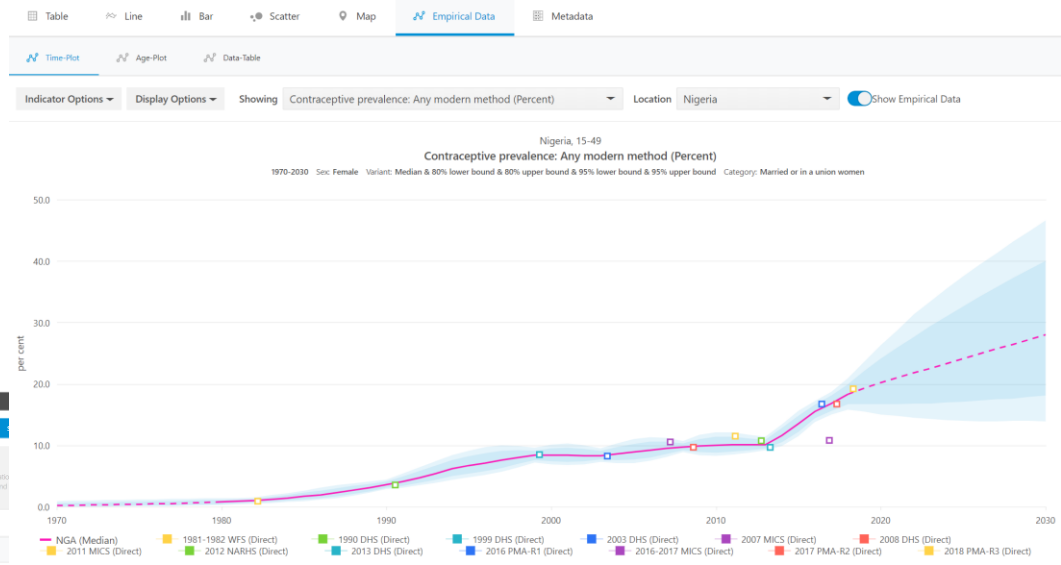
- DataCatalog: comprehensive **inventory for each country of primary data sources** (censuses, demographic surveys, etc.) providing data on demographic processes (fertility, mortality, population structure and dynamics, marital status and family planning) for all countries and areas, as of March 2020, more than 6750 entries
- DataArchive: a **virtual repository** of documents, tabular datasets and reports (potentially) for each data source, as of March 2020, more than 27,000 files
- DemoData: **SQL database to store in a structured and standardized way empirical data and demographic estimates** (with meta-information) on population, fertility, mortality and migration data from as many sources as possible

