The structure of the PIP data and how the API works

Michail Moatsos

01 Jun 2022

This document explains the structure of the PIP data and how the API can be accessed. A broader coverage of the methodology used by the World Bank to construct PIP can be found at https://worldbank.github.io/PIP-Methodology/index.html.

1 PIP data structure

PIP data provide a key set of identifiers and poverty/inequality related variables from 169 countries around the world, sparsely covering the period 1967 to 2021. Keep in mind that the data come in two flavors. One that includes only years for which there are available household surveys in the countries covered. And one that also includes estimates for some variables when countries do not have a household survey. The former has 2194 entries in total and the latter has 6682. The data in what follows use the dataset that only includes country-years (that is a combination of a particular country in a particular year) with household surveys.

Keep in mind that three countries (India, China, and Indonesia) have three household survey entries at each for which one exists. These three entries correspond to the Rural section of the country, the Urban section of the country and a National one which is a stitching of the two.

Also keep in mind that there are some countries that may have two different household surveys conducted in the same year, for example one being consumption based and the other income based. Therefore a combination of a country and a year do not uniquely identify a household survey in PIP.

There are five broad groups of variables available:

- Time/Space identifiers,
- Poverty and Inequality indicators,
- Household Survey metadata,
- Metadata variables and
- Auxiliary variables.

1.1 Time/Space Identifiers

These include information on region, country and year. Specifically:

- **region_name**:

```
table(HHS$region_name)
```

```
##
##           East Asia and Pacific         Europe and Central Asia
##                             260                             617
## Latin America and the Caribbean    Middle East and North Africa
##                             449                              66
##               Other high Income                      South Asia
##                             526                              63
##              Sub-Saharan Africa
```

```
##                             213
```

- **region_code**:

```
table(HHS$region_code)
```

```
##
## EAP ECA LAC MNA OHI SAS SSA
## 260 617 449  66 526  63 213
```

- **country_name** (only top 20 are shown):

```
head(sort(table(HHS$country_name),decreasing = T),20)
```

```
##
##          Indonesia              China             Poland            Romania
##                 87                 60                 46                 37
##             Brazil         Costa Rica          Argentina            Hungary
##                 36                 34                 32                 32
##      United States           Honduras Russian Federation     United Kingdom
##                 32                 31                 31                 29
##            Germany            Uruguay             Latvia             Panama
##                 28                 28                 27                 27
##           Colombia Dominican Republic        El Salvador          Lithuania
##                 26                 26                 26                 26
```

- **country_code** (closely following ISO3 standard):

```
sort(table(HHS$country_code),decreasing = T)
```

```
##
## IDN CHN POL ROU BRA CRI ARG HUN USA HND RUS GBR DEU URY LVA PAN COL DOM LTU PER
##  87  60  46  37  36  34  32  32  32  31  31  29  28  28  27  27  26  26  26  26
## SLV THA BLR EST GEO ECU ITA PRY SVK SWE UKR BEL BOL MDA SVN ARM AUT BGR ESP IRL
##  26  26  25  25  25  24  24  24  24  24  24  23  23  23  23  22  22  22  22  22
## ISR KGZ LUX NOR DNK FIN FRA IND NLD SRB CZE HRV KAZ TUR GRC PHL CAN MEX MKD CHE
##  22  22  22  22  21  21  21  21  21  21  20  20  20  20  19  19  18  18  18  17
## MNE PRT CYP CHL ISL ALB IRN MLT PAK VEN AUS MYS XKX CIV TWN VNM MNG NIC UGA BGD
##  17  17  16  15  15  14  14  14  13  13  12  12  12  11  11  11  10  10  10   9
## EGY ZMB LKA MDG NGA PSE BLZ GHA JAM JOR MRT NER TUN AZE BFA ETH GIN GTM KOR LAO
##   9   9   8   8   8   8   7   7   7   7   7   7   7   6   6   6   6   6   6   6
## MAR RWA SEN TJK ZAF BWA GNB KEN MLI MWI TZA BDI BEN BIH BTN CMR DJI FJI GMB LSO
##   6   6   6   6   6   5   5   5   5   5   5   4   4   4   4   4   4   4   4   4
## MDV MOZ NAM NPL SLE SWZ SYC TGO UZB AGO CPV DZA FSM HTI JPN LBR MUS STP TCD TLS
##   4   4   4   4   4   4   4   4   4   3   3   3   3   3   3   3   3   3   3   3
## TON WSM YEM ZWE ARE CAF COD COG COM GAB GUY IRQ KIR LCA MMR PNG SDN SLB SSD SYR
##   3   3   3   3   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## TTO VUT LBN MHL NRU SOM SUR TKM TUV
##   2   2   1   1   1   1   1   1   1
```

