Description of the Lakner-Milanovic (2013) World Panel Income Distribution (LM-WPID) database

The origin and logic of the database. The database has been created specifically for the paper "Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession", World Bank Policy Research Working Paper No. 6719, December 2013, published also in *World Bank Economic Review* (electronically available from 12 August 2015). Its origins, however, go back further in time, to the creation of the PovcalNet database by the World Bank Research Department in the early 1990s[1], and to the 1999 paper on global inequality by Milanovic.[2] The objective of the LM-WPID database was to create both a balanced and unbalanced panel of country-deciles covering the twenty year period 1988-2008, expressed in common currency and prices (2005 Purchasing Power Parity (PPP) dollars derived from the 2005 International Comparison Project). The database thus allows comparisons of average incomes by decile both across time and across countries. In other words, it allows the comparison of (say) the average real income of the 2nd decile of the US income distribution in 1988 with the average real income of the 8th decile of Brazilian income distribution in 2008.

Since income distribution data are not available for all countries annually, we derive global income distributions for five benchmark years at regular five-year intervals, starting with 1988 and ending in 2008. There are three additional adjustments that are made. First, because the data (household surveys) from individual countries are not always available at these precise benchmark years, we use information from household surveys that are within two years of a benchmark year. For example, for the benchmark year 1993, acceptable household surveys would come from all years between 1991 and 1995.[3] Second, we do not allow that a given country's surveys included in two adjacent benchmark years be unreasonably close or unreasonably far apart from each other in time. It would not be helpful if surveys representing a given country in benchmark years 1993 and 1998 came from years 1995 and 1996 (as according to the first adjustment they could). So, we

[1] See http://iresearch.worldbank.org/PovcalNet. The main persons associated with the development of PovcaNet are Shaohua Chen, Martin Ravallion and Qinghua Zhao.
[2] "True World Income Distribution, 1988 and 1993: First Calculations Based on Household Surveys Alone", World Bank Policy Research Working Papers Series No. 2244, November 1999. Later published in *Economic Journal*, vol. 112, January 2002.
[3] If we have several surveys for a given country that fulfil these criteria, we use the one closest to the benchmark year.

complement the first adjustment by the requirement that the two successive surveys used should be at least 3 years, but not more than 7 years, apart. Third, in accordance with the rest of the literature that uses household surveys to study global poverty and inequality, we use income and consumption surveys interchangeably. But in order to limit the bias (which is unavoidable if the survey instrument is not the same for all countries) and avoid any spurious changes over time for a given country, we require that a country be represented throughout all five benchmark years by either consumption or income surveys. In other words, we do not allow that a country be represented by an income survey in one year and a consumption survey in another.

Overall, our database includes 565 household surveys and each country-year observation is represented by the average income of ten income decile groups.[4] Each decile is weighted by its population. We thus measure interpersonal global income inequality where each individual is assigned the income of his or her national income decile.[5] National surveys collect information in terms of local currencies, which we convert into a common currency using first, within-country inflation to correct for changes in the price level between the survey year and 2005, and thus expressing everything in constant 2005 local currency units, and then using the 2005 PPP exchange rates to adjust for cost of living differences across countries.

Household surveys used. The data come from a number of sources. PovcalNet is the starting point of our database, contributing more than two-thirds of the surveys.[6] PovcalNet is the compilation of a large number of household surveys stored by the World Bank Research Department. It has been mostly used to compute estimates of world absolute poverty, and thus lacks data on rich countries. From PovcalNet we obtain average per capita incomes, already converted in 2005 $PPPs, and decile shares, which we combine to compute decile average incomes. Next, we merge it with the World Income Distribution (WYD) data.[7] PovcalNet and WYD provide almost 98 percent of all our data. We fill the remaining gaps with data from the

[4] In the main part of the paper, we treat urban and rural China, India and Indonesia as separate countries (i.e. in the data specify mysample=1). If users wish to use national distributions, they need to set mysample=0, which would results in 550 country-year observations.
[5] Population data are taken from the World Development Indicators (WDI).
[6] Data downloaded on 29 July 2012, which refers to the last PovcalNet update on 28 February 2012.
[7] Available at https://www.gc.cuny.edu/Page-Elements/Academics-Research-Centers-Initiatives/Centers-and-Institutes/Luxembourg-Income-Study-Center/Branko-Milanovic,-Senior-Scholar/Datasets.

Luxembourg Income Study (LIS), the British Household Panel Survey (BHPS), the European Union Survey of Income and Living Conditions (SILC) and data from country statistical offices.[8]

The variables. The database made available here consists of 16 variables. Their names and definitions are given below.

Variable name	Definition/remarks
contcod	Country code, 3 letter country code, separate code for urban and rural surveys
country	Name of country
region	8 geographical/income regions (defined in the paper)
year	Actual year in which survey is conducted (note: it need not be the same as the benchmark year)
inc_con	Binary variable indicating whether survey uses income (I) or consumption (C)
source	Source of household survey
mysample	If mysample==1, separate rural and urban distributions are used for China, India, Indonesia; if mysample==0, national distributions are used for these
bin_year	Benchmark years (1988, 1993, 1998, 2003, 2008)
group	Income decile groups (1 to 10)
pop	Population of country/decile in the benchmark year (WDI*)
RRinc	Average per capita income of country/decile expressed in 2005 PPP **(key variable for the analysis)**
RRmean	Average national per capita income expressed in 2005 PPP
cons_2005ppp_pc	Country's per capita household private consumption in 2005 PPP (WDI*)
gdp_2005ppp_pc	Country's per capita Gross Domestic Product in 2005 PPP (WDI*)
totpop	Country's total population in the benchmark year (WDI*)

*The WDI are updated periodically (including historic data), so these values may not be the most up-to-date.

How to refer to this database? The users should refer to the database as LM-WPID (abbreviation of Lakner-Milanovic (2013) World Panel Income Distribution) database and cite the Lakner-Milanovic paper. For further details about the database and the variables (including the creation of variables not listed here), the user is referred to the Lakner-Milanovic paper.

[8] The sources for the final database (country-years in parentheses) are: PovcalNet (379), WYD (173), LIS (8), SILC (2), and one survey each from BHPS, Statistics Finland, and Statistics Portugal. The BHPS data were accessed via the UK Data Service.