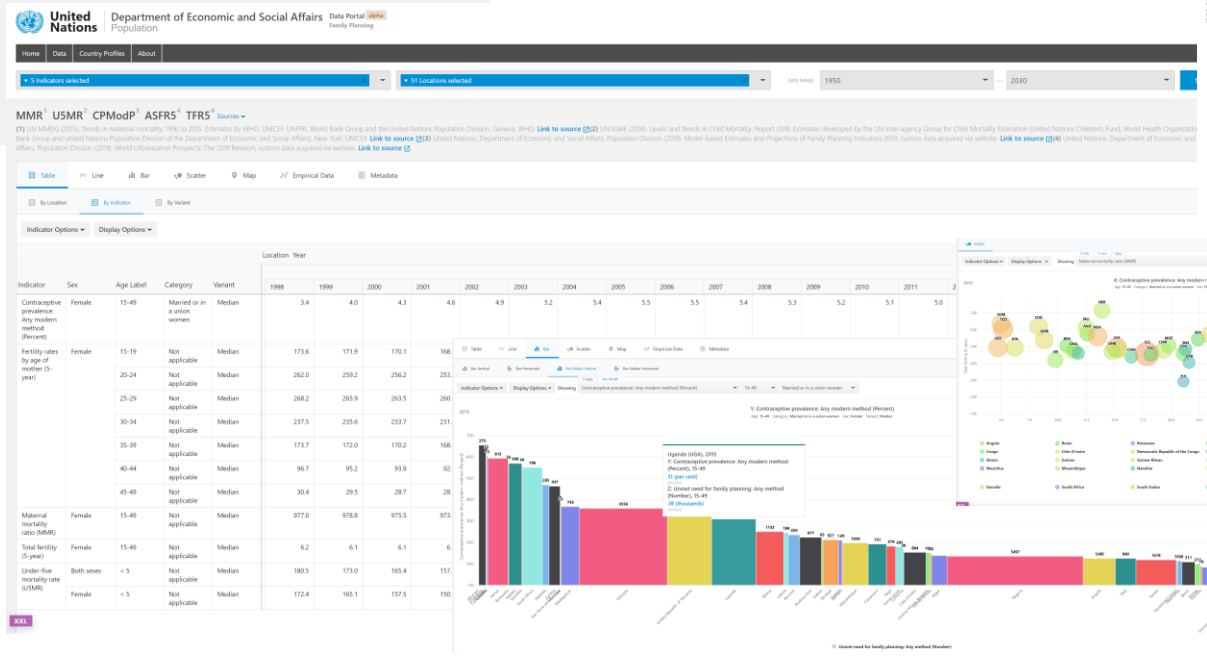


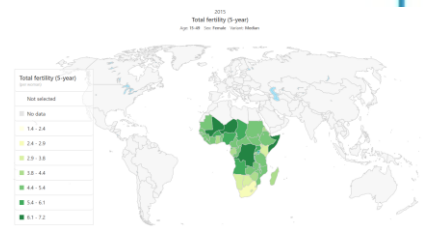
Data Portal: dissemination of estimates/projections and empirical data

