Exporting the PIP data

Michail Moatsos

09 Jun 2022

This document goes through and explains what each variable of the resulting dataset means (rather than the mechanics of how it was constructed). It includes summary statistics, and simple visualizations that give the gist of the variable(s). Exporting data is done in R using the script ExportPIP.R.

Note: that an initial (country level) export takes place before the main csv file. That initial export is described in PIP_Dataset_Construction.pdf.

1 Export Country Data

Unfortunately there is a ridiculous, yet necessary, number of variable in this dataset (at the moment 180), so the sub-section that explains them below is rather long, it also offers the summary statistics and the simple visualizations on the same spot, as promised above.

Before that, some general remarks about the general procedure are in order. The main thing I am doing is using gpinter to fit the distributions. This helps with estimating (on top of the available estimates from the data) the mean, the decile shares and median, as well as decile averages and thresholds.

Especially the median is estimated using two methods. One is with the use of gpinter, and the other is more directly linked to the data and we simply ask the value of the poverty line when the headcount ratio is at 0.5 (stored in MedianPerDay_est). When MedianPerDay_est is NA it means that there is no headcount for that HHS in the region of 0.495 and 0.505 (which are also used and their average is taken when there is no value at exactly 0.5), as there is a jump in the data as in the case of AUTnational1989income.

Note, that as the World Bank has informed us, for the imputed HHS the median values are not imputed, but instead something like the closest original value is reported instead. Therefore the medians reported from non-original HHS are deleted to avoid confusion.

Note: most variables are estimated only when the underlying distribution exceeds 98% (i.e. headcount ratio>=0.98).

1.1 Explanation and Presentation of Variables

Wherever you see X in the column names below it can take the values of 1, 1_9, 3_2, 5_5, 10, 20, 30, or 40 (where underscore is interpreted as a comma, so 1_9 is 1.9).

Wherever you see a Y in the column names below it can take the values of 40, 50, or 60.

1.1.1 Entity

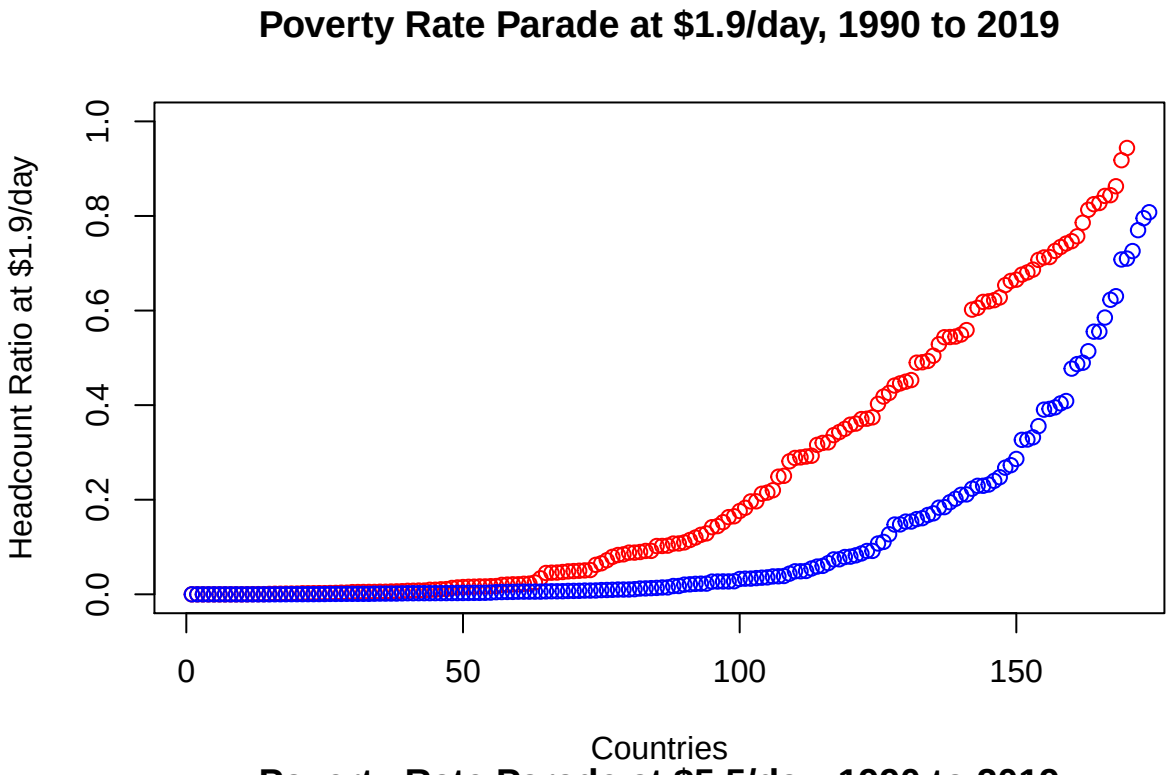
This is the ISO3 code of the country.

