

# WIID4 Version Note

December 19, 2018

The World Income Inequality Database (WIID) version 4 collects and stores information on income inequality for developed, developing, and transition countries.

Data on economic inequality (Gini coefficients and distributional shares) are obtained from historical repositories from the World Bank, Asian Development Bank and UNICEF, from the main available current databases, such as the Luxembourg Income Study (LIS), Socio-Economic Database for Latin America and the Caribbean (SEDLAC), Economic Commission for Latin America and the Caribbean (ECLAC), Eurostat, OECD and PovcalNet (World Bank), as well as from national statistical offices and from a large range of independent research papers.

WIID4 retains several elements of the concept of the previous version, WIID3.4, but it also includes many changes, both in the content and structure of the variables. The main changes are stated in this document. Users are kindly referred to the WIID4 User Guide for the complete description of the variables and the format of the database.

## New observations

WIID4 comprises 11,101 observations, whereas WIID3.4 had 8,817. The following table summarizes the total numbers of observations for different time periods:

Time span	Number of observations
Total observations	11,101
Before 1960	306
1960–69	670
1970–79	817
1980–89	1,381
1990–99	2,518
2000–09	3,147
2010–17	2,262

The database has 3,446 unique country-year observations covering 189 countries (including historical entities), up from the 182 in WIID3.4, providing almost worldwide coverage.<sup>1</sup> In several cases the database reports more than one observation for the same country and year. This occurs when these observations add relevant information. For example, they provide consistency with other time series for the same country, or they refer to different resource concepts, area coverage, equivalence scales, among others.

Some observations from the previous version have been deleted, mostly duplicate records with similar information available in the other observations. Most of these cases were observations originating from

---

<sup>1</sup> Exceptions being Libya, North Korea, some Gulf states (Bahrain, Brunei, United Arab Emirates, Kuwait), some microstates (e.g. Monaco, Vatican City, Grenada, San Marino) and certain specific areas (e.g. Western Sahara, French Guiana).

various World Bank sources.

Some cases where the reported values of the Gini coefficient were inconsistent with historical trends have been verified with the source and corrected accordingly.

## List of changes

### Approach to the grouped variables

The approach to the consolidated variables and the original full variables has been changed. Previously, information was provided fully in the original variables (e.g. *AreaCovr*) and then in a consolidated manner in the variables with the suffix *\_new* (e.g. *AreaCovr\_new*). Now, the logic has been altered to work to the other direction instead. In WIID4 the main variable, also by name, is the grouped variable, e.g. *scale*, and then the full is given in the detailed variable, e.g. *scale\_detailed*. This change has been applied as it is convenient for most users to just use the consolidated variables.

Categories for the grouped variables have been considerably reduced. This will allow most users to work immediately with the categories provided without creating their own mapping.

Categories for the detailed variables have been reduced and edited. In some cases this has been done by checking from the source and in some cases removing the duplicate or near-duplicate categories.

For the grouped and detailed variables, it is now much easier to follow which detailed values fall under which consolidated variable categories as the numerical values in the detailed variables are now clearly referring (with the first digit) to under which grouped category they fall into.

### Variable level changes

All variable names have been made lowercase and variable labels have been updated. The order of the variables has also been changed.

An identifier variable *id* has been added to the data.

Country code variables are now named *c3* and *c2* respectively.

The Gini coefficient variable is now named *gini\_reported*, to highlight the fact that it is presented as in the originating source.

Variables for the bottom five and top five percent of income earners have now been named *bottom5* and *top5* respectively, to avoid any confusion that the old convention (P5 and P95) might introduce.

Previous variables for welfare definition are now named as *resource* and *resource\_detailed*.

Previous variables for equivalence scale are now named as *scale* and *scale\_detailed*.

The variable for income sharing unit/statistical unit information is now known as *reference\_unit*.

The variable for the unit of analysis is now known as *sharing\_unit*.

Area coverage variables are now called *areacovr* and *areacovr\_detailed*. The value 'All' is in some cases, strictly speaking, 'Representative all' or 'Almost all', but can be used as nationally representative.

WIID4 does not carry a stand-alone variable for age coverage. It was mostly 'All' for the observations in the previous database version. In cases where it had other values those have now been incorporated into the variable for population coverage.

Population coverage information is now provided in the variables *popcovr* and *popcovr\_detailed*. The latter includes age coverage information in certain cases.

For regional information, in addition to the United Nations geoscheme variables *region\_un* and *region\_un\_sub*, we now also provide the World Bank classification with the variable *region\_wb*.

A new variable *gdp\_ppp\_pc\_usd2011* is introduced. The values for this variable derive from the World Development Indicators by the World Bank. Gross Domestic Product (GDP) is converted to United States Dollars (USD) using purchasing power parity rates and divided by total population. Data are in constant 2011 United States Dollar (USD).

Population variable *population* is now taken from the United Nations authored World Population Prospects instead of the Penn World Tables. For some historical entities we take the values from the French Institute for Demographic Studies (INED), for Kosovo from the World Development Indicators and for West Bank and Palestine from its statistical authority.

In addition to the quality variable, we now provide also a computed variable *quality\_score*, which is a first effort to systematically evaluate the observations. This aims at giving a sense of how much information is provided by each observation, under the understanding that the more information we have about the survey and methodology used to produce the estimates, the better. It also considers how close the estimates are from the standard ones used in the literature. It does not make any consideration, however, about the quality of the survey or the methodology. It is envisioned that together these two quality variables will enable the user to make better informed decisions, if they want to rule out some observations.

Variable *source* has been edited substantially. It is now a general source type indicator with numerical values.

The full information for the source is now given in the variable *source\_detailed*. This variable has been cleaned extensively.

*source\_comments* remains the same, but in many cases information from it has been moved into the new variable *survey*, which contains the name of the originating survey for many observations.