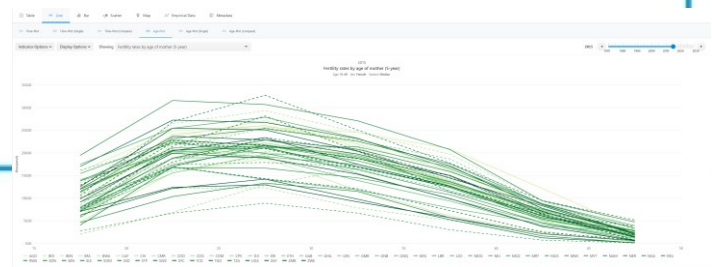




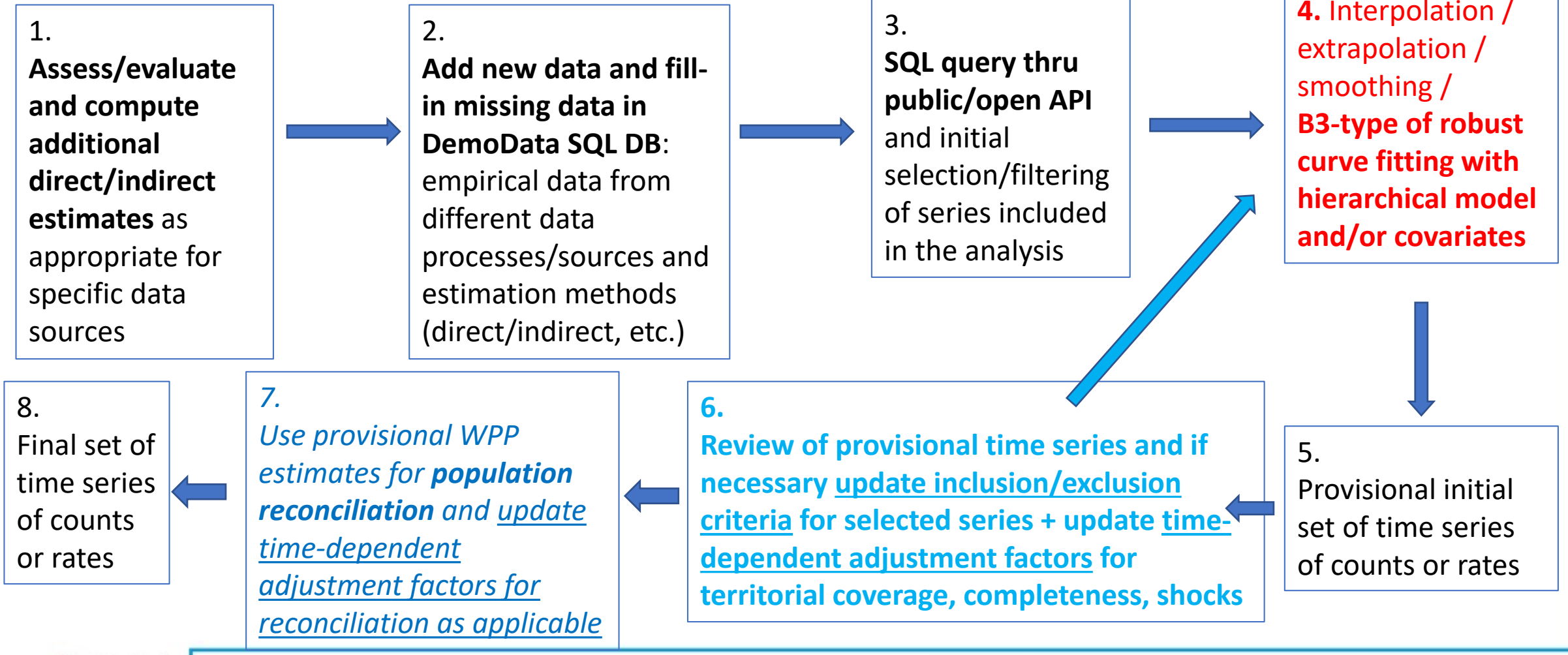




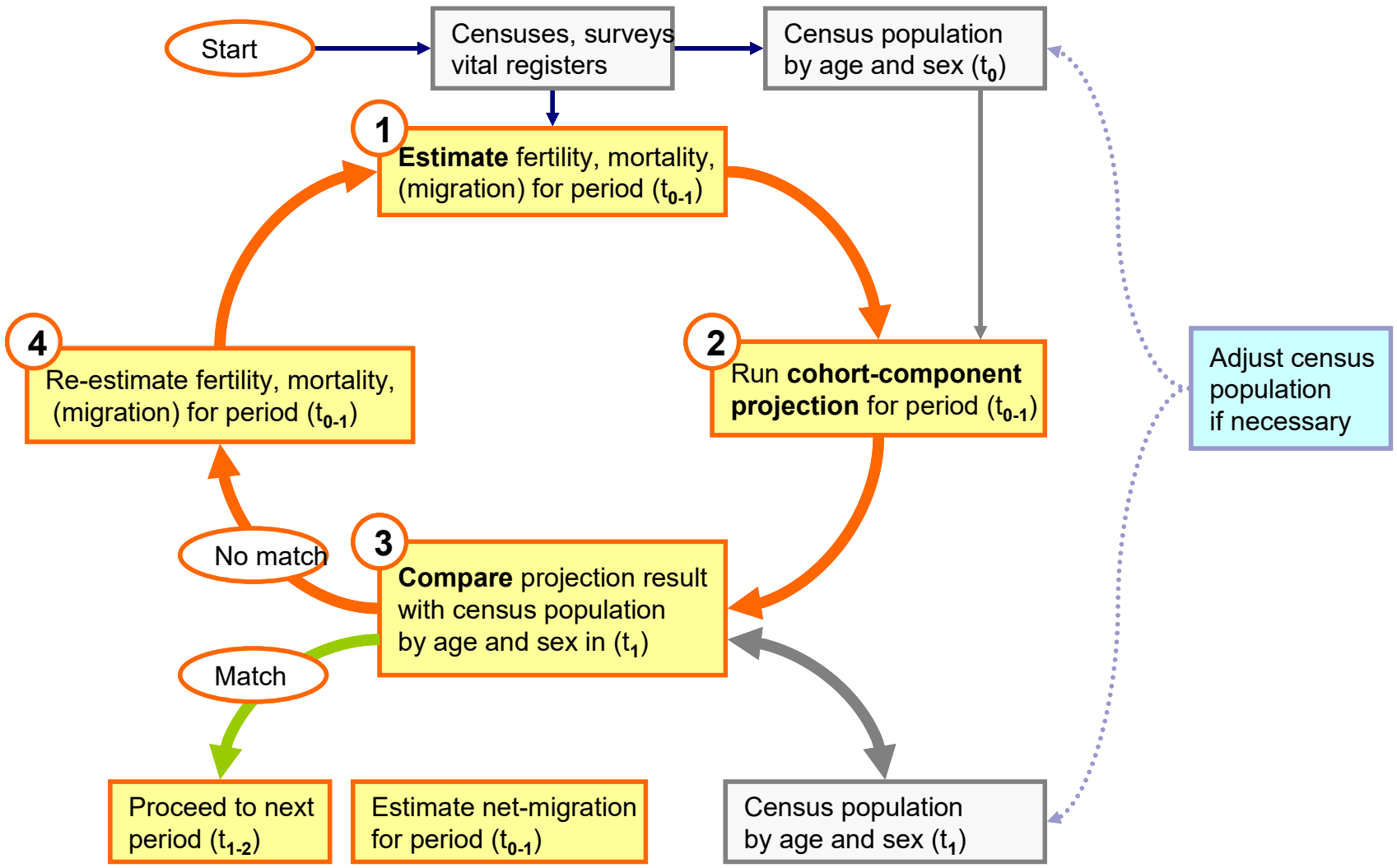




Country data -> provisional WPP estimates -> final … by demographic component for WPP2021



WPP estimation process for each country/area



WPP2021 Workflow

DataArchive

A file repository that contains the PDFs, Excel files, and other file types that contain the raw data (in unstructured format or as published or analyzed) from countries on population size and age structure, as well as fertility, mortality and migration

Relevant tables with metadata from the DataArchive are loaded into DemoData using interactive data entry processes

DemoData

A SQL database with an open API with a web front-end that holds data on total population, population by age and sex, births by sex and by age of mother, deaths by age and sex, TFR and ASFR, life table quantities, and migration flows. These data are accumulated from various sources including the DataArchive and data compilations like IGME estimates, DYB, etc.

EAGLE

Plots

A section of Eagle where estimates from DemoData are displayed next to estimates from the INPUTS and outputs of the Engine.

Tables

A section of Eagle containing tabulated outputs of the Engine.

ShortNotes

A web front-end to a SQL database (linked to DemoData) where analysts indicate for each demographic component which empirical data and estimation/ adjustment methods have been used/considered/excluded in producing estimates for WPP.

Engine

Estimates: 1x1 cohort-component population estimation framework
+ set of parameters (including sets of time series adjustment factors by demographic components)

Projections: Bayesian projection of demographic rates and application to cohort-component
+ set of parameters by country
+ start year for each components

OUTPUTS (data focused)