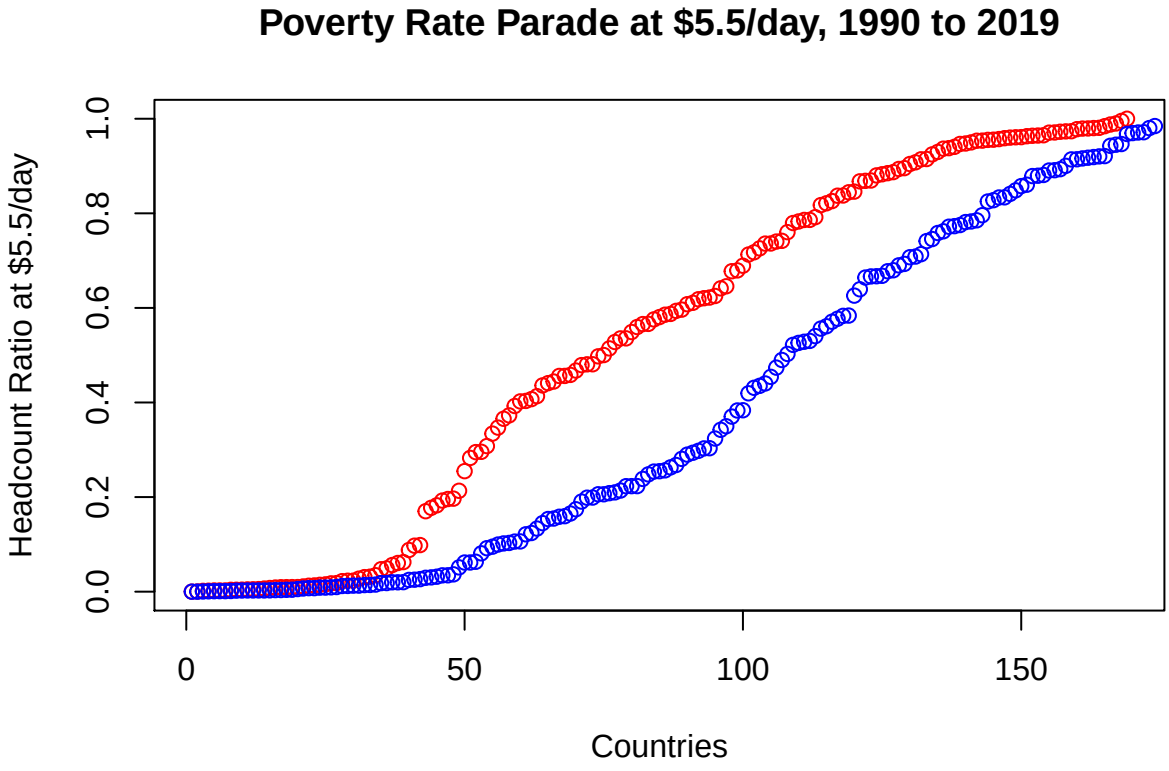
1.1.2 Year

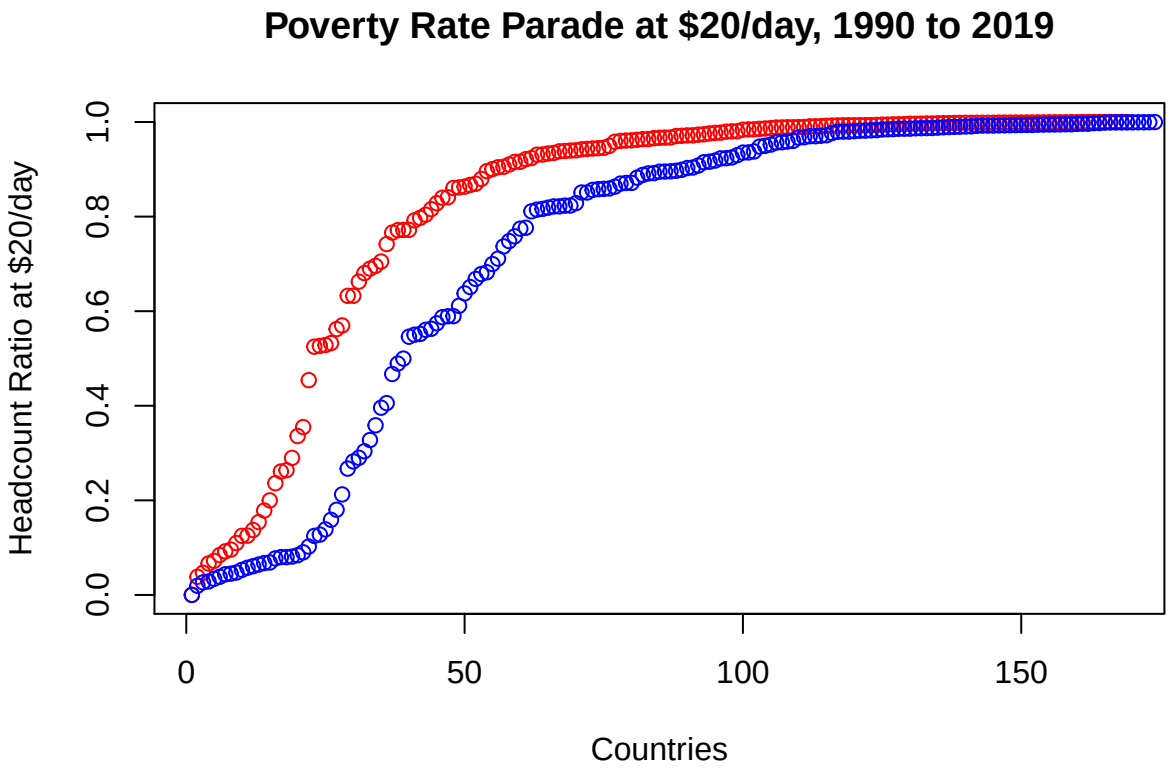
The survey_year from the household survey, as defined in PIP_Structure_and_how_the_API_works.pdf.

1.1.3 headcount_ratio_X_00

% of population living in households with consumption or income per person below the poverty line







1.1.4 poverty_gap_index_X_00

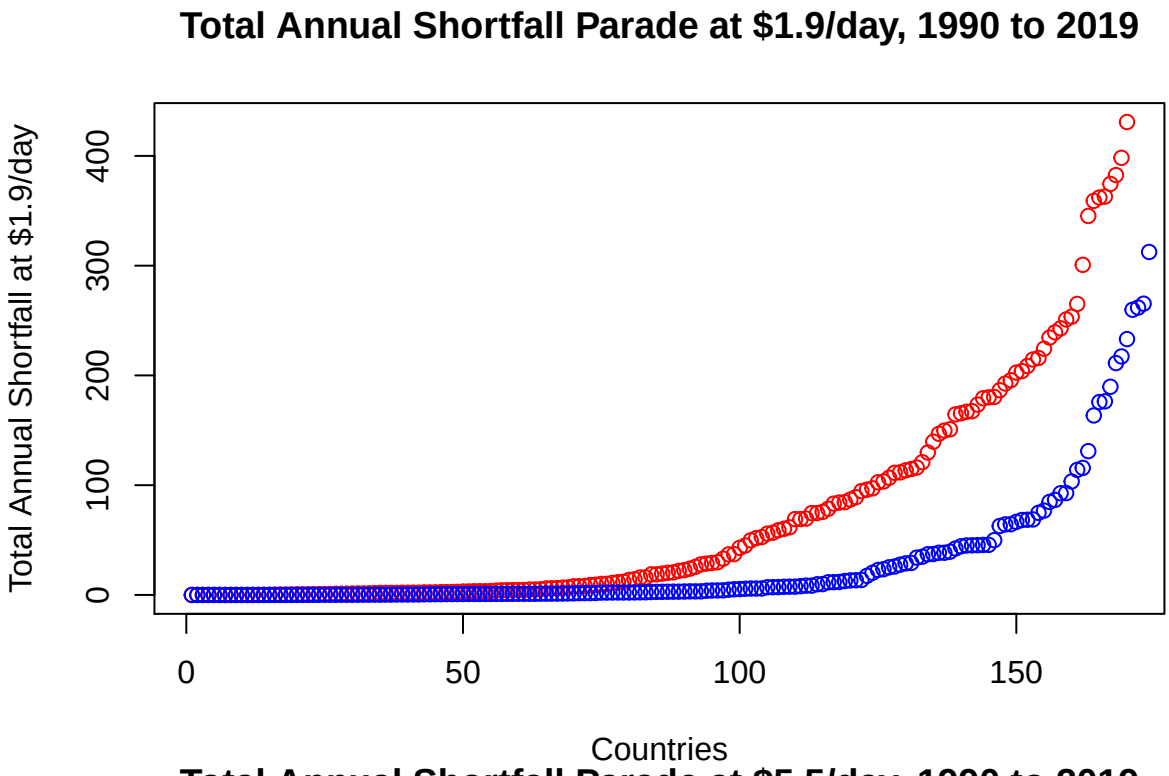
The mean shortfall of income from the poverty line. The mean is based on the entire population treating the nonpoor as having a shortfall of zero, and the shortfall is expressed as a proportion of the poverty line.