- **reporting_year**:

```
table(HHS$reporting_year)
```

```
##
## 1967 1969 1971 1974 1975 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987
##    1    1    1    2    2    3    1    5    4   13    2    6   13   14   17   31
## 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003
##   12   21   22   24   38   39   31   46   42   35   53   46   62   47   67   73
```

```
## 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
##   85   89   87   81   87   90   94   87   94   86   90   92   91   83   94   65
## 2020 2021
##   22    3
```

1.2 Poverty and Inequality indicators

Variable descriptions in this section are copied from PovcalNet, and "Statistical Measurement of Income Polarization. A cross-national comparison" by Axel Schmidt (2002).

- **poverty_line**, typically it has the value of 1.9 dollars corrected for purchasing power parity differences across countries.
- **headcount**, % of population living in households with consumption or income per person below the poverty line:
- **poverty_gap**, the mean shortfall of income from the poverty line. The mean is based on the entire population treating the nonpoor as having a shortfall of zero, and the shortfall is expressed as a proportion of the poverty line.
- **poverty_severity**, (poverty gap squared) the mean squared shortfall of income from the poverty line. The mean is based on the entire population treating the nonpoor as having a shortfall of zero, and the shortfall is expressed as a proportion of the poverty line (and then squared).
- **watts**, this is the mean across the population of the proportionate poverty gaps, as measured by the log of the ratio of the poverty line to income, where the mean is formed over the whole population, counting the nonpoor as having a zero poverty gap.
- **mean**, is the average daily household per capita income or consumption expenditure from the survey in 2011 PPP.
- **median**, is the median of daily household per capita income or consumption expenditure from the survey in 2011 PPP.
- **mld**, stands for the mean log deviation. This is an index of inequality, given by the mean across the population of the log of the overall mean divided by individual income.
- **gini**, a measure of inequality between 0 (everyone has the same income) and 100 (richest person has all the income).
- **polarization**, "Polarization deals with building homogeneous clusters that oppose each other. Maximum polarization is reached if half the population is penniless, while the others share the total income equally" (Schmidt, 2002). Increased polarization indicates a disappearing middle class. (Wolfson M. (1994) When inequalities diverge, The American Economic Review, 84, p. 353-358.). The World Bank does not offer an exact definition here.
- **decile1:decile10**, these variables represent the share of total income or consumption accruing to each decile of the population starting from the least well-off (corresponding to decile1) and ending with the most well-off (corresponding to decile10).

1.3 Household Survey metadata

Interestingly survey_year is not exactly the calendar year. For example there are survey years with a value of 2005.23 (Iran) or 2001.31 (Senegal), etc. Actually, about 11.21% of the entries do not exactly correspond to a calendar year.

- **reporting_level**: the coverage of the national level regions for each entry in the PIP database. This variable needs to be seen in combination with variable survey_coverage. survey_coverage reports the coverage of the actual survey, while reporting_level indicates the level at which the data from the actual survey are reported by PIP. For example, the survey_coverage of surveys from China, India,

and Indonesia are national, but in PIP they are offered on the rural, urban and national (aggregated) reporting_level.

```
table(HHS$reporting_level)
```

```
##
## national    rural    urban
##     2023       58      113
```

- **survey_acronym**: the acronym of the HHS used. The 16 most used are:

```
head(sort(table(HHS$survey_acronym),decreasing = T),16)
```

```
##
##   EU-SILC       HBS   SUSENAS      HIES      EHPM        EH       HIS CRHS-CUHS
##       481       257        87        66        47        43        40        36
##     ENAHO       EPH       ECH     EPHPM GSOEP-LIS      PNAD       SES     HICES
##        35        35        34        30        28        27        27        26
```

- **survey_coverage**: the coverage of the national level regions for each HHS.

```
table(HHS$survey_coverage)
```

```
##
## national    rural    urban
##     2135        2       57
```

- **survey_year**: the exact survey year. Do note that they do not always include integer year values. For example, see Solomon Islands has a survey_year value of 2012.79. This relates to the fact that data collection dates started 2012-10 and ended in 2013-11. The exact form of how the two are averaged/weighted is not directly clear from the available information at PIP.