Aggregates: SQL and R

CountryProfiles

A public display of WPP estimates/projections in tabular form with accompanying visualizations – part of DataPortal

Revision Comparison

A public comparison of major differences by country, and for each demographic component (part of DataPortal)

DataPortal

An interactive data query/browser and visualization (+ public API) for WPP results + empirical data (included/excluded) for key indicators

DataChecking

A set of tools in Excel with SQL queries used to detect errors or inconsistencies in the estimates prior to publication
+ R bulk comparison plots with empirical data

INPUTS

Total counts: From registers and admin. sources for births, deaths, and total pop (census and official estimates) and in selected age groups (infants based on immunizations, primary and secondary educ. stats, national ID and voter registrations)

1x1 time series of age-specific fertility rates

TFR and ASFRs by single year of age of mother from 10-49

VR

For good VR country-years: TFR and ASFR from VR, read from DemoData using inclusion/exclusion criteria from ShortNotes

B3 model

For non VR country-years: Smooth time series of TFR and ASFR. Input data are from DemoData, using inclusion/exclusion criteria from ShortNotes
+ extrapolations

Annual time series of sex ratio at birth

Read/computed from DemoData using inclusion/exclusion criteria from ShortNotes. Then B3 model for smooth time series.

1x1 time series of mortality (life tables)

By sex and for single years of age from 0 to 110+

VR

For good VR country-years: 1x1 life tables (as published complete or graduated from abridged) extended to 110+. These are read from DemoData using inclusion/exclusion criteria from ShortNotes. Smoothed over time/age, adjusted for consistency by sex, as needed.
+ Extrapolations as necessary to 1950.