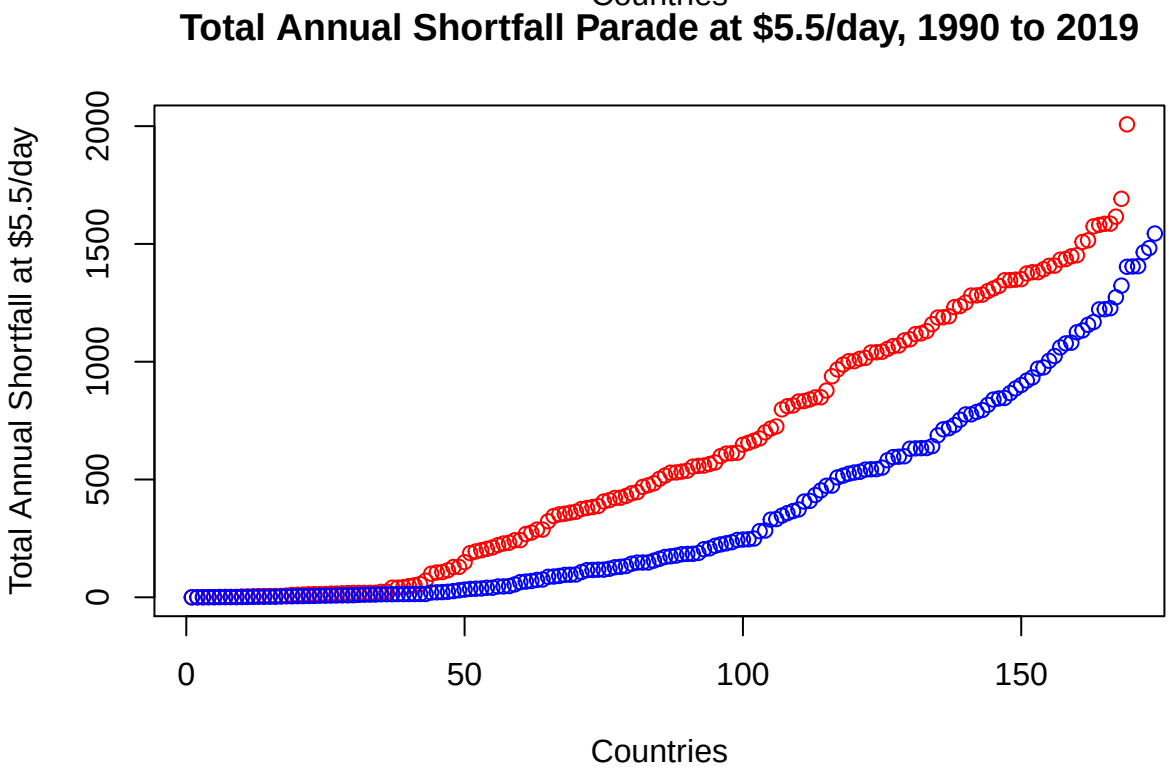
1.1.5 headcount_X_00

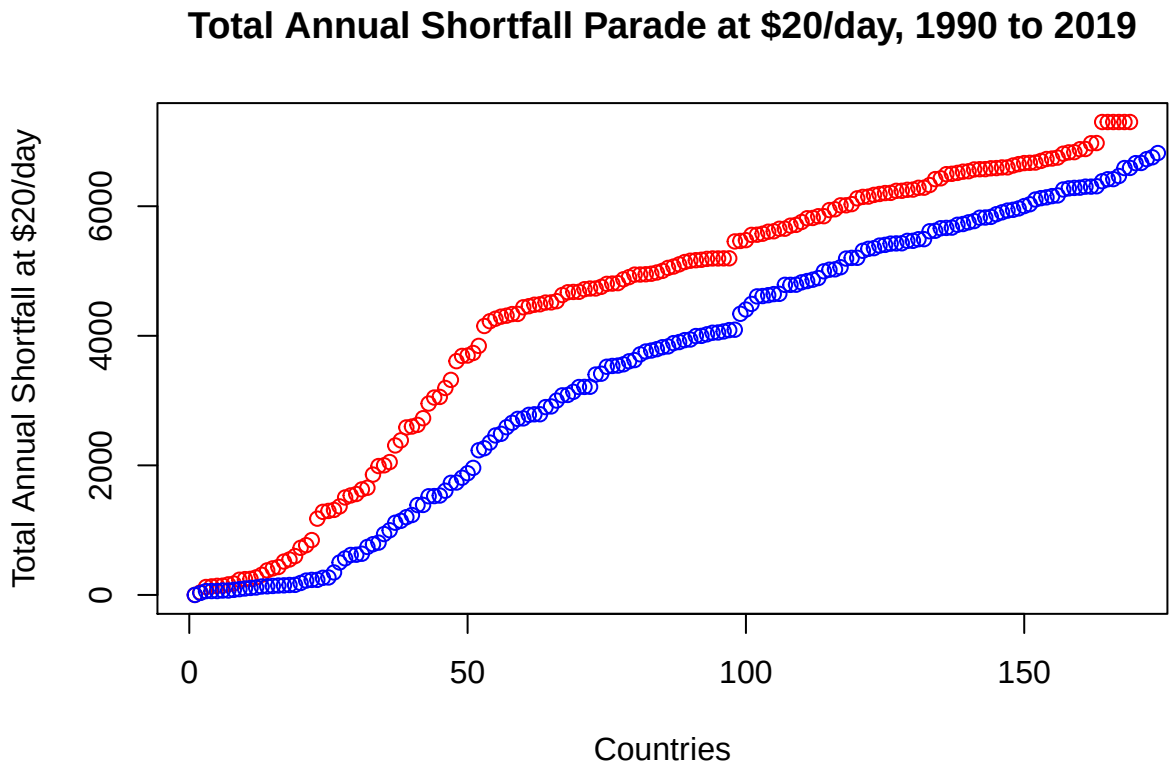
The number of people living in households with consumption or income per person below the poverty line.