- **welfare_type**: consumption or income based

```
table(HHS$welfare_type)
```

```
##
## consumption      income
##         960        1234
```

- **survey_comparability**: "The comparability metadata database is organized as follows. Each survey point (i.e., a combination of country, year, welfare and data type) has a corresponding value in the comparability column. Within the same country, all the survey points with the same value in the comparability variable are considered comparable or, at least, no substantial reason to break the series was found. The oldest comparable series in each country starts with the value zero (0) in the comparability variable. When comparability is broken, the value changes to one (1) for the year of the break and it goes on until the comparability is broken again in a subsequent year. The process repeats until the most recent surveys point available. In this way, the most recent comparable poverty series per country is such with the highest value in the comparability column." from the September 2019 PovcalNet Update : What's New available at https://openknowledge.worldbank.org/handle/10986/32478.

```
table(HHS$survey_comparability)
```

```
##
##   0   1   2   3   4   5   6
## 521 661 486 312 108  97   9
```

- **comparable_spell**: continuous years that are comparable based on the methodology used in the HHS. The 18 most common spells are shown here:

```
head(sort(table(HHS$comparable_spell),decreasing = T),18)
```

```
##
## 2003 - 2019 2004 - 2019 2003 - 2018 2006 - 2019 1998 - 2020 2002 - 2019
##         199         132          63          56          46          42
## 2002 - 2020 1999 - 2018 2003 - 2020 2004 - 2020 1990 - 2012 1991 - 2018
##          36          35          35          34          30          28
## 1981 - 2011 1991 - 2019 1997 - 2020 1979 - 2018 1998 - 2019 1989 - 2009
##          27          25          24          22          22          21
```

1.4 Metadata variables

- **is_interpolated**: only makes sense in the imputed/interpolated version of the PIP data which are available from API when setting the relevant flag to TRUE (fill_gaps=TRUE). Mind you that in that case one gets less entries with this indicator at FALSE. This is because the interpolated set of data only start in 1981, and some duplicated years are removed (for example, by countries that have both an income and a consumption HHS available in a year).

```
table(HHS2$is_interpolated)
```

```
##
## FALSE  TRUE
##  1809  4873
```

- **distribution_type**

```
table(HHS2$distribution_type)
```

```
##
## aggregate     group   imputed     micro     mixed
##       168       834        17      5366       297
```

- **estimation_type**

```
table(HHS2$estimation_type)
```

```
##
## extrapolation interpolation        survey
##          2652          2221          1809
```

1.5 Auxiliary variables

- **cpi**: chain consumption price index using 2011 as a reference year. It shows the volatility of prices across years.
- **ppp**: PPP exchange rates from the 2011 ICP round. Basically PPP exchange rates are very similar to the usual market exchange rates. We tend to use PPPs when we wish to compare living standards across countries. Their main advantage is that they correct the market exchange rates for non-tradable goods, since the market exchange rates are mainly representative of the tradable goods sector.
- **reporting_pop**: population of the particular country.
- **reporting_gdp**: gross domestic product per capita in PPP terms of the particular country.
- **reporting_pce**: private consumption/expenditure per capita in PPP terms of the particular country.

2 The PIP API

2.1 How the PIP API works?

The way I have collected the data from the PIP is through the API examples that they report on page: https://pip.worldbank.org/api.

The commands I use for getting the regional (aggregate) level data (in R) are as follows, where jjj stands for a particular poverty line values:

res = GET(paste0("https://api.worldbank.org/pip/v1/pip-grp?country=all&year=all&povline=", jjj,"&group_by=wb&format=json"))

Temp <- fromJSON(rawToChar(res$content))

Now Temp contains a snapshot of the entire database on the regional (aggregate) level.

A similar command gets you the data at the country level:

res = GET("https://api.worldbank.org/pip/v1/pip?country=all&year=all&povline=1.9")

And if you also want the interpolated data you can use:

res = GET("https://api.worldbank.org/pip/v1/pip?country=all&year=all&povline=1.9&fill_gaps=true")

All the above commands return the data in json format.

Some additional parameters, not used here, can be found at https://pip.worldbank.org/api.