B3 model

For non VR country-years: Smooth time series of q1, q5, and 35q15, eventually old-age mortality. Input data are from DemoData, using inclusion/exclusion criteria from ShortNotes

Model life tables

SVD comp with output of B3 models of mortality as inputs (as well as UNAIDS time series of HIV prevalence and ART coverage for AIDS countries)

1x1 sparse time series of population censuses

By sex and for single years of age from 0 to 110+. Read from DemoData and adjusted (for underenumeration, age heaping, etc.) according to WPP methods protocol

1x1 Migration

Residual from censuses or estimates from population registers
Smoothed age patterns
+ Model patterns
Informed by refugee flows, migrant flows

estimated from stocks, levels and trends in foreign work permits, population registers, etc.

An open suite of R packages and functions

- New 1x1 cohort-component population projection computational engine designed to work with a standard set of 1x1 inputs and outputs
  - R implementation – for deterministic projections/simulations
  - C implementation – for probabilistic projections/simulations
- [TBD: B3-type of robust time trend modelling for TFR/ASFR, and adult mortality]
- DDSQLTools: set of functions to query DemoData SQL database with API
- DemoTools: set of functions to evaluate, transform and adjust counts or rates
- DDM and FertEstR: set of functions to evaluate and adjust mortality and fertility data
- Ungroup and MortalityLaws: mortality graduation and extension at older ages
- SVDcomp: new expanded set of model life tables (including HIV and ART)
- Calibrated Splines: graduation of fertility age patterns
- [TBD: additional packages/functions to operationalize WPP method protocol]
- popReconstruct: probabilistic demographic estimation and population reconstruction
- BayesTFR, BayesLife/BayesLifeHIV, MortCast, BayesPop: probabilistic projections

Time table

- Feb-August 2020: upgrade of existing IT/data system
- 6-8 April 2020: EGM about WPP2021 and beyond

- Sept.-Dec. 2020: update of country data and WPP estimates using new 1x1 approach
- Jan. 2021: QA data checking and error detections
- Feb. 2021: probabilistic projections
- March 2021: computation of all derived outputs
- April-May 2021: preparation of reports + Data Portal & media launch
- June 2021: public launch

Roadmap toward a more open and reproducible WPP

Strategy = Five year plan: 3 tiers of countries into 3 steps…

Roadmap toward a more open and reproducible WPP

- Following the UN Statistics Division (UNSD) classification for the reliability of population estimates and vital statistics officially reported by national statistical authorities and used since the 1950s for the Demographic Yearbook (DYB)

  "Reliability of data: *Reliable mid-year population estimates are those that are based on a complete census (or a sample survey) and have been adjusted by a continuous population register or on the basis of the calculated balance of births, deaths and migration" (DYB 2017, p. 51)*

- Two key dimensions for the most recent population estimates:
  - Nature of the base measurement of the population (register, census, etc.) and Time elapsed since the last measurement
  - Method and quality of time adjustment by which the base figure was brought up to date (registers, vital registration (VR) + migrations or not, availability of multiple censuses with intercensal period ≤ 10-15 years, etc.)

  Note: both census in last 10 year and VR coverage (100% birth registration, and 80% death registration) are used for SDG indicator 17.19.2

Roadmap toward a more open and reproducible WPP

Based on these UNSD DYB criteria for the reliability of official population estimates for the period 1950-2020, countries/areas can be grouped into 3 tiers:

- Tier 1: countries/areas (n<50) with population registers or regular censuses and VR ≥ 99% since 1950
- Tier 2: countries/areas (n<100) with regular censuses and VR ≥ 60% since 1950 (not included in Tier 1)
- Tier 3: countries/areas (n<90) with no regular censuses and/or no VR or < 60% since 1950

[note the cut/off for VR is only for illustrative purpose]

Roadmap toward a more open and reproducible WPP

WPP revision	Tier 1: Countries with population registers or regular censuses and VR ≥ 99%	Tier 2: Countries with regular censuses and VR ≥ 60%	Tier 3: Countries with no regular censuses and/or no VR or < 60%
2021	Probabilistic estimates	**Deterministic estimates and reconciliation informed by statistical modeling of time trends**	**Deterministic estimates and reconciliation informed by statistical modeling of time trends**
2023	Probabilistic estimates	Probabilistic estimates (conditional on staff/resources availability)	**Deterministic estimates and reconciliation informed by statistical modeling of time trends**
2025	Probabilistic estimates	Probabilistic estimates (conditional on staff/resources availability)	Probabilistic estimates (conditional on staff/resources availability)