1.1.6 total_shortfall_annual_X_00

The amount of money (theoretically) needed to bring everyone up to the poverty line, expressed in annual terms.


Total Annual Shortfall Parade at $1.9/day, 1990 to 2019
Total Annual Shortfall at $1.9/day
0
100
200
300
400
0
50
100
150
Countries


Total Annual Shortfall Parade at $5.5/day, 1990 to 2019
Total Annual Shortfall at $5.5/day
0
500
1000
1500
2000
0
50
100
150
Countries



1.1.7 income_gap_ratio_X_00

"Mean distance below the poverty line as a proportion of the line, among the poor alone.", Ravallion (2016)

1.1.8 watts_index_X_00

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.

1.1.9 headcount_ratio_Y_median/poverty_gap_index_Y_median/headcount_Y_median /total_shortfall_annual_Y_median/income_gap_ratio_Y_median/watts_index_Y_median

As with the corresponding variables above but with the median as a poverty line. The are also available in their "_est" form where the MedianPerDay_est or the MedianPerDay_gp_est are used instead of the original meadian value.

1.1.10 MeanPerDay/MeanPerDay_est

The original mean value of the household survey. / The mean value of the household survey estimated using the gpinter library.

1.1.11 MedianPerDay/MedianPerDay_est/MedianPerDay_gp_est

The original median value of the household survey. / The median value of the household survey estimated from the data by finding where the headcount ratio becomes 0.5 (if the data jump that point, then the average at data points (0.499,0.501), or (0.498,0.502), or (0.497,0.503), or (0.496,0.504), or (0.495,0.505) are used instead; if those are also unavailable, then the value is NA. / The median value of the household survey estimated using the gpinter library.

1.1.12 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.

1.1.13 Population

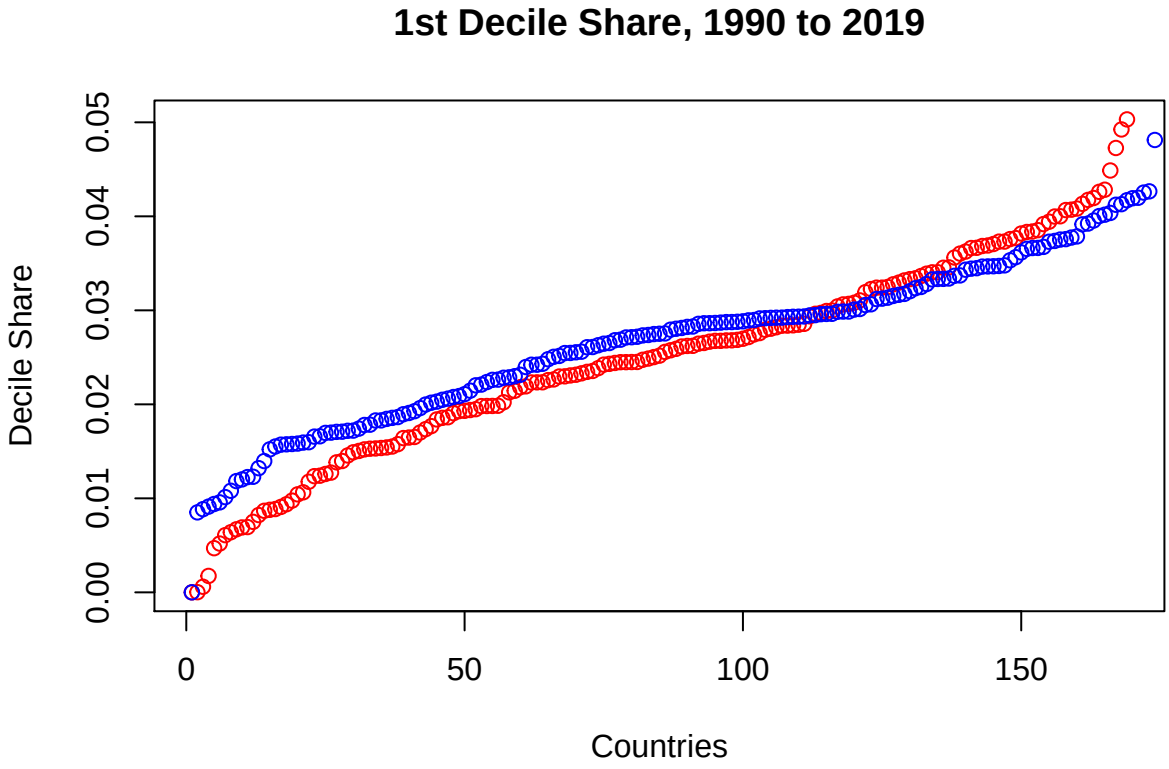
Total population for the corresponding country in the corresponding year.

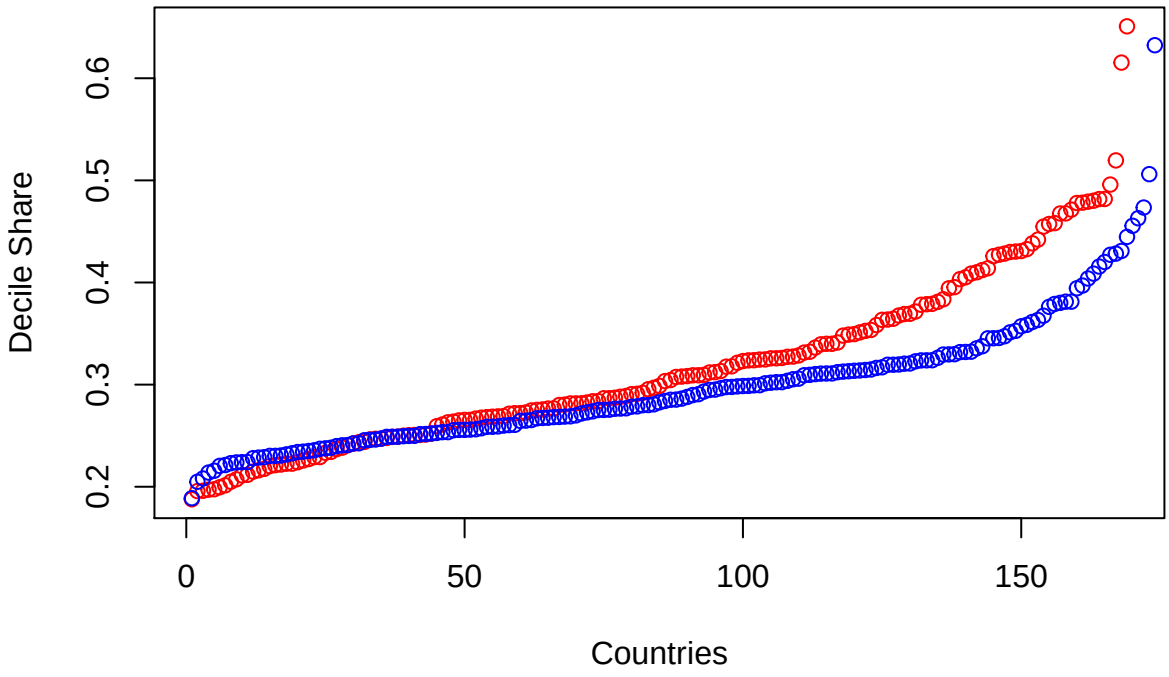
1.1.14 share_decile_1:10

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).

1.1.15 share_decile_1:10_est

As share_decile_1:10 above, but using the gpinter function to get their values.





###

average_decile_1:10_est

Average income/consumption for each of the aforementioned deciles. There are only available in their estimated forms using gpinter, as PIP does not provide this information.

1.1.16 threshold_decile_1:9_est

The lowest value of income/consumption at each of the 9 larger deciles (the first decile always has a threshold of 0). There are only available in their estimated forms using gpinter, as PIP does not provide this information.

1.1.17 Gini/Gini_est

a measure of inequality between 0 (everyone has the same income) and 100 (richest person has all the income).

1.1.18 Polarization/Polarization_est

"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.

1.1.19 MLD/MLD_est

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.

1.1.20 Palma (estimated only)

The Palma ratio is the share of all income received by the 10% people with highest disposable income divided by the share of all income received by the 40% https://data.oecd.org/inequality/income-inequality.htm

1.1.21 P90_P10_ratio (estimated only)

The P90/P10 ratio is the ratio of the upper bound value of the ninth decile (i.e. the 10% of people with highest income) to that of the first. (ibid)

1.1.22 P90_P50_ratio (estimated only)

The P50/P10 ratio is the ratio of median income to the upper bound value of the first decile. (ibid)

1.1.23 Entropy_0_5/Entropy_1_0/Entropy_1_5/Entropy_2_0 (estimated only)

The generalized entropy index has been proposed as a measure of income inequality in a population.[1] It is derived from information theory as a measure of redundancy in data. (wikipedia) This index, as well as the Atkinson and Theil indices below are evaluated at 4 parameter values as indicated by the corresponding variable names.

1.1.24 Atkinson_0_5/Atkinson_1_0/Atkinson_1_5/Atkinson_2_0 (estimated only)

The Atkinson index (also known as the Atkinson measure or Atkinson inequality measure) is a measure of income inequality developed by British economist Anthony Barnes Atkinson. The measure is useful in determining which end of the distribution contributed most to the observed inequality. (wikipedia)

1.1.25 Theil_0_5/Theil_1_0/Theil_1_5/Theil_2_0 (estimated only)

The Theil index is a statistic used to measure economic inequality. The Theil index measures an entropic "distance" the population is away from the "ideal" egalitarian state of everyone having the same income. The numerical result is in terms of negative entropy so that a higher number indicates more order that is further away from the "ideal" of maximum disorder. Formulating the index to represent negative entropy instead of entropy allows it to be a measure of inequality rather than equality. https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/theil-index.html

1.1.26 Var.Coeff (estimated only)

The coefficient of variation is the square root of the variance of the incomes divided by the mean income. It has the advantages of being mathematically tractable and its square is subgroup decomposable, but it is not bounded from above. (wikipedia)

1.1.27 survey_year/is_interpolated/distribution_type/reporting_level/estimation_type/welfare_type/survey_comparability/survey_acronym/reporting_gdp/reporting_pce

See PIP_Structure_and_how_the_API_works.pdf

1.1.28 MaxHeadcountRatio

The maximum atainable headcount ratio. Normally this should be equal to 1 (which means 100%). But there are some distributions like the first actual distribution for GNB and all imputations prior to that year, that only reach a headcount ratio of 0.821 or so.

1.1.29 IsSurveyYear

A flag indicating whether or not the particular entry is from a year with a survey or not

1.1.30 MonotonicityBreaks

The number of entries/rows where the headcount is dropping relative to the previous monotonic slice of the distribution. For example, if a distribution is monotonic (meaning constant or increasing) up to the poverty line of 1, and then at 1.001 it is dropping (very unclear why it happens, as it must not happen, yet it is...)

then until the headcount is back to the level achieved at poverty line equals 1, all points in between count as MonotonicityBreaks. This can happen again at say poverty line 13.5, and the additional breaks are added to the MonotonicityBreaks sum.

1.1.31 MonotonicityBreaksDistinct

See above. Here, MonotonicityBreaksDistinct only count the distinct cases where monotonicity breaks, but not the sum of affected points. The total of affected points can be found in MonotonicityBreaks.

1.1.32 RowsWithIncreasingHeadcount

Indicating the number of rows for a particular HHS where headcount is increasing. This is to be contrasted with the number of rows where the headcount remains the same. Keeping only the rows where headcount is increasing is not loosing any information vis-a-vis a version with all the 14920 rows kept at all requested poverty lines from the API.

1.1.33 DataframeRowsForGpinter

Indicates the number of rows of the HHS data that are given to the gpinter for fitting the distribution. Normally it is around 1000 points, one at each first decimal of a percentile. When this dataset is not fitted by gpinter (for various reasons, better see the comments in the script ExportPIP.R), then I keep only the percentiles above the threshold that makes it difficult for gpinter to fit the distribution. If again gpinter fails, then I feed gpinter with percentiles only, and if that fails too, then I provide only the odd percentiles.

1.1.34 ISO3Year

The combination of the entity's ISO3 code as it appears in PIP and the reporting_year variable.

2 Export Regional Data

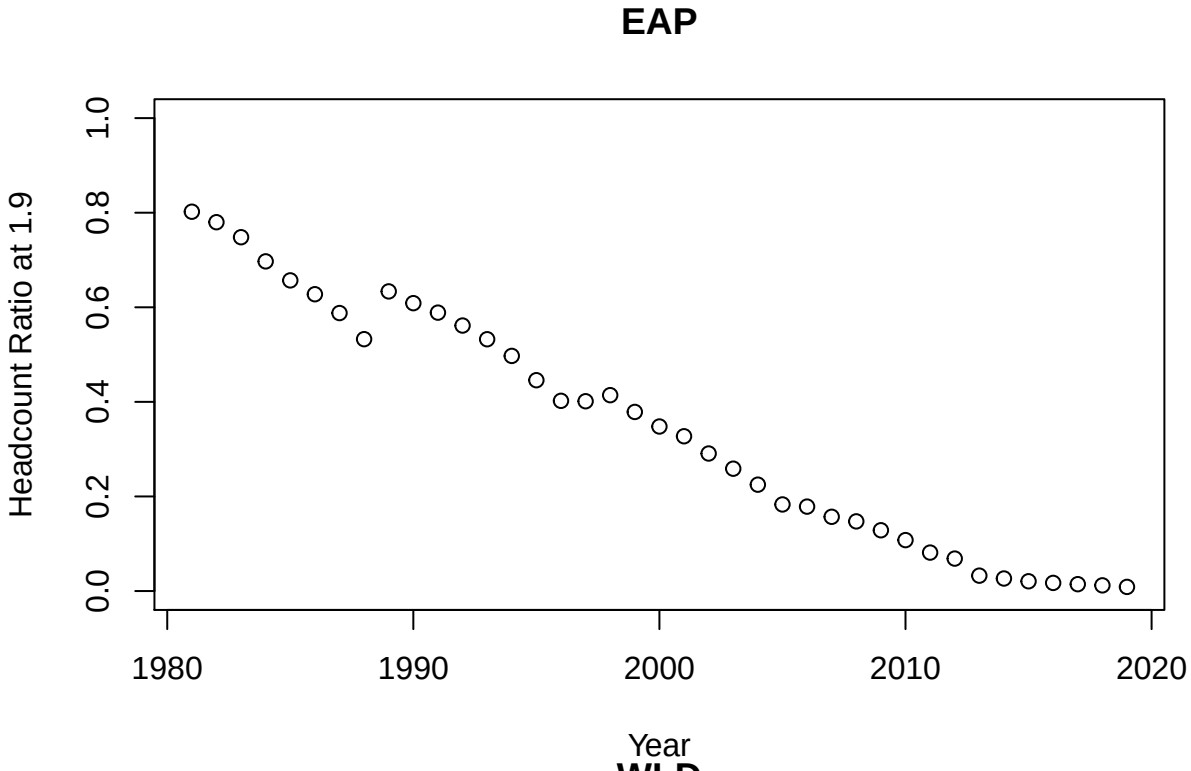
As above but the variables shown below are all NA, because PIP provides no data. The estimated version of those variables (with the exception of polarization) are present as non-NAs in the data.

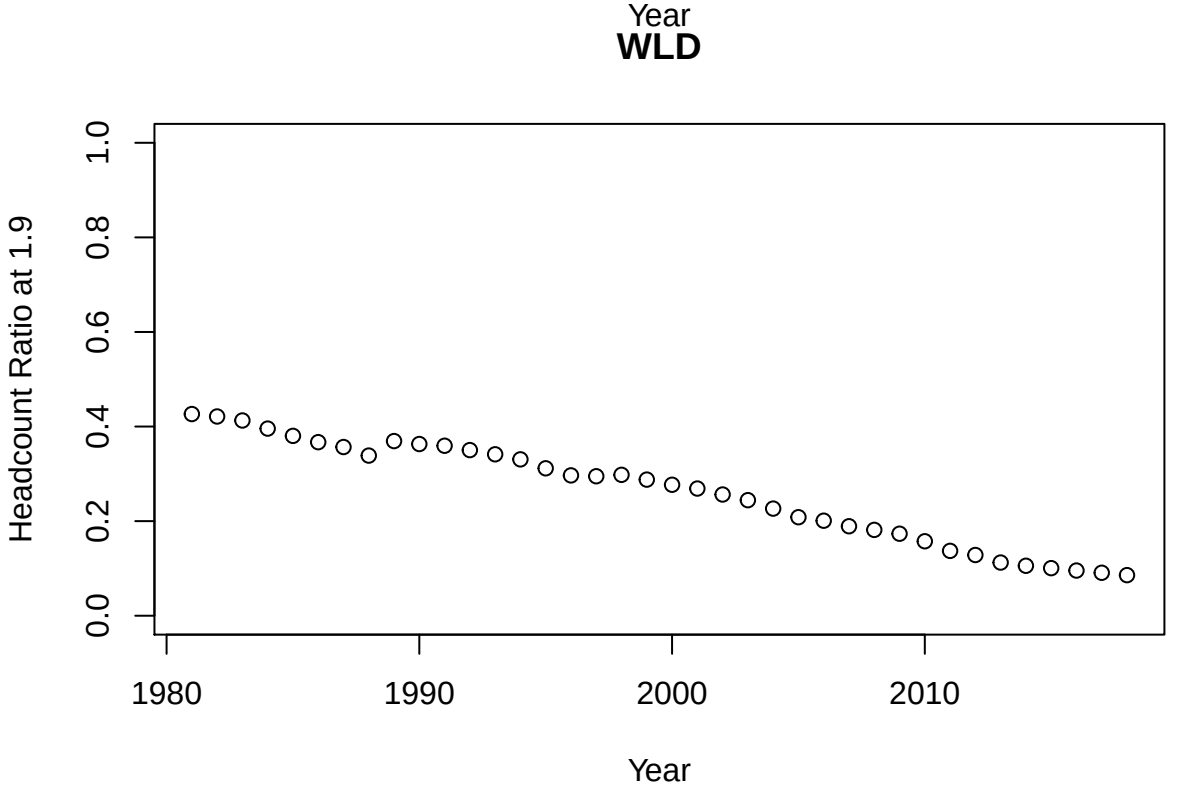
```
##  [1] "headcount_ratio_40_median"        "headcount_ratio_50_median"
##  [3] "headcount_ratio_60_median"        "poverty_gap_index_40_median"
##  [5] "poverty_gap_index_50_median"      "poverty_gap_index_60_median"
##  [7] "headcount_40_median"              "headcount_50_median"
##  [9] "headcount_60_median"              "total_shortfall_annual_40_median"
## [11] "total_shortfall_annual_50_median" "total_shortfall_annual_60_median"
## [13] "income_gap_ratio_40_median"       "income_gap_ratio_50_median"
## [15] "income_gap_ratio_60_median"       "watts_index_40_median"
## [17] "watts_index_50_median"            "watts_index_60_median"
## [19] "MedianPerDay"                     "PPP"
## [21] "share_decile_1"                   "share_decile_2"
## [23] "share_decile_3"                   "share_decile_4"
## [25] "share_decile_5"                   "share_decile_6"
## [27] "share_decile_7"                   "share_decile_8"
## [29] "share_decile_9"                   "share_decile_10"
## [31] "Gini"                             "Polarization"
## [33] "Polarization_est"                 "MLD"
## [35] "survey_year"                      "is_interpolated"
## [37] "distribution_type"                "reporting_level"
## [39] "estimation_type"                  "welfare_type"
## [41] "comparable_spell"                 "survey_comparability"
## [43] "survey_acronym"                   "IsSurveyYear"
```

```
## [45] "reporting_gdp"                    "reporting_pce"
```

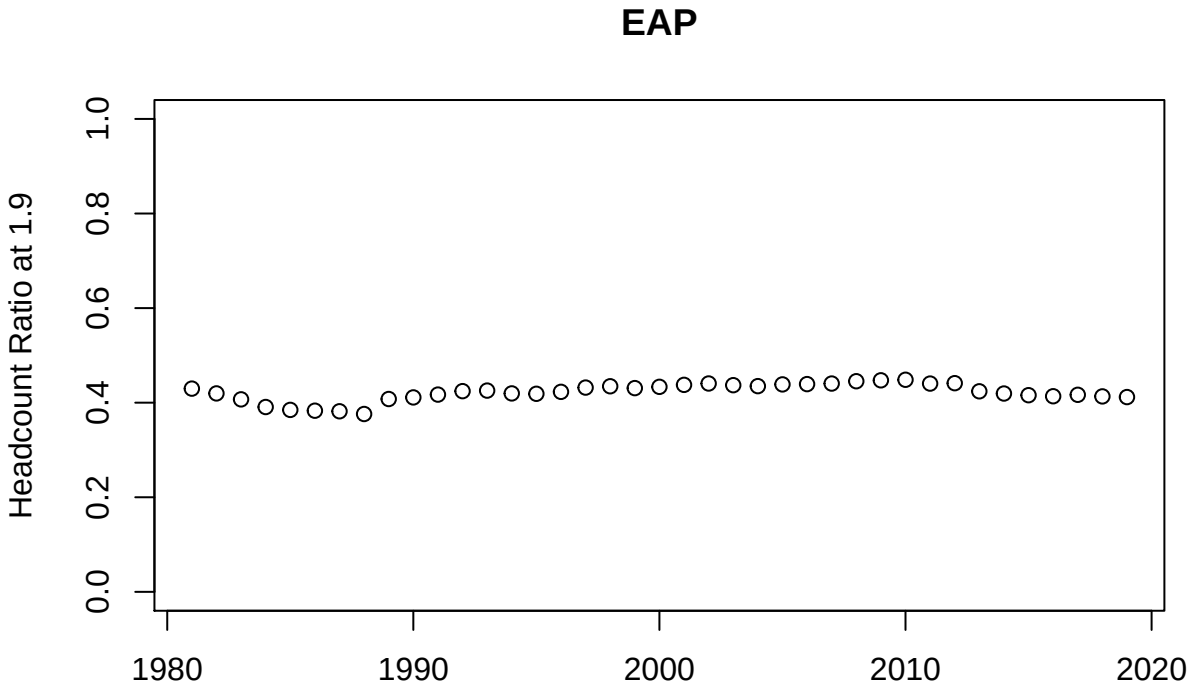
Some plots follow to give you a gist of what is in the data:

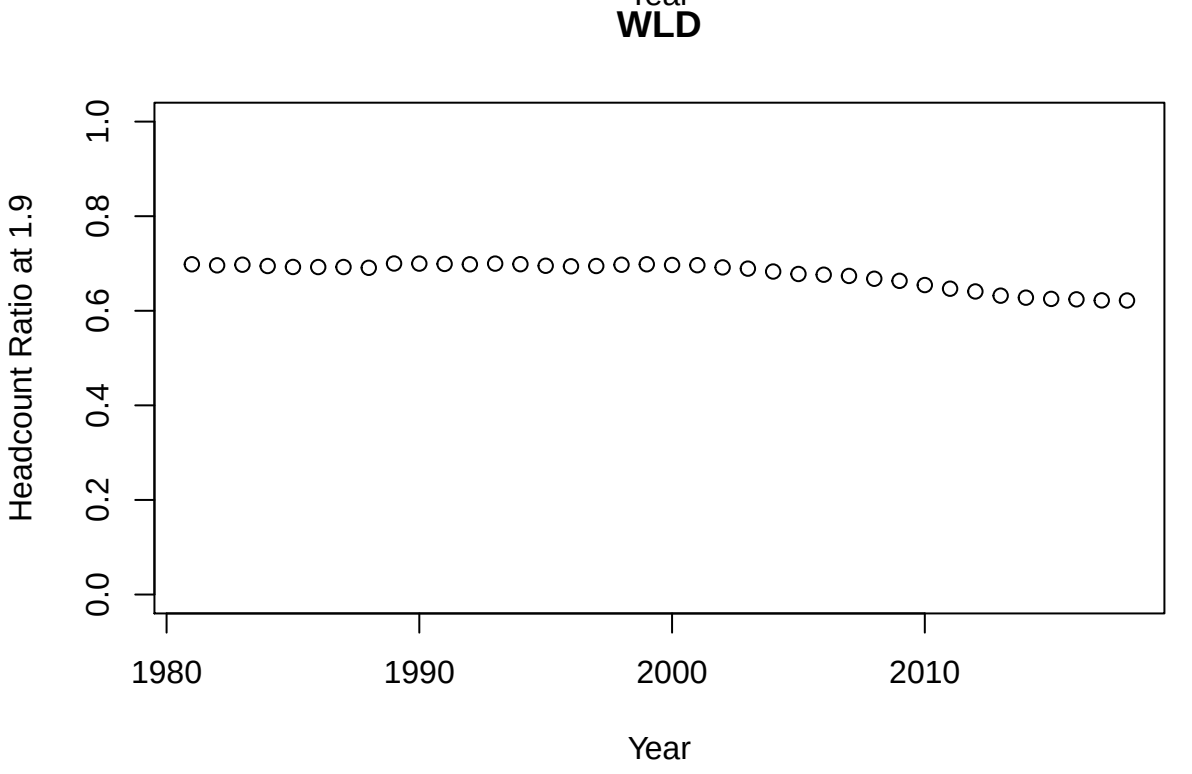
2.1 Regional And Global Headcount Ratios



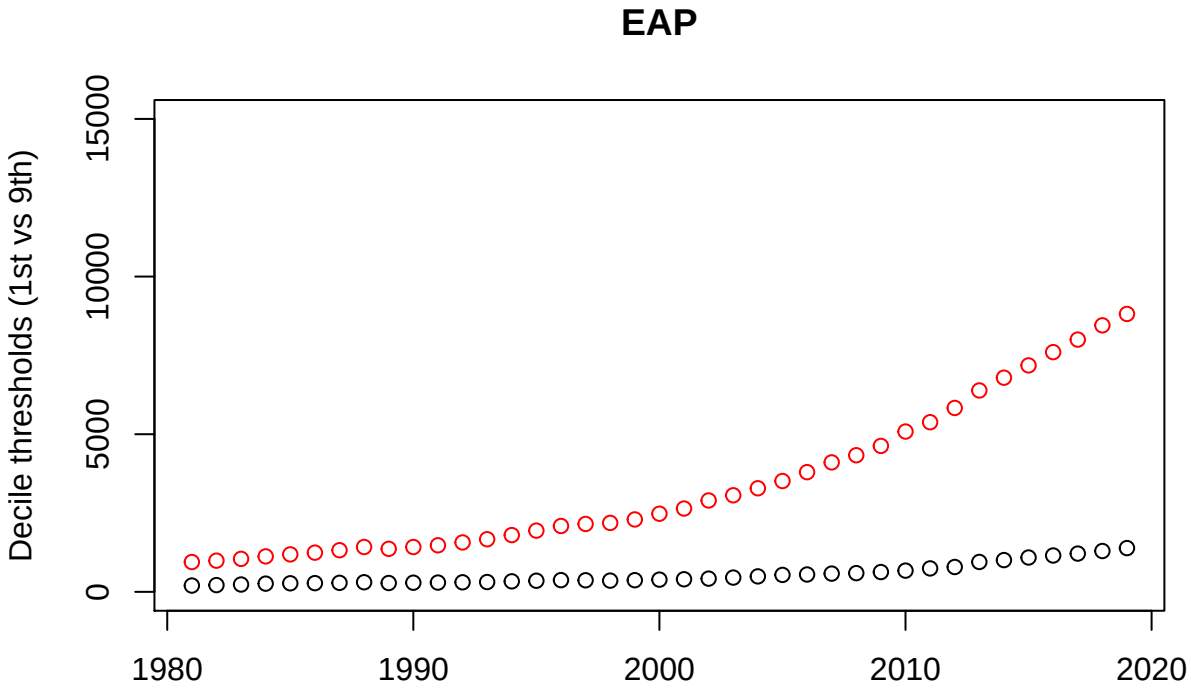


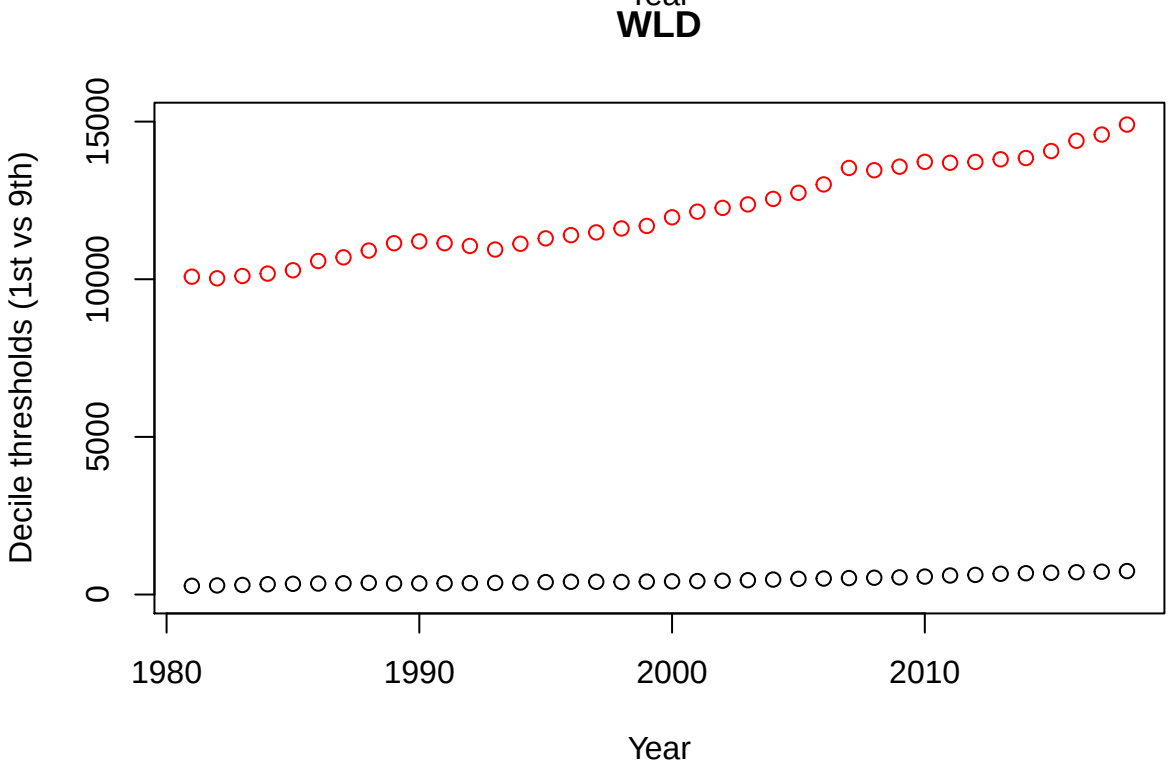
2.2 Regional And Global Gini





2.3 Regional And Global Decile Thresholds





2.4 Regional And Global Total Annual Shortfall